Homework 6

Liu Huihang
SA18017026
QQ: 184050473
MAIL: huihang@mail.ustc.edu.cn

Problem 1 *Let $X_1, \ldots, X_m i.i.d \sim F$, $Y_1, \ldots, Y_n i.i.d \sim G$ and i.i.d within each sample, then*

(1) get the U-statistic U_n with kernel $h(x_1, x_2, y_1, y_2) = I(x_1 < y_1, x_2 < y_2)$,
(2) get the limit distribution of U-statistic U_n with $m + n \to \infty$, $\frac{m}{n+m} \to p \in (0, 1)$,
(3) get the limit distribution of U-statistic U_n under null hypothesis $H_0 : F = G$.

Solution

(1) The kerne is $h(x_1, x_2, y_1, y_2) = I(x_1 < y_1, x_2 < y_2)$, which is of order 2 in both x and y. The corresponding U-statistic is

$$U_n = \frac{1}{\binom{m}{2}\binom{n}{2}} \sum_{i<k}^{m} \sum_{j<l}^{n} I(X_i < Y_j, X_k < Y_l)$$

(2) The projection of $U - \theta$ onto the set of all functions of the form $\sum_{i=1}^{m} k_i(X_i) + \sum_{j=1}^{n} l_j(Y_j)$ is given by

$$\hat{U} = \frac{2}{m} \sum_{i=1}^{m} h_{1,0}(X_i) + \frac{2}{n} \sum_{j=1}^{n} h_{0,1}(Y_j),$$

where the functions $h_{1,0}$ and $h_{0,1}$ are defined by

$$\begin{aligned} h_{1,0}(x) &= Eh(x, X_2, Y_1, Y_2) - \theta, \\ h_{0,1}(y) &= Eh(X_1, X_2, y, Y_2) - \theta. \end{aligned}$$

The sequence $\hat{U}$ is asymtotically normal by the central limit theorem. Then difference between $\hat{U}$ and $U_n - \theta$ is asymtotically negligible.

$$\begin{aligned} \theta &= Eh(X_1, X_2, Y_1, Y_2) \\ &= P(X_1 < Y_1)P(X_2 < Y_2) \\ &= (E[P(X_1 < y|Y_1 = y)])^2 \\ &= (E[F(Y_1)])^2 \\ &= \left(\int F(y)dG(y)\right)^2 \end{aligned}$$

Then calculate $\zeta_{1,0}$ and $\zeta_{0,1}$.

$$
\begin{aligned}
\zeta_{1,0} &= cov(h(X_1, X_2, Y_1, Y_2), h(X_1.X_3, Y_3, Y_4)) \\
&= cov(I(X_1 < Y_1, X_2 < Y_2), I(X_1 < Y_3, X_3 < Y_4)) \\
&= E[I(X_1 < Y_1, X_2 < Y_2)I(X_1 < Y_3, X_3 < Y_4)] \\
&\qquad - E[I(X_1 < Y_1, X_2 < Y_2)]E[I(X_1 < Y_3, X_3 < Y_4)] \\
&= E[I(X_1 < \min(Y_1, Y_3))I(X_2 < Y_2)I(X_3 < Y_4)] \\
&\qquad - E[I(X_1 < Y_1)I(X_2 < Y_2)]E[I(X_1 < Y_3)I(X_3 < Y_4)] \\
&= P(X_1 < \min(Y_1, Y_3))(P(X_1 < Y_1))^2 - (P(X_1 < Y_1))^4 \\
&= \left(\int F(y)dG(y)\right)^2 \int F(z)d(-G^2(z) + 2G(z)) - \left(\int F(y)dG(y)\right)^4
\end{aligned}
$$

and

$$
\begin{aligned}
\zeta_{0,1} &= cov(h(X_1, X_2, Y_1, Y_2), h(X_3.X_4, Y_1, Y_3)) \\
&= cov(I(X_1 < Y_1, X_2 < Y_2), I(X_3 < Y_1, X_4 < Y_3)) \\
&= E[I(X_1 < Y_1, X_2 < Y_2)I(X_3 < Y_1, X_4 < Y_3)] \\
&\qquad - E[I(X_1 < Y_1, X_2 < Y_2)]E[I(X_3 < Y1, X_4 < Y_3)] \\
&= E[I(Y_1 > \max(X_1, X_3))I(Y_2 > X_2)I(Y_3 > X_4)] \\
&\qquad - E[I(X_1 < Y_1)I(X_2 < Y_2)]E[I(X_3 < Y_1)I(X_4 < Y_3)] \\
&= P(Y_1 > \max(X_1, X_3))(P(X_1 < Y_1))^2 - (P(X_1 < Y_1))^4 \\
&= \left(\int F(y)dG(y)\right)^2 \int F^2(z)dG(z) - \left(\int F(y)dG(y)\right)^4
\end{aligned}
$$

