

VARIANCE COMPONENTS, FINITE POPULATIONS, AND EXPERIMENTAL INFERENCE

Prepared Under Contract No. DA-36-034-ORD-1517 (RD)  
(Experimental Designs for Industrial Research)

by

H. F. Smith

Institute of Statistics  
Mimeo Series No. 135  
July, 1955

# VARIANCE COMPONENTS, FINITE POPULATIONS AND EXPERIMENTAL INFERENCE

by

H. F. Smith

North Carolina State College

## Abstract

Recent development of models for analyses of variance has been concerned with the effect of postulating that an observed set of treatments is a sample from a finite class of treatment variants, and with consequent effects on definitions of variance components. In this paper the word population is used to imply an infinite population, the word universe to imply a finite population. The one extreme, infinite populations of all classifications, is Eisenhart's model II, commonly described as the "random model". The other extreme, observing all members of small universes, is commonly taken to be equivalent to Eisenhart's model I. These two do not however have the same philosophical background. Model I postulates that we wish to evaluate the means, parameters or statistics of location, for each observed treatment irrespective of any universe from which they may have been selected; it is described as a regression model, regression techniques being used to evaluate the statistics of location. The variance component model on the other hand postulates no interest in individual treatment means but only in the dispersion of the elements of their universe; it becomes similar to model I only incidentally, when a complete universe or sub-universe is observed, in the sense that we then have the means to completely describe the universe, including its dispersion, by an enumeration of its elements.

Section 1 notes that statistical analyses are formal. The parts into which we imagine an observation or variance to be divisible are not physical entities dictated by nature; they are defined empirically for purposes of statistical

description. The criterion for preference among alternative models is simplicity and convenience of consequent statistics: nothing more fundamental.

Sections 3 to 6 review standard models and their interpretation -- a setting of the stage for subsequent discussion.

Section 7 endeavours to tighten up the definition of some statistical concepts, because, although the topic may seem pedantic, a deal of controversy seems to trace back to small differences in usage and interpretation of words by different workers. Among other things the conclusion is reached that Tukey's  $k_2$  statistics for a universe cannot be defined as variances without consequent inconsistencies. They are therefore referred to throughout this paper as k-statistics. Their parametric values are denoted by the capital,  $K$ ; the greek letter, usual designation for a parameter, having been reserved for the analogous parameters of infinite populations.

Section 8 presents the general model for universes of any size. Analysis of variance in terms of generalized symmetric means then indicates how we can define components of mean squares which are "inherited on the average" (Tukey, 1950). They are therefore an extension to components of the generalized k-statistics defined by Tukey for moments of universes and samples from them. Like these the definition of sample values is independent of the size of parent universe and therefore is the same as for the variance components of a random (infinite population) model. Being invariant for alterations of hypothetical universe sizes they are christened canonical variance components.

Variance components as usually defined for finite universes are here distinguished as 'specific variance components'. Their formulae appear to be most easily derivable as a linear function of the canonical components (sec. 9). However since inferences may be made just as well, and often better, in terms of canonical components the specific ones may seldom be worth evaluating. Recollecting

that the criterion of a good statistical model is simplicity of the descriptive and test statistics to which it leads, a modified model is proposed (the canonical latent equation) such that the variances of its groups of elements may be the canonical variance components.

Sections 10 and 11 consider the interpretation of regression and mixed models as limiting cases of the generalized model. The most important point to emerge is that controversy over the mixed model derives from universal oversight that when one turns to examine the means of observed variants of what was defined as the random factor one has abandoned the mixed model as originally defined and is now applying regression interpretation to that factor. When that is realized the restrictions formerly laid down on the "fixed factor" are seen to be irrelevant to evaluating means of observed variants of the postulated "random factor"; and the debate is resolved. Section 12 maintains that a significance test, like statistical estimation discussed in sec. 7, always involves a hypothetical infinite population and that appropriate error mean squares are better taken to be those indicated by canonical, rather than by specific, variance components.

Section 13 examines the ideas promulgated by Kempthorne and Wilk for "logical derivation" of linear models and their analyses for experimental situations. It finds: (1) that their definition of the random variables generated by sampling elements of a universe lacks theoretical validity and merely produces an unnecessarily cumbersome algebraic system; (2) the randomization tests which they advocate do not (within the framework of the theory) produce the answers which an experimenter requires; and (3) insertion of unidentifiable interactions in the model introduces redundant complications. The last follows from recognizing that an interaction can properly be defined only for reproducible effects of variants which can be identified

in replications of the experimental units. Unidentifiable interactions are a part of experimental error and need to be considered only in so far as errors may be heterogeneous, for example to distinguish between main-plot and split-plot errors.

Section 14 (not yet written) presents some examples to illustrate the principles discussed.

VARIANCE COMPONENTS, FINITE POPULATIONS, AND EXPERIMENTAL INFERENCE

by

H. Fairfield Smith

North Carolina State College

Contents

0. Introduction.	1 - 2
1. Analyses are formal.	3 - 5
2. General notation and definitions.	6 - 8
3. Variance component models.	9 - 11
4. Regression on quantitative factors	12 - 16
5. Regression on qualitative factors.	17 - 19
6. Random model.	20 - 26
7. Some finite population theory	27 - 39
Appendix to Section 7. (Definition of parameters)	40 - 44
8. General model.	45 - 58
Appendix to Section 8. (Variances of mean squares)	59 - 66
9. The canonical variance model.	67 - 76
10. Regression as limit of general model.	77 - 78
11. The mixed model.	79 - 88
12. Error terms and significance tests.	89 - 95
13. Randomization tests and unidentified interactions.	96 - 104
14. Examples	
Summary	
Literature Cited	

## Introduction

In practice analysis of variance is not difficult to interpret relative to a given purpose even although there may be some ambiguity about the underlying model to best represent facts. The two ways in which analysis of variance may be used were expounded by Eisenhart (1947). Often however one may want (as it has been described) to view a given set of data "through either pair of spectacles" and the alternative may occasion no trouble. Nevertheless with usage being extended over ever increasing range of applications, many have felt it desirable to get the various models more rigorously formulated and synthesized. To this end, consideration has been given to interpretations centering around postulates that variants or levels of factors in a factorial experiment are representatives of limited (finite and perhaps small) universes of variants. Alternative postulates, sometimes implied rather than defined, have led to controversy on whether or not main effect mean squares should or should not contain interaction variance components in the mixed model and what coefficients should be associated with them in more general models.

The main purpose of this paper is to define a set of "canonical" variance components (Sec. 8) which depend only on the design of an experiment and are invariant under alternative postulates about the universes supposed to have been sampled. They have useful properties, in particular that there are unique unbiased estimators which are "inherited on the average" (Tukey, 1950), and they lend themselves to flexible interpretation of experimental data with universe postulates introduced only as part of the process of interpretation instead of as a part of the model. Their sample estimators are defined in the same way as for variance components in the "random" model, as is required by their invariance and

the condition that both should be asymptotically equivalent. For elementary exposition they may be regarded as parameters and estimators for hypothetical infinite populations, leading to inferences relative to finite universes with "finite population corrections" as developed by Yates and Zaccopani (1935), Cochran (1939), Hendricks (1947, 1951), Irwin and Kendall (1944), (Sec. 9).

Endeavor to obtain definitions which would be always consistent indicated that ambiguity is creeping into current usage of many terms such as parameter, variance, fixed effects. Indeed much contemporary discussion on analysis of variance and variance components seems to derive from varying use and interpretation of words; it poses, at least in part, a problem in semantics. Therefore the first seven sections endeavor to set down just what we mean by certain concepts. The later sections develop the application of canonical variance components to experimental inference.

1. Statistical analyses are formal.

Analyses of observations,  $y$ , into components according to a linear model such as

$$y_{ijk} = \mu + \alpha_i + \beta_j + \dots + \epsilon_{ijk} \quad (1.1)$$

or an analysis of variance symbolized by

$$\sigma_y^2 = \sigma_\alpha^2 + \sigma_\beta^2 + \dots + \sigma_\epsilon^2 \quad (1.2)$$

are purely formal, the respective components being defined solely for simplicity of statistical description. The respective components may be associated with causal influences; but they do not represent physically real or distinguishable parts of  $y$  or of  $\sigma_y^2$ ; nor even of the "response" to a treatment, which should naturally be evaluated as deviation from the yield with no treatment rather than from the mean for a set of treatments.

If we make chemical analyses of, for example, equal volumes of normal solutions of NaCl and of KCl, the analysis is dictated by nature - so much water, chlorine, and sodium or potassium. We can say that the solutions have the same amounts of water and chlorine in common but differ in their cations.

For contrast suppose hardness of rubber may be as indicated in Table 1 after vulcanizing with alternative qualities of sulphur,  $\alpha_i$ , and of carbon black,  $\beta_j$ , in all combinations. The results are to be expressed as a linear function representing contributions from each factor. A logical suggestion might be to define  $\mu$  = the base of reference = 1;  $\alpha$  = effect of sulphur = 6 for either type;  $\beta$  = effect of carbon black = 1;  $(\alpha\beta)_{ij}$  = interaction of  $\alpha_i$  with  $\beta_j = 1, 3, 2$  and  $4$  for  $ij = 11, 12, 21$  and  $22$  respectively, the gains over 9 being regarded as due to chemical interactions. But these definitions fail to simplify things in the way we want. The zero levels of either ingredient are likely to be uninteresting and unobserved; how much simpler to summarize the other four combinations by  $\mu = 11.5$ ,  $\alpha_i = +.5$ ,  $\beta_j = +1$ ,  $(\alpha\beta)_{ij} = 0$ , although chemically speaking these may be arbitrary definitions.

Table 1

	0	1	2
0	1	2	2
1	7	10	12
2	7	11	13

In that example we considered only description of certain fixed figures free of sampling variations which are the statistician's major concern. Consider next measuring sizes of seeds in pods of a legume. They are too multitudinous to be individually enumerated and we must be content with representative observations from which a picture of the whole may be formulated. Earlier statisticians described the complex by saying that seeds within a pod are "correlated" and sought to devise a measure of the correlation. Nowadays it is obviously more clearly and comprehensibly described by analysis of variance and its associated "linear model" for the  $j$ th seed in the  $i$ th pod

$$y_{ij} = \mu + \pi_i + \delta_{ij} \quad (1.3)$$

We speak of seeds in the same pod as having  $\mu + \pi_i$  "in common", but it is a very different communality from that exhibited in the chemical analysis -- the seeds are not physically divisible into corresponding parts, and definitions of the parts are arbitrary. A classical taxonomist would perhaps define  $\mu$  for a "type" pod, and  $\pi_i$  for a type position, say the proximal seed. A "finite populationist" describing a single tree would define  $\mu$  as the mean weight of all seeds in the tree,  $\mu + \pi_i$  as the mean seed weight in the  $i$ th pod, and his definition of the variance component  $\sigma^2$  would coincide exactly with the actually existing variance of pod means. The biometrician will usually prefer to define them for a hypothetical population which will not be exactly realizable in any observed tree. All are reasonable formulations and the biometrician's choice is determined only by statistical convenience. It is nothing more or less than a device to simplify distribution of correlated variates by division into parts whose distributions are uncorrelated.

Adopting a linear model is from the start arbitrary. The realistic analysis of yield of a cereal plant into components which have physical meaning -- ear number ( $e$ ), number of grains per ear ( $n$ ) and weight per grain ( $g$ ) -- is

$$y = e.n.g.$$

Treatment effects often combine in a similar or more complex way, yet for statistical purposes the empirical linear equation may be worth using. Analogous variants are possible for the variance equation (1.2) and will be illustrated in the sequel.

Statistical components, either of an observation  $y$  or of its variance  $\sigma_y^2$ , are dictated only in part by the structure of a population, sample or experiment. Detail of definitions is arbitrary and can be formulated to produce as much simplicity of description as possible.

## 2. General notation and definitions.

The structure of a population or experiment will mean the ways in which data may be classified according to factors, the number of variants of each of these, and whether they are "crossed", (e.g. potato varieties x insect sprays, temperature x concentration of a reagent) or "nested" (e.g. districts within states). When a factor is continuous (e.g. temperature) we shall speak of its levels; when discrete (e.g. varieties) of its variants.

There must always be a close relation between an experiment and the population about which it gives information. Ideally we should consider first the population structure and select an experimental design to match; frequently action precedes such thought and we reconstruct the population which a given design can reasonably represent. For our purpose the order is immaterial as we can take the one to dictate the structure of the other, except that the design need not state the number of variants of each factor in the population. Actually a given experiment can answer questions relative to a variety of populations, for example we can seek to evaluate yields for variants of one factor when crossed by certain selected variants of another or as averaged for a population of them. We shall therefore take the experimental design as basic structure, and suggest a general analysis which will depend only on that design and be readily adaptable to answering questions relative to any reasonable postulates about the sorts of populations which it might represent. These will be introduced only as secondary postulates with specific questions, and not be laid down as part of the initial model.

We speak of yield as the measure of an observation on each experimental unit or plot whether this be actually a yield as of a field crop or any other characteristic (e.g. hardness). We write yield as a linear function of elements (e.g. 1.1

or 2.1) which can be associated with each factor level or variant in a given treatment combination. A linear function is used as routine merely because it is the simplest mathematically. Since we always have at least as many elements as observations these equations are identities.

The number of elements into which we imagine any yield analyzable is dictated by the structure except that there may often be some discretionary choice about the number of interaction terms to be inserted, i.e. that are deemed worth considering, depending on what we already know about the response of yield to the treatments and how well it may be represented by a linear function with fewer than the maximum number of elements. Throughout this paper we assume "balanced" arrangements in both experiment and population. (The meaning of balanced in this context is defined by Crump, 1951).

For the sake of specificity consider a standard example which represents all main features, namely a two-factor experiment with a variants of factor *A* crossed by b variants of factor *B* and with d random replications of each treatment combination. The linear model is

$$y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \delta_{ijk} \quad (2.1)$$

$$i = 1, 2 \dots a \dots A \leq \infty$$

$$j = 1, 2 \dots b \dots B \leq \infty$$

$$k = 1, 2 \dots d \dots D \leq \infty$$

$B$  = the potential number of variants of *B* of which b are observed, etc

Elements indicated by the same letter will be called a group.

When dealing with finite populations, a given interaction element  $(\alpha\beta)_{ij}$  must be associated with a particular  $\alpha_i$  and  $\beta_j$ . Whether or not a similar condition

applies for  $\delta_{ijk}$  depends on circumstances: the postulated cause of deviation represented by this element and whether or not it can be randomized with the treatment combinations.

A subscript zero will indicate a mean of a finite universe, for example

$$y_{ij_0} = \sum^D y_{ijk}/D;$$

a subscript dot will indicate a mean of a sample

$$y_{ij.} = \sum y_{ijk}/d.$$

So far as these definitions have gone the elements are still indeterminate. Without altering  $y$  a constant could be added to every element of one group and subtracted from every element of another; or added to just one  $\alpha_i$  or  $\beta_j$  and subtracted from the corresponding sub-group of  $(\alpha\beta)_{ij}$ ; similarly for exchanges with the  $\delta$  group but with sub-groups depending on the randomization postulates. To obtain unique definition of the elements the location of each group or sub-group must be pinned and we can do this arbitrarily. For algebraic convenience we define  $\mu$  to be the mean of the whole complex population, and the means of every group or sub-group to be zero.

### 3. Classification of models.

Eisenhart (1947) gave a clear and explicit exposition of two alternative models underlying analyses of variance. I: The objective is to evaluate mean yields of an observed set of treatments; the analysis of variance is an algorithm for tests of significance for differences between these means, usually by groups. It is a standardized formulation of tests which follow from classical least squares procedure for estimating the means, the parameters of location, which can be formulated as regression coefficients. This is therefore described as the regression model. II: The objective is to estimate the contribution of each factor to variability in a complex population; an analysis of variance in the literal meaning of the words. In effect we now say that we are not interested in the yields associated with individual factor variants, but only in the dispersion of yields associated with a population of variants. Interpretation is simplest when we can take the observed variants to be a sample from a potentially infinite number; but this is not an essential part of the model. Even if every potential variant of a factor be observed the objective in view may require only an assessment of the variability for which that factor can be held responsible. For example in studying variability of a given kind of cloth we may observe every spindle in the looms which produce it and all the information we may need is how much of cloth variation is due to variability among the spindles irrespective of the effect of each one individually. This is described as: variance component analysis.

These two formulations have more recently been well described by Hoel (1954), under the names "linear hypothesis model" and "components of variance model", in a review which skillfully avoids complexities of the subject.

Tukey (1949) gave a more elaborate classification. Grump (1951) notes that its "general features...are common to any set of data arranged in a multiple classification and described by a linear model". However it was introduced under the section heading "Estimating effect variances", its exposition included the sentence "To each kind of effect there corresponds a corresponding effect variance or component of variance", and it is frequently taken as a classification of variance component models. Tukey writes of row, column and cell "effects" indiscriminately whether they are to be regarded as parameters of location or random variables, and the above quotations apparently relate to the stand taken in his 1950 paper "to define the variance of any finite set of numbers as  $k_2$ , whether the finite set is a sample or a population". (Contrast the practice of Anderson and Bancroft (1952) and of other writers who, when dealing with a regression or mixed model, write expectations of mean squares with an explicit function of "fixed effects" which they refrain from regarding as a variance.) Sec. 7 will demonstrate that there is justification for these views when finite universes are being considered, but that they lead to inconsistencies of terminology. They seem to have engendered a tendency for contemporary writers to discriminate between models I and II in terms of population sizes. Furthermore the place of the Eisenhart models in the Tukey classification is determined by their postulates of normality, postulates which were no more than riders to justify tests of significance. These two circumstances have had the unfortunate effect of making the more basic distinction between models I and II, relative to which population sizes and forms of distributions are mere incidentals.

The Tukey classification is primarily a classification of two-factor universes and of ways of sampling therefrom. It cuts across the distinction of models I and

II which this paper will endeavour to maintain. In so far that the regression model can be treated as a degeneration of the general model (Secs. 8-11) the two formulations inevitably merge to some extent. But in the viewpoint to be developed here, universe postulates will be regarded as secondary conditions to be imposed when asking specific questions of a set of data, not as an essential part of the initial model.

4. Regression on quantitative factors.

We consider first application of the regression model to an experiment with quantitative factors because its interpretation may aid in visualizing interpretation for the more general case. It differs from the general case in that the levels of each factor can be ordered a priori with a measurable distance between each on the scale of 'intensity' of each factor. Analysis can be, and most often is, carried through on the general model (2.1), as if each treatment were discrete. With factors at only two levels that is the only practicable procedure since two levels give no handle on the response surface (unless it can be assumed a priori to be a plane, rare for quantitative factors). With three levels per factor one usually proceeds similarly for initial analysis, later partitioning treatment sum of squares for linear and curvature effects which are simple functions of the treatment means (cf. Yates, 1937). To bring out the points to be made here assume many levels per factor and a standard regression approach to evaluate the response surface. Let  $x$  be the levels of  $A$  measured in any convenient units, similarly  $z$  for  $B$ , and write the regression.

$$y_{ijk} = m + a_1 f_1(x_i) + a_2 f_2(x_i) + \dots + b_1 g_1(z_j) + \dots + c_1 h_1(x_i, z_j) + \dots + d_{ij} + e_{ijk} \quad (4.1)$$

where  $f_p(x)$  are arbitrary functions of  $x$  chosen so that the surface may approximate to the anticipated form of response with as few terms as possible, commonly (in ignorance)  $x, x^2, \dots$ ;  $p \leq (a-1)$ ; similarly for  $g_q$  and  $h_r$ ;  $d_{ij}$  represent deviation of the means  $y_{ij.}$  from the regression surface, and  $e_{ijk} = y_{ijk} - \hat{y}_{ijk}$ . If we use polynomials of  $f_1$  and  $g_1$  and the maximum number of terms (viz.  $p = 1 \dots (a-1)$ ,  $q = 1 \dots (b-1)$ ,  $r = 1 \dots (p-1)(q-1)$ ), the regression surface passes identically through all observed means  $y_{ij.}$ , and  $d_{ij}$  are zero.

For simplicity of exposition suppose we fit a quadratic

$$\hat{y} = m + a_1x + b_1z + a_2x^2 + cxz + b_2z^2 \quad (4.2)$$

The associated analysis of variance will be as in table 4.1.

Table 4.1. Analysis of variance for a quadratic regression fitted to a two factor experiment with a,b levels and d replications of each treatment combination.

	d.f.	S.Sq.	M. Sq.
Quadratic regression	5	$S_a$	$M_a$
Deviations of $y_{ij}$ from regression	(ab-6)	$S_t - S_a = S_r$	$M_r$
Total between treatments	(ab-1)	$d \sum^{ab} (y_{ij} - y_{...})^2 = S_t$	
"Internal error"	ab(d-1)	$\sum^{abr} (y_{ijk} - y_{ij.})^2$	$M_e$

$M_r$  may be greater than  $M_e$  either because the regression fails adequately to describe treatment effects, or because experimental conditions have permitted errors to creep into  $y_{ij}$ , which are not fully represented in variation between replicates and hence in the "internal estimate of error" evaluated therefrom (in other words, replicate observations on the same treatment are correlated). The conventional 5 per cent significance level for  $F = M_r/M_e$  is not adequate to distinguish these because it is not sensitive to systematic deviations (cf. Yule and Kendall, 1937, Sec. 22.20); adequate statistical test requires fitting a further regression term. Sometimes graphical examination may be more expeditious to indicate if this is worth doing and useful in showing just what kind of deviations are occurring. If deviations appear to be at random and a few more terms cannot reduce  $M_r$  to equivalence with  $M_e$ , then  $M_r$  (termed by Edwards Deming the "external estimate of error") may be the

preferred estimate of error from which to evaluate precision of regression coefficients and "predictions". Inflation of  $M_p$  may not be due to a real error component but to failure to find a proper functional form for the surface which should be fitted, or to having deliberately used a simple form of surface to facilitate deductions to an adequate degree of approximation. (Box and Hader's "lack of fit"). In such cases  $M_p$  will not behave properly in accordance with statistical distribution theory of random variables, and the uses we make of it as an estimate of error variance may not be theoretically justifiable. We may nevertheless feel obliged to ignore these complications to get a working compromise in place of impracticable complexity. Consequent inaccuracy of tests may not be appreciably worse than results from failure of real data to conform with postulates of normality. Interaction mean squares in the general case will often be in that role.

Suppose we have obtained a regression with satisfactory fit. What we have done is to evaluate a surface which states the mean yield (for the average environment of the given experiment) for any treatment combination  $xz$ , including values intermediate between those observed, within the range of observation. (Some extrapolation may be considered depending on one's faith that the fitted surface represents a functional form. Purely empirical polynomials should not be extrapolated. They can do 'crazy' things at very short distances from the observed region.) In addition to being able to interpolate between observed points the regression improves accuracy of estimation everywhere within the region because estimates even at observed points are reinforced by interpolated information from adjacent points. A few coefficients may be of individual interest for testing some hypothesis; for example, if we know that the functional relation follows a quadratic form, the quadratic coefficients tell the rate of change slope, valid everywhere. But if the surface is curved the linear coefficients tell only the

slope at an arbitrary location where we have put the independent variables equal to zero. By and large the coefficients individually tell little, they are only stepping stones to estimates of mean yield at any point. If, as commonly, we have fitted orthogonal polynomials, an estimate which omits some effective terms, for example using only linear terms (equivalent to main effects) when quadratic ones are significant, is a mean yield of some subset of treatments and represents a point off the regression surface, the centre of a chord. Relative to what it represents its standard error is correctly evaluable from  $M_e$ , but what it represents may be of little interest. On the other hand if curvature is small deviation of a chord from the surface proper may be small relative to random error and we may choose to use it for simplicity, but there is no strictly valid answer to what standard error should be attached to such estimates. As described above some form of  $M_x$  may be used as an expedient. This interpretation applies to the  $\alpha_i$  and  $\beta_j$  as these would be evaluated for the general equation (2.1).

