LECTURE 11: CONFIDENCE SETS. BAYESIANS IN LARGE SAMPLES.

Michal Kolesár*

October 9, 2024

REFERENCE This is not really covered in HMC, apart from some basics in Chapter 4.2. Chapter 9 in CB is a much better reference. H22, Chapter 14 and 16.13.

In this lecture, we will study confidence sets. We observe data X, with distribution $F(\cdot \mid \theta)$, $\theta \in \Theta$. The parameter space Θ may be finite- or infinite-dimensional. We are interested in summarizing what we know about the parameter of interest $g(\theta)$ by constructing a data-dependent set $C(X)$ that contains $g(\theta)$ with large probability.

In standard problems, if $g(\theta)$ is a scalar, the set $C(X)$ will often take form of an interval, $[L(X), U(X)]$. One possibility is to set $L(X) = -\infty$ and $U(X) = \infty$. Such an interval will contain $g(\theta)$ with probability 1, a large probability indeed. Of course, the problem with this interval is that it is not informative about $g(\theta)$. So, we want the set $C(X)$ to simultaneously achieve two things: have a small volume (or be short, if it's an interval), and also contain $g(\theta)$ with a high probability. For a particular value of θ, the probability of containing $g(\theta)$, $P_\theta(g(\theta) \in C(X))$ is called the *coverage probability*. Similar to how we resolved the size-power tradeoff in testing, it is customary to bound the coverage probability, and subject to this bound, seek sets with small volume.

Definition 1. A set $C(X) \subseteq \Theta$ has *confidence level* $1-\alpha$ if $\inf_{\theta\in\Theta} P_\theta(\theta \in C(X)) \geq 1-\alpha$. In this case we call $C(X)$ a *confidence set* with level $1-\alpha$. If $C(X)$ takes the form of an interval, we call it a confidence interval (CI) with level $1-\alpha$.

So our task will be to find confidence sets that are informative (as judged by their average length), while maintaining a given confidence level. For particular values of θ, the coverage probability of the set $C(X)$ may be higher than $1-\alpha$, just as the rejection rate of a test may be lower than α for particular null values of θ. Note the randomness in the coverage comes from the data X: the set $C(X)$ changes in repeated sampling.

Next, we consider two small-sample methods for finding confidence sets. We then consider some large-sample approaches, and compare them to what a Bayesian would do.

*Email: mkolesar@princeton.edu.

1. TEST INVERSION AND THE PIVOT METHOD

For each possible parameter value $\theta_0 \in \Theta$, consider the problem of testing $H_0\colon g(\theta) = g_0$ against $H_1\colon g(\theta) \neq \theta_0$. Suppose that for each such hypothesis we have a test $\phi_{g_0}(X)$ of level α. Then the confidence set

$$C(X) = \{g_0 :\ H_0\colon g(\theta) = g_0 \text{ is not rejected}\}$$

has confidence level $1-\alpha$. Indeed, suppose that the true value of parameter is θ_0. Since the test $\phi_{g(\theta_0)}$ has level α, by construction, $E_{\theta_0}[\phi_{g(\theta_0)}(X)] = P_{\theta_0}(\phi_{g(\theta_0)}(X) = 1) \leq \alpha$. So, for any θ,

$$P_\theta(g(\theta) \in C(X)) = P_\theta(\phi_{g(\theta)}(X) = 0) \geq 1 - \alpha,$$

which implies $\inf_{\theta\in\Theta} P_\theta\{g(\theta) \in C(X)\} \geq 1-\alpha$. This procedure is called *test inversion*. One problem with test inversion is that sometimes a confidence set obtained via this procedure will consist of several disjoint intervals, which makes it a bit hard to interpret.

Conversely, if we have a way of constructing a confidence set with confidence level $1-\alpha$, we can use it for testing $H_0\colon g(\theta) = g_0$ against a two-sided alternative by letting $\phi_{g_0}(X) = 1$ if $g_0 \notin C(X)$. In other words, we accept the null if $g_0 \in C(X)$. This test will have level α.

