# ALTERNATIVE MODELS FOR THE ANALYSIS OF VARIANCE[1]

BY HENRY SCHEFFÉ[2]

*University of California, Berkeley*

**Summary.** The terminology is defined and illustrated in Section 1. A little historical background not very familiar to statisticians is sketched in Section 2. In Section 3 some difficulties about the formulation of random interactions are discussed. Section 4 deals with models reflecting a randomization in the experiment to assign the treatment combinations to finite populations of experimental units.

**1. Introduction.** A broad survey of the present state of the theory of alternative models in the analysis of variance would require a monograph and be soon outmoded. The selective approach of this paper has been determined mainly by the interests of the writer. His chief interests are in the mathematical models—their formulation, motivation, and statistical inference in them. The reader is referred to a useful survey by Crump (1951);[3] he will find little overlap between the two papers. In these papers there is little attempt to deal with the always important and often difficult problems of careful tailoring of the models to particular situations in the physical world; discussions of such problems may be found in the work of Kempthorne and Wilk cited in the References at the end.

The *analysis of variance* might be defined as a statistical method for analyzing observations assumed to be constituted of linear combinations, subject to a certain restriction to be stated below, of *effects*. (We use the terminology of "effects" to include what are usually called the "general mean," "main effects," "interactions," and "errors.") The effects—not directly observable quantities—are more or less idealized formulations of some properties of interest to the investigator in the phenomena underlying the observations. The purpose of the analysis is to make inferences about some of the effects, these inferences to be valid regardless of the magnitudes of certain other effects, which may be present in the linear combinations, and which we may be more desirous of "eliminating" than "assessing."

The theory of this method naturally has implications about how observations should be taken or an experiment planned, i.e., about *experimental design*. The term "experimental design" is used here in a broad sense to include, for example, the application described below of variance-components analysis to the non-

Received October 14, 1955.

[1] An address presented on September 1, 1955 at the Ann Arbor meeting of the Institute of Mathematical Statistics by invitation of the I.M.S. Committee on Special Invited Papers.

[2] This paper was prepared with the partial support of the Office of Naval Research. It may be reproduced in whole or in part for any purpose of the United States Government.

[3] Names followed by dates in parentheses are listed in the References at the end.

experimental science of astronomy to help decide on how many nights observations should be taken and how many per night.

The restriction mentioned above is that the coefficients in the linear combinations which give the observations in terms of the effects be integers; usually they are exclusively 0 and 1. (For an example where some are $-1$ see Scheffé (1952, Sec. 7); for an example where some are 2, see Kempthorne (1952, Sec. 6.8). If more than a few different integers occur as coefficients of the same effect it would be customary to say we have a case of analysis of covariance or of regression analysis instead of analysis of variance; it does not seem worthwhile to attempt to draw sharp dividing lines.)

Each effect is regarded as either an unknown constant or else as a random variable, the joint distribution of the random variables being in general not assumed completely known. If the effect is treated as an unknown constant it is called a *fixed effect* or *Model* I *effect*, otherwise it is called a *random effect* or *Model* II *effect*. (Some writers apply the terminology of "Model II effects" only to random effects which satisfy certain further distribution assumptions including independence and normality.) The diversity of models that have been considered for the analysis of variance arises from the possibilities of treating various effects as fixed or random and of making various distribution assumptions about the random effects. For simplicity we shall always assume that all random effects have finite variance (this might not be necessary in a nonparametric approach). We may assume all random effects to have zero means by introducing further fixed effects if necessary. We shall assume that there is at least one set of random effects equal in number to the number of observations, a different one of which appears in each observation, which is called a set of *errors*. There is usually one fixed effect that appears in every observation; if this is present we shall call it the *additive constant*—or the *general mean* if it is the mean of all the observations in some sense.

The equations expressing the observations as linear combinations of the effects will be called the *model equations*: Together with the distribution assumptions on the random effects and possible side conditions on the fixed effects they determine the *model*. The model will be called a *fixed-effects model* or *Model* I if the only random effects in the model equations are the error terms; it is called a *random-effects model* or a *variance-components model* or *Model* II if all effects, except the additive constant if there is one, are random effects. A case falling under neither of these categories is called a *mixed model*.

We now illustrate the terminology. Imagine an experiment in a factory to study the performance of $P$ machines and $Q$ workers. Suppose that all the machines, each run by a single worker, produce small parts of the same kind, that a large number is produced daily on each machine, and that there is considerable day-to-day variation for any worker (for some purposes we will treat the output as though it were a continuous random variable). Denote by $\mu_{pq}$ the "true" daily output of the $q$th worker on the $p$th machine; this differs from the observed output by an "error." We might regard $\mu_{pq}$ as an idealized long-term daily average for the $q$th worker on the $p$th machine after he has reached a relatively stable period following a learning stage. Our convention on subscripts throughout will be that subscripts $p$, $p'$, $p''$ range from 1 to $P$; $q$, $q'$, $q''$, from 1 to $Q$, etc.; and that when a subscript is replaced by a dot it indicates that the arithmetic mean has been taken over that subscript; thus $\mu_{p.} = \sum_q \mu_{pq}/Q$, $\mu_{..} = \sum_p \sum_q \mu_{pq}/(PQ)$. In the familiar way we define the general mean to be

$$\mu = \mu_{..}\,, \tag{1.1}$$

the main effect for the $p$th machine to be

$$\alpha_p = \mu_{p.} - \mu_{..}\,, \tag{1.2}$$

the main effect for the $q$th worker to be

$$\beta_q = \mu_{.q} - \mu_{..}\,, \tag{1.3}$$

and the interaction between the $p$th machine and the $q$th worker to be

$$\gamma_{pq} = \mu_{pq} - \mu_{p.} - \mu_{.q} + \mu_{..}\,, \tag{1.4}$$

so that

$$\mu_{pq} = \mu + \alpha_p + \beta_q + \gamma_{pq}\,, \tag{1.5}$$

where

$$\textstyle\sum_p \alpha_p = 0, \qquad \sum_q \beta_q = 0, \qquad \sum_p \gamma_{pq} = 0 \text{ for all } q, \ \sum_q \gamma_{pq} = 0 \text{ for all } p. \tag{1.6}$$

Suppose an experiment is contemplated in which each of the workers is put for $K$ days on each of the machines. For reasons to become clear later, let us now change from the subscripts $p$, $q$ to $i$, $j$. Then the output of the $j$th worker the $k$th day he is on the $i$th machine may be written

$$y_{ijk} = \mu_{ij} + e_{ijk}\,, \tag{1.7}$$

where $e_{ijk}$ is the "error." We shall assume the set of $IJK$ errors $\{e_{ijk}\}$ to be independently distributed with zero means and common variance $\sigma_e^2$. For some purposes a normality assumption on the errors may be added. We shall sometimes employ the jargon that the $I$ machines are the "$I$ levels of factor $A$" and the $J$ workers are the "$J$ levels of factor $B$" in the experiment.

The formulation of the interactions in the other models causes some difficulties we wish to postpone; so to keep all the illustrations simple at this point, let us assume the interactions between machines and workers are zero. With all $\gamma_{ij} = 0$ the model equation then becomes

$$y_{ijk} = \mu + \alpha_i + \beta_j + e_{ijk}\,, \tag{1.8}$$

and since

$$\textstyle\sum_i \alpha_i = 0, \tag{1.9}$$

$$\textstyle\sum_j \beta_j = 0, \tag{1.10}$$

and $E(e_{ijk}) = 0$, therefore $\mu$ is the general mean in an obvious sense. The effects are the terms of four kinds on the right of (1.8); all are fixed effects except the errors, so this is a fixed-effects model.

Suppose the $J$ workers in the experiment could be regarded as a random sample from a large pool of workers and that we were more interested in making inferences about this population of workers than about the $J$ individuals in the experiment. Idealizing the population of workers as infinite and assuming no interactions we would then be led to the model equation

$$(1.11) \qquad y_{ijk} = \mu + \alpha_i + b_j + e_{ijk},$$

where the random variables $\{b_j\}$, the worker effects, are independently and identically distributed. We again assume the errors $\{e_{ijk}\}$ to be independently distributed with zero means and equal variance, and we also assume them independent of the $\{b_j\}$. We may without loss of generality assume the $\{\alpha_i\}$ continue to satisfy (1.9), but the $\{b_j\}$ for the $J$ workers in the experiment of course no longer satisfy the analogue of (1.10). By adding $E(b_j)$ to $\mu$ we may redefine the $\{b_j\}$ and $\mu$ so that $E(b_j) = 0$ and $\mu$ is again the general mean. This example is one of a mixed model.

