南京大学
NANJING UNIVERSITY

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

15.1 卡方检验概述

卡方检验的基本原理

- **卡方分布**

设随机变量 $X_1, X_2, \cdots, X_n$ 相互独立,

且都服从标准正态分布$N(0, 1)$, 则

$$\sum_{i=1}^{n} X_i^2 \sim \chi^2(n)$$

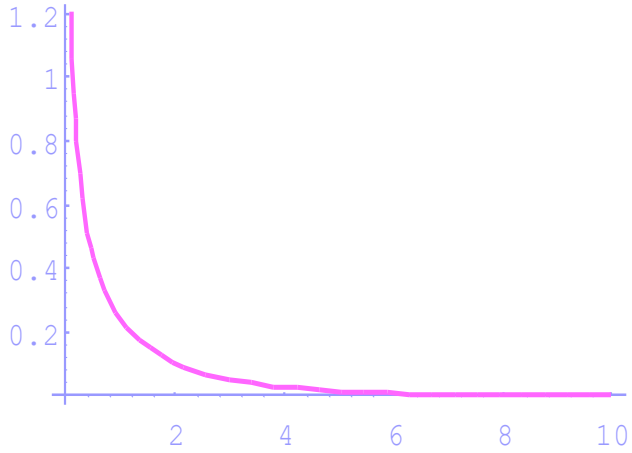
称 $\sum_{i=1}^{n} X_i^2$ 服从自由度为 n 的卡方分布。

卡方检验的基本原理

- 卡方分布

$n=1$ 时,其密度函数为

$$f(x)=\begin{cases}\dfrac{1}{\sqrt{2\pi}}x^{-\frac{1}{2}}e^{-\frac{x}{2}}, & x>0\\ 0, & x\le 0\end{cases}$$