Example 1. Let $X_1, \ldots, X_n$ be a random sample from distribution $\mathcal{N}(\theta, \sigma^2)$, σ^2 known. Let us use test inversion to construct a confidence set for θ of level $1-\alpha$. Consider the problem of testing the null hypothesis, $H_0\colon \theta = \theta_0$ against the alternative, $H_1\colon \theta \neq \theta_0$. The uniformly most powerful (UMP) unbiased test rejects if $|\overline{X}_n - \theta_0| \geq z_{1-\alpha/2}\sigma/\sqrt{n}$. Thus, $\theta \in C(X)$ iff

$$-z_{1-\alpha/2}\sigma/\sqrt{n} \leq \overline{X}_n - \theta \leq z_{1-\alpha/2}\sigma/\sqrt{n},$$

or, rewriting this, the confidence set is given by

$$\left[\overline{X}_n - z_{1-\alpha/2}\sigma/\sqrt{n}\ \ \overline{X}_n + z_{1-\alpha/2}\sigma/\sqrt{n}\right] \tag{1}$$

So in this case, we actually get an interval. Consider inverting the one-sided z-test, which rejects if $\overline{X}_n - \theta_0 \geq z_{1-\alpha}\sigma/\sqrt{n}$. In this case, $\theta \in C(X)$ iff

$$\overline{X}_n - z_{1-\alpha}\sigma/\sqrt{n} \leq \theta,$$

so the confidence interval will be one-sided, $[\overline{X}_n - z_{1-\alpha}\sigma/\sqrt{n}, \infty)$. ⊠

Example 2. Let $X_1, \ldots, X_n$ be a random sample from the distribution $\mathcal{N}(\mu, \sigma^2)$. Let us use the test inversion to construct a confidence set for σ^2 of level $1-\alpha$. Consider the problem of testing the null hypothesis, $H_0\colon \sigma^2 = \sigma_0^2$ against the alternative, $H_1\colon \sigma^2 \neq \sigma_0^2$. Under the null hypothesis, $(n-1)S_n^2/\sigma_0^2 \sim \chi^2(n-1)$. The test that accepts the null

hypothesis if and only if

$$\chi^{-2}_{\alpha/2}(n-1) \leq (n-1)S_n^2/\sigma_0^2 \leq \chi^{-2}_{1-\alpha/2}(n-1).$$

has size α. This test will accept the null hypothesis $\sigma^2 = \sigma_0^2$ if and only if

$$(n-1)S_n^2/\chi^{-2}_{1-\alpha/2}(n-1) \leq \sigma_0^2 \leq (n-1)S_n^2/\chi^{-2}_{\alpha/2}(n-1),$$

so again, the confidence set is an interval, which has coverage $1-\alpha$ independently of the value of μ. Notice the quantiles $\chi^{-2}_{\alpha/2}(n-1)$ and $\chi^{-2}_{1-\alpha/2}(n-1)$ get switched. ⊠

In the above examples, we first constructed a test with a given significance level, and then inverted it. A closely related approach is to find a pivotal statistic, and invert it. Recall from Lecture 8 that a pivot for a parameter $g(\theta)$ is a statistic $Q = q(X, g_0)$ such that when $g(\theta) = g_0$, its distribution doesn't depend on the value of g_0 or any other of any nuisance parameters that determine the distribution of the data X. In lecture 8, we used the pivot for constructing tests, but we can also use it for CI construction.

Example 3. In Lecture 8, we showed that the t-statistic, $T_n = (\overline{X}_n - \mu)/\sqrt{S_n^2/n} \sim t(n-1)$, where $X_i \sim \mathcal{N}(\mu, \sigma^2)$. If μ is a parameter of interest, then this is a pivot. Inverting the t-statistic gives us the usual CI, $\overline{X}_n \pm t_{1-\alpha/2}(n-1)S_n/\sqrt{n}$. ⊠

More generally, for any pivot Q, the distribution of Q is independent of the true parameter value, we can find numbers a and b such that $P_\theta\{a \leq q(X,\theta) \leq b\} = 1-\alpha$ for all $\theta \in \Theta$. Then we can construct a level α test by accepting the null hypothesis that $\theta = \theta_0$ if and only if $a \leq q(X, \theta_0) \leq b$. The confidence set will consist of all parameter values θ_0 which are accepted:

$$C(X) = \{\theta \in \Theta \colon a \leq q(X,\theta) \leq b\}.$$

2. PRATT'S THEOREM (OPTIONAL)

Intuitively, inverting powerful tests should yield short confidence intervals (confidence sets with small volume). Inverting tests that are UMP against a particular alternative θ_1 should yield confidence intervals that are as short as possible when θ_1 is the true parameter. The next result formalizes this intuition.

