

The spread of genes in random mating control populations

BY J. W. JAMES

Department of Animal Husbandry, University of Queensland, Brisbane

(Received 10 June 1960)

I. INTRODUCTION

The maintenance of unselected random mating control populations to serve as a control in the evaluation of experimental breeding procedures is now attracting considerable interest. The objective is to keep these populations genetically constant, but because they are necessarily limited in size, some genetic drift will occur. The magnitude of this random variation in gene frequency affects the efficiency of the population as a control.

Gowe, Robertson & Latter (1959) have recently discussed two designs for control populations and the genetic drift expected from either design. The 'pedigreed flock' is one in which every attempt is made to choose one male from each sire family and one female from each dam family as parents of the next generation. The 'random breeding flock' is one in which the required number of males and females are chosen at random from all available offspring.

Crow & Morton (1955) derived an expression for the sampling variance of gene frequency in a population, taking into account unequal sex ratio and variation in family size. Their definition of gene frequency as a weighted average of frequencies in the two sexes is not appropriate to experimental situations, because in experimental populations the frequency in each sex is of equal importance, regardless of the actual sex ratio. Latter (1959) derived an expression for genetic drift in random breeding flocks based on an unweighted average for the sexes, and this expression was used by Gowe, Robertson & Latter (1959) in their discussion.

Random changes in gene frequency can be regarded as occurring in two stages. Starting from a given group of breeding individuals, gene frequency may change through sampling of gametes from these individuals in the formation of offspring. Following this, gene frequency may change again when the next set of breeding individuals is chosen. The actual point of division here is somewhat arbitrary. This second sampling stage could, for example, be further subdivided, into a random change occurring when it is decided which individuals are to be included in the breeding operation, and a random change occurring when matings are made up. The magnitude of this last change would be affected by operational procedures, such as whether each individual of a given sex was allotted the same number of mates, whether the numbers of mates per individual were allotted at random, or whether mass matings were used. Changes of this nature can with advantage be assimilated into the first class, changes due to sampling of gametes, the differences

between the various procedures being then described by the distribution of family sizes. Changes in the second class are then regarded as complete when the individuals to be included in the breeding programme are chosen. General descriptions applying to all operational procedures are then possible.

In dealing with this second stage of sampling, Latter (1959) based his discussion on the assumption that each chosen individual is sampled at random from all surviving individuals of the same sex, i.e. sampling with replacement. In experimental populations, the actual procedure is to choose a set of parents at random from among all possible such sets. This is sampling without replacement, and can be expected to lead to a slightly smaller sampling variance, because in sampling with replacement there is a finite probability that any individual may be sampled more than once, whereas this is not true under sampling without replacement. In this paper, genetic drift when parents are sampled without replacement will be considered.

James & McBride (1958) analysed a poultry flock under selection by means of changes in 'percentages of genes' from various ancestors. Where pedigrees are available, this technique can be applied to control populations, with the difference that changes in percentages of genes are due to chance instead of selection. Since sampling of parents contributes to genetic drift, changes in percentages of genes will be related to genetic drift, and this relationship will also be investigated in this paper.

A model of a random mating population based on sampling without replacement is now put forward, and, on the basis of this model, equations are derived for the sampling variances of gene frequency, G_s and G_D , where G_s is the proportion of genes from a sire in the matings of his progeny, and G_D is the corresponding term for a dam.

II. DERIVATIONS

It is assumed that there are T individuals (male and female) available for choice as parents of the next generation. These T individuals are the offspring of m sires and f dams which have been mated at random, the i th sire, S_i , contributing k_i males and l_i females, K_i in all, and the j th dam, D_j , contributing a_j males and b_j females, A_j in all. Then $m\bar{k}$ ($=f\bar{a}$) is the number of males and $m\bar{l}$ ($=f\bar{b}$) is the number of females available, so that

$$m\bar{k} + m\bar{l} = f\bar{a} + f\bar{b} = m\bar{K} = f\bar{A} = T.$$

A set of M males is then chosen at random from among all possible such sets which may be formed from the $m\bar{k}$ males, and a set of F females similarly chosen from the $m\bar{l}$ females available. S_i contributes v_i sons and w_i daughters, and D_j contributes x_j sons and y_j daughters to these sets. Then

