```
b = 59.2590    1.8434
bint = 43.5293    74.9886
       0.8282     2.8587
(r 与 rint 略去)
s = 0.4808    14.8171    0.0014.
```

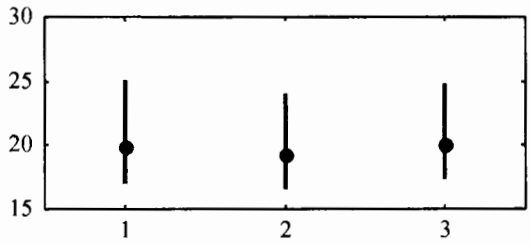

图 5－21

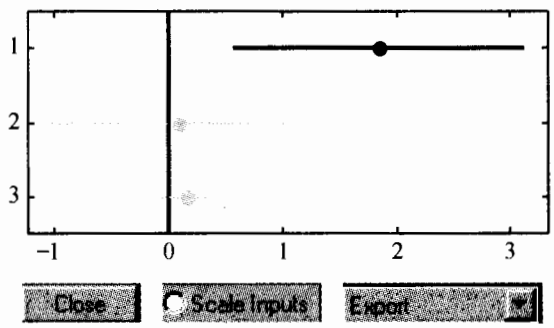

图 5－22

第五节 概率统计模型举例

一、气象观测站的优化

某地区有 12 个气象观测站，为了节省开支，计划减少气象观测站数目，已知该地区 12 个气象观测站的位置(图 5－23)，以及 10 年来各站测得的年降水量(表 5－1).

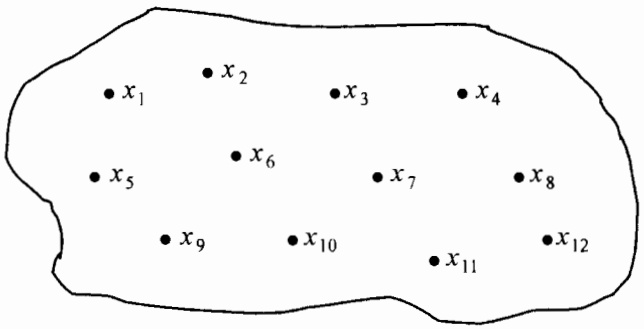

图 5－23　气象观测站分布图

表 5－1　12 个气象观测站测得的年降水量(mm)

年份	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}
1981	276.2	324.5	158.6	412.5	292.8	258.4	334.1	303.2	292.9	243.2	159.7	331.2
1982	251.6	287.3	349.5	297.4	227.8	453.6	321.5	451.0	466.2	307.5	421.1	455.1
1983	192.7	433.2	289.9	366.3	466.2	239.1	357.4	219.7	245.7	411.1	357.0	353.2
1984	246.2	232.4	243.7	372.5	460.4	158.9	298.7	314.5	256.6	327.0	296.5	423.0
1985	291.7	311.0	502.4	254.0	245.6	324.8	401.0	266.5	251.3	289.9	255.4	362.1
1986	466.5	158.9	223.5	425.1	251.4	321.0	315.4	317.4	246.2	277.5	304.2	410.7
1987	258.6	327.4	432.1	403.9	256.6	282.9	389.7	413.2	466.5	199.3	282.1	387.6
1988	453.4	365.5	357.6	258.1	278.8	467.2	355.2	228.5	453.6	315.6	456.3	407.2
1989	158.5	271.0	410.2	344.2	250.0	360.7	376.4	179.4	159.2	342.4	331.2	377.7
1990	324.8	406.5	235.7	288.8	192.6	284.9	290.5	343.7	283.4	281.2	243.7	411.1

应减少哪些观测站才能使所得到的降水量的信息保持足够大?

先做一些必要假设:

(1) 相近地域的气象特性具有较大的相似性和相关性,它们之间的影响可近似地看作是一种线性关系;

(2) 该地区的地理特性具有一定的均匀性,而不是表现为复杂多变的地理特性;

(3) 在距离较远的条件下,由于地形、环境等因素而造成不同区域年降水量相似的可能性很小,不同区域年降水量的差异主要与距离有关;

(4) 不考虑其他区域对本地区的影响.

其中,前两个假设是建模的关键,也是构成模型的前提和基础.

