SOME STATISTICAL ASPECTS OF PARTITIONING GENOTYPE-ENVIRONMENTAL COMPONENTS OF VARIABILITY

G. K. SHUKLA
A.R.C. Unit of Statistics, 21 Buccleuch Place, University of Edinburgh, Edinburgh EH8 9LN

Received 17.xii.71

1. Introduction

Recently, Freeman and Perkins (1971) examined some of the existing methods of partitioning genotype-environmental component of variability and their statistical validity. They have considered the usual practice of calculating the regression of genotype means on the environmental means calculated by taking the average of all genotypes in that environment, first used by Yates and Cochran (1938) and later used by Finlay and Wilkinson (1963) and since then used by several other authors of which references have been cited by Freeman and Perkins (1971). They have shown the statistical invalidity of using such regressions and their sums of squares for testing homogeneity. They have further suggested that some genotypes should be taken in each environment (not included in calculating the mean of each genotype) as a measure of environment and regression of genotype means should be calculated on the independent measure of environment thus making the procedure statistically more valid.

In the first part of this paper we have shown how by looking at the model in a different way one could draw statistically valid conclusions of certain hypotheses from the same analysis of Yates and Cochran (1938). In the second part of this paper we have suggested a method of estimating a component of genotype-environmental interaction corresponding to each genotype, thus giving a better measure of genotype stability. This paper is concerned with the presentation of practical methods rather than with statistical theory (to which references are given); it does not itself contain much that is new, but a more general statistical treatment is being prepared and will be published elsewhere.

2. Regression of genotype means on environmental means

We have used mostly the same model and notation as used by Perkins and Jinks (1968) for t genotypes, s environments and r replications of each genotype within each environment and the model could be represented as,

$$y_{ijk} = \mu + d_i + \epsilon_j + g_{ij} + e_{ijk}, \tag{1}$$

where μ is the grand mean, $d_i (i = 1, \ldots, t)$ the additive genetic contribution of the ith genotype, $\epsilon_j (j = 1, \ldots, s)$ the additive environmental contribution of the jth environment, g_{ij} the genotype-environment interaction of the ith genotype in the jth environment and $e_{ijk} (k = 1, \ldots, r)$ is the residual variation contributed by the kth replicate of the ith genotype in the jth environment. We shall assume that

$$E(\bar{e}_{ij}) = 0; \quad V(\bar{e}_{ij}) = \sigma_0^2, \tag{2}$$

for all i and j where $\bar{e}_{ij}$ is the mean of e_{ijk} over r replications and σ_0^2 is the within environment error variance for the mean of r replications. We shall further assume that environment effects ϵ_j's are random effects with population mean zero and variance σ_ϵ^2. A random sample of s environments has been selected from an infinite population of environments. We shall estimate σ_0^2 as usual by $\hat{\sigma}_0^2$, where

$$\hat{\sigma}_0^2 = \sum_i \sum_j \sum_k (y_{ijk} - \bar{y}_{ij})^2 / str(r-1), \tag{3}$$

with $st(r-1)$ degrees of freedom. Hereafter we shall work with $\bar{y}_{ij}$ (the mean of r replications of the ith genotype at the jth environment). In the present paper we shall confine ourselves to the genotype-environment $(G \times E)$ part of the analysis of variance.

Working with the means, the model in (1) can be written

$$\bar{y}_{ij} = \mu + d_i + \epsilon_j + g_{ij} + \bar{e}_{ij}. \tag{4}$$

Putting

$$g_{ij} = b_i' \epsilon_j + \eta_{ij} \quad \text{and} \quad \alpha_{ij} = \eta_{ij} + \bar{e}_{ij}$$

we obtain from (4)

$$\bar{y}_{ij} = \mu + d_i + \epsilon_j + b_i' \epsilon_j + \alpha_{ij}. \tag{5}$$

Putting $b_i = b_i' - \bar{b}'$ where $\bar{b}' = \sum_i b_i'/t$ we obtain from (5)

$$\bar{y}_{ij} = \mu + d_i + \epsilon_j(1 + \bar{b}') + b_i \epsilon_j + \alpha_{ij}. \tag{6}$$