Variance component analysis is (almost) never relevant to an experiment with quantitative factors. We cannot reasonably formulate a probability distribution of treatments over a surface. The variance of yield between arbitrarily selected points is meaningless. The only exception is for a factor like location on coils in the example given by Vaurio and Daniel; (1954), used by Wilk and Kempthorne (1955), and briefly described in Sec. 13 below; that is a "factor" which will not be at choice for future working but for which a given range must always be present in production. We can imagine location selectable at random from a uniform distribution over the length of a coil. Some systematic trend may be describable by regression, and further random variability representable by a variance component between randomly chosen locations. This is different from the ordinary quantitative

factor which is studied for the purpose of selecting that level at which operating conditions are best for most profitable yield, and for which variation between selected levels will be irrelevant. (We might be interested in variability which could be associated with a specified amount of variation of one factor as may occur under operating conditions of manufacture. But this is something different from the variance component between the arbitrarily selected levels observed in an experiment).

5. Regression model for qualitative factors.

In the general case for qualitative factors we term Eisenhart's I the "regression model" because the linear equation (2.1) can be written

$$y_{ijk} = \mu + \sum_{i=1}^a a_i x_i + \sum_{j=1}^b b_j z_j + \sum_{i=1}^a \sum_{j=1}^b (ab)_{ij} x_i z_j + \delta_{ijk} \quad (5.1)$$

where  $x, z$  are now "dummy" variables taking the values 0, 1 as required to reduce (5.1) to (2.1) (except that here we use roman letters for the regression coefficients in order to conform with notation for the general model and analogy to it as will be described in Sec. 10). The difference from (4.1) is that we are now fitting a "surface" to  $ab$  discrete points in  $(a+b)$  space, and interpolation between points has no meaning. Expectations of mean squares in the analysis of variance are commonly expressed as in Table 5.1, where  $\theta(a) = \sum a_i^2 / (a-1)$ , etc. The excess of mean squares over error is expressed in this form to indicate that they are functions of location parameters not subject to variance as in sampling random variables; but they can be read as proportional to the second moment of a restricted sub-universe of variants. As stated in Sec. 3 the analysis is an algorithm to indicate tests of significance for groups of regression coefficients; the  $\theta$  indicate how each term may be inflated if real effects are present.

Table 5.1

Factor	d.f.	E(M.Sq.)
$A$	$(a-1)$	$\sigma^2 + bd \theta(a)$
$B$	$(b-1)$	$\sigma^2 + ad \theta(b)$
$AB$	$(a-1)(b-1)$	$\sigma^2 + d \theta(ab)$
Error	$ab(d-1)$	$\sigma^2$

If the  $(ab)_{ij}$ 's can be assumed zero or negligible the 'curved' or 'crooked' surface becomes a 'plane' which can be evaluated with only  $(a+b-1)$  constants; in other words we can bring averages to bear to improve accuracy without thereby making projections off the plane. If there is a systematic pattern among the  $(ab)_{ij}$  it may be possible to reduce them to a fewer number of constants analogous to fitting a curved surface. If the  $(ab)_{ij}$  are not negligible, and whether systematic or erratic, averages represented by  $a_i$  and  $b_j$ , "main effects", are projections onto artificial points off the 'surface' in the same way as estimating points on a curved surface using only linear terms. They represent yields not realizable in practice, except perhaps by deliberate mixtures of treatments such as would not ordinarily be used. The "internal estimate of error" is technically valid for comparisons between such means interpreted strictly for what they are, averages of certain sets of treatments, but only circumstances can determine whether or not such comparisons are good ones to make.

Interaction elements play approximately the same role as coefficients for curvature in quantitative regression. Substantial ones indicate that treatment combinations should be individually considered. When they are relatively small the main effects may be read as general estimators for treatment effects over more or less wide ranges, analogous to using a linear regression for approximate description, or for interpolation, although the true relation may not be quite straight. The question of suitable error variance is then the same as for  $M_T$  in Sec. 4. The interaction mean square may be appropriate as an "external estimate of error", alias variance of deviations from the fitted linear 'regression'. But if the interactions do not occur at random and unpredictably with respect to future applications, the decision may be more a question of expediency than of

statistical theory. Preferred procedure may be suggested by proposed applications to future working rather than be dictated by the treatment combinations observed or the population which the experiment was originally designed to sample.

Main effects and interactions as originally defined for factorial experiments (Yates, 1935, 1937; Fisher, 1935) belong to the pure regression model. These definitions were explicitly formulated to summarize informatively and effectively contrasts between certain selected treatment combinations, equivalent to fitting a regression to describe gradients between a given set of points. As such they are correctly judged relative to the internal error which assesses precision of measurement of the mean yields of each treatment combination individually. If one wishes to go on to a wider class of hypothesis, to make inferences about related but unobserved treatments, the definitions and precision of their estimators relative to the wider field of inference, in effect extrapolation, must be reconsidered. We will return to these considerations under the heading of the generalized and mixed models. The regression model can be regarded as a limiting case of such more general hypotheses wherein the universe of study is defined to be just the set of treatments observed, but such definition will often be merely a conventional way of reverting attention to individual means.

Although the typical regression model has only one group of stochastic elements, sometimes, for example in a split-plot experiment, there may be random errors at two levels of magnitude. In such cases the model breaks into two regressions: that for main-plots has split-plot treatments balanced in every unit, that for split-plots has main-plot treatments as blocks, each with its respective single error group. Complications are possible but can usually be sorted out on general principles and need not be taken up in detail here.

6. The "random" model.

By the "random" model is usually implied a variance component analysis with the observed elements of each group assumed to be random samples from potentially infinite populations. With respect to these populations the individual elements of the observed groups are of little interest; the objective now is to assess the share of variance of the complex population,  $\sigma_y^2$ , which may be ascribable to each factor operating within it.

The linear equation (2.1) (except that in this case A, B, D do not exist) may be termed the latent equation to signify that values of the individual elements, usually described as random variables, are no longer of interest. The function of the latent equation is first to define the structure of the population and experiment, and assumptions about interactions; second to provide machinery to evaluate expectations of functions of the observations, particularly mean squares, as functions of distribution parameters of the random elements, and thence to define the variance components.

For a complex infinite population of the type postulated its total variance is the sum of variances of its composite elements. We therefore write the variance equation

$$\sigma_y^2 = \sigma_\alpha^2 + \sigma_\beta^2 + \sigma_{\alpha\beta}^2 + \sigma_\delta^2 \quad (6.1)$$

to indicate the parametric variance components which are to be evaluated.

Strictly speaking the letters of the linear equation are "random variables" only when regarded as functional forms. If the linear equation be read as representing a particular observation and set of elements, these are not now random variables, though it may be convenient to speak of them as such (Sec. 7.6). Furthermore the nature of a random variable is a function of the method of sampling.

The elements postulated in (2.1) are uncorrelated between groups only because of the balanced sampling design which has been specified. The elements of an unbalanced experiment will not in general be uncorrelated. Relative to unbalanced experiments the appropriate postulate about balance or unbalance in the hypothetical population, and thence the consequent correlation of elements in a sample and the appropriate definitions of the variance components, seem never to have been threshed out.

For balanced experiments and populations these postulates suffice for estimating the variance components by analysis of variance. To use maximum likelihood estimation we have to add more, we have to specify also the probability distribution of a given sample of  $y_j$  or equivalently, given the definition of the elements, to postulate their joint probability distribution. The complication over ordinary likelihood problems arises from the condition that observations are correlated within classes and that we have therefore to consider an  $n$ -variate probability function. The likelihood of a sample must be expressible as a function of the observations ( $y$ ), of the parameters (6.1), and of not more than an estimable number of nuisance parameters. The random elements may not appear.

For example consider the simplest case — a nested classification symbolized by the latent equation

$$\begin{aligned} y_{ij} &= \mu + \alpha_i + \delta_{ij} \\ i &= 1 \dots a \\ j &= 1 \dots d \end{aligned} \tag{6.2}$$

with the assumptions that  $\alpha$  is NID  $(0, \sigma_\alpha^2)$ ,  $\delta$  is NID  $(0, \sigma_\delta^2)$  for every  $i$ . The probability of the sample can be formulated as a probability element of an ad-variate normal distribution, with common variance  $(\sigma_\delta^2 + \sigma_\alpha^2)$ , and with covariances

$\sigma_{\alpha}^2$  between all pairs within the same class. Some heavy algebra then leads to the usual analysis of variance solution. Easier formulation is given by noting that the mean square within classes,  $s_{\delta}^2$ , must be a sufficient estimator for  $\sigma_{\delta}^2$ . (Recollect that we are limiting ourselves to balanced designs, i.e.,  $d$  is constant in all classes. If  $d_i$  were variable the mean square within classes does not contain all the information on  $\sigma_{\delta}^2$  and complications ensue.) We can then reduce all available information to that given by  $s_{\delta}^2$  and the  $y_{i.}$  which are NID ( $\mu$ ,  $(\sigma_{\alpha}^2 + \sigma_{\delta}^2/d)$ ). The log likelihood for these can then be written as

$$\ln L = \text{Const} + \frac{1}{2} a(d-1) \ln (s_{\delta}^2/\sigma_{\delta}^2) - \frac{a(d-1) s_{\delta}^2}{2\sigma_{\delta}^2} - \frac{a}{2} \ln (\sigma_{\alpha}^2 + \frac{\sigma_{\delta}^2}{d}) -$$

$$- \sum_i \frac{(y_{i.} - \mu)^2}{2(\sigma_{\alpha}^2 + \sigma_{\delta}^2/d)} \quad (6.3)$$

(cf. Anderson and Bancroft, pp 319-320, who outline a more complex case; and Cochran, 1937, who gives corresponding formulation when  $d$  and  $\sigma_{\delta}^2$  are both variable between classes.)

The direct likelihood approach has forced us back to considering correlation of grouped observations, a complexity which the linear model and analysis of variance approach was formulated to avoid. It does this in taking account of postulated or given forms of distribution within and between classes, extra information which the variance analysis does not need and does not use. Variance analysis estimation is a distribution-free technique.

Without stopping here to define a random variable (Sec. 7.5) notice that a simple random variable as usually considered is a function whose observed value on any event cannot be predicted in advance and we cannot in making another observation

demand that an element similar to one previously observed should appear again. Yet models (2.1) or (6.2) imply that all elements except  $\delta$  can be re-sampled at will. Imagine the  $\alpha_i$  of (6.2) as values corresponding to urns which can be specified by location, and the  $\delta_{ij}$  as balls in the  $i$ th urn. With respect to sampling the urns  $\alpha$  is a random variable which takes the value  $\alpha_i$  on the  $i$ th drawing. But after selecting the  $i$ th urn we return to it again and again, i.e. to the same  $\alpha_i$ , to sample various  $y_{ij}$ . With respect to the distribution of repeated observations from the same urn  $\alpha$  is a parameter. In the simplest case being considered here the duality is not troublesome, but in more complex cases it can become confusing when elements may seem to flip back and forth between parameters and random variables like the faces of a cube drawn on flat paper for which psychologists sometimes ask us to record how frequently they snap from lower to upper views and back again.

If the dual role of the  $\alpha$ 's be not watched one might be tempted to formulate the above likelihood problem by saying that the probability of the sample is given by the probability of selecting a  $\alpha$ 's multiplied by the conditional probability of then selecting  $n$   $\epsilon$ 's leading to

$$\ln L = \text{const} - a \ln \sigma_\alpha - \sum \alpha_i^2 / (2\sigma_\alpha^2) - n \ln \sigma_\delta - \sum \sum (y_{ij} - \mu - \alpha_i)^2 / (2\sigma_\delta^2) \quad (6.4)$$

Furthermore maximizing this with respect to variation of the  $\alpha_i$ , as well as of  $\mu$ ,  $\sigma_\alpha^2$  and  $\sigma_\delta^2$ , yields a solution (if between class mean square be not too small) which appears not nonsensical. At least it is consistent in that it predicts a set of  $\alpha_i$  whose second moment is  $\sigma_\alpha^2$ , whereas if we estimate  $\alpha_i$  as independent parameters the variance of these estimates will not be the  $\sigma_\alpha^2$  which we seek. The solution might however be described as 'schizophrenic'. In the third term of (6.4)

the  $\alpha_i$  are variates which the procedure is trying to draw together as estimators of their central parameter, zero; whereas in the last term they are parameters being spread out to minimize variances within urns. The procedure is irrational in attempting to treat  $\alpha_i$  simultaneously in two roles, and the fault is here easily exposed. But is not a similar misdemeanour in effect committed, albeit more indirectly, when one purports to estimate both a variance component for a larger population and the means of observed variants under guise of using the same linear model (Sec. 11)?

In my opinion Neyman and Scott (1948) also commit this misdemeanour, although it can be argued that they have evaded it by inserting only the  $\alpha_i$  without the parameters of their distribution in (6.4). When some parameters of the probability distribution of a sample occur only with a number of observations which remains finite as the sample is increased they term them "incidental parameters"; parameters occurring in the distribution of all observations are termed "structural". Their two main examples are: (1) The observations are normally distributed about a common mean  $\alpha$ , the structural parameter, but fall into groups of  $n_i$  observations with each group having its own variance  $\sigma_i^2$ , incidental parameters. Example (2) is similar but with common  $\sigma^2$  and variable means,  $\alpha_i$  (this being the case discussed above except that no distribution is postulated for the  $\alpha_i$ ). Kendall's (1952) maximum likelihood derivation of Kummell's solution for a linear functional relation,  $Y = \alpha_0 + \alpha_1 X$ , is similar. Here the structural parameters are  $\alpha_0$  and  $\alpha_1$ ; the groups have each two observations, the paired observations  $x_i, y_i$ , with the "true" values,  $X_i$ , as incidental parameters.

Neyman and Scott claim to have established two propositions:

- "(1) Maximum-likelihood estimates of the structural parameters relating to a partially consistent series of observations need not be consistent."

"(2) Even if the maximum-likelihood estimate of a structural parameter is consistent, if the series of observations is only partially consistent, the maximum-likelihood estimate need not possess the property of asymptotic efficiency."

They demonstrate proposition (1) with example (2), and vice versa. In both cases the failure depends on  $n_1$  remaining finite. The properties of maximum-likelihood are asymptotic ones, and they proceed to the limit by taking an indefinitely increasing number of sub-samples of  $n_1$  observations. My feeling is that this is an improper use of the asymptotic process. If the incidental parameters are parameters in any true sense it is, at least theoretically, possible to return again and again to resample the sub-populations of which they are parameters (cf. Secs. 7.18 and 13 below), hence the only valid asymptotic procedure is to allow all  $n_1$  to tend to infinity. In that case the usual maximum likelihood properties obtain and the propositions fail. If the set-up is imagined to be such that  $n_1$  cannot be increased, so that the only available approach to infinity is by increasing the number of groups, then the 'incidental parameters' must be random variables with their own distribution, and should appear as such in the likelihood function as in (6.3) as opposed to (6.4). Christening them incidental parameters is a dialectical evasion of ignorance about what probability distribution should be postulated for them and hence what parameters should be inserted in their place. The only other way round is to treat the incidental parameters as if from a finite universe whose parameters are the individual values which may occur (Sec. 7.11). In that case Sec. 7.12 shows that maximum likelihood cannot be used, and again the propositions fail because maximum likelihood estimates do not exist.

Some texts and lecture notes appear to imply that a variance component analysis

can be deduced from a regression model. Recognition of their incidental, almost accidental, association may therefore be advisable. Least squares estimation can be directed only to model I, i.e. to estimate parameters of location. Analysis of variance, without reference to regression, indicates an estimation procedure for variance components. It is not necessarily efficient, (except for normal distributions and balanced samples, for which it gives the maximum likelihood solution) but it is manageable where maximum likelihood solutions are impracticable or impossible. The best partition of the sum of squares of an unbalanced sample is however not obvious, and it is customary to use a quasi-regression approach merely to indicate a workable partition. It is not ideal because it does not lead to unique solutions independently of order of "elimination" of the (dummy) independent variables. But at least it is a pointer among innumerable possibilities. In principle the procedure is not necessarily restricted to partitions of sums of squares. We could consider expectations of all sorts of functions of observations; but quadratic forms are known to be most efficient for normal distributions, and among the many such which might be concocted those indicated by a quasi-regression analysis of variance seem reasonable.

7. Some basic theory relating to samples from finite populations.

1. Most of us approach variance component analysis from the side of elementary theory applicable to infinite populations as in Sec. 6. When we start postulating finite populations for some of the classifications consequent effects on theory are not at once obvious. Invariably at least one group of elements are random variables (elements from a hypothetical infinite population) and therefore so also are the observable  $y$ 's. Thence the elements of any group, although finite in number, can still be defined only asymptotically and more or less arbitrarily. In these circumstances one may overlook things which would be obvious in working with a simple finite population whose elements can be directly observed; and other subtleties which can usually be slurred over without noticeable consequence now become more critical for clear understanding. This section endeavours to bring these explicitly to attention and remove lurking ambiguities.

2. To avoid wearisome reiteration of 'infinite' and 'finite', population will be used to denote a hypothetical infinite population, universe will denote a potentially existent or postulated finite population.

3. Element will be used to imply the quantitative value of an individual, a fixed quantity.

4. The condition that a universe has just a definite number, say  $N$ , of elements distinguishes its study from other statistical methods more sharply than is perhaps generally realized. Firstly the number  $N$  has to be assumed known. (We exclude here investigations to discover  $N$ , for example number of fish in a lake, which introduce other matters with which we are not concerned.) Secondly all observations and statistics are going to be limited to combinations and functions of a selection from these  $N$  fixed quantities. Since  $N$  is finite the frequency distribution of the universe is necessarily discrete. The space denoting the values which the

elements may have (before we observe them and thus know what they are) may however be continuous, or if discrete is likely to have very many more points than are actually represented. Complete specification of the frequency distribution of a universe means specifying both the number and quantitative value of each type of element present. Since the number of types may be  $N$ , or nearly  $N$ , it will often be convenient to suppose each type to be present only once, and to imagine the universe distribution defined by listing every element individually, say  $X_1 \dots X_N$  (obviously with repetitions of the same numerical value as may be necessary). This distribution is of course not a probability distribution, though many writings appear to treat it as if it were so -- we get so accustomed to thinking of frequency distributions and probability distributions as synonymous.

5. A probability distribution is associated only with a random variable. In fact, conversely, a random variable (or variate) is usually defined (e.g. Kendall, 1943) as "a variable with which is associated a probability distribution". Feller (1950) emphasizes that a random variable has the characteristics of a function and is defined only in a given sample space. Cramer (1946) describes its observed values as determined by a "random experiment which may be repeated a large number of times under uniform conditions".

6. The act of sampling from a universe (the "random experiment") generates a random variable, say  $x$  (without subscript). An 'observation',  $x_1$ , is the value assumed by a random variable in a particular experiment (sampling). Equivalently it is the element sampled. An observation once taken is again a fixed quantity, sometimes regarded as a constant but (to me at least) to so regard it may be more confusing than helpful. A mathematician speaking of the function  $f(x) = x^2$  writes  $f(2) = 4$ , and tabulates the "function" for given arguments. Therefore following

Feller's description of a random variable as a function whose "independent variable (is) a point in sample space i.e. outcome of an experiment", to describe an observation as the random variable of an experiment seems permissible. An estimator is a random variable, an estimate is a value observed in a given experiment; either is called a statistic and this bracketing seems preferable to bracketing both an estimate and a parameter as constants. Furthermore the suggested usage is in effect adopted when a standard error is attached to an observation or estimate, and spoken of as its standard error.

7. The probability distribution of a random variable generated from a universe states the asymptotic relative frequency with which observed values may appear in infinitely many samplings of the same kind (with replacement -- the repetitions of a random experiment under uniform conditions). The universe dictates those points of the sample space whose probability is non-zero; the probability function is a function of the method of sampling. The probability distribution must specify both. If single elements are "sampled with equal probability" the probability distribution is

$$p(x) = \frac{1}{N}, \quad x = X_1 \dots X_N \quad (7.1)$$

It is easy to specify innumerable other methods of sampling to each of which would correspond a different function  $p(x)$ , for example sampling with probability proportional to size as is common in sample surveys.

8. The probability function  $p(x)$  has meaning and interpretation only with respect to the relative frequencies of a hypothetical infinite population of observations resulting from samplings with replacement - i.e. repetitions of a "random experiment" under uniform conditions (cf. Kendall, 1949). We thus commonly think of the observations, and also of the random variable itself, as being elements

of that population. In thinking of an ordinary mathematical function as a curve or surface in space we identify it with the infinity of points on that surface; so why not identify the function known as "random variable" with an infinite population of elements representing the values it can assume? The hypothetical population is a model which portrays its probability distribution as a function of a function,  $p(x(\text{expt.}))$ .

9. Statistical theory, while concerned also with other matters, is to a large degree synonymous with probability theory. In particular the theory of statistical estimation is essentially the theory of how to estimate the parameters of a probability distribution, that is of a hypothetical infinite population. Thus a statistical estimate always invokes the concept of an infinite population and can do not otherwise.

Tukey (1950) essentially reaches the same concept, but he does so empirically as "a price to pay for simplicity" and with a later exclamation mark. I come to it independently as a fundamental concept found necessary in trying to write down basic ideas which will remain logically consistent with each other under all circumstances.

