

DOI: 10.3724/SP.J.1006.2010.00191

重组自交家系群体 4 对主基因加多基因混合遗传模型分离分析方法的建立

王金社 李海旺 赵团结 盖钧铭*

南京农业大学大豆研究所 / 国家大豆改良中心 / 作物遗传与种质创新国家重点实验室, 江苏南京 210095

摘 要: 主基因加多基因混合遗传模型分离分析方法是基于数量性状表型数据的统计遗传分析方法, 该方法适于育种工作者利用杂种分离世代的数据对育种性状的遗传组成做出初步判断, 制订相应的育种策略, 也可用以校验 QTL 定位所揭示的性状遗传组成。重组自交家系群体(RIL)是一种永久性群体, 可以进行有重复的比较试验, 适合于环境影响较大的复杂性状的遗传研究。本研究以 RIL 群体为对象, 将遗传模型拓展到 4 对主基因, 建立相应的分离分析方法。新构建的模型共包括 4 对主基因和 4 对主基因加多基因两类共 15 个遗传模型, 通过极大似然法和 IECM 算法估算各种模型的分布参数, 由 AIC 值和一组适合性测验选取最佳遗传模型, 再由最小二乘法估计模型的遗传参数。通过模拟实验对所建立的模型进行验证, 模拟群体中一阶遗传参数的估计值与设定值之间有很好的致一致性。以大豆科丰 1 号×南农 1138-2 构成的 RIL 群体及其亲本棕榈酸含量的遗传(I-1 模型, 即 4 对加性-上位性主基因和加性-上位性多基因遗传模型)为例, 说明该方法的应用效果。

关键词: 重组自交家系群体(RIL); 主基因加多基因混合遗传; 分离分析

Establishment of Segregation Analysis of Mixed Inheritance Model with Four Major Genes plus Polygenes in Recombinant Inbred Lines Population

WANG Jin-She, LI Hai-Wang, ZHAO Tuan-Jie, and GAI Jun-Yi*

Soybean Research Institute of Nanjing Agricultural University / National Center for Soybean Improvement / National Key Laboratory for Crop Genetics and Germplasm Enhancement, Nanjing 210095, China

Abstract: The segregation analysis of major genes plus polygenes is a statistical method for genetic analysis of quantitative traits. This method is particularly valuable for breeders to use their data accumulated from segregation populations to estimate the genetic system of target traits which is necessary for designing breeding strategies and also useful for validating the results of QTL mapping. The recombinant inbred line (RIL) population is a permanent population, which is suitable for genetic analyses of complex traits and can be used in replicated experiments. For RIL, the analytical procedures of three major genes plus polygenes mixed inheritance models have been established and widely used in crops. There is an increasing demand on the genetic model expanding from three major genes plus polygenes to four or more major genes plus polygenes. Therefore, the objective of the present study was to establish the analytical procedures of segregation analyses for four major genes plus polygenes mixed inheritance model in RIL population. Fifteen genetic models with four additive and/or epistatic major genes including those without and with polygenes were established. The genetic models and their distribution parameters were solved and estimated with maximum likelihood method and IECM algorithm. The best model was chosen based on Akaike's Information Criterion (AIC) and a set of goodness of fit tests. The genetic parameters of the best model were estimated with the least square method. The established procedure was validated with a set of Monte Carlo simulation experiments. The results showed a relatively high accuracy and consistency for first order parameters between the simulated population and scheduled population. For demonstration of the usefulness of the established procedure, the data of palmitic acid content obtained from a RIL population NJRIKY (derived from Kefeng 1 × Nannong 1138-2) along with their P_1 and P_2 were analyzed. The results showed that the data fitted to Model I-1, i.e. four additive and epistasis major genes plus additive and epistatic polygenes mixed inheritance model.

Keywords: RIL population; Major gene plus polygenes mixed inheritance; Segregation analysis

本研究由国家重点基础研究发展规划(973 计划)项目(2006CB101708, 2009CB118404), 国家高技术研究发展计划(863 计划)项目(2006AA100104), 国家农业部公益性行业专项(200803060)和教育部高等学校创新引智计划项目(B08025)资助。

* 通讯作者(Corresponding author): 盖钧铭, E-mail: sri@njau.edu.cn, Tel: 025-84395405(需要分析软件者请与通讯作者联系。)

第一作者联系方式: E-mail: wangjinshe@gmail.com

Received(收稿日期): 2009-08-03; Accepted(接受日期): 2009-12-08.

自从 Nilsson-Ehle 提出数量性状的多基因假说以来, Fisher、Mather 和 Kempthorne 等对数量遗传学的研究都是建立在多基因基础之上的。1941 年, Mather 将控制质量性状基因与控制数量性状基因区别开来, 将控制质量性状基因称之为主基因, 将控制数量性状基因称为数量基因或多基因。然而育种实践表明有些现象与多基因假说不符。如有些性状的测度是数量化的但发现受主基因控制; 利用分子标记可以定位到控制数量性状的基因; 同一套遗传试验材料在不同环境下的遗传分析得出了不同的试验结果。根据上述遗传现象, 盖钧镒等^[1]提出了对数量性状遗传体系的新认识, 即泛主基因加多基因理论。该理论认为, 控制数量性状的基因数目有多有少, 各对基因效应大小不同且容易受环境影响。将效应较大的、在一般试验条件下可以检测出来的基因, 称为主效基因; 效应小的、在现有的试验条件下即使通过专门的技术仍然检测不出来的基因, 称为微效基因或多基因; 主基因与多基因是相对的。数量性状遗传体系可能仅由主基因组成, 也可能仅由多基因组成, 也可能由主基因和多基因共同组成; 将主基因和多基因同时组成的数量遗传体系称为主基因和多基因混合遗传体系, 或主基因加多基因混合遗传体系, 或主基因加微基因混合遗传体系^[1]。将纯主基因或纯多基因遗传模式看作主基因加多基因混合遗传体系的特例。

