Variance Components

P.J. Solomon

Volume 8, pp. 5685–5697

In

Encyclopedia Of Biostatistics
Second Edition
(ISBN 0-470-84907-X)

Edited by

Peter Armitage & Thedore Colton

Variance Components

In a **simple random sample**, one observation is made on each of a number of separate individuals and the variation is assumed to be represented by independent and identically distributed **random variables**, one for each individual. This forms the basis of **regression** and other models widely used in biostatistics. However, there are two ways in which the assumption of a single random component corresponding to each individual might fail to be adequate. In the first, the random variation may have a more complex structure arising from several identifiable sources. The variation is then considered to have multiple components, which we call *components of variance*. This is the classical field of variance components and has a long history dating from the nineteenth century. The second way in which the assumption can fail is when the parameters describing the systematic part of the variation may themselves change randomly, for example, between individuals or groups of individuals. This forms the basis of **hierarchical** or **multilevel modeling** in which the emphasis is on **computer intensive** methods for handling unbalanced or nonnormal data.

We begin this article with three examples to illustrate the key concepts and objectives involved in variance component analysis. Example 1 presents the simplest situation of the balanced one-way model. Example 2 describes a more complex model for microarray data, which involves *nesting* and *cross-classification* and helps distinguish these features. Examples 1 and 2 are classical variance component models. Example 3 outlines a linear **random effects** regression for a marker of HIV/**AIDS** disease and is an example of a multilevel model.

Example 1 *One-way balanced model.* Consider a group of patients, each of whom has a 'true' value of cholesterol say, or blood pressure, denoted by μ_j, $j = 1, \ldots, n_J$. For each patient, one measurement is made by a conditionally unbiased method; this means that for a given patient, μ_j has corresponding observation $Y_j = \mu_j + \varepsilon_j$, where the random term ε_j has mean zero and variance σ_ε^2. We call σ_ε^2 the *component of variance within patients*, which usually represents sampling or measurement error or some such.

Suppose now that the n_J patients are to be regarded as a random sample from a hypothetical infinite population of patients of true mean μ. This situation could arise, for example, in a **clinical trial** in which a homogeneous group of patients has been **randomized** to a treatment and interest centers on the efficacy of that treatment. The mean for patient j becomes a random variable, which can be written as the sum of the overall population mean, μ, and an independent random contribution from the patient, ξ_j. This gives $Y_j = \mu + \xi_j + \varepsilon_j$, where ξ_j has mean zero and variance σ_ξ^2. The latter is called the *component of variance between patients*. It follows that the variance of Y is the sum of two components, $\sigma_\xi^2 + \sigma_\varepsilon^2$, which are not separately estimable without either an external estimate of σ_ξ^2 from other studies or repeated measurements on each patient.

Suppose that several measurements are made on each patient for whom the response is assumed to remain stable. This gives observations

$$Y_j = \mu + \xi_j + \varepsilon_{js} \qquad (1)$$

in which n_S, $s = 1, \ldots, n_S$, repeat observations are nested within patients. This means that observation 1 on patient i is assumed to have no special connection with observation 1 on a different patient k, and so on. The simplest situation assumes that all the random variables ξ and ε are mutually uncorrelated, but such an assumption should not be made uncritically. For example, errors would be uncorrelated if repeated samples were taken from a patient, homogenized, then split into n_S subsamples. Many such considerations relate to the design of the investigation.

This is the balanced one-way model in which there are two components of variance, between-patients and within-patients, each with zero mean. The random variables are usually, although by no means necessarily, assumed independently normally distributed. Repeat observations for a randomly chosen patient are correlated in the one-way model with *intraclass* **correlation** *coefficient* $\rho = \sigma_\xi^2/(\sigma_\xi^2 + \sigma_\varepsilon^2)$. This is a dimensionless measure and such measures are in general useful for formal inference, such as in genetics, but the variance components themselves are more informative as a basis for comparing the spread between and within patients.

Some statisticians prefer to represent variance component models via **covariance matrices** rather than random variables. The covariance matrix of the full $n_J n_S \times 1$ random vector formed by stacking the rows of $\{Y_{js}\}$ into a single column is a block diagonal matrix of the form

$$(\tau_\xi J_{n_S} + \tau_\varepsilon I_{n_S}) \otimes I_{n_J} = \tau_\xi U_\xi + \tau_\varepsilon U_\varepsilon, \qquad (2)$$

where $\otimes$ denotes the Kronecker product [27], and I_{n_S} and J_{n_S} are the $n_S \times n_S$ identity matrix and the matrix all of whose elements are one, respectively; U_ξ, U_ε are associated matrices connected with indicator matrices defining the contribution of the component random variables to the observations (*see* **Matrix Algebra**). This formulation paves the way for a very general version with each separate component of variance identified with its own associated matrix. For interpretation and inference, however, we regard the representation in terms of component random variables as primary and this is the focus of the present article.

Example 2 *A model for cDNA microarray data* (*see* **DNA Sequences**). In cDNA microarrays, known single-stranded DNA clones are robotically spotted out and fixed onto a glass microscope slide. At the same time, two mRNA samples from the cell populations to be compared are reversed transcribed into cDNA and separately labeled with dyes, usually red (Cy5) and green (Cy3). The two labeled targets are mixed together and applied to the microarray slide. During hybridization, single strands in the target solution competitively combine with their complementary base-pair nucleotide sequences spotted on the slide. The relative intensities of red and green at a spot are extracted by image processing the scanned microarray images. The motivation for the technique is that the mRNA in the original cell sample reflects which genes are being used by the cell, and that the intensity ratio at a spot is a measure of the relative abundance of that gene in the two samples. The intensity ratios are usually adjusted for background noise on the slide, normalized to remove systematic sources of variation, transformed to log base 2 to induce approximate normality and additivity of effects, and denoted by the random variable M. For a detailed description of the biological and technical background, see [29].

