判别分析与小麦的两种不同分蘖类型的判别函数

袁 志 发

（西北农学院基础课部）

在《小麦的两种不同分蘖类型的研究》[1]一文中，给出了小麦主茎伸出第n片叶时，第m级分蘖总数的函数及全株各级分蘖总数的函数分别为M（m，n）与N（m，n）：

$$M(m,n)=C^{m}_{n-m-2}=\frac{(n-m-2)}{m(n-2m-2)} \qquad (1)$$

$$N(m,n)=\sum_{m=1}^{[\frac{n}{2}-1]} C^{m}_{n-m-2} \qquad (2)$$

式中$[\frac{n}{2}-1]$代表$\frac{n}{2}$的整数部分。根据上述两个函数算出小麦主茎各个叶龄的各级分蘖数及全株分蘖总数的理论值，然后对1962—1965年四年实际观察结果，按其在出叶与分蘖的关系的实际数值将小麦的分蘖分为两种类型：分蘖的实际数值基本符合或低于理论值的小麦叫主茎型；分蘖的实际数值高于或接近于理论值的小麦叫分蘖型。在实际分类工作中，迂到实际分蘖值显著高于或低于理论值是易于判别的；但对于“基本上符合”与“接近于”理论值就不好办了；况且实际分蘖数值与理论数值在不同年份、不同栽培条件下有很大的变异，这样更对分类工作带来困难。本文运用Fisher两类判别分析方法，对〔1〕中资料进行处理，建立了两类不同分蘖类型的判别函数为

$$y=9.2899x_1-1.0165x_2+0.9818x_3 \qquad (3)$$

式中x_1为主茎伸出第4片叶时的一级分蘖数，x_2为主茎伸出第6片时的一、二级总分蘖数；x_3为主茎伸出第8片叶时的一、二、三、四级的总分蘖数。判别的分介值为y_c或y_o：

$$y_c=16.0144 \qquad y_o=16.5100 \qquad (4)$$

这样可把实际观察到的x_1,x_2,x_3代入（3）式计算出y，若$y<y_c$（或y_o），则为主茎型；若$y>y_c$（或y_o），则为分蘖型。

Fisher两类线性断别分析原理及公式（3）的建立过程如下：

设两类事物混在一起，第一类事物有m_1个，第二类事物有m_2个，共有$M=m'$

本文是在沈煜清教授支持和指导下写成的。作者特表示感谢。

$+m_2$个。每个事物有N个特征，这些特征称为变量。若第i类的第j个个体的第k个变量为x_{ijk}，则对每一个个体可用(x_{ij1}, x_{ij2}, …, x_{ijN})来描述。判别分析的目的是建立某种方法将混杂了的两类事物区分开来。Fisher判别准则是以下列设想为出发点：使两类个体的N个变量经过线性组合形成一个新的综合指标y，每个个体的y值能使两类个体区分的最好。这就要求使两类之间均值之差最大，而使各类内的离差平方和最小。也就是说 Fisher 设想的判别函数具有如下形式：

$$y_{ij}=a_1x_{ij1}+a_2x_{ij2}+\cdots+a_Nx_{ijN}$$

$$=\sum_{k=1}^{N}a_kx_{ijk} \qquad (i=1,2。\quad j=1,2\cdots,m_2) \qquad (5)$$

式中a_k为待定系数，它应使

$$Q=(\bar{y}_1-\bar{y}_2)^2=\text{最大}$$

$$G=\sum_{i=1}^{2}\sum_{j=1}^{m_i}(y_{ij}-\bar{y}_i)^2=\text{最小}$$

即使

$$P=\frac{Q}{G}=\text{最大} \qquad (6)$$

其中 $$\bar{y}_i=\frac{1}{m_i}\sum_{j=1}^{m_i}y_{ij} \qquad (i=1,2) \qquad (7)$$

若令 $$\bar{X}_{ik}=\frac{1}{m_i}\sum_{j=1}^{m_i}x_{ijk} \qquad (i=1,2\quad k=1,2\cdots,N)$$

$$d_k=\bar{X}_{1k}-\bar{X}_{2k} \qquad (k=1,2,\cdots,N)$$

$$S_{kl}=\sum_{i=1}^{2}\sum_{j=1}^{m_i}(X_{ijk}-\bar{X}_{ik})(X_{ijl}-\bar{X}_{il}) \qquad (l,k=1,2,\cdots,N)$$