$$m\bar{v} = f\bar{x} = M,$$

$$m\bar{w} = f\bar{y} = F.$$

Latter (1959) has shown that if δq_m is the change in gene frequency from sampling offspring of males, and δq_f the change from sampling offspring of females, then

$$V(\delta q_m) = \frac{q(1-q)}{8} \left[\frac{1}{M} + \frac{1}{F} + \frac{mV(v)}{M^2} + \frac{mV(w)}{F^2} + \frac{2m\text{Cov}(vw)}{MF} \right],$$

with a corresponding equation for $V(\delta q_f)$. In these equations the symbols V and Cov stand for the variance and covariance respectively of the bracketed terms associated with them. Further, if Δq is the total change in gene frequency,

$$V(\Delta q) = \frac{1}{4} [V(\delta q_m) + V(\delta q_f)].$$

To predict the genetic drift from the model above, it is therefore necessary to obtain variances and covariances of v , w , x and y in terms of the parameters of the model. In what follows, a typical expression is derived, the other expressions being derived in the same way.

The probability that exactly r sons of S_i are chosen is given by

$$P(r) = \binom{k_i}{r} \binom{m\bar{k} - k_i}{m\bar{v} - r} / \binom{m\bar{k}}{m\bar{v}}$$

where $\binom{\alpha}{\beta} = \alpha!/\beta!(\alpha-\beta)!$, so that v_i has the hypergeometric distribution. If \bar{v}_i is the mean number of sons of S_i chosen, it follows that

$$\bar{v}_i = \frac{\bar{v}}{\bar{k}} k_i.$$

If $V(v_i)$ is the variance of number of sons of S_i chosen, it also follows that

$$\begin{aligned} V(v_i) &= \mathcal{E}(v_i - \bar{v}_i)^2 \\ &= \frac{k_i(m\bar{k} - k_i)(\bar{k} - \bar{v})\bar{v}}{\bar{k}^2(m\bar{k} - 1)}. \end{aligned}$$

This value of $V(v_i)$ is needed to find $V(v)$, the variance of numbers of sons chosen from different sires, i.e.

$$V(v) = \mathcal{E} \left[\frac{\sum_{i=1}^m (v_i - \bar{v})^2}{m} \right],$$

and since $v_i - \bar{v} = v_i - \bar{v}_i + \bar{v}_i - \bar{v}$,

$$V(v) = \mathcal{E} \left[\frac{\sum_{i=1}^m (v_i - \bar{v}_i)^2 + \sum_{i=1}^m (\bar{v}_i - \bar{v})^2}{m} \right]$$

because $(v_i - \bar{v}_i)$ and $(\bar{v}_i - \bar{v})$ are uncorrelated.

Hence, since $\bar{v}_i = \frac{\bar{v}}{\bar{k}} k_i$,

$$V(v) = \frac{1}{m} \sum_{i=1}^m V(v_i) + \frac{\bar{v}^2}{\bar{k}^2} \left[\sum_{i=1}^m \frac{(k_i - \bar{k})^2}{m} \right].$$

If $\left[\sum_{i=1}^m \frac{(k_i - \bar{k})^2}{m} \right]$ is written $V(k)$, the equation becomes

$$V(v) = \frac{1}{m} \sum_{i=1}^m V(v_i) + \frac{\bar{v}^2}{\bar{k}^2} V(k).$$

When the previously found expression for $V(v_i)$ is substituted, it turns out that

$$V(v) = \frac{\bar{v}}{m\bar{k} - 1} \left[(m-1)(\bar{k} - \bar{v}) + (m\bar{v} - 1) \frac{V(k)}{\bar{k}} \right].$$

By a similar argument,

$$\begin{aligned} \text{Cov}(vw) &= \mathcal{E} \left[\frac{\sum_{i=1}^m (v_i - \bar{v})(w_i - \bar{w})}{m} \right] \\ &= \frac{1}{m} \sum_{i=1}^m \text{Cov}(v_i, w_i) + \frac{\bar{v}\bar{w}}{\bar{k}\bar{l}} \text{Cov}(kl). \end{aligned}$$

