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

# Lecture 6 – January 25

*Lecturer: John Duchi* *Scribe: Matthew Tyler*

***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:**

- Finish Basics of Hypothesis Testing
- Likelihood Ratio Tests
- Wald Tests
- Rao/Score Tests

## 1 Asymptotics of Tests

**Goal:** Understand asymptotics of tests. Let $T_n$ be a sequence of tests, meaning $T_n = (X_1, \ldots, X_n)$ and $T_n$ either rejects or does not reject [the null hypothesis].

**Definition 1.1.** *For a sequence of tests $T_n$, the uniform asymptotic level of $T_n$ for null hypothesis $H_0 : \theta \in \Theta_0$ is*

$$\limsup_{n\to\infty} \sup_{\theta\in\Theta_0} P_\theta(T_n \text{ rejects}).$$

*The pointwise asymptotic level of $T_n$ is*

$$\sup_{\theta\in\Theta_0} \limsup_{n\to\infty} P_\theta(T_n \text{ rejects}).$$

**Remark** The pointwise asymptotic level of $T_n$ is never more than the uniform asymptotic level of $T_n$. However, we will usually only concern ourselves with the pointwise asymptotic level.

## 2 Generalized Likelihood Ratio Tests

**Goal:** Test $H_0 : \theta \in \Theta_0$ versus $H_1 : \theta \in \Theta$, assuming $\Theta_0 \subset \Theta$.

We make use of the following test statistic:

$$T(x) := \log \frac{\sup_{\theta\in\Theta} p_\theta(x)}{\sup_{\theta\in\Theta_0} p_\theta(x)}.$$

**Proposition 1** (Wilk's, simplified)**.** *Let* $\Theta_0 = \{\theta_0\}, \Theta \subseteq \mathbb{R}^d$ *be open. Let* $L_n(X;\theta) = \sum_{i=1}^n \ell_\theta(X_i) = \sum_{i=1}^n p_\theta(X_i)$*. Define* $\Delta_n := L_n(X;\hat{\theta}_n) - L_n(X;\theta_0) = T(X)$*. Then under the typical conditions for asymptotic efficiency of the MLE,*

$$2\Delta_n \xrightarrow[H_0]{d} \chi^2_d.$$

*Note* $\chi^2_d \overset{dist}{=} \|w\|_2^2$ *where* $w \sim \mathcal{N}(0, I_{d\times d})$*.*

**Proof** Under $H_0$, $\hat{\theta}_n \xrightarrow{p} \theta_0$. For large enough $n$,

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

where $\mathrm{Err}_{(i)} = O_p(||\hat{\theta}_n - \theta_0||)$. This was a Taylor approximation of the gradient of $L_n$. In addition, we take a second-order Taylor approximation of $L_n$:

$$L_n(X;\hat{\theta}_n) = L_n(X;\theta_0) + \nabla L_n(X;\theta_0)(\hat{\theta}_n - \theta_0) + \frac{1}{2}(\hat{\theta}_n - \theta_0)^T \nabla^2 L_n(X;\theta_0)(\hat{\theta}_n - \theta_0) + o_p(||\hat{\theta_n} - \theta_0||).$$

After substituting the first equation into the second,

$$\begin{aligned}\Delta_n &= L_n(X;\hat{\theta}_n) - L_n(X;\theta_0) \\ &= -\frac{1}{2}(\hat{\theta}_n - \theta_0)^T \nabla^2 L_n(X;\theta_0)(\hat{\theta}_n - \theta_0) + \sum_{i=1}^n (\hat{\theta}_n - \theta_0)\mathrm{Err}_{(i)}(\hat{\theta}_n - \theta_0) + o_p(1).\end{aligned}$$

Now let $w_n = \sqrt{n}(\hat{\theta}_n - \theta_0)$, so $w_n \xrightarrow[H_0]{d} \mathcal{N}(0, I_{\theta_0}^{-1})$. With this new notation,

$$\begin{aligned}\Delta_n &= -\frac{1}{2} w_n^T \underbrace{\left(\frac{1}{n}\nabla^2 L_n(X;\theta_0)\right)}_{\xrightarrow{p} -I_{\theta_0}} w_n + w_n^T \underbrace{\left(\frac{1}{n}\sum_{i=1}^n \mathrm{Err}_{(i)}\right)}_{\xrightarrow{p} 0} w_n + o_p(1) \\ &= \frac{1}{2} w_n^T I_{\theta_0} w_n + o_p(1) \xrightarrow{d} \frac{1}{2}\chi^2_d.\end{aligned}$$

Thus $2\Delta_n \xrightarrow{d} \chi^2_d$. □

**Remark**

- Could use likelihood ratio test for testing $H_0 : \theta = \theta_0$, but may require substantial computation; e.g., to get the MLE.
- Can we use simpler tests to get the same asymptotic $\chi^2$ behavior?
- Note that everything is quadratic. Let's just start with quadratics instead!

