Georgia Journal of Science

Volume 81 *No. 2 Scholarly Contributions from the Membership and Others* Article 11

2023

Making the Error Bar Overlap Myth a Reality: Comparative Confidence Intervals

Frank S. Corotto
University of North Georgia, frank.corotto@ung.edu

MAKING THE ERROR BAR OVERLAP MYTH A REALITY: COMPARATIVE CONFIDENCE INTERVALS

Frank S. Corotto
Department of Biology (retired)
University of North Georgia,
Dahlonega, Georgia, 30597
Frank.Corotto@ung.edu

ABSTRACT

Many interpret error bars to mean that if they do not overlap the difference is statistically "significant". This overlap rule is really an overlap myth; the rule does not hold true for any conventional type of error bar. There are rules of thumb for estimating *P* values, but it would be better to show error bars for which the overlap rule holds true. Here I explain how to calculate *comparative confidence intervals* which, when plotted as error bars, let us judge significance based on overlap or separation. Others have published on these intervals (the mathematical basis goes back to John Tukey) but here I advertise comparative confidence intervals in the hope that more people use them. Judging statistical "significance" by eye would be most useful when making multiple comparisons, so I show how comparative confidence intervals can be used to illustrate the results of Tukey test. I also explain how to use of comparative confidence intervals to illustrate the effects of multiple independent variables and explore the problems posed by heterogeneity of variance and repeated measures. When families of comparative confidence intervals are plotted around means, I show how box-and-whiskers plots make it easy to judge which intervals overlap with which. Comparative confidence intervals have the potential to be used in a wide variety of circumstances, so I describe an easy way to confirm the intervals' validity. When sample means are being compared to each other, they should be plotted with error bars that indicate comparative confidence intervals, either along with or instead of conventional error bars.

Keywords: confidence interval, confidence limit, error bar, inferential confidence interval, inferential confidence limit, Tukey

INTRODUCTION

A common myth is that when error bars for two samples do not overlap, the difference is statistically meaningful, a term I use in place of statistically *significant*. This overlap rule is really an overlap myth; the rule does not hold true for any type of conventional error bar. There are rules of thumb for estimating *P* values from error bars (Cumming et al. 2007), but it would be better to show bars for which that overlap rule holds true. We could quickly assess the statistical meaningfulness of a pattern.

If we want the overlap rule to hold true, what should we plot as error bars? John Tukey gave the answer (see Benjamini and Braun 2002) and suggested that *interference notches* would be a good way to show the intervals graphically (Tukey 1993). Others unknowingly repeated Tukey's work in different ways (Austin and Hux 2002; Knoll et al.

2011) with Schuun (1999) using the phrase *statistical significance bars* and Tryon (2001) *inferential confidence intervals* in place of Tukey's inference notches.

None of the proposed terms for these error bars is ideal. All confidence intervals are inferential, statistical "significance" is widely misunderstood (which is why I use *meaningfulness* instead)[1], and Tukey's notches cannot be created with spreadsheets. I propose *comparative confidence intervals* (CCIs), preceded by alpha as in *0.05 CCIs*. The use of alpha reminds us that CCIs are not conventional confidence intervals.

To facilitate the broader use of comparative confidence intervals, I show here how to calculate the CCIs, how the intervals can be used in a variety of settings, and how they can be validated. I also explain why box-and-whiskers plots are a good way to show CCIs, in place of Tukey's notches. Schunn (1999) touched on some of the topics I cover here, but his approach was mathematical. To make a better case for comparative confidence intervals, I use figures instead.

CONVENTIONAL CONFIDENCE INTERVALS

To understand how CCIs are calculated, we must first understand conventional confidence intervals. Conventional intervals are calculated by performing null hypothesis tests backwards, often single-sample *t* tests. To explain, consider how we perform a single-sample *t* test forward. We begin with a null hypothesis, such as *the population mean equals precisely zero (μ = 0)*. It is that numerical precision that makes a null hypothesis a null hypothesis (Fisher 1971; reviewed in section 3.3 of Corotto 2022), although that precision may be implied rather than explicit. With that null in mind, we formulate a prediction, such as *the sample mean will be zero to an infinite number of decimal places ($\bar{x}$ = 0)*. We collect our data and use numerator in the formula for *t* below to compare our outcome to our prediction.

$$t = \frac{|\bar{x} - \mu|}{SE}$$

We divide that numerator by standard error (*SE*) to solve for *t*. With *t* in hand, we use degrees of freedom to determine *P*. To calculate a confidence interval, we start with a *P* value, such as 0.05, use degrees of freedom to find the *t* value that goes with that *P* value, i.e., the *critical value* of *t*, and multiply that critical value of *t* by standard error. Since standard error is in the denominator of the formula for *t*, when we multiply the critical value of *t* by standard error, standard error cancels out. We are left with the numerator in the formula, which is half of the confidence interval. The mean would be shown plus and minus that half-interval. What does that half-interval show? It shows the numerator in the formula for *t* that corresponds to *P* = 0.05, i.e., the smallest difference between our prediction and our outcome that would yield the finding that *P* is less than or equal to 0.05.

