浙 江 林 学 院 学 报 2009，**26**(1)：122－126
Journal of Zhejiang Forestry College

哈迪-温伯格平衡检验及其基因频率的估计

管 宇，张香云，徐群芳，吴志松，顾光同，聂 文

（浙江林学院 理学院，浙江 临安 311300）

摘要：哈迪-温伯格平衡定律（Hardy-Weinberg equilibrium，HWE)是群体遗传学中最重要的原理，是群体有性繁殖上下代之间基因频率与基因型频率是否保持平衡的检验尺度。在HWE检验时通常需要对各基因频率作估计。研究了存在显隐性时基因频率的极大似然估计值的牛顿迭代算法，牛顿迭代算法的收敛速度快于EM算法；讨论了最小Pearson χ^2统计量估计与极大似然法估计的近似关系。推荐极大似然估计牛顿迭代法和最小Pearson χ^2估计法。参10

关键词：哈迪-温伯格平衡；HWE检验；基因频率；极大似然估计；最小Pearson χ^2估计
中图分类号：Q348；O29　　**文献标志码**：A　　**文章编号**：1000-5692(2009)01-0122-05

Hardy-Weinberg equilibrium testing and estimating of allele frequency

GUAN Yu，ZHANG Xiang-yun，XU Qun-fang，WU Zhi-song，GU Guang-tong，NIE Wen

（School of Sciences，Zhejiang Forestry College，Lin'an 311300，Zhejiang，China）

Abstract：As the most important principle in population genetics，Hardy-Weinberg equilibrium （HWE）is a rule to check whether observed genotypic frequencies and allele frequencies between parents and their offspring are in equilibrium in a population. During HWE testing，frequencies of those alleles must be estimated. When there were dominant and recessive genotypes，we used Newton alternate algorithm to get maximum likelihood estimate （MLE）of allele frequencies. The research findings indicated that Newton alternate algorithm had a faster convergence than method of maximum likelihood. An approximate formula about the least Pearson χ^2 statistic estimate and MLE was discussed. Newton alternate algorithm on MLE and the least Pearson χ^2 statistic estimate were recommended. ［Ch，10 ref.］

Key words：Hardy-Weinberg equilibrium(HWE)；HWE test；frequency of allele；maximum likelihood estimate；minimum Pearson χ^2 estimate

哈迪-温伯格平衡定律(Hardy-Weinberg equilibrium，HWE)亦称遗传学平衡定律，是群体遗传学中最重要的原理，是群体有性繁殖上下代之间基因频率与基因型频率是否保持平衡的检验尺度[1-10]。HWE检验方法主要有：一是针对大样本的拟合优度检验，如Pearson χ^2检验和似然比检验；二是针对小样本的精确检验。在HWE检验时通常需要对各基因频率作估计，当存在显隐性时，常用极大似然法估计基因频率，但没有直接的精确计算公式。文章介绍牛顿迭代算法，它比期望最大化(EM)算法收敛快。另外，极大似然法估计参数对应的Pearson χ^2统计量一般不会最小。文章讨论了两者之间的近似关系。最后通过例子进行数值计算，比较了极大似然估计牛顿迭代法和最小Pearson χ^2估计法的优劣。

1 存在显隐性时基因频率的极大似然估计

设等位基因A_1，A_2，…，A_k（$k\geqslant 2$），基因频率依次为p_1，p_2，…，p_k（其中$p_1+p_2+\cdots+p_k=1$；

收稿日期：2008-06-23；修回日期：2008-09-26
基金项目：浙江省自然科学基金资助项目(Y607480)；浙江省森林培育重中之重学科开放基金资助项目(200604)
作者简介：管宇，副教授，从事统计计算和生物数学等研究。E-mail：guanyu@zjfc.edu.cn

p_1，p_2，…，$p_k \geqslant 0$），共有 $k(k+1)/2$ 个基因型在一个大的随机群体中。通常为：

$$\begin{bmatrix} A_1 & A_2 & \cdots & A_k \\ p_1 & p_2 & \cdots & p_k \end{bmatrix}^2 = \begin{bmatrix} A_1A_1 & A_2A_2 & \cdots & A_kA_k & A_1A_2 & \cdots & A_iA_j & \cdots & A_{k-1}A_k \\ p_1^2 & p_2^2 & \cdots & p_k^2 & 2p_1p_2 & \cdots & 2p_ip_j & \cdots & 2p_{k-1}p_k \end{bmatrix}。$$

