Chapter 14: Repeated-measures designs

Oliver Twisted

Please, Sir, can I have some more … sphericity?

The following article is adapted from:

Field, A. P. (1998). A bluffer's guide to sphericity. *Newsletter of the Mathematical, Statistical and Computing Section of the British Psychological Society*, 6(1), 13–22.

The use of repeated measures, where the same subjects are tested under a number of conditions, has numerous practical and statistical benefits. For one thing, it reduces the error variance caused by between-group individual differences; however, this reduction of error comes at a price because repeated-measures designs potentially introduce covariation between experimental conditions (this is because the same people are used in each condition and so there is likely to be some consistency in their behaviour across conditions). In between-group ANOVA we have to assume that the groups we test are independent for the test to be accurate (Scariano & Davenport, 1987, have documented some of the consequences of violating this assumption). As such, the relationship between treatments in a repeated-measures design creates problems with the accuracy of the test statistic. The purpose of this article is to explain, as simply as possible, the issues that arise in analysing repeated-measures data with ANOVA: specifically, what is sphericity and why is it important?

What is sphericity?

Most of us are taught during our degrees that it is crucial to have homogeneity of variance between conditions when analysing data from *different* subjects, but often we are left to assume that this problem 'goes away' in repeated-measures designs. This is not so, and the assumption of sphericity can be likened to the assumption of homogeneity of variance in between-group ANOVA.

Sphericity (denoted by ε and sometimes referred to as *circularity*) is a more general condition of *compound symmetry*. Imagine you had a population covariance matrix $\mathbf{\Sigma}$, where:

$$\mathbf{\Sigma} = \begin{bmatrix} s_{11}^2 & a_{12} & a_{13} & \dots & a_{1n} \\ a_{21} & s_{22}^2 & a_{23} & \dots & a_{2n} \\ a_{31} & a_{32} & s_{33}^2 & \dots & a_{3n} \\ \dots & \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & a_{n3} & \dots & s_{nn}^2 \end{bmatrix} \qquad (1)$$

This matrix represents two things: (1) the off-diagonal elements represent the covariances between the treatments 1, …, *n* (you can think of this as the unstandardised correlation between each of the repeated-measures conditions); and (2) the diagonal elements signify the variances within each treatment. As such, the assumption of homogeneity of variance between treatments will hold when:

$$s_{11}^2 \approx s_{22}^2 \approx s_{33}^2 \approx \dots \approx s_{nn}^2 \qquad (2)$$

(i.e. when the diagonal components of the matrix are approximately equal). This is comparable to the situation we would expect in a between-group design. However, in repeated-measures designs there is the added complication that the experimental conditions covary with each other. The end result is that we have to consider the effect of these covariances when we analyse the data, and specifically we need to assume that all of the covariances are approximately equal (i.e. all of the conditions are related to each other to the same degree and so the effect of participating in one treatment level after another is also equal). Compound symmetry holds when there is a pattern of constant variances along the diagonal (i.e. homogeneity of variance — see **Error! Reference source not found.**equation (2)) and constant covariances off of the diagonal (i.e. the covariances between treatments are equal — see **Error! Reference source not found.**equation (3)). While compound symmetry has been shown to be a sufficient condition for conducting ANOVA on repeated-measures data, it is not a necessary condition.

$$a_{12} \approx a_{13} \approx a_{23} \approx \dots \approx a_{1n} \approx a_{2n} \approx a_{3n} \approx \dots \qquad (3)$$

Sphericity is a less restrictive form of compound symmetry (in fact much of the early research into repeated-measures ANOVA confused compound symmetry with sphericity). Sphericity refers to the equality of variances of the *differences* between treatment levels.

Whereas compound symmetry concerns the covariation between treatments, sphericity is related to the variance of the differences between treatments. So, if you were to take each pair of treatment levels, and calculate the differences between each pair of scores, then it is necessary that these differences have equal variances. Imagine a situation where there are 4 levels of a repeated-measures treatment (A, B, C, D). For sphericity to hold, one condition must be satisfied:

$$s^2_{\text{A–B}} \approx s^2_{\text{A–C}} \approx s^2_{\text{A–D}} \approx s^2_{\text{B–C}} \approx s^2_{\text{B–D}} \approx s^2_{\text{C–D}} \qquad (4)$$

Sphericity is violated when the condition in **Error! Reference source not found.**equation (4) is not met (i.e. the differences between pairs of conditions have unequal variances).

