IMPROVEMENT BY MEANS OF SELECTION

W. G. COCHRAN
JOHNS HOPKINS UNIVERSITY

1. Introduction

One of the principal techniques for improving quality is to select those members of a population that appear to be of high quality and to reject those that appear to be of low quality. Usually the selection is based on a number of measurements that have been made on the available candidates. In personnel selection the measurements are sometimes obtained by competitive examination, the hope being that persons who obtain high marks will have superior ability for performing the subsequent tasks. In a program for improving hogs, the choice of a sire for breeding may be made after a study of his own characteristics, for example, weight at 180 days, plus those of his first few offspring.

A common feature in most selection problems is that at the time of selection we cannot measure directly the quantity which we wish to improve. Thus when a promotion from one type of work to another is in question, for example from salesmanship to an administrative task, success in the old occupation may not guarantee success in the new one. Stated mathematically, the problem is to improve some quantity y by means of indirect selection that is made from a group of tests or measurements $x_1, x_2, \ldots, x_p$.

The mathematical foundation of most of the work that has been done thus far is Karl Pearson's memoir [1] of 1902. His primary interest was in the effects of *natural* selection on correlation and variability. On the assumption that y and the x's follow a multivariate normal distribution, he gave some important theorems about the means, variances and correlations of the variates after a selection based on the values of certain of the x's. Various applications of these and other results are dispersed in the literature on personnel selection and on plant and animal improvement [2], [3], [4].

The object of this paper is to present the principal mathematical results that are useful for setting up a selection program. This part is mainly expository in character, though a few results are given in a more general form than hitherto. In addition, we shall discuss some of the problems that are encountered when we come to apply the theory to selection in practice. Here there appears to be need for much further research.

2. Statement of the problem

We shall assume that y is a continuous variate. This is not always the case in practice, since the object of selection is sometimes to draw out those who possess a specific attribute. The same general approach is valid whether y is continuous

or discrete, though the details differ. Before selection, the variates $y, x_1, x_2, \ldots, x_p$ are assumed to follow a distribution whose frequency function is $f(y, x_1, x_2, \ldots, x_p)$. The decision to accept or reject a candidate is to be made by an objective rule that can be unambiguously applied as soon as the values of the x's are known for the candidate. That is, some region R in the sample space of the x's is chosen as the region of selection.

When we compare different rules for selection, it is natural to make the comparison subject to the condition that all rules operate with the same intensity of selection. In other words, if any rule is applied repeatedly to the parent population, it should in the long run select a fraction α and reject a fraction $(1 - \alpha)$ of the candidates. The size of the region R is therefore α.

It will be supposed that the specific purpose of selection is to maximize the mean value of y in the selected portion of the universe. This purpose has usually been taken for granted in applications, and seems a reasonable one to adopt, though cases can be imagined where a different objective would be more appropriate. For example, we might wish to maximize the probability that y exceeds some value y_0. This would in general require a different mathematical treatment and lead to a different rule of selection.

To simplify the notation, the symbol x will be used to denote the set of p variates $x_1, x_2, \ldots, x_p$, and dx to denote the product of their differentials $dx_1, dx_2, \ldots, dx_p$. Given the joint frequency function $f(y, x_1, x_2, \ldots, x_p) \equiv f(y, x)$, the problem is to find a region R in the sample space of the x's, such that

$$\frac{1}{\alpha} \int_{-\infty}^{\infty} dy \int_R y f(y, x)\, dx \tag{1}$$

is maximized subject to the restriction

$$\int_R f_1(x)\, dx = \alpha \tag{2}$$

where $f_1(x)$ is the joint frequency function of the x's.

3. The optimum rule for selection

If the regression $\eta(x)$ of y on the x's exists, the optimum rule is to select all members for which

$$\eta(x) \geqq k$$

where the value of k is chosen so that the frequency of selection is α. The result requires the assumption that the cumulative distribution function of $\eta(x)$ is continuous and strongly monotone, so that for any α, $(0 < \alpha < 1)$, there is one and only one $k(\alpha)$ for which

$$P\{\eta(x) \geqq k\} = \alpha .$$

To prove the result, we write

$$f(y, x) = \phi(y|x) f_1(x) ,$$

where $\phi(y|x)$ is the conditional frequency function of y, given the x's. The mean

value of y after selection is

$$\frac{1}{a}\int_{-\infty}^{\infty} dy \int_R y\phi(y \mid x) f_1(x)\, dx$$

$$= \frac{1}{a}\int_R f_1(x)\, dx \int_{-\infty}^{\infty} y\phi(y \mid x)\, dy$$

$$= \frac{1}{a}\int_R \eta(x) f_1(x)\, dx, \tag{3}$$

from the usual definition of a regression function. The problem is therefore to find a region R which maximizes (3) subject to (2).

This problem is analogous to that of finding the best critical region R for a test of significance of a null hypothesis H_0 against a single specified alternative H_1. In the analogy, $f_1(x)$ corresponds to p_0, the frequency function of a sample point given that H_0 holds, while the product $\eta(x)f_1(x)$ corresponds to p_1, the frequency function of a sample point given that H_1 holds. Neyman and Pearson [5] have shown that the best critical region R_0 is defined by

$$p_1 \geqq kp_0$$

where k is chosen so that the region is of size a. The corresponding region in our problem is

$$\eta(x) f_1(x) \geqq k f_1(x), \qquad \text{or} \qquad \eta(x) \geqq k. \tag{4}$$

Although the argument seems to apply without change, it may be worth repeating the principal part of the proof. Let R_0 be the region defined by (4) and let R_1 be any other region of size a in the sample space of the x's. If the two regions have a common part, denote this by R_{01}.

Since both regions are of size a, it is clear that

$$\int_{(R_0 - R_{01})} f_1(x)\, dx = \int_{(R_1 - R_{01})} f_1(x)\, dx. \tag{5}$$

Now

$$\int_{R_0} \eta(x) f_1(x)\, dx = \int_{R_{01}} \eta(x) f_1(x)\, dx + \int_{(R_0 - R_{01})} \eta(x) f_1(x)\, dx$$

$$\geqq \int_{R_{01}} \eta(x) f_1(x)\, dx + \int_{(R_0 - R_{01})} k f_1(x)\, dx$$

$$\geqq \int_{R_{01}} \eta(x) f_1(x)\, dx + \int_{(R_1 - R_{01})} k f_1(x)\, dx,$$

using (5). But in $(R_1 - R_{01})$, we have $k > \eta(x)$, so that

$$\int_{R_0} \eta(x) f_1(x)\, dx \geqq \int_{R_{01}} \eta(x) f_1(x)\, dx + \int_{(R_1 - R_{01})} \eta(x) f_1(x)\, dx$$

$$\geqq \int_{R_1} \eta(x) f_1(x)\, dx.$$

The equality will hold only if the region $(R_1 - R_{01})$ is empty, that is, if R_1 and R_0 coincide.

This result might of course be anticipated by elementary considerations. By a selection which operates entirely on the x's we cannot hope to influence the individual variations of y in arrays in which the x's are fixed: the most that we can hope to do is to choose arrays in which the mean value of y is relatively high. The result is a convenient one, since it implies that selection can be based on a single index by which the candidates are scored. The use of the regression as an index is well known for the case where all variates follow the multivariate normal distribution, but actually it does not require normality nor linearity of regression.

