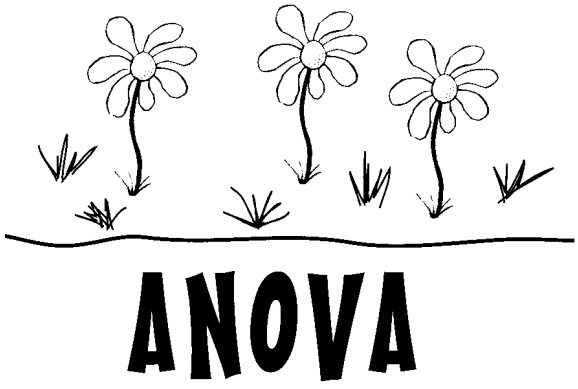
Repeated Measures ANOVA, R.I.P.?



Charles E. McCulloch

It is a difficult experiment to run and to analyze: What are the effects of alcohol on sleepiness and does a hormone, pregnenolone, which has been shown to enhance memory in rat experiments, help alleviate the sleepiness? Each person is tested under each of four conditions on four different visits in random order: a placebo for the drug and for the hormone, alcohol alone, hormone alone, and the combination. Each subject is also queried multiple times within a visit in the minutes after alcohol (or placebo) ingestion. Some subjects drop out of the protocol without completing all the conditions and some of the sleepiness scores are not recorded within a visit because of difficulties executing the protocol. How should the data be analyzed?

This is an opportunity for the professional statistician to pull out any of a number of impressive and more recent tools of the trade: generalized estimating equations, mixed model analyses, imputation, and inverse probability weighting. Gone are the Huynh-Feldt or Geisser-Greenhouse corrections, expected mean squares, figuring out the "correct" error term, and filling in missing data. Or are they? The investigator, after being led through the results of a SAS Proc MIXED analysis bemoans, "Can't you run a repeated measures analysis of variance? That is all I can really understand." Are the new methods overkill, or do they offer significant advantages?

Analysis of Variance

Analysis of variance (ANOVA) has a long history dating back to R.A. Fisher (1925). The key idea is to divide up the variability in a data set into interpretable components. Consider a simplification of the above scenario: We record Y_{ijk} = the average sleepiness score between 60 and 120 minutes after administration of alcohol or placebo for subject $i(= 1, 2, \ldots, n)$ under alcohol condition $j(= 1, 2)$ and pregnenolone condition $k(= 1, 2)$. One of the simplest models we might entertain for such a situation would be

$$Y_{ijk} = \mu + \alpha_j + \beta_k + \alpha\beta_{jk} + p_i + \varepsilon_{ijk}, \qquad (1)$$

where μ is the overall mean, α_j represents the alcohol effect, β_k the pregnenolone effect, $\alpha\beta_{jk}$ the interaction effect, p_i is the person effect, and ε_{ijk} is an error term. This model hypothesizes simple person effects that raise or lower (if the effect is negative) the average sleepiness in all four conditions. Interest focuses on the interaction, because the scientific question is whether pregnenolone helps to reduce the sleep-inducing effect of alcohol.

With a mean, error term and four explanatory factors in the model, the analysis of variance would partition the variability in Y_{ijk} into four sources: person, alcohol, pregnenolone, and the interaction. An ANOVA would generate a table as outlined in Table 1. A typical assumption is that the ε_{ijk} are all independent and follow a $N(0, \sigma^2)$ distribution and that the rest of the terms in (1) are fixed. If these assumptions are correct, the F-statistic of interest, namely F(Int)= MS(Int)/MS(Err), has an exact F-distribution with 1 and $3(n-1)$ degrees of freedom. So this forms a straightforward significance test of the null hypothesis of no interaction, with p-value given by $P\{F^1_{3(n-1)} \geq \mathrm{F(Int)}\}$, with F representing an F-distributed random variable with the given degrees of freedom.

