合理进行多元分析——定性资料的判别分析

胡纯严[1]，胡良平[1,2*]

（1. 军事科学院研究生院，北京 100850；
2. 世界中医药学会联合会临床科研统计学专业委员会，北京 100029
*通信作者：胡良平，E-mail：lphu927@163.com）

【摘要】 本文目的是介绍与定性资料判别分析有关的基本概念、计算方法、两个实例以及SAS实现。基本概念包括定性变量与分类变量、判别分析、最大似然法、先验概率和先验分布、后验概率和后验分布；计算方法涉及最大似然判别法和Bayes公式判别法；两个实例的资料分别是“各类阑尾炎在各细目上出现的频率”以及“一例患者的症状和体征资料”；借助SAS，对两个实例的数据分别进行了判别分析，并对SAS输出结果做出了解释。

【关键词】 定性资料；判别分析；最大似然法；先验概率；贝叶斯公式
中图分类号：R195.1 文献标识码：A doi：10.11886/scjsws20230830004

Reasonably carry out multivariate analysis：qualitative data discrimination analysis

Hu Chunyan[1], *Hu Liangping*[1,2*]

（1. *Graduate School, Academy of Military Sciences PLA China, Beijing* 100850, *China*;
2. *Specialty Committee of Clinical Scientific Research Statistics of World Federation of Chinese Medicine Societies, Beijing* 100029, *China*
*Corresponding author: *Hu Liangping*, *E-mail*: *lphu*927@163.*com*）

【Abstract】 The purpose of the paper was to introduce the basic concepts, calculation methods, two examples and SAS implementation related to the qualitative data discrimination analysis. Basic concepts included qualitative and categorical variables, discriminant analysis, maximum likelihood method, prior probability and prior distribution, posterior probability and posterior distribution. The calculation methods involved the maximum likelihood discriminant method and Bayes formula discriminant approach. The data in the two examples were “the frequency of various types of appendicitis in each detail” and “the symptoms and signs of a patient”. With the help of SAS software, the discriminant analysis was carried out on the data in the two examples, and an explanation was made for the output results of SAS.

【Keywords】 Qualitative data; Discriminant analysis; Maximum likelihood method; Prior probability; Bayes formula

气象预报员需要根据现有的和历史的多项气象资料来判断未来几天的天气情况；医生需要根据就诊者的主诉和各项检查的结果以及医生的经验，对该就诊者可能患何种疾病进行诊断；地质学工作者需要根据地质勘探的多种结果判断某地区是否存在矿产以及矿产种类。显然，在做出判断之前，判断者需要掌握大量有关资料（即多个自变量及其取值），再基于某种原理构造出一个判别公式，从而实现对未知类别的个体进行判别。判别分析方法有多种，本文仅介绍基于最大似然法和Bayes（贝叶斯）公式法进行判别的方法[1-2]。

1 基本概念

1.1 定性变量与分类变量

“性别与职业”这样的变量被称为定性变量，当它们及其取值同时出现时，就称由它们组成的全部信息为定性资料。一般来说，性别的取值分为“男性”与“女性”，而职业的取值常分为“工人”“农民”“商人”和“军人”等。从实用性角度看，有些定性变量又被称为分类变量，因为按照定性变量的不同水平，可以将总体中的全部受试对象划分为互不相交、也不重叠的若干个子群体，每个子群体被称为一个类别。例如，可将总体中的全部个体划分为男性群体与女性群体；可将总体中的全部个体划分为A、B、AB、O型血的4个群体。

1.2 判别分析

判别分析就是基于n个个体的多个特征变量（例如身高、体重、胸围等）和一个分类变量及其取值，构建一个判别公式，并根据该判别公式对任何一个未知类别的个体进行分类（前提条件是未知个体归属于判别公式所定义的类别之一）。

1.3 最大似然法

用日常语言表述，“最大似然法”中的“最大”顾

名思义为“取得最大值”之意；而“似然”则是“概率”的代名词，在统计学上，只有“事件或随机变量”的发生或取某数值的机会大小可以用“概率或几率”来度量，而在“最大似然法”中，所研究的对象都是“模型中待估计的参数（它们是未知的、但有确定的数值）”。换言之，研究“参数”最可能取什么值的方法被称为“最大似然法”，而不是“最大概率法”。文献[3]给出了用数学语言表述最大似然法的详细内容。