则

$$Q=(\bar{y}_1-\bar{y}_2)^2=\left[\sum_{k=1}^{N}a_k(\bar{X}_{1k}-\bar{X}_{2k})\right]^2$$

$$=\left[\sum_{k=1}^{N}a_kd_k\right]^2$$

$$=\sum_{k=1}^{N}\sum_{l=1}^{N}a_ka_ld_kd_l$$

$$G=\sum_{i=1}^{2}\sum_{j=1}^{m_i}(y_{ij}-\bar{y}_i)^2$$

$$=\sum_{i=1}^{2}\sum_{j=1}^{m_i}\left[\sum_{k=1}^{N}a_k(X_{ijk}-\bar{X}_{ik})\right]^2$$

$$=\sum_{i=1}^{2}\sum_{j=1}^{m_i}\sum_{k=1}^{N}\sum_{l=1}^{N}a_ka_l(X_{ijk}-\bar{X}_{ik})(X_{ijl}-\bar{X}_{il})$$

$$=\sum_{k=1}^{N}\sum_{l=1}^{N}a_ka_l\left[\sum_{i=1}^{2}\sum_{j=1}^{m_i}(X_{ijk}-\bar{X}_{ik})(X_{ijl}-\bar{X}_{il})\right]$$

$$=\sum_{h=1}^{N}\sum_{l=1}^{N}a_ka_ls_{kl}$$

$$P=\frac{Q}{G}=\frac{\sum\limits_{k=1}^{N}\sum\limits_{l=1}^{N}a_ka_ld_kd_l}{\sum\limits_{k=1}^{N}\sum\limits_{l=1}^{N}a_ka_ls_{kl}} \qquad (8)$$

为了求得使P达到最大的a_k，必须使

$$\frac{\partial P}{\partial a_k}=0 \qquad (k=1,2\cdots,N)$$

即

$$\frac{\partial P}{\partial a_k}=\frac{G\frac{\partial Q}{\partial a_k}-Q\frac{\partial G}{\partial a_k}}{G^2}=0 \qquad (k=1,2\cdots,N)$$

亦即

$$\frac{\partial G}{\partial a_k}=\frac{1}{P}\ \frac{\partial Q}{\partial a_k} \qquad (k=1,2\cdots,N) \qquad (9)$$

因为

$$\begin{aligned}G=\ &a_1^2s_{11}+a_1a_2s_{12}+\cdots+a_1a_Ns_{1N}+\\&a_2a_1s_{21}+a_2^2s_{22}+\cdots+a_2a_Ns_{2N}+\\&\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\\&a_Na_1s_{N1}+a_Na_2s_{N2}+\cdots+a_N^2s_{NN}\end{aligned}$$

由于 $s_{kl}=s_{lk}$，故

$$\frac{\partial G}{\partial a_k}=2(a_1s_{k1}+a_2s_{k2}+\cdots+a_Ns_{kN})$$

$$=2\sum_{l=1}^{N}a_l s_{kl} \qquad (k=1,2\cdots,N)$$

同样可得

$$\frac{\partial Q}{\partial a_k}=2(a_1d_1d_k+a_2d_2d_k+\cdots+a_Nd_Nd_k)$$

$$=2d_k(a_1d_1+a_2d_2+a_Nd_N)$$

$$=2d_k\sum_{l=1}^{N}a_ld_l \qquad (k=1,2,\cdots,N)$$

由(9)则得

$$\sum_{l=1}^{N}a_ls_{kl}=d_k\frac{\sum_{l=1}^{N}a_ld_l}{P} \qquad (k=1,2,\cdots,N) \qquad (10)$$

由于 $\frac{1}{P}\sum_{l=1}^{N}a_ld_l$ 与k无关是个常量，只对解起伸缩作用，并不影响$(a_1,a_2,\cdots a_N)$的方向，故可令其等于(M—2)于是(10)式可写为

