Computational Biology Group

D BSSE

ETH zürich

Introduction to Bayesian Statistics with R

2: Exercise solutions

Jack Kuipers

First we load the tidyverse and set a seed.

```
library(tidyverse); set.seed(42)
```

Exercise 2.1 - confidence intervals

The answers to the first part are in the lecture notes, so we focus on the following questions here.

Take a sample from a normal distribution extract its 95% confidence interval.

Sample from the same process a large number of times and see how often the sample means lie within the first confidence interval.

Is it 95%? How does this align with statement 6 if we replace "true mean" with "sample mean"?

Let's take our first sample and store its mean and confidence interval

```
test_sample <- rnorm(50, mean = -0.25, sd = 1)
(first_sample_mean <- mean(test_sample))
```

```
## [1] -0.2856718
```

```
(first_conf_int <- t.test(test_sample)$conf.int)
```

```
## [1] -0.61291775  0.04157418
## attr(,"conf.level")
## [1] 0.95
```

in this case we can see that the interval is shifted left compared to one centred at the mean.

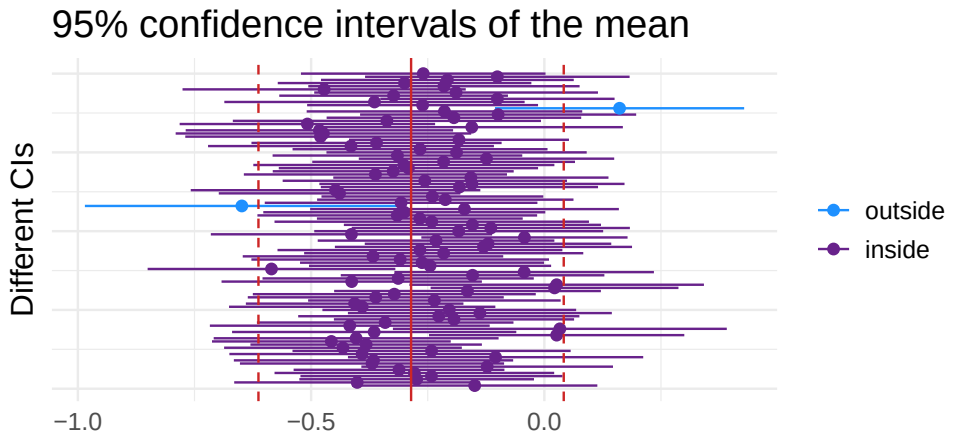
Let's move to 100 repetitions

```
conf_ints <- NULL
for (ii in 1:100) {
    test_sample <- rnorm(50, mean = -0.25, sd = 1)
    tt   <- t.test(test_sample, mu = -0.25) # test against true mean
    low  <- tt$conf.int[1]
    high <- tt$conf.int[2]
    me   <- tt$estimate
    inside <- me < first_conf_int[2] && me > first_conf_int[1]
    conf_ints  <- rbind(conf_ints,
```

```
                  data.frame(id = ii, low = low, high = high, inside = inside, me = me))
}
```

Let's plot them all and see which sample means lie in our original confidence interval:

```
ggplot(conf_ints) +
  geom_segment(aes(x = low, xend = high, y = id, yend = id, col = inside)) +
  geom_point(aes(x = me, y = id, col = inside), size = 2) + theme_minimal() +
  theme(legend.title = element_blank(), axis.text.y = element_blank()) +
  scale_color_manual(values = c("dodgerblue", "darkorchid4"),
                     label = c("outside", "inside")) +
  labs(x = "", y = "Different CIs") +
  ggtitle("95% confidence intervals of the mean") +
  geom_vline(aes(xintercept = first_sample_mean), col = "firebrick3") +
  geom_vline(aes(xintercept = first_conf_int[1]), col = "firebrick3", linetype=2) +
  geom_vline(aes(xintercept = first_conf_int[2]), col = "firebrick3", linetype=2) +
  theme(text = element_text(size = 14))
```



Here 98 of the new sample means were inside the original confidence interval.

For a larger number of repetitions, say a million:

```
n_reps <- 1e6
mean_inside <- rep(NA, n_reps)
for (ii in 1:n_reps) {
    tt <- t.test(test_sample <- rnorm(50, mean = -0.25, sd = 1), mu = -0.25)$estimate
    mean_inside[ii] <- tt < first_conf_int[2] && tt > first_conf_int[1]
}
round(100*mean(mean_inside), 2)
```

```
## [1] 97.5
```

and we actually get more than 95% of the new sample means lying in the original confidence interval.

