Probability and Statistics for Data Analysis

Assignment 1

Vasileios Galanos
MSc in Data Science(PT)
p3351902

October 27, 2019

Exercise 1.

1.1

Let S be the set of all people who have the pathological symptom.
We have that $P(S) = 0.1$.
Let A be the set of all people who have the disease A.
We have that $P(S|A) = 0.95$.
We also have that $P(A) = \frac{15}{1000} = 0.015$

The probability of symptom and the disease $P(S \cap A)$ can be calculated as follows:

$$P(S \cap A) = P(S|A)P(A) = 0.95 \times 0.015 = 0.01425$$

1.2

If S is independent of A then

$$P(S|A) = P(S) \implies 0.95 = 0.1$$

But this equality is not true, so we have proven that the argument is invalid too. The presence of the symptom is not independent from disease A.

1.3

We have already calculated the probability of the symptom **and** the disease as $P(S \cap A) = 0.01425$
The probability of the symptom **or** the disease is

$$P(S \cup A) = P(S) + P(A) - P(S \cap A) = 0.1 + 0.015 - 0.01425 = 0.10075$$

1.4

The probability of the disease but not the symptom can be calculated as follows:

$$P(A \cap \bar{S}) = P(A) - P(A \cap S) = 0.015 - 0.01425 = 0.00075$$

where $\bar{S}$ is the complementary set of S.

1.5

We calculate the conditional probability using the *Baye's rule* as follows:

$$P(A|S) = P(S|A)\frac{P(A)}{P(S)} = 0.95 \times \frac{0.015}{0.1} = 0.1425$$

1.6

The probability of the symptom given that a person is healthy is:

$$P(S|\bar{A}) = \frac{P(S \cap \bar{A})}{P(\bar{A})} = \frac{P(S) - P(S \cap A)}{1 - P(A)} = \frac{0.1 - 0.01425}{1 - 0.015} = \frac{0.08575}{0.985} \approx 0.08705$$

Exercise 2.

2.1

There are $\binom{3}{1} = 3$ ways of choosing one ball from a set of 3 balls of the same colour.
We want all ten balls to be of different colour so there are $\binom{3}{1}^{10}$ ways of choosing those 10 balls.

Also we have a total of $\binom{30}{10}$ ways of picking 10 out of 30 balls.
Finally the probability of picking all ten balls of different colour is

$$P(\text{all balls are of different colour}) = \frac{\binom{3}{1}^{10}}{\binom{30}{10}} = 0.001965351$$

Alternatively, we calculate the probability that all balls are of different colour as follows:
The probability of choosing the first ball correctly is $\frac{30}{30} = 1$, since we don't care about the colour on our first choice.

On our second choice, since we eliminated the colour of the first ball from our options, so we have a $\frac{27}{29}$ probability of choosing a ball of different colour.
Next for the third ball, we choose it correctly with a $\frac{24}{28}$ probability.
So the probability of choosing all ten balls of different colour is:

$$P(\text{all balls are of different colour}) = \frac{30}{30} \times \frac{27}{29} \times \frac{24}{28} \times \frac{21}{27} \times \frac{18}{26} \times \frac{15}{25} \times \frac{12}{24} \times \frac{9}{23} \times \frac{6}{22} \times \frac{3}{21} = 0.001965351$$

2.2

Let A: *the outcome contains all three balls of* ***at least one*** *colour*
Let A_1: *the outcome contains all three balls of colour* ***one***
Let A_2: *the outcome contains all three balls of colour* ***two***
...
Let A_{10}: *the outcome contains all three balls of colour* ***ten***

In that case the probability $P(A)$ that the outcome contains all three balls of **at least one** colour is:

$$\begin{aligned} P(A) &\overset{hint}{=\!=} \sum_{i=1}^{10} \left(\bigcup A_i \right) \\ &= \sum_{i=1}^{10} P(A_i) - \sum_{1 \le i < j \le 10} P(A_i \cap A_j) + \sum_{1 \le i < j < k \le 10} P(A_i \cap A_j \cap A_k) \\ &+ \cdots + (-1)^9 P(A_1 \cap A_2 \cap \cdots \cap A_{10}) \end{aligned} \tag{1}$$

