# Distribution Fitting and Point Estimation

Benoit Mialet

02/12/2021

Working directory setting:

```
setwd("D:/Formations/DSTI/2021 07 - Advanced stats and ML/assignment")
```

---

Observations of data1.txt dataset are associated to a point process. The time arrival between two points should be an exponential random variable with parameter $\lambda$. My objective is to find if it is the case with different methods.

To answer the question I will proceed as following:

1. Import data file into a dataframe, and look at the data.
2. Create a new variable *time*, based on the time arrival between two points. The objective will be to determine if the random variable *time* follows or not an exponential distribution.
3. For the random variable *time*, compute an estimator $\hat{\lambda}$ for the parameter $\lambda$ of a supposed exponential distribution. 2 methods will be used for this purpose:
   - Methods of Moments of orders $k = 1$ and $k = 2$
   - Method of Maximum Likelihood
4. Determine if the distribution of *time* fits an exponential distribution, with parameter $\hat{\lambda}$. 2 methods will be used for this purpose:
   - Kolmogorov-Smirnov test
   - Visualization of PDF and QQ-Plots, and comparison of the random variable *time* distribution and a theoretical exponential distribution with parameter $\hat{\lambda}$.
5. Conclude

## I) Data handling and first look to the data

I open *data1.txt* file and store it in a dataframe:

```
data1 <- read.table("data1.txt", header=FALSE, col.names = "point",
                    sep=";",dec=".", fileEncoding="latin1", check.names=FALSE)
```

I observe some basic information about the data: dimension of the dataframe, type (class) of variables, number of NA values, and first observations

```
str(data1)
```

```
## 'data.frame':    4766 obs. of  1 variable:
##  $ point: num  0.00601 0.14179 0.25937 0.82704 1.62859 ...
```

```
paste("Number of NA values : ", sum(is.na(data1)))
```

```
## [1] "Number of NA values :  0"
```

```
head(data1,10)
```

```
##          point
## 1  0.006005354
## 2  0.141786600
## 3  0.259368800
## 4  0.827037100
## 5  1.628588000
## 6  1.673124000
## 7  1.981014000
## 8  2.015298000
## 9  2.831079000
## 10 4.052456000
```

Here, variable *point* is numerical with no missing value.

## II) Computation of the *time* variable, observation of its metrics and its distribution

To study the the time arrival between points, I create a new dataframe *df_time* from data1 observations, based on the time difference between two successive points. I name this new variable "*time*" :

```
time <- {}
for (i in c(1:(length(data1$point)-1))) {
  time[i] <- data1$point[i+1]-data1$point[i]
}
df_time <- data.frame(time)
head(df_time,10)
```

```
##         time
## 1  0.1357812
## 2  0.1175822
## 3  0.5676683
## 4  0.8015509
## 5  0.0445360
## 6  0.3078900
## 7  0.0342840
## 8  0.8157810
## 9  1.2213770
## 10 0.1620000
```

I observe dimensions and some basic metrics of theses distances (*time*)

```
dim(df_time)
```