In a study of osteoarthritis, n bone samples from diseased patients are compared to n bone samples taken from the same site in nondiseased control cadavers. The aim of the investigation is to identify which genes are differentially expressed in the osteoarthritis and control bone samples. In a simplified situation, the patients are assumed to be homogeneous for the **risk factors** age and sex. There is no shortage of slides so each case i is hybridized with each control j, m times. Replicates are assumed to be independent. One model for the observed log intensity ratio for gene g is

$$M_{gijk} = \mu_g + \xi_{gi}^D + \varepsilon_{gi}^D - \xi_{gj}^N - \varepsilon_{gj}^N + \varepsilon_{gijk}, \qquad (3)$$

where μ_g represents the true mean difference in expression of gene g in the two samples and all the remaining terms are independent random variables with zero means. In particular, ξ_{gi}^D and ξ_{gj}^N are crossed random effects specific to the diseased and control individuals with variances $\sigma_{g\xi D}^2$ and $\sigma_{g\xi N}^2$, respectively. The random variable ε_{gi}^D is an error term with component of variance $\sigma_{g\varepsilon D}^2$ specifically associated with the ith diseased case and believed to arise from random errors accumulating through the mRNA extraction, amplification, and labeling steps prior to hybridization; $\sigma_{g\varepsilon N}^2$ is the analogous component of error for the jth control sample. Finally, ε_{gijk} is the measurement error associated with the hybridization, scanning and image processing of patient i with control j and is assumed to have variance $\sigma_{g\varepsilon}^2$ for gene g. The k replicates across slides are nested within the disease-control classification (i, j). The variance of M_{gijk} is then

$$\text{var}(M_{gijk}) = \sigma_{g\xi D}^2 + \sigma_{g\varepsilon D}^2 + \sigma_{g\xi N}^2 + \sigma_{g\varepsilon N}^2 + \sigma_{g\varepsilon}^2. \qquad (4)$$

In practice, it may not be feasible to estimate the separate components of variance in the model, not least because many sources of systematic and random variation in microarray experimentation are still not well understood. In this example, it would be adequate for determining differential expression to combine the sources of error into a single variance component term corresponding to the variability between log intensity ratios across slides for gene g. This illustrates an important general point that it is often adequate to use a model in which many sources of error are combined into a single variance term.

Microarray data analysis is receiving increasing attention from statisticians. Speed and Yang [44] are among the first researchers to critically examine the assumption of independent random variables and replication in this context.

Example 3 *A random effects regression model.* Suppose that a marker of disease progression such as log viral load or CD4 cell count in individuals infected with HIV varies roughly linearly over time in each individual. An initial analysis might be reasonably based on a linear regression with time, in which each individual j has intercept and slope parameters β_0 and β_1, that is,

$$Y_{jt} = \beta_0 + \beta_1 x_t + \varepsilon_{jt}. \tag{5}$$

(*see* **Nonlinear Mixed Effects Models for Longitudinal Data**).

However, a cohort of infected individuals would be very unlikely to have the same parameters. The next step might then be to regard the intercept and slope as responses regressed on individual characteristics, or to consider models in which the parameters themselves have random structure; that is, to model the slope for individual j as $\beta_{1j} = \beta_1 + \xi_{1j}$, where β_1 is the mean slope and ξ_{1j} is a random term, and similarly for the intercept, which we write as $\beta_{0j} = \beta_0 + \xi_{0j}$. In this model, interest focuses on the magnitudes of the random variation of individual responses about their regression line, in the variation in the intercepts and slopes, as well as on explanatory determinants of the regression parameters. The random effects themselves are often assumed to be normally distributed although it may not be possible to test the assumption, and it will nearly always be essential to allow these random terms to be correlated so that σ_ξ^2 denotes the covariance matrix of (ξ_{0j}, ξ_{1j}). These ideas generalize to nonnormal response data and to binary **logistic regression** models in particular.

There are only really two key ideas involved in these examples and in variance component problems generally. The first is the distinction between nesting and cross-classification. This is a qualitative rather than a statistical issue, and is to do with the design and logical structure of the data under study and not with any probabilistic or distributional model assumptions (*see* **Experimental Design**). The second key idea is statistical: are we going to treat the levels of factors as intrinsically interesting (i.e. as **fixed effects**) or are the factors to be regarded as random variables (i.e. as **random effects**) where interest might be in their variances? For example, in genetics, an investigator may want to partition the variability into environmental versus inherited components (*see* **Twin Analysis**).

Both dichotomies are subject-matter considerations. There are some general principles, which can be helpful in deciding whether a factor should be regarded as fixed or random. If the levels of a factor are treatments, for example, different therapies for breast cancer, they would usually be treated as fixed effects. Exceptions arise, such as in a clinical trial comparing the effects of many antibiotics.

The key to variance component analysis is to build models that represent different situations and explain levels of variability that are plausible approximations of what we actually observe. The motivation may be intrinsic interest in the variance components themselves, such as in a comparison of different measuring techniques, or on estimating the precision of the mean or other model parameters. Alternatively, the motivation may be the design of further studies via a *synthesis of variance*, which we discuss below.

History

The idea of partitioning variability can be traced at least as far back as Airy's interest in errors of measurement in astronomy [1]. The more recent systematic study of splitting variation into components dates from R.A. Fisher's introduction of the **analysis of variance**; his original motivation was to improve on the intraclass correlation. There followed periods of intense activity during the last century in biometrical genetics as described by Bulmer [6] (*see* **Polygenic Inheritance**), in the analysis of variability in industrial processes dating from the 1930s work in the cotton industry by Tippett [46] and in the wool industries by Daniels [12], and on error structures especially in randomized experimental designs in the 1950s [8]. Eisenhart made explicit the distinction between fixed and random interpretations of an analysis of variance and introduced this terminology [13].

