Poznań Studies in the Philosophy of the Sciences and the Humanities
2000, *vol.* 71, *pp.* 136–151

Robert B. Gardner and Michael Wooten

AN APPLICATION OF BAYES' THEOREM TO POPULATION GENETICS

ABSTRACT. In this paper, we derive several conditional probabilities with direct applications in population genetics. The probability that an n-degree relative shares a trait with a given individual is calculated. An example of the application of these probabilities to some previously published DNA fingerprint data is given in which a χ^2 test is performed to determine the degree of relatedness between two individuals.

1. Introduction

The purpose of this paper is to present the derivation of certain conditional probabilities which have applications in population genetics. We make extensive use of Bayes' Theorem and thereby avoid the use of transition matrices (derivations using transition matrices of some of our results have been performed by Li and Sacks 1954). We discuss the application of the derived conditional probabilities to the problem of determining the degree of relatedness between individuals based on their phenotypes and present an analysis of such a case using previously presented DNA fingerprint data.

A detailed account of the use of DNA fingerprinting in the estimation of relatedness was described by Lynch (1988). He showed that the proportion of shared bands is a poor estimate of relatedness unless the frequencies of the bands are near zero (Lynch 1988, Figure 2). This is not surprising since a band with a high relative frequency would be present in significant numbers of unrelated individuals. To compensate for this problem, it is necessary to make a correction in the probability of shared bands for different band frequencies and degrees of relationship. Calculations relative to these corrections are presented here and the results are applicable to any traits which undergo Mendelian inheritance.

2. Results

If a population is in Hardy-Weinberg equilibrium for a given trait, then certain conditional probabilities concerning the presence or absence of the trait in related individuals can be calculated. For example, if an individual demonstrates a trait, the probability that his or her offspring, sibling, cousin, etc. also shows this trait can be calculated.

Consider a dominant trait with allele frequency p. For an individual X, denote the homozygous dominant state as X_1, the heterozygous state as X_2 and the homozygous recessive state as X_3, then the different states of X have the following probabilities: $P(X_1) = p^2$, $P(X_2) = 2p - 2p^2$, $P(X_3) = 1 - 2p + p^2$. However, with a dominant trait, it is unlikely that a heterozygous and a homozygous dominant individual can be distinguished. If the presence of the trait is denoted as $X-$, then $P(X-) = P(X_1) + P(X_2) = 2p - p^2$.

Now, we can do calculations to determine the probabilities of a dominant trait appearing in an offspring of an individual. Denote the known individual as M and the offspring as D. For this calculation, we will need to consider the other parent of D, say F. First, we calculate the probability of the genotypes of the parents, given the genotype of parent M (so when this is nonzero, it will depend only on F_i). Next, we calculate the probabilities for the different genotypes of D given M_i and F_j. Then the probability of each genotype of D given the genotype of M can be derived. With the values in Table 1, we can calculate the probability of each possible genotype of D given any genotype of M as follows:

$$\begin{aligned}
P(D_1|M_1) &= \sum_j P(F_j)P(D_1|M_1 \text{ and } F_j) = p, \\
P(D_1|M_2) &= \sum_j P(F_j)P(D_1|M_2 \text{ and } F_j) = \frac{p}{2}, \\
P(D_1|M_3) &= \sum_j P(F_j)P(D_1|M_3 \text{ and } F_j) = 0, \\
P(D_2|M_1) &= \sum_j P(F_j)P(D_2|M_1 \text{ and } F_j) = 1 - p, \\
P(D_2|M_2) &= \sum_j P(F_j)P(D_2|M_2 \text{ and } F_j) = \frac{1}{2}, \\
P(D_2|M_3) &= \sum_j P(F_j)P(D_2|M_3 \text{ and } F_j) = p,
\end{aligned}$$

$$\begin{aligned}
P(D_3|M_1) &= \sum_j P(F_j)P(D_3|M_1 \text{ and } F_j) = 0, \\
P(D_3|M_2) &= \sum_j P(F_j)P(D_3|M_2 \text{ and } F_j) = \frac{1-p}{2}, \text{ and} \\
P(D_3|M_3) &= \sum_j P(F_j)P(D_3|M_3 \text{ and } F_j) = 1-p.
\end{aligned}$$

M_i	F_j	$P(D_1\|M_i \text{ and } F_j)$	$P(D_2\|M_i \text{ and } F_j)$	$P(D_3\|M_i \text{ and } F_j)$
M_1	F_1	1	0	0
M_1	F_2	1/2	1/2	0
M_1	F_3	0	1	0
M_2	F_1	1/2	1/2	0
M_2	F_2	1/4	1/2	1/4
M_2	F_3	0	1/2	1/2
M_3	F_1	0	1	0
M_3	F_2	0	1/2	1/2
M_3	F_3	0	0	1

Table 1. Conditional Probability of Genotype of Offspring D given the Genotypes of the Parents M and F.

