8.2: Large-Sample Confidence Intervals for a Population Mean and Proportion

Taylor

University of Virginia

Motivation

Last section we found intervals using an expression $h(X_1, \ldots, X_n, \theta)$. This quantity

- had its representation depend on only the parameter of interest θ
- had a known probability distribution (specifically, it did NOT depend on the unknown parameter) θ

Q: What if finding $h(X_1, \ldots, X_n, \theta)$ is difficult/impossible?
A: We can find an approximate probability distribution for it.

Exact Interval 1

We saw if $X_1, \ldots, X_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$, then our confidence interval for μ was based on the fact that

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

since $\bar{X} \sim \mathcal{N}(\mu, \frac{\sigma^2}{n})$

Exact Interval 2

Next section, we will see if $X_1, \ldots, X_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$, and we din't know σ^2, then our confidence interval for μ will be based on

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

This is because $\frac{\bar{X}-\mu}{s/\sqrt{n}} \sim t_{n-1}$

Motivation

What if our data are not randomly sampled from a normal distribution though (they could be strictly positive, they could have some skew in their histogram, etc.)

An easy and still highly accurate solution is to use the Central Limit Theorem (CLT). If $X_1, \ldots, X_n$ is a random sample from some other distribution, and if it has its first and second moments, then

$$\bar{X} \overset{approx}{\sim} \mathcal{N}(\mu, \frac{\sigma^2}{n})$$

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \overset{approx}{\sim} \mathcal{N}(0,1)$$

$$\frac{\bar{X} - \mu}{s/\sqrt{n}} \overset{\text{approx}}{\sim} \mathcal{N}(0,1)$$

if n is *large*

A generalization

Then they go on to generalize this a bit. They say we can make a confidence interval for θ with $\hat{\theta}$ with this:

$$P\left(-z_{\alpha/2} \leq \frac{\hat{\theta}-\theta}{\sqrt{\mathsf{Var}\left(\hat{\theta}\right)}} \leq z_{\alpha/2}\right) \approx 1-\alpha$$

if

1. $\hat{\theta}$ is approximately normal (CLT or MLE justification)
2. $\hat{\theta}$ is approximately unbiased
   - if CLT justification: compute the expectation and take the limit to find out
   - if MLE justification: asymptotic unbiasedness is guaranteed
3. the standard deviation of $\hat{\theta}$ is available: $\sigma_{\hat{\theta}}$
   - if CLT justification: compute the variance directly
   - if MLE justification: reciprocal of CRLB

Example

This example has its own section on page 395. Let $X \sim \text{Binomial}(n, p)$. Then we estimate p with $\hat{p} = \frac{X}{n}$.

We've already showed that $E\hat{p} = p$ and $SE[\hat{p}] = \sqrt{\frac{p(1-p)}{n}}$. By the previous slide, we can use that z-score-ish confidence interval.

$$P\left(-z_{\alpha/2} \leq \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \leq z_{\alpha/2}\right) \approx 1 - \alpha$$

Note, this interval is also suggested by the MLE properties: this estimator's variance achieved the CRLB, remember.

Example

I leave it to you to check that this is the large-sample (CLT) CI. Isolating p involves finding the roots of a quadratic polynomial.

$$\tilde{p} \pm z_{\alpha/2} \frac{\sqrt{\hat{p}\hat{q}/n + z^2_{\alpha/2}/4n^2}}{1 + z^2_{\alpha/2}/n}$$

where $\tilde{p} = \frac{\hat{p}+z^2_{\alpha/2}/2n}{1+z^2_{\alpha/2}/n}$ and $\hat{q} = 1 - \hat{p}$

Example 2

Let $X_1, \ldots, X_n \overset{iid}{\sim}$ Poisson(λ). By the CLT, $\frac{\bar{X}-\lambda}{\sqrt{\frac{\lambda}{n}}} \overset{approx}{\sim}$ Normal$(0,1)$

$$
\begin{aligned}
.95 &= P\left(-z_{.025} < \frac{\bar{X}-\lambda}{\sqrt{\frac{\lambda}{n}}} < z_{.025}\right) \\
&= P\left(-z_{.025}\sqrt{\frac{\lambda}{n}} < \bar{X}-\lambda < z_{.025}\sqrt{\frac{\lambda}{n}}\right) \\
&= P\left(z_{.025}^2 \frac{\lambda}{n} > (\bar{X}-\lambda)^2\right) \\
&= P\left(z_{.025}^2 \frac{\lambda}{n} > \bar{X}^2 + \lambda^2 - 2\lambda\bar{X}\right) \\
&= P\left(0 > -z_{.025}^2 \frac{\lambda}{n} + \bar{X}^2 + \lambda^2 - 2\lambda\bar{X}\right)
\end{aligned}
$$

