《传播统计学》

卡方检验

教师：林志良
邮箱：linzhl@nfu.edu.cn
个人网站：www.zhilianglin.com

目录

卡方检验介绍

- **作用**：确定两个**分类型变量**的关系
  - 例如：
    - 大学生性别与学生干部的关系。
    - 星座是否适配与恋爱结局的关系。
    - 文理科与当公务员的关系。
    - 政治身份（党员/民主党派/团员/群众）与对某政策的态度（支持/反对）。

卡方检验介绍

列联表

- 为了分析两个类别型变量之间的关系，我们常常先将数据整理成列联表（contingency table）的形式。

- 列联表是将两个以上的变量进行交叉分类的频数分布表。

卡方检验介绍

列联表

例如，为了研究性别和使用电脑系统的关系，将调查样本（n=300）按性别和电脑系统类型构建列联表。

性别与电脑类型的列联表

性别	电脑类型	
	Mac	Windows
女性	12	108
男性	24	156

卡方检验介绍

观察频数

列联表中的单元格中的数值我们称之为**观察频数（observed frequency）**

性别	电脑类型	
	Mac	Windows
女性	12	108
男性	24	156

卡方检验介绍

期望频数

$$f_e = \frac{RT \times CT}{n}$$

其中，

f_e = 给定单元格的期望频数

RT = 给定单元所在的列（Row）总和（Total）

CT = 给定单元所在的行（Column）总和（Total）

n = 样本量

卡方检验介绍

观察频数 vs. 期望频数

性别	电脑类型		
	Mac	Windows	
女性	观察频数 = 12 期望频数 = 14.4	观察频数= 108 期望频数 = 105.6	120
男性	观察频数= 24 期望频数 = 21.6	观察频数= 156 期望频数 = 158.4	180
	36	264	300

CT（行总和）

RT（列总和）

n（样本量）

其中一个期望频数的计算：

$$f_e = \frac{RT \times CT}{n} = \frac{36 \times 120}{300} = 14.4$$

卡方检验介绍

期望频数的含义

期望频数的含义为假设两分类变量没有关系，列联表中的频数的理论分布状况（理论频数）。

例如，上例中我们的样本量为300。其中，使用Windows的人数为264，占比88%；使用Mac的人数为36，占比12%. 样本中有女生120人，男生180人。如果我们知道性别和使用电脑系统没有关系……

