YIELD-STAND ANALYSES*

By R. G. D. Steel and W. T. Federer

Biometrics Unit, Department of Plant Breeding
Cornell University, Ithaca, N.Y.

1. Introduction

An experimenter is generally interested in both the stand and yield of his crops. When observations are made on both characteristics, the question of how such data are to be analyzed arises.

For sugar beet data, yield may be regarded as the product of the number or stand per plot and average weight per beet per plot. Thus, an analysis of yield is a form of joint analysis of the characters number and average weight; this analysis may be the most pertinent in terms of answering the questions posed in the planning of the experiment. On the other hand, an analysis of yield adjusted for stand, *i.e.*, an analysis of covariance with stand as the covariate, eliminates the effect on yield of variations in stand. This may be desirable if variation in stand is not a result of the treatments or if average weight is not affected by competition among plants within the plot. If differences among stand are expected to be attributable to treatments, the analysis of variance of yield may be what is desired; an analysis of covariance of yield adjusted for covariance on stand may eliminate, at least partially, true treatment differences.

The experimenter may be interested in an analysis of variance of stand itself. Finally, the experimenter may wish to consider analyses of variance of both yield and stand. If inferences are to be made from a joint consideration of the characters yield and stand, some rules of procedure are required. For example, in a more general case we might agree to declare the difference between the effects of two treatments to be significant if: (i) differences between corresponding treatment means for at least one character are significant, (ii) differences between corresponding treatment means for both characters are significant, or (iii) differences between the sums or some other combination of the means of stand and yield for the two treatments are significant. Other rules of procedure are possible.

* Paper No. 322, Department of Plant Breeding and No. 18, Biometrics Unit, Department of Plant Breeding, Cornell University, Ithaca, N.Y.

TABLE I

Number of beets, X_1, yield, X_2, and average weight per beet, X_3, together with ranks of treatment means for Snedecor's sugar beet example

Trt. No.	Fertilizer	Character*	Block 1	Block 2	Block 3	Block 4	Block 5	Block 6	Total	Mean	Rank
1	None ..	X_1	183	176	291	254	225	249	1378	229·7	2·5
		X_2	2·45	2·25	4·38	4·35	3·42	3·27	20·12	3·35	1
		X_3	·0134	·0128	·0151	·0171	·0152	·0131	·0867	·0144	1
2	P, superphosphate	X_1	356	300	301	271	288	258	1774	295·7	4
		X_2	6·71	5·44	4·92	5·23	6·74	4·74	33·78	5·63	4
		X_3	·0188	·0181	·0163	·0193	·0234	·0184	·1143	·0190	4
3	K, muriate of potash	X_1	224	258	244	217	192	236	1371	228·5	1
		X_2	3·22	4·14	2·32	4·42	3·28	4·00	21·38	3·56	3
		X_3	·0144	·0160	·0095	·0204	·0171	·0169	·0943	·0157	3
4	P+K ..	X_1	329	283	308	326	318	318	1882	313·7	5
		X_2	6·34	5·44	5·22	8·00	6·96	6·96	38·92	6·49	5
		X_3	·0193	·0192	·0169	·0245	·0219	·0219	·1237	·0206	6
5	P+N (N, sodium nitrate)	X_1	371	354	352	331	290	410	2108	351·3	6
		X_2	6·48	7·11	5·88	7·54	6·61	8·86	42·48	7·08	6
		X_3	·0175	·0201	·0167	·0228	·0228	·0216	·1215	·0202	5
6	K+N ..	X_1	230	221	237	193	247	250	1378	229·7	2·5
		X_2	3·70	3·24	2·82	2·15	5·19	4·13	21·23	3·54	2
		X_3	·0161	·0147	·0119	·0111	·0210	·0165	·0913	·0152	2
7	P+K+N ..	X_1	322	367	400	333	314	385	2121	353·5	7
		X_2	6·10	7·68	7·37	7·83	7·75	7·39	44·12	7·35	7
		X_3	·0189	·0209	·0184	·0235	·0247	·0192	·1256	·0209	7
	Total ..	X_1	2015	1959	2133	1925	1874	2106	12012	286·0	
		X_2	35·00	35·30	32·91	39·52	39·95	39·35	222·03	5·29	
		X_3	·1184	·1218	·1048	·1387	·1461	·1276	·7574	·0180	

† $\Sigma x_1^2 = 152{,}158{\cdot}00$; $\Sigma x_2^2 = 142{\cdot}4022$; $\Sigma x_3^2 = 0{\cdot}00055969$; $\Sigma x_1 x_2 = 4{,}163{\cdot}69$; $\Sigma x_1 x_3 = 5{\cdot}3445$

* Weight, X_2, and average weight, X_3, have been multiplied by a factor converting them to tons per acre. This will not affect conclusions drawn from any of the analyses, discriminant functions, or other techniques used here.

† Lower case letters are used when the effect of the mean has been eliminated,

e.g., $\Sigma x_1^2 = \Sigma X_1^2 - \frac{(\Sigma X_1)^2}{n}$.

In experiments where it is desired to make statements involving two or more characters jointly, the calculation of a correct probability usually poses difficult problems. For the case where neither stand nor yield is to be relegated to the role of a covariate, a bivariate analysis is appropriate whatever the degree of relation between the variables. By a bivariate analysis is meant one in which a set of paired means, a pair consisting of a mean for each character for the given treatment, is tested to determine whether the spread in a circular or elliptical area can be attributed to chance.