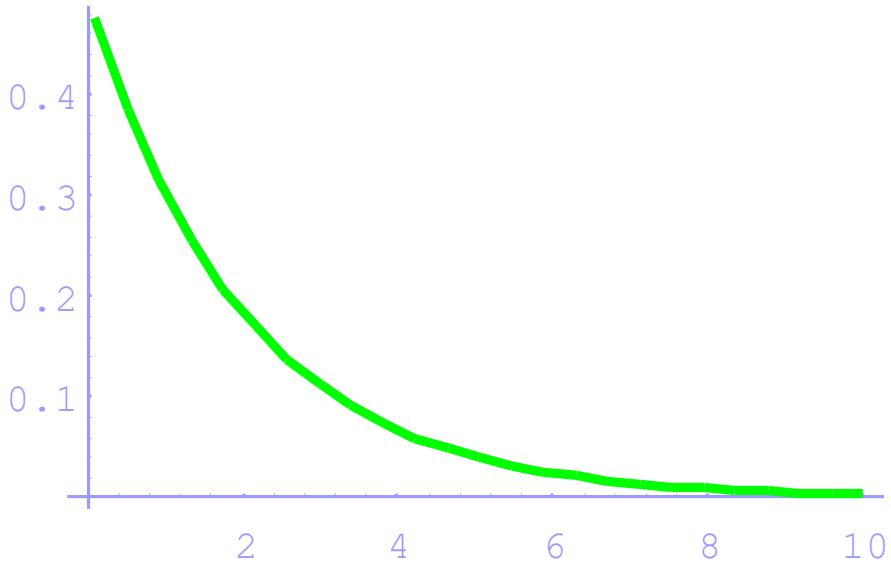
卡方检验的基本原理

- 卡方分布

$n = 2$ 时,其密度函数为

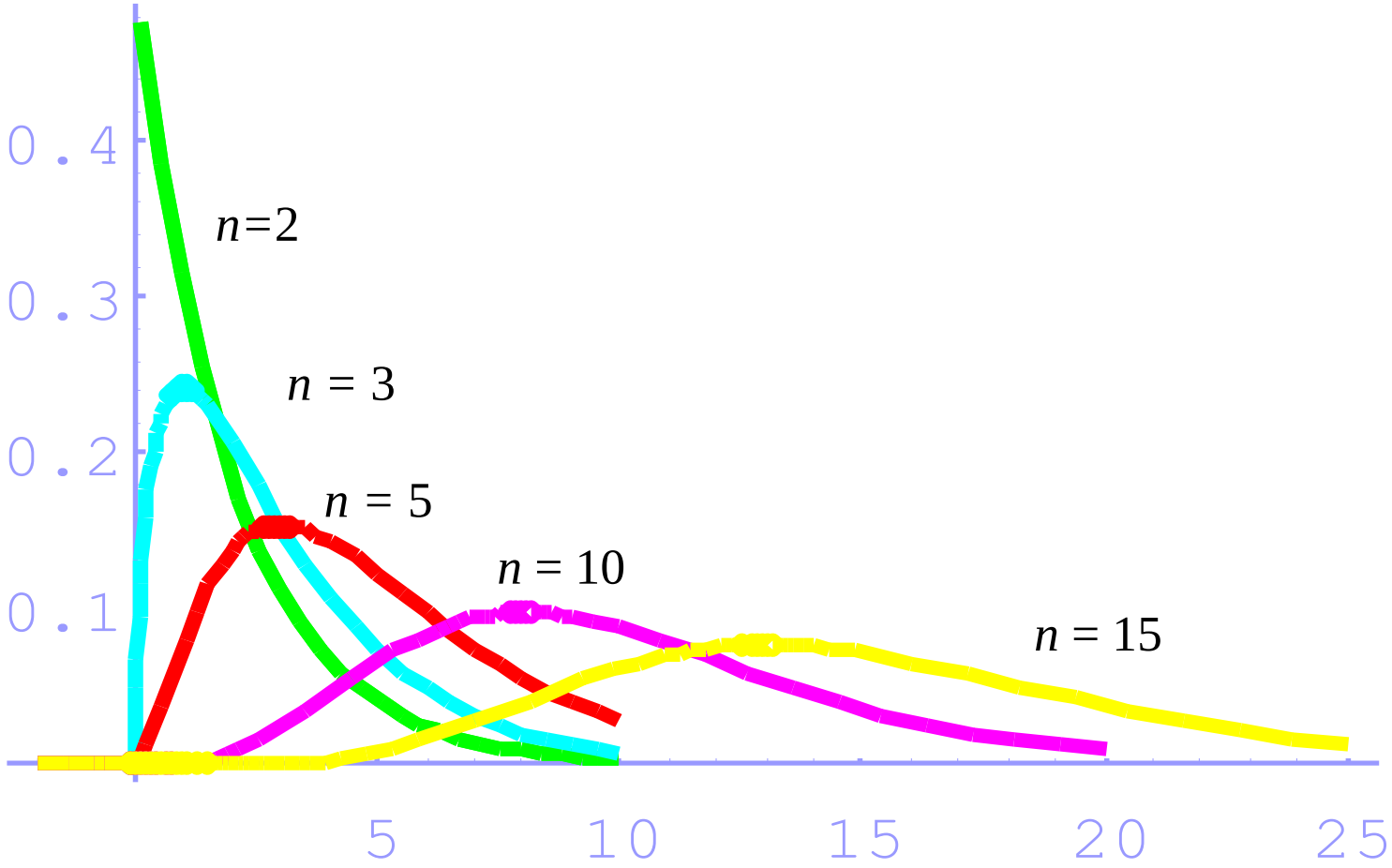
$$f(x) = \begin{cases} \frac{1}{2} e^{-\frac{x}{2}}, & x > 0 \\ 0, & x \le 0 \end{cases}$$

为参数为 2 的指数分布.



卡方检验的基本原理

- 卡方分布



卡方检验的基本原理

- 卡方检验的基本思想
  - 针对单个样本，考察各区间(连续变量)或各类别（无序分类变量）的观察频数与期望频数有无较大差别（例如，考察200704的性别分布是否均衡）
  - 针对两个或多个样本，用于区别样本来源的分类变量和另一个分类变量交叉形成单元格，考察单元格内的观察频数与期望频数有无较大差别（例如，考察4个不同时间点的性别分布是否相同）

卡方检验的基本原理

- **卡方检验的步骤**

1. 根据具体问题写出H_0：样本来自的总体的某一分类变量各类别概率分别为p_1，p_2，$\cdots$，p_k

2. 统计k个类别上的观察频数A_i，$i=1,2,\cdots,k$

3. 在H_0真条件下，计算n个样本落在各类别上的期望频数np_i，$i=1,2,\cdots,k$

卡方检验的基本原理

- **卡方检验的步骤**

4. 计算期望频数与观察频数的偏差和：

$$\chi^2=\sum_{i=1}^{k}\frac{(A_i-np_i)^2}{np_i}=\sum_{i=1}^{k}\frac{A_i^2}{np_i}-n$$

当H_0真，n充分大时$\chi^2\sim\chi^2(k-1-r)$，
r为计算np_i时用到的估计出的参数个数.

5. 计算$P\{\chi^2(k-1-r)\geq\chi^2\}$，即为$P$值，根据$P$值来判断是否拒绝$H_0$.

卡方检验的基本原理

- **卡方检验的样本量要求**

  在实际应用χ^2检验法时，要满足样本容量$n \geq 50$，期望频数$np_i \geq 5$。

卡方检验的基本原理

- SPSS中的相应功能
  - “分析”→“非参数检验”→“单样本”→“卡方检验”(检验某个分类变量各类别的出现概率是否等于指定概率)
  - “分析”→“非参数检验”→“旧对话框”→“卡方检验”(检验某个分类变量各类别的出现概率是否等于指定概率)
  - “分析”→“描述统计”→“交叉表格”→“统计”→“卡方”(检验两个分类变量是否独立，检验两种方法的结果是否一致)

15.2 单样本案例

检验的目的

- 从已知样本数据出发，来判断总体各取值水平出现的概率是否与已知概率相符
- 可看成是二项检验的推广
  - 二项检验：检验两个类别各自出现的概率是否与已知概率相符
  - 卡方检验：检验两个类别或多个类别各自出现的概率是否与已知概率相符

