STATS 200: Introduction to Statistical Inference **Autumn 2016**

Lecture 18 — Confidence intervals

We have seen how to understand the variability of an estimate $\hat{\theta}$ for a parameter θ, or of $g(\hat{\theta})$ for a quantity $g(\theta)$, in terms of its sampling distribution and its standard error. This understanding may be used to construct a confidence interval for θ or $g(\theta)$.

18.1 Exact confidence intervals

In a parametric model, let $g(\theta)$ be any quantity of interest (which might be the parameter θ itself). Informally, a **confidence interval** for $g(\theta)$ is a random interval calculated from the data that contains this value $g(\theta)$ with a specified probability. For example, a 90% confidence interval contains $g(\theta)$ with probability 0.9, and a 95% confidence interval contains $g(\theta)$ with probability 0.95. (If we construct 100 different 90% confidence intervals for θ using 100 independent sets of data, then we would expect about 90 of them to contain θ.)

More formally, what this means is the following: Let $X_1, \ldots, X_n$ be a sample of data. By *random interval*, we mean an interval whose lower and upper endpoints $L(X_1, \ldots, X_n)$ and $U(X_1, \ldots, X_n)$ are functions of the data $X_1, \ldots, X_n$. (Hence the interval is random in the same sense that the data itself is random—a different realization of the data leads to a different interval.) The interval $[L(X_1, \ldots, X_n), U(X_1, \ldots, X_n)]$ is a $100(1-\alpha)\%$ confidence interval for $g(\theta)$ if, for *all* $\theta \in \Omega$,

$$\mathbb{P}_\theta \left[L(X_1, \ldots, X_n) \leq g(\theta) \leq U(X_1, \ldots, X_n)\right] = 1 - \alpha,$$

where $\mathbb{P}_\theta$ denotes probability under $X_1, \ldots, X_n \overset{IID}{\sim} f(x|\theta)$.

A confidence interval for $g(\theta)$ is commonly constructed from an estimate of $g(\theta)$ and an estimate of the associated standard error:

Example 18.1. Consider data $X_1, \ldots, X_n \overset{IID}{\sim} \mathcal{N}(\mu, \sigma^2)$, where both μ and σ^2 are unknown. To construct a confidence interval for μ, consider the estimate $\bar{X}$. As $\bar{X} \sim \mathcal{N}(\mu, \sigma^2/n)$, the standard error of $\bar{X}$ is $\sigma/\sqrt{n}$, which we may estimate by $S/\sqrt{n}$ where

$$S = \sqrt{\frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2}.$$

Recall from Lecture 7 that when $X_1, \ldots, X_n \overset{IID}{\sim} \mathcal{N}(\mu, \sigma^2)$, the quantity

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} = \frac{\sqrt{n}(\bar{X} - \mu)}{S}$$

has a t-distribution with $n-1$ degrees of freedom. (In Lecture 7 we assumed $\mu = 0$, but the distribution of this quantity doesn't depend on μ.) Letting $t_{n-1}(\alpha/2)$ be the upper-$\alpha/2$

point of the t_{n-1} distribution and noting that $-t_{n-1}(\alpha/2)$ is then the lower-$\alpha/2$ point by symmetry, this means

$$\mathbb{P}_{\mu,\sigma^2}\left[-t_{n-1}(\alpha/2) \leq \frac{\sqrt{n}(\bar{X}-\mu)}{S} \leq t_{n-1}(\alpha/2)\right] = 1-\alpha.$$

The upper inequality above may be rearranged as

$$\bar{X} - \frac{S}{\sqrt{n}} t_{n-1}(\alpha/2) \leq \mu,$$

and the lower inequality may be rearranged as

$$\mu \leq \bar{X} + \frac{S}{\sqrt{n}} t_{n-1}(\alpha/2).$$

Hence

$$\mathbb{P}_{\mu,\sigma^2}\left[\bar{X} - \frac{S}{\sqrt{n}} t_{n-1}(\alpha/2) \leq \mu \leq \bar{X} + \frac{S}{\sqrt{n}} t_{n-1}(\alpha/2)\right] = 1-\alpha,$$

so $[\bar{X} - \frac{S}{\sqrt{n}} t_{n-1}(\alpha/2), \bar{X} + \frac{S}{\sqrt{n}} t_{n-1}(\alpha/2)]$ is a $100(1-\alpha)\%$ confidence interval for μ. We'll use the notation $\bar{X} \pm \frac{S}{\sqrt{n}} t_{n-1}(\alpha/2)$ as shorthand for this interval.

