非参数假设检验

参数模型—-总体的分布族的数学形式已知,只有少数几个实参数的值未知—事先有比较多的关于总体分布的信息

非参数方法—-事先只有很少的关于总体分布的信息，需要一种与总体分布族的具体数学形式无关的统计方法.

非参数方法的特点

回忆: 总体分位数定义如下: 对$0 < p < 1$, 若

$$F(\xi_p) = p$$

或者

$$F(\xi_p) < p \quad \text{但} \quad F(\xi_p + 0) \geq p,$$

则称ξ_p为总体X(或分布函数$F(x)$)的(下侧)p分位数, 1/2分位数称为中位数.

§符号检验

设$X \sim F(x)$为连续型随机变量，X的p分位数$t_p (0 < p < 1)$满足

$$F(t_p) = p.$$

$m_e = t_{1/2}$为中位数.
假设X的密度函数$p(x)$在点t_p处连续且$p(t_p) > 0$,那么此时t_p唯一.
考察假设检验问题

$$H_0 : t_p \leq t_0 \longleftrightarrow H_1 : t_p > t_0. \tag{1}$$

设$\widetilde{X}=(X_1,\cdots,X_n)$是来自总体$X$的一个样本，令

$$Y_i=\begin{cases}1, & X_i-t_0>0,\\ 0, & \text{其它}.\end{cases}$$

记$\theta=\mathsf{P}(Y_i=1)=\mathsf{P}(X_i-t_0>0)$. 则$Y_1,\cdots,Y_n$ i.i.d.$\sim B(1,\theta)$，并且$\sum_{i=1}^n Y_i\sim B(n,\theta)$. $\sum_{i=1}^n Y_i$恰好是$\{X_i-t_0;i=1,2,\cdots,n\}$中正值的个数$N^+$. 当$H_0$为真时，即$t_p\le t_0$时，有

$$\begin{aligned}\theta=&\mathsf{P}(X_i-t_0>0)=\mathsf{P}(X_i>t_0)\le\mathsf{P}(X_i>t_p)\\ =&1-\mathsf{P}(X_i\le t_p)=1-F(t_p)=1-p.\end{aligned}$$

反过来，若$\theta\le 1-p$，则$\mathsf{P}(X_i-t_0>0)\le\mathsf{P}(X_i-t_p>0)$，从而$t_p\le t_0$.

所以假设检验问题(1)等价于

$$H_0 : \theta \le 1 - p \longleftrightarrow H_1 : \theta > 1 - p. \tag{2}$$

问题(2)的检验统计量是$\sum_{i=1}^n Y_i = N^+$，拒绝域为

$$\{\tilde{x} : \sum_{i=1}^n y_i \ge C^*\} = \{\tilde{x} : N^+ \ge C^*\},$$

其中

$$\begin{aligned} C^* &= \min\{C : \mathsf{P}_{1-p}(N^+ \ge C) \le \alpha, C\text{为整数}\} \\ &= \min\left\{C : \sum_{i=C}^n \binom{n}{i}(1-p)^i p^{n-i} \le \alpha, C\text{为整数}\right\}. \end{aligned}$$

H_0	H_1	拒绝域
$t_p \le t_0$	$t_p > t_0$	$\{\tilde{x}: N^+ \ge C^*\}$, $C^* = \min\left\{C \in \mathcal{N}: \sum_{i=C}^{n} \binom{n}{i}(1-p)^i p^{n-i} \le \alpha\right\}$
$t_p \ge t_0$	$t_p < t_0$	$\{\tilde{x}: N^+ \le C^*\}$, $C^* = \max\left\{C \in \mathcal{N}: \sum_{i=0}^{C} \binom{n}{i}(1-p)^i p^{n-i} \le \alpha\right\}$
$t_p = t_0$	$t_p \ne t_0$	$\{\tilde{x}: N^+ \le C_1^*$ 或$N^+ \ge C_2^*\}$, $C_1^* = \max\left\{C \in \mathcal{N}: \sum_{i=0}^{C} \binom{n}{i}(1-p)^i p^{n-i} \le \alpha/2\right\}$ $C_2^* = \min\left\{C \in \mathcal{N}: \sum_{i=C}^{n} \binom{n}{i}(1-p)^i p^{n-i} \le \alpha/2\right\}$

§秩和检验(Wilcoxon Test)

一、基本概念

Definition

定义 设$x_1,\cdots,x_n$为两两互不相等的实数,若$x_1,\cdots,x_n$中恰有R_i个元素的值不超过x_i(包括x_i自己), 则称x_i在$(x_1,\cdots,x_n)$中的秩(rank)为R_i.
若将$x_1,\cdots,x_n$按从小到大排列成$x_{(1)}<x_{(2)}<\cdots<x_{(n)}$,若$x_i$的秩为$R_i$
则$x_i=x_{(R_i)}$, 反之亦然.

Definition

定义6.2.1 设$X_1, X_2, \cdots, X_n$为两两互不相等的一组样本,将其从小到大排列成$X_{(1)} < X_{(2)} < \cdots < X_{(n)}$，若$X_i = X_{(R_i)}$，则称$X_i$在样本$(X_1, \cdots, X_n)$中的秩(rank)为$R_i$.

显然,若$X_1, X_2, \cdots, X_n$为来自连续分布$F(x)$的一组样本,则以概率1保证$X_1, X_2, \cdots, X_n$是两两互不相等.

