Lecture 8 – February 2

Lecturer: John Duchi *Scribe: Dylan Greaves*

Warning: *these notes may contain factual errors*

Reading:

Outline:

- Finish U-statistics
- Testing and Confidence Intervals
- Duality between Testing and Confidence
- (Generalized) Likelihood Ratio Tests

Recap: Last lecture, we proved the following two results:

Claim 1. *Let $U_n = \binom{n}{r}^{-1} \sum_{|\beta|=r} h(X_\beta)$, and $h_c(x_1, \ldots, x_c) = \mathbb{E}[h(x_1, \ldots, x_c, X_{c+1}, \ldots, X_r)]$, then*

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

where $\zeta_1 = \mathrm{Var}(h_1)$.

Claim 2. *If $\mathcal{S}$ is a linear subspace and $\hat{S}_n$ is the projection of T_n on $\mathcal{S}$, then*

$$\frac{\mathrm{Var}(\hat{S}_n)}{\mathrm{Var}(T_n)} \to 1 \implies \frac{T_n - \mathbb{E}T_n}{\sqrt{\mathrm{Var}(T_n)}} - \frac{\hat{S}_n - \mathbb{E}\hat{S}_n}{\sqrt{\mathrm{Var}(\hat{S}_n)}} \xrightarrow{P} 0.$$

We will combine these two ideas to show the asymptotic normaility of U-statistics.

1 Asymptotic Normality of U-statistics

(Hajék) The main idea is to use projections onto sets of the form

$$\mathcal{S}_n = \{\sum_{i=1}^{n} g_i(X_i) : g_i(X_i) \in L_2(P)\}.$$

Theorem 3. *Let h be a symmetric kernel (function) of order r and let $\mathbb{E}h^2 < \infty$, U_n be the associated U-statistic, then*

$$\sqrt{n}(U_n - \theta) \xrightarrow{d} \mathsf{N}(0, r^2\zeta_1),$$

where $\theta = \mathbb{E}U_n = \mathbb{E}h(X_1, \ldots, X_n)$.

Proof Let $\hat{U}_n$ be defined as $\hat{U}_n = \sum_{i=1}^n \mathbb{E}[U_n - \theta | X_i]$, then $\hat{U}_n$ is the projection of $U_n - \theta$ onto $\mathcal{S}_n$. Let's compare the variances and expectations.

Let $\beta \subseteq [n]$, $|\beta| = r$, then

$$\mathbb{E}[h(X_\beta) - \theta | X_i] = \begin{cases} 0 & i \notin \beta \\ h_1(X_i) & i \in \beta \end{cases}.$$

Then

$$\begin{aligned}
\mathbb{E}[U_n - \theta | X_i] &= \binom{n}{r}^{-1} \sum_{|\beta|=r} \mathbb{E}[h(X_\beta) - \theta | X_i = x] \\
&= \binom{n}{r}^{-1} \sum_{|\beta|=r, i\in\beta} h_1(X_i) \\
&= \binom{n}{r}^{-1} \binom{n-1}{r-1} h_1(X_i) = \frac{r}{n} h_1(X_i)
\end{aligned}$$

It follows that $\hat{U}_n = \frac{r}{n} \sum_{i=1}^n h_1(X_i)$ and certainly $\sqrt{n}(\hat{U}_n - \theta) \xrightarrow{d} \mathsf{N}(0, r^2 \zeta_1)$.

Now apply ratio of variance condition (Claim 2), since

$$\begin{aligned}
\mathrm{Var}(U_n) &= \frac{r^2}{n} \zeta_1 + O(n^{-2}) \\
\mathrm{Var}(\hat{U}_n) &= \frac{r^2}{n} \zeta_1
\end{aligned}$$

we have $\frac{\mathrm{Var}(U_n)}{\mathrm{Var}(\hat{U}_n)} \to 1$ as $n \to \infty$, from which we conclude U_n and $\hat{U}_n$ have the same asymptotic behavior. □

2 Testing and Confidence Intervals

We've seen a number of scenarios where

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} \mathsf{N}(0, \Sigma).$$

Suppose we would like to make the following claim about the population parameter θ_0: "With reasonably high confidence, $\theta_0 \in \mathcal{C}_n$, where $\mathcal{C}_n \subseteq \mathbb{R}^d$ is a set."

