Asymptotics III: Bayes Inference and Large-Sample Tests

MIT 18.655

Dr. Kempthorne

Spring 2016

Outline

Consistency of Posterior Distribution

Framework

- $X_1, \ldots, X_n$ iid $P_\theta, \theta \in \Theta$.
- Θ (open) $\subset R$ or $\Theta = \{\theta_1, \ldots, \theta_k\}$ finite.
- Regular model with identifiable θ.

Consistency: Finite Θ

Posterior distribution of θ given $\mathbf{X}_n = (X_1, \ldots, X_n)$:

$$\pi(\theta' \mid \mathbf{X}_n) \equiv P[\theta = \theta' \mid X_1, \ldots, X_n],\ \theta' \in \Theta.$$

Definition: $\pi(\cdot \mid \mathbf{X}_n)$ is **consistent** if and only if for every $\theta' \in \Theta$,

$$P_{\theta'}[|\pi(\theta' \mid \mathbf{X}_n) - 1|] \geq \epsilon \to 0$$

for all $\epsilon > 0$.

Definition: $\pi(\cdot \mid \mathbf{X}_n)$ is **a.s. (almost surely) consistent** if and only if for every $\theta' \in \Theta$,

$$\pi(\theta' \mid \mathbf{X}_n) \xrightarrow{a.s. P_{\theta'}} 1.$$

Consistency of Posterior Distribution

Theorem 5.5.1 Let $\pi_j = P[\theta = \theta_j]$, $j = 1, \ldots, k$ denote the prior distribution of θ. Then

$$\pi(\cdot \mid \mathbf{X}_n) \text{ is consistent iff } \pi_j > 0, \text{ for all } \pi_j \in \Theta.$$

Proof:

- Let $p(x \mid \theta)$ denote the density/pmf function of a single X_i. The posterior distribution is given by:
$$\begin{aligned} \pi(\theta_j \mid X_1, \ldots, X_n) &= P[\theta = \theta_j \mid X_1, \ldots, X_n] \\ &= \frac{\pi_j \prod_{i=1}^n p(X_i \mid \theta_j)}{\sum_{a=1}^k \pi_a \prod_{i=1}^n p(X_i \mid \theta_a)} \end{aligned}$$
If any $\pi_j = 0$, then $\pi(\theta_j \mid \mathbf{X}_n) = 0$ for all n; i.e., the posterior is not consistent.
- Suppose all $\pi_j > 0$. For a fixed j, suppose θ_j is true, i.e., $\theta = \theta_j$.
We show that
$$\pi(\theta_j \mid \mathbf{X}_n) \longrightarrow 1 \text{ and } \pi(\theta_a \mid \mathbf{X}_n) \longrightarrow 0, \text{ for } a \neq j.$$

Consistency Theorem Proof

Proof (continued)

Evaluate the log of the posterior odds to the true θ :

$$
\begin{aligned}
\log\left[\frac{\pi(\theta_a \mid \mathbf{X}_n)}{\pi(\theta_j \mid \mathbf{X}_n)}\right] &= \log\left[\frac{\pi_a \prod_{i=1}^n p(X_i \mid \theta_a)}{\pi_j \prod_{i=1}^n p(X_i \mid \theta_j)}\right] \\
&= \log\left[\frac{\pi_a}{\pi_j}\right] + \log\left[\frac{\prod_{i=1}^n p(X_i \mid \theta_a)}{\prod_{i=1}^n p(X_i \mid \theta_j)}\right] \\
&= \log\left[\frac{\pi_a}{\pi_j}\right] + \sum_{i=1}^n \log\left[\frac{p(X_i|\theta_a)}{p(X_i|\theta_j)}\right] \\
&= n\left(\frac{1}{n}\log\left[\frac{\pi_a}{\pi_j}\right] + \frac{1}{n}\sum_{i=1}^n \log\left[\frac{p(X_i|\theta_a)}{p(X_i|\theta_j)}\right]\right) \\
&\longrightarrow n\left(\ 0 \quad + \quad E\left[\log\left[\frac{p(X_1|\theta_a)}{p(X_1|\theta_j)}\right]\right]\right) \\
&\longrightarrow \begin{cases} 0 & if \quad a = j \\ -\infty & if \quad a \neq j \end{cases}
\end{aligned}
$$

(Shannon's Inequality gives $E\left[\log\left[\frac{p(X_1|\theta_a)}{p(X_1|\theta_j)}\right]\right] < 0$, for $a \neq j$)

Outline

Asymptotic Normality of Posterior Distribution

Theorem 5.5.2 ("Bernstein/von Mises").