Much of the early work dealt with balanced data. Henderson, in a long series of papers starting in the 1950s, gave noniterative methods for handling unbalanced data based on equating suitable quadratic forms to their expectation [18, 19]. This more intuitive approach has now largely been replaced by **likelihood**-based methods. Hartley and Rao [17] gave a general matrix formulation and **maximum likelihood** estimation for the unbalanced linear model. The important subsequent generalization of maximum likelihood to REML (reduced, **restricted or**

residual maximum likelihood, which we discuss in the next section) for unbalanced data was developed in detail by Patterson and Thompson [30]. Searle et al. [38] provide a very detailed and systematic account of the normal theory formulations and the associated matrix algebra for balanced and unbalanced data. Rao gives a broad account of normal theory aspects too [35]. Rao and Kleffe [34] emphasize the point **estimation** of variance components using quadratic error **loss**, and we discuss this and other methods of estimation in the next section.

Variance component problems with discrete response data have a long history going back to the Lexis urn models of dispersion associated with the **binomial** distribution; see, for instance, [20] (*see* **Overdispersion**). For an early paper on the **beta-binomial distribution**, see [40]. Greenwood and Yule [16] derived the **negative binomial distribution** as a **Poisson distribution** with an additional source of variation in connection with an analysis of accidents to London bus drivers (*see* **Accident Proneness**). Anscombe [2] compared the theoretical properties of various methods of estimation of its parameters. Cox [9] proposed simple methods for variance components in multiplicative models for Poisson variables.

The literature on multilevel modeling has been steadily growing over the past decade and is now very large. See Goldstein [15] for a thorough discussion. In addition to the references already mentioned, Snijders and Bosker [41] contains important computational work and guidance for fitting random effects and other models and Verbeke and Molenberghs [48] give an extremely thorough account of **linear mixed** models. McCulloch and Searle [28] discuss **generalized, linear, and mixed models** as do Fahrmeir and Tutz [14]. Pinheiro and Bates [32] focus on nonlinear normal theory models (*see* **Nonlinear Mixed Effects Models for Longitudinal Data**).

Variance components arise implicitly or explicitly in many problems in sampling and experimental design. Important applications include industrial processes and reliability studies, genetics, animal and plant breeding, econometrics, the design and analysis of interlaboratory standardization trials, epidemiology, psychometric testing, and education. Khuri and Sahai [23] review developments in variance components analysis to the mid-1980s and include a comprehensive bibliography, and a recent issue of *Statistical Methods in Medical Research* was devoted to variance components [42].

Estimation

The most important and often most difficult issue in variance component problems is the appropriate formulation of a model, or equivalently, the formulation of an analysis of variance table. We begin with the simplest situation described in Example 1. It is well known from the analysis of variance that for balanced systems, there are parallel **orthogonal** decompositions of the data vector, of sums of squares of the components, and of the **degrees of freedom**. The observation vector is decomposed into orthogonal components as

$$Y_{js} = \overline{Y}_{..} + (\overline{Y}_{j.} - \overline{Y}_{..}) + (Y_{js} - \overline{Y}_{j.}), \qquad (6)$$

and if we write the data as one long vector, orthogonality implies that the cross-product terms on the right-hand side vanish.

It is conventional to write out the analysis of variance table for the components and this is shown in Table 1, in which MS denotes Mean Square. Roughly, the mean square measures the sum of squares per dimension for the component. The analysis of variance formulation is entirely structural and does not involve model or distributional assumptions. For interpretation, we bring in the probability model, although we still only need the theory of a simple random sample to derive the key properties, in particular, for equating mean squares to their expected values.

For the one-way balanced arrangement, the first important property is that $E(\mathrm{MS}_\varepsilon) = \sigma_\varepsilon^2$, which only concerns how repeat observations for an individual vary around the true mean for that individual. It is also straightforward to show that $E(\mathrm{MS}_\xi) =$

Table 1 Analysis of variance table for the one-way balanced variance component model

Source	SS	df	
Mean	$\Sigma_{j,s}\overline{Y}_{..}^2$	1	MS
Between individuals	$\Sigma_{j,s}(\overline{Y}_{j.} - \overline{Y}_{..})^2$	$n_J - 1$	MS_ξ
Within individuals	$\Sigma_{j,s}(Y_{js} - \overline{Y}_{j.})^2$	$n_J(n_S - 1)$	MS_ε
Total	$\Sigma_{j,s}Y_{js}^2$	$n_J n_S$	

$n_S\sigma_\xi^2 + \sigma_\varepsilon^2$, from which we deduce σ_ξ^2 is estimated via $(\mathrm{MS}_\xi - \mathrm{MS}_\varepsilon)/n_S$.

If we are interested in the overall mean μ, for instance, to compare the means in two or more groups treated in different ways, we want $E(\overline{Y}_{..}) = \mu$ and $\mathrm{var}(\overline{Y}_{..}) = \sigma_\xi^2/n_J + \sigma_\varepsilon^2/(n_J n_S)$. Hence, a pivot for the estimation of μ is

$$\frac{\overline{Y}_{..} - \mu}{\sqrt{\mathrm{MS}_\xi/(n_J n_S)}}.$$

Assuming the pivot is approximately normally distributed, we can also obtain (approximate) confidence limits for μ.

These estimates are sometimes called the least-squares–based estimators and are **unbiased** estimates of the variance components. The overall approach can be generalized to more complex situations in which estimating equations are formed by equating suitable functions of the data (here sums of squares) to their expectations under the assumed model (*see* **Estimating Functions**). Alternative (biased) estimators are given by the method of maximum likelihood and these are discussed below.

If we make the further assumption that all the random variables are independently normally distributed, several important properties follow that also extend to general balanced cases. The most important is that the two sums of squares and the sample mean are minimal sufficient statistics implying various strong optimum properties, and in particular, that as long as the model is adequate, all we need for analysis are the sums of squares and the mean. The assumption of normality should not be made uncritically however, and some effort should be expended on investigating the sensitivity of the conclusions. We discuss ways of assessing model adequacy later.

