Ann. Hum. Genet., Lond. (1972), **35**, 485
Printed in Great Britain

Extension of covariance selection mathematics

By GEORGE R. PRICE
The Galton Laboratory, University College London

This paper gives some extensions of the selection mathematics based on the covariance function published in Price (1970). Application of the mathematics to 'group selection' is briefly illustrated. More about applications will be shown in a later paper concerning 'Selection in populations with overlapping generations', which will be submitted to this journal. To facilitate reference in that paper, the equations in this paper are labelled with the letter 'A'.

The mathematics given here applies not only to genetical selection but to selection in general. It is intended mainly for use in deriving general relations and constructing theories, and to clarify understanding of selection phenomena, rather than for numerical calculation.

WEIGHTED STATISTICAL FUNCTIONS

In this paper we will be concerned with population functions and make no use of sample functions, hence we will not observe notational conventions for distinguishing population and sample variables and functions. We begin by defining notation for weighted statistical functions. Here we generalize and extend notation defined in Price (1971):

$$\text{ave}_w x = (\sum_i w_i x_i)/\sum_i w_i, \tag{A 1}$$

$$\text{cov}_w(x, y) = [\sum_i w_i(x_i - \text{ave}_w x)\,(y_i - \text{ave}_w y)]/\sum_i w_i, \tag{A 2}$$

$$\text{var}_w x = \text{cov}_w(x, x), \quad \sigma_w(x) = \sqrt{\text{var}_w x}, \tag{A 3}$$

$$\beta_w(x, y) = \text{cov}_w(x, y)/\text{var}_w y, \tag{A 4}$$

$$\rho_w(x, y) = \text{cov}_w(x, y)/[\sigma_w(x)\,\sigma_w(y)]. \tag{A 5}$$

Here w, x and y can be any variables, and summations are over all population members. We call w the 'weight' or 'weighting variable', and we can speak of these functions as 'weighted arithmetic means', 'weighted covariances', 'weighted variances', and so on. If all w_i are equal, the weighted functions become ordinary unweighted functions, hence what is said about weighted functions will apply to unweighted functions as particular cases. It should be noted that the value of a weighted function is unchanged if every weight, w_i, is multiplied by the same constant.

It should also be noted that the right sides of equations (A 1) to (A 5) are identical with standard expressions for grouped variables, where w_i = the number of individuals in the ith group. The sole difference is that we are not limiting w to integral values; but of course with grouped variables we can approximate weighting by fractional amounts as closely as desired by using groups of size cw_i, where c is a large constant. Accordingly, any identity that holds

with unweighted functions will hold with weighted functions, provided we weight each statistical function in the equation with the same set of w_i weights.

Two useful identities involving weighted statistical functions are:

$$(\sum_i w_i x_i y_i)/\sum_i w_i = \mathrm{cov}_w(x, y) + (\mathrm{ave}_w x)(\mathrm{ave}_w y), \tag{A 6}$$

$$\mathrm{ave}_w x = \bar{x} + \mathrm{cov}(w, x)/\bar{w}, \tag{A 7}$$

The first gives the relation between product moments around the origin and product moments around the means. The second shows the relation between the ordinary, unweighted arithmetic mean, $\bar{x}$, and the weighted mean, $\mathrm{ave}_w x$.

EQUATIONS FOR TYPE I SELECTION PROBLEMS

We will use p to represent the numerical value of some *property* of population members, and we define

$$\Delta p_i = p'_i - p_i, \quad \Delta(\mathrm{ave}_w p) = \mathrm{ave}_{w'} p' - \mathrm{ave}_w p. \tag{A 8}$$

We will call s a 'selection coefficient', and we define

$$s_i = w'_i/w_i. \tag{A 9}$$

We now expand $\mathrm{ave}_{w'} p'$ by means of (A 1), and substitute $p_i + \Delta p_i$ for p'_i and then substitute $w_i s_i$ for one of the w'_i pairs, giving

$$\mathrm{ave}_{w'} p' = (\sum_i w_i s_i p_i)/\sum_i w_i s_i + (\sum_i w'_i \Delta p_i)/\sum_i w'_i. \tag{A 10}$$

