The Annals of Statistics
2005, Vol. 33, No. 1, 1–53
DOI 10.1214/009053604000001048

DISCUSSION PAPER

ANALYSIS OF VARIANCE—WHY IT IS MORE IMPORTANT THAN EVER[1]

BY ANDREW GELMAN

Columbia University

Analysis of variance (ANOVA) is an extremely important method in exploratory and confirmatory data analysis. Unfortunately, in complex problems (e.g., split-plot designs), it is not always easy to set up an appropriate ANOVA. We propose a hierarchical analysis that automatically gives the correct ANOVA comparisons even in complex scenarios. The inferences for all means and variances are performed under a model with a separate batch of effects for each row of the ANOVA table.

We connect to classical ANOVA by working with finite-sample variance components: fixed and random effects models are characterized by inferences about existing levels of a factor and new levels, respectively. We also introduce a new graphical display showing inferences about the standard deviations of each batch of effects.

We illustrate with two examples from our applied data analysis, first illustrating the usefulness of our hierarchical computations and displays, and second showing how the ideas of ANOVA are helpful in understanding a previously fit hierarchical model.

1. Is ANOVA obsolete? What is the analysis of variance? Econometricians see it as an uninteresting special case of linear regression. Bayesians see it as an inflexible classical method. Theoretical statisticians have supplied many mathematical definitions [see, e.g., Speed (1987)]. Instructors see it as one of the hardest topics in classical statistics to teach, especially in its more elaborate forms such as split-plot analysis. We believe, however, that the ideas of ANOVA are useful in many applications of statistics. For the purpose of this paper, we identify ANOVA with the structuring of parameters into batches—that is, with variance components models. There are more general mathematical formulations of the analysis of variance, but this is the aspect that we believe is most relevant in applied statistics, especially for regression modeling.

Received November 2002; revised November 2003.

[1]Supported in part by the National Science Foundation with Young Investigator Award DMS-97-96129 and Grants SBR-97-08424, SES-99-87748 and SES-00-84368. A version of this paper was originally presented as a special Invited Lecture for the Institute of Mathematical Statistics.

AMS 2000 subject classifications. 62J10, 62J07, 62F15, 62J05, 62J12.

Key words and phrases. ANOVA, Bayesian inference, fixed effects, hierarchical model, linear regression, multilevel model, random effects, variance components.

We shall demonstrate how many of the difficulties in understanding and computing ANOVAs can be resolved using a hierarchical Bayesian framework. Conversely, we illustrate how thinking in terms of variance components can be useful in understanding and displaying hierarchical regressions. With hierarchical (multilevel) models becoming used more and more widely, we view ANOVA as more important than ever in statistical applications.

Classical ANOVA for balanced data does three things at once:

1. As exploratory data analysis, an ANOVA is an organization of an additive data decomposition, and its sums of squares indicate the variance of each component of the decomposition (or, equivalently, each set of terms of a linear model).
2. Comparisons of mean squares, along with F-tests [or F-like tests; see, e.g., Cornfield and Tukey (1956)], allow testing of a nested sequence of models.
3. Closely related to the ANOVA is a linear model fit with coefficient estimates and standard errors.

Unfortunately, in the classical literature there is some debate on how to perform ANOVA in complicated data structures with nesting, crossing and lack of balance. In fact, given the multiple goals listed above, it is not at all obvious that a procedure recognizable as "ANOVA" should be possible at all in general settings [which is perhaps one reason that Speed (1987) restricts ANOVA to balanced designs].

In a linear regression, or more generally an additive model, ANOVA represents a batching of effects, with each row of the ANOVA table corresponding to a set of predictors. We are potentially interested in the individual coefficients and also in the variance of the coefficients in each batch. Our approach is to use variance components modeling for all rows of the table, even for those sources of variation that have commonly been regarded as fixed effects. We thus borrow many ideas from the classical variance components literature.

As we show in Section 2 of this paper, least-squares regression solves some ANOVA problems but has trouble with hierarchical structures [see also Gelman (2000)]. In Sections 3 and 4 we present a more general hierarchical regression approach that works in all ANOVA problems in which effects are structured into exchangeable batches, following the approach of Sargent and Hodges (1997). In this sense, ANOVA is indeed a special case of linear regression, but only if hierarchical models are used. In fact, the batching of effects in a hierarchical model has an exact counterpart in the rows of the analysis of variance table. Section 5 presents a new analysis of variance table that we believe more directly addresses the questions of interest in linear models, and Section 6 discusses the distinction between fixed and random effects. We present two applied examples in Section 7 and conclude with some open problems in Section 8.

2. ANOVA and linear regression. We begin by reviewing the benefits and limitations of classical nonhierarchical regression for ANOVA problems.

2.1. *ANOVA and classical regression*: *good news.* It is well known that many ANOVA computations can be performed using linear regression computations, with each row of the ANOVA table corresponding to the variance of a corresponding set of regression coefficients.

2.1.1. *Latin square.* For a simple example, consider a Latin square with five treatments randomized to a 5×5 array of plots. The ANOVA regression has 25 data points and the following predictors: one constant, four rows, four columns and four treatments, with only four in each batch because, if all five were included, the predictors would be collinear. (Although not necessary for understanding the mathematical structure of the model, the details of counting the predictors and checking for collinearity are important in actually implementing the regression computation and are relevant to the question of whether ANOVA can be computed simply using classical regression. As we shall discuss in Section 3.1, we ultimately will find it more helpful to include all five predictors in each batch using a hierarchical regression framework.)

For each of the three batches of variables in the Latin square problem, the variance of the $J = 5$ underlying coefficients can be estimated using the basic variance decomposition formula, where we use the notation $\text{var}_{j=1}^{J}$ for the sample variance of J items:

$$\begin{aligned}\text{E(variance between the } \hat{\beta}_j\text{'s)} &= \text{variance between the true } \beta_j\text{'s} \\ &\quad + \text{estimation variance,} \\ \text{E}(\text{var}_{j=1}^{J} \hat{\beta}_j) &= \text{var}_{j=1}^{J} \beta_j + \text{E}\big(\text{var}(\hat{\beta}_j|\beta_j)\big), \\ \text{E}(V(\hat{\beta})) &= V(\beta) + V_{\text{estimation}}.\end{aligned} \tag{1}$$

One can compute $V(\hat{\beta})$ and an estimate of $V_{\text{estimation}}$ directly from the coefficient estimates and standard errors, respectively, in the linear regression output, and then use the simple unbiased estimate,

$$\widehat{V}(\beta) = V(\hat{\beta}) - \widehat{V}_{\text{estimation}}. \tag{2}$$

[More sophisticated estimates of variance components are possible; see, e.g., Searle, Casella and McCulloch (1992).] An F-test for null treatment effects corresponds to a test that $V(\beta) = 0$.

Unlike in the usual ANOVA setup, here we do not need to decide on the comparison variances (i.e., the denominators for the F-tests). The regression automatically gives standard errors for coefficient estimates that can directly be input into $\widehat{V}_{\text{estimation}}$ in (2).

2.1.2. *Comparing two treatments.* The benefits of the regression approach can be further seen in two simple examples. First, consider a simple experiment with 20 units completely randomized to two treatments, with each treatment applied to 10 units. The regression has 20 data points and two predictors: one constant and one treatment indicator (or no constant and two treatment indicators). Eighteen degrees of freedom are available to estimate the residual variance, just as in the corresponding ANOVA.

Next, consider a design with 10 pairs of units, with the two treatments randomized within each pair. The corresponding regression analysis has 20 data points and 11 predictors: one constant, one indicator for treatment and nine indicators for pairs, and, if you run the regression, the standard errors for the treatment effect estimates are automatically based on the nine degrees of freedom for the within-pair variance.

The different analyses for paired and unpaired designs are confusing for students, but here they are clearly determined by the principle of including in the regression all the information used in the design.

2.2. *ANOVA and classical regression*: *bad news.* Now we consider two examples where classical nonhierarchical regression *cannot* be used to automatically get the correct answer.

2.2.1. *A split-plot Latin square.* Here is the form of the analysis of variance table for a $5 \times 5 \times 2$ split-plot Latin square: a standard experimental design but one that is complicated enough that most students analyze it incorrectly unless they are told where to look it up. (We view the difficulty of teaching these principles as a sign of the awkwardness of the usual theoretical framework of these ideas rather than a fault of the students.)

Source	**df**
row	4
column	4
(A, B, C, D, E)	4
plot	12
$(1, 2)$	1
row $\times$ $(1, 2)$	4
column $\times$ $(1, 2)$	4
(A, B, C, D, E) $\times$ $(1, 2)$	4
plot $\times$ $(1, 2)$	12

In this example, there are 25 plots with five full-plot treatments (labeled A, B, C, D, E), and each plot is divided into two subplots with subplot varieties (labeled

1 and 2). As is indicated by the horizontal lines in the ANOVA table, the main-plot residual mean squares should be used for the main-plot effects and the sub-plot residual mean squares for the sub-plot effects.

It is not hard for a student to decompose the 49 degrees of freedom to the rows in the ANOVA table; the tricky part of the analysis is to know which residuals are to be used for which comparisons.

What happens if we input the data into the `aov` function in the statistical package `S-Plus`? This program uses the linear-model fitting routine `lm`, as one might expect based on the theory that analysis of variance is a special case of linear regression. [E.g., Fox (2002) writes, "It is, from one point of view, unnecessary to consider analysis of variance models separately from the general class of linear models."] Figure 1 shows three attempts to fit the split-plot data with `aov`, only the last of which worked. We include this not to disparage `S-Plus` in any way but just to point out that ANOVA can be done in many ways in the classical linear regression framework, and not all these ways give the correct answer.

At this point, we seem to have the following "method" for analysis of variance: first, recognize the form of the problem (e.g., split-plot Latin square); second, look it up in an authoritative book such as Snedecor and Cochran (1989) or Cochran and Cox (1957); third, perform the computations, using the appropriate residual mean squares. This is unappealing for practice as well as teaching and in addition contradicts the idea that, "If you know linear regression, you know ANOVA."

2.2.2. *A simple hierarchical design.* We continue to explore the difficulties of regression for ANOVA with a simple example. Consider an experiment on four treatments for an industrial process applied to 20 machines (randomly divided into four groups of 5), with each treatment applied six times independently on each of its five machines. For simplicity, we assume no systematic time effects, so that the six measurements are simply replications. The ANOVA table is then

Source	**df**
treatment	3
treatment × machine	16
treatment × machine × measurement	100

There are no rows for just "machine" or "measurement" because the design is fully nested.

Without knowing ANOVA, is it possible to get appropriate inferences for the treatment effects using linear regression? The averages for the treatments

```
> summary (aov (data ~ rows + columns + tABCDE + plots +
  t12*rows + t12*columns + t12*tABCDE + t12*plots))
            Df Sum Sq Mean Sq F value    Pr(>F)
rows         4 288.48   72.12  4.0283 0.0268475 *
columns      4 389.48   97.37  5.4387 0.0098253 **
tABCDE       4 702.28  175.57  9.8066 0.0009245 ***
plots       12 308.04   25.67  1.4338 0.2710432
t12          1 332.82  332.82 18.5898 0.0010110 **
rows:t12     4  74.08   18.52  1.0344 0.4291297
columns:t12  4  96.68   24.17  1.3500 0.3079352
tABCDE:t12   4  57.08   14.27  0.7971 0.5496092
Residuals   12 214.84   17.90


> summary (aov (data ~ rows + columns + tABCDE +
  t12*rows + t12*columns + t12*tABCDE + t12*plots + Error(plots)))
Error: plots
          Df Sum Sq Mean Sq F value Pr(>F)
rows       4 288.48   72.12  7.3592 0.2689
columns    4 389.48   97.37  9.9357 0.2331
tABCDE     4 702.28  175.57 17.9153 0.1752
plots     11 298.24   27.11  2.7666 0.4401
Residuals  1   9.80    9.80

Error: Within
            Df Sum Sq Mean Sq F value Pr(>F)
t12          1 332.82  332.82  7.3960 0.2243
rows:t12     4  74.08   18.52  0.4116 0.8059
columns:t12  4  96.68   24.17  0.5371 0.7559
tABCDE:t12   4  57.08   14.27  0.3171 0.8496
t12:plots   11 169.84   15.44  0.3431 0.8842
Residuals    1  45.00   45.00


> summary (aov (data ~ rows + columns + tABCDE +
  t12*rows + t12*columns + t12*tABCDE + Error(plots))
Error: plots
          Df Sum Sq Mean Sq F value   Pr(>F)
rows       4 288.48   72.12  2.8095 0.073984 .
columns    4 389.48   97.37  3.7931 0.032271 *
tABCDE     4 702.28  175.57  6.8395 0.004154 **
Residuals 12 308.04   25.67

Error: Within
            Df Sum Sq Mean Sq F value   Pr(>F)
t12          1 332.82  332.82 18.5898 0.001011 **
rows:t12     4  74.08   18.52  1.0344 0.429130
columns:t12  4  96.68   24.17  1.3500 0.307935
tABCDE:t12   4  57.08   14.27  0.7971 0.549609
Residuals   12 214.84   17.90
```

FIG. 1. *Three attempts at running the* `aov` *command in* `S-Plus`. *Only the last gave the correct comparisons. This is not intended as a criticism of* `S-Plus`; *in general, classical ANOVA requires careful identification of variance components in order to give the correct results with hierarchical data structures.*

$i = 1, \ldots, 4$ can be written in two ways:

$$\bar{y}_{i\cdot\cdot} = \tfrac{1}{30} \sum_{j=1}^{5} \sum_{k=1}^{6} y_{ijk} \tag{3}$$

or

$$\bar{y}_{i\cdot\cdot} = \tfrac{1}{5} \sum_{j=1}^{5} \bar{y}_{ij\cdot} \tag{4}$$

Formula (3) uses all the data and suggests a standard error based on 29 degrees of freedom for each treatment, but this would ignore the nesting in the design. Formula (4) follows the design and suggests a standard error based on the four degrees of freedom from the five machines for each treatment.

Formulas (3) and (4) give the same estimated treatment effects but imply different standard errors and different ANOVA F-tests. If there is any chance of machine effects, the second analysis is standard. However, to do this you must know to base your uncertainties on the "treatment × machine" variance, not the "treatment × machine × measurement" variance. An automatic ANOVA program must be able to automatically correctly choose this comparison variance.

Can this problem be solved using least-squares regression on the 120 data points? The simplest regression uses four predictors—one constant term and three treatment indicators—with 116 residual degrees of freedom. This model gives the wrong residual variance: we want the between-machine, not the between-measurement, variance.

Since the machines are used in the design, they should be included in the analysis. This suggests a model with 24 predictors: one constant, three treatment indicators, and 20 machine indicators. But these predictors are collinear, so we must eliminate four of the machine indicators. Unfortunately, the standard errors of the treatment effects in this model are estimated using the within-machine variation, which is still wrong. The problem becomes even more difficult if the design is unbalanced.

The appropriate analysis, of course, is to include the 20 machines as a variance component, which classically could be estimated using REML (treating the machine effects as missing data) or using regression without machine effects but with a block-structured covariance matrix with intraclass correlation estimated from data. In a Bayesian context the machine effects would be estimated with a population distribution whose variance is estimated from data, as we discuss in general in the next section. In any case, we would like to come at this answer simply by identifying the important effects—treatments and machines—without having to explicitly recognize the hierarchical nature of the design, in the same way that we would like to be able to analyze split-plot data without the potential mishaps illustrated in Figure 1.

3. ANOVA using hierarchical regression.

3.1. *Formulation as a regression model.* We shall work with linear models, with the "analysis of variance" corresponding to the batching of effects into "sources of variation," and each batch corresponding to one row of the ANOVA table. This is the model of Sargent and Hodges (1997). We use the notation $m = 1, \ldots, M$ for the rows of the table. Each row m represents a batch of J_m regression coefficients $\beta_j^{(m)}$, $j = 1, \ldots, J_m$. We denote the mth subvector of coefficients as $\beta^{(m)} = (\beta_1^{(m)}, \ldots, \beta_{J_m}^{(m)})$ and the corresponding classical least-squares estimate as $\hat{\beta}^{(m)}$. These estimates are subject to c_m linear constraints, yielding $(df)_m = J_m - c_m$ degrees of freedom. We label the constraint matrix as $C^{(m)}$, so that $C^{(m)}\hat{\beta}^{(m)} = 0$ for all m. For notational convenience, we label the grand mean as $\beta_1^{(0)}$, corresponding to the (invisible) zeroth row of the ANOVA table and estimated with no linear constraints.

