遗传 HEREDITAS (Beijing) 6(5): 42—48 1984

讲座

最大似然法及其应用

莫惠栋

(江苏农学院数量遗传研究室，扬州)

最大似然法是参数估计的一种重要方法。在遗传学研究中，广泛地应用于计数资料的总体成数估计。由于估计值以满足在观察结果中的出现概率最大为条件，故又称最大似然估计。

一、基本原理

设一总体的各个体可根据某些特征而分成 k 组，各组的理论频率为 $p_i(i=1,2,\cdots,k)$；而 p_i 又是将要估计的参数 p 的某种函数［即 $p_i=f(p)$，但其具体形式可随 i 的不同而不同，由有关专业知识确定］，并有 $\sum_1^k p_i=1$。则以容量 n 抽样，各组观察次数 $a_i\left(\sum_1^k a_i=n\right)$ 的概率分布为复项式(multinomial)：

$$(p_1+p_2+\cdots+p_k)^n \tag{1}$$

展开。而特定于某一观察的 $(a_1, a_2, \cdots, a_k)$ 组合的概率(似然率) L 则为：

$$L=\frac{n!}{a_1!a_2!\cdots a_k!}(p_1)^{a_1}(p_2)^{a_2}\cdots(p_k)^{a_k} \tag{2}$$

由 (2) 找出参数 p 的估计值 $\hat{p}$，使之满足 L 为最大，就是对 p 的最大似然估计。这个问题显然只是对方程 $dL/dp=0$ 求根。为便于微分，可先对(2)作对数变换，即有：

$$\ln L=C+a_1\ln p_1+a_2\ln p_2+\cdots+a_k\ln p_k \tag{3}$$

(3)中的 C 为常数项，在此

$$C=\ln\left(\frac{n!}{a_1!a_2!\cdots a_k!}\right),$$

因在微分时成为 0，可省略。这里 (3) 的 $\ln L$ 最大和(2)的 L 最大显然等价，故 p 的最大似然估计值 $\hat{p}$ 即方程：

$$\frac{d(\ln L)}{dp}=\sum_1^k a_j\frac{d(\ln p_j)}{dp}=\sum_1^k\frac{a_j}{p_j}\left(\frac{dp_j}{dp}\right)=0 \tag{4}$$

的根。根据 Rao-Cramér 不等式，不难证明，p 的抽样方差 V_p 渐近于：

$$\frac{1}{V_p}=-E\left(\frac{d^2(\ln L)}{dp^2}\right) \tag{5}$$

(5) 中的 $\left(\frac{d^2(\ln L)}{dp^2}\right)$ 为 $\ln L$ 对于 p 的二阶导数；E 为取期望，在我们讨论的范围内，即以 np_i 代 a_i。由于：

$$\frac{d^2(\ln L)}{dp^2}=-\sum_1^k\frac{a_j}{p_j^2}\left(\frac{dp_j}{dp}\right)^2+\sum_1^k\frac{a_j}{p_j}\left(\frac{d^2p_j}{dp^2}\right)$$

$$E\left(\frac{d^2(\ln L)}{dp^2}\right)=-n\sum_1^k\frac{1}{p_j}\left(\frac{dp_j}{dp}\right)^2+n\sum_1^k\frac{d^2p_j}{dp^2}=-n\sum_1^k\frac{1}{p_j}\left(\frac{dp_j}{dp}\right)^2$$

$$\left(\text{因为 }\sum_1^k\frac{d^2p_j}{dp^2}=\frac{d}{dp}\sum_1^k p_j=0\right)$$

所以(5)可变形为：

$$\frac{1}{V_p}=n\sum_1^k\left[\frac{1}{p_j}\left(\frac{dp_j}{dp}\right)^2\right]=n\sum_1^k i_j=I \tag{6}$$

Mo Huidong: The Method of Maximum Likelihood and Its Application in Genetics

(6)为 R. A. Fisher 曾予定义的“信息函数”。其中 I 称总信息量，$\sum_1^k i_j = \frac{I}{n}$（或记作 i_p）称单一观察信息量。$\sum_1^k i_j$ 愈大，表明样本中有关“p 的信息”愈多，于是 V_p 愈小，对 p 的估计愈可靠。当 $p_j = f(p)$ 的关系比较复杂时，由(6)求 V_p 将特别简便。

