OSCAR KEMPTHORNE

*Iowa State University*

**Introduction.** I wish first to make some very general comments.

The title is a well-posed question. Analysis of variance is, arguably, the most widely used of the very broad spectrum of processes that are used in applied statistics. If we attempt to use language reasonably, as mathematical statisticians, the phrase "analysis of variance" should mean analysis of $E[X - E(X)]^2$, where $X$ is a random variable. Is this what applied statisticians are doing when they obtain "analysis of variance" or "anovas?" I believe not. I have taught courses on linear models and analysis of variance for many years. I found that in a course of 30 lectures, I do not become involved in random variables until about the 20th lecture. If we accept Speed's picture in which variance of a random variable is the "true" or "real" underlying concept, I have obviously been doing the "wrong thing." But, obviously, I would not have been doing what I regard as the "wrong thing" over all those years.

This leads me to a general plaint about the field of statistics. Over the decades, a language has been developed which is strongly misleading when examined in the light of our general nontechnical language. Let me mention a few of our basic words. One, which is strongly related to the present topic, is the word "regression." In mathematical statistics, this relates to a multivariate random variate, say $(X, Y)$, with regression being a conditional expectation, but in many, perhaps most, uses of the term, $X$ is not a random variable but a controlled variable. Other words are "confidence" and "test," as in "tests of significance" or "tests of hypotheses." And, of course, our most basic word is "probability," in which we, all I think, have problems and our field is, perhaps, schizophrenic.

**The Fisher quotation.** Speed gives a very challenging beginning with his Fisher quotation. This surely needs examination. Fisher says analysis of variance is not a mathematical theorem, but "rather a convenient method of arranging the arithmetic." The remark by Fisher (made as a contribution to the discussion of a paper) seems reasonable at first sight. But I have to ask if it has any "real" content. What is "the arithmetic?" I would say, rather, that it is a convenient method of arranging a "species" of arithmetic. If we accept my modification, we then have to specify what "species" of arithmetic is involved. I shall discuss this, but at present merely point out that it seems to have no relationship to random variables and variance of random variables.

**What then is analysis of variance?** I hold the view that analysis of variance is, indeed, a species of "arithmetic" that is related to linear models without any concept of random variables. We have a vector, $y$, say $n \times 1$, that we are trying to explain by a linear model

$$y = X_1\beta_1 + X_2\beta_2 + \cdots + X_k\beta_k + \text{disturbance},$$

and we can construct orthogonal projection matrices projecting $y$ onto

$\mathscr{C}(X_1), \mathscr{C}(X_1, X_2), \ldots$ which we may denote by $P_1, P_{12}, \ldots$. Then we decompose $y$ as

$$y = P_1 y + (P_{12} - P_1)y + \cdots + P_0 y,$$

with the consequence that

$$\|y\|^2 = \|P_1 y\|^2 + \|(P_{12} - P_1)y\|^2 + \cdots,$$

which is a generalized Pythagorean theorem, and can be written as an ANOVA with "sources," "degrees of freedom," "sums of squares" and "mean squares."

The matrices $P_1, P_{12}, \ldots,$ are symmetric idempotent and if $\mathscr{C}(X) \subset \mathscr{C}(Z)$, $P_Z - P_X$ is also symmetric idempotent.

This is surely just "arithmetic," if we use the term arithmetic with some breadth.

Is there a mathematical theorem? I say: Of course, there is. The "theorem" is that we have analyzed or decomposed the vector $y$ into $(k + 1)$ vectors or parts that are orthogonal or perpendicular in the usual sense. We are using the particular inner product of two vectors, say, $y$ and $z$, to be $y'z$. We can, easily, develop another partition of $y$ using another defined inner product. Obviously, Speed knows this, so I am not being critical of him. I feel confident that Fisher made at least two *lapsi linguae*.

The "arithmetic" I indicate easily leads to standard presented theory associated with the model that what I call "disturbance" is a realization of a random variable, $E \sim N_n(0, \sigma^2 I)$, because then linear functions of the random vector, $Y$ say, are normally distributed and geometrically perpendicular linear functions are uncorrelated and independent, and one obtains the panorama of $t$-tests, $F$-tests, and so on.

I take the view then that what we call analysis of variance is really just a "species of arithmetic." The fact that this species of arithmetic leads to the various very pleasing theorems of mathematical statistics is pure serendipity. We need only consider that the disturbances are realizations of non-Gaussian random variables and note that we then do not have any "nice" theory.