The linear model is fit to the data points y_i, $i = 1, \ldots, n$, and can be written as

$$y_i = \sum_{m=0}^{M} \beta_{j_i^m}^{(m)}, \tag{5}$$

where j_i^m indexes the appropriate coefficient j in batch m corresponding to data point i. Thus, each data point pulls one coefficient from each row in the ANOVA table. Equation (5) could also be expressed as a linear regression model with a design matrix composed entirely of 0's and 1's. The coefficients β_j^M of the last row of the table correspond to the residuals or error term of the model. ANOVA can also be applied more generally to regression models (or to generalized linear models), in which case we could have any design matrix X, and (5) would be generalized to

$$y_i = \sum_{m=0}^{M} \sum_{j=1}^{J_m} x_{ij}^{(m)} \beta_j^{(m)}. \tag{6}$$

The essence of analysis of variance is in the structuring of the coefficients into batches—hence the notation $\beta_j^{(m)}$—going beyond the usual linear model formulation that has a single indexing of coefficients β_j. We assume that the structure (5), or the more general regression parameterization (6), has already been constructed using knowledge of the data structure. To use ANOVA terminology, we assume the sources of variation have already been set, and our goal is to perform inference for each variance component.

We shall use a hierarchical formulation in which each batch of regression coefficients is modeled as a sample from a normal distribution with mean 0 and its own variance σ_m^2:

$$\beta_j^{(m)} \sim \mathrm{N}(0, \sigma_m^2) \qquad \text{for } j = 1, \ldots, J_m \text{ for each batch } m = 1, \ldots, M. \tag{7}$$

We follow the notation of Nelder (1977, 1994) by modeling the underlying β coefficients as unconstrained, unlike the least-squares estimates. Setting the variances σ_m^2 to ∞ and constraining the $\beta_j^{(m)}$'s yields classical least-squares estimates.

Model (7) corresponds to exchangeability of each set of factor levels, which is a form of partial exchangeability or invariance of the entire set of cell means [see Aldous (1981)]. We do not mean to suggest that this model is universally appropriate for data but rather that it is often used, explicitly or implicitly, as a starting point for assessing the relative importance of the effects β in linear models structured as in (5) and (6). We discuss nonexchangeable models in Section 8.3.

One measure of the importance of each row or "source" in the ANOVA table is the standard deviation of its constrained regression coefficients, which we denote

$$s_m = \sqrt{\frac{1}{(df)_m}\beta^{(m)T}\left[I - C^{(m)}\left(C^{(m)T}C^{(m)}\right)^{-1}C^{(m)T}\right]\beta^{(m)}}, \tag{8}$$

where $\beta^{(m)}$ is the vector of coefficients in batch m and $C^{(m)}$ is the $c_m \times J_m$ full rank matrix of constraints (for which $C^{(m)}\beta^{(m)} = 0$). Expression (8) is just the mean square of the coefficients' residuals after projection to the constraint space. We divide by $(df)_m = J_m - c_m$ rather than $J_m - 1$ because multiplying by $C^{(m)}$ induces c_m linear constraints.

Variance estimation is often presented in terms of the superpopulation standard deviations σ_m, but in our ANOVA summaries we focus on the finite-population quantities s_m for reasons discussed in Section 3.5. However, for computational reasons the parameters σ_m are useful intermediate quantities to estimate.

3.2. *Batching of regression coefficients.* Our general solution to the ANOVA problem is simple: we treat *every* row in the table as a batch of "random effects"; that is, a set of regression coefficients drawn from a distribution with mean 0 and some standard deviation to be estimated from the data. The mean of 0 comes naturally from the ANOVA decomposition structure (pulling out the grand mean, main effects, interactions and so forth), and the standard deviations are simply the magnitudes of the variance components corresponding to each row of the table. For example, we can write the simple hierarchical design of Section 2.2.2 as

Source	**Number of coefficients**	**Standard deviation**
treatment	4	s_1
treatment × machine	20	s_2
treatment × machine × measurement	120	s_3

Except for our focus on s rather than σ, this is the approach recommended by Box and Tiao (1973) although computational difficulties made it difficult to implement at that time.

The primary goal of ANOVA is to estimate the variance components (in this case, s_1, s_2, s_3) and compare them to zero and to each other. The secondary goal is to estimate (and summarize the uncertainties in) the individual coefficients, especially, in this example, the four treatment effects. From the hierarchical model the coefficient estimates will be pulled toward zero, with the amount of shrinkage determined by the estimated variance components. But, more importantly, the variance components and standard errors are estimated from the data, without any need to specify comparisons based on the design. Thus, the struggles of Section 2.2 are avoided, and (hierarchical) linear regression can indeed be used to compute ANOVA automatically, once the rows of the table (the sources of variation) have been specified.

For another example, the split-plot Latin square looks like

Source	**Number of coefficients**	**Standard deviation**
row	5	s_1
column	5	s_2
(A, B, C, D, E)	5	s_3
plot	25	s_4
$(1, 2)$	2	s_5
row $\times$ $(1, 2)$	10	s_6
column $\times$ $(1, 2)$	10	s_7
(A, B, C, D, E) $\times$ $(1, 2)$	10	s_8
plot $\times$ $(1, 2)$	50	s_9

This is automatic, based on the principle that all variables in the design be included in the analysis. Setting up the model in this way, with all nine variance components estimated, automatically gives the correct comparisons (e.g., uncertainties for comparisons between treatments A, B, C, D, E will be estimated based on main-plot variation and uncertainties for varieties 1, 2 will be estimated based on sub-plot variation).

3.3. *Getting something for nothing*? At this point we seem to have a paradox. In classical ANOVA, you (sometimes) need to know the design in order to select the correct analysis, as in the examples in Section 2.2. But the hierarchical analysis does it automatically. How can this be? How can the analysis "know" how to do the split-plot analysis, for example, without being "told" that the data come from a split-plot design?

The answer is in two parts. First, as with the classical analyses, we require that the rows of the ANOVA be specified by the modeler. In the notation of (5) and (6),

the user must specify the structuring or batching of the linear parameters β. In the classical analysis, however, this is not enough, as discussed in Section 2.2.

The second part of making the hierarchical ANOVA work is that the information from the design is encoded in the design matrix of the linear regression [as shown by Nelder (1965a, b) and implemented in the software Genstat]. For example, the nesting in the example of Section 2.2.2 is reflected in the collinearity of the machine indicators within each treatment. The automatic encoding is particularly useful in incomplete designs where there is no simple classical analysis.

From a linear-modeling perspective, classical nonhierarchical regression has a serious limitation: each batch of parameters (corresponding to each row of the ANOVA table) must be included with no shrinkage (i.e., $\sigma_m = \infty$) or excluded ($\sigma_m = 0$), with the exception of the last row of the table, whose variance can be estimated. In the example of Section 2.2.2, we must either include the machine effects unshrunken or ignore them, and neither approach gives the correct analysis. The hierarchical model works automatically because it allows finite nonzero values for all the variance components.

The hierarchical regression analysis is based on the model of exchangeable effects within batches, as expressed in model (7), which is not necessarily the best analysis in any particular application. For example, Besag and Higdon (1999) recommend using spatial models (rather than exchangeable row and column effects) for data such as in the split-plot experiment described previously. Here we are simply trying to understand why, when given the standard assumptions underlying the classical ANOVA, the hierarchical analysis automatically gives the appropriate inferences for the variance components without the need for additional effort of identifying appropriate error terms for each row of the table.

3.4. *Classical and Bayesian interpretations.* We are most comfortable interpreting the linear model in a Bayesian manner, that is, with a joint probability distribution on all unknown parameters. However, our recommended hierarchical approach can also be considered classically, in which case the regression coefficients are considered as random variables (and thus are "predicted") and the variance components are considered as parameters (and thus "estimated"); see Robinson (1991) and Gelman, Carlin, Stern and Rubin [(1995), page 380]. The main difference between classical and Bayesian methods here is between using a point estimate for the variance parameters or including uncertainty distributions. Conditional on the parameters σ_m, the classical and Bayesian inferences for the linear parameters β_j^m are identical in our ANOVA models. In either case, the individual regression coefficients are estimated by linear unbiased predictors or, equivalently, posterior means, balancing the direct information on each parameter with the shrinkage from the batch of effects. There will be more shrinkage for batches of effects whose standard deviations σ_m are near zero, which will occur for factors that contribute little variation to the data.

When will it make a practical difference to estimate variance parameters Bayesianly rather than with point estimates? Only when these variances are hard to distinguish from 0. For example, Figure 2 shows the posterior distribution of the hierarchical standard deviation from an example of Rubin (1981) and Gelman, Carlin, Stern and Rubin [(1995), Chapter 5]. The data are consistent with a standard deviation of 0, but it could also be as high as 10 or 20. Setting the variance parameter to zero in such a situation is generally *not* desirable because it would lead to falsely precise estimates of the $\beta_j^{(m)}$'s. Setting the variance to some nonzero value would require additional work which, in practice, would not be done since it would offer no advantages over Bayesian posterior averaging.

It might be argued that such examples—in which the maximum likelihood estimate of the hierarchical variance is at or near zero—are pathological and unlikely to occur in practice. But we would argue that such situations will be common in ANOVA settings, for two reasons. First, when studying the many rows of a large ANOVA table, we expect (in fact, we hope) to see various near-zero variances at higher levels of interaction. After all, one of the purposes of an ANOVA decomposition is to identify the important main effects and interactions in a complex data set [see Sargent and Hodges (1997)]. Nonsignificant rows of the ANOVA table correspond to variance components that are statistically indistinguishable from zero. Our second reason for expecting to see near-zero variance components is that, as informative covariates are added to a linear

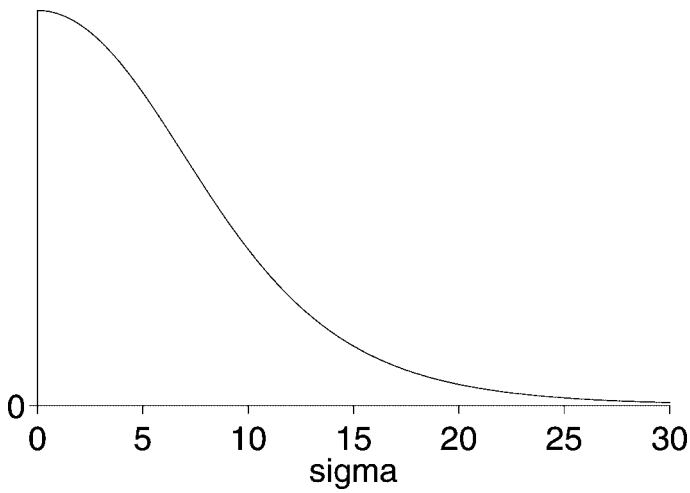

FIG. 2. *Illustration of the difficulties of point estimation for variance components. Pictured is the marginal posterior distribution for a hierarchical standard deviation parameter from Rubin* (1981) *and Gelman, Carlin, Stern and Rubin* [(1995), *Chapter* 5]. *The simplest point estimate, the posterior mode or REML estimate, is zero, but this estimate is on the extreme of parameter space and would cause the inferences to understate the uncertainties in this batch of regression coefficients.*

model, hierarchical variances decrease until it is no longer possible to add more information [see Gelman (1996)].

When variance parameters are not well summarized by point estimates, Bayesian inferences are sensitive to the prior distribution. For our basic ANOVA computations we use noninformative prior distributions of the form $p(\sigma_m) \propto 1$ (which can be considered as a degenerate case of the inverse-gamma family, as we discuss in Section 4.2). We further discuss the issue of near-zero variance components in Section 8.2.

3.5. *Superpopulation and finite-population variances.* For each row m of an ANOVA table, there are two natural variance parameters to estimate: the *superpopulation* standard deviation σ_m and the *finite-population* standard deviation s_m as defined in (8). The superpopulation standard deviation characterizes the uncertainty for predicting a new coefficient from batch m, whereas the finite-population standard deviation describes the existing J_m coefficients. The two variances can be given the same point estimate—in classical unbiased estimation $\mathrm{E}(s_m^2|\sigma_m^2) = \sigma_m^2$, and in Bayesian inference with a noninformative prior distribution (see Section 4.2) the conditional posterior mode of σ_m^2 given all other parameters in the model is s^2. The superpopulation variance has more uncertainty, however.

To see the difference between the two variances, consider the extreme case in which $J_m = 2$ [and so $(df)_m = 1$] and a large amount of data is available in both groups. Then the two parameters $\beta_1^{(m)}$ and $\beta_2^{(m)}$ will be estimated accurately and so will $s_m^2 = (\beta_1^{(m)} - \beta_2^{(m)})^2/2$. The superpopulation variance σ_m^2, on the other hand, is only being estimated by a measurement that is proportional to a χ^2 with one degree of freedom. We know much about the two parameters $\beta_1^{(m)}, \beta_2^{(m)}$ but can say little about others from their batch.

As we discuss in Section 6, we believe that much of the literature on fixed and random effects can be fruitfully reexpressed in terms of finite-population and superpopulation inferences. In some contexts (e.g., obtaining inference for the 50 U.S. states) the finite population seems more meaningful, whereas in others (e.g., subject-level effects in a psychological experiment) interest clearly lies in the superpopulation.

To keep connection with classical ANOVA, which focuses on a description—a variance decomposition—of an existing dataset, we focus on finite-population variances s_m^2. However, as an intermediate step in any computation—classical or Bayesian—we perform inferences about the superpopulation variances σ_m^2.

4. Inference for the variance components.

4.1. *Classical inference.* Although we have argued that hierarchical models are best analyzed using Bayesian methods, we discuss classical computations first, partly because of their simplicity and partly to connect to the vast literature on

the estimation of variance components [see, e.g., Searle, Casella and McCulloch (1992)]. The basic tool is the method of moments. We can first estimate the superpopulation variances σ_m^2 and their approximate uncertainty intervals, then go back and estimate uncertainty intervals for the finite-population variances s_m^2. Here we are working with the additive model (5) rather than the general regression formulation (6).

The estimates for the parameters σ_m^2 are standard and can be expressed in terms of classical ANOVA quantities, as follows. The sum of squares for row m is the sum of the squared coefficient estimates corresponding to the n data points,

$$SS_m = \sum_{i=1}^{n} (\hat{\beta}_{j_i^m}^{(m)})^2,$$

and can also be written as a weighted sum of the squared coefficient estimates for that row,

$$SS_m = n \sum_{j=1}^{J_m} w_j (\hat{\beta}_j^{(m)})^2,$$

where the weights w_j sum to 1, and

$$\text{for balanced designs:} \quad SS_m = \frac{n}{J_m} \sum_{j=1}^{J_m} (\hat{\beta}_j^{(m)})^2.$$

The mean square is the sum of squares divided by degrees of freedom,

$$MS_m = SS_m/(df)_m$$

and

$$\text{for balanced designs:} \quad MS_m = \frac{n}{J_m (df)_m} \sum_{j=1}^{J_m} (\hat{\beta}_j^{(m)})^2.$$

The all-important expected mean square, EMS_m, is the expected contribution of sampling variance to MS_m, and it is also $\mathrm{E}(MS_m)$ under the null hypothesis that the coefficients $\beta_j^{(m)}$ are all equal to zero. Much of the classical literature is devoted to determining EMS_m under different designs and different assumptions, and computing or approximating the F-ratio, MS_m/EMS_m, to assess statistical significance.

We shall proceed in a slightly different direction. First, we compute EMS_m under the general model allowing all other variance components in the model to be nonzero. (This means that, in general, EMS_m depends on variance components estimated lower down in the ANOVA table.) Second, we use the expected mean square as a tool to *estimate* variance components, not to test their statistical significance. Both these steps follow classical practice for random effects; our only

innovation is to indiscriminately apply them to *all* the variance components in a model, and to follow this computation with an estimate of the uncertainty in the finite-population variances s_m^2.