Theorem 2 (Pratt 1961). Let $X \sim f(x \mid \theta)$, let $C(X)$ be a confidence set for $\theta \in \mathbb{R}$ based on inverting tests $\phi_\theta(X)$, and let $\lambda(C(X)) = \int_\theta (1-\phi_\theta(X))d\theta$ denote its volume. Then, for any θ_1,

$$E_{\theta_1}[\lambda(C(X))] = \int_\Theta (1-\beta_\theta(\theta_1))\,\mathrm{d}\theta = \int_\Theta P_{\theta_1}(\theta \in C(X))\,\mathrm{d}\theta,$$

where $\beta_\theta(\theta_1) = E_{\theta_1}[\phi_\theta]$ denotes the power of the test ϕ_θ against θ_1. Furthermore, if the tests $\phi_\theta(X)$ are UMP against θ_1, and the expected length is finite, then $C(X)$ achieves the shortest expected length at θ_1 among all confidence sets with level $1-\alpha$.

Proof. The first result follows from changing the order of integration:

$$E_{\theta_1}[\lambda(C(X))] = E_{\theta_1}\left[\int_\Theta (1-\phi_\theta(X))\,\mathrm{d}\theta\right] = \int_\Theta \int_{\mathcal{X}} (1-\phi_\theta(x)) f(x \mid \theta_1)\,\mathrm{d}x\,\mathrm{d}\theta = \int_\Theta (1-\beta_\theta(\theta_1))\,\mathrm{d}\theta.$$

If $C(X)$ is based on inverting tests ϕ_θ that are UMP against θ_1, then for each $\theta \neq \theta_1$, such tests minimize $1-\beta_\theta(\theta_1)$, and therefore also minimize the integral. □

The second part, in spite of often being cited to justify the practice of constructing confidence sets by inverting tests, is not that useful since it only tells us about how to minimize the length of CI at a particular point, and often, this comes at the expense of terrible expected length at other points (generically, we won't simultaneously be able to minimize length at all θ). A more useful result would tell us how to construct a CI that minimizes its maximum length $\sup_\theta E_\theta[\lambda(C(X))]$, or average expected length, averaged over θ's using some weights $w(\theta)$, or how one can ever justify reporting one-sided CIs (which have infinite length and therefore Pratt's result doesn't apply). We'll leave those theorems for later courses.

Remark 3. The current theorem is useful, however, as a benchmark for the best possible performance of a CI in more complicated problems: if we construct a CI that is "close", in some sense, to the Pratt bound uniformly over θ, then we know one cannot improve much over such CI.

Example 1 (continued). Since CI in Example 1 is based on inverting UMP unbiased tests, the CI in eq. (1) is the CI with the shortest expected length among all intervals based on inverting unbiased tests. However, stated in this way, it's not clear why one should restrict attention to unbiased tests.

What if we don't restrict attention to unbiased tests? Then we can use Theorem 2 as a benchmark to establish for the best possible performance of any CI at some given θ_1. Specifically, by Theorem 2, the confidence set with the shortest expected length when $\theta = \theta_1$ is given by inverting UMP tests of the null $H_0\colon \theta = \theta_0$ against the alternative $H_1\colon \theta = \theta_1$. By the Neyman-Pearson lemma, these are one-sided z-tests that reject if $\overline{X}_n - \theta_0 \geq \zeta$ if $\theta_0 < \theta_1$, and reject if $\overline{X}_n - \theta_0 \leq -\zeta$ if $\theta_0 > \theta_1$, where $\zeta = z_{1-\alpha}\sigma/\sqrt{n}$. This means that the CI consists of all values of θ that either satisfy ($\overline{X}_n - \zeta < \theta$ and $\theta < \theta_1$), or else satisfy ($\overline{X}_n + \zeta \geq \theta$ and $\theta > \theta_1$). We may write this as

$$\{\theta\colon \overline{X}_n - \zeta < \theta < \theta_1 \quad \text{or} \quad \theta_1 \leq \theta \leq \overline{X}_n + \zeta\}$$

