Behav Res (2014) 46:1149–1151
DOI 10.3758/s13428-013-0441-z

BRIEF COMMUNICATION

Error bars in within-subject designs: a comment on Baguley (2012)

Denis Cousineau · Fearghal O'Brien

Published online: 30 January 2014

Abstract The problem of calculating error bars in within-subject designs has proven to be a difficult problem and has received much attention in recent years. Baguley (*Behavior Research Methods*, *44*, 158–175, 2012) recommended what he called the Cousineau–Morey method. This method requires two steps: first, centering the data set in a certain way to remove between-subject differences and, second, integrating a correction factor to debias the standard errors obtained from the normalized data set. However, within some statistical packages, it can be difficult to integrate this correction factor. Baguley (2012) proposed a solution that works well in most statistical packages in which the alpha level is altered to incorporate the correction factor. However, with this solution, it is possible to plot confidence intervals, but not standard errors. Here, we propose a second solution that can return confidence intervals or standard error bars in a mean plot.

Keywords Statistics · Statistical inference · Mean plots · Confidence intervals · Within-subject designs

Error bars in mean plots are now recognized as being as useful as *p* values, and many articles recommend that they be included (see, e.g., Loftus, 1996, and Wilkinson and the Task Force on Statistical Inference, 1999, among others). Error bars can represent standard errors or confidence intervals (CIs) for a certain level (typically, 95 % CI). Cumming and Finch (2005) provided rules of thumb (more precisely, *rules of eye*) to improve researchers' intuitive grasp of these aids. Still, it is common to view a mean plot (or any descriptive statistic plot, for that matter) in which there are no error bars. As an example, in the first issue of volume 75 of *Attention*, *Perception*, & *Psychophysics* published in 2013 (http://blog.apastyle.org/apastyle/2012/09/citinga-whole-periodical.html), for 57 figures presenting summary statistics, 16 (28 %) did not contain error bars. One possible reason that error bars are not used more frequently is that error bars depend not only on the data, but also on the researcher's objectives and the experimental design. The objective can be to compare means with a target value or to compare means with each other (called difference-adjusted intervals in Baguley, 2012; see also Franz & Loftus, 2012). The experimental design can be within subjects, between subjects, or a mix of the two. Whereas standard errors and CIs are well understood for between-subject designs (and implemented in most statistical packages), their implementation in within-subject designs and mixed designs is still debated (and no statistical package has these displaying options).

The computation of error bars for within-subject designs began with the seminal work of Loftus and Masson (1994). Recently, Baguley (2012) provided a review of the recent propositions. He suggested the use of what he called the Cousineau–Morey method when the researcher is interested in differences between means (and when sample sizes are not too small; for very small sample sizes, the method proposed by Loftus & Masson, 1994, which uses pooled variance estimates, should be preferred). The Cousineau–Morey method can be seen as a two-step method in which (1) the data are "normalized" in such a way that the between-subject differences are removed (Cousineau, 2005), and (2) a correction factor is used to correct the estimates (it depends only on the number of repeated measures) because the standard errors from this normalized set are biased downward (Morey, 2008).

In his review, Baguley (2012) provided one method to obtain CIs integrating the correction factor that works in most statistical packages. This solution is correct and fairly simple to use, and the author provided example code on how to implement it is SPSS and R. However, it can work only in conjunction with CIs. If standard errors are to be plotted, this solution is not applicable.

In what follows, we review Baguley's (2012) solution and propose an alternative approach of greater generalizability

D. Cousineau (✉)
École de psychologie, Université d'Ottawa, 136 Jean-Jacques Lussier, K1N 6N5 Ottawa, Canada
e-mail: denis.cousineau@uottawa.ca

F. O'Brien
School of Psychology, Aras an Phiarsigh, Trinity College, Dublin 2, Ireland
e-mail: obrienfk@tcd.ie

that can be used whether CIs or standard errors are wished for.