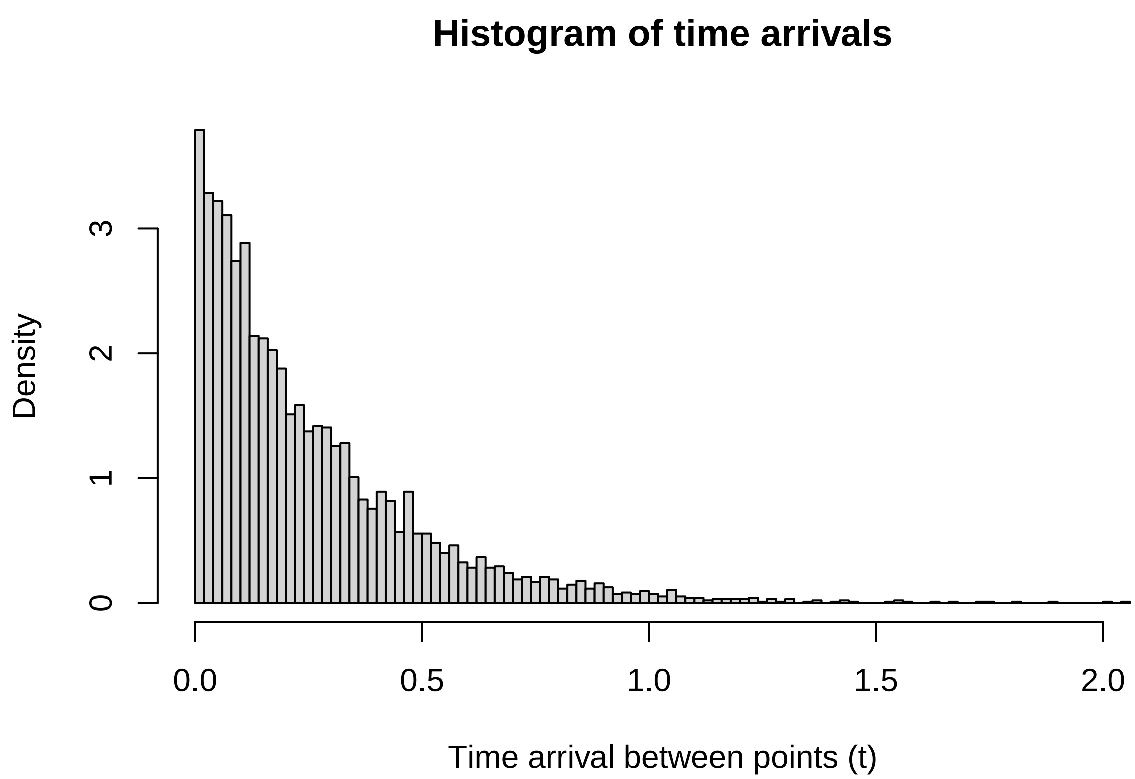
```
## [1] 4765    1
```

```
summary(df_time)
```

```
##       time
##  Min.   :0.0000
##  1st Qu.:0.0746
##  Median :0.1770
##  Mean   :0.2518
##  3rd Qu.:0.3490
##  Max.   :2.0425
```

I plot a histogram of distances to observe the shape of the probability density function of the variable *time* to study:

```
hist(
  df_time$time,
  breaks = 100,
  freq=FALSE,
  main ="Histogram of time arrivals",
  xlab="Time arrival between points (t)"
  )
```



The distribution of *time* looks like an exponential distribution. In the following, I will first estimate a parameter $\lambda$ before testing the goodness of fit of this distribution with an exponential one.

## II) Point estimation: parameter $\lambda$ of an exponential distribution

Let assume that all observations of the dataframe *df_time* $X_1, .., X_n$ are **independent and identically distributed** random variables. Expectation of a random variable $E[X_i]$ with an exponential distribution is by definition:

$$E[X_i] = E[X_1] = \frac{1}{\lambda}$$

With $\lambda \in \mathbb{R}$.

In order to provide some estimators of $\lambda$, I will use several methods in the following.

### II.1) Method of moment with k = 1:

I have:

$$E[X_1^1] = \frac{1}{\lambda}$$

By application of the Law of Large Numbers, an estimator $\hat{\lambda_n}$ for $\lambda$ will be a solution of:

$$g(\hat{\lambda_n}) = \frac{1}{\hat{\lambda_n}} = \frac{1}{n}\sum_{i=1}^{n}(X_i) = \overline{X_n}$$

I thus obtain:

$$\hat{\lambda_n} = \frac{1}{\overline{X_n}}$$

I can now compute the value of $\hat{\lambda_n}$:

```
lambda_hat_n <- 1/mean(df_time$time)
lambda_hat_n
```

```
## [1] 3.971419
```

### II.2) Method of moment with k = 2:

By definition, assuming that $X_1, .., X_n$ are independent and identically distributed random variables, I have by definition:

$$V[X_1] = E[X_1^2] - E[X_1]^2 E[X_1^2] = V[X_1] + E[X_1]^2$$

With $V[X_1] = \frac{1}{\lambda^2}$ and $E[X_1]^2 = \frac{1}{\lambda^2}$ (with $\lambda \in \mathbb{R}$).