Definition

定义6.2.2 设$X_1, X_2, \cdots, X_n$为来自单个总体的样本，或来自多个总体的合样本. 记R_i为X_i的秩，则称$R = (R_1, R_2, \cdots, R_n)$为$(X_1, X_2, \cdots, X_n)$的秩统计量(rank statistics)，其中R_i为X_i的秩. 由R导出的统计量也称为秩统计量. 基于秩统计量的检验方法称为秩检验(rank test).

Theorem

定理 设$X_1, X_2, \cdots, X_n$ 为来自连续型总体$X \sim F(x)$的简单随机样本，则$X_1, X_2, \cdots, X_n$的秩统计量R取$(1, 2, \cdots, n)$的任一置换的概率都是为$1/n!$.

证明：设$(k_1, k_2, \cdots, k_n)$为$(1, 2, \cdots, n)$的一个置换，记d_i 为k_i 的反变换，即：如果$k_j = i$ 则$d_i = j$. 由于$(X_{d_1}, X_{d_2}, \cdots, X_{d_n})$与$(X_1, X_2, \cdots, X_n)$同分布，所以

$$
\begin{aligned}
& \mathsf{P}\left(R_1 = k_1, R_2 = k_2, \cdots, R_n = k_n\right) \\
= & \mathsf{P}\left(X_1 = X_{(k_1)}, X_2 = X_{(k_2)}, \cdots, X_n = X_{(k_n)}\right) \\
= & \mathsf{P}\left(X_{(1)} = X_{d_1}, X_{(2)} = X_{d_2}, \cdots, X_{(n)} = X_{d_n}\right) \\
= & \mathsf{P}\left(X_{d_1} < X_{d_2} < \cdots < X_{d_n}\right) \\
= & \mathsf{P}\left(X_1 < X_2 < \cdots < X_n\right) = \mathsf{P}\left(R_1 = 1, R_2 = 2, \cdots, R_n = n\right) = \frac{1}{n!}.
\end{aligned}
$$

二、秩和检验

检验统计量

两连续总体$X \sim F(x)$, $Y \sim G(y)$, 独立. 为比较两总体,
取$X_1, X_2, \cdots, X_m$和$Y_1, Y_2, \cdots, Y_n$分别来自这两个总体的两个独立样本.

记Y_i在合样本$(X_1, X_2, \cdots, X_m, Y_1, Y_2, \cdots, Y_n)$中的秩为$R_i$, $i = 1, 2, \cdots, n$.

Wilcoxon提出把

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

作为检验统计量, 用于处理比较$F(x)$和$G(y)$大小的检验问题.

拒绝域的形式

若$F(x)>G(x)$, 则

$$\begin{aligned}\mathsf{P}(X>Y)=&\int\int_{x>y}dG(y)dF(x)=\int_{-\infty}^{\infty}(\int_{-\infty}^{x}dG(y))dF(x)\\ &=\int_{-\infty}^{\infty}G(x)dF(x)\\ &<\int_{-\infty}^{\infty}F(x)dF(x)=\int_0^1 zdz=1/2.\end{aligned}$$

这表明, 在$F(x)>G(x)$时, Y的取值偏大的可能性大, Y_i, $i=1,2\cdots,n$, 的秩应偏大, 从而W的值应偏大才合理.

若$F(x)<G(x)$, 则同理$\mathsf{P}(X>Y)>1/2$, Y的取值偏小的可能性大, Y_i, $i=1,2\cdots,n$, 的秩应偏小, 从而W的值应偏小才合理.

H_0	H_1	拒绝域
$F(x) \leq G(x)$	$F(x) > G(x)$	$W \geq c$
$F(x) \geq G(x)$	$F(x) < G(x)$	$W \leq d$
$F(x) = G(x)$	$F(x) \neq G(x)$	$W \leq d$ 或$W \geq c$.

临界值的确定

先来看看当$F(x)=G(x)$时，W的分布.

W服从离散型分布，

最小值为$1+2+\cdots+n=n(n+1)/2$,

最大值为$(m+1)+\cdots+(m+n)=n(n+1)/2+mn=n(2m+n+1)/2$.

由前面的定理知，

$$\mathsf{P}\{W=i\}=\frac{t_{m,n}(i)}{\binom{m+n}{n}},$$

$$i=n(n+1)/2, n(n+1)/2+1,\cdots,n(2m+n+1)/2,$$

其中，$t_{m,n}(i)$为在$1,2,\cdots,m+n$中不重复地取出n个数，其和恰好为i的组合种数.

秩和统计量的性质:

- 分别记$(X_1,\cdots,X_m)$和$(Y_1,\cdots,Y_n)$在合样本中的秩和为W_X、W_Y，则

$$W_X+W_Y=1+2+\cdots+(m+n)=(m+n)(m+n+1)/2.$$

用W_X作为检验统计量与用W_Y作为检验统计量都是可以的;

- 在$F(x)=G(x)$时候，W的分布关于$n(m+n+1)/2$对称,即 W 与$n(m+n+1)-W$ 同分布,

$$\mathsf{P}(W\le d)=\mathsf{P}(n(m+n+1)-W\le d)=\mathsf{P}(W\ge n(m+n+1)-d).$$

事实上，设$(a_1, a_2, \cdots, a_n)$是从$1, 2, \cdots, m+n$中不重复取出的和为i的n个数. 令$b_j = m+n+1-a_j$，则$(b_1, b_2, \cdots, b_n)$是从$1, 2, \cdots, m+n$中不重复取出的和为$n(m+n+1)-i$的n个数. 所以

$$t_{m,n}(i) = t_{m,n}(n(m+n+1)-i).$$

因此

$$\mathsf{P}(W=i) = \mathsf{P}(W = n(m+n+1)-i) = \mathsf{P}(n(m+n+1)-W=i),$$

$$\text{对任意的}i = \frac{n(n+1)}{2}, \frac{n(n+1)}{2}+1, \cdots, \frac{n(2m+n+1)}{2}.$$

- $\sup_{F(x)\le G(x)} \mathsf{P}\left\{W \ge c \middle| X_i \sim F, Y_j \sim G\right\} = \mathsf{P}_{F(x)=G(x)}\left\{W \ge c\right\}$.