Since we begin null hypothesis testing with our null hypothesis it follows that, if we conduct a null hypothesis test backwards, our final product should be a null hypothesis. Our final product, our confidence interval, *is* a null hypothesis. In the case of performing a single-sample *t* test backwards, our null is that the population mean lies

[1]I use *meaningfully different*, *statistically meaningful*, and *statistically different.* "We in the behavioral sciences should 'give' this word [significant] back to the general public."—R. Kline (2004). Kline would use "statistical difference", but it is awkward to turn that around and say that a difference is statistical.

within the confidence interval. If we set alpha to 0.05, over a lifetime of constructing 95% confidence intervals around sample means, the population means will be outside of those intervals 5% of the time. Similarly, if we set alpha to 0.05, over a lifetime of testing true null hypotheses, we will incorrectly reject 5% of them. But a null hypothesis is infinitely numerically precise. Where is the infinite numerical precision in an interval? The precision is in the confidence *limits*.

SIMULTANEOUS CONFIDENCE INTERVALS

Comparative confidence intervals would be most useful when there are multiple comparisons being made. We could easily assess the statistical meaningfulness of a pattern. When there are multiple comparisons, however, the dogma is that we cannot base our confidence intervals on *t* tests. To explain, suppose we collect samples A, B, and C and compare A with B, A with C, and B with C by performing three *t* tests. The cumulative or *familywise* error risk would be 0.14, not 0.05 (for why it is not 3 × 0.05 = 0.15, see Zar 2010, pp. 189,190). To keep familywise error at alpha, instead of performing *t* tests backwards to get our intervals, we can perform multiple comparisons tests backwards. Good multiple comparisons tests hold familywise error at alpha. The result would be *simultaneous* confidence intervals, simultaneous in that they have been corrected for multiple comparisons. Here I use Tukey-Kramer tests, because the Tukey test is highly regarded (Zar 2010, p. 232), and the Tukey-Kramer version allows sample size to vary. Also, I construct the intervals around the difference between means, not the means themselves.

To illustrate the calculation of simultaneous confidence intervals, I created eight samples with similar variances but different sample sizes and performed a 1-way ANOVA (Appendixes A and B). The denominator in the resulting *F* ratio (2.308) is variously termed *mean square error, MS error*, or simply *the error term*. We will use the last two. The error term is important later, but for now we need the degrees of freedom associated with it, which is 52 (Appendix B). We use those 52 degrees of freedom; the number of categories compared by the ANOVA (typically shown as *k* in tables), which is eight in this case; and alpha (we will use 0.05) to find the corresponding critical value of q (q_{CV}; use the table of critical values of q not t). In this case that critical value is 4.466. We calculate standard error with the Tukey-Kramer formula, which follows.

$$SE = \sqrt{\left(\frac{MS\ error}{2}\right)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

The two sample sizes are indicated by n_1 and n_2. We will use sample A and sample F (Appendix A) as an example.

$$SE = \sqrt{\left(\frac{2.308}{2}\right)\left(\frac{1}{5} + \frac{1}{7}\right)}$$

$$SE = 0.629$$

Standard error multiplied by q_{CV} yields a half simultaneous confidence interval of 2.809.

For both Tukey and Tukey-Kramer tests, the test statistic q is calculated with the formula below

$$q = \frac{|\bar{x}_A - \bar{x}_B|}{SE}$$

in which $\bar{x}_A$ and $\bar{x}_B$ are the two sample means. By performing a Tukey-Kramer test backwards, we have solved for the numerator in the formula for q that corresponds to $P = 0.05$. The difference between the two sample means ($\bar{x}_A - \bar{x}_B = -3$) plus and minus the half simultaneous confidence interval (2.809) constitutes the simultaneous confidence interval of –5.809 through –0.191.