The data used in this paper were obtained from Section 12·7, *Statistical Methods*, by George W. Snedecor. They are presented here in Table I with the addition of average weight per beet, X_3, and of ranks. Snedecor gives analyses of variance for stand and yield and an analysis of covariance with stand as the covariate.

2. Analyses of Variance

Univariate analyses of the data are given in Table II. The treatment mean square is highly significant in all cases. The discriminating

TABLE II

Analyses of variance of number of beets, X_1, *yield,* X_2, *and average weight per beet,* X_3

Source of variation	*d.f.*	Mean squares		
		X_1	X_2	X_3
Blocks ..	5	1,495	1·26	0·00003121†
Treatments ..	$f_t=6$	19,337†	18·81†	0·00004717†
Error ..	$f_e=30$	956	0·774	0·00000402
Discriminating ability ..		80·2%	82·9%	68·4%

† Observed F is greater than tabulated F at the 1% level.

ability is measured as the ratio of treatment sum of squares to (treatment + error) sum of squares, *i.e.*, it is the percentage of a total sum of squares attributable to a variable of classification, in this case treatment effects. Also, it is the square of a correlation coefficient. The squared correlation coefficient and the F-value are different measures of the ability to detect differences among treatment means. It can be shown that

$$\frac{\text{Treatment } SS}{\text{Treatment } SS + \text{Error } SS} = \frac{f_t F}{f_e + f_t F}$$

so that the one to one relationship between the two quantities is obvious. The squared correlation coefficient is sometimes more useful than F in comparing discriminating abilities.

The biologist should not be misled into using the variable which gives the best discrimination if it is not the most pertinent to the questions which led to the experiment. In particular, the use of a ratio like X_3 is often subject to criticism on valid grounds. For example, real differences due to treatments may be hidden in a ratio when there is a linear regression of yield on stand which does not pass through the origin; the same thing can happen with regression through the origin if treatments affect stand but not the ratio; at the same time, non-additivity of effects and heterogeneity of variance may result. Equally misleading conclusions may be drawn as the result of other situations.

Once the significance of differences among treatment means has been established, the location of significant differences, *i.e.*, discrimination among treatment means, becomes important and a number of techniques are available. While the non-orthogonal comparisons given for these data by Snedecor in Example 15.12 probably form the most satisfactory set, there are now valid methods of comparison for cases where no such relations exist among the treatments. We shall apply one of these methods.

Let us apply Tukey's *hsd* (honestly significant difference) procedure to yield, X_2. The procedure is outlined by Federer (1955) and consists of: (i) allotting treatment means to biological or other natural groups according to treatment. (We proceed as though there were a single natural group though this is not the case.) (ii) Choosing a significance level, say 5%. (iii) Computing and applying an *hsd*. For this, we require the standard deviation of a mean, *i.e.*, $s_{\bar{x}} = \sqrt{0 \cdot 774/6} = \cdot 36$, and a factor from a table by Pearson and Hartley (1954). The factor is $q_\alpha = 4 \cdot 46$ for 7 treatments and 30 *d.f.* in error. The

$$hsd = q_\alpha s_{\bar{x}} = 4 \cdot 46 \, (\cdot 36) = 1 \cdot 61$$

is applicable to the testing of linear combinations of treatment means where the sum of the coefficients is zero. The error rate is on a per-experiment basis, *i.e.*, when the null hypothesis is true, on the average one experiment out of 20 will contain at least one comparison which will be declared significant.

By this scheme, each of $\bar{x}_1$, $\bar{x}_6$ and $\bar{x}_3$ is declared significantly different from each of $\bar{x}_2$, $\bar{x}_4$, $\bar{x}_5$ and $\bar{x}_7$ since the smallest difference, $\bar{x}_2$

$-\bar{x}_3 = 5\cdot63 - 3\cdot56 = 2\cdot07$, is greater than the *hsd.* Also, $\bar{x}_2$ is declared significantly different from $\bar{x}_7$ since $\bar{x}_7 - \bar{x}_2 = 1\cdot72$. No other difference between two means is declared significant.

3. Analysis of Covariance

When the experimenter considers the analysis of covariance with stand as the covariate to be appropriate, he tests differences among treatment means after they have been adjusted to a common stand. This method is appropriate when differences among treatment means for stand are due to random sampling. However, in cases where non-random differences exist in the covariate due to treatment or environmental effects, covariance may not be applicable.

Table III contains pertinent sums of squares from the analyses of covariance for stand and yield and stand and average weight per beet.

Table III

Analyses of covariance

Source of variation	*d.f.*	Sums of squares adjusted for regression of X_1 on	
		X_2	X_3
Error ..	29	6·9969	·00010256
Treatment+Error ..	35	9·4655	·00013684
Adjusted treatment means	6	2·4686	·00003428
Discriminating ability ..		26·1 %	25·1 %

Discriminating ability is as previously defined. Only about one quarter of the adjusted (treatment + error) sum of squares can now be attributed to treatment effects. Loss of ability to discriminate was to be expected since there is a marked correlation between the paired treatment means X_1 with X_2 and X_1 with X_3. This is seen from ranks alone. Differences among treatment means are no longer significant, *F*-values being 1·71 and 1·61 respectively.

