On aspects of the deterministic theory and their possible stochastic extensions

Warren J Ewens

Marseille, 14 June 2012

General theme 1.

Deterministic evolutionary population genetics theory is currently motivated by the need to consider whole genome data in non-randomly-mating diploid populations, (particularly the human population) and has thus moved to the *whole-genome, non-random-mating diploid* case. This talk discusses aspects of this theory and considers possible extensions to the corresponding stochastic theory.

General theme 2. The Price equation.

The Price equation is part of the deterministic whole-genome theory of population genetics in which no assumption is made about the mating scheme. It is little-known in standard population genetics theory.

The Price equation is used to discuss:-

1. Evolution.

2. The correlation between relatives for a metrical trait.

3. Kin and group selection matters.

The Price equation

What is it? There are two versions.

1. The original (1970) Price version. This involves only allelic (gene) frequency changes from one generation to the next.

2. The later (1972) version, involving a generalization to changes in any character from one generation to the next.

What benefits are there in using the Price equation?

Answer #1 (Gardner, Current Biology, Vol 18)

“Price discovered an entirely novel approach to population genetics, and the basis for a general theory of selection – the Price equation. The Price equation has come to underpin several key areas in evolutionary theory.”

Answer # 2. (Grafen (2008)). The Price equation can be used to show how individuals in a population “solve the [evolutionary] optimization problem”.

(WJE: Whatever this expression means - See comments later on this claim).

More generally, enthusiasts for the equation claim that it explains everything about evolution.

Answer #3 (Ewens, Marseille, 2012)

The Price equation is an interesting result that allows one to do some “unusual” calculations. This arises because it looks at parent-offspring relationships in a way that is different from that used in classical population genetics.

q_g

The “Price” approach might also possibly be associated with coalescent calculations (see later).

However, claims for its usefulness are exaggerated. Also, several either wrong or trivial results have been derived from it. Further, aspects of the general form of the equation is suspect.

What is fitness?

The Price equation was originally written under the view that the fitness of any diploid individual is the actual number of genes at any locus that he passes on to the next generation (i.e. his number of offspring). Under this viewpoint I write w_{i} as the fitness of individual *i*, ($i = 1, 2, \ldots, N$).

In classical population genetics a fitness is a parameter, and is associated with a genotype. All individuals of the same genotype have the same fitness. So when considering classical population genetics I write w_{g} as the fitness of genotype *g* ($g = 1, 2, \ldots,$ approx 10^5000). (Why so many?)

There is of course an infinite population translation between the two approaches.

The original Price equation

This refers to the change in frequency of some gene *A* between a parental and a daughter generation.

Label individuals in the parent generation by i ($i = 1, 2, \ldots, \mathrm{N}$). Let q_{i} be the proportion of *A* genes in individual i ($q_{\mathrm{i}} = 0, 1/2$ or 1). The parental generation frequency of *A* is thus $\Sigma_{\mathrm{i}}\, q_{\mathrm{i}} /\mathrm{N}$.

Let individual i transmit w_{i} genes to the next generation. Let n_{i}' of these be *A* genes. Write

$$\Delta q_{\mathrm{i}} = n_{\mathrm{i}}'/w_{\mathrm{i}} - q_{\mathrm{i}}.$$

This is the difference between the frequency of the *A* gene transmitted by individual i and the frequency in that individual.

The between - generation change in the frequency of the *A* gene is

$$\Delta q = \sum_{i=1}^{N} \{n_{\mathrm{i}}^{'} /(N\overline{w}) - q_i / N\}$$

$$\sum_{i=1}^{N} \{(w_i - \overline{w})q_i + w_i \Delta q_i\}/(N\overline{w})$$

$$= \mathrm{cov}(w, q)/\overline{w} + E(w\Delta q) / \overline{w}.$$

Is this equation useful?

Finite N – thus it admits a stochastic version.

(As N increases, the mean of the second term approaches 0. Thus there is only one term on the RHS in the infinite population version.)

The focus on individuals instead of genotypes (as indicated by the notation) leads to the possibility of using the equation for kin and group selection questions and to discuss altruism.

Other uses – see later.

The general Price equation.

Consider some metrical trait, say z = (mature age) height. Let z_{i} be the height of individual i (i = 1, 2, …, N) in a parental generation.

