重慶工商大學
CHONGQING TECHNOLOGY AND BUSINESS UNIVERSITY

判别与分类

DISCRIMINATION AND CLASSIFICATION

数学与统计学院 杨炜明

什么是判别分析

在我们的日常生活和工作实践中，常常会遇到判别分析问题，即根据历史上划分类别的有关资料和某种最优准则，确定一种判别方法，判定一个新的样本归属哪一类。例如，某医院有部分患有肺炎、肝炎、冠心病、糖尿病等病人的资料，记录了每个患者若干项症状指标数据。现在想利用现有的这些资料找出一种方法，使得对于一个新的病人，当测得这些症状指标数据时，能够判定其患有哪种病。又如，在天气预报中，我们有一段较长时间关于某地区每天气象的记录资料（晴阴雨、气温、气压、湿度等），现在想建立一种用连续五天的气象资料来预报第六天是什么天气的方法。这些问题都可以应用判别分析方法予以解决。

- 把这类问题用数学语言来表达，可以叙述如下：设有n个样本，对每个样本测得p项指标（变量）的数据，已知每个样本属于k个类别（或总体）G_1，G_2，…，G_k中的某一类，且它们的分布函数分别为$F_1(x)$，$F_2(x)$，…，$F_k(x)$。我们希望利用这些数据，找出一种判别函数，使得这一函数具有某种最优性质，能把属于不同类别的样本点尽可能地区别开来，并对测得同样p项指标（变量）数据的一个新样本，能判定这个样本归属于哪一类。

重庆工商大学
CHONGQING TECHNOLOGY AND BUSINESS UNIVERSITY

判别分析内容很丰富，方法很多。判断分析按判别的总体数来区分，有两个总体判别分析和多总体判别分析；按区分不同总体所用的数学模型来分，有线性判别和非线性判别；按判别时所处理的变量方法不同，有逐步判别和序贯判别等。判别分析可以从不同角度提出问题，因此有不同的判别准则，如马氏距离最小准则、Fisher准则、平均损失最小准则、最小平方准则、最大似然准则、最大概率准则等等，按判别准则的不同又提出多种判别方法。本章仅介绍常用的几种判别分析方法：距离判别法、Fisher判别法、Bayes判别法和逐步判别法。

例 中小企业的破产模型

为了研究中小企业的破产模型，选定4个经济指标：

X_1总负债率（现金收益/总负债）

X_2收益性指标（纯收入/总财产）

X_3短期支付能力（流动资产/流动负债）

X_4生产效率性指标（流动资产/纯销售额）

对17个破产企业（1类）和21个正常运行企业（2类）进行了调查，得如下资料：

总负债率	收益性指标	短期支付能力	生产效率指标	类别
-0.45	-0.41	1.09	0.45	1
-0.56	-0.31	1.51	0.16	1
0.06	0.02	1.01	0.4	1
-0.07	-0.09	1.45	0.26	1
-0.1	-0.09	1.56	0.67	1
-0.14	-0.07	0.71	0.28	1
-0.23	-0.3	0.22	0.18	1
0.07	0.02	1.31	0.25	1
0.01	0	2.15	0.7	1
-0.28	-0.23	1.19	0.66	1
0.15	0.05	1.88	0.27	1
0.37	0.11	1.99	0.38	1
-0.08	-0.08	1.51	0.42	1
0.05	0.03	1.68	0.95	1
0.01	0	1.26	0.6	1
0.12	0.11	1.14	0.17	1
-0.28	-0.27	1.27	0.51	1

0.51	0.1	2.49	0.54	2
0.08	0.02	2.01	0.53	2
0.38	0.11	3.27	0.55	2
0.19	0.05	2.25	0.33	2
0.32	0.07	4.24	0.63	2
0.31	0.05	4.45	0.69	2
0.12	0.05	2.52	0.69	2
-0.02	0.02	2.05	0.35	2
0.22	0.08	2.35	0.4	2
0.17	0.07	1.8	0.52	2
0.15	0.05	2.17	0.55	2
-0.1	-1.01	2.5	0.58	2
0.14	-0.03	0.46	0.26	2
0.14	0.07	2.61	0.52	2
-0.33	-0.09	3.01	0.47	2
0.48	0.09	1.24	0.18	2
0.56	0.11	4.29	0.45	2
0.2	0.08	1.99	0.3	2
0.47	0.14	2.92	0.45	2
0.17	0.04	2.45	0.14	2
0.58	0.04	5.06	0.13	2

0. 04	0. 01	1. 5	0. 71	未知
-0. 06	-0. 06	1. 37	0. 4	未知
0. 07	-0. 01	1. 37	0. 34	未知
-0. 13	-0. 14	1. 42	0. 44	未知
0. 15	0. 06	2. 23	0. 56	未知
0. 16	0. 05	2. 31	0. 2	未知
0. 29	0. 06	1. 84	0. 38	未知
0. 54	0. 11	2. 33	0. 48	未知

距离判别法

一、马氏距离的概念

二、距离判别的思想及方法

三、判别分析的实质

一、马氏距离的概念

设 p 维欧氏空间 R^p 中的两点 $\mathbf{X}=(X_1,X_2,\cdots,X_p)'$ 和 $\mathbf{Y}=(Y_1,Y_2,\cdots,Y_p)'$，通常我们所说的两点之间的距离，是指欧氏距离，即

$$d^2(\mathbf{X},\mathbf{Y})=(X_1-Y_1)^2+\cdots+(X_p-Y_p)^2$$

在解决实际问题时，特别是针对多元数据的分析问题，欧氏距离就显示出了它的薄弱环节。

第一、设有两个正态总体，$X\sim N(\mu_1,\sigma^2)$ 和 $Y\sim N(\mu_2,4\sigma^2)$，现有一个样品位于如图 4.1 所示的 A 点，距总体 X 的中心 2σ 远，距总体 Y 的中心 3σ 远，那么，A 点处的样品到底离哪一个总体近呢？若按欧氏距离来量度，A 点离总体 X 要比离总体 Y “近一些”。但是，从概率的角度看，A 点位于 μ_1 右侧的 $2\sigma_x$ 处，而位于 μ_2 左侧 $1.5\sigma_y$ 处，应该认为 A 点离总体 Y “近一些”。显然，后一种量度更合理些。

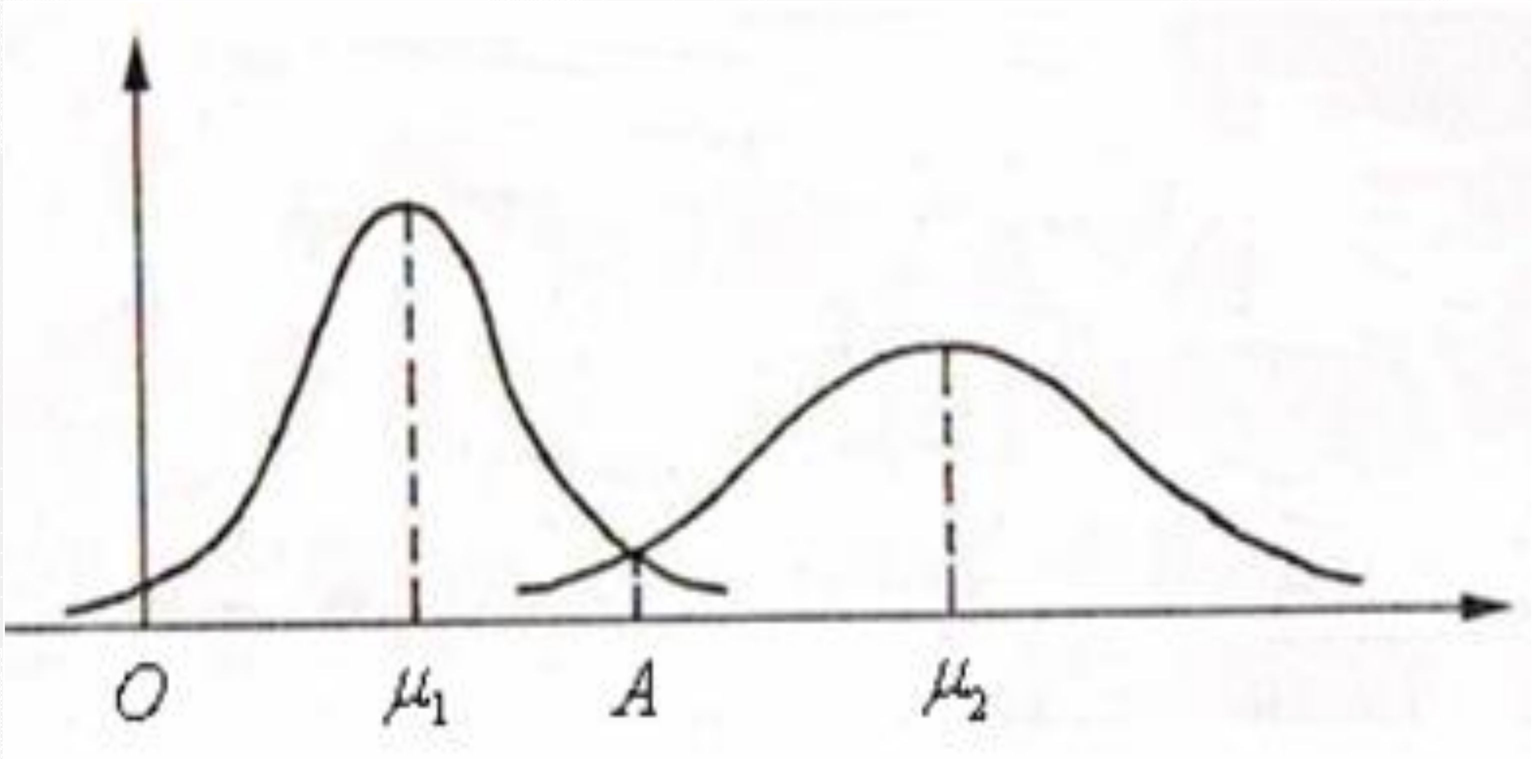


图4.1

第二、设有量度重量和长度的两个变量 X 与 Y，以单位分别为 kg 和 cm 得到样本 $A(0,5)$， $B(10,0)$， $C(1,0)$， $D(0,10)$。今按照欧氏距离计算，有