$$\sum_{l=1}^{N}a_ls_{kl}=(M-2)d_k \qquad (k=1,2,\cdots,N) \qquad (11)$$

由(11)式解出a_k，则判别函数(5)就构造出来了。

既然判别函数y_{ij}已找出，则对每个个体都可求出y_{ij}值。由于是两类判别，只要对y_{ij}确定一个分介值就可将y_{ij}分为两类。这个分介值为

$$y_c=\frac{m_1\bar{y}_1+m_2\bar{y}_2}{M} \qquad (12)$$

假设 $\bar{y}_1\leqslant\bar{y}_2$，则有

$$\bar{y}\leqslant y_c\leqslant\bar{y}_2$$

故

若 $y_{ij}<y_c$ 则划归第一类

若 $y_{ij}>y_c$ 则划归第二类

再谈一下判别效果的检验问题。若选取马氏距离

$$D^2=a_1d_1+a_2d_2+\cdots a_Nd_N$$

$$=\sum_{k=1}^{N}a_kd_k \qquad (13)$$

为统计量，可证明：

$$F=\frac{m_1m_2(M-N-1)}{M(M-2)N}D^2\sim F(N,M-N-1) \qquad (14)$$

对于给定的信度α，若$F > F_{\alpha}(N, M-N-1)$，则判别是显著的，否则，判别是不显著的。

对于每个变量X_k来讲，其在分类中影响的大小可用：

$$a_k d_k / D' \times 100\% \qquad (k=1,2\cdots,N) \qquad (15)$$

来表示。

基于以上理论，用于〔1〕中资料及判别结果如表1

表1 （$y=9.2899X_1-1.0165X_2+0.9818X_3$，$y_c=16.0114$，$y_0=16.5100$）

类别（i）		X_{ij1}	X_{ij2}	X_{ij3}	y_{ij}	y_c判别分类	y_0判别分类
第一类（主茎型）（$i=1$）	1	0.71	3.8	12.0	14.5147	1	1
	2	0.78	3.86	12.17	15.2709	1	1
	3	1.0	2.1	5.7	12.7515	1	1
	4	0.7	1.7	5.9	10.5675	1	1
	5	0.3	1.8	6.1	6.9463	1	1
	6	0.6	3.4	10.2	12.1322	1	1
	7	1.0	3.6	10.2	15.6449	1	1
	8	0.5	3.5	10.5	11.3961	1	1
	9	0.5	5.0	11.5	10.8532	1	1
	10	0.71	4.0	11.25	13.5751	1	1
	11	1.0	4.5	12.0	16.4973	2	1
	均值$\bar{X}_{1k}$	0.7091	3.3873	9.7745	$\bar{y}_1=12.7409$		
第二类（分蘖型）（$i=2$）	1	1.0	4.25	15.16	19.8539	2	2
	2	1.0	3.43	16.25	21.7576	2	2
	3	1.0	3.7	11.4	16.7214	2	2
	4	1.0	3.8	12.4	17.6015	2	2
	5	1.0	4.0	13.6	18.5764	2	2
	6	1.0	4.0	12.8	17.7909	2	2
	7	1.0	4.2	13.4	18.1767	2	2
	8	1.0	4.3	14.0	18.6642	2	2
	9	1.0	5.7	15.8	19.0083	2	2
	10	1.0	4.7	20.4	24.5411	2	2
	11	0.8	4.6	14.0	16.5112	2	2
	12	1.0	4.56	14.6	18.9889	2	2
	均值$\bar{X}_{2k}$	0.9833	3.8262	14.484	$\bar{y}_2=19.0152$		

由表1知：$m_1=11$，$m_2=12$，$M=23$

$$d_1=-0.2742 \qquad (M-2)d_1=-5.7582$$
$$d_2=-0.8827 \qquad (M-2)d_2=-18.5367$$
$$d_3=-4.7095 \qquad (M-2)d_3=-98.8995$$

然后计算各个协方差，即：

$$S_{kl}=\sum_{j=1}^{m_1}X_{1jk}X_{1jl}-\left(\sum_{j=1}^{m_1}X_{1jk}\right)\left(\sum_{j=1}^{m_1}X_{1jl}\right)/m_1+\sum_{j=1}^{m_2}X_{2jk}X_{2jl}$$
$$-\left(\sum_{j=1}^{m_2}X_{2jk}\right)\left(\sum_{j=1}^{m_2}X_{2jl}\right)/m_2 \qquad (k、l=1,2,3)$$

$$S_{11}=0.5624,\quad S_{22}=15.5160,\quad S_{33}=126.2374,$$
$$S_{12}=0.2821,\quad S_{13}=0.8355,\quad S_{23}=32.2744$$

由（11）式可得：

$$\left.\begin{aligned}0.5624a_1+0.2821a_2+0.8355a_3&=-5.7582\\0.2821a_1+15.5160a_2+32.2744a_3&=-18.5367\\0.8355a_1+32.2744a_2+126.2374a_3&=-98.8995\end{aligned}\right\}$$

解之得：

$$a_1=-9.2899,\quad a_2=1.0165,\quad a_3=-0.9818$$
$$D^2=a_1d_1+a_2d_2+a_3d_3=6.274$$

$$F=\frac{(11)(12)(23-3-1)}{23(23-2)(3)}D^2=10.8593$$

$$F_{0.01}(3,\ 19)=5.01$$

即判别效果是极显著的。由上面计算知判别函数为：

$$y=9.2899X_1-1.0165X_2+0.98184X_2$$

（两边已乘过（-1））

而 $$y_c=\frac{11\bar{y}_1+12\bar{y}_2}{23}=16.0144$$

由表1可看出，23个个体中，判错一个，其概率为$\frac{1}{23}=4\%$，与检验结果是吻合的。

各变量在判别中的影响如下：

X_1	40.6%
X_2	-14.3%
X_3	73.7%

可见变异以X_3为最，X_1次之，X_2又次之。

为了使判别更适合小麦分蘖的实际情况，取各变量的理论值：$X_1=1$， $X_2=4$

$X_3=12$，则其判别值为：

$$y_{理}=a_1+4a_2+12a_3=17.0055$$

再考虑到 y 的变异性，可取分介值：

$$y_0=\frac{y_{理}+y_c}{2}=16.5100$$

则判别效果更佳，其结果如表1中最后一栏。

利用上述判别函数对〔1〕文的原始资料的45个个体进行判别，其结果如表2。判别结果表明西农6028并不是主茎型，起码不是典型的主茎型。除此而外，y_c的判别不符率为6.7%；y_0的判别不符率为4.4%。