I thus obtain:

$$E[X_1]^2 = \frac{2}{\lambda^2}$$

By application of the Law of Large Numbers, an estimator $\hat{\lambda_{n,2}}$ for $\lambda$ will be a solution of:

$$g(\hat{\lambda_{n,2}}) = \frac{2}{\hat{\lambda^2_{n,2}}} = \frac{1}{n}\sum_{i=1}^{n}(X_i^2) = \overline{X_n}$$

I obtain:

$$\hat{\lambda_{n,2}} = \sqrt{\frac{2n}{\sum_{i=1}^{n}(X_i^2)}}$$

I thus compute $\sum_{i=1}^{n}(X_i^2)$ then $\hat{\lambda_{n,2}}$:

```
sum_Xi_squared <- 0
for (i in c(1:length(df_time$time))) {
  sum_Xi_squared <- sum_Xi_squared + df_time$time[i]**2
}

lambda_hat_n_2 <- sqrt(2*length(df_time$time)/sum_Xi_squared)
lambda_hat_n_2
```

```
## [1] 4.015886
```

### II.3) Method of Maximum Likelihood:

By definition, probability density function of a continuous random variable $X_i$ which has an exponential distribution is given by:

$$f_{Xi}(x_i) = \lambda.\mathrm{e}^{-\lambda x_i}\mathbb{1}_{[0;+\infty[}(x_i)$$

With $\lambda \in \mathbb{R}^{++}$.

An estimator $\hat{\lambda}$ of $\lambda$ with the method of maximum likelihood would be a solution of the following maximization problem:

$$\hat{\lambda_n} = \arg\max_{\lambda_n} L_{(\lambda_n;x_1,\ldots,x_n)}$$

I assume that all the terms $x_i$ are independent and identically distributed, I can thus write the likelihood function as:

$$\begin{aligned}
L_{(\lambda;x_1,\ldots,x_n)} &= \prod_{i=1}^{n} f_{Xi}(x_i;\lambda) \\
&= \prod_{i=1}^{n}\left(\lambda\mathrm{e}{-\lambda x_i}\mathbb{1}_{[0;+\infty[}(x_i)\right) \\
&= \lambda^n\mathrm{e}{-\lambda}\sum_{i=1}^{n} x_i\mathbb{1}_{[0;+\infty[}(min(x_i))
\end{aligned}$$

As in this case I am in the frame of an exponential distribution, I have $min(x_i) \geq 0$. I can thus consider only the left part of the equation, without the indicator function:

$$L_{(\lambda;x_1,\ldots,x_n)} = \lambda^n\mathrm{e}{-\lambda}\sum_{i=1}^{n} x_i$$

I can consider a log transformation of the exponential function. I have:

$$\begin{aligned} l_{(\lambda;x_1,\ldots,x_n)} &= \ln\left(L_{(\lambda;x_1,\ldots,x_n)}\right) \\ &= \ln\left(\lambda^n \mathrm{e}{-}\lambda\sum_{i=1}^{n} x_i\right) \\ &= n\ln(\lambda) - \lambda\sum_{i=1}^{n} x_i \end{aligned}$$

At a critical point, the derivative of the function will be equal to 0:

$$\frac{d}{d\lambda_n} l_{(\lambda_n;x_1,\ldots,x_n)} = 0$$

I thus have:

$$\begin{aligned} \frac{d}{d\lambda_n}\left(n\ln(\lambda_n) - \lambda_n\sum_{i=1}^{n} x_i\right) &= 0 \\ \frac{n}{\lambda_n} - \sum_{i=1}^{n} x_i &= 0 \end{aligned}$$

To verify that this solution (critical point) is associated to a maximum, I have to verify that its derivative is $< 0$:

$$\frac{d}{d\lambda_n}\left(\frac{n}{\lambda_n} - \sum_{i=1}^{n} x_i\right) < 0$$

I obtain:

$$\frac{-n}{\lambda_n^2} < 0$$

As $n \geq 0$ and $\lambda_n^2 \geq 0$, this equation is verified, and I am sure that the previous critical point is associated to a maximum. Thus, an estimator $\hat{\lambda_n}$ of $\lambda$ obtained by the maximum likelihood method is:

$$\hat{\lambda_n} = \frac{n}{\sum_{i=1}^{n} x_i}$$

The expression can also be written:

$$\hat{\lambda_n} = \frac{1}{\overline{X_n}}$$

This estimator is the same as the one obtained by method of moments with $k = 1$. In the following, I will thus only consider the latter estimator, and will thus compute values for both $\hat{\lambda_n}$ and $\hat{\lambda_{n,2}}$

