




Correlation-adjusted standard errors and confidence intervals for within-subject designs: A simple multiplicative approach

Denis Cousineau^a 

^aUniversité d'Ottawa

Abstract ■ In within-subject designs, the multiple scores of a given participant are correlated. This correlation implies that the observed variance can be partitioned into between-subject variance and between-measure variance. The basic confidence interval about the mean does not separate these two sources and is therefore of little use in within-subject designs. Two solutions have been proposed, one (Loftus and Masson) requires the computation of the interaction terms including the subject and all within-subject factors, the other (Cousineau and Morey) requires a two-step transformation of the data. As shown, these two methods are nearly equivalent. Herein, I present a correlation-adjusted method which requires the mean correlation across all pairs of measurements. This solution is shown to be similar to the other two for data satisfying the compound symmetry assumption. It is found to be too liberal for data having homogeneous correlations and heterogeneous variances but a Welch correction for heterogeneous variances can be used. Finally, it is inadequate for data that do not satisfy the compound symmetry assumption but satisfy the sphericity assumption. A statistical test of compound symmetry is discussed.

Keywords ■ Error bars, confidence intervals, correlations.

 denis.cousineau@uottawa.ca

 DC: 0000-0001-5908-0402

 10.20982/tqmp.15.3.p226

Acting Editor ■
Roland Pfister (Uni-
versität Würzburg)

Reviewers
■ Daniel W. Heck
(Philipps-Universität
Marburg)

■ Thom Baguley (Not-
tingham Trent Uni-
versity)

Introduction

Getting confidence intervals about means in within-subject designs has been a recurrent concern over the last decades. The difficulty comes from the fact that the basic confidence interval is meant to assess a plausible range of values for a mean, irrespective of other means in the sample. However, as argued in Baguley (2012), researchers in psychology are rarely interested in a single mean in isolation. Instead, they are more often interested in the relative positions between means. This change in perspective should be mirrored by changes in the ways that error bars are estimated (also see Cousineau, 2017; Pfister & Janczyk, 2013).

A first change, spearheaded by Goldstein and Healy

(1995), Baguley (2012) and Franz and Loftus (2012), is to use difference-adjusted intervals. When examining a mean relative to another mean, both have uncertainty (measured by within-condition variability). However, their relative positions also bring uncertainty. Assuming that the variances are homogeneous, it is sufficient to increase the length of the error bars by a factor of $\sqrt{2}$ to take into account this additional source of uncertainty. Herein, I will apply this correction factor to all the intervals examined, resulting in what is called "difference-adjusted intervals".¹ Difference-adjusted intervals are meaningful only for pairwise comparisons. For comparisons to a pre-specified value, use the unadjusted CI.

¹Note that Goldstein and Healy (1995) also suggested dividing the intervals by 2 so that it is the overlap between two CI that indicates an absence of difference between means. This is called a half-width interval. In Franz and Loftus (2012), both are used simultaneously with a correction factor of $\sqrt{2}/2 = 1/\sqrt{2}$, complicating the presentation.



Table 1 ■ Example data set from Loftus and Masson (1994). The average pair-wise correlation is .983.

Descriptive statistics	Conditions		
	1-sec	2-sec	5-sec
	10.	13.	13.
	6.	8.	8.
	11.	14.	14.
	22.	23.	25.
	16.	18.	20.
	15.	17.	17.
	1.	1.	4.
	12.	15.	17.
	9.	12.	12.
	8.	9.	12.
Sample size N	10		
Mean M_i	11.0	13.0	14.2
Standard deviation S_i	5.793	6.074	5.959
Correlations	1-sec	1	.985
	2-sec		1
	5-sec		

A second change which is necessary to assess relative positions of means is to take into account the experimental design. It is generally accepted that within-subject designs are more powerful than between-subject designs to assess differences in means. Consequently, this additional precision should be reflected in shorter intervals. The source of this additional statistical power is to be found in the correlation between the variables considered. Hence, I introduce a new proposal that I call a correlation-adjusted method to compute standard errors and confidence intervals.

In what follows, I first present the correlation-adjusted method. The advantage of the present proposal is that it only requires the computation of the average correlation between variables. Because correlations are useful statistics in repeated-measures designs, for example to assess statistical power (Goulet & Cousineau, 2019) or Cohen’s d (Goulet-Pelletier & Cousineau, 2018), the correlation matrix will often be assembled as part of routine data analysis. I then briefly overview two existing methods and show under what conditions all three solutions of computing difference-adjusted CIs are equivalent. As will be seen, all three methods are equivalent under compound symmetry, a technical term describing the structure of the covariance matrix whereby a common correlation between pairs of measures is assumed and a unique variance for all measures is assumed as well. It implies equality of the diagonal and equality of the off-diagonals in the covariance

matrix. In the last section, I perform systematic comparisons across methods under various covariance structures. The results confirm that the easier correlation-adjusted method returns reliable confidence intervals (CI) mainly under compound symmetry.

The correlation-adjusted method

The correlation-adjusted (CA) confidence interval is given by

$$M_i \pm \frac{S_i}{\sqrt{N_i}} \times \sqrt{1 - \bar{r}} \times t_{N_i-1} \quad (CA;1)$$

in which M_i is the mean of the measures in the i th condition ($i = 1, \dots, C$), S_i is the standard deviation of those measures, N_i is the number of measures, \bar{r} is the average correlation across all the pairs of measures from the C repeated measures, and finally, t_{N_i-1} (or in full, $t_{N_i-1}(1 - (1 - \gamma)/2)$) is the coverage factor in which γ —often 95%—is the proportion of coverage desired based on $N_i - 1$ degrees of freedom.

The second term in Eq. 1, $S_i/\sqrt{N_i} \times \sqrt{1 - \bar{r}} \times t_{N_i-1}$, is sometimes called the interval half-width. This is the same formula as the basic CI (see below) except for a multiplicative adjustment based on the correlation, $\sqrt{1 - \bar{r}}$. The term $S_i/\sqrt{N_i}$ is the standard error of the i th mean; by extension, $S_i/\sqrt{N_i} \times \sqrt{1 - \bar{r}}$ will be called the correlation-adjusted standard error of the mean. For a difference-adjusted CI, multiply the CI width by $\sqrt{2}$ (i.e., increase its length by 41%). All within-subject CI should also be difference-adjusted CI because it is contradictory to use information from the other measurements if the purpose is not to compare them.

As an example, consider the data in Table 1 taken from Loftus and Masson (1994); also see Nathoo, Kilshaw, and Masson (2018), Heck (2019). From the correlation matrix, we find that the average correlation, \bar{r} , is .983.²Based on $10 - 1 = 9$ degrees of freedom, the 95% coverage factor is $t_9(.975) = 2.262$. Thus, with $S_1 = 5.793$ (see Table 1), the standard error for Measure 1 is $5.793/\sqrt{10} \times \sqrt{1 - .983} = 0.24$ and its correlation-adjusted confidence interval width is $5.793/\sqrt{10} \times \sqrt{1 - .983} \times 2.262 = 0.54$. Multiplied by $\sqrt{2}$, we get the width of the correlation and difference adjusted CI: $0.54 \times 1.414 = 0.76$. The correlation and difference adjusted CI for Measure 1 is therefore

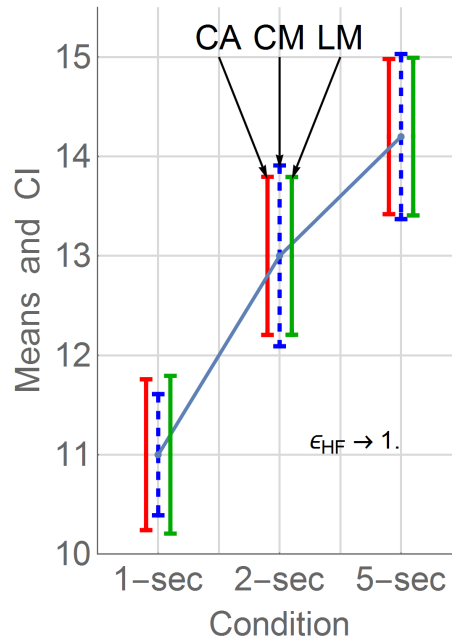
$$11.0 \pm 0.54 \times \sqrt{2} = [11.0 - 0.76, 11.0 + 0.76] \\ = [10.24, 11.76]$$

The CI widths for all three measures including the difference adjustment ($\sqrt{2}$) are given in Table 2. The average correlation can be obtained in R with the following instruc-

²We used the arithmetic mean. The assumption of homogeneous correlations is mute as to what method could be used to average the correlations. One reader suggested using the Fisher’s r -to- z transform before performing average, which returns a mean r of 0.984.