We find it more convenient to work with not the sums of squares or mean squares but with the variances of the batches of estimated regression coefficients, which we label as

$$V_m = \frac{1}{(df)_m} \sum_{j=1}^{J_m} (\hat{\beta}_j^{(m)})^2. \tag{9}$$

V_m can be considered a variance since for each row the J_m effect estimates $\hat{\beta}^{(m)}$ have several linear constraints [with $(df)_m$ remaining degrees of freedom] and must sum to 0. [For the "zeroth" row of the table, we define $V_0 = (\hat{\beta}_1^{(0)})^2$, the square of the estimated grand mean in the model.] For each row of the table,

$$\text{for balanced designs:} \quad V_m = \frac{J_m}{n} MS_m.$$

We start by estimating the superpopulation variances σ_m^2, and the constrained method-of-moments estimator is based on the variance-decomposition identity [see (1)]

$$\mathrm{E}(V_m) = \sigma_m^2 + EV_m,$$

where EV_m is the contribution of sampling variance to V_m, that is, the expected value of V_m if σ_m were equal to 0. EV_m in turn depends on other variance components in the model, and

$$\text{for balanced designs:} \quad EV_m = \frac{J_m}{n} EMS_m.$$

The natural estimate of the underlying variance is then

$$\hat{\sigma}_m^2 = \max(0, V_m - \widehat{EV}_m). \tag{10}$$

The expected value $\widehat{EV}_m$ is itself estimated based on the other variance components in the model, as we discuss shortly.

Thus, the classical hierarchical ANOVA computations reduce to estimating the expected mean squares EMS_m (and thus EV_m) in terms of the estimated variance components σ_m. For nonbalanced designs, this can be complicated compared to the Bayesian computation as described in Section 4.2.

For balanced designs, however, simple formulas exist. We do not go through all the literature here [see, e.g., Cornfield and Tukey (1956), Green and Tukey (1960) and Plackett (1960)]. A summary is given in Searle, Casella and McCulloch [(1992), Section 4.2]. The basic idea is that, in a balanced design, the effect estimates $\hat{\beta}_j^{(m)}$ in a batch m are simply averages of data, adjusted to fit a set of linear constraints. The sampling variance $\widehat{EV}_m$ in (10) can be written in terms

of variances σ_k^2 for all batches k representing interactions that include m in the ANOVA table. We write this as

$$\widehat{EV}_m = \sum_{k \in I(m)} \frac{J_m}{J_k} \sigma_k^2, \tag{11}$$

where $I(m)$ represents the set of all rows in the ANOVA table representing interactions that include the variables m as a subset. For example, in the example in Section 2.2.2, consider the treatment effects (i.e., $m = 1$ in the ANOVA table). Here, $J_1 = 4$, $n = 120$ and $\widehat{EV}_1 = \frac{4}{20}\sigma_2^2 + \frac{4}{120}\sigma_3^2$. For another example, in the split-plot latin square in Section 2.2.1, the main-plot treatment effects are the third row of the ANOVA table ($m = 3$), and $\widehat{EV}_3 = \frac{5}{25}\sigma_4^2 + \frac{5}{10}\sigma_8^2 + \frac{5}{50}\sigma_9^2$.

For balanced designs, then, variance components can be estimated by starting at the bottom of the table (with the highest-level interaction, or residuals) and then working upwards, at each step using the appropriate variance components from lower in the table in formulas (10) and (11). In this way the variance components σ_m^2 can be estimated noniteratively. Alternatively, we can compute the moments estimator of the entire vector $\sigma^2 = (\sigma_1^2, \ldots, \sigma_M^2)$ at once by solving the linear system $V = A\hat{\sigma}^2$, where V is the vector of raw row variances V_m and A is the square matrix with $A_{km} = \frac{J_m}{J_k}$ if $k \in I(m)$ and 0 otherwise.

The next step is to determine uncertainties for the estimated variance components. Once again, there is an extensive literature on this; the basic method is to express each estimate $\hat{\sigma}_m^2$ as a sum and difference of independent random variables whose distributions are proportional to χ^2, and then to compute the variance of the estimate. The difficulty of this standard approach is in working with this combination-of-χ^2 distribution.

Instead, we evaluate the uncertainties of the estimated variance components by simulation, performing the following steps 1000 times: (1) simulate uncertainty in each raw row variance V_m by multiplying by a random variable of the form $(df)_m/\chi^2_{(df)_m}$, (2) solve for $\hat{\sigma}^2$ in $V = A\hat{\sigma}^2$, (3) constrain the solution to be nonnegative, and (4) compute the 50% and 95% intervals from the constrained simulation draws. This simulation has a parametric bootstrap or Bayesian flavor and is motivated by the approximate equivalence between repeated-sampling and Bayesian inferences [see, e.g., DeGroot (1970) and Efron and Tibshirani (1993)].

Conditional on the simulation for σ, we can now estimate the finite-population standard deviations s_m. As discussed in Section 3.5, the data provide additional information about these, and so our intervals for s_m will be narrower than for σ_m, especially for variance components with few degrees of freedom. Given σ, the parameters $\beta_j^{(m)}$ have a multivariate normal distribution (in Bayesian terms, a conditional posterior distribution; in classical terms, a predictive distribution). The resulting inference for each s_m can be derived from (8), computing either by simulation of the β's or by approximation with the χ^2 distribution. Finally, averaging over the simulations of σ yields predictive inferences about the s_m's.

4.2. *Bayesian inference.* To estimate the variance components using Bayesian methods, one needs a probability model for the regression coefficients $\beta_j^{(m)}$ and the variance parameters σ_m. The standard model for β's is independent normal, as given by (7). In our ANOVA formulation (5) or (6), the regression error terms are just the highest-level interactions, $\beta_j^{(M)}$, and so the distributions (7) include the likelihood as well as the prior distribution. For generalized linear models, the likelihood can be written separately (see Section 7.2 for an example).

The conditionally conjugate hyperprior distributions for the variances can be written as scaled inverse-χ^2:

$$\sigma_m^2 \sim \text{Inv-}\chi^2(\nu_m, \sigma_{0m}^2).$$

A standard noninformative prior distribution is uniform on σ, which corresponds to each $\nu_m = -1$ and $\sigma_{0m} = 0$ [see, e.g., Gelman, Carlin, Stern and Rubin (1995)]. For values of m in which J_m is large (i.e., rows of the ANOVA table corresponding to many linear predictors), σ_m is essentially estimated from data. When J_m is small, the flat prior distribution implies that σ is allowed the possibility of taking on large values, which minimizes the amount of shrinkage in the effect estimates.

More generally, it would make sense to model the variance parameters σ_m themselves, especially for complicated models with many variance components (i.e., many rows of the ANOVA table). Such models are a potential subject of future research; see Section 8.2.

With the model as set up above, the posterior distribution for the parameters (β, σ) can be simulated using the Gibbs sampler, alternately updating the vector β given σ with linear regression, and updating the vector σ from the independent inverse-χ^2 conditional posterior distributions given β. The only trouble with this Gibbs sampler is that it can get stuck with variance components σ_m near zero. A more efficient updating reparameterizes into vectors γ, α and τ, which are defined as follows:

$$\begin{aligned} \beta_j^{(m)} &= \alpha_m \gamma_j^{(m)}, \\ \sigma_m &= \alpha_m \tau_m. \end{aligned} \tag{12}$$

The model can be then expressed as

$$\begin{aligned} y &= X(\alpha\gamma), \\ \gamma_j^{(m)} &\sim \text{N}(0, \tau_m^2) \qquad \text{for each } m, \\ \tau_m^2 &\sim \text{Inv-}\chi^2(\nu_m, \sigma_{0m}^2). \end{aligned}$$

The auxiliary parameters α are given a uniform prior distribution, and then this reduces to the original model [see Boscardin (1996), Meng and van Dyk (1997), Liu, Rubin and Wu (1998), Liu and Wu (1999) and Gelman (2004)]. The Gibbs sampler then proceeds by updating γ (using linear regression with

n data points and $\sum_{m=0}^{M} J_m$ predictors), α (linear regression with n data points and M predictors) and τ^2 (independent inverse-χ^2 distributions). The parameters in the original parameterization, β and σ, can then be recomputed from (12) and stored at each step.

Starting points for the Bayesian computation can be adapted from the classical point estimates for σ^2 and their uncertainties from Section 4.1. The only difficulty is that the variance parameters cannot be set to exactly zero. One reasonable approach is to replace any σ_m^2 of zero by a random value between zero and $|V_m - \widehat{EV}_m|$, treating this absolute value as a rough measure of the noise level in the estimate. Generalized linear models can be computed using this Gibbs sampler with Metropolis jumping for the nonconjugate conditional densities [see, e.g., Gelman, Carlin, Stern and Rubin (1995)] or data augmentation [see Albert and Chib (1993) and Liu (2002)]. In either case, once the simulations have approximately converged and posterior simulations are available, one can construct simulation-based intervals for all the parameters and for derived quantities of interest such as the finite-population standard deviations s_m defined in (8).

When we use the uniform prior density for the parameters σ_m, the posterior distributions are proper for batches m with at least two degrees of freedom. However, for effects that are unique or in pairs [i.e., batches for which $(df)_m = 1$], the posterior density for the corresponding σ_m is improper, with infinite mass in the limit $\sigma_j \to \infty$ [Gelman, Carlin, Stern and Rubin (1995), Exercise 5.8], and so the coefficients $\beta_j^{(m)}$ in these batches are essentially being estimated via maximum likelihood. This relates to the classical result that shrinkage estimation dominates least squares when estimating three or more parameters in a normal model [James and Stein (1961)].

5. A new ANOVA table. There is room for improvement in the standard analysis of variance table: it is read in order to assess the relative importance of different sources of variation, but the numbers in the table do not directly address this issue. The sums of squares are a decomposition of the total sum of squares, but the lines in the table with higher sums of squares are not necessarily those with higher estimated underlying variance components. The mean square for each row has the property that, if the corresponding effects are all zero, its expectation equals that of the error mean square. Unfortunately, if these other effects are *not* zero, the mean square has no direct interpretation in terms of the model parameters. The mean square is the variance explained per parameter, which is not directly comparable to the parameters s_m^2 and σ_m^2, which represent underlying variance components.

Similarly, statistical significance (or lack thereof) of the mean squares is relevant; however, rows with higher F-ratios or more extreme p-values do *not* necessarily correspond to batches of effects with higher estimated magnitudes. In summary, the standard ANOVA table gives all sorts of information, but nothing to directly compare the listed sources of variation.

Our alternative ANOVA table presents, for each source of variation m, the estimates and uncertainties for s_m, the standard deviation of the coefficients corresponding to that row of the table. In addition to focusing on estimation rather than testing, we display the estimates and uncertainties graphically. Since the essence of ANOVA is comparing the importance of different rows of the table, it is helpful to allow direct graphical comparison, as with tabular displays in general [see Gelman, Pasarica and Dodhia (2002)]. In addition, using careful formatting, we can display this in no more space than is required by the classical ANOVA table.

Figure 3 shows an example with the split-plot data that we considered earlier. For each source of variation, the method-of-moments estimate of s_m is shown by a point, with the thick and thin lines showing 50% and 95% intervals from the simulations. The point estimates are not always at the center of the intervals because of edge effects caused by the restriction that all the variance components be nonnegative. In an applied context it might make sense to use as point estimates the medians of the simulations. We display the moments estimates here to show the effects of the constrained inference in an example where uncertainty is large.

In our ANOVA table, the inferences for all the variance components are simultaneous, in contrast to the classical approach in which each variance component is tested under the model that all others, except for the error term, are zero. Thus, the two tables answer different inferential questions. We would argue that the simultaneous inference is more relevant in applications. However, if the classical p-values are of interest, they could be incorporated into our graphical display.

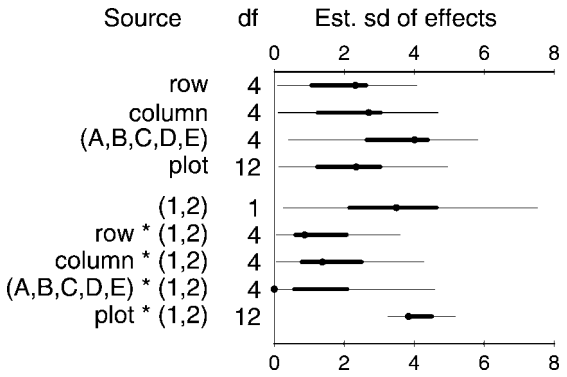

FIG. 3. *ANOVA display for a split-plot latin square experiment* (*cf. to the classical ANOVA, which is the final table in Figure* 1). *The points indicate classical variance component estimates, and the bars display* 50% *and* 95% *intervals for the finite-population standard deviations* σ_m. *The confidence intervals are based on simulations assuming the variance parameters are nonnegative*; *as a result, they can differ from the point estimates, which are based on the method of moments, truncating negative estimates to zero.*

6. Fixed and random effects. A persistent point of conflict in the ANOVA literature is the appropriate use of fixed or random effects, an issue which we must address since we advocate treating *all* batches of effects as sets of random variables. Eisenhart (1947) distinguishes between fixed and random effects in estimating variance components, and this approach is standard in current textbooks [e.g., Kirk (1995)]. However, there has been a stream of dissenters over the years; for example, Yates (1967):

> ... whether the factor levels are a random selection from some defined set (as might be the case with, say, varieties), or are deliberately chosen by the experimenter, does not affect the logical basis of the formal analysis of variance or the derivation of variance components.

Before discussing the technical issues, we briefly review what is meant by fixed and random effects. It turns out that different—in fact, incompatible—definitions are used in different contexts. [See also Kreft and de Leeuw (1998), Section 1.3.3, for a discussion of the multiplicity of definitions of fixed and random effects and coefficients, and Robinson (1998) for a historical overview.] Here we outline five definitions that we have seen:

1. Fixed effects are constant across individuals, and random effects vary. For example, in a growth study, a model with random intercepts α_i and fixed slope β corresponds to parallel lines for different individuals i, or the model $y_{it} = \alpha_i + \beta t$. Kreft and de Leeuw [(1998), page 12] thus distinguish between fixed and random coefficients.
2. Effects are fixed if they are interesting in themselves or random if there is interest in the underlying population. Searle, Casella and McCulloch [(1992), Section 1.4] explore this distinction in depth.
3. "When a sample exhausts the population, the corresponding variable is *fixed*; when the sample is a small (i.e., negligible) part of the population the corresponding variable is *random*" [Green and Tukey (1960)].
4. "If an effect is assumed to be a realized value of a random variable, it is called a random effect" [LaMotte (1983)].
5. Fixed effects are estimated using least squares (or, more generally, maximum likelihood) and random effects are estimated with shrinkage ["linear unbiased prediction" in the terminology of Robinson (1991)]. This definition is standard in the multilevel modeling literature [see, e.g., Snijders and Bosker (1999), Section 4.2] and in econometrics.

   In the Bayesian framework, this definition implies that fixed effects $\beta_j^{(m)}$ are estimated conditional on $\sigma_m = \infty$ and random effects $\beta_j^{(m)}$ are estimated conditional on σ_m from the posterior distribution.

Of these definitions, the first clearly stands apart, but the other four definitions differ also. Under the second definition, an effect can change from fixed to random with a change in the goals of inference, even if the data and design

are unchanged. The third definition differs from the others in defining a finite population (while leaving open the question of what to do with a large but not exhaustive sample), while the fourth definition makes no reference to an actual (rather than mathematical) population at all. The second definition allows fixed effects to come from a distribution, as long as that distribution is not of interest, whereas the fourth and fifth do not use any distribution for inference about fixed effects. The fifth definition has the virtue of mathematical precision but leaves unclear when a given set of effects should be considered fixed or random. In summary, it is easily possible for a factor to be "fixed" according to some of the definitions above and "random" for others. Because of these conflicting definitions, it is no surprise that "clear answers to the question 'fixed or random?' are not necessarily the norm" [Searle, Casella and McCulloch (1992), page 15].

One way to focus a discussion of fixed and random effects is to ask how inferences change when a set of effects is changed from fixed to random, with no change in the data. For example, suppose a factor has four degrees of freedom corresponding to five different medical treatments, and these are the only existing treatments and are thus considered "fixed" (according to definitions 2 and 3 above). Suppose it is then discovered that these are part of a larger family of many possible treatments, and so it is desired to model them as "random." In the framework of this paper, the inference about these five parameters $\beta_j^{(m)}$ and their finite-population and superpopulation standard deviations, s_m and σ_m, will not change with the news that they actually are viewed as a random sample from a distribution of possible treatment effects. But the superpopulation variance now has an important new role in characterizing this distribution. The difference between fixed and random effects is thus not a difference in inference or computation but in the ways that these inferences will be used. Thus, we strongly *disagree* with the claim of Montgomery [(1986), page 45] that in the random effects model, "knowledge about particular [regression coefficients] is relatively useless."