Certain exact inferential procedures for the three unknown parameters μ, σ_ξ^2, and σ_ε^2 follow from the assumption of normality. For example, a technical refinement of the pivot for μ is that it then has the **Student *t* distribution** with $n_J - 1$ degrees of freedom. However, only certain combinations of the parameters can be tackled by these procedures, which may not be of substantive interest. For example, we can obtain exact confidence limits for the ratio of variances $\sigma_\xi^2/\sigma_\varepsilon^2$, but not for σ_ξ^2 itself, which is of interest in comparing estimates from two or more similar sets of data, subject to checks of homogeneity. The safest general procedure for doing this is the use of **profile likelihood** or one of its generalizations. There are however simpler and essentially equivalent methods. For example, if T is an approximately unbiased estimate of a positive parameter θ with effective degrees of freedom d, then $\log T$ is approximately normally distributed around mean $\log\theta$ with variance $2/d$, and further issues of analysis are in a normal theory least-squares framework. See [11] for a more thorough discussion of these less standard procedures.

Example 4 *Angiogenesis microarray data.* In a collaborative study, the author has been investigating genes involved in the growth of blood vessels, a process known as angiogenesis. The ability to stimulate new blood vessel growth is a prerequisite for the expansion of a solid tumor and future anticancer treatments are postulated to involve therapy directed to both cancer cells and the expanding vascular system. COX2 (Prostaglandin endoperoxide synthase 2) is a gene known to regulate angiogenesis and cell migration, and served as a control gene in a cDNA microarray experiment comparing mRNA samples from time three hours with time zero. The microarray consisted of a subtracted library of 10 400 clones, each duplicated on the slide. The duplicate spots were printed next to each other and are therefore spatially correlated, but we will ignore this special feature of the data. Four slides were hybridized and we assume that the hybridized slides are independent.

The observed log intensity ratios for COX2 are given in Table 2, which illustrates the data structure for the simple one-way model with two replicate observations. Note that in general, the ordering of the observations within rows is arbitrary. In the notation of Table 1, $n_J = 4$ and $n_S = 2$. The appropriate analysis is based on the pivot for the mean, μ, which under the null hypothesis of no differential expression and the assumption that the log ratios are normally distributed, is t with 3 degrees of freedom.

Table 2 Log intensity ratios for a COX2 in a cDNA microarray experiment with four slides and duplicate spots within slides

Slide	Log ratios M	
1	3.5040	3.4757
2	3.7160	3.7896
3	3.6215	3.7496
4	2.9467	2.8873

Thus, $T = 3.4613/(0.3796/2) = 18.23$ on 3 degrees of freedom. The associated P value is 0.00036 with an estimated 95% confidence interval for the true mean difference in expression $(2.8572, 4.0654)$, indicating that COX2 is significantly upregulated at three hours.

We are ignoring here issues of multiple testing, which can be important in microarray experiments when many thousands of genes are analyzed simultaneously (*see* **Multiple Comparisons**).

Negative estimates: All variances are by definition nonnegative. However, the standard least-squares estimates of the upper variance component in the one-way balanced model are based on differences of mean squares and hence may sometimes be negative. The simplest way to deal with negative estimates arising from this and similar situations is to replace them by zero. For example, we would take $\max\{(\mathrm{MS}_\xi - \mathrm{MS}_\varepsilon)/n_S, 0\}$ as an estimate of σ_ξ^2. There are two qualifications to this recommendation. Firstly, if the mean square between individuals is substantially smaller than the mean square within individuals, this indicates that the data are inconsistent with the model and may be a warning that a systematic effect has been omitted. Alternatively, it may be a warning that important correlations between the random variables have been ignored. Secondly, in an analysis that synthesizes an estimate of σ_ξ^2 from several separate sets of data, such as in a **meta-analysis** of **case–control** studies, then negative values should be retained to avoid systematic error in the pooled estimate.

The procedures described so far extend directly to more complex models provided the data are balanced. In practice, however, data are often not balanced, either by design or as a result of various forms of missingness. The concepts involved are not affected by lack of balance, but the analytical details are. In particular, the decompositions for the balanced case no longer hold and the underlying algebra is more complicated. It is not always obvious how to find the variance estimates for more complicated models and general procedures are required. One very powerful procedure is maximum likelihood for which we find algebraically, or more commonly numerically, the combination of parameter values that maximize the likelihood.

Maximum likelihood and REML: It is well known that the maximum likelihood estimate of the variance in a simple random sample is biased, having divisor n_S rather than $n_S - 1$. In more complex models, the resulting estimates of variance may be entirely unsatisfactory especially if the number of **nuisance parameters** is large, and alternative methods of estimation need to be deployed. The most widely used method and preferred basis for the formal analysis of unbalanced normal models is REML, which maximizes the likelihood of judiciously chosen parts of the data, rather than that of all the data.

REML may be formulated as follows for the one-way analysis. We may apply an **orthogonal** transformation to each individual (or sample) to replace the n_S values by the quantity $\overline{Y}_{j.}\sqrt{n_S}$ and $n_S - 1$ variables, which are independently normally distributed with zero mean and variance σ^2. The contribution of the individual to the likelihood is thus the product of two factors, one depending on μ_j and σ^2, and the other depending only on σ^2 and involving the data only via $\sum(Y_{js} - \overline{Y}_{j.})^2$. In many problems, especially when little is known initially about μ_j, the first factor contains little or no information about σ^2. Thus, for inference about σ^2, we use only the second factor. This leads to a loglikelihood based on $n_J(n_S - 1)$ observations that are independently normally distributed with mean zero and variance σ^2. The corresponding maximum likelihood estimate then has the correct divisor, which is the degrees of freedom within individuals. The same idea can be applied to the general linear mixed model with fixed and random effects.

REML has the advantage of returning the usual least-squares estimates of the variance components for balanced data. It is a particular case of the use of **marginal likelihood** and **conditional** likelihood; see [21] for a general study of both. Barndorff-Nielsen and Cox [3] show that REML is a special case of modified profile likelihood.

Alternative methods of estimation: Powerful and efficient methods for model fitting are important. Indeed, the lack of such methods for unbalanced data held the subject of variance components back until relatively recently. A disadvantage of these developments, however, is that the relationship between the data and the conclusions can be obscure, and for complicated problems, simpler methods may be useful for conceptual clarity and interpretation.