如果不存在显隐性时，容易推得基因 A_j 频率的极大似然估计：

$$\hat{p_j} = \frac{N_j}{N}\ (j = 1,\ 2,\ \cdots,\ k)。\tag{1}$$

式(1)中，$N_j = n_{jj} + \sum_{i=1}^{k} n_{ij}$（$1 \leqslant j \leqslant k$），$n_{ij}$ 表示基因型 A_iA_j（$1 \leqslant i,\ j \leqslant k$）的实际观测数，及 $N = 2\sum_{1 \leqslant i \leqslant j \leqslant k} n_{ij}$。

当存在显隐性基因时，基因频率的估计会有些麻烦。设 k 个复等位基因共有 m 个表型（$m \leqslant k\ (k+1)/2$），第 i 个表型（$i = 1,\ 2,\ \cdots,\ m$）的样本观测数 n_i 相对应的基因型频率为 q_i，易知 q_i 等于 p_s^2 或 $2p_sp_t$，或 p_s^2 与若干个 $2p_sp_t$ 的和（$s \neq t$，s，$t = 1,\ 2,\ \cdots,\ k$）。完全数据 $n_i(i = 1,\ 2,\ \cdots,\ m)$ 的对数似然函数：

$$\ln L(p) = c + \sum_{i=1}^{m} n_i \ln q_i = c + \sum_{j=1}^{k} n'_j \ln p_j + \sum_{i=1}^{m} \delta_i n_i \ln r_i。$$

其中：c 是与基因频率无关的常数；n'_j 是基因 A_j-的样本观测数（$j = 1,2,\cdots,k$）；当第 i 个表型是杂合体时 $\delta_i = 1$，否则取 0；r_i 等于 q_i 除去某个 p_i（公因子）。

对数似然函数分别对 p_1，p_2，…，p_{k-1} 求偏导数：

$$\frac{\partial \ln L(p)}{\partial p_j} = \frac{n'_j}{p_j} - \frac{n'_k}{p_k} + \sum_i \left(\delta_{i1}^{(j)} \frac{n_i}{r_i} - \delta_{i2}^{(j)} \frac{2n_i}{r_i} - \delta_{i3}^{(j)} \frac{n_i}{r_i}\right)(j = 1,2,\cdots,k-1)。$$

其中：当 r_i 中含有 p_j 但不含有 p_k 时，$\delta_{i1}^{(j)} = 1$，否则取 0；当 r_i 中含有 $p_j + 2p_k$ 时，$\delta_{i2}^{(j)} = 1$，否则取 0；当 r_i 中不含有 p_j 但含有 $2p_k$ 时，$\delta_{i3}^{(j)} = 1$，否则取 0。因此，求极大似然估计相当于解非线性方程组：

$$\begin{cases} \dfrac{n'_j}{p_j} - \dfrac{n'_k}{p_k} + \displaystyle\sum_i \left(\delta_{i1}^{(j)} \frac{n_i}{r_i} - \delta_{i2}^{(j)} \frac{2n_i}{r_i} - \delta_{i3}^{(j)} \frac{n_i}{r_i}\right) = 0\ (j = 1,2,\cdots,k-1)。 \\ \displaystyle\sum_{l=1}^{k} p_l = 0 \end{cases}\tag{2}$$

这个方程组没有简单的代数解法。由于方程组左边式子中的分母都是 p_1，p_2，…，p_k 的一次式，求导非常方便，完全可利用牛顿法求解方程组。记 $k-1$ 元函数向量：

$$f(p) = \frac{\partial \ln L(p)}{\partial p_j} = \left(\frac{\partial \ln L(p)}{\partial p_1},\ \frac{\partial \ln L(p)}{\partial p_2},\ \cdots,\ \frac{\partial \ln L(p)}{\partial p_{k-1}}\right)^T。$$

其中向量 $p = (p_1,\ p_2,\ \cdots,\ p_{k-1})^T$，$p_k = 1 - (p_1 + p_2 + \cdots + p_{k-1})$。$f(p)$ 的雅可比矩阵：

$$\frac{\partial f(p)}{\partial p} = \frac{\partial^2 \ln L(p)}{\partial p^2} = \begin{bmatrix} \dfrac{\partial^2 \ln L(p)}{\partial p_1^2} & \dfrac{\partial^2 \ln L(p)}{\partial p_1 \partial p_2} & \cdots & \dfrac{\partial^2 \ln L(p)}{\partial p_1 \partial p_{k-1}} \\ \dfrac{\partial^2 \ln L(p)}{\partial p_2 \partial p_1} & \dfrac{\partial^2 \ln L(p)}{\partial p_2^2} & \cdots & \dfrac{\partial^2 \ln L(p)}{\partial p_2 \partial p_{k-1}} \\ \cdots & \cdots & \cdots & \cdots \\ \dfrac{\partial^2 \ln L(p)}{\partial p_{k-1} \partial p_1} & \dfrac{\partial^2 \ln L(p)}{\partial p_{k-1} \partial p_2} & \cdots & \dfrac{\partial^2 \ln L(p)}{\partial p_{k-1}^2} \end{bmatrix}。\tag{3}$$