在大样本时，p 的抽样分布逼近正态。因而有了 V_p，就可对 p 作出区间估计。

以上讨论的是估计一个成数的最大似然值。如要估计几个成数，即 $p_j = f(p, q, r, \cdots)$，而 p、q、r 等都是需要独立估计的。则可由 $\partial(\ln L)/\partial p = 0$，$\partial(\ln L)/\partial q = 0$，$\partial(\ln L)/\partial r = 0$ 等而组成的联立方程解出。这里并无新的原理，但所得方程组往往缺乏代数学的一般解法，需迭代试估。

二、 二项成数的最大似然估计

设某总体的各个体可不含糊地分成 A、B 两组，各具成数 p 和 $1-p=q$。则在随机观察 n 个个体时，一个特定的 A 组个体数 a_1、B 组个体数 a_2 的出现概率为：

$$L = \frac{n!}{a_1!a_2!}(p_1)^{a_1}(p_2)^{a_2} = \frac{n!}{a_1!a_2!}(p)^{a_1}(1-p)^{a_2}$$

或

$$\ln L = C + a_1 \ln p + a_2 \ln(1-p)$$

求导得：

$$\frac{d(\ln L)}{dp} = \frac{a_1}{p} - \frac{a_2}{1-p} = 0$$

即

$$a_1(1-p) - a_2 p = 0$$

因而对于 p（或 q）的最大似然估计为：

$$\left.\begin{aligned} \hat{p} &= \frac{a_1}{a_1 + a_2} = \frac{a_1}{n} \\ \text{或} \quad \hat{q} &= 1 - \hat{p} = \frac{a_2}{n} \end{aligned}\right\} \tag{7}$$

对于抽样方差 $V_p = V_q$，如根据(5)，则有：

$$\frac{1}{V_p} = -E\left(\frac{d^2(\ln L)}{dp^2}\right) = -E\left[\frac{-a_1}{p^2} - \frac{a_2}{(1-p)^2}\right] = \frac{np}{p^2} + \frac{n(1-p)}{(1-p)^2} = \frac{n}{p(1-p)}$$

$$\therefore V_p = \frac{\hat{p}(1-\hat{p})}{n} \tag{8}$$

如根据(6)，则就 A 组而言，$p_1 = p$，$dp_1/dp = 1$，$i_1 = 1/p$；就 B 组而言，$p_2 = 1-p$，$dp_2/dp = -1$，$i_2 = 1/(1-p)$。所以，$I = n(i_1 + i_2) = n\left(\frac{1}{p} + \frac{1}{1-p}\right) = \frac{n}{p(1-p)}$，$V_p = \frac{1}{I} = \frac{\hat{p}(1-\hat{p})}{n}$，结果同上。

以上的(7)和(8)，就是已广泛使用的二项成数及其方差。这里仅是证明其为最大似然估计。

例 1 棉花幼芽有黄(隐性)、绿(显性)两种。现以一具芽黄性状棉为母本，与绿芽性状棉间种，使天然杂交，再收取芽黄棉上的种子。次年检查了后裔幼苗 250 株 (n)，得具芽黄性状的 162 株 (a_1)。试以最大似然法估计自花授粉率。

这里后裔为芽黄株即自花授粉（含品种内授粉)株，因而由 (7) 和 (8) 得自花授粉率 $\hat{p}$ 及其抽样方差为：

$$\hat{p} = \frac{162}{250} = 0.648,$$

$$V_p = \frac{0.648 \times 0.352}{250} = 0.0009124$$

或置信系数为 0.95 的自花授粉率区间可估计为：

$$0.648 \mp 1.96 \times \sqrt{0.0009124} = 0.589 \sim 0.707$$

即 58.9—70.7%。

三、 基因频率的最大似然估计

一对等位基因有 AA、Aa 和 aa 三种基因型，在遗传平衡时，其频率依次为 $p_1 = p^2$，$p_2 = 2p(1-p) = 2pq$，$p_3 = (1-p)^2 = q^2$。这里的 p 和 q 分别为 A 和 a 基因的频率。如抽样观

察了 n 个个体，得 AA、Aa 和 aa 的个体数依次为 a、b、c 个 $(a+b+c=n)$，则获得该 (a, b, c) 组合的概率为：

$$L = \frac{n!}{a!b!c!}(p_1)^a(p_2)^b(p_3)^c$$
$$= \frac{n!}{a!b!c!}(p^2)^a(2pq)^b(q^2)^c$$
$$= \frac{n!2^b}{a!b!c!}(p)^{2a+b}(q)^{b+2c}$$
$$= \frac{n!2^b}{a!b!c!}(p)^{2a+b}(1-p)^{b+2c}$$

