附录 数理统计和线性代数基础

第一节 概率的基本定理

一、概率加法定理

若 2 个互斥事件 A 与 B 在 N 次试验中各出现了 n_A 与 n_B 次，那么它们的和事件 C（记为 $A+B$）在试验中出现了(n_A+n_B)次，这两事件的和 C 的概率为

$$P(C)=\frac{n_A+n_B}{N}=\frac{n_A}{N}+\frac{n_B}{N}=P(A)+P(B) \qquad \text{[A-1]}$$

因此 2 个互斥事件之和的概率等于 2 个事件的概率之和，称之为概率加法定理。加法定理还可推广到 n 个两两互斥事件，即 n 个互斥事件和的概率等于 n 个互斥事件概率之和。

例如，2 个纯合亲本 P1 和 P2 的基因型分别为 AA 和 aa，杂种 F_1 的基因型为 Aa，杂种 F_2 种 3 种基因型 AA、Aa 和 aa 的概率分别为 $\frac{1}{4}$ 、$\frac{1}{2}$ 和 $\frac{1}{4}$，若 A 对 a 呈完全显性，求从 F_2 群体中随机抽取一个个体表现为显性性状的概率为多少？因为个体的基因型为 AA 和 Aa 是两互斥事件，又 A 对 a 呈完全显性，因此显性性状为 AA 和 Aa 这 2 种基因型之和，所以随机抽取一株具有显性性状的概率 P(AA+Aa)=P(AA)+P(Aa)= $\frac{1}{4}+\frac{1}{2}=\frac{3}{4}$ 。

二、条件概率和概率乘法定理

对于事件 A 和 B，A（或 B）事件的概率不受 B（或 A）事件发生与否的影响，则称 A、B 二事件独立，否则称该二事件不独立。例如，田间有 20 株表现基本一致的小麦植株，其中 18 株结红粒，2 株结白粒，设甲和乙二人“从中任抽一株恰为白粒植株”的事件分别为 A 和 B。分两种情况讨论 2 个事件的概率，一种是甲抽取放回后，乙再抽取，显然，这时事件 A 和事件 B 发生的概率应相等，即 $P(A)=P(B)=\frac{2}{20}=\frac{1}{10}$，因为和甲一样，乙仍是从同样的 20 株小麦

中抽取，这时事件 *A* 的发生并不影响事件 *B* 的发生，因此 *A* 和 *B* 是独立的；另一种情况是，甲抽到后不放回乙就抽取，这时由于事件 *A* 发生，田间还剩下 19 株，其中结白粒的只有 1 株了，这时“乙从中任抽一株恰为白粒植株”的概率就等于 $\frac{1}{19}$，也就是说，事件 *A* 的发生影响了事件 *B* 的发生时，称这两事件不独立。

后一种情况实际上是在事件 *A* 已经发生的条件下，再来考查事件 *B* 的概率的。我们把在事件 *A* 已经发生的条件下事件 *B* 发生的概率称为条件概率，记为 *P*（*B*|*A*）。对于任意两个事件（不管独立与否）*A*、*B*，它们同时出现的概率 *P*(*AB*)等于 *A* 的概率 *P*(*A*)乘以在 *A* 已经发生的条件下 *B* 之概率 *P*（*B*|*A*）；或等于 *B* 的概率 *P*(*B*)乘以在 *B* 已经发生的条件下 *A* 之概率 *P*（*A*|*B*），即：

$$P(AB)=P(A)P(B|A)=P(B)P(A|B) \qquad \text{[A-2]}$$

由[A-2]可得条件概率为：

$$P(B\mid A)=\frac{P(AB)}{P(A)},\quad P(A\mid B)=\frac{P(AB)}{P(B)}$$

例如，一随机交配群体中 3 种基因型 AA、Aa 和 aa 的频率分别为 P、H 和 Q，求交配为 Aa×Aa 且子代基因型为 AA 的概率为多少？设交配类型 Aa×Aa 为事件 A，产生的 AA 子代基因型为事件 B，而 $P(A)=P(Aa)\times P(Aa)=H^2$，$P(B|A)=\frac{1}{4}$，根据 [A-2] 有 $P(AB)=P(A)P(B|A)=\frac{1}{4}H^2$。

若事件 *A* 与事件 *B* 相互独立，我们有 *P*(*B*|*A*)=*P*(*B*)或 *P*(*A*|*B*)=*P*(*A*)，则这二事件同时发生的概率 *P*(*AB*)等于事件 *A* 的概率 *P*(*A*)与事件 *B* 的概率 *P*(*B*)之乘积，即

$$P(AB)=P(A)P(B) \qquad \text{[A-3]}$$

第二节 离散型和连续型随机变量

用 X 表示随机现象的各种结果，X 取不同的数值就表示不同的事件发生，但是 X 究竟取什么值事先是不知道的，这样的变量称为随机变量。随机变量的取值都有确定的概率，根据随机变量的取值情况，可以把它分为两类：如果它的取值是有限个或可数个（可以按一定顺序一一列举出来），则称为离散型随机变量；如果它的取值为无穷不可数个，则称为连续型随机变量。

一、离散型随机变量的概率分布

离散型随机变量 X 只能取有限个或可数个值，设它的可能取值是 x_1，x_2，…，x_k，…，为了完全描述随机变量 X，只知道它的可能取值是远远不够的，更重要的是要知道它取各个值的概率，也就是要知道 $P(X=x_1)$，$P(X=x_2)$，…。设

$$P(X=x_k)=p_k \quad (k=1，2，\cdots) \qquad [A\text{-}4]$$

则 X 的可能取值及相应概率可列成下表，

表 A-1 离散型分布的概率分布表

X	x_1	x_2	…	x_k	…
P	p_1	p_2	…	p_k	…

表 A-1 称为随机变量 X 的概率分布表，它清楚而完整地表示了 X 取值的概率分布情况。离散型随机变量 X 有两个基本性质：

（1）$p_k \geq 0$ (k=1，2，…)；

（2）$\sum_{k=1}^{\infty} p_k = 1$

对于（1），这是显然的，因为任何概率都具有非负性。对于（2），因为随机变量 X 取遍所有可能的值时，就得到所有的基本事件，这些基本事件的和是一个必然事件。常见的两

个概率分布（离散型）是二项分布和泊松（Poisson）分布。

（1）二项分布。如果随机变量 X 的分布如下，

$$P（X=k）=C_n^k p^k q^{n-k} \quad (k=0，1，2，\cdots，n，0<p<1，q=1-p)$$

则称 X 服从二项分布（Binomial distribution）。这个分布也满足概率分布的两个基本性质：

（1）$P（X=k）=p_k \geq 0$；

（2）$\sum_{k=0}^{n} p_k = \sum_{k=0}^{n} C_n^k p^k q^{n-k} = (p+q)^n = 1$

如果一个试验是由 n 次独立试验构成的（n 次独立试验是指各次试验的结果相互独立），而单次试验中事件 A 发生的概率为 p ($0<p<1$)，不发生的概率为 q ($q=1-p$)，则在 n 次独立试验中事件 A 发生 k 次的概率为

$$P（A\text{ 发生 }k\text{ 次}）=C_n^k p^k q^{n-k} \quad (k=0，1，2，\cdots，n，0<p<1，q=1-p)$$

如果用 X 表示在 n 次独立试验中事件 A 发生的次数，则随机变量 X 服从二项分布。再如有一对等位基因 A 和 a，A 对 a 表现为显性，在亲本 AA 和 aa 杂交产生的 F_2 群体中将产生两种表现型，用 A-（由 AA 和 Aa 产生）和 aa 表示，频率分别为 0.75 和 0.25，在容量为 n 的一个 F_2 群体中，表现型 A-的个数服从二项分布。$n=10$ 时，表现型 A-的个体数的概率分布如图 A-1 所示。

（2）泊松（Poisson）分布。如果随机变量 X 的分布如下，

$$P（X=k）=\frac{\lambda^k}{k!}e^{-\lambda} \qquad (k=0，1，2，\cdots，\lambda>0)$$

则称 X 服从泊松分布。图 A-2 给出 $\lambda=1$ 时泊松分布的概率分布。

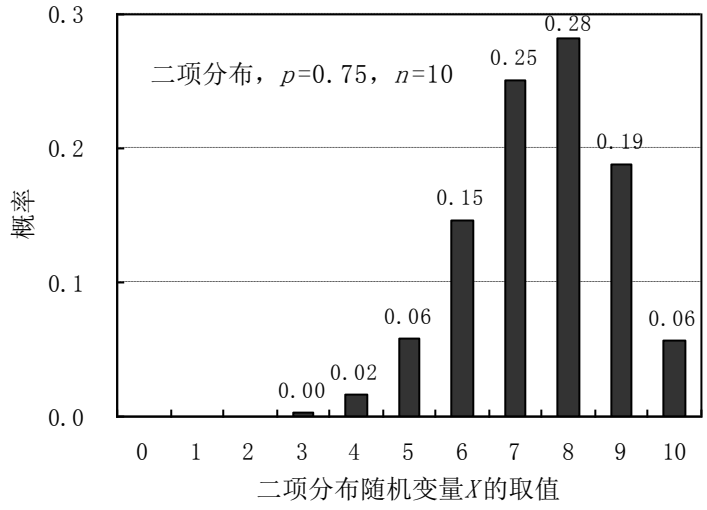

图 A-1 *p*=0.75，*n*=10 时二项分布的概率分布

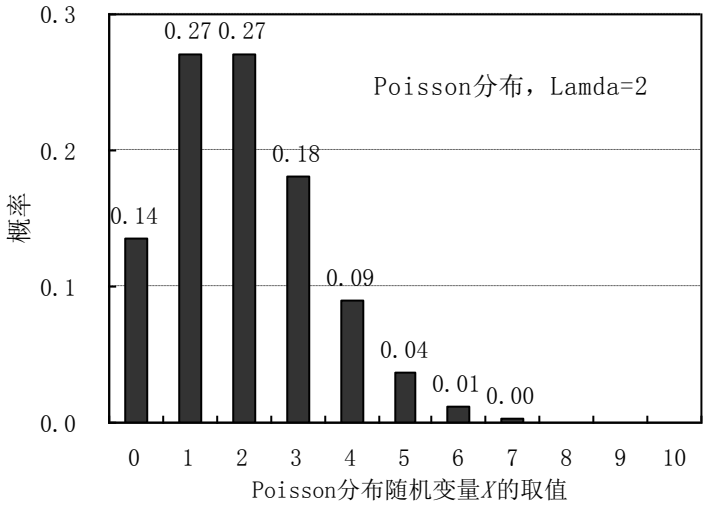

图 A-2 λ=2 时泊松分布的概率分布

二、连续型随机变量

研究随机变量，主要是了解它取值的统计规律。所谓随机变量 X 的统计规律，指的是它在各种各样的范围内取值的概率。例如 $P(x<a)$， $P(x\geq b)$， $P(a<x<b)$ 等等，这里 a，b 是任意实数。对于随机变量，若存在非负可积函数 $p(x)$ $(-\infty<x<+\infty)$，对于任意的 a，b（$a<b$）都有