We know, that it is impossible to have 4 triplets of the same colour when we choose 10 balls.
So every division of 4 or more sets is the empty space.
The equation **1** becomes:

$$P(A) = \sum_{i=1}^{10} P(A_i) - \sum_{1 \le i < j \le 10} P(A_i \cap A_j) + \sum_{1 \le i < j < k \le 10} P(A_i \cap A_j \cap A_k) \tag{2}$$

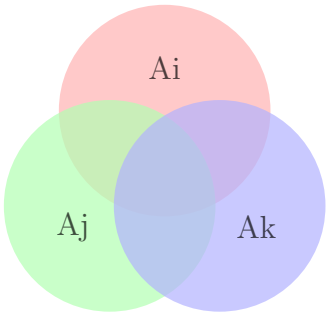


Figure 1: Logical relations between the sets A_i,A_j,A_k

Now we calculate the individual probabilities:

$$P(A_i) = \frac{\binom{3}{3}\binom{27}{7}}{\binom{30}{10}} = \frac{\binom{27}{7}}{\binom{30}{10}}$$

$$P(A_i \cap A_j) = \frac{\binom{3}{3}\binom{3}{3}\binom{24}{4}}{\binom{30}{10}} = \frac{\binom{24}{4}}{\binom{30}{10}}$$

$$P(A_i \cap A_j \cap A_k) = \frac{\binom{3}{3}\binom{3}{3}\binom{3}{3}\binom{21}{1}}{\binom{30}{10}} = \frac{\binom{21}{1}}{\binom{30}{10}}$$

Furthermore, we have $\binom{10}{2} = 45$ possible combinations of $A_i and A_j$ and $\binom{10}{3} = 120$ possible combinations of A_i, A_j, A_k .

Combining all of the above, we have:

$$P(A_i) = 10\frac{\binom{27}{7}}{\binom{30}{10}} - 45\frac{\binom{24}{4}}{\binom{30}{10}} + 120\frac{\binom{21}{1}}{\binom{30}{10}}$$

$$= \frac{10 \times 888030 - 45 \times 10626 + 120 \times 21}{30045015} \approx 0.2797352572$$

Exercise 3.

3.1

Using R we simulate the random experiment of Exercise 2 $m = 100000$ times.
For the event **2.1** we use the following code:

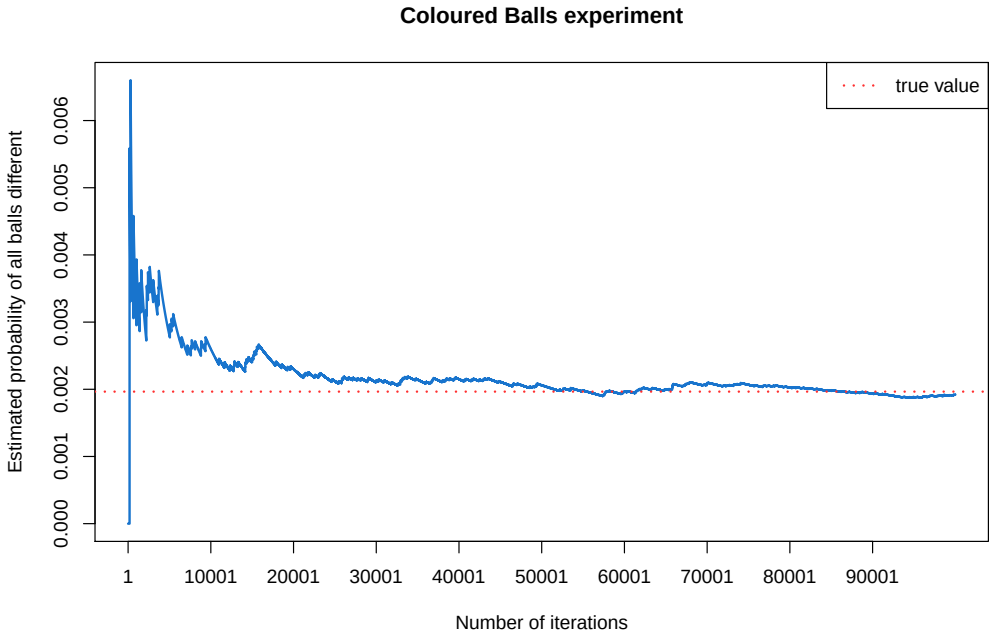
```
m<-100000
# Let c1 represent a ball of colour 1. In that way:
a <- c(rep('c1',3),rep('c2',3),rep('c3',3),rep('c4',3),rep('c5',3)
       ,rep('c6',3),rep('c7',3),rep('c8',3),rep('c9',3),rep('c10',3))

prob <- numeric(m)
sum<- 0
set.seed(111)

for (i in 1:m){
  s<- sample(a,10,replace = FALSE)
  if(length(unique(s)) == 10){
    sum <- sum + 1
  }
  prob[i] <- sum/i
}
sprintf("%.10f",sum/m)

#pdf(file = 'Ex3_1-figure.pdf', width = 9, height = 6)
plot(prob, xlab = 'Number of iterations',
     ylab = 'Estimated probability of all balls different',
     main = 'Coloured Balls experiment',
     col = 'dodgerblue3', lwd = 2, type = 'l'
     ,xaxt = "n")
axis(1, at=seq(1, m, by=10000))
abline(h = 0.001965351, col = 'firebrick1', lty = 3, lwd = 2)
legend('topright', 'true value', lty = 3, lwd = 2, col = 'firebrick1')
#
#dev.off()
```

