在基因分析中的一些设计问题

谢民育

(华中师范大学统计系　武汉)

摘　要　这篇文章介绍基因分析中的一些设计问题，主要涉及连锁分析和关联分析中的抽样效率。在连锁分析中，谢和李(2003)通过利用优良设计理论，获得了有效的抽样方法。在关联分析中，可用均匀设计方法选出固定的置换来替代传统的随机置换，从而更有效地获得关联性置换检验的 p-值估计。

关键词：优良设计，均匀设计，基因研究

1、引言

我们首先介绍一下基因遗传的一些概念和原理，然后讨论连锁分析和关联分析中的抽样效率。

每一种动植物的细胞（cell）都会有该物种特有的一些染色体（chromosomes）对。例如，人的每一个细胞通常有 23 对染色体，鼠 20 对，豌豆 7 对，玉米 10 对等。在高等动植物的染色体对中，有一对是性染色体（sex chromosomes），例如，人和果蝇的性染色体是由 X 和 Y 组成，雌性为 XX 型染色体，雄性为 XY 型染色体，染色体 Y 除了与雄性的生殖力有关外，几乎没有什么机能，因此常称为惰性染色体。这里只涉及活性染色体的研究，今后出现的染色体均指常染色体。在常染色体的研究中，男性和女性是平等的。

一对染色体可想象为两条平行线。染色体上一个给定的位置，好比两平行线上的一点或一段，叫做基因座（locus）。基因座上不同形式的 DNA 遗传变量叫做等位基因（allele），常用字母 A, a, B, b 或数字 1, 2, 3 等来表示。在分子遗传学中，基因（gene）这个术语既指基因座又指等位基因。图 1 给出了这些概念的几何直观：两平行线代表一对染色体，与两平行线垂直的直线标明了染色体上一个基因座，交点处的变量 X 和 Y 取等位基因 A, a, B, b 等。基因座上这样一对等位基因 XY 叫做基因型（genotype），等位基因的顺序并不影响基因型的类型，即当 X 和 Y 取不同的等位基因时，也可认为 XY 和 YX 为同一基因型。如果 X 和 Y 取相同的等位基因，则称此基因型是纯合的（homozygous）；如果取不同的等位基因，则称此基因型是杂合的（heterozygous）。假定一个基因座上有 m 个等位基因，那么，可能的纯合基因型为 m 个，杂合基因型为$\binom{m}{2}$个，总基因型为 $m+\frac{m(m-1)}{2}$ 个。基因型是生物机能的基本单位。

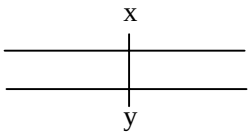

图 1.1　基因的几何解释

现在讨论两个基因座之间的关系，多个基因座的情况在原理上与两基因座的情况是一样的。如果两个基因座在同一对染色体上，且位置较近，则两基因座上等位基因的传递是不独立的，这是由染色体的交换现象所决定的，即在配子（精子或卵子）形成的过程中，两条染色体上的等位基因可能发生交换。假定一个基因座有等位基因 A 和 a，另一基因座有等位基因 B 和 b，如果一个体在前一基因座上有基因型 Aa，在后一基因座上有基因型 Bb，且 A 和 B 在同一染色体上，记该个体为

$$\frac{A\quad B}{a\quad b}$$

当交换发生偶数次时，可能形成的配子是 A B 或 a b，当交换发生奇数次时，可能形成的配子则为 A b 或 a B，如果用 θ 记两基因之间发生奇数次交换的概率，则以 $\frac{1-\theta}{2}$ 的概率形成配子 A B 或 a b，以 $\frac{\theta}{2}$ 的概率形成配子 A b 或 a B，θ 是 0 与 1/2 之间的一个值，称为重组率。$\theta=\frac{1}{2}$ 说明两基因座在不同的染色体对上，称两基因座不连锁；$\theta<\frac{1}{2}$ 则表明两基因座在同一对染色体上，称两基因座连锁；θ 越小，两基因座的位置越近，事实上通过图谱函数可将 θ 转换为两基因座之间的物理距离。

