非参数统计

王学钦

中国科学技术大学
管理学院

2021年6月9日





1 经验分布函数

2 经典秩检验

3 秩统计量的大样本性质



1 经验分布函数

2 经典秩检验

3 秩统计量的大样本性质



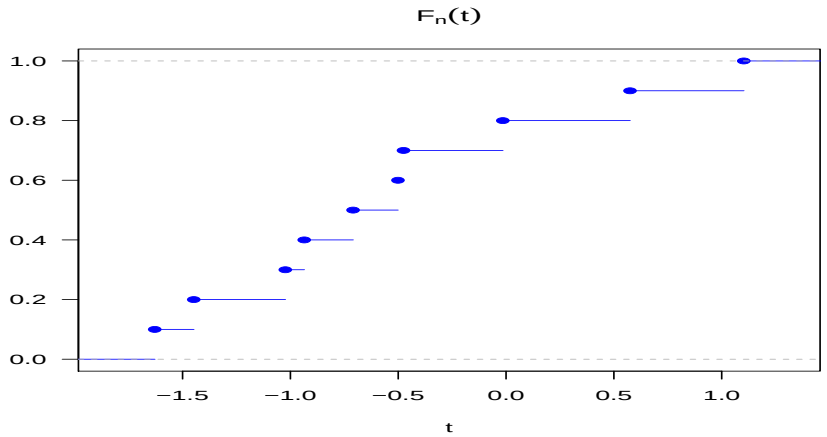
经验分布与秩方法

经典的非参数统计学是建立在经验分布函数和秩的基础上。经验分布函数、秩乃至于经验似然的思想对整个统计学来说也是至关重要的，而且它们的计算及其简单。

定义1.1

给定一组i.i.d.样本$X_1, X_2, \ldots, X_n \sim F$，$F$ 的经验分布函数(EDF)F_n定义为：

$$F_n(t) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \le t).$$



图：标准正态分布的经验分布函数



经验分布函数的性质

显然, $EF_n(t) = F(t),\ \ Var(F_n(t)) = F(t)(1-F(t)),\ \forall t \in R.$ 且:

- 由大数定理可知:

$$EF_n(t) \xrightarrow{a.s.} EI(X \le t) = F(t),\ \forall t \in R.$$

- 由中心极限定理可知:

$$\sqrt{n}\frac{F_n(t) - F(t)}{\sqrt{F(t)(1-F(t))}} \xrightarrow{L} N(0,1),\ \forall t \in R.$$

- **格里文科定理:**

$$P(\lim_{n\to\infty} \sup_{-\infty<t<\infty} |F_n(t) - F(t)| = 0) = 1.$$



单样本的Kolomogorov-Smirnov检验

设$X_1, X_2, \ldots, X_n$是来自连续分布F的独立随机样本，对于检验问题:

$$H_0 : F = F_0, \quad v.s. \quad F \neq F_0,$$

其中F_0是已知的连续分布函数.

设F_n为F的经验分布函数，则可取Kolomogorov-Smirnov统计量:

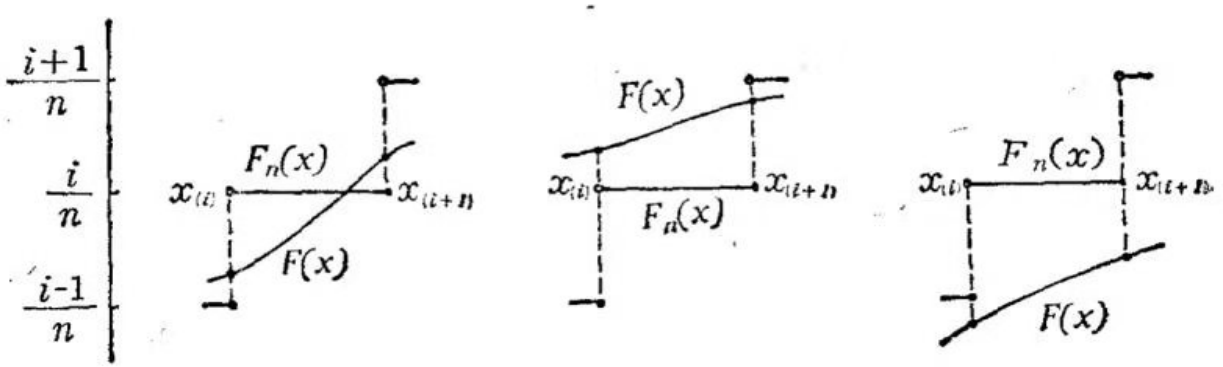
$$K_n := \sup_{-\infty<x<\infty} |F_n(x) - F_0(x)|$$

为检验统计量，且K_n较大时拒绝H_0。



K_n的计算公式可以简化为：

$$K_n = \max_{1\le i\le n}\left\{\max\left\{|F_0(X_{(i)}) - \frac{i-1}{n}|,\ |F_0(X_{(i)}) - \frac{i}{n}|\right\}\right\}$$

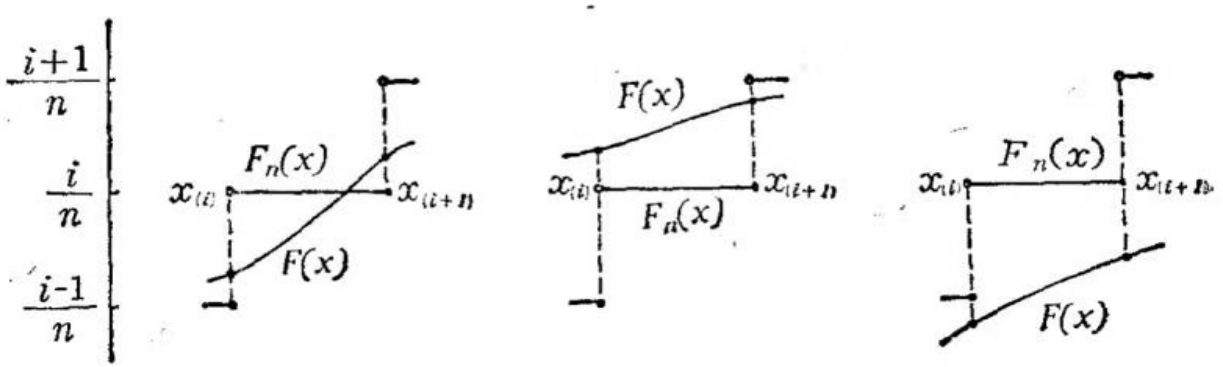


图：K_n计算示意图



K_n的计算公式可以简化为：

$$K_n = \max_{1\le i\le n}\left\{\max\left\{|F_0(X_{(i)}) - \frac{i-1}{n}|,\ |F_0(X_{(i)}) - \frac{i}{n}|\right\}\right\}$$



图：K_n计算示意图

定理1.1 (Kolomogorov-Smirnov定理)

假设理论分布F_0为连续分布函数，则在H_0成立及$n \to \infty$时，有

$$P(\sqrt{n}K_n < \lambda) \longrightarrow \sum_{j=1}^{\infty}(-1)^j \exp(-2j^2\lambda^2),\ \ \forall \lambda > 0.$$

单样本的Cramér-von Mises检验

对于检验问题:

$$H_0 : F = F_0, \quad v.s. \quad F \neq F_0.$$

亦可取Cramér-von Mises型统计量:

$$\omega_n^2 := n \int_{-\infty}^{\infty} (F_n(x) - F_0(x))^2 \psi(F_0(x)) dF_0(x) = n \int_0^1 (G_n(t) - t)^2 \psi(t) dt$$

为检验统计量，其中，$G_n(\cdot)$为$U(0,1)$的经验分布，$\psi(\cdot)$为非负权函数。

- 取$\psi(t) = 1$，得到Cramér-von Mises统计量:

$$W_n^2 = n \int_{-\infty}^{\infty} (F_n(x) - F_0(x))^2 dF_0(x) = n \int_0^1 (G_n(t) - t)^2 dt;$$

