生物统计基础 (P3)

统计 91 董晟渤, 2193510853

西安交通大学数学与统计学院

日期：2022 年 5 月

目录

1 方法: 两样本独立性检验

1.1 问题提出

假设我们有一批数据, 关于特征 A 和特征 B 可以对它们进行分类. 例如, A 和 $\bar{A}$ 可以表示受试者是否具有某项特征, 而 B 和 $\bar{B}$ 可以表示受试者是否患有某种疾病. 将数据整理, 并列出表格如**表 1**所示.

表 1: 数据关于特征 A 和特征 B 的分类

	B	$\bar{B}$	合计
A	n_{11}	n_{12}	$n_{1\cdot}$
$\bar{A}$	n_{21}	n_{22}	$n_{2\cdot}$
合计	$n_{\cdot 1}$	$n_{\cdot 2}$	n

我们希望能够探究: 特征 A 与特征 B 是否是独立的? 对于上面所举的例子而言, 就是探究受试者的某项特征是否会引起或者抑制某种疾病. 通常, 记

$$p_1 = \mathbb{P}\left(B|A\right), \quad p_2 = \mathbb{P}\left(B\middle|\bar{A}\right),$$

它们对应的估计为

$$\hat{p}_1 = \frac{n_{11}}{n_{11}+n_{12}} = \frac{n_{11}}{n_{1\cdot}}, \quad \hat{p}_2 = \frac{n_{21}}{n_{21}+n_{22}} = \frac{n_{21}}{n_{2\cdot}}.$$

如果 A 与 B 是独立的, 则一定有 $p_1 = p_2$, 从而要研究特征 A 与特征 B 是否是独立的, 所要检验的假设为

$$H_0 : p_1 = p_2.$$

本篇报告中, 将叙述对该检验的精确检验方法 (即 Fisher 精确检验) 和近似检验方法 (即 χ^2 检验), 并分别应用这两种检验方法于模拟数据和实际数据中, 探讨这两种方法的有效性.

1.2 Fisher 精确检验

假定样本量 n 已知, 并且**表 1**中的 $n_{1\cdot}, n_{2\cdot}, n_{\cdot 1}$ 和 $n_{\cdot 2}$ 均已知. 在 H_0 成立的情况下, 我们需要求出 $n_{11}, n_{12}, n_{21}, n_{22}$ 的零分布, 并基于该分布得到对于实际数据的 p 值的计算公式.

若 $p_1 = p_2$, 则不管是已知 A 还是 $\bar{A}$, 事件 B 发生的概率都相等, 且等于 $\mathbb{P}(B)$. 记 $p = \mathbb{P}(B)$. 首先, 从 $n_{1\cdot} = n_{11} + n_{12}$ 个满足特征 A 的数据中, 选出满足 n_{11} 个满足特征 B 的

数据的概率为

$$\binom{n_{11}+n_{12}}{n_{11}} \cdot p^{n_{11}} \cdot (1-p)^{n_{12}}.$$

另外一方面, 从 $n_{2\cdot} = n_{21} + n_{22}$ 个不满足特征 A 的数据中, 选出满足 n_{21} 个满足特征 B 的数据的概率为

$$\binom{n_{21}+n_{22}}{n_{21}} \cdot p^{n_{21}} \cdot (1-p)^{n_{22}}.$$

而对于所有数据而言, 从 n 个数据中, 选出 $n_{\cdot 1} = n_{11} + n_{21}$ 个满足特征 B 的数据的概率为

$$\binom{n_{11}+n_{12}+n_{21}+n_{22}}{n_{11}+n_{21}} \cdot p^{n_{11}+n_{21}} \cdot (1-p)^{n_{12}+n_{22}}.$$

在上面的基础上, 计算得到: 从 n 个数据中, 选出 n_{11} 个满足特征 AB, n_{12} 个满足特征 $A\bar{B}$, n_{21} 个满足特征 $\bar{A}B$, n_{22} 个满足特征 $\bar{A}\bar{B}$ 的概率

$$\begin{aligned}
\mathbb{P}(n_{11}n_{12}n_{21}n_{22}) &= \frac{\binom{n_{11}+n_{12}}{n_{11}}\binom{n_{21}+n_{22}}{n_{21}} \cdot p^{n_{11}+n_{21}} \cdot (1-p)^{n_{12}+n_{22}}}{\binom{n_{11}+n_{12}+n_{21}+n_{22}}{n_{11}+n_{21}} \cdot p^{n_{11}+n_{21}} \cdot (1-p)^{n_{12}+n_{22}}} \\
&= \frac{(n_{11}+n_{12})! \cdot (n_{21}+n_{22})! \cdot (n_{11}+n_{21})! \cdot (n_{12}+n_{22})!}{(n_{11}+n_{12}+n_{21}+n_{22})! \cdot n_{11}! \cdot n_{12}! \cdot n_{21}! \cdot n_{22}!} \\
&= \frac{n_{1\cdot}! \cdot n_{2\cdot}! \cdot n_{\cdot 1}! \cdot n_{\cdot 2}!}{n! \cdot n_{11}! \cdot n_{12}! \cdot n_{21}! \cdot n_{22}!}.
\end{aligned}$$

