(Relative) Efficiency of Estimators and Basic Tests using Fisher Information

John Duchi

Stats 300b – Winter Quarter 2021

Outline

- Efficiency of estimators
- Super-efficiency
- Some basic tests
- Confidence intervals
- Likelihood Ratios
- Standard tests: likelihood ratio, Wald, and Rao's Score

Reading: The references here are somewhat redundant to one another, but their union is more than sufficient:

1. Lehmann, *Elements of Large Sample Theory* Chs. 3.1, 3.2, 4.1.
2. Lehmann & Romano, *Testing Statistical Hypotheses* Ch. 12.4.
3. van der Vaart, *Asymptotic Statistics* Chs. 8.1, 8.2, 14.1–14.3.

Efficiency of estimators

Definition

We say an estimator $\widehat{\theta}_n$ is *efficient* for a parameter θ in the model $\{P_\theta\}$ with Fisher information I_θ if

$$\sqrt{n}(\widehat{\theta}_n - \theta) \stackrel{d}{\rightarrow} \mathcal{N}(0, I_\theta^{-1}).$$

Examples:

- Gaussian mean
- Poisson parameter estimation
- Regular exponential family MLEs.

Comparing estimators

Let $\widehat{\theta}_n$ and T_n be (sequences) of estimators of a parameter $\theta \in \mathbb{R}$, where we have

$$\sqrt{n}(\widehat{\theta}_n - \theta) \xrightarrow{d} \mathcal{N}(0, \sigma^2(\theta)).$$

Definition

If there is a sequence $m(n) \to \infty$ such that

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

then the limit (assuming it exists)

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

is the *asymptotic relative efficiency* (ARE) of $\widehat{\theta}_n$ to T_n.

Idea: relative sample size estimators require to get an estimate of the same "quality"

Confidence intervals

Constructing an interval

- Asymptotically normal estimate $\widehat{\theta}_n$, $\sqrt{n}(\widehat{\theta}_n - \theta) \xrightarrow{d} \mathcal{N}(0, \sigma^2(\theta))$
- Gaussian $1 - \alpha/2$ quantile $\mathbb{P}(|Z| \geq z_{1-\alpha/2}) = \alpha$
- Natural (Wald) confidence interval

$$C_n := \left[\widehat{\theta}_n - z_{1-\alpha/2}\sqrt{\frac{\sigma^2(\theta)}{n}}, \widehat{\theta}_n + z_{1-\alpha/2}\sqrt{\frac{\sigma^2(\theta)}{n}}\right]$$

satisfies $\lim_{n\to\infty} P_\theta(\theta \in C_n) = 1 - \alpha$

Comparing inverals: If ARE of $\widehat{\theta}_n$ to T_n is $A \in (0, \infty)$, when are intervals the same size?

From variance to relative efficiency

Lemma

If

$$\sqrt{n}(\widehat{\theta}_n - \theta) \xrightarrow{d} \mathcal{N}(0, \sigma^2(\theta)) \quad \textit{and} \quad \sqrt{n}(T_n - \theta) \xrightarrow{d} \mathcal{N}(0, \tau^2(\theta))$$

then the asymptotic relative efficiency (ARE) of $\widehat{\theta}_n$ *w.r.t.* T_n *is* $\frac{\tau^2(\theta)}{\sigma^2(\theta)}$

Super-efficiency and comparison of estimators

Food for thought: say estimators T_n, $\widehat{\theta}_n$ satisfy

$$\sqrt{n}(T_n - \theta) \xrightarrow{d} \mathcal{N}(0, \tau^2(\theta)) \quad \text{and} \quad \sqrt{n}(\widehat{\theta}_n - \theta) \xrightarrow{d} \mathcal{N}(0, \sigma^2(\theta))$$

where $\tau^2(\theta) \le \sigma^2(\theta)$ everywhere, and $\tau^2(\theta_0) < \sigma^2(\theta_0)$ for some θ_0

Definition

If the preceding occurs and $\sigma^2(\theta) = I_\theta^{-1}$, T_n is *super-efficient*.

