《新媒体数据运营与分析》

Excel (7): 统计分析-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**（男/女）与**学生干部**（干部/非干部）的关系。
    - **专业类型**（人文社科/理工科）与**考公**（考/不考）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

其中一个期望频数的计算：

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

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）

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，

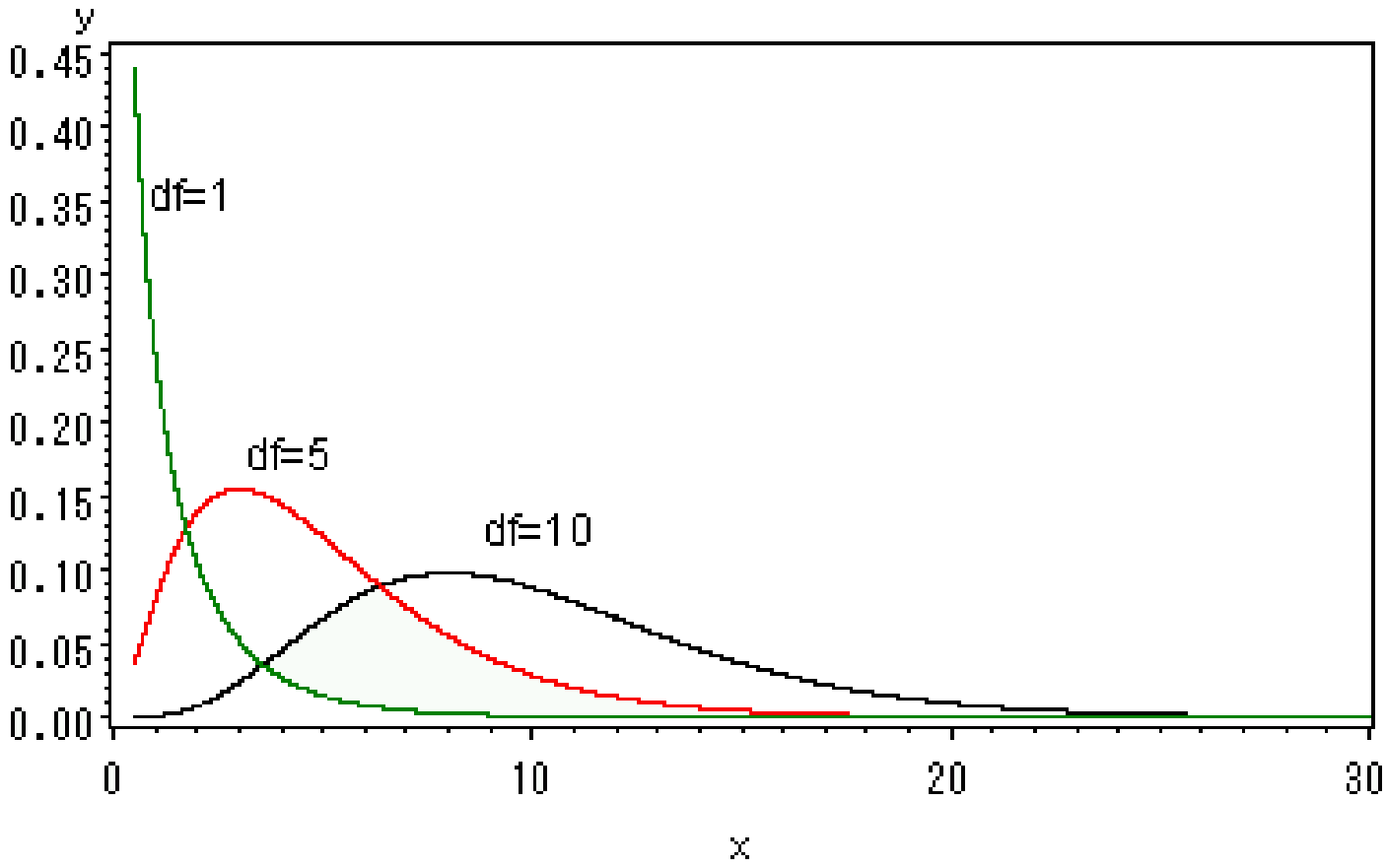
f_o = 观察频数

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

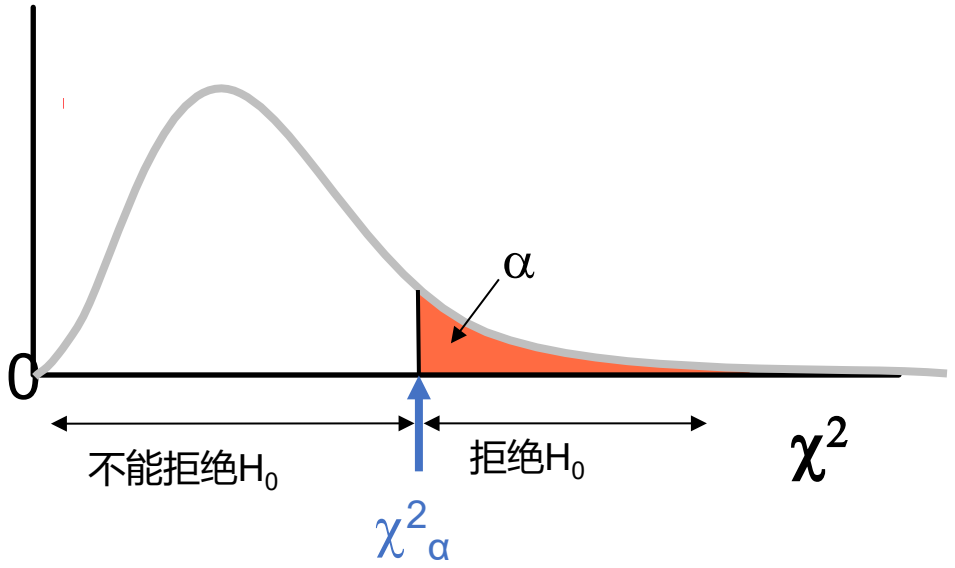
H_0: 两分类变量相互独立。

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