It would be appropriate to treat the machine effects also as random if the machines in the experiment were of the same make and model and could be regarded as a random sample from some population of machines which is of primary interest. If this population be also idealized as infinite, the machine effects are then independently and identically distributed random variables $\{a_i\}$, and again we may without loss of generality assume $E(a_i) = 0$. Since the random sample of workers is assumed to be selected independently of the random sample of machines, the set of worker effects $\{b_j\}$ is independent of the set $\{a_i\}$. We further assume the set of errors $\{e_{ijk}\}$ to be independent of the sets of effects $\{a_i\}$ and $\{b_j\}$, and again assume the $\{e_{ijk}\}$ to be independent with zero means and equal variance. We now have the model equation

$$(1.12) \qquad y_{ijk} = \mu + a_i + b_j + e_{ijk}$$

for a random-effects model or variance-components model. The latter terminology arises from the relation that the variance of an observation is now

$$(1.13) \qquad \sigma_y^2 = \sigma_A^2 + \sigma_B^2 + \sigma_e^2,$$

where $\sigma_y^2$, $\sigma_A^2$, $\sigma_B^2$, and $\sigma_e^2$ are the respective variances of the observations, the machine (factor $A$) effects, the worker (factor $B$) effects, and the errors, and the three terms on the right of (1.13) are appropriately called the *variance components*.

We see that in formulating a model one must ask for each factor whether one is interested individually in the particular levels occurring in the experiment or primarily in a population from which the levels in the experiment can be regarded as a sample: the main effects are accordingly treated as fixed or as random. (It is conceivable that for two different purposes the same data might be analyzed according to two different models in which the same main effects are regarded as fixed or as random effects.) Interactions between several factors are naturally treated as fixed if all these factors have fixed effects, and as random if one or more of these factors have random effects. The difficulty already alluded to and to be discussed in Section 3 concerns the kind of dependence assumptions to be made about the random interactions. While the decision as to whether the main effects of any factor, say $A$, are to be treated as fixed or random obviously affects the meaning of the main effects of $A$ and the interactions of sets of factors including $A$, it also affects the meaning of the main effects of the other factors and of the interactions of sets of factors not including $A$: This is because the latter main effects and interactions are defined as averages over the levels of $A$, and the decision determines whether these averages are taken over the particular levels of $A$ in the experiment, or over a population of levels, of which the levels in the experiment are a sample.

The assumption of independent errors made in this section is not appropriate if the errors arise from the random assignment of experimental units to treatment combinations from finite populations of units; this will be considered in Section 4.

We shall abide by the notational rule followed in this section of denoting fixed effects by Greek letters and random effects by Latin letters.

**2. Some history.** Fixed-effects models in which the covariance matrix of the errors is known up to a scalar factor are special (because of the restriction on the coefficients to be integers) cases of the models, sometimes called "linear hypothesis" models, used in the theory of least squares. It is well known that the theory of least squares was invented independently and published by Legendre (1806) and Gauss (1809; see also Plackett (1949)) in books on astronomical problems, so that to these problems we must ascribe the origin of the fixed-effects models. It is not so well known that astronomers, long before statisticians, also formulated variance-components models. For the references establishing this I am indebted to Dr. Churchill Eisenhart.

Very explicit use of a variance-components model for the one-way layout is made by Airy (1861, Part IV), with all the subscript notation[4] necessary for clarity. Suppose that on $I$ nights observations are made with a telescope on the same phenomenon, $J_i$ observations on the $i$th night. Airy assumes the following structure for the $j$th observation on the $i$th night:

$$y_{ij} = \mu + c_i + e_{ij}, \tag{2.1}$$

where $\mu$ is the general mean or "true" value, and the $\{c_i\}$ and $\{e_{ij}\}$ are random effects with the following meanings: He calls $c_i$ the "constant error," meaning it is constant on the $i$th night; we would call it the $i$th night effect; it is caused

[4] To conform to the notation of this paper I have only changed his capital letters (1861, Sec. 118; Sec. 133 in 3rd ed.) to lower case, and I have added the general mean $\mu$ since he writes the equations for the observations minus $\mu$ instead of for the observations.

by the "atmospheric and personal circumstances" peculiar to the $i$th night. The $\{e_{ij}\}$ for fixed $i$ we would call the errors about the (conditional) mean $\mu + c_i$ on the $i$th night. It is implied by Airy's discussion that he assumes all the $e_{ij}$ independently and identically distributed, similarly for the $c_i$, that the $\{e_{ij}\}$ are independent of the $\{c_i\}$, and that all have zero means. Let us denote the variances of the $\{e_{ij}\}$ and the $\{c_i\}$ by $\sigma_e^2$ and $\sigma_c^2$.

To decide about his equivalent of the hypothesis $\sigma_c^2 = 0$, Airy compares, as we would, a between-nights measure of variability with a within-nights measure, but he uses different measures than we would. For brevity suppose all $J_i = J$. From the observations on the $i$th night Airy estimates $\sigma_e$ by the r.m.s. (root-mean-square) estimate—note his use of $J - 1$ in the denominator—

$$\hat{\sigma}_{e,i} = [\textstyle\sum_j (y_{ij} - y_{i.})^2/(J - 1)]^{\frac{1}{2}}, \tag{2.2}$$

and he then takes the arithmetic average of these to get $\hat{\sigma}_e = \sum_i \hat{\sigma}_{e,i}/I$, where we would use the r.m.s. average. Actually he works with the "probable errors," which are a conventional constant, calculated from the normal distribution, times the r.m.s. estimates. For the between-nights measure he uses not a function of the between-nights sum of squares $J \sum_i (y_{i.} - y_{..})^2$ but the corresponding mean deviation from the mean,

$$d = I^{-1}\textstyle\sum_i | y_{i.} - y_{..} |. \tag{2.3}$$

Under the hypothesis $\sigma_c^2 = 0$ he calculates an approximate probable error for $d$ by replacing $y_{..}$ in (2.3) by $\mu$ (so the terms in (2.3) become independent) and the unknown $\sigma_e$ by $\hat{\sigma}_e$. If $d$ is less than the approximate probable error thus obtained he accepts $\sigma_c^2 = 0$, if $d$ is large compared with the approximate probable error (how large Airy does not specify, and he seems to despair of the possibility of a mathematical criterion), he rejects $\sigma_c^2 = 0$ and estimates $\sigma_c$ by a conventional constant times $d$. There is no attempt to correct this estimate of $\sigma_c$ for bias due to $\sigma_e$ inherent in the relation Var $(y_{i.}) = \sigma_c^2 + J^{-1}\sigma_e^2$; anyway, under his procedure Airy's estimate of $J^{-1}\sigma_e^2$ would be small compared with his estimate of $\sigma_c^2$. One wonders whether Airy used the mean deviation to measure the between-nights variation rather than the r.m.s. measure, as he did in (2.2), because he found it easier to approximate the probable error of the former.

Chauvenet (1863, Vol. 2, Art. 163, 164), while not writing model equations like (2.1) with all the subscripts, nevertheless implies such models and utilizes the consequences, such as Var $(y_{..}) = I^{-1}(\sigma_c^2 + J^{-1}\sigma_e^2)$ from (2.1). He concludes from this that there is no practical advantage in increasing $J$ beyond a certain point in such a case, and credits this idea to Bessel (1820, p. 166), saying that Bessel thought $J = 5$ sufficient for a certain situation. Chauvenet's reference to Bessel on this specific point ($J = 5$) is incorrect, but the page he cites does contain a formula for the probable error of a sum of independent random variables which could be the basis for such a conclusion. Probably Bessel made the remark elsewhere.

