Lecture 33: U-statistics and their variances

Let $X_1, \ldots, X_n$ be i.i.d. from an unknown population $P$ in a nonparametric family $\mathcal{P}$. If the vector of order statistic is sufficient and complete for $P \in \mathcal{P}$, then a symmetric unbiased estimator of any estimable $\vartheta$ is the UMVUE of $\vartheta$.

In a large class of problems, parameters to be estimated are of the form

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

with a positive integer $m$ and a Borel function $h$ that is symmetric and satisfies

$$E|h(X_1, \ldots, X_m)| < \infty$$

for any $P \in \mathcal{P}$.

It is easy to see that a symmetric unbiased estimator of $\vartheta$ is

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

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

**Definition 3.2.** The statistic $U_n$ in (1) is called a *U-statistic* with kernel $h$ of order $m$.

The use of U-statistics is an effective way of obtaining unbiased estimators. In nonparametric problems, U-statistics are often UMVUE’s, whereas in parametric problems, U-statistics can be used as initial estimators to derive more efficient estimators.

If $m = 1$, $U_n$ in (1) is simply a type of sample mean. Examples include the empirical c.d.f. evaluated at a particular $t$ and the sample moments $n^{-1} \sum_{i=1}^{n} X_i^k$ for a positive integer $k$.

Consider the estimation of $\vartheta = \mu^m$, where $\mu = EX_1$ and $m$ is a positive integer. Using $h(x_1, \ldots, x_m) = x_1 \cdots x_m$, we obtain the following U-statistic unbiased for $\vartheta = \mu^m$:

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

Consider the estimation of $\vartheta = \sigma^2 = \text{Var}(X_1)$. Since

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

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

$$U_n = \frac{2}{n(n-1)} \sum_{1 \leq i < j \leq n} \frac{(X_i - X_j)^2}{2} = \frac{1}{n-1} \left( \sum_{i=1}^{n} X_i^2 - n \bar{X}^2 \right) = S^2,$$
which is the sample variance.

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 \( \vartheta = E|X_1 - X_2| \):

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

which is known as Gini’s mean difference.

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

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

which is known as the one-sample Wilcoxon statistic.

If \( E[h(X_1, ..., X_m)]^2 < \infty \), then the variance of \( U_n \) in (1) with kernel \( h \) has an explicit form. To derive \( \text{Var}(U_n) \), we need some notation.

For \( k = 1, ..., m \), let

\[
h_k(x_1, ..., x_m) = E[h(X_1, ..., X_m)|X_1 = x_1, ..., X_k = x_k]
\]

\[
= E[h(x_1, ..., x_k, X_{k+1}, ..., X_m)].
\]

Note that \( h_m = h \).

It can be shown that

\[
h_k(x_1, ..., x_k) = E[h_{k+1}(x_1, ..., x_k, X_{k+1})].
\]

Define

\[
\tilde{h}_k = h_k - E[h(X_1, ..., X_m)],
\]

\( k = 1, ..., m \), and \( \tilde{h} = \tilde{h}_m \).

Then, for any \( U_n \) defined by (1),

\[
U_n - E(U_n) = \binom{n}{m}^{-1} \sum_{i_1 < ... < i_m} \tilde{h}(X_{i_1}, ..., X_{i_m}).
\]

**Theorem 3.4** (Hoeffding’s theorem). For a U-statistic \( U_n \) given by (1) with \( E[h(X_1, ..., X_m)]^2 < \infty \),

\[
\text{Var}(U_n) = \binom{n}{m}^{-1} \sum_{k=1}^{m} \binom{m}{k} \binom{n-m}{m-k} \zeta_k,
\]

where

\[
\zeta_k = \text{Var}(h_k(X_1, ..., X_k)).
\]

**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 \( X_1, \ldots, X_n \),

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

for \( k = 1, \ldots, m \).

