Lab 5: Confidence intervals

STAT218

The objective of this lab is to learn to compute confidence intervals for a population mean, and more specifically, to learn to adjust interval coverage by calculating appropriate critical values. Since you already learned to calculate an interval in the last lab, the basic mechanics of the arithmetic are familiar.

We’ll use data from a sample of 100 births in North Carolina in 2004. To change things up a little, the data is stored as a .csv file (not an .RData file). If you were to download and open this file on your computer, it would likely appear as a spreadsheet (try if you’re curious). Read in the data using the command below.

# read in data and preview  
ncbirths <- read\_csv('data/ncbirths.csv')  
head(ncbirths)

# A tibble: 6 × 4  
 mother.age weeks birth.weight sex   
 <dbl> <dbl> <dbl> <chr>   
1 36 39 7.69 male   
2 35 40 8.88 male   
3 40 40 9 female  
4 37 40 7.94 male   
5 35 28 1.63 female  
6 25 40 8.75 female

Recall that the general formula for an interval is:

Throughout this lab, you’ll manipulate the coverage by obtaining different values of the critical value . We’ll start the back-of-the-envelope approach following the empirical rule.

### Intervals using the empirical rule

The empirical rule allows us to construct intervals using whole number multiples of the standard error and obtain the following approximate coverages:

* gives 68% coverage
* gives 95% coverage
* gives 99.7% coverage

An approximate 95% confidence interval for the mean birth weight (lbs) in NC in 2004 is:

# retrieve variable of interest  
bweight <- ncbirths$birth.weight  
  
# interval ingredients  
bweight.mean <- mean(bweight)  
bweight.sd <- sd(bweight)  
bweight.n <- length(bweight)  
  
# standard error  
bweight.se <- bweight.sd/sqrt(bweight.n)  
  
# 95% interval using empirical rule  
bweight.mean + c(-2, 2)\*bweight.se

[1] 6.89267 7.46633

Following class discussion, we’d interpret this as follows:

With 95% confidence, the mean birth weight of babies born in North Carolina in 2004 is estimated to be between 6.893 and 7.466 lbs.

To compute a 68% interval, we need only change the critical value. All of the above remains the same except the last command, which we change to:

# 68% interval using empirical rule  
bweight.mean + c(-1, 1)\*bweight.se

[1] 7.036085 7.322915

Notice that the interval got narrower: a more precise estimate can be given at a reduced coverage rate (which of course means the estimate is wrong more often).

With 68% confidence, the mean birth weight of babies born in North Carolina in 2004 is estimated to be between 7.036 and 7.323 lbs.

|  |
| --- |
| Your turn 1 |
| Calculate and interpret a 99.7% confidence interval for the mean number of weeks at birth.  # retrieve variable of interest (no. weeks at birth)  # interval ingredients  # standard error  # 99.7% interval for mean number of weeks at birth using empirical rule |

Because we’re only changing the critical value here, let’s save some work and write a simple function to calculate an interval from a vector of values and a critical value. (You don’t need to understand the syntax or be able to write functions in R, as this is a programming technique, but it may interest you to see how it can be done.)

# run this before continuing, but ignore unless interested  
make\_ci <- function(vec, cval){  
 vec.mean <- mean(vec)  
 vec.mean.se <- sd(vec)/sqrt(length(vec) - 1)  
 interval <- vec.mean + c(-1, 1)\*cval\*vec.mean.se   
 names(interval) <- c('lwr', 'upr')  
 return(interval)  
}

We can use this function to compute an interval a bit more efficiently. Check that the following give the same intervals as obtained above using fully manual calculations.

