

# 第九章 分类数据分析

李德山

四川师范大学商学院

2022 年 4 月 15 日

# Contents

- ① 分类数据与  $\chi^2$  统计量
- ② 拟合优度检验
- ③ 列联分析：独立性检验
- ④ 列联表中的相关测量

## ① 分类数据与 $\chi^2$ 统计量

## ② 拟合优度检验

## ③ 列联分析：独立性检验

## ④ 列联表中的相关测量

# 分类数据

- 分类变量的结果表现为类别。例如：性别（男，女）
- 各类别用符号或数字代码来测度
- 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
- 对分类数据的描述和分析通常使用列联表
- 可使用  $\chi^2$  检验

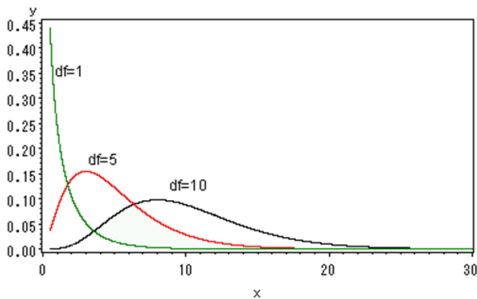
# $\chi^2$ 统计量

- 用于检验分类变量拟合优度
- 计算公式为

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

# $\chi^2$ 统计量

- 分布与自由度的关系



① 分类数据与  $\chi^2$  统计量

② 拟合优度检验

③ 列联分析：独立性检验

④ 列联表中的相关测量

# 拟合优度检验

- 【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，幸存者共 718 人，其中男性 374 人，女性 344 人，检验存活状况与性别是否有关。（ $\alpha=0.1$ ）



# 拟合优度检验

- 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

$\chi^2$  计算表

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

# 拟合优度检验

- 自由度的计算为  $df = R - 1$ ,  $R$  为分类变量类型的个数。经查分布表,
- 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  $df = 2 - 1 = 1$
- $\chi^2_{(0.1)}(1) = 2.706$ , 故拒绝  $H_0$ , 说明存活状况与性别显著相关

- ① 分类数据与  $\chi^2$  统计量
- ② 拟合优度检验
- ③ 列联分析：独立性检验
- ④ 列联表中的相关测量

## 列联表 contingency table

- 由两个以上的变量交叉分类的频数分布表
- 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
- 每种组合的观察频数用  $f_{ij}$  表示
- 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
- 一个  $r$  行  $c$  列的列联表称为  $r \times c$  列联表

# 列联表

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	$c_1$	$c_2$	...	$n$

# 独立性检验

- 【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？( $\alpha=0.05$ )

# 独立性检验

- 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  $H_0$ ，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

3×3 列联表期望值及  $\chi^2$  计算结果

行	列	$f_{..}$	$f_{.j}$	$f_{..} - f_{.j}$	$(f_{..} - f_{.j})^2$	$(f_{..} - f_{.j})^2 / f_{..}$
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

# 独立性检验

卡方检验

	值	df	渐进 Sig. (双侧)
Pearson 卡方	19.822 <sup>a</sup>	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。



## SPSS

- 【例】美国的 General Social Survey

		幸福状况			合计
		非常幸福	比较幸福	不太幸福	
婚姻状况	已婚	574	726	82	1382
	丧偶	70	149	59	278
	离异	83	292	79	454
	分居	14	73	30	117
	未婚	136	419	99	654
合计		877	1659	349	2885

## SPSS

- 二维列联表中的两个变量是否相互独立：卡方检验
- $H_0$ ：婚姻状况和幸福状况这两个变量相互独立；
- $H_1$ ：婚姻状况和幸福状况不相互独立。
- 与单个变量的  $\chi^2$  检验类似

## SPSS

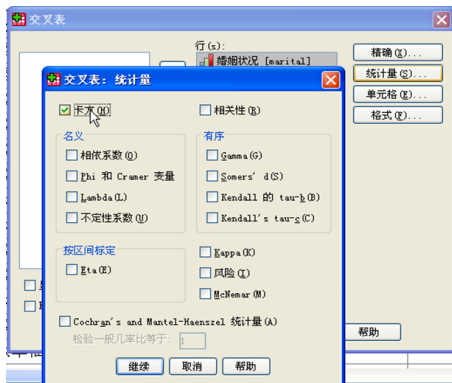
- 【例】美国的 General Social Survey

		幸福状况			合计
		非常幸福	比较幸福	不太幸福	
婚姻状况	已婚	574	726	82	1382
	丧偶	70	149	59	278
	离异	83	292	79	454
	分居	14	73	30	117
	未婚	136	419	99	654
合计		877	1659	349	2885

# SPSS

- “分析” → “描述统计” → “交叉表”，把“婚姻状况”设为行变量，把“幸福状况”设为列变量。接下来单击“统计量”，在弹出的对话框中选中“卡方”，单击“继续”；选择“单元格”，选中弹出对话框中的“期望值”，单击“继续”返回前一个是对话框，单击“确定”

