第九章

数量性状的主基因和多基因混合遗传模型

数量遗传研究中的假定

- 遗传学上的假定
- 统计学上的假定

遗传学方面的假定

- 多基因假说
- 亲本完全纯合
- 二倍体遗传、无复等位基因
- 无细胞质效应或无母体影响
- 理论群体，无突变、无迁移、无选择
- 在大多数情况下假定各位点基因效应相等
- 在大多数情况下假定基因间无连锁
- 在一些情况下假定无上位性作用（即基因间的相互作用）
- 无基因型和环境互作

统计学方面的假定：表型值分解的线性模型

- 个体或家系的表型观测值是基因型值、环境效应、基因型值与环境的互作效应和随机环境效应的线性总和，即

$$P = G + E + GE + \varepsilon$$

统计学方面的假定：基因型值分解的线性模型

- 基因型值又能进一步分解为加性效应、显性效应和上位性效应的线性相加，即

$$G = A + D + I$$

$$V_G = V_A + V_D + V_I$$

主基因和多基因混合遗传模型

- 主基因和多基因混合遗传模型将主基因效应和多基因效应纳入统计分析模型，即

$$P = m + t + c + \varepsilon$$

  - 其中m 为群体平均数，
  - t 为主基因效应，为固定效应
  - c 为多基因效应，为随机环境效应
  - ε为随机环境效应

随机效应的分布

- 正态分布惯用 $N(\mu,\sigma^2)$ 表示
  - 其中*N* 表示正态分布 (Normal distribution)
  - μ 为正态分布的均值
  - σ^2 为正态分布的方差

$$c \sim N(0,\sigma^2_{pg}) \quad \varepsilon \sim N(0,\sigma^2_{\varepsilon})$$

主基因存在时
分离世代的一些分布特征

- 最早考虑数量性状遗传体系中有无主基因的存在是直观地根据分离世代的分布图形作判断
- 纯粹多基因体系必然是呈现单峰对称的正态分布
- 出现多峰或偏离正态分布的单峰则认为可能存在有主基因

遗传方差的分解：主基因变异和多基因变异

- 总的表型变异可分解为主基因变异、多基因变异和随机环境变异。有

$$c+\varepsilon \sim N(0,\sigma^2) \qquad \sigma^2_{mg} = \frac{1}{2}a^2 + \frac{1}{4}d^2$$

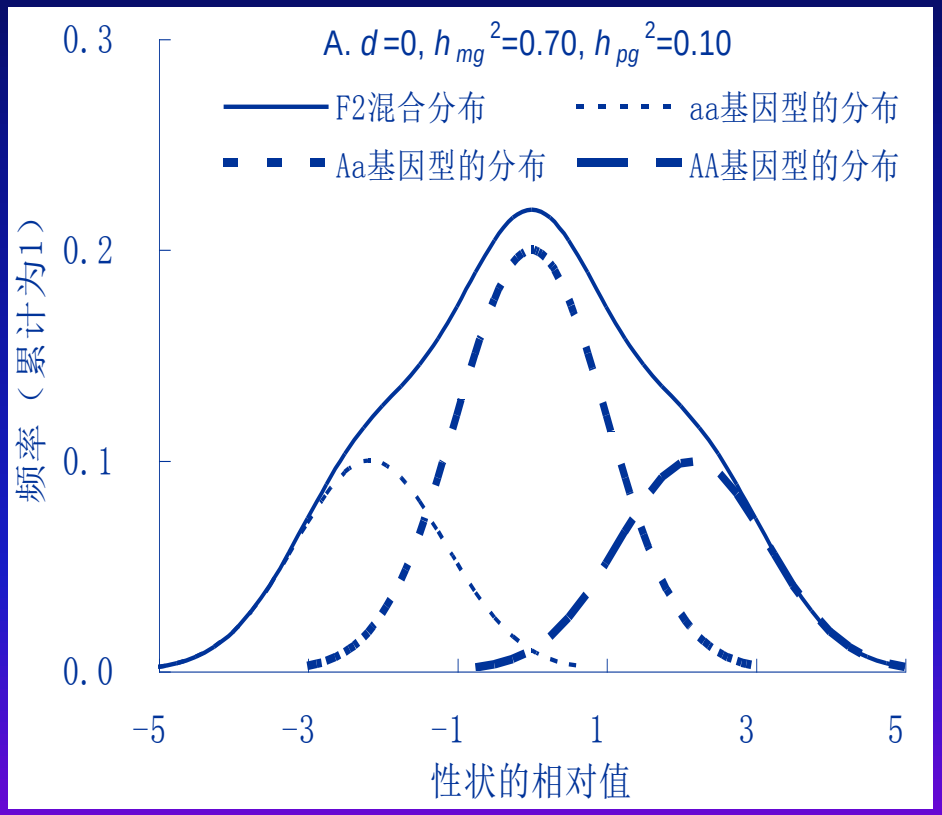
- 主基因遗传率 (major gene heritability)

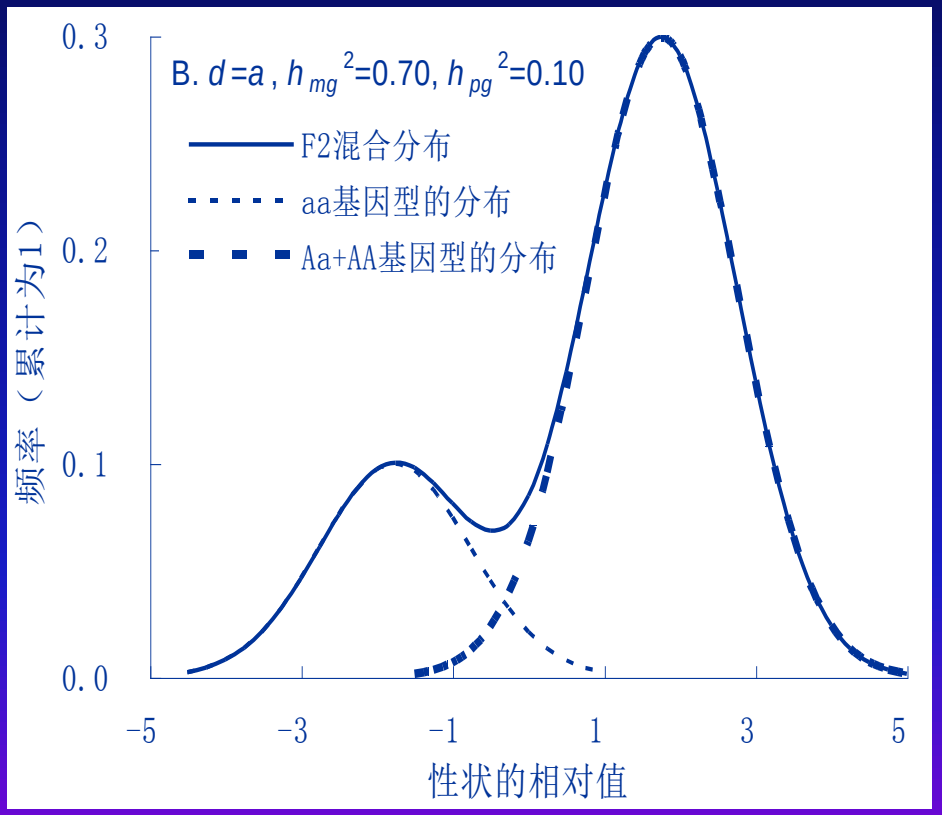
$$h^2_{mg} = \frac{\sigma^2_{mg}}{\sigma^2_{mg} + \sigma^2} \qquad a = \sqrt{\frac{4h^2_{mg}\sigma^2}{(2+r^2)(1-h^2_{mg})}}$$