$$P(a<x<b)=\int_a^b p(x)dx \qquad [A\text{-}5]$$

则称X为连续型随机变量，$p(x)$为随机变量X的分布密度，又称概率密度（Probability density）。作为分布密度$p(x)$，由定义不难推知它有以下基本性质：

（1） $p(x)\geq 0$；

（2） $\int_{-\infty}^{+\infty}p(x)dx=1$；

（3）对任意随机变量的取值a，$P(x=a)=\int_a^a P(x)dx=0$；而对于任意的$x_1<x_2$，有$P(x_1<x<x_2)=P(x_1\leq x<x_2)=P(x_1<x\leq x_2)=P(x_1\leq x\leq x_2)$。

下面介绍两个常见的连续分布。

（1）均匀分布。若随机变量X的分布密度为

$$p(x)=\begin{cases}\frac{1}{b-a}, & 当a\leq x\leq b时；\\ 0, & 其它。\end{cases}$$

则称X服从[a，b]区间上的均匀分布（Uniform distribution）。对于任意满足$a\leq c<d\leq b$的c和d，有

$$P(c<X<d)=\int_c^d p(x)dx=\frac{1}{b-a}\int_c^d dx=\frac{1}{b-a}(d-c)$$

上式表明X取值于[a,b]中任一小区间的概率与该区间的长度成正比，而与该区间的位置无关。在数值计算中，如果用X表示由于四舍五入小数点后第一位小数所引起的误差，则X可以看作是一个服从[-0.5，0.5]区间上的均匀分布的随机变量。又如在[a，b]区间中随机投点，如果用X表示点的坐标，则X也可以看作是一个服从[a，b]区间上的均匀分布的随机变量。

（2）正态分布。若随机变量X的分布密度为

$$p(x)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}}\qquad(-\infty<x<+\infty,\ \sigma^2>0)\qquad\text{[A-6]}$$

则称 X 服从正态分布 $N(\mu,\sigma^2)$，简记为 $X\sim N(\mu,\sigma^2)$，μ 称为正态分布的均值，σ^2 称为正态分布的方差。均值为 0、方差为 1 的正态分布 N（0，1）又称为标准正态分布。实际中，服从正态分布的随机变量是非常多的，例如测量误差、植物的高度、动物的体重、人的身长、健康人的红血球数目、年降水量、月平均温度、海洋的波浪高度等等都服从正态分布，在概率论和数理统计的理论研究和实际应用中正态随机变量起着特别重要的作用。

曲线 $p(x)$ 有以下特点：

（a）曲线呈钟形，以 x 轴为渐近线（$x\to\pm\infty$ 时）（图 A-3A）；

（b）曲线关于直线 $x=\mu$ 对称（图 A-3A）；

（c）$x=\mu$ 时曲线达到最高点，$x=\mu\pm\sigma$ 处有拐点（图 A-3A）；

（d）正态分布的密度曲线与 x 轴之间的总面积等于 1，而且曲线下介于 $x=x_1$ 到 $x=x_2$（$x_1<x_2$）之间的面积等于随机变量落入区间（x_1，x_2）的概率；

（e）带有任意参数 μ，σ^2 的一个正态随机变量总可以通过变量替换 $Y=\dfrac{X-\mu}{\sigma}$，使之变为分布密度为 $\dfrac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}$ $(-\infty<x<+\infty)$ 的随机变量来研究，即标准正态分布。X 的值落在区间 $(-\infty,x]$ 上的概率刚好等于 Y 的值落在区间（－∞，$\dfrac{x-\mu}{\sigma}$］上的概率。

由（e）可以看出，正态随机变量 Y 的参数是 μ＝0，σ^2＝1，它是最简单的正态随机变量。根据这个特点，只要借助于标准正态分布，带有任意参数 μ，σ^2 的正态分布随机变量落入某个范围的概率就都可以计算出来。（e）中的变换 $Y=\dfrac{X-\mu}{\sigma}$ 在数据的标准化转换中有重要的应用。

三、连续型随机变量的分布函数

如果 X 为一连续型随机变量，其分布密度为 $p(x)$ $(-\infty<x<\infty)$，则分布函数为

$$F(x) = P(X \le x) = \int_{-\infty}^{x} p(t)dt$$

$F(x)$的数值等于概率密度曲线下区间$(-\infty, x]$上的面积。图 A-3B 给出几个正态分布的概率分布函数曲线。分布函数在假设检验中有着广泛的应用。

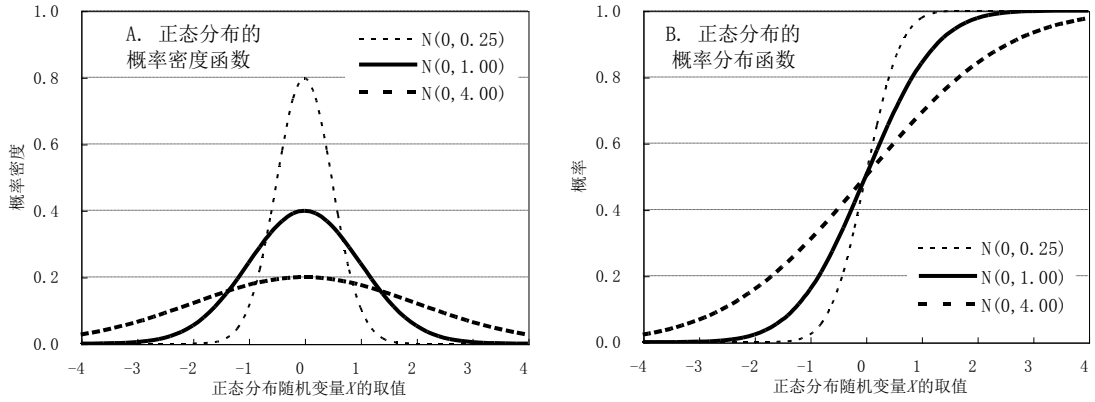

图 A-3 正态分布的概率密度和概率分布函数

四、随机变量的均值和方差

随机变量的概率分布完整地描述了随机变量的取值规律，而且往往依赖于少数的几个参数，这些参数称为随机变量的数字特征，于是确定了随机变量的数字特征，也就确定随机变量的分布。

（1）均值。随机变量的均值体现了随机变量取值的平均大小，均值又称数学期望或简称期望（Expectation），离散型随机变量的均值为

$$E(X) = \sum_{k=1}^{\infty} x_k p_k \qquad [A\text{-}7]$$

连续型随机变量的均值为

$$E(X) = \int_{-\infty}^{+\infty} xp(x)dx \qquad [A\text{-}8]$$

（2）方差。随机变量的取值具有一定的偶然性，但它总是以确定的概率围绕着它的均

值 $E(X)$取值，随机变量 X 与均值 $E(X)$的差 $X-E(X)$或大或小。已经知道随机变量的均值体现了随机变量取值的平均大小，但是只知道均值的大小是不够的，有时还需知道随机变量取值如何在均值周围变化，偏差 $X-E(X)$的大小是能够体现这一点的，它的大小反映了随机变量的取值是比较集中或是比较分散。为了描述这种情况，我们采用偏差平方的均值来衡量，并称它为随机变量 X 的方差（Variance），即

$$V(X)=E[X-E(X)]^2=E(X^2)-[E(X)]^2 \qquad \text{[A-9]}$$

方差大则说明取值分散，方差小则说明取值集中。一些常用分布的均值和方差列于表 A-2。

表 A-2 常用分布的概率函数、均值和方差

分布	概率函数	均值	方差	参数的取值范围
二项分布	$P(X=x)=C_n^x p^x q^{n-x}$，（x=0, 1, 2, …, n）	np	npq	$0<p<1$，q=1-p
泊松分布	$P(X=x)=\frac{\lambda^x}{x!}e^{-\lambda}$ （x=0, 1, 2, …）	λ	λ	λ>0
均匀分布	$p(x)=\frac{1}{b-a}$ （$a\le x\le b$）	$\frac{b+a}{2}$	$\frac{1}{12}(b-a)^2$	$b>a$
正态分布	$p(x)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$	μ	σ^2	μ 任意，$\sigma^2>0$

