UNDERSTANDING THE REPEATED-MEASURES ANOVA

REPEATED MEASURES ANOVA – Analysis of Variance in which subjects are measured more than once to determine whether statistically significant change has occurred, for example, from the pretest to the posttest. (Vogt, 1999)

- **REPEATED MEASURES (ANOVA)** – An ANOVA in which subjects are measured two or more times and the total variation is partitioned into three components: (1) variation among individuals; (2) variation among test occasions; and (3) residual variation (Hinkle, Wiersma, & Jurs, 2003).

REPEATED-MEASURES DESIGN – A research design in which subjects are measured two or more times on the dependent variable. Rather than using different subjects for each level of treatment, the subjects are given more than one treatment and are measured after each. This means that each subject will be its own control. This research design goes by several different names, including within-subjects ANOVA, treatments-by-subjects ANOVA, randomized-blocks ANOVA, one-way repeated-measures ANOVA and correlated groups design. (Vogt, 1999)

SPHERICITY ASSUMPTION – A statistical assumption important for repeated-measures ANOVAs. When it is violated, *F* values will be positively biased. Researchers adjust for this bias by raising the critical value of *F* needed to attain statistical significance. Mauchley's test for sphericity is the most common way to see whether the assumption has been met. (Vogt, 1999)

RESIDUAL VARIATION – Variation not due to either individuals or test occasions in repeated measures ANOVA (Hinkle, Wiersma, & Jurs, 2003).

ANOVA: SIMPLE REPEATED MEASURES designs involve measuring an individual two or more times on the dependent variable. For example, a researcher may test the same sample of individuals under different conditions or at different times. These people's scores comprise dependent samples. For example, learning experiments often involve measuring the same people's performance solving a problem under different conditions. Such a situation is analogous to, and an extension of, the design in which the dependent-samples *t* test was applied. With the one-way repeated-measures designs, each subject or case in a study is exposed to all levels of a qualitative variable and measured on a quantitative variable during each exposure. The qualitative variable is referred to as a repeated-measures factor or a within-subjects factor. The quantitative variable is called the dependent variable.

In repeated-measures analysis (also called a within-subjects analysis); scores for the same individual are dependent, whereas the scores for different individuals are independent. Accordingly, the partitioning of the variation in ANOVA needs to be adjusted so that appropriate *F* ratios can be computed. The total sums of squares (SS_T) is partitioned into three components: (1) the variation among individuals (SS_I); (2) the variation among test occasions (SS_O); and (3) the remaining variation, which is called the **residual variation** (SS_{Res}). The mean squares for these sources of variation are computed, as before, by dividing the sums of squares by their appropriate degrees of freedom. The mean square for the residual variation ($MS_{Res} = SS_{Res}/df_{Res}$) is used as the error term (the denominator of the *F* ratio) for testing the effect of test occasion, which is the effect of primary interest. It must be noted that there is no appropriate error term for testing the effect of differences among the individuals.

We seldom test the effect due to individuals in the repeated-measures design. The test occasions are the primary focus of this design. We actually have very little to gain be testing the individual effect. The main reason for obtaining the individual effect (SS_I) in the first place is to absorb the correlations between treatments and thereby remove individual differences from the error term (SS_{error}) producing a smaller MS_{Res}. A test on the individual effect, if it were significant, would merely indicate that people are different – hardly a momentous finding. Additionally, most statisticians agree that the residual error (MS_{Res}) would be an incorrect denominator in calculating an F ratio for the individual effect, and as such is typically not conducted.

To conduct a repeated-measures ANOVA in SPSS, we do not specify the repeated-measures factor and the dependent variable in the SPSS data file. Instead, the SPSS data file contains several quantitative variables. The number of quantitative variables is equal to the number of levels of the within-subjects factor. The scores on any one of these quantitative variables are the scores on the dependent variable for a single level of the within-subjects factor. Although we do not define the within-subjects factor in the SPSS data file, we specify it in the dialog box for the General Linear Model Repeated-Measures procedure. To define the factor, we give a name to the within-subjects factor, specify the number of levels of this factor, and indicate the quantitative variables in the data set associated with the levels of the within-subjects factor.

