

非参数统计

第二章 U 统计量

制作：邓明华

2023 秋季学期

本章目录

- ① U 统计量简介
- ② U 统计量的定义
- ③ U 统计量性质
- ④ U 统计量大样本特性

本节目录

- 1 U 统计量简介
- 2 U 统计量的定义
- 3 U 统计量性质
- 4 U 统计量大样本特性

U 统计量简介

- 基本理论由 W.Hoeffding 1948 年给出. W.Hoeffding. A class of statistics with asymptotically normal distribution. Ann. Math. Statist. 19: 293-325, 1948.
- 参考书
 - ① Denker, Manfred (1985) Asymptotic Distribution Theory in Nonparametric Statistics, Fr. Vieweg & Sohn, Braunschweig, Wiesbaden.
 - ② Lee, A. J. (1990) U-Statistics, Marcel Dekker Inc., New York.
 - Fraser, D. A. S. (1957) Nonparametric Methods in Statistics, John Wiley & Sons, New York.
 - ③ Serfling, R. J. (1980) Approximation Theorems of Mathematical Statistics, JohnWiley & Sons, New York.
 - ④ Lehmann, E. L. (1999) Elements of Large Sample Theory, Springer.

核的概念

- **定义 2.1:** 对分布族 $\mathcal{F}$ 的参数 θ , 如果存在样本量为 r 的样本 $X_1, \dots, X_r$ 以及统计量 $h(X_1, \dots, X_r)$, 使得

$$E_F h(X_1, \dots, X_r) = \theta$$

则称参数 θ 对分布族 $\mathcal{F}$ 是 **r 可估**的, 函数 $h(x_1, \dots, x_r)$ 称为**参数 θ 的核**。

- 例: 因为总体期望有无偏估计 X_1 , 于是总体期望是 1 可估的, X_1 是总体期望的核。

对称核

- 一般地，我们可以要求核对称。即对 $(1, 2, \dots, r)$ 的任何一个排列 $(i_1, i_2, \dots, i_r)$ ，都有

$$h(x_1, \dots, x_r) = h(x_{i_1}, x_{i_2}, \dots, x_{i_r})$$

- 如果核不对称，可以对称化，

$$h^*(x_1, \dots, x_r) = \frac{1}{r!} \sum_{(i_1, i_2, \dots, i_r)} h(x_{i_1}, \dots, x_{i_r})$$

本节目录

- 1 U 统计量简介
- 2 U 统计量的定义
- 3 U 统计量性质
- 4 U 统计量大样本特性

U 统计量的概念

- **定义 2.2:** 设 $X_1, \dots, X_n$ 是总体 $F(x) \in \mathcal{F}$ 的简单随机样本。 r 可估参数 θ 有对称核 $h(X_1, \dots, X_r)$. 则参数 θ 的 U 统计量定义为,

$$U(X_1, \dots, X_n) = \frac{1}{C_n^r} \sum_{(\beta_1, \dots, \beta_r)} h(X_{\beta_1}, \dots, X_{\beta_r})$$

特殊 U 统计量 (I)

- 设 $\mathcal{F}$ 是全体一阶矩存在的分布类, 其均值 $\theta = E_F(X)$, 则 $h(x) = x$ 即为核, 于是,

$$U(X_1, \dots, X_n) = \frac{1}{n} \sum_{i=1}^n X_i = \bar{X}$$

- 设 $\mathcal{F}$ 是全体 k 阶矩存在的分布类, 其 k 阶矩 $\mu_k = \int x^k dF(x)$ 是自由度为 1 的可估参数, 核函数为 $h(x) = x^k$, 于是,

$$U(X_1, \dots, X_n) = \frac{1}{n} \sum_{i=1}^n X_i^k$$

特殊 U 统计量 (II)

- 方差 σ^2 的估计 $\sigma^2 = E(X^2) - \mu_1^2$, 于是 σ^2 的核函数 $h_1(x_1, x_2) = x_1^2 - x_1 x_2$. 对称化,

$$\begin{aligned} h(x_1, x_2) &= \frac{1}{2}(h(x_1, x_2) + h(x_2, x_1)) \\ &= \frac{1}{2}(x_1^2 - x_1 x_2 + x_2^2 - x_2 x_1) \\ &= \frac{1}{2}(x_1 - x_2)^2 \end{aligned}$$

- U 统计量

$$\begin{aligned} U(x_1, \dots, x_n) &= \frac{1}{C_n^2} \sum_{i < j} \frac{1}{2} (x_i - x_j)^2 \\ &= \frac{1}{n(n-1)} \sum_{i < j} (x_i^2 - 2x_i x_j + x_j^2) \end{aligned}$$

特殊 U 统计量 (III)

- 接上页

$$\begin{aligned}U &= \frac{1}{n(n-1)} \left\{ \sum_{i \neq j} \frac{x_i^2 + x_j^2}{2} - \sum_{i \neq j} x_i x_j \right\} \\&= \frac{1}{n(n-1)} \left\{ \sum_{i \neq j} \frac{x_i^2 + x_j^2}{2} - (\sum x_i)^2 + \sum_{i=1}^n x_i^2 \right\} \\&= \frac{1}{n(n-1)} \left\{ \sum_i \sum_j \frac{x_i^2 + x_j^2}{2} - \sum_{i=j} \frac{x_i^2 + x_j^2}{2} - (\sum x_i)^2 + \sum_{i=1}^n x_i^2 \right\} \\&= \frac{1}{n(n-1)} \left\{ n \sum_{i=1}^n x_i^2 - (\sum_i x_i)^2 \right\} \\&= \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 = S^2\end{aligned}$$

