EXERCISE 6

WEIYU LI

1. Let $X_1, \ldots, X_m i.i.d. \sim F$**,** $Y_1, \ldots, Y_n i.i.d. \sim G$ **be two independent groups of samples.**

(1) Solve the U statistic U_n *with kernel* $h = I(x_1 < y_1, x_2 < y_2)$*.*

Solve. Since h is not symmetric, symmetrize it as $\tilde{h} = \frac{1}{2}\left(I(x_1 < y_1, x_2 < y_2) + I(x_2 < y_1, x_1 < y_2)\right)$ instead. Then the U statistic is

$$
\begin{aligned}
U_n &= \frac{1}{\binom{m}{2}\binom{n}{2}} \sum_{i_1 < i_2, j_1 < j_2} \frac{1}{2}\left(I(x_{i_1} < y_{j_1}, x_{i_2} < y_{j_2}) + I(x_{i_2} < y_{j_1}, x_{i_1} < y_{j_2})\right) \\
&= \frac{1}{m(m-1)n(n-1)} \sum_{i_1 \neq i_2, j_1 \neq j_2} I(x_{i_1} < y_{j_1}, x_{i_2} < y_{j_2}).
\end{aligned}
$$

□

(2) Solve the limit distribution of U_n *when* $m + n \to \infty$*,* $\frac{m}{n+m} \to p \in (0,1)$*.*

Solve. We use the notations in the slides. U_n estimates $\theta = E\tilde{h} = P(X < Y)^2$, where $X \sim F, Y \sim G$ are independent such that $P(X < Y) = \int (1-G)dF = \int FdG$. And due to the independence and symmetry, the "partial covariances" are

$$
\begin{aligned}
\zeta_{1,0} &= cov(\tilde{h}(X_1, X_2, Y_1, Y_2), \tilde{h}(X_1, X_2', Y_1', Y_2')) \\
&= \big(P(X < Y, X < Y') - \theta\big)\theta \\
&= \Big(\int (1-G)^2 dF - \theta\Big)\theta, \\
\zeta_{0,1} &= cov(\tilde{h}(X_1, X_2, Y_1, Y_2), \tilde{h}(X_1', X_2', Y_1, Y_2')) \\
&= \big(P(X < Y, X' < Y) - \theta\big)\theta \\
&= \Big(\int F^2 dG - \theta\Big)\theta.
\end{aligned}
$$

From the Theorem in P17 in Lec6.pdf, we derive that

$$
\sqrt{n+m}(U_n - \theta) \xrightarrow{d} N\big(0, 4(\frac{\zeta_{1,0}}{p} + \frac{\zeta_{0,1}}{1-p})\big).
$$

□

(3) Solve the asymptotic distribution of U_n *under* $H_0 : F = G$*.*

Solve. Under H_0, $P(X < Y) = \frac{1}{2}, \theta = \frac{1}{4}$ and $\zeta_{1,0} = \zeta_{0,1} = \left(\frac{1}{3} - \frac{1}{4}\right)\frac{1}{4} = \frac{1}{48}$. Therefore,

$$
\sqrt{n+m}(U_n - \frac{1}{4}) \xrightarrow{d} N\big(0, \frac{1}{12p(1-p)}\big).
$$

□

Date: 2019/10/21.

liweiyu@mail.ustc.edu.cn.

2. Suppose the distribution of X is symmetric with respect to the origin, $\sigma^2 = EX^2 > 0$, $EX^4 < \infty$. Consider the kernel $h(x,y) = xy + (x^2 - \sigma^2)(y^2 - \sigma^2)$.

(1) Prove its U statistic U_n has a degeneracy of order 1.

Proof. From the symmetry of the distribution of X, we have $EX = E(X^2 - \sigma^2) = 0$. Then we can derive $\zeta_1 = Var(h_1(x)) = 0$ since $h_1(x) = Eh(x,Y) = 0$.

Next, we show that $\zeta_2 > 0$, which completes the proof. Notice that the symmetric distribution also leads to $EX^3 = 0$. The definition gives that $h_2(x,y) = xy + (x^2 - \sigma^2)(y^2 - \sigma^2)$, thus

$$\begin{aligned}\zeta_2 = Var(h_2) &= E\left(X^2Y^2 + 2XY(X^2-\sigma^2)(Y^2-\sigma^2) + (X^2-\sigma^2)^2(Y^2-\sigma^2)^2\right)\\ &= \sigma^4 + (EX^4 - \sigma^4)^2 > 0.\end{aligned}$$

From the definition, we conclude that U_n has a degeneracy of order 1. □

(2) Solve λ_1, λ_2 and orthogonal $\phi_1(x), \phi_2(x)$, such that $h(x,y) = \lambda_1\phi_1(x)\phi_1(y) + \lambda_2\phi_2(x)\phi_2(y)$.