基因连锁分析的任务是通过对基因数据的分析去寻找基因座的位置。在寻找引起某种疾病的基因座（称为性状基因座）位置时，通常的方法是先定一个与之有关的已知基因座（称为标记基因座）作参照物，收集有关病人在标记基因座上的基因数据，通过这些数据来检验特征基因座是否在标记基因座附近，即检验假设

$$H_0:\theta=\frac{1}{2}\ \text{（无连锁）}\ \leftrightarrow\ H_1:\theta<\frac{1}{2}\ \text{（连锁）。}$$

如果能证实引起疾病的基因座就在标记基因座附近，这有利于病理的研究和疾病的治疗。

与连锁有关的另一概念是两基因座等位基因的关联性，量

$$\delta=P(AB)-P(A)P(B)$$

称为关联参数，如果 $\delta=0$，则有，

$$P(AB)=P(A)P(B)\ ,\quad P(Ab)=P(A)P(b)$$

$$P(aB)=P(a)P(B)\ ,\quad P(ab)=P(a)P(b)$$

即两位置等位基因的传递是相互独立的。于是，关联性检验就等价于联列表的独立性检验。关联性与连锁性有如下关系

$$\delta_n = (1-\theta)^n \delta_0 ,$$

其中δ_0是初始代的关联参数，δ_n是第 n 代的关联参数，θ是两基因座之间的重组率。当$n \to \infty$时，$\delta_n \to 0$，但当$\delta_0 \neq 0$和θ很小时，δ_n趋于零的速度很慢。上述关系是关联分析的理论基础。

这篇文章其余部分的安排如下：第二节通过建立检验设计的评判标准--设计的风险，在判决理论的框架下，找到了容许的 Minimax 设计，并证明了它在对称设计类还是最优的。第三节讨论了与关联性有关的联列表的独立性置换检验，我们试图用均匀设计选出固定的置换来代替传统的随机置换，从而更有效地获得检验的 p-值估计。

2、基于 Haseman-Elston 模型连锁分析的优良设计

基因连锁分析中著名的 Haseman-Elston 模型（Haseman 和 Elston，1972）尽管已建立了 30 年，但它现在仍然是基因连锁分析中的一个重要模型（Olson 等，1999）。谢和李（2003）针对这一模型参数的估计问题，给出了优良设计，下面介绍这方面的工作。

2.1、模型

Haseman-Elston 模型涉及到数量形状和同胞对方法。令x_1和x_2分别为同胞对中两同胞的数量性状值，例如体重。假定数量性状值有以下结构：

$$\begin{cases} x_1 = \alpha + v_1 + e_1 \\ x_2 = \alpha + v_2 + e_2 \end{cases} \qquad (1.1)$$

这里α是总体均值，v_i是基因的影响，e_i是环境的影响。为了方便，我们假定基因性状值v_i的性状基因座只有两个等位基因 B 和 b，此时v_i的分布可表为

基因型	BB	Bb	bb
基因性状值	a	d	-a
概率	p^2	2pq	q^2

这里 p 和 q=1-p 分别是等位基因 B 和 b 的概率（见 Olson 等，1999），进一步假定基因的影响和环境的影响是相互独立的，且

$$E(e_i)=0,\ E(e_i^{\ 2})<\infty,\ i=1,2$$

令 $y=(x_1-x_2)^2$，设 π_m 为标记基因座上同胞对共有同源等位基因（IBD）的数目，它可能的取值为 0，1 和 2。对无显性的情形 $(d=0)$，Haseman 和 Elston（1972） 建立了回归模型

$$E(y\mid\pi_m)=\sigma_e^{\ 2}+2\sigma_v^{\ 2}\psi+\sigma_v^{\ 2}(1-2\psi)\pi_m$$

