

Lecture 24: October 30

Lecturer: Siva Balakrishnan

Today we will continue discussing confidence sets and ways to construct them. We have discussed point estimation so far where the goal is to construct an estimate $\hat{\theta}(X_1, \dots, X_n)$ of some parameter θ after observing $\{X_1, \dots, X_n\}$.

As we discussed in the last lecture: the setting is that we have a statistical model (i.e. a collection of distributions) \mathcal{P} . Let $C_n(X_1, \dots, X_n)$ be a set constructed using the observed data X_1, \dots, X_n . This is a random set. C_n is a $1 - \alpha$ confidence set for a parameter θ if:

$$P(\theta \in C_n(X_1, \dots, X_n)) \geq 1 - \alpha, \quad \text{for all } P \in \mathcal{P}.$$

This means that no matter which distribution in \mathcal{P} generated the data, the interval guarantees the coverage property described above. Some people would refer to such intervals as *honest* confidence intervals to make explicit the fact that the coverage is uniform over the model.

At a high-level, the confidence interval gives us some idea of how precise our estimate of the unknown parameter is, i.e. a wide interval indicates that our (point) estimate is imprecise.

Example: Suppose that we considered, $X_1, \dots, X_n \sim U[0, \theta]$. Then we could construct the usual point estimate $\hat{\theta} = X_{(n)}$. We could perhaps consider two types of confidence intervals:

$$\begin{aligned} C_1 &= [a_1 \hat{\theta}, b_1 \hat{\theta}], \quad 1 \leq a_1 \leq b_1 \\ C_2 &= [\hat{\theta} + a_2, \hat{\theta} + b_2], \quad a_2, b_2 \geq 0. \end{aligned}$$

Let us try to calculate the coverage probabilities of these two types of intervals. As a preliminary observe that:

$$\mathbb{P}(\hat{\theta} \leq t) = \left(\frac{t}{\theta}\right)^n, \quad \text{for } 0 \leq t \leq \theta.$$

1. C_1 : We can compute that,

$$\begin{aligned} \mathbb{P}(\theta \in C_1) &= \mathbb{P}(\hat{\theta} \leq \theta/a_1, \hat{\theta} \geq \theta/b_1) \\ &= \mathbb{P}(\hat{\theta} \leq \theta/a_1) - \mathbb{P}(\hat{\theta} \leq \theta/b_1) \\ &= \left(\frac{1}{a_1}\right)^n - \left(\frac{1}{b_1}\right)^n. \end{aligned}$$

So we have that for instance choosing $a_1 = 1$, $b_1 = \left(\frac{1}{\alpha}\right)^{1/n}$ guarantees us that this confidence interval has coverage probability exactly $1 - \alpha$.

2. C_2 : Similarly we have that,

$$\begin{aligned}\mathbb{P}(\theta \in C_2) &= \mathbb{P}(\hat{\theta} \leq \theta - a_2, \hat{\theta} \geq \theta - b_2) \\ &= \mathbb{P}(\hat{\theta} \leq \theta - a_2) - \mathbb{P}(\hat{\theta} \leq \theta - b_2) \\ &= \left(\frac{\theta - a_2}{\theta}\right)^n - \left(\frac{\theta - b_2}{\theta}\right)^n.\end{aligned}$$

Notice that now the coverage probability depends on the unknown parameter θ (which is undesirable). Furthermore, if we choose any constants (a_2, b_2) (say depending only on the desired coverage probability α), then as $\theta \rightarrow \infty$ we have that the interval has coverage probability that tends to 0, i.e. the interval is not honest for any constants (a_2, b_2) .

We will now discuss a few different ways of constructing confidence intervals. Although superficially different most of these methods are roughly the same.

24.1 Inverting a test

We discussed this method in the last lecture. We suppose that we have a (family of) test(s) for the hypotheses:

$$\begin{aligned}H_0 &: \theta = \theta_0 \\ H_1 &: \theta \neq \theta_0.\end{aligned}$$

These tests have a rejection region and a corresponding acceptance region (where we fail to reject the null). Denote the acceptance region for the test of $\theta = \theta_0$ as $A(\theta_0)$. This is a subset of the sample space.

Given observed data $\{X_1, \dots, X_n\}$ we consider the random set:

$$C(X_1, \dots, X_n) = \{\theta_0 : \{X_1, \dots, X_n\} \in A(\theta_0)\}.$$

Our confidence set is simply the set of parameters θ_0 that we would fail to reject using our family of tests. If our family of tests has level α then the set $C(X_1, \dots, X_n)$ is a $1 - \alpha$ confidence set.