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

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

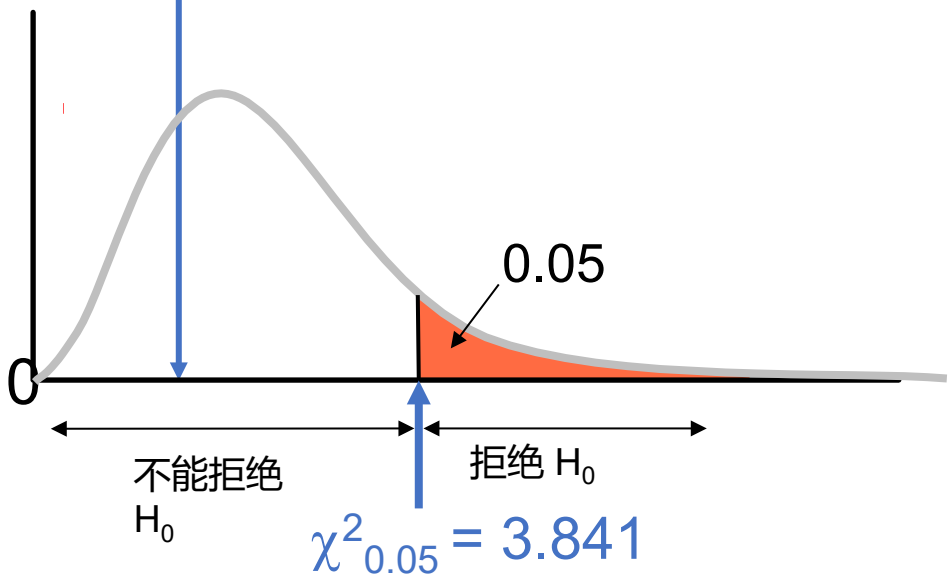
例题

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 $\alpha = 0.05$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。**
- **语法**：CHISQ.TEST(**观测频数表范围, 期望频数表范围**)
- **结果解读：**
  - **p值较小（通常 < 0.05）：**意味着观测到的数据与期望值有显著差异，可以拒绝原假设，说明变量之间存在关联。
  - **p值较大：**意味着观测值与期望值的差异不大，不能拒绝原假设，说明变量之间无显著关联。

软件实操

观察频数表

	Mac	Windows	总计	
女性	12	108	120	行总和（CT）
男性	24	156	180	行总和（CT）
总计	36	264	300	
	列总和（RT）	列总和（RT）		样本量（n）

