# 几个统计检验方法的实现与评价

统计 81 柴思楚 2183310799

2021 年 5 月 2 日

## 引言

实现 two sample independent test 检验 categorical 数据中两个变量的关系, 比较 chi-square test 和 Fisher exact test, 在模拟数据和一个实际生物数据上评价这两种方法。本文具体使用了带 Yates' continuity correction 的 Pearson's Chi-squared test。,

实现 two sample dependent test 检验 categorical 数据中两个变量的关系, 在模拟数据和一个实际生物数据上评价 McNemar test 方法。本文具体使用了带 continuity correction 的 McNemar's Chi-squared test。

实现 Wilcoxon rank sum test, 在模拟数据和实际数据上和 t-test 比较, 评价这两种方法。本文只考虑数据不打结的情况。

chi-square test, Fisher exact test, McNemar test 以如下列联表为基础

| | 条件 1 | | Total |
|---|---|---|---|
| 条件 2 | a | b | a+b |
| | c | d | c+d |
| Toal | a+c | b+d | N |

Wilcoxon rank sum test, t-test 以如下数据表为基础

| | 1 | 2 | | $n_1$ |
|---|---|---|---|---|
| A 组 | $a_1$ | $a_2$ | ... | $a_{n-1}$ |
| | 1 | 2 | | $n_2$ |
| B 组 | $b_1$ | $b_2$ | ... | $b_{n-2}$ |

# 1 two sample independent test

这一节中将实现和评价 chi-square test 与 Fisher exact test。

对于 chi-square test, 首先计算每一单元的期望值 $E_{ij}, i, j = 1, 2$

$$\begin{aligned} E_{11} = (a+c) \times \frac{a+b}{N}; E_{12} = (b+d) \times \frac{a+b}{N}; \\ E_{21} = (a+c) \times \frac{c+d}{N}; E_{22} = (b+d) \times \frac{c+d}{N} \end{aligned} \tag{1}$$

若至少有一个值不小于 5, 则 chi-square test 有可能不准。若没有一个值小于 5, 则其检验统计量为

$$\chi^2 = \sum_{i=1}^{2} \sum_{j=1}^{2} \frac{(|O_{ij} - E_{ij]}| - 0.5)^2}{E_{ij}} \tag{2}$$

其中 $O_{ij}$ 为观测值。

对于 Fisher exact-test, 首先计算当前列联表的数据, 有

$$p_0 = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{n!a!b!c!d!} \tag{3}$$

在满足 $a+b, c+d, b+c, b+d$ 不变的情况下, 改变各位置数据, 得到一系列 $p_i$, 保留小于 $p_0$ 的 $p_i^{'}$, 相加得 $p$ 值

$$p = p_0 + \sum_i p_i^{'} \tag{4}$$

## 1.1 模拟数据

**实验 1：两个班级考试及格率比较**

假设有 200 个新生入学, 他们被随机分为两组, 分别由 A, B 两位老师教授同一门课, 学期结束后参加同一场考试, 记录成绩及格与否, 以下是结果。（结果通过随机数生成, 假设 A 老师有 90% 的通过率, B 老师有 80% 的通过率）

| | teacher A | teacher B | Total |
|---|---|---|---|
| Pass | 92 | 81 | 173 |
| Fail | 8 | 19 | 27 |
| Total | 100 | 100 | 200 |

原假设为学生及格率与教师无关, 即教师 A 与教师 B 的班级及格率相同。

$$H_0 : \mu_1 = \mu_2$$

$$H_1 : \mu_1 \neq \mu_2$$

利用 chi-square test, 得到 $\chi^2 = 4.28, p = .038 < .05$, 故拒绝原假设, 因此学生及格率与教师有关。

利用 Fisher exact test, 得到 $p = .037 < .05$, 故拒绝原假设, 因此学生及格率与教师有关。

可以看出, 在较大样本量的情况下, chi-square test 和 Fisher exact test, 得到了相同的结果。

#### 实验 2：性别与现在是否在学习的关系

假设有一组学生样本首先被分为男女两组, 并根据是否正在学习进行统计, 以下是结果。

| | Men | Women | Total |
|---|---|---|---|
| Studying | 1 | 8 | 9 |
| Not-studying | 11 | 4 | 15 |
| Total | 12 | 12 | 24 |