- 取$\psi(t) = [t(1-t)]^{-1}$时，可得到Anderson-Darling 统计量:

$$A_n^2 = n \int_{-\infty}^{\infty} \frac{(F_n(x) - F_0(x))^2}{F_0(x)(1 - F_0(x))} dF_0(x) = n \int_0^1 \frac{(G_n(t) - t)^2}{t(1-t)} dt.$$



双样本的Kolomogorov-Smirnov检验与Cramér-von Mises 检验

假设变量X, Y为相互独立的随机变量，$X \sim F, Y \sim G$，
$X_1, X_2, \ldots, X_m$及$Y_1, Y_2, \ldots, Y_n$分别为来自X和Y的简单随机样本，考虑检验问题：

$$H_0 : F = G, \quad v.s. \ F \neq G.$$

相应的Kolomogorov-Smirnov检验与Cramér-von Mises检验如下：

1) Kolmogorov-Smirnov 检验：

$$K_{m,n} := \sqrt{\frac{mn}{m+n}} \sup_{t \in \mathbb{R}} |\hat{F}_X(t) - \hat{G}_Y(t)|.$$

2) Cramér-von Mises检验：

$$W^2_{m,n} := \frac{mn}{m+n} \int_{-\infty}^{\infty} (\hat{F}_X(t) - \hat{G}_Y(t))^2 d\hat{H}(t).$$

其中$\hat{F}_m, \hat{G}_n$分别为F, G的经验分布函数，$(m+n)\hat{H}(t) = m\hat{F}_X(t) + n\hat{G}_Y(t)$。



独立性检验

给定一组i.i.d.样本：$(X_1, Y_1), \ldots, (X_n, Y_n) \sim H$。假设其边际分布函数分别是$F$和$G$，而且$H_n, F_n$和$G_n$分别是$H, F$和$G$的经验分布函数。对于独立性检验问题：

$$H_0 : H(x, y) = F(x)G(y), \quad v.s. \quad H_1 : H(x, y) \neq F(x)G(y).$$

可选以下检验统计量。

- Hoeffding criterion:

$$T_n := \int (H_n(x, y) - F_n(x)G_n(y))^2 dH_n(x, y).$$

- Spearman's rho:

$$\rho_n := 1 - \frac{n \int (F_n(x) - G_n(y))^2 dH_n(x, y)}{n^2 - 1}.$$



秩方法

秩方法始于1904年Spearman提出的秩相关方法，但受到人们越来越大的重视是在1945年Wilcoxon提出了秩和检验方法之后. 秩和检验方法用于两个总体位置参数的比较问题，它在应用上有很大的意义，同时它还极大地推动了往后的秩方法的发展。

定义1.2

不妨假设独立同分布的样本$X_1,\ldots,X_n$的总体X是连续型随机变量，从而以概率1保证，样本$X_1,\ldots,X_n$互不相等。将样本从小到大排列成

$$X_{(1)} < X_{(2)} < X_{(3)} < \ldots < X_{(n)}.$$

*若$X_i = X_{(R_i)}$，则称$X_i (i = 1, 2, \ldots, n)$在$X_1, X_2, \ldots, X_n$中的秩为$R_i$。$R = (R_1, R_2, \ldots, R_n)$以及$R$任意的可测函数称为**秩统计量**。*



秩与经验分布函数(EDF)关系

给定一组i.i.d.样本$X_1, X_2, \ldots, X_n \sim F$，$F_n$ 是F 的EDF:

$$F_n(t) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \le t).$$

记R_i为X_i的秩，$i = 1, 2, \ldots, n$，则秩与EDF有如下关系：

$$R_j = nF_n(X_j) = \sum_{i=1}^{n} I(X_i \le X_j).$$



由样本的独立同分布性可知有

$$P(R_i = r) = 1/n, \qquad r = 1, 2, \ldots, n$$

及

$$P(R_i = r_1, R_j = r_2) = \frac{1}{n(n-1)}, \qquad r_1 \neq r_2, i \neq j.$$

从而有以下结论

定理1.2

$$E(R_i) = \frac{n+1}{2},$$
$$Var(R_i) = \frac{n^2-1}{12}$$
$$Cov(R_i, R_j) = -\frac{n+1}{12}.$$

证明：

$$
\begin{aligned}
E(R_i) &= (1+2+\ldots+n)\times\frac{1}{n} = \frac{n+1}{2},\\
Var(R_i) &= E(R_i^2) - (E(R_i))^2\\
&= (1^2+2^2+\ldots+n^2)\times\frac{1}{n} - (\frac{n+1}{2})^2\\
&= \frac{(n+1)(2n+1)}{6} - \frac{(n+1)^2}{4}\\
&= \frac{n^2-1}{12},\\
Cov(R_i,R_j) &= ER_iR_j - (ER_i)\times(ER_j)\\
&= \frac{1}{n(n-1)}\sum_{i\neq j} i\times j - \frac{(n+1)^2}{4}\\
&= \frac{1}{n(n-1)}(\sum_{i=1,j=1}^{n} i\times j - \sum_{i=1}^{n} i^2) - \frac{(n+1)^2}{4}\\
&= \frac{1}{n(n-1)}\Big(\frac{n^2(n+1)^2}{4} - \frac{n(n+1)(2n+1)}{6}\Big) - \frac{(n+1)^2}{4}\\
&= -\frac{n+1}{12}.
\end{aligned}
$$



秩得分函数

定义计分函数为$a(r), r = 1, 2, \ldots, n$，在$X_i$的秩为$R_i$时，将$X_i$的得分定义为$a(R_i)$，则类似有以下结论：

定理1.3

$$E(a(R_i)) = \overline{a},$$

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

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

其中$\overline{a} = \frac{1}{n}\sum_{i=1}^{n} a(i)$。



证明：

$$
\begin{aligned}
E(a(R_i)) &= (a(1)+a(2)+\ldots+a(n))\times\frac{1}{n}=\bar{a},\\
Var(a(R_i)) &= E(a(R_i)-E(a(R_i)))^2\\
&= E(a(R_i)-\bar{a})^2\\
&= \frac{1}{n}\sum_{i=1}^{n}(a(i)-\bar{a})^2,\\
Cov(a(R_i),a(R_j)) &= E(a(R_i)-Ea(R_i))(a(R_j)-Ea(R_j))\\
&= \frac{1}{n(n-1)}\sum_{i\neq j}(a(i)-\bar{a})(a(j)-\bar{a})\\
&= \frac{1}{n(n-1)}\Big[\sum_{i=1,j=1}^{n}(a(i)-\bar{a})(a(j)-\bar{a})-\sum_{i=1}^{n}(a(i)-\bar{a})^2\Big]\\
&= \frac{1}{n(n-1)}\Big[\Big(\sum_{i=1}^{n}(a(i)-\bar{a})\Big)^2-\sum_{i=1}^{n}(a(i)-\bar{a})^2\Big]\\
&= -\frac{1}{n(n-1)}\sum_{i=1}^{n}(a(i)-\bar{a})^2.
\end{aligned}
$$



1 经验分布函数

2 经典秩检验

3 秩统计量的大样本性质



1.Wilcoxon符号秩检验

假设i.i.d.样本$X_1, X_2, \ldots, X_n$来自总体F. 考虑对称中心θ的检验问题：

$$H_0 : \theta = 0 \quad v.s. \quad H_1 : \theta > 0. \tag{1}$$

选取符号秩和检验统计量：

$$W^+ := \sum_{i=1}^{n} R_i I(X_i > 0),$$

其中R_i为$|X_i|$在$(|X_1|, |X_2|, \ldots, |X_n|)$中的秩. 显然，$W^+$综合了样本数据是正数还是负数的信息及样本数据值的大小的信息。



零假设下W^+的期望与方差

为了得到零假设下W^+的期望与方差，先叙述一个定理理：

定理2.1

假设连续随机变量X的分布关于原点0对称，其分布函数为$F(x)$，令$u(x) = I(x > 0)$，则$|X|$ 与$u(X)$ 相互独立。

证明： 因为对任意$x \in R$，当X的分布关于原点0对称有

$$\begin{aligned} P(u(X) = 1, |X| \le x) &= P(X > 0, |X| \le x) \\ &= P(0 \le X \le x) \\ &= F(x) - F(0) \\ &= F(x) - \frac{1}{2} \\ &= \frac{1}{2}(2F(x) - 1) \\ &= P(u(X) = 1) \cdot P(|X| \le x). \end{aligned}$$