Figure 1 illustrates the results. The differences between every pair of sample means are plotted along with a family of simultaneous confidence intervals based on different alphas. For samples A and F, –3 is plotted along with bars that end at–5.809 and –0.191,

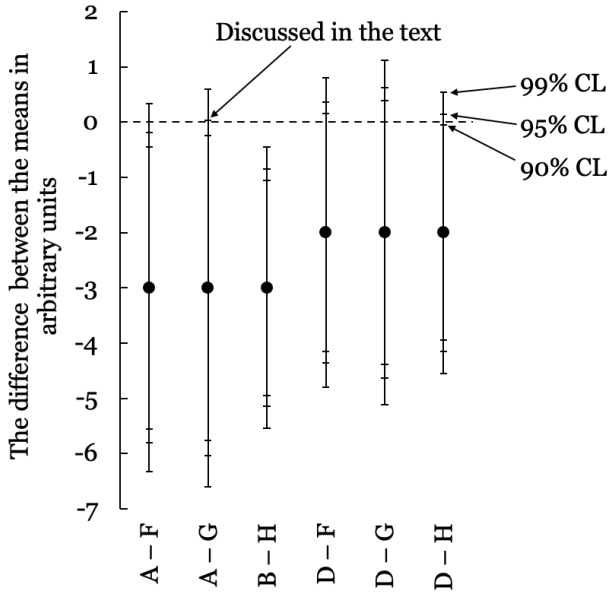

Figure 1. Some of the pairwise differences among the sample means in Appendix A, along with conventional simultaneous confidence intervals. CL = confidence limit.

the 95% simultaneous confidence limits. The fact that zero lies outside of the 95% simultaneous confidence interval but inside the 99% interval shows that P is less than 0.05 but greater than 0.01. The actual P value is 0.028 (Appendix C). The error bars illustrate the results of Tukey-Kramer tests.

COMPARATIVE CONFIDENCE INTERVALS

One problem with plots like Figure 1 is that we must think about what is being subtracted from what to interpret the signs of the outcomes. It is sample mean A minus sample mean F, so the negative difference means that F is greater than A, and not the other way around. Another problem is that, by showing the differences between the means, we cannot compare the means themselves by eye; larger patterns are obscured. It would be better to plot the means themselves along with comparative confidence intervals. To calculate the CCIs, we simply divide half simultaneous confidence intervals by two. Here is why.

Consider the comparison of samples A and G (Figure 1). The difference between the means is –3 and the upper 95% simultaneous confidence limit lies almost on zero. Suppose that limit was just beyond zero, i.e., $P = 0.05$, and the means themselves were

plotted rather than the difference between them. Those means would be separated by 3. If we want bars for which separation indicates that P is less than 0.05, how long should they be? They should be half the length of the bar extending from –3 to zero. To calculate comparative confidence intervals, we calculate half simultaneous confidence intervals, and divide by two. Then we plot the CCIs around means, not differences. The idea goes back to John Tukey. Benjamini and Braun (2002) describe his thoughts as follows:

> *If there exists a distance beyond which the two means are considered separated, then an effective graphical display involves drawing an allowance equal to plus or minus half that distance around the mean, and noting whether the allowances of the pair of means being compared overlap.*

In the case of samples A and G, the "distance beyond which the two means are considered separated" is the half simultaneous confidence interval of 3. We would plot sample means ±1.5. See also Figures 7 and 8 in Wainer (1996).

We will use sample A to show how CCIs are calculated. Because there is only one sample, we calculate standard error with the Tukey test's formula, which is as follows.

$$SE = \sqrt{\frac{MS\ error}{n}}$$

Here is the calculation for sample A.

$$SE = \sqrt{\frac{2.308}{5}}$$

$$SE = 0.679$$

If alpha is 0.05, and there are eight groups, the critical value of q is 4.466, as we saw earlier. That critical value multiplied by standard error yields a half simultaneous confidence interval of 3.032. We divide that half-interval by two and get 1.516. Bars that long would be plotted above and below sample A's mean of 2 to show 0.05 CCIs, i.e., the full interval would be 3.032 long. Comparative confidence intervals are actually half simultaneous confidence intervals.

THE SINGLE-SAMPLE TUKEY TEST

Let us indulge ourselves in a brief diversion and consider what type of null hypothesis test we just performed backwards. We used q's distribution and Tukey's formula for standard error, but our intent was to plot each CCI around a sample mean. The numerator for the Tukey test, which we might think we were solving for, has two sample means. Instead, we were solving for the numerator in the formula for a single sample t test, as shown earlier. Not only did we perform a null hypothesis test backwards, the test we performed backwards was the elusive single-sample Tukey test.

