Stats 300b: Theory of Statistics **Winter 2018**

Lecture 5 – January 23

Lecturer: John Duchi *Scribe: Zijun GAO*

Warning: *these notes may contain factual errors*

Reading:

- ELST (Lehmann) 3.1, 3.2, 4.1.
- TSH (Lehmann, Romano) 12.4.

Outline:

1. Efficiency of estimators
2. Testing
   (a) Confidence sets
   (b) Likelihood-ratios

Notation: In the lecture note, we use I_θ to denote the Fisher information with regard to parameter θ. Particularly, we let

$$I_\theta = \mathbb{E}\left[\nabla \ell_\theta \nabla \ell_\theta^T\right] = -\mathbb{E}\left[\nabla^2 \ell_\theta\right].$$

Recap: In the last class, we talked about asymptotic normality. Particularly, we have the following theorem with regard to MLE.

Theorem 1 (Asymptotic normality of MLE)**.** *If a family of distributions $\{\mathbb{P}_\theta\}$ are "nice", and let $\hat{\theta}_n$ be the MLE, then*

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow[\theta]{d} \mathsf{N}(0, I_\theta^{-1}).$$

Example 1 (Exponential family)**.** *Consider exponential family with densities and moment generating functions*

$$p_\theta(x) = \exp\{\theta^\top T(x) - A(\theta) - \log(k(x))\}, \quad A(\theta) = \log\left(\int \exp\{\theta^\top T(x)\} d\mu(x)\right).$$

Then for either moment estimators or MLE, we have

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow[\theta]{d} \mathsf{N}(0, (\nabla^2 A(\theta))^{-1}).$$

Remark In fact, when the curvature of the likelihood function is large, the estimation is usually easier.

1 Efficiency of estimators

1.1 Efficiency of estimators

Note: we do not introduce the rigorous version of efficiency of estimators here. We will discuss it at the end of the term.

Definition 1.1 (Efficiency). *An estimator T_n for θ is efficient in model $\{\mathbb{P}_\theta\}_{\theta\in\Theta}$ if*

$$\sqrt{n}(T_n-\theta)\xrightarrow[\theta]{d}\mathsf{N}(0,I_\theta^{-1}).$$

Example 2 (Gaussian sample mean). *Consider a family of Gaussian distributions $\{\mathsf{N}(\theta,1)\}_{\theta\in\mathbb{R}}$ and the sample mean estimator $T_n=\bar{X}$, then we have*

$$\sqrt{n}(T_n-\theta)\xrightarrow[\theta]{d}\mathsf{N}(0,1).$$

Note that $I_\theta=1$, therefore, T_n is efficient in model $\{N(\theta,1)\}_{\theta\in\mathbb{R}}$.

Example 3 (Poisson). *Consider a family of Poisson distributions $\{Poi(\lambda)\}_{\lambda\in\mathbb{R}^+}$, where*

$$\mathbb{P}_\lambda(x)=\frac{\lambda^x}{x!}\exp\{-\lambda\}=\exp\{x\log(\lambda)-\lambda-\log(x!)\}=\exp\{x\cdot\theta-\exp\{\theta\}-\log(x!)\}.$$

where $\theta=\log(\lambda)$. We then have the following quantities

$$A(\theta)=\exp\{\theta\},\quad \dot{A}(\theta)=\exp\{\theta\},\quad \ddot{A}(\theta)=\exp\{\theta\}.$$

Consider the moment estimator $T_n=\log(\bar{X})$, and we have

$$\sqrt{n}(T_n-\theta)\xrightarrow[\theta]{d}\mathsf{N}(0,\exp\{-\theta\}).$$

Therefore, T_n is efficient in model $\{Poi(\lambda)\}_{\lambda\in\mathbb{R}^+}$.

1.2 Asymptotic relative efficiency (ARE)

Definition 1.2 (Asymptotic relative efficiency (ARE)). *Let $\hat{\theta}_n$ and T_n be estimators of parameter $\theta\in\mathbb{R}$. Assume that*

$$\sqrt{n}(\hat{\theta}_n-\theta)\xrightarrow[\theta]{d}\mathsf{N}(0,\sigma^2(\theta)).$$

Let $m(n)\to\infty$ such that

$$\sqrt{n}(T_{m(n)}-\theta)\xrightarrow[\theta]{d}\mathsf{N}(0,\sigma^2(\theta)).$$

The asymptotic relative efficiency of $\hat{\theta}_n$ with respect to T_n is

$$\lim_{n\to\infty}\inf\ \frac{m(n)}{n}.$$

We have several interpretations of the definition of ARE. For simplicity, in explanations we assume that $\text{ARE}=\lim_{n\to\infty}m(n)/n$ exists and equals some constant c.

- Sample size: if $c\geq 1$, then T_n requires $c\cdot n$ samples compared to n samples to get an estimate of the same "quality" as $\hat{\theta}_n$.