- 因此 U 统计量是均值和方差估计的推广。

特殊 U 统计量 (IV)

- 符号秩统计量

$$\begin{aligned} W^+ &= \sum_{i \leq j} \psi(x_i + x_j) \\ &= \sum_i \psi(x_i) + \sum_{i < j} \psi(x_i + x_j) \end{aligned}$$

- 令核函数

$$h_1(x_1) = \psi(x_1), h_2(x_1, x_2) = \psi(x_1 + x_2)$$

- 于是

$$\begin{aligned} U_1(x_1, \dots, x_n) &= \frac{1}{n} \sum_{i=1}^n \psi(x_i) \\ U_2(x_1, \dots, x_n) &= \frac{1}{C_n^2} \sum_{i < j} \psi(x_i + x_j) \\ W^+ &= nU_1 + \frac{n(n-1)}{2} U_2 \end{aligned}$$

特殊 U 统计量 (V)

$$\begin{aligned}\theta_1 &= \int h_1(x) dF(x) = \int_{x>0} dF(x) \\ &= P(x > 0) = 1 - F(0)\end{aligned}$$

$$\begin{aligned}\theta_2 &= \int h_2(x_1, x_2) dF(x_1) dF(x_2) = \int_{x_1+x_2>0} dF(x_1) dF(x_2) \\ &= \int_{-\infty}^{+\infty} dF(x_1) \int_{-x_1}^{+\infty} dF(x_2) \\ &= \int_{-\infty}^{+\infty} (1 - F(-x_1)) dF(x_1) \\ &= P(x_1 + x_2 > 0) = E_F(1 - F(-x))\end{aligned}$$

本节目录

- 1 U 统计量简介
- 2 U 统计量的定义
- 3 U 统计量性质**
- 4 U 统计量大样本特性

U 统计量的性质

- **定理 2.1:** 设 $X_1, \dots, X_n$ 是来自分布 $F(x)$ 的样本, $F(x)$ 取自分布族 $\mathcal{F} = \{F(x, \theta) | \theta \in \Theta\}$. θ 是 r 可估参数, 核为 $h(x_1, \dots, x_r)$, $U(X_1, \dots, X_n)$ 为 U 统计量, 则有

$$E(U(X_1, \dots, X_n)) = \theta,$$

$$Var(U(X_1, \dots, X_n)) = \frac{1}{C_n^r} \sum_{c=0}^r C_r^c C_{n-r}^{r-c} \xi_c$$

其中

$$\xi_c = Cov[h(X_1, \dots, X_c, X_{c+1}, \dots, X_r), \\ h(X_1, \dots, X_c, X_{r+1}, \dots, X_{2r-c})].$$

定理 2.1 证明

- 期望

$$\begin{aligned} & E_F U(x_1, \dots, x_n) \\ &= \frac{1}{C_n^r} \sum_{(\beta_1, \dots, \beta_r)} E_F[h(x_{\beta_1}, \dots, x_{\beta_r})] \\ &= \frac{1}{C_n^r} \sum_{(\beta_1, \dots, \beta_r)} \theta = \theta \end{aligned}$$

- 方差

$$\begin{aligned} \text{Var}(U) &= E\left[\frac{1}{C_n^r} \sum_{(\beta_1, \dots, \beta_r)} (h(X_{\beta_1}, \dots, X_{\beta_r}) - \theta)\right]^2 \\ &= \left(\frac{1}{C_n^r}\right)^2 \sum_{(\alpha_1, \dots, \alpha_r)} \sum_{(\beta_1, \dots, \beta_r)} E_F[(h(X_{\alpha_1}, \dots, X_{\alpha_r}) - \theta)(h(X_{\beta_1}, \dots, X_{\beta_r}) - \theta)] \\ &= \left(\frac{1}{C_n^r}\right)^2 \sum_{(\alpha_1, \dots, \alpha_r)} \sum_{(\beta_1, \dots, \beta_r)} \text{Cov}[h(X_{\alpha_1}, \dots, X_{\alpha_r}), h(X_{\beta_1}, \dots, X_{\beta_r})] \\ &= \left(\frac{1}{C_n^r}\right)^2 \sum_{c=0}^r C_n^r C_r^c C_{n-r}^{r-c} \xi_c = \frac{1}{C_n^r} \sum_{c=0}^r C_r^c C_{n-r}^{r-c} \xi_c \end{aligned}$$

定理 2.1 注

- 特别地

$$\begin{aligned}\xi_0 &= \text{Cov}[h(X_1, \dots, X_r), h(X_{r+1}, \dots, X_{2r-1})] = 0 \\ \xi_r &= \text{Cov}[h(X_1, \dots, X_r), h(X_1, \dots, X_r)] \\ &= \text{Var}[h(X_1, \dots, X_r)]\end{aligned}$$

