ADS2 - Bayesian Inference, model solution

ADS2 (based on Simon's solution)

Semester 2, 2023/24

These solutions are just model ones. Thus, you might have developed your own strategy that gives comparable results but uses more optimal code. Or your code may be very different from the one suggested here. It is absolutely fine as soon as our results are comparable.

Is the Guinness factory adding enough barley?

Calculations for two hypotheses being equally correct

The Guinness beer factory requires 50 g of barley per pint of beer. They examined 50 pints and found an average barley content of 46 g per pint.

Determine the posterior probability that enough barley is being added to each pint and the Bayes Factor for these two hypotheses.

We do not suspect anybody. Thus, the probabilities of either hypothesis are equal. The given prior probabilities and data likelihoods are listed here:

```
# Prior probabilities of seeing your results
P_h1 <- 0.5 # Enough barley was added to each pint of beer
P_h2 <- 0.5 # NOT enough barley was added to each pint of beer

# Likelihoods
P_data_h1 <- 0.7
P_data_h2 <- 0.4
```

Bayes factor is the ratio of the posterior odds to prior odds[1]. Its main use is that it can help you to compare the *likelihood* of one hypothesis over the other given the data you have[2]. After certain math transformations, you can get the following definition of Bayes factor:

$$BF_{(H_1:H_2)} = \frac{PO_{(H_1:H_2)}}{O_{(H_1:H_2)}} = \;...\; =$$

$$= (\frac{P_{(data\ \mid\ H_1)}}{P_{(data\ \mid\ H_2)}} \times \frac{P_{(H_1)}}{P_{(H_2)}}) \;/\; \frac{P_{(H_1)}}{P_{(H_2)}} =$$

$$= \frac{P_{(data\ \mid\ H_1)}}{P_{(data\ \mid\ H_2)}}$$

where $O_{(H_1:H_2)} = \frac{P_{(H_1)}}{P_{(H_2)}}$ is the prior odds, $PO_{(H_1:H_2)} = \frac{P_{(H_1\ \mid\ data)}}{P_{(H_2\ \mid\ data)}}$ is the posterior odds, and $BF_{(H_1:H_2)}$ is Bayes factor.

[1] **Merlise Clyde, Mine Çetinkaya-Rundel, Colin Rundel, David Banks, Christine Chai, Lizzy Huang. An Introduction to Bayesian Thinking: A Companion to the Statistics with R Course. GitHub; 2022, p. 52**

[2] https://www.statology.org/bayes-factor/

```r
P_data_h1/P_data_h2 # Bayes factor
```

```
## [1] 1.75
```

```r
Bayes_f <- P_data_h1/P_data_h2
```

We are interested in the probability of hypothesis 1 (there is enough barley in the beer) being true given the mean barley content is 46 g. Thus, we are interested in $P_{(H1\ |\ mean=46g)}$.

```r
(P_h1/P_h2) # Ratio of priors
```

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

```r
# Given that Ratio of posteriors is the product of Bayes factor x Ratio of priors
Bayes_f * (P_h1/P_h2) # Ratio of posteriors
```

```
## [1] 1.75
```