Based on the value of $\overline{X}_n$, it may be the case that either both sets of inequalities may hold at once, or else one of them cannot hold. Breaking down the possible cases, we obtain the CI

$$\begin{cases} [\overline{X}_n - \zeta, \theta_1] & \text{if } \overline{X}_n + \zeta \leq \theta_1, \\ [\overline{X}_n - \zeta, X + \zeta] & \text{if } \overline{X}_n - \zeta \leq \theta_1 \leq \overline{X}_n + \zeta, \\ [\theta_1, \overline{X}_n + \zeta] & \text{if } \overline{X}_n - \zeta \geq \theta_1, \end{cases} \tag{2}$$

It then follows by some algebra (fill it in!) that the expected length of the CI when $\theta = \theta_1$ is

$$\frac{2\sigma}{\sqrt{n}}\left[z_{1-\alpha}(1-2\alpha) + E[(Z + z_{1-\alpha})\,\mathbb{1}\{Z > z_{1-\alpha}\}]\right] = \frac{2\sigma}{\sqrt{n}}\left[z_{1-\alpha}(1-\alpha) + \varphi(z_{1-\alpha})\right].$$

where Z is standard normal and φ is the standard normal probability density function (PDF). In contrast, the length of the CI in eq. (1) is $2\sigma/\sqrt{n} \cdot z_{1-\alpha/2}$. Thus, when $\theta = \theta_1$, the efficiency of the CI in eq. (1) is

$$\frac{z_{1-\alpha}(1-\alpha) + \varphi(z_{1-\alpha})}{z_{1-\alpha/2}},$$

which equals 84.99% when $\alpha = 0.05$. Thus, the CI in eq. (1) is "nearly" efficient in the sense that it is not possible to improve upon in (in terms of expected length) by more than 15% even if we correctly "direct power" at the right θ.

Note that the interval in eq. (2) can be arbitrarily long if $\overline{X}_n$ is far from θ_1. So this interval is not something that we'd want to use in practice: we get a CI that is slightly shorter when $\overline{X}_n$ is close to θ_1 at the expense of being very long when $\overline{X}_n$ is far from θ_1. ⊠

3. ASYMPTOTIC THEORY FOR INTERVAL CONSTRUCTION

Instead of inverting small-sample valid tests, we can also invert tests which have a large-sample justification. Here we consider inverting the trinity of likelihood-based tests: the Wald, Lagrange multiplier (LM) and likelihood ratio (LR) tests.

Let $X_1, \ldots, X_n$ be a random sample from distribution $f(x \mid \theta)$ with $\theta \in \Theta$. Under some regularity conditions,

$$\sqrt{n}(\hat{\theta}_{ML} - \theta) \Rightarrow N(0, \mathcal{I}^{-1}(\theta)).$$

For any function $g\colon \Theta \to \mathbb{R}$, under some regularity conditions, by delta-method

$$\sqrt{n}(g(\hat{\theta}_{ML}) - g(\theta)) \Rightarrow N(0, \nabla g(\theta)'\mathcal{I}^{-1}(\theta)\nabla g(\theta)).$$

We can consistently estimate the asymptotic variance $(g'(\theta))^2\mathcal{I}^{-1}(\theta)$ by

$$\hat{V} = \nabla g(\hat{\theta}_{ML})' \left(-\nabla^2 \ell_n(\hat{\theta}_{ML})/n\right)^{-1} \nabla g(\hat{\theta}_{ML}).$$

Then, by Slutsky theorem,

$$\frac{\sqrt{n}(g(\hat{\theta}_{ML}) - g(\theta))}{\hat{V}^{1/2}} \Rightarrow \mathcal{N}(0,1),$$

so that we can construct an approximate CI for $g(\theta)$ as

$$g(\hat{\theta}_{ML}) \pm z_{1-\alpha/2}\hat{V}^{1/2}/\sqrt{n}.$$

The quantity $\hat{V}^{1/2}/\sqrt{n}$ is called the *standard error*.