- 而且

$$\begin{aligned}\xi_c &= \text{Cov}[h(X_1, \dots, X_c, X_{c+1}, \dots, X_r), h(X_1, \dots, X_c, X_{r+1}, \dots, X_{2r-c})] \\ &\leq \sqrt{\text{Var}(h(X_1, \dots, X_c, X_{c+1}, \dots, X_r))} \\ &\quad \times \sqrt{\text{Var}(h(X_1, \dots, X_c, X_{r+1}, \dots, X_{2r-c}))} \\ &= \text{Var}(h(X_1, \dots, X_r)) = \xi_r\end{aligned}$$

定理 2.1 注 (续)

$$\begin{aligned}\xi_c &= E[(h(X_1, \dots, X_c, X_{c+1}, \dots, X_r) - \theta)(h(X_1, \dots, X_c, X_{r+1}, \dots, X_{2r-c}) - \theta)] \\&= E_{X_1, \dots, X_c} \{ E_{X_{c+1}, \dots, X_r} [h(X_1, \dots, X_c, X_{c+1}, \dots, X_r) - \theta] \\&\quad E_{X_{r+1}, \dots, X_{2r-c}} [h(X_1, \dots, X_c, X_{r+1}, \dots, X_{2r-c}) - \theta] \} \\&= E_{X_1, \dots, X_c} [E_{X_{c+1}, \dots, X_r} (h(X_1, \dots, X_c, X_{c+1}, \dots, X_r) - \theta)^2] \\&\geq 0\end{aligned}$$

- 实际上, 有些书中直接用条件期望定义函数

$$\begin{aligned}&h_c(t_1, \dots, t_c) \\&= E_{X_{c+1}, \dots, X_r} [h(X_1, \dots, X_c, X_{c+1}, \dots, X_r) | X_1 = t_1, \dots, X_c = t_c]\end{aligned}$$

则 $\xi_c = \text{Var}(h_c(X_1, X_2, \dots, X_c))$.

条件期望 (补充)

- 为简单起见, 不妨设密度函数存在。设随机变量 (ξ, η) 的联合概率密度为 $p(x, y)$, 则边缘密度为

$$p_{\xi}(x) = \int_{-\infty}^{+\infty} p(x, y) dy, \quad p_{\eta}(y) = \int_{-\infty}^{+\infty} p(x, y) dx$$

- 条件概率密度 $p(y|x)$ 定义为

$$p(y|x) = \frac{p(x, y)}{p_{\xi}(x)}$$

- 条件数学期望

$$E[\eta|\xi = x] = E_{(P_{y|\xi=x})}(\eta) = \int_{-\infty}^{+\infty} yp(y|x) dy$$

条件期望公式

- 条件期望公式 $E(E[\eta | \xi = x]) = E(\eta)$.
- 证明:

$$\begin{aligned} E(\eta) &= \int_{-\infty}^{+\infty} yp_{\eta}(y) dy \\ &= \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} yp(x, y) dx dy \end{aligned}$$

$$\begin{aligned} E(\eta | \xi = x) &= \int_{-\infty}^{+\infty} yp(y|x) dy \\ &= \int_{-\infty}^{+\infty} y \frac{p(x, y)}{p_{\xi}(x)} dy \end{aligned}$$

$$\begin{aligned} &E(E(\eta | \xi = x)) \\ &= \int_{-\infty}^{+\infty} \left[\int_{-\infty}^{+\infty} y \frac{p(x, y)}{p_{\xi}(x)} dy \right] p_{\xi}(x) dx \\ &= \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} yp(x, y) dx dy \end{aligned}$$

定理 2.1 应用 (I)

- 例 1 求 $\bar{X}$ 的方差。 $h(x) = x$, $r = 1$, $\xi_1 = \text{Var}(X_1) = \sigma^2$, 则

$$\text{Var}(U) = \frac{1}{C_n^1} C_{n-1}^{r-1} \xi_1 = \frac{\xi_1}{n} = \frac{\sigma^2}{n}$$

- 例 2 求 S^2 的方差。 $r = 2$, $h(x_1, x_2) = (x_1 - x_2)^2/2$.

$$\begin{aligned} h_1(x_1) &= E[(X - x_1)^2/2] \\ &= E\left(\frac{X^2}{2} - Xx_1 + \frac{x_1^2}{2}\right) \\ &= \frac{x_1^2}{2} - \mu x_1 + \frac{1}{2}(\sigma^2 + \mu^2) \\ &= \frac{\sigma^2}{2} + \frac{1}{2}(x_1 - \mu)^2 \end{aligned}$$

定理 2.1 应用 (II)

$$\xi_1 = \text{Var}(h_1(X)) = \text{Var}\left(\frac{(X - \mu)^2}{2}\right) = \frac{1}{4}[E(X - \mu)^4 - \sigma^4]$$

$$\begin{aligned}\xi_2 &= \text{Var}(h(x_1, x_2)) = \text{Var}\left(\frac{(X_1 - X_2)^2}{2}\right) \\&= \frac{1}{4}\{E((X_1 - X_2)^4) - E^2((X_1 - X_2)^2)\} \\&= \frac{1}{4}\{E((X_1 - \mu)^4 - 4(X_1 - \mu)^3(X_2 - \mu) + 6(X_1 - \mu)^2(X_2 - \mu)^2 \\&\quad - 4(X_1 - \mu)(X_2 - \mu)^3 + (X_2 - \mu)^4) \\&\quad - E^2((X_1 - \mu)^2 - 2(X_1 - \mu)(X_2 - \mu) + (X_2 - \mu)^2)\} \\&= \frac{1}{4}\{2E(X - \mu)^4 + 6\sigma^4 - (2\sigma^2)^2\} \\&= \frac{1}{2}(E(X - \mu)^4 + \sigma^4)\end{aligned}$$

