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

Lecture 18 – March 8

Lecturer: John Duchi *Scribe: Fan Zhang*

Warning: *these notes may contain factual errors*

Reading:

1. Erich Leo Lehmann, *Testing Statistical Hypotheses*, Chapter 12.3

2. Van der Vaart, *Asymptotic Statistics*, Chapter 6.

Outline:

1. Absolute continuity of measures.

2. Contiguity and Asymptotics; Le Cam's Lemmas.

3. Distance for distributions.

1 Recapitulation

For asymptotic test, we want to understand efficiency / sample size requirement. Suppose that $\theta_n \to 0$, and T_n is a sequence of tests that satisfies

$$\sqrt{n}\left(\frac{T_n - \mu(\theta_n)}{\sigma(\theta_n)}\right) \xrightarrow{d} \mathsf{N}(0, 1).$$

We have null hypothesis $H_0 : \theta = 0$ versus alternative hypothesis $H_1 : \theta > 0$. The null hypothesis H_0 is rejected for large $T_n - \mu(0)$. The performance of the test is governed by the slope $\frac{\mu'(0)}{\sigma(0)}$.

Today we are going to discuss how to get limit for sequence of distributions Q_n based on nearby distributions P_n, when P_n and Q_n are "getting close".

2 Absolute Continuity

We begin with the notion that allows us to change measures in a non-asymptotic setting.

Definition 2.1. *Probability measure Q is* absolutely continuous *with respect to probability measure P, written $Q \ll P$, if $P(A) = 0$ implies $Q(A) = 0$ for any set A.*

If $Q \ll P$, then the Radon-Nikodym theorem says that there exists a nonnegative measurable function g, denoted $\frac{dQ}{dP}$, such that $\mathbb{E}_Q[f] = \mathbb{E}_P[fg] = \int fgdP = \int f\frac{dQ}{dP}dP$ for all f integrable with respect to Q. Thus, given likelihood ratio $g = \frac{dQ}{dP}$ and distribution P, we know everything about Q.

Note: Let $p = \frac{dP}{d\mu}$, $q = \frac{dQ}{d\mu}$ for a dominating distribution μ such that $\mu \gg P$, $\mu \gg Q$, then $Q \ll P$ $\Leftrightarrow Q(p > 0) = 0 \Leftrightarrow \int (\frac{q}{p}) dP = \int (\frac{q}{p}) \mathbf{1}_{\{p>0\}} dP = 1$.

The following corollary is another formulation of some of these ideas.

Corollary 1. *Let M be the joint measure (law) of the pair $(X, L) := \left(X, \frac{dQ}{dP}\right)$ under distribution P. (So M is defined on $\mathcal{X} \times \mathbb{R}_+$). Then $L \geq 0$, $\mathbb{E}_M[L] = 1$, and $Q(B) = \mathbb{E}_P\left[\mathbf{1}_B(X)\frac{dQ}{dP}\right] = \mathbb{E}_P[\mathbf{1}_B(X)L] = \int_{B\times\mathbb{R}_+} r dM(x, r)$. Moreover, $\mathbb{E}_Q[f(X, L)] = \int_{\mathcal{X}\times\mathbb{R}_+} f(x, r) r dM(x, r)$ for all bounded measurable function f, i.e. (X, L) has density $r dM(x, r)$ under distribution Q.*

Idea: If we know law of X and L under P, then we can construct law of X under Q. Can we do this in an asymptotic sense? We would like to find an asymptotic version of Corollary 1. This is going to allow us to transfer power of calculations under a sequence P_n to an alternative Q_n, if Q_n is "asymptotically absolutely continuous" with respect to P_n.

3 Contiguity and Asymptotics; Le Cam's Lemmas.

The notion that will allow to perform the asymptotic versions of the previous calculations is contiguity.