10. The bridge to inference about an existent or potentially existent universe involves (1) assuming that sampling procedure has done its duty and (2) invoking the a priori theory of probability to formulate a known relation between characteristic:

of the universe and parameters of the probability distribution of a random variable generated by the postulated sampling procedure (cf. Kendall, 1949). \*

11. I do not know any satisfactory definition of 'parameters', text books seem to evade stating one. Most often we think of them as the constants  $\theta$  in expressing a probability function as  $f(x; \theta_1 \dots \theta_p)$ . However the population moments are commonly referred to as parameters and the usage seems justified since they are functions of  $\theta$ , and conversely the  $\theta$  can be expressed as functions of any  $p$  moments. Similarly for any other suitable set of population characteristics. Less common usage, but one which seems to be necessary, is that the basic parameters of a distribution

---

\* Kendall, in this interesting and stimulating paper "On the reconciliation of theories of probability", demonstrates that the frequentist and non-frequentist theories of probability must each invoke the other at some point enroute from theory to application. At the same time (among other statements which might be disputed) when criticising maximum likelihood and confidence intervals for ignoring a priori probability he seems to forget the "bridge". His example is to suppose observation of 1000 births of which 600 were male. He asks: "Are we then to conclude that the sex ratio lies between .59 and .61 when from tremendous previous experience we know it to be close to .51?" But the question skips a chasm and is irrelevant. We do not do an experiment to ascertain something already known. It would be futile to observe 1000 births to modify an estimate already based on millions. We do an experiment to get information about something unknown. If there are other samples which we know to be properly drawn from the same population (and this "knowledge" is not contradicted by evidence of heterogeneity), they are really all one experiment and estimation properly proceeds from all pooled together. Estimation from a single sample may have several objectives. We may know the population but have used a new sampling procedure, or novitiate observers, and require to check the "bridge". We may think it might be from a certain population or a similar one, and assuming the bridge sound, wish to check if the supposition is reasonable. We may know it is from a population sui generis, perhaps an inbred one affected by sex-linked genes whose frequency is to be estimated. To begin by seeking an estimate already biased by a prior guess, about what population the sample might have come from, would be poor procedure. The need is to get an independent estimate about the population which this sample in fact represents, free of prejudicial suppositions. Only then can we objectively consider relations to hypothetical possibilities. That, maximum likelihood and confidence intervals very properly do; how can objective reasoning proceed otherwise? Use of a priori knowledge or supposition requires that one should be more than ordinarily sure that all bridges lead to the same island.

derived from a universe are the  $N$  values  $X_1 \dots X_N$ ; all of them are necessary to completely specify any such distribution (with some possible compression if several  $X_i$  have the same numerical value).

12. The probability function of the distribution of samples from a universe is (usually) a function of  $N$  only; the parameters  $X_i$  appear in the place of what would usually be a specification of known sample space. (Some sampling schemes, for example sampling with probability proportional to size, may introduce the  $X_i$  into  $p(x)$ ; but it is of little help since they still remain in specification of those parts of sample space for which  $p(x) \neq 0$ .) Two consequences follow: (1) We can never estimate all the parameters from a sample of less than the universe itself. "To achieve reasonable simplicity it is necessary to describe the probability distributions rather summarily by a few 'typical values' " (quoted from Feller, 1950, p. 171, with alteration of two words, the context is different). For the groups of elements with which analysis of variance is concerned the mean (location) is given by definition as zero; the only typical value with which we concern ourselves is the "variance" or its analogue. (2) Since neither the specific parameters nor that single function of them ('variance') to be estimated, enters into the probability function,  $p(x)$ , the method of maximum likelihood cannot be used to indicate a preferred estimator, nor to provide an absolute measure of efficiency for any other estimator. The estimation problem is not, as usual, to estimate the probability function; but to estimate points of sample space to which the probability function applies, or a function of them. Herein lies a major difference distinguishing problems of estimating characteristics of a universe from those relating to parameters of probability distributions more commonly discussed.

I have never before seen this explicitly stated. It is indeed obvious. But it seems worth stating to clarify consequences of imposing finite universe postulates on our usual approach to analysis of variance. For example: on first reading Tukey's (1950) classification of models for data in two-way tables one can be irritated by his vague use of "effects", ignoring whether these are to be regarded as "fixed effects" to be individually evaluated, or as "random variables" whose dispersion only is of interest. On formulating these consequences of postulating a finite universe one realizes excuse for what at first seemed careless terminology.

It might seem simpler to say that, whereas ordinary statistical theory states the sample space a priori and the probability function contains the parameters, universe theory postulates the probability function as given and has a sample space defined by the parameters (either those of the sample distribution itself or of the parent universe). Such formulation however would imply more fundamental divergence between "ordinary" and "universe" theory than does the above formulation, and would not, I guess, be acceptable to theorists. (But see appendix to this section.)

13. Unless stated otherwise a (single) sample is assumed to be drawn as a whole i.e. without replacement. The observations in such a sample are of course not independent. Consequently for most purposes it is convenient to regard the whole sample as one observation of a single random variable. For generality it may be described as a vector; but often the relevant random variable may be merely a single statistic of interest, e.g. the sample mean.

14. 'Variance', a measure of a random variable's state of being variant', is defined by

$$\text{var}(x) = \sigma^2 = \mu_2 = E(x-E(x))^2 \quad (7.2)$$

The definition involves the probability distribution of  $x$  and thus, strictly interpreted, applies only to random variables. Since, as the second moment, it is a parameter of the probability distribution it may be said to be the 'variance of the (hypothetical) population'.

15. A universe, being an assembly of fixed quantities, does not have a variance. Symbolized by mass points along a line it does have a second moment, which, by analogy with measures of dispersion found effective in describing probability distributions, is naturally adopted as a measure of the dispersion of its elements.

Inevitably this second moment gets called a 'variance', although the usage seems inadvisable not only etymologically and semantically but because it helps to blur distinctions made above.\* It may be excused on the grounds that it states the variance of a random variable generated by sampling single elements with equal probability, and this may be regarded as the most primitive sampling distribution derivable from the universe.

16. The 'variance of a sample' is justifiable on the same lines, only more so if we agree to regard an observation as a 'random variable of an experiment'. It also has the excuse that estimates of a parameter commonly go by the same name and are differentiated only symbolically. Variance of a sample is seldom wanted for itself alone, i.e., merely as description of n given quantities, but has interest

---

\* Edwards Deming (1950, p.56) notes that, on the analogy to physics from which many of our terms derive, the "moment" should be defined as  $\sum (X-\bar{X})^2$ , and  $\sum (X-\bar{X})^2/N$  should be termed the "moment coefficient". However accepted usage has gone too far to retract, and a total and a mean being effectively the same statistic a student may usefully be encouraged so to think of them (provided he does not mix them in computing!). Clearly the moment is a characteristic of the universe, not of its individual (fixed) elements; therefore, for logic as well as for brevity, and despite leading texts, the phrase "moment (variance) of a universe" seems preferable to "variance of X in the universe".

Only the struggle for consistency throughout this paper has sensitized me to such shades of phrasing. At other times I might have been numbered among those who say: "Admittedly 'variance' is not etymologically applicable to a universe; but the idea of using its moment to measure spread of its elements is the same, so why quibble on a name?". However as a general rule careful use of analogies should be guarded. For example: the probability of a discrete distribution should not be termed a "density". To do so destroys the whole utility of our analogy of space and probability measures to volume and mass. Yet this erroneous use of a physical term is becoming regrettably common in texts and handbooks. The distinction is not trivial. My first meeting with that error in an excellent and authoritative text, before my own understanding had crystallized, cost many hours trying to rationalize it before explaining the analogy and its uses to students reading that text. In a similar way many students have lost time worrying over "variance of a universe", though this one is less critical and does not do comparable damage.

mainly as an estimate of a population parameter. Restricting oneself for the moment to samples of  $n$  independent observations, the true (population) mean being unknown, the logical definition of sample variance seems to be

$$s^2 = k_2 = \sum_{i=1}^n (x_i - \bar{x})^2 / (n-1) \quad (7.13)$$

A surprising number of writers still insist on  $m_2 = \sum (x - \bar{x})^2 / n$ . The same writers would object that, ever since Gauss, it has been wrong to describe variance about a regression on  $p$  independent variables as other than  $\sum (x - \hat{x})^2 / (n-p)$ . Since  $\bar{x}$  is essentially a regression coefficient on a constant independent variable they are thence being inconsistent. If required  $m_2$  is properly describable as the second moment of the sample.

17. When, as happens within samples from a universe, observations are not independent, proper usage is more problematic. Tukey (1950) has given the most emphatic statement of the inclination of many authorities by choosing for convenience "to define the variance of any finite set of numbers as  $k_2$ , whether the finite set is a sample or a population". (It was my own choice and practice until endeavouring to find consistent terminology for this paper.) The convenience of course arises from the fact that  $k_2$  is "inherited on the average" (Tukey, 1950) a stronger quality than being unbiased, meaning that it is an unbiased estimator of  $K_2 = \sum (X - \bar{X})^2 / (N-1)$  independently of  $N$ . (I follow Wishart, 1952, in using capital letters to designate universe homologues of sample statistics. Cochran, 1953, and Hansen et al, 1953, use  $K_2$  under designation  $S^2$  without other name. Wishart has given the name "generalized k-statistics" to Tukey's k-statistics for a universe.) These properties give  $K_2$ , whatever its name, a central place in universe theory; but it is not the variance by definition (7.2) of any random variable derivable from the universe by ordinary sampling. The variance of single observations on

repeated sampling is, by (7.2),  $M_2 = \sum (X - \bar{X})^2 / N$ , and cannot be arbitrarily redefined without becoming inconsistent with the variance of means of samples of  $n$  observations which is  $K_2 \left( \frac{1}{n} - \frac{1}{N} \right) = M_2$  when  $n = 1$ .

Furthermore, if we form a sum of two elements drawn singly from each of two universes then the variance of such sums is

$$\text{var } (x+y) = E ((x+y) - E (x+y))^2 = M_2 (x) + M_2 (y) = M_2 (x+y)$$

where the last term is the second moment of the universe of  $N_x N_y$  possible sums.

Contrariwise  $K_2 (x+y)$  is the average for all possible samples of the  $k_2$  determined within samples of  $n$  sums drawn without replacement;

$$K_2(x+y) = K_2(x) + K_2(y) = \text{Ave} \sum_{n}^{n} (x+y-\bar{x}-\bar{y})^2 / (n-1) \quad n \leq \min (N_x, N_y) \quad (7.4)$$

It is not a  $K_2$  for any universe of sums and, while extremely useful, is not to be interpreted in the same way as a sum of variances.

Other inconsistencies will appear below. Therefore, perhaps rather ruefully since one might like so to describe it to experimenters, one seems forced to conclude that  $K_2$  cannot properly be called a variance. In some contexts it may for expediency be regarded as a variance of a hypothetical infinite population, but in general distinction may be advisable.

18. Observations in samples from a universe have a peculiarly ambivalency which is related to ambiguity often seen in interpretations of analyses of variance. Each observation,  $x_i$ , is immediately an evaluation, without error, of one of the parameters,  $X_j$ , of the universe. Nevertheless, in absence of the whole  $N$ , the sub-class of  $n$  are still random variables in the sense that their values could not be predicted, and the summary statistic  $k_2 = \sum_{n}^{n} (x-\bar{x})^2 / (n-1)$  is a random variable and an unbiased estimator of  $K_2$  of the universe. Except in that its distribution depends on the ratio  $n:N$ , in such a way that its dispersion tends to zero as  $n \rightarrow N$ ,

it is not essentially different from the variance of a sample from an infinite population, despite the fact that the individual elements composing it are, in a sense, parameters themselves.

19. The concept that both the observed sample and the universe may be regarded as samples from a hypothetical infinite population with variance  $K_2$  and unspecified mean was used by Cochran (1939) in first introducing that parameter, and has been advocated by Hendricks (1947, 1951). Their objective was to evaluate the variance of means of observed samples,  $\bar{x}_n$ , and it is easy to show that if we imagine  $\bar{X}$  to be distributed about the population  $\mu$  with variance  $K_2/N$ , similarly  $\bar{x}_n$  with variance  $K_2/n$ , and if the sample of  $n$  is a sub-sample of the super-sample of  $N$ , then it follows that the mean square deviation of  $\bar{x}_n$  about  $\bar{X} = E(\bar{x}_n - \bar{X})^2 = K_2(\frac{1}{n} - \frac{1}{N})$ . In recent years sample survey workers seem to have dropped this approach, presumably for reasons similar to those outlined above. However it may sometimes (Sec. 9) appear convenient to think of what we shall designate as  $K$ (parameters) or  $k$ (statistics) with greek subscripts as variances of hypothetical infinite populations.

20. Like "variance", the terms "correlation" or "covariance" and "independence" are properly applicable only to random variables. In a universe with multiple classifications the analogue of covariance is the product moment; the analogue of independence is a certain symmetry of associated elements. Consider a fixed universe of  $N = ABD$  elements corresponding to the standard model

$$y_{ijk} = m + a_i + b_j + (ab)_{ij} + d_{ijk}$$

$$i = 1 \dots A, j = 1 \dots B, k = 1 \dots D$$

We may picture the universe space as  $N$  points representing all possible values of  $y$  plotted against co-ordinate axes, one for each group of elements. In most cases

$d_{ijk}$  will have a distribution which is the same for all  $ij$  (often it will be specified only as a random variable with a probability distribution, pictured as an infinite population) and may be said to be "independent" of the other elements. If not, considerations regarding it will be similar to those for  $(ab)_{ij}$ . Imagine three co-ordinate axis for  $(a)$ ,  $(b)$  and  $(ab)$ , functional forms which may take values  $(a)_i$ ,  $(b)_j$ ,  $(ab)_{ij}$ . The sub-scripts with brackets refer to values as opposed to elements; we want to picture many elements of each group having the same value. In the space thus formed plot a point for every  $y' = a_i + b_j + ab_{ij}$  which may occur in the universe of such  $y'$ . Two groups will be said to be "independent in the universe" if the conditional distributions of such points for values of one group, given any single value of the other, are the same.

21. By definition of a balanced universe every  $a_i$  occurs equally often with every  $b_j$ ; they are therefore independent in the universe. Furthermore by the method of sampling to form a balanced experiment they are also independent in experiments -- i.e. independent in the universe sense, regarding observations as fixed quantities which is the reason why  $\sigma_a^2$  does not appear in the mean square for  $\beta$  and conversely. They are not only uncorrelated in the general sense, their sampling correlations are identically zero.

22. Since a given treatment combination determines a particular  $ab_{ij}$  with a prescribed  $a_i$  and  $b_j$ ,  $(ab)$  will not usually be independent of  $(a)$  and  $(b)$ . But it may be if there are many elements  $a_i$  and  $b_j$  having the same value, and for every subset of these with given  $(a)_i(b)_j$  the associated  $(ab)_{ij}$  have the same distribution. The condition is unlikely to occur in a real finite universe, but this is what we in effect assume for "large" universes or populations in the random model. The question is often raised as to how  $(\alpha\beta)$  can be independent of  $\alpha$  and  $\beta$  even for

an infinite universe. We commonly assume each group to be normally distributed, and since by definition they are uncorrelated, independence must also be implied. The postulate of normality necessarily invokes the concept of an infinite population, and the point that seems often to be missed is that that concept implies, not only an infinity of elements for each group as a whole, but also infinitely many elements at every pair of  $\alpha$  and  $\beta$  values corresponding to the conditional distribution of  $(\alpha\beta)$ . What we are really doing is to postulate that the relative probabilities with which given values can appear in samples can be approximated by sampling distributions for which these populations of elements are only conceptual models to aid thought.

23. In a universe of a finite number of elements,  $(ab)_{ij}$  will usually not be independent of  $a_i$  and  $b_j$ . But if the distribution of  $(ab)_{ij}$  is effectively at random, that is shows no systematic pattern relative to the magnitudes of associated  $(a)_i$  and  $(b)_j$ , it may be said to be 'practically independent'. (We can picture this by saying that if we section the  $(a)(b)$  plane in squares and consider histograms for the associated frequencies of  $(ab)_{ij}$ , then all of these conditional histograms should be approximately similar and with equal, zero, means.) In any case, by definition, the universe product moments  $\sum_{AB} a_i (ab)_{ij}$  and  $\sum_{AB} b_j (ab)_{ij}$  are zero; and the random variables formed by simple sampling are uncorrelated. The conditional covariance of say  $a_i$  with  $(ab)_{ij}$  for a given sub-set of  $b_j$  is however not necessarily zero. Attention to this point may be relevant for some problems which are being considered by one of my colleagues. In this paper we will always assume averaging to be over the whole universe.

Appendix to Section 7

Definition of Parameters

Subsequent to writing Section 7 it was brought to my attention that Kendall and Sundrum (1953) have given reasons for preferring to restrict the definition of parameters to the constants commonly designated  $\theta_i$  in the formulation of a probability function of a variate as  $f(x; \theta_1, \dots, \theta_p)$ . Their primary problem was to define "non-parametric hypotheses" and "distribution-free inference"; consequently their definition is conditioned by that context with arguments like the following.

"7. Consider the hypotheses

(a) the population is normal

(b) the population has finite cumulants but  $\mu_r = 0$  for  $r \geq 2$ .

The first is not a parametric hypothesis under any ordinary interpretation. For consistency, therefore, the second cannot be parametric for it is completely equivalent to the first. This reinforces our conclusions that the parameters must be finite in number to constitute a parametric hypothesis."

Their paragraphs 8 and 9 then find "that the question whether a given hypothesis is parametric or not has to be decided in the light of alternatives against which it is to be tested." Whence

"10. We therefore define a statistical hypothesis as parametric if

(a) it makes an assertion about a distribution;

(b) it specifies the distribution completely except for the values of a finite number of parameters;

(c) it is considered against alternatives of the same distributional form which, therefore, differ only in the values of the parameters involved.

If a hypothesis is not parametric it may be said to be non-parametric.

- "11. It is important to confine the adjectives "parametric" and "non-parametric to statistical hypotheses. They should not be applied to statistics, tests or types of inference. This may sound rather austere, but we have found a great deal of confusion arising from the use of phrases like 'non-parametric tests' and 'non-parametric inference'.
- "12. It remains open whether, in view of the subtlety and relativity of the expressions, it is worth while continuing to make a distinction between parametric and non-parametric hypotheses. On the whole we think that it is worth preserving but we should not quarrel with anyone who took the opposite view or who felt that some other classification was more desirable."

These arguments thus relate mainly to defining the adjectival usage, "parametric hypothesis", but recognising a logical need to embrace the noun, earlier paragraphs state:

- "2. It is necessary in the first place to define what is meant by a 'parameter in relation to a statistical distribution. Traditionally, the term, which was borrowed from pure mathematics, denotes a variable, appearing in the equation defining the distribution, which can be regarded as having one of a set of values. Thus, the quantities  $m$  and  $\sigma$  in

$$dF = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{1}{2} (x-m)^2 / \sigma^2} dx, \quad -\infty \leq x \leq \infty \quad (1)$$

are parameters; and so is the quantity  $\theta$  in

$$dF = dx, \quad \theta \leq x \leq \theta + 1. \quad (2)$$

"3. But statistical practice has not stopped here. For example, it is sometimes said that the mean is a parameter of the normal distribution (1). This is inaccurate; what should be said is that the parameter  $m$  is equal to the mean. The point of the distinction is that if we allow ourselves to refer to measures of location, dispersion, etc., as parameters, we get into difficulties. In fact, all the moments of the normal distribution would then be parameters (though, of course, functionally dependent); and every distribution would have an infinite number of parameters. We should have to admit the median, the quantiles, the individual frequencies of a discrete distribution and perhaps even the distribution function itself as parameters. This is not at all what is intended, and we conclude that the practice of referring to the summarizing measures of a distribution as parameters is to be condemned."

Their arguments are weighty but they create difficulty for other contexts. Fisher made a fundamental clearing of a statistical tangle by emphasizing the importance of distinguishing between a sample estimate and the population value which it estimates: in that context the contrast statistic versus parameter is firmly entrenched in our terminology. Sample moments, medians, etc., will continue to be statistics (or estimators) whether or not we can make an assertion about the form of the parent distribution. But if Kendall and Sundrum's restriction on parameters is to be maintained what shall we call their antitheses?

Neither least squares estimation of location 'parameters' nor analysis of variance estimation of variance components depends on any postulate about the form of distribution of the observations. Therefore on the restricted definition neither the elements of our linear model nor the components of the variance equation

can be called parameters. 'Things to be estimated' is intolerable. 'Elements' or 'typical values' do not convey the required distinction of population values. 'Population' is an unsatisfactory adjective. 'True' regression or variance equation implies an attempt at colloquial explanation unsatisfactory for a technical term. To think of the elements of these equations as other than parameters is now difficult

In a brief chance conversation before I knew of his paper - much too brief for thorough consideration - Professor Kendall expressed the opinion that, following his definition of "parameters", the probability distribution of samples from a finite universe has no parameters, and that the elements of the universe define the sample space. Contrariwise Sec. 7.12 had taken the view that sample space is always definable a priori to cover all possible outcomes of an experiment, that the probability distribution defines the relative frequency with which each point in sample space may be observed (including those which have "probability measure zero" for a given population), and that all population characteristics required completely to define the probability distribution must be parameters. Suppose an experiment is a sample survey to determine the distribution of some 'event' in a district. Several districts and the same district in different months are a class of similar experiments. Would it not be contrary to established usage to postulate a different sample space for each of these "experiments"? Is not specification of the potential outcome more closely related to the  $\theta$  of the second example of Kendall and Sundrum's par. 2 than to a specification of sample space?

For these reasons it seems more consistent with custom to define a parameter as any characteristic of a population, including both general characteristics such as mean and variance and all those required to specify a probability distribution. To do otherwise would seem to require unpleasant periphrases in the contexts where

the word has been used in this paper; and these usages may be considered to have historical precedence over endeavours to distinguish between parametric and non-parametric hypothesis. If it is now impracticable to have different words for the general and restricted classes presumably we will, as so often, have to get along with two usages depending on context for differentiation. It may not be too difficult once the alternative definitions have been explicitly formulated.

8. Generalized variance model

The generalized model postulates that the complex population of  $y$  is built up with only a finite number of elements in one or more (possibly all) groups of the linear model. Let the specific latent equation be

$$y_{ijk} = m + a_i + b_j + (ab)_{ij} + d_{ijk} \quad (8.1)$$

$$i = 1 \dots a \dots A \leq \infty$$

$$j = 1 \dots b \dots B \leq \infty$$

$$k = 1 \dots d \dots D \leq \infty$$

$$ABD = N$$

Some readers may object to the duplicate roles for the same letter, e.g.  $a$  = the sample number of elements  $a_i$ . But if one notes the convention that an element will always be indicated with a subscript and a letter (other than  $m$ ) without subscript will denote a frequency (lower case for a sample, capital for a universe) the mnemonic value outweighs risk of confusion.

As description of the (complex) universe of  $y$  the elements of (8.1) are not random variables. Relative to a sample they may be interpreted either as random variables or as possible values which such random variables may take. For precision notice that each group of elements corresponds to only one random variable; the position might be most clearly exhibited by writing the random variable for the first group as  $a(i)$  to indicate a function which takes on the value  $a_i$  at the  $i$ th sampling of factor  $A$ .

Suppose first that  $A, B, D$  are all finite, and that, case (i), there is a nested sub-universe of  $d_{ijk}$  for each  $ij$  combination. To obtain unique definition of the elements we impose the restrictions

$$\sum_i^A a_i = \sum_j^B b_j = \sum_i^A (ab)_{ij} = \sum_j^B (ab)_{ij} = \sum_{ijk}^D d_{ijk} = 0.$$

More commonly, case (ii), there may be only one universe of  $d_{ijk}$  whose elements associate at random with all treatments. Here we will consider them as subject to only the one restriction,  $\sum_{ijk}^N d_{ijk} = 0$ , although more generally "block" restriction may be imposed. Case (i) is commonly described as "nested sampling"; case (ii) represents a "completely randomized experiment".