We prefer to sidestep the overloaded terms "fixed" and "random" with a cleaner distinction by simply renaming the terms in definition 1 above. We define effects (or coefficients) in a multilevel model as *constant* if they are identical for all groups in a population and *varying* if they are allowed to differ from group to group. For example, the model $y_{ij} = \alpha_j + \beta x_{ij}$ (of units i in groups j) has a constant slope and varying intercepts, and $y_{ij} = \alpha_j + \beta_j x_{ij}$ has varying slopes and intercepts. In this terminology (which we would apply at any level of the hierarchy in a multilevel model), varying effects occur in batches, whether or not the effects are interesting in themselves (definition 2), and whether or not they are a sample from a larger set (definition 3). Definitions 4 and 5 do not arise for us since we estimate all batches of effects hierarchically, with the variance components σ_m estimated from data.

7. Examples. We give two examples from our own consulting and research where ANOVA has been helpful in understanding the structure of variation in a

dataset. Section 7.1 describes a multilevel linear model for a full-factorial dataset, and Section 7.2 describes a multilevel logistic regression.

From a classical perspective of inference for variance components, these cases can be considered as examples of the effectiveness of automatically setting up hierarchical models with random effects for each row in the ANOVA table. From a Bayesian perspective, these examples demonstrate how the ANOVA idea—batching effects into rows and considering the importance of each batch—applies outside of the familiar context of hypothesis testing.

7.1. *A five-way factorial structure*: *Web connect times.* Data were collected by an Internet infrastructure provider on connect times—the time required for a signal to reach a specified destination—as processed by each of two different companies. Messages were sent every hour for 25 consecutive hours, from each of 45 locations to four different destinations, and the study was repeated one week later. It was desired to quickly summarize these data to learn about the importance of different sources of variation in connect times.

Figure 4 shows a classical ANOVA of logarithms of connect times using the standard factorial decomposition on the five factors: destination ("to"), source ("from"), service provider ("company"), time of day ("hour") and week. The data have a full factorial structure with no replication, so the full five-way interaction, at the bottom of the table, represents the "error" or lowest-level variability. The ANOVA reveals that all the main effects and almost all the interactions are statistically significant. However, as discussed in Section 5, it is difficult to use these significance levels, or the associated sums of squares, mean squares or F-statistics, to *compare* the importance of the different factors.

Figure 5 shows the full multilevel ANOVA display for these data. Each row shows the estimated finite-population standard deviation of the corresponding group of parameters, along with 50% and 95% uncertainty intervals. We can now immediately see that the lowest-level variation is more important in variance than any of the factors except for the main effect of the destination. `Company` has a large effect on its own and, perhaps more interestingly, in interaction with `to`, `from`, and in the three-way interaction.

The information in the multilevel display in Figure 5 is *not* simply contained in the mean squares of the classical ANOVA table in Figure 4. For example, the effects of `from * hour` have a relatively high estimated standard deviation but a relatively low mean square (see, e.g., `to * week`).

Figure 5 does not represent the end of any statistical analysis; for example, in this problem the analysis has ignored any geographical structure in the "to" and "from" locations and the time ordering of the hours. As is usual, ANOVA is a tool for data exploration—for learning about which factors are important in predicting the variation in the data—which can be used to construct useful models or design future data collection. The linear model is a standard approach to analyzing factorial data; in this context, we see that the multilevel ANOVA display,

Source	Df	Ss	Ms	Fstat	Pvalue
to	3	31193.62	10397.87	26660.68	0.00
from	44	5635.24	128.07	328.39	0.00
company	1	1027.44	1027.44	2634.40	0.00
hour	24	128.74	5.36	13.75	0.00
week	1	3.76	3.76	9.64	0.00
to * from	132	669.56	5.07	13.01	0.00
to * company	3	497.03	165.68	424.80	0.00
to * hour	72	44.00	0.61	1.57	0.00
to * week	3	14.59	4.86	12.47	0.00
from * company	44	1029.74	23.40	60.01	0.00
from * hour	1056	1793.35	1.70	4.35	0.00
from * week	44	426.40	9.69	24.85	0.00
company * hour	24	29.32	1.22	3.13	0.00
company * week	1	13.73	13.73	35.20	0.00
hour * week	24	43.20	1.80	4.62	0.00
to * from * company	132	487.21	3.69	9.46	0.00
to * from * hour	3168	1326.40	0.42	1.07	0.02
to * from * week	132	162.25	1.23	3.15	0.00
to * company * hour	72	38.60	0.54	1.37	0.02
to * company * week	3	6.54	2.18	5.59	0.00
to * hour * week	72	25.91	0.36	0.92	0.66
from * company * hour	1056	745.65	0.71	1.81	0.00
from * company * week	44	139.37	3.17	8.12	0.00
from * hour * week	1056	782.30	0.74	1.90	0.00
company * hour * week	24	24.51	1.02	2.62	0.00
to * from * company * hour	3168	1339.13	0.42	1.08	0.01
to * from * company * week	132	117.49	0.89	2.28	0.00
to * from * hour * week	3168	1308.72	0.41	1.06	0.05
to * company * hour * week	72	31.62	0.44	1.13	0.22
from * company * hour * week	1056	528.34	0.50	1.28	0.00
to * from * company * hour * week	3168	1235.54	0.39		

FIG. 4. *Classical ANOVA table for a* $4 \times 45 \times 2 \times 25 \times 2$ *factorial data structure. The data are logarithms of connect times for messages on the World Wide Web.*

which focuses on variance components, conveys more relevant information than does the classical ANOVA, which focuses on null hypothesis testing.

Another direction to consider is the generalization of the model to new situations. Figure 5 displays uncertainty intervals for the finite-population standard deviations so as to be comparable to classical ANOVA. This makes sense when comparing the two companies and 25 hours, but the "to" sites, the "from" sites and the weeks are sampled from a larger population, and for these generalizations, the superpopulation variances would be relevant.

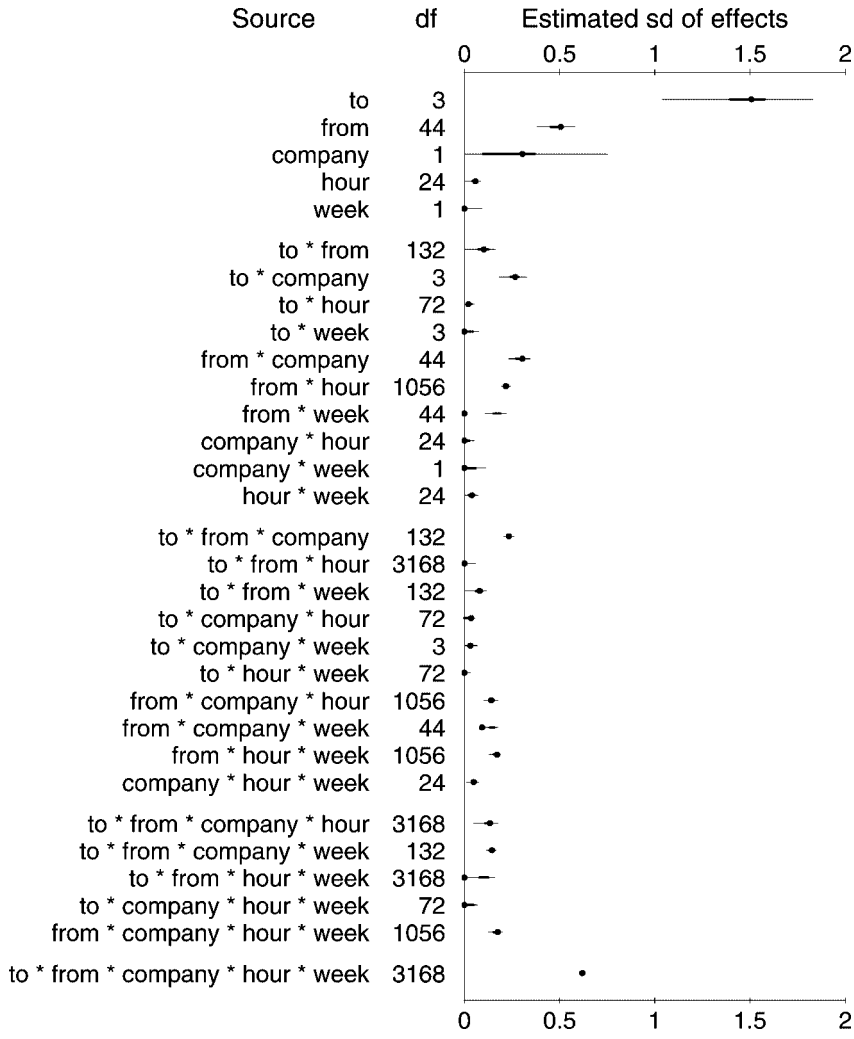

FIG. 5. *ANOVA display for the World Wide Web data* (*cf. to the classical ANOVA in Figure* 4). *The bars indicate* 50% *and* 95% *intervals for the finite-population standard deviations* s_m, *computed using simulation based on the classical variance component estimates. Compared to the classical ANOVA in Figure* 4, *this display makes apparent the magnitudes and uncertainties of the different components of variation. Since the data are on the logarithmic scale, the standard deviation parameters can be interpreted directly. For example,* $s_m = 0.20$ *corresponds to a coefficient of variation of* $\exp(0.2) - 1 \approx 0.2$ *on the original scale, and so the unlogged coefficients* $\exp(\beta_j^{(m)})$ *in this batch correspond to multiplicative increases or decreases in the range of* 20%.

7.2. *A multilevel logistic regression model with interactions*: *political opinions.* Dozens of national opinion polls are conducted by media organizations before every election, and it is desirable to estimate opinions at the levels of individual states as well as for the entire country. These polls are generally based on national

random-digit dialing with corrections for nonresponse based on demographic factors such as sex, ethnicity, age and education [see Voss, Gelman and King (1995)]. We estimated state-level opinions from these polls, while simultaneously correcting for nonresponse, in two steps. For any survey response of interest:

1. We fit a regression model for the individual response given demographics and state. This model thus estimates an average response θ_j for each cross-classification j of demographics and state. In our example, we have sex (male/female), ethnicity (black/nonblack), age (four categories), education (four categories) and 50 states; thus 3200 categories.
2. From the Census, we get the adult population N_j for each category j. The estimated average response in any state s is then $\theta_s = \sum_{j \in s} N_j \theta_j / \sum_{j \in s} N_j$, with each summation over the 64 demographic categories in the state.

We need a large number of categories because (a) we are interested in separating out the responses by state, and (b) nonresponse adjustments force us to include the demographics. As a result, any given survey will have few or no data in many categories. This is not a problem, however, if a multilevel model is fit, as is done automatically in our ANOVA procedure: each factor or set of interactions in the model, corresponding to a row in the ANOVA table, is automatically given a variance component.

As described by Gelman and Little (1997) and Bafumi, Gelman and Park (2002), this inferential procedure works well and outperforms standard survey estimates when estimating state-level outcomes. For this paper, we choose a single outcome—the probability that a respondent prefers the Republican candidate for President—as estimated by a logistic regression model from a set of seven CBS News polls conducted during the week before the 1988 Presidential election. We focus here on the first stage of the estimation procedure—the inference for the logistic regression model—and use our ANOVA tools to display the relative importance of each factor in the model.

We label the survey responses y_i as 1 for supporters of the Republican candidate and 0 for supporters of the Democrat (with undecideds excluded) and model them as independent, with $\Pr(y_i = 1) = \text{logit}^{-1}((X\beta)_i)$. The design matrix X is all 0's and 1's with indicators for the demographic variables used by CBS in the survey weighting: sex, ethnicity, age, education and the interactions of sex $\times$ ethnicity and age $\times$ education. We also include in X indicators for the 50 states and for the four regions of the country (northeast, midwest, south and west). Since the states are nested within regions (which is implied by the design matrix of the regression), no main effects for states are needed. As in our general approach for linear models, we give each batch of regression coefficients an independent normal distribution centered at zero and with standard deviation estimated hierarchically given a uniform prior density.

We fit the model using the Bayesian software Bugs [Spiegelhalter, Thomas, Best and Lunn (2002)], linked to R [R Project (2000) and Gelman (2003)] where

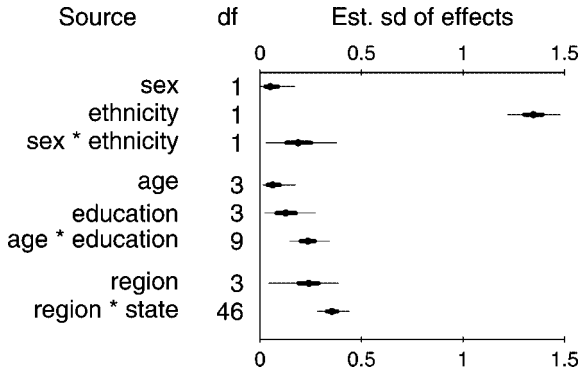

FIG. 6. *ANOVA display for the logistic regression model of the probability that a survey respondent prefers the Republican candidate for the* 1988 *U.S. Presidential election, based on data from seven CBS News polls. Point estimates and error bars show posterior medians,* 50% *intervals and* 95% *intervals of the finite-population standard deviations* s_m*, computed using Bayesian posterior simulation. The demographic factors are those used by CBS to perform their nonresponse adjustments, and states and regions are included because we were interested in estimating average opinions by state. The large effects for ethnicity and the general political interest in states suggest that it might make sense to include interactions*; *see Figure* 7.

we computed the finite-sample standard deviations and plotted the results. Figure 6 displays the ANOVA table, which shows that ethnicity is by far the most important demographic factor, with state also explaining quite a bit of variation.

The natural next step is to consider interactions among the most important effects, as shown in Figure 7. The `ethnicity * state * region` interactions are surprisingly large: the differences between African-Americans and others vary dramatically by state. As with the previous example, ANOVA is a useful tool in understanding the importance of different components of a hierarchical model.

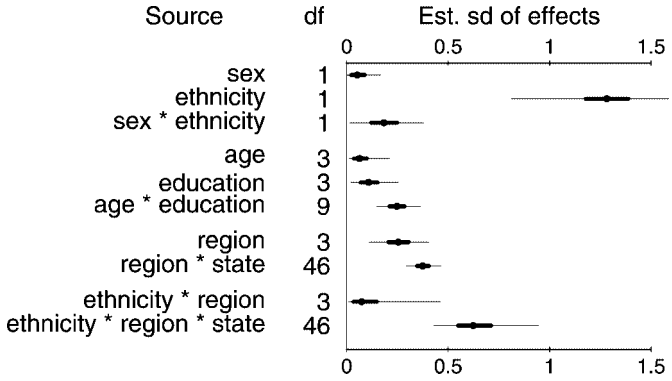

FIG. 7. *ANOVA display for the logistic regression model for vote preferences, adding interactions of ethnicity with region and state. Compare to Figure* 6.

8. Discussion. In summary, we have found hierarchical modeling to be a key step in allowing ANOVA to be performed reliably and automatically. Conversely, the ideas of ANOVA are extremely powerful in modeling complex data of the sort that we increasingly handle in statistics—hence the title of this paper. We conclude by reviewing these points and noting some areas for further work.

8.1. *The importance of hierarchical modeling in formulating and computing ANOVA.* Analysis of variance is fundamentally about multilevel modeling: each row in the ANOVA table corresponds to a different batch of parameters, along with inference about the standard deviation of the parameters in this batch. A crucial difficulty in classical ANOVA and, more generally, in classical linear modeling, is identifying the correct variance components to use in computing standard errors and testing hypotheses. The hierarchical data structures in Section 2.2 illustrate the limitations of performing ANOVA using classical regression.

However, as we discuss in this paper, assigning probability distributions for all variance components automatically gives the correct comparisons and standard errors. Just as a design matrix corresponds to a particular linear model, an ANOVA table corresponds to a particular multilevel batching of random effects. It should thus be possible to fit any ANOVA automatically without having to figure out the appropriate error variances, even for notoriously difficult designs such as split-plots (recall Figure 1).