$$=g_0(u,\psi)+g_1(u,\psi)\pi_m$$

（1.3）

这里 $\sigma_e^{\ 2}$ 是 $e=e_1-e_2$ 的方差，$\sigma_v^{\ 2}=2pqa^2$ 是性状基因座上基因的方差，$\psi=\theta^2+(1-\theta)^2$ 是性状基因座和标记基因座之间的重组率 θ 的函数，$u=(\sigma_e^{\ 2},\sigma_v^{\ 2})\in R_+^2=(0,\infty)\times(0,\infty)$，$g_0(u,\psi)=\sigma_e^{\ 2}+2\sigma_v^{\ 2}\psi$ 和 $g_1(u,\psi)=\sigma_v^{\ 2}(1-2\psi)$。模型（1.3）是不满足标准回归假设的，因为 y 关于 π_m 的条件方差

$$\sigma^2(u,\psi,\pi_m)=Var(y\mid\pi_m)=2\sigma_e^{\ 4}+2\sigma_v^{\ 2}(4\sigma_e^{\ 2}+\sigma_v^{\ 2}\psi+a^2)\psi$$

$$+\sigma_v^{\ 2}(1-2\psi)(4\sigma_e^{\ 2}+\sigma_v^{\ 2}+a^2)\pi_m-2\sigma_v^{\ 4}(1-2\psi)^2\pi_m(2-\pi_m)$$

（1.4）

是非齐性的。

标记基因座和性状基因座之间的重组率 θ 属于[0，1/2]。如果两个基因座在不同的染色体对上，或者虽然在同一对染色体上但相隔很远，那幺其中一个基因座上基因的分离与另一个基因座上基因的分离相互独立，θ=1/2，此时称两个基因座是无连锁的。如果两个基因座相邻近，就会出现两个基因座上基因的共分离现象，θ<1/2，此时称两个基因座是连锁的。两个基因座相隔越近，共分离现象发生的概率越大，θ 的值就越小。两个基因座之间的重组率 θ 完全刻画了它们之间的连锁性（Olson 等，1999）。在基因连锁分析中，人们的主要兴趣是检验标记基因座是否与性状基因座相连锁，即检验假设 θ=1/2 和对立假设 $0\le\theta<1/2$。注意到在模型（1.3）中参数 ψ 是 θ 的严格下降函数，且 $\psi\in[\frac{1}{2},1]$，于是关于 θ 的假设等价于关于 ψ 的零假设 $\psi=1/2$ 和对立假设 $1/2<\psi\le1$。容易看出模型（1.3）中的参数

$g_1(u,\psi)=\sigma_v^{\ 2}(1-2\psi)\le 0$，$\forall u=(\sigma_e^{\ 2},\sigma_v^{\ 2})'\in R_+^2$，$\forall \psi\in\Psi=[\frac{1}{2},1]$，　(1.5)

且等号成立的充要条件是$\psi=1/2$，这一事实确保了关于ψ的假设可转化为下面关于模型(1.3)中参数 g_1的零假设：

H_0：$g_1=0$　　(1.6)

和对立假设

H_1：$g_1<0$。　　(1.7)

上述假设的一个自然的检验统计量就是 g_1的最小二乘估计 $\hat{g}_1$。很多作者，比如说，Carey 和 Williason(1991)、Eaves 和 Meyer(1994)、 Risch 和 Zhang(1995,1996)、Zhang 和 Risch(1996)、Zhao(1997)等,基于与模型（1.3）不同的模型，研究了结构（1.1）的抽样方案。实验费用为什么是昂贵的原因能够在这些文章中找到，这里强调的是如何去优化检验 $\hat{g}_1$ 的效率问题。

试验设计的目的就是试验条件的有效选择，以优化所要进行的统计推断，因此设计的优良性与试验的目标有关。在点估计中，许多优良标准，比如 D-优，A-优和 E-优已经建立。这里需要为检验建立优良设计标准，注意到在给定检验的两类错误概率的条件下，一个设计所需的抽样数越小，这个设计就越有效，于是可从抽样数大小的角度来定义设计的风险。仿判决理论，能够建立容许性、极大极小性和设计子类中最优性标准。在这些标准下，找到一个容许的极大极小设计，且证明了这个设计在基因连锁分析中不仅是容许的和极大极小的，而且在一个合理的设计子类中是最优的。通过利用这个优良设计，基因连锁分析中检验所需的样本数减少了一半。

