05: Bayes' Rule

Exercise 5.1 [Iterative application of Bayes rule, and seeing how posterior probabilities change with inclusion of more data]

Suppose that the same randomly selected person as in Table 5.4 gets re-tested after the first test result was positive, and on the retest the result is negative. When taking into account the results of both tests, what is the probability that the person has the disease? Hint: for the prior probability of the retest, use the posterior computer from the Table 5.4 ***TLDR:*** *first test positive; second test negative; find probability diseased.*

```
library("dplyr")
prior = data.frame(pPresent = 0.001, pAbsent = 1-0.001)
table5.4 = data.frame(
           pPresent = c(0.99*prior$pPresent,(1-0.99)*prior$pPresent),
           pAbsent = c(0.05*prior$pAbsent, (1-0.05)*prior$pAbsent),
           row.names = c("Positive","Negative")
           )
posterior1 = table5.4["Positive",] / sum(table5.4["Positive",])
prior = posterior1
table5.4 = data.frame(
           pPresent = c(0.99*prior$pPresent,(1-0.99)*prior$pPresent),
           pAbsent = c(0.05*prior$pAbsent, (1-0.05)*prior$pAbsent),
           row.names = c("Positive","Negative")
           )
posterior2 = table5.4["Negative",] / sum(table5.4["Negative",])
posterior2
```

```
##              pPresent   pAbsent
## Negative 0.0002085862 0.9997914
```

Exercise 5.2

A. Suppose that the population consists of 100,000 people. Compute how many people would be expected to fall into each cell of table 5.4

```
popSize = 100000
prior = data.frame(pPresent = 0.001, pAbsent = 1-0.001)
table5.4freq = data.frame(
           fPresent = c(popSize*0.99*prior$pPresent,popSize*(1-0.99)*prior$pPresent),
           fAbsent = c(popSize*0.05*prior$pAbsent,popSize*(1-0.05)*prior$pAbsent),
           row.names=c("Positive","Negative"))
table5.4freq = rbind(table5.4freq,
                     data.frame(
                       fPresent = sum(table5.4freq$fPresent),
                       fAbsent = sum(table5.4freq$fAbsent),
                       row.names=c('rowTotal')
                     ))
table5.4freq$colTotal = table5.4freq$fPresent + table5.4freq$fAbsent
table5.4freq
```

```
##          fPresent fAbsent colTotal
## Positive       99    4995     5094
## Negative        1   94905    94906
## rowTotal      100   99900   100000
```

Notice the frequencies on the lower margin of the table. They indicate that out of 100,000 people, only 100 have the disease, while 99,900 do not have the disease. These marginal frequencies instantiate the prior probability that p(pPresent)=0.001. Notice also the cell frequencies in the column fPresent, which indicate that of 100 people with the disease, 99 have a positive test result and 1 has a negative test result. These cell frequencies instantiate the hit rate of 0.99. Your job for this part of the exercise is to fill in the frequencies of the remain cells of the table. ***TLDR:*** *The frequencies instantiate the probabilities already laid out.*

B. Take a good look at the frequencies in the table you just computed for the previous part. These are the so-called *"natural frequencies"* of the events, as opposed to the somewhat unintuitive expression in terms of conditional probabilities. From cell frequencies alone, determine the proportion of people who have the disease given that their test result is positive. Before computing the exact answer arithmetically first give a rough intuitive answer merely by looking at the relative frequencies in the row "Positive".

Intuitive guess from table: approx 1 in 50 ~= 0.02

Calculated result:

```
table5.4freq["Positive","fPresent"]/table5.4freq["Positive","colTotal"]
```

```
## [1] 0.01943463
```

Does your intuitive answer match the intuitive answer you provided when originally reading about Table5.4? Probably not. Your intuitive answer here is probably much coser to the correct answer. Now compute the exact answer arithmetically. It should match the result from applying Bayes' rule in Table5.4 ***TLDR:*** *Frequencies and the table format help our intuition.*

C. Now we'll consider a related representation of the probabilities in terms of natural frequencies, which is especially useful when we accumulate more data. This type of representation is called a "Markov" representation by Krauss, Martingnon, andHoffrage (1999). Suppose now we start with a population of N=10,000,000 people. We expect 99.9% of them (9,990,000) of them not to have the disease,and just 0.1% of them (10000) to have the disease. Now consider how many people we expect to test positive. Of the 10,000 people who have the disease 99% will be expected to test positive (9,900). Of the 9,990,000 people who do not have the disease 5% (499,500) will be expected to test positive. Now consider re-testing everyone who has tested positive on the first test. How many of them are expected to show a negative result on the retest?

