合理进行多元分析——定性资料对应分析和Shannon信息量分析

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对应分析和Shannon信息量分析有关的基本概念、计算方法、两个实例以及SAS实现。基本概念包括列联表与Burt表、边缘概率、行剖面与列剖面、信息和信息量、熵；计算方法涉及定性资料对应分析和Shannon信息量分析；两个实例分别为“某医院3年间不同季节4种甲状腺疾病的检出情况”和“不同专业学生的4种气质类型分布”；借助SAS软件，对两个实例中的数据分别进行了定性资料对应分析和Shannon信息量分析，并对SAS输出结果做出了解释。

【关键词】 列联表；边缘概率；行剖面；信息量；信息熵

中图分类号：R195.1 文献标识码：A doi：10.11886/scjsws20230726002

Reasonably carry out multivariate analysis: qualitative data correspondence analysis and Shannon information quantity 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*: *lphu927@163.com*)

【Abstract】 The purpose of this article was to introduce the basic concepts, calculation methods, two examples and SAS implementation related to the qualitative data correspondence analysis and Shannon information quantity analysis. Basic concepts included contingency table and Burt table, marginal probability, row profile and column profile, information and information volume, entropy. The calculation method involved the qualitative data correspondence analysis and Shannon information quantity analysis. The two examples were "detection of 4 kinds of thyroid diseases in different seasons in a hospital in 3 years" and "distribution of 4 temperament types of students in different majors". With the help of SAS software, qualitative data correspondence analysis and Shannon information quantity analysis were carried out on the data in the two examples, and an explanation was made for the SAS output results.

【Keywords】 Contingency table; Marginal probability; Row profile; Information volume; Information entropy

二维列联表资料是一种最常见的定性资料，卡方检验是处理这种资料的最常见的统计分析方法。然而，卡方检验的结果不能明确回答两个属性变量各水平之间是否存在一定程度的关联性问题。本文介绍定性资料对应分析和Shannon信息量分析，这两种分析方法在一定程度上弥补了卡方检验的不足。此外，定性资料对应分析还可以用于分析Burt表资料，以一种简化的方式实现对高维列联表资料的分析。

1 基本概念

1.1 列联表与Burt表

大样本定性资料通常以列联表的形式呈现。所谓列联表，就是将定性的原因和结果变量分别放置在表格的左边和表头上，各行与各列分别代表定性变量的具体水平，而行与列交叉处则是相应位置上出现的频数。当列联表中只有2个定性变量时，就称为二维列联表；当列联表中有k个($k\geq3$)定性变量时，就称为高维列联表[1]。在高维列联表资料中，若将横向和纵向上的多个定性变量分别进行水平组合，形成一个复合型定性变量，此时，就将高维表转化成为二维表了。在统计学上，称此种列联表为Burt表[2]。

1.2 边缘概率

在二维列联表中，分别求出各行与各列频数的合计，再求出总合计N。若分别用各行合计频数除

以N，就得到各行的频率，被称为行边缘概率；若分别用各列合计频数除以N，就得到各列的频率，被称为列边缘概率[3]。

1.3 行剖面与列剖面

在二维列联表中，设横向变量为A、纵向变量为B；又设A有n个水平、B有m个水平，若以各行上合计频数为分母，分别以各行上每个频数为分子，求出各行上m个相对数。第$i(i=1,2,\cdots,n)$行上m个相对数构成的一个行向量，被称为一个“行剖面”；同理，可得到第$j(j=1,2,\cdots,m)$列上n个相对数构成的一个列向量，被称为一个“列剖面”[4]。

1.4 信息和信息量

早年的信息与消息是同义词，而现今人们通常把信息看作由语言、文字、图象表示的新闻、消息或情报。信息是人类认识世界、改造世界的知识源泉。人类社会发展的速度在一定程度上取决于人类对信息利用的水平。信息、物质和能量被称为构成系统的三大要素。系统的状态往往具有多样性，例如生物多样性、环境多样性、人类社会活动的多样性等。信息是人们在认识多样性问题中所获得知识的反映，而知识总是与事物存在的多种状态及每个状态发生的可能性（随机性）有关。信息论中的信息是描述系统状态多样性丰富度的一个概念。信息量是指信息含量的多少，用来定量地描述信息。信息的获得与情况不确定度的减少相关。信息获得愈多，不确定度愈少；信息获得足够，不确定度为零[5]。

