A Contribution to the Genetic Epidemiology of Structured Populations

Alan E. Stark*

School of Mathematics and Statistics FO7, University of Sydney, NSW, 2006, Australia

Abstract: A mating system, previously derived, which is more general than random mating is defined by the gene frequency *q* and a parameter *F* which measures divergence from Hardy-Weinberg proportions commonly used in genetic analysis. *F* can be viewed as the average *coefficient of inbreeding* in a population, the use emphasized here. Also it can characterize the variation in gene frequency in a stratified population. Taking *q* as fixed, the distribution of *F* over values admissible under the general mating system is derived by simulation. The mating system may be seen to be based on indifference as to choice of mates. This is the first object of the paper. The second uses the derived distribution of *F* to make a Bayesian estimate of *F* from a single sample of genotypic counts. Such an estimate has a number of uses in genetic analysis.

Keywords: Genetic Equilibrium, Hardy-Weinberg Law, Mate choice indifference, Inbreeding coefficient, Bayesian estimation.

INTRODUCTION

When analysing a population, at the outset we often assume a state of equilibrium. In genetic analysis this is often Hardy-Weinberg equilibrium (HWE), expressed in the set of proportions

$$\{q^2, 2q(1-q), (1-q)^2\} \quad (1)$$

In (1) *q* is the frequency (proportion) of the first of two alleles, here denoted as *U* and *T*, in the population (see Edwards [1], Mayo [2] and Russell [3]). We restrict ourselves to a single autosomal locus with these two alleles. More generally, the population will have genotypic proportions, as given, in different notation, by Morton [4]:

$$P(UU) = q^2(1 - F) + qF \quad (2)$$

$$P(UT) = 2q(1 - q)(1 - F) \quad (3)$$

$$P(TT) = (1 - q)^2(1 - F) + (1- q)F \quad (4)$$

Morton, page 109, introduces F as follows: "We seek a single parameter, called the (coefficient of) *inbreeding* F, not dependent on the gene frequencies (although its range is so dependent), which will predict genotype and mating type frequencies in populations not necessarily panmictic (randomly mating)." [4]

Bittles and Black [5] state "It ... is not surprising that the prevailing Western public and medical opinion with regard to consanguinity is largely negative." As can be seen by comparing (1) and (2), when *q* is very small and F is positive, the frequency of type *UU* may be considerably raised relatively under consanguinity (F > 0). Thus a deleterious trait which is recessively inherited will have raised incidence compared with a population in which F = 0. Bittles and Black point out that there may be countervailing benefits from consanguineous unions in some communities [5].

At the time of publication (2010), Bittles and Black noted that "close-kin marriage continues to be preferential in many major populations". They cite Strømme *et al*. who write "Progressive encephalopathy (PE) is a heterogeneous group of individually rare diseases, many with an autosomal recessive mode of inheritance. We estimated the increased risk of PE associated with consanguinity. ... The population attributable risk due to parental consanguinity was 50.3% in the Pakistani sub-population (of Oslo)." [6] Consanguineous marriage was defined as a union between partners who were first cousins or more closely related.

Bittles and Black have a section on consanguinity and "complex diseases". In many of these, as the label suggests, the aetiology is unresolved. These authors refer to some of the difficulties of interpretation and give a wide-ranging review of these questions.

Risch [7], writing at the time when the entire DNA sequence of Man was about to be revealed, discussed the problem of unravelling the genetic basic of complex diseases. One such, HBSL, is reported in Taft *et al*. [8] An account of this for the general reader is given by Kaminsky and the parents of one of the affected children, Stephen and Sally Damiani [9]. OMIM states:

*Address correspondence to this author at the School of Mathematics and Statistics, F07, University of Sydney, NSW 2006, Australia; Tel: +61 2 9351 2222; Fax: +61 2 9351 8938; E-mail: alans@exemail.com.au

E-ISSN: 1929-6029/15 © 2015 Lifescience Global

"HBSL is caused by homozygous or compound heterozygous mutation in the DARS gene (603084) on chromosome 2q21.3" and further, "The transmission pattern in the families with HBSL reported by Taft *et al*. was consistent with autosomal recessive inheritance." [15] In the light of this it is interesting to note that both Sally and Stephen Damiani have Armenian heritage, as stated on page 154 of their book *Cracking the Code*. This is relevant to another use of *F* as a measure of variability of gene frequencies between sub-populations as well as an average coefficient of inbreeding.

