BDA - Assignment 1

Anonymous

Contents

Loaded packages

```r
library(aaltobda)
library(stats)
library(ggplot2)
library(tinytex)
```

Exercise 1) Basic probability theory notation and terms

- **probability:** is a numerical quantity, defined on a set of 'outcomes,' that are nonnegative, additive over mutually exclusive outcomes, and sum to 1 over all possible mutually exclusive outcomes, and is used as the fundamental measure or yardstick of uncertainty in Bayesian statistics.

** Sidenote: It can also be described as a quantitative measure taking values between zero and one describing the chance of an event occurring, or the uncertainty related to an event.

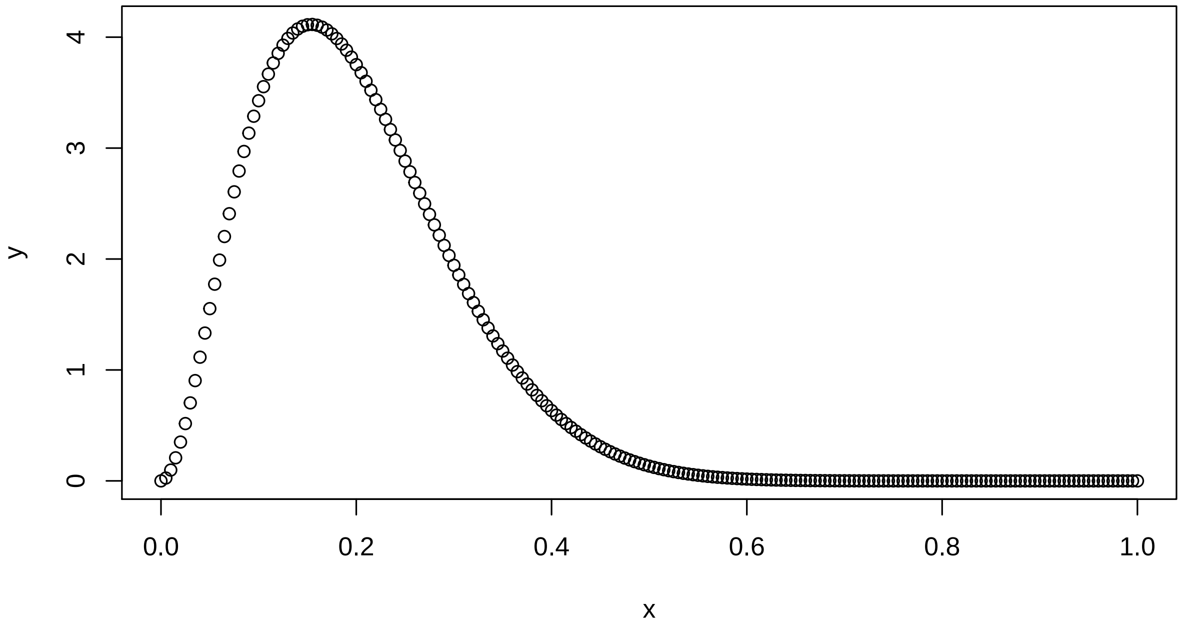
- **probability mass:** describes the amount of probability related to an event.
- **probability density:** describes the distribution of probability mass in the sample space.
- **probability mass function (pmf):** is a nonnegative function giving the probability of an outcome for a discrete random variable so that the function values sum upto one over all possible outcomes.
- **probability density function (pdf):** is a nonnegative function which can be interpreted to describe the relative plausibility of events for a continuous random variable.
- **probability distribution:** is a function relating probabilities to the possible outcomes in the sample space of a random variable, thus describing the chance of an event.
- **discrete probability distribution:** is a noncontinuous probability distribution defined by a probability mass function of a discrete random variable describing the chances of different outcomes.

