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

Lecture 5January 22

Lecture 5 – January 22

Lecturer: John Duchi *Scribe: Yujia Jin*

Warning: *these notes may contain factual errors*

Reading: Elements of Large Sample Theory Ch. 3.1, 3.2, 4.1 and Testing Statistical Hypotheses Ch. 12.4.

Outline:

- Efficiency Estimators.
- Tests (beginning ideas in asymptotic regime)
  - confidence intervals
  - likelihood ratio, tests

1 Recap

Asymptotic Normality

If family $\{P_\theta\}_{\theta\in\Theta}$ is nice enough and $\hat{\theta}_n$ is MLE, $\sqrt{n}(\hat{\theta}_n - \theta^*) \xrightarrow{d} N(0, I_{\theta^*}^{-1}), I_\theta = \mathbb{E}_\theta[\nabla l_\theta \nabla l_\theta^{\mathrm{T}}]$.

Exponential Family

$$P_\theta(x) = \exp(\theta^{\mathrm{T}} T(x) - A(\theta))$$
$$A(\theta) = \log \int \exp(\langle \theta, T(x)\rangle) d\mu(x)$$
$$\sqrt{n}(\hat{\theta}_n - \theta) \to^d N(0, \nabla^2 A(\theta)^{-1})$$

Here $\hat{\theta}_n$ can either be MLE or moment-matching estimator (equivalent).

2 Efficiency of Estimators

Definition 2.1. *An estimator $\hat{\theta}_n$ is efficient for parameter θ if $\sqrt{n}(\hat{\theta}_n - \theta) \to^d N(0, I_\theta^{-1})$.*

Example 1:

- Gaussian:

  mean is efficient

- Poisson:

$$x \in \mathbb{N} = \{0, 1, 2, \cdots\},$$
$$p_\lambda(x) = \frac{\lambda^x e^{-x}}{x!} = \exp(x\log\lambda - \lambda - \log(x!)).$$

  Thus, $\theta = \log\lambda, i.e. \lambda = e^\theta$.

We can write $p_\theta(x) = \exp(\theta x - e^\theta - \log(x!))$. We have the properties: $A(\theta) = e^\theta, A'(\theta) = A''(\theta) = e^\theta$. And $\hat{\theta}_n$ satisfies $P_n X = A'(\theta) = e^\theta = \mathbb{E}_\theta[x]$ or $\theta = \log P_n(X)$. By δ-method, we know $\sqrt{n}(\hat{\theta}_n - \theta) = \sqrt{n}(\log P_n X - log\lambda) \xrightarrow{d} N(0, \frac{1}{\lambda}) = N(0, e^{-\theta}) = N(0, A''(\theta)^{-1})$.

♣

3 Comparing Estimators

Definition 3.1. *Let $\hat{\theta}_n$ and T_n be sequences of estimators of $\theta \in R$. Assume $\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} N(0, \sigma^2(\theta))$ and for some $m(n) \to \infty$, $\sqrt{n}(T_{m(n)} - \theta) \xrightarrow{d} N(0, \sigma^2(\theta))$.*

Then the asymptotic relative efficiency (ARE) of $\hat{\theta}_n$ to T_n is

$$ARE := \liminf_{n\to\infty} \frac{m(n)}{n}.$$

Remark If ARE of $\hat{\theta}_n$ vs. T_n is $c \geq 0$, then to get an estimate of θ of some "quality" as $\hat{\theta}_n$ (i.e. error scaling like $\sqrt{\frac{\sigma^2(\theta)}{n}}$), T_n requires sample size C-times larger than $\hat{\theta}_n$.

3.1 Confidence Interval Intuition

If ARE of $\hat{\theta}_n$ vs. T_n is $c \geq 0$ (more strictly here we assume $\lim_{n\to\infty} \frac{m(n)}{n} = c$), let $z_{1-\alpha/2}$ be level α confidence interval for $N(0,1)$, i.e. $Z \sim N(0,1), P(|Z| \geq z_{1-\alpha/2}) = \alpha$. Now we consider sets:

$$\begin{aligned}
c_\theta &= (\hat{\theta}_n - z_{1-\alpha/2}\sqrt{\frac{\sigma^2(\theta)}{n}}, \hat{\theta}_n + z_{1-\alpha/2}\sqrt{\frac{\sigma^2(\theta)}{n}}) \\
c_T &= (T_m - z_{1-\alpha/2}\sqrt{\frac{\sigma^2(\theta)}{m} \cdot \frac{m}{m^{-1}(m)}}, T_m + z_{1-\alpha/2}\sqrt{\frac{\sigma^2(\theta)}{m} \cdot \frac{m}{m^{-1}(m)}}) \\
&\approx (T_m - z_{1-\alpha/2}\sqrt{\frac{\sigma^2(\theta)}{m}c}, T_m + z_{1-\alpha/2}\sqrt{\frac{\sigma^2(\theta)}{m}c})
\end{aligned}$$

where $m^{-1}(m) = n$ s.t. $m(n) = m$ and we assume in addition that it exists.

For both, we have $\lim_{n\to\infty} P(\theta \in C) = 1 - \alpha$ by definition of Asymptotic Normality.

Proposition 1. *Let*

$$\begin{aligned}
\sqrt{n}(\hat{\theta}_n - \theta) &\xrightarrow{d} N(0, \sigma^2(\theta)), \\
\sqrt{n}(T_n - \theta) &\xrightarrow{d} N(0, \tau^2(\theta)).
\end{aligned}$$

Then the ARE of $\hat{\theta}_n$ w.r.t T_n is $\frac{\tau^2(\theta)}{\sigma^2(\theta)}$.

