Varieties of Confidence Intervals

Denis Cousineau

École de Psychologie, Université d'Ottawa, Canada

ABSTRACT

Error bars are useful to understand data and their interrelations. Here, it is shown that confidence intervals of the mean (CI_Ms) can be adjusted based on whether the objective is to highlight differences between measures or not and based on the experimental design (within- or between-group designs). Confidence intervals (CIs) can also be adjusted to take into account the sampling mechanisms and the population size (if not infinite). Names are proposed to distinguish the various types of CIs and the assumptions underlying them, and how to assess their validity is explained. The various CIs presented here are easily obtained from a succession of multiplicative adjustments to the basic (unadjusted) CI width. All summary results should present a measure of precision, such as CIs, as this information is complementary to effect sizes.

KEYWORDS

statistics, methods,
precision,
confidence intervals

VARIETIES OF CONFIDENCE INTERVALS

Error bars have an important role to play in describing results and their precision and, to a lesser extent, in assessing whether the results meet the researcher's expectations or if they are at odds with them (Cumming, 2014; Loftus, 1993, 1996; Wilkinson & the Task Force on Statistical Inference, 1999). However, error bars come in many different types, and there is some confusion in the literature as to (a) when to use error bars, (b) which one to depict, and (c) how to interpret them. The answers to these questions are straightforward. Concerning the answer to question (a): Error bars should always be present on any plot showing summary results. There should be no exception, and editors should request them prior to publication (Fidler, Thomason, Cumming, Finch, & Leeman, 2004). The answer to question (b) is: Error bars are meant to provide some representation of the magnitude of probable error around a result. Two simple statistics can be used to that end: the standard error (*SE*) or a confidence interval (CI). Other, more advanced statistics can also be used (based, e.g., on Bayesian credible intervals, tolerance intervals, or likelihood regions, see, e.g., Lee, 2012; Wiens & Nilsson, 2016). Which one is chosen ultimately rests on what the authors are trying to convey as a result. However, unless there is a specific reason to prefer a different measure, error bars should preferably represent 95% CIs, as argued by, among others, Baguley (2012b), Cumming (2014), Franz and Loftus (2012), and Loftus (1996). The last answer regarding question (c), the interpretation of CIs: CIs (unlike other types of error bars) must all be interpreted in the same fashion—if a given value is within the interval of a result, the two can be informally assimilated as being comparable. This is the *golden rule of confidence intervals* and all CIs should obey this rule. Although the name of the rule is my proposal, this rule is found in many sources (e.g., DeGroot, 1989, p. 337; Neyman, 1937, p. 348).

Keep in mind that CIs are not magical wands. They are only meant to better qualify effect sizes, facilitate the detection of patterns of results, and, to a lesser extent, to attract attention to odd results or deviations that are surprisingly large. When they are correctly used, they are powerful tools to understand the results (Cumming & Fidler, 2009). Sadly, there has been some confusion in the recent years on how to interpret them (e.g., Belia, Fidler, Williams, & Cumming, 2005; Cumming & Finch, 2005) or even if they should be interpreted at all (Hoekstra, Morey, Rouder, & Wagenmakers, 2014; Morey, Hoekstra,

Corresponding author: Denis Cousineau, École de Psychologie, Université d'Ottawa, 136 Jean-Jacques Lussier, K1N 6N5 Ottawa, Canada.
E-mail: denis.cousineau@uottawa.ca

DOI • 10.5709/acp-0214-z

Rouder, Lee, & Wagenmakers, 2016; but see Miller & Ulrich, 2016) to the point that they are sometimes reported in figures but ignored in the text (Fidler et al., 2004).

The truth is that CIs are reliable as long as they are (a) built from adequate assumptions, (b) given correct information (sample size, population size, experimental design, sampling mechanism), and (c) used for the purpose they were built for (estimation of a quantity or comparison with another estimate). Although CI formulas are derived using mathematical arguments, it is easy to validate a confidence interval of the mean (CI_M) using random number generators: Generate a dataset from a simulated population with a known mean and verify that the population mean is contained within the bounds of the γ-level CI (often, γ is 95%). Sometimes, it will not be within the bounds, but over many replications the proportion of times it is will be γ.

Formally defined, a 95% CI is made in a way that in the long run, 95 out of 100 replications will return an interval which indeed contains the true population mean. Remember, however, that for a given CI, a Type I error is always a possibility.

In this article, I concentrate mostly on CI_Ms. I argue that there are different types of CI_Ms to serve the researcher's objective (compare a result to a fixed value or to other results), to match the experimental design (within-subject or between-groups), and to reflect the sampling mechanism used (simple randomized sampling or cluster randomized sampling). To avoid confusion, I propose specific names to distinguish the types of CIs. What is less known is that most CI_Ms are based on assumptions. I will highlight these assumptions and indicate how or if they can be assessed from visual inspection. I will briefly discuss the difference between the formula-based CI_M and the bootstrap CI_M. CIs are not just for mean results, they exist for any summary statistics, and I will present examples along with the relevant literature.

This article is not about the aesthetic of plots and error bars. There are discussions as to whether summary statistics are better represented by histograms or by dots and whether the extremities of error bars should be signaled by a crossbar or not. In the present article, I chose to use dots and no crossbars (see, e.g., Baguley, 2012a, and discussions linked to that web page), but the quality of a good plot is ultimately evaluated by how well it reveals the important effects. Hence, it may be necessary to try various layouts and various aesthetics to find out which one works best.

COMPUTING CONFIDENCE INTERVALS: TWO BASIC ADJUSTMENTS

Most researchers know the usual CI of the mean given by

$$[M - t_\gamma \times SE_M, M + t_\gamma \times SE_M] \qquad (1)$$

in which M is the mean of a set of observations, SE_M is the SE of that mean, and t_γ is a multiplier read from a Student t distribution with degrees of freedom given by $n - 1$ (n being the number of observations) and coverage level γ, where γ is commonly 95% (oftentimes noted in full as $t_{(1-\gamma)/2,\, n-1}$).

The *SE* of the mean is an indication of how much a sample mean is expected to vary from the population mean. All descriptive statistics have an *SE* (see later) and *SE*s are often used as a yardstick to compute CIs—as in Equation 1. The *SE* of the mean is given by

$$SE_M = s/\sqrt{n} \qquad (2)$$

where s is an estimate of the population *SD* obtained by computing the sample *SD*.

What is less known is that this type of CI has a very limited scope: It cannot be easily used to compare a mean to another mean, and it is useless for that purpose in repeated-measures designs. In this section, adjustments to Equation 1 are presented so that CI_M can be used for comparison purposes in between-group and within-subject designs.

Confidence Intervals and the Researcher's Objective

The CI_M of Equation 1 is based on the assumption that the mean will be examined in isolation. If it is compared, it is compared to fixed values—to a hypothesized population mean, for example. This fixed value has no uncertainty attached to it; hence, there is just one source of error, the sampling error of the group.

If one group mean is compared to another group mean, both the position of each mean and the relative position of one mean with respect to the other mean are uncertain. Consequently, the *SE* of a difference between two means is larger than the *SE* of the difference between one mean and a fixed value. Expanding the length of the CI_M compensates for the fact that both quantities are based on samples and, consequently, that their difference contains a larger amount of uncertainty.

How much to expand the CI depends on the variances in each condition to be compared. However, if the variances are fairly homogeneous across conditions, a simple solution exists because the sum of two identical variances amounts to multiplying a common variance by two. Consequently, the CI must be √2 ≈ 1.41 times wider (i.e., increased by 41%).[1] Thus, when the purpose of a CI is to compare a mean to other means and variances are considered homogeneous, the CI_M is given by

$$[M - t_\gamma \times \sqrt{2} \times SE_M, M + t_\gamma \times \sqrt{2} \times SE_M] \qquad (3)$$

Alternatively, if the variances are not homogeneous, use the *SE* of a difference (SE_D) instead of $\sqrt{2} \times SE_M$, which is based on the pooled *SD*:

$$SE_D = \sqrt{2} \times s_p/\sqrt{\tilde{n}} \qquad (4)$$

where $\tilde{n}$ is the harmonic mean of the groups' sample sizes; see Pfister and Janczyk (2013). If the variances are homogeneous, whether the pooled *SD* (s_p) is used (as recommended by Loftus & Masson, 1994) or each group's *SD* (s) is used (as recommended by Cousineau, 2005) is a matter of taste. If you choose both s_p and $\tilde{n}$, all the error bars will be of the same length. On the other hand, if you take *SD*s and sample sizes from each group separately, the error bars will most likely be different. Note that Equation 3 is identical to Equation 1 in all points except for the adjustment factor √2 . Such a *difference-adjusted* CI_M can only be interpreted with respect to differences between sample means using the golden rule.