基于上述认识, 南京农业大学大豆研究所数量遗传课题组建立了一套数量性状主基因加多基因混合遗传模型分离分析方法体系^[1-8]。该体系包括 5 个主要环节: (1) 每个主基因型的分布为受多基因和环境修饰的正态分布, 分离世代的分布为若干个正态分布的混合分布; (2) 建立一个或多个分离世代联合的各种可能遗传模型的似然函数, 其中包含各个成分分布的频率、平均数和方差等有待根据试验数据做出估计的分布参数。已推导出这些待估分布参数的遗传模型有 7 类, 即 1 对主基因(A)模型、2 对主基因(B)模型、多基因(C)模型、1 对主基因加多基因(D)模型、2 对主基因加多基因(E)模型、3 对主基因(F)模型、3 对主基因加多基因(G)模型, 每类模型中又按分离世代类型和主基因及多基因的加、显、上位性效应的有无、相对大小而进一步分为若干个模型; (3) 采用 IECM 算法估计混合分布中各成分分布的分布数; (4) 对观察数据进行各种可能遗传模型的极大似然分析, 利用 AIC 准则, 似然比检验以及一组适合性测验的结果从各种可能模型中选出最佳遗

传模型及其相应的成分分布参数; (5) 按入选的遗传模型及其相应的遗传参数与成分分布参数间的关系式, 由入选模型的成分分布参数通过最小二乘法估计相应的遗传参数。

主基因加多基因混合遗传模型分离分析方法是基于数量性状表型数据的遗传分析方法, 不需分子标记数据, 适于育种工作者利用杂种分离世代的数据对育种性状的遗传组成做出初步判断, 制订相应的育种策略, 也可用以校验 QTL 定位所揭示的性状遗传组成的结果。迄今, 该方法已得到广泛应用, 在大田和园艺作物上发表相关论文 200 余篇, 为育种工作者提供了有意义的遗传信息^[9-15]。

重组自交家系群体(RIL)是一种永久性群体, 可以进行有重复的比较试验, 尤其适合于环境影响较大的复杂性状的遗传研究。前人已建立了 3 对主基因加多基因混合遗传模型的分离分析方法^[1-3]。为充分提取遗传数据的信息, 许多研究工作者提议, 需要将遗传模型扩展到更多主基因数。本研究以 RIL 群体为对象, 将遗传模型扩展到 4 对主基因加多基因混合遗传模型(包括 4 对主基因(H)和 4 对主基因加多基因(I)两类), 并建立相应的分离分析方法。

1 方法推演

主基因加多基因混合遗传模型的基本假定是, 亲本完全纯合; 所研究的植物为二倍体, 且无复等位基因存在; 无母体效应; 所研究的群体不存在突变、迁移和选择的影响; 不考虑基因间连锁^[1]。

1.1 遗传模型及分布函数

假定试验 RIL 群体的样本组成为 $Y=(y_1, y_2, \dots, y_n)$, 则对应的统计模型为:

$$y_i = \mu + c_i + p_i + e_i \quad (1)$$

其中 y_i 为第 i 个家系的表型值, c_i 为主基因遗传效应, p_i 为多基因遗传效应(c_i 和 p_i 的取值详见附录表 A), e_i 为误差。

1.1.1 重组自交家系群体的遗传模型及其混合分布

根据上述假定, RIL 群体表现为 k 个正态分布的混合分布, 即:

$$x_i \sim \sum_{j=1}^k \pi_j N(\mu_j, \sigma_j^2) \quad (2)$$

由此构建的样本似然函数为:

$$L(Y|\theta) = \prod_{i=1}^n \sum_{j=1}^k \omega_{ij} f(x_i, \mu_j, \sigma_j^2) \quad (3)$$

其中, n 为 RIL 群体的家系数, k 为成分分布数, 不同遗传模型 k 的取值不同(表 1), ω_{ij} 为第 i 个个体属于第 j 个分布中的后验概率, $f(x_i, \mu_j, \sigma^2)$ 表示均值为 μ_j , 方差为 σ^2 的正态分布密度函数。多基因不存在时, σ^2 代表环境方差; 多基因存在时, σ^2 代表多基因方差与环境方差的混合方差。分布参数与遗传参数间的关系因遗传模型的不同而不同。为了更好地划分环境误差, 进一步拓展出包含亲本和重组自交系群体的联合世代分析方法。

1.1.2 P_1 、 P_2 和重组自交家系联合世代的遗传模型及其混合分布 根据分离分析方法的基本假定, P_1 和 P_2 群体表现为单一正态分布, RIL 群体表现为 k 个正态分布的混合分布,

$$\begin{aligned} x_{1\bullet} &\sim N(\mu_1, \sigma_1^2), \quad x_{2\bullet} \sim N(\mu_2, \sigma_2^2), \\ x_{3\bullet} &\sim \sum_{j=1}^k \pi_j N(\mu_{3j}, \sigma_{3j}^2) \end{aligned} \quad (4)$$

由此构建的样本似然函数为:

$$\begin{aligned} L(\mathbf{Y}|\boldsymbol{\theta}) &= \prod_{i=1}^{n_1} f(x_{1i}, \mu_1, \sigma^2) \prod_{i=1}^{n_2} f(x_{2i}, \mu_2, \sigma^2) \\ &\quad \prod_{i=1}^{n_3} \sum_{j=1}^k \omega_{ij} f(x_{3i}, \mu_{3j}, \sigma_{3j}^2) \end{aligned} \quad (5)$$

其中, n_1 为亲本 P_1 的观测值的个数, n_2 为亲本 P_2 的观测值的个数, n_3 为 RIL 群体的家系数, k 同模型(3), x_{1i} 为亲本 P_1 的第 i 个观测值, x_{2i} 为亲本 P_2 的第 i 个观测值, x_{3i} 为 RIL 群体的第 i 个家系的观测值的平均值, μ_1 和 μ_2 分别为亲本 P_1 和亲本 P_2 的分布均值, σ^2 为环境方差, μ_{3j} 为 RIL 群体的第 j 个成分分布的均值, σ_{3j}^2 为第 j 个成分分布的分布方差, 对于 RIL 群体可以假定各成分分布方差相等($\sigma_{3j}^2 = \sigma_3^2$), 当多基因不存在时 $\sigma_3^2 = \sigma^2$, 当多基因存在时 $\sigma_3^2 = \sigma_{pg}^2 + \sigma^2$, 其中 σ_{pg}^2 代表多基因遗传方差。分布参数与遗传参数间的关系因遗传模型不同而不同。

1.2 4 对主基因的非连锁遗传模型

构建的遗传模型主要包括主基因的加性效应、上位性效应, 部分模型中还包括微效基因的加性效应和上位性效应。根据主基因加性效应、主基因位点之间的互作关系、多基因效应是否存在以及多基因之间是否存在互作, 共构建了 15 个遗传模型。例如, I-1 模型是主基因加性上位性和多基因加性上位性遗传模型, 该模型包括互不相等的主基因加性效应(用符号分别表示为 d_a, d_b, d_c, d_d), 主基因之间