而

$$\ln L = C + (2a+b)\ln p + (b+2c)\ln(1-p)$$
$$\frac{d(\ln L)}{dp} = \frac{(2a+b)}{p} - \frac{(b+2c)}{1-p} = 0$$

即

$$(1-p)(2a+b) - p(b+2c) = 0$$
$$2a + b - 2np = 0$$
$$(\because\ 2a+2b+2c = 2n)$$

所以 A 基因频率 p 的最大似然估计为：

$$\hat{p} = \frac{2a+b}{2n} \tag{9}$$

或 a 基因频率 q 的最大似然估计为：

$$\hat{q} = 1 - \hat{p} = (b+2c)/2n$$

对于(9)的抽样方差 V_p，据(5)可得：

$$\frac{1}{V_p} = -E\left(\frac{d^2(\ln L)}{dp^2}\right)$$
$$= -E\left[-\frac{2a+b}{p^2} - \frac{b+2c}{(1-p)^2}\right]$$
$$= \frac{2np}{p^2} + \frac{2n(1-p)}{(1-p)^2} = \frac{2n}{p(1-p)}$$

故

$$V_p = \frac{\hat{p}(1-\hat{p})}{2n} \tag{10}$$

如根据 (6)，则可先在表 1 求单一观察信息：

表 1 (9) 的单一观察信息 $\sum_1^k i_j$

组　别	期望频率 p_j	$\frac{dp_j}{dp}$	$i_j = \frac{1}{p_j}\left(\frac{dp_j}{dp}\right)^2$
1. AA	$p_1 = p^2$	$2p$	$i_1 = 4p^2/p^2 = 4$
2. Aa	$p_2 = 2p(1-p)$	$2(1-2p)$	$i_2 = \frac{4(1-2p)^2}{2p(1-p)} = \frac{2(1-2p)^2}{p(1-p)}$
3. aa	$p_3 = (1-p)^2$	$-2(1-p)$	$i_3 = 4(1-p)^2/(1-p)^2 = 4$
总　和	1	0	$\sum_1^k i_j = 8 + \frac{2(1-2p)^2}{p(1-p)}$

因此有

$$I = 2n\left[4 + \frac{(1-2p)^2}{p(1-p)}\right] = \frac{2n}{p(1-p)},$$
$$V_p = \hat{p}(1-\hat{p})/2n,$$

结果同(10)。

例 2 人类的 M-N 血型有 M、MN 和 N 三种，设其基因型为 MM、MN 和 NN。现测定 1029(n) 人，得三种人数依次为 342(a)、500(b) 和 187(c)。试估计 M 基因频率 p（或 N 基因频率 $q = 1-p$）。

(9) 给出 M 基因频率的最大似然估计为：

$$\hat{p} = \frac{2\times 342 + 500}{2\times 1029} = 0.5753$$

(或 N 基因频率 $\hat{q} = 1 - \hat{p} = 0.4247$)。其抽样方差为：

$$V_p = \frac{0.5753\times 0.4247}{2\times 1029} = 0.0001187$$

故 M 基因频率可表示为：

$$\hat{p} \pm \sqrt{V_p} = 0.5753 \pm 0.0109。$$

这里的方法可推广应用于某些复等位基因的频率估计。例如有等位基因 A_1、A_2、A_3，需分别估计其频率 p、q 和 r。由于遗传平衡时，

表 2　相引连锁的配子和后裔基因型及频率

♀ \ ♂	AB $\frac{1}{2}(1-p)$	Ab $\frac{1}{2}p$	aB $\frac{1}{2}p$	ab $\frac{1}{2}(1-p)$
AB, $\frac{1}{2}(1-p)$	AABB $\frac{1}{4}(1-p)^2$	AABb $\frac{1}{4}p(1-p)$	AaBB $\frac{1}{4}p(1-p)$	AaBb $\frac{1}{4}(1-p)^2$
Ab, $\frac{1}{2}p$	AABb $\frac{1}{4}p(1-p)$	AAbb $\frac{1}{4}p^2$	AaBb $\frac{1}{4}p^2$	Aabb $\frac{1}{4}p(1-p)$
aB, $\frac{1}{2}p$	AaBB $\frac{1}{4}p(1-p)$	AaBb $\frac{1}{4}p^2$	aaBB $\frac{1}{4}p^2$	aaBb $\frac{1}{4}p(1-p)$
ab, $\frac{1}{2}(1-p)$	AaBb $\frac{1}{4}(1-p)^2$	Aabb $\frac{1}{4}p(1-p)$	aaBb $\frac{1}{4}p(1-p)$	aabb $\frac{1}{4}(1-p)^2$

