



非参数统计分析

作者：Lollins

时间：December 14, 2023



改变人生的事情，你必须冒险；意义非凡的事情，大多碰巧发生；不重要的事，才有周全的计划。

前言

非参数统计分析笔记，一些图片的代码在 `code` 文件夹下。

Lollins

December 14, 2023

目录

第 1 章 绪论	1
1.1 序	1
1.1.1 非参数统计概念及学习意义	1
1.1.2 非参数统计的历史及发展	1
1.2 引言	1
1.2.1 参数统计方法与非参数统计方法的区别	1
1.2.2 非参数统计方法的特点	1
第 2 章 描述性统计	2
2.1 图表法	2
2.2 数值方法	2
2.2.1 表示中心位置的数值	2
2.2.2 表示离散程度的数值	3
2.2.3 标准误	3
2.2.4 偏度	3
2.2.5 峰度	3
第 3 章 符号检验法	5
3.1 符号检验	5
3.1.1 具体操作方法	5
3.1.2 注意事项	5
3.1.3 中位数的估计	5
3.2 符号检验在定性数据分析中的应用	6
3.3 成对数据的比较问题	6
第 4 章 符号秩和检验法	7
4.1 对称中心为原点的检验问题	7
4.1.1 符号秩和检验统计量 W^+	7
4.1.2 符号秩和检验	7
4.2 符号秩和检验统计量 W^+ 的性质	8
4.2.1 概率分布	8
4.2.2 W^+ 分布的对称性	8
4.3 符号秩和检验统计量 W^+ 的渐进正态性	8
4.3.1 期望与方差	8
4.3.2 W^+ 渐进正态性	9
4.4 平均秩法	9
4.4.1 定义	9
4.4.2 性质	9
4.5 对称中心的点估计	9
第 5 章 两样本问题	11
5.1 Mood 中位数检验法 (2×2 列联表检验法)	11
5.1.1 Mood 中位数检验法	11

5.1.2	大样本情形	11
5.2	Wilcoxon 秩和检验法	11
5.2.1	秩	11
5.2.2	Wilcoxon 秩和检验统计量的性质	12
5.2.3	Wilcoxon 秩和检验的备择假设	13
5.2.4	Wilcoxon 秩和检验的平均秩	14
5.2.5	位置参数差的检验与估计	14
5.3	Mann-Whitney U 检验	14
5.3.1	U 统计量	14
5.3.2	Mann-Whitney U 统计量 (W_{xy}) 和 Wilcoxon 秩和检验统计量 (W_y)	15
5.3.3	Mann-Whitney U 统计量的性质	16
5.4	两样本尺度参数的秩检验方法	16
5.4.1	尺度参数	16
5.4.2	尺度参数的检验问题	17
第 6 章	多样本问题	18
6.1	Kruskal-Wallis 检验法	18
6.1.1	Kruskal-Wallis 检验	18

第1章 绪论

1.1 序

1.1.1 非参数统计概念及学习意义

1.1.1.1 意义

1.1.1.2 概念

- **参数统计方法**：数据样本被视为从分布族的某个参数族抽取出来的总体的代表，未知的仅仅是总体分布具体数值，这样推断问题就转化为分布族的若干未知参数的估计问题，用样本来对这些参数进行估计或进行假设检验，从而得知背后的分布，这类推断方法称为参数统计方法。
- **非参数统计方法**：不假定总体分布的具体形式，尽量从数据（或样本）本身获得所需要的信息，通过估计而获得分布的结构，并逐步建立对事物的数学描述和统计模型的方法。

1.1.2 非参数统计的历史及发展

1.2 引言

1.2.1 参数统计方法与非参数统计方法的区别

- **参数统计方法**：假定总体的分布形式，既利用样本的数据信息，又利用产生数据总体的信息，是一个有效的数据分析方法，针对性强，但可能出现大的错误。
- **非参数统计方法**：不假定总体的分布形式，更接近大多数实际情况，故不会出现大的错误。

1.2.2 非参数统计方法的特点

- (1) 有广泛的适用性（广）
- (2) 样本方法是非参数统计的基本方法（样本）
- (3) 计算简单（简）
- (4) 良好的稳定性（稳）

第2章 描述性统计

定义 2.1 (描述性统计)

是在对产生数据的总体的分布不作任何假设的情况下，整理数据、显示数据、分析数据，将数据中有用的信息提取出来的统计方法。本章介绍常用的描述性统计方法：表格法、图形法和数值方法。



2.1 图表法

表格法、图形法描述统计数据主要是频数（率）分布表和直方图。

2.2 数值方法

数值方法主要是用数值来表示数据的中心位置和离散程度等的方法。

2.2.1 表示中心位置的数值

我们要求数据的中心位置满足这样一个**条件**：它到各个数据点的距离的和比较小。表示中心位置的数值有平均数、中位数、众数、切尾平均数。

2.2.1.1 平均数

如果用平方值距离法，则点 a 到各数据点 x_1, x_2, \dots, x_n 的距离的和可以用 $\sum_{i=1}^n (x_i - a)^2$ 来衡量。平均数 $\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$ 满足条件：

$$\sum_{i=1}^n (x_i - \bar{x})^2 = \min_a \sum_{i=1}^n (x_i - a)^2 \quad (2.1)$$

上式表示平均数这一点到各个数据点的平方值距离和最短。所以在平方值距离方法下，数据中心位置的代表是平均数。

2.2.1.2 中位数

如果用绝对值距离法，则点 a 到各数据点 x_1, x_2, \dots, x_n 的距离的和可以用 $\sum_{i=1}^n |x_i - a|$ 来衡量，中位数 me 满足条件：