Since males and females are chosen independently, $\text{Cov}(v_i, w_i) = 0$, and

$$\text{Cov}(vw) = \frac{\bar{v}\bar{w}}{\bar{k}\bar{l}} \text{Cov}(kl).$$

Since there are similar expressions for the variances and covariances of other terms, the expected genetic drift in any particular situation may be worked out. It would, however, be useful to express the results in a form which may be fairly generally applied. For this purpose it is convenient to assume that at the time of sampling a random individual has equal probabilities of being either male or female. The result obtained is still too cumbersome to be convenient, so to reduce it to a more manageable form it is further assumed that the actual numbers of males and females equal their expected value of $\frac{1}{2}T$. This allows a very considerable simplification, and the error introduced by the approximation will be discussed later. It is possible under these conditions to make the substitutions

$$\begin{aligned} V(k) &= \frac{1}{4}[V(K) + \bar{K}], \\ \text{Cov}(kl) &= \frac{1}{4}[V(K) - \bar{K}], \\ m\bar{k} &= \frac{1}{2}T. \end{aligned}$$

When this is done, it turns out that

$$V(v) = \frac{M}{T-2} \left[\left(\frac{m-1}{m} \right) \left(\frac{T-2M}{m} \right) + \left(\frac{M-1}{m} \right) \left(\frac{T}{m} R_m + 2 \right) \right]$$

and
$$\text{Cov}(vw) = \frac{MF}{m^2} R_m,$$

where
$$R_m = \frac{V(K) - \bar{K}}{\bar{K}^2}.$$

It should be noted that the size and structure of the breeding flock are assumed predetermined, as is the case in most experimental control populations. This means that, for example, \bar{v} is fixed, with no sampling variation, so that no degree of freedom is used in its estimation, all degrees of freedom remaining for estimation of $V(v)$.

It is convenient to express genetic drift in terms of effective population size, N_E , defined by the equation

$$V(\Delta q) = \frac{q(1-q)}{2N_E}.$$

After algebraic manipulation, it turns out that, for the present model,

$$\begin{aligned} \frac{1}{N_E} = & \frac{T-1}{T-2} \left[\frac{1}{4M} + \frac{1}{4F} + \frac{R_m}{4m} + \frac{R_f}{4f} \right] - \frac{T}{T-2} \left[\frac{F+M}{4FM} \left(\frac{1+R_m}{4m} + \frac{1+R_f}{4f} \right) \right] \\ & + \frac{1}{T-2} \left[\frac{f+m}{4fm} - \frac{F+M}{4FM} \right] \end{aligned} \tag{1}$$

where
$$R_f = \frac{V(A) - \bar{A}}{\bar{A}^2}.$$

Of particular interest is the case where the numbers of sires and dams remain constant, when

$$\frac{1}{N_E} = \frac{T-1}{T-2} \left[\frac{1+R_m}{4M} + \frac{1+R_f}{4F} \right] - \frac{T}{T-2} \left[\frac{F+M}{4FM} \left(\frac{1+R_m}{4M} + \frac{1+R_f}{4F} \right) \right].$$

Denoting the effective size obtained by Latter for sampling with replacement as N_L ,

$$\frac{1}{N_L} = \frac{1+R_m}{4M} + \frac{1+R_f}{4F},$$

so that N_E is slightly larger than N_L .

If all individuals of the same sex have equal reproductive capacities, family size is binomially distributed, and $R_m = R_f = -\frac{1}{T}$. In this case, $N_L \sim \frac{4FM}{F+M}$. Thus when numbers of males and females remain constant, and reproductive capacities are equal for members of the same sex,

$$\frac{1}{N_E} = \frac{T-1}{T-2} \left(\frac{F+M}{4FM} \right) \left[1 - \frac{F+M}{4FM} - \frac{1}{T} \right].$$

Denoting $\frac{4FM}{F+M}$ as N_W (after Wright) it can be seen that when T and N_W are fairly large,

$$N_E \sim N_W + 1.$$

The effective size under sampling without replacement is about one more than under sampling with replacement.

