I05 - Confidence intervals

STAT 401 (Engineering) - Iowa State University

February 16, 2018

Confidence intervals

Definition

The coverage of an interval estimator is the probability the interval will contain the true value of the parameter *when the data are considered to be random*. If an interval estimator has $100(1-a)\%$ coverage, then we call it a $100(1-a)\%$ confidence interval and $1-a$ is the confidence level.

Normal model

If $Y_i \stackrel{ind}{\sim} N(\mu, \sigma^2)$ and we assume the default prior $p(\mu, \sigma^2) \propto 1/\sigma^2$, then a $100(1-a)\%$ credible interval for μ is given by

$$\overline{y} \pm t_{n-1,a/2} s/\sqrt{n}.$$

When the data are considered random

$$T = \frac{\overline{Y} - \mu}{S/\sqrt{n}} \sim t_{n-1}(0,1)$$

thus the probability μ is within our credible interval is

$$\begin{aligned}
&P\left(\overline{Y} - t_{n-1,a/2}S/\sqrt{n} < \mu < \overline{Y} + t_{n-1,a/2}S/\sqrt{n}\right) \\
&= P\left(-t_{n-1,a/2} < \frac{\overline{Y}-\mu}{S/\sqrt{n}} < t_{n-1,a/2}\right) \\
&= P\left(-t_{n-1,a/2} < T < t_{n-1,a/2}\right) \\
&= 1 - a.
\end{aligned}$$

Thus, this $100(1-a)\%$ credible interval is also a $100(1-a)\%$ confidence interval.

Data example

Recall the corn yield example from I04 with 9 randomly selected fields in Iowa whose sample average yield is 205 and sample standard deviation is 21. Then a 95% confidence interval for the mean corn yield on Iowa farms is

$$205 \pm 2.31 \times 21/\sqrt{9} = (189, 221).$$

This confidence interval tells us nothing about the true mean yield of fields in Iowa. Instead, it tells us that if we use this procedure repeatedly on different data sets, then 95% of the time, the interval will contain the true parameter. But because this confidence interval happens to correspond to a credible interval, it does tell us what we should believe about the true mean yield.

Data example - R code

```
y %>% round(1)

[1] 182.1 203.7 231.8 177.4 198.4 202.6 214.2 195.2 239.7

n <- length(y)
ybar <- mean(y)
s <- sd(y)
a <- .05
t_crit <- qt(1-a/2, df = n-1)
L <- ybar - t_crit*s/sqrt(n)
U <- ybar + t_crit*s/sqrt(n)
c(L,U) %>% round

[1] 189 221
```

Sampling distribution

Definition

The sampling distribution of a statistic is the distribution of the statistic *when the data are considered random*.

Example

If $Y_i \overset{ind}{\sim} N(\mu, \sigma^2)$, then

$$T = \frac{\overline{Y} - \mu}{S/\sqrt{n}} \sim t_{n-1}(0,1)$$

where $\overline{Y}$ and S are the sample average and sample standard deviation over n observations.

Most of the time we don't know the sampling distribution of the statistic.

Approximate sampling distributions

If the estimator $\hat{\theta} = \hat{\theta}(Y)$ is based on an average or a sum, then the Central Limit Theorem tells us what its approximate sampling distribution is, i.e.

$$\hat{\theta} \dot{\sim} N(E[\hat{\theta}], Var[\hat{\theta}]).$$

Example

If $Y_i \overset{ind}{\sim} N(\mu, \sigma^2)$, then

$$\hat{\mu} = \overline{Y} \dot{\sim} N(E[\overline{Y}], Var[\overline{Y}])$$

where

$$\begin{aligned} E[\overline{Y}] &= \mu \\ Var[\overline{Y}] &= \sigma^2/n. \end{aligned}$$

Standard error

Definition

A standard error of an estimator is an *estimate* of the standard deviation of the sampling distribution of the estimator.

Example

If $Y_i \overset{ind}{\sim} N(\mu, \sigma^2)$ and $\hat{\mu} = \overline{Y}$, then

$$Var[\hat{\mu}] = Var[\overline{Y}] = \sigma^2/n \quad \text{and} \quad \sqrt{Var[\hat{\mu}]} = \sigma/\sqrt{n}.$$

We have the following *consistent* estimators

$$S^2 \xrightarrow{P} \sigma^2, \quad S \xrightarrow{P} \sigma, \quad \text{and} \quad S/\sqrt{n} \xrightarrow{P} \sigma/\sqrt{n}$$

and thus

$$SE[\hat{\mu}] = \sqrt{\widehat{Var[\hat{\mu}]}} = S/\sqrt{n}.$$

where S is the sample standard deviation.

Approximate confidence intervals

If an estimator has an asymptotic normal distribution, then we can construct an approximate $100(1-a)\%$ confidence interval for $E[\hat{\theta}]$ using

$$\hat{\theta} \pm z_{a/2} SE[\hat{\theta}].$$

where $SE(\hat{\theta}) = \widehat{\sqrt{Var[\hat{\theta}]}}$ is the standard error of the estimator.