应用的场景

- 检验骰子各面出现的概率是否各等于1/6
- 国家人口老龄化问题是否更严重了（老年人出现的概率是否提高了）
- 某产品的市场占有份额是否较以前更大
- ……

案例

- 数据文件：CCSS_Sample.sav
- 要求：检验2007年4月份的性别分布是否均衡
- 分析：因为性别是二分类变量，可以采用二项分布检验，也可以采用卡方检验

案例

- 卡方检验步骤
  - **H_0: 男性比例为0.5(p_1), 女性比例为0.5(p_2)**

    H_1: 不满足“男性比例为0.5(p_1), 女性比例为0.5(p_2)”
  - **统计男性的观察频数A_1和女性的观察频数A_2**
  - **计算男性的期望频数np_1和女性的期望频数np_2**
  - **计算观察频数和期望频数的偏差和:**

$$\chi^2=\sum_{i=1}^{2}\frac{(A_i-np_i)^2}{np_i}=\sum_{i=1}^{2}\frac{A_i^2}{np_i}-n$$

  - **计算样本事件以及更极端事件的概率P，若$P\le\alpha$，拒绝 H_0，否则不拒绝 H_0**

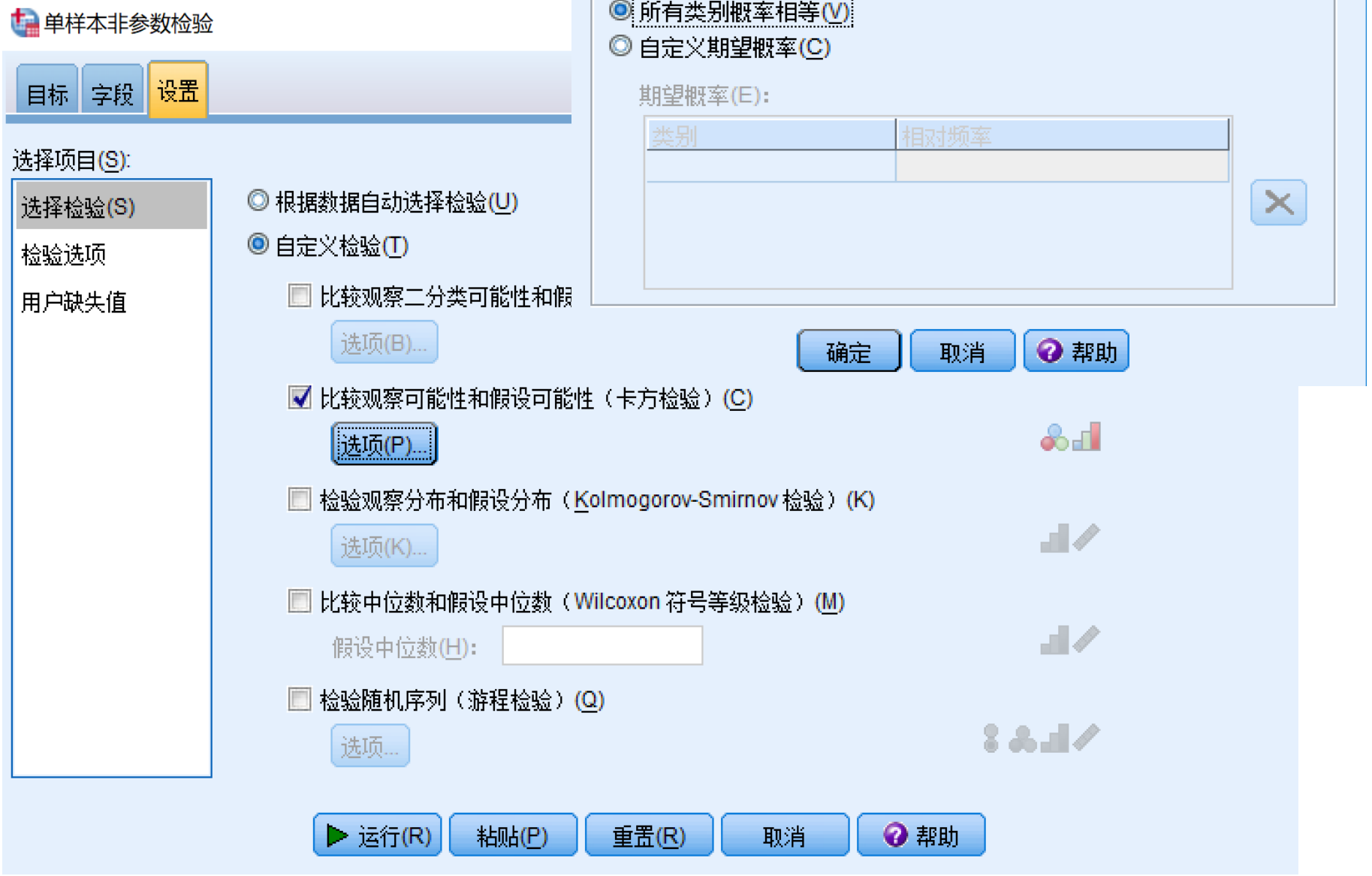
案例

- 实现过程
  - 选择个案(200704)
  - “分析” →“非参数检验” →“单样本”
  - “目标”选项卡，“自动比较观察数据和假设数据”
  - “字段”选项卡中，将“S2”选入“检验字段”框
  - “设置”选项卡中，选择“卡方检验”，再点击下方的“选项”按钮，选择“所有类别概率相等”