Equation 1 is the CI_M when it is meant to be compared to a fixed point; Equation 3 is the CI_M when the researcher's objective is to compare one mean to other means. This adjustment was used by Hollands and Jarmasz (2010) to rephrase the golden rule: “the difference between the means of two conditions is significant if it exceeds half the total length of the CI […] multiplied by a factor of √2” (p. 135; Loftus & Masson, 1994, report a similar rule). What truly differentiates the two types of CI$_M$s in Equations 1 and 3 is the objective. This distinction was also present in Goldstein and Healy (1995), Franz and Loftus (2012), and Baguley (2012b), among others. When the term √2 is omitted, the proportion of CI$_M$ of future replications containing the true population difference is not 95% but only 83.4%, as the error bars are too short. This problem was first raised by Estes (1997) and explored by Cumming, Williams, and Fidler (2004).

It may seem counterintuitive that the error bars for differences are longer than the error bars of each mean taken individually. If the observer was to use such bars to estimate the population true mean, it is as if precision had been lost. However, remember that difference-adjusted CI$_M$s are meant to assess differences, not single means in isolation. It is therefore important that the type of CI$_M$ pictured is clearly indicated.

As an example, suppose that one member of a research group is in charge of collecting the data from a treatment group, with the hope that this group’s mean score is different from 100. After collecting the data and generating a plot showing the 95% CI$_M$ as per Equation 1, she finds that the mean seems different from 100. Indeed, the observed mean is 105.0; the 95% CI$_M$ ranges from 100.9 to 109.1 (the raw data for this example and most of the following ones are available as supplementary material so that readers can replicate the computations). If she runs a *t* test with the null hypothesis $H_0 : \mu = 100$, she finds that the null hypothesis is rejected at the .05 level, Hedge’s $g = 0.50$, $t(24) = 2.5$, $p = .02$.

A colleague measures the control group with the hope that it has a mean close to 100. He finds that the control group has a mean of precisely 100.0 (not significantly different from 100, needless to say). The CI obtained from Equation 1 is [95.8, 104.2] and does not include the mean of the treatment group.

If they merge the datasets, they will be surprised to find that a two-sample *t* test indicates no significant difference at the .05 level, $g = 0.50$, $t(48) = 1.76$, $p = .085$. The left panel of Figure 1 shows the plot they produced (in both groups, the *SD*s are approximately 10.0).

Because their objective is to compare both groups, they increase the length of both CIs by a factor of 1.41. Figure 1, middle panel, shows the results using CI$_M$ based on the SE_D (Equation 3). Here, because one mean is included in the CI of the other mean, the difference between them can informally be assimilated to an absence of difference, congruent with the result of the *t* test.

Alternatively, and as recommended by many, for example, Cumming (2014) and Franz and Loftus (2012), they could have made a plot of the difference in mean score, as shown in the last panel of Figure 1. This approach is explained fully in Pfister and Janczyk (2013). However, for designs with multiple groups, the number of pairwise differences increases very rapidly. For three or four groups, it is still possible to show on a single plot all the pairwise differences; one example is illustrated in Figure 2. Beyond that, the benefit of the pairwise difference plot is dubious, as seen if you compare the left panels of Figure 2 with the right panels.

One critique that can be addressed to these adjusted CIs is that they do not provide an estimate of the population mean for a given group. This critique is relatively correct. However, in Psychology, it can be argued that we are rarely interested in estimating a population mean in isolation. As Loftus and Masson (1994) put it, “in psychological experiments, it is rare […] for one to be genuinely interested in inferring the specific value of a population mean. More typically, one is interested in inferring the *pattern* formed by a *set* of population means” (p. 480, the authors’ emphasis).

Because an absolute estimate of mean performance is utopian, psychologists spend considerable time and resources measuring control groups, placebo groups, pre-treatment scores, and other forms of baseline scores, separately for any new experiment. These design requirements should be mirrored by equivalent estimates meant to highlight patterns of results. This is the purpose of the adjusted CIs.

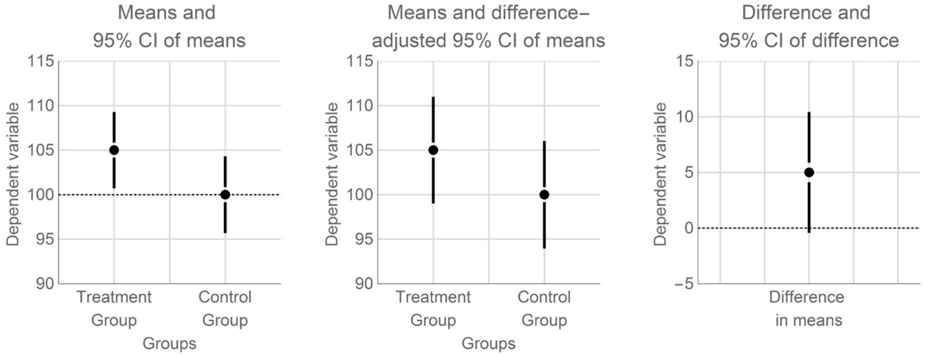

FIGURE 1.

Example mean plots from two independent groups. Left: The error bars show the 95% CI of the means (CI$_M$s); middle: The error bars show the difference-adjusted 95% CI$_M$s; right: The difference between groups is shown, with the 95% CI of the difference. The raw data are available in the supplementary material.

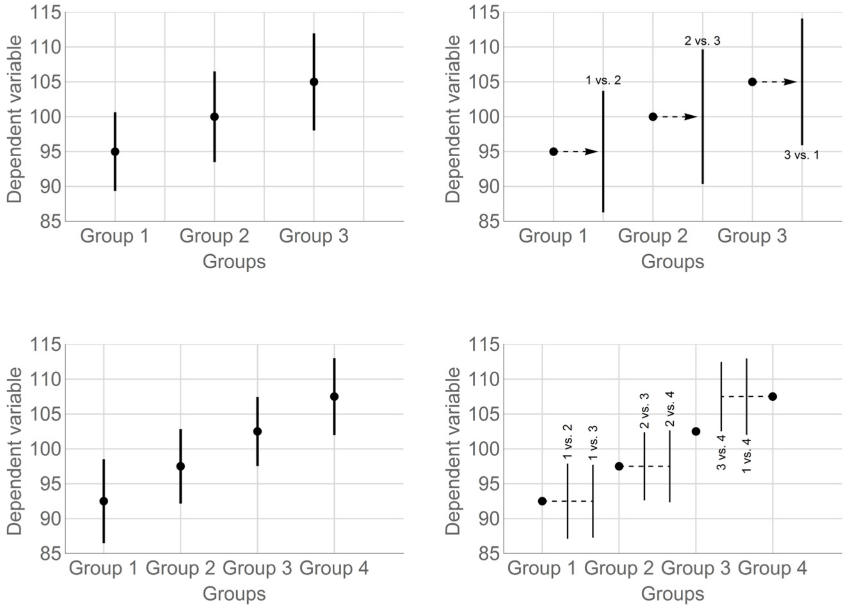

FIGURE 2.

Example mean plots for a three-groups design (top) and a four-groups design (bottom) with error bars showing difference-adjusted 95% CI of the means (CI_Ms) (left) and 95% CI of the difference for all pairwise differences (right).

Confidence Intervals and the Experimental Design

Experimental designs can be divided as to whether they are between-groups or within-subject (mixed designs will be discussed in a later section). In a within-subject (or repeated-measures) design, the participants' scores are typically positively correlated. By considering such correlations in the participants' scores, it is possible to evaluate differences between two means more precisely, a fact little known (Belia et al., 2005).

In a two repeated-measures design, the correction in length is equal to $\sqrt{1-r}$ when the variances are homogeneous, in which *r* is Pearson's correlation, such that the CI_M for a repeated-measures design is

$$[\ \mathbf{M} - t_\gamma \times \sqrt{1-r} \times \sqrt{2} \times \frac{s}{\sqrt{n}}, \mathbf{M} + t_\gamma \times \sqrt{1-r} \times \sqrt{2} \times \frac{s}{\sqrt{n}}\ ] \tag{5a}$$

An alternative way to understand the correlation adjustment is to note that in Equation 2, the square root of the sample size is replaced by $\sqrt{n}/\sqrt{1-r}$ to obtain Equation 5a, so that the ratio $n/1-r$ can be termed *the effective sample size*. The stronger the correlation is, the more accurate the regression slope is. Consequently, the difference between the two means is estimated as if we had measured a larger sample. With a sample size of 25 and a correlation of .8, for example, the effective sample size is five times larger than the true sample size (as $n/1-r$ = 25/0.2 = 125).

When there are more than two measurements, there is no universally accepted way to get a CI_M adjusted to within-subject correlations. The difficulty owes to the fact that the variances are not perfectly identical between groups and the correlations are not perfectly identical between pairs of groups. One method (Bakeman & McArthur, 1996; Cousineau, 2005; see Morey, 2008, for the appropriate correction for bias) is to obtain a transformed dataset *Z* derived from the original data set, such that within-subject correlation is removed. Then, the CI_M is obtained as usual using the *SE* from the transformed data set rather than from the original data set.