To see this observe that since our test controls the Type I error we have that for any parameter θ_0 ,

$$\mathbb{P}_{\theta_0}(\{X_1, \dots, X_n\} \notin A(\theta_0)) \leq \alpha,$$

so with probability at least $1 - \alpha$ we have that, $\{X_1, \dots, X_n\} \in A(\theta_0)$ and hence that $\theta_0 \in C(X_1, \dots, X_n)$.

We can also construct tests using confidence intervals, i.e. consider the test that rejects the null hypothesis $\theta = \theta_0$ if $\theta_0 \notin C(X_1, \dots, X_n)$, then if $C(X_1, \dots, X_n)$ is a $1 - \alpha$ confidence interval this test has level α , i.e.

$$\mathbb{P}_{\theta_0}(\text{reject null } \theta = \theta_0) = \mathbb{P}_{\theta_0}(\theta_0 \notin C(X_1, \dots, X_n)) \leq \alpha.$$

We did a simple example of this in the last lecture (constructing a confidence interval for a Gaussian mean) so let us try a more complex example today.

Example: Suppose we observe $X_1, \dots, X_n \sim \text{Exp}(\lambda)$, and want to construct a confidence interval for λ .

As our test, suppose we use the LRT, i.e. we define the likelihood ratio:

$$\Lambda = \frac{\lambda_0^n \exp(-\lambda_0 \sum_i X_i)}{(\frac{1}{\bar{X}})^n \exp(-n)}.$$

The acceptance region has the form:

$$A(\lambda_0) = \left\{ \{X_1, \dots, X_n\} : \left(\lambda_0 \sum X_i \right)^n \exp(-\lambda_0 \sum X_i) \geq k_\alpha(\lambda_0) \right\},$$

where $k_\alpha(\lambda_0)$ needs to be chosen appropriately to control the Type I error. Observe that since $X_i \sim \text{Exp}(\lambda_0)$, $\lambda_0 X_i \sim \text{Exp}(1)$, so $k_\alpha(\lambda_0)$ does not depend on λ_0 . Once we determine the cut-off we would obtain the confidence interval by collecting:

$$C(X_1, \dots, X_n) = \left\{ \lambda : \left(\lambda \sum X_i \right)^n \exp(-\lambda \sum X_i) \geq k_\alpha \right\},$$

which is an expression that can be solved numerically. Determining k_α and then finding the confidence set can be quite tedious to do exactly (see the Casella and Berger book) and an alternative would be to use large-sample (asymptotic approximations).

24.2 Inverting Probability Inequalities

In some simple cases, we can use tail bounds to derive confidence intervals. These typically have the advantage of being exact, finite-sample intervals. However, they are rarely used in practice for many reasons including: (1) we do not always have tail bounds for estimators of interest (2) there are usually imprecisely known constants in tails bounds (3) related to (2) they are often very conservative (i.e. the intervals are often too wide to be useful).

Here are a couple of examples:

Example 24.1 Let $X_1, \dots, X_n \sim \text{Bernoulli}(p)$. By Hoeffding's inequality:

$$\mathbb{P}(|\hat{p} - p| > \epsilon) \leq 2e^{-2n\epsilon^2}.$$

Let

$$\epsilon_n = \sqrt{\frac{1}{2n} \log \left(\frac{2}{\alpha} \right)}.$$

Then

$$\mathbb{P} \left(|\hat{p} - p| > \sqrt{\frac{1}{2n} \log \left(\frac{2}{\alpha} \right)} \right) \leq \alpha.$$

Hence, $\mathbb{P}(p \in C) \geq 1 - \alpha$ where $C = (\hat{p} - \epsilon_n, \hat{p} + \epsilon_n)$.

Example 24.2 Let $X_1, \dots, X_n \sim F$. Suppose we want a **confidence band** for F . We can use VC theory. Remember that

$$\mathbb{P} \left(\sup_x |F_n(x) - F(x)| > \epsilon \right) \leq 2e^{-2n\epsilon^2}.$$

Let

$$\epsilon_n = \sqrt{\frac{1}{2n} \log \left(\frac{2}{\alpha} \right)}.$$

Then

$$\mathbb{P} \left(\sup_x |F_n(x) - F(x)| > \sqrt{\frac{1}{2n} \log \left(\frac{2}{\alpha} \right)} \right) \leq \alpha.$$

Hence,

$$P_F(L(t) \leq F(t) \leq U(t) \text{ for all } t) \geq 1 - \alpha$$

for all F , where

$$L(t) = \hat{F}_n(t) - \epsilon_n, \quad U(t) = \hat{F}_n(t) + \epsilon_n.$$

We can improve this by taking

$$L(t) = \max \left\{ \hat{F}_n(t) - \epsilon_n, 0 \right\}, \quad U(t) = \min \left\{ \hat{F}_n(t) + \epsilon_n, 1 \right\}.$$

24.3 Pivots

This is something that is known as pivotal inference in statistics. This idea is really motivated by our first example where we constructed two different confidence intervals for the uniform parameter θ . One of them was well-behaved (i.e. we could easily get it to have the right coverage probability) and the other one failed (i.e. we needed to set the length of the interval to depend on the unknown parameter θ).

