THE UNIVERSITY OF TEXAS AT AUSTIN
Department of Statistics and Data Sciences
College of Natural Sciences

# SDS 384 11: Theoretical Statistics

## Lecture 8: U Statistics

Purnamrita Sarkar
Department of Statistics and Data Science
The University of Texas at Austin

# U Statistics

- We will see many interesting examples of U statistics.
- Interesting properties
    - Unbiased
    - Reduces variance
    - Concentration (via McDiarmid)
    - Asymptotic variance
    - Asymptotic distribution

# An estimable parameter

- Let $\mathcal{P}$ be a family of probability measures on some arbitrary measurable space.
- We will now define a notion of an an estimable parameter. (coined "regular parameters" by Hoeffding.)
- An estimable parameter $\theta(P)$ satisfies the following.

**Theorem (Halmos)**

*$\theta$ admits an unbiased estimator iff for some integer $m$ there exists an unbiased estimator of $\theta(P)$ based on $X_1, \ldots, X_m \overset{iid}{\sim} P$ that is, if there exists a real-valued measurable function $h(X_1, ..., X_m)$ such that*

$$\theta = Eh(X_1, \ldots, X_m).$$

*The smallest integer $m$ for which the above is true is called the degree of $\theta(P)$.*

# U statistics

- The function $h$ may be taken to be a symmetric function of its arguments.
- This is because if $f(X_1, \ldots, X_m)$ is an unbiased estimator of $\theta(P)$, so is
$$h(X_1, \ldots, X_m) := \frac{\sum_{\pi \in \Pi_m} f(X_{\pi_1}, \ldots, X_{\pi_m})}{m!}$$
- For simplicity, we will assume $h$ is symmetric for our notes.

# U Statistics (Due to Wassily Hoeffding in 1948)

**Definition**

Let $X_i \overset{iid}{\sim} f$, let $h(x_1, \ldots, x_r)$ be a symmetric kernel function and $\Theta(F) = E[h(x_1, \ldots, x_r)]$. A U-statistic $U_n$ of order $r$ is defined as

$$U_n = \frac{\sum_{\{i_1,\ldots,i_r\} \in \mathcal{I}_r} h(X_{i_1}, X_{i_2}, \ldots, X_{i_r})}{\binom{n}{r}},$$

where $\mathcal{I}_r$ is the set of subsets of size $r$ from $[n]$.

# Sample variance as an U-Statistic

**Example**

The sample variance is an U-statistic of order 2.

**Proof.**

Let $\theta(F) = \sigma^2$.

$$\begin{aligned}\sum_{i \neq j}^{n} (X_i - X_j)^2 &= 2n \sum_i X_i^2 - 2 \sum_{i,j} X_i X_j \\ &= 2n \sum_i X_i^2 - 2n^2 \bar{X}^2 \\ &= 2n(n-1) \frac{\sum_i X_i^2 - n\bar{X}^2}{n-1}\end{aligned}$$

$$U_n := \frac{\sum_{i<j}^{n} (X_i - X_j)^2 / 2}{n(n-1)/2} = 2s_n^2$$

$\square$

# Sample variance as U-statistic

- Is its expectation the variance?
- $\frac{1}{2}E[(X_1 - X_2)^2] = \frac{1}{2}E(X_1 - \mu - (X_2 - \mu))^2 = \sigma^2$

# U-statistics examples: Wilcoxon one sample rank statistic

**Example**

$U_n = \sum_i R_i 1(X_i > 0)$, where $R_i$ is the rank of $X_i$ in the sorted order $|X_1| \leq |X_2| \ldots$.

- This is used to check if the distribution of $X_i$ is symmetric around zero.
- Assume $X_i$ to be distinct.
- $R_i = \sum_{j=1}^{n} 1(|X_j| \leq |X_i|)$

# U-statistics examples: Wilcoxon one sample rank statistic

**Example**

$T_n = \sum_i R_i 1(X_i > 0)$, where $R_i$ is the rank of $X_i$ in the sorted order $|X_1| \le |X_2| \ldots$.

$$
\begin{aligned}
T_n = \sum_i R_i 1(X_i > 0) &= \sum_{i=1}^n \sum_{j=1}^n 1(|X_j| \le |X_i|) 1(X_i > 0) \\
&= \sum_{i=1}^n \sum_{j=1}^n 1(|X_j| \le X_i) 1(X_i \neq 0) = \sum_{i \neq j}^n 1(|X_j| < X_i) + \sum_{i=1}^n 1(X_i > 0) \\
&= \sum_{i<j} 1(|X_j| < X_i) + \sum_{i<j} 1(|X_i| < X_j) + \sum_{i=1}^n 1(X_i > 0) \\
&= \sum_{i<j} 1(X_i + X_j > 0) + \sum_{i=1}^n 1(X_i > 0) = \binom{n}{2} U_2 + nU_1
\end{aligned}
$$

- Asymptotically dominated by the first term, which is an U statistic.
- Why isn't it a U statistic?

