BDA - Assignment 1

Anonymous

Contents

Loaded Packages

```
library(aaltobda)

#library(markmyassignment)
```

Exercise 1)

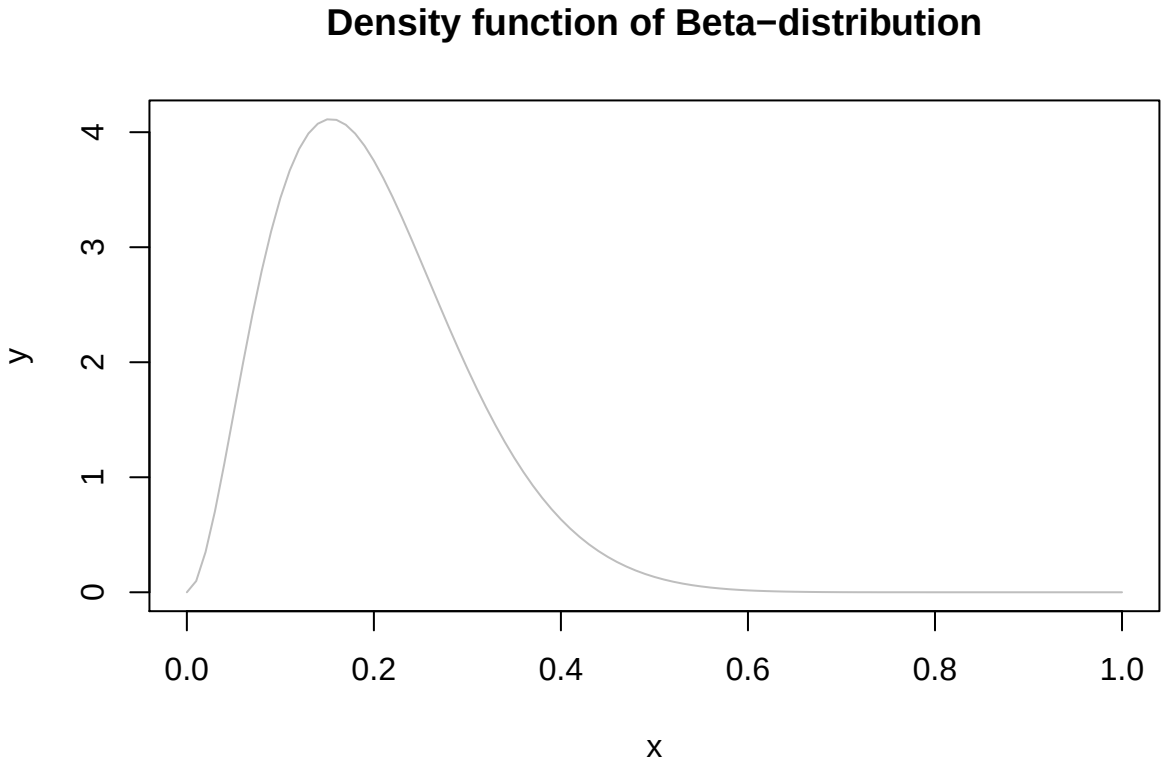
- Probability: given the sample space of an experiment, Ω, the probability in Ω, p, is a map such that $p : A \rightarrow p(A)$, where A is an event and $p(A) \in \mathbb{R}$, such that $0 \leq p(A) \leq 1$, $p(\Omega) = 1$ and, given $A_1, A_2..., A_n$, mutually exclusive events, then, $p(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} p(A_i)$.
- Probability mass: value of the probability mass function at a given point.
- Probability density: value of the probability density function at a given point.
- Probability mass function (pmf): this function (p) provides the probability that the random discrete variable is equal to a given sample: $p : \Omega \rightarrow [0, 1]$, where Ω is the sample space. From a mathematical point of view: $p(w) = Pr[X = w]$
- Probability density function (pdf): this function represents the relative probability at any sample that the value of the random continuous variable is equal to the given sample: $p(A) = \int_A f(t)dt$
- Probability distribution: this function provides the probability of an experiment for each of the potential outcomes. The probability distribution models the probability mass function or the probability density function.
- Discrete probability distribution: it is a probability distribution that can take just some certain values (it is not continuous). It represents the probability for each of the values that the discrete variable can take.
- Continuous probability distribution: it is a probability distribution that represents the probability of the variable taking any value.

