BDA - Assignment 1

9/15/2021

Contents

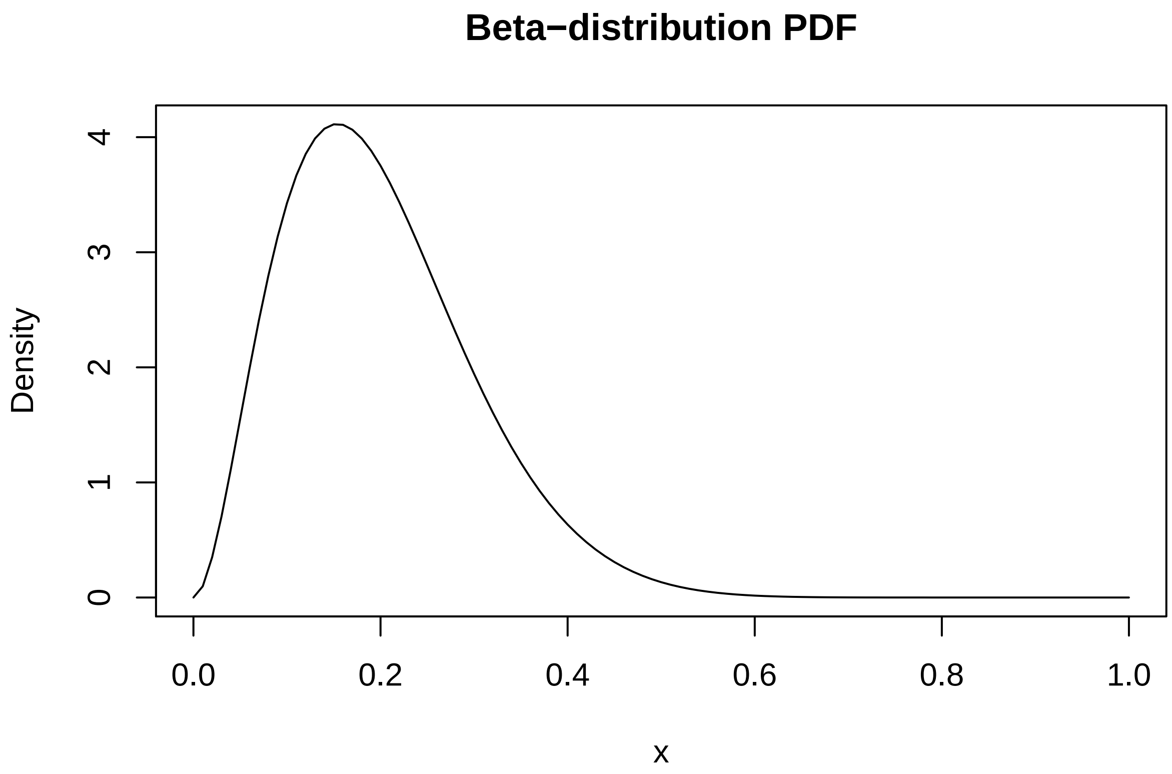
1. Basic probability theory notation and terms

- probability: The process used to predict what will happen (in the future).
- probability mass: The probability that a discrete random variable will have exactly some value.
- probability density: The relative (to other points) likelihood that a continuous random variable would take a given value.
- probability mass function (pmf): A function of the probability mass which gives us the probability that the random variable takes on a certain discrete value.
- probability density function (pdf): A function of the probability density which gives us the probability of the random variable falling in some range.
- probability distribution: The mathematical function which takes in the sample space of the variable and outputs the probabilities of events.
- discrete probability distribution: A probability distribution of a random variable that only take on discrete values.
- continuous probability distribution: A probability distribution of a random variable of real values.
- cumulative distribution function (cdf): The cumulative distribution function gives the probability that $P(X \leq x)$, i.e. the probability that the random variable takes on at least the value x.
- likelihood: The probability of seeing a certain outcome of a underlying defined model.