Tukey (1949), Bennett and Franklin (1954) and Wilk and Kempthorne (1955), define the universe  $K_2$ -parameters (which they designate  $\sigma^2$ )

$$K_2 = \sum_i^A a_i^2 / (A-1), \quad K_b = \sum_j^B b_j^2 / (B-1), \quad K_{ab} = \sum_{ij}^{AB} (ab)_{ij}^2 / (A-1)(B-1)$$

$$K_d(i) = \sum_{ijk}^N d_{ijk}^2 / AB(D-1) \quad \text{or} \quad K_d(ii) = \sum_{ijk}^N d_{ijk}^2 / (N-1) \quad (8.2)$$

To simplify notation a  $K$  or  $k$  without numerical subscript is to be read as  $K_2$  or  $k_2$ . ( $K$ 's with other subscripts will be needed only in the appendix.) Definitions (8.2) will be called "specific variance components". They will also be referred to as  $K(r)$ , indicating  $K$  for "roman elements", both for brevity in distinction from  $K(\gamma)$  for "greek elements" to be defined later and as reminder that they are  $K$ -parameters rather than variances (Sec. 7.17).

If the universe of every group is finite and there are separate sets of  $D$  units for each  $ij$  combination (case (i)) the elements are defined as unique means of a complex universe which is potentially observable. In case (ii) the elements are defined only as averages over all possible randomizations with  $d_{ijk}$ . We can imagine them as means of a conceptual universe of  $A^2 B^2 D$  elements, that is with all  $N$  conceptually possible replications on every  $ij$  combination. However such universe

is not potentially observable: therefore, we are justified in describing such averages as "expectations", indicated by the operator E, implying the limits in probability, random sampling with equal probability being assumed.

More usually the size of at least one group, usually D, will be imagined as tending to infinity. The elements are then defined only as asymptotic limits in probability. The K are therefore, in general, functions of hypothetical quantities. ( $K_{ab}$  is an extension of Tukey's statistics for a universe with two-way restrictions.  $K_d(i)$ , being merely an average for AB sub-universes, involves no essential extension.)

The authors quoted above state the expectations of mean squares of analyses of variance in terms of K(r). These expectations are shown in Table 8.1 for nested classification based on the model

$$y_{ijk} = m + a_i + b_{j(i)} + d_{k(ij)} \quad (8.3)$$

where  $b_{j(i)}$  indicates nesting in classification i, etc. (notation introduced by Bennett and Franklin); and in table 8.2 for the model (8.1). Table 8.2 is presented as for case (i), for which the last term of (8.1) might better be written as  $d_{k(ij)}$  in conformity with (8.3); if  $d_{ijk}$  randomize with all treatments (case ii) the factors d/D are to be deleted.

Table 8.1 Nested Classification

	d.f.	E(MSq) in terms of K(r)			E(MSq) in terms of K( )		
		$K_d$	$K_b$	$K_a$	$K_\delta$	$K_\beta$	$K_\alpha$
A	(a-1)	$(1-\frac{d}{D})$	$d(1-\frac{b}{B})$	bd	1	d	db
B	a(b-1)	$(1-\frac{d}{D})$	d		1	d	
D	ab(d-1)	1			1		

Table 8.2 Crossed Classification (error terms nested)

	d.f.	$K_d$	$K_{ab}$	$K_b$	$K_a$	$K_\delta$	$K_{\alpha\beta}$	$K_\beta$	$K_\alpha$
<i>A</i>	(a-1)	$(1-\frac{d}{D})$	$d(1-\frac{b}{B})$	-	db	1	d	-	db
<i>B</i>	(b-1)	$(1-\frac{d}{D})$	$d(1-\frac{a}{A})$	da		1	d	da	
<i>AB</i>	(a-1)(b-1)	$(1-\frac{d}{D})$	d			1	d		
Error	ab(d-1)	1				1			

Delete d/D if  $d_{ijk}$  randomize with treatments.

The coefficients of the specific variance components in the expectations of mean squares can be evaluated in several ways. Bennett and Franklin, following suggestions by Tukey, have formulated an algorithm for most ordinary experimental designs. It does not, however, cover all situations. A general and simple method derives them from the coefficients for  $K(\gamma)$  as described later. Wilk and Kempthorne give a general procedure (discussed in Sec. 13 below) but one which is too complex, with consequent risk of algebraic errors, to be suitable for regular use. Main effect and error mean squares can be fairly directly evaluated in terms of Tukey's k-statistics (with a slight complication for randomized errors which do not form with the treatment effects simple "random sums" as defined by him); but they are not adapted to evaluate the interaction terms since  $K_{ab}$  as defined in (8.2) is not a member of that class of parameters. However going back to his "brackets" or "symmetric means" yields a straight-forward and general method which will be developed here because we shall want to refer to it later.

For initial simplicity consider the following abbreviated models:

$$y_{ij} = m + t_i + d_{ij} \tag{8.4}$$

where  $t_i$  represent a simple set of treatments, and  $d_{ij}$  may be read as  $d_j(i)$  to imply nested classification (case i), or as  $d_k$ ,  $k = 1 \dots (ad) \dots (AD) = N$ , to imply a completely randomized experiment (case ii): and

$$y_{ij} = m + a_i + b_j + ab_{ij} \quad (8.5)$$

which is (8.1) without its error element. For brevity, to save writing formulae twice, the range of i ('rows') will be taken for both cases as  $1 \dots a \dots A$ , though to conform with our usual notation these will represent  $t, T$  when applied to (8.4). The range of j is  $1 \dots d \dots D$  in (8.4),  $1 \dots b \dots B$  in (8.5); for formulae which apply to both models we will write  $c, C$  to represent either case as required. No other indication will be given as to whether formulae apply to one or either model.

Let  $[x_{ij}]$  be an  $A \times C$  matrix representing the elements of a two-way universe. Initially we take  $x_{ij}$  equivalent to  $y_{ij}$  but use a different letter because it will later be set equal to elements. Relative to (8.4) rows represent treatments and order within rows is at random. An experiment is formed by sampling  $d$  elements from each of  $a$  rows. There are  $\binom{A}{a} \binom{D}{d}^a$  possible samples. Relative to (8.5) rows and columns represent crossed variants of two factors. An experiment is formed by sampling  $a$  rows and  $b$  columns. There are  $\binom{A}{a} \binom{B}{b}$  ways of drawing such a sample. In each case every sample arrangement is assumed to be equally probable.

We want to express analyses of variance of samples or universes in terms of symmetric means analogous to these defined by Tukey (1950); but now products can be formed in more than one way with different implications. A more extended notation is therefore required. The following angular brackets are the Tukey-Hooke notation for "generalized symmetric means", abbreviated "g.s.m.". Square brackets are used to indicate sums in conformity with the notation of David and Kendall (1949).

Define:

$$\sum_{i=1}^a \sum_{j=1}^c x_{ij} = [1] = ac \langle 1 \rangle$$

$$\sum_{i=1}^a \sum_{j=1}^c x_{ij}^2 = [2] = ac \langle 2 \rangle$$

$$\sum_{i=1}^a \sum_{j=1}^{c(c-1)} x_{ij} x_{ij'} = \begin{bmatrix} 1 & 1 \\ \cdot & \cdot \end{bmatrix} = ac(c-1) \langle \begin{smallmatrix} 11 \\ \cdot \cdot \end{smallmatrix} \rangle \quad j \neq j'$$

$$\sum_{i=1}^{a(a-1)} \sum_{j=1}^b x_{ij} x_{i'j} = \begin{bmatrix} 1 & \cdot \\ 1 & \cdot \end{bmatrix} = a(a-1)b \langle \begin{smallmatrix} 1 \cdot \\ 1 \cdot \end{smallmatrix} \rangle \quad i \neq i'$$

$$\sum_{i=1}^{a(a-1)} \sum_{j=1}^{b(b-1)} x_{ij} x_{i'j'} = \begin{bmatrix} 1 & \cdot \\ \cdot & 1 \end{bmatrix} = a(a-1)b(b-1) \langle \begin{smallmatrix} 1 \cdot \\ \cdot 1 \end{smallmatrix} \rangle$$

$$\sum_{i=1}^{a(a-1)} \sum_{j,k=1}^{d^2} x_{ij} x_{i'k} = \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 1 & \cdot \\ 1 & \cdot \end{bmatrix} + \begin{bmatrix} \cdot & \cdot \\ \cdot & 1 \end{bmatrix} = a(a-1)d^2 \langle \begin{smallmatrix} 1 \\ 1 \end{smallmatrix} \rangle$$

Let  $[pq]^*$  and  $\langle pq \rangle^*$  be the analogous functions for the universe, that is replacing  $a, c$  by  $A, C$ ; for example

$$\sum_{i=1}^A \sum_{j=1}^C x_{ij}^2 = [2]^* = AC \langle 2 \rangle^* .$$

In Tukey's phrase each symmetric mean, represented by angular brackets, is "inherited on the average", that is

$$\text{ave} \langle pq \rangle = \langle pq \rangle^*$$

For example, consider  $\langle \begin{smallmatrix} 11 \\ \cdot \cdot \end{smallmatrix} \rangle$ . When summed over all possible samples symmetry requires that every ordered pair of elements must appear an equal number of times, namely  $\binom{A-1}{a-1} \binom{C-2}{c-2}$  = the number of samples which can contain a given row and a given pair of columns. Therefore

$$\text{ave} \langle \begin{smallmatrix} 11 \\ \cdot \cdot \end{smallmatrix} \rangle = \frac{\binom{A-1}{a-1} \binom{C-2}{c-2} \left[ \begin{smallmatrix} 1 & 1 \\ \cdot & \cdot \end{smallmatrix} \right]^*}{\binom{A}{a} \binom{C}{c} ac(c-1)} = \frac{\left[ \begin{smallmatrix} 1 & 1 \\ \cdot & \cdot \end{smallmatrix} \right]^*}{AC(C-1)} = \langle \begin{smallmatrix} 11 \\ \cdot \cdot \end{smallmatrix} \rangle^*$$

Similarly for the others. Note that the various types of products do not occur an equal number of times over all samples and do not have equal expectations.

From elementary algebraic relations

$$\left( \sum_a^a \sum_j^d x_{ij} \right)^2 = [1]^2 = [2] + \begin{bmatrix} 1 & 1 \\ \cdot & \cdot \end{bmatrix} + \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

$$\sum_a^a \left( \sum_j^d x_{ij} \right)^2 = [2] + \begin{bmatrix} 1 & 1 \\ \cdot & \cdot \end{bmatrix}$$

whence

$$adx_{..}^2 = \langle 2 \rangle + (d-1) \langle \begin{smallmatrix} 11 \\ \cdot\cdot \end{smallmatrix} \rangle + d(a-1) \langle \begin{smallmatrix} 1 \\ 1 \end{smallmatrix} \rangle$$

$$d \sum_i^a x_{i.}^2 = a \langle 2 \rangle + a(d-1) \langle \begin{smallmatrix} 11 \\ \cdot\cdot \end{smallmatrix} \rangle$$

$$\sum_j^d \sum_i^a x_{ij}^2 = ad \langle 2 \rangle$$

Taking differences and dividing by degrees of freedom leads to Table 8.3 for nested classification. Similar algebra using  $\langle \begin{smallmatrix} 1\cdot \\ 1\cdot \end{smallmatrix} \rangle$  and  $\langle \begin{smallmatrix} 1\cdot \\ \cdot 1 \end{smallmatrix} \rangle$  leads to Table 8.4 for the crossed classification.

Table 8.3

---

Treatments M. Sq.	$M_t = \langle 2 \rangle = \langle \begin{smallmatrix} 11 \\ \cdot\cdot \end{smallmatrix} \rangle + d( \langle \begin{smallmatrix} 11 \\ \cdot\cdot \end{smallmatrix} \rangle - \langle \begin{smallmatrix} 1 \\ 1 \end{smallmatrix} \rangle )$
Deviations M. Sq.	$M_d = \langle 2 \rangle - \langle \begin{smallmatrix} 11 \\ \cdot\cdot \end{smallmatrix} \rangle$

---

Table 8.4

---

<i>A</i> M. Sq.	$M_a = M_{ab} + b( \langle \begin{smallmatrix} 11 \\ \cdot\cdot \end{smallmatrix} \rangle - \langle \begin{smallmatrix} 1\cdot \\ \cdot 1 \end{smallmatrix} \rangle )$
<i>B</i> M. Sq.	$M_b = M_{ab} + a( \langle \begin{smallmatrix} 1\cdot \\ 1\cdot \end{smallmatrix} \rangle - \langle \begin{smallmatrix} 1\cdot \\ \cdot 1 \end{smallmatrix} \rangle )$
<i>AB</i> M. Sq.	$M_{ab} = \langle 2 \rangle - \langle \begin{smallmatrix} 11 \\ \cdot\cdot \end{smallmatrix} \rangle - \langle \begin{smallmatrix} 1\cdot \\ 1\cdot \end{smallmatrix} \rangle + \langle \begin{smallmatrix} 1\cdot \\ \cdot 1 \end{smallmatrix} \rangle$

---

To obtain expectations note first that on squaring the linear models, or any averages of them used in computing sums of squares, and averaging over all samples, the expectations of products between any two groups of elements vanish. We can therefore find expectations by letting  $x_{ij}$  stand for single groups of elements in turn and replacing brackets by their universe values.

Letting  $x_{ij} = t_i$ , since this is constant within rows

$$\langle \begin{matrix} 11 \\ \dots \\ \end{matrix} \rangle = \langle 2 \rangle$$

and the contribution of  $t_i$  to  $M_d$  vanishes. Also  $\langle \begin{matrix} 1 \\ 1 \end{matrix} \rangle$  is a simple mean product for a one way universe. Therefore  $\langle \begin{matrix} 11 \\ \dots \\ \end{matrix} \rangle - \langle \begin{matrix} 1 \\ 1 \end{matrix} \rangle = \langle 2 \rangle - \langle \begin{matrix} 1 \\ 1 \end{matrix} \rangle$  is a simple  $k_2$ -statistic as defined by Tukey, in our notation  $k_t$ , with expectation  $K_t$ .

Letting  $x_{ij} = d_k$  (case ii) the matrix is degenerate since there is no restriction of rows, and mean products in any direction must have the same expectation, namely  $-\langle 2 \rangle^*/(N-1)$  since  $[1]^* = 0$ . Therefore

$$E(M_t) = E(M_d) = \langle 2 \rangle^* - \langle 11 \rangle^* = \langle 2 \rangle^* N/(N-1) = K_d.$$

In case (i) (nested)  $\sum_D d_{ij} = 0$  within every row; therefore

$$\langle \begin{matrix} 11 \\ \dots \\ \end{matrix} \rangle^* = -\langle 2 \rangle^*/(D-1)$$

and

$$\langle \begin{matrix} 1 \\ 1 \end{matrix} \rangle^* = 0$$

Therefore

$$E(M_d) = \langle 2 \rangle^* - \langle \begin{matrix} 11 \\ \dots \\ \end{matrix} \rangle^* = \langle 2 \rangle^* D/(D-1) = K_d$$

$$E(M_t) = E(M_d) + d(\langle \begin{matrix} 11 \\ \dots \\ \end{matrix} \rangle^* - \langle \begin{matrix} 1 \\ 1 \end{matrix} \rangle^*) = \langle 2 \rangle^* (D-d)/(D-1) = (1 - \frac{d}{D})K_d$$

Extension to triple classification gives the formulae in Table 8.1. If the  $t$  treatments be regarded as made up of a  $\times$   $b$  crossed variants of two factors it is

not difficult to show that the expectations of the  $d_{ijk}$  terms are maintained in analysis of the two factors, and hence the factors for  $K_d$  in Table 8.2, the remainder of which we proceed to examine on model (8.5) ignoring error terms.

As before for  $t_i$ ,  $a_i$  and  $b_j$  being constant in rows and columns respectively they vanish from  $M_{ab}$  (Table 8.4), and  $\langle \begin{smallmatrix} 11 \\ \cdot\cdot \end{smallmatrix} \rangle - \langle \begin{smallmatrix} 1\cdot \\ \cdot 1 \end{smallmatrix} \rangle$  and  $\langle \begin{smallmatrix} 1\cdot \\ 1\cdot \end{smallmatrix} \rangle - \langle \begin{smallmatrix} \cdot 1 \\ \cdot 1 \end{smallmatrix} \rangle$  either vanish or are simple k statistics with expectations  $K_a$  and  $K_b$  respectively. Letting

$x_{ij} = ab_{ij}$  we have

$$\sum^A (\sum^B ab_{ij})^2 = 0 = \langle 2 \rangle^* + \left[ \begin{smallmatrix} 1 & 1 \\ \cdot & \cdot \end{smallmatrix} \right]^*$$

$$\sum^B (\sum^A ab_{ij})^2 = 0 = \langle 2 \rangle^* + \left[ \begin{smallmatrix} 1 & \cdot \\ 1 & \cdot \end{smallmatrix} \right]^*$$

$$\left( \sum^A \sum^B ab_{ij} \right)^2 = 0 = \langle 2 \rangle^* + \left[ \begin{smallmatrix} 1 & 1 \\ \cdot & \cdot \end{smallmatrix} \right]^* + \left[ \begin{smallmatrix} 1 & \cdot \\ 1 & \cdot \end{smallmatrix} \right]^* + \left[ \begin{smallmatrix} 1 & \cdot \\ \cdot & 1 \end{smallmatrix} \right]^*$$

which lead to

$$\langle \begin{smallmatrix} 11 \\ \cdot\cdot \end{smallmatrix} \rangle^* = -\langle 2 \rangle^* / (B-1)$$

$$\langle \begin{smallmatrix} 1\cdot \\ 1\cdot \end{smallmatrix} \rangle^* = -\langle 2 \rangle^* / (A-1)$$

$$\langle \begin{smallmatrix} 1\cdot \\ \cdot 1 \end{smallmatrix} \rangle^* = +\langle 2 \rangle^* / (A-1)(B-1)$$

and substituting these as expectations for the brackets in Table 8.4 (and multiplying by number of replications) leads to the coefficients of  $K_{ab}$  in Table 8.2. These results are of course already known, having been given by the authors quoted above.

The following indicates, without detailed working, extension of this form of analysis to three-way universes. Consider a three-factor experiment without replication:

$$y_{ijk} = m + a_i + b_j + d_k + ab_{ij} + ad_{ik} + bd_{jk} + abd_{ijk} \quad (8.7)$$

with our usual convention on numbers of variants.

Since we cannot conveniently write a three-way matrix on plane paper the preceding notation is inconvenient. Imagine the  $A$ -variants designated by rows,  $B$  variants by columns (horizontal), and  $D$  variants by verticals. The various sets of products of pairs of elements will be designated as follows:  $\underline{A}$  will indicate a pair of elements in the same row,  $\underline{a}$  a pair each of which comes from a different row; similarly  $\underline{B}$ ,  $\underline{b}$  for within and between columns;  $\underline{D}$  and  $\underline{d}$  for within and between verticals. Since these letters appear overworked it might at this stage seem advisable to use  $R$ ,  $C$ , and  $V$ ; but the reason is that they will lead to a simple algorithm for the mean squares. They will be distinguished by their associated brackets. We then define symmetric sums and means as follows, giving just a few examples:

$$\begin{aligned} [ABD] &= \sum^a \sum^b \sum^d x_{ijk}^2 = abd \langle ABD \rangle = abd \langle 2 \rangle \\ [aBD] &= \sum^{a(a-1)} \sum^b \sum^c x_{ijk} x_{i'jk} = a(a-1)bd \langle aBD \rangle \\ [Abd] &= \sum^a \sum^{b(b-1)} \sum^{c(c-1)} x_{ijk} x_{ij'k'} = ab(b-1)d(d-1) \langle Abd \rangle \\ [abd] &= \sum^{a(a-1)} \sum^{b(b-1)} \sum^{c(c-1)} x_{ijk} x_{i'j'k'} = a(a-1)b(b-1)d(d-1) \langle abd \rangle \end{aligned} \quad (8.8)$$

The rule for the number of terms in each sum (including permutations of each pair of elements) is easily seen from the examples. Then following the same procedure as above, the analysis of variance in terms of g.s.m. is found to be as in Table 8.5 which shows the multiples of symmetric means to be added to the mean squares indicated in the second column to give the total mean square for each row. It is

now visible that the above notation yields an algorithm for the formation of the mean squares similar to the well known one for treatment effects of a factorial experiment. For example, with letters inside angular brackets indicating symmetric mean products, those outside being replication multipliers:

$$MSq(ABD) = \langle (A-a)(B-b)(D-d) \rangle \tag{8.9}$$

$$MSq(A) = bd \langle (A-a)bd \rangle + MSq(AB) + MSq(AD) - MSq(ABD)$$

We next consider the model (8.1) a two factor experiment with replication. So far as concerns the analysis of variance it makes no difference whether we consider the error terms,  $d_{ijk}$ , nested within each treatment combination or associated at random with all treatments. The difference lies in restrictions which only come into play when we want to evaluate the expectations in terms of  $K_d$ . The difference from the foregoing is that since elements can be assigned to verticals at random there is no difference in the products within and between verticals when associated with between rows or columns. Thus in place of, for example,  $\langle aBD \rangle$  and  $\langle aBd \rangle$  we get

$$[aB.] = [aBD] + [aBd] = a(a-1)bd^2 \langle aB. \rangle$$

etc.; but  $\langle ABD \rangle$  and  $\langle ABd \rangle$  remain distinct, being respectively squares and products. These amalgamations lead to Table 8.6. Expectations of the g.s.m. for each group of elements leads directly to Table 8.2 without the composite inference previously used.

Table 8.5 Analysis of Variance of three-factor experiment

	M.Sq. +	Mean Squares							
		< ABD >	< ABd >	< AbD >	< Abd >	< aBD >	< aBd >	< abD >	< abd >
A	(AB + AD - ABD)	-	-	-	bd	-	-	-	-bd
B	(AB + BD - ABD)	-	-	-	-	-	ad	-	-ad
D	(AD + BD - ABD)	-	-	-	-	-	-	ab	-ab
AB	(ABD)	-	d	-	-d	-	-d	-	d
AD	(ABD)	-	-	b	-b	-	-	-b	b
BD	(ABD)	-	-	-	-	a	-a	-a	a
	-	1	-1	-1	1	-1	1	1	-1

Table 8.6 Analysis of Variance of two-factor experiment with replication: model (8.1)

	M.Sq. +	Mean Squares				
		< ABD >	< ABd >	< Ab. >	< aB. >	< ab. >
A	(D) + (AB)	-	-	bd	-	-bd
B	(D) + (AB)	-	-	-	ad	-ad
AB	(D)	-	d	-d	-d	d
D	-	1	-1	-	-	-

In Tables 8.1 and 8.2 the expectations are also written as they would be evaluated for a completely random model, the components which would thus be indicated being collectively termed  $k(\gamma)$  and  $K(\gamma)$ . By comparison with later tables we see that their definitions in terms of g.s.m. are: for one-way classification, from Table 8.3,

$$k_{\tau} = \langle \begin{smallmatrix} 11 \\ \dots \\ 1 \end{smallmatrix} \rangle - \langle \begin{smallmatrix} 1 \\ 1 \end{smallmatrix} \rangle ;$$

for a two factor analysis, from Table 8.4 or 8.6,

$$\begin{aligned}
 k_{\alpha} &= \langle \begin{smallmatrix} 11 \\ \cdot\cdot \end{smallmatrix} \rangle - \langle \begin{smallmatrix} 1\cdot \\ \cdot 1 \end{smallmatrix} \rangle \text{ or more generally } \langle Ab. \rangle - \langle ab. \rangle \\
 k_{\beta} &= \langle \begin{smallmatrix} 1\cdot \\ 1\cdot \end{smallmatrix} \rangle - \langle \begin{smallmatrix} 1\cdot \\ \cdot 1 \end{smallmatrix} \rangle \text{ or more generally } \langle aB. \rangle - \langle ab. \rangle \\
 k_{\beta\beta} &= \langle 2 \rangle - \langle \begin{smallmatrix} 11 \\ \cdot\cdot \end{smallmatrix} \rangle - \langle \begin{smallmatrix} 1\cdot \\ 1\cdot \end{smallmatrix} \rangle + \langle \begin{smallmatrix} 1\cdot \\ \cdot 1 \end{smallmatrix} \rangle \text{ or more generally } \langle ABd \rangle - \langle Ab. \rangle - \\
 &\qquad \qquad \qquad \langle aB. \rangle + \langle ab. \rangle
 \end{aligned}
 \tag{8.10}$$

and so on for more complex cases. Hooke (1954) obtained Table 8.4, and starting from the linear functions of g.s.m. (8.10) developed a family of statistics and parameters which he christened "bipolykays" to be used as mechanism for evaluating sampling moments of mean squares and specific variance components,  $k(r)$ , for model (8.5). He evidently regards  $k(r)$  as basic definitions of variance components for finite models.

