生物统计基础 (P2)

统计 91 董晟渤, 2193510853

西安交通大学数学与统计学院

日期：2022 年 4 月

目录

Part I

任务 1: Wilcoxon 符号秩检验

1 方法: Wilcoxon 符号秩检验

1.1 检验问题

Wilcoxon 符号秩检验是用于验证成对数据是否来自同一分布的非参数检验方法.

- **相较于 t 检验**, Wilcoxon 符号秩检验作为一种非参数检验方法, 具有不需要假定总体的分布的优点;
- **相较于符号检验**, 同为非参数检验方法, Wilcoxon 符号秩检验除了考虑符号以外, 进一步考虑了差距的大小, 是符号检验的推广.

本篇报告中, 将叙述进行 Wilcoxon 符号秩检验的过程, 并且使用 MATLAB 实现该过程.

设样本量为 n, 已知成对样本 $(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$, 考虑原假设

$$H_0 : x_1, x_2, \cdots, x_n \text{ 和 } y_1, y_2, \cdots, y_n \text{ 来自同一个分布.}$$

记每对数据的差值 $d_i := x_i - y_i (1 \le i \le n)$, 其可以用来比较两组不同的数据. 在符号检验中, 考虑的是 d_i 的符号, 更准确地说, 检验使用的统计量 C 为使得 $d_i > 0$ 的 i 的个数. 这样的检验方法方便对比两组数据之间的差异, 然而不能精确衡量差异的大小. 为了改进符号检验的方法, Wilcoxon 提出了符号秩检验方法.

1.2 符号秩统计量

首先, 引入秩的概念和基本性质.

定义 1.1 (秩)**.** 设 $x_1, x_2, \cdots, x_n$ 是简单随机样本, $x_{(1)} \le x_{(2)} \le \cdots \le x_{(n)}$ 是次序统计量, 则 x_i 在次序统计量中的位置称为 x_i 的**秩**, 记作 R_i.

考虑到 $x_1, x_2, \cdots, x_n$ 是简单随机样本, 因此 $R_1, R_2, \cdots, R_n$ 是 $1, 2, \cdots, n$ 的一个随机排列, 从而对任意的 $1 \le i \le n$, R_i 是一个取值为 $1, 2, \cdots, n$ 的离散型均匀随机变量.

命题 1.1. *设 $x_1, x_2, \cdots, x_n$ 是简单随机样本, 对 $1 \le i \le n$, R_i 是 x_i 的秩.*

(1) 对任意的 $1 \le i \le n$, $\mathbb{E}(R_i) = \dfrac{n+1}{2}$;

(2) 对任意的 $1 \le i \le n$, $\mathrm{Var}(R_i) = \dfrac{n^2-1}{12}$;

(3) 对任意的 $1 \le i, j \le n$, $\mathrm{Cov}(R_i, R_j) = -\dfrac{n+1}{12}$.

接下来, 引入本篇报告中所需要使用的统计量. 设 $(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$ 是成对样本, 记 $d_i := x_i - y_i (1 \le i \le n)$. 我们将 $|d_1|, |d_2|, \cdots, |d_n|$ 从小到大进行排列, 考虑 $|d_i|$ 的秩 R_i. 此时 R_i 越大, $|d_i|$ 就越大. 直觉上, 在符号检验考虑 $d_i > 0$ 的个数的基础上, 我们可以再考虑 $|d_i|$ 的秩, 引入统计量

$$W^+ := \sum_{i=1}^{n} R_i \cdot I_{\{d_i > 0\}},$$

表示使得 $d_i > 0$ 的 d_i 所对应的 $|d_i|$ 的秩之和, 这相比于 C, 能够更精确地衡量两组数据之间的差异. W^+ 被称为**符号秩统计量**.

根据定义, W^+ 的取值为 $0, 1, \cdots, \dfrac{n(n+1)}{2}$. 当 $x_1, x_2, \cdots, x_n$ 和 $y_1, y_2, \cdots, y_n$ 来自同一个分布, 也即 H_0 成立时, 能够证明以下结论:

命题 1.2. 设样本量为 n, W^+ 为符号秩统计量.

(1) 对任意的 $0 \le d \le \dfrac{n(n+1)}{2}$, $\mathbb{P}(W^+ = d) = \dfrac{t(n,d)}{2^n}$, 其中 $t(n,d)$ 表示从 $1, 2, \cdots, n$ 中任取数, 使得它们的和为 d 的个数;

(2) $\mathbb{E}(W^+) = \dfrac{n(n+1)}{4}$, $\mathrm{Var}(W^+) = \dfrac{n(n+1)(2n+1)}{24}$.

