STAT 620: Asymptotic Statistics **Spring 2022**

Lecture: Feb 24

Lecturer: Xianyang Zhang

1 Some Examples

1.1 Variance

Following from the previous lecture, we take $h(X_1, X_2) = \frac{1}{2}(X_1 - X_2)^2$. We have seen that $E[h(X_1, X_2)] = \text{var}(X_1)$. Now, note that

$$\begin{aligned} U_n &= \frac{1}{\binom{n}{2}} \sum_{i<j} \frac{1}{2}(X_i - X_j)^2 \\ &= \frac{1}{2\binom{n}{2}} \sum_{i=1}^n \sum_{j=1}^n \frac{1}{2}(X_i - X_j)^2 \\ &= \frac{1}{2n(n-1)} \sum_{i=1}^n \sum_{j=1}^n (X_i - \bar{X} - (X_j - \bar{X}))^2 \\ &= \frac{1}{2n(n-1)} \sum_{i=1}^n \sum_{j=1}^n [(X_i - \bar{X})^2 + (X_j - \bar{X})^2 - 2(X_i - \bar{X})(X_j - \bar{X})] \\ &= \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2. \end{aligned}$$

So U_n turns out to be the well-known unbiased estimator of variance.

1.2 Gini's mean difference

Take $h(X, Y) = |X - Y|$. Parameter $\theta = E[h(X, Y)] = E|X - Y|$ which is a measure of variability.

1.3 Signed rank test

It is used to test the location of a distribution. Here $h(X_1, X_2) = 1\{X_1 + X_2 > 0\}$ and so $\theta = P(X_1 + X_2 > 0)$. The corresponding U-statistic is as follows

$$U_n = \frac{1}{\binom{n}{2}} \sum_{i<j} 1\{X_i + X_j > 0\}.$$

Signed rank test statistic W_n^+ is defined as follows: Let $R_1^+, R_2^+, \ldots, R_n^+$ denote the ranks of $|X_1|, \ldots, |X_n|$. Here $R_i^+ = k$ if $|X_i|$ is the k^{th} smallest observation. Let

$$W_n^+ = \sum_{i=1}^n R_i^+ 1\{X_i > 0\}.$$

U-statistic and signed rank test is connected as follows: Suppose $|X_i| \neq |X_j|$ for any $i \neq j$. Then (exercise left to the reader),

$$W_n^+ = \binom{n}{2} U_n + \sum_{i=1}^{n} 1\{X_i > 0\}.$$

For large n the first term dominates, so asymptotically W_n^+ behaves like $n^2 U_n/2$. Note that if the distribution is continuous and symmetric around zero, then

$$\theta = P(X_1 > -X_2) = \frac{1}{2}.$$

Suppose we want to test the null that the distribution is symmetric around zero (i.e., $\theta = 1/2$) versus the alternative that $\theta > 1/2$. The signed rank test rejects null if W_n^+ is too large, and this is asymptotically equivalent to the test that rejects if U_n is too large.

1.4 Kendall's tau

Kendall's tau is a nonparametric measure of monotone dependence. Given the pairs of observations $(X_1, Y_1), \ldots, (X_n, Y_n)$, Kendall's tau is defined as

$$\tau = \frac{2}{n(n-1)} \sum_{i<j} \{21\{(Y_i - Y_j)(X_i - X_j) > 0\} - 1\}.$$

We think of $[21\{(Y_i - Y_j)(X_i - X_j) > 0\} - 1]$ as $h((X_i, Y_i), (X_j, Y_j))$. The intuition here is that If X and Y are positively (negatively) correlated then τ is closer to $+1(-1)$.

Assume that X_i and Y_j are independent for all i, j. Let $\tilde{X}_{ij} = X_i - X_j$ and $\tilde{Y}_{ij} = Y_i - Y_j$. We now analyze the correlation in τ's expression. By the independence between the two samples, we have

$$P((Y_i - Y_j)(X_i - X_j) > 0) = P(\tilde{X}_{ij}\tilde{Y}_{ij} > 0) = P(\tilde{X}_{ij} > 0)P(\tilde{Y}_{ij} > 0) + P(\tilde{X}_{ij} < 0)P(\tilde{Y}_{ij} < 0).$$

Since $\tilde{X}_{ij} \stackrel{d}{=} -\tilde{X}_{ij}, \tilde{Y}_{ij} \stackrel{d}{=} -\tilde{Y}_{ij}$, we obtain

$$P((Y_i - Y_j)(X_i - X_j) > 0) = \frac{1}{2} \cdot \frac{1}{2} + \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{2}$$

and thus $E[\tau] = 0$.