For the unbalanced one-way arrangement, the two simplest procedures are to base the estimation of the upper-level variance component σ_ξ^2 on either the unweighted sum of squares $\sum(\overline{Y}_{j.} - \overline{Y}^{(u)}_{..})^2$, where $\overline{Y}_{j.}$ is the mean of the r_j responses for individual j

and $\overline{Y}^{(u)}_{..}$ is the unweighted average of these means, or on the usual analysis of variance sum of squares $\sum r_j(\overline{Y}_{j.} - \overline{Y}^{(r)}_{..})^2$, where $\overline{Y}^{(r)}_{..}$ is the average of the $\overline{Y}_{j.}$ weighted by the group size. The idea is to decide informally whether the upper or lower component of variance is dominant and to use the unweighted or standard analysis of variance sum of squares as a basis for examining the upper-level component of variance. The same idea can be extended to general models. These simpler approaches are related to the various methods of estimation proposed by Henderson and are described in detail in [38].

An important special case is when $T_1, \ldots, T_{n_J}$ are estimates of a parameter θ obtained from independent sets of data, each with its own internal estimate of error. For example, in combining the results of a number of case–control studies, θ could be the log **odds ratio** for treatment versus control after adjustment by maximum likelihood logistic regression for imbalance with respect to **explanatory variables**, which might be different in the different studies. Note that it is not necessary that the same model is fitted to each group of data, only that the parameter θ has the same interpretation. The estimates may vary more than would be expected on the basis of internal error and it may not be feasible to explain the extra variability as systematic. In this case, we may represent the additional variability as random, and in particular, take as a reasonable approximation $T_j = \theta + \xi_j + \varepsilon_j$, where the ξ and ε are approximately normally distributed and independent. The idea is that a simple analysis helps decide whether a component of variance σ_ξ^2 is necessary, whether there are **outlying** groups, and which of the weighted or unweighted estimates of θ are likely to have high efficiency. Similar arguments apply if θ is a vector. Cox [10] outlines the approach and Cox and Solomon [11] give details of these simpler procedures in applications to nonnormal response data and random effects logistic regression.

We mention briefly another class of methods for estimating variance components known as minimum norm quadratic unbiased estimators (MINQUE) described in detail in [34]. In this and related criteria, low moment assumptions are made about the component random variables and attention focusses on quadratic point estimates that satisfy conditions such as unbiasedness and minimum variance (*see* **Minimum Variance Unbiased (MVU) Estimator**).

Synthesis of Variance

This refers to the process of putting the variance components back together with a view to determining what the variability would be in different sampling situations, or the variance that should be attached to a nonstandard type of comparison. Calculations of this sort are particularly important in designs of systems to achieve a balance between the number of groups or individuals that need to be studied and the number of replicates within each individual.

The simplest example is to estimate the variance of a mean if n_{S_1} repeat observations are to be made on each of n_{J_1} individuals. This is $(\sigma_\varepsilon^2 + n_{S_1}\sigma_\xi^2)/(n_{J_1}n_{S_1})$, which can be estimated. We may be interested to know how much better will the precision be if we take three or four repeat observations on each individual rather than one, say. If the different individuals are very different, that is, σ_ξ^2 is large, then there is little point in replicating more within individuals. But if σ_ε^2 is large relative to σ_ξ^2, there will be an $n_{J_1}n_{S_1}$ effect and increasing the number of replicates will improve the precision of the overall mean.

The estimated synthesized variance of the mean under the new design with n_{S_1} rather than n_S repeated observations within each individual is

$$\frac{1}{n_{S_1}n_{J_1}}\left(\mathrm{MS}_\varepsilon + n_{S_1}\frac{\mathrm{MS}_\xi - \mathrm{MS}_\varepsilon}{n_S}\right),$$

where the observed mean squares and degrees of freedom are those from the original data.

Example 4 revisited: *Angiogenesis microarray data*. The estimated components of variance for the gene COX2 in Example 4 are 0.3731 for the between-slide component and 0.0131 for the within-slide component. In view of the considerations outlined above, increasing the number of replicate spots within a slide would have little impact on the precision as compared with increasing the number of slides hybridized in the experiment.

Components of Covariance and Regression

In the simplest case of a number of groups with a regression of Y, say, on X within each group, work on multilevel modeling has tended to stress the effect of the correlation and additional variation within groups on the regression coefficient of Y on X and its standard error. But if the groups represent **bivariate** populations, there are two regression coefficients, one within groups and another that would be defined by a scatterplot of the means of X and Y. In an early exposition of **analysis of covariance**, Pearson stressed the distinction between these coefficients [31].

To formulate the issues explicitly, consider the bivariate one-way balanced model in which each observation Y_{js} is a 1×2 row vector giving an $n_J n_S \times 2$ data matrix Y. The pairs of observations are

$$Y_{js} = \mu_Y + \xi_j^Y + \varepsilon_{js}^Y,$$

$$X_{js} = \mu_X + \xi_j^X + \varepsilon_{js}^X, \qquad (7)$$

for which there are four variances $\sigma^2_{Y\xi}$, $\sigma^2_{X\xi}$, $\sigma^2_{Y\varepsilon}$ and $\sigma^2_{X\varepsilon}$ as well as covariances $\text{cov}(\xi_j^Y, \xi_j^X)$ and $\text{cov}(\varepsilon_{js}^Y, \varepsilon_{js}^X)$. In the general case with p response variables, each observed random variable is replaced by a set of p components.

As explained above, we can view the bivariate decomposition in two different ways. If Y and X are treated on an equal footing, we have two covariance matrices for the interpretation of associations at the two different levels. Sometimes it may be helpful to estimate separately the two **correlations** $\text{corr}(\xi_j^Y, \xi_j^X)$ and $\text{corr}(\varepsilon_j^Y, \varepsilon_j^X)$. The second possibility is that X should be considered as explanatory to the response Y. Then there are two regression coefficients of Y on X, namely, $\beta_{\xi,YX}$ and $\beta_{\varepsilon,YX}$, regression coefficients from the between- and within-group structure, respectively.