（3）协方差。随机变量 X 和 Y 间的协方差定义为

$$Cov(X,Y)=E\{[X-E(X)]\times[Y-E(Y)]\}=E(XY)-E(X)E(Y) \qquad \text{[A-10]}$$

如果$Cov(X,Y)=0$，称随机变量 X 和 Y 是相互独立的，此时，$E(XY)=E(X)E(Y)$。

第三节 极大似然估计和统计假设检验

一、极大似然估计的定义

极大似然方法是统计中最重要、应用最广泛的方法之一，该方法最初由德国数学家 Gauss 于 1821 年提出，但未得到重视。R.A. Fisher 在 1922 年再次提出了极大似然的思想并探讨了它的统计性质，从而导致了极大似然方法的广泛研究和应用。

在概率统计中，概率密度函数 $f(X,\theta)$ 是最常用的，θ 为分布参数，例如对于正态分布来说，$\theta=(\mu,\sigma^2)$。当 θ 已知时，$f(X,\theta)$ 反映了密度函数怎样随 X 变化，当固定 X 而把 $f(X,\theta)$ 看成是 θ 的函数时，概率函数又称为似然函数，它反映了 X 对 θ 的解释能力，所以概率函数和似然函数可以说是一回事，只是着眼点不同：前者是固定 θ 而看成是 X 在样本空间上的函数，后者则固定 X 而看成是 θ 的函数。这种差别在统计上的意义如下：若把参数 θ 和样本 X 分别看成是“原因”和“结果”，定义了 θ 的值，就完全确定了样本分布，也就确定了得到种种结果（X）的机会的大小；从另一方面看，当有了结果（样本）X 时，可问“当参数 θ 取各种不同的值（原因）时，导致这个结果（X）的可能性有多大？”对这个问题的回答就引出了似然函数的概念，由于统计推断是由样本推断参数，这种看法就可以作为极大似然统计推断方法的哲理基础，基于每个参数值的“似然性”去进行统计推断这一原则，叫做似然原则（Likelihood principle）。似然原则的一项重要应用是参数极大似然估计方法的提出。

设总体 X 的分布密度函数为 $f(X,\theta)$，θ 为待估计的分布参数，x_1，x_2，…，x_n 是总体 X 的一个随机样本，则样本似然函数定义为

$$L(x_1,x_2,\cdots,x_n;\theta)=\prod_{i=1}^{n}f(x_i,\theta) \qquad \text{[A-11]}$$

若 $\hat{\theta}=\hat{\theta}(x_1,x_2,\cdots,x_n)$ 是一个统计量，满足条件：

$$L(x_1,x_2,\cdots,x_n;\hat{\theta})=\sup\{L(x_1,x_2,\cdots,x_n;\theta)\}$$

即 $L(x_1,x_2,\cdots,x_n;\hat{\theta})$ 是函数 $L(x_1,x_2,\cdots,x_n;\theta)$ 的一个上确界，则称 $\hat{\theta}$ 是 θ 的极大似然估计。

按照上面对“似然性”的解释，θ 的极大似然估计 $\hat{\theta}(X)$ 就是在已得样本 X 的情况下，似然性最大的那个 θ 值。$\hat{\theta}$ 的确定要解一个极值问题，有时求解极值问题很困难，不得不采用数值方法或迭代方法（如 EM 算法，详见混合遗传模型一章）。在 $X=(x_1,\cdots,x_n)$ 为简单随机样本而正态总体分布有概率密度函数 $f(X,\theta)$ 时，似然函数为：

$$L(X;\theta)=\prod_{i=1}^{n} f(x_i,\theta)=\prod_{i=1}^{n}[\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}]$$

$$=\frac{1}{(2\pi\sigma^2)^{n/2}}\exp\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\}$$

$L(X;\theta)$ 的对数（称为对数似然函数）在使用上较方便，极大似然估计可利用对数似然函数等价地定义为：

$$\ln L(X;\hat{\theta}(X))=\sup\{\ln L(X;\theta)\}=\sup\{-\frac{n}{2}\ln(2\pi\sigma^2)-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\}$$

极大似然估计一般通过求解似然函数的偏导并令偏导数等于零来获得，称方程

$$\frac{\partial \ln L(X;\theta)}{\partial\theta}=0$$

为似然方程。当 $f(X,\theta)$ 为正态分布的概率密度函数时，似然方程是：

$$\frac{\partial \ln L}{\partial\mu}=\frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu)=0$$

$$\frac{\partial \ln L}{\partial\sigma^2}=\frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i-\mu)^2-\frac{n}{2\sigma^2}=0$$

由这两个方程求解 μ 和 σ^2，从而有：

$$\hat{\mu}=\frac{1}{n}\sum_{i=1}^{n}x_i,\quad \hat{\sigma}^2=\frac{1}{n}\sum_{i=1}^{n}(x_i-\hat{\mu})^2$$

统计上还可证明 $\hat{\mu}$ 是 μ 的无偏估计，而 $\hat{\sigma}^2$ 是有偏的，即：

$$E(\hat{\mu})=\mu,\quad E(\hat{\sigma}^2)=\frac{n}{n-1}\sigma^2\neq\sigma^2$$

因此在生物统计中，将 $\overline{X}=\frac{1}{n}\sum_{i=1}^{n}x_i$ 作为均值的估计，并称为样本均值；$S^2=V_X=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2$ 作为方差的估计值，被称为样本方差，它是 σ^2 的无偏估计。

二、极大似然估计的统计性质

以下只给出极大似然估计的一些统计性质，具体证明请参照有关数理统计方面的专著。

定理 1 设 $X=(x_1,\cdots,x_n)$ 是来自总体{ $f(X,\theta)$，$\theta\in\Theta$ }的样本，T是 θ 的充分统计量，如果 θ 的极大似然估计存在，则它是 T的函数。

定理 2 如果 Rao-Cramer 不等式中的条件成立， T是 θ 的有效估计，则似然方程具有唯一解 T，同时 T也是极大似然估计。

定理 3 设母体的分布函数为 $f(X,\theta)$，θ 是在某个开区间 Θ 上取值的实参数（单参数情形），且满足一定的条件（略），则：

（1）似然方程当 $n\to\infty$ 时，其概率趋近于 1 存在一致性解；

（2）似然方程的一致性解渐近于正态分布 $N(\theta,\ \frac{1}{nI(\theta)})$。

显然，定理 3 中的结论（2）可用于估计极大似然估计的方差。

三、样本均值、方差和协方差的估算

（1）利用简单随机样本的估计。设 x_1，x_2，…，x_n 是总体 X 的一个简单随机样本，y_1，y_2，…，y_n 是总体 Y 的一个简单随机样本，则总体 X 和总体 Y 的均值和方差的估计值分别为：

$$\overline{X}=\frac{1}{n}\sum_{i=1}^{n}x_i\ ,\ \ V_X=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\overline{X})^2$$

$$\overline{Y}=\frac{1}{n}\sum_{i=1}^{n}y_i\ ,\ \ V_Y=\frac{1}{n-1}\sum_{i=1}^{n}(y_i-\overline{Y})^2$$

总体 X 和总体 Y 间的协方差的估计值为

$$Cov_{XY}=Cov(X,Y)=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\overline{X})(y_i-\overline{Y})$$

方差估计值的平方根又称标准差（Standard deviation），即 $SD_X=\sqrt{V_X}$ 。

（2）利用总体数据的估计。如果 x_1，x_2，…，x_n 包含了总体 X 的所有的取值，取值频率分别为 f_1，f_2，…，f_n，则均值和方差的估计值为：

$$\overline{X}=\sum_{i=1}^{n}f_ix_i\ ,\ \ V_X=\sum_{i=1}^{n}f_i(x_i-\overline{X})^2=\sum_{i=1}^{n}f_ix_i^2-\overline{X}^2 \qquad \text{[A-12]}$$

类似地，如果 x_1，x_2，…，x_n 包含了总体 X 的所有取值，y_1，y_2，…，y_n 包含了总体 Y 的所有取值，频率分别为 f_1，f_2，…，f_n，则 X 和 Y 间协方差的估计值为

$$Cov_{XY}=Cov(X,Y)=\sum_{i=1}^{n}f_i(x_i-\overline{X})(y_i-\overline{Y})=\sum_{i=1}^{n}f_ix_iy_i-\overline{X}\overline{Y} \qquad \text{[A-13]}$$

上述均值和方差的估计值在数量遗传中的应用十分广泛，使用时，一定要区分清楚所采用的数据是一个样本，还是总体本身。

（3）均值方差的估计。均值的方差一般用样本方差除以样本量估计，即，

$$V_{\bar{X}} = \frac{V_X}{n}$$

在尺度检验时，常要估计平均数的方差。

四、似然比检验

极大似然法的一大优点在于对参数模型给出一种统一的检验方法，即似然比检验（Likelihood ratio test，LRT），并且检验统计量在一般条件下有统一的渐近 χ^2 分布，因而实际中得到广泛应用。其应用条件是：若一个模型是另一个模型的特殊情形，则可利用似然比检验来比较这个模型与另一个模型是否有显著差异。假定某一模型(用 H_1 表示)与它的特例模型(用 H_0 表示)相差 k 个独立的限制条件(或相差 k 个可估遗传参数)，那么似然比统计量 λ 渐近服从自由度为 k 的 χ^2 分布，即

$$\lambda = 2[L(X;\hat{\theta}_1) - L(X;\hat{\theta}_0)] \sim \chi^2_{(k)} \qquad \text{[A-14]}$$

其中 $\hat{\theta}_1$ 和 $\hat{\theta}_0$ 分别是 H_1 和 H_0 下模型参数的极大似然估计值，L 是对数似然函数。

五、分布的适合性检验

在主基因和多基因混合遗传分析中，通过 AIC 准则选择一个至几个较低 AIC 值的模型或确定混合分布中成分分布个数后，育种工作者关心的是所选择模型是否能解释所获得的遗传数据，或者说期望分布与观测分布间是否一致，这就需进行进一步的检验，这种检验称为适合性检验（Test of fitness）。适合性检验的常用方法包括均匀性检验、Smirnov 检验和 Kolmogorov 检验，共有五个统计量：U_1^2、U_2^2、U_3^2（均匀性检验）、nW^2（Smirnov 检验）和 D_n（Kolmogorov 检验），前三者是对分布的某一数字特征进行检验，即检验分布的平均数、二阶原点矩和二阶中心矩，后两者是对分布整体进行检验，因此，后两者比前三者的功效更高。