Because we are only concerned with the presence or absence of the trait, the following are obtained:

$$\begin{aligned}
P(D-|M-) &= P((D_1 \text{ or } D_2)|(M_1 \text{ or } M_2)) \\
&= \frac{P((D_1 \text{ or } D_2) \text{ and } (M_1 \text{ or } M_2))}{P(M_1 \text{ or } M_2)} \\
&= P((D_1 \text{ and } M_1) \text{ or } (D_1 \text{ and } M_2) \text{ or } (D_2 \text{ and } M_1) \text{ or} \\
&\quad (D_2 \text{ and } M_2))/P(M_1 \text{ or } M_2) \\
&= \{P(M_1)[P(D_1|M_1) + P(D_2|M_1)] + P(M_2)[P(D_1|X_2) \\
&\quad +P(D_2|M_2)]\}/(P(M_1) + P(M_2)) \\
&= \frac{1+p-p^2}{2-p}, \\
P(D-|M_3) &= P(D_1 \text{ or } D_2|M_3)) = P(D_1|M_3) + P(D_2|M_3) = p, \\
P(D_3|M-) &= \frac{P(M_1)P(D_3|M_1) + P(M_2)P(D_3|M_2)}{P(M_1) + P(M_2)} = \frac{1-2p+p^2}{2-p},
\end{aligned}$$

$$P(D_3|M_3) = 1 - p.$$

Now, consider the probability of the presence of a trait in a parent M given the presence of the trait in the offspring D. In this case, also, we have to consider the other parent, F. The calculations are similar to those above and the results are the same, establishing a parent/offspring symmetry. The calculations require the entries of Table 2.

M_i	F_j	$P(M_i \text{ and } F_j\|D_1)$	$P(M_i \text{ and } F_j\|D_2)$	$P(M_i \text{ and } F_j\|D_3)$
M_1	F_1	p^2	0	0
M_1	F_2	$p(1-p)$	$p^2/2$	0
M_1	F_3	0	$p(1-p)/2$	0
M_2	F_1	$p(1-p)$	$p^2/2$	0
M_2	F_2	$(1-p)^2$	$p(1-p)$	p^2
M_2	F_3	0	$(1-p)^2/2$	$p(1-p)$
M_3	F_1	0	$p(1-p)/2$	0
M_3	F_2	0	$(1-p)^2/2$	$p(1-p)$
M_3	F_3	0	0	$(1-p)^2$

Table 2. Conditional Probabilities of Parental Genotypes, M_i and F_j, given the Genotype of an Offspring, D.

For the calculation of the same types of probabilities for two individuals that are siblings, say S^1 and S^2, we use the values from Tables 1 and 2 to get:

$$\begin{aligned}
P(S_1^1|S_1^2) &= \sum_{i,j} P(M_i \text{ and } F_j|S_1^2)P(S_1^1|M_i \text{ and } F_j) = \frac{1+2p+p^2}{4}, \\
P(S_1^1|S_2^2) &= \sum_{i,j} P(M_i \text{ and } F_j|S_2^2)P(S_1^1|M_i \text{ and } F_j) = \frac{p+p^2}{4}, \\
P(S_1^1|S_3^2) &= \sum_{i,j} P(M_i \text{ and } F_j|S_3^2)P(S_1^1|M_i \text{ and } F_j) = \frac{p^2}{4}, \\
P(S_2^1|S_1^2) &= \sum_{i,j} P(M_i \text{ and } F_j|S_1^2)P(S_2^1|M_i \text{ and } F_j) = \frac{1-p^2}{2}, \\
P(S_2^1|S_2^2) &= \sum_{i,j} P(M_i \text{ and } F_j|S_2^2)P(S_2^1|M_i \text{ and } F_j) = \frac{1+p-p^2}{2},
\end{aligned}$$