Note that the *correlation-adjusted* CI_M must always be difference-adjusted as well, as implicitly, the two groups are compared in getting a correlation. Thus, the *SE* of *Z* must be increased by a √2 factor as well. In general, for two or more repeated measures, the CI_M is given by

$$[\,\mathbf{M} - t_\gamma \times \sqrt{2} \times SE_Z, \mathbf{M} + t_\gamma \times \sqrt{2} \times SE_Z\,] \tag{5b}$$

It is thus a correlation-adjusted as well as a difference-adjusted CI of the mean.

Cousineau and O'Brien (2014) give more details on how to compute the transformed data set *Z*. Masson and Loftus (2003; see also Loftus & Masson, 1994) provide an alternative approach. Both methods are identical when variances and correlation are truly homogeneous between measurements. Baguley (2012b) and Franz and Loftus (2012) evaluated these and other propositions.

As another example, a researcher gets data from a sample of 25 participants in a repeated-measures design (for example a pre-post design). The mean of the first measurement is 105.0 and the mean of the second measurement is 100.0. Both *SD*s are near 15.0. The error bars obtained from Equation 1 are shown in the left panel of Figure 3. There does not seem to be any difference between the two measures, yet a paired *t* test indicates a strong and significant difference (Cohen's d_z = 0.58, *t*[24] = 2.90, *p* = .008). The researcher, remembering that his objective is to compare the two measures, may switch to a difference-adjusted CI_M (Equation 3), but things would get worse as CI_Ms for differences are √2 times longer as seen in Figure 3, second panel. The apparent inconsistency between the CI and the statistical test owes to the fact that this is a repeated-measures design: Participants' scores are correlated. In the present dataset, the correlation between the pairs of scores is .84, so that $\sqrt{1-0.84} = 0.4$. Hence, in this case, the CI_M taking into account correlation should be 40% the length of the error bars, based on independent samples (more than halved). The third panel of Figure 3 shows the resulting, correct CI_M. If you prefer the paired difference CI, it is shown in the last panel of Figure 3.

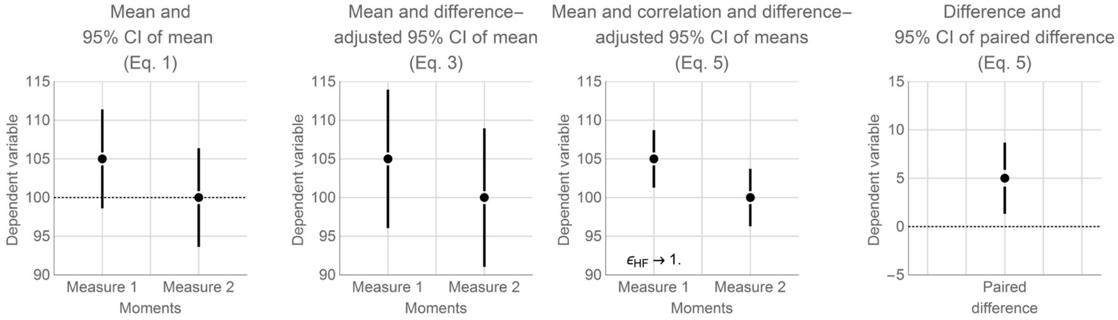

FIGURE 3.

Example mean plots for two repeated measures. First panel: The error bars show the 95% CI of the means (CI_Ms); second panel: The error bars show the difference-adjusted 95% CI_Ms; third panel: the error bars show the correlation and difference-adjusted 95% CI_Ms; fourth panel: paired difference and 95% CI of the paired difference. ε_{HF} is discussed later in the text. The raw data are available in the supplementary material.

When the data are correlated, the CIs are shortened as within-subject correlation is used to better estimate the difference across means; the more positively correlated the data are, the shorter the CI_M becomes. In the unlikely event that the data are negatively correlated, the CI_M is expanded by the correlation adjustment.

Naming Convention

At this time, the three types of CI_M (unadjusted, difference-adjusted, and correlation- and difference-adjusted) have no distinct names. It is therefore difficult in a figure caption to figure out which type is plotted. A common statement is “the error bars are corrected for within-subject variability” followed by a reference, for example, “Loftus and Masson (1994).” I propose the following three labels:

- *CIs of the means* (Equation 1)
- *Difference-adjusted CIs of the means* (Equation 3)
- *Correlation- and difference-adjusted CIs of the means* (Equation 5)

Loftus and Masson (1994), contrary to Baguley (2012b), recommend the use of the pooled *SE* so that the label representing their approach could be *correlation and difference-adjusted pooled 95% CIs of the mean*. As typically, the purpose of mean plots is to compare means to other means, the difference-adjusted CI_Ms would be used most often. If unadjusted CI_Ms are used and there exists a conventional reference point, that point of reference could be present on the plot with a dashed line, for example, as was done in the first panels of Figures 1 and 3; no reference point should be shown when difference-adjusted CI_Ms are depicted.

Rouder and Morey (2005) suggested the expressions *arelational* (unadjusted, Equation 1) and *relational* (all the other CI_Ms proposed here). These authors noted that “there are many advantages to arelational CIs: They provide a rough guide to variability in data, a coarse view of replicability of patterns and a quick check of the heterogeneity of variances. Arelational CIs, however, do not reflect between-group information and cannot be used for direct comparisons” (p. 77).

Pfister and Janczyk (2013) also proposed a naming convention which applies when the difference between two means is plotted. They coined the expressions *CI of means* (unadjusted CI_M), *CI of differences* (for between-group difference in means), and *CI of paired difference* (for within-subject difference in means). The naming convention is important so that the type of CI shown on plots can be identified unambiguously.

In addition to naming the CI_Ms, it is useful to have a uniform way of reporting them when the authors want to write down the CI_M. Following Cumming (2014) and the American Psychological Association Publication Manual (2009), brackets should be used to denote 95% CIs. The notation $M \pm \mathrm{CI}_M$ should not be used, as CIs are not always symmetrical. CIs are symmetrical for central tendencies (the mean, the median, the geometric and the harmonic means) and some nonparametric statistics of dispersion (median absolute deviation and interquartile range). However, in general, they are not symmetrical, as, for example, the CI of the *SD* and the CI of the kurtosis. Conversely, *SE*s are always symmetrical, so the notation $M \pm SE_M$ makes sense and should be used exclusively to report *SE*s.

Finally, *SE* should not be used for the length of the error bars in plots. They are not easy to interpret (but see Cumming & Finch, 2001, 2005) and the fact that *SE*s are always symmetrical may yield a false impression. For example, suppose a group of 20 data has an *SD* of 12.33. The *SE* of the *SD* in that case is 2.00. The 95% CI of that *SD* is [9.38, 18.0]. There is no single number which added to and subtracted from 12.33 can yield this interval. Further, the asymmetry in precision would go unnoticed if *SE* were reported.

CONFIDENCE INTERVALS AND HIDDEN ASSUMPTIONS

CIs are well known (albeit not universally used). However, one thing that might be less known is that CI estimates are not assumption free. On the one hand, the use of the *SE* of the mean, $SE_M = s/\sqrt{n}$, rests on few, quite general assumptions. CIs, on the other hand, are based on the assumption that the means are normally distributed. Indeed, to obtain a CI, the *SE* is multiplied by a *t* value which is based on this assumption. Owing to the central limit theorem, large-sample means should meet this assumption; for smaller samples, one safeguard, prior to drawing a mean plot, could be to run a test of normality (e.g., a Kolmogorov-Smirnov test) or tests for null skewness and null kurtosis and fail to reject the null (if the sample size is small) or find mild deviations (if the sample size is moderate; Rochon, Gondan, & Kieser, 2012).

Likewise, and as we saw, the difference-adjusted CI_M is based on the homogeneity of variances assumption. This assumption can be checked visually when the groups are of the same size: As a given CI_M is based on the *SD* of that group of data only, the length of the error bars should all be of a comparable size. If there are important differences in length, then there is certainly a problem with the assumption of homogeneity of variances. Figure 4, left panel, shows an example with equal sample sizes. As the CI_Ms are of very different lengths, it can be inferred that the variances are not homogeneous, and therefore the difference-adjusted CI_M should not be relied upon strongly. As a rule of thumb and for samples of moderate sizes, if the variance in one group is twice the variance in another group, Levene's test will likely detect heterogeneity (and indeed it did in Figure 4, left panel: $F = 6.72$, $p = .013$). In terms of CI_Ms, as they are based on *SD*s (square roots of variances), a CI_M which is 40% longer than another one suggests heterogeneity of variances.

