Lec 6: U Statistics (cont'd)

Weiping Zhang

from «Asymptotic Statistics»
by A. W. van der Vaart

Review of U-Statistic

A U-statistic of order r with kernel h under sample $X_1, \ldots, X_n$ i.i.d $\sim F$ is

$$U_n = \frac{1}{\binom{n}{r}} \sum_{\beta} h(X_{\beta_1}, \ldots, X_{\beta_r})$$

where h is symmetric in its arguments.

- Variance:

$$Var(U_n) = \frac{1}{\binom{n}{r}} \sum_{c=1}^{r} \binom{r}{c} \binom{n-c}{r-c} \zeta_c$$

where $\zeta_c = Var_F(h_c(X_1, \ldots, X_c))$ with $h_c(x_1, \ldots, x_c) = Eh(x_1, \ldots, x_c, X_{c+1}, \ldots, X_r)$.

Asymptotic Normality of U-Statistic

If $Eh^2 < \infty$, then the Hájek projection of $U_n - \theta$ is

$$\hat{U} = \frac{r}{n}\sum_{i=1}^{n}(h_1(X_i) - \theta),$$

where $\theta = EU_n$. Furthermore,

$$\sqrt{n}(U_n - \theta) \rightsquigarrow N(0, r^2\zeta_1).$$

Hoeffding Decomposition

- The Hájek projection gives a best approximation by a sum of functions of one X_i at a time.
- The approximation can be improved by using sums of functions of two, or more, variables. This leads to the *Hoeffding decomposition*.
- Because a projection onto a sum of orthogonal spaces is the sum of the projections onto the individual spaces, it is convenient to decompose the proposed projection space into a sum of orthogonal spaces.

Hoeffding Decomposition

- Given independent variables $X_1, \dots, X_n$ and a subset $A \subset \{1, \dots, n\}$, let H_A denote the set of all square-integrable random variables of the type

$$g_A(X_i : i \in A),$$

for measurable functions g_A of $|A|$ arguments such that

$$E[g_A(X_i : i \in A)|X_j : j \in B] = 0, \quad \textit{every} \quad B : |B| < |A|.$$

(Define $E(T|\emptyset) = ET$)

- By the independence of $X_1, \dots, X_n$, the condition in the last display is automatically valid for any $B \subset \{1, \dots, n\}$ that does not contain A.
- Consequently, the spaces H_A when A ranges over all subsets of $\{l, ..., n\}$, are pairwise orthogonal.

Hoeffding Decomposition

- The condition reflects the intention to build approximations of increasing complexity by projecting a given variable in turn onto the spaces

$$\Big[1\Big], \quad \Big[\sum_i g_{[i]}(X_i)\Big], \quad \Big[\sum\sum_{i<j} g_{[i,j]}(X_i, X_j)\Big], \quad \ldots,$$

  where $g_{[i]} \in H_{[i]}, g_{[i,j]} \in H_{[i,j]}$ and so forth. Each new space is chosen orthogonal to the preceding spaces.
- Let $P_A T$ denote the projection of T onto H_A. Then, by the orthogonality of the H_A, the projection onto the sum of the first r spaces is the sum $\sum_{|A|<r} P_A T$ of the projections onto the individual spaces.

Hoeffding Decomposition

- The projection onto the sum of the first two spaces is the Hádjek projection.
- The projections of zero, first, and second order can be seen to be

$$
\begin{aligned}
P_{\emptyset} T &= ET \\
P_{[i]} T &= E(T|X_i) - ET \\
P_{[i,j]} T &= E(T|X_i, X_j) - E(T|X_i) - E(T|X_j) + ET
\end{aligned}
$$

The general formula given by the following lemma should not be surprising.