Hodge's super-efficient estimator

Assume $X_i \overset{\text{iid}}{\sim} \mathcal{N}(\theta, 1)$ and define

$$T_n := \begin{cases} \overline{X}_n & \text{if } |\overline{X}_n| \geq n^{-1/4} \\ 0 & \text{otherwise} \end{cases}$$

Lemma

Hodge's estimator is super-efficient, as

$$\sqrt{n}(T_n - \theta) \xrightarrow[P_\theta]{d} \begin{cases} \mathcal{N}(0,1) & \textit{if } \theta \neq 0 \\ 0 & \textit{if } \theta = 0. \end{cases}$$

Testing

The scientific method: We propose a hypothesis, develop an experiment to test the hypothesis, and then either (i) reject the hypothesis or state that (ii) the hypothesis remains consistent with the data. (There is no truth.)

Strong inference consists of applying the following steps to every problem in science, formally and explicitly and regularly:

1. Devising alternative hypotheses
2. Devising a crucial experiment (or several), with alternative possible outcomes, each of which will, as nearly as possible, exclude one or more of the hypotheses;
3. Carrying out the experiment so as to get a clean result

John Platt, "Strong Inference," *Science* 1964.

Testing and confidence intervals

Constructing a confidence region: Given

$$\sqrt{n}(\widehat{\theta}_n - \theta_0) \xrightarrow{d} \mathcal{N}(0, I_{\theta_0}^{-1})$$

would like to say "with reasonably high confidence, $\theta_0 \in C_n$" for some set C_n. (This isn't the scientific method.)

Example (Wald confidence ellipse)

If $\sqrt{n}(\widehat{\theta}_n - \theta_0) \xrightarrow{d} \mathcal{N}(0, I_{\theta_0}^{-1})$ and I_θ is continuous,

$$C_{n,\gamma} := \left\{\theta : (\theta - \widehat{\theta}_n)^T I_{\widehat{\theta}_n}(\theta - \widehat{\theta}_n) \leq \frac{\gamma}{n}\right\}$$

(we'll modify notation slightly later) gives a confidence set with

$$P_\theta(\theta \in C_{n,\gamma}) \to \mathbb{P}(\|W\|_2^2 \leq \gamma) \text{ for } W \sim \mathcal{N}(0, I_d)$$

Duality: testing and confidence regions

Conjecture a model P_{θ_0} is "true" and then obtain

$$P_{\theta_0}\begin{pmatrix}\text{see data as extreme as}\\ \text{what we have seen}\end{pmatrix} \le \alpha$$

Definition (p-value)

Let $H_0 : \{P_\theta : \theta \in \Theta_0\}$. The *p-value* associated with a sample X_1^n is

$$\sup_{\theta\in\Theta_0} P_\theta \left(\text{Data at least as extreme as } X_1^n \text{ observed}\right).$$

Example (Normal observations)

For $H_0 : \{X_i \overset{\text{iid}}{\sim} \mathcal{N}(0,1)\}$, standard p-value is $P_0(|Z| \ge \sqrt{n}|\overline{X}_n|)$.

Neyman-Pearson tests

For a point null and alternative

$$H_0 : P_0 \quad \text{and} \quad H_1 : P_1$$

the test maximizing power subject to a level constraint is likelihood ratio test: for

$$T(x) := \log \frac{dP_1(x)}{dP_0(x)}$$

we

$$\begin{aligned} \text{Accept } H_1, \text{ reject } H_0 \quad &\text{if } T(x) > t \\ \text{Accept } H_0, \text{ reject } H_1 \quad &\text{if } T(x) < t \\ \text{balance/randomize} \quad &\text{if } T(x) = t. \end{aligned}$$

Asymptotic level of a test

- Setting: model family $\{P_\theta\}_{\theta \in \Theta}$,
- Null $H_0 : \theta \in \Theta_0 \subset \Theta$ (often $\Theta_0 = \{\theta_0\}$ is point null)
- T_n is sequence of test statistics that may reject null.