The Gaussian error model [and, indeed, the notion that our data (the vector $y$) are a realization of a random variable] is an idea that is a huge leap from the data and the arithmetic. This idea must come from data analysis and from the arithmetic we do via residual plots or (with the more recent buzz word) diagnostics.

**The problem with the "prescription" I give.** Obviously, in making the decomposition I give in the previous description, we have to choose an order of the part-model matrices, $X_1, X_2, \ldots, X_k$. Given potential explanatory matrices $X_1, X_2, \ldots, X_k$, there are, of course, $k!$ possible orders, so there are $k!$ decompositions of the data vector, $y$, and there are, in general, $k!$ ANOVAs. There are, of course, cases in which $X_1, X_2, \ldots, X_k$ have special structures, as in data arising from planned "orthogonal experiments" in which there is only one particular ANOVA, or all of the interesting $k!$ ANOVAs are the same (this is related to orthogonality of the partitions which the matrices $X_1, X_2, \ldots, X_k$ specify as incidence matrices).

It is in this connection that I am inclined to fault Speed's exposition. We are not given any warning that this problem arises in data analysis. It seems, rather, that Speed assumes a data-model structure in which this problem does not occur.

It is this problem (perhaps more than any other) that the applied statistician, the data analyst, meets. And it is this problem that has produced a massive literature of both theoretical and applied orientation.

**The simple example of Speed.** We have scalar data, $\{y_{ij}\colon\ i = 1(1)m,\ j = 1(1)n\}$ with the factor with levels indexed by $j$ being nested by the factor with levels indexed by $i$. Then, with standard notation,

$$y_{ij} = y_{\cdot\cdot} + (y_{i\cdot} - y_{\cdot\cdot}) + (y_{ij} - y_{i\cdot}),$$

and then, *because of arithmetic*, and a species of orthogonality:

$$\sum_{ij} y_{ij}^2 = \sum_{ij} y_{\cdot\cdot}^2 + \sum_{ij} (y_{i\cdot} - y_{\cdot\cdot})^2 + \sum_{ij} (y_{ij} - y_{i\cdot})^2$$

or

$$\sum_{ij} (y_{ij} - y_{\cdot\cdot})^2 = n\sum_{i} (y_{i\cdot} - y_{\cdot\cdot})^2 + \sum_{ij} (y_{ij} - y_{\cdot\cdot})^2.$$

This arithmetic identity has no necessary connection with random variables. It is "pure arithmetic." We can agree with Fisher that "analysis of variance" is just a convenient way of "arranging the arithmetic."

In this arithmetic, to rephrase a phrase of Speed, *there is no variance in sight*. Speed knows and describes this. So I am not being critical of him. That this decomposition of $\{y_{ij}\}$ can be expressed as

$$y = S_0 y + S_1 y + S_2 y,$$

when $S_0$, $S_1$, $S_2$ are symmetric, idempotent, pairwise orthogonal matrices is, I wish to insist, merely a matter of arithmetic. The "orthogonality" arises because, for instance,

$$\sum_{ij} (y_{i\cdot} - y_{\cdot\cdot})(y_{ij} - y_{i\cdot}) = \sum_{i} (y_{i\cdot} - y_{\cdot\cdot}) \sum_{j} (y_{ij} - y_{i\cdot}) = 0,$$

which is surely a matter of elementary arithmetic.

That the identity gives above for $\{y_{ij}\}$ can be extended to a wide variety of "balanced" data structures was exposited by Zyskind (1962). Then with balanced sampling from balanced populations, Zyskind gave expectations of terms that appear in the resulting ANOVA of the sample. The population factors could, clearly, have an infinite number of levels. I must, I think, express a point of criticism for Speed's omission of this work.

**Analysis of variance of a random variable.** Clearly, this is the thrust of Speed's paper and I have no criticisms of the material he gives, though I do wonder how many users of ANOVA will approach his paper (with its title) and become very frustrated. Such readers must realize that Speed's paper is a paper on variances and covariances of random variables.

Speed adjoins the idea that the data are realizations of random variables to the arithmetic.