UNDERSTANDING ONE-WAY REPEATED-MEASURES ANOVA

In many studies using the one-way repeated-measures design, the levels of a within-subject factor represent multiple observations on a scale over time or under different conditions. However, for some studies, levels of a within-subjects factor may represent scores from different scales, and the focus may be on evaluating differences in means among these scales. In such a setting the scales must be commensurable for the ANOVA significance tests to be meaningful. That is, the scales must measure individuals on the same metric, and the difference scores between scales must be interpretable.

In some studies, individuals are matched on one or more variables so that individuals within a set are similar on a matching variable(s), while individuals not in the same set are dissimilar. The number of individuals within a set is equal to the number of levels of a factor. The individuals within a set are then observed under various levels of this factor. The matching process for these designs is likely to produce correlated responses on the dependent variable like those of repeated-measures designs. Consequently, the data from these studies can be analyzed as if the factor is a within-subjects factor.

SPSS conducts a standard univariate F test if the within-subjects factor has only two levels. Three types of tests are conducted if the within-subjects factor has more than two levels: the standard univariate F test, alternative univariate tests, and multivariate tests. All three types of tests evaluate the same hypothesis – the population means are equal for all levels of the factor. The choice of what test to report should be made prior to viewing the results.

> **The standard univariate ANOVA *F* test** is not recommended when the within-subjects factor has more then two levels because on of its assumptions, the sphericity assumption is commonly violated, and the ANOVA F test yields inaccurate p values to the extent that this assumption is violated.

The alternative univariate tests take into account violations of the sphericity assumption. These tests employ the same calculated *F* statistic as the standard univariate test, but its associated *p* value potentially differs. In determining the *p* value, an epsilon statistic is calculated based on the sample data to assess the degree that the sphericity assumption is violated. The numerator and denominator degrees of freedom of the standard test are multiplied by epsilon to obtain a corrected set of degrees of freedom for the tabled *F* value and to determine its *p* value.

The multivariate test does not require the assumption of sphericity. Difference scores are computed by comparing scores from different levels of the within-subjects factor. For example, for a within-subjects factor with three levels, difference scores might be computed between the first and second level and between the second and third level. The multivariate test then would evaluate whether the population means for these two sets of difference scores are simultaneously equal to *zero*. This test evaluates not only the means associated with these two sets of difference scores, but also evaluates whether the mean of the difference scores between the first and third levels of the factor is equal to zero as well as linear combinations of these difference scores.

The SPSS Repeated-Measures procedure computes the difference scores used in the analysis for us. However, these difference scores do not become part of our data file and, therefore, we may or may not be aware that the multivariate test is conducted on these difference scores. Applied statisticians tend to prefer the multivariate test to the standard or the alternative univariate test because the multivariate test and follow-up tests have a close conceptual link to each other.

If the initial hypothesis that the means are equal is rejected and there are more than two means, then follow-up tests are conducted to determine which of the means differs significantly from each other. Although more complex comparisons can be performed, most researchers choose to conduct pairwise comparisons. These comparisons may be evaluated with SPSS using the paired-samples *t* test procedure, and a Bonferroni approach or the Holm's Sequential Bonferroni procedure, can be used to control for Type I error across the multiple pairwise tests.

NULL AND ALTERNATIVE HYPOTHESIS

The null hypothesis would be that there is no difference among the treatments; that is,

H_0: $\mu_1 = \mu_2 = ... = \mu_K$

The alternative hypothesis would be that there is at least a difference between two of the occasions, that is,

H_a: $\mu_i \neq \mu_k$ for some i, k

THE RM ANOVA SUMMARY TABLE

The degrees of freedom associated with the repeated-measures design are as follows:

$$df_{\mathrm{I}} = n - 1$$

$$df_{\mathrm{O}} = K - 1$$

$$df_{\mathrm{Res}} = (K - 1)(n - 1)$$

$$df_{\mathrm{T}} = N - 1$$

The effect of interest is the test occasion and is tested using the following F ratio:

$$F_O = \frac{MS_O}{MS_{\mathrm{Res}}}$$

The ANOVA summary table for the general simple repeated-measures ANOVA is as follows:

Source	SS	df	MS	F
Individuals	SS_{I}	$n - 1$	$\frac{SS_I}{(n-1)}$	
Occasions	SS_{O}	$K - 1$	$\frac{SS_O}{(K-1)}$	$\frac{MS_O}{MS_{\mathrm{Res}}}$
Residual (Error)	SS_{Res}	$(K - 1)(n - 1)$	$\frac{SS_{\mathrm{Res}}}{(K-1)(n-1)}$	
Total	SS_{T}	$N - 1$		

Recall that, there is no appropriate error term for testing the effect of differences among the individuals. As such, there is no F ratio calculated for the individuals.

