

# 第15章 无序分类变量的统计推断

## —卡方检验

# 学习目标

- 了解卡方检验的**基本原理**
- 能检验**某个分类变量**各类别的出现概率是否等于指定概率
- 能检验**某两个分类变量**是否相互独立
- 能在**控制住某种或某几种分类因素**的情况下，检验另两个分类变量是否相互独立
- 检验**某两种方法**的结果是否一致

# 主要内容

- 15.1 卡方检验概述
- 15.2 单样本案例
- 15.3 两样本案例
- 15.4 卡方检验的事后两两比较
- 15.5 确切概率法和蒙特卡洛法
- 15.6 两分类变量间关联程度的度量
- 15.7 一致性检验与配对卡方检验
- 15.8 分层卡方检验



南京大學  
NANJING UNIVERSITY

## 15.6 两分类变量间关联程度的度量

- 在案例“检验不同收入级别家庭的轿车拥有率是否相同”中，检验结果如下：

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	71.134 <sup>a</sup>	1	.000	.000	.000
Continuity Correction <sup>b</sup>	69.848	1	.000		
Likelihood Ratio	80.146	1	.000		
Fisher's Exact Test					
Linear-by-Linear Association	71.062	1	.000		
N of Valid Cases	989				

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 87.05.

b. Computed only for a 2x2 table

似然比卡方： $H_0$  行变量与列变量之间相互独立，此处应拒绝 $H_0$

能否对涉及的两个分类变量间的关联强度进行度量呢？

## 相对危险度(Relative Risk, RR)

$$RR = \frac{P_t(\text{实验组人群反应阳性概率})}{P_c(\text{对照组人群反应阳性概率})} = \frac{a / n_t}{c / n_c}$$

分子 $P_t$ 为实验组人群(收入 < 48000元)反映阳性(拥有汽车)的概率,  
分母 $P_c$ 为对照组人群(收入  $\geq$  48000元)反映阳性(拥有汽车)的概率,

$n_t$ 为实验组总人数,  $a$ 为实验组反映阳性人数,  
 $n_c$ 为对照组总人数,  $c$ 为对照组反映阳性人数,

# 相对危险度(Relative Risk, RR)

01. 是否拥有家用轿车 \* 家庭收入2级 Crosstabulation

			家庭收入2级		Total
			Below 48,000	Over 48,000	
01. 是否拥有家用轿车	有	Count	32	225	257
		Expected Count	87.1	169.9	257.0
		% within 家庭收入2级	9.6%	34.4%	26.0%
	没有	Count	303	429	732
		Expected Count	247.9	484.1	732.0
		% within 家庭收入2级	90.4%	65.6%	74.0%
Total	Count		335	654	989
	Expected Count		335.0	654.0	989.0
	% within 家庭收入2级		100.0%	100.0%	100.0%

$$RR = \frac{32 / 335}{225 / 654} = \frac{9.6\%}{34.4\%} = 0.278$$

# 相对危险度(Relative Risk, RR)

RR用于反映实验因素(收入<48000元)与反映阳性(拥有汽车)的关联程度,

$$RR \in [0, \infty)$$

RR=1时, 表明实验因素与反映阳性无关联,

RR < 1时, 表明实验因素导致反映阳性的发生率降低,

RR > 1时, 表明实验因素导致反映阳性的发生率增加,

$$RR = \frac{32 / 335}{225 / 654} = \frac{9.6\%}{34.4\%} = 0.278, \text{ 表明收入} < 48000 \text{元导致}$$

轿车拥有率降低, 且收入 < 48000元的轿车拥有率是收入  $\geq 48000$ 元轿车拥有率的0.278倍.