When every subject is tested under each of the conditions and there are no missing data, the calculations leading to Table 1, though somewhat tedious, can be performed with a hand calculator. There is a certain tidiness in showing how all the variability in the data is divided up into pieces, no matter what order factors are entered into the model, the additional amount attributable to the model upon entering a new explanatory factor is given by the sum of squares (SS) in Table 1 for that factor. Also, the sums of squares for the various factors add up to SS(Tot), the "total" sum of squares.

When there is unbalancedness to the data in the sense that not every subject has data for each condition, then the calculations are much more difficult and essentially require

Table 1— Analysis of Variance for Model (1)

Source	Sums of Squares	d.f.	Mean Square	F-statistic
Person	SS(Person)	n – 1	MS(Person)	MS(Person)/MS(Err)
Alcohol	SS(Alch)	1	MS(Alch)	MS(Alch)/MS(Err)
Pregnenolone	SS(Preg)	1	MS(Preg)	MS(Preg)/MS(Err)
Interaction	SS(Int)	1	MS(Int)	F(Int)= MS(Int)/MS(Err)
Error	SS(Err)	3(n – 1)	MS(Err)	
Total	SS(Tot)	4n – 1		

a computer. The F-statistic of interest still has an exact F-distribution, but two other features are lost. First, division of the total variability is not so neat: The increase in variance explained when adding a term to the model depends on which terms have already been included in the model. Second, the exact hypothesis being tested by the interaction F-statistic is difficult to express in easy-to-understand terms (Searle, 1970).

Repeated Measures ANOVA

Even with balanced data another feature of (1) is unattractive. One would typically argue that the multiple measurements taken on the same person should be modeled as correlated, but if all the quantities excepting the ε_{ijk} are fixed, then there is no correlation between observations taken on the same person. To illustrate this, under model (1) the covariance between two different observations taken on the ith person is given by

$$\begin{aligned}\operatorname{cov}(Y_{ijk}, Y_{ij'k'}) &= \operatorname{cov}(\mu + \alpha_j + \beta_k + \alpha\beta_{jk} + p_i + \varepsilon_{ijk}, \\ &\quad \mu + \alpha_{j'} + \beta_{k'} + \alpha\beta_{j'k'} + p_i + \varepsilon_{ij'k'}) \\ &= \operatorname{cov}(\varepsilon_{ijk}, \varepsilon_{ij'k'}) \\ &= 0 \text{ for } j \neq j' \text{ or } k \neq k'.\end{aligned}$$

The next-to-last equality follows because the fixed terms do not contribute covariance, and the last equality is true because the ε_{ijk} are assumed to be independent.

A related idea is that the typical scientific experiment aims to present conclusions that apply more generally than to just the subjects included in the experiment. While subjects are rarely a true random sample from a larger population of subjects, they are often selected so that some generalization to a larger population is possible. In other words, they can be regarded as a random sample from some reasonable population of subjects. If this is the case, then the effects associated with a person, p_i, can also be regarded as a random sample from the collection of subject effects associated with that reasonable population of subjects. In this case, we would call the person effect a random effect and the model would be termed a mixed model because it has both fixed and random effects.

Technically we would quantify this as saying that the p_i are independent and identically distributed with some variance, denoted by σ_p^2, and (without loss of generality) mean zero. This additional assumption does induce a correlation between measurements taken on the same person:

$$\begin{aligned}\operatorname{cov}(Y_{ijk}, Y_{ij'k'}) &= \operatorname{cov}(\mu + \alpha_j + \beta_k + \alpha\beta_{jk} + p_i + \varepsilon_{ijk}, \\ &\quad \mu + \alpha_{j'} + \beta_{k'} + \alpha\beta_{j'k'} + p_i + \varepsilon_{ij'k'}) \\ &= \operatorname{cov}(p_i + \varepsilon_{ijk}, p_i + \varepsilon_{ij'k'}) \\ &= \sigma_p^2 \text{ for } j \neq j' \text{ or } k \neq k'.\end{aligned}$$

Similar calculations show that the variance of Y_{ijk} is given by $\sigma_p^2 + \sigma^2$.