Example 2

$$\cdots = P\left(0 > -z_{.025}^2 \frac{\lambda}{n} + \bar{X}^2 + \lambda^2 - 2\lambda\bar{X}\right)$$

$$= P\left(0 > \lambda^2 [1] + \lambda \left[-\frac{z_{.025}^2}{n} - 2\bar{X}\right] + [\bar{X}^2]\right)$$

$$= P\left(0 > a\lambda^2 + b\lambda + c\right)$$

$$= P\left(-\frac{b}{2a} - \frac{\sqrt{b^2 - 4ac}}{2a} < \lambda < -\frac{b}{2a} + \frac{\sqrt{b^2 - 4ac}}{2a}\right)$$

So our confidence interval is $\left[-\frac{b}{2a} - \frac{\sqrt{b^2-4ac}}{2a}, -\frac{b}{2a} + \frac{\sqrt{b^2-4ac}}{2a}\right]$ or

$$\left[-\frac{1}{2}\left[-\frac{z_{.025}^2}{n} - 2\bar{X}\right] \pm \frac{\sqrt{\left[-\frac{z_{.025}^2}{n} - 2\bar{X}\right]^2 - 4\bar{X}^2}}{2}\right]$$

An easier way...

And if we don't want to use $\sigma_{\hat{\theta}}$, then we can plug in its estimate $\sqrt{\widehat{\text{Var}\left(\hat{\theta}\right)}} = \widehat{\sigma_{\hat{\theta}}}$.

1. This will always be justified in this class, and we will never prove why.
2. We are using a CLT-like justification

$$P\left(-z_{\alpha/2} \leq \frac{\hat{\theta} - \theta}{\widehat{\sigma_{\hat{\theta}}}} \leq z_{\alpha/2}\right) \approx 1 - \alpha$$

...valid under 'general conditions'

Example

This is an adjustment of the last example. $X \sim \text{Binomial}(n, p)$ still, and we're still estimating p with $\hat{p} = \frac{X}{n}$. Recall $SE[\hat{p}] = \sqrt{\frac{p(1-p)}{n}}$. Now we just use *estimated* standard error (we replace all the ps with $\hat{p}$s): $\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

$$P\left(-z_{\alpha/2} \leq \frac{\hat{p} - p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} \leq z_{\alpha/2}\right) \approx 1 - \alpha$$

...and this one is a lot easier to work with (no quadratic formula or anything)

Overview of Approximate Intervals

If $\hat{\theta}$ is approximately normal, we can start with either

1. $P(-z_{\alpha/2} \leq \frac{\hat{\theta}-\theta}{\sqrt{\mathrm{Var}(\hat{\theta})}} \leq z_{\alpha/2})$
2. $P(-z_{\alpha/2} \leq \frac{\hat{\theta}-\theta}{\widehat{\sigma_{\hat{\theta}}}} \leq z_{\alpha/2})$

Two justifications for $\hat{\theta}$ being normal:

1. $\hat{\theta}$ looks like a sample mean (CLT)
2. $\hat{\theta}$ is the MLE estimator

Two ways to handle the standard error

- exact standard error and CRLB reciprocal: $\sqrt{\mathrm{Var}\left(\hat{\theta}\right)}$
- estimated standard error with plug-in estimates: $\sqrt{\widehat{\mathrm{Var}\left(\hat{\theta}\right)}}$

Definition

This chapter also shows you how to compute **one-sided confidence intervals/bounds**.

If you want a lower bound, you isolate μ inside the parentheses in this expression:

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

If you want an upper bound, you isolate μ using this:

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

...so the quantiles change, and you only need one of them instead of two.

Definition

Here we write out the **large-sample upper confidence bound for** μ

$$(-\infty, \bar{x} + z_\alpha \frac{s}{\sqrt{n}}]$$

and the **large-sample lower confidence bound for** μ

$$[\bar{x} - z_\alpha \frac{s}{\sqrt{n}}, \infty)$$

for a reference.