Lecture 18: U- and V-statistics

U-statistics

Let $X_1, ..., X_n$ be i.i.d. from an unknown population P in a nonparametric family $\mathscr{P}$.

If the vector of order statistic is sufficient and complete for $P \in \mathscr{P}$, then a symmetric unbiased estimator of an estimable ϑ is the UMVUE of ϑ.

In many problems, parameters to be estimated are of the form

$$\vartheta = E[h(X_1, ..., X_m)]$$

with a positive integer m and a Borel function h that is symmetric and satisfies $E|h(X_1, ..., X_m)| < \infty$ for any $P \in \mathscr{P}$.

An effective way of obtaining an unbiased estimator of ϑ (which is a UMVUE in some nonparametric problems) is to use

$$U_n = \binom{n}{m}^{-1} \sum_c h(X_{i_1}, ..., X_{i_m}), \qquad (1)$$

where $\sum_c$ denotes the summation over the $\binom{n}{m}$ combinations of m distinct elements $\{i_1, ..., i_m\}$ from $\{1, ..., n\}$.

Definition 3.2

The statistic in (1) is called a *U-statistic* with kernel h of order m.

Examples

Consider the estimation of μ^m, where $\mu = EX_1$ and m is an integer > 0. Using $h(x_1,...,x_m) = x_1 \cdots x_m$, we obtain the following U-statistic for μ^m:

$$U_n = \binom{n}{m}^{-1} \sum_c X_{i_1} \cdots X_{i_m}.$$

Consider next the estimation of

$$\sigma^2 = [\,\mathrm{Var}(X_1) + \mathrm{Var}(X_2)]/2 = E[(X_1 - X_2)^2/2],$$

we obtain the following U-statistic with kernel $h(x_1,x_2) = (x_1 - x_2)^2/2$:

$$U_n = \frac{2}{n(n-1)} \sum_{1 \le i < j \le n} \frac{(X_i - X_j)^2}{2} = \frac{1}{n-1}\left(\sum_{i=1}^n X_i^2 - n\bar{X}^2\right) = S^2,$$

which is the sample variance.

Examples

In some cases, we would like to estimate $\vartheta = E|X_1 - X_2|$, a measure of concentration.
Using kernel $h(x_1, x_2) = |x_1 - x_2|$, we obtain the following U-statistic unbiased for $\vartheta = E|X_1 - X_2|$:

$$U_n = \frac{2}{n(n-1)} \sum_{1 \le i < j \le n} |X_i - X_j|,$$

which is known as *Gini's mean difference*.

Let $\vartheta = P(X_1 + X_2 \le 0)$.
Using kernel $h(x_1, x_2) = I_{(-\infty,0]}(x_1 + x_2)$, we obtain the following U-statistic unbiased for ϑ:

$$U_n = \frac{2}{n(n-1)} \sum_{1 \le i < j \le n} I_{(-\infty,0]}(X_i + X_j),$$

which is known as the *one-sample Wilcoxon statistic*.

Variance of a U-statistic

The variance of a U-statistic U_n with kernel h has an explicit form.
For $k = 1, ..., m$, let

$$\begin{aligned} h_k(x_1, ..., x_k) &= E[h(X_1, ..., X_m)|X_1 = x_1, ..., X_k = x_k] \\ &= E[h(x_1, ..., x_k, X_{k+1}, ..., X_m)] \\ \tilde{h}_k &= h_k - E[h(X_1, ..., X_m)] \end{aligned}$$

For any U-statistic with kernel h,

$$U_n - E(U_n) = \binom{n}{m}^{-1} \sum_c \tilde{h}(X_{i_1}, ..., X_{i_m}). \quad (2)$$

Theorem 3.4 (Hoeffding's theorem)

For a U-statistic U_n with $E[h(X_1, ..., X_m)]^2 < \infty$,

$$\mathrm{Var}(U_n) = \binom{n}{m}^{-1} \sum_{k=1}^{m} \binom{m}{k} \binom{n-m}{m-k} \zeta_k,$$

where $\zeta_k = \mathrm{Var}(h_k(X_1, ..., X_k))$.

