Lecture 34: The projection method

Since $\mathcal{P}$ is nonparametric, the exact distribution of any U-statistic is hard to derive.
We study asymptotic distributions of U-statistics by using the method of *projection*.

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 $\psi_n(X_i) = E(T_n|X_i)$.
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 $E[\psi_n(X_i)] = E[E(T_n|X_i)] = E(T_n)$.
If $E(T_n^2) < \infty$ and $\mathrm{Var}(\psi_n(X_i)) > 0$, then

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

by the CLT.
Let $\check{T}_n$ be the projection of T_n on $X_1, ..., X_n$.
Then

$$T_n - \check{T}_n = T_n - E(T_n) - \sum_{i=1}^{n} [\psi_n(X_i) - E(T_n)]. \tag{2}$$

If we can show that $T_n - \check{T}_n$ has a negligible order of magnitude, then we can derive the asymptotic distribution of T_n by using (1)-(2) and Slutsky's theorem.
The order of magnitude of $T_n - \check{T}_n$ can be obtained with the help of the following lemma.

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

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

From Definition 3.3 with $Y_i = X_i$ and $k_n = n$,

$$\mathrm{Var}(\check{T}_n) = n\mathrm{Var}(E(T_n|X_i)).$$

The result follows from

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

This method of deriving the asymptotic distribution of T_n is known as the method of projection and is particularly effective for U-statistics.
For a U-statistic U_n, one can show (exercise) that

$$\check{U}_n = E(U_n) + \frac{m}{n}\sum_{i=1}^{n} \tilde{h}_1(X_i), \tag{3}$$

where $\check{U}_n$ is the projection of U_n on $X_1, ..., X_n$ and $\tilde{h}_1(x) = h_1(x) - E[h(X_1, ..., X_m)]$, $h_1(x) = E[h(x, X_2, ..., X_m)]$.
Hence

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

and, by Corollary 3.2 and Lemma 3.1,

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

If $\zeta_1 > 0$, then (1) holds with $\psi_n(X_i) = mh_1(X_i)$, which leads to the result in Theorem 3.5(i) stated later.

If $\zeta_1 = 0$, then $\tilde{h}_1 \equiv 0$ and we have to use another projection of U_n.
Suppose that $\zeta_1 = \cdots = \zeta_{k-1} = 0$ and $\zeta_k > 0$ for an integer $k > 1$.
Consider the projection $\check{U}_{kn}$ of U_n on $\binom{n}{k}$ random vectors $\{X_{i_1}, ..., X_{i_k}\}$, $1 \le i_1 < \cdots < i_k \le n$.
We can establish a result similar to that in Lemma 3.1 and show that

$$E(U_n - \check{U}_n)^2 = O(n^{-(k+1)}).$$

Also, see Serfling (1980, §5.3.4).

With these results, we obtain the following theorem.

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 $\sum_{j=1}^{\infty} \lambda_j^2 = \zeta_2$.

We have actually proved Theorem 3.5(i).
A proof for Theorem 3.5(ii) is given in Serfling (1980, §5.5.2).
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.

If $\zeta_1 > 0$, it follows from Theorem 3.5(i) and Corollary 3.2(iii) that

$$\text{amse}_{U_n}(P) = m^2\zeta_1/n = \text{Var}(U_n) + O(n^{-2}).$$

By Proposition 2.4(ii), $\{n[U_n - E(U_n)]^2\}$ is uniformly integrable.

If $\zeta_1 = 0$ but $\zeta_2 > 0$, it follows from Theorem 3.5(ii) that $\text{amse}_{U_n}(P) = EY^2/n^2$, where Y denotes the random variable on the right-hand side of (4).
The following result provides the value of EY^2.

Lemma 3.2. Let Y be the random variable on the right-hand side of (4). Then $EY^2 = \frac{m^2(m-1)^2}{2}\zeta_2$.
Proof. Define

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

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 $\text{Var}(\chi^2_{1j}) = 2$, $EY_k = 0$ for any k and

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

It follows from Corollary 3.2(iii) and Lemma 3.2 that

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

if $\zeta_1 = 0$.
Again, by Proposition 2.4(ii), the sequence $\{n^2[U_n - E(U_n)]^2\}$ is uniformly integrable.

We now apply Theorem 3.5 to the U-statistics in Example 3.11.
For $U_n = \frac{2}{n(n-1)}\sum_{1\le i<j\le n} X_i X_j$, $\zeta_1 = \mu^2\sigma^2$.
Thus, if $\mu \neq 0$, the result in Theorem 3.5(i) holds with $\zeta_1 = \mu^2\sigma^2$.
If $\mu = 0$, then $\zeta_1 = 0$, $\zeta_2 = \sigma^4 > 0$, and Theorem 3.5(ii) applies.
However, it is not convenient to use Theorem 3.5(ii) to find the limiting distribution of U_n.
We may derive this limiting distribution using the following technique, which is further discussed in §3.5.
By the CLT and Theorem 1.10,

$$n\bar{X}^2/\sigma^2 \to_d \chi^2_1$$

when $\mu = 0$, 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 $\mu = 0$, this implies that the right-hand side of (4) is $\sigma^2(\chi_1^2 - 1)$, i.e., $\lambda_1 = \sigma^2$ and $\lambda_j = 0$ when $j > 1$.

For the one-sample Wilcoxon statistic, $\zeta_1 = \mathrm{Var}(F(-X_1)) > 0$ unless F is degenerate.
Similarly, for Gini's mean difference, $\zeta_1 > 0$ unless F is degenerate.
Hence Theorem 3.5(i) applies to these two cases.