Note that this confidence set is constructed based on inverting the second version of the Wald statistic that we considered in the previous lecture note,

$$\tilde{\zeta}_{Wald}(g_0) = \frac{n(g(\hat{\theta}_{ML}) - g_0)^2}{\hat{V}}.$$

Digression (Uniform size control). Because the Wald test $\phi_{g_0,Wald}(X_1,\ldots,X_n) = \mathbb{1}\{\tilde{\zeta}_{Wald} \geq z^2_{1-\alpha/2}\}$ controls size, so that

$$\lim_{n\to\infty} E_{\theta:\, g(\theta)=g_0}[\phi_{g_0,Wald}] \leq \alpha,$$

it follows that for each θ, $\lim_{n\to\infty} P_\theta(g(\theta) \in C(X)) = 1-\alpha$, and therefore that

$$\inf_\theta \lim_{n\to\infty} P_\theta(g(\theta) \in C(X)) = 1-\alpha$$

(the CI achieves nominal coverage in large samples pointwise in θ). However, it does not immediately follow that

$$\lim_{n\to\infty} \inf_{\theta\in\Theta} P_\theta(g(\theta) \in C(X)) = 1-\alpha$$

(i.e. that for n large enough, the level is approximately $1-\alpha$, this means that the CI achieves nominal coverage in large samples *uniformly* in θ), although it's possible to show that under additional regularity conditions (which, however, fail in some important models). ⊠

Example 4. Let $X_1,\ldots,X_n$ be a random sample from Bernoulli(θ). Suppose we want to construct a confidence set for the odds ratio $g(\theta) = \theta/(1-\theta)$. We have $\hat{\theta}_{ML} = \overline{X}_n$, and

$$\sqrt{n}(\hat{\theta}_{ML} - \theta) \Rightarrow N(0, \theta(1-\theta)).$$

In addition,

$$g'(\theta) = \frac{(1-\theta)+\theta}{(1-\theta)^2} = \frac{1}{(1-\theta)^2}.$$

By the delta method,

$$\sqrt{n}(g(\hat{\theta}_{ML}) - g(\theta)) \Rightarrow N(0, \theta/(1-\theta)^3).$$

So, $\hat{V} = \hat{\theta}_{ML}/(1-\hat{\theta}_{ML})^3$. Thus, an approximate confidence interval for $\theta/(1-\theta)$ is

$$\left[\frac{\hat{\theta}_{ML}}{1-\hat{\theta}_{ML}} - z_{1-\alpha/2}\sqrt{\frac{\hat{\theta}_{ML}}{(1-\hat{\theta}_{ML})^3 n}}, \frac{\hat{\theta}_{ML}}{1-\hat{\theta}_{ML}} + z_{1-\alpha/2}\sqrt{\frac{\hat{\theta}_{ML}}{(1-\hat{\theta}_{ML})^3 n}}\right]. \tag{3}$$

⊠

In addition to the Wald statistic, we can invert tests based on the LM and the LR statistics as well. However, these confidence sets are usually more involved.

Example 4 (continued). Let's consider a CI for θ. The CI based on inverting a Wald test is given by

$$\hat{\theta}_{ML} \pm z_{1-\alpha/2}\sqrt{\hat{\theta}_{ML}(1-\hat{\theta}_{ML})/n}.$$

To derive a CI based on inverting the LR test, note that the joint log-likelihood is

$$\ell_n(\theta) = \log\left(\theta^{\sum X_i}(1-\theta)^{n-\sum X_i}\right) = n\overline{X}_n \log(\theta) + n(1-\overline{X}_n)\log(1-\theta).$$

So

$$\begin{aligned}\zeta_{LR}(\theta) &= 2n\left[\overline{X}_n \log(\hat{\theta}_{ML}/\theta) + (1-\overline{X}_n)\log((1-\hat{\theta}_{ML})/(1-\theta))\right] \\ &= 2n\left[\overline{X}_n \log(\overline{X}_n/\theta) + (1-\overline{X}_n)\log((1-\overline{X}_n)/(1-\theta))\right]\end{aligned}$$

This yields the CI

$$\{\theta \in [0,1]\colon 2n\left[\overline{X}_n \log(\overline{X}_n/\theta) + (1-\overline{X}_n)\log((1-\overline{X}_n)/(1-\theta))\right] \le z^2_{1-\alpha/2}\} \tag{4}$$

It is the solution to a nonlinear inequality.