For repeated-measures designs, the correlation-adjusted CI_Ms are based on the sphericity assumption; loosely speaking, this is similar to a homogeneity of correlation assumption (Baguley, 2004; Lane, 2016, are more precise). This is true whether a method based on separate estimates (such as Cousineau, 2005; Morey, 2008) or based on a pooled estimate (such as Loftus & Masson, 1994) is used. Sadly, this assumption cannot be verified visually with error bars. For example, the CI_Ms may be of very different lengths and yet sphericity still holds (Baguley, 2004; Huynh, 1978). One solution is to compute epsilon (ε), a measure of sphericity whose value is between $1 / (J - 1)$ and 1, where J is the number of repeated measures; ε of 1 means that the data are perfectly spherical, that is, that the CI_Ms are accurate. Some authors consider that εs above .9 indicate a mild deviation from perfect sphericity (see Field, 2013; Tabachnick & Fidell, 1996, for more on this measure). The ε measure was originally created by Greenhouse and Geisser (1959); Huynh and Feldt (1976) provided a formula corrected for bias.[2] Figure 4, right panel, shows the means for three measurements (an example based on Baguley, 2004). Visually, the CI_Ms are of unequal length, but this information in not relevant as this is a repeated-measures design. The Huynh-Feldt ε is 1, which indicates that the sphericity assumption holds for this data set (and indeed, a Mauchly test of sphericity does not reject the null hypothesis of sphericity, $W = 0.947$, $\chi^2[2] = 1.24$, $p = .55$). Because it is possible to have a visually educated guess with regards to homogeneity of variances in between-subjects designs, I suggest that the Huynh-Feldt ε (ε_{HF}) be always visible on a mean plot showing repeated measures (when there are three or more repeated measures; with only two measures, sphericity always holds; Lane, 2016).

If there is any problem with the assumptions (normality and either homogeneity of variances for between-group designs or sphericity for within-subject designs), the assumption-based CI_Ms might nevertheless be used as visual tools to provide rough intuitions on the results. However, if statistical inference is important, they should not be used. Alternatively, it is also possible to use bootstrap estimates of CI_M (Efron & Tibshirani, 1993). The basic algorithm for bootstrap estimation is simple:[4]

Given a sample of size n:

1. Subsample the sample, extracting n data with replacement from the original sample.

2. Compute on this subsample the statistic desired (e.g., the mean for a CI_M).

3. Repeat Steps 1 and 2 a very large number of times (e.g., 10,000 times).

4a. Finally, obtain the CI_M by locating the bounds within which a proportion γ of the subsample statistics are located.

4b. Alternatively, if you want *SE* instead, compute the *SD* of all the subsample statistics.

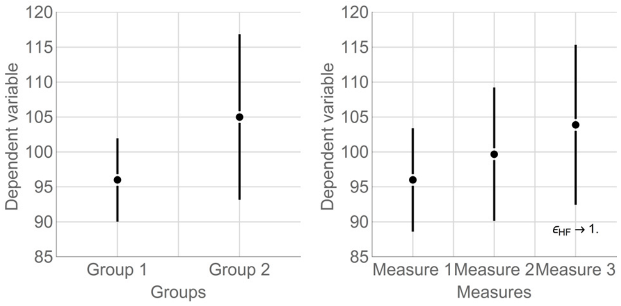

FIGURE 4.

Example mean plots for two experiments with sample size 25 per condition. Left: means of two independent groups with error bars showing difference-adjusted 95% CI of the means (CI_Ms). The two groups have different variances, as evidenced by the error bars of unequal length. Right: means from three measures with error bars showing correlation- and difference-adjusted 95% CI_Ms. Although the measures' variances are different, the data do not violate the sphericity assumption, as evidenced by a Huynh-Feldt ε of 1. In the left panel, we can test the difference in means using the Welch test, a *t* test whose degrees of freedom are corrected to handle heterogeneity of variances; the difference is borderline not significant, $g = 0.40$, $t(35.2) = 2.01$, $p = .052$. In the right panel, the analysis of variance (ANOVA) is significant, $\eta^2 = 0.12$, $F(2, 48) = 3.34$, $p = .04$. Post hoc analyses show that the difference between Measure 1 and Measure 3 (8 points of separation) is the only significant one ($p = .026$).

Bootstrap estimates should be based on a large number of subsamples (minimally 10,000, but more if your platform can run it); as a consequence, they are slower to obtain than the formula-based intervals.

Bootstrap CIs are based on fairly mild assumptions about the underlying population distribution (e.g., Shao & Tu, 1995).[3] The sample should be reasonably large, although there is no explicit prescription as to what *large* means precisely. One safe rule is to at least match the sample size recommended from power computations (Mayr, Erdfelder, Buchner, & Faul, 2007). When the assumptions are met, bootstrap CI returns on average the same interval as the formula-based CI. One disadvantage of bootstrap estimates is that their exact value is different every time they are computed. This is why they must be based on a large number of subsamples. With 10,000 subsamples, the first two digits should be stable, so do not report bootstrap estimates with more than two significant digits or increase the number of subsamples. More sophisticated bootstrap algorithms have been developed (see, e.g., *BCa*; Efron, 1987; or *ABC*; DiCiccio & Efron, 1996).

Figure 5 shows simulated data with three groups, in which black lines show the formula-based CI$_M$ and gray lines show the bootstrap-based CI$_M$. The data were simulated from a normal distribution with means of 97, 100, and 103 and a common *SD* of 15. For a large sample (200 in Figure 5, right panel), the difference between the two types of approach to estimating CI$_M$ is immaterial.

COMPUTING CONFIDENCE INTERVALS: ADVANCED ADJUSTMENTS

All the CI and *SE* formulas given in the present article are valid for experimental designs examining a population of infinite size using simple randomized sampling. Yet Little (2004) strongly encouraged researchers to incorporate the sampling mechanisms in their models. Consequently, this information should also be incorporated in the CI by using sampling adjustments. Here, I illustrate how this can be done when the population is not so large as to be considered infinite, when a different sampling mechanism is used, or both.

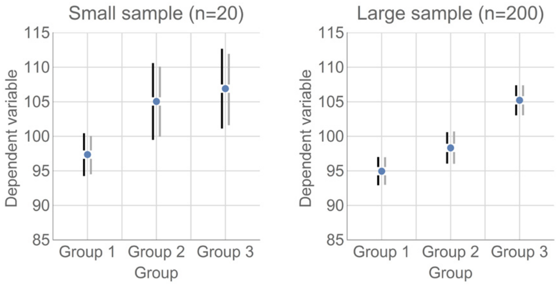

FIGURE 5.

Example mean plots from three independent groups with error bars showing difference-adjusted 95% CI of the mean (CI$_M$s) obtained from formula-based estimates (Equation 3; black bars) or from bootstrap estimates (gray bars). Left: a small sample (*n* of 20 per group); right: a large sample (*n* of 200 per group). The raw data for the left panel are available in the supplementary material.

Confidence Intervals and the Population Size

When the sample represents a sizeable proportion of the whole population, it is not possible to consider the population as infinite. Examples where the population cannot be considered infinite include: a study of employees within a given company, the LGBT community in a linguistic minority, or students' achievements in public schools. Regarding the last example, the Austrian government aims to assess 20% of the population every year.

As discussed in Cochran (1953), when the sample size exceeds 5% of the population size, a finite population correction must be applied to the sample estimates of variability (see also Thompson, 2012). In the following example, let *n* denote the sample size and *N* denote the population size. The adjustment is based on the proportion of elements not sampled from the population, 1− *n*/*N* so that the CI$_M$ adjusted for population size becomes

$$\left[\mathbf{M} - t_\gamma \times \sqrt{1-\frac{n}{N}} \times SE_\mathbf{M}, \mathbf{M} + t_\gamma \times \sqrt{1-\frac{n}{N}} \times SE_\mathbf{M}\right] \qquad (6)$$

in the case where there are no other adjustments. As *n* tends to *N*, there is less and less uncertainty in the estimated variance of the population so that the adjustment factor tends to zero and the CIs shrink to null.

The adjustments for finite sample size can be used jointly with the correlation adjustment and the difference adjustment.

Confidence Intervals and the Sampling Method

In simple randomized sampling, all the participants are chosen randomly from the studied population with an equal chance of being selected. Other sampling techniques exist, such as cluster randomized sampling and stratified sampling (Kish, 1965; Thompson, 2012). Cluster randomized sampling is often used in educational psychology and consists, for example, of picking whole classes from schools. The children are not selected with equal chances; the classes are. Stratified sampling is often used for survey studies and consists in selecting individuals, such that the sample is representative of the population on certain control variable(s), on age categories, for example.

Regarding cluster randomized sampling, Cousineau and Laurencelle (2015) provided a *cluster-adjusted* CI$_M$. It requires an estimate of the intraclass correlation. The cluster adjustment can be used in conjunction with difference and correlation adjustments. Likewise, Lai, Kwok, Hsiao, and Cao (in press) argue that the correction for cluster randomized sampling can be used in conjunction with the correction for finite population size. The detailed computation of this adjustment is given in Appendix A.

For stratified sampling techniques and other sampling techniques, the expression of *SE*s and CIs are not agreed-upon and most require numerical algorithms so that a simple adjustment does not seem possible at this time.