Figure 1 ■ Plot of the mean results for Table 1 data along with difference-adjusted 95% confidence intervals based on three methods. In red (left-most error bars): the Correlation-adjusted method (CA); in blue (central error bars): the Cousineau-Morey method (CM); in green (leftmost error bars): the Loftus and Masson method (LM). The Huyndt-Felt epsilon is 1.0 indicating spherical data; the Welch factor is 0.999, indicating homogeneous variances; the data do not reject compound symmetry (Winer’s test $M = 2.55, \chi^2(4) = 2.12, p = 0.713$ and do not reject sphericity (Mauchly’s test $W = 0.816, \chi^2(2) = 1.622, p = 0.444$). The basic error bars are not shown; they are roughly ten times longer.



tions, assuming that X is a data frame containing only the repeated measures:

```
r <- cor(X)
rbar <- mean(r[upper.tri(r)])
```

Figure 1 shows the error bars including the difference adjustment.

Comparison to previous methods

The basic CI is based on the sample standard deviation and the sample size. The confidence interval is given by

$$M \pm \frac{S}{\sqrt{N}} \times t_{N-1} \quad (\text{Basic; } 2)$$

This confidence interval can be derived from the two-sided, one sample t-test, as its rule

$$\text{Reject } H_0 \text{ if } \frac{M - \mu_0}{S/\sqrt{N}} > t_{N-1}$$

can be reformulated into

$$\text{Do not reject } H_0 \text{ if } M - \frac{S}{\sqrt{N}} \times t_{N-1} < \mu_0 < M + \frac{S}{\sqrt{N}} \times t_{N-1}$$

where μ_0 is a hypothesized population mean under H_0 . The examination of within-subject design error bars and confidence intervals was initiated by Loftus and Masson’s seminal paper (LM; 1994). After examining many possible solutions, they recommended the use of a standard error based on the subject \times within-subject conditions interaction term,

$$S_{LM} = \sqrt{\frac{SS_{S \times C}}{(C - 1)(N - 1)}}$$

The quantity S_{LM}^2 is also noted $MS_{S \times C}$ in Loftus and Masson (1994), their Eq. 2. The sum of square, taken from a within-subject design analysis of variance (ANOVA), is given by

$$SS_{S \times C} = \sum_{i=1}^C \sum_{j=1}^N (\mathbf{X}_{ij} - \bar{\mathbf{X}}_{.j} - \bar{\mathbf{X}}_{i.} + \bar{\bar{\mathbf{X}}})^2.$$

where $\bar{\bar{\mathbf{X}}}$ is the grand mean. As seen, double-centering is used (Abdi, 2010), that is, the scores are centered relative to the subject mean ($\bar{\mathbf{X}}_{.j}$) and relative to the condition



Table 2 ■ Standard errors and difference-adjusted 95% Confidence intervals widths based on three methods. Basic: the basic method; CA: Correction-adjusted method; CM: Cousineau and Morey method; LM: Loftus and Masson method. The method LM returns a unique width for all conditions because it is based on a pooled standard error.

	Equation	SE			CI width		
		1-sec	2-sec	5-sec	1-sec	2-sec	5-sec
Basic	(2)	1.83	1.92	1.88	5.860	6.144	6.029
CA	(1)	0.237	0.249	0.244	0.759	0.795	0.780
CM	(3)	0.191	0.284	0.260	0.609	0.909	0.831
LM	(5)	0.248			0.737		

mean (\bar{X}_i). The Loftus and Masson confidence interval for within-subject design is thus

$$M \pm \frac{S_{LM}}{\sqrt{N}} \times t_{(C-1)(N-1)} \quad (\text{LM}; 3)$$

Cousineau (2005), complemented by Morey (2008), proposed a different approach (named CM in Baguley, 2012) whereby the data are first transformed and then basic confidence intervals are obtained from the transformed dataset. It is a two-step approach involving, first, a subject-centering transformation and second, a rescaling:

$$Y_{ij} = X_{ij} - \bar{X}_j + \bar{X} \quad (4a)$$

$$Z_{ij} = \sqrt{\frac{C}{C-1}} (Y_{ij} - \bar{Y}_{.j}) + \bar{Y}_{.j} \quad (4b)$$

(Cousineau & O'Brien, 2014).³ Finally, the confidence interval is obtained with the basic method

$$M \pm \frac{S_Z}{\sqrt{N}} \times t_{N-1}. \quad (\text{CM}; 5)$$

The reason for this two-step approach is pragmatic: any graphing software can draw within-subject error bars from **Z** if it can draw the basic error bars from **X** (as is the case for most statistics software with graphing capabilities).

The second step is meant to increase error variance by $C/(C-1)$ because as shown in Morey (2008), the variability estimated from **Y** is biased downward. The method is thus equivalent to

$$M \pm \sqrt{\frac{C}{C-1}} \times \frac{S_Y}{\sqrt{N}} \times t_{N-1} \quad (5')$$

³Confusions abound as to how the transformation of Eq. (4a) should be called. Loftus and Masson (1994), Bakeman and McArthur (1996), Morey (2008) and others call it a "normalization". However the transformation does not make the data more normally distributed. "Standardization" is found in O'Brien and Cousineau (2014). I now think that "Subject-centering transformation" is probably the most accurate label.

⁴This method, noted NKM herein, is given by $S_{NKM} = \sqrt{SS_{S \times C} / (C(N-1))}$ and $M \pm S_{NKM} / \sqrt{N} \times t_{C(N-1)}$. Heck (2019) recommends the use of subsampling with Monte Carlo Markov Chains —MCMC— to counteract the bias. The source of the problem is that it is derived from a maximum likelihood estimate of variance, which is known to be biased, underestimating the population variance. Countering the bias is achieved with subsampling or with the correction factor $C/(C-1)$ (Morey, 2008). As observed by Heck (2019), p. 29, Nathoo et al.'s (2018) method is no longer different from LM method once a proper unbiased approach is used (subsampling or correction factor). Note that the Nathoo et al.'s method also has the largest degrees of freedom for the coverage factor.

It can be shown that both LM and CM methods are nearly identical except for two variations: (i) LM uses a pooled standard deviation whereas CM uses distinct standard deviations for each condition (see Appendix A for a demonstration); (ii) LM confidence intervals use larger degrees of freedom $(C-1)(N-1)$ whereas CM uses $N-1$. Consequently, when comparing CI width, the coverage factor is not the same in both methods and

$$\frac{t_{(C-1)(N-1)}}{t_{N-1}} \quad (6)$$

is the difference between the widths of the confidence intervals of each method. As an example, this ratio is 0.928 for 95% coverage of 3 measures with 10 subjects, as in Table 1. Thus, LM CI is roughly 7% shorter than CM CI for such a small sample size. For larger sample sizes, there is no sizeable difference.