- F2群体的分布密度函数为

$$g(x) = \frac{1}{4} f(x;-a,\sigma^2) + \frac{1}{2} f(x;d,\sigma^2) + \frac{1}{4} f(x;a,\sigma^2)$$

一对主基因和多基因混合遗传模型下F2群体及其成分分布





多基因变异的进一步分解：多基因的加性方差和显性方差

- 多基因遗传率 (polygene heritability)
- 多基因的加性方差、显性方差及遗传率为

$$h_{pg}^2 = \frac{\sigma_{pg}^2}{\sigma_{mg}^2 + \sigma_{pg}^2 + \sigma_{\varepsilon}^2} = \frac{\frac{1}{2}\sum a^2 + \frac{1}{4}\sum d^2}{\frac{1}{2}a^2 + \frac{1}{4}d^2 + \frac{1}{2}\sum a^2 + \frac{1}{4}\sum d^2 + V_E}$$

$$\sum a^2 = \frac{4h_{pg}^2\sigma^2}{(1-h_{mg}^2)(2+\bar{r}^2)}$$

$$\sum d^2 = \bar{r}^2 \times \frac{4h_{pg}^2\sigma^2}{(1-h_{mg}^2)(2+\bar{r}^2)}$$

- F2群体中的环境变异为 $V_E = \sigma^2 - \sigma_{pg}^2$

环境变异的进一步分解：家系间环境变异和家系内环境变异

- 环境方差可分解为家系间方差和家系内方差两部分，在F2世代中，环境方差为二者之和，即，

$$V_E = V_{Ec} + V_{Ew}$$

- 当有一对主基因存在时，F2:3家系平均数的分布为三个正态分布以1:2:1的比例混合分布。
- 当有一对主基因存在时，重组近交家系 (RIL，recombination inbred lines) 群体或加倍单倍体群体 (DH，doubled haploids) 是由两个正态成分分布按1:1比例构成的混合分布。