减少观测站的个数，必将损失得到的信息量，但却由此可节约开支，因而最优的结果应该是站数较少的情况下，仍能得到足够大的信息量，在站数和信息量这两个相互制约的因素之间，应主要考虑信息量，因为信息量减少到一定的程度，就会使气象观测失去意义，因而问题的关键就在于怎样减少观测站个数，使信息量不小于一定值.

信息量是一个比较模糊的概念，要保证一定的信息量就必须弄清楚如何利用观测数据分析得到气象信息.气象部门的数据主要用于预报，因而，分析数据的变化规律便成为问题的关键所在.影响气象的因素很多，在气象观测中，一般应比较全面地观测各种因素，从而汇总出具有一定特点、一定代表性的观测站的数据.大气系统由于其自身的规律，地理位置上相近的区域在气象上往往具有一定的相似性，各气象因素之间往往存在一些客观联系，根据这些知识，去掉的观测站应使剩下观测点能在一定程度上反映其原有信息量.于是我们可以在原始数据中将能反映同一规律，即相关性、相似性好的几个站中去掉多余的部分，使剩下的站可以反映它们的共同特点，保留原始数据中与其他站联系不大的站.保留下来的站中一个观测点的观测值实际上是作为相似区域或相近区域的代表值而使用的.因此，除考虑观测站的特色外，还应该注意到一个观测站所代表的区域大小.因为去掉的站是相关性好的，所以去掉的站可以用剩下的站来表示，而且误差较小.

气象部门用观测站测得的实验数据来估计一个地区内降水量的分布，通常用观测到的降水量数据求一种结构函数 $b(L)$.结构函数 $b(L)$反映了该区域降雨量随地区分布的基本规律，从理论上讲，结构函数 $b(L)$是指地面上任意距离为 L 的两点的降水量之差的平均值.实践中因为只给出 n 个站的数据($n=12$)，故近似求 $b(L)$的方法是：求出任意两个站间的距离及相应的两站间的降水量之差，得到 C_n^2 个距离值 $l_i(i=1,2,\cdots,C_n^2)$及相应的降水量之差的绝对值 $f_i(i=1,2,\cdots,C_n^2)$.

令
$$a=\min_{i=1,2,\cdots,C_n^2}(l_i),b=\max_{i=1,2,\cdots,C_n^2}(l_i),$$
将区间$[a,b]m$ 等分，设落在第 k 段小区间$[a+(k-1)(b-a)/m,a+k(b-a)/m]$内的 l_i 值共有 j 个，记为 $l_{i1},l_{i2},\cdots,l_{ij}$.则在各区间中点 $L_k=a+\left(k-\frac{1}{2}\right)\frac{b-a}{m}$处，结构函数的值为

$$b(L_k)=\frac{f_{k1}+f_{k2}+\cdots+f_{kj}}{j}=\frac{\sum_{i=1}^{j}f_{ki}}{j},(k=1,2,\cdots,m),$$

用折线将 $b(L_k)$的值连起来，即得到连续的 $b(L)$曲线.对于给出的数据，可先

求得两个站每年的降水量之差，再求 f_i 的值. 可取 $m=8$，把 L 分成[10,20]，[20,30]，[30,40]，[40,50]，[50,60]，[60,70]，[70,80]，[80,90]8 个区间，求得每个区间中点的 $b(L_k)$值(表 5－2).

由表 5－2 可以看出，除了有两个点例外，其余点的 $b(L)$值大体是呈递增趋势，即距离越远降水量之差越大. 为了更好地反映这种单调递增的趋势，可引入修正的结构函数 $b'(L)$

$$b'(L_k)=\frac{\sum_{j\geqslant k} b(L_j)}{9-k}$$

于是得表 5－3 及 $b(L_k)$，$b'(L_k)$与 L_k 的曲线(图 5－24).

