**36-755: Advanced Statistical Theory** **Fall 2016**

# Lecture 25: November 28

*Lecturer: Alessando Rinaldo* *Scribes: Octavio Mesner*

**Note**: *LaTeX template courtesy of UC Berkeley EECS dept.*



## 25.1 U-Statistics

Let $X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} P$ on $(\mathcal{X}, \mathcal{B})$ and let $h : \mathcal{X} \to \mathbb{R}$ (called a kernel) be symmetric in its arguments.

$$\mathbb{E}\left[h(X_1, \ldots, X_m)\right] = \theta(P)$$

for $m$ fixed and $n > m$.
A U-statistic of order $m$ is

$$U_n = \frac{1}{\binom{n}{m}} \sum_{i_1 \leq i_2 \leq \ldots \leq i_m} h(X_{i_1}, X_{i_2}, \ldots, X_{i_m}),$$

a summation over all $m$-subsets of $\{1, \ldots, n\}$. $\mathbb{E}[U_n] = \theta$. The goal with U-statistics is to estimate our parameter without bias and with least variance

**Examples:**

1. Mean $\theta(P) = \mathbb{E}[X]$. $h(x) = x, m = 1, U_n = \frac{1}{n}\sum_i X_1$. Similarly, if $\theta = \mathbb{E}[X^k], h(x) = x^k$.

2. Variance $\theta(P) = V[X] = \frac{1}{2}\mathbb{E}\left[(X_1 - X_2)^2\right]$ for $X_1, X_2 \overset{\text{iid}}{\sim} P$, $h(x_1, x_2) = \frac{1}{2}(x_1 - x_2)^2$ then

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

3. Wilcoxon Sign Rank Test. Assume $P$ has a continuous cdf. Let $\theta = \mathbb{P}(X_1 > 0)$. If $P$ is symmetric then $\theta = \frac{1}{2}$.

$$U_n = \frac{1}{n} \sum_{i=1}^{n} 1\left\{X_1 > 0\right\}$$

Instead we use Wilcoxon test

$$T^+ = \sum_{i=1}^{n} R_i^+ 1\left\{X_i > 0\right\}$$

where $R_1^+, R_2^+, \ldots, R_n^+$ are the ranks of $|X_1|, \ldots, |X_n|$ in increasing order.

$$R_i^+ = \sum_j 1\left\{|X_j| \leq |X_i|\right\}$$

Using some algebra,

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

where $h_1(x_1, x_2) = \binom{n}{2} 1\{x_1 + x_2 > 0\}$ and $h_2(x_1) = n1\{x_1 > 0\}$. $T^+$ is the sum of an order 2 and order 1 U-statistic.

4. Kendall's tau. We observe $n$ i.i.d. pairs $(X_1, Y_1), \ldots, (X_n, Y_n)$ from some continuous $P$ on $\mathbb{R}^2$. Kendall's tau statistic is

$$\tau = \frac{4}{n(n-1)} \left[ \sum_{i<j} 1\{(Y_j - Y_i)(X_j - X_i) > 0\} \right] - 1.$$

It computes the fraction of concordant pairs where $(X_i, Y_i)$ are concordant if $(Y_j - Y_i)(X_j - X_i) > 0$. If $X \perp Y$ then $\mathbb{E}[\tau] = 0$ and if $\tau = \pm 1$ then there is some monotonic function $f$ such that $Y = f(X)$. This is a U-statistic of order 2 with kernel

$$h\left(\binom{x_1}{y_1}, \binom{x_2}{y_2}\right) = 2 \times 1\{(y_2 - y_1)(x_2 - x_1) > 0\} - 1 = 4 \times 1\{x_1 < x_2, y_1 < y_2\} - 1$$

**Naïve Approach** sample size: $n$, order of U-statistic: $m$ (fixed). Split the sample $(X_1, \ldots, X_n)$ into $\lfloor \frac{n}{m} \rfloor$ non-overlapping blocks of size $m$, evaluate $h$ on each block and then average to obtain an estimator with variance $= \frac{m}{n} V[h(X_1, \ldots, X_n)]$.

### 25.1.1 Variance of $U_n$

Assume that $V[h(X_1, \ldots, X_m)] \leq \infty$, For $c = 0, \ldots, m$, let

$$h_c(x_1, \ldots, x_c) = \mathbb{E}\left[h(x_1, \ldots, x_c, X_{c+1}, \ldots, X_m)\right], \text{ where } X_{c+1}, \ldots, X_m \overset{\text{iid}}{\sim} P$$

Then

$$h_c(x_1, \ldots, x_c) = \mathbb{E}\left[h(x_1, \ldots, X_m) | X_1 = x_1, \ldots, X_c = x_c\right]$$

Because of independence

- Set $h_0 = \theta$ and $h_n$
- Notice that

$$\mathbb{E}\left[h_c(X_1, \ldots, X_c)\right] = \mathbb{E}\left[\mathbb{E}\left[h(X_1, \ldots, X_m) | X_1, \ldots, X_c\right]\right] = \mathbb{E}\left[h(X_1, \ldots, X_m)\right] = \theta$$

- Set $\zeta_c = V[h_c(X_1, \ldots, X_c)]$ and $\zeta_0 = 0$.