Finally, Nathoo et al. (2018) proposed a new interval derived from Bayesian arguments. However, Heck (2019) observed that this interval suffers from bias. Thus, this method will not be discussed further.⁴

All the formulas are summarized in Table 3. When comparing the methods, I do not consider the differences in degrees of freedom (e. g., Eq. 6) as they introduce fairly negligible differences only.

In what follows, I compare the methods when various assumptions are not met (the sphericity assumption or the homogeneity of variances plus the homogeneity of correlations assumptions jointly called the compound symmetry assumption). Along the way, I demonstrate that under compound symmetry, CA is an unbiased estimator of the same CI than CM and LM. But first, I briefly consider the paired-sample design (only two repeated measures); it



Table 3 ■ Overview of the confidence intervals in within-subject designs

Method	Eq.	Formula	in which
Basic	(2)	$M_i \pm \frac{S_i}{\sqrt{N}} \times t_{N-1}$	$S_i^2 = \frac{1}{N-1} \sum_{j=1}^N (\mathbf{X}_{ij} - \bar{\mathbf{X}}_i)^2$
CA	(1)	$M_i \pm \frac{S_i}{\sqrt{N}} \times \sqrt{1 - \bar{r}} \times t_{N-1}$	
CM	(3)	$M_i \pm \frac{S_{\mathbf{Z},i}}{\sqrt{N}} \times t_{N-1}$	$S_{\mathbf{Z},i}^2 = \frac{C}{C-1} S_{\mathbf{Y},i}^2 = \frac{C}{C-1} \frac{1}{N-1} \sum_{j=1}^N ((\mathbf{X}_{ij} - \bar{\mathbf{X}}_j + \bar{\bar{\mathbf{X}}}) - \bar{\mathbf{X}}_i)^2$
LM	(5)	$M_i \pm \sqrt{\frac{SS_{S \times C}}{(C-1)(N-1)}} \times \frac{1}{\sqrt{N}} \times t_{(C-1)(N-1)}$	$\frac{SS_{S \times C}}{(C-1)(N-1)} = S_{\mathbf{Z},pooled}^2 = \frac{1}{C-1} \sum_{i=1}^C S_{\mathbf{Y},i}^2$

Note. All these intervals should be difference-adjusted, i.e., the length increased by $\sqrt{2}$ but this adjustment is not shown in the table. The symbol \bar{r} is the mean pairwise correlation; $SS_{S \times C}$ is the interaction sum of square; C is the number of repeated measures and N is the number of participants. The mean in the *i*th condition is noted $\bar{\mathbf{X}}_i$, or M_i and the grand mean is noted $\bar{\bar{\mathbf{X}}}$ whereas the standard deviation in the *i*th condition is noted S_i .

will set the stage for useful concepts. Because one method uses a pooled standard error and the others, distinct standard errors, equivalence between the two methods is said to be in the root mean squared sense: If one takes the standard errors from CM, squares them and sums them, it will return the same total as if the LM standard errors are squared and summed. In validating the methods, the prime requisite for a valid CI is that over multiple replications, at least $\gamma \times 100\%$ of the intervals include the true population parameter. Its complement, $1 - \gamma$, is akin to type-I error rate if we use CI to reject or not the true situation.

Correlation-adjusted method in the paired-sample design

In the paired-sample design, there are only two repeated measures. The CI interval is then taken directly from the paired-sample *t*-test. This test is often seen as

$$\text{Reject } H_0 \text{ if } \frac{|M_1 - M_2|}{S_D / \sqrt{N}} > t_{N-1} \tag{8a}$$

where S_D is the standard deviation of the differences between the pairs of scores (Pfister & Janczyk, 2013). What is less known is that it can also be expressed as

$$\text{Reject } H_0 \text{ if } \frac{|M_1 - M_2|}{\sqrt{2} S_{pool} \sqrt{1 - r} / \sqrt{N}} > t_{N-1} \tag{8b}$$

where $S_{pool} = \sqrt{(S_1^2 + S_2^2)/2}$ comes from averaging the variances (the squared standard deviations). Consequently, the paired-sample *t*-test is identical to the two-independent sample *t*-test except for a correction factor $\sqrt{1 - r}$ (Cousineau, 2010; Afshartous & Preston, 2010).

Demonstrating the equivalence between the two versions of the test (Eqs. 8a and 8b) requires the homogeneity of variances assumption. If the two measures' variances σ_1^2 and σ_2^2 are equal, say, to σ , then the well-known relation $\sigma_D^2 = \sigma_1^2 + \sigma_2^2 - 2\sigma_{12}$ becomes $\sigma_D^2 = \sigma^2 + \sigma^2 - 2\rho_{12}\sigma \times \sigma = 2\sigma^2(1 - \rho_{12})$ in which σ_{12} is the covariance, σ is the common standard deviation and ρ_{12} is the correlation between the two measures. Using S_{pool} to estimate σ and Pearson's *r* to estimate ρ_{12} from the observed data, we get that $S_D = \sqrt{2} S_{pool} \sqrt{1 - r}$. Reorganizing the rule, we find

$$\text{Do not reject } H_0 \text{ if } M_2 - \frac{\sqrt{2} S_{pool} \sqrt{1 - r}}{\sqrt{N}} t_{N-1} < M_1 < M_2 + \frac{\sqrt{2} S_{pool} \sqrt{1 - r}}{\sqrt{N}} t_{N-1} \tag{9a}$$



or equivalently

$$\text{Do not reject } H_0 \text{ if } M_1 - \frac{\sqrt{2} S_{pool}\sqrt{1-r}}{\sqrt{N}} t_{N-1} < M_2 < M_1 + \frac{\sqrt{2} S_{pool}\sqrt{1-r}}{\sqrt{N}} t_{N-1}. \quad (9b)$$

Two observations are noteworthy. First, the difference-adjustment correction $\sqrt{2}$ appears naturally, which shows its logical necessity. Second, it illustrates well the golden rule of confidence intervals: "If one mean is included within the confidence interval of the other mean, the two can be considered comparable" (Cousineau, 2017). This rule is valid whichever mean M_1 or M_2 is used to compare to a CI.

The equivalence between the correlation-adjusted test (Eq. 8b) and the test of the differences (Eq. 8a) is critically based on the homoscedasticity assumption (i. e., the assumption of equal variances). When variances are heterogeneous, a correction to the degrees of freedom has been proposed for the two independent-sample design by Welch (Welch, 1938; Derrick, Toher, & White, 2016). It involves substituting the degree of freedom ν_{homo} with ν_{hetero} where

$$\nu_{\text{homo}} = N_1 + N_2 - 2 \quad (10a)$$

$$\nu_{\text{hetero}} = \frac{(S_1^2/N_1 + S_2^2/N_2)^2}{\left(\frac{S_1^2}{N_1}\right)^2/(N_1 - 1) + \left(\frac{S_2^2}{N_2}\right)^2/(N_2 - 1)}. \quad (10b)$$

When samples are of the same size ($N_1 = N_2 = N$), as is the case for paired samples, the above expressions can be simplified greatly into

$$\nu_{\text{homo}} = 2(N - 1) \quad (10a')$$

$$\nu_{\text{hetero}} = (1 + W_f)(N - 1) \quad (10b')$$

where W_f , that I call the Welch factor, is given by

$$W_f = \frac{\widetilde{V}^2}{\widehat{V}^2}$$

in which \widetilde{V}^2 is the harmonic mean of the squared variances observed in the two samples and \widehat{V}^2 is the geometric mean of the same squared variances. This factor ranges from 0 (very different variances) to 1 (identical variances). The Welch factor is a measure of the discrepancy between the variances; it is applicable to more than two samples when sample sizes are all equal (although informal simulations suggests that N in Eq. 10b' can be replaced with

\widetilde{N}). The Welch correction preserves the type-I error rate but is known to result in a less powerful test of means.

