VARIANCES OF VARIANCE COMPONENTS: I. BALANCED DESIGNS[1]

By John W. Tukey

Princeton University

1. Summary. Analyses of variance are sometimes intended to reveal information about means (when tests of significance and, better, confidence procedures are appropriate). At other times analyses of variance have the purpose indicated by their name: to estimate the sizes of the various components contributed to the over-all variance from the corresponding sources. If we make certain assumptions of independence and normality for all of the quantities involved, it is easy to obtain formulas for the variances of the natural estimates of these variance components. The utility of these estimates can be called in question on the grounds of three sorts of assumptions: of certain amounts of independence, of infinite populations, of normality of distribution. This paper treats of the case where the latter two of these assumptions are removed, leaving only the customary (and dangerous) independence assumptions (as do the next two papers in this series).

The treatment makes intensive use of polykays (which were introduced in [1], although that name was not used, and discussed in [2]) and is applied specifically to balanced single and double classifications, to Latin squares, and to balanced incomplete blocks. A general definition of balance for an analysis of variance situation is given, and the general application of the technique to balanced situations is set forth. An application to a less simple example of a balanced single classification concludes the paper.

2. Introduction to polykays. In order to deal easily and effectively with problems involving random samples from finite populations, the writer emphasized in [1] certain homogeneous polynomial symmetric functions of a finite set of numbers. These are of two sorts: (i) the brackets or symmetric means, exemplified by

$$\langle 12 \rangle = \frac{\sum^{\neq} x_i x_j^2}{n(n-1)},$$

where the summation is over the $n(n-1)$ pairs (i, j), with $i \neq j$; and (ii) the parentheses or polykays, exemplified by

$$(12) = k_{12} = k_1 k_2 - \frac{1}{n} k_3 = \langle 1 \rangle \langle 2 \rangle - \langle 111 \rangle.$$

Each set can be expressed linearly in terms of the other with constant coefficients. (Elsewhere, [1], [2], we use (12) as an alternate to k_{12}, but in the present

Received April 6, 1955.

[1] Prepared in connection with research sponsored by the Office of Naval Research and based on part of Memorandum Report 45, "Finite Sampling Simplified," Statistical Research Group, Princeton University, which was written while the author was a Fellow of the John Simon Guggenhein Foundation.

paper it will be simpler not to use this alternative notation.) The facts about polykays which we shall need are the following:

(A) Any polynomial symmetric function of a finite set of numbers can be expressed linearly in the polykays of that set.

(B) The average, over all random samples drawn from a finite population of numbers, of any polynomial in the values of the sample can be expressed linearly in the polykays of the population with coefficients which do not involve the size of the population.

(C) If adding a constant to all the numbers of the set in (A) or the finite population in (B) leaves the polynomial invariant, then the coefficients of all polykays with one or more subscripts "1" vanish.

(D) Any polynomial function of several finite sets of numbers which is symmetric in each of the sets separately can be expressed in terms of products of polykays from the various sets, the polykays of each set entering, at most, linearly.

(E) The average, over all sets of random samples from the respective finite populations of numbers, of any polynomial in the values of these samples can be expressed in terms of products of polykays from the various sets, the polykays of each set entering at most linearly, with coefficients which do not involve the sizes of the populations.

(F) If adding a constant to all the numbers, of one set in (D) or of one finite population in (E), leaves the polynomial invariant, then the coefficients of all products involving a polykay of the corresponding set, or population, which has one or more subscripts "1" vanish.

(G) The following formula holds:

$$[k_2]^2 = k_{22} + \frac{1}{n} k_4 + \frac{2}{n-1} k_{22},$$

where n is the size of the set, or population, for which k_2, k_{22}, k_4 are some of the polykays.

(H) For a set made up of $n - 1$ zeros and one (nonzero) value, t, all brackets and polykays with more than one index vanish, and the rest are given by:

$$k_p = \langle p \rangle = \frac{t^p}{n}.$$

The proofs of most of these statements can be easily disposed of by simple argument or by reference.

Thus, (A) implies (B), and (D) implies (E), because the average of a polynomial over all random samples is a symmetric polynomial in the values of the finite population. Every symmetric polynomial can be written linearly in terms of symmetric means, and every symmetric mean can be written linearly in terms of polykays (the actual formulas for degree $\leqq 4$, the highest with which we shall be concerned here, are given in [2]), so that (A) holds. A similar argument disposes of (D). The argument establishing (C) and (F) is given in Section 11. (G) appears in [1], page 516. And, finally, if only one value is nonzero, all symmetric means with two or more indices vanish, and since the expression of a polykay in terms of symmetric means involves only symmetric means with at least as many indices (see [2]; the actual formulas for degree $\leqq 4$ also appear in [1]), the same is true of polykays. The values of the one-index brackets and the polykays then follow by direct calculation.