同样假定 H_0 成立, 在得知了 W^+ 的分布、期望和方差之后, 可以建立关于 W^+ 的中心极限定理, 方便用于大样本情形下的近似检验.

定理 1.3. 当 $n \to \infty$ 时, 有

$$\frac{W^+ - \dfrac{n(n+1)}{4}}{\sqrt{\dfrac{n(n+1)(2n+1)}{24}}} \xrightarrow{L} \mathcal{N}(0,1).$$

相应地, 考虑 $d_i < 0$ 的情形, 我们可以引入符号秩统计量

$$W^- := \sum_{i=1}^{n} R_i \cdot I_{\{d_i < 0\}},$$

并且注意到 $W^+ + W^- = \dfrac{n(n+1)}{2}$. 当 H_0 成立时, 根据对称性, 上述命题和定理对 W^- 也是成立的.

1.3 符号秩检验

借助符号秩统计量, 我们可以解决最开始的检验问题. 设 $(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$ 是成对样本, 考虑原假设

$$H_0 : x_1, x_2, \cdots, x_n \text{ 和 } y_1, y_2, \cdots, y_n \text{ 来自同一个分布.}$$

记 $d_i := x_i - y_i (1 \leq i \leq n)$, 并记 $|d_i|$ 的秩为 R_i, 考虑符号秩统计量

$$W^+ := \sum_{i=1}^{n} R_i \cdot I_{\{d_i > 0\}}.$$

样本量较小时

若样本量 n 较小, 可以使用 W^+ 的精确分布. 设显著性水平为 α, W^+ 的精确分布的 $\dfrac{\alpha}{2}$ 分位数和 $1 - \dfrac{\alpha}{2}$ 分位数分别为 $W^+_{\frac{\alpha}{2}}$ 和 $W^+_{1-\frac{\alpha}{2}}$, 则检验的拒绝域为

$$\left\{W^+ \leq W^+_{\frac{\alpha}{2}}\right\} \cup \left\{W^+ \geq W^+_{1-\frac{\alpha}{2}}\right\};$$

设代入真实数据计算得到的符号秩统计量为 W^+_0, 则检验的 p 值

$$p = 2 \min \left\{ \mathbb{P}\left(W^+ \leq W^+_0\right), \mathbb{P}\left(W^+ \geq W^+_0\right), \frac{1}{2} \right\}.$$

样本量较大时

若样本量 n 较大, 方便起见, 可以使用 W^+ 的近似分布, 进行近似检验. 考虑统计量

$$t = \frac{\left| W^+ - \dfrac{n(n+1)}{4} \right| - \dfrac{1}{2}}{\sqrt{\dfrac{n(n+1)(2n+1)}{24}}},$$

设显著性水平为 α, 标准正态分布 $\mathcal{N}(0,1)$ 的 $\dfrac{\alpha}{2}$ 分位数和 $1 - \dfrac{\alpha}{2}$ 分位数分别为 $u_{\frac{\alpha}{2}}$ 和 $u_{1-\frac{\alpha}{2}}$, 并且注意到 $u_{1-\frac{\alpha}{2}} = -u_{\frac{\alpha}{2}}$, 因此检验的拒绝域为

$$\left\{t \geq u_{1-\frac{\alpha}{2}}\right\};$$

设代入真实数据计算得到的统计量为 t_0, $u \sim \mathcal{N}(0,1)$, 则检验的 p 值

$$p = \mathbb{P}\left(|u| \geq t_0\right) = 2 \cdot (1 - \Phi(t_0)).$$

通常, 当 $n \geq 20$ 时可采取近似检验的方法.

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 $n = 15$, $x_1, x_2, \cdots, x_n$ 来自正态总体 $\mathcal{N}(1, 1)$, $y_1, y_2, \cdots, y_n$ 来自正态总体 $\mathcal{N}(1.5, 1)$, 生成的数据如**图 1**和**表 1**所示.