Returning to the correlation-adjusted confidence intervals (Eq. 1), an alternate proposal could be to replace the coverage factor t_{N-1} by $t_{(N-1) \times W_f}$ when variances are unequal (according to a Levene test, for example, but see Rochon, Gondan & Keiser, 2012). The value $(N - 1) \times W_f$ is always smaller or equal to $N - 1$.

Comparing the methods

To assess the various methods, it is necessary to consider the various structures that the covariance matrix can take. The simplest structure is called compound symmetry (e.g., Winer, Brown, & Michels, 1991). Under this structure, variances are homoscedastic (they are homogeneous for all the variables measured) and correlations are homosociotic (they are homogeneous for all pairs of variables measured).⁵ This structure is but one possible structure; Figure 2 proposes a general classification.

The sphericity assumption in ANOVAs (which stipulates that the variance of the differences between all the pairs of variables is constant) is true when compound symmetry is true; it can also be true if a very specific pattern of variance/covariance is found under heteroscedasticity and heterosocioticity. There exists a well-known test of sphericity (Mauchly, 1940). Less known, there also exists a test of compound symmetry (Winer et al., 1991) described in Appendix B.

Comparing the methods under compound-symmetric covariances

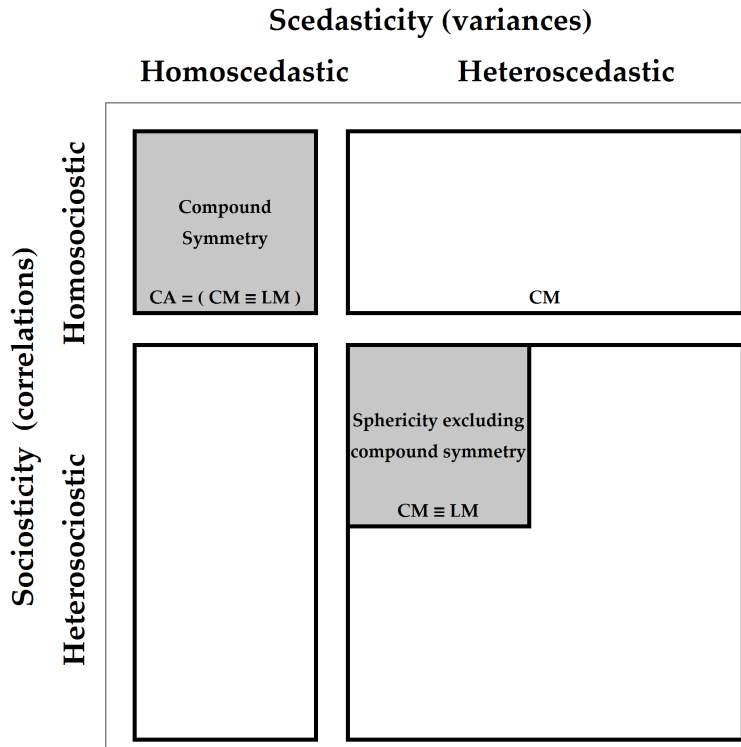
Under compound symmetry, the variance-covariance matrix contains a common variance along the main diagonal (say σ^2), and a common covariance outside of the main diagonal, obtained from a common correlation (say, ρ , so that the covariances are all equal to $\rho \times \sigma^2$). The best estimate of σ is the pooled standard deviation S_{pool} whereas an estimate of ρ is the average observed pairwise correlation,

$$\bar{r} = \frac{1}{C(C-1)/2} \sum_{i=1}^C \sum_{j>i}^C r_{ij}$$

⁵I forged this neologism from the Latin root of association, *ad socius*, literally *together linked*.



Figure 2 ■ Figure 2. Structure of the variance/covariance matrices, divided in four families based on scedasticity (homogeneous or heterogeneous) and sociosticity (homogeneous or heterogeneous). In the lower part of the boxes are methods that are adequate. The grey area represents sphericity. The symbol \equiv denotes identical methods in the root mean squared sense and $=$ denotes methods that are equivalent in distribution.



in which C is the number of measures, and $C(C - 1)/2$ returns the number of unique pairwise correlations in the matrix. It excludes the main diagonal (because the correlation of a variable with itself, r_{ii} is always 1) and the lower triangular region of the covariance matrix (because it mirrors the upper triangular region). Because the CA method is constructed from the compound symmetry assumption, it is valid in this scenario. Further, under compound symmetry, both CM and CA methods are equivalent in distribution. To demonstrate this equivalence, I first show that the transformation from \mathbf{X} to \mathbf{Y} (subject-centering transformation, eq. 4a) results in a variance/covariance matrix whose correlations are all equal to a constant $-1/(C - 1)$ (see Appendix C). Consequently, under the CA method, the correction factor $\sqrt{1 - r}$ results in a correction factor of

$$\sqrt{1 - r} = \sqrt{1 - \left(-\frac{1}{C - 1}\right)} = \sqrt{\frac{C}{C - 1}}$$

which is exactly the conversion factor from $S_{\mathbf{Y},i}$ to $S_{\mathbf{Z},i}$ (Appendix A and Eq. 5'). Therefore, CA, CM and LM stan-

dard errors are all unbiased estimates of σ_e/\sqrt{N} under compound symmetry. The only difference is that LM pools the estimates so that a unique error bar width is used; in CA and CM, error bars differ slightly owing to small deviations to compound symmetry in the sample. The data from Table 1 showed a situation where compound symmetry is not reject (Welch factor is virtually equal to 1, $W_f = 0.999$ according to the Winer's test $M = 2.55$, $\chi^2(4) = 2.12, p = 0.713$, but keep in mind that power is probably very weak in this situation). CM, CA and LM error bars are quite comparable, as seen in Figure 1.

To confirm the overall absence of difference between CA, CM, and LM methods under compound symmetry, I ran simulations described in Appendix D (source code available on the Open Science Framework, <https://osf.io/zfbyb8/>). In a nutshell, I generated random covariance matrices with the only restriction that they had to satisfy a certain covariance structure. From it, I generated random multinormal data. Finally, I estimated the proportion of accurate coverage from the three methods. For compound symmetry



simulations, the assumption is met by sampling a single correlation and a single variance duplicated in the matrix (see Appendix D for more details; note that Appendix B explores deviations from compound symmetry).

The results are shown in Table 4 for the case of 3 and 5 repeated measures. As shown in the table, under compound symmetry, the coverage of all three methods is just a little above 95% on average and the standard deviations in the estimated interval widths are very small (0.3%) indicating correct coverage and very homogeneous results across covariance matrices.

This result is expected, but it provides a baseline to assess the importance of deviations for other variance/covariance matrices. In the subsequent simulations, two times 0.3% will be used as a threshold to identify inadequate methods.

Spherical covariance structure

In this second set of simulations, the covariance matrices meet the sphericity assumption. They could by chance also satisfy the compound symmetry assumption but this is quite improbable so I did not control for this possibility. Both CM and LM methods are based on difference scores and therefore, they should be adequate methods here. However, CA method is based on more stringent assumptions and therefore should not be a proper method in this scenario.

Simulations confirmed these predictions. The CA 95% confidence intervals have an average coverage of about 94%, below the threshold set above. Further, variability for this coverage is about 10 times larger. This large variability suggests that other factors not captured by the CA methods should influence its confidence interval width.