If the analysis of covariance had indicated significant differences among adjusted treatment means, discrimination would be especially time-consuming because of the significant differences among means for the covariate. When no such differences are present, methods such as that due to Tukey as illustrated in Section 2 or that due to Duncan as illustrated in Section 5 can be applied with $s_{\bar{x}}$ calculated as

$$\sqrt{\frac{s_{y \cdot x}^2}{n}\left[1 + \frac{t_{xx}}{E_{xx}}\right]}$$

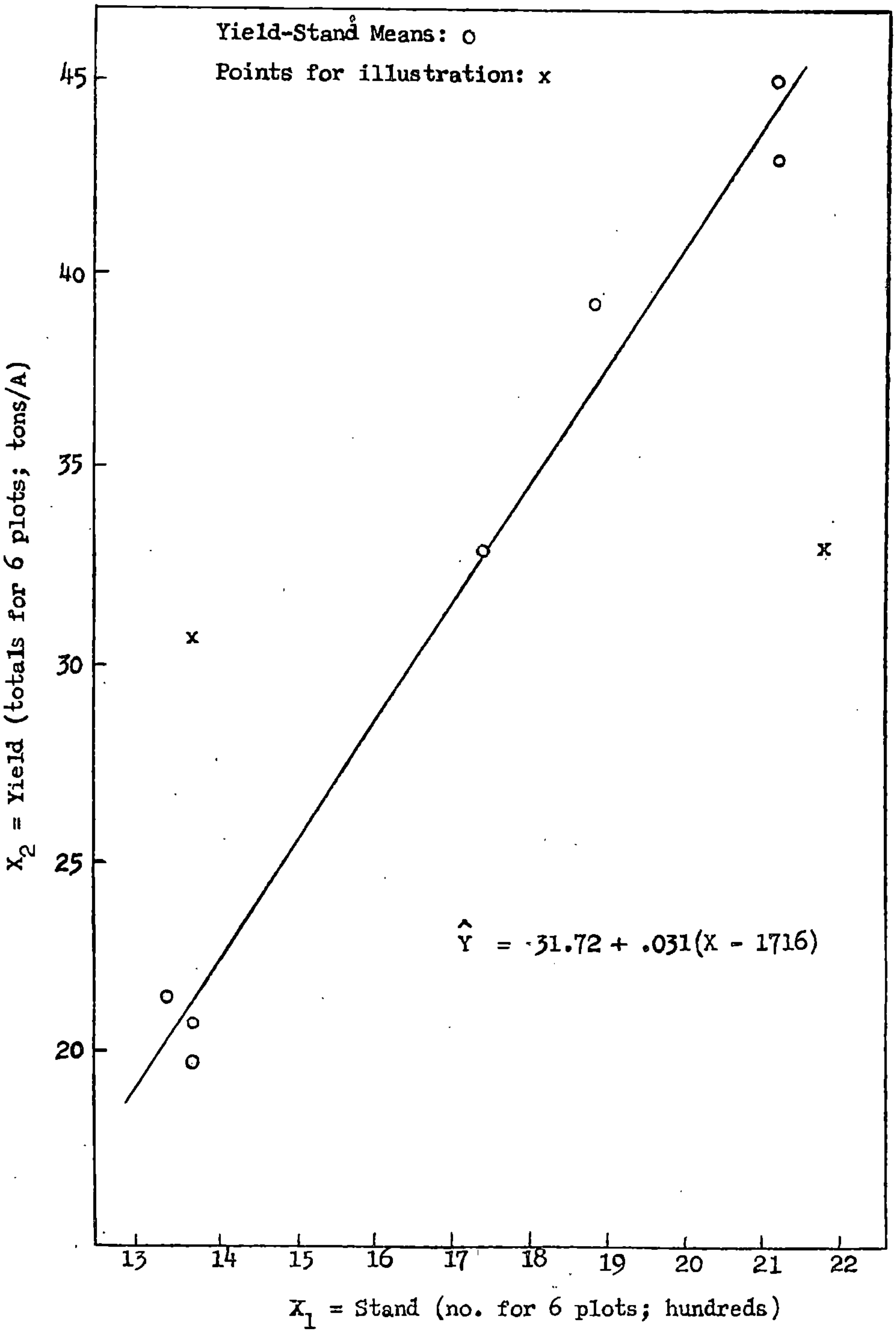
Yield-Stand Means: o
Points for illustration: x
45
40
35
30
25
20
X_2 = Yield (totals for 6 plots; tons/A)
$\hat{Y}$ = 31.72 + .031(X - 1716)
13
14
15
16
17
18
19
20
21
22
X_1 = Stand (no. for 6 plots; hundreds)

FIG. 1

where $s_{y.x}^2$ is the error variance adjusted for regression, t_{xx} and E_{xx} are respectively the treatment mean square and the error sum of squares for the covariate and n is the number of observations per mean [see Finney (1946)].

4. Bivariate Analysis of Variance

When the experimenter wishes to use the information in both variables without relegating one to the position of a covariate and also to take cognizance of the degree of relationship between the variables, he may use a bivariate analysis of variance.

To grasp the ideas underlying a bivariate analysis, consider Fig. 1. This is a plot of the paired treatment means (X_1, X_2), shown by dots, and two additional points marked x to be used for illustration. If all points were to lie more or less "closely" together in an ellipse or circle, intuition would suggest no differences among the paired treatment means. Here the seven points representing the observed pairs of means appear to be extended in a manner strongly indicative of real differences. Further, it would appear that these points lie reasonably "close" to a line. This suggests the possibility that some sort of a standard deviation appropriate to the line can be used to discriminate among such paired treatment means. If the points had shown "too much" scatter from the line as when the two extra points are considered as part of the data, then the line itself and a single standard deviation appropriate to it would seem inadequate for the best discrimination.

