Lab 5: Confidence intervals

STAT218

The focus of this lab is on confidence intervals, with a particular emphasis on understanding the meaning of coverage (*i.e.*, confidence level) and exploring this concept through simulation. There are three main learning objectives:

1. Explore the sampling variability of intervals
2. Demonstrate the under-coverage of the normal model for small samples
3. Calculate an interval using the t model

We will use the same total cholesterol variable from the NHANES data and continue to pretend that these 3,179 observations form a population for which we have complete data. You'll need to load the dataset and extract the cholesterol variable.

```
# load dataset
load('data/nhanes.RData')

# extract cholesterol variable
cholesterol <- nhanes$TotChol
```

Remember, we're pretending that `cholesterol` contains all population values.

Exploring interval coverage

First you'll draw a single sample and calculate an interval estimate for the mean. The commands below will do this for you with no changes. Your goal will be to obtain an interval using these commands and compare your result with those of your groupmates. Since each of you is drawing a sample at random, your results will differ slightly.

```
# fix sample size
n <- 35

# draw a sample of size 35
my_samp <- sample(cholesterol, size = n)
```

```r
# calculate mean and standard error
xbar <- mean(my_samp)
xbar.se <- sd(my_samp)/sqrt(n)

# calculate the interval
xbar.interval <- xbar + c(-1, 1)*2*xbar.se

# display
xbar.interval
```

```
[1] 4.433772 5.095942
```

> **Your turn**
>
> Determine whether your interval captured the parameter of interest. Compute the population mean and compare with the interval estimate.
>
> ```r
> # does the interval contain the population mean?
> ```

Now imagine your group was really big, say, 10,000 students, and you all generated one interval each. The proportion of you who captured the population mean measures the **coverage** of the interval. We will simulate this in class to determine the coverage of the interval you just calculated.

Simulating deviations

The multiplier $2 \times SE$ comes from the normal model for the sampling distribution of the point estimate $\bar{x}$. This particular multiplier is intended to achieve 95% coverage — so in theory, 95 out of every 100 intervals will capture the population mean. The normal model can be expressed as a model for the scaled deviations $\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$, where the scaling factor uses the population standard deviation $\sigma/\sqrt{n}$. A nominal 95% coverage arises from the assumption that 95% of these deviations are between -2 and 2. But when we calculate an interval, we're actually applying the normal model to $\frac{\bar{x}-\mu}{s_x/\sqrt{n}}$, where the deviations have been scaled by the standard error $s_x/\sqrt{n}$. Are 95% of these deviations between -2 and 2?

The commands below allow you to simulate scaled deviations of sample means from a large number of samples, using both (a) population standard deviation and (b) standard error as scaling factors. Don't worry about understanding the codes too much — as long as you understand the output, that is sufficient. Your task is to run these commands and compare the coverage of the normal model when SE is used in place of SD.

```r
# fix sample size
n <- 5

# function to simulate scaled deviation of one sample mean
sim_dev <- function(n, data){
  samp <- sample(data, size = n)
  xbar <- mean(samp)
  xbar.se <- sd(samp)/sqrt(n)
  xbar.sd <- sd(data)/sqrt(n)
  mu <- mean(data)
  dev <- c(se = (xbar - mu)/xbar.se,
           sd = (xbar - mu)/xbar.sd)
  return(dev)
}

# repeat many simulations
nsim <- 1000
sim.devs <- sapply(1:nsim, function(i){sim_dev(n, data = cholesterol)}) |>
  t() |>
  as.data.frame()

# how many are between -2 and 2 using SD?
coverage.sd <- sum(abs(sim.devs$sd) < 2)/nsim
coverage.sd
```

```
[1] 0.954
```

```r
# using SE?
coverage.se <- sum(abs(sim.devs$se) < 2)/nsim
coverage.se
```

```
[1] 0.866
```

Your turn

Run the codes above and discuss with your group:

- What did these commands do? Affirm your understanding of this simulation by listing the steps performed.
- Do either, none, or both scaled deviations achieve the nominal coverage of 95%?

- If you change the sample size, do you get a different answer?

Calculating an interval

In light of the under-coverage resulting from the normal model for small sample sizes, the t model provides a better interval in practice.

$$\bar{x} \pm c \times SE(\bar{x})$$

In the expression above, c is a "critical value" obtained from the t model with $n-1$ degrees of freedom; its exact value depends on the desired coverage. The commands below show you how to calculate each piece and form an interval.

```
# fix sample size and desired coverage
n <- 10
coverage <- 0.95

# draw one sample
samp <- sample(cholesterol, size = n)

# interval ingredients
xbar <- mean(samp)
xbar.se <- sd(samp)/sqrt(n)
c.val <- qt((1 - coverage)/2, df = n - 1, lower.tail = F)

# calculate interval
xbar + c(-1, 1)*c.val*xbar.se
```

```
[1] 4.12674 5.44326
```

Your answer will differ slightly from this result, since you will draw a different sample when you run these commands.

Your turn

Using the same sample above, compute a 90% confidence interval for mean total HDL cholesterol. Discuss with your group and interpret the interval in context.

```
# adjust coverage and calculate a new critical value

# calculate the interval
```