$$\sum_{i=1}^n |x_i - me| = \min_a \sum_{i=1}^n |x_i - a| \quad (2.2)$$

上式表示中位数这一点到各个数据点的绝对值距离和最短。所以在绝对值距离方法下，数据中心位置的代表是中位数。

注：

- 中位数是非线性规划选址问题的解；
- 中位数不受极大（小）的影响，有时能较好地表示数据的中心位置。

2.2.1.3 众数

众数：一组数据中出现频数最高的数据。

注：

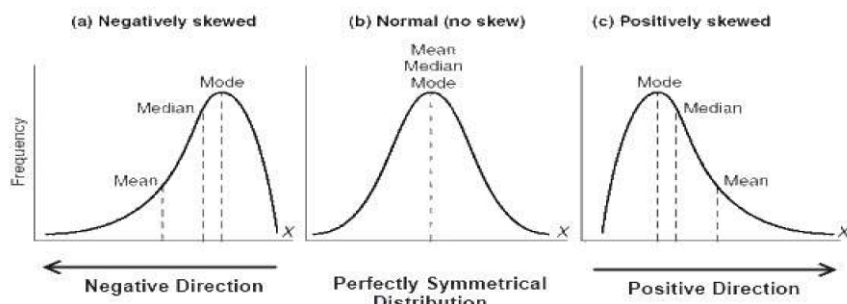


图 2.1

- 众数也能描述数据的中心位置。特别是定性数据；
- 一组数据有偏时，若数据右偏 (Positively Skewed)，通常有 $mo < me < \bar{x}$ ，若数据左偏 (Negatively Skewed)，通常有 $\bar{x} < me < mo$ ，见图2.1。

2.2.1.4 切尾平均数

设 $X_{(1)}, \dots, X_{(n)}$ 是来自总体 X 的简单随机样本 X_1, \dots, X_n 的次序统计值，称

$$T_{nk} = \frac{1}{n - 2k} (x_{(k+1)} + \dots + x_{(n-k)}) \quad (2.3)$$

为原样本的切尾均值。

2.2.2 表示离散程度的数值

样本方差、标准差、全距（范围）、四分位数间距。

2.2.3 标准误

$$se = \frac{s}{\sqrt{n}}, s \text{ 为样本标准差} \quad (2.4)$$

2.2.4 偏度

偏度反映单峰分布对称性，常用 β_s 表示总体偏度，

$$\beta_s = E\left[\left(\frac{x - \mu}{\sigma}\right)^3\right] = \frac{\mu_3}{\sigma^3}, \text{ 其中 } \mu_3 = E(x - \mu)^3 \quad (2.5)$$

注：对称分布的偏度 $\beta_s = 0$ ；反之不成立，即 $\beta_s = 0$ ，不一定是对称分布。

样本偏度用 b_s 表示，

$$b_s = \frac{m_3}{m_2^{3/2}}, \text{ 其中 } m_j = \frac{1}{n} \sum_i (x_i - \bar{x})^j \quad (2.6)$$

注： $b_s > 0$ 时，倾向于认为数据分布右偏； $b_s < 0$ 时，倾向于认为数据分布左偏； $b_s \approx 0$ 时，倾向认为数据分布是对称的。

2.2.5 峰度

峰度反映分布峰的尖峭程度，常用 β_k 表示总体峰度。

$$\beta_k = E\left[\left(\frac{x - \mu}{\sigma}\right)^4\right] = \frac{\mu_4}{\sigma^4} \quad (2.7)$$

注：若 $X \sim N(\mu, \sigma^2)$ ，则 $\beta_k = 3$ 。当 $\beta_k > 3$ 时，该分布具有过度的峰度 (厚尾分布)，当 $\beta_k < 3$ 时，该分布具有不足的峰度 (薄尾分布)，

样本峰度用 b_k 表示,

$$b_k = \frac{m_4}{(m_2)^2} \quad (2.8)$$

第3章 符号检验法

在非参数检验中，总体的中心位置的数通常用中位数表示，本章主要讨论中位数、p分位数检验问题的符号检验方法，中位数的点估计、区间估计等。

3.1 符号检验

3.1.1 具体操作方法

符号检验问题的原假设和备择假设有三种情况。这三种情况的原假设 H_0 都是 $me = me_0$ ，其中 me_0 是给定的常数，备择假设 H_1 分别是 $me > me_0$, $me < me_0$ 和 $me \neq me_0$ 。

由于 $P(X = me) = 0$ ，所以不妨假设样本单元 x_1, x_2, \dots, x_n 都不等于 me_0 。符号检验的检验统计量为

$$S^+ = \#G = \#\{x_i : x_i - me_0 > 0, i = 1, 2, \dots, n\}, \quad (3.1)$$

记号 $\#$ 表示计数，即 S^+ 是集合 G 中元素的个数。 S^+ 也可以等价的表示为

$$S^+ = \sum_{i=1}^n u_i, u_i = \begin{cases} 1, & x_i - me_0 > 0 \\ 0, & \text{否则} \end{cases}, i = 1, 2, \dots, n \quad (3.2)$$

由于在 $me = me_0$ 时， $S^+ \sim b(n, \frac{1}{2})$ 。

考虑备择假设 $H_1 : me > me_0$ ，我们用 p 值来度量 S^+ 是否足够大，让我们拒接原假设。 p 值等于二项分布 $b(n, \frac{1}{2})$ 的随机变量大于等于 S^+ 的概率 $P(b(n, \frac{1}{2}) \geq S^+)$ ， p 值越小， S^+ 越大。

如果 p 值 $\leq \alpha$ ，则在显著性水平 α 下拒接原假设，认为备择假设 H_1 成立；如果 p 值 $> \alpha$ ，则在显著性水平 α 下不拒绝原假设。

