What repeated measures analysis of variances really tells us

Younsuk Lee

¹Department of Anesthesiology and Pain Medicine, Dongguk University Ilsan Hospital, ²D/M Statistics Institute, Dongguk University, Goyang, Korea

This article examined repeated measures analysis of variance (RMANOVA). Within-subjects repeated measurements are unavoidable during clinical and experimental investigation, and between- and within-subject variability should be treated separately. Only through proper use and meticulous interpretation can ethical and scientific integrity be guaranteed. The philosophical background of, and knowledge pertaining to, RMANOVA are described in the first half of this text. The sphericity assumption and associated issues are discussed in the latter half. The final section provides a summary measure analysis, which was neglected by P value-dependent interpreters.

Key Words: Data interpretation, Repeated measurements, Sphericity condition.

Introduction

Readers frequently encounter repeated measures analysis of variance (RMANOVA) when browsing the medical literature. In the field of anesthesiology, we measure blood pressure, cardiac outputs, and pain scores repeatedly at different time intervals. We can also measure blood pressure at different sites: radial and femoral, and right and left. Although RMANOVA represents a major analytical method for repeated measures (RM) data, it is frequently misused or misinterpreted due its complexity. Several articles have been published in medical journals focusing on the analysis of RM data [1,2]. However, despite the quality of these reports, readers of the Korean Journal of Anesthesiology (KJA),

as well as potential authors, remain uncertain with regard to understanding and practically applying RMANOVA. This article focuses on three learning objectives: (1) the pitfalls of erroneously applying simple analysis of variance (ANOVA) to RM data instead of RMANOVA; (2) the obligatory sphericity assumption of RMANOVA, including adjustments and workarounds; and (3) summary measures analysis.

New readers may find it difficult to understand the statistical jargon employed in this article; therefore, several of the key technical terms and abbreviations are defined and listed presently:

- RM data is that in which two or more observations are made within an experimental unit. In the KJA, an experimental unit typically refers to a single human or animal subject. Repeated observations can occur temporally or spatially. Longitudinal data represents a special form of RM data, in which repeated observations are made over long period of time.
- RMANOVA is a distinct type of ANOVA associated with within-subject variability. Some statisticians use RMANOVA instead of univariate ANOVA to assess subject effects. In that context, RMANOVA can be considered a univariate rather than multivariate approach.
- The sum of squares (SS), which measures the variability (uncertainty, error) of data, is calculated as the sum of the squares of the distances between each observation and the mean.

Corresponding author: Younsuk Lee, M.D., Ph.D.
Department of Anesthesiology and Pain Medicine, Dongguk University Ilsan Hospital, D/M Statistics Institute, Dongguk University, 27, Dongguk-ro, Ilsandong-gu, Goyang 410-773, Korea
Tel: 82-31-961-7872, Fax: 82-31-961-7864
E-mail: ylee@dongguk.edu
ORCID: <http://orcid.org/0000-0003-2488-5926>

Received: May 18, 2015.
Revised: June 22, 2015.
Accepted: June 30, 2015.

Korean J Anesthesiol 2015 August 68(4): 340-345
<http://dx.doi.org/10.4097/kjae.2015.68.4.340>

© This is an open-access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

Copyright © the Korean Society of Anesthesiologists, 2015

Online access in <http://ekja.org>

- $SS_{\text{something}}$ denotes the variability explained by something known: e.g., SS_{time} , SS_{group} , and SS_{subject} . The total sum of squares, SS_{total} , is the sum of all the SS components of a dataset. If we have groups A, B, and C, then the notations can be simplified as SS_A , SS_B , and SS_C .
- Readers should be aware that certain statistical reports use another convention, i.e., “something-SS” or “SS-something”, which may be denoted as SST or TSS instead of SS_{total} .
- Mean squares (MS) indicates the average of the SS. MS is estimated by dividing SS by the degrees of freedom (d.f.). The ratio of each MS per MS_{error} is called the F value.
- $Y \sim X$ denotes that “Y is modeled as X” (according to the convention of Wilkinson and Rogers, 1973) [3], which is equivalent to “Y is explained by X”. When the right-hand side of the equation is empty, $Y \sim 1$ equates to “Y is modeled as an intercept,” or “Y is explained by nothing,” which accords with the null hypothesis.
- $A : B$ denotes the interaction between conditions A and B.