8.2. *Estimation and hypothesis testing for variance components.* This paper has identified ANOVA with estimation in variance components models. As discussed in Section 3.5, uncertainties can be much lower for finite-population variances s_m^2 than for superpopulation variances σ_m^2, and it is through finite-population variances that we connect to classical ANOVA, in which it is possible to draw useful inferences for even small batches (as in our split-plot Latin square example).

Hypothesis testing is in general a more difficult problem than estimation because many different possible hypotheses can be considered. In some relatively simple balanced designs, the hypotheses can be tested independently; for example, the split-plot Latin square allows independent testing of row, column and treatment effects at the between- and within-plot levels. More generally, however, the test of the hypothesis that some $\sigma_m = 0$ will depend on the assumptions made about the variance components lower in the table. For example, in the factorial analysis of the Internet data in Section 7.1, a test of the `to * from` interaction will depend on the estimated variances for all the higher-level lower interactions including `to * from`, and it would be inappropriate to consider only the full five-way interaction as an "error term" for this test (since, as Figures 4 and 5 show, many of the intermediate outcomes are both statistically significant and reasonably large). Khuri, Mathew and Sinha (1998) discuss some of the options in testing

for variance components, and from a classical perspective these options proliferate for unbalanced designs and highly structured models.

From a Bayesian perspective, the corresponding step is to model the variance parameters σ_m. Testing for null hypotheses of zero variance components corresponds to hierarchical prior distributions for the variance components that have a potential for nonnegligible mass near zero, as has been discussed in the Bayesian literature on shrinkage and model selection [e.g., Gelman (1992), George and McCulloch (1993) and Chipman, George and McCulloch (2001)]. In the ANOVA context such a model is potentially more difficult to set up since it should ideally reflect the structure of the variance components (e.g., if two sets of main effects are large, then one might expect their interaction to be potentially large).

8.3. *More general models.* Our model (7) for the linear parameters corresponds to the default inferences in ANOVA, based on computations of variances and exchangeable coefficients within each batch. This model can be expanded in various ways. Most simply, the distributions for the effects in each batch can be generalized beyond normality (e.g., using t or mixture distributions), and the variance parameters can themselves be modeled hierarchically, as discussed immediately above.

Another generalization is to nonexchangeable models. A common way that nonexchangeable regression coefficients arise in hierarchical models is through group-level regressions. For example, the five rows, columns and possibly treatments in the Latin square are ordered, and systematic patterns there could be modeled, at the very least, using regression coefficients for linear trends. In the election survey example, one can add state-level predictors such as previous Presidential election results. After subtracting batch-level regression predictors, the additive effects for the factor levels in each batch could be modeled as exchangeable. This corresponds to analysis of covariance or contrast analysis in classical ANOVA. Our basic model (6) sets up a regression at the level of the data, but regressions on the hierarchical coefficients (i.e., contrasts) can have a different substantive interpretation as interblock or contextual effects [see Kreft and de Leeuw (1998) and Snijders and Bosker (1999)]. In either case, including contrasts adds another twist in that defining a superpopulation for predictive purposes now requires specifying a distribution over the contrast variable (e.g., in the Latin square example, if the rows are labeled as $-2, -1, 0, 1, 2$, then a reasonable superpopulation might be a uniform distribution on the range $[-2.5, 2.5]$).

More complex structures, such as time-series and spatial models [see Ripley (1981) and Besag and Higdon (1999)], or negative intraclass correlations, cannot be additively decomposed in a natural way into exchangeable components. One particularly interesting class of generalizations of classical ANOVA involves the nonadditive structures of interactions. For example, in the Internet example in Section 7.1 the coefficients in any batch of two-way or higher-level interactions

have a natural gridded structure that is potentially more complex than the pure exchangeability of additive components [see Aldous (1981)].

8.4. *The importance of the ANOVA idea in statistical modeling and inference.* ANOVA is more important than ever because it represents a key idea in statistical modeling of complex data structures—the grouping of predictor variables and their coefficients into batches. Hierarchical modeling, along with the structuring of input variables, allows the modeler easily to include hundreds of predictors in a regression model (as with the examples in Section 7), as has been noted by proponents of multilevel modeling [e.g., Goldstein (1995), Kreft and de Leeuw (1998) and Snijders and Bosker (1999)]. ANOVA allows us to understand these models in a way that we cannot by simply looking at regression coefficients, by generalizing classical variance components estimates [e.g., Cochran and Cox (1957) and Searle, Casella and McCulloch (1992)]. The ideas of the analysis of variance also help us to include finite-population and superpopulation inferences in a single fitted model, hence unifying fixed and random effects. A future research challenge is to generalize our inferences and displays to include multivariate models of coefficients (e.g., with random slopes and random intercepts, which will jointly have a covariance matrix as well as individual variances).

Acknowledgments. We thank Hal Stern for help with the linear model formulation; John Nelder, Donald Rubin, Iven Van Mechelen and the editors and referees for helpful comments; and Alan Edelman for the data used in Section 7.1.

REFERENCES

ALBERT, J. H. and CHIB, S. (1993). Bayesian analysis of binary and polychotomous response data. *J. Amer. Statist. Assoc.* **88** 669–679.

ALDOUS, D. J. (1981). Representations for partially exchangeable arrays of random variables. *J. Multivariate Anal.* **11** 581–598.

BAFUMI, J., GELMAN, A. and PARK, D. K. (2002). State-level opinions from national polls. Technical report, Dept. Political Science, Columbia Univ.

BESAG, J. and HIGDON, D. (1999). Bayesian analysis of agricultural field experiments (with discussion). *J. R. Stat. Soc. Ser. B Stat. Methodol.* **61** 691–746.

BOSCARDIN, W. J. (1996). Bayesian analysis for some hierarchical linear models. Ph.D. dissertation, Dept. Statistics, Univ. California, Berkeley.

BOX, G. E. P. and TIAO, G. C. (1973). *Bayesian Inference in Statistical Analysis*. Addison–Wesley, Reading, MA.

CARLIN, B. P. and LOUIS, T. A. (1996). *Bayes and Empirical Bayes Methods for Data Analysis*. Chapman and Hall, London.

CHIPMAN, H., GEORGE, E. I. and MCCULLOCH, R. E. (2001). The practical implementation of Bayesian model selection. In *Model Selection* (P. Lahiri, ed.) 67–116. IMS, Beachwood, Ohio.

COCHRAN, W. G. and COX, G. M. (1957). *Experimental Designs*, 2nd ed. Wiley, New York.

CORNFIELD, J. and TUKEY, J. W. (1956). Average values of mean squares in factorials. *Ann. Math. Statist.* **27** 907–949.

DEGROOT, M. H. (1970). *Optimal Statistical Decisions*. McGraw-Hill, New York.

EFRON, B. and TIBSHIRANI, R. (1993). *An Introduction to the Bootstrap*. Chapman and Hall, New York.

EISENHART, C. (1947). The assumptions underlying the analysis of variance. *Biometrics* **3** 1–21.

FOX, J. (2002). *An R and S-Plus Companion to Applied Regression*. Sage, Thousand Oaks, CA.

GELMAN, A. (1992). Discussion of "Maximum entropy and the nearly black object," by D. Donoho et al. *J. Roy. Statist. Soc. Ser. B* **54** 72–73.

GELMAN, A. (1996). Discussion of "Hierarchical generalized linear models," by Y. Lee and J. A. Nelder. *J. Roy. Statist. Soc. Ser. B* **58** 668.

GELMAN, A. (2000). Bayesiaanse variantieanalyse. *Kwantitatieve Methoden* **21** 5–12.

GELMAN, A. (2003). Bugs.R: Functions for running WinBugs from R. Available at www.stat.columbia.edu/~gelman/bugsR/.

GELMAN, A. (2004). Parameterization and Bayesian modeling. *J. Amer. Statist. Assoc.* **99** 537–545.

GELMAN, A., CARLIN, J. B., STERN, H. S. and RUBIN, D. B. (1995). *Bayesian Data Analysis*. Chapman and Hall, London.

GELMAN, A. and LITTLE, T. C. (1997). Poststratification into many categories using hierarchical logistic regression. *Survey Methodology* **23** 127–135.

GELMAN, A., PASARICA, C. and DODHIA, R. M. (2002). Let's practice what we preach: Turning tables into graphs. *Amer. Statist.* **56** 121–130.

GEORGE, E. I. and MCCULLOCH, R. E. (1993). Variable selection via Gibbs sampling. *J. Amer. Statist. Assoc.* **88** 881–889.

GOLDSTEIN, H. (1995). *Multilevel Statistical Models*, 2nd ed. Arnold, London.

GREEN, B. F. and TUKEY, J. W. (1960). Complex analyses of variance: General problems. *Psychometrika* **25** 127–152.

JAMES, W. and STEIN, C. (1961). Estimation with quadratic loss. *Proc. Fourth Berkeley Symp. Math. Statist. Probab.* **1** 361–379. Univ. California Press, Berkeley.

JOHNSON, E. G. and TUKEY, J. W. (1987). Graphical exploratory analysis of variance illustrated on a splitting of the Johnson and Tsao data. In *Design, Data and Analysis by Some Friends of Cuthbert Daniel* (C. Mallows, ed.) 171–244. Wiley, New York.

KHURI, A. I., MATHEW, T. and SINHA, B. K. (1998). *Statistical Tests for Mixed Linear Models*. Wiley, New York.

KIRK, R. E. (1995). *Experimental Design*: *Procedures for the Behavioral Sciences*, 3rd ed. Brooks/Cole, Belmont, MA.

KREFT, I. and DE LEEUW, J. (1998). *Introducing Multilevel Modeling*. Sage, London.

LAMOTTE, L. R. (1983). Fixed-, random-, and mixed-effects models. In *Encyclopedia of Statistical Sciences* (S. Kotz, N. L. Johnson and C. B. Read, eds.) **3** 137–141. Wiley, New York.

LIU, C. (2002). Robit regression: A simple robust alternative to logistic and probit regression. Technical report, Bell Laboratories.

LIU, C., RUBIN, D. B. and WU, Y. N. (1998). Parameter expansion to accelerate EM—the PX-EM algorithm. *Biometrika* **85** 755–770.

LIU, J. and WU, Y. N. (1999). Parameter expansion for data augmentation. *J. Amer. Statist. Assoc.* **94** 1264–1274.

MENG, X.-L. and VAN DYK, D. (1997). The EM algorithm—an old folk-song sung to a fast new tune (with discussion). *J. Roy. Statist. Soc. Ser. B* **59** 511–567.

MONTGOMERY, D. C. (1986). *Design and Analysis of Experiments*, 2nd ed. Wiley, New York.

NELDER, J. A. (1965a). The analysis of randomized experiments with orthogonal block structure. I. Block structure and the null analysis of variance. *Proc. Roy. Soc. London Ser. A* **283** 147–162.

NELDER, J. A. (1965b). The analysis of randomized experiments with orthogonal block structure. II. Treatment structure and the general analysis of variance. *Proc. Roy. Soc. London Ser. A* **283** 163–178.
NELDER, J. A. (1977). A reformulation of linear models (with discussion). *J. Roy. Statist. Soc. Ser. A* **140** 48–76.
NELDER, J. A. (1994). The statistics of linear models: Back to basics. *Statist. Comput.* **4** 221–234.
PLACKETT, R. L. (1960). Models in the analysis of variance (with discussion). *J. Roy. Statist. Soc. Ser. B* **22** 195–217.
R PROJECT (2000). The R project for statistical computing. Available at www.r-project.org.
RIPLEY, B. D. (1981). *Spatial Statistics*. Wiley, New York.
ROBINSON, G. K. (1991). That BLUP is a good thing: The estimation of random effects (with discussion). *Statist. Sci.* **6** 15–51.
ROBINSON, G. K. (1998). Variance components. In *Encyclopedia of Biostatistics* (P. Armitage and T. Colton, eds.) **6** 4713–4719. Wiley, Chichester.
RUBIN, D. B. (1981). Estimation in parallel randomized experiments. *J. Educational Statistics* **6** 377–401.
SARGENT, D. J. and HODGES, J. S. (1997). Smoothed ANOVA with application to subgroup analysis. Technical report, Dept. Biostatistics, Univ. Minnesota.
SEARLE, S. R., CASELLA, G. and MCCULLOCH, C. E. (1992). *Variance Components*. Wiley, New York.
SNEDECOR, G. W. and COCHRAN, W. G. (1989). *Statistical Methods*, 8th ed. Iowa State Univ. Press, Ames, IA.
SNIJDERS, T. A. B. and BOSKER, R. J. (1999). *Multilevel Analysis*. Sage, London.
SPEED, T. P. (1987). What is an analysis of variance? (with discussion). *Ann. Statist.* **15** 885–941.
SPIEGELHALTER, D., THOMAS, A., BEST, N. and LUNN, D. (2002). BUGS: Bayesian inference using Gibbs sampling, version 1.4. MRC Biostatistics Unit, Cambridge, England. Available at www.mrc-bsu.cam.ac.uk/bugs/.
VOSS, D. S., GELMAN, A. and KING, G. (1995). Pre-election survey methodology: Details from eight polling organizations, 1988 and 1992. *Public Opinion Quarterly* **59** 98–132.
YATES, F. (1967). A fresh look at the basic principles of the design and analysis of experiments. *Proc. Fifth Berkeley Symp. Math. Statist. Probab.* **4** 777–790. Univ. California Press, Berkeley.

DEPARTMENT OF STATISTICS
COLUMBIA UNIVERSITY
NEW YORK, NEW YORK 10027
USA
E-MAIL: gelman@stat.columbia.edu

DISCUSSION

BY TUE TJUR

Copenhagen Business School

I am not sure that I agree with the statement that ANOVA is more important than ever. On the contrary, I think that the development during the last 40–50 years has somewhat scaled down the importance of this topic, by separating the

computational aspect from the model aspect. Many of the concepts in classical ANOVA for balanced designs are related to the computations. Of course, it is still easier to do the computations in the balanced case, and balancedness also implies other advantages such as maximal efficiency and exact distributions instead of approximations for the mixed models. But the availability of methods for handling of linear models and mixed models in unbalanced designs has changed the focus. Today, I believe, we are more inclined to think of these models as examples (though very important examples) of statistical models, whereas in the classical approach one could hardly mention ANOVA and multiple regression in the same course or textbook.

To me, modern statistics [as opposed, e.g., to the approach taken by Cochran and Cox (1957)] is characterized by the ultimate focusing on the statistical model as the central object. And this brings me to the main topic of my comment to this article, which is the theory or method presented in Section 3. I must admit that I am rather confused here and that I have not been able to understand much of what is going on. The reason for this is, as I see it, that it is not clear at all what the statistical model is. The basic idea seems to be to let all effects enter formally as random effects. But since the method is claimed to be able to handle fixed effects as well (and even to make the comparisons automatically with the correct standard deviations), there must be something I have missed. The random model is only a tool, it is obviously not the model we want to analyze.

Intuitively, it is not difficult to see that there is some element of truth in this approach. For example, in the machine-treatment example, where treatments are confounded with machines, it is certainly correct that the interesting part of the analysis is equivalent to a simple one-way analysis of the 20 machine averages. But what is not at all clear to me is what the method actually does (a detailed operational description), when it works and why it works. I can see no way of proving this without an explicit statement of the model(s) that we actually want to analyze. My guess is that the validity of this method (whatever it is) can only be proved under assumptions about balancedness and orthogonality. Even here there may be problems, since it is not obvious how a phenomenon like partial confounding of a treatment effect with a block effect can be handled. Probably by the introduction of pseudo factors, but where do they come in?

This is all rather negative, and I would have liked to be more positive because I think one of the declared purposes, to make split-plot and other analyses more understandable to students, is an important one. However, my experience here is that the best way of making these things understandable is to focus on the model rather than the design. The analysis of a split-plot design should, in my opinion, be regarded as no more and no less than the analysis of a mixed model. The implications of balancedness (considerable simplification of the computations, exact confidence intervals for contrasts, exact distributions of test statistics, etc.) are important, but irrelevant to the understanding of the statistical model and the interpretation of its parameters.

REFERENCE

COCHRAN, W. G. and COX, G. M. (1957). *Experimental Designs*, 2nd ed. Wiley, New York.