2.2、优良标准

在模型（1.3）中参数向量 g_1 的最小二乘估计 $\hat{g}_1$ 依赖与试验区域 T={0,1,2}中 n 个点 $t_1,t_2,\cdots,t_n$上的相互独立的观察值。这样 n 个点形成了一个设计，它可以用概率测度$\xi=\sum_{i=0}^{2}\delta_{\{i\}}n_i/n$来描述，这里 0,1,2 是 $t_1,t_2\cdots t_n$ 中所有可能的不同的点，非负整数 n_0、n_1和 n_2 是相应点的重复观察数，$\delta_{\{i\}}$附质量 1 于点 i 上。上述设计的概念已推广到允许用任意的权$\omega_i\ge 0$来代替n_i/n，其中$\sum_{i=0}^{2}\omega_i=1$。具有权$\omega_i>0$的点 i 是设计ξ的一个支撑点。

为了保证统计量 $\hat{g}_1$是 g_1的无偏估计，仅限考虑

$$D=\{\xi \sqcup c=(0,-1)'\in L[M(\xi)]\} \tag{2.1}$$

中的设计，这里 $M(\xi)=\int f(\pi_m)f(\pi_m)'d\xi(\pi_m)$ 是设计 ξ 的信息矩阵，$f(\pi_m)=(1,\pi_m)', L[M(\xi)]$ 是 $M(\xi)$ 的列向量所生成的线性空间。此时统计量 $-\hat{g}_1=c'\hat{g}$ 近似服从均值为 $-g_1=c'g$ 和方差为

$$Var(c'\hat{g})=\frac{1}{n}c'M^{-}(\psi)M_{(u,\psi)}M^{-}(\psi)c \tag{2.2}$$

的正态分布，其中 $M^{-}(\xi)$ 是信息矩阵 $M(\xi)$ 的广义逆，

$$M_{(u,\psi)}(\xi)=\int\sigma^2(u,\psi,\pi_m)f(\pi_m)f(\pi_m)'d\xi(\pi_m)。 \tag{2.3}$$

由（1.4）知

$$Var_{H_0}(y\mid t)=\sigma^2(u,\frac{1}{2},\pi_m)=\sigma_0^2(u) \tag{2.4}$$

与 π_m 无关，即模型（1.3）在 H_0 下满足等方差性的标准回归假设。但在对立假设下，由（1.4）知，模型（1.3）不满足等方差性的假设。（2.4）确保了在 H_0 下统计量 $c'\hat{g}$ 是 $c'g$ 的最优线性无偏估计且近似服从正态分布 $N(0,\sigma_0^2(u)c'M^{-}(\xi)c/n)$。

设 α 和 β 分别为关于假设（1.6）对（1.7）检验的犯第一类错误和第二类错误的概率，于是有

$$\alpha=P_{H_0}\left(\frac{c'\hat{g}}{\sqrt{Var_{H_0}(c'\hat{g})}}>z_{1-\alpha}\right)=1-\Phi(z_{1-\alpha}) \tag{2.5}$$

和

$$\beta=P_{H_1}\left(\frac{c'\hat{g}-c'g}{\sqrt{Var_{H_1}(c'\hat{g})}}<\frac{z_{1-\alpha}\sqrt{Var_{H_0}(c'\hat{g})}-c'g}{\sqrt{Var_{H_1}(c'\hat{g})}}\right)$$

$$=\Phi\left(\frac{z_{1-\alpha}\sqrt{\sigma_0{}^2(u)c'M^{-}(\xi)c}-n^{1/2}c'g}{\sqrt{c'M^{-}(\xi)M_{(u,\psi)}(\xi)M^{-}(\xi)c}}\right)=\Phi(z_\beta) \tag{2.6}$$