The average height in this generation is $N^{-1}\Sigma\, z_{\mathrm{i}}$.

Write the average height of the offspring of individual i as $z_i + \overline{d}_i$.

The average height of individuals in the offspring generation is then

$$[\sum_{i=1}^{N} w_i]^{-1}\ [\sum_{i=1}^{N} w_i(z_i + \overline{d}_i)].$$

The change $\Delta\overline{z}$ in average height between parental and daughter generations is the difference between these two quantities.

(There is a problem here - an offspring is the product of two parents, so how is the accounting done? This seems unresolved.)

Writing

$$\sum_{i=1}^{N} w_i / N = \overline{w},$$

$$N^{-1} \sum_{i=1}^{N} (w_i - \overline{w}) z_i \text{ as } \operatorname{cov}(w, z),$$

$$N^{-1} \sum_{i=1}^{N} w_i \overline{d}_i = E(w\overline{d}),$$

this is then

$$\Delta \overline{z} = (\overline{w})^{-1}[\operatorname{cov}(w, z) + E(w\overline{d})].$$

This is the (general form of the) Price equation.

The Price equation uses an “accounting system”, and therefore also a notation system, that is different from the classical one.

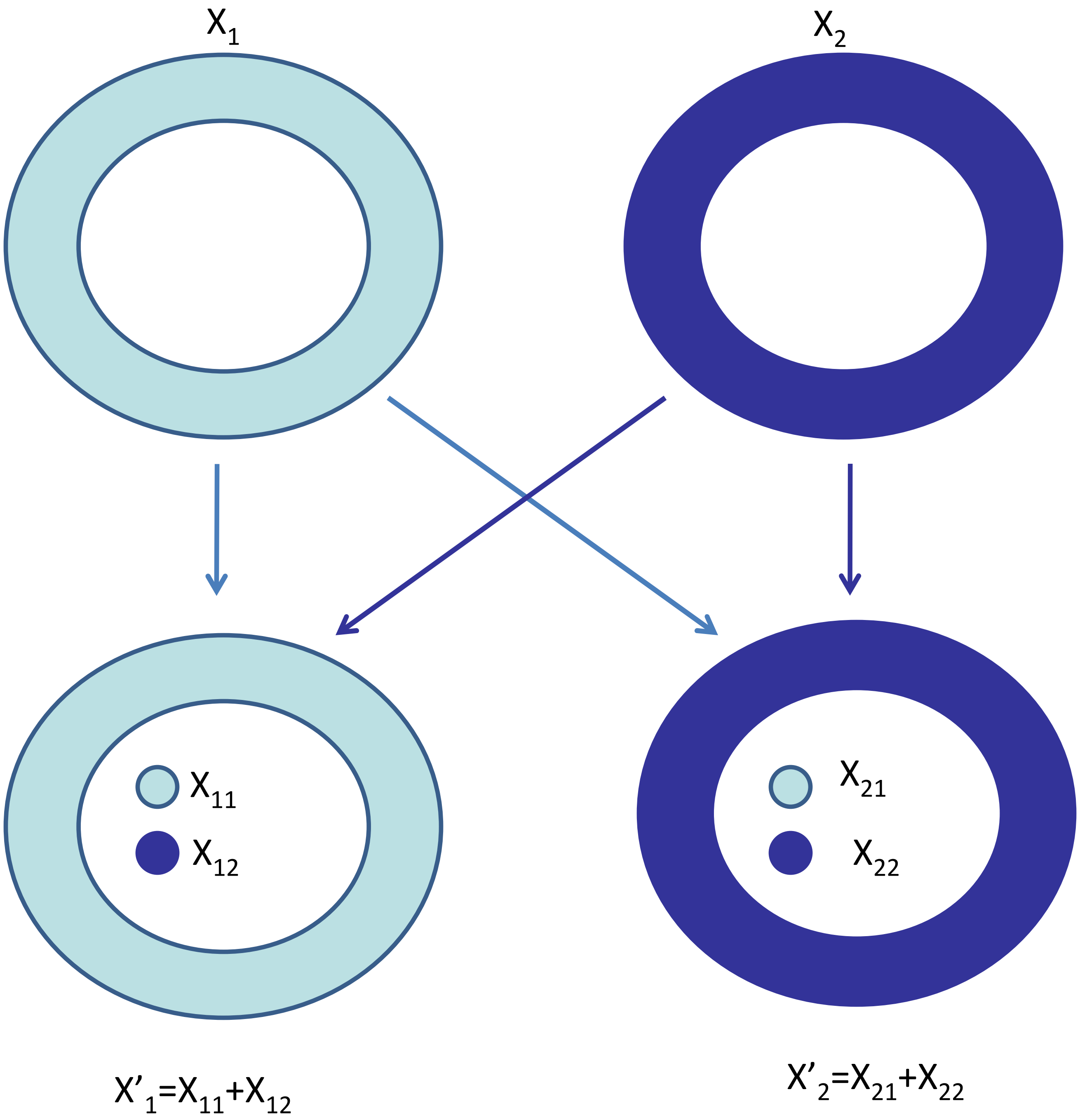
Considering for example genotype frequencies, we have