The proportion of genes, G_{S_i} , from S_i in the matings of his offspring is, according to the definition of James & McBride (1958), given by $G_{S_i} = \frac{v_i}{4M} + \frac{w_i}{4F}$.

The variances of G_S and G_D , the proportions of genes from different sires and dams in the matings of their offspring, are given by

$$V(G_S) = \frac{V(v)}{16M^2} + \frac{V(w)}{16F^2} + \frac{2\text{Cov}(vw)}{16MF},$$

$$V(G_D) = \frac{V(x)}{16M^2} + \frac{V(y)}{16F^2} + \frac{2\text{Cov}(xy)}{16MF}.$$

It then follows that

$$V(\delta q_m) = \frac{q(1-q)}{2} \left[\frac{1}{4M} + \frac{1}{4F} + 4m V(G_S) \right],$$

$$V(\delta q_f) = \frac{q(1-q)}{2} \left[\frac{1}{4M} + \frac{1}{4F} + 4f V(G_D) \right],$$

which gives the relation between genetic drift and changes in proportions of genes.

It seems unnecessary to give explicit results for variances of G_S and G_D , but when T is fairly large, and the breeding flock has constant size and structure, it turns out that the equations

$$V(G_S) = \frac{1}{4M} \left[\frac{1}{N_W} + \frac{R_m}{M} \right],$$

$$V(G_D) = \frac{1}{4F} \left[\frac{1}{N_W} + \frac{R_f}{F} \right],$$

are approximately correct.

The error introduced by the assumption that \bar{k} and \bar{l} equal their expectation of $T/2m$ will obviously depend on by how much the assumption is incorrect. It can be shown by tedious algebra that the error in $\frac{1}{N_E}$ is roughly $\frac{T}{16} \left[\frac{m\bar{l} - m\bar{k}}{(m\bar{l})(m\bar{k})} \right]^2$, when $m = M$, $f = F$, and $R_m = R_f = 0$. This should provide a rough guide to the more general case.

Since the probabilities of each sex are assumed equal, $m\bar{k}$ (or $m\bar{l}$) is distributed approximately normally with mean $\frac{1}{2}T$ and variance $\frac{1}{4}T$ if $T > 10$. Thus about $\frac{2}{3}$ of the values of $m\bar{k}$ will lie within the range $\frac{1}{2}T \pm \frac{1}{2}\sqrt{T}$. Substituting $m\bar{k} = \frac{1}{2}T \pm \frac{1}{2}\sqrt{T}$ in the above expression, we obtain as the approximate error $1/(T-1)^2$. Since N_E will be at most of order T , it would seem that provided T is of the order of 50, the approximation will be reliable.

III. UNIVERSITY OF QUEENSLAND CONTROL POULTRY FLOCK

The theory developed here has been applied to the University of Queensland control poultry flock, which was formed in 1955 from a strain of Australorps under selection for egg production since 1953. Since 1955 it has been maintained by mating randomly chosen males and females, full pedigrees being kept. Females have been chosen by use of a table of random numbers, as have males since 1957. Previously the first males picked up from the yards had been used. It has been customary to allot approximately the same number of females as mates for each male.

Table 1 gives the number of parents used in each year. In 1955 only ten males were used at any one time, but one male died during the breeding season and was replaced.

Table 1. *Numbers of males and females used*

Year	1955	1956	1957	1958	1959
♂	11	12	12	10	10
♀	63	84	80	60	60

From the family sizes for sires and dams (K and A in the present notation) values of R_m and R_f have been calculated for each year and are given in Table 2. Both are consistently greater than zero, and this may be interpreted as evidence of the action of natural selection. In 1957 a deficiency of riboflavin in the diet of the chicks caused a heavy mortality, and both R_m and R_f increased, suggesting that mortality was not random over all families. The high value of R_m for 1955 was caused by the cockerel's death mentioned above.