The  $k(\gamma)$ , now extended to more than two dimensions, are all linear functions of g.s.m. with coefficients l independently of the sizes of samples and universes. They are therefore inherited on the average. Contrariwise the  $k(r)$  are linear functions with coefficients which involve both the observed sample size and the postulated group universe sizes. These introduce nuisance factors depending on universe postulates which are nearly always dubious, they destroy inheritance on the average, and make the sampling distributions clumsier than those of  $k(\gamma)$ . The linear model whence all derive is a quite arbitrary approximation to reality. If without making a model substantially more arbitrary we can get rid of these troubles we may be well served. Defining variances of the group universes by (8.2), instead of as the more classic and basic second moments, was already a move to gain some of the simplicity which was introduced by Tukey's generalized  $k$ -statistics, or "polykays", to study of simple universes. The logical continuation

is to consider adopting the  $k(\gamma)$  as basic descriptive statistics for complex universes. Subsequent sections consider their use in interpreting experiments. Since their formulae are invariant for finite or infinite group populations I shall term them canonical variance components.

Since comments both in the literature and at statistical meetings indicate that many feel confused about consequences of the interaction elements  $(ab)_{ij}$  being not independent of  $a_i$  and  $b_j$ , notice in passing that independence is not assumed in these definitions. Notice also that, although in a small universe the conditional arrays of  $(ab)_{ij}$  for given  $i$  are unlikely to be similarly distributed, and conversely for given  $j$ , nevertheless the mean second moment of the  $i$  arrays must be identically equal to the mean second moment of the  $j$  arrays by virtue of the definitions,

$$\text{Aver}(M_2|a_i) = \frac{1}{A} \sum^A \left( \sum^B (ab)_{ij}^2 / B \right) = \frac{1}{B} \sum^B \left( \sum^A (ab)_{ij}^2 / A \right) = \text{Aver}(M_2|b_j).$$
 This remains true for  $K(\gamma)$ , but is not true of the mean  $K_2(r)$  within arrays unless  $A = B$ .

Appendix to Section 8

Variations of mean squares

The theme of this paper is concerned with estimating variance components and, for that purpose, with expectations of mean squares. To consider distribution of estimated variances has not been part of its purpose. Tukey (1950) indicated that his generalized k-statistics could be applied to that problem apparently implying that association of elements of the linear models could be treated as randomized sums. The combination of treatment and error elements is, however, not quite of that form since (1) with nested universes the error mean to associate with each treatment comes from a different sub-universe, and (2) with randomized errors the association of several  $d_{ij}$  with each  $t_{i.}$ , leads to a greater multiplicity of arrangements than for simple randomized sums. In an unpublished report Tukey (1950a) sidesteps these complications without mentioning them and obtains variances of specific variance components for cases where groups associate at random by an ingenious method of inference from particular cases. Before that report was available to me I had obtained variances of mean squares for one-way classifications, model (8.4), by direct application of his formulae for randomized sums. The procedure seems of some interest for its incidental evaluation of generalized k-statistics of means and relative to a speculation by Wishart (1952). We consider here only the treatment mean square which is denoted  $dv_t$ ,  $v_t$  being a  $k_2$  statistic determined from  $t$  class means,  $y_{i.} = m + t_i + d_{i.}$ ,

$$v_t = \sum_{i=1}^t (y_{i.} - y_{..})^2 / (t-1)$$

Case (i), nested sub-universes, that is each set of  $d$  values of  $d_{ij}$  contributing to  $d_{i.}$  is from a different sub-universe. Initially we assume that  $d_{ij}$  have the same distribution in each of the  $T$  sub-universes, with  $K$ -parameters  $K_2$ ,  $K_{22}$  and  $K_4$  as defined by Tukey. (Since we now have to use subscripts to distinguish different orders of the parameters, these parameters will be written  $K_2(d)$  etc. when they have to be distinguished from those for the  $t$  group.) Then the moments of  $d_{i.}$ ,  $M_r = E(d_{i.})^r$ , are:

$$M_2 = \alpha K_2$$

$$M_4 = d\alpha(\alpha^3 + D^{-3})K_4 + 3\alpha^2 K_{22} \quad (A1)$$

where  $\alpha = \left(\frac{1}{d} - \frac{1}{D}\right)$

as may be derived from formulae given by Wishart (1952). We define also

$$K_4(d_{i.}) = M_4 - 3M_2^2 = \alpha\left(\frac{1}{d^2} - \frac{6\alpha}{D}\right)K_4 - \frac{6\alpha^2}{D-1}K_{22} \quad (A2)$$

because the  $d_{i.}$  are in effect sampled with replacement and are therefore as if from an infinite population with  $K_r(d_{i.})$  equivalent to ordinary cumulants and  $K_{22}(d_{i.}) = K_2^2(d_{i.}) = M_2^2$ . Note however that the  $K$  on right hand sides of (A1) and (A2) are defined for the universes of  $d_{ij}$  for which  $K_2^2 = \frac{D+1}{D-1}K_{22} + \frac{1}{D}K_4$ .

Now  $(y_{i.} - m) = t_i + d_{i.}$  is a "random sum" in Tukey's sense, with the  $t_i$  selected (without replacement) from a universe of  $T$  elements,  $d_{i.}$  selected from an infinite population. It follows from Tukey (1950) p. 507 that

$$E(v_t) = K_2(t) + M_2 \quad (A3)$$

as already given in Sec. 8. And extending Tukey's argument as indicated by him, pp. 517-8,

$$\begin{aligned} \text{var}(v_t) = & \left(\frac{1}{t} - \frac{1}{T}\right)K_4(t) + 2\left(\frac{1}{t-1} - \frac{1}{T-1}\right)K_{22}(t) + \frac{4K_2(t) M_2}{t-1} \\ & + \frac{1}{t}K_4(d_{i.}) + \frac{2M_2^2}{t-1} \end{aligned} \quad (A4)$$

This formula is equivalent to, for n random pairs of, say, a and b,

$$\text{var}(k_2(a+b)) = \text{var}(k_2(a)) + \text{var}(k_2(b)) + 4\text{cov}^2(ab)/(n-1) \quad (A5)$$

where  $\text{cov}^2(ab) = E \left[ \frac{\sum_{i=1}^n (a_i - a_{.})(b_i - b_{.})}{(n-1)} \right]^2$  = the variance of the covariance since by definition of the random sum  $E(\text{cov}) = 0$ . (Note however that the last term of (A5) is not  $2\text{cov}(k_2(a), k_2(b))$ , whose expectation is zero, but a term which arises from random association of sets of a and b and still exists even though the whole universes be used so that either or both of  $\text{var}(k_2(a))$  and  $\text{var}(k_2(b))$  may be zero.)

If we put  $t_i = \text{constant} = 0$ , only the last two terms of (A4) survive, representing  $\text{var}(k_2(d_{i.}))$ , and is the usual formula for variance of a sample variance from an infinite population as clearly it should be since in effect the same population of  $d_{i.}$  is being sampled with replacement. In terms of the K-parameters of the original  $d_{ij}$

$$\text{var}(k_2(d_{i.})) = \frac{1}{t} \left[ \frac{1}{d^2} - \frac{6\alpha}{D+1} \right] K_4 + 2\alpha^2 \left[ \frac{1}{t-1} - \frac{3}{t(D+1)} \right] K_2^2 \quad (A6)$$

Wishart (1952, p.8) notes the natural extension of moments of generalized k-statistics to cumulants but seems dubious about it. He writes: "... if we accept

Tukey's concept of an infinite population of samples of size  $n$  from a population of size  $N$  ... we might define a cumulant [in the manner of (A2)] ... But this question really needs further consideration." In Section 7.9 it was noted that all such statistics are random variables formed by the act of sampling and that the distribution of a random variable is inevitably that of a hypothetical infinite population; that this is fundamental and not merely a device. Therefore the use of  $K_4(d_{i.})$  as defined in (A2), and  $M_2^2$  for  $K_{22}$ , in the above formulae appear to need no further justification. However in view of Wishart's doubt I have checked the usage in two ways. First by deriving  $\text{var}(k_2(d_{i.}))$  by expansion and averaging over all possible combinations; cf. equation (A8). Secondly the complete formulae (A4) and (A6) were empirically checked by forming all possible values of  $k_2(y_{i.})$  from a universe of  $t_i = -1, 0, 0, 1$ , combined with all combinations of  $d_{i.} = -1, -1, 1, 1$ , which in turn can be derived either as singletons ( $d = 1$ ) from similar  $d_{ij}$  universes, or as samples of  $d = 3$  from universes of  $d_{ij} = -3, -3, 3, 3$ .

If the distribution of  $d_{ij}$ , and thence also of  $d_{i.}$ , differs from class to class, expanding  $k_2(d_{i.})$  and its square in terms of symmetric polynomials of the variables and evaluating their expectations leads to

$$E(k_2) = \bar{M}_2 = \sum^T M_{2i} / T \quad (A7)$$

as would be expected, where  $M_{2i}$  = the second moment of  $d_{i.}$  for samples from the  $i$ -th universe =  $\mathcal{K}_{2i}$ ; and to

$$\text{var}(k_2) = \frac{1}{t}(\bar{M}_4 - 3\bar{M}_2^2) + \frac{2\bar{M}_2^2}{t-1} - \beta \text{var}(M_2) \quad (A8)$$

$$= \frac{1}{t}(\bar{M}_4 - 3\bar{M}_2^2) + \frac{2\bar{M}_2^2}{t-1} + \left(\frac{t-3}{t(t-1)} - \beta\right) \text{var}(M_2) \quad (A9)$$

where  $\text{var}(M_2) = \overline{M_2^2} - \overline{M_2}^2$  and  $\beta = \frac{t^2 - 2t + 3}{t(t-1)(T-1)}$ . The first two terms are the same as (A6), or the last two terms of (A4) if  $M_{2i}$  is constant. (A9) is equivalent to using the average of  $K_1(d_{1.})$  evaluated for each universe; but clearly we do better to use (A8), with  $\overline{M_2^2}$  in stead of  $M_2^2$ , since  $\beta$  is negligible if  $T$  is large, and then (A8) is little affected by variation of  $M_{2i}$ .

Further complications may arise when we consider the effect on  $\text{var}(v_t)$  of allowing variable types of universes of  $d_{1j}$ . Writing

$$k_2(t) = k', \quad k_2(d_{1.}) = k'', \quad K_2(t) = E(k') = K',$$

$$\text{cov} = \sum_{t_1}^t (t_1 - t_.) d_{1.} / (t-1),$$

one finds

$$\text{var}(v_t) = \text{var}(k') + \text{var}(k'') + 4E \text{cov}^2 + 2(E(k'k'') - K'M_2) \quad (A10)$$

$$+ 4E(k'' \text{cov}) + 4E(k' \text{cov}).$$

The first three terms are the standard formula (A5). The last is zero because along with any given set of  $t_1$  ave  $\sum d_{1.} = 0$ . The other two terms are not necessarily zero:  $k'$  and  $k''$  will be correlated if  $t_1^2$  is correlated with  $M_{2i}$ ;  $k''$  and  $\text{cov}$  will be correlated if  $t_1$  is correlated with skewness of the distributions of  $d_{1.}$ . However in practical situations, e.g. Wilk and Kempthorne's second example, such correlations can safely be assumed to be negligible; so that we need consider only average moments of the error universes along with standard formulae as if all their distributions were the same.

Case (ii): a single universe of  $N = DT$  experimental units with deviations  $d_j$ ,  $j = 1 \dots N$ , randomizable with all treatments, and with parameters  $K_2, K_{22}, K_1$ .

Consider first a single partition of the  $N$  elements into sets of  $d$  elements each, giving a set of  $U = N/d$  values of  $d_{i.}^1$  of which  $t$  may be associated with the  $t_i$  to form a set of  $y_{i.}^1$ . The  $t$  values of  $(y_{i.}^1 - m) = (t_i + d_{i.}^1)$  thus formed are a sample of "randomized sums" as defined by Tukey (1950), the paired elements being selected from two universes of  $T$  and  $U$  elements respectively with parameters  $K_2(t)$  and,  $K_2^1(d_{i.}^1)$ . There are  $W = \frac{N!}{(d!)^U U!}$  possible sets of  $d_{i.}^1$ , and averaging over all such sets

$$\text{ave } K_2^1(d_{i.}^1) = \frac{\sum_{W} \sum_{U} \sum_{d} (\sum d_{ik}^1)^2}{W(U-1)d^2}$$

When we expand

$$(\sum_{d} d_{ik}^1)^2 = \sum_{d} d_{ik}^2 + \sum_{d(d-1)} d_{ik}^1 d_{ik}^1$$

and sum over the  $W$  arrangements, every possible  $d_j^2$  and product  $d_j d_{j'}$ , ( $j = 1 \dots N, j \neq j'$ ) occurs an equal number of times. Therefore, writing

$$\text{Ave}(d_j^2) = \sum_{N} d_j^2 / N \text{ and } \text{Ave}(d_j d_{j'}) = \sum_{N(N-1)} d_j d_{j'} / N(N-1)$$

we obtain

$$\text{Ave } K_2^1(d_{i.}^1) = \frac{WU \left[ d \text{ Ave}(d_j^2) + d(d-1) \text{ Ave}(d_j d_{j'}) \right]}{W(U-1)d^2}$$

But since  $\sum_{N} d_j = 0$ , we have  $-\sum_{N(N-1)} d_j d_{j'} = \sum_{N} d_j^2 = (N-1)K_d$ , and substituting these and  $U = N/d$  leads to

$$\text{Ave } K_2^1(d_{i.}^1) = K_d / d \tag{All}$$

Extending this argument yields

$$\text{Ave } K'_{22}(d_{i.}) = K_{22}/d^2$$

$$\text{Ave } K'_4(d_{i.}) = K_4/d^3 \tag{A12}$$

Tukey's (1950) argument then gives

$$\begin{aligned} \text{Ave var}'(v_t) = & \left(\frac{1}{t} - \frac{1}{T}\right)K_4(t) + 2\left(\frac{1}{t-1} - \frac{1}{T-1}\right)K_{22}(t) + \frac{4K_2(t) \cdot K_2(d)}{(t-1)d} \\ & + \left(\frac{1}{t} - \frac{1}{U}\right) \frac{K_4(d)}{d^3} + 2\left(\frac{1}{t-1} - \frac{1}{U-1}\right) \frac{K_{22}(d)}{d^2} \end{aligned} \tag{A13}$$

for the average variance of random sums formed within sets. To this has to be added the variance of  $K'_2(d_{i.})$  between sets (which is independent of  $t_i$ ). By

Tukey's relations

$$(K'_2(d_{i.}))^2 = \frac{U+1}{U-1} K'_{22}(d_{i.}) + \frac{1}{U} K'_4(d_{i.})$$

The average value follows from (A12).

Similarly

$$(\text{Ave } K'_2(d_{i.}))^2 = \frac{1}{d^2} (K_2(d))^2 = \frac{1}{d^2} \left( \frac{N+1}{N-1} K_{22}(d) + \frac{1}{N} K_4(d) \right).$$

Thence

$$\begin{aligned} \text{var}(K'_2(d_{i.})) &= \text{Ave}(K'_2(d_{i.}))^2 - (\text{Ave } K'_2(d_{i.}))^2 \\ &= \frac{2}{d^2} \left( \frac{1}{U-1} - \frac{1}{N-1} \right) K_{22}(d) \end{aligned}$$

Adding this to (A13) and substituting  $n = dt$ ,  $N = dU$ ,

$$\begin{aligned} \text{var}(v_t) = & \left(\frac{1}{t} - \frac{1}{T}\right)K_4(t) + 2\left(\frac{1}{t-1} - \frac{1}{T-1}\right)K_{22}(t) + \frac{4K_2(t)K_2(d)}{(t-1)d} \\ & + \frac{1}{d^2} \left(\frac{1}{n} - \frac{1}{N}\right)K_4(d) + 2\left(\frac{1}{t-1} - \frac{1}{N-1}\right)K_{22}(d) \quad (A14) \end{aligned}$$

The formula was empirically checked on a numerical example. Subsequently it has been found to agree with Tukey's (1950a, unpublished) formulae which were obtained in different manner for variances and covariances of the components.

Case (ii) is more generally applicable than case (i), and is much the more important one. Although at first sight it appears more complex, owing to there being no single universe of  $d_1$  from which to form random sums with  $t_1$ , it is interesting that it works out more neatly. Furthermore no secondary complications arise from variable distributions of several error universes.

9. The canonical variance model

Define a canonical latent equation

$$y = \mu + \alpha + \beta + (\alpha\beta) + \delta \quad (9.1)$$

(In mathematics "canonical" implies merely a "standard" form.) It is written without subscripts to imply random variables in the functional sense. Subscripts may be added to indicate elements of hypothetical populations, but without bothering about how to define individual elements (9.1) already serves the first function of a latent equation: it defines the structure of the population and sample, and directs that a sum of squares, either of a sample or universe, be analyzed as in tables 9.1. Since  $k(\gamma)$  of a sample, table 9.1 (a), are inherited on the average (Sec. 8), their expectations are the analogous functions  $K(\gamma)$  of the universe, Table 9.1 (b) or (c) depending on error term postulates.  $K(r)$  are by definition mean squares of the universe. Table 9.1 (b) or (c) therefore provides an explicit linear transformation between the two classes of parameters, namely:

$$\begin{aligned} K_a &= K_\alpha + K_{\alpha\beta}/B + K_\delta/BD \\ K_b &= K_\beta + K_{\alpha\beta}/A + K_\delta/AD \\ K_{ab} &= K_{\alpha\beta} + K_\delta/D \\ K_d &= K_\delta \end{aligned} \quad (9.2)$$

and conversely

$$\begin{aligned} K_\delta &= K_d \\ K_{\alpha\beta} &= K_{ab} - K_d/D \\ K_\beta &= K_b - K_{ab}/A \\ K_\alpha &= K_a - K_{ab}/B \end{aligned} \quad (9.3)$$

with terms in  $1/D$  deleted if error terms randomize with treatments. (Here and in expectations of mean squares in terms of  $K(\gamma)$  we will make the convention that

$D \rightarrow \infty$  for randomized error terms, as well as for postulated infinite error populations, in order to save re-writing separate formulae for the nested and randomized postulates. It is justifiable by noting that when errors randomize we

Table 9.1

		d.f.		Mean Squares			
(a) Sample				$k_{\delta}$	$k_{\alpha\beta}$	$k_{\beta}$	$k_{\gamma}$
<i>A</i>	(a-1)	$bdv_a$	= 1		d		bd
<i>B</i>	(b-1)	$adv_b$	= 1		d	ad	
<i>AB</i>	(a-1)(b-1)	$dv_{ab}$	= 1		d		
Error	ab(d-1)	$v_o$	= 1				
<hr/>							
(b) Universe ( all groups finite and sub-groups of $d_{ijk}$ nested)							
				$K_{\delta}$	$K_{\alpha\beta}$	$K_{\beta}$	$K_{\gamma}$
<i>A</i>	(A-1)	$BDK_a$	= 1		D		BD
<i>B</i>	(B-1)	$ADK_b$	= 1		D	AD	
<i>AB</i>	(A-1)(B-1)	$DK_{ab}$	= 1		D		
Error	AB(D-1)	$K_d$	= 1				
<hr/>							
(c) Universe (if $D \rightarrow \infty$ , or $d_{ijk}$ randomize with other elements: analysis of $E(y_{ij0})$ ).							
				$K_{\delta}$	$K_{\alpha\beta}$	$K_{\beta}$	$K_{\gamma}$
<i>A</i>	(A-1)	$BK_a$	=		1		B
<i>B</i>	(B-1)	$AK_b$	=		1	A	
<i>AB</i>	(A-1)(B-1)	$K_{ab}$	=		1		

in effect define treatment means as expectations of infinitely many samplings from the error universe with replacement, and the  $K(\gamma)$  are defined in terms of such expectations. But the convention does not carry over to higher moments of k-statistics, nor to sampling variances of class means, e.g. equations (11.6).)

Substituting (9.3) in Table 9.1 (a) leads at once to the expectations of mean squares in terms of  $K(r)$  as given in Table 8.2. This seems to be the simplest method for getting the coefficients of  $K(r)$  in such analyses - if they be required and considered worth getting.

Any universe is usually an arbitrary postulate made to focus interest on some region of temporary interest. We may today be concerned with variation between spindles of a particular loom manufacturing a particular type of cloth. But there are other looms and more spindles of which our observations might reasonably be a sample and to which we might at another time wish to extend our inferences. Conversely we might be interested to consider variation in a smaller loom, or in cloth made using only part of the observed loom. All such changes of postulates mean changes in the definitions of  $k(r)$  and  $K(r)$  which are tied to each postulate in turn. If having evaluated  $K_a$  for universes B and D, we want to evaluate  $K'_a$  for universes B' and D', we have so to speak to undo the old postulates and substitute the new with formulae such as:

$$\hat{K}'_a = k'_a = k_a + \left(\frac{1}{B'} - \frac{1}{B}\right)k_{ab} + \frac{1}{B'}\left(\frac{1}{D'} - \frac{1}{D}\right)k_d \quad (9.4)$$

The canonical  $k(\gamma)$  statistics are by definition invariant under such changes of postulates. They remind us that  $K'_a$  are not so much fixed parameters as functions of means associated with certain groupings; in terms of  $k(\gamma)$  the formula for  $k'_a$  involves only the new postulates without regard to prior ones.

$$k'_a = k(y_{i_0'0'}) = k_a + k_{a\beta}/B' + k_{a\delta}/B'D' \quad (9.5)$$

We write (9.4) and (9.5) in terms of estimated  $k$ 's because theoretically the  $K$  which they estimate may alter with the particular factor variants supposed included in a universe, although the estimating  $k$  do not so change, being dependent only on the sample whatever universe it be supposed to represent. These conditions give  $k(\gamma)$  a more fundamental position than the  $k(r)$  which depend on mutable postulates.