HOW TO PLOT COMPARATIVE CONFIDENCE INTERVALS

When using comparative confidence intervals, we must assess the degree to which error bars overlap with other error bars. This can be difficult if families of CCIs are plotted that correspond to different alphas. One way to reduce visual clutter is to plot just two intervals. I prefer 0.05 and 0.15 CCIs. (Critical values for 0.15 are available at the bottom

of this article's web page at the Georgia Journal of Science.) I chose 0.05 because it is a conventional alpha and 0.15 to hold the risk of a type III (or type S) error around 5% (sections 4.9 and 4.12 of Corotto 2022). Another reason to use 0.15 is to highlight close calls. If the 0.05 CCIs overlap, but the 0.15 CCIs do not, *P* is between 0.05 and 0.15; it may be worth increasing the sample size or conducting another study to further investigate. You, of course, might choose lower alphas. We might instead plot one set of CCIs along with another quantity we cannot bear to part with, such as standard error or conventional confidence intervals (the latter to make the American Psychological Association happy [APA 2020]). We might even plot CCIs based on a multiple comparisons test with CCIs based on uncorrected *t* tests. There are heretics who argue against correcting for multiple comparisons (reviewed by Hurlbert and Lombardi 2012).

Another way to reduce clutter is to use box-and-whiskers plots. In Figure 2, the boxes show the 0.15 CCIs and the whiskers 0.05. For example, if we compare sample A with sample E, the whiskers do not overlap, and $P = 0.022$ (Appendix C). If we compare samples D and H, the whiskers overlap, but the boxes do not, and $P = 0.085$.

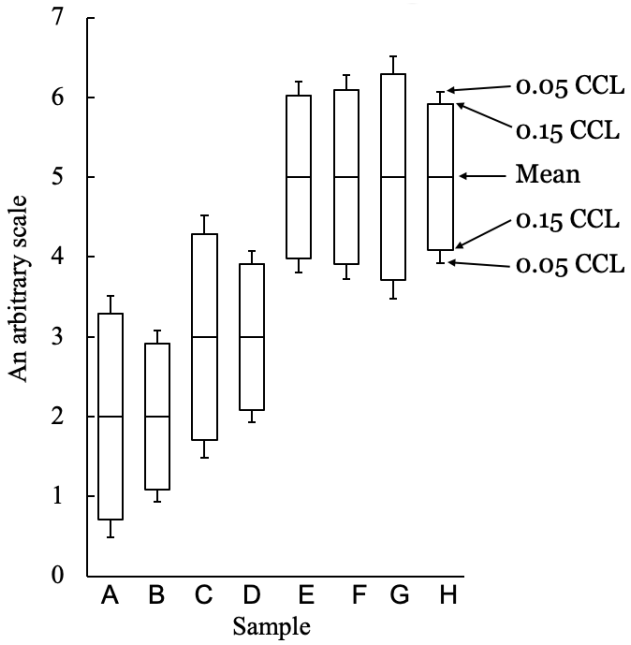

Figure 2. The means in Appendix A, along with comparative confidence intervals. CCL = comparative confidence limit.

DIFFERING SAMPLE SIZES

When solving for *P*, the Tukey-Kramer formula is used to calculate standard error when sample sizes differ. This means standard error may vary depending on which sample is being compared to which. For example, in the comparison of sample A with sample F, we found that standard error was 0.629. For samples A and G, however, standard error is 0.679. Consequently, the conventional simultaneous intervals for the comparison of sample A with F are smaller than those for the comparison of A with G (Figure 1). To calculate the comparative confidence intervals in Figure 2, I used the Tukey test's formula to calculate standard error, with *n* varying according to each sample. Do the resulting intervals reflect the *P* values that would be obtained by performing Tukey-Kramer tests in which standard error varies? For example, can the CCIs for sample A be compared to those for both F and G when Tukey-Kramer tests for those comparisons would entail the

use of different standard errors? To find out, I plotted the *P* values obtained from some of the Tukey-Kramer tests (Appendix C) as a function of the distance between the 0.05 and 0.15 CCIs, choosing results for which sample size varies and *P* values are low. The CCIs reflect the *P* values almost perfectly (Figure 3). Although sample A, for example, was