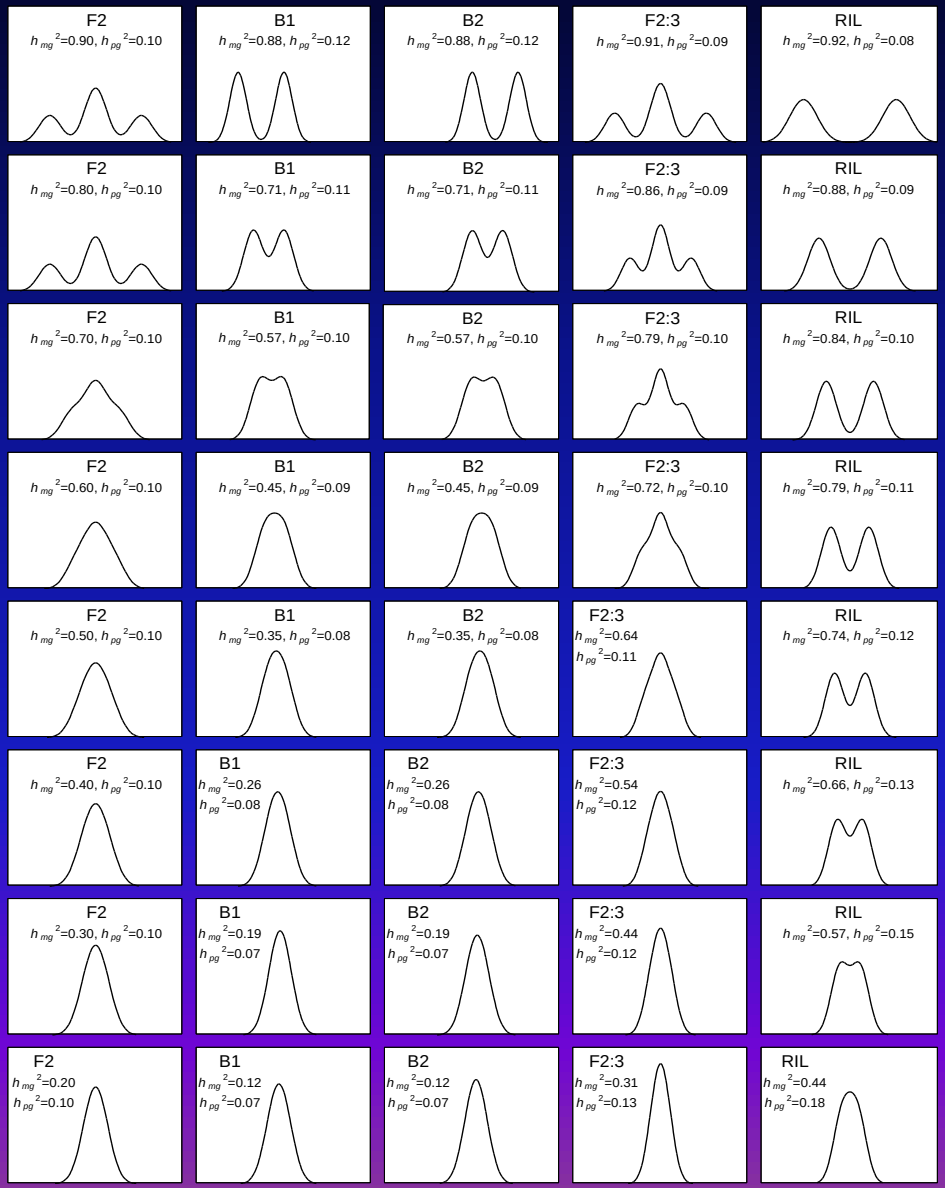
F2:3家系世代中AA家系、Aa家系和aa家系平均的平均数和方差

遗传参数	AA家系	aa家系	Aa家系（n为家系大小）
家系平均数的平均数	$m+a$	$m-a$	$m+\frac{1}{2}d$
家系平均数的理论方差	$\frac{1}{2}\sum a^2+\frac{1}{16}\sum d^2+V_{Ec}$		$\frac{1}{2}\sum a^2+\frac{1}{16}\sum d^2+V_{Ec}$
家系内的方差	$\frac{1}{4}\sum a^2+\frac{1}{8}\sum d^2+V_{Ew}$		$\frac{1}{2}a^2+\frac{1}{4}d^2+\frac{1}{4}\sum a^2+\frac{1}{8}\sum d^2+V_{Ew}$
家系平均数方差的估计（每个家系有n个观测值）	$\frac{1}{2}\sum a^2+\frac{1}{16}\sum d^2+V_{Ec}+\frac{1}{n}\left(\frac{1}{4}\sum a^2+\frac{1}{8}\sum d^2+V_{Ew}\right)$		$\frac{1}{2}\sum a^2+\frac{1}{16}\sum d^2+V_{Ec}+\frac{1}{n}\left(\frac{1}{2}a^2+\frac{1}{4}d^2+\frac{1}{4}\sum a^2+\frac{1}{8}\sum d^2+V_{Ew}\right)$

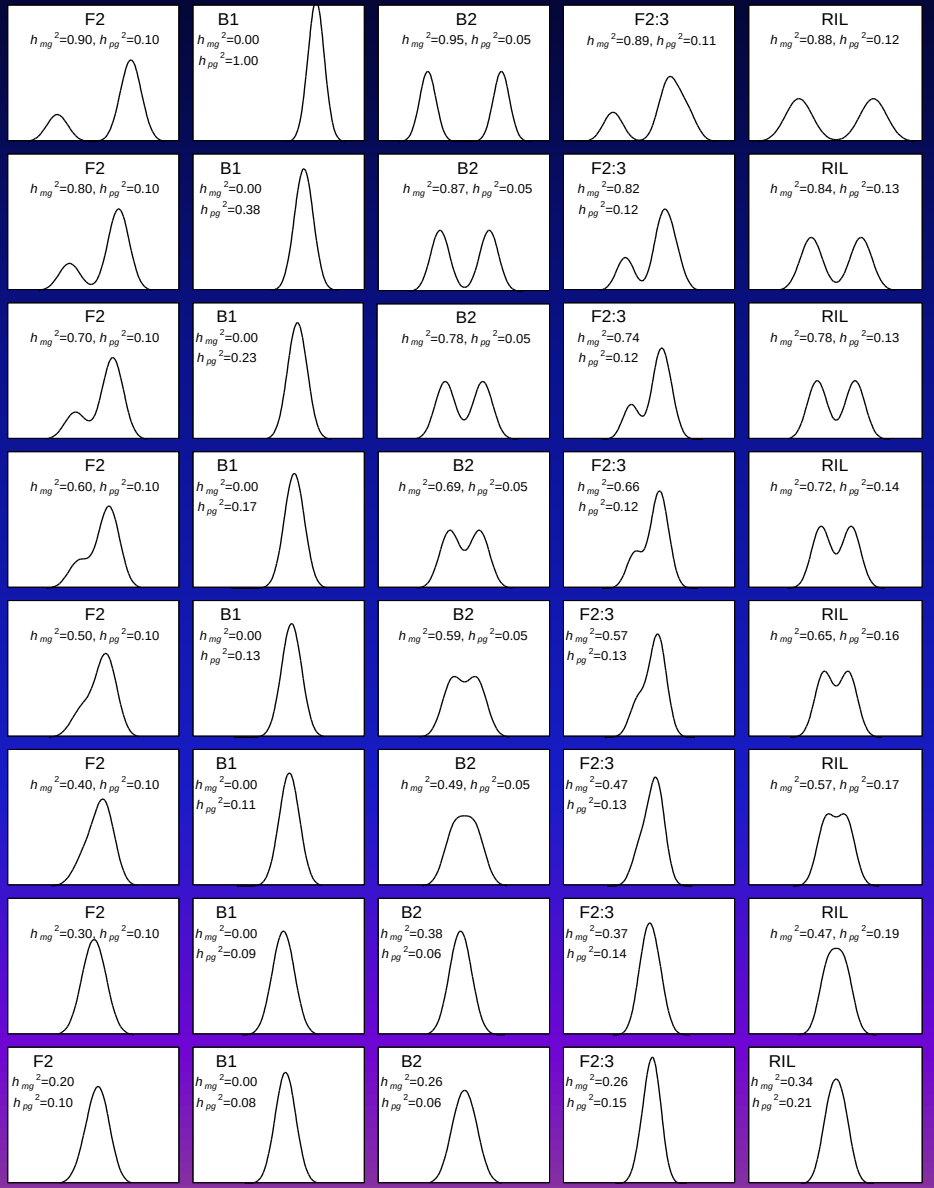
一对主基因和多基因混合遗传模型下各个分离世代的方差组分

世代	主基因方差（σ^2_{mg}）	多基因方差（σ^2_{pg}）	环境方差（σ^2_e）（n为家系大小）
F_2	$\frac{1}{2}a^2+\frac{1}{4}d^2$	$\frac{1}{2}\sum a^2+\frac{1}{4}\sum d^2$	$V_{Ew}+V_{Ec}$
B_1	$\frac{1}{4}(a-d)^2$	$\frac{1}{4}\sum a^2+\frac{1}{4}\sum d^2-\frac{1}{2}\sum ad$	$V_{Ew}+V_{Ec}$
B_2	$\frac{1}{4}(a+d)^2$	$\frac{1}{4}\sum a^2+\frac{1}{4}\sum d^2-\frac{1}{2}\sum ad$	$V_{Ew}+V_{Ec}$
$F_{2:3}$	$\frac{1}{2}a^2+\frac{1}{16}d^2$	$\frac{1}{2}\sum a^2+\frac{1}{16}\sum d^2$	$\frac{1}{n}(\frac{1}{4}\sum a^2+\frac{1}{8}\sum d^2+V_{Ew})+V_{Ec}$
RIL（或DH）	a^2	$\sum a^2$	$\frac{1}{n}V_{Ew}+V_{Ec}$