1.4 先验概率和先验分布

先验概率是指在某随机事件发生前，依据经验或预试验的结果，预测该随机事件发生的概率；而对于一个随机变量而言，在试验或观察开始前，依据经验或预试验的结果，预测该随机变量可能具有何种概率分布，此分布即为先验分布。

$$\begin{cases} L_1 = P(S_1|A_1)P(S_2|A_1)\cdots P(S_m|A_1) = \prod_{j=1}^{m} P(S_j|A_1) \\ L_2 = P(S_1|A_2)P(S_2|A_2)\cdots P(S_m|A_2) = \prod_{j=1}^{m} P(S_j|A_2) \\ \cdots \\ L_g = P(S_1|A_g)P(S_2|A_g)\cdots P(S_m|A_g) = \prod_{j=1}^{m} P(S_j|A_g) \end{cases} \quad (1)$$

式(1)中，各$P(S_j|A_G)$(G=1,2,…,g;j=1,2,…,m)表示A_G类中因素X_j取值为S_j的条件概率，$\prod$为连乘符号。

如果各$P(S_j|A_G)$已知，则对于某一个体，就可按式(1)求得$L_1, L_2, \cdots, L_g$的值，挑选最大的L值，假定为L_f($f \in \{1, 2, \ldots, g\}$)，则判断该个体属于$L_f$类，这就是最大似然判别法。在实际工作中，各$P(S_j|A_G)$可根据一个足够大的样本的原始资料，算得各频率作为其估计值。

当m和g较大时，计算$L_1, L_2, \cdots, L_g$虽然不难，但得到的结果却全是小数，且位数较多。为此，可设法简化其计算。事实上，在式(1)中，我们只关心$L_1, L_2, \cdots, L_g$的相对大小，而不关心其具体数值。故将$L_G = \prod_{j=1}^{m} P(S_j|A_G)$两边取对数，见式(2)。

$$\lg L_G = \lg \prod_{j=1}^{m} P(S_j|A_G) = \sum_{j=1}^{m} \lg P(S_j|A_G) \quad (2)$$

要比较各L_G的大小，只需比较各$\lg L_G$的大小。与此同时，乘法计算已简化为加法计算。上式如果取近似值，还可进一步简化运算。经过一系列的代数运算和变形[4]，可以将式(2)转变成式(3)的形式。

$$H_G = \sum_{j=1}^{m} 10 \times [\lg P(S_j|A_G) + 1] \quad (G=1,2,\ldots,g) \quad (3)$$

1.5 后验概率和后验分布

基于先验概率或先验分布、试验或观察的结果，并按照特定问题所对应的计算公式，求出某随机事件或随机变量发生的概率或概率分布，它们被称为后验概率或后验分布。

2 计算方法

2.1 最大似然判别法

最大似然判别法建立在独立事件概率乘法定理的基础上。假设检测了受试对象的m个变量（或因素）$X_1, X_2, \cdots, X_m$，并已知受试对象可能的分类有$A_1, A_2, \cdots, A_g$类，则当某个个体的各因素取值为$S_1, S_2, \cdots, S_m$时，可根据概率论中独立事件的乘法定理求得似然值，见式(1)。

不同的$P(S_j|A_G)$值均可转换为相应的得分，利用这种得分可以大大简化运算，并使计算结果更加直观。通过比较各H_G值的大小，即可做出判断。

2.2 Bayes公式判别法

此法与最大似然判别法大同小异，例如，仍以$X_1, X_2, \cdots, X_m$表示因素，以$A_1, A_2, \cdots, A_g$表示类别，当某个体各因素值分别为$S_1, S_2, \cdots, S_m$时，该个体属于A_G类的概率$P(A_G|S_1S_2\cdots S_m)$可按式(4)算出。

$$P(A_G|S_1S_2\cdots S_m) = \frac{P(A_G)P(S_1|A_G)P(S_2|A_G)\cdots P(S_m|A_G)}{\sum_{G=1}^{g} P(A_G)P(S_1|A_G)P(S_2|A_G)\cdots P(S_m|A_G)} \quad (4)$$

式(4)中，$P(S_j|A_G)$意义同前，$P(A_G)$(G=1,2,…,g)则是最大似然判别法中没有的，称为先验概率，它是A_G类疾病发生的概率。

算得各$P(A_G|S_1S_2\cdots S_m)$后，比较其大小，如$P(A_f|S_1S_2\cdots S_m)$($f \in \{1, 2, \ldots, g\}$)最大，就判断该个体属于$A_f$类。本法不仅能作判别，由于求得了判为$A_f$类的后验概率，所作判断的可靠性就有了一个数量表示，而且应用了先验概率，还能提高鉴别的灵敏度。但有时先验概率不易得知。在大样本情况下，可用构成比代替各$P(A_G)$。有时也可将先验概