6 种基因型及其频率可分成如下三组：

基因型：A_1A_1，　$A_1A_2+A_1A_3$，　$A_2A_2+A_2A_3+A_3A_3$；

期望频率：p^2，　$2p(q+r)$，　$(q+r)^2$；

因而，有了 A_1A_1、$(A_1A_2+A_1A_3)$ 和 $(A_2A_2+A_2A_3+A_3A_3)$ 的观察数 a、b 和 c，就可令 $(q+r)=1-p$，从而由 (9) 得 $\hat{p}$。同样，若归纳成如下三组：

基因型：A_2A_2，　$A_1A_2+A_2A_3$，　$A_1A_1+A_1A_3+A_3A_3$；

期望频率：q^2，　$2q(p+r)$，　$(p+r)^2$；

并令 $(p+r)=1-q$，就可由 (9) 得 $\hat{q}$。而 r 则可从 $\hat{r}=1-\hat{p}-\hat{q}$ 得出。

四、连锁基因交换率的最大似然估计

在试验可以直接计数重组个体时，交换率 p 即重组个体占总个体数的成数，其算式即(7)和(8)，不需复述。这里要讨论的是由自交的 F_2 代估计交换率的最大似然方法。这在难以获得大量回交后裔的生物上(如稻、麦)特别有用。

设基因 A-a、B-b 为相引连锁，且 A 对 a、B 对 b 为显性，交换率为 p。则 $\frac{AB}{ab}$ 基因型产生重组配子 Ab 和 aB 的频率各为 $\frac{1}{2}p$，产生连锁配子 AB 和 ab 的频率各为 $\frac{1}{2}(1-p)$。雌、雄配子随机结合，后裔各种基因型及其频率列于表 2。

表 2 共 9 种基因型，在存在显性时，只能区别 4 种表型：A-B-，具频率：

$$p_1=3\left[\frac{1}{4}(1-p)^2\right]+4\left[\frac{1}{4}p(1-p)\right]+2\left[\frac{1}{4}p^2\right]=\frac{1}{2}+\frac{1}{4}(1-p)^2;$$

A-bb 和 aaB-，各具频率：

$$p_2=p_3=\frac{1}{4}p^2+2\left[\frac{1}{4}p(1-p)\right]=\frac{1}{4}-\frac{1}{4}(1-p)^2;$$

aabb，具频率：$p_4=\frac{1}{4}(1-p)^2$。

设上述 4 种表型的观察数依次为 a_1、a_2、a_3、a_4，并有 $\sum_1^4 a_i=n$，则由(3)得：

$$\ln L=a_1\ln\left[\frac{1}{2}+\frac{(1-p)^2}{4}\right]+(a_2+a_3)\ln\left[\frac{1}{4}-\frac{(1-p)^2}{4}\right]+a_4\ln\left[\frac{(1-p)^2}{4}\right]$$

现在要求得使 $\ln L$ 为最大的 p 值。为便于运算，令

$$K=(1-p)^2 \qquad (11)$$

于是有：

$$\ln L=a_1\ln\left[\frac{1}{2}+\frac{K}{4}\right]+(a_2+a_3)\ln\left[\frac{1}{4}-\frac{K}{4}\right]$$

表 3 (13)的单一观察信息 $\sum_{1}^{4} i_j$

组　别	期望频率 p_j	$\frac{dp_j}{dp}$	$i_j=\frac{1}{p_j}\left(\frac{dp_j}{dp}\right)^2$
1. A_B_	$p_1=\frac{1}{2}+\frac{1}{4}(1-p)^2$	$-\frac{2}{4}(1-p)$	$i_1=\frac{(1-p)^2}{2+(1-p)^2}=\frac{K}{2+K}$
2. A_bb	$p_2=\frac{1}{4}-\frac{1}{4}(1-p)^2$	$\frac{2}{4}(1-p)$	$i_2=\frac{(1-p)^2}{1-(1-p)^2}=\frac{K}{1-K}$
3. aaB_	$p_3=\frac{1}{4}-\frac{1}{4}(1-p)^2$	$\frac{2}{4}(1-p)$	$i_3=\frac{(1-p)^2}{1-(1-p)^2}=\frac{K}{1-K}$
4. aabb	$p_4=\frac{1}{4}(1-p)^2$	$-\frac{2}{4}(1-p)$	$i_4=\frac{4(1-p)^2}{4(1-p)^2}=1$
总　和	n	0	$\sum_{1}^{4} i_j=\frac{K}{2+K}+\frac{2K}{1-K}+1$