4. The gain in y due to selection

We may choose the scales so that in the original population all variates have zero means. Since $E(y) = E(\eta)$ in the unselected population, it follows that η also has zero mean. Hence the mean values of y and η after selection are the increases or gains in these variates due to the selection.

From (3) we see that the gain in y, $G(y)$, is the same as that in η. This result can be put in another form that is sometimes of interest. In measuring the gain in a variate, we often express it as a fraction of the standard deviation of the variate in the original population. This device converts the gain to a type of standard scale which is invariant under any linear transformation of the units in which measurements are recorded. In standard units,

$$\frac{G(y)}{\sigma_y} = \left(\frac{\sigma_\eta}{\sigma_y}\right)\frac{G(\eta)}{\sigma_\eta}.$$

But in the original population,

$$\operatorname{cov}(y\eta) = \int\int y\eta f(y, x)\, dy dx = \int \eta f_1(x)\, dx \int y\phi(y \mid x)\, dy$$

$$= \int \eta^2 f_1(x)\, dx = \sigma_\eta^2.$$

Hence $\rho_{y\eta} = \sigma_\eta/\sigma_y$ and we have

$$\frac{G(y)}{\sigma_y} = \rho_{y\eta}\frac{G(\eta)}{\sigma_\eta}. \tag{6}$$

In standard units, the gain in y is a fraction $\rho_{y\eta}$ of that in η, where $\rho_{y\eta}$ is the correlation coefficient between y and η.

It would be interesting to have a simple expression which gives the gain due to indirect selection in terms of that due to direct selection on y of the same intensity, but I have been unable to discover one. It is of course easy to show that indirect selection cannot be superior to direct selection. A very simple result which connects the two gains is obtained if the variates follow the multivariate normal distribution, as discussed in the next section.

5. Results when all variates follow a multivariate normal distribution

In this case, which has been assumed in most applications, the results can be made more specific. In particular, η is a linear function of the x's and is normally

distributed in the original population. Hence

$$\frac{G(\eta)}{\sigma_\eta} = \frac{1}{a\sqrt{2\pi\sigma_\eta^2}} \int_k^\infty \eta\, e^{-\eta^2/2\sigma_\eta^2} d\eta = \frac{1}{a\sqrt{2\pi}}\, e^{-k^2/2\sigma_\eta^2} = \frac{z(a)}{a}$$

where $z(a)$ is the ordinate of the normal frequency function at the point k/σ_η at which a fraction a of the total area lies above the ordinate. This gives

$$\frac{G(y)}{\sigma_y} = \rho_{y\eta} \frac{z(a)}{a}. \tag{7}$$

If we were able to select directly on y, the gain in y would be $z(a)/a$. Thus the gain due to indirect selection is a fraction $\rho_{y\eta}$ of that due to direct selection with the same intensity of selection. The correlation $\rho_{y\eta}$ is the multiple correlation coefficient between y and the x's.

The following are the chief properties of the distribution of y after selection.

Frequency function: For $y_1 = y/\sigma_y$

$$f(y_1) = \frac{1}{a\sqrt{2\pi}}\, e^{-y_1^2/2} \int_{\frac{t-\rho y_1}{\sqrt{1-\rho^2}}}^\infty e^{-u^2/2} du \tag{8}$$

where t is the point on the abscissa of the normal curve above which a fraction a of the area lies, and ρ denotes $\rho_{y\eta}$.

Mean:

$$G(y) = \rho \frac{z}{a} \sigma_y \tag{9}$$

Variance:

$$V(y) = \sigma_y^2 \left[1 - \rho^2 \frac{z}{a}\left(\frac{z}{a} - t\right)\right] \tag{10}$$

Correlation between y and η:

$$\rho' = \rho \sqrt{\frac{1 - \frac{z}{a}\left(\frac{z}{a} - t\right)}{1 - \rho^2 \frac{z}{a}\left(\frac{z}{a} - t\right)}}. \tag{11}$$

The frequency function is positively skewed to a marked degree if ρ is high and a is small: otherwise skewness is only moderate and the general appearance is similar to that of a normal curve. Both the variance and the correlation between y and η are reduced by the selection.

Table I gives numerical data on some properties of the distribution of y after selection, for several values of ρ and a. Before selection, y is normally distributed with mean zero and unit standard deviation.

The values shown after selection are the mean, standard deviation, and the correlation with η. As the intensity of selection increases, the mean increases, the s.d. decreases (rather slowly), and the correlation between y and η decreases. The effects are in the same direction as ρ increases, except that ρ' increases with ρ.

6. Selection in two stages

Sometimes the measurements that seem useful for selection, because they are thought to be correlated with y, become available at different times. For example, in the selection of a hog as a sire, his weight at 180 days is known before we have records on the performance of his offspring, although the latter records seem more relevant to the purpose at hand. With dairy cows that are being selected for milk yield, each successive lactation provides new data. Consequently a selection program may involve repeated selections as more measurements accumulate.

TABLE I

PROPERTIES OF THE DISTRIBUTION OF y AFTER SELECTION IMPOSED ON η

Per Cent Selected 100 α	Mean			S.D.			ρ'		
	ρ			ρ			ρ		
	.5	.8	.95	.5	.8	.95	.5	.8	.95
50	.40	.64	.76	.92	.77	.65	.33	.63	.88
25	.64	1.02	1.21	.89	.72	.56	.27	.55	.83
10	.88	1.40	1.67	.89	.68	.50	.23	.48	.78
5	1.03	1.65	1.96	.89	.67	.47	.21	.44	.75
1	1.33	2.13	2.53	.88	.65	.43	.18	.38	.69

To consider selection in two stages, suppose that the variates $x_1, x_2, \ldots, x_q$, $(q < p)$ are known when the first selection is to be made, while the remaining variates $x_{q+1}, \ldots, x_p$ do not become known until the second selection is made. If the frequencies of selection α_1, α_2 at the two stages have been decided in advance, the optimum rule for selection is given by the previous theory. At the first selection, we use as a selection index the regression $\eta_1(x_1, x_2, \ldots, x_q)$ of y on $x_1, x_2, \ldots, x_q$. This regression is, by definition,

$$\eta_1(x_1, x_2, \ldots, x_q) = \int \ldots \int y f(y, x)\, dy dx_{q+1} \ldots dx_p,$$

where the integration for y and $x_{q+1}, \ldots, x_p$ extends over all the sample space. We select whenever $\eta_1 \geqq k_1$, where k_1 is chosen so that the frequency of selection is α_1.

At the second stage, the optimum index is the regression $\eta_2(x)$ of y on all the variables, in that fraction α_1 of the sample space which remains after the first selection. Since, however, the first selection operates purely on the x's and does not alter the frequency distribution of y in arrays in which all x's are fixed, $\eta_2(x)$ is exactly the same function as $\eta(x)$, the regression in the original population. We select whenever $\eta_2 \geqq k_2$, where k_2 satisfies the equation

$$\alpha_2 = \frac{1}{\alpha_1} \int_{\substack{\eta_1 \geqq k_1 \\ \eta \geqq k_2}} f_1(x)\, dx.$$

If all variates have zero means in the original population, the gain in y due to the two stage selection may be written formally as

$$G(y) = \frac{1}{a_1 a_2} \int_{\substack{\eta_1 \geqq k_1 \\ \eta \geqq k_2}} y f(y, x)\, dy dx .$$

The extension to three or more stages of selection is easily made.