A function $Q(X_1, \dots, X_n, \theta)$ is a *pivot* if the distribution of Q does not depend on θ . For example, if $X_1, \dots, X_n \sim N(\theta, 1)$ then

$$\bar{X}_n - \theta \sim N(0, 1/n)$$

so $Q = \bar{X}_n - \theta$ is a pivot.

Let a and b be such that

$$P_\theta(a \leq Q(X, \theta) \leq b) \geq 1 - \alpha$$

for all θ . We can find such an a and b because Q is a pivot. It follows immediately that

$$C(X) = \{\theta : a \leq Q(X, \theta) \leq b\}$$

has coverage $1 - \alpha$.

Example 24.3 Let $X_1, \dots, X_n \sim N(\mu, \sigma^2)$. (σ known.) Then

$$Z = \frac{\sqrt{n}(\bar{X} - \mu)}{\sigma} \sim N(0, 1).$$

We know that

$$P(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}) = 1 - \alpha$$

and so

$$P\left(-z_{\alpha/2} \leq \frac{\sqrt{n}(\bar{X} - \mu)}{\sigma} \leq z_{\alpha/2}\right) = 1 - \alpha.$$

Thus

$$C = \bar{X} \pm \frac{\sigma}{\sqrt{n}} z_{\alpha/2}.$$

If σ is unknown, then this becomes

$$C = \bar{X} \pm \frac{S}{\sqrt{n}} t_{n-1, \alpha/2}$$

because

$$T = \frac{\sqrt{n}(\bar{X} - \mu)}{S} \sim t_{n-1}.$$

We can similarly re-interpret the uniform interval we created in terms of a pivot.

Example 24.4 Let $X_1, \dots, X_n \sim \text{Uniform}(0, \theta)$. Let $Q = X_{(n)}/\theta$. Then

$$\mathbb{P}(Q \leq t) = \prod_i \mathbb{P}(X_i \leq t\theta) = t^n$$

so Q is a pivot. Let $c = \alpha^{1/n}$. Then

$$\mathbb{P}(Q \leq c) = \alpha.$$

Also, $\mathbb{P}(Q \leq 1) = 1$. Therefore,

$$\begin{aligned} 1 - \alpha &= \mathbb{P}(c \leq Q \leq 1) = \mathbb{P}\left(c \leq \frac{X_{(n)}}{\theta} \leq 1\right) \\ &= \mathbb{P}\left(\frac{1}{c} \geq \frac{\theta}{X_{(n)}} \geq 1\right) \\ &= \mathbb{P}\left(X_{(n)} \leq \theta \leq \frac{X_{(n)}}{c}\right) \end{aligned}$$

so a $1 - \alpha$ confidence interval is

$$\left(X_{(n)}, \frac{X_{(n)}}{\alpha^{1/n}}\right).$$

24.4 Large Sample Intervals

The Wald Interval. We know that, under regularity conditions,

$$\frac{\hat{\theta}_n - \theta}{\text{se}} \rightsquigarrow N(0, 1)$$

where $\hat{\theta}_n$ is the mle and $\text{se} = 1/\sqrt{I_n(\hat{\theta})}$. So this is an asymptotic pivot and an approximate confidence interval is

$$\hat{\theta}_n \pm z_{\alpha/2} \text{se}.$$

By the delta method, a confidence interval for $\tau(\theta)$ is

$$\tau(\hat{\theta}_n) \pm z_{\alpha/2} \text{se}(\hat{\theta}) |\tau'(\hat{\theta}_n)|.$$

The Likelihood-Based Confidence Set. We saw an example before of trying to exactly invert the LRT. This is often difficult to do. Let's consider inverting the asymptotic LRT. We test

$$H_0 : \theta = \theta_0 \quad \text{versus} \quad H_1 : \theta \neq \theta_0.$$

Let k be the dimension of θ . We don't reject if

$$-2 \log \left(\frac{L(\theta_0)}{L(\hat{\theta})} \right) \leq \chi_{k,\alpha}^2$$

that is, if

$$\frac{L(\theta_0)}{L(\hat{\theta})} > e^{-\chi_{k,\alpha}^2/2}.$$

So, the set of non-rejected nulls is

$$C_n = \left\{ \theta : \frac{L(\theta)}{L(\hat{\theta})} > e^{-\frac{\chi_{k,\alpha}^2}{2}} \right\}.$$

This confidence set has a somewhat pleasing interpretation that it is just a collection of parameters that have high likelihood. Then

$$P_\theta(\theta \in C) \rightarrow 1 - \alpha$$

for each θ .