As will be clear later, F (*F* in our notation) does not fix mating type frequencies as well as genotype frequencies of the 'standing' population. Figure **5** in Stark and Seneta (2014) demonstrates this fact [10]. When F = 0, the frequencies in (2)-(4) reduce to those in (1). As Morton points out, F (*F*) can take negative values subject to restrictions decribed below [4]. However, for the case of inbreeding, F is positive.

The object of this paper is twofold: to use our model of genetic equilibrium to derive a simple prior distribution of *F*; to use it to make a Bayesian estimate of *F* from a sample of genotypic counts. We have given the model earlier but for the reader's convenience repeat the outline here. Next is given the method of calculating the prior distribution of *F* and finally the Bayesian estimation of the posterior distribution using a sample of counts.

THE GENERAL MATING EQUILIBRIUM MODEL

We deal only with a single autosomal locus with two alleles U and T with frequencies in the population q and p $(q+p=1)$. Throughout q remains constant because this is guaranteed by the nature of the selected mating system. A set of frequencies of genotypes $\{UU,UT,TT\}$ can be represented in terms of q and a measure of departure from Hardy-Weinberg (HW) form F as, say, $\boldsymbol{a}' = \{q^2+Fpq, 2pq-2Fpq, p^2+Fpq\}$. These will vary according to F and will be denoted generally by $\{f_0, f_1, f_2\}$, $(f_0+f_1+f_2=1)$, that is $f_0 = q^2+Fpq$, etc.

The population is maintained in discrete generations according to the mating scheme

$$\begin{bmatrix} UU\times UU & UU\times UT & UU\times TT \\ UT\times UU & UT\times UT & UT\times TT \\ TT\times UU & TT\times UT & TT\times TT \end{bmatrix}$$

with commensurate pairing frequencies given by the matrix

$$C = \begin{bmatrix} f_{00} & f_{01} & f_{02} \\ f_{10} & f_{11} & f_{12} \\ f_{20} & f_{21} & f_{22} \end{bmatrix}.$$

C is symmetric, that is $f_{ij} = f_{ji}$, with row and column sums $\{f_0, f_1, f_2\}$. This triple of sums is the parental frequency distribution.

Below we use C in the extended (row vector) form

$$u' = \{f_{00}, f_{01}, f_{02}, f_{10}, f_{11}, f_{12}, f_{20}, f_{21}, f_{22}\}.$$

To follow the progression of generations we need Mendel's coefficients of heredity given in matrix form by

$$M = \begin{bmatrix} 1 & 1/2 & 0 & 1/2 & 1/4 & 0 & 0 & 0 & 0 \\ 0 & 1/2 & 1 & 1/2 & 1/2 & 1/2 & 1 & 1/2 & 0 \\ 0 & 0 & 0 & 0 & 1/4 & 1/2 & 0 & 1/2 & 1 \end{bmatrix}.$$

Then the frequency distribution of juveniles is calculated from

$$j' = (Mu)'$$

which in detail is

$$j = \left\{ \begin{matrix} f_{00} + \dfrac{f_{01}+f_{10}}{2} + \dfrac{f_{11}}{4}, \dfrac{f_{01}}{2} + \\ f_{02} + \dfrac{f_{10}+f_{11}+f_{12}}{2} + f_{20} + \dfrac{f_{21}}{2}, \dfrac{f_{11}}{4} + \dfrac{f_{12}+f_{21}}{2} + f_{22} \end{matrix} \right\}'.$$

The population is in equilibrium, that is: the distribution of juveniles is the same as that of adults, if and only if matrix C has, in addition to the properties given above, the special property

$$f_{11} = 4f_{02} = 4f_{20}. \qquad (5)$$

The notation used here is a modified version of that given in Stark & Seneta [10, 11].