2. Basic computer skills

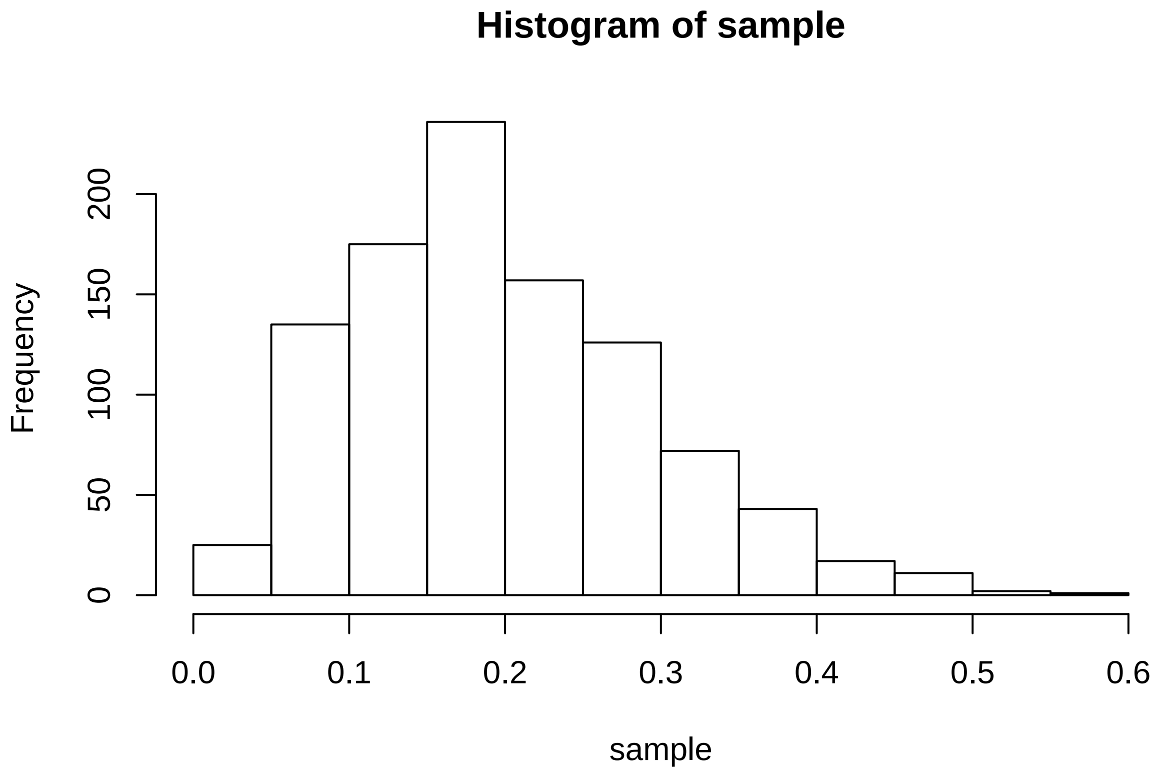
a)

```r
rm(list = ls())
mu = 0.2
sigma2=0.01
alpha=mu*((mu*(1-mu)/sigma2)-1)
beta=alpha*(1-mu)/mu
x = seq(from=0,to=1, by=0.01)
betapdf = dbeta(x,alpha,beta)
plot(x, betapdf,
     type='l',
     main='Beta-distribution PDF',
     ylab= 'Density')
```



b)

```
n = 1000
sample = rbeta(n,alpha,beta)
hist(sample)
```



We can see that the density function has a very similar shape to the histogram.

c)

```
mean = mean(sample)
variance = var(sample)
paste0("Sample mean: ", mean)
```

```
## [1] "Sample mean: 0.197832207641283"
```

```
paste0("Sample variance: ",variance)
```

