The above interpretation of the results is, of course, not necessarily correct, as the change in the activity of the population may have been due to genetic drift rather than selection. This explanation is less likely, and can only be tested by further study of the population, and of other populations.

Acknowledgments.—I am very grateful to Dr M. J. Kearsey for his advice, and to Professor P. L. Broadhurst for his comments on the manuscript. This work was carried out during tenure of a studentship from the Science Research Council, and was further supported by a grant from the Medical Research Council to Professors P. L. Broadhurst and J. L. Jinks for work on psychogenetics.

5. REFERENCES

- ANGUS, J. 1974a. Genetic control of activity, preening, and the response to a shadow stimulus in *Drosophila melanogaster*. *Behav. Genet.* 4, 317-329.
- ANGUS, J. 1974b. Measurement of locomotor activity in *Drosophila*. *Anim. Behav.*, in press.
- BARNES, B. W. 1968. Stabilising selection in *Drosophila melanogaster*. *Heredity*, 23, 433-442.
- CRAWFORD-SIDEBOTHAM, J. T. 1970. Differential susceptibility of species of slugs to metaldehyde/bran and to methiocarb baits. *Oecologia (Berl.)*, 5, 305-324.
- HAY, D. A. 1972. Genetical and maternal determinants of the activity and preening behaviour of *Drosophila melanogaster* reared in different environments. *Heredity*, 28, 311-336.
- KENDALL, M. G., AND STUART, M. A. 1961. *The Advanced Theory of Statistics Vol. 2, Inference and Relationship*. Griffin, London.
- MATHER, K. 1960. Evolution in polygenic systems. *Int. Colloquium on Evolution and Genetics, Rome, Acad. Naz. dei Lincei*.
- MATHER, K. 1966. Variability and selection. *Proc. Roy. Soc. B.*, 164, 328-340.
- MATHER, K. 1967. Complementary and duplicate gene interactions in biometrical genetics. *Heredity*, 22, 97-103.
- ROBERTS, R. C. 1967. Some evolutionary implications of behavior. *Canad. J. Genet. and Cytol.*, 9, 419-435.

THE BIAS IN THE REGRESSION OF Δq ON q

G. C. KIRBY

Department of Genetics, University of Adelaide, Adelaide, South Australia

Received 20.xi.73

SUMMARY

It is shown that when the regression of Δq (the change in gene frequency) on q (gene frequency) is calculated from data obtained from a sequence of censuses of one population, the regression coefficient does not provide an unbiased estimate of the amount of immigration or selection maintaining an equilibrium gene frequency in the population. Also, the equilibrium gene frequency estimated from the regression will be biased towards the observed gene frequency in the first census.

THE relationship between the gene frequency in a population at time $t(q_t)$ and the change in gene frequency (Δq_t) over the next time interval is useful for locating gene frequency equilibria in theoretical models of population genetics (Li, 1955). At an equilibrium point, $\Delta q_t = 0$. If there is a stable equilibrium gene frequency ($\bar{q}$) maintained by recurrent mutation, frequency dependent selection or immigration from a source population with constant

gene frequency, then the expected change in gene frequency over one generation is

$$\Delta q_t = x(\bar{q} - q_t) \quad (1)$$

where x is a constant which is a measure of the "force" by which the gene frequency is returned to $\bar{q}$ after perturbation. If the equilibrium is maintained by heterosis, then a similar equation will be valid except that x will be approximately constant only when q_t is near $\bar{q}$ (Kimura and Weiss, 1964). The simplicity of this approach to the determination of a gene frequency equilibrium and the "force" maintaining the equilibrium in theoretical models suggests that a similar approach may perhaps be useful when examining data from natural populations with the intention of demonstrating selection pressures or gene flow. When q_t has been estimated in a population on many successive censuses then the regression of Δq_t on q_t can be calculated to obtain an equation similar to equation (1). The slope of the regression, b , will be an estimate of $-x$, and an estimate of $\bar{q}$ is obtained by solving the regression equation for $\Delta q = 0$. This regression procedure has been noted by Williamson (1972) (although he plots $\Delta q/q(1-q)$ against q , which is preferable if heterosis is suspected) and applied by Tamarin and Krebs (1969, 1973) to *Microtus* population data.