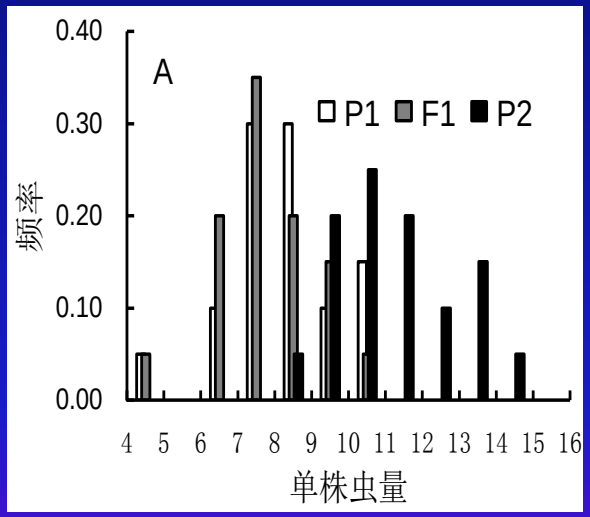
一对加性主基因和多基因混合遗传模型

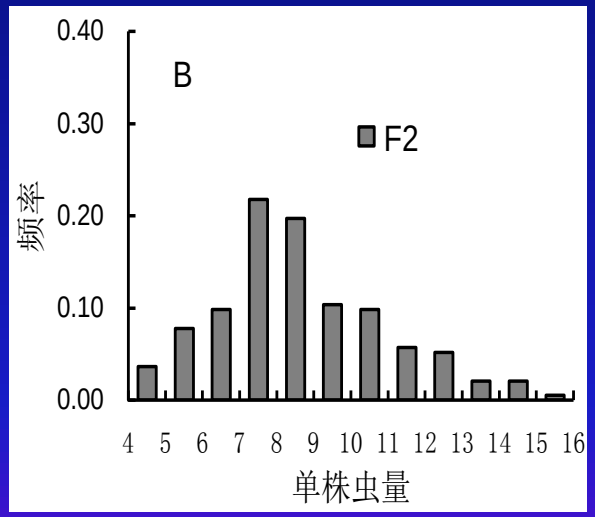


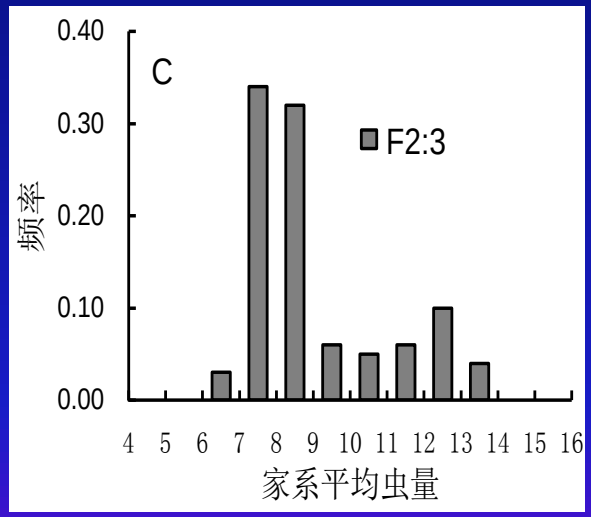
一对显性主基因和多基因混合遗传模型



两个大豆亲本1138-2 (P1) 和邳县天鹅旦 (P2) 及其杂交衍生的一些世代中豆秆蝇虫量的次数分布图



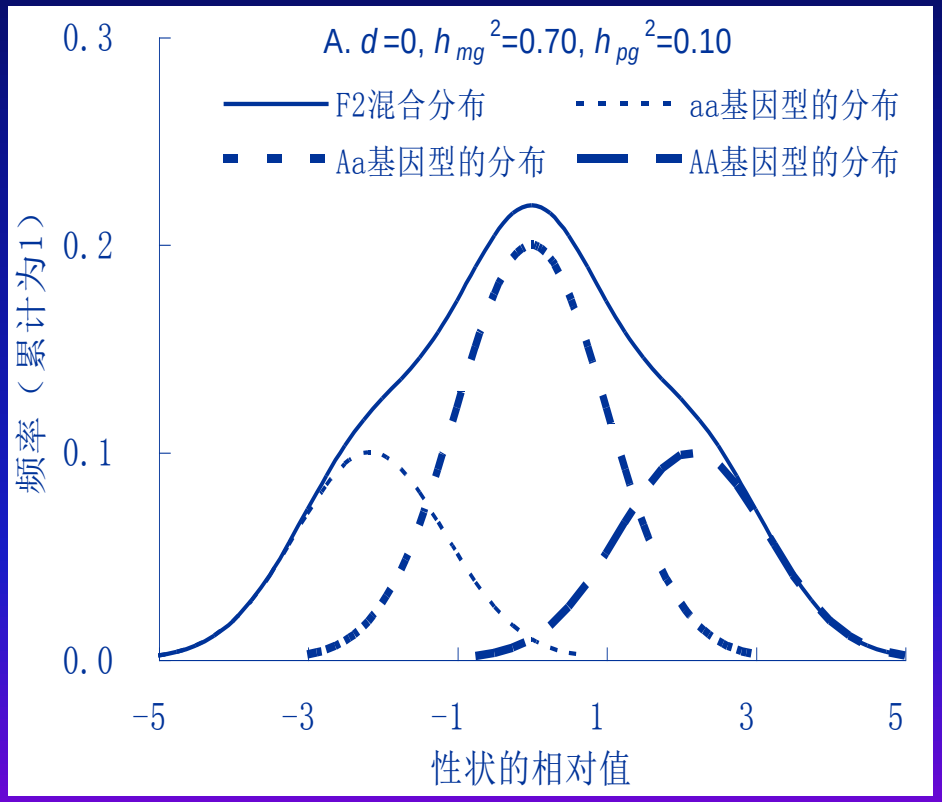


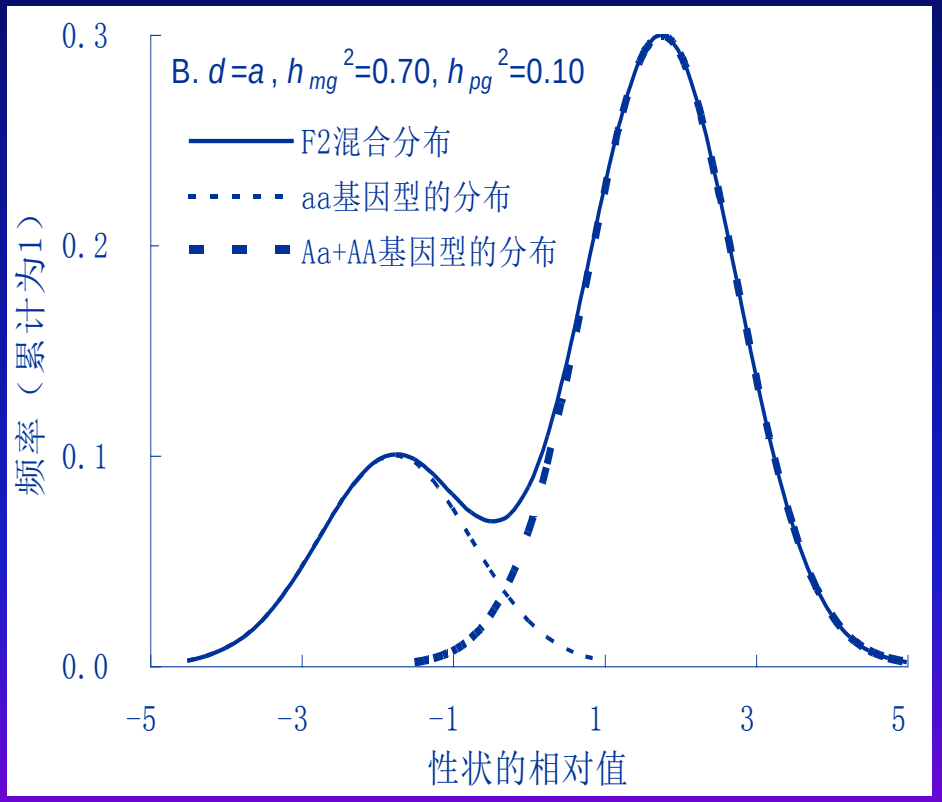


与次数分布图类似的理论曲线

图编号	F_2		$F_{2:3}$ 家系		B_1		B_2		RIL	
	h^2_{mg}	h^2_{pg}	h^2_{mg}	h^2_{mg}	h^2_{mg}	h^2_{pg}	h^2_{mg}	h^2_{pg}	h^2_{mg}	h^2_{pg}
图 9-2 第 4 行	0.60	0.10	0.66	0.12	0.00	0.17	0.69	0.05	0.72	0.14
图 9-2 第 5 行	0.50	0.10	0.57	0.13	0.00	0.13	0.59	0.05	0.65	0.16

混合遗传模型的分离分析方法





混合分布的基本概念

- 混合遗传模型下概率密度函数

$$p(x)=\pi_1 f_1(x)+\pi_2 f_2(x)+\cdots+\pi_k f_k(x)$$

- 混合分布的密度函数

$$p(x|\Phi)=\pi_1 f_1(x|\theta_1)+\pi_2 f_2(x|\theta_2)+\cdots+\pi_k f_k(x|\theta_k)=\sum_{j=1}^{k}\pi_j f_j(x|\theta_j)$$

- 正态分布时的密度函数

$$p(x|\Phi)=\pi_1\frac{1}{\sqrt{2\pi}\sigma_1}e^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}}+\pi_2\frac{1}{\sqrt{2\pi}\sigma_2}e^{-\frac{(x-\mu_2)^2}{2\sigma_2^2}}+\cdots+\pi_k\frac{1}{\sqrt{2\pi}\sigma_k}e^{-\frac{(x-\mu_k)^2}{2\sigma_k^2}}$$

$$=\pi_1 f_1(x;\mu_1,\sigma_1^2)+\pi_2 f_2(x;\mu_2,\sigma_2^2)+\cdots+\pi_k f_k(x;\mu_k,\sigma_k^2)$$

$$=\sum_{j=1}^{k}\pi_j f(x;\mu_j,\sigma_j^2)$$

样本及其似然函数

- 混合总体样本的似然函数

$$L_0(\Phi)=\prod_{i=1}^{n}p(x_i\mid\Phi)=\prod_{i=1}^{n}\sum_{j=1}^{k}\pi_j f_j(x_i\mid\theta_j)=\prod_{i=1}^{n}\sum_{j=1}^{k}\pi_j f_j(x_i;\mu_j,\sigma_j^2)$$