Definition

The *uniform asymptotic level* of T_n for null H_0 is

$$\limsup_{n \to \infty} \sup_{\theta_0 \in \Theta_0} P_{\theta_0}(T_n \text{ rejects}).$$

The *pointwise asymptotic level* of T_n for null H_0 is

$$\sup_{\theta_0 \in \Theta_0} \limsup_{n \to \infty} P_{\theta_0}(T_n \text{ rejects})$$

Typically, we want asymptotically level α tests (α small)

Three standard tests

- Generalized likelihood ratio test
- Wald test
- Rao's score test

Generalized likelihood ratio test

In model family $\{P_\theta\}_{\theta \in \Theta}$, testing

$$H_0 : \theta \in \Theta_0 \quad \text{vs} \quad H_1 : \theta \in \Theta$$

Analogue of likelihood ratio test:

$$T(x) := \log \frac{\sup_{\theta \in \Theta} p(x; \theta)}{\sup_{\theta \in \Theta_0} p(x; \theta)} = \log \frac{p(x; \widehat{\theta}_{\mathsf{mle}})}{\sup_{\theta \in \Theta_0} p(x; \theta)}$$

Wilks' Theorem

- $\Theta_0 = \{\theta_0\}$ is point null, $\theta_0 \in \operatorname{int}\Theta \subset \mathbb{R}^d$
- Log-likelihood $L_n(X;\theta) := \sum_{i=1}^n \ell_\theta(X_i) = \sum_{i=1}^n \log p_\theta(X_i)$
- MLE $\widehat{\theta}_n = \operatorname{argmax}_\theta L_n(\theta)$

Theorem (Wilks, simplified)

Define

$$\Delta_n := L_n(X;\widehat{\theta}_n) - L_n(X;\theta_0).$$

Then (under typical smoothness assumptions)

$$2\Delta_n \xrightarrow[P_{\theta_0}]{d} \chi^2_d.$$

Wald tests

- Insight: everything looks like quadratics (in classical case)
- Recall Wald confidence ellipse (for γ to be specified)

$$C_n := \left\{\theta : (\theta - \widehat{\theta}_n)^T I_{\widehat{\theta}_n}(\theta - \widehat{\theta}_n) \le \frac{\gamma}{n}\right\}$$

- Convergence under null $H_0 : P_{\theta_0}$ when $I_{\theta_0} \succ 0$,

$$n(\theta_0 - \widehat{\theta}_n)^T I_{\widehat{\theta}_n}(\theta_0 - \widehat{\theta}_n) \xrightarrow[H_0]{d} \|W\|_2^2 \stackrel{\text{dist}}{=} \chi_d^2, \quad W \sim \mathcal{N}(0, I_d)$$

Definition (Wald test of point null $\theta = \theta_0$)

Let $u_{d,\alpha}^2$ be the α quantile of a χ_d^2 R.V., $\mathbb{P}(\|W\|_2^2 \le u_{d,\alpha}^2) = \alpha$ for $W \sim \mathcal{N}(0, I_d)$. The *Wald test* at asymptotic level α is

$$T_n := \begin{cases} \text{Reject} & \text{if } \theta_0 \notin C_{n,\alpha} \\ \text{Don't reject} & \text{if } \theta_0 \in C_{n,\alpha} \end{cases}$$

where $C_{n,\alpha}$ is Wald confidence ellipse with $\gamma = u_{d,\alpha}^2$.

What about nuisance parameters (composite nulls)?

Example (Normal sample, unknown variance)

Say $X_i \stackrel{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$ and $H_0 : \mu = 0$, but σ^2 unspecified

Idea: essentially, estimate the nuisance parameters

Setting: I_θ exists and is invertible, so MLE (usually) satisfies

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

Insight: asymptotics of sub-vectors are immediate

Notation for Wald test with nuisances

For $v \in \mathbb{R}^d$ and $\Sigma \in \mathbb{R}^{d \times d}$, let $[v]_{1:k}$ be the first k components of v and $\Sigma^{(k)}$ be the k-by-k principal submatrix

$$[v]_{1:k} = \begin{bmatrix} v_1 \\ \vdots \\ v_k \end{bmatrix}, \quad \Sigma = \begin{bmatrix} \Sigma^{(k)} & \cdots \\ \vdots & \ddots \end{bmatrix}, \; \Sigma^{(k)} \in \mathbb{R}^{k \times k}$$