So the simple act of assuming that the subjects in the experiment are a sample from a larger population of subjects, and coupling it with model (1) so as to make inferences to that population of people, generates an equal correlation among all observations (four in our example) between the observations taken on the same person, where the correlation is given by

$$\begin{aligned}\operatorname{corr}(Y_{ijk}, Y_{ij'k'}) &= \frac{\operatorname{cov}(Y_{ijk}, Y_{ij'k'})}{\mathrm{SD}(Y_{ijk})\mathrm{SD}(Y_{ij'k'})} \\ &= \frac{\sigma_p^2}{\sqrt{(\sigma_p^2 + \sigma^2)(\sigma_p^2 + \sigma^2)}} \\ &= \frac{\sigma_p^2}{(\sigma_p^2 + \sigma^2)}\end{aligned}$$

With this change to the model, with its induced correlation, how should the data be analyzed? One strategy is to start with the ANOVA given in Table 1 and ask what modifications are necessary. When the data are balanced the F-statistics for testing the alcohol, pregnenolone or interaction terms still have exact F-distributions, perhaps surprisingly.

However, when the data are unbalanced, the F-statistic no longer follows an exact F-distribution; in fact, the approximation can be quite poor (Tietjen, 1974). One strategy to get around such a situation would be simply to evaluate the empirical distribution of the F-statistic by a technique such as the bootstrap. Then the issue would be efficiency, which I touch on below.

Another issue is whether the simple equal correlation case is adequate to describe the 4×4 correlation matrix of sleepiness score across the four conditions within a person. If the correlations are not all equal, then, even when the data are balanced, the distributions of the F-statistics are no longer exactly F-distributions. To improve the agreement with the reference F-distribution, various correction factors have been suggested, such as the Geisser-Greenhouse correction (Greenhouse and Geisser 1957) and the Huynh-Feldt correction (Huynh and Feldt 1976).

Another workaround when the repeated measurements on a subject do not follow an equal correlation structure is to analyze the data as multivariate data (in our case, regarding the four observations on a subject as a multivariate response vector), and make no assumptions about the form of the correlation matrix. However, it can be wasteful to estimate all the elements of a variance covariance matrix if, in fact, the equal variance, equal correlation assumption is a good one. For example, with four repeated measures there would be the four variances and six covariances as opposed to the single variance and covariance in the simpler structure. With four repeated measurements this might not be a bad tradeoff to gain freedom from assumptions, but it would be a much poorer choice with many more repeated measurements. Also, typical analysis software drops all the observations on a person when any one of them is missing. This is both wasteful and can give misleading results if those with some missing data are different from those with complete data (which seems almost always to be the case in practice).

The situation gets worse in complicated problems with multiple random factors, such as a split plot experiment (Federer 1955). Legions of books (for example, Searle 1970) give rules for calculating expected mean squares for help in figuring out "proper" error terms to form the denominator of the F-statistic. With some complicated balanced designs and most unbalanced designs, it can be difficult or impossible to figure out a reasonable test statistic.

Maximum Likelihood Estimation of Mixed Models

Suppose we are willing to assume that the p_i are normally distributed. If we interpret equation (1) as the conditional distribution of the data given the p_i, then the distributional model for the data is completely specified. With the p_i normally distributed and the conditional distribution normal, the marginal distribution of the data is multivariate normal, with the correlation structure being given by (4), (see, for example Searle, Casella, and McCulloch 1992). We can then write down the likelihood and perform maximum likelihood estimation.

In general it is not possible to write the maximum likelihood estimators in closed form, but with modern computational capabilities it is generally not too difficult to evaluate them numerically. This is typically performed by an iterative algorithm that starts with reasonable guesses and improves them until no further improvement is seen. Inferences are based on the large sample properties of the maximum likelihood estimators, although some software has incorporated improvements to the small sample inferences. For example, the Kenward-Roger (1997) improvement on the approximation to the distribution of F-like statistics is available in SAS (Proc MIXED, SAS Ver. 8 and later, SAS Institute, Cary, NC). So testing the null hypothesis of no interaction, as is of interest in the alcohol and pregnenolone example, is straightforward using the programs.