表 5－2　$b(L_k)$的值

L_k	15	25	35	45	55	65	75	85
$b(L_k)$	23.61	25.46	19.46	26.45	31.35	24.11	43.43	55.53

表 5－3　$b'(L_k)$的值

L_k	15	25	35	45	55	65	75	85
$b'(L_k)$	31.16	32.24	33.37	36.15	38.58	41.02	49.48	55.53

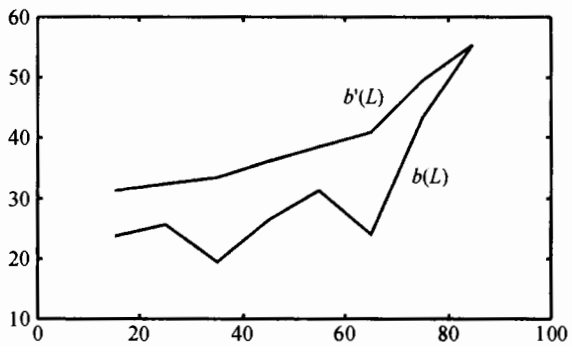

图 5－24　$b(L)$与 $b'(L)$变化趋势图

假如用一点估计另一点的误差允许范围为 10%，由数据得该地区年平均降水量为 320 mm，故用一点作为另一点近似值时，降水量之差最大应为 $\Delta=320\times 10\%=32$ mm. 在 $b'(L)$曲线上，当 $b'(L)=32$ mm 时，$L_0=22.78$ mm. 于是用一点估计另一点时，这两点的距离就不能大于 L_0，否则其误差就会超过 10%，于是问题变为剩余多少个站才足以保证该地区内任一点的降水量都可以用最近站的降水量来估计.

不妨将该地区近似为正方形，记为 Q，且观测站均匀分布．以 n 个站为中心把 Q 分成 n 个面积为 d^2 的小方格，若 A 为该区域内任一点，并设 $l_{\min}$ 为与 A 相距最近的站到 A 的距离，则必有

$$\max_{A\in Q} l_{\min}(A)\leqslant L_0=22.78,$$

易得

$$\max_{A\in Q} l_{\min}(A)=\frac{\sqrt{2}}{2}d,$$

即

$$\frac{\sqrt{2}}{2}d\leqslant L_0=22.78,\ \text{故}\ d\leqslant 32.21.$$

计算得该区域面积 $S=8\ 000$ 单位，每个小方格面积为 d^2，故应剩余的站数为

$$n=\frac{S}{d_{\max}^2}=\frac{8\ 000}{32.21^2}=7.7\approx 8,$$

即保留 8 个站时，仍能保证该地区内任意一点都有与之足够靠近的一个或多个站可以预报该点的降水量，因此这 8 个观测站所提供的数据足以为该地区提供足够的信息量．

为了确定从 12 个站中去掉哪 4 个站，且仍能使信息量尽可能多地保留下来，可用如下算法：

第一步　对每个站确定一集合 $A_i=\{x_j\mid d(x_j,x_i)\leqslant d_0\}$，即与 x_i 距离小于 d_0 的所有 x_j 的集合，取 $d_0=31.5$；

第二步　对每个集合 $A_i(i=1,2,3,\cdots,n)$，以 x_i 为因变量，其他所有 $x_j\in A_i$ 为自变量作多元线性回归，得回归方程 $e(i)$ $(i=1,2,3,\cdots,n)$，并求出每个方程的残差平方和 $S_{残}$ 及回归方程的显著性检验值 F_i；

第三步　显著水平取 0.1，对回归方程 $e(i)$ 检验，若 $F_i>F_a$，则表示 $e(i)$ 有显著的线性关系，对线性关系不显著的 $e(k)$，检验 A_k 中的每一个自变量的显著性，剔除不显著的自变量后得新的回归方程 $e(k)$，直到所有方程均显著或 A_k 为空集为止；

第四步　对有显著线性关系的回归方程 $e(i)$，判断应去掉哪个变量（即观测站），设 x_j 在 $e(i)$ 中的系数为 β_{ij}，则对每一变量定义权重 W_j

$$W_j=\sum_{i=1}^{k}\frac{|\beta_{ij}|}{\sum_{r=1}^{12}|\beta_{ir}|}\ (k\ \text{为回归方程的数目}),$$

选出权重最大的变量 x_l，同时考虑这个站若一开始认为是不应该去掉的，则选权重次之的变量，这样的变量 x_l 就是应该去掉的站；

第五步　去掉 x_l 后，利用剩余变量重新定义 A_i，并回到第一步，直到取掉四个

变量(站)为止.

对此可作如下分析：

(1) 当两站距离增大时,两点间相互影响变弱,模型中以 d_0 为界限,考虑一个站与其它站影响时,仅考虑与之距离小于 d_0 的站,将会大大简化运算.