$$AB=\sqrt{10^2+5^2}=\sqrt{125};$$

$$CD=\sqrt{1^2+10^2}=\sqrt{101}$$

如果我们将长度单位变为 mm，那么，有

$$AB=\sqrt{10^2+50^2}=\sqrt{2600};$$

$$CD=\sqrt{1^2+100^2}=\sqrt{10001}$$

量纲的变化，将影响欧氏距离计算的结果。

- 为此，我们引入一种由印度著名统计学家马哈拉诺比斯（Mahalanobis, 1936）提出的“马氏距离”的概念。

- **设 $\mathbf{X}$ 和 $\mathbf{Y}$ 是来自均值向量为 $\boldsymbol{\mu}$，协方差为 $\boldsymbol{\Sigma}(>0)$ 的总体 G 中的 p 维样本，则总体 G 内两点 $\mathbf{X}$ 与 $\mathbf{Y}$ 之间的马氏距离定义为**

$$D^2(\mathbf{X},\mathbf{Y})=(\mathbf{X}-\mathbf{Y})'\boldsymbol{\Sigma}^{-1}(\mathbf{X}-\mathbf{Y}) \qquad \textbf{(4.2)}$$

定义点 $\mathbf{X}$ 到总体 G 的马氏距离为

$$D^2(\mathbf{X},G)=(\mathbf{X}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{X}-\boldsymbol{\mu}) \qquad \textbf{(4.3)}$$

这里应该注意到，当 $\boldsymbol{\Sigma}=\mathbf{I}$（单位矩阵）时，即为欧氏距离的情形。

重庆工商大学 CHONGQING TECHNOLOGY AND BUSINESS UNIVERSITY

例：

$$\mu_1=\begin{bmatrix}0\\0\end{bmatrix},\quad \Sigma_1=\begin{bmatrix}1&0\\0&1\end{bmatrix},\quad x_1=\begin{bmatrix}1\\1\end{bmatrix}$$

$$d^2{}_{x_1u_1}=\begin{bmatrix}1&1\end{bmatrix}\begin{bmatrix}1&0\\0&1\end{bmatrix}^{-1}\begin{bmatrix}1\\1\end{bmatrix}=2$$

$$\mu_2=\begin{bmatrix}0\\0\end{bmatrix},\quad \Sigma_2=\begin{bmatrix}1&0\\0&100\end{bmatrix},\quad x_2=\begin{bmatrix}1\\10\end{bmatrix}$$

$$d^2{}_{x_2u_2}=\begin{bmatrix}1&10\end{bmatrix}\begin{bmatrix}1&0\\0&100\end{bmatrix}^{-1}\begin{bmatrix}1\\10\end{bmatrix}=2$$

$$\mu_3=\begin{bmatrix}0\\0\end{bmatrix},\quad \Sigma_3=\begin{bmatrix}1 & 0.5\\0.5 & 1\end{bmatrix},\quad x_3=\begin{bmatrix}1\\1\end{bmatrix}$$

$$d^2{}_{x_3u_3}=\begin{bmatrix}1 & 1\end{bmatrix}\begin{bmatrix}1 & 0.5\\0.5 & 1\end{bmatrix}^{-1}\begin{bmatrix}1\\1\end{bmatrix}=1.3333$$

$$\mu_4=\begin{bmatrix}0\\0\end{bmatrix},\quad \Sigma_4=\begin{bmatrix}1 & 0.9\\0.9 & 1\end{bmatrix},\quad x_4=\begin{bmatrix}1\\1\end{bmatrix}$$

$$d^2{}_{x_4u_4}=\begin{bmatrix}1 & 1\end{bmatrix}\begin{bmatrix}1 & 0.9\\0.9 & 1\end{bmatrix}^{-1}\begin{bmatrix}1\\1\end{bmatrix}=1.0526$$

17个破产企业4项指标的均值和协方差

$$\mu_1 = \begin{bmatrix} -0.07941 \\ -0.08882 \\ 1.34882 \\ 0.43000 \end{bmatrix} \quad \Sigma_1 = \begin{bmatrix} 0.0527808824 & 0.0340117647 & 0.0433257353 & 0.0040562500 \\ 0.0340117647 & 0.0251110294 & 0.0333014706 & 0.0031437500 \\ 0.0433257353 & 0.0333014706 & 0.2187610294 & 0.0403875000 \\ 0.0040562500 & 0.0031437500 & 0.0403875000 & 0.0503375000 \end{bmatrix}$$

待判点到此类的距离

$$x = \begin{bmatrix} 0.04 \\ 0.01 \\ 1.5 \\ 0.71 \end{bmatrix}$$

$$d_1^2 = (\mathbf{x} - \mu_1)' \Sigma_1^{-1} (\mathbf{x} - \mu_1) = 2.0618$$

重慶工商大學
CHONGQING TECHNOLOGY AND BUSINESS UNIVERSITY

21个正常企业4项指标的均值和协方差

$$\mu_2=\begin{bmatrix}0.22571\\0.00524\\2.67286\\0.44095\end{bmatrix}\quad \Sigma_2=\begin{bmatrix}0.050925714 & 0.025623571 & 0.094172857 & -0.007260714\\0.025623571 & 0.056596190 & 0.020054286 & -0.006345238\\0.094172857 & 0.020054286 & 1.207031429 & 0.043737143\\-0.007260714 & -0.006345238 & 0.043737143 & 0.028219048\end{bmatrix}$$

待判点到此类的距离 $x=\begin{bmatrix}0.04\\0.01\\1.5\\0.71\end{bmatrix}$

$$d_2^2=(\mathbf{x}-\mu_2)'\Sigma_2^{-1}(\mathbf{x}-\mu_2)=4.998$$

如果假定它们有相同的协方差

$$\Sigma=\begin{bmatrix} 0.073991252 & 0.035845804 & 0.172219346 & -0.001322119 \\ 0.035845804 & 0.043697866 & 0.056863158 & -0.001808819 \\ 0.172219346 & 0.056863158 & 1.192172688 & 0.044788620 \\ -0.001322119 & -0.001808819 & 0.044788620 & 0.037051565 \end{bmatrix}$$

$$d_1^2=(\mathbf{x}-\mu_1)'\Sigma^{-1}(\mathbf{x}-\mu_1)=2.7163$$

$$d_2^2=(\mathbf{x}-\mu_2)'\Sigma^{-1}(\mathbf{x}-\mu_2)=4.155$$

二、距离判别的思想及方法

1、两个总体的距离判别问题

- 问题：设有协方差矩阵$\sum$相等的两个总体G_1和G_2，其均值分别是μ_1和μ_2，对于一个新的样品X，要判断它来自哪个总体。
- 一般的想法是计算新样品 X 到两个总体的马氏距离 D^2（X，G_1）和D^2（X，G_2），并按照如下的判别规则进行判断

$$\begin{cases} \mathbf{X} \in G_1, & \text{如果} \quad D^2(\mathbf{X}, G_1) \le D^2(\mathbf{X}, G_2) \\ \mathbf{X} \in G_2, & \text{如果} \quad D^2(\mathbf{X}, G_1) > D^2(\mathbf{X}, G_2) \end{cases}$$

- 这个判别规则的等价描述为：求新样品X到G_1的距离与到G_2的距离之差，如果其值为正，X属于G_2；否则X属于G_1。

- 我们考虑

$$
\begin{aligned}
&D^2(\mathbf{X},G_1)-D^2(\mathbf{X},G_2)\\
&=(\mathbf{X}-\boldsymbol{\mu}_1)'\boldsymbol{\Sigma}^{-1}(\mathbf{X}-\boldsymbol{\mu}_1)-(\mathbf{X}-\boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}(\mathbf{X}-\boldsymbol{\mu}_2)\\
&=\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X}-2\mathbf{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1+\boldsymbol{\mu}_1'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1-(\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X}-2\mathbf{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_2+\boldsymbol{\mu}_2'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_2)\\
&=2\mathbf{X}'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_2-\boldsymbol{\mu}_1)+\boldsymbol{\mu}_1'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1-\boldsymbol{\mu}_2'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_2\\
&=2\mathbf{X}'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_2-\boldsymbol{\mu}_1)+(\boldsymbol{\mu}_1+\boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)\\
&=-2\left(\mathbf{X}-\frac{\boldsymbol{\mu}_1+\boldsymbol{\mu}_2}{2}\right)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)\\
&=-2(\mathbf{X}-\bar{\boldsymbol{\mu}})'\boldsymbol{\alpha}=-2\boldsymbol{\alpha}'(\mathbf{X}-\bar{\boldsymbol{\mu}})
\end{aligned}
$$

其中 $\bar{\boldsymbol{\mu}}=\frac{1}{2}(\boldsymbol{\mu}_1+\boldsymbol{\mu}_2)$ 是两个总体均值的平均值，$\boldsymbol{\alpha}=\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)$，记 $W(\mathbf{X})=\boldsymbol{\alpha}'(\mathbf{X}-\bar{\boldsymbol{\mu}})$