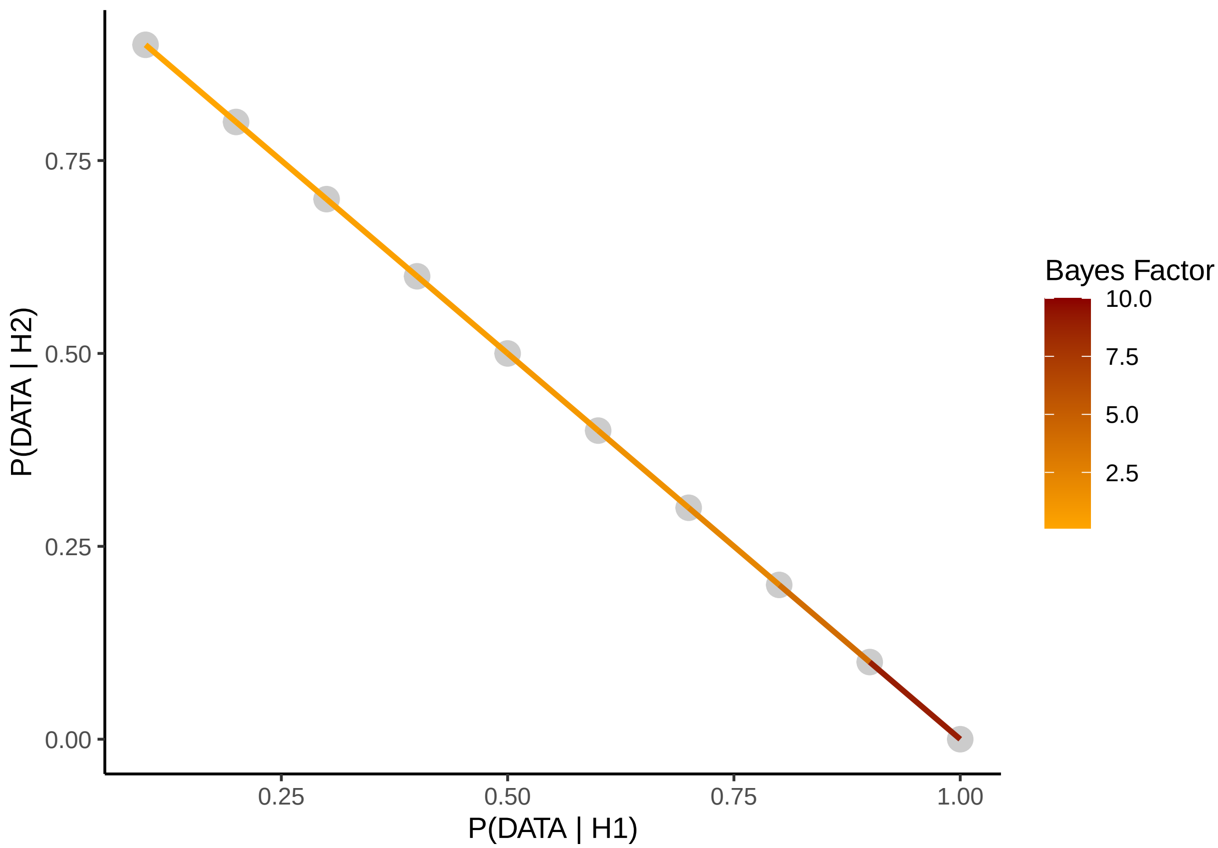
Originally, we were thinking that either hypothesis is correct ($\frac{P_{(H_1)}}{P_{(H_2)}} = \frac{0.5}{0.5} = 1$). The ratio of posteriors says that given the data we have the probability of H1 (enough barley is added) being true is 1.75 times higher than of H2 (NOT enough barley is added). Or $P_{(H1)}$ is 1.75 times more likely to be correct. From this, a simple arithmetic suggests that $P_{(H1\ |\ mean=46g)} = 0.64$.

One question that may arise is where to get these likelihoods, $P_{(data\ |\ H_n)}$, that are used to calculate Bayes factor. In this example, you are already provided with it. But normally, you would have to calculate it by yourself. For example, like it is described in the latter section, **A dicy game**.

The relationship of the Bayes factor with the probabilities of each hypothesis being correct

```r
bfs = data.frame('H1' = seq(0.1, 1, 0.1),
                 'H2' = 1 - seq(0.1, 1, 0.1)) %>%
  mutate('bf' = H1 / H2) %>%
  mutate('bf' = ifelse(bf == 'Inf', 10, bf))

ggplot(bfs, aes(H1, H2)) +
  theme_classic() +
  geom_point(size = 4, alpha = 0.2) +
  geom_line(aes(colour = bf), size = 1) +
  labs(x = 'P(DATA | H1)', y = 'P(DATA | H2)') +
  scale_color_gradient(low = 'orange',
                       high = 'darkred',
                       name = 'Bayes Factor')
```



Calculations for two hypotheses NOT being equally correct

Let's discuss a scenario when we do suspect that employees steal from the factory. Thus, the probability of H1 ($P_{(H1)}$, enough barley is being added) is lower than of H2 ($P_{(H2)}$). At the same time, the probabilities of the data being correct under each hypothesis **are not changed** ($P_{(mean=46g \mid H1)}$ or $P_{(mean=46g \mid H2)}$).

```
P_h1 <- 0.2 # You may use your own number here
P_h2 <- 1 - P_h1
P_data_h1 <- 0.7
P_data_h2 <- 0.4

# Bayes factor
Bayes_f <- P_data_h1/P_data_h2
Bayes_f
```

```
## [1] 1.75
```

```
# Posteriors
Ratio_posteriors <- Bayes_f * (P_h1/P_h2)
Ratio_posteriors
```

```
## [1] 0.4375
```

The ratio of posteriors for this scenario says that given the data we have the probability of H1 (enough barley is added) being true is more than 2 times lower than of H_2 (NOT enough barley is added). From this, a simple arithmetic suggests that $P_{(H1 \mid mean=46g)} = 0.304$.