Other covariance structures

To have an indication of what might influence CA coverage, I also ran simulations under homosociostic and heteroscedastic covariance matrices (the upper right quadrant of Figure 2). Because CA assumes homogeneous variances, results should be poor and indeed, they turned out to be the worst, with a coverage factor of about 90% for 95% confidence intervals. Using the Welch factor to correct the degrees of freedom ($(N - 1) \times W_f$ instead of $(N - 1)$ as usual), coverage returned to 95% (95.0 for three repeated measures and 95.7% for five repeated measures); however, variability in these mean estimates was huge (8.4% and 9.3% respectively, almost 30 times larger than in the baseline simulation) so that this additional correction factor is not a safe approach. Unexpectedly, the CM method was adequate with this covariance structure (coverage of .953 with a variability of 0.024 with 3 measures; 0.954 with a variability of 0.025 for 5 measures) whereas LM was not

(coverage of 0.945 and 0.944 for 3 and 5 measures respectively).

In the last two covariance structures identified in Figure 2 (heterosociostic & homoscedastic and heterosociostic & heteroscedastic possibly including spherical matrices by chance), none of the methods were adequate, returning a coverage slightly below the desired threshold. This is not that surprising: these situations require multivariate techniques which are not embedded in the present methods.

Discussion

I presented a new method, the correlation-adjusted method, to compute within-subject standard errors and confidence intervals. I showed that—under compound symmetry—this method is equivalent in distribution to two former methods, the Cousineau-Morey and Loftus and Masson's methods. However, the correlation-adjusted method is not adequate when the data do not meet the compound symmetry assumption. Hence, Winer's test of compound symmetry, akin to Mauchly's test of sphericity, was discussed and documented.

The CA method is advantageous in one aspect: it is based on the mean correlations across replicated measures. Correlation is an intuitive measure, and most researchers know roughly the amount of correlation to expect between two repeated measures. For example, if the correlation is known to be approximately 0.75, then the basic CI should be reduced by half (as $\sqrt{1 - 0.75} = 0.5$). Thus, a mere pocket calculator is all that is needed to apply this correction. Placing correlation to the upfront is not a bad practice. Furthermore, Goulet and Cousineau (2019) explained how correlation is involved in statistical power.

However, the CA method is limited to data whose covariance matrix satisfies the compound symmetry assumption. Hence, a Winer test of compound symmetry could be performed prior to using this method. If compound symmetry is rejected, a Mauchly test of sphericity could be performed and if it does not lead to rejection, CM or LM can be used (whether one uses CM or LM is just a matter of preference as they are equivalent in the root mean squared sense). Note that (a) these tests may have limited statistical power; (b) some authors warn against the use of assumption-checking tests as it alters the error rates of null hypothesis statistical testing (e.g., Rochon, Gondan, & Kieser, 2012).

I provided code to perform the Winer test in R; we are also finishing a graphing module in R to plot descriptive statistics and corresponding error bars under any of the methods outlined here (Cousineau, Goulet, & Harding, submitted). It also includes the effect of sampling method (Cousineau & Laurencelle, 2015) and extends to other descriptive statistics (Harding, Tremblay, & Cousineau, 2014,



Table 4 ■ Coverage of the various 95% confidence interval methods under 2 covariance matrix structures. Between parentheses is the standard deviation of the coverage estimates across 1000 random covariance matrices. In light gray are cells which do not satisfy a proper coverage of 95%. The highlighted cells are more than 2 baseline standard deviations below the desired 95% coverage.

Covariance matrix structure	95% confidence interval method		
	CA	CM	LM
When there are 3 repeated measures			
Compound symmetry	0.951 (0.003)	0.952 (0.003)	0.952 (0.003)
Spherical	0.937 (0.034)	0.952 (0.003)	0.952 (0.003)
When there are 5 repeated measures			
Compound symmetry	0.952 (0.003)	0.952 (0.003)	0.953 (0.003)
Spherical	0.942 (0.032)	0.952 (0.003)	0.954 (0.003)

2015).

A safe approach would be to uniquely use CM (or LM) intervals and ignore CA intervals altogether. As a matter of fact, in the universe of random covariance matrices, those that are meeting the sphericity assumption are more numerous than those satisfying the more restrictive compound symmetry assumption. On the other hand, CA is the first method not based on data transformation.

To conclude, and paraphrasing Amrhein, Greenland, and McShane (2019), I believe that confidence intervals should be renamed. These authors proposed compatibility intervals whereas I already suggested precision intervals (Cousineau, 2017). This change in name is called for to reduce the blind reliance on these intervals and progressively switch to stronger argumentations and prospective explanations when discussing results.

Authors’ note

I would like to thank Thom Baguley, Étienne Dumesnil, Marc-André Goulet, Bradley Harding, Daniel Heck and Michael Masson for their comments on an earlier version of this text. This research was supported in part by the *Conseil pour la recherche en sciences naturelles et en génie du Canada*.

References

Abdi, H. (2010). The greenhouse-geisser correction. In). E. of Research Design . Thousand Oaks (Ed.), *Neil salkind (eds (pp. 1–10). Encyclopedia of Research Design . Thousand Oaks: Sage.*

Afshartous, D., & Preston, R. A. (2010). Confidence intervals for dependent data: Equating non-overlap with statistical significance. *Computational Statistics and Data Analysis, 54*, 2296–2305. doi:10.1016/j.csda.2010.04.011

Amrhein, V., Greenland, S., & McShane, B. (2019). Scientists rise up against statistical significance. *Nature, 567*, 305–307.

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

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

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. (2010). *Panorama des statistiques pour psychologues*. Bruxelles: Les éditions de Boeck Université.

Cousineau, D. (2017). Varieties of confidence intervals. *Advances in Cognitive Psychology, 13*, 140–155. doi:10.5709/acp-02140z

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

Cousineau, D., & O’Brien, F. (2014). Error bars in within-subject designs: A comment on baguley (2012). *Behavior Research Methods, 46*, 1149–1159. doi:10.3758/s13428-013-0441-z

Derrick, D., Toher, D., & White, P. (2016). Why welch’s test is type i error robust. *The Quantitative Methods for Psychology, 12*, 30–38. doi:10.20982/tqmp.12.1.p030

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



- 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](https://doi.org/10.2307/2983411)
- Goulet-Pelletier, J.-C., & Cousineau, D. (2018). A review of effect sizes and their confidence intervals, part i: The cohen's d family. *The Quantitative Methods for Psychology*, *14*, 242–265. doi:[10.20982/tqmp.14.4.p242](https://doi.org/10.20982/tqmp.14.4.p242)
- Goulet, M.-A., & Cousineau, D. (2019). The power of replicated measures to increase statistical power. *Advances in Methods and Practices in Psychological Sciences, Online, online*, 1–15. doi:[10.1177/2515245919849434](https://doi.org/10.1177/2515245919849434)
- 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](https://doi.org/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](https://doi.org/10.20982/tqmp.11.1.p032)
- Heck, D. W. (2019). Accounting for estimation uncertainty and shrinkage in bayesian within-subject intervals: A comment on nathoo, kilshaw, and masson (2018). *Journal of Mathematical Psychology*, *88*, 27–31. doi:[10.1016/j.jmp.2018.11.002](https://doi.org/10.1016/j.jmp.2018.11.002)
- 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](https://doi.org/10.3758/BF03210951)
- Mauchly, J. W. (1940). Significance test for sphericity of a normal n-variate distribution. *The Annals of Mathematical Statistics*, *11*, 200–209.
- 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.p061](https://doi.org/10.20982/tqmp.04.2.p061)
- Nathoo, F. S., Kilshaw, R. E., & Masson, M. E. J. (2018). A better (bayesian) interval estimate for within-subject designs. *Journal of Mathematical Psychology*, *86*, 1–9. doi:[10.1016/j.jmp.2018.07.005](https://doi.org/10.1016/j.jmp.2018.07.005)
- 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.
- 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](https://doi.org/10.2478/v10053-008-0133-x)
- 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](https://doi.org/10.1186/1471-2288-12-81)
- Roy, S. N. (1954). *A new test of compound symmetry (technical report series no. mimeograph series no. 97)*. Chapel Hill: Institute of statistics.
- Votaw, D. F., Jr. (1948). Testing compound symmetry in a normal multivariate distribution. *Annals of Mathematical Statistics*, *19*, 447–473. doi:[10.1214/aoms/1177730145](https://doi.org/10.1214/aoms/1177730145)
- Wallenstein, S., & Fleiss, J. L. (1979). Repeated measurements analyses of variance when the correlations have a certain pattern. *Psychometrika*, *44*, 229–233. doi:[10.1007/BF02293973](https://doi.org/10.1007/BF02293973)
- Welch, B. L. (1938). The significance of the difference between two means when the population variances are unequal. *Biometrika*, *29*, 350–362. doi:[10.2307/2332010](https://doi.org/10.2307/2332010)
- Wilks, S. S. (1938). The large-sample distribution of the likelihood ratio for testing composite hypotheses. *Annals of Mathematical Statistics*, *9*, 60–62. doi:[10.1214/aoms/1177732360](https://doi.org/10.1214/aoms/1177732360)
- Winer, B. J., Brown, D. R., & Michels, K. M. (1991). *Statistical principles in experimental design*. New York: McGraw-Hill.

