# SDS 384 11: Theoretical Statistics

## Lecture 10: U Statistics cont.

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 (done)
  - Reduces variance (done)
  - Concentration (via McDiarmid) (done)
  - Asymptotic variance
  - Asymptotic distribution

## Normal Convergence of U statistics-proof

- Trick: find some $\hat{U}$ such that $\hat{U}$ is asymptotically equivalent to $U$.

- Make sure $\hat{U}$ is easy to analyze.

**Theorem**

If $X_n \xrightarrow{d} X$ and $|Y_n - X_n| \xrightarrow{P} 0$, then $Y_n \xrightarrow{d} X$.

- In our case we will use $\hat{U}$ as a sum of functions of $X_i$

- Then use CLT on $\hat{U}$

- We will find the functions using Hájek projections.

## Hájek Projections – Setup

- Let $\{X_1, \ldots, X_n\}$ be independent random vectors.
- Consider a linear space $\mathcal{S}$ of random variables.
  - E.g. $\mathcal{S}$ can be the set of all random variables of the form

  $$\sum_{i=1}^{n} g_i(X_i)$$

  - $g_i$ are arbitrary measurable functions $g_i : \mathbb{R}^d \to \mathbb{R}$ with $E[g_i(X_i)^2] < \infty$, for $i \in [n]$
- $ES^2 < \infty, \forall S \in \mathcal{S}$
- Consider a random variable $T$ with $E[T^2] < \infty$

## Hájek projections

- Define by the projection $\hat{S} = \arg \inf_{S \in \mathcal{S}} E[(T - S)^2]$

**Theorem**

$\hat{S}$ is a projection of $T$ onto a linear space $\mathcal{S}$ with finite second moments, iff, $\hat{S} \in \mathcal{S}$ and

$$E[(T - \hat{S})S] = 0, \qquad \text{For every } S \in \mathcal{S}. \qquad \text{Orthogonality}$$

Every two projections of $T$ onto $\mathcal{S}$ are equal a.s. If $\mathcal{S}$ contains the constant variables, then $E[T] = E[\hat{S}]$ and $\text{cov}(T - \hat{S}, S) = 0$ for every $S \in \hat{S}$.

## Hájek projections

**Proof.**

- First note that

$$E(T - S)^2 = E[(T - \hat{S})^2] + 2E[(T - \hat{S})(\hat{S} - S)] + E[(S - \hat{S})^2]$$

- If the orthogonality condition is satisfied, then the middle term is zero.

- So $E(T - S)^2 \geq E(T - \hat{S})^2$, and this inequality is strict unless $E(\hat{S} - S)^2 = 0$. This proves uniqueness.

$\square$

## Hájek projections-converse

**Proof.**

- For any number $\alpha$

$$E(T - \hat{S} - \alpha S)^2 = E[(T - \hat{S})^2] - 2\alpha E[(T - \hat{S})S] + \alpha^2 E[S^2]$$

- If $\hat{S}$ is the projection, then $\forall \alpha$ and $\forall S \in \mathcal{S}$,

$$\alpha^2 E[S^2] - 2\alpha E[(T - \hat{S})S] \geq 0$$

- So for $\alpha > 0$, $E[(T - \hat{S})S] \leq \alpha E[S^2]/2$
- for $\alpha < 0$, $E[(T - \hat{S})S] \geq -|\alpha| E[S^2]/2$
- So the orthogonality condition must hold.

$\square$

- If constants are in $\mathcal{S}$, then the orthogonality condition with $S = 1$ gives $E[T] = E[\hat{S}]$.
- So, $\mathrm{cov}(T - \hat{S}, S) = E[(T - \hat{S})S] - E[T - \hat{S}]E[S] = 0$
- The first term is zero using orthogonality.
- The second term is zero because $E[T] = E[\hat{S}]$.
- Hájek projections do not always exist, i.e. the $\inf\limits_{S \in \mathcal{S}}$ may not be achievable.
- However it is typically easy to establish existence directly

## Projections and asymptotic equivalence

- By the orthogonality, we have $E[T^2] = E[(T - \hat{S})^2] + E[\hat{S}^2]$
- If $\mathcal{S}$ contains constants, then $E[T] = E[\hat{S}]$
- So $\text{var}(T) = \text{var}(T - \hat{S}) + \text{var}(\hat{S})$
- So if $\mathcal{S}$ has constants, and $\text{var}(T) = \text{var}(\hat{S})$, then $\hat{S} = T$ a.s.
- What if the variances are not equal, but almost (or asymptotically) equal?

8

## Projections and asymptotic equivalence

### Theorem

*Consider linear spaces of random variables with finite second moment $S_n$ that contains constants. Let $T_n$ be random variables with projections $\hat{S}_n$ onto $S_n$. If $var(T_n)/var(S_n) \to 1$, then,*

$$\frac{T_n - E[T_n]}{sd(T_n)} - \frac{\hat{S}_n - E[\hat{S}_n]}{sd(\hat{S}_n)} \xrightarrow{P} 0,$$