相互作用的上位性效应(用符号分别表示为 $i_{ab}, i_{ac}, i_{ad}, i_{bc}, i_{bd}, i_{cd}$)和多基因的加性效应和上位性效应(用符号分别表示为 $[d], [i]$), 受这种遗传模型控制性状的分离群体是由 16 个不同的正态分布组成的混合分布。根据主基因的加性效应有 2 对相等、3 对相等或 4 对主基因的加性效应相等以及主基因之间存在着不同的互作关系, 构成了不同的遗传模型, 分离群体的成分分布数有不同程度的变化。另外, 为进一步鉴别是否存在多基因效应, 构建了包含多基因效应的 I 模型和不包含多基因效应的 H 模型。

构建的遗传模型是建立在基因之间不存在连锁关系的假定之上的。因此, 所构建的两类 15 个遗传模型, 包括受 4 对不连锁主基因控制的性状的所有遗传模型。表 1 和表 2 分别列出了具体的遗传模型的代号、遗传模型的名称、模型对应的分离群体的成分分布数、模型中独立的遗传参数的个数和模型中待估的遗传参数。

1.3 样本似然函数中成分分布参数极大似然估计的 IECM 算法

采用 IECM 算法获得样本似然函数中分布参数的极大似然估计值。IECM 算法包括 E 步骤和迭代 EM 步骤, E 步骤的完全数据对数似然函数的期望函数为:

$$\begin{aligned} L_c(\mathbf{Y}|\boldsymbol{\theta}) &= \sum_{i=1}^{n_1} \log f(x_{1i}; \mu_1, \sigma^2) + \sum_{i=1}^{n_2} \log f(x_{2i}; \mu_2, \sigma^2) + \\ &\quad \sum_{i=1}^{n_3} \sum_{t=1}^k \omega_{it} \log f(x_{3i}; \mu_{3t}, \sigma_{3t}^2) \end{aligned} \quad (6)$$

其中, ω_{it} 为第 i 个家系第 t 个分布的 Bayes 后验概率。迭代 CM 步骤是分步骤计算 $L_c(\mathbf{Y}|\boldsymbol{\theta})$ 的条件极大值和极大值点。

CM₁ 步为在固定多基因方差组分 σ_{30}^2 和误差方差 σ^2 条件下求分步平均数的条件极大似然估计值。

CM₂ 步为在固定迭代 CM₁ 步中获得的分步平均数和 σ^2 条件下求多基因方差组分 σ_{30}^2 的条件极大似然估计值。

CM₃ 步为在固定迭代 CM₁ 和迭代 CM₂ 步骤得到的分步平均数和多基因方差组分下求 σ^2 的条件极大似然估计值。

分步平均数间无约束条件的模型可直接分步骤对 $L_c(\mathbf{Y}|\boldsymbol{\theta})$ 求偏导数得到新一轮参数估计值; 分步平均数间有约束条件的模型可利用 Lagrange 函数分步

表 1 各遗传模型中样本似然函数的成分分布数

Table 1 Number of sample likelihood function component distributions under variance inheritance models

模型代号 Model code	模型名称 Model name	成分分布数 (k) Number of component distributions
H-1	主基因加性上位性模型 Major genes, additive and epistasis model	16
H-2	主基因加性模型 Major genes, additive model	16
H-3	主基因等加性模型 Major genes, equal additive model	5
H-4	主基因, 2 基因等加性模型 Major genes, two gene equal additive model	9
H-5	主基因, 3 基因等加性模型 Major genes, three gene equal additive model	8
I-1	主基因加性上位性-多基因加性上位性模型 Major genes additive-epistasis-polygenes additive-epistasis model	16
I-2	主基因加性上位性-多基因加性模型 Major genes additive-epistasis-polygenes additive model	16
I-3	主基因加性-多基因加性上位性模型 Major genes additive-polygenes additive epistasis model	16
I-4	主基因加性-多基因加性模型 Major genes additive-polygenes additive model	16
I-5	主基因等加性-多基因加性上位性模型 Major genes equal additive-polygenes additive-epistasis model	5
I-6	主基因等加性-多基因加性模型 Major genes equal additive-polygenes additive model	5
I-7	主基因加性效应有 2 对相等-多基因加性上位性模型 Two of major genes equally additive-polygenes additive epistasis model	9
I-8	主基因加性有 2 个相等-多基因加性模型 Major genes 2 equal additive-polygenes additive model	9
I-9	主基因加性效应有 3 个相等-多基因加性上位性模型 Major genes 3 equal additive-polygenes additive-epistasis model	8
I-10	主基因加性效应有 3 个相等-多基因加性模型 Major genes 3 equal additive-polygenes additive model	8

表 2 各遗传模型中样本似然函数的可估遗传参数

Table 2 Estimated parameters of sample likelihood functions under variance mixed genetic models

模型代号 Model code	独立参数个数 Independent parameters	一阶遗传参数 First order parameters	二阶参数 Second order parameters
H-1	12	$m, d_a, d_b, d_c, d_d, i_{ab}, i_{ac}, i_{ad}, i_{bc}, i_{bd}, i_{cd}$	σ^2
H-2	6	m, d_a, d_b, d_c, d_d	σ^2
H-3	3	$m, d_a=d_b=d_c=d_d$	σ^2
H-4	4	$m, d_a=d_b, d_c=d_d$	σ^2
H-5	4	$m, d_a=d_b=d_c, d_d$	σ^2
I-1	15	$m, d_a, d_b, d_c, d_d, i_{ab}, i_{ac}, i_{ad}, i_{bc}, i_{bd}, i_{cd}, [d], [i]$	σ_{30}^2, σ^2
I-2	14	$m, d_a, d_b, d_c, d_d, i_{ab}, i_{ac}, i_{ad}, i_{bc}, i_{bd}, i_{cd}, [d]$	σ_{30}^2, σ^2
I-3	9	$m, d_a, d_b, d_c, d_d, [d], [i]$	σ_{30}^2, σ^2
I-4	8	$m, d_a, d_b, d_c, d_d, [d]$	σ_{30}^2, σ^2
I-5	6	$m, d_a=d_b=d_c=d_d, [d], [i]$	σ_{30}^2, σ^2
I-6	5	$m, d_a=d_b=d_c=d_d, [d]$	σ_{30}^2, σ^2
I-7	7	$m, d_a=d_b, d_c=d_d, [d], [i]$	σ_{30}^2, σ^2
I-8	6	$m, d_a=d_b, d_c=d_d, [d]$	σ_{30}^2, σ^2
I-9	7	$m, d_a=d_b=d_c, d_d, [d], [i]$	σ_{30}^2, σ^2
I-10	6	$m, d_a=d_b=d_c, d_d, [d]$	σ_{30}^2, σ^2