By reading this article, readers will learn typical conventions useful for interpreting full-length statistical reports; the information contained herein should act as a bridge toward understand-

ing complex theory. All statistics were estimated using the R: A Language and Environment for Statistical Computing (ver. 3.2.0; R Foundation for Statistical Computing, Vienna, Austria). An additional library “car” (An R Companion to Applied Regression, 2nd Edition; J. Fox and S. Weisberg) was used for Mauchly’s test. The complete computational procedures undertaken are attached in the appendices in R script format. The datasets introduced herein are real but have been modified slightly to aid understanding.

Major Differences between ANOVA and RMANOVA

A total of 16 boys and 11 girls were enrolled in a study conducted at a university dental hospital in North Carolina. Radiographic distances (mm) between the pituitary and pterygomaxillary fissure were measured repeatedly for each subject, at 8, 10, 12, and 14 years of age [4]. For simplicity, the girls’ data are focused on herein, and are referred to as the “girls dataset” (Table 1).

Total Uncertainty Explained by No Factors

The initial estimation begins with a null hypothesis, e.g., “the dental measurements (in girls) were totally unexplainable,” or “the dental measurements (in girls) were explained by no factors.” Total SS (SS_T) = the SS of the error (SS_E) and is computed by:

$$\begin{aligned} SS_T &= (X_{1,1} - \bar{X})^2 + (X_{1,2} - \bar{X})^2 + (X_{1,3} - \bar{X})^2 + \dots \\ &= (21 - 22.6)^2 + (20 - 22.6)^2 + (22 - 22.6)^2 + (23 - 22.6)^2 + \dots \\ &= 247.29 \end{aligned}$$

$X_{i,j}$ denotes the distance on the j^{th} occasion in the i^{th} subject and $\bar{X}$ denotes the mean distance. SS_T is also computed using a simple ANOVA table that includes “nothing” as an explanatory variable (Table 2a).

ANOVA Model of the Effect of Age

It is intuitive to hypothesize that dental distances increase with age. The effect of age can be estimated and is denoted by

Table 1. Dental Measurements (mm) in the “Girls Dataset” (n = 11)

Subject	Age 8	Age 10	Age 12	Age 14
F01	21.0	20.0	21.5	23.0
F02	21.0	21.5	24.0	25.5
F03	20.5	24.0	24.5	26.0
F04	23.5	24.5	25.0	26.5
F05	21.5	23.0	22.5	23.5
F06	20.0	21.0	21.0	22.5
F07	21.5	22.5	23.0	25.0
F08	23.0	23.0	23.5	24.0
F09	20.0	21.0	22.0	21.5
F10	16.5	19.0	19.0	19.5
F11	24.5	25.0	28.0	28.0

The data for the 11 girls were retrieved from the full dataset (16 boys and 11 girls), originally introduced by Potthoff RF, Roy S. A generalized multivariate analysis of variance model useful especially for growth curve problems. *Biometrika*. 1964;51:313-26. With permission from Oxford University Press (3660581189264).

Table 2. Sum of Squares for Two ANOVA Models of the “Girls Dataset” (n = 11)

	d.f.	SS	MS	F	P value
(a) Null model. “The distances are explained by no factors.”					
Error	43	247.3	5.751	-	-
(b) ANOVA model of the effect of age. “The distances are explained by the effect of age” (misleading)					
Age	3	50.65	16.884	3.435	0.0258
Error	40	196.94	4.683	NA	-

ANOVA: analysis of variance, d.f.: degrees of freedom; SS: sum of squares, MS: mean squares.

SS_{age} ; this approach would be incorrect unless treated as a within-subjects effect. In this model, SS_T is given by the sum of SS_{age} = 50.65 and SS_E = 196.7, such that SS_T = 247.29 of the null model (Table 2b).

ANOVA Model of the Effects of Age, Gender, and Their Interactive Effect

Similar to the girls dataset, in the full dental measurements dataset, SS_T is given by the sum of SS_{age} , SS_{gender} , and $SS_{age:gender}$ (Table 3a). The models estimated thus far all exclude the effect of subject. Because the measurements for each subject were repeated four times, the SS values should have comprised $SS_{within-subject}$ and $SS_{between-subject}$. Statistics are not correct here for effects that are repeated within-subjects, such as age and the age : gender interaction. The value of SS_T = 917.7 after summing all of the SS components.