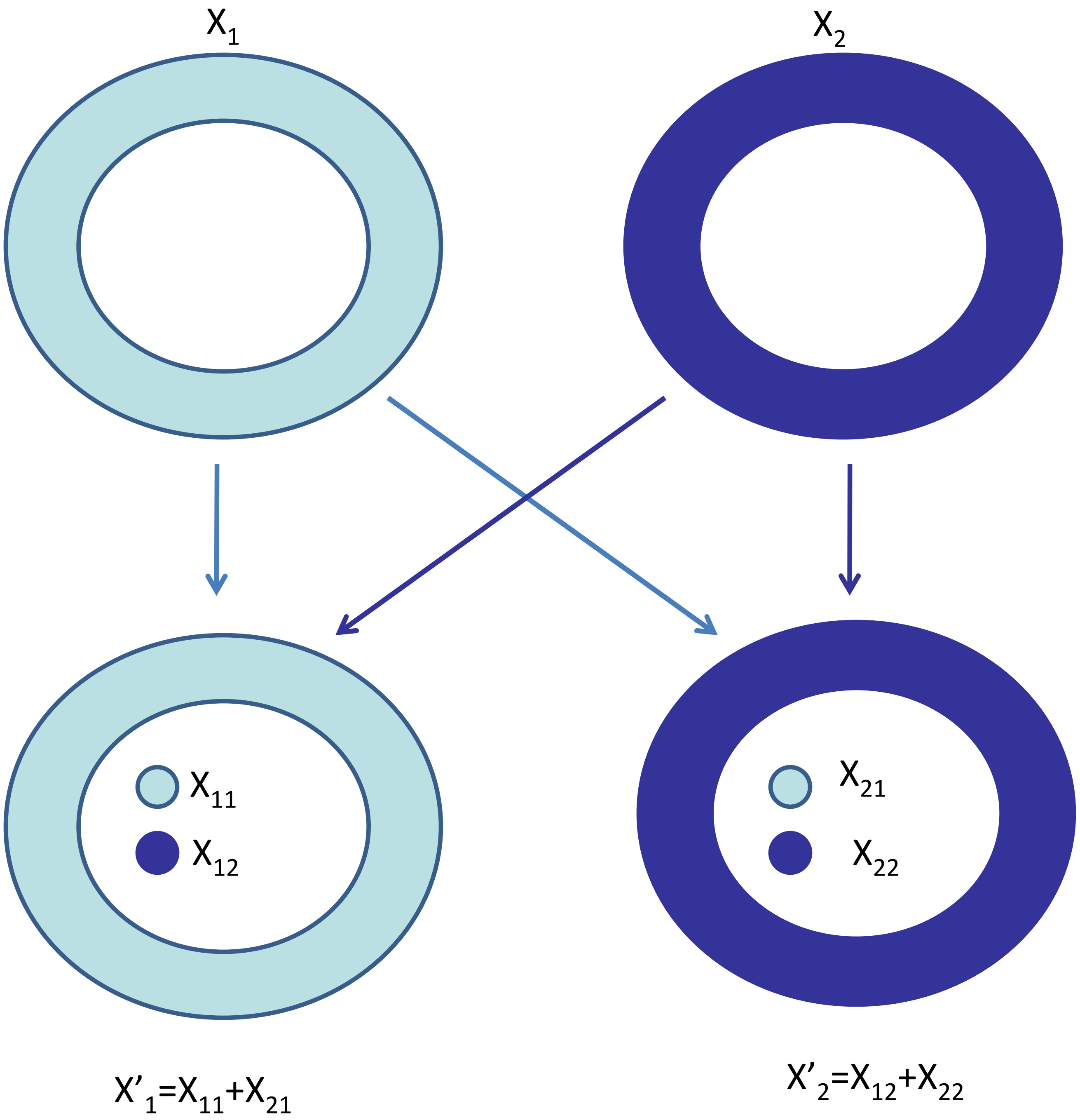
Classical population genetics notation:

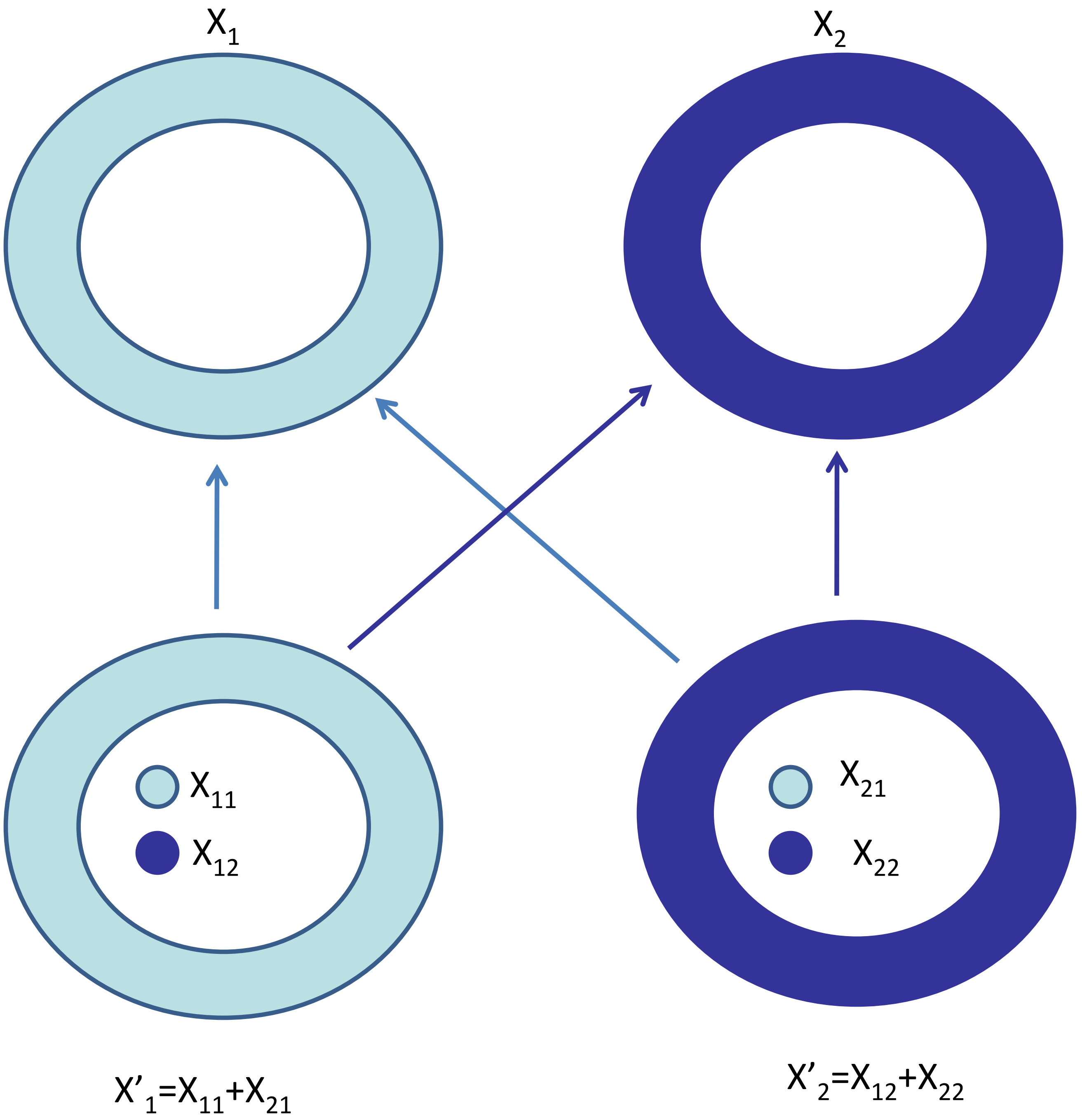
Offspring generation frequency of genotype “g” is x_g'.

