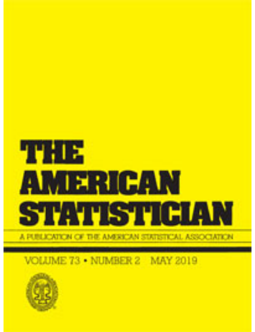


THE AMERICAN STATISTICIAN

A PUBLICATION OF THE AMERICAN STATISTICAL ASSOCIATION

VOLUME 73 • NUMBER 2 MAY 2019

The American Statistician

ISSN: 0003-1305 (Print) 1537-2731 (Online) Journal homepage: https://www.tandfonline.com/loi/utas20

A Primer on Visualizations for Comparing Populations, Including the Issue of Overlapping Confidence Intervals

Tommy Wright, Martin Klein & Jerzy Wieczorek

To cite this article: Tommy Wright, Martin Klein & Jerzy Wieczorek (2019) A Primer on Visualizations for Comparing Populations, Including the Issue of Overlapping Confidence Intervals, The American Statistician, 73:2, 165-178, DOI: 10.1080/00031305.2017.1392359

To link to this article: https://doi.org/10.1080/00031305.2017.1392359

Accepted author version posted online: 30 Oct 2017.
Published online: 22 Jun 2018.

THE AMERICAN STATISTICIAN
2019, VOL. 73, NO. 2, 165–178: Statistical Practice
https://doi.org/10.1080/00031305.2017.1392359

Taylor & Francis
Taylor & Francis Group

A Primer on Visualizations for Comparing Populations, Including the Issue of Overlapping Confidence Intervals

Tommy Wright[a,b], Martin Klein[a,c], and Jerzy Wieczorek[d]

[a]Center for Statistical Research and Methodology, U.S. Bureau of Census, Washington, DC; [b]Mathematics and Statistics Department, Georgetown University, Washington, DC; [c]Department of Mathematics and Statistics, University of Maryland Baltimore County, Baltimore, MD; [d]Department of Mathematics and Statistics, Colby College, Waterville, ME

ABSTRACT

In comparing a collection of K populations, it is common practice to display in one visualization confidence intervals for the corresponding population parameters $\theta_1, \theta_2, \dots, \theta_K$. For a pair of confidence intervals that do (or do not) overlap, viewers of the visualization are cognitively compelled to declare that there is not (or there is) a statistically significant difference between the two corresponding population parameters. It is generally well known that the method of examining overlap of pairs of confidence intervals should not be used for formal hypothesis testing. However, use of a single visualization with overlapping and nonoverlapping confidence intervals leads many to draw such conclusions, despite the best efforts of statisticians toward preventing users from reaching such conclusions. In this article, we summarize some alternative visualizations from the literature that can be used to properly test equality between a pair of population parameters. We recommend that these visualizations be used with caution to avoid incorrect statistical inference. The methods presented require only that we have K sample estimates and their associated standard errors. We also assume that the sample estimators are independent, unbiased, and normally distributed.

ARTICLE HISTORY
Received September 2016
Revised September 2017

KEYWORDS
Confidence intervals; Overlap; Two-sample test; Visualizations

1. Introduction

Assume K independently sampled populations with associated cumulative distribution functions $F_1(y), \dots, F_K(y)$. Let θ_k be a real-valued characteristic (parameter) related to $F_k(y)$, for $k = 1, \dots, K$. While the values of $\theta_1, \dots, \theta_K$ are unknown, it is desired to compare (and possibly rank) the K populations based on these unknown values. If $Y_{k1}, \dots, Y_{kn_k}$ is a probability sample of size n_k from the kth population and the statistic $\hat{\theta}_k = \hat{\theta}_k(Y_{k1}, \dots, Y_{kn_k})$ is an estimator of θ_k for $k = 1, \dots, K$, we compare the K populations based on the observed values, $\hat{\theta}_1, \dots, \hat{\theta}_K$. In this article, K estimates and K associated estimated standard errors form the basis for each of the methods. Knowledge of the specific complex sampling design and estimation methodology for each population is not required.

In Section 2, we give five visual methods that present statistical uncertainty in the comparisons by visually comparing pairs of U.S. states. Pairs of any states k and k' are visually compared using differences $\hat{\theta}_k - \hat{\theta}_{k'}$ where $k \neq k'$ for $k, k' = 1, 2, \dots, K$. The visual methods make use of shading (I); use of comparisons of differences to zero (II); use of "comparison intervals" (III); use of "adjustment" of confidence level (IV); and use of "two-tiered error bars" (V). These methods construct and display confidence intervals and hypothesis tests for individual parameters for each population, and for the pairwise difference in the parameters for two populations. To facilitate the comparison discussion, we rank the populations based on the computed values of $\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_K$ and indicate the estimated rank of population k by $\hat{r}_k$, where $\hat{r}_k = 1$ corresponds to the smallest $\hat{\theta}_k$ and $\hat{r}_k = K$ corresponds to the largest $\hat{\theta}_k$.

In the context of the U.S. Census Bureau's *American Community Survey (ACS)*, we discuss these methods based on observed sample estimates during 2011 of θ_k, the mean travel time (in minutes) to work of workers 16 years and over who did not work at home (henceforth "mean travel time to work") for state k (including Washington, DC) where $k = 1, 2, \dots, 51$. The ACS's sampling design is basically a national stratified random sample with sampling and estimation following a finite population design-based framework. Paraphrasing *https://www.census.gov/programs-surveys/acs/about.html*,

> *the ACS provides data every year that help determine how hundreds of billions of dollars in federal and state funds are distributed each year. Currently, over 3,500,000 households are contacted each year to provide data for various geographic levels. The ACS questionnaire asks about: age, sex, race, family and relationships, income and benefits, health insurance, education, veteran status, disabilities, where people work and how they get there, and where they live and how much they pay for some essentials.*

1.1. The Issue of Overlapping Confidence Intervals

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

$$\left(\hat{\theta}_k - z_{\frac{\alpha}{2}}\,\mathrm{SE}_k\,,\ \hat{\theta}_k + z_{\frac{\alpha}{2}}\,\mathrm{SE}_k\right), \tag{1}$$

CONTACT Tommy Wright tommy.wright@census.gov Center for Statistical Research and Methodology (CSRM), U.S. Census Bureau, Washington, DC 20233.

$\hat{r}_k$	*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

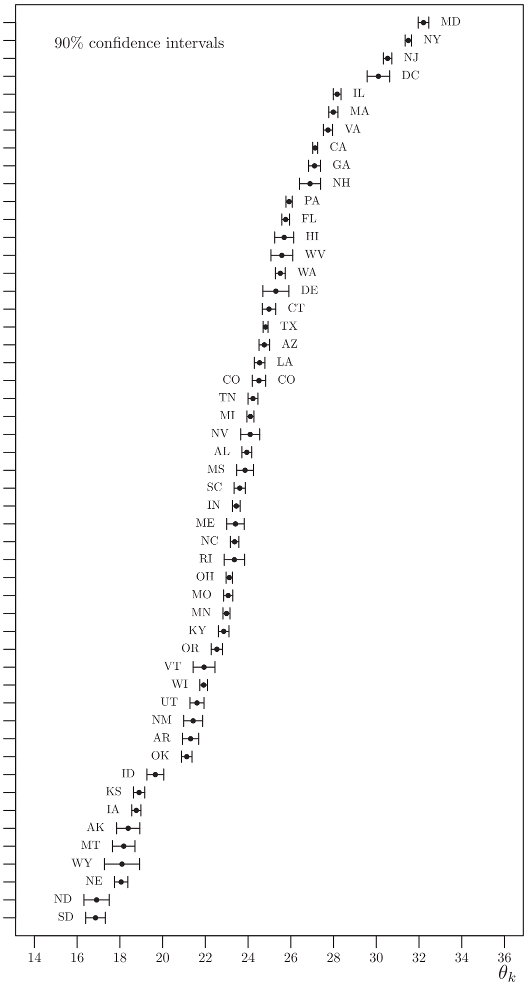


Figure 1. A 90% confidence interval for θ_k for each state for mean travel time to work (in minutes). (*Data Source:* 2011 American Community Survey.)

where $\text{SE}_k = \sqrt{\widehat{\text{var}}(\hat{\theta}_k)}$ is the standard error, $z_{\frac{\alpha}{2}} = \Phi^{-1}(1 - \frac{\alpha}{2})$, and Φ is the standard normal cumulative distribution function.

Our initial focus is illustrated in Figure 1 which ranks the 51 states (including Washington, DC) and provides a 90% confidence interval for each state's mean travel time to work for 2011 (θ_k). There is strong temptation in viewing such a visual to conclude that there are (are not) differences between states that have confidence intervals that do not (do) overlap. In fact, for two independently sampled populations, Cumming and Finch (2005) gave an approximate rule for using the amount of observed overlap between two confidence intervals to judge statistical significance. However, it is well known (e.g., Schenker and Gentleman 2001; Wright, Klein, and Wieczorek 2014) that the method of examining overlap of pairs of confidence intervals should not be used to formally test $H_0 : \theta_k = \theta_{k'}$ versus $H_A : \theta_k \neq \theta_{k'}$. Specifically in our setting, we show below that CLAIM 1 is true, but CLAIM 2 is false.

CLAIM 1 (True): If a $100(1-\alpha)$% confidence interval for θ_k does not overlap a $100(1-\alpha)$% confidence interval for $\theta_{k'}$, then the $100(1-\alpha)$% confidence interval for $\theta_k - \theta_{k'}$ does not contain 0.

CLAIM 2 (False): If a $100(1-\alpha)$% confidence interval for θ_k does overlap a $100(1-\alpha)$% confidence interval for $\theta_{k'}$, then the $100(1-\alpha)$% confidence interval for $\theta_k - \theta_{k'}$ does contain 0.

As noted, it is common practice to present one plot showing the 51 90% confidence intervals as in Figure 1 where each 90% confidence interval is computed as in (1). 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 correctly inferring that 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 $\alpha = 0.10$ (see Section 1.2); merely comparing the 90% confidence interval of θ_k with the 90% confidence interval for $\theta_{k'}$ is not for $\alpha = 0.10$. The approach of looking at overlapping and nonoverlapping intervals as might be done with Figure 1 as opposed to the approach of constructing a confidence interval for a difference are not equivalent approaches as we now demonstrate.