(2) 在决定去掉哪个站时,用步骤四中的权重 W.考虑这样一个事实,设回归方程为 $y=\sum \beta_i x_i+\beta_0$,$|\beta_i|$愈小,反映了 x_i 对 y 的作用影响愈小.若 $\beta_i=0$,说明 x_i 与 y 无关,若删去 x_i,则 x_i 所代表的信息不能用其它 x 替代,故应保留 x_i.

(3) 对去掉的变量,因保留了与之相关性好的其它变量,故仍可用剩余变量估计它.

(4) 第三步的剔除变量的处理,是因为集合 A 中的元素虽然距离小于等于 d_0,但相关性未必好,因此用 A_i 的所有元素线性回归时,仍可能存在与 x_i 相关性很差的变量,使显著性下降,因此去掉 A_i 中与 x_i 相关性差的变量以提高显著性.

通过以上计算,可知应去掉 x_6,x_7,x_{10}和 x_{12}四个观测站,去掉四个站前后各站管辖区域如图 5－25 所示

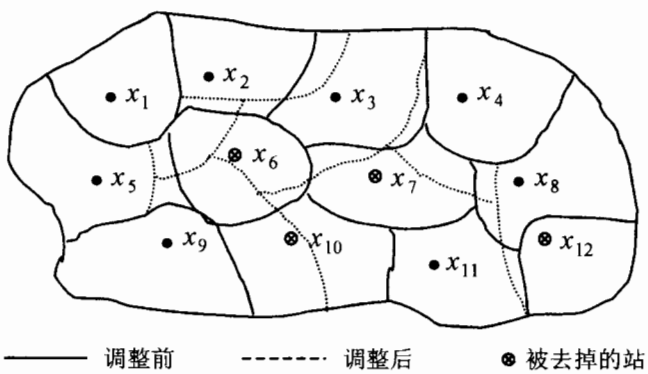

图 5－25　站点分布及管辖区域图

再用去掉四个站后剩余的站将降水量估计出来,用最小二乘法求得：

$x_{10}=219.2+0.178x_2-0.07x_3-0.137x_4+0.199x_5-0.372x_9+0.489x_{11}$,

残差平方和为 1 248.2,显著性检验值为 $F=11.315>F(6,3)=5.28$;

$x_7=86.95+0.136x_2+0.363\ x_3+0.288\ x_4-0.001x_5$,

残差平方和为 2 388.8,显著性检验值为 $F=5.568>F(4,5)=3.52$;

$x_{12}=208.3+0.416x_8-0.188\ 9\ x_9+0.373\ x_{11}$，

残差平方和为 2737.8，显著性检验值为 $F=6.918>F(3,6)=3.29$；

$x_6=302.9-0.1x_4-0.641x_5+0.029\ x_9+0.726\ x_{11}$，

残差平方和为 9 386.7，显著性检验值为 $F=9.36>F(4,5)=3.52$.

对于这个结果可以粗略验证如下

根据给出的数据，求两两相关系数，有 3 个站与 x_{12} 相关系数大于 0.5；有 5 个站与 x_{10} 相关系数大于 0.4；有 3 个站与 x_6 相关系数大于 0.4；x_7 与 x_3 的相关系数为 0.83. 因此，这四个站与其它站的相关系数是很高的. 各年降水量的标准差见表 5－4

表 5－4 各年降水量的标准差

站 x	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}
标准差	95.1	76.8	102.7	60.7	89.3	89.4	36.1	80.7	103.8	54.3	82.1	34.9

对于观测站来讲，标准差大，包含的信息量也大，对 12 个站的标准差排序，发现 x_6，x_7，x_{10} 和 x_{12} 分别排在第 9，2，3 和 1 位，即它们的标准差相对较小，因而所包含的信息也较少，算得的结果是令人满意的.