Fisher's (1918) basic paper on population genetics, which introduces the terms "variance" and "analysis of variance," employs variance-components models, and they are of course behind his (1925, Sec. 40) treatment of the intraclass correlation coefficient. First to add to Fisher's analysis of variance tables the very useful column of expected mean squares for variance-components models appears to have been Tippett (1931, Table XXIV; in (1929) he calculated but did not table them). While a mixed model is implied by Fisher's (1935, Sec. 65) discussion of varietal trials in a randomly selected set of locations, and by Yates' (1935a) analysis of the split-plot design, the first explicit model equation the writer has found for this case is in a paper on mental tests by Jackson (1939), where the score of the $j$th individual on the $i$th trial of a test is assumed to have the structure (1.11), with the subscript $k$ suppressed, the trial effects being treated as fixed and the "individual" effects as random. Interaction effects which are clearly labeled as random effects were introduced into the variance components model equation (1.12) by Crump (1946). The terminology of "Model I" and "Model II" is due to Eisenhart (1947). Basic work of Tukey will be discussed in Section 3, and of Fisher and Neyman in Section 4. In textbooks, alternative models for the analysis of variance were introduced by Mood (1950) and developed at length by Kempthorne (1952), and Anderson and Bancroft (1952).

**3. Treatment of random interactions.** The examples of fixed-effects models, mixed models, and random-effects models given at the end of Section 1 all refer to an $I \times J$ two-way layout with $K(K \geqq 1)$ observations per cell, where $y_{ijk}$ denotes the $k$th observation in the $i$, $j$-cell. Let us define the usual sums of squares, namely, those for rows (or $A$),

$$SS_A = JK \textstyle\sum_i (y_{i..} - y_{...})^2, \tag{3.1}$$

for columns (or $B$),

$$SS_B = IK \textstyle\sum_j (y_{.j.} - y_{...})^2, \tag{3.2}$$

for interactions (or $A \times B$),

$$SS_{AB} = K \textstyle\sum_i \sum_j (y_{ij.} - y_{i..} - y_{.j.} + y_{...})^2, \tag{3.3}$$

for error,

$$SS_e = \textstyle\sum_i \sum_j \sum_k (y_{ijk} - y_{ij.})^2, \tag{3.4}$$

and a "pooled" error sum of squares,

$$SS_{pe} = SS_{AB} + SS_e. \tag{3.5}$$

As long as interactions are omitted from the model equations, no great differences appear among the three models formulated in connection with (1.8), (1.11), and (1.12). The expected values of the above sums of squares in the three models may actually be expressed by the same formulas by the usual device of suitably defining the symbols $\sigma_A^2$ and $\sigma_B^2$ : When the levels of factor $A$ have random main effects $\{a_i\}$, as in (1.12), we define

$$\sigma_A^2 = \text{Var}\,(a_i); \tag{3.6}$$

likewise

$$\sigma_B^2 = \text{Var}\,(b_j) \tag{3.7}$$

in (1.11) and (1.12). When the levels of factor $A$ have fixed main effects $\{\alpha_i\}$, as in (1.8) and (1.11), we define $\sigma_A^2$ to be the following function of the fixed effects:[5]

$$\sigma_A^2 = (I - 1)^{-1}\textstyle\sum_i \alpha_i^2 ; \tag{3.8}$$

likewise

$$\sigma_B^2 = (J - 1)^{-1}\textstyle\sum_j \beta_j^2 \tag{3.9}$$

in (1.8). Then for all three models

$$E(MS_A) = JK\sigma_A^2 + \sigma_e^2, \tag{3.10}$$

$$E(MS_B) = IK\sigma_B^2 + \sigma_e^2, \tag{3.11}$$

$$E(MS_{AB}) = E(MS_e) = E(MS_{pe}) = \sigma_e^2, \tag{3.12}$$

where the mean square $MS_x$ is defined as the corresponding $SS_x$ in (3.1) to (3.5) divided by the number of *d.f.* (degrees of freedom), namely, $I - 1$, $J - 1$, $(I - 1)(J - 1)$, $IJ(K - 1)$, $IJK - I - J + 1$, for $x = A$, B, AB, $e$, $pe$, respectively. (In statements involving $MS_e$ it is assumed that $K > 1$.)

If we add the normality assumption (namely, that all random effects are normally distributed) we get exact $F$-tests of the hypotheses $H_A{:}\sigma_A^2 = 0$ and $H_B{:}\sigma_B^2 = 0$ by employing in the usual way the statistics $MS_A/MS_{pe}$ and $MS_B/MS_{pe}$. The power functions however are quite different under the three models, involving, for example, only central $F$-distributions under the random-effects model and only noncentral (this term includes "central") $F$-distributions, not central under the alternatives, for the fixed-effects model. These results are all earily verified by substituting $y_{ijk}$ from the model equations into (3.1) to (3.5), simplifying, and then applying well-known "linear hypothesis" theory. When an $F$-test rejects, one would usually proceed differently in the three models: thus, if $H_A$ is rejected one might use a multiple comparison method on the $\{\alpha_i\}$ if they are fixed effects, and an interval estimate of $\sigma_A^2$ (an exact solution is not at present available) or of $\sigma_A^2/\sigma_e^2$ if the $\{a_j\}$ are random.

The consequences on statistical inference are much more divergent when interaction terms are included in the three models. For the fixed-effects model we denote the interactions by $\{\gamma_{ij}\}$, so that the model equation becomes

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + e_{ijk}, \tag{3.13}$$

[5] Similar definitions with the denominators increased by unity were used by Daniels (1939).

where the $\{\alpha_i\}$, $\{\beta_j\}$, and $\{\gamma_{ij}\}$ satisfy (1.6) with $p$, $q$ replaced by $i$, $j$. The expected mean squares for this model are shown in column (i) of Table 1, where $\sigma_{AB}^2$ is defined to be the following function of the fixed interactions,

$$\sigma_{AB}^2 = (I-1)^{-1}(J-1)^{-1}\textstyle\sum_i \sum_j \gamma_{ij}^2 . \tag{3.14}$$

This is under the assumption that the $\{e_{ijk}\}$ are independently distributed with zero means and equal variances $\sigma_e^2$; if we add the normality assumption we get the well-known theory of estimation and testing for this model; in particular, $MS_A$, $MS_B$, and $MS_{AB}$ are all tested against $MS_e$.

For the mixed model it seems inescapable (also for the random-effects model) to regard the interaction between the $j$th level of the column factor and the $i$th level of the row factor as a random effect, since the $j$th level of the column factor is chosen at random from a population of levels. Let us denote it by $c_{ij}$, so that we have the model equation

$$y_{ijk} = \mu + \alpha_i + b_j + c_{ij} + e_{ijk} \tag{3.15}$$

for the mixed model. (Similarly we write the model equation

$$y_{ijk} = \mu + a_i + b_j + c_{ij} + e_{ijk} \tag{3.16}$$

for the random-effects model.) What should we assume about the distribution of the random variables $\{c_{ij}\}$? The easiest thing is to assume them independently and identically distributed, with zero means, and independent of the $\{e_{ijk}\}$ and the $\{b_j\}$ (and of the $\{a_i\}$ in the random effects model). But the assumption that the $\{c_{ij}\}$ are independent of the $\{b_j\}$ (and of the $\{a_j\}$) is hard to accept: Thus in the above example of the mixed model for machines and workers, it is not unreasonable to assume that the $J$ workers are chosen independently from a population; but then to assume that after a certain worker is included in the experiment, the interaction between him and any one of the $I$ machines should be independent of the worker (or at least of the worker's main effect) and of the machine seems to violate the very notion of interaction between worker and machine. (A similar objection would apply in the random-effects model.)

