统计学与 R 语言

第 23 讲　卡方检验

张敬信

2022 年 5 月 16 日

哈尔滨商业大学

- **参数检验：**要求样本来自的总体分布已知，对总体参数进行估计；能充分利用数据信息，检验效能高；但对数据要求高，适用范围窄。
- **非参数检验：**不对总体参数进行推断，不受总体分布限制，适用范围广，对数据要求不高，检验效能相对较低，难充分利用数据信息。

分类数据分析是分析频数，卡方检验是最常用的分析分类数据频数的非参数检验方法。

一. 卡方检验原理

卡方检验，是针对无序分类变量的非参数检验，其理论依据是：实际观察频数 f_0 与理论频数 f_e（又称期望频数）之差的平方再除以理论频数所得的统计量，近似服从 χ^2 分布：

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

卡方检验的一般是用来检验无序分类变量的实际观察频数和理论频数分布之间是否存在显著差异，要求：

- 分类变量相互排斥，互不包容；
- 观测相互独立；
- 样本容量不宜太小，理论频数 ≥ 5，否则需要进行校正（合并单元格或校正卡方值）

卡方检验常用于：

- **拟合优度检验**：检验数据”符合”概率模型的”程度”，比如分类变量各类的出现概率是否等于指定概率；
- **同质性检验**：检验两组频数是否来自同一总体，若是，则每一类出现的概率应该是差不多的；检验两种方法的结果是否一致，例如两种方法对同一批人进行诊断，其结果是否一致；
- **独立性/关联性检验**：检验两个分类变量是否相互独立；
- **配对卡方检验**：检验具有相关性的两组比例有无差异；
- **比例趋势检验**：检验不同组别的变化趋势有无差异。

二. 卡方拟合优度检验

- 即检验数据" 符合" 概率模型的" 程度" 如何?

例 1 单比例检验

例 1 男女占比假设某学校总体男女占比为 60% 和 40%, 若随机抽样抽到男生 53 人，女生 47 人，那么该样本能否代表总体？或者说数据是否很好地符合" 男生 60%, 女生 40%"的概率模型?

$$H_0 : p_m = 0.60, \quad H_1 : p_m \neq 0.60$$

记抽样总人数为 n, 则抽取的男生人数 Y_m 服从二项分布 $B(n, p_m)$, 女生人数 Y_f 服从 $B(n, 1 - p_m)$.

对足够大的样本（$np_m \geq 5,\ n(1-p_m) \geq 5$），由中心极限定理，

$$Z = \frac{Y_m - np_m}{\sqrt{np_m(1-p_m)}}$$

近似服从 $N(0,1)$, 从而 Z^2 近似服从自由度为 1 的 χ^2 分布。可进一步计算统计量卡方值（$p_f = 1 - p_m$）：

$$Z^2 = \frac{(Y_m - np_m)^2}{np_m} + \frac{(Y_f - np_f)^2}{np_f} := Q_1$$

注： Y_m, Y_f 分别就是男生类、女生类的观测频数；np_m, np_f 分别就是男生类、女生类的期望频数。

- 手动计算

```r
y1 = 53; y2 = 47
n = y1 + y2
pm = 0.6; pf = 1 - pm
Q1 = (y1 - n*pm)^2 / (n * pm) + (y2 - n*pf)^2 / (n * pf)
Q1                                  # 卡方统计量
#> [1] 2.04
alpha = 0.05
qchisq(1-alpha, 1)                  # 临界值
#> [1] 3.84
1 - pchisq(Q1, 1)                   # P 值
#> [1] 0.153
```

- 直接用 rstatix::prop_test() 函数

```
library(rstatix)
prop_test(53, 100, 0.60, correct = FALSE)
#> # A tibble: 1 x 5
#>       n statistic    df     p p.signif
#> * <dbl>     <dbl> <int> <dbl> <chr>
#> 1   100      2.04     1 0.153 ns
# 或者用精确二项检验
binom_test(x = 53, n = 100, p = 0.60)
#> # A tibble: 1 x 6
#>       n estimate conf.low conf.high     p p.signif
#> * <dbl>    <dbl>    <dbl>     <dbl> <dbl> <chr>
#> 1   100     0.53    0.428     0.631 0.154 ns
```