Maximum likelihood methods inherently are based on sufficient statistics and use the full amount of information available in the data, whereas ANOVA-based methods may work from statistics that engender some loss of information (Scheffe, 1959).

Commercial implementations of such models typically allow very flexible specification of the variance covariance structure in the data, handle arbitrarily unbalanced data with aplomb, and protect against some forms of bias due to missing data (Laird 1988).

The maximum likelihood methods have further advantages in more easily allowing inference about the variances and covariances and generating minimum mean-squared error-predicted values of the random effects. In some situations the primary focus is on the correlation; for example, in Fisher et al. (2002) the primary focus was on whether a form of high blood pressure tended to be aggregated in families, suggesting a genetic component to the condition. In other situations, the primary focus is on predicting the random effects. For example, Austin et al. (2003) describe 7 models for predicting mortality across a number of hospitals. A random effect is placed in the model to capture the mortality rate associated with an individual hospital. Interest focuses on identifying those hospitals with extremely high or low hospital effects, corresponding to ones (after adjustment for other covariates, such as age and health of the patients) that have especially high or low mortality rates.

I performed a simulation experiment to demonstrate some of these points. A dataset of $n = 20$ subjects was simulated from model (1) with complete, balanced data: Each subject had four observations, one from each condition; 40% of that design was selected for elimination to create an unbalanced dataset from the balanced one. Then 20,000 replications were performed for both the balanced and unbalanced versions of each dataset under each of four scenarios for the interaction term: twice in the null hypothesis case where the interaction was zero and one each where the interaction effects were size two and three. The error variance was set to one, and the person-to-person variance, σ_p^2, was set to two, generating a within person correlation of 2/3. The simulation was conducted in Stata Version 8.0 (Stata Corp., College Station, TX) using its ANOVA and xtreg (mle) commands.

The first 20,000 replications under the null hypothesis case were used to calibrate the size of the test in case the distribution was not exactly an F-distribution (from the repeated measures ANOVA) or Z-distribution (from the maximum likelihood analysis). The calibrated cutoffs for statistical significance were actually close to that predicted by theory. Using those cutoffs, interaction effects of size 0, 2, and 3 were simulated to gauge the standard errors of the estimates and the power of the tests. The simulation error of the average estimate is less than 0.01 in each case, and the simulation error of the power (or size) is less than 0.004 in each case.

Table 2—Simulation Comparison of Repeated Measures ANOVA (RMA) and Maximum Likelihood Estimation (MLE) for Model (1) with n = 20

Data Layout		Interaction				
	Method	True	Ave Est	SD	Ave SE	Prob Reject
Balanced Data						
	RMA	0	0.0043	1.00	0.96	0.050
	MLE	0	0.0043	1.00	0.95	0.050
	RMA	2	2.001	1.00	1.00	0.50
	MLE	2	2.001	1.00	0.95	0.50
	RMA	3	3.000	1.00	0.99	0.84
	MLE	3	3.000	1.00	0.95	0.84
Unbalanced Data						
	RMA	0	-0.0055	1.38	1.38	0.050
	MLE	0	-0.0049	1.30	1.21	0.048
	RMA	2	2.00	1.39	1.38	0.29
	MLE	2	2.00	1.30	1.21	0.32
	RMA	3	3.00	1.39	1.37	0.55
	MLE	3	3.00	1.30	1.21	0.62

A total of 20,000 replications were performed for each true value of the interaction. Simulation errors for the average of the estimate (Ave Est) are 0.01 or less in each case and the simulation error for the power (or size) estimates are 0.004 or less in each case. The table also reports the standard deviation (SD) of the estimator across the 20,000 replications, the average of the model-based standard errors (Ave SE) and the probability of rejecting the hypothesis of interaction equal to zero (Prob Reject).