Theorem

Let $X_1, \ldots, X_n$ be independent variables, and let T be an arbitrary random variable with $ET^2 < \infty$. Then the projection of T onto H_A is given by

$$P_A T = \sum_{B \subset A} (-1)^{|A|-|B|} E(T|X_i : i \in B).$$

if $T \perp H_B$ for every subset $B \subset A$ of a given set A, then $E(T|X_i : i \in A) = 0$. Consequently, the sum of the spaces H_B with $B \subset A$ contains all square-integrable functions of $(X_i : i \in A)$.

Proof of Hoeffding Decomposition

Proof Abbreviate $E(T|X_i : i \in A)$ to $E(T|A)$ and $g_A(X_i : i \in A)$ to g_A. By the independence of $X_1, \ldots, X_n$ if follows that $E[E(T|A)|B] = E[T|A \cap B]$ for every subsets A and B of $\{1, \ldots, n\}$. Thus, for $P_A T$ as defined in the theorem and a set C strictly contained in A,

$$E[P_A T|C] = \sum_{B \subset A} (-1)^{|A|-|B|} E[T|B \cap C]$$

$$= \sum_{D \subset C} \sum_{j=0}^{|A|-|D|} (-1)^{|A|-|D|-j} \binom{|A| - |D|}{j} E[T|D]$$

By the binomial formula, the inner sum is zero for every D. Thus the left side is zero. In view of the form of $P_A T$ it was not a loss of generality to assume that $C \subset A$. Hence $P_A T$ is contained in H_A.

Proof of Hoeffding Decomposition

Next we verify the orthogonality relationship. For any measurable function

$$E(T - P_A T)g_A = E(T - E(T|A))g_A \\ - \sum_{B \subset A, B \neq A} (-1)^{|A|-|B|} EE(T|B)E(g_A|B).$$

This is zero for any $g_A \in H_A$. This concludes the proof that $P_A T$ is as given.

Proof of Hoeffding Decomposition

- We prove the second assertion of the theorem by induction on $r = |A|$. If $T \perp H_\emptyset$, then $E(T|\emptyset) = ET = 0$. Thus the assertion is true for $r = 0$.
- Suppose that it is true for $0, \ldots, r-1$, and consider a set A of r elements. If $T \perp H_B$ for every $B \subset A$, then certainly $T \perp H_C$ for every $C \subset B$. Consequently, the induction hypothesis shows that $E(T|B) = 0$ for every $B \subset A$ of $r-1$ or fewer elements. The formula for $P_A T$ now shows that $P_A T = E(T|A)$. By assumption the left side is zero. This concludes the induction argument.
- The final assertion of the theorem follows if the variable $T_A := T - \sum_{B \subset A} P_B T$ is zero for every T that depends on $(X_i : i \in A)$ only. But in this case T_A depends on $(X_i : i \in A)$ only and hence equals $E(T_A|A)$, which is zero, because $T_A \perp H_B$ for every $B \subset A$.

Hoeffding Decomposition

If $T = T(X_1, \ldots ..X_n)$ is permutation-symmetric and $X_1 \ldots ..X_n$ are independent and identically distributed, then the Hoeffding decomposition of T can be simplified to

$$T = \sum_{r=0}^{n} \sum_{|A|=r} g_r(X_i : i \in A)$$

for

$$g_r(x_1, \ldots, x_r) = \sum_{B \subset \{1,\ldots,r\}} (-1)^{r-|B|} \mathrm{E}\, T(x_i \in B, X_i \notin B)$$

The inner sum in the representaion of T is for each r a U-statisic of order r, with degenerate kemel. All terms in the sum are orthogonal, hence the variance of T can be found as $\mathrm{var}(T) = \sum_{r=1}^{n} \binom{n}{r} E g_r^2(X_1, \ldots, X_r)$.

