非参数统计
第一章 计数统计量和秩统计量

制作：邓明华

2023 秋季学期

本章目录

本节目录

适应任意分布的统计量

- 定义：设随机变量 $X_1, X_2, \cdots, X_n$ 是来自总体 $F(x)$ 的样本，一切可能的 $F(x)$ 组成分布类 $\mathcal{F}$. 如果统计量 $T(X_1, \cdots, X_n)$ 对任意的 $F(x) \in \mathcal{F}$ 均有相同的分布，则称 T 关于 $\mathcal{F}$ 是适应任意分布的。

- 说明：$\mathcal{F}$ 可“大”可“小”，对于象正态分布这样的分布函数族，T 检验统计量也是 d-free 的。与秩相关的统计量对函数 $F(x)$ 要求最少 (连续函数)，对应于“大”的函数族。

例子 (d-free 统计量)

- 设 $X_1, X_2, \cdots, X_n \sim N(\mu, \sigma^2 2)$，分布类 $\mathcal{F} = \{N(\mu, \sigma^2) | \sigma^2 > 0\}$，$T$ 统计量

$$T(X_1, \cdots, X_n) = \frac{(\overline{X} - \mu_0)}{S/\sqrt{n}}$$

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

- 统计量 T 对于分布族 $\mathcal{F}$ 是适应任意分布的。因为对一切 $\sigma^2 > 0$, T 的分布均为自由度为 $n-1$ 的 t 分布。

本节目录

计数统计量

- 设 X 是随机变量，对于给定的实数 θ_0, 定义随机变量 $\Psi = \psi(X > \theta_0)$, 其中

$$\psi(x) = \begin{cases} 1, & x > 0 \\ 0, & x \leq 0, \end{cases}$$

  称随机变量 Ψ 为 X 按 θ_0 分段的 计数统计量。
- **定理 1.1：** 设随机变量 $X_i \sim F_i(x)$ 相互独立，如果 θ_0 是所有分布 $F_i(x)$ 的 p_0 分位点，即 $F_i(\theta_0) = p_0$, $i = 1, \cdots, n$, 则 $\Psi_i = \psi(X_i - \theta_0)$ 相互独立同分布，其共同分布是参数为 $(1 - p_0)$ 的二点分布。
- 基于定理 1.1, 可以对分布的 p_0 分位点值进行检验，就是下面讨论的符号检验。

符号检验

- 设简单随机样本 $X_1, X_2, \cdots, X_n \sim F(x)$, $F(x)$ 在 θ_0 点连续
- 考虑符号检验统计量

$$B = \sum_{i=1}^{n} \psi(X_i - \mu_0)$$

- 如果 p_0 分位数等于 μ_0, 则 $F(\mu_0) = p_0$, $B \sim \mathcal{B}(n, 1 - p_0)$ 为二项分布;
- 如果 p_0 分位数大于等于 μ_0, 则 $F(\mu_0) < p_0$, 从而 $B \sim \mathcal{B}(n, 1 - p_0 + [p_0 - F(\mu_0)])$;
- 如果 p_0 分位数小于 μ_0, 则 $F(\mu_0) > p_0$, 从而 $B \sim \mathcal{B}(n, 1 - p_0 - [F(\mu_0) - p_0])$. 也就是说分位数点检验将与二项检验等价。

符号检验

- 记常数 $b_l(\alpha, p)$, $b_r(\alpha)$ 使得

$$P\{\mathcal{B}(n,p) \le b_l(\alpha)\} \le \alpha,\ P\{\mathcal{B}(n,p) \ge b_r(\alpha)\} \le \alpha,$$

- 实践表明，当 n 较小时（比如 $n < 20$），可以直接去查二项分布表格得到相应的临界值 $b_l(p,\alpha)$, $b_r(p,\alpha)$. 当 $n \ge 20$ 时，要采用“连续性校正”的正态近似

$$b_l(p,\alpha) = np - z_\alpha\sqrt{np(1-p)} + 0.5,$$
$$b_r(p,\alpha) = np + z_\alpha\sqrt{np(1-p)} - 0.5.$$

符号检验

- 双边检验

$$H_0：p_0分位数等于\mu_0, \leftrightarrow H_1：p_0分位数不等于\mu_0$$

的否定域为

$$B \in \left\{0, 1, \cdots, b_l(1-p_0, \frac{\alpha}{2})\right\} \cup \left\{b_r(1-p_0\frac{\alpha}{2}), b_r(1-p_0, \frac{\alpha}{2})+1, \cdots, n\right\};$$

- 单边右侧检验

$$H_0：p_0分位数小于等于\mu_0, \leftrightarrow H_1：p_0分位数大于\mu_0$$

的否定域为

$$B \in \left\{b_r(1-p_0, \alpha), b_r(1-p_0, \alpha)+1, \cdots, n\right\};$$

- 单边左侧检验

$$H_0：p_0分位数大于等于\mu_0, \leftrightarrow H_1：p_0分位数小于\mu_0$$

的否定域为

$$B \in \left\{0, 1, \cdots, b_l(1-p_0, \alpha)\right\}.$$

符号检验的 p 值

- 对于双边检验，

$$p = \begin{cases} 2P(\mathcal{B}(n, 1-p_0) \le B_0), & B_0 \le n(1-p_0), \\ 2P(\mathcal{B}(n, 1-p_0) \ge B_0)), & B_0 \ge n(1-p_0). \end{cases}$$

- 当 $n < 20$ 时，上述概率可以通过二项分布表格查出，而当 $n > 20$ 时，一般采用“连续性校正”的正态近似进行计算

$$p = \begin{cases} 2P\left\{ Z \ge \dfrac{B_0 - n(1-p_0) - 0.5}{\sqrt{np_0(1-p_0)}} \right\}, & B_0 \ge n(1-p_0), \\ 2P\left\{ Z \le \dfrac{B_0 - n(1-p_0) + 0.5}{\sqrt{np_0(1-p_0)}} \right\}, & B_0 \le n(1-p_0). \end{cases}$$

其中 Z 为标准正态分布随机变量。

符号检验的 p 值

- 对于单边右侧检验，

$$p = Pr(\mathcal{B}(n, 1-p_0) \geq B_0)$$

当 $n < 20$ 时，上述概率可以通过二项分布表格查出，而当 $n > 20$ ，一般采用“连续性校正”的正态近似进行计算

$$p = Pr\left\{ Z \geq \frac{B_0 - 0.5 - n(1-p_0)}{\sqrt{np_0(1-p_0)}} \right\}.$$

其中 Z 为标准正态分布随机变量。

- 类似地，对于单边左侧检验，

$$p = P\{\mathcal{B}(n, 1-p_0) \leq B_0\}$$

当 $n < 20$ 时，上述概率可以通过二项分布表格查出，而当 $n > 20$ 时，一般采用“连续性校正”的正态近似进行计算

$$p = P\left\{ Z \leq \frac{B_0 - n(1-p_0) + 0.5}{\sqrt{np_0(1-p_0)}} \right\}$$

其中 Z 为标准正态分布随机变量。

一个例子

- 联合国人员在 66 个大城市上的生活花费指数（以纽约市 1996 年 12 月为 100）从小到大排列如下

66	75	78	80	81	81	82	83	83	83	83
84	85	85	86	86	86	86	87	87	88	88
88	88	88	89	89	89	89	90	90	91	91
91	91	92	93	93	96	96	96	97	99	100
101	102	103	103	104	104	104	105	106	109	109
110	110	110	111	113	115	116	117	118	155	192

