Stats 300b: Theory of Statistics **Winter 2017**

Lecture 6– January 26

Lecturer: John Duchi *Scribe: Yan Chen, Xiaoyan Han*

Warning: *these notes may contain factual errors*

Reading: VDV Chapter 12

Outline:

- Efficiency of Estimators
  - Super Efficiency
- U-Statistics (VDV Chapter 12)
  - Definitions
  - Examples
  - Variance

1 Efficiency of Estimators

Recall: If $\hat{\theta}_n$ and T_n are estimators/statistics such that $\sqrt{n}\left(\hat{\theta}_n - \theta\right) \xrightarrow{d} N\left(0, \sigma^2\right)$ for some $m(n) \to \infty$, and $\sqrt{n}\left(T_{m(n)} - \theta\right) \xrightarrow{d} N\left(0, \sigma^2\right)$, then the ARE (Asymptotic Relative Efficiency) of $\hat{\theta}_n$ w.r.t. T_n is $\liminf_{n\to\infty} \frac{m(n)}{n}$.

Observation 1. *(Pitman Efficiency): Say that* $\sqrt{n}\left(\hat{\theta}_n - \theta\right) \xrightarrow{d} N\left(0, \sigma^2(\theta)\right)$ *and* $\sqrt{n}\left(T_n - \theta\right) \xrightarrow{d} N\left(0, \tau^2(\theta)\right)$*. Then the ARE of* $\hat{\theta}_n$ *w.r.t.* T_n *is* $\frac{\tau^2(\theta)}{\sigma^2(\theta)}$ *in the one-dimensional case. (In higher dimensions, it is roughly* $tr\left(\tau^2(\theta)\left(\sigma^2(\theta)\right)^{-1}\right)$*).*

Proof Let $m(n) = \left\lceil \frac{\tau^2}{\sigma^2} n \right\rceil$. Then,

$$\sqrt{n}\left(T_{m(n)} - \theta\right) = \underbrace{\sqrt{\frac{n}{m(n)}}}_{\to \frac{\sigma}{\tau}} \underbrace{\sqrt{m(n)}\left(T_{m(n)} - \theta\right)}_{\xrightarrow{d} N(0,\tau^2)} \underbrace{\xrightarrow{d}}_{\text{Slutsky's}} N\left(0, \sigma^2(\theta)\right)$$

Roughly, if $\tau^2 > \sigma^2$, we prefer $\hat{\theta}_n$ to T_n because T_n requires $\frac{\tau^2}{\sigma^2}$ times the sample size that $\hat{\theta}_n$ does for a similar quality. □

1.1 Super-Efficiency

Definition 1.1 (Comparison of Estimators and Super-Efficiency)**.** *Recall if $I(\theta)$ is Fisher-Information for the model $\{P_\theta\}_{\theta\in\Theta}$ and if $\sqrt{n}\left(\hat{\theta}_n-\theta\right)\underset{P_\theta}{\overset{d}{\to}}N\left(0,I(\theta)^{-1}\right)$, then $\hat{\theta}_n$ is efficient. If $\sqrt{n}\left(\hat{\theta}_n-\theta\right)\underset{P_\theta}{\overset{d}{\to}}N\left(0,\sigma^2(\theta)\right)$, where $\sigma^2(\theta)\le I(\theta)^{-1}$ and there is some θ_0 such that $\sigma^{-1}(\theta_0)<I(\theta)^{-1}$, then we call $\hat{\theta}_n$ super-efficient.*

Example: (Hodges Estimator for Gaussian Mean)

For $X_i\overset{i.i.d}{\sim}N(\theta,1)$ and $\hat{\theta}_n=\bar{X}_n=\frac{1}{n}\sum_{i=1}^n X_i$, let $T:=\begin{cases}\bar{X}_n & \text{if } \left|\bar{X}_n\right|\ge n^{-1/4}\\ 0 & \text{if } \left|\bar{X}_n\right|<n^{-1/4}\end{cases}$.

If $\theta=0$, then

$$P_\theta\left(\sqrt{n}T_n=0\right)=P_\theta\left(\left|\bar{X}_n\right|\le n^{-1/4}\right)=P_\theta\left(\left|\underbrace{\sqrt{n}\bar{X}_n}_{N(0,1)}\right|\le n^{1/4}\right)\to 1\text{ as }n\to\infty.$$

Therefore, $\sqrt{n}\left(T_n\right)\underset{P_0}{\overset{d}{\to}}0$.