率取为相等，此时 Bayes 公式判别法与最大似然判别法所得结果就完全相同。

与最大似然判别法类似，也可简化 Bayes 公式判别法的计算。式(4)中，各 $P(A_G|S_1S_2\cdots S_m)$ 的分母相同，比较各 $P(A_G|S_1S_2\cdots S_m)$，只需比较各 $P(A_G)P(S_1|A_G)P(S_2|A_G)\cdots P(S_m|A_G)$ 的大小即可。如果记 $L_G^*=P(A_G)P(S_1|A_G)P(S_2|A_G)\cdots P(S_m|A_G)$，并对其两边取对数，得到式(5)。

$$\lg L_G^*=\lg[P(A_G)P(S_1|A_G)P(S_2|A_G)\cdots P(S_m|A_G)]$$
$$=\lg P(A_G)+\sum_{j=1}^{m}\lg P(S_j|A_G) \quad (5)$$

于是，比较各 L_G^* 的大小，只需比较各 $\lg L_G^*$ 的大小。此时，乘法计算已简化为加法计算。如果取近似值，还可进一步简化运算。经过一系列的代数运算和变形[4]，可将式(5)转变成式(6)的形式。

$$H_G^*=10\times[\lg P(A_G)+1]+\sum_{j=1}^{m}10\times[\lg P(S_j|A_G)+1] \quad (6)$$

式(6)中，$G=1,2,\cdots,g$。通过比较各 H_G^* 值的大小，即能做出判断。

3 实例与SAS实现

3.1 问题与数据结构

3.1.1 2个实际问题及数据

【例1】欲作阑尾炎四种类型的鉴别诊断，测得5 668例阑尾炎患者的症状发生频率见表1[5]。据此资料，寻找数据内部已经存在的某种规律，以便对阑尾炎的具体分类做出判断。为了实现此目的，请帮助研究者建立一种判断依据。

【例2】设现有一例阑尾炎患者各项症状及其细目如下，试根据例1所建立的判断依据判别该患者患的阑尾炎最可能属于哪种类型。上腹开始痛，恶心(+)呕吐(-)，24小时内正常排便，仅右下腹部有压痛，肌紧张(-)反跳痛(+)，体温37.3℃，白血球7 200×10⁹/L，这7种症状对应的细目编号从上到下依次为：3、7、9、12、15、18、20。

表1 各类阑尾炎在各细目上出现的频率

Table 1 Frequency of various types of appendicitis in various categories

体征、症状及化验项目	细 目	细目编号	频率(%)			
			卡他性	蜂窝组织炎性	坏疽性	腹膜炎
腹痛开始部位(X_1)	右下腹	1	57	34	35	21
	下腹	2	15	13	12	27
	上腹	3	12	35	35	34
	脐周围	4	12	10	9	6
	全腹	5	4	8	9	12
恶心呕吐(X_2)	恶心(-)呕吐(-)	6	73	33	8	13
	恶心(+)呕吐(-)	7	16	30	37	22
	呕吐(+)	8	11	37	55	65
排便情况(X_3)	24小时内正常排便	9	72	45	35	22
	24小时内无正常便或腹泻	10	20	40	55	34
	腹泻且有里急后重感	11	8	15	10	44
腹部压痛范围(X_4)	仅右下腹	12	95	93	81	9
	更广泛	13	5	7	19	91
肌肉紧张和反跳痛(X_5)	肌紧张(+)反跳痛(+)	14	8	39	79	96
	肌紧张(-)反跳痛(+)	15	70	34	12	3
	肌紧张(-)反跳痛(-)	16	22	27	9	1
体温(X_6)	<37℃	17	61	32	18	10
	37~38℃	18	31	57	59	46
	>38℃	19	8	11	23	44
白血球(X_7)	<10 000×10⁹/L	20	70	16	6	12
	10 000×10⁹/L~15 000×10⁹/L	21	22	56	33	31
	>15 000×10⁹/L	22	8	28	61	57