Table 2 reports the results of the simulation. The top half of the table reports the balanced data results. In that case, the performance of the estimators (as judged by the average, standard deviation, size, or power) is virtually identical for the repeated measures ANOVA (RMA) and the maximum likelihood fit to the mixed model (MLE).

With unbalanced data there are some differences. The standard deviation of the MLE is somewhat smaller, and the power somewhat higher than the RMA method, reflective of the fact that the MLE uses all the information in the data. Of note also is that the average of the model-based standard errors for the MLE is somewhat off compared to its actual standard deviation. This is true in both the balanced and unbalanced data situations, which is not surprising given the small number of subjects and the fact that the standard errors for the MLEs are based on large sample approximations.

Comparisons

What are the advantages and disadvantages of each approach? Repeated measures ANOVA has the advantage of being less computationally intensive and (at least for those familiar with it) more understandable. When the data are balanced, the F-statistics often follow an exact *F*-distribution. On the down side, it estimates variances and covariances by the method of moments, which is usually less efficient than maximum likelihood and makes it more difficult to make inferences about the correlations in the data. When the data are not balanced, tests are approximate (often to an unknown degree) and may be difficult to specify correctly.

Maximum likelihood estimation of mixed models has the advantages of automatic generation of test statistics, natural handling of inference about the correlation, a wide array of available correlation structures, and gives better predictions of the random effects. Maximum likelihood methods also generalize naturally to non-normally distributed outcomes (see, for example, McCulloch and Searle 2000), unlike repeated measures ANOVA. Its disadvantages include the need for iterative computation, which can sometimes fail to converge (invariably with a cryptic error message!), dependence on large sample approximations, and less familiarity to those who were raised on ANOVA.

Maximum likelihood routines for correlated data, linear models are widely available in commercial and open-source software, including SAS (Proc MIXED), SPSS (Linear Mixed Model), Stata (new in Version 9), and S-Plus or R (lme).

Conclusions

If the data are from a simple design and very close to balanced, it is nice to be able to use simpler techniques that have exact distributions under exact balance and exact normality. In such a case, repeated measures ANOVA is the approach of choice. However, my opinion is that, in the vast majority of cases, the newer mixed model algorithms are by far the better choices. These methods allow full and efficient use of the data, do not require manual specification of test statistics, allow very flexible correlation structures, allow inferences about the correlation and covariance structure, and generate better predicted values than ANOVA-based methods.

References

Austin, P., Alter, D., and Tu, J. 2003. The use of fixed- and random-effects models for classifying hospitals as mortality outliers: A Monte Carlo assessment. *Medical Decision Making*, 23: 526–539.

Federer, W.T., 1955. *Experimental Design: Theory and Application*. New York: MacMillan.

Fisher, N, Hurwitz, S., Jeunemaitre, X., Hopkins, P.N., Hollenberg, N.K., and Williams, G.H. 2002. Familial aggregation of low-renin hypertension. *Hypertension*, 39: 914–918.

Fisher, R.A. 1935. *Statistical Methods for Research Workers*. London: Oliver and Boyd.

Greenhouse, S. W., and Geisser, S. 1959. On methods in the analysis of profile data. *Psychometrika*, 32: 95–112.

Huynh, H., and Feldt, L. S. 1976. Estimation of the box correction for degrees of freedom from sample data in the randomized block and split plot designs, *Journal of Educational Statistics*, 1: 69–82.

Laird, N. 1988. Missing data in longitudinal studies. *Statistics in Medicine*, 7: 305–315.

Kenward, M.G., and Roger, J.H. 1997. Small sample inference for fixed effects from restricted maximum likelihood, *Biometrics*, 53: 983–997.

McCulloch, C. E., and Searle, S. R. 2000. *Generalized, Linear, and Mixed Models*. New York: Wiley.

Scheffe, H. 1959. *The Analysis of Variance*. New York: Wiley.

Searle, S. R. (1970). Linear Models, New York: Wiley.

Searle, S. R., Casella, G., and McCulloch, C. E. 1992. *Variance Components*. New York: Wiley.

