Selection and Covariance

THIS is a preliminary communication describing applications to genetical selection of a new mathematical treatment of selection in general.

Gene frequency change is the basic event in biological evolution. The following equation (notation to be explained), which gives frequency change under selection from one generation to the next for a single gene or for any linear function of any number of genes at any number of loci, holds for any sort of dominance or epistasis, for sexual or asexual reproduction, for random or non-random mating, for diploid, haploid or polyploid species, and even for imaginary species with more than two sexes

$$\Delta Q = \mathrm{Cov}(z,q)/\bar{z} \qquad (1)$$

The equation easily translates into regression coefficient (β_{zq}) or correlation coefficient (ρ_{zq}) form

$$\Delta Q = \beta_{zq}\sigma_q^2/\bar{z} = \rho_{zq}\sigma_z\sigma_q/\bar{z}$$

It thus has great transparency, making it useful as a tool in qualitative evolutionary reasoning. It can also be applied to non-genetical selection. For example, if students' expectations of passing a certain course vary with IQ and if student IQs do not change appreciably during the course, then equation 1 (with its variables suitably redefined) will give the difference in mean IQ between students entering the course and those completing it (and equation 4 below will apply if IQs do change during the course).

Derivation is as follows. Let P_1 and P_2 be populations of a single species, such that P_1 contains all parents of P_2 members and P_2 consists of all offspring of P_1 members. Let the number of P_1 members be N. We label these with identification numbers $i = 1, 2, \ldots, N$, assigned in any order. Let n_z be the zygotic ploidy of the species for gene A; let g_i be the dose of gene A in individual i (for example, if $n_z = 2$, $g_i = 0$, 1, or 2 according to whether i lacks gene A, is heterozygous for A or is homozygous for A); let q_i be the frequency of gene A in individual i, defined by $q_i = g_i/n_z$; and let Q_1 be the frequency of gene A in population P_1

$$Q_1 = \Sigma g_i/n_z N = \Sigma n_z q_i/n_z N = \bar{q} \qquad (2)$$

where the summations are taken over all members of P_1 ($i = 1$ to N) and $\bar{q}$ is the arithmetic mean in population P_1 (that is, $\bar{q}$ is a population variable even though I use sample variable notation).

Now we turn attention to offspring. A gamete from a P_1 member that contributes genes to a P_2 member will be termed a "successful gamete". Let n_G be the gametic ploidy for gene A; let z_i be the number of successful gametes produced by individual i (= the number of i's offspring); let g_i' be the number of A genes in the set of all of i's successful gametes; let q_i' be the frequency of gene A in this set of gametes, defined by $q_i' = g_i'/z_i n_G$ if $z_i \neq 0$, $q_i' = q_i$ if $z_i = 0$; let $\Delta q_i = q_i' - q_i$; and let Q_2 be the frequency of gene A in population P_2. The following can be seen to hold

$$\begin{aligned} Q_2 &= (\Sigma g_i')/\Sigma z_i n_G = (\Sigma z_i n_G q_i')/\Sigma z_i n_G = \Sigma z_i q_i'/N\bar{z} \\ &= \Sigma z_i q_i/N\bar{z} + \Sigma z_i \Delta q_i/N\bar{z} = [\bar{z}\cdot\bar{q} + \mathrm{Cov}(z,q)]/\bar{z} + \Sigma z_i \Delta q_i/N\bar{z} \\ &= \bar{q} + \mathrm{Cov}(z,q)/\bar{z} + \Sigma z_i \Delta q_i/N\bar{z} \end{aligned} \qquad (3)$$

where the summations are taken over all P_1 members, $\bar{z}$ is the arithmetic mean of z in P_1 and $\mathrm{Cov}(z,q)$ is the covariance (or first order central product moment) of z and q in population P_1. Subtraction of equation 2 from equation 3 gives

$$\Delta Q = Q_2 - Q_1 = \mathrm{Cov}(z,q)/\bar{z} + \Sigma z_i \Delta q_i/N\bar{z} \qquad (4)$$

If meiosis and fertilization are random with respect to gene A, the summation term at the right will be zero except for statistical sampling effects ("random drift"), and these will tend to average out to give equation 1.