这里 Φ 是标准正态分布，$z_{1-\alpha}$ 是它的 $1-\alpha$ 分位点。由于 $\sigma_0^2(u)$ 未知，故由（2.5）式导出的拒绝域 $(z_{1-\alpha}\sqrt{\sigma_0^2(u)c'M^{-}(\xi)c/n},\infty)$ 是不确定的，但可以用 $\sigma_0^2(u)$ 的估计来代替 $\sigma_0^2(u)$，从而得到一个确定的拒绝域，就象基因连

锁分析中所做的那样（见 Haseman 和 Elson, 1972、Olson 等, 1999 和 Stoesz, 1997）。在 H_1 下，由（2.6）式有

$$\frac{z_{1-\alpha}\sqrt{\sigma_0{}^2(u)c'M^-(\xi)c}-n^{1/2}c'g}{\sqrt{c'M^-(\xi)M_{(u,\psi)}(\xi)M^-(\xi)c}}=z_\beta,$$

解此方程，得到所需抽样数

$$n=\frac{1}{(c'g)^2}\left(z_{1-\alpha}\sqrt{\sigma_0^2(u)c'M^-(\xi)c}-z_\beta\sqrt{c'M^-(\xi)M_{(u,\psi)}(\xi)M^-(\xi)c}\right)^2。\quad(2.7)$$

注意到 n 越小，设计 ξ 就越有效，因此追求能够极小化（2.7）的设计。由于在（2.7）式中 $z_{1-\alpha}>0, z_\beta<0$ 和 $(c'g)^2>0$ 是与设计无关的，因此设计极小化（2.7）式的充要条件是它极小化

$$R(\xi,u,\psi)=\frac{\omega\sigma_0(u)\sqrt{c'M^-(\xi)c}+(1-\omega)\sqrt{c'M^-(\xi)M_{(u,\psi)}(\xi)M^-(\xi)c}}{Q(u,\psi)},\quad(2.8)$$

这里 $\omega=\dfrac{|z_{1-\alpha}|}{|z_{1-\alpha}|+|z_\beta|}=\dfrac{|z_\alpha|}{|z_\alpha|+|z_\beta|}$ 是两类错误概率 α 和 β 的一个比较量度，$Q(u,\psi)$ 是 u 和 ψ 的取正值的函数，它不依赖于设计，选择函数 $Q(u,\psi)$ 使得 $R(\xi,u,\psi)$ 在 $D\times R_+^2\times[\frac{1}{2},1]$ 上有界。仿决策理论，称函数 $R(\xi,u,\psi)$ 为设计 ξ 的风险，现在优良设计标准可建立如下：

定义 2.1 对检验 $c'\hat{g}$ 来说，一个设计 ξ_1 称为优于另一个设计 ξ_2，如果

$$R(\xi_1,u,\psi)\le R(\xi_2,u,\psi)，\ \forall u\in R_+^2,\forall\psi\in(\frac{1}{2},1]$$

且严格不等号至少在某一点 (u,ψ) 处成立。如果没有优于 ξ 的设计，则称 ξ 是检验 $c'\hat{g}$ 的容许设计。

定义 2.2 一个设计 ξ^* 称为检验 $c'\hat{g}$ 的极大极小（Minimax）设计，如果

$$\sup_{(u,\psi)\in R_+^2\times(\frac{1}{2},1]} R(\xi^*,u,\psi)=\inf_{\xi\in D}\sup_{(u,\psi)\in R_+^2\times(\frac{1}{2},1]} R(\xi,u,\psi).$$

定义 2.3 令 D_0 为 D 的一个子集，一个设计 $\xi^*\in D_0$ 称为检验 $c'\hat{g}$ 的 D_{0^-} 优设计，如果对每一个 $\xi\in D_0$，都有

$$R(\xi^*,u,\psi)\le R(\xi,u,\psi),\forall(u,\psi)\in R_+^2\times(\frac{1}{2},1]。$$