若北京的生活消费指数是 99, 问北京是否超出了全球 60% 的水平。

- 这是单边左侧检验问题，希望验证 60% 分位点小于北京的消费指数 99

$$H_0 : 60\%\text{分位数大于等于}\mu_0, \leftrightarrow H_1 : 60\%\text{ 位数小于}\mu_0.$$

- 统计量 $B = \sum_{i=1}^{n} \psi(X_i - \mu_0)$, B 的观测值为 $B_0 = 23$. 转换为二项单边检验问题。现在 $n(1 - p_0) = 26.4 > B_0$. 于是对应的单边左侧检验的 p 值由二项分布 $\mathcal{B}(66, 0.4)$ 给出，

一个例子

- 用 R 可以精确计算这个截尾概率，得到 p 值为 0.2345。

```
> pbinom(23,66,0.4)
[1] 0.2345
```

- 用基于连续性校正得到 p 值为 0.2331，与精确计算略有差异，

```
> pnorm((23-66*0.4+0.5)/sqrt(66*0.4*0.6))
[1] 0.2331
>  pnorm((23-66*0.4)/sqrt(66*0.4*0.6))
[1] 0.1965
```

- 也可以直接用 R 程序进行二项检验，给出的结果与精确计算一致

```
>  binom.test(23,66,0.40, alternative="less")
   Exact binomial test
data:  23 and 66
number of successes = 23, number of trials = 66, p-value = 0.2345
alternative hypothesis: true probability of success is less than 0.4
95 percent confidence interval:
 0.0000 0.4563
sample estimates:
probability of success  0.3485
```

本节目录

秩统计量

- **定义：** 设简单随机样本 $X_1, X_2, \cdots, X_n \sim F(x)$, 将观测样本按照升序排列 $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$, 则每个观测样本在序中的位置称为该样本的秩 (Rank), 即

$$X_{R_i} = X_i, R_i = \sum_{j=1}^{n} I(X_j \leq X_i)$$

- 如果观测值没有重复，则样本的秩向量唯一确定

X_i	30	23	20	38	69	19	26	52	79
R_i	5	3	2	6	8	1	4	7	9

例子

- 用 R 随机生成 6 个均匀分布的随机数，然后取其秩 (Rank)，结果如下：

```
> x<-runif(6,0,1)
> x
[1] 0.5137 0.3082 0.8298 0.1186 0.6511 0.5593
> rank(x)
[1] 3 2 6 1 5 4
```

结

- 观测值出现重复时，原秩定义不唯一

X_i	30	23	38	69	19	23	52	79	
R_i	5	4	2	6	8	1	3	7	9
R_i	5	3	2	6	8	1	4	7	9

- 此时称有结存在。
- 修正方法：将不唯一的秩定义为它们的平均。

结

- 定义: 将样本从小到大排列

$$X_{(1)} = \cdots = X_{(\tau_1)} < X_{(\tau_1+1)} = \cdots = X_{(\tau_1+\tau_2)}$$
$$< \cdots < X_{(\tau_1+\cdots+\tau_{g-1}+1)} = \cdots = X_{(\tau_1+\cdots+\tau_{g-1}+\tau_g)}$$

其中 g 为结点个数, τ_i 为结点长度, $\tau = (\tau_1, \cdots, \tau_g)$ 称为结长统计量。

- 每个结的秩:

$$r_i = \frac{1}{\tau_i}\sum_{k=1}^{\tau_i}(\tau_1 + \cdots + \tau_{i-1} + k)$$
$$= \tau_1 + \cdots + \tau_{i-1} + \frac{\tau_i + 1}{2}$$

结的例子

- 用 R 随机生成 10 个正态分布 $N(5, 4^2)$ 样本，然后对其取整，就有可能取值相同的元素，然后对其取秩。R 程序运行结果如下：

```
> x<-rnorm(10,5,4)
> x
 [1]  1.0896  1.8771  7.3279  6.2951  0.5642 11.0862
 [7] -3.0668  7.1709  9.7685  8.9747
> y<-floor(x)
> y
 [1]  1  1  7  6  0 11 -4  7  9  8
> rank(y)
 [1]  3.5  3.5  6.5  5.0  2.0 10.0  1.0  6.5  9.0  8.0
```

秩概率分布 (I)

- **定理 1.2：** 对于简单随机样本，秩 $R=(R_1,\cdots,R_n)$ 在集合 $\mathcal{R}=\{(r_1,\cdots,r_n)|(r_1,\cdots,r_n)$ 是$(1,2,\cdots,n)$ 上的一个排列$\}$。则 R 在 $\mathcal{R}$ 上均匀分布，即

$$P(R=r)=\frac{1}{n!}$$

- 说明：
  1. 均匀分布是随机抽样的体现。
  2. 仅依赖秩 R 的统计量关于连续函数分布构成的分布类是 Distribution-free 的。

秩概率分布 (II)

- **定理 1.3:** 秩 $R = (R_1, \cdots, R_n)$ 边缘分布也是均匀分布，即

$$Pr(R_i = r_i) = \frac{1}{n}$$

$$Pr(R_i = r_i, R_j = r_j) = \frac{1}{n(n-1)}$$

- 证明思路：证明所有 R_i 与 R_1 同分布，而且

$$\{R_i = r\} \cap \{R_j = r\} = \emptyset$$

秩概率分布 (III)

- **定理 1.4:** 秩 $R = (R_1, \cdots, R_n)$ 的均值，方差和协方差如下

$$\begin{cases} E(R_i) = \dfrac{n+1}{2} \\ Var(R_i) = \dfrac{n^2-1}{12} \\ cov(R_i, R_j) = -\dfrac{n+1}{12} \end{cases}$$

秩概率分布 (IV)

- 证明思路：方差计算中，展开

$$Var(R_i) = \frac{1}{n}\sum_{i=1}^{n}(i - \frac{n+1}{2})^2$$

需要利用公式

$$\sum_{i=1}^{n} i^2 = \frac{n(n+1)(2n+1)}{6}$$

- 对协方差的计算中，利用等式 $\sum_{i=1}^{n}(i - \frac{n+1}{2}) = 0$ 进行平方和展开，将 (3) 中协方差计算公式转化

$$cov(R_i, R_j) = \frac{1}{n(n-1)}\sum_{i\neq j}(i - \frac{n+1}{2})(j - \frac{n+1}{2})$$

$$= -\frac{1}{n(n-1)}\sum_{i=1}^{N}(i - \frac{n+1}{2})^2$$

Wilcoxon 统计量

- 设 $X_1, \cdots, X_m \sim F(x)$, $Y_1, \cdots, Y_n \sim G(x)$, $F(x)$, $G(x)$ 连续。考虑检验问题：随机变量 Y 是否大于随机变量 X, 即

  $H_0 : F(x) = G(x), \forall x, \Leftrightarrow H_1 : F(x) \geq G(x)$ 且在某些点不等式严格成立

- Wilcoxon 统计量：将 $X_1, \cdots, X_m$, $Y_1, \cdots, Y_n$ 共 $m+n$ 个观测值混合在一起排序，产生秩向量 $R = (Q_1, \cdots, Q_m; R_1, \cdots, R_n)$. 定义 Wilcoxon 秩和统计量为

$$W = \sum_{i=1}^{n} R_i$$

Wilcoxon 秩和统计量

- 参数检验：正态分布下 T 检验；
- 直观意义：Y 随机大于 X，则相应的秩倾向于偏大，秩和也倾向于偏大；
- Wilcoxon 秩和检验对应于参数统计的 T 检验；