3. The variance of a sample variance. If $x_1, x_2, \cdots, x_n$ are a sample of n, their variance

$$s^2 = \frac{1}{n-1}\sum(x_i - \bar{x})^2$$

is one of the most familiar statistics. Its variance in sampling, first from infinite and later from finite populations, has been derived by many writers. A derivation using polykays is presented in [1], page 517. In principle, analogous processes can be used for the variances of more complex expressions, but the algebra can be avoided by taking another path. We illustrate this path now for the variance of the variance.

We deal, then, with a sample of size n and polykays k_1, k_{11}, k_2 from a population of size N and polykays k_1', k_{11}', k_2'. We have $s^2 \equiv k_2$ and

$$\text{var } s^2 \equiv \text{var } k_2 \equiv \text{ave } k_2^2 - (\text{ave } k_2)^2, \tag*{1}$$

where "ave" and "var" refer to average values and variances for all samples of n from the population. Now ave k_2^2 is a homogeneous polynomial of degree 4. Moreover, adding a constant to all x's leaves k_2, and also k_2^2, invariant. According to (C), therefore, we can express ave k_2^2 linearly in terms of population polykays which do not involve any index "1", and the coefficients will not involve N. Hence, ave k_2^2 is of the form

$$\phi_1(n)k_4' + \psi_1(n)k_{22}',$$

while $(\text{ave }\{k_2\})^2 = (k_2')^2$ is of the form

$$\left(\phi_2(n) + \frac{1}{N}\right)k_4' + \left(\psi_2(n) + \frac{2}{N-1}\right)k_{22}'$$

(where actually $\phi_2(n) = 0, \psi_2(n) = 1$). Hence,

$$\text{var }(k_2) = \left(\phi(n) - \frac{1}{N}\right)k_4' + \left(\psi(n) - \frac{2}{N-1}\right)k_{22}',$$

where $\phi(n) = \phi_1(n) - \phi_2(n), \psi(n) = \psi_1(n) - \psi_2(n)$.

Now consider the case $n = N$, where the sample consists of the whole population, and k_2 is constant. We have

$$0 = \left(\phi(n) - \frac{1}{n}\right)k_4' + \left(\psi(n) - \frac{2}{n-1}\right)k_{22}',$$

and since k_4' and k_{22}' do not satisfy any linear identity, we must have

$$\phi(n) - \frac{1}{n} = 0, \qquad \psi(n) - \frac{2}{n-1} = 0,$$

so that the variance of k_2 is

$$\text{var}(k_2) = \left(\frac{1}{n} - \frac{1}{N}\right) k_4' + \left(\frac{2}{n-1} - \frac{2}{N-1}\right) k_{22}',$$

as before.

This method of evaluating both "finite population corrections" and some of the other coefficients will extend easily to the standard analysis of variance situations, as we shall see. To it we shall need to add the use of *minimal unit populations*, by which we mean finite populations whose values are all zero except for one value of unity and whose size is as small as possible for the situation considered. This population has vanishing polykays except for

$$k_p' = 1/n, \qquad p = 1, 2, \cdots .$$

This is the opposite of the infinite normal population for which

$$k_2' = \sigma^2, \qquad k_4' \equiv 0, \qquad k_{22}' = \sigma^4, \qquad \cdots .$$

4. The balanced single classification. We now tackle the simplest model which we know how to specify for the balanced single classification, namely, the model with two finite populations:

$$x_{ij} = \mu + \eta_i + \omega_{ij}, \qquad i = 1, 2, \cdots, c, \qquad j = 1, 2, \cdots, r;$$

$$\{\eta_i\} \text{ sampled from } n, k_1, k_{11}, \cdots,$$

$$\{\omega_{ij}\} \text{ sampled from } N, K_1, K_{11}, \cdots;$$

sampling independent, order randomized.

(We shall omit the primes from both sets of the polykays for simplicity and convenience.)

Let A be an estimate of the "between" variance k_2, which is a homogeneous quadratic function of the x_{ij} and is unbiased in mean. Let B be an estimate of the "within" variance K_2 with similar properties. Then, we may conclude that

$$\text{var}(A), \qquad \text{cov}(A, B), \qquad \text{var}(B),$$

all have the form

$$\alpha k_4 + \beta k_{22} + \gamma k_2 K_2 + \delta K_4 + \epsilon K_{22} .$$

Our task is to determine the three sets of α, β, γ, δ, and ϵ.

Just as in the second method of treating the variance of the variance, the sizes of populations can only enter through the correction terms arising from

$$k_2^2 = k_{22} + \frac{1}{n} k_4 + \frac{2}{n-1} k_{22},$$

$$k_2 K_2 = k_2 K_2,$$

$$K_2^2 = K_{22} + \frac{1}{N} K_4 + \frac{2}{N-1} K_{22}$$

In particular, the terms in n and N are the same for any quadratic unbiased estimates.