Under this easy but unrealistic assumption on the $\{c_{ij}\}$ the expected mean

TABLE 1

*Expected values of mean squares*

| Mean square | Expected value in | | |
|---|---|---|---|
| | (i) | (ii) | (iii) |
| | Fixed-effects model | Mixed or random-effects model with independent interactions | Mixed model with dependent interactions |
| $MS_A$ | $JK\sigma_A^2 + \sigma_e^2$ | $JK\sigma_A^2 + K\sigma_{AB}^2 + \sigma_e^2$ | $JK\sigma_A^2 + K\sigma_{AB}^2 + \sigma_e^2$ |
| $MS_B$ | $IK\sigma_B^2 + \sigma_e^2$ | $IK\sigma_B^2 + K\sigma_{AB}^2 + \sigma_e^2$ | $IK\sigma_B^2 + \sigma_e^2$ |
| $MS_{AB}$ | $K\sigma_{AB}^2 + \sigma_e^2$ | $K\sigma_{AB}^2 + \sigma_e^2$ | $K\sigma_{AB}^2 + \sigma_e^2$ |
| $MS_e$ | $\sigma_e^2$ | $\sigma_e^2$ | $\sigma_e^2$ |

squares turn out to be those listed in column (ii) of Table 1; here $\sigma_{AB}^2$ denotes Var $(c_{ij})$. The table suggests that in the mixed or random-effects models, unlike the fixed-effects model, the mean squares for main effects should be tested against the interaction mean square. It is easy to show these procedures (as well as testing $MS_{AB}$ against $MS_e$) give exact $F$-tests of the respective hypotheses under the normality assumption (that all random effects are jointly normal). Again the power in every case involves only noncentral $F$-distributions (all central for the random-effects model).

Distribution assumptions on the random interactions $\{c_{ij}\}$ that seem acceptable to the writer may be reached by following a trail broken by Tukey[6] (1949). Let us consider first the case of a finite number of machines and men, not all of which are going to be included in the experiment. If $\mu_{pq}$ is the "true" output of the $q$th worker on the $p$th machine, there is no question about how to define the main effects $\{\alpha_p\}$, $\{\beta_q\}$, and interactions $\{\gamma_{pq}\}$; they are defined by equations (1.2), (1.3), and (1.4). Now let us conceive of the mixed model as a limiting case where the number $Q$ of workers in the population becomes infinite and only a sample of $J$ workers is included in the experiment, but all the machines are included, so that $p = i$, $P = I$. Then the role of $\mu_{pq}$ is played by $m(i, x)$, where $x$ labels the worker in the population, and $m(i, x)$ is his "true" output on the $i$th machine. It will be convenient to denote by $\mathcal{P}$ the population distribution of $x$, even though it does not enter the calculations directly.[7]

Clearly the analogue of the resolution (1.5) is

$$m(i, x) = \mu + \alpha_i + b(x) + c_i(x), \tag{3.17}$$

where

$$\mu = m(., .), \tag{3.18}$$

$$\alpha_i = m(i, .) - m(., .), \tag{3.19}$$

$$b(x) = m(., x) - m(., .), \tag{3.20}$$

$$c_i(x) = m(i, x) - m(i, .) - m(., x) + m(., .), \tag{3.21}$$

and where replacing $i$ by a dot in $m(i, x)$ signifies that the arithmetic average has been taken over $i$ for $i = 1, 2, \cdots, I$, and replacing $x$ by a dot means the expectation has been taken over $x$ with respect to $\mathcal{P}$. We may call $\{\alpha_i\}$ and $b(x)$ the main effects in the population, and $\{c_i(x)\}$ the population interactions, and we note they satisfy the conditions

$$\textstyle\sum_i \alpha_i = 0, \quad E(b(x)) = 0, \quad \sum_i c_i(x) = 0 \text{ (all } x), \quad E(c_i(x)) = 0 \text{ (all } i). \tag{3.22}$$

[6] Tukey did not publish his results in a journal and they were independently found by Wilk and Kempthorne, and Cornfield.

[7] The reader interested in these points will easily supply the mathematical assumptions under which $\mathcal{P}$ is a probability distribution on a probability space of points $x$, the $I$ functions $m(i, x)$ are random variables with finite variance, and also the appropriate assumptions on the product space of points $(y, x)$ for the random effects model below.

The random variables $b(x), c_1(x), \cdots, c_I(x)$ are not independent; their variances and covariances are functions of the covariance matrix $(\sigma_{ii'})$ of the $I$ random variables $\{m(i, x)\}$. If

$$\sigma_{ii'} = \text{Cov}\,(m(i, x), m(i', x)), \tag{3.23}$$

then

$$\text{Var}\,(b(x)) = \sigma_{..}, \tag{3.24}$$

$$\text{Cov}\,(c_i(x), c_{i'}(x)) = \sigma_{ii'} - \sigma_{i.} - \sigma_{i'.} + \sigma_{..}, \tag{3.25}$$

$$\text{Cov}\,(b(x), c_i(x)) = \sigma_{i.} - \sigma_{..}. \tag{3.26}$$

Suppose now a random sample of $J$ workers is taken from the population. If they are labeled by $x_1, \cdots, x_J$, so that the $\{x_j\}$ are independently distributed according to $\mathcal{P}$, then the "true" output of the $j$th worker in the experiment on the $i$th machine will be $m(i, x_j)$, which we shall write $m_{ij}$. We shall assume the observation $y_{ijk}$ to be of the form

$$y_{ijk} = m_{ij} + e_{ijk}, \tag{3.27}$$

where the set $\{e_{ijk}\}$ is independent of the set $\{m_{ij}\}$. We shall also assume the $\{e_{ijk}\}$ to be independent with zero means and equal variance $\sigma_e^2$. Writing $b_j = b(x_j)$ and $c_{ij} = c_i(x_j)$, we have from (3.27) and (3.17),

$$y_{ijk} = \mu + \alpha_i + b_j + c_{ij} + e_{ijk}, \tag{3.28}$$

where the $\{e_{ijk}\}$ are independent of the $\{b_j\}$ and $\{c_{ij}\}$, all have zero means, and the $J$ sets $\{b_j, c_{1j}, \cdots, c_{Ij}\}$ are independently and identically distributed like the random vector $(b(x), c_1(x), \cdots, c_I(x))$ whose covariance matrix is given by (3.24), (3.25), (3.26), the elements of the underlying covariance matrix $(\sigma_{ii'})$ being regarded as unknown parameters.

The appropriate definition of the symbols $\sigma_B^2$ and $\sigma_{AB}^2$ is suggested by starting from the customary definitions for the finite set $\{\mu_{pq}\}$, namely,

$$\sigma_B^2 = (Q-1)^{-1}\textstyle\sum_q \beta_q^2, \tag{3.29}$$

$$\sigma_{AB}^2 = (I-1)^{-1}(Q-1)^{-1}\textstyle\sum_i \sum_q \gamma_{iq}^2, \tag{3.30}$$

and going to the limit; the result is

$$\sigma_B^2 = \text{Var}\,(b(x)), \tag{3.31}$$

$$\sigma_{AB}^2 = (I-1)^{-1}\textstyle\sum_i \text{Var}\,(c_i(x)). \tag{3.32}$$

For details and discussion of this and the other results we shall now briefly mention for this model, and for citations of related work, the reader is referred to Scheffé (1956). With these definitions the expected mean squares are those in column (iii) of Table 1. Under the normality assumption the tests of the hypotheses $\sigma_B^2 = 0$ and $\sigma_{AB}^2 = 0$ suggested by the table, namely, those based on $MS_B/MS_e$ and $MS_{AB}/MS_e$, turn out to be exact $F$-tests. However under the hypothesis $\sigma_A^2 = 0$ (i.e., $\alpha_1 = \alpha_2 = \cdots = \alpha_I = 0$), the statistic $MS_A/MS_{AB}$

suggested by the table does not have an $F$-distribution; an exact test of this hypothesis and an associated multiple-comparison method may be based on Hotelling's $T^2$ statistic.