For a bivariate analysis, the necessary numerical values are the sums of squares and cross-products of the analysis of covariance.

The various mean products of Table IV are obtained by dividing sums of products by appropriate degrees of freedom. The *multivariance components* in the last column are derived in the usual manner, *i.e.*, a single position within a matrix is considered at a time. For example, $77 \cdot 0015 = (1494 \cdot 5140 - 955 \cdot 5033)/7$, and $96 \cdot 1558 = \frac{1}{6}(599 \cdot 6750 - 22 \cdot 7400)$, where 7 and 6 are the numbers of observations in the block and treatment totals. Multivariance components are useful in much the same manner as are ordinary variance components.

To test the significance of a source of variation in a bivariate analysis, it is necessary to compute the statistic U, defined as

$$U = \frac{\begin{vmatrix} E_{11} & E_{12} \\ E_{21} & E_{22} \end{vmatrix}}{\begin{vmatrix} E_{11} + T_{11} & E_{12} + T_{12} \\ E_{21} + T_{21} & E_{22} + T_{22} \end{vmatrix}}$$

TABLE IV

Bivariate analysis of variance of X_1 *and* X_2

Source of variation	*d.f.*	Sum of products	Mean product	Multivariance components
Total ..	41	$\begin{pmatrix} 152,158\cdot00 & 4,163\cdot69 \\ 4,163\cdot69 & 142\cdot4022 \end{pmatrix}$		
Replicate or block	5	$\begin{pmatrix} 7,472\cdot57 & -116\cdot56 \\ -116\cdot56 & 6\cdot3134 \end{pmatrix}$	$\begin{pmatrix} 1,494\cdot5140 & -23\cdot3120 \\ -23\cdot3120 & 1\cdot2627 \end{pmatrix}$	$\begin{pmatrix} 77\cdot0015 & -6\cdot5789 \\ -6\cdot5789 & 0\cdot0698 \end{pmatrix}$
Treatment ..	$f_t=6$	$\begin{pmatrix} 116,020\cdot33 & 3,598\cdot05 \\ 3,598\cdot05 & 112\cdot8562 \end{pmatrix}$	$\begin{pmatrix} 19,336\cdot7217 & 599\cdot6750 \\ 599\cdot6750 & 18\cdot8094 \end{pmatrix}$	$\begin{pmatrix} 3,063\cdot5364 & 96\cdot1558 \\ 96\cdot1558 & 3\cdot0058 \end{pmatrix}$
Error ..	$f_e=30$	$\begin{pmatrix} 28,665\cdot10 & 682\cdot20 \\ 682\cdot20 & 23\cdot2326 \end{pmatrix}$	$\begin{pmatrix} 955\cdot5033 & 22\cdot7400 \\ 22\cdot7400 & 0\cdot7744 \end{pmatrix}$	

$$U = \frac{\begin{vmatrix} E_{11} & E_{12} \\ E_{21} & E_{22} \end{vmatrix}}{\begin{vmatrix} E_{11}+T_{11} & E_{12}+T_{12} \\ E_{21}+T_{21} & E_{22}+T_{22} \end{vmatrix}} = \frac{\begin{vmatrix} 28,665.10 & 682\cdot20 \\ 682\cdot20 & 23\cdot23 \end{vmatrix}}{\begin{vmatrix} 144,685\cdot43 & 4,280\cdot25 \\ 4,280\cdot25 & 136\cdot09 \end{vmatrix}} = \frac{200,493\cdot4330}{1,369,700\cdot1062} = 0\cdot14637761$$

where E_{ij} is error sum of products and T_{ij} is treatment sum of products.

$$F = \frac{1-\sqrt{U}}{\sqrt{U}}\left(\frac{2\,(f_e-1)}{2f_t}\right) = \frac{\cdot6174\,(58)}{\cdot3826\,(12)} = \frac{35\cdot8092}{4\cdot5912} = 7\cdot80 \text{ with 12 and 58 degrees of freedom.}$$

$$= \frac{E_{11}E_{22} - E_{12}^2}{(E_{11} + T_{11})(E_{22} + T_{22}) - (E_{12} + T_{12})^2}$$

where E_{11} and T_{11} are error and treatment sums of squares for X_1, E_{22} and T_{22} are similar sums of squares for X_2, and $E_{12} = E_{21}$ and $T_{12} = T_{21}$ are error and treatment sums of products of X_1 and X_2.

The numerical value of the criterion U lies between zero and one, with values near one supporting the null hypothesis and values near zero indicating significant departures. The null hypothesis states that there are no differences within the set of pairs of means, *i.e.*, that sampling is from a single bivariate population. The quantity $\sqrt{U}$ has been shown by Wilks (1935, 1946) to have a beta distribution with parameters p and q as used by Pearson (1932) equal to (residual *d.f.* $-$ 1) and variety *d.f.* respectively. The square-root is related to Snedecor's F and follows:—

$$F(n, m) = \frac{1 - \sqrt{U}}{\sqrt{U}} \cdot \frac{m}{n}$$

with $n = 2$ (variety *d.f.*) $= 12$ and $m = 2$ (residual *d.f.* $- 1) = 58$ as *d.f.* to be used in entering the F-table. There is little doubt of the existence of real differences among the seven pairs of treatment means.