- **continuous probability distribution:** is a probability distribution defined by a probability density function of a continuous random variable describing the plausibility of different outcomes.
- **cumulative distribution function (cdf):** is a function describing the cumulated probability up to some value x by giving the probability that a random variable may have a value less than or equal to x.
- **likelihood:** is a numeric quantity describing the plausibility of observations with respect to a probabilistic model

Exercise 2) Basic computer skills

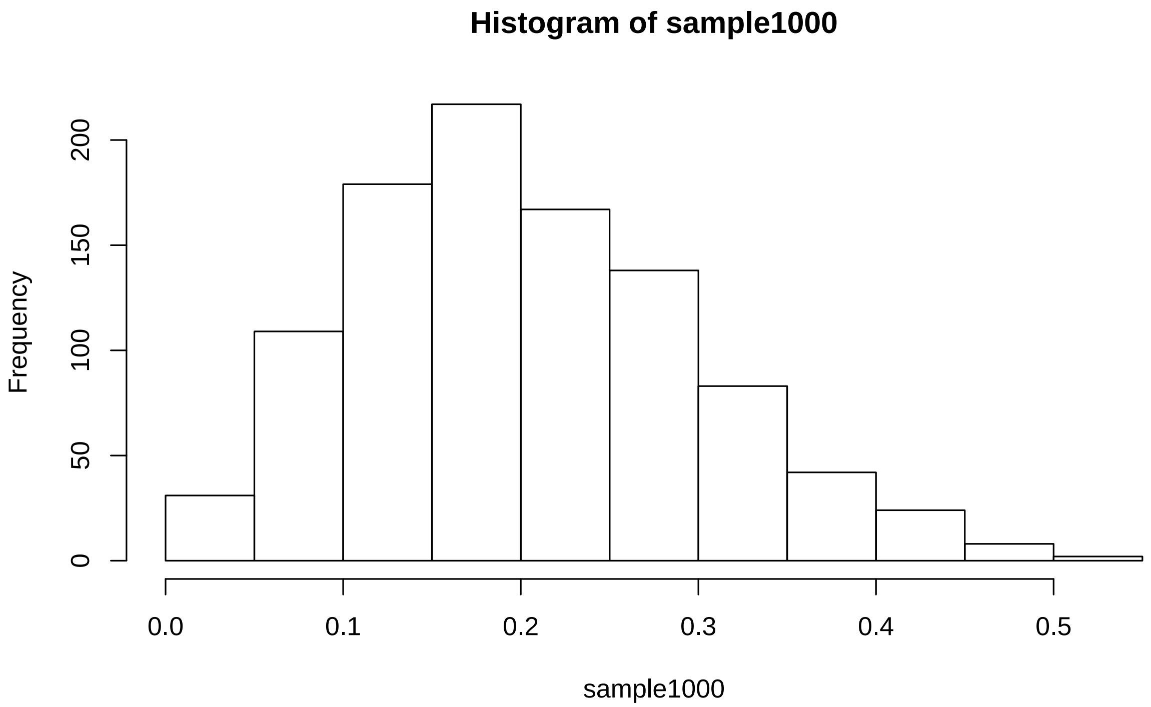
a)

```
mean = 0.2
var = 0.01
a = mean*((mean*(1-mean)/var)-1)
b = a*(1-mean)/mean
x = seq(0, 1, 0.005)
y = dbeta(x, a, b)
plot(x,y)
```



b)

```
sample1000 = rbeta(1000, a, b)
hist(sample1000)
```



Compare visually to the density function: The graphs from a) and b) have relatively quite similar shape.

c)

```
sample1000.mean <-mean(sample1000)
sample1000.var <-var(sample1000)
```

Verify that they match (roughly) to the true mean and variance of the distribution

```
lables=c("Sample Mean", "True Mean", "Sample Variance", "True Variance")
values=cbind(sample1000.mean, mean, sample1000.var, var)
rownames(values)="Value"
colnames(values)=lables
values
```

```
##       Sample Mean True Mean Sample Variance True Variance
## Value     0.20265       0.2     0.009020676          0.01
```

As we can see, the sample mean and true mean are quite close to each other. Sample Variance and True Variance are also relatively similar to each other.

d)

```r
quantile(sample1000, probs = c(0.025, 0.975))
```

```
##       2.5%      97.5% 
## 0.04634475 0.41204840
```

Based on the sample, the 95% probability interval is [0.05168726, 0.42538985].

