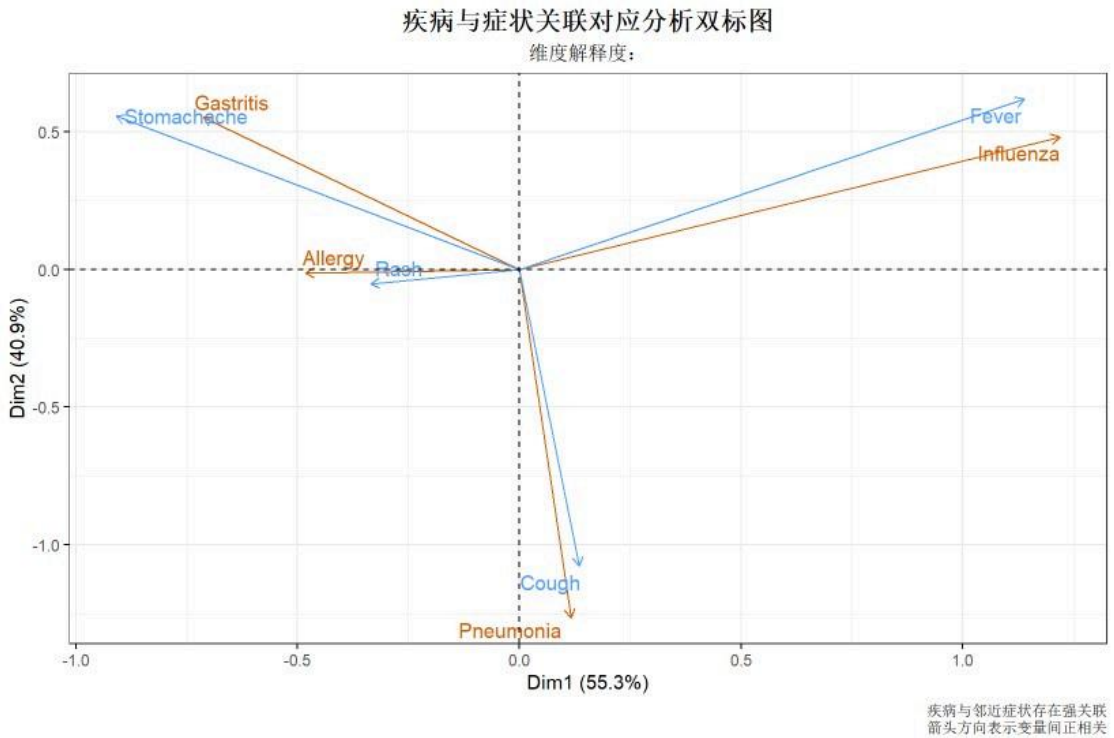
第10章 对应分析及R使用

武慧
wuh@hit.edu.cn

经济管理学院
哈尔滨工业大学（威海）

1

疾病与症状关联可视化



如何从医疗数据中挖掘疾病与症状的典型组合？

- 疾病（橙色）：Influenza：流感，Pneumonia：肺炎，Gastritis：胃炎，Allergy：过敏。
- 症状（蓝色）：Fever：发热、发烧，Cough：咳嗽，Stomachache：胃痛，Rash：皮疹。
- 箭头显示变量间相关关系分布。

2

本章内容

1. 对应分析的基础概念
2. 对应分析的基本原理
2. 对应分析的R实现

3

1. 对应分析的基础概念

4

回顾列联表分析

- 列联表（也称为交叉表）是一个二维矩阵，每一列和每一行都代表一个分类变量的不同类别，表格中的每个单元格包含了对应类别组合的观测频次。
- 列联表常用于分析分类数据，特别是用于检验不同类别变量之间是否有统计上的关联。
- 例如，调查了100名消费者了解他们的性别和是否喜欢某品牌的饮料的关联，可以构建一个2×2列联表：