ASSUMPTIONS FOR REPEATED-MEASURES ANOVA

Several standard univariate assumptions underlie the use of simple repeated-measures ANOVA.

1. The sample was randomly selected from the population.

   The cases represent a random sample from the population, and there is no dependency in the scores between participants.

   The only type of dependency that should exist among dependent variable scores is the dependency introduced by having the same individuals produce multiple scores. Even this type of dependency introduced by the within-subjects factor is limited and must conform to the sphericity assumption. The results for a one-way, within subjects ANOVA should not be trusted if the scores between individuals are related.

2. The dependent variable is normally distributed in the population.

   The dependent variable is normally distributed in the population for each level of the within-subjects factor.

   In many applications with a moderate or larger sample size, the one-way repeated-measures ANOVA may yield reasonably accurate *p* values even when the normality assumption is violated. A commonly accepted value for a moderate sample size is 30 subjects. Larger sample sizes may be required to produce relatively valid *p* values if the population distribution is substantially non-normal. In addition, the power of this test may be reduced considerably if the population distribution is non-normal and, more specifically, thick-tailed or heavily skewed.

3. The population variances for the test occasions are equal. The population correlation coefficients between pairs of test occasion scores are equal.

   The population variance of difference scores computed between any two levels of a within-subjects factor is the same value regardless of which two levels are chosen.

   This assumption is sometimes referred to as the sphericity assumption or as the homogeneity-of-variance-of-differences assumption. The sphericity assumption is meaningful only if there are more than two levels of a within-subjects factor.

   If this assumption is violated, the *p* values associated with the standard within-subjects ANOVA cannot be trusted. However, other methods do not require the sphericity assumption. Two approaches are alternative univariate methods that correct the degrees of freedom to take into account violation of this assumption, as well as the multivariate approach that does not require the sphericity assumption.

Basically, ***sphericity*** refers to the equality of the variances of the differences between levels of the repeated measures factor. In other words, we calculate the differences between each pair of levels of the repeated measures factor and then calculate the variance of these difference scores. Sphericity requires that the variances for each set of difference scores be equal. Put simplistically, we assume that the relationship between pairs of pairs of groups is similar (i.e., the level of dependence between pairs of groups is roughly equal). When this assumption is not met, the Type I error rate can be seriously affected. However, we can make an appropriate correction by changing the degrees of freedom from $K - 1$ and $(n - 1)(K - 1)$ to 1 and $n - 1$, respectively. This change provides a conservative test of the null hypothesis.

The assumptions of repeated-measures ANOVA are similar to those for the one-way ANOVA, which include independence of observations (unless the dependent data comprise the "within-subjects" or "repeated measures" factor), normality, and homogeneity of variances. However, in addition to variances, which involve deviations from the mean of each person's score on one measure, the repeated measures design includes more than one measure for each person. Thus, covariances, which involve deviations from the mean of each of two measures for each person, also exist, and these covariances need to meet certain assumptions as well. The homogeneity assumption for repeated measures designs, known as sphericity, requires equal variances and covariances for each level of the within subjects variable. Another way of thinking about sphericity is that, if one created new variables for each pair of within subjects variable levels by

subtracting each person's score for one level of the repeated measures variable from that same person's score for the other level of the within subject variable, the variances for all these new difference scores would be equal. Unfortunately, it is rare for behavioral science data to meet the sphericity assumption, and violations of this assumption can seriously affect results. However, fortunately, there are good ways of dealing with this problem – either by adjusting the degrees of freedom or by using a multivariate approach to repeated measures. One can test for the sphericity assumption using the Mauchly's test, the Box test, the Greenhouse-Geisser test, and/or the Huynh-Feldt tests. Even though the repeated measures ANOVA is fairly robust to violations of normality, the dependent variable should be approximately normally distributed for each level of the independent variable.

