第 6 章　分类数据分析

内容提要

6.1 分类数据与χ^2统计量

分类数据

对事物进行分类的结果。

类型

- 无序分类：性别，…
- 有序分类：产品等级，…

分析

- 单一分类变量的特征分析

  集中趋势、离散趋势（第4章）
- 某一分类变量观测频数与期望频数的**相似性**：拟合优度检验
- 两个分类变量的**相关性**：列联表-独立性检验

χ^2统计量

- **重要性**

  拟合优度检验和列联表分析的基础
- **公式**

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$$

其中：f_o-观察频数；f_e-期望频数

- **说明**

  观测频数与期望频数越接近，$(f_o - f_e)$绝对值越小，χ^2越小，越不显著。

6.2 拟合优度检验

原理

依据总体分布状况，计算出分类变量中各个类别的期望频数，与观察频数进行对比，**判断期望频数与观察频数之间是否存在显著差异**。

适用

分类变量**观测值**与**期望值**之间**相似性**判断

拟合优度检验公式

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e} \sim \chi^2(R-1)$$

说明

- f_o-观察频数； f_e-期望频数
- R为分类变量的类型个数。比如性别有男女2个类别，R=2
- 观测频数与期望频数差异越小（越相似），χ^2越不显著。

例子：拟合优度检验

例子

1912年4月15日，豪华巨轮泰坦尼克号与冰山相撞沉没。当时船上共有共2208人，其中男性1738人，女性470人。海难发生后，幸存者为718人，其中男性374人，女性344人，以$\alpha = 0.1$的显著性水平检验：存活状况与性别是否有关。

问题:为什么存活状况与性别显著相关？

拟合优度检验

- **假设**

$\boldsymbol{H_0}$:观察频数与期望频数一致

$\boldsymbol{H_1}$:观察频数与期望频数不一致

- **检验过程**

χ^2 计算表

		步骤一	步骤二	步骤三
f_0	f_e	$f_0 - f_e$	$(f_0 - f_e)^2$	$(f_0 - f_e)^2 / f_e$
374	565	-191	36481	64.6
344	153	191	36481	238.4
	步骤四	$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e} = 303$		

查询χ^2分布表知：临界值$\chi^2_{0.1}(1) = 2.706$，$\chi^2 > \chi^2_{0.1}(1)$，故拒绝H_0，不得不接受H_1，即存活状况与性别显著相关。

6.3 列联分析：独立性检验

列联表的定义

由两个以上的分类变量交叉分类的频数分布表。

说明

- 行变量的类别用r表示：r_i表示第i个类别
- 列变量的类别用c表示：c_j表示第j个类别
- f_{ij}表示第i行第j列的观察频数
- 一个r行c列的列联表称为r×c列联表

列联分析的原理

将两个以上分类变量交叉形成列联表，检验列联表的**行变量**和**列变量**是否**相互独立**。（独立性检验）

行(r_i)	列(c_j)			合计
	j =1	j = 2	…	
i =1	f_{11}	f_{12}	…	r_1
i = 2	f_{21}	f_{22}	…	r_2
⋮	⋮	⋮	⋮	⋮
合计	c_1	c_2	…	n

例子：列联分析-独立性检验

例子

一种原料来自三个不同的地区，原料质量被分成三个不同等级。从这批原料中随机抽取500件进行检验，结果如下表所示，要求检验各个地区和原料质量之间是否存在依赖关系？（$\alpha = 0.05$）

表 9—3　　原料抽样的结果

	一级	二级	三级	合计
甲地区	52	64	24	140
乙地区	60	59	52	171
丙地区	50	65	74	189
合计	162	188	150	500

列联分析：独立性检验

- 假设

H_0：地区与原料等级之间是独立的（不存在依赖关系）

H_1：地区与原料等级之间是不独立（存在依赖关系）

- 检验过程

表 9—4　　3×3 列联表期望值及 χ^2 计算结果

行	列	f_o	f_e	$f_o - f_e$	$(f_o - f_e)^2$	$(f_o - f_e)^2 / f_e$
1	1	52	45.36	6.64	44.09	0.97
1	2	64	52.64	11.36	129.05	2.45
1	3	24	42.00	−18	324	7.71
2	1	60	55.40	4.60	21.16	0.38
2	2	59	64.30	−5.3	28.09	0.44
2	3	52	51.30	0.7	0.49	0.01
3	1	50	61.24	−11.24	126.34	2.06
3	2	65	71.06	−6.06	36.72	0.52
3	3	74	56.70	17.30	299.29	5.28
						19.82