- **使用Windows和Mac系统的女生的人数应该分别是多少?**

  使用Windows的女生人数：120*88%=105.6；使用Mac的女生人数： 120*12%=14.4

- **使用Windows和Mac系统的男生的人数应该分别是多少?**

  使用Windows的男生人数：264*88%=158.4；使用Mac的男生人数： 264*12%=21.6

卡方检验介绍

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

其中，

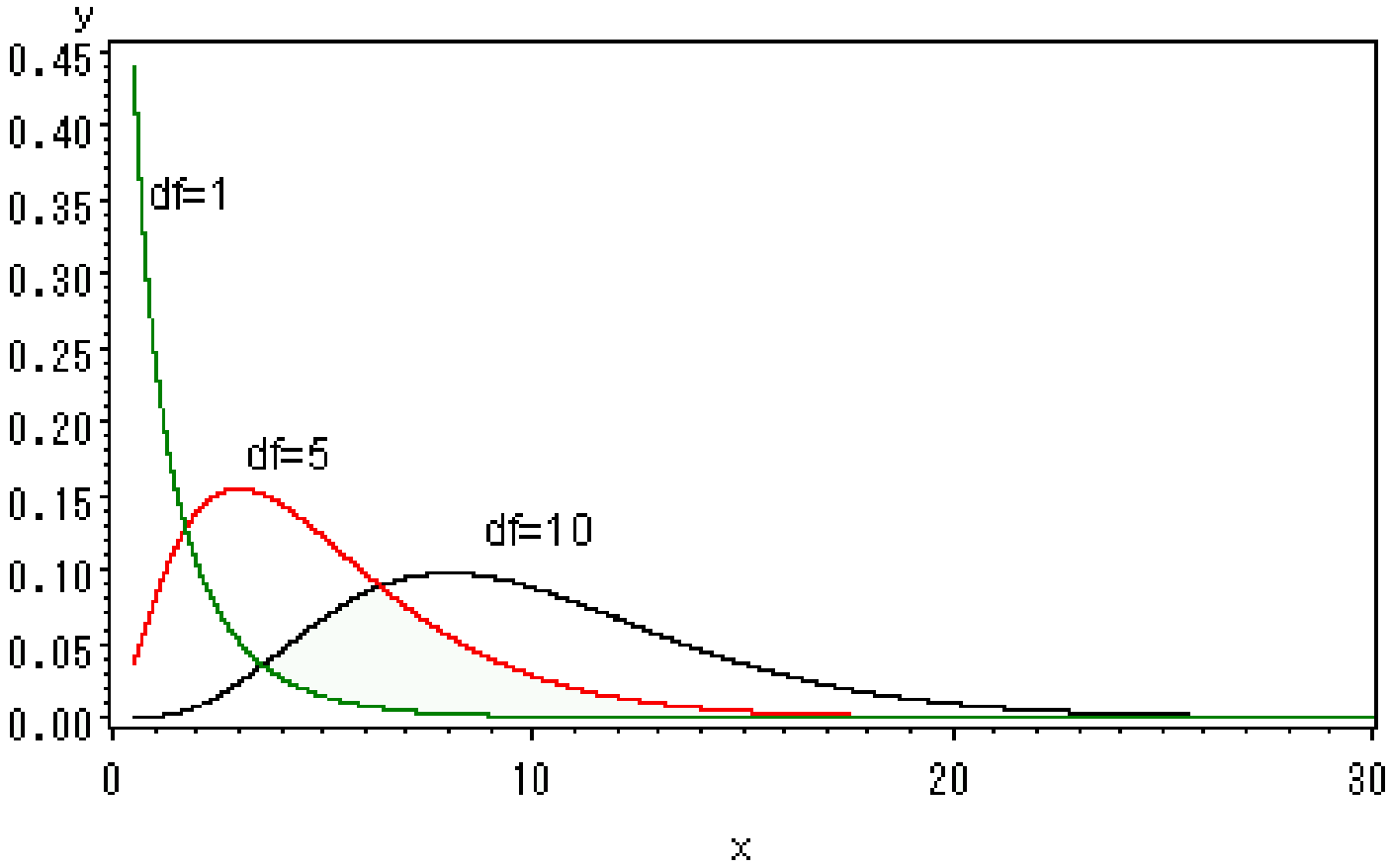
f_o = 观察频数

f_e = 期望频数

卡方检验的自由度为df = (R-1)(C-1)，R和C分别是两个分类型变量的类别量

卡方检验介绍

卡方分布



卡方分布与自由度的关系

卡方检验介绍

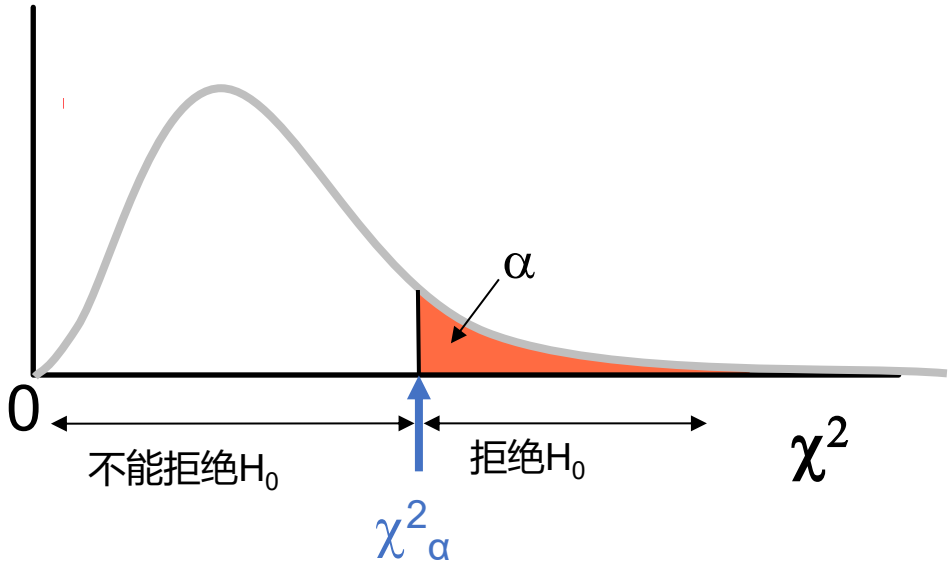
H_0: 两分类变量相互独立。

H_1 :两分类变量不是相互独立。

“独立”是一个专业术语，是没有关联的意思。

χ^2 检验统计量服从卡方分布

如果检验统计量大于临界值（ $\chi^2 > \chi^2_\alpha$ ），拒绝 H_0,否则则不能拒绝 H_0



卡方检验介绍

卡方检验的适用条件

使用卡方检定需要注意：

- 1、少于20%的期望频数小于5
- 2、每一格期望频数大于等于1

方可使用

例题

性别与电脑类型

性别	电脑类型		
	Mac	Windows	
女性	观察频数 = 12 期望频数 = 14.4	观察频数= 108 期望频数 = 105.6	120
男性	观察频数= 24 期望频数 = 21.6	观察频数= 156 期望频数 = 158.4	180
	36	264	300

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

$$= \frac{(12-14.4)^2}{14.4} + \frac{(108-105.6)^2}{105.6} + \frac{(24-21.6)^2}{21.6} + \frac{(156-158.4)^2}{158.4} = 0.7576$$