Suppose as an illustration that on a large sample of subjects of stable health and in a narrow age range, measurements are made of blood pressure, Y, and sodium (Na) intake, X. For each subject, the observations are repeated some months later. If we ignore possible time trends, we may consider a one-way analysis. The regression coefficient $\beta_{\varepsilon,YX}$ measures the mean increase in blood pressure Y when the Na intake, X, of a particular subject varies by one unit, for example, 10 mg per day. By contrast $\beta_{\xi,YX}$ is the average difference in the mean blood pressure of two different subjects whose long-run mean Na intakes differ by 10 mg per day. The naive interpretation of $\beta_{\xi,YX}$ would imply that if subjects changed their long-run mean Na intake by 10 mg per day, then there would be a corresponding change in long-run mean blood pressure as determined by $\beta_{\xi,YX}$. The naive interpretation of $\beta_{\varepsilon,YX}$ would imply that individuals increasing their Na intake by 10 mg per day would on average have an increase in blood pressure determined by the regression coefficient.

In an observational study, both interpretations involve substantial assumptions and would be quite speculative. If individuals had been randomized to Na levels on the other hand, the interpretation of the regression coefficients would be unambiguous. In the absence of randomization, however, there may be explanatory variables, observed or unobserved, and long-run features of individuals that are themselves explanatory to both Y and X. These arguments extend to more complicated structures. The difficulties of applying aggregate-level conclusions to individuals in this way is often called **ecological**.

Empirical Bayes

When a frequency probability analysis is based on empirical data with structural assumptions, for example, that certain terms in a regression are random, we call the analysis **empirical Bayes**. No special conceptual issues to do with defining a suitable **prior** probability and so forth are involved.

Classical empirical Bayes analysis proceeds as follows. Consider the univariate one-way model again where, as before, the ξ_j and the ε_{js} are independently normally distributed with zero mean. There are three unknown parameters in the model, $\theta = (\mu, \sigma_\xi^2, \sigma_\varepsilon^2)$. Suppose θ is known and that interest is in the mean of the first group, $\mu + \xi_1$; ξ_1 is an unobserved random variable, which itself partly determines the distribution of the observations. It is therefore appealing, and can be justified formally from various points of view that information about ξ_1 is best summarized by its conditional distribution given the data. This is derived by **Bayes's theorem**. We can show [11, Chapter 3] that the required conditional distribution is normal with mean of the form of an optimally weighted mean obtained from combining the information from the data $\overline{y}_{1.}$ and that from the distribution of ξ_1 around μ, denoted $\tilde{\xi}_1$. In effect, the sample mean $\overline{y}_{1.}$ is shrunk towards the general mean (*see* **Shrinkage**). By the same argument, the estimate of any contrast is obtained by shrinking the sample contrast towards zero (*see* **Shrinkage Estimation**).

It can be shown that if in the originating model, the random variables are not normally distributed, then the above estimates are in a sense the best linear estimates. Viewed as a point predictor of ξ_1, the quantity $\tilde{\xi}_1$ has a property summarized in the term *best linear unbiased predictor* (BLUP; see [37]).

Nonnormal Models

There are broadly parallel developments for the Poisson and binomial distributions with one extra level of variability to those discussed above for continuous random variables. An alternative approach to analysis may be to use (approximate) weighted least-squares methods on the basis of an empirical **transform** of the response variable, for example, the square root or logarithm of Poisson variables and the empirical logistic, probit, or log–log transform of binomial variables.

There may be some loss of efficiency in these approximate procedures. But a more general discussion of variance component models for **generalized linear models** and normal theory **nonlinear regression** is difficult, primarily due to the fact that formally efficient methods of estimation involve high-dimensional integration. The general form of the full likelihood is given by

$$\int \operatorname{lik}(\theta \mid \xi; y)\, \mathrm{d}F(\xi; \tau),$$

where $F(\xi; \tau)$ is the distribution function of ξ depending on parameters τ, which are typically components of variance and their generalizations. In cases without **time series** or similar structure, ξ will consist of independent components so that $F(\xi; \tau)$ factorizes into a product component by component. The integral will factorize into subintegrals but, even so, the dimension of each may be large.

In the special case in which, given the random terms ξ, the observations have an **exponential family** distribution, we obtain a **generalized linear mixed model**. In the simplest formulation, the random effects ξ_j are independently and identically normally distributed with zero mean and q-dimensional variance matrix $D(\tau)$, where τ is a vector of unknown variance components; D is often called the *dispersion matrix*. The conditional independence of the observations within an individual or cluster allows us to write the exact marginal likelihood

$$\operatorname{lik}(\beta, \tau; y) = \prod_{j=1}^{n_J} \int \prod_{s=1}^{r_j} f(y_{js}|\xi_j; \beta) g(\xi_j; \tau)\, \mathrm{d}\xi_j, \tag{8}$$

where g is the link function for the generalized linear model.

Formal inference can be based on maximum likelihood or on **Bayesian** considerations, and there are currently three ways to approach the **numerical integration** problem. The most direct and appealing method is direct or preferably adaptive quadrature. The second, applicable when the integrals can be resolved into a sequence of one-dimensional integrals, is to use an analytical approximation, usually based on a few terms of a Laplace expansion [5, 39, 43]. Such expansions are based on the idea that integrals involving an exponential of a function are dominated by behavior of that function near its maximum. This method can sometimes yield relatively simple interpretable results. Calculation of higher terms in the expansions may be feasible, especially if aided by **computerized algebra**. Higher terms are important to give at least a partial check on the adequacy of the approximations but there is often some uncertainty about the range of applicability of the approximations.

The third method is **Markov chain Monte Carlo (MCMC)**. In the Bayesian version, a **Markov chain** is defined, which has as its equilibrium distribution, the posterior distributions of interest. The chain is then simulated a very large number of times and if the realizations appear to have converged to stationarity, the frequency distribution of realized values, excluding a run-in period, is used to estimate the posterior distributions. MCMC is a powerful and general technique but there is the possibility, in theory at least, that apparent convergence to a stationary state is illusory. Some protection can be achieved by starting the **simulations** from very different initial states.