Corollary

If $\Sigma_\theta = I_\theta^{-1}$, so $\Sigma^{(k)} = (I_\theta^{-1})^{(k)}$, then (under typical smoothness conditions)

$$\sqrt{n}([\widehat{\theta}_n]_{1:k} - [\theta]_{1:k}) \overset{d}{\to} \mathcal{N}\left(0, \Sigma^{(k)}\right)$$

and

$$n([\widehat{\theta}_n]_{1:k} - [\theta]_{1:k})^\top (\Sigma_{\widehat{\theta}_n}^{(k)})^{-1} ([\widehat{\theta}_n]_{1:k} - [\theta]_{1:k}) \xrightarrow[P_\theta]{d} \chi_k^2.$$

A "reduction" in information

Lemma

For symmetric block matrix

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix},$$

$M = A^{-1}$ *satisfies* $M_{11} = (A_{11} - A_{12}A_{22}^{-1}A_{21})^{-1}$

- Apply with $A = I_\theta$, $\Sigma = M = I_\theta^{-1}$
- Get "reduced" information $(I_\theta)_{11} - (I_\theta)_{12}(I_\theta)_{22}^{-1}(I_\theta)_{21}$

Wald test with nuisance parameters

- Composite null on $\theta \in \mathbb{R}^d$

$$H_0 : \{\theta_1 = \theta_1^0, \ldots, \theta_k = \theta_k^0, \quad \theta_{k+1}, \ldots, \theta_d \text{ unrestricted}\}.$$

- Confidence ellipse

$$C_{n,\alpha} := \Big\{\theta \in \mathbb{R}^d :$$

$$([\theta]_{1:k} - [\widehat{\theta}_n]_{1:k})^\top (\Sigma_{\widehat{\theta}_n}^{(k)})^{-1} ([\theta]_{1:k} - [\widehat{\theta}_n]_{1:k}) \le \frac{u_{k,\alpha}^2}{n} \Big\}$$

- Wald test at (pointwise) asymptotic level α is

$$T_n := \begin{cases} \text{Reject} & \text{if } \theta_0 \notin C_{n,\alpha} \\ \text{Don't reject} & \text{if } \theta_0 \in C_{n,\alpha} \end{cases}$$

Wald test comments and example

- Actually need to use $\Sigma^{(k)}_{\widehat{\theta}_n}$ to get a consistent Fisher information estimate

Example (Gaussian mean, unknown covariance)

For null $H_0 : \{\mathcal{N}(\theta, \Sigma), \theta = 0, \Sigma \succ 0\}$,

$$C_{n,\alpha} := \left\{ \theta \in \mathbb{R}^d : \theta^\top \widehat{\Sigma}^{-1} \theta \le \frac{u_{d,\alpha}^2}{n} \right\}$$

and

$$\mathbb{P}(\overline{X}_n \in C_{n,\alpha}) \to \alpha.$$

Rao's score test

- an asymptotic test that doesn't rely on MLE computation
- use limits of score under θ, $\sqrt{n} P_n \nabla \ell_\theta \xrightarrow{d}_{P_\theta} \mathcal{N}(0, I_\theta)$
- under null $H_0 : \theta = \theta_0 \in \mathbb{R}^d$,

$$n P_n \nabla \ell_{\theta_0}^\top I_{\theta_0}^{-1} \nabla \ell_{\theta_0} \underset{H_0}{\xrightarrow{d}} \chi_d^2$$

Definition (Rao test)

The *Rao test* of asymptotic level α rejects $H_0 : \theta = \theta_0$ when

$$P_n \nabla \ell_{\theta_0}^\top I_{\theta_0}^{-1} \nabla \ell_{\theta_0} \geq \frac{u_{d,\alpha}^2}{n}.$$

- strong connections to optimality (revisit later)
- analogues for composite nulls to other cases