- $\mathbf{X}_n = (X_1, \ldots, X_n)$ where the X_i are iid $P_{\theta_0}, \theta_0 \in \Theta$.
- $\hat{\theta}_n = \hat{\theta}_n(\mathbf{X}_n)$ is the MLE of θ_0
- Regularity conditions are satisfied such that
$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{\mathcal{L}} N(0, \mathrm{I}^{-1}(\theta_0)).$$
- The prior distribution on Θ has density $\pi(\cdot)$ which is continuous and positive at all $\theta' \in \Theta$.
- Consider the scaled version of the posterior distribution:
$$\mathcal{L}\left(\sqrt{n}(\theta - \hat{\theta}) \mid \mathbf{X}_n\right)$$

Under sufficient regularity conditions:
$$\mathcal{L}\left(\sqrt{n}(\theta - \hat{\theta}) \mid \mathbf{X}_n\right) \longrightarrow N(0, \mathrm{I}^{-1}(\theta_0))$$
i.e.,
$$\pi\left(\sqrt{n}(\theta - \hat{\theta}) \leq x \mid \mathbf{X}_n\right) \longrightarrow \Phi(x\sqrt{\mathrm{I}(\theta_0)})$$

Bernstein / Von Mises Theorem

Proof:

- To compute the asymptotic distribution of $\sqrt{n}(\theta - \hat{\theta}(\mathbf{X}_n))$, define
$$t = \sqrt{n}(\theta - \hat{\theta}(\mathbf{X}_n))$$
so that
$$\theta = \hat{\theta}(\mathbf{X}_n) + \tfrac{t}{\sqrt{n}}.$$
- The posterior density of t given $\mathbf{X}_n$ is
$$\begin{aligned} q_n(t) &\propto \pi(\hat{\theta}(\mathbf{X}_n) + \tfrac{t}{\sqrt{n}}) \prod_{i=1}^n p(X_i \mid \hat{\theta}(\mathbf{X}_n) + \tfrac{t}{\sqrt{n}}) \\ &= c_n^{-1} \pi(\hat{\theta}(\mathbf{X}_n) + \tfrac{t}{\sqrt{n}}) \prod_{i=1}^n p(X_i \mid \hat{\theta}(\mathbf{X}_n) + \tfrac{t}{\sqrt{n}}) \end{aligned}$$
where $c_n = \int_{-\infty}^{\infty} \pi(\hat{\theta}(\mathbf{X}_n) + \tfrac{t}{\sqrt{n}}) \prod_{i=1}^n p(X_i \mid \hat{\theta}(\mathbf{X}_n) + \tfrac{t}{\sqrt{n}}) dt.$
- Divide numerator and denominator of $q_n(t)$ by
$$\prod_{i=1}^n p(X_i \mid \hat{\theta}(\mathbf{X}_n))$$

Bernstein / Von Mises Theorem

Proof (continued)

$$\begin{aligned} q_n(t) &= c_n^{-1}\pi(\hat{\theta}(\mathbf{X}_n)+\tfrac{t}{\sqrt{n}})\prod_{i=1}^n p(X_i \mid \hat{\theta}(\mathbf{X}_n)+\tfrac{t}{\sqrt{n}}) \\ &= c_n^{-1}\pi(\hat{\theta}+\tfrac{t}{\sqrt{n}})exp\{\sum_{i=1}^n log(p(X_i \mid \hat{\theta}+\tfrac{t}{\sqrt{n}}))\} \\ &= d_n^{-1}\pi(\hat{\theta}+\tfrac{t}{\sqrt{n}})exp\{\sum_{i=1}^n \ell(X_i \mid \hat{\theta}+\tfrac{t}{\sqrt{n}}) - \ell(X_i,\hat{\theta})\} \end{aligned}$$

where

$$d_n = \int_{-\infty}^{\infty}\pi(\hat{\theta}+\tfrac{t}{\sqrt{n}})exp\{\sum_{i=1}^n \ell(X_i \mid \hat{\theta}+\tfrac{t}{\sqrt{n}}) - \ell(X_i,\hat{\theta})\}dt$$

Claims

- $d_n q_n(t) \xrightarrow{P_{\theta_0}} \pi(\theta_0)exp\{-\frac{t^2 \mathrm{I}(\theta_0)}{2}\}$
- $d_n \xrightarrow{P_{\theta_0}} \pi(\theta_0)\int_{-\infty}^{\infty} exp\{-\frac{s^2 \mathrm{I}(\theta_0)}{2}\}ds = \frac{\pi(\theta_0)\sqrt{2\pi}}{\sqrt{\mathrm{I}(\theta_0)}}$

which give:

$$q_n \xrightarrow{P_{\theta_0}} \sqrt{\mathrm{I}(\theta_0)}\phi(t\sqrt{\mathrm{I}(\theta_0)}).$$

Theorem follows by Scheffe's Theorem (B.7.6).

