Computational Biology Group

D BSSE

ETH zürich

Introduction to Bayesian Statistics with R

2: Exercise solutions

Jack Kuipers

28 November 2022

First we load the tidyverse and set a seed.

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

Exercise 2.1 - confidence intervals

Take a sample from a normal distribution extract its 95% confidence interval. Is the true mean inside this confidence interval?

Repeat this procedure a large number of times. How often is the true mean in the confidence interval?

Is your result in line with questions 4 and 5 of the confidence interval quiz?

For a single sample

```
test_sample <- rnorm(50, mean = -0.25, sd = 1)
t.test(test_sample)$conf.int
```

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

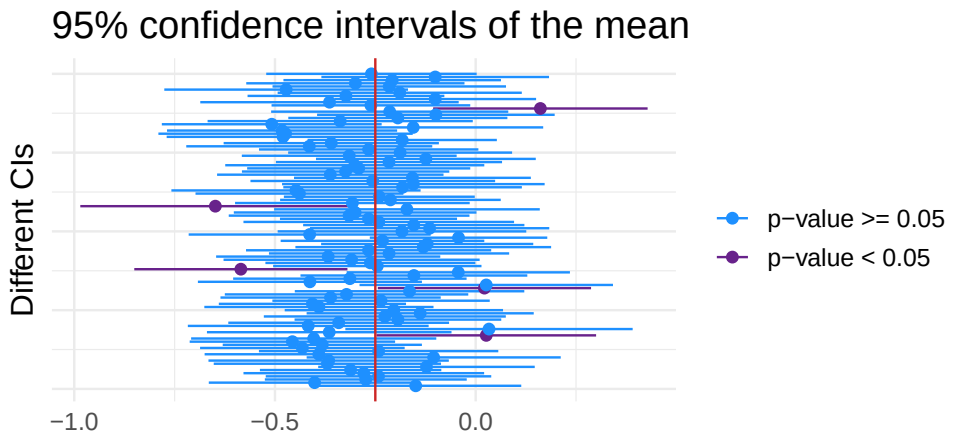
we can see that the true mean is in the confidence interval.

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]
    pval <- tt$p.value
    me   <- tt$estimate
    conf_ints  <- rbind(conf_ints,
                  data.frame(id = ii, low = low, high = high, pval = pval, me = me))
}
```

Remembering that for two-sided tests, if a p-value (with a null of the true mean) is significant at a significance level α, the $1-\alpha$ confidence interval does *not* contain the true mean, we colour by significance and plot:

```
ggplot(conf_ints) +
  geom_segment(aes(x = low, xend = high, y = id, yend = id, col = pval < 0.05)) +
  geom_point(aes(x = me, y = id, col = pval < 0.05), size = 2) + theme_minimal() +
  theme(legend.title = element_blank(), axis.text.y = element_blank()) +
  scale_color_manual(values = c("dodgerblue", "darkorchid4"),
                     label = c("p-value >= 0.05", "p-value < 0.05")) +
  labs(x = "", y = "Different CIs") +
  ggtitle("95% confidence intervals of the mean") +
  geom_vline(aes(xintercept = -0.25), col = "firebrick3") +
  theme(text = element_text(size = 14))
```



Here 5 were significant so 95 percent of the confidence intervals included the true mean.

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)
    mean_inside[ii] <- tt$conf.int[1] < -0.25 && tt$conf.int[2] > -0.25
}
round(100*mean(mean_inside), 2)
```

```
## [1] 94.99
```

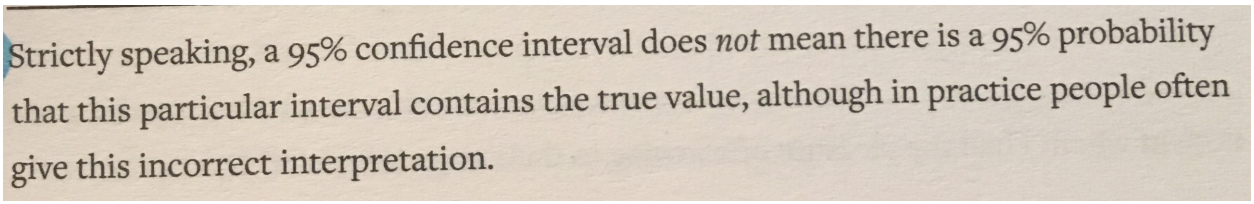
we indeed get around 95% of the confidence intervals including the true mean of $-\frac{1}{4}$.