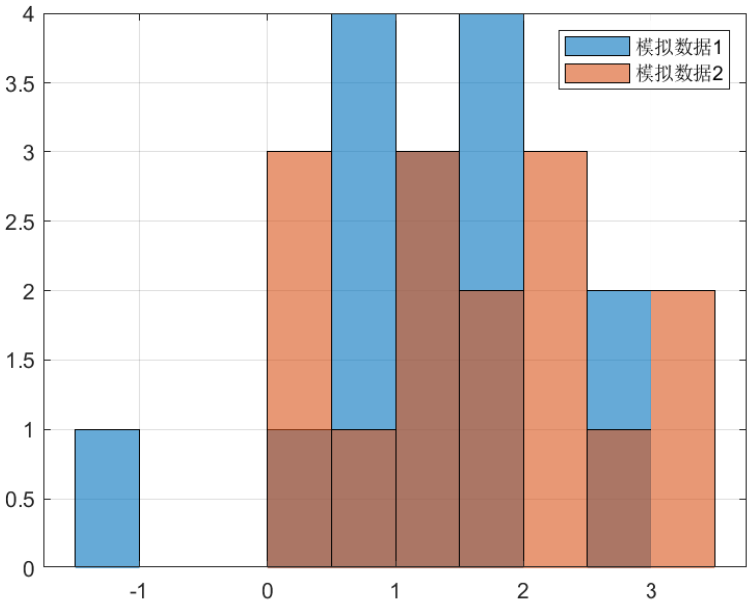

图 1: 成对模拟数据的直方图

表 1: 成对模拟数据

i	x_i	y_i	i	x_i	y_i	i	x_i	y_i
1	0.5070	0.1730	6	1.1093	2.3123	11	1.5265	3.0163
2	0.8193	0.0590	7	2.8140	2.0455	12	0.7397	1.4674
3	1.0458	1.9018	8	1.3120	0.4484	13	1.6001	3.1360
4	0.9362	2.9702	9	2.8045	1.8975	14	1.5939	1.0749
5	1.6113	1.1732	10	0.2769	0.7481	15	-1.1860	2.0894

2.2 检验的结果

记 $d_i : x_i - y_i (1 \le i \le n)$, 则 $d_i, |d_i|$ 和 R_i 的值如**表 2**所示.

表 2: 成对模拟数据

i	d_i	$\lvert d_i \rvert$	R_i	i	d_i	$\lvert d_i \rvert$	R_i	i	d_i	$\lvert d_i \rvert$	R_i
1	0.3340	0.3340	1	6	-1.2030	1.2030	11	11	-1.4897	1.4897	12
2	0.7603	0.7603	6	7	0.7685	0.7685	7	12	-0.7277	0.7277	5
3	-0.8560	0.8560	8	8	0.8637	0.8637	9	13	-1.5359	1.5359	13
4	-2.0340	2.0340	14	9	0.9070	0.9070	10	14	0.5190	0.5190	4
5	0.4381	0.4381	2	10	-0.4712	0.4712	3	15	-3.2755	3.2755	15

由**表 2**得知, 满足 $d_i > 0$ 的有 $d_1, d_2, d_5, d_7, d_8, d_9, d_{14}$, 对应的秩分别为 $1, 6, 2, 7, 9, 10, 4$, 将数据代入统计量, 计算得到

$$W_0^+ = 1 + 6 + 2 + 7 + 9 + 10 + 4 = 39.$$

若进行精确检验, 计算得 $\dfrac{n(n+1)}{4} = 60 > 39 = W_0^+$, 因此检验的 p 值

$$p = 2 \cdot \mathbb{P}(W^+ \le W_0^+) = 0.2523;$$

若进行近似检验, 计算得此时 $t_0 = 1.1643$, 因此检验的 p 值

$$p = 2 \cdot (1 - \Phi(t_0)) = 0.2443.$$

两种检验的 p 值是接近的. 若取显著性水平 $\alpha = 0.05$, 则接受原假设, 也即认为 x_i 和 y_i 来自同一分布. 根据**图 1**, 发现此时数据 $x_1, x_2, \cdots, x_n$ 和 $y_1, y_2, \cdots, y_n$ 较为接近, 难以区分.

2.3 重复模拟的结果

为了验证检验方法, 修改样本量和总体的参数, 每次进行 5 次模拟.

表 3: 重复模拟的结果, 样本量 $n = 15$, 总体分别为 $\mathcal{N}(1,1)$ 和 $\mathcal{N}(1.5,1)$

模拟次数	1	2	3	4	5
精确检验的 p 值	0.2523	0.0180	0.2292	0.0553	0.1514
近似检验的 p 值	0.2443	0.0214	0.2220	0.0571	0.1475

表 3列出了样本量 $n=15$, 总体分别为 $\mathcal{N}(1,1)$ 和 $\mathcal{N}(1.5,1)$ 的重复模拟结果. 此时 p 值的变化较大, 受到随机性的影响明显, 有可能接受原假设, 也有可能拒绝原假设.