Limiting Posterior Distributions: Examples

Posterior Distribution of Normal Mean

- $X_1, \ldots, X_n$ iid $N(\theta_0, \sigma^2)$ with σ^2 known.
- Prior distribution: $\theta \sim N(\eta, \tau^2)$.
- Posterior distribution:

$$\pi(\theta \mid \mathbf{X}_n) = N(\eta_n, \tau_n^2),$$

where

$$\tau_n^{-2} = \tau^{-2} + \frac{n}{\sigma^2}$$
$$\eta_n = w_n\eta + (1 - w_n)\overline{X}, \text{ with } w_n = \frac{\sigma^2}{n\tau^2+\sigma^2}$$

Note:

- $\eta_n \longrightarrow \hat{\theta} = \overline{X}$, $\tau_n^2 \longrightarrow 0$, and $\overline{X} \xrightarrow{P_{\theta_0}} \theta$, so

$$\pi(\theta \mid \mathbf{X}_n) \xrightarrow{P_{\theta_0}} \text{point-mass at } \theta = \theta_0.$$

- A posteriori,

$$\begin{array}{rcl} \sqrt{n}(\theta - \hat{\theta}) & \sim & N(\sqrt{n}w_n(\eta - \overline{X}), n(\frac{n}{\sigma^2} + \frac{1}{\tau^2})^{-1}) \\ & \longrightarrow & N(0, \mathrm{I}^{-1}(\theta_0)) = N(0, \sigma^2) \end{array}$$

Limiting Posterior Distributions: Examples

Posterior Distribution of Success Probability in Bernoulli Trials

- $X_1, \ldots, X_n$ iid $Bernoulli(\theta_0)$.
- $S_n = \sum_1^n X_i \sim Binomial(n, \theta_0)$.
- Prior distribution: $\theta \sim Beta(r, s)$.
- Posterior distribution $\theta \mid S_n \sim Beta(r^*, s^*)$, where $r^* = S_n + r$, and $s^* = s + (n - S_n)$.
- By Problem 5.3.20, if $r^* \to \infty$ and $s^* \to \infty$ such that $r^*/(r^* + s^*) \to \theta_0 \in (0, 1)$, then the $Beta(r^*, s^*)$ r.v. θ:

$$P\left[\sqrt{r^* + s^*}\frac{(\theta - r^*/(r^*+s^*))}{\sqrt{\theta_0(1-\theta_0)}}\right] \longrightarrow N(0, 1).$$

This is easily shown to be equivalent to

$$\sqrt{n}(\theta - \overline{X}) \xrightarrow{\mathcal{L}} N(0, \theta_0(1 - \theta_0)) = N(0, \mathrm{I}^{-1}(\theta_0))$$

Outline

Mutual Optimality of Bayes and MLE Procedures

Theorem 5.5.3 Under the conditions of the previous theorems, let $\hat{\theta}$ be the MLE of θ and let $\hat{\theta}^*$ be the median of the posterior distribution of θ. Then

(i). From a frequentist point of view, i.e., given P_θ:

$$\sqrt{n}(\hat{\theta}^* - \hat{\theta}) \xrightarrow{a.s.P_\theta} 0, \text{ for all } \theta$$
$$\hat{\theta}^* = \theta + \frac{1}{n}\sum_{i=1}^n \mathrm{I}^{-1}(\theta)\frac{\partial \ell}{\partial \theta}(X_i, \theta) + o_{P_\theta}(n^{-1/2})$$
$$\sqrt{n}(\hat{\theta}^* - \theta) \xrightarrow{\mathcal{L}} N(0, \mathrm{I}^{-1}(\theta)).$$

(ii). From a Bayesian point of view, i.e., for $\pi(\theta \mid X_1, \ldots, X_n)$:

$$E[\sqrt{n}(|\theta - \hat{\theta})| - |\theta - \hat{\theta}^*|) \mid X_1, \ldots, X_n] = o_P(1), \text{ and}$$

$$E[\sqrt{n}(|\theta - \hat{\theta})| - |\theta|) \mid X_1, \ldots, X_n] = \min_d \left(E[\sqrt{n}(|\theta - d)| - |\theta|) \mid X_1, \ldots, X_n]\right) + o_P(1).$$