But let us take the usual estimates obtained from an analysis of variance of A and B. Then, we know that the η_i vanish identically from B, and hence that k_4, k_{22}, and k_2 cannot appear in var (B) or cov (A, B). Thus, we have

$$\begin{aligned}
\text{var }(A) &= \left(\alpha_1 - \frac{1}{n}\right) k_4 + \left(\beta_1 - \frac{2}{n-1}\right) k_{22} + \gamma_1 k_2 K_2 + \delta_1 K_4 + \epsilon_1 K_{22}, \\
\text{cov }(A, B) &= \delta_2 K_4 + \epsilon_2 K_{22}, \\
\text{var }(B) &= \left(\delta_3 - \frac{1}{N}\right) K_4 + \left(\epsilon_3 - \frac{2}{N-1}\right) K_{22},
\end{aligned}$$

where α_1, β_1, $\cdots$, ϵ_3 are independent of n and N.

Now take $\omega_{ij} = 0$, that is, take $K_4 = K_{22} = K_2 = 0$ and $n = c$, so that every η_i is always used. Then A is constant, and we see that $\alpha_1 = 1/c$, $\beta_1 = 2/(c - 1)$.

Start again, take the η's $= 0$ and take a minimal unit population (of size rc) for the ω's. Then, one and only one x will be unity, the others will be zero. A and B will be constant, but $K_4 \neq 0$ and k_4, k_{22}, k_2, K_{22} all vanish. Hence,

$$\delta_1 = \delta_2 = \delta_3 - \frac{1}{rc} = 0.$$

We have now reduced our variances to the form

$$\begin{aligned}
\text{var }(A) &= \left(\frac{1}{c} - \frac{1}{n}\right) k_4 + \left(\frac{2}{c-1} - \frac{2}{n-1}\right) k_{22} + \gamma_1 k_2 K_2 + \epsilon_1 K_{22}, \\
\text{cov }(A, B) &= \epsilon_2 K_{22}, \\
\text{var }(B) &= \left(\frac{1}{rc} - \frac{1}{N}\right) K_4 + \left(\epsilon_3 - \frac{2}{N-1}\right) K_{22}.
\end{aligned}$$

We have four more coefficients to determine. We could find them in two steps by (i) introducing minimal unit populations for both η's and ω's, considering the two cases which arise, and finding γ_1; and then (ii) letting the η's vanish and introducing a minimal population with *two* nonzero elements for the ω's so as to determine the remaining coefficients. It seems simplest, however, to fall back on normal theory.

It is well known that when η's and ω's are drawn from (infinite) normal populations, the means squares are distributed like multiples of chi-square. Hence, we have

	Mean	Variance	Covariance
Between.....	$K_2 + rk_2$	$2(K_2 + rk_2)^2 \,/\, (c - 1)$	0
Within.......	K_2	$2K_2^2 \,/\, c(r - 1)$	

The within component is the within mean square, and for infinite populations, $K_2^2 = K_{22}$, so that we have found that

$$\epsilon_3 = \frac{2}{c(r-1)}.$$

Observing that the between variance component is

$$\frac{1}{r}\,(\text{MS between} - \text{MS within}),$$

we find the covariance between the two components to be

$$-\frac{1}{r}\operatorname{var}\{\text{MS within}\} = -\frac{2}{cr(r-1)}K_2^2,$$

so that

$$\epsilon_2 = \frac{2}{cr(r-1)},$$

and the variance of the between component to be

$$\frac{1}{r^2}\frac{2(K_2 + rk_2)^2}{c-1} + \frac{2K_2^2}{c(r-1)} = \frac{2}{c-1}k_2^2 + \frac{4}{r(c-1)}k_2K_2 + \frac{2}{r^2}\left(\frac{1}{c-1} + \frac{1}{c(r-1)}\right)K_2^2,$$

so that, again, since $K_2^2 = K_{22}$ for the normal distribution,

$$\beta_1 = \frac{2}{c-1},\ \gamma_1 = \frac{4}{r(c-1)},\ \epsilon_1 = \frac{2}{r^2}\left(\frac{1}{c-1} + \frac{1}{c(r-1)}\right) = \frac{2(rc-1)}{cr^2(c-2)(r-1)},$$

Our final results, then, are

$$\operatorname{var}(\text{between}) = \left(\frac{1}{c} - \frac{1}{n}\right)k_4 + \left(\frac{2}{c-1} - \frac{2}{n-1}\right)k_{22} + \frac{4}{r(c-1)}k_2K_2 + \frac{2(rc-1)}{r^2c(r-1)(c-1)}K_{22},$$

$$\operatorname{cov} = -\frac{2}{rc(r-1)}K_{22},$$

$$\operatorname{var}(\text{within}) = \left(\frac{1}{rc} - \frac{1}{N}\right)K_4 + \left(\frac{2}{c(r-1)} - \frac{2}{N-1}\right)K_{22}.$$

These are reasonably simple formulas and are entirely free of assumptions of normality of distribution and infinity of population. They retain, however, an assumption of an independence character, namely, that the η's and ω's are independently drawn and allotted.