Appropriate distribution assumptions about the interactions in the random-effects model may be motivated in a similar way. Let the machines be labeled by $y$ with population distribution $\mathcal{Q}$. We assume $y$ and $x$ independent, corresponding to independent choices of machine and worker. Let the random variable $m(y, x)$ be the "true" output of the worker labeled $x$ on the machine labeled $y$. Define the general mean, main effects for machines, main effects for workers, and interactions, all in the population, respectively by

$$\mu = m(.,.), \tag{3.33}$$

$$a(y) = m(y,.) - m(.,.), \tag{3.34}$$

$$b(x) = m(.,x) - m(.,.), \tag{3.35}$$

$$c(y,x) = m(y,x) - m(y,.) - m(.,x) + m(.,.), \tag{3.36}$$

where replacing $x$ or $y$ by a dot in $m(y, x)$ indicates the expected value has been taken over $x$ or $y$ with respect to $\mathcal{P}$ or $\mathcal{Q}$, respectively. The analogue of (1.5) is now

$$m(y,x) = \mu + a(y) + b(x) + c(y,x). \tag{3.37}$$

If the $I$ machines randomly selected for the experiment are labeled by $y_1, \cdots, y_I$, the $J$ workers by $x_1, \cdots, x_J$, so that the $\{x_i\}$ and $\{y_j\}$ are independently distributed according to $\mathcal{P}$ and $\mathcal{Q}$, respectively, then the "true" output of the $j$th worker on the $i$th machine in the experiment will be $m(y_i, x_j)$. Write $m_{ij} = m(y_i, x_j)$, $a_i = a(y_i)$, $b_j = b(x_j)$, $c_{ij} = c(y_i, x_j)$. Then assuming as in (3.27) that the errors $\{e_{ijk}\}$ are independent of the $\{m_{ij}\}$ and are independently distributed with zero means and common variance $\sigma_e^2$, we get the model equation

$$y_{ijk} = \mu + a_i + b_j + c_{ij} + e_{ijk}, \tag{3.38}$$

and may verify that all the random effects have zero means, the $a_1, \cdots, a_I$, $b_1, \cdots, b_J$ are independent, and the $\{c_{ij}\}$ are *uncorrelated* with each other and with the $\{a_i\}$ and $\{b_j\}$. The expected mean squares are then those in column (ii) of Table 1. If we now add the assumption that all the random effects $\{a_i\}$, $\{b_j\}$, $\{c_{ij}\}$ are jointly normal, this forces the $\{c_{ij}\}$ to be *independent* of the $\{a_i\}$ and $\{b_j\}$ as well as among themselves, and we are back to the random-effects model discussed earlier—with perhaps a little less aversion to the independence assumptions on the interactions—or a little more suspicion about the innocuousness of the normality assumption!

This approach to the random-effects model is easily extended to more factors, and it is again found that *all* the random effects in the model equation are uncorrelated. The extension along the above lines (involving $T^2$) of the mixed model to more factors is in progress; expected mean squares have already been given by several writers, including Tukey (1949), Wilk and Kempthorne (1953–1955), and Bennett and Franklin (1954).

In applying models like those discussed in this section one cannot evade the question as to the effect of the inevitable violations of some of the following three assumptions made on the errors: (i) independence, (ii) equal variance, and (iii) normality. Of these, (i) is the most difficult to discuss. The effects on the $F$-tests of certain kinds of correlation in the fixed-effects model have been studied by Daniels (1938) for the one-way layout and by Box (1954) for the two-way layout, and we merely mention that they can be serious. The effect on point estimation and tests, of correlation of errors due to the random assignment of treatment combinations to finite populations of experimental units will be treated in Section 4. That violation of (ii) should not seriously affect the $F$-tests in the case of balanced designs is suggested by approximations by Daniels (1938) and Horsnell (1953), and exact small-sample calculations by Box (1954). Such insensitivity of the $F$-tests to variance heterogeneity would then carry over to the multiple-comparisons methods associated with the $F$-test (Scheffé (1953)), although single inferences using the $t$-distribution and based on the assumption of variance homogeneity could be extremely misleading. As long as we limit ourselves to calculating expected mean squares, violation of (iii) is of course of no effect. Work of many writers,[8] including E. S. Pearson (1931), Box (1953), and Box and Andersen (1954) leads to the generalization that non-normality should have little effect on the validity of inferences about fixed effects but may play havoc with inferences about random effects. There is again the comforting consideration that multiple-comparison methods associated with the $F$ or $T^2$ tests (Scheffé 1956)) should share with these tests their insensitivity to violation of (iii).

Tukey's (1949, 1951) and Wilk and Kempthorne's (1953–1955) extensive work on alternative models is in a more general form than we have considered, in that, if a factor appears in the experiment at $n$ levels, the levels are treated as a random sample (without replacement) from a finite population of $N$ levels. This would be of interest, for example, in the illustration of machines and men if it were desired to make inferences about a finite population of $N$ machines in the factory but it is feasible to include only a sample of $n$ of them in the experiment. The above treatment of a factor with random main effects is then included as a limiting case for $N \to \infty$. Tukey regards the case of a factor with fixed main effects as that where $n = N$. This imposes a certain symmetry on the levels which usually does not correspond to the situation where the model is applied, and was not assumed in the above treatment of the mixed model associated with (3.28), but which does not affect the expected values of the

[8] Extensive moment calculations designed to assess the effects of violations of (ii) and (iii) on $F$-tests have been published by David and Johnson (1951a, 1951b, 1951c, 1952) but the numerical tables promised by them have not yet appeared, except for Horsnell's (1953) paper which utilizes their calculations.

usual mean squares, which are obviously symmetric in the levels (for further discussion, see Scheffé (1956)). The rules for calculating expected mean squares for models where the levels are sampled from finite populations may be found also in the book by Bennett and Franklin (1954, p. 414). Most of Tukey's (1951) work is concerned with moments, higher than the first, of the mean squares and estimated variance components; in this normality is not assumed, but the various sets (for each combination of factors) of random interactions are assumed independent among themselves, of each other, and of the main effects. Some results on moments have been obtained by Hooke (1954a, 1954b) for Tukey's (1949) more realistic formulation of interactions in which the interaction between the levels of two factors depends on the levels obtained in the sampling.

**4. Randomization models.** An important case of the two-way layout is the *randomized blocks design*. To this the example of machines and workers we have carried along thus far is not adapted, and we consider another: Suppose $I$ treatments of a crop are compared on $J$ blocks of $I$ plots each. The essential feature of the design is that in each block the $I$ treatments are assigned to the $I$ plots at random by use of a table of random numbers, or coin tossing, or urn drawing, etc., the randomizations in the $I$ blocks being independent. For this design the following model was formulated by Neyman (1935, pp. 110–112, 145–150): In each block number the plots with $i' = 1, 2, \cdots, I$. Let $\mu_{iji'}$ be the "true" yield under the $i$th treatment on the $j, i'$ plot ($i'$th plot of the $j$th block); this conceptual quantity is regarded as the expected value *if* the $i$th treatment were applied in the $j, i'$ plot. In a thought experiment involving a sequence of repetitions under the same conditions, the observed yields of the $i$th treatment in the $j, i'$ plot would differ from $\mu_{iji'}$ on any particular trial by a *technical error* $e_{iji'}$; this is regarded as a random variable and, by definition of $\mu_{iji'}$, $E(e_{iji'}) = 0$. The randomization by which the treatments are assigned to the plots is independent of the set $\{e_{iji'}\}$. We may write[9]

$$\mu_{iji'} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{iji'}, \tag{4.1}$$

where the general mean $\mu$, treatment main effects $\{\alpha_i\}$, block main effects $\{\beta_j\}$, and treatment-block interactions $\{\gamma_{ij}\}$ are defined in terms of the $\{\mu_{ij.}\}$ as in (1.1) to (1.4), where $\mu_{pq}$ is replaced by $\mu_{ij.}$, and

$$\epsilon_{iji'} = \mu_{iji'} - \mu_{ij.}. \tag{4.2}$$

We note the $\{\epsilon_{iji'}\}$ satisfy $\sum_{i'} \epsilon_{iji'} = 0$ for all $i, j$. The $\epsilon_{iji'}$ is the unit (plot) effect of the $j, i'$ plot specific to the $i$th treatment, and within the $j$th block. If $y_{ij}$ denotes the observation on the $i$th treatment in the $j$th block in the experiment, then

$$y_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \tilde{e}_{ij} + e_{ij}, \tag{4.3}$$