This statement of the problem is not very realistic for most applications. It is more likely that only the product $a_1 a_2$, that is, the desired frequency of survivors of both selections, would be decided in advance. For given $a_1 a_2$, the question as to the optimum values of a_1 and a_2 is often asked in practice. In any specific case, this question can be answered from the equations above by trial and error, inserting various values of a_1 and a_2 to see which give the greatest gain in y. There does not seem to be a useful general solution in functional terms.

Even this form of the problem may not be what is wanted. For a given value of the product $a_1 a_2$, the cost of a selection program may vary according to the values of a_1 and a_2. If we decide to retain a group of hogs until information on their progeny is obtained rather than to sell them, we have to feed them in the intervening period with perhaps no compensating increase in their saleable value. The desirability of reserving judgment on a dairy cow for several lactations will depend on the amount of profit which her milk yields. Thus two stage selection problems usually have to be considered in the light of a specific cost situation, with the object of maximizing the gain in y for a given outlay.

As before, results become more definite if the variates follow a multivariate normal distribution. In this event, y, η_1 and η_2 are all normally distributed and jointly follow a trivariate normal. If the original multivariate normal distribution is given, the covariance matrix of this trivariate normal can be found. Since the gain in y is determined solely by the parameters of the trivariate distribution, there is no loss of generality in confining discussion to the trivariate distribution. In the original population, we may assume that all variates have zero means and unit variances. The parameters ρ_1, ρ_2 and ρ will denote the *simple* correlations between y and η_1, y and η_2, and η_1 and η_2, respectively.

The points of truncation k_1, k_2 satisfy the equations

$$a_1 = \frac{1}{\sqrt{2\pi}} \int_{k_1}^{\infty} e^{-\eta_1^2/2} d\eta_1 ; \tag{12}$$

$$a_1 a_2 = \frac{1}{2\pi\sqrt{1-\rho^2}} \int_{k_1}^{\infty} d\eta_1 \int_{k_2}^{\infty} e^{-[1/2(1-\rho^2)]\,\{\eta_1^2 - 2\rho\eta_1\eta_2 + \eta_2^2\}} d\eta_2 . \tag{13}$$

They can be found from the tables of the univariate and bivariate normal [6], respectively.

We now find the gain in y due to selection on η_1, followed by selection on η_2. Write

$$y = \beta_1\eta_1 + \beta_2\eta_2 + e ,$$

where $(\beta_1\eta_1 + \beta_2\eta_2)$ is the multiple regression of y on the η's in the original popula-

tion. If E' denotes a mean value in the selected part of the universe, we have

$$E'(y) = \beta_1 E'(\eta_1) + \beta_2 E'(\eta_2) + E'(e).$$

But the distribution of e is independent of that of the η's, so that $E'(e) = 0$. Hence

$$G(y) = \beta_1 G(\eta_1) + \beta_2 G(\eta_2), \tag{14}$$

so that we need only find $G(\eta_1)$ and $G(\eta_2)$. It is simpler to find $G(\eta_1 - \rho\eta_2)$ as follows.

$$\begin{aligned}
&a_1 a_2 G(\eta_1 - \rho\eta_2) \\
&\quad = \frac{1}{2\pi\sqrt{1-\rho^2}} \int_{k_2}^{\infty} d\eta_2 \int_{k_1}^{\infty} (\eta_1 - \rho\eta_2)\, e^{-[1/2(1-\rho^2)]\,\{\eta_1^2 - 2\rho\eta_1\eta_2 + \eta_2^2\}} d\eta_1 \\
&\quad = \frac{\sqrt{1-\rho^2}}{2\pi} \int_{k_2}^{\infty} e^{-[1/2(1-\rho^2)]\,\{k_1^2 - 2\rho k_1\eta_2 + \eta_2^2\}} d\eta_2 \\
&\quad = \frac{\sqrt{1-\rho^2}}{2\pi} \int_{k_2}^{\infty} e^{-[(\eta_2 - k_1\rho)^2/2(1-\rho^2)] - k_1^2/2} d\eta_2 \\
&\quad = (1-\rho^2)\left(\frac{e^{-k_1^2/2}}{\sqrt{2\pi}}\right)\left(\frac{1}{\sqrt{2\pi}} \int_{\frac{k_2 - \rho k_1}{\sqrt{1-\rho^2}}}^{\infty} e^{-t^2/2} dt\right) \\
&\quad = (1-\rho^2)\, z(k_1)\, I\left(\frac{k_2 - \rho k_1}{\sqrt{1-\rho_2}}\right) = (1-\rho^2)\, z_1 I_2, \text{ say},
\end{aligned}$$

where z and I denote the ordinate and the incomplete area of the normal curve, respectively. A corresponding equation holds for $G(\eta_2 - \rho\eta_1)$. Solving the two equations, we find

$$a_1 a_2 G(\eta_1) = z_1 I_2 + \rho z_2 I_1 \qquad a_1 a_2 G(\eta_2) = \rho z_1 I_2 + z_2 I_1 . \tag{14.1}$$

Hence, from (14)

$$\begin{aligned}
a_1 a_2 G(y) &= \beta_1 a_1 a_2 G(\eta_1) + \beta_2 a_1 a_2 G(\eta_2) \\
&= (\beta_1 + \rho\beta_2)\, z_1 I_2 + (\rho\beta_1 + \beta_2)\, z_2 I_1 .
\end{aligned}$$

But

$$\rho_1 = \operatorname{cov}\{y\eta_1\} = \operatorname{cov}\{(\beta_1\eta_1 + \beta_2\eta_2)\,\eta_1\} = \beta_1 + \rho\beta_2$$

and similarly for ρ_2. This gives for the final result

$$G(y) = \frac{\rho_1 z_1 I_2 + \rho_2 z_2 I_1}{a_1 a_2} \tag{15}$$

where it is to be noted that ρ_1, ρ_2 and ρ are *simple* correlation coefficients in the original population, and

$$I_1 = I\left(\frac{k_1 - \rho k_2}{\sqrt{1-\rho^2}}\right); \qquad I_2 = I\left(\frac{k_2 - \rho k_1}{\sqrt{1-\rho^2}}\right).$$

If y does not have unit standard deviation, the only change needed is to multiply the right side of (15) by σ_y.

In this proof no use has been made of the fact that η_2 is the population regression of y on the x's. The result therefore holds for two stage selection on *any* pair of

variates η_1 and η_2, provided that η_1, η_2 and y follow a trivariate normal. In practice the second selection is sometimes based only on those variates that were not available in time for the first selection, presumably because it is considered that there will be little further gain in bringing into the second index variates that have already been used.

If η_2 is the population regression of y on the x's, the proof can be shortened a little. From equation (6) of section (4)

$$\frac{G(y)}{\sigma_y} = \rho_{y\eta_2}\frac{G(\eta_2)}{\sigma_{\eta_2}} = \rho_2\frac{G(\eta_2)}{\sigma_{\eta_2}}.$$

This equation holds for any kind of selection based on the x's, so that it applies to two stage selection. But from (14.1)

$$a_1a_2\frac{G(\eta_2)}{\sigma_{\eta_2}} = \rho z_1I_2 + z_2I_1.$$

Hence

$$\frac{G(y)}{\sigma_y} = \frac{\rho_2(\rho z_1I_2 + z_2I_1)}{a_1a_2}. \tag{15.1}$$

This result is equivalent to (15). For if η_2 is the population regression, the partial regression of y on η_1, holding η_2 constant, is zero. Hence $\rho_1 = \rho_2\rho$, which makes (15) reduce to (15.1).