$$\begin{aligned}
P(S^1_2|S^2_3) &= \sum_{i,j} P(M_i \text{ and } F_j|S^2_3)P(S^1_2|M_i \text{ and } F_j) = \frac{2p-p^2}{2}, \\
P(S^1_3|S^2_1) &= \sum_{i,j} P(M_i \text{ and } F_j|S^2_1)P(S^1_3|M_i \text{ and } F_j) = \frac{1-2p+p^2}{4}, \\
P(S^1_3|S^2_2) &= \sum_{i,j} P(M_i \text{ and } F_j|S^2_2)P(S^1_3|M_i \text{ and } F_j) = \frac{2-3p+p^2}{4}, \\
P(S^1_3|S^2_3) &= \sum_{i,j} P(M_i \text{ and } F_j|S^2_3)P(S^1_3|M_i \text{ and } F_j) = \frac{4-4p+p^2}{4}.
\end{aligned}$$

And so for siblings:

$$\begin{aligned}
P(S^1-|S^2-) &= \{P(S^2_1)[P(S^1_1|S^2_1)+P(S^1_2|S^2_1)]+P(S^2_2)[P(S^1_1|S^2_2)+ \\
&\quad P(S^1_2|S^2_2)]\}/(P(S^2_1)+P(S^2_2)) \\
&= \frac{4+5p-6p^2+p^3}{4(2-p)}, \\
P(S^1_3|S^2-) &= \frac{P(S^2_1)P(S^1_3|S^2_1)+P(S^2_2)P(S^1_3|S^2_2)}{P(S^2_1)+P(S^2_2)} \\
&= \frac{4-9p+6p^2-p^3}{4(2-p)}, \\
P(S^1-|S^2_3) &= P(S^1_1|S^2_3)+P(S^1_2|S^2_3) = \frac{4p-p^2}{4}, \\
P(S^1_3|S^2_3) &= \frac{4-4p+p^2}{4}.
\end{aligned}$$

In fact, $P(S^1-|S^2)$ has already appeared in the literature (Jeffreys *et al.* 1985).

If two (non-inbreed) individuals are related, then they will be related by a series of offspring to parent steps followed by a series of parent to offspring steps. There may also be a sibling step at the apex of this path. If the number of such steps is n, then these are called *n-degree relatives.* We show that the calculations of the probabilities allow a commuting of the sibling step with the parent/offspring steps by showing that the numbers are the same for a niece/nephew as they are for an aunt/uncle. So a combination of a parent/offspring step and a sibling step is needed. This corresponds to comparing, for example, a niece, N, and an aunt, A.

Again, comparisons between these two individuals are symmetric and one gets

$$
\begin{aligned}
P(A-|N-) &= P(N-|A-) = \frac{1+5p-5p^2+p^3}{2(2-p)}, \\
P(A_3|N-) &= P(N_3|A-) = \frac{3-7p+5p^2-p^3}{2(2-p)}, \\
P(A-|N_3) &= P(N-|A_3) = \frac{3p-p^2}{2}, \\
P(A_3|N_3) &= P(N_3|A_3) = \frac{2-3p+p^2}{2}.
\end{aligned}
$$

Calculation of a grandparent/grandchild pair yields the same results.

Surprisingly, even though the probabilities for parents and offspring were different from those of siblings, the grandparent/grandchild probabilities are the same as the aunt/uncle and niece/nephew probabilities. So, if we know the relationship between two individuals (provided they are not siblings), then we can calculate the relevant probabilities by simply using a series of parent/offspring steps, where the sibling step, if present, counts the same as a parent/offspring step.

