Stats 300b: Theory of Statistics **Winter 2019**

Lecture 7– January 29

Lecturer: John Duchi *Scribes: Kevin Han, Yuchen Wu*

Warning: *these notes may contain factual errors*

Reading: Elements of Large Sample Theory Ch. 3.1, 3.2, 4.1, VDV Chapter 11, 12

Outline:

- Finish "basic" tests
- U-Statistics
  - Definitions
  - Examples
  - Variance calculation

1 Recap: Wald, Likelihood Ratio Tests

Goal: For fixed $\alpha > 0$, find regions $\mathcal{C}_n$ such that for H_0: $\{\theta \in \Theta_0\}$,

$$\sup_{\theta \in \Theta_0} \limsup_{n \to \infty} P_\theta(T_n \notin \mathcal{C}_n) \leq \alpha$$

How to deal with nuisance/composite nulls, e.g.

$$\Theta_0 = \{\theta : [\theta]_{1:k} = \left[\theta^0\right]_{1:k}, \theta_{k+1}, \cdots, \theta_d \text{ unspecified}\}$$

1.1 Wald Test

Let $\Sigma^{(k)}(\theta) =$ First $k \times k$ block of $I(\theta)^{-1}$.

$$\mathcal{C}_{n,\alpha} = \left\{\theta \in \mathbb{R}^d : ([\theta]_{1:k} - [\theta^0]_{1:k})^T \left[\Sigma^{(k)}(\theta^0)\right]^{-1} ([\theta]_{1:k} - [\theta^0]_{1:k}) \leq u_{k,\alpha}^2/n\right\}$$

where $u_{k,\alpha}^2$ was quantile of χ_k^2, i.e. $P(\|w\|_2^2) \geq u_{k,\alpha}^2 = \alpha$ for $w \sim \mathcal{N}(0, I_{k\times k})$.

$$T_n := \begin{cases} \text{Reject} & \text{if } \hat{\theta}_n \notin C_{n,\alpha} \\ \text{Don't Reject} & \text{otherwise} \end{cases}$$

2 Rao Test (Score Test)

We know the (limiting) distribution of $P_n \nabla \ell_\theta = P_n \dot{\ell}_\theta = \frac{1}{n}\sum_{i=1}^n \nabla \ell_\theta(X_i)$ under P_θ, i.e.

$$\sqrt{n}(P_n \dot{\ell}_\theta) \xrightarrow[P_\theta]{d} \mathcal{N}(0, I_\theta).$$

In $H_0 : \theta = \theta_0 \in \mathbf{R}^d$, then

$$n \left(P_n \nabla \ell_{\theta_0}\right)^T I_{\theta_0}^{-1} \left(P_n \nabla \ell_{\theta_0}\right) \xrightarrow[H_0]{d} \chi_d^2.$$

Definition 2.1. *Rao test is to define rejection region*

$$\left(P_n\dot{\ell}_{\theta_0}\right)^T I_{\theta_0}^{-1}\left(P_n\dot{\ell}_{\theta_0}\right) \geq \frac{u_{d,\alpha}^2}{n}$$

Immediately, we have

$$\lim_{n\to\infty} P_{H_0}(\text{reject}) = \alpha$$

Remark

- All of these tests (related score/asymptotic normality) strongly related to optimality. In future, we compute powers under alternatives of form

  $$H_0 : \theta = \theta_0 \quad H_{1,n} : \theta = \theta_0 + \frac{h}{\sqrt{n}}$$

  Look at Power$(h) := \lim_{n\to\infty} P_{H_{1,n}}(T_n \text{ rejects})$

- Rao test has analogues for composite nulls

3 U-Statistics

3.1 Introduction

Suppose we have a function h of k variables, want to estimate $\theta := \mathbb{E}\left[h(X_1, \cdots, X_k)\right]$ where X_i are independent. How should we estimate θ given $\{X_i\}_{i=1}^n$?
Example: $P(X_1 \geq X_2 + t) \quad h(y,z) = \mathbb{1}(y \geq z + t)$ ♣