Proof of CLAIM 1: Let the $100(1-\alpha)$% confidence intervals for θ_k and $\theta_{k'}$ be

$$\left(\hat{\theta}_k - z_{\frac{\alpha}{2}}\text{SE}_k\,,\ \hat{\theta}_k + z_{\frac{\alpha}{2}}\text{SE}_k\right) \text{ and } \left(\hat{\theta}_{k'} - z_{\frac{\alpha}{2}}\text{SE}_{k'}\,,\ \hat{\theta}_{k'} + z_{\frac{\alpha}{2}}\text{SE}_{k'}\right),$$

respectively. Also let the $100(1-\alpha)$% confidence interval for $\theta_k - \theta_{k'}$ be

$$\left((\hat{\theta}_k - \hat{\theta}_{k'}) - z_{\frac{\alpha}{2}}\sqrt{(\mathrm{SE}_k)^2 + (\mathrm{SE}_{k'})^2},\ (\hat{\theta}_k - \hat{\theta}_{k'}) + z_{\frac{\alpha}{2}}\sqrt{(\mathrm{SE}_k)^2 + (\mathrm{SE}_{k'})^2}\right),$$

where $\hat{\theta}_k$ and $\hat{\theta}_{k'}$ are independent estimators.

Case 1. $\hat{\theta}_k < \hat{\theta}_{k'}$: Assume that the intervals for θ_k and $\theta_{k'}$ do not overlap. Because $\hat{\theta}_k < \hat{\theta}_{k'}$, this implies the following sequence of inequalities: $\hat{\theta}_k + z_{\frac{\alpha}{2}}\mathrm{SE}_k < \hat{\theta}_{k'} - z_{\frac{\alpha}{2}}\mathrm{SE}_{k'}$; $(\hat{\theta}_k - \hat{\theta}_{k'}) + z_{\frac{\alpha}{2}}(\mathrm{SE}_k + \mathrm{SE}_{k'}) < 0$; and $(\hat{\theta}_k - \hat{\theta}_{k'}) + z_{\frac{\alpha}{2}}\sqrt{(\mathrm{SE}_k)^2 + (\mathrm{SE}_{k'})^2} < 0$. The last inequality follows because $\sqrt{(\mathrm{SE}_k)^2 + (\mathrm{SE}_{k'})^2} \le \mathrm{SE}_k + \mathrm{SE}_{k'}$, and it implies that the $100(1-\alpha)\%$ confidence interval for $\theta_k - \theta_{k'}$ does not contain 0.

Case 2. $\hat{\theta}_k > \hat{\theta}_{k'}$: The proof is similar to that in Case 1. Hence CLAIM 1 is shown.

Demonstration that CLAIM 2 is false by counterexample: From Figure 1 for Colorado and Michigan, $\hat{\theta}_{\mathrm{CO}} = 24.51$, $\mathrm{SE}_{\mathrm{CO}} = 0.19$, $\hat{\theta}_{\mathrm{MI}} = 24.11$, and $\mathrm{SE}_{\mathrm{MI}} = 0.10$. It follows that (i) the usual 90% confidence interval for θ_{CO} is (24.20, 24.82); (ii) the usual 90% confidence interval for θ_{MI} is (23.95, 24.27); and (iii) the usual 90% confidence interval for $\theta_{\mathrm{CO}} - \theta_{\mathrm{MI}}$ is (0.05, 0.75) which does not contain 0. However, the individual 90% confidence intervals do overlap, and CLAIM 2 has been shown to be false.

1.2. General Setting for Each of the Five Visual Methods

For population k, let $\hat{\theta}_k$ have estimated standard error $\mathrm{SE}(\hat{\theta}_k) = \mathrm{SE}_k$ for $k = 1, \ldots, K$ and assume $E(\hat{\theta}_k) = \theta_k$. In this article, we treat the SE_k estimates as known constants. Let k^* be a specific reference population among the K populations.

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

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

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

$$H_0 : \theta_k = \theta_{k^*} \qquad \text{versus} \qquad H_A : \theta_k \neq \theta_{k^*}, \tag{3}$$

reject H_0 in favor of H_A if (2) does not contain 0; otherwise, do not reject H_0. For the applications in this article, the sample survey sizes n_k are sufficiently large to support the assumption of normality for $\hat{\theta}_k$ for all k.

Finally, in most of the article we assume independent sample estimators and standard errors and focus on methods that do not restrict the ranges of the plotted intervals. We discuss dependent estimators in general in Section 3, but one important special case is addressed by Baguley (2012), whose graphical methods account for the dependence induced by estimating within-subjects and between-subjects effects in classical ANOVA designs. Also, we assume no restrictions on the allowable ranges of our parameters, but it may be desirable to construct "range-preserving" confidence intervals, for example, to ensure that the intervals are nonnegative when a parameter is known to be nonnegative. Noguchi and Marmolejo-Ramos (2016) proposed a way to adjust a pair of range-preserving confidence intervals so that checking for (non) overlap is equivalent to significance testing.

1.3. Multiple Comparisons

For the methods discussed in Section 2, our illustrative examples use a demi-Bonferroni (Almond et al. 2000) correction to perform one-to-many multiple comparisons adjustments. By "demi-Bonferroni" we mean that we only correct for 50 comparisons between a reference state and all others, not for all $\binom{51}{2}$ possible pairwise comparisons. For instance, journalists writing a story on travel time to work for their state k^* may want to compare state k^* with all others, ignoring any comparisons that exclude k^*.

More powerful alternatives to the Bonferroni correction exist, but many of them assume equal variances between groups. When that assumption is warranted, as in the simple one-way ANOVA setting, our figures could be remade using other procedures, such as Dunnett's for one-to-many comparisons, or using a variant of Tukey–Kramer's Honestly Significant Difference procedure for all-to-all comparisons as in Gabriel (1978). However, in this article's example dataset and those of many other sample surveys, the equal variance assumption is implausible. Also, for one-to-many comparisons using significance tests alone, Figures 2 and 3 could be remade using Holm's step-down procedure, which allows for unequal variances and has more power than our demi-Bonferroni adjustment. However, Holm's procedure does not directly translate into interval estimates that we could use for later figures. For these reasons, we choose to consistently illustrate every method in this article using the same, simple, generally-applicable Bonferroni approach. Also, we use demi-Bonferroni in particular because the figures in Sections 2.2 and 2.3 are only suitable for one-to-many comparisons. When using the methods of Sections 2.1, 2.4, or 2.5, it would be possible to use full Bonferroni or other all-to-all multiple comparisons adjustment.

2. Visual Comparisons of Pairs

2.1. Comparing One Reference State With Each of the Others Using Shading

Figure 2, a "shaded column plot," gives an estimated ranking of the $K = 51$ states (including Washington, DC) based on point estimates $\hat{\theta}_k$ for the kth state's mean travel time to work with associated standard error SE_k for $k = 1, \ldots, 51$ using 2011 American Community Survey data. For example, the 2011 ACS estimate $\hat{\theta}_k$ of mean travel time to work for California (CA) is 27.14 min with standard error $\mathrm{SE}_k = 0.07$ min. California has estimated rank $\hat{r}_k = 44$.

For reference state $k^* \equiv$ Colorado (CO), and using a demi-Bonferroni correction for each of the 50 tests (3) comparing θ_{k^*} with each θ_k for $k \neq k^*$, the shaded (both heavy and light shading) states in the column (Figure 2) are statistically significantly different from the reference state CO, while the nonshaded states in the column are not. The level of significance for each

$\hat{r}_k$	*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 2. Shaded column plot. Shaded states do (unshaded states do not) differ from the reference state Colorado for mean travel time to work (in minutes). Significance level for each pair being compared is 0.002. The familywise (or overall) significance level for all pairs simultaneously being compared is 0.10. (*Data Source:* 2011 American Community Survey.)

test is $\frac{\alpha}{50} = 0.002$ (note $\frac{0.002}{2} = 0.001$ and $z_{0.001} = 3.1$), and the familywise (or overall) level of significance for the collection of 50 tests in the column is $\alpha = 0.10$. Where statistical packages may use an asterisk or star (*) to indicate a statistically significant difference from 0, we use shading in Figure 2 to indicate a statistically significant difference from the reference state.

The full "shaded columns plot" (also called "panty-hose plot") in Figure 3 (inspired by Almond et al. 2000) gives the overall visualization for all states where each column presents the 50 tests using demi-Bonferroni corrections for the reference state noted at the very bottom of the column.

The "letter display" of Piepho (2004) is an alternative visual form closely related to our Figures 2 and 3. Where we use white shading within a column to indicate a group of rows that are not significantly different from the reference state, a letter display would repeat a given letter within a column to indicate rows that are not significantly different from one another. Letter displays are designed for all-to-all comparisons, in which case Piepho (2004) also provided an algorithm to condense the display by removing redundant columns. This generalizes the familiar "lines display" for all-to-all comparisons, which only works when the states can be ordered such that not-significantly-different states are always adjacent. In contrast, our Figure 3 shows one-to-many comparisons for a different reference population in each column, so it would not make sense to condense the columns. Unlike the letter display, Figures 2 and 3 provide heavy and light shading to remind the reader which statistically significant differences are higher or lower. Finally, either our Figure 3 or a letter display (as appropriate) could be placed alongside a plot of ordinary confidence intervals like Figure 1, showing both the original intervals and the significance of comparisons in one display.

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

Using the same setup as in Section 1.2, Figure 4 gives 50 confidence intervals for the difference $\theta_k - \theta_{k^*}$ for reference state $k^* \equiv$ Colorado and $k \neq k^*$. We use a demi-Bonferroni correction as noted in Section 2.1. The bold intervals show the states that are statistically significantly different from Colorado (CO), while the nonbold intervals show the states that are not statistically significantly different from CO. Figures 2 and 4 both compare CO with each of the other 50 states and Washington, DC.

