Psychonomic Bulletin & Review
2010, 17 (1), 135-138
doi:10.3758/PBR.17.1.135

Notes and Comment

Revisiting confidence intervals for repeated measures designs

Justin G. Hollands and Jerzy Jarmasz
Defence Research and Development Canada,
Toronto, Ontario, Canada

Loftus and Masson (1994) proposed a method for computing confidence intervals (CIs) in repeated measures (RM) designs and later proposed that RM CIs for factorial designs should be based on number of observations rather than number of participants (Masson & Loftus, 2003). However, determining the correct number of observations for a particular effect can be complicated, given that its value depends on the relation between the effect and the overall design. To address this, we recently defined a general number-of-observations principle, explained why it obtains, and provided step-by-step instructions for constructing CIs for various effect types (Jarmasz & Hollands, 2009). In this note, we provide a brief summary of our approach.

Confidence intervals (CIs) can be used to draw statistical inferences about differences between conditions in an experiment. Indeed, CIs bear a systematic relationship to significance tests. The difference between the means of two conditions is significant if it exceeds half the total length of the CI (called the *margin of error*) multiplied by a factor of $\sqrt{2}$ (Loftus & Masson, 1994). A corresponding visual rule of thumb states that when the margins of error for two means overlap by less than half, the difference is significant. Thus, graphing CIs serves as a highly useful technique for determining whether the various conditions involved in an effect are different from one another.

The procedures for computing CIs for independent measures (IMs; also called *between-subjects*) designs are well known (e.g., Kirk, 1982; Loftus & Loftus, 1988). In contrast, methods for computing CIs in repeated measures (RMs; also called *within-subjects*) designs have a shorter history. For a long time, there were no published methods for computing RM CIs. However, in 1994, Loftus and Masson published a landmark article in *Psychonomic Bulletin & Review*, showing how RM CIs could be computed, how they are related to RM significance tests, and how they can be used for making inferences about RM effects. This article has served as an important reference for the large number of experimental psychologists who use RM designs on a frequent basis.

In IM designs, a CI can be computed around the mean for each condition, and the size of the CI is affected by the number of participants contributing to each mean. This assumes one observation per participant. For RM designs, a participant serves in multiple conditions, thus generating multiple observations. How does one determine the number of observations for a particular effect in an RM design? The literature on RM CIs has not provided a comprehensive guide to computing the number of observations for factorial designs. Indeed, as Cumming and Finch (2005) noted, it is unclear how well CIs can be effectively used by researchers to understand effects within complex experimental designs. In RM designs, the number of observations is affected by the number of participants, the nature of the effect, and the overall design of the experiment. Thus, there has been a need to clarify the procedure for computing the number of observations for a variety of designs and effects.

In a recent article (Jarmasz & Hollands, 2009), we have specified the procedures for computing CIs across the range of effects found in factorial RM and IM–RM designs, following the Loftus and Masson (1994; Masson & Loftus, 2003) approach. We reviewed the use of CIs for inferential purposes, explained the difficulties in correctly obtaining the number of observations, and developed a general method for obtaining the number of observations for various effects, which we coined the *number-of-observations* principle. We also provided examples illustrating the use of the principle to compute CIs for various effects. The purpose of the present brief note is to provide a synopsis of the key implications for the psychological researcher.

The Importance of Number of Observations

Loftus and Masson (1994) showed that CIs can be constructed for statistical inference involving RM effects by using the following formula:

$$\mathrm{CI} = M_i \pm t_{\mathrm{critical}} \times \sqrt{\frac{MS_{\mathrm{e}}}{N}}, \qquad (1)$$

where M_i is the sample mean for condition i, t_{critical} is a Student's t value, MS_{e} is the mean square error for the effect in question, and N is the number of participants in the experiment. The value of t_{critical} is determined by the degrees of freedom associated with the MS_{e} term and the confidence level selected for the CI (typically set to 95%). This equation is appropriate for one-way RM designs. However, in a factorial RM design, *the number of observations contributing to each mean is not simply the number of participants*. For any given level of a specific factor,

J. G. Hollands, justin.hollands@drdc-rddc.gc.ca

the number of observations is affected by the number of levels of all other factors considered in the analysis.