Table 2. *Non-randomness of family size*

Year	1955	1956	1957	1958
R_m	0.2346	0.0411	0.1638	0.0756
R_f	0.4530	0.2595	0.8013	0.3845
T	698	555	299	677

T is the total number of offspring available for sampling, and R_m and R_f are defined in the text.

From the values of $v, w \dots$ in each year the proportions of genes from each male and female in the matings of their offspring have been calculated. For convenience calculations have been made on the percentages of genes γ_s and γ_D . Thus $V(\gamma_s)$ and $V(\gamma_D)$ are 10^4 times $V(G_s)$ and $V(G_D)$ respectively, and their values in different years are given in Table 3, together with their expected values based on the exact predictive formulae of the previous section.

From these formulae, $V(\gamma_D)$ should be roughly M/F times $V(\gamma_s)$. This holds reasonably for expected values, but not at all well for the actual values. It will also be noted that both $V(\gamma_s)$ and $V(\gamma_D)$ are less than expected in three of the years, while in 1956 both are more than expected. This suggests some systematic bias in sampling, but there seems to be no feature of the sampling procedure which would lead to bias.

Table 3. *Variances of gene percentages*

Year	1955	1956	1957	1958
$V(\gamma_S)$	9.0524	7.4617	3.7905	4.3750
$\mathcal{E}\{V(\gamma_S)\}$	9.6859	5.3029	8.3949	8.4349
$V(\gamma_D)$	0.9926	0.8349	0.8116	1.1690
$\mathcal{E}\{V(\gamma_D)\}$	1.2096	0.7951	1.2062	1.4556

γ_S and γ_D are the percentages of genes, from sires and dams respectively, represented in the matings of their progeny, and their expected variances are as calculated from the formulae given in the text.

The actual effective size of the flock for each year, N_E , has been calculated from the values of M , F , v , w , etc. The values of N_E estimated from equation (1), $\mathcal{E}(N_E)$, have also been calculated. For comparison N_L , the effective size given by Latter's formula, and N_W , the effective size given by the well-known formula due to Wright, have also been calculated; and all four sets of values are presented in Table 4.

Table 4. *Effective population sizes*

Year	1955	1956	1957	1958
N_E	35.6	35.8	36.2	38.5
$\mathcal{E}(N_E)$	33.1	40.0	29.5	31.5
N_L	32.3	39.1	28.5	30.6
N_W	42.0	41.7	34.3	34.3

N_E is the actual effective size as calculated from the data. $\mathcal{E}(N_E)$ is the effective size given by the equation derived in the text, N_L and N_W being the values given by the formulae of Latter and Wright.

It can be seen that, as expected, $\mathcal{E}(N_E)$ in each case is roughly $N_L + 1$, but that actual values of N_E have differed from their expectations by as much as 7. In 1957 and 1958 the values of N_E are considerably above expectation, though in these years the selection of breeders was made by using a table of random numbers, instead of the earlier practice of choosing the first ones picked out from the rearing-yards. It would thus seem that the deviations from expectation must be attributed to sampling, as the sex ratio among available offspring is near equality. Such deviations must always be expected, as only the average situation is described by the predictive equation.

The average inbreeding coefficient has been calculated for the flock in each generation, and the increase has averaged 1.27% per generation, compared with an average increase of about 1.37% expected from the effective numbers.

IV. DISCUSSION

A control population may be considered to have two sizes: (a) its breeding size, N_E ; and (b) its census size, T .

The breeding size determines its efficiency in maintaining genetic stability, while its census size determines its efficiency as an estimator of environmental variations because the sampling variance of the mean is proportional to $1/T$.

These two functions are not entirely independent, for genetic drift will be confounded with environmental changes, so that N_E also affects the efficiency of environmental corrections. On the other hand, N_E is not independent of T , though if T is at all large, this effect is likely to be negligible.

The variance in the control flock mean expected from genetic drift is V_g/N_E , where V_g is the genetic variance, while the sampling variance is V_p/T , where V_p is the phenotypic variance. Thus the relative importance of genetic drift and sampling of phenotypes as factors reducing efficiency is Th^2/N_E . If phenotype can be measured in one sex only, this ratio becomes $Th^2/2N_E$, with equality of sex ratio.