3.1.2 注意事项

在实际问题中，可能出现一些观察值正好等于 me_0 ，这时有以下两种处理方法：

- 1、将这些正好等于 me_0 的观察值去掉，并相应的减少样本容量 n 的值。
- 2、（不常用，不写了）

3.1.3 中位数的估计

3.1.3.1 点估计

引理 3.1

设 x_1, x_2, \dots, x_n 是来自总体 X 的样本， t_p 为总体 X 的 p 分位数， m_{np} 为样本的 p 分位数，则

$$P(\lim_{n \rightarrow \infty} m_{np} = t_p) = 1 \quad (3.3)$$

根据引理3.1，我们可以结论

$$\hat{t}_p = m_{np} = \begin{cases} x_{([np]+1)}, & np \text{ 为非整数} \\ \frac{1}{2}(x_{(np+1)} + x_{(np)}), & np \text{ 为整数} \end{cases} \quad (3.4)$$

3.1.3.2 区间估计

设 x_1, x_2, \dots, x_n 是来自总体 X 的样本, $S^+ = \#\{x_i : x_i - \text{me}_0 > 0, i = 1, 2, \dots, n\} \sim b(n, \frac{1}{2})$ 那么有

$$P(x_{(r)} \leq \text{me} \leq x_{(n-r+1)}) = 1 - P(\text{me} < x_{(r)}) - P(\text{me} > x_{(n-r+1)}) = 1 - \sum_{i=0}^{r-1} \binom{n}{i} \left(\frac{1}{2}\right)^n \quad (3.5)$$

注: 层数 r 越大, 置信区间越短, 置信水平越低 (置信水平为 $1 - \alpha$)。

3.2 符号检验在定性数据分析中的应用

根据中心极限定理, 当 n 很大时, 且 $S^+ \sim b(n, p)$, 那么 $z = \frac{S^+ - np}{\sqrt{np(1-p)}} \sim N(0, 1)$ 。

对于 $x \sim b(n, p)$, 做连续性修正:

- 1、 $P(X \leq k) \approx \Phi\left(\frac{k + \frac{1}{2} - np}{\sqrt{np(1-p)}}\right), P(X < k) \approx \Phi\left(\frac{k - \frac{1}{2} - np}{\sqrt{np(1-p)}}\right)$
- 2、 $P(X \geq k) \approx \Phi\left(\frac{np - k + \frac{1}{2}}{\sqrt{np(1-p)}}\right), P(X > k) \approx \Phi\left(\frac{np - k - \frac{1}{2}}{\sqrt{np(1-p)}}\right)$

3.3 成对数据的比较问题

定义 3.1 (配对数据)

两样本间配偶成对, 每一对样本除随机给予的不同处理外, 其他试验条件尽量一致。



第4章 符号秩和检验法

本章主要讨论对称中心的检验及估计问题。

4.1 对称中心为原点的检验问题

4.1.1 符号秩和检验统计量 W^+

符号检验统计量

$$S^+ = \sum_{i=1}^n u_i, u_i = \begin{cases} 1, & x_i > 0, \\ 0, & \text{否则,} \end{cases} i = 1, 2, \dots, n. \quad (4.1)$$

注： S^+ 仅使用样本数据量的正负信息，未使用样本数据量的大小信息。

符号秩和统计量

设 $|x_1|, |x_2|, \dots, |x_n|$ 互不相等，由大到小排列为 $z_{(1)} < z_{(2)} < \dots < z_{(n)}$ ，若 $|x_i| = z_{(R_i)}$ ，则称 $|x_i|$ 的秩为 $R_i, R_i = 1, 2, \dots, n$ 。符号秩和统计量为

$$W^+ = \sum_{i=1}^n u_i R_i \quad (4.2)$$

此处的 u_i 定义与式4.1中相同。

注： W^+ 不仅使用样本数据量的符号信息，还是使用了样本数据量的大小信息。

在表4.1中给出了 10 个观察值以及它们的 10 个观察值的符号，绝对值和绝对值的秩。这 10 个观察值的符号

表 4.1: 10 个观察值的符号，绝对值和绝对值的秩

观察值	-7.6	-5.5	4.3	2.7	-4.8	2.1	-1.2	-6.6	-3.3	-8.5
符号	-	-	+	+	-	+	-	-	-	-
绝对值	7.6	5.5	4.3	2.7	4.8	2.1	1.2	6.6	3.3	8.5
绝对值的秩	9	7	5	3	6	2	1	8	4	10

检验统计量 $S^+ = 3$ ，符号秩和统计量 $W^+ = 5 + 3 + 2 = 10$ 。

4.1.2 符号秩和检验

检验统计量： W^+ ，原假设 $H_0: \theta = 0$

1、备择假设 $H_1: \theta > 0$ ，若备择假设 H_1 成立，则 $\forall a > \theta$ ，有 $P(x > a) > P(x < -a)$ 。如图4.1所示，代码见 im4_1.r。

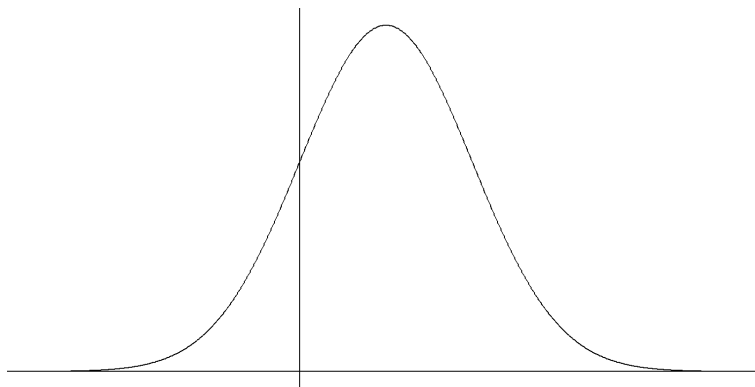


图 4.1

给定置信水平 α , 拒绝域为 $W^+ \geq c$, 其中

$$c = \inf\{c^* : P(W^+ \geq c^*) \leq \alpha\}$$

2、备择假设 $H_1 : \theta < 0$, 拒绝域为 $W^+ \leq d$, 其中

$$d = \sup\{d^* : P(W^+ \leq d^*) \leq \alpha\}$$

3、备择假设 $H_1 : \theta \neq 0$, 拒绝域为 $W^+ \geq c$ 或 $W^+ \leq d$, 其中

$$c = \inf\{c^* : P(W^+ \geq c^*) \leq \alpha/2\}, d = \sup\{d^* : P(W^+ \leq d^*) \leq \alpha/2\}.$$

4.2 符号秩和检验统计量 W^+ 的性质

4.2.1 概率分布

命题 4.1

令 $S = \sum_{i=1}^n iu_i$, 则在总体关于原点对称时, W^+ 和 S 同分布, 即 $W^+ \stackrel{d}{=} S$.