定理2.2 (零假设下W^+的期望与方差)

假设i.i.d.样本$X_1, X_2, \ldots, X_n$来自总体$X \sim F$,

$$W^+ := \sum_{i=1}^{n} R_i I(X_i > 0),$$

其中R_i为$|X_i|$在$(|X_1|, |X_2|, \ldots, |X_n|)$中的秩.则当$X$分布关于$0$点对称时有:

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

证明: 由前面定理2.1可知, 在X分布关于0点对称条件下有$(|X_1|, \ldots, |X_n|)$与$(u(X_1), \ldots, u(X_n))$相互独立, 而$R_1, \ldots R_n$为$(|X_1|, \ldots, |X_n|)$的函数, 因此, 可知$(u(X_1), \ldots, u(X_n))$与$(R_1, \ldots, R_n)$相互独立, 记$u_i = u(X_i)$, 则有:



$$E(W^+) = E(\sum_{i=1}^{n} R_i u_i) = nE(R_1 u_1) = nE(R_1)E(u_1) = n \cdot \frac{1}{2} \cdot \frac{n+1}{2} = \frac{n(n+1)}{4},$$

$$Var(W^+) = nVar(u_1 R_1) + n(n-1)Cov(u_1 R_1, u_2 R_2).$$

而其中

$$\begin{aligned} Var(u_1 R_1) &= E(u_1)^2 E(R_1)^2 - (E(u_1)E(R_1))^2 \\ &= \frac{1}{2}\frac{(n+1)(2n+1)}{6} - (\frac{n+1}{4})^2 \\ &= \frac{(n+1)(5n+1)}{48}, \end{aligned}$$

$$\begin{aligned} Cov(u_1 R_1, u_2 R_2) &= E(u_1)E(u_2)E(R_1 R_2) - (E(u_1 R_1)E(u_2 R_2))^2 \\ &= \frac{1}{4} \cdot \frac{1}{n(n-1)} \sum_{i \neq j} ij - \frac{(n+1)^2}{16} \\ &= \frac{1}{4}\frac{(1+2+\ldots+n)^2 - (1^2+2^2+\ldots+n^2)}{n(n-1)} - \frac{(n+1)^2}{16} \\ &= -\frac{n+1}{48}. \end{aligned}$$



因此，

$$Var(W^+) = n\frac{(n+1)(5n+1)}{48} + n(n-1)\frac{-(n+1)}{48}$$
$$= \frac{n(n+1)(2n+1)}{24}.$$

定理2.3 (零假设下W^+的精确分布)

如果总体的分布是关于原点0对称的连续分布，则W^+的概率分布为：

$$P(W^+ = d) = P(\sum_{i=1}^{n} R_i u_i = d) = \frac{t_{n,d}}{2^n}, \quad d = 0, 1, 2, \ldots, \frac{n(n+1)}{2}.$$

其中$t_{n,d}$表示从$1, 2, \ldots, n$这n个数中任取若干个数（包括一个都不取），其和恰为d的取法种数。

从而可知检验(1)的p值为

$$p = \sum_{d \geq W^+} \frac{t_{n,d}}{2^n}.$$



2. Wilcoxon-Mann-Whitney检验

假设样本$X_1, X_2, \ldots, X_m$和$Y_1, Y_2, \ldots, Y_n$分别来自相互独立的连续随机变量总体X和Y。分别记X和Y的分布函数为F和G,且有$G(y) = F(y-a)$. 考虑检验问题:

$$H_0 : a = 0 \quad v.s. \quad H_1 : a > 0. \tag{2}$$

选取秩和检验统计量:

$$W_y := \sum_{j=1}^{n} R_j,$$

其中R_j为$Y_j (j = 1, 2, \ldots, n)$在联合样本$(X_1, X_2, \ldots, X_m, Y_1, Y_2, \ldots, Y_n)$中的秩。



定理2.4 (零假设下W_y的期望与方差)

当假设$H_0: a = 0$即$F = G$成立时，有

$$E(W_y) = n(m+n+1)/2,$$
$$Var(W_y) = mn(m+n+1)/12.$$

证明: 由于当$a = 0$，即$F = G$成立时，$X_1, X_2, \ldots, X_m, Y_1, Y_2, \ldots, Y_n$为来自同一分布的i.i.d. 样本，因此由定理1.2可知

$$E(R_i) = \frac{m+n+1}{2}, \quad Var(R_i) = \frac{(m+n)^2-1}{12}, \quad Cov(R_i, R_j) = -\frac{m+n+1}{12}.$$

从而可知

$$E(W_y) = E(\sum_{j=1}^{n} R_j) = nE(R_1) = n(m+n+1)/2,$$

$$Var(W_y) = Var(\sum_{j=1}^{n} R_j) = \sum_{j=1}^{n} Var(R_j) + n(n-1)\sum_{i \neq j} Cov(R_i, R_j)$$
$$= n\frac{(m+n)^2-1}{12} - n(n-1)\frac{m+n+1}{12} = \frac{mn(m+n+1)}{12}.$$



定理2.5 (零假设下W_y的精确分布)

如果$F = G$同为连续分布，则W_y的概率分布为：

$$P(W_y = d) = P(\sum_{j=1}^{n} R_j) = \frac{t_{m,n}(d)}{C_{m+n}^{n}}, \quad d = \frac{n(n+1)}{2}, \ldots, \frac{n(n+1)}{2} + mn.$$

其中$t_{m,n}(d)$表示从$1, 2, \ldots, m+n$这$m+n$个数中任取n个数，其和恰为d的取法种数。

从而可知检验(2)的p值为：

$$p = \sum_{d \geq W_y} \frac{t_{m,n}(d)}{C_{m+n}^{n}}.$$



3. Spearman秩相关系数

假设有二元随机向量(X,Y)，其联合分布为F，边缘分布分别为F_X, F_Y。
设$\{(X_1,Y_1),(X_2,Y_2),\ldots,(X_n,Y_n)\}$是来自$F$的样本，考虑独立性检验问题：

$$H_0: F(x,y) = F_X(x)F_Y(y) \ \ v.s. \ \ H_1: F(x,y) \neq F_X(x)F_Y(y)$$

记$R^X := (R_1^X, R_2^X, \ldots, R_n^X)$，$R^Y := (R_1^Y, R_2^Y, \ldots, R_n^Y)$，其中$R_i^X (i = 1,2,\ldots,n)$为$X_i$的秩，$R_i^Y (i = 1,2,\ldots,n)$为$Y_i$的秩，则$(X,Y)$的样本Spearman秩相关系数为

$$r_S := r_{R^X R^Y} = \frac{\sum_{i=1}^n (R_i^X - \overline{R^X})(R_i^Y - \overline{R^Y})}{\sqrt{\sum_{i=1}^n (R_i^X - \overline{R^X})^2}\sqrt{\sum_{i=1}^n (R_i^Y - \overline{R^Y})^2}} = 1 - \frac{6\sum_{i=1}^n d_i^2}{n(n^2-1)}$$

其中$d_i = R_i^X - R_i^Y$，$i = 1,2,\ldots,n$。
Spearman秩相关系数为秩的Pearson相关系数，它反映了X,Y之间单调关系的强弱，显然有

$$-1 \leq r_S \leq 1,$$

且$r_S = 1$相且仅当两变量存在严格单增的函数关系，$r_S = -1$相且仅当两变量存在严格单减的函数关系。



定理2.6 (零假设下r_S的期望与方差)

若X, Y独立，则有

$$E(r_S) = 0, \quad Var(r_S) = \frac{1}{n-1}.$$

证明： 注意到r_S的期望与方差取决于$\sum_i d_i^2$，而

$$\sum_i d_i^2 = \sum_i (R_i^X - R_i^Y)^2 = \frac{n(n+1)(2n+1)}{3} - 2\sum_i R_i^X R_i^Y.$$