Incorporating the correction factor via the alpha level

We explain the approach by looking at the CI equation, assuming that a repeated measures design was used, in which the participants are measured *J* times:

$$\mathrm{CI}_{1-\alpha} = \overline{\mathrm{X}}_{.j} \pm \mathrm{SE}_{\mathrm{Y}.j} \times \sqrt{\frac{J}{J-1}} \times t_{n-1}(\alpha/2). \quad (1)$$

In this equation, $\overline{\mathrm{X}}_{.j}$ represents the mean obtained in the *j*th level of the treatment, the ratio $\sqrt{J/(J-1)}$ is the correction factor, and $\mathrm{SE}_{\mathrm{Y}.j}$ is the standard error of the mean for that level obtained from the normalized data set. The standard error is computed as usual, $\mathrm{SE} = s/\sqrt{n}$, in which *s* is the standard deviation of the scores and *n* is the number of subjects. To obtain the normalized data, use the following transformation:

$$\mathrm{Y}_{sj} = \mathrm{X}_{sj} - \overline{\mathrm{X}}_{s.} + \overline{\mathrm{X}}_{..}, \quad (2)$$

in which Y_{sj} is the transformed score for subject *s* in condition *j*, X_{sj} is the original score of the *s*th participant in the *j*th condition, $\overline{\mathrm{X}}_{s.}$ is the mean for the participant across the conditions, and $\overline{\mathrm{X}}_{..}$ is the overall mean.

The difficulty is that prepackaged software does not allow the introduction of a correction factor when performing a plot with error bars. In fact, the only quantity that can be specified is the alpha level. To get around this difficulty, the solution proposed was to consider simultaneously the last two terms in Eq. 1 and find an adjusted alpha level α^* so that the result would correspond to the desired value:

$$t_{n-1}(\alpha^*/2) = \sqrt{\frac{J}{J-1}} \times t_{n-1}(\alpha/2). \quad (3a)$$

In Eq. 3a, *t* is a quantile function (given a probability level, it returns the critical value); the inverse of a quantile function is the cumulative probability function (given a critical value, it returns the probability of the occurrence of this value or less), which will be noted here as t^{-1}. Since the inverse exists, it is possible to isolate α^* in the above, and we find:

$$\alpha^* = 2t_{n-1}^{-1}\left(\sqrt{\frac{J}{J-1}} \times t_{n-1}(\alpha/2)\right). \quad (3b)$$

Hence, if a plot of the normalized data set is requested with CIs of an alpha level given by α^*, the result will be CIs integrating the correction factor.

The above is precisely the solution proposed by Baguley (2012). Although it works fine, it is, however, impossible to make a plot of standard errors, since they are not corrected; this approach leaves SE_{Y} unaffected.

An alternative approach

In what follows, we present an alternative approach that consists of performing a second "normalization" of the data set, with the purpose of reducing the standard error to the correct level. Assuming that the set **Y** was obtained from **X**, we now go from **Y** to a new data set **Z**. The last data set will incorporate the correction factor so that any plot of the means on **Z** will draw proper error bars, be it standard errors or CIs. In the case of CIs, it is not necessary with this approach to alter the alpha level.

The new set of transformed data **Z** can be obtained from **Y** with:

$$\mathrm{Z}_{sj} = \sqrt{\frac{J}{J-1}} \times \left(\mathrm{Y}_{sj} - \overline{\mathrm{Y}}_{.j}\right) + \overline{\mathrm{Y}}_{.j}, \quad (4)$$

where Z_{sj} is the new score of participant *s* in the condition *j*, $\overline{\mathrm{Y}}_{.j}$ is the mean for the *j*th condition across participants, and $\sqrt{J/(J-1)}$ is the correction factor described above (Morey, 2008). As a result of this transformation,

$$\mathrm{SE}_{\mathrm{Z}} = \mathrm{SE}_{\mathrm{Y}} \times \sqrt{\frac{J}{J-1}}. \quad (5)$$

This approach corrects for bias by reducing the spread of the data. It consists of first centering the data at zero, changing the spread of the data using the correction factor, and finally undoing the centering. With this manipulation, the means of **Z** are the same as the means of the original data **X**, but the spread has been modified (by Eqs. 2 and 4). As such, a mean plot on **Z** displaying error bars will show the correct within-subject error bars of the data **X**.