There are also at least two other approaches to these problems. Lee and Nelder [25, 26] study a notion of h-likelihood in which, in effect, realized values of individual random variables representing portions of variability are treated like unknown parameters. This is likely to be effective when there is substantial information about each such realized value. Another mode of analysis called *penalized* *quasi-likelihood* concentrates on the underlying estimating equations and their justification in a broader setting than a fully parametric one [4, 5, 24, 45] (*see* **Penalized Maximum Likelihood**; **Quasi-likelihood**). Rabe-Hesketh et al. [33] provide a valuable comparison of methods for estimation in generalized linear mixed models.

In **survival** or **event history** data, a random term for each individual with an associated variance component is often called *frailty*. The terminology arises from applications in which the randomly occurring events are failures or adverse reactions of some kind. See **Frailty** for a detailed discussion of these and related models.

Model Assessment and Prediction

Although there is a very large literature on formal and informal tests of model adequacy, little of it is directly relevant to variance component models (*see* **Model Checking**). The most important type of failure of a model stems from omitting a substantial effect, for example, treating a cross-classification as if nested. This destroys the independence assumptions underlying the discussion and is likely to be detected by anomalous behavior of the mean squares, possibly leading to substantial negative estimates of variance.

Outlying observations or individuals can influence the usual quadratic estimates of variance. For example, in the one-way arrangement, an anomalous single observation has a large effect on the estimated component of variance within individuals, but relatively little effect on the estimated component between individuals. In more complex situations, the distinction between outliers at the different levels becomes harder to detect empirically. **Robust** methods provide one way of dealing with outlying observations but do not retain key parameter properties, which are central to variance component analysis, in particular, the additivity of variance as a parameter.

Other important departures from the standard formulation include nonnormality of one or more of the component random variables, or dependence between the variability within an individual and the individual mean. Mild nonnormality of the variances within or between individuals may be of relatively minor concern, but the dependence described above may lead to inappropriate predictions or supplementary analyses. Solomon and Cox [43] suggested a formal analysis in which the nonnormal variances and dependence features are separated.

Further, special topics on model criticism and improvement include the prediction of exceedances, the analysis of **panel** data, fitting more elaborate models, transformations, and study of the distributional form of the underlying random variables. Many of these methods and ideas are discussed in detail in [11]. An important general point when assessing model adequacy is that the analysis should focus on issues that are of substantive importance. For example, discriminating between heterogeneous variances versus constant variances with differing individual means is only worth attempting if the distinction can be given a physical interpretation.

Generalizations and Further Topics

There are many additional areas of current work related to variance components including the following.

Measurement error models: The main emphasis in this article has been on the estimation of variance components as parameters of intrinsic interest. One situation where the real emphasis lies elsewhere and the components of variance are of concern because they affect this primary aspect, is the effect of measurement error in explanatory variables on regression analysis (*see* **Errors in the Measurement of Covariates**). **Measurement error models** have a long history and an extensive literature; see, for example, [7, 36] for a recent application.

Design of investigations: The objective of variance components analysis is the study of patterns of variation as they exist rather than the assessment of interventions under controlled conditions, which is the purpose of formal design of experiments. However, many of the general principles of **experimental design**, and especially those common with the principles of sampling (*see* **Sample Surveys in the Health Sciences**), apply. Khuri [22] gives a systematic review and bibliography of work on design for the estimation of variance components, and Cox and Solomon [11, Chapter 3] present some new ideas.

Finite population aspects: Occasionally, the individuals are not regarded as individuals or as sampled from an infinite population but as from an existing **finite population** of known size, or in particular, as forming the whole of the population in which variation is to be assessed. The finite population variance component is relevant only in very special situations, and in some industrial problems in particular. The importance of distinguishing between finite and infinite populations when defining variance components was first stressed by Daniels in the context of studies of variation in industrial processes [12]. Tukey [47] extended these ideas to sampling a finite population. A formulation relevant to the industrial context is outlined in [11].

Synthesis of studies: In many fields of application, the synthesis of information from several studies is crucial. Variation between studies and interactions of such variation with the treatment effects under investigation may involve representation by components of variance. In biostatistics, the term *overview* or **meta-analysis** is often used and is an integral part of evidence-based medicine. A representation in terms of random effects would only be indicated if no direct explanation of important observed variation in treatment effect is apparently available, such as nonconstancy of the treatment effect being confined to certain contrasts. The use of variance components in meta-analysis is not without controversy.

Acknowledgments

Example 2 is based on joint work in progress with G. Glonek and N. Fazzalari. Example 4 is from collaborative work with C. Hahn, J. Gamble and A. Tsykin. I am grateful to Sir David Cox, FRS, for helpful comments.

References

[1] Airy, G.B. (1861). *On the Algebraic and Numerical Theory of Errors of Observation and the Combination of Observations*. McMillan, London.

[2] Anscombe, F.J. (1950). Sampling theory of the negative binomial and log series distributions, *Biometrika* **37**, 358–382.

[3] Barndorff-Nielsen, O.E. & Cox, D.R. (1994). *Inference and Asymptotics*. Chapman & Hall, London.

[4] Breslow, N.E. & Clayton, D.G. (1993). Approximate inference in generalized linear mixed models, *Journal of American Statistical Association* **88**, 9–25.

[5] Breslow, N.E. & Lin, X. (1995). Bias correction in generalized linear mixed models with a single component of dispersion, *Biometrika* **82**, 81–91.

[6] Bulmer, M.G. (1980). *The Mathematical Theory of Quantitative Genetics*. Oxford University Press, Oxford.

[7] Carroll, R.J., Ruppert, D. & Stefanski, L.A. (1995). *Measurement Error in Nonlinear Models*. Chapman & Hall, London.

[8] Cornfield, J. & Tukey, J.W. (1956). Average values of mean squares in factorials, *Annals of Mathematical Statistics* **27**, 907–949.

[9] Cox, D.R. (1955). Some statistical methods connected with series of events (with discussion), *Journal of the Royal Statistical Society B* **17**, 129–164.