3.1.2 对数据结构的分析

例1的表1实际上是对资料的一种压缩的表达形式，它最主要的特点是：在同一张表中表达了同一批受试对象多个方面的情况。在本质上，表1资料属于结果变量为四值名义变量(即4种类型的阑尾炎)且含有7个自变量或称影响因素(即体征、症状及化验项目)的资料，具体的“细目”就是各自变量的具体取值，例如，腹痛开始部位 X_1 就有“右下

腹、下腹、上腹、脐周围、全腹”5个取值。例2沿用了例1的资料，此处不再赘述。

3.2 用SAS实现统计分析

3.2.1 分析例1的资料

设所需要的SAS程序如下[6-7]：

```
DATA a1;
infile 'D:\MXWTTJXS\lanweiyan.txt';
%LET v=7;
%LET m=4;
ARRAY c(&v);
ARRAY y(&v,&v,&v);
c(1)=5; c(2)=3;
c(3)=3; c(4)=2;
c(5)=3; c(6)=3;
c(7)=3;
DO i=1 TO &v;
DO k=1 TO c(i);
DO j=1 TO &m;
INPUT a @@;
b=10*(log10(a/100)+1);
y(i,k,j)=round(b,1);
OUTPUT;
END;
END;
END;
ODS HTML; FILE PRINT;
PUT @25'各种症状在各细目上的得分值';
PUT @10'----------
----------';
PUT @35 '得分值';
PUT @10'症状的细目' @22'----------
----------';
PUT @22'类型:卡他性' @35'蜂窝组织' @45'
坏疽性' @55'腹膜炎';
PUT @10'----------
----------';
DO i=1 TO &v;
DO k=1 TO c(i);
PUT @10 'x(' i ',' k ')'
y(i,k,1) 25-27 y(i,k,2) 35-37 y(i,k,3) 45-
47 y(i,k,4) 55-57;
END; END;
PUT @10'----------
----------';
RUN;
ODS HTML CLOSE;
```

【SAS程序说明】将表1中的22行4列数据输入计算机，创建一个名为lanweiyan.txt的文本格式文件，存放在D盘MXWTTJXS文件夹内。程序中定义了两个宏变量，其中v=7代表症状个数，m=4代表阑尾炎分类数，这两个数需要根据拟解决的不同问题而作相应调整。根据具体情况，还需调整的数据是各种症状之下的细目数，它们对应的语句是“$c(1)=5;c(2)=3;c(3)=3;c(4)=2;c(5)=3;c(6)=3;c(7)=3;$”，其含义是：症状1有5个细目、症状2有3个细目、……、症状7有3个细目。

【SAS输出结果及解释】各种症状在各细目上的得分见表2。

表2 各种症状在各细目上的得分

Table 2 Score values of various symptoms on each detail item

症状的细目	得分			
	卡他性	蜂窝组织炎性	坏疽性	腹膜炎
$x(1,1)$	8	5	5	3
$x(1,2)$	2	1	1	4
$x(1,3)$	1	5	5	5
$x(1,4)$	1	0	0	-2
$x(1,5)$	-4	-1	0	1
$x(2,1)$	9	5	-1	1
$x(2,2)$	2	5	6	3
$x(2,3)$	0	6	7	8
$x(3,1)$	9	7	5	3
$x(3,2)$	3	6	7	5
$x(3,3)$	-1	2	0	6
$x(4,1)$	10	10	9	0
$x(4,2)$	-3	-2	3	10
$x(5,1)$	-1	6	9	10
$x(5,2)$	8	5	1	-5
$x(5,3)$	3	4	0	-10
$x(6,1)$	8	5	3	0
$x(6,2)$	5	8	8	7
$x(6,3)$	-1	0	4	6
$x(7,1)$	8	2	-2	1
$x(7,2)$	3	7	5	5
$x(7,3)$	-1	4	8	8

基于表2中的结果，只要给定某患阑尾炎患者在表1中第一列上7种“体征、症状及化验项目”的具体细目，就可从表2中找到相应的行，该行中的4个“得分值”就是可能患阑尾炎的4种类型所对应的分值。将7种“体征、症状及化验项目”的具体细目对应的7行按“列”分别相加，就得到患4种阑尾炎的总分，判定该患者患有总分最高的一类阑尾炎。