定理 2.1 应用 (III)

- 于是

$$\begin{aligned} \text{Var}(S^2) &= \frac{1}{C_n^2} [C_2^1 C_{n-2}^1 \xi_1 + C_2^2 C_{n-2}^0 \xi_2] \\ &= \frac{2}{n(n-1)} [2(n-2)\xi_1 + \xi_2] \\ &= \frac{2}{n(n-1)} \left[2(n-2) \times \frac{1}{4} (E(X-\mu)^4 - \sigma^4) \right. \\ &\quad \left. + \frac{1}{2} (E(X-\mu)^4 + \sigma^4) \right] \\ &= \frac{1}{n} E(X-\mu)^4 - \frac{n-3}{n(n-1)} \sigma^4 \end{aligned}$$

本节目录

- 1 U 统计量简介
- 2 U 统计量的定义
- 3 U 统计量性质
- 4 U 统计量大样本特性**

大样本性质

- **定理 2.2:** 设 $X_1, \dots, X_n \sim F(x)$, $F(x) \in \mathcal{F}$ 是分布函数。设 θ 是由分布函数 $F(x)$ 确定的未知参数。若 θ 是 r 可估参数, 对称核 $h(x_1, \dots, x_r)$, $U(X_1, \dots, X_n)$ 是 θ 的 U 统计量, 若

$$E[h^2(X_1, \dots, X_r)] < +\infty$$

则

$$\lim_{n \rightarrow +\infty} n \text{Var}[U(X_1, \dots, X_n)] = r^2 \xi_1$$

其中

$$\xi_c = \text{Cov}[h(X_1, \dots, X_c, X_{c+1}, \dots, X_r), \\ h(X_1, \dots, X_c, X_{r+1}, \dots, X_{2r-c})].$$

定理 2.2 证明 (I)

- 前面证明 $0 \leq \xi_c \leq \xi_r$, $c = 1, \dots, r$. 由已知条件 $\xi_r \leq +\infty$, 则 ξ_c 均有限, $c = 1, \dots, r$.

$$\begin{aligned} n \operatorname{Var}(U) &= n \sum_{c=1}^r \frac{C_r^c C_{n-r}^{r-c}}{C_n^r} \xi_c \\ &= n \sum_{c=1}^r \frac{C_r^c \frac{(n-r)!}{(r-c)!(n-2r+c)!}}{\frac{n!}{(n-r)!r!}} \xi_c \\ &= n \sum_{c=1}^r C_r^c \frac{r!}{(r-c)!} \frac{[(n-r)!]^2}{n!(n-2r+c)!} \xi_c \\ &= \sum_{c=1}^r C_r^c \frac{r!}{(r-c)!} \frac{(n-r)(n-r-1) \cdots (n-2r+c+1)}{(n-1)(n-2) \cdots (n-r+1)} \end{aligned}$$

定理 2.2 证明 (II)

- 当 $c = 2, 3, \dots, r$ 时, 分子项数比分母项数至少少 1, 故极限为 0;
- 当 $c = 1$ 时, 分子项数与分母项数相同, 且极限为 1.

$$\lim_{n \rightarrow +\infty} n \text{Var}(U) = \frac{C_r^1 \cdot r!}{(r-1)!} \xi_1 = r^2 \xi_1$$

- 因此, 当 $h(X_1, \dots, X_r)$ 的二阶矩存在时, $U(X_1, \dots, X_n)$ 均方收敛于 θ , 即 U 是 θ 的相合估计。

大样本性质

- **定理 2.3 (Hoeffding 定理):** 设 $X_1, X_2, \dots, X_n \sim F(x)$ 是简单随机样本, θ 是 $F(x)$ 的 r 可估参数, 其对称核为 $h(x_1, \dots, x_r)$. 若

$$E[h^2(x_1, \dots, x_r)] < +\infty$$

且 $\xi_1 > 0$, 统计量

$$\sqrt{n}(U(X_1, \dots, X_n) - \theta) \xrightarrow{d} N(0, r^2 \xi_1)$$

即有极限分布为 $N(0, r^2 \xi_1)$.

- 这个定理的证明需要引进一个新的概念——投影。

投影的定义

- **定义 2.3:** 设随机变量 $X_1, \dots, X_n$ 相互独立同分布, 分布函数为 $F(t)$. 假设统计量 $W(X_1, \dots, X_n)$ 对变元对称且 $E(W) = 0$. 对于某个实函数 $K(t)$, 考虑集合

$$V = \left\{ v \mid v = \sum_{i=1}^n K(X_i) \right\}$$

则 W 在 V 上的投影定义为

$$v^* = \sum_{i=1}^n K^*(X_i)$$

其中 $K^*(t) = E(W | X_1 = t)$

投影的直观意义

- 用 $h_1(t)$ 来近似 U 统计量。
- 从定理 2.2 的证明过程我们可以看出，U 统计量的方差在样本量 n 趋于无穷时，最高阶在 ξ_1 上，而 ξ_1 是 $h_1(t)$ 的方差，因此，真正影响 U 的极限行为的是 $h_1(t)$ 。
- 下面的定理揭示其作为投影的特性。