18.2 Asymptotic confidence intervals

In the previous example, we were able to construct an exact confidence interval because we knew the exact distribution of $\sqrt{n}(\bar{X}-\mu)/S$, which is t_{n-1} (and which does not depend on μ and σ^2). Suppose that we had forgotten this fact. If n is large, we could have still reasoned as follows: By the Central Limit Theorem, as $n \to \infty$,

$$\sqrt{n}(\bar{X}-\mu) \to \mathcal{N}(0,\sigma^2)$$

in distribution. By our addendum at the end of Lecture 10, $S^2 \to \sigma^2$ in probability (meaning, the sample variance S^2 is consistent for σ^2). Then, applying the Continuous Mapping Theorem and Slutsky's Lemma,

$$\frac{\sqrt{n}(\bar{X}-\mu)}{S} = \frac{\sigma}{S} \times \frac{\sqrt{n}(\bar{X}-\mu)}{\sigma} \to \mathcal{N}(0,1)$$

in distribution, so

$$\mathbb{P}_{\mu,\sigma^2}\left[-z(\alpha/2) \leq \frac{\sqrt{n}(\bar{X}-\mu)}{S} \leq z(\alpha/2)\right] \to 1-\alpha$$

as $n \to \infty$. Rearranging the inequalities above in the same way as the previous example yields a $100(1-\alpha)\%$ **asymptotic confidence interval** $\bar{X} \pm \frac{S}{\sqrt{n}} z(\alpha/2)$ for μ. We expect this interval to be accurate (meaning its coverage of μ is close to $100(1-\alpha)\%$) for large n—indeed, for large n, $z(\alpha/2) \approx t_{n-1}(\alpha/2)$ because the t_{n-1} distribution is very close to the standard normal distribution, so that this interval is almost the same as the exact interval of the previous example.

This method may be applied to construct an approximate confidence interval from any asymptotically normal estimator, as we will see in the following examples.

Example 18.2. Let $X_1, \ldots, X_n \overset{IID}{\sim}$ Poisson(λ). To construct an asymptotic confidence interval for λ, let's start with the estimator $\hat{\lambda} = \bar{X}$. By the Central Limit Theorem,

$$\sqrt{n}(\hat{\lambda} - \lambda) \to \mathcal{N}(0, \lambda).$$

We don't know the variance λ of this limiting normal distribution, but we can estimate it by $\hat{\lambda}$. By the Law of Large Numbers, $\hat{\lambda} \to \lambda$ in probability as $n \to \infty$, i.e. $\hat{\lambda}$ is consistent for λ. Then by the Continuous Mapping Theorem and Slutsky's Lemma,

$$\frac{\sqrt{n}(\hat{\lambda} - \lambda)}{\sqrt{\hat{\lambda}}} = \frac{\sqrt{\lambda}}{\sqrt{\hat{\lambda}}} \times \frac{\sqrt{n}(\hat{\lambda} - \lambda)}{\sqrt{\lambda}} \to \mathcal{N}(0, 1),$$

so

$$\mathbb{P}_\lambda \left[-z(\alpha/2) \leq \frac{\sqrt{n}(\hat{\lambda} - \lambda)}{\sqrt{\hat{\lambda}}} \leq z(\alpha/2) \right] \to 1 - \alpha.$$

Rearranging these inequalities yields the asymptotic $100(1 - \alpha)\%$ confidence interval $\hat{\lambda} \pm \sqrt{\frac{\hat{\lambda}}{n}} z(\alpha/2)$.

For various values of λ and n, the table below shows the simulated true probabilities that the 90% and 95% confidence intervals constructed in this way cover λ:

	Desired coverage: 90%			Desired coverage: 95%		
	$\lambda = 0.1$	$\lambda = 1$	$\lambda = 5$	$\lambda = 0.1$	$\lambda = 1$	$\lambda = 5$
$n = 10$	0.63	0.91	0.90	0.63	0.93	0.95
$n = 30$	0.79	0.89	0.90	0.80	0.93	0.95
$n = 100$	0.91	0.90	0.90	0.93	0.94	0.95

(Meaning, we simulated $X_1, \ldots, X_n \overset{IID}{\sim}$ Poisson(λ), computed the confidence interval, checked whether it contained λ, and repeated this $B = 1{,}000{,}000$ times. The table reports the fraction of simulations for which the interval covered λ.) We observe that coverage is closer to the desired levels for larger values of n, as well as for larger values of λ. For small n and/or small λ, the normal approximation to the distribution of $\hat{\lambda}$ is inaccurate, and the simulations show that we underestimate the variability of $\hat{\lambda}$.