Example:

$$\text{Var}\,(X) = \frac{1}{2}\mathbb{E}\left[(X_1 - X_2)^2\right] = \mathbb{E}\left[X^2\right] - (\mathbb{E}X)^2. \quad X_i\text{'s are i.i.d.}$$

$$h(x,y) = \frac{1}{2}(x-y)^2$$

♣

To do this, use U-statistics.
Developed by Hoeffding (1940s-ish), one of fathers of nonparametric statistics. Idea is to develop more "robust" tests, e.g. of location, that don't make parametric modeling assumptions. e.g. want more robustness than something like,

$$X \sim N(\theta_1, 1) \quad \text{and} \quad Y \sim N(\theta_2, 1), \quad \text{is } \theta_1 \underset{>}{<} \theta_2 \text{ ?}$$

Allow us to abstract away many annoying details, still perform inference, testing, estimation.

3.2 Definitions

Definition 3.1 (U-Statistic)**.** *For $X_i \overset{i.i.d}{\sim} P$, denote $\theta(P) := E_P[h(X_1, ..., X_r)]$. A U-statistic is a random variable of the form*

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

where $h: X^r \to \mathbb{R}$ is a symmetric (kernel) function, β ranges over all size r subsets of $[n] := \{1, ..., n\}$, and $X_\beta := (X_{i_1}, ..., X_{i_r})$ for $\beta = (i_1, ..., i_r)$.

Remark The U in "U-statistics" is because $\mathbb{E}_P[U_n] = \theta(P) := \mathbb{E}[h(X_1, ..., X_r)]$, so U_n is **unbiased**.

Why use a U-statistic at all? Why not use

$$h(X_1, X_2, ..., X_r)$$

or

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

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

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

By Rao-Blackwellization, we know that for *any* convex (loss) function L and any r.v. Z_n such that $\mathbb{E}[Z_n|(X_{(i)})_{1 \le i \le n}] = U_n$,

$$\mathbb{E}[L(Z_n)] \ge \mathbb{E}[L(U_n)].$$

3.3 Examples

Example (Sample Variance)**:** Consider $h(x, y) = \frac{1}{2}(x-y)^2$. Then $\mathbb{E}[h(X_1, X_2)] = \frac{1}{2}\left(\mathbb{E}\left[X_1^2\right] + \mathbb{E}\left[X_2^2\right]\right) - \mathbb{E}[X_1, X_2] = \mathrm{Var}(X)$. When we have more than two samples, we use

$$\begin{aligned} U_n &= \frac{1}{\binom{n}{2}} \sum_{1 \le i < j \le n} \frac{1}{2}(X_i - X_j)^2 \\ &= \frac{2}{n(n-1)} \sum_{1 \le i < j \le n} \frac{1}{2}(X_i - X_j)^2 \\ &= \frac{1}{2n(n-1)} \sum_{1 \le i < j \le n} (X_i - X_j)^2 \end{aligned}$$

♣

Example (Gini's Mean-Difference)**:** $h(x, y) = |x - y|$ and $\mathbb{E}[U_n] = \mathbb{E}[|X_1 - X_2|]$. ♣

Example (Quantiles, $r = 1$)**:**

$$\theta(P) = P(X \leq t) \text{ and } h(X) = \mathbf{1}\{X \leq t\}$$

This is a first order U-statistic. ♣

Example (Signed Rank Statistic)**:** Provide information about location of distributions

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

This means $h(x,y) = \mathbf{1}\{x+y>0\}$ and $U_n = \frac{1}{\binom{n}{2}} \sum_{i<j} \mathbf{1}\{X_i + X_j > 0\}$. ♣

Definition 3.2 (Two-sample U-Statistic)**.** *Given two samples $\{X_1,...,X_n\}$and $\{Y_1,...,Y_n\}$, $N = n+m$, a two-sample U-statistic is a random variable of the form*

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

where $h: X^r \times Y^s \to \mathbb{R}$. h is symmetric in its first r arguments and in its last s arguments.

Big Use: Are samples coming from same distribution or not?