As seen, considerations related to sampling methods are easily handled using additional adjustments that are simply multiplied to the CI length.

VARIOUS CONSIDERATIONS

Visualizing Confidence Intervals in Mixed Designs

The fact that CIs are different in between-groups designs and in within-subject designs is problematic for mixed designs where both types of CI_Ms coexist. In this case, the researcher may choose to plot just one type of CI_M, the one which captures the results he or she wants to concentrate on. If the researcher wants to show both types of CI_Ms, Baguley (2012b) proposed the use of *two-tiered error bars*. These error bars are drawn with two sets of aesthetics: The ones delimited with a cross-line (often the shortest) are correlation- and difference-adjusted CI_Ms (and related to the within-subject results); the ones without cross-lines (often the longest) are the difference-adjusted CI_Ms (and related to the between-subjects results). Although this solution is ingenious, the plots are often harder to interpret. The CIs are meant to synthetize results so that they are more easily apprehended. Multiplying the number of bars only achieves the opposite effect. As a general rule, the number of error bars should be kept to a minimum. If both between-groups and within-subject CIs are important, consider presenting two distinct plots.[5]

Software for Computing Confidence Intervals for the Means and Other Statistics

Typically used summary statistics, not just means, all have *SE*s and CIs (see Harding, Tremblay, & Cousineau, 2014, for a review). Hence, all summary plots should be drawn with some measure of dispersion around them, the conventional measure being 95% CIs. As an example, Figure 6 illustrates 95% CIs for nine descriptive statistics, including robust and nonparametric statistics (the median, the median absolute deviation, and the Pearson skew; Daszykowski, Kaczmarek, Vander Heyden, & Walczak, 2007; Harding, Tremblay, & Cousineau, 2015; Siegel & Castellan, 1988).

At this time, there is no statistical package that implements the adjustments to CI_Ms of the means. SPSS can only draw unadjusted CIs for many descriptive statistics; an extension to SPSS (O'Brien & Cousineau, 2014) implements both correlation-adjusted and difference-adjusted CI_Ms. Likewise, R has no standard commands to draw adjusted CI_Ms, but Baguley (2012b) programmed commands to that end and Kelley (2017) made the MBESS R library with CIs for a few statistics such as effect sizes. A standalone application, MorePower, can compute CIs for a few within-subject and between-subjects designs (Campbell & Thompson, 2012). Finally, a Mathematica package, available from the author, is briefly described in Supplementary Material.

There are a number of references in which the computation of CIs are given and described. Beaulieu-Prévost (2006) reported how to compute the unadjusted CI_M as well as the difference-adjusted CI_M; also, the CI of the Pearson's correlation *r* is given. Finally, CIs for a proportion *p* and for difference-adjusted proportions are given. Cumming and Fidler (2009) reported some of the above, and also CIs for the Hedges' effect size *g*. Harding et al. (2014) reviewed *SE*s and CIs for an exhaustive list of descriptive statistics: (central tendency) mean, median, geometric and harmonic means, (dispersion) variance, *SD*, median absolute deviation and interquartile range, (shape) Fisher skew, and kurtosis. Harding et al. (2015) gave *SE*s and CIs for the Pearson skew.

Bootstrap estimates are fairly easy to obtain in SPSS with the module BOOTSTRAP, sold separately from SPSS (version 19 and above) or the module GSD (Harding & Cousineau, 2016). Otherwise, Weaver and Koopman (2014) showed how to bootstrap estimates of CIs for Pearson's correlation with SPSS; Hallgren (2013) showed how to perform bootstrapping in general in the R environment. Finally, Hélie (2006) provided a general introduction to the topic of model selection using bootstrap.

Few commands provide the full flexibility needed to plot any summary statistics in conjunction with any type of CIs. I hope that this situation will change rapidly so that researchers are encouraged to plot adjusted CIs routinely.

GENERAL DISCUSSION

All the CIs reviewed here are summarized in Algorithm 1. Also, the relevant formulas are provided in Appendix A. They all obey the golden rule of interpretation for CIs: If a given value is within the interval of a result, the two can be informally assimilated as being comparable.

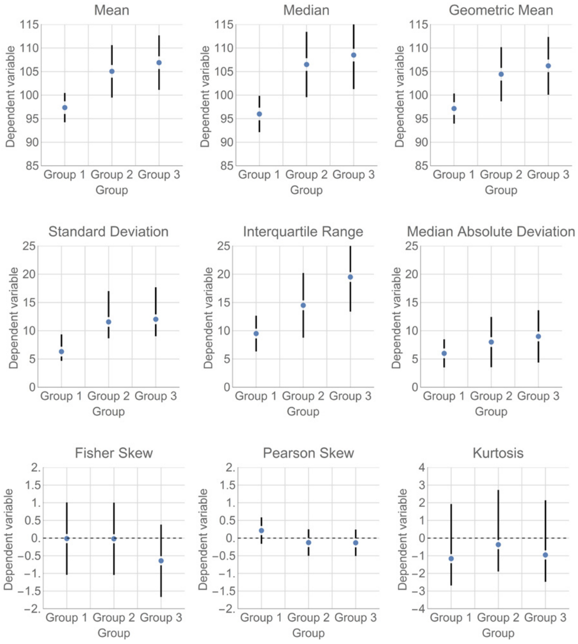

FIGURE 6.

Plots of various statistics from fictitious data as a function of group with error bars showing 95% CIs. The CIs are asymmetrical for *SD* and kurtosis. The first six are difference-adjusted; the last three shows unadjusted CIs in which zero is the reference. The same data set is used in all panels and were used in Figure 5, left panel (so that the first panel shows the same results in both figures).

By making all CIs follow the same and unique interpretative rule, researchers might start relying on these statistics more frequently, more consistently, and more confidently.

```
Algorithm 1
Steps to compute SEs and CIs of means

1- Are the data from a within-subject design or mixed design?
   Yes: decorrelate the data within each group (Equation A4).
2- Compute SEs for each group and each measure (Equation 2).
   Do you want to pool the SEs? Yes: use Equation A5.
   Do you want to pool the sample sizes? Yes: use harmonic mean.
3- Do you want to show CI instead of SE?
   Yes: Choose your confidence level (typically 95%) and get $t_\gamma$
    then multiply $SE$ by the multiplier $t_\gamma$ (Equation A3).
4- Purpose: Will comparisons be made to other sample means?
   Yes: Use difference adjustment (Equation A6).
5- Sampling mechanisms.
   a- Is the population of finite size? Yes: Equation A7.
   b- Is the sample obtained from cluster randomized sampling?
   Yes: use Equations A8 and A9.
6- Place the CI about the mean (Equation A1).
```

Some have argued that Equations 1, 3, and 5 (5a or 5b) are not three different types of CIs, but just one type of CI for three different statistics (Equation 1 is the CI_M of a single mean, Equation 3 is akin to the CI_M for the difference between two independent means, and Equation 5b is the CI_M for the within-subject difference in means). I do not object to this point of view and if it is more intuitive to the readers, please make these the labels by which you identify the intervals in your future communications. The only thing that really matters is that anything having the name *confidence interval* should be interpreted in a consistent and universal fashion, that is, according to the golden rule.

CIs should be part of any plots or listed in tables of results whenever a summary statistic is reported. There exists a CI for any statistic you may want to report and many can be found in the literature. Although CIs are not always clearly understood in very formal ways (see Belia et al., 2005; Cumming et al., 2004; Hoekstra et al., 2014), I believe that they are more intuitive than other kinds of statistical information. See, among others, Loftus (1996) for a similar point of view. If we can agree on the golden rule and make sure that all CIs plotted conform to it consistently and systematically, intuition regarding them should improve. Previous texts have not sought to enforce uniformity by discussing error bars based on *SE* or by promoting half-length intervals. *Half-length CIs* were suggested by Baguley (2012b), Franz and Loftus (2012), and Goldstein and Healy (1995), by which the length of the difference-adjusted CI_M is divided by 2. Such half-length CIs must be interpreted differently as it is the presence of overlap between error bars that signals comparable means. This is unfortunate; if we want researchers to develop the correct automatisms when facing error bars, we must devise intervals that are to be interpreted consistently (Shiffrin & Schneider, 1977).

CIs are the result of solid mathematical arguments. They provide an interval which likely contains the population means. Indeed, just to take an example, 95% of the 95% CIs of the means do contain the population mean. There is no guarantee that one specific CI_M contains the population mean, but we may have a certain confidence that this is the case (Miller & Ulrich, 2016).

Note that a CI is accurate only if the assumptions are correct, only if the experimental design and sampling methods are inscribed in it, and only if it is used for the correct objective. If any of these elements are changed, the CI length will change accordingly (as was shown in Morey et al., 2016). It is not a demonstration that CIs are fallacious; it is a demonstration that CIs must be informed as accurately and as completely as possible.