Example 1: Suppose $\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow[\theta_0]{d} \mathsf{N}(0, I_{\theta_0}^{-1})$. Say I_θ is continuous in θ and I_θ is invertible. Let

$$C_{n,\gamma} := \{\theta \in \mathbb{R}^d : (\theta - \hat{\theta}_n)^T I_{\hat{\theta}_n} (\theta - \hat{\theta}_n) \le \frac{\gamma}{n}\}.$$

For $\theta = \theta_0$, we have

$$\begin{aligned}
n(\theta_0 - \hat{\theta}_n)^T I_{\hat{\theta}_n}(\theta_0 - \hat{\theta}_n) &= (\sqrt{n}(\hat{\theta}_n - \theta_0))^T (I_{\theta_0} + o_P(1))(\sqrt{n}(\hat{\theta}_n - \theta_0)) \\
&= (\underbrace{\sqrt{n}(\hat{\theta}_n - \theta_0)}_{\xrightarrow{d} \mathsf{N}(0, I_{\theta_0}^{-1})})^T I_{\theta_0}(\sqrt{n}(\hat{\theta}_n - \theta_0)) + o_P(1) \\
&\xrightarrow{d} Z^T I_{\theta_0} Z && Z \sim \mathsf{N}(0, I_{\theta_0}^{-1}) \\
&\overset{d}{\equiv} \|W\|_2^2 \overset{d}{\equiv} \chi_d^2 && W \sim \mathsf{N}(0, I_{d\times d})
\end{aligned}$$

Then

$$\begin{aligned}
\mathbb{P}_{\theta_0}(\theta_0 \in C_{n,\gamma}) &= \mathbb{P}_{\theta_0}((\theta_0 - \hat{\theta}_n)^T I_{\hat{\theta}_n}(\theta_n - \hat{\theta}_n) \leq \frac{\gamma}{n}) \\
&\to \mathbb{P}(\|W\|_2^2 \leq \gamma) && W \sim \mathsf{N}(0, I_{d\times d}) \\
&= \mathbb{P}(\chi_d^2 \leq \gamma).
\end{aligned}$$

$C_{n,\gamma}$ is pivotal, since it doesn't depend on the parameter θ_0. If for some level $\alpha < 1$ you want that $P_{\theta_0}(\theta_0 \in C_{n,\gamma}) \to \alpha$, take $C_n = C_{n,\gamma}$, where γ is chosen such that $P(\chi_d^2 \leq \gamma) = \alpha$. ♣

3 Dual Problem to Confidence Sets

The typical approach to hypothesis testing is the following: Can we reject some type of null hypothesis, that is supposingly conjecture $H_0 : P_{\theta_0}$? Can we get results like

$$P_{\theta_0}(\text{data at least as "extreme" as what we got}) \leq \alpha?$$

It's questionable whether this is even a reasonable thing to do, since this is a ill-formed definition – "extreme" is vague. One might also take philosophical issue with this approach, since the only conclusions that result from it are negative statements – "this null hypothesis doesn't explain the world." While this may be troubling, it's worthwhile to note that this is also nature of the scientific method: scientific hypotheses are never proven "true," prevailing hypotheses are only held until they are falsified by new observations (e.g. Michelson-Morley experiment (1887) and the æther drag hypothesis).

Definition 3.1 (p-value)**.** *Let $H_0 : \{P_\theta : \theta \in \Theta_0\}$. The p-value associated with a sample $X_1, \ldots, X_n$ is defined to be*

$$\sup_{\theta \in \Theta_0} P_\theta(\textit{data as extreme as } X_1, \ldots, X_n \textit{ observed})$$

Example 2: Let $H_0 : X_i \overset{\text{iid}}{\sim} \mathsf{N}(0, 1)$. The standard p-value is given by

$$P_0(|\bar{Z}| > |\hat{\theta}|),$$

where $\hat{\theta}_n = \frac{1}{n}\sum_{i=1}^n X_i$. ♣

How can we understand these and develop a few tests with reasonable properties?

4 Generalized Likelihood Ratio Tests

Recall the classic Neyman-Pearson setup with simple null and alternative:

$$H_0 : p_0$$
$$H_1 : p_1$$

The test that is the "best" (maximizes power at all levels) is the likelihood ratio test, where if $L_1(x) = \log dP_1(x)$, $L_0(x) = \log dP_0(x)$, and $T(x) = L_1(x) - L_0(x) = \log \frac{dP_1}{dP_0}(x)$, the most powerful test is given by

$$\begin{cases} \text{accept } H_1/\text{reject } H_0 & T > t \\ \text{accept } H_0/\text{reject } H_1 & T < t \\ \text{balance} & T = t \end{cases}$$

for some t.

