Assignment 1 - Bayesian data analysis

anonymous

1 General information

```
library(aaltobda)
```

2 Basic probability theory notation and terms

probability: how likely is that a particular event will occur, bounded between 0 and 1.

probability mass: probability assigned to discrete values of a distribution. How likely a set (1 or more) of possible values of a discrete variable distribution are.

probability density: probability assigned to continuous values of a distribution. How likely a range of possible values of a continuous variable distribution are.

probability mass function: function that assigns to each value x of the discrete distribution the corresponding probability.

probability density function: function that assigns to each value x of the continuous distribution the corresponding density.

probability distribution: General concept including discrete and continuous variables. A function that returns the probability of occurrence for any value of the distribution.

discrete probability distribution: Probability distribution for a variable that can only take discrete values.

continuous probability distribution: Probability distribution for a variable that take continuous values.

cumulative distribution function (cdf): Function that returns for any value the probability that an occurrence of the corresponding random variable will have a value equal or less to it.

likelihood: probability to see some value(s) of a random variable given some parameter that specifies it's distribution.

3 Basic computer skills

Do some setup here. Explain shortly what you do.

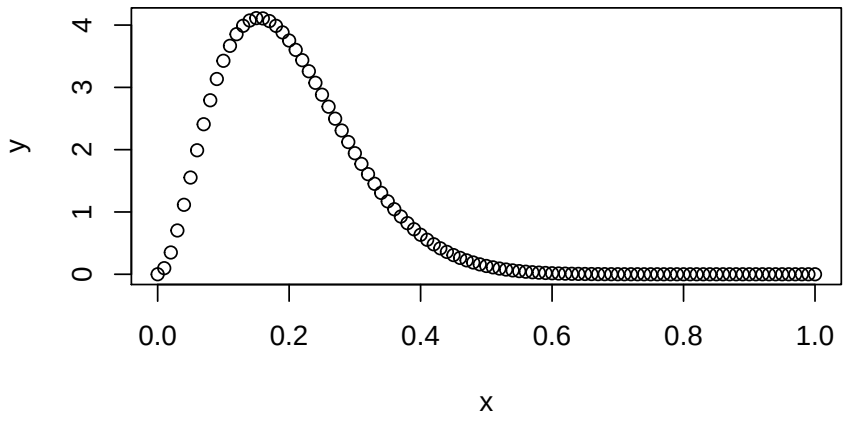
```
# Defining mean and variance  and deriving beta distribution parameters
mu=0.2
sigma2= 0.01

shape1 = mu*(mu*(1-mu)/sigma2 - 1)
shape2 = shape1*(1-mu)/mu
```

3.1 (a)

Plot the PDF here. Explain shortly what you do.

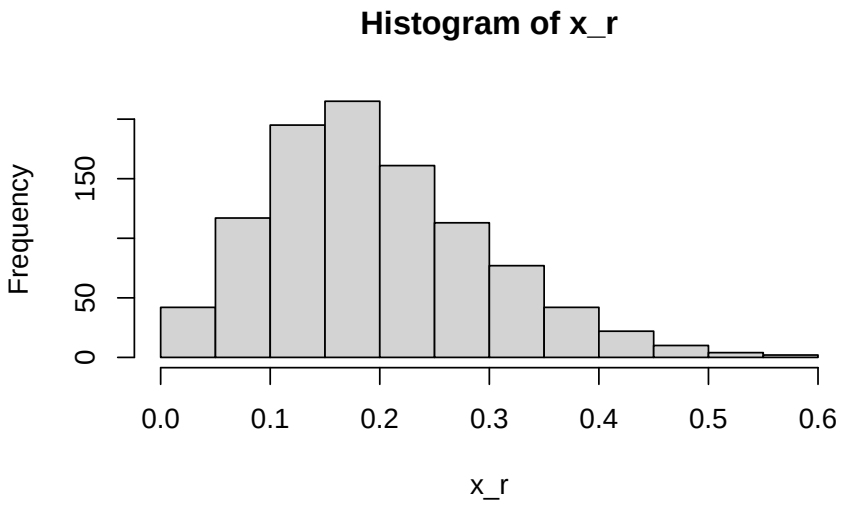
```
# we define the range of the x axis we want to plot
x = seq(0,1, 0.01)
# we get the density for each value of the range X given the parameters defined above and
y = dbeta(x, shape1=shape1, shape2=shape2)
plot(x,y)
```