在零假设下，$(R_1^X, R_2^X, \ldots, R_n^X)$与$(R_1^Y, R_2^Y, \ldots, R_n^Y)$ 独立，从而有

$$\sum_i R_i^X R_i^Y \overset{D}{=} \sum_i i R_i^X$$

由定理1.2 可知：

$$E(\sum_i d_i^2) = \frac{n(n+1)(2n+1)}{3} - 2\sum_i iE(R_i^X) = \frac{n(n+1)(2n+1)}{3} - 2\frac{(n+1)^2}{4}$$

$$= \frac{n(n^2-1)}{6}$$

$$\Longrightarrow E(r_S) = 0.$$



$$
\begin{aligned}
Var(\sum_i d_i^2) &= 4Var(\sum_i iR_i^X) = 4\sum_i i^2 D(R_i^X) + 4\sum_{i\neq j} ijCov(R_i^X, R_j^Y) \\
&= 4\frac{n^2-1}{12}\sum_i i^2 - 4\frac{n+1}{12}\sum_{i\neq j} ij \\
&= \frac{(n^2-1)n(n+1)(2n+1)}{18} - \frac{n(n+1)^2(3n+2)(n-1)}{36} \\
&= \frac{n^2(n+1)^2(n-1)}{36}.
\end{aligned}
$$

$$
\begin{aligned}
\Longrightarrow Var(r_S) &= \frac{36}{n^2(n^2-1)^2} Var(\sum_i d_i^2) \\
&= \frac{36}{n^2(n^2-1)^2} \times \frac{n^2(n+1)^2(n-1)}{36} \\
&= \frac{1}{n-1}.
\end{aligned}
$$



4. Kendall'sτ 系数

假设有二元随机向量(X,Y)，其联合分布为F，边缘分布分别为F_X, F_Y。$\{(X_1,Y_1),(X_2,Y_2),\ldots,(X_n,Y_n)\}$是来自$F$的样本。$X,Y$ 的Kendall'sτ系数定义为

$$\tau = P((X_1-X_2)(Y_1-Y_2)>0) - P((X_1-X_2)(Y_1-Y_2)<0).$$

相应的样本无偏估计为

$$\tau_n = \frac{1}{C_n^2}\sum_{i<j} sgn(X_i-X_j)(Y_i-Y_j),$$

其中$sgn(x) := I(x>0) - I(x<0)$。



5. Kruskal-Wallis单因素方差分析

假定k个独立样本有连续分布函数$F_1, F_2, \ldots, F_k$, X_{ij}, $i = 1, 2, \ldots, k$, $j = 1, 2, \ldots, n_i$为来自第i个总体的第j观测，数据格式如下：

表：完全随机设计数据形态

观测次数	总体1	总体2	$\cdots$	总体k
1	X_{11}	X_{21}	$\cdots$	X_{k1}
2	X_{12}	X_{22}	$\cdots$	X_{k2}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	
n_i	X_{1n_1}	X_{2n_2}	$\cdots$	X_{kn_k}

考虑检验问题：

$$H_0: F_1 = F_2 = \ldots = F_k, \quad v.s. \quad H_1: F_i(x) = F(x + \theta_i), \ i = 1, 2, \ldots, k. \tag{3}$$



对表1中观测全部混在一起，记R_{ij}为观测X_{ij}在全样本中的秩，从而可得与表1相应的秩表：

表：完全随机设计数据的秩

观测次数	总体1	总体2	$\cdots$	总体k
1	R_{11}	R_{21}	$\cdots$	R_{k1}
2	R_{12}	R_{22}	$\cdots$	R_{k2}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	
n_i	R_{1n_1}	R_{2n_2}	$\cdots$	R_{kn_k}
秩和	$R_{1\cdot}$	$R_{2\cdot}$	$\cdots$	$R_{k\cdot}$

其中$R_{i\cdot} = \sum_{j=1}^{n_i} R_ij$。记$n = n_1 + n_2 + \ldots + n_k$,

$$\overline{R}_{i\cdot} = R_{i\cdot}/n_i,$$

$$R_{\cdot\cdot} = \sum_{i,j} R_{ij} = 1 + 2 + \ldots + n = n(n+1)/2.$$



定理2.7 (零假设下平均秩$\overline{R}_{i\cdot}$的期望与方差)

当k个总体之间分布无差异时，有

$$E(\overline{R}_{i\cdot}) = \frac{n+1}{2}, \quad Var(\overline{R}_{i\cdot}) = \frac{(n-n_i)(n+1)}{12n_i}, \quad Cov(\overline{R}_{i\cdot}, \overline{R}_{j\cdot}) = -\frac{n+1}{12}.$$

显然，

$$\sum_i \sum_j R_{ij}^2 = 1^2 + 2^2 + \cdots + n^2 = \frac{n(n+1)(2n+1)}{6}.$$

因此可得混合数据各秩的总离差平方和为：

$$\begin{aligned} SST &= \sum_{i=1}^{k} \sum_{j=1}^{n_i} (R_{ij} - R_{\cdot\cdot})^2 \\ &= \sum_i \sum_j R_{ij}^2 - R_{\cdot\cdot}^2/n \\ &= \frac{n(n+1)(2n+1)}{6} - \frac{1}{4} n(n+1)^2 \\ &= n(n^2-1)/12. \end{aligned}$$



其总方差估计值（总均方）为

$$MST = SST/(n-1) = n(n+1)/12.$$

各组样本之间的离差平方和为：

$$SST = \sum_{i=1}^{k} n_i(\overline{R}_{i\cdot} - R_{\cdot\cdot})^2 = \sum_{i=1}^{k} R_{i\cdot}^2/n_i - R_{\cdot\cdot}^2/n = \sum_{i=1}^{k} R_{i\cdot}^2/n_i - n(n+1)^2/4.$$

检验(3)的Kruskal-Wallis的H值定义为：

$$\begin{aligned} H = SST/MST &= \frac{\sum_{i=1}^{k} R_{i\cdot}^2/n_i - n(n+1)^2/4}{n(n+1)/12} \\ &= \frac{12}{n(n+1)} \sum_{i=1}^{k} R_{i\cdot}^2/n_i - 3(n+1). \end{aligned}$$

当H较大时，拒绝H_0。在零假设下，H近似服从自由度为$k-1$的$\chi^2(k-1)$分布



6. 正态记分检验

- **基本思想:** 利用由秩对应到不同分布的分位点之间的单调关系不变的性质，以正态分布作为参照，将秩转化成相应的正态分布的分位点，从而将依赖于秩的检验，化为对分位点大小的检验。
- 假设$X_1, X_2, \cdots, X_n$是来自连续分布$F(\theta, x)$的简单随机样本，记M为分布F的中位数，$R_1, R_2, \ldots, R_n$为相应的秩，对于检验问题:

$$H_0: M = M_0, \quad v.s.\ M \neq M_0, \tag{4}$$

其中M_0为某一常数。

若记r_i为$|X_i - M_0|$的秩，则其相应的符号正态记分为

$$s_i = \Phi^{-1}\Big(\frac{1}{2}\Big[1 + \frac{r_i}{n+1}\Big]\Big) sgn(X_i - M_0).$$

其中$sgn(x) = I(x > 0) - I(x < 0)$, $\Phi^{-1}(\cdot)$为标准正态分布函数的反函数。



记

$$W = \sum_{i=1}^{n} s_i,$$

显然，W为正负记分和之差，对于检验(4)可选正态记分检验统计量：

$$T = \frac{W}{\sqrt{\sum_{i=1}^{n} s_i^2}},$$

且T较大时，拒绝H_0。

在总体分布接近正态或大样本情形下，T近似服从标准正态分布。

注： 若记$\Phi_+(x) = 2\Phi(x) - 1 = P(|X| \le x)$，则有

$$\Phi_+^{-1}(\frac{i}{n+1}) = \Phi^{-1}\Big(\frac{1}{2}\Big[1 + \frac{i}{n+1}\Big]\Big),$$