Clearly, variances and covariances of components for at least certain balanced designs can most easily be found by combining (1) the finite population terms, (2) the effects of minimal unit populations taken one at a time, (3) normal theory. We shall do this in a few cases.

5. Simplest double classification. We now go on to the simplest row-by-column model, one without explicitly identified interaction, where

$$x_{ij} = \mu + \eta_i + \xi_j + \omega_{ij}, \qquad i = 1, 2, \cdots, c, \qquad j = 1, 2, \cdots, r;$$

$$\eta_i \text{ from } n, k_1, k_{11}, \cdots,$$

$$\xi_j \text{ from } n^*, K_1^*, k_{11}^*, \cdots,$$

$$\omega_{ij} \text{ from } N, K_1, K_{11}, \cdots;$$

independently and randomly sampled and arranged.

We have components for columns (associated with the η's), rows (associated with the ξ's), and residual (associated with the ω's; also called "interaction," "discrepance," or "error"). Making use of the same principle for allotting finite population terms and zero coefficients, we may write down the general structure of their variances and covariances in the following forms:

$$\text{var}\{\text{cols}\} = \left(\alpha_1 - \frac{1}{n}\right) k_4 + \left(\beta_1 - \frac{2}{n-1}\right) k_{22} + \gamma_1 k_2 K_2 + \delta_1 K_4 + \epsilon_1 K_{22},$$

$$\text{var}\{\text{rows}\} = \left(\alpha_2 - \frac{1}{n'}\right) k_4^* + \left(\beta_2 - \frac{2}{n'-1}\right) k_{22}^* + \gamma_2 k_2^* K_2 + \delta_2 K_4 + \epsilon_2 K_{22},$$

$$\text{cov}\{\text{rows, cols}\} = \delta_3 K_4 + \epsilon_3 K_{22},$$

$$\text{var}\{\text{res}\} = \left(\delta_4 - \frac{1}{N}\right) K_4 + \left(\epsilon_4 - \frac{2}{N-1}\right) K_{22},$$

$$\text{cov}\{\text{cols, res}\} = \delta_5 K_4 + \epsilon_5 K_{22},$$

$$\text{cov}\{\text{rows, res}\} = \delta_6 K_4 + \epsilon_6 K_{22}.$$

If any one population is a minimum unit population (that is, if $n = c$, $n' = r$, or $N = rc$) and the others are constant, then all parts of the analysis of variance, and hence all of the estimates of variance components, are easily seen to be constant under randomization. Hence, we see that

$$\alpha_1 = 1/c, \qquad \alpha_2 = 1/r,$$

$$\delta_1 = \delta_2 = \delta_3 = \delta_4 - 1/N = \delta_5 = \delta_6 = 0.$$

Now, taking normal theory, we have for the mean squares

	Mean	Variance	Covariances •
MS {cols}....	$K_2 + rk_2$	$2(K_2 + rk_2)^2 \,/\, (c - 1)$	0
MS {rows}...	$K_2 + ck_2^*$	$2(K_2 + ck_2^*)^2 \,/\, (r - 1)$	0
MS {res}.....	K_2	$2K_2^2 \,/\, (r - 1)(c - 1)$	0

From this, we easily derive the corresponding table for estimates of variance components:

Component	Mean	Variance	Covariances
col.....	k_2	$\frac{1}{r^2}\left(\frac{2(K_2 + rk_2)^2}{(c - 1)} + \frac{2K_2^2}{(r - 1)(c - 1)}\right)$	
			$\frac{1}{rc}\frac{2K_2^2}{(r - 1)(c - 1)}$
rows...	k_2^*	$\frac{1}{c^2}\left(\frac{2(K_2 + ck_2^*)^2}{(r - 1)} + \frac{2K_2^2}{(r - 1)(c - 1)}\right)$	
			$-\frac{1}{r}\frac{2K_2^2}{(r - 1)(c - 1)}$
			$-\frac{1}{c}\frac{2K_2^2}{(r - 1)(c - 1)}$
res.....	K_2	$2K_2^2/(r - 1)(c - 1)$	

whence we see that

$$\beta_1 = \frac{2}{c - 1}, \qquad \beta_2 = \frac{2}{r - 1},$$

$$\gamma_1 = \frac{4}{r(c - 1)}, \qquad \gamma_2 = \frac{4}{c(r - 1)},$$

$$\epsilon_1 = \frac{2}{r^2}\left(\frac{1}{c - 1} + \frac{1}{(r - 1)(c - 1)}\right) = \frac{2}{r(r - 1)(c - 1)},$$

$$\epsilon_2 = \frac{2}{c^2}\left(\frac{1}{r - 1} + \frac{1}{(r - 1)(c - 1)}\right) = \frac{2}{c(r - 1)(c - 1)},$$