Proof Let $m(n) = \lceil \frac{\tau^2}{\sigma^2} n \rceil$, then

$$\sqrt{n}(T_{m(n)} - \theta) = \sqrt{\frac{n}{m(n)}}\sqrt{m(n)}(T_{m(n)} - \theta) \xrightarrow{d} N(0, \sigma^2(\theta))$$

by noticing that $\sqrt{\frac{n}{m(n)}} \to \frac{\sigma}{\tau}$ and $\sqrt{m(n)}(T_{m(n)} - \theta) \xrightarrow{d} N(0, \tau^2(\theta))$. □

So if $\tau^2 \geq \sigma^2$, we prefer $\hat{\theta}_n$ to T_n.

3.2 Comparison of Estimators

Definition 3.2. *Suppose we have $T_n, \hat{\theta}_n$ s.t. $\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} N(0, \sigma^2(\theta))$, $\sqrt{n}(T_n - \theta) \xrightarrow{d} N(0, \tau^2(\theta))$ and $\tau^2(\theta) \le \sigma^2(\theta)$ everywhere, with $\tau^2(\theta) < \sigma^2(\theta)$ strictly for some θ_0. If $\sigma^2(\theta) = I^{-1}(\theta)$, then T_n is super-efficient.*

Example 2: Hodge's counterexample/super-efficient estimator.

Let $X_i \xrightarrow{i.i.d} N(\theta, 1)$, $\hat{\theta}_n := \frac{1}{n}\sum_{i=1}^n X_i$. Define

$$T_n := \begin{cases} \overline{X_n} & \text{if} \overline{X_n} \ge n^{-\frac{1}{4}} \\ 0 & \text{otherwise} \end{cases}$$

What is the limiting distribution?

When $\theta = 0$,

$$P_\theta(\sqrt{n}T_n = 0) = P_\theta(|\overline{X_n}| < n^{-\frac{1}{4}}) = P_\theta(|\sqrt{n}\overline{X_n}| < n^{\frac{1}{4}}) \to 1$$

since $\sqrt{n}\overline{X_n} \sim N(0, 1)$. Thus we have

$$\sqrt{n}(T_n - d) \xrightarrow{d} 0.$$

When $\theta \ne 0$,

$$\begin{aligned} \sqrt{n}(T_n - \theta) &= \sqrt{n}(\overline{X_n} - \theta)\mathbb{1}(|\overline{X_n}| \ge n^{-\frac{1}{4}}) + \sqrt{n}(0 - \theta)\mathbb{1}(|\overline{X_n}| \le n^{-\frac{1}{4}}) \\ &= \sqrt{n}(\overline{X_n} - \theta) + O_p(1) \xrightarrow{d} N(0, 1). \end{aligned}$$

as $\mathbb{1}((|\overline{X_n}| \ge n^{-\frac{1}{4}})) \to 1$ and $\mathbb{1}(|\overline{X_n}| \le n^{-\frac{1}{4}}) \to 0$ enentually. ♣

Remark Is it good? (See Homework.) This relates to why fisher couldn't prove efficiency is efficiency and what optimality is meant for estimation.

4 Testing

Definition 4.1. *A scientific method: propose a hypothesis → develop experiment → when fail, reject; otherwise, cannot reject.*

Remark The philosophy here is we are not able to verify but only to falsify.

We've seen many situations where we have some type of asymptotic normality: $\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N(0, I_{\theta_0}^{-1})$. Suppose we'd like to say with reasonably high confidence, $\theta_0 \in C_n$. (C_n here is some given set; not scientific method).

Example 3:

If $\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N(0, I_{\theta_0}^{-1})$ and I_θ is continuous in θ, let's try $C_{n,r} := \{\theta : (\theta - \hat{\theta}_n)^{\mathrm{T}} I_{\hat{\theta}_n})(\theta - \hat{\theta}_n) \le \frac{r}{n}\}$. Then we have:

$$\begin{aligned} n(\theta_0 - \hat{\theta}_n)^{\mathrm{T}} I_{\hat{\theta}_n}(\theta_0 - \hat{\theta}_n) &= \sqrt{n}(\theta_0 - \hat{\theta}_n)^{\mathrm{T}} I_{\hat{\theta}_n})\sqrt{n}(\theta_0 - \hat{\theta}_n) \\ &= \sqrt{n}(\theta_0 - \hat{\theta}_n)^{\mathrm{T}}(I_{\theta_0} + o_p(1))\sqrt{n}(\theta_0 - \hat{\theta}_n) \\ &\xrightarrow{d} z^T I_{\theta_0} z, \text{where } z \sim N(0, I_{\theta_0}^{-1}) \\ &\stackrel{d}{=} \|w\|_2^2, \text{where } w \sim N(0, I_{\theta_0}) \\ &\stackrel{d}{=} \chi_d^2. \text{(chi-square with } d\text{-degree of freedom)} \end{aligned}$$

Then $P_{\theta_0}(\theta_0 \in C_{n,r}) \to P(\|w\|_2^2 \le r)$ by definition. ♣

Dual Problem: (Science!)

Can we reject some type of null hypothesis? If we conjecture P_{θ_0} is true, where are we confident it is false?

Want: P_{θ_0}(see data as extreme as what we've observed) $\le \alpha$, then reject θ_0.

Definition 4.2. *(p-value) Let* $H_0 = \{P_\theta \text{ s.t. } \theta \in \Theta_0\}$. *The p-value associated with sample* $X_1, X_2, \cdots, X_n$ *is* $\sup_{\theta \in \Theta_0} P_\theta(\textit{Data as extreme as } X_1, \cdots, X_n \textit{ observed})$.