表 4: 重复模拟的结果, 样本量 $n=15$, 总体分别为 $\mathcal{N}(1,1)$ 和 $\mathcal{N}(2,1)$

模拟次数	1	2	3	4	5
精确检验的 p 值	0.0053	0.0067	0.0053	0.0301	0.0412
近似检验的 p 值	0.0083	0.0098	0.0083	0.0332	0.0438

表 4列出了样本量 $n=15$, 总体分别为 $\mathcal{N}(1,1)$ 和 $\mathcal{N}(2,1)$ 的重复模拟结果. 这时候 p 值稳定地小于 0.05, 该检验方法能够很好地得出结论.

表 5: 重复模拟的结果, 样本量 $n=50$, 总体分别为 $\mathcal{N}(1,1)$ 和 $\mathcal{N}(1.5,1)$

模拟次数	1	2	3	4	5
近似检验的 p 值	0.2711	0.1346	0.0026	0.5888	0.1296

表 6: 重复模拟的结果, 样本量 $n=50$, 总体分别为 $\mathcal{N}(1,1)$ 和 $\mathcal{N}(2,1)$

模拟次数	1	2	3	4	5
近似检验的 p 值	8.1980e-5	1.3993e-5	1.0886e-6	0.0171	1.2813e-05

对应于**表 3**和**表 4**, **表 5**和**表 6**列出了将样本量设置为 $n=50$ 时重复模拟的结果. 考虑到样本量较大, 精确检验情形的计算较为复杂, 因此计算 p 值时只使用了近似检验的方法. 当总体分别为 $\mathcal{N}(1,1)$ 和 $\mathcal{N}(1.5,1)$ 时, p 值受到随机性的影响仍然较大; 而当总体分别为 $\mathcal{N}(1,1)$ 和 $\mathcal{N}(2,1)$ 时, 计算得到的 p 值极小, 从而可以做出接受原假设的决定.

3 应用 2: 真实数据

3.1 真实数据的选取

选取的真实数据来自课本. 为了对比母乳喂养和奶粉喂养婴儿的中耳积液持续时间, 实验数据如**表 7**所示.

表 7: 成对真实数据

序号	母乳喂养	奶粉喂养	序号	母乳喂养	奶粉喂养
1	20	18	11	17	186
2	11	35	12	12	29
3	3	7	13	52	39
4	24	182	14	14	15
5	7	6	15	12	21
6	28	33	16	30	28
7	58	223	17	7	8
8	7	7	18	15	27
9	39	57	19	65	77
10	17	76	20	10	12

此时, $n = 10$, 考虑到样本并不一定是正态分布总体, t 检验并不是有效的. 在这里, 我们使用非参数检验的方法来解决该问题.

3.2 检验的结果

记 $d_i : x_i - y_i (1 \leq i \leq n)$, 则 $d_i, |d_i|$ 和 R_i 的值如**表 8**所示.

表 8: 成对真实数据

i	d_i	$\lvert d_i \rvert$	R_i	i	d_i	$\lvert d_i \rvert$	R_i
1	2	2	7	11	-169	169	20
2	-24	24	16	12	-17	17	14
3	-4	4	8	13	13	13	13

4	-158	158	18	14	-1	1	3
5	1	1	4	15	-9	9	10
6	-5	5	9	16	2	2	6
7	-165	165	19	17	-1	1	2
8	0	0	1	18	-12	12	12
9	-18	18	15	19	-12	12	11
10	-59	59	17	20	-2	2	5

由**表 8**得知, 满足 $d_i > 0$ 的有 d_1, d_5, d_{13}, d_{16}, 对应的秩分别为 7, 4, 13, 6, 将数据代入统计量, 计算得到

$$W_0^+ = 7 + 4 + 13 + 6 = 30.$$

考虑到 $n = 20$, 我们可以分别求出精确检验和近似检验的 p 值, 并进行对比.

- 对于精确检验, p 值为 0.0037;
- 对于近似检验, p 值为 0.0054.

注意到精确检验和近似检验的 p 值都极小, 从而拒绝原假设, 也即认为这母乳喂养和奶粉喂养对婴儿的中耳积液持续时间的影响有显著的差异.

Part II

任务 2: Wilcoxon 秩和检验

4 方法: Wilcoxon 秩和检验

4.1 检验问题

Wilcoxon 秩和检验是另外一种用于验证数据是否来自同一分布的非参数检验方法. 相较于符号秩检验, Wilcoxon 秩和检验可以用于处理非成对数据的检验, 拓宽了研究的对象. 本篇报告剩下的部分, 将介绍进行 Wilcoxon 秩和检验的过程, 并且使用 MATLAB 实现该过程.