注: 总体 X 的分布关于原点对称时, u_1, u_2, \dots, u_n 相互独立同分布, 且 $P(u_i = 0) = P(u_i = 1) = \frac{1}{2}, i = 1, 2, \dots, n$. 故 $S = \sum_{i=1}^n iu_i$ 为离散型分布, 它的取值范围为 $0, 1, \dots, \frac{n(n+1)}{2}$, 并且

$$P(S = d) = P\left(\sum_{i=1}^n iu_i = d\right) = \frac{t_n(d)}{2^n}, d = 0, 1, 2, \dots, \frac{n(n+1)}{2} \quad (4.3)$$

其中 $t_n(d)$ 表示从 $1, 2, \dots, n$ 中任取若干个, 其和恰为 d 的取法数量 (其中 $t_n(d) = t_n(\frac{n(n+1)}{2} - d)$).

4.2.2 W^+ 分布的对称性

命题 4.2

在总体的分布关于原点 0 对称时, W^+ 服从对称分布, 对称中心为 $0, 1, \dots, \frac{n(n+1)}{2}$ 的中点 $\frac{n(n+1)}{4}$.

注: 当 $n \leq 30$ 时, 可查表得到符号秩和检验临界值 c_α , 使 $P(W^+ \geq c_\alpha) = \alpha$, 故当 $d = \frac{n(n+1)}{2} - c_\alpha$ 时, $P(W^+ \leq d_\alpha) = \alpha$.

4.3 符号秩和检验统计量 W^+ 的渐进正态性

4.3.1 期望与方差

H_0 成立时, $W^+ \stackrel{d}{=} S = \sum_{i=1}^n iu_i$, $u_i \sim \begin{pmatrix} 0 & 1 \\ \frac{1}{2} & \frac{1}{2} \end{pmatrix}$. 再根据 $E(u_i) = \frac{1}{2}, D(u_i) = \frac{1}{4}$, 求得 S 的期望与方差为

$$\begin{aligned} E(S) &= \frac{1}{2} \sum_{i=1}^n i = \frac{n(n+1)}{4}, \\ D(S) &= \frac{1}{4} \sum_{i=1}^n i^2 = \frac{n(n+1)(2n+1)}{24}. \end{aligned} \quad (4.4)$$

由于 W^+ 与 S 有相同的分布, 所以我们求得了 W^+ 的均值与方差。

命题 4.3

在总体分布关于原点 0 对称时,

$$\begin{aligned} E(W^+) &= \frac{n(n+1)}{4}, \\ D(W^+) &= \frac{n(n+1)(2n+1)}{24}. \end{aligned} \quad (4.5)$$

**4.3.2 W^+ 渐进正态性**

由 liapunov 中心极限定理知 S 渐进服从正态分布, 而 W^+ 与 S 有相同的分布, 所以 W^+ 也有渐进正态性。

命题 4.4

如果总体的分布关于原点 0 对称, 则在样本容量 n 趋于无穷大时, W^+ 也有渐进正态性, 即

$$\frac{W^+ - E(W^+)}{\sqrt{D(W^+)}} = \frac{W^+ - n(n+1)/4}{\sqrt{n(n+1)(2n+1)/24}} \xrightarrow{L} N(0, 1) \quad (4.6)$$

该渐进正态性简记为

$$W^+ \sim N\left(\frac{n(n+1)}{4}, \frac{n(n+1)(2n+1)}{24}\right) \quad (4.7)$$

**4.4 平均秩法****4.4.1 定义****定义 4.1**

设 x_1, x_2, \dots, x_n 为取自总体 X 的样本, 其中相等的 x_i 组成一个结, 结中 x_i 的个数称为该结的结长 $\tau(\geq 2)$, 结的个数记为 g 。



秩的定义方式: 随机秩, 平均秩。

4.4.2 性质**命题 4.5**

若总体 X 的分布关于原点对称, 有结数据取平均秩, 则

$$\begin{aligned} E(W^+) &= \frac{n(n+1)}{4}, \\ D(W^+) &= n(n+1)((2n+1)/24 - \sum_{j=1}^g (\tau_j^3 - \tau_j)/48. \end{aligned} \quad (4.8)$$



注: 有结数据取平均秩, W^+ 依旧服从渐进正态分布。

4.5 对称中心的点估计

- 1、样本的均值估计对称中心 θ 。
- 2、样本的中位数估计对称中心 θ 。
- 3、样本的切尾均值估计对称中心 θ 。
- 4、Winsort 化样本的均值估计对称中心 θ 。

定义 4.2

设 x_1, x_2, \dots, x_n 的次序统计量为 $x_{(1)}, x_{(2)}, \dots, x_{(n)}$, 称

$$W_{nk} = \frac{1}{n} \left(\sum_{i=k+1}^{n-k} x_{(i)} + kx_{(k+1)} + kx_{(n-k)} \right) \quad (4.9)$$

为对称中心的 Winsort 化均值估计。



注: Winsort 化均值估计为切尾均值的一个修正, 它加重了端头值在估计中的权重。

5、Hodges Lehmann(H-L) 估计

H-L 估计对称中心步骤如下:

(i) 先构造统计量 $T = T(x_1, x_2, \dots, x_n)$ 满足一下性质:

- $\theta = 0$ 时, T 的分布关于某点 c 对称, 且与 x 分布函数 $F(x)$ 。
- 任意 $x_1, x_2, \dots, x_n \in R$ 时, $T(x_1 + \theta, x_2 + \theta, \dots, x_n + \theta)$ 关于 θ 非降。

(ii) 定义:

$$\begin{aligned} \hat{\theta}_1 &= \sup\{a : T(x_1 - a, x_2 - a, \dots, x_n - a) > c\} \\ \hat{\theta}_2 &= \sup\{a : T(x_1 - a, x_2 - a, \dots, x_n - a) < c\} \end{aligned} \quad (4.10)$$

一般有 $\hat{\theta}_1 \leq \hat{\theta}_2$ 。

(iii) 令 $\theta = \frac{\hat{\theta}_1 + \hat{\theta}_2}{2}$ 作为对称中心 θ 的估计。

常用对称中心 H-L 估计量如下:

- (1) 当 T 统计量为 $T = \frac{\sqrt{n}\bar{x}}{S}$ 时, $\hat{\theta} = \bar{x}$;
- (2) 当 T 统计量为 $T = S^+ = \{x_i > 0, i = 1, 2, \dots, n\}$ 时, $\hat{\theta} = m_n$ (中位数);
- (3) 当 T 统计量为 $T = W^+$ 时, 将 $\{\frac{x_i + x_j}{2}, 1 \leq i \leq j \leq n\}$, 共有 $N = \frac{n(n+1)}{2}$ 个值, 从小到大排序为 $W_{(1)}^+ \leq W_{(2)}^+ \leq \dots \leq W_{(N)}^+$, 则称对称中心 θ 的 H-L 估计为 $\{\frac{x_i + x_j}{2}, 1 \leq i \leq j \leq n\}$ 的中位数。

第5章 两样本问题

5.1 Mood 中位数检验法 (2×2 列联表检验法)

5.1.1 Mood 中位数检验法

样本 x_1, \dots, x_m 和 y_1, \dots, y_n 分别来自相互独立的连续型总体 X 和 Y , 分别记其中位数为 me_x, me_y 。 ($H_0 : me_x = me_y$)

首先将样本 x_1, \dots, x_m 和 y_1, \dots, y_n 合在一起, 并从小到大排列, 计算混合样本中位数 m_n , 得四格表

表 5.1

	$\leq m_n$	$\geq m_n$	合计
X 样本	N_{11}	N_{12}	N_{1+}
Y 样本	N_{21}	N_{22}	N_{2+}
	N_{+1}	N_{+2}	N

1、备择假设为 $H_1 : me_x > me_y$

当 N_{11} 较小时, 拒绝 H_0 , 检验 p 值为

$$\sum_{k \leq N_{11}} P(k, N_{1+}, N_{+1}, N)$$

2、备择假设为 $H_1 : me_x < me_y$

当 N_{11} 较大时, 拒绝 H_0 , 检验 p 值为

$$\sum_{k \geq N_{11}} P(k, N_{1+}, N_{+1}, N)$$

其中 $P(k, N_{1+}, N_{+1}, N) = \frac{\binom{N_{+1}}{N_{11}} \binom{N_{+2}}{N_{12}}}{\binom{N}{N_{1+}}}$ 。

5.1.2 大样本情形

当样本容量较大时, 超几何分布可以近似服从正态分布, 过程与上一章大样本情形类似, 此处过程省去。

5.2 Wilcoxon 秩和检验法

5.2.1 秩

定义 5.1

设 x_1, \dots, x_N 是取自总体 X 的简单随机样本, 我们定义 x_i 的秩 R_i 为

$$R_i = \sum_{j=1}^N I_{(x_j \leq x_i)} \quad (5.1)$$

定义 5.2

设 x_1, \dots, x_N 是取自总体 X 的简单随机样本, R_i 为 x_i 的秩, 则 $R = (R_1, R_2, \dots, R_N)$ 或部分分量 $(R_1, R_2, \dots, R_m) (1 \leq m \leq N)$ 或由 R 构成的统计量统称为秩统计量。

命题 5.1

对于简单随机样本 x_1, \dots, x_N , 秩统计量 $R = (R_1, R_2, \dots, R_N)$ 等可能的取 $(1, 2, \dots, N)$ 的任意 $N!$ 个排列之一, 且 R 是由在 $(1, 2, \dots, N)$ 的所有可能的排列组成的空间 R 上的均匀分布, 即

$$P(R = (r_1, r_2, \dots, r_N)) = \frac{1}{N!} \quad (5.2)$$

注: 对于简单随机样本, R 的边缘分布也是均匀分布, 如

$$\begin{aligned} P(R_i = r) &= \frac{1}{N}, r = 1, 2, \dots, N \\ P(R_i = r_1, R_j = r_2) &= \frac{1}{N(N-1)}, r_1(\text{或} r_2) = 1, 2, \dots, N, r_1 \neq r_2 \end{aligned} \quad (5.3)$$

定理 5.1

对 $\forall i = 1, 2, \dots, N$, 有

$$E(R_i) = \frac{N+1}{2}, V(R_i) = \frac{N^2-1}{12} \quad (5.4)$$

定理 5.2

对 $\forall i \neq j$, 有

$$\text{Cov}(R_i, R_j) = -\frac{N+1}{12} \quad (5.5)$$

5.2.2 Wilcoxon 秩和检验统计量的性质

设两样本 x_1, x_2, \dots, x_m 和 $y_1, y_2, \dots, y_n (m \geq n)$, 样本容量 $N = m + n$ 。

Wilcoxon 秩和检验原假设 $H_0: X$ 与 Y 同分布。 H_0 成立时,

$$P(R_1 = r_1, R_2 = r_2, \dots, R_n = r_n) = \frac{1}{N(N-1) \cdots (N-n+1)}$$

其中 (r_1, r_2, \dots, r_n) 是从 $1, 2, \dots, N$ 中取出的 n 个数的一个排列。

记 Y 样本 y_1, y_2, \dots, y_n 的秩和为 W_y , 即

$$W_y = \sum_{j=1}^n R_j \quad (5.6)$$

1、概率分布

W_y 服从离散型分布, 最小值为 $1 + 2 + \cdots + n = \frac{n(n+1)}{2}$, 最大值为 $(m+1) + (m+2) + \cdots + (m+n) = mn + \frac{n(n+1)}{2}$ 。