*where $sd(X)$ is $\sqrt{var(X)}$.*

## Projections and asymptotic equivalence-proof

**Proof.**

- We will prove convergence in second mean.
- Let $D_n = \dfrac{T_n - E[T_n]}{\text{sd}(T_n)} - \dfrac{\hat{S}_n - E[\hat{S}_n]}{\text{sd}(\hat{S}_n)}$
- $E[D_n] = 0$
- So the variance calculation gives:

$$\text{var}(D_n) = 2 - 2\frac{\text{cov}(T_n, \hat{S}_n)}{\text{sd}(T_n)\text{sd}(\hat{S}_n)}$$

$$= 2 - 2\frac{\text{cov}(T_n - \hat{S}_n, \hat{S}_n) + \text{var}(\hat{S}_n)}{\text{sd}(T_n)\text{sd}(\hat{S}_n)}$$

$$= 2 - 2\frac{\text{var}(\hat{S}_n)}{\text{sd}(T_n)\text{sd}(\hat{S}_n)} \to 0$$

□

## How to get a Hájek projection

- Let $\{X_1, \ldots, X_n\}$ be independent random vectors.
- Consider a linear space $\mathcal{S}$ of random variables.

    - E.g. $\mathcal{S}$ can be the set of all random variables of the form $\sum_{i=1}^{n} g_i(X_i)$.

    - $g_i$ are arbitrary measurable functions $g_i : \mathbb{R}^d \to \mathbb{R}$ with $E[g_i(X_i)^2] < \infty$, for $i \in [n]$

**Theorem**

*The Hájek projection of an arbitrary random variable $T(X_1, \ldots, X_n)$ with finite second moment onto $\mathcal{S}$ is given by*

$$\hat{S} = \sum_{i=1}^{n} E[T|X_i] - (n-1)E[T].$$

## How to get a Hájek projection

**Proof.**

- First note that $\hat{S} \in \mathcal{S}$
- All that remains is to check the orthogonality condition.
- 

$$E[(T - \hat{S})S] = E[(T - \hat{S}) \sum_i g_i(X_i)]$$
$$= \sum_i E[(T - \hat{S})g_i(X_i)]$$
$$= \sum_i E_{X_i} E[(T - \hat{S})g_i(X_i)|X_i]$$
$$= \sum_i E g_i(X_i) E[T - \hat{S}|X_i]$$

- But $E[\hat{S}|X_i] = E[\sum_j E[T|X_j]|X_i] - (n-1)E[T] = E[T|X_i]$. $\qquad \square$

## What if $X_i$'s are iid?

- If $X_1, \ldots, X_n$ are iid,

- So in this case,
$$E[T|X_i = x] = E[T(X_1, \ldots, X_{i-1}, x, X_i, \ldots)]$$
$$= E[T(x, X_2, \ldots, X_n)]$$

- Thus the Hájek projections can be computed by taking a projection on a smaller set $\mathcal{S}' \subset \mathcal{S}$

- $\mathcal{S}'$ contains random variables of the form $\sum_{i=1}^{n} g(X_i)$ where $g$ is some arbitrary measurable function with $E[g(X_i)^2] < \infty$

## Normal Convergence of U statistics-proof

- Recall $U := \dfrac{1}{\binom{n}{r}} \displaystyle\sum_{S \in \mathcal{I}_r} h(X_S)$

- Define the Hájek projection as

$$
\begin{aligned}
\hat{U} &:= \sum_{i=1}^{n} E[U - \theta | X_i] \\
&= \frac{1}{\binom{n}{r}} \sum_{i=1}^{n} \sum_{S \in \mathcal{I}_r} E[h(X_S) - \theta | X_i]
\end{aligned}
$$

- Note that

$$
E[h(X_S) - \theta | X_i = x] = \begin{cases} E[h(x, X_2, \ldots, X_r)] - \theta =: g(x) & \text{When } i \in S \\ 0 & \text{o.w.} \end{cases}
$$

## Normal Convergence of U statistics-proof

- Define the Hájek projection as

$$\hat{U} := \sum_{i=1}^{n} E[U - \theta | X_i]$$

$$= \frac{1}{\binom{n}{r}} \sum_{i=1}^{n} \sum_{S \in \mathcal{I}_r} E[h(X_S) - \theta | X_i]$$

$$= \frac{1}{\binom{n}{r}} \sum_{i=1}^{n} \sum_{S \in \mathcal{I}_r : X_i \in S} E[h(X_S) - \theta | X_i]$$

$$= \frac{1}{\binom{n}{r}} \sum_{i=1}^{n} \binom{n-1}{r-1} g(X_i)$$

$$= \frac{r}{n} \sum_{i=1}^{n} g(X_i)$$

## Normal Convergence of U statistics-proof

- Ok. So we got a projection. Now we need to move to asymptotics

- So let us calculate the variance of $\hat{U}$

$$
\begin{aligned}
\operatorname{var}(\hat{U}) &= \frac{r^2}{n}\operatorname{var}(g(X_1)) \\
&= \frac{r^2}{n}\operatorname{var}(E[h(X_S)|X_1]) = \frac{r^2}{n}\xi_1
\end{aligned}
$$

- Now CLT gives, $\sqrt{n}(\hat{U} - \theta) \xrightarrow{d} N(0, r^2\xi_1)$

## Normal Convergence of U statistics-proof

- $\sqrt{n}(\hat{U} - \theta) \xrightarrow{d} N(0, r^2 \xi_1)$

- We already proves $\dfrac{\text{var}(U)}{\text{var}(\hat{U})} \to 1$

- So $\sqrt{n}(\hat{U} - (U - \theta)) \xrightarrow{P} 0$

- So $\sqrt{n}(U - \theta) \xrightarrow{d} N(0, r^2 \xi_1)$