This alternative approach is easy to implement, as long as it is possible to manipulate data sets to normalize them in various ways (a graphical user interface for SPSS generating mean plots with within-subject error bars, described in O'Brien and Cousineau, 2014, uses this approach).

Discussion

The present comment discussed a simple approach to obtaining CIs appropriate for within-subject designs. It is adequate for obtaining standard errors or CIs (contrary to the approach suggested in Baguley, 2012). Another advantage of the present approach is that it can be used to obtain difference-adjusted intervals (Baguley, 2012, Franz & Loftus, 2012, Tryon, 2001).[1] These intervals are corrected by an additional

[1] We Thank Thom Baguley for pointing this out.

correction factor, $1/\sqrt{2}$, so that if 95 % CIs are drawn, the means are not different at a decision threshold of .05 if the CI in one condition contain the mean of another condition; conversely, the means are different if the CI in one condition does not contain the mean of another condition. This additional correction factor can easily be integrated in Eq. 4 using $\sqrt{J/(2(J-1))}$ instead of $\sqrt{J/(J-1)}$.

The Cousineau–Morey approach introduced an accessible way to plot error bars of various kinds in mean plots when repeated measure designs are used. Still, the discussion is far from over. First, as Franz and Loftus (2012) correctly noted, such an approach requires that the sphericity assumption be valid (the same is true for some of the propositions in Loftus & Masson, 1994). Hence, a mean plot of within-subject design data should always report a measure of sphericity such as the Huynh–Feldt epsilon (1976), although one should beware, since some popular statistical packages compute this statistic incorrectly (see Lecoutre, 1991). This measure of sphericity should be above 0.70 at the very least. See Franz and Loftus (2012) for alternative propositions.

Second, mixed designs involve both within- and between-group treatments. In this case, we end up with two different standard errors and, consequently, two different CIs depending on whether the conditions are compared across measures or across groups. Baguley (2012) suggested the use of two-tiered error bars in which, ticks show both error bars. This solution has the advantage that if the ticks are equal for the between and within error bars, it implies that there is no correlation between the participants' scores. However, the presence of two sets of ticks on each error bar could potentially be misleading.

Other alternatives were discussed in Franz and Loftus (2012); at some point, completeness of the picture must be weighed against parsimony of the representation. As has been pointed out by Loftus and Masson (1994), Baguley (2012), and others, these error bars are not exactly equivalent to a statistical test and are not meant to replace them. We can use error bars to complement statistical tests if the patterns are very clear (large effects) or for uninteresting/noncritical hypotheses. Future discussions are required to decide precisely what is expected from error bars.

Author Notes We would like to thank Bradley Harding, Christophe Tremblay, Thom Baguley, and an anonymous reviewer for their comments on an earlier version of this text.

References

Baguley, T. (2012). Calculating and graphing within-subject confidence intervals for ANOVA. *Behavior Research Methods, 44,* 158–175.

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,* 71–75.

Cumming, G., & Finch, S. (2005). Inference by eye: Confidence intervals and how to read pictures of data. *American Psychologist, 60,* 170–180.

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.

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.

Lecoutre, B. (1991). A correction for the epsilon_tilde approximate test in repeated measures designs with two or more independent groups. *Journal of Educational Statistics, 16,* 371–372.

Loftus, G. R. (1996). Psychology will be a much better science when we change the way we analyze data. *Current Directions in Psycholgical Science, 5,* 161–171.

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

Morey, R. D. (2008). Confidence Intervals from Normalized Data: A correction to Cousineau (2005). *Tutorials in Quantitative Methods for Psychology, 4,* 61–64.

O'Brien, F., & Cousineau, D. (2014) Representing error bars in within-subject designs in typical software packages. *The Quantitative Methods for Psychology,* 58–70.

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.

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