命题 5.2

当 H_0 成立时,

$$\begin{aligned} P(W_y = d) &= P\left(\sum_{j=1}^n R_j = d\right) = \frac{t_{m,n}(d)}{C_N^n} \\ P(W_y \leq d) &= P\left(\sum_{j=1}^n R_j \leq d\right) = \frac{\sum_{j \leq d} t_{m,n}(j)}{C_N^n} \end{aligned} \quad (5.7)$$

其中 $t_{m,n}(d)$ 表示从 $1, 2, \dots, N$ 中任取 n 个数, 其和恰为 d 的取法总数。

2、对称性

假设从 $1, 2, \dots, N$ 中任取 n 个数为 a_1, a_2, \dots, a_n , 其和为 d , 若令 $b_i = N+1-a_i$, 则 $1 \leq b_i \leq N, i = 1, 2, \dots, n$, 其和为 $n(N+1) - d$, 则 $t_{m,n}(d) = t_{m,n}(n(N+1) - d)$ 。

故有以下结论：

$$\begin{aligned} P(W_y = d) &= P(W_y = n(N+1) - d) \\ P(W_y \leq d) &= P(W_y \geq n(N+1) - d) \end{aligned} \quad (5.8)$$

其中 $d = \frac{n(n+1)}{2}, 1 + \frac{n(n+1)}{2}, \dots, mn + \frac{n(n+1)}{2}$ 。

特别地，我们可以推出

$$\begin{aligned} P(W_y = n(N+1)/2 - d) &= P(W_y = n(N+1)/2 + d) \\ P(W_y \leq n(N+1)/2 - d) &= P(W_y \geq n(N+1)/2 + d) \end{aligned} \quad (5.9)$$

命题 5.3

当 H_0 成立时， W_y 服从对称分布，对称中心为 $\frac{n(N+1)}{2}$ 。

3、 W_y 的期望和方差

命题 5.4

当 H_0 成立时，

$$\begin{aligned} E(W_y) &= \frac{n(N+1)}{2} \\ D(W_y) &= \frac{mn(N+1)}{12} \end{aligned} \quad (5.10)$$

4、渐进正态性

命题 5.5

H_0 成立时，若 $\min\{m, n\} \rightarrow \infty$ ，且 $\frac{m}{N} \rightarrow \lambda \in (0, 1)$ ， λ 为常数，则

$$\frac{W_y - E(W_y)}{\sqrt{D(W_y)}} = \frac{W_y - n(N+1)/2}{\sqrt{mn(N+1)/12}} \xrightarrow{L} N(0, 1). \quad (5.11)$$

5.2.3 Wilcoxon 秩和检验的备择假设

原假设 H_0 : X 和 Y 同分布。而 Wilcoxon 秩和检验的备择假设有四种定量描述方法。

- 1、备择假设： $H_1 : P(X > Y) > \frac{1}{2}; P(X > Y) < \frac{1}{2}; P(X > Y) \neq \frac{1}{2}$ 。
- 2、设总体 X 和 Y 的分布函数、密度函数为 $F(x), G(x), f(x), g(x)$ ，则 $H_1 : F < G; F > G; F \neq G$ 。

定理 5.3

设总体 X 和 Y 相互独立， $\forall c \in \mathbb{R}$ ，都有 $F(c) < G(c)$ ，则 $P(X > Y) > \frac{1}{2}$ 。

- 3、若 $X+a$ 和 Y 同分布，则 a 为位置参数， $H_1 : a > 0; a < 0; a \neq 0$ 。

定理 5.4

$X+a$ 与 Y 同分布，当且仅当 $\forall c \in \mathbb{R}$ ，有 $G(c) = F(c-a)$ 。

- 4、 $H_1 : me_x > me_y; me_x < me_y; me_x \neq me_y$ 。

定理 5.5

- (1) $\forall c \in \mathbb{R}$ ，都有 $F(c) < G(c)$ ，则 $me_x > me_y$ 。
- (2) 假设 $X+a \stackrel{d}{=} Y$ 或 $\forall c \in \mathbb{R}$ ，都有 $F(c-a) = G(c)$ ，则 $me_x + a = me_y$ 。

注：以上四种备择假设的关系： $3 \rightarrow 2 \rightarrow 1, 3 \rightarrow 2 \rightarrow 4$ 。

5.2.4 Wilcoxon 秩和检验的平均秩

$W_y = \sum_{i=1}^n a(R_i)$, 其中 $a(r), r = 1, 2, \dots, n$ 为计分函数。结长为 1 时, $a(r) = r$; 结长大于 1 时, $a(r)$ 为结长的平均秩。

1、计分函数 $a(R_i)$ 的性质

$$\begin{aligned} E(a(R_i)) &= \bar{a} \\ D(a(R_i)) &= \frac{1}{N} \sum_{i=1}^N (a(i) - \bar{a})^2 \\ \text{Cov}(a(R_i), a(R_j)) &= -\frac{1}{N(N-1)} \sum_{i=1}^N (a(i) - \bar{a})^2 \end{aligned} \quad (5.12)$$

其中 $\bar{a} = \frac{\sum_{i=1}^N a(i)}{N}$ 。

定理 5.6

在 X 和 Y 同分布时, 有

$$\begin{aligned} E\left(\sum_{i=1}^n a(R_i)\right) &= n\bar{a} \\ D\left(\sum_{i=1}^n a(R_i)\right) &= \frac{nm}{N(N-1)} \sum_{i=1}^n (a(i) - \bar{a})^2 \end{aligned} \quad (5.13)$$