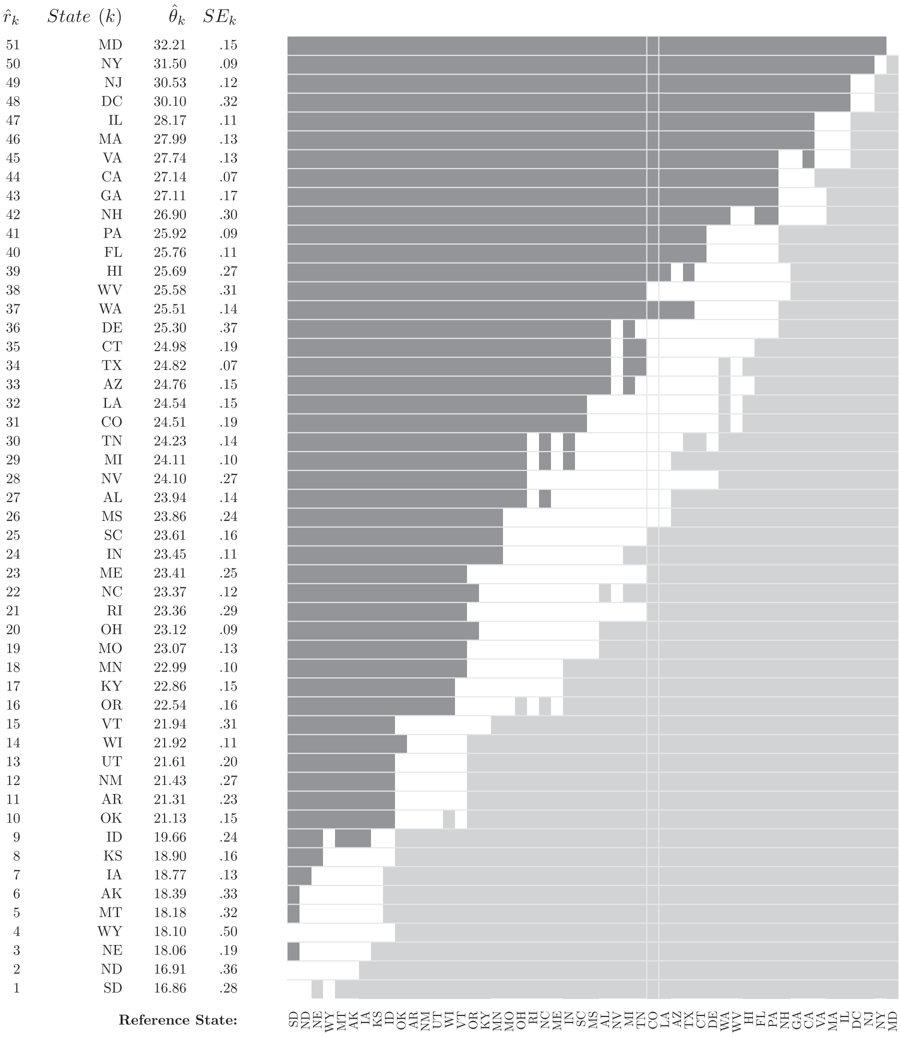
Section 2.4 gives a method of adjusting the level of the individual confidence intervals that leads to a test of significance level α based on the overlap/nonoverlap of one confidence interval with another confidence interval. Before that, we consider the overlap/nonoverlap of one confidence interval 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 in (1), 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. We proceed as in Almond et al. (2000).

If $\hat{\theta}_{k^*}$ is normal, a $100(1-\alpha)\%$ confidence interval for θ_{k^*} is given by (see also (1))

$$\left(\hat{\theta}_{k^*} - z_{\frac{\alpha}{2}}\mathrm{SE}_{k^*},\ \hat{\theta}_{k^*} + z_{\frac{\alpha}{2}}\mathrm{SE}_{k^*}\right). \tag{4}$$



$\hat{r}_k$	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 3. Fifty-one shaded columns plot. In each column, shaded states do (unshaded states do not) differ from reference state for mean travel time to work (in minutes). Significance level for each pair being compared is 0.002. For each column, the familywise (or overall) significance level for all pairs simultaneously being compared is 0.10. (*Data Source:* 2011 American Community Survey.)

Now consider another population, say k, where $\hat{\theta}_k < \hat{\theta}_{k^*}$. 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 (4) 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

$$\begin{aligned}(d_{k(\mathrm{low})}, d_{k(\mathrm{high})}) &= \Big((\hat{\theta}_{k^*} - z_{\frac{\alpha}{2}}\mathrm{SE}_{k^*}) - (\hat{\theta}_k + w_k)\,,\\ &\qquad (\hat{\theta}_{k^*} + z_{\frac{\alpha}{2}}\mathrm{SE}_{k^*}) - (\hat{\theta}_k - w_k)\Big)\\ &= \Big((\hat{\theta}_{k^*} - \hat{\theta}_k) - (z_{\frac{\alpha}{2}}\mathrm{SE}_{k^*} + w_k)\,,\\ &\qquad (\hat{\theta}_{k^*} - \hat{\theta}_k) + (z_{\frac{\alpha}{2}}\mathrm{SE}_{k^*} + w_k)\Big) \qquad (5)\end{aligned}$$

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 (2))

$$\begin{aligned}\Big((\hat{\theta}_{k^*} - \hat{\theta}_k) - z_{\frac{\alpha}{2}}\sqrt{(\mathrm{SE}_{k^*})^2 + (\mathrm{SE}_k)^2}\,,\ (\hat{\theta}_{k^*} - \hat{\theta}_k)\\ + z_{\frac{\alpha}{2}}\sqrt{(\mathrm{SE}_{k^*})^2 + (\mathrm{SE}_k)^2}\Big). \qquad (6)\end{aligned}$$

Equating results in (5) and (6) gives

$$z_{\frac{\alpha}{2}}\mathrm{SE}_{k^*} + w_k = z_{\frac{\alpha}{2}}\sqrt{(\mathrm{SE}_{k^*})^2 + (\mathrm{SE}_k)^2} \qquad (7)$$

or equivalently

$$w_k = z_{\frac{\alpha}{2}}\sqrt{(\mathrm{SE}_{k^*})^2 + (\mathrm{SE}_k)^2} - z_{\frac{\alpha}{2}}\mathrm{SE}_{k^*}. \qquad (8)$$

Note: When $\hat{\theta}_{k^*} < \hat{\theta}_k$, we have a figure analogous to Figure 5 with a new definition of $d'_{k(\mathrm{low})}$ and $d'_{k(\mathrm{high})}$ given in (5′) as

$$\begin{aligned}(d'_{k(\mathrm{low})}, d'_{k(\mathrm{high})}) &= \Big((\hat{\theta}_k - w_k) - (\hat{\theta}_{k^*} + z_{\frac{\alpha}{2}}\mathrm{SE}_{k^*})\,,\\ &\qquad (\hat{\theta}_k + w_k) - (\hat{\theta}_{k^*} - z_{\frac{\alpha}{2}}\mathrm{SE}_{k^*})\Big). \qquad (5')\end{aligned}$$

If $(\hat{\theta}_k - w_k, \hat{\theta}_k + w_k)$ and (4) do not overlap as in Figure 5, both $d_{k(\mathrm{low})}$ and $d_{k(\mathrm{high})}$ are positive; the confidence interval in (5) does not contain zero; 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 (4) do overlap, $d_{k(\mathrm{low})}$ is negative and $d_{k(\mathrm{high})}$ is positive; the confidence interval in (6) will contain zero; hence θ_k and θ_{k^*} are not significantly different at level α.

$\hat{r}_k$	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

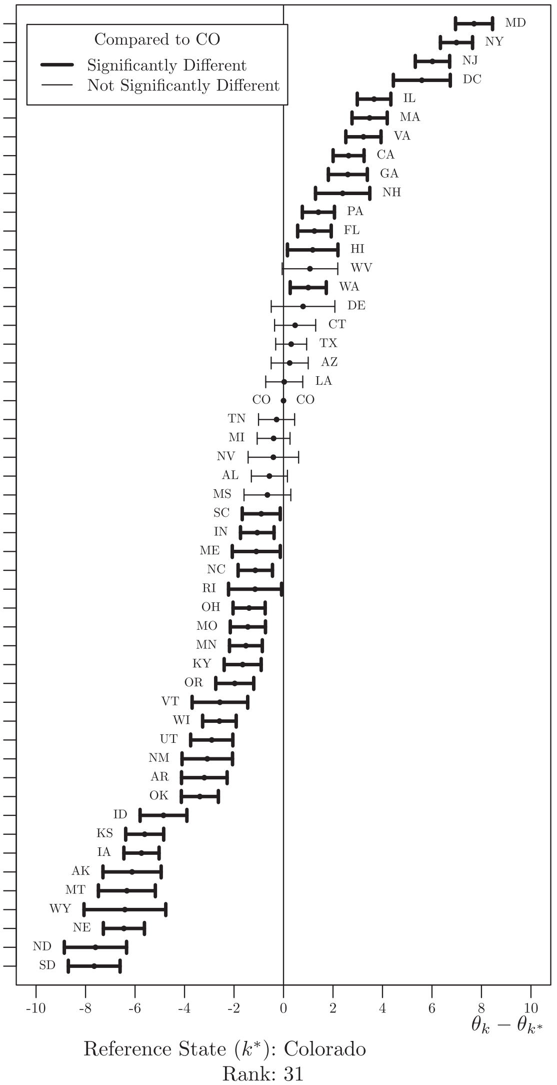


Figure 4. Fifty different $100(1 - 0.002)\% = 99.8\%$ confidence intervals for $\theta_k - \theta_{k^*}$ with reference state $k^* \equiv$ Colorado for mean travel time to work (in minutes). Overall $\alpha = 0.10$ for the collection of 50 tests. (*Data Source:* 2011 American Community Survey.)

Relative to $\hat{\theta}_{k^*}$, we refer to $(\hat{\theta}_k - w_k, \hat{\theta}_k + w_k)$ 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 for $K = 2$ (where $\hat{\theta}_k < \hat{\theta}_{k^*}$), three of many possibilities are shown in Figure 6. In each case, the distance from $\hat{\theta}_k$ to each bar is w_k. In Figure 6(a), populations k^* and k are significantly different at level α. In Figure 6(b) or 6(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 demi-Bonferroni correction for 50 separate tests of hypotheses where Colorado state's mean travel time to work is compared with each of the other $K - 1 = 50$ states' mean travel time. The level of significance for each test is then $\frac{\alpha}{50} = 0.002$, and the familywise (or overall) level of significance for the collection of 50 tests is $\alpha = 0.10$.

From the testing in Figure 7 at overall level $\alpha = 0.10$, Colorado's mean travel time to work is significantly different from all states except Mississippi (MS), Alabama (AL), Nevada (NV),

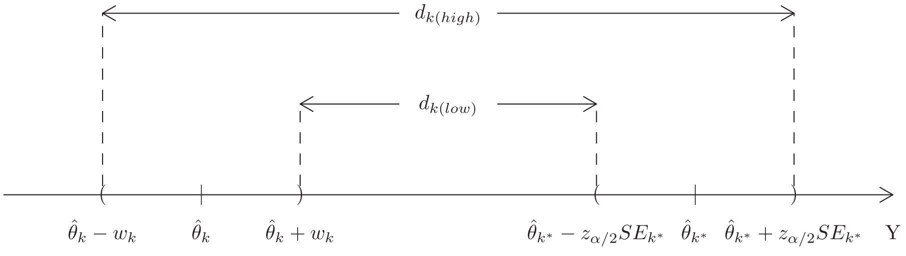


Figure 5. Illustration of motivation for method of Almond et al. (2000).