Thus,

$$
\sqrt{m+n}(U_n - \theta) \rightsquigarrow N\left(0, \frac{4\zeta_{1,0}}{p} + \frac{4\zeta_{0,1}}{1-p}\right)
$$

where $\theta = \left(\int F(y)dG(y)\right)^2$, $\zeta_{1,0} = \left(\int F(y)dG(y)\right)^2 \int F(z)d(-G^2(z)+2G(z)) - \left(\int F(y)dG(y)\right)^4$ and $\zeta_{0,1} = \left(\int F(y)dG(y)\right)^2 \int F^2(z)dG(z) - \left(\int F(y)dG(y)\right)^4$.

(3) Under $H_0 : F = G$, $\theta = 1/4$, $\zeta_{1,0} = \frac{1}{4*12}$ and $\zeta_{0,1} = \frac{1}{4*12}$.

$$
\sqrt{m+n}\left(U_n - \frac{1}{2}\right) \rightsquigarrow N\left(0, \frac{1}{12p} + \frac{1}{12(1-p)}\right)
$$

■

Problem 2 *Suppose the distribution of X is symmetric about zero with variance $\sigma^2 > 0$ and $EX^4 < \infty$, consider kernel $h(x, y) = xy + (x^2 - \sigma^2)(y^2 - \sigma^2)$, then*

(1) prove that the U-statistic U_n with kernel $h(x, y)$ has a degeneracy of order 1,

(2) get λ_1, λ_2 and orthogonal functions $\Phi_1(x), \Phi_2(x)$, such that $h(x, y) = \lambda_1\varphi_1(x)\varphi_1(y) + \lambda_2\varphi_2(x)\varphi_2(y)$,

(3) get the limit distribution of nU_n.

Solution

(1) Firstly, we obtain U-statistic U_n,

$$U_n = \frac{1}{\binom{n}{2}} \sum_{i<j}^{n} X_i X_j + (X_i^2 - \sigma^2)(X_j^2 - \sigma^2)$$

with $X_1, \dots, X_n$ *i.i.d* and $E[X_i] = \mu = 0, E[X_i^3] = 0$.

Then, we obtain θ,

$$\begin{aligned}\theta &= E[h(X_1, X_2)] \\ &= E[X_1 X_2 + (X_1^2 - \sigma^2)(X_2^2 - \sigma^2)] \\ &= \mu^2 + \mu^4.\end{aligned}$$

Therefore, U_n is an unbiased estimator of $\mu^2 + \mu^4$.

To prove the U-statistic U_n with kernel $h(x, y)$ has a degeneracy of order 1, it's sufficient to show that $\zeta_1 = 0$ and $\zeta_2 > 0$.

Since $h_1(x) = E[xX_2 + (x^2 - \sigma^2)(X_2^2 - \sigma^2)] = x\mu + (x^2 - \sigma^2)\mu^2$ and

$$\begin{aligned}\zeta_1 &= \mathrm{var}[h_1(X_1)] = \mu^2(\mathrm{var}(X_1 + \mu(X_1^2 - \sigma^2))) = 0 \\ \zeta_2 &= \mathrm{var}[h(X_1, X_2)] = \mathrm{var}[X_1 X_2 + (X_1^2 - \sigma^2)(X_2^2 - \sigma^2)] \\ &= \sigma^4 + \left(E[X_1^4] - \sigma^4\right)^2 > 0\end{aligned}$$

so that the degeneracy is of order 1.