骤确定迭代 CM($i=1,2,3$)步骤中 $L_c(Y|\theta)$ 的条件极值。

(1) 根据样本观测值选择一组分布参数的初始值;

(2) 计算混合群体中样本观测值的后验概率 ω_{it} , 从而得到完全数据的对数似然函数的数学期望 $L_c(Y|\theta)$ (E 步骤);

(3) 分步骤对 $L_c(Y|\theta)$ 求条件极值, 用迭代方法得到分布平均数、多基因方差组分和环境方差的估计(ICM 步骤);

(4) 将得到的估计值作为初始值重复进行(2)和(3)步骤, 直到达到预定的精度为止。

1.4 遗传模型的选择和模型的适合性检验

在得到各遗传模型下分布参数的极大似然估计值后, 可以计算各模型的 AIC 值:

$$AIC = -2 \ln L(Y|\theta) + 2N(k) \quad (7)$$

其中 $N(k)$ 为各模型中独立参数的个数。AIC 值最小的模型应该为最适合的遗传模型。

为确保所选择的遗传模型准确, 对利用 AIC 准则所选择的遗传模型做适合性测验, 如果对所选模型进行适合性测验的结果比其他近似模型的结果好, 则通过 AIC 准则选择的模型是符合所研究性状的最优遗传模型, 否则考虑用其他近似的模型对性状进行分析。适合性检验采用均匀性检验、Smirnov 检验和 Kolmogorov 检验(详见附录)以确定期望均方样本分布间的适合性。

1.5 遗传参数的估计

在最优最适遗传模型下一阶遗传参数的估计是通过最小二乘法利用分布平均数的估计值估计得到的。计算公式见附录。

主基因遗传方差的估计有如下方法: (1) 由分布平均数和分布权重估计,

$$\sigma_{mg}^2 = \sum_{j=1}^k (\pi_j \mu_j^2) - \left[\sum_{j=1}^k (\pi_j \mu_j) \right]^2 \quad (8)$$

其中 σ_{mg}^2 代表主基因遗传方差; k 代表成分分布数; π_j ($j=1,2,\dots,k$) 代表分离群体中第 j 个成分分布在所有分布中所占的比例; μ_j 代表第 j 个成分分布的分布均值。

(2) 由公式 $\sigma_{mg}^2 = \sigma_p^2 - \sigma_3^2$ 计算获得, 其中 σ_p^2 代表分离群体的表型方差。获得主基因遗传方差和表型方差后, 利用公式 $h_{mg}^2 = \sigma_{mg}^2 / \sigma_p^2$ 获得主基因遗传率的估计。

如果试验数据是不包含亲本的单个重组自交家系群体, 可通过家系重复试验提供误差方差 σ_e^2 的估

计, 可得到多基因遗传方差 $\sigma_{pg}^2 = \sigma_3^2 - \sigma_e^2$ 。进而可以获得多基因的遗传率估计值: $h_{pg}^2 = \sigma_{pg}^2 / \sigma_p^2$ 。如果是含有亲本的联合世代分析, 可以通过估计亲本的方差以获得对误差方差 σ_e^2 的估计, 从而获得多基因遗传率的估计。

1.6 计算程序的编制

根据 1.3 所述参数估计方法, 利用 MATLAB 编写相应的计算程序。程序包括 RIL 群体原来推导的 7 类模型和表 1 中列出的两类遗传模型。计算结果包括算法中涉及的分布参数估计结果、极大似然值、AIC 值和各个遗传模型适合性检验的结果。当选到最优遗传模型后, 计算程序可以根据所选的遗传模型估计出相应的一阶遗传参数。

2 遗传模型的模拟验证

为验证模型的正确性, 以最复杂模型 I-1 为基础做模拟实验。根据 I-1 模型设定遗传参数和误差方差, 利用 Monte Carlo 模拟方法分别抽样得到亲本 P_1 、 P_2 和 RIL 群体的表型数据, 利用所编写的分析软件进行分析, 将计算得到的遗传参数的结果和设定的结果进行比较, 以检验该方法的正确性和准确性。

假定某一数量性状受 4 对主基因加性上位性加多基因加性上位性控制。遗传参数设定如下, 群体均值设为 $m = 10.0$, 主基因的加性效应分别设为 $d_a = 5.0$, $d_b = 3.0$, $d_c = 2.5$, $d_d = 1.5$, 主基因之间的上位性效应分别设为 $i_{ab} = -1.0$, $i_{ac} = -2.0$, $i_{ad} = 1.4$, $i_{bc} = 4.0$, $i_{bd} = 0.5$, $i_{cd} = -3.5$, 多基因的加性效应和上位性效应分别设为 $[d] = 2.2$, $[i] = 1.8$, 误差方差设为 $\sigma^2 = 8.0$ 。模拟实验按照 3 次重复的随机区组试验进行设计, 区组效应分别设为 -0.20 , 0.15 , 0.10 。模拟数据包括重组自交家系的两个亲本世代的表型和重组自交家系群体的表型, 两个亲本的样本大小分别设为 50, 重组自交家系群体包括 480 个家系。模拟的重组自交系群体中, 16 种基因型按等比例均匀分布。为衡量分析结果的准确性, 分别做了 5、10、100 次的模拟实验。对 3 组模拟实验所估计的遗传参数做 t 测验, t 测验的无效假设为 $H_0: \hat{g}_i = g_i$, 备择假设为 $H_1: \hat{g}_i \neq g_i$, 其中 \hat{g} 表示遗传参数的估计值, g_i 表示遗传参数的设定值, 其中 $i = 1, 2, \dots, 14$, 表示不同的遗传参数。