Definition 3.1. *A sequence $\{Q_n\}$ of distributions is* contiguous *with respect to $\{P_n\}$, written $Q_n \lhd P_n$, if $P_n(A_n) \to 0$ implies $Q_n(A_n) \to 0$ for any sequence of sets A_n. Sequences $\{Q_n\}$ and $\{P_n\}$ are* mutually contiguous*, written $Q_n \lhd\rhd P_n$, if $Q_n \lhd P_n$ and $P_n \lhd Q_n$.*

Below we will characterize contiguity with conditions on the limits of density representations of P_n and Q_n. Because P_n and Q_n need not be absolutely continuous with respect to each other, nor are we a priori provided some mutually dominating measure, the following observations are useful.

Observation 2. *Suppose that there exist a dominating distribution μ s.t. $p_n = \frac{dP_n}{d\mu}$ and $q_n = \frac{dQ_n}{d\mu}$. Let $L_n = \frac{dQ_n}{dP_n}$, then we always have $\mathbb{E}_{P_n}[L_n] = \int \frac{q_n}{p_n} dP_n = \int \frac{q_n}{p_n} \mathbf{1}_{\{p_n>0\}} dP_n = Q_n(p_n > 0) \leq 1$. Thus, under P_n, the sequence L_n is tight, which implies that L_n always has a weakly convergent subsequence.*

We are now ready to state alternative characterizations of contiguity.

Lemma 3 (Le Cam's First Lemma, or "Limits determine contiguity")**.** *The following are equivalent:*

1. $Q_n \lhd P_n$.
2. *If* $L_n^{-1} \xrightarrow[Q_n]{d} U$ *along a subsequence, then* $\mathbb{P}(U > 0) = 1$.
3. *If* $L_n \xrightarrow[P_n]{d} L$ *along a subsequence, then* $\mathbb{E}[L] = 1$.
4. $T_n \xrightarrow{P_n} 0$ *implies* $T_n \xrightarrow{Q_n} 0$.

Proof Idea *4⇒1*: Take A_n such that $P_n(A_n) \to 0$. Define $T_n = \mathbf{1}_{A_n}$. Then certainly $T_n \overset{P_n}{\to} 0$, so $T_n \overset{Q_n}{\to} 0$. That is, $Q_n(A_n) \to 0$.

The second claim is basically saying that limit $\mathbb{E}_{P_n}[L_n] = 1$, so everything with $dP_n = 0$ satisfies $dQ_n = 0$.

The third claim says that $(\frac{p_n}{q_n})q_n > 0$ is eventually going to happen, i.e. p_n can't be 0 if $q_n > 0$ for large n. $\square$

The proof can be found in Van der Vaart Chapter 6.

The following two examples will be useful in later developments.

Example 1 (Asymptotic log normality)**:** Suppose we have $\log \frac{dP_n}{dQ_n} \xrightarrow[Q_n]{d} \mathsf{N}(\mu, \sigma^2)$. Then, by the continuous mapping theorem, we have $\frac{dP_n}{dQ_n} \xrightarrow[Q_n]{d} \exp(\mathsf{N}(\mu, \sigma^2))$ is greater than 0 with probability 1. Applying the second characterization of LeCam's First Lemma implies $Q_n \triangleleft P_n$. On the other hand, based on our knowledge of the moment generating function for normal random variables, $\mathbb{E}[\exp(\mathsf{N}(\mu, \sigma^2)] = \exp\left(\mu + \frac{1}{2}\sigma^2\right)$. Applying the third characterization of LeCam's First Lemma gives that $P_n \triangleleft Q_n$ if and only if $\exp\left(\mu + \frac{1}{2}\sigma^2\right) = 1$. That is, $\mu = -\frac{1}{2}\sigma^2$. ♣

Example 2 (Smooth likelihoods under local alternatives)**:** Suppose $\{P_\theta\}_{\theta\in\Theta}$ has densities p_θ, and p_θ is smooth enough in θ that $\log p_\theta$ has a Taylor expansion around $\theta_0 \in \operatorname{int}\Theta$. That is,

$$\log p_{\theta_0+h} = \ell_{\theta_0} + h^T\nabla\ell_{\theta_0} + \frac{1}{2}h^T\nabla^2\ell_{\theta_0}h + O(\|h\|^3)$$

where $\ell_\theta = \log p_\theta$.