$$+a_4\ln\left[\frac{K}{4}\right]$$

$$\frac{d(\ln L)^*}{dK}=\frac{a_1}{2+K}-\frac{(a_2+a_3)}{1-K}+\frac{a_4}{K}=0$$

即

$$nK^2+(2n-3a_1-a_4)K-2a_4=0$$

$$\therefore K=\frac{-(2n-3a_1-a_4)+\sqrt{(2n-3a_1-a_4)^2+8na_4}}{2n} \tag{12}$$

$$\hat{p}=1-\sqrt{K} \tag{13}$$

这就是交换率 p 的最大似然估计。下面根据(6)导出(13)的抽样方差，表 3 列出其单一观察信息。

据之可得：

$$I=n\left(\frac{K}{2+K}+\frac{2K}{1-K}+1\right)=\frac{n[K(1-K)+2K(2+K)+(2+K)(1-K)]}{(2+K)(1-K)}$$

$$\therefore V_p=\frac{(2+K)(1-K)}{n[K(1-K)+2K(2+K)+(2+K)(1-K)]} \tag{14}$$

当两对基因为相斥连锁时，仿表 2 格式可得表型 A_B_、A_bb、aaB_ 和 aabb 的期望频率依次为：$\frac{1}{2}+\frac{p^2}{4}$、$\frac{1}{4}-\frac{p^2}{4}$、$\frac{1}{4}-\frac{p^2}{4}$和$\frac{p^2}{4}$。

故

$$\ln L=a_1\ln\left[\frac{1}{2}+\frac{p^2}{4}\right]+(a_2+a_3)\left[\frac{1}{4}-\frac{p^2}{4}\right]+a_4\left[\frac{p^2}{4}\right]$$

若令

$$K=p^2 \tag{15}$$

同样得(12)和(14)。而相斥连锁基因的交换率则为：

$$\hat{p}=\sqrt{K} \tag{16}$$

这里注意：K 的取值区间，在(13)中为 $1>K>0.25$，在 (16) 中为 $0<K<0.25$。$K=0.25$ 时为独立遗传。

例 3 以紫花长花粉粒 (BBLL) 与红花园花粉粒 (bbll) 香豌豆杂交，在 F_2 代检查 6952 株 (n)，得 4 种表型的植株数为：

B_L_	B_ll	bbL_	bbll
4831	390	393	1338

这是一个相引连锁资料，试据之求基因交换率。

这里 $(2n-3a_1-a_4)=2\times6952-3\times4831-1338=-1927$。根据 (12)、(13) 和 (14)可得：

$$K=\frac{1927+\sqrt{1927^2+8\times6952\times1338}}{2\times6952}$$

* 根据 (4)，这里应为 $d(\ln L)/dp$。因 $K=(1-p)^2$，$dK/dp=-2(1-p)$，$\frac{d(\ln L)}{dp}=\frac{d(\ln L)}{dK}\cdot\frac{dK}{dp}=-2(1-p)\frac{d(\ln L)}{dK}=0$。这里 $-2(1-p)$ 为常数，故由 $d(\ln L)/dK=0$ 求 p 不影响结果。

$= 0.7743$

$$\hat{p} = 1 - \sqrt{0.7743} = 0.1201,$$

$$V_p = 1.76702 \times 10^{-5}$$

所以交换率（即产生重组配子 Bl 和 bL 的频率）为：

$0.1201 \pm \sqrt{1.76702 \times 10^{-5}} = 12.01 \pm 0.42\%$。

例 4 以正常穗矮生（AAdd）和疏穗非矮生（aaDD）水稻杂交，在 F_2 得 4 种表型的株数为：

基因型	A_D_	A_dd	aaD_	aadd	总和
植株数	647	273	288	11	1219

这是相斥连锁，现估计其交换率。

这里 $(2n - 3a_1 - a_4) = 2 \times 1219 - 3 \times 647 - 11 = 486$。根据(12)、(16)和(14)得：