If $\theta\neq 0$, then

$$\sqrt{n}\left(T_n-\theta\right)=\underbrace{\sqrt{n}\left(\bar{X}_n-\theta\right)}_{\overset{d}{\to}N(0,1)}\underbrace{\mathbf{1}\left\{\left|\bar{X}_n\right|\ge n^{-1/4}\right\}}_{\overset{a.s}{\to}1}+\sqrt{n}\left(0-\theta\right)\underbrace{\mathbf{1}\left\{\left|\bar{X}_n\right|<n^{-1/4}\right\}}_{\overset{a.s}{\to}0}\overset{d}{\to}N(0,1)$$

So $\sqrt{n}\left(T_n-\theta\right)\overset{d}{\to}\begin{cases}0 & \text{if }\theta=0\\ N(0,1) & \text{o.w.}\end{cases}$.

This is a bad estimator. We will explore this in the homework.

♣

2 U-Statistics

2.1 Definitions

Suppose I have $h: X^r\to\mathbb{R}$ and want to estimate $\theta=E\left[h\left(X_1,...,X_r\right)\right]$ where the X_i are independent. Given a sample $X_1,...,X_n$, how should I estimate θ? (Why care about this?)

Example:

Observe that

$$\mathrm{Var}(X)=E\left[X_1^2\right]-E\left[X_1X_2\right]=\frac{1}{2}E\left[\left(X_1-X_2\right)^2\right].$$

So,

$$h\left(X_1,X_2\right)=\frac{1}{2}\left(X_1-X_2\right)^2$$

♣

Remark Always, without loss of generality, we assume h is symmetric.

I should estimate θ with with U-Statistics (Hoeffding 1940s). It allows us to (1) abstract away annoying details and still perform inference and (2) develop statistics and tests that do not depend on parametric assumptions (non-parametric) while things are parametric.

Definition 2.1 (U-Statistics)**.** *Let $h: \ X^r \to \mathbb{R}$ be symmetric. (Kernel Function). For $X_i \overset{i.i.d}{\sim} P$, define $\theta(P) := E_P[h(X_1, ..., X_r)]$ and*

$$U_n := \frac{1}{\binom{n}{r}} \sum_{|\beta|=r, \beta \subset [n]} h(X_\beta)$$

where β ranges over size r subsets of $[n] = \{1, ..., n\}$, $X_\beta = (X_{i_1}, ..., X_{i_r})$ for $\beta = (i_1, ..., i_r)$.

Remark $\quad E_P[U_n] = \theta(P)$, so it is unbiased.

Why use an U-statistic at all? Why not

$$\frac{1}{\binom{n}{r}} \sum_{\ell=1}^{\frac{n}{r}} h\left(X_{\ell(r-1)+1}, ..., X_{\ell r}\right)?$$

Let $\left\{X_{(1)}, ..., X_{(n)}\right\}$ be the sample with "index" information removed. (e.g. Order Statistics. Generally a histogram. For EE, called "type" of the sample.) Then, under $X_i \overset{i.i.d}{\sim} P$, then $\left\{X_{(i)}\right\}_{i=1}^n$ is a sufficient statistic for everything. Observe that

$$E\left\{h(X_1, ..., X_r) | X_{(1)}, ..., X_{(n)}\right\} = U_n := \frac{1}{\binom{n}{r}} \sum_{|\beta|=r, \beta \subset [n]} h(X_\beta)$$

By Rao-Blackwellization,

$$\mathrm{Var}(U_n) \leq \mathrm{Var}\left(\underbrace{\text{Other Unbiased Estimators}}_{\text{of form Avg}\left(h\left(X_\beta\right),\ |\beta|=r\right)}\right).$$