分别对 3 组模拟实验做次数分布图(图 1)。根据图 1 的结果发现, 分离群体呈多个不同的分布组成的混合分布。利用已经开发的软件对 3 次实验的模拟

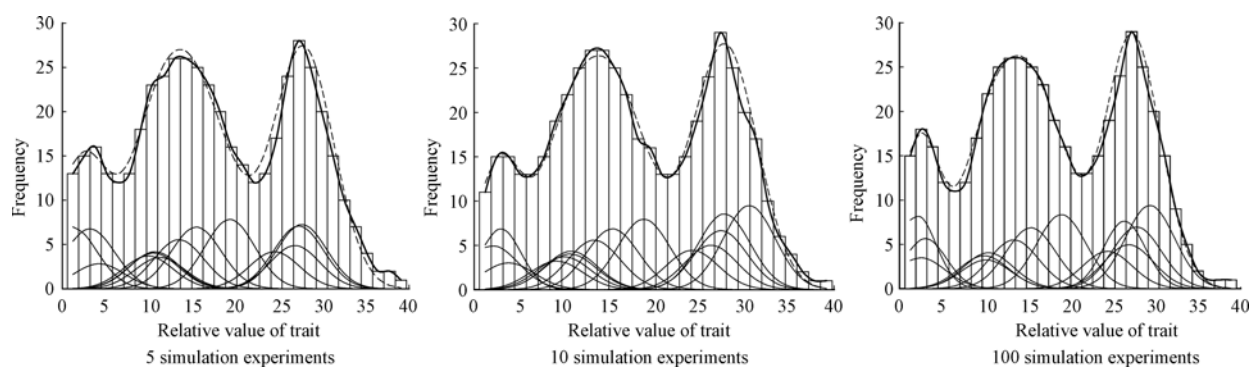


图1 3组模拟实验的模拟分布及其理论分布图

Fig. 1 Simulated and theoretical distributions obtained from three sets of simulation experiments

各图中,上部为群体总分布,其中柱形图为样本分布,虚线表示拟合分布,实线表示理论分布;下部为群体的成分分布。

In the figures, the upper set of curves are the distributions of the population where histogram denotes sample distributions, dotted line denotes fitted distributions and solid line denotes theoretical distributions; the lower set of curves are the component distributions of the population.

数据进行分析,根据 AIC 值和适合性检验的结果 3 次实验都以 I-1 模型为最优最适遗传模型,即 4 对主基因等加性上位性加多基因加性上位性遗传模型。利用估算出的成分分布平均值、成分分布比例和成分分布方差做了拟合分布(图 1),结果表明,参数估计结果和理论分布基本吻合,且随着模拟实验次数增加吻合程度逐渐提高。

根据表 3 和图 1 的结果可知,随着模拟实验次数增加,遗传参数的估计值越接近设定值,群体模拟分布与理论分布接近重合。模拟结果一阶遗传参数估计值的准确度较高,二阶遗传参数估计值的准确度较低,当模拟实验次数达到 100 次时,一阶遗传参数估计值的准确度至少在 93% 以上,其中部分参数的准确度在 99% 以上。各组实验 t 测验的结果表明一阶遗传参数的估计值与设定值之间均无显著差异,只有当模拟次数 100 时误差方差估计值显著偏离设定值。其实际离差并不比其他两组实验大,因为样本大,标准误小,增大了测验的灵敏度,还可能与极大似然法对遗传方差的有偏估计有关。

3 遗传模型的应用实例

本例的实验数据是由国家大豆改良中心提供的大豆 RIL 群体的棕榈酸含量。试验材料为科丰 1 号×南农 1138-2 衍生的 184 个重组自交家系(NJRIKY)及其亲本。于 2004 年在国家大豆改良中心江浦试验站进行试验,采用随机区组设计,设 3 重复。研究的性状为 NJRIKY 群体及其亲本的棕榈酸含量的遗传规律。

将 3 个区组的棕榈酸含量取平均值后用主基因加多基因混合遗传模型分析方法进行分析。选择适

合棕榈酸遗传规律的最佳遗传模型,计算模型中各个遗传参数的大小。计算得到全部遗传模型的 AIC 值和似然函数值(表 4),通过适合性检验和似然比检验结合图 2 的拟合效果,选择 I-1 为最佳遗传模型,即棕榈酸的遗传受 4 对加性-上位性主基因和加性-上位性多基因遗传模型控制。模型 I-1 中各个遗传参数的估计结果列于表 5。由表 5 可知,棕榈酸含量的 4 对主基因的加性遗传效应均为负效应,而各个主基因位点之间的上位性效应均为正效应。控制棕榈酸含量的多基因加性效应比主基因的上位性效应大,多基因的上位性效应为负。控制棕榈酸含量的主基因遗传率为 65.38%,多基因遗传率为 34.62%,由此可知棕榈酸含量的遗传主要受主基因效应控制,但微效多基因效应不可忽略。

4 讨论

育种实践和 QTL 定位的结果表明,很多数量性状的遗传受到 4 对甚至更多的主基因控制。章元明等^[3]构建的遗传模型只包含 3 对主基因,不能满足育种研究的需要。本文分析的大豆 NJRIKY 群体棕榈酸含量的遗传规律就是这样一个例子。郑永战等^[11]利用未拓展之前的主基因加多基因混合遗传模型分析 NJRIKY 群体棕榈酸含量的遗传规律,发现棕榈酸遗传受 3 对主基因加多基因遗传模型控制,其中两对主基因为等加性。本研究通过新建立的 4 对主基因加多基因混合遗传模型对同一群体数据分析发现大豆棕榈酸的遗传受 4 对加性-上位性主基因和加性-上位性多基因遗传模型控制,说明构建 4 对主基因及 4 对主基因加多基因遗传模型的必要性。