**Lemma 25.1** *For $(i_1, \ldots, i_m)$ and $(j_1, \ldots, j_m)$, $m$-subsets of $\{1, \ldots, n\}$, we have*

$$Cov[h(X_{i_1}, \ldots, X_{i_m}), h(X_{j_1}, \ldots, X_{j_m})] = \zeta_c$$

*where $c = |\{i_1, \ldots, i_m\} \cap \{j_1, \ldots, j_m\}|$*

**Proof:** Without loss of generality, assume $c > 0$ and the first $c$ terms are common among the $(i_1 < \ldots < i_m)$ and $(j_1 < \ldots < j_m)$. Let $X_1, \ldots X_m, X'_{c+1}, \ldots, X'_m \overset{\text{iid}}{\sim} P$. Calculating the covariance, we have

$$\begin{aligned}
&\text{Cov}\left[h\left(X_1, \ldots, X_c, X_{c+1}, \ldots, X_m\right), h\left(X_1, \ldots, X_c, X'_{c+1}, \ldots, X'_m\right)\right] \\
&= \mathbb{E}\left[\mathbb{E}(h(X_1, \ldots, X_m) - \theta)\, h(X_1, \ldots, X_c, X'_{c+1}, \ldots, X'_m) | X_1, \ldots, X_c\right] \\
&= \mathbb{E}\left[(h_c(X_1, \ldots, X_c) - \theta)^2\right] \\
&= V\left[(h_c(X_1, \ldots, X_c) - \theta)^2\right] \\
&= V\left[h_c(X_1, \ldots, X_c)\right] = \zeta_c
\end{aligned}$$

■

The same argument shows that

$$\text{Cov}\left[h_c(X_1, \ldots, X_c), h(X_1, \ldots, X_m)\right] = \zeta_c.$$

Now, using Cauchy-Schwartz inequality, $\zeta_c \leq \sqrt{\zeta_c}\sqrt{\zeta_m}$ so that $\zeta_c \leq \zeta_m \forall c$. We can also show that $0 = \zeta_0 \leq \zeta_1 \leq \zeta_2 \leq \zeta_m$. In fact, $0 \leq \frac{\zeta_c}{c} \leq \frac{\zeta_d}{d}$ for $1 \leq c \leq d \leq m$ Hoeffding (1948).

**Theorem 25.2**

$$V[U_n] = \binom{n}{m}^{-1} \sum_{c=1}^{m} \binom{m}{c}\binom{n-m}{m-c}\zeta_c.$$

*In particular, as $n \to \infty$, (m fixed)*

$$V[U_n] = \frac{m^2}{n}\zeta_1 + o(n^{-2}) \text{ and } V[U_n] \downarrow \frac{m^2}{n}\zeta_1.$$

*For finite samples,*

$$\frac{m^2}{n}\zeta_1 \leq V[U_n] \leq \frac{m^2}{n}\zeta_m = \frac{m}{n}V[h(X_1, \ldots, X_m)].$$

**Remarks**

1. We assume that $\zeta_1 > 0$. It may be the case that $\zeta_1 = 0$, in which case, $U_n$ is degenerate.
2. If $m$ is allowed to grow with $n$, there are few results in the literature.

**Proof:** Start with

$$\begin{aligned}
V[U_n] &= V\left[\binom{n}{m}^{-1} \sum_{i_1 < \ldots < i_m} h(X_{i_1}, \ldots, X_{i_m})\right] \\
&= \binom{n}{m}^{-2} \sum_{i_1 < \ldots < i_m} \sum_{j_1 < \ldots < j_m} \text{Cov}\left[h(X_{i_1}, \ldots, X_{i_m}), h(X_{j_1}, \ldots, X_{j_m})\right]
\end{aligned}$$

If $|\{i_1, \ldots, i_m\} \cap \{j_1, \ldots, j_m\}| = 0$ then the covariance is 0. Otherwise, if $|\{i_1, \ldots, i_m\} \cap \{j_1, \ldots, j_m\}| = c$ then the covariance is $\zeta_c$.
There are $\binom{n}{m}\binom{m}{c}\binom{n-m}{m-c}$ number of pairs $\{i_1 < \ldots < i_m\}$ and $\{j_1 < \ldots < j_m\}$ with $c$ common elements,

$c = 1, \ldots, m$.
So,

$$\begin{aligned} V[U_n] &= \binom{n}{m}^{-2} \sum_{c=0}^{m} \binom{n}{m} \binom{m}{c} \binom{n-m}{m-c} \zeta_c \\ &= \binom{n}{m}^{-1} \sum_{c=0}^{m} \binom{m}{c} \binom{n-m}{m-c} \zeta_c \\ &= \sum_{c=1}^{m} \frac{(m!)^2}{c!(m-c)!} \frac{(n-m)(n-m-1)\cdots(n-2m+c+1)}{n(n-1)\cdots(n-m+1)} \zeta_c \end{aligned}$$

as $n \to \infty$ the terms in the sum with $c = 1$ is of order $o(n^{-c})$. ∎

Next time, we will show that

$$\sqrt{n}(U_n - \theta) \xrightarrow{D} \mathcal{N}(0, m^2 \zeta_1)$$

# References

[H84] W. Hoeffding, "A class of statistics with asymptotically normal distribution," *The annals of mathematical statistics*, 1948, pp. 293–325.