For specific applications, (15) can be computed from tables of the univariate normal distribution. In the case of three stage selection $G(y)$ can be expressed in terms of functions of the univariate and bivariate normals: the area of the trivariate distribution is however needed for reading k_3. The results (15) have been given in another form by Perotti [7].

7. An application to plant selection

As mentioned previously, the formulae in the preceding section are most likely to be useful in connection with specific applications. An example may help to clarify the procedure. In a program for finding superior varieties of a crop, it is quite common to start with a large number of varieties. A replicated field trial is conducted each year for several years, and at the end of each year some varieties are discarded. Suppose that we have a two year program, and at the end of two years we wish to retain only 1/24 of the original number of varieties, so that $a_1a_2 = 1/24$. The same number of plots is available for experimentation each year. It follows that if $a_1 = 1/a$, the varieties that survive to the second year can be tested in a times as many replications as the varieties in the first year. The problem is to find the best value of a_1 or a.

We will assume that we are trying to improve only a single characteristic of the crop, such as yield per acre, and are oblivious to all others, though this is somewhat of an oversimplification. Let y denote the true yielding ability of a variety. From the first year's experiment, we obtain an estimate $x_1 = y + e_1$, where x_1 is the observed mean yield of the variety and e_1 is the experimental error of x_1. At the end of the first year, a selection is made by means of x_1. For any variety which sur-

vives this, we obtain a second estimate $x_2 = y + e_2$ at the end of the second year. If experimental errors per plot are the same in the two years, we will have $\sigma^2_{e_2} = \sigma^2_{e_1}/a$.

The variates y, e_1 and e_2 are assumed to be normally and independently distributed, with zero means. Consequently, y, x_1 and x_2 follow a trivariate normal. It is easy to verify that, apart from a constant factor, the multiple regression of y on the x's is $(x_1 + ax_2)/(a + 1)$. This can be interpreted as the unweighted mean of all available observations on the yield, the apparent weight a arising because x_2 is based on a times as many replicates as x_1. From the general rule, selection at the end of the second year should be based on $(x_1 + ax_2)/(a + 1)$. Thus we may take

$$\eta_1 = x_1 = y + e_1; \qquad \eta_2 = \frac{(x_1 + ax_2)}{a+1} = y + \frac{e_1 + ae_2}{a+1}.$$

If $u = \sigma^2_{e_1}/\sigma^2_y$, it is found that

$$\rho_1 = \rho_{y\eta_1} = \frac{1}{\sqrt{1+u}}; \qquad \rho_2 = \rho_{y\eta_2} = \sqrt{\frac{a+1}{a+1+u}};$$

$$\rho = \rho_{\eta_1\eta_2} = \sqrt{\frac{a+1+u}{(a+1)(1+u)}}.$$

The symbol u measures the ratio of the error variance to the true genetic variance of a varietal mean in the first year. Calculations were made for $u = 1, 3, 15$ and 63. These values make the correlations ρ_1, between y and η_1, 0.707, 0.5, 0.25 and 0.125 respectively. The values of a were 1, 2, 3, 4, 6, 8, 12 and 24. The method of computation is first to find k_1 and k_2 from equations (12) and (13) of the previous section. The gain in y is then computed from equation (15). Results appear in table II.

TABLE II

GAIN IN y FOR VARIOUS METHODS OF TWO STAGE SELECTION

a_1	a_2	$u=1$ $\rho_1=.707$	$u=3$ $\rho_1=.5$	$u=15$ $\rho_1=.25$	$u=63$ $\rho_1=.125$
1	$\frac{1}{24}$	1.745	1.352	0.733	0.375
$\frac{1}{2}$	$\frac{1}{12}$	1.867	1.507	0.858	0.452
$\frac{1}{3}$	$\frac{1}{8}$	1.902	1.592	0.948	0.501
$\frac{1}{4}$	$\frac{1}{6}$	1.936	1.637	0.996	0.532
$\frac{1}{6}$	$\frac{1}{4}$	1.947	1.649	1.035	0.564
$\frac{1}{8}$	$\frac{1}{3}$	1.935	1.630	1.032	0.572
$\frac{1}{12}$	$\frac{1}{2}$	1.867	1.529	0.970	0.547
$\frac{1}{24}$	1	1.511	1.069	0.534	0.267

The best selection intensity for the first year appears to be fairly independent of the relative amounts of genetic and experimental error variation, the optimum being in the neighborhood of $a_1 = 1/6$ in all cases. Since the maxima are moderately flat, the use of equal selection rates of approximately 1/5 in both years would be a good simple rule.

The comparison of the optimum with the last line of the table is of interest. With $a_1 = 1/24$, all the selection is made in the first year, there being no need for a trial in the second year. When genetic variance is relatively low ($u = 15$ and $u = 63$), the optimum gain is about double that when $a_1 = 1/24$. This suggests that the

return per unit of work is about the same in the two years. As genetic variance increases, there is a diminishing return from the second year's work. In the limiting case $u = 0$ where there is no environmental variance (not shown in table II) there would be no return from a second year's work, since the maximum attainable gain $z(a)/a$, in this case 2.138, would be reached by a selection in the first year.

Comparison of the optimum with the top line of table II is also of interest. When $a_1 = 1$, all the selection is made at the end of the second year. The comparison is therefore one of a single versus a two stage selection, for the same total number of plots. The increase due to two stage selection is worthwhile though not spectacular. Since the gains for $a_1 = 1$ could have been obtained in the first year by growing double the number of plots, it might be questioned whether the increase from two stage selection compensates for a year's delay. On this issue the mathematical assumptions are too simple for application to plant breeding practice, where the true yielding ability of a variety would be expected to vary to some extent from year to year. Intuitively, this suggests that the advantage of a two year trial would be greater than that revealed by the comparison above.[1]

8. The construction of selection indices

The preceding theory assumes a knowledge of the exact form of the joint frequency distribution of y and the x's. In practice, such knowledge is rarely available. Instead, the general functional form of the joint distribution is assumed, with certain parameters introduced into the expression in order to give some degree of flexibility to the assumptions. The values of these parameters are estimated from some initial data. In practically all applications with which I am familiar, the functional form assumed has been the multivariate normal distribution. In this section the construction of selection indices for this distribution will be described: comment is reserved for later sections.

In some applications it is possible to obtain an initial sample in which the values of y and the x's are measured. For instance, if a group of tests is to be used to select personnel for some type of work, we might have available, from past data, records which show the performance of a sample of people both in the tests and in the subsequent work. In this event the procedure is the familiar one of calculating the least squares regression of y on the x's from this sample. The sample regression function $Y = \sum b_i x_i$ is taken as the selection index, and used to score new candidates.