原假设为男生和女生同等可能的正在学习。

$$H_0 : \mu_1 = \mu_2$$
$$H_1 : \mu_1 < \mu_2$$

有一个单元期望值小于 5, 应该用 Fisher exact test, 得到 $p = .0009 < .05$, 故拒绝原假设, 因此男生学习的可能性较女生更低。

### 1.2 实际生物数据

#### 实验 3：154 例骨巨细胞瘤患者术后复发的比较[1]

骨巨细胞瘤 (giant cell tumor of bone, GTC) 是一种好发于长骨干骺端的原发性骨肿瘤, 占原发骨肿瘤的 3%∼5%, 占原发良性骨肿瘤的 15% 。目前 GTC 组织发生关系尚不明确, 组织学与临床过程间的关系亦不清楚, 且 20% ∼30% 患者有持续进展的潜在恶性, 组织学上没有恶变就可发生转移, 因此很难预测 GTC 患者的预后。大宗资料报道 GTC 术后复发率可达 4%∼50% 。本研究中, 回顾性分析 2004 年 1 月—2017 年 1 月西安交通大学医学院附属红会医院收治的 154 例四肢 GTC 患者的临床及随访资料, 探讨 GTC 的外科治疗方法及其临床疗效, 评估影响复发的因素及辅助治疗的必要性。

观察患者术后骨愈合情况, 并发症发生情况。定期复查的 X 线片, 观察病灶转移、复发情况。术后 6 个月根据骨骼肌肉系统肿瘤协会 (Musculoskeletal Tumor Society,MSTS) 制定术后重建功能评定标准评定临床疗效。

以下是不同性别的病人复发情况的比较。

| 影响因素 | 例数 | 复发 | 未复发 |
|---|---|---|---|
| 性别 | | | |
| 男 | 79 | 9 | 70 |
| 女 | 75 | 7 | 68 |

以 $p < .05$ 为差异有统计学意义。

利用 chi-square test, 得到 $\chi^2 = 0.024, p = .88 > .05$, 说明性别对术后复发率没有影响。

利用 Fisher exact test, 得到 $p = .79 > 0.5$, 说明性别对术后复发率没有影响。

类似的, 对这个项目其他的 categorical 数据进行统计分析, 得到以下结果。

| 影响因素 | 例数 | 复发 | 未复发 | Chi-square test | | Fisher exact test |
|---|---|---|---|---|---|---|
| | | | | $\chi^2$值 | $p$值 | $p$值 |
| 性别 | | | | | | |
| 男 | 79 | 9 | 70 | 0.176 | .68>.05 | .79>.05 |
| 女 | 75 | 7 | 68 | | | |
| 肿瘤部位 | | | | | | |
| 上肢 | 39 | 4 | 35 | ! 0.000 | 1>.05 | 1>.05 |
| 下肢 | 115 | 12 | 103 | | | |
| 手术方式 | | | | | | |
| A 组 | 64 | 7 | 57 | ! 0.000 | 1>.05 | 1>.05 |
| B 组 | 48 | 5 | 43 | | | |

[1] A 组为病灶刮除 + 植骨/骨水泥填充

[2] B 组为病灶刮除 + 辅助治疗 + 植骨/骨水泥填充

[3] ! 表示这一组列联表有至少一个单元期望值小于 5, 使用 chi-square test 有可能不准。

可以看出, 不同性别、肿瘤部位及手术方式的患者术后局部复发率比较, 差异均无统计学意义 ($p$值均 $> .05$)。

## 2 two sample dependent test

这一节将实现 McNemar test。对于 dependent 数据, 当 $b + c > 20$ 时, 其检验统计量为

$$\chi^2 = \frac{(|b - c| - 1)^2}{b + c} \tag{5}$$

## 2.1 模拟数据

例如, 研究人员正在测试一种新药物, 并记录该药物是否起作用（“是”）或不起作用(“否”)。建立一个表格, 上面列出了服药前后的人数。

| | | After | | |
|---|---|---|---|---|
| | | No | Yes | Total |
| Before | No | 80 | 100 | 180 |
| | Yes | 10 | 110 | 120 |
| | Total | 90 | 210 | 300 |