Now, consider an individual, G^n, that is an n-degree relative of X (e.g. if $n=1$ then G^1 is a parent, sibling or offspring; if $n=2$ then G^2 is a grandparent, a grandchild, an aunt, a niece, etc.). Then the following probabilities are derived, which have been established for $n=1$ (with the relationship "sibling" being a special case) and can be established in general by induction.

$$
\begin{aligned}
P(G_1^n|X_1) &= \frac{p+(2^{n-1}-1)p^2}{2^{n-1}}, \\
P(G_2^n|X_1) &= \frac{1+(2^n-3)p-(2^n-2)p^2}{2^{n-1}}, \\
P(G_3^n|X_1) &= \frac{(2^{n-1}-1)-(2^n-2)p+(2^{n-1}-1)p^2}{2^{n-1}}, \\
P(G_1^n|X_2) &= \frac{p+(2^n-2)p^2}{2^n}, \\
P(G_2^n|X_2) &= \frac{1+(2^{n+1}-4)p-(2^{n+1}-4)p^2}{2^n}, \\
P(G_3^n|X_2) &= \frac{(2^n-1)-(2^{n+1}-3)p+(2^n-2)p^2}{2^n},
\end{aligned}
$$

$$\begin{aligned}
P(G_1^n|X_3) &= \frac{2^{n-1}-1}{2^{n-1}}p^2,\\
P(G_2^n|X_3) &= \frac{(2^n-1)p-(2^n-2)p^2}{2^{n-1}},\\
P(G_3^n|X_3) &= \frac{2^{n-1}-(2^n-1)p+(2^{n-1}-1)p^2}{2^{n-1}}.
\end{aligned}$$

From which:

$$\begin{aligned}
P(G^n - |X-) &= \{P(X_1)[P(G_1^n|X_1)+P(G_2^n|X_1)]+P(X_2)[P(G_1^n|X_2)\\
&\quad +P(G_2^n|X_2)]\}/[P(X_1)+P(X_2)]\\
&= \frac{1+(2^{n+1}-3)p-(2^{n+1}-3)p^2+(2^{n-1}-1)p^3}{2^{n-1}(2-p)}, \qquad (1)\\
P(G^n - |X_3) &= P(G_1^n|X_3)+P(G_2^n|X_3)\\
&= \frac{(2^n-1)p-(2^{n-1}-1)p^2}{2^{n-1}}, \qquad (2)\\
P(G_3^n|X-) &= \frac{P(X_1)P(G_3^n|X_1)+P(X_2)p(G_3^n|X_2)}{P(X_1)+P(X_2)}\\
&= \left\{(2^n-1)-(5\times 2^{n-1}-3)p+(2^{n+1}-3)p^2-\right.\\
&\quad \left.(2^{n-1}-1)p^3\right\}/(2^{n-1}(2-p)),\\
P(G_3^n|X_3) &= \frac{2^{n-1}-(2^n-1)p+(2^{n-1}-1)p^2}{2^{n-1}}.
\end{aligned}$$

Additional observations can be made about these formulae. Notice that if a limit as n approaches infinity is taken, in each of the above formulae, the probability of the presence or absence of the trait is:

$$\begin{aligned}
P(G^\infty - |X-) &= 2p-p^2,\\
P(G^\infty - |X_3) &= 2p-p^2,\\
P(G_3^\infty|X-) &= (1-p)^2,\\
P(G_3^\infty|X_3) &= (1-p)^2.
\end{aligned}$$

This is expected since one assumes mating with random unrelated individuals. Notice that if we take $p=0$, then $P(G^n - |X-) = \left(\frac{1}{2}\right)^n$ which is simply the coefficient of relationship (Lynch 1988; Wright 1922).

S_i^2	$P(S_j^2\|S_1^1)$	$P(S_j^2\|S_2^1)$	$P(S_j^2\|S_3^1)$
S_1^2	$(1+2p+p^2)/4$	$(p+p^2)/4$	$p^2/4$
S_2^2	$(1-p^2)/2$	$(1+p-p^2)/2$	$(2p-p^2)/2$
S_3^2	$(1-2p+p^2)/4$	$(2-3p+p^2)/4$	$(4-4p+p^2)/4$

Table 3. Conditional Probability of the Genotype of an Individual, S_j^2, given the Genotype of a sibling, S^1.