$$\epsilon_3 = \frac{2}{c(c - 1)r(r - 1)},$$

$$\epsilon_4 = \frac{2}{(r - 1)(c - 1)},$$

$$\epsilon_5 = \frac{-2}{r(r - 1)(c - 1)},$$

$$\epsilon_6 = \frac{-2}{c(r - 1)(c - 1)},$$

thus completing the calculation. The final answers are, then,

$$\begin{aligned}
\operatorname{var}\{\text{cols}\} &= \left(\frac{1}{c} - \frac{1}{n}\right) k_4 + \left(\frac{2}{c-1} - \frac{2}{n-1}\right) k_{22} + \frac{4}{(c-1)r}\, k_2 K_2 \\
&\qquad + \frac{2}{(c-1)r(r-1)}\, K_{22};
\end{aligned}$$

$$\begin{aligned}
\operatorname{var}\{\text{rows}\} &= \left(\frac{1}{r} - \frac{1}{n}\right) k_4^* + \left(\frac{2}{r-1} - \frac{2}{n'-1}\right) k_{22}^* + \frac{4}{c(r-1)}\, k_2^* K_2 \\
&\qquad + \frac{2}{c(c-1)(r-1)}\, K_{22},
\end{aligned}$$

$$\operatorname{cov}\{\text{cols, rows}\} = \frac{2}{c(c-1)r(r-1)}\, K_{22},$$

$$\operatorname{var}\{\text{res}\} = \left(\frac{1}{cr} - \frac{1}{N}\right) K_4 + \left(\frac{2}{(c-1)(r-1)} - \frac{2}{N-1}\right) K_{22},$$

$$\operatorname{cov}\{\text{cols, res}\} = -\frac{2}{(c-1)r(r-1)}\, K_{22},$$

$$\operatorname{cov}\{\text{rows, res}\} = -\frac{2}{c(c-1)(r-1)}\, K_{22}.$$

6. The Latin square. The next example in order of complexity is the Latin square, for whose side we use k to avoid confusion with n. Since rows, columns, and treatments enter symmetrically in a fully randomized model of the sort we are discussing, we need not treat them separately.

We can write down the formulas almost at once by analogy with those just given. The main effect considered (columns, rows, or treatments) is sampled from n, k_1, k_{11}, $\cdots$, while the cell contribution is sampled from N, K_1, K_{11}, $\cdots$. Then,

$$\begin{aligned}
\operatorname{var}\{\text{main}\} &= \left(\frac{1}{k} - \frac{1}{n}\right) k_4 + \left(\frac{2}{k-1} - \frac{2}{n-1}\right) k_{22} + \frac{4}{k(k-1)}\, k_2 K_2 \\
&\qquad + \frac{1}{k^2(k-2)}\, K_{22},
\end{aligned}$$

$$\operatorname{var}\{\text{res}\} = \left(\frac{1}{k^2} - \frac{1}{N}\right) K_4 + \left(\frac{2}{(k-1)(k-2)} - \frac{2}{N-1}\right) K_{22},$$

$$\operatorname{cov}\{\text{main, main}^*\} = \frac{2}{k^2(k-1)(k-2)}\, K_{22},$$

$$\operatorname{cov}\{\text{main, res}\} = -\frac{2}{k(k-1)(k-2)}\, K_{22}.$$

7. Balanced incomplete blocks. A moment's computation shows that a single minimal unit population, with the others constant, still leads to constant analyses of variance. Hence, the terms in K_4 still vanish, while those in k_4 and k_4^* have their usual simple form.

If we have

b blocks, v varieties, r replications,

and hence

vr/b plots per block,

and if the

varieties are from n, k_1, k_{11}, $\cdots$,

blocks are from n^*, k_1^*, k_{11}^*, $\cdots$, and

fluctuations are from N, K_1, K_{11}, $\cdots$,

and if we recall that the analysis of variance runs

	DF	AvMS
Varieties................	$v - 1$	$K_2 + rk_2$
Blocks..................	$b - 1$	$K_2 + \frac{vr}{b}k_2^*$
Residue.................	$vr - v - b + 2$	K_2