事实上，记$W = W(\widetilde{X}, \widetilde{Y})$，则$W$为$Y_1, \cdots, Y_n$的非降函数，为$X_1, \cdots, X_m$的非增函数. 不妨设，$Y_i = G^{-1}(U_i)$，其中$U_1, \cdots, U_n$ i.i.d.$\sim U(0,1)$.
当$F(x) \le G(x)$时，$F^{-1}(x) \ge G^{-1}(x)$. 因此

$$\begin{aligned} &\mathsf{P}(W(\widetilde{X}, \widetilde{Y}) \ge c) = \mathsf{P}(W(\widetilde{X}, G^{-1}(U_1), \cdots, G^{-1}(U_n)) \ge c) \\ &\le \mathsf{P}(W(\widetilde{X}, F^{-1}(U_1), \cdots, F^{-1}(U_n)) \ge c). \end{aligned}$$

而$F^{-1}(U_i) \sim F$. 结论得证.

- $\sup_{F(x)\ge G(x)} \mathsf{P}\{W \le d \big| X_i \sim F, Y_j \sim G\} = \mathsf{P}_{F(x)=G(x)}\{W \le d\}.$

临界值的确定

小样本情形:

当m和n不太大时，可以计算出当$F(x)=G(x)$时，

$$\mathsf{P}\{W \geq c_\alpha\} = \sum_{i \geq c_\alpha} \mathsf{P}\{W = i\} \leq \alpha$$

成立的最小整数,即为这里的临界值c_α(也可参见附表12).

由

$$\mathsf{P}\{W \leq d\} = \mathsf{P}\{W \geq n(n+m+1) - d\},$$

又可以得到

$$\mathsf{P}\{W \leq d_\alpha\} \leq \alpha$$

的临界值$d_\alpha = n(n+m+1) - c_\alpha$.

Example

例6.3.1 某种羊毛在进行某种工艺处理之前与处理之后,各随机抽取一个样本,测得其含脂率如下:

处理前: 0.20, 0.24, 0.66, 0.42, 0.12;

处理后: 0.13, 0.07, 0.21, 0.08, 0.19.

问该处理后含脂率是否下降($\alpha = 0.05$).

解: 第一步: 提出假设

设$X \sim F(x)$和$Y \sim G(x)$分别表示处理前、后羊毛的含脂率. 由于

$$F(x) > G(x) \Rightarrow \mathsf{P}(X > Y) < 1/2,$$

$$F(x) < G(x) \Rightarrow \mathsf{P}(X > Y) > 1/2.$$

这说明当$F(x) > g(x)$时, 相对于X而言, Y的值偏大;当$F(x) < g(x)$时, 相对于X而言, Y的值偏小.

所以"处理后含脂率没有下降"可用"$F(x) \geq G(x)$"表示;"处理后含脂率下降"可用"$F(x) < G(x)$"表示. 这样我们要检验

$$H_0 : F(x) \geq G(x) \longleftrightarrow H_1 : F(x) < G(x).$$

(若$G(x) = F(x-\theta)$,则上述假设检验可写为:

$$H_0 : \theta \geq 0 \longleftrightarrow H_1 : \theta < 0.$$

)

第二步: 给出检验统计量与拒绝域的形式. 设$X_1,\cdots,X_m$和$Y_1,\cdots,Y_n$分别来自$F(x)$和$G(x)$的两个样本. 记Y_i在合样本$(X_1,\cdots,X_m,Y_1,\cdots,Y_n)$中的秩为$R_i$, $i=1,2,\cdots,n$.我们采用的是秩和检验, 取检验统计量为

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

那么拒绝域的形式为$\{W\leq d\}$.

第三步: 计算出临界值, 以确定拒绝域.

由$m = n = 5$, $\alpha = 0.05$. 查表,得

$$\mathsf{P}\{W \geq 36\} \leq 0.05.$$

从而$d = n(m+n+1) - 36 = 19$. 所以

$$\mathsf{P}\{W \leq 19\} = \mathsf{P}\{W \geq 36\} \leq 0.05.$$

拒绝域为

$$\{w \leq 19\}.$$

第四步：求W的观察值，从而做出判断.
将两个样本观察值合起来,按从小到大排序得

X			0.12			0.20		0.24	0.42	0.66
Y	0.07	0.08		0.13	0.19		0.21			
秩	1	2	3	4	5	6	7	8	9	10

处理后羊毛含脂率观察值相对应的秩和为

$$w = 1 + 2 + 4 + 5 + 7 = 19.$$

由$w \le 19$，所以拒绝原假设，即认为处理后羊毛含脂率下降了.
(两点说明)