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e} = 19.82$$

χ^2 的自由度 $= (R-1)(C-1) = 4$

令 $\alpha = 0.05$，查表知：$\chi^2_{0.05}(4) = 9.488$

由于 $\chi^2 > \chi^2_{0.05}(4)$，故拒绝 H_0，接受 H_1，即地区和原料等级之间存在依赖关系，原料的质量受地区的影响。

6.4 列联表中的相关测量

φ相关系数

公式 $\varphi = \sqrt{\frac{\chi^2}{n}}$

说明

- 适合2×2的列联表
- n为列联表总频数
- 取值: [0,1]
- φ=0，不相关
- φ=1，完全相关
- φ取值可能为负值，正负号没有意义

列联相关系数(C系数)

公式 $\mathbf{C} = \sqrt{\frac{\chi^2}{\chi^2+n}}$

说明

- 适合大于2×2的列联表
- n为列联表总频数
- 取值: [0,1)
- C=0，不相关
- C越大，相关程度越高
- 行数和列数越大，c越大

V相关系数

公式 $\mathbf{V} = \sqrt{\frac{\chi^2}{n*min[(R-1),(C-1)]}}$

说明

- 适合所有列联表
- n为列联表总频数
- R为行数，C列数
- 取值: [0,1]
- V=0，不相关
- V=1，完全相关
- 当R=2或C=2时，V=φ

例子：列联表中的相关测量

例子

一种原料来自三个不同的地区，原料质量被分成三个不同等级。从这批原料中随机抽取500件进行检验，结果如下表所示。分别计算φ相关系数、C系数和V相关系数，并分析相关程度。

表 9—3　　原料抽样的结果

	一级	二级	三级	合计
甲地区	52	64	24	140
乙地区	60	59	52	171
丙地区	50	65	74	189
合计	162	188	150	500

分析

已知n=500，$\chi^2 = 19.82$，列联表为3×3

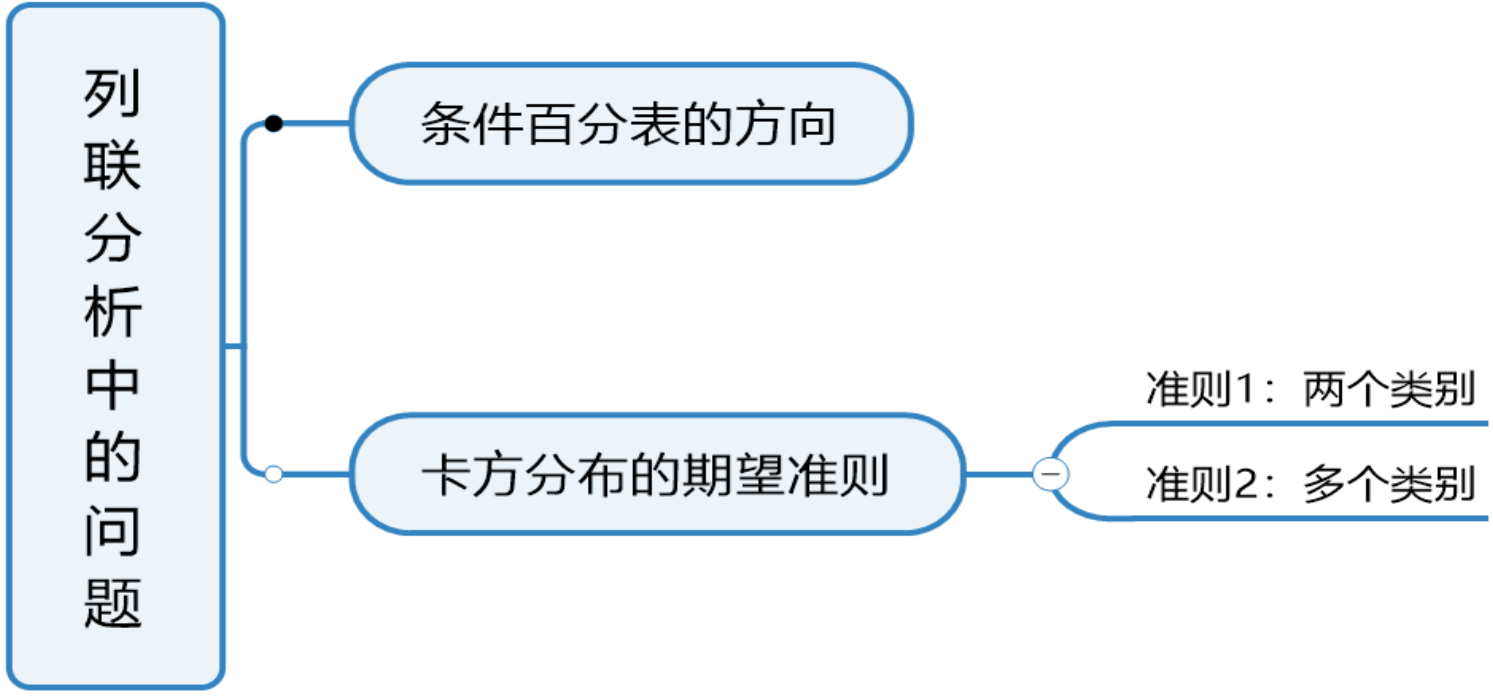
$$\varphi = \sqrt{\frac{\chi^2}{n}} = \sqrt{\frac{19.82}{500}} = 0.199$$