# 95% interval  
make\_ci(bweight, cval = 2)

lwr upr   
6.891225 7.467775

# 68% interval  
make\_ci(bweight, cval = 1)

lwr upr   
7.035362 7.323638

Try it yourself to get the hang of using this function.

|  |
| --- |
| Your turn 2 |
| Use make\_ci(...) to compute 95% and 99.7% confidence intervals for the mean number of weeks at birth.  # 99.7% interval for mean number of weeks at birth, using make\_ci(...)  # 95% interval for mean number of weeks at birth, using make\_ci(...) |

### Intervals using critical values

If you want to construct an interval with a coverage other than 68%, 95%, or 99.7%, you’ll need to use a different critical value. Instead of a whole number, you’ll need the th quantile from the model where:

This is perhaps a little more complex than it looks. You could probably determine which quantile to use in your head – the quantile you want is just the midpoint between your coverage level and 1. Consider the following examples:

* for a 95% interval, use
* for an 80% interval, use
* for a 99% interval, use

|  |
| --- |
| Your turn 3 |
| Which quantile would you use for…   1. A 96% confidence interval? 2. An 85% confidence interval? 3. A 98% confidence interval? |

#### Calculating quantiles

The qt(...) function in R will calculate quantiles for you. It takes two arguments: which quantile you want () and the degrees of freedom for the model. The degrees of freedom is one less than the sample size in this case (). The following commands illustrate the calculation.

# for 80% interval from n = 15 observations, use this quantile  
qt(p = 0.9, df = 14)

[1] 1.34503

# for a 92% interval from n = 30 observations, use this quantile  
qt(p = 0.96, df = 29)

[1] 1.814238

|  |
| --- |
| Your turn 4 |
| Calculate quantiles for the following scenarios:   1. 90% interval from 27 observations 2. 95% interval from 18 observations 3. 99% interval from 51 observations   # quantile for a 90% interval from 27 observations  # quantile for a 95% interval from 18 observations  # quantile for a 99% interval from 51 observations |

These quantiles are the critical values you’d use to construct an interval with the specified coverage and number of observations.

#### Constructing intervals

If we want a confidence interval for the mean with a specific coverage, first determine which quantile is needed as above and compute it, and then construct the interval as usual using that quantile as the critical value.

For example, if we want confidence intervals for the mean birth weight:

# 98% ci for mean birth weight  
crit.val <- qt(p = 0.99, df = 99)  
make\_ci(bweight, cval = crit.val)

lwr upr   
6.838671 7.520329

# 95% ci for mean birth weight  
crit.val <- qt(p = 0.975, df = 99)  
make\_ci(bweight, cval = crit.val)

lwr upr   
6.893499 7.465501

# 90% ci for mean birth weight  
crit.val <- qt(p = 0.95, df = 99)  
make\_ci(bweight, cval = crit.val)

lwr upr   
6.940175 7.418825

As a matter of interest, note that the critical value for the 95% interval is 1.9842. Technically, this is the value that provides an interval with 95% coverage; the approximation of 2 provided by the empirical rule is just that – an approximation.

|  |
| --- |
| Your turn 5 |
| Construct and interpret a 99% confidence interval for the mean number of weeks at birth.  # 99% ci for mean number of weeks at birth |

Now that you have a sense of the critical value calculation, it’s helpful to remind yourself how to make the interval fully from scratch.

|  |
| --- |
| Your turn 6 |
| Repeat the previous calculation, but without using the make\_ci(...) function. You should obtain exactly the same numerical result.  ## repeat last calculation but fully 'by hand'  # point estimate and standard error  # critical value  # interval |

### Working backwards to determine coverage

Now let’s try doing the above backwards. If you’re given a confidence interval and you know the summary statistics, you can figure out the interval coverage by solving for the critical value and using the pt(...) function. Really, you only need to know the standard error (or sample size and standard deviation) to do this.

First, find the margin of error by taking half the interval width.

Then, divide by the standard error to solve for the critical value:

Lastly, find the coverage as the area of the sampling distribution below the critical value:

In R:

# example interval for mean birth weight  
bweight.ci <- c(7.030077, 7.328923)  
  
# margin of error (half the width)  
bweight.ci.me <- diff(bweight.ci)/2  
  
# divide out standard error to get critical value  
crit.val <- bweight.ci.me/bweight.se  
  
# coverage  
pt(q = crit.val, df = bweight.n - 1)

[1] 0.85

Take a moment to align the R commands with the calculations shown above and check to make sure you see how this is consistent with the way we formed the interval in the first place. Then try it on your own.

|  |
| --- |
| Your turn 7 |
| Determine the coverage for the interval below for the mean number of weeks at birth.  # interval for mean number of weeks at birth bweeks.ci <- c(37.79485, 39.30515)  # margin of error (half the width)  # divide out standard error to get critical value  # coverage |