二、人类 ABO 血型的交配类型及其后代频率

一般认为，人类的 ABO 血型是由三个等位基因所决定的：A(产生抗原 A)对 a(不能产生任何抗原)为显性；另一等位基因 a′(产生抗原 B)对 a 也是显性；但 A 与 a′ 间没有显性(等显性)，并且各自产生其抗原，从而导致杂合子 Aa′ 同时具有两种抗原，因而在人类群体中有四种可辨认的表型(血型). 设 A，a′ 和 a 的基因频率分别为 p，q，r，则基因型频率见表 5－5

表 5－5 人类 ABO 血型的频率

基因型	Aa′	AA，Aa	a′a′，a′a	aa
血型	AB	A	B	O
频率	$2pq$	p^2+2pr	q^2+2qr	r^2

例如，设 $p=0.22$，$q=0.16$，$r=0.62$，则 AB 型所占比例为 7.04%，A 型占 32.12%，B 型占 22.40%，O 型占 38.44%. 其中 A 型中纯合体为 4.84%，杂合体为 27.28%；而 B 型中纯合体为 2.56%，杂合体为 19.84%. 可见，在人类群体中，多数的 A 和 B 都是杂合的. 有人曾调查了 6 000 个中国人的血型频率，结果见表 5－6

表 5－6　中国人的血型频率

血型	AB	A	B	O	合 计
人数	607	1 920	1 627	1 846	6 000
频率	0.101 16	0.320 00	0.271 16	0.307 66	1.000 00

由表 5－5 可知

$$A+O=(p+r)^2,\quad B+O=(q+r)^2,\quad O=r^2,$$

于是可得

$$p=1-\sqrt{B+O},q=1-\sqrt{A+O},r=\sqrt{O},$$

根据这个公式，我们便可以利用 O，A，B 和 AB 血型的频率来求基因频率 p，q，r. 从理论上来讲，$p+q+r=1$，但在实际情况中，根据抽样估计出的基因频率之和并不等于 1，虽然在大群体中它们往往很接近于这个数. 令 $d=1-(p+q+r)$，则 d 的方差为

$$D(d)=\frac{pq}{2G(1-p)(1-q)},$$

其中 G 为样本中的个体数，利用这个方差可以检验 $p+q+r$ 距离 1 的离差的显著性. 为了使这三个基因频率之和为 1，Bernstein 对这三个基因频率进行了修正，修正后的频率用 p'，q' 和 r' 表示，则有：

$$p'=p(1+d/2),q'=q(1+d/2),r'=(r+d/2)(1+d/2),$$

注意到 $p+q+r+d=1$，即 $p+q+r+d/2=1-d/2$，则修正后三个基因频率之和为

$$p'+q'+r'=(1+d/2)(1-d/2)=1-d^2/4,$$

若 d 值很小，则这个值非常接近于 1. 如果所求的和与 1 还有相当的差距，则可重复上述步骤进行进一步的修正，但一般只需修正一次就足够精确了. 如对表 5－6 有：$p=0.239\ 19$，$q=0.207\ 75$，$r=0.554\ 68$，$p+q+r=1.001\ 62$；修正后 $p'=0.239\ 00$，$q'=0.207\ 58$，$r'=0.553\ 42$，$p'+q'+r'=1.000\ 00$. 可见这种修正方法是简单实用的，但其惟一的缺点是难于测定所修正的基因频率的准确方差.

为了估计基因频率的方差，用最大似然估计法是必要的. 设 (AB) 为 AB 型个体的观察数，并将 $(AB)+(A)$ 写为 $(AB+A)$，在忽视不牵涉基因频率的项目后，对数似然函数为

$$L=(AB)\ln(pq)+(A)\ln(p^2+2pr)+(B)\ln(q^2+2qr)+(O)\ln r^2,$$

整理得

$L=(AB+A)\ln p+(AB+B)\ln q+(A)\ln(p+2r)+(B)\ln(q+2r)+2(O)\ln r$，将 p 和 q 作为两个独立的待估参数，$r=1-p-q$，对 p 和 q 求偏导数有

$$\begin{cases}\dfrac{\partial L}{\partial p}=\dfrac{(AB+A)}{p}-\dfrac{(A)}{p+2r}-\dfrac{2(B)}{q+2r}-\dfrac{2(O)}{r},\\ \dfrac{\partial L}{\partial q}=\dfrac{(AB+B)}{q}-\dfrac{2(A)}{p+2r}-\dfrac{(B)}{q+2r}-\dfrac{2(O)}{r},\end{cases}$$

p 和 q 的估计应使：$\dfrac{\partial L}{\partial p}=0,\dfrac{\partial L}{\partial q}=0$. 这是一个非线性代数方程组，一般用迭代法求解. 可求信息矩阵