案例


单样本非参数检验
目标
字段
设置
使用预定义角色(U)
使用自定义字段分配(C)
字段:
排序:
无(N)
月份
ID
S0. 城市
S3. 年龄
S4. 学历
S5. 职业
S7. 婚姻状况
S9. 家庭月收入
C0. 请问您的家庭目前有下列还
C0. 请问您的家庭目前有下列还
C0. 请问您的家庭目前有下列还
O1. 是否拥有家用轿车
A3. 首先，请问与一年前相比，
全部
检验字段(T):
S2. 性别
运行(R)
粘贴(P)
重置(R)
取消
帮助

案例


单样本非参数检验
目标
字段
设置
选择项目(S):
选择检验(S)
检验选项
用户缺失值
根据数据自动选择检验(U)
自定义检验(T)
比较观察二分类可能性和假
选项(B)...
比较观察可能性和假设可能性（卡方检验）(C)
选项(P)...
检验观察分布和假设分布（Kolmogorov-Smirnov 检验）(K)
选项(K)...
比较中位数和假设中位数（Wilcoxon 符号等级检验）(M)
假设中位数(H):
检验随机序列（游程检验）(Q)
选项...
运行(R)
粘贴(P)
重置(R)
取消
帮助
卡方检验选项
选择检验选项
所有类别概率相等(V)
自定义期望概率(C)
期望概率(E):
类别
相对频率
确定
取消
帮助

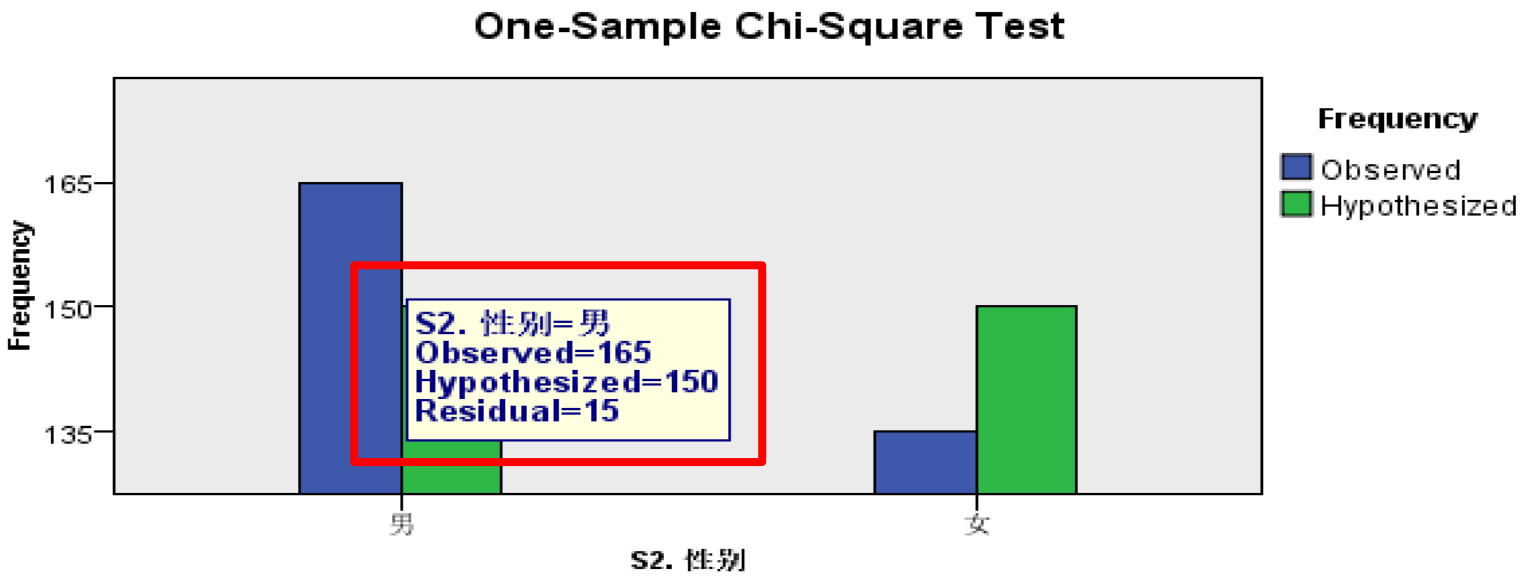
案例

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The categories of S2. 性别 occur with equal probabilities.	One-Sample Chi-Square Test	.083	Retain the null hypothesis.

Asymptotic significances are displayed. The significance level is .05.