Proof

Consider two sets $\{i_1, ..., i_m\}$ and $\{j_1, ..., j_m\}$ of m distinct integers from $\{1, ..., n\}$ with exactly k integers in common.
The number of distinct choices of two such sets is $\binom{n}{m}\binom{m}{k}\binom{n-m}{m-k}$.
By the symmetry of $\tilde{h}_m$ and independence of $X_1, ..., X_n$,

$$E[\tilde{h}(X_{i_1}, ..., X_{i_m})\tilde{h}(X_{j_1}, ..., X_{j_m})] = \zeta_k$$

for $k = 1, ..., m$.
Then, by (2),

$$\begin{aligned} \mathrm{Var}(U_n) &= \binom{n}{m}^{-2} \sum_c \sum_c E[\tilde{h}(X_{i_1}, ..., X_{i_m})\tilde{h}(X_{j_1}, ..., X_{j_m})] \\ &= \binom{n}{m}^{-2} \sum_{k=1}^{m} \binom{n}{m}\binom{m}{k}\binom{n-m}{m-k}\zeta_k. \end{aligned}$$

This proves the result.

Corollary 3.2

Under the condition of Theorem 3.4,

(i) $\frac{m^2}{n}\zeta_1 \leq \mathrm{Var}(U_n) \leq \frac{m}{n}\zeta_m$;

(ii) $(n+1)\,\mathrm{Var}(U_{n+1}) \leq n\,\mathrm{Var}(U_n)$ for any $n > m$;

(iii) For any fixed m and $k = 1,...,m$, if $\zeta_j = 0$ for $j < k$ and $\zeta_k > 0$, then

$$\mathrm{Var}(U_n) = \frac{k!\binom{m}{k}^2\zeta_k}{n^k} + O\left(\frac{1}{n^{k+1}}\right).$$

For any fixed m, if $\zeta_j = 0$ for $j < k$ and $\zeta_k > 0$, then the mse of U_n is of the order n^{-k} and, therefore, U_n is $n^{k/2}$-consistent.

Example 3.11

Consider $h(x_1,x_2) = x_1x_2$, the U-statistic unbiased for μ^2, $\mu = EX_1$.
Note that $h_1(x_1) = \mu x_1$, $\tilde{h}_1(x_1) = \mu(x_1 - \mu)$,
$\zeta_1 = E[\tilde{h}_1(X_1)]^2 = \mu^2\,\mathrm{Var}(X_1) = \mu^2\sigma^2$, $\tilde{h}(x_1,x_2) = x_1x_2 - \mu^2$, and
$\zeta_2 = \mathrm{Var}(X_1X_2) = E(X_1X_2)^2 - \mu^4 = (\mu^2+\sigma^2)^2 - \mu^4$.
By Theorem 3.4, for $U_n = \binom{n}{2}^{-1}\sum_{1\leq i<j\leq n} X_iX_j$,

$$
\begin{aligned}
\mathrm{Var}(U_n) &= \binom{n}{2}^{-1}\left[\binom{2}{1}\binom{n-2}{1}\zeta_1 + \binom{2}{2}\binom{n-2}{0}\zeta_2\right] \\
&= \frac{2}{n(n-1)}\left[2(n-2)\mu^2\sigma^2 + (\mu^2+\sigma^2)^2 - \mu^4\right] \\
&= \frac{4\mu^2\sigma^2}{n} + \frac{2\sigma^4}{n(n-1)}.
\end{aligned}
$$

Next, consider $h(x_1,x_2) = I_{(-\infty,0]}(x_1+x_2)$, which leads to the one-sample Wilcoxon statistic.

Note that $h_1(x_1) = P(x_1 + X_2 \le 0) = F(-x_1)$, where F is the c.d.f. of P.

Then $\zeta_1 = \mathrm{Var}(F(-X_1))$.

Let $\vartheta = E[h(X_1,X_2)]$.

Then $\zeta_2 = \mathrm{Var}(h(X_1,X_2)) = \vartheta(1-\vartheta)$.

