RESEARCH REPORT SERIES
(Statistics #2014-12)

Ranking Populations Based on Sample Survey Data

Tommy Wright
Martin Klein
Jerzy Wieczorek[1]

[1]Currently a PhD student in statistics at Carnegie Mellon University

Center for Statistical Research & Methodology
Research and Methodology Directorate
U.S. Census Bureau
Washington, D.C. 20233

Report Issued: November 14, 2014
Revised and Reissued: January 13, 2015

Ranking Populations Based on Sample Survey Data

Tommy Wright, Martin Klein, and Jerzy Wieczorek

Abstract

Assume K populations with associated respective real-valued parameters $\theta_1, \theta_2, ..., \theta_K$. While the values of $\theta_1, \theta_2, ..., \theta_K$ are unknown, we seek to rank the K populations from smallest to largest based on estimates of these unknown values. If the statistic $\hat{\theta}_k$ is an estimator of θ_k for $k = 1, 2, 3, ..., K$ based on sample survey data, it is a common practice to rank the K populations based on the ranking of the observed values, $\hat{\theta}_1, \hat{\theta}_2, ..., \hat{\theta}_K$, that is,

$$\hat{\theta}_{(1)} \leq \hat{\theta}_{(2)} \leq \cdots \leq \hat{\theta}_{(k)} \leq \cdots \leq \hat{\theta}_{(K)}.$$

For example, the U. S. Census Bureau's American Community Survey (ACS) produced 85 different (explicit) rankings of the $K = 51$ states (actually 50 states and Washington, D.C.) based on observed sample estimates during 2011. One of those rankings ranks the states based on $\hat{\theta}_k$, the estimated mean travel time to work for workers 16 years and over who did not work at home (minutes) for state k, where $k = 1, 2, 3, ..., 51$. Because rankings based on the observed values of the statistics $\hat{\theta}_1, \hat{\theta}_2, ..., \hat{\theta}_K$ can vary depending on the variability among the possible samples that could be observed, some statement of uncertainty should accompany the presentation of each reported ranking. Assuming that a nation's official statistics should be *widely understood* and *robust*, this paper reports concepts and empirical results of some methods for stating uncertainty in rankings using ACS data. Beginning with pair-wise comparisons, we limit our focus to some practices, assisted by visualizations, found in the literature from classical central limit theorem based methods and the bootstrap (nonparametric/parametric). We demonstrate using discussion, some theory, real data, and visualizations that all presented methods (4 methods comparing a pair of populations using normal theory and 3 uncertainty measures and their estimates for the estimated ranks using the bootstrap) are simple and easy to implement and that they can be easily explored and tested, especially by national statistical agencies that release rankings of K populations based on sample survey data. All that is needed are the K sample estimates and their associated standard errors.

KEY WORDS: Bootstrap; Nonparametrics; Official statistics; Uncertainty in rankings.

1. INTRODUCTION

Our main objective is to push a conversation and call national statistical agencies' attention to the need to express uncertainty in rankings based on data from sample surveys. Specially, we assume that a statistical agency has released K estimates $\hat{\theta}_1, ..., \hat{\theta}_K$ and associated standard errors $SE_1, ..., SE_K$ that are based on data from K independent sample surveys in K populations. We share some simple and easy to use methods for presenting uncertainty in rankings of K populations or governmental units - an activity that dates back at least to Aristotle (fourth century B.C. in his *Politeiai*) who did some rankings of 158 Greek city-states in terms of their forms of government (Larsen and Marx, 2012). The desire to rank (either explicitly or implicitly) units such as states based on data from sample surveys is ubiquitous.

Tommy Wright is Chief of the Center for Statistical Research and Methodology (CSRM), U. S. Census Bureau, Washington, D.C., 20233 and adjunct faculty at Georgetown University (E-mail: tommy.wright@census.gov). Martin Klein is Research Mathematical Statistician, CSRM, U. S. Census Bureau, Washington, D.C. 20233 and adjunct faculty at UMBC (E-mail: martin.klein@census.gov). Jerzy Wieczorek, formerly with the U. S. Census Bureau) is a graduate student in statistics at Carnegie Mellon University, Pittsburgh, PA (E-mail: jerzy@cmu.edu). The views expressed are those of the authors and not necessarily those of the U. S. Census Bureau. 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.

More formally, assume K populations with associated independent continuous random variables $Y_1, Y_2, ..., Y_K$ and respective cumulative distribution functions $F_1(y), F_2(y), ..., F_K(y)$. Let θ_k be a real-valued characteristic (parameter) related to $F_k(y)$, for $k = 1, 2, ..., K$. While the values of $\theta_1, \theta_2, ..., \theta_K$ are unknown, it is desired to rank the K populations from smallest to largest based on these unknown values, i.e., based on

$$\theta_{(1)} \leq \theta_{(2)} \leq \cdots \leq \theta_{(k)} \leq \cdots \leq \theta_{(K)}. \tag{1}$$

If $Y_{k1}, Y_{k2}, ..., Y_{ki}, ..., Y_{kn_k}$ is a random sample of size n_k from the k^{th} population where the statistic $\hat{\theta}_k = \hat{\theta}_k(Y_{k1}, Y_{k2}, ..., Y_{kn_k})$ is an estimator of θ_k for $k = 1, 2, 3, ..., K$, it is common practice to rank the K populations based on the observed ranking of the values, $\hat{\theta}_1, \hat{\theta}_2, ..., \hat{\theta}_K$, i.e.,

$$\hat{\theta}_{(1)} \leq \hat{\theta}_{(2)} \leq \cdots \leq \hat{\theta}_{(k)} \leq \cdots \leq \hat{\theta}_{(K)}. \tag{2}$$

For example, the U. S. Census Bureau's American Community Survey (ACS) produces an explicit ranking of the $K = 51$ states (actually 50 states and Washington, D.C.) based on observed sample estimates during 2011 of θ_k the mean travel time to work for workers 16 years and over who did not work at home (minutes) for state k, where $k = 1, 2, 3, ..., 51$. A listing by topic of all 85 rankings from the 2011 ACS is given in Appendix A. Even when estimates are given in a table without an explicit ranking, users will, without exception, compare states looking for smallest, largest, and how states stand relative to each other in terms of their estimates. This occurs all of the time, and we refer to such tables as providing "implicit" rankings.

Because rankings based on the observed values of the statistics $\hat{\theta}_1, \hat{\theta}_2, ..., \hat{\theta}_K$ can vary depending on the variability among the possible samples that could be observed, some statement of uncertainty should accompany the presentation of each reported ranking. While the ACS's sampling design is basically a national stratified random sample with sampling and estimation following a finite population design-based framework, for simplicity, we assume the sample from each state is a random sample, and we will use some ACS data throughout to present some examples of methods for expressing uncertainty in the rankings.

A nation's official statistics should be *widely understood* and *robust*, among many other properties. By widely understood, we mean that the concepts and methods that form the basis for these statistics should be sufficiently simple to be understood by many, especially by those who use the data for making decisions for a nation's people and economy. By robust, we mean that the methods should be valid and applicable in many situations, and that they should be free from very strict and specific assumptions.

These two desired characteristics drive what we present in this paper, and they are shared by classical probability design-based sampling methods (e.g., Cochran, 1977; Fuller, 2009; Lohr, 2010) that are commonly used by national statistical agencies around the world. Models can play a supporting and complementary role, and this is often reflected in model-based sampling (e.g., Valliant, Dorfman, and Royall, 2000) and model-assisted (Särndal, Swensson, and Wretman, 2003) approaches. We also see opportunities in applying methods from classical nonparametrics (e.g., see Hollander and Wolfe, 1999).

Hollander and Wolfe (1999) note, "... a *nonparametric procedure* is a statistical procedure that has certain desirable properties that hold under relatively mild assumptions regarding the underlying population (Y) from which the data are obtained". The related term *distribution-free* refers to statistical procedures where the relevant probability statements are independent of the actual population distributions. In many cases, methods are based on the ranks of the observed sample values instead of the observed sample values themselves. Hollander and Wolfe (1999) present several advantages of nonparametric methods:

- *few assumptions* about the underlying population(s).
- user can often determine things *exactly* (e.g., exact p-values for tests, exact coverage probabilities for confidence intervals, ...).
- often *easier to apply.*
- often *easier to understand.*
- often *only slightly less efficient* than their normal theory competitors when the underlying populations are normal.
- can be *"mildly or wildly" more efficient* than these competitors when the underlying populations are not normal.
- relatively *insensitive to outlying observations.*
- *applicable in many situations* where normal theory procedures cannot be utilized.
- often *valid in many complicated situations* where the distribution theory needed to support *parametric methods is intractable.*

In this paper, we present elementary methods, many nonparametric, using software we have developed with visualizations, which we believe offer some tools for stating uncertainty when releasing rankings to wide audiences. We include and discuss known pair-wise comparison procedures based on normal theory/central limit theorem and the Bootstrap (Efron, 1979; Efron and Gong, 1983). Specifically, we present seven (7) simple and useful methods where we assume a collection of K populations (states) with K independent sample survey estimates and associated estimates of standard error (Wright, Klein, and Wieczorek, 2013). These K estimates and K estimated standard errors form the basis for each of these methods in Sections 2 and 3. Knowledge of the specific complex sampling design and estimation methodology for each population is not required. In Subsection 3.2, where we consider the nonparametric bootstrap, we also assume that we have access to the detailed reported microdata from each sample respondent. In Section 2, we present four (4) methods that focus on visually comparing pairs of states using normal theory presenting uncertainty in the estimated ranking through the use of confidence intervals and hypothesis tests for individual parameters for each population (state) in a ranking, and for the pairwise difference in the parameters for two states: (M1) comparing one reference state with each of the other states; (M2) comparing one reference state with each of the other states showing confidence intervals for differences; (M3) comparing one reference state using its confidence interval with each of the other states using their "comparison intervals"; and (M4) comparing a pair of states by presenting overlapping/non-overlapping confidence intervals appropriately for each state in the pair. In Section 3, we present three (3) uncertainty measures and their estimates for the estimated ranks using the bootstrap (parametric as well as nonparametric) for: (M5) a collection of K confidence intervals for the unknown true ranks; (M6) a collection of K estimates of the probabilities that the estimated rank for a specific state is within c units of the true rank of that state, where c is a positive real number; and (M7) joint probabilities on estimated ranks for all states. Section 3 concludes with some simulation results. In Section 4, we present some concluding remarks.

Future research will consider more from the classical nonparametrics literature, the vast *ranking and selection* literature (e.g., Panchapakesan, 2006), as well as additional parametric methods, including models. Results presented for states extend to populations in general.

Before proceeding, we give an overview of the American Community Survey which is conducted by the U. S. Census Bureau. (http://www.census.gov/acs/www/about_the_survey/american_community_survey/ gives many more details.) “The American Community Survey (ACS) is an ongoing sample survey that provides data every year - giving communities the current information they need to plan investments and services. Information from the sample survey generates data that help determine how more than $400 billion in federal and state funds are distributed each year. (Currently, over 3,500,000 households are contacted each year by Internet, mail, telephone, and face-to-face in a nationwide stratified probability sample survey to provide data for various geographic levels, including national and state, as well as lower levels.) To help communities, state governments, and federal programs, the ACS questionnaire asks about: age, sex, race, family and relationships, income and benefits, health insurance, education, veteran status, disabilities, where you work and how you get there, and where you live and how much you pay for some essentials. All this detail is combined into (statistical estimates) that are used to help decide everything from school lunch programs to new hospitals.”

1.1. An Overview of the Literature

Many of the papers in this overview are highlighted in Frey (2008). While not intended to be comprehensive, our literature review seeks to give a sampling of previous related work. We believe that none directly address our focus to reach wide audiences with uncertainty measurement for rankings based on probability sample data that are released by national statistical agencies.

1.1.1. Classical Ranking and Selection and Nonparametric Methods

Mosteller(1948) presents a nonparametric test for deciding if one of K populations (identical in shape) is stochastically larger than the others. Given K equal size random samples from each of the K populations, one first determines which of the K samples contains the largest observation. The test statistic T rejects the null hypothesis of identical locations for the K populations if T is large, where T is the number of observations in the sample (containing the largest observation) that exceed all of the observations from all other samples.

Paulson(1949) assumes K normal populations with common variance σ^2 and equal size random samples n. If $\hat{\theta}_{max}$ is the largest of the K sample means, each of the remaining populations is classified in a “superior” group if its sample mean is not smaller than $\hat{\theta}_{max} - \frac{\lambda\sigma}{\sqrt{n}}$ where λ is a critical value. Probabilities of two types of error are considered with this procedure.

Paulson(1952a) assumes the same conditions as in Paulson(1949), but lets $\hat{\theta}_1$ be the sample mean of a control population and lets the $K-1$ other populations represent experimental treatments. The population associated with $\hat{\theta}_{max}$ is the “best” population if $\hat{\theta}_{max} - \hat{\theta}_1 > \lambda\sigma\sqrt{\frac{2}{n}}$ where λ is chosen to satisfy certain probability of an error.

Paulson(1952b) extends the problem considered by Mosteller(1948) where each population is normal. His procedure concludes that the population corresponding to $\hat{\theta}_{max}$ is stochastically larger than the others if

$$\frac{n(\hat{\theta}_{max} - \hat{\hat{\theta}})}{\sqrt{\sum_{k=1}^{K}\sum_{i=1}^{n}(y_{ki} - \hat{\hat{\theta}})^2}} > \lambda_\alpha$$

where λ_α is a critical value and $\hat{\hat{\theta}} = \frac{\sum_k \hat{\theta}_k}{K}$. Optimality for this procedure is shown.

In what many consider the major seminal paper in the ranking and selection literature, Bechhofer(1954) presents, among several procedures, a procedure for ranking K populations where the ranking is based on the observed sample means. He considers the probability of a correct ordering when the distance between any two of the ordered true means is at least some positive value Δ.

Specifically, Bechhofer assumes K independent normal random variables Y_k associated with K populations, where $k = 1, 2, ..., K$. The means θ_k are unknown, and the variances σ_k^2 are known and may be equal or unequal. The precise ranking of the means

$$\theta_{(1)} \leq \theta_{(2)} \leq \cdots \leq \theta_{(K)}$$

is unknown. On the basis of random samples of sizes $n_1, n_2, ..., n_K$, the desire is to make inferences on the true ranking based on the observed sample means $\hat{\theta}_1, \hat{\theta}_2, ..., \hat{\theta}_K$.

A *very general goal* of Bechhofer (1954) is to find the s groups of means where we have

> "The K_s (largest means), the K_{s-1} (second largest means), the K_{s-2} (third largest means), etc, and finally the K_1 (smallest means)."

Note that $K_1, K_2, ..., K_{s-2}, K_{s-1}, K_s$ ($s \leq K$) are all positive integers and $\sum_{g=1}^{s} K_g = K$. Thus Bechhofer notes that the probability of a correct ranking associated with the very general goal is given by

$$\begin{aligned} P[&max\{\hat{\theta}_{(1)}, ..., \hat{\theta}_{(K_1)}\} < min\{\hat{\theta}_{(K_1+1)}, ..., \hat{\theta}_{(K_1+K_2)}\}, \\ &max\{\hat{\theta}_{(K_1+1)}, ..., \hat{\theta}_{(K_1+K_2)}\} < min\{\hat{\theta}_{(K_1+K_2+1)}, ..., \hat{\theta}_{(K_1+K_2+K_3)}\}, \\ &\vdots \\ &max\{\hat{\theta}_{(K-K_s-K_{s-1}+1)}, ..., \hat{\theta}_{(K-K_s)}\} < min\{\hat{\theta}_{(K-K_s+1)}, ..., \hat{\theta}_{(K)}\}]. \end{aligned}$$

It is worth noting that if $s = K$ and $K_1 = K_2 = \cdots = K_s = 1$, then the very general goal reduces to the *very specific goal* of finding

$$\theta_{(1)} \leq \theta_{(2)} \leq \cdots \leq \theta_{(K)},$$

and the probability of a correct ranking associated with the very specific goal is

$$P[\hat{\theta}_{(1)} \leq \hat{\theta}_{(2)} \leq \cdots \leq \hat{\theta}_{(K)}].$$

Let $E[\hat{\theta}_{(k)} - \hat{\theta}_{(k-1)}] = \theta_{(k)} - \theta_{(k+1)} = \delta_{k,k+1}$ for $k = 2, ..., K$. Assuming that all sample means $\hat{\theta}_k$ have the same variance, that is

$$\frac{\sigma^2_{(k)}}{n_{(k)}} = \sigma^2_{\hat{\theta}}$$

for $k = 1, 2, ..., K$, then the probability of a correct ranking associated with the very specific goal above becomes

$$P[\theta_{(1)} \leq \theta_{(2)} \leq \cdots \leq \theta_{(K)}] = \frac{\sqrt{K}}{\pi^{\frac{K-1}{2}}} \int_{\frac{-\delta_{K,K-1}}{\sqrt{2}\sigma_{\hat{\theta}}}}^{+\infty} \int_{\frac{-\delta_{K-1,K-2}}{\sqrt{2}\sigma_{\hat{\theta}}}}^{+\infty} \cdots \int_{\frac{-\delta_{2,1}}{\sqrt{2}\sigma_{\hat{\theta}}}}^{+\infty} e^{-\frac{1}{2}z' P_2^{-1} z} dz_1 dz_2 \cdots dz_{K-1}$$

where $P_2 = [\rho_{kk'}]$ is the $K-1$ by $K-1$ correlation matrix with

$$\rho_{kk'} = \begin{cases} 1 & \text{for} \quad k = k' \\ -\frac{1}{2} & \text{for} \quad |k - k'| = 1 \\ 0 & \text{for} \quad |k - k'| > 1 \end{cases}$$

for $k, k' = 1, 2, ..., K-1$ and z' is the row vector $(z_1, z_2, ..., z_{K-1})$.

Assume that the experimenter is able to specify desired values of $\delta_{k,k-1}$, say $\delta^*_{k,k-1}$, which are the smallest values of the $\delta_{k,k-1}$ which are "worth detecting" a difference of this size between the k^{th} largest mean and the $(k-1)^{th}$ mean. If the $\delta^*_{k,k-1}$ values are arbitrarily small, then the probability of a correct ranking given by the multiple integral can be made close to $\dfrac{1}{2}$. If the $\delta^*_{k,k-1}$ values are very large, then the probability of a correct ranking can be made close to 1. It is also clear that very large values of n_k $(k = 1, 2, ..., K)$ will make the probability given by the multiple integral near 1 by $\sigma^2_{\hat{\theta}}$. Thus the problem becomes one of finding "the smallest $n = \displaystyle\sum_{k=1}^{K} n_k$ which will guarantee a specified probability $\gamma < 1$ of a correct ranking whenever $\delta_{k,k-1} \geq \delta^*_{k,k-1}$ for $k = 1, 2, ..., K$. By considering equivalent integrals and some simplifying assumptions, tables are given for some small values of K.

Other special cases of the very general goal include finding the r populations with the largest (or smallest) means θ_k; and procedures for ranking the r populations with the largest (or smallest) means θ_k.

Gupta (1965) presents a procedure that selects a subset of the K populations such that there is a high probability that the population with the highest mean is in the subset. The populations in the subset are determined by those whose sample means exceed $\hat{\theta}_{max} - d$, where d is a critical value.