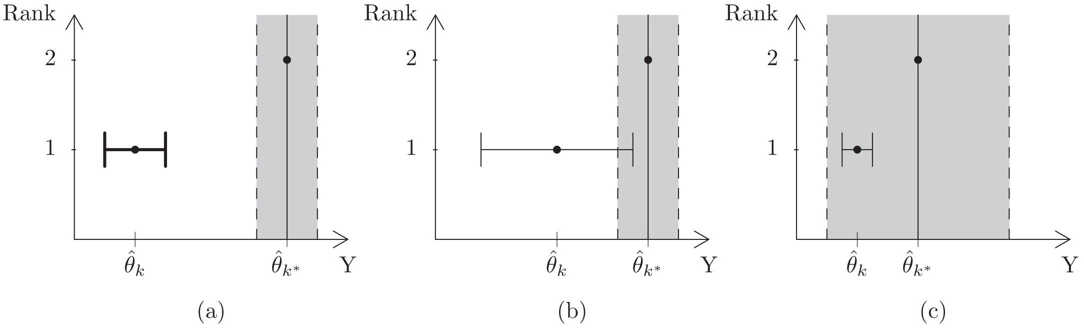


Figure 6. For $K = 2$, three of many possibilities for method of Almond et al. (2000). In each given possibility, the confidence interval for θ_{k^*} is shown with shading; the comparison interval for θ_k is shown with a bar. (a) Both estimates are precise relative to the difference in their means, and their difference is statistically significant. (b)–(c) One estimate is precise but the other is not, and their difference is not statistically significant.

Michigan (MI), Tennessee (TN), Louisiana (LA), Arizona (AZ), Texas (TX), Connecticut (CT), Delaware (DE), and West Virginia (WV). (Note the same comparison results for Colorado in Figures 2, 3, and 4.) The interval around Colorado (the reference state) that corresponds to the shaded strip is an approximate 99.8% confidence interval (demi-Bonferroni-corrected) for Colorado's mean travel time to work during the year 2011. In other words, Colorado's displayed 99.8% confidence interval follows from applying a demi-Bonferroni correction to a 90% confidence interval. The interval around each of the other states, say k, represents the comparison interval $(\hat{\theta}_k - w_k, \hat{\theta}_k + w_k)$ with w_k in (8).

Figures 4 and 7 provide the same information regarding comparing Colorado (θ_{k^*}) to the other states, but not the same

$\hat{r}_k$	*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

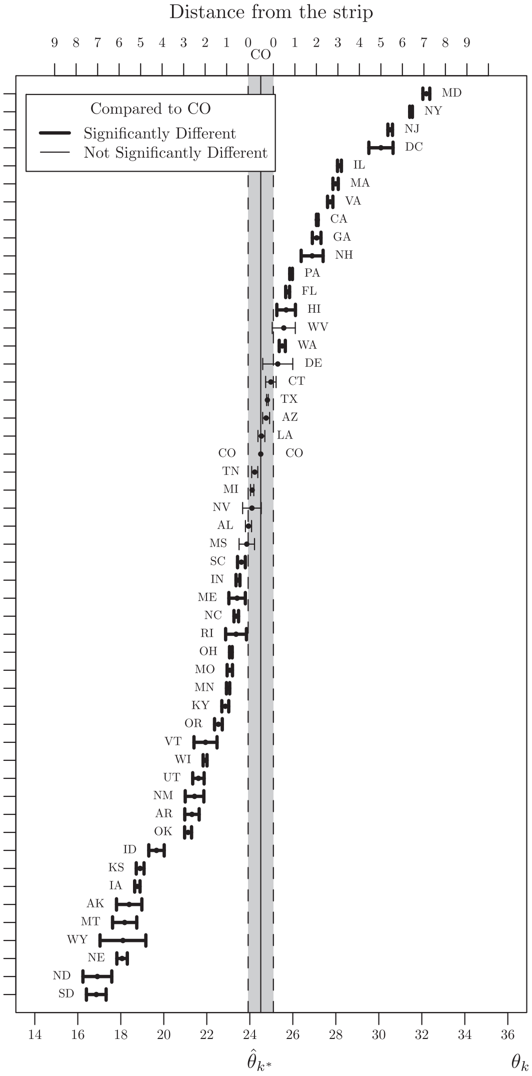


Figure 7. Comparisons with reference state Colorado using overlapping intervals for mean travel time to work (in minutes). A $100(1 - \alpha)\% = 99.8\%$ confidence interval for Colorado (θ_{k^*}) is shown with shading; the other intervals are comparison intervals. Significance level of test for each state being compared with Colorado is 0.002. Overall $\alpha = 0.10$ for the collection of 50 tests. (*Data Source:* 2011 American Community Survey.)

information about Colorado itself. In Figure 7, the usual 99.8% confidence interval for the reference state Colorado (θ_{k^*}) is shown explicitly; it is not shown in Figure 4. The (demi-Bonferroni-corrected) "comparison intervals" are not usual confidence intervals, and their endpoints or widths should not be interpreted separately from the reference state's interval. Finally, each state's comparison interval (θ_k), along with the 99.8% confidence interval (θ_{k^*}) for the reference state Colorado, provides the usual test of $H_0 : \theta_k = \theta_{k^*}$ via the 99.8% confidence interval for $\theta_k - \theta_{k^*}$ as given by (2).

It is possible to visually obtain the $100(1 - \alpha)\%$ confidence intervals for $\theta_k - \theta_{k^*}$ given in Figure 4 by using the confidence interval and comparison intervals in Figure 7.

For states k with $\hat{\theta}_k < \hat{\theta}_{k^*}$: The value of $d_{k(\text{high})}$ is obtained by measuring the distance between the upper bound of the confidence interval and the lower bound of the comparison interval for θ_k; $d_{k(\text{high})}$ equals this distance. The value of $d_{k(\text{low})}$ is obtained by measuring the distance between the lower bound of the confidence interval and the upper bound of the comparison interval for θ_k; if the comparison interval does not overlap the confidence interval, then $d_{k(\text{low})}$ equals this distance; if the intervals overlap, then $d_{k(\text{low})}$ equals the negative of this distance. The 99.8% confidence interval for $\theta_{k^*} - \theta_k$ is $(d_{k(\text{low})}, d_{k(\text{high})})$. The upper axis in Figure 7 is helpful for measuring these distances. Note that the width of the 99.8% confidence interval for θ_{CO} is 1.2. Using (5), we illustrate with three examples from Figure 7.

Example for $\theta_{SD} - \theta_{CO}$: $d_{\text{SD(high)}} = 7.50 + 1.20 = 8.70$; $d_{\text{SD(low)}} = 6.60$; the 99.8% confidence interval for $\theta_{\text{CO}} - \theta_{\text{SD}}$ is (6.60, 8.70); and for $\theta_{\text{SD}} - \theta_{\text{CO}}$, it is (−8.70, −6.60) as shown in Figure 4.

Example for $\theta_{MS} - \theta_{CO}$: $d_{\text{MS(high)}} = 0.4 + 1.2 = 1.6$; $d_{\text{MS(low)}} = -0.3$; the 99.8% confidence interval for $\theta_{\text{CO}} - \theta_{\text{MS}}$ is (−0.3, 1.6); and for $\theta_{\text{MS}} - \theta_{\text{CO}}$, it is (−1.6, 0.3). See Figure 4.

Example for $\theta_{TN} - \theta_{CO}$: $d_{\text{TN(high)}} = 1.01$; $d_{\text{TN(low)}} = -0.45$; the 99.8% confidence interval for $\theta_{\text{CO}} - \theta_{\text{TN}}$ is (−0.45, 1.01); and for $\theta_{\text{TN}} - \theta_{\text{CO}}$, it is (−1.01, 0.45). See Figure 4.

For states k with $\hat{\theta}_{k^*} < \hat{\theta}_k$: The value of $d'_{k(\text{high})}$ is obtained by measuring the distance between the upper bound of the θ_k comparison interval and the lower bound of the confidence interval; $d'_{k(\text{high})}$ equals this distance. The value of $d'_{k(\text{low})}$ is obtained by measuring the distance between the lower bound of the θ_k comparison interval and the upper bound of the confidence interval; if the comparison interval does not overlap the confidence interval, then $d'_{k(\text{low})}$ equals this distance; if the intervals overlap, then $d'_{k(\text{low})}$ equals the negative of this distance. The 99.8% confidence interval for $\theta_k - \theta_{k^*}$ is $(d'_{k(\text{low})}, d'_{k(\text{high})})$. We illustrate with three examples from Figure 7. For the shown precision, we use (5′).

Example for $\theta_{AZ} - \theta_{CO}$: $d'_{\text{AZ(high)}} = 1.0$; $d'_{\text{AZ(low)}} = -0.5$; and the 99.8% confidence interval for $\theta_{\text{AZ}} - \theta_{\text{CO}}$ is (−0.5, 1.0). See Figure 4.

Example for $\theta_{DE} - \theta_{CO}$: $d'_{\text{DE(high)}} = 1.20 + 0.88 = 2.08$; $d'_{\text{DE(low)}} = -0.50$; and the 99.8% confidence interval for $\theta_{\text{DE}} - \theta_{\text{CO}}$ is (−0.50, 2.08). See Figure 4.

Example for $\theta_{MD} - \theta_{CO}$: $d'_{\text{MD(high)}} = 1.20 + 7.25 = 8.45$; $d'_{\text{MD(low)}} = 6.95$; the 99.8% confidence interval for $\theta_{\text{MD}} - \theta_{\text{CO}}$ is (6.95, 8.45). See Figure 4.

When comparison intervals in Figure 7 do or do not overlap the confidence interval for the reference state, we are compelled to make a correct inference of comparison, but we have one interval which is a usual confidence interval while the other comparison intervals are not. The method presented in Section 2.4 is an attempt to make use of usual confidence intervals for all states in the visual display.

2.4. *Comparing Two States by Presenting Appropriate Overlapping/Nonoverlapping Confidence Intervals for Each State in the Pair*

Goldstein and Spiegelhalter (1996) argued for the use of intervals in conveying uncertainty explicitly in estimates or estimated ranks. Two procedures for deriving intervals are given: (i) the usual confidence intervals around estimated means of each population and (ii) a proposal by Goldstein and Healy (1995) described next.