When a modification of this method was applied to data from three house-mouse populations that were studied by mark-recapture techniques for a year, calculated values of b yielded estimates of x that were too large (0.48 to 1.3) to be accepted (Kirby, unpublished). This observation suggested that the regression might be biased so as to produce estimates of x that are much too large and it is easy to show that this is the case. Consider a locus with two neutral alleles in a small population which receives immigrants at a steady rate m per generation from a source population with constant gene frequency $\bar{q}$. Then the gene frequency in generation $t+1$ (q_{t+1}) will be given by

$$q_{t+1} = q_t + m(\bar{q} - q_t) + \delta q_t \quad (2)$$

where δq_t is the change in gene frequency due to random drift.

If the regression of Δq_t on q_t is calculated, the slope will be

$$b = \frac{\text{cov}(q_t, \Delta q_t)}{\text{var}(q_t)} = \frac{\text{cov}(q_t, q_{t+1})}{\text{var}(q_t)} - 1 \quad (3)$$

$$\doteq r(q_t, q_{t+1}) - 1 \quad (4)$$

where $\text{cov}(q_t, \Delta q_t)$ is the covariance of q_t and Δq_t , $\text{var}(q_t)$ is the variance of q_t and $r(q_t, q_{t+1})$ is the correlation between q_t and q_{t+1} . From equation (4), the value of b is approximately limited to the range 0 to -2 , that is, the estimates of m (derived from $-b$) will be restricted to the range 0 to 2. When there is no correlation between q_t and q_{t+1} , m will be $+1$, but in data collected from natural populations a positive correlation between successive gene frequencies would be expected and m will usually be less than 1.0. If equation (2) is substituted into equation (3), it is possible to obtain

$$b = -m + \frac{\text{cov}(q_t, \delta q_t)}{\text{var}(q_t)}. \quad (5)$$

Thus b will decrease with increasing m , and the value of b may also be affected by the size of the deme (as $E(\delta q_t)^2$ is proportional to $1/(\text{deme size})$) and the length of time over which the censuses are made, as $\text{var}(q_t)$ will increase with time. The second term in the R.H.S. in equation (5) represents the bias in the estimation of $-m$ by b . As $E(\delta q_t)^2$ is proportional to $q_t(1-q_t)$, $\text{cov}(q_b, \delta q_t)$ will not usually be zero.

For a given population, it may be possible to estimate the magnitude of this bias $\frac{\text{cov}(q_b, \delta q_t)}{\text{var}(q_t)}$ and hence to compensate for it, if the bias is influenced in a simple manner by only a few parameters. With this possibility in mind, a number of computer simulations were made of a haploid population with non-overlapping generations. Immigrants entered the population at rate m per generation, where m is the probability of a gene in the next generation coming from a source population with gene frequency 0.5 instead of from the genes in the present generation (sampling with replacement). In the first set of simulations three population sizes (100, 200 and 400 genes) and three immigration rates (0, 10 and 20 per cent) were used and simulations were run for three different time periods (10, 20 and 40 generations) with q_t and Δq_t calculated every generation. Each simulation began with $q_0 = 0.5$ and for each combination of population size, m and time period (NGENS), 100 replicate simulations were made with different pseudo-random numbers in order that each mean estimate of b should have a reasonably small standard error. The results showed that $-b$ appeared to increase linearly with m and decrease non-linearly with increasing time periods. However, population size had no detectable effect. Since the deviation between the initial and equilibrium gene frequencies may also affect b , this possibility was checked because in real populations it is unlikely that q_0 will be the same as $\bar{q}$. A second set of simulations was made with conditions the same as before except that only one population size (100 genes) was used and simulations started at a range of gene frequencies (0.1 to 0.9). Part of the results are given in table 1 (as the results for $q_0 > 0.5$ are a reflection of those for $q_0 < 0.5$, they have been left out). It is clear that $-b$ is a highly biased estimator of m even for long time periods and that the bias depends upon q_0 , m and NGENS in a complex fashion. The bias is most severe for small NGENS and always declines as NGENS increases. However, the effect upon the bias of altering q_0 depends upon m because the bias tends to decrease slightly as q_0 increases when $m = 0$, but tends to increase with q_0 when $m = 0.2$. This leads to the unexpected finding that when q_0 and m are small, an increase in m may lead to a decrease in $-b$. When q_0 is near 0.5 and especially for large NGENS, an increase in m will give an almost equivalent increase in $-b$. For some intermediate situations (*e.g.* $q_0 = 0.3$, NGENS = 10) a change in m has no statistically significant effect on b . Further simulations of populations with overlapping generations showed that if there is a lag between the time of mating and the time at which the progeny enter the observable population (this will occur, for example, in populations of animals where trapping misses the juveniles or plants with variable seed dormancies) then b is affected by the lag and the turnover of the population between censuses.