5. Further Tests in a Bivariate Analysis, the Discriminant Function

As with a significant F, the problem of discrimination among these unlike pairs now arises. To obtain further information on the nature of the differences, consider a determinantal equation in U, namely

$$\begin{vmatrix} (E_{11} + T_{11})U - E_{11} & (E_{12} + T_{12})U - E_{12} \\ (E_{21} + T_{21})U - E_{21} & (E_{22} + T_{22})U - E_{22} \end{vmatrix} = 0.$$

For our data on (X_1, X_2), we have

$$\begin{vmatrix} 144{,}685\cdot43\,U - 28{,}665\cdot10 & 4{,}280\cdot25\,U - 682\cdot20 \\ 4{,}280\cdot25\,U - 682\cdot20 & 136\cdot09\,U - 23\cdot23 \end{vmatrix} = 0,$$

or

$$1{,}369{,}700\cdot1062\,U^2 - 1{,}422{,}102{,}\cdot8979\,U + 200{,}493\cdot4330 = 0.$$

The roots are $U_1 = \cdot87001$ and $U_2 = \cdot16825$ with $U = U_1U_2 = (\cdot87001)(\cdot16825) = \cdot1464$.

While the exact distributions of these roots are obtainable, a chi-square approximation seems reasonably useful especially where significance is so pronounced, and is more easily obtained. This approximation is also available for U. The approximations are:

for U:

$$\chi^2 \{(pf_t)\, df\} = -\left\{f_e + f_t - \frac{p + f_t + 1}{2}\right\} \log_e U$$

$$= -\, 2\cdot 30259 \left(30 + 6 - \frac{2 + 6 + 1}{2}\right) \log_{10} (\cdot 14637761)$$

$$= 77\cdot 82.$$

for U_2:

$$\chi^2 \{(p + f_t - 1)\, df\} = -\left(f_e + f_t - \frac{p + f_t + 1}{2}\right) \log_e U_2$$

$$= -\, 2\cdot 30259 \left(30+6- \frac{2+6+1}{2}\right) \log_{10} (\cdot 16825)$$

$$= 72\cdot 18 \text{ with } 7\ d.f.$$

and for U_1:

$$\chi^2 \{(p + f_t - 3)\, df\} = -\left(f_e + f_t - \frac{p + f_t + 1}{2}\right) \log_e U_1$$

$$= -\, 2\cdot 30259 \left(30+6- \frac{2+6+1}{2}\right) \log_{10} (\cdot 87001)$$

$$= 5\cdot 64 \text{ with } 5\ d.f.$$

where f_t = treatment degrees of freedom, f_e = error degrees of freedom, and p = number of characteristics. These results are summarized in Table V.

The single significant root is quite helpful in that it tells us that a single linear function may be used to discriminate among paired treatment means. In other words, some line as in Fig. 1 together with an appropriate standard deviation as a measuring stick, can adequately

TABLE V

Roots of a determinantal equation

Root	*d.f.*	$\chi^2 =$	Prob. of greater $\chi^2 =$	$R^2 = 1 - U$
$U_1 = \cdot 8700$ ..	$p + f_t - 3 = 5$	5·64	Approx. = ·40	·1300
$U_2 = \cdot 1682$ ..	$p + f_t - 1 = 7$	72·18	< ·0001	·8318
$U = \cdot 1464$ ..	$pf_t = 12$	77·82	< ·0001	

The chi-squares for the two roots, U_1 and U_2, add to the chi-square value for U.

locate significant differences among the set of seven points. Notice the pronounced linear relation between treatment means for stand and yield. If both roots are significant, a second discriminant function is required and the problems of interpretation are increased. The alternative would be some bivariate criterion.

The values U_1 and U_2 contain still more information. They are associated with discriminant functions of which only one, that for U_2, appears to be useful. These two functions are uncorrelated. The complements of the U-values $1 - U_1$ and $1 - U_2$, are the squares of so-called *canonical correlations*, measures of the dependence of the data, after removal of replicate effects, on variety effects. Thus $R_2{}^2 = 1 - U_2 = \cdot 8318$ says that 83% of the (treatments + error) sum of squares can be accounted for by treatments if the appropriate linear combination of stand and yield is used; and no other single linear function can be found which will do as well. We have already seen in Table II that yield alone does virtually as well, the difference being trivial.

In order to find the discriminant function implied by U_2, we solve the equations:

$$[(1 - U_2)(E_{11} + T_{11}) - T_{11}]\, a_1 + [(1 - U_2)(E_{12} + T_{12}) - T_{12}]\, a_2 = 0$$

and

$$[(1 - U_2)(E_{12} + T_{21}) - T_{21}]\, a_1 + [(1 - U_2)(E_{22} + T_{22}) - T_{22}]\, a_2 = 0$$

or

$$4{,}321\cdot 77\, a_1 - 37\cdot 95\, a_2 = 0 \text{ and } -37\cdot 95\, a_1 + \cdot 336\, a_2 = 0,$$

for a_1 and a_2. The equations give $a_2/a_1 = 37\cdot 95/\cdot 336 = 113$ or $a_2 = 113a_1$. The discriminant function is, then, $X_1 + 113X_2$.