If  $\text{Ave}(a_i^2) = \sum a_i^2/A = K_a(A-1)/A$ , etc., be termed variances, distinction from sampling variances of statistics sometimes becomes a matter of some subtlety since we find ourselves applying the same name,  $\text{var}(a_i)$ , both to  $\text{Ave}(a_i^2)$  and to  $\mathcal{E}(\hat{a}_i - a_i)^2 = \mathcal{E}(y_{i..} - y_{i00})^2$  = the sampling variance of the estimate of  $a_i$  on repeated sampling from the subuniverse with mean  $(m+a_i)$ . In an earlier draft I tried to distinguish the two types of variance by different symbols,  $\text{var}$  and  $V$ , but the endeavour failed of consistency because all variances are essentially sampling variances (sec. 7.15). When the former is called a variance it means  $\mathcal{E}(a_i^2)$  for repeated samples of one with replacement from the universe of  $a_i$ . Thus the only basic distinction between the two, when both are regarded as variances, lies in the universes being sampled and not in the type of 'variance'. Therefore it seems best to distinguish  $\text{Ave}(a_i^2)$  as the second moment of the universe  $a_i$  or as proportional to the  $K$ -parameter.

For simplicity in the following discussion assume randomized errors. It is much the commoner and more important case. We then have

$$K_d = K_\delta \text{ and } K_{ab} = K_{\alpha\beta}$$

Recollect that a variance component model says that we are not interested in individual elements but only in dispersions which may be associated with variants of each factor. When interaction exists effects are a function of both factors acting jointly, and effects of one factor are not definable in isolation from the other. If we choose to associate certain magnitudes of effects and variances with factors singly and in interaction the partition can be only empirical and to some degree dependent on formal definition of what we shall mean by such phrases. At the initial stage of defining a latent equation the specific latent equation (8.1) seems simplest and most straightforward. But it is only empirical and if at a later stage of analysis it leads to non-simple statistics we can go back to consider

modified models which may be simpler over the whole run of deductions. Since  $k(\gamma)$  are substantially simpler statistics than  $k(r)$ , and form a complete set, it is reasonable to consider formulating a model for which they may themselves be interpreted as the dispersion components, not merely as a stepping stone to the specific components.

Consider the general variance equation (1.2) or (6.1). In a complex finite universe variances are still additive in the same way, but the K-parameters of either type are not. For a two-way universe:

$$M_2(y_{ij}) = \sum_{i=1}^A \sum_{j=1}^B (y_{ij} - y_{00})^2 / AB$$

$$= M_2(a) + M_2(b) + M_2(ab) \tag{9.6}$$

$$= (1 - \frac{1}{A})K_a + (1 - \frac{1}{B})K_b + (1 - \frac{1}{A})(1 - \frac{1}{B})K_{ab} \tag{9.7}$$

$$= M_2(\alpha) + M_2(\beta) + M_2(\alpha\beta) \tag{9.8}$$

$$= (1 - \frac{1}{A})K_{\alpha} + (1 - \frac{1}{B})K_{\beta} + (1 - \frac{1}{AB})K_{\alpha\beta} \tag{9.9}$$

(9.6) follows directly from the definition of elements in the specific latent equation, (9.7) gives its parts in terms of K(r) as these were defined by (8.2). The definitions of K( $\gamma$ ) necessitate (9.9). Assume that its parts are the variances of the elements of (9.1). Remembering the relation between K-parameters and variances we see that  $K_{\alpha\beta}$  appears as the K-parameter for a simple universe of AB elements, that is for a universe with only one location restriction, in contrast to the (A + B - 1) restrictions on  $ab_{ij}$ . We have merely to allow that the mean  $y$  for all variants of one factor with a single variant or subset of variants of the other factor may still contain a small part derived by definition from interaction. For example

$$y_{i0} = \mu + \alpha_i + \alpha\beta_{i0}$$

Since elements of  $y_{ij}$  are anyway empirical and imaginary (Sec. 1) this is a small price to pay for simplification of the variance analysis and its estimators.

Equating the specific (8.1) and canonical (9.1) latent equations

$$y_{ij} = m + a_i + b_j + ab_{ij} = \mu + \alpha_i + \beta_j + \alpha\beta_{ij}$$

and retaining arbitrary restrictions,  $\alpha_0 = \beta_0 = \alpha\beta_{00} = 0$ , the relations between the

two types of elements are seen to be

$$\begin{aligned}
 m &= y_{00} = \mu \\
 a_i &= (y_{i0} - y_{00}) = \alpha_i + \alpha\beta_{i0} \\
 b_j &= (y_{0j} - y_{00}) = \beta_j + \alpha\beta_{0j} \\
 ab_{ij} &= (y_{ij} - y_{i0} - y_{0j} + y_{00}) = \alpha\beta_{ij} - \alpha\beta_{i0} - \alpha\beta_{0j}
 \end{aligned} \tag{9.10}$$

The canonical elements are not uniquely defined but we do not require that they should be. It was part of the model to state that we are not interested in the elements individually but only in their dispersions. These and their estimators are uniquely and explicitly defined by Tables 9.1 and equations (9.8) and (9.9).

If we ask for estimates of the elements we alter the model by raising them from the status of random variables to that of location parameters (Sec. 7.9). Suppose we may wish to make this extension: the canonical model is over-determinate in the sense that it now has more parameters than observations available for estimating them, a not uncommon situation, for example: factor analysis. We can add the requirements that  $M_2(\alpha) = (1 - \frac{1}{A})K_\alpha$ ,  $M_2(\alpha\beta_{i0}) = (1 - \frac{1}{B})K_{\alpha\beta}$ , etc., but innumerable solutions are still possible. In practice one would postulate  $\alpha\beta_{i0}$  etc. = 0 and estimate  $\hat{\alpha}_i = (y_{i.} - y_{..}) = \hat{a}_i$ . If we could make observations such that the element estimators  $\hat{a}_i$  conformed to the dispersion requirement of the model to yield  $E(k_2(\hat{a}_i)) = K_a$  there would be some attraction in retaining the specific model. But in practice this never happens since  $E(k_2(\hat{a}_i)) = K_a + (\frac{1}{b} - \frac{1}{B})K_{ab} + K_d/bd$ . There is therefore no lost advantage in accepting similar estimators for  $\alpha_i$  subject to  $E(k_2(\hat{\alpha}_i)) = K_\alpha + K_{\alpha\beta}/b + K_d/bd$ . However, as indicated, this is all supererogatory since it really means a change from the variance component model to the regression model.

Error terms add little complication. With randomized errors (9.10) are defined by expectations of the  $y$  means and otherwise remain unaltered. With finite nested universes of experimental units we similarly remove the restrictions  $d_{ij0} = 0$  and add means  $\delta_{i00}$  etc. to the right hand sides of (9.10). They make a nice distinction between the two cases by reminding us that in nested universes the experimental unit are integral parts of the treatments and not experimental errors in the true sense. They are fundamentally similar to the interaction effects as discussed above since treatment effects cannot now be isolated from the associated units. A term  $(1 - \frac{1}{N})K_6$  now stands to be added to equation (9.9) irrespective of assumptions on randomization; contrariwise the term to be added to (9.5) for specific elements is not invariant for the alternative assumptions.

The purpose of the foregoing is to develop the parameters and statistics  $K(\gamma)$  and  $k(\gamma)$  as a complete and sufficient set for describing the composition of dispersal of a complex finite universe, without need to invoke hypothetical infinite population which some workers seem to regard as undesirable abstractions; and to indicate that their latent model is no more empirical than the usual specific model. All that has been done is to relax restrictions on the sub-group means of  $ab_{ij}$ , allowing them to take arbitrary values with variance inversely proportional to numbers in each sub-group. Such relaxation is permissible because the original restrictions, pinning these means at zero, were imposed on the specific latent equation only with intent to simplify algebra (Sec. 2); they need not be retained when found not to serve that purpose to best advantage.

Having thus relaxed restrictions on sub-group means it is reasonable to consider the effect of similar relaxation on all group means,  $\alpha_0$ ,  $\beta_0$ ,  $\alpha\beta_{00}$ , etc. Allow them to take arbitrary values conformable with the universe being a super-sample

from hypothetical infinite populations in all groups, subject only to the restriction that the expectations of every group, and covariances of pairs of groups are zero. The extension is reasonable because postulated universes are almost invariably restricted segments of potentially larger populations. This being done the canonical model (9.1) becomes identical with the general "random" model,  $K( )$  can be interpreted as variance components of the infinite population, and inferences about a restricted universe (super-sample) of which an observed sample is a part, follow from well known relations between statistics of a sample and of a sub-sample. Doing this adds obvious further terms to equations (9.8), for example

$$m = \mu + \alpha_0 + \beta_0 + \alpha\beta_{00}$$

$$a_i = \alpha_i - \alpha_0 + \alpha\beta_{i0} - \alpha\beta_{00}, \text{ etc.} \quad (9.8a)$$

With this convention the canonical latent equation most easily performs its second function: to provide machinery for evaluating expectations of mean squares and of sampling variances. Examples will be given in Sec. 11.

Defining 'canonical variance components' in the way suggested has then the following advantages over the 'specific variance components' (equations 8.2):

1. Being generalized k-statistics which are inherited on the average they are unique unbiased estimators of the universe K-parameters independently of size of sample and of universe postulates. This is probably the chief advantage. Postulates about universe sizes are nearly always introduced with a note of apology and uncertainty for the dubious numbers assumed. At other times they depend on, and may alter with, different sorts of questions to be asked of the same data (Sec. 11). Consequently they are better treated as tentative hypotheses several of which may be considered. What we need are statistics which can be estimated once and for all from each experiment and used as

required to indicate conditions under each such hypothesis which may seem of interest. The  $k(\gamma)$  supply this requirement.

2. If a universe be everywhere finite the canonical components are just as unique and concretely defined in terms of the universe elements as are the specific components. If, as is usual, any group is postulated as infinite or randomizable the specific components themselves are defined only in terms of asymptotic or expected values. They are therefore not really any more concrete than are the canonical components.
3. The expectations of mean squares can be evaluated simply and unambiguously by well-known rules. For a balanced experiment they can be written down at sight by Lee Crump's (1946) rule. Although Bennett and Franklin have given a reasonable algorithm for the specific components, it still remains more onerous if only because one has to evaluate different coefficients for each row.
4. For a balanced experiment the coefficients for a given canonical component are constant for all rows in which it occurs. This both simplifies writing down the expectations and facilitates solution of equations to estimate the components.
5. Any variance which may be required is expressible as a simple linear function of the canonical components. The specific components of course also yield linear combinations; but, except for the particular sample size observed and the postulated universe sizes on which their computation has been based, their coefficients are more complicated (examples in Sec. 11).
6. Interaction was originally, and is still, basically defined with respect to means in a factorial experiment. In that context it is essentially symmetric with respect to both (or all) factors involved. The interaction of  $A$  with  $B$  is identically the same as the interaction of  $B$  with  $A$ . Although a trivial

point it seems aesthetically desirable that its analogue in variance components should be similarly symmetric. The canonical components meet this pleasure, the specific ones do not.

7. Although contrary to fashionable current practice, Sec. 12 will argue that the canonical components are more meaningful for interpreting most experiments, and more relevant to defining tests of significance, than are specific components.

10. Regression as a limiting case of the generalized model.

Table 10.1, derived from table 8.2, gives the analysis of variance for a two-factor experiment with randomized replications in terms of specific variance components. If we now let  $a = A$  and  $b = B$  the factors

Table 10.1

Classification	d.f.	$K_d$	E(M.Sq.)		
			$K_{ab}$	$K_b$	$K_a$
<i>A</i>	(a-1)	1	$d(1 - b/B)$		db
<i>B</i>	(b-1)	1	$d(1 - a/A)$	da	
<i>AB</i>	(a-1)(b-1)	1	d		
Error	ab(d-1)	1			

for  $K_{ab}$  in *A* and in *B* become zero. Since then also the definitions of the  $K(r)$ , equations (8.2), become identical with those of  $\theta$  in Sec. 5, Table 10.1 becomes the same as Table 5.1 for the regression model.

Sec. 7.11 has noted that the parameters of a universe are an enumeration of its elements. With a universe of observable elements there is, of course, no estimation problem if all are observed; here the elements of each universe are central parameters of other universes or populations and an estimating problem still remains. But with representation of all elements of such a universe all its specific parameters become estimable, which in effect is the same as saying that the universe stands to be described by evaluating the means of its sub-classes individually, exactly what the regression approach does. Accordingly the regression model is commonly presented as the general analysis with the whole universe of elements observed, these elements then being commonly described as "fixed effects". The only confusing feature is that students often feel ambiguity about what is then meant by the universe of treatments. Few experiments cover all variants of a factor

whose evaluation would be useful if resources permitted. To say that the potential universe is observed is false, consequently the idea is often introduced with appearance of apology for an artificial postulate, although what is intended is not that  $ab$  has been increased to  $AB$  but that a sub-universe is selected for study which by definition reduces  $AB$  to  $ab$ . Although reasonable the appearance of artificiality is often difficult to avoid, and one finds students thoroughly confused with the query: "When is a variable 'fixed' or 'random'?" The answer is that "fixed effect" (rarely defined) implies a focus of interest. To give an element that name is to say that one is interested in determining its value, as a parameter in its own right, independently of, and not merely as a representative of, the class of similar elements whose potential number is then irrelevant, that is that it is to be treated as a regression coefficient and not as a random element from a universe. The definition relative to a restricted sub-universe is however useful in directing attention to the composition of means to which a regression surface is fitted.

Sec. 5 noted that, if we want to make a simplifying approximation by fitting fewer regression coefficients than are required for complete fitting to all class means, the variance of deviations about such simplified regression surface may be acceptable as an external estimate of error which may be preferred to the internal estimate with respect to deductions using the approximating surface. Analogous considerations in terms of the generalised model are most easily formulated in terms of canonical variance components for which the interaction components will not disappear from the main effect mean squares. We will return to this in the next section.

11. The "mixed model"

Relation of the "mixed model" to the regression and random models can be exhibited by splitting the linear model into two parts. Suppose  $a_i$  to be "fixed" effects;  $\beta_j$  to be "random". We can write the regression equation:

$$y_{ijk} = \mu + a_i + \varepsilon_{ijk} \quad (11.1)$$

$$i = 1 \dots a;$$

a canonical latent equation for the random variables:

$$\varepsilon = \beta + (\alpha\beta) + \delta; \quad (11.2)$$

and a variance equation:

$$\sigma_\varepsilon^2 = \sigma_\beta^2 + \sigma_{\alpha\beta}^2 + \sigma_\delta^2 \quad (11.3)$$

Equation (11.1) points up the composite random variable,  $\varepsilon$ , which determines deviations about the "regression". However, since groups of  $\varepsilon_{ijk}$  contain the same  $\beta_j$  and  $(\alpha\beta)_{ij}$  they are no longer independent and the regression cannot be fitted and interpreted quite so simply as formerly. The usual analysis of variance in effect fits  $b$  parallel regressions, one for each  $\beta$ -variant observed, minimising the sum of squares of the independent deviations  $(\alpha\beta)_{ij} + \delta_{ij}$ , that is the interaction sum of squares which accordingly becomes the error component of  $MSq(A)$ .

Current custom has endeavoured to avoid these complications by treating the case according to the general model for a universe with  $a = A$ . Interpretations of  $A$  effects is then straightforward. For example a treatment contrast  $(a_i - a_{i'})$  is estimated by  $(y_{i..} - y_{i'..})$ . The average  $\beta$  effects cancel from this difference and its sampling error depends on the interaction meansquare  $(A\beta)$  which measures deviations about the partial regressions. Following the canonical formulation its

sampling variance is

$$\begin{aligned}
 V(\hat{a}_i - \hat{a}_{i'}) &= E(y_{i..} - y_{i'..} - y_{i00} + y_{i'00})^2 \\
 &= E \left[ (\alpha\beta_{i.} - \alpha\beta_{i'0}) - (\alpha\beta_{i'1.} - \alpha\beta_{i'1'0}) + (\delta_{i..} - \delta_{i'00}) \right. \\
 &\quad \left. - (\delta_{i'1..} - \delta_{i'1'00}) \right]^2 \\
 &= 2 \left[ K_{\alpha\beta} \left( \frac{1}{b} - \frac{1}{B} \right) + K_{\delta} \left( \frac{1}{bd} - \frac{1}{BD} \right) \right] \quad (11.4)
 \end{aligned}$$

Finite population corrections apply to each bracket since the elements forming each universe mean include the corresponding sample means, and bracket pairs are uncorrelated on the average. The term in  $1/D$  is not here to be deleted for randomized errors. For the usual mixed model  $B$  and  $D$  tend to infinity and (11.4) becomes  $2 MSq(AB)/bd$  as is well known. A contrast between two  $A$ -variants for a given  $\beta_j$  is equivalent to considering a restricted sub-universe with  $b = B = 1$ ; (11.4) then yields its error variance as  $2K_{\delta}/d$  (when  $D$  is infinite) as is also well known from elementary considerations. A "main  $A$  effect" as usually defined for a factorial experiment is to consider the average contrast in a sub-universe with  $B = b$ , and thence with variance  $2K_{\delta}/bd$ , and so on. The absolute mean for treatment  $A_i$ , namely  $y_{i00} = m + a_i$ , is estimated by  $y_{i..}$ . This estimate being derived from the single average regression, instead of from contrasts within the partial regressions, its discrepancy contains a component due to observing only a sample of  $\beta$ -variants and its sampling variance must accordingly include  $K_{\beta}$ :

$$(y_{i..} - y_{i00}) = (\beta_{.} - \beta_0) + (\alpha\beta_{i.} - \alpha\beta_{i0}) + (\delta_{i..} - \delta_{i00}) \quad (11.5)$$

whence its variance is

$$\begin{aligned}
 \text{var}(y_{i..}) &= E(y_{i..} - y_{i00})^2 \\
 &= (K_{\beta} + K_{\alpha\beta}) \left( \frac{1}{b} - \frac{1}{B} \right) + K_{\delta} \left( \frac{1}{bd} - \frac{1}{N} \right) \quad (11.6)
 \end{aligned}$$

where  $N$  is the universe of elements from which the  $bd \delta_{ijk}$ 's are sampled.  $BD$  for nested classification,  $ABD$  for randomized error.

In terms of  $K(r)$  (11.5) would be, for nested sampling

$$(K_b + \frac{A-1}{A}K_{ab})(\frac{1}{b} - \frac{1}{B}) + K_d (\frac{1}{d} - \frac{1}{D}) \frac{1}{b} \quad (11.6a)$$

The factor  $(A-1)/A$  enters because of the nuisance that the average  $K_2(ab)$  within a row for given  $a_i$ , viz.

$$\frac{1}{A} \sum \frac{\sum_{j=1}^B ab_{ij}^2}{B-1} \text{ is not equal to } K_{ab}.$$

The last term follows from  $d_{i..}$  being an unrestricted mean of  $b$   $d_{ij}$ 's each with variance  $(\frac{1}{d} - \frac{1}{D})K_d$ . If error terms randomize the form of the last term has to be altered to that in (11.6). Similar modifications apply for (11.4) in terms of  $K(r)$ . Using canonical components variances can be written down at sight from equations like (11.5) paying attention only to the number of elements entering into each mean, of sample or of universe. With specific components the sampling structure must also be considered in determining the forms of the coefficients,

Interpretations concerning the factor  $\beta$  whose variants are postulated as "random" has occasioned considerable controversy. Both parties to the controversy appear to be operating within the framework of the general model, but evaluate mean square ( $\beta$ ) differently in terms of variance components; in effect one group evaluate specific components with  $a = A$ , the other evaluate canonical components. Both then assume that the appropriate error term, both for testing significance of mean square ( $\beta$ ) and for attaching to means of  $\beta$ -variants, is an estimate of the part of  $E(\text{mean square } \beta)$  additional to that part respectively labelled  $\sigma_b^2$

( $K_b$  or  $K_\beta$  in our notation). At this stage both seem to be overlooking precise postulates of the model and that although the mixed model can for many purposes be interpreted as a limiting case of the general model, the two are not identical.

We shall return to this after first examining the initial difference between the two interpretations of the mean square.

Although involving some repetition of Section 9, it seems worth while to try to elucidate the apparent discrepancy of models because it has occasioned considerable debate. One group (for example Mood, 1950; Hald, 1952, Mentzer, 1953; Scheffé, 1954) note that postulating an infinite population of  $\beta_j$  necessarily implies an infinite population of  $(\alpha\beta)_{ij}$ . Exactly how they reason next is not always clear. Usually both are assumed to be normally distributed, whence the definitional condition that they are uncorrelated implies independence, and a bivariate normal population. Whether or not normality be assumed, they do assume, perhaps without full consideration, that the population of  $(\alpha\beta)_{ij}$  is infinite for given  $\beta$  as well as for given  $a_i$ , and that the sample of  $a$  or  $A$  values associated with a given  $\beta_j$  is from such conditional population with expectation zero but not necessarily with sample mean zero. This leads to defining variance components as "canonical components", Table 9.1, except that  $K_\alpha$  may be written as  $\theta(\alpha)$  or  $\theta(a)$ .

The other group (for example Anderson and Bancroft, 1952; Kempthorne, 1952; Bennett and Franklin, 1954), avoiding assumption of independence, assume sampling whole columns from an  $A \times B$  array of  $(ab)_{ij}$  with  $B \rightarrow \infty$ , and the restriction  $\sum_i^A (ab)_{ij} = 0$  in every column. The effect is most easily seen by starting from the general model for finite universes, elements being defined by the specific latent equation (8.1), and analysis of variance as in Table 10.1. Putting  $B = \infty$  and  $a = A$  causes  $K_{ab}$  to drop out of  $E(MSq(\beta))$ . Exactly what happens as  $B$  tends to infinity is not discussed, being apparently assumed obvious. But most users of analysis of variance are not mathematicians, so let us consider what may be supposed to happen.

The postulated distribution of  $(ab)_{ij}$  appears as a (or A) distinct continuous distributions, one for each row (assumed similar, and normal if one wishes to add that postulate), but linked in such manner that a selection from any one row determines a column of elements, one from every row, satisfying the above restriction. Imagine a column of A elements for every j, the number of columns increasing indefinitely. There is no intention to suggest the distributions in every column to be the same apart from randomization with rows; first because such formulation would be contradicted by facts; second because it would imply the row distributions to be made up of the same A elements (with repetitions), and they could not be continuous, let alone normal. At first sight therefore this model seems to preclude independence of b and (ab). But a continuous distribution of b implies not only infinitely many elements in the whole population, but also infinitely many at every b value (as distinct from  $b_j$  as an element, Sec. 7.22), and therefore also infinitely many superimposed column distributions to form collectively a 'b-array'. The progression goes little further by allowing that all such b-arrays may have similar distributions, b and (ab) then being independent, although this is not necessary.\* In the limit the b-array distributions may be supposed to become continuous. We thus reach a continuous bivariate distribution with the variates (b and ab) uncorrelated and perhaps independent. The difference left from the previous formulation is that sampling from any b-array (given  $b_j$ ) is made conditional on choosing sets of a (or A) elements  $(ab)_{ij}$  with means identically zero.