“Price” notation:

Average offspring generation frequency of individuals who are the offspring of genotype “g” parents is x_{g}'.


X_1
X_2
X_{11}
X_{12}
X_{21}
X_{22}
$X'_1=X_{11}+X_{12}$
$X'_2=X_{21}+X_{22}$


X_1
X_2
X_{11}
X_{12}
X_{21}
X_{22}
$X'_1=X_{11}+X_{21}$
$X'_2=X_{12}+X_{22}$


X_1
X_2
X_{11}
X_{12}
X_{21}
X_{22}
$X'_1=X_{11}+X_{21}$
$X'_2=X_{12}+X_{22}$

Thus the Price approach with reversed arrows looks like a coalescent approach.

It would be desirable to construct a stochastic theory based on the Price equation in connection with the coalescent.

In view of this, and for other reasons, it is interesting to consider further aspects of the Price equation and its relation to classical population genetics theory for the whole-genome, non-random-mating diploid case.

To discuss potential merits of the Price approach we consider a parental and a daughter generation at the same time points in their respective life cycles. We choose the time of conception of both, and call these time points A and C respectively. It is also necessary to consider the time when the parental generation reproduces (time point B).

------A--------------B--C------>

We use a single dash (') to denote changes (Δ) in any quantity between time points A and B, a double dash ('') to denote changes between time points B and C, and Δ to denote the inter-generational change between time points A and C.

Thus $\Delta = \Delta' + \Delta''$

What merits are there in using the Price equation approach? Consider the frequency of the allele *A* at locus A. In classical theory this changes between time-points A and B, due to selection. However it does NOT change between time-points A and B *within an individual*. Further, the mean frequency with which an individual transmits this allele to his/her offspring is the same as the frequency within that individual him/herself. This makes the second term in the Price equation “close to” zero.

The within generation change Δ' depends *only* on (viability) selection, and has nothing to do with recombination phenomena or the mating scheme. It has be called the “selection change”. (This however is not quite accurate – see later.)

The B to C change Δ'' does *not* depend on selection, but *does* depend on recombination phenomena and the mating scheme. It can be called the “transmission” change. Thus another frequently-used notation is

$$\Delta = \Delta_{\text{selection}} + \Delta_{\text{transmission}}$$

What can be said, in the classical theory, about these changes, if we make no assumptions about the mating scheme or the recombination structure?

The classical theory.

Let “g” index whole genome genotypes. Write w_g as the fitness of genotype g and x_g the “time-point A” frequency of this genotype.

The population mean fitness at time point A is $\overline{w} = \sum_g x_g w_g$

Because of the different fitnesses of the various genotypes, the population frequency of genotype g at time point B has changed to

$$x_g^{'} = x_g w_g / \overline{w}.$$

Thus the change in mean fitness between time points A and B is

$$\Delta'(\overline{w}) = \sum_g x_g (w_g - \overline{w})^2 = \sigma_w^2 / \overline{w}.$$

The variance σ_w^2 is the TOTAL variance in fitness in the parental generation at the time of its conception. (This is a standard result.)

However, we cannot say *anything* about the change in mean fitness between time points B and C.

Consider next some arbitrary character *z*. All individuals of genotype *g* have character value z_g. The respective means of the character within the parental generation at time points A and B are

$$\bar{z} = \sum_g x_g z_g \qquad \text{and} \qquad \bar{z}' = \sum_g x_g z_g w_g / \bar{w}$$

The change in the mean of the character within the parental generation (i.e. between time points A and B) is

$$\Delta' \bar{z} = \bar{z}' - \bar{z} = \sum_g x_g z_g \{(w_g / \bar{w}) - 1\}$$

$$= (\bar{w})^{-1} \sum_g x_g z_g (w_g - \bar{w}).$$

(Mis)using the word covariance, this gives

$$\Delta' \bar{z} = (\bar{w})^{-1} \operatorname{cov}(z, w)$$

The right hand term is the first term on the right-hand side of the classical population genetics Price “covariance” equation, and also in the Price equation itself.