注：2.1 假定对每个 $(u,\pi_m)\in U\times T,$，$\sigma^2(u,\psi,\pi_m)$ 在 $\frac{1}{2}$ 点连续，且对任意 $u\in U$，（2.8）式中函数 $Q(u,\psi)$ 在 $\frac{1}{2}$ 也连续，在这两个假定下，定义 2.1~2.3 中 $(\frac{1}{2},1]$ 可用 $[\frac{1}{2},1]$ 来代替。

2.3、优良设计

设计 $$\xi=\frac{1}{2}\delta_{\{0\}}+\frac{1}{2}\delta_{\{2\}} \qquad (3.1)$$

就是 2.2 节所述的各种优良性标准下的优良设计，具体地有

定理 3.1 设计（3.1）关于检验 $c'\hat{g}$ 是容许的。

定理 3.2 设计（3.1）关于检验 $c'\hat{g}$ 是极大极小的。

设计（3.1）不仅是容许极大极小的，而且在对称设计类

$$D_s=\{\xi\in D:\xi(\pi_m)=\xi(2-\pi_m)\}$$

中是最优的，即我们有

定理 3.3 设计（3.1）关于检验 $c'\hat{g}$ 是 D_s －优的。

注意到 $p(\pi_m=0)=p(\pi_m=2)=\frac{1}{4}$ 和 $p(\pi_m=1)=\frac{1}{2}$（见 Olson 等，1999），于是由大数定理知随机抽样设计在大样本下可近似地表示为

$$\eta=\frac{1}{4}\delta_{\{0\}}+\frac{1}{2}\delta_{\{1\}}+\frac{1}{4}\delta_{\{2\}}。 \qquad (3.2)$$

设计（3.2）是对称设计类中的成员，且是 Haseman-Elston 模型(1.3)中经常使用的设计（见 Olson 等，1999)。定理 3.3 表明优良设计(3.1)要优于常用的随机抽样设计(3.2)。由(2.7)知设计(3.1)和(3.2)的样本大小分别为

$$n_{\xi}=\frac{1}{g_1^2(u,\psi)}\left[z_{1-\alpha}\sqrt{\sigma_0^2(u)}-2^{-1/2}z_{\beta}\sqrt{\sigma^2(u,\psi,0)+\sigma^2(u,\psi,2)}\right]^2$$

和

$$n_{\eta}=\frac{2}{g_1^2(u,\psi)}\left[z_{1-\alpha}\sqrt{\sigma_0^2(u)}-2^{-1/2}z_{\beta}\sqrt{\sigma^2(u,\psi,0)+\sigma^2(u,\psi,2)}\right]^2,$$

容易看出优良设计(3.1)的样本数只是常用设计(3.2)的样本数的一半，因此新的设计(3.1)显着地改善了原设计(3.2)的功效。

3、 独立性置换检验中 p-值的计算

置换检验是基于 Fisher 的随机化原则（Fisher，1951）而提出的一种检验方法。其目的是想把通常基于正态独立性和等方差假设的方差分析方法置于更现实的基础上。然而，要具体实现这个检验却非易事，文献中常用两种方法来实现置换检验：(1）用连续分布来逼近置换带来的离散分布，然后根据连续分布定出否定域的界限，从而得到一个大样本的置换检验（见陈，1981)；(2）用 Monte-Carlo 方法估计出检验的 p-值，从而获得一个小样本的置换检验，它在基因研究中有着广泛的应用（见 Lange，1997)。这一节我们说明可用均匀设计方法代替 Monte-Carlo 方法,从而更有效的给出 p-值估计。

为了叙述的简单，我们用随机向量 $x=(x_1,x_2,\ldots,x_m)$ 来描述 m 个因子的联列表模型，其中 x_i 可能的取值为 $1,2,\ldots,q_i$， $i=1,2,\ldots,m$。记

$J=\{I=(i_1,i_2,\ldots,i_m)'\,|\,1\le i_j\le q_j\}$， $P_I=P(x=I)$ 和 $p_{ji_j}=P(x_j=i_j)$。