How is sphericity measured?

The simplest way to see whether or not the assumption of sphericity has been met is to calculate the differences between pairs of scores in all combinations of the treatment levels. Once this has been done, you can simply calculate the variance of these differences. For example, Table 1 shows data from an experiment with 3 conditions (for simplicity there are only 5 scores per condition). The differences between pairs of conditions can then be calculated for each subject. The variance for each set of differences can then be calculated. We saw above that sphericity is met when these variances are roughly equal. For this data, sphericity will hold when:

$$s^2_{\text{A–B}} \approx s^2_{\text{A–C}} \approx s^2_{\text{B–C}}$$

where:

$$\begin{aligned} s^2_{\text{A–B}} &= 15.7 \\ s^2_{\text{A–C}} &= 10.3 \\ s^2_{\text{B–C}} &= 10.3 \end{aligned}$$

As such,

$$s^2_{\text{A–B}} \neq s^2_{\text{A–C}} = s^2_{\text{B–C}}$$

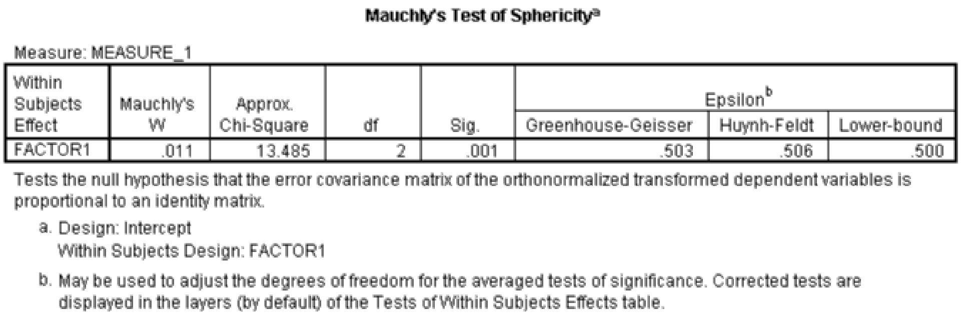
Condition A	Condition B	Condition C	A−B	A−C	B−C
10	12	8	−2	2	5
15	15	12	0	3	3
25	30	20	−5	5	10
35	30	28	5	7	2
30	27	20	3	10	7
		Variance:	**15.7**	**10.3**	**10.3**

Table 1: Hypothetical data to illustrate the calculation of the variance of the differences between conditions.

So there is at least some deviation from sphericity because the variance of the differences between conditions A and B is greater than the variance of the differences between conditions A and C, and between B and C. However, we can say that this data has *local circularity* (or local sphericity) because two of the variances are identical. This means that for any multiple comparisons involving these differences, the sphericity assumption has been met (for a discussion of local circularity see Rouanet & Lépine, 1970). The deviation from sphericity in the data in Table 1 does not seem too severe (all variances are *roughly* equal). This raises the issue of how we assess whether violations from sphericity are severe enough to warrant action.

Assessing the severity of departures from sphericity

Luckily the advancement of computer packages makes it needless to ponder the details of how to assess departures from sphericity. SPSS produces a test known as Mauchly's test, which tests the hypothesis that the variances of the differences between conditions are equal. Therefore, if Mauchly's test statistic is significant (i.e. has a probability value less than .05) we must conclude that there are significant differences between the variance of differences, ergo the condition of sphericity has not been met. If, however, Mauchly's test statistic is non-significant (i.e. $p > .05$) then it is reasonable to conclude that the variances of differences are not significantly different (i.e. they are roughly equal). So, in short, if Mauchly's test is significant then we must be wary of the *F*-ratios produced by the computer.



Mauchly's Test of Sphericity[a]

Measure: MEASURE_1

Within Subjects Effect	Mauchly's W	Approx. Chi-Square	df	Sig.	Epsilon[b]		
					Greenhouse-Geisser	Huynh-Feldt	Lower-bound
FACTOR1	.011	13.485	2	.001	.503	.506	.500

Tests the null hypothesis that the error covariance matrix of the orthonormalized transformed dependent variables is proportional to an identity matrix.

a. Design: Intercept
Within Subjects Design: FACTOR1

b. May be used to adjust the degrees of freedom for the averaged tests of significance. Corrected tests are displayed in the layers (by default) of the Tests of Within Subjects Effects table.