Now we multiply the first term on the right of (A 10) by $\Sigma w_i/\Sigma w_i$, convert by (A 6), and subtract $\mathrm{ave}_w p$ from both sides, to give

$$\Delta(\mathrm{ave}_w p) = \mathrm{cov}_w(s, p)/\mathrm{ave}_w s + \mathrm{ave}_{w'}(\Delta p). \tag{A 11}$$

This equation is an identity for all values of w_i, p_i, w'_i and p'_i. It is the basic equation for what will be called 'Type I' selection problems. What we mean by a Type I selection problem, how we apply equation (A 11) to such problems, and what the primed variables represent will be made clear shortly, but first two possible ways of simplifying (A 11) will be explained.

The first is simply a matter of notation. It is often preferable to think of the covariance in (A 11) as involving, not the selection coefficient s_i, but the relative selection coefficient, $s_i/\mathrm{ave}_w s$. We will use a tilde to symbolize such relative variables, involving a variable divided by its population mean, so that we define $\tilde{s}_i = s_i/\mathrm{ave}_w s$. Hence (A 11) can be rewritten as

$$\Delta(\mathrm{ave}_w p) = \mathrm{cov}_w(\tilde{s}, p) + \mathrm{ave}_{w'}(\Delta p). \tag{A 12}$$

A further simplification that is permissible in many applications is to omit the Δp term, giving the very simple form

$$\Delta(\mathrm{ave}_w p) = \mathrm{cov}_w(\tilde{s}, p). \tag{A 13}$$

This holds if $p'_i = p_i$ for all i. Also, if the expected value $\mathscr{E}(p'_i) = p_i$ for all i, and if the population size is large, then $\mathrm{ave}_{w'}(\Delta p)$ will generally be negligible relative to the covariance term so that (A 13) will hold as a very good approximation.

Still further simplification occurs with selection problems where all the w_i are equal. This reduces the covariance to the unweighted form $\mathrm{cov}(\bar{s}, p)$, giving selection equations like equations (1) and (4) of Price (1970).

A GROUP SELECTION EXAMPLE

The best way to explain the meaning and uses of the basic Type I selection equations is through an example. Currently 'group selection' is a highly controversial topic in evolutionary biology. (There is an enormous literature on group selection, of which I will cite just one sample paper: Wright, 1949.) Our example will be to derive a 'group selection' equation covering the limiting case of reproductively isolated groups with no intergroup migration. We imagine a large population, Π, that is subdivided into a number of subpopulations or groups. For simplicity we will assume that different generations do not overlap. Let n_i = the population of group G_i in the parent generation, and let n_i' = the population of group G_i in the offspring generation. Let $s_i = n_i'/n_i$ = the mean number of offspring per parent generation member of group G_i (when each parent is given credit for half of each offspring conceived). Let p_i = the population frequency of gene A in group G_i in the parent generation, and let p_i' = the frequency in the offspring generation. Similarly, let P = the population frequency of gene A in the total population Π in the parent generation, and let P' = the frequency in Π in the offspring generation. Our problem is to determine the change $\Delta P = P' - P$.

It can easily be seen that

$$P = \mathrm{ave}_n\, p, \quad P' = \mathrm{ave}_{n'}\, p', \tag{A 14}$$

so that $\Delta P = \Delta(\mathrm{ave}_n\, p)$. Consequently, since (A 12) is an identity for all w, p, w' and p' (provided that $s_i = w_i'/w_i$), it follows that

$$\Delta P = \mathrm{cov}_n(\bar{s}, p) + \mathrm{ave}_{n'}(\Delta p). \tag{A 15}$$

Here n is of course playing the role of w in equation (A 12). It should be noted that (A 15) does not depend upon any assumptions about mechanisms of heredity or anything else of that sort, but holds because it is an identity no matter how n, p, n' and p' are defined (provided that $s_i = n_i'/n_i$ in accordance with equation A 9).