[9] Neyman's notation has been modified to that of this paper; he used a single term $X_{..}(k)$ for our $\mu + \alpha_i$, another $B_i(k)$, for our $\beta_j + \gamma_{ij}$, and his $u_{ij}(k)$ is our $\epsilon_{iji'}$

where $\tilde{e}_{ij}$ and $e_{ij}$ are respectively the $\epsilon_{iji'}$ and $e_{iji'}$ for which $i' = i'(i, j)$ is the plot number to which the $i$th treatment got assigned in the $j$th block. Neyman called the $\{\tilde{e}_{ij}\}$ the "soil errors"; we shall call them the *unit errors.* Employing a convenient notation of Kempthorne (1952), the unit error $\tilde{e}_{ij}$ and technical error $e_{ij}$ may be written

$$(4.4) \qquad \tilde{e}_{ij} = \sum_{i'} d^j_{ii'}\epsilon_{iji'}, \qquad e_{ij} = \sum_{i'} d^j_{ii'}e_{iji'},$$

where the $\{\epsilon_{iji'}\}$ are regarded as unknown constants and the $\{d^j_{ii'}\}$ are $I^2J$ random variables taking on only the values 0 and 1. Their joint distribution, determined by the randomization scheme described above for assigning the treatments to the plots, is evidently the following: For different $j$ the $J$ sets of $I^2$ variables $\{d^j_{ii'}\}$ are independent. For fixed $j$, think of the set $\{d^j_{ii'}\}$ arranged in an $I \times I$ square with $d^j_{ii'}$ in the $i$th row and $i'$th column; then the possible values for the set are the $I!$ ones in which there is exactly one 1 in each row and column and 0's elsewhere, and these $I!$ values are taken on with equal probability. The $\{d^j_{ii'}\}$ are independent of the $\{e_{iji'}\}$; because of this and $E(d^j_{ii'}) = 1/I$ it follows from (4.4) that

$$(4.5) \qquad E(\tilde{e}_{ij}) = 0, \qquad E(e_{ij}) = 0.$$

Neyman showed that an unbiased estimate of any treatment difference $\alpha_i - \alpha_{i''}$ is $y_{i.} - y_{i''.}$; more generally, the same is true for the estimate

$$(4.6) \qquad \hat{\varphi} = \sum_i \lambda_i y_{i.}$$

of any contrast $\varphi = \sum_i \lambda_i \alpha_i (\sum_i \lambda_i = 0)$, from (4.3), (4.5), and (1.6).

Denote the sums of squares for treatments, for blocks, and for interactions (usually called "for error") by $SS_A$, $SS_B$, and $SS_{AB}$; they are defined by (3.1), (3.2), and (3.3) with $K = 1$ and the subscript $k$ deleted. The calculation of the expected values of the corresponding mean squares $MS_A$, $MS_B$, and $MS_{AB}$ is greatly simplified if we can assume the set of technical errors $\{e_{ij}\}$ to be independently and identically distributed, and independently of the set of unit errors $\{\tilde{e}_{ij}\}$. A sufficient condition for this is that the set $\{e_{iji'}\}$ be independently and identically distributed, and this we shall assume—until we discuss tests below.

Neyman calculated[10] $E(MS_A)$ and $E(MS_{AB})$ under some further simplifying assumptions about the $\{\epsilon_{iji'}\}$. The general values without further assumptions follow from formulas first given by Kempthorne (1952, p. 148; the technical errors[11] are assumed negligible there). Resolve the unit effect $\epsilon_{iji'} = \mu_{iji'} - \mu_{ij.}$ of the $j$, $i'$ plot specific to the $i$th treatment into a unit main effect within the $j$th block,

$$(4.7) \qquad \xi_{ji'} = \mu_{.ji'} - \mu_{.j.},$$

[10] Professor O. Kempthorne pointed out to me a slip in Neyman's calculation; its effect is the loss of the term $\sigma^2_{AB}$ from $E(MS_{AB})$.

[11] Technical errors are included in randomization models by Wilk (1955b).

plus a treatment-unit interaction within the $j$th block,

$$\eta_{iji'} = \mu_{iji'} - \mu_{.ji'} - \mu_{ij.} + \mu_{.j.}\,, \tag{4.8}$$

and define the symbols $\sigma_U^2$ and $\sigma_{AU}^2$ (corresponding to a "units factor" and to its interaction with the "treatments factor") as

$$\sigma_U^2 = J^{-1}(I-1)^{-1}\textstyle\sum_j \sum_{i'} \xi_{ji'}^2\,, \tag{4.9}$$

$$\sigma_{AU}^2 = J^{-1}(I-1)^{-2}\textstyle\sum_i \sum_j \sum_{i'} \eta_{iji'}^2\,. \tag{4.10}$$

Then with $\sigma_A^2$, $\sigma_B^2$, $\sigma_{AB}^2$ defined by (3.8), (3.9), (3.14), and

$$\sigma_e^2 = \operatorname{Var}(e_{ij}), \tag{4.11}$$

the desired formulas are

$$E(MS_A) = J\sigma_A^2 + \sigma_U^2 + I^{-1}(I-2)\sigma_{AU}^2 + \sigma_e^2\,, \tag{4.12}$$

$$E(MS_B) = I\sigma_B^2 + I^{-1}(I-1)\sigma_{AU}^2 + \sigma_e^2\,, \tag{4.13}$$

$$E(MS_{AB}) = \sigma_{AB}^2 + \sigma_U^2 + I^{-1}(I-2)\sigma_{AU}^2 + \sigma_e^2\,. \tag{4.14}$$

It is easy to derive an expression for the variance of the estimated contrast (4.6) in terms of the unknown parameters $\{\epsilon_{iji'}\}$ and $\sigma_e^2$, but there exists no unbiased estimate of it. An overestimate can however be obtained by estimating the contrast separately for each block by

$$\hat{\varphi}_j = \textstyle\sum_i \lambda_i y_{ij}\,, \tag{4.15}$$

and using the sample variance of these $J$ estimates,

$$s^2 = (J-1)^{-1}\textstyle\sum_j (\hat{\varphi}_j - \hat{\varphi}_.)^2. \tag{4.16}$$

From (4.3), $\hat{\varphi}_j = \varphi + \theta_j + u_j$, where $\theta_j = \sum_i \lambda_i \gamma_{ij}$, and $u_j = \sum_i \lambda_i(\tilde{e}_{ij} + e_{ij})$, the $u_j$ having zero means, and being independent since they are calculated from different blocks. Since $\hat{\varphi} = \hat{\varphi}_.$, therefore $\operatorname{Var}(\hat{\varphi}) = J^{-2}\sum_j \operatorname{Var}(u_j)$. Then $s^2/J$ is an overestimate of $\operatorname{Var}(\hat{\varphi})$ in the sense that $E(s^2/J) \geqq \operatorname{Var}(\hat{\varphi})$, since $E(s^2/J) = J^{-1}(J-1)^{-1}\sum_i \theta_i^2 + \operatorname{Var}(\hat{\varphi})$. Clearly $s^2/J$ is an unbiased estimate if $\sigma_{AB}^2 = 0$.

The problem of statistical tests under the randomization model associated with (4.3) is complicated. Let us call *normal-theory model* that in which the terms $\{e_{ij}\}$ in (4.3) are independently normally distributed with zero means and equal variance $\sigma_e^2$, while the $\tilde{e}_{ij}$ are always zero (which is equivalent to all $\epsilon_{iji'} = 0$). The usual $F$-test for treatment effects is then a test of the hypothesis $\sigma_A^2 = \sigma_{AB}^2 = 0$ in the normal-theory model. Its power is usually considered against alternatives with $\sigma_{AB}^2 = 0$, in which case the power can be expressed in terms of the noncentral $F$-distribution. The randomization model seems very far removed from the normal-theory model if the unit effects $\{\epsilon_{iji'}\}$ are not small compared with the $\sigma_e^2$ characterizing the technical errors. Nevertheless, following Fisher (1935, Sec. 21) it has become a common belief among statisticians that referring the usual $F$-statistic (in the above case, $MS_A/MS_{AB}$) to the tables valid under the normal-theory model gives a good approximation in some sense to the exact test under the randomization model, which we shall describe in a moment and shall call the permutation test based on the $F$-statistic. The writer has had difficulty in trying to formulate clearly the sense in which this approximation is expected to hold.

The null hypothesis for the permutation test specifies that there is no difference whatever among the treatments, that is, under the above randomization model that $\sigma_A^2 = \sigma_{AB}^2 = \sigma_{AU}^2 = 0$ and the joint distribution of the $IJ$ random variables $\{e_{iji'}\}$ where $i' = i'(i, j)$ depends on the assignment of treatments to plots, is the same for every one of the $(I!)^J$ assignments. This is equivalent to the hypothesis that the joint distribution of the $IJ$ observations under any of the assignments is the same for all the assignments.