- 来自各成分分布样本的似然函数

$$L_1(\Phi)=L_0(\Phi)\prod_{j=1}^{k}\prod_{h=1}^{n_j}f_j(x_{jh}\mid\theta_j)$$

- 分类数据是独立的样本似然函数

$$L_2(\Phi)=L_1(\Phi)\prod_{j=1}^{k}\pi_j^{n_j}$$

混合分布中所包含的成分分布个数的估计

- AIC (Akaike's information criterion) 准则
- 熵最大原理 (Principle of entropy maximization)
- 熵与最大似然函数的关系:

$$-2B(f;g)+C=-2L(Y|\theta)+2K$$

- 适合度 (Goodness-of-fit) 和节省 (Parsimony)
- 极大似然原理 (Maximum likelihood principle)

EM算法

- EM算法 (Expectation and maximization)
- ECM算法 (Expectation and conditional maximization)
- 迭代ECM (Iterated ECM)

确定混合分布中其它参数的极大似然估计的EM算法

- E步骤：$L_C(\theta)$ 在初始值 $\theta^{(0)}$ 下的期望值 $Q(\theta,\theta^{(0)})$

$$Q(\theta,\theta^{(0)}) = E(L_C(\theta)\mid X;\theta^{(0)}) = \sum_{i=1}^{n}\sum_{j=1}^{k} w_{ij}^{(0)}[\ln\pi_j + \ln f_j(x_i;\theta_j)]$$

$$w_{ij}^{(0)} = \Pr(x_i \in G_j \mid x_i;\theta^{(0)}) = \frac{\pi_j^{(0)} f(x_i;\theta_j)}{\sum_{t=1}^{k}\pi_t^{(0)} f(x_i;\theta_t)}$$

- M步骤：$Q(\theta,\theta^{(0)})$ 的极大值点由下式确定

$$\pi_j = \frac{1}{n}\sum_{i=1}^{n} w_{ij}^{(0)}$$

$$\sum_{i=1}^{n}\sum_{j=1}^{k} w_{ij}^{(0)} \frac{\partial \ln f_j(x_i;\theta)}{\partial\theta} = 0$$

EM算法的优点

- M步骤在正态混合分布的情况下期望函数的极大值点可以用数学式子明确表示出来，而一般情况下企图通过对似然函数求导数来获得极大似然估计的明显数学表示几乎是不可能的；
- EM迭代过程中，似然函数是单调增加的，即：$L(\theta^{(t+1)}) \geq L(\theta^{(t)})$，$t \geq 0$表示第$t$次迭代，这意味着不论对于怎样的初始值，EM算法最终总能获得一个极大值点。