THE STATISTICS GROUP
COPENHAGEN BUSINESS SCHOOL
SOLBJERG PLADS 3
DK-2000 FREDERIKSBERG
DENMARK
E-MAIL: tt.mes@cbs.dk

DISCUSSION

BY PETER MCCULLAGH[1]

University of Chicago

Factorial models and analysis of variance have been a central pillar of statistical thinking and practice for at least 70 years, so the opportunity to revisit these topics to look for signs of new life is welcome. Given its lengthy history, ANOVA is an unpromising area for new development, but recent advances in computational techniques have altered the outlook regarding spatial factors, and could possibly affect what is done in nonspatial applications of factorial designs. In common with a large number of statisticians, Gelman regards analysis of variance as an algorithm or procedure with a well-defined sequence of computational steps to be performed in fixed sequence. The paper emphasizes tactics, how to "perform ANOVA," how to "set up an ANOVA," how to compute "the correct ANOVA," what software to use and how to use it to best effect. The "solution to the ANOVA problem" proffered in Section 3.2 emphasizes once again, how to do it in the modern hierarchical style. Were it not for the recommendation favoring shrinkage, one might have expected a more accurate descriptive title such as the Joy of ANOVA.

I admire the breezy style, the fresh approach and the raw enthusiasm of this paper. It contains perhaps three points with which I agree, namely the importance of ANOVA, the usefulness of thinking in terms of variance components and a passage in Section 3.4 on near-zero estimated variance components. How we could agree on these points and disagree on nearly everything else takes a good deal of explanation. My own philosophy is that it is usually best to begin with the question or questions, and to tailor the analyses to address those questions. Generally speaking, one expects the answer to depend on the question, and it is unreasonable to ask that the analysis, or even a major part of it, should be the same for all questions asked. In my experience, routine statistical questions

[1]Supported in part by NSF Grant 03-05009.

are less common than questionable statistical routines, so I am loath to make pronouncements about what is and what is not relevant in applied statistics. In one of his least convincing passages Gelman argues that the new methodology does the right thing automatically, even for complicated designs. I am inclined to regard this claim either as a regrettable rhetorical flourish, or as a self-fulfilling statement *defining* the class of designs and factors with which the paper is concerned. In the latter case, there is little left to discuss, except to protest that large segments of analysis of variance and factorial design have been overlooked.

The phrase "random coefficient model" or "varying coefficient model" is one that ought to trigger alarm bells. If x is temperature in °C and x' is temperature in °F, the linear models

$$\beta_0 + \beta_1 x \quad \text{and} \quad \beta_0' + \beta_1' x'$$

are equivalent in the sense that they determine the same subspace and thus the same set of probability distributions. Consider now the model in which $\beta_1 \sim N(\bar{\beta}_1, \sigma_1^2)$ and the corresponding one in which $\beta_1' \sim N(\bar{\beta}_1', \tau_1^2)$. On the observation space, the implied marginal covariances are

$$\sigma^2 I_n + \sigma_1^2 (x x^\top) \quad \text{and} \quad \sigma^2 I_n + \tau_1^2 (x' x'^\top),$$

two linear combinations of matrices spanning different spaces. In other words, these random-coefficient formulations do not determine the same set of distributions. It is only in very special circumstances that a random-effects model constructed in this way makes much sense. Making sense is a property that is intuitively obvious: mathematically it means that the model is a group homomorphism or representation.

Gelman's paper is concerned almost exclusively with simple factorial designs in which the factor effects are plausibly regarded as exchangeable. A batch is not a set of regression coefficients as suggested in Section 3.2, but a set of *effects*, one effect for each factor level, and one set or batch for each factor or interaction. The preceding paragraph shows why the distinction between coefficient and effect matters. If batch were synonymous with subset, the new term would be redundant, so it appears that the effects in a batch are meant to be random. In Section 6, a batch of effects is defined as a set of random variables, which are then treated as exchangeable without comment, as if no other option exists. The grouping by batches is determined by factor levels, which is automatic for simple factorial designs, nested or crossed. However, this is not necessarily the case for more general factorial structures such as arise in fertility studies [Cox and Snell (1981), pages 58–62], tournament models [Joe (1990)], origin-destination designs [Stewart (1948)], import-export models, citation studies [Stigler (1994)] or plant breeding designs in which the same factor occurs twice.

In virtually all of the literature on factorial design and analysis of variance, effects are either fixed or random. No other types are tolerated, and all random

effects are independent and identically distributed, as in Section 6 of the present paper. This regrettable instance of linguistic imperialism makes it difficult to find a satisfactory term for random effects in which the components are random but not independent. Clarity of language is important, and in this instance the jargon has developed in such a way that it has become a major obstacle to communication. My own preference is to address matters of terminology, such as treatment and block factors, fixed and random effects, and so on, by what they imply in a statistical model, as described in the next two paragraphs. The alternative to these definitions is the linguistic quagmire so well documented by Gelman in Section 6.

A treatment factor or classification factor A is a list such that $A(i)$ is the level of factor A on unit i. Usually, the set of levels is finite, and the information may then be coded in an indicator matrix $X = X(A)$, one column for each level. By contrast, a block factor E is an equivalence relation on the units such that $E_{ij} = 1$ if units i, j are in the same block, and zero otherwise. A treatment or classification factor may be converted into a block factor by the forgetful transformation $E = XX^\top$ in which the names of the factor levels are lost. A block factor cannot be transformed into a treatment factor because the blocks are unlabelled. A factor may occur in a linear model in several ways, the most common of which are additively in the mean and additively in the covariance

$$Y \sim N(X\beta, \sigma^2 I_n) \quad \text{or} \quad Y \sim N(\mathbf{1}\mu, \sigma^2 I_n + \sigma_b^2 E). \tag{1}$$

Traditionally, the terms "fixed-effects model" and "random-effects model" are used here, but this terminology is not to be encouraged because it perpetuates the myth that random effects are necessarily independent and identically distributed. Note that I_n is the equivalence relation corresponding to units, and $\sigma^2 I_n$, the variance of the exchangeable random unit effects, is included in both models.

Suppose now that two factors A, B are defined on the same set of units, and that these factors are crossed, $A.B$ denoting the list of ordered pairs. The corresponding block factors may be denoted by E_A, E_B and E_{AB}. Two factors may occur in a linear model in several ways, the conventional factorial models for the mean being denoted by

$$1, \qquad A, \qquad B, \qquad A + B, \qquad A.B,$$

with a similar list of linear block-factor models for the covariances

$$I, \qquad I + E_{AB}, \qquad I + E_A, \qquad I + E_B,$$

$$I + E_A + E_B, \qquad I + E_A + E_B + E_{AB}.$$

Here $A + B$ denotes the vector space of additive functions on the factor levels, whereas $I + E_A + E_B$ denotes the set of nonnegative combinations of three matrices in which the coefficients are called variance components. If a factor occurs in the model for the mean, the associated variance component is not identifiable. For example, if the model for the mean includes $A.B$, a so-called

nonrandom interaction, the variance components associated with E_A, E_B, E_{AB} are not identifiable. However, if the variance model includes the interaction E_{AB}, the additive model $A + B$ for the mean is ordinarily identifiable. These are mathematical statements concerning the underlying linear algebra. Philosophical pronouncements such as "if one main effect is random, the interaction is also random" have no place in the discussion.

The subspace $A \subset \mathcal{R}^n$ determined by a factor is of a very special type: it is also a ring, closed under functional multiplication, with **1** as identity element. A factorial model is also a special type of vector subspace of functions on the units. Each is a representation of the product symmetric group in the tensor product space that is also closed under deletion of levels [McCullagh (2000)]. Each of the variance-component models listed above is also a representation in the same sense, but one in which the subspace consists of certain symmetric functions on ordered pairs of units, that is, symmetric matrices. Specifically, each exchangeable variance-components model is a *trivial* representation in the space of symmetric matrices that is closed under deletion of levels. By contrast, a Taguchi-type model in which the variance depends on one or more factor levels is a representation, but not a trivial representation. This may not be a helpful statement for most student audiences, but it does serve to emphasize the point that factorial subspaces are determined by groups and representations. ANOVA decomposition requires one further ingredient in the form of an inner product on the observation space.

If the term "classical linear regression model" implies independence of components, as Gelman's usage in Section 3.3 suggests, then most of the factorial models described above are not classical. On the other hand, they have been a part of the literature in biometry and agricultural field trials for at least 70 years, so they are not lacking in venerability. For clarity of expression, the term "neoclassical" is used here to include models of the above type, linear in the mean and linear in the covariance. The prefix "neo-" refers to more recent versions, including certain spatial models, spline-smoothing models and Taguchi-type industrial applications in which the primary effect of so-called noise factors [Wu and Hamada (2000)] is on variability. A pure variance-components model is one in which the model for the mean is trivial, that is, the constant functions only. The simplest neoclassical procedure for estimation and prediction is first to compute the variance components using residual maximum likelihood, then to compute regression coefficients by weighted least squares, and then to compute predicted values and related summary statistics. For prediction to be possible, the model must be a family of processes.

The main thrust of Gelman's paper as I understand it is to argue that ANOVA should be performed and interpreted in the context of an additive variance-components model rather than an additive factorial model for the mean. This is the special neoclassical model in which the subspace for the mean is the one-dimensional vector space of constant functions, and all effects and interactions are included as block factors in the covariance function. A joint prior distribution

on the variance components avoids the discontinuity associated with near-zero estimated variance components. Individual treatment effects do not occur as parameters in this model, but they may be estimated by prediction, that is, by computing the conditional mean for a new unit having a given factor level, or the difference between conditional means for two such units. When the factor levels are numerous or nonspecific [Cox (1984)], or ephemeral or faceless [Tukey (1974)], this approach is uncontroversial, and indeed, strongly recommended. However, numerous examples exist in which one or more factors have levels that are not of this type, where inference for a specific treatment contrast or a specific classification contrast is the primary purpose of the experiment. Exchangeability is simply one of many modeling options, sensible in many cases, debatable in others, and irrelevant for the remainder. To my mind, Gelman has failed to make a convincing case that additive models for the mean should be abandoned in favor of a scheme that "automatically gets it right" by assuming that every factor has levels whose effects are exchangeable.

In applications where the factor levels have a spatial or temporal structure, it is best to replace the equivalence matrix E in (1) by a more suitable covariance matrix or generalized covariance function, justifying the neoclassical label. As an extreme example, consider a quantitative covariate, which is simply a factor taking values in the real line. The neoclassical Gaussian model with stationary additive random effects has the form

$$E(Y_i) = \beta_0 + \beta_1 x_i,$$
$$\operatorname{cov}(Y_i, Y_j) = \sigma^2 \delta_{ij} + \sigma_s^2 K(|x_i - x_j|),$$

in which K is a covariance function or generalized covariance function. Exchangeability implies $K(x, x') = \text{const} + \delta(x, x')$, but the more usual choices are Brownian motion in which $K(x, x') = -|x - x'|$, or integrated Brownian motion with $K(x, x') = |x - x'|^3$. The latter is a spline-smoothing model having the property that the predicted mean $E(Y(i^*)|\text{data})$ for a new unit such that $x(i^*) = x$ is a cubic spline in x [Wahba (1990)]. This example may seem far removed from the sorts of factorial designs discussed in the paper, but factor levels are frequently ordered or partially ordered, in which case the argument for exchangeability of effects is not compelling. In principle, one may construct a similar covariance function for the effects of a conventional factor whose levels are ordered or partially ordered. Another option is to assume that the departures from linearity are exchangeable.

In time-series analysis, the spectrum determines a decomposition of the total sum of squares into components, two degrees of freedom for each Fourier frequency. Although there are no factors with identifiable levels in the conventional sense, by any reasonable interpretation of the term, this decomposition is an analysis of variance. In fact the key computational idea in the fast Fourier transform has its roots in Yates' algorithm for 2^n factorial designs, so the

similarities are more than superficial. With this in mind, it is hard to understand Gelman's claim in Section 8 that analysis of variance is fundamentally about multilevel modeling. The canonical decomposition of the whole space as the direct sum of two-dimensional subspaces, one for each frequency, is a consequence of stationarity, or the group of translations. Any connection with exchangeability or the batching of coefficients is purely superficial.

Gelman's paper is a courageous attempt to reformulate a central part of applied statistics emphasizing Bayesian hierarchical modeling. Anyone who has taught factorial design and analysis at the graduate level will understand the constant difficult and sometimes painful struggle to achieve a reasonable and balanced attitude to the subject with its myriad and varied applications. Initially one tries to distill rules and extract common threads from typical applications, only to find later that all applications are atypical in one way or another. My own experience is that the state of this battle evolves as a process: it may converge, but it is not convergent to a fixed attitude or state of understanding. What seems important at one time often declines into insignificance later. It is clear that Gelman has thought hard about factorial models and ANOVA, and his views have evolved through consulting and teaching over a period of 10–15 years. My hope is that he will continue to think hard about the topic, and my prediction is that his views will continue to evolve for some time to come.

REFERENCES

COX, D. R. (1984). Interaction (with discussion). *Internat. Statist. Rev.* **52** 1–31.

COX, D. R. and SNELL, E. J. (1981). *Applied Statistics*: *Principles and Examples*. Chapman and Hall, London.

JOE, H. (1990). Extended use of paired comparison models with application to chess rankings. *Appl. Statist.* **39** 85–93.

MCCULLAGH, P. (2000). Invariance and factorial models (with discussion). *J. R. Stat. Soc. Ser. B Stat. Methodol.* **62** 209–256.

STEWART, J. Q. (1948). Demographic gravitation: Evidence and application. *Sociometry* **11** 31–58.

STIGLER, S. M. (1994). Citation patterns in the journals of statistics and probability. *Statist. Sci.* **9** 94–108.

TUKEY, J. W. (1974). Named and faceless values: An initial exploration in memory of Prasanta C. Mahalanobis. *Sankhyā Ser. A* **36** 125–176.

WAHBA, G. (1985). A comparison of GCV and GML for choosing the smoothing parameter in the generalized spline smoothing problem. *Ann. Statist.* **13** 1378–1402.

WAHBA, G. (1990). *Spline Models for Observational Data*. SIAM, Philadelphia.

WU, C. F. J. and HAMADA, M. (2000). *Experiments*: *Planning, Analysis, and Parameter Design Optimization*. Wiley, New York.

DEPARTMENT OF STATISTICS
UNIVERSITY OF CHICAGO
5734 UNIVERSITY AVENUE
CHICAGO, ILLINOIS 60637-1514
USA
E-MAIL: pmcc@galton.uchicago.edu

DISCUSSION

BY JOOP HOX AND HERBERT HOIJTINK

Utrecht University

Bayesian inference and fixed and random effects. Professor Gelman writes "Bayesians see analysis of variance as an inflexible classical method." He adopts a hierarchical Bayesian framework to "identify ANOVA with the structuring of parameters into batches." In this framework he sidesteps "the overloaded terms fixed and random" and defines effects "as constant if they are identical for all groups in a population and varying if they are allowed to differ from group to group." Applying this approach to his first example (a Latin square with five treatments randomized to a 5×5 array of plots), variance components have to be estimated for row, column and treatment effects.

In our opinion, his approach provides an insightful connection between analysis of variance and hierarchical modeling. It renders an informative and easy to interpret display of variance components that is a nice alternative for traditional analysis of variance. However, we wonder whether sidestepping the terms fixed and random is always wise. Furthermore, currently his approach is rather descriptive, and does not contain truly Bayesian inference. Both points will be briefly discussed in the sequel.

To look into the question of fixed versus random and the use of hierarchical modeling, we carried out a small experiment. We constructed a dataset for the example in Section 2.2.2: 20 machines randomly divided into four treatment groups, with six outcome measures for each machine. We asked a statistician who is very skilled in multilevel analysis to analyze these data. The result was a hierarchical multivariate data structure with six outcomes nested within 20 machines, and the treatments coded as dummy variables at the machine level. Variance components were estimated for machines and measures. The treatment effects were tested by constraining all treatments to be equal and using a likelihood-ratio test.

Comparing this procedure with the discussion of this example in Gelman's paper shows that this is not what he had in mind. It certainly contradicts the notion implied in Sections 3.2 and 3.3 that using hierarchical modeling, so to speak, automagically leads to a correct model. In fact, the multilevel analysis approach outlined above makes sense if we assume that the four treatments exhaust all treatments we are interested in. If we assume that there is a population of treatments, or that variations in implementation can lead to different outcomes, we can structure the data as a three-level model, with outcome measures nested within machines nested within treatments, and estimate a variance for the treatments. But even in this case one may ask if this variance is an interesting number to estimate.

We would probably be more interested in the actual treatment effects, or in their differences.