性别＼是否喜欢饮料	喜欢	不喜欢	总计
男性	30	20	50
女性	40	10	50
总计	70	30	100

回顾列联表分析步骤

(1) 构建列联表：首先需要收集数据并将其按照类别分组，形成列联表。

(2) 计算卡方检验统计量：用于检验两个分类变量之间是否有显著的关联（原假设：变量之间相互独立）。卡方检验基于观察频数与期望频数之间的差异来计算卡方统计量：

$$\chi^2 = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

其中，O_{ij} 是观察频数，E_{ij} 是期望频数。

$$E_{ij} = \frac{(\text{行总和}) \times (\text{列总和})}{\text{总样本数}}$$

(3) 计算p值：通过卡方统计量和自由度（自由度=（行数-1）×（列数-1））查找卡方分布表，获得p值。p值用于判断是否拒绝原假设。

6

回顾列联表分析步骤：示例

- 构建列联表：

性别＼是否喜欢饮料	喜欢	不喜欢	总计
男性	30	20	50
女性	40	10	50
总计	70	30	100

- 计算期望频数：

$$E_{ij} = \frac{(\text{行总和}) \times (\text{列总和})}{\text{总样本数}}$$

性别＼是否喜欢饮料	喜欢	不喜欢	总计
男性	35	15	50
女性	35	15	50
总计	70	30	100

回顾列联表分析步骤：示例

- 计算卡方统计量：$\chi^2 = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$

对于“男性喜欢”的单元格：$\frac{(30-35)^2}{35} = \frac{25}{35} \approx 0.714$

对于“男性不喜欢”的单元格：$\frac{(20-15)^2}{15} = \frac{25}{15} \approx 1.667$

对于“女性喜欢”的单元格：$\frac{(40-35)^2}{35} = \frac{25}{35} \approx 0.714$

对于“女性不喜欢”的单元格：$\frac{(10-15)^2}{15} = \frac{25}{15} \approx 1.667$

因此，卡方统计量为：

$$\chi^2 = 0.714 + 1.667 + 0.714 + 1.667 = 4.762$$

8

回顾列联表分析步骤：示例

- 计算p值：

计算自由度：$\mathrm{df} = (\text{行数} - 1) \times (\text{列数} - 1) = (2-1) \times (2-1) = 1$

根据自由度为1的卡方分布，查找卡方值4.762对应的p值：0.029。这个 p 值表示在零假设成立（即性别和是否喜欢该品牌饮料没有关系）下，观察到这么大的卡方统计量的概率。

- 由于p值0.029小于显著性水平0.05，我们拒绝原假设，这意味着性别和是否喜欢该品牌饮料之间存在显著的关联。

9

对应分析

- 在分析列联表时，常常使用卡方检验来检验两个分类变量之间是否存在统计学上的显著关联。
- 尽管卡方检验可以告诉我们两个变量是否独立，但它并不能提供关于这些变量之间如何相互关联的详细信息。
- 对应分析是一种基于列联表数据的降维技术，能够在低维空间中可视化分类变量的关系。它通过将高维的列联表数据转换成二维或三维的坐标系来帮助我们理解变量之间的关系。

10

引例

假设我们有一项市场研究数据，数据集记录了顾客群体与不同品牌（如A、B、C、D等）之间的购买行为。我们希望通过对应分析找出顾客偏好背后的模式，并了解顾客群体与品牌之间的关联性。数据以列联表的形式呈现，列联表反映了不同顾客群体在不同品牌上的购买数量。

顾客群体\品牌	品牌A	品牌B	品牌C	品牌D
年轻人	50	30	20	10
中年人	20	40	30	10
老年人	10	20	50	20

11

单选题 1分

在列联表中，行和列分别表示的是：

A 不同时间点的观测数据

B 两个变量的类别或分组

C 数值型数据和分类数据

D 变量的回归系数和残差

12

单选题 1分

卡方检验的主要目的是：

A 比较不同组之间的平均值

B 检查两个或更多类别变量之间是否存在显著的关联

C 计算数据的标准差

D 确定数据分布的类型

13

单选题 1分

在进行卡方检验时，期望频数的计算公式为：

A 期望频数 = 实际频数 / 总样本数

B 期望频数 = 列总频数 × 总样本数

C 期望频数 = 行总频数 × 列总频数

D 期望频数 = (行总频数 × 列总频数) / 总样本数

14

单选题 1分

在对应分析中，如何理解“对应关系”？

A 变量之间的线性关系

B 分类变量之间的相互依赖关系

C 两个变量的直接因果关系

D 对象的相似性

15

2. 对应分析的基本原理

16

对应分析简介

- 对应分析是是一种用于分析和可视化分类数据之间关系的多变量统计方法，尤其适用于分析列联表（也称为交叉表）中的频数数据。
- 它通过将数据从高维空间映射到低维空间（通常是二维或三维），帮助我们理解分类变量之间的关系，揭示潜在的模式和关联。

对应分析步骤

（1）构建列联表：

$$O = \begin{pmatrix} O_{11} & O_{12} & \cdots & O_{1n} \\ O_{21} & O_{22} & \cdots & O_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ O_{m1} & O_{m2} & \cdots & O_{mn} \end{pmatrix}$$

