# 第六章 列联分析

## 6.1 类别数据与列联表

《泰坦尼克号》这部电影在 1997 年上映，一上映就爆火全世界。而其是根据真实事件改编的，早在 1912 年巨轮泰坦尼克号就不小心和冰山碰撞到一起，导致坠船。当时的泰坦尼克号伤亡表如下图所示

| 年龄/性别 | 舱位/身份 | 获救 | 罹难 | 总计 |
|---|---|---|---|---|
| 儿童 | 头等舱 | 5 | 1 | 6 |
| | 二等舱 | 24 | 0 | 24 |
| | 三等舱 | 27 | 52 | 79 |
| 女人 | 头等舱 | 140 | 4 | 144 |
| | 二等舱 | 80 | 13 | 93 |
| | 三等舱 | 76 | 89 | 165 |
| | 船员 | 20 | 3 | 23 |
| 男人 | 头等舱 | 57 | 118 | 175 |
| | 二等舱 | 14 | 154 | 168 |
| | 三等舱 | 75 | 387 | 462 |
| | 船员 | 192 | 693 | 885 |

。

我们都知道统计有类别数据和数值数据之分。对于类别数据，最终结果也显示为数值，但不同数值描述的对象特征不同。例如，在泰坦尼克号伤亡情况分析的例子中，如果要讨论死亡是否与性别有关，可以将成人群体分为男性和女性两类，将男性标记为 1，将女性标记为 0。要研究死亡是否与船舱有关，也可以将乘客分为头等舱乘客（标记为 1）、二等舱乘客（标记为 2）、三等舱乘客（标记为 3）。显然，对上述问题的分析是以统计数据的综合分类为基础的。另外，为了方便后续分析工作的开展，选择有效的方式组织数据也是必要的。这种有效的方式就是所谓的列联表。

列联表是由两个以上的变量进行交叉分类的频数分布表。如下图

| 人员状况 | 头等舱 | 二等舱 | 三等舱 | 总计 |
|---|---|---|---|---|
| 获救 | 202 | 118 | 178 | 498 |
| 罹难 | 123 | 167 | 528 | 818 |
| 总计 | 325 | 285 | 706 | 1316 |

下一小节我们就来介绍卡方检验

## 6.2 Pearson 的卡方检验

皮尔逊在 1899 提出了卡方统计量，可以写成

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

下面我们来举一个基于列联表的例子

据悉，女性怀孕期间饮酒或吸烟会对胎儿产生不良影响。有些人认为饮酒和吸烟之间有某种联系，比如一般酗酒的人都有吸烟的爱好。了解两者之间的关系对于研究孕妇的相关行为对胎儿可能产生的影响至关重要。1984 年，研究人员对 452 名母亲进行了调查，根据他们得知怀孕前的酒精和香烟摄入量，得出了如下表（表 7）所示的序列表。请问喝酒和抽烟有关系吗？

表 7 吸烟与饮酒的关系图

| | | 尼古丁摄入（mg/d） | | | |
|---|---|---|---|---|---|
| | | 0 | 1~15 | ≥16 | 总计 |
| 酒精摄入（oz/d） | 0 | 105 | 7 | 11 | 123 |
| | 0.01~0.10 | 58 | 5 | 13 | 76 |
| | 0.11~0.99 | 84 | 37 | 42 | 163 |
| | ≥1.00 | 57 | 16 | 17 | 90 |
| | 总计 | 304 | 65 | 83 | 452 |

当使用列联表作为基础进行假设检验时，通常假设原假设 H0 为两个因素间不存在联系、彼此独立，这是因为列联表由两个因素的横向与纵向交叉而成。同样，另一种选择假设 H1 的原始假设是错误的。在掷骰子的例子中，H0 确定了所有可以输出的概率，那时 H0 只指定了概率之间的关系。对于饮酒和吸烟关系的例子，我们可以提出以下原假设和预备假设。