- Cumulative distribution function (cdf): it enables the generalization of the probability mass function to continuous domains. The cumulative distribution function of the variable X is a monotonic function as: $F(q) = P(X \leq q)$. The probability of an interval is the difference of two cdf: $P(a < X < b) = F(b) - F(a)$. The derivative of the cumulative density function is the probability density function.
- Likelihood: According to the Bayes' Rule expression: $P(y|x) = \frac{P(x|y)P(y)}{P(x)} \rightarrow posterior = \frac{likelihood \cdot prior}{evidence}$, the likelihood represents, given the cause (y), the probability of the effect (x). In many occasions the observer only knows the outcomes of the experiment but they do not know the parameters underlying the stochastic process. The likelihood enables the computation of those parameters by maximizing its function.

Exercise 2)

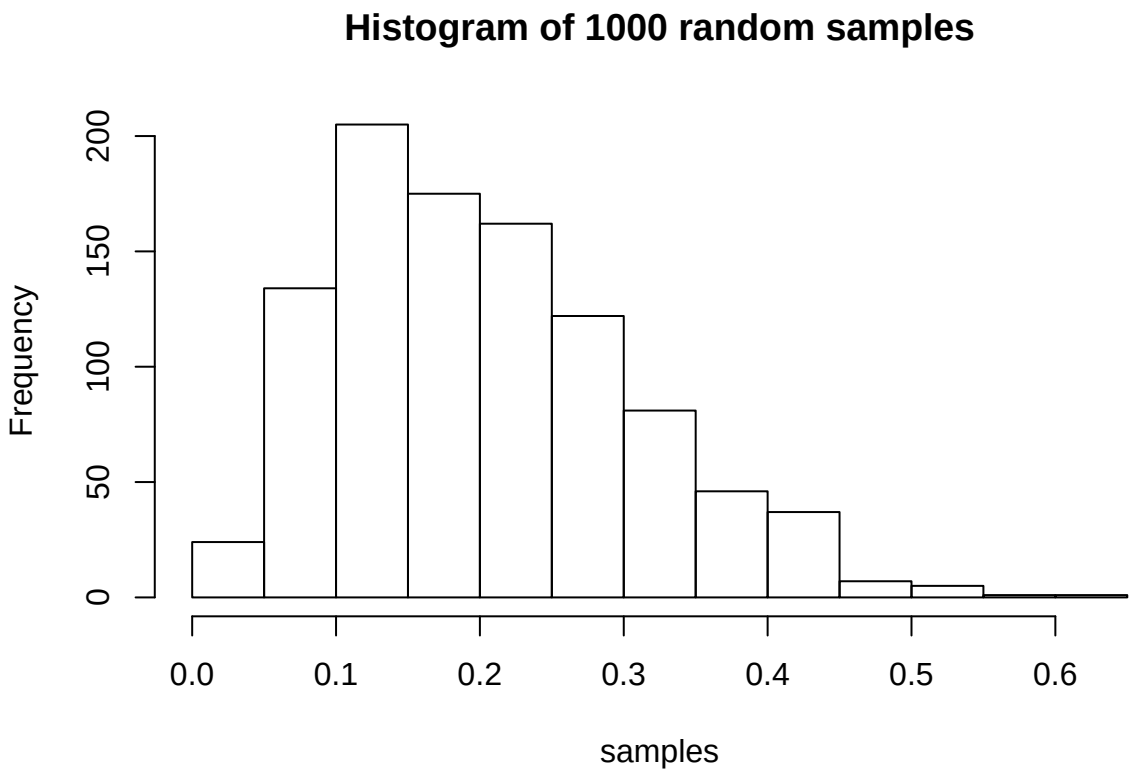
a)

```
x <- seq(from = 0, to = 1, by = 0.01)
u <- 0.2
s <- 0.01
alpha <- u*(u*(1-u)/s-1)
beta <- alpha*(1-u)/u
y <- dbeta(x, alpha, beta)
plot(x, y, type = "l", col = "grey", main = "Density function of Beta-distribution")
```



b)

```
samples <- rbeta(n = 1000, alpha, beta)
hist(samples, main = "Histogram of 1000 random samples")
```



As it is depicted in the previous figure, the histogram has a very similar shape to the Density function of the Beta-distribution. The greater the number of samples, the most accurate the shape is.

c)

```
#Mean from the drawn sample
mean(samples)
```

```
## [1] 0.2035585
```

```
#Variance from the drawn sample
var(samples)
```

```
## [1] 0.01065767
```

d)

Quantiles are points determining a range of the probability distribution (e.g, the quantile corresponding to 25% is the point for which the data of the distribution is less or equal than 0.25). The central 95% probability interval of the distribution from the drawn samples is the following:

$$p(x < a) = p(x > b) = 0.025$$

```
a <- quantile(samples, probs = 0.025)
b <- quantile(samples, probs = 0.975)
a
```

```
##      2.5%
## 0.0519577
```

```
b
```

```
##     97.5%
## 0.4334861
```