The model in (5) is reparameterised in (6) in such a way that $\sum_i b_i = 0$. When all b_i's are 0 (or b_i' are equal) then the model in (6) becomes as in (7),

$$\bar{y}_{ij} = \mu + d_i + \epsilon_j(1 + b') + \alpha_{ij}. \tag{7}$$

Thus the problem of testing the equality of all b_i''s becomes the problem of testing model (7) against model (6). This is equivalent to testing the presence of the non-additivity term $b_i \epsilon_j$ in (6) when ϵ_j are taken as fixed effects. The test for presence of non-additivity of this type was given by Mandel (1961), which is a generalisation of Tukey (1949). Same results hold good even if ϵ_j are taken as random effects. He has estimated b_i by $\hat{b}_i$ and calculated the sum of squares due to non-additivity (S) as follows:

$$\hat{b}_i = \frac{\sum_j (\bar{y}_{ij} - \bar{y}_{.j})(\bar{y}_{.j} - \bar{y}_{..})}{\sum_j (\bar{y}_{.j} - \bar{y}_{..})^2},$$

$$S = \sum_i \hat{b}_i^2 \sum_j (\bar{y}_{.j} - \bar{y}_{..})^2, \tag{8}$$

and " Balance " = Interaction Sum of Square $(G \times E) - S$.

The sum of squares in (8) is similar to that calculated by Yates and Cochran (1938) and the same as that due to heterogeneity in regression in Freeman and Perkins (1971) table 2, with z_j in Freeman and Perkins' notations replaced by $\bar{y}_{.j} - \bar{y}_{..}$. The estimated regression coefficients, $\hat{b}_i$, here are similar to those given by Yates and Cochran (1938) except that we have regressed deviation of genotype means from the environmental

means rather than genotype means. The sum of squares S indeed is a ratio of quartic terms and quadratic terms in the y's.

In the absence of interaction ($g_{ij} = 0$ for all i and j) it can be shown that S/σ_0^2 and " Balance "$/\sigma_0^2$ are both independently distributed as χ^2 on $(t-1)$ and $(t-1)(s-2)$ degrees of freedom, respectively. However, in the presence of interaction S/σ_0^2 and " Balance "$/\sigma_0^2$ are not distributed as χ^2 even if all $b_i = 0$, though they are independently distributed of each other. In the presence of interaction the appropriate test statistic for all $b_i = 0$ will be F' as given in (9).

$$F' = \frac{S/(t-1)}{\text{" Balance "}/(t-1)(s-2)} \tag{9}$$

F' will be distributed as F on $(t-1)$ and $(t-1)(s-2)$ degrees of freedom. The same test was proposed by Perkins and Jinks (1968) and this statistic gives correct probability level. Equality of any two b_i's could be tested by doing the similar analysis for any two genotypes of interest. The above argument can be generalised for other arrangements of genotypes within and between environments and also in the presence of non-orthogonality in the data (Milliken and Graybill, 1970).

To use the $\hat{b}_i$'s in the usual sense of regression would not be valid, but they give rough guidance about the relation of genotype means to environmental means. The $\hat{b}_i$'s are biased estimators of the b_i's. In general this bias will be small when σ_e^2 is large but could be corrected in the way suggested by Tai (1971). The partition of sum of squares into two components is also only approximate, but this may be quite satisfactory for practical purposes. More efficient estimates of b_i's and ϵ_j's, and their sum of squares could be obtained by fitting the model in (6) by a non-linear least squares method as suggested by Elston (1961) and Tai (1971) and approximate tests could be obtained. Under these circumstances, independent measure of environment based on more genotype means may not be worthwhile.

