Theor Appl Genet (1983) 66:23–28

A comparison of exact and sequential methods in multi-stage index selection *

A. M. Saxton

Department of Animal Science, P.O. Box 5127, North Carolina State University, Raleigh, NC 27650, USA

Received February 5, 1982; Accepted February 15, 1983
Communicated by D. Van Vleck

Summary. The theory of sequential multi-stage index selection makes an implicit assumption that the correlation between indices at different stages is zero. This assumption was shown to result in errors in the estimation of genetic gain and in the proportion of the population selected by truncating the joint distribution of the indices. Knowledge of the means and volumes of truncated multivariate normal distributions was used to correct these estimates. Effects of selection intensity and the correlation between the first and second stage indices (ϱ) on the accuracy of the approximate sequential method were examined. Computational constraints limited this analysis to two-stage index selection procedures. The sequential method performed well for ϱ less than 0.6 but accuracy deteriorated rapidly as ϱ increased beyond this value. The effect of selection intensity on accuracy was smaller than ϱ. On a percentage basis, errors in actual percent selected and underestimation of genetic gain increased with selection intensity while overestimation decreased. The types of errors which occur and their magnitude depend on the intensity of first stage selection.

Key words: Independent culling – Multi-stage selection – Sequential selection

Introduction

Multiple stage selection procedures are an important technique in animal breeding, and are often used in practice without regard for the theoretical implications. For example, an initial culling often occurs before putting animals on a more complex test procedure, due to limited facilities. The effect of this initial culling on the variances and on the prediction of gains could be significant. The initital culling should then be considered as part of the selection program, and a two-stage selection program results. While the theoretical foundations of multi-stage selection are still being developed, many situations can be analyzed with present knowledge.

The idea of several stages of selection appears to have begun with the work of Dickerson and Hazel (1944), who explored the use of individual or family performance information in the first stage and progeny test information in the second stage for improving a single trait. Jain and Amble (1962) extended single-trait selection to an arbitrary number of stages with no restriction on the type of information used at each stage, using the theory developed by Cochran (1951). Namkoong (1970) considered the problem of optimum allocation of selection intensity in two stages of selection for a single trait, with the addition of a cost function to account for the costs of obtaining the second stage information.

When two traits are being selected, the theory developed for independent culling by Young and Weiler (1960) can be used. Predicted gains are calculated by assuming a simultaneous truncation selection on each of the two traits, but in practice the selections are often performed at different times, particularly if the traits are measured at different ages. Young (1964) extended the theory to permit selection of more than two traits and to allow, within each stage, selection of more than one trait. Young (1964) also introduced the idea of using part and whole indices for the two traits to be independently culled, and stated that expected gain would be greatest when all intensity

* Paper No. 8157 of the Journal Series of the North Carolina Agricultural Research Service, Raleigh, NC, USA

was on the second stage, where the whole index was being used as the criterion for selection. The use of such part and whole indices is expected to be more efficient than the usual independent culling method because the indices use all available information.

In contrast to the 'exact' solution given by independent culling theory, several authors have developed an approximate sequential method for selection in several stages. Cotterill and James (1981) put independent culling selection in a sequential two-stage form and considered multiple sources of information at the second stage. Cunningham (1975) used the part and whole index concept of Young (1964) to develop a method for sequential multi-stage index selection. Using Cunningham's algorithm, Bartels et al. (1980) found that the relative efficiency of multi-stage index selection could be greater than the classical single-stage selection index. This result is contrary to the intuitive expectation of Young (1964).

The idea of sequential selection was developed to ease the computational burden created by multivariate distributions, but results in estimates of genetic gain which are known to be approximations. The present paper will examine how close these approximations are by comparing the algorithm of Cunningham (1975) with exact results obtained from a modified computer program for independent culling selection (Saxton 1982).

Multi-stage index selection

For convenience, the theory of sequential and exact selection methods will be briefly reviewed. Following the development of Cunningham (1975) let

P be the n by n phenotypic variance-covariance matrix, let
G be the n by m covariance matrix between the n traits in the index and the m traits in the aggregate genotype (H), and let
C be the m by m genotypic variance-covariance matrix.

Selection procedure 4 suggested by Cunningham (1975) will be used throughout this paper. This procedure selects in the first stage upon an index (I_1) incorporating a subset of the n traits, without loss of generality assumed to be the first r traits. In the second stage selection is on an index (I_2) incorporating all n traits.