Hence, for U_n being the one-sample Wilcoxon statistic,

$$
\mathrm{Var}(U_n) = \frac{2}{n(n-1)}\left[2(n-2)\zeta_1 + \vartheta(1-\vartheta)\right].
$$

If F is continuous and symmetric about 0, then ζ_1 can be simplified as

$$
\zeta_1 = \mathrm{Var}(F(-X_1)) = \mathrm{Var}(1-F(X_1)) = \mathrm{Var}(F(X_1)) = \tfrac{1}{12}
$$

Asymptotic distributions of U-statistics

For nonparametric $\mathscr{P}$, the exact distribution of U_n is hard to derive. We study the method of *projection*, which is particularly effective for studying asymptotic distributions of U-statistics.

Definition 3.3

Let T_n be a given statistic based on $X_1,...,X_n$. The projection of T_n on k_n random elements $Y_1,...,Y_{k_n}$ is defined to be

$$\check{T}_n = E(T_n) + \sum_{i=1}^{k_n} [E(T_n|Y_i) - E(T_n)].$$

Let $\check{T}_n$ be the projection of T_n on $X_1,...,X_n$, and $\psi_n(X_i) = E(T_n|X_i)$.
If T_n is symmetric (as a function of $X_1,...,X_n$), then $\psi_n(X_1),...,\psi_n(X_n)$ are i.i.d. with mean $E[\psi_n(X_i)] = E(\check{T}_n) = E[E(T_n|X_i)] = E(T_n)$.
If $E(T_n^2) < \infty$ and $\mathrm{Var}(\psi_n(X_i)) > 0$, then, by the CLT,

$$\frac{1}{\sqrt{n\,\mathrm{Var}(\psi_n(X_1))}} \sum_{i=1}^{n} [\psi_n(X_i) - E(T_n)] \to_d N(0,1) \tag{3}$$

If we can show $T_n - \check{T}_n$ has a negligible order, then we can derive the asymptotic distribution of T_n by using (3) and Slutsky's theorem.

Lemma 3.1

Let T_n be a symmetric statistic with $\mathrm{Var}(T_n) < \infty$ for every n and $\check{T}_n$ be the projection of T_n on $X_1, ..., X_n$.
Then $E(T_n) = E(\check{T}_n)$ and

$$E(T_n - \check{T}_n)^2 = \mathrm{Var}(T_n) - \mathrm{Var}(\check{T}_n).$$

Proof

Since $E(T_n) = E(\check{T}_n)$,

$$E(T_n - \check{T}_n)^2 = \mathrm{Var}(T_n) + \mathrm{Var}(\check{T}_n) - 2\,\mathrm{Cov}(T_n, \check{T}_n)$$

$$\begin{aligned}
\mathrm{Cov}(T_n, \check{T}_n) &= E(T_n \check{T}_n) - [E(T_n)]^2 \\
&= nE[T_n E(T_n|X_i)] - n[E(T_n)]^2 \\
&= nE\{E[T_n E(T_n|X_i)|X_i]\} - n[E(T_n)]^2 \\
&= nE\{[E(T_n|X_i)]^2\} - n[E(T_n)]^2 \\
&= n\,\mathrm{Var}(E(T_n|X_i)) \;=\; \mathrm{Var}(\check{T}_n)
\end{aligned}$$

For a U-statistic U_n, one can show (exercise) that

$$\check{U}_n = E(U_n) + \frac{m}{n}\sum_{i=1}^{n} \tilde{h}_1(X_i),$$

where $\check{U}_n$ is the projection of U_n on $X_1,...,X_n$ and

$$\tilde{h}_1(x) = h_1(x) - E[h(X_1,...,X_m)], \quad h_1(x) = E[h(x,X_2,...,X_m)].$$

Hence, if $\zeta_1 = \mathrm{Var}(\tilde{h}_1(X_i)) > 0$,

$$\mathrm{Var}(\check{U}_n) = m^2\zeta_1/n$$

and, by Corollary 3.2 and Lemma 3.1,

$$E(U_n - \check{U}_n)^2 = O(n^{-2}).$$

This is enough for establishing the asymptotic distribution of U_n.