To derive a CI based on inverting the LM test, note

$$\mathcal{S}_n(\theta) = n\left(\frac{\overline{X}_n}{\theta} - \frac{1-\overline{X}_n}{1-\theta}\right) = n\frac{\overline{X}_n - \theta}{\theta(1-\theta)}, \qquad \mathcal{I}(\theta) = \frac{1}{\theta(1-\theta)}.$$

Thus,

$$\zeta_{LM}(\theta_0) = \mathcal{S}_n(\theta_0)\mathcal{I}(\theta_0)^{-1}\mathcal{S}_n(\theta_0)/n = \frac{n(\overline{X}_n - \theta)^2}{\theta(1-\theta)}.$$

We know that $\zeta_{LM}(\theta) \Rightarrow \chi^2_1$. So, the confidence set based on inverting the *LM* test is

$$\begin{aligned}&\left\{\theta \in [0,1]\colon n(\overline{X}_n - \theta)^2 \le z^2_{1-\alpha/2}\theta(1-\theta)\right\} \\ &\qquad = \left\{\theta \in [0,1]\colon (1 + z^2_{1-\alpha/2}/n)\theta^2 - (2\overline{X}_n + z^2_{1-\alpha/2}/n)\theta + \overline{X}^2_n \le 0\right\}.\end{aligned} \tag{5}$$

It is the solution to a quadratic inequality.

What about CIs for $g(\theta) = \theta/(1-\theta)$? The LR statistic is invariant to reparametrization. Further, since the restricted value of θ under the null $H_0\colon g(\theta) = g_0$ is $\theta_0 = g^{-1}(g_0) = g_0/(1+g_0)$, the CI is given by

$$\{g_0 \in (0,\infty)\colon \zeta_{LR}(g^{-1}(g_0)) \le z^2_{1-\alpha/2}\}.$$

Similarly, for the LM test, since the restricted value of θ under the null is $g_0/(1+g_0)$, the

CI is given by

$$\{g_0 \in (0,\infty)\colon \zeta_{LM}(g^{-1}(g_0)) \leq z^2_{1-\alpha/2}\} \\ = \left\{g_0 \in (0,\infty)\colon n(\overline{X}_n(1+g_0)-g_0)^2 \leq z^2_{1-\alpha/2} g_0\right\}.$$

Notice that, for both the LM and LR sets, the confidence set is a *projection* of the confidence set for θ. That is, we could equivalently first construct a CI for θ based on eq. (4) or eq. (5), and then collect all values of $g(\theta)$ that fall inside this CI. We would end up with the same result. However, projecting the Wald CI for θ yields the CI

$$\left[\frac{\hat{\theta}_{ML} - z_{1-\alpha/2}\sqrt{\hat{\theta}_{ML}(1-\hat{\theta}_{ML})/n}}{1-\hat{\theta}_{ML}+z_{1-\alpha/2}\sqrt{\hat{\theta}_{ML}(1-\hat{\theta}_{ML})/n}}, \frac{\hat{\theta}_{ML} + z_{1-\alpha/2}\sqrt{\hat{\theta}_{ML}(1-\hat{\theta}_{ML})/n}}{1-\hat{\theta}_{ML}-z_{1-\alpha/2}\sqrt{\hat{\theta}_{ML}(1-\hat{\theta}_{ML})/n}}\right],$$

which is different from the Wald CI for $g(\theta)$ we derived in eq. (3). This is a consequence of the Wald statistic not being invariant to reparametrization. ⊠

[[TODO: Plot them! Perhaps also compare with the optimal CI, based on inverting UMP unbiased test]].

4. BERNSTEIN-VON MISES

While the form of the confidence sets differs in small samples, one can show that CIs based on inverting Wald, LM and LR tests are all asymptotically equivalent. How do they compare to Bayesian credible intervals?

Theorem 4. Suppose that the conditions for asymptotic normality of $\hat{\theta}_{ML}$ hold. Consider the posterior distribution for the rescaled and recentered parameter $\zeta = \sqrt{n}(\theta - \theta_0)$ under a continuous prior distribution $\Pi(\theta)$, such that the prior density $\pi(\theta)$ is continuous at θ_0. Then the posterior distribution for ζ converges in probability to the distribution $\mathcal{N}(\sqrt{n}(\hat{\theta}_{ML}-\theta_0), \mathcal{I}(\theta_0)^{-1})$.