1.5 熵

设X是一个离散随机变量，它有m个可能的取值，记作$a_1,\cdots,a_m$，它们出现的概率分别为$p(a_1),\cdots,p(a_m)$。于是，统计学家用下式来定义熵，见式(1)。

$$H(X)=-\sum_{j=1}^{m}p\left(a_j\right)log_2p\left(a_j\right) \quad (1)$$

在信息论中，统计学家采用式(1)来度量随机变量X的平均信息量。

2 计算方法

2.1 定性资料对应分析

设拟分析的定性资料是一个二维列联表（包括标准的二维列联表和Burt表），则可以参照定量资料对应分析中的变量变换方法[6-7]，对表中的频数进行变换，基于变换后的数据构造矩阵Z，进而基于Z矩阵构造出两个协方差S_R和S_Q，分别对它们进行因子分析。在两次因子分析中，都取前两个公因子，以两个公因子为坐标轴，构成一个二维直角坐标系。可以证明，基于前述两个协方差矩阵导出的两个二维直角坐标系是重合的。于是，二维列联表横向上定性变量的各水平点（可视为“样品”点）与纵向上定性变量的各水平点（可视为“变量”点）可以呈现在同一个二维直角坐标系内[8-9]。

2.2 Shannon信息量分析

在热力学中，“熵”是系统无序度大小的度量。1948年，Shannon把熵的概念引入信息论中，他所定义的信息熵，实际上就是平均信息量。熵是系统的无序度的度量，而获得信息却使不确定度（熵）减少[5]。

对于只取有限个状态的随机变量$X=\{x_1,x_2,\cdots,x_n\}$，形成了一个状态空间，其状态称为信息符号。信息符号x_i出现的概率为$P_i(i=1,2,\cdots,n)$，即X的概率向量为$P=(P_1,P_2,\cdots,P_n)'$。包含信息符号出现概率的状态空间，称为信源，一般用以下方式表示，见式(2)。

$$[X,P]\text{或}X:\begin{bmatrix}x_1,x_2,...,x_n\\p_1,p_2,...,p_n\end{bmatrix} \quad (2)$$

式(2)中，$P_i\geq0,\sum_{i=1}^{n}P_i=1$。

可以证明，信息符号x_i的信息量是其概率的单调递减函数$f(P_i)$，见式(3)。

$$f(P_i)=-\log_bP_i \quad (3)$$

式(3)中，b的取值决定了信息量的单位，b=2，e，10，信息量的单位分别为bit（比特）、nat（奈特）和hart（哈特）。它们的换算关系见式(4)和式(5)。

$$1\text{ hart}=3.32\text{ bit} \quad (4)$$

$$1\text{ nat}=1.44\text{ bit} \quad (5)$$

如何定义信源式(2)中的整个信息量？Shannon的定义为各信息符号信息量的平均信息量（即信息熵），用$S(X)$表示，见式(6)。

$$S(X)=-\sum_{i=1}^{n}P_i\log_bP_i \quad (6)$$

通常情况下，均以nat为单位，见式(7)。

$$S(X)=-\sum_{i=1}^{n}P_i\ln P_i \quad (7)$$

由式(3)、式(6)和式(7)可以看出，Shannon信息量仅与信源的概率向量$P=(P_1,P_2,\cdots,P_n)'$有关，而与信息符号的具体状态获取值无关。

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】某医院观察了3年间不同季节中4种甲状腺疾病的检出情况，结果见表1[5]。试对此资料进行对应分析。

表1 某医院3年间不同季节4种甲状腺疾病的检出情况

Table 1 Detection of four thyroid disease in different seasons in a hospital in three years

甲状腺疾病分类	检出例数				
	季节：春季(C)	夏季(X)	秋季(Q)	冬季(D)	合计
甲亢(K)	411	451	294	284	1 440
亚甲炎(Y)	249	329	331	204	1 113
甲低(L)	60	61	59	52	232
甲状腺瘤(W)	45	50	46	40	181
合计	765	891	730	580	2 966

【例2】某大学对计算机专业、金融专业、传媒专业各50名学生进行心理测试，并判断每名学生属于哪一种典型气质类型，结果见表2[5]。试对此资料进行Shannon信息量分析。

表2 不同专业的大学生4种气质类型分布

Table 2 Distribution of four temperament types among college students of different majors