大约等于$E|X|_{(i)}$，即它和期望正态记分相近。



1 经验分布函数

2 经典秩检验

3 秩统计量的大样本性质



定义3.1

令$X_1, X_2, \ldots, X_n$ 是来自分布$F(\theta, x)$的简单随机样本，$\psi(x_1, x_2, \ldots, x_k)$ 是关于$(x_1, x_2, \ldots, x_k)$ 排列对称的k元函数，称具有形式：

$$U_n := \frac{1}{C_n^k} \sum_{(n,k)} \psi(X_{i_1}, X_{i_2}, \cdots, X_{i_k}) \tag{5}$$

的统计量U_n为以ψ为核的自由度为k 的U统计量，其中$\sum_{(n,k)}$ 表示对$1, 2, \cdots, n$的所有组合$(i_1, \cdots, i_k)$ 求和。

记

$$\psi_c(x_1, x_2, \cdots, x_c) := E(\psi(x_1, x_2, ..., x_c, X_{c+1}, ..., X_k)$$
$$\sigma_c^2 := Var(\psi_c(X_1, X_2, \cdots, X_c))$$

U统计量的H-分解公式是研究U统计量渐近理论的出发点，它的表述如下：



定理3.1 (*U*-统计量的*H*-分解)

记$\theta = E\psi(X_1, X_2, \cdots, X_k)$，且令

$$h^{(1)}(x_1) := \psi_1(x_1) - \theta,$$
$$h^{(2)}(x_1, x_2) := \psi_2(x_1, x_2) - (h^{(1)}(x_1) + h^{(1)}(x_2)) - \theta,$$
$$h^{(c)}(x_1, x_2, \cdots, x_c) := \psi_c(x_1, x_2, \cdots, x_c) - \sum_{j=1}^{c-1} \sum_{(c,j)} h^{(j)}(x_{i_1}, x_{i_2}, \cdots, x_{i_j}) - \theta, c = 3, 4, \cdots, k.$$

我们有：$Eh^{(j)}(X_1, X_2, \cdots, X_j) = 0, j = 1, 2, \cdots, k$，及

$$U_n = \theta + \sum_{j=1}^{k} C_k^j H_n^{(j)} \tag{6}$$

其中$H_n^{(j)} = \frac{1}{C_n^j} \sum_{(n,j)} h^{(j)}(X_{s_1}, X_{s_2}, \cdots, X_{s_j})$ 为以$h^{(j)}(X_1, X_2, \cdots, X_j)$为核的自由度为$j$的$U$统计量。

证明：

$$
\begin{aligned}
U_n &= \frac{1}{C_n^k}\sum_{(n,k)}\psi(X_{i_1},X_{i_2},\cdots,X_{i_k}) \\
&= \frac{1}{C_n^k}\sum_{(n,k)}(\theta+\sum_{j=1}^{k-1}\sum_{(k,j)}h^{(j)}(X_{s_1},X_{s_2},\cdots,X_{s_j})+h^{(k)}(X_{i_1},X_{i_2},\cdots,X_{i_k})) \\
&= \theta+\frac{1}{C_n^k}\sum_{(n,k)}\sum_{j=1}^{k}\sum_{(k,j)}h^{(j)}(X_{s_1},X_{s_2},\cdots,X_{s_j}) \\
&= \theta+\sum_{j=1}^{k}\frac{1}{C_n^k}\sum_{(n,k)}\sum_{(k,j)}h^{(j)}(X_{s_1},X_{s_2},\cdots,X_{s_j}) \\
&= \theta+\sum_{j=1}^{k}\frac{1}{C_n^k}C_{n-j}^{k-j}\sum_{(n,j)}h^{(j)}(X_{s_1},X_{s_2},\cdots,X_{s_j}) \\
&= \theta+\sum_{j=1}^{k}C_k^j\frac{1}{C_n^j}\sum_{(n,j)}h^{(j)}(X_{s_1},X_{s_2},\cdots,X_{s_j}) \qquad (\text{利用等式} C_n^k C_k^j = C_{n-j}^{k-j}C_n^j) \\
&= \theta+\sum_{j=1}^{k}C_k^j H_n^{(j)}.
\end{aligned}
$$



U统计量的H-分解(6)说明任何一个自由度为k的U统计量都可以表示成k个U统计量的线性组合，这些U 统计量自由度分别为$1,2,\cdots,k$。下面的结论说是了这些U统计量之间的特殊关系。

定理3.2

令$h_c^{(j)}(x_1,x_2,\cdots,x_c) := Eh^{(j)}(x_1,x_2,\cdots,x_c,X_{c+1},\cdots,X_j)$，则对任意$c=1,2,\cdots,j-1, j=1,2,\cdots,k$ 我们有:

$$h_c^{(j)}(x_1,x_2,\cdots,x_c)=0.$$



性质3.1

由定理3.2结论可知有下面三点成立：

- 记$\delta_j^2 := Var(h^{(j)}(X_1, X_2, \cdots, X_j))$, 则

$$Var(H_n^{(j)}) = \frac{1}{C_n^j} Var(h^{(j)}(X_1, X_2, \cdots, X_c)) = \frac{1}{C_n^j}\delta_j^2 = O(n^{-j}).$$

- 对于$j < j'$, 令S_1 与S_2 分别为$\{1, 2, \cdots, n\}$ 的包含j个与j'个元素的子集，则有

$$Cov(h^{(j)}(S_1), h^{(j')}(S_2)) = E[E(h^{(j)}(S_1)h^{(j')}(S_2)|S_1 \cap S_2)] = 0,$$

- 对任意$j = 1, 2, \cdots, k$，若S_1 与S_2 为$\{1, 2, \cdots, n\}$的包含j个元素的两个不同子集，则同样有: $Cov(h^{(j)}(S_1), h^{(j)}(S_2)) = 0$。

性质3.1说明U统计量组$\{H_n^{(1)}, H_n^{(2)}, \cdots, H_n^{(k)}\}$，它们是两两互不相关的，且对于同一个$U$ 统计量$H_n^{(j)}$的核$h^{(j)}$而言，只要$S_1 \neq S_2$，$h^{(j)}(S_1)$、$h^{(j)}(S_2)$亦是不相关的。



定理3.3 (单样本U-统计量的渐近分布)

对于U统计量$U_n = \frac{1}{C_n^k}\sum_{(n,k)}\psi(X_{i_1}, X_{i_2}, \cdots, X_{i_k})$，记$\theta = E\psi(X_1, X_2, \cdots, X_k)$，及$\psi_1(X_1) := E(\psi(X_1, X_2, \ldots, X_n)|X_1)$, 如果$\sigma_1^2 = Var(\psi_1(X_1)) < \infty$，则有

$$\sqrt{n}(U_n - \theta) \xrightarrow{L} N(0,\ k^2\sigma_1^2).$$

证明： 根据H-分解公式可知:

$$U_n = \theta + kH_n^{(1)} + \sum_{j=2}^{k} C_k^j H_n^{(j)} \Longrightarrow \sqrt{n}(U_n - \theta) = \sqrt{n}kH_n^{(1)} + \sqrt{n}(\sum_{j=2}^{k} C_k^j H_n^{(j)}),$$

其中

$$Var(\sum_{j=2}^{k} C_k^j H_n^{(j)}) = \sum_{j=2}^{k}(C_k^j)^2 Var(H_n^{(j)}) = O(n^{-2}).$$

这说明$\sqrt{n}\sum_{j=2}^{k} C_k^j H_n^{(j)} \xrightarrow{P} 0$，及$\sqrt{n}kH_n^{(1)} = \sqrt{n}k\frac{1}{n}\sum_i h^{(1)}(X_1) \xrightarrow{L} N(0, k^2\sigma_1^2)$。
也就是

$$\sqrt{n}(U_n - \theta) \xrightarrow{L} N(0, k^2\sigma_1^2)\text{”}$$



例： 假设有二元随机向量(X,Y)，其联合分布为F，边缘分布分别为F_X,F_Y。设$\{(X_1,Y_1),(X_2,Y_2),\ldots,(X_n,Y_n)\}$是来自$F$的样本，记$X,Y$的样本方差及$(X,Y)$的样本协方差、样本相关系数分别为

$$S_X^2=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2,\quad S_Y^2=\frac{1}{n-1}\sum_{i=1}^{n}(Y_i-\bar{Y})^2,$$

$$S_{XY}=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y}),$$

$$r_{XY}=\frac{S_{XY}}{S_XS_Y}.$$

则当X,Y独立时，r_{XY}有渐近分布：

$$\sqrt{n}r_{XY}\xrightarrow{L}N(0,\ 1).$$

证明： 容易验证S_{XY}可写成U统计量形式：

$$S_{XY}=\frac{1}{C_n^2}\sum_{i<j}\frac{1}{2}(X_i-X_j)(Y_i-Y_j)$$

即S_{XY}是以$\psi((x_1,y_1),(x_2,y_2))=\frac{1}{2}(x_1-x_2)(y_1-y_2)$为核的自由度为2的$U$统计量。