In plant and animal selection, on the other hand, the value of y cannot be observed directly, and a more ingenious approach is needed. The situation is that y is a linear function $\sum a_i \xi_i$ of the important genetic characteristics ξ_i of the candidate, where the a_i are known weights. In the example cited by Fairfield Smith [8], to whom this approach is due, the a_i were determined by the relative economic values of improvements in the several characteristics. Alternatively, if we are interested only in a single characteristic, we take all a's zero except one. The variates ξ_i are presumed to follow a multivariate normal distribution. From experiments, we can

[1] An interesting discussion of the problem of single stage selection for sugar beets has been given by Y. Tang [13].

observe estimates x_i of the ξ_i, where

$$x_i = \xi_i + \epsilon_i ,$$

ϵ_i being the experimental error. The variates ϵ_i are also assumed to follow a multivariate normal distribution, and the joint distributions of the ξ_i and ϵ_i are taken as independent of one another. These assumptions are sufficient to ensure that y and the x's follow a multivariate normal. Consequently, from our previous theory, the best selection index for y that can be constructed from the x's is the population regression of y on the x's.

Let τ_{ij}, γ_{ij}, and ϵ_{ij} be the population covariances of x_ix_j, $\xi_i\xi_j$, and $\epsilon_i\epsilon_j$ respectively, so that

$$\tau_{ij} = \gamma_{ij} + \epsilon_{ij} . \tag{16}$$

Then if $\eta = \sum \beta_i x_i$ is the population regression of y on the x's, the β_i satisfy the equations

$$\sum_i \beta_i \tau_{ij} = \operatorname{cov}(x_j y) = \sum_k a_k \gamma_{jk} . \tag{17}$$

The theory above requires a knowledge of the parameters τ_{ij}, γ_{jk}. Estimates of these are obtained from an initial experiment. In this, we take a random sample of n members of the population, for example, n varieties of a crop, and conduct a replicated experiment in which all p measurements x_i are made on every plot. The design of the experiment may take various forms, but we will suppose that it is arranged in randomized blocks, with the following analysis of covariance between x_i and x_j.

	d.f.	Mean square	Unbiased estimate of
Between members (varieties)	$(n - 1)$	l_{ij}	$\tau_{ij} = \gamma_{ij} + \epsilon_{ij}$
Experimental error	$(n - 1)(m - 1)$	e_{ij}	ϵ_{ij}

This analysis is in terms of a varietal mean over the m replications. For simplicity, we have supposed that the selection is to be made from these means, so that the experimental error of a mean is the quantity ϵ_i which enters into the theoretical argument. From the analysis we can substitute unbiased estimates of τ_{ij} and γ_{jk} into (17). For the coefficients b_i in the estimated selection index $\sum b_i x_i$, this gives

$$\sum_i b_i l_{ij} = \sum_k a_k (l_{jk} - e_{jk}) , \qquad \text{for } j = 1, 2, \ldots, p . \tag{18}$$

Equations (18) are linear in the weights b_i, and their arithmetic solution proceeds by exactly the same methods as for a set of normal equations in least squares. As Bartlett [9] has pointed out, the sampling distribution of the b_i appears to be much more complicated than in least squares, because the right hand side of (18) consists of a linear function of covariances. This means that in any discussion of the sampling errors of a selection index that is computed from an initial sample, the situation in which y cannot be measured will require a separate investigation from that in which y can be measured.

The equations are slightly more elaborate if the means x_i from which the selection is to be made are based on m' replicates, whereas the means in the initial experiment by which the index was constructed are based on m replicates, $(m \neq m')$. The extension to this case has been given by Nanda [10]. In other applications, particularly in animal selection, a more complex analysis of covariance may be necessary to estimate the unknown covariances: on the other hand, the exact values of some of the correlation coefficients can be predicted from Mendelian theory, so that less remains to be estimated from the initial sample. The general structure of the equations of estimation is similar to that in equations (18), though the differences mentioned above become important in any investigation of sampling error theory.

9. The problem of constructing a best index from an initial sample

The procedure of substituting sample values for the unknown population parameters seems natural enough that it might be taken for granted as the obvious practical method for constructing a selection index. It does, however, raise the question: in what sense is a selection index $Y = \sum b_i x_i$, constructed from a sample, the "best" index? This is, of course, a more complex problem than the construction of a selection index when the values of the β_i are given. I have not been able to find any rule for constructing a "best" sample index, and I am doubtful whether one exists, unless a very specialized meaning of the word "best" is adopted.

In particular, in the case where the b_i are obtained from a least squares regression, the index $\sum b_i x_i$ does not have any obvious optimum properties, though this does not mean that an alternative which is superior can be found. In this connection, it may be noted that in practice we tend to act as if we distrust the least squares formula to some extent. It is a common procedure, starting with, say, 7 x-variables, to include in the index only those whose observed partial correlations with y in the initial sample are "large" enough, in some sense of this term. Some investigators retain only those x's whose partial correlations with y are statistically significant: others prefer a more flexible rule. Although such procedures are sometimes justified on the grounds that we want to keep the index simple, it is also thought, I believe, that the inclusion of the rejected x's would actually weaken the index, because the deleterious effect of sampling errors in the weights b_i would more than offset any gain that there might be in the correlation with y. When the β_i are known, on the other hand, the optimum rule shows that it pays to include any x whose partial correlation with y in the population differs from zero.

An introduction to the problem of finding a "best" sample index will be given for the case in which the initial sample is random, of size n, and provides data on y and the x's. An underlying multivariate normal distribution for y and the x's is assumed. Without loss of generality, we may suppose that *in the population* the only variable that is correlated with y is x_1, and that all x's are independently distributed with unit variance. Thus

$$(19) \qquad y = \beta_1 x_1 + e_1, \quad (\beta_1 > 0); \qquad \eta = \beta_1 x_1; \qquad \rho = \rho_{y\eta} = \frac{\beta_1}{\sigma_y}.$$

Consider any linear index $I = \sum w_i x_i$, where the w_i are known numbers. This

index is used repeatedly to select or reject new candidates, the w's remaining unchanged throughout. For these new candidates, the variates y and the x's are presumed to follow the same multivariate normal from which the initial sample was drawn. Hence for the new candidates the joint distribution of y and I is a bivariate normal in which

$$\text{(20)} \qquad \operatorname{cov}(yI) = w_1\beta_1: \qquad \sigma_I^2 = \sum_{i=1}^{p} w_i^2: \qquad \rho_{yI} = \frac{w_1\beta_1}{\sigma_y\sqrt{\sum w_i^2}}$$

since the x's are independently distributed with unit variance. Thus the average increase in y due to selection on I is

$$\text{(21)} \qquad G(y) = \rho_{yI}\frac{z(a)}{a}\sigma_y = \frac{w_1\beta_1}{\sqrt{\sum w_i^2}} \cdot \frac{z(a)}{a},$$

where the population over which the correlation is taken consists of an infinite number of selections made by the same index I.

If we are seeking the values of the w_i which give the best index, we might be inclined to choose them so as to maximize $G(y)$ in (21). This approach is fruitless. From the extreme right side of (21), we see that the maximizing values are $w_i = 0$, $i \geqq 2$, while w_1 can take any positive value. This solution confirms the general theorem that x_1, or any positive multiple of it, is the best index. But like the general theorem it requires knowledge of the β_i, since it could not be used unless we knew that $\beta_i = 0$, $i \geqq 2$.

Evidently we cannot get to the heart of the problem by considering the repeated use of a fixed index calculated from a single initial sample. It seems necessary to consider a two stage population in which (i) initial samples are repeatedly drawn, (ii) from each sample an index I is calculated by some rule and (iii) each index is then used repeatedly for selection. In such a population, the correlation ρ_{yI} will follow some frequency distribution. A rule for constructing w's which maximize the average ρ of ρ_{yI} might reasonably be considered the "best" rule, since it would also maximize the average $G(y)$ in equation (21). If we are willing to confine our attention to rules for which the w_i are linear functions of the y's in the initial sample, the frequency distribution of ρ_{yI}, though complex, does not look unmanageable, but I have not been able to express it in any form in which the consequences of different rules can be studied.