则判别规则可表示为

$$\begin{cases}\mathbf{X}\in G_1, & \text{如果} \quad W(\mathbf{X})\geq 0\\ \mathbf{X}\in G_2, & \text{如果} \quad W(\mathbf{X})<0\end{cases}$$

这里称 $W(\mathbf{X})$ 为两总体距离判别的判别函数，由于它是 $\mathbf{X}$ 的线性函数，故又称为线性判别函数，$\boldsymbol{\alpha}$ 称为判别系数。

重慶工商大學
CHONGQING TECHNOLOGY AND BUSINESS UNIVERSITY

例　在企业的考核中，可以根据企业的生产经营情况把企业分为优秀企业和一般企业。考核企业经营状况的指标有：

资金利润率=利润总额/资金占用总额

劳动生产率=总产值/职工平均人数

产品净值率=净产值/总产值

三个指标的均值向量和协方差矩阵如下。现有二个企业，观测值分别为

（7.8，39.1，9.6）和（8.1，34.2，6.9），问这两个企业应该属于哪一类？

变量	均值向量		协方差矩阵		
	优秀	一般			
资金利润率	**13.5**	**5.4**	**68.39**	**40.24**	**21.41**
劳动生产率	**40.7**	**29.8**	**40.24**	**54.58**	**11.67**
产品净值率	**10.7**	**6.2**	**21.41**	**11.67**	**7.90**

$$\Sigma^{-1}=\begin{bmatrix}0.119337 & -0.02753 & -0.28276\\ -0.02753 & 0.033129 & 0.025659\\ -0.28276 & 0.025659 & 0.854988\end{bmatrix}$$

$$\mu_1 - \mu_2 = \begin{bmatrix} 8.1 \\ 10.9 \\ 4.5 \end{bmatrix} \qquad (\mu_1 + \mu_2)/2 = \begin{bmatrix} 9.45 \\ 35.25 \\ 8.45 \end{bmatrix}$$

$$\text{判别函数的系数}\Sigma^{-1}(\mu_1 - \mu_2) = \begin{bmatrix} -0.60581 \\ 0.25362 \\ 1.83679 \end{bmatrix}$$

$$\text{判别函数的常数项}\ \left(\frac{\mu_1 + \mu_2}{2}\right)' \Sigma^{-1}(\mu_1 - \mu_2)$$

$$= \begin{bmatrix} 9.45 & 35.25 & 8.45 \end{bmatrix} \begin{bmatrix} -0.60581 \\ 0.25362 \\ 1.83679 \end{bmatrix} = 18.73596$$

线性判别函数：

$$y = -0.60581x_1 + 0.25362x_2 + 1.83679x_3 - 18.73596$$

$$y_1 = -0.60581\times 7.8 + 0.25362\times 39.1 + 1.83679\times 9.6 - 18.73596$$
$$= 4.0892 > 0\text{(第一个新企业属于一类）}$$

$$y_2 = -0.60581\times 8.1 + 0.25362\times 34.2 + 1.83679\times 6.9 - 18.73596$$
$$= -2.2956 < 0\text{(第二个新企业属于二类）}$$

在实际应用中，总体的均值和协方差矩阵一般是未知的，可由样本均值和样本协方差矩阵分别进行估计。

设 $\mathbf{X}_1^{(1)},\cdots,\mathbf{X}_{n_1}^{(1)}$ 来自总体 G_1 的样本，$\mathbf{X}_1^{(2)},\cdots,\mathbf{X}_{n_2}^{(2)}$ 是来自总体 G_2 的样本，μ_1 和 μ_2 的一个无偏估计分别为

$$\bar{\mathbf{X}}^{(1)}=\frac{1}{n_1}\sum_{i=1}^{n_1}\mathbf{X}_i^{(1)} \quad 和 \quad \bar{\mathbf{X}}^{(2)}=\frac{1}{n_2}\sum_{i=1}^{n_2}\mathbf{X}_i^{(2)}$$

$\mathbf{\Sigma}$ 的一个联合无偏估计为 $\hat{\mathbf{\Sigma}}=\dfrac{1}{n_1+n_2-2}(\mathbf{S}_1+\mathbf{S}_2)$

这里 $\mathbf{S}_\alpha=\sum_{i=1}^{n_\alpha}(\mathbf{X}_i^{(\alpha)}-\bar{\mathbf{X}}^{(\alpha)})(\mathbf{X}_i^{(\alpha)}-\bar{\mathbf{X}}^{(\alpha)})', \quad \alpha=1,2$

此时，两总体距离判别的判别函数为 $\hat{W}(\mathbf{X})=\hat{\boldsymbol{\alpha}}'(\mathbf{X}-\bar{\mathbf{X}})$

其中 $\bar{\mathbf{X}}=\dfrac{1}{2}(\bar{\mathbf{X}}^{(1)}+\bar{\mathbf{X}}^{(2)})$，$\hat{\boldsymbol{\alpha}}=\hat{\mathbf{\Sigma}}^{-1}(\bar{\mathbf{X}}^{(1)}-\bar{\mathbf{X}}^{(2)})$。这样，判别规则为

$$\begin{cases}\mathbf{X}\in G_1, & 如果 \quad \hat{W}(\mathbf{X})\ge 0\\ \mathbf{X}\in G_2, & 如果 \quad \hat{W}(\mathbf{X})<0\end{cases} \qquad (4.7)$$

- 这里我们应该注意到:

（1） 当 $p=1$， G_1 和 G_2 的分布分别为 $N(\mu_1,\sigma^2)$ 和 $N(\mu_2,\sigma^2)$ 时， μ_1,μ_2,σ^2 均为已知，且 $\mu_1<\mu_2$，则判别系数为 $\alpha=\dfrac{\mu_1-\mu_2}{\sigma^2}<0$，判别函数为

$$W(x)=\alpha(x-\bar{\mu})$$

判别规则为

$$\begin{cases} x\in G_1, & \text{如果} \quad x\le\bar{\mu} \\ x\in G_2, & \text{如果} \quad x>\bar{\mu} \end{cases}$$

（2） 当$\boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$，$\boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2$时，我们采用下式作为判别规则的形式。选择判别函数为

$$\begin{aligned} W^*(\mathbf{X}) &= D^2(\mathbf{X}, G_1) - D^2(\mathbf{X}, G_2) \\ &= (\mathbf{X}-\boldsymbol{\mu}_1)'\boldsymbol{\Sigma}_1^{-1}(\mathbf{X}-\boldsymbol{\mu}_1) - (\mathbf{X}-\boldsymbol{\mu}_2)'\boldsymbol{\Sigma}_2^{-1}(\mathbf{X}-\boldsymbol{\mu}_2) \end{aligned}$$

它是 $\mathbf{X}$ 的二次函数，相应的判别规则为

$$\begin{cases} \mathbf{X} \in G_1, & \text{如果} \quad W^*(\mathbf{X}) \leq 0 \\ \mathbf{X} \in G_2, & \text{如果} \quad W^*(\mathbf{X}) > 0 \end{cases}$$

例1:

	x1	x2	x3
高水平	76	99	5374
	79.5	99	5359
	78	99	5372
	72.1	95.9	5242
	73.8	77	5370
中水平	71.2	93	4250
	75.3	94.9	3412
	70	91.2	3390
	72.8	99	2300
	62.9	80.6	3799
待判	68.5	79.3	1950
	69.9	96.9	2840
	77.6	93.8	5233
	69.3	90.3	5158

变量	均值向量		协方差矩阵的估计		
	高	中			
x1	75.88	70.44	153.8	218.95	-5558.75
x2	93.98	91.74	218.95	695.35	-14484.25
x3	5343.4	3430.2	-5558.75	-14484.25	2625465

$$\mu_1 - \mu_2 = \begin{bmatrix} 5.44 \\ 2.51 \\ 1913.2 \end{bmatrix} \qquad (\mu_1 + \mu_2)/2 = \begin{bmatrix} 73.16 \\ 95.725 \\ 4386.8 \end{bmatrix}$$

判别函数的系数$\Sigma^{-1}(\mu_1 - \mu_2) = \begin{bmatrix} 0.0651 \\ 0.0013 \\ 0.0009 \end{bmatrix}$

判别函数的常数项（$\frac{\mu_1 + \mu_2}{2})' \ \Sigma^{-1}(\mu_1 - \mu_2)$

$$= \begin{bmatrix} 73.16 & 95.725 & 4386.8 \end{bmatrix} \begin{bmatrix} -0.60581 \\ 0.25362 \\ 1.83679 \end{bmatrix} = 8.7161$$

线性判别函数：

$$y = 0.0651x_1 + 0.0013x_2 + 0.0009x_3 - 8.7161$$

$$y_1 = 0.0651\times 68.5 + 0.0013\times 79.3 + 0.0009\times 1950 - 8.7161$$

$= -2.4502 < 0$(第一个样本属于中水平）

$$y_2 = -0.60581\times 69.9 + 0.25362\times 96.9 + 1.83679\times 2840 - 8.7161$$

$= -1.5582 < 0$(第二个样本属于中水平）

$$y_3 = -0.60581\times 77.6 + 0.25362\times 93.8 + 1.83679\times 5233 - 8.7161$$

$= 1.0297 > 0$(第三个样本属于高水平）

$$y_4 = -0.60581\times 69.3 + 0.25362\times 90.3 + 1.83679\times 5158 - 8.7161$$

$= 0.4195 > 0$(第四个样本属于高水平）

重庆工商大学 CHONGQING TECHNOLOGY AND BUSINESS UNIVERSITY

(二) 多总体的距离判别法

随着计算机计算能力的增强和计算机的普及，距离判别法的判别函数也在逐步改进，一种等价的距离判别为:

设有个K总体，分别有均值向量$\boldsymbol{\mu}_{\mathrm{i}}$ (i=1,2,…,k)和协方差阵$\Sigma_{\mathrm{i}}=\Sigma$，各总体出现的先验概率相等。又设Y是一个待判样品。则与的距离为（即判别函数）

$$d^2(\mathbf{y},G_i)=(\mathbf{y}-\mu_i)'\Sigma^{-1}(\mathbf{y}-\mu_i)$$

$$=\mathbf{y}'\Sigma^{-1}\mathbf{y}-2\mathbf{y}'\Sigma^{-1}\mu_i+\mu_i'\Sigma^{-1}\mu_i'$$

上式中的第一项$\mathbf{y}'\Sigma^{-1}\mathbf{y}$与$i$无关，则舍去，得一个等价的函数

$$g_i(Y)=-2\mathbf{y}'\Sigma^{-1}\mu_i+\mu_i'\Sigma^{-1}\mu_i'$$

重慶工商大學
CHONGQING TECHNOLOGY AND BUSINESS UNIVERSITY

将上式中提-2，得

$$g_i(Y) = -2(\mathbf{y}'\Sigma^{-1}\mu_i - 0.5\mu_i'\Sigma^{-1}\mu_i')$$

令 $$f_i(Y) = (\mathbf{y}'\Sigma^{-1}\mu_i - 0.5\mu_i'\Sigma^{-1}\mu_i')$$

则距离判别法的判别函数为：

令 $$f_i(Y) = (\mathbf{y}'\Sigma^{-1}\mu_i - 0.5\mu_i'\Sigma^{-1}\mu_i')$$

判别规则为 $f_l(y) = \max\limits_{1 \le i \le k} f_i(x)$，则$y \in G_l$

注：这与前面所提出的距离判别是等价的.

$$d^2(\mathbf{y}, G_i) = (\mathbf{y} - \mu_i)'\Sigma_2^{-1}(\mathbf{y} - \mu_i)\text{最小} \Leftrightarrow$$

$$f_i(Y) = (\mathbf{y}'\Sigma^{-1}\mu_i - 0.5\mu_i'\Sigma^{-1}\mu_i')\text{最大}$$

对判别效果做出检验

1、错判概率

由上面的分析可以看出，马氏距离判别法是合理的，但是这并不意谓着不会发生误判。

两总体分别服从$N(\mu_1,\sigma^2)$和$N(\mu_2,\sigma^2)$，其线性判别函数为：

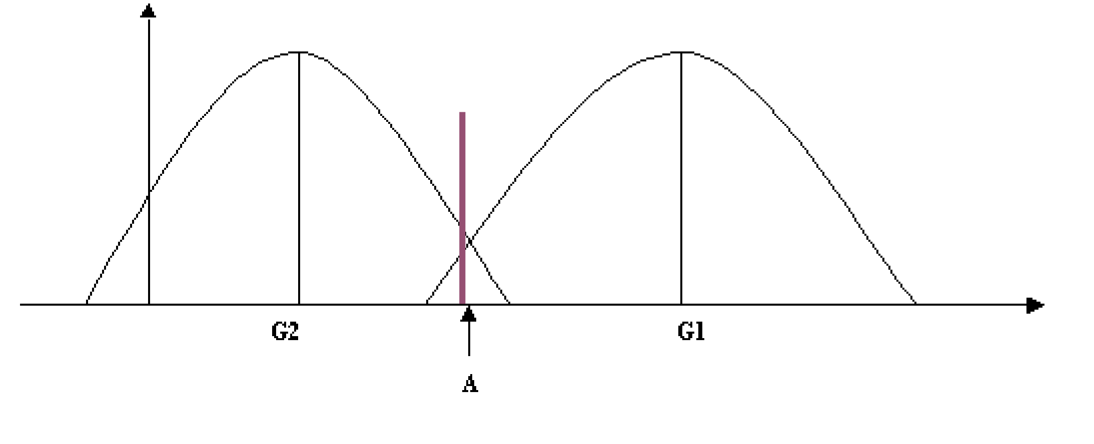
$$W(y)=(y-\bar{\mu})\frac{1}{\sigma^2}(\mu_1-\mu_2)$$，其中

$$\bar{\mu}=\frac{\mu_1+\mu_2}{2}$$。不失一般性设$\mu_1>\mu_2$，这种情况下线性判别函数 W（y）的符号取决于 $y>\bar{\mu}$ 还是 $y\leq\bar{\mu}$。

如果 $y>\bar{\mu}$，则$\mathbf{y}\in G_1$

如果 $y<\bar{\mu}$，则$\mathbf{y}\in G_2$。

重慶工商大學
CHONGQING TECHNOLOGY AND BUSINESS UNIVERSITY



错判概率：$P(X_2 > \overline{\mu}) = P(X_2 - \mu_2 > \dfrac{\mu_1 + \mu_2}{2} - \mu_2)$

$= P(X_2 - \mu_2 > \dfrac{\mu_1 - \mu_2}{2})$ $= P(\dfrac{X_2 - \mu_2}{\sigma} > \dfrac{\mu_1 - \mu_2}{2\sigma})$

$= 1 - \Phi(\dfrac{\mu_1 - \mu_2}{2\sigma})$

三、判别分析的实质

我们知道，判别分析就是希望利用已经测得的变量数据，找出一种判别函数，使得这一函数具有某种最优性质，能把属于不同类别的样本点尽可能地区别开来。为了更清楚的认识判别分析的实质，以便能灵活的应用判别分析方法解决实际问题，我们有必要了解“划分”这样概念。

设R_1，R_2，…，R_k是p维空间R^p的k个子集，如果它们互不相交，且它们的和集为R^p，则称

R_1，R_2，…，R_k为R^p的一个划分。

重慶工商大學 CHONGQING TECHNOLOGY AND BUSINESS UNIVERSITY

在两个总体的距离判别问题中，利用$W(\mathbf{X})=\boldsymbol{\alpha}'(\mathbf{X}-\bar{\boldsymbol{\mu}})$可以得到空间$R^p$的一个划分

$$\begin{cases} R_1=\{\mathbf{X}:W(\mathbf{X})\geq 0\} \\ R_2=\{\mathbf{X}:W(\mathbf{X})<0\} \end{cases} \qquad (4.11)$$

新的样品$\mathbf{X}$落入R_1推断$\mathbf{X}\in G_1$，落入R_2推断$\mathbf{X}\in G_2$。

这样我们将会发现，判别分析问题实质上就是在某种意义上，以最优的性质对p维空间R^p构造一个“划分”，这个“划分”就构成了一个判别规则。这一思想将在后面的各节中体现的更加清楚。

贝叶斯（Bayes）判别法

一、Bayes判别的基本思想

二、Bayes判别的基本方法

距离判别只要求知道总体的数值特征，不涉及总体的分布函数，当参数未知和协方差时，就用样本的均值和协方差矩阵来估计。距离判别方法简单实用，但没有考虑到每个总体出现的机会大小，即先验概率，没有考虑到错判的损失。

贝叶斯判别法正是为了解决这两个问题提出的判别分析方法。

重慶工商大學 CHONGQING TECHNOLOGY AND BUSINESS UNIVERSITY

办公室新来了一个雇员小王，小王是好人还是坏人大家都在猜测。按人们主观意识，一个人是好人或坏人的概率均为0.5。坏人总是要做坏事，好人总是做好事，偶尔也会做一件坏事，一般好人做好事的概率为0.9，坏人做好事的概率为0.2，一天，小王做了一件好事，小王是好人的概率有多大，你现在把小王判为何种人。

$$P(\text{好人}/\text{做好事})$$

$$=\frac{P(\text{好人})P(\text{做好事}/\text{好人})}{P(\text{好人})P(\text{做好事}/\text{好人})+P(\text{坏人})P(\text{做好事}/\text{坏人})}$$

$$=\frac{0.5\times 0.9}{0.5\times 0.9+0.5\times 0.2}=0.82$$

$$P(\text{坏人}/\text{做好事})$$

$$=\frac{P(\text{坏人})P(\text{做好事}/\text{坏人})}{P(\text{好人})P(\text{做好事}/\text{好人})+P(\text{坏人})P(\text{做好事}/\text{坏人})}$$

$$=\frac{0.5\times 0.2}{0.5\times 0.9+0.5\times 0.2}=0.18$$

距离判别简单直观，很实用，但是距离判别的方法把总体等同看待，没有考虑到总体会以不同的概率（先验概率）出现。一个好的判别方法，要考虑到各个总体出现的先验概率，**Bayes**判别就具有这些优点，其判别效果更加理想，应用也更广泛。

贝叶斯公式是一个我们熟知的公式

$$P(B_i \mid A) = \frac{P(A \mid B_i)P(B_i)}{\Sigma P(A \mid B_i)P(B_i)}$$

重慶工商大學
CHONGQING TECHNOLOGY AND BUSINESS UNIVERSITY

一、Bayes判别的基本思想

问题：设有 k 个总体 $G_1, G_2, \cdots, G_k$，其各自的分布密度函数 $f_1(\mathbf{x}), f_2(\mathbf{x}), \cdots, f_k(\mathbf{x})$ 互不相同的，假设 k 个总体各自出现的概率分别为 $q_1, q_2, \cdots, q_k$（先验概率），$q_i \geq 0$，$\sum_{i=1}^{k} q_i = 1$。假设已知若将本来属于 G_i 总体的样品错判到总体 G_j 时造成的损失为 $C(j \mid i)$，$i, j = 1, 2, \cdots, k$。在这样的情形下，对于新的样品 $\mathbf{X}$ 判断其来自哪个总体。