Mutual Optimality of Bayes and MLE Procedures

Significant Results

- Bayes estimates for a wide variety of loss functions and priors are asymptotically efficient in the sense being asymptotically unbiased with minimum asymptotic variance.
- Maximum-likelihood estimates are asymptotically equivalent in a Bayesian sense to the Bayes estimate for a variety of priors and loss functions.
  E.g., the Bayesian posterior median with $L(\theta, d) = |\theta - d|$,
  the Bayesian posterior mean with $L(\theta, d) = |\theta - d|^2$.

Bayes Credible Regions

Theorem 5.5.4 Under the conditions of the previous theorems, consider

- The **Bayes Credible Region**:
$$C_n(X_1, \dots, X_n) = \{\theta : \pi(\theta \mid X_1, \dots, X_n) \geq c_n\},$$
where c_n is chosen so that $\pi(C_n \mid X_1, \dots, X_n) = 1 - \alpha$.
- For $\gamma : 0 < \gamma < 1$, the level $(1 - \gamma)$ **Asymptotically Optimal Interval Estimate** based on $\hat{\theta}$, given by
$$Interval_n(\gamma) = [\hat{\theta} - d_n(\gamma), \hat{\theta} + d_n(\gamma)]$$
where $d_n(\gamma) = [\Phi^{-1}(1 - \gamma/2)] \times (\frac{1}{\sqrt{n}\sqrt{[\mathrm{I}(\theta_0)]}})$.

Then, for every $\epsilon > 0$, and every θ:
$$P_\theta\left[Interval_n(\alpha + \epsilon) \subset C_n(X_1, \dots, X_n) \subset Interval_n(\alpha - \epsilon)\right] \longrightarrow 1$$

Outline

Likelihood Ratio Test

Likelihood Ratio Test Statistic

- $\mathbf{X}_n = (X_1, \ldots, X_n)$ iid $P_\theta, \theta \in \Theta$.
- Testing null vs alternative hypotheses:
$$H : \theta \in \Theta_0 \text{ vs } K : \theta \notin \Theta_0.$$
- Likelihood ratio statistic:
$$\lambda(\mathbf{x}_n) = \frac{\sup_{\theta \in \Theta} p(\mathbf{x}_n \mid \theta)}{\sup_{\theta \in \Theta_0} p(\mathbf{x}_n \mid \theta)}$$
Standard transformation:
$$2 \log \lambda(\mathbf{x}_n) = 2[\ell_n(\hat{\theta} \mid \mathbf{x}_n) - \ell_n(\hat{\theta}_0 \mid \mathbf{x}_n)]$$
where $\hat{\theta}(\mathbf{x}_n)$ is the MLE (over all Θ) and $\hat{\theta}_0(\mathbf{x}_n)$ is the MLE under $H : \theta \in \Theta_0$.

Theorem 6.3.1 Given suitable assumptions (e.g. Theorem 6.2.2), if $\Theta \subset R^r$, and $H : \theta = \theta_0$ is true, then
$$2 \log \lambda(\mathbf{x}) = 2[\ell_n(\hat{\theta} \mid \mathbf{x}) - \ell_n(\theta_0)] \xrightarrow{\mathcal{L}} \chi^2_r,$$

Theorem 6.2.2 Proof

- By Theorem 6.2.2. Given suitable assumptions, the MLE $\hat{\theta}(\mathbf{x}_n)$ satisfies
$$\hat{\theta}(\mathbf{x}_n) = \theta + \frac{1}{n}\sum_{i=1}^{n} \mathrm{I}^{-1}(\theta) D\ell(X_i, \theta) + o_P(n^{-1/2})$$
so that
$$\sqrt{n}(\hat{\theta}(\mathbf{x}_n) - \theta) \xrightarrow{\mathcal{L}} N(0, \mathrm{I}^{-1}(\theta)).$$