Appendix A: Equivalence between Loftus and Masson's method and Cousineau and Morey's method

To demonstrate the equivalence in the root mean squared sense between the two methods, note that

$$\begin{aligned}
 S_{\mathbf{Y},i}^2 &= \frac{1}{N-1} \sum_{j=1}^N (\mathbf{Y}_{ij} - \bar{\mathbf{Y}}_i.)^2 \\
 &= \frac{1}{N-1} \sum_{j=1}^N (\mathbf{X}_{ij} - \bar{\mathbf{X}}_{.j} + \bar{\mathbf{X}} - \bar{\mathbf{X}}_i.)^2
 \end{aligned}$$



because $\bar{Y}_i = \bar{X}_i$, i.e., the means within conditions are unaltered by the transformation. Pooling all the conditions, we get

$$S_{Y,pool}^2 = \frac{1}{C} \sum_{i=1}^C S_{Y,i}^2$$

$$= \frac{1}{C(N-1)} \sum_{i=1}^C \sum_{j=1}^N (\mathbf{x}_{ij} - \bar{\mathbf{x}}_{.j} - \bar{\mathbf{x}}_i + \bar{\bar{\mathbf{x}}})^2$$

Consequently,

$$S_{Z,pool}^2 = \frac{C}{C-1} \left(\frac{1}{C(N-1)} \sum_{i=1}^C \sum_{j=1}^N (\mathbf{x}_{ij} - \bar{\mathbf{x}}_{.j} - \bar{\mathbf{x}}_i + \bar{\bar{\mathbf{x}}})^2 \right)$$

$$= \frac{1}{(C-1)(N-1)} \sum_{i=1}^C \sum_{j=1}^N (\mathbf{x}_{ij} - \bar{\mathbf{x}}_{.j} - \bar{\mathbf{x}}_i + \bar{\bar{\mathbf{x}}})^2$$

$$= \frac{SS_{S \times C}}{(C-1)(N-1)}$$

Another way to observe this equality is to average the error variances (the squared SE) of the CM method. We get exactly the squared SE of the LM method. In Table 2 for example, the CM standard errors are 0.191, 0.284 and 0.260, so that $\sqrt{(0.191^2 + 0.284^2 + 0.260^2)/3} = \sqrt{0.0616} = 0.248$ is the SE of the LM method.

Appendix: B statistical assessment of compound symmetry

In this appendix, I present a technique to assess the significance of deviation to compound symmetry. I examined a few proposals (Votaw, 1948; Roy, 1954; a likelihood ratio test based on Wilks, 1938, and an F test) but found that Winer (1971, reedited Winer et al., 1991, p. 517) was the most powerful test. Sadly, this test (actually, the correction factor; see below) is given without mathematical justification. Hence, I ran extensive tests to determine that it surpasses the other alternative tests.

The null hypothesis in which the covariance matrix, noted Σ , satisfies compound symmetry can be formalized as

$$H_0 : \Sigma = \bar{V} \mathbf{I}_q + \bar{r} \bar{V} (\mathbf{1}_q - \mathbf{I}_q)$$

in which \bar{V} is the average variance, \bar{r} is the average pairwise correlation, q is the number of repeated-measure variables, \mathbf{I}_q is the identity matrix of size $q \times q$, and $\mathbf{1}_q$ is a matrix containing only 1's of size $q \times q$. Note that in the main text, I use C to denote the number of repeated-measure variables; here I keep q to follow Winer et al.'s notation.

The test requires the following quantities

$$M = -(N-1) \ln \left(\frac{|S_1|}{|S_0|} \right)$$

$$C = \frac{\nu + 2}{\nu} \times \frac{q + 1}{q - 1} \times \frac{2q - 3}{6(N - 1)}$$

in which N is the sample size, and ν is the degree of freedom of the test, that is, $\nu = q(q + 1)/2 - 2$. The expression $|S_1|$ is the determinant of the observed covariance matrix whereas $|S_0|$ is the determinant of the covariance matrix based on the null hypothesis. The quantity M is equivalent to twice the log-likelihood ratio, $2(\ln l_1 - \ln l_0)$, assuming a multinormal distribution with means set to the observed means in both case. As shown in Wilks (1938), this ratio has asymptotically a chi-square distribution with the degree of freedom equal to the difference in the number of free parameters. For the numerator, the upper triangular half of the covariance matrix (of size $q(q + 1)/2$) and the mean vector (of size q) contain free parameters; for the null model, the means, a common variance and a common covariance are free parameters (hence, $q + 2$ free parameters). The difference, $q(q + 1)/2 - 2$, is therefore the degree of freedom and asymptotically



Listing 1 ■ Listing 1. R code to perform a test of compound symmetry using Winer’s corrected chi-square test. This function requires a data frame as input with only the repeated-measure variables; it assesses the significance of the null hypothesis that the covariance matrix is compound symmetric. This test is found without demonstration in Winer, Brown, and Michels (1991), p. 517.

```
WinerCompoundSymmetryTest <- function(X) {
  # This function requires a data frame X as input with only the
  # repeated-measure variables. It assesses the significance of the
  # null hypothesis that the covariance matrix is compound symmetric.
  # This test is given without demonstration in
  # Winer, Browns, & Michels, 1991, p. 517.

  # Get basic descriptive statistics
  q <- length(X)
  n <- dim(X)[1]
  S1 <- cov(X)

  # get H0 statistics
  vbar <- mean(diag(S1))
  cbar <- mean(S1[upper.tri(S1)])
  S0 <- vbar * diag(q) + (1-diag(q)) * cbar

  # the chi-square test corrected for small sample;
  # M is a shortcut for the likelihood ratio
  # cf is a correction factor for small samples
  # df is the degree of freedom of the test distribution
  M <- -(n-1) * log( det(S1) / det(S0) )
  cf <- (q * (q+1)^2 * (2*q-3)) / (6 * (n-1) * (q-1) * (q^2 + q - 4))
  df <- q*(q+1)/2-2
  W <- M * (1 - cf)
  pW <- 1-pchisq(W, df )

  cat("M_=" , M, " , _W(", df, ") _=", W, " , _p_=", pW, "\n", sep = " ")
}
```

$M \approx \chi^2(\nu)$. This test performs generally well, but requires large samples to have a type-I error rate close to α . For small samples (e.g., 16 observations), the error rate inflates to close to twice the value of α .