从文献报道,主基因加多基因混合遗传模型分

表 3 模拟数据的遗传参数估计
Table 3 Estimates of genetic parameters of simulation data

遗传参数 Genetic value	设定值 Set value	5 次模拟 5 times simulation					10 次模拟 10 times simulation					100 次模拟 100 times simulation				
		平均值	准确度	最小值	最大值	<i>P</i> 值	平均值	准确度	最小值	最大值	<i>P</i> 值	平均值	准确度	最小值	最大值	<i>P</i> 值
		Mean	Accuracy	Min.	Max.	<i>P</i> -value	Mean	Accuracy	Min.	Max.	<i>P</i> -value	Mean	Accuracy	Min.	Max.	<i>P</i> -value
<i>m</i>	10.0	10.77	92.30	9.08	11.83	0.17	10.53	94.70	8.54	11.41	0.09	9.95	99.50	7.85	11.92	0.50
<i>d_a</i>	5.0	5.38	92.40	4.99	6.04	0.12	5.21	95.80	4.42	5.89	0.22	5.10	98.00	3.00	6.87	0.23
<i>d_b</i>	3.0	2.75	91.67	2.36	3.41	0.26	2.82	94.00	2.03	3.50	0.29	3.07	97.67	0.97	4.84	0.40
<i>d_c</i>	2.5	2.58	96.80	2.19	3.24	0.70	2.49	99.60	1.70	3.17	0.95	2.59	96.40	0.49	4.36	0.28
<i>d_d</i>	1.5	1.36	90.67	0.97	2.02	0.50	1.34	89.33	0.55	2.02	0.35	1.58	94.67	-0.52	3.35	0.34
<i>i_{ab}</i>	-1.0	-1.04	96.00	-1.43	-0.38	0.84	-1.02	98.00	-1.81	-0.34	0.91	-0.93	93.00	-3.03	0.84	0.40
<i>i_{ac}</i>	-2.0	-1.99	99.50	-2.38	-1.33	0.97	-2.15	92.50	-2.94	-1.47	0.38	-1.94	97.00	-4.04	-0.17	0.48
<i>i_{ad}</i>	1.4	1.24	88.57	0.85	1.90	0.45	1.41	99.29	0.62	2.09	0.95	1.48	94.29	-0.62	3.25	0.34
<i>i_{bc}</i>	4.0	4.31	92.25	3.92	4.97	0.18	3.72	93.00	2.79	4.26	0.06	4.04	99.00	1.94	5.81	0.65
<i>i_{bd}</i>	0.5	0.35	70.00	-0.04	1.01	0.48	0.60	80.00	-0.31	1.16	0.91	0.48	96.00	-1.50	2.37	0.23
<i>i_{cd}</i>	-3.5	-3.45	98.57	-3.84	-2.79	0.81	-3.54	98.86	-4.33	-2.86	0.81	-3.50	100.00	-5.60	-1.73	0.96
[<i>d</i>]	2.2	1.76	80.00	1.37	2.42	0.08	2.36	92.73	1.57	3.04	0.34	2.20	100.00	0.10	3.97	0.96
[<i>i</i>]	1.8	2.06	85.56	1.67	2.72	0.25	1.59	88.33	0.61	2.08	0.14	1.92	93.33	-0.18	3.69	0.14
σ^2	5.0	6.41	71.80	4.37	7.23	0.16	5.73	85.40	4.64	7.02	0.06	5.39	92.20	3.63	7.58	0.03

准确度表示均值与设定值的偏离程度; *P* 值表示估计值与设定值相等的概率。
Accuracy is the deviation of the estimate from the theoretical value; *P* is the probability of no difference between the estimate and theoretical values.

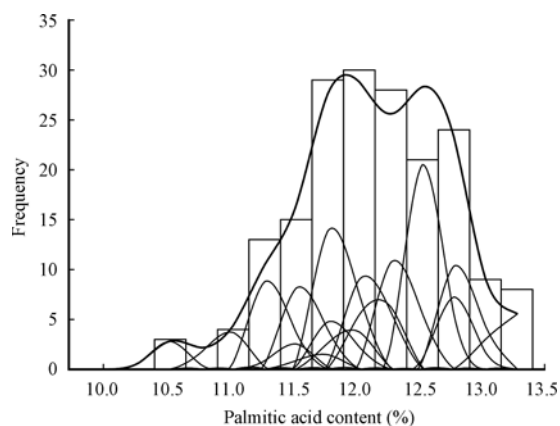


图 2 科丰 1 号×南农 1138-2 RIL 群体的棕榈酸含量的次数分布及拟合分布

Fig. 2 Frequency distribution and fitted distribution of palmitic acid content in NJRIKY derived from a cross Kefeng 1 × Nannong 1138-2

析方法的分析结果与 QTL 定位的结果有相对一致性。王春娥等^[12]利用主基因加多基因遗传模型分析发现, 大豆干豆腐得率的遗传符合两对具有累加作用的连锁主基因加多基因混合遗传模型。通过 CIM

法的 QTL 定位结果显示, 在大豆遗传图谱的 C2 连锁群上 STAS815T~A676I 标记区间存在与干豆腐得率相关的两个紧密连锁的 QTL。刘莹等^[13]在分析大豆耐旱性相关的根系性状时发现, 主基因加多基因混合遗传模型方法和 QTL 定位方法均能检测到相应的性状主要受效应较大的一个数量性状位点控制。王芳等^[14]利用主基因加多基因混合遗传模型分析发现, 大豆在全淹情况下苗期存活率的遗传受 3 对等加性主基因遗传模型控制, 利用 CIM 和 MIM 方法同时检测到 3 个 QTL, 分别位于大豆遗传图谱 A1、D1a 和 G 连锁群上。以上研究结果说明主基因加多基因混合遗传模型方法和 QTL 定位结果一定程度上可以相互验证, 二者的分析结果具有相对一致性。至于本研究所扩展的 4 对主基因加多基因遗传模型分离分析与 QTL 定位结果间有无相对一致性, 还有待后续的研究。

本研究的 4 对主基因加多基因遗传模型中暂只考虑了主基因间无连锁的情况, 在此基础上有必要进一步建立主基因间有连锁的遗传模型。鉴于受成

表 4 棕榈酸含量在不同遗传模型的 AIC 值和极大似然值

Table 4 AIC and maximum likelihood values of palmitic acid content under various mixed genetic models

模型 Model	极大似然值 ML	AIC 值 AIC	模型 Model	极大似然值 ML	AIC 值 AIC	模型 Model	极大似然值 ML	AIC 值 AIC
A-0	-1209.46	2438.92	D-0	-791.18	1594.37	F-3	-782.49	1584.98
A-1	—	—	D-1	-791.18	1594.37	F-4	-768.87	1562.73
B-1-1	-1980.82	4011.64	E-1-1	-782.53	1635.06	G-1	-736.96	1568.92
B-1-2	-827.53	1705.06	E-1-2	-782.51	1625.02	G-2	-737.15	1504.30
B-1-3	-3980.76	8001.53	E-1-3	-783.25	1616.50	G-3	-782.63	1595.25
B-1-4	-2501.38	5042.77	E-1-4	-782.65	1615.30	G-4	-772.49	1579.98
B-1-5	-4544.86	9129.71	E-1-5	-782.41	1629.81	H-1	-686.57	1313.14
B-1-6	-783.25	1606.50	E-1-6	-782.41	1624.81	H-2	-657.11	1464.21
B-1-7	-790.92	1606.84	E-1-7	-792.28	1634.56	H-3	-781.84	1483.69
B-1-8	-790.92	1611.84	E-1-8	-790.92	1631.84	H-4	-726.44	1382.88
B-1-9	-790.92	1621.84	E-1-9	-790.92	1631.84	H-5	-739.34	1418.68
B-2-1	-598.19	1256.38	E-2-1	-746.35	1572.69	I-1	-653.06	1186.12
B-2-2	-790.33	1630.66	E-2-2	-2215.03	4500.05	I-2	-653.32	1246.65
B-2-3	-785.13	1610.26	E-2-3	-784.20	1628.40	I-3	-649.20	1258.40
B-2-4	-756.26	1552.53	E-2-4	-2754.61	5579.22	I-4	-710.14	1380.28
B-2-5	-786.67	1623.34	E-2-5	-786.67	1643.34	I-5	-769.78	1679.56
B-2-6	-782.38	1614.75	E-2-6	-782.35	1634.70	I-6	-778.33	1466.66
B-2-7	-794.42	1628.83	E-2-7	-790.06	1650.13	I-7	-737.24	1414.48
B-2-8	-792.33	1624.67	E-2-8	-791.18	1642.37	I-8	-776.60	1503.21
B-2-9	-791.18	1622.37	E-2-9	-791.18	1642.37	I-9	-740.02	1420.04
C-0	-791.18	1552.00	F-1	-737.43	1564.87	I-10	-768.15	1466.30
C-1	-791.18	1541.95	F-2	-731.66	1493.32			