This produces the following figure:



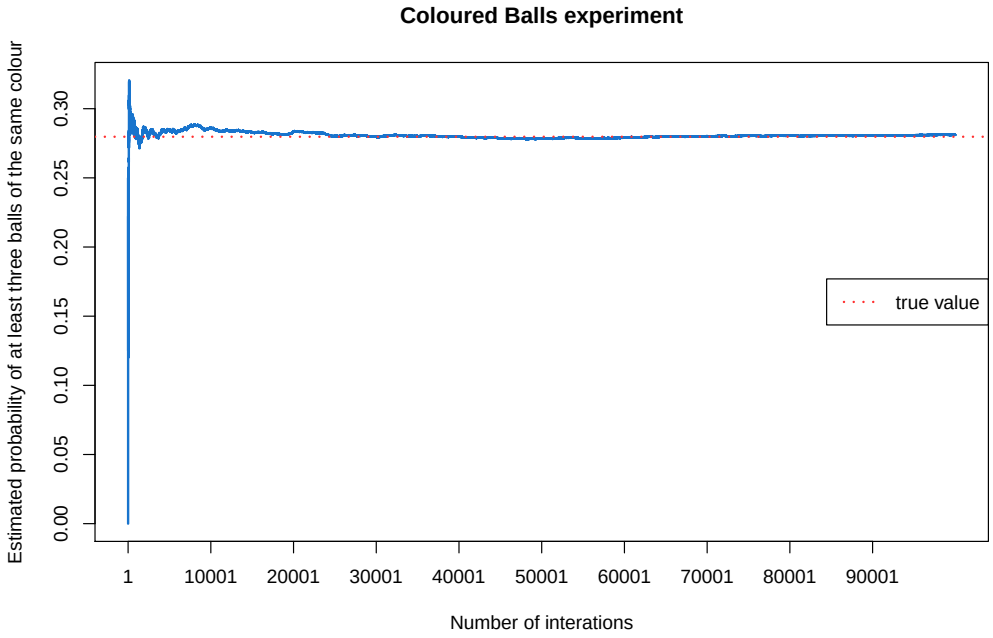
In a similar way for the event **2.2**:

```
sum<- 0
m<- 100000
prob <- numeric(m)

for (i in 1:m){
  s<- sample(a,10,replace = FALSE)
  if(sum(table(s) == 3) >= 1){
    sum <- sum + 1
  }
  prob[i] <- sum/i
}
sprintf("%.10f",sum/m)

#pdf(file = 'Ex3_23-figure.pdf', width = 9, height = 6)
plot(prob, xlab = 'Number of interations',
     ylab = 'Estimated probability of at
      least three balls of the same colour',
     main = 'Coloured Balls experiment',
     col = 'dodgerblue3', lwd = 2, type = 'l'
     ,xaxt = "n")
axis(1, at=seq(1, m, by=10000))
abline(h = 0.2797352572, col = 'firebrick1', lty = 3, lwd = 2)
legend('right', 'true value', lty = 3, lwd = 2, col = 'firebrick1')
#dev.off()
```

This produces the following figure:



Exercise 4.