$$I=\begin{bmatrix}I_{pp} & I_{pq}\\ I_{pq} & I_{qq}\end{bmatrix},$$

其中 $I_{pp}=-\dfrac{\partial^2 L}{\partial p^2},I_{qq}=-\dfrac{\partial^2 L}{\partial q^2},I_{pq}=-\dfrac{\partial^2 L}{\partial p\partial q}$，则

$$\begin{cases}I_{pp}=\dfrac{(AB+A)}{p^2}+\dfrac{(A)}{(p+2r)^2}+\dfrac{4(B)}{(q+2r)^2}+\dfrac{2(O)}{r^2},\\ I_{qq}=\dfrac{(AB+B)}{q^2}+\dfrac{4(A)}{(p+2r)^2}+\dfrac{(B)}{(q+2r)^2}+\dfrac{2(O)}{r^2},\\ I_{pq}=\dfrac{2(A)}{(p+2r)^2}+\dfrac{2(B)}{(q+2r)^2}+\dfrac{2(O)}{r^2},\end{cases}$$

则 $\boldsymbol{I}^{-1}$ 为 p 和 q 的协方差阵，$\boldsymbol{I}^{-1}$ 的四个元素之和为 r 的方差. 设

$$\boldsymbol{I}^{-1}=\boldsymbol{V}=\begin{bmatrix}V_{pp} & V_{pq}\\ V_{pq} & V_{qq}\end{bmatrix},C_p=\frac{\partial L}{\partial p},C_q=\frac{\partial L}{\partial q},$$

则 $\delta_p=C_pV_{pp}+C_qV_{pq},\delta_q=C_pV_{pq}+C_qV_{qq},\delta_r=-\delta_p-\delta_q$，

给定初值 p_0，q_0 和 r_0 后，可求出 $\delta_p^{(0)},\delta_q^{(0)}$ 和 $\delta_r^{(0)}$. 令

$$\begin{cases}p_k=p_{k-1}+\delta_p^{(k-1)},\\ q_k=q_{k-1}+\delta_q^{(k-1)},\quad (k=1,2,\cdots),\\ r_k=r_{k-1}+\delta_r^{(k-1)}\end{cases}$$

直到 $\delta_p^{(k)},\delta_q^{(k)}$ 和 $\delta_r^{(k)}$ 均小于给定误差，则最后一次求出的 p,q 和 r 就是它们的最大似然估计，$\boldsymbol{V}$ 就是估计值的方差.

现在可以写出六种基因型的二十一种不同类型的交配及它们相应的后代基因型频率(表 5-7，表中将它们集中为十个可辨别的表型交配，并将六种基因型后代集中为四种血型).

把十六种交配按照一个亲本的表型合并成四群，就可以得到 ABO 血型的亲子组合频率(表 5-8).

表 5－7　ABO 血型交配类型及其后代的频率

交配类型	交配频率	后代			
		O	A	B	AB
O×O	r^4	r^4			
O×A	$2pr^2(p+2r)$	$2pr^3$	$2pr^2(p+r)$		
O×B	$2qr^2(q+2r)$	$2qr^3$		$2qr^2(q+r)$	
O×AB	$4pqr^2$		$2pqr^2$	$2pqr^2$	
A×A	$p^2(p+2r)^2$	p^2r^2	$p^2(p+r)(p+3r)$		
A×B	$2pq(p+2r)(q+2r)$	$2pqr^2$	$2pqr(p+r)$	$2pqr(q+r)$	$2pq(p+r)(q+r)$
B×B	$q^2(q+2r)^2$	q^2r^2		$q^2(q+r)(q+3r)$	
A×AB	$4p^2q(p+2r)$		$2p^2q(p+2r)$	$2p^2qr$	$2p^2q(p+r)$
B×AB	$4pq^2(q+2r)$		$2pq^2r$	$2pq^2(q+2r)$	$2pq^2(q+r)$
AB×AB	$4p^2q^2$		p^2q^2	p^2q^2	$2p^2q^2$
总计	1.00	r^2	p^2+2pr	q^2+2qr	$2pq$

表 5－8　ABO 血型的亲子组合频率

母亲	孩子			
	AB	A	B	O
AB	$pq(p+q)$	$pq(p+r)$	$pq(q+r)$	
A	$pq(p+r)$	$p(p^2+3pr+r^2)$	pqr	pr^2
B	$pq(q+r)$	pqr	$q(q^2+3qr+r^2)$	qr^2
O		pr^2	qr^2	r^3