[10] Cox, D.R. (1998). Components of variance: a miscellany, *Statistical Methods in Medical Research* **7**, 3–12.

[11] Cox, D.R. & Solomon, P.J. (2002). *Components of Variance*. Chapman & Hall/CRC, Boca Raton.

[12] Daniels, H.E. (1939). The estimation of components of variance, *Supplement of Journal of Royal Statistical Society* **6**, 186–197.

[13] Eisenhart, C. (1947). The assumptions underlying the analysis of variance, *Biometrics* **47**, 1–21.

[14] Fahrmeir, L. & Tutz, G. (1994). *Multivariate Statistical Modelling Based on Generalized Iinear Models*. Springer, New York.

[15] Goldstein, H. (1995). *Multilevel Statistical Models*. Wiley, New York.

[16] Greenwood, M. & Yule, G.U. (1920). An enquiry into the nature of frequency distributions representative of multiple happenings with particular reference to the occurrence of multiple attacks of disease or of repeated accidents, *Journal of Royal Statistical Society* **83**, 255–279.

[17] Hartley, H.O. & Rao, J.N.K. (1967). Maximum-likelihood estimation for the mixed analysis of variance model, *Biometrika* **54**, 93–108.

[18] Henderson, C.R. (1953). Estimation of variance and covariance components, *Biometrics* **9**, 226–252.

[19] Henderson, C.R. (1975). Best linear unbiased estimation and prediction under a selection model, *Biometrics* **31**, 423–447.

[20] Johnson, N.L., Kotz, S. & Kemp, A.W. (1993). *Univariate Discrete Distributions*, 2nd Ed. John Wiley & Sons, New York.

[21] Kalbfleisch, J.D. & Sprott, D.A. (1970). Application of likelihood methods to models involving large numbers of parameters (with discussion), *Journal of Royal Statistical Society B* **32**, 175–208.

[22] Khuri, A.I. (2000). Designs for variance component estimation: past and present, *International Statistical Review* **68**, 311–322.

[23] Khuri, A.I. & Sahai, H. (1985). Variance components analysis: a selective bibliography and survey, *International Statistical Review* **53**, 279–300.

[24] Laird, N. (1978). Empirical Bayes methods for two-way contingency tables, *Journal of American Statistical Association* **65**, 581–590.

[25] Lee, Y. & Nelder, J.A. (1996). Hierarchical generalized linear models (with discussion), *Journal of Royal Statistical Society B* **58**, 619–678.

[26] Lee, Y. & Nelder, J.A. (2001). Hierarchical generalized linear models: a synthesis of generalized linear models, random-effect models and structural dispersions, *Biometrika* **88**, 987–1006.

[27] McCullagh, P. (1987). *Tensor Methods in Statistics*. Chapman & Hall, London.

[28] McCulloch, C.E. & Searle, S.R. (2001). *Generalized, Linear, and Mixed Models*. Wiley, New York.

[29] Nguyen, D.V., Arpat, A.B., Wang, N. & Carroll, R.J. (2002). DNA microarray experiments: biological and technical aspects, *Biometrics* **58**, 701–717.

[30] Patterson, H.D. & Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal, *Biometrika* **58**, 545–554.

[31] Pearson, E.S. (1932). Discussion of paper by B.H. Wilsdon, *Supplement of the Journal of the Royal Statistical Society* **1**, 200–202.

[32] Pinheiro, J.C. & Bates, D.M. (2000). *Mixed-effects Models in S and S-PLUS*. Springer, New York.

[33] Rabe-Hesketh, S., Skrondal, A. & Pickles, A. (2002). Reliable estimation of generalized linear mixed models using adaptive quadrature, *The Stata Journal* **2**, 1–21.

[34] Rao, P.S.R.S. (1997). *Variance Component Estimation*. Chapman & Hall, London.

[35] Rao, C.R. & Kleffe, J. (1988). *Estimation of Variance Components and Applications*. North-Holland, Amsterdam.

[36] Reeves, G.K., Cox, D.R., Darby, S.C. & Whitley, E. (1998). Some aspects of measurement error in explanatory variables for continuous and binary regression models, *Statistics in Medicine* **17**, 2157–2177.

[37] Robinson, G.K. (1991). That BLUP is a good thing: the estimation of random effects (with discussion), *Statistical Science* **6**, 15–51.

[38] Searle, S.R., Casella, G. & McCulloch, C.E. (1992). *Variance Components*. Wiley, New York.

[39] Shun, Z. (1997). Another look at the salamander mating data: a modified Laplace approximation approach, *Journal of American Statistical Association* **92**, 341–349.

[40] Skellam, J.G. (1948). A probability distribution derived from the binomial by regarding the probability of success as variable between sets of trials, *Journal of Royal Statistical Society B* **10**, 257–261.

[41] Snijders, T. & Bosker, R. (1999). *Multilevel Analysis: an Introduction to Basic and Advanced Multilevel Modelling*. Sage, London.

[42] Solomon, P.J. ed. (1998). Five papers on variance components in medical research, *Statistical Methods in Medical Research* **7**, 1–84.

[43] Solomon, P.J. & Cox, D.R. (1992). Nonlinear component of variance models, *Biometrika* **79**, 1–11.

[44] Speed, T.P. & Yang, Y.W. (2003). Direct and indirect hybridizations for cDNA microarray experiments, *Sankya Series A* **64**, 707–721.

[45] Stiratelli, R., Laird, N. & Ware, J. (1984). Random effects models for serial observations with binary responses, *Biometrics* **40**, 961–971.

[46] Tippett, L.H.C. (1931). *Methods of Statistics*. Matthew Norgate, London.

[47] Tukey, J.W. (1950). Some sampling simplified, *Journal of American Statistical Association* **45**, 501–519, Reprinted in *The collected works of* John W. Tukey, Vol. 7. Wadsworth, Pacific Grove.

[48] Verbeke, G. & Molenberghs, G. (2000). *Linear Mixed Models for Longitudinal Data*. Springer, New York.

P.J. SOLOMON