Example (Mann-Whitney Statistic)**:** Test difference in locations of X and Y

$$U_{n,m} = \frac{1}{nm} \sum_{i=1}^{n} \sum_{j=1}^{m} \mathbf{1}\{X_i \leq Y_j\},$$
$$\mathrm{E}(U_N) = \mathrm{P}(X \leq Y),$$
$$\text{NULL:} H_0 = \{P(X \leq Y) = \frac{1}{2}\}, \text{i.e. same location}$$

♣ **Game Plan:** Can we get asymptotics of these U-statistics under appropriate distributions?

The answer is yes. Project out annoying(lower order) terms, see what is left(iid sums).

3.4 Variance of U-Statistics(Hoeffding)

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

Definition 3.3. *Assume that $E\left[|h|^2\right] < \infty$, $X_i \sim P$, iid, for any $c < r$. Define*

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

Remark

1. $h_0 = E[h(X_1,...,X_r)] = \theta(P)$

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 3.4.

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

Then define

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

(Note that $\hat{h}(x_{1:r}) = h(x_{1:r}) - \theta(P)$*.)*

Consider Variances: Fix $A, B \subset [n]$, $|A| = |B| = r$, let $|A \cap B| = c$

Define: $\zeta_C = \mathbb{E}\left[\hat{h}(X_A)\hat{h}(X_B)\right]$

Claim: $\zeta_C = \mathbb{E}\left[\hat{h}_C(x1:C)^2\right] = \text{Var}(\hat{h}_C)$

Proof Using the symmetry of h,

$$\begin{aligned}
\mathbb{E}\left[\hat{h}(X_A)\hat{h}(X_B)\right] &= \mathbb{E}\left[\hat{h}(X_{A\setminus S}, X_S)\hat{h}(X_{B\setminus S}, X_S)\right] \\
&= \mathbb{E}\left[\mathbb{E}[\hat{h}(X_{A\setminus S}, X_S) \mid X_S] \cdot \mathbb{E}[\hat{h}(X_{B\setminus S}, X_S) \mid X_S]\right] \quad \text{(since } X_{A\setminus S},\ X_{B\setminus S} \text{ indep.)} \\
&= \mathbb{E}\left[\hat{h}_c(X_S) \cdot \hat{h}_c(X_S)\right] \\
&= \zeta_c.
\end{aligned}$$

□

Now let's compute the variance of U_n

Theorem 1. *Let* U_n *be an* r^{th} *order U-statistic. Then*

$$\text{Var}U_n = \frac{r^2}{n}\zeta_1 + O(n^{-2}).$$

Proof There are $\binom{n}{r}\binom{r}{c}\binom{n-r}{r-c}$ ways to select a pair of subsets of $[n]$, each of size r, with c common elements. Hence,

$$\begin{aligned}
U_n - \theta &= \binom{n}{r}^{-1} \sum_{|B|=r} \hat{h}(X_B), \\
\text{Var}U_n &= \binom{n}{r}^{-2} \sum_{|A|=r} \sum_{|B|=r} \mathbb{E}\left[\hat{h}(X_A)\hat{h}(X_B)\right] \\
&= \binom{n}{r}^{-2} \sum_{c=1}^{r} \binom{n}{r}\binom{r}{c}\binom{n-r}{r-c}\zeta_c \\
&= \sum_{c=1}^{r} \frac{r!^2}{c!(r-c)!^2} \frac{(n-r)(n-r-1)\dots(n-2r+c+1)}{n(n-1)\dots(n-r+1)}\zeta_c.
\end{aligned}$$

For fixed c, $\frac{(n-r)(n-r-1)\ldots(n-2r+c+1)}{n(n-1)\ldots(n-r+1)}$ has $r-c$ terms in the numerator and r terms in the denominator. Hence,

$$\begin{aligned}
\text{Var}U_n &= r^2 \frac{(n-r)(n-r-1)\ldots(n-2r+2)}{n(n-1)\ldots(n-r+1)}\zeta_1 + \sum_{c=2}^{r} O\left(\frac{n^{r-c}}{n^r}\right)\zeta_c \\
&= r^2\left[\frac{1}{n} + O(n^{-2})\right]\zeta_1 + O(n^{-2}) \\
&= \frac{r^2}{n}\zeta_1 + O(n^{-2}).
\end{aligned}$$

□