Treating the treatment effects as fixed versus random requires knowledge about the actual design of the study, and a decision on how we should view these treatments. Our point here is that none of this comes automagically. We agree with Gelman that, once such decisions are made, the hierarchical modeling framework is both elegant and powerful. By way of illustration: all models discussed by Gelman for these data can be analyzed using the software MLwiN [Goldstein et al. (1998)]. Given the small sample size, maximum likelihood estimation is not attractive, but MLwiN includes a fully Bayesian inference option. So, at least one widely available multilevel program can be used to analyze these data correctly—after we have specified what we regard as "correct."

Our second point is that, to us, truly Bayesian inference is inseparably connected to the use of informative prior information (excluding the "I know nothing" kind of prior information) and the use of Bayes theorem to quantify the support in the data for different sources of prior information or competing models, that is, to compute posterior probabilities. Consider, for example, the five treatments in Gelman's first "25 plots Latin square" example. Even without the treatment labels, we can come up with several prior expectations that could be interesting in this context. Researcher A evaluated the five treatments and came up with the following model for the treatment effects β:

$$M_A: \quad \beta_1 < \beta_2 < \beta_3 < \beta_4 < \beta_5.$$

Researcher B has a different evaluation and renders the following prior expectation:

$$M_B: \quad \{\beta_1, \beta_4\} < \{\beta_2, \beta_3, \beta_5\}.$$

These models imply that the treatment effect is not (to use Gelman's terminology) a varying effect. Stated otherwise, the assumption that the five treatment effects come from the same distribution does not hold. It is also clear that the treatment effects are not constant, that is, equal for all treatment groups.

This problem can be solved by reinstating the term fixed effect and defining it as a varying effect with components that do not come from the same distribution. However, then the data cannot be analyzed in the framework proposed by Gelman. This does not imply that we disqualify his approach; we only want to stress again that there are situations in which the terms fixed and random effects (varying effects that do not and do come from the same distribution) are still appropriate.

Evaluation of models A and B (for simplicity ignoring the row and column effects of the Latin square design) is possible in a Bayesian framework. First of all, a prior distribution has to be specified for each model. If σ^2 denotes the residual variance of a one-way ANOVA with five groups, this prior could have the form

$$g(\theta|M_m) = g(\beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \sigma^2|M_m) \propto \prod_{i=1}^{5} N(\beta_i|0, 1000)\chi^{-2}(\sigma^2|1, 10)I_{M_m},$$

where i denotes the treatment, the indicator function has the value 1 if the treatment effects are in accordance with the restrictions imposed by model $m = A, B$ and 0 otherwise, and $N(\cdot)$ and $\chi^{-2}(\cdot)$ denote uninformative normal and scaled inverse chi-square distributions, respectively. Note that the resulting prior distributions are informative because the prior expectations formulated by researchers A and B are included using the indicator function. Note also that otherwise the prior is uninformative, and does not differ between treatment effect parameters and competing models. Subsequently, Bayes theorem can be used to compute the posterior probability of models A and B:

$$P(M_m|y) \propto P(y|M_m)P(M_m),$$

where

$$P(y|M_m) = \int_\theta P(y|\theta)g(\theta|M_m)\,d\theta,$$

and y is a vector containing the treatment effects for each of the 25 plots.

In our opinion this approach is truly Bayesian because prior knowledge is formalized in prior distributions and subsequently evaluated using posterior probabilities. This is lacking in Gelman's approach. His prior distribution is uninformative, and there are neither competing models nor different sources of prior information that are evaluated using posterior probabilities.

The remaining question is whether it is a problem that Gelman changes from fixed/random to constant/varying, and, whether it is a problem that his prior distribution is uninformative and that there is no inference in the sense that posterior probabilities are computed. Both the answers no and yes are possible. No, because the approach proposed is valuable in itself. Yes, because (as is hopefully illustrated by the examples given) fixed effects are not necessarily treated optimally if Professor Gelman's approach is used. Also yes, because the mainly descriptive framework presented by Professor Gelman can potentially be modified such that competing models/prior information can be evaluated in a Bayesian manner. Consider once more the Latin square example with row, column and treatment effects. A first model could state that the variance of the row and column effects is zero; a second model that the variance of the row and column effects is smaller than the variance of the treatment effects; and, a third model that the variance of the treatment effects is zero. Potentially, the Bayesian approach can be used to compute posterior probabilities for each of these models. The main problem is the specification of prior distributions. As has been illustrated, this is fairly easy for inequality constrained models. The construction of priors for the comparison of models with zero and nonzero variance components is less straightforward. Bluntly fixing a variance at zero for one model, and giving it an uninformative prior distribution for another model, will lead to an analogue of Lindley's paradox [Lee (1997), pages 130 and 131]. The solution here might be prior distributions based on training data [Berger and Pericchi (1996)], or informative prior distributions.

REFERENCES

BERGER, J. O. and PERICCHI, L. (1996). The intrinsic Bayes factor for model selection and prediction. *J. Amer. Statist. Assoc.* **91** 109–122.

GOLDSTEIN, H., RASBASH, J., PLEWIS, I., DRAPER, D., BROWNE, W., YANG, M., WOODHOUSE, G. and HEALY, M. (1998). *A User's Guide to MLwiN*. Institute of Education, Univ. London.

LEE, P. M. (1997). *Bayesian Statistics*: *An Introduction*. Arnold, London.

P.O. BOX 80140
NL-3508 TC UTRECHT
THE NETHERLANDS
E-MAIL: J.J.C.M.Hox@fss.uu.nl

DISCUSSION
FROM ANOVA TO VARIANCE COMPONENTS

BY ALAN M. ZASLAVSKY

Harvard University

Andrew Gelman's contribution shifts the focus of "Analysis of Variance" (ANOVA) from the limited sense in which it has been commonly used in classical statistics, as a method of testing, to the broader framework of estimation and inference. The term more commonly used in this sense, "variance components modeling," also captures the same spirit. The essential idea is that of constructing distributions by using ideas of exchangeability; each variance component corresponds to a collection of exchangeable effects. This extremely powerful approach to linking the scientific structure of a dataset with a model has been and will continue to be widely applicable.

The transition from ANOVA to variance components modeling shifts attention from decomposition of the variance of the *sample* as in classical ANOVA to decomposition of the variance of the *population*. This shift in focus is appropriate to a world in which scientific questions become increasingly complex and are less frequently answerable through simple designed experiments.

As Gelman notes, in a Bayesian framework the estimation of variance components is relatively automatic; attention can be focused on defining sensible models rather than on constructing designs that can be analyzed easily.

Finally, Gelman's discussion of the manifold definitions of "fixed and random effects" is itself worth the price of admission.

1. Testing and prior distributions. Gelman reserves the issue of model specification, specifically testing of variance components, to a short section (Section 8.2) toward the end of the article (alluding to it only briefly in Section 3.4).

Testing variance components is inherently different from testing a regression coefficient because the null hypothesis $\sigma^2 = 0$ is on the boundary of the parameter space. The difficulties this causes for hypothesis testing in the likelihood framework are well known.

In a Bayesian setting, we might distinguish two purposes of hypothesis testing: determining whether a scientifically interesting conclusion can be drawn with adequate certainty to be worth reporting, and selecting models (omitting unneeded model effects). For "scientific" testing of a regression coefficient we might select a locally uniform prior and then see whether we can at least be adequately confident of its sign, that is, is either $P(\beta > 0)$ or $P(\beta < 0)$ a posteriori close enough to 1? For a variance component, the boundary problem prevents defining a locally uniform prior or applying this "two-sided" approach. (Scale-invariant improper priors typically yield degenerate posteriors in variance components models.)

A "model-mixing" approach combines a point mass at the null with a proper distribution over the remainder of the distribution. I find this unsatisfactory as a default solution, especially in the context of independent priors on the magnitudes of variance components, because it requires informative prior beliefs about both the probability of the null hypotheses $\sigma_m^2 = 0$ (and the various combinations of nulls) and the scale of the variance component if nonzero.

We might avoid the scaling problem by defining prior distributions for *relative* variances $\sigma_m^2 / \sum_{m'} \sigma_{m'}^2$ rather than *absolute* variances σ_m^2. It is more natural to combine this prior with point masses for submodels because the prior probabilities for the submodels are relative to the distribution of a variable that is always scaled on $(0, 1)$; the notion of a "small enough to be scientifically uninteresting" component is also more readily interpretable on a relative scale. Such a prior could also accommodate prior information about relative magnitudes of variance components as suggested by Gelman. Note that Gelman's suggestion of independent uniform priors on each component implies a uniform prior on these relative variance components, that is, a Dirichlet$(\delta, \delta, \ldots, \delta)$ prior with $\delta = 1$, conditional on the sum of the variance components (the marginal variance of the data in an additive model). A prior belief that the variance components should be nearly equal suggests a similar prior with $\delta \gg 1$, and if we believe that a few components should predominate, then we might assume $0 < \delta < 1$. At least in additive models, this prior specification allows us to separate specification of the prior for the marginal variance of the data from that for the ratios of components; such a separation is more difficult with independent priors for the different components.

I find posterior predictive tests [Rubin (1984) and Meng (1994)] a more satisfactory way to test variance components than model mixing: we fix a component at zero, and then by simulating data from its predictive distribution determine whether the observed value of a statistic related to that component in some monotone way is consistent with the predictions of the constrained model. Indeed, the sums-of-squares statistics of the classical ANOVA table are suitable

for such a test. The boundary problem is not an issue with this approach since the reference distribution is determined by simulation rather than by asymptotics, and indeed no prior distribution is required for the variance component being tested. I would conjecture that for balanced data, the sums-of-squares statistics are optimal for posterior predictive testing of the null hypothesis on the corresponding variance components. An interesting research direction would be to prove this conjecture or find a superior statistic for the balanced case, and then to identify better statistics for posterior predictive testing with unbalanced data.

We might also be interested in testing as a means of selecting a model with fewer nuisance parameters, specifically by reducing the number of variance components. For this objective a conservative approach would incline toward retaining as many components as possible, but with a prior distribution that allows their estimates to stay close to zero if there is little or no evidence for nonzero values, possibly by using a small value of δ in a "relative" Dirichlet prior. Sensitivity of inferences of interest to the choice of prior might indicate that the data cannot unequivocally answer the questions of interest.

2. Variance components as a focus of scientific research. Much of the applied multilevel modeling literature treats variance components as nuisance parameters, putting the primary emphasis on estimation and testing of regression coefficients (representing scientifically interesting systematic relationships) or of functions of random effects (small area estimation in survey sampling and official statistics, profiling in health care, "league tables" for schools). An important exception is genetics, in which variance components are the basis for calculations of heritability. In my own applied research, I have found that variance components are also an inherently interesting object of inference. Two examples follow.

An analysis of predictors of administration of clinically appropriate chemotherapy for colorectal cancer estimated a residual variance component for hospital effects, after controlling for measured hospital and patient characteristics [Ayanian et al. (2003)]. To explain the importance of this variation to clinical readers, we noted that the difference between a moderately above-average and a moderately below-average hospital (1 SD above or below average) was about as large as the effect of the most important patient characteristic identified in the model. The large magnitude of this residual variation suggests that measurement of additional hospital characteristics might yield a scientific payoff. Furthermore, quality improvement activities might be directed to bringing lower-performing hospitals closer to the practices of their better-performing peers, consistent with arguments that substantial unexplained variation in rates of use of a medical procedure is in itself evidence of poor quality [Wennberg and Gittelsohn (1982)].

Samples of members of Medicare managed care health plans (private organizations that contract with the U.S. government to provide health care to elderly or disabled individuals) have been administered a survey annually for the last eight years to assess various aspects of the services they receive [Zaslavsky, Zaborski

and Cleary (2004)]. The effects of measured characteristics of individual members or plans are fairly small, and the effects of an individual's characteristics are of little interest because the primary objective of the survey is to evaluate health care systems, not the predictors of an individual's reported experiences. These data were modeled with variance components for three levels of nested geographical units (region, state, Metropolitan Statistical Area or MSA) and for the organizational unit (the health plan). For ratings of "the plan" (primarily reflecting the quality of customer service interactions), the majority of variance (excluding the large bottom-level individual component) was explained by the organizational unit. However, the explainable variance for ratings of doctors was mainly attributable to geographical variation, with a smaller component attributable to the health plan. We interpreted this finding as reflecting the fact that the health plans have more control over customer services provided directly by the plan than over health care. The latter is largely provided by doctors and hospitals that are organizationally independent of the plans and might contract with multiple plans; other studies have shown substantial geographical variation in their practice patterns. This finding has implications for quality improvement, suggesting that interventions to improve quality of care might have to be directed to health care providers in an area rather than trying to identify and improve lower-performing health plans. Similar patterns were identified for other quality dimensions measured in the survey. A further analysis estimated variance components for the geographical units, the plan organization, time (year of survey administration) and interactions of these effects. The time effects were interesting in evaluating the extent to which relative changes in quality might be detected between consecutive years, while the plan by MSA interaction was useful for deciding whether to generate separate estimates by geographical area within large plans serving extensive areas. Estimation of these complex models was made possible by the unusual size of the survey datasets (over 700,000 respondents).

These examples illustrate that despite their relative unfamiliarity in many fields, variance components can be interpreted to nonstatisticians in a scientifically meaningful way.

3. Miscellaneous comments.

Finite population variance components. Gelman correctly notes that inference can be made for both finite-population and superpopulation variances, and that the distinction between these two targets of inference corresponds to the distinction between fixed and random effects. In my experience, estimation of a finite-population variance is relatively rarely of interest. When we are really concerned about a specific set of units (such as alternative treatments in an experiment), we are likely to estimate rankings of and differences among those units as a basis for future action. On the other hand, when estimation of the variance component is intended as part of an inference about a more general law (as is common

in econometric analyses of U.S. state data), we are likely to think of the finite population as part of a larger hypothetical population even if (as the 50 states) they in fact constitute the entire population.

Method of moments. Gelman draws out the connection between classical ANOVA and method-of-moments variance components estimators. Because these estimators are essentially linear combinations of variance statistics that can be directly calculated from the data, they have substantial heuristic value, since maximum-likelihood or Bayesian estimation in complex problems can be too much of a "black box" to yield adequate direct insight into the connection between the data and the parameter estimates. On the other hand, the simple decomposition of Gelman's equation (1) only applies when there are unbiased estimators with independent errors, which is not likely to be the case for complex models with unbalanced datasets.

REFERENCES

AYANIAN, J. Z., ZASLAVSKY, A. M., FUCHS, C. S., GUADAGNOLI, E., CREECH, C. M., CRESS, R. D., O'CONNOR, L. C., WEST, D. W., ALLEN, M. E., WOLF, R. E. and WRIGHT, W. E. (2003). Use of adjuvant chemotherapy and radiation therapy for colorectal cancer in a population-based cohort. *J. Clinical Oncology* **21** 1293–1300.

MENG, X.-L. (1994). Posterior predictive p-values. *Ann. Statist.* **22** 1142–1160.

RUBIN, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Ann. Statist.* **12** 1151–1172.

WENNBERG, J. E. and GITTELSOHN, A. (1982). Variations in medical care among small areas. *Scientific American* **246**(4) 120–134.

ZASLAVSKY, A. M., ZABORSKI, L. B. and CLEARY, P. D. (2004). Plan, geographical, and temporal variation of consumer assessments of ambulatory health care. *Health Services Res.* **39** 1467–1485.

DEPARTMENT OF HEALTH CARE POLICY
HARVARD UNIVERSITY MEDICAL SCHOOL
180 LONGWOOD AVENUE
BOSTON, MASSACHUSETTS 02115-5899
USA
E-MAIL: zaslavsk@hcp.med.harvard.edu

REJOINDER

BY ANDREW GELMAN

Columbia University

ANOVA is more important than ever because we are fitting models with many parameters, and these parameters can often usefully be structured into batches. The

essence of "ANOVA" (as we see it) is to compare the importance of the batches and to provide a framework for efficient estimation of the individual parameters and related summaries such as comparisons and contrasts.

Classical ANOVA is associated with many things, including linear models, F-tests of nested and nonnested interactions, decompositions of sums of squares and hypothesis tests. Our paper focuses on generalizing the assessment, with uncertainty bounds, of the importance of each row of the "ANOVA table." This falls in the general category of data reduction or model summary, and presupposes an existing model (most simply, a linear regression) and an existing batching of coefficients (or more generally "effects," as noted by McCullagh) into named batches.