Why, then are statements 4 and 5 of the confidence interval quiz false? The pedantic answer is that μ is fixed. Whether μ lies between 0.1 and 0.4 is deterministic and either true or false. Is it not a probabilistic question, and we cannot answer it with a probability. Note that in all the simulations above we knew the true mean, but in the quiz setting we only know the sample mean. From the sample mean and without knowing the true mean, we don't know if our particular confidence interval included the true mean or not. A weak analogy would be rolling a die. Before rolling it, the chance you get a 6 is $\frac{1}{6}$.

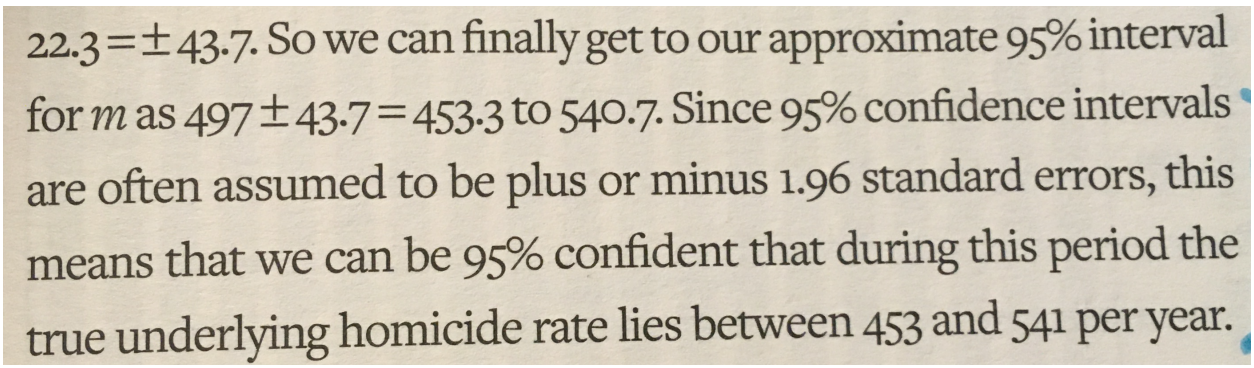
```
die_sample <- sample(6, 1)
```

Once rolled, you get a 5 and it either is a 6 or isn't (not in this case). Before running our experiment and getting the data, the probability the confidence interval containing the true mean is 95%. For the particular random realisation of the confidence interval we cannot say.

The misinterpretation of confidence intervals is fairly common. For example, despite this correct note in a book from a statistics professor


Strictly speaking, a 95% confidence interval does *not* mean there is a 95% probability that this particular interval contains the true value, although in practice people often give this incorrect interpretation.

their own incorrect description 7 pages later of the confidence interval computed from real data, ironically highlights their point about misinterpretation


22.3 = ± 43.7. So we can finally get to our approximate 95% interval for *m* as 497 ± 43.7 = 453.3 to 540.7. Since 95% confidence intervals are often assumed to be plus or minus 1.96 standard errors, this means that we can be 95% confident that during this period the true underlying homicide rate lies between 453 and 541 per year.

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

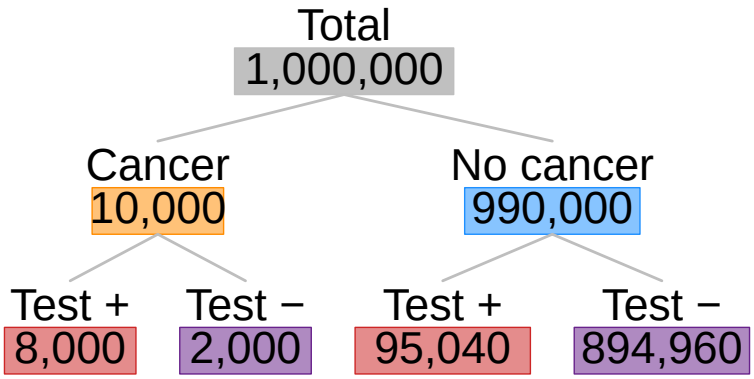
$$P(+) = P(+ \mid C)P(C) + P(+ \mid \neg C)P(\neg C) = 0.8 \times 0.01 + 0.096 \times 0.99 = 0.10304$$

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

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