事实上由于

$$\begin{aligned}
\sum_{i<j}(X_i - X_j)(Y_i - Y_j) &= \frac{1}{2}\sum_i\sum_j(X_i - X_j)(Y_i - Y_j) \\
&= \frac{1}{2}\sum_i\sum_j[(X_i - \bar{X}) - (X_j - \bar{X})][(Y_i - \bar{Y}) - (Y_j - \bar{Y})] \\
&= \frac{1}{2}[\sum_i\sum_j(X_i - \bar{X})(Y_i - \bar{Y}) + \sum_i\sum_j(X_j - \bar{X})(Y_j - \bar{Y}) \\
&\quad - \sum_i\sum_j(X_i - \bar{X})(Y_j - \bar{Y}) - \sum_i\sum_j(X_j - \bar{X})(Y_i - \bar{Y})] \\
&= \frac{1}{2} * 2n\sum_i(X_i - \bar{X})(Y_i - \bar{Y}) \\
&= n\sum_i(X_i - \bar{X})(Y_i - \bar{Y}).
\end{aligned}$$

因此可得:

$$\frac{1}{C_n^2}\sum_{i<j}\frac{1}{2}(X_i - X_j)(Y_i - Y_j) = \frac{1}{n-1}\sum_i(X_i - \bar{X})(Y_i - \bar{Y}) = S_{XY}.$$



显然 $ES_{XY} = \frac{1}{2}E(X_1 - X_2)(Y_1 - Y_2) = Cov(X, Y)$。另外，由于

$$\psi_1((x, y)) := \frac{1}{2}E(X - x)(Y - y) = \frac{1}{2}[(x - \mu_x)(y - \mu_y).$$

因此

$$\begin{aligned}\sigma_1^2 = Var(\psi_1(X_1, Y_1)) &= \frac{1}{4}Var((X - \mu_x)^2(Y - \mu_Y)) \\ &= \frac{1}{4}(E(X - \mu_x)^2(Y - \mu_Y)^2 - Cov^2(X, Y)),\end{aligned}$$

从而当X, Y 独立时，有$\sigma_1^2 = \frac{1}{4}Var(X)Var(Y)$。因此，由定理3.3可知

$$\sqrt{n}S_{XY} \xrightarrow{L} N(0, Var(X)Var(Y)).$$

此外，由于

$$S_X^2 \xrightarrow{P} Var(X), \quad S_Y^2 \xrightarrow{P} Var(Y),$$

因此，根据Slutsky 公式可知当X, Y独立时，

$$\sqrt{n}r_{XY} = \sqrt{n}\frac{S_{XY}}{S_X S_Y} \xrightarrow{L} N(0, 1).$$



定义3.2

令$X_1, X_2, \cdots, X_m$ 和$Y_1, Y_2, \ldots, Y_n$ 分别来自相互独立的总体X和Y。
设$X_1, X_2, \ldots, X_{k_1}$给定时，$\phi(X_1, X_2, \ldots, X_{k_1}; Y_1, Y_2, \ldots, Y_{k_2})$ 为$Y_1, Y_2, \ldots, Y_{k_2}$的对称函数，而在$Y_1, Y_2, \ldots, Y_{k_2}$ 给定时，$\phi(X_1, X_2, \ldots, X_{k_1}; Y_1, Y_2, \ldots, Y_{k_2})$
为$\phi(X_1, X_2, \ldots, X_{k_1})$ 的对称函数，称具有形式:

$$U_{m,n} := \frac{1}{C_m^{k_1} C_n^{k_2}} \sum_{(m,k_1)} \sum_{(n,k_2)} \phi(X_{i_1}, X_{i_2}, \ldots, X_{i_{k_1}}; Y_{j_1}, Y_{j_2}, \ldots, Y_{j_{k_2}}) \tag{7}$$

的统计量$U_{m,n}$为以ϕ为核的自由度为(k_1, k_2) 的双样本U 统计量。其中$\sum_{(n,k)}$ 表示对$1, 2, \ldots, n$ 的所有组合$(i_1, \ldots, i_k)$ 求和。



定理3.4 (双样本U-统计量的H-分解)

对于U-统计量:

$$U_{m,n} := \frac{1}{C_m^{k_1} C_n^{k_2}} \sum_{(m,k_1)} \sum_{(n,k_2)} \phi(X_{i_1}, X_{i_2}, \ldots, X_{i_{k_1}}; Y_{j_1}, Y_{j_2}, \ldots, Y_{j_{k_2}})$$

若记

$$\begin{aligned}
h^{(0,0)} &= EU_{m,n} = \theta \\
h^{(1,0)}(x) &= E\phi(x, X_2, \ldots, X_{k_1}; Y_1, Y_2, \ldots, Y_{k_2}) - \theta \\
h^{(0,1)}(y) &= E\phi(X_1, X_2, \ldots, X_{k_1}; y, Y_2, \ldots, Y_{k_2}) - \theta \\
h^{(c,d)}(x_1, x_2, \ldots, x_c; y_1, y_2, \ldots, y_d) &= \phi(x_1, \ldots, x_c, X_{c+1}, \ldots, X_{k_1}; y_1, \ldots, y_d, Y_{d+1}, \ldots, Y_{k_2}) \\
&\quad - \sum_{\{(s,t)|s\in\{0,1,\ldots,c\}, t\in\{0,1,\ldots,d\}\}-\{(s=c,t=d)\}} \sum_{(c,s)} \sum_{(d,t)} h^{(s,t)}(x_{i1}, \ldots, x_{is}; y_{j1}, \ldots, y_{jt})
\end{aligned}$$

则$U_{m,n}$有分解式:

$$U_{m,n} = \sum_{c=0}^{k_1} \sum_{d=0}^{k_2} C_{k_1}^c C_{k_2}^d H_{m,n}^{(c,d)}, \tag{8}$$

其中$H_{m,n}^{(c,d)} = \frac{1}{C_m^c C_n^d} \sum_{(m,c)} \sum_{(n,d)} h^{(c,d)}(X_{i1}, \ldots, X_{ic}; Y_{j1}, \ldots, Y_{jd})$为自由度为$(c,d)$的双样本$U$-统计量。

类似定理6及性质3.2有关系式

- $Eh^{(c,d)}(X_1, X_2, \ldots, X_c; Y_1, Y_2, \ldots, Y_d) = 0,\ \forall c = 0, \ldots, k_1; d = 0, \ldots, k_2;$
- $Cov(h^{(c,d)}(S_1, S_2), h^{(c',d')}(S_1', S_2')) = 0,\ \forall c, d, c', d'$及集合$S_1, S_2, S_1', S_2'$，只有当$c = d', d = d', S_1 = S_1', S_2 = S_2'$时，才有$Cov(h^{(c,d)}(S_1, S_2), h^{(c',d')}(S_1', S_2')) \neq 0;$
- 各$H_{m,n}^{(c,d)}$互不相关，且同一$H_{m,n}^{(c,d)}$中各项核之间亦不相关，从而有

$$\begin{aligned} Var(H_{m,n}^{(c,d)}) &= \frac{1}{(C_m^c C_n^d)^2} \sum_{(m,c)} \sum_{(n,d)} Var(h^{(c,d)}(X_{i1}, \ldots, X_{ic}; Y_{j1}, \ldots, Y_{jd})) \\ &= \frac{1}{(C_m^c C_n^d)^2} C_m^c C_n^d \delta_{c,d}^2 \\ &= \frac{1}{C_m^c C_n^d} \delta_{c,d}^2 = O(\frac{1}{m^c n^d}) \end{aligned}$$