大样本情形

Theorem

定理 假设$F(x) \equiv G(x)$. 则

$$EW = \frac{n(m+n+1)}{2}, \tag{1}$$

$$Var\{W\} = \frac{mn(m+n+1)}{12}. \tag{2}$$

当$m, n \to \infty$时

$$W^* = \frac{W - n(m+n+1)/2}{\sqrt{mn(m+n+1)/12}} \xrightarrow{D} N(0,1). \tag{3}$$

我们只证明(1), (2). 由于R_i $(i=1,\cdots,n)$的分布为点集$(1,2,\cdots,m+n)$上的均匀分布, 所以

$$\mathsf{E}(R_i)=\sum_{j=1}^{m+n} j/(m+n)=(m+n+1)/2.$$

$$\begin{aligned}\mathsf{Var}(R_i)=&\sum_{j=1}^{m+n} j^2/(m+n)-\big[E(R_i)\big]^2\\ =&(m+n+1)(m+n-1)/12.\end{aligned}$$

由于R_i与R_j $(i \neq j, i, j = 1, \cdots, n)$的联合分布为点集

$$\{(k, l) : k \neq l, k, l = 1, \cdots, m+n\}$$

上的均匀分布, 所以

$$Cov(R_i, R_j) = \sum_{k \neq l} (kl)/[(m+n)(m+n-1)] - \mathsf{E}(R_i)\mathsf{E}(R_j)$$

$$= -(m+n+1)/12.$$

从而$\mathsf{E}(W) = \mathsf{E}(\sum_{i=1}^n R_i) = n(m+n+1)/2$,

$$\begin{aligned}\mathsf{Var}(W) =& \mathsf{Var}(\sum_{i=1}^n R_i) = \sum_{i=1}^n \mathsf{Var}(R_i) + 2\sum_{i<j} \mathsf{Cov}(R_i, R_j) \\ =& mn(m+n+1)/12.\end{aligned}$$

由(3), 当m,n充分大(当大于7时,一般就可以用),在$F(x)\equiv G(x)$的条件下,

$$W^* \sim_{\text{近似}} N(0,1),$$

取之为检验统计量.

H_0	H_1	拒绝域
$F(x)\le G(x)$	$F(x)>G(x)$	$W^*\ge u_\alpha$
$F(x)\ge G(x)$	$F(x)<G(x)$	$W^*\le -u_\alpha$
$F(x)=G(x)$	$F(x)\ne G(x)$	$\lvert W^*\rvert\ge u_{\alpha/2}$.

Matlab program

```
x=[0.20,0.24,0.66,0.42,0.12];
y=[0.13,0.07,0.21,0.08,0.19];
[p,h,stat] = ranksum(x,y,'alpha',0.10)
p = 0.0952
h = 1
stat = ranksum: 36
```

ranksum

Wilcoxon rank sum test for equal medians

Syntax

h = ranksum(x,y)

[p,h] = ranksum(x,y)

[p,h] = ranksum(x,y,'alpha',alpha)

[p,h,stats] = ranksum(...)

Description

h = ranksum(x,y) performs a two-sided rank sum test of the hypothesis that two independent samples, in the vectors x and y, come from distributions with equal medians, and returns the p-value from the test. p is the probability of observing the given result, or one more extreme, by chance if the null hypothesis is true, i.e., the medians are equal. Small values of p cast doubt on the validity of the null hypothesis. The two sets of data are assumed to come from continuous distributions that are identical except possibly for a location shift, but are otherwise arbitrary. x and y can be different lengths. The Wilcoxon rank sum test is equivalent to the Mann-Whitney U test.

[p,h] = ranksum(x,y) returns the result of the hypothesis test, performed at the 0.05 significance level, in h. If h = 0, then the null hypothesis, i.e., medians are equal, cannot be rejected at the 5% level. If h = 1, then the null hypothesis can be rejected at the 5% level.

[p,h] = ranksum(x,y,'alpha',alpha) returns the result of the hypothesis test performed at the significance level alpha.

[p,h] = ranksum(...,'method', method) computes the p-value using an exact algorithm, if you set method to 'exact' or a normal approximation, if you set method to 'approximate'. If you omit this argument, ranksum uses the exact method for small samples and the approximate method for larger samples.

[p,h,stats] = ranksum(...) returns stats, a structure with one or two fields. The field 'ranksum' contains the value of the rank sum statistic. If the sample size is large, then p is calculated using a normal approximation and the field 'zval' contains the value of the normal (Z) statistic.

Example

This example tests the hypothesis of equal medians for two independent unequal-sized samples. The theoretical distributions are identical except for a shift of 0.25.

x = unifrnd(0,1,10,1);

y = unifrnd(.25,1.25,15,1);

[p,h] = ranksum(x,y,0.05)

p = 0.0375

h = 1

§成对数据的检验问题
—–一样本问题中的非参数假设检验

成对数据问题

Example

例5.2.5 今有两台测量材料中某金属含量的光谱仪A和B，为鉴定它们的质量有无显著差异，对金属含量不同的9件材料样品进行测量，得到9对观察值为

i	1	2	3	4	5	6	7	8	9
u(单位：%)	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	1.00
v(单位：%)	0.10	0.21	0.52	0.32	0.78	0.59	0.68	0.77	0.89

问根据实验结果，能否判断这两台光谱仪的质量有无显著的差异($\alpha = 0.01$)？

Example

例6.2.3 工厂的两个化验室，每天同时从工厂的冷却水中取样,测量水中的含氯量(ppm).

i	1	2	3	4	5	6
x_i(A)	1.15	1.86	0.76	1.82	1.14	1.65
y_i(B)	1.00	1.90	0.90	1.80	1.20	1.70
i	7	8	9	10	11	
x_i(A)	1.92	1.01	1.12	0.90	1.40	
y_i(B)	1.95	1.02	1.23	0.97	1.52	

问：两个化验室测定的结果之间有无显著的差异($\alpha = 0.10$)?