原假设为应用该药物先后没有区别。

$$H_0 : \mu_1 = \mu_2$$

$$H_1 : \mu_1 \neq \mu_2$$

利用 McNemar test, 得到 $\chi = 72.01, p \ll .05$, 故拒绝原假设, 因此该药物有作用。

## 2.2 实际生物数据

### 实验 4：探究多发性骨髓瘤患者巩固治疗的效果[2]

新诊断的多发性骨髓瘤患者自体造血干细胞移植后, 硼替佐米-沙利度胺-地塞米松在巩固治疗方面优于沙利度胺-地塞米松

2006 年 5 月至 2008 年 4 月期间, 在意大利马拉蒂 • 埃马多罗 • 米洛玛网络 73 个中心, 有 480 名患者参加了一项研究 (包括了本文实验在内的多个实验)。传统上, 对多发性骨髓瘤移植合格患者最重要的巩固疗法是高剂量的美法仑加自体干细胞移植 (autologous stem cell transplantation, ASCT)。在过去的几年中, 随着将新药加入 ASCT 巩固治疗方案, 适合移植后的多发性骨髓瘤患者的治疗范式不断发展。硼替佐米-沙利度胺-地塞米松 (VTD) 和沙利度胺-地塞米松 (TD) 被尝试纳入 ASCT 的治疗方案中。ASCT 前后达到完全缓解/完整响应 (complete response, CR) 或至少达到很好的部分缓解 (VGPR) 是长期临床结果最重要的预测指标之一。

本实验将对使用 VTD 或 TD 进行巩固治疗后升高或降低其反应状态的患者进行分析, 采用 CR 指标作为药物有效与否 (CR 有效, <CR 无效) 的判别标准。以下是实验结果。

| 巩固疗法之前的反应指标 | 巩固疗法之后的反应指标 | | | | | |
|---|---|---|---|---|---|---|
| | VTD | | | TD | | |
| | CR | <CR | Total | CR | <CR | Total |
| CR | 72 | 6 | 78 | 59 | 6 | 65 |
| <CR | 25 | 57 | 82 | 16 | 80 | 96 |
| Total | 97 | 63 | 160 | 75 | 86 | 161 |

对于 VTD 的疗效, 利用 McNemar test, 得到 $\chi^2 = 10.45, p \ll .05$, 拒绝原假设, 证实了 VTD 巩固疗法对 CR 发生率的有利影响。

对于 TD 的疗效, 利用 McNemar test, 得到 $\chi^2 = 3.68, p = .055 > .05$, 说明 TD 对巩固疗法的效果不显著。

# 3 Wilcoxon rank sum test

t-test 适用于比较两遵循正态分布的数据样本差异的方法。Wilcoxon rank sum test 与 t-test 相对应。当“总体正态”这一前提不成立时, 不能用 t-test, 可以用秩和检验法。

本文对样本量较大 ($n_1, n_2$>10) 的两独立样本的比较应用用 Wilcoxon rank sum test, 当 $n \to \infty$ 时, 秩和 $R$ 就趋向正态分布, 此时, 秩和 $R$ 的分布接近正态分布。若检验假设成立, 则两组的秩和不应相差太大。本文只考虑没有样本数据相同的情况。

具体方法是：

1. 建立假设

$$\begin{aligned} H_0 &: A = B \\ H_a &: A \neq B \end{aligned} \tag{6}$$

2. 将所有样本从小到大排序, 编秩并求秩和。

3. 求均值和标准差

$$\begin{aligned} \mu &= \frac{n_1 (n_1 + n_2 + 1)}{2} \\ \sigma &= \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}} \end{aligned} \tag{7}$$