# 3 Wald Tests

**Definition 3.1.** *A Wald confidence ellipse is*

$$C_{n,\alpha} = \{\theta \in \mathbb{R}^d : (\theta - \hat{\theta}_n)^T I_{\hat{\theta}_n}(\theta - \hat{\theta}_n) \leq U^2_{d,\alpha}/n\}$$

*where $U^2_{d,\alpha}$ is [uniquely] determined by $\mathbb{P}(\chi^2_d \geq U^2_{d,\alpha}) = \alpha$).*

**Remark** We showed last lecture that If $\hat{\theta}_n$ is an efficient estimator for $\theta_0$ then $n(\hat{\theta}_n - \theta_0) I_{\hat{\theta}_n}(\hat{\theta}_n - \theta_0) \xrightarrow[P_\theta]{d} \chi^2_d \overset{\text{dist}}{=} \|w\|_2^2, w \sim \mathcal{N}(0, I_{d\times d})$.

**Definition 3.2.** *A Wald test of point null $\theta = \theta_0$ (against $\theta \neq \theta_0$) is*

$$\begin{aligned} T_n(X) : &= \begin{cases} \textit{Reject} & \textit{if } \theta_0 \notin C_{n,\alpha} \\ \textit{Don't Reject} & \textit{otherwise} \end{cases} \\ &= \textit{Reject if } (\theta_0 - \hat{\theta}_n)^T I_{\hat{\theta}_n}(\theta_0 - \hat{\theta}_n) > U^2_{d,\alpha}/n. \end{aligned}$$

**Proposition 2.** *For testing $H_0 : \theta = \theta_0$, a Wald test is asymptotically level $\alpha$.*

**Proof** Immediate from earlier results. □

**Remark**

- For the Fisher Information, we can replace $I_{\hat{\theta}_n}$ with $I_{\theta_0}$ and the asymptotic level is the same.
- Works with any efficient estimator — not just the MLE.
- One weakness is that likelihood ratio and Wald tests can only handle point nulls. What if we have nuisance parameters?

**Example 1:** $X_i \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$. $H_0 = \{\mu = 0, \overbrace{\sigma^2 \geq 0}^{\text{"nuisance parameter"}}\}$. None of the results we have gathered so far apply in this case. ♣

Let us now consider smooth problems with $I(\theta) \in \mathbb{R}^{d\times d}$. Define $\Sigma(\theta) := I(\theta)^{-1}$. Assume efficient estimators $\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow[P_\theta]{d} \mathcal{N}(0, \Sigma(\theta)$. We will consider the case where we only care about estimating functions of $\theta$, usually just certain coordinates. Define

$$[v]_{1:k} = \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_k \end{bmatrix}.$$

That is, just the first $k$ coordinates of $v \in \mathbb{R}^d$, $k \leq d$.

Similarly, define $\Sigma^{(k)} \in \mathbb{R}^{k\times k}$ as the leading principal minor (of order $k$). Specifically,

$$\Sigma = \begin{bmatrix} \Sigma^{(k)} & \cdots \\ \vdots & \ddots \end{bmatrix}.$$

Then automatically due to the properties of the multivariate normal,

$$\sqrt{n}([\hat{\theta}_n]_{1:k} - [\theta_0]_{1:k}) \xrightarrow{d} \mathcal{N}(0, \Sigma^{(k)}(\theta_0)).$$

Note that $\Sigma^{(k)}(\theta)$ acts as the inverse Fisher Information for the first $k$ coordinates.