H0：吸烟和饮酒之间没有关系，即两者是独立的。

H1：原假设是错误的。

下面我们来看饮酒与吸烟的期望数据

| | | 尼古丁摄入（mg/d) | | | |
|---|---|---|---|---|---|
| | | 0 | 1~15 | ≥16 | 总计 |
| 酒精摄入（oz/d) | 0 | 82.73 | 17.69 | 22.59 | 123 |
| | 0.01~0.10 | 51.12 | 10.93 | 13.96 | 76 |
| | 0.11~0.99 | 109.63 | 23.44 | 29.93 | 163 |
| | ≥1.00 | 60.53 | 12.94 | 16.53 | 90 |
| | 总计 | 304 | 65 | 83 | 452 |

再考虑一下用于检查 x 2 分布的自由度，对于列联表而言，般计算公式为 df=（r-1)。×（c-1）=rc-（r+c-1）其中 r 表示行数，c 表示列数，因此 rc 是表中列出的类别总数。r 同时给出了 r 个限制条件，列数 c 同时给出了 c 个限制条件。然而总行和＝总列和＝表中数值总和所以在计算行限制和列限制给定的限制条件的数量时，会有重复计算，我们必须去掉。最后一个限制是 r+c-1。针对当

前所讨论的问题，自由度为

、

$$df=（r-1）\times（c-1）=（4-1）\times（3-1）=6$$

我们通过查表法可以查到卡方 6 的临界值，得到的 P 值小于 0.001。我们用 R 语言来计算 P 值

```
> pchisq(42.252,6,lower.tail = F)
[1] 1.639671e-07
```

我们通过计算的 P 值可以拒绝原假设，即饮酒和吸烟之间是有关联的。

### 6.3 列联分析应用条件

列联分析是一种用来分析两个变量之间关系的统计方法，所以他也有应用条件，一般来说列联分析适用于以下几种情况：

变量类型：列联分析适用于分析分类变量之间的关系，可以是名义变量或有序变量。

样本量：所用样本的规模应足够大，通常建议每一格中样本数不应小于 5，而且百分之 80 的期望频数都应该大于 5。

式样设计：列联分析需要在受试者间进行比较，因此，应该优化试验设计，尽可能排除可能造成偏差的因素。

数据收集：收集到的数据应该准确、可靠、完整。

变量独立：列联分析是在变量相互独立的基础上运作的，因此，两个变量之间不应该有重叠或重复的项目。

我们来看一个例子，我们想知道阿尔兹海默患者和铝元素的摄入是否存在关联，我们这里统计了普通人和阿尔兹海默患者中铝元素的摄入量，如下表（表 8）

表 8

| | 含铝抗酸剂 | | | |
|---|---|---|---|---|
| | 无 | 低 | 中 | 高 |
| 阿尔兹海默患者 | 112 | 3 | 5 | 8 |
| 控制组 | 114 | 9 | 3 | 2 |

下面我们将用代码来进行列联分析

```
> aluminum.by.alzheimkers<-matrix(c(112,3,5,8,114,9,3,2),nrow = 2,byrow = TRUE)  %该矩阵为两行四列
> (a.by.a.test<-chisq.test( aluminum.by.alzheimkers))
        Pearson's Chi-squared test

data:  aluminum.by.alzheimkers
X-squared = 7.1177, df = 3, p-value = 0.06824
```

由于我们前面说过列联分析有频数要求，所以我们在 rstudio 中用代码来检验一下所有期望的频数。

```
> a.by.a.test$expected
     [,1] [,2] [,3] [,4]
[1,]  113    6    4    5
[2,]  113    6    4    5
```

这里的频数有 25%小过了 5，所以算出来的 P 可信度没有这么高。所以我们用另

一种更加可靠的方法来算出更加精准的 P。

```
> chisq.test(aluminum.by.alzheimkers,simulate.p.value = TRUE) %我们将模拟方法来计算 P 值
        Pearson's Chi-squared test with simulated p-value (based on
        2000 replicates)

data:  aluminum.by.alzheimkers
X-squared = 7.1177, df = NA, p-value = 0.06197
```

由上述程序可得 P 的值为 0.06197，所以阿尔兹海默症与铝元素的摄入实际上是有相互联系的。