投影定理

- 定理 2.4: 设统计量 $W(X_1, \dots, X_n)$ 是独立同分布随机变量 $X_1, \dots, X_n$ 的对称统计量, $E(W) = 0$. 令

$$v^* = \sum_{i=1}^n K^*(X_i),$$

- 则 v^* 是 W 在 V 上的投影, 满足

$$\forall v \in V, E[W - v^*]^2 \leq E[W - v]^2$$

定理 2.4 证明 (I)

- 证明:

$$\begin{aligned} E(W - v)^2 &= E[(W - v^*) + (v^* - v)] \\ &= E(W - v^*)^2 + E(v^* - v)^2 + 2E[(W - v^*)(v^* - v)] \end{aligned}$$

只需证明交叉项为 0.

$$\begin{aligned} &E\{(W - v^*)(v^* - v)\} \\ &= \sum_{i=1}^n E\{(W - v^*)(K^*(X_i) - K(X_i))\} \\ &= \sum_{i=1}^n E_{X_i}\{E_{X_{[-i]}}[(W - v^*)(K^*(X_i) - K(X_i))|X_i]\} \\ &= \sum_{i=1}^n E_{X_i}\{(K^*(X_i) - K(X_i))E_{X_{[-i]}}[(W - v^*)|X_i]\} \end{aligned}$$

定理 2.4 证明 (II)

- 由 K^* 的定义,

$$\begin{aligned} & E_{X_{[-i]}} [(W - v^*) | X_i] \\ &= E(W | X_i) - \sum_{j=1}^n E_{X_{[-i]}} (K^*(X_j) | X_i) \end{aligned}$$

- 此时有两种情况

- ① 当 $j = i$ 时, $E_{X_{[-i]}} (K^*(X_i) | X_i) = K^*(X_i)$.
- ② 当 $j \neq i$ 时, 由 X_i 和 X_j 独立, 条件期望成为无条件期望

$$E_{X_{[-i]}} (K^*(X_j) | X_i) = E_{X_j} (K^*(X_j))$$

定理 2.4 证明 (III)

- 于是

$$\begin{aligned} & E_{X_{[-i]}} [(W - v^*) | X_i] \\ &= K^*(X_i) - K^*(X_i) - (n-1)E_{X_j}(K^*(X_j)) \\ &= -(n-1)E_{X_j}(K^*(X_j)) \\ &= -(n-1)E_{X_j}(E(W | X_j)) = -(n-1)E(W) = 0 \end{aligned}$$

- 那么

$$\begin{aligned} & E\{(W - v^*)(v^* - v)\} \\ &= \sum_{i=1}^n E_{X_i}\{(K^*(X_i) - K(X_i))E_{X_{[-i]}}[(W - v^*) | X_i]\} = 0 \end{aligned}$$

投影的计算

- **定理 2.5:** 若 $U(X_1, \dots, X_n)$ 是自由度为 r 的可估参数 θ 的 U 统计量, 其对称核为 $h(x_1, \dots, x_r)$. 则 $U(X_1, \dots, X_n) - \theta$ 在 U 上的投影为,

$$v^* = \frac{r}{n} \sum_{i=1}^n [h_1(X_i) - \theta]$$

其中 $h_1(t) = E[h(X_1, \dots, X_r) | X_1 = t]$.

定理 2.5 证明

- 按定义

$$\begin{aligned} K^*(t) &= E\{U(X_1, \dots, X_n) - \theta | X_1 = t\} \\ &= E\left\{ \frac{1}{C_n^r} \sum_{(\beta_1, \dots, \beta_r)} h(X_{\beta_1}, \dots, X_{\beta_r}) - \theta | X_1 = t \right\} \\ &= \frac{1}{C_n^r} \{ C_{n-1}^r \theta + C_{n-1}^{r-1} h_1(t) \} - \theta \\ &= \frac{r}{n} h_1(t) + \frac{n-r}{n} \theta - \theta \\ &= \frac{r}{n} [h_1(t) - \theta] \end{aligned}$$

投影计算: 例子 (I)

- 例 1: σ^2 的估计 S^2 的投影计算。此时, $r = 2$,
 $h(x_1, x_2) = (x_1 - x_2)^2/2$,

$$\begin{aligned}h_1(t) &= E\left(\frac{(X_1 - X_2)^2}{2} \mid X_1 = t\right) \\&= E\left(\frac{(X - t)^2}{2}\right) \\&= \frac{1}{2}(\sigma^2 + (\mu - t)^2)\end{aligned}$$

于是

$$v^* = \frac{1}{n} \sum_{i=1}^n [(X_i - \mu)^2 - \sigma^2]$$

投影计算: 例子 (II)

- 例 2: 符号统计量 $W^+ = nU_1 + n(n-1)/2 U_2$,
 - ① U_1 : 核 $h(x) = \psi(x)$, $r=1$, $\theta_1 = p(X > 0)$;
 - ② U_2 : 核 $h(x_1, x_2) = \psi(x_1 + x_2)$, $r=2$, $\theta_2 = p(X_1 + X_2 > 0)$;
- 对于 U_2 ,

$$\begin{aligned} h_1(t) &= E[\psi(X_1 + X_2) | X_1 = t] \\ &= E[\psi(t + X_1)] = p(X_1 + t > 0) = 1 - F(-t) \end{aligned}$$