In sequential selection, a proportion of the population, p_1, is selected by truncation selection on I_1, using the usual univariate selection theory. Variances and covariances are adjusted for this first stage selection (Cochran 1951). Truncation selection on I_2 is then performed, selecting a fraction p_2 of the remaining population, such that $p_1p_2 = S$, where S is the desired fraction of the population to be selected. Univariate theory is used for this second stage selection also, and thus the complexities of truncated multivariate normal distributions are avoided. This requires an assumption that the bivariate normality between I_2 and the aggregate genotype is not seriously affected by selection on I_1 or equivalently, that the phenotypic correlation between I_1 and I_2 is zero. Total gain is calculated by adding the estimates of gain from the two stages.

Table 1. Variances and covariances among the selection indices (I_1, I_2) and the aggregate genotype (H)

Definition	Formula
Variance of H	$v' C v$
Variance of I_1	$b_1' P_1 b_1$
Variance of I_2	$b' P b$
Covariance between I_1 and I_2	$b' P^* b_1$
Covariance between I_1 and H	$b_1' G_1 v$
Covariance between I_2 and H	$b' G v$

P^* is the n by r submatrix of P

Exact results are obtained by using independent culling selection on the two indices. For the first stage, index weights are given by

$$b_1 = P_1^{-1} G_1 v,$$

where P_1 is the r by r matrix consisting of the first r rows and columns of P, G_1 is the r by m matrix containing the first r rows and all m columns of G and v is the vector of economic weights. The second stage index weights are of course given by

$$b = P^{-1} G v.$$

Parameters needed for independent culling can now be calculated according to the formulae in Table 1. Finally, define the fraction selected as

$$S = \lambda (2\Pi)^{-1} \int_{t_2}^{\infty} \int_{t_1}^{\infty} \exp [-\lambda^2 (x^2 - 2 \varrho xy + y^2)/2] \, dx \, dy$$

where

$\lambda = (1-\varrho^2)^{-1/2}$,
ϱ = correlation between I_1 and I_2,
t_1 = truncation point for I_1 and
t_2 = truncation point for I_2.

A computer program was written which sequentially took values of t_1 over its possible range, calculated the t_2 which results in selecting a fraction S of the population (Saxton 1982) and then computed the gain in the aggregate genotype resulting from truncating I_1 and I_2 at t_1 and t_2, respectively (Tallis 1961). These results are exact, given the common assumptions of initial multivariate normality and an infinite population.

The exact and sequential methods differ only in the assumption of normality in the second stage population required by the sequential method. Therefore, it is expected that only parameters which influence this normality will affect the accuracy of the sequential method. These parameters are the correlation between I_1 and I_2 and the intensity of selection.

Examples

The first example illustrates the types of errors which can occur in using the sequential method. Data given in Cunningham (1975) are used and it is assumed that 6% of the population is to be selected. There are five traits available and the first stage index is based on the first three traits. Table 2 gives two points of comparison between the exact algorithm and the sequential method of Cunningham (1975). Method 1 is the usual single-

Table 2. Comparison of approximate sequential and exact methods using data from Cunningham (1975). The two-stage methods involve selecting the first three traits at stage one and all traits at stage two. See text for further discussion

Method	Truncation points		% selected		SD of indices		Gain in aggregate genotype ΔH
	t_1	t_2	p_1	S	σ_{I_1}	σ_{I_2}	
1. Single-stage index	–	1.555	–	6.0	–	4.21	8.36
2. Two-stage							
Sequential	0.305	1.003	38.0	15.3	2.21	2.48	7.88
Exact	0.305	1.555	38.0	6.0	4.05	4.21	8.36
3. Alternative two-stage							
Sequential	1.255	−0.184	10.5	10.4	1.68	2.03	8.41
Exact	1.255	1.525	10.5	6.0	4.05	4.21	8.33

stage selection index, which will be used as a standard for comparing the two-stage selection procedures. Method 2 assumes a first-stage truncation value of 0.305 σ_{I_1}. This method duplicates Cunningham's selection procedure 4 results, but shows that the sequential approach fails to select the desired proportion of the population ($\varrho = 0.96$). Genetic gains are underestimated.

The truncation value of 1.255 σ_{I_1} in Method 3 was chosen as the point which maximized the genetic gain predicted by the sequential calculations. This was obtained by scanning the range of possible values for t_1. Again, the sequential algorithm failed to select the desired proportion of the population, but genetic gains were overestimated, in fact exceeding the classical selection index.