RMANOVA Model of the Effects of Age, Gender, and Their Interactive Effect

We will now discuss RMANOVA (Table 3b). The ANOVA table divides sources of variability into two categories: within- and between-subjects. The effects of age (SS_{age} = 237.19), the age : gender interaction ($SS_{age:gender}$ = 13.99), and its error term (SS_W = 148.13) comprise the within-subject variability ($SS_{within-subject}$). The effects of gender (SS_{gender} = 140.5) and its error term ($SS_{between}$ = 377.9) comprise the between-subject variability ($SS_{between-subject}$).

$$SS_T = SS_{within-subject} + SS_{between-subject} = 399.31 + 518.4 = 917.7$$

$$SS_{within-subject} = SS_{age} + SS_{age:gender} + SS_W \\ = 237.19 + 13.99 + 148.13 = 399.31$$

$$SS_{between-subject} = SS_{gender} + SS_{between} = 140.5 + 377.9 = 518.4$$

Perceptive readers may note that the absolute SS values equate to those of the simple ANOVA models described in the previous section; the resulting value of SS_T is always constant within a dataset. Changes in F values affect the calculation of P values. In the final RMANOVA model, the result of this is that the P values are either lower or higher than those listed in Table 3A, which indicates that, if RMANOVA is not used, a simple ANOVA will inflate Type I error (false-positives) in between-subject effects and Type II error (false-negative decision) in within-subject effects. A graphical approach may aid the reader in understanding the concept that total variability is comprised of several different sources of variability, denoted by the areas of the rectangles (SS; Fig. 1).

Sphericity Assumption

In simple terms, the variances of the differences between all combinations of measurements should be equal when using univariate RMANOVA. This is referred to as the sphericity (or circular) assumption. Sphericity, of the RM data of the covariance matrix, is strongly assumed for within-subject RMANOVA statistics. In cases that violate the sphericity assumption, within-subject RMANOVA statistics are meaningless. Given its name, i.e., "sphericity," readers may expect to encounter a relatively complicated algebraic concept, such that plain English is used to aid understanding in the discussion below.

Violations of sphericity may be evaluated using the sphericity test developed by Mauchly, which can be performed easily, or even automatically, in the majority of statistical software packages. Mauchly's test with a $P > 0.05$ (or 0.10 depending on your *a priori* assessment of the data) allows us to interpret the results of RMANOVA. Returning to the girls dataset, six pairwise differences were calculated (Table 4): 10–8, 12–8, ... , 14–12. The

Table 3. ANOVA Tables for The Full Dental Measurements Dataset (n = 27)

	d.f.	SS	MS	F	P value
(a) ANOVA model of the effects of age, gender, and their interactive effect (misleading)					
Age	3	237.19	79.06	15.030	3.79e-08
Gender	1	140.5	140.5	26.702	1.22e-06
Age : gender	3	14.0	4.66	0.887	0.451
Error	100	526.0	5.26	-	-
(b) RMANOVA model of the effects of age, gender, and their interactive effect (within-subject variability is estimated)					
Within-subjects					
Age	3	237.19	79.06	40.032	1.49e-15
Age : gender	3	13.99	4.66	2.362	0.0781
Error	75	148.13	1.98	-	-
Between-subjects					
Gender	1	140.5	140.5	9.292	0.005375
Error	25	377.9	15.12	-	-

ANOVA: analysis of variance, RMANOVA: repeated measures analysis of variance; d.f.: degrees of freedom, SS: sum of squares, MS: mean squares.

Fig. 1. Graphical representation of the concept of analysis of variance (ANOVA). The designated variabilities reduce total variability, and the areas of the rectangles denote the amount of variability. (A) ANOVA model of the effects of age, gender, and their interactive effect. (B) Repeated measures ANOVA model of the effects of age, gender, and their interactive effect. The effects of age, and the age : gender interaction, are estimated within-subjects.

Table 4. Pairwise Differences in the “Girls Dataset” (n = 11)

	10–8	12–8	14–8	12–10	14–10	14–12
F01	–1.0	0.5	2.0	1.5	3.0	1.5
F02	0.5	3.0	4.5	2.5	4.0	1.5
F03	3.5	4.0	5.5	0.5	2.0	1.5
F04	1.0	1.5	3.0	0.5	2.0	1.5
F05	1.5	1.0	2.0	–0.5	0.5	1.0
F06	1.0	1.0	2.5	0.0	1.5	1.5
F07	1.0	1.5	3.5	0.5	2.5	2.0
F08	0.0	0.5	1.0	0.5	1.0	0.5
F09	1.0	2.0	1.5	1.0	0.5	–0.5
F10	2.5	2.5	3.0	0.0	0.5	0.5
F11	0.5	3.5	3.5	3.0	3.0	0.0
Variance	1.4	1.4	1.7	1.2	1.4	0.6

variances of the pairwise differences ranged from 0.60 to 1.74, which appears relatively wide; however, the Mauchly statistic (W) = 0.69, and the estimated P = 0.67, indicating that the girls dataset satisfies the sphericity assumption. A favorable result was