案例



Total N	300
Test Statistic	3.000
Degrees of Freedom	1
Asymptotic Sig. (2-sided test)	.083

1. There are 0 cells (0%) with expected values less than 5. The minimum expected value is 150.

15.3 两样本案例

检验的目的

- 针对两个或多个样本，比较他们所属总体的某个分类变量各类别的发生率/构成比是否相同
- 注意:不注重发生率/构成比的具体取值，而注重是否有显著差异

应用的场景

- 不同收入级别家庭的轿车拥有率是否相同
- 不同地区家庭的轿车拥有率是否相同
- ……

案例

- 数据文件：CCSS_Sample.sav
- 要求：检验不同收入级别家庭的轿车拥有率是否相同
- 分析：有两个收入级别，涉及到了两个样本，轿车拥有率是一个构成比，因此采用卡方检验方法

案例

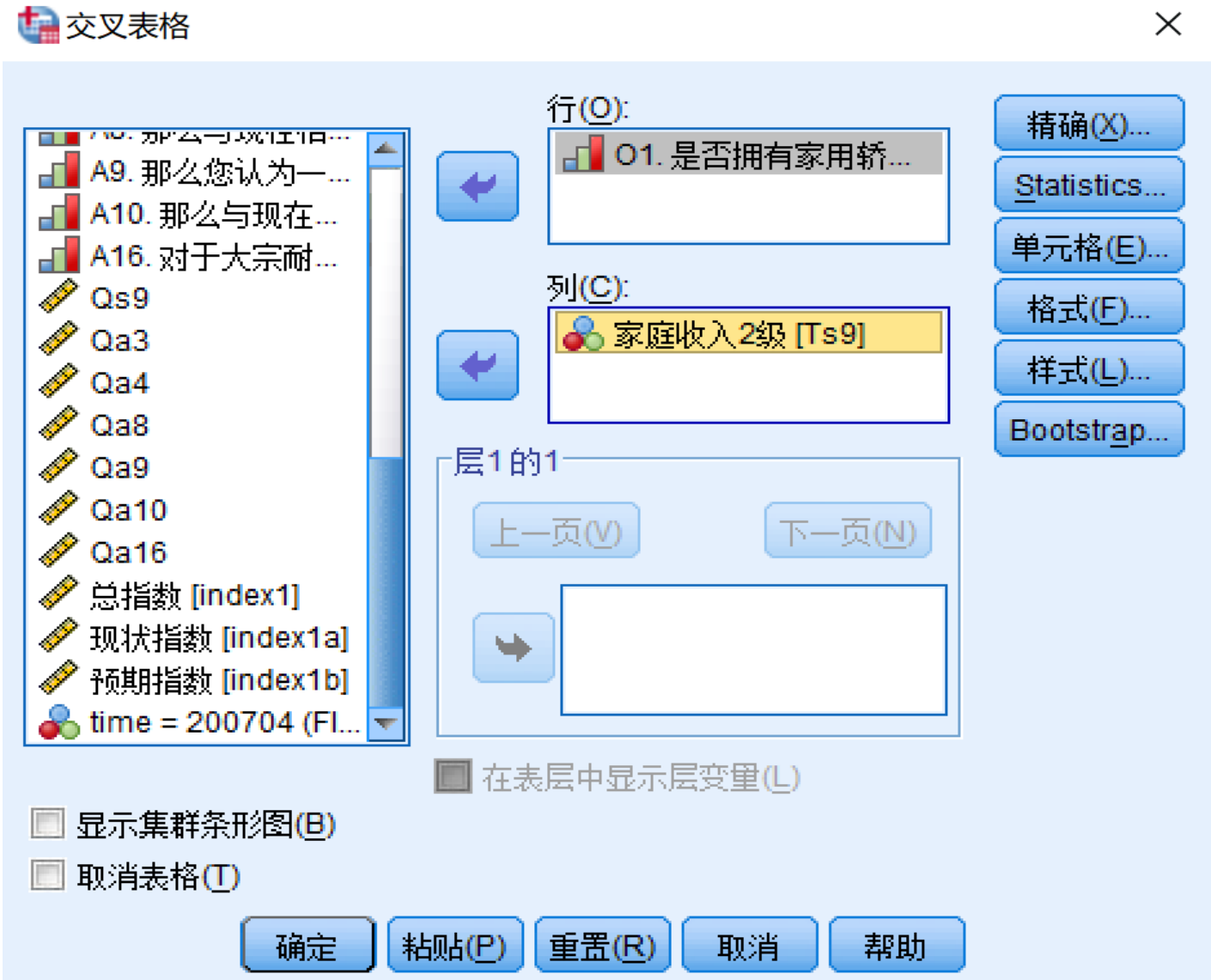
- 卡方检验步骤
  - H_0: 不同收入级别家庭的轿车拥有率相同

    H_1: 不同收入级别家庭的轿车拥有率不相同
  - 统计不同收入级别家庭拥有和不拥有轿车的观察频数
  - 计算不同收入级别家庭拥有和不拥有轿车的期望频数
  - 计算所有观察频数和期望频数的偏差和
  - 在 H_0为真条件下，偏差和较小是大概率事件，偏差和较大是小概率事件，如果 P值 $\leq \alpha$，出现小概率事件，则拒绝 H_0，否则不拒绝