（1）均匀性检验。设 $F(x)$为概率分布函数，x_1，x_2，…，x_n 为样本观测值。若 $F(x)$与总体分布间无显著差异，则 $F(x)$是[0，1]上的均匀分布，利用自由度均为 1 的 χ^2 统计量U_1^2、U_2^2

和U_3^2分别检验 $F(x_i)$的平均数是否等于$\frac{1}{2}$，二阶原点矩是否等于$\frac{1}{3}$，二阶中心矩是否等于$\frac{1}{12}$，以达到检验 $F(x)$是否是均匀分布的目的：

$$U_1^2=\frac{12}{n}[\sum F(x_i)-\frac{n}{2}]^2 \sim \chi_{k=1}^2$$

$$U_2^2=\frac{45}{4n}[\sum F^2(x_i)-\frac{n}{3}]^2 \sim \chi_{k=1}^2$$

$$U_3^2=\frac{180}{n}\{\sum[F(x_i)-\frac{1}{2}]^2-\frac{n}{12}\}^2 \sim \chi_{k=1}^2$$

其中 $F(x)$是 $P(x<x_i)$，即 $x<x_i$ 的概率；若总体为 k 个分布的混合，则为$\sum_{t=1}^{k}\pi_t P_t(x<x_i)$。

（2）Smirnov 检验。记 $F_n^*(x)$ 为经验分布函数，按观测值数值大小顺序排成 $x_{(1)}$，$x_{(2)}$，…，$x_{(n)}$，称其为顺序统计量；$F_0(x)$ 为通过模型得到的总体分布，即期望分布。Smirnov（1938）提出利用统计量

$$nW^2=n\int_{-\infty}^{+\infty}[F_n(x)-F_0(x)]^2 dF_0(x)=\frac{1}{12n}+\sum[F(x_{(r)})-\frac{r-0.5}{n}]^2$$

作适合性检验，并证明了 nW^2 的极限分布。Marshall（1958）证明了 nW^2 达到它的极限分布的速度非常快。当 n=3 时，nW^2 已接近它的极限分布。一些显著水平下的临界值列于表 A-3。

表 A-3 Smirnov 检验统计量 nW^2 的临界值表

α	0.10	0.05	0.01	0.001
临界值	0.347	0.461	0.743	1.168

（3）Kolmogorov 检验。Kolmogorov (1933)提出适合性检验的另一统计量 D_n：

$$D_n=Sup|F_n^*(x)-F_0(x)|=\max_{-\infty<x<+\infty}|F_n^*(x)-F_0(x)|$$

$$=\max_{1\le i\le n}\{|\widetilde{F}_n^*(x_{(i)})-F_0(x_{(i)})|,|\widetilde{F}_n^*(x_{(i-1)})-F_0(x_{(i)})|\}$$

其中$1 \le i \le n$，$\widetilde{F}_n^*(x_{(i)}) = i/n$。当样本量 n 较小时，其临界值通过查表 A-4 获得；当 n 大于 10 时，$D_{n,0.05} \approx 1.358/\sqrt{n}$，$D_{n,0.01} \approx 1.628/\sqrt{n}$。

表 A-4 Kolmogorov 检验统计量 D_n 的临界值表

α	0.90	0.75	0.50	0.25	0.10	0.05	0.01
D_α	0.575	0.678	0.830	1.02	1.23	1.36	1.63

六、遗传数据的数据变换

在主基因和多基因混合遗传分离分析方法的导出过程中，有一个很重要的假定，即认为同质遗传群体为单一正态分布，而分离群体为多个正态分布的混合。因此实际应用中应对这一假设进行检验，采用的方法是看亲本和 F_1 这些非分离群体是否服从正态分布，如果亲本和 F_1 经适合性测验具有正态分布，那么我们一般认为在主基因和多基因混合遗传模型下分离群体服从正态混合分布，不需要做数据变换；否则，应考虑遗传数据的一些分布特征而采用适当的变换，以满足分析过程中正态性和正态混合性的假定。常用的变换方法有以下几种。

（1）平方根变换 $y = \sqrt{x}$。如果 X 服从泊松分布，这时分布的方差等于分布的期望（即 $\sigma^2 = \mu$），则做平方根变换后的分布接近于正态分布。一些取值较低的百分数数据（0<x<20%）或处理平均数与均方成比例的数据，也可以用平方根变换后进行分析，但当 x 都很小时，变换 $y = \sqrt{x}$ 又会使方差发生变化，这时采用 $y = \sqrt{x + 0.5}$ 较为合适。

（2）反正弦变换 $y = \sin^{-1}\sqrt{x}$。如果 X 服从两项分布，这时 $\sigma^2 = \mu(1-\mu)$，则反正弦变换后的分布接近于正态分布。注意这时二项分布的数据是以百分数来表示的。

（3）对数变换 $y = \log x$。如果变量的方差与处理平均数的平方成比例时，可做对数变换。如果变量包含有 0 值，可考虑采用 $y = \log(1+x)$。在生物学研究中，有些连续性变量在小值方向的变动有限（如不小于 0），在大值方向却可以有较大的变异，因而可以期望是以左偏分布，并且各处理均方随平均数的增加而增加，则做对数变换也可以改进数据的正态性。

(4) Cox-Box 变换。设 $y=\begin{cases}\dfrac{x^{\lambda}-1}{\lambda}, & \lambda\neq 0\\ \log x, & \lambda=0\end{cases}$

其中 λ 为待定参数。常见的估计参数 λ 方法有 Atkinson 估计和极大似然估计。

第四节 矩阵理论及其应用

一、矩阵理论

（1）矩阵的定义。若有 $m\times n$ 个元素（Element）$a_{ij}(i=1,2,\cdots,m;\ j=1,2,\cdots,n)$ 排成 m 行 n 列，即

$$\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n}\\ a_{21} & a_{22} & \cdots & a_{2n}\\ \vdots & \vdots & \ddots & \vdots\\ a_{m1} & a_{m2} & \cdots & a_{mn}\end{bmatrix}$$

则称之为一 m 行 n 列（或 $m\times n$ 阶）矩阵（Matrix），常用 $\mathbf{A}$、$\mathbf{A}_{m\times n}$ 或 $[a_{ij}]_{m\times n}$ 表示。当一个矩阵的行列数相等时，该矩阵又称为一个方阵（Square matrix）。例如，$\mathbf{A}=\begin{bmatrix}1 & 4\\ 2 & 5\\ 3 & 6\end{bmatrix}$ 为一 3×2 阶矩阵，$\mathbf{B}=\begin{bmatrix}1 & 3\\ 0 & 2\end{bmatrix}$ 为一 2 阶方阵。

当一个矩阵的行数为 m、列数为 1 时，又称为一个 m 维向量（Vector）。如果一个方阵除对角线上的元素外均为 0，则称该方阵为一对角阵（Diagonal matrix），记为

$$\mathbf{D}=diag(d_1,d_2,\cdots,d_n)=\begin{bmatrix} d_1 & \cdots & 0 & 0\\ 0 & d_2 & \cdots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \cdots & d_n\end{bmatrix}$$

如果一个对角阵的对角元素均为 1，则称为单位矩阵（Identity matrix），常用 **I** 表示。

（2）矩阵的相等。有两个同阶矩阵，设 $\mathbf{A}=[a_{ij}]_{m\times n}$ 和 $\mathbf{B}=[b_{ij}]_{m\times n}$，如果一切对应元素均相等，即 $a_{ij}=b_{ij}(i=1,2,\cdots,m;\ j=1,2,\cdots n)$，则称矩阵 **A** 和矩阵 **B** 相等，记为 **A**=**B**。

（3）矩阵的加减法。两个同阶矩阵可以相加减，设 $\mathbf{A}=[a_{ij}]_{m\times n}$ 和 $\mathbf{B}=[b_{ij}]_{m\times n}$，则，

$$\mathbf{C}=\mathbf{A}+\mathbf{B}=[a_{ij}+b_{ij}]_{m\times n}$$

$$\mathbf{D}=\mathbf{A}-\mathbf{B}=[a_{ij}-b_{ij}]_{m\times n}$$

例如，$\mathbf{A}=\begin{bmatrix}1 & 4\\ 2 & 5\\ 3 & 6\end{bmatrix}$ 和 $\mathbf{B}=\begin{bmatrix}0 & 2\\ 1 & -2\\ 5 & -1\end{bmatrix}$，则

$$\mathbf{C}=\mathbf{A}+\mathbf{B}=\begin{bmatrix}1+0 & 4+2\\ 2+1 & 5-2\\ 3+5 & 6-1\end{bmatrix}=\begin{bmatrix}1 & 6\\ 3 & 3\\ 8 & 5\end{bmatrix}$$

$$\mathbf{D}=\mathbf{A}-\mathbf{B}=\begin{bmatrix}1-0 & 4-2\\ 2-1 & 5+2\\ 3-5 & 6+1\end{bmatrix}=\begin{bmatrix}1 & 2\\ 1 & 7\\ -2 & 7\end{bmatrix}$$

矩阵的加法满足交换律和结合律，即 **A**+**B**=**B**+**A**，**A**+(**B**+**C**)=(**A**+**B**)+**C**=**A**+**B**+**C**。