Example 24.5 Let $X_1, \dots, X_n \sim \text{Bernoulli}(p)$, let $S = \sum X_i$. Using the Wald statistic

$$\frac{\hat{p} - p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} \rightsquigarrow N(0, 1)$$

so an approximate confidence interval is

$$\hat{p} \pm z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}.$$

Using the LRT we get

$$C = \left\{ p : -2 \log \left(\frac{p^S (1-p)^{n-S}}{\hat{p}^S (1-\hat{p})^{n-S}} \right) \leq \chi_{1,\alpha}^2 \right\}.$$

These intervals are different but, for large n , they are nearly the same.

24.5 Multiple Confidence Intervals

When we have multiple parameters we might want to construct intervals C_1, \dots, C_d such that we control:

$$\mathbb{P}(\exists j \text{ such that } \theta_j \notin C_j) \leq \alpha.$$

In this case, we could just Bonferroni correct our confidence intervals, i.e. we could construct confidence intervals which have coverage probability α/d , and these would then have the desired property. This is analogous to controlling the FWER.

It is worth pondering what the analog of FDR control might be for confidence intervals (maybe look up False Coverage-statement Rate).

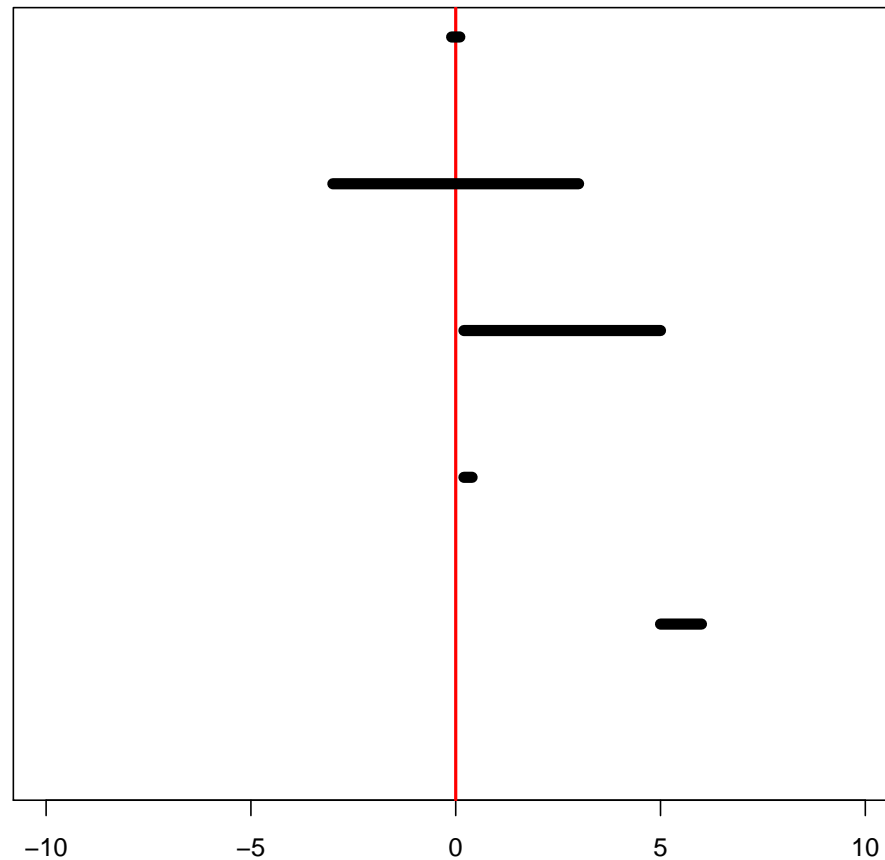


Figure 24.1: Five examples: 1. Not significant, precise. 2. Not significant, imprecise. 3. Barely significant, imprecise. 4. Barely significant, precise. 5. Significant and precise.

24.6 Tests Versus Confidence Intervals

Confidence intervals are more informative than tests. Intuitively, p-values are more informative than an accept/reject decision because it summarizes all the significance levels for which we would reject the null hypothesis. Similarly, a confidence interval is more informative than a test because it summarizes all the parameters for which we would (fail to) reject the null hypothesis. More practically, a confidence interval tells us something about the “effect size” as well as something about the uncertainty in our estimate of the “effect size”.

Look at Figure 24.1. Suppose we are testing $H_0 : \theta = 0$ versus $H_1 : \theta \neq 0$. We see 5 different confidence intervals. The first two cases (top two) correspond to not rejecting H_0 . The other three correspond to rejecting H_0 . Reporting the confidence intervals is much more informative than simply reporting “reject” or “don’t reject.”