Thus this must be the “selection” term (Δ') in the Price equation, and thus the second term on the right-hand side of the Price equation must be the “transmission” term (Δ'').

What can be said in general about the “transmission term” Δ'' in both the classical and the “Price” theory?

Nothing.

Unless further assumptions are made about recombination and the mating scheme, the Price equation and the classical theory are in general useless in discussing inter-generational changes that depend on genotype frequencies.

They are both “dynamically insufficient”.

Thus if we define evolution minimally as a process relating whole-genome properties of a parental and a daughter generation, both the Price equation and the classical theory usually cannot even say anything about evolutionary properties from one generation to the next, unless further assumptions are made.

What then can we salvage from all this? There are two answers to this question.

First, if the first term on the right-hand side of the Price equation provides useful information about the action of selection (more important, if it captures all the evolutionary change in the character that can be assigned to selection) then that might be useful.

Second, *in the classical theory and the deterministic version of the Price equation there are no changes in allelic frequencies between time points B and C, whatever the mating scheme and recombination structure might be.*

Therefore the *only* transmission changes that we can calculate, and thus the only *inter*-generational changes (Δ) that we can calculate, must be those that depend only on *allelic frequencies*.

This motivated a focus on allelic frequencies and the Fundamental Theorem of Natural Selection (Fisher).

This leads to the concept of the “average effect” of an allele and, in turn, to a (brief) discussion of the Fundamental Theorem of Natural Selection (FTNS).

The average effects for fitness of all the alleles at all the loci in the genome are defined by a least-squares procedure. The “alpha version” definition of the average effect of allele A_{aj} at gene locus *j* in the genome is α_{aj}. These average effects are defined by minimizing

$$\sum_{g=1}^{G} g_s \{w_s - \overline{w} - \sum^{(g)} \alpha_{aj}\}^2,$$

where the outer summation is over all *G* genotypes in the genome, and the inner (*g*) summation is over all alleles at all loci occurring within genotype *g*, with allele α_{aj} counted in as many times (0, 1 or 2) as the allele A_{aj} occurs in genotype *g*.

The α_{aj} values are given implicitly as the solution of the matrix equation

$$\mathrm{A}\alpha = \overline{w}\,\Delta$$

for a (known) matrix A. Δ is the vector of inter- (and intra-) generational changes in allelic frequencies.

The sum of squares removed by fitting the α_{aj} values is the (whole genome) additive genetic variance σ_A^2, whose value is 2 $\overline{w}\,\alpha'\Delta$.

It is important to note that the values of the average effects depend the allelic frequencies, and thus change as these frequencies change.

We now think of the fitness of genotype *g* not as w_g, but instead as the "allele-derived" linear combination

$$\hat{w}_g = \overline{w} + \sum^{(g)} \alpha_{aj},$$

where the (*g*) summation is as before. Note that using this does not change the "time-point A" population mean fitness, since

$$\sum_{g=1}^{G} \hat{w}_g x_g = \overline{w}.$$

Although the α_{aj} values depend on allelic frequencies and thus change between time-points A and B as these frequencies change, we consider the time point A to B "partial" change in mean fitness, *where we ignore changes in the* α_{aj} *values*. This partial change is

$$\sum_{g=1}^{G} \{\Delta' x_g\}(\overline{w} + \sum^{(g)} \alpha_{aj}).$$

This is, exactly, $\sigma_A^2 / \overline{w}$, where σ_A^2 is the whole-genome additive genetic variance in fitness.

Since this change depends on changes in gene frequencies only, it is thus also an inter-generational change. That is to say

$$\Delta_P \overline{w} = \sigma_A^2 / \overline{w}. \qquad \text{(P = partial.)}$$

This is the (adult version of) the Fundamental Theorem of Natural Selection. It is an exact whole-genome result, independent of the mating scheme.

A central point is that it is the *additive genetic variance,* not the total variance, that is relevant to the evolutionary process. This makes sense – it is the variance in fitness attributable to “genes within genotypes”.

What are some of the misinterpretations of the FTNS given by devotes of the Price equation?