Consider again the statement that 83% of the (treatments + error) sum of squares of the variable $X_1 + 113 X_2$, called a canonical variable, can be attributed to treatments. An analysis of variance of this variable can be found directly from the coefficients 1 and 113 and the original bivariate analysis by evaluating two expressions, of which the one for error SS is:

$$(1 \quad 113) \begin{pmatrix} 28,665\cdot 10 & 682\cdot 20 \\ 682\cdot 20 & 23\cdot 2326 \end{pmatrix} \begin{pmatrix} 1 \\ 113 \end{pmatrix} = 479,499\cdot 3694.$$

For treatments, the corresponding sum of squares is 2,370,240·4478. The ratio of treatment to (treatment + residual) sum of squares is ·8318.

To actually discriminate among treatments for the canonical variable, we require new treatment means as well as the new error term. These are:

Treatment:	None	P	K	$P+K$	$P+N$	$K+N$	$P+K+N$
Number:	1	2	3	4	5	6	7
Mean:	608·2	931·9	630·8	1047·1	1151·3	629·7	1184·0

For example, $608\cdot 2 = 229\cdot 7 + 113\,(3\cdot 35)$. To discriminate, an $s_{\bar{x}}$ is required. The error sum of squares has already been calculated. Had the usual discriminant function analysis been applied to the data and had the results been presented in analysis of variance form, treatments would have 7 *d.f.* and error 29 *d.f.* The loss of a *d.f.* from error can be rationalized on the basis that the number 113 was estimated from the data. Let us use $30 - 1 = 29$ *d.f.* for our error mean square. Then,

$$s_{\bar{x}} = \sqrt{\frac{479,499\cdot 3694}{29 \times 6}} = 52\cdot 50.$$

We now use Duncan's (1955) New Multiple Range Test to discriminate among the treatment means. This test, like Tukey's in Section 2, is not the most appropriate procedure for such a set of treatments but is used here for illustrative purposes. First, rank the treatment means. The means, with treatment numbers in parentheses, are given in Table VI.

TABLE VI

Ranked treatment means with corresponding treatment numbers

608·2 (1)	629·7 (6)	630·8 (3)	931·9 (2)	1047·1 (4)	1151·3 (5)	1184·0 (7)

The spacing corresponds roughly to the separation

Secondly, enter Duncan's (1955) table of Significant Studentized Ranges for a 5% level test at row $n_2 = 29$ *d.f.* and obtain significant ranges from columns $p = 3, 4, 5, 6, 7$ and 8 but refer to them as sizes $2, 3 \ldots, 7$. This is a temporary expedient suggested by the authors; in the analysis of variance of a single variable or of a linear function such as $X_1 + aX_2$ with the coefficient a not supplied by the data, tabulated p-values of $2, 3 \ldots, 7$ would be used. The expedient is suggested since an extra *d.f.* appears in treatment sum of squares in the analysis of variance presentation referred to. The values are given in Table VII. The values of range($s_{\bar{x}}$) are called *shortest significant ranges*.

TABLE VII

Significant Studentized ranges and shortest significant ranges

Actual No. of means	2	3	4	5	6	7
Tabulated No. of means	3	4	5	6	7	8
Significant ranges ..	3·04	3·12	3·20	3·25	3·29	3·32
Range($s_{\bar{x}}$) ..	159·6	163·8	168·0	170·6	172·7	174·3

Differences are tested in the order largest *minus* smallest, largest *minus* second smallest, ..., largest *minus* second largest, second largest *minus* smallest, ..., second smallest *minus* smallest. Here, the order is (7) — (1), (7) — (6), (7) — (3), (7) — (2), (7) — (4), (7) — (5), (5) — (1), (5) — (6), (5) — (3), (5) — (2), (5) — (4), (4) — (1), (4) — (6), (4) — (3), (4) — (2), (2) — (1), (2) — (6), (2) — (3), (3) — (1), (3) — (6) and (6) — (1).

A difference is significant only if it exceeds the corresponding least significant range *except* that "no difference between two means can be declared significant if they are between two other means with a non-significant range." "Between" may be interpreted as including one of the means in the non-significant range. As soon as a non-significant range is found, it is convenient to underline these means and those between. The procedure is indicated in Table VI. Tables VI and VII may be combined to give a single worksheet.

Means (4), (5) and (7) are underlined in Table VI because (7) — (4) = 137·77 is smaller than 163·8, the least significant range for three means. Any two means not underlined with the same line are declared significantly different.

Duncan's procedure has an error rate that is on a per-single-*d.f.* basis. It is thus seen to be a different procedure from Tukey's. By Duncan's method, we find $\bar{x}_2$ significantly different from $\bar{x}_5$. This was not significant using X_2 alone and Tukey's procedure. Detection of this difference is likely due to the difference in the two authors' methods of setting error rates rather than to the introduction of X_1 into the discriminant function.

6. An Alternative Discriminant Function

Smith (1936-37) first described the application of a discriminant function or selection index for plant selection. He pointed out that in selecting for quantitative characters such as yield, the differences due to genotype are masked by environmental effects. Plant breeders attempt to select breeding material on the basis of observable characteristics which they believe are associated with the desired character since the actual worth of each of the observable characteristics is usually unknown. Smith suggested that the discriminant function approach be used to best indicate the "genetic value" of a line.

It may be assumed that the true genotype of the plant is measured by

$$\theta = \Sigma c_i \xi_i,$$

where the c_i are assigned values representing relative values of the observed characters, X_i, whose true values are ξ_i.