$$K = \frac{-486 + \sqrt{486^2 - 8 \times 1219 \times 11}}{2 \times 1219}$$

$$= 0.041042$$

$$\hat{p} = \sqrt{0.041042} = 0.2026,$$

$$V_p = 7.41919 \times 10^{-4}$$

故基因交换率（即重组配子 AD 和 ad 的发生频率）估计为：

$0.2026 \pm \sqrt{7.41919 \times 10^{-4}} = 20.26 \pm 2.72\%$。

五、 其他成数的最大似然估计

在遗传学研究中需要估计成数的场合很多，不可能一一列述。以下举出两例以示一斑。

例 5 在甘蓝（*B. oleracea*）中发现有 3 对双价染色体减数分裂时有次级联会（secondary association）现象。一个共 337 个（n）细胞的观察结果为：

次级联会的染色体对数	0	1	2	3
观察到的细胞数	32	103	122	80

要是这 3 对染色体是随机次级联会的，具频率 p，则上述 4 种情况的期望频率应为 $(q + p)^3$ 展开（这里 $q = 1 - p$，因 $p + q = 1$），即各组频率依次为：

$$p_1 = (1 - p)^3, \quad p_2 = 3p(1 - p)^2,$$

$$p_3 = 3p^2 \times (1 - p), \quad p_4 = p^3$$

因此需首先估计 p。根据(3)和(4)，在此有：

$$\ln L = 32\ln(1 - p)^3 + 103\ln[3p(1 - p)^2] + 122\ln[3p^2(1 - p)] + 80\ln p^3$$

$$\frac{d(\ln L)}{dp} = \frac{-96}{1 - p} + \frac{103(1 - 3p)}{p(1 - p)} + \frac{122(2 - 3p)}{p(1 - p)} + \frac{240}{p}$$

$$= \frac{587 - 1011p}{p(1 - p)} = 0$$

$$\therefore \quad \hat{p} = 587/1011 = 0.5806$$

将此 $\hat{p} = 0.5806$ 代 p，即可估计各期望次数 $n\hat{p}_i$。如“0”的期望次数为 $n(1 - \hat{p})^3 = 337 \times (1 - 0.5806)^3 = 24.86$，“1”的期望次数为 $3n\hat{p}(1 - \hat{p})^2 = 3 \times 337 \times 0.5806 \times (1 - 0.5806)^2 = 103.25$ 等。以实际次数对期望次数作 χ^2 测验，就可作出次级联会是否随机的推断。这里可得 $\chi^2 = 8.105$，它 $> \chi^2_{0.05,2}$。因此，次级联会可能有着某种机制，而不是随机的。

这里的抽样方差也可根据 (6) 而方便地导出。其结果为：

$$I = n\left[9(1 - p) + \frac{3(1 - 3p)^2}{p} + \frac{3(2 - 3p)^2}{(1 - p)} + 9p\right]$$

$$= 3n\left[\frac{1}{p(1 - p)}\right]$$

$$V_p = \frac{p(1 - p)}{3n}$$

例 6 人类的 **ABO** 血型一般以 3 个等位基因解释，即基因 A 和 a′ 对 a 都是显性，但 A 对 a′ 互不为显性（共显性）。其基因型、血型和遗传平衡群体的期望频率为：

基因型	Aa′ AA 和 Aa	a′a′ 和 a′a	aa
血　型	**AB**	**A**	**B**	**O**
频　率	$2pq$	$p^2 + 2pr$	$q^2 + 2qr$	r^2

以上 p、q、r 依次为 A、a′ 和 a 基因的频率，并有 $p + q + r = 1$。现有我国 6000 人（n）的调查结果为：**AB** 型 607 人，**A** 型 1920 人，**B** 型 1627 人，**O** 型 1846 人。为估计 p、q 和 r，根据(3)，这里有：

$$\ln L = 607\ln(2pq) + 1920\ln(p^2 + 2pr)$$

$$+1627\ln(q^2+2qr)+1846\ln r^2$$
$$=607(\ln 2+\ln p+\ln q)+1920[\ln p+\ln(p+2r)]+1627[\ln q+\ln(q+2r)]+2\times 1846r$$
$$=2527\ln p+2234\ln q+1920\ln(p+2r)+1627\ln(q+2r)+3692\ln r+C=2527\ln p+2234\ln q+1920\ln(2-p-2q)+1627\ln(2-2p-q)+3692\ln(1-p-q)+C$$

