BDA - Assignment 1

Contents

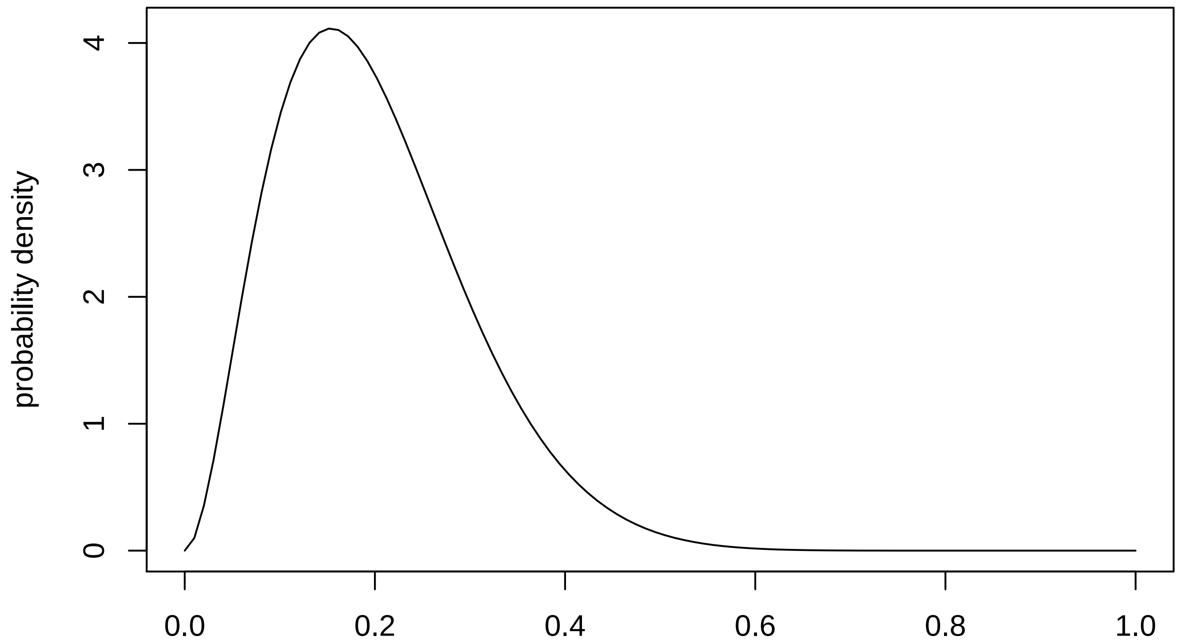
Exercise 1.

Probability is the likelihood or chance of an event occurring.
A probability mass function (PMF) gives probabilities for discrete random variables.
So a probability mass is the probability for a discrete random variable.
Probability density is the density of a probability distribution of a continuous variable. Accumulation of probability density gives probability.
Probability density function (PDF) is a statistical expression that defines a probability distribution for a continuous random variable as opposed to a discrete random variable.
A probability distribution is a function that describes the likelihood of obtaining the possible values that a random variable can assume.
A discrete distribution describes the probability of occurrence of each value of a discrete random variable.
So then Continuous probability distribution is a type of distribution that deals with continuous types of data or random variables.
A likelihood is a function of parameters within the parameter space that how probable a given set of observations is for different values of statistical parameters.

Exercise 2.

a)

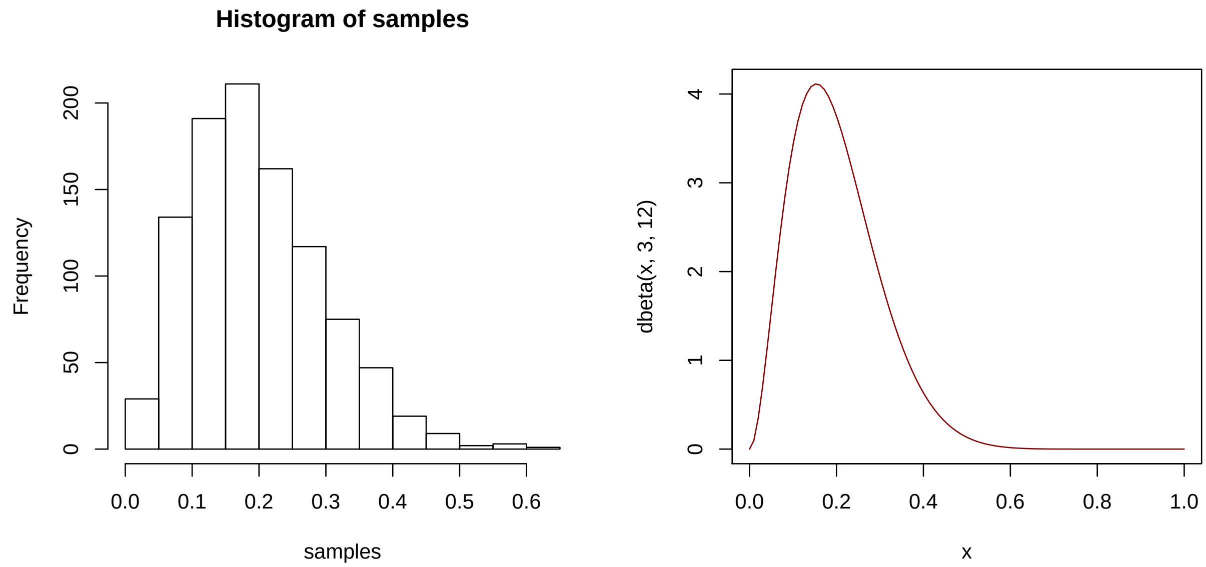
```
x<-seq(0,1,length=100)
plot(x,dbeta(x,3,12),"l",sub="density function of Beta-distribution with
     alpha=3 and beta=12",ylab="probability density")
```