重庆工商大学 CHONGQING TECHNOLOGY AND BUSINESS UNIVERSITY

下面我们对这一问题进行分析。首先应该清楚$C(i\mid i)=0$、$C(j\mid i)\geq 0$，对于任意的$i,j=1,2,\cdots,k$成立。设k个总体$G_1,G_2,\cdots,G_k$相应的p维样本空间为$R_1,R_2,\cdots,R_k$，即为一个划分，故我们可以简记一个判别规则为$R=(R_1,R_2,\cdots,R_k)$。从描述平均损失的角度出发，如果原来属于总体G_i且分布密度为$f_i(\mathbf{x})$的样品，正好取值落入了R_j，我们就将会错判为属于G_j。

故在规则 R 下，将属于 G_i 的样品错判为 G_j 的概率为

$$P(j \mid i,R)=\int_{R_j} f_i(\mathbf{x})d\mathbf{x} \qquad i,j=1,2,\cdots,k \qquad i\neq j$$

如果实属 G_i 的样品，错判到其它总体 $G_1,\cdots,G_{i-1},G_{i+1}\cdots,G_k$ 所造成的损失为 $C(1\mid i),\cdots,C(i-1\mid i),C(i+1\mid i)\cdots,C(k\mid i)$，则这种判别规则 R 对总体 G_i 而言，样品错判后所造成的平均损失为

$$r(i\mid R)=\sum_{j=1}^{k}[C(j\mid i)P(j\mid i,R)] \qquad i=1,2,\cdots,k$$

其中 $C(i\mid i)=0$

由于 k 个总体 $G_1, G_2, \cdots, G_k$ 出现的先验概率分别为 $q_1, q_2, \cdots, q_k$，则用规则 R 来进行判别所造成的总平均损失为

$$
\begin{aligned}
g(R) &= \sum_{i=1}^{k} q_i r(i, R) \\
&= \sum_{i=1}^{k} q_i \sum_{j=1}^{k} C(j \mid i) P(j \mid i, R) \qquad (4.12)
\end{aligned}
$$

所谓 Bayes 判别法则，就是要选择 $R_1, R_2, \cdots, R_k$，使得（4.12）式表示的总平均损失 $g(R)$ 达到极小。

设每一个总体 G_i 的分布密度为 $f_i(\mathbf{x})$，$i=1,2,\cdots,k$，来自总体 G_i 的样品 $\mathbf{X}$ 被错判为来自总体 G_j（$i,j=1,2,\cdots,k$）时所造成的损失记为 $C(j\,|\,i)$，并且 $C(i\,|\,i)=0$。那么，对于判别规则 $R=(R_1,R_2,\cdots,R_k)$ 产生的误判概率记为 $P(j\,|\,i,R)$，有

$$P(j\,|\,i,R)=\int_{R_j} f_i(\mathbf{x})d\mathbf{x}$$

如果已知样品 X 来自总体G_i的先验概率为q_i，则在规则R下，可知误判的总平均损失为

$$g(R)=\sum_{i=1}^{k}q_i\sum_{j=1}^{k}C(j\,|\,i)P(j\,|\,i,R)$$

$$=\sum_{i=1}^{k}q_i\sum_{j=1}^{k}C(j\,|\,i)\int_{R_j}f_i(\mathbf{x})d\mathbf{x}$$

$$=\sum_{j=1}^{k}\int_{R_j}(\sum_{i=1}^{k}q_iC(j\,|\,i)f_i(\mathbf{x}))d\mathbf{x} \qquad \text{（4.13）}$$

令 $\sum_{i=1}^{k}q_iC(j\,|\,i)f_i(\mathbf{x})=h_j(\mathbf{x})$，那么，（4.13）式为

$$g(R)=\sum_{j=1}^{k}\int_{R_j}h_j(\mathbf{x})d\mathbf{x}$$

如果空间 R^p 有另一种划分 $R^* = (R_1^*, R_2^*, \cdots, R_k^*)$ ，则它的总平均损失为

$$g(R^*) = \sum_{j=1}^{k} \int_{R_j^*} h_j(\mathbf{x}) d\mathbf{x}$$

那么，在两种划分下的总平均损失之差为

$$g(R) - g(R^*) = \sum_{i=1}^{k} \sum_{j=1}^{k} \int_{R_i \cap R_j^*} [h_i(\mathbf{x}) - h_j(\mathbf{x})] d\mathbf{x} \qquad (4.14)$$

由 R_i 的定义，在 R_i 上 $h_i(\mathbf{x}) \le h_j(\mathbf{x})$ 对一切 j 成立，故（4.14）式小于或等于零，这说明 $R_1, R_2, \cdots, R_k$ 确能使总平均损失达到极小，它是 Bayes 判别的解。

重慶工商大學
CHONGQING TECHNOLOGY AND BUSINESS UNIVERSITY

这样，我们以 Bayes 判别的思想得到的划分 $R=(R_1,R_2,\cdots,R_k)$ 为

$$R_i=\{\mathbf{x}\mid h_i(\mathbf{x})=\min_{1\le j\le k}h_j(\mathbf{x})\}\qquad i=1,2,\cdots,k\qquad(4.15)$$

具体说来，当抽取了一个未知总体的样本值 $\mathbf{X}$，要判断它属于哪个总体，只要前计算出 k 个按先验分布加权的误判平均损失

$$h_j(\mathbf{x})=\sum_{i=1}^{k}q_iC(j\mid i)f_i(\mathbf{x}))\qquad j=1,2,\cdots,k\qquad(4.16)$$

然后再比较这 k 个误判平均损失 $h_1(\mathbf{x}),h_2(\mathbf{x}),\cdots,h_k(\mathbf{x})$ 的大小，选取其中最小的，则判定样品 $\mathbf{X}$ 来自该总体。

这里我们看一个特殊情形，当$k=2$时，由（4.16）式得

$$h_1(\mathbf{x}) = q_2 C(1|2) f_2(\mathbf{x})) \qquad h_2(\mathbf{x}) = q_1 C(2|1) f_1(\mathbf{x}))$$

从而

$$R_1 = \{\mathbf{x} \mid q_2 C(1|2) f_2(\mathbf{x}) \le q_1 C(2|1) f_1(\mathbf{x})\}$$
$$R_2 = \{\mathbf{x} \mid q_2 C(1|2) f_2(\mathbf{x}) > q_1 C(2|1) f_1(\mathbf{x})\}$$

若令

$$V(\mathbf{x}) = \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})}, \qquad d = \frac{q_2 C(1|2)}{q_1 C(2|1)}$$

则判别规则可表示为

$$\begin{cases} \mathbf{x} \in G_1, & \text{当} V(\mathbf{x}) \ge d \\ \mathbf{x} \in G_2, & \text{当} V(\mathbf{x}) < d \end{cases} \qquad (4.17)$$

设有总体 $G_i(i=1,2,\cdots,k)$ ， 具有概率密度函数 $f_i(x)$。并且根据以往的统计分析，知道 G_i 出现的概率为 q_i 。即当样本 x_0 发生时，求他属于某类的概率。由贝叶斯公式计算后验概率，有：

$$P(G_i \mid x_0)=\frac{q_i f_i(x_0)}{\Sigma q_j f_j(x_0)}$$

判别规则

$$P(G_l \mid x_0)=\frac{q_l f_l(x_0)}{\Sigma q_j f_j(x_0)}=\max_{1\le i\le k}\frac{q_i f_i(x_0)}{\Sigma q_j f_j(x_0)}$$

则 x_0 判给 G_l。在正态的假定下， $f_i(x)$ 为正态分布的密度函数。

下面讨论总体服从正态分布的情形

$$q_l f_l(x_0) = \max_{1 \le i \le k} q_i f_i(x_0),$$ 则 x_0 判给 G_l。

$$若 f_i(x) = \frac{1}{(2\pi)^{p/2}|\Sigma_i|^{1/2}} \exp[-\frac{1}{2}(x-\mu^{(i)})'\Sigma_i^{-1}(x-\mu^{(i)})]$$

$$则, q_i f_i(x) = q_i \frac{1}{(2\pi)^{p/2}|\Sigma_i|^{1/2}} \exp[-\frac{1}{2}(x-\mu^{(i)})'\Sigma_i^{-1}(x-\mu^{(i)})]$$

17个破产企业4项指标的均值和协方差

$N_4(\mu_1, \Sigma_1)$.

$$\mu_1 = \begin{bmatrix} -0.07941 \\ -0.08882 \\ 1.34882 \\ 0.43000 \end{bmatrix} \quad \Sigma_1 = \begin{bmatrix} 0.0527808824 & 0.0340117647 & 0.0433257353 & 0.0040562500 \\ 0.0340117647 & 0.0251110294 & 0.0333014706 & 0.0031437500 \\ 0.0433257353 & 0.0333014706 & 0.2187610294 & 0.0403875000 \\ 0.0040562500 & 0.0031437500 & 0.0403875000 & 0.0503375000 \end{bmatrix}$$

$$f_1(x) = \frac{1}{(2\pi)^{\frac{4}{2}} |\Sigma_1|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu_1)' \Sigma_1^{-1} (x-\mu_1) \right) \qquad q_1 = \frac{17}{17+21}$$

$$x = \begin{bmatrix} 0.04 \\ 0.01 \\ 1.5 \\ 0.71 \end{bmatrix}$$

0.0082

21个正常企业4项指标的均值和协方差

$N_4(\mu_2, \Sigma_2)$.