Gupta and McDonald (1970) present nonparametric procedures for selecting a subset of the K populations such that there is a high probability that the population with the highest mean is in the subset. They combine the K samples and produce one overall ranking from 1 to n $(= \displaystyle\sum_{k=1}^{K} n_k)$. Then for each of the K populations, a mean rank score is computed, that is H_k is the average of the ranks for population k in the combined ranking, for $k = 1, 2, ..., K$. One selection rule consists of selecting all populations for which H_k is within a certain distance of $H \equiv max\{H_1, H_2, ..., H_K\}$;

a second rule consists of selecting all populations for which H_k is at least some stated number c.

McDonald (1973) gives nonparametric procedures similar to Gupta and McDonald (1970) where a subset of the K populations is identified. The data are assumed in n blocks where each block is of size K and contains one sample observation from each of the K populations. Inside each block, the observations are ranked. For population k, T_k is the sum of its ranks across the n blocks for $k = 1, 2, ..., K$. One rule puts population k in the subset of populations if $T_k > T_{max} - d$ for $T_{max} = max\{T_1, T_2, ..., T_K\}$ and d is a critical value. Another rule puts population k in the subset of populations if $T_k > c$, where c is a critical value.

McDonald (1979) applies some of his earlier results to identify subsets of states in the United States that have the highest and lowest traffic fatality rates.

Dudewicz (1980) summarizes the literature and notes that when ranking larger groups of populations, the chance of a correct decision is likely to be close to 0.

1.1.2. Bayesian Ranking Methods

Assuming the usual Bayesian setup of sample data and 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 suggests that ranking on posterior means can lead to "very poor results" (Frey, 2008).

Govindarajulu and Harvey (1974) "... consider Bayesian approaches to the problem of ranking K parameters. They take the joint posterior distribution for the parameters as given, and they focus on ways of moving from that posterior distribution... They point out that simply choosing the ranking with the highest posterior probability may not be an ideal approach, even if it were possible" (Frey, 2008).

Louis (1984) argues that any ranking of populations based on θ_k should consider the collection or ensemble $\{\theta_1, \theta_2, ..., \theta_K\}$ and not the θ_k individually. While not specifically focusing on ranking populations based on θ_k or estimates $\hat{\theta}_k$, Louis (1984) does focus on estimating the collection of parameters. More specifically, the paper focuses on estimating the histogram of the θ_k in a Bayesian setting.

He assumes $\theta_1, \theta_2, ..., \theta_K$ are *iid* $N(\mu, \tau^2)$ as prior distributions. Assume $\hat{\theta}_1, ..., \hat{\theta}_K$ are independent, and $\hat{\theta}_k$ given θ_k is $N(\theta_k, 1)$ for $k = 1, 2, ..., K$. It follows that the posterior distribution of θ_k given $\hat{\theta}_k$ is normal with posterior mean

$$E(\theta_k|\hat{\theta}_k) = \mu + D(\hat{\theta}_k - \mu) = \mu + (\frac{\tau^2}{1+\tau^2})(\hat{\theta}_k - \mu) = (\frac{1}{1+\tau^2})\mu + (\frac{\tau^2}{1+\tau^2})\hat{\theta}_k$$

and posterior variance

$$V(\theta_k|\hat{\theta}_k) = \frac{\tau^2}{1+\tau^2}.$$

Also the posterior distributions of $\theta_1, \theta_2, ..., \theta_K$ are independent.

Now for $\bar{\theta} = \frac{\sum_k \theta_k}{K}$ and $\bar{\hat{\theta}} = \frac{\sum_k \hat{\theta}_k}{K}$, note that we have conditionally on $\hat{\theta}_1, \hat{\theta}_2, ..., \hat{\theta}_K$, the posterior expected sample mean is

$$E[\bar{\theta}|\hat{\theta}_1, \hat{\theta}_2, ..., \hat{\theta}_K] = \mu + \frac{\tau^2}{1+\tau^2}(\bar{\hat{\theta}} - \mu)$$

and the posterior expected sample variance is

$$E\left[\frac{\sum_k (\theta_k - \bar{\theta})^2}{K-1}|\hat{\theta}_1, \hat{\theta}_2, ..., \hat{\theta}_K\right] = (\frac{\tau^2}{1+\tau^2})[1 + (\frac{\tau^2}{1+\tau^2})S^2]$$

where $S^2 = \frac{\sum_k (\hat{\theta}_k - \bar{\hat{\theta}})^2}{K-1}$.

It is worth noting that because the marginal distribution of $\hat{\theta}_k$ is $N(\mu, 1+\tau^2)$, then $\bar{\hat{\theta}} \xrightarrow{P} \mu$, and $S^2 \xrightarrow{P} 1+\tau^2$ as $K \to \infty$. Hence the conditional expectation of $\bar{\theta}$ converges to

$$E(\bar{\theta}|\hat{\theta}_1, \hat{\theta}_2, ..., \hat{\theta}_K) = \mu + \frac{\tau^2}{1+\tau^2}(\bar{\hat{\theta}} - \mu) \xrightarrow{P} \mu,$$

and the conditional variance converges to

$$E\left[\frac{\sum_k (\theta_k - \bar{\theta})^2}{K-1}|\hat{\theta}_1, \hat{\theta}_2, ..., \hat{\theta}_K\right] = \frac{\tau^2}{1+\tau^2}[1+(\frac{\tau^2}{1+\tau^2})S^2] \xrightarrow{P} (\frac{\tau^2}{1+\tau^2})[1+(\frac{\tau^2}{1+\tau^2})(1+\tau^2)] = \frac{\tau^2}{1+\tau^2}[1+\tau^2] = \tau^2$$

Now with the summed squared error loss

$$SSEL = \sum_{k=1}^{K} (\hat{\theta}_k - \theta_k)^2,$$

the posterior means are the standard Bayes estimates

$$E(\theta_k|\hat{\theta}_k) = \hat{\theta}_k^B = \mu + (\frac{\tau^2}{1+\tau^2})(\hat{\theta}_k - \mu).$$

We see that the *mean of the standard Bayes estimates* above is

$$\bar{\hat{\theta}}^B = \frac{1}{K}\sum_{k=1}^{K} E(\theta_k|\hat{\theta}_k) = \frac{1}{K}\sum_{k=1}^{K} \hat{\theta}_k^B$$

which *is equal* to the posterior expected sample mean $E[\bar{\theta}|\hat{\theta}_1, \hat{\theta}_2, ..., \hat{\theta}_K]$. However, the *variance of the standard Bayes estimates* $E(\theta_k|\hat{\theta}_k)$ is

$$\frac{1}{K-1}\sum_{k=1}^{K} (\hat{\theta}_k^B - \bar{\hat{\theta}}^B)^2 = (\frac{\tau^2}{1+\tau^2})^2 S^2$$

which *is not equal* to the posterior expected sample variance $E\left[\dfrac{\sum_k(\theta_k-\bar{\theta})^2}{K-1}|\hat{\theta}_1,\hat{\theta}_2,...,\hat{\theta}_K\right]$.

Thus from

$$\frac{1}{K-1}\sum_{k=1}^{K}(\hat{\theta}_k^B-\bar{\hat{\theta}}^B)^2=(\frac{\tau^2}{1+\tau^2})^2S^2$$

and

$$E\left[\frac{\sum_k(\theta_k-\bar{\theta})^2}{K-1}|\hat{\theta}_1,\hat{\theta}_2,...,\hat{\theta}_K\right]=(\frac{\tau^2}{1+\tau^2})[1+(\frac{\tau^2}{1+\tau^2})S^2],$$

we see that the standard Bayes estimates *shrink* too far toward the prior mean because $0\le\dfrac{\tau^2}{1+\tau^2}\le 1$.

Thus a modified estimator is considered and given by

$$\hat{\theta}_k^L=\zeta+A(\hat{\theta}_k-\zeta)$$

where $A=D^{\frac{1}{2}}\left[\dfrac{1+S^2D}{S^2}\right]^{\frac{1}{2}}$, $\zeta=\dfrac{(1-D)\mu+\bar{\hat{\theta}}(D-A)}{1-A}$, and $D=\dfrac{\tau^2}{1+\tau^2}$.

Thus we see that the *mean of the estimates* $\hat{\theta}_k^L$ is

$$\bar{\hat{\theta}}=\mu+(\frac{\tau^2}{1+\tau^2})(\bar{\hat{\theta}}-\mu)$$

which *is equal* to the posterior expected sample mean $E[\bar{\theta}|\hat{\theta}_1,\hat{\theta}_2,...,\hat{\theta}_K]$. Also, the *variance of the estimates* $\hat{\theta}_k^L$ is

$$\frac{1}{K-1}\sum_{k=1}^{K}(\hat{\theta}_k^L-\bar{\hat{\theta}}^L)^2=(\frac{\tau^2}{1+\tau^2})[1+(\frac{\tau^2}{1+\tau^2})S^2]$$

which *is equal* to the posterior expected sample variance $E\left[\dfrac{\sum_k(\theta_k-\bar{\theta})^2}{K-1}|\hat{\theta}_1,\hat{\theta}_2,...,\hat{\theta}_K\right]$.

Note that A converges to

$$A=D^{\frac{1}{2}}\left[\frac{1+S^2D}{S^2}\right]^{\frac{1}{2}}\xrightarrow{P}D^{\frac{1}{2}}\left[\frac{1+(1+\tau^2)D}{(1+\tau^2)}\right]^{\frac{1}{2}}=D^{\frac{1}{2}}[\frac{1+\tau^2}{1+\tau^2}]^{\frac{1}{2}}=D^{\frac{1}{2}}$$

.

In practice, replacing ζ by μ and A by $D^{\frac{1}{2}}$ in the modified estimator $\hat{\theta}_k^L$ gives a "histogram" of estimates $\hat{\theta}_k^L$ that is very good for estimating the "histogram" of the parameters θ_k.

Berger and Deely (1988) present a Bayesian approach to ranking K normal means with noninformative priors and base inferences on the joint posterior distribution. They suggest taking the highest mean as the one with the highest posterior probability of being the largest. They note that using the ranking of the posterior means can lead to unreasonable results.

Shen and Louis (1998) point out that the goal or feature of interest of an investigation determines the solution. For example, they note, "(1) if unit-specific parameters are themselves the feature of interest, their posterior means are the optimal estimates, (2) If the ranks of the unit-specific parameters are the target feature..., the conditional expected ranks or a discretized version of them are optimal..., (and) (3) If the feature of interest is the histogram or empirical distribution function of the unit-specific parameters..., then the conditional expected empirical distribution function or a discretized version of it is optimal...No single set of estimates can simultaneously optimize the three inferential goals..." Thus Shen and Louis (1998) set out to produce "*triple goal estimates*; (1) those producing a histogram that is a good estimate of the parameter histogram, (2) (those) with induced ranks that are good estimates of the parameter ranks; and (3) (those) with good performance in estimating unit-specific parameters".

Seeking these *triple goal estimates*, Shen and Louis consider three candidates: (1) posterior means (PM); (2) the constrained Bayes estimates (CB) of Louis (1984); and (3) what they call GR estimates which "optimize on estimation of the empirical distribution function (G_K) and ranks (R)". For the GR estimates, one , according to Frey (2008), "...chooses individual parameter estimates that are consistent with the estimates of the ranks and the distribution function for the parameters". Shen and Louis provide theoretical results and simulation results that favor the triple goal estimates GR.

Klein and Wright (2001) present empirical results comparing several ranking procedures in a Bayesian setting.

Bootstrap Methods

Hall and Miller (2009) note that "...The bootstrap is a popular approach to developing ... a measure (of authority for a ranking of performance measures θ_k for $k = 1, 2, ..., K$ institutions, e.g., local governments, or health providers, or universities)".

Model: If r_k is the rank of the k^{th} population in an ordering of the K parameters $\theta_{(1)} \leq \theta_{(2)}, \leq \cdots \leq \theta_{(K)}$, then $\hat{r}_k$ is the estimated rank of the k^{th} population in an empirical ordering of the K sample estimates $\hat{\theta}_{(1)} \leq \hat{\theta}_{(2)} \leq \cdots \leq \hat{\theta}_{(K)}$ for $k = 1, 2, ..., K$. The researchers consider two cases; "fixed K" and "large K". An example of "large K" is the expression-level data on K genes over individuals and K can range from 5,000 to 20,000 genes to be ordered. Our interest is in the "fixed K" case.

Basic Bootstrap Methodology: It is assumed that there is a sample of size n independently from each of the K populations and that $\hat{r}_k$ is as noted above. Using the sample data from the k^{th} population, a standard bootstrap sample of size n is selected for $k = 1, 2, ..., K$. Using the new independent bootstrap samples, we compute $\hat{\theta}_1^*, \hat{\theta}_2^*, ..., \hat{\theta}_K^*$ and use these to compute the bootstrap estimates $\hat{r}_1^*, \hat{r}_2^*, ..., \hat{r}_K^*$. We repeat the bootstrap process a large number of times (this will be described in greater detail in Section 3 of this paper), ultimately obtaining a distribution of $\hat{r}_k^*$ for $k = 1, 2, ..., K$. This n-out-of-n conventional bootstrap approach can produce inconsistency; they observe that the distribution of an empirical rank may not converge in the usual sense; the estimation may converge in distribution, but not in probability. However, "the m-out-of-n bootstrap (where $m < n$) can improve performance and produce statistical consistency, but it requires empirical choice of m; (they) suggest a tuning solution to this problem." Though they use a nonparametric bootstrap, they note that the conclusions also apply to parametric bootstrap methods. Theory is provided. Some results do not require that the samples from the population be independent.

Hall and Miller (2010) study the phenomenon that certain "highly ranked" populations with *light-tailed distributions* such as normal or exponential tend to keep their ranks "...even when the data on which the rankings are based are extensively revised, and even when a large number of new (populations) are added to the investigation..."

2. VISUALLY COMPARING PAIRS OF STATES USING NORMAL THEORY

2.1. Comparing One Reference State with Each of the Other States

For population k, let $\hat{\theta}_k$ be as defined in Section 1, and let the estimated standard error be denoted by $SE(\hat{\theta}_k) = SE_k$ for $k = 1, 2, ..., K$. In this paper, we treat the SE_k estimates as though they were known constants. Let k^* be a specific reference population among the K populations with estimate $\hat{\theta}_{k^*}$ and standard error SE_{k^*}.

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

$$\left((\hat{\theta}_k - \hat{\theta}_{k^*}) - z_{\frac{\alpha}{2}} \sqrt{(SE_k)^2 + (SE_{k^*})^2} \ , \ (\hat{\theta}_k - \hat{\theta}_{k^*}) + z_{\frac{\alpha}{2}} \sqrt{(SE_k)^2 + (SE_{k^*})^2} \right) \tag{3}$$

where $z_{\frac{\alpha}{2}} = \Phi^{-1}(1 - \frac{\alpha}{2})$, and Φ is the cumulative standard normal distribution. To test

$$H_0 : \theta_k = \theta_{k^*} \qquad vs \qquad H_A : \theta_k \neq \theta_{k^*} \tag{4}$$

at significance level α, it is enough to compare the interval in (3) with zero (0). If the interval in (3) does not contain 0, we reject H_0 in favor of H_A; otherwise, we do not reject H_0.

Figure 1 gives an estimated ranking of the $K = 51$ states (including Washington, D.C.) based on point estimates $\hat{\theta}_k$ for the k^{th} state's mean travel time to work of workers 16 years and over who did not work at home with associated SE_k for $k = 1, 2, ..., 51$ using 2011 American Community Survey data. For example, the 2011 ACS estimate $\hat{\theta}_k$ of mean travel time to work of workers 16 years and over who did not work at home for California (CA) is 27.14 minutes with standard error $SE_k = .07$ minutes. Among the 51 states and based on the estimates, California has estimated rank $\hat{r}_k = 44$. We will formally define $\hat{r}_k$ in the section on bootstrapping and ranking (Section 3).

For the reference state $k^* \equiv$ Colorado (CO), and using a Bonferroni correction for each of the 50 tests comparing $\hat{\theta}_{k^*}$ with each $\hat{\theta}_k$ for $k \neq k^*$, we see that the shaded (both heavy and light shading) states in the column (Figure 1) are statistically significantly different from CO, while the non-shaded states in the column are not statistically significantly different from CO. The level of significance for each test is $\frac{\alpha}{50} = .002$ (note $\frac{.002}{2} = .001$ and $z_{.001} = 3.1$), and the family-wide (or overall) level of significance for the collection of 50 tests in the column is $\alpha = .10$.

$\hat{r}_k$	*State (k)*	$\hat{\theta}_k$	SE_k
51	MD	32.21	.15
50	NY	31.50	.09
49	NJ	30.53	.12
48	DC	30.10	.32
47	IL	28.17	.11
46	MA	27.99	.13
45	VA	27.74	.13
44	CA	27.14	.07
43	GA	27.11	.17
42	NH	26.90	.30
41	PA	25.92	.09
40	FL	25.76	.11
39	HI	25.69	.27
38	WV	25.58	.31
37	WA	25.51	.14
36	DE	25.30	.37
35	CT	24.98	.19
34	TX	24.82	.07
33	AZ	24.76	.15
32	LA	24.54	.15
31	CO	24.51	.19
30	TN	24.23	.14
29	MI	24.11	.10
28	NV	24.10	.27
27	AL	23.94	.14
26	MS	23.86	.24
25	SC	23.61	.16
24	IN	23.45	.11
23	ME	23.41	.25
22	NC	23.37	.12
21	RI	23.36	.29
20	OH	23.12	.09
19	MO	23.07	.13
18	MN	22.99	.10
17	KY	22.86	.15
16	OR	22.54	.16
15	VT	21.94	.31
14	WI	21.92	.11
13	UT	21.61	.20
12	NM	21.43	.27
11	AR	21.31	.23
10	OK	21.13	.15
9	ID	19.66	.24
8	KS	18.90	.16
7	IA	18.77	.13
6	AK	18.39	.33
5	MT	18.18	.32
4	WY	18.10	.50
3	NE	18.06	.19
2	ND	16.91	.36
1	SD	16.86	.28

Reference state: CO

Figure 1: Shaded States Do (Unshaded States Do Not) Differ from the Reference State Colorado for Mean Travel Time to Work of Workers 16 Years and Over Who Did Not Work at Home (Minutes). Significance Level for Each Pair Being Compared Is .002. For the Column with Some Shading, the Family-wide (or Overall) Significance Level for All Pairs Simultaneously Being Compared Is .10. For the USA, $\hat{\theta} = 25.51$ and $SE = .02$. (*Data Source:* 2011 American Community Survey)

Figure 2 (inspired by a display in Almond, Lewis, Tukey, and Yan, 2000) gives the overall visualization for all states where each column presents the 50 tests for the reference state noted at the very bottom of the column. For the United States, $\hat{\theta} = 25.51$ and $SE = 0.02$.

2.2. Comparing One Reference State with Each of the Other States Showing Confidence Intervals for Differences

Using the same setup as in Subsection 2.1, Figure 3 gives 50 confidence intervals of the difference $\theta_k - \theta_{k^*}$ for reference state $k^* \equiv$ Colorado and $k \neq k^*$. We use a Bonferroni correction for the tests as noted in Subsection 2.1. The level of significance for each test is .002, and the family-wide (or overall) level of significance for the collection of 50 tests is $\alpha = .10$. 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. Figures 1 and 3 both compare Colorado (CO) with each of the other 50 states.