Figure 1~~2~~: Output of Mauchly’s test from SPSS version 7.0

Figure 1~~Figure 2~~ shows the result of Mauchly’s test on some fictitious data with three conditions (A, B and C). The result of the test is highly significant, indicating that the variance between the differences were significantly different. The table also displays the degrees of freedom (the *df* are simply $N - 1$, where N is the number of variances compared) and three estimates of sphericity (see section on correcting for sphericity).

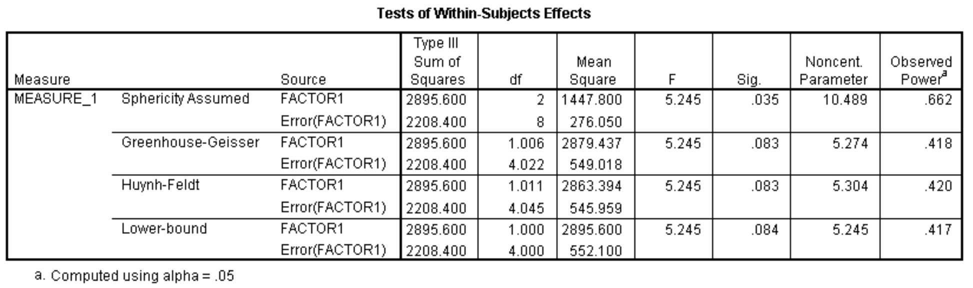
What is the effect of violating the assumption of sphericity?

Rouanet and Lépine (1970) provided a detailed account of the validity of the *F*-ratio when the sphericity assumption does not hold. They argued that there are two different *F*-ratios that can be used to assess treatment comparisons. The two types of *F*-ratio were labelled F' and F'' respectively. F' refers to an *F*-ratio derived from the mean squares of the comparison in question and the interaction of the subjects with that comparison (i.e. the specific error term for each comparison is used — this is the *F*-ratio normally used). F'' is derived not from the specific error mean square but from the total error mean squares for all of the repeated-measures comparisons. Rouanet and Lépine (1970) argued that F' is less powerful than F'' and so it may be the case that this test statistic misses genuine effects. In addition, they showed that for F' to be valid the covariation matrix, Σ, must obey local circularity (i.e. sphericity must hold for the *specific comparison in question*), and Mendoza, Toothaker, and Crain (1976) have supported this by demonstrating that the *F* ratios of an $L \times J \times K$ factorial design with two repeated-measures are valid only if local circularity holds. F' requires only *overall* circularity (i.e. the whole data set must be circular), but because of the non-reciprocal nature of circularity and compound symmetry, F'' does not require compound symmetry whilst F' does. So, given that F' is the statistic generally used, the effect of violating sphericity is a loss of power (compared to when F'' is used) and a test statistic (F-ratio) which simply cannot be validity compared to tabulated values of the F-distribution.

Correcting for violations of sphericity

If data violates the sphericity assumption there are a number of corrections that can be applied to produce a valid F-ratio. SPSS produces three corrections based upon the estimates of sphericity advocated by Greenhouse and Geisser (1959) and Huynh and Feldt (1976). Both of these estimates give rise to a correction factor that is applied to the degrees of freedom used to assess the observed value of F. How each estimate is calculated is beyond the scope of this article; for our purposes all we need know is that each estimate differs slightly from the others. The Greenhouse–Geisser estimate (usually denoted as $\hat{\varepsilon}$) varies between $1/(k - 1)$ (where k is the number of repeated-measures conditions) and 1. The closer that $\hat{\varepsilon}$ is to 1, the more homogeneous are the variances of differences, and hence the closer the data are to being spherical. Figure 1 shows a situation with three conditions and hence the lower limit of $\hat{\varepsilon}$ is .5; it is clear that the calculated value of $\hat{\varepsilon}$ is .503 which is very close to .5 and so represents a substantial deviation from sphericity. Huynh and Feldt (1976) reported that when $\hat{\varepsilon} > .75$ too many false null hypotheses fail to be rejected (i.e. the test is too conservative) and Collier, Baker, Mandeville, & Hayes (1967) showed that this was also true with $\hat{\varepsilon}$ as high as .90. Huynh and Feldt, therefore, proposed a correction to $\hat{\varepsilon}$ to make it less conservative (usually denoted as $\tilde{\varepsilon}$). However, Maxwell and Delaney (1990) report that $\tilde{\varepsilon}$ actually overestimates sphericity. Stevens (1992) therefore recommends taking an average of the two and adjusting the *df* by this averaged value. Girden (1992) recommends that when $\hat{\varepsilon} > .75$ the *df* should be corrected using $\tilde{\varepsilon}$; if $\hat{\varepsilon} < 0.75$, or nothing is known about sphericity at all, then the conservative $\hat{\varepsilon}$ should be used to adjust the *df*.