Winer introduced a correction factor which alters the likelihood ratio so that the test statistic is adequate for all sample sizes (Winer et al., 1991). The quantity

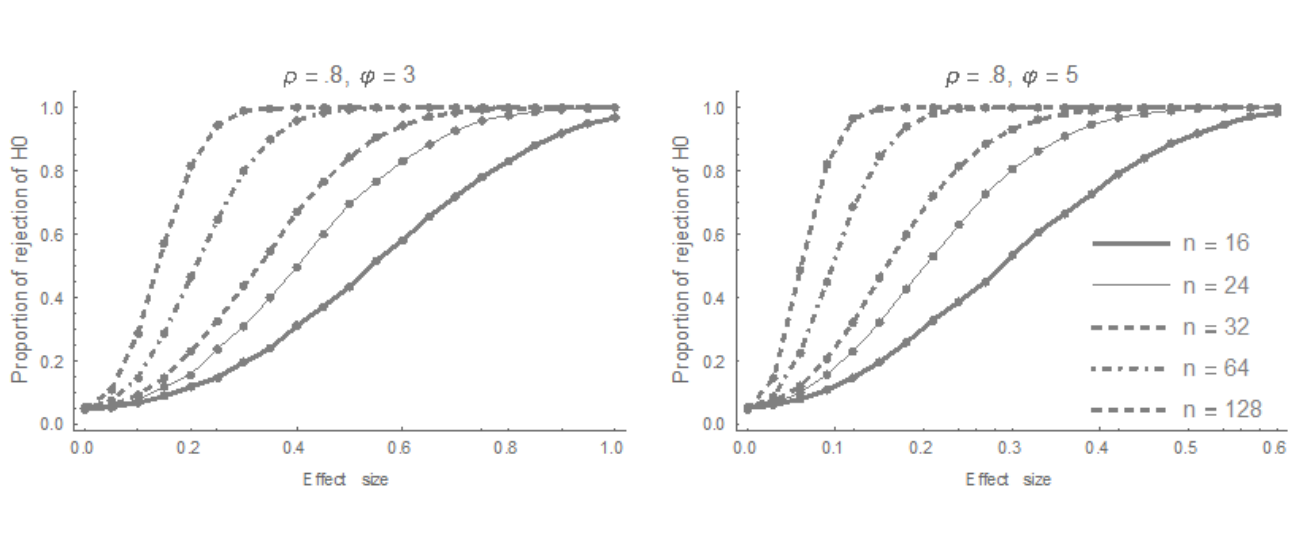
$$(1 - C)M \sim \chi^2(\nu)$$

follow a chi-square distribution with degree of freedom given by $q(q + 1)/2 - 2$. Listing 1 provides R code to perform the Winer test given a data frame.

To evaluate the test, I ran systematic tests. In one simulation, I used a common variance for all variables but altered the correlations so that pairs of adjacent variables (e.g., variables 1 and 2) were sampled from a simulated population with a true correlation of ρ ; pairs of variables two positions remote (e.g., variables 1 and 3) had a weaker population correlation of $\rho 2^{-\delta}$, and so on as we consider pairs further away ($\rho i^{-\delta}$ where i denotes the separation between variables). The parameter δ is the effect size: when set to zero, there is no decay of ρ and consequently, the covariance matrix is compound symmetric; when δ is 1, there is a rapid decay of the correlations and consequently, an important deviation to compound symmetry. Wallenstein and Fleiss (1979) examined a closely-related covariance structure, called a simplex configuration (the difference being that here, the decay follows a power curve whereas in Wallenstein and Fleiss, the decay in correlation is exponential).



Figure 3 ■ Power curves of the Winer’s test for samples of sizes 16 to 128 (lines) and for an effect size ranging from nil to a strong decay of the pairwise correlation (horizontal axis). Left panel: three repeated measures are generated; right panel: five repeated measures are generated. In both, the base correlation is .8.



I generated samples of size 16, 24, 32, 64, and 128. I also varied the number of variables and the true value of ρ . In each cell, I generated 10,000 multivariate samples and ran 10,000 Winer tests with a decision threshold of .05. I counted the proportion of test suggesting a rejection of the null hypothesis. When δ is zero, the rate of rejections represents the type-I error rate; when δ is larger than zero, the rate of rejections represents the statistical power of the test.

Figure 3 shows some of the results. As seen, the power curves all start very close to .05, and that, irrespective of the sample sizes tested, of the number of repeated-measure variables, and of the true correlation ρ . On average across all the simulations explored, the type-I error rate was 4.91%, very close to 5%. As usual, statistical power increases faster with larger sample sizes.

In a second simulation, I let ρ be a constant. However, I manipulated the variance. The standard deviation of the variable in the center of the covariance matrix was 15 (or the two variables in the center if there is an even number of variables); for every variable away from that central variable(s), standard deviation was decreased by δ . Thus, for δ of zero, there is no alteration in the variances and the matrix is compound symmetric. Everything else is as in the first simulation.

Figure 4 shows some of the results. Again, all the power curve starts very close to the α level (mean across the simulations tested of 4.94%). Statistical power increases rapidly, more so when the effect size is larger and when sample size is larger. Also, power increases faster when the number of variables is larger. This last result is not quite surprising as there are more raw data when the number of variables increases.

In summary, the Winer test of compound symmetry was found to be an excellent way to assess deviation to compound symmetry with larger W indicating larger deviations. The significance of W can be evaluated reliably with a p value taken from a chi-square distribution.

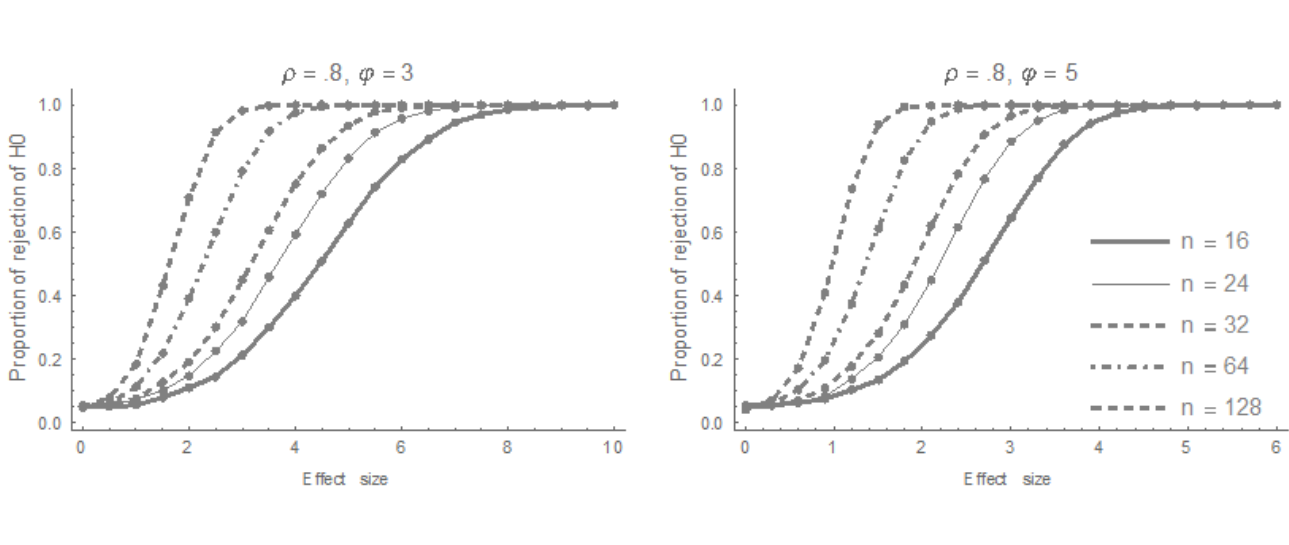
Appendix C: Demonstration that subject-centering transformation results in pairwise correlations of $-1/(C - 1)$ under compounds symmetry

I begin by assuming a hierarchical linear model

$$X_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}$$



Figure 4 ■ Power curves of the Winer’s test for samples of sizes 16 to 128 (lines) and for an effect size on variances ranging from nil to a rapid decrease of the variances (horizontal axis). Left panel: three repeated measures are generated; right panel: five repeated measures are generated. In both, the base correlation is .8.



in which

$$\alpha_i \sim \mathcal{N}(0, \sigma_a^2)$$

$$\varepsilon_{ij} \sim \mathcal{N}(0, \sigma_e^2)$$

where $\alpha_i, i = 1 \dots N$ is the subject effect and $\varepsilon_{ij}, j = 1 \dots C$ is random error; both are independent; β_j is the effect of the j th condition, assumed a fixed effect.