Consider the pair of populations k and k' with parameters θ_k and $\theta_{k'}$. For given α, we want to determine α_A so that the following statement is true: "a $100(1 - \alpha_A)\%$ confidence interval for θ_k overlaps a $100(1 - \alpha_A)\%$ confidence interval for $\theta_{k'}$ if and only if a $100(1 - \alpha)\%$ confidence interval for $\theta_k - \theta_{k'}$ contains 0"; equivalently, "a $100(1 - \alpha_A)\%$ confidence interval for θ_k does not overlap a $100(1 - \alpha_A)\%$ confidence interval for $\theta_{k'}$ if and only if a $100(1 - \alpha)\%$ confidence interval for $\theta_k - \theta_{k'}$ does not contain 0." Goldstein and Healy (1995) showed how to do this when comparing one pair of estimates, as well as how to approximate it when comparing several pairs of estimates. We explain and illustrate their method. Assume K independently normally distributed $\hat{\theta}_k$ with standard error SE_k for $k = 1, \ldots, K$.

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 level α.

Let the $100(1 - \alpha_A)\%$ confidence interval for θ_k be $(\hat{\theta}_k - z_{\frac{\alpha_A}{2}} \text{SE}_k, \hat{\theta}_k + z_{\frac{\alpha_A}{2}} \text{SE}_k)$ and the $100(1 - \alpha_A)\%$ confidence interval for $\theta_{k'}$ be $(\hat{\theta}_{k'} - z_{\frac{\alpha_A}{2}} \text{SE}_{k'}, \hat{\theta}_{k'} + z_{\frac{\alpha_A}{2}} \text{SE}_{k'})$. If $|\hat{\theta}_k - \hat{\theta}_{k'}| > z_{\frac{\alpha_A}{2}} (\text{SE}_k + \text{SE}_{k'})$, then we have two cases: (i) $\hat{\theta}_k - \hat{\theta}_{k'} > z_{\frac{\alpha_A}{2}} (\text{SE}_k + \text{SE}_{k'})$ or (ii) $-(\hat{\theta}_k - \hat{\theta}_{k'}) > z_{\frac{\alpha_A}{2}} (\text{SE}_k + \text{SE}_{k'})$.

(i) In the first case, $\hat{\theta}_k - \hat{\theta}_{k'} > z_{\frac{\alpha_A}{2}} (\text{SE}_k + \text{SE}_{k'})$ is equivalent to $\hat{\theta}_k - z_{\frac{\alpha_A}{2}} \text{SE}_k > \hat{\theta}_{k'} + z_{\frac{\alpha_A}{2}} \text{SE}_{k'}$. So the $100(1 - \alpha_A)\%$ confidence intervals for θ_k and $\theta_{k'}$ do not overlap.

(ii) In the second case, $-(\hat{\theta}_k - \hat{\theta}_{k'}) > z_{\frac{\alpha_A}{2}} (\text{SE}_k + \text{SE}_{k'})$ is equivalent to $\hat{\theta}_{k'} - z_{\frac{\alpha_A}{2}} \text{SE}_{k'} > \hat{\theta}_k + z_{\frac{\alpha_A}{2}} \text{SE}_k$. So the $100(1 - \alpha_A)\%$ confidence intervals for $\theta_{k'}$ and θ_k do not overlap.

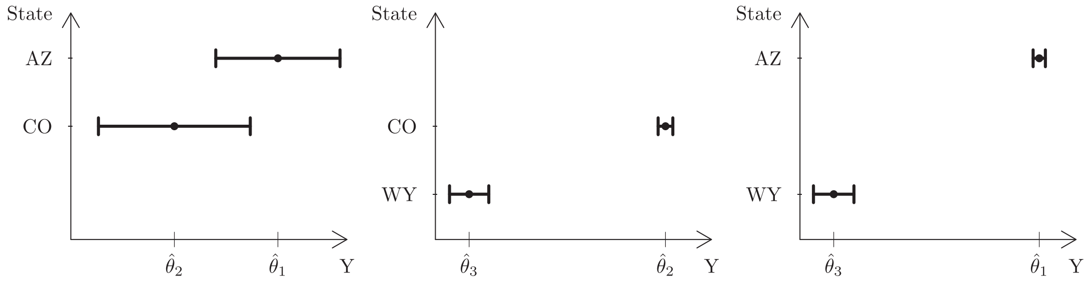
Thus,

$$|\hat{\theta}_k - \hat{\theta}_{k'}| > z_{\frac{\alpha_A}{2}} (\text{SE}_k + \text{SE}_{k'}) \tag{9}$$

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

If α_A is given, Equation (10) below gives the probability of a Type I error, which we call $\gamma_{kk'}$. We can set $\gamma_{kk'}$ equal to a chosen α and use (10) to determine the appropriate α_A using (11) below.

Specifically, let $(\text{SE}_{kk'})^2 \equiv \text{var}(\hat{\theta}_k - \hat{\theta}_{k'}) = (\text{SE}_k)^2 + (\text{SE}_{k'})^2$. The probability of the event in (9) under the hypothesis $\theta_k = \theta_{k'}$,



(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 $\alpha = 0.10$ if $100(1-\alpha_A)\%$ confidence intervals in each pair do not overlap. (*Data Source:* 2011 American Community Survey.)

that is, probability of a Type I error, is

$$\begin{aligned}\gamma_{kk'} &= P\left(|\hat{\theta}_k - \hat{\theta}_{k'}| > z_{\frac{\alpha_A}{2}}(\mathrm{SE}_k + \mathrm{SE}_{k'})\right)\\ &= 2P\left(\frac{(\hat{\theta}_k - \hat{\theta}_{k'}) - 0}{\mathrm{SE}_{kk'}} > z_{\frac{\alpha_A}{2}}\frac{(\mathrm{SE}_k + \mathrm{SE}_{k'})}{\mathrm{SE}_{kk'}}\right)\\ &= 2\left(1 - \Phi(z_{\frac{\alpha_A}{2}}\frac{(\mathrm{SE}_k + \mathrm{SE}_{k'})}{\mathrm{SE}_{kk'}})\right).\end{aligned} \tag{10}$$

Thus (10) relates $\gamma_{kk'}$ and $z_{\frac{\alpha_A}{2}}$ (hence α and α_A) for given values of SE_k and $\mathrm{SE}_{k'}$. So if we want the probability of a Type I error $\gamma_{kk'}$ to be equal to a specific value, say α, then we can 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. In practice, we set $\gamma_{kk'}$ equal to a chosen α, and determine the appropriate α_A given SE_k and $\mathrm{SE}_{k'}$ using

$$z_{\frac{\alpha_A}{2}}\frac{\mathrm{SE}_k + \mathrm{SE}_{k'}}{\mathrm{SE}_{kk'}} = z_{\frac{\alpha}{2}}. \tag{11}$$

Using values of $\hat{\theta}_k$ and SE_k from Figure 1 for Arizona (AZ: $\hat{\theta}_1 = 24.76$; $\mathrm{SE}_1 = 0.15$), Colorado (CO: $\hat{\theta}_2 = 24.51$; $\mathrm{SE}_2 = 0.19$), and Wyoming (WY: $\hat{\theta}_3 = 18.10$; $\mathrm{SE}_3 = 0.50$), we illustrate the method of Goldstein and Healy (1995).

Example, comparing the pair of states AZ and CO: Let $\alpha = 0.10$. Determine α_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 level α. Note that $\frac{(\mathrm{SE}_1+\mathrm{SE}_2)}{\mathrm{SE}_{12}} = 1.40$. For $\alpha = 0.10$, $z_{0.05} = 1.645$. Hence by (11) and solving $z_{\frac{\alpha_A}{2}} 1.40 = 1.645$, $z_{\frac{\alpha_A}{2}} = 1.17$ which implies $\alpha_A = 0.242$. Thus, the $100(1-0.242)\% = 76\%$ confidence interval for θ_1 by (1) is (24.62, 24.98). Similarly, a 76% confidence interval for θ_2 is (24.28 , 24.72). Note that they overlap (Figure 8(a)). Note also for $\alpha = 0.10$, that a 90% confidence interval for $\theta_1 - \theta_2$ by (2) is $(-0.10, 0.70)$ which includes 0. So we would not be able to say that the populations are significantly different at $\alpha = 0.10$. This is consistent with the 76% confidence intervals for θ_1 and θ_2 which overlap.

Example, comparing the pair of states WY and CO: For Colorado (θ_2) and Wyoming (θ_3) and $\alpha = 0.10$, we are led to $100(1-\alpha_A)\% = 80\%$ confidence intervals for θ_2 and θ_3, respectively, as (24.26, 24.74) and (17.46, 18.74) which do not overlap. We infer that θ_2 and θ_3 are different at $\alpha = 0.10$. See Figure 8(b). A 90% confidence interval for $\theta_3 - \theta_2$ is $(-7.28, -5.52)$ which does not contain 0.

Example, comparing the pair of states AZ and WY: For Arizona (θ_1) and Wyoming (θ_3) and $\alpha = 0.10$, we are led to 81% confidence intervals for θ_1 and θ_3, respectively, as (24.56, 24.96) and (17.44, 18.76) which do not overlap. We infer that θ_1 and θ_3 are different at $\alpha = 0.10$. See Figure 8(c). 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′: When there are more than two populations, Goldstein and Healy (1995) proposed to select α_A so that the average value of $\gamma_{kk'}$ over all (k, k') is a predetermined value α. Thus, we compute α_A such that

$$\alpha = \frac{1}{\binom{K}{2}}\sum_{1\le k<k'\le K} 2\left[1 - \Phi\left(z_{\frac{\alpha_A}{2}}\frac{\mathrm{SE}_k + \mathrm{SE}_{k'}}{\mathrm{SE}_{kk'}}\right)\right]. \tag{12}$$

Equation (12) can be solved numerically using an iterative procedure. Based on (11), Goldstein and Healy (1995) suggested that a starting point is to choose $z_{\frac{\alpha_A}{2}}$ such that

$$z_{\frac{\alpha_A}{2}}\left(\frac{1}{\binom{K}{2}}\sum_{1\le k<k'\le K}\frac{\mathrm{SE}_k + \mathrm{SE}_{k'}}{\mathrm{SE}_{kk'}}\right) = z_{\frac{\alpha}{2}}. \tag{13}$$

Finally, the confidence interval for the kth population is $(\hat{\theta}_k - z_{\frac{\alpha_A}{2}}\mathrm{SE}_k\ ,\ \hat{\theta}_k + z_{\frac{\alpha_A}{2}}\mathrm{SE}_k)$.