Wilcoxon 秩和统计量概率分布 (I)

- **定理 1.5:** 在 $H_0 : F(x) = G(x), \forall x$ 下，Wilcoxon 秩和统计量 W 的分布为

$$P(W = d) = \frac{t_{m,n}(d)}{C_{m+n}^{n}}$$

其中 $d = n(n+1)/2, \cdots, m + n(n+1)/2$. $t_{m,n}(d)$ 表示从 $1, \cdots, m+n$ 中取 n 个数和为 d 的所有可能取法。

- 说明：在定理 1.2 下，以上定理实际上是一个平凡的组合计算，如何求 $t_{m,n}(d)$ 才是关键。

Wilcoxon 秩和统计量概率分布 (II)

- 关于 d 的取值问题，最小对应于选取的是 $m+n$ 个数中的前 n 个，和为 $n(n+1)/2$; 最大对应于选取的是 $m+n$ 个数中的后 n 个，和为 $m+n(n+1)/2$.
- 下面看 $tm,n(d)$ 的计算：看混合样本中最大的样本，其秩为 $m+n$, 这个样本只有两种可能。如果它属于总体 X，没有被 W 记入，此时还是在 $m+n-1$ 个数中选取，有 $t_{m-1,n}(d)$ 种组合方式; 如果这个样本属于总体 Y，被 W 记入，此时只要再选 $n-1$ 个数使其和为 $d-m-n$ 即可。故可以递推计算 $t_{m,n}(d)$.

Wilcoxon 秩和统计量概率分布 (III)

- 有如下递推关系式

$$\begin{cases} t_{m,n}(d) = t_{m-1,n}(d) + t_{m,n-1}(d-m-n) \\ t_{i,0}(0) = 1, i = 1, \cdots, m; \\ t_{i,0}(d) = 0, d \neq 0, i = 1, \cdots, m; \\ t_{0,j}(\dfrac{j(j+1)}{2}) = 1 \\ t_{0,j}(d) = 0; d \neq \dfrac{j(j+1)}{2}, j = 1, \cdots, n. \end{cases}$$

- 对小样本用以上公式计算，大样本时需要有下一章介绍的正态近似进行近似计算。

Wilcoxon 秩和统计量的性质

- **定理 1.7：** 在 H_0 $F(x) = G(x), \forall x$ 下，Wilcoxon 秩和统计量 W 满足

$$\begin{cases} E(W) = \dfrac{n(m+n+1)}{2} \\ Var(W) = \dfrac{mn(m+n+1)}{12} \end{cases}$$

  且 W 的分布关于 $n(m+n+1)/2$ 对称。
- 直接计算可以得到

Mann-Whitney 统计量

- 考察不同总体观测样本对之间的差异, $\psi(Y_i - X_j)$，如果 Y 随机大于 X，则多数观测值对取正。将所有观测对的符号相加在一起，得到的统计量就是 Mann-Whitney 统计量，即

$$U = \sum_{i=1}^{m} \sum_{j=1}^{n} \psi(Y_j - X_i)$$

两个统计量之间的关系

- Wilcoxon 统计量 W 和 Mann-Whitney 统计量 U 满足如下关系

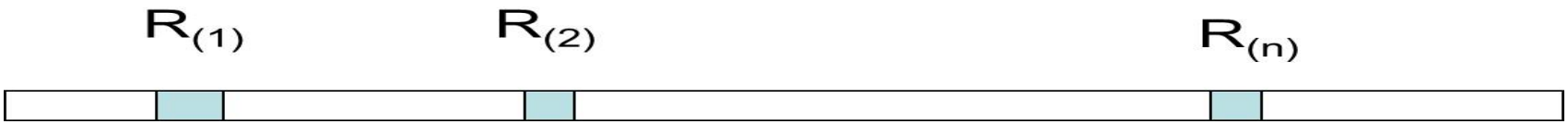
$$W = U + \frac{n(n+1)}{2}$$

- 证明思路：将 Y 排序

$$Y_{(1)} \leq Y_{(2)} \leq \cdots \leq Y_{(n)}$$

设它们在混合样本中的序为 $R_{(1)}, \cdots, R_{(n)}$, 那么可以数出在这些样本前面的 X 样本的个数。它们对应于 U 统计量的定义。

两统计量之间的关系



$$\#\{X_i < Y_{(1)}, i = 1, \cdots, m\} = R_{(1)} - 1$$

$$\cdots\cdots\cdots\cdots$$

$$\#\{X_i < Y_{(j)}, i = 1, \cdots, m\} = R_{(j)} - j$$

$$\cdots\cdots\cdots\cdots$$

$$\#\{X_i < Y_{(n)}, i = 1, \cdots, m\} = R_{(n)} - n$$

$$U = \sum_{j=1}^{n}(R_{(j)} - j) = \sum_{j=1}^{n} R_j - \sum_{j=1}^{n} j$$

$$= W - \frac{n(n+1)}{2}$$

两样本位置检验

- 设 $\omega(m,n,\alpha)$ 是参数为 n,m 的 Wilcoxon 秩和统计量的上分位点

$$P\{W(m,n)>\omega(m,n,\alpha)\}\le\alpha.$$

  由对称性，相应的下 $1-\alpha$ 上分位数为 $n(n+m+1)-\omega(m,n,\alpha)$.
- 双侧检验的拒绝域为

$$\left\{X:\ W<n(n+m+1)-\omega(m,n,\frac{\alpha}{2})\right\}\cup\left\{X:\ W>\omega(m,n,\frac{\alpha}{2})\right\};$$

- 单边左侧检验的拒绝域为

$$\left\{X:\ W<n(n+m+1)-\omega(m,n,\alpha)\right\};$$

  单边右侧检验的拒绝域为

$$\left\{X:\ W>\omega(m,n,\alpha)\right\}.$$

两样本位置检验 p 值

- 设当前的 Wilcoxon 秩和统计量的观测值为 W_0
- 双侧检验的 p 值为

$$p = \begin{cases} 2P\{W(m,n) \geq W_0\}, & W_0 > \dfrac{n(n+m+1)}{2}, \\ 2P\{W(m,n) \leq W_0\}, & W_0 < \dfrac{n(n+m+1)}{2}; \end{cases}$$

- 单边左侧检验的 p 值为

$$p = P\{W(m,n) \leq W_0\};$$

- 单边右侧检验的 p 值为

$$p = P\{W(m,n) \geq W_0\};$$

其中 $W(m,n)$ 为参数为 m,n 的 Wilcoxon 秩和随机变量。

两样本位置检验 p 值

- 一般地，当 $0 \le n, m \le 20$ 时，可以借助于 Wilcoxon 秩和统计量的分位数表得到上述的截尾概率。当样本量 n 或者 m 超过 20 时，可以下一章的结论，用正态近似计算上述截尾概率，

$$
\begin{aligned}
& P(W(m,n) \ge W_0) \\
=& P\left(Z \ge \frac{W_0 - 0.5 - \dfrac{n(n+m+1)}{2}}{\sqrt{\dfrac{nm(n+m+1)}{12} - \dfrac{nm}{12(n+m)(n+m+1)} \sum_{i=1}^{g} (\tau_i^3 - \tau_i)}} \right) \\
& P(W(m,n) \le W_0) \\
=& P\left(Z \le \frac{W_0 + 0.5 - \dfrac{n(n+m+1)}{2}}{\sqrt{\dfrac{nm(n+m+1)}{12} - \dfrac{nm}{12(n+m)(n+m+1)} \sum_{i=1}^{g} (\tau_i^3 - \tau_i)}} \right)
\end{aligned}
$$