Identity (5) allows for non-random mating (NRM) as well as random mating (RM).

A schematic illustration of the admissible region is given in Figure **1**. The details are explained fully in Stark & Seneta [10] and Stark [12]. For a fixed value of *q*, points within the region are given by the set of coordinates $\{F, f_{11}, f_{01}\}$. Table **1** gives the coordinates

of points in Figure **1**. The admissible set of points are within the region defined by vertices Q V Z D E A.

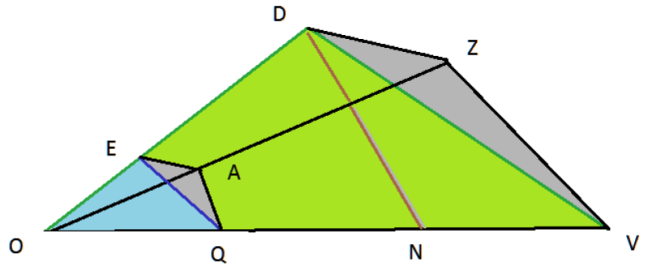

Figure 1: Schematic illustration of the bounding region of admissible sets of F, f_{11} and f_{01} for $1/4 < q < 1/2$.

The region defined by vertices O,Q,A, and E are not part of the admissible region. The coordinates of the vertices are given in Table **1**. Other points of reference, not given in Table **1**, are: O $(-q/p,0,0)$; B $((p-2q)/(3p),0,0)$; N $((2p-q)/(3p),0,0)$.

Table 1: The Coordinates of the Vertices of the Admissible Region as Functions of q

Vertex	F	f_{11}	f_{01}
A	$-(p-q)^2/(4pq)$	0	$q-1/4$
V	1	0	0
D	$(p-2q)/(3p)$	$4q/3$	0
E	$-(1-4q+6q^2)/(6pq)$	$2(4q-1)/3$	0
Z	$(2p-q)/(3p)$	0	$2q/3$
Q	$(3q-1)/(3q)$	0	0

THE SAMPLING PROCESS

As noted in the previous section, each point in the admissible space corresponds to a mating system in equilibrium. The distribution of genotypes in the parental and equally the offspring generations are identified by *q* and *F*. Since *q* can be regarded as fixed, the 'universe' of systems satisfying the conditions of pair formation is described by the variation in *F*. In this section we seek the distribution of *F* by sampling the universe and find it to be of simple form.

The sampling process starts from an arbitrary value of *q*, taken without loss of generality to be in the interval $0 < q \le 1/2$. First, a value of *F*, the parameter of divergence from Hardy-Weinberg form, is selected from the uniform distribution over the interval $-q/p \le F \le 1$. This leads to the population genotypic distribution

$$\{f_0 = q^2 + Fpq, f_1 = 2pq - 2Fpq, f_2 = p^2 + Fpq\}$$

Then f_{11} is chosen randomly from the uniform distribution over the interval $0 \le f_{11} \le f_1$, leading to $f_{02} = f_{11}/4$ and $f_{20} = f_{02}$. Then, if $f_{02} \le f_0$, f_{01} is chosen from the uniform distribution over the interval $0 \le f_{01} \le f_0 - f_{02}$, otherwise discard the sample. Continuing, calculate $f_{00} = f_0 - f_{01} - f_{02}$, $f_{10} = f_{01}$, $f_{12} = f_1 - f_{10} - f_{11}$, $f_{21} = f_{12}$ and $f_{22} = f_2 - f_{20} - f_{21}$. If $f_{22} < 0$, discard. Finally, if $f_{12} \ge 0$, retain the sample and add to the pool of admissible systems.