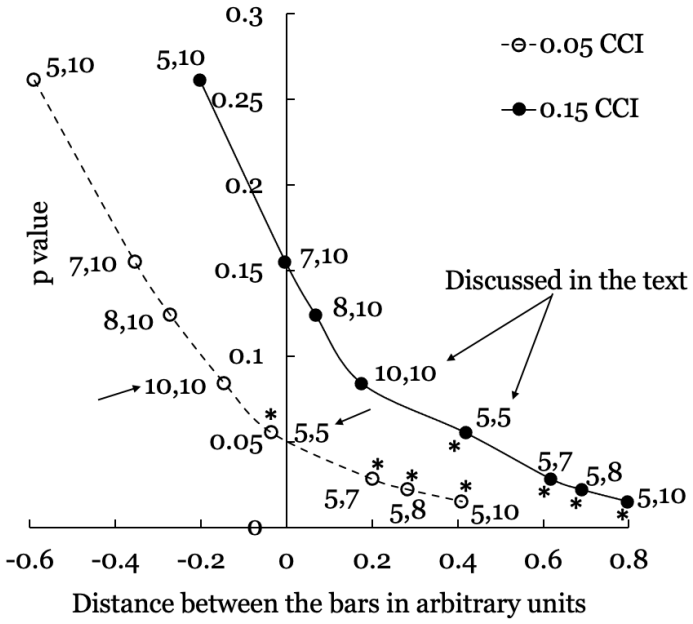

Figure 3. The distance between bars representing comparative confidence intervals reflects *P* values calculated with the Tukey-Kramer method. Negative *x*-values represent overlap of the bars. Asterisks indicate comparisons of sample A with other samples. Comparative confidence intervals are shown in Figure 2. Numbers indicate sample sizes for each pair of samples being compared. For both curves, the samples being compared are as follows, upper left to lower right: C and H; F and D; E and D; D and H; A and G; A and F; A and E; and A and H. CCI: comparative confidence interval.

compared to four different samples with four different sample sizes, the data for those comparisons line up with the rest (see the asterisks in Figure 3), and the curves as a whole have *y*-intercepts of 0.05 and 0.15. Large differences in sample size create slight anomalies. In the case where both sample sizes are 10, the two curves come close to each other while, when both sample sizes are five, they are farther apart (indicated by arrows in Figure 3). No doubt this is because different degrees of freedom correspond to different distributions of *q*. When the size of each sample is used to calculate standard error, CCIs reflect the results of Tukey-Kramer tests, though the relation between *P* and the degree of overlap or separation is complicated by the use of several distributions.

OTHER TESTS FOR OTHER SITUATIONS

Because comparative confidence intervals can be obtained by performing Tukey tests backwards, any way we perform null hypothesis tests forward, to obtain *P*, we can perform backwards, to get CCIs. If we were to compare a number of samples to a single reference (such as a control) and not to each other, we could calculate the intervals by performing Dunnett's test backwards. Dunnett's test is more powerful than Tukey's. What is important is that the critical values are based on the degrees of freedom associated with *MS error* and the total number of groups, since those are the values that would be used to conduct Dunnett's tests in the forward direction, and standard error should be based on mean square error as well. If there are only two samples being compared, comparative confidence intervals are not too important—there is no larger pattern to assess—but we could perform two-sample *t* tests backwards to get comparative confidence intervals. It would just be a matter of using total degrees of freedom to find the critical value of *t* and

using pooled variance to calculate standard error, since pooled variance is the equivalent of *MS error*.

If sample size varies, standard error can be calculated for each sample based on each sample's size, i.e., as was done for the intervals shown in Figure 2. This method works just as well for intervals based on Dunnett's test and two-sample *t* tests as for intervals based on Tukey tests (the outcomes are similar to what is shown in Figure 3). Note that for both *t* tests and Dunnett's test, the error term must be multiplied by two when calculating standard error. This is not the case when calculating standard error to obtain Tukey-based CCIs.

MAIN EFFECTS AND INTERACTIONS

Sometimes null hypothesis tests only tell us what is already obvious once we plot our data. Where these tests are particularly helpful is when there are multiple independent variables, i.e., a factorial design. Independent variables can have effects on their own, *main effects*, and they can affect each other; they can *interact*. It is often hard to judge by eye whether such an interaction is statistically meaningful or created by sampling error. We need to calculate *P*. To illustrate how we can use CCIs to show these *P* values, imagine we are testing three brands of tire, at the front and rear positions, and determining their longevity. If every possible combination of independent variables is represented, we have a factorial design (Figure 4).