其中 Z 为标准正态分布随机变量，而式中加减 0.5 是所谓的“连续性修正”。

一个例子

- 两个班同学进行某项体育测试，成绩如下：

甲班	2.4	6.2	9.9	6.4	6.1	10.6	9.1	15.3
	14.8	6.7	6.7	10.6	5.0	3.6	18.6	1.8
	2.6	1.0	3.2	5.9	4.0			
乙班	14.8	10.6	12.7	16.9	7.6	7.3	12.5	14.2
	7.9	11.3	5.6	12.9	12.6	16.0	8.3	6.3
	16.1	2.1	10.6	9.0	11.4	17.7	5.6	4.2
	7.2	11.8	5.6					

请问两个班成绩的分布是否相同?

一个例子

-
- 这是一个双边检验问题

$$H_0 : F(x) = G(x), \leftrightarrow \ H_1 : F(x) \neq G(x).$$

将两个数据混合排序，得到混合秩如下表

甲班	2.4	6.2	9.9	6.4	6.1	10.6	9.1	15.3	14.8	6.7	6.7
Q_i	4	16	28	18	15	30.5	27	43	41.5	19.5	19.5
	10.6	5.0	3.6	18.6	1.8	2.6	1.0	3.2	5.9	4.0	
Q_i	30.5	10	7	48	2	5	1	6	14	8	
乙班	14.8	10.6	12.7	16.9	7.6	7.3	12.5	14.2	7.9	11.3	5.6
R_i	41.5	30.5	38	46	23	22	36	40	24	33	12
	12.9	12.6	16.0	8.3	6.3	16.1	2.1	10.6	9.0	11.4	17.7
R_i	39	37	44	25	17	45	3	30.5	26	34	47
	5.6	4.2	7.2	11.8	5.6						
R_i	12	9	21	35	12						

一个例子 (续)

- 数据中有 4 个结长超过 1 的，分别为 $3, 2, 3, 2$, 对应的观测为 $5.6, 6.7, 10.6, 14.8$.
- 对 Y 中的秩求和得到 $W_0 = 782.5$, $m = 21$, $n = 27$. 考虑到 R 中实际是 Mann-Whitney 统计量 U 进行的两样本检验，故 $U_0 = W_0 - \frac{n(n+1)}{2} = 404.5$. 于是 p 值为

$$p = 2P(U(21, 27) \geq U_0) = 0.0110,$$

$$p \approx 2P\left(Z \geq \frac{W_0 - 0.5 - \frac{n(n+m+1)}{2}}{\sqrt{\frac{nm(n+m+1)}{12} - \frac{nm}{12(n+m)(n+m+1)}\sum_i (t_i^3 - \tau_i)}}\right) = 0.0122$$

- 在 R 中，Wilcoxon 分布实际上是 Mann-Whitney 统计量 U 的分布,

```
> 2*(1-pwilcox(404.5,21,27))
[1] 0.0110
> 2*(1-pnorm((782-27*49/2)/sqrt(21*27*49/12-21*27*60/(12*48*49))))
[1] 0.0122
```

一个例子 (续)

- 直接采用 R 的函数 wilcoxon.test() 对两个总体进行 Wilcoxon 秩和检验，得到如下结果：

```
> x<-c(2.4,6.2,9.9,6.4,6.1,10.6,9.1,15.3,
+      14.8,6.7,6.7,10.6,5.0,3.6,
+      18.6,1.8,2.6,1.0,3.2,5.9,4.0)
> y<-c(14.8,10.6,12.7,16.9,7.6,7.3,12.5,14.2,7.9,
+    11.3,5.6,12.9,12.6,16.0,8.3,6.3,16.1,2.1,10.6,
+    9.0, 11.4,17.7,5.6, 4.2, 7.2, 11.8, 5.6)
> wilcox.test(y,x,alternative="two.sided")
        Wilcoxon rank sum test with continuity correction
data:  y and x
W = 404.5, p-value = 0.0122
alternative hypothesis: true location shift is not equal to 0
Warning message:
In wilcox.test.default(y, x, alternative = "two.sided"):
  无法精确计算带连结的p值
```

- 比较两个计算结果，此时用正态近似计算出来的 p 值误差很小。

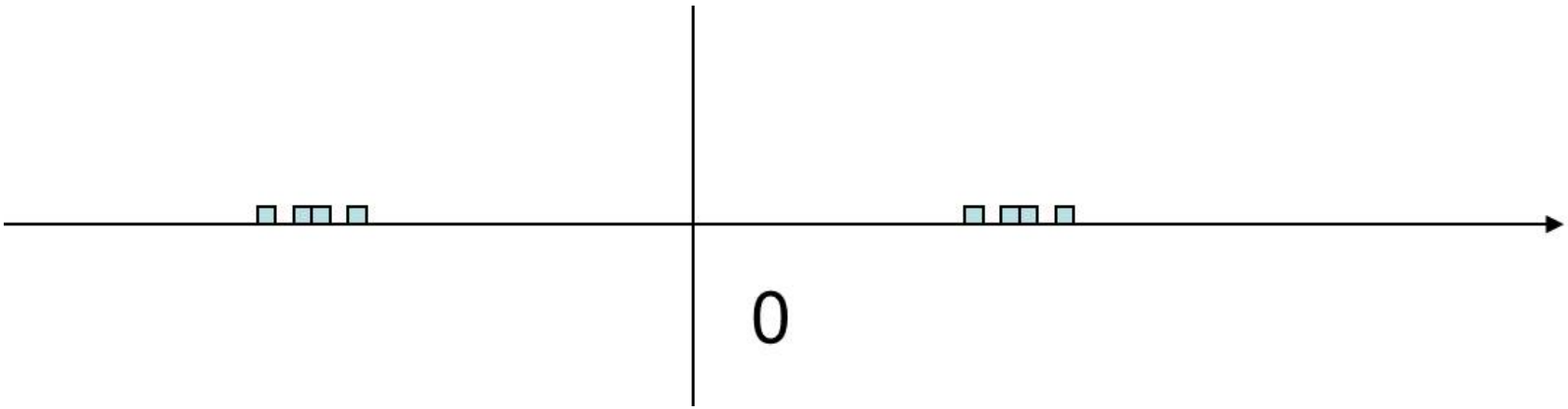
本节目录

Wilcoxon 符号秩检验

- 如何对对称分布的中位数进行检验?
- 对中位数值可采用符号检验。
- 对称性刻划: 正负样本到原点的距离。故用观测值的绝对值进行排序得到秩向量。
- 把正、负样本当成两个总体，将负样本取绝对值后和正样本进行位置比较。
- 将符号统计量和绝对值对应的秩统计量相结合，得到符号秩统计量。