设两组样本的样本量分别为 m 和 n, 且 $m < n$, 并设已知第一组样本为 $x_1, x_2, \cdots, x_m$, 第二组样本为 $y_1, y_2, \cdots, y_n$, 考虑原假设

$$H_0 : x_1, x_2, \cdots, x_m \text{ 和 } y_1, y_2, \cdots, y_n \text{ 来自同一个分布.}$$

此时样本量并不相同, 无法对这两组样本进行作差, 需要另外寻找合适的统计量来进行检验. 为了解决该问题, Wilcoxon 进一步提出了秩和检验的方法.

4.2 秩和统计量

在上一部分中, 已经介绍了秩的概念和符号秩统计量. 在这里, 统计量的构造仍然需要使用到秩. 将已知的样本放在一起, 记作 $x_1, x_2, \cdots, x_m, y_1, y_2, \cdots, y_n$. 若 H_0 成立, 则上述样本是容量为 $m+n$ 的简单随机样本. 对于 $1 \le i \le m$, 记 x_i 在上述样本中的秩为 R_i; 对于 $1 \le j \le n$, 记 y_j 在上述样本中的秩为 R_{m+j}, 考虑 $x_1, x_2, \cdots, x_m$ 的秩的和, 也即

$$W := \sum_{i=1}^{m} R_i.$$

若 H_0 成立, 则 W 不会太大, 也不会太小, 因此其可以作为用于检验的统计量, 并称为**秩和统计量**.

根据定义, W 的取值为 $\frac{m(m+1)}{2}, \frac{m(m+1)}{2}+1, \cdots, \frac{m(m+1)}{2}+mn$. 当 H_0 成立时, 根据命题1.1, 能够证明以下结论:

命题 4.1. *设样本量分别为 m 和 n, W 为秩和统计量.*

(1) 对任意的 $\dfrac{m(m+1)}{2} \le d \le \dfrac{m(m+1)}{2} + mn$, $\mathbb{P}(W^+ = d) = \dfrac{m! \cdot n! \cdot s(m,n,d)}{(m+n)!}$, 其中 $s(m,n,d)$ 表示从 $1, 2, \cdots, m+n$ 中任取 m 个数, 使得它们的和为 d 的个数;

(2) $\mathbb{E}(W) = \dfrac{m(m+n+1)}{2}$, $\mathrm{Var}(W) = \dfrac{mn(m+n+1)}{12}$.

同样假定 H_0 成立, 类似符号秩检验, 可以建立 W 的中心极限定理:

定理 4.2. 当 $n \to \infty$ 时, 有

$$\frac{W - \dfrac{m(m+n+1)}{2}}{\sqrt{\dfrac{mn(m+n+1)}{12}}} \xrightarrow{L} \mathcal{N}(0,1).$$

另外, 考虑 $y_1, y_2, \cdots, y_n$ 的秩的和, 也即

$$W' := \sum_{j=1}^{n} R_{m+j},$$

则有 $W + W' = \dfrac{(m+n)(m+n+1)}{2}$, 可以类似地给出 W' 相关的结论.

4.3 秩和检验

基于秩和统计量, 我们可以解决所提出的检验问题. 设两组样本的样本量分别为 m 和 n, 且 $m < n$, 并设第一组样本为 $x_1, x_2, \cdots, x_m$, 第二组样本为 $y_1, y_2, \cdots, y_n$, 考虑原假设

$$H_0 : x_1, x_2, \cdots, x_m \text{ 和 } y_1, y_2, \cdots, y_n \text{ 来自同一个分布.}$$

记 x_i 在 $x_1, x_2, \cdots, x_m, y_1, y_2, \cdots, y_n$ 中的秩为 R_i, 考虑统计量

$$W := \sum_{i=1}^{m} R_i.$$

样本量较小时

若样本量 m 或 n 较小, 可以用 W 的精确分布. 设显著性水平为 α, W 的精确分布的 $\dfrac{\alpha}{2}$ 分位数和 $1 - \dfrac{\alpha}{2}$ 分位数分别为 $W_{\frac{\alpha}{2}}$ 和 $W_{1-\frac{\alpha}{2}}$, 则检验的拒绝域为

$$\left\{W \le W_{\frac{\alpha}{2}}\right\} \cup \left\{W \ge W_{1-\frac{\alpha}{2}}\right\};$$