The only arbitrary aspect of CIs is the coverage level γ used to compute t_γ. The purpose of this quantity is to provide a reasonably large coverage for the interval. On the one hand, too narrow an interval could yield the impression that a study is hardly replicable (even if replications are scarce within Psychology; see Cousineau, 2014; Jasny, Chin, Chong, & Vignieri, 2011; Makel, Plucker, & Hegarty, 2012; Pashler & Wagenmakers, 2012). On the other hand, too wide an interval would bring little information with respect to the true characteristic(s) of a population. A conventional level is required; Cumming (2014; also see publication policy of the *Psychological Science* journal regarding statistics) argued that a 95% coverage level is a reasonable position (Marmolejo-Ramos & Cousineau, 2016).

Finally, keep in mind that ultimately, good science should return short CIs. Being able to assess patterns of means is important, as argued in the Introduction. However, being able to assess results with high precision is also, if not more, important.

Along this document, I made a few recommendations that I reiterate here:

1. always show or list CIs whenever results based on summary statistics are given;

2. use formula-based CIs if the assumptions are not rejected by the data of if it is conventional to do so in the area of research; use bootstrap CIs otherwise;

3. prefer difference-adjusted CI_Ms if focus is on the pattern of results; if unadjusted CI_Ms are given and there is a conventional reference value, provide the reference value on the plot (e.g., with a dashed line);

4. in the text, use the notation [low, high] for 95% CI_Ms. Use the notation ± to denote *SE*s.

5. in plots showing means in a within-subject design, provide the Huynh-Feldt ε so that readers can assess whether the sphericity assumption holds or not. In between-subjects designs, the reader can assess the homogeneity of variances assumption visually by comparing the length of the error bars.

6. if half-length CIs are used, clearly identify this fact and give the rule for interpreting these.

As mentioned by Belia et al. (2005), "better guidelines for researchers and less ambiguous graphical conventions are needed before the advantages of CIs for research communication can be realized" (p. 389). I hope that this article is one step further in that direction.

FOOTNOTES

[1] The result can also be seen in the denominator of the two-sample t test. Some authors write the two-sample t test as Reject H_0 if

$$|M_1 - M_2| \,/\, \left(s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\right) > t_\gamma$$

where n_1 and n_2 are the two groups' sample sizes and s_p is the pooled *SD*. However, note that $\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$ can be simplified into $\sqrt{2}/\sqrt{\tilde{n}}$, where $\tilde{n}$ is the harmonic mean of the number of participants in the two groups so that the t test becomes Reject H_0 if $|M_1 - M_2| \,/\, \left(\sqrt{2}s_p/\sqrt{\tilde{n}}\right) > t_\gamma$. Using this formulation, the √2 adjustment is evident.

[2] A warning to SPSS users wishing to compute the Huynh-Feldt ε: Consult Lecoutre, 1991, and Dalgaard, 2007, pp. 3-4.

[3] To be formal, this approach is called a *non-parametric bootstrap estimation*. Bootstraps which incorporate some properties of the population distribution are called *parametric bootstrap estimations*.

[4] One restriction to bootstrap estimation is that this method cannot be used to estimate lower bound or upper bound parameters. The core of bootstrapping is that it should be possible to underestimate the true parameter on some subsamples, and overestimate the true parameter on other subsamples. With boundary parameters, such as an upper bound, it is not possible to overestimate this parameter using observed data so that nonparametric bootstrap is not applicable (Bickel & Freedman, 1981).

[5] Thanks to an anonymous reviewer for suggesting this solution.

AUTHOR NOTE

I would like to thank Thom Baguley, Geoff Cumming, Stéphane Dufau, Bradley Harding, Rink Hoekstra, Roland Pfister, Cynthia Wan, and two anonymous reviewers for their comments on an earlier version of this text. Requests for reprint should be addressed to Denis Cousineau using e-mail at: denis.cousineau@uottawa.ca.

REFERENCES

American, Psychological Association (2009). *Publication Manual of the American Psychological Association, Sixth Edition*. ISBN-13: 978-1433805622

Baguley, T. (2004). *An introduction to sphericity*. Retrieved from http://homepages.gold.ac.uk/aphome/spheric.html.

Baguley, T. (2012a). *The aesthetics of error bars*, [Internet resource]. Retrieved from http://psychologicalstatistics.blogspot.fr/2012/05/aesthetics-of-error-bars.html, last consulted 29/June/2015.

Baguley, T. (2012b). Calculating and graphing within-subject confidence intervals for ANOVA. *Behavior Research Methods, 44*, 158-175. doi: 10.3758/s13428-011-0123-7 WWW

Bakeman, R., & McArthur, D. (1996). Picturing repeated measures: Comments on Loftus, Morrison and others. *Behavior Research Methods, Instruments, & Computers, 28*, 584-589. doi: 10.3758/BF03200546

Beaulieu-Prévost, D. (2006). Confidence Intervals: From tests of statistical significance to confidence intervals, range hypotheses and substantial effects. *Tutorials in Quantitative Methods for Psychology, 2*, 11-19. doi: 10.20982/tqmp.02.1.p011

Belia, S., Fidler, F., Williams, J., & Cumming, G. (2005). Researchers misunderstand confidence intervals and standard error bars. *Psychological Methods, 10*, 389-396. doi: 10.1037/1082-989-X.10.4.389 WWW

Bickel, P., & Freedman, D. (1981). Some asymptotic theory for the bootstrap. *Annals of Statistics, 9*, 1196–1217.

Campbell, J. I. D., & Thompson, V. A. (2012). MorePower 6.0 for ANOVA with relational confidence intervals and Bayesian analysis. *Behavior Research Methods, 44*, 1255-1265. doi: 10.3758/s13428-012-0186-0 WWW

Cochran, W. G. (1953). *Sampling techniques*. New York, NY: Wiley.

Cousineau, D. (2005). Confidence intervals in within-subject designs: A simpler solution to Loftus and Masson's method. *Tutorials in Quantitative Methods for Psychology, 1*, 42-45. doi: 10.20982/tqmp.01.1.p042

Cousineau, D. (2014). Restoring confidence in psychological science findings: A call for direct replication studies. *The Quantitative Methods for Psychology, 10*, 77-79. doi:10.20982/tqmp.10.2.p077

Cousineau, D., & Laurencelle, L. (2015). A correction factor for the impact of cluster randomized sampling and its applications. *Psychological Methods, 21*, 121-135. doi: 10.1037/met0000055 WWW

Cousineau, D., & O'Brien, F. (2014). Error bars in within-subject designs: A comment on Baguley (2012). *Behavior Research Methods, 46*, 1-3. doi: 10.3758/s13428-013-0441-z WWW

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science, 25*, 7-29. doi:10.1177/0956797613504966

Cumming, G., & Fidler, F. (2009). Confidence intervals: Better answers to better questions. *Journal of Psychology, 217*, 15-26. doi: 10.1027/0044-3409.217.1.15

Cumming, G., & Finch, S. (2001). A primer on the understanding, use, and calculation of confidence intervals that are based on central and noncentral distributions. *Educational and Psychological Measurement, 61*, 532-574. doi: 10.1177/00131640121971374

Cumming, G., & Finch, S. (2005). Inference by eye: Confidence intervals and how to read pictures of data. *American Psychologist, 60*, 170-180. doi: 10.1037/0003-066X.60.2.170 WWW

Cumming, G., Williams, J., & Fidler, F. (2004). Replication and researchers' understanding of confidence intervals and standard error bars. *Understanding Statistics, 3*, 299-311. doi: 10.1207/s15328031us0304_5

Dalgaard, P. (2007). New functions for multivariate analysis. *R News, 7*, 2-7.

Daszykowski, M., Kaczmarek, K., Vander Heyden, Y., & Walczak, B. (2007). Robust statistics in data analysis - A review: Basic concepts. *Chemometrics and Intelligent Laboratory Systems, 85*, 203-219. doi: 10.1016/j.chemolab.2006.06.016

DeGroot, M. H. (1989). *Probability and statistics* (2nd ed.). Menlo Park, CA: Addison-Wesley.

DiCiccio, T. J., & Efron, B. (1996). Bootstrap confidence intervals. *Statistical Science, 11*, 189-228. doi: 10.1016/0167-7152-(88)90042-9

Efron, B. (1987). Better bootstrap confidence intervals. *Journal of the American Statistical Association, 82*, 171-185. doi: 10.2307/2289144

Efron, B., & Tibshirani, R. (1993). *An introduction to the bootstrap*. New York, NY: Chapman & Hall.

Estes, W. K. (1997). On the communication of information by displays of standard errors and confidence intervals. *Psychonomic Bulletin & Review, 4*, 330-341. doi:10.3758/BF03210790

Fidler, F., Thomason, N., Cumming, G., Finch, S., & Leeman, J. (2004). Editors can lead researchers to confidence intervals, but can't make them think: Statistical reform lessons from medicine. *Psychological Science, 15*, 119-126. doi: 10.1111/j.0963-7214.2004.01502008.x WWW

Field, A. (2013). *Discovering statistics using IBM SPSS Statistics*. New York, NY: Sage.