We will illustrate this advice by finding $z_{\frac{\alpha_A}{2}}$ simultaneously for the three pairs (AZ, CO), (WY, CO), and (AZ, WY) so that the average significance level across all three pairs is $\alpha = 0.10$. Note that for the various pairs we have the values in Table 1.

The average value of 1.40, 1.29, and 1.25 is 1.313. Using (13), we want $z_{\frac{\alpha_A}{2}}$ such that $z_{\frac{\alpha_A}{2}}(1.313) = 1.645$ or equivalently $z_{\frac{\alpha_A}{2}} = 1.25$. For $z_{\frac{\alpha_A}{2}} = 1.25$, $100(1-\alpha_A)\% = 100(1-2(.1056))\% \approx 79\%$. The 79% confidence intervals are given in Table 2.

For $z_{\frac{\alpha_A}{2}} = 1.25$, the level of significance for testing each pair is by (10) $\gamma_{kk'} = 2P(Z > z_{\frac{\alpha_A}{2}}\frac{\mathrm{SE}_k + \mathrm{SE}_{k'}}{\mathrm{SE}_{kk'}})$. We have the values in Table 3.

Table 1. Values of $\frac{\mathrm{SE}_k+\mathrm{SE}_{k'}}{\mathrm{SE}_{kk'}}$ for three pairs.

Pairs	$\frac{\mathrm{SE}_k+\mathrm{SE}_{k'}}{\mathrm{SE}_{kk'}}$
(AZ, CO)	1.40
(WY, CO)	1.29
(AZ, WY)	1.25

Table 2. 79% Confidence intervals for each state.

State(k)	79% confidence intervals for θ_k (see Figure 9)
AZ	$24.76 \pm 1.25(.15) \Rightarrow (24.57, 24.95)$
CO	$24.51 \pm 1.25(.19) \Rightarrow (24.27, 24.75)$
WY	$18.10 \pm 1.25(.50) \Rightarrow (17.48, 18.73)$

Table 3. Values of $\gamma_{kk'}$ for three pairs.

Testing pair	$\gamma_{kk'}$
(AZ, CO)	0.0802
(WY, CO)	0.1074
(AZ, WY)	0.1188

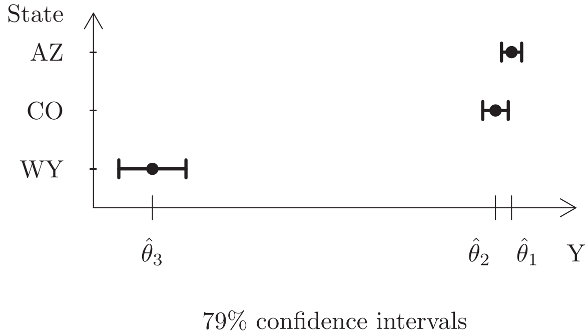


Figure 9. $100(1-\alpha_A)\% = 79\%$ confidence intervals for three states: For any pair, declare states k and k' statistically significantly different at "average significance level" $\alpha = 0.10$ if the 79% confidence intervals for the pair k and k' do not overlap. (*Data Source:* 2011 American Community Survey.)

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

$$\gamma_{kk'} = 2P(Z > 1.25(1.40)) = 2P(Z > 1.75)$$
$$= 2(0.0401) = 0.0802.$$

Also note that the average of the levels of significance is $(0.0802 + 0.1074 + 0.1188)/3 = 0.1021 \approx \alpha$, so in this case no further search is needed. Furthermore, the $100(1-\gamma_{kk'})\%$ confidence intervals for the differences are given in Table 4.

Figure 9 permits comparisons for three pairs of states while Figure 10 permits comparisons for all pairs of states. Again, no further search is needed after direct use of (13) on all pairs of states, which gives an average of the levels of significance of $0.1009 \approx \alpha$.

Unlike Figure 1, the visual display in Figure 10 leads to valid statistical inferences. In Figure 10, the 77.49% confidence intervals for Iowa and Idaho do not overlap, and by (13), we would correctly declare that Iowa and Idaho differ for an average significance level of $\alpha = 0.10$. In Figure 1, the 90% confidence intervals for Iowa and Idaho do not overlap, and by CLAIM 1, we would correctly declare that Iowa and Idaho differ for significance level of $\alpha = 0.10$. Note that the confidence intervals for Iowa and Kansas overlap in both Figures 1 and 10. So from Figure 10, we could say that Iowa and Kansas are not different for an average statistical significance level of $\alpha = 0.10$, but Figure 1 does not allow the analogous inference. Further, the intervals for Colorado and Delaware do not overlap in Figure 10 while they do overlap in Figure 1. Hence from Figure 10, we can correctly say that Colorado and Delaware differ for an average statistical significance level of $\alpha = 0.10$.

2.5. Comparing Two States by Presenting Appropriate Overlapping/Nonoverlapping Confidence Intervals for Each State in the Pair, Overlaid With Usual Confidence Intervals

The method of Section 2.4 is designed for making a single comparison valid at average significance level $\alpha = 0.10$, when the particular comparison is not known in advance. By showing intervals with a single nonstandard confidence level, for example, 77.49% as in Figure 10, we risk misleading readers who expect to see 90% confidence intervals and who judge individual estimates' precision this way. This problem can be addressed by using *two-tiered error bars* (Cleveland 1994). An "outer tier" runs the full length of the wider interval, with no cross-bars at the end, while an "inner tier" runs between cross-bars somewhere along the interval's length. As an illustration, Figure 11 shows two-tiered error bars for each estimate and brings the visualizations in Figures 1 and 10 together in one visualization. Each error bar's inner tier (between the cross-bars) of Figure 11 shows the same 77.49% confidence interval as in Figure 10. Each outer tier (the full width beyond the cross-bars) shows the original 90% confidence interval of Figure 1.

On the other hand, if we expect the reader to make a collection of comparisons, we may want a multiple comparisons correction. For example, we could use a Bonferroni correction to control the familywise error over all possible $\binom{51}{2} = 1275$ tests comparing two populations. Alternatively, to be comparable with Sections 2.1 through 2.3, we could use a demi-Bonferroni correction for 50 tests of hypotheses, with the intent of comparing a reference state with each of the other 50 states. Using a level of significance of $\alpha_A/50 = (1 - 0.7749)/50 = 0.0045$ for each test, we plot a $(1 - 0.0045) \times 100\% = 99.55\%$ confidence interval for each state. This admits a Goldstein and Healy (1995)-style average significance level of $0.1/50 = 0.002$. The value 0.002 includes a demi-Bonferroni correction for 50 tests. The two-tiered error bars using a demi-Bonferroni correction for 50 tests of hypotheses are shown in Figure 12. In Figure 12, note the reversed role of the inner and outer tiers relative to Figure 11: each error bar's inner part (between the cross-bars) shows the same 90% confidence interval as in Figure 1. Each outer tier (the full width beyond the cross-bars) shows the 99.55% confidence interval just described: 50-way demi-Bonferroni-corrected versions of the error bars from Figure 10. In this way, the outer tier

Table 4. Confidence intervals for differences.

Pair	$\gamma_{kk'}$	$z_{\frac{\gamma_{kk'}}{2}}$	$100(1-\gamma_{kk'})\%$	$100(1-\gamma_{kk'})\%$ Confidence interval for $\theta_k - \theta_{k'}$
(AZ, CO)	0.0802	1.75	92%	$(24.76 - 24.51) \pm 1.75\sqrt{(0.15)^2 + (0.19)^2} \Rightarrow (-0.17, 0.67)$
(WY, CO)	0.1074	1.61	89%	$(18.10 - 24.51) \pm 1.61\sqrt{(0.50)^2 + (0.19)^2} \Rightarrow (-7.29, -5.53)$
(AZ, WY)	0.1188	1.56	88%	$(24.76 - 18.10) \pm 1.56\sqrt{(0.15)^2 + (0.50)^2} \Rightarrow (6.14, 7.18)$

$\hat{r}_k$	*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

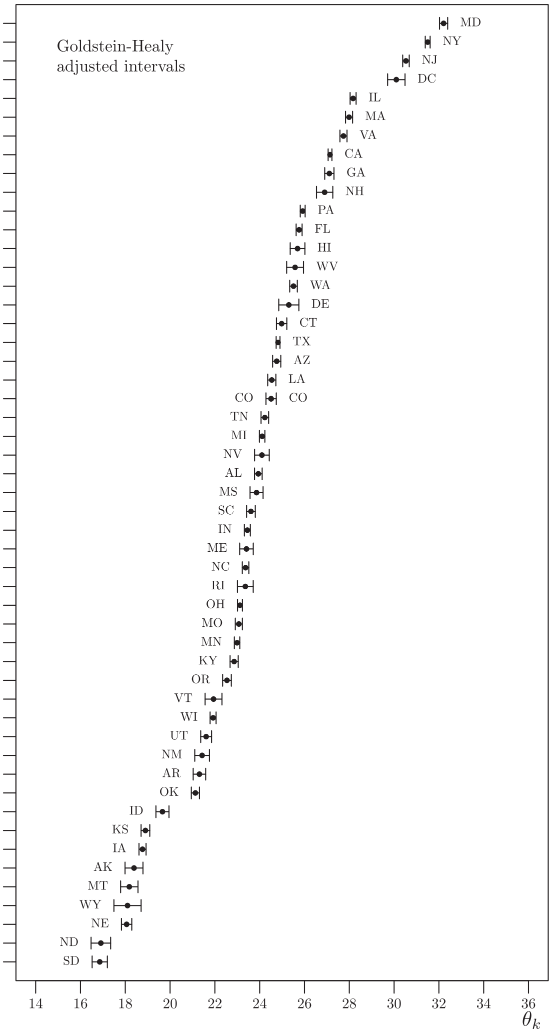


Figure 10. A $100(1-\alpha_A)\% = 77.49\%$ confidence interval for each state: For any pair, declare states k and k' statistically significantly different at "average significance level" $\alpha = 0.10$ if $100(1-\alpha_A)\%$ confidence intervals for the pair k and k' do not overlap. (*Data Source:* 2011 American Community Survey.)

allows up to 50 comparisons of the sort described in Section 2.4, controlled to a familywise average significance level of $\alpha = 0.10$. Meanwhile, readers can judge the precision of each individual estimate using the inner tier's usual 90% confidence intervals.