设代入真实数据计算得到的秩和统计量为 W_0, 则检验的 p 值

$$p = 2\min\left\{\mathbb{P}\left(W \le W_0\right), \mathbb{P}\left(W \ge W_0\right), \frac{1}{2}\right\}.$$

样本量较大时

若样本量 m 和 n 较大, 方便起见, 可以用 W 的近似分布, 进行近似 u 检验. 考虑统计量

$$t=\frac{\left|W-\frac{m(m+n+1)}{2}\right|-\frac{1}{2}}{\sqrt{\frac{mn(m+n+1)}{12}}}$$

设显著性水平为 α, 标准正态分布 $\mathcal{N}(0,1)$ 的 $\frac{\alpha}{2}$ 分位数和 $1-\frac{\alpha}{2}$ 分位数分别为 $u_{\frac{\alpha}{2}}$ 和 $u_{1-\frac{\alpha}{2}}$, 并且注意到 $u_{1-\frac{\alpha}{2}}=-u_{\frac{\alpha}{2}}$, 因此检验的拒绝域为

$$\left\{t \geq u_{1-\frac{\alpha}{2}}\right\};$$

设代入真实数据计算得到的统计量为 t_0, $u \sim \mathcal{N}(0,1)$, 则检验的 p 值

$$p=\mathbb{P}\left(|u| \geq t_0\right)=2 \cdot\left(1-\Phi\left(t_0\right)\right).$$

通常, 当 $m, n \geq 10$ 时可采取近似检验的方法.

5 应用 1: 模拟数据

5.1 模拟数据的生成

生成模拟数据时, 设 $m = 5, x_1, x_2, \cdots, x_m$ 来自正态总体 $\mathcal{N}(1, 1)$; $n = 10, y_1, y_2, \cdots, y_n$ 来自正态总体 $\mathcal{N}(1.5, 1)$, 生成的数据如**图 2**和**表 9**所示.

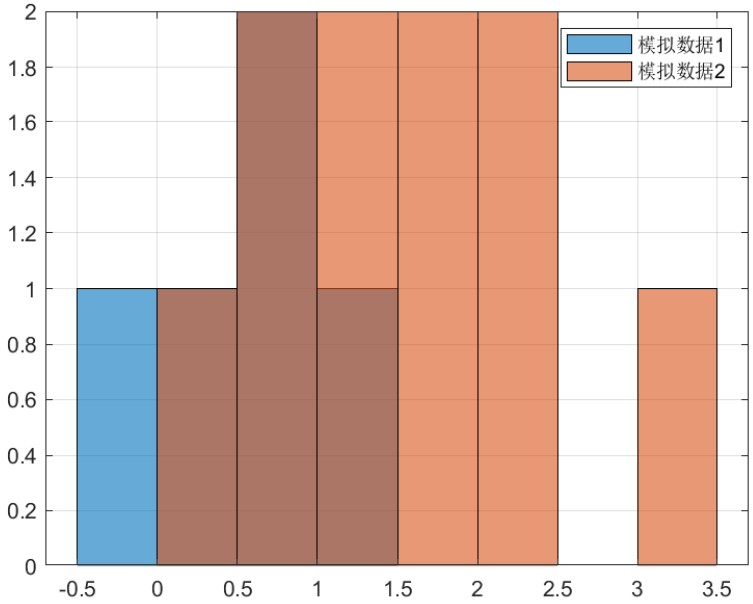

图 2: 非成对模拟数据的直方图

表 9: 非成对模拟数据

i	x_i	j	y_j	j	y_j
1	0.5019	1	1.3157	6	2.3706
2	-0.4214	2	1.9020	7	1.8308
3	0.7292	3	2.0392	8	0.1521
4	1.4397	4	0.7664	9	3.0479
5	0.4939	5	1.2316	10	0.8834

5.2 检验的结果

记 x_i 在 $x_1, x_2, \cdots, x_m, y_1, y_2, \cdots, y_n$ 中的秩为 R_i, y_j 在 $x_1, x_2, \cdots, x_m, y_1, y_2, \cdots, y_n$ 中的秩为 R_{m+i}, 则 R_i 的值如**表 10**所示.