Now consider certain special cases where inbreeding is present. If two siblings S^1 and S^2 have an offspring D then by referring to Tables 1 and 3 (replacing M and F by S^1 and S^2 in Table 1) we get:

$$\begin{aligned}
P(D_1|S_1^1) &= \sum_j P(S_j^2|S_1^1)P(D_1|S_1^1 \text{ and } S_j^2) = \frac{1+p}{2}, \\
P(D_1|S_2^1) &= \sum_j P(S_j^2|S_2^1)P(D_1|S_2^1 \text{ and } S_j^2) = \frac{1+2p}{8}, \\
P(D_1|S_3^1) &= \sum_j P(S_j^2|S_3^1)P(D_1|S_3^1 \text{ and } S_j^2) = 0, \\
P(D_2|S_1^1) &= \sum_j P(S_j^2|S_1^1)P(D_2|S_1^1 \text{ and } S_j^2) = \frac{1-p}{2}, \\
P(D_2|S_2^1) &= \sum_j P(S_j^2|S_2^1)P(D_2|S_2^1 \text{ and } S_j^2) = \frac{1}{2}, \\
P(D_2|S_3^1) &= \sum_j P(S_j^2|S_3^1)P(D_2|S_3^1 \text{ and } S_j^2) = \frac{p}{2}, \\
P(D_3|S_1^1) &= \sum_j P(S_j^2|S_1^1)P(D_3|S_1^1 \text{ and } S_j^2) = 0, \\
P(D_3|S_2^1) &= \sum_j P(S_j^2|S_2^1)P(D_3|S_2^1 \text{ and } S_j^2) = \frac{3-2p}{8}, \text{ and} \\
P(D_3|S_3^1) &= \sum_j P(S_j^2|S_3^1)P(D_3|S_3^1 \text{ and } S_j^2) = \frac{2-p}{2}.
\end{aligned}$$

From which:

$$\begin{aligned}
P(D-|S^1-) = {} & \{P(S_1^1)[P(D_1|S_1^1) + P(D_2|S_1^1)] + P(S_2^1)[P(D_1|S_2^1) + \\
& \{P(D_2|S_2^1)]\} / (P(S_1^1) + P(S_2^1))
\end{aligned}$$

$$\begin{aligned}
&= \frac{5+p-2p^2}{4(2-p)}, \\
P(D-|S_3^1) &= P(D_1|S_3^1)+P(D_2|S_3^1)=\frac{p}{2}, \\
P(D_3|S^1-) &= \frac{P(S_1^1)P(D_3|S_1^1)+P(S_2^1)P(D_3|S_2^1)}{P(S_1^1)+P(S_2^1)}=\frac{3-5p+2p^2}{4(2-p)}, \\
P(D_3|S_3^1) &= \frac{2-p}{2}.
\end{aligned}$$

Now consider a similar case, where an individual X mates with an n-degree relative G^n producing an offspring D. See Tables 1 and 4 (replacing M and F by X and G^n in Table 1). These follow:

$$\begin{aligned}
P(D_1|X_1) &= \sum_j P(G_j^n|X_1)P(D_1|X_1 \text{ and } G_j^n)=\frac{1+(2^n-1)p}{2^n}, \\
P(D_1|X_2) &= \sum_j P(G_j^n|X_2)P(D_1|X_2 \text{ and } G_j^n)=\frac{1+(2^{n+1}-2)p}{2^{n+2}}, \\
P(D_1|X_3) &= \sum_j P(G_j^n|X_3)P(D_1|X_3 \text{ and } G_j^n)=0, \\
P(D_2|X_1) &= \sum_j P(G_j^n|X_1)P(D_2|X_1 \text{ and } G_j^n)=\frac{(2^n-1)-(2^n-1)p}{2^n}, \\
P(D_2|X_2) &= \sum_j P(G_j^n|X_2)P(D_2|X_2 \text{ and } G_j^n)=\frac{1}{2}, \\
P(D_2|X_3) &= \sum_j P(G_j^n|X_3)P(D_2|X_3 \text{ and } G_j^n)=\frac{(2^n-1)p}{2^n}, \\
P(D_3|X_1) &= \sum_j P(G_j^n|X_1)P(D_3|X_1 \text{ and } G_j^n)=0, \\
P(D_3|X_2) &= \sum_j P(G_j^n|X_2)P(D_3|X_2 \text{ and } G_j^n) \\
&= \frac{(2^{n+1}-1)-(2^{n+1}-2)p}{2^{n+2}}, \text{ and} \\
P(D_3|X_3) &= \sum_j P(G_j^n|X_3)P(D_3|X_3 \text{ and } G_j^n)=\frac{2^n-(2^n-1)p}{2^n}.
\end{aligned}$$