3. Components of interaction sum of squares

The characterisation of genotypes on the basis of regression coefficients may not be very effective when only a small fraction of the interaction sum of squares ($G \times E$) can be attributed to heterogeneity among the regressions. It might be then of great interest to partition $G \times E$ into t components, one corresponding to each genotype, as mentioned by Baker (1969). Put

$$g_{ij} + \bar{e}_{ij} = v_{ij}.$$

Let us further assume that

$$E(v_{ij}) = 0;\ V(v_{ij}) = \sigma_i^2;\ E(v_{ij}, v_{i'j'}) = 0 \text{ for } i \neq i' \text{ or } j \neq j';$$
$$V(g_{ij}) = \sigma_i'^2;\ E(g_{ij}\bar{e}_{ij}) = 0;\ i = 1, \ldots, t. \tag{10}$$

Then,

$$\sigma_i^2 = \sigma_i'^2 + \sigma_0^2.$$

In the above expression σ_i^2 could be taken as the sum of two components, *viz.* within environmental variance (σ_0^2) and between environmental variance ($\sigma_i'^2$) of the ith genotype (after correcting for additive common effect of

environment ϵ_j), and we shall name it the " stability variance " of the ith genotype. We shall call a genotype stable if its stability variance (σ_i^2) is equal to within environmental variance (σ_0^2) which means that $\sigma_i'^2 = 0$. Relatively large values of σ_i^2 will indicate more instability of genotype.

Estimation of σ_i^2 is analogous to the problem of estimating heterogeneous variances in a two-way classification when they change in one way considered by Ehrenberg (1950) and later by Russell and Bradley (1958). Rao (1970) has further generalised the above prcoedure for any classification and also considered some optimum properties of the above estimators. Without going into detail, we give the unbiased estimators of σ_i^2, denoted by $\hat{\sigma}_i^2$, as

$$\hat{\sigma}_i^2 = \frac{1}{(s-1)(t-1)(t-2)} [t(t-1) \sum_j (\bar{y}_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}_{..})^2 - \sum_i \sum_j (y_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}_{..})^2]$$

$$= \frac{1}{(s-1)(t-1)(t-2)} [t(t-1) \sum_j (u_{ij} - \bar{u}_{i.})^2 - \sum_i \sum_j (u_{ij} - \bar{u}_{i.})^2] \qquad (11)$$

where

$$u_{ij} = \bar{y}_{ij} - \bar{y}_{.j} \text{ and } \bar{u}_{i.} = \sum_j u_{ij}/s.$$

They are obtained as linear combinations of squares of residuals

$$(\bar{y}_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}_{..}),$$

therefore, they are independent of μ, d_i and variance of ϵ_j. It is not difficult to verify that they are unbiased estimators of σ_i^2. Under the assumption of symmetrical distribution of σ_i^2's, Rao (1970) has proved that on average they have minimum variance among all possible quadratic unbiased estimators (MINQUE) of σ_i^2. It is not difficult to see that their mean is the same as the mean sum of squares ($G \times E$). Therefore, by multiplying each $\hat{\sigma}_i^2$ by $(t-1)(s-1)/t$ we shall obtain t components of $G \times E$, one corresponding to each genotype. These components are not statistically independent; as they are differences of two sums of squares, they can be negative, but negative estimates of variances are not uncommon in variance components problems.

The essential difference between the present approach and Baker's (1969) approach is this that his method estimates $((t-2)\sigma_i^2 + \bar{\sigma}^2)/t$ while the above method estimates σ_i^2 where

$$\bar{\sigma}^2 = \frac{1}{t} \sum_{i=1}^{t} \sigma_i^2.$$

The same is true for deviation from regression component.

The variance of $\hat{\sigma}_i^2$ is not only a function of σ_i^2 but of variances of other genotypes $\sigma_j^2 (j \neq i)$ taken in trial. Such estimators are only available when $t \geqq 3$. The problem of testing homogeneity of σ_i^2's has been considered by Russell and Bradley (1958), Johnson (1962), Han (1969) and Shukla (1971). The method proposed by Shukla could be easily applied for testing the homogeneity of all the variances or any pair of them.

It might be of some interest to test whether certain genotypes are stable or not. Johnson (1962) suggested a test criterion based on the ratio $\hat{\sigma}_i^2/\hat{\sigma}_0^2$.