假定对 x 进行的 n 次独立的观察中， $x=I$ 的次数为 $n_I^0, I\in J$。现要根据这批观察值来检验 $x_1,x_2,\ldots,x_m$ 的独立性假设，即检验假设

$$H_0：\ P_I=\prod_{j=1}^{m}p_{ji_j}，\ \forall I\in J\quad 。$$

当 n 充分大时，常用 Pearson χ^2 统计量

$$K_n=\sum_{I\in J}\frac{[n_I^0-n\prod_{j=1}^{m}(\frac{n_{ji_j}^0}{n})]^2}{n\prod_{j=1}^{m}(\frac{n_{ji_j}^0}{n})}$$

来检验假设 H_0，其中 $n_{ji_j}^0=\sum_{t\neq j}\sum_{i_t=1}^{q_t}n_I^0$。此时 K_n 渐进于自由度为 $\prod_{j=1}^{m}(q_j-1)$ 的 $\chi^2_{\prod_{j=1}^{m}(q_j-1)}$ 分布。

值得注意的是即使 n 很大，但当 $\prod_{j=1}^{m}q_j$ 较大时，联列表还是非常的稀疏。对稀疏的联列表采用 $\chi^2_{\prod_{j=1}^{m}(q_j-1)}$ 分布来确定否定域的界限的检验将是可疑的（Lange,1997）。一种有效的处理稀疏联列表的方法就是下面将要介绍的精确置换检验。

基于观察值 $\{n_I^0,I\in J\}$，给出一个 $m\times n$ 的矩阵 $A=(a_{ij})$ 它的第 i 行与第 i 个因子相对立，它的第 j 列与第 j 次观察相对应，矩阵 A 中的第 i 行将重复出现第 i 个因子的第 k 个水平 n_{jk}^0 次，$1\le k\le q_j$，这里我们称矩阵 A 为观察矩阵。例如考虑一个因子有两个水平，两个因子均有三个水平的三因子情况，记为 2×3^2。假定 4 次独立观察所得的数据为

$$n_{112}^0=2，n_{221}^0=1，n_{233}^0=1$$

则 3×4 观察矩阵为

$$\begin{pmatrix}1&1&2&3\\1&1&2&3\\2&2&1&3\end{pmatrix}\quad 。$$

对矩阵 A 的每一行进行任意置换，由于每一行有 $n!$ 个不同的置换，

于是共有$(n!)^m$个置换。记这些置换分别为$\pi_l, l=1,2,\ldots,(n!)^m$。它们将会导致出$(n!)^m$个置换矩阵：$\pi_1(A),\ldots\pi_{(n!)^m}(A)$。如果赋予每个矩阵$1/(n!)^m$的概率，这样就得到了一个置换矩阵的离散“均匀”分布，这里给均匀二字打上引号的意义是指它不是真正的均匀分布。事实上在$(n!)^m$个置换矩阵$\pi_l(A)$中仅有$\prod_{j=1}^{m}\binom{n}{\{n_{jk}^0\}}$个不同的矩阵，其中$\binom{n}{\{n_{jk}^0\}}=\frac{n!}{\prod_{j=1}^{m}n_{jk}^0!}$。

让 T 是一个定义在置换矩阵上的检验统计量，例如 T 可取为 Pearson χ^2 统计量K_n，那么 T 以相等的概率取到$T[\pi_1(A)],\ldots,T[\pi_{(n!)^m}(A)]$，即

$$p(T=T[\pi_l(A)])=1/(n!)^m。$$

对于给定的显着性水平α，当 p-值

$$p(T\geq T(A))=\sum\frac{1}{(n!)^m}I_{\{T[\pi_l(A)]\geq T(A)\}}$$

不超过α时，则拒绝假设H_0，这就是文献中常称的精确置换检验，这里精确二字是指 p-值是一个可算出的精确值。这里$I_{\{T[\pi_l(A)]\geq T(A)\}}$是示性函数，即当$T[\pi_l(A)]\geq T(A)$时为 1，否则为 0。然而，要具体实现上述精确置换检验却远非易事，这是因为当 n 和 m 不太小时，p-值的计算量非常之大。于是一个自然的想法就是通过从置换矩阵中随机地抽取$A_1,A_2,\ldots,A_r$ r 个样本，然后用样本示性函数的均值

