CAMBRIDGE
UNIVERSITY PRESS

Article

C. C. Li and Quasi-Random Mating

Alan E. Stark

School of Mathematics and Statistics FO7, The University of Sydney, Sydney, New South Wales, Australia

Abstract

A simple model by which Hardy-Weinberg proportions are attained in a single generation while maintaining gene frequencies is stated and illustrated. The title 'Quasi-random mating' is proposed. Confusion about the Hardy-Weinberg principle can be avoided only if there is clear separation between the basic deterministic model and factors influencing a population's structure. Eighty years passed before C. C. Li coined the term 'pseudo-random mating'. The lesson taught by Li has not been taken on board.

By a sleight of hand, G. H. Hardy (1908) indicated how the proportions $\{q^2, p^2, 2pq\}$ could be produced and maintained by 'random mating' while keeping the gene frequencies unchanged. Later, it was perceived that Wilhelm Weinberg (1908) had the same idea. Hardy (1908) and Weinberg (1908) have been cited or honored countless times. Hardy has been portrayed in a feature film as the eccentric professor of mathematics. Weinberg is less well known in the English-speaking world. An account of his life and work is given by Sperlich and Früh (2015).

C. C. Li (1988) proved that Hardy-Weinberg proportions can be maintained by what he called 'pseudo-random' mating. Kimura (1988, pp. 87–91) has a section entitled 'Gene Frequency and Mating System'. He is critical of the way the Hardy-Weinberg principle is treated in textbooks, pointing out that its most useful application is relating gene and genotypic frequencies rather than emphasizing how it explains stability. The preface to the book is dated February 1988, the same year in which Li (1988) appeared. Kimura's explanation of the Hardy-Weinberg principle is conventional and there is no way of knowing whether he would have changed it in the light of Li's finding.

A search on the internet using the phrase 'Hardy-Weinberg principle' yields similar responses to that by Wikipedia (2023). After giving the array $\{q^2, 2pq, p^2\}$, Wikipedia states that the array is used primarily to test for population stratification and other forms of nonrandom mating. The inference is that if mating is not 'random', frequencies will not follow the above array. Evidently there is no impetus in the genetics community to incorporate the fact shown by Stark (2006) that the array can be produced in one round of nonrandom mating while keeping the gene frequencies constant. The point of this note is to restate the model in the simplest form with a numerical example. A suggested title for model (1) defined in the next section is 'quasi-random mating'.

The final section comments briefly on the use of the Hardy-Weinberg principle in genetic association studies that exploit the notion called Mendelian randomization.

The Basic Model — 'Quasi-Random Mating'

The usual entry to population genetics theory begins with the Hardy-Weinberg law. Consider an autosomal locus with two alleles A and B and genotypes AA, BB and AB numbered 1, 2 and 3. Mating pairs are formed in the current generation to produce offspring in the next. The proportions of the mating pairs are given symbolically in Table 1. The elements c_{ij} are non-negative and symmetrical in value ($c_{ij} = c_{ji}$) and sum to 1.

Malécot (1969) is the English version of *Les Mathématiques de l'Hérédité* (published in 1948), which was one of the first systematic introductions to population genetics theory. In 1948, Malécot was still not aware of Weinberg (1908) and referred to Hardy's (1908) law.

Malécot's account is faultless but, being expressed in probabilistic terms, it obscures the fact that Hardy's model is deterministic. This would not create a problem except that it has led to the construction of an elaborate edifice in which the original model is embellished with the details of real populations.

There is a further problem in that Hardy's model is incomplete. Li (1988) shows that Hardy-Weinberg proportions can be maintained by nonrandom mating, which he calls 'pseudo-random mating'. This property is implicit in a formula given by Stark (1980).

Stark (2006) shows that Hardy-Weinberg proportions can be reached in one generation from any genotypic distribution, assuming that males and females are equally distributed, as is now demonstrated. Suppose that the genotypic proportions are

$$G_1 = q^2 + Fpq;\ G_2 = p^2 + Fpq;\ G_3 = 2pq(1 - F),$$

F measures departure from Hardy-Weinberg form and the gene frequencies are $q = (2G_1 + G_3)/2; p = (2G_2 + G_3)/2 = 1 - q$.