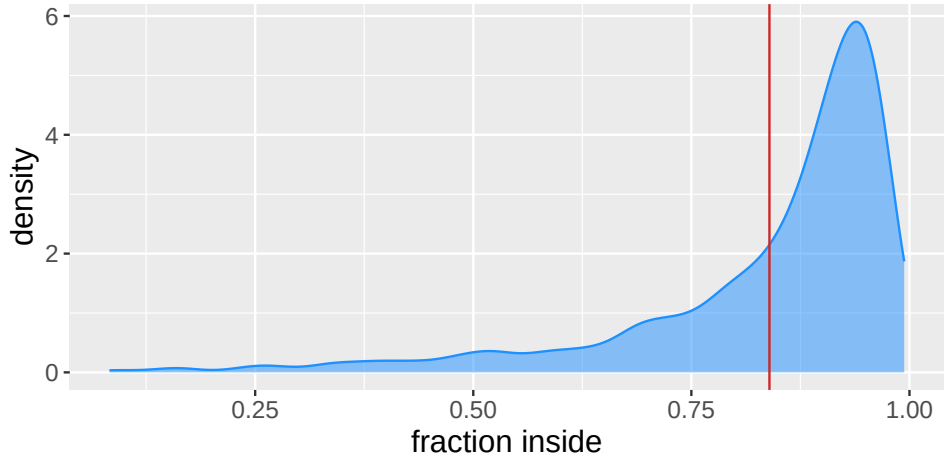
In fact this percentage will depend on the original confidence interval, if its mean is to the left or right of the real mean and if its width is over or underestimated. Let's generate some statistics

```
n_reps <- 1e3
inside_means <- rep(NA, n_reps)
for (jj in 1:n_reps) {
  test_sample <- rnorm(50, mean = -0.25, sd = 1)
  first_conf_int <- t.test(test_sample)$conf.int
  n_reps_inner <- 1e3
```

```
  mean_inside <- rep(NA, n_reps_inner)
  for (ii in 1:n_reps) {
    tt <- t.test(test_sample <- rnorm(50, mean = -0.25, sd = 1), mu = -0.25)$estimate
    mean_inside[ii] <- tt < first_conf_int[2] && tt > first_conf_int[1]
  }
  inside_means[jj] <- mean(mean_inside)
}
```

and make a kernel density plot of the different estimates

```
ggplot(data.frame(inside_means = inside_means), aes(x = inside_means)) +
  geom_density(colour = "dodgerblue", fill = "dodgerblue", alpha = 0.5) +
  geom_vline(aes(xintercept = mean(inside_means)), col = "firebrick3") +
  theme(text = element_text(size = 14)) + xlab("fraction inside")
```



As you can see, in general they are not at 95% with a mean (red line) at 83.95%. Of course if we know the true mean (and the standard error), we can construct an interval which will contain 95% of sample means, but from one sample we do not know where we are compared to the true mean or how many other sample means will align with that particular one. So for question 6, we cannot make such a quantitative statement about sample means, let alone the true mean (as in the wording of question 6) which is not even a random variable.

Bonus Exercise 2.2 - a testing example

*This example comes from **Eddy (1982)**, and asked of medical doctors to see if they can get the right ballpark probability in the end:*

- *1% of women at age forty who participate in routine screening have breast cancer.*
- *80% of women with breast cancer will get positive mammographies.*
- *9.6% of women without breast cancer will also get positive mammographies.*

A woman in this age group had a positive mammography in a routine screening. What is the probability that she actually has breast cancer?

We can plug all the information into Bayes theorem, using + to signify a positive test and C to indicate having breast cancer.

$$P(C \mid +) = \frac{P(+ \mid C)P(C)}{P(+)} = \frac{0.8 \times 0.01}{P(+)} = \frac{0.008}{P(+)}$$

To proceed we use the expanded version of the denominator in terms of the two possible cancer states

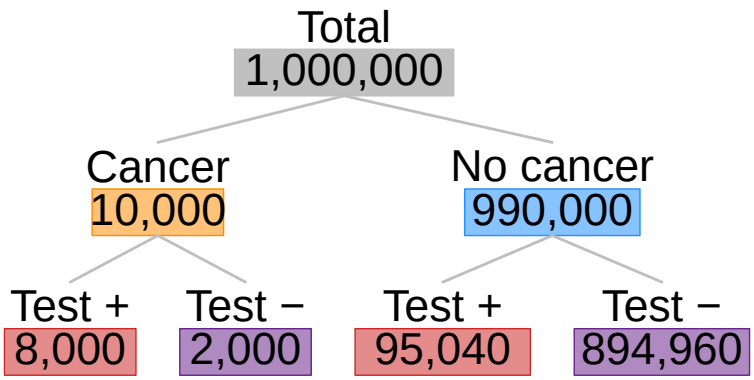
$$P(+) = P(+ \mid C)P(C) + P(+ \mid \neg C)P(\neg C) = 0.8 \times 0.01 + 0.096 \times 0.99 = 0.10304$$

and obtain

$$P(C \mid +) = \frac{0.008}{0.10304} = 0.078$$

or around 8%. This is quite far from a lot of people's intuitive answer of around 80% in line with the true positive rate of the test, and illustrates the dangers of "inverting" conditional probabilities in your head.

A possibly easier way to get to the solution is to imagine a large population which we can separate into the four categories according to the probabilities above:



Among those with a positive test: 8,000 have cancer, while 95,040 don't so

$$P(C \mid +) = \frac{8,000}{8,000 + 95,040} = 0.078$$