It is difficult to derive the exact distribution of $\hat{\sigma}_i^2$, but when t is large the variance of $\hat{\sigma}_i^2$ is approximated as

$$V(\hat{\sigma}_i^2) \cong 2\sigma_i^4/(s-1). \tag{12}$$

The above expression helps us in obtaining an approximate distribution. When t is large, $(s-1)\hat{\sigma}_i^2/\sigma_i^2$ will be approximately distributed as χ^2 on $(s-1)$ degrees of freedom; thus, under the hypothesis that $\sigma_i'^2 = 0$, F^* will have an approximate F distribution on $(s-1)$ and $st(r-1)$ degrees of freedom where

$$F^* = \hat{\sigma}_i^2/\hat{\sigma}_0^2. \tag{13}$$

When $\hat{\sigma}_i^2$ is negative or less than $\hat{\sigma}_0^2$ then $\sigma_i'^2$ may be taken equal to zero as usual.

4. Further extension of model

For further progress in the interpretation of instability, we shall reconsider the model in (5). To keep the treatment general, we replace ϵ_j by z_j in the non-additive term $b_i'\epsilon_j$ and rewrite the model as

$$y_{ij} = \mu + d_i + \epsilon_j' + b_i z_j + \alpha_{ij}; \tag{14}$$

where z_j is a measure of some characteristic of jth environment; by taking deviation from the mean we can make $\sum_j z_j = 0$, and $\epsilon_j' = \epsilon_j + \bar{b}' z_j$.

We shall further assume that

$$V(\alpha_{ij}) = s_i^2; \quad (i = 1, \ldots, t),$$

and then discuss the method of estimation of s_i^2. The usual estimator of b_i, by the method of unweighted least squares, can be obtained as

$$\hat{b}_i = \sum_j \frac{(u_{ij} - \bar{u}_{i.})z_j}{\sum_j z_j^2}. \tag{15}$$

Using methods as in Section 3, unbiased estimators of s_i^2 for extended model in (14) could be obtained as $\hat{s}_i^2$:

$$\hat{s}_i^2 = \frac{t}{(t-2)(s-2)}\left[S_i - \sum_i \frac{S_i}{t(t-1)}\right] \tag{16}$$

where

$$S_i = \sum_{j=1}^{s} (u_{ij} - \bar{u}_{i.} - \hat{b}_i z_j)^2.$$

It is apparent that the model in (14) is just the extension of the model in (7) to take into account a covariate z_j. The estimators obtained in (16) are quadratic (in y's) estimators of s_i^2 and have the properties of MINQUE estimators. When t is large, the variance of $\hat{s}_i^2$ can be approximated by

$$V(\hat{s}_i^2) = \frac{2s_i^4}{(s-2)} \tag{17}$$

and their distribution can be approximated by χ^2 on $(s-2)$ degrees of freedom. An approximate significance test against $\hat{\sigma}_0^2$ is possible as in (13). If some of the genotypes become stable after taking the covariate into account, it may be inferred that the instability was introduced by the linear effect of the covariate and such information may be useful. The above approach could also be extended to more than one covariate.

5. Relationship between regression approach and the "stability variance" approach

The definition of stability is similar to Baker (1969) and Eberhart and Russell (1966). A significant departure of the regression of a genotype from zero will be indicated by a relatively high "stability variance", but a regression coefficient of zero need not mean that the particular genotype is stable. A zero regression will be obtained if there is no linear relationship between genotype mean and environmental mean, yet the "stability variance" (σ_i^2) may be greater than σ_0^2.

Once some of the genotypes are found unstable, it may be of interest to examine further the reasons for instability. The approach of Section 4 may be followed if observations are available on covariates which are likely to affect the genotypes differentially. We can examine the effect on stability variance of linear regression on environmental means by the method in the previous section. To examine any effect of differential fertility we have used $z_j = \bar{y}_{.j} - \bar{y}_{..}$ as used by many other authors mentioned in Sections 1 and 2. It must be noted here that the estimators of s_i^2 obtained by putting $z_j = \bar{y}_{.j} - \bar{y}_{..}$ in (16) will not be quadratic estimators of y's and therefore the optimum properties described in Section 4 may not hold. Again as mentioned earlier the effect of departure from optimality may be small when σ_e^2 is large. The effect of such differential regression on the stability of genotypes could be tested as above under the assumption that z_j's are constant. The estimation of individual s_i^2 is analogous to what Perkins and Jinks (1968) have suggested by the mean sum of squares $\sum_j \delta_{ij}^2/(s-2)$ and Baker (1969) by deviation from regression sum of squares but the above approach has an advantage as they are unbiased estimates of s_i^2 (free from any other nuisance parameters) and the mean of $\hat{s}_i^2$ is the same as the mean sum of squares of departure from regressions ("Balance") and this could be taken as equivalent to dividing the "Balance" into components corresponding to each genotype.