- 于是 $U_2 - \theta_2$ 的投影为

$$v^* = \frac{2}{n} \sum_{i=1}^n [1 - F(-X_i) - \theta_2]$$

Hoeffding 定理

- 再次回到定理 2.3。
- **定理 2.3 (Hoeffding 定理):** 设简单随机样本 $X_1, X_2, \dots, X_n \sim F(x)$, θ 是 $F(x)$ 的 r 可估参数, 其对称核为 $h(x_1, \dots, x_r)$, 若

$$E[h^2(x_1, \dots, x_r)] < +\infty,$$

且 $\xi_1 > 0$, 那么统计量 $\sqrt{n}[U(x_1, \dots, x_r) - \theta]$ 的极限分布为 $N(0, r^2\xi_1)$, 即

$$\sqrt{n}[U(x_1, \dots, x_r) - \theta] \xrightarrow{d} N(0, r^2\xi_1), \quad n \rightarrow +\infty$$

Hoeffding 定理说明

- 注意到定理中有条件 $\xi_1 > 0$, 有时候 ξ_1 可能为 0, 这时候对应于一种退化情形。针对退化情形, 可以考虑对应于第一个不为 0 的 ξ_k , 称之为退化阶数。
- 退化情形可能发生。一个例子是 $h(x_1, x_2) = x_1 x_2$, $h_1(t) = \mu t$, 其中 μ 为总体均值。于是

$$\xi_1 = \text{Var}(\mu X) = \mu^2 \sigma^2$$

其中 σ^2 为总体方差。当 $\mu = 0$ 时出现退化。但此时 $\xi_2 = \text{Var}(X_1 X_2) = \sigma^4 > 0$, 则此核函数的退化阶数为 2。

定理 2.3(Hoeffding 定理) 证明思路

- 对投影 v_* 用中心极限定理得到正态极限分布 (独立同分布变量和)
- 只要 U 与投影 v_* 足够“近”，应用 Slutsky 定理，得到 U 的极限分布。于是只要验证 Slutsky 定理的条件即可。

Hoeffding 定理证明 (I)

- 证明：首先看投影

$$v_n^* = \frac{r}{n} \sum_{i=1}^n [h_1(X_i) - \theta]$$

$$E(h_1(X_i) - \theta) = 0$$

$$\text{Var}(h_1(X_i)) = \xi_1 < \xi_r = E[h(X_1, \dots, X_r) - \theta]^2$$

$$\text{Var}(v_n^*) = \left(\frac{r}{n}\right)^2 \sum_{i=1}^n \text{Var}(h_1(X_i)) = \frac{r^2}{n} \xi_1$$

- 即 v_n^* 是二阶矩存在的独立同分布随机变量之和，由中心极限定理得到其有正态逼近。

Hoeffding 定理证明 (II)

- 其次验证 Slutsky 定理条件, 即证明

$$E\{\sqrt{n}(U_n - \theta - v_n^*)\}^2 \rightarrow 0$$

- 实际上

$$\begin{aligned} & nE\{(U_n - \theta - v_n^*)\}^2 \\ &= nE(U_n - \theta)^2 + nE(v_n^*)^2 - 2nE[(U_n - \theta)v_n^*] \\ &= 2r^2\xi_1 + o(1) - 2nE[(U_n - \theta)v_n^*] \end{aligned}$$

- 上式中的交叉项

$$\begin{aligned} & nE\{(U_n - \theta)v_n^*\} \\ &= \frac{r}{C_n^r} \sum_{i=1}^n \sum_{(\beta_1, \dots, \beta_r)} E\{[h(X_{\beta_1}, \dots, X_{\beta_r}) - \theta][h_1(X_i) - \theta]\} \end{aligned}$$

Hoeffding 定理证明 (III)

- 当 $i \notin (\beta_1, \dots, \beta_r)$ 时, 上式中两项独立, 期望为 0;
- 当 $i \in (\beta_1, \dots, \beta_r)$ 时,

$$\begin{aligned} & E\{(h(X_{\beta_1}, \dots, X_{\beta_r}) - \theta)(h_1(X_i) - \theta)\} \\ &= E_{X_i} E\{(h(X_{\beta_1}, \dots, X_{\beta_r}) - \theta)(h_1(X_i) - \theta) | X_i\} \\ &= E_{X_i} \{(h_1(X_i) - \theta)^2\} = \xi_1 \end{aligned}$$

$$\begin{aligned} nE\{(U_n - \theta)v_n^*\} &= \frac{r}{C_n^r} \sum_{i=1}^n C_{n-1}^{r-1} \xi_1 \\ &= n \cdot \frac{r}{\frac{n!}{r!(n-r)!}} \cdot \frac{(n-1)!}{(r-1)!(n-r)!} \xi_1 = r^2 \xi_1 \end{aligned}$$

- 综上可得 U 统计量和投影满足 Slutsky 定理条件。

Slusky 定理 (补充)