The convergence in the theorem is in total variation distance, that is $\sup_A |\Pi(\zeta \in A \mid X) - P(A \in \mathcal{N}(\sqrt{n}(\hat{\theta}_{ML}-\theta_0), \mathcal{I}(\theta_0)^{-1})) \xrightarrow{p} 0$. This notion of convergence is needed because both the posterior distribution, and the normal approximating distributions are random: the former depends on the data, and the latter depends on $\hat{\theta}_{ML}$, which is random. To talk about two random distributions being close in probability, we need to convert the difference between them into a single number, which is what the total variation distance does. If the two distributions are continuous with PDFs g_n and f_n, respectively, then the statement is equivalent to $\int |g_n(x) - f_n(x)| dx \xrightarrow{p} 0$.

Proof. We prove the result pointwise, and show that the ratio of PDFs at a single point ζ converges to one in probability, rather showing the stronger result that the integral of the absolute values of the differences converges to zero in probability. By the change of variables formula, the posterior

density for ζ is given by $\pi(\theta_0 + \zeta/\sqrt{n} \mid X)/\sqrt{n}$, where $\pi(\theta \mid X)$ is the posterior for θ. If we scale the posterior by the posterior at θ_0, we get

$$\frac{\pi(\theta_0 + \zeta/\sqrt{n} \mid X)}{\pi(\theta_0 \mid X)} = \frac{\mathcal{L}_n(\theta_0 + \zeta/\sqrt{n})\pi(\theta_0 + \zeta/\sqrt{n})}{\mathcal{L}_n(\theta_0)\pi(\theta_0)}.$$

Since π is continuous at θ_0, $\pi(\theta_0 + \zeta/\sqrt{n})/\pi(\theta_0) \to 1$. Furthermore, the log-likelihood difference satisfies, by a Taylor expansion,

$$\ell_n(\theta_0 + \zeta/\sqrt{n}) - \ell_n(\theta_0) = \zeta' \mathcal{S}_n(\theta_0)/\sqrt{n} + \frac{1}{2n}\zeta' \nabla^2 \ell_n \zeta,$$

where each element of the second derivative $\nabla^2 \ell_n$ are evaluated at intermediate values of θ that lie between θ_0 and $\theta_0 + \zeta/\sqrt{n}$. By the usual hand-waving, since for a fixed θ, $\frac{1}{n}\nabla^2 \ell_n(\theta) \xrightarrow{p} -\mathcal{I}(\theta)$, this suggests

$$\ell_n(\theta_0 + \zeta/\sqrt{n}) - \ell_n(\theta_0) = \zeta' \mathcal{S}_n(\theta_0)/\sqrt{n} - \frac{1}{2}\zeta' \mathcal{I}(\theta_0)\zeta + \text{remainder terms}. \tag{6}$$

On the other hand, if we evaluate the normal density $\mathcal{N}(\sqrt{n}(\hat{\theta}_{ML} - \theta_0), \mathcal{I}(\theta_0)^{-1})$ at a point ζ, the log-density is, given by

$$-\frac{1}{2}(\sqrt{n}(\hat{\theta}_{ML} - \theta_0) - \zeta)' \mathcal{I}(\theta_0)(\sqrt{n}(\hat{\theta}_{ML} - \theta_0) - \zeta).$$

plus a constant. In the proof of asymptotic normality in Lecture 6, we saw that up to remainder terms, $\sqrt{n}(\hat{\theta}_{ML} - \theta_0)$, equals $\mathcal{I}(\theta_0)^{-1}\mathcal{S}_n(\theta_0)/\sqrt{n}$. Plugging this into the previous display and observing that the term $-\frac{1}{2}\mathcal{S}_n(\theta_0)'/\sqrt{n}\mathcal{I}(\theta_0)^{-1}\mathcal{S}_n(\theta_0)/\sqrt{n}$ is just a constant (since it doesn't depend on ζ), we obtain that, up to remainder terms and constants the previous display equals

$$\zeta' \mathcal{S}_n(\theta_0)/\sqrt{n} - \frac{1}{2}\zeta' \mathcal{I}(\theta_0)\zeta,$$

which matches eq. (6). So the difference in the log densities at ζ converges to zero in probability.□

This is a remarkable result. The first remarkable thing is that the limit distribution does not depend on the prior: the prior gets dominated by the likelihood, since as $n \to \infty$, whatever one's prior beliefs are, they get dominated by the data. Loosely speaking, we have that for n large, the posterior for θ is given by

$$\theta \mid X \approx \mathcal{N}(\hat{\theta}, \mathcal{I}(\theta_0)^{-1}/n), \tag{7}$$

while in Lecture 6, we have proved that $\hat{\theta}_{ML} \mid \theta \approx \mathcal{N}(\theta, \mathcal{I}(\theta_0)^{-1}/n)$.