数据结构:

$$X_i = \mu_i + \xi_i, \quad Y_i = \mu_i + \eta_i, \quad i = 1, 2, \cdots, n.$$

X_i, Y_i—观测到的样本,

μ_i—真值;

ξ_i, η_i–测量误差(假设$\xi_i, i.i.d.$, $\eta_i, i.i.d.$, 且相互独立).

检验问题:ξ_i 的分布$F(x)$与η_i的分布$G(x)$是否相同, 即

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

记

$$Z_i = X_i - Y_i = \xi_i - \eta_i.$$

Z_i是成对数据统计推断的出发点.

若$F(x) = G(x)$, 则$Z_i = \xi_i - \eta_i$与$\eta_i - \xi_i = -Z_i$同分布.即Z_i的分布关于原点对称.

如果Z_i, $i.i.d.$服从正态分布, 则可以采用t检验.

如果Z_i分布未知呢? 就要采用非参数检验. (假设Z_i服从连续分布)

成对数据的符号检验法:

令$N^+=\sum_{i=1}^n I\{Z_i>0\}$为$\{Z_i\}$中正值的个数. 则在$H_0: F(x)=G(x)$为真时, N^+的值既不偏大,也不偏小. 故拒绝域为

$$D=\{N^+\geq c \text{ 或} N^+\leq d\}.$$

另一方面, 当H_0为真时, Z_i的分布关于原点对称, 所以$N^+\sim B(n,1/2)$(假设$F(x)$和$G(x)$均为连续的). 所以临界值c和d由下式确定:

$$d=\max_{d'}\{d'\geq 0, d'\text{为整数}: \sum_{k=0}^{d'}\binom{n}{k}\cdot\left(\frac{1}{2}\right)^n\leq \alpha/2\},\quad c=n-d.$$

在例6.2.3中, $n = 11$, $\alpha = 0.10$.

$$\sum_{k=0}^{2} \binom{11}{k} \cdot \left(\frac{1}{2}\right)^{11} = 0.0327 \le 0.05$$

$$\sum_{k=0}^{3} \binom{11}{k} \cdot \left(\frac{1}{2}\right)^{11} = 0.113 > 0.05.$$

所以$d = 2$, $c = 11 - d = 9$. 水平为$\alpha = 0.10$的符号检验的拒绝域为

$$\{N^+ \le 2 \text{ 或} N^+ \ge 9\}.$$

例6.2.3

i	1	2	3	4	5	6
x_i(A)	1.15	1.86	0.76	1.82	1.14	1.65
y_i(B)	1.00	1.90	0.90	1.80	1.20	1.70
z_i	0.15	-0.04	-0.16	0.02	-0.06	-0.05
i	7	8	9	10	11	
x_i(A)	1.92	1.01	1.12	0.90	1.40	
y_i(B)	1.95	1.02	1.23	0.97	1.52	
z_i	-0.03	-0.01	-0.11	-0.07	-0.12	

$N^+ = 2$. 因此在水平$\alpha = 0.10$下,拒绝H_0,认为这两个化验室测定的结果之间有显著的差异.

成对数据的符号秩和检验法:

符号检验只考虑$\{Z_i\}$的符号，把具体数值部分都丢弃了，信息损失过多.

符号秩和检验法：检验统计量为:

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

其中

$$V_i = \begin{cases} 1, & Z_i > 0 \\ 0, & \text{otherwise}, \end{cases}$$

R_i为Z_i在$(|Z_1|, \cdots, |Z_n|)$中的秩. 对于假设检验问题$H_0: F(x) = G(x) \longleftrightarrow H_1: F(x) \neq G(x)$,其拒绝域为

$$\{W^+ \leq d \text{ 或} W^+ \geq c\}.$$

可以证明在原假设$H_0 : F(x) = G(x)$为真时，W^+的分布为:

$$\mathsf{P}(W^+ = i) = \frac{t_n(i)}{2^n}, \quad i = 0, 1, \cdots, n(n+1)/2, \qquad (**)$$

(证明见附录)其中，$t_n(i)$为在$1, \cdots, n$中不重复地任意取若干个数(允许取0个数)其和恰好为i的方法种数;而

$$\mathsf{P}(W^+ \le d) = \mathsf{P}(W^+ \ge n(n+1)/2 - d).$$

即W^+与$n(n+1)/2 - W^+$同分布.

当n不大时，可查表

$$\mathsf{P}(W^+ \ge c) \le \alpha/2, \quad d = n(n+1)/2 - c.$$

例6.2.3 由$n = 11$, $\alpha/2 = 0.05$, 查表得$c = 53$, 故$d = 11(11+1)/2 - c = 13$.所以其拒绝域为$\{W^+ \leq 13$ 或$W^+ \geq 53\}$.

i	1	2	3	4	5	6
z_i	0.15	-0.04	-0.16	0.02	-0.06	-0.05
Rank of $\lvert z_i \rvert$	10	4	11	2	6	5
i	7	8	9	10	11	
z_i	-0.03	-0.01	-0.11	-0.07	-0.12	
Rank of $\lvert z_i \rvert$	3	1	8	7	9	

现$W^+ = 10 + 2 = 12$.由于$W^+ < 13$. 因此在水平$\alpha = 0.10$下,拒绝H_0,认为这两个化验室测定的结果之间有显著的差异.