2、 W_y 的数字特征及渐进分布

$$\begin{aligned} E(\alpha(R_i)) &= \alpha = \frac{n(N+1)}{2} \\ D(\alpha(R_i)) &= \frac{nm(N+1)}{12} - \frac{nm}{12N(N-1)} \sum_{j=1}^g (\tau_j^3 - \tau_j) \end{aligned} \quad (5.14)$$

5.2.5 位置参数差的检验与估计

若 $\exists a$, 对 $\forall c \in \mathbb{R}$, 都有 $G(c) = F(c-a)$, 则 $X+a$ 与 Y 同分布。

1、位置参数差的检验

$$H_0 : a = \eta \text{ vs } H_1 : a < \eta; a \neq \eta; a > \eta$$

若 H_0 成立, 则 $X + \eta$ 与 Y 同分布。

2、位置参数差的估计

(1) 点估计:

- (a) 样本均值差估计位置参数差: $\hat{a} = \bar{y} - \bar{x}$;
- (b) 样本中位数差估计位置参数差: $\hat{a} = me_y - me_x$;
- (c) H-L 估计: $\hat{a} = me(Y - X)$, 即 $\{y_j - x_i, i = 1, \dots, m, j = 1, \dots, n\}$ 的中位数。

(2) 区间估计。

5.3 Mann-Whitney U 检验

5.3.1 U 统计量

1、单样本 U 统计量

定义 5.3

设 x_1, x_2, \dots, x_n 为取自总体 X 的样本, h 为 m 元函数 ($m \leq n$), 若 $h(x_1, x_2, \dots, x_m)$ 为总体分布参数 θ 的无偏估计, 即

$$E(h(x_1, x_2, \dots, x_m)) = \theta \quad (5.15)$$

则称 $U_n = U_n(x_1, x_2, \dots, x_n) = \frac{1}{A_n^m} \sum_{1 \leq i_1 \neq i_2 \neq \dots \neq i_m \leq n} h(x_{i_1}, \dots, x_{i_m})$ 为 U 统计量, 或称其是以函数 h 为核的基于样本 x_1, x_2, \dots, x_n 的参数 θ 的 U 统计量。



注:

- (i) U 统计量是 θ 的无偏估计, 即 $E(U_n) = \theta$;
- (ii) 若核函数 h 为对称核函数, 即任一 $(1, 2, \dots, m)$ 的排列 $\alpha_1, \alpha_2, \dots, \alpha_m$, 有 $h(x_{\alpha_1}, x_{\alpha_2}, \dots, x_{\alpha_m}) = h(x_1, x_2, \dots, x_m)$, 则 U 统计量可以简写为:

$$U_n = \frac{1}{C_n^m} \sum_{1 \leq i_1 \neq i_2 \neq \dots \neq i_m \leq n} h(x_{i_1}, \dots, x_{i_m}) \quad (5.16)$$

- (iii) 若核函数 h 不是对称核函数, 可以构造等价的对称核函数

$$h^*(x_1, \dots, x_m) = \frac{1}{m!} \sum_{1 \leq i_1 \neq i_2 \neq \dots \neq i_m \leq m} h(x_{i_1}, \dots, x_{i_m}) \quad (5.17)$$

其中 (i_1, i_2, \dots, i_m) 为 $(1, 2, \dots, m)$ 的任意排列。

2、两样本 U 统计量

定义 5.4

设 x_1, x_2, \dots, x_m 和 y_1, y_2, \dots, y_n 分别为取自分布为 $F(x)$ 的总体 X 和分布为 $G(y)$ 的总体 Y 的样本, h 为 $m_1 + m_2$ 元函数。若 $h(x_1, x_2, \dots, x_{m_1}, y_1, y_2, \dots, y_{m_2})$ 为总体分布参数 θ 的无偏估计, 即 $E(h(x_1, x_2, \dots, x_{m_1}, y_1, y_2, \dots, y_{m_2})) = \theta(F, G)$, 则以 h 为核基于两样本 $(x_1, x_2, \dots, x_m, y_1, y_2, \dots, y_n)$ 的参数 θ 的 U 统计量为

$$U_{mn} = \frac{1}{A_m^{m_1} A_n^{m_2}} \sum_{(1 \leq i_1 \neq \dots \neq i_{m_1} \leq m)} \sum_{(1 \leq j_1 \neq \dots \neq j_{m_2} \leq n)} h(x_{i_1}, \dots, x_{i_{m_1}}, y_{j_1}, \dots, y_{j_{m_2}}) \quad (5.18)$$



注:

- 1. U 统计量是 θ 的无偏估计, 即 $E(U_{mn}) = \theta$;
- 2. 若核函数 h 为对称核函数, 即任一 $(1, 2, \dots, m_1)$ 的排列 $(\alpha_1, \alpha_2, \dots, \alpha_{m_1})$ 和 $(1, 2, \dots, m_2)$ 的排列 $(\beta_1, \beta_2, \dots, \beta_{m_2})$, 有

$$U_{mn} = \frac{1}{C_m^{m_1} C_n^{m_2}} \sum_{(1 \leq i_1 \neq \dots \neq i_{m_1} \leq m)} \sum_{(1 \leq j_1 \neq \dots \neq j_{m_2} \leq n)} h(x_{i_1}, \dots, x_{i_{m_1}}, y_{j_1}, \dots, y_{j_{m_2}}) \quad (5.19)$$

- 3. 若核函数 h 不是对称核函数, 构造等价的对称核函数

$$h^*(x_1, \dots, x_{m_1}, y_1, \dots, y_{m_2}) = \frac{1}{m_1! m_2!} \sum_{(1 \leq i_1 \neq \dots \neq i_{m_1} \leq m_1)} \sum_{(1 \leq j_1 \neq \dots \neq j_{m_2} \leq m_2)} h(x_{i_1}, \dots, x_{i_{m_1}}, y_{j_1}, \dots, y_{j_{m_2}}) \quad (5.20)$$