The interval is [0.0519577, 0.4334861]

Exercise 3)

Considering the following variables:

$$P\text{: positive test result.}$$

$$\bar{P}\text{: negative test result.}$$

$$L\text{: lung cancer.}$$

$$\bar{L}\text{: no lung cancer.}$$

Expressing the data provided by the exercise using the previously defined variables: The probability that the test gives a positive result when the subject has lung cancer is 95%:

$$p(P|L) = 98\% = 0.98$$

The probability that the test gives a negative result when the subject does not have lung cancer is 96%:

$$p(\bar{P}|\bar{L}) = 96\% = 0.96$$

The probability of having lung cancer is 0.1%:

$$p(L) = 0.1\% = 0.001$$

According to the previous expressions, the probability of a false negative result and the that of a false positive can be computed as following: Probability of a false negative result, meaning that the test provides a negative result when the patient has lung cancer:

$$p(\bar{P}|L) = 1 - p(P|L) = 1 - 0.98 = 0.02 = 2\%$$

Probability of a false positive result, meaning that the test provides a positive result when the patient does not have lung cancer:

$$p(P|\bar{L}) = 1 - p(\bar{P}|\bar{L}) = 1 - 0.96 = 0.04 = 4\%$$

The following "matrix" summarizes the data:

True / Predicted	Positive	Negative
Positive	TP: 0.98	FN: 0.02
Negative	FP: 0.04	TN: 0.96

Analogously, the probability of not having lung cancer is:

$$p(\bar{L}) = 1 - p(L) = 1 - 0.001 = 0.999 = 99.9\%$$

The probability of being tested positive can be computed with Using the Bayes' Rule, the following computa-

tions can be performed: Probability of having lung cancer when tested positive:

$$p(L/P) = \frac{p(P/L)p(L)}{p(P)} = \frac{0.98 \cdot 0.001}{p(P)}$$

Probability of being tested positive can be computed considering as mutually exclusive events being tested positive and having lung cancer and being tested positive and not having lung cancer:

$$p(P) = p(P|L)p(L) + p(P|\bar{L})p(\bar{L}) = 0.98 \cdot 0.001 + 0.04 \cdot 0.999 = 0.041 = 4.1\%$$

The probability of having lung cancer when being tested positive can be computed as following:

$$p(L|P) = \frac{p(P|L)p(L)}{p(P)} = \frac{0.98 \cdot 0.001}{0.041} \approx 0.024 = 2.4\%$$

The probability of not having lung cancer when being tested positive would be:

$$p(\bar{L}|P) = \frac{p(P|\bar{L})p(\bar{L})}{p(P)} = \frac{0.04 \cdot 0.999}{0.041} \approx 0.976 = 97.6\%$$

The probability of not having lung cancer when being tested negative would be:

$$p(\bar{L}|\bar{P}) = \frac{p(\bar{P}|\bar{L})p(\bar{L})}{p(\bar{P})} = \frac{0.96 \cdot 0.999}{1 - 0.041} \approx 0.999 = 99.9\%$$

The probability of having lung cancer when being tested negative would be:

$$p(L|\bar{P}) = \frac{p(\bar{P}|L)p(L)}{p(\bar{P})} = \frac{0.02 \cdot 0.001}{1 - 0.041} \approx 0.000021 = 0.0021\%$$

Finally, the joint probability can be computed:

$$p(P, L) = p(P|L)p(L) = 0.98 \cdot 0.001 = 0.00098 = 0.098\%$$

As a conclusion, I would recommend the researches to improve their test before releasing it into the market. The probability of being tested positive while having lung cancer is very high. However, the test fails in many occasions, since the probability of having lung cancer while being tested positive is extremely low (2.4%). This means that even if the subject was tested positive, it is highly probable that they will not have lung cancer (97.6%). On the other hand, the error when testing negative is low, the probability of not having lung cancer when the result was negative is very high, over 99%.

Exercise 4)

The probability of choosing the box C is the amount left after subtracting the probabilities of choosing A and B from the total: $p(C) = 1 - p(A) - p(B)$. The variable "boxes" was defined as following:

Box	Red	White
A	2	5
B	4	1
C	1	3

Since the events of extracting a red ball from each of the boxes are mutually exclusive: $R = RA + RB + RC$ (RA, RB and RC are mutually exclusive), then, by using the Law of the total probability:

$p(B) = \sum_{i=1}^{n} p(B/A_i)p(A_i)$, the following expression is obtained:

$$p(R) = p(R|A)p(A) + p(R|B)p(B) + p(R|C)p(C) = \frac{2}{7} \cdot 0.4 + \frac{4}{5} \cdot 0.1 + \frac{1}{4} \cdot 0.5 = 0.3193$$

, where p(R) is the probability of choosing a red ball.