These simulation results show that many factors affect the bias in the estimation of m from b and in any realistic situation it will probably not be possible to compensate for this bias. This conclusion also applies to

TABLE 1

Estimation of the coefficient of the regression of Δq on q for various initial gene frequencies and immigration rates

q_0	m	NGENS		
		10	20	40
0.1	0	-0.4508 ± 0.0294	-0.2357 ± 0.0235	-0.1981 ± 0.0137
	0.1	-0.2434 ± 0.0185	-0.1822 ± 0.0112	-0.1391 ± 0.0061
	0.2	-0.2679 ± 0.0131	-0.2544 ± 0.0083	-0.2212 ± 0.0067
0.2	0	-0.4114 ± 0.0329	-0.2398 ± 0.0213	-0.1147 ± 0.0093
	0.1	-0.3508 ± 0.0272	-0.2107 ± 0.0146	-0.1487 ± 0.0081
	0.2	-0.3514 ± 0.0209	-0.2572 ± 0.0104	-0.2443 ± 0.0079
0.3	0	-0.3702 ± 0.0301	-0.2179 ± 0.0195	-0.1220 ± 0.0097
	0.1	-0.3825 ± 0.0305	-0.2310 ± 0.0150	-0.1921 ± 0.0103
	0.2	-0.4164 ± 0.0242	-0.3034 ± 0.0142	-0.2734 ± 0.0106
0.4	0	-0.3980 ± 0.0318	-0.1831 ± 0.0148	-0.1222 ± 0.0085
	0.1	-0.4660 ± 0.0279	-0.2712 ± 0.0168	-0.1917 ± 0.0090
	0.2	-0.5404 ± 0.0305	-0.3432 ± 0.0190	-0.2810 ± 0.0114
0.5	0	-0.4195 ± 0.0284	-0.2185 ± 0.0179	-0.1217 ± 0.0101
	0.1	-0.4244 ± 0.0283	-0.3095 ± 0.0193	-0.2147 ± 0.0112
	0.2	-0.5631 ± 0.0329	-0.3911 ± 0.0208	-0.2761 ± 0.0106

q_0 = Initial gene frequency; m = immigration rate from population with gene frequency of 0.5; NGENS = number of generations over which the regression is calculated. For each estimate, 100 replicate simulations were made and the standard error of each estimate is given.

estimates of selection coefficients derived from b (Tamarin and Krebs, 1969); such estimates will be biased and no valid conclusions can be drawn from them.

Does the regression also yield biased estimates of the equilibrium gene frequency? Simulation results indicate that it does. In the second set of simulations described above, the estimated equilibrium gene frequency ($\hat{q}$) was obtained for each simulation from the equation,

$$\hat{q} = \bar{q}_t - \bar{\Delta q}_t / b \quad (6)$$

where $\bar{q}_t$ and $\bar{\Delta q}_t$ are the mean values of q_t and Δq_t respectively. Some results are given in table 2. When $m = 0$, $\hat{q}$ provides an apparently unbiased estimate of q_0 , but when there is immigration, $\hat{q}$ appears to be an estimate of $\bar{q}$. However, for small m and NGENS the estimate $\hat{q}$ is biased towards q_0 and hence this method will not yield an unbiased estimate of $\bar{q}$ unless m , or NGENS, or both are sufficiently large.