$$\mu_2 = \begin{bmatrix} 0.22571 \\ 0.00524 \\ 2.67286 \\ 0.44095 \end{bmatrix} \quad \Sigma_2 = \begin{bmatrix} 0.050925714 & 0.025623571 & 0.094172857 & -0.007260714 \\ 0.025623571 & 0.056596190 & 0.020054286 & -0.006345238 \\ 0.094172857 & 0.020054286 & 1.207031429 & 0.043737143 \\ -0.007260714 & -0.006345238 & 0.043737143 & 0.028219048 \end{bmatrix}$$

$$f_2(x) = \frac{1}{(2\pi)^{\frac{4}{2}} |\Sigma_2|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu_2)' \Sigma_2^{-1} (x-\mu_2) \right) \qquad q_2 = \frac{21}{17+21}$$

$$x = \begin{bmatrix} 0.04 \\ 0.01 \\ 1.5 \\ 0.71 \end{bmatrix}$$

0.001

上式两边取对数并去掉与i无关的项，则等价的判别函数为：

$$z_i(x)=\ln(q_i f_i(\mathbf{x}))$$

$$=\ln q_i-\frac{1}{2}\ln|\Sigma_i|-\frac{1}{2}(x-\mu^{(i)})'\Sigma_i^{-1}(x-\mu^{(i)})]$$

问题转化为若 $Z_l(x)=\max\limits_{1\le i\le k}[Z_i(x)]$ ，则判 $x\in G_l$ 。

$$z_1(x)=\ln(q_1 f_1(\mathrm{x}))$$

$$=\ln q_1-\frac{1}{2}\ln|\Sigma_1|-\frac{1}{2}(x-\mu^{(1)})'\Sigma_1^{-1}(x-\mu^{(1)})$$

$$=-1.1329$$

$$z_2(x)=\ln(q_2 f_2(\mathrm{x}))$$

$$=\ln q_2-\frac{1}{2}\ln|\Sigma_2|-\frac{1}{2}(x-\mu^{(2)})'\Sigma_2^{-1}(x-\mu^{(2)})$$

$$=-3.1903$$

当协方差阵相等 $\Sigma_1 = \cdots \Sigma_k = \Sigma$

去掉相同的部分得：

$$\boldsymbol{m}_i(\mathbf{x}) = \ln \boldsymbol{q}_i - \frac{1}{2}\boldsymbol{\mu}^{(\mathrm{i})\prime}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}^{(\mathrm{i})} + \boldsymbol{\mu}^{(\mathrm{i})\prime}\boldsymbol{\Sigma}^{-1}\mathbf{x}$$

问题转化为若 $m_l(x) = \max\limits_{1\le i\le k}[m_i(x)]$，则判 $x \in G_l$ 。

当先验概率相等，$q_1 = \cdots = q_k = \dfrac{1}{k}$

去掉相同的部分得：

有 $\quad m_i(x) = -\dfrac{1}{2}\boldsymbol{\mu}^{(i)\prime}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}^{(i)} + \boldsymbol{\mu}^{(i)\prime}\boldsymbol{\Sigma}^{-1}\mathbf{x}$

完全成为距离判别法 。

例1:

	x1	x2	x3
高水平	76	99	5374
	79.5	99	5359
	78	99	5372
	72.1	95.9	5242
	73.8	77	5370
中水平	71.2	93	4250
	75.3	94.9	3412
	70	91.2	3390
	72.8	99	2300
	62.9	80.6	3799
待判	68.5	79.3	1950
	69.9	96.9	2840
	77.6	93.8	5233
	69.3	90.3	5158

变量	均值向量		协方差矩阵的估计		
	高	中			
x1	75.88	70.44	153.8	218.95	-5558.75
x2	93.98	91.74	218.95	695.35	-14484.25
x3	5343.4	3430.2	-5558.75	-14484.25	2625465

$$q_1 = q_2 = 0.5$$

$$m_i(x) = \ln q_i - \frac{1}{2}\mu^{(i)'}\Sigma^{-1}\mu^{(i)} + \mu^{(i)'}\Sigma^{-1}x$$

费雪（Fisher）判别法

一、Fisher判别的基本思想

二、 Fisher判别函数的构造

一、Fisher判别的基本思想

从 k 个总体中抽取具有 p 个指标的样品观测数据，借助方差分析的思想构造一个线性判别函数

$$U(\mathbf{X}) = c_1X_1 + c_2X_2 + \cdots + c_pX_p = C'\mathbf{X} \qquad (4.19)$$

其中系数 $\mathbf{u} = (u_1, u_2, \cdots, u_p)'$ 确定的原则是使得总体之间区别最大，而使每个总体内部的离差最小。有了线性判别函数后，对于一个新的样品，将它的 p 个指标值代入线性判别函数（4.19）式中求出 $U(\mathbf{X})$ 值，然后根据判别一定的规则，就可以判别新的样品属于哪个总体。

重庆工商大学 CHONGQING TECHNOLOGY AND BUSINESS UNIVERSITY

已知 A、B 两类观察对象，A 类有 n_{A} 例，B 类有 n_{B} 例，分别记录了 $X_1, X_2, \cdots, X_p$ 个观察指标，称为判别指标或变量。Fisher 判别法就是找出一个线性组合

$$Z = C_1 X_1 + C_2 X_2 + \cdots + C_p X_p$$

Fisher 准则：使得综合指标 Z 在 A 类的均数 $\bar{Z}_A$ 与在 B 类的均数 $\bar{Z}_B$ 的差异 $|\bar{Z}_A - \bar{Z}_B|$ 尽可能大，而两类内综合指标 Z 的变异 $S_A^2 + S_B^2$ 尽可能小，即使得 λ 达到最大。

$$\lambda = \frac{|\bar{Z}_A - \bar{Z}_B|}{S_A^2 + S_B^2}$$

判别系数 C 可通过对 λ 求导，由下列方程组解出

$$\begin{cases} S_{11}C_1 + S_{12}C_2 + \cdots + S_{1m}C_m = D_1 \\ S_{21}C_1 + S_{22}C_2 + \cdots + S_{2m}C_m = D_2 \\ \cdots \\ S_{m1}C_1 + S_{m2}C_2 + \cdots + S_{mm}C_m = D_m \end{cases}$$

式中 $D_j = \bar{X}_j^{(A)} - \bar{X}_j^{(B)}$， $\bar{X}_j^{(A)}, \bar{X}_j^{(B)}$ 分别是 A 类和 B 类第 j 个指标的均数 $(j = 1, 2, \cdots, p)$；

S_{ij} 是 $X_1, X_2, \cdots, X_p$ 的合并协方差阵的元素。

$$S_{ij} = \frac{\sum (X_i^{(A)} - \bar{X}_i^{(A)})(X_j^{(A)} - \bar{X}_j^{(A)}) + \sum (X_i^{(B)} - \bar{X}_i^{(B)}))(X_j^{(B)} - \bar{X}_j^{(B)})}{n_A + n_B - 2}$$

式中 $X_i^{(A)}, X_i^{(B)}, X_j^{(A)}, X_j^{(B)}$ 分别为 X_i 和 X_j 于 A 类和 B 类的观察值。

2.判别规则　建立判别函数后,按公式(18-1)逐例计算判别函数值 Z_i，进一步求 Z_i 的两类均数 $\bar{Z}_A$ 、$\bar{Z}_B$ 与总均数 $\bar{Z}$，按下式计算判别界值:

$$Z_c = \frac{\bar{Z}_A + \bar{Z}_B}{2} \qquad (18\text{-}5)$$

判别规则：

$$\begin{cases} Z_i > Z_c, & \text{判为A类} \\ Z_i < Z_c, & \text{判为B类} \\ Z_i = Z_c, & \text{判为任意一类} \end{cases} \qquad (18\text{-}6)$$

例18-1　收集了22例某病患者的三个指标（X_1，X_2，X_3）的资料列于表18-1，其中前期患者（A）类12例，晚期患者（B）类10例。试作判别分析。

表18-1　22例患者三项指标观察结果（Z_c=-0.147）

类别	编号	观察值 X_1	观察值 X_2	观察值 X_3	Z	Fisher 判别结果
A	1	23	8	0	0.19	A
A	2	-1	9	-2	2.73	A
A	3	-10	5	0	1.83	A
A	4	-7	-2	1	-0.28	B
A	5	-11	3	-4	2.72	A
A	6	-10	3	-1	1.69	A
A	7	25	9	-2	0.91	A
A	8	-19	12	-3	4.98	A
A	9	9	8	-2	1.81	A
A	10	-25	-3	-1	1.39	A
A	11	0	-2	2	-1.09	B
A	12	-10	-2	0	0.25	A
B	13	9	-5	1	-2.07	B
B	14	2	-1	-1	-0.05	A
B	15	17	-6	-1	-2.22	B
B	16	8	-2	1	-1.33	B
B	17	17	-9	1	-3.53	B
B	18	0	-11	3	-3.43	B
B	19	-9	-20	3	-4.82	B
B	20	-7	-2	3	-0.91	B
B	21	-9	6	0	1.98	A
B	22	12	0	0	-0.84	B

（1）计算变量的类均数及类间均值差D_j，计算结果列于表18-2。

表18-2　变量的均数及类间均值差

类别	例数	$\bar{X}_1$	$\bar{X}_2$	$\bar{X}_3$
A	12	−3	4	−1
B	10	4	−5	1
类间均值差D_j		−7	9	−2

（2）计算合并协方差矩阵: 按公式（18-4），例如:

$$S_{11}=\frac{[(23+3)^2+(-1+3)^2+\cdots+(-10+3)^2]+[(9-4)^2+(2-4)^2+\cdots+(12-4)^2]}{12+10-2}=175.3$$

得到合并协方差阵，代入公式（18-3）得

$$\mathbf{S}=\begin{pmatrix}175.3 & 20.3 & -2.3\\ 20.3 & 38.2 & -5.8\\ -2.3 & -5.8 & 2.7\end{pmatrix} \longrightarrow \begin{cases}175.3C_1+20.3C_2-2.3C_3=-7\\ 20.3C_1+38.2C_2-5.8C_3=9\\ -2.3C_1-5.8C_2+2.7C_3=-2\end{cases}$$