Exercise 3) Bayes' theorem

Bayes' Theorem: $P(A|B) = \frac{P(B|A) \times P(A)}{P(B)}$

Let's assign:

A = Test gives positive. Therefore, A^C = Test gives negative

B = Subject has lung cancer. Therefore, B^C = Subject does not have lung cancer

We have:

$P(A|B) = 0.98 \rightarrow P(A^C|B) = 1 - 0.98 = 0.02$

$P(A^C|B^C) = 0.96 \rightarrow P(A|B^C) = 1 - 0.96 = 0.04$

$P(B) = \frac{1}{1000} = 0.001 \rightarrow P(B^C) = 1 - 0.001 = 0.999$

Thus:

$P(A) = P(A|B) \times P(B) + P(A|B^C) \times P(B^C) = 0.98 \times 0.001 + 0.04 \times 0.999 = 0.04094$

```r
0.98*0.001+0.04*0.999
```

```
## [1] 0.04094
```

$\rightarrow P(A^C) = 1 - 0.04094 = 0.95906$

```r
1-(0.98*0.001+0.04*0.999)
```

```
## [1] 0.95906
```

Using the Bayes' theorem, we have:

$P(B|A) = \frac{P(A|B) \times P(B)}{P(A)} = \frac{0.98 \times 0.001}{0.04094} = 0.02393747$

```r
(0.98*0.001)/(0.98*0.001+0.04*0.999)
```

```
## [1] 0.02393747
```

$\rightarrow P(B^C|A) = 1 - 0.02393747 = 0.9760625$

```
1-((0.98*0.001)/(0.98*0.001+0.04*0.999))
```

```
## [1] 0.9760625
```

$P(B|A^C) = \frac{P(A^C|B)\times P(B)}{P(A^C)} = \frac{0.02\times 0.001}{0.95906} = 2.085375e - 05$

```
(0.02*0.001)/0.95906
```

```
## [1] 2.085375e-05
```

$\rightarrow P(B^C|A^C) = 1 - (2.085375e - 05) = 0.9999791$

```
1-((0.02*0.001)/0.95906)
```

```
## [1] 0.9999791
```

As we can see there will be a massive amount of false positives (when a test says positive but the underlying condition is the subject not having a cancer, meaning that the test falsely says positive when it should not). In this case, the false positive rate is equal to $P(B^C|A) = 1 - 0.0239375 = 0.9760625$.

A false negative results when a test says negative but the underlying condition is the subject having a cancer, meaning that the test falsely says negative when it should not. In this case, the false negative rate is relatively low, $P(B|A^C) = 2.085375e - 05$. This low false negative rate may not pose a big problem .

Overall, I would advice them to not use this method for detecting lung cancer. It would be really expensive because of the high false positive rate, which leads to unnecessary administer medication. In general, high false positive rate is typically bad and undesirable in tests. Besides, though the false negative rate is relatively low, it shows that still there are possibilities that cancer cannot be detected correctly by using this test.

Exercise 4) Bayes' theorem

```
boxes <- matrix(c(2,4,1,5,1,3), ncol = 2,
dimnames = list(c("A", "B", "C"), c("red", "white")))
boxes
```

```
##   red white
## A   2     5
## B   4     1
## C   1     3
```

Let R be an an event picking up a red ball.

a)

The probability of picking up a red ball can be obtained using the total law of probability:

$P(R) = P(R|A)\times P(A) + P(R|B)\times P(B) + P(R|C)\times P(C) = \frac{2}{7}\times 0.4 + \frac{4}{5}\times 0.1 + \frac{1}{4}\times 0.5 = 0.3192857$

```
# This is P(R|A), P(R|B), P(R|C)
boxes[, "red"]/rowSums(boxes)
```

```
##         A         B         C 
## 0.2857143 0.8000000 0.2500000
```

```
prob_red <- sum(boxes[, "red"]/rowSums(boxes)*c(0.4, 0.1, 0.5))
#0.4 = P(A), 0.1 = P(B), 0.5 = P(C)
prob_red
```

```
## [1] 0.3192857
```