A dicy game

This task can be expressed as a set of Bernoulli trials with the success – we get sixes, and the failure – we get another value. This scenario can be modeled by the binomial probability:

$$P_{(0\ \le\ x\ \le\ 20)} = \binom{20}{x} \cdot model^x \cdot (1 - model)^{20-x}$$

The problem here is to find out whether the die used by the friend has only one side with the 6 or it has several sides with sixes. Thus, the set of hypotheses includes:

- H_1: Their die has 1 six
- H_2: Their die has 2 sixes
- ...
- H_6: Their die has 6 sixes

Ok, the last hypothesis can be rejected as there were 13 non-sixes out of 20 rolls. But we still have 5 hypotheses. We need to calculate the posterior probability of each hypothesis being correct *given* the respective data (7 sixes out of 20 trials). We need to use Bayes' formula:

$$P_{(Hypothesis\ |\ Evidence)} = \frac{P_{(Evidence\ |\ Hypothesis)} \cdot P_{(Hypothesis)}}{P_{(Evidence)}}$$

Or shorter:

$$P_{(H\ |\ D)} = \frac{P_{(D\ |\ H)} \cdot P_{(H)}}{P_{(D)}}$$

whereH_n is n sixes and the probability of getting six, respectively; D is the observed data; P_{H_n} is the prior probability of the respective hypothesis; $P_{(H_n|D)}$ is the posterior probability for each hypothesis; $P_{(D\ |\ H_n)}$ is the likelohood of seing your observed results according to the respective hypothesis; P_D is the probability of getting the observed results according to any hypothesis $P_D = \sum P_{H_n} \times P_{(D\ |\ H_n)}$.

You hypotheses must be transformed from the trivial form (see the list above) to the statistical form: the probability of getting six.

- H_1: Their die has 1 six => the chance of getting 6 is 1/6
- H_2: Their die has 2 sixes => the chance of getting 6 is 2/6
- ...
- H_5: Their die has 5 sixes => the chance of getting 6 is 5/6

Now about the priors. It is a tricky part that you must be able to justify. In this example, we assume that anything is possible. It is a good assumption when you have no prior data about the discussed phenomenon. Deciding whether it is a reasonable assumption in this scenario is up to you, but we will accept exactly this model of our prior beliefs: we are not sure which hypothesis is correct, so we give them equal credibility: 0.2.

```r
## 1 vs 5 sixes per die
Probabilities_sixes <- c(1 / 6, 2 / 6, 3 / 6, 4 / 6, 5 / 6)

P_givenData_Expectation <- dbinom(x = 7, size = 20, prob = Probabilities_sixes)
```

```r
dice <- cbind(Probabilities_sixes, P_givenData_Expectation)

P_hypotheses <- rep(0.2, 5) # As far as we are not sure in any hypothesis, we set it to 0.2

P_givenData <- sum(P_hypotheses * dice[, 2])
```

Let's list what we have and calculate our posterior beliefs:

```r
Probabilities_sixes # Statistical models for our data == chances of success
```

```
## [1] 0.1666667 0.3333333 0.5000000 0.6666667 0.8333333
```

```r
P_hypotheses # Our prior beliefs in each hypothesis
```

```
## [1] 0.2 0.2 0.2 0.2 0.2
```

```r
P_givenData # The probability of getting 7 sixes out of 20 trials according to each hypothesis
```

```
## [1] 0.05695742
```

```r
P_givenData_Expectation # The likelihood of each hypothesis
```

```
## [1] 2.588206e-02 1.821288e-01 7.392883e-02 2.845762e-03 1.656452e-06
```

```r
(P_givenData_Expectation[1] * P_hypotheses[1]) / P_givenData # The probability of H1 being correct:
```

```
## [1] 0.09088213
```

Ref: https://allendowney.github.io/BiteSizeBayes/04_dice.html"'

Originally created by Simon (Jingyuan Chen) on April 18$^{\text{th}}$, 2022,

Updated by Dmytro Shytikov in 2024