TLDR: *Testing those who already tested positive, now using Markov representation.*

```
N = 10000000
nPresent = N*0.001 #10000
nPresPos = nPresent*0.99 #9900
nPresPosNeg = nPresPos*0.01 #99
nAbsent = N-nPresent #9990000
nAbsPos = nAbsent*0.05 #499500
nAbsPosNeg = nAbsPos*0.95 #474525
nAbsPosNeg + nPresPosNeg
```

```
## [1] 474624
```

D. Use the diagram in the previous part to answer this: what proportion of people who test positive at first and then negative on the retest, actually have the disease? In other words, of the number of people at the bottom of the diagram in the previous part, what proportion of them are in the left branch of the tree?

```
nPresPosNeg / (nAbsPosNeg + nPresPosNeg)
```

```
## [1] 0.0002085862
```

How does the result compare with your answer to exercise 1? Agrees exactly ***TLDR:*** *Can use this to work out probability of having the disease, given that the tests were positive then negative. Agrees with exercise 1.*

Note: in real life, it's likely the test results won't be independent on repetition…

Exercise 5.3

Consider again the disease and diagnostic test of the previous two exercises. **A.** Suppose that a person selected at random from the population gets the test and it comes back negative. Compute the probability that the person has the disease.

```
table5.4freq["Negative","fPresent"]/table5.4freq["Negative","colTotal"]
```

```
## [1] 1.053674e-05
```

B. The person then gets re-tested, and on the second test the result is positive. Compute the probability that the person has the disease.

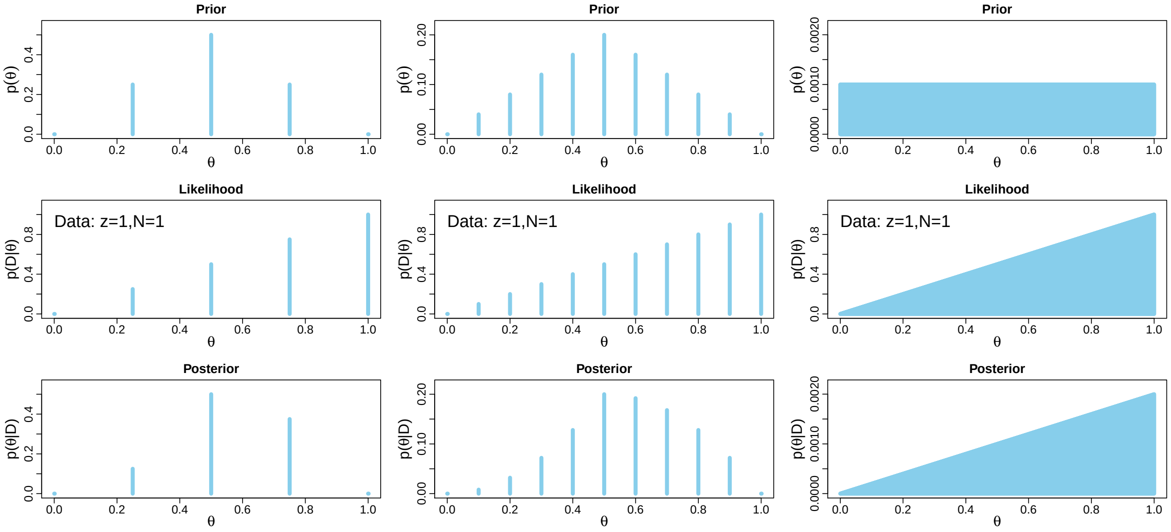
```
N = 10000000
nPresent = N*0.001 #10000
nPresNeg = nPresent*0.01
nPresNegPos = nPresNeg*0.99
nAbsent = N-nPresent #9990000
nAbsNeg = nAbsent*0.95
nAbsNegPos = nAbsNeg*0.05
nPresNegPos / (nPresNegPos + nAbsNegPos)
```