$$\frac{1}{r}\sum_{i=1}^{r}I_{\{T(A_i)\geq T(A)\}}$$

来估计 p-值。

当 n 和 m 大小适中时，随机抽样方法需要很大的样本量 r 才能保证 p-值的估计达到一定的精度，这将导致上式的计算量非常大。本文利用方和王提出的均匀设计思想（见方，1980、王和方，1981），特别是新近发展的离散偏差方法（见 Hickernell 和 Liu，2002、方等，2003），从置换矩阵中选出均匀散布在置换矩阵中的 s 个确定的置换矩阵来代替随机抽取的 r 个矩阵，并希望在同样的 p-值估计精度下，均匀抽样数 s 要远远小于随机抽样数，这样就有效地减少 p-值估计的计算量。

参考文献

[1] Blackwelder W C, Elston R C. Power and robustness of sib-pair linkage test and extension to larger sibships. Commun Stat Thory Methods, 1982, 11: 449-484

[2] Carey G, Williamson J. Linkage analysis of quantitative trait: increased power by using selected samples. Am J Hum Genet, 1991,49: 786-796

[3] Eaves L, Meyer J. Locating human quantitative trait loci: guidelines for the selection of sibling pairs for genotyping. Behav Genet, 1994, 24:443-455

[4] Fang, K.T., Lin, D.K.J. and Liu, M.Q. Optimal mixed-level supersaturated design. Metrika, to appear.

[5] Fisher, The Design of Experiments, Hafner Publishing Co.Inc.1951

[6] Haseman J K, Elston R C. The investigation of linkage between a quantitative trait and a marker locus. Behav Genet, 1972, 2: 3-19

[7] Hickernell, F.J. and Liu, M.G. Uniform designs limit aliasing. Biometrika, 89, 893-904, 2002

[8] Lang, K. Mathematical and Statistical Methods for Genetic Analysis, New York: Springer-Verlag, 1997

[9] Olson J M, Witte J S, Elston R C. Tutorial in biostatistics: genetic mapping of complex traits. Statist Med, 1999, 18: 2961-2981

[10] Pukelsheim F. Optimal Design of Experiments. New York: John Wiley, 1993

[11] Risch N, Zhang H. Extreme discordant sib pairs for mapping quantitative trait loci in humans. Science, 1995, 268:1584-1589

[12] Risch N, Zhang H. Mapping quantitative trait loci with extreme discordant sib pairs: sample size considerations. Am J Hum Genet, 1996, 58: 836-843

[13] Stoesz M R, Cohen J C, Mooser V, et al. Extension of the Haseman-Elston method to multiple alleles and multiple loci: theory and practice for candidate genes. Ann Hum Genet, 1997, 61: 263-274

[14] Zhang H, Risch N. Mapping quantitative trait loci in humans using extreme concordant sib pair: selected sampling by parental phenotypes. Am J Hum Genet, 1996, 59: 951-957

[15] Zhao H, Zhang H, Rotter J I. Cost-effective sib-pair designs in the mapping of quantitative-trait loci. Am J Hum Genet, 1996, 60: 1211-1221

[16] 陈希孺，数理统计引论，北京：科学出版社，1982.

[17] 方开泰，Experimental design by uniform distribution, Acta Mathematical Applicatae sinica, 3: 363-372, 1980.

[18] 王元和方开泰, A note on uniform distribution and experimental design Kexue Tongbao, 26: 485-489. 1981

[19] 谢民育和李照海，假设检验的优良设计及其在基因连锁分析中的应用，中国科学（A 辑），vol.33(4): 397-408, 2003