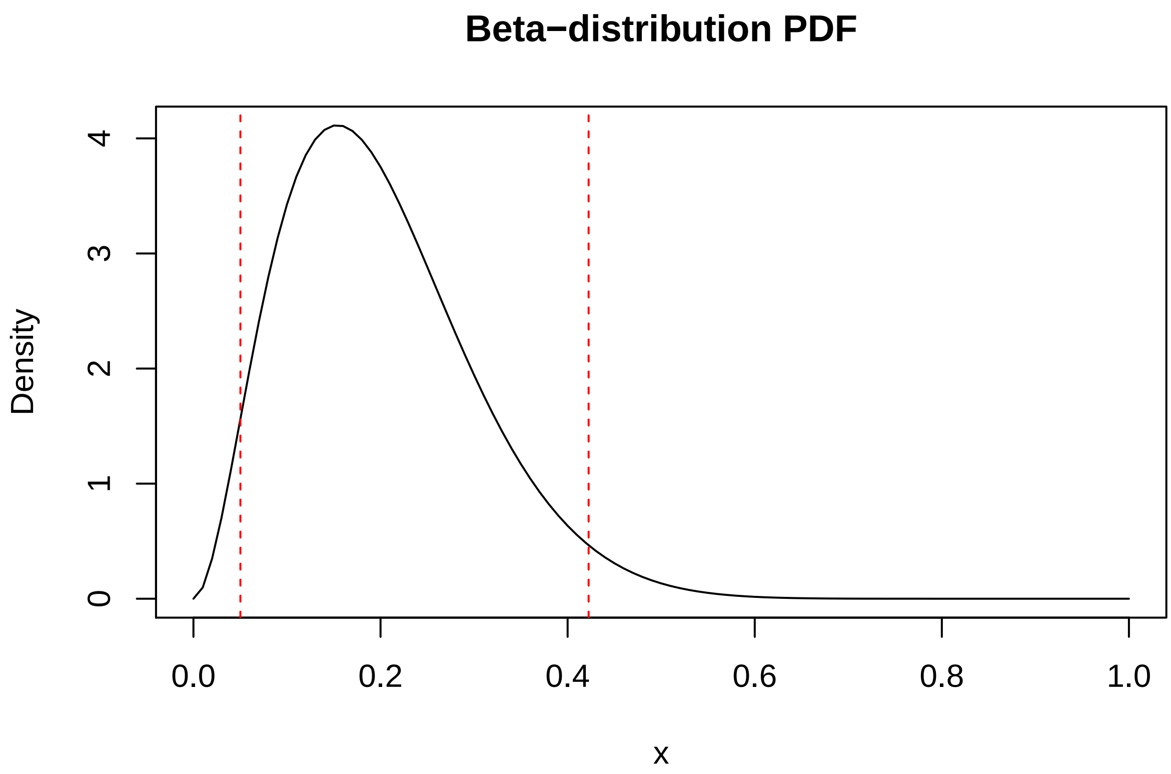
```
## [1] "Sample variance: 0.00906469894581722"
```

We can see that the mean and the true mean are roughly the same The true variance σ^2 and the sample variance are also very close to each.

d)

```
CI95=quantile(sample,probs=c(0.025,0.975))
plot(x,betapdf,
     type='l',
     main='Beta-distribution PDF',
     ylab= 'Density')
abline(v = CI95[[1]],
     lty = 2,
     col='red')
```

```
abline(v = CI95[[2]],
       lty = 2,
       col='red')
```



The central 95% probability interval is approximately between 0.05 and 0.42. The dashed lines highlight the interval edges on the density function plot.