10. The loss due to the use of a least squares index

Some insight into the nature of the distribution of ρ_{yI} in the two stage population can be obtained if I is the least squares index $Y = \sum b_i x_i$. Consider the conditional distribution of ρ_{yY} in initial samples for which the values of the x's are fixed. For any fixed set of these x's, b_i is normally distributed. The mean value of b_1 is β_1 and the mean value of any other b_i is zero. Since, from the right side of (20),

$$\text{(22)} \qquad \rho_{yY}\sigma_y = \frac{b_1\beta_1}{\sqrt{\sum b_i^2}},$$

it follows that $\rho_{yY}\sigma_y$ is distributed as the ratio of a normal variate to the square root of a noncentral quadratic form in normal variables.

From standard regression theory, the covariance matrix of the b_i in the conditional distribution is $\sigma_e^2 S^{ij}$, where S^{ij} is the inverse of the matrix $S_{ij} = \sum x_i x_j$, taken over the initial sample. Thus in the conditional distribution the b_i have somewhat different variances, and are correlated to some extent.

Consider now an averaging of these conditional distributions over all possible initial samples. In this population all b_i have the same variance $\sigma_e^2/(n - p - 1)$. For by a familiar transformation the quantity S^{ii} may be written $1/S_{i\cdot jk}\ldots$, where $S_{i\cdot jk}\ldots$ is the sum of squares of deviations of x_i from its linear regression on the $(p - 1)$ other x-variables. Since the x's are normally and independently distributed, $S_{i\cdot jk}\ldots$ is distributed as χ^2 with $(n - p + 1)$ degrees of freedom. But if χ^2 has ν degrees of freedom, the average value of $1/\chi^2$ is known to be $1/(\nu - 2)$, from which the result follows. Further, in the unconditional distribution b_i and b_j may easily be shown to have zero covariance. The b's are not normally distributed, the distribution of $(b_i - \beta_i)$ being similar to Student's t-distribution.

An approximation to the distribution of ρ_{yY} may be obtained by regarding the b_i as normally and independently distributed with the same variance $\sigma_e^2/(n - p - 1)$, where

$$\rho_{yY} = \frac{b_1\beta_1}{\sigma_y\sqrt{\sum b_i^2}}. \tag{23}$$

Write $z_i = b_i/\sigma_b$, so that the z_i have unit variance. The mean value of z_1 is $\beta_1\sqrt{n - p - 1}/\sigma_e$. But

$$\beta_1 = \rho\sigma_y, \text{ from (19)}; \qquad \sigma_e^2 = \sigma_y^2(1 - \rho^2), \tag{24}$$

where ρ is the population multiple correlation coefficient between y and the x's. Hence $E(z_1) = \rho\sqrt{n - p - 1}/\sqrt{1 - \rho^2}$.

Under these assumptions, the quantity

$$t = \frac{z_1\sqrt{p - 1}}{\sqrt{\sum_2^p z_i^2}}$$

follows the noncentral t-distribution, with $(p - 1)$ degrees of freedom, and parameter $\tau = \rho\sqrt{n - p - 1}/\sqrt{1 - \rho^2}$. Thus from (23),

$$\rho_{yY} = \left(\frac{\beta_1}{\sigma_y}\right)\frac{z_1}{\sqrt{\sum_1^p z_i^2}} = \frac{\rho t}{\sqrt{t^2 + (p - 1)}}. \tag{25}$$

The form of the result (25) is of some interest. It shows that the correlation between y and Y equals the correlation between y and η, multiplied by a fraction which cannot exceed unity. The average value of this fraction therefore represents that fraction of the possible gain in y (possible if the β's were known) which will actually be attained. The quantities n and ρ enter into the result only in the form

$\rho\sqrt{n-p-1}/\sqrt{1-\rho^2}$. For a given number p of x-variates it follows that the initial sample size n must be much larger when ρ is small than when ρ is large if the same fractional loss in $G(y)$ is to be sustained.

We now give some calculations from which the size of the fractional loss in $G(y)$ can be estimated in specific cases. For $(p-1)$ degrees of freedom, the frequency distribution of the noncentral t may be written

$$(26) \qquad f(t_1)\,dt_1 = \frac{e^{-\tau^2/2}}{\sqrt{\pi}\left(\frac{p-3}{2}\right)!}\sum_{r=0}^{\infty}\frac{2^{r/2}\left(\frac{p+r-2}{2}\right)!}{r!\,(1+t_1^2)^{(p+r)/2}}(t_1\tau)^r dt_1,$$

where $t_1 = t/\sqrt{p-1}$, $\tau = \rho\sqrt{n-p-1}/\sqrt{1-\rho^2}$, $(p-1) > 0$.

The variate in which we are interested is $z = t_1/\sqrt{1+t_1^2}$. A routine transformation gives

$$(27) \quad \phi(z)\,dz = \frac{e^{-\tau^2/2}}{\sqrt{\pi}\left(\frac{p-3}{2}\right)!}\sum_{r=0}^{\infty}\frac{2^{r/2}\left(\frac{p+r-2}{2}\right)!}{r!}(\tau z)^r(1-z^2)^{(p-3)/2}dz.$$

From term by term integration, the mean value of z is found to be

$$(28) \quad E(z) = \frac{\bar{\rho}}{\rho} = \frac{\tau}{\sqrt{2}}\,e^{-\tau^2/2}\frac{\left(\frac{p-1}{2}\right)!}{\left(\frac{p}{2}\right)!}\left\{1+\frac{(p+1)}{(p+2)}\left(\frac{\tau^2}{2}\right)+\frac{(p+1)(p+3)}{(p+2)(p+4)}\frac{1}{2!}\left(\frac{\tau^2}{2}\right)^2+\cdots\right\}.$$

In table III the values of $E(z)$ are given for $p = 2, 3, 4$ and 5, and a series of values of the parameter $\tau^2/2$. This seems the most useful form for a succinct presentation, since by interpolation the reader can compute $\bar{\rho}/\rho$ for any case in which he is interested.

Example 1. Suppose that a regression on 4 x-variates is computed from an initial sample of size 10. If $\rho = 0.6$, what is the expected correlation between y and Y? In this case

$$\tfrac{1}{2}\tau^2 = \frac{1}{2}\frac{0.36}{0.64}(10-4-1) = 1.4.$$

By interpolation in the column $p = 4$, we find $\bar{\rho} = 0.64\rho$. Hence, on the average, the correlation between y and Y is $(0.64)(0.6) = 0.38$.

Example 2. In a selection index based on 5 x-variates, it is confidently expected that the true multiple correlation coefficient will be at least 0.7. An initial random sample for constructing the index is to be drawn. It is desired to take this large enough so that the fraction of the potential gain in y that is lost through errors in the b's will not exceed 5 percent. How large must n be?

We want $\bar{\rho}/\rho$ to be at least 0.95. From table III it is clear that a larger sample is needed for $\rho = 0.7$ than for any higher ρ. Hence we make the estimate for $\rho = 0.7$.