What should analysis of variance-of-a-random-variable mean? The simple example is that in which

$$\sigma^2 = \sigma_b^2 + \sigma_w^2,$$

so that one does indeed "analyze" $\sigma^2$. A more complicated example occurs in quantitative genetic theory with variance $\sigma^2$ decomposed or "analyzed" thusly:

$$\sigma^2 = \sigma_A^2 + \sigma_D^2 + \sigma_{AA}^2 + \cdots + \sigma_E^2.$$

In these cases, one really is analyzing $\sigma^2$, that is, decomposing $\sigma^2$ into parts that have individual roles. There are examples galore of such analysis of $\sigma^2$.

**ANOVAs for infinite arrays.** That if we have "*balanced*" infinite arrays and "balanced" finite samples from such arrays, we are led to ANOVAs with "nice" expectations of mean squares is clear from Zyskind's (1962) work. It is interesting that the standard random effects models arise naturally as spectral decompositions of infinite (*but* balanced) arrays of multi-indexed random variables. I surmise that we should not be surprised by this. The covariance of the results of two samples from a finite population of size $N$ is $-1/(N-1)$, which will be zero if $N$ is indefinitely large. The covariance structures can be regarded as given ("out of thin air") or as being derived by random sampling. Which view one takes can be regarded as a matter of taste [*de gustibus*...]. I prefer the latter view.

**Classical ANOVA.** In this section, Speed brings to bear (what are to me) rather modern ideas of combinatoric algebra. This is rather formidable, though it seems that the whole matter goes back to the basic idea of the nesting of factors. This idea seems rather simple in that a factor is a partition, and factor 2 with partition $\pi_2$ is nested by factor 1 with partition $\pi_1$ if partition $\pi_2$ is equal to the product partition $\pi_1 \times \pi_2$. Again, I am of the opinion that it is "the arithmetic" that underlies the whole story and is the basis. One obtains random variables by some variety of *balanced* sampling.

I am impressed with how Speed and his collaborators have tied the very elementary approach I indicate with the modern ideas of algebra. My hope is that this path of development will lead to improved ideas of data analysis, which I regard as the "stuff" of statistics.

**What is ANOVA?** It will have become clear that my view is that ANOVA is a species of arithmetic. That this species has very interesting relationships to variability and covariability of random variables is, I think, very much a matter of serendipity. So I find myself in agreement with the Fisher quotation (provided it is "fixed up").

In saying this, I am not being at all critical of Speed's paper on the theory of variance of a random variable.

Perhaps I can summarize my view by the statement: "Analysis of variance" is not "analysis of 'variance'" but is analysis of variability and covariability of given *data*.

The notion that actual data are a realization of a random variable is nearly at the end of a data analysis; it is the outcome of various data analyses, of which ANOVA is a beginning component. Speed seems, however, to take this notion as the beginning of theory of analysis of variance. It is in this respect that I disagree with him.

Another comment is that analysis of variance is not necessarily a decomposition of a sum of squares into quadratic forms, as we can see with the model: $y_{ij} \doteq \mu a_i b_j$, first explored by Fisher and Mackenzie (1923) and developed over the past, say 20 years, by various workers. Fisher and Mackenzie gave an "analysis of variance." However, distribution theory of the constituent parts of the ANOVA is very difficult and quite different from what I would call a linear analysis of variance.

One can suggest, I think, that the basic theorems of ANOVA are nothing but Bessel's inequality and Parseval's theorem in finite-dimensional Euclidean spaces. These are surely "mathematical theorems."

Also, the main mathematical difference between ANOVA as arithmetic and ANOVA as a decomposition of variance is the use of different inner products for the two cases. It is then entirely natural that there is a close relationship, especially under equiprobable probability sampling.

The entirely natural, to me, even though somewhat sophisiticated, way to "integrate" the whole area is by use of partition theory and orthogonal partitions. So I am sympathetic to the introduction of combinatorial algebra, many of the ideas which were used by workers in the past without their knowing so [cf., *le bourgeois gentilhomme*].

## REFERENCES

FISHER, R. A. and MACKENZIE, W. A. (1923). Studies in crop variation. II. The manurial response of different potato varieties. *J. Agri. Sci.* **13** 311–320.

ZYSKIND, G. (1962). On structure, relation, sigma and expectations of mean squares. *Sankhyā Ser. A* **24** 115–148.

STATISTICAL LABORATORY
AND DEPARTMENT OF STATISTICS
SNEDECOR HALL
IOWA STATE UNIVERSITY
AMES, IOWA 50011