- **Slusky 定理 (1915):** X_n 和 Y_n 是随机变量序列, 如果 X_n 依分布收敛于 X , Y_n 依概率收敛到常数 b , 则 $X_n + Y_n$ 依分布收敛到 $X + b$, $X_n Y_n$ 依分布收敛到 Xb .
- Slusky 定理首先由前苏联经济学家 Eugen Slusky(1880-1948) 提出, 后来由英国经济学家 John Hicks(1904-1989) 和 Roy Allen(1906-1983) 加以推广。

<https://www.statisticshowto.com/sluskys-theorem/>

Wilcoxon 符号秩统计量的极限分布 (I)

- 由前面的计算, Wilcoxon 符号秩统计量可以表示为

$$W^+ = nU_1 + n(n-1)/2 U_2$$

- 在 H_0 下, $E(U_1) = E(U_2) = 1/2$, 且

$$E(W^+) = \frac{n(n+1)}{4} = nE(U_1) + \frac{n(n-1)}{2}E(U_2)$$

$$(W^+ - EW^+) = n(U_1 - E(U_1)) + \frac{n(n-1)}{2}(U_2 - EU_2)$$

$$\begin{aligned} & \sqrt{n} \left(\frac{W^+ - EW^+}{C_n^2} \right) \\ &= \frac{2\sqrt{n}}{n-1} \cdot (U_1 - EU_1) + \sqrt{n}(U_2 - EU_2) \rightarrow N(0, 4\xi_1) \end{aligned}$$

Wilcoxon 符号秩统计量的极限分布 (II)

- 下面计算 U_2 的 ξ_1 , 由前面计算, 在零假设下 (即 $F(x)$ 关于 0 点对称), $h_1(t) = F(t)$, 于是

$$\begin{aligned}\xi_1 &= \text{Var}(h_1(X)) = E(h_1(X))^2 - [Eh_1(X)]^2 \\&= \int_{-\infty}^{+\infty} F^2(t) dF(t) - \left(\int_{-\infty}^{+\infty} F(t) dF(t) \right)^2 \\&= \frac{1}{3} - \left(\frac{1}{2} \right)^2 = \frac{1}{12}\end{aligned}$$

Wilcoxon 符号秩统计量的极限分布 (III)

- 于是 $\frac{\sqrt{n}}{C_n^2} (W^+ - E(W^+)) \rightarrow N(0, \frac{1}{3})$, 即

$$\frac{W^+ - E(W^+)}{\sqrt{\frac{1}{3n} \cdot C_n^2}} \rightarrow N(0, 1)$$

- 注意到 $Var(W^+) = \frac{n(n+1)(2n+1)}{24}$,

$$\lim_{n \rightarrow +\infty} \frac{\sqrt{Var(W^+)}}{\sqrt{\frac{1}{3n} \cdot C_n^2}} = \lim_{n \rightarrow +\infty} \frac{\sqrt{\frac{n(n+1)(2n+1)}{24}}}{\sqrt{\frac{1}{3n} \cdot \frac{n(n-1)}{2}}} = 1$$

- 得到

$$\frac{W^+ - E(W^+)}{\sqrt{Var(W^+)}} \rightarrow N(0, 1), n \rightarrow +\infty$$

两样本核函数

- **定义 2.4 (两样本核):** 设简单随机样本 $X_1, \dots, X_m \sim F(x)$, $Y_1, \dots, Y_n \sim G(y)$, 参数 θ 称为对分布族 $\mathcal{F} = (F, G)$ 是 (r, s) 可估的, 如果存在函数 $h(x_1, \dots, x_r; y_1, \dots, y_s)$, 使得

$$E_{(F,G)}(h(X_1, \dots, X_r; Y_1, \dots, Y_s)) = \theta,$$

$h(x_1, \dots, x_m; y_1, \dots, y_n)$ 称为核函数。

- 不妨设核函数 $h(\vec{t}_1; \vec{t}_2)$ 对 $\vec{t}_1$ 和 $\vec{t}_2$ 的置换对称, 否则可以对称化。

两样本 U 统计量

- **定义 2.5 (两样本 U 统计量):** 设简单随机样本 $X_1, \dots, X_m \sim F(x)$, $Y_1, \dots, Y_n \sim G(y)$, 参数 θ 是 (r, s) 可估的, 其对称核 $h(x_1, \dots, x_r; y_1, \dots, y_s)$, 由其形成的 U 统计量定义为

$$U(X_1, \dots, X_m; Y_1, \dots, Y_n) = \frac{1}{C_m^r \cdot C_n^s} \sum_{(\alpha_1, \dots, \alpha_r)} \sum_{(\beta_1, \dots, \beta_s)} h(X_{\alpha_1}, \dots, X_{\alpha_r}; Y_{\beta_1}, \dots, Y_{\beta_s})$$

Wilcoxon 符号秩的极限分布 (I)

- 例 1: $\theta = P(X < Y)$ 是 $(1, 1)$ 可估参数, 其对称核为 $\psi(y_1 - x_1)$, 对应的 U 统计量为

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

- 即为 Mann-Whitney 统计量

两样本 U 统计量性质

- 类似于单样本的 U 统计量计算, 可以计算出两样本 U 统计量的方差,

$$\text{Var}(U) = \frac{1}{C_m^r C_n^s} \sum_{c=0}^r \sum_{d=0}^s C_r^c C_{m-r}^{r-c} C_s^d C_{m-s}^{s-d} \xi_{c,d}$$