Since the entry 0.95 lies outside the limits of the table, we use the approximation

in the footnote. We want

$$\frac{1}{1+\frac{4}{\tau^2}} \geqq (.95)^2 = 0.9025 .$$

This leads to

$$\tau^2 = \frac{(.49)(n-6)}{.51} \geqq \frac{4}{\frac{1}{.9025}-1} = 37$$

and hence to $n = 45$.

To give more concrete results, the values of $\bar{\rho}/\rho$ are shown in table IV for initial samples of sizes 10 and 30, and for $\rho = 0.9, 0.7, 0.5$ and 0.3.

TABLE III*

VALUES OF $\bar{\rho}/\rho$

$\frac{1}{2}\tau^2 = \frac{\rho^2(n-p-1)}{2(1-\rho^2)}$	p = NUMBER OF x-VARIATES			
	2	3	4	5
0.2	.377	.324	.288	.262
0.4	.510	.441	.394	.360
0.6	.597	.521	.469	.430
0.8	.661	.581	.526	.484
1.0	.710	.629	.572	.528
1.5	.793	.714	.656	.611
2.0	.844	.770	.714	.670
3.0	.900	.838	.788	.747
4.0	.928	.876	.833	.796
5.0	.944	.900	.863	.830

* For values of $\frac{1}{2}\tau^2$ outside these limits, use the following approximations:

$$(\tfrac{1}{2}\tau^2 > 5): \qquad \frac{\bar{\rho}}{\rho} = \frac{1}{\sqrt{1+\frac{(p-1)}{\tau^2}}}$$

$$(\tfrac{1}{2}\tau^2 < 0.2): \qquad \frac{\bar{\rho}}{\rho} = \frac{\left(\frac{p-1}{2}\right)!}{\left(\frac{p}{2}\right)!}\frac{\tau}{\sqrt{2}}$$

The results illustrate the fact that when an index Y is computed by multiple regression from a small sample, the correlation between y and Y is likely to be substantially less than the multiple correlation coefficient between y and η. The gain in y following selection on Y is reduced in the same ratio. The loss in correlation increases with the addition of each independent variate, and when measured as a fraction of ρ, as in table IV, it increases rapidly as ρ becomes small. Since the results are derived from an approximation to the distribution of ρ_{yY}, not too much reliance can be placed on individual figures. But there seems little doubt that in small samples the addition of an extra x-variable to the index will not increase the gain due to selection unless it produces at least a moderate increase in ρ.

The results also suggest, as would be expected, that the decrease in the improvement in y can be avoided by the choice of an initial sample which is large enough.

When ρ is as high as 0.9, the initial sample may be quite small, at least if the index contains no more than 5 x-variables. Tables III and IV enable rough estimates to be made of the size of sample needed in specific cases. It should be noted that these tables apply to a *random* sample of size n. It is not uncommon to choose an initial sample in which the variation among the x's is substantially larger than that in a random sample of the same size. The purpose of this device is to decrease the sampling errors of the b's, and its effect is to make the appropriate size of sample for reading tables III and IV larger than the actual size.

TABLE IV

VALUES OF $\bar{\rho}/\rho$ FOR INITIAL SAMPLES OF SIZE 10, 30

ρ	$n=10$ $p=$ No. of x-Variates				$n=30$ $p=$ No. of x-Variates			
	2	3	4	5	2	3	4	5
0.9	.98	.96	.93	.90	1.00	.99	.99	.98
0.7	.91	.83	.74	.66	.98	.96	.94	.92
0.5	.74	.63	.54	.45	.94	.88	.84	.80
0.3	.67	.38	.31	.26	.77	.67	.61	.56

11. The effect of discarding variates from the index

As has been mentioned, the practice of discarding from the index those x-variates which appear to have little partial correlation with y may be regarded as an attempt to avoid some of the decrease in correlation which results from errors in the weights b_i. The practice seems obviously sound if we can be sure that the initial sample informs us correctly which variates to discard. In a small initial sample, however, the sample partial correlations may not be close to the corresponding population correlations, and the process of discarding is itself subject to errors. Precise information about the effects of such errors on ρ_{yY} would be worth having. In particular, it is relevant to discover whether the reduction in $G(y)$ through errors in discarding is as great as the reduction incurred if we do not discard. Such an investigation encounters intricate mathematics. Some results will be given for a simple case in which the analysis is not difficult.

We adopt the same mathematical framework as in the previous section. That is, in the two stage population the b_i are assumed as an approximation to be normally and independently distributed with variances $\sigma_e^2/(n - p - 1)$. Only x_1 is correlated with y, so that all β's except β_1 are zero. The index Y is to contain only one x. From any initial sample we select that x_i for which the corresponding b_i is greatest.

This method of discarding differs from methods that are common in practice in three respects: (i) we retain only one x_i, whereas all x's which have significant partial correlations with y are usually retained; (ii) in any specific initial sample, the conditional variances of the b_i will not all be the same, and it would be more customary to retain that x_i for which b_i gives the highest value of Student's t; (iii) in our problem we either retain an index x_1 which is actually the optimum index, or one which is of no value at all for selection, whereas in practice the choice would

probably lie among a number of variates each of which was of some value, although none was optimum. These deviations from practice were accepted in order to simplify the analysis. It does not seem that they distort the essentials of the problem unduly.

The frequency distribution of ρ_{yY} has only two values, ρ if x_1 is chosen and 0 otherwise. Hence we need consider only the probability that x_1 is chosen, that is, the probability that b_1 is the greatest (algebraically) of the b's. As before, write $z_i = b_i\sqrt{n-p-1}/\sigma_e$, so that the z_i have unit variances. The mean value of z_1 is $\tau = \beta_1\sqrt{n-p-1}/\sigma_e$, or $\rho\sqrt{n-p-1}/\sqrt{1-\rho^2}$. All other z's have zero means.

The method of calculating the probability P that z_1 is the greatest, which was suggested by J. W. Tukey, was to use the formula

$$P = 1 - (p-1)P_i + \frac{(p-1)(p-2)}{2!}P_{ij} - \frac{(p-1)(p-2)(p-3)}{3!}P_{ijk} + \ldots$$

where p is the number of x-variates from which the winner is chosen, and P_{ijk}, for example, is the probability that any three *specified* variables x_u, $u \geqq 2$, will exceed x_1. By symmetry, this probability is the same for any choice of the three. P_i is read from tables of the univariate normal distribution. P_{ij} is obtained by noting that

TABLE V

MEAN VALUES OF $\bar{\rho}/\rho$ FOR TWO METHODS OF CONSTRUCTING AN INDEX

ρ	ONLY THE "BEST" VARIABLE RETAINED						ALL x-VARIABLES RETAINED					
	$n'=8$			$n'=32$			$n'=8$			$n'=32$		
	p			p			p			p		
	2	3	4	2	3	4	2	3	4	2	3	4
0.9	1.00	1.00	1.00	1.00	1.00	1.00	.98	.97	.96	1.00	1.00	.99
0.7	.98	.95	.94	1.00	1.00	1.00	.92	.87	.83	.98	.97	.95
0.5	.88	.80	.72	.99	.98	.97	.77	.69	.63	.95	.91	.87
0.3	.74	.60	.49	.89	.82	.76	.51	.44	.39	.80	.72	.67

$(z_2 - z_1)$ and $(z_3 - z_1)$ follow a bivariate normal with means $-\tau$, variances 2, and correlation $+\frac{1}{2}$. The probability that both variates exceed zero is read from tables of the bivariate normal distribution. The higher P's necessitate numerical integration.