Now suppose that the sets of a elements of  $(ab)_{ij}$  can be generated by random sampling from an infinite population of  $(\alpha\beta)_{ij}$  with variance  $K_{\alpha\beta} = K_{ab}$ , followed

---

\* This formulation indicates that for two variates, each normally distributed and uncorrelated, it is possible to formulate a non-continuous bivariate distribution in which the two variates are not independent, and the regressions not necessarily linear.

by subtracting the mean of each set:  $(ab)_{ij} = (\alpha)_{ij} - (\alpha)_{.j}$  where  $(\alpha)_{.j} = \sum_{i=1}^a (\alpha)_{ij}/a$ . Since the whole analysis is formal (Sec. 1) we can further imagine  $b_j$  to be made up of two uncorrelated parts:  $b_j = \beta_j + (\alpha)_{.j}$ . Given a set of  $\beta_j$  and  $(\alpha)_{ij}$  the  $b_j$  and  $(ab)_{ij}$  are uniquely determined, but not conversely: cf. equations (9.10). If we could observe without error the elements of real complete universes retention of the consequently uniquely defined  $b_j$  in the linear model might be preferred. But this rarely happens. Invariably any observation contains random variation and individual elements can be evaluated only as estimates of statistical parameters, variance between such estimates being greater than that between the postulated parameters. Estimates of  $b_j$  and  $\beta_j$  will be the same; the only practical effect of postulating one rather than the other in the model is to alter the definition of what is measured by variance between the estimates ( $y_{.j}$ ), proportionate to  $MSq(\ )$ . Since anyway this is going to have a component  $K_\beta$ , to postulate the remainder as having a component  $K_{\alpha}$  is of no consequence if by so doing we can slightly simplify general procedures. Using the transformation

$$\begin{aligned}
 b_j &= \beta_j + (\alpha)_{.j} \\
 (ab)_{ij} &= (\alpha)_{ij} - (\alpha)_{.j}
 \end{aligned}
 \tag{11.7}$$

the variance relations are

$$\text{var}(b) = K_b = K_\beta + K_{\alpha}/a
 \tag{11.8}$$

$$\text{var}(ab) = K_{\alpha} (1 - 1/a)
 \tag{11.9}$$

These are consistent with all postulates.  $K_b$  is equal to  $\text{var}(b)$  because the population of  $b_j$  has been postulated as infinite.  $\text{Var}(ab)$  follows from (11.7) by the usual argument for variance about a sample (or universe) mean, and corresponds with the original definition of  $K_{ab}$  which differs from  $\text{var}(ab)$  since  $a$  is finite (Sec. 7.17).

The difference between the two groups of writers is thus seen to lie merely in the linear transformation (11.7) of the nominal components, the first group using  $\beta_j + (\alpha\beta)_{ij}$ , the second  $b_i + (ab)_{ij}$  in their linear model. The transformation consists in taking a part  $(\alpha\beta)_{.j}$  out of  $b_j$  and adding it to  $(ab)_{ij}$ , the part being randomly selected subject to having variance  $K_{\alpha\beta}/a = K_{ab}/a$  and being uncorrelated with  $\beta_j$ . The transformation is permissible since the analysis of  $y_{ijk}$  into component elements is from the start formal. In recombining parts and their variances to answer specific questions both formulations must lead to the same answers. The more flexible canonical formulation seems preferable for basic analysis because the restriction  $a = A$  may be undesirable with respect to defining  $K_b$  and  $K_{ab}$ .

Equations (11.1) to (11.3) define a mixed model as regression on  $A$ -variants while  $B$ -variants are a random selection. That is, it diverts attention to mean yields of each observed  $A$ -variant, and says that we are to evaluate only dispersion of  $B$ -effects without regard to their individual values. The first of these introduces the distinction between the mixed model and a true general model. The former does not say that the whole universe of  $A$ -variants is observed. Secondly when we turn to consider individual  $B$  effects we abandon the model originally postulated, and change over to regression on the  $B$ -variants. With this change of model the background universe of  $A$  variants should be reconsidered. The restriction  $A = a$  was only a device for directing attention to means of observed  $A$ -variants; relative to assessing  $B$ -effects it may be unnecessarily restrictive and artificial. We are now free to consider defining  $B$ -effects relative to any  $A$  universe which may seem appropriate for this new purpose. Similarly the specific variance component  $K_b$  can be defined for a general model with any  $AB$  universe

independently of the limiting case. The model used by the first group of writers above in effect uses this freedom to define  $\beta$ -effects for an infinite population of  $A$ -variants—nominally they do it only for an infinite  $(\alpha\beta)_{ij}$  universe for every  $j$ , but it comes to the same thing since the  $\alpha_i$  cancel out whether this number be assumed infinite or restricted.

The mixed model is commonly assumed when one of the factors is quantitative (for example crop varieties by levels of a fertilizer), the quantitative factor being taken as the fixed one. Sec. 4 argued that variance component analysis is not suited to quantitative factors whose observed levels cannot rationally be regarded as a random sample from a universe of levels; so what should we do when a quantitative factor is crossed by a (random) qualitative one?

Some writers (for example, Bennett and Franklin, pp. 369-370) argue that assumption of a continuous relation between two variables is a technical decision, and that statistical inference can be applicable only to observed levels of a quantitative factor. But it is futile to suggest that we should concern ourselves only with such isolated observations. The purpose of research is to seek general laws of as wide applicability as possible (which these writers, despite their disclaimer, of course proceed to do), and the province of statistics is the whole chain of inductive reasoning from observations to general law. When dealing with quantitative factors we have no parent frequency distribution and no random sample, therefore the bridge from particular to general cannot be based on the consequences of random sampling from a frequency distribution. The analogous function as stepping stone from particular to general, and an essential part of the inductive procedure, is now played by the assumption of a continuous relationship. Either to disclaim onus for the assumption, or to restrict statistical inference to the observed

points, is pedantic and unreasonable.

Unless for some reason "external" error is greater than "internal" as indicated by replications (in which case the "interaction" is properly a random error rather than a real interaction of the two factors) interaction with a quantitative factor would not ordinarily occur at random. Desirable analysis is therefore to subdivide it in search of systematic components. Practicable procedure may be to fit regressions on levels ( $x$ ) of the quantitative factor for each qualitative variant ( $\beta_j$ ) and to study variation of the regression coefficients, (examples , Sec. ).

I do not believe that fixed definitions of main effects etc., should be laid down. A research worker should rely on his wits according to circumstances of each case. Suppose the regression on  $x$  may be adequately fitted by a quadratic polynomial for each  $\beta_j$ :

$$y'_{ij} = y_{ij} - \delta_{ij} = c_{0j} + c_{1j}x_i + c_{2j}x_i^2 \quad (11.8)$$

Dropping subscript  $i$  gives the functional or interpolation form for all  $x$  at given  $\beta_j$ . Suppose  $x_i$  to be measured as deviations from its mean ( $\bar{x} = 0$ ). (Since this paper deals only with balanced experiments we are supposing the same set of  $x_i$  for every  $\beta_j$ ; but similar considerations carry over, with some complications, to more general cases.) Modern practice would usually fit orthogonal polynomials, the quadratic form being

$$y'_j = y_{.j} + c_{1j}x + c_{2j}(x^2 - \bar{x}^2) \quad (11.9)$$

Contrasts between  $y_{.j}$ 's correspond to the usual definition of main effects, but these values do not fall on the curves. In effect they estimate  $y'_j$  at

$$x_{0j} = -c_{1j}/(2c_{2j}) \pm (c_{1j}^2/4c_{2j}^2 + \bar{x}^2)^{1/2}$$

Therefore unless  $c_{1j}/c_{2j}$  is constant for all  $j$  (or can be supposed constant but for experimental error), the usual definition of main effects is making comparison of

estimates at different  $x$  levels for each  $\beta_j$ . In many circumstances or for some purposes these comparisons may sufficiently well evaluate average  $\beta_j$  effects, but the point may occasionally deserve more consideration relative to individual circumstances than it usually receives.

When a quadratic function is adequate, it is evident from (11.8) or (11.9) that  $c_{1j}$  measure the slopes of all regressions at the same arbitrary (but not necessarily most interesting) level,  $x = 0$ . If higher order polynomials have to be fitted, considerations similar to those for  $y_{.j}$  would apply to the linear coefficients of the orthogonal forms; and so on as the order of polynomial is increased.

When using the average regression to describe average  $A$  effects, that is average variation with the quantitative factor, the standard errors of its coefficients will depend on their variation with the  $\beta$ -variants. Variance of the constant term,  $y_{...}$ , evidently involves variation of the "main" effects,  $y_{.j}$ , as in equation (11.6); while the other coefficients correspond to contrasts like  $(\alpha_1 - \alpha_{i_1})$  and their variances will involve only the formal interactions (in addition to experimental error) analogous to equation (11.4).

12. Error terms and tests of significance.

The chief use to which specific variance components have been put is to indicate appropriate error variances for testing significance of any given mean square in an analysis of variance. We exclude here interpretation of main effects and interactions according to their classical definition. That was noted in section 5 to be a pure regression formulation, defining contrasts between selected treatment combinations without references to potential universes which they might represent. We have since noted that that position can be reproduced as a limiting case of the general model by defining sub-universes containing only the observed treatments, but that this is little more than a convention to link the regression (attention to individual means) with the general model. The idea here is that we are interested in dispersion of real potential universes, and the question is what measure of dispersion do we really want to evaluate and test, independently of individual contrasts which can be handled as in Section 11. The question is one of considerable current debate.

The prevailing trend, as illustrated by Bennett and Franklin, seems to be to recommend formulating models according to the specific latent equation, specifying the potential universes of each group, and seeking to interpret the analysis of variance as indicated by Table 10.1. Some recent writers imply, somewhat dogmatically, that error variances against which each mean square of an analysis of variance should be weighed, must be determined according to its composition as evaluated for specific components with finite population correction factors. In fact, when it comes to examples, they almost invariably make arbitrary decisions that the variants of some factors will be considered individually, others will be taken as samples of indefinitely large universes. In practice the general model is rarely used except in this conventional degenerate form which might be better

replaced by a less rigid mixed regression and random model. The only example known to me with a realistic finite universe is that of parts of a loom given by Daniels (1939). Furthermore, as illustrated in discussion of contrasts in Section 11 such arbitrary conventional partitions merely tie one's hands unnecessarily with respect to evaluating comparisons among those variants which were postulated as random. Compare example , Section .

However, suppose that universe sizes for each group can be postulated, giving rise to a genuine finite universe model. To follow the advocated procedure means: firstly that a different error variance has to be computed for each mean square to be tested; secondly that these error variances will be linear functions of observed mean squares, with ensuing complications including the chore of computing quasi-degrees-of-freedom numbers for approximating  $F'$  tests. Are such troubles really necessary or worth while? Do they gain anything?

If interactions can be assumed zero so that the interaction  $K$  can be dropped the main effect  $K$ 's become effectively the same in both specific and standard formulations and no problem arises - except by mischance when an interaction which should be zero accidentally appears significant. If  $A$  be really inert  $B$  cannot behave differently with different  $A$ -variants. Therefore if the interaction be significant and we conclude that it is real we must be concluding that both factors produce effects. To test the main effect of  $A$  for the null hypothesis that it produces no effect is then pointless. The problem therefore is, assuming interaction present, what do we really want to test under the name of main effect variance component? The 'specific school' say that they want to test for differences among the universe means of  $A$ -variants; but if interactions are real it would be very surprising if they should exactly balance to make these means equal, so we can safely bet, without test, that they do differ and the test is redundant. It tests for  $b_k + b_k/B$  being greater than zero, and if  $b/B$  is not too small it

could be just another test for the interaction. Of more practical interest is to ask whether one variant is consistently more effective than another over and above such average difference as might be anticipated from chance arrangement of interaction effects. In practical applications one will normally use only one treatment combination; rarely will future action produce the average of a universe of crossed treatments. The question therefore is, the cross variant to be used in a future application being unknown, what can we say about average response that may be anticipated whatever one be used and supposed selected by chance? (If we can specify a particular cross variant to be used we have to consider individual treatment combinations instead of average main effects.) All of which merely says that to test for evidence that  $K_{\alpha}$  is greater than zero will generally be of more interest than to test for  $K_{\alpha}$ . For example: suppose the  $A$ -variants are drugs of a certain chemical family,  $B$ -variants are varieties of a genus of cocci bacteria;  $y$  being some measure of health of test animals. The universes of these factors may have reasonable assignable sizes. The action of the drugs may be fairly specific. If we can have accurate diagnosis of the particular variety of coccus to be treated we choose the drug according to the best specific combination. But if, from lack of time or of facilities, the particular variety is unknown, or is one which was not included in the experiment, we must choose the drug which does best on the average and the measure of confidence in its effectiveness depends on the interaction variance. The dispersion of average  $A$  effects per single chance cross variant is evaluated by  $K_{\alpha}$ . Although dealing with finite universes, the universe means whose spread is measured by  $K_{\alpha}$  are irrelevant to practical application.

A similar point of view was expressed, all too briefly, by Yates (1946, pp. 17 and 42) relative to sample surveys. Before doing a survey we can safely say that no two cities or states, etc. would show the same mean for any character in a

complete census. A test for that is of no interest. If we make a test for significance of a difference between classes of a sample survey it can only be to answer the question: is there some influence operating to make the individuals of one state by and large different from those of another state? If we could sample again and again in unlimited time while general environment remains the same, or if we could subject innumerable different samples of individuals to the environment of two states, is there some factor operating differently in each state so that a difference similar to that observed would be consistently reproduced? In other words a test of significance, like random variables, always has in its background hypothetical infinite populations. Tests of significance should be those indicated by a model allowing infinitely repeated sampling. (Cf. also Deming and Stephan, 1941, on interpreting censuses as samples.)

Some difficulty of interpretation occurs when  $A$  is not significantly greater than  $AB$  and  $AB$  is not significantly greater than Error, but  $A$  is significantly greater than Error. No rigid rule can be given, some discretion should be allowed depending on what ancillary information may be available as to most reasonable interpretation. As a rough general rule we might pool  $A$  and  $AB$  as a portmanteau test for magnitude of  $A$  effects whether due to interactions or consistently similar with all  $b_j$ . (Whether or not  $B$  should also be thrown in the pool again depends on circumstances. If  $B$  is clearly significant it would be kept out since the problem then relates only to  $A$ . If it also is insignificant relative to  $AB$  it would go in to test whether or not treatments in general are having effects irrespective of combinations. And of course conversely for testing  $B$  if  $A$  is or is not independently significant.) If the pooled test is significant we may conclude that effects are produced though we remain uncertain whether they are mainly interactions are consistent across the board, or are a mixture of the two.

If a main effect is not clearly significant relative to interaction, while differences between treatments at large are big enough to merit attention, the indication is that we cannot categorically recommend one quality of a factor as best irrespective of crossed variants.  $K_{\alpha\beta}$  being large relative to  $K_{\alpha}$  (what we mean by relative depends on circumstances and the economics of the case) is a warning that decision about best operating procedure requires that we study combinations individually.

Canonical variance components ( $K(\gamma)$ ) may be negative. A negative sample value is not necessarily to be regarded as an accident of sampling and interpreted as zero.  $A$  being significantly less than  $AB$  is evidence that some form of compensation is taking place, evidence which will usually be worth having. Appearance of a  $k(\gamma)$  with negative sign and appreciable numerical magnitude will bring it to attention; whereas analysis based on specific components would tend to overlook it. (Cf. Yates and Zecopanay, 1935; dealing with special conditions they termed the effect "competition".)

In view of these considerations, recent worry over what are appropriate error terms, entangled by certain postulates about potential universes and complicated by finite population adjustment factors, seems to be creating unnecessary complications for routine working of statistical tools. Attention to practical operating conditions to which consequent recommendations will be applied, and to the potential infinite resamplings throughout which a reported effect may be expected, may usually indicate the straightforward classical tests to be relevant.

In general usage the conventional significance levels,  $P = .05, .01, .001$ , are arbitrary and do not have to be taken as literally exact. The vaguer terminology—significant, highly significant, very highly significant—is, I believe, intended to remove the statements from the aura of accuracy implied by precise

figures and to indicate merely a rougher classification of "degrees of belief" which may suffice to accept or reject a hypothesis for practical working. Inferences from experimental data will usually be made in circumstances which may approximate more or less to one of the following three 'climates':

- (1) We have to deal with a problem in pure science. We know we cannot make a perfectly exact statement, the whole inferential apparatus—linear models, etc.—leads only to approximate description. The point at which we stop saying 'Get more evidence' and rest with 'This is accurate enough for the time being', depends on contentment, perhaps forced by available resources, which will be determined by a subjective degree of belief or confidence in the results and only roughly conditioned by a nominal  $P$  value.
- (2) Immediate action is required and we have to advise on what seems best from available evidence. This will be the treatment which was best in our sample and has to be recommended whether the significance of its advantage over second best was at a .01, .3 or only .9 significance level.
- (3) We have to weigh benefit of a new process against cost of bringing it into operation. For this decision we do need accurate probability assessment, but the relevant null hypothesis will rarely be that the specific variance component  $K_a = 0$ . The probabilities to be assessed will have to be determined in light of all the circumstances of each individual problem, in particular with respect to influences bearing on future operation which will not usually be those of an arbitrary universe supposed sampled by an experiment and belonging more to history of the experiment than to future application.\*

---

\* For a sophisticated discussion of these points and an approach towards a definitive decision theory see Lindley (1953) and discussion on that paper.

Interpretation according to the specific model and its concomitant Table 10.1 may be correct for some purposes. But research means keeping alert to alternatives. Those who advocate the specific model seem to be tying themselves too rigidly to arbitrary or dubious universe postulates for all interpretations from a given experiment. The chief thing is to avoid being dogmatic. The over-riding advantage of the canonical analysis is to remain free from such entanglements, while still allowing them to be easily introduced at a later stage as and when they may be relevant to answer specific questions.

13. Randomization tests and unidentifiable interactions.

Wilk and Kempthorne (1955) have endeavoured to work out a "logical derivation of linear models for experimental situations" such that, given the pattern by which observations are taken, the model can be written by objective rules, whence may follow without ambiguity the expectations of mean squares and estimates of error variances. For brevity that paper will be referred to as WK. Their formulation appears intriguing; but deeper inspection shows it to be unsatisfactory at three points: inference from randomization tests is too restrictive, their algebraic mechanism to represent random sampling is computationally cumbersome and wraps the theory in a dialectic haze, and their treatment of unidentifiable interactions drags in redundant complications; besides that it suffers from the rigidity in interpreting mixed models which has been indicated above to be undesirable.

Kempthorne (1952) stated a preference whenever possible to base significance tests on randomization without appeal to infinite population theory. WK note, what no one will dispute, "that it is possible to draw inferences statistically only about the population from which samples are drawn according to probability considerations", and that the parent universe of a randomization test is the particular set of plots or experimental units on which an experiment was done. They continue to the logical conclusion about randomization tests: "extensions of such inferences to wider circumstances cannot be assessed probability wise". In other words a statistician using randomization tests can act only as historian evaluating what might have been on a certain past occasion if the dice rolled different ways. They recognize that this is not the information wanted, and are endeavouring to emphasize "the tenuous relation between the physical situation and mathematical or statistical abstractions". The stand seems to veer close to encasing the statistician in an

ivory tower where he can say: "I guarantee that arrows shot in my tower will hit their mark in the way I say; but I wash my hands of any responsibility for what they will do when used in the big world outside."

An inference which is applicable only to a historical group of experimental units which will never be used again is of no interest to anyone. An experiment is useful only in so far that we can expect phenomena similar to those which it displays to be reproduced in a wider sphere. As far as possible a scientist should state only that of which he can be sure. But there comes a point where some speculation must be risked to gain objectives, and a statistician's duty would seem to include sharing responsibility for inferential jumps rather than putting them entirely on the shoulders of those who may be ill equipped to assess the risks. One of my early clients was a chemist who at first did not like probability statements. "To heck with your five per cents: is this result right, yes or no?" If I retreated even further to say: "There is a five per cent chance that the conclusion is wrong for the particular hundred specimens you observed in the laboratory. I can give you no answer at all about what will happen in the factory"; he would presumably reply: "In other words, I must judge for myself as of yore. Therefore you are quite useless to me, I need not consult you again"; and he would appear to be justified.

No one will disguise that there is a difficult jump from an experiment to factory or farm conditions. Whatever pains we take to sample realistic circumstances we can never be sure that future commercial production will draw its materials from a supply exactly similar to a population we sampled. But to retreat to the point of saying that the universe we sampled is only the  $N$  units actually observed and nothing more is unpractical and stultifying. We must endeavour to arrange that

experimental material will be a sample of a realistically large bulk of material available for commercial use, describe as well as possible the bulk which was properly sampled and for which inferences are valid, and leave to the farmer or production manager only to decide whether or not the resources available to him are reasonably similar or differ in a way for which rational allowance may be made. I would go further to say that (with rare exceptions) economics prohibit doing an experiment unless its inferences may be applicable to a bulk of material so vastly greater than that used in the experiment that it constitutes virtually an infinite population. In other words infinite population theory is almost always applicable to the error elements of an experiment. We do not hide that it is often difficult to define just what population we did sample, the difficulty is especially great and is well recognized in agricultural experimentation. We can but do our best to state what sort of fields and weathers our experiments reasonably represent.

These aspects of randomization tests were recognized in the discussion given by Welch (1937), Pitman (1938), Pearson (1937), and Johnson (1948). Pearson notes a further point, namely that randomization tests are distribution-free only with respect to control of type I error: the optimum choice of critical region, or of the criterion to be randomized, depends on the form of the population sampled. He notes further, as illustrated by the work of Tedin on blank experiments, that an experimenter is rather more concerned with repetitions which will belong to different randomization sets. "Some of these distributions ... would be biased in one way, some in another, so that when they are all combined together the resulting  $z$ -distribution should approach that of normal theory. From each randomization set the experimenter is concerned in fact with only one value of  $z$ , and this has been selected at random ... ; consequently from the point of view of his long run

experience, the appropriate probability distribution for him to use would appear to be that of normal theory." "Possibly we have here another instance of the difference ... between regarding a test as giving essentially a rule to be applied and justified by long-run experience, rather than a probability measure associated with an isolated experiment." "The conception of randomization illustrated in the examples given above is both exceedingly suggestive and often practically useful, but perhaps it should be described as a valuable device rather than a fundamental principle. Its adoption, when it can be followed by the calculation necessary to determine what I have described as the class I elements, ensures accuracy in the determination of the probability level of a test criterion, but without the aid of some further principle it cannot help us to decide which of a number of alternative tests to choose. It seems hardly possible to build the methods of statistics into a consistent whole without facing squarely the why of that choice."

That to postulate an infinite population of experimental units simplifies statistical treatment is a welcome sequel; it is not the reason for making the postulate.