Gardner, in "The Price equation", Current Biology Vol 18, 2008:-

"The [FTNS] is easily proven using the Price equation. It is:

$$\Delta\overline{w} = \sum_g p_g (w_g - \overline{w})^2 = \sigma_w^2 / \overline{w}.$$

This is dreadful.

(i) It is the standard intra-generational (time point A to time-point B) result, and is thus is not relevant to evolution.

(ii) It uses the total variance in fitness, and not the additive genetic variance in fitness.

(iii) It implies that the mean fitness is non-decreasing from one generation to the next, whereas it can and often does decrease.

Another erroneous statement of the FTNS (Rice, *Evolutionary Theory*) deriving from the Price equation is:-

$$\Delta\overline{W} = \sigma_A^2 / \overline{W} + \overline{\delta}_T(W).$$

This states that the total change in mean fitness between parental and daughter generations is the additive genetic variance divided by the mean fitness, plus a further term (whose value cannot be known in practice) that describes the change in mean fitness “due to transmission”

This comment implies that the first term is the change in mean fitness between time points A and B.

It follows that this equation is not correct. The change in mean fitness between time points A and B is (as indicated previously)

$$\Delta'\overline{w} = \sum_g p_g(w_g - \overline{w})^2 = \sigma_w^2 / \overline{w}.$$

Another aspect of the FTNS is that a basic assumption is that the environment changes rapidly, so that only “one generation to the next” results (such as the FTNS) are useful.

In particular, stationarity results are not useful.

Consider now the earlier claim that the first term on the right-hand side of the Price equation term does not necessarily capture all the evolutionary change in the character that can be assigned to selection.

Suppose that the character "z_{g}" is the linear approximation to the fitness of genotype *g* via summation of average effect, a central concept in the FTNS. That is, (see a previous slide),

$$z_g = \overline{w} + \sum^{(g)} \alpha_{aj}.$$

Then the first term in the Price equation becomes, with z_g defined as $z_g = \overline{w} + \sum^{(g)} \alpha_{aj}$,

$$\sum_g \Delta'(x_g)\{\overline{w} + \sum^{(g)} \alpha_{aj}\}$$

But this is the partial increase in mean fitness within the parental generation between time points A and B, not the total increase between those time points. Thus the first term on the right-hand side of the Price equation captures only a part of the change in mean fitness that is due to natural selection.

The following “optimization” principle is claimed to arise using Price equation methods.

If all linear combinations of changes in the gene frequencies of all genes at all loci in the genome are zero, then by definition the individuals in the population have “solved the optimization problem” and “there is no scope for selection”.

This is the same as saying: If, for all genes at all loci in the genome the daughter generation frequencies are the same as the parental generation frequencies, then individuals in the population have “solved the optimization problem”.

This is pathetic. It says nothing.

Further problems with this “optimizing” concept:

1. Fitnesses 1.0 (A_1A_1) , 0.9 (A_1A_2), 1.0 (A_2A_2), both gene frequencies = ½, random mating. There is no change in gene frequencies. But the equilibrium is unstable, so there is scope for increasing mean fitness. (Consider also stochastic changes.)

2. If gene frequencies do not change between generations *t* and *t*+1, is it true that they will not change between generations *t*+1 and *t*+2?

3. Even when gene frequencies do not change under natural selection, genotype frequencies can be changing. Thus under natural selection the population is still “continuing to optimize”.

In the classical theory, what is optimal about natural selection in a Mendelian population?

Natural selection acts in such a way that the gene frequency changes brought about by natural selection are such that, for a given increase in mean fitness between parent and daughter generations, they maximize that part of the increase in mean fitness that is due to “genes within genotypes”.

(An incorrect optimizing principle found often in the literature is that, under natural selection, gene frequencies change in such a way as to maximize the rate of increase in mean fitness. This is not even true in the one-locus random-mating case.)

(i) The Price equation involves parent/offspring (P/O) covariances. How does it relate to standard population genetics covariance theory?

Standard P/O covariances are very complex. They rely on sums over all loci of the additive genetic variance at each locus, sums of all possible additive × additive genetic variances, sums of all possible additive × additive × additive genetic variances, …, assume no fitness differentials, and are complicated by the mating structure. In general they are not really known. (Problems associated with the "missing heritability".)

The Price approach might help here, since it deals naturally with covariances (and thus correlations).