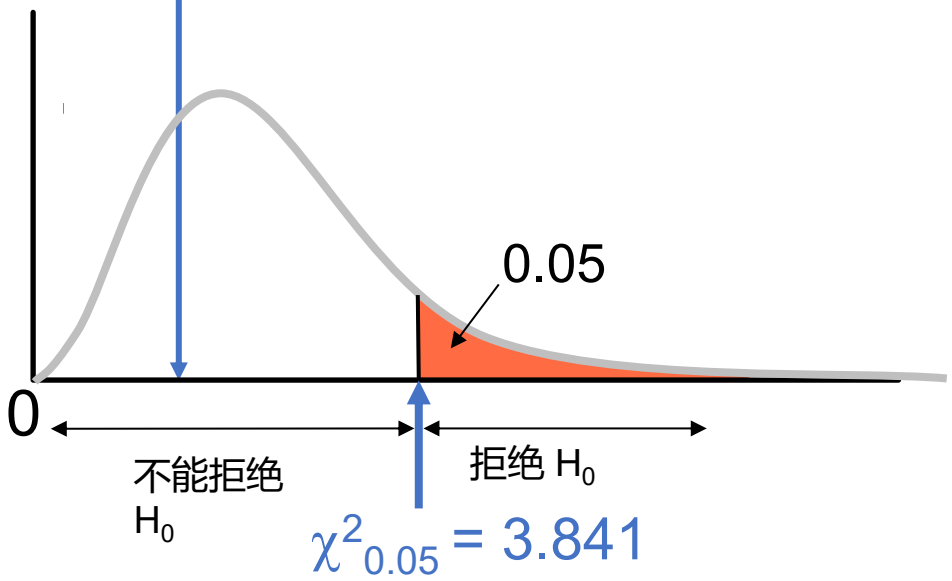
例题

H_0: 性别与使用电脑系统两变量相互独立。

H_1:性别与使用电脑系统两变量不是相互独立。

检验统计量 $\chi^2 = 0.7576$; 临界值为 $\chi^2_{0.05} = 3.841$

自由度df = (R – 1)(C - 1) = (2-1)(2-1) = 1



$\chi^2 = 0.7576 < \chi^2_{0.05} = 3.841$,
因此不能拒绝H_0，因此在 α = 0.05的水平下我们不能认为性别与使用的电脑类型存在相关

软件实操

如果使用Excel做卡方检验，需要手动整理观察频数表和期望频数表，然后使用函数CHISQ.TEST计算卡方检验的p值。

=CHISQ.TEST(B2:C3,F2:G3)

	A	B	C	D	E	F	G	H	I
		Mac	Windows			Mac	Windows		
	女性	12	108		女性	14.4	105.6		0.384088
	男性	24	156		男性	21.6	158.4		
		观察频数表				期望频数表			

软件实操


分析(A)
图形(G)
实用程序(U)
扩展(X)
窗口(W)
报告(P)
描述统计(E)
贝叶斯统计信息(B)
表(B)
比较平均值(M)
一般线性模型(G)
广义线性模型(Z)
混合模型(X)
相关(C)
回归(R)
对数线性(O)
神经网络(W)
分类(F)
降维(D)
刻度(A)
非参数检验(N)
时间序列预测(T)
生存分析(S)
多重响应(U)
缺失值分析(Y)...
多重插补(T)
复杂抽样(L)
模拟(I)...
质量控制(Q)
空间和时间建模(S)...
直销(K)
频率(F)...
描述(D)...
探索(E)...
交叉表(C)...
比率(R)...
P-P 图...
Q-Q 图...
Q2_选


交叉表
行(O):
性别
列(C):
电脑类型
层 1 / 1
上一个(V)
下一个(N)
在表层中显示层变量(L)
显示簇状条形图(B)
禁止显示表(T)
精确(X)...
统计(S)...
单元格(E)...
格式(F)...
样式(L)...
自助抽样(A)...
确定
粘贴(P)
重置(R)
取消
帮助

交叉表：统计
卡方(H)
相关性(R)
名义
列联系数(O)
Phi 和克莱姆 V
Lambda
不确定性系数(U)
有序
Gamma
萨默斯 d(S)
肯德尔 tau-b
肯德尔 tau-c
按区间标定
Eta
Kappa
风险(I)
麦克尼马尔(M)
柯克兰和曼特尔-亨塞尔统计(A)
检验一般比值比等于(T):
1
继续(C)
取消
帮助

软件实操

SPSS提供了多种卡方检验的结果，我们一般只要选择看第一行的就行了。

结论：性别与电脑类型无显著关联

（ $\chi^2 = 0.758, p = 0.384 > 0.05$ ）

个案处理摘要

	个案					
	有效		缺失		总计	
	N	百分比	N	百分比	N	百分比
性别 * 电脑类型	300	100.0%	0	0.0%	300	100.0%

性别 * 电脑类型 交叉表

计数

		电脑类型		
		1.00	2.00	总计
性别	1.00	12	108	120
	2.00	24	156	180
总计		36	264	300

卡方值

p值

卡方检验

	值	自由度	渐进显著性（双侧）	精确显著性（双侧）	精确显著性（单侧）
皮尔逊卡方	.758[a]	1	.384		
连续性修正[b]	.475	1	.491		
似然比	.772	1	.380		
费希尔精确检验				.469	.247
线性关联	.755	1	.385		
有效个案数	300				

a. 0 个单元格 (0.0%) 的期望计数小于 5。最小期望计数为 14.40。

b. 仅针对 2x2 表进行计算

练习

性别与存活状况

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α=0.05的显著性水平检验存活状况与性别是否有关。


谢谢！