牛顿法求解方程组迭代公式：

$$p^{(t+1)} = p^{(t)} - \left(\frac{\partial f(p^{(t)})}{\partial p^{(t)}}\right)^{-1} f(p^{(t)})(t = 0,\ 1,\ 2,\ \cdots)。\tag{4}$$

牛顿迭代法式(4)具有至少二阶收敛速度，一般迭代少数几次即达到数值计算要求，它是数值求解方程的最主要方法。本算法主要的计算量应该是求 $k-1$ 阶雅可比矩阵的逆矩阵，另外牛顿法要求

初值应该在真值附近。

关于初始值的估计我们设计如下：①从只在某一表型中出现的基因(不妨设之 A_1)开始，若该表型是纯合体，则$\hat{p}_1 = \sqrt{n_{11}/n}$；若该表型是杂合体，则$\hat{p}_1 = n_{11}/2n$。估算所有这样的基因频率。②考虑 2 个或 2 个以上表型的基因（不妨设之 A_2)，但要求这些表型中除基因 A_2 外其他的都是前面已估计过的基因；如此循环，最终基因频率估计一般地总体上能在真值附近。

使用牛顿法的两大关键点是一、二阶导数和初始值问题，基因频率问题的对数似然函数的一阶、二阶偏导数非常容易给出一般性的公式，且初始值也容易得出一般算式，因此，牛顿法估计基因频率的程序可方便地编写和运行。

极大似然估计值也可考虑利用 EM 算法进行迭代计算估计[10]。据实际数值计算发现，牛顿法收敛速度明显快于 EM 算法(参看例 1)。

2 最小 Pearson χ^2 估计

由于对数似然比检验时需要求对数，Pearson χ^2 检验统计量仅涉及简单算术运算，因此人们拟合检验时总是首选 Pearson χ^2 检验。另一方面，参数极大似然估计是充分统计量且有大样本下的相合性、有效性等优良统计性质，估计参数时人们首选极大似然估计。非常遗憾，当参数取极大似然估计值时，对数似然比检验统计量达最小值，但 Pearson χ^2 检验统计量通常不是最小值。这样就有了参数的最小 Pearson χ^2 估计，即参数的估计值使得 Pearson χ^2 检验统计量达到最小值。

Pearson χ^2 检验统计量：

$$\chi^2 = \sum_{i=1}^{m} \frac{(n_i - nq_i)^2}{nq_i} = \sum_{i=1}^{m} \frac{n_i^2}{nq_i} - 2n + n = \sum_{i=1}^{m} \frac{n_i^2}{nq_i} - n。 \tag{5}$$

求偏导数：

$$\frac{\partial \chi^2}{\partial p_j} = -\sum_{i=1}^{m} \frac{n_i^2}{nq_i^2}\frac{\partial q_i}{\partial p_j} = -n\sum_{i=1}^{m} \frac{n_i}{nq_i}\frac{\partial q_i}{\partial p_j} - \sum_{i=1}^{m} \frac{n_i}{nq_i^2}\frac{\partial q_i}{\partial p_j}(n_i - nq_i)$$

$$= -n\frac{\partial \ln L(p)}{\partial p_j} - \sum_{i=1}^{m} \frac{n_i}{nq_i^2}\frac{\partial q_i}{\partial p_j}(n_i - nq_i), (j = 1,2,\cdots,k-1)。$$

当原假设为真时，即满足 Hardy-Weinberg 平衡或实验数据与理论模型相吻合时，n_i 与 nq_i 非常接近($i = 1,2,\cdots,m$)，$\left|\sum_{i=1}^{m} \frac{n_i^2}{nq_i^2}\frac{\partial q_i}{\partial p_j}(n_i - nq_i)\right| << \left|n\frac{\partial \ln L(p)}{\partial p_j}\right|$。换句话说，当理论模型假设正确时，参数的极大似然估计值与最小 Pearson χ^2 估计值非常接近。经数值模拟运算，在 Pearson χ^2 拟合检验时，用极大似然估计法与最小 Pearson χ^2 法估计参数在大多数情形下会得到相同的结论。显然，利用牛顿法也能计算得参数的最小 Pearson χ^2 估计值，但运算过程要略比求极大估计值麻烦。