5.3.2 Mann-Whitney U 统计量 (W_{xy}) 和 Wilcoxon 秩和检验统计量 (W_y)

1、Mann-Whitney U 统计量

$$\Phi(x_i, y_j) = \begin{cases} 1, & x_i - y_j < 0 \\ 0, & \text{其他} \end{cases} \quad (5.21)$$

则 $W_{xy} = \sum_{i=1}^m \sum_{j=1}^n \Phi(x_i, y_j)$ 。

2、 W_{xy} 和 W_y

定理 5.7

W_{xy} 和 W_y 仅相差一个常数, 即: $W_{xy} = W_y - \frac{n(n+1)}{2}$; $W_{yx} = W_x - \frac{m(m+1)}{2}$ 。



注: “用 Mann-Whitney U 统计量作检验统计量” 等价于 “用 Wilcoxon 秩和统计量作检验统计量”。

5.3.3 Mann-Whitney U 统计量的性质**1、小样本情形****命题 5.6**

若原假设 H_0 成立, 则 W_{xy} 服从对称分布, 分布中心为 $\frac{mn}{2}$ 。由此可以推导出如下结论:

$$\begin{aligned} P(W_y \leq d_\alpha) &= \alpha \\ P(W_{xy} \leq d_\alpha - \frac{n(n+1)}{2}) &= \alpha \\ P(W_{xy} \geq mn - d_\alpha + \frac{n(n+1)}{2}) &= \alpha \end{aligned}$$

**2、大样本情形****命题 5.7**

若原假设 H_0 成立, 则有

$$\begin{aligned} EW_{xy} &= EW_y - n(n+1)/2 = mn/2 \\ DW_{xy} &= DW_y = mn(N+1)/12 \end{aligned} \quad (5.22)$$

且若 $\min\{m, n\} \rightarrow \infty$, 且 $\frac{m}{N} \rightarrow \lambda \in (0, 1)$, λ 为常数, 则 W_{xy} 有渐进正态性。

**5.4 两样本尺度参数的秩检验方法****5.4.1 尺度参数****1、定义****定义 5.5**

设总体 X 和 Y 的分布函数分别为 $F(x)$ 和 $G(x)$, 若 $F(0) = G(0) = \frac{1}{2}$, 且对任意实数 c , 有 $G(c) = F(\frac{c}{b})$, 则称 b 为 X 与 Y 的尺度参数 ($b > 0$)。

**定理 5.8**

设总体 X 和 Y 的分布函数分别为 $F(x)$ 和 $G(x)$, 若 $F(0) = G(0) = \frac{1}{2}$, 且对任意实数 c , 都有 $G(c) = F(\frac{c}{b}) \iff bX$ 与 Y 同分布。

**2、尺度参数 b 取值大小的意义**

- (1) $b > 1$, 因为 $p(Y > c) = p(bX > c) = p(X > c/b)$, 则若 $b > 1$
 - 当 $c > 0$ 时, $p(Y > c) > p(X > c)$;
 - 当 $c < 0$ 时, $p(Y < c) > p(X < c)$ 。
- (2) $b < 1$, 分析与上面的类似。

3、

若 b 为 X 与 Y 的尺度参数, 则有

$$\sigma_y^2 = b^2 \sigma_x^2, \text{IQR}_y = b(\text{IQR}_x) \quad (5.23)$$

5.4.2 尺度参数的检验问题

$$H_0 : b = 1 \text{ vs } H_1 : b > 1; b < 1; b \neq 1$$

定义 5.6

积分函数 $\alpha(r), r = 1, 2, \dots, n$, y_i 的秩为 R_i 时, y_i 的得分为 $\alpha(R_i)$ 。



1、Mood 检验

取 $\alpha(r) = (r - \frac{N+1}{2})^2, r = 1, 2, \dots, N$, 此时 $\alpha(r)$ 为单谷函数, 记 $M_y = \sum_{i=1}^n \alpha(R_i)$ 。则当 H_0 成立时, 有

$$E(M_y) = \frac{n(N^2 - 1)}{12}, D(M_y) = \frac{nm(N+1)(N^2 - 4)}{180} \quad (5.24)$$

且若 $\min m, n \rightarrow \infty$, 且 $\frac{m}{N} \rightarrow \lambda \in (0, 1)$, λ 为常数, 则 M_y 有渐进正态性。

2、Siegal-Turkey 检验

3、Ansari-Bradley 检验

取 $\alpha(r)$ 为单峰函数, 令 $\alpha(r) = \frac{N+1}{2} - \|r - \frac{N+1}{2}\|, r = 1, 2, \dots, N$, 记 $A = \sum_{i=1}^N \alpha(R_i)$ 。

4、Klotz 检验

记标准正态分布的分布函数为 $\Phi(x)$, 取 $\Phi(x)$ 的反函数 $\Phi^{-1}(x)$, 取 $\alpha(r) = [\Phi^{-1}(\frac{r}{N+1})]^2, r = 1, 2, \dots, N$, 则 $\alpha(r)$ 为单谷函数, 记 $K_y = \sum_{i=1}^N \alpha(R_i)$ 。

第 6 章 多样本问题

6.1 Kruskal-Wallis 检验法

6.1.1 Kruskal-Wallis 检验

1、H 的分析

- 组间平方和: $SSB = \sum_{j=1}^k n_j (\bar{x}_j - \bar{x})^2$;
- 组内平方和: $SSW = \sum_{i=1}^k \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2$;
- 总平方和: $SST = \sum_{i=1}^k \sum_{j=1}^{n_i} (x_{ij} - \bar{x})^2$ 。

Bibliography

- [1] 孙山泽. 非参数统计讲义. 北京大学出版社
- [2] 陈希孺. 非参数统计. 中国科学技术大学出版社
- [3] 李裕奇. 非参数统计方法. 西南交通大学出版社