3.2.2 分析例2的资料

设所需SAS程序如下[6-7]：

```
data score;
do r=1 to 22;
input ka feng huai fu @@;
if r=3 or r=7 or r=9 or r=12 or
r=15 or r=18 or r=20 then output; end;
cards;
8 5 5 3
（此处省略多行数据）
-1 4 8 8
;
run;
proc summary data=score;
var ka feng huai fu;
output out=sum_value sum=sum_ka sum_feng sum_huai sum_fu;
run;
proc print data=sum_value;
var sum_ka sum_feng sum_huai sum_fu;
run;
```

【SAS输出结果及解释】本病例与卡他性、蜂窝组织炎性、坏疽性、腹膜炎对应的四列的合计得分值分别为43、42、32、14，故判断此患者所患的阑尾炎类型最可能为卡他性或蜂窝组织炎性。

4 讨论与小结

4.1 讨论

最大似然判别法和Bayes公式判别法均适用于处理原因变量为定性变量的判别分析问题。实际使用中需注意以下问题：①最大似然判别法建立在独立事件概率乘法定理的基础上，它要求各因素（症状）X_j之间必须互相独立，各分类A_G之间必须互斥，而这些要求在临床上不易严格满足，特别是各X_j之间互相独立的条件更不符合临床实际，故此数学模型只能认为是近似的。②两种方法均须用频率来估计条件概率。此时要求样本足够大，如果样本不够大，所求出的得分值就不够稳定，近似程度亦较差。得分值可以简化计算，但它是近似值，所以会存在一定的误差。③由于最大似然判别法和Bayes公式判别法通常仅关心H_G值或H_G^*值的相对大小，选最大值作为判断依据。如果最大值与次大值相差较大，做出的判断较为可靠；但若相差不大，就必须谨慎，需进一步观察。④在评价最大似然判别法和Bayes公式判别法的效果时，不但要看总的符合率或正确率，还要考虑误诊和漏诊对各病的危害性。⑤如果某X_j的各类疾病的得分相差较大，则表明该因素对这几类疾病的鉴别价值大；反之，可考虑舍弃。

4.2 小结

本文介绍了与定性资料判别分析有关的基本概念、计算方法、两个实例以及用SAS实现计算的方法。基本概念包括定性变量与分类变量、判别分析、最大似然法、先验概率和先验分布、后验概率和后验分布；计算方法涉及最大似然判别法和Bayes公式判别法；两个实例的资料分别是“各类阑尾炎在各细目上出现的频率”和“一例患者的症状和体征资料”；借助SAS，对两个实例的数据分别进行了判别分析，并对SAS输出结果做出了解释。

参考文献

[1] 茆诗松. 统计手册[M]. 北京：科学出版社，2003：351-429，539-549.
Mao SS. Statistical manual[M]. Beijing: Science Press, 2003: 351-429, 539-549.

[2] 孙尚拱. 实用多变量统计方法与计算程序[M]. 北京：北京医科大学、中国协和医科大学联合出版社，1990：100-108.
Sun SG. Practical multivariate statistical methods and calculation programs[M]. Beijing: Beijing Medical University and China Union Medical University Joint Publishing House, 1990: 100-108.

[3] 胡良平，胡纯严，鲍晓蕾. 应用数理统计[M]. 北京：电子工业出版社，2015：185-192.
Hu LP, Hu CY, Bao XL. Applied mathematical statistics[M]. Beijing: Publishing House of Electronics Industry, 2015: 185-192.

[4] 胡良平. 面向问题的统计学：(2)多因素设计与线性模型分析[M]. 北京：人民卫生出版社，2012：579-591.
Hu LP. Problem-oriented statistics: (2) multifactor design and linear model analysis[M]. Beijing: People's Medical Publishing House, 2012: 579-591.

[5] 郭祖超. 医用数理统计方法[M]. 3版. 北京：人民卫生出版社，1988：498.
Guo ZC. Medical mathematical statistical methods[M]. 3rd edition. Beijing: People's Medical Publishing House, 1988: 498.

[6] 朱世武. SAS编程技术教程[M]. 北京：清华大学出版社，2007：185-217.
Zhu SW. SAS programming technology tutorial[M]. Beijing: Tsinghua University Press, 2007: 185-217.

[7] 胡良平，胡纯严. SAS语言基础与高级编程技术[M]. 北京：电子工业出版社，2014：23-278.
Hu LP, Hu CY. Fundamentals of SAS language and advanced programming techniques[M]. Beijing: Publishing House of Electronics Industry, 2014: 23-278.

（收稿日期：2023-08-30）
（本文编辑：陈 霞）