3.2 (b)

Draw samples and plot the histogram here. Explain shortly what you do.

```
# random sampling 1000 value from the beta distribution with parameters defined above
x_r = rbeta(1000, shape1, shape2)
hist(x_r)
```



3.3 (c)

Compute the sample mean and variance here. Explain shortly what you do.

```
# taking the mean and variance from the sample we took in 3b
print(mean(x_r))
```

```
[1] 0.1972324
```

```
print(var(x_r))
```

```
[1] 0.009991785
```

3.4 (d)

Compute the central interval here. Explain shortly what you do.

```
# we get the central 95% of values from the sample we took in 3b. We leave 2.5% in each t
quantile(x_r, probs=c(0.025, 0.975))
```

```
     2.5%      97.5%
0.04305331 0.42978581
```

4 Bayes' theorem 1

Compute the quantities needed to justify your recommendation here. Explain shortly what you do.
We have the accuracy numbers in the general population but those have to be read with care given how unlikely is to have lung cancer in the general population. This means that false positives can be high in nominal quantities compared to true positives despite only being 4% of false positive rate.
What we want to inspect using Bayes theorem what is the probability of actually having lung cancer if the results is positive.
For that we will use Bayes Theorem and we need to pre compute some probabilities such as: * p_pos_given_notcancer * p_notcancer * p_pos

```
#setup
p_pos_given_cancer = 0.98
p_neg_given_notcancer = 0.96
p_cancer = 0.001

#pre compute
p_pos_given_notcancer = 1 - p_neg_given_notcancer
p_notcancer = 1 - p_cancer
p_pos = p_pos_given_cancer*p_cancer + p_pos_given_notcancer * p_notcancer

# Bayes theorem
p_cancer_given_pos = p_pos_given_cancer * p_cancer / p_pos
```

The probability of having cancer given a positive result is approximately 0.024, which is really low and further analysis should be done in terms of the costs (human and money related) of telling someone they have cancer with such a low confidence and all the further expenses related to that result.

5 Bayes' theorem 2

You will need to change the numbers to the numbers in the exercise.

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

5.1 (a)

Keep the below name and format for the function to work with `markmyassignment`:

```
p_red <- function(boxes) {
  # probability of picking each box
  box_prob = c(0.4, 0.1, 0.5)
  # total amount of balls per box
  box_balls = rowSums(boxes)
  # probability of picking a red ball in each box
  box_red_prob = boxes[,1] / box_balls
  # total probability of picking red. red probability by box times probability of each bo
  red_prob = sum(box_red_prob * box_prob)
  red_prob
}
```

The probability of picking a red ball is around 0.319.

5.2 (b)

Keep the below name and format for the function to work with `markmyassignment`:

```
p_box <- function(boxes) {
    box_prob = c(0.4, 0.1, 0.5)
    # total amount of balls per box
    box_balls = rowSums(boxes)
    # probability of picking a red ball in each box
    box_red_prob = boxes[,1] / box_balls
    # Bayes rule for each box
    result = box_red_prob * box_prob / p_red(boxes)
    result
}
```

The probabilities for each box given a red ball was picked are 0.358, 0.251, 0.391.

```r
for_report =p_box(boxes)
boxes_names = c('A','B','C')
max_prob_box = boxes_names[which.max(for_report)]
```

Given that, the most probable box is C

6 Bayes' theorem 3

You will need to change the numbers to the numbers in the exercise.

```r
fraternal_prob = 1/150
identical_prob = 1/400
```

Keep the below name and format for the function to work with `markmyassignment`:

```r
p_identical_twin <- function(fraternal_prob, identical_prob) {
  # We can think this a p_identical_twin given it's a twin of same gender
  # We apply Bayes rule
  p_samegender_given_identical = 1
  fraternal_given_twin = fraternal_prob / (fraternal_prob + identical_prob)
  identical_given_twin =  1 - fraternal_given_twin
  # we assume fraternal and identical are independent
  p_samegender = 0.5*fraternal_given_twin + 1*identical_given_twin

  p_samegender_given_identical * identical_given_twin / p_samegender

}
```

The probability Elvis had an identical twin brother is around 0.429