## 优势比(Odds Ratio, OR)

$$OR = \frac{a / b (\text{反应阳性人群中实验因素有无的比例})}{c / d (\text{反应阴性人群中实验因素有无的比例})} = \frac{ad}{bc}$$

$a$ 为反应阳性组实验因素阳性人数,

$b$ 为反应阳性组实验因素阴性人数,

$c$ 为反应阴性组实验因素阳性人数,

$d$ 为反应阴性组实验因素阴性人数,

# 优势比(Odds Ratio, OR)

01. 是否拥有家用轿车 \* 家庭收入2级 Crosstabulation

			家庭收入2级		Total
			Below 48,000	Over 48,000	
01. 是否拥有家用轿车	有	Count	<i>a</i> 32	<i>b</i> 225	257
		Expected Count	87.1	169.9	257.0
		% within 家庭收入2级	9.6%	34.4%	26.0%
	没有	Count	<i>c</i> 303	<i>d</i> 429	732
		Expected Count	247.9	484.1	732.0
		% within 家庭收入2级	90.4%	65.6%	74.0%
Total	Count		335	654	989
	Expected Count		335.0	654.0	989.0
	% within 家庭收入2级		100.0%	100.0%	100.0%

$$OR = \frac{a/b}{c/d} = \frac{32/225}{303/429} = \frac{32 \times 429}{225 \times 303} = \frac{32/303}{225/429} = \frac{9.6\% / 90.4\%}{34.4\% / 65.6\%} = 0.201$$

# 优势比(Odds Ratio, OR)

OR用于反映实验因素(收入<48000元)与反映阳性(拥有汽车)的关联程度,

$$OR \in [0, \infty)$$

OR=1时, 表明实验因素与反映阳性无关联,

OR < 1时, 表明实验因素导致反映阳性的发生率降低,

OR > 1时, 表明实验因素导致反映阳性的发生率增加,

$$OR = \frac{32 / 225}{303 / 429} = 0.201, \text{ 表明收入} < 48000 \text{ 元导致}$$

轿车拥有率降低.

# 案例：计算家庭收入级别和轿车拥有情况的关联程度

- 数据文件：CCSS\_Sample.sav
- 实现过程
  - “分析” → “描述统计” → “交叉表格”
  - “行” 列表框：选入家庭收入级别Ts9(实验因素)
  - “列” 列表框：选入是否拥有家庭轿车O1(反应阳性或阴性)
  - 点击“统计”按钮：选中“风险”复选框

# 案例：计算家庭收入级别和轿车拥有情况的关联程度

交叉表格

×

行(O):

家庭收入 2级 [Ts9]

列(C):

O1. 是否拥有家用轿...

层1的1

上一页(V) 下一页(N)

☐ 在表层中显示层变量(L)

☐ 显示集群条形图(B)

☐ 取消表格(T)

确定 粘贴(P) 重置(R) 取消 帮助

精确(X)... Statistics... 单元格(E)... 格式(F)... 样式(L)... Bootstrap...

交叉表格: 统计

×

☐ 卡方(H) ☐ 相关性(R)

名义

☐ 相依系数(O) ☐ Phi 和 Cramer V ☐ Lambda ☐ 不确定性系数(U)

有序

☐ 伽玛(G) ☐ Somers' d ☐ Kendall's tau-b ☐ Kendall's tau-c

按区间标定

☐ Eta ☒ 风险(I) ☐ Kappa ☐ McNemar

☐ Cochran's and Mantel-Haenszel 统计

检验一般几率比等于(T): 1

继续 取消 帮助

# 案例：计算家庭收入级别和轿车拥有情况的关联程度

实验组

Risk Estimate

	Value	95% Confidence Interval	
		Lower	Upper
Odds Ratio for 家庭收入2级 (Below 48,000 / Over 48,000)	.201	.135	.300
For cohort O1. 是否拥有家用轿车 = 有	.278	.196	.392
For cohort O1. 是否拥有家用轿车 = 没有	1.379	1.291	1.472
N of Valid Cases	989		

OR=0.201, 95%的置信区间内也没有1, 说明收入级别与是否拥有轿车有显著关联, 且中低收入家庭拥有轿车的优势比为0.201。

RR=0.278, 此时阳性反应为“有轿车”, 95%的置信区间内也没有1, 说明收入级别与是否拥有轿车有显著关联, 且低收入家庭轿车拥有率是中高收入家庭轿车拥有率的0.278倍。

RR=1.379, 此时阳性反应为“没有轿车”, 95%的置信区间内也没有1, 说明收入级别与是否拥有轿车有显著关联, 且低收入家庭轿车不拥有率是中高收入家庭轿车不拥有率的1.379倍。