As usual under this framework, the correlation is $\rho = \sigma_a^2 / (\sigma_a^2 + \sigma_e^2)$ and the variance of the measurements is $\sigma^2 = \sigma_a^2 + \sigma_e^2$. Finally, the covariance matrix of \mathbf{X} is

$$\Sigma_{\mathbf{X}} = \mathbf{I}_C (\sigma_a^2 + \sigma_e^2) + (\mathbf{1}_C - \mathbf{I}_C) \sigma_a^2$$

where \mathbf{I}_C returns the identity matrix of size $C \times C$ and $\mathbf{1}_C$ returns a matrix filled with 1s of size $C \times C$.

To get the subject-centering transformation, one approach is to have a transformation matrix, here defined as

$$\mathbf{K} = \mathbf{I}_C - \frac{1}{C} \mathbf{1}_C$$

such that

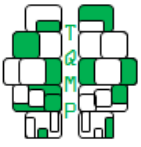
$$\mathbf{Y} = \mathbf{K} \cdot \mathbf{X}$$

is the subject-centered data (add the grand mean $\bar{\bar{\mathbf{X}}}$ to get \mathbf{Y} as defined in the main text, Eq. 4a). From this transformation, we know that

$$\Sigma_{\mathbf{Y}} = \mathbf{K} \cdot \Sigma_{\mathbf{X}} \cdot \mathbf{K}^T$$

from which we derive that

- a) in the main diagonal of the covariance matrix, $\Sigma_{\mathbf{Y}}(ii) = \frac{C-1}{C} \sigma_e^2$
- b) off the main diagonal, $\Sigma_{\mathbf{Y}}(ij, j \neq i) = -\frac{1}{C} \sigma_e^2$.



Consequently, the correlation ρ_{ij} is

$$\begin{aligned} \rho_{ij} &= \frac{\Sigma_Y(ij)}{\sqrt{\Sigma_Y(ii)} \times \sqrt{\Sigma_Y(jj)}} \\ &= \frac{-\sigma_e^2/C}{(C-1)\sigma_e^2/C} \\ &= -\frac{1}{C-1} \end{aligned}$$

■

Appendix D: Simulations to assess coverage of the confidence intervals and the generation of random data

Simulations were used to test the confidence intervals. Herein, the normality assumption is satisfied so that the present confidence intervals should behave identically to a test of null hypothesis. In particular, in the absence of a difference, a 95% confidence interval should contain 95% of the times the null hypothesis. When applied to multiple groups (or multiple measures here), the difference-adjusted 95% confidence interval of one mean should include 95% of the time the other means. This is what I called coverage in the main text. Coverage should not be less than the desired confidence interval for an interval to be a valid confidence interval.

Assuming multinormality, the only two parameters required are the vector of means (all equal when there is no true population difference) and the covariance matrix. In a typical simulation, I generated a random covariance matrix (under a given scenario of scedasticity and sociosticity; see next). I then generated 1000 data sets each with 64 simulated participants, computed the error bars for the first measure, and checked that the last measure was included within that confidence interval (these conditions were chosen arbitrarily). From those 1000 simulations, I got one estimate of the proportion of coverage of the method used. I repeated the process a 1000 times to have a thousand random covariance matrices tested. The mean coverage as well as the standard deviation in the coverage is finally retained.

To generate compound symmetric covariance matrices, I used

$$\begin{aligned} \rho &\sim \mathcal{U}(-1/(C-1), +1)_1 \\ \sigma &\sim \mathcal{U}(0, 25)_1 \\ \Sigma_{C \times C} &= \mathbf{I}_C \sigma^2 + (\mathbf{1}_C - \mathbf{I}_C) \sigma^2 \rho \end{aligned}$$

in which $\mathcal{U}(low, high)_1$ returns a uniform random number between low and high. The subscript 1 at the end of the notation is used to highlight that a single scalar is returned; in the subsequent scenarios it can return vectors or matrices of random numbers as well. Correlation is above $-1/(C-1)$ to make sure that all the matrices are positive definite. The symbol \sim is used to denote one possible realization of the random number generator. As usual, \mathbf{I}_C is the $C \times C$ identity matrix and $\mathbf{1}_C$ is a $C \times C$ matrix filled with 1s.

Homosociostic and heteroscedastic covariance matrices are obtained from

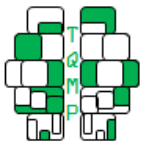
$$\begin{aligned} \rho &\sim \mathcal{U}(-1/(C-1), +1)_1 \\ \sigma_C &\sim \mathcal{U}(0, 25)_C \\ r_{C \times C} &= \mathbf{I}_C + (\mathbf{1}_C - \mathbf{I}_C) \rho \\ \Sigma_{C \times C} &= \sigma \cdot \sigma^T \times r \end{aligned}$$

where \times is the point-to-point multiplication.

Heterosociostic and homoscedastic covariance matrices are obtained from

$$\begin{aligned} \Sigma_{C \times C} &\sim \mathcal{U}(-25, +25)_{C \times C} \\ \Sigma &= \Sigma^T \cdot \Sigma \\ \sigma_{C \times C} &= \sqrt{Diag(\Sigma) \cdot Diag(\Sigma)^T} \\ \Sigma_{C \times C} &= \Sigma / \sigma^2 \times \mathcal{U}(0, 25)_1^2 \end{aligned}$$

where $Diag(\Sigma)$ is a vector containing the main diagonal of Σ . On step 2, the outer product of Σ is used to get a positive definite matrix.



Spherical covariance matrices are obtained from the H matrices described in Winer et al. (1991), p. 241:

$$\begin{aligned}a_C &\sim \mathcal{U}(-75, +75)_C \\ \mathbf{A}_{C \times C} &= \{a_C; a_C; \dots; a_C\}_{C \times C} \\ l &\sim \mathcal{U}(-75, +75)_1 \\ \Sigma &= \mathbf{A} + \mathbf{A}^T + \mathbf{I}_C l\end{aligned}$$

Finally, heterosociostic and heteroscedastic covariance matrices are obtained with

$$\begin{aligned}\Sigma &\sim \mathcal{U}(-25, +25)_{C \times C} \\ \Sigma &= \Sigma^T \cdot \Sigma\end{aligned}$$

Citation

Cousineau, D. (2019). Correlation-adjusted standard errors and confidence intervals for within-subject designs: A simple multiplicative approach. *The Quantitative Methods for Psychology*, 15(3), 226–241. doi:[10.20982/tqmp.15.3.p226](https://doi.org/10.20982/tqmp.15.3.p226)

Copyright © 2019, *Cousineau*. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Received: 01/01/2019 ~ Accepted: 21/08/2019