ON A MATCHING PROBLEM ARISING IN GENETICS

By Howard Levene

Columbia University

1. Summary. A statistic useful for detecting deviations from the Hardy-Weinberg equilibrium in population genetics is discussed. Both exact and asymptotic distributions are given and a special case where there is misclassification is discussed. The distribution obtained also arises from a certain card matching problem.

2. Introduction. A system of multiple alleles behaves as follows under Mendelian inheritance: There are r distinct forms or alleles, $a_1, \cdots, a_r$, of a given gene. A given individual contains two genes and can be represented as a_i/a_j. If $i = j$ the individual is called a homozygote; if $i \neq j$ it is called a heterozygote. The representation a_i/a_j is called the genotype. In reproduction each gamete produced by an a_i/a_j individual contains one gene which has a probability $1/2$ of being a_i and $1/2$ of being a_j. In fertilization a paternal and a maternal gamete fuse to form a new individual which contains two genes, giving the well-known Mendelian ratios. We now consider a large *random breeding* population of N individuals. This will contain $2N$ genes, of which the proportion q_i will be of type $a_i (i = 1, \cdots, r; \Sigma q_i = 1)$. The probability that a random individual from the next generation will be a_i/a_j is $q_i^2 (i = j)$ or $2q_iq_j (i \neq j)$, which are known as the Hardy-Weinberg equilibrium probabilities. The statistical problem arose in testing (by means of a sample of n individuals) the hypothesis that this Hardy-Weinberg ratio holds against the alternative hypothesis that disturbing forces decrease the number of homozygotes. The actual data has been discussed elsewhere [1].

3. The sample distribution of number of homozygotes. We shall assume throughout this paper that N is so large that random fluctuations in the population proportions from generation to generation can be ignored. Let $x_{ij} (i \leq j = 1, \cdots, r)$ be the number of a_i/a_j individuals in the sample, and let $y_i = x_{ii} + \Sigma_{j=1}^{r} x_{ij}$ be the number of a_i genes in the sample. We have $\Sigma\Sigma x_{ij} = n$ and $\Sigma y_i = 2n$. Let $h = \Sigma x_{ii}$ be the number of homozygotes, and $z = n - h$ be the number of heterozygotes in the sample. The probability of the observed sample is

$$P = \frac{n!}{\prod_{j \leq i} x_{ij}!} \prod_{i=1}^{r} (q_i^{2x_{ii}}) \prod_{i<j} (2q_iq_j)^{x_{ij}}$$

$$= \frac{n!2^z}{\prod_{j \leq i} x_{ij}!} \prod_{i=1}^{r} q_i^{y_i}. \tag{1}$$

Since the q_i are unknown we use the conditional probability when $y_1, \cdots, y_r$ are held constant. Whenever we use the word "conditional" hereafter, this condition will be understood. The conditional probability is

$$K' \frac{n!2^z}{\prod_{i \le j} x_{ij}!}, \quad \text{where}$$

$$\frac{1}{K'} = \sum' \frac{n!2^z}{\prod_{i \le j} x_{ij}!}; \tag{2}$$

where the summation Σ' is over all non-negative integral values of the x_{ij} subject to the condition

$$x_{ii} + \Sigma_j x_{ij} = y_i \qquad (i = 1, \cdots, r).$$

Consider

$$\left(\sum_1^r t_i\right)^{2n} = (\sum t_i^2 + 2\sum_{i<j} t_i t_j)^n \tag{3}$$

$$= \sum^* \frac{n!2^z}{\prod_{i \le j} x_{ij}!} \prod_{i=1} t_i^{x_{ii}+\sum_j x_{ij}}, \tag{4}$$

where the summation Σ^* is over all non-negative values of the x_{ij} subject to the condition $\Sigma_{i \le j} x_{ij} = n$. Evidently $1/K'$ is the coefficient of $\Pi t_i^{y_i}$ in (4); but this must equal the coefficient of this term in the left member of (3); and thus $1/K' = (2n)!/\Pi y_i!$. Hence the conditional probability of the observed sample is

$$P = \frac{n! \prod y_i!}{(2n)!} \cdot \frac{2^z}{\prod_{i \le j} x_{ij}!}. \tag{5}$$

For any function $u(x_{11}, \cdots, x_{1r}, \cdots, x_{rr})$ we will now let $E(u)$ and $\sigma^2(u)$ denote the *conditional* mean and variance of u for fixed y_i, and will refer to them simply as the mean and variance. We first obtain the sth factorial moment of x_{ii}, that is $E(x_{ii}^{(s)})$, where $x^{(s)} = x(x-1)\cdots(x-s+1)$. Consider

$$\sum' \frac{2^z n!}{\prod_{j \le k} x_{jk}!} x_{ii}^{(s)} = n^{(s)} \sum' \frac{2^z (n-s)!}{\prod_{j \le k} x'_{jk}!}, \tag{6}$$

where $x'_{jk} = x_{jk}$ except that $x'_{ii} = x_{ii} - s$, and Σ' has the same meaning as in (2). The right member of (6) is evaluated exactly as before, giving