注：这里是为了让结果一致，强行设置不修正，正常做修正更好。

(2) 多比例检验

单比例检验涉及两个类别的比例，只需要检验一个比例。这可以推广到更多 k 个类别：

类别	1	$\cdots$	$k-1$	k
观测频数	Y_1	$\cdots$	Y_{k-1}	$n-\sum_{i=1}^{k-1} Y_i$
期望频数	np_1	$\cdots$	np_{k-1}	np_k

其卡方统计量为：

$$Q_{k-1}=\sum_{i=1}^{k}\frac{(Y_i-np_i)^2}{np_i}$$

近似服从 $\chi^2(k-1)$.

例 2 巧克力糖果

- 检验抽样频数是否符合理论比例?

类别	棕色	黄色	橘色	绿色	咖色	总计
观测频数	224	119	130	48	59	580
理论比例	0.4	0.2	0.2	0.1	0.1	1

- 手动计算

```
x = c(224,119,130,48,59)
n = sum(x)
p = c(0.4,0.2,0.2,0.1,0.1)
fe = n * p                          # 理论频数
```

```r
Q4 = sum((x - fe)^2 / fe)
Q4
#> [1] 3.78
k = 5
qchisq(1-alpha, k-1)            # 临界值
#> [1] 9.49
1 - pchisq(Q4, k-1)             # P 值
#> [1] 0.436
```

- 直接用 rstatix::chisq_test()函数

```
chisq_test(x, p = p)
#> # A tibble: 1 x 6
#>       n statistic     p    df method          p.signif
#> * <int>     <dbl> <dbl> <dbl> <chr>           <chr>
#> 1     5      3.78 0.436     4 Chi-square test ns
```

注 1：前文比例检验相当于是检验参数已知的概率分布，若参数未知，也可以用卡方拟合优度检验：先估计出未知参数，再依次计算各类别概率、期望频数、卡方统计量，检验显著性。但是注意，卡方统计量的自由度需要额外减去未知参数个数。

注 2：上述方法直接适用于离散分布。对于连续分布，额外需要先将观测值切分成若干小段（类别），统计小段内的频数。

具体原理和案例，可参阅STAT 415: Introduction to Mathematical Statistics.

三. 卡方同质性检验

二维**列联表**，是将两个分类变量，分别放在行和列，交叉单元格为相应类别的交叉频数。

变量	B_1	B_2	$\cdots$	B_k	总计
A_1	$Y_{11}\,(p_{11})$	$Y_{12}\,(p_{12})$	$\cdots$	$Y_{1k}\,(p_{1k})$	$Y_{1\cdot}\,(p_{1\cdot})$
A_2	$Y_{21}\,(p_{21})$	$Y_{22}\,(p_{22})$	$\cdots$	$Y_{2k}\,(p_{2k})$	$Y_{2\cdot}\,(p_{2\cdot})$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
A_h	$Y_{h1}\,(p_{h1})$	$Y_{h2}\,(p_{h2})$	$\cdots$	$Y_{hk}\,(p_{hk1})$	$Y_{h\cdot}\,(p_{h\cdot})$
总计	$Y_{\cdot 1}\,(p_{\cdot 1})$	$Y_{\cdot 2}\,(p_{\cdot 2})$	$\cdots$	$Y_{\cdot k}\,(p_{\cdot k})$	n

注： $p_{ij} = P(A_i \cap B_j)$, $p_{i\cdot} = P(A_i)$, $p_{\cdot j} = P(B_j)$.

卡方同质性检验，是检验列联表两行的列类别频数分布是否同质（没有差别）。

例 3 检验主治医师与实习医师的不必要输血是否同质

- 列联表数据（交叉人数）如下：

医师	经常	偶尔	很少	从不	总计
主治医师	2 (4.1%)	3 (6.1%)	31 (63.3%)	13 (26.5%)	49
实习医师	15 (21.1%)	28 (39.4%)	23 (32.4%)	5 (7.0%)	71
总计	17	31	54	18	120

检验两种医师不必要输血的比例（这两行百分数）是否同质?