最小 Pearson χ^2 估计法具体可按以下步骤操作：①首先利用第 1 节介绍的方法得基因频率的初始值$(p_1,p_2,\cdots,p_k)$，计算其卡方值$\chi^2 = \sum_{i=1}^{m} \frac{(n_i - nq_i)^2}{nq_i} = \sum_{i=1}^{m} \frac{n_i^2}{nq_i} - n$。若$\chi^2 \leqslant \chi_\alpha^2$（通常取 $\alpha = 0.05$)，则输出结果：基因频率估计值$(p_1,p_2,\cdots,p_k)$和哈迪-温伯格平衡定律满足。若$\chi^2 > \chi_\alpha^2$，则进入第②步。②计算最小 Pearson χ^2 值。可直接调用现有求极值程序：如科学计算软件 Matlab 中的指令 “fminsearch(χ^2, $[p_1,p_2,\cdots,p_k]$)”(输入函数式和自变量的初始值)非常方便地得到最小卡方值和对应的自变量值；也可利用 Basic 或 C 等语言按现成算法自己编写程序。

这里：步骤①实是一般文献中的方法；Hardy-Weinberg 平衡定律满足，此时的基因频率估计值虽不一定最合适，但至少在概率意义下与真值的偏差不会大；增加步骤②的目的是减小假设检验时犯第 1 类(弃真)错误的概率。如果需要尽可能高的拟合度，则直接做第②步计算最小 Pearson χ^2 值及其相应的频率估计值。

3　各算法的数值运算比较

例 1　某地有 6 000 人，其中 AB，A，B 和 O 型血型的频数分别是 627，1 900，1 627 和 1 846。(文献[10]中原始数据中数目分别为：607，1 920，1 627，1 846，其χ^2检验值不到 0.4，很小，各方法输出结果差异较小。改动数据使χ^2值不太小，这样各种算法的输出结果就有了较明显差异)。

3.1　Bernstein 法

对于 ABO 血型系统中等位基因频率的估计，Bernstein 公式无疑是最简便的方法，当然，误差也不小。

$$p_A = (1 - \sqrt{(n_O + n_B)/n})/D,\ p_B = (1 - \sqrt{(n_O + n_A)/n})/D,\ p_O = (\sqrt{n_O/n})/D,$$

$$D = 2 - \sqrt{(n_O + n_A)/n} - \sqrt{(n_O + n_B)/n} + \sqrt{n_O/n}。$$

计算得：$p_A = 0.238\ 3\cdots$，$p_B = 0.209\ 0\cdots$，$p_O = 0.552\ 6\cdots$。相应的χ^2值为 2.037 2…，相应的χ^2 p 值为 0.153 4…。

几乎所有的高精度数值计算都是通过迭代来完成的，初始值的好坏有时会影响到迭代的收敛性。Bernstein 公式简单，是迭代初始值的首选者。

3.2　最小 Pearson χ^2 估计法

利用计算软件 Matlab，仅需编写简单的 3 条指令(初始值 $p_A = 0.3$，$p_B = 0.2$)：

fun = inline('(627^2/2/p(1)/p(2) + 1900^2/(p(1) + 2*(1 - p(1) - p(2)))/p(1) + 1627^2/(p (2) + 2*(1 - p(1) - p(2)))/p(2) + 1846^2/(1 - p(1) - p(2))/(1 - p(1) - p(2 /6000 - 6000', 'p');

[p, g] = fminsearch(fun, [0.3, 0.2]), P_value=1-chi2cdf(g, 1)。

输出结果(改写成文字形式)：

当 p_A= 0.238 768 817 964 01 和 p_B = 0.209 523 809 184 32 时，χ^2 值取最小值 2.004 455 896 821 13，相应的卡方 p 值为 0.156 837 559 639 45。

3.3　极大似然估计法(Matlab)

利用计算软件 Matlab，初始值 $p_A = 0.3$，$p_B = 0.2$，需要编写 5 条指令(程序，略)。输出结果(改写成文字形式)：

当 p_A = 0.238 772 519 587 07 和 p_B = 0.209 459 739 436 86 时，χ^2 值取最小值 2.004 512 737 426 92，相应的卡方 p 值为 0.156 831 680 697 21。

3.4　极大似然估计法(牛顿算法)