Franz, V. H., & Loftus, G. R. (2012). Standard errors and confidence intervals in within-subjects designs: Generalizing Loftus and Masson (1994) and avoiding the biases of alternative accounts. *Psychonomic Bulletin & Review, 19*, 395-404. doi:10.3758/s13423-012-0230-1 WWW

Goldstein, H., & Healy, M. J. R. (1995). The graphical presentation of a collection of means. *Journal of the Royal Statistical Society A, 158*, 175-177. doi: 10.2307/2983411

Greenhouse, S. W., & Geisser, S. (1959). On methods in the analysis of profile data. *Psychometrika, 24*, 95-112.

Hallgren, K. A. (2013). Conducting simulation studies in the R programming environment. *Tutorials in Quantitative Methods for Psychology, 9*, 43-60. WWW

Harding, B., & Cousineau, D. (2016). GSD: An SPSS extension command for sub-sampling and bootstrapping datasets. *The Quantitative Methods for Psychology, 12*, 145-153. doi: 10.20982/tqmp.12.2.p145

Harding, B., Tremblay, C., & Cousineau, D. (2014). Standard errors: A review and evaluation of standard error estimators using Monte Carlo simulations. *The Quantitative Methods for Psychology, 10*, 107-123. doi: 10.20982/tqmp.10.2.p107

Harding, B., Tremblay, C., Cousineau, D. (2015). The standard error of the Pearson skew. *The Quantitative Methods for Psychology, 11*, 32-37. doi: 10.20982/tqmp.11.1.p032

Hélie, S. (2006). An introduction to model selections: Tools and algorithms. *Tutorials in Quantitative Methods for Psychology, 2*, 1-10. doi: 10.20982/tqmp.02.1.p001

Hoekstra, R., Morey, R. D., Rouder, J. N., & Wagenmakers, E. J. (2014). Robust misinterpretation of confidence invervals. *Psychonomic Bulletin & Review, 21*, 1157-1164. doi: 10.3758/s13423-013-0572-3 WWW

Hollands, J. G., & Jarmasz, J. (2010). Revisiting confidence intervals for repeated measures designs. *Psychonomic Bulletin & Review, 17*, 135-138. doi: 10.3758/PBR.17.1.135 WWW

Huynh, H. (1978). Some approximate tests for repeated measurement designs. *Psychometrika, 43*, 161-175. doi: 10.1007/BF02293860

Huynh, H., & Feldt, L. S. (1976). Estimation of the Box correction for degrees of freedom from sample data in randomized block and split-splot designs. *Journal of Educational Statistics, 1*, 69-82. doi: 10.3102/10769986001001069

Jasny, B. R., Chin, G., Chong, L., & Vignieri, S. (2011, December). Again, and again, and again. *Science, 334*(6060), 1225-1225. doi: 10.1126/science.334.6060.1225

Kelley, K. (2017). MBESS (version 4.2.0) [Computer software]. Retrieved from http://nd.edu/~kkelley/site/MBESS.html.

Kish, L. (1965). *Survey sampling*. New York, NY: John Wiley & Sons Inc.

Lai, M. H. C., Kwok, O., Hsiao, Y. Y., & Cao, Q. (in press). Finite population correction for two-level hierarchical linear models. *Psychological Methods*. WWW

Lane, D. (2016). The assumption of sphericity in repeated-measures designs: What it means and what to do when it is violated. *The Quantitative Methods for Psychology, 12*, 114-122. doi: 10.20982/tqmp.12.2.p114

Lecoutre, B. (1991). A correction for the $\tilde{\varepsilon}$ approximate test in repeated measures designs with two or more independent groups. *Journal of Educational Statistics, 16*, 371-372. doi: 10.3102/10769986016004371

Lee, P. M. (2012). *Bayesian statistics: An introduction* (4th ed.). New York, NY: Wiley-Blackwell.

Little, R. J. (2004). To model or not to model? Competing modes of inference for finite population sampling. *Journal of the American Statistical Association, 99*, 546-556. doi: 10.1198/016214504000000467

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. (1996). Psychology will be a much better science when we change the way we analyse data. *Current Directions in Psychological Science, 5*, 161-171. doi:10.1111/1467-8721.ep11512376

Loftus, G. R., & Masson, M. E. J. (1994). Using confidence intervals in within-subject designs. *Psychonomic Bulletin & Review, 1*, 476-490. doi: 10.3758/BF03210951 WWW

Makel, M. C., Plucker, J. A., & Hegarty, B. (2012). Replications in psychology research: How often do they really occur? *Perspectives on Psychological Science, 7*, 537-542. doi: 10.1177/1745691612460688 WWW

Marmolejo-Ramos, F., & Cousineau, D. (2016). Perspectives on the use of null hypothesis statistical testing. Part II: Is null hypothesis statistical testing an irregular bulk of masonry? *Educational and Psychological Measurement*, online first. doi: 10.1177/0013164416667987

Masson, M. E. J., & Loftus, G. R. (2003). Using confidence intervals for graphically-based data interpretation. *Canadian Journal of Experimental Psychology, 57*, 203-220. doi: 10.1037/h0087426 WWW

Mayr, S., Erdfelder, E., Buchner, A., & Faul, F. (2007). A short tutorial of GPower. *Tutorials in Quantitative Methods for Psychology, 3*, 51-59. doi: 10.20982/tqmp.03.2.p051

Miller, J., & Ulrich, R. (2016). Interpreting confidence intervals: A comment on Hoekstra, Morey, Rouder, and Wagenmakers. *Psychonomic Bulletin & Review, 23*, 124-130. doi:10.3758/s13423-015-0859-7 WWW

Morey, R. D. (2008). Confidence intervals from normalized data: A correction to Cousineau (2005). *Tutorials in Quantitative Methods for Psychology, 4*, 61-64. doi:10.20982/tqmp.04.2.p062

Morey, R. D., Hoekstra, R., Rouder, J. N., Lee, M. D., & Wagenmakers, E.-J. (2016). The fallacy of placing confidence in confidence intervals. *Psychonomic Bulletin & Review, 23*, 103-123. doi: 10.3758/s13423-015-0947-8 WWW

Neyman, J. (1937). Outline of a theory of statistical estimation based on the classical theory of probability. *Philosophical Transactions of the Royal Society of London. Series A, Mathematical and Physical Sciences, 236*, 333–380.

O'Brien, F., & Cousineau, D. (2014). Representing error bars in within-subject designs in typical software packages. *The Quantitative Methods for Psychology, 10*, 56-67. doi: 10.20982/tqmp.10.1.p056 WWW

Pashler, H., & Wagenmakers, E.-J. (2012). Editor's introduction to the special section on replicability in psychological science: A crisis of confidence? *Perspectives on Psychological Science, 7*, 528-530. doi: 10.1177/1745691612465253 WWW

Pfister, R., & Janczyk, M. (2013). Confidence intervals for two sample means: Calculation, interpretation, and a few simple rules. *Advances in Cognitive Psychology, 9*, 74-80. doi: 10.2478/v10053-008-0133-x WWW

Rochon, J., Gondan, M., & Kieser, M. (2012). To test or not to test: Preliminary assessment of normality when comparing two independent samples. *BMC Medical Research Methodology, 12*, 1-11. doi: 10.1186/1471-2288-12-81

Rouder, J. N., & Morey, R. D. (2005). Relational and arelational confidence intervals: A comment on Fidler, Thomason, Cumming, Finch, and Leeman (2004). *Psychological Science, 16*, 77-79. doi: 10.1111/j.0956-7976.2005.00783.x WWW

Siegel, S., & Castellan, N. J. (1988). *Nonparametric statistics for the behavioral sciences*. New York, NY: McGraw Hill.

Shao, J., & Tu, D. (1995). *The jackknife and bootstrap*. Berlin, Germany: Springer.

Shiffrin, R. M., & Schneider, W. (1977). Controlled and automatic human information processing: II Perceptual learning, automatic attending, and a general theory. *Psychological Review, 84*, 127-190. doi: 10.1037/0033-295X.84.2.127

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86*, 420-428. doi: 10.1037/0033-2909.86.2.420 WWW

Tabachnick, B. G., & Fidell, L. S. (1996). *Using multivariate statistics* (3rd ed.). New York, NY: Harper Collins.

Thompson, S. K. (2012). *Sampling* (3rd edition). New York, NY: Wiley.