The permutation test based on the $F$-statistic is made as follows: Consider the group $G$ of all $m$ permutations of the observations which leaves their distribution invariant under the null hypothesis (in the above case $G$ consists of the $m = (I!)^J$ permutations obtained by making all possible $I!$ permutations within each block). If these $m$ permutations are made on the observations actually obtained in the experiment and the $F$-statistic calculated as though the permuted observations had been obtained, a set of $m$ (in general not distinct) values of $F$ will be generated. The idea is to reject the hypothesis at the $\alpha$ level of significance if the value of $F$ actually obtained lies among the $\alpha m$ largest of the $m$ values—some obvious qualifications have to be made because $\alpha m$ may not be an integer and there is trouble about the $m$ values not being distinct (for further details see Scheffé (1943) or Hoeffding (1952)). For any fixed set of observations there is thus determined a "significance level" $F_\alpha$ for the statistic $F$ so that we reject if $F > F_\alpha$, but $F_\alpha$ is a random variable depending on the outcome of the experiment, and we write $F_\alpha = F_\alpha(y)$ to indicate this. In most of the potential applications of the permutation test, the value of $F_\alpha(y)$ is extremely tedious to calculate. The evidence that the usual $F$-test approximates the permutation test is of three kinds:

First, numerical examples have been published where for particular sets of observations it transpires that $F_\alpha(y)$ is close to the value in the $F$-tables; see for example Eden and Yates[12] (1933), Fisher (1935, Sec. 21), Welch (1937, p. 31), Pitman (1938, p. 334), and Kempthorne (1952, p. 132).

Second, there are moment calculations, up to fourth-order moments, made on a transform of the $F$-statistic which has the incomplete beta distribution under the hypothesis in the normal-theory model. These were made for ran-

[12] Their results can be regarded as a comparison with values in the $F$-tables of estimates of $F_\alpha(y)$ obtained by empirical sampling of the permutation distribution of the $F$-statistic, for various levels $\alpha$ and a single set of "observations" $y$ (not the actual observations but averages of sets of 8 observations in a uniformity trial in randomized blocks; also, they use $z = \frac{1}{2}\log F$ instead of $F$. This paper is clarified by a discussion between Yates (1935b, pp. 164, 165) and Neyman.

domized blocks by Welch (1937) and Pitman (1938), who worked under the above model with the restriction that the technical errors were assumed identically zero. It is easy to remove this restriction by a conditional probability argument about the probability of rejecting the hypothesis when true. A measure of the magnitude of the unit effects $\{\xi_{ji'}\}$ in the $j$th block is $\sigma_{U,j}^2 = (I - 1)^{-1}\sum_{i'} \xi_{ji'}^2$. Pitman's calculations on the first four moments indicate that if the $\{\sigma_{U,j}^2\}$ do not differ too widely for the $J$ blocks, then $F_\alpha(y)$ should be close to the value in the $F$-tables. This is for the case of zero technical errors: under this assumption and the null hypothesis all the parameters can be calculated exactly from the observations, and in particular $\sigma_{U,j}^2$ becomes

$$(I - 1)^{-1}\sum_i (y_{ij} - y_{i.})^2. \tag{4.17}$$

If we do not assume zero technical errors, then the condition of not too great difference of the $\{\sigma_{U,j}^2\}$ characterizing the blocks is replaced by the same condition on the functions (4.17) of the observations. Welch (1937) and Pitman (1938) also made moment calculations for the Latin square in the same papers; the calculation for the first moment had been published earlier by Yates (1933). The approximation of the $F$-test to the permutation test for randomized blocks can be improved by adjusting as follows the numbers of *d.f.* with which the $F$-tables are entered: The first moment of the above-mentioned transform of the $F$-statistic in its permutation distribution does not depend on the observations and is the same as under the normal-theory model and the null hypothesis; the second moment is determined by the quantities (4.17). If, as suggested by Welch and Pitman, we choose the numbers of *d.f.* of an approximating incomplete beta distribution to give the correct first two moments, this is equivalent to referring the $F$-statistic to the $F$-tables with the same numbers of *d.f.* This correction to the numbers of *d.f.* is given by Box and Andersen (1954), as well as a similar correction for the one-way layout.

Third, there are some asymptotic calculations. As the number $J$ of blocks increases with the number $I$ of treatments fixed, the limiting distribution of the $F$-statistic under the normal-theory model is chi-square with $I - 1$ *d.f.* Wald and Wolfowitz (1944) showed that as $J$ increases with fixed $I$, if the sequence of observations satisfies certain restrictions, then the permutation distribution of the $F$-statistic has the same limiting form. Hoeffding (1952) proved that as $J$ increases with fixed $I$, then under certain assumptions on the sequence of distributions of the observations, the random variable "significance level" $F_\alpha(y)$ of the permutation test approaches a constant in probability. With this he was able to show that the permutation test had in a certain sense asymptotically the same power as the usual $F$-test against alternatives of the normal-theory model. Of course, what we would like to know more about is the power of the usual $F$-test against the alternatives allowed by the randomization model. An asymptotic calculation similar to Wald and Wolfowitz's just mentioned was carried out for the one-way layout by Silvey (1954).

Randomization models have been formulated and expected mean squares calculated for many other designs by Kempthorne (1952, 1955), Wilk (1955a, 1955b) and Wilk and Kempthorne (1953–1955, 1955). Fisher (1926) was the pioneer in emphasizing the importance of randomization and in conceiving of permutation tests (1925, Sec. 24; 1935, Sec. 21). In introducing randomized blocks Fisher (1926) did not formulate explicitly a model like Neyman's, above, and one might infer he had in mind a more restricted one, in particular with $\sigma_{AB}^2 = 0$, since he claims that "One way of making sure that a valid estimate of error will be obtained is to arrange the plots deliberately at random. . . ." The mathematical model for the completely randomized experiment was given by Neyman (1923)[13] under the restriction of zero technical errors.

The writer has had the benefit of solicited comments on this paper from the following persons, and has improved the paper by incorporating many of their suggestions, for which he is very grateful: Professors William G. Cochran, Oscar Kempthorne, William Kruskal, H. Fairfield Smith, John W. Tukey, and Martin B. Wilk.

## REFERENCES

G. B. Airy (1861), *On the Algebraical and Numerical Theory of Errors of Observations and the Combination of Observations*, Macmillan and Co., Ltd., Cambridge and London.

R. L. Anderson and T. A. Bancroft (1952), *Statistical Theory in Research*, McGraw-Hill Book Company, Inc., New York.

C. A. Bennett and N. L. Franklin (1954), *Statistical Analysis in Chemistry and the Chemical Industry*, John Wiley & Sons, New York.

I. Bessel (1820), "Beschreibung des auf der Königsberger Sternwarte aufgestellten Reichenbachschen Meridiankreises, dessen Anwendung und Genauigkeit, imgleichen der Repoldschen Uhr," *Astronomisches Jahrbuch für das Jahr* 1823, Berlin.

G. E. P. Box (1953), "Non-normality and tests on variances," *Biometrika*, Vol. 40, pp. 318–335.

G. E. P. Box (1954), "Some theorems on quadratic forms applied in the study of analysis of variance problems. I. Effect of inequality of variance in the one-way classification," *Ann. Math. Stat.*, Vol. 25, pp. 290–302; "II. Effects of inequality of variance and of correlation in the two-way classification," *Ann. Math. Stat.*, Vol. 25, pp. 484–498.

G. E. P. Box and S. L. Andersen (1954), *Robust Tests for Variance and Effect of Non-normality and Variance Heterogeneity on Standard Tests*, Inst. of Stat. Mimeo. Series No. 101. Published as "Permutation theory in the derivation of robust criteria and the study of departures from assumption" in *J. Roy. Stat. Soc.*, Series B, Vol. 17 (1955), pp. 1–34.

W. Chauvenet (1863), *A Manual of Spherical and Practical Astronomy*, 2 Vols., J. B. Lippincott Company, Philadelphia.

---

[13] In this paper Neyman uses an urn model for the completely randomized experiment with zero technical errors to prove that the usual estimates of variety differences are unbiased, and to calculate the variances of these estimates. He compares his results with "normal theory" calculations, concluding that "normal theory" may often lead wrongly to the conclusion that the estimates are not significantly different from zero. He is pessimistic about the effects of assuming in practice that a certain correlation coefficient is unity—this corresponds to assuming no plot × variety interactions. Results of a sampling experiment with an artificial set of plot effects and identical varieties are given and show good agreement with "normal theory." Other parts of this early paper of Neyman's are of historical interest, in showing that the evolution of his ideas on statistical inference passed through a Bayesian stage.

