Day 17 - Confidence intervals

Learning outcomes and terms to know

By now, we have a good understanding of how to test specific research questions via hypothesis testing. In the Coke vs. Pepsi example, our small p-value provided strong evidence that π is greater than 0.5. But *how much* greater? What is our best estimate of π, the true probability of correctly identifying Coke? To answer such a question, we create a **confidence interval**, which represents a range of plausible values for the parameter.

Simulation-based estimation for a single proportion (confidence intervals)

Confidence intervals

A _____ represents a range of plausible values for an unknown parameter.

Bootstrapping

In order to obtain a _____ for the true population proportion, we do not need to assume the null hypothesis is true. Instead, we just want to understand the variability of the sample statistic.

To obtain the sampling distribution of the sample proportion, we can bootstrap resample from the observed data. To do so, we:

- resample with replacement from the original data n times
- compute the sample proportion of successes for each bootstrap resample
- repeat this process many times (i.e. 1000 or 10000)

Let's keep working with the data from the first day of class, when 23/32 students correctly identified Coke.

Creating the bootstrap distribution by hand

1. Describe how we could use labeled cards to generate the bootstrap distribution. Be sure to describe: how you would label the cards and how many cards you need, how you would draw the cards, and what you would calculate for each simulation.

   > We would need 32 cards, of which 23 would be labelled "correct" and 9 "incorrect." For each simulation, draw 32 cards, each time putting the card back in the deck. Compute the proportion of the 32 draws resulting in "correct."

2. Use R to generate the sampling distribution of the sample proportion by bootstrap resampling and use the distribution to find a 95% confidence interval.

```
# create vector of observations
obs <- c(rep("correct", 23), rep("incorrect", 9))

# do the bootstrapping
bootstrap <- matrix(NA, nrow = 1000, ncol = 1)
for(i in 1:1000){
  bootstrap[i,1] <- mean(sample(obs, size = length(obs), replace = T) == "correct")
}
int <- quantile(bootstrap, c(0.025, 0.975))
int
```

```
   2.5%   97.5%
0.56250 0.84375
```

3. Interpret the interval in the context of the problem.

   > **!** Important
   >
   > Generally, your interpretation of a confidence interval should read like: We are ___% confident that the [**parameter**] is in the interval (L, U).

   > We are 95% confident that the true probability of correctly identifying Coke (in a blind taste test against Pepsi) is between 0.562 and 0.844.