4.1

The subspace of the random experiment for the first time the two players flip coins is

$$S_0 = \{H_A H_B, H_A T_B, T_A H_B, T_A T_B\}$$

where $H_A H_B$ is the event that both players A and B obtained *heads* .

Player A plays first so he wins in two out of four possible events: $\{H_A H_B, H_A T_B\}$. Player B wins only in the event $\{T_A H_B\}$.
When both players obtain *tails* $\{T_A T_B\}$ they repeat the process until one of them wins, so the second time the subspace of the experiment given that they both obtained tails in the first coin toss is

$$S_1 = \{T H_A T H_B, T H_A T T_B, T T_A T H_B, T T_A T T_B\}$$

and the N_{th} time

$$S_N = \{T^N H_A T^N H_B, T^N H_A T^N T_B, T^N T_A T^N H_B, T^N T_A T^N T_B\}$$

Finally the **sample space** of the random experiment is the union of all subspaces for every round they flip a coin.Thus,

$$S = \{S_0 \cup S_1 \cup S_2 \cup \cdots \cup S_N\}$$

$$= \sum_{i=0}^{N} \left(\bigcup S^i \right) = \sum_{i=0}^{N} \left(\bigcup \{T^i H_A T^i H_B, T^i H_A T^i T_B, T^i T_A T^i H_B, T^i T_A T^i T_B\} \right)$$

4.2

They first time they flip coins, Player A wins with a probability p.
If he obtains *tails* with a probability of $1-p$ then his only chance of winning is for Player B to obtain *tails* too, so they both repeat the process in the next round until one of them wins the game e.t.c

Specifically,

$$\begin{aligned} P(\textit{Player A wins}) =& p + (1-p)(1-p)p + (1-p)(1-p)(1-p)(1-p)p \dots \\ &= \sum_{k=0}^{\infty} p\left[(1-p)^2\right]^k \end{aligned}$$

This is a known geometric series $\sum_{k=0}^{\infty} ax^k$ that converges to $\frac{a}{1-x}$ if $|x| < 1$. In our case $0 < p < 1$, so the probability of A wins converges to

$$P(\textit{Player A wins}) \longrightarrow \frac{p}{1-(1-p)^2} = \frac{p}{p(2-p)} = \frac{1}{(2-p)}$$

Interestingly, if our coin was fair ($p = \frac{1}{2}$) then Player A would have a $\frac{2}{3}$ probability of winning

4.3

To show that $\forall p \in (0,1)$, $P(\textit{Player A wins}) > \frac{1}{2}$ we simply need to prove that

$$\frac{1}{(2-p)} > \frac{1}{2} \qquad \forall p \in (0,1)$$

We know that $2 - p > 0 \quad \forall p \in (0,1)$, so

$$2 > 2 - p \Rightarrow -p < 0 \Rightarrow p > 0, \qquad \textit{which is true}$$

Exercise 5.