其中$n_1$ 为较小的样本容量, 即$n_1 \leq n_2$

5. 求检验统计量。当 $n1 = n2$ 时, 取 $R$ 较小的一组。当 $R_1 \neq \frac{n_1(n_1+n_2+1)}{2}$ 时, 计算

$$T = \left[ \left| R_1 - \frac{n_1 (n_1 + n_2 + 1)}{2} \right| - \frac{1}{2} \right] / \sqrt{\left( \frac{n_1 n_2}{12} \right) (n_1 + n_2 + 1)} \tag{8}$$

当 $R_1 = \frac{n_1(n_1+n_2+1)}{2}$ 时, $T = 0$。

4. 根据 $T$ 求 $p$, 如果 $p \leq 0.05$, 则拒绝原假设, A 组与 B 组有显著区别；如果 $p > 0.05$, 则拒绝原假设, A 组与 B 组没有显著区别。

## 3.1 模拟数据

**实验 5：两组依正态分布的随机数的比较**

随机生成 A, B 两组数据, 方差均为 1, A 组均值为 1, B 组均值有变化, 从而构造出两组可以比较的样本。分别对它们进行 Wilcoxon rank sum test 和 t-test。

I：A 组均值为 10, B 组均值为 10

A[11.161 9.2 10.736 9.585 10.255 9.225 9.606 8.479 8.974 10.028]

B[ 8.835 11.46 9.43 9.799 8.523 10.281 10.587 9.383 9.083 12.072]

进行 Wilcoxon rank sum test, 得到 $p = .79 > .05$, 故接受原假设, 两组数据没有显著区别。

进行 t-test, 得到 $p = .52 > .05$, 故接受原假设, 两组数据没有显著区别。

II：A 组均值为 10, B 组均值为 11

A[10.829 10.594 10.136 10.348 9.64 9.661 11.139 11.211 9.316 10.706]

B[10.62 11.335 10.528 11.247 11.171 11.748 10.43 14.265 11.306 9.436]

进行 Wilcoxon rank sum test, 得到 $p = .0539 > .05$, 故接受原假设, 两组数据没有显著区别。

进行 t-test, 得到 $p = .00224 < .05$, 故拒绝原假设, 两组数据有显著区别。

III：A 组均值为 10, B 组均值为 12

A[ 9.728 10.094 9.99 10.765 9.533 11.196 10.159 9.308 11.519 10.758]

B[12.182 12.36 12.691 10.945 11.395 13.668 10.844 13.123 11.465 12.081]

进行 Wilcoxon rank sum test, 得到 $p = .001 < .05$, 故拒绝原假设, 两组数据有显著区别。

进行 t-test, 得到 $p \ll .05$, 故拒绝原假设, 两组数据有显著区别。

综合来看 t-test 所得 p 值相对 Wilcoxon rank sum test 较小, 从而有可能出现通过 Wilcoxon rank sum test 没有表现出区别的两组数据在 t-test 中有显著区别。