Table 5. Mauchly's Test of Sphericity for the “Girls Dataset” (n = 11)

	Mauchly's test statistic W	P value
Original data	0.69474	0.6746
Modified data*	0.14898	0.0056

*Produced by multiplying the distances at 12 years of age by 2 (arbitrarily).

expected for this dataset because there was no reasonable basis on which to assume the presence of another factor aside from age over the 2-year periods.

To enhance the reader's understanding of the concept of sphericity, the girls dataset was modified arbitrarily by multiplying the values obtained at 12 years of age by 2. Therefore, the Mauchly statistic W = 0.15, and P = 0.006, which proves that the dataset violates the assumption. This arbitrary modification illustrates the relative rigidity of the sphericity assumption (Table 5). Because conditions between the repeated measurements should be uniform, we cannot anticipate that the assumption will be satisfied, especially when two or more conditions with brief intervals

are added to a single RM dataset (e.g., administration of drugs and attempted endotracheal intubation). Such designs represent a substantial proportion of typical anesthesiology study designs.

Several “quick-and-dirty” adjustment procedures are available for RM data that violate the sphericity assumption, known as sphericity adjustments. Software packages usually provide factors (ϵ , epsilon) that adjust for degrees of freedom (d.f.) with respect to within-subject RMANOVA statistics. These include the Greenhouse-Geisser ($\hat{\epsilon}$) and Huynh-Feldt ($\tilde{\epsilon}$) adjustment factors. By definition, the true ϵ values = 1, such that the sphericity assumption is fully satisfied. In the modified girls dataset described above, the Greenhouse-Geisser value $\hat{\epsilon}$ was estimated at 0.47, and the Huynh-Feldt $\tilde{\epsilon}$ = 0.53. The effect of age has d.f. values of 3 (numerator) and 75 (denominator), such that the Huynh-Feldt adjusted d.f. values were as follows:

$$\begin{aligned}\text{d.f. (numerator)} &= 3 \times 0.53 = 1.05 \\ \text{d.f. (denominator)} &= 75 \times 0.53 = 39.75\end{aligned}$$

Workarounds for RMANOVA

If the repetition has a single factor (e.g., only the time-based repetition), the calculation and interpretation of Mauchly's statistic would be easier. However, if there are more than two repetition factors, or they are nested, such calculations are rendered more difficult. Statisticians use two distinct methods to work around any violation of sphericity: multivariate analysis of variance (MANOVA) and mixed-effect modeling (MEM). Although MANOVA and MEM require more statistical knowledge, MANOVA is highly resistant to the violation of any assumption during the analysis of RM data, and MEM is a highly flexible method that uses user-defined variance structures; therefore, researchers should be familiar with both methods. We must also be aware that the editors of one international anesthesiol-

Table 6. Results of a T-Test Applied to Summary Measures of Dental Distance in 27 Children

	Boys	Girls	P value
Mean distance (mm)	25.0 (1.8)	22.6 (2.1)	0.01
Maximum distance (mm)	27.8 (2.2)	24.1 (2.4)	0.00

Data are presented as means (SD).

ogy journal recommend MEM as the method of choice for the analysis of RM data [5]. The use of MEM should be confined to studies in which an effect of subject represents the primary concern [6].

Summary Measure Analysis

Because the statistics-heavy results and numerous P values generated by RMANOVA often confuse researchers, they sometimes fail to notice straightforward values within their data. Everitt and Rabe-Hesketh (2001) [7], and Frison and Pocock (1992) [8], suggested that researchers should extract more direct values from RM data, such as the overall mean, maximum (minimum) value, time to maximum (minimum) response, regression slope, and time to reach a particular value. Despite a lack of consensus regarding a gold standard summary measure, after identifying data-by-data analysis becomes more straightforward, such that a t-test or simple ANOVA can be applied. In our full dental dataset, the individual mean distances and maximum distances can be calculated readily and compared between genders using a t-test (Table 6).

Declaration of Interest

Y.L. is the Editor-in-Chief of the KJA.