## SPSS



## SPSS

			幸福状况			
			非常幸福	比较幸福	不太幸福	合计
婚姻状况	已婚	计数	574	726	82	1382
		期望的计数	420.1	794.7	167.2	1382.0
	丧偶	计数	70	149	59	278
		期望的计数	84.5	159.9	33.6	278.0
	离异	计数	83	292	79	454
		期望的计数	138.0	261.1	54.9	454.0
	分居	计数	14	73	30	117
		期望的计数	35.6	67.3	14.2	117.0
	未婚	计数	136	419	99	654
		期望的计数	198.8	376.1	79.1	654.0
合计		计数	877	1659	349	2885
		期望的计数	877.0	1659.0	349.0	2885.0

## SPSS

	值	df	渐进 Sig. (双侧)
<b>Pearson 卡方</b>	<b>225.274<sup>a</sup></b>	<b>8</b>	<b>.000</b>
似然比	230.166	8	.000
线性和线性组合	137.494	1	.000
有效案例中的 N	2885		
a. 0 单元格(.0%)的期望计数少于 5。最小期望计数为 14.15。			

- $\chi^2$  统计量的值为 225.274，相应的  $p$  值为 0.000。由于  $p$  值远远小于通常使用的显著性水平，因此检验的结论是拒绝原假设，不能认为婚姻状况和幸福状况相互独立。

- ① 分类数据与  $\chi^2$  统计量
- ② 拟合优度检验
- ③ 列联分析：独立性检验
- ④ 列联表中的相关测量



# 列联表中的相关测量

- 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
- 列联表变量的相关属于品质相关
- 列联表相关测量的统计量主要有
  - $\varphi$  相关系数
  - 列联相关系数
  - $\nu$  相关系数

## $\varphi$ 相关系数

- 测度  $2 \times 2$  列联表中数据相关程度
- 对于  $2 \times 2$  列联表,  $\varphi$  系数的值在  $0 \sim 1$  之间
- $\varphi$  相关系数计算公式为

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

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

- $n$  为实际频数的总个数, 即样本容量

# $\varphi$ 相关系数

- 一个简化的  $2 \times 2$  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	$a + c$	$b + d$	$n$

## $\varphi$ 相关系数

- 列联表中每个单元格的期望频数分别为

$$\begin{aligned} e_{11} &= \frac{(a+b)(a+c)}{n} & e_{21} &= \frac{(a+c)(c+d)}{n} \\ e_{12} &= \frac{(a+b)(b+d)}{n} & e_{22} &= \frac{(b+d)(c+d)}{n} \end{aligned}$$

- 将各期望频数代入  $\chi^2$  的计算公式得

$$\begin{aligned} \chi^2 &= \frac{(a-e_{11})^2}{e_{11}} + \frac{(b-e_{12})^2}{e_{12}} + \frac{(c-e_{21})^2}{e_{21}} + \frac{(d-e_{22})^2}{e_{22}} \\ &= \frac{n(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)} \end{aligned}$$

## $\varphi$ 相关系数

- 将  $\chi^2$  入  $\varphi$  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,  $\varphi$  的符号没有实际意义, 故取绝对值即可

## 列联相关系数

- 用于测度大于  $2 \times 2$  列联表中数据的相关程度
- 计算公式为

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

- $C$  的取值范围是  $0 \leq C < 1$
- $C = 0$  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

## $\nu$ 相关系数

- 计算公式为

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

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

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

## $\varphi$ 、 $C$ 、 $\nu$ 的比较

- 同一个列联表， $\varphi$ 、 $C$ 、 $\nu$  的结果会不同
- 不同的列联表， $\varphi$ 、 $C$ 、 $\nu$  的结果也不同
- 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】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  $\varphi$ 、 $C$ 、 $\nu$  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

- 解：已知  $n = 500$ ,  $\chi^2 = 19.82$ , 列联表为  $3 \times 3$

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

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

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

- 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

对称度量

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

## 列联分析中应注意的问题

- 条件百分表的方向，一般是自变量  $X$  放在列的位置，更好展现因果关系。
- 但也有例外，为了满足分析的需要，抽样扩大了因变量某项内容的样本量。
- 卡方分布的期望准则：
  - 1 如果只有两个类别，则每个单元的期望频数必须是大于等于 5，才可以用卡方检验。
  - 2 如果有两个以上类别，那么应有 80% 的单元的期望频数大于 5，才可以用卡方检验。
  - 解决办法：合并。

## 参考资料

- 贾俊平.《统计学》(第八版) [M]. 北京: 中国人民大学出版社, 2021。
- 刘杨.《统计学》[M]. 北京: 中国统计出版社, 2010。

Q&A

THANK YOU