The probability of picking a red ball is approximately 0.3192857.

b)

Using Bayes' Theorem, we have:

$P(A|R) = \frac{P(R|A) \times P(A)}{P(R)} = \frac{\frac{2}{7} \times 0.4}{\frac{2}{7} \times 0.4 + \frac{4}{5} \times 0.1 + \frac{1}{4} \times 0.5} = 0.3579418$

$P(B|R) = \frac{P(R|B) \times P(B)}{P(R)} = \frac{\frac{4}{5} \times 0.1}{\frac{2}{7} \times 0.4 + \frac{4}{5} \times 0.1 + \frac{1}{4} \times 0.5} = 0.2505593$

$P(C|R) = \frac{P(R|C) \times P(C)}{P(R)} = \frac{\frac{1}{4} \times 0.5}{\frac{2}{7} \times 0.4 + \frac{4}{5} \times 0.1 + \frac{1}{4} \times 0.5} = 0.3914989$

```
# Bayes' Rule
prob_boxes <- (boxes[, "red"]/rowSums(boxes)*c(0.4, 0.1, 0.5))/prob_red
#0.4 = P(A), 0.1 = P(B), 0.5 = P(C)
prob_boxes
```

```
##         A         B         C 
## 0.3579418 0.2505593 0.3914989
```

If a red ball was picked, it most probably came from box C.

Exercise 5) Bayes' theorem

Let's assign:

F = Fraternal Twins

I = Idential Twins

TB = Twin Brother

We have:

$P(F) = \frac{1}{150}$

$P(I) = \frac{1}{400}$

Since:

Fraternal twins have two fertilized eggs and then could be of different sex

Identical twins have single egg divides into two separate embryos, so both have the same sex

Thus, we have:

$P(TB|F) = \frac{1}{4}$ (since fraternal twins can be either two boys or two girls or 1 girl and 1 boy or 1 boy and 1 girl)

$P(TB|I) = \frac{1}{2}$ (since identical twins can be either two boys or two girls)

```
fraternal_prob <- 1/150
identical_prob <- 1/400
prob_twin_brother_given_frat <- 1/4
prob_twin_brother_given_iden <- 1/2
```

We have:

P(F and TB) $= P(F) \times P(TB|F) = \frac{1}{150} \times \frac{1}{4} = \frac{1}{600} = 0.001666667$

```
prob_fraternal_twin_brother <- prob_twin_brother_given_frat*fraternal_prob
prob_fraternal_twin_brother
```

```
## [1] 0.001666667
```

P(I and TB) $= P(I) \times P(TB|I) = \frac{1}{400} \times \frac{1}{2} = \frac{1}{800} = 0.00125$

```
prob_identical_twin_brother <- prob_twin_brother_given_iden*identical_prob
prob_identical_twin_brother
```

```
## [1] 0.00125
```

P(TB) $= P(F) \times P(TB|F) + P(I) \times P(TB|I) = \frac{1}{150} \times \frac{1}{4} + \frac{1}{400} \times \frac{1}{2} = 0.002916667$

```
prob_twin_brother <- prob_fraternal_twin_brother+prob_identical_twin_brother
prob_twin_brother
```

```
## [1] 0.002916667
```

Thus, we have:

$P(I|TB) = \frac{P(TB|I) \times P(I)}{P(TB)} = \frac{\frac{1}{2} \times \frac{1}{400}}{\frac{1}{4} \times \frac{1}{150} + \frac{1}{2} \times \frac{1}{400}} = 0.4285714$

```
# Bayes' Rule
prob_identical_twin_given_twin_brother <- prob_identical_twin_brother/prob_twin_brother
prob_identical_twin_given_twin_brother
```

```
## [1] 0.4285714
```

So, the probability that Elvis was an identical twin is approximately 0.4285714.