This function cannot be evaluated directly because only the phenotypic performance and not the genotypic performance is observed. Let the phenotypic value be represented by the equation

$$Y = \Sigma b_i X_i.$$

The problem is to find values of b_i such that the function Y will detect best those lines which have the greatest genotypic value θ; that is, the b_i are to be such that the regression of Y on θ will be maximum. If the line variances and covariances are denoted by f_{ij}, the error variances and covariances by e_{ij}, and if $g_{ij} = f_{ij} - e_{ij}$ represents a multiple of the multivariance component which is an estimate of the component due to genotype, maximization of the regression of Y on θ results in the following equations:—

$$b_1 f_{11} + b_2 f_{12} + \ldots + b_p f_{1p} = A_1$$

$$\ldots\ldots\ldots$$

$$b_1 f_{1p} + b_2 f_{2p} + \cdots + b_p f_{pp} = A_p,$$

where

$$A_1 = c_1 g_{11} + c_2 g_{12} + \cdots + c_p g_{1p}$$

$$\ldots\ldots\ldots$$

$$A_p = c_1 g_{1p} + c_2 g_{2p} + \cdots + c_p g_{pp}.$$

The A_j are computed from the data after the c_i have been decided upon. Goulden (1952) suggests that the c_i be set equal to the reciprocal of some multiple of the mean for a given character.

Though not developed for this purpose, the above procedure may be used directly on the sugar beet data after the relative worth of the two characters, stand and yield per plot, is decided upon. Since stand is such an important part of yield in sugar beets, it might be advisable to give the variables equal coefficients, that is set $c_1 = 1 = c_2$. Any values that appear reasonable to the experimenter may be used. From Table IV, the $p = 2$ equations involving b_1 and b_2 are:—

$$19{,}336\cdot7217b_1 + 599\cdot6750b_2 = 3{,}063\cdot5364c_1 + 96\cdot1558c_2 = 3{,}159\cdot6922.$$

$$599\cdot6750b_1 + 18\cdot8094b_2 = 96\cdot1558c_1 + 3\cdot0058c_2 = 99\cdot1616.$$

Solution of these equations gives

$$b_1 = -\,0\cdot008000436 \text{ and } b_2 = 5\cdot526984463,$$

or in relative values

$$b_1\sqrt{E_{11}/f_e} = -\,0\cdot008000436\sqrt{955\cdot5035} = -\,0\cdot2473 \text{ and}$$

$$b_2\sqrt{E_{22}/f_e} = 5\cdot526984463\,\sqrt{0\cdot7744} = 4\cdot8637.$$

Thus, yield is about 20 times more important than stand in discriminating among the 7 treatments. The discriminant function is

$$Y = -\,X_1 + 690\cdot8\,X_2,$$

where the coefficients are divided by $0\cdot008000436$ to give X_1 a coefficient of $-\,1$, a more convenient form for the discriminant function.

The ability of this function to discriminate among transformed treatment means is found as in Section 5. We need treatment and (treatment + error) sums of squares as if an analysis of variance of values of $-\,X_1 + 690\cdot8X_2$ has been made. The treatment sum of squares is 49,000,456·74 (see Section 5 for computing method). The (treatment + error) sum of squares is 61,834,415·87. The ratio of treatment to (treatment + error) sum of squares, *i.e.*, the proportion of this total sum of squares attributable to or explained by treatment effects, is $R^2 = \cdot7924$.

This figure of about 79% had to be less than the corresponding figure of about 83% for the previous discriminant function for that was the maximum possible value for a linear function. However, the figures are about the same on a practical basis.

7. The Bivariate Analysis For X_1 and X_3

The bivariate analysis of average weight per beet and stand per plot is given in Table VIII. Again only one of the roots is significant and a single discriminant function is adequate. The discriminant function implicit in $U_2 = \cdot 1774$ is $X_1 + 7820X_3$. The discriminating ability is $R^2 = 1 - \cdot 1774 = \cdot 8226$, *i.e.*, about 82% of the total sum of squares for treatments and error can be explained by treatment effects when the variable is $X_1 + 7820X_3$. No other linear function can do better. In this case introduction of X_1 shows a marked improvement over use of X_3 alone where discriminating ability was about 64%.

If H. F. Smith's (1936–37) discriminant function is calculated, the two equations involving b_1 and b_2 for $c_1 = 1 = c_2$ are:

$$19{,}336 \cdot 7b_1 + \cdot 9156b_2 = 3{,}063 \cdot 53c_1 + \cdot 1486c_2 = 3{,}063 \cdot 6786 \text{ and}$$

$$\cdot 9156b_1 + \cdot 00004717b_2 = \cdot 1486c_1 + \cdot 00000719c_2 = \cdot 14860719,$$

and the solution for the b's gives:

$$b_1 = \cdot 114514341 \text{ and } b_2 = 927 \cdot 629246,$$

or in relative values

$$b_1 \sqrt{E_{11}/f_e} = \cdot 114514341 \sqrt{955 \cdot 5} = 3 \cdot 5398 \text{ and}$$

$$b_2 \sqrt{E_{22}/f_e} = 927 \cdot 629246 \sqrt{\cdot 00000402} = 1 \cdot 8599.$$

The discriminant function is

$$Y = X_1 + 8101X_3.$$

In the analysis, stand per plot is about twice as important as average weight per beet in discriminating among the 7 treatment means. Here, the discriminating ability is virtually the same since the function $X_1 + 7820X_3$ is so little different.