References

1. Park E, Cho M, Ki CS. Correct Use of Repeated Measures Analysis of Variance. *Korean J Lab Med* 2009; 29: 1-9.
2. Kim HY. Statistical notes for clinical researchers: A one-way repeated measures ANOVA for data with repeated observations. *Restor Dent Endod* 2015; 40: 91-5.
3. Wilkinson GN, Rogers CE. Symbolic description of factorial models for analysis of variance. *Applied Statistics* 1973; 22: 392-9.
4. Potthoff RF, Roy S. A generalized multivariate analysis of variance model useful especially for growth curve problems. *Biometrika* 1964; 51: 313-26.
5. Dexter F. Checklist for Statistical Topics in Anesthesia & Analgesia Reviews. *Anesth Analg* 2011; 113: 216-9.
6. Ma Y, Mazumdar M, Memtsoudis SG. Beyond Repeated measures ANOVA: advanced statistical methods for the analysis of longitudinal data in anesthesia research. *Reg Anesth Pain Med* 2012; 37: 99-105.
7. Everitt B, Rabe-Hesketh S. *Analyzing medical data using S-PLUS*. New York, Springer Verlag. 2001. pp. 149-51.
8. Frison L, Pocock SJ. Repeated measures in clinical trials: Analysis using mean summary statistics and its implications for design. *Statist Med* 1992; 11: 1685-704.

Appendix

The following R code was created on 13-May-2015 and finally verified on 21-Jun-2015 on a 64-bit x86 Darwin platform.

The library nlme is installed to read dataset Orthodont.
The library car is installed to perform Mauchly's test.

```
library(nlme)
data(Orthodont)
Orthodont$age <- factor(Orthodont$age)
```

the girls dataset from the full dental dataset

```
OrthodontF <- Orthodont[Orthodont$Sex == "Female", c("Subject", "age", "distance")]
distanceMat <- matrix(OrthodontF$distance, ncol=4, byrow=T)
rownames(distanceMat) <- c("F01", "F02", "F03", "F04", "F05", "F06", "F07", "F08", "F09", "F10", "F11")
colnames(distanceMat) <- c("Age8", "Age10", "Age12", "Age14")
print(distanceMat) # Generating Table 1
```

Extract 11 girls data

```
girlsDental <- Orthodont[Orthodont$Sex == "Female", ]
```

Models

```
girls0.aov <- aov(distance ~ 1, girlsDental)
summary(girls0.aov) # Table 2a
girlsWithAge.aov <- aov(distance ~ age, girlsDental)
summary(girlsWithAge.aov) # Table 2b
colnames(Orthodont)[4] <- "gender"
orth.aov <- aov(distance ~ age * gender, Orthodont)
summary(orth.aov) # Table 3a
orthW.aov <- aov(distance ~ age * gender + Error(Subject), Orthodont)
summary(orthW.aov$Within) # Table 3b upper
summary(orthW.aov$Subject) # Table 3b lower
```

Pairwise difference table

```
diffMat <- cbind(distanceMat[, "Age10"] - distanceMat[, "Age8"],
  distanceMat[, "Age12"] - distanceMat[, "Age8"],
  distanceMat[, "Age14"] - distanceMat[, "Age8"],
  distanceMat[, "Age12"] - distanceMat[, "Age10"],
  distanceMat[, "Age14"] - distanceMat[, "Age10"],
  distanceMat[, "Age14"] - distanceMat[, "Age12"])
colnames(diffMat) <- c("10-8", "12-8", "14-8", "12-10", "14-10", "14-12")
diffAndVar <- rbind(diffMat, apply(diffMat, 2, var))
rownames(diffAndVar)[nrow(diffAndVar)] <- "Variations"
print(diffAndVar) # Table 4
```

Mauchly's test

```
library(car) # for Mauchly's test
rfactor <- factor(c(8, 10, 12, 14))
girls.mlm <- lm(distanceMat ~ 1)
girls.aov <- Anova(girls.mlm, idata = data.frame(rfactor),
  idesign = ~ rfactor, type="III")
summary(girls.aov, multivariate = F) # Table 5
```

Summary Measures

```
distances <- matrix(Orthodont$distance, ncol=4, byrow=T,
  dimnames = list(gender = c(rep("boy", 16), rep("girl", 11)),
  age = c(8, 10, 12, 14)))
means <- apply(distances, 1, mean)
maxes <- apply(distances, 1, max)
t.test(means ~ rownames(distances)) # Table 6
t.test(maxes ~ rownames(distances)) # Table 6
```