(AvMS = average mean square), then we see that

$$\text{var}\,\{\text{vars}\} = \left(\frac{1}{v} - \frac{1}{n}\right) k_4 + \left(\frac{2}{v-1} + \frac{2}{n-1}\right) k_{22} + \frac{4}{r(v-1)}\, k_2 K_2 + \frac{2(vr - b + 1)}{r^2(v-1)(vr - v - b + 2)}\, K_{22},$$

$$\text{var}\,\{\text{blocks}\} = \left(\frac{1}{b} - \frac{1}{n^*}\right) k_4^* + \left(\frac{2}{b-1} - \frac{2}{n^*-1}\right) k_{22}^* + \frac{4b}{vr(b-1)}\, k_2^* K_2 + \frac{2(vr - v + 1)b^2}{v^2r^2(b-1)(vr - b - v + 2)}\, K_{22},$$

$$\text{var}\,\{\text{res}\} = \left(\frac{1}{vr} - \frac{1}{N}\right) K_4 + \left(\frac{2}{vr - v - b + 2} - \frac{2}{N-1}\right) K_{22},$$

$$\text{cov}\,\{\text{vars, blocks}\} = \frac{2b}{vr^2(vr - v - b + 2)}\, K_{22},$$

$$\text{cov}\,\{\text{vars, res}\} = -\frac{2}{r(vr - v - b + 2)}\, K_{22},$$

$$\text{cov}\,\{\text{blocks, res}\} = -\frac{2b}{vr(vr - v - b + 2)}\, K_{22}.$$

8. The general notion of balance. We are familiar with the idea of an analysis of variance separated into "lines" such that there are one or more kinds of contribution peculiar to each line. The notion of a line A falling "below" a line B is clearly understood by most expert practitioners, although it is not often discussed in print. For our present purposes, it will suffice to say that A is *below* B if the variance component corresponding to A appears in the average value of the mean square for B with a nonzero coefficient, but the converse is not true. (Besides "A below B" and "B below A", we could have "A beside B," when neither variance component appears in the other AvMS, or we could have "A intertwined with B," when both variance components appear in the other AvMS.) We shall be dealing both with individual lines and with groups of lines, which by convenient analogy we call *paragraphs*.

In any specific analysis of variance which does not involve intertwined lines, if we fix our attention on a specific line, we can divide all the lines into three paragraphs:

(a) The upper paragraph, containing lines above and beside the chosen line,

(b) The chosen line,

(c) The lower paragraph, containing lines below the chosen line.

Some of these paragraphs may be empty. This division is based on average values of mean squares. (The implied inequalities need not carry over completely to individual values of mean squares.) But stronger conditions may hold, as in the examples discussed above. In particular,

(1) An arbitrary change in the contributions associated with the given line may have no effect on the mean squares in the upper paragraph in each and every particular analysis of variance.

(2) If all the contributions associated with the given line vanish except for one, and the contributions from the upper paragraph all vanish, then the mean squares in the upper paragraph may not depend on the location of the one nonzero contribution from the given line.

Furthermore, if both of these contingencies occur, we shall say that the analysis of variance is *balanced* with respect to the given line. An analysis of variance balanced with respect to all of its lines is *balanced*. (Since the analyses of variance usually called "balanced" possess these properties, this definition is an extension of previous usage.)

9. Balanced cases in general. We can easily write down the variances and covariances of any variance component in such an analysis. It will involve three types of terms:

(1) Terms in the k_4 and k_{22} of the corresponding contributions,

(2) Cross-terms of the form k_2K_2, where K_2 refers to some line from the lower paragraph,

(3) Terms in the K_{22}'s for these lower lines.

(There will be no terms in the K_4's for lower lines because of condition (2) of the last Section applied to these lines.) Suppose there were no errors, no lower contributions, then we should have measured the upper contributions of interest exactly, and must face a variance of the form

$$\left(\frac{1}{b}-\frac{1}{n}\right)k_4+\left(\frac{2}{b-1}-\frac{2}{n-1}\right)k_{22},$$

where we have investigated b out of n cases. But these must be just the terms of the first type, since the others vanish with the errors and lower contributions. They will always be easy to write down.

The other terms are those which we found from normal theory. Let us illustrate in an imaginary example. Let us suppose that some design leads to an analysis of variance of the following shape:

Item	DF	AvMS
a	3	$\sigma^2+3\sigma_1^2+7\sigma_2^2$
b	7	$\sigma^2+4\sigma_1^2$
c	8	σ^2

Clearly, the mean-square component estimating σ_2^2 must be

$$\frac{4\text{MS}_a-3\text{MS}_b-\text{MS}_c}{28},$$

and its normal theory variance is to be found from the chi-squared variance of

$$2\,\frac{(\text{average})^2}{\text{degrees of freedom}}$$

applied to each mean square, which yields

$$2\,\frac{4^2}{28^2}\,\frac{(\sigma^2+3\sigma_1^2+7\sigma_2^2)^2}{3}+\frac{3^2}{28^2}\,\frac{(\sigma^2+4\sigma_1^2)^2}{7}+\frac{1^2}{28^2}\,\frac{(\sigma^2)^2}{8}$$

$$=\frac{1}{392}\left\{6.7444\,\sigma^4+42.29\,\sigma^2\sigma_1^2+7.467\sigma^2\sigma_2^2+68.57\sigma_1^4+224\sigma_1^2\sigma_2^2+261.3\sigma_2^4\right\}.$$

These, then, are the terms of types (2) and (3), which can be easily written down in almost any balanced case.