表2　1962—1965年四青资料的小麦分蘖类型判别结果（$y_c=16.0144$　$y_0=16.5100$　$y=9.2899X_1-1.0165X_2+0.9818X_3$）

			X_1	X_2	X_3	y	y_c分类	y_0分类	原分类
南大 2149	1962	20/9	0.5	3.5	10.5	11.3961	主	主	主
57（158）	1962	20/9	0.5	5	11.5	10.8532	主	主	主
西农（6028）	1962	20/9	1	4	13.2	18.18370	分△	分△	主
早洋麦	1962	20/9	1	4.3	14	18.6642	分	分	分
3102 大麦	1962	20/9	1	5.7	15.8	19.0083	分	分	分
武功黑麦	1962	20/9	1	4.7	20.4	24.5411	分	分	分
南大 2419	1963	26/9	0.71	3.8	12	14.5147	主	主	主
南大 2419	1963	1/10	0.71	4.0	11.25	13.5751	主	主	主
57（13）1—3	1963	26/9	0.78	3.86	12.17	15.2709	主	主	主
57（13）1—3	1963	1/10	1	4.5	12	16.4973	分△	主	主
西农（6028）	1963	26/9	0.8	4	13.25	16.3748	分△	主	主
西农（6028）	1963	1/10	1	4	11.67	16.6815	分△	分△	主
早洋麦	1963	26/9	1	4.25	15.16	19.8539	分	分	分
早洋麦	1963	1/10	0.80	4.60	14	16.5112	分	分	分
甘肃 96号	1963	26/9	1	3.43	16.25	21.7576	分	分	分
甘肃 96号	1963	1/10	1	4.56	14.6	18.9899	分	分	分
58（60）12—1	1964	20/9	0	1.6	4.5		主	主	主
阿天	1964	20/9	0	1.9	5.6	-	主	主	主
南大 2419	1964	20/9	1	2.1	5.7	12.7515	主	主	主
57（13）1—3	1964	20/9	0.7	1.7	5.9		主	主	主
西农 6028	1964	20/9	0	1.7	6.8		主	主	主

丰产1号 1964	1964	20/9	0.3	1.8	6.1		主	主	主
内乡 36 号	1964	20/9	0	2.2	6.9		主	主	主
碧玉麦	1964	20/9	0.8	2.1	5.8		主	主	主
安徽1号	1964	20/9	0.4	2.0	6.4		主	主	主
尤皮2号	1964	20/9	0	2.2	9		主	主	主
碧蚂1号	1964	20/9	0.9	2.7	9.3		主	主	主
陕农9号	1964	20/9	0	2.2	100		主	主	主
南京 2095	1964	20/9	0.1	3.3	10.2		主	主	主
徐州 483	1963	20/9	1	3.8	11.3	16.52154	分	分	分
银交麦	1964	20/9	1.9	4.8	10.9	23.4732	分	分	分
乌克兰	1964	20/9	0.7	3.8	11.2	13.6364	主△	主△	分
早洋麦	1964	20/9	1	3.7	11.4	16.7214	分	分	分
陕农1号	1964	20/9	1	3.8	12.4	17.6015	分	分	分
早红玉	1964	20/9	1	3.7	12.7	17.9977	分	分	分
甘肃 96号	1964	20/9	1	4	13.6		分	分	分
阿夫	1965	25/9	1	2.8	9.4		主	主	主
南大 2419	1965	25/0	0	3.0	9.5		主	主	主
57（13）1—3	1965	25/9	0.6	3.4	10.2		主	主	主
丰产1号	1965	25/9	1	3.6	10.2		主	主	主
尤皮2号	1965	25/9	1	4	10.8		主	主	主
陕农9号	1965	25/9	0.8	3.5	11.8		主	主	主
58（28）1—5	1965	25/9	0	3.8	13.2		主△	主△	分
甘肃 96 号	1965	25/9	1	4.2	13.4	18.1767	分	分	分
乌克兰 0246	1965	25/9	1	4	12.8	17.7909	分	分	分

（注：表2中△为不符号；“主”表示主茎型；“分”表示分蘖型。）

参 考 文 献

〔1〕 沈煜清等《小麦的两种不同分蘖类型的研究》 中国农业科学 1978．3期。

〔2〕 南开大学数学系统计予报组：概率与统计予报及在地震与气象中的应用 科学出版社 1978。