U-statistics

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 an estimable \( \vartheta \) is the UMVUE of \( \vartheta \).

In many 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 \( P \in \mathcal{P} \).

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

\[
U_n = \left( \binom{n}{m} \right)^{-1} \sum_{\mathcal{C}} h(X_{i_1}, \ldots, X_{i_m}), \tag{1}
\]

where \( \sum_{\mathcal{C}} \) 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 in (1) is called a *U-statistic* with kernel $h$ of order $m$.

Examples

Consider the estimation of $\mu^m$, where $\mu = E X_1$ and $m$ is an integer $> 0$. Using $h(x_1, \ldots, x_m) = x_1 \cdots x_m$, we obtain the following U-statistic for $\mu^m$:

$$U_n = \left( \binom{n}{m} \right)^{-1} \sum c X_{i_1} \cdots X_{i_m}.$$  

Consider next the estimation of

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

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

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

which is known as Gini’s mean difference.

Let \( \varrho = P(X_1 + X_2 \leq 0) \).

Using kernel \( h(x_1, x_2) = I_{(-\infty,0]}(x_1 + x_2) \), we obtain the following U-statistic unbiased for \( \varrho \):

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

which is known as the one-sample Wilcoxon statistic.
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

$$h_k(x_1, ..., x_k) = E[h(X_1, ..., X_m)|X_1 = x_1, ..., X_k = x_k] = E[h(x_1, ..., x_k, X_{k+1}, ..., X_m)]$$

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

For any U-statistic with kernel $h$,

$$U_n - E(U_n) = \left(\begin{array}{c} n \\ m \end{array}\right)^{-1} \sum_c \tilde{h}(X_{i_1}, ..., X_{i_m}).$$ (2)

Theorem 3.4 (Hoeffding’s theorem)

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

$$\text{Var}(U_n) = \left(\begin{array}{c} n \\ m \end{array}\right)^{-1} \sum_{k=1}^{m} \left(\begin{array}{c} m \\ k \end{array}\right) \left(\begin{array}{c} n-m \\ m-k \end{array}\right) \zeta_k,$$

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

Consider two sets \( \{i_1, \ldots, i_m\} \) and \( \{j_1, \ldots, j_m\} \) of \( m \) distinct integers from \( \{1, \ldots, n\} \) 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 (2),

\[
\text{Var}(U_n) = \binom{n}{m}^{-2} \sum_c \sum_c E[\tilde{h}(X_{i_1}, \ldots, X_{i_m}) \tilde{h}(X_{j_1}, \ldots, X_{j_m})]
\]

\[
= \binom{n}{m}^{-2} \sum_{k=1}^m \binom{n}{m} \binom{m}{k} \binom{n-m}{m-k} \zeta_k.
\]

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

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

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 \( h(x_1, x_2) = x_1 x_2 \), the U-statistic unbiased for \( \mu^2, \mu = EX_1. \)

Note that \( h_1(x_1) = \mu x_1, \hat{h}_1(x_1) = \mu (x_1 - \mu), \)

\( \zeta_1 = E[\hat{h}_1(X_1)]^2 = \mu^2 \text{Var}(X_1) = \mu^2 \sigma^2, \)

\( \hat{h}_1(x_1, x_2) = x_1 x_2 - \mu^2, \) and \( \zeta_2 = \text{Var}(X_1 X_2) = E(X_1 X_2)^2 - \mu^4 = (\mu^2 + \sigma^2)^2 - \mu^4. \)

By Theorem 3.4, for \( U_n = \binom{n}{2}^{-1} \sum_{1 \leq i < j \leq n} X_i X_j, \)
\[
\text{Var}(U_n) = \left( \frac{n}{2} \right)^{-1} \left[ \left( \frac{2}{1} \right) \left( \frac{n-2}{1} \right) \zeta_1 + \left( \frac{2}{2} \right) \left( \frac{n-2}{0} \right) \zeta_2 \right]
\]
\[
= \frac{2}{n(n-1)} \left[ 2(n-2)\mu^2 \sigma^2 + (\mu^2 + \sigma^2)^2 - \mu^4 \right]
\]
\[
= \frac{4\mu^2 \sigma^2}{n} + \frac{2\sigma^4}{n(n-1)}.
\]

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

\[
\zeta_1 = \text{Var}(F(-X_1)) = \text{Var}(1 - F(X_1)) = \text{Var}(F(X_1)) = \frac{1}{12}.
\]
Asymptotic distributions of U-statistics

For nonparametric $P$, 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 $X_1, \ldots, X_n$. The projection of $T_n$ on $k_n$ random elements $Y_1, \ldots, Y_{k_n}$ is defined to be

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

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

$$
\frac{1}{\sqrt{n \text{Var}(\psi_n(X_1))}} \sum_{i=1}^{n} [\psi_n(X_i) - E(T_n)] \rightarrow^d N(0, 1)
$$

(3)
If we can show $T_n - \tilde{T}_n$ 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 $\text{Var}(T_n) < \infty$ for every $n$ and $\tilde{T}_n$ be the projection of $T_n$ on $X_1, \ldots, X_n$. Then $E(T_n) = E(\tilde{T}_n)$ and

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

**Proof**

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

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

$$\text{Cov}(T_n, \tilde{T}_n) = E(T_n \tilde{T}_n) - [E(T_n)]^2$$

$$= nE[T_n E(T_n | X_i)] - n[E(T_n)]^2$$

$$= nE\{E[T_n E(T_n | X_i) | X_i]\} - n[E(T_n)]^2$$

$$= nE\{[E(T_n | X_i)]^2\} - n[E(T_n)]^2$$

$$= n \text{Var}(E(T_n | X_i)) = \text{Var}(\tilde{T}_n)$$
For a U-statistic $U_n$, one can show (exercise) that

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

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

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

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

$$\text{Var}(\tilde{U}_n) = \frac{m^2 \zeta_1}{n}$$

and, by Corollary 3.2 and Lemma 3.1,

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

This is enough for establishing the asymptotic distribution of $U_n$.

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

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

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

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

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

$$n[U_n - E(U_n)] \xrightarrow{d} \frac{m(m-1)}{2} \sum_{j=1}^{\infty} \lambda_j (\chi_{1j}^2 - 1),$$

where $\chi_{1j}^2$'s are i.i.d. random variables having the chi-square distribution $\chi_1^2$ and $\lambda_j$'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 $EY^2 = \frac{m^2(m-1)^2}{2} \zeta_2.$