（4）矩阵与常数相乘。设 $\mathbf{A}=[a_{ij}]_{m\times n}$ 为一 $m\times n$ 阶矩阵，λ 为常数，则规定 λ 与 **A** 的乘积为 $\lambda\mathbf{A}=[\lambda a_{ij}]_{m\times n}$。容易证明矩阵与常数相乘时满足：

$$(\alpha+\beta)\mathbf{A}=\alpha\mathbf{A}+\beta\mathbf{A},\quad \alpha(\mathbf{A}+\mathbf{B})=\alpha\mathbf{A}+\alpha\mathbf{B}$$

（5）矩阵的乘法。当矩阵 **A** 的列数等于矩阵 **B** 的行数时，还可定义矩阵间的乘法。$\mathbf{A}=[a_{ij}]_{m\times n}$ 为一个 $m\times n$ 阶矩阵，$\mathbf{B}=[b_{jk}]_{n\times p}$ 为一个 $n\times p$ 阶矩阵，则 **A** 与 **B** 的乘积为一个

$m\times p$ 阶矩阵，记为 $\mathbf{C}=\mathbf{AB}=[c_{ik}]_{m\times p}$，其中 $c_{ik}=\sum_{j=1}^{n}a_{ij}\times b_{jk}$。例如，$\mathbf{A}=\begin{bmatrix}1 & 4\\ 2 & 5\\ 3 & 6\end{bmatrix}$ 为一个 3×2 阶矩阵，$\mathbf{B}=\begin{bmatrix}-1 & 0\\ 3 & 2\end{bmatrix}$ 为一个 2 阶方阵，$\mathbf{x}=\begin{bmatrix}0.5\\ 1.5\end{bmatrix}$ 为一 2 维向量，则 **AB** 为一个 3×2 阶矩阵，**Ax** 为一 3 维向量，即

$$\mathbf{AB}=\begin{bmatrix}1\times(-1)+4\times 3 & 1\times 0+4\times 2\\ 2\times(-1)+5\times 3 & 2\times 0+5\times 2\\ 3\times(-1)+6\times 3 & 3\times 0+6\times 2\end{bmatrix}=\begin{bmatrix}11 & 8\\ 13 & 10\\ 15 & 12\end{bmatrix}$$

$$\mathbf{Ax}=\begin{bmatrix}1\times 0.5+4\times 1.5\\ 2\times 0.5+5\times 1.5\\ 3\times 0.5+6\times 1.5\end{bmatrix}=\begin{bmatrix}6.5\\ 8.5\\ 10.5\end{bmatrix}$$

矩阵乘法不满足交换律，但满足结合律和分配律，即：

AB≠BA，**(AB)C=A(BC)=ABC**，**A(B+C)=AB+AC**

（6）矩阵的转置。设 $\mathbf{A}=[a_{ij}]_{m\times n}$ 为一个 $m\times n$ 阶矩阵，将它的行列交换，则变成一个 $n\times m$ 阶矩阵，称该矩阵为 **A** 的转置（Transposition），记为 $\mathbf{A}'$ 或 $\mathbf{A}^T$。例如，

$$\mathbf{A}=\begin{bmatrix}1 & 4\\ 2 & 5\\ 3 & 6\end{bmatrix},\quad \mathbf{A}'=\begin{bmatrix}1 & 2 & 3\\ 4 & 5 & 6\end{bmatrix}$$

若 **A** 为一方阵，并且有 $\mathbf{A}'=\mathbf{A}$，则称 **A** 为对称矩阵（Symmetric matrix）。例如 $\mathbf{B}=\begin{bmatrix}-1 & 3\\ 3 & 2\end{bmatrix}$ 为一对称矩阵。对于乘积矩阵的转置有 $(\mathbf{AB})'=\mathbf{B}'\mathbf{A}'$。

（7）逆矩阵。设 $\mathbf{A}=[a_{ij}]_{n\times n}$ 为一个 *n* 阶方阵，如果它是非奇异的（其行列式不为 0，即 $|\mathbf{A}|\neq 0$），则存在唯一的一个 *n* 阶方阵 $\mathbf{A}^{-1}$，使得 $\mathbf{A}^{-1}\mathbf{A}=\mathbf{A}\mathbf{A}^{-1}=\mathbf{I}$，则称 $\mathbf{A}^{-1}$ 为 **A** 的逆矩阵。

例如，$\mathbf{A}=\begin{bmatrix}1 & 2\\ 3 & 4\end{bmatrix}$，$|\mathbf{A}|=1\times4-2\times3\neq0$，有唯一$\mathbf{A}^{-1}=\begin{bmatrix}-2.0 & 1.0\\ 1.5 & -0.5\end{bmatrix}$，使得 $\mathbf{A}^{-1}\mathbf{A}=\mathbf{A}\mathbf{A}^{-1}=\mathbf{I}$。

对于乘积矩阵的逆矩阵有$(\mathbf{AB})^{-1}=\mathbf{B}^{-1}\mathbf{A}^{-1}$。

二、最小二乘估计的矩阵形式

数量遗传学研究中使用的模型多为线性模型，最小二乘法是估算线性模型中有关参数的一个有效方法，因此是数量遗传学研究中的基本方法之一。

（1）最小二乘估计的基本原理。考虑一个线性模型

$$y=b_0+b_1x_1+b_2x_2+\cdots+b_nx_n+e \qquad \text{[A-15]}$$

这个模型依赖于 n+1 个参数 b_0，b_1，…，b_n，一般来说这些参数事先是不知道的，要解决的问题是根据 m 组观测值（y_i，x_{i1}，x_{i2}，…，x_{in}）（i=1，2，…，m）将它们估计出来，即：

$$y_1=b_0+b_1x_{11}+b_2x_{12}+\cdots+b_nx_{1n}+e_1$$

$$y_2=b_0+b_1x_{21}+b_2x_{22}+\cdots+b_nx_{2n}+e_2$$

…

$$y_m=b_0+b_1x_{m1}+b_2x_{m2}+\cdots+b_nx_{mn}+e_m$$

在此线性模型中，b_0 称为回归截距，b_j（j=1，2，…，n）称为回归系数，e_i（i=1，2，…，m）为剩余误差。一般来说有 $m>n$。定义离差（即剩余误差）平方和 Q 为

$$Q=\sum_{i=1}^{m}e_i^2=\sum_{i=1}^{m}[y_i-(b_0+b_1x_{i1}+b_2x_{i2}+\cdots+b_nx_{in})]^2$$

则称使得 Q 最小的估计值为最小二乘估计，最小二乘估计可从方程组 $\frac{\partial Q}{\partial b_j}=0$（$j$=0，1，2，…，

n）求得。若记

$$\mathbf{y}_{m\times 1}=\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix},\quad \mathbf{X}_{m\times(n+1)}=\begin{bmatrix} 1 & x_{11} & \cdots & x_{1n} \\ 1 & x_{21} & \cdots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{m1} & \cdots & x_{mn} \end{bmatrix},$$

$$\mathbf{b}_{(n+1)\times 1}=\begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_n \end{bmatrix},\quad \mathbf{e}_{m\times 1}=\begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_m \end{bmatrix}$$

则以上线性模型可用矩阵形式表示为

$$\mathbf{y}=\mathbf{X}\mathbf{b}+\mathbf{e} \qquad \text{[A-16]}$$

X 称为设计矩阵或发生矩阵（Design matrix 或 Incidence matrix）。可以证明参数向量 **b** 的最小二乘估计满足以下正规方程

$$\mathbf{X}'\mathbf{X}\mathbf{b}=\mathbf{X}'\mathbf{y}$$

如果 $\mathbf{X}'\mathbf{X}$ 的逆矩阵存在，则 **b** 的最小二乘估计 $\hat{\mathbf{b}}$ 为

$$\hat{\mathbf{b}}=(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \qquad \text{[A-17]}$$

（2）加权最小二乘估计。如果以上线性模型中，每一组观测数据（y_i，x_{i1}，x_{i2}，…，x_{in}）（i=1，2，…，m）所占的比重是不同的，也就是说每一组数据的误差 e_i 的方差 σ_i^2 是不同的，为了得到各参数的理想的最小二乘估计，就需要把各组观测数据变换为等方差数据。为此，只要把各组数据除以各自的标准差 σ_i 即可，由变换后的数据得到的最小二乘估计称为加权最小二乘估计。记 **W** 为方差的倒数构成的对角阵，即

$$\mathbf{W}=diag(\frac{1}{\sigma_1^2},\frac{1}{\sigma_2^2},\cdots,\frac{1}{\sigma_m^2})=\begin{bmatrix}\frac{1}{\sigma_1^2} & 0 & \cdots & 0\\ 0 & \frac{1}{\sigma_2^2} & \cdots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \cdots & \frac{1}{\sigma_m^2}\end{bmatrix}$$

则加权最小二乘估计的正规方程为

$$\mathbf{X'WXb}=\mathbf{X'Wy}$$

三、世代矩阵

（1）自交后代的基因型频率。近亲繁殖在动植物育种上有重要地位，以植物上的自交为例，假定某一基因位点有两个等位基因 A 和 a，三种基因型 AA、Aa 和 aa，在自交过程中，纯合体 AA 和 aa 的后代与亲代有相同的基因型，杂合体 Aa 产生出 $\frac{1}{2}$ 的纯合体和 $\frac{1}{2}$ 的杂合体。现以 f_1 表示群体中纯合体的频率，f_2 表示杂合体的频率，上标(i)表示繁殖的代数，假定基因型间不存在选择（不同基因型有相同的生存力和繁殖力），则繁殖一代后的纯合体的频率 $f_1^{(1)}$ 和杂合体的频率 $f_2^{(1)}$ 分别为：

$$f_1^{(1)}=f_1^{(0)}+\tfrac{1}{2}f_2^{(0)},\quad f_2^{(1)}=\tfrac{1}{2}f_2^{(0)}$$