Comment [RL1]:



Tests of Within-Subjects Effects

Measure		Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Noncent. Parameter	Observed Power^a
MEASURE_1	Sphericity Assumed	FACTOR1	2895.600	2	1447.800	5.245	.035	10.489	.662
		Error(FACTOR1)	2208.400	8	276.050				
	Greenhouse-Geisser	FACTOR1	2895.600	1.006	2879.437	5.245	.083	5.274	.418
		Error(FACTOR1)	2208.400	4.022	549.018				
	Huynh-Feldt	FACTOR1	2895.600	1.011	2863.394	5.245	.083	5.304	.420
		Error(FACTOR1)	2208.400	4.045	545.959				
	Lower-bound	FACTOR1	2895.600	1.000	2895.600	5.245	.084	5.245	.417
		Error(FACTOR1)	2208.400	4.000	552.100				

a. Computed using alpha = .05

Figure 2~~3~~: Output of epsilon corrected F-values from SPSS version 7.0.

Figure 2~~Figure 3~~ shows a typical ANOVA table for a set of data that violated sphericity (the same data used to generate Figure 1~~Figure 2~~). The table in Figure 2~~Figure 3~~ shows the *F*-ratio and associated degrees of freedom when sphericity is assumed; as can be seen, this results in a significant *F*-statistic, indicating some difference(s) between the means of the three conditions. Underneath are the corrected values (for each of the three estimates of sphericity). Notice that in all cases the *F*-ratios remain the same, it is the degrees of freedom that change (and hence the critical value of *F*). The degrees of freedom are corrected by the estimate of sphericity. How this is done can be seen in Table 2. The new degrees of freedom are then used to ascertain the critical value of *F*. For this data this results in the observed *F* being non-significant at $p < .05$. This particular data set illustrates how important it is to use a valid critical value of *F*, it can mean the difference between a statistically significant result and a non-significant result. More importantly, it can mean the difference between making a Type I error and not.

Estimate of Sphericity Used	Value of Estimate	Term	*df*	Correction	New *df*
None		Effect Error	2 8		
	0.503	Effect Error	2 8	0.503 × 2 0.503 × 8	1.006 4.024
	0.506	Effect Error	2 8	0.506 × 2 0.506 × 8	1.012 4.048

Table 2: Shows how the sphericity corrections are applied to the degrees of freedom.

Multivariate vs. univariate tests

A final option, when you have data that violates sphericity, is to use multivariate test statistics (MANOVA) because they are not dependent upon the assumption of sphericity (see O'Brien & Kaiser, 1985). There is a trade-off of test power between univariate and multivariate approaches, although some authors argue that this can be overcome with suitable mastery of the techniques (O'Brien & Kaiser, 1985). MANOVA avoids the assumption of sphericity (and all the corresponding considerations about appropriate *F* ratios and corrections) by using a specific error term for contrasts with 1 *df* , and hence each contrast is only ever associated with its specific error term (rather than the pooled error terms used in ANOVA). Davidson

(1972) compared the power of adjusted univariate techniques with those of Hotelling's T^2 (a MANOVA test statistic) and found that the univariate technique was relatively powerless to detect small reliable changes between highly correlated conditions when other less correlated conditions were also present. Mendoza, Toothaker, and Nicewander (1974) conducted a Monte Carlo study comparing univariate and multivariate techniques under violations of compound symmetry and normality and found that 'as the degree of violation of compound symmetry increased, the empirical power for the multivariate tests also increased. In contrast, the power for the univariate tests generally decreased' (p. 174). Maxwell and Delaney (1990) noted that the univariate test is relatively more powerful than the multivariate test as *n* decreases and proposed that 'the multivariate approach should probably not be used if *n* is less than *a* + 10 (*a* is the number of levels for repeated-measures)' (p. 602). As a general rule it seems that when you have a large violation of sphericity ($\varepsilon < 0.7$) and your sample size is greater than *a* + 10 then multivariate procedures are more powerful, whilst with small sample sizes or when sphericity holds ($\varepsilon > 0.7$) the univariate approach is preferred (Stevens, 1992). It is also worth noting that the power of MANOVA increases and decreases as a function of the correlations between dependent variables (Cole, Maxwell, Arvey, & Salas, 1994) and so the relationship between treatment conditions must be considered also.