Table 1. Symbolic mating proportions reproducing offspring

Male\Female	AA	BB	AB	Total
AA	c_{11}	c_{12}	c_{13}	G_1
BB	c_{21}	c_{22}	c_{23}	G_2
AB	c_{31}	c_{32}	c_{33}	G_3
Total	G_1	G_2	G_3	1

Table 2. Mating proportions for parameters $q = 1/4$, $F = 1/3$, $h = 1/20$, elements to be divided by 512. Hardy-Weinberg proportions in offspring are $\{1, 9, 6\}/16$.

Male\Female	AA	BB	AB	Total
AA	9	41	14	64
BB	41	201	78	320
AB	14	78	36	128
Total	64	320	128	512

The mating frequencies are

$$c_{ij} = G_i G_j (1 + h e_i e_j / v) \tag{1}$$

where $e_1 = p(F-1)/(q+Fp); e_2 = q(F-1)/(p+Fq); e_3 = 1$; and $v = pq(1-F^2)/((q+Fp)(p+Fq))$.

The gene frequencies are not changed through the action of (1). Subject to constraints, h can be chosen over a wide range, allowing uncountable possibilities for varying the mating regime but still producing, in one generation, offspring distributed according to the Hardy-Weinberg formulae:

$$H_1 = q^2; H_2 = p^2; H_3 = 2pq. \tag{2}$$

This can be verified by calculating the offspring frequencies by applying Mendel's rule to (1):

$$AA\, c_{11} + c_{13} + c_{33}/4;\ BB\ c_{22} + c_{23} + c_{33}/4;\ AB\ 2c_{12} + c_{13} + c_{23} + c_{33}/2$$

Table 2 illustrates the model for $q = \frac{1}{4}$, $F = \frac{1}{3}$, $h = 1/20$. The offspring distribution is $\{1/16, 9/16, 6/16\}$.

The Hardy-Weinberg model, as explained by Hardy (1908), produced the equilibrium distribution characterized in the notation used here by

$$4H_1 \cdot H_2 = (H_3)^2 \tag{3}$$

Hardy used expression (3) simply as a shorthand for (2), which does not convey information about $\{c_{ij}\}$. Malécot (1969, p. 14) identifies (3) as 'Hardy's Law'. The set $\{G_1, G_2, G_3\}$ conforms to (3) if and only if $F = 0$.

Mendelian Randomization

In many studies, counts of genotypes have produced proportions approximately in Hardy-Weinberg form. As a result it is used as a convenient benchmark for assessing the validity of data. Often the inference has been drawn that the mating regime of the population is 'random'. The object of this paper is to stress that there is an uncountable number of ways, other than 'random mating', but close to random mating, which can produce the Hardy-Weinberg distribution. Taking the value $h = 0$ in (1) specifies what is given the label 'random mating'. Taking negative and positive values of h near to zero provides mating regimes close to 'random mating' with Hardy-Weinberg frequencies in offspring for any starting structure.

Rodriguez et al. (2009) give an example of the epidemiological concept known as Mendelian randomization (MR). They state: 'A particular genetic feature of randomly breeding populations is that of Hardy-Weinberg equilibrium (HWE)' (p. 506). 'In a very large (outbred) population there should be exact HWE at the point of conception' (p. 512). They claim that MR permits causal inference between exposures and a disease. They suggest that property (3) could be used to construct a test for agreement with the Hardy-Weinberg distribution (p. 506).

In studies such as Gu et al. (2000), the expectation is that a locus will have approximate Hardy-Weinberg proportions so that a nonsignificant test result in the control group assures a valid comparison with affected subjects.

Gu et al. (2000) classified 1032 subjects with respect to the *CYP2A6* locus, noting those who possessed, or did not possess, the *160H* allele. Possessing the *160H* allele was associated with later age to begin smoking and greater likelihood to quit smoking. From the point of view of this paper the authors validated their findings by comparing counts of the *160H* allele with predictions based on the Hardy-Weinberg formulae (distribution [2]).

