Tutorials in Quantitative Methods for Psychology
2005, Vol. 1(1), p. 42-45.

Confidence intervals in within-subject designs: A simpler solution to Loftus and Masson's method

Denis Cousineau
Université de Montréal

Within-subject ANOVAs are a powerful tool to analyze data because the variance associated to differences between the participants is removed from the analysis. Hence, small differences, when present for most of the participants, can be significant even when the participants are very different from one another. Yet, graphs showing standard error or confidence interval bars are misleading since these bars include the between-subject variability. Loftus and Masson (1994) noticed this fact and proposed an alternate method to compute the error bars. However, i) their approach requires that the ANOVA be performed first, which is paradoxical since a graph is an aid to decide whether to perform analyses or not; ii) their method provides a single error bar for all the conditions, masking information such as the heterogeneity of variances across conditions; iii) the method proposed is difficult to implement in commonly-used graphing software. Here we propose a simple alternative and sow how it can be implemented in SPSS.

Consider the results shown in Figure 1 where mean results from a 2 × 5 experiment are shown. The error bars show the standard error in each condition, measured on 16 participants per point. If confidence intervals had been shown, the error bars would have been about twice their actual sizes! By looking at this figure, we have no doubt that it is only noise. Yet, have a look at the ANOVA table: the effects and the interaction are all highly significant! How can this be?

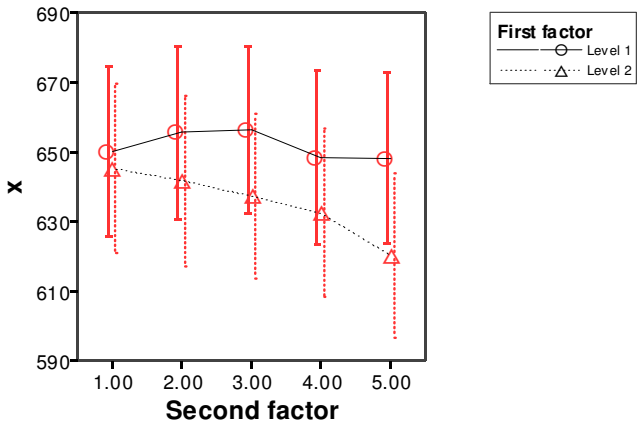
The present data are simulated. However, we obtained similar results in Paradis and Cousineau (in preparation). This kind of situation was first noted by Loftus and Masson (1994).

The cause of the discrepancy between the figure and the analyses is not obvious. It is not a problem with homogeneity of variances (all the variances are homogeneous and spherical, Tabachnik & Fidell, 1996, Mauchly's $W = 0.74$, $p > .50$ for factor 2 and $W = 0.68$, $p > .50$ for the interaction; this test cannot be performed for factor 1 since it has only two levels but we generated the data such that it is also homogeneous). The Greenhouse-Geiser and the Huynh-Feldt epsilons are close to 1 so that we don't need to use corrections (Huynh, 1978, Rouanet and Lepine, 1970). Using multivariate tests (such as Hotelling's T or Wilks λ) does not change the results in any way.

The inconsistency between the graph and the analyses comes from the fact that we are using a repeated-measure design. All the participants are measured in each of the 10 combinations of the factor 1 and factor 2 levels. Hence, it is possible to assess whether a given participant systematically scores high or systematically scores low. In fact, that is what happens in the present case. Figure 2 shows the results for each individual participant. As seen, there is a tremendous amount of difference between each participant. Hence, we can safely conclude that the participants differ significantly (in fact, this information is indeed provided by most statistical software, $F(1, 15) = 710$, $p < .001$). However, in general, we don't care about this: in psychology, it is a plain fact that most humans differ. What we really want to know is if the factors influence the results. By looking carefully at the second condition of the factor 1 (the right panel of Figure 2), we see that for most of the participants, scores decrease when going from the first level of factor 2 to the fifth.

Correspondance for this article sould be sent to Denis Cousineau, Psychologie, Université de C.P. 6128, succ. Centre-ville, Montréal, QC, CANADA, H2T 2H5 or using email at denis.cousineau@umontreal.ca. The author wishes to thank David Paradis and Dominic Charbonneau for challenging results which initiated this research. Support for this research was provided by the Conseil pour la recherche en sciences naturelles et en génie du Canada.

Figure 1: Fictitious results of an experiment with two factors, the first with two levels and the second with 5 levels. Error bars show the mean ± 1 standard error.



Therefore, if we could ignore the relative position of the participants, the trend would be very clear.

Figure 3 shows exactly that. Each participant's scores were adjusted so that its relative position is no longer present. As seen on the right panel, the downward trend is very clear whereas in the other condition on the left, there is no visible trend. Hence, we should observe an interaction.