In factorial designs, using the number of participants, rather than the number of observations, to construct CIs produces CIs that are too big. That is, they do not correspond to the results from a significance test. However, until RM CIs were developed, many researchers used the number of participants to compute CIs. Perhaps a greater familiarity with IM CIs led psychologists to use number of participants for the value of *N*. Indeed, although Masson and Loftus (2003) properly defined *N* as number of observations, some of the CIs for RM factorial designs in their example used the number of participants instead. This was corrected later (Masson, 2004).

The logic of computing CIs on the basis of number of observations extends to IM effects and IM × RM interactions in mixed IM–RM designs. For IM effects, the number of observations per participant is determined by the number of levels of all RM factors in the analysis. For IM × RM interactions, the number of observations per participant is determined by the number of levels of all RM factors in the analysis not involved in the interaction. Again, our experience has been that this is not well understood, and the approach for determining the correct number of observations is not being followed when CIs are plotted for mixed IM–RM designs.

In general, to obtain the CIs for a given effect (main effect or interaction) in a factorial RM design, the MS_e for that effect must be divided by the number of observations. The number of observations for a given effect is the number of participants multiplied by the product of the number of levels of all other RM factors in the design. Multiplying the number of participants by the number of levels of all RM factors in the design and then dividing the result by the number of levels associated with the effect produces the same result. Thus, we define an effect's *number of observations* as the number of participants multiplied by the product of the number of levels of all factors, divided by the number of levels for the effect of interest. We can compute CIs for an RM Effect R from a factorial RM ANOVA as follows:

$$\mathrm{CI} = M_i \pm t_{\mathrm{critical}} \times \sqrt{\frac{MS_{\mathrm{R\times S}}}{N \times \frac{L}{r}}}, \qquad (2)$$

where $MS_{\mathrm{R\times S}}$ is the MS_e for R, *L* is the product of the levels of all RM factors in the analysis, and *r* is the number of levels for R. If Effect R is an interaction, then *r* represents the product of the number of levels of all the factors involved in the interaction.

Thus, Equation 2 shows that $N \times (L/r)$ is the number of observations for Effect R. It is only when $r = L$ (as occurs for the highest level interaction or a one-way design) that the CI will equal the value obtained by dividing by *N* only.

As an example, if you have a data set organized for a 2 × 3 × 3 RM ANOVA, the data set will contain scores for all RM factors; that is, it will contain $2 \times 3 \times 3 = 18$ observations for each participant. If you are interested in computing a CI for the 3 × 3 interaction, the full data set contains the number of cells for the effect (9) times the number of observations (2) per participant. Thus, computing CIs for this effect involves dividing its MS_e by the relevant number of observations (twice the number of participants), this last value being nine times smaller than the full data set. In contrast, if you are interested in computing the CI for the main effect having two levels, the number of observations per participant will be $18/2 = 9$, or half the size of the full data set. Thus, the number of observations per participant represents the number of levels of factors in the overall design not considered within (collapsed across) the effect of interest. It is not the number of levels of factors within the effect itself. This is counterintuitive, since one might think of the number of observations as the number of cells for the effect of interest times the number of participants. Thus, multiplying *N* by the number of cells for the effect of interest and using this value as the number of observations per participant will yield incorrect CIs. Similarly, computing the total number of conditions in the design and dividing by the number of nuisance variables (variables not involved in the effect of interest) will also produce incorrect CIs.

More generally, the number of observations for a particular effect is determined by the relation of that effect to the entire design. The relation of the effect to the larger design is important because it is this relation that affects the size of the MS_e term computed in the ANOVA. The interested reader is referred to Jarmasz and Hollands (2009), where we have provided a more detailed treatment.