Genetic gains predicted by the sequential method appear to be affected by several sources of bias, including errors in the proportion of the population selected, loss of normality, and correlation between estimates of gains in the two stages which makes the estimate of genetic gain, $i_1 \sigma_{I_1} + i_2 \sigma_{I_2}$, incorrect. These biases can all be attributed to the correlation between I_1 and I_2. The effect that these biases have across the range of possible selection schemes is shown in Fig. 1 (A). The points in Table 2 are just points along these curves. The results show that the sequential method can underestimate or overestimate the true genetic gain by as much as 9% and 1%, respectively.

Still assuming 6% of the population is to be selected, the pair of lines in Fig. 1 (B) gives the predicted genetic gains for the two methods when the order of trait selection is reversed. That is, the last two traits are selected in the first stage index instead of the first three traits. The last two traits have a lower correlation with the aggregate genotype, as shown by the much reduced gain when all selection pressure is put on the first stage index ($t_1 = 1.55$). It follows that the correlation between this first stage index and I_2 will be reduced, and in fact the correlation dropped from 0.96 to 0.80. The reduced correlation explains the observed improvement in the accuracy of the gains predicted by the sequential method.

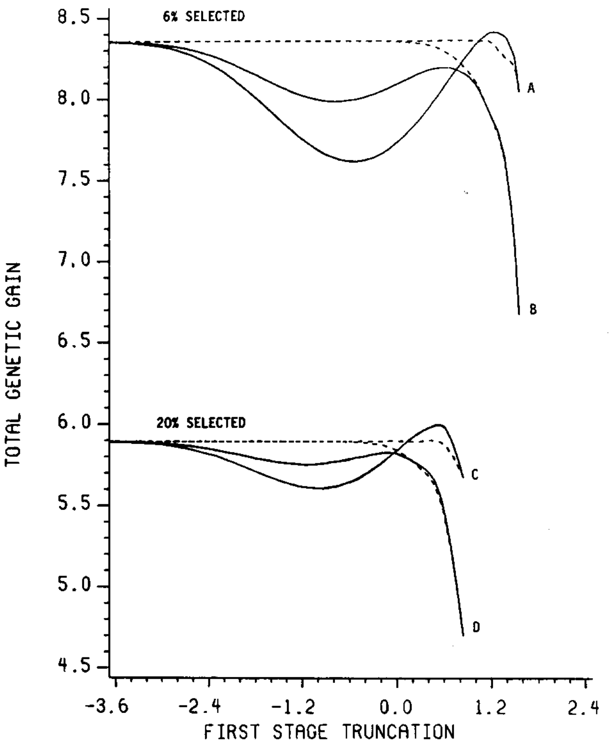

Fig. 1. The approximate sequential and exact selection methods are compared for various selection schemes. Curves show the effect of the amount of selection pressure applied at the first stage on total genetic gain. Curves ending at points *A* and *C* are for a first-stage index containing the first three traits, while curves *B* and *D* represent the results when the first-stage index contains the last two traits. Solid lines are gains predicted by the approximate sequential method, dotted lines are gains predicted by the exact method

Table 3. The effect of selection intensity and the correlation between I_1 and I_2 (ϱ) on the accuracy of the approximate sequential method