Examples

Example Consider a U-statistic of order 2.

$$U_n = \frac{1}{\binom{n}{2}} \sum_{i<j} h(X_i, X_j)$$

The Hoeffding decomposition is:

$$U_n = U + \frac{2}{n} \sum_{i} h_1(X_i) + \frac{1}{\binom{n}{2}} \sum_{i<j} h_2(X_i, X_j),$$

where

$$U = EU_n = Eh(X_1, X_2), h_1(x) = Eh(x, X_2) - U,$$
$$h_2(x, y) = h(x, y) - h_1(x) - h_1(y) - U.$$

Two-sample U-statistic

- Suppose the observations consist of two independent samples $X_1, \dots X_m$ and $Y_1, \dots, Y_n$, i.i.d. within each sample, from possibly different distributions.
- Let $h(x_1, \dots, x_r, y_1, \dots, y_s)$ be a known function that is permutation symmetric in $x_1, \dots, x_r$ and $y_1, \dots, y_s$ separately.
- A *two-sample U-statistic* with kernel h has the form

$$U_{mn} = \frac{1}{\binom{m}{r}\binom{n}{s}} \sum_{\alpha} \sum_{\beta} h(X_{\alpha_1}, \dots, X_{\alpha_r}, Y_{\beta_1}, \dots, Y_{\beta_s}),$$

where α and β range over the collections of all subsets of r different elements from $\{1, 2, \dots, m\}$ and of s different elements from $\{1, 2, \dots, n\}$, respectively.

- Clearly, U_{mn} is an unbiased estimator of the parameter

$$\theta = Eh(x_1, \ldots, x_r, y_1, \ldots, y_s)$$

- The sequence U_{mn} can be shown to be asymptotically normal by the same arguments as for one-sample U-statistics.
- Here we let both $m \to \infty$ and $n \to \infty$, in such a way that the number of X_i and Y_j are of the same order. Specifically, if $N = m + n$ is the total number of observations we assume that, as $m, n \to \infty$,

$$\frac{m}{N} \to \lambda, \quad \frac{n}{N} \to 1 - \lambda, 0 < \lambda < 1$$

- The projection of $U_{mn} - \theta$ onto the set of all functions of the form $\sum_{i=1}^{m} k_i(X_i) + \sum_{j=1}^{n} l_j(Y_j)$ is given by

$$\hat{U} = \frac{r}{m}\sum_{i=1}^{m} h_{1,0}(X_i) + \frac{s}{n}\sum_{j=1}^{n} h_{0,1}(Y_j),$$

  where the functions $h_{1,0}$ and $h_{0,1}$ are defined by

$$h_{1,0}(x) = Eh(x, X_2, \ldots, X_r, Y_1, \ldots, Y_s) - \theta,$$
$$h_{0,1}(y) = Eh(X_1, X_2, \ldots, X_r, y, \ldots, Y_s) - \theta$$

- This follows, as before, by first applying the Hádjek projection, and next expressing $E(U|X_i)$ and $E(U|Y_j)$ in the kernel function.

- If the kernel is square-integrable, then the sequence $\hat{U}$ is asymptotically normal by the central limit theorem. The difference between $\hat{U}$ and $U - \theta$ is asymptotically negligible.

Theorem

If $Eh^2(X_1, X_2, \ldots, X_r, Y_1, \ldots, Y_s) < \infty$, then the sequence $\sqrt{N}(U_{mn} - \theta - \hat{U})$ converges in probability to zero. Consequently, the sequence $\sqrt{N}(U_{mn} - \theta)$ converges in distribution to the normal law with mean zero and variance $r^2\zeta_{1,0}/\lambda + s^2\zeta_{0,1}/(1-\lambda)$, where, with the X_i being i.i.d variables independent of the i.i.d variables Y_j, and

$$\zeta_{c,d} = cov(h(X_1, X_2, \ldots, X_r, Y_1, \ldots, Y_s)),$$
$$h(X_1, X_2, \ldots, X_c, X'_{c+1}, \ldots, X'_r, Y_1, \ldots, Y_d, Y'_{d+1}, \ldots, Y'_s)$$

Proof.