# Kendal's Tau

**Example**

Let $P_1 = (X_1, Y_1)$ and $P_2 = (X_2, Y_2)$ be two points. $P_1$ and $P_2$ are called concordant if the line joining them (call this $P_1P_2$) has a positive slope and discordant if it has a negative slope. Kendal's tau is defined as:

$$\tau := P(P_1P_2 \text{ has +ve slope}) - P(P_1P_2 \text{ has -ve slope})$$

- This is very much like a correlation coefficient, i.e. lies between $-1, 1$
- Its zero when $X, Y$ are independent, and $\pm 1$ when $Y = f(X)$ is a monotonically increasing (or decreasing) function.

## Kendal's Tau

- Define $h(P_1, P_2) = \begin{cases} 1 & \text{If } P_1, P_2 \text{ is concordant} \\ -1 & \text{If } P_1, P_2 \text{ is discordant} \end{cases}$

- Now define $h(P_1, P_2) = sgn(X_1 - X_2)(Y_1 - Y_2)$

- So $U = \frac{\sum_{i<j} h(P_i, P_j)}{\binom{n}{2}}$ is an U statistic which computes Kendals Tau, and it has order 2.

# More novel examples

**Example (Gini's mean difference/ mean absolute deviation)**

Let $\theta(F) := E[|X_1 - X_2|]$; the corresponding U statistic is
$U_n = \frac{\sum_{i<j} |x_i - x_j|}{\binom{n}{2}}$.

**Example (Quantile Statistic)**

Let $\theta(F) := P(X_1 \leq t) = E[1(X_1 \leq t)]$; the corresponding U statistic is
$U_n = \frac{\sum_i 1(X_i \leq t)}{n}$.

# Properties of the U-statistic

- The U is for unbiased.
- Note that $E[U] = Eh(X_1, \ldots, X_r)$
- $\text{var}(U(X_1, \ldots, X_n)) \leq \text{var}(h(X_1, \ldots, X_r))$ (Rao Blackwell theorem)
    - Just $h(X_1, \ldots, X_r)$ is an unbiased estimator of $\theta(F)$.
    - But averaging over many subsets reduces variance.

# Properties of U-statistics

- Let $X_{(1)} \dots, X_{(n)}$ denote the order statistics of the data.
- The empirical distribution puts $1/n$ mass on each data point.
- So we can think about the U statistic as

$$U_n = E[h(X_1, \dots, X_r)|X_{(1)}, \dots, X_{(n)}]$$

- We also have:

$$\begin{aligned} E[(U-\theta)^2] &= E\left[\left(E[h(X_1,\dots,X_r)-\theta|X_{(1)},\dots,X_{(n)}]\right)^2\right] \\ &\le E[E[(h(X_1,\dots,X_r)-\theta)^2|X_{(1)},\dots,X_{(n)}]] \\ &= \mathrm{var}(h(X_1,\dots,X_r)) \end{aligned}$$

- Rao-Blackwell theorem says that the conditional expectation of any estimator given the sufficient statistic has smaller variance than the estimator itself.
- For $X_1, \dots, X_n \overset{iid}{\sim} P$, the order statistics are sufficient. (why?)

# Concentration

- Consider a U statistic of order 2 $U = \frac{\sum_{i<j} h(X_i, X_j)}{\binom{n}{2}}$.
- How does $U$ concentrate around its expectation?
- Recall McDiarmid's inequality?

**Theorem**

*Let $f : \mathcal{X}^n \to \mathbb{R}$ satisfy the following bounded difference condition $\forall x_1, \ldots, x_n, x_i' \in \mathcal{X}$:*

$$|f(x_1, \ldots, x_{i-1}, x_i, x_{i+1}, \ldots, x_n) - f(x_1, \ldots, x_{i-1}, x_i', x_{i+1}, \ldots, x_n)| \leq B_i,$$

*then,* $P(|f(X) - E[f(X)]| \geq t) \leq 2\exp\left(-\frac{2t^2}{\sum_i B_i^2}\right)$

# Concentration

Consider a U statistic of order 2. $U = \frac{\sum_{i<j} h(X_i, X_j)}{\binom{n}{2}}$.