EM的不足之处

- 收敛速度较慢
- 且收敛速度对初始值的选择有较大的依赖性
- EM算法最终总能获得一个局部极大值点

EM算法的具体过程

- 假定 $\phi^{(0)}=(\pi_{1,}^{(0)}\cdots,\pi_{k,}^{(0)},\mu_1^{(0)},\cdots,\mu_k^{(0)},\sigma_1^{2(0)},\cdots,\sigma_k^{2(0)})$ 是初始值，$f(x;\mu,\sigma^2)$ 表示正态分布的密度函数，则在E-步骤中

$$w_{ij}^{(0)}=\Pr(x_i\in G_j\mid x_i;\phi^{(0)})=\frac{\pi_j^{(0)}f(x_i;\mu_j^{(0)},\sigma_j^{2(0)})}{\sum_{t=1}^{k}\pi_t^{(0)}f(x_i;\mu_t^{(0)},\sigma_t^{2(0)})}$$

- 对M-步骤有

$$\pi_j^{(1)}=\frac{1}{n}\sum_{i=1}^{n}w_{ij}^{(0)}$$

$$\mu_j^{(1)}=\frac{\sum_{i=1}^{n}w_{ij}^{(0)}x_i}{n\pi_j^{(0)}}$$

$$\sigma_j^{2(1)}=\frac{\sum_{i=1}^{n}w_{ij}^{(0)}(x_i-\mu_j^{(0)})^2}{n\pi_j^{(0)}}$$

分离世代个体的后验概率

- 后验概率

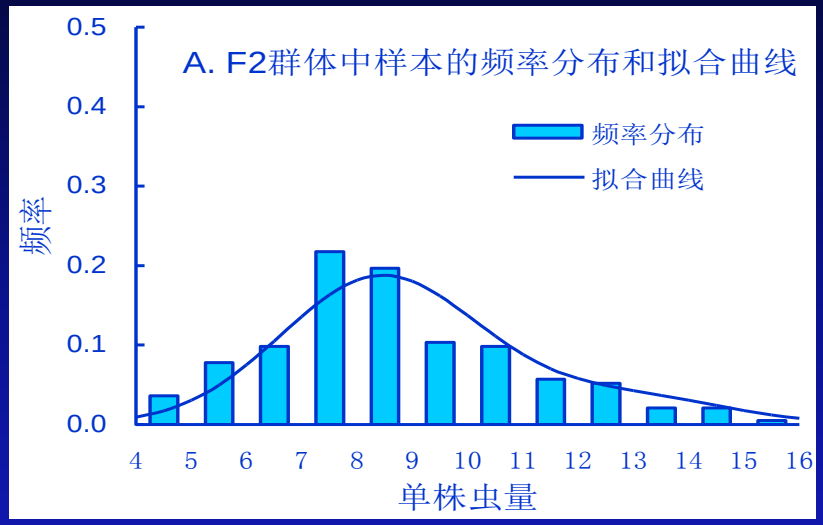
$$w_{ij} = \Pr(x_i \in G_j \mid x_i;\phi) = \frac{\pi_j f(x_i;\mu_j,\sigma_j^2)}{\sum_{t=1}^{k} \pi_t f(x_i;\mu_t,\sigma_t^2)}$$

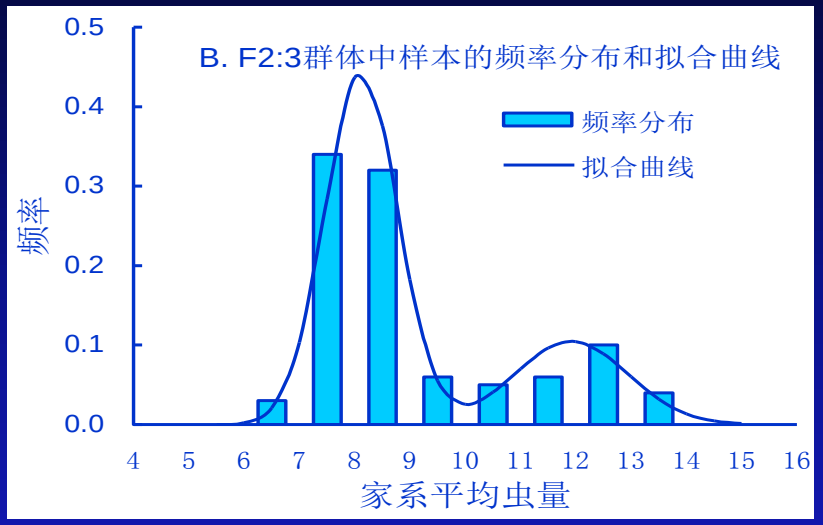
  - 后验概率对主基因型判定过程是一种概率的判定，也就是说存在误判的可能。尤其当两个成分分布重叠严重时，对主基因型判定的可信度将更低。