It has been shown by Gowe, Robertson & Latter (1959) that if N_P is the effective size of a pedigree flock (P),

$$\frac{1}{N_P} = \frac{3}{16M} + \frac{1}{16F}$$

and that N_P is considerably greater than N_R , the effective size of a similarly constituted random breeding flock (R). It is therefore likely that experimenters will choose P flocks in preference to R flocks where this is practical. The experimenter may choose to do this in two ways. He may set aside a fixed amount of facilities for his control and use a P flock to reduce genetic drift, or he may decide on a tolerable value of N_E and use a P flock to reduce the facilities needed. The alternatives may be illustrated by reference to the University of Queensland control poultry flock.

For the first alternative we may suppose 10 males and 60 females used for either R or P flocks. Then $N_R = 34.3$ and $N_P = 50.5$. T may be taken for convenience as $3F$, here 180. Then since the variance of errors of estimation is $\frac{V_g}{N_E} + \frac{V_p}{T}$, the gain in efficiency is not 47%, but, if $h^2 = 0.2$, about 21%. For the second alternative, suppose N_E is set at 35 and the male/female ratio is again 1/6. For an R flock, we have $M = 10$, $F = 60$, $T = 180$, and for a P flock we have $M = 7$, $F = 42$, $T = 126$. Again taking $h^2 = 0.2$, the efficiency of an R flock is 20% greater than that of a P flock. A variation of this approach is to set a fixed error variance, say that obtained by an R flock with $M = 10$ and $F = 60$. A P flock is then constructed to match this error. Under the above conditions it turns out that about 50 females are needed for breeding and about 150 phenotypes must be measured. The reduction in facilities needed is less than would be expected from consideration of genetic drift alone. In these examples it has been assumed that $R_m = R_f = 0$.

The above considerations apply only to changes from one generation to the next. Usually the statistic of interest will be $\bar{p}_t - \bar{p}_0$, the change in population mean over t generations, as this will be compared with the corresponding change in the mean of the selected population. The variance of the error of estimation associated with this measure of environmental change is given by

$$V(\bar{p}_t - \bar{p}_0) = V_p \left[\frac{th^2}{N_E} + \frac{2}{T} \right].$$

The number of measured individuals will usually lie between F and $20F$, and when heritability is not high, the major component of the error at the lower end of this range will arise from phenotypic sampling, while at the other end of the range genetic drift will be most important. At the lower end of the range it may pay to devote facilities to increasing the number of individuals measured, rather than to reducing genetic drift.

For traits with high heritabilities more attention should be paid to genetic drift, but it should be stressed that it is not the only feature which should be considered. The efficient design of a control flock requires that the effects of T , t , h^2 and N_E in giving rise to error should all be taken into consideration.

SUMMARY

1. The effect of genetic sampling, when this sampling is without replacement, on variation in gene frequency is studied, and equations describing the genetic drift are derived. The effective size turns out to be about one greater than under sampling with replacement.

2. The relation between 'spread of genes' and genetic drift is worked out.

3. The University of Queensland control poultry flock is analysed by these methods.

4. The design of control populations is discussed with particular reference to the relative importance of genetic drift and phenotypic sampling.

Part of this work was done while the author held a fellowship supported by the Rural Credits Development Fund of the Reserve Bank of Australia. I wish to thank Dr G. McBride for his helpful comments, and the referee, whose criticisms have been most valuable.

REFERENCES

- CROW, J. F. & MORTON, N. E. (1955). Measurement of gene frequency drift in small populations. *Evolution*, **9**, 202-214.
- GOWE, R. S., ROBERTSON, A. & LATTEr, B. D. H. (1959). Environment and poultry breeding problems. 5. The design of poultry control strains. *Poult. Sci.* **38**, 462-471.
- JAMES, J. W. & McBRIDE, G. (1958). The spread of genes by natural and artificial selection in a closed poultry flock. *J. Genet.* **56**, 55-62.
- LATTEr, B. D. H. (1959). Genetic sampling in a random mating control population of constant size and sex-ratio. *Aust. J. biol. Sci.* **12**, 500-505.