Comment [RL2]:

Multiple comparisons

So far, I have discussed the effects of sphericity on the omnibus ANOVA. As a final flurry some discussion of the effects on multiple comparison procedures is warranted. Boik (1981) provided an estimable account of the effects of non-sphericity on *a priori* tests in repeated-measures designs, concluded that even very small departures from sphericity produce large biases in the *F*-test, and recommends against using these tests for repeated-measures contrasts. When experimental error terms are small, the power to detect relatively strong effects can be as low as .05 (when sphericity = .80). He argues that the situation for *a priori* comparisons cannot be improved and concludes by recommending a multivariate analogue. Mitzel and Games (1981) found that when sphericity does not hold ($\varepsilon < 1$) the pooled error term conventionally employed in pairwise comparisons resulted in non-significant differences between two means declared significant (i.e. a lenient Type I error rate) or undetected differences (a conservative Type I error rate). They therefore recommended the use of separate error terms for each comparison. Maxwell (1980) systematically tested the power and alpha levels for 5 *a priori* tests under repeated-measures conditions. The tests assessed were Tukey's wholly significant difference (WSD) test which uses a pooled error term, Tukey's procedure but with a separate error term with either $n - 1$ df (labelled SEP1) or $(n - 1)(k - 1)$ df (labelled SEP2), Bonferroni's procedure (BON), and a multivariate approach — the Roy–Bose simultaneous confidence interval (SCI). Maxwell tested these *a priori* procedures, varying the sample size, number of levels of the repeated factor and departure from sphericity. He found that the multivariate approach was always 'too conservative for practical use' (p. 277) and this

Comment [RL3]:

was most extreme when *n* (the number of subjects) is small relative to *k* (the number of conditions). Tukey's test inflated the alpha rate as the covariance matrix departs from sphericity, and even when a separate error term was used (SEP1) alpha was slightly inflated as *k* increased whilst SEP2 also led to unacceptably high alpha levels. The Bonferroni method, however, was extremely robust (although *slightly* conservative) and controlled alpha levels regardless of the manipulation. Therefore, in terms of Type I error rates the Bonferroni method was best. In terms of test power (the Type II error rate) for a small sample (*n* = 8) WSD was the most powerful under conditions of non-sphericity. This advantage was severely reduced when *n* = 15. Keselman and Keselman (1988) extended Maxwell's work and also investigated unbalanced designs. They too used Tukey's WSD, a modified WSD (with non-pooled error variance), Bonferroni *t*-statistics, and a multivariate approach, and looked at the same factors as Maxwell (with the addition of unequal samples). They found that when unweighted means were used (with unbalanced designs) none of the four tests could control the Type I error rate. When weighted means were used only the multivariate tests could limit alpha rates, although Bonferroni *t* statistics were considerably better than the two Tukey methods. In terms of power they concluded that 'as the number of repeated treatment levels increases, BON is substantially more powerful than SCI' (p. 223).

So, in terms of these studies, the Bonferroni method seems to be generally the most robust of the univariate techniques, especially in terms of power and control of the Type I error rate.

Conclusion

It is more often the rule than the exception that sphericity is violated in repeated-measures designs. For this reason, all repeated-measures designs should be exposed to tests of violations of sphericity. If sphericity is violated then the researcher must decide whether a multivariate or univariate analysis is preferred (with due consideration to the trade-off between test validity on the one hand and power on the other). If univariate methods are chosen then the omnibus ANOVA must be corrected appropriately, depending on the level of departure from sphericity. Finally, if pairwise comparisons are required the Bonferroni method should probably be used to control the Type I error rate. Finally, ensure that the group sizes are equal otherwise even the Bonferroni technique is subject to inflations of alpha levels.

References

Mitzel, H. C., & Games, P. A. (1981). Circularity and multiple comparisons in repeated measures designs. *British Journal of Mathematical and Statistical Psychology*, *34*, 253–259.

Stevens, J. (1992). *Applied multivariate statistics for the social sciences* (2nd edition). Hillsdale, NJ: Erlbaum.

All other references can be found at the back of the book.