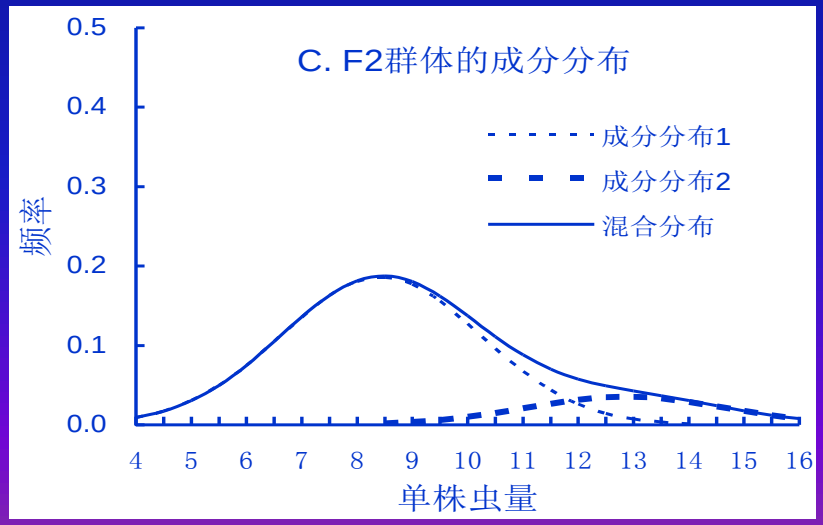
简单分离分析
（Simple segregation analysis）

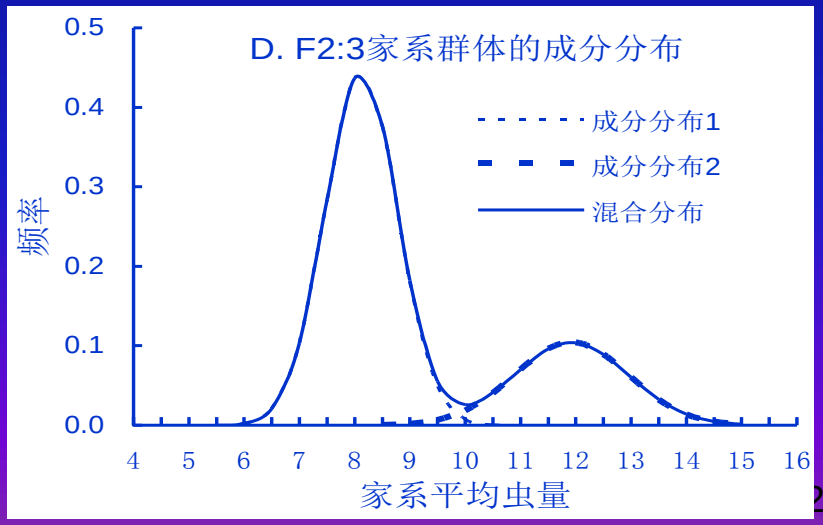
- 将分离世代群体视为混合分布对其进行分解
- 通过比较不同成分分布的AIC值确定成分分布的个数
- 根据选定的成分分布个数估计出的成分分布的权重做分离比的适合性检验
- 根据分离比的适合性检验确定主基因的对数和显隐性关系
- 根据主基因的对数确定其基因型与成分分布的对应关系
- 根据成分分布的均值与主基因的遗传效应间的关系估计主基因的遗传效应

大豆亲本1138-2和邳县天鹅旦衍生的F2和F2:3世代中虫量的频率分布和拟合曲线









不同成分分布个数时的极大似然函数值和AIC值

成分分布数	1	2	3	4
独立参数的个数	2	5	8	11
对数似然函数的极大值	-249.55	-214.40	-212.75	-208.96
AIC 值	503.10	438.80	441.49	439.82

成分分布	成分分布的权重	成分分布的均值	成分分布的方差
1	0.28 (0.04)	12.08 (0.12)	0.44 (0.17)
2	0.72 (0.04)	8.62 (0.09)	0.60 (0.09)

多个分离世代的联合分离分析

- 联合分离分析 (Joint segregation analysis)
- 联合分离分析中所包含的群体及其有关符号

世代	P_1	F_1	P_2	B_1	B_2	F_2
代号	1	2	3	4	5	6
样本容量	n_1	n_2	n_3	n_4	n_5	n_6
样本观测值	x_{1i}	x_{2i}	x_{3i}	x_{4i}	x_{5i}	x_{6i}
群体平均数	μ_1	μ_2	μ_3	μ_4	μ_5	μ_6
群体方差	σ_ε^2	σ_ε^2	σ_ε^2	σ_4^2	σ_5^2	σ_6^2

零（空）模型

- 零模型 (NULL)：各世代的变异完全是随机环境效应引起的，因此每个世代群体都服从相同的正态分布。即

$$\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5 = \mu_6$$

$$\sigma_4^2 = \sigma_5^2 = \sigma_6^2 = \sigma_\varepsilon^2$$

- 零模型可在选择其它模型时作为参照。

一个主基因位点 (A-a) 和多基因的混合遗传模型 (MX1)

- MX1-AD-ADI：加-显性主基因和加-显-位性多基因的混合遗传
- MX1-AD-AD：加-显性主基因和加-显性多基因的混合遗传
- MX1-A-AD：加性主基因和加-显性多基因的混合遗传
- MX1-D-AD：显性主基因和加-显性多基因的混合遗传
- MX1-ND-AD：负向显性主基因和加-显性多基因的混合遗传

一个主基因位点（A-a) 的遗传模型（1MG)

- 在模型MX1中，如果没有多基因的存在，则模型简化为一个基因位点的主基因遗传 (1MG)
- 模型分类：
  - 加-显性主基因模型 (1MG-AD)
  - 加性主基因模型 (1MG-A)
  - 显性主基因模型 (1MG-D)
  - 负向显性主基因模型 (1MG-ND)

两个主基因位点（A-a和B-b）的遗传模型（2MG）

- 加性—显性—上位性模型 (2MG-ADI)
- 加性—显性模型 (2MG-AD)
- 加性模型 (2MG-A)
- 等加性模型 (2MG-EA)
- 完全显性模型 (2MG-D)
- 等完全显性模型 (2MG-ED)

多基因模型（PG）

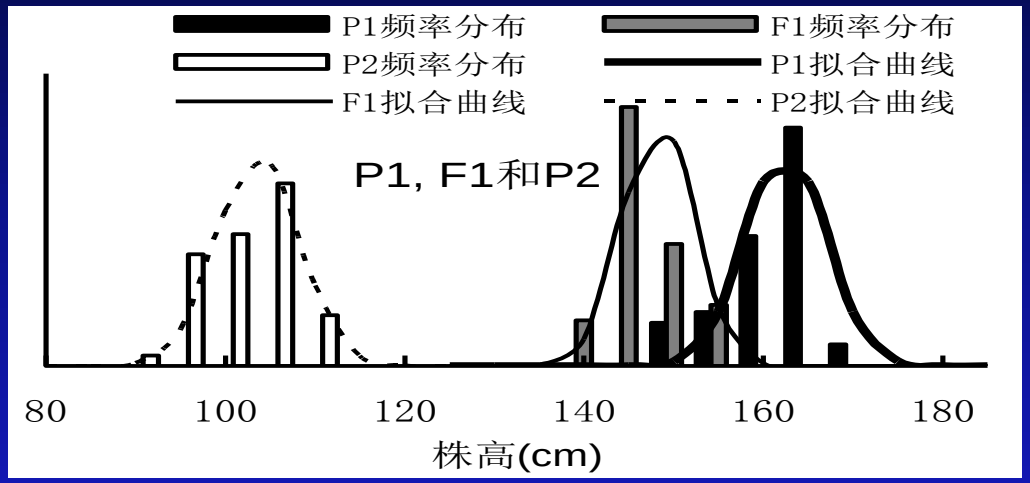
- 若控制数量性状基因的效应较大，则可通过分离分析把它鉴定为主基因。若无主基因存在， 数量性状在分离世代表现为单一正态分布或近似为单一正态分布。

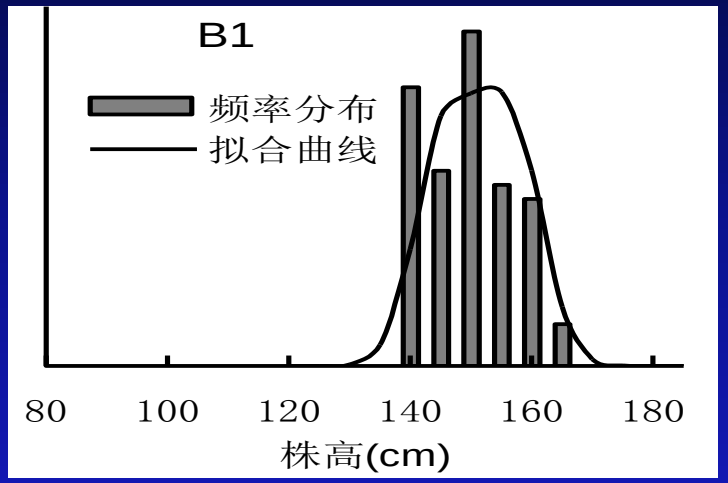
- 加性-显性-上位性多基因模型 (PG-ADI)
- 加性-显性多基因模型 (PG-AD)