$$E(x_{ii}^{(s)}) = \frac{n^{(s)} (y_i)^{(2s)}}{(2n)^{(2s)}}. \tag{7}$$

From this expression we obtain

$$E(x_{ii}) = \frac{y_i(y_i - 1)}{4n - 2} = nf_i^2 + O(1), \tag{8}$$

and

$$\sigma^2(x_{ii}) = \frac{n^{(2)}y_i^{(4)}}{(2n)^{(4)}} + \frac{ny_i^{(2)}}{(2n)^{(2)}} - \left[\frac{ny_i^{(2)}}{(2n)^{(2)}}\right]^2 = nf_i^2(1 - f_i)^2 + O(1), \tag{9}$$

where $f_i = y_i/2n$ is the sample estimate of q_i. Similarly

$$E(x_{ii}^{(s)}x_{jj}^{(t)}) = \frac{n^{(s+t)}y_i^{(2s)}y_j^{(2t)}}{(2n)^{(2s+2t)}} \tag{10}$$

giving

$$\sigma(x_{ii}, x_{jj}) = \frac{n^{(2)}y_i^{(2)}y_j^{(2)}}{(2n)^{(4)}} - \frac{y_i^{(2)}y_j^{(2)}}{4(2n-1)^2} \tag{11}$$

Other moments can be similarly evaluated, in particular $E(x_{ij}) = y_iy_j/(2n - 1)$.

4. Asymptotic distribution of number of homozygotes. From (8), (9), and (11) we may easily obtain

$$E(h) = \Sigma E(x_{ii}) = (C - 2n)/(4n - 2), \tag{12}$$

$$\sigma^2(h) = \Sigma\sigma^2(x_{ii}) + 2\sum\sum_{i<j}\sigma(x_{ii}, x_{jj}) \tag{13}$$

$$= \frac{1}{4n^2}\left\{C(n+2) + C^2\left(\frac{2n+5}{8n^2}\right) - D\left(\frac{n+2}{n}\right)\right\} - \tfrac{1}{2} + O\left(\frac{1}{n}\right), \tag{14}$$

where $C = \Sigma y_i^2$ and $D = \Sigma y_i^3$. The formula (14) is a close approximation to (13) and is easily computed. From (5) by means similar to those classically used to prove asymptotic normality of the binomial distribution we can prove asymptotic normality of the conditional distribution of h; more precisely, if $n \to \infty$ and $y_i/n \to$ constant $(i = 1, \cdots, r)$, then

$$\text{Prob}\left\{\frac{h - E(h)}{\sigma(h)} \leq t\right\} \to \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{t} e^{-x^2/2}\,dx. \tag{15}$$

5. Effect of misclassification. There is a further complication in the particular case reported in [1]. All individuals of genotype a_i/a_i are correctly classified, but an individual of genotype a_i/a_j $(i \neq j)$ has a known probability $p/2$ of being classified a_i/a_i and an equal probability of being classified a_j/a_j. As a result, the observed proportion of homozygotes is a biased estimate of the proportion in the population. Let h, x_{ij}, y_i denote the true sample values, and let h', x'_{ij}, y'_i denote the recorded sample values. Then $h^* = h' - e$, where $e = (n - h')\cdot p/(1 - p)$, will give an unbiased estimate, *i.e.* $E(h^*) = E(h)$. In order to use h^* we must have its (conditional) variance. Since $h^* = np/(1 - p) + h'/(1 - p)$,

$$\sigma_{h^*}^2 = [1/(1 - p)]^2\sigma_{h'}^2.$$

Let $h - h' = \epsilon$, then for large fixed $(n - h)$, ϵ is approximately normally distributed with mean $(n - h)p$ and variance

$$(n - h)p(1 - p) = [n - E(h)]p(1 - p)[1 + O_p(1/\sqrt{n})].$$

Neglecting the remainder term in this variance, ϵ and h have a joint normal distribution with parameters that are easily calculated. We thus have

$$\sigma_{h'}^2 = \sigma_h^2 + \sigma_\epsilon^2 + 2\sigma(h, \epsilon), \quad \text{or} \quad \sigma_{h'}^2 = [n - E(h)]p(1 - p) + (1 - p)^2\sigma_h^2,$$

giving

$$\sigma_{h*}^2 = \sigma_h^2 + [n - E(h)]p/(1 - p). \tag{16}$$

In [1] σ_{h*}^2 was given as $\sigma_h^2 + e$ for the sake of simplicity. This would tend to be smaller than (16), but only negligibly so. Strictly speaking the calculation of $E(h)$ and σ_h^2 from (12) and (14) requires a knowledge of the true y_i, but the observed y_i' are unbiased estimates of the y_i and their use should cause no serious trouble.

6. Combinatorial statement of the problem. This problem can also be expressed as one of card matching as follows: A deck contains $2n$ cards of r different suits; with y_i cards of the ith suit ($i = 1, \cdots, r$). We draw n pairs of cards at random without replacement, exhausting the deck. What is the distribution of h, the number of twins (pairs in which both members are of the same suit). If $z = n - h$, the probability of exactly h twins is given by (5), and in the limit h is normally distributed with mean given by (12) and variance given by (14). The card matching problem does not involve the notion of conditional probability. By introducing variables u_α equal to one if the αth pair is a twin and zero otherwise, the moments of h can also be obtained without using generating functions.

REFERENCE

[1] Theodosius Dobzhansky and Howard Levene, "Genetics of natural populations. XVII. Proof of operation of natural selection in wild populations of *Drosophila pseudoobscura*," *Genetics*, Vol. 33 (1948), pp. 537–547.