The repeated-measure ANOVA got those results right because it first starts computing the between-subject sum of square (SS_S, Keppel, 1973) and removes it from the total sum of square before partitioning the remaining sum of square in the usual manner (main effects, interaction, and error terms). Hence, a great deal of variability is removed, giving more chance to the F ration to exceed the critical value.

Because the mean square of error $MS_e = \sqrt{(MS_e / dl_e)}$ is a direct measure of the unexplained variation, Loftus and Masson (1994) suggested to use this term for the computation of the error bar. Indeed, the ANOVA test uses the quantity $\sqrt{(MS_e / dl_e)}$so that a confidence interval is $t_{\alpha, dle} \times \sqrt{(MS_e / dl_e)}$ where a is the confidence level (often 5%) and t is obtained in a Student table with dl_e degrees of freedom. By extension with other tests, we can equate the standard error to the $\sqrt{(MS_e / dl_e)}$ term as it is the part that does not depend on a confidence level.

Although Loftus and Masson solution is sound in providing standard errors and confidence intervals exempted of between-subject differences, it has three limitations:

First, the analyses must be performed first, in order to get the $\sqrt{(MS_e / dl_e)}$ term that is used as the error bars in the graphs. This is paradoxical since graphs should precede any analysis, as their purpose is to help anticipate the results of the analyses. In addition, in factorial designs, there are more than 1 error terms so which one to use is ambiguous. For instance, if you expect a main effect and no interaction, use the error term associated with that effect whereas if you expect an interaction, use the interaction error term…

Second, Loftus and Masson's method provides the size for a unique error bar that will be applied to all the points. The (omnibus) ANOVA do use a single error term per effect, and in this respect, the graph is congruent with the analysis. However, we may want to look at other information on the graph. For instance, are the variances homogeneous across levels of the factors? By using a single error bar, this information is lost.

Thirdly, most plotting software either computes error bars automatically (but then, they wrongly include between-subject differences) or they must be provided manually. This last technique takes a lot more time, requiring multiple steps and manual interventions.

In the following, we present an alternative technique that solves all three limitations. Further, it can be implemented easily in most statistical software. We show how using SPSS 13.

Consider the data from three participants presented in Table 2. As seen by the marginal means, there seems to be an effect of the manipulation. However, there are also large differences between the participants. For instance, the first one is on average 55 ms faster than the mean of the group. Hence, if we add 55 ms to all the obtained performance, we would "erase" the particularities of that participant. The participant mean (noted $\overline{X}_1$) minus the group mean (noted

Table 1. Results of a 2 × 5 experiment with the first factor having two levels and the second factor having 5 levels

Effect name	SS	dl	MS	F	
Factor 1	10621	1	10621	76.8	***
Error	2073	15	135		
Factor 2	11784	4	8196	16.4	***
Error	4378	60	72.9		
Interaction	2250	4	562	6.52	***
Error	5171	60	86.2		

***: $p < .001$

Figure 2: The individual results of the 16 simulated participants of Figure 1. Left panel is for the first level of the first factor.

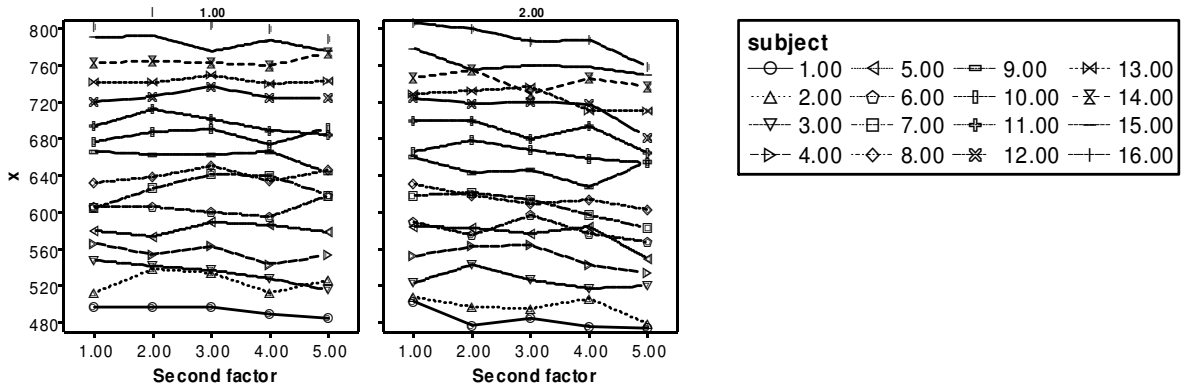


Table 2. Hypothetical results from a repeated-measure experiment with one factor having three levels.