表 10: 非成对模拟数据

i	x_i	R_i	j	y_j	R_{m+j}	j	y_j	R_{m+j}
1	0.5019	4	1	1.3157	9	6	2.3706	14
2	-0.4214	1	2	1.9020	12	7	1.8308	11
3	0.7292	5	3	2.0392	13	8	0.1521	2
4	1.4397	10	4	0.7664	6	9	3.0479	15
5	0.4939	3	5	1.2316	8	10	0.8834	7

由**表 10**, 将 x_i 的秩加在一起, 得

$$W_0 = 4 + 1 + 5 + 10 + 3 = 23.$$

若进行精确检验, 计算得 $\dfrac{m(m+n+1)}{2} = 40 > 23 = W_0$, 因此检验的 p 值

$$p = 2 \cdot \mathbb{P}(W \leq W_0) = 0.0400;$$

若进行近似检验, 计算得此时 $t_0 = 2.0208$, 因此检验的 p 值

$$p = 2 \cdot (1 - \Phi(t_0)) = 0.0433.$$

两种检验的 p 值是接近的. 若取显著性水平 $\alpha = 0.05$, 则拒绝原假设, 也即认为 x_i 和 y_i 来自不同分布, 这和**图 2**的结果是相符的.

5.3 重复模拟的结果

为了验证检验方法, 同样修改样本量和总体的参数, 每次进行 5 次模拟.

表 11: 重复模拟的结果, 样本量分别为 $m = 5$ 和 $n = 10$, 总体分别为 $\mathcal{N}(1, 1)$ 和 $\mathcal{N}(1.5, 1)$

模拟次数	1	2	3	4	5
精确检验的 p 值	0.0400	0.7679	0.5135	0.0280	0.8591
近似检验的 p 值	0.0433	0.7595	0.5006	0.0321	0.8542

表 11列出了样本量分别为 $m=5$ 和 $n=10$, 总体分别为 $\mathcal{N}(1,1)$ 和 $\mathcal{N}(1.5,1)$ 的重复模拟结果. 此时 p 值的变化较大, 受到随机性的影响明显, 有一定的可能接受原假设, 但是在几次模拟中, p 值会非常大.

表 12: 重复模拟的结果, 样本量分别为 $m=5$ 和 $n=10$, 总体分别为 $\mathcal{N}(1,1)$ 和 $\mathcal{N}(3,1)$

模拟次数	1	2	3	4	5
精确检验的 p 值	0.0027	0.0127	0.0047	0.0013	0.0013
近似检验的 p 值	0.0059	0.0169	0.0085	0.0040	0.0040

表 12列出了样本量分别为 $m=5$ 和 $n=10$, 总体分别为 $\mathcal{N}(1,1)$ 和 $\mathcal{N}(3,1)$ 的重复模拟结果. 此时 p 值较小, 每次模拟都能做出拒绝原假设的决定.

表 13: 重复模拟的结果, 样本量分别为 $m=30$ 和 $n=50$, 总体分别为 $\mathcal{N}(1,1)$ 和 $\mathcal{N}(1.5,1)$

模拟次数	1	2	3	4	5
近似检验的 p 值	0.0022	0.0014	0.0411	0.0532	0.0667

表 14: 重复模拟的结果, 样本量分别为 $m=30$ 和 $n=50$, 总体分别为 $\mathcal{N}(1,1)$ 和 $\mathcal{N}(3,1)$

模拟次数	1	2	3	4	5
近似检验的 p 值	4.3376e-07	4.2087e-10	9.5102e-09	1.0849e-07	2.5227e-10

对应于**表 11**和**表 12**, **表 13**和**表 14**列出了将样本量设置为 $m=30$ 和 $n=50$ 时重复模拟的结果. 考虑到样本量较大, 精确检验情形的计算较为复杂, 因此计算 p 值时只使用了近似检验的方法. 当总体分别为 $\mathcal{N}(1,1)$ 和 $\mathcal{N}(1.5,1)$ 时, p 值受到随机性的影响已经很小了, 基本上不会超过 0.1; 而当总体分别为 $\mathcal{N}(1,1)$ 和 $\mathcal{N}(3,1)$ 时, 计算得到的 p 值极小, 从而可以做出接受原假设的决定.

6 应用 2: 真实数据

6.1 真实数据的选取

选取的真实数据来自课本. 假设我们想比较两个不同医院相同疾病的患者的住院时间, 结果如下:

- 第一个医院, 住院时间分别为 21, 10, 32, 60, 8, 44, 29, 5, 13, 26, 33;
- 第二个医院, 住院时间分别为 86, 27, 10, 68, 87, 76, 125, 60, 35, 73, 96, 44, 238.

此时, $m = 11$, $n = 13$, 考虑到样本并不一定是正态分布总体, t 检验并不是有效的. 在这里, 我们使用非参数检验的方法来解决该问题.