期望频数表

	Mac	Windows
女性	14.4	105.6
男性	21.6	158.4

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

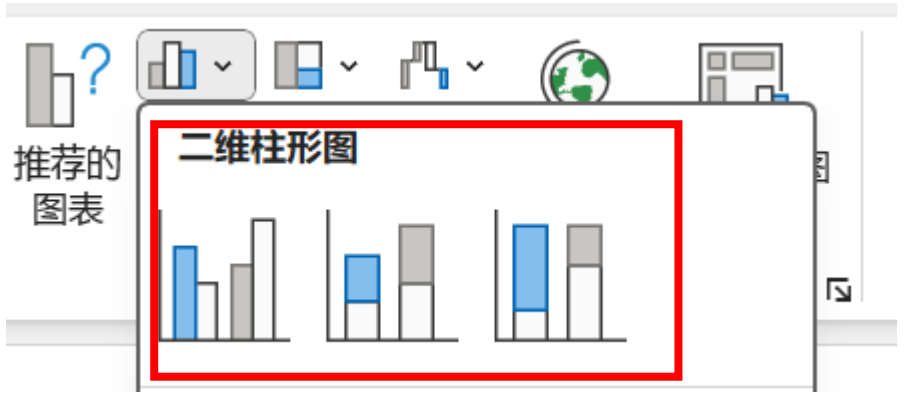
卡方检验

卡方检验的p值： 0.384088249 =CHISQ.TEST(B25:C26,K25:L26)

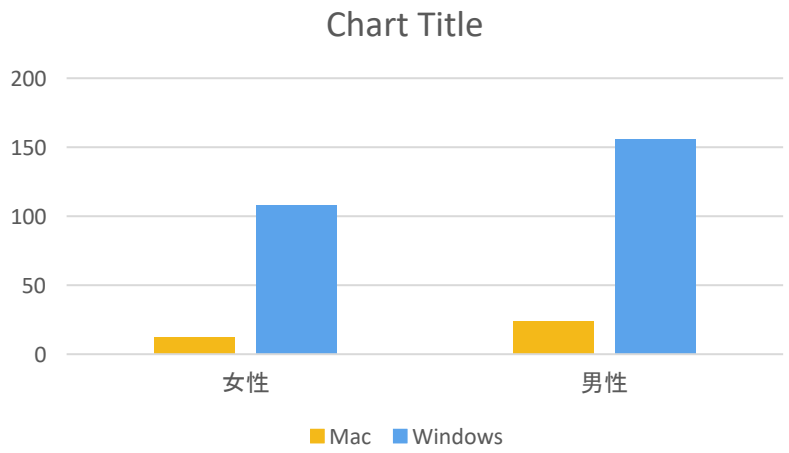
p < 0.05 → 有显著关联
p > 0.05 → 无显著关联

软件实操

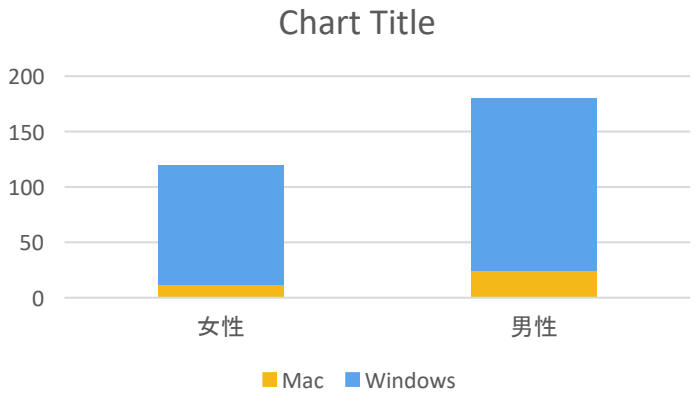
对于两个分类型变量关系的可视化，我们可以选用柱状图



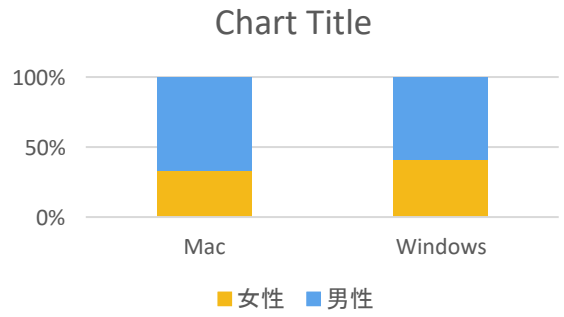
簇状柱状图



堆叠柱状图



百分比堆积柱状图



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。

参考资料

- Lizongzhang的个人空间-合集_·Excel 数据分析实战
- 《Data Visualization in Excel》(Jonathan Schwabish)


谢谢！