The function that computes the probability for each of the boxes, given that a red ball was extracted is based on the Bayes' Rule, so:

$$p(A|R) = \frac{p(R|A)p(A)}{p(R)} = \frac{\frac{2}{7} \cdot 0.4}{0.3193}$$

$$p(B|R) = \frac{p(R|B)p(B)}{p(R)} = \frac{\frac{4}{5} \cdot 0.1}{0.3193}$$

$$p(C|R) = \frac{p(R|C)p(C)}{p(R)} = \frac{\frac{1}{4} \cdot 0.5}{0.3193}$$

CODE

```
boxes <- matrix(c(2,4,1,5,1,3), ncol = 2, dimnames = list(c("A", "B", "C"),
c("red", "white")))
p_A <- 0.4
p_B <- 0.1
p_C <- 1-p_A-p_B

#Function
#p_red computes the probability of getting a red ball.
#It requires the matrix that contains the data from the boxes
p_red <- function(boxes) {
  p_R <- boxes[1,1]/(boxes[1,1]+boxes[1,2])*p_A + boxes[2,1]/(boxes[2,1]+boxes[2,2])*p_B +
boxes[3,1]/(boxes[3,1]+boxes[3,2])*p_C
  return(p_R)
}


#Function
#p_box computes the probability of each box given that the extracted ball is red.
#It requires the matrix that contains the data from the #boxes
p_box <- function(boxes) {
  #Probability that the box is A given that the extracted ball was red
  p_ar = (boxes[1,1]/(boxes[1,1]+boxes[1,2])*p_A)/(boxes[1,1]/(boxes[1,1]+boxes[1,2])*p_A+
boxes[2,1]/(boxes[2,1]+boxes[2,2])*p_B + boxes[3,1]/(boxes[3,1]+boxes[3,2])*p_C)

  #Probability that the box is B given that the extracted ball was red
  p_br = (boxes[2,1]/(boxes[2,1]+boxes[2,2])*p_B)/(boxes[1,1]/(boxes[1,1]+boxes[1,2])*p_A+
boxes[2,1]/(boxes[2,1]+boxes[2,2])*p_B + boxes[3,1]/(boxes[3,1]+boxes[3,2])*p_C)

  #Probability that the box is C given that the extracted ball was red
  p_cr = (boxes[3,1]/(boxes[3,1]+boxes[3,2])*p_C)/(boxes[1,1]/(boxes[1,1]+boxes[1,2])*p_A+
boxes[2,1]/(boxes[2,1]+boxes[2,2])*p_B + boxes[3,1]/(boxes[3,1]+boxes[3,2])*p_C)

  results <- c(p_ar, p_br, p_cr)
  return(results)
}
```

```r
#Probability of getting a red ball
p_red <- p_red(boxes)
p_red
```

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

```r
#Probability of each of the boxes knowing that the extracted ball was red
p_box <- p_box(boxes)
p_box
```

```
## [1] 0.3579418 0.2505593 0.3914989
```

Exercise 5)

The function provides the probability that Elvis had an identical twin brother. To compute, there must be taken into account that the probability of being both boys when identical twins is 50% (Identical twins have the same gender), while the probability of being both boys when fraternal twins is 25% (fraternal twins might have different gender, so there are 4 possible combinations b-b, b-g, g-g, g-b). With the previous information, the following joint probabilities can be computed:

$$p(identicaltwins, bothboys) = p(identicaltwins)p(bothboys|identicaltwins) = \frac{1}{400} \cdot 0.5$$

$$p(fraternaltwins, bothboys) = p(fraternaltwins)p(bothboys|fraternaltwins) = \frac{1}{125} \cdot 0.25$$

Finally, knowing that Elvis had a twin brother, the probability that he was an identical twin is:

$$p(identicaltwin|bothboys) = \frac{p(identicaltwin, bothboys)}{p(bothboys)}$$

```r
fraternal_prob <- 1/150
identical_prob <- 1/400

#Function that computes the probability of having an identical twin
p_identical_twin <- function(fraternal_prob, identical_prob){
  p_identical_twinbrother <- identical_prob * 1/2
  p_fraternal_twinbrother <- fraternal_prob * 1/4

  p_cond_identical_twinbrother = p_identical_twinbrother/(p_identical_twinbrother +
p_fraternal_twinbrother)

  return(p_cond_identical_twinbrother)
}

p_identical_twin <- p_identical_twin(fraternal_prob, identical_prob)
p_identical_twin
```

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