其中 m 是行类别的数目，n 是列类别的数目，

O_{ij} 是第 i 行和第 j 列的观测频数（即实际观察到的频率）。

18

对应分析步骤

（2）标准化频数矩阵：将列联表中的每个元素O_{ij}转化为相对频率P_{ij}，即标准化后的矩阵。标准化矩阵$\boldsymbol{P}$的每个元素为：

$$P_{ij} = \frac{O_{ij}}{N} \qquad N = \sum_{i=1}^{m}\sum_{j=1}^{n} O_{ij}$$

通过将数据标准化为相对频率，消除了行和列类别总和的影响，提高了类别间的相对重要性，便于揭示各类别之间的相对关系和结构。

19

对应分析步骤

（3）计算行和列的边际分布

$$r_i = \sum_{j=1}^{n} P_{ij} = \frac{\sum_{j=1}^{n} O_{ij}}{N}$$

$$c_j = \sum_{i=1}^{m} P_{ij} = \frac{\sum_{i=1}^{m} O_{ij}}{N}$$

对应分析步骤

（4）构造标准化残差矩阵：

$$Z_{ij} = \frac{P_{ij} - e_{ij}}{\sqrt{e_{ij}}} \qquad e_{ij} = r_i \cdot c_j$$

- 残差矩阵$\boldsymbol{Z} = (Z_{ij})$反映了每个单元格的实际观察值和期望值之间的标准化偏差。
- 它表示了行和列变量之间的关联性。较大的残差绝对值通常表明该单元格的观测值与期望值之间的差异较大，可能反映了某种显著的关联。

对应分析步骤

（5）将标准化残差矩阵Z进行奇异值分解（SVD）：

$$Z = U\Sigma V^T$$

- U 是 $m \times r$ 的矩阵，包含了行类别的主成分（每个行类别在新空间中的坐标）。
- Σ 是 $r \times r$ 的对角矩阵，包含了奇异值，表示不同维度的贡献度。
- V 是 $n \times r$ 的矩阵，包含了列类别的主成分（每个列类别在新空间中的坐标）。

- 其中，r是矩阵Z的秩，通常$r \leq \min(m, n)$。
- 通过奇异值分解，我们能够得到最重要的主成分，这些主成分能够解释数据中大部分的变异性。

22

对应分析步骤

（6）计算主成分坐标（标准化）：对应分析的目标是将行和列的类别映射到低维空间中。通过奇异值分解，我们可以得到行和列的坐标：

行的坐标（主成分）由矩阵 $U\Sigma$ 给出。

列的坐标（主成分）由矩阵 $V\Sigma$ 给出。

- 每个行和列的点可以在二维或三维空间中表示出来，反映它们之间的相似性或关联性。
- 在这个图中，距离较近的点表示这两个类别之间的关系较强，而距离较远的点则表示它们之间的关系较弱。

23

单选题 1分

对应分析的主要目的是

A 识别变量之间的线性关系

B 分析多个分类变量之间的关系并降维展示

C 测量变量的标准差

D 对样本数据进行回归分析

24

单选题 1分

在对应分析中，"列联表"的作用是：

A 展示每个类别的频率分布

B 提供不同变量之间的交叉频次数据

C 计算每个变量的平均值

D 描述数据的相关性

25

单选题 1分

以下哪项描述最适合对应分析的主要用途?

A 用于分析两个定量变量之间的关系

B 用于分析时间序列数据的趋势

C 用于探索分类数据之间的结构和关系

D 用于预测连续型因变量

26

单选题 1分

对应分析的图表中，点的距离可以用来衡量：

A 变量之间的线性关系

B 行和列类别之间的相似度或关联强度

C 卡方检验的统计量

D 观测数据的离散程度

27

3. 对应分析的R实现

28

创建列联表

```
data <- matrix(c(50,30,20,10,
    20,40,30,10,
    10,20,50,20), nrow = 3, byrow = TRUE)
colnames(data) <- c("品牌A","品牌B","品牌C","品牌D")
rownames(data) <- c("年轻人", "中年人", "老年人")
```

```
> data
       品牌A 品牌B 品牌C 品牌D
年轻人    50    30    20    10
中年人    20    40    30    10
老年人    10    20    50    20
```