**Theorem**

*If* $|h(X_1, X_2)| \leq B$ *a.s., then,*

$$P(|U - E[U]| \geq t) \leq 2\exp\left(-\frac{nt^2}{8B^2}\right).$$

# Concentration

**Proof.**

- Consider two samples $X, X'$ which differ in the $i^{th}$ coordinate.
- We have:
$$\begin{aligned} |U(X) - U(X')| &\leq \frac{\sum_{j \neq i} |h(X_i, X_j) - h(X_i, X_j')|}{\binom{n}{2}} \\ &\leq \frac{4B}{n} \end{aligned}.$$
- Now we have:

$$P(|U - E[U]| \geq t) \leq 2 \exp\left(-\frac{nt^2}{8B^2}\right).$$

□

# Concentration

Now consider a U statistic of order $r$. $U = \frac{\sum_{i \in \mathcal{I}_r} h(X_{i_1}, \ldots, X_{i_r})}{\binom{n}{r}}$.

**Theorem**

*If* $|h(X_{i_1}, \ldots, X_{i_r})| \leq B$ *a.s., then,*

$$P(|U - E[U]| \geq t) \leq 2 \exp\left(-\frac{nt^2}{2r^2B^2}\right).$$

# Concentration

**Proof.**

- Consider two samples $X, X'$ which differ in the first coordinate.
- Let $\mathcal{I}_{r-1}$ is the set of $r-1$ subsets from $2, \ldots, n$.
- We have:

$$|U(X) - U(X')| \leq \frac{\sum_{j \in \mathcal{I}_{r-1}} |h(X_1, X_{j_1}, \ldots, X_{j_r}) - h(X_1, X'_{j_1}, \ldots, X'_{j_r})|}{\binom{n}{r}} \\ \leq \frac{2B\binom{n-1}{r-1}}{\binom{n}{r}} = \frac{2rB}{n}.$$

- Now we have:

$$P(|U - E[U]| \geq t) \leq 2 \exp\left(-\frac{nt^2}{2r^2B^2}\right).$$

□

# Hoeffding's bound from his 1963 paper

Now consider a U statistic of order $r$. $U = \dfrac{\sum_{i \in \mathcal{I}_r} h(X_{i_1}, \ldots, X_{i_r})}{\binom{n}{r}}$.

**Theorem**

*If* $|h(X_{i_1}, \ldots, X_{i_r})| \leq B$ *a.s., then,*

$$P(|U - E[U]| \geq t) \leq 2 \exp\left(-\frac{\lfloor n/r \rfloor t^2}{2B^2}\right).$$

- What are we missing?

# Lets start with Markov

- First note that if I can write $U - E[U] = \sum_i p_i T_i$ where $\sum_i p_i = 1$,
- Then,

$$\begin{aligned} P(U - E[U] \geq t) &\leq E[\exp(\lambda \sum_i p_i (T_i - t))] \\ &\leq \sum_i p_i E[\exp(\lambda (T_i - t))] \end{aligned}$$

- So, if $T_i$ is a sum of independent random variables, we can plug in previous bounds into the above.
- But how can we write the U statistics as a sum of such $T_i$'s?

# Lets do a bit of combinatorics

- For simplicity assume that $n = kr$.
- Write $V(X_1, \ldots, X_n) = \dfrac{h(X_1, \ldots, X_r) + \cdots + h(X_{(k-1)r+1}, \ldots, X_{kr}))}{k}$
- Note that $U = \dfrac{\sum_{\pi \in \Pi} V(X_{\pi_1}, \ldots, X_{\pi_n})}{n!}$
- So set $T_\pi = V(X_{\pi_1}, \ldots, X_{\pi_n}) - E[.]$.
- Since $V$ is an average of $k = n/r$ **independent** random variables, using Hoeffding's inequality we have

$$E[\exp(\lambda(T_i - t))] \leq \exp(-\lambda t + \lambda^2 B^2/2k) \leq \exp(-kt^2/2B^2)$$

- Since each $V_\pi$ behave stochastically equivalently, we can take the $\lambda$ the same everywhere.

# Variance of U statistic

Next time!