其中$\delta_{c,d}^2 = Var(h^{(c,d)}(X_1, \ldots, X_c; Y_1, \ldots, Y_d))$。



定理3.5 (双样本U-统计量的渐近分布)

设$E\phi(X_1, X_2, \ldots, X_{k_1}; Y_1, Y_2, \ldots, Y_{k_2}) = \theta$，且记

$$\phi_{c_1,c_2}(X_1, \ldots, X_{c_1}; Y_1, \ldots, Y_{c_2}) := E(\phi(X_1, \ldots, X_{k_1}; Y_1, \ldots, Y_{k_2})|X_1, \ldots, X_{c_1}; Y_1, \ldots, Y_{c_2})$$

及

$$\sigma^2_{c_1c_2} := Var[\phi_{c_1,c_2}(X_1, \ldots, X_{c_1}; Y_1, \ldots, Y_{c_2})].$$

如果$\sigma^2_{10}, \sigma^2_{01} < \infty$，且当$m, n \longrightarrow \infty, m/(m+n) \longrightarrow p$，则有

$$\sqrt{m+n}(U_{m,n} - \theta) \xrightarrow{L} N\Big(0, \ \frac{k_1^2\sigma^2_{10}}{p} + \frac{k_2^2\sigma^2_{01}}{1-p}\Big).$$



证明： 由$U_{m,n}$的H-分解式(8)可知：

$$\begin{aligned} U_{m,n} &= \sum_{c=0}^{k_1}\sum_{d=0}^{k_2} C_{k_1}^{c} C_{k_2}^{d} H_{m,n}^{(c,d)} \\ &= \theta + k_1 H_{m,n}^{(1,0)} + k_2 H_{m,n}^{(0,1)} + R_{m,n}, \end{aligned}$$

其中

$$H_{m,n}^{(1,0)} = \frac{1}{m}\sum_{i=1}^{m} h^{(1,0)}(X_i); \quad H_{m,n}^{(0,1)} = \frac{1}{n}\sum_{j=1}^{n} h^{(0,1)}(Y_j); \quad R_{m,n} = \sum_{c+d\geq 2} C_{k_1}^{c} C_{k_2}^{d} H_{m,n}^{(c,d)}.$$

而

$$Var(R_{m,n}) = \sum_{c+d\geq 2} (C_{k_1}^{c} C_{k_2}^{d})^2 \frac{1}{C_m^c C_n^d}\delta_{c,d}^2 = O(\frac{1}{m+n}).$$

由中心极限定理可知

$$\sqrt{m} H_{m,n}^{(1,0)} \xrightarrow{L} N(0, Var(h^{(1,0)}(X))$$

$$\sqrt{n} H_{m,n}^{(0,1)} \xrightarrow{L} N(0, Var(h^{(0,1)}(Y)).$$



$H_{m,n}^{(1,0)}, H_{m,n}^{(0,1)}$相互独立，因此有

$$\begin{aligned}\sqrt{m+n}(U_{m,n}-\theta) &= \sqrt{m+n}(k_1 H_{m,n}^{(1,0)} + k_2 H_{m,n}^{(0,1)} + R_{m,n}) \\ &= \sqrt{\frac{m+n}{m}} k_1 \sqrt{m} H_{m,n}^{(1,0)} + \sqrt{\frac{m+n}{n}} k_1 \sqrt{n} H_{m,n}^{(0,1)} + \sqrt{m+n} R_{m,n} \\ &= \sqrt{\frac{m+n}{m}} k_1 \sqrt{m} H_{m,n}^{(1,0)} + \sqrt{\frac{m+n}{n}} k_1 \sqrt{n} H_{m,n}^{(0,1)} + O(1) \\ &\xrightarrow{L} N\Big(0, \quad \frac{k_1^2}{p} Var(h^{(1,0)}(X)) + \frac{k_2^2}{1-p} Var(h^{(0,1)}(Y))\Big).\end{aligned}$$

显然$Var(h^{(1,0)}(X)) = \sigma_{10}^2, Var(h^{(0,1)}(Y)) = \sigma_{01}^2$，从而可得：

$$\sqrt{m+n}(U_{m,n}-\theta) \xrightarrow{L} N\Big(0, \quad \frac{k_1^2\sigma_{10}^2}{p} + \frac{k_2^2\sigma_{01}^2}{1-p}\Big).$$



定理3.6 (零假设下W^+的渐近分布)

当F为对称的连续分布函数时，Wilcoxon秩统计量$W^+ = \sum_{i=1}^n R_i I(X_i > 0)$有渐近正态性：

$$\frac{W^+ - E(W^+)}{\sqrt{Var(W^+)}} = \frac{W^+ - n(n+1)/4}{\sqrt{n(n+1)(2n+1)/24}} \xrightarrow{L} N(0,1), n \longrightarrow \infty.$$

证明： 由于

$$\begin{aligned}\{X_i + X_j > 0\} =& \{X_i + X_j > 0, X_i > X_j\} + \{X_i + X_j > 0, X_j > X_i\} \\ =& \{X_i > 0, X_j > -X_i, X_j < X_i\} + \{X_j > 0, X_j > X_i, X_i > -X_j\} \\ =& \{X_i > 0, |X_j| < X_i\} + \{X_j > 0, |X_i| < X_j\},\end{aligned}$$



因此，有

$$\sum_{1\le i\le j\le n} I\{X_i+X_j>0\} = \sum_{i<j} I\{X_i+X_j>0\} + \sum_{i=j} I\{X_i+X_j>0\}$$

$$= \sum_{i=1}^{n} I\{X_i>0\} + \sum_{i<j} I\{X_i>0\}I\{|X_j|<X_i\} + \sum_{i<j} I\{X_j>0\}I\{|X_i|<X_j\}$$

$$= \sum_{i=1}^{n} I\{X_i>0\} + \sum_{i<j} I\{X_i>0\}I\{|X_j|<X_i\} + \sum_{i>j} I\{X_i>0\}I\{|X_j|<X_i\}$$

$$= \sum_{i=j}^{n} I\{X_i>0\}I\{|X_j|\le X_i\} + \sum_{i\ne j} I\{X_i>0\}I\{|X_j|\le X_i\}$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{n} I\{X_i>0\}I\{|X_j|\le X_i\}$$

$$= \sum_{i=1}^{n} I\{X_i>0\}\sum_{j=1}^{n} I\{|X_j|\le X_i\}$$

$$= \sum_{i=1}^{n} R_i I\{X_i>0\} = W^+,$$



即有

$$W^+ = \sum_{1\le i\le j\le n} I\{X_i + X_j > 0\} = C_n^2 U_n^{(1)} + nU_n^{(2)}.$$

其中$U_n^{(1)} = \frac{1}{C_n^2}\sum_{(n,2)} I\{X_i + X_j > 0\}, U_n^{(2)} = \frac{1}{n}\sum_i I\{X_i > 0\}$. 易知$U_n^{(1)}, U_n^{(2)}$均为$U$统计量，对于$U_n^{(1)}$，由于

$$h^{(1)}(x) := EI\{X + x > 0\} = P(X > -x) = 1 - F(-x),$$

所以当X分布对称时有:

$$h^{(1)}(x) = F(x), \quad EI\{X_1 + X_2 > 0\} = \frac{1}{2}.$$

从而$\sigma_1^2 = var(h^{(1)}(X)) = var(F(X)) = \frac{1}{12}$，由$U$-统计量理论定理3.3可得在原假设$H_0 : \theta = 0$成立的条件下:

$$\sqrt{n}(U_n^{(1)} - \frac{1}{2}) \xrightarrow{L} N(0, \frac{1}{3}).$$



即

$$\sqrt{n}\Big(\frac{1}{C_n^2}(W^+ - nU_n^{(2)}) - \frac{1}{2}\Big) = \frac{\sqrt{n}}{C_n^2}(W^+ - nU_n^{(2)} - \frac{1}{2}C_n^2) \xrightarrow{L} N(0, \frac{1}{3}),$$