Selection intensity (%)	ϱ	Range[a]	Errors[b]			Observed[c] selection intensity (%)	Cor (E, S)[d]
			% under	% over	% excess		
0.1	0.1	450.3 – 45.2	0.0	0.2	0.0	0.12	1.00
	0.3	150.4 – 45.2	0.3	0.0	0.0	0.17	1.00
	0.6	75.1 – 45.2	3.2	0.0	0.0	0.31	0.99
	0.8	56.4 – 45.2	8.6	0.0	0.0	0.53	0.60
	0.9	50.1 – 45.2	13.0	0.2	0.0	0.83	– 0.01
	0.95	47.5 – 45.2	17.0	0.2	0.0	1.21	– 0.15
	0.99	45.6 – 45.2	18.9	0.6	0.5	2.08	– 0.12
5	0.1	276.7 – 27.7	0.0	0.0	0.0	5.4	1.00
	0.3	92.2 – 27.7	0.1	0.0	0.0	6.4	1.00
	0.6	46.1 – 27.7	1.8	0.0	0.0	8.4	1.00
	0.8	34.6 – 27.7	4.8	0.3	0.0	10.8	0.89
	0.9	30.7 – 27.7	7.3	0.5	0.0	13.1	0.30
	0.95	29.1 – 27.7	9.0	0.9	0.2	15.3	– 0.07
	0.99	27.9 – 27.7	10.6	1.6	1.6	18.8	– 0.19
20	0.1	187.8 – 18.8	0.0	0.0	0.0	20.9	1.00
	0.3	62.6 – 18.8	0.1	0.0	0.0	22.9	1.00
	0.6	31.3 – 18.8	0.9	0.2	0.0	26.5	1.00
	0.8	23.5 – 18.8	2.4	0.6	0.0	30.4	0.96
	0.9	20.9 – 18.8	3.8	1.2	0.0	33.8	0.55
	0.95	19.8 – 18.8	4.6	1.8	1.0	36.6	0.03
	0.99	19.0 – 18.8	5.5	2.7	2.6	40.9	– 0.14
60	0.1	86.4 – 8.6	0.0	0.0	0.0	60.7	1.00
	0.3	28.8 – 8.6	0.0	0.1	0.0	62.2	1.00
	0.6	14.4 – 8.6	0.0	0.3	0.0	65.2	1.00
	0.8	10.8 – 8.6	0.0	1.6	0.7	68.2	0.98
	0.9	9.6 – 8.6	0.2	3.1	2.4	70.6	0.78
	0.95	9.1 – 8.6	0.3	4.0	4.0	72.5	0.33
	0.99	8.7 – 8.6	0.4	5.4	5.4	75.2	0.06

[a] Range in predicted genetic gain (exact method) from all weight on I_2 to all weight on I_1

[b] Maximum observed (across t_1) deviation of the sequential from the exact method as a percentage of the exact predicted gains; % under = maximum underestimation; % over = maximum overestimation; % excess is the maximum gain predicted by the sequential method in excess of the theoretical limit (all weight on I_2) as a percentage of this limit. If the maximum predicted gain did not exceed the limit, a value of zero is given

[c] Maximum percent actually selected by a particular sequential selection program

[d] Correlation between the exact and sequential predicted gain curves

The results for 20% selected (Fig. 1, C and D) show a pattern similar to the results above, but the biases in predicted gain were reduced on a percentage basis. Errors ranged from –5% to +2% deviation from the exact results. Selection of the desired proportion was also improved.

For a more detailed analysis of the effect that ϱ and selection intensity have on the accuracy of the sequential method, a deterministic simulation was done on a simple set of values given by

$$Pb=\begin{pmatrix}10 & c\\ c & 10\end{pmatrix} \quad b=\begin{pmatrix}20\\ 20\end{pmatrix}=Gv.$$

Multivariate normality and an infinite population size were assumed. The value of c was chosen to induce the desired correlation between I_1 (the first trait) and I_2. As expected, the actual values for the phenotypic variances and Gv vector did not affect the results. This was verified numerically by examining several other values for these parameters and also by examples of three trait systems. For each combination of ϱ and selection intensity, one computer run was made. Each run generated genetic gain predictions by the exact and sequential methods across the range of possible first stage truncation values. Comparisons between the two curves were then made.

Results of the comparison are given in Table 3. The range of genetic gain predicted by the exact method for each problem is presented, going from all weight on I_2 (the theoretical maximum) to all weight on I_1. These genetic gain curves have forms identical to those seen in Fig. 1 for the exact method. Gains predicted by the sequential method also follow the pattern in Fig. 1, with first a possibility of underestimation, then overestimation and perhaps even predicted gains exceeding the theoretical maximum. The curves in Fig. 1 show that the magnitude of these errors depends on the first stage truncation point. The maximum observed value for each type of error is given in Table 3. This is one way of characterizing the accuracy of the sequential method.

Maximum error values given in Table 3 indicate that underestimation is more important at high selection intensities while overestimation occurs more frequently at low selection intensities. Both under and overestimation increased with ϱ. Predicted gains exceeding the theoretical maximum only occurred for large ϱ and increased on a percentage basis as selection intensity decreased. The sequential method never selected less than the desired proportion of the population, but could select more than the specified percent for intermediate values of t_1. Errors in this category increased with ϱ and selection intensity. The values in Table 3 for errors of estimation and selection represent the worst case found across the range of t_1 values. Thus in practice, any particular selection program with a given t_1 will generally have errors smaller than these values. Also it should be noted that underestimation and overestimation will not occur simultaneously in the same selection program.