注：当总体分布为连续型时,差值Z_i中以概率1没有等于0的.但在实际操作中,Z_i的观测值可能会出现0.此时利用上面两种方法进行检验时,往往先将那些差值为0的样本去掉.

Example

习题6-1

解 检验问题为:

$$H_0: \text{甲品种不是对乙品种的改良} \longleftrightarrow H_1: \text{甲品种是对乙品种的改良}.$$

令8块土地上甲、乙两种作物产量分别为$(X_1, \cdots, X_8)$和$(Y_1, \cdots, Y_8)$. 记

$$Z_i = X_i - Y_i, \quad i = 1, 2, \cdots, 8.$$

(1) 已知Z_i 独立同分布，服从正态分布，记为$N(\mu, \sigma^2)$. 则假设检验问题为

$$H_0: \mu \leq 0 \longleftrightarrow H_1: \mu > 0.$$

取检验统计量为

$$T(\widetilde{Z}) = \frac{\overline{Z} - 0}{S_Z/\sqrt{n}}.$$

那么拒绝域为

$$D = \{\widetilde{z} : T(\widetilde{z}) > t_\alpha(n-1)\}.$$

现$\alpha = 0.05$, $n = 8$, 查表得$t_{0.05}(7) = 1.8946$. 从而拒绝域为

$$D = \{\widetilde{z} : T(\widetilde{z}) > 1.8946\}.$$

现在由样本求得$\overline{z} = 17.375$, $s_Z^2 \approx 452.56$, 故得检验统计量的值为2.31. 由于$2.31 > 1.8946$, 即$\widetilde{z} \in D$,所以拒绝H_0, 从而认定甲品种是对乙品种的改良.

(2) 要用符号检验法或者符号秩和检验法进行检验时，需先"去零". 注意到$Z_7 = 0$, 故此处真正的样本数为7.

符号检验法 取检验统计量为$N^+ = \sum_{i=1}^{7} I\{Z_i > 0\}$, 则拒绝域为

$$D = \{N^+ \geq c\}.$$

其中临界值c由下式确定:

$$c = \min_{c'}\{0 \leq c' \leq 7, c'\text{为整数} : \sum_{k=c'}^{7} \binom{7}{k} \cdot \left(\frac{1}{2}\right)^7 \leq \alpha\}.$$

现$\alpha = 0.05$, 可得$c = 7$, 即拒绝域为$D = \{N^+ \geq 7\}$. 现在由样本求得$N^+ = 6 < 7$, 即$\tilde{z} \notin D$, 故保留H_0, 从而认定甲品种不是对乙品种的改良.

符号秩和检验法 取检验统计量为:

$$W^+ = \sum_{i=1}^{7} V_i R_i,$$

其中

$$V_i = \begin{cases} 1, & Z_i > 0 \\ 0, & \text{otherwise}, \end{cases}$$

R_i为Z_i在$(|Z_1|, \cdots, |Z_7|)$中的秩. 拒绝域为$D = \{W^+ \geq c\}$. 现$\alpha = 0.05$, $n = 7$, 查表得$c = 25$, 即拒绝域为$D = \{W^+ \geq 25\}$. 现在由样本求得$W^+ = 1 + 3 + 4 + 5 + 6 + 7 = 26 > 25$, 即$\tilde{z} \in D$, 故拒绝$H_0$, 从而认定甲品种是对乙品种的改良.

大样本情形

Theorem

假设$F(x) \equiv G(x)$. 则

$$EW^+ = \frac{n(n+1)}{4},$$

$$Var\{W^+\} = \frac{n(n+1)(2n+1)}{24}.$$

当$n \to \infty$时

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

(证明见附录)

当n充分大(一般$n>10$时即可),当$F(x)=G(x)$时,

$$(W^+)^* \sim_{\text{近似}} N(0,1).$$

『附录』W^+的精确分布(**)和渐近分布的证明：由于Z_i的分布是连续型的，$\mathsf{P}(Z_i = 0) = 0$. 在$F(x) = G(x)$下，Z_i的分布是对称的，所以

$$\mathsf{P}(V_i = 1) = \mathsf{P}(Z_i > 0) = 1/2.$$

又对于$z > 0$,

$$\begin{aligned}
&\mathsf{P}(|Z_i| < z, V_i = 1) = \mathsf{P}(|Z_i| < z, Z_i > 0) \\
=&\mathsf{P}(|-Z_i| < z, -Z_i > 0) = \mathsf{P}(|Z_i| < z, Z_i < 0) \\
=&\mathsf{P}(|Z_i| < z)/2 = \mathsf{P}(|Z_i| < z)\mathsf{P}(V_i = 1),
\end{aligned}$$