ML: maximum likelihood estimate; AIC: Akaike's Information Criterion.

表 5 棕榈酸含量的遗传参数估计值
Table 5 Estimates of genetic parameters of palmitic acid content

一阶参数 First order parameters	估计值 Estimate	二阶参数 Second order parameters	估计值 Estimate
m	12.14	σ_p^2	0.52
d_a	-0.48	σ_{mg}^2	0.34
d_b	-0.26	σ_{pg}^2	0.18
d_c	-0.14	σ_e^2	0.02
d_d	-0.05	h_{mg}^2 (%)	65.38
i_{ab}	0.05	h_{pg}^2 (%)	34.62
i_{ac}	0.05		
i_{ad}	0.13		
i_{bc}	0.02		
i_{bd}	0.07		
i_{cd}	0.03		
$[d]$	0.24		
$[i]$	-0.06		

分分布数和待估参数过多的限制, 需要找出更好的参数估计的方法以确保参数估计的准确性。

5 结论

在前人研究的基础上, 将重组自交系(RIL)群体主基因+多基因混合遗传模型拓展到 4 对主基因, 建立了相应的分离分析方法。新构建的遗传模型共包括 4 对主基因和 4 对主基因+多基因两类共 15 个遗传模型, 通过极大似然法和 IECM 算法估算各种模型的分布参数, 由 AIC 值和一组适合性测验选取最佳遗传模型, 再由最小二乘法估计最佳模型的遗传参数。通过模拟实验验证了方法的正确性, 并以大豆重组自交家系群体及其亲本棕榈酸含量数据说明了该方法的应用效果。

References

- [1] Gai J-Y(盖钧铭), Zhang Y-M(章元明), Wang J-K(王健康). The Genetic System of Plant Quantitative Traits (植物数量性状遗传体系). Beijing: Science Press, 2003. pp 30–150 (in Chinese)
- [2] Wang J-K(王健康), Gai J-Y(盖钧铭). Identification of major gene and polygene mixed inheritance model and estimation of genetic parameters of a quantitative trait from F_2 progeny. *Acta Genet Sin* (遗传学报), 1997, 24(5): 432–440 (in Chinese with English abstract)
- [3] Zhang Y-M(章元明), Gai J-Y(盖钧铭), Wang Y-J(王永军). An expansion of joint segregation analysis of quantitative trait for using P_1 , P_2 and DH or RIL populations. *Hereditas* (遗传), 2001, 23(5): 467–470 (in Chinese with English abstract)
- [4] Zhang Y-M(章元明), Gai J-Y(盖钧铭). Identification of mixed major genes and polygene inheritance model of quantitative traits by using DH or RIL population. *Acta Genet Sin* (遗传学报), 2000, 27(7): 634–640 (in Chinese with English abstract)
- [5] Gai J Y, Wang J K. Identification and estimation of QTL model and effects. *Theor Appl Genet*, 1998, 97: 1162–1168
- [6] Gai J-Y(盖钧铭), Wang J-K(王健康). Identification of major gene and polygene mixed inheritance model from backcrosses or $F_{2.3}$ families. *Acta Agron Sin* (作物学报), 1998, 24(4): 402–409 (in Chinese with English abstract)
- [7] Wang J-K(王健康), Gai J-Y(盖钧铭). Identification of major gene and polygene mixed inheritance model of quantitative traits by using joint analysis of P_1 , F_1 , P_2 , F_2 and $F_{2.3}$ generations. *Acta Agron Sin* (作物学报), 1998, 24(6): 651–659 (in Chinese with English abstract)
- [8] Gai J-Y(盖钧铭), Zhang Y-M(章元明), Wang J-K(王健康). A joint analysis of multiple generations for QTL models extended to mixed two major genes plus polygene. *Acta Agron Sin* (作物学报), 2000, 26(4): 385–391 (in Chinese with English abstract)
- [9] Gai J-Y(盖钧铭). Quantitative trait genetic research: The method of major genes plus polygene segregation analysis. *Sci Res* (科学前沿), 2006, 1(1): 85–92 (in Chinese with English abstract)
- [10] He X-H(何小红), Gai J-Y(盖钧铭). Segregation analysis of quantitative traits in backcross inbred line population. *Acta Agron Sin* (作物学报), 2006, 32(2): 210–216 (in Chinese with English abstract)
- [11] Zheng Y-Z(郑永战), Gai J-Y(盖钧铭), Zhou R-B(周瑞宝), Tian S-J(田少君), Lu W-G(卢为国), Li W-D(李卫东). Inheritance of fat and fatty acid composition contents in soybean. *Soybean Sci* (大豆科学), 2007, 26(6): 801–806 (in Chinese with English abstract)
- [12] Wang C-E(王春娥), Gai J-Y(盖钧铭). Inheritance and QTL mapping of tofu and soymilk output in soybean. *Sci Agric Sin* (中国农业科学), 2008, 41(5): 1274–1282 (in Chinese with English abstract)
- [13] Liu Y(刘莹), Gai J-Y(盖钧铭), Lü H-N(吕慧能), Wang Y-J(王永军), Chen S-Y(陈受宜). Identification of drought tolerant germplasm and inheritance and QTL mapping of related root traits in soybean (*Glycine max* L. Merr.). *Acta Genet Sin* (遗传学报), 2005, 32(8): 855–863 (in Chinese with English abstract)
- [14] Wang F(王芳), Zhao T-J(赵团结), Yu D-Y(喻德跃), Chen S-Y(陈受宜), Gai J-Y(盖钧铭). Inheritance and QTL analysis of submergence tolerance at seedling stage in soybean [*Glycine max* L. Merr.]. *Acta Agron Sin* (作物学报), 2008, 34(5): 748–753 (in Chinese with English abstract)
- [15] Liu Z-X(刘章雄), Wang S-C(王守才). Studies of genetic analysis and SSR linked marker location of gene resistance to southern rust in inbred line P(25) of maize. *Acta Genet Sin* (遗传学报), 2003, 30(8): 706–710 (in Chinese with English abstract)