Assuming that $\hat{\theta}_k$ is normally distributed, a $100(1-\alpha)\%$ confidence interval for θ_k is given by

$$\left(\hat{\theta}_k - z_{\frac{\alpha}{2}} SE_k \ , \ \hat{\theta}_k + z_{\frac{\alpha}{2}} SE_k\right). \tag{5}$$

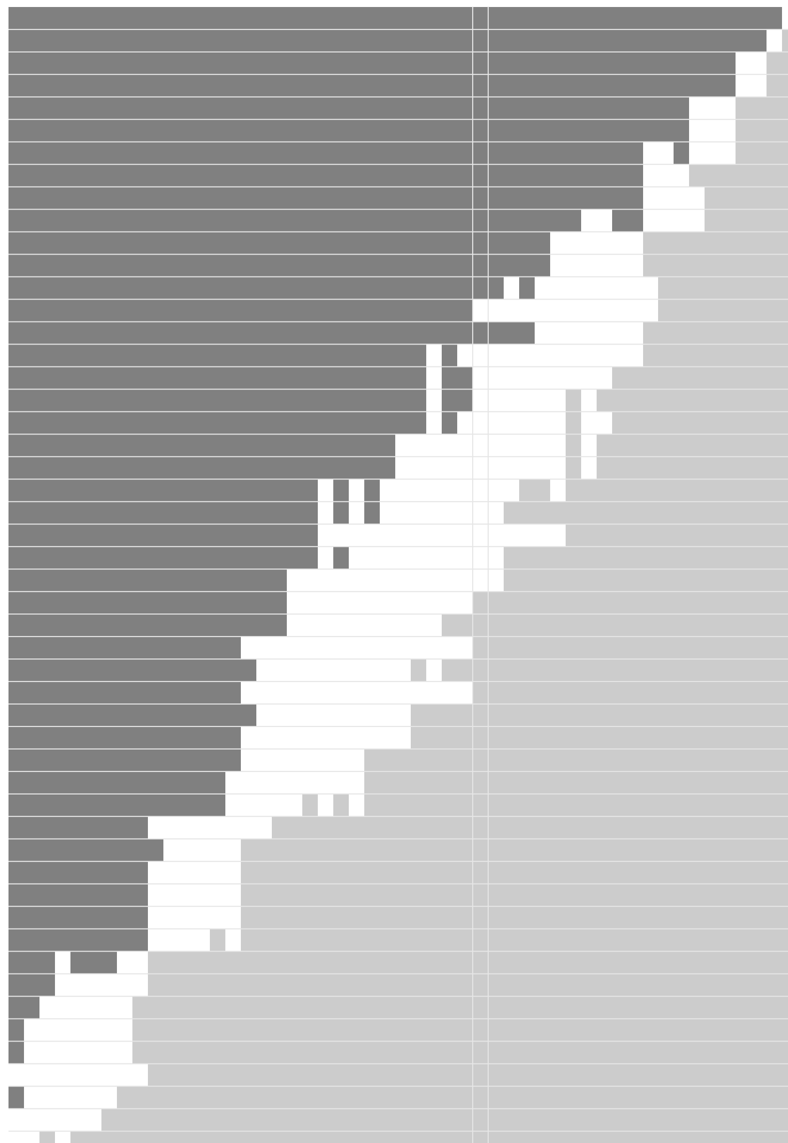
It is common practice to present one plot showing the fifty-one 90% confidence intervals similar to what is given in Figure 4 where each 90% confidence interval is computed as in (5). Incorrectly, some infer that overlapping confidence intervals for θ_k and $\theta_{k'}$ imply no statistically significant differences for θ_k and $\theta_{k'}$ at level α, while nonoverlapping intervals for θ_k and $\theta_{k'}$ imply statistically significant differences in θ_k and $\theta_{k'}$ for $k \neq k'$ at level α. In comparing populations k and k', the approach of considering a 90% confidence interval for the difference $\theta_k - \theta_{k'}$ is appropriate for $\alpha = .10$; merely comparing the 90% confidence interval of θ_k with the 90% confidence interval for $\theta_{k'}$ is not for $\alpha = .10$ (Schenker and Gentleman, 2001). In particular, the methods are not equivalent. For example, Schenker and Gentleman (2001) show that if a 90% confidence interval for θ_k does not overlap a 90% confidence interval for $\theta_{k'}$ and we use this to reject the hypothesis $H_0 : \theta_k = \theta_{k'}$, then this implies that we would reject the same hypothesis using the usual test (i.e., reject H_0 when the 90% confidence interval for $\theta_k - \theta_{k'}$ does not contain 0). This is okay. However, if the 90% confidence interval for $\theta_k - \theta_{k'}$ does not contain 0, it is not always true that the 90% confidence interval for θ_k does not overlap a 90% confidence interval for $\theta_{k'}$. In Subsection 2.4, we consider the overlap/nonoverlap of one confidence interval with another confidence interval. Before that and in Subsection 2.3, we consider the overlap/nonoverlap of one confidence interval with a "comparison interval".

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

Given a reference state k^* with a $100(1-\alpha)\%$ confidence interval for θ_{k^*} as given in (5), it is possible to construct an interval $(\hat{\theta}_k - w_k, \hat{\theta}_k + w_k)$ for state $k \neq k^*$ such that when the two intervals overlap, θ_k and θ_{k^*} are not statistically significantly different at level α, whereas if the two intervals do not overlap, then θ_k and θ_{k^*} are statistically significantly different. In this section, we discuss construction of the interval $(\hat{\theta}_k - w_k, \hat{\theta}_k + w_k)$, as presented in Almond, Lewis, Tukey, and Yan (2000).

For population k, let $\hat{\theta}_k$ be as defined in Section 1, and let the standard error be denoted by $SE(\hat{\theta}_k) = SE_k$ for $k = 1, 2, ..., K$. Let k^* be a specific reference population among the K populations with estimate $\hat{\theta}_{k^*}$ and standard error SE_{k^*}. Assuming $\hat{\theta}_{k^*}$ is normally distributed, a

$\hat{r}_k$	*State (k)*	$\hat{\theta}_k$	SE_k
51	MD	32.21	.15
50	NY	31.50	.09
49	NJ	30.53	.12
48	DC	30.10	.32
47	IL	28.17	.11
46	MA	27.99	.13
45	VA	27.74	.13
44	CA	27.14	.07
43	GA	27.11	.17
42	NH	26.90	.30
41	PA	25.92	.09
40	FL	25.76	.11
39	HI	25.69	.27
38	WV	25.58	.31
37	WA	25.51	.14
36	DE	25.30	.37
35	CT	24.98	.19
34	TX	24.82	.07
33	AZ	24.76	.15
32	LA	24.54	.15
31	CO	24.51	.19
30	TN	24.23	.14
29	MI	24.11	.10
28	NV	24.10	.27
27	AL	23.94	.14
26	MS	23.86	.24
25	SC	23.61	.16
24	IN	23.45	.11
23	ME	23.41	.25
22	NC	23.37	.12
21	RI	23.36	.29
20	OH	23.12	.09
19	MO	23.07	.13
18	MN	22.99	.10
17	KY	22.86	.15
16	OR	22.54	.16
15	VT	21.94	.31
14	WI	21.92	.11
13	UT	21.61	.20
12	NM	21.43	.27
11	AR	21.31	.23
10	OK	21.13	.15
9	ID	19.66	.24
8	KS	18.90	.16
7	IA	18.77	.13
6	AK	18.39	.33
5	MT	18.18	.32
4	WY	18.10	.50
3	NE	18.06	.19
2	ND	16.91	.36
1	SD	16.86	.28



Reference state: SD ND NE WY MT AK IA KS ID OK AR NM UT WI VT OR KY MN MO OH RI NC ME IN SC MS AL NV MI TN CO LA AZ TX CT DE WA WV HI FL PA NH GA CA VA MA IL DC NJ NY MD

Figure 2: In Each Column, Shaded States Do (Unshaded States Do Not) Differ from Reference State for Mean Travel Time to Work of Workers 16 Years and Over Who Did Not Work at Home (Minutes). Significance Level for Each Pair Being Compared Is .002. For Each Column with Some Shading, the Family-wide (or Overall) Significance Level for All Pairs Simultaneously Being Compared Is .10. For the USA, $\hat{\theta} = 25.51$ and $SE = .02$. (*Data Source:* 2011 American Community Survey)

$100(1-\alpha)\%$ confidence interval for θ_{k^*} is given by (see also (5))

$$\left(\hat{\theta}_{k^*} - z_{\frac{\alpha}{2}} SE_{k^*} \;\; , \;\; \hat{\theta}_{k^*} + z_{\frac{\alpha}{2}} SE_{k^*}\right). \tag{6}$$

Now consider another population, say k, where $\hat{\theta}_k < \hat{\theta}_{k^*}$. (What follows also holds in an analogous way if $\hat{\theta}_{k^*} < \hat{\theta}_k$.) We want to find the width w_k such that the interval $(\hat{\theta}_k - w_k, \hat{\theta}_k + w_k)$ overlaps the interval in (6) if and only if θ_k and θ_{k^*} are not significantly different at level α. In other words, referring to Figure 5, we want

$$\begin{aligned}(d_{k(low)}, d_{k(high)}) &= \left((\hat{\theta}_{k^*} - z_{\frac{\alpha}{2}} SE_{k^*}) - (\hat{\theta}_k + w_k) \;\; , \;\; (\hat{\theta}_{k^*} + z_{\frac{\alpha}{2}} SE_{k^*}) - (\hat{\theta}_k - w_k)\right) \\ &= \left((\hat{\theta}_{k^*} - \hat{\theta}_k) - (z_{\frac{\alpha}{2}} SE_{k^*} + w_k) \;\; , \;\; (\hat{\theta}_{k^*} - \hat{\theta}_k) + (z_{\frac{\alpha}{2}} SE_{k^*} + w_k)\right)\end{aligned} \tag{7}$$

to be a $100(1-\alpha)\%$ confidence interval for the difference $\theta_{k^*} - \theta_k$. But a $100(1-\alpha)\%$ confidence interval for $\theta_{k^*} - \theta_k$ is given by (see also (3))

$$\left((\hat{\theta}_{k^*} - \hat{\theta}_k) - z_{\frac{\alpha}{2}}\sqrt{(SE_{k^*})^2 + (SE_k)^2} \;\; , \;\; (\hat{\theta}_{k^*} - \hat{\theta}_k) + z_{\frac{\alpha}{2}}\sqrt{(SE_{k^*})^2 + (SE_k)^2}\right). \tag{8}$$

Equating results in (7) and (8) gives

$$z_{\frac{\alpha}{2}} SE_{k^*} + w_k = z_{\frac{\alpha}{2}}\sqrt{(SE_{k^*})^2 + (SE_k)^2} \tag{9}$$

or equivalently

$$w_k = z_{\frac{\alpha}{2}}\sqrt{(SE_{k^*})^2 + (SE_k)^2} - z_{\frac{\alpha}{2}} SE_{k^*}. \tag{10}$$

If the situation is as shown in Figure 5 where $(\hat{\theta}_k - w_k, \hat{\theta}_k + w_k)$ and (6) do not overlap, both $d_{k(low)}$ and $d_{k(high)}$ are positive; the confidence interval in (7) does not contain zero; and hence θ_k and θ_{k^*} are significantly different at level α. In the cases where $(\hat{\theta}_k - w_k, \hat{\theta}_k + w_k)$ and (6) do overlap, the signs of $d_{k(low)}$ and $d_{k(high)}$ will differ; the confidence interval in (7) will contain zero; and hence θ_k and θ_{k^*} are not significantly different at level α.

Relative to $\hat{\theta}_{k^*}$, we refer to the interval $(\hat{\theta}_k - w_k, \hat{\theta}_k + w_k)$ as a "θ_{k^*} *comparison interval for* θ_k." The comparison interval for θ_k is not a confidence interval, while the interval for θ_{k^*} is a confidence interval.

Thus considering only the specific reference population k^* and the other population k, that is $K = 2$ populations, we could have one of the possibilities shown in the graphical display of Figure 6 (where $\hat{\theta}_k < \hat{\theta}_{k^*}$) . In each case, the length of each bar around $\hat{\theta}_k$ is $2w_k$.

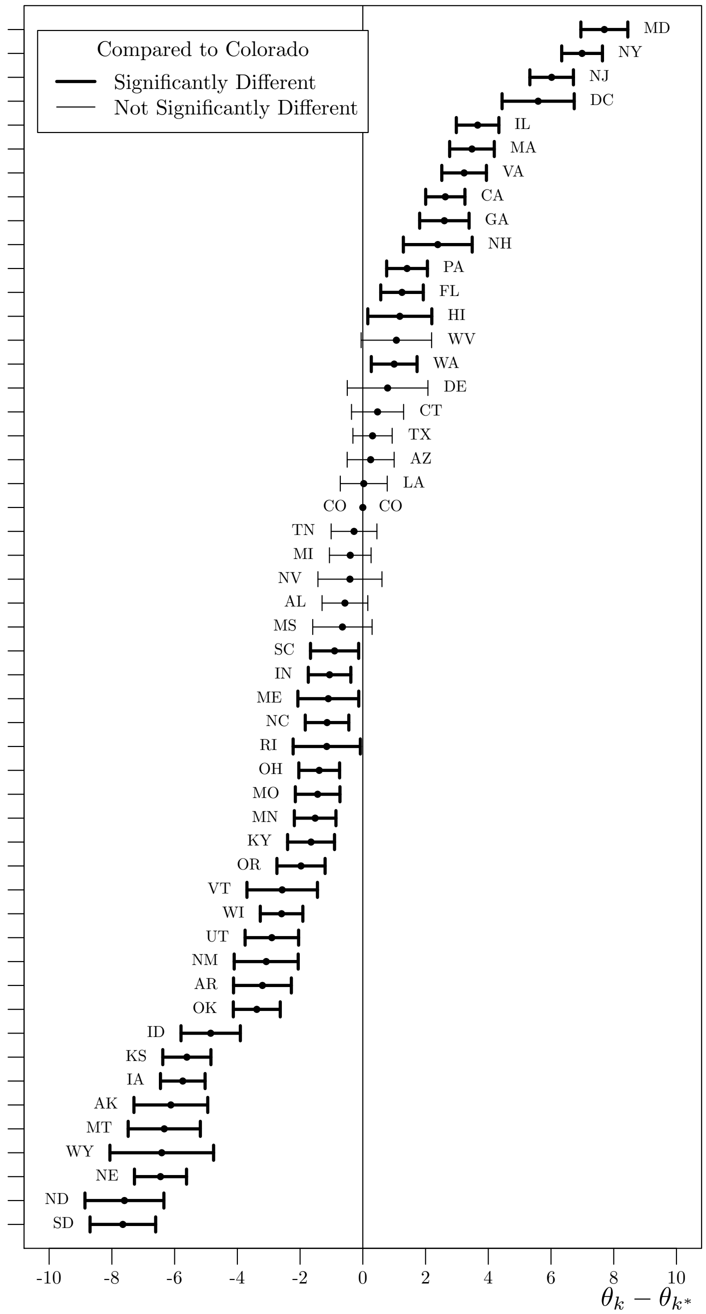
In Figure 6 (a), populations k^* and k are significantly different at level α. In Figures (b) or (c), populations k^* and k are not significantly different at level α.

Figure 7 shows a typical visualization where $K = 51$, and the reference population (workers who live in Colorado) has rank 31 based on the sample estimates. Figure 7 makes use of a Bonferroni correction for 50 separate tests of hypotheses where Colorado state's mean travel time is compared with each of the other $K-1=50$ states's mean travel time. The level of significance for each test is $\dfrac{\alpha}{50}=.002$, and the family-wide (or overall) level of significance for the collection of 50 tests is $\alpha=.10$.

From the overall testing in Figure 7 at overall level $\alpha=.10$, we see that Colorado's mean travel time to work of workers 16 years and over who did not work at home is significantly different from all of the states except Mississippi(MS), Alabama (AL), Nevada (NV), Michigan(MI), Tennessee (TN), Louisiana(LA), Arizona(AZ), Texas(TX), Connecticut(CT), Delaware(DE), and West Virginia(WV). The same comparison results for Colorado are also shown in Figures 1, 2, and 3. The interval around Colorado (the reference state) that corresponds to the shaded strip is an approximate 99.8% confidence interval for Colorado's mean travel time to work of workers 16 years and over who did not work at home during the year 2011. The intervals around each of the other states, say k, represents the comparison interval $(\hat{\theta}_k-w_k,\hat{\theta}_k+w_k)$ where w_k is given in (10). It is important to note that the interval $(\hat{\theta}_k-w_k,\hat{\theta}_k+w_k)$ is not a confidence interval for θ_k.

While visually different, Figures 3 and 7 provide the same information regarding comparing Colorado (θ_{k^*}) to the other states. In Figure 7, the usual 99.8% confidence interval for the reference state Colorado (θ_{k^*}) is shown explicitly; the (Bonferroni-corrected) "comparison intervals" are not usual confidence intervals; and each state comparison interval (θ_k) with the 99.8% confidence interval for the reference state Colorado provides the usual test of $H_0:\theta_k=\theta_{k^*}$ by use of the 99.8% confidence interval $\theta_k-\theta_{k^*}$. On the other hand, all of the intervals in Figure 3 are really the usual 99.8% confidence intervals for $\theta_k-\theta_{k^*}$, but we do not see the 99.8% confidence interval for θ_{k^*}, i.e., for the reference state Colorado.