Bosco et al. (2012) is an example of taking a simple test of concordance of a set of counts with hypothetical distribution (2) and building an elaborate theory with no obvious advantage to applied population genetics. The authors pursue a will-o'-the-wisp: 'In order to identify the properties of the equilibrium state revealed by the system's time series one should apply dynamical criteria and not statistical ones' (p .9). Although Bosco et al. cite Li (1988) and Stark (2006), the messages of Li and Stark are not reflected in their analysis.

It is ironic that much of the lip service paid to Hardy's law is poorly directed, as Salanti et al. (2005) show in detail. The authors evaluated dozens of genetic association studies published in high-prestige journals. They conclude that 'testing and reporting for HWE is often neglected and deviations are rarely admitted in the published reports. Moreover, power is limited for HWE testing in most current genetic association studies' (p. 840).

Fisher (1922, p. 324) uses criterion (3) of the previous section in deriving the equilibrium of a locus under selection, showing clearly how he perceived that Hardy's (1908) paper had removed any doubts about how a population's genetic composition could be maintained. Charlesworth (2022) acknowledges the huge contribution of Fisher (1922) but points out two errors, subsequently resolved, which do not diminish the achievement of that paper.

In that paper, Fisher refers to quantitative genetics theory developed by himself in 1918 that gives insight to the correlation between relatives for traits such as human stature. The variance and the correlation between parents, the tendency referred to as homogamy by Fisher, are central to the dissection of such traits.

Sella and Barton (2019) describe the use of genomewide association studies (GWASs) in humans to analyze the genetic basis of complex (quantitative) traits. Their article is wide ranging, taking in many facets, as would be expected after a century of intensive research on wild and commercial species. The following quotation illustrates the debt owed to Fisher (1918):

With numerous loci affecting a trait, how should we think about the relationship between an individual's genotypes at these loci and the person's trait value? In principle, quantitative genetics can describe any relationship between genotype, environment, and phenotype. The variance of a trait due to genetics (V_G) can be partitioned into additive (V_A), dominance (V_D), and a combined epistatic component (V_I), which itself can be partitioned into two-locus (V_{AA}, V_{AD} and V_{DD}) and multilocus components (V_{DDD}, etc.); higher-order terms in this expansion are defined through the residuals of lower-order ones. Fisher introduced this expansion in his seminal 1918 paper, showing how in principle the components can be estimated from the phenotypic correlations among relatives. (p. 464)

Clark (2023) uses the theory in a study of social status in English pedigrees over a long period. He found three notable results: strong persistence of social status across family trees; decline in correlation with genetic distance in the lineage is unchanged over the period 1600–2022; the correlations follow those of a simple genetic model of additive genetic determination of status.

Genealogies, including 422,215 individuals born in the period 1600–2022, were assembled. Six measures of social status, one of which is literacy, were scored. Correlations for the measures were calculated for relatives up to fourth cousins.

For the birth period 1725–1869, the correlation between relatives for literacy decreased from .407 for full sibs to .146 for fourth cousins. Measures such as these are explained by Clark (2023) in terms of m, the correlation between parents, and h^2, a measure of heritability for the trait.

Clark (2023) gives a table (Table A6, p. 32) of implied underlying phenotype correlation in marriage scores for the period 1837–2022. In five adjoining intervals over this period Clark gives the correlation between marriage partners as .480, .464, .384, .346, and .275. These were based on the score of the groom and an imputed measure of the bride using her father's score. The relevance of Clark's study for this paper is that choice of mates in humans is far from 'random mating'.

A book review by Coop and Przeworski (2022) includes the following:

> The author, Dr. Kathryn Paige Harden, is a Professor of Psychology at the University of Texas, Austin, who specializes in behavioral genetics. Her book starts from the premise that human behaviors, and in particular educational attainment, are 'heritable,' i.e., that within a study sample, some fraction of the phenotypic variance is explained by differences in genotypes. (p. 846)

In brief, Coop and Przeworski (2022) conclude that this view is not justified by current understanding. One suspects that they may have a similar view of Clark's (2023) findings with respect to social status. Coop and Przewoski part company from Harden when Harden claims that a (Mendelian) lottery is a perfect metaphor for genetic inheritance. This gets into the difficult area of group comparisons such as comparing IQ scores in different racial groups.