A characteristic of WK's formulation is the algebraic method which they use to represent random sampling. Cornfield (1944) suggested writing a function of a sample from a universe as a function of all the elements  $X_i$  in the universe with dummy variables:

$$\begin{aligned}\delta_i &= 1 \text{ if the } i\text{th element is included in the sample} \\ &= 0 \text{ otherwise}\end{aligned}$$

For example, the mean of a sample of observations from a universe of  $N$  elements is written

$$\bar{x} = \frac{1}{n} \sum_{i=1}^N \delta_i X_i$$

Expectations of  $\bar{X}$  and of functions of it, such as its variance, are then derived by operating as if  $\delta_i$  were the random variable. Cornfield presented the scheme merely as a "device that has been found useful in the derivation of expected values and variances of statistics ... Its only advantage is that it reduces the manipulations ... to a simple algebraic routine." What is being done is to recognize that over all possible randomizations every  $X_i$  must appear an equal number of times, therefore that the expectation must be a function of the average  $X$  (or  $X^2$ , etc.) multiplied by some constant which can be obtained by averaging the constant coefficients  $\delta$  without bothering about individual  $X_i$  values. It merely lends an algebraic formulation to precisely the same argument as was used in section 8 to evaluate expectations of generalized product means. The random variable produced by the act of sampling is a variable vector of  $n$  elements  $(X_1 \dots X_n)$ . The formulation writes the function to be considered, for example the sum, as the matrix product of two vectors  $(\delta_1 \dots \delta_N) (X_1 \dots X_N)'$ . Then over all possible randomizations the vector of elements remains constant and we have to average only the simpler vector of zeros and ones.

Kempthorne (1952, sec. 8.2) has put this device to more extended usage. He postulates that a randomized block experiment is performed on a particular set of  $bt$  plots each with its own fixed error  $e_{ij}$ ,  $i = 1 \dots b$ ,  $j = 1 \dots t$ . He writes the linear model,

$$y_{ik} = \mu + b_i + t_j + \sum_j \delta_{ij}^k e_{ij} \quad (4)$$

where  $\delta_{ij}^k$  is equal to unity if treatment  $k$  occurs on plot  $j$  in the  $i$ th block and is zero otherwise." This is already more complicated than Cornfield's formulation because the sample (experiment) has  $bt$  of these vectors instead of only one; an extension which appears unavoidable in view of the more complex sampling pattern

and variety of quadratic forms in the analysis of variance.

However, more important in Kempthorne's use of the device is the theoretical role which he assigns to it. He continues: "The random error attached to any observed yield is the whole expression  $\sum_j \delta_{ij}^k e_{ij}$ . Any particular  $e_{ij}$  is a fixed variable which we do not know. The random variable in the expression (4) is the term  $\delta_{ij}^k \dots$ ". Wilk and Kempthorne (1955) extend the device to all elements of the general linear model, that for the two-factor experiment, which has been used as standard example in above sections, being written.

$$y_{i^*j^*f} = \mu + \sum_i \alpha_i^{i^*} a_i + \sum_j \beta_j^{j^*} b_j + \sum_{ij} \alpha_i^{i^*} \beta_j^{j^*} (ab)_{ij} + \sum_k \rho_k^{i^*j^*f} e_k$$

summations being over the whole of respective group universes and the greek letters being unity when relevant universe elements indicated by subscripts  $i, j, k$  are the same as those indicated by the sample subscripts  $i^*, j^*, f$ , and zero otherwise. On introducing this formulation they remark: "The quantities (greek letters) can be treated as random variables because random methods of selection and allocation are employed." But caution which that sentence might apply is dispersed by the next paragraph which continues: "Of course the random variables in the statistical model are the  $\alpha_i^{i^*}, \beta_j^{j^*}$  and  $\rho_k^{i^*j^*f}$ , which take on the values 0 and 1 with known probabilities. All other quantities in the model are fixed, unknown, parameters. The formulation appears to be an ingenious method of making the transfer from universe elements as fixed constants to their appearance as random variables in a sample. But if it were true that  $\alpha_i^{i^*}$  etc. are the random variables,  $a_i$  etc. are the constants, the expectations of quadratic forms should appear as variances of  $\alpha$  multiplied by functions of  $a_i$  as constants. Quite the opposite is produced like a conjuring trick whose deception the authors have apparently overlooked. The expectations which come out of the hat are variances of the universe elements with

the nominal variances and covariances of the  $\alpha$ 's as constants ! The definitions of "variances" used by the authors are the  $K_2$  parameters of the group universes (8.2); whether one chooses to regard these as proportionate to the second moments of the universes, or to variances of random variables produced by the act of sampling (sec. 7.15) is incidental. The switch of notation reveals that the only variables being studied are the  $a_i$ , the  $\alpha$ 's are dummy constants of the same nature as the  $x$ 's in the regression equation (5.1), and are no more random variables than the universe elements. What the model really postulates is A fixed vectors (0...1...0), the act of random sampling picks a of these vectors, which is identically the same as saying that it selects a of the  $a_i$ . The only thing that has been done is to write the element  $a_i$  as a matrix product (0...1...0)( $a_i \dots a_i \dots a_A$ )'. Presentation of the individual  $\alpha_i^{i*}$  as basic random variables independently of the  $a_i$  is a dialectical artifact which does nothing but confuse the issues. The only excuse for using them is as mathematical operators if they simplify algebra. This they do not do; they lead to very complex algebra, which can be evaded either by using Tukey's "brackets" or the canonical variance components  $K(\gamma)$  followed by substituting the simple relations between  $K(\gamma)$  and  $K(r)$ .

The last feature of WK's formulation to be discussed here is their insistence that potential interaction of treatments with experimental units should be explicitly recognized in the linear model. The problem seems to have been first stated by Neyman et al (1935). It has been treated by Anderson and Bancroft (1952, Chap. 23), in the slightly disguised form of discussing whether blocks should be treated as random or fixed effects.

WK illustrate on an example by Vaurio and Daniel (1954) which has also been discussed by Scheffé (1954). It concerns the effect of different methods of

annealing tinned coils. Material for observation is taken from certain prescribed locations (head, middle and tail) on each coil which are regarded as subsidiary treatments. The anneal treatments must be applied to whole coils, random samples of  $c$  coils being selected for each treatment. Comparisons between locations are within coils. A surprising feature of all three discussions of this experiment is that they all treat coils formally (like locations) as a treatment. Although surely they must have recognized the circumstance none of the three presents the experiment in the form which most clearly reveals its structure, as a split-plot experiment. The coils are main-plots on which anneal treatments are tested, segments of the coils are split-plots for evaluation of location effects. An agriculturist would write the model as

$$y_{ijk} = \mu + \alpha_i + \delta_{ij} + \lambda_k + (\alpha\lambda)_{ik} + \epsilon_{ijk}$$

$\delta_{ij}$  being the main-plot error and  $\epsilon_{ijk}$  being the split-plot error. (The main-plots are here completely randomized for treatments without blocking.) It is true that the theoretical effects  $(y_{ij} - y_{i'j})$  and  $(y_{ij} - y_{ij'})$  may differ (theoretical because two different anneals,  $i$  and  $i'$ , cannot be applied to the same coil), and we want this to be recognized by writing the main plot part of the model as  $\mu + \alpha_i + \gamma_{(i)j} + (\alpha\gamma)_{ij}$ . But once started on such refinements surely they ought also to recognize  $\delta_{ij}$  to represent all sources of random experimental variation affecting whole plots independently of the intrinsic nature of the coil itself? That they do not do so illustrates the impracticability of carrying models to the objective detail for which they ask and write as if they were achieving. Apart from their theoretical argument the point is trivial since  $\gamma_{(i)j}$  and  $\delta_{ij}$  are completely confounded. For our present purpose we can ignore the split-plots which are irrelevant to main treatment comparisons, the mean of split-plot elements

over variation of k becoming part of  $\delta_{ij}$  to give the main plot model

$$y_{ij} = m + a_i + c_{(i)j} + (ac)_{ij} + d_{ij}$$

which I write with roman letters to indicate the specific latent equation. After the preceding exposition I dispense with the greek letter dummy variables of WK's formulation.

Treating the coils as experimental units,  $c_{(i)j} + d_{ij}$  is identical with the  $e_k$ ,  $(ac)_{ij}$  with the  $n_{ijk}$ , of the earlier part of WK's paper. Following the first part of this section it would be superfluous in practice to consider the coils as anything except a sample from a large population of coils. But, to follow out the theory, suppose that a finite universe of coils is being sampled. Only the regression view of anneal effects is of interest so a is read as A without thereby implying that we are saying anything about a potential universe of anneal treatments.

The analysis of variance is

	d.f.	M.Sq.	E(M.Sq.)			
			$K_\delta$	$K_\alpha$	$K_\gamma$	$K_\lambda$
Between anneals	(A - 1)	$\langle 2 \rangle - \langle 11 \rangle + c(\langle 11 \rangle - \langle \frac{1}{.} \frac{1}{.} \rangle)$	1	1	1	c
Coils within anneals	A(c - 1)	$\langle 2 \rangle - \langle 11 \rangle$	1	1	1	

Formulation of the mean squares in terms of brackets exhibits both the sample values and their expectations. The coefficients of the standard components,  $K(\gamma)$ , are obvious, or follow from Crump's rules remembering that coil effects include (AC). Alternatively they follow from their definitions in terms of brackets (8.13) after expanding

$$\langle 2 \rangle - \langle \frac{1}{.} \frac{1}{.} \rangle = (\langle 2 \rangle - \langle \frac{1}{.} \frac{1}{.} \rangle - \langle \frac{1}{.} \frac{1}{.} \rangle + \langle \frac{1}{.} \frac{1}{.} \rangle) + (\langle \frac{1}{.} \frac{1}{.} \rangle - \langle \frac{1}{.} \frac{1}{.} \rangle).$$

The specific components (8.2) follow from the usual relations

$$K_a = K_\alpha + K_\lambda / C, \quad K_c = K_\gamma + K_\lambda / A.$$

Alternatively one can consider the theoretical A x C matrices for each group of elements, with the usual restrictions, to obtain:

for  $a_i$ , constant in rows,

$$\langle 2 \rangle = \langle \begin{matrix} 11 \\ \cdot\cdot \end{matrix} \rangle$$

$$\langle \begin{matrix} 1\cdot \\ \cdot 1 \end{matrix} \rangle = \langle \begin{matrix} 1\cdot \\ 1\cdot \end{matrix} \rangle = - \langle 2 \rangle / (A-1)$$

For  $c_j$ , constant in columns,

$$\langle 2 \rangle = \langle \begin{matrix} 1\cdot \\ 1\cdot \end{matrix} \rangle$$

$$\langle \begin{matrix} 1\cdot \\ \cdot 1 \end{matrix} \rangle = \langle \begin{matrix} 11 \\ \cdot\cdot \end{matrix} \rangle = - \langle 2 \rangle / (C-1)$$

for  $(ac)_{ij}$ ,

$$\langle \begin{matrix} 11 \\ \cdot\cdot \end{matrix} \rangle = - \langle 2 \rangle / (C-1)$$

$$\langle \begin{matrix} 1\cdot \\ 1\cdot \end{matrix} \rangle = - \langle 2 \rangle / (A-1)$$

$$\langle \begin{matrix} 1\cdot \\ \cdot 1 \end{matrix} \rangle = \langle 2 \rangle / (A-1)(C-1)$$

the  $\langle \begin{matrix} 1\cdot \\ \cdot 1 \end{matrix} \rangle$  values being however here redundant since the sample contains only one element per column and does not evaluate these products. Either substitution leads at once to

	$K_d$	$K_{ac}$	$K_c$	$K_a$
Anneals	1	$(1 - \frac{1}{A} - \frac{C}{C})$	1	C
Coils	1	$(1 - 1/A)$	1	

which are the first two rows of WK's table 4, except that they have illogically redefined  $K_{ac}$  as  $\sum_{i,j}^{AC} (ac)_{ij}^2 / A(C-1)$ , a form which implies absence of the restriction  $\sum_{i=0}^A (ac)_{i=0}$ , which is none the less imagined, whereas the above table uses the denominator  $(A-1)(C-1)$  in accordance with (8.2).

Is this formulation in terms of hypothetical specific components of any use? One result is that Wilk (discussion at the Montreal meeting) claimed that estimates of variance components are always biased by treatment-unit interaction. But that deduction is seen to depend merely on the definition of what is to be estimated. The canonical components are, as always, unbiased. So that detail is easily bypassed by those who will agree that the  $k(\gamma)$  are anyway more convenient statistics.

Evidently  $K_c$  and  $K_{ac}$  cannot be separated and  $K_a$  cannot be estimated. But these sub-divisions are academic. In this example a variance component between treatments is irrelevant because we are concerned only with treatment means. But even suppose we were concerned with a universe of treatments, still there can be no universe for which  $K_a$  is a variance component. Distinction of  $K_c$  and  $K_{ac}$  is equally an artifact. The practical situation is that coils annealed in one way will have a certain mean and distribution; annealed in another way another mean and distribution. In the models under discussion the distributions are supposed to be similar and only the means have to be differentiated. The worth to a factory of a difference of means depends on their separation relative to the total variability of coils with a given treatment, measured by  $K_\gamma + K_{\alpha\gamma}$ . (We omit  $K_\delta$  because, subject to an irreducible processing variability which is inseparable from  $K_\gamma$  and must be included therewith, it can be freely modified according to number of specimens observed from each coil.) Similarly a test of significance between two experimentally observed means depends on the total variance within treatments, the variance of a difference between two treatment means is  $\frac{2}{c} (K_\delta + K_{\alpha\gamma} + K_\gamma)$ , independently of any  $C$ .

The finite population correction factors appearing with  $K_{ac}$  in the analysis of variance are therefore irrelevant for any practical consideration. This does

not say that an interaction should be ignored just because it cannot be estimated; only that it is of no interest when it is an integral part of the only variability which does interest us. Interactions of this sort exist merely as part of a formal analysis (sec. 1). I assert that interactions of treatments with unidentifiable random variables have no meaning, not even an academic one. Behind the WK argument lies the idea of a pure scientific investigation as to how a treatment operates. But there cannot be any "how" attached to a non-repeatable random effect. Scientific investigation depends on being able to repeat something definable. An interaction can have meaning only between identifiable characters. Suppose anneal effect is affected by manganese content: we can investigate the anneal and manganese interaction either by introducing different manganese contents as a deliberate treatment or by a covariance study on manganese contents as they appear by chance. Either way the manganese content gives an identifiable link between one coil and another; and the interaction has meaning. But so long as there is no such identifiable link the interaction coil x treatment is a non-reproducible random effect and a formal abstraction which can be nothing more than an unidentifiable part of coil variability.

In a similar manner Anderson and Bancroft (Chap. 8) consider block x treatment interaction composed of both error and "real" interaction, and what should be the error variance for treatment effects according as blocks are regarded as "random" or "fixed". Block x treatment interaction is measurable by adding replication within blocks; but it is never identifiable in the above sense. The fact that internal replication is almost never used is prima facie evidence that experimenters instinctively feel little interest in the matter however it may be debated theoretically. A block is defined as a particular group of experimental units as used

on a particular occasion. It is an experimental unit standing in the same relation to ultimate units as main-plots to split-plots. Hence block x treatment interaction is the same as treatment x experimental unit interaction merely pushed back to a larger unit. We may be able to repeat some characteristic nominally used in forming blocks, but the whole complex of a given block as experimental unit is not repeatable. Fertility of a lump of soil changes from year to year, different plants are growing on it, and are affected by weather conditions and their incidence with plant phases. Even if we could be interested in particular areas of land, identification of location x treatment interaction needs replication in years; it is not synonymous with block x treatment interaction in one year. If blocks for a nutrition experiment are formed by grouping animals according to age, evidence of varying response to age may be worth looking for, but needs to be evaluated against treatment x block interaction for blocks of similar age. Comparison to variation of duplicates within an oven batch may show treatment x batch interaction, but to link it to a treatment x oven interaction we need replication of batches. Any character of experimental interest occurring between blocks in effect gives the experiment a split-plot design, and analysis proceeds accordingly. A block in its general sense being unrepeatable, nominal interaction with it is an unreproducible chance effect. I therefore assert that unidentified treatment x block interaction is an integral part of experimental error.

Note in passing that the inferred population is not one of similar blocks, characterized by the mean and variance of their absolute yields. It is a population having similar mean and variance of differences (cf. variances (11.4) and (11.6)). With further experimentation we may be able to identify parts of treatment x block interaction - by location, by weather, by age, by oven, etc. - but

this is rather beside the point. On the above definition of a block we can never identify the whole of a block x treatment interaction. Being unreproducible we are led to describe blocks always as "random", never as "fixed" effects.

When the population sampled cannot be rigorously defined this point of view will not satisfy those who repudiate inference to an undefined population. It seems however to be a lesser evil than retreating to the extent of regarding inference as being only for the observed set of experimental units.

## REFERENCES

- Anderson, R. L. and Bancroft, T. A. 1952 Statistical theory in research. McGraw-Hill: New York
- Bennett, C. A. and Franklin, N. L. 1954 Statistical analysis in chemistry and the chemical industry. Wiley: New York.
- Cochran, W. G. 1937 Problems arising in the analysis of a series of similar experiments. J. Roy. Stat. Soc. Suppl. 4:102-118.
- Cochran, W. G. 1939 The use of analysis of variance in enumeration of sampling. J. Am. Stat. Assoc. 34:492-510.
- Cochran, W. G. 1953 Sample survey techniques. Wiley: New York.
- Cornfield, J. 1944 On samples from finite populations. J. Am. Stat. Assoc. 39:236-239.
- Cramer, H. 1946 Mathematical methods of statistics. Princeton Univ. Press.
- Crump, S. L. 1946 The estimation of variance components in analysis of variance. Biometrics Bull. 2:7-11.
- Crump, S. L. 1951 The present status of variance component analysis. Biometrics 7:1-16
- Daniels, H. E. 1939 The estimation of components of variance. J. Roy. Stat. Soc. Suppl. 6:186-197.
- David, F. N. and Kendall, M. G. 1949-53 Tables of symmetric functions. Biometrika 36:431-449; 38:435-462; 40:427-446.
- Deming, W. E. 1950 Some theory of sampling. Wiley: New York.
- Deming, W. E. and Stephan, F. F. 1941 On the interpretation of censuses as samples. J. Am. Stat. Assoc. 36:45-49.
- Eisenhart, G. 1947 The assumptions underlying the analysis of variance. Biometrics 3:1-21.
- Feller, W. 1950 An introduction to probability theory and its application. Wiley: New York.
- Fisher, R. A. 1925 Statistical methods for research workers. Oliver and Boyd: Edinburgh.
- Fisher, R. A. 1935 Design of experiments. Oliver and Boyd: Edinburgh.
- Hald, A. 1952 Statistical theory with engineering applications. Wiley: New York.
- Hanson, M. H., Hurwitz, W. N., and Madow, W. G. 1953 Sample survey methods. II. Theory. Wiley: New York.
- Hendricks, W. A. 1948 Mathematics of sampling. Va. Ag. Expt. Sta.: Special Tech. Bull.
- Hendricks, W. A. 1951 Variance components as a tool for the analysis of sample data. Biometrics 7:97-101.

- Hoel, P. G. 1954 Introduction to mathematical statistics. 2nd Ed. Wiley: New York.
- Irwin, J. and Kendall, M. G. 1944 Sampling moments of moments for a finite population. *Ann. Eng.* 12:135-142.
- Johnson, N. L. 1948 Alternative systems in the analysis of variance. *Biometrika* 35:80-87.
- Kaplan, E. L. 1952 Tensor notation and the sampling cumulants of k-statistics. *Biom.* 39:319-323.
- Kempthorne, O. 1952 The design and analysis of experiments. Wiley: New York.
- Kendall, M. G. 1943 The advanced theory of statistics. Vol. I. Griffin. London.
- Kendall, M. G. 1949 On the reconciliation of theories of probability. *Biom.* 36: 101-116.
- Kendall, M. G. 1951 Regression, structure and functional relationship. *Biom.* 38: 11-25.
- Kendall, M. G. and Sundrum, R. M. 1953 Distribution-free methods and order properties. *Rev. Intern. Stat. Instit.* 3:124-134.
- Lindley, D. V. 1953 Statistical inference. *J. Roy. Stat. Soc. B* 15:30-65-76.
- Mentzer, E. G. 1953 Tests by the analysis of variance. Wright Air Development Center Tech. Report 53-23.
- Mood, A. M. 1950 Introduction to the theory of statistics. McGraw-Hill, New York.
- Neyman, J. 1935 Statistical problems in agricultural experimentation. *J. Roy. Stat. Soc. Suppl.* 2:108-144-154-180.
- Neyman, J. and Scott, E. 1948 Consistent estimates based on partially consistent observations. *Econometrica* 16:1-32.
- Pearson, E. S. 1937 Some aspects of the problem of randomization. *Biom.* 29: 53-64.
- Pitman, E. S. G. 1938 Significance tests which may be applied to samples from any populations. III The analysis of variance test. *Biom.* 29:322-335.
- Scheffe, H. 1954 Statistical methods for evaluation of several sets of constants and sources of variability. *Chem. Eng. Progress* 50:200-205.
- Smith, H. F. 1951 The analysis of variance with unequal but proportionate numbers of observations in the sub-classes of a two-way classification. *Biometrics* 7:70-74.
- Tukey, J. W. 1949 Dyadic anova, an analysis of variance for vectors. *Human Biology* 21:65-110.
- Tukey, J. W. 1950 Some sampling simplified. *J. Am. Stat. Assoc.* 45:501-519.
- Wilk, M. B. 1955. The randomization analysis of a generalized randomized block design. *Biom.* 42:70-79.

- Vaurio, V. W. and Daniel, C. 1954 Evaluation of several sets of constants and several sources of variability. Chem. Eng. Progress, 50:81-86.
- Weloh, B. L. 1937 On the  $\chi^2$ -test in randomized blocks and latin squares. Biom. 29:21-52.
- Wishart, J. 1952 Moment coefficients of the k-statistics in samples from a finite population. Biom. 39:1-13.
- Yates, F. 1935 Complex experiments. J. Royl Stat. Soc. Suppl. 2:181-247.
- Yates, F. 1937 The design and analysis of factorial experiments. Imp. Bur. Soil Sc. Tech. Comm. 35.
- Yates, F. A review of recent statistical developments in sampling and sampling surveys. J. Roy. Stat. Soc. 109:12-30-43.
- Yates, F. and Zaccopani, I. 1935 The estimation of the efficiency of sampling, with special reference to sampling for yield in cereal experiments. J. Ag. Sci. 25:545-577.
- Yule, G. U. and Kendall, M. G. 1950 An introduction to the theory of statistics. 11th. Ed. Griffin. London.

#### UNPUBLISHED REPORTS

- Hooke, R. 1953 Sampling from a matrix, with applications to the theory of testing. Statistical Research Group, Princeton Univ., Memo. Rept. 53
- Hooke, R. 1954 Moments of moments in matrix sampling: An extension of polykays. ibid. 55.
- Hooke, R. 1954 The estimation of polykays in the analysis of variance. ibid. 56.
- Tukey, J. W. 1949a Interaction in a row-by-column design. ibid. 19.
- Tukey, J. W. 1950a Finite sampling simplified. ibid. 45.
- Wilk, M. B. and Kempthorne, O. 1955 The logical derivation of linear models and their use in selecting the appropriate error term in the analysis of variance. IV Fixed, mixed and random models.