For $X_1, \dots, X_n \overset{i.i.d.}{\sim} P_{\theta_0}$, a Taylor expansion gives

$$\begin{aligned}
\log \frac{dP_{\theta_0+h/\sqrt{n}}}{dP_{\theta_0}}(X_1, \cdots, X_n) &= \sum_{i=1}^n \left(\ell_{\theta_0+h/\sqrt{n}}(X_i) - \ell_{\theta_0}(X_i)\right) \\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^n h^T\nabla\ell_{\theta_0}(X_i) + \frac{1}{2n}h^T\sum_{i=1}^n\nabla^2\ell_{\theta_0}(X_i)h + o_{P_{\theta_0}}(1) \\
&= \underbrace{\frac{1}{\sqrt{n}}\sum_{i=1}^n h^T\nabla\ell_{\theta_0}(X_i)}_{\xrightarrow{d}\mathsf{N}(0,h^TI_\theta h)} - \frac{1}{2}h^TI_{\theta_0}h + o_{P_{\theta_0}}(1) \\
&\xrightarrow{d} \mathsf{N}\left(-\frac{1}{2}h^TI_{\theta_0}h, h^TI_{\theta_0}h\right).
\end{aligned}$$

Surprisingly, this is exactly the condition for mutual contiguity from Example 1! Thus $P^n_{\theta_0} \triangleleft\triangleright P^n_{\theta_0+h/\sqrt{n}}$. Local alternatives and the null are mutually contiguous! ♣

In analogy with Corollary 1 and in light of the previous example, we now wonder whether mutual contiguity tells us something about limit distributions under alternatives. The next theorem is the beginning of an answer.

Theorem 4 (Le Cam)**.** *Let P_n, Q_n be distributions on $X_n \in \mathcal{X}$, $L_n = \frac{dQ_n}{dP_n}$. If $Q_n \triangleleft P_n$ and $(X_n, L_n) \xrightarrow[P_n]{d} (X, L)$ with joint measure M on $\mathcal{X} \times \mathbb{R}_+$, then $(X_n, L_n) \xrightarrow[Q_n]{d} W$, where W is a distribution with density $rdM(x, r)$ on $\mathcal{X} \times \mathbb{R}_+$.*

Written differently, if $f : \mathcal{X} \times \mathbb{R}_+ \mapsto \mathbb{R}$ is a bounded continuous function and $\mathbb{E}_{P_n}[f(X_n, L_n)] \longrightarrow \int f(x, r)dM(x, r)$, then $\mathbb{E}_{Q_n}[f(X_n, L_n)] \longrightarrow \int f(x, r)dM(x, r)$.

If we define $Q(B) := \mathbb{E}_M[\mathbf{1}_B(X)L]$ then Q is a probability measure, and $X_n \xrightarrow[Q_n]{d} Z$ where $Z \sim Q$.

Proof (See Van der Vaart Chapter 6). □

The following example is an important application of Theorem 4.

Lemma 5 (Le Cam's Third Lemma)**.** *If* $\left(X_n, \log \frac{dQ_n}{dP_n}\right) \xrightarrow[P_n]{d} \mathsf{N}\left(\begin{pmatrix} \mu \\ -\frac{1}{2}\sigma^2 \end{pmatrix}, \begin{pmatrix} \Sigma & \tau \\ \tau^T & \sigma^2 \end{pmatrix}\right)$, *then* $X_n \xrightarrow[Q_n]{d} \mathsf{N}(\mu + \tau, \Sigma)$.