G_j^n	$P(G_j^n \mid X_1)$	$P(G_j^n \mid X_2)$	$P(G_j^n \mid X_3)$
G_1^n	$\frac{p+(2^{n-1}-1)p^2}{2^{n-1}}$	$\frac{p+(2^n-2)p^2}{2^n}$	$\frac{(2^{n-1}-1)p^2}{2^{n-1}}$
G_2^n	$\frac{1+(2^n-3)p-(2^n-2)p^2}{2^{n-1}}$	$\frac{1+(2^{n+1}-4)p-(2^{n+1}-4)p^2}{2^n}$	$\frac{(2^n-1)p-(2^n-2)p^2}{2^{n-1}}$
G_3^n	$\frac{(2^{n-1}-1)-(2^n-2)p+(2^{n-1}-1)p^2}{2^{n-1}}$	$\frac{(2^n-1)-(2^{n+1}-3)p+(2^n-2)p^2}{2^n}$	$\frac{2^{n-1}-(2^n-1)p+(2^{n-1}-1)p^2}{2^{n-1}}$

Table 4. Conditional Probability of Genotype of n-degree Relatives, X and G^n, given the Genotype of X.

And hence:

$$\begin{aligned}
P(D-|X-) &= \{P(X_1)[P(D_1|X_1)+P(D_2|X_1)]+P(X_2)[P(D_1|X_2)+\\
&\quad P(D_2|X_2)]\}\,/(P(X_1)+P(X_2))\\
&= \frac{(2^{n+1}+1)+(2^{n+1}-3)p-(2^{n+1}-2)p^2}{2^{n+1}(2-p)},\\
P(D-|X_3) &= P(D_1|X_3)+P(D_2|X_3)=\frac{(2^n-1)p}{2^n},\\
P(D_3|X-) &= \frac{P(X_1)P(D_3|X_1)+P(X_2)P(D_3|X_2)}{P(X_1)+P(X_2)}\\
&= \frac{(2^{n+1}-1)-(2^{n+2}-3)p+(2^{n+1}-2)p^2}{2^{n+1}(2-p)},\\
P(D_3|X_3) &= \frac{2^n-(2^n-1)p}{2^n}.
\end{aligned}$$

Notice that for $n=1$, this also reduces to the case of mated siblings. Also, for any particular type of inbreeding, say continued brother/sister mating, it is possible to calculate these type probabilities. The calculations (although tedious) would employ the methods used here, that is, the construction of tables of the nine possible genotypes of the parents with the corresponding conditional probabilities.

3. An Application

Jeffreys *et al.* (1985) presented the following example of the use of DNA fingerprints in determining relationships: "The case concerned a Ghanaian boy born in the United Kingdom who emigrated to Ghana to join his father and subsequently returned alone to the United Kingdom to be reunited with his mother, brother and two sisters. However, there was evidence to suggest that a substitution might have occurred, either for an unrelated boy, or a son of a sister of the mother... As a result, the returning boy was not granted residence in the United Kingdom." Conventional genetic markers indicated that the woman and boy were related (with 99% probability), but could not determine whether the woman was the boy's mother or aunt. DNA fingerprints were produced from blood DNA samples taken from the boy, the mother, her three other children, and an unrelated individual. The father was unavailable. Based on the probability of unrelated individuals sharing a band, the allele frequency for a band was calculated to be $p = 0.14$ (all bands were assumed to have the same frequency). The mother and boy were found to share 25 maternal specific bands. Based on this, it was calculated that the probability of these two being unrelated was $(0.26)^{25} = 2 \times 10^{-15}$ (an allele with frequency 0.14 will appear in an individual with probability $2(0.14) - (0.14)^2 = 0.26$). The corresponding probability of the mother actually being the aunt of the boy was said to be 6×10^{-6}. This latter calculation, however, was found to be erroneous and declared irrelevant by Hill (1986) who approached this same problem from a maximum likelihood viewpoint, and reached the same conclusion. The boy was granted residence in the United Kingdom. The method used by Jeffreys *et al.* (1985) is very restricted and not suitable for general use in testing specified relationships.

We can re-analyze this case using the established probabilities. It was remarked that there are two relationships to be tested, that of aunt/nephew ($n = 2$) and mother/son ($n = 1$). Using equations (1) and (2) with $p = 0.14$ for all bands, as in the paper, we can generate expected numbers of shared bands and do a χ^2 test on these two relationships and on the possibility of no relation between the mother and boy. We get the results of Table 5 and find that, of these three possibilities, we fail to reject only the option of a mother/son relationship.