这便是我们所要求的 $n_{11}, n_{12}, n_{21}, n_{22}$ 的零分布. 当 $n_{11} = x$ 时, $n_{12} = n_{1\cdot} - x$, $n_{21} = n_{\cdot 1} - x$, $n_{22} = n_{2\cdot} - n_{21} = n_{2\cdot} - n_{\cdot 1} + x$, 也即只要 n_{11} 确定, 其余的数量也会被确定. 在上面的结果的基础上, 进一步有

$$\mathbb{P}(n_{11} = x) = \frac{n_{1\cdot}! \cdot n_{2\cdot}! \cdot n_{\cdot 1}! \cdot n_{\cdot 2}!}{n! \cdot x! \cdot (n_{1\cdot} - x)! \cdot (n_{\cdot 1} - x)! \cdot (n_{2\cdot} - n_{\cdot 1} + x)!}.$$

这便是用于检验原假设的 n_{11} 的零分布.

基于上面的结果, 若 $n_{11} = a$, 则检验的 p 值

$$p = 2\min\left\{\mathbb{P}(n_{11} \le a), \mathbb{P}(n_{11} \ge a), \frac{1}{2}\right\}.$$

记 $k = \min\{n_{1\cdot}, n_{\cdot 1}\}$ 为 n_{11} 的最大可能取值, 则检验的 p 值

$$p = 2\min\left\{\sum_{x=0}^{a}\mathbb{P}(n_{11} = x), \sum_{x=a}^{k}\mathbb{P}(n_{11} = x), \frac{1}{2}\right\}.$$

1.3 χ^2 拟合优度检验

上面的检验是精确的检验, 在计算上常常不太方便. 当样本量较大时, 可以考虑进行 χ^2 拟合优度检验. 考虑到 $\Omega = AB \cup A\bar{B} \cup \bar{A}B \cup \bar{A}\bar{B}$, 记

$$p_{11} = \mathbb{P}(AB), \quad p_{12} = \mathbb{P}\left(A\bar{B}\right), \quad p_{21} = \mathbb{P}\left(\bar{A}B\right), \quad p_{22} = \mathbb{P}\left(\bar{A}\bar{B}\right),$$

它们对应的估计为

$$\hat{p}_{11} = \frac{n_{1\cdot} \cdot n_{\cdot 1}}{n^2}, \quad \hat{p}_{12} = \frac{n_{1\cdot} \cdot n_{\cdot 2}}{n^2}, \quad \hat{p}_{21} = \frac{n_{2\cdot} \cdot n_{\cdot 1}}{n^2}, \quad \hat{p}_{22} = \frac{n_{2\cdot} \cdot n_{\cdot 2}}{n^2}.$$

进行 χ^2 拟合优度检验时, 考虑的统计量为

$$\chi^2 = \frac{(n_{11} - np_{11})^2}{np_{11}} + \frac{(n_{12} - np_{12})^2}{np_{12}} + \frac{(n_{21} - np_{21})^2}{np_{21}} + \frac{(n_{22} - np_{22})^2}{np_{22}},$$

若代入 $\hat{p}_{11}, \hat{p}_{12}, \hat{p}_{21}$ 和 $\hat{p}_{22}$, 则

$$\begin{aligned}
\chi^2 &= \sum_{i=1}^{2}\sum_{j=1}^{2} \frac{(n_{ij} - n\hat{p}_{ij})^2}{n\hat{p}_{ij}} \\
&= \sum_{i=1}^{2}\sum_{j=1}^{2} \frac{(n \cdot n_{ij} - n_{i\cdot} \cdot n_{\cdot j})^2}{n \cdot n_{i\cdot} \cdot n_{\cdot j}} \\
&= \frac{n \cdot (n_{11}n_{22} - n_{12}n_{21})^2}{n_{1\cdot} \cdot n_{2\cdot} \cdot n_{\cdot 1} \cdot n_{\cdot 2}}.
\end{aligned}$$

当 H_0 成立时, 有 $\chi^2 \xrightarrow{L} \chi^2(1)$, 因此当样本量充分大时, 可以进行近似检验. 设代入实际数据计算得到的统计量为 χ_0^2, 则检验的 p 值

$$p = \mathbb{P}(\chi^2 \geq \chi_0^2).$$

2 应用 1: 模拟数据

2.1 模拟数据的生成

生成模拟数据时, 设 $n = 30$. 首先通过二项分布 $\mathcal{B}(n, p_0)$, 生成符合特征 A 和 $\bar{A}$ 的数据 (这里的 p_0 不重要), 并设符合特征 A 和 $\bar{A}$ 的数据共有 $n_{1\cdot}$ 和 $n_{2\cdot}$ 个. 接下来, 分别通过二项分布 $\mathcal{B}(n_{1\cdot}, p_1)$ 和 $\mathcal{B}(n_{2\cdot}, p_2)$, 生成 A 和 $\bar{A}$ 中符合特征 B 的数据 (这里的 p_1 和 p_2 即为我们需要检验的参数). 设置 $p_0 = 0.5, p_1 = 0.3, p_2 = 0.6$, 生成的数据如**表 2**所示.