A demi-Bonferroni correction can cause the inner and outer tiers to swap roles between figures, as in Figures 11 and 12, although not within a single figure. For example, if a reader planned in advance to compare Colorado and Maine, Figure 10 shows that the 77.49% confidence intervals for Colorado and Maine do not overlap, so they differ at an average significance level of $\alpha = 0.10$. One can also see this from Figure 11 which shows the two-tiered error bars for both states. The inner tiers do not overlap, showing they differ at an average significance level of $\alpha = 0.10$, but we also see the outer tiers giving the original 90% confidence intervals.

On the other hand, if a reader is comparing Colorado to all other 50 states, Figure 12 shows that the 99.55% confidence intervals for Colorado and Maine do overlap, so they do not differ at a *familywise* average significance level of $\alpha = 0.10$. Note that in this case, the combination of demi-Bonferroni correction with average significance level turns out to be more conservative for Colorado than using a demi-Bonferroni correction alone. In Figure 12, Colorado's outer-tier intervals overlap with Hawaii, Maine, and Rhode Island, although Colorado was significantly different from these states in Figures 2, 3, 4, and 7.

Baguley (2012) suggested similar two-tiered error bars specifically for within-subject ANOVA designs, although the idea applies in a much wider setting. Two-tiered error bars could also be presented in a different visual form: Gelman and Hill (2007) plotted the outer tier as a thin line and the inner tier as a thicker line overlaid on top, instead of using cross-bars as in our Figures 11 and 12. Either of these forms (cross-bars vs. thick/thin lines) could be more visually distinct than the other, depending on the visual medium, for example, when a printout is photocopied or slides are projected at low resolution.

3. The Case of Dependent Estimators

Throughout, we have assumed that the parameter estimators $\hat{\theta}_1, \ldots, \hat{\theta}_K$ are independently distributed, and we have used this assumption in the construction of all visualizations. Alternatively, suppose that $\hat{\theta}_1, \ldots, \hat{\theta}_K$ are jointly distributed as K-dimensional multivariate normal. As before, for $k = 1, \ldots, K$, we assume $E(\hat{\theta}_k) = \theta_k$ and the standard error of

$\hat{r}_k$	*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

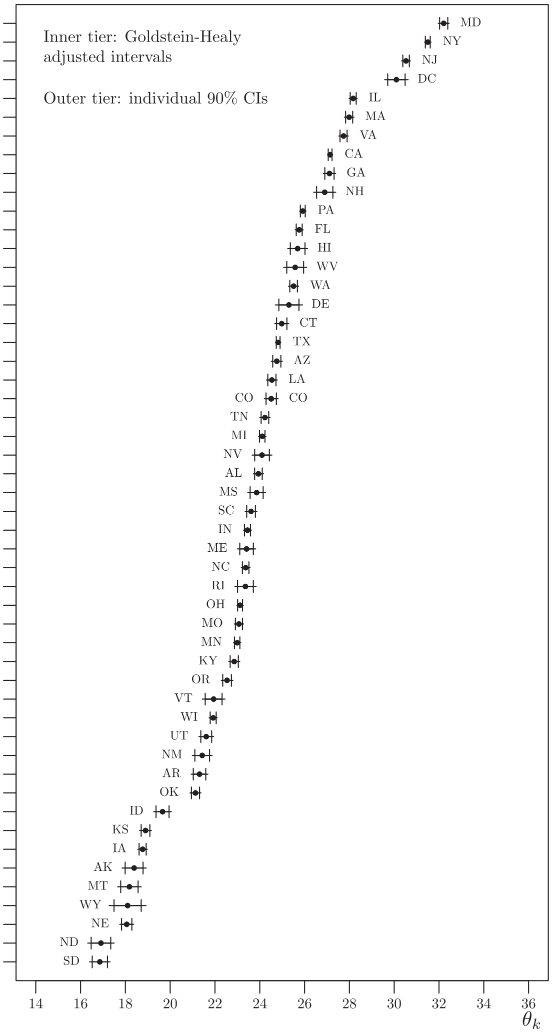


Figure 11. Two-tiered error bars showing a $100(1-\alpha_A)\% = 77.49\%$ confidence interval (inner) and a $100(1-\alpha)\% = 90\%$ confidence interval (outer) for each state: For any pair, declare states k and k' statistically significantly different at "average significance level" $\alpha = 0.10$ if $100(1-\alpha_A)\%$ confidence intervals (inner) for the pair k and k' do not overlap. Note that the tiers swap roles in Figure 12, where the individual intervals become inner tiers. (*Data Source:* 2011 American Community Survey.)

$\hat{\theta}_k$ is a known constant SE_k, but we now assume that the correlation between $\hat{\theta}_k$ and $\hat{\theta}_{k'}$ equals a known constant $\rho_{kk'}$ for each $k \neq k'$. In this situation, it is still true that a $100(1-\alpha)\%$ confidence interval for θ_k is given by (1). Because of dependence among the estimators, a $100(1-\alpha)\%$ confidence for the difference $\theta_k - \theta_{k'}$, $k \neq k'$, is no longer given by (2), but is now given by

$$\left((\hat{\theta}_k - \hat{\theta}_{k'}) - z_{\frac{\alpha}{2}} \text{SE}(\hat{\theta}_k - \hat{\theta}_{k'}) \,,\ (\hat{\theta}_k - \hat{\theta}_{k'}) + z_{\frac{\alpha}{2}} \text{SE}(\hat{\theta}_k - \hat{\theta}_{k'}) \right), \tag{14}$$

where $\text{SE}(\hat{\theta}_k - \hat{\theta}_{k'}) = \sqrt{(\text{SE}_k)^2 + (\text{SE}_{k'})^2 - 2\rho_{kk'}(\text{SE}_k)(\text{SE}_{k'})}$. Referring to CLAIM 1 and CLAIM 2 from Section 1, with the $100(1-\alpha)\%$ confidence interval for $\theta_k - \theta_{k'}$ now given by (14), we observe that it is still the case that CLAIM 1 is true, while CLAIM 2 in general is false.

Proof of CLAIM 1: Let the $100(1-\alpha)\%$ confidence intervals for θ_k and $\theta_{k'}$ be $(\hat{\theta}_k - z_{\frac{\alpha}{2}}\text{SE}_k, \hat{\theta}_k + z_{\frac{\alpha}{2}}\text{SE}_k)$ and $(\hat{\theta}_{k'} - z_{\frac{\alpha}{2}}\text{SE}_{k'}, \hat{\theta}_{k'} + z_{\frac{\alpha}{2}}\text{SE}_{k'})$, respectively. Also, let the $100(1-\alpha)\%$ confidence interval for $\theta_k - \theta_{k'}$ be (14). For $-1 \leq \rho_{kk'} \leq 1$, we observe that

$$\begin{aligned}&\sqrt{(\text{SE}_k)^2 + (\text{SE}_{k'})^2 - 2\rho_{kk'}(\text{SE}_k)(\text{SE}_{k'})} \\ &\quad \leq \sqrt{(\text{SE}_k)^2 + (\text{SE}_{k'})^2 + 2(\text{SE}_k)(\text{SE}_{k'})} = (\text{SE}_k + \text{SE}_{k'}).\end{aligned} \tag{15}$$

Case 1. $\hat{\theta}_k < \hat{\theta}_{k'}$: Assume the intervals for θ_k and $\theta_{k'}$ do not overlap. Because $\hat{\theta}_k < \hat{\theta}_{k'}$, this implies: $\hat{\theta}_k + z_{\frac{\alpha}{2}}\text{SE}_k < \hat{\theta}_{k'} - z_{\frac{\alpha}{2}}\text{SE}_{k'} \implies (\hat{\theta}_k - \hat{\theta}_{k'}) + z_{\frac{\alpha}{2}}(\text{SE}_k + \text{SE}_{k'}) < 0$. Combining the previous inequality with (15), we find that $(\hat{\theta}_k - \hat{\theta}_{k'}) + z_{\frac{\alpha}{2}}\sqrt{(\text{SE}_k)^2 + (\text{SE}_{k'})^2 - 2\rho_{kk'}(\text{SE}_k)(\text{SE}_{k'})} \leq (\hat{\theta}_k - \hat{\theta}_{k'}) + z_{\frac{\alpha}{2}}(\text{SE}_k + \text{SE}_{k'}) < 0$. Thus the confidence interval (14) does not contain 0.

Case 2. $\hat{\theta}_{k'} < \hat{\theta}_k$: Assume the intervals for θ_k and $\theta_{k'}$ do not overlap. Because $\hat{\theta}_{k'} < \hat{\theta}_k$, this implies: $\hat{\theta}_{k'} + z_{\frac{\alpha}{2}}\text{SE}_{k'} < \hat{\theta}_k - z_{\frac{\alpha}{2}}\text{SE}_k \implies 0 < (\hat{\theta}_k - \hat{\theta}_{k'}) - z_{\frac{\alpha}{2}}(\text{SE}_k + \text{SE}_{k'})$. Combining the previous inequality with (15), we find that $0 < (\hat{\theta}_k - \hat{\theta}_{k'}) - z_{\frac{\alpha}{2}}(\text{SE}_k + \text{SE}_{k'}) \leq (\hat{\theta}_k - \hat{\theta}_{k'}) - z_{\frac{\alpha}{2}}\sqrt{(\text{SE}_k)^2 + (\text{SE}_{k'})^2 - 2\rho_{kk'}(\text{SE}_k)(\text{SE}_{k'})}$. Thus the confidence interval (14) does not contain 0.

$\hat{r}_k$	*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

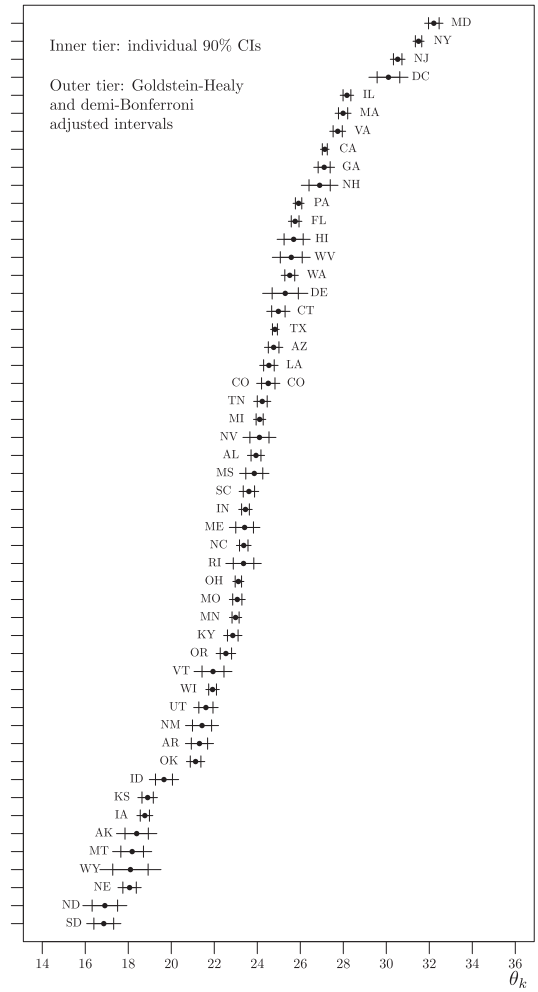


Figure 12. Two-tiered error bars (demi-Bonferroni corrected for 50 tests) showing a $100(1-\alpha_A/50)\% = 99.55\%$ confidence interval (outer) and a $100(1-\alpha)\% = 90\%$ confidence interval (inner) for each state: For any pair, declare states k and k' statistically significantly different at familywise "average significance level" $\alpha = 0.10$ if $100(1-\alpha_A/50)\%$ confidence intervals (outer) for the pair k and k' do not overlap, for up to 50 such tests. Note that the tiers swap roles in Figure 11, where the individual intervals become outer tiers. (*Data Source:* 2011 American Community Survey.)