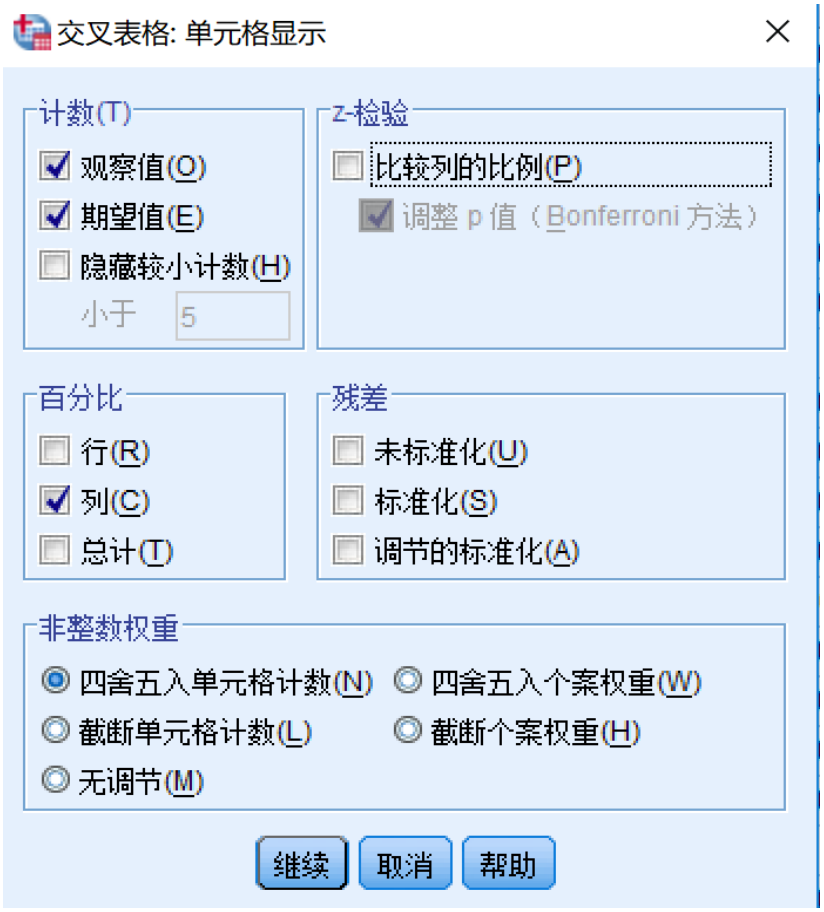
案例

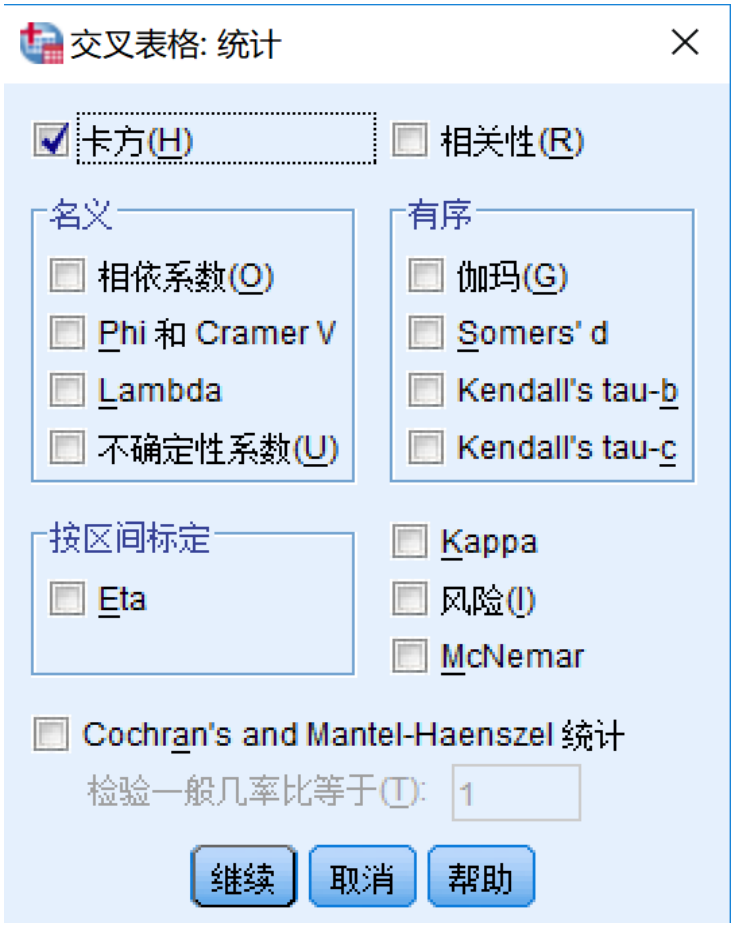
- 实现过程
  - “分析”→“描述统计”→“交叉表格”
  - 将“O1”选入“行”，将“Ts9”选入“列”
  - 点击“单元格”按钮，选中“期望值”，选中“列百分比”
  - 点击“统计”按钮，选中“卡方”复选框