The Poisson distribution with parameter λ is defined by the following equation

$$P(X = x|\lambda) = \frac{e^{-\lambda}\lambda^x}{x!} \quad , x \in 0, 1, 2, 3, ...$$

In our case, the average number of incoming calls to a call center in one minute is 10. ($\lambda = 10$). So we have:

$$P(X = x|10) = \frac{e^{-10}10^x}{x!}$$

5.1

Now we can calculate the probability as follows:

$$P(\text{No calls arrive within one minute}) = P(X = 0) = \frac{e^{-10}10^0}{0!} = e^{-10} \approx 0.0000453999$$

5.2

In the same way,

$$P(\text{10 calls arrive within one minute}) = P(X = 10) = \frac{e^{-10}10^{10}}{10!} \approx 0.1251100357$$

5.3

Now to calculate the probability of at least 10 calls within one minute, we must calculate the probability density function as follows:

$$P(\text{at least 10 calls arrive within one minute}) = P(X \geq 10) = P(X > 9)$$
$$= 1 - P(X \leq 9) = 1 - \sum_{x=0}^{9} \frac{e^{-10}10^x}{x!} \approx 0.5420702855$$

5.4

$$P(\text{10 calls arrive within two minutes}) =$$

$$P(X = 10 \mid \lambda = 20) = \frac{e^{-20}20^{10}}{10!} \approx 0.0058163065$$

5.5

Now, we want to calculate the probability of at least 10 calls arrive within one minute, conditional on the event that at least one call arrives (within one minute)

$$P(X \geq 10 | X \geq 1) = \frac{P(X \geq 10 \cap X \geq 1)}{P(X \geq 1)} = \frac{P(X \geq 10)}{P(X \geq 1)}$$

$$= \frac{P(X > 9)}{P(X > 0)} = \frac{1 - P(X \leq 9)}{1 - P(X \leq 0)} = \frac{1 - \sum_{x=0}^{9} \frac{e^{-10}10^{x}}{x!}}{1 - \frac{e^{-10}10^{0}}{0!}} \approx 0.5420948966$$

Exercise 6.

We know that our random variable (X)) has the probability density function:

$$f_X(x) = xe^{-x}I_{(0,\infty)}(x)$$

6.1

The probability that a call lasts at most one minute is

$$\begin{aligned} P(X \leq 1) &= \int_0^1 xe^{-x}dx = \int_0^1 x(-e^{-x})'dx \overset{IBP}{=\!=} (-xe^{-x})|_0^1 - \int_0^1 -e^{-x}x'dx \\ &= -e^{-1} - \int_0^1 (e^{-x})'dx = -e^{-1} - (e^{-x})|_0^1 = 1 - 2e^{-1} \approx 0.2642411 \end{aligned}$$

6.2

The probability that a call lasts at least two minutes is

$$\begin{aligned} P(X \geq 2) &= 1 - P(X \leq 2) = 1 - \int_0^2 xe^{-x}dx \\ &\overset{IBP}{=\!=} 1 - [(-xe^{-x})|_0^2 - \int_0^2 -e^{-x}x'dx] = 1 - [-2e^{-2} - (e^{-x})|_0^2] \\ &= 1 - (1 - 3e^{-2}) = 3e^{-2} \approx 0.4060058 \end{aligned}$$

6.3

We calculate the mean of call duration as follows

$$E(X) = \int_0^\infty xf_x(x)dx = \int_0^\infty x^2e^{-x}dx$$

This is the gamma function $\Gamma(3)$. *Note:* $(\Gamma(\alpha) = \int_0^\infty x^{\alpha-1}e^{-x}dx)$

$$\Rightarrow E(X) = \Gamma(3) = 2! = 2$$

To calculate the variance we must first calculate the metric $E(g(x)), \quad g(x) = X^2$.

$$E(g(x)) = \int_0^\infty g(x) f_x(x) dx = \int_0^\infty x^3 e^{-x} dx = \Gamma(4) = 6$$

Now we can calculate the variance as follows:

$$Var(x) = E(X^2) - [E(X)]^2 = 6 - 4 = 2$$

6.4

We know that the cost of a call with duration x is equal to

$$c(x) = \begin{cases} 2 & , 0 < x \leq 3 \\ 2 + 6(x-3) & , x > 3 \end{cases}$$

The average duration of a call is **2** minutes (< 3), thus the cost of a call with an average duration is **2** euros.