The argument is similar to the one given previously for one-sample U-statistics. The variances of U_{mn} and its projection are given by

$$Var(\hat{U}) = \frac{r^2}{m}\zeta_{1,0} + \frac{s^2}{n}\zeta_{0,1}$$
$$Var(U_{mn}) = \frac{1}{\binom{m}{r}^2\binom{n}{s}^2}\sum_{c=0}^{r}\sum_{d=0}^{s}\binom{m}{r}\binom{r}{c}\binom{m-r}{r-c}\binom{n}{s}\binom{s}{d}\binom{n-s}{s-d}\zeta_{c,d}$$

It can be checked from this that both the sequence $NVar(\hat{U})$ and the sequence $NVar(U_{mn})$ converge to the number $r^2\zeta_{1,0}/\lambda + s^2\zeta_{0,1}/(1-\lambda)$. $\square$

Example(Mann-Whitney statistic). The kernel for the parameter $\theta = P(X < Y)$ is $h(x,y) = 1(X < Y)$, which is of order 1 in both x and y. The corresponding U-statistic is

$$U = \frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} 1(X_i \leq Y_j)$$

The statistic mnU is known as the Mann-Whitney statistic and is used to test for a difference in location between the two samples. A large value indicates that the Y_j are "stochastically larger" than the X_i.

If the X_i and Y_j have cumulative distribution functions F and G, respectively, then the projection of $U - \theta$ can be written

$$\hat{U} = -\frac{1}{m}\sum_{i=1}^{m}[G(X_i-) - EG(X_i-)] + \frac{1}{n}\sum_{j=1}^{n}[F(Y_j) - EF(Y_j)]$$

In particular, under the null hypothesis that the pooled sample $X_1, \ldots, X_m$, $Y_1, \ldots, Y_n$ is i.i.d. with continuous distribution function $F = G$ the sequence

$$\sqrt{12mn/N}(U - \frac{1}{2}) \rightsquigarrow N(0,1)$$

Degenerate U-Statistics

- When using U-statistics for testing hypotheses, it occasionally happens that at the null hypothesis, the asymptotic distribution has variance zero
- Let $h_c(x_1, \ldots, x_c) = Eh(x_1, \ldots, x_c, X_{c+1}, \ldots, X_r)$, and $\zeta_c = Var(h_c(X_1, \ldots, X_c))$.
- We say that a U-statistic has a degeneracy of order k if $\zeta_1 = \cdots = \zeta_k = 0$ and $\zeta_{k+1} > 0$
- To present the ideas, we restrict attention to kernels with degeneracy of order 1, for which $\zeta_1 = 0$ and $\zeta_2 > 0$.

Example Consider the following U-statistics

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

with $X_1, \ldots, X_n$ *i.i.d* and $EX_i = \mu$, $Var(X_i) = \sigma^2$. Therefore, U_n is an unbiased estimator of μ^2.
Since $h_1(x_1) = E(x_1 X_2) = x_1 \mu$ and

$$\zeta_1 = Var(h_1(X_1)) = \mu^2 \sigma^2$$

we have

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

But suppose that $\mu = 0$ under the null hypothesis. Then the limiting variance is zero, so that this theorem is useless for finding cutoff points for a test of the null hypothesis.

Assuming $\sigma^2 > 0$, we have $\zeta_2 = Var(X_1 X_2) = \sigma^4 > 0$, so that the degeneracy is of order 1. To find the asymptotic distribution of U_n for a sample $X_1, X_2, \ldots$ from a distribution with mean 0 and variance σ^2, we rewrite U_n as follows

$$
\begin{aligned}
U_n &= \frac{1}{\binom{n}{2}} \sum_{i<j} X_i X_j \\
&= \frac{1}{n(n-1)} \sum_{i \neq j} X_i X_j \\
&= \frac{1}{n-1} \Big[\Big(\frac{1}{\sqrt{n}} \sum_i X_i \Big)^2 - \frac{1}{n} \sum_i X_i^2 \Big]
\end{aligned}
$$

From the central limit theorem we have $\frac{1}{\sqrt{n}} \sum_i X_i \rightsquigarrow N(0, \sigma^2)$, and from the law of large numbers we have $\frac{1}{n} \sum_i X_i^2 \to \sigma^2$.