Counterexample to CLAIM 2: Suppose $\hat{\theta}_k = 24.51$, $\text{SE}_k = 0.19$, $\hat{\theta}_{k'} = 24.11$, and $\text{SE}_{k'} = 0.10$. If $\rho_{kk'} = 0$, then this is the same example used in Section 1 to show that CLAIM 2 is false under the assumption that the parameter estimators are independent. There we noted that the usual 90% confidence interval for θ_k is (24.20, 24.83), the usual 90% confidence interval for $\theta_{k'}$ is (23.95, 24.27), and hence these intervals overlap. Now if, for example, $\rho_{kk'} = 0.7$ then the 90% confidence interval (14) for $\theta_k - \theta_{k'}$ is (0.17, 0.63), thus contradicting CLAIM 2. In fact, one may notice that for these particular values of $\hat{\theta}_k$, SE_k, $\hat{\theta}_{k'}$, and $\text{SE}_{k'}$, the 90% confidence interval (14) for $\theta_k - \theta_{k'}$ will contain 0 if and only if $\rho_{kk'} < [(\text{SE}_k)^2 + (\text{SE}_{k'})^2 - \frac{(\hat{\theta}_k - \hat{\theta}_{k'})^2}{z^2_{\frac{\alpha}{2}}}]/(2\text{SE}_k\text{SE}_{k'}) \approx -0.343$.

Schenker and Gentleman (2001) observed that the $100(1-\alpha)\%$ confidence intervals $(\hat{\theta}_k - z_{\frac{\alpha}{2}}\text{SE}_k, \hat{\theta}_k + z_{\frac{\alpha}{2}}\text{SE}_k)$ and $(\hat{\theta}_{k'} - z_{\frac{\alpha}{2}}\text{SE}_{k'}, \hat{\theta}_{k'} + z_{\frac{\alpha}{2}}\text{SE}_{k'})$ overlap if and only if the interval

$$\left((\hat{\theta}_k - \hat{\theta}_{k'}) - z_{\frac{\alpha}{2}}(\text{SE}_k + \text{SE}_{k'})\ ,\ (\hat{\theta}_k - \hat{\theta}_{k'}) + z_{\frac{\alpha}{2}}(\text{SE}_k + \text{SE}_{k'})\right) \tag{16}$$

contains 0. Intervals (14) and (16) are both centered at $\hat{\theta}_k - \hat{\theta}_{k'}$, but from (15), it is seen that in general interval (16) is wider than the $100(1-\alpha)\%$ confidence interval (14). However, when $\rho_{kk'} = -1$, the $100(1-\alpha)\%$ confidence interval in (14) is equivalent to the interval in (16). Hence, in the special case where $\rho_{kk'} = -1$, both CLAIM 1 and CLAIM 2 are true. We refer to Schenker and Gentleman (2001), Afshartous and Preston (2010), and Baguley (2012) for more discussion on the effects of correlation. While the visualizations presented in Sections 2.1– 2.5 assume that the estimators $\hat{\theta}_1, \ldots, \hat{\theta}_K$ are independent, with appropriate adjustments, the visualizations can also be constructed in the case of dependent estimators when the $100(1-\alpha)\%$ confidence interval for $\theta_k - \theta_{k'}$ is given by (14). The method of Goldstein and Healy (1995) presented in Section 2.4 has been extended to the case of dependent sample means by Afshartous and Preston (2010).

4. Concluding Comments

The visual methods of Section 2 are simple and easy to use. They are widely applicable and, with care, can be widely understood. For implementation, they mainly require K sample estimates

$\hat{\theta}_k$ and their associated standard errors SE_k, for $k = 1, \dots, K$. Theory exists to support their use. The methods presented require normality on $\hat{\theta}_k$, which can be justified in the ACS and other sample surveys where sample sizes are sufficiently large.

Each of the five visual methods has different ideal use-cases. A summary follows:

(i) The method of Section 2.1 (Figures 2 and 3) is a compact display if we only care about the statistical significance of differences and not their magnitudes.

(ii) The method of Section 2.2 (Figure 4) is helpful if we care about the size and precision of differences from a reference state, but not the precision of the reference state itself.

(iii) The method of Section 2.3 (Figure 7) is effective when we care about precision of one reference state and the statistical significance (but not concretely the estimated precision) of its differences with other states. Because the method of Section 2.3 involves "comparison intervals" which are not confidence intervals, we hesitate to recommend this method for general audiences to avoid confusion where one might refer to them incorrectly as confidence intervals. Our reason for including comparison intervals is to be instructive. First, when a comparison interval does or does not overlap the confidence interval for the reference state, we are compelled to make a correct inference of comparison. Second, unlike Figure 4, Figure 7 with comparison intervals shows the confidence interval of the reference state.

(iv) The method of Section 2.4 (Figure 10) is valuable when we care about the precisions of each state individually as well as the "average significance" of their differences. On the other hand, the individual precisions will be overestimated without a careful reading which recognizes that these are not the usual 90% or 95% confidence intervals.

(v) The method of Section 2.5 (Figures 11 and 12) indicates both the individual precisions, in familiar 90% confidence interval format, as well as the "average significance" of their differences. Specifically, in Figure 11, we visually link the original 90% confidence intervals of Figure 1 together with the adjusted 77.49% confidence intervals of Figure 10. In Figure 12, we take Figure 11 one step further by bringing in a demi-Bonferroni correction for 50 tests.

Finally, in many analyses we also have the original observations, not only their sample estimates and standard errors. In such cases, it may be useful to enhance the methods of Sections 2.3– 2.5 by overlaying each population's interval with a dot plot of its raw data, perhaps jittered for legibility. This approach would inform readers about the original data's distribution, not just the statistical summaries.

5. R Software for Figures

All figures in this article were made in R (R Core Team 2017). We also use the `tikzDevice` package (Sharpsteen and Bracken 2016) to allow our figures' text to match the article's LaTeX typesetting and fonts. We have collected our dataset, plotting functions, and example code into an R package, `RankingProject` (Wieczorek 2017), which is available online at the CRAN repository, *https://cran.r-project.org/package=RankingProject*. Our functions make it easy to produce equivalent figures for any dataset of K sample estimates and their standard errors. This package also contains a vignette which reproduces all major figures in our article.

Acknowledgments

The views expressed are those of the authors and not necessarily those of the U.S. Bureau of the Census. The authors are grateful to their colleagues (Derrick Simmons, Jun Shao, Carolina Franco, Josh Tokle, Pat Hunley, and Sarah Wilson) who read various drafts and to the referees and editors for many helpful suggestions.

References

Afshartous, D., and Preston, R. A. (2010), "Confidence Intervals for Dependent Data: Equating Non-Overlap with Statistical Significance," *Computational Statistics and Data Analysis*, 54, 2296–2305. [177]

Almond, R. G., Lewis, C., Tukey, J. W., and Yan, D. (2000), "Displays for Comparing a Given State to Many Others," *The American Statistician*, 54, 89–93. [167,168]

Baguley, T. (2012), "Calculating and Graphing Within-Subject Confidence Intervals for ANOVA," *Behavior Research Methods*, 44, 158–175. [167,175,177]

Cleveland, W. S. (1994), *The Elements of Graphing Data* (2nd ed.), Summit, NJ: Hobart Press. [174]

Cumming, G., and Finch, S. (2005), "Inference by Eye: Confidence Intervals, and How to Read Pictures of Data," *American Psychologist*, 60, 170–180. [166]

Gabriel, K. R. (1978), "A Simple Method of Multiple Comparisons of Means," *Journal of the American Statistical Association*, 73, 724–729. [167]

Gelman, A., and Hill, J. (2007), *Data Analysis Using Regression and Multilevel/Hierarchical Models*, New York: Cambridge University Press. [175]

Goldstein, H., and Healy, M. J. R. (1995), "The Graphical Presentation of a Collection of Means," *Journal of the Royal Statistical Society*, Series A, 158, 175–177. [172,173,174,177]

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, 159, 385–443. [172]

Noguchi, K., and Marmolejo-Ramos, F. (2016), "Assessing Equality of Means Using the Overlap of Range-Preserving Confidence Intervals," *The American Statistician*, 70, 325–334. [167]

Piepho, H.-P. (2004), "An Algorithm for a Letter-Based Representation of All-Pairwise Comparisons," *Journal of Computational and Graphical Statistics*, 13, 456–466. [168]

R Core Team (2017), *R: A Language and Environment for Statistical Computing*, Vienna, Austria: R Foundation for Statistical Computing. Available at *https://www.R-project.org/* [178]

Schenker, N., and Gentleman, J. F. (2001), "On Judging the Significance of Differences by Examining the Overlap Between Confidence Intervals," *The American Statistician*, 55, 182–186. [166,177]

Sharpsteen, C., and Bracken, C. (2016), *`tikzDevice`: R Graphics Output in LaTeX Format. R Package Version 0.10-1*. Available at *https://github.com/yihui/tikzDevice* [178]

Wieczorek, J. (2017), *`RankingProject`: The Ranking Project: Visualizations for Comparing Populations, R Package Version 0.1.1*. Available at *https://cran.r-project.org/package=RankingProject* [178]

Wright, T., Klein, M., and Wieczorek, J. (2014), "Ranking Populations Based on Sample Survey Data," Research Report Series (Statistics # 2014-12), Center for Statistical Research and Methodology, U.S. Bureau of the Census, Washington, DC. [166]