案例


交叉表格
A9. 那么您认为一...
A10. 那么与现在...
A16. 对于大宗耐...
Qs9
Qa3
Qa4
Qa8
Qa9
Qa10
Qa16
总指数 [index1]
现状指数 [index1a]
预期指数 [index1b]
time = 200704 (Fl...
行(O):
O1. 是否拥有家用轿...
列(C):
家庭收入2级 [Ts9]
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)
精确(X)...
Statistics...
单元格(E)...
格式(F)...
样式(L)...
Bootstrap...
确定
粘贴(P)
重置(R)
取消
帮助

案例


交叉表格: 单元格显示
计数(T)
观察值(O)
期望值(E)
隐藏较小计数(H)
小于 5
z-检验
比较列的比例(P)
调整 p 值（Bonferroni 方法）
百分比
行(R)
列(C)
总计(T)
残差
未标准化(U)
标准化(S)
调节的标准化(A)
非整数权重
四舍五入单元格计数(N)
四舍五入个案权重(W)
截断单元格计数(L)
截断个案权重(H)
无调节(M)
继续
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
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

案例

O1. 是否拥有家用轿车 * 家庭收入2级 Crosstabulation

			家庭收入2级		
			Below 48,000	Over 48,000	Total
O1. 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%

案例

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	71.134[a]	1	.000		
Continuity Correction[b]	69.848	1	.000		
Likelihood Ratio	80.146	1	.000		
Fisher's Exact Test				.000	.000
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

15.4 卡方检验的事后两两比较

检验的目的

- 当对3个及3个以上的样本检验某一分类变量各类别的构成比是否相同时，假如检验的结果是构成比存在显著差异，此时若想了解究竟是哪两个样本有显著差异，可通过卡方检验的两两比较来实现。

案例

- 数据文件：**CCSS_Sample.sav**
- 要求：检验北京、上海、广州三地家庭的轿车拥有率是否相同
- 分析：有3个地区，涉及到了3个样本，轿车拥有率是一个构成比，因此采用卡方检验方法。如果检验结果是存在显著差异，可进一步检验哪两个地区有显著差异

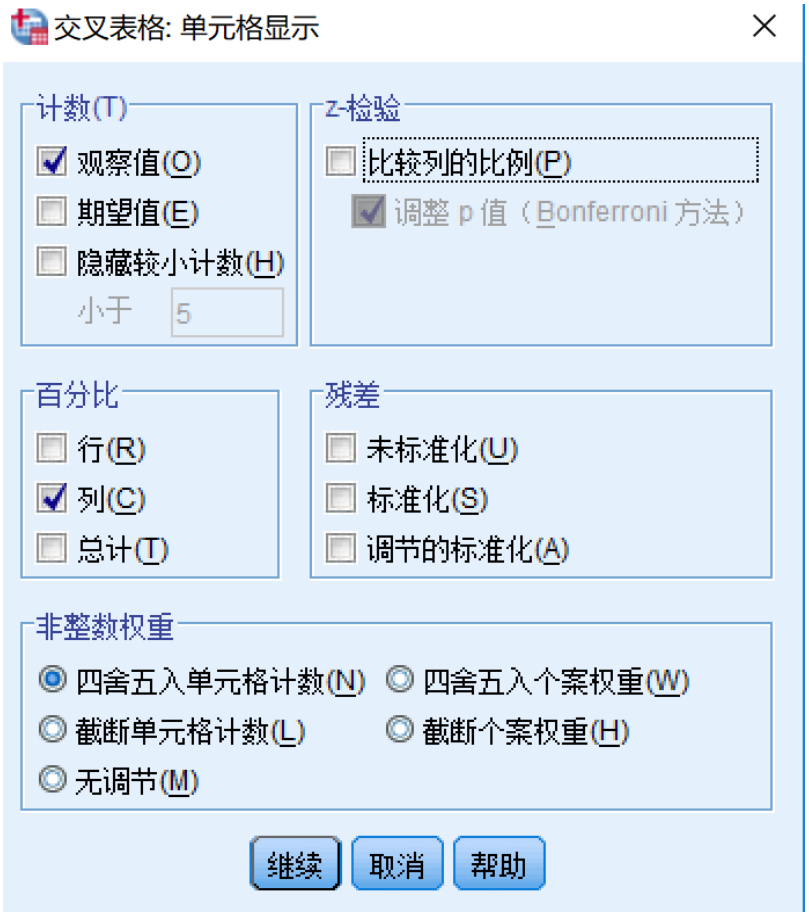
案例

- 实现过程
  - “分析”→“描述统计”→“交叉表格”
  - 将“O1”选入“行”，将“S0”选入“列”
  - 点击“单元格”按钮，选中“列百分比”
  - 点击“统计”按钮，选中“卡方”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):
O1. 是否拥有家用轿...
列(C):
S0. 城市 [s0]
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)
精确(X)...
Statistics...
单元格(E)...
格式(F)...
样式(L)...
Bootstrap...
确定
粘贴(P)
重置(R)
取消
帮助

案例


交叉表格: 单元格显示
计数(T)
观察值(O)
期望值(E)
隐藏较小计数(H)
小于 5
z-检验
比较列的比例(P)
调整 p 值（Bonferroni 方法）
百分比
行(R)
列(C)
总计(T)
残差
未标准化(U)
标准化(S)
调节的标准化(A)
非整数权重
四舍五入单元格计数(N)
四舍五入个案权重(W)
截断单元格计数(L)
截断个案权重(H)
无调节(M)
继续
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
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