附录

1 遗传模型中主基因和多基因效应分解的系数矩阵

表 A 遗传模型中主基因和多基因效应分解的系数矩阵
Table A Resolved coefficient matrix of major genes and polygene under mixed inheritance models

	c_i 分解系数加费 Coefficient matrix of c_i													p_i 系数 Coefficient matrix of p_i	
	d_a	d_b	d_c	d_d	d_1	d_2	d	i_{ab}	i_{ac}	i_{ad}	i_{bc}	i_{bd}	i_{cd}	$[d]$	$[i]$
H-1	1	1	1	1	0	0	0	1	1	1	1	1	1	0	0
H-2	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0
H-3	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
H-4	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0
H-5	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0
I-1	1	1	1	1	0	0	0	1	1	1	1	1	1	1	1
I-2	1	1	1	1	0	0	0	1	1	1	1	1	1	1	0
I-3	1	1	1	1	0	0	0	0	0	0	0	0	0	1	1
I-4	1	1	1	1	0	0	0	0	0	0	0	0	0	1	0
I-5	0	0	0	0	0	0	1	0	0	0	0	0	0	1	1
I-6	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0
I-7	0	0	0	0	1	1	0	0	0	0	0	0	0	1	1
I-8	0	0	0	0	1	1	0	0	0	0	0	0	0	1	0
I-9	0	0	0	0	1	1	0	0	0	0	0	0	0	1	1
I-10	0	0	0	0	1	1	0	0	0	0	0	0	0	1	0

2 遗传模型的似然函数和遗传模型中参数的估算方法

遗传模型的样本似然函数如下：

$$L(Y|\theta) = \prod_{i=1}^{n_1} f(x_{1i}, \mu_1, \sigma^2) \prod_{i=1}^{n_2} f(x_{2i}, \mu_2, \sigma^2) \prod_{i=1}^{n_3} \sum_{j=1}^k \omega_{ij} f(x_{3i}, \mu_{3j}, \sigma_{3j}^2)$$

$L(Y|\theta)$ 表示待估参数向量为 θ , RIL 表型值为 Y 的似然函数。其中, n_1 为亲本 P_1 的观测值的个数, n_2 为亲本 P_2 的观测值的个数, n_3 为 RIL 群体的家系个数, k 表示不同遗传模型中设定的 RIL 群体的成分分布数, x_{1i} 为亲本 P_1 的第 i 个观测值, x_{2i} 为亲本 P_2 的第 i 个观测值, x_{3i} 为 RIL 群体的第 i 个家系的观测值的平均值, μ_1 和 μ_2 分别为亲本 P_1 和亲本 P_2 的分布均值, σ^2 为环境方差, μ_{3j} 为 RIL 群体的第 j 个成分分布的均值, σ_{3j}^2 为第 j 个成分分布的分布方差, $f(x_{1i}, \mu_1, \sigma^2)$ 表示均值为 μ_1 , 方差为 σ^2 的正态分布密度函数。 $f(x_{2i}, \mu_2, \sigma^2)$ 表示均值为 μ_2 , 方差为 σ^2 的正态分布密度函数。 $f(x_{3i}, \mu_{3j}, \sigma_{3j}^2)$ 表示均值为 μ_{3j} , 方差为 σ_{3j}^2 的正态分布密度函数。对于 RIL 群体可以假定各成分分布方差相等 ($\sigma_{3j}^2 = \sigma_3^2$), 当多基因不存在时 $\sigma_3^2 = \sigma^2$, 当多基因存在时 $\sigma_3^2 = \sigma_{pg}^2 + \sigma^2$, 其中 σ_{pg}^2 代表多基因遗传方差。分布参数与遗传参数间的关系因遗传模型不同而不同。

采用最小二乘法对不同遗传模型中的遗传参数进行估计, 具体的计算公式如下：

$$G = (X'X)^{-1} X'\mu$$

G 表示待估计的遗传参数构成的列向量, 不同遗传模型 G 包含的遗传参数不同。 X 表示成分分布均值与遗传参数的系数矩阵, X 的大小随着遗传模型的不同而不同。 μ 表示由成分分布均值构成的列向量, μ 的大小随着遗传模型的成分分布数的多少而变化。符号“ $'$ ”表示矩阵转置, 符号“ $^{-1}$ ”表示矩阵求逆。

3 样本适合性检验统计量

设 $F(x)$ 为概率分布函数, X_1, X_2, \dots, X_n 为样本观测值, 则 $F(X_i)$ 是 $[0, 1]$ 上的均匀分布, 利用以下 3 个自由度为 1 的 χ^2 统计量, 可以检验 $F(X_i)$ 是否是 $[0, 1]$ 上的均匀分布:

$$U_1^2 = 12 \left[\sum_{i=1}^n F(X_i) - n/2 \right]^2 / n \sim \chi_{df=1}^2$$

$$U_2^2 = (45/4) \left[\sum_{i=1}^n F(X_i)^2 - n/3 \right]^2 / n \sim \chi_{df=1}^2$$

$$U_3^2 = 180 \left[\sum_{i=1}^n (F(X_i) - 0.5)^2 - n/12 \right]^2 / n \sim \chi_{df=1}^2$$

Smirnov 检验的统计量为:

$${}_nW^2 = n \int_{-\infty}^{\infty} [F_n^*(X) - F_0(X)]^2 dF_0(X) = \frac{1}{12n} + \sum_{r=1}^n \left[F(X_{(r)}) - \frac{r-0.5}{n} \right]^2$$

其中 $F_n^*(X)$ 为经验分布函数, $X_{(r)}$ ($r=1, 2, \dots, n$) 为顺序统计量, $F_0(X)$ 为期望分布。

Kolmogorov 检验的统计量为:

$$D_n = \sup \left| F_n^*(x) - F_0(x) \right|$$

其中 $F_n^*(X)$, $F_0(X)$ 同 Smirnov 检验。