With a factorial design, the averages for each combination of every independent variable are referred to as cell means, because the averages occupy cells in the matrix that illustrates the factorial design, e.g., brand A went an average of 40,000 miles in the front position (Figure 4). If we pool the data across the rows or columns, we can calculate *marginal* means that illustrate the main effects of each independent variable. For example, the average longevity of brand A is the average of its two cell means, 35,000 and 40,000 miles, or 37,500 miles, shown in the bottom margin in Figure 4. Similarly, we can pool sample sizes and illustrate them in the margins too. Understanding cell and marginal means and sample sizes will help us understand how to use comparative confidence intervals when there is a factorial design.

In the case of tire brand and position, we would analyze the results with a 2-way ANOVA, because there are two independent variables. The ANOVA would generate *F* ratios and *P* values for both of the main effects (*tire* and *position*) and also for the interaction. If there is a statistically meaningful main effect of *tire*, we might plot the marginal means of the three brands along with comparative confidence intervals to illustrate which brand is statistically different from which. The CCIs would be based on whatever test we would use to compare the three brands. Here Tukey-Kramer tests would be appropriate because sample size varies. We would use the number of groups being compared (three) and the degrees of freedom associated with *MS error* to find the critical value of *q*. To calculate standard error, we would use *MS error* for variance; we are assuming equal variances, so the best estimate is that error term; and the marginal samples sizes for each group, e.g., 39, 40, and 40 in the example shown in Figure 4. The resulting CCIs would be plotted around the marginal means for *tire* to illustrate the results of Tukey-Kramer tests.

There would be no reason to investigate the main effect of position, since our interest is in tire brand but, if we did want to plot the marginal means for *front* and *rear*, we could base our comparative confidence intervals on a two-sample *t* test. To calculate

		Brand of car tire			
		A	B	C	Marginal means and sample sizes
Position	Front	$\bar{x}$ = 40k n = 20	$\bar{x}$ = 50k n = 20	$\bar{x}$ = 50k n = 20	$\bar{x}$ = 46.7k n = 60
	Rear	$\bar{x}$ = 35k n = 19	$\bar{x}$ = 40k n = 20	$\bar{x}$ = 45k n = 20	$\bar{x}$ = 40k n = 59
	Marginal means and sample sizes	$\bar{x}$ = 37.5k n = 39	$\bar{x}$ = 45k n = 40	$\bar{x}$ = 47.5k n = 40	

Figure 4. A factorial design in which the longevity of three brands of car tire are compared at the front and rear positions. Longevity is in thousands (k) of miles. Sample size is indicated by *n*.

standard error, we would use *MS error* for variance and the marginal sample sizes of 59 and 60. If there were three positions, as would be the case when towing a small trailer, then we might base our CCIs on Tukey-Kramer tests, not *t* tests.

If the interaction is statistically meaningful, it means that we can exclude chance as the sole cause of a difference among differences. For example, there is a greater difference between brands A and B when they are at the front position than when they are at the rear. Is that difference in differences statistically meaningful? Is that why *P* is less than or equal to alpha for the interaction? To find out, we might plot CCIs based on two sets of multiple comparisons tests, one for *front* and one for *rear*; or plot CCIs based on three two-sample *t* tests, one for each brand. The latter makes more sense since our interest is in *tire*. Either way, the CCIs would be plotted around the cell means to illustrate the interaction. Error bars must always be explained, so we would make clear that the error bars can only be used for comparing across the brands within each position or vise versa.

AREAS FOR FUTURE STUDY

I know of two situations in which there are problems with comparative confidence intervals. One is when there is heterogeneity of variance. The other can arise when there are repeated measures. Repeated measures require the data to display *sphericity*: all samples must correlate to each other to the same degree. Concerns regarding both variance and sphericity are often addressed the same way, by testing the null hypothesis that variances are uniform and sphericity is perfect. This practice presents two problems. Because a null hypothesis is infinitely numerically precise, many nulls cannot be correct (reviewed by Nickerson 2000, Hurlbert and Lombardi 2009, and by authors they cite). These two null hypotheses in particular—that of uniform variance and perfect sphericity—must be wrong. Also, if we find that *P* is greater than alpha and decide that the data are good enough to proceed, we are asking if a deviation from perfection is large enough to be important. A null hypothesis test cannot answer that question. See also O'Brien and Kaiser (1985, pp. 318, 331).