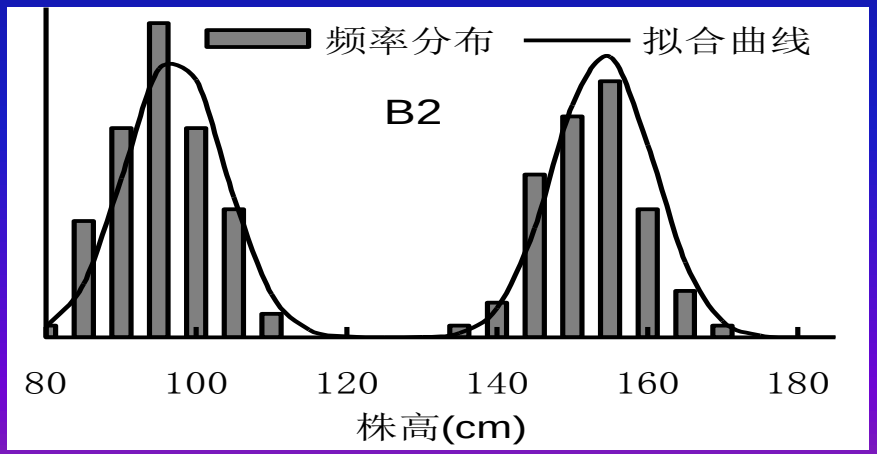
联合分离分析的一般过程

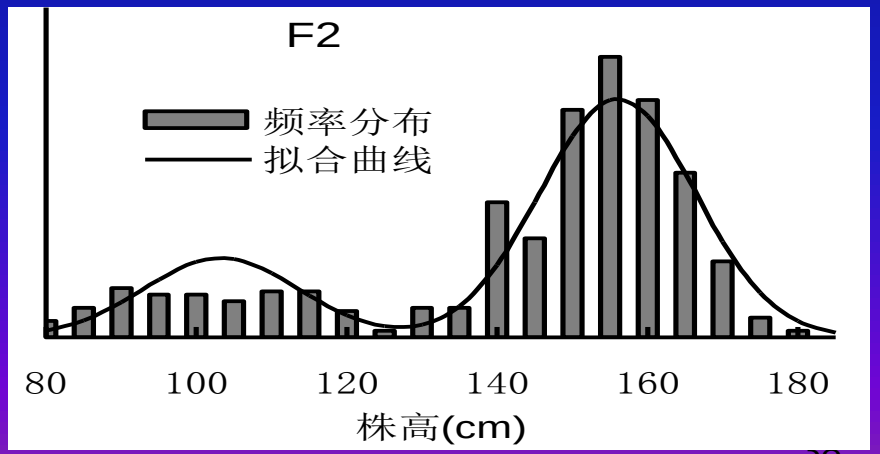
- 假定二倍体核遗传、不存在母体效应、主基因和多基因间无互作和连锁、配子或合子无选择，每一个分离世代的分布是混合分布；
- 设定多种可能的主基因和多基因混合遗传模型，包括 NULL模型和1MG模型，建立各种遗传模型下的极大似然函数；
- 通过EM算法计算出各种可能模型下的成分分布参数及相应的似然函数值和AIC值；
- 根据期望熵最大为最优假定的原则，从各种模型中选出最优模型及其相应的成分分布参数；
- 由最优模型的分布参数估计遗传参数；
- 按Bayes方法计算每一个体属于各种主基因型的后验概率，由概率值大小判别主基因型的归属。

水稻杂交组合南京6号 (P1) ×广丛 (P2) 中各世代的频率分布和拟合曲线









不同模型的AIC值

遗传模型	独立变量个数	最大对数似然函数值	AIC值	按似然函数排序	按AIC值排序
1MG-AD	4	-3596.38	7230.75	MX1	MX1
1MG-A	3	-3984.39	7976.79	MX1-AD-ADI	MX1-AD-ADI
1MG-D	3	-3618.97	7243.94	MX1-A-AD	MX1-A-AD
1MG-ND	3	-4122.60	8251.19	2MG-ADI	2MG-ADI
2MG-ADI	8	-3521.80	7063.61	2MG-AD	2MG-AD
2MG-AD	6	-3538.69	7089.38	1MG-AD	1MG-AD
2MG-A	4	-3983.97	7975.95	2MG-D	1MG-D
2MG-EA	4	-4093.22	8192.44	1MG-D	2MG-D
2MG-D	3	-3618.85	7245.71	PG-ADI	PG-ADI
PG-ADI	7	-3782.60	7583.21	MX1-AD-AD	MX1-AD-AD
PG-AD	6	-3804.39	7620.77	PG-AD	PG-AD
MX1	10	-3487.85	6993.71	2MG-A	2MG-A
MX1-AD-ADI	8	-3512.90	7043.81	1MG-A	1MG-A
MX1-AD-AD	7	-3801.36	7618.73	2MG-EA	2MG-EA
MX1-A-AD	7	-3514.13	7044.26	1MG-ND	1MG-ND
MX1-D-AD	7	-4160.12	8336.24	MX1-D-AD	MX1-D-AD

遗传参数的估计

一阶参数		估 计 值	二阶参数	估计值		
		（cm）		B_1	B_2	F_2
主	a	29.09	σ_p^2	53.11	853.26	563.86
基	d	24.22	σ_{mg}^2	28.15	809.12	456.93
因	d/a	0.83	σ_{pg}^2	5.61	24.79	87.58
多	$\sum a$	0.33	σ_ε^2	19.35	19.35	19.35
基	$\sum d$	-49.25	h_{mg}^2 (%)	53.00	94.82	82.03
因	$\sum aa$	-15.96	h_{pg}^2 (%)	10.56	2.90	15.53
	$\sum ad$	-7.27	成分分布 1	N(156.64, 24.96)	N(154.17, 44.14)	N(156.68, 106.93)
	$\sum dd$	24.41	成分分布 2	N(146.01, 24.96)	N(97.13, 44.14)	N(155.70, 106.93)
			成分分布 3			N(103.25, 106.93)

观测频率分布和理论拟合曲线