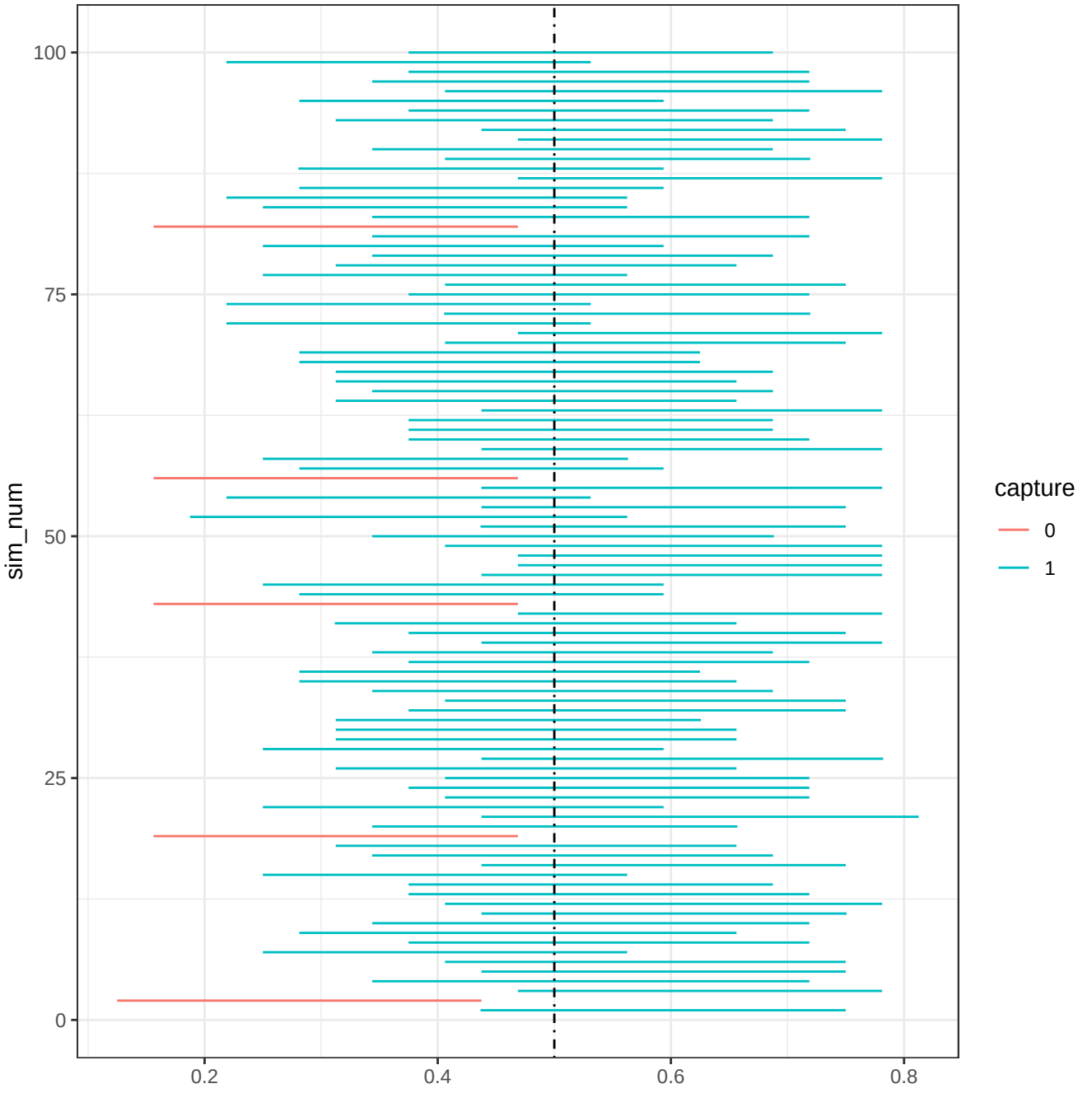
Meaning of confidence

Let's work together to use simulation to understand the meaning of confidence. In order to understand what confidence means, we have to *assume we actually know* π. For now, let's assume that $\pi = 0.5$ (i.e., we are just guessing).

```
nsims <- 100
n <- 32
cis <- matrix(NA, nsims, 2)
true_pi <- .5
for(i in 1:nsims){
  ncorrect <- rbinom(1, n, true_pi)
  ncorrect_vec <-c(
    rep("yes", ncorrect),
    rep("no", n - ncorrect)
  )

  boot <- replicate(1000, expr = {
    tmp <- sample(ncorrect_vec, size = 32, replace = T)
    mean(tmp == "yes")
  })
  ci <- quantile(boot, c(0.025, 0.975))
  cis[i,] <- ci
}
ci_tbl <- tibble(
  sim_num = 1:nsims,
  lwr = cis[,1],
  upr = cis[,2]
)  %>%
  mutate(capture = ifelse(lwr < true_pi & upr > true_pi, 1, 0) %>%
           factor(., levels = c(0, 1)))

ci_tbl %>%
  ggplot() +
  geom_linerange(aes(xmin = lwr, xmax = upr, y = sim_num, col = capture)) +
  theme_bw() +
  geom_vline(aes(xintercept = true_pi), linetype = "dotdash")
```


sim_num
100
75
50
25
0
0.2
0.4
0.6
0.8
capture
0
1

The meaning of confidence

When we say 95% confidence, we mean that if we were to repeatedly sample from the population of interest, calculate a statistic, and use that statistic to generate a 95% confidence interval for the parameter, then 95% of those intervals would capture the parameter *in the long run*.

Therefore, we say that confidence describes *confidence in the process of creating intervals*, not in a single interval. The probability that a single interval contains the parameter is either 0 or 1; and we will never know which!