符号秩统计量

- 定义：简单随机样本 $X_1, \cdots, X_n \sim F(x)$, $F(x)$ 连续，关于 0 点对称，计数统计量 $\Psi_i = \psi(X_i)$, 随机变量 $|X_1|, \cdots, |X_n|$ 对应的秩向量记为 $R^+ = (R_1^+, \cdots, R_n^+)$, R_i^+ 称为 X_i 的绝对秩，则 $\Psi_i R_i^+$ 称为 X_i 的符号秩。



绝对值和符号统计量

- **定理 1.8：** 若 X 是连续型随机变量，分布关于 0 点对称，则 $|X|$ 与其符号统计量 $\Psi(X)$ 相互独立。
- 证明：

$$
\begin{aligned}
& Pr\big(\Psi(X)=1,|X|\le t\big) \\
=& Pr\big(X>0,|X|\le t\big) \\
=& Pr\big(0<X\le t\big) \\
=& \frac{1}{2}\big[Pr\big(0<X\le t\big)+Pr\big(-t<X\le 0\big)\big] \\
=& \frac{1}{2}Pr\big(|X|\le t\big) \\
=& Pr\big(\Psi(X)=1\big)Pr\big(|X|\le t\big)
\end{aligned}
$$

Distribution-free

- **定理 1.9：** 若简单随机样本 $X_1, \cdots, X_n \sim F(x)$, $F(x)$ 连续，关于 0 点对称，其绝对秩向量 $R^+ = (R_1^+, \cdots, R_n^+)$, 计数统计量 $\Psi_i = \psi(Xi)$, 则 $\Psi_1, \cdots, \Psi_n$, R^+ 相互独立；Ψ 服从参数为 1/2 的两点分布，R^+ 在 $(1, \cdots, n)$ 的排列空间上均匀分布。
- 定理理解：对称性和样本随机性决定了 Ψ 服从两点分布；由对称性把样本镜像到正半轴后随机排序，故 R^+ 仍然在排列空间上均匀分布。
- 定理说明符号秩统计量对于关于 0 点对称的分布族 Distribution-free.

符号秩统计量概率分布 (I)

- **定理 1.10：** 设简单随机样本 $X_1, \cdots, X_n \sim F(x)$, $F(x)$ 连续，关于 0 点对称。令计数统计量 $\Psi_i = \psi(X_i)$, 样本中正数个数可以表示为

$$Q = \sum_{i=1}^{n} \Psi_i$$

当 $Q = q$ 时，$X_1, \cdots, X_n$ 中正样本对应的绝对秩记为 $S_1 < \cdots < S_q$, 则

$$\begin{aligned} & Pr(Q = q, S_1 = t_1 \cdots, S_q = t_q) \\ = & \begin{cases} (\frac{1}{2})^n, & q = 0, 1, \cdots, n, t_1 \leq \cdots \leq t_q, \\ 0, & \text{其它.} \end{cases} \end{aligned}$$

定理 1.10 证明

$$
\begin{aligned}
&Pr\big(Q=q, S_1=t_1, \cdots, S_q=t_q\big) \\
&=\sum^{C_n^q} Pr\big(\psi_{d_1}=1, \cdots, \psi_{d_q}=1; \psi_{d_{q+1}}=0, \cdots, \psi_{d_n}=0; S_1=t_1, \cdots, S_q=t_q\big) \\
&=\sum^{C_n^q} Pr\big(\psi_{d_1}=1, \cdots, \psi_{d_q}=1; \psi_{d_{q+1}}=0; \cdots, \psi_{d_n}=0; S_1=t_1, \cdots, S_q=t_q\big) \\
&=\sum^{C_n^q} Pr\big(S_1=t_1, \cdots, S_q=t_q | \psi_{d_1}=1, \cdots, \psi_{d_q}=1; \psi_{d_{q+1}}=0, \cdots, \psi_{d_n}=0\big) \\
&\qquad\qquad Pr\big(\psi_{d_1}=1, \cdots, \psi_{d_q}=1; \psi_{d_{q+1}}=0, \cdots, \psi_{d_n}=0\big) \\
&=\sum^{C_n^q} \sum^{q!(n-q)!} Pr\big((R_{d_1}^+, \cdots R_{d_q}^+)=t, (R_{d_{q+1}}^+, \cdots R_{d_n}^+)=r | \psi_{d_1}=1, \cdots, \psi_{d_q}=1; \\
&\qquad\qquad \psi_{d_{q+1}}=0, \cdots, \psi_{d_n}=0\big)(\frac{1}{2})^n \\
&=\sum^{C_n^q} \sum^{q!(n-q)!} Pr\big((R_{d_1}^+, \cdots, R_{d_n}^+)=(t,r)\big) \cdot (\frac{1}{2})^n \\
&=(\frac{1}{2})^n \sum^{C_n^q} \sum^{q!(n-q)!} \frac{1}{n!} = (\frac{1}{2})^n
\end{aligned}
$$

Wilcoxon 符号秩统计量概率分布 (I)

- **定理 1.11：** 设简单随机样本 $X_1, \cdots, X_n \sim F(x)$. $F(x)$ 连续，关于 0 点对称，相应的符号秩统计量为 $\Psi_1 R_1^+, \cdots, \Psi_n R_n^+$, 则 $W^+ = \sum_{i=1}^n \Psi_i R_i^+$ 的概率分布为

$$Pr(W^+ = k) = \begin{cases} \frac{C_n(k)}{2^n}, & k = 0, 1, \cdots, \frac{n(n+1)}{2}, \\ 0, & \text{其它}. \end{cases}$$

  其中 $C_n(k)$ 表示集合内元素和恰为 k 的 $\{1, 2, \cdots, n\}$ 的子集个数。

- 定理证明：由定理 1.10,

$$\begin{aligned} & Pr(W^+ = k) \\ =& \sum_{q=0}^{n} \sum_{t_1+\cdots+t_q=k} Pr\{Q = q, S_1 = t_1, \cdots, S_q = t_q\} \\ =& \frac{C_n(k)}{2^n} \end{aligned}$$

Wilcoxon 符号秩统计量概率分布 (II)

- 递推公式计算 $C_n(k)$,

$$C_n(k) = C_{n-1}(k-n) + C_{n-1}(k)$$
$$C_1(0) = C_1(1) = 1, C_1(d) = 0, d \neq 0, 1$$

- 递推公式推导，看最后一个元素 n 是否包含在子集之中：若其在子集中，只要从前 $n-1$ 个中取子集使元素和为 $k-n$; 若其不在子集中，要从前 $n-1$ 个中取子集使元素和为 k.

Wilcoxon 符号秩统计量

- 如果用 w_j 表示绝对秩为 j 的样本所对应的样本的符号统计量，即

$$w_j = \Psi(X_{R_j^+})$$

- 于是 Wilcoxon 符号秩统计量可以表示为

$$W^+ = \sum_{j=1}^{n} j w_j$$

用生成函数求 W+ 概率分布

- 按照定义

$$E(\exp(tjw_j)) = \frac{1}{2}\exp(0) + \frac{1}{2}\exp(tj)$$
$$= \frac{1}{2}(1 + e^{tj})$$

- 进一步有

$$M(t) = E(\exp(tW^+)) = E(\exp(\sum_{j=1}^{n} tjw_j))$$
$$= \prod_{j=1}^{n} E(\exp(tjw_j)) = \frac{1}{2^n}\prod_{j=1}^{n}(1 + e^{tj})$$

- 若 $M(t) = \alpha_0 + \alpha_1 e^t + \alpha_2 e^{2t} + \cdots$, 则 $Pr(W^+ = k) = \alpha_k$.