density function of Beta−distribution with
alpha=3 and beta=12

b)

```
par(mfcol=c(1,2))
samples<-rbeta(1000,3,12)
hist(samples)
plot(x,dbeta(x,3,12),"l",col = "dark red")
```



c)

```
mean(samples)
```

```
## [1] 0.1975164
```

```
var(samples)
```

```
## [1] 0.009854737
```

d)

```
quantile(samples,probs=c(0.025,0.975))
```

```
##       2.5%      97.5%
## 0.04786022 0.42105764
```

Exercise 3.

Basically we have several given probabilities: $P(Positive|Cancer) = 0.98$, $P(Negative|\overline{Cancer}) = 0.96$ and $P(Cancer) = 0.001$. Taking about success rate, actually if we think over it carefully again, the success rate should be $P(success) = P(cancer|positive) = \frac{P(positive|cancer)*P(cancer)}{P(positive)}$ where $P(positive) = P(cancer) * P(positive|cancer) + P(\overline{cancer}) * (1 - P(negative|\overline{cancer}) = 0.04094$. So then $P(success) \approx 0.024 = 2.4\%$ which is much lower than what the researchers expected. Based on the formula, my suggestion is that they should lower down the rate of false postive, it is because that in most cases the examinees don't have cancer so decrease the number of false positive will increase the success rate much more than increasing the ability to distinguish cancer patients from examinees.

Exercise 4.

Formula used in a): $P(red) = P(A) * P(red|A) + P(B) * P(red|B) + P(C) * P(red|C)$
Formula used in b): $P(A|red) = \frac{P(red|A)*P(A)}{P(red)}$ $P(B|red) = \frac{P(red|B)*P(B)}{P(red)}$ $P(C|red) = \frac{P(red|C)*P(C)}{P(red)}$

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

p_red<-function(boxes){
pred1=boxes[1,1]/(boxes[1,1]+boxes[1,2])
pred2=boxes[2,1]/(boxes[2,1]+boxes[2,2])
pred3=boxes[3,1]/(boxes[3,1]+boxes[3,2])
return(pred1*0.4+pred2*0.1+pred3*0.5)
}

p_box<-function(boxes){
pred1=boxes[1,1]/(boxes[1,1]+boxes[1,2])
pred2=boxes[2,1]/(boxes[2,1]+boxes[2,2])
pred3=boxes[3,1]/(boxes[3,1]+boxes[3,2])
pred=pred1*0.4+pred2*0.1+pred3*0.5
pbox1=(0.4*boxes[1,1]/(boxes[1,1]+boxes[1,2]))/pred
pbox2=(0.1*boxes[2,1]/(boxes[2,1]+boxes[2,2]))/pred
pbox3=(0.5*boxes[3,1]/(boxes[3,1]+boxes[3,2]))/pred
return(c(pbox1,pbox2,pbox3))
}

p_red(boxes)
```

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

```r
p_box(boxes)
```

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

Exercise 5.

Formula used: $P(identical\ twins|twin\ boys) = \frac{P(twin\ boys|identical\ twins)*P(identical\ twins)}{P(twin\ boys)}$

$$P(twin\ boys) = 0.5 * P(identical\ twins) + 0.25 * P(fraternal\ twins)$$

```r
p_identical_twin<-function(fraternal_prob, identical_prob){
return((identical_prob*0.5)/(identical_prob*0.5+fraternal_prob*0.25))
}
p_identical_twin(1/150, 1/400)
```

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