Tietjen, G.L. 1974. Exact and approximate tests for unbalanced random effects designs. *Biometrics*, 30: 573–581.

Comment: Anova as a Tool for Structuring and Understanding Hierarchical Models

Andrew Gelman

I agree with McCulloch that hierarchical models (which consider the persons in the experiment as a random sample from a hypothetical population) are a good idea for repeated measures data. As McCulloch points out, this assumption is typically well motivated by the goal of extrapolating the experimental findings to the population. He also explains why classical ANOVA is not the best tool for exploring such data.

However, if we think about ANOVA more broadly—as a way of structuring statisical analyses rather than as one particular set of computations—I believe a unification is possible that will give us the benefits of hierarchical modeling (efficient estimation, even under imbalance, missing data, nonnormality, and other realistic data conditions, as discussed by McCulloch), while also preserving the benefits of ANOVA (the summary of a complicated model in terms of batches of coefficients and variance parameters). In my own areas of application, I have not found much use for F-tests and p-values, but I have found concepts, such as decomposition of degrees of freedom, and estimation of the importance of different components of variation, to be helpful. In the embrace of maximum likelihood (or, more generally, Bayesian) estimation, I do not want to lose these helpful summaries.

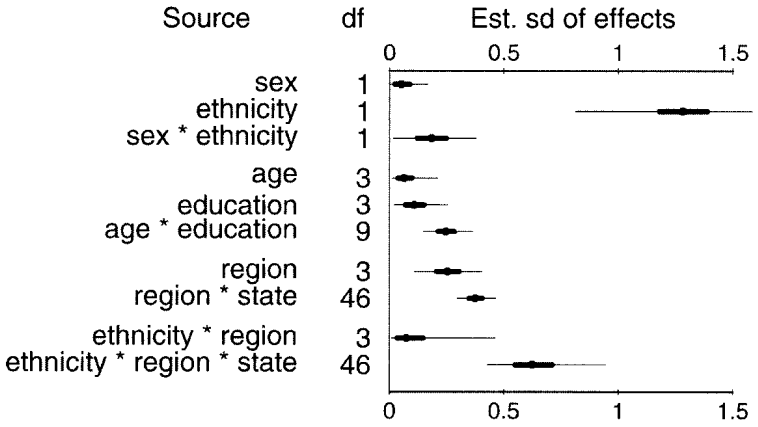


Figure 1. ANOVA display for a multilevel logistic regression model fit to survey data on voter preferences. The estimate, 50% interval, and 95% interval are shown for the finite-sample standard deviation of each batch of effects in the model. From Gelman (2005).

So, how can we get the most out of ANOVA in a likelihood/Bayesian modeling context? Each row of the ANOVA table corresponds to, and labels, a different batch of coefficients in the linear model. For example, McCulloch's Table 1 has five rows, and his equation (1) has five subsetted coefficients. I would like to see, for each row of the ANOVA table, an estimate of the standard deviation of its batch of coefficients. This idea is discussed fully in Gelman (2005); for an example, see Figure 1.

Thus, I agree with McCulloch's point that hierarchical models and likelihood-based estimation can work better than classical ANOVA, especially in complex settings; but I would like to reserve a role for the concepts of ANOVA to help us understand fitted models. The goal here is not to test hypotheses of zero effects but rather to summarize the importance of each batch of coefficients.

One of the most important advantages of model-summary tools is that they can facilitate the fitting of multiple models. With regard to McCulloch's particular application discussed here, I would be interested in seeing interactions between person and alcohol and between person and pregnenolone. These batches of interactions would represent random samples from the interactions in the entire population and thus would have additional variance components, each estimated from n-1 degrees of freedom. An ANOVA display along the lines of Figure 1. could help us understand such a model and could lead to further investigation of treatment effects of interest. Hierarchical estimation tools of the sort discussed by McCulloch are crucial in allowing us to fit these models.

Reference

Gelman, A. 2005. Analysis of variance: Why it is more important than ever (with discussion). *Annals of Statistics*, 33: 1-53.