We define a Type I selection problem as a problem where the aim is to determine the change in an arithmetic mean (weighted or unweighted) caused by selection. We use primes in the sense of 'successor' or 'later' to indicate variables *after* the operation of selection. In general, w is some sort of measure of *amount* or 'value', p can be any finite quantitative property, and s_i, for values of s that do not exceed unity, can be thought of as describing the fraction of set member i that is 'selected'. The first term on the right in equations (A 11), (A 12) and (A 15) gives the change in weighted mean p caused by selection; the second term gives the change in mean p due to 'property change'. It should be noted that this mathematical treatment of selection implies a one–one relation between pre-selection and post-selection set members, with a common system of i index numbering. Such one–one relation always 'exists' in any case of selection, though sometimes it 'exists' only in a mathematical sense. It should also be mentioned that the 'Type I' category is a far broader problem category than one might at first assume.

INTERPRETATIONS AND EXTENSIONS

Returning to the group selection example, we can use equation (1) of Price (1970) to evaluate the change, Δp_i, in gene A frequency within each group. In our present notation that equation can be written as

$$\Delta p = \text{cov}(\tilde{s},q) = \text{cov}(\tilde{z},q). \tag{A 16}$$

Here p is population gene frequency, q is 'individual gene frequency' (see Price, 1970, 1971) z is the number of offspring conceived by a particular individual, when we credit each parent with a full offspring, and s is half of z (giving each parent credit for half of each offspring). Substitution from (A 16) into (A 15) gives

$$\Delta P = \text{cov}_n(\tilde{s},p) + \text{ave}_{n'}[\text{cov}(\tilde{z},q)]. \tag{A 17}$$

Here s has the meaning defined for equation (A 15), and we have expressed (A 16) in the z form in order to avoid confusion between the (A 15) and (A 16) uses of s. In (A 17) the $\text{cov}_n(\tilde{s},p)$ term can be thought of as representing group selection (or, more precisely, *inter*group selection), for it shows what part of the change ΔP is due to differences among the mean fecundities of the groups G_i. The last term shows the part of the change ΔP that is caused by *intra*group selection due to differences among individuals within a group. We can speak of (A 17) as a 'two-level' selection equation involving two different 'levels' of selection. The procedure involved in going from (A 15) to (A 17) shows how a property change (ΔP, Δp or Δq) at one selection level can often be interpreted as due to selection at a lower level.

Covariance treatment gives maximum simplicity with Type I equations, but conversion to regression or correlation form may make them more intuitively understandable. Equation (A 13) can be expanded by (A 4) or (A 5), giving

$$\Delta(\text{ave}_w p) = \beta_w(\tilde{s},p)\,\text{var}_w p = \rho_w(\tilde{s},p)\,\sigma_w(\tilde{s})\,\sigma_w(p), \tag{A 18}$$

and similar expansions can be used with other forms such as (A 11). Equations (A 18) can be interpreted in terms of the following pattern:

$$\text{effect of selection} = \text{intensity of selection} \times \text{variation on which selection acts.}$$

Possible measures of the 'variation on which selection acts' are $\text{var}_w p$, $\sigma_w(p)$, or $\sigma_w(\tilde{s}).\sigma_w(p)$. Possible measures of the 'intensity of selection' are $\beta_w(\tilde{s},p)$, $\beta_w(\tilde{s},p)\,\sigma_w(p)$, or $\rho_w(\tilde{s},p)$. These measure selection intensity in relation to property p, whereas Haldane's (1954) measure of the 'intensity of natural selection' is a measure of the range of s_i or z_i values existing in a population, without regard to how these variables relate to any property. Thus the types of measure, though similar in name, are fundamentally different in the selection characteristics that they measure.

Once again returning to the group selection example, we use the first of the two (A 18) expansions to put (A 17) into this form:

$$\Delta P = \beta_n(\tilde{s},p)\,\text{var}_n p + \text{ave}_{n'}[\beta(\tilde{z},q)\,\text{var}\,q]. \tag{A 19}$$