Wilcoxon 符号秩统计量

- **定理 1.12：** 设简单随机样本 $X_1, \cdots, X_n \sim F(x)$, $F(x)$ 连续，关于 0 点对称，$W^+ = \sum_{i=1}^{n} \Psi_i R_i^+$ 是 Wilcoxon 符号秩统计量，则

$$\begin{cases} E(W^+) = \dfrac{n(n+1)}{4} \\ Var(W^+) = \dfrac{n(n+1)(2n+1)}{24} \end{cases}$$

  且 W 的分布关于 $n(n+1)/4$ 对称。

- 定理 1.12 证明：均值

$$\begin{aligned} E(W^+) &= E(\sum_{j=1}^{n} \Psi_i R_i^+) = \sum_{i=1}^{n} E(\Psi_i) E(R_i^+) \\ &= \sum_{i=1}^{n} \frac{1}{2}(\frac{n+1}{2}) = \frac{n(n+1)}{4} \end{aligned}$$

定理 1.12 证明

- 方差

$$
\begin{aligned}
Var(\Psi_i R_i^+) &= E\Big[(\Psi_i R_i^+ - E(\Psi_i)E(R_i^+))^2\Big] \\
&= E\Big[(\Psi_i R_i^+ - \Psi_i E(R_i^+) + \Psi_i E(R_i^+) - E(\Psi_i)E(R_i^+))^2\Big] \\
&= E\Big[\Psi_i^2 (R_i^+ - E(R_i^+))^2\Big] + (E(R_i^+))^2 E\Big[(\Psi_i - E(\Psi_i))^2\Big] \\
&= E(\Psi_i^2) E\Big[(R_i^+ - E(R_i^+))^2\Big] + (E(R_i^+))^2 E\Big[(\Psi_i - E(\Psi_i))^2\Big] \\
&= \frac{1}{2} Var(R_i^+) + (E(R_i^+))^2 Var(\Psi_i) \\
&= \frac{1}{2} * \frac{(n-1)(n+1)}{12} + (\frac{n+1}{2})^2 * \frac{1}{4} \\
&= \frac{(n+1)(5n+1)}{48}
\end{aligned}
$$

定理 1.12 证明

- 协方差

$$
\begin{aligned}
&\operatorname{cov}(\Psi_i R_i^+, \Psi_i R_j^+) \\
&= E\big((\Psi_i R_i^+ - E(\Psi_i)E(R_i^+))(\Psi_j R_j^+ - E\Psi_j E(R_j^+))\big) \\
&= E(\Psi_i)E(\Psi_j)\operatorname{cov}(R_i^+, R_j^+) \\
&= \frac{1}{2} * \frac{1}{2} * (-\frac{n+1}{12}) \\
&= -\frac{n+1}{48}
\end{aligned}
$$

- 于是

$$
\begin{aligned}
Var(W^+) &= \sum_{i=1}^{n} Var(\Psi_i R_i^+) + \sum_{i \neq j} \operatorname{cov}(\Psi_i R_i^+, \Psi_j R_j^+) \\
&= n * \frac{(n+1)(5n+1)}{48} - \frac{n(n-1)(n+1)}{48} \\
&= \frac{n(n+1)(2n+1)}{24}
\end{aligned}
$$

定理 1.12 证明

- 因为分布 $F(x)$ 关于 0 点对称，从而

$$\Psi_i \overset{d}{=} 1 - \Psi_i$$

- 于是

$$\begin{aligned} W^+ &= \sum_{i=1}^{n} \Psi_i R_i^+ \\ &\overset{d}{=} \sum_{i=1}^{n} (1 - \Psi_i) R_i^+ \\ &\overset{d}{=} \frac{n(n+1)}{2} - W^+ \end{aligned}$$

Walsh 平均和 Wilcoxon 符号秩

- Walsh 平均：设 $X_1, \cdots, X_n$ 是 $F(x)$ 的简单随机样本，计算任意两个数的平均，得到的 $n(n+1)/2$ 个数据，这组数称之为 Walsh 平均，考虑其符号可以得到 Wilcoxon 符号秩统计量的另一个表达方式，

$$W^+ = \#\{\frac{X_i + X_j}{2} > 0, i \leq j = 1, 2, \cdots, n\}$$

  即 W^+ 是 Walsh 平均值中正数的个数。

- 这一结论类似与 Wilcoxon 秩和统计量用 Mann-Whitney 统计量的表达方式。
- 实际上，如果把 Wilcoxon 符号秩统计量看成对称分布的正样本和负样本的 Wilcoxon 秩和统计量，那么 Mann-Whitney 统计量中考虑 Y_j 和 X_i 的大小关系 $\psi(Y_j - X_i)$ 就转换成考虑 Walsh 平均值的符号 $\psi(X_j + X_i)$.

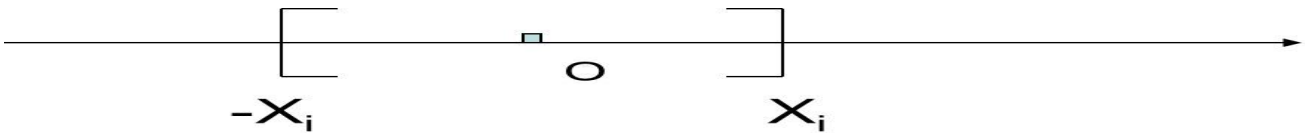
结论的证明

- 首先容易验证：正样本 X_i 的绝对秩 R_i^+ 是落入区间 $[-X_i, X_i]$ 的样本点的个数。
- 显然落入区间 $[-X_i, X_i]$ 的样本 X_j 都满足 $X_j + X_i > 0$, 于是

$$R_i^+ = \#\{X_j + X_i > 0 | j \leq R_i\}$$

- 而对于负样本，

$$\#\{X_j + X_i > 0 | j \leq R_i\} = 0$$



结论证明 (II)

$$
\begin{aligned}
W^+ &= \sum_{i=1}^{n} \Psi_i R_i^+ \\
&= \sum_{i=1}^{n} \sum_{j \le R_i} \Psi\left(X_{(R_i)} + X_{(j)}\right) \\
&= \sum_{i=1}^{n} \sum_{j \le i} \Psi\left(X_{(i)} + X_{(j)}\right) \\
&= \sum_{i=1}^{n} \sum_{j \le i} \Psi\left(X_i + X_j\right)
\end{aligned}
$$

对称分布中位数检验

- 设简单随机样本 $X_1, \cdots, X_n \sim F(x-\theta)$, $F(t)$ 关于 0 点对称;
- 若存在临界值 $\omega^+(n, \frac{\alpha}{2})$, 使得 $P\left\{W^+ > \omega^+(n, \frac{\alpha}{2})\right\} > \frac{\alpha}{2}$, 以其作为右侧临界点; 由对称性, 左侧临界点为 $\frac{n(n+1)}{2} - \omega^+(n, \frac{\alpha}{2})$
- 双边检验

$$H_0 : \mu = \mu_0, \leftrightarrow H_1 : \mu \neq \mu_0$$

的拒绝域为

$$\left\{W^+ > \omega^+(n, \frac{\alpha}{2})\right\} \cup \left\{W^+ < \frac{n(n+1)}{2} - \omega^+(n, \frac{\alpha}{2})\right\}$$