## IV) Distribution fitting: comparison of time arrivals distribution with an exponential distribution

### IV.1) Goodness of fit: Kolmogorv-Smirnov test

I compute a Kolmogorv-Smirnov test to compare both distributions, with a Null hypothesis $\mathcal{H}_0$ stating that the observed distribution is not different from an exponential distribution with parameter lambda_hat_n. I will arbitrary set a risk level $\alpha = 0.05$. I will perform one test for each estimator of $\lambda$ I previously computed:

```r
ks.test(df_time$time, "pexp", lambda_hat_n)
```

```
## Warning in ks.test(df_time$time, "pexp", lambda_hat_n): ties should not be
## present for the Kolmogorov-Smirnov test
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  df_time$time
## D = 0.011477, p-value = 0.5568
## alternative hypothesis: two-sided
```

```r
ks.test(df_time$time, "pexp", lambda_hat_n_2)
```

```
## Warning in ks.test(df_time$time, "pexp", lambda_hat_n_2): ties should not be
## present for the Kolmogorov-Smirnov test
```

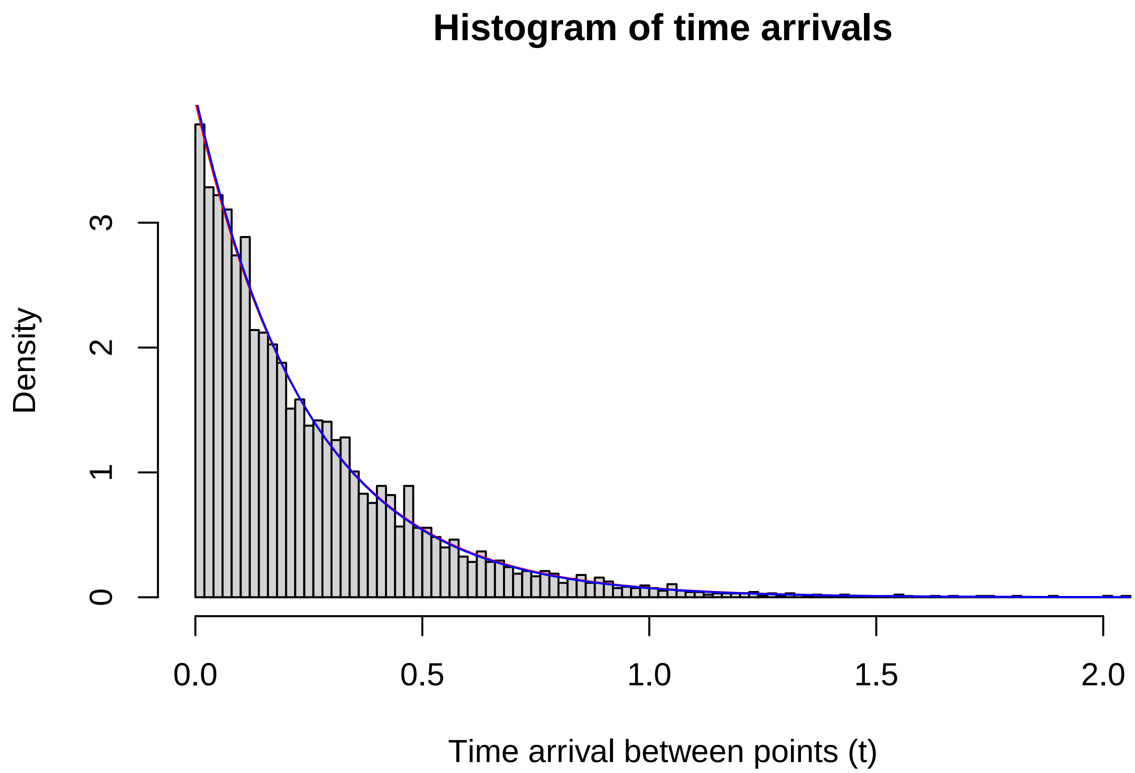
```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  df_time$time
## D = 0.015573, p-value = 0.1981
## alternative hypothesis: two-sided
```

With p-values 0.1981 and $0.5568 > \alpha$, I can accept the null hypothesis and assume that time arrivals follows an exponential distribution, for all $\lambda$ estimators previously computed.