$$H_0：\text{两行内容同质，即} p_{1j} = p_{2j}, \quad j = 1, \cdots, 4$$

列联表里的数据就是观测频数，再来计算理论频数。

这里的自由度是 $(2-1)*(4-1)=3$, 也就是说只有 3 个单元格需要计算，其余单元格均可由总计依次做差得到：

医师	经常	偶尔	很少	从不	总计
主治医师	6.942	12.658	22.05		49
实习医师					71
总计	17	31	54	18	120

其中，

$$49 \cdot \frac{17}{120} = 6.942, \quad 49 \cdot \frac{31}{120} = 12.658, \quad 49 \cdot \frac{54}{120} = 22.05$$

上表就是理论频数，从而可以计算卡方统计量：

$$\begin{aligned}\chi^2(3) &= \sum_{i=1}^{h}\sum_{j=1}^{k}\frac{(y_{ij}-n_i\hat{p}_j)^2}{n_i\hat{p}_j}\\ &= \frac{(2-6.942)^2}{6.942}+\cdots+\frac{(5-10.65)^2}{10.65}\\ &= 31.881\end{aligned}$$

- 手动计算

```
xtab = matrix(c(2,3,31,13,
                15,28,23,5), nrow = 2, byrow = TRUE)
dimnames(xtab) = list(c(" 主治医师"," 实习医师"),
                      c(" 经常"," 偶尔"," 很少"," 从不"))
xtab
#>          经常 偶尔 很少 从不
#> 主治医师    2    3   31   13
#> 实习医师   15   28   23    5
n = sum(xtab)                    # 总和
rs = rowSums(xtab)               # 行和
cs = colSums(xtab)               # 列和
```

```r
fe = rbind(rs[1] * cs / n, rs[2] * cs / n)
fe                                          # 理论频数
#>       经常 偶尔 很少  从不
#> [1,]  6.94 12.7 22.1  7.35
#> [2,] 10.06 18.3 31.9 10.65
Q3 = sum((xtab - fe)^2 / fe)                # 卡方统计量
Q3
#> [1] 31.9
h = 2; k = 4
f = (h-1)*(k-1)
qchisq(1-alpha, f)                          # 临界值
#> [1] 7.81
1 - pchisq(Q3, f)                           # P 值
#> [1] 5.54e-07
```

- 直接用 rstatix::chisq_test() 函数计算

```
rlt = chisq_test(xtab)
rlt
#> # A tibble: 1 x 6
#>       n statistic           p    df method           p.sign
#> * <dbl>     <dbl>       <dbl> <int> <chr>            <chr>
#> 1   120      31.9 0.000000554     3 Chi-square test ****
expected_freq(rlt)                  # 提取期望频数
#>             经常 偶尔 很少  从不
#> 主治医师  6.94 12.7 22.1  7.35
#> 实习医师 10.06 18.3 31.9 10.65
```

注意：若单元格的频数过小（< 5），不适合做卡方检验，可以将多个小类别适当合并，或者做 Fisher 精确检验 `rstatix::`fisher_test()‘:

```
fisher_test(xtab, detailed = TRUE)
#> # A tibble: 1 x 5
#>       n           p method                  alternative p.signi
#> * <dbl>       <dbl> <chr>                   <chr>       <chr>
#> 1   120 0.00000015 Fisher's Exact test two.sided   ****
```

四. 卡方独立/关联性检验

卡方独立/关联性检验，就是分析列联表中行变量和列变量是否相互独立（有无关联）。

$$H_0 = \text{变量}A\text{与变量}B\text{相互独立（无关联）}$$

即 $P(A_i \cap B_j) = P(A_i)P(B_j),\ \forall i, j.$

卡方检验统计量：

$$\chi^2 = \sum_{j=1}^{k}\sum_{i=1}^{h} \frac{(Y_{ij} - Y_{i\cdot}Y_{\cdot j}/n)^2}{Y_{i\cdot}Y_{\cdot j}/n}$$

近似服从自由度为 $(h-1)(k-1)$ 的卡方分布。

卡方独立性/关联性的检验统计量，与卡方同质性的检验统计量是等价的！

计算过程和结果完全相同！

例 4 检验地区和原料质量是否存在关联关系

地区/质量	一级	二级	三级	总计
甲	52	64	24	140
乙	60	59	52	171
丙	50	65	74	189
总计	162	188	150	500