南京大學  
NANJING UNIVERSITY

## 15.7 一致性检验与配对卡方检验

# Kappa一致性检验

- 问题1

- 应用两种诊断方法对若干个对象进行疾病轻度、中度、重度的诊断，这些数据可看成是配对数据，希望检验两种诊断方法的结果是否一致

- 问题2

- 两个不同的裁判员对若干个事物进行差、中、好级别的评价，这些数据可看成是配对数据，希望检验两个裁判员的裁判结果是否一致



# Kappa一致性检验

- 一致性与相关性的区别
  - 相关不一定一致
  - 例如，诊断方法1分别诊断为轻度、中度、重度疾病的患者，诊断方法2一律分别诊断为中度、重度、轻度患者。则两种方法的诊断结果不一致，但的确存在关联。

# Kappa一致性检验

- 案例

- 数据文件: site.sav

- 要求: 检验两位顾问的评价结果是否一致

count	cons1	cons2
6.00	差	差
5.00	中	差
1.00	好	差
2.00	中	中
2.00	中	好
4.00	好	好


该数据是频数数据，应  
首先将频数设置为权重



# Kappa一致性检验


- 实现过程

- “数据” → “加权个案”
- “分析” → “统计描述” → “交叉表格”
- “行” 列表框: cons1
- “列” 列表框: cons2
- 点击 “统计量” 按钮: 选择 “Kappa” 复选框

# Kappa一致性检验

 加权个案 ×

 顾问一的评价 [cons1]  
 顾问二的评价 [cons2]

☐ 请勿对个案加权(D)  
☒ 加权个案(W)  
频率变量(F):  
 频数 [count]

当前状态:加权个案count

确定

粘贴(P)

重置(R)

取消

帮助

# Kappa一致性检验

交叉表格

频数 [count]

行(O):  
顾问一的评价 [cons1]

列(C):  
顾问二的评价 [cons2]

层1的1  
上一页(V) 下一页(N)

☐ 在表层中显示层变量(L)

☐ 显示集群条形图(B)  
☐ 取消表格(T)

确定 粘贴(P) 重置(R) 取消 帮助

精确(X)...  
Statistics...  
单元格(E)...  
格式(F)...  
样式(L)...  
Bootstrap...

交叉表格: 统计

☐ 卡方(H)  
名义——  
☐ 相依系数(O)  
☐ Phi 和 Cramer V  
☐ Lambda  
☐ 不确定性系数(U)

☐ 相关性(R)  
有序——  
☐ 伽玛(G)  
☐ Somers' d  
☐ Kendall's tau-b  
☐ Kendall's tau-c

☒ Kappa  
☐ Eta  
☐ 风险(I)  
☐ McNemar

☐ Cochran's and Mantel-Haenszel 统计  
检验一般几率比等于(T): 1

继续 取消 帮助

# Kappa一致性检验

顾问一的评价 \* 顾问二的评价 Crosstabulation

Count

		顾问二的评价			Total
		差	中	好	
顾问一的评价	差	6	0	0	6
	中	5	2	2	9
	好	1	0	4	5
Total		12	2	6	20

# Kappa一致性检验

Symmetric Measures

		Value	Asymp. Std. Error <sup>a</sup>	Approx. T <sup>b</sup>	Approx. Sig.
Measure of Agreement	Kappa	.429	.131	3.333	.001
N of Valid Cases		20			

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

**Kappa  $\geq 0.75$ : 一致性较好**

**Kappa  $\in [0.4, 0.75)$ : 一致性一般**

**Kappa  $< 0.4$ : 一致性较差**

拒绝 $H_0$ ，此处 $H_0$ :两位顾问的评价结果不一致

# 配对卡方检验

- Kappa检验：检验两种结果是否一致，会利用列联表中的所有数据
- 配对卡方检验：检验两种结果是否有差异，仅利用非主对角线上的数据

顾问一的评价 \* 顾问二的评价 Crosstabulation

Count		顾问二的评价			Total
		差	中	好	
顾问一的评价	差	6	0	0	6
	中	5	2	2	9
	好	1	0	4	5
Total		12	2	6	20