专 业	人数				
	气质类型：	多血质	胆汁质	抑郁质	黏液质
计算机		16	13	7	14
金融		12	15	10	13
传媒		18	9	8	14
合计		46	37	25	42

3.1.2 对数据结构的分析

例1中，甲状腺疾病分类和季节是两个不同的属性变量，前者可以被称为原因变量，但后者不应被称为结果变量，它只是人们关注的一种情境。表中的数据是两个属性变量不同水平组合下的“人数”，这种表为二维列联表。

例2中，专业和气质类型是两个不同的属性变量，前者可以被称为原因变量，后者可以被视为结果变量。表中的数据是两个属性变量不同水平组合下的“人数”，它也是一个二维列联表。

3.1.3 创建SAS数据集

分析例1资料，设所需SAS数据步程序如下：

```
data a1;
input disease $ C X Q D;
cards;
K 411 451 294 284 1440
Y 249 329 331 204 1113
L 60 61 59 52 232
W 45 50 46 40 181
;
run;
```

【SAS程序说明】disease代表“疾病类型”，K、Y、L、W分别代表“甲亢”“亚甲炎”“甲低”和“甲状腺瘤”；C、X、Q、D分别代表“春”“夏”“秋”“冬”。每个属性变量的每个水平都用一个字母表示，代表两个属性变量各水平的字母不应重复，以便在二维图上呈现两个属性变量不同水平组合下的关联性。

分析例2资料，设所需要的SAS数据步程序如下：

```
%let nr=3;
%let nc=5;
data a1;
do a=1 to &nr;
do b=1 to &nc;
input f @@;
output;
end;
end;
cards;
16 13 7 14
12 15 10 13
18 9 8 14
;
run;
```

【SAS程序说明】首先利用宏变量“nr”和“nc”分别指定列联表中行变量和列变量的水平数。通过数据步建立原始sas数据集“a1”，利用do→end循环语句和input→output语句，输入变量a、b、f，分别读入行变量、列变量、频数变量。

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过程步程序如下[2]：

```
proc corresp data= a1 OUTC=aaa;
var C X Q D;
id disease;
run;
%plotit(data=aaa, datatype=corresp, tsize=0.5,
```

color='black', href=0, vref=0)

【SAS输出结果及解释】各种疾病在两个公因子上的负荷见表3。在以dim1为横轴、以dim2为纵轴的直角坐标系内，每种疾病就是一个点，如“甲亢”点的坐标为(-0.103,-0.011)，显然，该点在第三象限。在四种疾病对应的四个点中，任何两点之间的欧氏距离都可以计算出来，并被标记在直角坐标系相应的位置上。

表3 各种疾病在两个公因子上的负荷

Table 3 Load of each disease on two common factors

疾 病	dim1	dim2
甲亢(K)	-0.103	-0.011
亚甲炎(Y)	0.126	-0.017
甲低(L)	0.016	0.094
甲状腺瘤(W)	0.023	0.072

与各种疾病对应的3个统计量的计算结果见表4。“质量”为每种疾病上两个公因子贡献率之和，此值接近1，表明对应的疾病信息由两个公因子就可很好地反映出来；“密度”为原始数据中各行数据之和占总合计的百分比；“惯量”为每种疾病对总特征值0.012贡献的比例。

表4 与各种疾病对应的3个统计量的计算结果

Table 4 Calculation results of three statistics corresponding to each disease

疾 病	质 量	密 度	惯 量
甲亢(K)	1.000	0.486	0.421
亚甲炎(Y)	1.000	0.375	0.491
甲低(L)	0.992	0.078	0.058
甲状腺瘤(W)	0.969	0.061	0.029

每个公因子在每种疾病上的贡献率见表5。各列数值之和为1。显然，“甲亢”和“亚甲炎”对第一公因子贡献最大；“甲低”和“甲状腺瘤”对第二公因子贡献最大。

表5 各公因子在各种疾病上的贡献率

Table 5 Contribution rates of common factors in different diseases

疾 病	dim1	dim2
甲亢(K)	0.461	0.054
亚甲炎(Y)	0.535	0.088
甲低(L)	0.002	0.590
甲状腺瘤(W)	0.003	0.269