Null Hypothesis	χ^2	Conclusion
M and X are unrelated	49.4	reject
M and X are aunt/nephew	12.5	reject
M and X are mother/son	.75	fail to reject

Table 5. χ^2 Values for the Three Plausible Null Hypotheses for the Information of Jeffreys *et al.* (1985). M is the Alleged Mother and X the Alleged Son.

4. Conclusion

In conclusion, we have presented conditional probabilities on individuals sharing traits which undergo Mendelian inheritance given that the degree of relatedness of the individuals is known. An application of these conditional probabilities to the estimation of degree of relatedness using a χ^2 test has been given. An alternative approach to estimating relatedness would be to calculate likelihoods of certain degrees of relationships, given data sets. It is our belief that these methods would be valid and unambiguous in establishing close degrees of relatedness ($n = 1$ or $n = 2$) and when average allele frequencies are low (say 15% or less). However, these procedures may become inaccurate when applied to more distant relationships or when average allele frequencies are high (Lewin 1989; Lynch 1988). If the phenotypic data of several individuals chosen from a single population are available, this problem could possibly be overcome by piecing together a pedigree using only the close relationships. There can be a high degree of confidence in these close relationships, and therefore a fairly high degree of confidence in the derived pedigree.

The accurate derivation of pedigrees would be especially important in selective breeding programs in small populations or endangered species. Also, inbreeding could be controlled in such a situation if relatedness between individuals were known.

Acknowledgements. Support for this research was provided by the Alabama Agricultural Experiment Station (AAES no. 15-913078). Comments by S. Dobson, C. Guyer and G. Hepp on an earlier draft of this manuscript are gratefully acknowledged.

REFERENCES

Hill, W. (1986). DNA Fingerprint Analysis in Immigration Test-Case. *Nature* **322**, 290–291.

Jeffreys, A., F. Brookfield, and R. Semeonoff (1985). Positive Identification of an Immigration Test-Case using Human DNA Fingerprints. *Nature* **317**, 818–819.

Lewin, R. (1989). Limits to DNA Fingerprinting. *Science* **243**, 1549–1551.

Li, C. and L. Sacks (1954). The Derivation of Joint distribution and Correlation between Relatives by the use of Stochastic Matrices. *Biometrics* **10**, 347–360.

Lynch, M. (1988). Estimation of Relatedness by DNA Fingerprinting. *Molecular Biology and Evolution* **5**(5), 584–599.

Wright, S. (1922). Coefficients of Inbreeding and Relationship. *American Naturalist* **56**, 330–338.

Robert Gardner
Department of Mathematics
Department of Physics and Astro.
East Tennessee State University
Johnson City, Tennessee 37614
e-mail: gardnerr@etsu.edu

Michael Wooten
Department of Zoology
and Wildlife Science
AL Ag. Experiment Station
Auburn University
Auburn, Alabama 36830-5414
e-mail:mwooten@ag.auburn.edu

Commentary by Lynne Seymour

Bayes' Theorem is an elementary probabilistic mechanism with such broad implications that an entire branch of statistical inference has arisen from it. The authors have used this mechanism to simplify some basic probability calculations in population genetics. Why wasn't this done in the first place? Almost surely because the scientist who first derived those probabilities was unaware of the tools which could simplify his work.

This is a common occurrence at the interface of mathematics with the sciences. In particular, it happens in statistics, albeit to a slightly lesser extent since scientists tend to use statistics in their everyday research. Even so, few scientists truly understand the statistical procedures they use — procedures on which they base the conclusions of their livelihoods! The χ^2 test that the authors used will be employed here to highlight the logic inherent in a test of hypotheses, and to underscore how probability can be invaluable in interpreting the hypothesis test.

First, what is a statistical test of hypotheses? In general, it is a logical procedure which is derived under stated conditions. Null and alternative hypotheses are clearly stated. A random variable — a function of data called the *test statistic* — is developed and its distribution is derived and/or approximated assuming that the null hypothesis is true. A value for the test statistic is calculated using the actual data, and the likelihood of seeing this value from the test statistic's distribution under the null hypothesis is evaluated. If this value is judged to be unlikely under the null hypothesis, then the null hypothesis is rejected at an explicitly stated *significance level.* This significance level, usually taken to be 5%, is interpreted as the largest chance one is willing to tolerate of wrongly rejecting the null hypothesis. If any of the assumptions under which the test is derived are violated, the results of the test cannot be trusted.