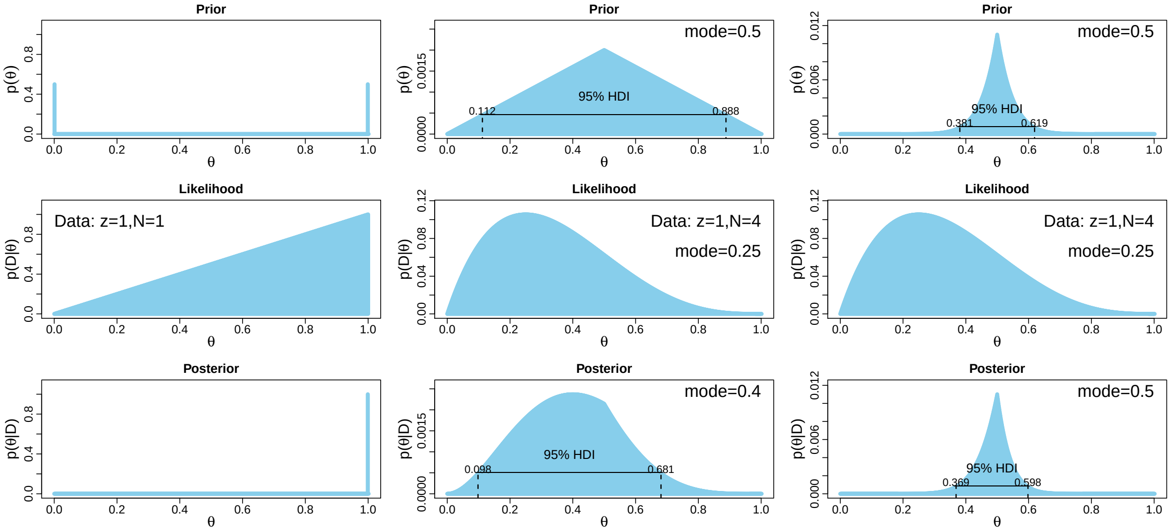
```
## [1] 0.0002085862
```

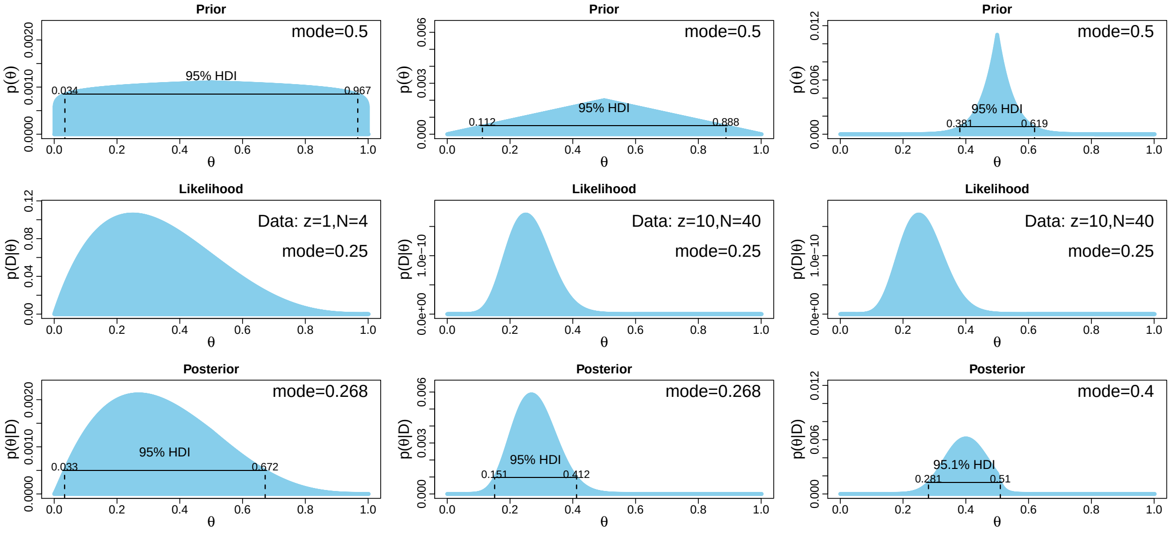
How does your answer compare to your answer to Exercise 5.1? Exactly the same. ***TLDR:*** *Reversing the order (negative then positive) gives same result.*