各种疾病的坐标对特征值贡献最多的标志见表6。贡献少、中、多分别用0、1、2表示。

每种疾病对两个公因子各自的贡献率见表7。各行数值之和近似为1，因为只用了两个主要的公因子。由各行数值可看出，4种疾病都可以由这两个公因子比较好地反映出来。

同理，可以解释关于列变量(本例为“季节”)的

表6 各种疾病的坐标对特征值贡献最多的标志

Table 6 Indicators with the most contribution of coordinates to eigenvalues by disease

疾 病	dim1	dim2	最 佳
甲亢(K)	1	0	1
亚甲炎(Y)	1	0	1
甲低(L)	0	2	2
甲状腺瘤(W)	0	2	2

表7 每种疾病对两个公因子各自的贡献率

Table 7 Contribution rates of each disease to two common factors

疾 病	dim1	dim2
甲亢(K)	0.988	0.012
亚甲炎(Y)	0.983	0.017
甲低(L)	0.027	0.965
甲状腺瘤(W)	0.090	0.878

类似输出结果。因篇幅所限，此处从略，仅扼要说明如下：“春季”和“秋季”对第一公因子贡献较大；“夏季”和“冬季”对第二公因子贡献最大。绘制出反映本资料中各“疾病”与各“季节”之间关联性的二维图形，见图1。由图1可看出，K与C接近，意味着甲亢(K)易发于春季(C)；W与D接近，意味着甲状腺瘤(W)易发于冬季(D)；Y与Q接近，意味着亚甲炎(Y)易发于秋季(Q)。

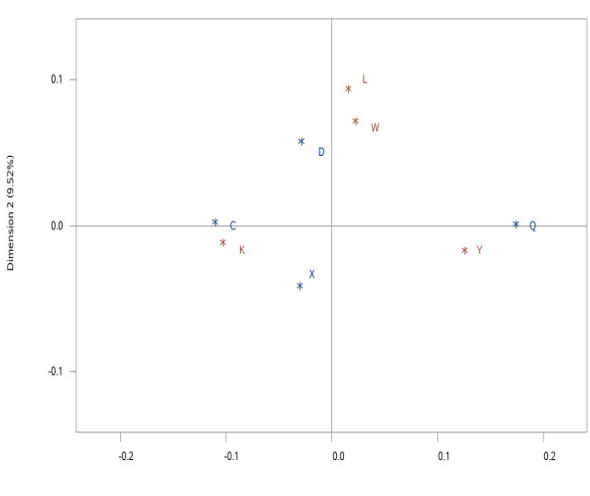

图1 “疾病”与“季节”之间的关联性

Figure 1 Association between "disease" and "season"

3.2.2 分析例2的资料

设所需要的SAS过程步程序如下[2]：

```
proc freq data=a1;
tables a*b/out=a2(drop=count percent) outpct noprint;
weight f;
run;
data a3(drop=pct_row pct_col);
```

```
set a2;
row=-(pct_row*log(pct_row/100)/100);
col=-(pct_col*log(pct_col/100)/100);
run;
proc sort data=a3;
by a b;
run; data a4(keep=a row_entropy);
set a3;
by a b;
if first. a then row_entropy=0;
row_entropy+row;
if last. a;
run;
proc sort data=a3;
by b a ;
run;
data a5(keep=b col_entropy);
set a3;
by b a ;
if first. b then col_entropy=0;
col_entropy+col;
if last. b;
run;
%macro print (dataset,title);
proc print data=&dataset noobs;
title &title;
run;
%mend;
%print(a4,"行变量的信息熵");
%print(a5,"列变量的信息熵");
```

【SAS程序说明】利用freq过程计算每个单元格对应的行百分比和列百分比。接下来的几个数据步和过程步，用于计算行变量和列变量不同取值水平所对应的Shannon信息量。最后使用1个名为“print”的宏，方便打印最终结果。该程序应用于其他类似的数据分析时，仅需修改第一个数据步程序前两行变量“nr”和“nc”的具体取值，并用新数据替换现在数据步中的原始数据即可。

【SAS输出结果及解释】行变量信息熵的计算结果见表8。表8是行变量“a(专业)”不同水平对应的信息熵“row_entropy(行信息熵)”的结果，熵值的大小关系为：a2>a1>a3，即“金融”优于“计算机”优于“传媒”。

列变量信息熵的计算结果见表9。表9是列变量“b(气质类型)”不同水平对应的信息熵“col_entropy(列信息熵)”的结果，熵值的大小关系为：b4>b3>b1>b2，即“黏液质”优于“抑郁质”优于“多血质”优于“胆汁质”。