MULTIVARIATE ASSUMPTIONS

The multivariate test is conducted on difference scores, and, therefore, the assumptions underlying the multivariate test concern these difference scores. The number of variables with difference scores is equal to the number of levels of the within-subjects factor minus 1. Although the difference-score variables may be computed in a number of ways, typically we will compute them by subtracting the scores associated with one level of the within-subjects factor from the scores for an adjacent level of the within-subjects factor.

1. The individual cases represent a random sample from the population, and the difference scores for any one subject are independent from the scores for any other subject.

   The test should not be used if the independence assumption is violated.

2. The difference scores are multivariately normally distributed in the population.

   If the difference scores are multivariately normally distributed, each difference score is normally distributed, ignoring the other difference scores. Also, each difference score is normally distributed at every combination of values of the other difference scores. To the extent that population distributions are not multivariate normal and sample sizes are small, the *p* values may be invalid. In addition, the power of the tests may be reduced considerably if the population distributions are not multivariate normal and, more specifically, thick-tailed or heavily skewed.

EFFECT SIZE

Pairwise Effect: $$ES = \frac{\overline{X}_i - \overline{X}_k}{\sqrt{MS_{\text{Res}}}}$$

The omnibus (overall) effect size reported for the standard univariate approach is a partial eta square (η^2) and may be calculated using the following equation:

$$\text{Partial } \eta^2_{\text{factor}} = \frac{\text{Sum of Squares}_{\text{factor}}}{\text{Sum of Squares}_{\text{factor}} + \text{Sum of Squares}_{\text{Error}}}$$

The omnibus effect size for the multivariate test associated with Wilk's lambda (Λ) is the multivariate eta square:

$$\text{Multivariate } \eta^2 = 1 - \Lambda$$

Both of these statistics (partial eta-square and multivariate eta-square) range from 0 to 1. A 0 indicates no relation between the repeated-measures factor and the dependent variable, while a 1 indicates the strongest possible relationship.

SAMPLE APA RESULTS SECTION

The mean time taken to complete the computer statistics test was 178.13 minutes on the first test; 99.63 minutes on the second test; and 88.75 minutes on the third and final test. The one-way repeated-measures ANOVA shows that these times are significantly different, $F(2, 14) = 25.83$, $p < .001$, partial $\eta^2 = .77$. Repeated-measures t-tests (using a Bonferroni adjustment, $\alpha = .05/3 = .017$) showed that subjects were significantly slower on the first test than they were on the second and third tests (test 1 versus test 2: $t(7) = 5.58$, $p < .01$, $ES = 2.89$; test 1 versus test 3: $t(7) = 5.33$, $p < .01$, $ES = 3.29$), but that there was no further reduction in completion time between the second and third tests (test 2 versus test 3: $t(7) = 1.28$, $p = .243$). It appears that practice produces an initial rapid improvement in subjects' speed of performing a task with SPSS, but that additional practice leads to little or no further improvement.

POINTS TO REMEMBER

- Rejecting the null hypothesis means that the population means for the test occasions are not equal.
- The probability that the observed differences in the means of the test occasions would have occurred by chance if the null hypothesis were true (the population means are equal) is less than the established alpha.
- When the null hypothesis is rejected, it is necessary to conduct a post hoc multiple comparison analysis in order to determine which pairs or combinations of means differ.

REFERENCES

Green, S. B., & Salkind, N. J. (2003). *Using SPSS for Windows and Macintosh: Analyzing and Understanding Data* (3rd ed.). Upper Saddle River, NJ: Prentice Hall.

Hinkle, D. E., Wiersma, W., & Jurs, S. G. (2003). *Applied Statistics for the Behavioral Sciences* (5th ed.). Boston, MA: Houghton Mifflin Company.

Howell, D. C. (2007). *Statistical Methods for Psychology* (6th ed.). Belmont, CA: Thomson Wadsworth.

Leech, N. L., Barrett, K. C., & Morgan, G. A. (2005). *SPSS for Intermediate Statistics: Use and Interpretation* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Vogt, W. P. (1999). *Dictionary of Statistics and Methodology: A Non-Technical Guide for the Social Sciences* (2nd ed.). Thousand Oaks, CA: Sage Publications.