Cavalli-Sforza and Bodmer give data relating to the MN blood-group locus: 47 M, 52 MN, and 12 N individuals [13, p. 43]. There are 76 N genes from a total 222 and so the frequency of gene N is *q* = 38/111.

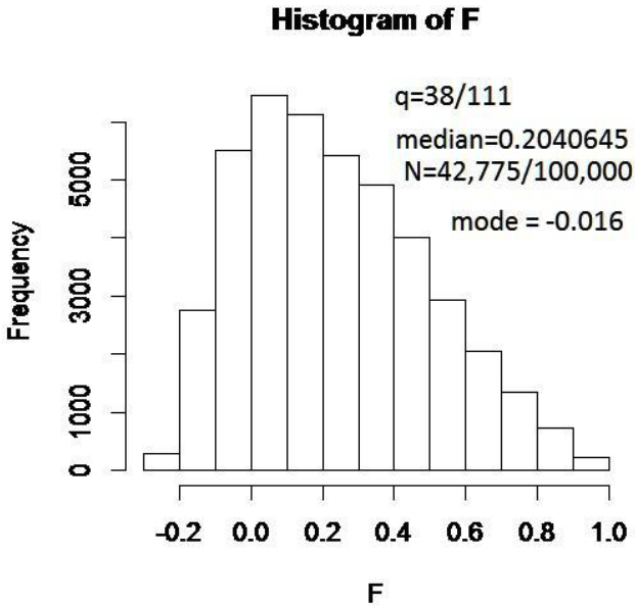

Figure 2: Empirical distribution of F for q = 38/111.

Figure **2** shows the result of sampling from a population with gene frequency *q* = 38/111. The result supports our conjecture that the set of admissible systems can be characterized by specifying *q* and the distribution of *F* as triangular with base consisting of the interval $F_\varpi \le F \le 1$, where $F_\varpi = -(1-4q+6q^2)/(6pq)$, and height $2/(1-F_\varpi)$, since the area of the triangle is unity.

We denote the observed median value of *F* in the sample as F_χ. Then the mode of *F* is calculated by

$$F_\eta = 1 - \frac{2(1-F_\chi)^2}{1-F_\varpi} \quad (6)$$

The triangular distribution derived by repeated sampling when *q* = 38/111 is shown in Figure **3**. It is a simpler and more defensible choice of prior than the one given in Stark [12].

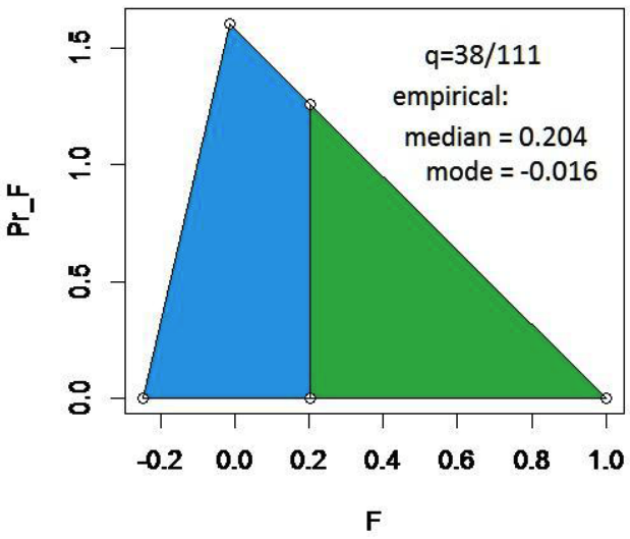

Figure 3: Constructed distribution of F for $q = 38/111$. (Pr_F stands for probability density of *F*).

We applied the same sampling process to several values of *q* and calculated the modes of *F* from equation (6). The results are displayed in Figure **4** which shows that modes close to zero, some even negative, apply over an interval of *F*.