- 相应地, 单边左侧检验

$$H_0 : \mu \geq \mu_0, \leftrightarrow H_1 : \mu < \mu_0$$

的拒绝域为

$$\left\{W^+ < \frac{n(n+1)}{2} - \omega^+(n, \alpha)\right\};$$

- 单边右侧检验

$$H_0 : \mu \leq \mu_0, \leftrightarrow H_1 : \mu > \mu_0$$

的拒绝域为

$$\left\{W^+ > \omega^+(n, \alpha)\right\}.$$

p 值

- 设观测值为 W_0^+, 那么双侧检验的 p 值为

$$p = \begin{cases} 2P\{W^+(n) \geq W_0^+\}, & W_0^+ \geq \dfrac{n(n+1)}{4}, \\ 2P\{W^+(n) \leq W_0^+\}, & W_0^+ \leq \dfrac{n(n+1)}{4}; \end{cases}$$

- 单边左侧检验的 p 值为

$$p = P\{W^+(n) \leq W_0^+\};$$

- 单边右侧检验的 p 值为

$$p = P\{W^+(n) \geq W_0^+\},$$

其中 $W^+(n)$ 是样本量为 n 的符号秩和随机变量。

p 值

- 一般地，当 $n \leq 50$ 时，上述截尾概率都可以通过查找符号秩和随机变量的分位数表得到上述临界值，当 $n > 50$ 时，可以基于定理下一章的定理进行正态近似，

$$P\{W^+(n) \geq W_0^+\} = P\left(Z \geq \frac{W_0^+ - 0.5 - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24} - \frac{1}{48}\sum_{i=1}^{g}(\tau_i^3 - \tau_i)}}\right),$$

$$P\{W^+(n) \leq W_0^+\} = P\left(Z \leq \frac{W_0^+ + 0.5 - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24} - \frac{1}{48}\sum_{i=1}^{g}(\tau_i^3 - \tau_i)}}\right),$$

其中，Z 为标准正态分布随机变量，而加减 0.5 是所谓的“连续性修正”。

一个例子

- 学院为了解邮件系统垃圾邮件情况，收集了 20 位老师每周收到的垃圾邮件数目，得到如下数据，

310	350	370	270	389	400	415	420	400	290
295	325	340	298	365	375	250	385	263	440

- 经验上可以认为垃圾邮件数服从中心为 μ 的对称分布，能否认为学院邮件系统每人每周收到垃圾邮件数的中心位置为 320.

一个例子

- 这是对称中心的检验问题

$$H_0 : \mu = 320, \leftrightarrow H_1 : \mu \neq 320.$$

X_i	310	350	370	270	389	400	415	420	400	290
Y_i	-10	30	50	-50	69	80	95	100	80	-30
R_i^+	2	6.5	9.5	9.5	14	16.5	18	19	16.5	6.5
X_i	295	325	340	298	365	375	250	385	263	440
Y_i	-25	5	20	-22	45	55	-70	65	-57	120
R_i^+	5	1	3	4	8	11	15	13	12	20

根据上表计算得到 Wilcoxon 符号秩 $W^+ = 156$. 查表得到 Wilcoxon 符号秩检验统计量的上分位点 $\omega^+(20, 0.025) = 157$, $\omega^+(20, 0.05) = 149$. 所以能够拒绝单边右侧检验零假设，但不能拒绝双边检验零假设。

一个例子 (p 值)

- 由于 $W_0^+ > \dfrac{n(n+1)}{4} = 105$ 现在直接调用 R 程序计算截尾概率，得到单边 p 值为 0.0266，而双边 p 值为单边 p 值的两倍 0.0532.
- 该数据存在结，结数为 17, 结长为 $\tau = (1,1,1,1,1,2,1,2,1,1,1,1,1,2,1,1,1)$. 如果用正态近似去逼近 p 值，近似值为

$$
\begin{aligned}
p &= P\left(Z \geq \frac{W_0^+ - 0.5 - \dfrac{n(n+1)}{4}}{\sqrt{\dfrac{n(n+1)(2n+1)}{24} - \dfrac{\sum_i(\tau_i^3-\tau_i)}{48}}}\right) \\
&= P\left(Z \geq \frac{156 - 0.5 - \dfrac{20\times 21}{4}}{\sqrt{\dfrac{20\times 21\times 41}{24} - \dfrac{3\times(2^3-2)}{48}}}\right) \\
&= 0.0297
\end{aligned}
$$

- 上述截尾概率的计算可以由 R 程序完成，

```
> 1-psignrank(156,20)
[1] 0.0266
> 1-pnorm((155.5-20*21/4)/sqrt(20*21*41/24-3*(8-2)/48))
[1] 0.0297
```

一个例子 (p 值)

- R 也可以直接进行 Wilcoxon 符号秩检验，程序与运行结果如下：

```
> x<-c(310,350,370,270,389,400,415,420,400,290,295,
+      325,340, 298,365,375,250, 385, 263,440)
> y=x-320
> wilcox.test(y,alternative="two.sided")
        Wilcoxon signed rank test with continuity correction
data:  y
V = 156, p-value = 0.05932
alternative hypothesis: true location is not equal to 0
Warning message:
In wilcox.test.default(y, alternative = "two.sided") : 无法
```

- 这个结果非常接近前面前面按照结修正的单边 p 值 0.0297 的两倍，表面了 R 中 Wilcoxon 符号秩检验缺省进行了结矫正。

本节目录

条件的适应任意分布统计量

- **定理 1.13：** 给定一组有序样本 $(x_{(1)}, \cdots, x_{(n)})$ 后，$(X_1, \cdots, X_n)$ 在 $(x_{(1)}, \cdots, x_{(n)})$ 的 $N!$ 个排列上均匀分布。
- 平均秩处理结 (以 Wilcoxon 秩和统计量为例)：$X_1, \cdots, X_m \sim F(x)$, $Y_1, \cdots, Y_n \sim F(x - \Delta)$. 考虑假设检验问题

$$H_0 : \Delta = 0, \Leftrightarrow H_1 : \Delta > 0.$$

在零假设 H_0 下，条件在 X 和 Y 的一组取值上，秩统计量 (Q, R) 在混合排序定义的秩排列空间上均匀分布，从而可以给出条件否定域。

定理 1.13 的证明

- 对于 $(1, 2, \cdots, N)$ 的任一排列 $r = (r_1, \cdots, r_N)$ 和 N 维实数集合 $A \subset \{(x_1, \cdots, x_N) \,|\, x_1 \leq \cdots \leq x_N\}$

$$\begin{aligned} & Pr(R = r, X^0 \in A) \\ &= Pr(X_1 = X_{(r_1)}, \cdots, X_N = X_{(r_N)}, X^0 \in A) \\ &= Pr((X_{d_1}, \cdots, X_{d_N}) \in A) \\ &= Pr((x_1, \cdots, x_N) \in A) \end{aligned}$$

其中 $d_j = i$ 当 $ri = j$, 对一切 i, j 成立. 于是

$$\begin{aligned} Pr(X^0 \in A) &= \sum_{(r_1, \cdots, r_N)} P(R = r, X^0 \in A) \\ &= N! Pr(x_1, \cdots, x_N) \in A) \end{aligned}$$