### IV.2) Histogram and quantiles visualization

In order to visualize and confirm this result, I plot a graph comparing histogram of time arrivals with the probability density functions of exponential distributions with parameters $\hat{\lambda_n}$ (in red) and $\hat{\lambda_{n,2}}$ (in blue):

```r
hist(
  df_time$time, breaks = 100,
  freq=FALSE,
  main ="Histogram of time arrivals",
  xlab="Time arrival between points (t)"
  )
curve(dexp(x, rate = lambda_hat_n), from = 0, col = "red", add = TRUE)
curve(dexp(x, rate = lambda_hat_n_2), from = 0, col = "blue", add = TRUE)
```



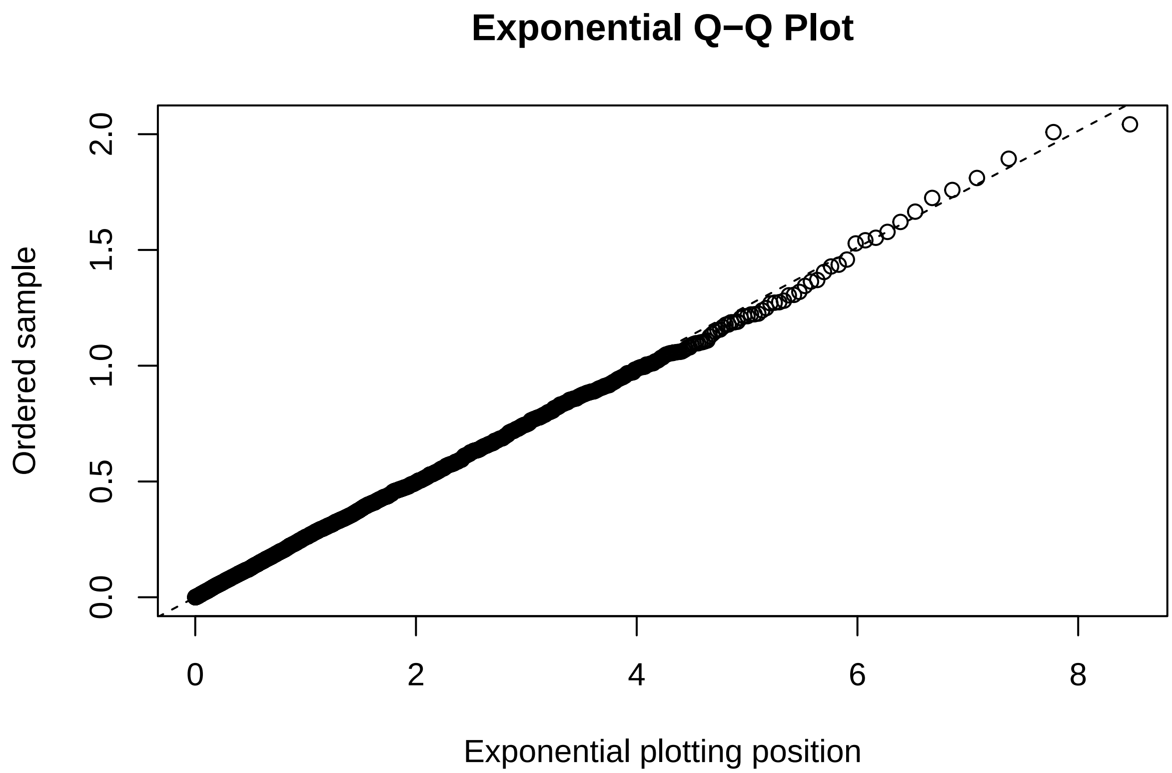

The curves seems indeed to fit the distribution in both cases.

I now visualize the Q-Q Plot to compare Quantile values of the *time* variable with quantile values of an exponential distribution:

```
library(SMPracticals)
```

```
## Warning: package 'SMPracticals' was built under R version 4.0.5

## Loading required package: ellipse

## Warning: package 'ellipse' was built under R version 4.0.5

##
## Attaching package: 'ellipse'

## The following object is masked from 'package:graphics':
##
##     pairs
```

```
# QQ-PLOT
qqexp(df_time$time, main = "Exponential Q-Q Plot",
       plot.it = TRUE, line = TRUE)
```




Again, the distribution of variable *time* seems to really fit an exponential distribution.

## V) Conclusion

In conclusion, considering result of the Kolmogorv-Smirnov test and by visualization of its histogram and its Q-Q Plot, I can assume that the time arrival between two points is an exponential random variable with parameter $\lambda$. I computed two estimators $\hat{\lambda_n}$ and $\hat{\lambda_{n,2}}$.

I can thus assume that observations from data1.txt dataset are associated to a Poisson process with parameter $\lambda$.