6.5

We calculate the average cost of a call as follows:

$$E(c(x)) = \int_0^\infty c(x) f_x(x) dx = \int_0^3 2xe^{-x} dx + \int_3^\infty [2 + 6(x-3)] xe^{-x} dx \quad (1)$$

We will first calculate $\int_0^3 xe^{-x} dx$ as:

$$\begin{aligned} \int_0^3 xe^{-x} dx &= (-xe^{-x})|_0^3 - \int_0^3 (-e^{-x} dx) = (-3e^{-3}) + \int_0^3 (-e^{-x})' dx \\ &= -3e^{-3} + (-e^{-x})|_0^3 = -3e^{-3} + (-e^{-3} + 1) \\ &= 1 - 4e^{-3} \end{aligned} \quad (2)$$

Secondly we calculate $\int_3^\infty [2+6(x-3)]xe^{-x}dx$ as follows:

$$\begin{aligned}\int_3^\infty [2+6(x-3)]xe^{-x}dx &= 2\int_3^\infty xe^{-x}dx + 6\int_3^\infty (x-3)xe^{-x}dx \\ &= 2\int_3^\infty xe^{-x}dx + 6\int_3^\infty x^2e^{-x}dx - 18\int_3^\infty xe^{-x}dx \\ &= 6\int_3^\infty x^2e^{-x}dx - 16\int_3^\infty xe^{-x}dx \end{aligned} \tag{3}$$

- $$\begin{aligned}\int_3^\infty xe^{-x}dx &= (-xe^{-x})|_3^\infty - \int_3^\infty (-e^{-x})dx = \lim_{b\to\infty}\frac{b}{e^b} + 3e^{-3} + \int_3^\infty e^{-x}dx \\ &= 3e^{-3} + \int_3^\infty (-e^{-x})'dx = 3e^{-3} + (-e^{-3x}) = 3e^{-3} - \lim_{b\to\infty}\frac{1}{e^b} + e^{-3} \\ &= 4e^{-3} \end{aligned} \tag{4}$$

- $$\begin{aligned}\int_3^\infty x^2e^{-x}dx &= (-x^2e^{-x})|_3^\infty - \int_3^\infty 2x(-e^{-x})dx \\ &= -\lim_{b\to\infty}\frac{b^2}{e^b} + 3^2e^{-3} + 2\int_3^\infty xe^{-x}dx = 9e^{-3} + 2\times 4e^{-3} = 17e^{-3} \end{aligned} \tag{5}$$

Thus equation **(3)** becomes:

$$\overset{4,5}{\Longrightarrow} 6\times 17e^{-3} - 16\times 4e^{-3} = 38e^{-3} \tag{6}$$

In conclusion equation **(1)** gives us the average cost of a call:

$$\begin{aligned}\overset{2,6}{\Longrightarrow} E(c(x)) &= \int_0^3 2xe^{-x}dx + \int_3^\infty [2+6(x-3)]xe^{-x}dx = 2\times(1-4e^{-3}) + 38e^{-3} \\ &= 2 - 30e^{-3} \approx 3.4936120510 \end{aligned}$$

6.6

The (upper) α quartile of an absolutely continuous distribution is a number c_α such that:

$$\begin{aligned}
P(X > c_\alpha) &= \alpha \Rightarrow \\
1 - P(X \leq c_\alpha) &= \alpha \Rightarrow \\
P(X \leq c_\alpha) &= 1 - \alpha \Rightarrow \\
\int_0^{c_\alpha} xe^{-x}dx &= 1 - \alpha \Rightarrow \\
(-xe^{-x})|_0^{c_\alpha} - \int_0^{c_\alpha} (-e^{-x})dx &= 1 - \alpha \Rightarrow \\
-c_\alpha e^{-c_\alpha} + (-e^{-x})|_0^{c_\alpha} &= 1 - \alpha \Rightarrow \\
1 - (c_\alpha + 1)e^{-c_\alpha} &= 1 - \alpha \Rightarrow \\
(c_\alpha + 1)e^{-c_\alpha} &= \alpha
\end{aligned}$$