As an overall indicator of the correspondence between the two methods, values of Cor(E, S) were calculated. These values are simply the Pearson correlation between points taken at equal intervals along the curves of predicted genetic gain. The points started from $t_1 = -3.6$ and ranged up to the point where all weight was on I_1. At least 30 points were included in each comparison. The results give a clear picture of the decrease in accuracy of the sequential method as ϱ increases.

It should be emphasized that these results are expected to hold for all problems with the given ϱ and selection intensity. However, for numerical reasons, there is some variation in the percent error and Cor(E, S) statistics presented here. Variation in Cor(E, S) can be attributed to the choice of points for calculation of the correlation. Variance in percent errors is caused by inexact determination of the first stage truncation where the maximum percent error occurs. The percentages and Cor(E, S) can effectively be regarded as having SE of 0.1 and 0.05 units respectively.

Discussion

In multi-stage selection there are more questions to be answered than simply estimation of genetic gain. Given a total selection intensity, a more basic question is how to partition the selection intensity among the stages. Optimal methods balance the economic savings from culling in the first stage with the loss in total genetic gain caused by this initial selection. Optimal methods also choose variables to be selected in the first stage which maximize the correlation between I_1 and I_2, within biological limits. As can be seen in Fig. 1, as ϱ increases higher selection intensity can be placed on the first stage without loss of total genetic gain. Cotterill and James (1981) also recognized the importance of order of trait selection.

The sequential method predicted genetic gains quite well below values for ϱ of 0.6. Above this value, errors in prediction got large, but it is an individual decision as to the significance of these errors. Since in practice all predicted gains are approximations and since the values in Table 3 are maximum errors, it could be argued that, in general, the gains predicted by the approximate sequential method are adequate for practical use. Note that predicted gain larger than theoretically possible, as observed by Bartels et al. (1980), was confirmed to be a possibility if the sequential method is used.

Of more concern is the difficulty which would occur in trying to choose an optimal multi-stage selection program. If the approximate sequential results were taken at face value, the optimal first stage truncation would generally be large. Depending on the economic parameters of the selection program, it is conceivable that the sequential method could lead to a program significantly different from the theoretical optimum.

The sequential method also leads to truncation values which may not select the desired proportion of the population. As much as twenty times the desired proportion can be selected, but recall that these errors are maximum errors. At the end points, where most of the selection pressure is on I_1 or I_2, the sequential method does well, since essentially univariate selection is being performed.

Arguments concerning the computational simplicity of the sequential method are of little importance for the two stage case since computers can perform all the calculations. However, if more than two stage programs are contemplated, the exact computations would have to be done by hand. This is due to the lack of generally available subroutines for calculating normal density volumes with three or more dimensions.

The purpose here is not to recommend one method over the other, but rather to point out when and what types of errors can arise from the approximation. Below a correlation between I_1 and I_2 of 0.6, the differences are very small. Above this value caution should be exercised when using the approximation, particularly at higher selection intensities.

Acknowledgement. I would like to thank Dr. E. J. Eisen for helpful discussions concerning this problem and reviewers for improving the presentation.

References

Bartels VS, Bruns E, Glodek P (1980) Anwendungsmöglichkeiten der Mehrstufenselektion bei der Züchtung von Göttinger Miniaturschweinen. Z Tierz Züchtungsbiol 97: 101–115

Cochran WG (1951) Improvement by means of selection. In: Newman J (ed) Proc 2nd Berkeley Symp Math Stat Prob, Berkley, Calif, pp 449–470

Cotterill PP, James JW (1981) Optimising two-stage independent culling selection in tree and animal breeding. Theor Appl Genet 59:67–72

Cunningham EP (1975): Multi-stage index selection. Theor Appl Genet 46:55–61

Dickerson GE, Hazel LN (1944) Effectiveness of selection on progeny performance as a supplement to earlier culling in livestock. J Agric Res 69:459–476

Jain JP, Amble VN (1962) Improvement through selection at successive stages. J Indian Soc Agric Stat 14:88–109

Namkoong G (1970) Optimum allocation of selection intensity in two stages of truncation selection. Biometrics 26: 465–476

Saxton AM (1982) A note on a computer program for independent culling. Anim Prod 35:295–297

Tallis GM (1961) The moment generating function of the truncated multi-normal distribution. J R Stat Soc B 23: 223–229

Young SSY, Weiler H (1960) Selection for two correlated traits by independent culling levels. J Genet 57:329–338

Young SSY (1964) Multi-stage selection for genetic gain. Heredity 19:131–145