第 9 章　分类数据分析

PowerPoint

第 9 章　分类数据分析

9.1 分类数据与χ^2统计量
9.2 拟合优度 检验
9.3 列联分析：独立性检验
9.4 列联表中的相关度量
9.5 列联分析中应注意的问题

学习目标

1. 理解分类数据与χ^2 统计量
2. 掌握拟合优度检验及其应用
3. 掌握独立性检验及其应用
4. 列联表中的相关度量
5. 掌握测度列联表中的相关性

9.1 分类数据与列联表

9.1.1 分类数据

9.1.2 χ^2统计量

作者：贾俊平，中国人民大学统计学院

分类数据

作者：贾俊平，中国人民大学统计学院

分类数据

1. 分类变量的结果表现为类别
   - 例如：性别 (男, 女)
2. 各类别用符号或数字代码来测度
3. 使用分类或顺序尺度
   - 你吸烟吗?
     - 1.是；2.否
   - 你赞成还是反对这一改革方案?
     - 1.赞成；2.反对
4. 对分类数据的描述和分析通常使用列联表
5. 可使用 χ^2检验

χ^2 统计量

作者：贾俊平，中国人民大学统计学院

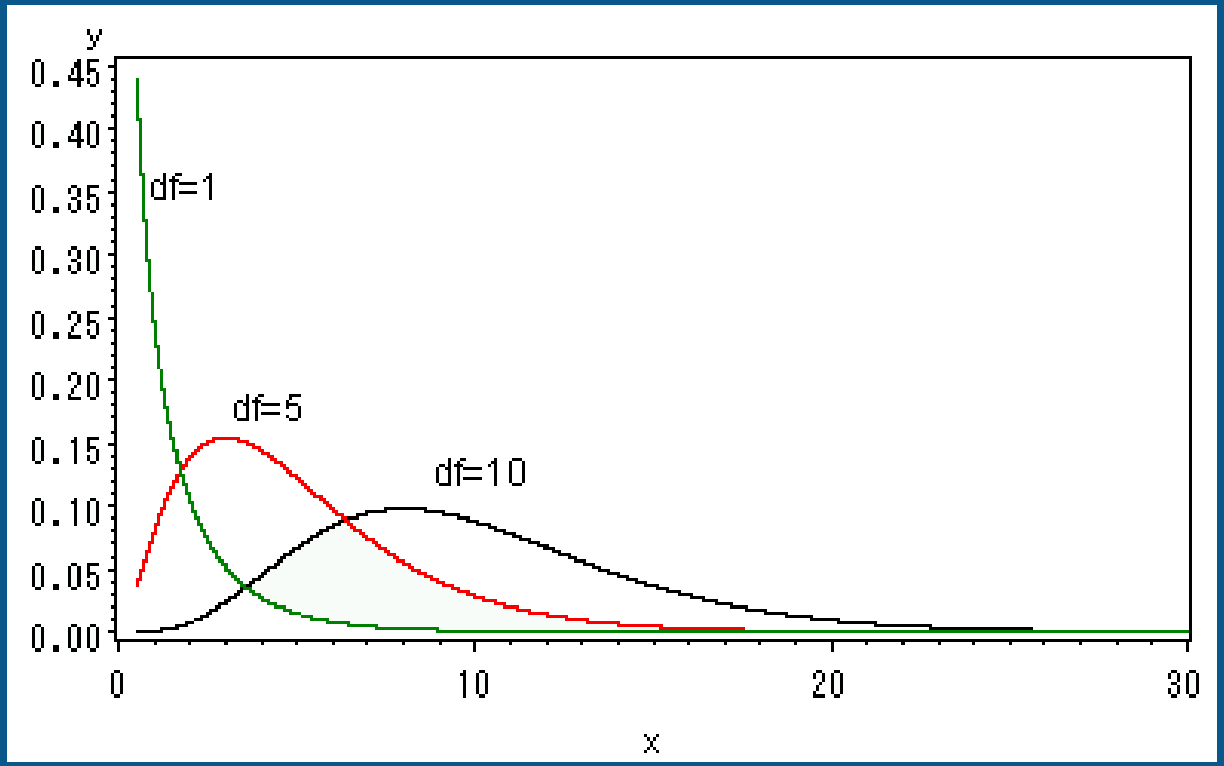
χ^2 统计量

1. 用于检验分类变量拟合优度
2. 计算公式为

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

χ^2 统计量

分布与自由度的关系



9.2 拟合优度检验

拟合优度检验
(例题分析)

【例】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，以的显著性水平检验存活状况与性别是否有关。 (α =0.05)

拟合优度检验
(例题分析)

解：要回答观察频数与期望频数是否一致，检验如下假设：

H_0：观察频数与期望频数一致

H_1：观察频数与期望频数不一致

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 $\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e} = 303$

拟合优度检验
(例题分析)

自由度的计算为df=R-1，R为分类变量类型的个数。在本例中，分类变量是性别，有男女两个类别，故R=2，于是自由度df=2-1=1，经查分布表，$\chi^2_{(0.1)}(1)=2.706$，故拒绝H_0，说明存活状况与性别显著相关

9.3 列联分析：独立性检验

9.3.1 列联表

9.3.2 独立性检验

列联表
(contingency table)

1. 由两个以上的变量交叉分类的频数分布表
2. 行变量的类别用 r 表示， r_i 表示第 i 个类别
3. 列变量的类别用 c 表示， c_j 表示第 j 个类别
4. 每种组合的观察频数用 f_{ij} 表示
5. 表中列出了行变量和列变量的所有可能的组合，所以称为列联表
6. 一个 r 行 c 列的列联表称为 $r \times c$ 列联表