表8 行变量信息熵

Table 8 Row variable information entropy

专 业	行信息熵
1	1.347
2	1.376
3	1.331

表9 列变量的信息熵

Table 9 Column variable information entropy

气质类型	列信息熵
1	1.085
2	1.077
3	1.088
4	1.097

4 讨论与小结

4.1 讨论

处理二维列联表资料最常见的统计分析方法是卡方检验，其目的是回答列联表中两个属性变量之间是否互相独立。若假设检验的结果为拒绝独立性，就意味着两个属性变量之间存在一定程度的关联。至于这种关联性的具体情况，卡方检验的结果无法给出明确判断。定性资料对应分析的结果能比较明确地显示行变量的某些水平与列变量的某些水平之间的关联性，但只能通过图形上的坐标点之间的接近程度来呈现，缺乏检验统计量来精确地度量。Shannon信息量分析可以显示各属性变量各水平所包含的信息量大小，但没有直接建立两个属性变量各水平之间的对应关系。因此，该方法无法明确解释哪些行与哪些列之间存在关联性。

此外，定性资料对应分析还可用于分析Burt表资料。也就是说，它可以回答两个复合型定性变量各水平之间是否存在一定程度的关联性，即以一种简化的方式实现对高维列联表资料的关联性分析。

4.2 小结

本文介绍了与定性资料对应分析和Shannon信息量分析有关的基本概念、计算方法、两个实例以及使用SAS实现计算的方法。基本概念包括列联表与Burt表、边缘概率、行剖面与列剖面、信息和信息量、熵；计算方法涉及定性资料对应分析和Shannon信息量分析；两个实例分别为“某医院3年间不同季

节4种甲状腺疾病的检出情况”和“不同专业学生的4种气质类型分布”；借助SAS软件，对两个实例中的数据分别进行了定性资料对应分析和Shannon信息量分析，并对SAS输出结果做出了解释。

参考文献

[1] Bishop YMM, Fienberg SE, Holland PW. 离散多元分析 理论与实践[M]. 张尧庭，译. 北京：中国统计出版社，1998：10-65.
Bishop YMM, Fienberg SE, Holland PW. Discrete multivariate analysis theory and practice [M]. Zhang YT, Translated. Beijing: China statistics Press, 1998: 10-65.

[2] SAS Institute Inc. SAS/STAT®15.1 user's guide[M]. Cary, NC: SAS Institute Inc, 2018: 2514-2578, 2997-3216.

[3] 高惠璇. 应用多元统计分析[M]. 北京：北京大学出版社，2005：324-342.
Gao HX. Applied multivariate statistical analysis[M]. Beijing: Peking University Press, 2005: 324-342.

[4] 何晓群. 多元统计分析[M]. 2版. 北京：中国人民大学出版社，2008：227-254.
He XQ. Multivariate statistical analysis[M]. 2nd edition. Beijing: China Renmin University Press, 2008: 227-254.

[5] 胡良平. 面向问题的统计学：(3)试验设计与多元统计分析[M]. 北京：人民卫生出版社，2012：286-302.
Hu LP. Problem-oriented statistics: (3) experimental design and multivariate statistical analysis[M]. Beijing: People's Mental Publishing House, 2012: 286-302.

[6] Johnson RA, Wichern DW. 实用多元统计分析[M]. 6版. 北京：清华大学出版社，2008：481-538.
Johnson RA, Wichern DW. Applied multivariate statistical analysis[M]. 6th edition. Beijing: Tsinghua University Press, 2008:481-538.

[7] 王静龙. 多元统计分析[M]. 北京：科学出版社，2008：360-375.
Wang JL. Multivariate statistical analysis[M]. Beijing: Science Press, 2008: 360-375.

[8] 李卫东. 应用多元统计分析[M]. 北京：北京大学出版社，2008：239-258.
Li WD. Applied multivariate statistical analysis[M]. Beijing: Peking University Press, 2008: 239-258.

[9] 余锦华，杨维权. 多元统计分析与应用[M]. 广州：中山大学出版社，2005：232-250.
Yu JH, Yang WQ. Multivariate statistical analysis and application [M]. Guangzhou: Sun Yat-sen University Press, 2005: 232-250.

（收稿日期：2023-07-26）
（本文编辑：陈 霞）