### 实验 6：乳腺癌易感基因 (BRCA) 和乳腺癌的关联

乳腺癌 1 型易感性蛋白是一种蛋白质, 在人类中被 BRCA1 编码的基因。BRCA1 是人类肿瘤抑制基因（也称为看门人基因）, 负责修复 DNA。

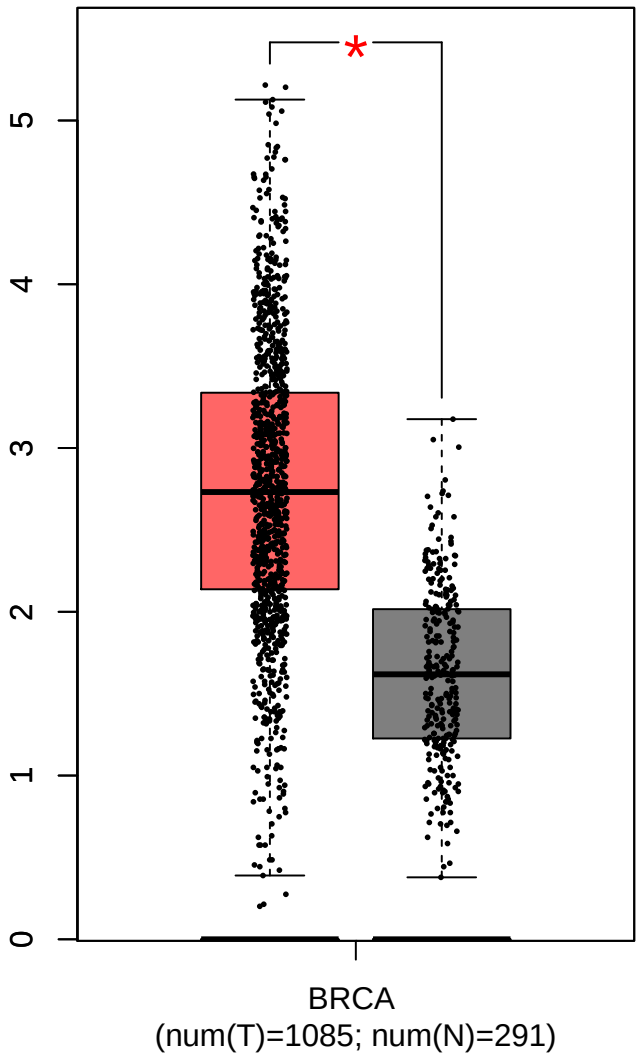
如果 BRCA1 或 BRCA2 本身被 BRCA 突变损坏, 受损的 DNA 没有得到正确修复, 这增加了患乳腺癌的风险。BRCA1 和 BRCA2 被描述为" 乳腺癌易感基因" 和" 乳腺癌易感蛋白"。主要等位基因具有正常的肿瘤抑制功能, 而这些基因的高渗透突变会导致肿瘤抑制功能丧失, 这与乳腺癌风险增加有关。[3]

简单来说, 在患有乳腺癌的病人体重能检出更高的 BRCA 表达。本文实验以罹患乳腺癌的病人与正常人体内的 BRCA1 表达量为例, 利用 Wilcoxon rank sum test 和 t-test 证明 BRCA1 与乳腺癌的关联性。

1. 下载实验数据[4, 5], 做 $\log_2(count + 1)$ 的数据处理。做出箱型图, 列出部分数据

Cancer[4.569 2.744 1.607 2.794 2.896 0.717 0.598 2.425 3.351 3.307 0.685 3.569]

Normal[ 1.078 1.821 2.559 2.545 0.135 1.948 0.764 0.716 1.155 0.802]




2. 原假设：BRCA1 表达量与乳腺癌无关备择假设：BRCA1 表达量与乳腺癌有关, 且在患者体内有更多表达。

$$\begin{aligned} H_0 &: C = N \\ H_a &: C > N \end{aligned} \tag{9}$$

3. 由于数据量巨大, 选取 200 个患者数据, 50 个健康人数据。进行 Wilcoxon rank sum test, $p = 7.176215e-11 \ll .05$, 则拒绝原假设, 患者与健康人体内 BRCA1 表达量有显著区别。

进行 t-test, $p = 3.432696e-08 \ll .05$, 则拒绝原假设, 患者与健康人体内 BRCA1 表达量有显著区别。

4. 结论, BRCA1 基因显著影响了乳腺癌的发病率。

# 参考文献


[1] 同志超, 周海振, 陈博, 等. 四肢骨巨细胞瘤的外科治疗分析 [J]. 中华解剖与临床杂志, 23(3): 234-239.

[2] Cavo M, Pantani L, Petrucci M T, et al. Bortezomib-thalidomide-dexamethasone is superior to thalidomide-dexamethasone as consolidation therapy after autologous hematopoietic stem cell transplantation in patients with newly diagnosed multiple myeloma[J]. Blood, 2012, 120(1): 9-19.

[3] https://en.wikipedia.org/wiki/BRCA1

[4] https://gdc-hub.s3.us-east-1.amazonaws.com/download/TCGA-BRCA.GDC_phenotype.tsv.gz

[5] https://gdc-hub.s3.us-east-1.amazonaws.com/download/TCGA-BRCA.htseq_counts.tsv.gz


# 项目地址

https://github.com/cimutianxin/two-sample-test-2021-4