We thank the discussants for pointing out that more work needs to be done to generalize these ideas beyond classical regression settings with exchangeable batches of parameters. In this rejoinder, we review the essentials of our approach and then address some specific issues raised in the discussions.

1. General comments. McCullagh states that we regard "analysis of variance as an algorithm or procedure with a well-defined sequence of computational steps to be performed in fixed sequence." We appreciate this comment, especially in light of Tjur's complaint that it is not clear what our statistical model is. We would like to split the difference and say they are both right: our procedure is indeed performed in a fixed sequence, and the first step is to take a statistical model that must be specified from the outside.

A statistical model is usually taken to be summarized by a likelihood, or a likelihood and a prior distribution, but we go an extra step by noting that the parameters of a model are typically batched, and we take this batching as an essential part of the model. If a model is already set up in a fully Bayesian form, our ANOVA step is merely to summarize each batch's standard deviation (whether superpopulation or finite-population; this depends on the substantive context, as discussed by Zaslavsky and in our Section 3.5). If only a likelihood is specified, along with a batching of parameters, we recommend fitting a multilevel model with a variance parameter for each batch, to be estimated from data. Yes, this is an automatic step, and yes, this can be inappropriate in particular cases, but we think it is a big step forward from the current situation in which the analyst must supply redundant information to avoid making inappropriate variance comparisons. Tjur also recognizes our goal of making split-plot and other analyses more understandable to students. More generally, we want to set up a framework where nonstudents can get the correct (classical) answer too (and avoid difficulties such as illustrated in Figure 1)!

Our procedure gives an appropriate answer in a wide range of classical problems, and we find the summary in terms of within-batch standard deviations to be more relevant than the usual ANOVA table of sums of squares, mean squares and F-tests. None of the discussants disputes either of these points, but they

all would like to go beyond classical linear models with balanced designs. We provide a more general example in Section 7.2 of our paper (an unbalanced logistic regression) and discuss other generalizations in Section 8.3, but we accept the point that choices remain when implementing ANOVA ideas in nonexchangeable models.

2. The model comes first. The discussants raised several important points that we agree with and regret not emphasizing enough in the paper. First, all the discussants, but especially Tjur and McCullagh, emphasize that the model comes first, and the model should ideally be motivated by substantive concerns, not by mathematical convenience and not by the structure of the data or the design of data collection. As noted above, our conception of ANOVA is a way of structuring inferences given that a model has already been fit and that its parameters are already structured into batches. As McCullagh points out, such batches should not necessarily be modeled exchangeably; we defend our paper's focus on exchangeable batches as they are an extremely important special case and starting point (we assume that the coauthor of an influential book on generalized linear models will appreciate the importance of deep understanding of a limited class of models), but note in Section 8.3 that more can be done.

3. ANOVA is not just for linear models. Our paper emphasized simple models in order to respond to the unfortunate attitude among many statisticians and econometricians that ANOVA is just a special case of linear regression. Sections 2 and 3 of our paper demonstrate that ANOVA can only be thought of this way if "linear regression" is interpreted to include multilevel models. But ANOVA applies in much more general settings.

The vote-preference example of Section 7.2 is closer to our usual practice, which is to use ANOVA ideas to structure and summarize hierarchical models that have dozens of parameters. In this example, we did not fit a multilevel model because of any philosophical predilections or because we had any particular interest in finite populations, superpopulations or variance parameters. Rather, we sought to capture many different patterns in the data (in particular, state effects to allow separate state estimates, and demographic effects to allow poststratification adjustment for survey nonresponse). The multilevel model allows more accurate inferences—the usual partial pooling or Bayesian rationale [see Park, Gelman and Bafumi (2004)]—and ANOVA is a convenient conceptual framework for understanding and comparing the multiplicity of inferences that result. Compare Figures 6 and 7 to the usual tables of regression coefficients (in this case, with over 50 or 100 parameters) to see the practical advantages of our approach.

Various complications arose naturally in the model fitting stage. For example, state effects started out as exchangeable and then we put in region indicators as state-level predictors. We are currently working on extending these models to time series of opinion polls as classified by states and demographics.

So, even in the example of our paper, the modeling is not as automatic as our paper unfortunately made it to appear. What was automatic was the decision to estimate variance parameters for all batches of parameters and to summarize using the estimated standard deviations. A small contribution, but one that moves us from a tangle of over seventy logistic regression parameters (with potential identifiability problems if parameters are estimated using maximum likelihood or least squares) to a compact and informative display that is a starting point to more focused inferential questions and model improvements.

As the discussants emphasize, in a variety of important application areas we can and should go beyond linear models or even generalized linear models, to include nonlinear, nonadditive and nonexchangeable structures. We have found the method of structuring parameters into batches to be useful in many different sorts of models, including nonlinear differential equations in toxicology, where population variability can be expressed in terms of a distribution of person-level parameters [e.g., Gelman, Bois and Jiang (1996)] and Boolean latent-data models in psychometrics, which have batches of parameters indexed by individuals, situations and psychiatric symptoms [e.g., Meulders et al. (2001)]. We cite our own work here to emphasize that we certainly were not trying to suggest that the analysis of variance be restricted to linear models. With modern Bayesian computation, a great deal more is possible, as Hox and Hoijtink point out (and as they have demonstrated in their own applied work). We recommend that practitioners consider ANOVA ideas in summarizing their inferences in these multilevel settings.

4. ANOVA as a supplement to inferences about quantities of interest. In a discussion of the example of our Section 2.2.2 (to which we shall return below), Hox and Hoijtink point out that in any specific application an applied researcher will typically be interested in particular treatment effects, or comparisons of treatment effects, rather than in variance components. We agree and thank these discussants for emphasizing this point. As with classical ANOVA, our goal in summarizing variance components is to understand the model as a whole—how important is each source of variation in explaining the data?—as a prelude or accompaniment to more focused inferences. ANOVA may be "more important than ever" but it is intended to add perspective to, not to take the place of, inference for quantities of substantive interest.

To put it another way: if you are already fitting a statistical model, its parameters can probably be grouped into batches, and it is probably interesting to compare the magnitude of the variation of the parameters in each batch. Recent statistical research has revealed many sorts of useful densely parameterized models, including hierarchical regressions, splines, wavelets, mixture models, image models, and so on. However, it can be tricky to understand such models or compare them when they are fit to different datasets. A long list of parameter estimates and standard errors will not necessarily be helpful, partly for simple

reasons of graphical display, and partly because an ensemble of point estimates will not capture the variance of an ensemble of parameters [Louis (1984)]. The two examples provided by Zaslavsky illustrate ways in which inferences for variance components can be relevant in applied settings.

Tjur asks about partial confounding and other unbalanced designs. We would simply handle these using Bayesian inference. For example, Section 7.2 gives an example of an unbalanced design. Our paper discussed classical estimates for balanced designs, to connect to classical ANOVA and provide fast calculations for problems like the Internet example, but more generally one can always use full Bayesian computations, as pointed out by Hox and Hoijtink.

5. Estimation and hypothesis testing. As Zaslavsky notes, our treatment of ANOVA focuses on estimation of variance components (and, implicitly, of individual coefficients and contrasts via shrinkage estimation), rather than on hypothesis testing. In the application areas in which we have worked, interest has lain in questions of the form, "How important are the effects of factor X?," rather than "Does factor X have an effect?"; see Figure A for an example. (We acknowledge McCullagh's point that our focus is the product of our experiences in environmental, behavioral and social sciences; in other fields, such as genetics, hypotheses of zero effects are arguably more relevant research questions.)

In settings where hypothesis testing is desired, we agree with Zaslavsky that posterior predictive checking is the best approach, since it allows a hypothesis about any subset of parameters to be tested while accounting for uncertainty in the estimation of the other parameters in the model. Posterior predictive checking can also be applied to the multilevel model as a whole to test assumptions such as additivity, linearity and normality.

6. Finite-population and superpopulation summaries. Zaslavsky points out that, in settings where one is interested in generalizing or predicting for new groups, superpopulation summaries are most relevant. We emphasized finite-population summaries in our paper so as to provide more continuity with classical ANOVA. For example, with only five treatment levels, nonzero superpopulation variances are inherently difficult to estimate, a problem that is somewhat ducked by the usual classical analysis which focuses on testing hypotheses of zero variance.

A related issue arises in hierarchical regression models, where the concept of a "contrast" in ANOVA plays the role of a finite-population regression coefficient, while the coefficient in the group-level regression has a superpopulation interpretation. For instance, in the Latin square example shown in Figure 3, suppose we are interested in the linear contrast of treatments A, B, C, D, E. The finite-population contrast is $-2 \cdot \beta_1 + (-1) \cdot \beta_2 + 0 \cdot \beta_3 + 1 \cdot \beta_4 + 2 \cdot \beta_5$, whereas the superpopulation contrast is the appropriately scaled coefficient of $(-2, 1, 0, 1, 2)$ included as a treatment-level predictor in the multilevel model.

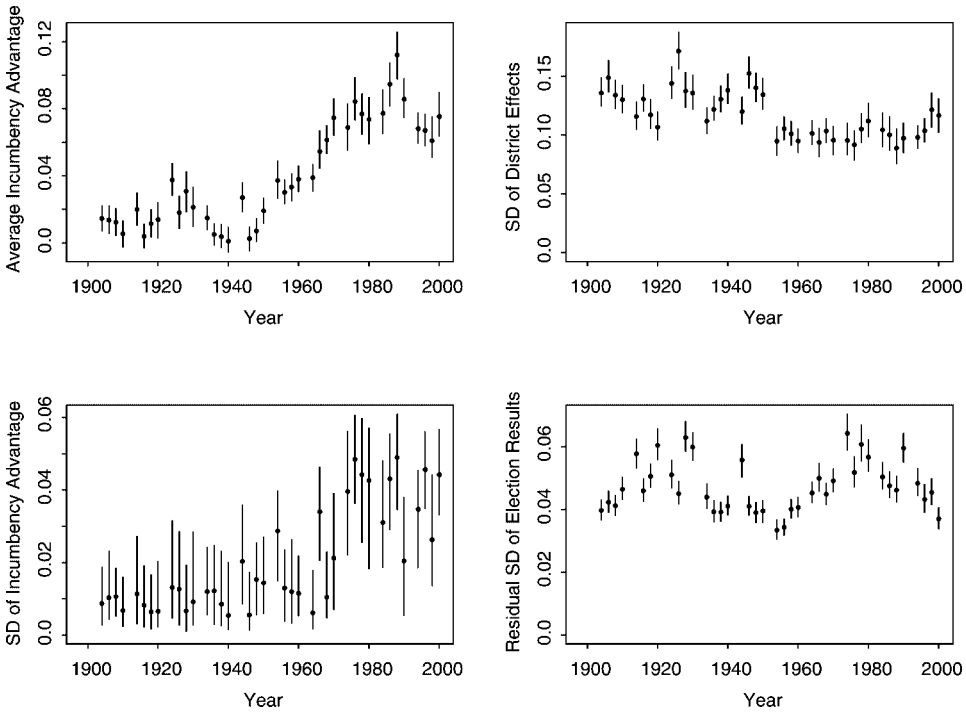

FIG. A. *Estimates and* 95% *intervals for an average treatment effect and three variance parameters for a hierarchical model of elections for the U.S. House of Representatives, fit separately to pairs of successive elections from the past century. The graphs illustrate how we are interested in estimating the magnitude of each source of variation, not simply testing whether effects or variance components equal zero. From Gelman and Huang* (2005).

A key technical contribution of our paper is to disentangle modeling and inferential summaries. A single multilevel model can yield inference for finite-population and superpopulation inferences. For example, in the example of Section 2.2.2, the structure of the problem implies a model with treatment and machine effects, as noted by Hox and Hoijtink. These authors state that their preferred procedure is "not what [we] had in mind," but they do not fully state what their model is. The key question is: what is the population distribution for the four treatment effect parameters? Our recommendation is to fit a normal distribution with mean and standard deviation as hyperparameters estimated from the data. This is the "superpopulation" standard deviation in the terminology of our Section 3.5; fitting the model would also give inferences for the individual treatment effects and their standard deviation. Hox and Hoijtink question whether the variance of the treatment effects is "an interesting number to estimate"; as we note above in our discussion, we agree with this point but find the general comparison of all the variance parameters to be a useful overall summary (as

illustrated by the ANOVA graphs in our paper) without being a replacement for the estimation of individual treatment effects.

To continue with Hox and Hoijtink's discussion of our example: we are not sure what analysis they are suggesting in place of our recommended hierarchical model. One possibility is least-squares estimation for the treatment effect parameters, which would correspond to our hierarchical model with a variance parameter preset to infinity. This seems to us to be inferior to the more general Bayesian approach of treating this variance as a hyperparameter and estimating it from data, and it would also seem to contradict Hox and Hoijtink's opposition to noninformative prior distributions later in their discussion. Another possibility would be a full Bayesian approach with a more informative hyperprior distribution than the uniform distribution that we use. We agree that in the context of any specific problem, a better prior distribution (or, for that matter, a better likelihood) should be available, but we find the normal model useful as a default or starting point.

7. Fixed and random effects. We suspect that statisticians are generally unaware of the many conflicting definitions of the terms "fixed" and "random"; in fact, a reviewer of an earlier version of this paper criticized the multiple definitions in Section 6 as "straw men," which is why we went to the trouble of getting references for each. We are glad that Zaslavsky liked our discussion of fixed and random effects and that McCullagh recognized the "linguistic quagmire."

Hox and Hoijtink would like to define a fixed effect as "a varying effect with components that do not come from the same distribution." This distinction may be important, but we are not hopeful that they will be successful in establishing a new meaning to an already overloaded expression that has at least five other existing interpretations in the statistical literature! Is the phrase "fixed effect" so important that it is worth fighting over this patch of linguistic ground? We use the terms "constant" and "varying" effects because they are unambiguous statements about parameters in a model, and we have the need to communicate with researchers in a wide range of substantive fields. If Hox and Hoijtink find it useful to label sets of effects that are batched but do not come from a common distribution, we recommend they use an unambiguous phrase such as "differently distributed effects" that communicates the concept directly.

Tjur states that our "basic idea seems to be to let all effects enter formally as random effects." We are disappointed to see that he seems to have skipped over Section 6 of our paper! The term "random effect" has no clear (let alone "formal") definition, so we certainly do not consider it to be any part of our basic idea! On the contrary, our basic idea is to recognize that the parameters in a model are not simply a long undifferentiated vector but can be usefully grouped into batches, which in fact are already specified in the classical ANOVA table.

8. Summary: why is ANOVA important now? First, as noted above, if you are already fitting a complicated model, your inferences can be better understood using the structure of that model. We have presented a method for doing so in the context of batches of exchangeable parameters, and we anticipate future developments in other classes of models such as discussed by McCullagh.

Second, if you have a complicated data structure and are trying to set up a model, it can help to use multilevel modeling—not just a simple units-within-groups structure but a more general approach with crossed factors where appropriate. This is the way that researchers in psychology use ANOVA, but they are often ill-served by the classical framework of mean squares and F-tests. We hope that our estimation-oriented approach will allow the powerful tools of Bayesian modeling to be used for the applied goals of inference about large numbers of structured parameters.

ADDITIONAL REFERENCES

GELMAN, A., BOIS, F. Y. and JIANG, J. (1996). Physiological pharmacokinetic analysis using population modeling and informative prior distributions. *J. Amer. Statist. Assoc.* **91** 1400–1412.

GELMAN, A. and HUANG, Z. (2005). Estimating incumbency advantage and its variation, as an example of a before/after study. *J. Amer. Statist. Assoc.* To appear.

LOUIS, T. A. (1984). Estimating a population of parameter values using Bayes and empirical Bayes methods. *J. Amer. Statist. Assoc.* **79** 393–398.

MEULDERS, M., DE BOECK, P., VAN MECHELEN, I., GELMAN, A. and MARIS, E. (2001). Bayesian inference with probability matrix decomposition models. *J. Educational and Behavioral Statistics* **26** 153–179.

PARK, D. K., GELMAN, A. and BAFUMI, J. (2004). Bayesian multilevel estimation with poststratification: State-level estimates from national polls. *Political Analysis* **12** 375–385.

DEPARTMENT OF STATISTICS
COLUMBIA UNIVERSITY
NEW YORK, NEW YORK 10027
USA
E-MAIL: gelman@stat.columbia.edu