Biostatistics Homework 2

Lumi (12112618)

2023-03-03

```
library(ggplot2)

setwd('/Users/lumizhang/Documents/sustech/biology/classes/Biostatistics/Homework/hw2')
```

1 Lots of Coins

1.1

$$\begin{aligned} P(\text{getting a head}) &= P(\text{head}|\text{HH coin}) \cdot P(\text{HH coin}) + P(\text{head}|\text{HT coin}) \cdot P(\text{HT coin}) \\ &= 1 \times \frac{1}{9} + \frac{1}{2} \times \frac{8}{9} \\ &= \frac{5}{9} \end{aligned}$$

1.2

On the first four times of flip, we all get heads, that changes the possibility of the coin having two heads. It is much more likely than $\frac{1}{9}$ that the coin is a fair coin. So the new probability can be calculated with:

$$\begin{aligned} P(\text{HH}|\text{4H}) &= \frac{P(\text{4H}|\text{HH}) \cdot P(\text{HH})}{P(\text{4H})} \\ &= \frac{P(\text{4H}|\text{HH}) \cdot P(\text{HH})}{P(\text{4H}|\text{HT}) \cdot P(\text{HT}) + P(\text{4H}|\text{HH}) \cdot P(\text{HH})} \\ &= \frac{1 \times \frac{1}{9}}{1 \times \frac{1}{9} + (\frac{1}{2})^4 \times \frac{8}{9}} = \frac{2}{3} \\ P(\text{HT}|\text{4H}) &= 1 - P(\text{HH}|\text{4H}) = 1 - \frac{2}{3} = \frac{1}{3} \end{aligned}$$

Thus we can get the probability of getting another head, based on our knowledge that the first four tosses are all heads.

$$\begin{aligned} P(\text{another head}) &= P(\text{HH}|\text{4H}) \cdot P(\text{H}|\text{HH}) + P(\text{HT}|\text{4H}) \cdot P(\text{H}|\text{HT}) \\ &= \frac{2}{3} \times 1 + \frac{1}{3} \times \frac{1}{2} \\ &= \frac{5}{6} \end{aligned}$$

2 Detecting Virus with our Kit

2.1 & 2.2

$$P(E_1) = 0.01 \times 0.9 + 0.99 \times 0.05 = 0.0585 P(E_2) = 0.01 \times 0.95 + 0.99 \times 0.1 = 0.1085$$

2.3

E_1 and E_2 are not independent, because sadly, the test positive result of any one of the kit increases the probability that the person have actually get the virus. When he do a second test, the kit is more likely to show a positive result.

$$P(E_1 \cap E_2) \neq P(E_1) \cdot P(E_2)$$

2.4

Under the condition that the person is or is not carrying the virus, events E_1 and E_2 are actually independent. So:

$$\begin{aligned} P(E_3|E_1 \cap E_2) &= \frac{P(E_1 \cap E_2|E_3) \cdot P(E_3)}{P(E_1 \cap E_2)} \\ &= \frac{P(E_1|E_3) \cdot P(E_2|E_3) \cdot P(E_3)}{P(E_1 \cap E_2|E_3) \cdot P(E_3) + P(E_1 \cap E_2|E_3') \cdot P(E_3')} \\ &= \frac{0.9 \times 0.95 \times 0.01}{0.9 \times 0.95 \times 0.01 + 0.05 \times 0.1 \times 0.99} \\ &= \frac{19}{30} = 0.633... \end{aligned}$$

From this problem, we can also see that:

$$P(E_1 \cap E_2) = 0.0135 P(E_1) \cdot P(E_2) \neq 0.0135$$

So E_1 and E_2 are not independent.