$\hat{r}_k$	*State (k)*	$\hat{\theta}_k$	SE_k
51	Maryland	32.21	.15
50	New York	31.50	.09
49	New Jersey	30.53	.12
48	District of Columbia	30.10	.32
47	Illinois	28.17	.11
46	Massachusetts	27.99	.13
45	Virginia	27.74	.13
44	California	27.14	.07
43	Georgia	27.11	.17
42	New Hampshire	26.90	.30
41	Pennsylvania	25.92	.09
40	Florida	25.76	.11
39	Hawaii	25.69	.27
38	West Virginia	25.58	.31
37	Washington	25.51	.14
36	Delaware	25.30	.37
35	Connecticut	24.98	.19
34	Texas	24.82	.07
33	Arizona	24.76	.15
32	Louisiana	24.54	.15
31	Colorado	24.51	.19
30	Tennessee	24.23	.14
29	Michigan	24.11	.10
28	Nevada	24.10	.27
27	Alabama	23.94	.14
26	Mississippi	23.86	.24
25	South Carolina	23.61	.16
24	Indiana	23.45	.11
23	Maine	23.41	.25
22	North Carolina	23.37	.12
21	Rhode Island	23.36	.29
20	Ohio	23.12	.09
19	Missouri	23.07	.13
18	Minnesota	22.99	.10
17	Kentucky	22.86	.15
16	Oregon	22.54	.16
15	Vermont	21.94	.31
14	Wisconsin	21.92	.11
13	Utah	21.61	.20
12	New Mexico	21.43	.27
11	Arkansas	21.31	.23
10	Oklahoma	21.13	.15
9	Idaho	19.66	.24
8	Kansas	18.90	.16
7	Iowa	18.77	.13
6	Alaska	18.39	.33
5	Montana	18.18	.32
4	Wyoming	18.10	.50
3	Nebraska	18.06	.19
2	North Dakota	16.91	.36
1	South Dakota	16.86	.28



Reference State (k^*): Colorado
Rank: 31

Figure 3: Fifty Different $100(1 - .002)\% = 99.8\%$ Confidence Intervals for $\theta_k - \theta_{k^*}$ with Reference State $k^* \equiv$ Colorado for Mean Travel Time to Work of Workers 16 Years and Over Who Did Not Work at Home (Minutes). Overall $\alpha = .10$ for The Collection of Fifty Tests. (*Data Source:* 2011 American Community Survey)

$\hat{r}_k$	*State (k)*	$\hat{\theta}_k$	SE_k
51	Maryland	32.21	.15
50	New York	31.50	.09
49	New Jersey	30.53	.12
48	District of Columbia	30.10	.32
47	Illinois	28.17	.11
46	Massachusetts	27.99	.13
45	Virginia	27.74	.13
44	California	27.14	.07
43	Georgia	27.11	.17
42	New Hampshire	26.90	.30
41	Pennsylvania	25.92	.09
40	Florida	25.76	.11
39	Hawaii	25.69	.27
38	West Virginia	25.58	.31
37	Washington	25.51	.14
36	Delaware	25.30	.37
35	Connecticut	24.98	.19
34	Texas	24.82	.07
33	Arizona	24.76	.15
32	Louisiana	24.54	.15
31	Colorado	24.51	.19
30	Tennessee	24.23	.14
29	Michigan	24.11	.10
28	Nevada	24.10	.27
27	Alabama	23.94	.14
26	Mississippi	23.86	.24
25	South Carolina	23.61	.16
24	Indiana	23.45	.11
23	Maine	23.41	.25
22	North Carolina	23.37	.12
21	Rhode Island	23.36	.29
20	Ohio	23.12	.09
19	Missouri	23.07	.13
18	Minnesota	22.99	.10
17	Kentucky	22.86	.15
16	Oregon	22.54	.16
15	Vermont	21.94	.31
14	Wisconsin	21.92	.11
13	Utah	21.61	.20
12	New Mexico	21.43	.27
11	Arkansas	21.31	.23
10	Oklahoma	21.13	.15
9	Idaho	19.66	.24
8	Kansas	18.90	.16
7	Iowa	18.77	.13
6	Alaska	18.39	.33
5	Montana	18.18	.32
4	Wyoming	18.10	.50
3	Nebraska	18.06	.19
2	North Dakota	16.91	.36
1	South Dakota	16.86	.28

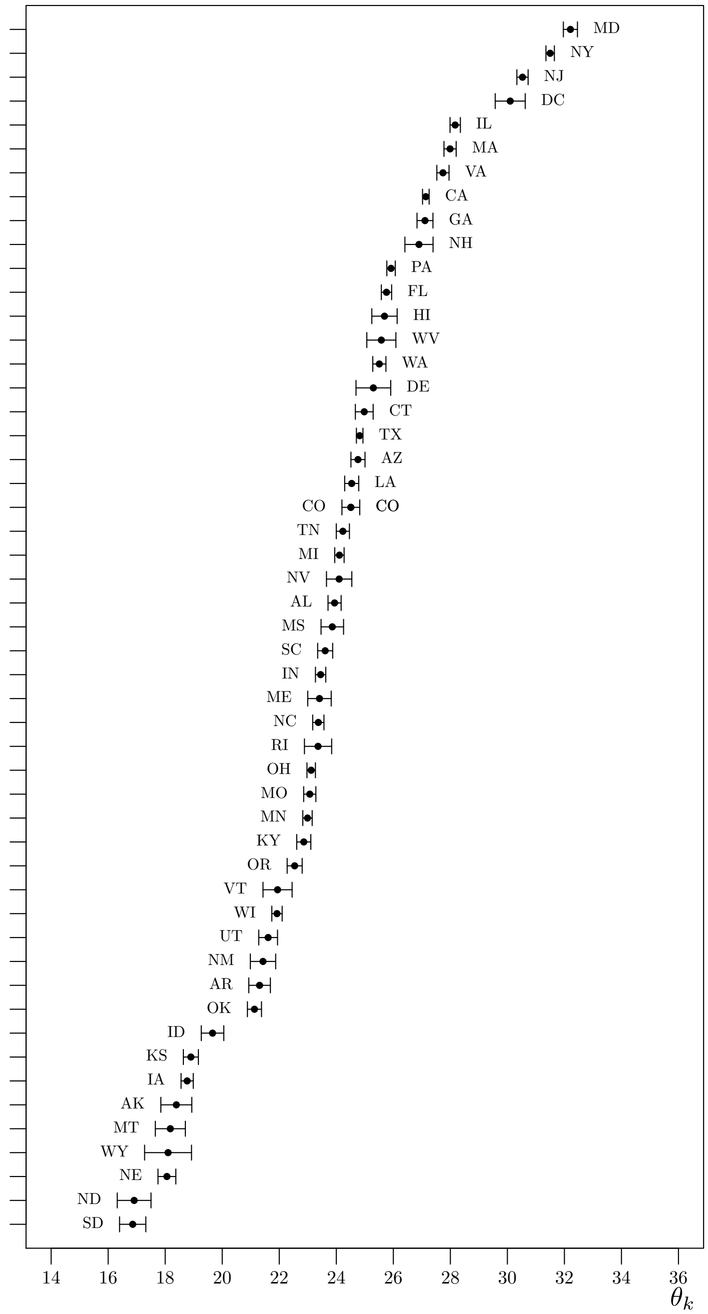


Figure 4: 90% Confidence Interval for θ_k for Each State for Mean Travel Time to Work of Workers 16 Years and Over Who Did Not Work at Home (Minutes). (*Data Source:* 2011 American Community Survey)

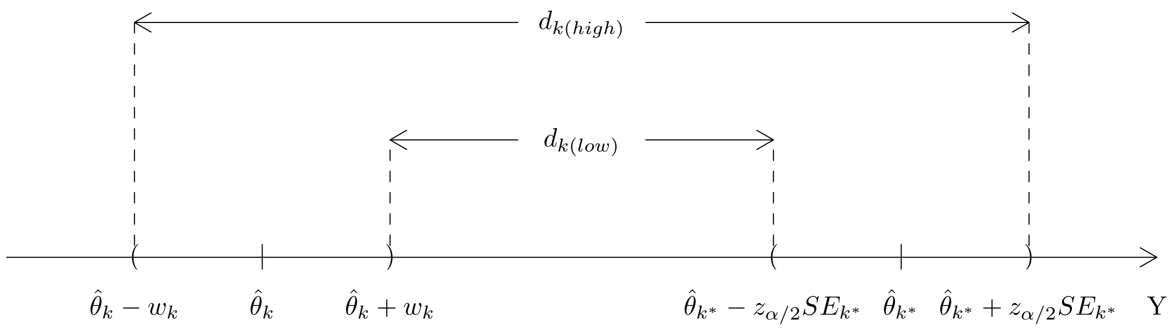


Figure 5: Illustration of Motivation for Method of Almond, Lewis, Tukey, and Yan(2000).

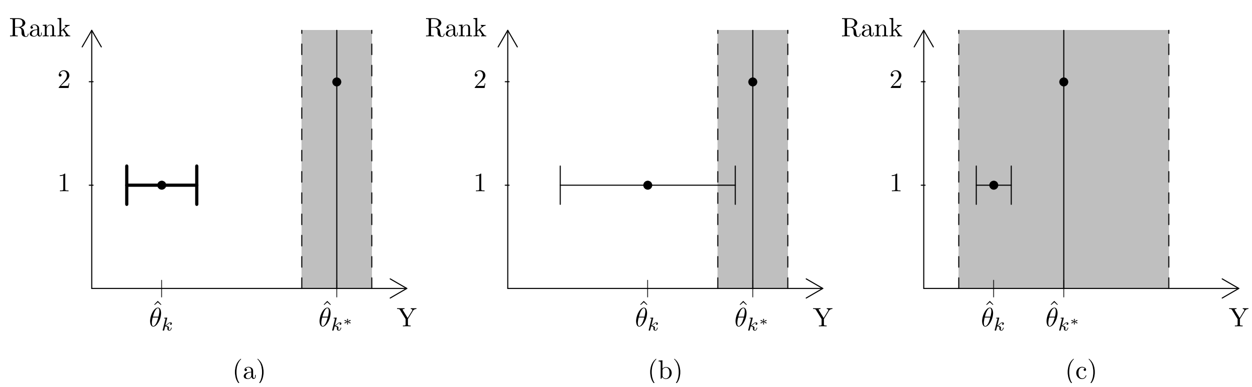


Figure 6: For $K = 2$, Three Possibilities for Method of Almond, Lewis, Tukey, and Yan (2000).

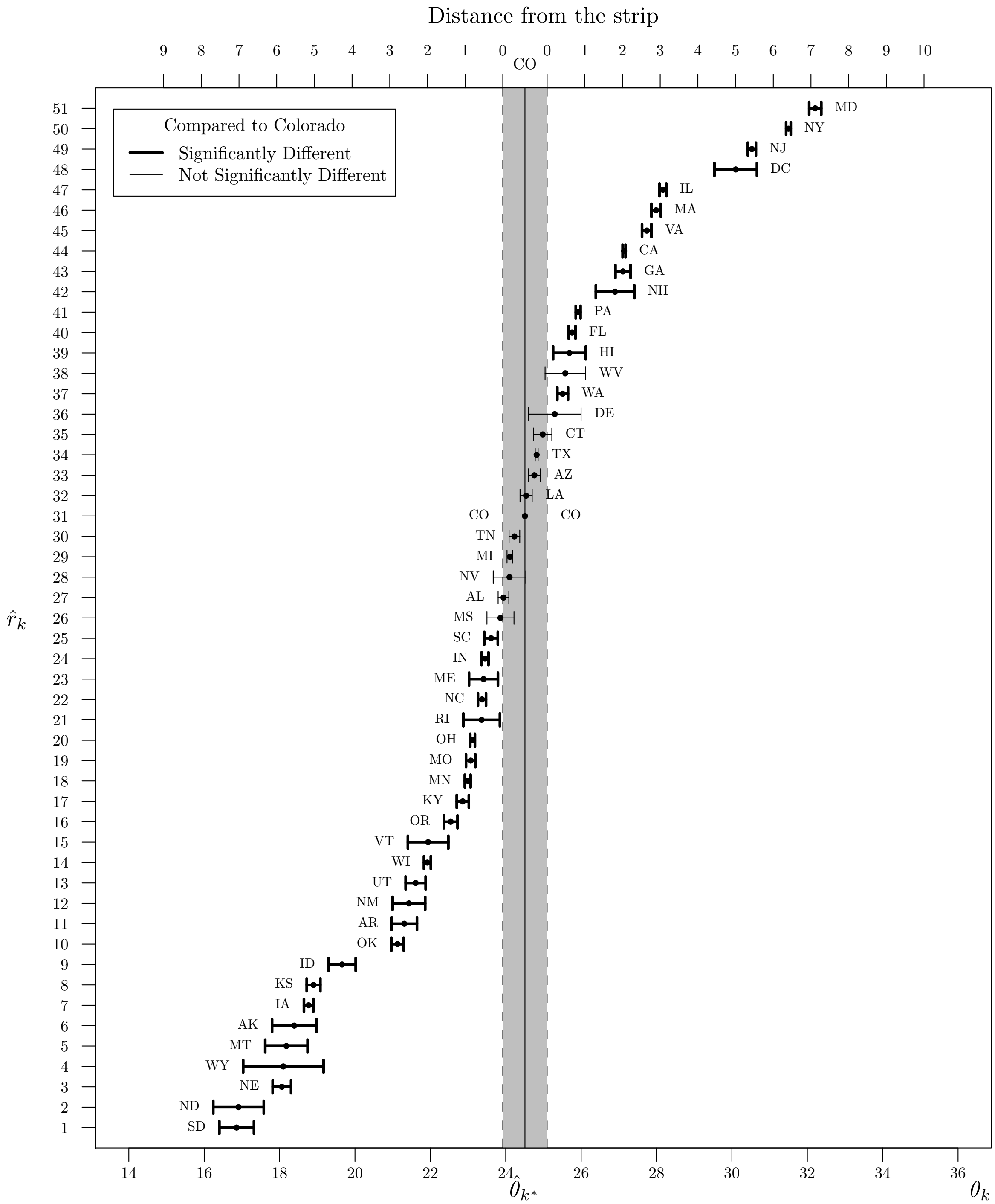


Figure 7: Comparisons with Reference State Colorado Using Overlapping Intervals for Mean Travel Time to Work of Workers 16 Years and Over Who Did Not Work at Home (Minutes). Significance Level for Each State Being Compared with Colorado is .002. (*Data Source:* 2011 American Community Survey)

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

Goldstein and Spiegelhalter (1996) and discussants of their paper provide extensive details highlighting the great difficulty in assessing the uncertainty in ranking tables (or "league tables"). In particular, they focus on some statistical issues where institutions are ranked based on quantitative performance indicators. Acknowledging three broad categories of performance indicators (input, process, and output), they focus on output indicators for ranking education and health institutions, e.g., school examinations results or hospital operative mortality. Arguing that the context of outcome indicators should take account of institutional circumstances, models play a major role in their presentation. Much attention is devoted to ensuring that data are adjusted so that institutions are "comparative." Whatever approach is taken, they argue strongly for the use of intervals in conveying uncertainty explicitly in estimates or estimated ranks. Two procedures for deriving intervals are given: (i) one that makes use of usual confidence intervals around estimated means of each institution and (ii) another that was proposed by Goldstein and Healy (1995) which we describe in what follows.

Consider the pair of populations k and k' with parameters θ_k and $\theta_{k'}$. If a $100(1-\alpha)\%$ confidence interval for θ_k does not overlap an independent $100(1-\alpha)\%$ confidence interval for $\theta_{k'}$, we could declare that θ_k and $\theta_{k'}$ are statistically significantly different, but the level of significance is not α. However, the level of significance can be determined. If the desire is that the level of significance be α, we can adjust the confidence coefficient to a value, say $100(1-\alpha_A)\%$, such that if the $100(1-\alpha_A)\%$ confidence interval for θ_k does not overlap an independent $100(1-\alpha_A)\%$ confidence interval for $\theta_{k'}$, then we can declare θ_k and $\theta_{k'}$ as statistically significantly different at significance level α. Goldstein and Healy (1995) show how to do this when comparing one pair of estimates as well as what one might do to find a common α_A in the case of comparing several pairs of estimates. We give details of and illustrate their method.

We continue to assume that we have K independently normally distributed estimators $\hat{\theta}_k$ with standard error SE_k for $k = 1, 2, 3, ..., K$.

Comparing One Pair of Populations k and k'

When comparing one pair of populations k and k', we want to determine an adjusted value α_A for a desired significance level α such that when the $100(1-\alpha_A)\%$ confidence interval for θ_k does not overlap the $100(1-\alpha_A)\%$ confidence interval for $\theta_{k'}$ we can correctly declare θ_k and $\theta_{k'}$ statistically significantly different at significance level α.

Let the $100(1-\alpha_A)\%$ confidence interval for θ_k be $\left(\hat{\theta}_k - z_{\frac{\alpha_A}{2}} SE_k \;\; , \;\; \hat{\theta}_k + z_{\frac{\alpha_A}{2}} SE_k\right)$ and the $100(1-\alpha_A)\%$ confidence interval for $\theta_{k'}$ be $\left(\hat{\theta}_{k'} - z_{\frac{\alpha_A}{2}} SE_{k'} \;\; , \;\; \hat{\theta}_{k'} + z_{\frac{\alpha_A}{2}} SE_{k'}\right)$. In each case, $z_{\frac{\alpha_A}{2}}$ is the value associated with the standard normal such that $P(Z \geq z_{\frac{\alpha_A}{2}}) = \frac{\alpha_A}{2}$. If $|\hat{\theta}_k - \hat{\theta}_{k'}| > z_{\frac{\alpha_A}{2}}(SE_k + SE_{k'})$, then we have two cases: (i) $\hat{\theta}_k - \hat{\theta}_{k'} > z_{\frac{\alpha_A}{2}}(SE_k + SE_{k'})$ or (ii) $-(\hat{\theta}_k - \hat{\theta}_{k'}) > z_{\frac{\alpha_A}{2}}(SE_k + SE_{k'})$.

(i) In the first case, $\hat{\theta}_k - \hat{\theta}_{k'} > z_{\frac{\alpha_A}{2}}(SE_k + SE_{k'})$ is equivalent to $\hat{\theta}_k - z_{\frac{\alpha_A}{2}} SE_k > \hat{\theta}_{k'} + z_{\frac{\alpha_A}{2}} SE_{k'}$. Hence the confidence interval for θ_k is completely above the confidence interval for $\theta_{k'}$, and thus they do not overlap.

(ii) In the second case, $-(\hat{\theta}_k - \hat{\theta}_{k'}) > z_{\frac{\alpha_A}{2}}(SE_k + SE_{k'})$ is equivalent to $\hat{\theta}_{k'} - z_{\frac{\alpha_A}{2}} SE_{k'} > \hat{\theta}_k + z_{\frac{\alpha_A}{2}} SE_k$. Hence the confidence interval for $\theta_{k'}$ is completely above the confidence interval for θ_k, and thus they do not overlap.

Thus

$$|\hat{\theta}_k - \hat{\theta}_{k'}| > z_{\frac{\alpha_A}{2}}(SE_k + SE_{k'}) \tag{11}$$

if and only if the $100(1-\alpha_A)\%$ confidence intervals for θ_k and $\theta_{k'}$ do not overlap.

Next, let $(SE_{kk'})^2 \equiv Var(\hat{\theta}_k - \hat{\theta}_{k'}) = (SE_k)^2 + (SE_{k'})^2$. Then the probability of the event in (11) under the hypothesis $\theta_k = \theta_{k'}$, which is the probability of a Type I error, is

$$\begin{aligned}
\gamma_{kk'} &= P\left(|\hat{\theta}_k - \hat{\theta}_{k'}| > z_{\frac{\alpha_A}{2}}(SE_k + SE_{k'})\right) = 2P\left(\hat{\theta}_k - \hat{\theta}_{k'} > z_{\frac{\alpha_A}{2}}(SE_k + SE_{k'})\right) \\
&= 2P\left(\frac{(\hat{\theta}_k - \hat{\theta}_{k'}) - 0}{SE_{kk'}} > z_{\frac{\alpha_A}{2}} \frac{(SE_k + SE_{k'})}{SE_{kk'}}\right) \\
&= 2P\left(Z > z_{\frac{\alpha_A}{2}} \frac{(SE_k + SE_{k'})}{SE_{kk'}}\right) \\
&= 2\left(1 - \Phi(z_{\frac{\alpha_A}{2}} \frac{(SE_k + SE_{k'})}{SE_{kk'}})\right)
\end{aligned} \tag{12}$$

where Φ is the cumulative standard normal distribution. Thus (12) relates $\gamma_{kk'}$ and $z_{\frac{\alpha_A}{2}}$ (hence α and $z_{\frac{\alpha_A}{2}}$) for given values of SE_k and $SE_{k'}$. So if we want the probability of a Type I error $\gamma_{kk'}$ 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_k and $\theta_{k'}$ do not overlap we can correctly say that θ_k and $\theta_{k'}$ are statistically significantly different at significance level α.