- 其中

$$\xi_{c,d} = \text{Cov}[h(X_1, \dots, X_c, X_{c+1}, \dots, X_r; Y_1, \dots, Y_d, Y_{d+1}, \dots, Y_s), \\ h(X_1, \dots, X_c, X_{r+1}, \dots, X_{2r-c}; Y_1, \dots, Y_d, Y_{s+1}, \dots, Y_{2s-d})]$$

大样本性质 (I)

- **定理 2.6:** 设简单随机样本 $X_1, \dots, X_m \sim F(x)$, $Y_1, \dots, Y_n \sim G(y)$. $U(X_1, \dots, X_m; Y_1, \dots, Y_n)$ 是 (r, s) 可估参数 θ 的 U 统计量, 其对称核为 $h(x_1, \dots, x_r; y_1, \dots, y_s)$. 若

$$E[h(X_1, \dots, X_r; Y_1, \dots, Y_s)]^2 < +\infty,$$

又当 $N = n + m \rightarrow +\infty$ 时, 有 $m/(n + m) \rightarrow \lambda$, $0 < \lambda < 1$, 则

$$\begin{aligned} & \lim_{N \rightarrow +\infty} N \text{Var}[U(X_1, \dots, X_m; Y_1, \dots, Y_n)] \\ &= \frac{r^2 \xi_{1,0}}{\lambda} + \frac{s^2 \xi_{0,1}}{1 - \lambda} \end{aligned}$$

大样本性质 (II)

- **定理 2.7 (Lehmann-Hoeffding 定理):** 设简单随机样本 $X_1, \dots, X_m \sim F(x)$, $Y_1, \dots, Y_n \sim G(y)$. $U(X_1, \dots, X_m; Y_1, \dots, Y_n)$ 是 (r, s) 可估参数 θ 的 U 统计量, 其对称核为 $h(x_1, \dots, x_r; y_1, \dots, y_s)$. 若

$$E[h(X_1, \dots, X_r; Y_1, \dots, Y_s)]^2 < +\infty,$$

又当 $N = n + m \rightarrow +\infty$ 时, 有 $m/(n + m) \rightarrow \lambda$, $0 < \lambda < 1$, 则当 $r^2\xi_{10}/\lambda + s^2\xi_{01}/(1 - \lambda) > 0$ 时, 有渐近近似

$$\sqrt{N}[U(X_1, \dots, X_m; Y_1, \dots, Y_n) - \theta] \xrightarrow{d} N\left(0, \frac{r^2\xi_{10}}{\lambda} + \frac{s^2\xi_{01}}{(1 - \lambda)}\right).$$

Wilcoxon 秩和统计量的极限分布 (I)

- 由前面的结论

$$W = \frac{n(n+1)}{2} + \sum_{i=1}^m \sum_{j=1}^n \psi(Y_j - X_i) = \frac{n(n+1)}{2} + mnU$$

- 在零假设下, 即 $X_1, \dots, X_m; Y_1, \dots, Y_n$ 独立同分布 $F(x)$ 时

$$\begin{aligned}\xi_{1,0} &= Cov[\psi(Y_1 - X_1), \psi(Y_2 - X_1)] \\ &= P(X_1 < Y_1, X_1 < Y_2) - [P(X_1 < Y_1)]^2 \\ &= \frac{1}{3} - \left(\frac{1}{2}\right)^2 = \frac{1}{12}\end{aligned}$$

$$\begin{aligned}\xi_{0,1} &= Cov[\psi(Y_1 - X_1), \psi(Y_1 - X_2)] \\ &= P(X_1 < Y_1, X_2 < Y_1) - [P(X_1 < Y_1)]^2 \\ &= \frac{1}{3} - \left(\frac{1}{2}\right)^2 = \frac{1}{12}\end{aligned}$$

Wilcoxon 秩和统计量的极限分布 (II)

- 由 Lehmann-Hoeffding 定理, 在 H_0 下,

$$\sqrt{N}(U - \frac{1}{2}) \rightarrow N(0, \frac{1}{12\lambda} + \frac{1}{12(1-\lambda)})$$

- 即

$$\sqrt{12\lambda(1-\lambda)N} \cdot (U - \frac{1}{2}) \rightarrow N(0, 1)$$

- 而

$$\begin{cases} W = mnU + \frac{n(n+1)}{2} \\ E(W) = \frac{n(N+1)}{2}, \text{Var}(W) = \frac{mn(N+1)}{12} \end{cases}$$

Wilcoxon 秩和统计量的极限分布 (III)

- 于是

$$\begin{aligned}\frac{W - E(W)}{\sqrt{\text{Var}(W)}} &= \frac{mnU + \frac{n(n+1)}{2} - \frac{n(n+m+1)}{2}}{\sqrt{\frac{mn(N+1)}{12}}} \\&= \frac{mn(U - \frac{1}{2})}{\sqrt{\frac{mn(N+1)}{12}}} \\&= \frac{\sqrt{\frac{12mn}{N(N+1)}}}{\sqrt{12\lambda(1-\lambda)}} \cdot \left\{ \sqrt{12\lambda(1-\lambda)N} \cdot (U - \frac{1}{2}) \right\} \\&\rightarrow N(0, 1)\end{aligned}$$

作业

- P47 2;
- P47 3;
- P48 4;
- P48 5;