The equivalence between the distributions arises because by a Taylor expansion, in the neighborhood of the true θ_0 (for a fixed ζ), the log-likelihood is approximately quadratic in large samples; meaning that it is approximately Gaussian; as a result, the posterior is approximately Gaussian. On the other hand, as we saw in the proof of asymptotic normality of maximum likelihood estimator (MLE), $\sqrt{n}(\hat{\theta}_{ML} - \theta_0)$ will approximately equal the inverse information times the score, and the score is also Gaussian in large samples by the central limit theorem (CLT).

The second remarkable thing is that as a consequence of eq. (7), whatever Bayes

decision one makes based on the posterior, so long as the decision is sufficiently smooth in the posterior, the Bayes decision will in large samples converge to a Bayes rule based on Equation (7). In particular:

1. Bayes posterior mean and posterior median estimators will in large samples be equivalent to $\hat{\theta}_{ML}$. They inherit all the large-sample optimality properties of MLE.
2. Bayes credible sets will in large samples be equivalent to $\hat{\theta}_{ML} \pm z_{1-\alpha/2}\mathcal{I}(\theta_0)^1/n$. In particular, they will be equivalent to CIs based on inverting Wald, LM, or LR tests.

Therefore, from a frequentist perspective, Bayes procedures are optimal in large samples. Vice versa, a Bayesian reading frequentist papers will be able to approximately interpret likelihood-based CIs as credible sets, and MLEs as approximate posterior means.

ONE-PARAGRAPH SUMMARY OF THIS COURSE In Lecture 7, we showed that in finite-samples admissible rules are Bayes, and the reverse also. Theorem 4 tells us that Bayes rules are optimal in large samples, and equivalent to likelihood-based inference. In regular parametric models, there is no wedge in large samples between optimal frequentist and Bayesian methods. But Bayesian methods have the additional advantage that they are optimal in finite samples in terms of average risk. Furthermore, in regular parametric models, estimation and inference just boil down to inference on θ based on a single normal observation $\mathcal{N}(\theta, \mathcal{I}(\theta_0)^{-1}/n)$. This is the basis for the large-sample optimality statements about likelihood-based inference.

This raises the question of why bother with frequentist inference at all. In fully parametric models that are regular, there is arguably very little reason. But more generally, I would argue that there are two reasons for using frequentist methods, in addition to the reason that not all models are regular. The first is that often, making the model fully parametric is hard. For instance, in a regression, one may want to allow for clustered and heteroskedastic errors, but modeling the heteroskedasticity is hard business. As you'll see in the second half of the course, frequentists can appeal to large-sample results to avoid such modeling. The second reason is that we are often worried about model misspecification, and we want to make sure that our procedures are interpretable even if the model is wrong. In fact, this is in my view the main reason why linear regression is useful and should be used—more on that in 539B. In contrast, allowing for misspecification in a Bayesian world is quite a bit more involved.

5. (OPTIONAL) EXCITING NEW DEVELOPMENTS

This course has covered "standard" results in statistical theory. There are many recent results that are quite exciting, and that thing about statistics in a very different way. I will give here three such results. If you manage to figure out how to make use of these in econometrics, I think you'll have a thesis topic sorted out!

1. E-values. For instance, https://doi.org/10.1214/23-STS894
2. The HulC https://doi.org/10.1093/jrsssb/qkad134 and universal inference https://doi.org/10.1073/pnas.1922664117
3. Conformal prediction. For instance, https://doi.org/10.1007/978-3-031-06649-8

REFERENCES

Pratt, John W. 1961. "Length of Confidence Intervals." *Journal of the American Statistical Association* 56, no. 295 (September): 549–567. https://doi.org/10.2307/2282079.