3 Weird DIE Again

3.1

```
die_a <- c(1,2,3,4,5)
die_b <- c(1,2,3,4,5)
sample_space <- as.data.frame(expand.grid(die_a, die_b))
colnames(sample_space) <- c("die_a", "die_b")
sample_space
```

```
##    die_a die_b
## 1      1     1
## 2      2     1
## 3      3     1
## 4      4     1
## 5      5     1
## 6      1     2
## 7      2     2
## 8      3     2
## 9      4     2
## 10     5     2
## 11     1     3
## 12     2     3
## 13     3     3
## 14     4     3
## 15     5     3
```

```
## 16     1     4
## 17     2     4
## 18     3     4
## 19     4     4
## 20     5     4
## 21     1     5
## 22     2     5
## 23     3     5
## 24     4     5
## 25     5     5
```

```
sample_space$sum <- sample_space$die_a + sample_space$die_b
sample_space$A <- sample_space$sum == 10
sample_space$B <- sample_space$die_a == 5 | sample_space$die_b == 5
sample_space$C <- sample_space$die_a == 1 | sample_space$die_b == 1
sample_space
```

```
##    die_a die_b sum     A     B     C
## 1      1     1   2 FALSE FALSE  TRUE
## 2      2     1   3 FALSE FALSE  TRUE
## 3      3     1   4 FALSE FALSE  TRUE
## 4      4     1   5 FALSE FALSE  TRUE
## 5      5     1   6 FALSE  TRUE  TRUE
## 6      1     2   3 FALSE FALSE  TRUE
## 7      2     2   4 FALSE FALSE FALSE
## 8      3     2   5 FALSE FALSE FALSE
## 9      4     2   6 FALSE FALSE FALSE
## 10     5     2   7 FALSE  TRUE FALSE
## 11     1     3   4 FALSE FALSE  TRUE
## 12     2     3   5 FALSE FALSE FALSE
## 13     3     3   6 FALSE FALSE FALSE
## 14     4     3   7 FALSE FALSE FALSE
## 15     5     3   8 FALSE  TRUE FALSE
## 16     1     4   5 FALSE FALSE  TRUE
## 17     2     4   6 FALSE FALSE FALSE
## 18     3     4   7 FALSE FALSE FALSE
## 19     4     4   8 FALSE FALSE FALSE
## 20     5     4   9 FALSE  TRUE FALSE
## 21     1     5   6 FALSE  TRUE  TRUE
## 22     2     5   7 FALSE  TRUE FALSE
## 23     3     5   8 FALSE  TRUE FALSE
## 24     4     5   9 FALSE  TRUE FALSE
## 25     5     5  10  TRUE  TRUE FALSE
```

```
a <- sum(sample_space$A == TRUE)
b <- sum(sample_space$B == T)
c <- sum(sample_space$C == T)
ab <- sum(sample_space$A == T & sample_space$B == T)
ac <- sum(sample_space$A == T & sample_space$C == T)
sprintf("ab : %i/25, ac : %i/25", ab, ac)
```

```
## [1] "ab : 1/25, ac : 0/25"
```

We can see from the table and the printed results that

$$P(A \cap B) \neq P(A) \cdot P(B)$$
and
$$P(A \cap C) \neq P(A) \cdot P(B)$$

So A and B are not independent, and A and C are also not independent. Actually, A and C are mutually exclusive, so they are strongly dependent on each other.

3.2

```
sample_space$D <- sample_space$sum == 7
sample_space$E <- abs(sample_space$die_a - sample_space$die_b) == 1
sample_space$F <- sample_space$die_a < sample_space$die_b
e <- sum(sample_space$E == T)
f <- sum(sample_space$F == T)
ef <- sum(sample_space$E == T & sample_space$F == T)
sample_space[sample_space$F == T, c("E","F")]
```

```
##        E    F
## 6   TRUE TRUE
## 11 FALSE TRUE
## 12  TRUE TRUE
## 16 FALSE TRUE
## 17 FALSE TRUE
## 18  TRUE TRUE
## 21 FALSE TRUE
## 22 FALSE TRUE
## 23 FALSE TRUE
## 24  TRUE TRUE
```

$$P(E|F) = \frac{4}{10} \neq P(E) = \frac{8}{25}$$

So E and F are not independent.

```
sample_space[sample_space$D == T, c("E", "F")]
```

```
##        E     F
## 10 FALSE FALSE
## 14  TRUE FALSE
## 18  TRUE  TRUE
## 22 FALSE  TRUE
```

$$P(E|F) = 0.5 = P(E), P(F) \neq 0$$

So they are independent under the condition that D is true.