	Condition			
Participant	1	2	3	*Mean*
1	550	580	610	*580*
2	605	635	655	*635*
3	660	690	710	*690*
Mean	*605*	*635*	*655*	*635*

$\overline{X}$) indicates the amount of correction to apply to the performance obtained in the *i*th condition by the *j*th participant (noted X_{ij}). Overall, if we create a new variable Y and let

$$Y = X_{ij} - \overline{X}_1 + \overline{X} \quad (1)$$

for all conditions *i* and all participants *j*, all the individual differences will be erased. Figure 3 was made using *Y*s instead of *X*s. Further, a graph of Y as a function of the conditions can be made showing means and error bars automatically. Figure 4 shows the results from the fictitious experiment. It is now evident from inspection of the graph that an interaction is present.

In SPSS 13, computing the mean for each participant is performed using the following syntax:

```
Aggregate outfile=* mode=addvariables
        /break = subject
        /x.subj = mean(x).
```

where **x** is the name of the column containing the dependant variable and **subject** is the name of the column containing subject identification.

Figure 3: The individual results of the 16 simulated participants of Figure 1 after the individual differences were removed.

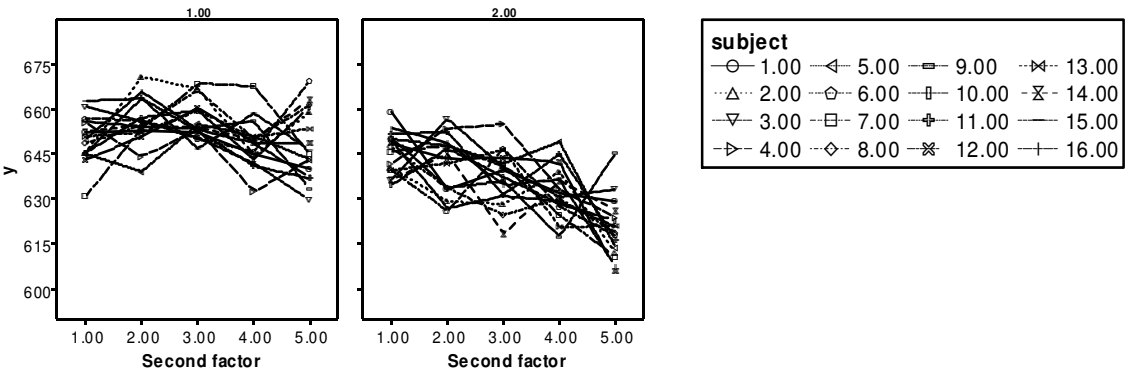
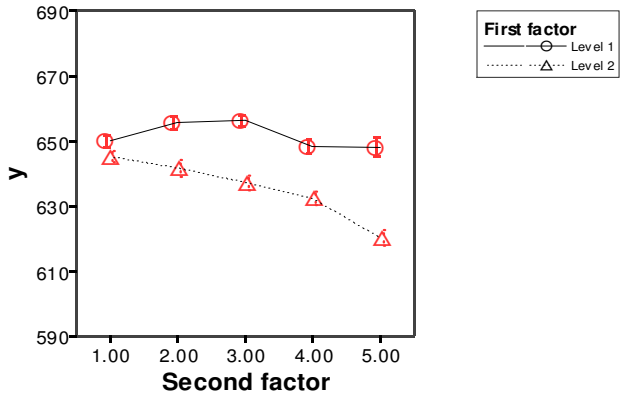


Figure 4: Same as in Figure 1 except that the error bars does not include variability associated with between-subject differences.



To compute the overall mean, you need to segment the data by groups. In a full within-subject design, there is no group, so you need to assign all the participants to the same group, e. g. group 1. This is done with:

Compute group = 1.

Then, you can compute the mean for each group with:

Aggregate outfile=* mode=addvariables
/break = group
/x.group = mean(x).

Finally, computing Y as in Eq. 1 is done with:

Compute y = x – x.subj + x.group.

That is all you need. You can now realize you graphs on *Y* to have the correct error bars. *Y* is only useful for graphing purposes; for the analyses, continue to use the original data contained in the column *X*.

You will find on the journal's web site the data used in the Figures along with a complete syntax file for SPSS 13.0 that computes Y, makes the graphs, and after reorganizing the data so that they are in distinct columns, performs the ANOVA.

References

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

Keppel, G. (1973). *Design and analysis: A researcher's handbook*. Englewood Cliffs, NJ: Prentice-Hall, inc.

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

Rouanet, H., Lépine, D. (1970). Comparison between treatments in a repeated-measurement design: Anova and multivariate methods. *British Journal of Mathematical and Statistical Psychology*, *23*, 147-163.

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

Received July 16, 2005.
Accepted August 16, 2005