As an illustration of these principles, we shall write down the variance of a main-effect variance component in a three-way (balanced) analysis with replication, the shape of which is

	DF	AvMS
Treatments	$t-1$	$\sigma^2+p\sigma_{crt}^2+cp\sigma_{tr}^2+rp\sigma_{ct}^2+crp\sigma_t^2$
Rows	$r-1$	$\sigma^2+p\sigma_{crt}^2+cp\sigma_{tr}^2+tp\sigma_{cr}^2+ctp\sigma_r^2$
T × R	$(t-1)(r-1)$	$\sigma^2+p\sigma_{crt}^2+cp\sigma_{rt}^2$
T × C	$(t-1)(c-1)$	$\sigma^2+p\sigma_{crt}^2+rp\sigma_{ct}^2$
R × C	$(r-1)(c-1)$	$\sigma^2+p\sigma_{crt}^2+tp\sigma_{cr}^2$
T × R × C	$(r-1)(c-1)(t-1)$	$\sigma^2+p\sigma_{crt}^2$
Replication	$rct(p-1)$	σ^2

Clearly, σ_t^2 is estimated from

$$\frac{\text{MS}T - \text{MS}(T \times R) - \text{MS}(T \times C) + \text{MS}(T \times R \times C)}{crp},$$

and hence the variance of estimate is

$$\left(\frac{1}{t} - \frac{1}{n}\right) k_4 + \left(\frac{2}{t-1} - \frac{2}{n-1}\right) k_{22} + \frac{4\sigma_t^2\sigma_{ct}^2}{c(t-1)} + \frac{4\sigma_t^2\sigma_{rt}^2}{r(t-1)} + \frac{4\sigma_t^2\sigma_{rct}^2}{rc(t-1)}$$
$$+ \frac{4\sigma_t^2\sigma^2}{rcp(t-1)} + \frac{2\sigma_{ct}^4}{c(c-1)(t-1)} + \frac{2\sigma_{rt}^4}{r(r-1)(t-1)} + \frac{2\sigma_{ct}^2\sigma_{rt}^2}{r^2c^2(t-1)}$$
$$+ \frac{4\sigma_{ct}^2\sigma_{crt}^2}{c(c-1)(t-1)r} + \frac{4\sigma_{ct}^2\sigma^2}{c(c-1)(t-1)rp} + \frac{4\sigma_{rt}^2\sigma_{crt}^2}{r(r-1)(t-1)c}$$
$$+ \frac{4\sigma_{rt}^2\sigma^2}{r(r-1)(t-1)cp} + \frac{2\sigma_{crt}^4}{c(c-1)r(r-1)(t-1)}$$
$$+ \frac{4\sigma_{crt}^2\sigma^2}{c(c-1)r(r-1)(t-1)p} + \frac{2\sigma^4}{c(c-1)r(r-1)(t-1)p^2},$$

where σ_{ct}^4, σ_{rt}^4, σ_{crt}^4, and σ^4 are understood to stand for the appropriate polykays (with subscripts "22") in the event any of these populations are finite.

10. A further example. Let us show that the generality of the concept of balance is rather wider than one might suppose at first glance. Consider the case of a single classification with equal numbers of cases in each column, with each column subject to different fluctuations. The formal model runs as follows:

$$x_{ij} = \mu + \eta_j + \omega_{ij}, \qquad i = 1, 2, \cdots, c, \qquad j = 1, 2, \cdots, r;$$

$\{\eta_j\}$ sampled from $n, k_1, k_{11}, \cdots,$

$\{\omega_{ij}\}$, for each i, sampled from $N_i, K_{1,i}, K_{11,i}, \cdots;$

independently and randomly sampled and arranged.

It is easy to verify that the conventional analysis, which does not take account of the differences in fluctuation between columns, is balanced and that the lines, degrees of freedom, and average mean squares are (writing σ_η^2 for k_2 and σ_i^2 for $K_{2,i}$) as follows:

Line	DF	AvMS
Between	$c - 1$	$\frac{1}{c}(\sigma_1^2 + \sigma_2^2 + \cdots + \sigma_c^2) + \sigma_\eta^2$
Within Col 1	$r - 1$	σ_1^2
Within Col 2	$r - 1$	σ_2^2
....	...	...
Within Col c	$r - 1$	σ_c^2

The estimate of σ_η^2 will be found from

$$\frac{(\text{MS between}) - \frac{1}{c}\sum(\text{MS } \mathit{within}_i)}{r}$$

and the corresponding variance is given by

$$\left(\frac{1}{c} - \frac{1}{n}\right) k_4 + \left(\frac{2}{c-1} - \frac{2}{n-1}\right) k_{22} + \frac{4}{r(c-1)c}\, k_2 \sum_i K_{2,i}$$
$$+ \frac{2}{(c-1)c^2r^2} \sum_{i \neq j} K_{2,i}K_{2,j} + \left(\frac{2}{(c-1)r^2c^2} + \frac{2}{c^2(r-1)r^2}\right) \sum_i K_{22,i}\,.$$