29

执行对应分析

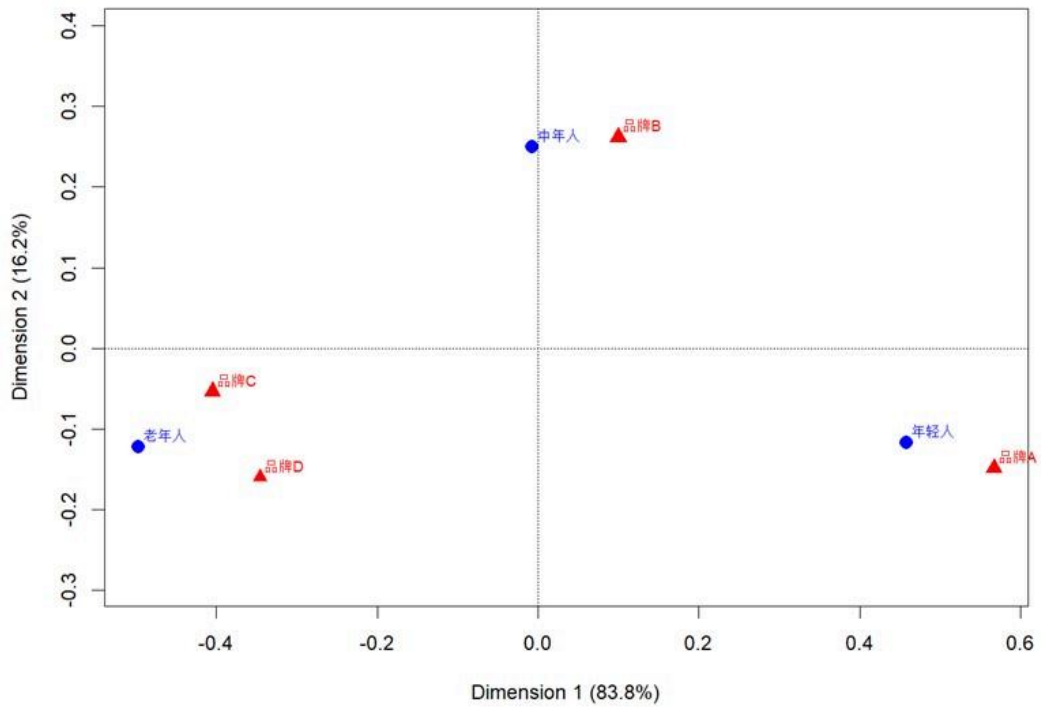
```
library(ca)
# 执行对应分析
ca_result <- ca(data)
# 输出奇异值
ca_result$sv
# 输出行类的标准坐标
ca_result$rowcoord
# 输出列类的标准坐标
ca_result$colcoord
```

```
> # 输出奇异值
> ca_result$sv
[1] 0.3924510 0.1726133
> # 输出行类的标准坐标
> ca_result$rowcoord
                Dim1       Dim2
年轻人  1.16698824 -0.6755148
中年人 -0.01813445  1.4490242
老年人 -1.26555262 -0.7059579
> # 输出列类的标准坐标
> ca_result$colcoord
              Dim1       Dim2
品牌A  1.4438490 -0.8584820
品牌B  0.2540506  1.5176094
品牌C -1.0315148 -0.3092152
品牌D -0.8805249 -0.9246193
```

30

可视化分析结果

```
plot(ca_result) # 绘制二维图
```



31

单选题 1分

在R中，进行对应分析时，ca()函数的输入通常是：

A 数字型数据矩阵

B 列联表（或交叉表）

C 时间序列数据

D 单一变量数据集

32

单选题 1分

在R中使用ca()函数进行对应分析时，如何绘制二维可视化图（即双重图）？

A 使用plot(ca)函数

B 使用biplot(ca)函数

C 使用ca.plot()函数

D 使用ca$plot()函数

33

单选题 1分

如果在对应分析的散点图中，两个点的距离非常近，这意味着：

A 这两个类别之间的关系较弱

B 这两个类别之间没有关系

C 这两个类别之间的关系较强

D 这两个类别属于同一个类别

34

单选题 1分

在对应分析中，如果行和列类别的点在散点图上分布得很分散，这意味着：

A 行和列类别之间没有显著关系

B 行和列类别之间有强关系

C 行和列类别之间有不确定关系

D 行和列类别的数据不完整

35

本章小结

- 对应分析的基础概念：列联表、卡方检验、对应分析的概念。
- 对应分析的基本原理：一种用于分析两个或多个分类变量之间的关系的多元统计方法，通过将高维的列联表数据降维到低维空间，揭示变量间的关联结构和分布模式。
- 对应分析的R实现：ca::ca。通过对应分析的结果，可以绘制散点图，其中行和列类别的坐标通过低维空间中的点表示。类别之间的关系通过点之间的距离来体现。