Recently Tai (1971) has worked with the above problem. The difference between our method and his method is this, that he has considered the model under certain side conditions on the interaction and we have not imposed any such conditions on them. It would not be very justifiable to impose any condition on interaction while estimating the individual component. The definition of stability is also different. According to his definition of stability one should have $b_i' = -1$ and $s_i^2 = \sigma_0^2$. Our definition of stability coincides with his definition of average stability ($\alpha_i = 0$; $\lambda_i = 1$ in Tai, 1971, notations). By our definition of stability we only mean that the performance of a genotype is sum of additive genotypic effect, additive environmental effect and a random error without any interaction between genotype and environment.

Prediction of the expected performance can also be made with reasonable accuracy for a given environment, if either the interaction is not present or most of it can be accounted by linear regression term (Jinks and Perkins, 1970).

6. Numerical example

For illustration purposes we have considered the data analysed by Yates and Cochran (1938). We shall only consider the part of the table dealing with $G \times E$.

Table 1

Variety × Place totals over the years

	Places						
Varieties	1	2	3	4	5	6	Total
Manchusia	161·7	247·0	185·4	218·7	165·3	154·6	1132·7
Svansota	187·7	257·5	182·4	183·3	138·9	143·8	1093·6
Velvet	200·1	262·9	194·9	220·2	165·8	146·3	1190·2
Tribi	196·9	339·2	271·2	266·3	151·2	193·6	1418·4
Peatland	182·5	253·8	219·2	200·5	184·4	190·1	1230·5
Total	928·9	1360·4	1053·1	1089·0	805·6	828·4	6065·4

Table 2 gives the values of u_{ij} obtained from table 1.

Table 2

u_{ij}'s and regression coefficients

	Places						
Varieties	1	2	3	4	5	6	$\hat{b}_i$
Manchusia	−24·08	−25·08	−25·22	0·90	4·18	−11·08	−0·156
Svansota	1·92	−14·58	−28·22	−34·50	−22·22	−21·88	−0·014
Velvet	14·32	−9·18	−15·72	2·40	4·68	−19·38	−0·054
Tribi	11·12	67·12	60·58	48·50	−9·92	27·92	0·609
Peatland	−3·28	−18·28	8·58	−17·30	23·28	24·42	−0·385

Table 3

$u_{ij} - \hat{b}_i z_j$

	Places					
Varieties	1	2	3	4	5	6
Manchusia	−26·14	−14·18	−23·90	3·34	−2·23	−16·77
Svansota	1·69	−13·60	−28·10	−34·28	−22·79	−22·39
Velvet	13·43	−5·41	−15·26	3·24	2·46	−21·35
Tribi	21·11	24·55	55·44	38·99	15·08	50·17
Peatland	−9·59	8·63	11·83	−11·29	7·47	10·37

To obtain the component of variances on the same unit as the sum of squares (units of single plot) in the Analysis of Variance (table 4) we have divided them by 6.

Comparison of $\hat{\sigma}_i^2$'s with $\hat{\sigma}_0^2$ shows that Tribi and Peatland are unstable. Further regression analysis shows that Tribi remains unstable, though, its variability has reduced considerably. Peatland becomes stable after taking covariate into consideration. Similar conclusions were drawn by Yates and

Cochran (1938) but the above type of analysis in general may be advantageous.