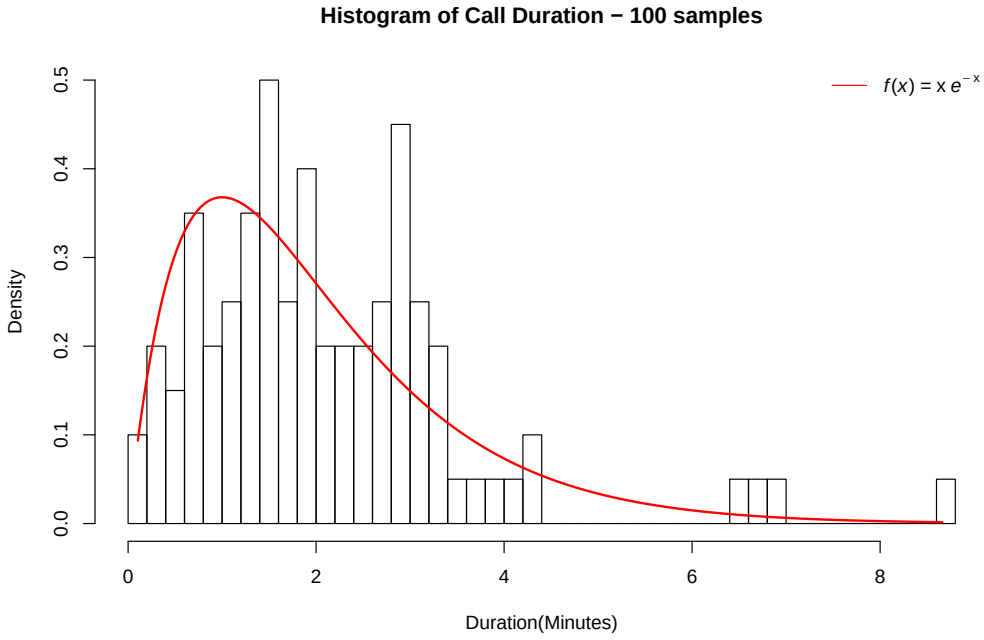
Finally, the 1st quartile corresponds to $\alpha = 0.75$, thus $c_{0.75} \approx 0.961279$. The 2nd quartile (median) corresponds to $\alpha = 0.5$, thus $c_{0.5} \approx 1.67835$. And the 3rd quartile corresponds to $\alpha = 0.25$, thus $c_{0.25} \approx 2.69263$.

Exercise 7.