It is important to note for any given values of SE_k and $SE_{k'}$, that by the triangle inequality, $\sqrt{(SE_k)^2 + (SE_{k'})^2} \leq SE_k + SE_{k'}$ and hence

$$1 \leq \frac{SE_k + SE_{k'}}{\sqrt{(SE_k)^2 + (SE_{k'})^2}} = \frac{SE_k + SE_{k'}}{SE_{kk'}} . \tag{13}$$

We also note that

$$\begin{aligned}
0 &\leq (SE_k - SE_{k'})^2 \\
0 &\leq (SE_k)^2 - 2(SE_k)(SE_{k'}) + (SE_{k'})^2 \\
(SE_k)^2 + 2(SE_k)(SE_{k'}) + (SE_{k'})^2 &\leq 2(SE_k)^2 + 2(SE_{k'})^2 \\
(SE_k + SE_{k'})^2 &\leq 2\left((SE_k)^2 + (SE_{k'})^2\right) \qquad (14) \\
\frac{SE_k + SE_{k'}}{\sqrt{(SE_k)^2 + (SE_{k'})^2}} &\leq \sqrt{2} \\
\frac{SE_k + SE_{k'}}{SE_{kk'}} &\leq \sqrt{2}.
\end{aligned}$$

Hence by (13) and (14), we have

$$1 \leq \frac{SE_k + SE_{k'}}{SE_{kk'}} \leq \sqrt{2}.$$

Thus given α_A, the probability of a Type I error $\gamma_{kk'} = 2\left(1 - \Phi(z_{\frac{\alpha_A}{2}} \frac{(SE_k + SE_{k'})}{SE_{kk'}})\right)$ is bounded above by $2\left(1 - \Phi(z_{\frac{\alpha_A}{2}})\right)$ and bounded below by $2\left(1 - \Phi(z_{\frac{\alpha_A}{2}}\sqrt{2})\right)$.

In practice, we set $\gamma_{kk'}$ equal to a chosen α, and determine the appropriate α_A given SE_k and $SE_{k'}$ using

$$z_{\frac{\alpha_A}{2}} \frac{SE_k + SE_{k'}}{SE_{kk'}} = z_{\frac{\alpha}{2}}. \qquad (15)$$

Using estimates from Figure 1 of the 2011 mean travel time to work of workers 16 years and over who did not work at home for Arizona (AZ), Colorado (CO), and Wyoming (WY), given for easy reference in the table below, we illustrate the method of Goldstein and Healy (1995).

State(k)	$\hat{\theta}_k$	SE_k
AZ(1)	24.76	.15
CO(2)	24.51	.19
WY(3)	18.10	.50

Example: Comparing the Pair of States AZ and CO

Let $\alpha = .10$. We want to determine the confidence coefficient $100(1-\alpha_A)\%$ such that if the $100(1-\alpha_A)\%$ confidence interval for Arizona's θ_1 does not overlap the $100(1-\alpha_A)\%$ confidence interval for Colorado's θ_2, then we can correctly declare θ_1 and θ_2 are statistically significantly different at significance level α.

Note that $z_{\frac{\alpha_A}{2}} \dfrac{(SE_1 + SE_2)}{SE_{12}} = z_{\frac{\alpha_A}{2}} 1.40$. So for $\alpha = .10$, $z_{.05} = 1.645$. Hence by (15) and solving $1.645 = z_{\frac{\alpha_A}{2}} 1.40$, we see that $z_{\frac{\alpha_A}{2}} = 1.17$, and hence $\alpha_A = .242$. Thus the $100(1 - .242)\% = 76\%$ confidence interval for θ_1 is $(24.76 - (1.17)(.15)\ ,\ 24.76 + (1.17)(.15)) = (24.62, 24.98)$. Similarly, a 76% confidence interval for θ_2 is $(24.28\ ,\ 24.72)$. Note that they overlap. See Figure 8(a).

Note also for $\alpha = .10$, that a 90% confidence interval for $\theta_1 - \theta_2$ is

$$\left((24.76 - 24.51) - 1.645\sqrt{(.15)^2 + (.19)^2}\ ,\ (24.76 - 24.51) + 1.645\sqrt{(.15)^2 + (.19)^2}\right) = (-.1\ ,\ .70)\,.$$

Because this interval does include 0, we would not be able to say that the populations are statistically significantly different at $\alpha = .10$. This is consistent with the conclusion from the previous paragraph where the 76% confidence intervals for θ_1 and θ_2 overlap.

Example: Comparing the Pair of States WY and CO

Again, let $\alpha = .10$. We want to determine the confidence coefficient $100(1 - \alpha_A)\%$ such that if the $100(1 - \alpha_A)\%$ confidence interval for Wyoming's θ_3 does not overlap the $100(1 - \alpha_A)\%$ confidence interval for Colorado's θ_2, then we can correctly declare θ_3 and θ_2 are statistically significantly different at significance level α.

Note that $z_{\frac{\alpha_A}{2}} \dfrac{(SE_3 + SE_2)}{SE_{32}} = z_{\frac{\alpha_A}{2}} 1.29$. So for $\alpha = .10$, $z_{.05} = 1.645$. Hence by (15) and solving $1.645 = z_{\frac{\alpha_A}{2}} 1.29$, we see that $z_{\frac{\alpha_A}{2}} = 1.28$, and hence $\alpha_A = .2006$. Thus the $100(1 - .2006)\% = 80\%$ confidence interval for θ_3 is $(18.10 - (1.28)(.50)\ ,\ 18.10 + (1.28)(.50)) = (17.46\ ,\ 18.74)$. Similarly, an 80% confidence interval for θ_2 is $(24.26\ ,\ 24.74)$. Note that they do not overlap. See Figure 8(b).

Note also for $\alpha = .10$, that a 90% confidence interval for $\theta_3 - \theta_2$ is

$$\left((18.10 - 24.51) - 1.645\sqrt{(.19)^2 + (.50)^2}\ ,\ (18.10 - 24.51) + 1.645\sqrt{(.19)^2 + (.50)^2}\right) = (-7.28\ ,\ -5.52)\,.$$

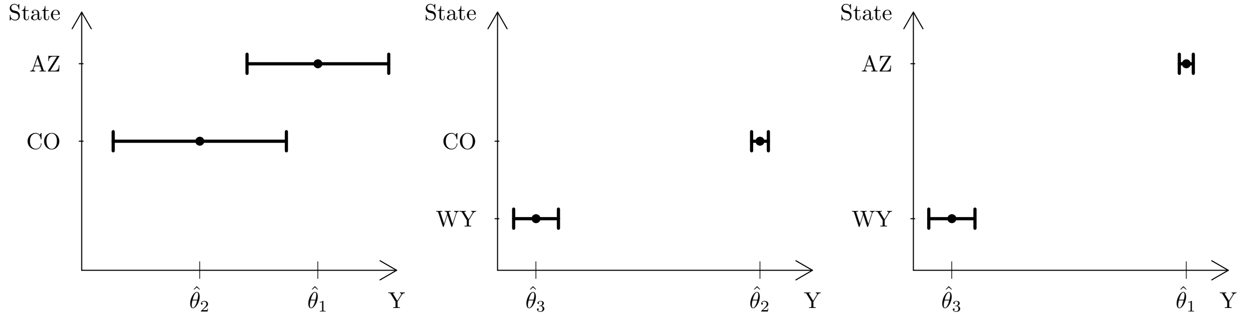
Because this interval does not include 0, we would be able to say that the populations are statistically significantly different at $\alpha = .10$. This is consistent with the conclusion from the previous paragraph where the 80% confidence intervals for θ_2 and θ_3 do not overlap.

Example: Comparing the Pair of States AZ and WY

For the pair of states Arizona (θ_1) and Wyoming (θ_3), we analogously determine for $\alpha = .10$ that we have 81% confidence intervals for θ_1 and θ_3 respectively as (24.56, 24.96) and (17.44, 18.76) which do not overlap. Thus we would infer that θ_1 and θ_3 are different at $\alpha = .10$. See Figure 8(c). Note also that a 90% confidence interval for $\theta_1 - \theta_3$ is (6.57, 6.75) which does not contain 0.

Comparing All Pairs of Populations k and k'

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 $\gamma_{kk'}$ over all (k, k') is a predetermined value,



(a) 76% confidence intervals (b) 80% confidence intervals (c) 81% confidence intervals

Figure 8: $100(1-\alpha_A)\%$ Confidence Intervals for Three Separate Pairs: Declare states k and k' statistically significantly different at significance level $\alpha = .10$ if $100(1-\alpha_A)\%$ confidence intervals in each pair do not overlap. (*Data Source:* 2011 American Community Survey)

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 $z_{\frac{\alpha_A}{2}} \dfrac{SE_k + SE_{k'}}{SE_{kk'}}$ taken over all the pairs (k, k'). The confidence interval for the k^{th} (population) is then given by $\left(\hat{\theta}_k - z_{\frac{\alpha_A}{2}} SE_k \;,\; \hat{\theta}_k + z_{\frac{\alpha_A}{2}} SE_k\right)$."

We will illustrate this advice by finding a $z_{\frac{\alpha_A}{2}}$ simultaneously for the three pairs (AZ, CO), (WY, CO), and (AZ, WY) so that the average significance level across all three pairs is $\alpha = .10$. Note that for the various pairs we have the following values.

Pairs	$\dfrac{SE_k + SE_{k'}}{SE_{kk'}}$
(AZ, CO)	1.40
(WY, CO)	1.29
(AZ, WY)	1.25

Now the average value of 1.40, 1.29, and 1.25 is 1.313. Hence, we want $z_{\frac{\alpha_A}{2}}$ such that $z_{\frac{\alpha_A}{2}}(1.313) = 1.645$ or equivalently $z_{\frac{\alpha_A}{2}} = 1.253$. For $z_{\frac{\alpha_A}{2}} = 1.25$, $100(1-\alpha_A)\% = 100(1-2(.1052))\% \approx 79\%$. Thus the 79% confidence intervals are given below.

State(k)	79% Confidence Intervals for θ_k
AZ	24.76 ± 1.25(.15) = (24.57, 24.95)
CO	24.51 ± 1.25(.19) = (24.27, 24.75)
WY	18.10 ± 1.25(.50) = (17.48, 18.73)

These 79% confidence intervals are shown in Figure 9.

For $z_{\frac{\alpha_A}{2}} = 1.253$, the level of significance for testing each pair is by (12) $\gamma_{kk'} = 2P\left(Z > z_{\frac{\alpha_A}{2}} \dfrac{SE_k + SE_{k'}}{SE_{kk'}}\right)$

Testing Pair	$\gamma_{kk'}$
(AZ, CO)	.0802
(WY, CO)	.1052
(AZ, WY)	.1164

For example, for the pair (AZ, CO),

$$\gamma_{kk'} = 2P(Z > 1.253(1.40)) = 2P(Z > 1.754) = 2(.0401) = .0802.$$

Also note that the average of the levels of significance is $(.0802 + .1052 + .1164)/3 = .1006 \approx \alpha$. Furthermore, the $100(1-\gamma_{kk'})\%$ confidence intervals for the differences are given in the table below.

Pair	$\gamma_{kk'}$	$z_{\frac{\gamma_{kk'}}{2}}$	$100(1-\alpha)\%$	$100(1-\alpha)\%$ Confidence Interval for $\theta_k - \theta_{k'}$
(AZ, CO)	.0802	1.754	92%	$(24.76 - 24.51) \pm 1.754\sqrt{(.15)^2 + (.19)^2} = (-.17, .67)$
(WY, CO)	.1052	1.616	89%	$(18.10 - 24.51) \pm 1.616\sqrt{(.50)^2 + (.19)^2} = (-7.27, -5.55)$
(AZ, WY)	.1164	1.566	88%	$(24.76 - 18.10) \pm 1.566\sqrt{(.15)^2 + (.50)^2} = (5.85, 7.47)$

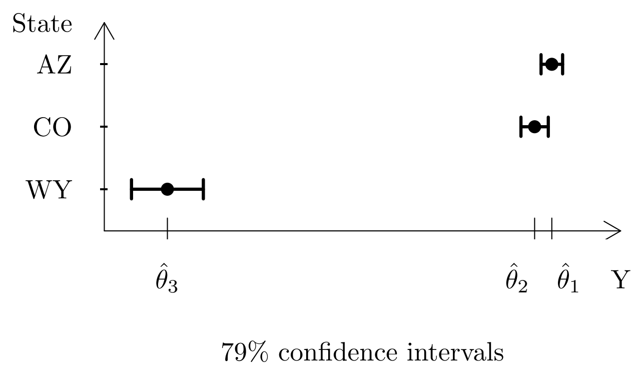


Figure 9: $100(1-\alpha_A)\% = 79\%$ Confidence Intervals for Three States: For any pair, declare states k and k' statistically significantly different at "average significance level" $\alpha = .10$ if the 79% confidence intervals for the pair k and k' do not overlap. (*Data Source:* 2011 American Community Survey)

In the discussion leading to the display in Figure 9, we considered the comparisons for three pairs of states. In a similar way, Figure 10 presents the comparisons for all pairs of states. For example, the 77.49% confidence intervals for Iowa and Kansas overlap; hence we would would not be able to say that Iowa and Kansas differ for an average significance level of $\alpha = .10$. 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 $\alpha = .10$.

$\hat{r}_k$	*State (k)*	$\hat{\theta}_k$	SE_k
51	Maryland	32.21	.15
50	New York	31.50	.09
49	New Jersey	30.53	.12
48	District of Columbia	30.10	.32
47	Illinois	28.17	.11
46	Massachusetts	27.99	.13
45	Virginia	27.74	.13
44	California	27.14	.07
43	Georgia	27.11	.17
42	New Hampshire	26.90	.30
41	Pennsylvania	25.92	.09
40	Florida	25.76	.11
39	Hawaii	25.69	.27
38	West Virginia	25.58	.31
37	Washington	25.51	.14
36	Delaware	25.30	.37
35	Connecticut	24.98	.19
34	Texas	24.82	.07
33	Arizona	24.76	.15
32	Louisiana	24.54	.15
31	Colorado	24.51	.19
30	Tennessee	24.23	.14
29	Michigan	24.11	.10
28	Nevada	24.10	.27
27	Alabama	23.94	.14
26	Mississippi	23.86	.24
25	South Carolina	23.61	.16
24	Indiana	23.45	.11
23	Maine	23.41	.25
22	North Carolina	23.37	.12
21	Rhode Island	23.36	.29
20	Ohio	23.12	.09
19	Missouri	23.07	.13
18	Minnesota	22.99	.10
17	Kentucky	22.86	.15
16	Oregon	22.54	.16
15	Vermont	21.94	.31
14	Wisconsin	21.92	.11
13	Utah	21.61	.20
12	New Mexico	21.43	.27
11	Arkansas	21.31	.23
10	Oklahoma	21.13	.15
9	Idaho	19.66	.24
8	Kansas	18.90	.16
7	Iowa	18.77	.13
6	Alaska	18.39	.33
5	Montana	18.18	.32
4	Wyoming	18.10	.50
3	Nebraska	18.06	.19
2	North Dakota	16.91	.36
1	South Dakota	16.86	.28

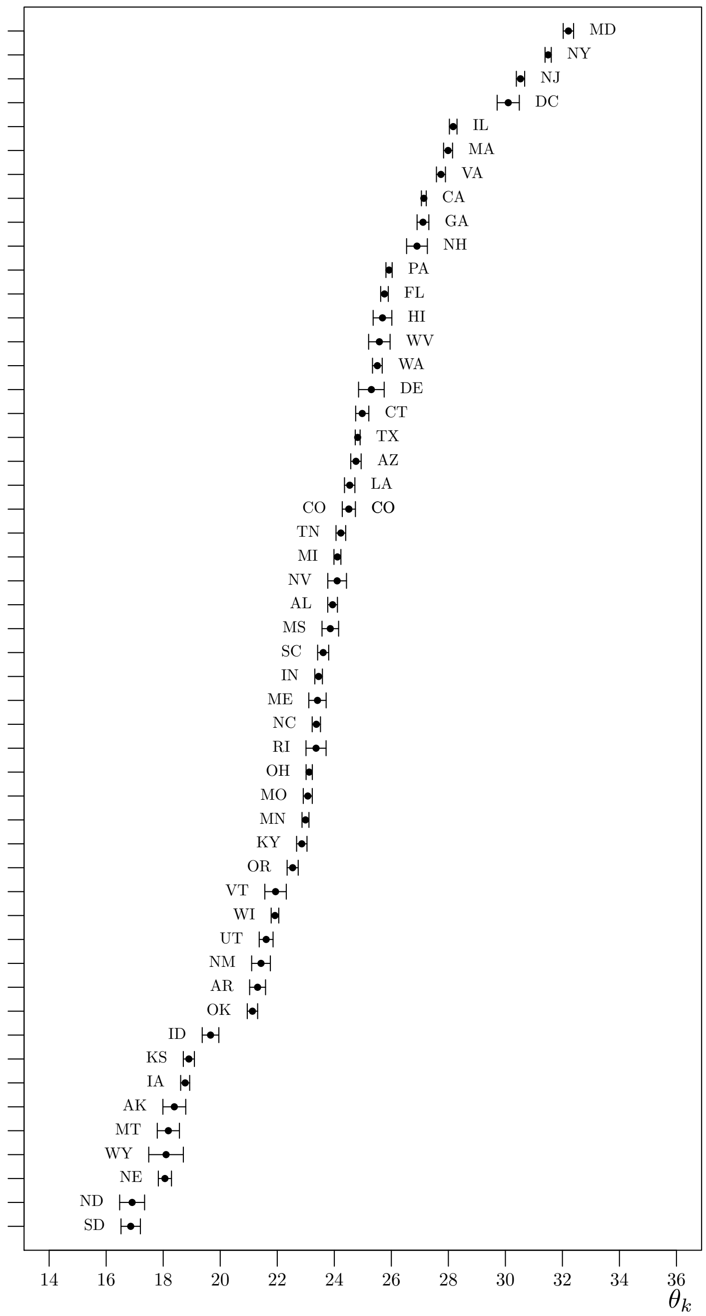


Figure 10: $100(1-\alpha_A)\% = 77.49\%$ Confidence Intervals for Each State: For any pair, declare states k and k' statistically significantly different at "average significance level" $\alpha = .10$ if $100(1-\alpha_A)\%$ confidence intervals for the pair k and k' do not overlap. (*Data Source:* 2011 American Community Survey)

3. BOOTSTRAP AND RANKING

3.1. Some Uncertainty Measures for Estimated Ranks

In the previous section, we presented uncertainty in the estimated ranking through confidence intervals and hypothesis tests for individual θ_k's, and for the pairwise differences $\theta_k - \theta_{k'}$. Alternatively, one may consider the individual ranks as the parameters of interest, and inferences can be drawn on them directly. The unknown true ranks are denoted by $r_1, r_2, ..., r_K$, 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 k^{th} smallest population as

$$r_k = \sum_{k'=1}^{K} I(\theta_{k'} \leq \theta_k) = 1 + \sum_{k':k' \neq k} I(\theta_{k'} \leq \theta_k), \quad \text{for} \quad k = 1, 2, ..., K. \tag{16}$$

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

$$\hat{r}_k = 1 + \sum_{k':k' \neq k} I(\hat{\theta}_{k'} \leq \hat{\theta}_k), \quad \text{for} \quad k = 1, 2, ..., K. \tag{17}$$

Naturally, uncertainty in the estimators $\hat{\theta}_1, \hat{\theta}_2, ..., \hat{\theta}_K$ is propagated to the estimated ranks. Therefore, an easily understandable measure of uncertainty should accompany a released ranking. Some uncertainty measures that may be useful for this purpose are as follows.