TABLE 4

Analysis of variance table and " stability variances " (units of single plot)

Source	D.F.	S.S.	M.S.	F
Places	5	7072·92	—	—
Varieties	4	1770·28	—	—
$V \times P$	20	1477·84	73·89	—
$\hat{\sigma}_1^2$	—	—	25·88	1·11
$\hat{\sigma}_2^2$	—	—	19·60	0·84
$\hat{\sigma}_3^2$	—	—	22·73	0·98
$\hat{\sigma}_4^2$	—	—	225·53	9·69**
$\hat{\sigma}_5^2$	—	—	75·68	3·25**
Source	D.F.	S.S.	M.S.	F
Heterogeneity	4	773·16	193·29	—
Balance	16	704·69	44·04	—
$\hat{s}_1^2$	—	—	34·10	1·46
$\hat{s}_2^2$	—	—	40·48	1·74
$\hat{s}_3^2$	—	—	42·78	1·84
$\hat{s}_4^2$	—	—	79·70	3·42*
$\hat{s}_5^2$	—	—	23·27	1·00
$\hat{\sigma}_0^2$	216	—	23·28	—

7. SUMMARY

1. The usual regression approach of explaining genotype-environment interaction has been considered by using a non-additive model and the statistical validity of the analysis has been discussed.

2. Alternative approach of dividing genotype-environmental interaction into components, one corresponding to each genotype has been proposed and the optimum properties have been discussed.

3. The alternative approach has been extended to take into account a covariate.

4. The relationship of new approach to the regression approach has been discussed.

5. A numerical example has been given as an illustration.

Acknowledgment.—I am very grateful to Professor D. J. Finney and Mr H. D. Patterson for their valuable suggestions and help. I am grateful to referees for their valuable suggestions.

8. REFERENCES

BAKER, R. J. 1969. Genotype-environment interaction in yield of wheat. *Can. J. Plant Sci.*, *49*, 743-751.

EBERHART, S. A., AND RUSSELL, W. L. 1966. Stability parameters for comparing varieties. *Crop Sci.*, *6*, 36-40.

EHRENBERG, A. S. C. 1950. The unbiased estimation of heterogeneous error variances. *Biometrika, 37,* 347-357.

ELSTON, R. C. 1961. On additivity in the analysis of variance. *Biometrics, 17,* 209-219.

FINLAY, K. W., AND WILKINSON, G. N. 1963. The analysis of adaption in a plant breeding programme. *Aust. J. Agric. Res., 14,* 742-754.

FREEMAN, G. H., AND PERKINS, JEAN M. 1971. Environmental and genotype-environmental components of variability. VIII. Relations between genotypes grown in different environments and measure of these environments. *Heredity, 26,* 15-23.

JINKS, J. L., AND PERKINS, JEAN, M. 1970. Environmental and genotype-environmental components of variability. VII. Simultaneous prediction across environments and generations. *Heredity, 25,* 475-480.

HAN, C. P. 1969. Testing the homogeneity of variances in a two-way classification. *Biometrics, 25,* 153-158.

JOHNSON, N. L. 1962. Some notes on the investigation of heterogeneity in interaction. *Trabajos de Estadistics, XIII,* 183-199.

MANDEL, J. 1961. Non-additivity in two-way analysis of variance. *J. Am. Stat. Assoc., 56,* 878-888.

MILLIKEN, G. A., AND GRAYBILL, F. A. 1970. Extension of general linear hypothesis. *J. Am. Stat. Assoc., 65,* 797-807.

PERKINS, JEAN M., AND JINKS, J. L. 1968. Environmental and genotype environmental components of variability. III. Multiple lines and crosses. *Heredity, 23,* 239-256.

RAO, C. R. 1970. Estimation of heteroscedastic variances in linear models. *J. Am. Stat. Assoc., 65,* 161-172.

RUSSELL, T. S., AND BRADLEY, R. A. 1958. One-way variances in a two-way classification. *Biometrika, 45,* 111-129.

SHUKLA, G. K. 1971. An invariant test for the homogeneity of variances in a two-way classification. To appear in *Biometrics.*

TAI, G. C. C. 1971. Genotype stability analysis and its application to potato regional trials. *Crop Science, 11,* 184-190.

TUKEY, J. W. 1949. One degree of freedom of non-additivity. *Biometrics, 5,* 232-242.

YATES, F., AND COCHRAN, W. G. 1938. The analysis of group experiments. *J. Agric. Sci., 28,* 556-580.