Varying Variance

Although Tukey tests are highly regarded, they are not robust when variance differs among samples, especially when sample size varies as well (Zar 2010, p. 231). If variances are not "similar", to use Zar's word (Zar 2010, p. 232), one option is to transform the data to achieve homogenous variances. We could plot means and CCIs of the transformed data. Transformation can change the nature of the question, however. For example, an interaction following a log transformation indicates proportional differences, rather than absolute differences. Transformation can also fail to equalize variances. In the case of a rank transformation, the more the samples overlap, the less successful the transformation (section 13.4 in Corotto 2022).

Repeated Measures

A repeated measures ANOVA removes some of the variation among subjects from the analysis. This reduces the error term and increases statistical power. Consequently, when there is a mixed design, e.g., one within-subjects factor (the repeated measure) and one among-subjects factor, there will be two error terms: one for each factor. When calculating standard error for CCIs, we must use the correct error term depending on what is being compared to what (Loftus and Masson 1994). The error term for the repeated measure is usually not indicated by *MS error*, but by *MS remainder* or something else.

But problems arise when there is a lack of sphericity, and there is no transformation to correct for this issue. The severity of the problem can be gauged with the Greenhouse-Geisser method, the Huynh-Feldt method, and others. Those methods produce a statistic, epsilon, which ranges from zero to one, with one indicating perfect sphericity. I know of no rule of thumb for deciding if epsilon is large enough to indicate that sphericity is satisfactory. It is common to correct the ANOVA by multiplying both of the F ratio's degrees of freedom by epsilon. The more severe the problem, the lower the value of epsilon, and the greater the correction. One method that *might* work to create CCIs would be to multiply degrees of freedom by epsilon when finding the critical value of q. The CCIs would be corrected just like the ANOVA. Unfortunately, I have not seen this method in the literature, and I lack the expertise to test it.

SUMMARY

Much of what I have discussed here has been described before. From what I can tell, the strategies for addressing different sample sizes and the problems with Tukey tests are mine, as is my advocacy for box-and-whiskers plots and my suggestion of the validation strategy illustrated in Figure 3. Because comparative confidence intervals are calculated by performing null hypothesis tests backwards, the intervals have the potential to be based on tests other than those I discussed. When basing CCIs on other tests, the intervals can be validated with the analysis illustrated in Figure 3.

Confidence intervals are "the best reporting strategy" according to the American Psychological Association (APA 2020). Conventional intervals that flank sample means provide a range of likely values for population means. When samples are compared, however, the relations among the sample means can be more important that the means themselves. When means are being compared to each other, comparative confidence intervals should be plotted, along with or instead of conventional intervals.

Null hypothesis testing has been debated for decades. In fields in which it is termed null hypothesis *significance* testing, always initialized to NHST, the practice of making thoughtless yes-or-no decisions based on *P* values was once rampant. With comparative confidence intervals, we can practice thoughtless NHST. We can break out the T-square and see precisely what overlaps with what. At the other end of the spectrum, Loftus (1993) encouraged plotting means with standard error and abandoning null hypothesis tests. With CCIs, we can take Loftus's advice to an extreme. We can take in the big picture and never think about *P* values. Most of us will choose some strategy in between NHST and Loftus's, and CCIs will serve us well. Comparative confidence intervals make the APA's "best reporting strategy" even better, or at least more appropriate for making multiple comparisons.

REFERENCES

APA; American Psychological Association. 2020. Publication Manual of the American Psychological Association, 6th ed, p. 88.

Austin, P.C. and J.E. Hux. 2002. A brief note on overlapping confidence intervals. Journal of Vascular Surgery, 36, 194–195. doi:10.1067/mva.2002.125015.

Benjamini,Y. and H. Braun. 2002. John W. Tukey-Kramer's contributions to multiple comparisons. The Annals of Statistics, 30, 1576–1594. projecteuclid.org/euclid.aos/1043351247.

Corotto, F.S. 2022. Wise Use of Null Hypothesis Tests: A Practitioner's Handbook. Elsevier/Academic Press. See section 4.6.

Cumming, G., F. Fidler, and D.L. Vaux. 2007. Error bars in experimental biology. The Journal of Cell Biology, 177, 7–11. doi:10.1083/jcb.200611141.

Fisher, R. 1971. The Design of Experiments, 9th ed. Hafner Publishing Company, Inc., p. 16.

Hurlbert, S.H. and C.M. Lombardi. 2009. Final collapse of the Neyman-Pearson decision theoretic framework and rise of the neoFisherian. Annales Zoologici Fennici, 46(5), 311–349.