重慶工商大學 CHONGQING TECHNOLOGY AND BUSINESS UNIVERSITY

解此正规方程得

$$C_1=-0.070，\quad C_2=0.225，\quad C_3=-0.318$$

判别函数为

$$Z=-0.070X_1+0.225X_2-0.318X_3。$$

逐例计算判别函数值 Z_i 列于表 18-1 中的 Z 列，同时计算出 $\bar{Z}_A=1.428$、$\bar{Z}_B=-1.722$ 与总均数 $\bar{Z}=-0.004$。

（3）确定界值，进行两类判别：按公式（18-5）计算 $Z_c=(1.428-1.722)/2=-0.147$ ，将 $Z_i>-0.147$ 判为 A 类，$Z_i<-0.147$ 判为 B 类。判别结果列于表 18-1 的最后一列，有 4 例错判。

二、判别效果的评价　　　用误判概率 P 衡量

方法：回顾性：样本回代。必须做，但效果差。

回顾性误判概率估计往往夸大判别效果。

前瞻性：验证样本。

刀切法：

步骤 ①顺序剔除一个样品，用余下的 N-1 个样品建立判别函数；

②用判别函数判别剔除的样品；

③重复上两步 N 次，计算误判概率。

此法优点：充分利用了样本的信息建立和验证判别函数。本例刀切法误判概率估计值为 $6/22 = 27.3\%$ 。

要求判别函数的误判概率小于 0.1 或 0.2 才有应用价值。

重庆工商大学
CHONGQING TECHNOLOGY AND BUSINESS UNIVERSITY

2.Logistic回归

logistic回归模型

工作目的：因变量为分类变量的回归模型

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i$$

普通回归：对回归模型中的自变量、回归系数以及残差项的取值都没有任何限制，作为自变量函数的因变量就必须能够在 $(-\infty, +\infty)$ 范围内自由取值。

如果因变量只取分类值，或者只取两类值（0、1），就会严重违反因变量为连续型变量的假设。

重慶工商大學
CHONGQING TECHNOLOGY AND BUSINESS UNIVERSITY

设因变量 y 只取0、1两个数值的虚拟变量，是一个两点分布变量，在给定的条件下，记概率

$$\mathrm{P}(y_i=1|x_{i1},x_{i2},\cdots,x_{ip})=p_i$$

$$\mathrm{P}(y_i=0|x_{i1},x_{i2},\cdots,x_{ip})=1-p_i$$

$$\mathrm{E}(y_i|x_{i1},x_{i2},\cdots,x_{ip})=1\times p_i+0\times(1-p_i)$$

$$=\mathrm{P}\left(y_i=1|x_{i1},x_{i2},\cdots,x_{ip}\right)$$

$$\mathrm{E}(y_i\,|\,x_{i1},x_{i2},\cdots,x_{ip})=\beta_0+\beta_1x_{i1}+\beta_2x_{i2}+\cdots+\beta_px_{ip}$$

这时所建立的线性回归模型就是在分析当自变量变化时，概率是如何变化的。

$$\mathrm{P}(y_i=1|x_{i1},x_{i2},\cdots,x_{ip})$$

问题：

（1）事件概率值

$$\mathrm{E}(y_i \mid x_{i1}, x_{i2}, \cdots, x_{ip}) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}$$

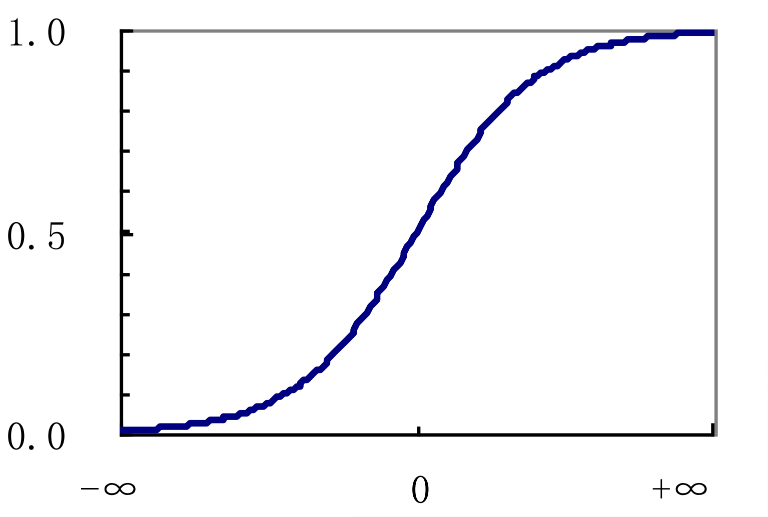
的取值只能在[0，1]范围内；

（2）随机误差项有**不变方差**的假设已不能再维持。

因此，在因变量为定性变量的情况下，普通线性回归模型是不适用的。

考虑采用非线性模型 （S型曲线）

1838年由比利时学者Verhulst首次提出。1920年美国学者 Bearl & Reed在研究果蝇的繁殖中发现和使用该函数，并在人口估计和预测中推广使用



$[0,1]$

logistic函数的

$$f(x)=\frac{e^x}{1+e^x}$$

用 $p_i=\mathrm{P}(y_i=1|x_{i1},x_{i2},\cdots,x_{ip})$ 作为因变量，得到logistic回归模型

$$p_i=\frac{\exp(\beta_0+\beta_1x_{i1}+\beta_2x_{i2}+\cdots+\beta_px_{ip})}{1+\exp(\beta_0+\beta_1x_{i1}+\beta_2x_{i2}+\cdots+\beta_px_{ip})}$$

$$\ln\frac{p_i}{1-p_i}=\beta_0+\beta_1x_{i1}+\beta_2x_{i2}+\cdots+\beta_px_{ip}$$

重慶工商大學 CHONGQING TECHNOLOGY AND BUSINESS UNIVERSITY

将logistic回归模型转换为线性模型

Logit（logistic probability unit）变换：

定义：

$$\text{Logit}(p_i)=\ln\frac{p_i}{1-p_i}$$

得到：

$$\text{Logit}(p_i)=\alpha+\beta_1 x_{i1}+\beta_2 x_{i2}+\cdots+\beta_p x_{ip}+\varepsilon_i$$

Logit变换的特点：

$$p_i\in[0,1]$$

$$\frac{p_i}{1-p_i}\in[0,+\infty)$$

$$\text{Logit}(p_i)=\ln\frac{p_i}{1-p_i}\in(-\infty,+\infty)$$

较好地克服了概率因变量的取值受到限制的困难。

重慶工商大學
CHONGQING TECHNOLOGY AND BUSINESS UNIVERSITY

关于 $\mathrm{Logit}(p_i)$解释

定义：对数发生比

$$\mathrm{Logit}(p_i) = \ln\frac{p_i}{1-p_i}$$

$\mathrm{Logit}(p_i)$ 与 p_i 之间具有同向变化的规律。

logistic回归模型解释了：对数发生比与自变量集合之间的线性回归关系。

模型求解：极大似然法

记得到一个实际观测值 $y_i(i=1,2,\cdots,n)$ **的概率为**

$$\mathrm{P}(y_i)=p_i^{y_i}(1-p_i)^{1-y_i}$$

则 $y_1,y_2,\cdots,y_n$ **的似然函数为** $L=\prod_{i=1}^{n}\mathrm{P}(y_i)=\prod_{i=1}^{n}p_i^{y_i}(1-p_i)^{1-y_i}$

两边取对数： $\ln L=\sum_{i=1}^{n}[y_i\ln p_i+(1-y_i)\ln(1-p_i)]=$

$$\sum_{i=1}^{n}[y_i\ln\frac{p_i}{1-p_i}+\ln(1-p_i)]$$

最后得到：

$$\ln L=\sum_{i=1}^{n}[y_i(\alpha+\beta_1x_{i1}+\cdots+\beta_px_{ip})-\ln(1+\exp(\alpha+\beta_1x_{i1}+\cdots+\beta_px_{ip}))]$$

当使得 $\ln L$ 取得最大值时，参数估计值即为所求。

重慶工商大學
CHONGQING TECHNOLOGY AND BUSINESS UNIVERSITY

Logistic回归模型

- 其中， $\beta_0,\beta_1,\cdots,\beta_k$ 是待估参数。根据上式可以得到优势的值：

$$\frac{p}{1-p}=e^{\beta_0+\beta_1 x_1+\cdots+\beta_k x_k}$$

- 可以看出，参数 β_i是控制其它x时 x_i 每增加一个单位对优势产生的乘积效应。
- 概率p的值：

$$p=\frac{e^{\beta_0+\beta_1 x_1+\cdots+\beta_k x_k}}{1+e^{\beta_0+\beta_1 x_1+\cdots+\beta_k x_k}}$$

含有名义数据的logit

- 有些协变量为定量数据，logistic回归模型的协变量可以是定性名义数据。这就需要对名义数据进行赋值。
- 通常某个名义数据有k个状态，则定义变量 $M_1,\cdots,M_{k-1}$ 代表前面的k-1状态，最后令k-1变量均为0或-1来代表第k个状态。
- 如婚姻状况有四种状态：未婚、有配偶、丧偶和离婚，则可以定义三个指示变量M_1、M_2、M_3，用(1,0,0)、(0,1,0)、(0,0,1)、(0,0,0)或(-1,-1,-1)来对以上四种状态赋值。