记 $\mathbf{f}^{(0)}=\begin{bmatrix}f_1^{(0)}\\ f_2^{(0)}\end{bmatrix}$，$\mathbf{f}^{(1)}=\begin{bmatrix}f_1^{(1)}\\ f_2^{(1)}\end{bmatrix}$，$\mathbf{T}=\begin{bmatrix}1 & \frac{1}{2}\\ 0 & \frac{1}{2}\end{bmatrix}$，则有：$\mathbf{f}^{(1)}=\mathbf{T}\mathbf{f}^{(0)}$。因此繁殖 m 代后有：$\mathbf{f}^{(m)}=\mathbf{T}^m\mathbf{f}^{(0)}$。$\mathbf{T}$ 称为世代矩阵（Generation matrix）或转移矩阵（Transition matrix）。

对于阶数较大的方阵 $\mathbf{T}$，要得到 $\mathbf{T}^m$ 不是很容易，而对角阵的 m 次方却容易算得，即

$$\mathbf{\Lambda}=diag(\lambda_1,\lambda_2,\cdots,\lambda_n),\quad \mathbf{\Lambda}^m=diag(\lambda_1^m,\lambda_2^m,\cdots,\lambda_n^m)$$

数学上已证明对于任意 n 阶方阵 $\mathbf{T}$，存在一个可逆矩阵 $\mathbf{C}=(c_{ij})_{n\times n}$，使得 $\mathbf{CTC}^{-1}$ 为一对角阵，即

$$\mathbf{CTC}^{-1}=\mathbf{\Lambda}=diag(\lambda_1,\lambda_2,\cdots,\lambda_n)，或\mathbf{CT}=\mathbf{\Lambda C}$$

对角元素 λ 由特征方程（Characteristic equation）$|(\mathbf{T}-\lambda\mathbf{I})|=0$ 确定，称为矩阵 **T** 的特征根（characteristic roots or eigenvalues）。在上面的例子中，特征方程为

$$|(\mathbf{T}-\lambda\mathbf{I})|=\begin{vmatrix}1-\lambda & \frac{1}{2}\\ 0 & \frac{1}{2}-\lambda\end{vmatrix}=(1-\lambda)(\tfrac{1}{2}-\lambda)=0$$

从而得到两个特征根 $\lambda_1=1$，$\lambda_2=\frac{1}{2}$。由 $\mathbf{CT}=\mathbf{\Lambda C}$ 得到一个 $\mathbf{C}=\begin{bmatrix}1 & 1\\ 0 & 1\end{bmatrix}$，其逆矩阵 $\mathbf{C}^{-1}=\begin{bmatrix}1 & -1\\ 0 & 1\end{bmatrix}$。如果设 $\mathbf{g}=\mathbf{Cf}$，则有

$$\mathbf{g}^{(1)}=\mathbf{Cf}^{(1)}=\mathbf{CTf}^{(0)}=\mathbf{CTC}^{-1}\mathbf{g}^{(0)}=\mathbf{\Lambda g}^{(0)}$$

因此

$$\mathbf{g}^{(m)}=\mathbf{\Lambda}^m\mathbf{g}^{(0)}=\begin{bmatrix}1 & 0\\ 0 & (\frac{1}{2})^m\end{bmatrix}\mathbf{Cf}^{(0)}=\begin{bmatrix}1 & 0\\ 0 & (\frac{1}{2})^m\end{bmatrix}\times\begin{bmatrix}1 & 1\\ 0 & 1\end{bmatrix}\begin{bmatrix}f_1^{(0)}\\ f_2^{(0)}\end{bmatrix}=\begin{bmatrix}f_1^{(0)}+f_2^{(0)}\\ (\frac{1}{2})^t f_2^{(0)}\end{bmatrix}$$

$$\mathbf{f}^{(m)}=\mathbf{C}^{-1}\mathbf{g}^{(m)}=\begin{bmatrix}1 & -1\\ 0 & 1\end{bmatrix}\begin{bmatrix}f_1^{(0)}+f_2^{0}\\ (\frac{1}{2})^m f_2^{(0)}\end{bmatrix}=\begin{bmatrix}f_1^{(0)}+f_2^{(0)}-(\frac{1}{2})^m f_2^{(0)}\\ (\frac{1}{2})^m f_2^{(0)}\end{bmatrix}$$

（2）重组近交家系群体中的重组率。设有两个基因位点（A-a 和 B-b）间一次交换的重组率为 *r*，两亲本的基因型假定为 AABB 和 aabb，由二者杂交产生的 F_2 群体经连续自交产生的重组近交家系群体（RIL，recombination inbred lines）中有 4 种基因型 AABB、AAbb、aaBB 和 aabb，有两种是亲本基因型，另外两种是重组型。如果在 RIL 产生的过程中不存在选择，则 4 种基因型的比例可用 $\frac{1}{2}(1-R)$，$\frac{1}{2}R$，$\frac{1}{2}R$ 和 $\frac{1}{2}(1-R)$ 表示，其中 *R* 表示 RIL 群体中的重组率，RIL 产生的过程中有多于一次的交换机会，因此一般说来有 $R>r$。

设杂交从 AABB×aabb 的 F_1 代开始，在以后的自交世代中，有 9 种可能的基因型，这些基因型可分为 5 类：（1）AABB 和 aabb，（2）AAbb 和 aaBB，（3）AABb、aaBb、AaBB 和

Aabb，（4）AB/ab 和（5）Ab/aB，各类中不同基因型的频率相等。从各类基因型产生的配子类型可以发现世代矩阵 **T** 为

$$\mathbf{T}=\begin{bmatrix}1 & 0 & \frac{1}{4} & \frac{1}{2}(1-r)^2 & \frac{1}{2}r^2\\ 0 & 1 & \frac{1}{4} & \frac{1}{2}r^2 & \frac{1}{2}(1-r)^2\\ 0 & 0 & \frac{1}{2} & 2r(1-r) & 2r(1-r)\\ 0 & 0 & 0 & \frac{1}{2}(1-r)^2 & \frac{1}{2}r^2\\ 0 & 0 & 0 & \frac{1}{2}r^2 & \frac{1}{2}(1-r)^2\end{bmatrix}=\begin{bmatrix}\mathbf{I}_{2\times2} & \mathbf{R}_{2\times3}\\ \mathbf{O}_{3\times2} & \mathbf{Q}_{3\times3}\end{bmatrix}$$

其中 **I**、**R**、**O** 和 **Q** 为矩阵 **T** 的分块矩阵。随机过程中，类型（1）和（2）称为吸收态（Absorbing states），一旦进入，将不会再转移到其它状态；类型（3）、（4）和（5）称为瞬时态（Transient states）。利用随机过程的有关理论可以证明，最终由类型 *j*+2（*j*=1，2，3）（瞬时态）进入类型 *i*（*i*=1，2）（吸收态）的概率由矩阵

$$\mathbf{R}(\mathbf{I}-\mathbf{Q})^{-1}$$

中的元素（*i*，*j*）表示。

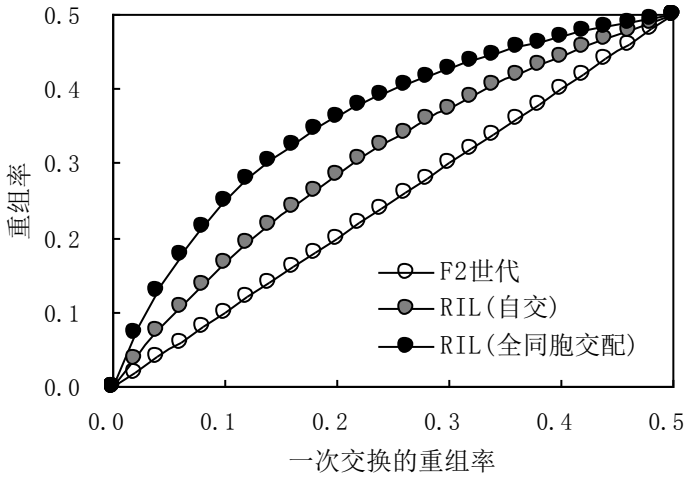

图 A-4 不同群体的重组率

群体 F_1 中只有类型（4），经计算可得由 F_1 中的类型（4）进入类型（1）的概率为 $\frac{1}{1+2r}$，进入类型（2）的概率为 $\frac{2r}{1+2r}$。因此得到通过 F_2 连续自交的 RIL 的重组率 *R* 为

$$R=\frac{2r}{1+2r} \qquad \text{[A-18]}$$

同样的原理可以得到通过全同胞交配的 RIL 的重组率 R 为

$$R=\frac{6r}{1+4r}$$

图 A-4 给出 F2 群体、自交 RIL 群体和全同胞交配 RIL 群体中的重组率。

第五节 随机向量和随机矩阵

一、随机向量和随机矩阵

矩阵代数是分析线性模型的有力工具，用 **x** 表示由 n 个随机变量构成的列向量，称为随机向量（Random vector），即

$$\mathbf{x}=[x_1,x_2,\cdots,x_n]'$$

有时我们需要构建一个 **x** 的线性组合随机变量 **z**，即

$$\mathbf{z}=\sum_{i=1}^{n}a_ix_i=\mathbf{a}'\mathbf{x}$$

其中 **a** 为一常数向量。用 **x** 表示由 n 维随机向量，用 **y** 表示由 m 维随机向量，记 $\mathbf{V}_{\mathbf{x}}=Cov(\mathbf{x})$，$\mathbf{V}_{\mathbf{xy}}=Cov(\mathbf{x},\mathbf{y})$，则：

$$V(\mathbf{a}'\mathbf{x})=\mathbf{a}'\mathbf{V}_{\mathbf{x}}\mathbf{a}$$

$$Cov(\mathbf{a}'\mathbf{x},\mathbf{b}'\mathbf{y})=\mathbf{a}'\mathbf{V}_{\mathbf{xy}}\mathbf{b}$$