(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 $P(|\hat{r}_k - r_k| \leq c)$ 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).

In evaluating the usefulness of measures (a) - (c), two important questions emerge: how are the quantities computed (by the statistical agency) and how are they interpreted (by the data user)? These questions are addressed in the following two subsections.

3.2. Bootstrap Estimation

The bootstrap (Efron, 1979) provides a straightforward way to compute/estimate the uncertainty measures (a) - (c) of Subsection 3.1. The bootstrap has been used previously in ranking problems, for instance, by Barker, Smith, Gerzoff, Luman, McCauley, and Strine (2005); Goldstein and Speigelhalter (1996); Hall and Miller (2009, 2010); and Klein and Wright (2011). As discussed in Shao and Tu (1995), the bootstrap is a computer intensive statistical method that has broad applications. We consider both the nonparametric bootstrap and the parametric bootstrap.

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

$$\hat{F}_k(y) = \frac{1}{n_k}\sum_{i=1}^{n_k} I(Y_{ki} \le y), \quad \text{for} \quad k = 1, 2, ..., K. \tag{18}$$

Note that the empirical distribution function places equal probability on each of the observed data points $Y_{k1}, Y_{k2}, ..., Y_{kn_k}$. An estimate of a quantity such as $P\left(|\hat{r}_k - r_k| \le c\right)$ 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, measures such as (a) - (c) in Subsection 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 $Y^*_{k1}, Y^*_{k2}, ..., Y^*_{kn_k}$ as a simple random sample with replacement from $Y_{k1}, Y_{k2}, ..., Y_{kn_k}$. Do this independently for each $k = 1, 2, ..., K$.

Step 2.
(a) Compute the bootstrap analog of $\hat{\theta}_k$ which is defined as $\hat{\theta}^*_k = \hat{\theta}_k(Y^*_{k1}, Y^*_{k2}, ..., Y^*_{kn_k})$ for $k = 1, 2, ..., K$.

(b) Compute the bootstrap analog of $\hat{r}_k$ which is defined as $\hat{r}^*_k = 1 + \sum_{k':k'\neq k} I(\hat{\theta}^*_{k'} \le \hat{\theta}^*_k)$ for $k = 1, 2, ..., K$.

Step 3. Repeat Steps 1 and 2 a total of B times where B is sufficiently large (say $B = 10000$) 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 collection of bootstrap replications of ranks obtained using the algorithm above, a bootstrap estimate of $P\{|\hat{r}_k - r_k| \le c\}$ is obtained as

$$\hat{P}_{boot}\{|\hat{r}_k - r_k| \le c\} = \frac{1}{B}\sum_{b=1}^{B} I\{|\hat{r}^*_{k,b} - \hat{r}_k| \le c\}, \tag{19}$$

and a bootstrap estimate of $P\left(|\hat{r}_1 - r_1| \le c, |\hat{r}_2 - r_2| \le c, ..., |\hat{r}_K - r_K| \le c\right)$ is obtained as

$$\hat{P}_{boot}\left(|\hat{r}_1 - r_1| \le c, ..., |\hat{r}_K - r_K| \le c\right) = \frac{1}{B}\sum_{b=1}^{B} I\{|\hat{r}^*_{1,b} - \hat{r}_1| \le c, ..., |\hat{r}^*_{K,b} - \hat{r}_K| \le c\}. \tag{20}$$

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

$$[\hat{r}_k^{*(\frac{\alpha}{2})}, \ \hat{r}_k^{*(1-\frac{\alpha}{2})}] \tag{21}$$

where $\hat{r}_k^{*(\frac{\alpha}{2})}$ and $\hat{r}_k^{*(1-\frac{\alpha}{2})}$ denote, respectively, the empirical $\frac{\alpha}{2}$ and $1 - \frac{\alpha}{2}$ quantiles of the bootstrap replications $\hat{r}^*_{k,1}, \hat{r}^*_{k,2}, ..., \hat{r}^*_{k,B}$. The confidence interval (21) is called the *bootstrap percentile interval* (Efron, 1981).

Parametric Bootstrap. Sometimes we know that the sampling distribution of each of $\hat{\theta}_1, \hat{\theta}_2, ..., \hat{\theta}_K$ is well approximated by a normal distribution, and $SE_1, SE_2, ..., SE_K$, the estimated standard errors (treated as known constants) of $\hat{\theta}_1, ..., \hat{\theta}_K$, are provided. In such a situation, 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. Specifically, the parametric bootstrap algorithm proceeds as follows.

Step 1. Draw $\hat{\theta}_k^*$ from $N(\hat{\theta}_k, (SE_k)^2)$, independently for $k = 1, 2, ..., K$.

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

Step 3. Repeat Steps 1 and 2 a total of B times where B is sufficiently large (say $B = 10000$) 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.

Once we have obtained the bootstrap replications of the ranks using this procedure, estimates of the various uncertainty measures are obtained using the estimators (19) - (21). 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 $Y_{k1}, Y_{k2}, ..., Y_{kn_k}$. 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) 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. Of course, the sampling distribution of each $\hat{\theta}_k$ must be approximately normal for this procedure to be valid.

Remark 1: In the spirit of (19) - (21), we can also use the results from the previous subsection to compute various *"nonparametric or parametric bootstrap estimates of probabilities,"* such as:

P(estimated rank of state k is r_0) $= \mathrm{P}(\hat{r}_k = r_0)$ where $r_0 = 1, 2, ..., 51$;

P(estimated rank of state k among 5 highest ranks) $= \mathrm{P}(\hat{r}_k \in \{47, 48, 49, 50, 51\})$;

P(estimated rank of state k among 10 lowest ranks) $= \mathrm{P}(\hat{r}_k \in \{1, 2, ..., 10\})$;

P(estimated rank of state k is between ranks r_0 and r_{00} inclusively) $= \mathrm{P}(r_0 \leq \hat{r}_k \leq r_{00})$;

P(estimated ranks of states k and k' among 4 highest) $= \mathrm{P}(\hat{r}_k \in \{48, ..., 51\},\ \hat{r}_{k'} \in \{48, ..., 51\})$; and

P(estimated rank of state k is higher than estimated rank of state k') $= \mathrm{P}(\hat{r}_k > \hat{r}_{k'})$.

Remark 2: 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 $F_k(y|\varphi)$ for $F_k(y)$, $k = 1, 2, ..., K$, where φ is an unknown parameter vector and $F_k(y|\varphi_k)$ is known when the value of φ 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 $\hat{\varphi}_k$ is an appropriate estimate of φ_k. To be clear, this alternative form of the bootstrap requires a model assumption of the data, and not an assumption about the estimates $\hat{\theta}_k$ where the central limit theorem is more likely to apply.

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

In this subsection, we apply the parametric bootstrap procedure of Subsection 3.2 to the estimated ranking of the $K = 51$ states, based on point estimates of mean travel time to work of workers 16 years and over who did not work at home for each state. As discussed previously, these estimates are based on 2011 ACS data. Using the parametric bootstrap, we estimate the uncertainty measures (a) - (c) from Subsection 3.1; the results are reported in Tables 1 and 2 for $B = 100000$. Below we describe how one can interpret the information contained in these tables.

Consider the probabilities of the form $P\{|\hat{r}_k - r_k| \leq c\}$. Table 1 contains parametric bootstrap estimates of these probabilities for each state $k = 1, ..., K$, and for $c = 0, 1, 2, 3$. To interpret these quantities, notice that the event $|\hat{r}_k - r_k| \leq c$ is of course, equivalent to the event $\hat{r}_k - c \leq r_k \leq \hat{r}_k + c$, and hence $P\{|\hat{r}_k - r_k| \leq c\} = P\{\hat{r}_k - c \leq r_k \leq \hat{r}_k + c\}$. Therefore, noting that $1 \leq r_k \leq K$, one can think of

$$[\max\{\hat{r}_k - c, 1\}, \min\{\hat{r}_k + c, K\}] \tag{22}$$

as a confidence interval for the unknown rank r_k, where the bootstrap estimated probabilities in Table 1 give estimates of the confidence coefficient of the interval for $c = 0, 1, 2, 3$. (More technically, because r_k can only take values $\{1, 2, ..., K\}$, ignoring any complications due to ties, we should write (22) as the set $[\max\{\hat{r}_k - c, 1\}, \min\{\hat{r}_k + c, K\}] \cap \{1, 2, ..., K\}$, but for ease of presentation, we usually do not do so.) As an illustration, suppose we want a 0.90 level confidence interval for Nebraska's rank (whose estimate is $\hat{r}_k = 3$). From Table 1, we find that the estimates of $P\{|\hat{r}_k - r_k| \leq c\}$ 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. Let us also look at South Dakota, whose estimated rank is $\hat{r}_k = 1$. In this case, the estimates of $P\{|\hat{r}_k - r_k| \leq c\}$ are 0.54, 0.99, 1.00, and 1.00 for $c =$ 0, 1, 2, and 3, respectively. Thus $[1, 2]$ is an approximate level 0.99 confidence interval for the rank of South Dakota while the singleton set $\{1\}$ only has an estimated confidence coefficient of 0.54. Therefore, even though South Dakota has an estimated rank of 1, it seems to be more reasonable to conclude that its rank could be either 1 or 2.

Next consider the quantities $\hat{r}_k^{*(.05)}$ and $\hat{r}_k^{*(.95)}$, also displayed in the last two columns of Table 1 for each state $k = 1, ..., K$. 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_k. Thus, based on this method, we find that a 0.90 level confidence interval for the rank of Nebraska is $[3, 6]$, which is of course, 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 estimated rank $\hat{r}_k = 3$ is much closer to those states with point estimated ranks $\hat{r}_{k'} = 4, 5$, or 6 than those states with point estimated ranks $\hat{r}_{k'} = 1$ or 2. So the symmetric $\hat{r}_k \pm 2$ confidence interval is quite different from the equal tail bootstrap percentile interval. For South Dakota, the 0.90 bootstrap percentile confidence interval is $[1, 2]$, which in this case is the same as the approximate 0.99 level confidence interval for South Dakota's rank that was reported in the previous paragraph.

Finally, let us examine the estimates of the joint probability $P(|\hat{r}_1 - r_1| \leq c, |\hat{r}_2 - r_2| \leq c, ..., |\hat{r}_K - r_K| \leq c)$ presented in Table 2 for $c = 0, 1, 2, ..., 8$. One can interpret these 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:

$$[\max\{\hat{r}_1 - c, 1\}, \min\{\hat{r}_1 + c, K\}] \times \cdots \times [\max\{\hat{r}_K - c, 1\}, \min\{\hat{r}_K + c, K\}]. \tag{23}$$

For example, we see from Table 2 that with $c = 5$ the estimated confidence coefficient of the above region (23) 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.

In summary, the uncertainty measures presented in Tables 1 and 2 have fairly straightforward interpretations in terms of marginal or joint confidence statements. Furthermore, the estimates can be readily computed using the parametric bootstrap. A nonparametric bootstrap procedure could also be used for estimation, and all interpretations would remain the same. Thus, the quantities presented in the Tables 1 and 2 are promising measures for statistical agencies to use for conveying the uncertainty in an estimated ranking. Notice that we have two reasonable methods for obtaining an approximate confidence interval on an individual rank r_k, namely, (i) take the interval as (22) and use the bootstrap to estimate the confidence coefficient, and (ii) the bootstrap percentile confidence interval given by (21). The question of which of these intervals is preferable requires further investigation and we will not pursue it here.

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

Estimated				$\widehat{P}_{\text{boot}}\{\lvert\hat{r}_k - r_k\rvert \leq c\}$				90% Confidence Interval	
Rank($\hat{r}_k$)	State(k)	$\hat{\theta}_k$	SE_k	$c = 0$	$c = 1$	$c = 2$	$c = 3$	$\hat{r}_k^{*(.05)}$	$\hat{r}_k^{*(.95)}$
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	0.52	1.00	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 $P(|\hat{r}_1 - r_1| \leq c, |\hat{r}_2 - r_2| \leq c, ..., |\hat{r}_K - r_K| \leq c)$

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

3.4. Bootstrap-An Empirical Investigation Using PUMS Data

Continuing to use 2011 American Community Survey data, we further illustrate the discussion of the bootstrap with an empirical investigation of three bootstrap variations, two of which are nonparametric and one which is parametric. Before considering these three bootstrap variations, in Figure 11 we share a brief look at the unweighted state distributions of reported travel times from sample respondents using 2011 ACS sample data available in a publically available microdata file referred to as the *Public Use Microdata Sample* (PUMS). The distributions are unweighted, and $\hat{r'}_k$ is determined by the ordered values of $\hat{\theta}'_k$ as calculated using ACS PUMS data. In an attempt to emphasize that the results in Section 3.4 are based on PUMS data, we use a prime ($'$) on $\hat{r}'_k$ and $\hat{\theta}'_k$ to distinguish them from all other results in this paper. As a consequence, the numerical results in this subsection differ slightly from those of all other sections of this paper. For ease of reading, some notation in this subsection may appear to be the same as in the other subsections, but the notation should be thought of as in the PUMS data context.

Nonparametric

(i) Ignoring Sampling Weights: For each state $k = 1, 2, ..., 51$, we take B simple random samples with replacement of size n_k from state k's PUMS and calculate from the b^{th} sample the *unweighted mean*

$$\hat{\theta}^{*(u)}_{kb} = \frac{1}{n_k} \sum_{i=1}^{n_k} Y^*_{kib} \quad . \tag{24}$$

This approach completely ignores the original sampling design and the accompanying sampling weights, with any adjustments.

(ii) Using Sampling Weights: Again, for each state $k = 1, 2, ..., 51$, we take B simple random samples with replacement of size n_k from state k's PUMS. For the i^{th} individual in the b^{th} sample from state k, we record not only the individual travel times Y^*_{kib}, but we also record the associated sampling weights w^*_{kib}. Then for the b^{th} sample, we calculate the *weighted mean*

$$\hat{\theta}^{*(w)}_{kb} = \frac{\sum_{i=1}^{n_k} w^*_{kib} Y^*_{kib}}{\sum_{i=1}^{n_k} w^*_{kib}} \quad . \tag{25}$$

This approach does not ignore the original sampling design and the accompanying sampling weights, and it is closer to the advice in the literature, see e.g., Shao and Tu(1995).

Parametric

Sampling Weights Incorporated in Estimates: For each state $k = 1, 2, ..., 51$, we take B draws from a normal distribution whose mean and standard deviation are determined by the PUMS based data mean $\hat{\theta}'_k$ and SE'_k, and we note the result of the b^{th} draw from $N(\hat{\theta}'_k, SE'_k)$ by

$$\hat{\theta}^{*(p)}_{kb} \quad . \tag{26}$$

Of the three approaches that we consider in this section, the parametric one retains the least information about the underlying data, but it also is the least computationally intensive.

Now, we can compare the three approaches in several ways:

- For a given state k, we present the distributions of $\hat{\theta}_{kb}^{*(u)}, \hat{\theta}_{kb}^{*(w)}$, and $\hat{\theta}_{kb}^{*(p)}$ across the replications $b = 1, 2, ..., B$. Specifically, we seek to answer the question: Are either $\hat{\theta}_{kb}^{*(p)}$ or $\hat{\theta}_{kb}^{*(u)}$ distributions good approximations to the *"more appropriate"* bootstrap distribution of $\hat{\theta}_{kb}^{*(w)}$?

- Alternately and according to each of the three approaches, for each bootstrap replication $b = 1, 2, ..., B$, rank the 51 states. We use the distribution of B sets of 51 ranks for each approach to make inferences about the true ranks, and determine whether these inferences differ much across the three approaches. For example, are the bootstrap 90% confidence intervals (percentile method) as in (21) for the state ranks similar across the three methods? Again, are the results of the less-computationally-intensive parametric bootstrap or unweighted nonparametric bootstrap similar to the *"more appropriate"* weighted nonparametric bootstrap?

If the parametric bootstrap's results are not practically different from the weighted nonparametric results, it is the preferred method because it requires less computation, less code to debug, and no need to access or store the microdata.

Remark: We believe that differences among the three methods should not be judged by statistical significance because the only relevant error is the Monte Carlo error due to using a finite B, which can be made arbitarily small with large enough B. Instead, we must decide how much of a difference among the approaches is *practically* significant. For example, we may decide that if all states's bootstrap 90% confidence intervals for ranks differ by only one or two ranks from method to method, this difference is not practically significant.

In Figure 12, we compare the results of the three bootstrap distributions using mean travel time to work data during 2011 from the ACS PUMS for the state of Colorado. For Colorado, the parametric bootstrap distribution of $\hat{\theta}_{kb}^{*(p)}$ is clearly a better fit to the bootstrap distribution of $\hat{\theta}_{kb}^{*(w)}$ than the distribution of $\hat{\theta}_{kb}^{*(u)}$ using unweighted PUMS data. The same holds for all states as shown in Figure 13, which shows the three bootstrap distributions for each of the 51 states.

In Table 3, we see, for the most part, that each state's 90% confidence intervals (percentile method) of rank are practically the same across the three approaches, but there are differences. More specifically, the confidence intervals based on $\hat{\theta}_{k}^{*(w)}$ (weighted PUMS data) and $\hat{\theta}_{k}^{*(p)}$ (parametric using PUMS based weighted estimates) are quite similar, and they tend to differ from those confidence intervals based on $\hat{\theta}_{k}^{*(u)}$ (unweighted PUMS data), see for example, the three 90% bootstrap percentile confidence intervals in Table 3 for the states AK, VT, WI, and ME. We believe that Table 3 provides evidence against the use of unweighted PUMS data and an argument for the use of the parametric approach.