故可得独立估计 p 和 q 的方程组：

$$\frac{\partial(\ln L)}{\partial p}=\frac{2527}{p}-\frac{1920}{2-p-2q}-\frac{2\times 1627}{2-2p-q}-\frac{3692}{1-p-q}=0$$

$$\frac{\partial(\ln L)}{\partial q}=\frac{2234}{q}-\frac{2\times 1920}{2-p-2q}-\frac{1627}{2-2p-q}-\frac{3692}{1-p-q}=0$$

而 $\hat{r}$ 则由 $1-\hat{p}-\hat{q}$ 得出。解上述方程组结果是：$\hat{p}=0.23900$，$\hat{q}=0.20758$，$\hat{r}=0.55342$。由之可进而得到 **AB**、**A**、**B** 和 **O** 血型的期望人数依次为 595.34、1929.93、1637.08 和 1837.64，具 $\chi^2=0.38$。这在 $df=1$ 时，$P>0.5$。所以人类的 **ABO** 血型为遗传平衡群体。

六、估计效率的比较

总体的某一成数 p，常常可能用不同的方法估计。方法的效率决定于单一观察所能提供的有关 p 的信息量。由 (6) 可知，在 n 一定时，该信息量愈大，则 V_p 愈小，估计愈精确。所以不同估计方法的相对效率 RE 可由各方法的单一观察信息量 $i_p=\sum_1^k i_j$ 之比得出，即：

$$RE=i_{p_1}/i_{p_2} \tag{17}$$

例 7 以 (13) 估计相引连锁基因的交换率时，单一观察的信息量（表 3）可记为：

$$i_{p_1}=\frac{(1-p)^2}{2+(1-p)^2}+\frac{2(1-p)^2}{1-(1-p)^2}+1 \tag{18}$$

现有人提出根据双隐性个体频率 (a_4/n) 的估计方法：由于 aabb 的期望频率为 $\frac{1}{4}(1-p)^2$（见表 2），故由 $\frac{1}{4}(1-p)^2=\frac{a_4}{n}$ 得 p 值为：

$$\hat{p}=1-2\sqrt{\frac{a_4}{n}} \tag{19}$$

(19) 的单一观察信息量（表 4）为：

$$i_{p_2}=\frac{(1-p)^2}{4-(1-p)^2}+1=\frac{4}{4-(1-p)^2} \tag{20}$$

表 4 (19) 的单一观察信息量

组　别	期望频率 p_j	$\frac{dp_j}{dp}$	$i_j=\frac{1}{p_j}\left(\frac{dp_j}{dp}\right)^2$
1. 非 aabb	$p_1=1-\frac{1}{4}(1-p)^2$	$\frac{2}{4}(1-p)$	$i_1=\frac{4(1-p)^2}{4[4-(1-p)^2]}=\frac{(1-p)^2}{4-(1-p)^2}$
2. aabb	$p_2=\frac{1}{4}(1-p)^2$	$-\frac{2}{4}(1-p)$	$i_2=\frac{4(1-p)^2}{4(1-p)^2}=1$

表 5 相引连锁时，由(13)和(19)估计基因交换率的效率比较

交换率 p	(13) 的 i_{p_1}	(19) 的 i_{p_2}	$RE=\frac{i_{p_1}}{i_{p_2}}$
0.5	1.7778	1.0667	1.667
0.4	2.2775	1.0989	2.073
0.3	3.1184	1.1396	2.736
0.2	4.7980	1.1905	4.030
0.1	9.8146	1.2539	7.827
0.01	99.8314	1.3245	75.373
0.001	999.8331	1.3324	750.400

根据 (18) 和 (20)，可得不同 p 时 (13) 和 (19) 对于估计 p 的效率比较于表 5。表 5 说明，不论 p 取何值，(19) 的含 p 信息都小于 (13)；在 p 小时则尤为突出。例如，在 $p=0.1$ 时，(13) 从 $n=100$ 的观察中所能提取的 p 的信息，等价于 (19) 从 $n=782.7$ 个个体所得的信息；而 $p=0.001$ 时，(13) 的 $n=100$ 将与 (19) 的 $n=75040$ 等价。(19) 对于相引连锁基因交换率的估计是极其低效的，应予抛弃。