表 2: 模拟数据

	B	$\bar{B}$	合计
A	4	8	12
$\bar{A}$	13	5	18
合计	17	13	30

2.2 检验的结果

对于**表 2**中的模拟数据, 首先进行 Fisher 精确检验. 此时 $n_{11} = 4$, 而 n_{11} 的最大取值为 12, 计算得

$$p = \mathbb{P}(n_{11} \leq 4) = 0.0830.$$

此时, 若取显著性水平为 0.05, 则接受原假设, 也即认为 $p_1 = p_2$, 或者说认为 A 与 B 不相关; 若取显著性水平为 0.10, 则拒绝原假设, 认为 A 与 B 相关.

接下来, 考虑进行 χ^2 拟合优度检验, 计算得

$$\chi_0^2 = \frac{30 \times (4 \times 5 - 8 \times 13)^2}{12^2 \times 18^2 \times 17^2 \times 13^2} = 4.4344,$$

据此求出

$$p = \mathbb{P}(\chi^2 \geq \chi_0^2) = 0.0352.$$

此时, 若取显著性水平为 0.05, 则拒绝原假设, 也即认为 $p_1 \neq p_2$, 或者说认为 A 与 B 相关.

2.3 重复模拟的结果

同样选定 $n = 30$, 进行多次模拟, 得到的 p 值如**表 3**所示.

表 3: $n = 30$, 重复模拟的结果

检验类型	1	2	3	4	5
Fisher 精确检验	0.0830	0.5471	0.2470	0.0101	0.1552
χ^2 拟合优度检验	0.0352	0.3457	0.1269	0.0037	0.0730

根据**表 3**结果, 在进行 Fisher 精确检验时, p 值会比 χ^2 拟合优度检验更大, 从而更容易接受原假设. 若将样本量设置为 $n = 100$, 再次进行模拟, 所得的结果如**表 4**所示.

表 4: $n = 100$, 重复模拟的结果

检验类型	1	2	3	4	5
Fisher 精确检验	0.0122	0.1887	0.0084	0.0172	0.0142
χ^2 拟合优度检验	0.0070	0.1277	0.0049	0.0100	0.0082

当 $n = 100$ 时, 所得的 p 值都较小, 得到的结论几乎都是拒绝原假设, 认为 A 与 B 相关.

3 应用 2: 真实数据

3.1 真实数据的选取

数据来自课本, 比较了乳腺癌患者与对照组的初产年龄, 见**表 5**.

表 5: 乳腺癌患者与对照组的初产年龄的数据

	初产年龄 ≥ 30	初产年龄 ≤ 29	合计
乳腺癌患者	683	2537	3220
非乳腺癌患者	1498	8747	10245
合计	2181	11284	13465

3.2 检验的结果

此时 $n = 13465$, 无法进行精确检验, 因为 13465! 的计算较为麻烦. 在这里, 我们采用 χ^2 检验的方式. 计算得

$$\chi_0^2 = \frac{13465 \times (683 \times 8747 - 2537 \times 1498)^2}{3220^2 \times 10245^2 \times 2181^2 \times 11284^2} = 78.3698,$$

而

$$p = \mathbb{P}(\chi^2 \geq \chi_0^2) \approx 0,$$

当显著性水平 $\alpha = 0.05$ 时, 拒绝原假设, 也即认为乳腺癌与初产年龄有关. 更具体地说, 在初产年龄 ≥ 30 的受试者中, 有大约 31.32% 患有乳腺癌; 而在初产年龄 ≤ 29 的受试者中, 只有大约 22.48% 患有乳腺癌, 据此得出结论, 初产年龄 ≥ 30 的妇女更易患乳腺癌.

附录

A 代码 1: Fisher 精确检验的零分布

以下代码存储在fisherpdf.m 中.

```
function p = fisherpdf(x, n10, n20, n01, n02, n)
    p = (factorial(n10) .* factorial(n20) .* factorial(n01) .* factorial
        (n02)) ...
         ./(factorial(n) .* factorial(x) .* factorial(n10 - x) .*
            factorial(n01 - x) .* factorial(n20 - n01 + x));
end
```

B 代码 2: 两样本独立性检验

以下代码存储在main.m 中.

```
%% 初始化

clc;
clear;
close;

%% 数据

n = 100;
n1 = binornd(n, 0.5); n2 = n - n1;
n11 = binornd(n1, 0.3); n12 = n1 - n11;
n21 = binornd(n2, 0.6); n22 = n2 - n21;

%% 精确检验

k = min(n11 + n12, n11 + n21);
prob = fisherpdf(0 : k, n11 + n12, n21 + n22, n11 + n21, n12 + n22, n);
p1 = 2 * min([sum(prob(1 : n11 + 1)), sum(prob(n11 + 1 : k + 1)), 0.5]);

%% 近似检验
```

```
chi2 = (n * (n11 * n22 - n21 * n12) ^ 2)/((n11 + n12) * (n21 + n22) * (
   n11 + n21) * (n12 + n22));
p2 = 1 - chi2cdf(chi2, 1);
```