以 Bernstein 公式计算得初始值，牛顿法(公式和程序相对前面 3 种算法较复杂，略)迭代 3 次后，双精度下数值不再发生变化。其中 3 次迭代 p_A 值分别是 0.238 743 914 343 03，0.238 744 148 258 46，0.238 744 148 258 52。第 3 次迭代后的χ^2值为 2.004 586 448 805 64，χ^2 p 值为0.156 824 057 208 10。如果初始值取 $p_A = 0.3$，$p_B = 0.2$，则需 5 次迭代即可。

3.5　极大似然估计法(EM 算法)

以 Bernstein 公式计算得初始值，EM 算法(公式和程序较牛顿法简单些，略)迭代 15 次后，双精度下数值不再发生变化。其中迭代到第 5 次时 p_A 的值是 0.238 744 083 924 78，第 15 次时才等于 0.238 744 148 258 52。如果初始值取 $p_A = 0.3$，$p_B = 0.2$，则需 17 次迭代方可。

因此，在基因频率的估计时，如果需要完全编程计算的话，极大似然估计牛顿迭代法应该是最好的，因为收敛较快；如果有现成的极值之类函数可供调用，那么最小 Pearson χ^2 估计法无疑是最实用最简便且精度高的算法。

参考文献：

[1] LI C C. *Population Genetics*[M]. Chicago：University of Chicago Press，1955.

[2] LI C C. Pseudo-random mating population[J]. *Genetics*，1988，**119**：731 – 737.

[3] CHEN J J，DUAN T，SINGLE R，*et al*. Hardy-Weinberg testing of single homozygous genotype[J]. *Genetics*，2005，**170**：

1439 – 1442.
[4] STARK A E. A clarification of the Hardy-Weinberg Law[J]. *Genetics*, 2006, **174**: 1695 – 1697.
[5] WIGGINTON J E, DUAN T, CUTLER D J, *et al*. A note on exact test of Hardy-Weinberg equilibrium[J]. *Am J Hum Genet*, 2005, **76**: 887 – 893.
[6] NADER E, DEVRIM B. A new method of Hardy-Weinberg equilibrium and ordering populations[J]. *J Genet*, 2007, **86**: 1 – 7.
[7] ANDRE R, MICHAEL J S, *et al*. Hardy-Weinberg equilibrium diagnostics[J]. *Theor Popul Biol*, 2002, **62**: 251 – 257.
[8] 黄代新, 杨庆恩. 卡方检验和精确检验在 HWE 检验中的应用[J]. 法医学杂志, 2004, **20** (2): 116 – 119.
HUANG Daixin, YANG Qingen. Application of chi-square test and exact test in Hardy-Weinberg equilibrium testing[J]. *J Foren Med*, 2004, **20** (2): 116 – 119.
[9] 汪小龙, 袁志发, 郭满才, 等. 最大信息熵原理与群体遗传平衡[J]. 遗传学报, 2002, **29** (6): 562 – 564.
WANG Xiaolong, YUAN Zhifa, GUO Mancai, *et al*. Maximum entropy principle and population genetic equilibrium[J]. *Acta Genet Sin*, 2002, **29** (6): 562 – 564.
[10] 李照海, 覃红, 张洪. 遗传学中的统计方法[M]. 北京: 科学出版社, 2006, 1 – 27.

第 6 届全国林木遗传育种大会在浙江林学院举行

2008 年 11 月 7 – 10 日, 由中国林学会林木遗传育种分会主办, 浙江林学院林业与生物技术学院以及浙江省林业厅种苗站、中国林科院亚热带林业研究所、浙江省林业科学研究院、浙江森禾种业股份有限公司等单位共同承办的第 6 届全国林木遗传育种大会在浙江林学院举行。

中国林学会常务副秘书长李岩泉、国家林业局种苗总站副站长刘红、浙江省林业厅副厅长吴鸿、东北林业大学校长杨传平、南京林业大学副校长施季森以及浙江林学院周国模、方伟、张立钦等校领导出席大会; 来自中国 30 个省(市、自治区)、中国香港特别行政区及美国、澳大利亚等国的遗传育种界资深专家共 300 余名代表参加了大会。

大会审议通过了沈熙环教授代表第 5 届林学会林木育种分会作的 6 年来分会工作报告, 充分肯定了第 5 届中国林学会林木育种分会所做的工作, 明确了下一阶段工作方向。大会围绕当前林木遗传育种的进展和存在问题作了研讨。22 位专家进行了专题发言, 就我国林木育种事业可持续发展战略、常规育种的特点和成功的关键技术、现代生物技术在学科中的应用等方面的问题作了阐述与交流。

大会共收到论文摘要 205 篇, 交流报告 35 篇。大会还选举产生了新一届委员会。

新闻中心