Many common statistical tests have "shorthand" names by which they are invoked. The authors, in applying their newly-derived probabilities to an already-published example, state that they "do a χ^2 test" to draw a conclusion. Which χ^2 test would that be? A statistician knows that there are a great many of them. An elementary text which assumes only a modest knowledge of high school algebra (Anderson *et al.* 1994) lists three χ^2 tests in the index. One is for the variance of a single population (which the reader can rule out fairly quickly); another is for independence (which, upon close inspection of the author's Table 5, may be ruled out); and yet another is for goodness of fit (which is the one used).

Consider the first of the χ^2 *goodness-of-fit* tests which the authors used. The null hypothesis is H_0 : *M and X are unrelated*, and the alternative is H_1 : *M and X are related.* The data indicates whether a matched pair of bands is the same or different. The test statistic requires that there be k categories into which each observation falls exactly once. Hence the $k = 2$ categories may be labeled *same bands* and *different bands.* The test statistic is given by $\sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$, where O_i is the observed number of bands in the ith category, computed assuming H_0 is true. Under H_0, this test statistic has a χ^2 distribution with $k - 1 = 1$ degree(s) of freedom, denoted $\chi^2(1)$. (Note that Table 5 does not specify the degrees of freedom for the χ^2 distribution.) From the $\chi^2(1)$ table, it becomes apparent that the conclusions given in Table 5 are not only at the 5% significance level, but certainly at any reasonable significance level.

Most statistical hypothesis tests today cite a *p-value* rather than a significance level (in fact, statistical analysis packages such as SAS report only p-values when performing statistical tests of hypotheses). The p-value is the tail probability associated with the test statistic: *i.e.*, it is the conditional probability of obtaining a test statistic more in line with the alternative hypothesis than the one observed, given that H_0 is true. The traditional significance level is the largest such probability one is willing to tolerate when rejecting H_0.

The p-values associated with the $\chi^2(1)$ statistics given in Table 5 are summarized below (they were calculated using the function CHIDIST in Excel). These p-values add an exclamation point to the example which cannot be communicated by the raw χ^2 statistics. Roughly, based on the evidence presented, they say: If one rejects that M and X are unrelated, the chance of being wrong is negligible; if one rejects that M and X are aunt/nephew, there is a .04% chance of being wrong; and if one rejects that M and X are mother/son, there is a 38.64% chance of being wrong.

Null Hypothesis	$\chi^2(1)$	p-value
M and X are unrelated	49.40	2.09×10^{-12}
M and X are aunt/nephew	12.50	0.000407
M and X are mother/son	0.75	0.386476

REFERENCES

Anderson, D., D. Sweeney, and T. Williams (1994). *Introduction to Statistics: Concepts and Applications* 3rd ed. St. Paul: West Publishing.

Lynne Seymour
Department of Statistics
University of Georgia
Athens, GA 30602-1952
e-mail: seymour@stat.uga.edu

Gardner and Wooten's Reply to Seymour

We wish to thank Dr. Seymour for her commentary. One of us (Gardner) is a pure mathematician, while the other (Wooten) is a population geneticist. Our grasp of the subtleties of statistics is, no doubt, not optimal. We do enthusiastically agree, however, that probability is "invaluable" in any statistical process.

We also agree that the assumptions of a statistical test must not be violated (of course). We should probably emphasize that our χ^2 test of Section 3 assumes that the relevant bands are inherited independently (that is, there is no linkage between the bands). We think this a reasonable assumption and note that the two previous analyzes of this case (by Jeffreys *et al.* and Hill) have also (implicitly) assumed no linkage between the bands.

The use of p-values, as mentioned by Dr. Seymour, is certainly preferable to our "significance level" approach (we chose our approach due to our limited access to the appropriate software). We thank her for the "exclamation point"!

In conclusion, we mention that an additional problem with our particular approach to the application of the derived probabilities lies in the area of variance. If there is a large amount of variance within a population, then the recognition of distant relatives will be difficult due to the "noise" produced by the variance. This is addressed in the Lynch paper referenced in our article.