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

Outline:

• Testing (continued)
  – Likelihood Ratio Tests (a.k.a. Wilks tests)
  – Wald Tests

1 Introduction

The p-value is a probability under the null of observing data ”at least as extreme” as what you actually saw.

For a given level $\alpha$, we find a confidence set $C_{n,\alpha}$ such that $P_{H_0}(X_1, \ldots, X_n \in C_{n,\alpha}) \geq 1 - \alpha$. If $X_1, \ldots, X_n \notin C_{n,\alpha}$, we reject the null. In general, any set $C_n$ such that we can compute $P_{H_0}(X_1, \ldots, X_n \in C_n)$ can function as a confidence set.

Example 1: To test $H_0 : X_i \overset{iid}{\sim} P_0 = \mathcal{N}(0, 1)$. The ”natural” p-value is $P_{0}(\bar{Z} \geq |\hat{\theta}|)$, where $\hat{\theta} = \frac{1}{n} \sum_{i=1}^{n} X_i$, and $\bar{Z} = \frac{1}{n} \sum_{i=1}^{n} Z_i$ for $Z_i \overset{iid}{\sim} P_0$.

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:

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

and we reject the null if $T(x)$ is big (which indicates that $\Theta$ is much more likely than $\Theta_0$).
Proposition 1 (Wilks’, simplified). Let $\Theta_0 = \{\theta_0\}, \Theta \subseteq \mathbb{R}^d$ be open. Let $L_n(X; \theta) = \sum_{i=1}^n \ell_{\theta}(X_i) = \sum_{i=1}^n \log p_{\theta}(X_i)$. Define $\Delta_n := L_n(X; \hat{\theta}_n) - L_n(X; \theta_0) = T(X)$, where $\hat{\theta}_n := \arg\max_{\theta \in \Theta} L_n(X; \theta)$. Then under typical smoothness conditions (such as consistency and asymptotic normality) of the MLE,

$$2\Delta_n \overset{d}{\rightarrow} \chi^2_d.$$ 

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

Hence we obtain confidence regions for level $\alpha$ tests:

Reject if $T(X) = \Delta_n \geq u_{d,\alpha}$, where $P(\chi^2_d \geq 2u_{d,\alpha}) \leq \alpha$.

Proof Under $H_0$, $\hat{\theta}_n \overset{P}{\rightarrow} \theta_0$. For large enough $n$,

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

where $\text{Err}(i) = O_p(||\hat{\theta}_n - \theta_0||)$. 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,

$$\Delta_n = L_n(X; \hat{\theta}_n) - L_n(X; \theta_0)$$

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

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

$$\Delta_n = -\frac{1}{2}w_n^T \left( \frac{1}{n} \nabla^2 L_n(X; \theta_0) \right) w_n + w_n^T \left( \frac{1}{n} \sum_{i=1}^n \text{Err}(i) \right) w_n + o_p(1)$$

$$= \frac{1}{2}w_n^T I_{\theta_0} w_n + o_p(1) \overset{d}{\rightarrow} \frac{1}{2} \chi^2_d.$$ 

Thus $2\Delta_n \overset{d}{\rightarrow} \chi^2_d$. \qed

Remark

- Could use likelihood ratio test for testing $H_0 : \theta = \theta_0$, but may require substantial computation; e.g., to get the MLE under $H_0$.

- 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 - Wald tests do this.
3 Wald Tests

Definition 3.1. A Wald confidence ellipse is

\[ C_{n,r} = \{ \theta \in \mathbb{R}^d : (\theta - \hat{\theta}_n)^T I_{\hat{\theta}_n} (\theta - \hat{\theta}_n) \leq r/n \} \]

where \( \hat{\theta}_n \) is the estimator of \( \theta \).

Remark We have shown that for a point null \( H_0 : \{ P_{\theta_0} \} \) we have

\[ n(\hat{\theta}_n - \theta_0) I_{\theta_0} (\hat{\theta}_n - \theta_0) \xrightarrow{d} \chi^2_d \text{dist} = \| w \|_2^2, w \sim N(0, I_{d \times d}). \]

Definition 3.2. A Wald test of point null \( \theta = \theta_0 \) (against \( \theta \neq \theta_0 \)) is constructed as follows: Let

\[ C_{n,\alpha} = \{ \theta \in \mathbb{R}^d : (\theta - \theta_0)^T I_{\hat{\theta}_n} (\theta - \theta_0) \leq u^2_{d,\alpha}/n \} \]

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

\[ T_n(X) : = \begin{cases} \text{Reject} & \text{if } \hat{\theta}_n \notin C_{n,\alpha} \\ \text{Don't Reject} & \text{otherwise} \end{cases} \]

\[ = \text{Reject iff } (\theta_0 - \hat{\theta}_n)^T I_{\hat{\theta}_n} (\theta_0 - \hat{\theta}_n) > u^2_{d,\alpha}/n. \]

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

Proof Immediate from earlier results. \( \square \)

Remark

- For the Fisher Information, we can replace \( I_{\hat{\theta}_n} \) with \( I_{\theta_0} \) and the asymptotic level is the same.

- One weakness is that likelihood ratio and Wald tests can only handle point nulls. What if we have a composite null, e.g. if we have nuisance parameters?

Example 2: \( X_i \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2) \). \( H_0 = \{ \mu = 0, \quad \sigma^2 \geq 0 \} \). 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 the MLE \( \sqrt{n}(\hat{\theta}_n - \theta_0) \overset{d}{\rightarrow} \mathcal{N}(0, \Sigma(\theta)) \). 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,

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

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 $M_{11} = (A_{11} - A_{12}A_{22}^{-1}A_{21})^{-1}$.

When $\hat{\theta}_n$ is the MLE of $\theta$, 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 $H_0 : \{ \theta \in \mathbb{R}^d : [\theta]_{1:k} = [\theta_0]_{1:k}, \theta_{k+1}, \ldots, \theta_d \text{ unspecified} \}$. Define the acceptance region as

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

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

$$T_n := \begin{cases} 
\text{Reject} & \text{if } \hat{\theta}_n \notin C_{n,\alpha} \\
\text{Don’t Reject} & \text{otherwise} 
\end{cases}.$$

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,

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

Remark

- Cannot substitute $\theta_0$ for $\hat{\theta}_n$ in $I_{\hat{\theta}_n}$ because we must estimate the nuisance parameters.