If $\zeta_1 = 0$ but $\zeta_2 > 0$, then we can show that

$$E(U_n - \check{U}_n)^2 = O(n^{-3}).$$

One may derive results for the cases where $\zeta_2 = 0$, but the case of either $\zeta_1 > 0$ or $\zeta_2 > 0$ is the most interesting case in applications.

Theorem 3.5

Let U_n be a U-statistic with $E[h(X_1,...,X_m)]^2 < \infty$.

(i) If $\zeta_1 > 0$, then

$$\sqrt{n}[U_n - E(U_n)] \to_d N(0, m^2\zeta_1).$$

(ii) If $\zeta_1 = 0$ but $\zeta_2 > 0$, then

$$n[U_n - E(U_n)] \to_d \frac{m(m-1)}{2} \sum_{j=1}^{\infty} \lambda_j(\chi^2_{1j} - 1), \tag{4}$$

where χ^2_{1j}'s are i.i.d. random variables having the chi-square distribution χ^2_1 and λ_j's are some constants (which may depend on P) satisfying $\sum_{j=1}^{\infty} \lambda_j^2 = \zeta_2$.

Lemma 3.2

Let Y be the random variable on the right-hand side of (4).
Then $EY^2 = \frac{m^2(m-1)^2}{2}\zeta_2$.

It follows from Corollary 3.2(iii) and Lemma 3.2 that if $\zeta_1 = 0$, then

$$\mathrm{amse}_{U_n}(P) = \frac{m^2(m-1)^2}{2}\zeta_2/n^2 = \mathrm{Var}(U_n) + O(n^{-3})$$

Proof of Lemma 3.2.

Define

$$Y_k = \frac{m(m-1)}{2}\sum_{j=1}^{k}\lambda_j(\chi^2_{1j} - 1), \quad k = 1, 2, \ldots.$$

It can be shown (exercise) that $\{Y_k^2\}$ is uniformly integrable.
Since $Y_k \to_d Y$ as $k \to \infty$, $\lim_{k\to\infty} EY_k^2 = EY^2$ (Theorem 1.8(viii)).
Since χ^2_{1j}'s are independent chi-square random variables with $E\chi^2_{1j} = 1$ and $\mathrm{Var}(\chi^2_{1j}) = 2$, $EY_k = 0$ for any k and

$$\begin{aligned} EY_k^2 &= \frac{m^2(m-1)^2}{4}\sum_{j=1}^{k}\lambda_j^2\,\mathrm{Var}(\chi^2_{1j}) \\ &= \frac{m^2(m-1)^2}{4}\left(2\sum_{j=1}^{k}\lambda_j^2\right) \\ &\to \frac{m^2(m-1)^2}{2}\zeta_2. \end{aligned}$$

A statistic closely related to U-statistic is described as follows.

V-statistics

Let $X_1, ..., X_n$ be i.i.d. from P.
For every U-statistic U_n as an estimator of $\vartheta = E[h(X_1, ..., X_m)]$, there is a closely related *V-statistic* defined by

$$V_n = \frac{1}{n^m} \sum_{i_1=1}^{n} \cdots \sum_{i_m=1}^{n} h(X_{i_1}, ..., X_{i_m}). \tag{5}$$

As an estimator of ϑ, V_n is biased; but the bias is small asymptotically.
For a fixed n, V_n may be better than U_n in terms of the mse.

Proposition 3.5

Let V_n be defined by (5).

(i) Assume that $E|h(X_{i_1}, ..., X_{i_m})| < \infty$ for all $1 \le i_1 \le \cdots \le i_m \le m$.
Then the bias of V_n satisfies

$$b_{V_n}(P) = O(n^{-1}).$$

Proposition 3.5 (continued)

(ii) Assume that $E[h(X_{i_1},...,X_{i_m})]^2 < \infty$ for all $1 \le i_1 \le \cdots \le i_m \le m$. Then the variance of V_n satisfies

$$\mathrm{Var}(V_n) = \mathrm{Var}(U_n) + O(n^{-2}),$$

where U_n is the U-statistic corresponding to V_n.

Theorem 3.16

Let V_n be a V-statistic with $E[h(X_{i_1},...,X_{i_m})]^2 < \infty$ for all $1 \le i_1 \le \cdots \le i_m \le m$.