4 What's a CD-ROM?

4.1

Number of defected old CD-ROMs are : $500 \times 15\% = 75$

Number of defected new CD-ROMs are : $1500 \times 5\% = 75$

So,

the probability of buying two successive broken old CD-ROMs are $\frac{75}{500} \times \frac{74}{499} = \frac{111}{4990}$
the probability of buying two successive broken old CD-ROMs are $\frac{75}{1500} \times \frac{74}{1499} = \frac{37}{14990}$
so the probability of event A that both CD-ROMs being defected is:

$$P(A) = P(\text{old}) \cdot \frac{111}{4990} + P(\text{new}) \cdot \frac{37}{14990} \approx 0.0123$$

4.2

$$P(\text{old}|A) = \frac{P(A|\text{old}) \cdot P(\text{old})}{P(A)} = \frac{\frac{111}{4990} \times 0.5}{0.0123} = \frac{4497}{4996} \approx 0.9$$

5 Activating Transcription Factor 1

First we look at the RAS $\rightarrow$ PI3K $\rightarrow$ AKT $\rightarrow$ TF1 pathway. The probability of this pathway succeeding is

$$0.9 \times 0.8 \times 0.9 = 0.648$$

Then we look at the RAS $\rightarrow$ RAF $\rightarrow$ MEK $\rightarrow$ ERK1 or ERK2 $\rightarrow$ TF1 pathway. The probability of this pathway succeeding is

$$0.8 \times 0.9 \times (1 - (1 - 0.9 \times 0.8)(1 - 0.8 \times 0.8)) = 0.647424$$

So in order for the whole RTK $\rightarrow$ TF1 to succeed, at least one of the two pathways should succeed.

$$P(\text{TF1}) = 0.9 \times (1 - (1 - 0.648)(1 - 0.647424)) \approx 0.7883$$

6 Weird Signals

6.1

The k-th signal is an independent event. So the probability of event $A = \{\text{receive correctly}\}$ is

$$P(A) = p \cdot (1 - \epsilon_0) + (1 - p)\epsilon_1$$

6.2

Because receiving the signal are independent events, we know that

$$\begin{aligned} P(1011|1011) &= P(1|1) \cdot P(0|0) \cdot P(1|1) \cdot P(1|1) \\ &= (1 - \epsilon_1)(1 - \epsilon_0)(1 - \epsilon_1)(1 - \epsilon_1) \end{aligned}$$

6.3

Event A = three zeros. Event B = two zeros.

$$\begin{aligned} P(A) &= (1 - \epsilon_0)^3 \\ P(B) &= (1 - \epsilon_0)^2 \cdot \epsilon_0 \cdot \binom{3}{1} = 3(1 - \epsilon_0)^2\epsilon_0 \\ P(\text{correct}) &= P(A) + P(B) = (1 - \epsilon_0)^2(1 + 2\epsilon_0) \end{aligned}$$

6.4

We can see the improvement from this:

$$\frac{P(\text{correct}|3\text{times})}{P(\text{correct}|1\text{time})} = \frac{(1-\epsilon_0)^2(1+2\epsilon_0)}{1-\epsilon_0} = (1-\epsilon_0)(1+2\epsilon_0) = 1+\epsilon_0-2\epsilon_0^2$$

Because we want an improvement, we want $1+\epsilon_0-2\epsilon_0^2$ to be larger than 1. So:

$$\begin{aligned} 1+\epsilon_0-2\epsilon_0^2 &> 1 \\ \epsilon_0 &> 2\epsilon_0^2 \\ 0 < \epsilon_0 &< \frac{1}{2} \end{aligned}$$

6.5

$$\begin{aligned} P(0|101) &= \frac{P(101|0)\cdot P(0)}{P(101|0)\cdot P(0) + P(101|1)\cdot P(1)} \\ &= \frac{\epsilon_0^2(1-\epsilon_0)p}{\epsilon_0^2(1-\epsilon_0)p + (1-\epsilon_1)^2\epsilon_1(1-p)} \end{aligned}$$

7 Happy Prison

The important thing here is that out of three prisoners, two will be chosen. So no matter if our lovely prisoner of interest is chosen to be released or not, there will always be another prisoner from the other two prisoners to be chosen. In other words $P(B|A) = P(B|A')$. So

$$P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A')P(A')} = P(A)$$

So event A and event B are independent.

8 Searching in a Smarter Way

From intuition, the fact that we didn't find the object in cell i must cause a reduction of the probability that the object is in cell i, and an increase in the probability of finding the object in other cells.