同理可得,$\mathsf{P}(|Z_i| < z, V_i = 0) = \mathsf{P}(|Z_i| < z)\mathsf{P}(V_i = 0)$. 所以$V_i$与$|Z_i|$独立，从而$V_1, \cdots, V_n$ 与$|Z_1|, \cdots, |Z_n|$独立，故

$$V_1, \cdots, V_n \quad \text{i.i.d. 且与} \quad R_1, \cdots, R_n \text{ 独立.}$$

显然对任意的实数$a_1, \cdots, a_n$ 和$(1, \cdots, n)$的置换$(k_1, \cdots, k_n)$, $\sum_j V_j a_j$ 与$\sum_j V_j a_{k_j}$ 同分布. 所以给定$R_1, \cdots, R_n$后, $W^+ = \sum_{j=1}^n V_j R_j$ 与$\sum_{j=1}^n V_j R_{(j)}$ $= \sum_{j=1}^n jV_j$ 同分布. 从而

$$W^+ \text{ 与 } \zeta = \sum_{j=1}^n jV_j \text{ 同分布.}$$

$\sum_{j=1}^n jV_j = i$ 意味着在$1, \cdots, n$中不重复地任意取若干个数(允许取0个数)其和恰好为i. (**)得证.

在$F(x)=G(x)$的条件下，

$$\begin{aligned}\mathsf{E}W^+=&\mathsf{E}\zeta=\sum_{j=1}^n j\mathsf{E}V_j=n(n+1)/4\\ \mathsf{Var}W^+=&\mathsf{Var}\zeta=\sum_{j=1}^n j^2\mathsf{Var}V_j=\sum_{j=1}^n j^2/4=\frac{n(n+1)(2n+1)}{24}.\end{aligned}$$

由于

$$
\begin{aligned}
&\frac{1}{(\mathsf{Var}\zeta)^{3/2}}\sum_{j=1}^{n} j^3\mathsf{E}|V_j-\mathsf{E}V_j|^3\\
\le&\left(\frac{24}{n(n+1)(2n+1)}\right)^{3/2}\sum_{j=1}^{n} j^3\le 12^{3/2}\frac{n^4}{n^{9/2}}\to 0,\\
&as\quad n\to\infty.
\end{aligned}
$$

由Lyapunov中心极限定理，

$$
\frac{\zeta-\mathsf{E}\zeta}{\sqrt{\mathsf{Var}\zeta}}\xrightarrow{D}N(0,1).
$$

从而

$$
(W^+)^*\xrightarrow{D}N(0,1).
$$

§两样本置换检验

我们要检验的问题是总体X和Y的效果一样, 即

$$H_0 : X \text{ 和} Y \text{ 同分布.}$$

设

$$X_1, \cdots, X_m \text{ i.i.d., 来自总体} X,$$
$$Y_1, \cdots, Y_n \text{ i.i.d., 来自总体} Y,$$

两组样本独立, 样本均值的差为$g = \overline{X} - \overline{Y}$. 将两组样本合在一起, 组成的合样本$\{Z_1, \cdots, Z_{m+n}\}$.

如果H_0成立, 那么把$Z_1, \cdots, Z_{m+n}$做任意一个置换$Z_{a_1}, \cdots, Z_{a_{m+n}}$, 这一新的样本与$Z_1, \cdots, Z_{m+n}$是同分布的. 把前$m$个看作是来自$X$的样本, 后$n$个作为来自$Y$的样本, 计算新的$g$的值,

$$g^* = \overline{X^*} - \overline{Y^*} = \frac{1}{m}\sum_{i=1}^{n} Z_{a_i} - \frac{1}{n}\sum_{j=m+1}^{m+n} Z_{a_j},$$

它等于$Z_1, \cdots, Z_{m+n}$中取m个值的平均减去剩下的n个的平均. 一共有$N = \binom{m+n}{m}$种情形: $g_1^*, \cdots, g_N^*$. 在H_0下,每种情形是等可能出现的, 概率均为$1/N$. 将其绝对值按大到小排列,

不妨仍令其为

$$|g_1^*| \geq |g_2^*| \geq \cdots \geq |g_N^*|.$$

如果H_0成立，那么$|g|$的取值不会太大，也就是说当$|g|$在上述排列中位置考靠前就应拒绝原假设. 找m使得，$|g| = |g_m^*|$. 则

$$p - value = \frac{m}{N}.$$

$\frac{m}{N} \leq \alpha$ 时拒绝原假设.

如果原假设改为X 不大于Y, 即

$$H_0 : F(x) \geq G(x).$$

则将$g_1^*, \cdots, g_N^*$的值按大到小排列, 不妨令其为

$$g_1^* \geq g_2^* \geq \cdots \geq g_N^*.$$

找m使得, $g = g_m^*$. 则

$$p - value = \frac{m}{N}.$$

§置换检验

Example

例6.2.6 为比较A, B两种施肥方法何种为优, 选择15块一样大的地, 把每块分成形状大小一样的两块小块,随机地将其中的一块分给A, 另一小块给B. 收获时得到各块的产量如下表. 要检验

$$H_0: A, B\text{的效果一样}.$$

i	1	2	3	4	5	6	7	8
x_i(A)	188	96	168	176	153	172	177	163
y_i(B)	139	163	160	160	147	149	149	122
$x_i - y_i$	49	-67	8	16	6	23	28	41
i	9	10	11	12	13	14	15	
x_i(A)	146	173	186	168	177	184	96	
y_i(B)	132	144	130	144	102	142	144	
$x_i - y_i$	14	29	56	24	75	60	-48	