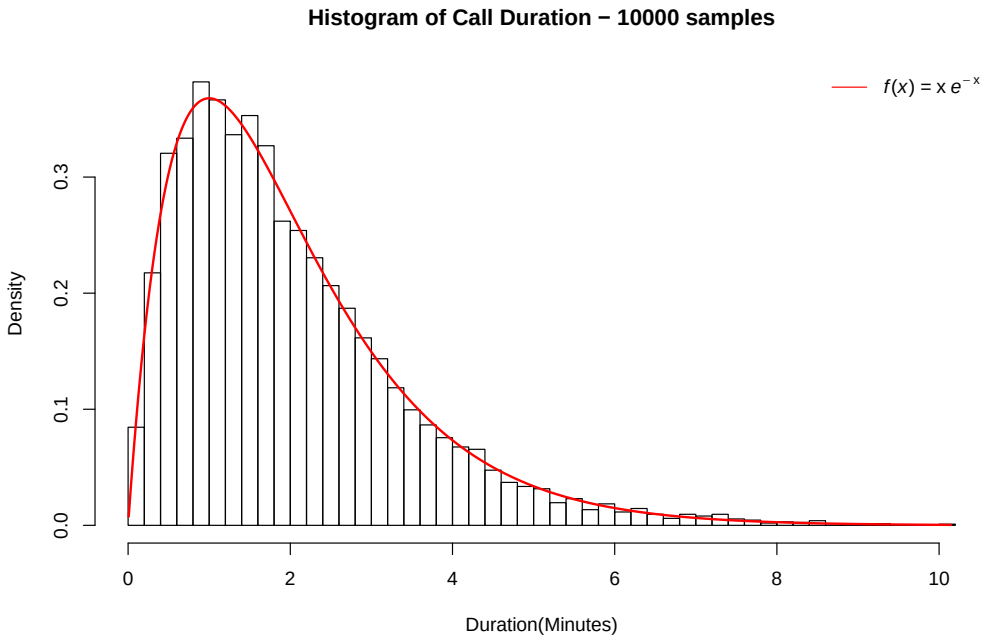
7.1

We process two datasets of 100 and 10000 randomly sampled calls from the call center of Exercise 6. For the first dataset we plot the histogram using the following R code

```
x_100 <- read.table('..\\datasets\\call_duration_100.txt')
x_100 <- unlist(x_100, use.names = FALSE)
#pdf(file = 'Ex7_1-figure_100.pdf', width = 9, height = 6)
hist(x_100 , 40, freq = FALSE, xlab = "Duration(Minutes)"
     ,main='Histogram of Call Duration - 100 samples')
xSeq <- seq(min(x_100),max(x_100), length = 1000)
points(xSeq, xSeq *exp(-xSeq ), type = "l", col = "red", lwd = 2)
legend("topright", lty = 1, col = 2, bty = "n",
 legend=bquote(italic(f)*"("*italic(x)*") ="~x~italic(e)^{- ~x}))
```



Similarly, for the second dataset we get the corresponding histogram:



7.2

(a) We estimate the probability that a call lasts at most one minute for the two datasets in R as follows:

```
> n_100 <- length(x_100)
> n_10000 <- length(x_10000)
> sprintf("%.4f", sum((x_100 <= 1)*1)/n_100 )
[1] "0.2000"
> sprintf("%.4f", sum((x_10000 <= 1)*1)/n_10000 )
[1] "0.2676"
```

(b) We estimate the probability that a call lasts at least two minutes for the two datasets in R as follows:

```
> sprintf("%.4f",sum((x_100 >= 2)*1)/n_100)
[1] "0.4500"
> sprintf("%.4f",sum((x_10000 >= 2)*1)/n_10000)
[1] "0.4034"
```

(c) We estimate the mean and variance of call duration as follows:

```
> sprintf("%.4f",mean(x_100))
[1] "2.1412"
> sprintf("%.4f",mean(x_10000))
[1] "1.9923"
> sprintf("%.4f",var(x_100))
[1] "2.0987"
> sprintf("%.4f",var(x_10000))
[1] "2.0095"
```

(d) We calculate how much the call costs for all values of the sample:

```
cost_100 <- numeric(n_100)

for (i in 1:n_100)
{
  if (x_100[i] <=3){ cost_100[i] <- 2 }
  else{ cost_100[i] <- 2 + 6*(x_100[i] - 3) }
}
```

Similarly we calculate for the other dataset and then we get the following results for the cost of a call with an average duration:

```
> sprintf("%.4f",cost_100[mean(x_10000)])
[1] "2.0000"
> sprintf("%.4f",cost_10000[mean(x_10000)])
[1] "2.0000"
```

(e) The average cost of a call for the two datasets is:

```
> sprintf("%.4f",mean(cost_100))
[1] "3.4515"
> sprintf("%.4f",mean(cost_10000))
[1] "3.4909"
```

(f) We estimate the 1st and 3rd quartile of call duration for the two datasets as follows:

```
> quantile(x_100, prob = 1:3/4)
     25%      50%      75%
1.215675 1.877632 2.834025
> quantile(x_10000, prob = 1:3/4)
     25%      50%      75%
0.948882 1.660063 2.676362
```

7.3

Comparing the theoretic and sample evaluated statistics, it is obvious that using the big sample(10000 samples), we get a more accurate estimate for all the metrics compared to the theoretical values.
For the small dataset, while it is not entirely inaccurate, it certainly diverges from the truth. That can be seen clearly in the following table:

Theoretical vs Sample Statistics			
Statistic	Theoretical	Sample - 100	Sample - 10000
P(a)	0.2642	0.2	0.2676
P(b)	0.4060	0.45	0.4034
mean	2	2.1412	1.9923
variance	2	2.0987	2.0095
call with average	2	2	2
average cost	3.4936	3.4515	3.4909
1st quartile	0.9612	1.2156	0.9488
3rd quartile	2.6926	2.8340	2.6763