**Lemma 3** (Schur Complement)**.** *Suppose*

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}, \quad A = A^T, \quad A \succ 0.$$

*If $M = A^{-1}$, then $M_{11} = \left(A_{11} - A_{12}A_{22}^{-1}A_{21}\right)^{-1}$.*

When $\hat{\theta}_n$ is efficient for $\theta$, then

$$n([\hat{\theta}_n]_{1:k} - [\theta_0]_{1:k})^T \left[\Sigma^{(k)}(\hat{\theta}_n)\right]^{-1} ([\hat{\theta}_n]_{1:k} - [\theta_0]_{1:k}) \xrightarrow{d} \chi^2_k,$$

where

$$\left[\Sigma^{(k)}(\hat{\theta}_n)\right]^{-1} = I_{11}(\hat{\theta}_n) - I_{12}(\hat{\theta}_n)I_{22}(\hat{\theta}_n)^{-1}I_{21}(\hat{\theta}_n).$$

Now we can design a Wald-type test of these composite nulls with nuisance parameters.

**Definition 3.3** (Wald Test, Composite)**.** *Let $H_0 : \{\theta \in \mathbb{R}^d : [\theta]_{1:k} = [\theta_0]_{1:k}\}$. Define the acceptance region as*

$$C_{n,\alpha} = \left\{\theta \in \mathbb{R}^d : ([\theta]_{1:k} - [\theta_0]_{1:k})^T \left[\Sigma^{(k)}(\theta_0)\right]^{-1} ([\theta]_{1:k} - [\theta_0]_{1:k}) \leq U^2_{k,n}/n\right\}$$

*where $U^d_{k,n}$ is [uniquely] determined by $\mathbb{P}(\chi^2_k \geq U^2_{k,n}) = \alpha$. The Wald test for composite nulls is given by*

$$T_n := \begin{cases} \textit{Reject} & \textit{if } \hat{\theta_n} \notin C_{n,\alpha} \\ \textit{Don't Reject} & \textit{otherwise} \end{cases}.$$

**Proposition 4.** *If $\hat{\theta}_n$ is efficient for $\theta$ in model $\{P_\theta\}_{\theta\in\Theta}$, then $T_n$ is pointwise asymptotic level $\alpha$. That is,*

$$\sup_{\theta\in\Theta_0} \limsup_{n\to\infty} P_\theta(T_n \textit{ rejects}) = \alpha.$$

**Remark**

- Cannot substitute $\theta_0$ for $\hat{\theta}_n$ in $I_{\hat{\theta}_n}$ because we must estimate the nuisance parameters.
- In terms of convergence in distribution, it is not necessary to estimate the nuisance parameters efficiently — since they only appear in the estimated Fisher Information. An estimator which is only consistent is sufficient to apply Slutsky's lemma and get the desired asymptotic testing level.

**Example 2:** Let $H_0 = \{\mathcal{N}(\mu, \Sigma) : \Sigma \succ 0, \theta = 0\}$. Then

$$C_{n,\alpha} = \{\theta \in \mathbb{R}^d : \theta^T\hat{\Sigma}^{-1}\theta \leq U^2_{d,\alpha}/n\}$$

where $\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X}_n)(X_i - \bar{X}_n)^T$. Then $P_{\theta,\Sigma}(\bar{X}_n \in C_{n,\alpha}) \to \alpha$ as $n \to \infty$. The estimate of $\hat{\Sigma}$ only needs to be consistent. ♣

# 4 Rao/Score Tests

**Goal:** If the MLE is difficult to compute, we can still do asymptotic testing.

We know the asymptotics of $P_n \nabla \ell_\theta = \frac{1}{n} \sum_{i=1}^n \nabla \ell_\theta(X_i)$ under $P_\theta$. That is,

$$\sqrt{n} \left( P_n \nabla \ell_\theta \right) \xrightarrow[P_\theta]{d} \mathcal{N}(0, I_\theta).$$

Thus, for point nulls $H_0 : \theta = \theta_0$,

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

Naturally, the Rao test rejection rule

$$n \left( P_n \nabla \ell_{\theta_0} \right)^T I_{\theta_0}^{-1} \left( P_n \nabla \ell_{\theta_0} \right) > U^2_{d,\alpha}$$

has asymptotic level $\alpha$.

**Remark**

- There exist strong connections between good tests and notions of optimality in estimation. We will explore this more later.
- Analogues of Rao and Generalized Likelihood Ratio tests exist for nuisance parameters and composite nulls, but they are similar to the extension of the Wald test and so will not be covered here.