- 故

$$Pr(R = r, X^0 \in A) = \frac{1}{N!} Pr(X^0 \in A) = Pr(R = r) Pr(X^0 \in A)$$

本节目录

带结的 Wilcoxon 秩和统计量

- **定理 1.14:** 设 $X_1, \cdots, X_m \sim F(x)$, $Y_1, \cdots, Y_n \sim F(x-\Delta)$ 为独立的简单随机样本, 若观测数据中结长为 g, 结向量为 $(\tau_1, \cdots, \tau_g)$. 在零假设 H_0 下, 条件在一组观测值上, 采用平均秩, 则秩和统计量 $W = \sum_{i=1}^{n} R_n$ 满足,

$$\begin{cases} E(W) = \dfrac{n(N+1)}{2} \\ Var(W) = \dfrac{nm(N+1)}{12} - \dfrac{nm}{12N(N-1)} \displaystyle\sum_{j=1}^{g} (\tau_j^3 - \tau_j) \end{cases}$$

定理 1.14 证明 (I)

- 首先引进计分函数 $a(r)$, $r = 1, 2, \cdots, n$. 对于第 j 个结中的全部, 计分函数取平均秩, 即:

$$a(r) = \tau_1 + \cdots + \tau_{j-1} + \frac{\tau_j + 1}{2}$$

$$\tau_1 + \cdots + \tau_{j-1} + 1 \leq r \leq \tau_1 + \cdots + \tau_{i-1} + \tau_j$$

- 其次对计分函数有

$$E(a(R_i)) = \overline{a} = \frac{\sum_{i=1}^{N} a(i)}{N}$$

$$Var(a(R_i)) = \frac{1}{N} \sum_{i=1}^{N} (a(i) - \overline{a})^2$$

$$Cov(a(R_i), a(R_j)) = -\frac{1}{N(N-1)} \sum_{i=1}^{N} (a(i) - \overline{a})^2$$

定理 1.14 证明 (II)

- 对于计分函数下的秩和统计量 $W = \sum_{i=1}^{n} a(R_i)$,

$$E(W) = E(\sum_{i=1}^{n} a(R_i)) = n\bar{a}$$

$$Var(W) = E\big[\big(\sum_{i=1}^{n}(a(R_i) - \bar{a})\big)^2\big]$$

$$= \sum_{i=1}^{n} Var(a(R_i)) + 2 \sum_{1 \le i < j \le n} Cov(a(R_i), a(R_j))$$

$$= \big(\frac{n}{N} - \frac{n(n-1)}{N(N-1)}\big) \sum_{i=1}^{N}(a(i) - \bar{a})^2$$

$$= \frac{nm}{N(N-1)} \sum_{i=1}^{N}(a(i) - \bar{a})^2$$

定理 1.14 证明 (III)

- 对于平均秩计分函数，

$$\overline{a} = \frac{1}{N}\sum_{i=1}^{N} a(i) = \frac{N+1}{2}$$

$$\sum_{i=1}^{N}(a(i)-\overline{a})^2 = \sum_{i=1}^{N} a(i)^2 - N\overline{a}^2$$

- 因此，只要计算在同一个结中平均秩的平方和与无结时秩的平方和的差别即可。

定理 1.14 的证明 (IV)

- 同一个结内的平均秩平方和，

$$[r+\frac{\tau+1}{2}]^2+\cdots+[r+\frac{\tau+1}{2}]^2=\tau[r+\frac{\tau+1}{2}]^2$$

- 无结时的秩平方和，

$$[r+1]^2+\cdots+[r+\tau]^2=\tau r^2+r\tau(\tau+1)+\frac{\tau(\tau+1)(2\tau+1)}{6}$$

- 两者之间的差别

$$\frac{(\tau^3-\tau)}{12}$$

定理 1.14 的证明 (V)

- 于是

$$E(W) = n\bar{a} = \frac{n(N+1)}{2}$$

$$Var(W) = \frac{nm}{N(N-1)} \sum_{i=1}^{N} (a(i) - \bar{a})^2$$

$$= \frac{nm}{N(N-1)} \left(1^2 + 2^2 + \cdots + N^2 - N * (\frac{N+1}{2})^2 - \sum_{j=1}^{g} (\tau_j^3 - \tau_j)\right)$$

$$= \frac{nm(N+1)}{12} - \frac{nm}{12N(N-1)} \sum_{j=1}^{g} (\tau_j^3 - \tau_j)$$

带结的 Wilcoxon 符号秩和统计量

- **定理 1.15:** 设 $X_1, \cdots, X_n \sim F(x)$ 是简单随机样本, $F(x)$ 连续关于 0 点对称。设观测数据的绝对值存在 g 个结, 结向量为 $(\tau_1, \cdots, \tau_g)$. 那么条件在一组观测值上, 符号秩统计量 $W^+ = \sum_{i=1}^{n} \Psi_i R_i$ 满足,

$$\begin{cases} E(W^+) = \dfrac{n(n+1)}{4} \\ Var(W^+) = \dfrac{n(n+1)(2n+1)}{24} - \displaystyle\sum_{j=1}^{g} \dfrac{(\tau_j^3 - \tau_j)}{12} \end{cases}$$

定理 1.15 的证明 (I)

- 首先引进计分函数 $a(r)$, $r = 1, 2, \cdots, n$. 对于第 j 个结中的全部，计分函数取平均秩，即：

$$a(r) = \tau_1 + \cdots + \tau_{j-1} + \frac{\tau_j + 1}{2}$$

$$\tau_1 + \cdots + \tau_{j-1} + 1 \leq r \leq \tau_1 + \cdots + \tau_{i-1} + \tau_j$$

- 其次有如下等式成立

$$\begin{cases} W^+ = \sum_{i=1}^{n} \Psi_i a(R_i) \stackrel{d}{=} \sum_{i=1}^{n} \Psi_i a(i) \\ E(W^+) = \sum_{i=1}^{n} a(i)/2 \\ Var(W^+) = \sum_{i=1}^{n} a(i)^2/4 \end{cases}$$

定理 1.15 的证明 (II)

- 于是只看一个结内的运算

$$[r+\frac{\tau+1}{2}]+\cdots+[r+\frac{\tau+1}{2}]=\tau[r+\frac{\tau+1}{2}]$$
$$[r+\frac{\tau+1}{2}]^2+\cdots+[r+\frac{\tau+1}{2}]^2=\tau[r+\frac{\tau+1}{2}]^2$$

- 没有结时的运算结果,

$$[r+1]+\cdots+[r+\tau]=\tau[r+\frac{\tau+1}{2}]$$
$$[r+1]^2+\cdots+[r+\tau]^2=\tau r^2+r\tau(\tau+1)+\frac{\tau(\tau+1)(2\tau+1)}{6}$$

定理 1.15 的证明 (III)

- 平方和在前后两次运算中的差别在于少了

$$\frac{(\tau^3 - \tau)}{12}$$

- 于是有结时

$$\begin{aligned} Var(W^+) &= \sum_{i=1}^{n} a(i)/4 \\ &= \frac{1}{4}\left(1^2 + 2^2 + \cdots + n^2 - \sum_{j=1}^{g} \frac{(\tau_j^3 - \tau_j)}{12}\right) \\ &= \frac{n(n+1)(2n+1)}{24} - \sum_{j=1}^{g} \frac{(\tau_j^3 - \tau_j)}{48} \end{aligned}$$

作业

- P27. 2
- P27. 3
- P27. 4
- P27. 5