(i) If $\zeta_1 = \mathrm{Var}(h_1(X_1)) > 0$, then $\sqrt{n}(V_n - \vartheta) \to_d N(0, m^2\zeta_1)$.

(ii) If $\zeta_1 = 0$ but $\zeta_2 = \mathrm{Var}(h_2(X_1, X_2)) > 0$, then

$$n(V_n - \vartheta) \to_d \frac{m(m-1)}{2} \sum_{j=1}^{\infty} \lambda_j \chi^2_{1j},$$

where χ^2_{1j}'s and λ_j's are the same as those in Theorem 3.5.

Discussion

- Theorem 3.16 shows that if $\zeta_1 > 0$, then the amse's of U_n and V_n are the same.
- If $\zeta_1 = 0$ but $\zeta_2 > 0$, then an argument similar to that in the proof of Lemma 3.2 leads to

$$\begin{aligned}\mathrm{amse}_{V_n}(P) &= \frac{m^2(m-1)^2\zeta_2}{2n^2} + \frac{m^2(m-1)^2}{4n^2}\left(\sum_{j=1}^{\infty}\lambda_j\right)^2 \\ &= \mathrm{amse}_{U_n}(P) + \frac{m^2(m-1)^2}{4n^2}\left(\sum_{j=1}^{\infty}\lambda_j\right)^2\end{aligned}$$

  (see Lemma 3.2).
- Hence U_n is asymptotically more efficient than V_n, unless $\sum_{j=1}^{\infty}\lambda_j = 0$.

Example.

Let $X_1,...,X_n$ be i.i.d. from a population with mean μ, variance σ^2, and finite 4th moment.

To estimate μ^2, the U-statistic and the corresponding V-statistic are

$$U_n = \frac{2}{n(n-1)} \sum_{i<j} X_i X_j, \qquad V_n = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} X_i X_j = \bar{X}^2$$

We now compare U_n and V_n.
Note that $\zeta_1 = \mu^2 \sigma^2$.
If $\mu \neq 0$, by the CLT and delta-method,

$$\sqrt{n}(V_n - \mu^2) = \sqrt{n}(\bar{X}^2 - \mu^2) \to_d N(0, 4\mu^2\sigma^2)$$

For U_n, the result in Theorem 3.5(i) holds with $\zeta_1 = \mu^2\sigma^2$, i.e.,

$$\sqrt{n}(U_n - \mu^2) \to_d N(0, 2^2\zeta_1) = N(0, 4\mu^2\sigma^2)$$

Thus, U_n and V_n are asymptotically the same.

Now consider $\mu = 0$.

Note that $\zeta_1 = 0$, $\zeta_2 = \sigma^4 > 0$, and Theorems 3.5(ii) and 3.16(ii) apply.

However, it is not convenient to use Theorems 3.5(ii) and 3.16(ii) to find the limiting distributions of U_n and V_n.

For V_n, by the CLT and Theorem 1.10,

$$nV_n/\sigma^2 = n\bar{X}^2/\sigma^2 \to_d \chi_1^2$$

where χ_1^2 is a random variable having the chi-square distribution χ_1^2.
Note that

$$\frac{n\bar{X}^2}{\sigma^2} = \frac{1}{\sigma^2 n}\sum_{i=1}^n X_i^2 + \frac{(n-1)U_n}{\sigma^2}.$$

By the SLLN,

$$\frac{1}{\sigma^2 n}\sum_{i=1}^n X_i^2 \to_{a.s.} 1.$$

An application of Slutsky's theorem leads to

$$nU_n/\sigma^2 \to_d \chi_1^2 - 1.$$

Since $\mu = 0$, by Theorem 3.5(ii),

$$nU_n \to_d \sum_{j=1}^\infty \lambda_j(\chi_{1j}^2 - 1)$$

which implies that $\lambda_1 = \sigma^2$ and $\lambda_j = 0$ when $j > 1$.

The amse of U_n is $2\sigma^4/n^2$ whereas the amse of V_n is $3\sigma^4/n^2$.