案例

O1. 是否拥有家用轿车 * S0. 城市 Crosstabulation

			S0. 城市			Total
			100北京	200上海	300广州	
O1. 是否拥有家用轿车	有	Count	118	87	107	312
		% within S0. 城市	31.4%	22.5%	28.1%	27.3%
	没有	Count	258	300	274	832
		% within S0. 城市	68.6%	77.5%	71.9%	72.7%
Total		Count	376	387	381	1144
		% within S0. 城市	100.0%	100.0%	100.0%	100.0%

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	7.810[a]	2	.020
Likelihood Ratio	7.901	2	.019
Linear-by-Linear Association	1.017	1	.313
N of Valid Cases	1144		

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 102.55.

拒绝三个地区家庭的轿车拥有率相同

案例

- 进一步比较哪两个地区的轿车拥有率不同
  - “分析”→“描述统计”→“交叉表格”
  - 将“O1”选入“行”，将“S0”选入“列”
  - 点击“单元格”按钮，选中“列百分比”，选中“z-检验”下的“比较列的比例”及其下方的“调整P值”
  - 点击“统计”按钮，选中“卡方”复选框

案例


交叉表格: 单元格显示
计数(T)
观察值(O)
期望值(E)
隐藏较小计数(H)
小于
5
z-检验
比较列的比例(P)
调整 p 值（Bonferroni 方法）
百分比
行(R)
列(C)
总计(T)
残差
未标准化(U)
标准化(S)
调节的标准化(A)
非整数权重
四舍五入单元格计数(N)
四舍五入个案权重(W)
截断单元格计数(L)
截断个案权重(H)
无调节(M)
继续
取消
帮助

案例

O1. 是否拥有家用轿车 * S0. 城市 Crosstabulation

			S0. 城市			Total
			100北京	200上海	300广州	
O1. 是否拥有家用轿车	有	Count	118a	87b	107a, b	312
		% within S0. 城市	31.4%	22.5%	28.1%	27.3%
	没有	Count	258a	300b	274a, b	832
		% within S0. 城市	68.6%	77.5%	71.9%	72.7%
Total		Count	376	387	381	1144
		% within S0. 城市	100.0%	100.0%	100.0%	100.0%

Each subscript letter denotes a subset of S0. 城市 categories whose column proportions do not differ significantly from each other at the .05 level.

表中以APA下标格式标注出各组的两两比较结果：

(1) 三个地区构成了两个同质组：a和b，同质组内的城市间无显著差异，不同同质组的城市间有显著差异

(2) 同质组a：北京、广州

(3) 同质组b：上海、广州

可知，北京与上海两地轿车拥有率具有显著差异。

15.5 确切概率法和蒙特卡洛法

计算P值的方法

	Value
Pearson Chi-Square	71.134[a]
Continuity Correction[b]	69.848
Likelihood Ratio	80.146
Fisher's Exact Test	
Linear-by-Linear Association	71.062
N of Valid Cases	989

a. 0 cells (0.0%) have expected count less

b. Computed only for a 2x2 table

渐进法

Yates校正法

确切概率法

除此之外，还有蒙特卡洛法

渐进法

- 默认选项
- 只计算基于检验统计量的渐近分布的近似概率值，而不计算确切概率
- 样本容量较大且P值远离检验水准时，对于结果的判断没有影响，且可以节省计算时间

Yates校正法

- 应用条件：
  - 样本量大于40
  - 所有单元格的期望频数均大于1
  - 只有1/5以下的单元格的期望频数小于5且大于1

确切概率法

- 计算确切的P值，而非近似值
- 如果渐进法或Yates校正法计算的P值接近检验水准，则需要利用确切概率法计算确切的P值
- 该方法的计算量很大，SPSS仅对四格表默认直接输出确切P值

蒙特卡洛法

- 是介于渐进法和确切概率法的中间方法，计算出的P值比渐进法准确，同时又没有确切概率法那么大的计算量
- 实现过程

  “分析”→“描述统计”

  →“交叉表格”→“精确”

蒙特卡洛法

- 考察北京、上海、广州三地家庭的轿车拥有率是否相同，使用蒙特卡洛法计算P值。

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Monte Carlo Sig. (2-sided)			Monte Carlo Sig. (1-sided)		
				Sig.	99% Confidence Interval		Sig.	99% Confidence Interval	
					Lower Bound	Upper Bound		Lower Bound	Upper Bound
Pearson Chi-Square	7.810[a]	2	.020	.020[b]	.016	.023			
Likelihood Ratio	7.901	2	.019	.019[b]	.016	.023			
Fisher's Exact Test	7.876			.019[b]	.016	.023			
Linear-by-Linear Association	1.017[c]	1	.313	.332[b]	.319	.344	.17[b]	.161	.180
N of Valid Cases	1144								

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 102.55.

b. Based on 10000 sampled tables with starting seed 2000000.

c. The standardized statistic is 1.009.

THE END