8. Concluding Remarks

The bivariate analysis is easily expanded to a multivariate analysis which can deal with more than two characters. A character, *e.g.*, yield, measured in several years or at several locations may be treated as several characters and a multivariate analysis of the resulting data may be performed. Tukey (1949) gives an example for an annual crop where the variables are yield in each of two years. Steel (1955) gives an example for a perennial crop where the same plots are harvested in successive years. These are alternatives to analyses of variance with years as sources of variation.

TABLE VIII

Bivariate analysis of variance for X_1 and X_3

Source of variation	d.f.	Sum of products	Mean product	Multivariance components
Total ..	41	$\begin{pmatrix} 152{,}158 & 5\cdot3445 \\ 5\cdot3445 & 0\cdot00055969 \end{pmatrix}$		
Replicate or block	5	$\begin{pmatrix} 7{,}473 & -0\cdot8685 \\ -0\cdot8685 & 0\cdot00015605 \end{pmatrix}$	$\begin{pmatrix} 1{,}494\cdot6 & -0\cdot1737 \\ -0\cdot1737 & 0\cdot00003121 \end{pmatrix}$	$\begin{pmatrix} 77\cdot01 & -0\cdot0282 \\ -0\cdot0282 & 0\cdot00000388 \end{pmatrix}$
Treatment ..	$f_t=6$	$\begin{pmatrix} 116{,}020 & 5\cdot4937 \\ 5\cdot4937 & 0\cdot00028303 \end{pmatrix}$	$\begin{pmatrix} 19{,}336\cdot7 & 0\cdot9156 \\ 0\cdot9156 & 0\cdot00004717 \end{pmatrix}$	$\begin{pmatrix} 3{,}063\cdot53 & 0\cdot1486 \\ 0\cdot1486 & 0\cdot00000719 \end{pmatrix}$
Error ..	$f_e=30$	$\begin{pmatrix} 28{,}665 & 0\cdot7193 \\ 0\cdot7193 & 0\cdot00012061 \end{pmatrix}$	$\begin{pmatrix} 955\cdot5 & 0\cdot0240 \\ 0\cdot0240 & 0\cdot00000402 \end{pmatrix}$	

$U = 0\cdot14848482$; $F\,(12, 58) = 7\cdot71$.

$19\cdot79928440\, U^2 - 20\cdot08277665\, U + 2\cdot93989316 = 0$; $U_2 = \cdot1774$; $U_1 = \cdot8369$.

For U_2: $\chi^2 = 70\cdot04$ with 7 *d.f.* and for U_1: $\chi^2 = 7\cdot21$ with 5 *d.f.*

The multivariate analysis of data where the different variables are the same characteristic, unlike the analysis of variance, requires no assumption about homogeneity of error variance from year to year or location to location. When more than one root of the *U*-equation is significant, more than one discriminant function is necessary; they are calculated in the same manner as here. A test of significance for the comparison of the discriminant function determined by the data and any proposed discriminant function is given by Fisher (1940).

The procedures of Duncan and Tukey do not require the calculation of an initial *F*-value and are more widely applicable than indicated here [see Duncan (1955) and Federer (1955)]. They can be used for tests involving linear functions of the observations. These procedures together with others are discussed in Chapter II of Federer's *Experimental Design* where necessary tables are available. Duncan's tables are also available in his paper (1955).

9. Summary

The analysis of variance, of covariance and a bivariate analysis are presented for an example in Chapter 12 of Snedecor's *Statistical Methods*. Methods and suggestions for testing differences in a set of means are given. In particular, the discriminant function for which the ratio of treatment sum of squares to (treatment + error) sum of squares is a maximum is calculated. The coefficients of the variables in this discriminant function are obtained from the data. A discriminant function with coefficients based on information outside the experiment is also calculated.

Tukey's *hsd* test and Duncan's Multiple Range Test are used to illustrate methods of testing all possible differences between pairs of treatment means, pairs of means adjusted for a covariate, and pairs of mean values of a discriminant function.

Literature Cited

Duncan, D. B. .. "Multiple range and multiple F Tests," *Biometrics*, 1955, **11**, 1–42.

Federer, W. T. .. *Experimental Designs—Theory and Application*, New York, N.Y., The Macmillan Company, 1955.

Finney, D. J. .. "Standard errors of yields adjusted for regression on an independent measurement," *Biom. Bull.*, 1946, **2**, 53–55.

Fisher, R. A. .. "The precision of discriminant functions," *Ann. Eugenics*, 1940, **10**, 422–29.

Goulden, C. H. .. *Methods of Statistical Analysis*, Chap. 17, Ed. 2, New York, N.Y., John Wiley & Sons, Inc., 1952.

Pearson, K. .. *Tables of the Incomplete Beta Function*, Cambridge, England, Cambridge University Press, 1932.

Pearson, E. S. and Hartley, H. O. *Biometrika Tables for Statisticians*, Vol. 1, Cambridge, England, Cambridge University Press, 1954.

Smith, H. F. .. "A discriminant function for plant selection," *Ann. Eugenics*, 1936–37, **7**, 240–50.

Steel, R. G. D. .. "An analysis of perennial crop data," *Biometrics*, 1955, **11**, 201–12.

Tukey, J. W. .. "Dyadic anova, an analysis of variance for vectors," *Human Biology*, 1949, **21**, 65–110.

Wilks, S. S. .. "Test criteria for statistical hypothesis involving several variables," *Jour. Am. Stat. Assoc.*, 1935, **30**, 549–60.

———— .. *Mathematical Statistics*, Princeton, N.J., Princeton University Press, 1946.