该二维列联表即观测频数表，接着直接把前面的计算过程照搬过来即可。

- 手动计算

```r
xtab = matrix(c(52,64,24,
                60,59,52,
                50,65,74), nrow = 3, byrow = TRUE)
dimnames(xtab) = list(c(" 甲"," 乙"," 丙"),
                      c(" 一级"," 二级"," 三级"))
xtab
#>    一级 二级 三级
#> 甲   52   64   24
#> 乙   60   59   52
#> 丙   50   65   74
n = sum(xtab)                    # 总和
rs = rowSums(xtab)               # 行和
cs = colSums(xtab)               # 列和
```

```r
fe = rbind(rs[1] * cs / n, rs[2] * cs / n, rs[3] * cs / n)
fe                                          # 理论频数
#>       一级 二级 三级
#> [1,] 45.4 52.6 42.0
#> [2,] 55.4 64.3 51.3
#> [3,] 61.2 71.1 56.7
Q4 = sum((xtab - fe)^2 / fe)                # 卡方统计量
Q4
#> [1] 19.8
h = 3; k = 3
f = (h-1)*(k-1)
qchisq(1-alpha, f)                          # 临界值
#> [1] 9.49
1 - pchisq(Q4, f)                           # P 值
#> [1] 0.000541
```

- 直接用 rstatix::chisq_test() 函数计算

```
rlt = chisq_test(xtab)
rlt
#> # A tibble: 1 x 6
#>       n statistic        p    df method          p.signif
#> * <dbl>     <dbl>    <dbl> <int> <chr>           <chr>
#> 1   500      19.8 0.000541     4 Chi-square test ***
expected_freq(rlt)                 # 提取期望频数
#>    一级 二级 三级
#> 甲 45.4 52.6 42.0
#> 乙 55.4 64.3 51.3
#> 丙 61.2 71.1 56.7
```

五. 配对卡方检验

检验配对比例有无差异，此时组间不再相互独立，而是具有了相关性，适合用 McNemar 卡方检验，H_0：组间比例无差异.

例如，施加干预前后，同一组人吸烟与不吸烟人数有无差异：

```
xtab = matrix(c(25, 6,
                21, 10), nrow = 2, byrow = TRUE)
dimnames(xtab) = list(
  intervention = c("before", "after"),
  smoker = c("yes", "no"))
```

```
xtab
#>              smoker
#> intervention yes no
#>       before  25  6
#>       after   21 10
mcnemar_test(xtab)
#> # A tibble: 1 x 6
#>       n statistic    df       p p.signif method
#> * <dbl>     <dbl> <dbl>   <dbl> <chr>    <chr>
#> 1    62      7.26     1 0.00705 **       McNemar test
```

六. 卡方（比例）趋势检验

可用来检验不同组别的变化趋势是否相同，$H0$：比例趋势相同.

例如，不同年龄段的肾结石比例变化趋势：

```
xtab = matrix(c(384, 536, 335,
                951, 869, 438), nrow = 2, byrow = TRUE)
dimnames(xtab) = list(
  stone = c("yes", "no"),
  age = c("30-39", "40-49", "50-59"))
xtab
#>       age
#> stone 30-39 40-49 50-59
#>   yes   384   536   335
#>   no    951   869   438
```

```r
# 对比组间的比例
prop_trend_test(xtab)
#> # A tibble: 1 x 6
#>       n statistic        p p.signif    df method
#> * <dbl>     <dbl>    <dbl> <chr>    <dbl> <chr>
#> 1  3513      49.7 1.78e-12 ****         1 Chi-square trend
```

本篇主要参阅 (张敬信, 2022), STAT 415: Introduction to Mathematical Statistics, (贾俊平, 2018)，模板感谢 (黄湘云, 2021), (谢益辉, 2021).

参考文献

张敬信 (2022). *R* 语言编程：基于 *tidyverse*. 人民邮电出版社, 北京.

谢益辉 (2021). *rmarkdown: Dynamic Documents for R*.

贾俊平 (2018). 统计学. 中国人民大学出版社, 北京, 7 edition.

黄湘云 (2021). *Github: R-Markdown-Template*.