Stark (2023) presents a different approach to maintaining a population's genetic structure and Hardy-Weinberg equilibrium, which is the main focus of this paper.

References

Bosco, F., Castro, D., & Briones, M. R. S. (2012). Neutral and stable equilibria of genetic systems and the Hardy–Weinberg principle: Limitations of the chi-square test and advantages of auto-correlation functions of allele frequencies. *Frontiers in Genetics*, *3*, 1–9. https://doi.10.3389/fgene.2012.00276

Charlesworth, B. (2022). Fisher's historic 1922 paper *On the dominance ratio*. *Genetics*, *220*, 1–8. https://doi.org/10.1093/genetics/iyac006

Clark, G. (2023). The Inheritance of Social Status: England, 1600–2022 [Discussion Paper 17835]. The Centre for Economic Policy Research (CEPR). https://repec.cepr.org/repec/cpr/ceprdp/DP17835.pdf

Coop, G., & Przeworski, M. (2022). Lottery, luck, or legacy. A review of 'The Genetic Lottery: Why DNA matters for social equality' (by Kathryn Paige Harden). *Evolution*, *76*, 846–853. https://doi.10.1111/evo.14449

Fisher, R. A. (1918). The correlation between relatives on the supposition of Mendelian inheritance. *Transactions of the Royal Society of Edinburgh*, *52*, 399–433.

Fisher, R. A. (1922). On the dominance ratio. *Proceedings of the Royal Society of Edinburgh*, *42*, 321–341. https://doi.10.1017/S0370164600023993

Gu, D. F., Hinks, L. J., N. E. Morton, N. E. & Day, I. N. M. (2000). The use of long PCR to confirm three common alleles at the *CYP2A6* locus and the relationship between genotype and smoking habit. *Annals of Human Genetics*, *64*, 383–390. https://doi.10.1046/j.1469-1809.2000.6450383.x

Hardy, G. H. (1908). Mendelian proportions in a mixed population. *Science*, *28*, 49–50. https://doi.10.1126/science.28.706.49

Kimura, M. (1988). *My thoughts on biological evolution*. Springer.

Li, C. C. (1988). Pseudo-random mating populations. In celebration of the 80th anniversary of the Hardy-Weinberg law. *Genetics*, *119*, 731–737. https://doi.10.1093/genetics/119.3.731

Malécot, G. (1969). *The mathematics of heredity*. W. H. Freeman and Company.

Rodriguez, S., Gaunt, T. R., & Day, I. N. M. (2009). Hardy-Weinberg equilibrium testing of biological ascertainment for Mendelian randomization studies. *American Journal of Epidemiology*, *169*, 505–514. https://doi.10.1093/aje/kwn359

Salanti, G., Amountza, G., Ntzani, E. E., & Ioannidis, J. P. A. (2005). Hardy–Weinberg equilibrium in genetic association studies: An empirical evaluation of reporting, deviations, and power. *European Journal of Human Genetics*, *13*, 840–848. https://doi.10.1038/sj.ejhg.5201410

Sella, G., & Barton, N. H. (2019). Thinking about the evolution of complex traits in the era of genome-wide association studies. *Annual Review of Genomics and Human Genetics*, *20*, 461–493. https://doi.10.1146/annurev-genom-083115-022316

Sperlich, D., & Früh, D. (2015). *Wilhelm Weinberg: Der zweite Vater des Hardy-Weinberg-Gesetzes*. Basilisken-Presse.

Stark, A. E. (1980). Inbreeding systems: Classification by a canonical form. *Journal of Mathematical Biology*, *10*, 305. https://doi.10.1007/BF00276989.

Stark, A. E. (2006). A clarification of the Hardy-Weinberg law. *Genetics*, *174*, 1695–1697. https://doi.10.1534/genetics.106.057042

Stark, A. E. (2023). Stable populations and Hardy-Weinberg equilibrium. *Hereditas*, *160*, 19. https://doi.org/10.1186/s41065-023-00284-x

Weinberg, W. (1908). Über den Nachweis der Vererbung beim Menschen. *Jahresh Ver Vaterl Naturkd*, *64*, 368–382.

Wikipedia. (2023). http://en.wikipedia.org/wiki/Hardy-Weinberg_principle