Weaver, B., & Koopman, R. (2014). An SPSS macro to compute confidence intervals for pearson's correlation. *The Quantitative Methods for Psychology, 10*, 29-39. doi: 10.20982/tqmp.10.1.p029

Wiens, S., & Nilsson, M. E. (2016). Performing contrast analysis in factorial designs: From NHST to confidence intervals and beyond. *Educational and Psychological Measurement*, online first. doi: 10.1177/0013164416668950

Wilkinson, L., & the Task, Force on Statistical Inference (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54*, 594-604. doi: 10.1037/h0027060

RECEIVED 19.12.2016 | ACCEPTED 11.04.2017

APPENDIX A

SUMMARY OF THE FORMULAS

In this appendix, L is used to denote the length of the CI from the mean, that is, the distance from one bound to the mean. CIs of means are always symmetrical so that both arms of the error bar are of equal length. Thus, in what follows:

$$\text{CI} = [M - \text{L}, M + \text{L}] \qquad \text{(A1)}$$

Confidence Interval Base-length in Between-Group Designs

Given a set of observations in a certain group or condition containing n observations, compute

$$SE_M = s/\sqrt{n} \qquad \text{(A2)}$$

$$\text{L} = SE_M \times t_\gamma \qquad \text{(A3)}$$

in which s is the SD of the observations, t_γ is a multiplier based on the confidence level desired and on the degrees of freedom n – 1.

Confidence Interval Base-Length in Within-Subject Designs

Here, X_{sj} is the score of subject s for the j^{th} measure; $\overline{X}_{s.}$ is the mean score of subject s and $\overline{X}_{..}$ is the grand mean. Finally, J is the number of repeated measures. In a mixed design, apply the transformations for all groups separately.

$$Y_{sj} = X_{sj} - \overline{X}_{s.} + \overline{X}_{..} \qquad \text{(A4a)}$$

$$Z_{sj} = \sqrt{\frac{J}{J-1}} \times (Y_{sj} - \overline{Y}_{.j}) + \overline{Y}_{.j} \qquad \text{(A4b)}$$

then compute L from the transformed dataset Z using Equations A2 and A3.

Pooling the Standard Deviations

If you choose to pool the SDs (as recommended by Loftus and Masson, 1994), replace s in Equation A2 with s_p:

$$s_p = \sqrt{\frac{\sum_{i=1}^{J} d_{fi} s_i^2}{\sum_{i=1}^{J} d_{fi}}} \qquad \text{(A5)}$$

where s_i is the SD of the data (raw if between-groups data or transformed if within-subject data) in condition i, J is the number of groups or measurements, and $df_i = n_i - 1$ is the degree of freedom for measurement i. This is a simple weighted average of the (squared) SD.

Difference Adjustment

$$\text{Multiply L by } \sqrt{2} \qquad \text{(A6)}$$

Finite-Population Adjustment

$$\text{Multiply } L \text{ by } \sqrt{1 - \frac{n}{N}} \qquad \text{(A7)}$$

This adjustment will shorten the length of the error bars. As N tends to infinity, the term $\sqrt{1 - n/N}$ tends to 1 so that for large N relative to n, this adjustment can be ignored.

Cluster Adjustment

Suppose that the group contains k clusters of m subjects ($k \times m = n$). The intra-class correlation, noted by ρ, must be estimated first (Shrout and Fleiss, 1979). The adjustment factor is given by λ

$$\lambda = \frac{1 + (n-1)\rho}{1 - \frac{n-1}{kn-1}\rho} \qquad \text{(A8)}$$

$$\text{Multiply } L \text{ by } \lambda \qquad \text{(A9)}$$

The value of λ is always larger than 1, reflecting the well-known fact that cluster randomized samples have less precision than simple randomized samples (Kish, 1965; Cousineau & Laurencelle, 2015).

APPENDIX B

Using a *Mathematica* Package to Make Summary Statistic Plots With Error Bars

These are commands to make a summary statistic plot using the package `MeanPlot` for Mathematica, available from the author. The options controlling aesthetics are not presented. Note that *Mathematica* is case sensitive. All the data files used next are tab-separated text files containing information using one line per subject. Only the information to be plotted must be present in the file and group membership must always be in the first column(s).

LOADING THE PACKAGE

Load the package; you must specify as the second parameter the location of the file “MeanPlot.m” on your computer, doubling the backslash if your operating system requires such character.

```
Needs[
    "MeanPlot`",
    "C:\\Users\\DenisCousineau\\Documents\\
MeanPlot.m"
]
```

MAKING FIGURE 1

Figure 1 is done in two different ways, first using unadjusted CIs, second, using difference-adjusted CIs.

```
X = Import["DataFigure1.tsv"];

MeanPlot[X,
  BetweenSubjectFactors -> {{"Groups", {1 ->
 "Treatment group",
      2 -> "Control group"}}},
  ErrorBarContent -> CI,
  PlotRange -> {90, 115}
]
MeanPlot[X,
    BetweenSubjectFactors -> {{"Groups", {1 ->
    "Treatment group",
        2 -> "Control group"}}},
    ErrorBarContent -> CI,
    Adjustments -> {Objective -> Difference},
    PlotRange -> {90, 115}
]
```

In the above, replace `CI` with `SE` to use *SE*s for the length of the error bars. MeanPlot default is to plot CIs so that the option `ErrorBarContent -> CI` is optional and omitted in the following. The confidence level can be adjusted by adding the option `Gamma -> level` (default is 0.95). The option `Adjustments -> {Objective -> Single}` is equivalent to no adjustments and is the default.

MAKING FIGURE 3

Figure 3 is based on a 2-measure within-subject design so that the data are organized in two columns.

```
X = Import["DataFigure3.tsv"];
MeanPlot[X,
    WithinSubjectFactors -> {{"Moments", {1 ->
    "Moment 1",
        2 -> "Moment 2"}}},
    Adjustments -> {Objective -> Difference,
    RepeatedMeasures -> CM},
    PlotRange -> {90, 115}
]
```

Instead of the Cousineau-Morey method (CM), it is possible to take the Loftus and Masson’s pooled estimate by replacing `CM` with `LM` in the above instruction.

MAKING FIGURE 5

To make Figure 5, replace the regular (assumption-based) CIs (`ErrorBarContent -> CI`) for an approach using bootstrap (with 10,000 subsamples with `ErrorBarContent -> (CI[Mean,#, .95, Algorithm-> {Bootstrap, 10000}]&))`.

```
X = Import["DataFigure5.tsv"];

MeanPlot[X,
    BetweenSubjectFactors -> {{"Groups", {1 ->
    "Group 1",
        2 -> "Group 2", 3 -> "Group 3"}}},
    ErrorBarContent -> (CI[Mean, #, .95,
        Algorithm -> {Bootstrap, 10000}] &),
    Adjustments -> {Objective -> Difference},
    PlotRange -> {85, 115},
    PlotLabel -> "Bootstrap-based"
]
```

MAKING FIGURE 6

In Figure 6, different summary statistics than the default (Mean) can be specified using the `SummaryStatistic` option. Here, the data of Figure 5 are used again.

```
X = Import["DataFigure5.tsv"];
MeanPlot[X,
    BetweenSubjectFactors -> {{"Moments", {1 ->
"Group 1",
        2 -> "Group 2", 3 -> "Group 3"}}},
    SummaryStatistic -> Median,
    Adjustments -> {Objective -> Difference},
    PlotRange -> {85, 115},
    PlotLabel -> "Median"
]
```

Instead of `Median`, the identifiers `Mean`, `StandardDeviation`, `Variance`, `HarmonicMean`, `GeometricMean`, `MedianDeviation`, `InterquartileRange`, `SkewnessU`, `SkewnessP`, and `KurtosisU`. can be used; other functions can also be defined by the user. The option `SummaryStatistic -> Mean` can be omitted as was done in the previous examples as it is the default.

APPENDIX C

Datasets Used for the Figures

Data set for Figure 1: A 2-group design (*n* is 25 in each group).

1	117
1	103
1	113
1	101
1	104
1	114
1	111
1	103
1	110
1	118
1	103
1	113
1	119
1	92
1	98
1	93
1	111
1	103
1	105
1	109
1	117
1	107
1	92
1	82
1	87
2	118
2	107
2	102
2	99
2	93
2	110
2	83
2	88
2	96
2	111
2	97
2	102
2	103
2	109
2	85
2	89
2	93
2	98
2	101
2	101
2	91
2	126
2	92
2	103
2	103

Data set for Figure 3: A repeated-measures design with two measurements (*n* is 25)

128	105
96	96
102	88
88	80
83	90
99	86
126	122
129	140
103	94
125	115
100	100
88	94
91	86
115	111
95	89
112	111
109	110
92	104
116	101
85	80
108	103
116	96
115	111
123	115
81	73

Dataset for Figure 5: A three-group design (*n* is 20 in each group).

1	93
1	102
1	105
1	103
1	86
1	102
1	92
1	90
1	108
1	94
1	94
1	98
1	95
1	105
1	93
1	95
1	105
1	102
1	97
1	88
2	107
2	113
2	81
2	107
2	122
2	97
2	120
2	111
2	98
2	99
2	89
2	95
2	108
2	100
2	95
2	108
2	99
2	121
2	125
2	106
3	121
3	116
3	87
3	117
3	103
3	93
3	120
3	107
3	119
3	110
3	106
3	120
3	84
3	118
3	104
3	92
3	112
3	107
3	113
3	89