2.2 Examples

Example (Sample Variance)**:** Consider $h(X_1, X_2) = \frac{1}{2}(X_1 - X_2)^2$ and $E[h(X_1, X_2)] = \frac{1}{2}\left(E\left[X_1^2\right] + E\left[X_2^2\right]\right) - E[X_1, X_2] = \mathrm{Var}(X)$

$$\begin{aligned} U_n &= \frac{1}{\binom{n}{2}} \sum_{1 \leq i < j \leq n} \frac{1}{2}(X_i - X_j)^2 \\ &= \frac{1}{2n(n-1)} \sum_{i,j} (X_i - X_j)^2 \\ &= \frac{1}{2n(n-1)} \sum_{i,j} \left((X_i - \bar{X}_n) - (X_j - \bar{X}_n)\right)^2 \\ &= \frac{1}{2n(n-1)} \sum_{i,j} \left((X_i - \bar{X}_n)^2 + (X_j - \bar{X}_n)^2\right) \\ &= \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2 \end{aligned}$$

♣

Example (Gini's Mean-Difference)**:**

$$h(X_1, X_2) = |X_1 - X_2| \text{ and } E[U_n] = E[|X_1 - X_2|]$$

♣

Example (Quantiles)**:**

$$\theta(P) = P(X \leq t) = \int_{-\infty}^{t} dp \text{ and } h(X) = \mathbf{1}\{X \leq t\}$$

This is a 1st order U-statistic.

♣

Example (Signed Rank Statistic)**:** Suppose we want to know if central location of P is 0 or not. (Even if $E[X]$ does not exist.)

$$\theta(P) := P(X_1 + X_2 > 0)$$

Remark If X_1, X_2 are symmetric about 0 and $\theta(P) = \frac{1}{2}$, then $h(X_1, X_2) = \mathbf{1}\{X_1 + X_2 > 0\}$ and $U_n = \frac{1}{\binom{n}{2}} \sum_{i<j} \mathbf{1}\{X_i + X_j > 0\}$.

♣

Definition 2.2 (Two-sample U-Statistic)**.** *For the samples* $\{X_1, ..., X_n\}$ *and* $\{Y_1, ..., Y_n\}$*,*

$$U = \frac{1}{\binom{n}{r}\binom{m}{s}} \sum_{|\alpha|=s, \alpha\subset[m]} \sum_{|\beta|=r, \beta\subset[n]} h(X_\beta, Y_\alpha)$$

where $h: X^r \times Y^s \to \mathbb{R}$*.* h *is symmetric in* X *and* Y *variables individually.*

Example (Mann-Whitney Statistic)**:**

Is X stochastically dominated by Y? Or are the at the same location?

$$\theta(P) = P(X \leq Y)$$
$$h(X, Y) = \mathbf{1}\{X \leq Y\}$$
$$U_{n,m} = \frac{1}{nm} \sum_{i=1}^{n} \sum_{j=1}^{m} \mathbf{1}\{X_i \leq Y_j\}$$

♣

2.3 Variance of U-Statistics

This is a precursor to asymptotic normality because "1st order terms" dominate everything else.

Definition 2.3. *Assume that* $E\left[|h|^2\right] < \infty$ *for any* $c < r$. *Define*

$$h_c\left(X_1, ..., X_c\right) := E\left[h\left(\underbrace{X_1, ..., X_c}_{fixed}, \underbrace{X_{c+1}, ..., X_r}_{i.i.d\ P}\right)\right].$$

Remark

1. $h_0 = E\left[h\left(X_1, ..., X_r\right)\right] = \theta\left(P\right)$
2. $E\left[h_c\left(X_1, ..., X_c\right)\right] = E\left[h\left(X_1, ..., X_r\right)\right] = \theta\left(P\right)$

Definition 2.4.

$$\hat{h}_c : = h_c - E\left[h_c\right] = h_c - \theta\left(P\right)$$
$$E\left[\hat{h}_c\right] = 0$$

Then define

$$\zeta_c := Var\left(h_c\left(X_1, ..., X_c\right)\right) = E\left[\hat{h}_c^2\right]$$

Goal: Write $Var\left[U_n\right]$ in terms of $\zeta_c' s$ for $c = 1, 2, ..., r$. Note the following:
If $\beta = \{i_1, ..., i_r\}$, $\alpha = \{i_1', ..., i_r'\}$, $S = \alpha \cap \beta$, $|S| = c$, then

$$E\left[\hat{h}\left(X_\beta\right)\hat{h}\left(X_\alpha\right)\right] = \zeta_c.$$