列联表的结构
($r \times c$ 列联表的一般表示)

列(c_j) / 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

f_{ij} 表示第 i 行第 j 列的观察频数

独立性检验
(例题分析)

【例】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，结果如表9-3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？(α =0.05)

解：H_0：地区和原料等级之间是独立的（不存在依赖关系）

H_1：地区和原料等级之间不独立 （存在依赖关系）

$\chi 2$ 0.05(4)=9.488故拒绝*H0*，接受*H1*，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

独立性检验
(例题分析)

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

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

独立性检验
(例题分析)

卡方检验

	值	df	渐进 Sig. (双侧)
Pearson 卡方	19.822[a]	4	.001
似然比	20.732	4	.000
有效案例中的 N	500		

a. 0 单元格(.0%)的期望计数少于 5。最小期望计数为 42.00。

9.4 列联表中的相关测量

9.4.1 φ 相关系数

9.4.2 列联相关系数

9.4.3 V 相关系数

作者：贾俊平，中国人民大学统计学院

列联表中的相关测量

1. 品质相关
   - 对品质数据(分类和顺序数据)之间相关程度的测度
2. 列联表变量的相关属于品质相关
3. 列联表相关测量的统计量主要有
   - φ 相关系数
   - 列联相关系数
   - V 相关系数

φ 相关系数 (correlation coefficient)

1. 测度2×2列联表中数据相关程度
2. 对于2×2 列联表，φ系数的值在0～1之间
3. φ相关系数计算公式为

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

式中：$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(f_{ij} - e_{ij})^2}{e_{ij}}$

n为实际频数的总个数，即样本容量

φ 相关系数
(原理分析)

一个简化的 2×2 列联表

因素 Y	因素 X		合计
	x_1	x_2	
y_1	a	b	$a+b$
y_2	c	d	$c+d$
合计	$a+c$	$b+d$	n

φ 相关系数 (原理分析)

- 列联表中每个单元格的期望频数分别为

$$e_{11}=\frac{(a+b)(a+c)}{n} \qquad e_{21}=\frac{(a+c)(c+d)}{n}$$

$$e_{12}=\frac{(a+b)(b+d)}{n} \qquad e_{22}=\frac{(b+d)(c+d)}{n}$$

- 将各期望频数代入 χ^2 的计算公式得

$$\chi^2=\frac{(a-e_{11})^2}{e_{11}}+\frac{(b-e_{12})^2}{e_{12}}+\frac{(c-e_{21})^2}{e_{21}}+\frac{(d-e_{22})^2}{e_{22}}$$

$$=\frac{n(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)}$$

φ 相关系数 (原理分析)

- 将χ^2入φ相关系数的计算公式得

$$\varphi = \sqrt{\frac{\chi^2}{n}} = \frac{ad - bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}$$

  - ad 等于 bc， $\varphi = 0$，表明变量X 与 Y 之间独立
  - 若 b=0 ，c=0，或a=0 ，d=0，意味着各观察频数全部落在对角线上，此时$|\varphi|$ =1, 表明变量X 与 Y 之间完全相关
- 列联表中变量的位置可以互换，φ的符号没有实际意义，故取绝对值即可

列联相关系数 (coefficient of contingency)

1. 用于测度大于2×2列联表中数据的相关程度
2. 计算公式为

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

- C 的取值范围是 $0 \le C < 1$
- $C = 0$表明列联表中的两个变量独立
- C 的数值大小取决于列联表的行数和列数，并随行数和列数的增大而增大
- 根据不同行和列的列联表计算的列联系数不便于比较

V 相关系数
(V correlation coefficient)

1. 计算公式为

$$V=\sqrt{\frac{\chi^2}{n\min\left[(r-1),(c-1)\right]}}$$

式中：$\min\left[(r-1),(c-1)\right]$表示取$(r-1),(c-1)$中较小的一个

2. *V* 的取值范围是 0≤*V*≤1
3. *V* = 0表明列联表中的两个变量独立
4. *V*=1表明列联表中的两个变量完全相关
5. 不同行和列的列联表计算的列联系数不便于比较
6. 当列联表中有一维为2，min[(*r*-1),(*c*-1)]=1,此时*V*=φ

φ、C、V的比较

1. 同一个列联表，φ、C、V的结果会不同
2. 不同的列联表，φ、C、V的结果也不同
3. 在对不同列联表变量之间的相关程度进行比较时，不同列联表中的行与行、列与列的个数要相同，并且采用同一种系数

列联表中的相关测量
(例题分析)

【例】一种原料来自三个不同地区，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。分别计算φ系数、C系数和V系数，并分析相关程度

地区	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

列联表中的相关测量
(例题分析)

解：已知n=500，χ^2＝19.82，列联表为3×3

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

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

$$V = \sqrt{\frac{\chi^2}{n \min\left[(r-1),(c-1)\right]}} = \sqrt{\frac{19.82}{500 \times 2}} = 0.141$$

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

列联表中的相关测量
(例题分析)

对称度量

		值	近似值 Sig.
按标量标定	φ	.199	.001
	Cramer的 V	.141	.001
	相依系数	.195	.001
有效案例中的 N		500	

本章小结

1. 拟合优度检验
2. 独立性检验
3. 测度列联表中的相关性

结　束