$\hat{r}'_k$	$State_k$	$Median_k$	$\hat{\theta}'_k$	Max_k
51	Maryland	29	32.33	146
50	New York	25	31.61	142
49	District of Columbia	30	30.65	136
48	New Jersey	25	30.63	145
47	Illinois	20	28.08	143
46	Massachusetts	20	28.00	144
45	Virginia	20	27.85	150
44	Georgia	20	27.23	148
43	California	20	27.17	142
42	New Hampshire	20	27.11	152
41	Pennsylvania	20	26.04	152
40	Florida	20	25.81	125
39	Hawaii	20	25.55	107
38	Washington	20	25.46	144
37	Delaware	20	25.41	144
36	West Virginia	20	25.20	148
35	Connecticut	20	24.80	144
34	Texas	20	24.78	146
33	Colorado	20	24.70	163
32	Arizona	20	24.69	120
31	Louisiana	20	24.37	163
30	Tennessee	20	24.36	122
29	Michigan	20	24.19	150
28	Alabama	20	24.16	126
27	Nevada	20	24.16	154
26	Mississippi	20	23.84	162
25	Maine	20	23.72	157
24	South Carolina	20	23.55	129
23	Rhode Island	20	23.52	140
22	North Carolina	20	23.41	143
21	Indiana	20	23.40	123
20	Missouri	20	23.23	147
19	Ohio	20	23.19	128
18	Minnesota	20	23.12	151
17	Kentucky	20	22.86	129
16	Oregon	17	22.63	125
15	Vermont	20	22.07	153
14	Wisconsin	20	22.04	148
13	Arkansas	15	21.59	131
12	Utah	15	21.41	138
11	New Mexico	15	21.35	119
10	Oklahoma	15	21.28	139
9	Idaho	15	19.66	131
8	Kansas	15	19.02	132
7	Iowa	15	18.60	133
6	Montana	15	18.26	145
5	Alaska	10	18.20	145
4	Nebraska	15	18.11	132
3	North Dakota	10	17.87	163
2	Wyoming	10	17.47	156
1	South Dakota	15	17.29	129

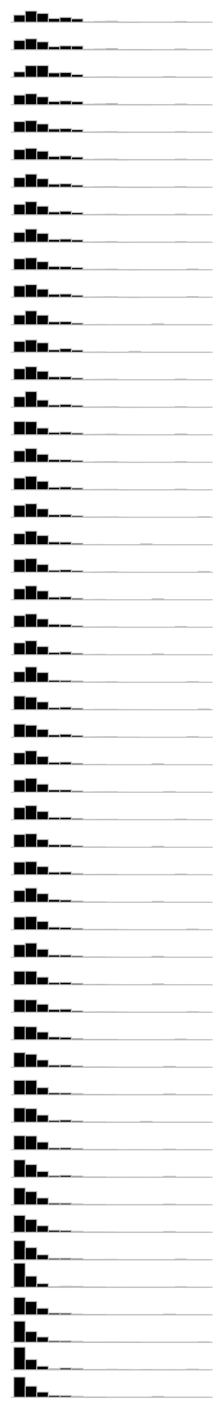


Figure 11: Unweighted PUMS Data Distributions for All 51 States for Mean Travel Time to Work of Workers 16 Years and Over Who Did Not Work at Home (Minutes). Every State Had At least One Individual with a Travel to Work Time of One Minute in the PUMS. (*Data Source:* 2011 American Community Survey PUMS)

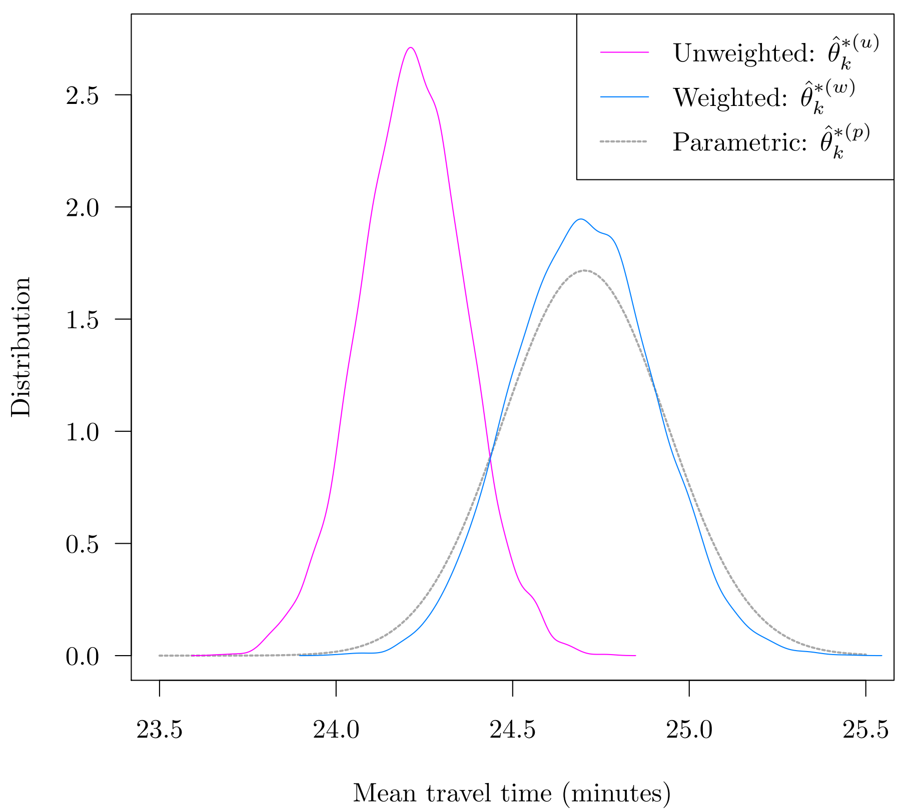


Figure 12: Three Bootstrap Distributions for Colorado for Mean Travel Time to Work of Workers 16 Years and Over Who Did Not Work at Home (Minutes). (*Data Source:* 2011 American Community Survey PUMS)

$\hat{r}'_k$	*State (k)*	$\hat{\theta}'_k$	SE'_k
51	Maryland	32.33	.18
50	New York	31.61	.12
49	District of Columbia	30.65	.43
48	New Jersey	30.63	.18
47	Illinois	28.08	.14
46	Massachusetts	28.00	.15
45	Virginia	27.85	.15
44	Georgia	27.23	.16
43	California	27.17	.08
42	New Hampshire	27.11	.46
41	Pennsylvania	26.04	.14
40	Florida	25.81	.11
39	Hawaii	25.55	.37
38	Washington	25.46	.16
37	Delaware	25.41	.49
36	West Virginia	25.20	.40
35	Connecticut	24.80	.21
34	Texas	24.78	.08
33	Colorado	24.70	.23
32	Arizona	24.69	.16
31	Louisiana	24.37	.19
30	Tennessee	24.36	.17
29	Michigan	24.19	.14
28	Alabama	24.16	.18
27	Nevada	24.16	.29
26	Mississippi	23.84	.28
25	Maine	23.72	.43
24	South Carolina	23.55	.16
23	Rhode Island	23.52	.38
22	North Carolina	23.41	.15
21	Indiana	23.40	.13
20	Missouri	23.23	.18
19	Ohio	23.19	.11
18	Minnesota	23.12	.17
17	Kentucky	22.86	.21
16	Oregon	22.63	.20
15	Vermont	22.07	.46
14	Wisconsin	22.04	.19
13	Arkansas	21.59	.31
12	Utah	21.41	.22
11	New Mexico	21.35	.29
10	Oklahoma	21.28	.22
9	Idaho	19.66	.29
8	Kansas	19.02	.23
7	Iowa	18.60	.19
6	Montana	18.26	.41
5	Alaska	18.20	.44
4	Nebraska	18.11	.30
3	North Dakota	17.87	.70
2	Wyoming	17.47	.46
1	South Dakota	17.29	.43

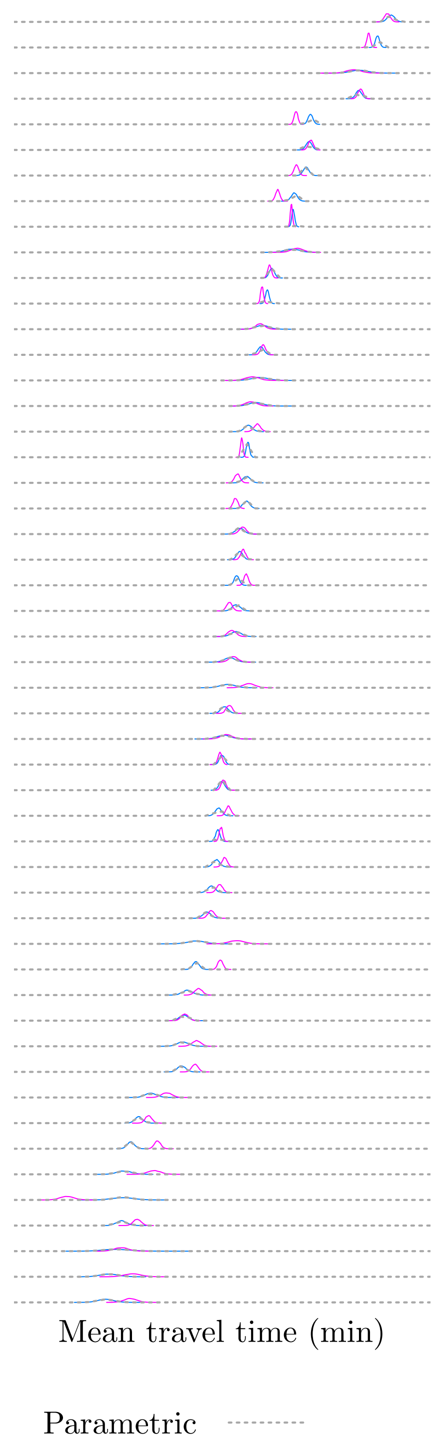


Figure 13: Three Bootstrap Distributions for All 51 States for Mean Travel Time to Work of Workers 16 Years and Over Who Did Not Work at Home (Minutes). (*Data Source:* 2011 American Community Survey PUMS)

Table 3: Three Different State 90% Bootstrap Confidence Intervals for Ranks Using ACS PUMS

$\hat{r}'_k$	$State_k$	$\hat{\theta}'_k$	SE'_k	Unweighted	Weighted	Parametric
51	MD	32.33	.18	[51, 51]	[51, 51]	[51, 51]
50	NY	31.61	.12	[50, 50]	[50, 50]	[50, 50]
49	DC	30.65	.43	[48, 49]	[48, 49]	[48, 49]
48	NJ	30.63	.18	[48, 49]	[48, 49]	[48, 49]
47	IL	28.08	.14	[44, 46]	[45, 47]	[45, 47]
46	MA	28.00	.15	[47, 47]	[45, 47]	[45, 47]
45	VA	27.85	.15	[44, 46]	[45, 47]	[44, 47]
44	GA	27.23	.16	[42, 42]	[42, 44]	[42, 44]
43	CA	27.17	.08	[43, 44]	[42, 44]	[42, 44]
42	NH	27.11	.46	[43, 46]	[42, 44]	[42, 45]
41	PA	26.04	.14	[40, 41]	[40, 41]	[39, 41]
40	FL	25.81	.11	[37, 40]	[38, 41]	[38, 41]
39	HI	25.55	.37	[35, 40]	[36, 41]	[36, 41]
38	WA	25.46	.16	[37, 40]	[36, 39]	[36, 39]
37	DE	25.41	.49	[30, 39]	[33, 41]	[32, 41]
36	WV	25.20	.40	[31, 38]	[32, 39]	[32, 40]
35	CT	24.80	.21	[35, 39]	[31, 36]	[31, 36]
34	TX	24.78	.08	[29, 32]	[32, 36]	[32, 36]
33	CO	24.70	.23	[25, 31]	[30, 36]	[30, 36]
32	AZ	24.69	.16	[25, 29]	[30, 36]	[31, 36]
31	LA	24.37	.19	[28, 34]	[27, 33]	[27, 33]
30	TN	24.36	.17	[29, 34]	[27, 32]	[27, 32]
29	MI	24.19	.14	[31, 35]	[26, 31]	[26, 31]
28	AL	24.16	.18	[20, 26]	[25, 31]	[25, 31]
27	NV	24.16	.29	[21, 29]	[25, 32]	[25, 32]
26	MS	23.84	.28	[21, 30]	[22, 30]	[22, 29]
25	ME	23.72	.43	[29, 37]	[19, 30]	[18, 30]
24	SC	23.55	.16	[20, 26]	[20, 26]	[21, 26]
23	RI	23.52	.38	[15, 27]	[18, 27]	[17, 28]
22	NC	23.41	.15	[15, 19]	[19, 25]	[19, 25]
21	IN	23.40	.13	[16, 22]	[19, 25]	[19, 24]
20	MO	23.23	.18	[20, 26]	[17, 23]	[17, 23]
19	OH	23.19	.11	[15, 20]	[18, 22]	[18, 22]
18	MN	23.12	.17	[17, 23]	[17, 22]	[17, 22]
17	KY	22.86	.21	[15, 20]	[16, 19]	[16, 20]
16	OR	22.63	.20	[14, 15]	[15, 17]	[15, 17]
15	VT	22.07	.46	[20, 34]	[11, 16]	[11, 17]
14	WI	22.04	.19	[15, 20]	[13, 15]	[13, 15]
13	AR	21.59	.31	[11, 13]	[10, 14]	[10, 15]
12	UT	21.41	.22	[10, 10]	[10, 13]	[10, 13]
11	NM	21.35	.29	[11, 13]	[10, 14]	[10, 14]
10	OK	21.28	.22	[11, 13]	[10, 13]	[10, 13]
9	ID	19.66	.29	[8, 9]	[8, 9]	[8, 9]
8	KS	19.02	.23	[5, 7]	[7, 8]	[7, 8]
7	IA	18.60	.19	[7, 9]	[5, 7]	[5, 8]
6	MT	18.26	.41	[6, 9]	[3, 7]	[2, 7]
5	AK	18.20	.44	[1, 1]	[2, 8]	[2, 7]
4	NE	18.11	.30	[3, 5]	[2, 6]	[2, 6]
3	ND	17.87	.70	[2, 4]	[1, 8]	[1, 8]
2	WY	17.47	.46	[2, 6]	[1, 5]	[1, 5]
1	SD	17.29	.43	[2, 5]	[1, 4]	[1, 4]

3.5. Simulation Study for Evaluating Parametric Bootstrap for Estimating Uncertainty in Ranking

We continue to make use of the same ACS data as used in Tables 1 and 2. For our simulation study to evaluate the parametric bootstrap, fix values for the unknown parameters $\theta_1, ..., \theta_K$, and of the known parameters $SE_1, ..., SE_K$. These known values are given in columns 3 and 4 of Table 4. Compute the ranks as $r_k = 1 + \sum_{k' \neq k} I(\theta_{k'} \leq \theta_k)$, for $k = 1, ..., K$. The ranks r_k represent the truth, and they are given in column 1 of Table 4.

Now to evaluate the parametric bootstrap method for ranking described in Subsection 3.2,

1. Draw $\hat{\theta}_1, ..., \hat{\theta}_K$ independently such that $\hat{\theta}_k \sim N(\theta_k, (SE_k)^2)$, for $k = 1, ..., K$.
2. Compute the estimated ranks $\hat{r}_k = 1 + \sum_{k' : k' \neq k} I(\hat{\theta}_{k'} \leq \hat{\theta}_k)$, for $k = 1, ..., K$.
3. Compute a set of bootstrap replications of the ranks as follows.
   (a) Draw $\hat{\theta}^*_k \sim N(\hat{\theta}_k, (SE_k)^2)$, independently for $k = 1, ..., K$.
   (b) Compute $\hat{r}^*_k = 1 + \sum_{k' \neq k} I(\hat{\theta}^*_{k'} \leq \hat{\theta}^*_k)$, for $k = 1, ..., K$.
   (c) Repeat Steps (a) and (b) a total of B times to get $(\hat{r}^*_{1,1}, ..., \hat{r}^*_{K,1}), ..., (\hat{r}^*_{1,B}, ..., \hat{r}^*_{K,B})$.
4. Define the $100(1-\alpha)\%$ bootstrap percentile confidence interval for r_k as $[\hat{r}_k^{*\left(\frac{\alpha}{2}\right)}, \hat{r}_k^{*\left(1-\frac{\alpha}{2}\right)}]$ where $\hat{r}_k^{*\left(\frac{\alpha}{2}\right)}$ and $\hat{r}_k^{*\left(1-\frac{\alpha}{2}\right)}$ denote the empirical $\frac{\alpha}{2}$ and $1 - \frac{\alpha}{2}$ quantiles of the bootstrap replications $\hat{r}^*_{k,1}, ..., \hat{r}^*_{k,B}$.
5. Compute $C_k = I\left[\hat{r}_k^{*\left(\frac{\alpha}{2}\right)} \leq r_k \leq \hat{r}_k^{*\left(1-\frac{\alpha}{2}\right)}\right]$, for $k = 1, ..., K$.
6. Repeat steps 1 - 5 a total of M times to get $(C_{1,1}, ..., C_{K,1}), ..., (C_{1,M}, ..., C_{K,M})$. The simulation based estimate of coverage probability of the bootstrap percentile confidence interval for r_k is $\overline{C}_k = \frac{1}{M}\sum_{i=1}^{M} C_{k,i}$, for $k = 1, ..., K$. These values are given in the last column of Table 4.

From the last column of Table 4, we see that our estimates of coverage probability of the bootstrap percentile confidence interval for r_k (i.e., $\bar{C}_k$) is at least 90% for $k = 1, 2, ..., 51$. Thus based on simulated estimates of coverage probability, the bootstrap percentile confidence interval for r_k is extremely good, for $k = 1, 2, ..., 51$.

Table 4: Simulated Coverage Probabilities of the 90% Boostrap Percentile Interval with $B = 1000$ and $M = 10000$

r_k	State	θ_k	SE_k	$\bar{C}_k$
51	Maryland	32.21	0.15	100.00
50	New York	31.50	0.09	100.00
49	New Jersey	30.53	0.12	99.81
48	District of Columbia	30.10	0.32	99.81
47	Illinois	28.17	0.11	99.70
46	Massachusetts	27.99	0.13	99.74
45	Virginia	27.74	0.13	99.92
44	California	27.14	0.07	92.28
43	Georgia	27.11	0.17	98.33
42	New Hampshire	26.90	0.30	97.98
41	Pennsylvania	25.92	0.09	95.15
40	Florida	25.76	0.11	92.39
39	Hawaii	25.69	0.27	95.10
38	West Virginia	25.58	0.31	95.24
37	Washington	25.51	0.14	97.36
36	Delaware	25.30	0.37	95.49
35	Connecticut	24.98	0.19	98.25
34	Texas	24.82	0.07	99.48
33	Arizona	24.76	0.15	97.18
32	Louisiana	24.54	0.15	97.43
31	Colorado	24.51	0.19	95.96
30	Tennessee	24.23	0.14	95.98
29	Michigan	24.11	0.10	96.51
28	Nevada	24.10	0.27	94.18
27	Alabama	23.94	0.14	98.24
26	Mississippi	23.86	0.24	95.44
25	South Carolina	23.61	0.16	96.66
24	Indiana	23.45	0.11	94.41
23	Maine	23.41	0.25	94.41
22	North Carolina	23.37	0.12	97.06
21	Rhode Island	23.36	0.29	90.80
20	Ohio	23.12	0.09	98.61
19	Missouri	23.07	0.13	96.67
18	Minnesota	22.99	0.10	96.71
17	Kentucky	22.86	0.15	97.49
16	Oregon	22.54	0.16	99.85
15	Vermont	21.94	0.31	94.33
14	Wisconsin	21.92	0.11	99.43
13	Utah	21.61	0.20	98.18
12	New Mexico	21.43	0.27	97.20
11	Arkansas	21.31	0.23	96.91
10	Oklahoma	21.13	0.15	96.80
9	Idaho	19.66	0.24	100.00
8	Kansas	18.90	0.16	97.16
7	Iowa	18.77	0.13	99.36
6	Alaska	18.39	0.33	94.22
5	Montana	18.18	0.32	95.95
4	Wyoming	18.10	0.50	95.30
3	Nebraska	18.06	0.19	84.15
2	North Dakota	16.91	0.36	96.74
1	South Dakota	16.86	0.28	95.72

3.6. Additional Examples of Parametric Bootstrap Estimates of Probabilities

To illustrate the great wealth of estimates of measures possible with rankings, the following estimated probabilities were computed using a parametric bootstrap with $B = 100,000$.