Using the first of the three suggested interpretations of the 'variation on which selection acts' and the interpretation of the 'intensity of selection' that belongs with it, we can interpret (A 19) as saying the following: the change in gene A frequency in the total population is equal to the intensity, $\beta_n(\tilde{s},p)$, of intergroup selection on gene A multiplied by the intergroup variance

of gene A population frequencies, plus the weighted mean of the products of intragroup selection intensities multiplied by intragroup variances of gene A individual frequencies. The current 'group selection controversy' hinges on the question of whether the intergroup variance, $\mathrm{var}_n\, p$, is likely to be of significant magnitude under realistic natural conditions, so that there can be significant increases ($\Delta P \gg 0$) in the frequency of a gene that is 'group-benefiting' ($\beta_n(\tilde{s},p) > 0$) but not 'individual-benefiting' ($\beta(\tilde{z},q) \leqslant 0$). Thus (A 19) 'maps' transparently the different factors in group selection, which may (hopefully) help to clarify the main issues in the controversy. (Of course, group selection is just being used here as an example, and it is not the intention of this paper to try to 'settle' the group selection controversy.)

Another useful extension of the Type I selection equations is to continuous time models. Let us suppose that in a selection problem unprimed variables p and w apply to a certain time t, and primed p' and w' apply to a later time $t' = t+\Delta t$. We define a new variable, r, which can be called the 'selection rate coefficient', defined as

$$r_i = (w_i'-w_i)/w_i\Delta t = (s_i-1)/\Delta t. \tag{A 20}$$

Hence $s_i = 1+r_i\Delta t$. Making this substitution in the first term on the right of (A 11), we obtain

$$\mathrm{cov}_w(s,p)/\mathrm{ave}_w\, s = \mathrm{cov}_w(r,p)\,\Delta t/(1+\mathrm{ave}_w r\Delta t),$$

so that (A 11) can be put into this form:

$$\Delta(\mathrm{ave}_w\, p)/\Delta t = \mathrm{cov}_w(r,p)/(1+\mathrm{ave}_w r\Delta t)+\mathrm{ave}_{w'}(\Delta p/\Delta t). \tag{A 21}$$

Now suppose that we set $\Delta t = dt$. In this case (A 20) becomes

$$r_i = d\log w_i/dt. \tag{A 22}$$

and the denominator $1+\mathrm{ave}_w r\Delta t$ becomes $1+\mathrm{ave}_w r\,dt = 1$, so that (A 21) becomes

$$d(\mathrm{ave}_w\, p)/dt = \mathrm{cov}_w(r,p)+\mathrm{ave}_{w'}(dp/dt). \tag{A 23}$$

In applying (A 22) and (A 23) in population genetics, where changes in numbers w_i or n_i will be discontinuous, we can use expected values rather than actual values, or if the population is very large, we can imagine the discontinuously varying p and w replaced by continuous variables. In applications where the (A 12) simplification is permissible – which should be true of the majority of population genetics applications – equation (A 23) takes the pleasingly simple form

$$d(\mathrm{ave}_w\, p)/dt = \mathrm{cov}_w(r,p). \tag{A 24}$$

SUMMARY

General equations are derived for changes in weighted population means caused by selection. Some new ways of measuring 'selection intensity' are suggested. Application of the selection equations in population genetics is illustrated by a simplified treatment of 'group selection'.

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

REFERENCES

HALDANE, J. B. S. (1954). The measurement of natural selection. *Proceedings of the 9th International Congress of Genetics*, part I, pp. 480–7.

PRICE, G. R. (1970). Selection and covariance. *Nature, London* **227**, 520–1.

PRICE, G. R. (1971). Extension of the Hardy–Weinberg Law to assortative mating. *Annals of Human Genetics* **34**, 455–8.

WRIGHT, S. (1949). Adaptation and selection. In G. L. Jepsen, E. Mayr and G. G. Simpson (ed), *Genetics, Paleontology, and Evolution*. New York: Atheneum.

Note added in proof. I have recently learned that Dr Alan Robertson published covariance selection equations similar to equation (1) of Price (1970) in 1966 in a paper on 'A mathematical model of the culling process in dairy cattle' in *Animal Production* **8**, 95–108. I believe that the present extensions are new. It should also be mentioned that equation 5.11.6 of J. F. Crow and M. Kimura (1970), *An Introduction to Population Genetics Theory* (New York, Evanston, and London: Harper & Row), has some similarity to my equation (A 17), though their equation involves variance rather than covariance and fitnesses instead of gene frequencies.