Hurlbert, S.H. and C.M. Lombardi. 2012. Lopsided reasoning on lopsided tests and multiple comparisons. Australian and New Zealand Journal of Statistics, 54, 23–42.

Kline, R. 2004. Beyond Significance Testing: Reforming Data Analysis Methods in Behavioral Research. American Psychological Association. See p. 87.

Knoll, M.J., W.R. Pestman, and D.E. Grobbee. 2011. The (mis)use of overlap of confidence intervals to assess effect modification. European Journal of Epidemiology, 26, 253–254. doi:10.1007/s10654-011-9563-8.

Loftus, G.R. 1993. A picture is worth a thousand *P* values: on the irrelevance of hypothesis testing in the microcomputer age. Behavior Research Methods, Instruments, & Computers, 25, 250–256. doi:10.3758/BF03204506.

Loftus, G.R. and M.E.J. Masson. 1994. Using confidence intervals in within-subjects designs. Psychonomic Bulletin & Review, 1, 476–490. doi:10.3758/BF03210951.

Nickerson, R.S. 2000. Null hypothesis significance testing: a review of an old and continuing controversy. Psychological Methods, 5(2), 241–301.

O'Brien, R. and M. Kaiser. 1985. MANOVA method for analyzing repeated measures designs: an extensive primer. Psychological Bulletin, 97(2), 316–333.

Schunn, C.D. 1999. Statistical significance bars (SSB): A way to make graphs more interpretable. http://www.lrdc.pitt.edu/schunn/ssb/SSB.rtf.

Tryon, W.W. 2001. Evaluating statistical difference, equivalence, and indeterminacy using inferential confidence intervals: an integrated alternative method of conducting null hypothesis statistical tests. Psychological Methods, 6, 371–386. doi:10.1037/1082-989X.6.4.371.

Tukey, J. 1993. Graphic comparisons of several linked aspects: alternatives and suggested principles. Journal of Computational and Graphical Statistics, 2(1), 1–33.

Wainer, H. 1996. Depicting error. The American Statistician, 50(2), 101–111.

Zar. J. 2010. Biostatistical Analysis, 5th ed. Prentice Hall.

Appendix A

Eight Samples with Similar Variances but Differing Sample Sizes

	Sample							
	A	B	C	D	E	F	G	H
	0	0	1	1	3	3	3	3
	1	1	2	2	4	4	4	4
	2	2	3	3	5	5	5	5
	3	3	4	4	6	6	6	6
	4	4	5	5	7	7	7	7
		0		1	4	3		3
		1		2	5	7		4
		2		3	6			5
		3		4				6
		4		5				7
mean	2	2	3	3	5	5	5	5
variance	2.500	2.222	2.500	2.222	1.714	3.000	2.500	2.222
SEM[a]	0.679	0.480	0.679	0.480	0.537	0.574	0.679	0.480
n[b]	5	10	5	10	8	7	5	10
df[c]	4	9	4	9	7	6	4	9

[a]SEM = standard error of the mean

[b]*n* = sample size

[c]*df* = degrees of freedom.

Appendix B

ANOVA Table for the Samples in Appendix A

Source	Sum of the squares	df	Mean square	*F*	*P*
Corrected model	101.250[a]	7	14.464	6.268	< 0.001
Intercept	770.642	1	770.642	333.945	< 0.001
Among groups	101.250	7	14.464	6.268	< 0.001
Within groups	120.000	52	2.308		
Total	1065.000	60			
Corrected total	221.250	59			

Note. The output was generated by SPSS except that the *P* values were reported as .000.

[a]r^2 = 0.458, adjusted r^2 = 0.385.

Appendix C

Pairwise Comparisons of All Samples in Appendix A. *P* values were calculated with Tukey-Kramer tests.[a]

Pair	*P* value	Pair	*P* value	Pair	*P* value
A vs B	0.999	B vs E	0.003	D vs E	0.124
A vs C	0.966	B vs F	0.005	D vs F	0.155
A vs D	0.928	B vs G	0.015	D vs G	0.261
A vs E	0.022	B vs H	0.001	D vs H	0.084
A vs F	0.028	C vs D	0.999	E vs F	0.999
A vs G	0.055	C vs E	0.308	E vs G	0.999
A vs H	0.015	C vs F	0.341	E vs H	0.999
B vs C	0.928	C vs G	0.440	F vs G	0.999
B vs D	0.818	C vs H	0.261	F vs H	0.999
				G vs H	0.999

[a]Results were obtained from SPSS. Values of 1.000 were changed to 0.999.