含有名义数据的logit

各年龄组和各类婚姻状况的死亡人数

年龄	未婚	有配偶	丧偶	离婚
25～29	349.1	417.1	4.1	11.4
30～34	329.2	877.8	11.6	25.5
35～39	213.9	1268.8	16.0	26.0
40～44	127.5	1299.2	31.5	27.9
45～49	68.7	1357.1	45.4	26.3
50～54	86.4	2107.4	130.1	38.3
55～59	99.4	4255.2	446.1	77.4
60～64	92.9	5868.7	1082.7	117.6
65～69	119.5	7240.7	2351.4	159.6
≥70	434.8	20271.0	29842.0	414.3
合计	1921.4	44963.0	33959.9	924.3

含有名义数据的logit

例：某地25岁及以上人中各类婚姻状况居民的死亡情况见表，试建立死亡率关于年龄和婚姻状况的logit模型。

$$\ln\frac{p}{1-p}=\mu+\beta A+\gamma_1 M_1+\gamma_2 M_2+\gamma_3 M_3$$

其中，A表示年龄(取中值)，M_1、M_2、M_3表示婚姻状况

于是，估计的logit方程为：

$$\ln\frac{\hat{p}}{1-\hat{p}}=-11.536+0.124A+0.711M_1-0.423M_2+0.021M_3$$

含有有序数据的logit

- Logit模型的协变量也可以是有序数据
- 对有序数据的赋值可以按顺序用数0,1,2,3,4分别表示

【例】某地某年各类文化程度的死亡人数见表，试建立logit模型。

- 建立死亡率关于年龄和文化程度的logit模型

$$\ln\frac{p}{1-p}=\mu+\beta A+\gamma E$$

- 其中A为年龄，E为文化程度

重慶工商大學
CHONGQING TECHNOLOGY AND BUSINESS UNIVERSITY

含有有序数据的logit

表6.5　某校男、女教职工工龄和文化程度的统计结果

性别	升职	工龄	文化程度：中学	大学	硕士	博士
男	有	≤5	1	2	6	13
		6－15	3	7	13	20
		16－29	7	14	20	25
		≥30	15	20	23	31
	无	≤5	49	97	142	185
		6－15	96	140	182	176
		16－29	141	183	170	137
		≥30	179	157	117	101
女	有	≤5	3	5	7	12
		6－15	5	10	10	14
		16－29	9	13	15	15
		≥30	14	16	16	11
	无	≤5	198	207	189	189
		6－15	179	236	163	137
		16－29	193	184	147	91
		≥30	186	151	83	41

含有有序数据的logit

于是，估计的logit方程为：

$$\ln\frac{p}{1-p}=-11.637+0.124A-0.164E$$

其中，年龄的系数0.124，说明年龄越大死亡率会越高；

文化程度的系数-0.164，说明文化程度与死亡率呈负相关，文化程度越高，死亡率越低。

重庆工商大学
CHONGQING TECHNOLOGY AND BUSINESS UNIVERSITY

多项logit模型

前面讨论的logit模型为二分数据的情况，有时候响应变量有可能取三个或更多值，即多类别的属性变量。

根据响应变量类型的不同，分两种情况：

- 响应变量为定性名义变量；
- 响应变量为定性有序变量；

当名义响应变量有多个类别时，多项logit模型应采取把每个类别与一个基线类别配成对，通常取最后一类为参照，称为基线-类别logit.

重慶工商大學
CHONGQING TECHNOLOGY AND BUSINESS UNIVERSITY

多项logit模型

预测变量为x的基线-类别logit模型为：

$$\ln\left(\frac{\pi_j}{\pi_J}\right)=\alpha_j+\beta_j x,\ j=1,\cdots,J-1$$

模型共有J-1个方程，每个方程有不同的参数，这些效应依据与基线配对的类别而变化；

软件可以同时拟合模型中的所有方程；

不管哪个类别作为基线，对于同一对类别都会有相同的参数估计；即基线类别的选择是任意的；

重庆工商大学
CHONGQING TECHNOLOGY AND BUSINESS UNIVERSITY

多项logit模型

【例】研究三个学校、两个课程计划对学生偏好何种学习方式的影响。调查数据见表:

其中，三个学校对应两个哑变量x_1和x_2，两个课程计划为常规(x_3=1)和附加(x_3=0)，学习方式分为：自修(y=1)、小组(y=2)、上课(y=3)

从题目可以看出，响应变量是学习方式有三类，属于多项逻辑斯蒂回归问题。于是，建模为：

$$\begin{cases} \ln \dfrac{p_1}{p_3} = \beta_{10} + \beta_{11}x_1 + \beta_{12}x_2 + \beta_{13}x_3 \\ \ln \dfrac{p_2}{p_3} = \beta_{20} + \beta_{21}x_1 + \beta_{22}x_2 + \beta_{23}x_3 \end{cases}$$

重庆工商大学
CHONGQING TECHNOLOGY AND BUSINESS UNIVERSITY

多项logit模型

表 6.9 学校、课程计划和学习方式

学校 (x_1x_2)	课程计划 x_3	学习方式 y			合计
		$y=1$	$y=2$	$y=3$	
(1 0)	$x_3=0$	5	12	50	67
	$x_3=1$	10	17	26	53
(0 1)	$x_3=0$	16	12	36	74
	$x_3=1$	21	17	26	64
(0 0)	$x_3=0$	12	12	20	44
	$x_3=1$	15	15	16	46

多项logit模型

- 应用统计软件可以得到模型的参数估计和回归方程：

$$\begin{cases} \ln\dfrac{p_1}{p_3} = -0.593 - 1.134x_1 + 0.618x_3 \\ \ln\dfrac{p_2}{p_3} = -0.603 + 0.635x_3 \end{cases}$$

- 然后，将x1和x3的取值代入上式，可以进一步对三个属性之间的关系加以分析。

  学校2与学校3的学生在自修与上课两种学习方式上偏好相同；
  学校1比学校2和3更偏好上课(1.727>0.593);
  课程计划中，常规课程与附加课程相比，常规课程学生更偏好自修；
  小组与上课相比，三个学校没有差别；常规课程学生更偏好小组学习。

多项logit模型

- 当响应变量为定性有序变量时，多项logit模型的处理会与名义变量有所不同。
- 有序响应变量的累积logit模型
  - 当变量为有序变量时，logit可以利用这一点，得到比基线-类别有更简单解释的模型；
  - Y的累积概率是指Y落在一个特定点的概率，对结果为类别j时，其累积概率为：

$$P(Y \le j) = \pi_1 + \cdots + \pi_j, \; j = 1, \cdots, J$$

  - 累积概率满足：$P(Y \le 1) \le \cdots \le P(Y \le J) = 1$
  - 累积概率的模型并不利用最后一个概率，因为它必然等于1

多项logit模型

【例】研究性别和两种治疗方法(传统疗法与新疗法)对某种疾病疗效的影响，84个病人的数据见表。

由题知，疗效是一个有序变量，包括显著、较有效和无效三个值，需要建立累积logit模型。

表 6.10　性别、治疗方法和疗效

性别 x_1	治疗方法 x_2	疗效			合计
		显著	较有效	无效	
男 $x_1=0$	新疗法 $x_2=1$	5	2	7	14
	传统疗法 $x_2=0$	1	0	10	11
女 $x_1=1$	新疗法 $x_2=1$	16	5	6	27
	传统疗法 $x_2=0$	6	7	19	32

多项logit模型

- 令p_1,p_2,p_3分别表示疗效的三种情况出现的概率，在对性别和疗法赋值后，则累积logit模型为：

$$\begin{cases} \ln \dfrac{p_1}{1-p_1} = \beta_{10} + \beta_1 x_1 + \beta_2 x_2 \\ \ln \dfrac{p_1+p_2}{1-(p_1+p_2)} = \beta_{20} + \beta_1 x_1 + \beta_2 x_2 \end{cases}$$

- 其中，与基线-类别logit不同的是，参数 β_1 描述了变量 x_1 对响应变量落在类j或小于j的对数优势的效应，且对所有(J-1)个累积logit都是相等的；β_2 的情况类似。
- 以上性质决定了在其他变量不变的情况下，x_1 每增加一个单位，响应变量在任意给定类别下的优势比将为 e^{β} 。
- 这一相同的比例(β)适用于每个累积概率,称为比例优势假设.

重庆工商大学 CHONGQING TECHNOLOGY AND BUSINESS UNIVERSITY

多项logit模型

- 应用统计软件，可以得到以上模型的参数估计和回归方程：

$$\begin{cases} \ln \dfrac{p_1}{1-p_1} = 0.449 + 1.319x_1 + 1.797x_2 \\ \ln \dfrac{p_1+p_2}{1-(p_1+p_2)} = 1.303 + 1.319x_1 + 1.797x_2 \end{cases}$$

- 统计分析结论如下：
  - 女性比男性的疗效好，其优势比为： $e^{1.319} = 3.798$
  - 新疗法比传统疗法好，其优势比为： $e^{1.797} = 6.032$

本次问卷中的案例 (以食堂满意度为例)

- 一般为多项逻辑模型，且响应变量为有序变量。

食堂环境对于选择该食堂就餐的影响程度					
性别E	年级(M1,M2)	影响程度			合计
		影响较大p1	有影响p2	无影响p3	
男E=0	2010级(1,0)				
	2011级(0,1)				
	2012级(-1,-1)				
女E=1	2010级(1,0)				
	2011级(0,1)				
	2012级(-1,-1)				

$$\begin{cases} \ln\dfrac{p_1}{1-p_1}=\beta_{10}+\beta_1x_1+\beta_2x_2 \\ \ln\dfrac{p_1+p_2}{1-(p_1+p_2)}=\beta_{20}+\beta_1x_1+\beta_2x_2 \end{cases}$$

重慶工商大學 CHONGQING TECHNOLOGY AND BUSINESS UNIVERSITY

重慶工商大学
CHONGQING TECHNOLOGY AND BUSINESS UNIVERSITY