# 配对卡方检验

- 案例

数据文件: site.sav

要求: 检验两位顾问的评价结果是否有差异

count	cons1	cons2
6.00	差	差
5.00	中	差
1.00	好	差
2.00	中	中
2.00	中	好
4.00	好	好

该数据是频数数据，应  
首先将频数设置为权重

# 配对卡方检验

- 实现过程

- “数据” → “加权个案”
- “分析” → “统计描述” → “交叉表格”
- “行” 列表框： cons1
- “列” 列表框： cons2
- 点击 “统计量” 按钮： 选择 “麦克尼玛尔” 复选框

# 配对卡方检验

交叉表格

频数 [count]

行(O):

顾问一的评价 [cons1]

列(C):

顾问二的评价 [cons2]

层1的1

上一页(V)

下一页(N)

☐ 在表层中显示层变量(L)

☐ 显示集群条形图(B)

☐ 取消表格(T)

确定

粘贴(P)

重置(R)

取消

帮助

精确(X)...

Statistics...

单元格(E)...

格式(F)...

样式(L)...

Bootstrap...

交叉表格: 统计

☐ 卡方(H)

☐ 相关性(R)

名义

☐ 相依系数(O)

☐ Phi 和 Cramer V

☐ Lambda

☐ 不确定性系数(U)

有序

☐ 伽玛(G)

☐ Somers' d

☐ Kendall's tau-b

☐ Kendall's tau-c

按区间标定

☐ Eta

☐ Kappa

☐ 风险(I)

☒ McNemar

☐ Cochran's and Mantel-Haenszel 统计

检验一般几率比等于(T): 1

继续

取消

帮助

# 配对卡方检验

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
McNemar-Bowker Test	8.000	3	.046
N of Valid Cases	20		

拒绝 $H_0$ ，此处 $H_0$ :两位顾问  
的评价结果无差异

# 配对卡方检验

- 说明：
  - 在应用中，对于一致性较好，即绝大多数数据都在主对角线上的列联表，麦克尼玛尔检验可能会失去实用价值



南京大學  
NANJING UNIVERSITY

## 15.8 分层卡方检验

- 之前已经检验发现家庭收入级别的确会影响家庭轿车的拥有情况（从总体上），但可能还有别的分类变量对轿车拥有情况产生影响，例如“城市”（构成一个混杂因素）
  - 不同城市的轿车拥有情况存在差异，那么收入级别对轿车拥有的影响在不同城市间是否也存在差异？
  - 如果收入级别对轿车拥有的影响在不同城市间存在差异，此时不应当考虑将不同城市的数据结合起来得到一个总的分析结果
  - 如果收入级别对轿车拥有的影响在不同城市间没有差异，此时可以考虑将不同城市的数据结合起来得到一个总的分析结果，但应在控制城市的混杂作用后计算校正后的RR或OR

# 案例

- 数据文件：CCSS\_Sample.sav
- 要求：在控制城市影响的前提下得到更准确的家庭收入级别和轿车拥有情况的关联程度测量指标
- 实现过程
  - “分析” → “统计描述” → “交叉表格”
  - “行”列表框：家庭收入级别Ts9
  - “列”列表框：是否拥有轿车O1
  - “层”列表框：城市S0
  - 点击“统计量”按钮：选择“风险”复选框、“柯克兰和曼特尔-亨塞尔统计”复选框



# 案例

交叉表格

月份 [time]

ID [id]

S2. 性别 [s2]

S3. 年龄 [s3]

S4. 学历 [s4]

S5. 职业 [s5]

S7. 婚姻状况 [s7]

S9. 家庭月收入 [s9]

C0. 请问您的家庭...

C0. 请问您的家庭...

C0. 请问您的家庭...

A3. 首先, 请问与...

A3a. 您为什么这...

A3a. 您为什么这...

行(O):

家庭收入2级 [Ts9]

列(C):

O1. 是否拥有家用轿...