- Confidence interval: we want $1-\alpha$ confidence intervals I_α for θ such that

$$\mathbb{P}\left(\theta \in I_\alpha\right) \longrightarrow \alpha \in (0,1).$$

Let $z_{1-\alpha/2}$ satisfy

$$\mathbb{P}\left(|Z| \geq z_{1-\alpha/2}\right) = \alpha,$$

where $Z \sim \mathsf{N}(0,1)$. Then the valid confidence intervals of $\hat{\theta}_n$ and T_n are:

$$C_{\hat{\theta}_n} : \left(\hat{\theta}_n - z_{1-\alpha/2}\sqrt{\frac{\sigma^2(\theta)}{n}}, \hat{\theta}_n + z_{1-\alpha/2}\sqrt{\frac{\sigma^2(\theta)}{n}}\right),$$

$$C_{T_n} : \left(T_n - z_{1-\alpha/2}\sqrt{\frac{c\cdot\sigma^2(\theta)}{n}}, T_n + z_{1-\alpha/2}\sqrt{\frac{c\cdot\sigma^2(\theta)}{n}}\right).$$

We compare the lengths of the intervals

$$\frac{\text{length}(C_{T_n})}{\text{length}(C(\hat{\theta}_n))} = \sqrt{c}.$$

Similar to the explanation using confidence interval, we have the following proposition regarding variances of estimators.

Proposition 2 (Asymptotic variances)**.** *Suppose $\hat{\theta}_n$ and T_n are estimators of θ such that*

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow[\theta]{d} \mathsf{N}(0, \sigma^2(\theta)),$$

$$\sqrt{n}(T_n - \theta) \xrightarrow[\theta]{d} \mathsf{N}(0, \tau^2(\theta)).$$

Then the ARE of $\hat{\theta}_n$ with respect to T_n is $\sigma^2(\theta)/\tau^2(\theta)$.

Proof Let $m(n) = \lceil n\cdot\tau^2/\sigma^2\rceil$. Then

$$\sqrt{n}\left(T_{m(n)} - \theta\right) = \underbrace{\sqrt{\frac{n}{m(n)}}}_{\to\sigma(\theta)/\tau(\theta)} \cdot \underbrace{\sqrt{m(n)}\cdot\left(T_{m(n)} - \theta\right)}_{\xrightarrow[\theta]{d}\mathsf{N}(0,\tau^2(\theta))} \xrightarrow{d} \mathsf{N}(0,\sigma^2(\theta)).$$

Thus, ARE is $m(n)/n = \tau^2(\theta)/\sigma^2(\theta)$. □

1.3 Super-efficiency

Definition 1.3 (Super-efficiency)**.** *Let $\sigma^2(\theta) = I_\theta^{-1}$, T_n and $\hat{\theta}_n$ be estimators of parameter θ such that*

$$\sqrt{n}\left(T_n - \theta\right) \xrightarrow[\theta]{d} \mathsf{N}\left(0, \tau^2(\theta)\right),$$

$$\sqrt{n}\left(\hat{\theta}_n - \theta\right) \xrightarrow[\theta]{d} \mathsf{N}\left(0, \sigma^2(\theta)\right),$$

where

$$\tau^2(\theta) \leq \sigma^2(\theta), \text{ for all } \theta \in \Theta, \quad \tau^2(\theta) > \sigma^2(\theta), \text{ for some } \theta \in \Theta.$$

Then we say that T_n is super-efficient.

We are interested in two questions:

1. Does super-efficient estimator exist?
2. If there are super-efficient estimators, are they good?

The answer to the first question is affirmative.

Example 4 (Hodge's estimator for Gaussian mean)**.** *For* $X_i \overset{\text{iid}}{\sim} \mathsf{N}(\theta, 1)$ *and* $\hat{\theta}_n = \frac{1}{n}\sum_{i=1}^n X_i$*. Let*

$$T_n := \begin{cases} \hat{\theta}_n, & \left|\hat{\theta}_n\right| \geq n^{-1/4} \\ 0, & \left|\hat{\theta}_n\right| < n^{-1/4} \end{cases}.$$

We show that T_n *is super-efficient.*

- *If* $\theta = 0$*, then*

$$P_0\left(\sqrt{n}\ T_n = 0\right) = P_0(|\hat{\theta}_n| < n^{-1/4}) = P_0(|\underbrace{\sqrt{n}\ \hat{\theta}_n}_{N(0,1)}| < n^{-1/4}) \to 1 \text{ as } n \to \infty,$$

  which gives us that $\sqrt{n}\ T_n \overset{d}{\underset{0}{\to}} 0$*.*

- *If* $\theta \neq 0$*, then*

$$\sqrt{n}\,(T_n - \theta) = \underbrace{\sqrt{n}\left(\hat{\theta}_n - \theta\right)}_{\overset{d}{\to} N(0,1)} \mathbf{1}_{\left\{|\hat{\theta}_n| \geq n^{-1/4}\right\}} + \sqrt{n}\,(0 - \theta)\,\mathbf{1}_{\left\{|\hat{\theta}_n| < n^{-1/4}\right\}} \overset{d}{\to} N(0,1),$$

  where we use $\mathbf{1}_{\left\{|\hat{\theta}_n| \geq n^{-1/4}\right\}} \to 1$ *a.s. Then we have* $\sqrt{n}\ (T_n - \theta) \overset{d}{\underset{\theta}{\to}} \mathsf{N}(0,1)$*,* $\theta \neq 0$*.*

Remark For the second question, we show in homework that the Hodge's estimator is unsatisfactory in some sense.