Then, by (5),

\[
\text{Var}(U_n) = \left( \frac{n}{m} \right)^{-2} \sum_{c} \sum_{c} E[\tilde{h}(X_{i_1}, \ldots, X_{i_m})\tilde{h}(X_{j_1}, \ldots, X_{j_m})]
\]

\[
= \left( \frac{n}{m} \right)^{-2} \sum_{k=1}^{m} \binom{n}{m} \binom{m}{k} \binom{n-m}{m-k} \zeta_k.
\]

This proves the result.

**Corollary 3.2.** Under the condition of Theorem 3.4,

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

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

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

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

It follows from Corollary 3.2 that a U-statistic \( U_n \) as an estimator of its mean is consistent in mse (under the finite second moment assumption on \( h \)).

In fact, 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 first \( h(x_1, x_2) = x_1x_2 \), which leads to a U-statistic unbiased for \( \mu^2, \mu = EX_1 \).

Note that \( \tilde{h}_1(x_1) = \mu x_1, \tilde{h}_1(x_1) = \mu(x_1 - \mu), \zeta_1 = E[\tilde{h}_1(X_1)]^2 = \mu^2 \text{Var}(X_1) = \mu^2 \sigma^2, \)

\( \tilde{h}_2(x_1, x_2) = x_1x_2 - \mu^2 \), and \( \zeta_2 = \text{Var}(X_1X_2) = E(X_1X_2)^2 - \mu^4 = (\mu^2 + \sigma^2)^2 - \mu^4 \).

By Theorem 3.4, for \( U_n = \binom{n}{2}^{-1} \sum_{1 \leq i < j \leq n} X_iX_j \),

\[
\text{Var}(U_n) = \binom{n}{2}^{-1} \left[ \binom{2}{1} \binom{n-2}{1} \zeta_1 + \binom{2}{2} \binom{n-2}{0} \zeta_2 \right]
\]

\[
= \frac{2}{n(n-1)} \left[ 2(n-2)\mu^2 \sigma^2 + (\mu^2 + \sigma^2)^2 - \mu^4 \right]
\]

\[
= \frac{4\mu^2 \sigma^2}{n} + \frac{2\sigma^4}{n(n-1)}.
\]

Comparing \( U_n \) with \( \bar{X}^2 - \sigma^2/n \) in Example 3.10, which is the UMVUE under the normality and known \( \sigma^2 \) assumption, we find that

\[
\text{Var}(U_n) - \text{Var}(\bar{X}^2 - \sigma^2/n) = \frac{2\sigma^4}{n^2(n-1)}.
\]
Next, consider $h(x_1, x_2) = I_{(-\infty,0]}(x_1 + x_2)$, which leads to the one-sample Wilcoxon statistic. Note that $h_1(x_1) = P(x_1 + X_2 \leq 0) = F(-x_1)$, where $F$ is the c.d.f. of $P$. Then $\zeta_1 = \text{Var}(F(-X_1))$.

Let $\vartheta = E[h(X_1, X_2)]$.
Then $\zeta_2 = \text{Var}(h(X_1, X_2)) = \vartheta(1 - \vartheta)$.

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

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

If $F$ is continuous and symmetric about 0, then $\zeta_1$ can be simplified as

$$\zeta_1 = \text{Var}(F(-X_1)) = \text{Var}(1 - F(X_1)) = \text{Var}(F(X_1)) = \frac{1}{12},$$

since $F(X_1)$ has the uniform distribution on $[0,1]$.

Finally, consider $h(x_1, x_2) = |x_1 - x_2|$, which leads to Gini’s mean difference. Note that

$$h_1(x_1) = E|x_1 - X_2| = \int |x_1 - y|dP(y),$$

and

$$\zeta_1 = \text{Var}(h_1(X_1)) = \int \left[ \int |x - y|dP(y) \right]^2 dP(x) - \vartheta^2,$$

where $\vartheta = E|X_1 - X_2|$.