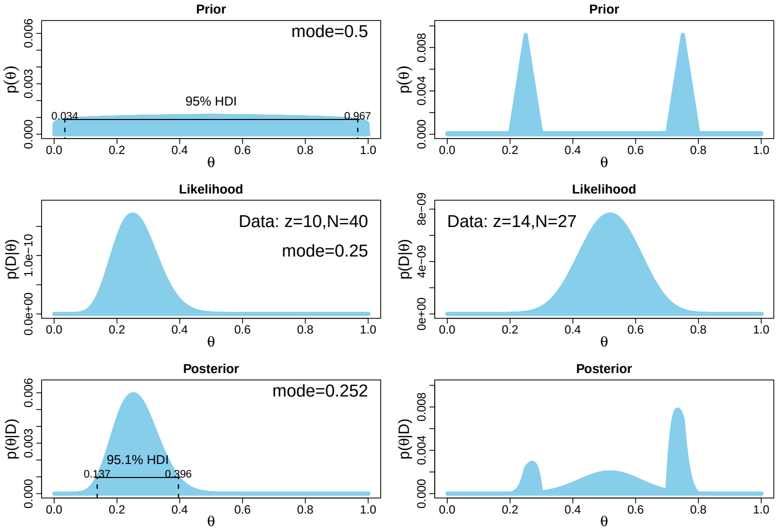
Exercise 5.4

TLDR: *Helpful diagrams for intuition of Bayesian updates.*


Prior
p(θ)
θ
Likelihood
Data: z=1,N=1
p(D|θ)
Posterior
p(θ|D)
0.0
0.2
0.4
0.6
0.8
1.0
0.10
0.20
0.00
0.0000
0.0010
0.0020


Prior
mode=0.5
95% HDI
0.112
0.888
0.381
0.619
Likelihood
Data: z=1,N=1
Data: z=1,N=4
mode=0.25
Posterior
mode=0.4
95% HDI
0.098
0.681
mode=0.5
0.369
0.598
p(θ)
p(D|θ)
p(θ|D)
θ
0.0015
0.0000
0.012
0.006
0.000
0.12
0.08
0.04
0.00


Prior
mode=0.5
95% HDI
0.034
0.967
0.112
0.888
0.381
0.619
Likelihood
Data: z=1,N=4
mode=0.25
Data: z=10,N=40
mode=0.25
Posterior
mode=0.268
95% HDI
0.033
0.672
mode=0.268
0.151
0.412
mode=0.4
95.1% HDI
0.281
0.51
p(θ)
p(D|θ)
p(θ|D)
θ
0.0000
0.0010
0.0020
0.000
0.003
0.006
0.012
0.0e+00
1.0e−10


Prior
mode=0.5
95% HDI
0.034
0.967
p(θ)
0.000
0.003
0.006
0.0
0.2
0.4
0.6
0.8
1.0
θ
Prior
p(θ)
0.000
0.004
0.008
Likelihood
Data: z=10,N=40
mode=0.25
p(D|θ)
0.0e+00
1.0e−10
Likelihood
Data: z=14,N=27
p(D|θ)
0e+00
4e−09
8e−09
Posterior
mode=0.252
95.1% HDI
0.137
0.396
p(θ|D)
0.000
0.003
0.006
Posterior
p(θ|D)
0.000
0.004
0.008