S. L. Crump (1946), "The estimation of variance components in analysis of variance," *Biometrics*, Vol. 2, pp. 7–11.

S. L. Crump (1951), "The present status of variance component analysis," *Biometrics*, Vol. 7, pp. 1–16.

H. E. Daniels (1938), "The effects of departures from ideal conditions other than non-normality on the $t$ and $z$ tests of significance," *Proc. Cambridge Philos. Soc.*, Vol. 34, pp. 321–328.

H. E. Daniels (1939), "The estimation of components of variance," *J. Roy. Stat. Soc., Suppl.*, Vol. 6, pp. 186–197.

F. N. David and N. L. Johnson (1951a), "The effects of non-normality on the power function of the $F$-test in the analysis of variance," *Biometrika*, Vol. 38, pp. 43–57.

F. N. David and N. L. Johnson (1951b), "A method of investigating the effect of non-normality and heterogeneity of variance on tests of the general linear hypothesis," *Ann. Math. Stat.*, Vol. 22, pp. 382–392.

F. N. David and N. L. Johnson (1951c), "The sensitivity of analysis of variance tests to random between groups variation," *Trabajos de Estadistica*, Vol. 2, pp. 179–188.

F. N. David and N. L. Johnson (1952), "Extension of a method of investigating the properties of analysis of variance tests to the case of random and mixed models," *Ann. Math. Stat.*, Vol. 23, pp. 594–601.

T. Eden and F. Yates (1933), "On the validity of Fisher's $z$-test when applied to an actual sample of non-normal data," *J. Agricultural Science*, Vol. 23, pp. 6–16.

C. Eisenhart (1947), "The assumptions underlying the analysis of variance," *Biometrics*, Vol. 3, pp. 1–21.

R. A. Fisher (1918), "The correlation between relatives on the supposition of Mendelian inheritance," *Trans. Roy. Soc. Edinburgh*, Vol. 52, pp. 399–433.

R. A. Fisher (1925), *Statistical Methods for Research Workers*, Oliver and Boyd, Ltd., Edinburgh.

R. A. Fisher (1926), "The arrangement of field experiments," *J. Ministry Agr.*, Vol. 33, pp. 503–513. Reprinted as Paper 17 in *Contributions to Mathematical Statistics*, John Wiley & Sons, New York, 1950.

R. A. Fisher (1935), *The Design of Experiments*, Oliver and Boyd, Ltd., Edinburgh.

K. F. Gauss (1809), *Theoria motus corporum coelestrium in sectionibus conicis solem ambientium*, Perthes and Besser, Hamburg.

W. Hoeffding (1952), "The large sample power of tests based on permutation of observations," *Ann. Math. Stat.*, Vol. 23, pp. 169–192.

R. Hooke (1954a), *Moments of Moments in Matrix Sampling—An Extension of Polykays*, Memorandum Report 55, Statistical Research Group, Princeton University, 25 + 6 pp. Results of this and the next paper are given in a paper by Hooke to appear in *Ann. Math. Stat.*

R. Hooke (1954b), *The Estimation of Polykays in the Analysis of Variance*, Memorandum Report 56, Statistical Research Group, Princeton University, 11 + 18 pp.

G. Horsnell (1953), "The effect of unequal group variances on the $F$-test for the homogeneity of group means," *Biometrika*, Vol. 40, pp. 128–136.

R. W. B. Jackson (1939), "The reliability of mental tests," *British J. Psych.*, Vol. 29, pp. 267–287.

O. Kempthorne (1952), *The Design and Analysis of Experiments*, John Wiley & Sons, New York.

O. Kempthorne (1955), "The randomization theory of experimental inference," *J. Amer. Stat. Assoc.*, Vol. 50, pp. 946–967.

A. M. Legendre (1806), *Nouvelles méthodes pour la détermination des orbites des comètes; avec un supplément contenant divers perfectionnements de ces méthodes et leur application aux deux comètes de* 1805, Courcier, Paris.

A. M. Mood (1950), *Introduction to the Theory of Statistics*, McGraw-Hill Book Company, Inc., New York.

J. NEYMAN (1923), "Sur les applications de la théorie des probabilités aux expériences agricoles. Essay des principes," *Roczniki Nauk Rolniczych*, Vol. 10, pp. 1–51. Polish with German summary.

J. NEYMAN (1935), with cooperation of K. IWASZKIEWICZ and S. KOLODZIEJCZYK, "Statistical problems in agricultural experimentation," *J. Roy. Stat. Soc., Suppl.*, Vol. 2, pp. 107–154.

E. S. PEARSON (1931), "The analysis of variance in cases of nonnormal variation," *Biometrika*, Vol. 23, pp. 114–133.

E. J. G. PITMAN (1938), "Significance tests which may be applied to samples from any populations. III. The analysis of variance test," *Biometrika*, Vol. 29, pp. 322–335.

R. L. PLACKETT (1949), "A historical note on the method of least squares," *Biometrika*, Vol. 36, pp. 458–460.

H. SCHEFFÉ (1943), "Statistical inference in the non-parametric case," *Ann. Math. Stat.*, Vol. 14, pp. 305–332.

H. SCHEFFÉ (1952), "An analysis of variance for paired comparisons," *J. Amer. Statist. Assoc.*, Vol. 47, pp. 381–400.

H. SCHEFFÉ (1953), "A method for judging all contrasts in the analysis of variance," *Biometrika*, Vol. 40, pp. 87–104.

H. SCHEFFÉ (1956), "A mixed model for the analysis of variance," *Ann. Math. Stat.*, Vol. 27, pp. 23–36.

S. D. SILVEY (1954), "The asymptotic distribution of statistics arising in certain non-parametric tests," *Proc. Glasgow Math. Assoc.*, Vol. 2, pp. 47–51.

L. H. C. TIPPETT (1929), "Statistical methods in textile research. The analysis of complex variations," *Shirley Inst. Memoirs*, Vol. 8, pp. 175–196.

L. H. C. TIPPETT (1931), *The Methods of Statistics*, Williams and Norgate, Ltd., London.

J. W. TUKEY (1949), *Interaction in a Row-by-Column Design*, Memorandum Report 18, Statistical Research Group, Princeton University, 14 pp. The results will be included in a paper by Cornfield and Tukey to appear in *Ann. Math. Stat.*

J. W. TUKEY (1951), *Finite Sampling Simplified. Polykays, Cumulants of Variances, and Cumulants of Variance Components*, Memorandum Report 45, Statistical Research Group, Princeton University, 112 pp. The table of contents is on p. 9.

A. WALD AND J. WOLFOWITZ (1944), "Statistical tests based on permutations of the observations," *Ann. Math. Stat.*, Vol. 15, pp. 358–372.

B. L. WELCH (1937), "On the $z$-test in randomized blocks and Latin squares," *Biometrika*, Vol. 29, pp. 21–52.

M. B. WILK (1955a), "The randomization analysis of a generalized randomized block design," *Biometrika*, Vol. 42, pp. 70–79.

M. B. WILK (1955b), *Linear Models and Randomized Experiments*, Ph.D. thesis, Iowa State College.

M. B. WILK AND O. KEMPTHORNE (1953–1955), a series of reports beginning December, 1953, on models for the analysis of variance, Analysis of Variance Project, Statistical Laboratory, Iowa State College.

M. B. WILK AND O. KEMPTHORNE (1955), "Fixed, mixed, and random models," *J. Amer. Statist. Assoc.*, Vol. 50, pp. 1144–1167.

F. YATES (1933), "The formation of Latin squares for use in field experiments," *Empire J. Experimental Agric.*, Vol. 1, pp. 235–244.

F. YATES (1935a), "Complex experiments," *J. Roy. Stat. Soc., Suppl.*, Vol. 2, pp. 181–247.

F. YATES (1935b), Discussion of Neyman's 1935 paper, *J. Roy. Stat. Soc., Suppl.*, Vol. 2. pp. 161–166.