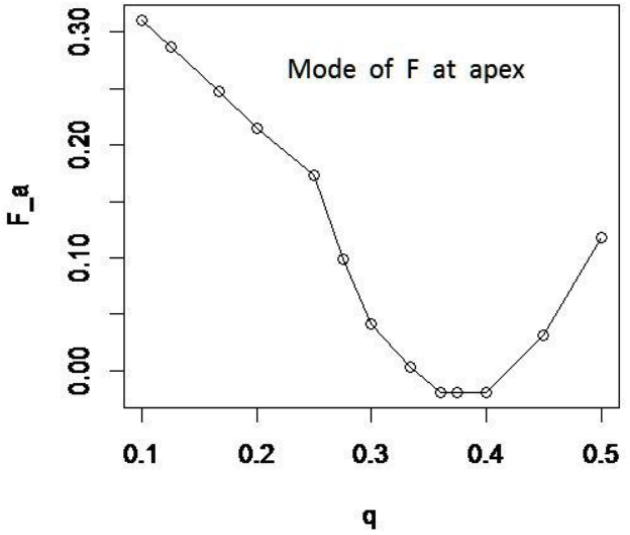

Figure 4: Modes of F for a selection of values of *q*. F_a stands for the calculated mode of the empirical distribution of *F* for assumed values of *q*.

THE BAYESIAN ESTIMATE OF *F*

The probability density function (pdf) of the distribution depicted in Figure **3** is simple to calculate. The length of the base is $1 - F_\omega$, where $F_\omega = (4q - 1 - 6q^2)/(6pq)$, in this case. Therefore the height of the function at the apex $\acute{\alpha} = 2/(1 - F_\omega)$, since the area between the pdf and zero is unity. The slope of the pdf for values of *F* between F_ω and F_η is $\acute{\epsilon} = \acute{\alpha}/(F_\eta - F_\omega)$; the value of the pdf in this interval is $\acute{\epsilon}.(F - F_\omega) = 2(F - F_\omega)/((1 - F_\omega)(F_\eta - F_\omega))$. The slope of the pdf for values of *F* between F_η and 1 is $\acute{\imath} = \acute{\alpha}/(F_\eta - 1)$; the value of the pdf in this interval is $\acute{\imath}.(F - 1) = 2(F - 1)/((1 - F_\omega)(F_\eta - 1))$. We denote the composite pdf from these two intervals as P(*F*), $F_\omega \le F \le 1$.

The pdf can be used to calculate a Bayesian estimate of *F*. If the value of the fixation index is *F*, the genotypic proportions in the population are

$$\{f_0 = q^2 + Fpq,\ f_1 = 2pq - 2Fpq,\ f_2 = p^2 + Fpq\}$$

Denoting the genotypic counts by $\{n_{UU}, n_{UT}, n_{TT}\}$, the (conditional) probability of observing these counts is

$$C(F) = \frac{n!}{n_{UU}! \times n_{UT}! \times n_{TT}!} \times f_0^{n_{UU}} \times f_1^{n_{UT}} \times f_2^{n_{TT}}$$

where *n* is the sample size.

If P(*F*).*dF* is the prior probability that *F* lies in an infinitesimal interval containing *F* , then, applying the formula of Bayes [14], the posterior probability that it is in that interval is

$$\text{P'}(F).dF = \frac{\text{P}(F).dF \times C(F)}{\int \text{P}(F) \times C(F).dF} \quad . \tag{7}$$

The posterior distribution of *F* from (7) for the above counts is displayed in Figure **5**.