(2) We take $A(x_1, x_2) = h(x_1, x_2) - \theta$, where $\theta = Eh(X_1, X_2) = 0$. For $k = 1, 2$,

$$E[A(x, X_2)\phi_k(X_2)] = \lambda_k \phi_k(x). \tag{1}$$

We choose $\phi_1(x) = x/\sigma, \phi_2(x) = (x^2 - \sigma^2)/\sqrt{EX^4 - \sigma^4}$, which satisfy $E\phi_j(X_1) = 0$ and

$$E\phi_j(X_1)\phi_k(X_1) = \begin{cases} 1, & j = k \\ 0, & j \neq k \end{cases}$$

So ϕ_1 is orthogonal with ϕ_2. Thus $\lambda_1 = \sigma^2, \lambda_2 = EX^4 - \sigma^4$ by (1).

(3) U_n is the U-statistic associated with a symmetric kernel of degree 2, degeneracy of order 1 and expectation 0. Then

$$nU_n \rightsquigarrow \sigma^2(Z_1^2 - 1) + (EX^4 - \sigma^4)(Z_2^2 - 1),$$

where Z_1, Z_2 are independent $N(0, 1)$.

■

Problem 3 (Prove the Hoeffding decomposition in page 13) *Let's consider the following U-statistic of order 2,*

$$U_n = \frac{1}{\binom{n}{2}} \sum_{i<j} h(X_i, X_j)$$

The Hoeffding decomposition is:

$$U_n = U + \frac{2}{n}\sum_i h_1(X_i) + \frac{1}{\binom{n}{2}}\sum_{i<j} h_2(X_i, X_j)$$

where

$$\begin{aligned} U &= EU_n = Eh(X_1, X_2), \\ h_1(x) &= Eh(x, X_2) - U, \\ h_2(x,y) &= h(x,y) - h_1(x) - h_1(y) - U. \end{aligned}$$

Solution

According to Problem 4 in the following, we can obtain the Hoeffding decomposition directly,

$$U_n = \sum_{c=0}^{2}\sum_{|A|=c}\sum_{B\subset\{1,c\}} P_A U_n$$

Additional, we have

$$\begin{aligned} P_\emptyset h &= Eh, \\ P_{[i]}h &= E(h|X_i) - Eh, \\ P_{[i,j]}h &= E(h|X_i, X_j) - E(h|X_i) - E(h|X_j) + Eh. \end{aligned}$$

Thus,

$$U_n = P_\emptyset h + \frac{2}{n}\sum_i P_{[i]}h + \frac{1}{\binom{n}{2}}\sum_{i<j} P_{[i,j]}h$$

which leads to

$$U_n = U + \frac{2}{n}\sum_i h_1(X_i) + \frac{1}{\binom{n}{2}}\sum_{i<j} h_2(X_i, X_j)$$

where

$$\begin{aligned} U &= Eh(X_1, X_2), \\ h_1(x) &= Eh(x, X_2) - U, \\ h_2(x,y) &= h(x,y) - h_1(x) - h_1(y) - U. \end{aligned}$$

■

Problem 4 (Prove the T decomposition in page 12) *If $T = T(X_1, ...X_n)$ is permutation symmetric and $X_1, ...X_n$ are independent and identically distributed, then the Hoeffding decomposition of T can be simplified to*

$$T = \sum_{r=0}^{n}\sum_{|A|=r} g_r(X_i : i \in A),$$

where

$$g_r\left(x_1,\ldots,x_r\right)=\sum_{B\subset\{1,\ldots,r\}}(-1)^{r-|B|}\,\mathrm{ET}\left(x_i\in B,X_i\notin B\right).$$

Solution

By the theorem in slide page 8, the projection of T onto H_A is given by

$$P_AT=\sum_{B\subset A}(-1)^{|A|-|B|}E\left(T|X_i:i\in B\right).$$

We have

$$\begin{aligned}T&=\sum_{r=0}^{n}\sum_{|A|=r}P_AT,\\&=\sum_{r=0}^{n}\sum_{|A|=r}\sum_{B\subset A}(-1)^{r-|B|}E\left(T|X_i:i\in B\right),\end{aligned}$$

because $T=T(X_1,.....X_n)$ is permutation symmetric, we have

$$T=\sum_{r=0}^{n}\sum_{|A|=r}\sum_{B\subset\{1,\ldots,r\}}(-1)^{r-|B|}E\left(T|X_i:i\in B\right).$$

This leads to the assertion. ■