层1的1

上一页(V)

下一页(N)

S0. 城市 [s0]

☐ 在表层中显示层变量(L)

☐ 显示集群条形图(B)

☐ 取消表格(T)

确定

粘贴(P)

重置(R)

取消

帮助

交叉表格: 统计

☐ 卡方(H)

☐ 相关性(R)

名义

☐ 相依系数(O)

☐ Phi 和 Cramer V

☐ Lambda

☐ 不确定性系数(U)

有序

☐ 伽玛(G)

☐ Somers' d

☐ Kendall's tau-b

☐ Kendall's tau-c

按区间标定

☐ Kappa

☒ 风险(I)

☐ McNemar

☒ Cochran's and Mantel-Haenszel 统计

检验一般几率比等于(T): 1

继续

取消

帮助

/私...	已婚	6000-7999元	2 无
职员...	已婚	99拒绝回答	2 无
职员...	已婚	5000-5999元	2 无

# 案例

Risk Estimate

S0. 城市		Value	95% Confidence Interval	
			Lower	Upper
100北京	Odds Ratio for 家庭收入2级 (Below 48,000 / Over 48,000)	.156	.075	.326
	For cohort O1. 是否拥有家用轿车 = 有	.231	.121	.440
	For cohort O1. 是否拥有家用轿车 = 没有	1.477	1.308	1.666
	N of Valid Cases	319		
200上海	Odds Ratio for 家庭收入2级 (Below 48,000 / Over 48,000)	.089	.031	.251
	For cohort O1. 是否拥有家用轿车 = 有	.123	.046	.328
	For cohort O1. 是否拥有家用轿车 = 没有	1.384	1.261	1.519
	N of Valid Cases	337		
300广州	Odds Ratio for 家庭收入2级 (Below 48,000 / Over 48,000)	.333	.189	.586
	For cohort O1. 是否拥有家用轿车 = 有	.434	.275	.683
	For cohort O1. 是否拥有家用轿车 = 没有	1.302	1.151	1.474
	N of Valid Cases	333		
Total	Odds Ratio for 家庭收入2级 (Below 48,000 / Over 48,000)	.201	.135	.300
	For cohort O1. 是否拥有家用轿车 = 有	.278	.196	.392
	For cohort O1. 是否拥有家用轿车 = 没有	1.379	1.291	1.472
	N of Valid Cases	989		

# 案例

## 层间OR值差异性的检验

Tests of Homogeneity of the Odds Ratio

	Chi-Squared	df	Asymp. Sig. (2-sided)
Breslow-Day	6.165	2	.046
Tarone's	6.161	2	.046

拒绝 $H_0$ ，此处 $H_0$ ：行列间的联系强度在3个城市间相同。  
既然拒绝了 $H_0$ ，因此不再考虑将不同城市的数据结合起来得到一个总的分析结果

# 案例

假如上页ppt上的P值>0.5, OR值齐性, 可继续观察下面的分层卡方检验结果 (控制了分层因素)

Tests of Conditional Independence

	Chi-Squared	df	Asymp. Sig. (2-sided)
Cochran's	72.397	1	.000
Mantel-Haenszel	70.879	1	.000

Under the conditional independence assumption, Cochran's statistic is asymptotically distributed as a 1 df chi-squared distribution, only if the number of strata is fixed, while the Mantel-Haenszel statistic is always asymptotically distributed as a 1 df chi-squared distribution. Note that the continuity correction is removed from the Mantel-Haenszel statistic when the sum of the differences between the observed and the expected is 0.

拒绝 $H_0$ , 此处 $H_0$ : 行列变量相互独立,  
说明行列变量间有关联

# 案例

Mantel-Haenszel Common Odds Ratio Estimate

Estimate			.195
ln(Estimate)			-1.636
Std. Error of ln(Estimate)			.206
Asymp. Sig. (2-sided)			.000
Asymp. 95% Confidence Interval	Common Odds Ratio	Lower Bound	.130
		Upper Bound	.292
	ln(Common Odds Ratio)	Lower Bound	-2.040
		Upper Bound	-1.232

The Mantel-Haenszel common odds ratio estimate is asymptotically normally distributed under the common odds ratio of 1.000 assumption. So is the natural log of the estimate.

校正后的OR值，即去除了不同分层的混杂效应后，与中高收入家庭相比，中低收入家庭拥有轿车的优势比为**0.195**，或者说概率是前者的大约**1/5**

**THE END**