We can prove this in a mathematical way (although mine might be wrong, but the results turned out to be ok). So, suppose event X_i means not finding the object in cell i and Π_i means the event of finding the object in the cell i (we use Π instead of π to differentiate these two).

$$\begin{aligned} P(\Pi_i|X_i) &= \frac{P(X_i|\Pi_i)P(\Pi_i)}{P(X_i)} \\ &= \frac{(1-p_i)\pi_i}{1-\pi_i p_i} \end{aligned}$$

We can prove that the updated information in the cell i actually leads to a decrease in probability of finding the object inside the cell. In other words $P(\Pi_i|X_i) < \pi_i$, since p_i doesn't change.

$$\frac{P(\Pi_i|X_i)}{\pi_i} = \frac{1-p_i}{1-\pi_i p_i} < 1, \text{thus } P(\Pi_i|X_i) < \pi_i$$

Because $\pi_i < 1$ so, $p_i > \pi_i p_i$, $1-\pi_i p_i > 1-p_i$. Which fits our guess that the updated probability should be smaller.

As for other cells, the possibility should be larger, $P(\Pi_j|X_i) > \pi_j$.

$$P(\Pi_j|X_i) = \frac{P(X_i|\Pi_j)P(\Pi_j)}{P(X_i)} = \frac{\pi_j}{1-\pi_i p_i}$$

Because $1 - \pi_i p_i$ is smaller than 1 but greater that 0, we can see that $P(\Pi_j|X_i) > \pi_j$.

The results are consistent with our intuitions.

Optional Java Practice

I have written a java program that can perform the function of Bayesian search on any size-determined area. The code can be found on my Github page, feel free to copy it and play.

The program will ask you whether you want to print the possibility of finding the thing if the thing is in the cell, whether you want to set the position of the thing (if not the position will be randomly generated), the size of the matrix, and the π and p values in each of the cells (you can use the auto generated values with $\pi = \frac{1}{n^2}$ and $p = 0.5$).