This comes from the fact that if $\hat{\theta} \dot{\sim} N(E[\hat{\theta}], SE[\hat{\theta}])$, then

$$\begin{aligned}
&P\left(\hat{\theta} - z_{a/2}SE(\hat{\theta}) < E[\hat{\theta}] < \hat{\theta} + z_{a/2}SE(\hat{\theta})\right)\\
&= P\left(-z_{a/2} < \frac{\hat{\theta}-E[\hat{\theta}]}{SE(\hat{\theta})} < z_{a/2}\right)\\
&= P\left(-z_{a/2} < \frac{\hat{\theta}-E[\hat{\theta}]}{\sqrt{Var[\hat{\theta}]}} < z_{a/2}\right)\\
&\approx P\left(-z_{a/2} < Z < z_{a/2}\right)\\
&= 1-a.
\end{aligned}$$

Normal example

If $Y_i \stackrel{ind}{\sim} N(\mu, \sigma^2)$ and we have the estimator $\hat{\mu} = \overline{Y}$, then

$$\begin{aligned} E[\hat{\mu}] &= \mu \\ SE[\hat{\mu}] &= S/\sqrt{n} \end{aligned}$$

Thus an approximate $100(1-a)\%$ confidence interval for $\mu = E[\hat{\mu}]$ is

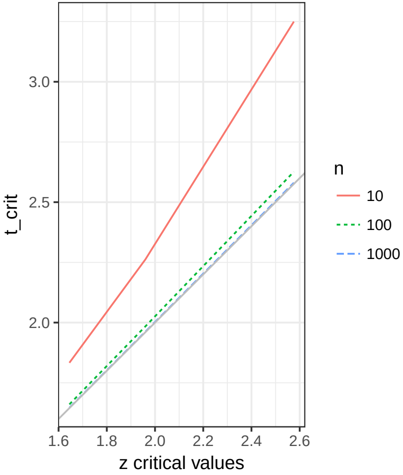
$$\hat{\mu} \pm z_{a/2} SE[\hat{\mu}] = \overline{Y} \pm z_{a/2} S/\sqrt{n}.$$

Note that this is almost identical to the exact $100(1-a)\%$ confidence interval for μ,

$$\overline{Y} \pm t_{n-1,a/2} S/\sqrt{n}$$

and when n is large $z_{a/2} \approx t_{n-1,a/2}$.

T critical values vs Z critical values



Binomial example

Suppose $Y \sim Bin(n, \theta)$ and we are interested in a confidence interval for θ. An unbiased estimator for θ is

$$\hat{\theta} = \frac{Y}{n}$$

since $E[\hat{\theta}] = \theta$. The variance of this estimator *when the data are considered random* is

$$Var[\hat{\theta}] = \frac{\theta(1-\theta)}{n}.$$

A standard error for this estimator is

$$SE[\hat{\theta}] = \sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}}$$

since $\hat{\theta} = Y/n$ is a consistent estimator for θ

Approximate confidence interval for binomial proportion

If $Y \sim Bin(n, \theta)$, then an approximate $100(1-a)\%$ confidence interval for θ is

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

where $\hat{\theta} = Y/n$.

Example

The Gallup poll is a poll of approximately 1,500 (randomly selected?) U.S. adults. In a poll dated 2017/02/19, 32.1% of respondents indicated that they were engaged at work. Thus an approximate 95% confidence interval for the proportion of all U.S. adults is

$$0.321 \pm 1.96 \times \sqrt{\frac{.321(1-.321)}{1500}} = (0.30, 0.34).$$

But the confidence interval actually says nothing about where the true proportion is, nor where we should believe it is. Instead it says something about how many of our intervals would cover the truth if we were to repeat this procedure over and over.

Confidence interval summary

Model	Estimator	Confidence Interval	Type
$Y_i \overset{ind}{\sim} N(\mu,\sigma^2)$	$\hat\mu = \overline{y}$	$\hat\mu \pm t_{n-1,a/2} s/\sqrt{n}$	exact
$Y_i \overset{ind}{\sim} N(\mu,\sigma^2)$	$\hat\mu = \overline{y}$	$\hat\mu \pm z_{a/2} s/\sqrt{n}$	approximate
$Y_i \overset{ind}{\sim} Ber(\theta)$	$\hat\theta = \overline{y}$	$\hat\theta \pm z_{a/2}\sqrt{\hat\theta(1-\hat\theta)/n}$	approximate
$Y \sim Bin(n,\theta)$	$\hat\theta = y/n$	$\hat\theta \pm z_{a/2}\sqrt{\hat\theta(1-\hat\theta)/n}$	approximate

Bayesian credible intervals based on the priors we have discussed are generally approximate confidence intervals.

Approximate here means that the coverage of the interval procedure will get closer and closer to the desired probability, i.e. 100(1-a)%, as the sample size gets larger.