6.2 检验的结果

记 x_i 在 $x_1, x_2, \cdots, x_m, y_1, y_2, \cdots, y_n$ 中的秩为 R_i, y_j 在 $x_1, x_2, \cdots, x_m, y_1, y_2, \cdots, y_n$ 中的秩为 R_{m+i}, 则 R_i 的值如**表 15**所示.

表 15: 非成对真实数据

i	x_i	R_i	j	y_j	R_{m+j}	j	y_j	R_{m+j}
1	21	6	1	86	20	12	44	13
2	10	4	2	27	8	13	238	24
3	32	10	3	10	3			
4	60	16	4	68	17			
5	8	2	5	87	21			
6	44	14	6	76	19			
7	29	9	7	125	23			
8	5	1	8	60	15			
9	13	5	9	35	12			
10	26	7	10	73	18			
11	33	11	11	96	22			

计算得

$$W_0 = 6 + 4 + 10 + 16 + 2 + 14 + 9 + 1 + 5 + 7 + 11 = 85.$$

考虑到 $m, n > 10$, 但是样本量并不大, 我们可以分别求出精确检验和近似检验的 p 值, 并进行对比.

- 对于精确检验, p 值为 0.0015;
- 对于近似检验, p 值为 0.0026.

注意到精确检验和近似检验的 p 值都极小, 从而拒绝原假设, 也即认为这两个医院的住院时间有显著差异.

附录

A 代码 1: 符号秩统计量的分布

以下代码存储在wppdf.m 中.

```
function p = wppdf(n, d)
  count = 0;
  for i = 1 : n
    count = count + size(find(sum(nchoosek(1 : n, i), 2) == d), 1);
  end
  p = count / (2 ^ n);
end
```

B 代码 2: Wilcoxon 符号秩检验

以下代码存储在main.m 中, 与wppdf.m 同目录.

```
%% 初始化

clc;
clear;
close;

%% 数据

% load data_1.mat;
x = normrnd(1, 1, 1, 15); % 模拟数据1
y = normrnd(1.5, 1, 1, 15); % 模拟数据2
d = x - y; % 作差
n = size(x, 2);

%% 绝对值的秩

e = abs(d);
r = zeros(1, n);
for i = 1 : n
```

```
  j = find(e == max(e), 1);
  e(j) = -1;
  r(j) = n - i + 1;
end

%% 符号秩检验

W = r * (d > 0)'; % W统计量
t = (abs(W - (n * (n + 1)) / 4) - 1/2) / sqrt((n * (n + 1) * (2 * n + 1)
    ) / 24); % t统计量
p1 = 2 * (1 - normcdf(t)); % 近似检验的p值

if n <= 20
  p2 = 0;
  if W >= n * (n + 1) / 4
    for i = W : n * (n + 1) / 2
      p2 = p2 + wppdf(n, i);
    end
  else
    for i = 0 : W
      p2 = p2 + wppdf(n, i);
    end
  end
  p2 = 2 * p2; % 精确检验的p值
end
```

C 代码 3: 秩和统计量的分布

以下代码存储在wpdf.m 中.

```
function p = wpdf(m, n, d)
  p = size(find(sum(nchoosek(1 : m + n, m), 2) == d), 1) / combntns(m +
      n, m);
end
```

D 代码 4: Wilcoxon 秩和检验

以下代码存储在`main.m` 中, 与`wpdf.m` 同目录.

```
%% 初始化

clc;
clear;
close;

%% 数据

% load data_2.mat;
x = normrnd(1, 1, 1, 5); % 模拟数据1
y = normrnd(1.5, 1, 1, 10); % 模拟数据2
m = size(x, 2);
n = size(y, 2);

%% 样本的秩

z = [x, y];
r = zeros(1, m + n);
for i = 1 : m + n
  j = find(z == max(z));
  z(j) = -1;
  r(j) = m + n - i + 1;
end

%% 秩和检验

W = sum(r(:, 1 : m)); % W统计量
t = (abs(W - (m * (m + n + 1)) / 2) - 1/2) / sqrt((m * n * (m + n + 1))
    / 12); % t统计量
p1 = 2 * (1 - normcdf(t)); % 近似检验的p值

if m <= 10 || n <= 10
  p2 = 0;
```

```
  if W >= m * (m + n + 1) / 2
    for i = W : m * (m + 1) / 2 + m * n
      p2 = p2 + wpdf(m, n, i);
    end
  else
    for i = 0 : W
      p2 = p2 + wpdf(m, n, i);
    end
  end
  p2 = 2 * p2; % 精确检验的p值
end
```