Example 18.3. More generally, let $\{f(x|\theta) : \theta \in \Omega\}$ be any parametric model satisfying the regularity conditions of Theorem 14.1, where θ is a single parameter. To obtain a confidence interval for θ, consider the MLE $\hat{\theta}$, which satisfies

$$\sqrt{n}(\hat{\theta} - \theta) \to \mathcal{N}(0, I(\theta)^{-1})$$

as $n \to \infty$. We may estimate $I(\theta)$ by the plugin estimator $I(\hat{\theta})$. If $I(\theta)$ is continuous in θ and $\hat{\theta}$ is consistent for θ, then the Continuous Mapping Theorem implies $I(\hat{\theta}) \to I(\theta)$ in probability, and hence

$$\sqrt{nI(\hat{\theta})}(\hat{\theta} - \theta) = \frac{\sqrt{I(\hat{\theta})}}{\sqrt{I(\theta)}} \times \sqrt{nI(\theta)}(\hat{\theta} - \theta) \to \mathcal{N}(0, 1).$$

So

$$\mathbb{P}_\theta\left[-z(\alpha/2) \leq \sqrt{nI(\hat{\theta})}(\hat{\theta}-\theta) \leq z(\alpha/2)\right] \to 1-\alpha,$$

and rearranging yields the asymptotic $100(1-\alpha)\%$ confidence interval $\hat{\theta} \pm \frac{1}{\sqrt{nI(\hat{\theta})}} z(\alpha/2)$. This is oftentimes called the **Wald interval** for θ.

Example 18.4. Let $X_1, \ldots, X_n \overset{IID}{\sim} \text{Bernoulli}(p)$. Suppose we wish to construct a confidence interval for the log-odds $g(p) = \log \frac{p}{1-p}$. In Lecture 17, we showed using the delta method that

$$\sqrt{n}(g(\hat{p}) - g(p)) \to \mathcal{N}\left(0, \frac{1}{p(1-p)}\right),$$

where $\hat{p} = \bar{X}$. Since $\hat{p} \to p$ in probability, by the Continuous Mapping Theorem and Slutsky's Lemma,

$$\sqrt{n\hat{p}(1-\hat{p})}(g(\hat{p}) - g(p)) = \frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{p(1-p)}}\sqrt{np(1-p)}(g(\hat{p}) - g(p)) \to \mathcal{N}(0,1),$$

so

$$\mathbb{P}_p\left[-z(\alpha/2) \leq \sqrt{n\hat{p}(1-\hat{p})}(g(\hat{p}) - g(p)) \leq z(\alpha/2)\right] \to 1-\alpha.$$

An asymptotic $100(1-\alpha)\%$ confidence interval for the log-odds $g(p) = \log \frac{p}{1-p}$ is then

$$[L(\hat{p}), U(\hat{p})] := \left[\log \frac{\hat{p}}{1-\hat{p}} - \sqrt{\frac{1}{n\hat{p}(1-\hat{p})}} z(\alpha/2),\ \log \frac{\hat{p}}{1-\hat{p}} + \sqrt{\frac{1}{n\hat{p}(1-\hat{p})}} z(\alpha/2)\right].$$

If we wish to obtain a confidence interval for the odds $\frac{p}{1-p}$ rather than the log-odds, note that $\mathbb{P}[L(\hat{p}) \leq \log \frac{p}{1-p} \leq U(\hat{p})] = \mathbb{P}[e^{L(\hat{p})} \leq \frac{p}{1-p} \leq e^{U(\hat{p})}]$, so that $[e^{L(\hat{p})}, e^{U(\hat{p})}]$ is a confidence interval for the odds. This interval is *not* symmetric around the estimate $\frac{\hat{p}}{1-\hat{p}}$, and is different from what we would have obtained if we instead applied the delta method directly to $g(p) = \frac{p}{1-p}$. The interval $[e^{L(\hat{p})}, e^{U(\hat{p})}]$ for the odds is typically used in practice—the distribution of $\log \frac{\hat{p}}{1-\hat{p}}$ is less skewed than that of $\frac{\hat{p}}{1-\hat{p}}$ for small to moderate n, so the normal approximation and resulting confidence interval are more accurate if we consider odds on the log scale.

Let us caution that in the construction of these asymptotic confidence intervals, a number of different approximations are being made:

- The true distribution of $\sqrt{n}(\hat{\theta} - \theta)$ is being approximated by a normal distribution.
- The true variance of this normal distribution, say $I(\theta)^{-1}$, is being approximated by a plugin estimate $I(\hat{\theta})^{-1}$.
- In the case where we are interested in $g(\theta)$ and g is a nonlinear function, the value $g(\hat{\theta})$ is being approximated by the Taylor expansion $g(\theta) + (\hat{\theta} - \theta)g'(\theta)$. (This is what is done in the delta method.)

These approximations are all valid in the limit $n \to \infty$, but their accuracy is not guaranteed for the finite sample size n of any given problem. Coverage of asymptotic confidence intervals should be checked by simulation, as Example 18.2 illustrates that they might be severely overconfident for small n.