IM Factors in Mixed IM–RM Designs

Multiple IM factors do not have these multiplying effects on the size of the MS_e. In IM factorial designs, there is a single MS_e: the mean square within, or MS_W. In between/within, or mixed, IM–RM designs, the addition of RM factors will inflate the MS_W for any IM factor. The following formula should be used to compute CIs for an IM Factor A in a mixed design:

$$\mathrm{CI} = M_i \pm t_{\mathrm{critical}} \times \sqrt{\frac{MS_{\mathrm{S/A}}}{n_i \times L}}, \qquad (3)$$

where M_i is the mean for the relevant level of Factor A, n_i is the number of participants serving in each level of Factor A, and *L* is the product of the number of levels of all RM factors in the analysis.

Mixed IM × RM interactions. Following Estes (1997), Jarmasz and Hollands (2009) provided a detailed treatment on how to compute CIs for mixed IM × RM interactions. When comparing levels of the RM factor within an IM condition, one starts with the number of participants in each condition, n_i. The scaling procedures required by any RM factors not involved in the interaction can then be applied. This leads to the following formula for the CIs for this type of interaction:

$$\mathrm{CI} = M_i \pm t_{\mathrm{critical}} \times \sqrt{\frac{MS_{\mathrm{R\times S}}}{n_i \times \frac{L}{r}}}. \qquad (4)$$

Note the use of n_i (as opposed to N); note also that $MS_{\mathrm{R\times S}}$ and r still refer to the MS_{e} and the number of levels of the related RM effect, respectively.

When levels of the IM factor are compared within an RM condition, CIs should be based on a variance estimator called the *pooled mean square within cells*, a weighted average of the MS_{W} for the IM factor and the MS_{e} of the RM factor ($MS_{\mathrm{R\times S}}$; see Estes, 1997). Thus,

$$MS_{\mathrm{WC}} = \frac{MS_{\mathrm{W}} + (r-1)MS_{\mathrm{R\times S}}}{r}, \qquad (5)$$

where MS_{WC} is the pooled mean square within cells. The degrees of freedom for MS_{WC} is also a weighted average:

$$df\left(MS_{\mathrm{WC}}\right) = \frac{df\left(MS_{\mathrm{W}}\right) + \left(r-1\right)df\left(MS_{\mathrm{R\times S}}\right)}{r}. \qquad (6)$$

MS_{WC} then replaces $MS_{\mathrm{R\times S}}$ in Equation 4. The value of t_{critical} is determined using $df(MS_{\mathrm{WC}})$.

An alternative approach has been developed by Masson and Loftus (2003). This approach plots effect sizes for main effects and interactions, each with associated CIs based on contrasts. Interested readers are encouraged to consider both approaches.

Caveats

The Masson and Loftus (2003) approach is one of several possible methods for treating CIs in RM designs. Blouin and Riopelle (2005) have described a mixed model methodology, which involves the use of the SAS MIXED procedure (SAS Institute, 1999). Tryon (2001) has also described an inferential CI approach that yields results equivalent to standard null hypothesis significance testing procedures (although he only considered two-condition experiments using *t* tests). In our view, there is a need to consolidate these various approaches. An integrative treatment is beyond the scope of this brief note, but we make a few summary points here.

First, it is important to distinguish between CIs used for estimating a parameter value and CIs used to draw inferences about differences between means. When one compares differences, there is a root-2 adjustment of the CI value. When one does this, it is easy to draw inferences about a pair of means: If the CIs overlap, the means are not different; if they do not overlap, they are different. The Blouin and Riopelle (2005) and Tryon (2001) approaches take this into account, but the Masson and Loftus (2003) approach does not. Instead, Masson and Loftus favor an approach based on patterns of means, rather than individual comparisons. They advocate that graphically displaying a set of means along with a representation of statistical error (CIs) helps the researcher identify how the means in a set relate to each other, which is more meaningful than making binary decisions about particular pairs of means within the set. We cannot resolve this debate here but note that the CIs proposed in Jarmasz and Hollands (2009) can simply be multiplied by $\sqrt{2}\ /\ 2$, which should yield results similar to those obtained with the approaches favored by Tryon or Blouin and Riopelle and allow direct comparison of pairs of means. In a sense, this is a graphical representation of the rule of thumb described earlier.