One example is shown here:

```
Welcome to the Bayesian Search Game!
The object of interest is hidden in a grid of cells.
Cell i, j means the cell in the i-th row, j-th column
By Lumi, 2023

Do you want to print probability of finding p (y/n):
y
Do you want to set the position of interest (y/n):
y
Enter the size of the cells:
5
Enter the position of interest:
3 1
Use preset values? (y/n)
n
Enter your values of pi and p for each cell:
0.005 0.7 0.002 0.6 0.001 0.3 0.007 0.6 0.007 1
0 0.1 0.001 0.7 0.002 0.8 0.52 0.9 0.002 0.8
0 0 0.001 0.8 0.002 0.9 0 0 0.004 0.6
0.005 1 0.43 0.4 0 0 0 0 0.002 0.7
0.001 0.7 0.003 0.8 0.002 0.7 0.002 0.3 0.001 0.9
Initial cell:
0.0050 (0.70) | 0.0020 (0.60) | 0.0010 (0.30) | 0.0070 (0.60) | 0.0070 (1.00) |
0.0000 (0.10) | 0.0010 (0.70) | 0.0020 (0.80) | 0.5200 (0.90) | 0.0020 (0.80) |
0.0000 (0.00) | 0.0010 (0.80) | 0.0020 (0.90) | 0.0000 (0.00) | 0.0040 (0.60) |
0.0050 (1.00) | 0.4300 (0.40) | 0.0000 (0.00) | 0.0000 (0.00) | 0.0020 (0.70) |
0.0010 (0.70) | 0.0030 (0.80) | 0.0020 (0.70) | 0.0020 (0.30) | 0.0010 (0.90) |

Checking cell 1 3
0.0094 (0.70) | 0.0038 (0.60) | 0.0019 (0.30) | 0.0132 (0.60) | 0.0132 (1.00) |
0.0000 (0.10) | 0.0019 (0.70) | 0.0038 (0.80) | 0.0977 (0.90) | 0.0038 (0.80) |
0.0000 (0.00) | 0.0019 (0.80) | 0.0038 (0.90) | 0.0000 (0.00) | 0.0075 (0.60) |
0.0094 (1.00) | 0.8083 (0.40) | 0.0000 (0.00) | 0.0000 (0.00) | 0.0038 (0.70) |
0.0019 (0.70) | 0.0056 (0.80) | 0.0038 (0.70) | 0.0038 (0.30) | 0.0019 (0.90) |
```

```
Checking cell 3 1
0.0139 (0.70) | 0.0056 (0.60) | 0.0028 (0.30) | 0.0194 (0.60) | 0.0194 (1.00) |
0.0000 (0.10) | 0.0028 (0.70) | 0.0056 (0.80) | 0.1444 (0.90) | 0.0056 (0.80) |
0.0000 (0.00) | 0.0028 (0.80) | 0.0056 (0.90) | 0.0000 (0.00) | 0.0111 (0.60) |
0.0139 (1.00) | 0.7167 (0.40) | 0.0000 (0.00) | 0.0000 (0.00) | 0.0056 (0.70) |
0.0028 (0.70) | 0.0083 (0.80) | 0.0056 (0.70) | 0.0056 (0.30) | 0.0028 (0.90) |

Checking cell 3 1
0.0195 (0.70) | 0.0078 (0.60) | 0.0039 (0.30) | 0.0273 (0.60) | 0.0273 (1.00) |
0.0000 (0.10) | 0.0039 (0.70) | 0.0078 (0.80) | 0.2025 (0.90) | 0.0078 (0.80) |
0.0000 (0.00) | 0.0039 (0.80) | 0.0078 (0.90) | 0.0000 (0.00) | 0.0156 (0.60) |
0.0195 (1.00) | 0.6028 (0.40) | 0.0000 (0.00) | 0.0000 (0.00) | 0.0078 (0.70) |
0.0039 (0.70) | 0.0117 (0.80) | 0.0078 (0.70) | 0.0078 (0.30) | 0.0039 (0.90) |

Checking cell 3 1
0.0257 (0.70) | 0.0103 (0.60) | 0.0051 (0.30) | 0.0359 (0.60) | 0.0359 (1.00) |
0.0000 (0.10) | 0.0051 (0.70) | 0.0103 (0.80) | 0.2668 (0.90) | 0.0103 (0.80) |
0.0000 (0.00) | 0.0051 (0.80) | 0.0103 (0.90) | 0.0000 (0.00) | 0.0205 (0.60) |
0.0257 (1.00) | 0.4766 (0.40) | 0.0000 (0.00) | 0.0000 (0.00) | 0.0103 (0.70) |
0.0051 (0.70) | 0.0154 (0.80) | 0.0103 (0.70) | 0.0103 (0.30) | 0.0051 (0.90) |

Checking cell 1 3
0.0338 (0.70) | 0.0135 (0.60) | 0.0068 (0.30) | 0.0473 (0.60) | 0.0473 (1.00) |
0.0000 (0.10) | 0.0068 (0.70) | 0.0135 (0.80) | 0.0351 (0.90) | 0.0135 (0.80) |
0.0000 (0.00) | 0.0068 (0.80) | 0.0135 (0.90) | 0.0000 (0.00) | 0.0270 (0.60) |
0.0338 (1.00) | 0.6272 (0.40) | 0.0000 (0.00) | 0.0000 (0.00) | 0.0135 (0.70) |
0.0068 (0.70) | 0.0203 (0.80) | 0.0135 (0.70) | 0.0135 (0.30) | 0.0068 (0.90) |

Checking cell 3 1
The object of interest is found after 5 rounds!
It is in 3 1
The final cell looks like
0.0338 (0.70) | 0.0135 (0.60) | 0.0068 (0.30) | 0.0473 (0.60) | 0.0473 (1.00) |
0.0000 (0.10) | 0.0068 (0.70) | 0.0135 (0.80) | 0.0351 (0.90) | 0.0135 (0.80) |
0.0000 (0.00) | 0.0068 (0.80) | 0.0135 (0.90) | 0.0000 (0.00) | 0.0270 (0.60) |
0.0338 (1.00) | 0.6272 (0.40) | 0.0000 (0.00) | 0.0000 (0.00) | 0.0135 (0.70) |
0.0068 (0.70) | 0.0203 (0.80) | 0.0135 (0.70) | 0.0135 (0.30) | 0.0068 (0.90) |

Congratulations!
```

We can see that when the cell (1,3) have a high possibility of having the object and a high possibility of finding the object, there is a great chance of finding the thing there. If it is not found in cell (1,3), the chance of the thing being there dramatically declined.

Although there is a great possibility of the object being in our cell of interest (3,1), the actual possibility of finding it there is low (0.4). So if you didn't find it in (3,1), the possibility won't change a lot.

Another interesting find is that, if you set the possibility of the thing being in the cell to zero, the Bayesian search method will always ignore the cell and never look into it, which is quite sad.

All in all, it's a fun practice.