Sketch of Proof An equivalent result to the Theorem 4 is that

$$\mathbb{E}_{P_n}[f(X_n, \log L_n)] \longrightarrow \int f(x,r)dM(x,r)$$

implies

$$\mathbb{E}_{Q_n}[f(X_n, \log L_n)] \longrightarrow \int f(x,r)e^r dM(x,r).$$

Let $(\tilde{X}, \tilde{Z})$ has density $e^r dM(x,r)$, then for $\lambda \in \mathbb{R}_+^d$ and $i = \sqrt{-1}$,

$$\begin{aligned}
\mathbb{E}[\exp(i\lambda^T \tilde{X})] &= \int e^{i\lambda^T x} e^r dM(x,r) \\
&= \mathbb{E}[\exp\left(i\lambda^T X + Z\right)] \qquad\qquad (X,Z) \sim M \\
&= \text{characteristic function of } (X,Z) \text{ at } t = \begin{pmatrix} \lambda \\ -i \end{pmatrix} \\
&= \exp\left(i(\mu+\tau)^T\lambda - \frac{1}{2}\lambda^T \Sigma \lambda\right) \\
&= \text{characteristic function of } \mathsf{N}(\mu+\tau, \Sigma),
\end{aligned}$$

which completes the proof. □

Example 3: Suppose usual parametric situation and that

$$T_n = -\frac{1}{\sqrt{n}} \sum_{i=1}^n I_{\theta_0}^{-1} \dot{\ell}_{\theta_0}(X_i) + o_{P_{\theta_0}}(1).$$

Under alternative distribution $P_{\theta_0 + h/\sqrt{n}}$, consider joint distribution

$$\begin{aligned}
&\left(T_n, \log\left(\frac{dP^n_{\theta_0+\frac{h}{\sqrt{n}}}}{dP^n_{\theta_0}}\right)\right) \\
&= \left(-\frac{1}{\sqrt{n}} \sum_{i=1}^n I_{\theta_0}^{-1} \dot{\ell}_{\theta_0}(X_i) + o_{P_{\theta_0}}(1), h^T\left(\frac{-1}{\sqrt{n}} \sum_{i=1}^n \dot{\ell}_{\theta_0}(X_i)\right) - \frac{1}{2} h^T I_{\theta_0} h + o_{P_{\theta_0}}(1)\right) \\
&\xrightarrow[P_{\theta_0}]{d} \mathsf{N}\left(\begin{pmatrix} 0 \\ -\frac{1}{2}h^T I_{\theta_0} h \end{pmatrix}, \begin{pmatrix} I_{\theta_0} & h \\ h^T & h^T I_{\theta_0} h \end{pmatrix}\right),
\end{aligned}$$

where we use

$$\mathrm{Cov}_{P_{\theta_0}}\left(\frac{1}{\sqrt{n}} \sum_{i=1}^n \dot{\ell}_{\theta_0}(X_i)^T h, \frac{1}{\sqrt{n}} \sum_{i=1}^n I_{\theta_0}^{-1} \dot{\ell}_{\theta_0}(X_i)\right) = h$$

to compute the covariance matrix. Therefore, by Le Cam's Third Lemma,

$$T_n \underset{P_{\theta_0+\frac{h}{\sqrt{n}}}}{\xrightarrow{d}} \mathsf{N}\left(h, I_{\theta_0}^{-1}\right).$$

Consequently, if you have estimator $\hat{\theta}_n$ with typical expansion, i.e.

$$\sqrt{n}\left(\hat{\theta}_n - \theta_0\right) = -\frac{1}{\sqrt{n}}\sum_{i=1}^{n} I_{\theta_0}^{-1}\dot{\ell}_{\theta_0}(X_i) + o_{P_{\theta_0}}(1),$$

then

$$\sqrt{n}\left(\hat{\theta}_n - \left(\theta_0 + \frac{h}{\sqrt{n}}\right)\right) \underset{P_{\theta_0+\frac{h}{\sqrt{n}}}}{\xrightarrow{d}} \mathsf{N}\left(0, I_{\theta_0}^{-1}\right).$$

♣