The special case with $K_{2,i} = K_2$, $K_{22,i} = K_{22}$, $N_i = N$ leads to

$$\left(\frac{1}{c} - \frac{1}{n}\right) k_4 + \left(\frac{2}{c-1} - \frac{2}{n-1}\right) k_{22} + \frac{4}{r(c-1)}\, k_2 K_2 + \frac{4}{cr^2}\,(K_2)^2$$
$$+ \frac{2}{cr^2}\left(\frac{1}{c-1} + \frac{1}{r-1}\right) K_{22}\,,$$

and when we use

$$K_2^2 = K_{22} + \frac{1}{N}\, K_4 + \frac{2}{N-1}\, K_{22}\,,$$

this becomes

$$\left[\left(\frac{1}{c} - \frac{1}{n}\right) k_4 + \left(\frac{2}{c-1} - \frac{2}{n-1}\right) k_{22} + \frac{4}{r(c-1)}\, k_2 K_2\right.$$
$$\left. + \frac{2(rc-1)}{(c-1)c(r-1)r^2}\, K_{22}\right] + \left[\frac{2}{cr^2}\left(\frac{1}{N}\, K_4 + \frac{2}{N-1}\, K_{22}\right)\right],$$

where the first bracket is the same as for the simple single classification model and the second bracket expresses the result of separating the fluctuations into the separate columns.

11. Proof of disappearance. It was asserted in (C) of Section 2 that if a polynomial is invariant under translation, its expansion will not involve any polykays with unit parts (indices "1"). We now proceed to give a proof.

Altering a finite set of numbers by translation through δ is equivalent to randomly pairing them with a set of numbers all of which equal δ. The polykays for this new set are

$$k_1' = \delta, \qquad k_{11}' = \delta^2, \qquad k_{111}' = \delta^3, \cdots,$$

and all others are zero. In accordance with the pairing rules (see [2], [1]), the effect of translation is then to alter the polykays of the original set as follows:

$$\begin{aligned}
&k_1 \rightarrow k_1 + \delta\\
&k_{11} \rightarrow k_{11} + 2\delta k_1 + \delta^2,\\
&k_2 \rightarrow k_2\,,\\
&k_{111} \rightarrow k_{111} + 3\delta k_{11} + 3\delta^2 k_1 + \delta^3,\\
&k_{12} \rightarrow k_{12} + \delta k_2\,,\\
&k_3 \rightarrow k_3\,.
\end{aligned}$$

Now if the invariant polynomial is

$$c_1 k_1 + c_{11} k_{11} + c_2 k_2 + c_{111} k_{111} + c_{12} k_{12} + c_3 k_3 + \cdots$$

before translation, it is increased by

$$\begin{aligned}
&c_1\delta + c_{11}(2\delta k_1 + \delta^2) + c_{111}(3\delta k_{11} + 3\delta^2 k_1 + \delta^3) + c_{12}\delta k_2 + \cdots\\
&\qquad = \delta(c_1 + 2c_{11}k_1 + 3c_{111}k_{11} + c_{12}k_2 + \cdots) + \delta^2(c_{11} + 3c_{111}k_1 + \cdots),
\end{aligned}$$

and since this must vanish for all δ, we have

$$\begin{aligned}
c_1 + 2c_{11}k_1 + 3c_{111}k_{11} + c_{12}k_2 + \cdots &= 0,\\
c_{11} + 3c_{111}k_1 + \cdots &= 0,
\end{aligned}$$

and so on. Now, the original finite set was at our disposal, and if we multiply each element in it by ϵ, we shall multiply each of its polykays by ϵ raised to the degree of the polykay. The last set of equations become

$$\begin{aligned}
c_1 + \epsilon(2k_1c_{11}) + \epsilon^2(3c_{111}k_{11} + c_{12}k_2) + \epsilon^3(4c_{1111}k_{111} + 2c_{112}k_{12} + c_{13}k_3) + \cdots &= 0,\\
c_{11} + \epsilon(c_{111}k_1) + \epsilon^2(6c_{1111}k_{11} + c_{112}k_2) + \cdots &= 0;\\
\cdots \qquad\qquad \cdots \qquad\qquad & \cdots
\end{aligned}$$

whence, comparing coefficients along a triangular path, we deduce the vanishing of c_1, c_{11}, c_{111}, c_{12}, c_{1111}, c_{112}, c_{13}, and so on. Hence we have the result stated.

REFERENCES

[1] John W. Tukey, "Some sampling simplified," *J. Amer. Stat. Assoc.*, Vol. 45 (1950), pp. 501–519.

[2] John W. Tukey, "Keeping moment-like sampling computations simple," *Ann. Math. Stat.*, Vol. 27 (1956), pp. 37–54.