即

$$2n^{-\frac{3}{2}}\Big(W^+ - \frac{n(n-1)}{4}\Big) - 2n^{-\frac{1}{2}}U_n^{(2)} \xrightarrow{L} N(0, \frac{1}{3}),$$

而$2n^{-\frac{1}{2}}U_n^{(2)} \xrightarrow{P} 0$，故可得:

$$2n^{-\frac{3}{2}}\Big(W^+ - \frac{n(n-1)}{4}\Big) \xrightarrow{L} N(0, \frac{1}{3}).$$

即

$$2\sqrt{3}n^{-\frac{3}{2}}\Big(W^+ - \frac{n(n-1)}{4}\Big) \xrightarrow{L} N(0, 1).$$

因此，当X分布对称时有:

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



因此，在显著性水平α下检验问题(1)的渐近拒绝域为

$$R = \left\{ \frac{W^+ - \frac{n(n+1)}{4}}{\sqrt{n(n+1)(2n+1)/24}} > Z_\alpha \right\}.$$

其中Z_α为标准正态分布的上α分位数。



定理3.7 (零假设下W_y的渐近分布)

当假设$H_0 : F = G$成立时，Wilcoxon-Mann-Whitney统计量$W_y = \sum_{j=1}^{n} R_j$有渐近正态性：

$$\frac{W_y - E(W_y)}{\sqrt{Var(W_y)}} = \frac{W_y - n(m+n+1)/2}{\sqrt{nm(m+n+1)/12}} \xrightarrow{L} N(0,1).$$

证明： 我们将从U-统计量的角度来证明该结论. 若记

$$U_{m,n} := \frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} I(X_i < Y_j),$$

则易知$U_{m,n}$为自由度为(1,1)的双样本U-统计量，且有关系式：

$$mnU_{m,n} = \sum_{i=1}^{m} \sum_{j=1}^{n} I(X_i < Y_j) = W_y - \frac{n(n+1)}{2}. \tag{9}$$



当$F = G$时，有$EU_{m,n} = P(X_1 < Y_1) = 1/2$，且有

$$\sigma_{10}^2 = Var(P(X_1 < Y_1|X_1)) = Var(1 - G(X_1)) = Var(G(X_1)) = Var(F(X_1)) = 1/12,$$
$$\sigma_{01}^2 = Var(P(X_1 < Y_1|Y_1)) = Var(F(Y_1)) = Var(G(Y_1)) = 1/12.$$

因此，由双样本U-统计量的渐近性质定理3.5可知

$$\sqrt{m+n}(U_{m,n} - \frac{1}{2}) \xrightarrow{L} N(0, \frac{\sigma_{10}^2}{p} + \frac{\sigma_{01}^2}{1-p}).$$

其中$\frac{m}{m+n} \longrightarrow p$，即

$$\sqrt{m+n}(U_{m,n} - \frac{1}{2}) \xrightarrow{L} N(0, \frac{1}{12p(1-p)}).$$

由W_y与$U_{m,n}$的关系式(9)可得：

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



定理3.8 (零假设下Spearman秩相关系数的渐近分布)

在大样本情况下，当$H_0: X, Y$ 独立时，$r_S = r_{R^X R^Y} = \frac{\sum_{i=1}^n (R_i^X - \overline{R}^X)(R_i^Y - \overline{R}^Y)}{\sqrt{\sum_{i=1}^n (R_i^X - \overline{R}^X)^2} \sqrt{\sum_{i=1}^n (R_i^Y - \overline{R}^Y)^2}}$ 有渐近正态性：

$$\sqrt{n-1} r_S \xrightarrow{L} N(0,1).$$

证明：当X, Y独立时，R^X, R^Y独立，因此，由相关系数的渐近正态性可知：有

$$\sqrt{n} r_{R^X R^Y} \xrightarrow{L} N(0,1),$$

即：

$$\sqrt{n-1} r_S \xrightarrow{L} N(0,1).$$

故在α显著性水平下，H_0的渐近拒绝域为：

$$R = \left\{ |\sqrt{n-1} r_S| > Z_{\alpha/2} \right\},$$

其中$Z_{\alpha/2}$为标准正态分布的上$\alpha/2$分位数。



定理3.9 (零假设下Kendall'τ 系数的渐近分布)

假设有二元随机向量(X,Y)，其联合分布为F，边缘分布分别为F_X, F_Y。$\{(X_1,Y_1),(X_2,Y_2),\ldots,(X_n,Y_n)\}$ 是来自F的样本。当X,Y独立时，样本Kendall's τ 系数

$$\tau_n = \frac{1}{C_n^2}\sum_{i<j} sgn(X_i - X_j)(Y_i - Y_j),$$

有渐近正态性：

$$\sqrt{n}\tau_n \xrightarrow{L} N(0, 4/9).$$

证明： 显然，τ_n 为自由度为2的单样本U-统计量。且当X,Y独立时，有

$$\phi_1((X_1,Y_1)) := E[sgn(X_1 - X_2)(Y_1 - Y_2)|(X_1,Y_1)] = (1 - 2F_X(X_1))(1 - 2F_Y(Y_1)).$$



因此，有

$$\begin{aligned} E\phi_1(X_1, Y_1) &= 0, \\ Var(\phi_1(X_1, Y_1)) &= E(1 - 2F_X(X_1))^2 E(1 - 2F_Y(Y_1))^2 \\ &= \int_0^1 (1-2u)^2 du \int_0^1 (1-2v)^2 dv \\ &= \frac{1}{9}. \end{aligned}$$

从而由单样本U-统计量的渐近正态性定理3.3 可知

$$\sqrt{n}\tau_n \xrightarrow{L} N(0, 4/9).$$

故在α显著性水平下，$H_0 : X, Y$独立的渐近拒绝域为：

$$R = \left\{ \frac{3}{2} \sqrt{n}|\tau_n| > Z_{\alpha/2} \right\},$$

其中$Z_{\alpha/2}$为标准正态分布的上$\alpha/2$分位数。



经验分布与秩方法小结

基于经验分布函数与秩方法的检验统计量计算其简单. 但是

- 问题1: 经验分布函数是比较大小的，而比较大小是一维空间特有的。
- 问题2: 古典的多元经验分布函数遭受维数灾难的。
- 问题3: 基于经验分布函数的统计方法效能“偏低”。

? 目标：我们能否提供一个新的框架处理新数据？如果是，是什么？



不可思议的事实！

Wilcoxon-Mann-Whitney 检验(W) 相对于两样本Student's t 检验(T)的渐近相对有效性(ARE) :

$$\inf \mathrm{ARE}(W/T) = 0.864, \text{but } \sup \mathrm{ARE}(W/T) = \infty.$$

Van der Waerden 检验(V):

$$S_{n,m} = \sum_{j=1}^{m} \Phi^{-1}\left(\frac{S_j}{n+m+1}\right),$$

Φ 是标准正态分布。

$$\inf \mathrm{ARE}(V/T) = 1.$$

等式成立当且仅当样本的分布为正态分布。



不可思议的事实！

Wilcoxon-Mann-Whitney 检验(W) 相对于两样本Student's t 检验(T)的渐近相对有效性(ARE) :

$$\inf \mathrm{ARE}(W/T) = 0.864, \text{but } \sup \mathrm{ARE}(W/T) = \infty.$$

Van der Waerden 检验(V):

$$S_{n,m} = \sum_{j=1}^{m} \Phi^{-1}\left(\frac{S_j}{n+m+1}\right),$$

Φ 是标准正态分布。

$$\inf \mathrm{ARE}(V/T) = 1.$$

等式成立当且仅当样本的分布为正态分布。



不可思议的事实！

Wilcoxon-Mann-Whitney 检验(W) 相对于两样本Student's t 检验(T)的渐近相对有效性(ARE) :

$$\inf \mathrm{ARE}(W/T) = 0.864, \text{but } \sup \mathrm{ARE}(W/T) = \infty.$$

Van der Waerden 检验(V):

$$S_{n,m} = \sum_{j=1}^{m} \Phi^{-1}\left(\frac{S_j}{n+m+1}\right),$$

Φ 是标准正态分布。

$$\inf \mathrm{ARE}(V/T) = 1.$$

等式成立当且仅当样本的分布为正态分布。