- The Taylor expansion of $\ell_n(\theta)$ about $\hat{\theta}(\mathbf{x}_n)$ evaluated at $\theta = \theta_0$ gives
$$\begin{aligned} 2\log\lambda(\mathbf{x}) &= 2[\ell_n(\hat{\theta} \mid \mathbf{x}) - \ell_n(\theta_0 \mid \mathbf{X})] \\ &= n(\hat{\theta}(\mathbf{x}_n) - \theta_0)^T \mathrm{I}_n(\theta^*)(\hat{\theta}(\mathbf{x}_n) - \theta_0) \end{aligned}$$
where $I_n(\theta) = || - \frac{1}{n}\sum_{i=1}^{n} \frac{\partial}{\partial\theta_k}\frac{\partial}{\partial\theta_j} \log p(X_i \mid \theta)||$,
the $r \times r$ matrix: $I_n(\theta) \xrightarrow{\mathcal{P}_{\theta_0}} \mathrm{I}(\theta_0)$

Theorem 6.2.2 Proof (continued)

- With $\mathbf{V} \sim N(0, \mathrm{I}^{-1}(\theta_0))$,
$$\begin{aligned} 2\log\lambda(\mathbf{x}) &= 2[\ell_n(\hat{\theta} \mid \mathbf{x}) - \ell_n(\theta_0 \mid \mathbf{X})] \\ &= n(\hat{\theta}(\mathbf{x}_n) - \theta_0)^T \mathrm{I}_n(\theta^*)(\hat{\theta}(\mathbf{x}_n) - \theta_0) \\ &\xrightarrow{\mathcal{L}} \mathbf{V}^T \mathrm{I}(\theta_0)\mathbf{V} \end{aligned}$$
and by Corollary $B.6.2$
$$\mathbf{V}^T \mathrm{I}(\theta_0)\mathbf{V} \sim \chi^2_r.$$

Theorem 6.3.2 Given suitable assumptions (e.g. Theorem 6.2.2), if $\Theta \subset R^r$, and $H : \theta \in \Theta_0$ with Θ_0 of dimension $q < r$, then
$$2\log\lambda(\mathbf{x}) = 2[\ell_n(\hat{\theta} \mid \mathbf{x}) - \ell_n(\hat{\theta}_0 \mid \mathbf{X})] \xrightarrow{\mathcal{L}} \chi^2_{r-q}.$$

Outline

The Wald Test

The asymptotic level-α Wald Test of the simple hypothesis

$$H : \theta = \theta_0 \text{ vs } K : \theta \neq \theta_0$$

rejects H when

$$W_n(\theta_0) = n(\hat{\theta}(\mathbf{x}_n) - \theta_0)^T \mathrm{I}(\theta_0)(\hat{\theta}(\mathbf{x}_n) - \theta_0) \geq C^*,$$

where the critical value C^* is such that $P(\chi^2_r > C^*) = 1 - \alpha$.

- Under the assumptions of Theorem 6.2.2
$$\sqrt{n}(\hat{\theta}(\mathbf{x}_n) - \theta) \xrightarrow{\mathcal{L}} N(0, \mathrm{I}^{-1}(\theta)).$$
- By Slutsky's theorem:
$$n(\hat{\theta}(\mathbf{x}_n) - \theta)^T \mathrm{I}(\theta)(\hat{\theta}(\mathbf{x}_n) - \theta) \xrightarrow{\mathcal{L}} \mathbf{V}^T \mathrm{I}(\theta)\mathbf{V}$$
where $\mathbf{V} \sim N_r(0, \mathrm{I}^{-1}(\theta))$.

The Wald Test extends to apply to a composite null hypothesis $H : \theta \in \Theta_0 \subset R^q$. If the MLE $\hat{\theta}_0(\mathbf{x}_n)$ under the null is consistent, then it can replace θ_0 in the Wald Test statistic which is asymptotically χ^2_{r-q} under H, where q is the dimensionality of Θ_0.

Outline

The Rao Score Test

- Simple hypothesis $H : \theta = \theta_0$.
- Apply the Central Limit Theorem to the maximum-likelihood contrast function, evaluated at $\theta = \theta_0$:
$$\psi_n(\theta_0) = \frac{1}{n}\sum_{i=1}^n D_\theta \ell_n(\theta_0) \xrightarrow{\mathcal{L}} N(0, \mathrm{I}(\theta_0)),$$
when H is true.
- It follows taht under H
$$R_n(\theta_0) = n\psi_n^T(\theta_0)\mathrm{I}^{-1}(\theta_0)\psi_n(\theta_0) \xrightarrow{\mathcal{L}} \chi_r^2.$$

The asymptotic level-α Rao Score Test rejects H when
$$R_n(\theta_0) \geq C^*$$
where $C^* : P(\chi_r^2 > C^*) = 1 - \alpha$.

Notes:

- The Rao Score Test does not require the MLE**!!**
- Extension to composite null hypothesis H only requires MLE under H (see Theorem 6.3.5).