4.1 Generalized LRT

We now consider a more general scenario with composite null and alternative. Suppose

$$H_0 : \theta \in \Theta_0$$
$$H_1 : \theta \in \Theta$$

where usually $\Theta_0 \subseteq \Theta$. Define

$$T(x) = \log \frac{\sup_{\theta \in \Theta} p(x, \theta)}{\sup_{\theta \in \Theta_0} p(x, \theta)} = \frac{p(x, \hat{\theta}_{\text{MLE}})}{\sup_{\theta \in \Theta_0} p(x, \theta)}.$$

The Generalized LRT rejects H_0 if $T(x) > t$.

Suppose $\{P_\theta\}_{\theta \in \Theta}$ is nice enough that the MLE is asymptotically normal,

$$I_{\theta_0} = \mathbb{E} \nabla \ell_{\theta_0} \ell_{\theta_0}^T = -\mathbb{E} \nabla^2 \ell_{\theta_0},$$

and

$$\left\|\nabla^2 \ell_\theta(x) - \nabla^2 \ell_{\theta'}(x)\right\|_{\text{op}} \le M(x) \left\|\theta - \theta'\right\|,$$

where $\mathbb{E}_\theta M^2(X) < \infty$. Then we have the following asymptotic result:

Proposition 4 (Wilk's Theorem). *Let $\Theta_0 = \{\theta_0\}$ be a point null, $\Theta = \mathbb{R}^d$. Let $L_n(x, \theta) = \sum_{i=1}^n \ell_\theta(x_i) = \sum_{i=1}^n \log p_\theta(x_i)$ and $T_n(x) = L_n(x, \hat{\theta}_{MLE}) - L_n(x, \theta_0)$. Then*

$$2T_n(X) \xrightarrow[\theta_0]{d} \chi_d^2,$$

where $X = (X_1, \dots, X_n)$ and $X_i \overset{\text{iid}}{\sim} P_{\theta_0}$.

Proof Let $\hat{\theta}_n = \operatorname{argmax}_{\theta \in \Theta} L_n(X, \theta) = \hat{\theta}_{\text{MLE}}$.
Under H_0, $\hat{\theta}_{\text{MLE}} - \theta_0 \xrightarrow[\theta_0]{P} 0$ and $\sqrt{n}(\hat{\theta}_{\text{MLE}} - \theta_0) \xrightarrow{d} \mathsf{N}(0, I_{\theta_0}^{-1})$.

By Taylor's Theorem, we have

$$0 = \nabla L_n(X, \hat{\theta}_n) = \nabla L_n(X, \theta_0) + \nabla^2 L_n(X, \theta_0)(\hat{\theta}_n - \theta_0) + \sum_{i=1}^{n} \mathrm{Err}_i^{(1)}(\hat{\theta} - \theta),$$

where $\left\| \mathrm{Err}_i^{(1)} \right\|_{\mathrm{op}} \leq M(X_i) \left\| \hat{\theta}_n - \theta_0 \right\|$.
Similarly,

$$L_n(X, \hat{\theta}_n) = L_n(X, \theta_0) + \nabla L_n(X, \theta_0)^T(\hat{\theta}_n - \theta_0) + \frac{1}{2}(\hat{\theta}_n - \theta_0)^T \nabla^2 L_n(X, \theta_0)(\hat{\theta}_n - \theta_0) + \sum_{i=1}^{n} (\hat{\theta}_n - \theta_0)^T \mathrm{Err}_i^{(2)}(\hat{\theta}_n - \theta_0)$$

where $\left\| \mathrm{Err}_i^{(2)} \right\|_{\mathrm{op}} \leq M(X_i) \left\| \hat{\theta}_n - \theta_0 \right\|$.
Substituting the first equation into the second, and letting $\mathrm{Err}_i = \mathrm{Err}_i^{(2)} - \mathrm{Err}_i^{(1)}$, we have

$$T(X) = L_n(X, \hat{\theta}_n) - L_n(X, \theta_0) = -\frac{1}{2}(\hat{\theta}_n - \theta_0)^T \nabla^2 L_n(X, \theta_0)(\hat{\theta}_n - \theta_0) + \sum_{i=1}^{n} (\hat{\theta}_n - \theta_0)^T \mathrm{Err}_i(\hat{\theta}_n - \theta_0).$$

Since $\frac{1}{n}\sum_{i=1}^{n} \nabla^2 \ell_{\theta_0}(X_i) \xrightarrow{P} -I_{\theta_0}$, $\sum_{i=1}^{n} \mathrm{Err}_i \xrightarrow{P} 0$, and $\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow[H_0]{d} \mathsf{N}(0, I_{\theta_0}^{-1})$, it follows that

$$2T(X) = \sqrt{n}(\hat{\theta}_n - \theta_0)^T I_{\theta_0}(\hat{\theta}_n - \theta_0) + o_P(1) \xrightarrow{d} \chi_d^2.$$

$\square$