The probability P is shown on the *left* side of table V for $p = 2, 3, 4$; $\rho = 0.9$, 0.7, 0.5 and 0.3; and $n' = (n - p - 1) = 8, 32$. Since ρ_{yY} takes only the two values ρ and 0, the quantity P is also the mean value of $\bar{\rho}/\rho$, the same quantity as tabulated in tables III and IV. Fixed values of n' rather than n were used for convenience in calculation.

For the higher values of ρ, there is relatively little chance of failing to pick the best variate even with $n' = 8$, that is, with sample sizes of the order of 12. For $\rho = 0.5$ there is an appreciable loss in correlation with the smaller sample, but practically none with the larger sample. For $\rho = 0.3$ there is some loss even with the larger sample, which has about 36 observations.

The right side of table V shows the average value of $\tilde{\rho}/\rho$ when all x-variates are retained, as calculated by the approximate method given in the previous section. So far as it goes, the comparison supports the practice of attempting to discard the x-variates that do not seem to contribute to the correlation, since in all cases in which there is any appreciable loss, it is greater on the right side of the table.

12. Further problems

There is evidently much to be learned about the properties of a least squares index computed from an initial sample, as it affects the gain in y that may be expected from the use of the index for selection. The analysis should also be undertaken for indices that are computed by the use of estimated components of variance. The work of Bartlett [9] and Nanda [10] on this problem indicates its complexity and shows that the effective size of the initial sample is determined mainly by the number of varieties, rather than by the amount of replication for each variety. In this section a few additional problems are mentioned.

One concerns selection from nonnormal populations. The general theorem on optimum selection does not require normality. The two principal results which it provides are (i) $\eta(x)$ is the best selection index and (ii) the gain in y is equal to that in η. Consequently, in setting up a selection program in a population that is nonnormal, we should attempt to find out the shape of the regression $\eta(x)$ of y on the x's, and to study the frequency distribution of $\eta(x)$ in the unselected population. Although the formula $z(a)\sigma_\eta/a$ for the gain in η due to selection holds only if η is normally distributed, the correct formula is easily found if the frequency distribution of η is known.

In view of the widespread assumption of normality in applications, an investigation of the consequences of this assumption in nonnormal populations would also be worthwhile. In general, a linear index will not be the best index, and predictions of the expected gain in y, based on normal theory, are likely to be in error. Unfortunately it cannot be taken for granted that a moderate departure from normality will have little effect. This may be so if selection is not intense and y has only a small correlation with the x's, so that progress is slow. But in intense selection the gains depend primarily on the shapes of the tails of frequency distributions. As is well known, a frequency curve which looks quite similar to the normal curve may differ greatly in its tail. A combination of theoretical investigations with sampling experiments on natural populations is suggested.

Secondly, how accurately can the gain due to selection be estimated from given initial data? This question is important for policy making in plant and animal improvement, particularly at the present time, when the prospects of a steady increase in the world production of food are the subject of much study. Often, a program of selection is only one of a number of feasible means for improving quality or quantity, and its expected gains must be compared with similar estimates for other approaches. For a multivariate normal population, the expected gain in y is $\rho_{y\eta}\sigma_y z(a)/a$. The standard error of the estimate of this quantity from an initial sample has been given by Nanda [10], following earlier work by Bartlett [9], for the type of estimation that arises in plant selection. For practical purposes this

standard error must be regarded as a lower limit to the effective error, since disturbances due to nonnormality and to time changes in the population undergoing selection will presumably be present. Moreover, the estimated gain itself is the gain that would be attained if the true regression were known, and some reduction in this estimate to take account of sampling errors in the index may be required.

A third problem of a more specialized type arises because it is not always practicable to use a selection index in the best way. The optimum rule is to select all candidates for which $\eta > k$. When selection is made from small samples to fulfill some specific purpose, the number of candidates for which $\eta > k$ will vary from sample to sample. In some cases, however, each sample must provide a known quota of successful candidates. Consequently, we impose the restriction that the number selected from the i-th sample is to be r_i. This restriction decreases the expected gain in y and changes the mathematical aspects of the problem. The changes are easily made if there is only a single stage of selection. For the multivariate normal case, with r_i constant, the result is to replace the factor $z(a)/a$ by the factor $a(r, n)$, which is the average value of the largest r out of a standardized normal sample of size n. The extension to two stage sampling presents difficulties.

In this paper we have considered that the purpose of selection is to maximize the mean value of y while retaining a specified fraction a of the members of the original population. Birnbaum [11] and Birnbaum and Chapman [12] investigate the related problem of maximizing the fraction of the population that is retained, subject to the condition that the mean value of y in the selected universe has some preassigned value. For a multivariate normal population, they show that truncation by means of the linear regression of y on the x's is optimum for this problem also.

In conclusion, these problems may leave the impression, not incorrectly, that more issues have been raised than solved in this paper. The topic of selection appears to be one where the applications have run somewhat ahead of their theoretical basis, and it may be anticipated that any new advances in theory will quickly be utilized.

REFERENCES

[1] K. PEARSON, "On the influence of natural selection on the variability and correlation of organs," *Roy. Soc. London Phil. Trans.*, A, Vol. 200 (1903), pp. 1–66.

[2] J. L. LUSH, "Family merit and individual merit as basis for selection," *Amer. Naturalist*, Vol. 81 (1947), pp. 241–261 and 362–379.

[3] G. E. DICKERSON and L. N. HAZEL, "Effectiveness of selection on progeny performance as a supplement to earlier culling in livestock," *Jour. Agr. Res.*, Vol. 69 (1944), pp. 459–476.

[4] W. T. FEDERER and G. F. SPRAGUE, "A comparison of variance components in corn yield trials," *Jour. Amer. Soc. Agronomy*, Vol. 39 (1947), pp. 453–463.

[5] J. NEYMAN and E. S. PEARSON, "On the problem of the most efficient tests of statistical hypotheses," *Roy. Soc. London Phil. Trans.*, A, Vol. 231 (1933), pp. 289–337.

[6] K. PEARSON, *Tables for Statisticians and Biometricians*, Part 2, Cambridge University Press, Cambridge, 1931.

[7] J. M. PEROTTI, "Mean improvement in a normal variate under direct and indirect selection," M. Sc. Thesis, Iowa State College, 1943.

[8] H. FAIRFIELD SMITH, "A discriminant function for plant selection," *Annals of Eugenics*, Vol. 7 (1936), pp. 240–250.

[9] M. S. BARTLETT, "The standard errors of discriminant function coefficients," *Jour. Roy. Stat. Soc., Suppl.*, Vol. 6 (1939), pp. 169–173.

[10] D. N. NANDA, "The standard errors of discriminant function coefficients in plant breeding experiments," *Jour. Roy. Stat. Soc.*, B, Vol. 11 (1949), pp. 283–290.

[11] Z. W. BIRNBAUM, "Effect of linear truncation on a multinormal population," *Annals of Math. Stat.*, Vol. 21 (1950), pp. 272–279.

[12] Z. W. BIRNBAUM and D. G. CHAPMAN, "On optimum selections from multinormal populations," *Annals of Math. Stat.*, Vol. 21 (1950), pp. 443–447.

[13] Y. TANG, "Certain statistical problems arising in plant breeding," *Biometrika*, Vol. 30 (1938), pp. 29–56.