$$\sum_i (x_i - y_i) = 314.$$

如果H_0成立，每块内$x_i - y_i$的值的不一样，并非由于A, B的效果不同，而是由于其两小块地的差别. 但是由于随机化的结果，每一小块有相等可能(1/2)分给A或者B. 因此，如在第一块，依随机化的结果不同，$x_1 - y_1$可以是49，也可以是−49，要看较好的那块是派给A还是B. 这样一来,这个试验的全部可能的$\sum_i (x_i - y_i)$值有2^{15}个(相等的分别算)

$$\pm(49) \pm (-67) \pm (8) \cdots \pm (60) \pm (-48),$$

实际得出的$\sum_i(x_i-y_i)=314$是2^{15} 个值中的一个. 用x_j记每个可能的值, $j=1,\cdots,2^{15}$, 共有2^{15}个. 将它们的绝对值从大到小排列

$$x_{(1)},x_{(2)},\cdots,x_{(2^{15})}.$$

在H_0成立的前提下, 这些值等可能出现,每个出现的概率为$1/2^{15}$.
找m, 使得$314=x_{(m)}$. 若314在上述排列中位置靠前, 说明原假设不成立. 从而314及比对H_0更不利的值, 在H_0成立时出现的机会只有

$$p_{314}=\frac{m}{2^{15}}.$$

这就是p-value. 本例中可以计算得到$p_{314}<0.0001$. 因此,即使在$\alpha=0.0001$显著型水平下也有理由否定H_0. 由于$314>0$, 结论是:有显著的证据表明A优于B.

§数据的初步考察与处理

游程总数检验

(用于检验某一批数据是否符合随机性原则)

$$H_0: \text{样本}\widetilde{X} = (X_1, X_2, \cdots, X_n)\text{符合随机性原则}.$$

Definition

给定一个由0和1两个元素组成的序列; 称以0为界限的一串1,或以1为界限的一串0为一个游程. 并用R记为序列的游程总数.

如: 0 1 0 1 1 1 0 1 1 0 1 0 0

此序列有5个0游程, 4个1的游程, $R = 9$.

设$x_1, x_2, \cdots, x_n$为人们依时间先后顺序抽取得到的一个样本观测值, m_e为其样本中位数. 令

$$y_i = \begin{cases} 1, & x_i \geq m_e \\ 0, & \text{otherwise}, \end{cases} \qquad i = 1, 2, \cdots, n.$$

则$\{y_1, y_2, \cdots, y_n\}$是由0和1组成的一个序列. 记其中0和1的个数分别为n_1和n_2, 则$n_1 + n_2 = n$. 显然, 对于$n > 1$,

R的最小值是2;

R的最大值是

$$\begin{cases} n_1 + n_2, & n_1 = n_2 \\ 2 \cdot \min(n_1, n_2) + 1, & \text{otherwise}. \end{cases}$$

若游程总数R过大, 即0和1呈周期性变化;
若游程总数R过小, 即序列的前一部分0(或1)占多数, 后一部分1(或0)占多数;
这样我们就认为样本受到了其它非随机性因素的干扰. 故当H_0成立时, R值不应偏大, 也不应偏小, 从而拒绝域为

$$W = \{\widetilde{X} : R \geq c \ \text{ 或 } \ R \leq d\},$$

其中

$$c = \inf_{c'}\{c' : \mathsf{P}(R \geq c' | H_0) \leq \alpha/2\},$$
$$d = \sup_{d'}\{d' : \mathsf{P}(R \leq d' | H_0) \leq \alpha/2\}.$$

Theorem

设$\{y_1, y_2, \cdots, y_n\}$是由0和1组成的一个序列. 记其中0和1的个数分别为n_1和n_2, 则在H_0成立时,

$$P(R = 2k) = 2\binom{n_1 - 1}{k - 1} \cdot \binom{n_2 - 1}{k - 1} \Big/ \binom{n_1 + n_2}{n_1},$$

$$P(R = 2k + 1) = \frac{\binom{n_1-1}{k-1} \cdot \binom{n_2-1}{k} + \binom{n_1-1}{k} \cdot \binom{n_2-1}{k-1}}{\binom{n_1+n_2}{n_1}}.$$

n_1, n_2较小时, 可查表

n_1, n_2较大时, 可以利用其近似正态性

大样本情形：如果$n_0, n_1 \to \infty$，且$n_0/n_1 \to C$，则在H_0成立时，

$$\left(R - \frac{2n_0n_1}{n_0+n_1}\right) \Big/ \left[\frac{2n_0n_1}{n_0+n_1}\frac{1}{\sqrt{n_0+n_1}}\right] \xrightarrow{D} N(0,1).$$

游程检验也可以用于两个分布函数的检验

设$X \sim F(x)$, $\widetilde{X} = (X_1, X_2, \cdots, X_{n_1})$为来自$X$的样本;
设$Y \sim G(y)$, $\widetilde{Y} = (Y_1, Y_2, \cdots, Y_{n_2})$为来自$Y$的样本.

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

将$\widetilde{X}$, $\widetilde{Y}$组合成一个合样本,由小到大排列, 记为$Z_1 \leq Z_2 \leq \cdots \leq Z_{n_1+n_2}$. 令

$$W_i = \begin{cases} 1, & 若Z_i是来自\widetilde{Y}, \\ 0, & 若Z_i是来自\widetilde{X}. \end{cases}$$

当H_0成立时, $\{X_i\}$与$\{Y_j\}$应能充分混和, 从而$\{W_i\}_{i=1}^{n_1+n_2}$的游程应较大, 所以拒绝域为$W = \{R \leq c\}$.