2 Testing and Confidence Sets

Scientific method Scientific method is of following procedures:

1. Propose a hypothesis;
2. Develop experiments;
3. Observe data that would invalidate the hypothesis;
4. Reject the hypothesis or the hypothesis remains consistent with the observed data.

Remark Scientific hypotheses are never proven "true". Prevailing hypotheses are held until they are falsified by new observations

2.1 Confidence Sets

In many situations we have

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

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

Example 5. *If $\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow[\theta_0]{d} \mathsf{N}(0, I_{\theta_0}^{-1})$, I_θ is invertible and continuous in θ. Let*

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

For $\theta = \theta_0$, let $Z \sim \mathsf{N}(0, I_{\theta_0}^{-1})$, $W \sim \mathsf{N}(0, I_d)$, we have

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

We use the continuous mapping theorem and Slutsky's theorem in the proof. Then implied by the convergence according to distribution, we have

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

By choosing a proper threshold γ, we obtain a confidence set of $\hat{\theta}_n$ with confidence level α.

2.2 Testing: Dual Problem to Confidence Sets

The typical approach of hypothesis testing is the following: can we reject some type of null hypothesis, that is, conjecture some model $\mathbb{P}_{\theta_0}$ to be "true"?

Definition 2.1 (Testing)**.** *We hope for results like*

$$\mathbb{P}_{\theta_0}(\textit{data at least as "extreme" as what we got}) \leq \alpha.$$

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 the nature of the scientific method.

Definition 2.2 (Ill-formed definition: p-values)**.** *Let $H_0 : \{\mathbb{P}_\theta : \theta \in \Theta_0\}$. The p-value associated with a sample $X_1, \ldots, X_n$ is defined to be*

$$\sup_{\theta \in \Theta_0} \mathbb{P}_\theta(\textit{data as extreme as } X_1, \ldots, X_n \textit{ observed}).$$

Example 6 (Gaussian sample mean)**.** *Let* $H_0 : X_i \overset{\text{iid}}{\sim} N(0,1)$*. The standard p-value is given by*

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

where the expectation is taken over $Z_i \overset{\text{iid}}{\sim} \mathsf{N}(0,1)$ *(and here* $\hat{\theta}_n = \frac{1}{n}\sum_{i=1}^n X_i$ *is treated as a fixed value).*

2.3 Likelihood-ratio Test

Likelihood-ratio test is the starting point for understanding "asymptotics" of testing.

2.3.1 Neyman-Pearson Test

Recall the classic Neymann-Pearson test with simple null and alternative:

$$H_0 : p_0 \;\leftrightarrow\; H_1 : p_1$$

The test that is the "best" (maximizes power at all levels) is the likelihood-ratio test. Let x be sample values and

$$T(x) = \log\left(\frac{d\mathbb{P}_1(x)}{d\mathbb{P}_0(x)}\right),$$

the most powerful test is given by

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

where $t \in \mathbb{R}$ is determined by the confidence level.

2.3.2 Generalized Likelihood-ratio Tests

We now generalize the likelihood-ratio test to composite null and alternative. Goal: to test

$$H_0 : \theta \in \Theta_0 \;\leftrightarrow\; H_1 : \theta \in \Theta,$$

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

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

The Generalized likelihood-ratio test follows similar rules in Neyman-Pearson test.

To determine the threshold t given some confidence level, we have the following proposition. Suppose that $\{P_\theta\}_{\theta\in\Theta}$ is nice enough that the MLE is asymptotically normal, and assume the Lipschitz-continuous condition of $\nabla^2\ell_{\theta_0}$

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

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

Proposition 1 (Wilk's Theorem)**.** *Let* $\Theta_0 = \{\theta_0\}$ *be a point null,* $\Theta \subset \mathbb{R}^d$*. Let*

$$L_n(x,\theta) = \sum_{i=1}^{n} \ell_\theta(x_i) = \sum_{i=1}^{n} \log p_\theta(x_i), \quad T_n(x) = L_n(x, \hat{\theta}_{MLE}) - L_n(x, \theta_0).$$

Then

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

where $X = (X_1, \ldots, X_n)$ *and* $X_i \overset{\text{iid}}{\sim} \mathbb{P}_{\theta_0}$*.*

Many thanks to the great jobs of the 300B scribers last winter!