Therefore by Slutsky's Theorem, we have

$$nU_n \rightsquigarrow (Z^2 - 1)\sigma^2$$

where $Z \sim N(0, 1)$.

Example Suppose now that $h(x_1, x_2) = af(x_1)f(x_2) + bg(x_1)g(x_2)$, where $f(x)$ and $g(x)$ are orthonormal functions of mean zero; that is, $Ef^2(X) = Eg^2(X) = 1$, $Ef(X)g(X) = 0$ and $Ef(X) = Eg(X) = 0$. Then $h_1(x_1) = Eh(x_1, X_2) = 0$, so that $\zeta_1 = 0$ and

$$\begin{aligned}\zeta_2 &= a^2 Var(f(X_1)f(X_2)) + 2abCov(f(X_1)f(X_2), g(X_1)g(X_2)) \\ &\quad + b^2 Var(g(X_1)g(X_2)) \\ &= a^2 + b^2\end{aligned}$$

so the degeneracy is of order 1 (assuming $a^2 + b^2 > 0$). To find the asymptotic distribution of U_n, we have

$$
\begin{aligned}
(n-1)U_n &= \frac{1}{n}\sum_{i\neq j}[af(X_i)f(X_j) + bg(X_i)g(X_j)] \\
&= a\Big[\Big(\frac{1}{\sqrt{n}}\sum f(X_i)\Big)^2 - \frac{1}{n}\sum f^2(X_i)\Big] \\
&+ b\Big[\Big(\frac{1}{\sqrt{n}}\sum g(X_i)\Big)^2 - \frac{1}{n}\sum g^2(X_i)\Big] \\
&\rightsquigarrow a(Z_1^2 - 1) + b(Z_2^2 - 1)
\end{aligned}
$$

where Z_1 and Z_2 are independent $N(0,1)$.

The General Case

- The above example is indicative of the general result for kernels with degeneracy of order 1. This is due to a result from the Hilbert-Schmidt theory of integral equations: For given i.i.d. random variables, X_1 and X_2, any symmetric, square integrable function, $A(x_1, x_2)$, $(A(x_1, x_2) = A(x_2, x_1)$ and $EA(X_1, X_2)^2 < \infty)$, admits a series expansion of the form,

$$A(x_1, x_2) = \sum_{k=1}^{\infty} \lambda_k \varphi_k(x_1) \varphi_k(x_2) \tag{1}$$

where the λ_k are real numbers, and the φ_k are an orthonormal sequence,

$$E\varphi_j(X_1)\varphi_k(X_1) = \begin{cases} 1, & if\ j = k, \\ 0, & if j \neq k \end{cases}$$

- The λ_k are the eigenvalues, and the $\varphi_k(x)$ are corresponding eigenfunctions of the transformation, $g(x) \mapsto EA(x, X_1)g(X_1)$. That is, for all k,

$$EA(x, X_2)\varphi_k(X_2) = \lambda_k \varphi_k(x) \tag{2}$$

- Equation (2) is to be understood in the L_2 sense, that

$$\sum_{k=1}^{n} \lambda_k \varphi_k(X_1)\varphi_k(X_2) \to A(X_1, X_2), \textit{in quadratic mean}$$

  Stronger conditions on A are required to obtain convergence a.s.

- In our problem, we take $A(x_1, x_2) = h(x_1, x_2) - \theta$, where $\theta = Eh(X_1, X_2)$. This is a symmetric square integrable kernel, but we are also assuming $\zeta_1 = Var(h_1(X)) = 0$, where $h_1(x) = Eh(x, X_2)$.