3. Bayes' theorem

```
rm(list = ls())
P_test_pos_con_cancer_pos = 0.98
P_test_neg_con_cancer_neg = 0.96
P_cancer_pos = 1/1000
P_cancer_neg = 1 - P_cancer_pos
P_test_neg_con_cancer_pos = 1 - P_test_pos_con_cancer_pos
P_test_pos_con_cancer_neg = 1 - P_test_neg_con_cancer_neg
P_test_pos_and_cancer_pos = P_test_pos_con_cancer_pos * P_cancer_pos
P_test_neg = P_test_neg_con_cancer_neg * P_cancer_neg + P_test_neg_con_cancer_pos * P_cancer_pos
P_test_pos = 1-P_test_neg
P_cancer_pos_con_test_neg = (P_test_neg_con_cancer_pos * P_cancer_pos) /  P_test_neg
P_cancer_pos_con_test_pos = P_test_pos_con_cancer_pos * P_cancer_pos/ P_test_pos
cat(paste0("P(cancer=positive|test=neg)=",round(P_cancer_pos_con_test_neg*100,4),"%\n"),sep="")
```

```
## P(cancer=positive|test=neg)=0.0021%
```

```
cat(paste0("P(cancer=positive|test=positive) = ",round(P_cancer_pos_con_test_pos*100,4),"%"),sep="")
```

```
## P(cancer=positive|test=positive) = 2.3937%
```

We compute the following numbers: P(cancer=positive|test=neg) = 0.002% P(cancer=positive|test=positive) = 2.3%. This means that so called false negatives are extremely rare, but true positives are also very low. This means that very few of those who test positive actually have cancer, only 2,3%, meaning that it will send approximately 50 times too many people to further tests. Thus, the test does not do what it's planned to do, i.e. reduce the amount of expensive tests, very well.

4. Bayes' theorem

a)

```
rm(list = ls())
p_red = function(boxes){
  P_A = 0.4
  P_B = 0.1
  P_C = 1 - P_A - P_B
  P_red_con_A = boxes[1,1]/rowSums(boxes)[1]
  P_red_con_B = boxes[2,1]/rowSums(boxes)[2]
  P_red_con_C = boxes[3,1]/rowSums(boxes)[3]
  res = P_red_con_A*P_A + P_B*P_red_con_B + P_red_con_C * P_C
  return(res)
}

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

```
##         A
## 0.3192857
```

Using the law of total probability and bayes' rule we can calculate that there is a approximately 28.4% of picking a red ball.

b)

```
p_box = function(boxes){
  P_A = 0.4
  P_B = 0.1
  P_C = 1 - P_A - P_B
  P_red_res = p_red(boxes = boxes)
  P_red_con_A = boxes[1,1]/rowSums(boxes)[1]
  P_red_con_B = boxes[2,1]/rowSums(boxes)[2]
  P_red_con_C = boxes[3,1]/rowSums(boxes)[3]
  P_A_con_red = P_red_con_A * P_A / P_red_res
  P_B_con_red = P_red_con_B * P_B / P_red_res
  P_C_con_red = P_red_con_C * P_C / P_red_res
  res = c(P_A_con_red,P_B_con_red,P_C_con_red)
  return(res)
}
p_box(boxes = boxes)
```

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

Using Bayes' rule we can compute the different probabilities of which box we have likely picked when we have picked a red ball. Based on our result, visible above, we can conclude that it's most probable to be box C with a probability of 39,14% .

5. Bayes' theorem

```
p_identical_twin = function(fraternal_prob, identical_prob){
  P_boy = 1/2
  P_ident_and_twin_boy = P_boy*identical_prob
  P_frat_and_twin_boy = P_boy*P_boy*fraternal_prob
  P_ident_con_twin_boy = P_ident_and_twin_boy/(P_ident_and_twin_boy+P_frat_and_twin_boy)
  return(P_ident_con_twin_boy)
}
P_res = p_identical_twin(fraternal_prob = 1/150, identical_prob = 1/400)
paste0("Probability of identical twin: ", round(P_res,4)*100,"%")
```

```
## [1] "Probability of identical twin: 42.86%"
```

We can see that by using Bayes' rule we can compute the probability of Elvis being a identical twin is approximately 42.86%.