通过常数矩阵 $\mathbf{A}_{k\times n}$ 还能构建一个 k 维随机变量向量 $\mathbf{y}=\mathbf{A}_{k\times n}\mathbf{x}$。一般地，对于由 $n\times k$ 个随机变量构成一个 $n\times k$ 阶随机矩阵（Random matrix）**X**，通过矩阵 $\mathbf{A}_{m\times n}$ 和 $\mathbf{B}_{k\times l}$，构建出另一个 $m\times l$ 阶随机矩阵 **Y**，

$$\mathbf{Y}_{m\times l}=\mathbf{A}_{m\times n}\mathbf{X}_{n\times k}\mathbf{B}_{k\times l}$$

随机矩阵 **X** 的期望矩阵等于各元素期望值构成的矩阵表示，则对两个同阶的随机矩阵 **X** 和 **Z**，有

$$E(\mathbf{X}+\mathbf{Z})=E(\mathbf{X})+E(\mathbf{Z})$$

对 $\mathbf{Y}_{m\times l}=\mathbf{A}_{m\times n}\mathbf{X}_{n\times k}\mathbf{B}_{k\times l}$ 有

$$E(\mathbf{Y})=\mathbf{A}E(\mathbf{X})\mathbf{B}$$

二、二次型

对 *n* 维向量 **x** 和 *n* 阶方阵 **A**， $\mathbf{x}'\mathbf{A}\mathbf{x}$ 为一常数，即

$$\mathbf{x}'\mathbf{A}\mathbf{x}=\sum_{i=1}^{n}\sum_{j=1}^{n}a_{ij}x_i x_j$$

称 $\mathbf{x}'\mathbf{A}\mathbf{x}$ 为二次型（Quadratic form 或 Quadratic product）。二次型的更一般形式为 $\mathbf{x}'\mathbf{A}\mathbf{y}$，

$$\mathbf{x}'\mathbf{A}\mathbf{y}=(\mathbf{x}_{n\times 1})'\mathbf{A}_{n\times m}\mathbf{y}_{m\times 1}$$

对二次型 $\mathbf{x}'\mathbf{A}\mathbf{y}$ 有

$$\mathbf{x}'\mathbf{A}\mathbf{y}=(\mathbf{x}'\mathbf{A}\mathbf{y})'=\mathbf{y}'\mathbf{A}'\mathbf{x}$$

记 $\boldsymbol{\mu}_{\mathbf{x}}=E(\mathbf{x})$， $\mathbf{V}_{\mathbf{x}}=Cov(\mathbf{x})$， $\boldsymbol{\mu}_{\mathbf{y}}=E(\mathbf{y})$， $\mathbf{V}_{\mathbf{y}}=Cov(\mathbf{y})$， $\mathbf{V}_{\mathbf{xy}}=Cov(\mathbf{x},\mathbf{y})$，则：

$$E(\mathbf{x}'\mathbf{A}\mathbf{x})=\mathrm{tr}(\mathbf{A}\mathbf{V})+\boldsymbol{\mu}_{\mathbf{x}}'\mathbf{A}\boldsymbol{\mu}_{\mathbf{x}}$$

$$E(\mathbf{x}'\mathbf{A}\mathbf{y})=\mathrm{tr}(\mathbf{A}\mathbf{V}_{\mathbf{xy}})+\boldsymbol{\mu}_{\mathbf{x}}'\mathbf{A}\boldsymbol{\mu}_{\mathbf{y}}$$

$$V(\mathbf{x}'\mathbf{A}\mathbf{x}) = 2\mathrm{tr}(\mathbf{A}\mathbf{V}_{\mathbf{x}})^2 + 4(\boldsymbol{\mu}_{\mathbf{x}})'\mathbf{A}\mathbf{V}_{\mathbf{x}}\mathbf{A}\boldsymbol{\mu}_{\mathbf{x}}$$

$$Cov(\mathbf{x}, \mathbf{x}'\mathbf{A}\mathbf{x}) = 2\mathbf{V}_{\mathbf{x}}\mathbf{A}\boldsymbol{\mu}_{\mathbf{x}}$$

$$Cov(\mathbf{x}'\mathbf{A}\mathbf{x}, \mathbf{x}'\mathbf{B}\mathbf{x}) = 2\mathrm{tr}(\mathbf{A}\mathbf{V}_{\mathbf{x}}\mathbf{B}\mathbf{V}_{\mathbf{x}}) + 4(\boldsymbol{\mu}_{\mathbf{x}})'\mathbf{A}\mathbf{V}_{\mathbf{x}}\mathbf{B}\boldsymbol{\mu}_{\mathbf{x}}$$

其中 $tr(\mathbf{M})$ 表示方阵 **M** 的迹 $\sum M_{ii}$，即 **M** 的对角元素之和。

仍用 **x** 表示 *n* 维随机向量，**V** 表示 **x** 的协方差矩阵，即

$$V_{ij} = Cov(x_i, x_j) = Cov(x_j, x_i) = V_{ji}$$

则线性组合随机变量 $y = \sum c_k x_k = \mathbf{c}'\mathbf{x}$ 的方差为

$$V(y) = Cov(\sum_{i=1}^{n} c_i x_i, \sum_{j=1}^{n} c_j x_j) = \sum_{i=1}^{n}\sum_{j=1}^{n} Cov(c_i x_i, c_j x_j) = \mathbf{c}'\mathbf{V}\mathbf{c}$$

对线性组合随机变量 $\mathbf{a}'\mathbf{x}$ 和 $\mathbf{b}'\mathbf{x}$，协方差为

$$Cov(\mathbf{a}'\mathbf{x}, \mathbf{b}'\mathbf{x}) = \mathbf{a}'\mathbf{V}\mathbf{b}$$

记 $\mathbf{V} = Cov(\mathbf{x}, \mathbf{x})$，则线性组合随机向量 $\mathbf{y}_{l\times 1} = \mathbf{A}_{l\times n}\mathbf{x}_{n\times 1}$ 和 $\mathbf{z}_{m\times 1} = \mathbf{B}_{m\times n}\mathbf{x}_{n\times 1}$ 的协方差矩阵为

$$Cov(\mathbf{y}, \mathbf{z}) = Cov(\mathbf{A}\mathbf{x}, \mathbf{B}\mathbf{x}) = \mathbf{A}\mathbf{V}\mathbf{B}'$$

三、多元正态分布

考虑 *n* 个独立的正态随机变量，$x_i \sim N(\mu_i, \sigma_i^2)$（*i*=1，2，…，*n*），它们的联合概率密度函数 $p(x)$ 是所有一元概率密度函数的乘积，即

$$p(\mathbf{x})=\prod_{i=1}^{n}[\frac{1}{\sqrt{2\pi}\sigma_i}e^{-\frac{(x_i-\mu_i)^2}{2\sigma_i^2}}]=(2\pi)^{-n/2}\prod_{i=1}^{n}\sigma_i\exp\{-\sum_{i=1}^{n}\frac{(x_i-\mu_i)^2}{2\sigma_i^2}\}$$

我们记

$$\mathbf{V}=diag(\sigma_1^2,\sigma_2^2,\cdots,\sigma_n^2)=\begin{bmatrix}\sigma_1^2 & 0 & \cdots & 0\\ 0 & \sigma_2^2 & \cdots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \cdots & \sigma_n^2\end{bmatrix},\quad \boldsymbol{\mu}=\begin{bmatrix}\mu_1\\ \mu_2\\ \vdots\\ \mu_n\end{bmatrix},\quad \mathbf{x}=\begin{bmatrix}x_1\\ x_2\\ \vdots\\ x_n\end{bmatrix}$$

则有：

$$|\mathbf{V}|=\prod_{i=1}^{n}\sigma_i^2$$

$$(\mathbf{x}-\boldsymbol{\mu})'\mathbf{V}^{-1}(\mathbf{x}-\boldsymbol{\mu})=\sum_{i=1}^{n}\frac{(x_i-\mu_i)^2}{\sigma_i^2}$$

因此联合概率密度函数 $p(\mathbf{x})$ 还可用矩阵表示为

$$p(\mathbf{x})=p(\mathbf{x},\boldsymbol{\mu},\mathbf{V})=(2\pi)^{-n/2}|\mathbf{V}|^{-1/2}\exp\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})'\mathbf{V}^{-1}(\mathbf{x}-\boldsymbol{\mu})\}$$

一般地，**x** 中的元素可能是相关的，上式中的对称相关矩阵不一定是对角阵，**x** 的联合概率密度函数仍用上面的公式表示，称 **X** 服从多元正态分布（Multivariate normal distribution），记为

$$\mathbf{X}\sim \mathrm{MVN}_n(\boldsymbol{\mu},\mathbf{V})$$

第六节 线性模型和方差分析

这里用两元素（用 A 和 B 表示）的因子设计说明方差分析中的线性模型，假定因素 A 有 *m* 个水平，因素 B 有 *n* 个水平，y_{ijk} 表示因子组合 $\mathrm{A}_i\mathrm{B}_j$ 下的第 *k* 个观测值。

一、线性模型的建立

用 μ_{ij} 表示因子组合 $\mathrm{A}_i\mathrm{B}_j$ 的理论值，则观测值 y_{ijk} 可先分解为

$$y_{ijk} = \mu_{ij} + \varepsilon_{ijk}$$

其中 ε_{ijk} 为试验误差，相互间独立，且服从均值为 0、方差为 σ_ε^2 的正态分布，即 $\varepsilon_{ijk} \sim N(0,\sigma_\varepsilon^2)$。进一步，设定：

$$\mu = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}\mu_{ij}\ ,\quad \mu_{i\bullet} = \frac{1}{n}\sum_{j=1}^{n}\mu_{ij}\ ,\quad \mu_{\bullet j} = \frac{1}{m}\sum_{i=1}^{m}\mu_{ij}$$