However, $\hat{q}$ may be a better estimate of $\bar{q}$ than an average of gene frequencies over all censuses. The two most likely averages to be considered are (i) the mean over all censuses, which will be very similar to $\bar{q}_t$ except that it will also include the last census, which is left out of the calculation of $\bar{q}_t$; (ii) if there are overlapping generations then many individuals will be censused more than once and it will be preferable to count each individual censused only once and calculate a mean frequency over all individuals. If $\hat{q}$ is biased towards q_0 then from equation (6) (noting that usually b is negative and $\bar{\Delta q}_t$ is positive or negative as q_0 is less than or greater than $\bar{q}$), $\bar{q}_t$ is biased towards q_0 even more than $\hat{q}$. It is not easy to construct a simple argument on the bias of the two averages mentioned above, but

TABLE 2
*Estimation of equilibrium gene frequency ($\hat{q}$) for various initial gene frequencies
 and immigration rates, when $\bar{q} = 0.5$*

q_0	m	NGENS		
		10	20	40
0.1	0	0.0927 $\pm$ 0.0131	0.1462 $\pm$ 0.0322	0.0757 $\pm$ 0.0130
	0.1	0.4239 $\pm$ 0.0557	0.0281 $\pm$ 0.3643	0.4858 $\pm$ 0.0087
	0.2	0.6274 $\pm$ 0.1143	0.4893 $\pm$ 0.0067	0.5057 $\pm$ 0.0044
0.2	0	0.4060 $\pm$ 0.1954	0.0927 $\pm$ 0.0449	0.1773 $\pm$ 0.0470
	0.1	0.3285 $\pm$ 0.0612	0.4863 $\pm$ 0.0369	0.6159 $\pm$ 0.0884
	0.2	0.4698 $\pm$ 0.0197	0.5074 $\pm$ 0.0096	0.4947 $\pm$ 0.0039
0.3	0	0.2487 $\pm$ 0.1026	0.3395 $\pm$ 0.0486	0.3094 $\pm$ 0.0438
	0.1	0.3829 $\pm$ 0.0247	0.3824 $\pm$ 0.0684	0.4955 $\pm$ 0.0097
	0.2	0.4862 $\pm$ 0.0170	0.4900 $\pm$ 0.0163	0.5001 $\pm$ 0.0039
0.4	0	0.2870 $\pm$ 0.0689	0.3697 $\pm$ 0.0346	0.3735 $\pm$ 0.0429
	0.1	0.4625 $\pm$ 0.0261	0.5000 $\pm$ 0.0129	0.4838 $\pm$ 0.0075
	0.2	0.4407 $\pm$ 0.0551	0.4283 $\pm$ 0.0687	0.4965 $\pm$ 0.0038
0.5	0	0.4641 $\pm$ 0.0735	0.5405 $\pm$ 0.0559	0.4824 $\pm$ 0.0358
	0.1	0.4991 $\pm$ 0.0156	0.4701 $\pm$ 0.0275	0.5026 $\pm$ 0.0085
	0.2	0.4968 $\pm$ 0.0077	0.5264 $\pm$ 0.0122	0.4990 $\pm$ 0.0041

q_0 = Initial gene frequency; m = immigration rate; NGENS = number of generations over which the regression is calculated. For each estimate, 100 replicate simulations were run and the standard error of each estimate is given.

by analogy with the bias of $\bar{q}_t$ it is likely that they are more biased than $\hat{q}$. If this assertion is true it implies that $\hat{q}$, biased though it may be, is the best estimate of $\bar{q}$ available until an improved estimator can be derived using a modification of this method or a new method.

To conclude, these results show that the regression of Δq on q should not be used to estimate immigration rates or selection pressures in populations and that the equilibrium gene frequencies estimated from the regression will be biased towards the initial gene frequency in the sequence of censuses.

Acknowledgments.—Professor J. H. Bennett and Dr O. Mayo made useful comments on the manuscript. I was supported by a University of Adelaide Research Grant Scholarship.

REFERENCES

- KIMURA, M., AND WEISS, G. H. 1964. The stepping stone model of population structure and the decrease of genetic correlation with distance. *Genetics*, 49, 561-576.
- LI, C. C. 1955. The stability of an equilibrium and the average fitness of a population. *Amer. Natur.*, 89, 281-295.
- TAMARIN, R. H., AND KREBS, C. J. 1969. *Microtus* population biology. II. Genetic changes at the transferrin locus in fluctuating populations of two vole species. *Evolution*, 23, 183-211.
- TAMARIN, R. H., AND KREBS, C. J. 1973. Selection at the transferrin locus in cropped vole populations. *Heredity*, 30, 53-62.
- WILLIAMSON, M. 1972. *The Analysis of Biological Populations*. Edward Arnold, London.