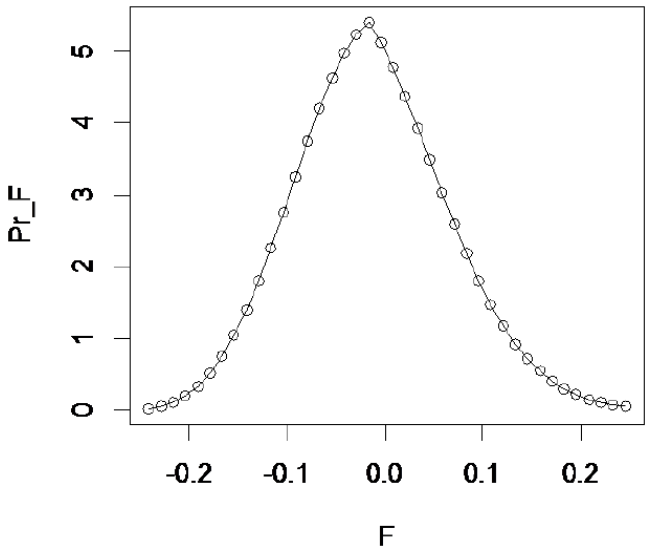

Figure 5: The posterior distribution of F computed from genotypic counts {12, 52, 47} from which $q = 38/111$ (Pr_F stands for probability density of *F*).

REFERENCES

[1] Edwards AWF. G.H. Hardy (1908) and Hardy-Weinberg equilibrium. Genetics 2008; 179: 1143-1150. http://dx.doi.org/10.1534/genetics.104.92940

[2] Mayo O. A century of Hardy-Weinberg equilibrium. Twin Res Hum Genet 2008; 11: 249-246. http://dx.doi.org/10.1375/twin.11.3.249

[3] Russell PJ. iGenetics: A Molecular Approach. 2nd ed. San Francisco, CA: Pearson Education, Inc., publishing as Benjamin Cummings 2006.

[4] Morton NE. Outline of Genetic Epidemiology. Basel: S. Karger 1982.

[5] Bittles AH, Black ML. Consanguinity, human evolution, and complex diseases. PNAS 2010; 107(suppl 1): 1779-1786. http://dx.doi.org/10.1073/pnas.0906079106

[6] Strømme P, Suren P, Kanavin ØJ, *et al*. Parental consanguinity is associated with a seven-fold increased risk of progressive encephalopathy: A cohort study from Oslo, Norway. Eur J Paediatr Neurol 2010; 14: 138-145. http://dx.doi.org/10.1016/j.ejpn.2009.03.007

[7] Risch NJ. Searching for genetic determinants in the new millennium. Nature 2000; 405: 847-856. http://dx.doi.org/10.1038/35015718

[8] Taft RJ, Vanderver A, Leventer RJ, *et al*. Mutations in DARS cause hypomyelination with brain stem and spinal cord involvement and leg spasticity. Am J Hum Genet 2013; 92: 774-780. http://dx.doi.org/10.1016/j.ajhg.2013.04.006

[9] Kaminski L, Damiani S, Damiani S. Cracking the Code. North Sydney NSW, Australia: Random House Australia Pty Ltd. 2015.

[10] Stark AE, Seneta E. Hardy-Weinberg equilibrium as foundational. Int J Stat Med Res 2014; 3:198-202. http://dx.doi.org/10.6000/1929-6029.2014.03.02.12

[11] Stark AE, Seneta E. A reality check on Hardy-Weinberg. Twin Res Hum Genet 2013; 16: 782-789. http://dx.doi.org/10.1017/thg.2013.40

[12] Stark AE. Estimation of divergence from Hardy-Weinberg form. Twin Res Hum Genet 2015; in press.

[13] Cavalli-Sforza LL, Bodmer WF. The Genetics of Human Populations. San Francisco: W. H. Freeman and Company; 1971.

[14] Bayes T. An essay towards solving a problem in the doctrine of chances. Phil Trans Roy Soc 1763; liii: 370-418.

[15] OMIM. http://www.omim.org/entry/615281?search=dars&highlight=dars

Received on 19-06-2015 Accepted on 17-07-2015 Published on 19-08-2015

http://dx.doi.org/10.6000/1929-6029.2015.04.03.5