Five points about equation 1 will be briefly explained. First, equation 1 in its regression coefficient form can be visualized in terms of a linear regression line fitted to a scatter diagram of z against q. (A linear regression line is the best construction in terms of the population effect ΔQ, even if it gives a poor fit in terms of individual points.) Since the regression line has slope β_{zq}, gene frequency change due to selection is exactly proportional to the slope. Therefore, at any step in constructing hypotheses about evolution through natural selection—for example, about why human canines do not protrude, why deer antlers are annually shed and renewed, why parrots mimic, why dolphins play—one can visualize such a diagram and consider whether the slope really would be appreciably non-zero under the assumptions of the theory. If there is no slope, then there is no frequency change except by Δq effects, and the hypothesis is probably wrong.

Second, equation 1 fails if gene A ploidy is not the same in each P_1 member. Suppose, for example, that the A locus is in X but not Y chromosomes in a species with XX females and XY males. Then Q_1 is redefined as $Q_1 = (\Sigma g_i)/\Sigma n_i$, where n_i is A locus ploidy in individual i (that is, $n_i = 1$ if i is male, or 2 if i is female); and Q_2, q_i, and q'_i are redefined in corresponding ways. If P_1 and P_2 have sex ratios of unity (as is commonly the case at conception), then the following can be derived

$$\Delta Q = \frac{2}{3}\,\mathrm{Cov}(z,q)_F/\bar{z}_F + \tfrac{1}{3}\mathrm{Cov}(z,q)_M/\bar{z}_M \qquad (5)$$

where $\mathrm{Cov}(z,q)_F$ is the z,q covariance and $\bar{z}_F$ is the mean in P_{1F}, the female subset of P_1, and $\mathrm{Cov}(z,q)_M$ and $\bar{z}_M$ apply to the male subset, P_{1M}.

Third, the specifications that were stated for P_1 and P_2 imply a "discrete generations model". This was done solely in order to simplify this preliminary report. Actually equation 1 can be applied to species with overlapping, interbreeding generations, and it is not necessary that P_2 should contain all offspring of P_1 members, nor that P_1 should contain all parents of P_2 members. Departure from the "all parents" condition, however, requires reinterpretation of what ΔQ means, and departure from the "all offspring" condition (meaning all zygotes conceived) must be done with insight to avoid introducing post-conceptual selection on P_2 (for post-conceptual selection would require the use of equation 4 instead of equation 1).

Fourth, as an example of how multiple gene functions can be handled, let us suppose that a regression analysis has given the relation

$$\varphi_i \approx 2{\cdot}3 + 1{\cdot}2q_{iA} - 0{\cdot}7q_{iB} + 0{\cdot}5q_{iC}$$

for the effects of genes A, B, and C on character φ. Then we may decide to define

$$\begin{aligned} q_i &= 2{\cdot}3 + 1{\cdot}2q_{iA} - 0{\cdot}7q_{iB} + 0{\cdot}5q_{iC} \\ \Delta Q = Q_2 - Q_1 &= (2{\cdot}3 + 1{\cdot}2Q_{2A} - 0{\cdot}7Q_{2B} + 0{\cdot}5Q_{2C}) \\ &\quad - (2{\cdot}3 + 1{\cdot}2Q_{1A} - 0{\cdot}7Q_{1B} + 0{\cdot}5Q_{1C}) \end{aligned}$$

and equation 1 will hold for these multiple gene functions or for any other linear function of q and Q for any numbe of genes, if it holds for each gene separately.

Fifth, it seems surprising that so simple a relation as equation 1 has not (to my knowledge) been recognized before. Probably this is because selection mathematics has largely been limited to genetical selection in diploid species, where covariance takes so simple a form that its implicit presence is hard to recognize (whereas if man were tetraploid, covariance would have been recognized long ago); and because, instead of using subscripts as "names" of individuals (as I have done), the usual practice in gene frequency equations is to use subscripts only as names of gene or genotype types, which makes the mathematics seem quite different. Recognition of covariance (or regression or correlation) is of no advantage for numerical calculation, but of much advantage for evolutionary reasoning and mathematical model building.

Some genetical selection cases (such as group selection) and many forms of non-genetical selection require more complex mathematics than that given here. I plan to discuss these and other matters in papers now in preparation.

I thank Professor C. A. B. Smith for help, and the Science Research Council for financial support.

George R. Price

Galton Laboratory,
University College,
Stephenson Way,
London NW1.