可以看出，μ 即所有组合的平均数，$\mu_{i\bullet}$ 为因素 A 的第 i 个水平的平均数，$\mu_{\bullet j}$ 为因素 B 的第 j 个水平的平均数。设定因子水平 A_i 的效应值为 $a_i = \mu_{i\bullet} - \mu$，$\mathrm{B}_j$ 的效应值为 $b_j = \mu_{\bullet j} - \mu$，显然因子 A 的所有效应值之和为 0，因子 B 的所有效应值之和也为 0，因子组合 $\mathrm{A}_i\mathrm{B}_j$ 的理论值 μ_{ij} 可进一步分解为

$$\mu_{ij} = \mu + a_i + b_j + [(\mu_{ij} - \mu) - a_i - b_j]$$

上式中，$(\mu_{ij} - \mu)$ 为组合 $\mathrm{A}_i\mathrm{B}_j$ 的效应值，因此 $[(\mu_{ij} - \mu) - a_i - b_j]$ 为组合 $\mathrm{A}_i\mathrm{B}_j$ 的效应值减去 A_i 和 B_j 的效应值，它衡量的是 A_i 和 B_j 搭配时的互作效应，用符号 $(ab)_{ij}$ 表示。容易验证

$$\sum_{i=1}^{m}(ab)_{ij} = \sum_{j=1}^{n}(ab)_{ij} = 0$$

根据试验材料的不同，模型[A-11]的效应可以是固定的，称为固定模型（Fixed model）；也可以是随机的，称为随机模型（Random model）；当然还会出现有些效应是固定效应，有些效应是随机效应的情形，称为混和模型（Mixed model）。判断模型中的效应应该视为固定的还是随机的，可借助以下两个原则。

（1）当因子的水平是完全可以控制的时候，因子效应视为固定效应；当因子的水平不是完全可以控制的时候，因子效应视为随机效应。

（2）当试验个体是人为指定的时候，我们不一定要把结论推广到试验群体以外的其它群体，这时试验个体的效应为固定效应；当试验个体是随机挑选的一个样本的时候，我们希望从样本推断总体的性质，试验个体的效应为随机效应。

这样就得到固定效应线性模型的完整表述：

$$y_{ijk}=\mu+a_i+b_j+(ab)_{ij}+\varepsilon_{ijk} \qquad \text{[A-19]}$$

$$\sum_{i=1}^{m}a_i=\sum_{j=1}^{n}b_j=0\,,\ \ \sum_{i=1}^{m}(ab)_{ij}=\sum_{j=1}^{n}(ab)_{ij}=0$$

$$a_i\sim(0,\sigma_{\mathrm{A}}^2)\,,\ \ b_j\sim(0,\sigma_{\mathrm{B}}^2)\,,\ \ (ab)_{ij}\sim(0,\sigma_{\mathrm{AB}}^2)\,,\ \ \varepsilon_{ijk}\sim N(0,\sigma_{\varepsilon}^2)$$

其中 $i=1,\cdots,m$ 表示因子 A 的第 *i* 个水平，$j=1,\cdots,n$ 表示因子 B 的第 *j* 个水平，$k=1,\cdots,r$ 表示第 *k* 个重复。

在随机模型中，各效应不再是一个数值，而是一个随机变量。随机效应线性模型的完整表述为：

$$y_{ijk}=\mu+a_i+b_j+(ab)_{ij}+\varepsilon_{ijk} \qquad \text{[A-20]}$$

$$a_i\sim(0,\sigma_{\mathrm{A}}^2)\,,\ \ b_j\sim(0,\sigma_{\mathrm{B}}^2)\,,\ \ (ab)_{ij}\sim(0,\sigma_{\mathrm{AB}}^2)\,,\ \ \varepsilon_{ijk}\sim N(0,\sigma_{\varepsilon}^2)$$

二、线性模型的方差分析

固定模型[A-11]和随机模型[A-12]有相同的自由度分解和平方和的分解，因此也有相同的均方，所不同的是期望均方这一项（表 A-5）。表 A-5 中平方和的计算如下。设：

$$\bar{y}=\frac{1}{mnr}\sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{k=1}^{r}y_{ijk}\,,\ \ \bar{y}_{ij\bullet}=\frac{1}{r}\sum_{k=1}^{r}y_{ijk}\,,$$

$$\bar{y}_{i\bullet\bullet}=\frac{1}{nr}\sum_{j=1}^{n}\sum_{k=1}^{r}y_{ijk}\,,\ \ \bar{y}_{\bullet j\bullet}=\frac{1}{mr}\sum_{i=1}^{m}\sum_{k=1}^{r}y_{ijk}$$

则从平方和可分解为

$$SS_{\mathrm{T}}=\sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{k=1}^{r}(y_{ijk}-\bar{y})^2=SS_{\mathrm{A}}+SS_{\mathrm{B}}+SS_{\mathrm{AB}}+SS_{\mathrm{E}}$$

其中

$$SS_{\mathrm{A}}=nr\sum_{i=1}^{m}(\bar{y}_{i\bullet\bullet}-\bar{y})^2\ ,\quad SS_{\mathrm{B}}=mr\sum_{j=1}^{n}(\bar{y}_{\bullet j\bullet}-\bar{y})^2\ ,$$

$$SS_{\mathrm{AB}}=r\sum_{i=1}^{m}\sum_{j=1}^{n}(\bar{y}_{ij\bullet}-\bar{y})^2\ ,\quad SS_{\varepsilon}=SS_{\mathrm{T}}-SS_{\mathrm{A}}-SS_{\mathrm{B}}-SS_{\mathrm{AB}}$$

表 A-5 中均方由平方和除以自由度获得。根据表 A-5 中的期望均方，对固定模型来说，各方差的估计量分别为：

$$\sigma_{\mathrm{A}}^2=\frac{1}{nr}(MS_{\mathrm{A}}-MS_{\varepsilon})$$

$$\sigma_{\mathrm{B}}^2=\frac{1}{mr}(MS_{\mathrm{B}}-MS_{\varepsilon})$$

$$\sigma_{\mathrm{AB}}^2=\frac{1}{r}(MS_{\mathrm{AB}}-MS_{\varepsilon})$$

$$\sigma_{\varepsilon}^2=MS_{\varepsilon}$$

方差σ_{A}^2、σ_{B}^2和σ_{AB}^2显著性检验的 F 统计量分别为：

$$F_{\mathrm{A}}=\frac{MS_{\mathrm{A}}}{MS_{\varepsilon}}\sim F[(m-1),mn(r-1)]$$

$$F_{\mathrm{B}}=\frac{MS_{\mathrm{B}}}{MS_{\varepsilon}}\sim F[(n-1),mn(r-1)]$$

$$F_{\mathrm{AB}}=\frac{MS_{\mathrm{AB}}}{MS_{\varepsilon}}\sim F[(m-1)(n-1),mn(r-1)]$$

表 A-5 两因子试验设计的方差分析

方差来源	自由度	平方和	均方	期望均方	
				固定模型	随机模型
因子 A	m-1	SS_{A}	$MS_{\mathrm{A}}= SS_{\mathrm{A}}/(m\text{-}1)$	$nr\sigma_{\mathrm{A}}^2+\sigma_{\varepsilon}^2$	$nr\sigma_{\mathrm{A}}^2+r\sigma_{\mathrm{AB}}^2+\sigma_{\varepsilon}^2$
因子 B	n-1	SS_{B}	$MS_{\mathrm{B}}=SS_{\mathrm{B}}/(n\text{-}1)$	$mr\sigma_{\mathrm{B}}^2+\sigma_{\varepsilon}^2$	$mr\sigma_{\mathrm{B}}^2+r\sigma_{\mathrm{AB}}^2+\sigma_{\varepsilon}^2$
A×B	$(m\text{-}1)(n\text{-}1)$	SS_{AB}	$MS_{\mathrm{AB}}=SS_{\mathrm{AB}}/[(m\text{-}1)(n\text{-}1)]$	$r\sigma_{\mathrm{AB}}^2+\sigma_{\varepsilon}^2$	$r\sigma_{\mathrm{AB}}^2+\sigma_{\varepsilon}^2$
随机误差	$mn(r\text{-}1)$	SS_{ε}	$MS_{\varepsilon}= SS_{\varepsilon}/[mn(r\text{-}1)]$	σ_{ε}^2	σ_{ε}^2

对随机模型来说，各方差的估计量分别为：

$$\sigma_{\mathrm{A}}^2=\tfrac{1}{nr}(MS_{\mathrm{A}}-MS_{\mathrm{AB}})$$

$$\sigma_{\mathrm{B}}^2=\tfrac{1}{mr}(MS_{\mathrm{B}}-MS_{\mathrm{AB}})$$

$$\sigma_{\mathrm{AB}}^2=\tfrac{1}{r}(MS_{\mathrm{AB}}-MS_{\mathrm{E}})$$

$$\sigma_{\varepsilon}^2=MS_{\varepsilon}$$

方差σ_{A}^2、σ_{B}^2和σ_{AB}^2显著性检验的 F 统计量分别为：

$$F_{\mathrm{A}}=\frac{MS_{\mathrm{A}}}{MS_{\mathrm{AB}}}\sim F[(m-1),(m-1)(n-1)]$$

$$F_{\mathrm{B}}=\frac{MS_{\mathrm{B}}}{MS_{\mathrm{AB}}}\sim F[(n-1),(m-1)(n-1)]$$

$$F_{\mathrm{AB}}=\frac{MS_{\mathrm{AB}}}{MS_{\varepsilon}}\sim F[(m-1)(n-1),mn(r-1)]$$

其它试验设计的线型模型和方程分析表可参照上述过程去建立。