Second, when there are multiple means associated with an effect, the number of possible comparisons also increases, leading to alpha inflation and the familywise error problem. When multiple comparisons are desired, there are, of course, a number of possibilities, such as the Bonferroni correction, detailed in most experimental design texts (e.g., Kirk, 1982). Using a sequential technique, the Benjamini–Hochberg procedure has been recently shown to yield greater power than the Bonferroni approach (Thissen, Steinberg, & Kuang, 2002). Note that the Masson and Loftus (2003) approach to CIs has the advantage that it focuses on the pattern of means and does not advocate multiple comparisons, circumventing the need for them.

Third, the mixed model methodology of Blouin and Riopelle (2005) appears to use the same number of observation values as we propose, at least for the examples provided in their article. This implies that if one runs SAS MIXED on data from RM or IM–RM designs, one should obtain the correctly sized CIs scaled to our number-of-observations principle, although we have not tested this hypothesis empirically.

Fourth, the advantage of the Masson and Loftus (2003) approach is that it is simple to apply when one has balanced data without violations of sphericity and that it maps in a more straightforward manner to well-known ANOVA procedures. The advantage of the Blouin and Riopelle (2005) mixed model approach, as we see it, is that the statistical model takes into account factors such as unequal sample sizes or inherent covariance relations among repeated measures. However, the details of the mixed model approach are complex and beyond the scope of this brief note. Readers are referred to Blouin and Riopelle for a thorough treatment.

Finally, we note that the CI, as discussed in this article, is not a *credible interval* as the term is used in Bayesian statistics. It should not be interpreted as such. The credible interval incorporates context from the prior distribution, whereas CIs are based on the data only (Lee, 1997).

Conclusions

Applying the number-of-observations principle involves determining how many participants contribute to the effect (total number of participants for pure RM effects, number of participants per condition for IM effects and IM × RM interactions). Then, number of participants is multiplied by the product of the levels of the remaining RM factors in the analysis. The appropriate MS_{e} must also be selected.

In a reference table, Jarmasz and Hollands (2009, Table 4) summarized the key values used to compute CIs for each possible effect in each type of factorial design involving RM factors. We hope that this will provide researchers with a simple reference tool that they can consult when constructing CIs.

Our aim in this effort was to provide psychological researchers with a principle to compute CIs in factorial designs involving RM effects. We hope that this makes it easier for researchers to plot the CIs associated with any particular effect in any design involving RM factors.

AUTHOR NOTE

We thank the reviewers of this article and other related manuscripts for critical comments that shaped some of the ideas expressed herein. Correspondence concerning this article should be addressed to J. G. Hollands, Defence Research and Development Canada–Toronto, 1133 Sheppard Avenue West, P.O. Box 2000, Toronto, ON, M3M 3B9 Canada (e-mail: justin.hollands@drdc-rddc.gc.ca).

REFERENCES

Blouin, D. C., & Riopelle, A. J. (2005). On confidence intervals for within-subjects designs. *Psychological Methods*, **10**, 397-412.

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

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

Jarmasz, J., & Hollands, J. G. (2009). Confidence intervals in repeated measures designs: The number of observations principle. *Canadian Journal of Experimental Psychology*, **63**, 124-138.

Kirk, R. E. (1982). *Experimental design: Procedures for the behavioral sciences* (2nd ed.). Monterey, CA: Brooks/Cole.

Lee, P. M. (1997). *Bayesian statistics: An introduction* (2nd ed.). London: Arnold.

Loftus, G. R., & Loftus, E. F. (1988). *Essence of statistics* (2nd ed.). New York: Knopf.

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

Masson, M. E. J. (2004). Correction to Masson and Loftus (2003). *Canadian Journal of Experimental Psychology*, **58**, 289.

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

SAS Institute (1999). *SAS/STAT user's guide* (Version 8, Vol. 2). Cary, NC: SAS Institute.

Thissen, D., Steinberg, L., & Kuang, D. (2002). Quick and easy implementation of the Benjamini–Hochberg procedure for controlling the false positive rate in multiple comparisons. *Journal of Educational & Behavioral Statistics*, **27**, 77-83.

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.

(Manuscript received November 3, 2008;
revision accepted for publication September 19, 2009.)