1.5 Two-sample U-statistic

Given two samples $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$, a two-sample U-statistic is given by

$$U_{n,m} = \frac{1}{\binom{n}{r}\binom{m}{s}} \sum_{|\beta|=r, \beta \subseteq [n]} \sum_{|\alpha|=s, \alpha \subseteq [m]} h(X_\beta, Y_\alpha)$$

where $h : \mathbb{R}^r \times \mathbb{R}^s \to \mathbb{R}$, and h is symmetric in the first r and last s arguments. One notable example is the Mann-Whitney statistic. For $\theta = P(X \leq Y)$ and $h(X, Y) = 1\{X \leq Y\}$, we have

$$U_{n,m} = \frac{1}{mn} \sum_{i=1}^{n} \sum_{j=1}^{m} 1\{X_i \leq Y_j\}.$$

2 Variance of U-statistic

For $c \leq r$ and deterministic $x_1, \ldots, x_c$, define $h_c(x_1, \ldots, x_c) := E[h(x_1, \ldots, x_c, X_{c+1}, \ldots, X_r)]$ where $X_{c+1}, \ldots, X_r \sim^{i.i.d} P$. It's easy to see

1. $h_0 = E[h(X_1, \ldots, X_r)] = \theta$.
2. $E[h_c(X_1, \ldots, X_c)] = E[E[h(X_1, \ldots, X_c, X_{c+1}, \ldots, X_r)|X_1, \ldots, X_c]] = E[h(X_1, \ldots, X_r)] = \theta$.

Let $\hat{h}_c(X_1, \ldots, X_c) := h_c(X_1, \ldots, X_c) - \theta$ and thus $E[\hat{h}_c(X_1, \ldots, X_c)] = 0$. Also let $\xi_c := \text{var}(h_c(X_1, \ldots, X_c))$ and $\xi_0 = 0$.

Goal:

Write $\text{var}(U_n)$ in terms of ξ_c.

2.1 Lemma

If $\alpha, \beta \subseteq [n]$ and $s = \alpha \cap \beta$ with $c = |s|$, then $E[\hat{h}_{|\alpha|}(X_\alpha)\hat{h}_{|\beta|}(X_\beta)] = \xi_c$.

2.2 Proof

We start with a standard fact, for $s \subseteq [n]$ with $|s| = c \leq r$, we have $E[\hat{h}_r(X_s, X_{-s})|X_s] = \hat{h}_c(X_s)$. Now,

$$\begin{aligned}
E[\hat{h}_{|\alpha|}(X_\alpha)\hat{h}_{|\beta|}(X_\beta)] &= E[\hat{h}_{|\alpha|}(X_{\alpha\setminus s}, X_s)\hat{h}_{|\beta|}(X_{\beta\setminus s}, X_s)] \\
&= E[E[\hat{h}_{|\alpha|}(X_{\alpha\setminus s}, X_s)\hat{h}_{|\beta|}(X_{\beta\setminus s}, X_s)|X_s]] \\
&= E[E[\hat{h}_{|\alpha|}(X_{\alpha\setminus s}, X_s)|X_s]E[\hat{h}_{|\beta|}(X_{\beta\setminus s}, X_s)|X_s]] \\
&= E[\hat{h}_c(X_s)\hat{h}_c(X_s)] \\
&= var(\hat{h}_c(X_s)) = \xi_c.
\end{aligned}$$

2.3 Theorem

Let U_n be an r^{th} order U-statistic. Then $var(U_n) = \frac{r^2}{n}\xi_1 + O(\frac{1}{n^2})$.

Proof: We know $U_n - \theta = \frac{1}{\binom{n}{r}} \sum_{|\beta|=r, \beta\subseteq[n]} \hat{h}(X_\beta)$. Note that here we have dropped the subscript $|\beta|$, that is, $\hat{h} \equiv \hat{h}_{|\beta|}$. We have

$$\begin{aligned}
var(U_n) &= \frac{1}{\binom{n}{r}^2} \sum_{|\beta|=r, \beta\subseteq[n]} \sum_{|\alpha|=r, \alpha\subseteq[n]} E[\hat{h}(X_\alpha)\hat{h}(X_\beta)] \\
\text{(by combinatorics)} &= \frac{1}{\binom{n}{r}^2} \sum_{c=1}^{r} \binom{n}{r}\binom{r}{c}\binom{n-r}{r-c}\xi_c \\
&= \frac{1}{\binom{n}{r}^2} \binom{n}{r}\binom{r}{1}\binom{n-r}{r-1}\xi_1 + \frac{1}{\binom{n}{r}^2} \sum_{c=2}^{r} \binom{n}{r}\binom{r}{c}\binom{n-r}{r-c}\xi_c \\
&= \frac{1}{\binom{n}{r}^2} \binom{n}{r}\binom{r}{1}\binom{n-r}{r-1}\xi_1 + O(\frac{1}{n^2}) \\
&= \left(\frac{r^2}{n} + o(1)\right)\xi_1 + O(\frac{1}{n^2}).
\end{aligned}$$