Solve. First, we observe that $\phi_1(x) = c_1 x$, $\phi_2(x) = c_2(x^2 - \sigma^2)$ are orthogonal from the symmetry of X. To calculate the constants c_1, c_2, the norm-1 property of the basis gives that

$$\begin{aligned}1 &= E\phi_1(X)\phi_1(X) = c_1^2 EX^2 = c_1^2\sigma^2,\\ 1 &= E\phi_2(X)\phi_2(X) = c_2^2 E(X^2-\sigma^2)^2 = c_2^2(EX^4 - \sigma^4).\end{aligned}$$

The solutions are $c_1 = \frac{1}{\sigma}$, $c_2 = \frac{1}{\sqrt{EX^4-\sigma^4}}$. Therefore, we obtain

$$\lambda_1 = \sigma^2, \lambda_2 = EX^4 - \sigma^4, \text{ and } \phi_1(x) = \frac{1}{\sigma}x, \phi_2(x) = \frac{1}{\sqrt{EX^4-\sigma^4}}(x^2 - \sigma^2).$$

□

(3) Solve the asymptotic distribution of nU_n.
Solve. Notice that $\theta = Eh(X,Y) = 0$, we can directly derive from the Theorem in Page 29 of Lec6.pdf that

$$nU_n \to \lambda_1(Z_1^2 - 1) + \lambda_2(Z_2^2 - 1) = \sigma^2(Z_1^2 - 1) + (EX^4 - \sigma^4)(Z_2^2 - 1),$$

where Z_1, Z_2 *i.i.d* $\sim N(0,1)$. □

3. Prove the Hoeffding decomposition in the Example in Page 13 of Lec6.pdf. That is, the Hoeffding decomposition of $U_n = \frac{1}{\binom{n}{2}}\sum_{i<j} h(X_i, X_j)$ 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 $U = EU_n = Eh(X_1, X_2)$, $h_1(x) = Eh(x, X_2) - U$, $h_2(x,y) = h(x,y) - h_1(x) - h_1(y) - U$.

Proof. From Page 7 in Lec6.pdf, the Hajek projections of the first two orders include all the information from U_n, to say,

$$U_n = P_\emptyset U_n + \sum_i P_{[i]} U_n + \sum_{i<j} P_{[i,j]} U_n,$$

where

$$\begin{aligned}
P_\emptyset U_n =& \frac{1}{\binom{n}{2}} \sum_{i<j} EU_n = U, \\
P_{[i]} U_n =& E(U_n|X_i) - U \\
=& \frac{1}{\binom{n}{2}} \left[\sum_{j\neq i} E\left[h(X_i, X_j)|X_i\right] + \sum_{j,k\neq i, j<k} U \right] - U \\
=& \frac{2}{n} \left[E\left[h(X_i, X_j)|X_i\right] - U \right] \\
=& \frac{2}{n} h_1(X_i), \\
P_{[i,j]} U_n =& E(U_n|X_i, X_j) - E(U_n|X_i) - E(U_n|X_j) + EU_n \\
=& \frac{1}{\binom{n}{2}} \left[h(X_i, X_j) - (n-2)Eh(x, X_2)|_{x=X_i} - (n-2)Eh(X_1, x)|_{x=X_j} + (\binom{n}{2} - 1 - 2(n-2))U \right] \\
& - \frac{2}{n} h_1(X_i) - U - \frac{2}{n} h_1(X_j) - U + U \\
=& \frac{1}{\binom{n}{2}} \left[h(X_i, X_j) - h_1(X_i) - h_1(X_j) - U \right].
\end{aligned}$$

□

4. Prove the decomposition of T in Page 12 of Lec6.pdf. That is, if $T = T(X_1, \ldots, X_n)$ is permutation-symmetric and X_i are i.i.d., then

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

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

Proof. From the Theorem in Page 8 of Lec6.pdf, we have

$$\begin{aligned}
T(x_1, \ldots, x_n) =& \sum_{A\subset\{1,\ldots,n\}} P_A T \\
=& \sum_{r=0}^{n} \sum_{|A|=r} \sum_{B\subset A} (-1)^{r-|B|} ET(x_i \in B, X_i \notin B).
\end{aligned}$$

For any $A = \{a_1, \ldots, a_r\}$, there exists a permutation σ, such that $\sigma(i) = a_i, i \in \{1, \ldots, r\}$. Then $ET(x_i \in B, X_i \notin B) = ET(x_{\sigma(i)} \in B, X_{\sigma(i)} \notin B) = ET(x_i \in \sigma^{-1}B, X_i \notin \sigma^{-1}B)$, where $\sigma^{-1}B \subset \sigma^{-1}A = \{1, \dots, r\}$ and summing over $B \subset A$ is equivalent to summing $\tilde{B} = \sigma^{-1}B$ over $\{1, \dots, r\}$. Therefore,

$$\begin{aligned} T(x_1, \dots, x_n) &= \sum_{r=0}^{n} \sum_{|A|=r} \sum_{\tilde{B} \subset \{1,\dots,r\}} (-1)^{r-|\tilde{B}|} ET(x_i \in \tilde{B}, X_i \notin \tilde{B}) \\ &= \sum_{r=0}^{n} \sum_{|A|=r} g_r(x_1, \dots, x_r). \end{aligned}$$

$\square$