$\hat{P}$(estimated rank of Colorado is 31)

$$\begin{aligned} &= \hat{P}(\hat{r}_{31} = 31) \\ &= \frac{1}{B}\sum_{b=1}^{B} I(\hat{r}^*_{31,b} = 31) \\ &= 0.367 \end{aligned}$$

$\hat{P}$(estimated rank of District of Columbia is among the 5 highest ranks)

$$\begin{aligned} &= \hat{P}(\hat{r}_{48} \in \{47, 48, 49, 50, 51\}) \\ &= \frac{1}{B}\sum_{b=1}^{B} I(\hat{r}^*_{48,b} \in \{47, 48, 49, 50, 51\}) \\ &= 1.000 \end{aligned}$$

$\hat{P}$(estimated rank of Kansas is among the 10 lowest ranks)

$$\begin{aligned} &= \hat{P}(\hat{r}_{8} \in \{1, 2, ..., 10\}) \\ &= \frac{1}{B}\sum_{b=1}^{B} I(\hat{r}^*_{8,b} \in \{1, 2, ..., 10\}) \\ &= 1.000 \end{aligned}$$

$\hat{P}$(estimated rank of Arkansas is among the 10 lowest ranks)

$$\begin{aligned} &= \hat{P}(\hat{r}_{11} \in \{1, 2, ..., 10\}) \\ &= \frac{1}{B}\sum_{b=1}^{B} I(\hat{r}^*_{11,b} \in \{1, 2, ..., 10\}) \\ &= 0.222 \end{aligned}$$

$\hat{P}$(estimated rank of Colorado is between 29 and 32)

$$\begin{aligned} &= \hat{P}(29 \leq \hat{r}_{31} \leq 32) \\ &= \frac{1}{B}\sum_{b=1}^{B} I(29 \leq \hat{r}^*_{31,b} \leq 32) \\ &= 0.820 \end{aligned}$$

$\hat{P}$(estimated ranks of District of Columbia and New Jersey are among the 4 highest)

$$\begin{aligned} &= \hat{P}(\hat{r}_{48} \in \{48, ..., 51\}, \hat{r}_{49} \in \{48, ..., 51\}) \\ &= \frac{1}{B}\sum_{b=1}^{B} I(\hat{r}^*_{48,b} \in \{48, ..., 51\},\ \hat{r}^*_{49,b} \in \{48, ..., 51\}) \\ &= 1.000 \end{aligned}$$

$\hat{P}$(estimated rank of Delaware is higher than estimated rank of Connecticut)

$$\begin{aligned} &= \hat{P}(\hat{r}_{36} > \hat{r}_{35}) \\ &= \frac{1}{B}\sum_{b=1}^{B} I(\hat{r}^*_{36,b} > \hat{r}^*_{35,b}) \\ &= 0.777 \end{aligned}$$

4. CONCLUDING COMMENTS

The methods presented in Section 2 and the uncertainty measures in Section 3 are simple and easy to use. They are robust and can be widely understood. For implementation, they mainly require K sample estimates $\hat{\theta}_k$ and their associated standard errors SE_k, for $k = 1, 2, ..., K$. Theory exists to support their use. Further research will reveal more theoretical properties of these methods and evaluate them empirically. They will also be compared with each other as well as with other methods. Perhaps most importantly, we believe that this paper and others will advance the need for national statistical agencies to express uncertainly in released rankings based on data from sample surveys.

Disclaimer: 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.

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.

Bechhofer, R. E. (1954). "A Single-sample Multiple Decision Procedure for Ranking Means of Normal Populations with Known Variances", *Annals of Mathematical Statistics, 25*, 16-29.

Berger, J. O. and Deely, J. (1988). "A Bayesian Approach to Ranking and Selection of Related Means with Alternatives to Analysis-of-Variance Methodology, *Journal of the American Statistical Association, 83*, 364-373.

Bonferroni, C. E. (1935). "Il Calcolo delle Assicurazioni su Gruppi di Teste", in *Studi in Onore del Professore Salvatore Ortu Carboni*, Rome, Italy, 13-60.

Bonferroni, C. E. (1936). "Teoria Statistica delle Classi e Calcolo delle Probabilita", *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze, 8*, 3-62.

Cochran, W. G. (1977). *Sampling Techniques (3rd Edition)*, New York, NY: John Wiley & Sons.

Dudewicz, E. J. (1980). “Ranking (Ordering) and Selection: An Overview of How to Select the Best”, *Technometrics, 22*, 113-119.

Dunn, O. J. (1961). “Multiple Comparisons among Means”, *Journal of the American Statistical Association, 56*, 52-64.

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.

Efron, B. and Gong, G. (1983). “A Leisurely Look at the Bootstrap, the Jackknife, and Cross-Validation”, *The American Statistician, Vol. 37, No. 1*, 36-48.

Efron, B. and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*, London: Chapman and Hall.

Frey, J. (2008). “A Review of Methods for Ranking”, Unpublished Report Prepared under Contract with U. S. Census Bureau, Washington, D.C.

Fuller, W. A. (2009). *Sampling Statistics*, Hoboken, NJ: John Wiley & Sons.

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.

Govindarajulu, Z. and Harvey, C. (1974). “Bayesian Procedures for Ranking and Selection Problems”, *Annals of the Institute of Statistical Mathematics, 26*, 35-53.

Gupta, S. S. (1965). “On Some Multiple Decision (Selection and Ranking) Rules”, *Technometrics, 7*, 225-245.

Gupta, S. S. and McDonald, G. C. (1970). “On Some Classes of Selection Procedures Based on Ranks”, 491-514 in Puri, M. L. (ed.) *Nonparametric Techniques in Statistical Inference*, Cambridge: Cambridge University Press.

Hall, P. and Miller, H. (2009). “Using the Bootstrap to Quantify the Authority of An Empirical Ranking”, *The Annals of Statistics, Vol 37, No. 6B*, 3929-3959.

Hall, P. and Miller, H. (2010). “Modeling the Variability of Rankings”, *The Annals of Statistics, Vol 38*, 2652-2677.

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.

Larsen, R. J. and Marx, M. L. (2012). *An Introduction to Mathematical Statistics and Its Applications*, Boston, MA: Prentice Hall.

Lohr, S. L. (2010). *Sampling: Design and Analysis (2^{nd} Edition)*, Boston, MA: Brooks/Cole.

Louis, T. A. (1984). "Estimating a Population of Parameter Values Using Bayes and Empirical Bayes Methods", *Journal of the American Statistical Association, 79*, 393-398.

Mann, H. B. and Whitney, D. R. (1947). "On a Test of Whether One of Two Random Variables is Stochastically Larger than the Other", *Annals of Mathematical Statistics, 18,* 50-60.

McDonald, G. C. (1973). "The Distribution of Some Rank Statistics with Applications in Block Design Selection Problems", *Sankhya A, 35*, 187-204.

McDonald, G. C. (1979). "Nonparametric Selection Procedures Applied to State Traffic Fatality Rates", *Technometrics, 21*, 515-523.

Mosteller, F. (1948). "A k-sample Slippage Test for an Extreme Population", *Annals of Mathematical Statistics, 19*, 58-65.

Panchapakesan, S.(2006), "Ranking and Selection Procedures", in *Encyclopedia of Statistical Sciences, Vol 10*, Hoboken, NJ: John Wiley & Sons, 6907-6915.

Paulson, E. (1949). "A Multiple Decision Procedure for Certain Problems in the Analysis of Variance", *Annals of Mathematical Statistics, 20*, 95-98.

Paulson, E. (1952a). "On the Comparison of Several Experimental Categories with a Control", *Annals of Mathematical Statistics, 23*, 239-246.

Paulson, E. (1952b). "An Optimum Solution to the k-sample Slippage Problem for the Normal Distribution", *Annals of Mathematical Statistics, 23*, 610-616.

Särndal, C.-E., Swensson, B., and Wretman, J. (2003). *Model Assisted Survey Sampling*, New York, NY: Springer.

Schenker, N. and Gentleman, J. F. (2001). "On Judging the Significance of Differences by Examining the Overlap between Confidence Intervals", *The American Statistician, 55:3*, 182-186.

Shen, W. and Louis, T. A. (1998). "Triple-goal Estimates in Two-statge Hierarchical Models", *Journal of the Royal Statistical Society, Series B, 60*, 455-471.

Shao, J. and Tu, D. (1995). *The Jackknife and Bootstrap*, New York, NY: Springer-Verlag.

Valliant, R., Dorfman, A. H., and Royall, R. M. (2000). *Finite Population Sampling and inference: A Prediction Approach*, New York, NY: John wiley & Sons.

Wilcoxon, F. (1945). “Individual Comparisons by Ranking Methods”, *Biometrics, 1*, 80-83.

Wright, T., Klein, M., and Wieczorek, J. (2013). “An Overview of Some Concepts for Potential Use in Ranking Populations Based on Sample Survey Data”, *Proceedings of the 2013 World Congress of Statistics*, International Statistical Institute, Hong Kong, China.

APPENDIX A

Listing of 85 Ranking Tables Based on the 2011 American Community Survey

R0201	Percent of the Total Population Who Are White Alone
R0202	Percent of the Total Population Who Are Black or African American Alone
R0203	Percent of the Total Population Who Are American indian and Alaska Native Alone
R0204	Percent of the Total Population Who Are Asian Alone
R0205	Percent of the Total Population Who Are Native Hawaiian and Other Pacific Islander Alone
R0206	Percent of the Total Population Who Are Some Other Race Alone
R0207	Percent of the Total Population Who Are Two or More Races
R0208	Percent of the Total Population Who Are Two or More Races Excluding Some Other Race
R0209	Percent of the Total Population Who Are White Alone, Not Hispanic or Latino
R0501	Percent of People Who Are Foreign Born
R0502	Percent of People Born in Europe
R0503	Percent of People Born in Asia
R0504	Percent of People Born in Latin America
R0505	Percent of People Born in Mexico
R0601	Percent of the Native Population Born in Their State of Residence (including Puerto Rico)
R0701	Percent of People 1 Year and Over Who Lived in a Different House in Either the U.S. or Puerto Rico 1 Year Ago
R0702	Percent of People 1 Year and Over Who Lived in a Different House within the Same State (including Puerto Rico) 1 Year Ago
R0703	Percent of People 1 Year and Over Who Lived in a Different State (including Puerto Rico) 1 Year Ago
R0801	Mean Travel Time to Work of Workers 16 Years and Over Who Did Not Work at Home (Minutes)
R0802	Percent of Workers 16 Years and Over Who Traveled to Work by Car, Truck, or Van-Drove Alone
R0803	Percent of Workers 16 Years and Over Who Traveled to Work by Car, Truck, or Van-Carpooled
R0804	Percent of Workers 16 Years and Over Who Traveled to Work by Public Transportation (Excluding Taxicab)
R0805	Percent of Workers 16 Years and Over Who Worked Outside County of Residence
R1001	Percent of Grandparents Responsible for Their Grandchildren
R1101	Percent of Households That Are Married-Couple Families
R1102	Percent of Households That Are Married-Couple Families With Own Children Under 18 Years
R1103	Percent of Households With One or More People Under 18 Years
R1104	Percent of Households With One or More People 65 Years and Over
R1105	Average Household Size
R1106	Percent of Households That Are Multigenerational
R1201	Percent of Men 15 Years and Over Who Were Never Married
R1202	Percent of Women 15 Years and Over Who Were Never Married
R1203	Ratio of Unmarried Men 15 to 44 Years per 100 Unmarried Women 15 to 44 Years
R1204	Median Age at First Marriage for Men
R1205	Median Age at First Marriage for Women
R1251	Marriage Rate per 1,000 Women 15 Years and Over
R1252	Marriage Rate per 1,000 Men 15 Years and Over
R1253	Divorce Rate per 1,000 Women 15 Years and Over
R1254	Divorce Rate per 1,000 Men 15 Years and Over
R1303	Women 15 to 50 Years Old Who Had a Birth in the Past 12 Months (per 1,000 Women)
R1304	Total Fertility Rate of Women 15 to 50 Years Old Who Had a Birth in the Past 12 Months
R1501	Percent of People 25 Years and Over Who Have Completed High School (includes Equivalency)
R1502	Percent of People 25 Years and Over Who Have Completed a Bachelor's Degree
R1503	Percent of People 25 Years and Over who Have Completed an Advanced Degree

APPENDIX A (continued)

Listing of 85 Ranking Tables Based on the 2011 American Community Survey

R1601	Percent of People 5 Years and Over Who Speak a Language Other Than English at Home
R1602	Percent of People 5 Years and Over Who Speak Spanish at Home
R1603	Percent of People 5 Years and Over Who Speak English Less Than "Very Well"
R1701	Percent of People Below Poverty Level in the Past 12 Months (For Whom Poverty Status Is Determined)
R1702	Percent of Related Children Under 18 Years Below Poverty Level in the Past 12 Months
R1703	Percent of People 65 Years and Over Below Poverty Level in the Past 12 Months
R1704	Percent of Children Under 18 Years Below Poverty Level in the Past 12 Months (For Whom Poverty Status Is Determined)
R1810	Percent of People with a Disability
R1811	Employment to Population Ratio for People with a Disability
R2407	Percent of Civilian Employed Population 16 Years and Over in Computer, Engineering, and Science Occupations
R2408	Percent of Civilian Employed Population 16 Years and Over in HealthCare Practitioners and Technical Occupations
R1801	Median Household Income (in 2006 Inflation-Adjusted Dollars)
R1902	Median Family Income (in 2006 Inflation-Adjusted Dollars)
R1903	Percent of Households with Retirement Income
R1904	Percent of Households with Cash Public Assistance Income
R2001	Median Earnings for Male Full-Time, Year-Round workers (in 2006 Inflation-Adjusted Dollars)
R2002	Median Earnings for Female Full-Time, Year-Round Workers (in 2006 Inflation-Adjusted Dollars)
R2101	Percent of the Civilian Population 18 Years and Over Who Are Veterans
R2301	Percent of People 16 to 64 Years Who Are in the Labor Force (including Armed Forces)
R2302	Percent of Children Under 6 Years Old with All Parents in the Labor Force
R2303	Employment/Population Ratio for the Population 16 to 64 Years Old
R2304	Percent of Married-Couple Families with Both Husband and Wife in the Labor Force
R2401	Percent of Civilian Employed Population 16 Years and Over in Management, Business, and Financial Occupations
R2403	Percent of Civilian Employed Population 16 Years and Over in Service Occupations
R2404	Percent of Civilian Employed Population 16 Years and Over in the Manufacturing Industry
R2405	Percent of Civilian Employed Population 16 Years and Over in the Information Industry
R2406	Percent of Civilian Employed Population 16 Years and Over Who Were Private Wage and Salary Workers
R2501	Percent of Housing Units That Are Mobile Homes
R2502	Percent of Housing Units That Were Built in 2005 or Later
R2503	Percent of Housing Units That Were Built in 1939 or Earlier
R2504	Percent of Occupied Housing Units That Were Moved Into in 2005 or Later
R2505	Percent of Occupied Housing Units with Gas as Principal Heating Fuel
R2506	Percent of Occupied Housing Units with Electricity as Principal Heating Fuel
R2507	Percent of Occupied Housing Units with Fuel Oil, Kerosene, Etc. as Principal Heating Fuel
R2509	Percent of Occupied Housing Units with 1.01 or More Occupants per Room
R2510	Median Housing Value of Owner-Occupied Housing Units (Dollars)
R2511	Median Monthly Housing Costs for Owner-Occupied Housing Units with a Mortgage (Dollars)
R2512	Percent of Occupied Housing Units That Are Owner-Occupied
R2513	Percent of Mortgaged Owners Spending 30 Percent or More of Household Income on Selected Monthly Owner Costs
R2514	Median Monthly Housing Costs for Renter-Occupied Housing Units (Dollars)
R2515	Percent of Renter-Occupied Units Spending 30 Percent or More of Household Income on Rent and Utilities
R2701	Percent Without Health Insurance Coverage
R2702	Percent of Children Without Health Insurance Coverage

APPENDIX B

Multiple Comparisons and Bonferroni Correction

In each case of Figure 1, there is only one test, and the significance level of each test is α. In Figure 1, where the focus is on the reference population of Colorado with estimated rank 31, there are actually fifty different tests, population of state with rank k *vs* reference population of Colorado with estimated rank 31 for $k \neq 31$. If we want the overall level of the collection of fifty tests to be α, some adjustment is needed for the level of significance for each of the fifty separate tests. The *Bonferroni correction* provides some guidance, and we give a few details in the remainder of this appendix.

Assume a family-wide or collection of M tests (independent or dependent) of hypotheses:

$$\begin{array}{cccc} \text{Test 1} & H_0(1) & vs & H_A(1) \\ \text{Test 2} & H_0(2) & vs & H_A(2) \\ \vdots & & \vdots & \\ \text{Test } M & H_0(M) & vs & H_A(M) \end{array}$$

Let α be given where $0 < \alpha < 1$. Assume the probabilities of type one error for the tests separately are $\dfrac{\alpha}{M}$ so that

$$\begin{array}{rcl} P(\text{reject} \quad H_0(1) \mid H_0(1) \quad \text{true}) & \leq & \frac{\alpha}{M} \\ P(\text{reject} \quad H_0(2) \mid H_0(2) \quad \text{true}) & \leq & \frac{\alpha}{M} \\ & \vdots & \\ P(\text{reject} \quad H_0(M) \mid H_0(M) \quad \text{true}) & \leq & \frac{\alpha}{M} \end{array}$$

Thus the level of statistical significance for the m^{th} test is $\dfrac{\alpha}{M}$ for $m = 1, 2, 3, ..., M$. Hence the overall level of statistical significance for the collection of M tests simultaneously is α because by Boole's Law, we have

$P(\text{reject at least one of the } M \text{ tests given it is true})$

$$\begin{aligned} &= P([reject\ H_0(1)|H_0(1) \text{ true}] \ or\ [\text{reject } H_0(2)|H_0(2) \text{ true}] \ or\ \cdots \ or\ [\text{reject } H_0(M)|H_0(M) \text{ true}]) \\ &\leq P(reject\ H_0(1)|H_0(1) \text{ true}) + P(\text{reject } H_0(2)|H_0(2) \text{ true}) + \cdots + P(\text{reject } H_0(M)|H_0(M) \text{ true}) \\ &\leq \tfrac{\alpha}{M} + \tfrac{\alpha}{M} + \cdots + \tfrac{\alpha}{M} \\ &= \alpha. \end{aligned}$$

Thus when testing all of the M hypotheses simultaneously at significance level α, one can test each one of the hypotheses at significance level $\dfrac{\alpha}{M}$.

This method of multiple comparisons is referred to as the *Bonferroni correction* (Bonferroni, 1935, 1936; Dunn, 1961). When testing a collection of hypotheses simultaneously, we reject the entire collection of null hypotheses if the null hypothesis for at least one of them is rejected. When we increase the number of hypotheses being tested simultaneously, we are more likely to have a type one error if each hypothesis is tested at the same level α. By decreasing the level of each separate test to $\frac{\alpha}{M}$, we are maintaining the overall level of significance at α, and it is in this sense that we think of the Bonferroni correction.