$$\mathbf{c} = \sqrt{\frac{\chi^2}{\chi^2 + n}} = \sqrt{\frac{19.82}{19.82 + 500}} = 0.195$$

$$\mathbf{V} = \sqrt{\frac{\chi^2}{n * min[(R-1),(C-1)]}} = \sqrt{\frac{19.82}{500 * 2}} = 0.141$$

结论：三个系数均不高，表明产地和原料等级之间的相关程度不高。

6.5 列联分析中应注意的问题



条件百分表的方向

- 通常，列联表中的行列位置是任意的
- 例外：若列联变量之间存在**因果关系**，则**自变量**放在**列**的位置，**因变量**放在**行**的位置。

表 9—12　　职业与价值取向

价值取向 Y	职业 X	
	制造业	服务业
物质报酬	105	45
%	72	56
人情关系	40	35
%	28	44
合计	145	80
%	100	100

χ^2分布的期望值准则

- 一般原则：利用χ^2分布进行独立性检验时，样本必须足够大，每个单元格中期望频数不能太小。
- **准则1**：若分类变量有**两个类别**，则**期望频数**必须**大于等于5**时，才能应用χ^2检验。

表 9-15　说明表（一）

以往病史	f_o	f_e
未患过肝炎	532	531
患过肝炎	4	5

分析：符合准则1，χ^2检验可用。

χ^2分布的期望值准则

- **准则2**：若分类变量的**超过2个类别，多于20%类型**的**期望频数小于5**时，不能应用卡方检验。

表 9-16　说明表（二）

类别	f_o	f_e
A	28	26
B	49	47
C	18	23
D	6	4
E	92	88
F	20	25
合计	213	213

分析：符合准则2，χ^2检验可用。

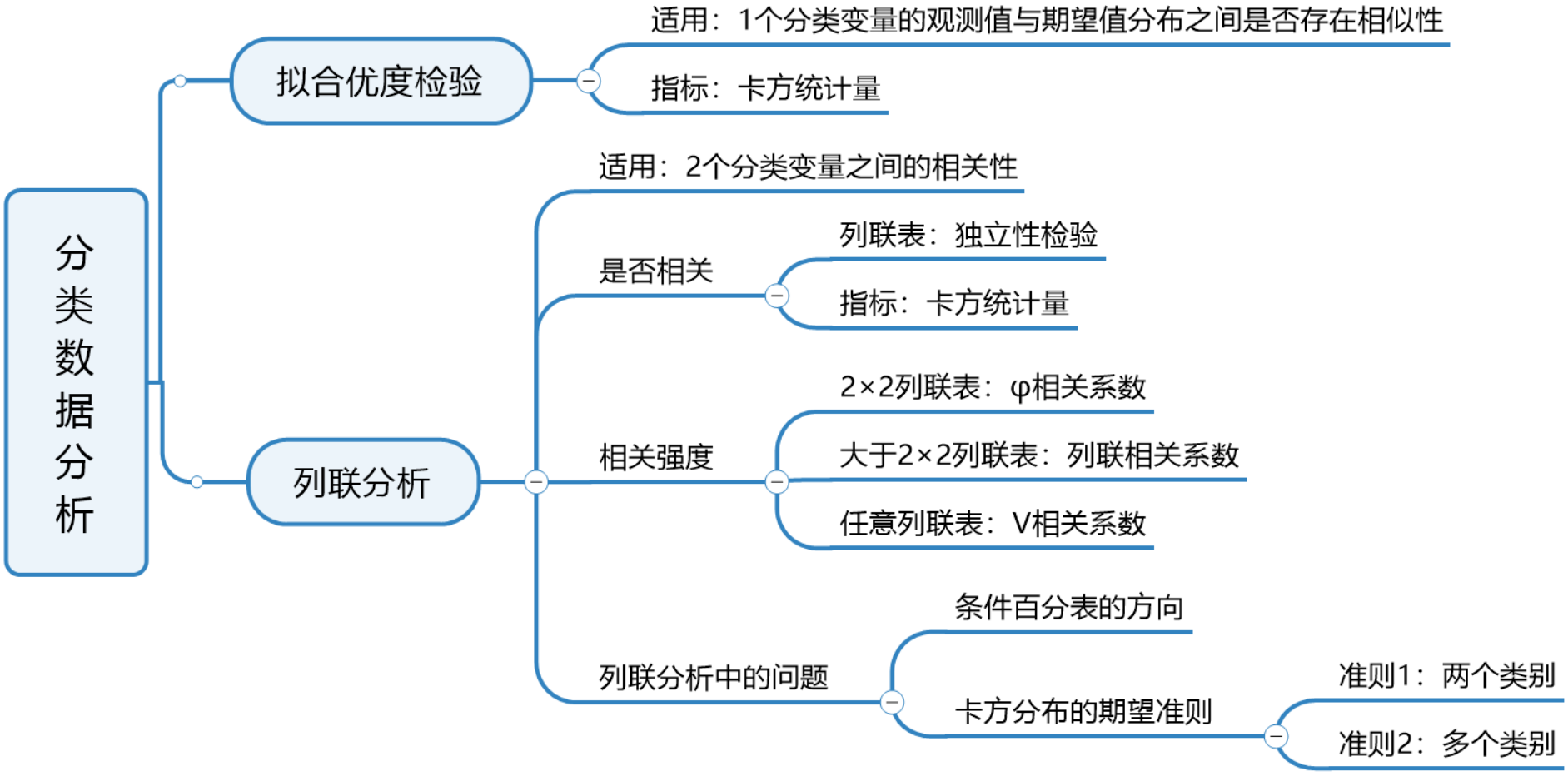
表 9-17　说明表（三）

类别	f_o	f_e
A	30	32
B	110	113
C	86	87
D	23	24
E	5	2
F	5	4
G	4	1
合计	263	263

分析：不符合准则2，χ^2检验不可用

解决：合并E,F,G三种类型

本章小结


分类数据分析
拟合优度检验
适用：1个分类变量的观测值与期望值分布之间是否存在相似性
指标：卡方统计量
列联分析
适用：2个分类变量之间的相关性
是否相关
列联表：独立性检验
指标：卡方统计量
相关强度
2×2列联表：φ相关系数
大于2×2列联表：列联相关系数
任意列联表：V相关系数
列联分析中的问题
条件百分表的方向
卡方分布的期望准则
准则1：两个类别
准则2：多个类别