- Note $Eh_1(X) = \theta$, but since $Var(h_1(X)) = 0$, we have $h_1(x) \equiv \theta$ a.s. Now replace x in (2) by X_1 and take expectations on both sides. We obtain

$$\begin{aligned} \lambda_k E\varphi_k(X_1) &= E[(h(X_1, X_2) - \theta)\varphi_k(X_2)] \\ &= E[E(h(X_1, X_2) - \theta | X_2)\varphi_k(X_2)] \\ &= E[(h_1(X_2) - \theta)\varphi_k(X_2)] = 0. \end{aligned}$$

- Thus all eigenfunctions corresponding to nonzero eigenvalues have mean zero. Now we can apply the method of above example, to find the asymptotic distribution of $n(U_n - \theta)$.

Theorem

Let U_n be the U-statistic associated with a symmetric kernel of degree 2, degeneracy of order 1, and expectation θ. Then $n(U_n - \theta) \rightsquigarrow \sum_{j=1}^{\infty} \lambda_j(Z_j^2 - 1)$, where $Z_1, Z_2, \ldots$ are independent $N(0,1)$ and $\lambda_1, \lambda_2, \ldots$ are the eigenvalues satisfying (1) with $A(x_1, x_2) = h(x_1, x_2) - \theta$.

- For h having degeneracy of order 1 and arbitrary degree $r \geq 2$, the corresponding result gives the asymptotic distribution of $n(U_n - \theta)$ as

$$\binom{r}{2} \sum_{j=1}^{\infty} \lambda_j (Z_j^2 - 1),$$

  where the λ_j are the eigenvalues for the kernel $h_2(x_1, x_2) - \theta$. (See Serfling (1980) or Lee (1990).)
- For many kernels, there are just a finite number of nonzero eigenvalues.

Comparing U- and V- statistics

- $r = 1$: $U = \frac{1}{n}\sum_{i=1}^{n} h(X_i) = V$
- $r = 2$:

$$U = \frac{1}{\binom{n}{2}} \sum_{i<j} h(X_i, X_j)$$

On the other hand,

$$V = \frac{1}{n^2}\sum_i \sum_j h(X_i, X_j) = \frac{1}{n^2}\sum_i \Big[\sum_{j \neq i} h(X_i, X_j) + h(X_i, X_i)\Big]$$

$$= \frac{1}{n^2}\sum_i \sum_{j \neq i} h(X_i, X_j) + \frac{1}{n^2}\sum_i h(X_i, X_i)$$

$$= \frac{n(n-1)}{n^2} U + \frac{1}{n^2}\sum_i h(X_i, X_i) \to U$$

Moreover, $Eh(X_1, X_2) = \theta$,

$$EV = \frac{n-1}{n} EU + \frac{1}{h} Eh(X_1, X_1)$$
$$= \theta + \frac{1}{n}[Eh(X_1, X_1) - \theta]$$

thus, $bias \to 0$.

Theorem

Let $r = 2$, $\zeta_i = Eh(X_1, \ldots, X_i)$ and suppose $0 < \zeta_1 < \infty$, $\zeta_2 < \infty$, then U- and V-statistics have the same asymptotic distribution,

$$\sqrt{n}(V - \theta) \rightsquigarrow N(0, 4\zeta_1),$$

Proof.
Since $\sqrt{n}(U-\theta) \rightsquigarrow N(0, 4\zeta_1)$, and

$$
\begin{aligned}
\sqrt{n}(V-\theta) &= \sqrt{n}\Big[\frac{n-1}{n}U + \frac{1}{n^2}\sum_i h(X_i, X_i) - \theta\frac{n-1+1}{n}\Big] \\
&= \sqrt{n}\Big[\frac{n-1}{n}(U-\theta) + \frac{1}{n^2}\sum_i h(X_i, X_i) - \theta\frac{1}{n}\Big] \\
&= \frac{n-1}{n}\sqrt{n}(U-\theta) + \frac{1}{\sqrt{n}}\frac{1}{n}\sum_i (h(X_i, X_i) - \theta)\Big] \\
&\rightsquigarrow N(0, 4\zeta_1).
\end{aligned}
$$

$\square$

Conclusion: U- and V-statistics are asymptotically equivalent. The V-statistic is a more intuitive estimator, the U-statistic is more convenient for proofs (and unbiased).