SI-M1-TD3

Guillaume Metzler

10/25/2021

L'objectif de ce TD est de travailler sur la notion de normalité d'un échantillon, qui est une hypothèse importante sur les données pour l'utilisation de certains tests paramétriques lorsque les échantillons sont de tailels inférieures à 30.

Pour tester la normalité des données, on peut avoir recours au test de Shapiro(-Wilk) qui se révèle particulièrement efficace lorsque notre échantillon a une taille comprise entre 15 et 30 environ

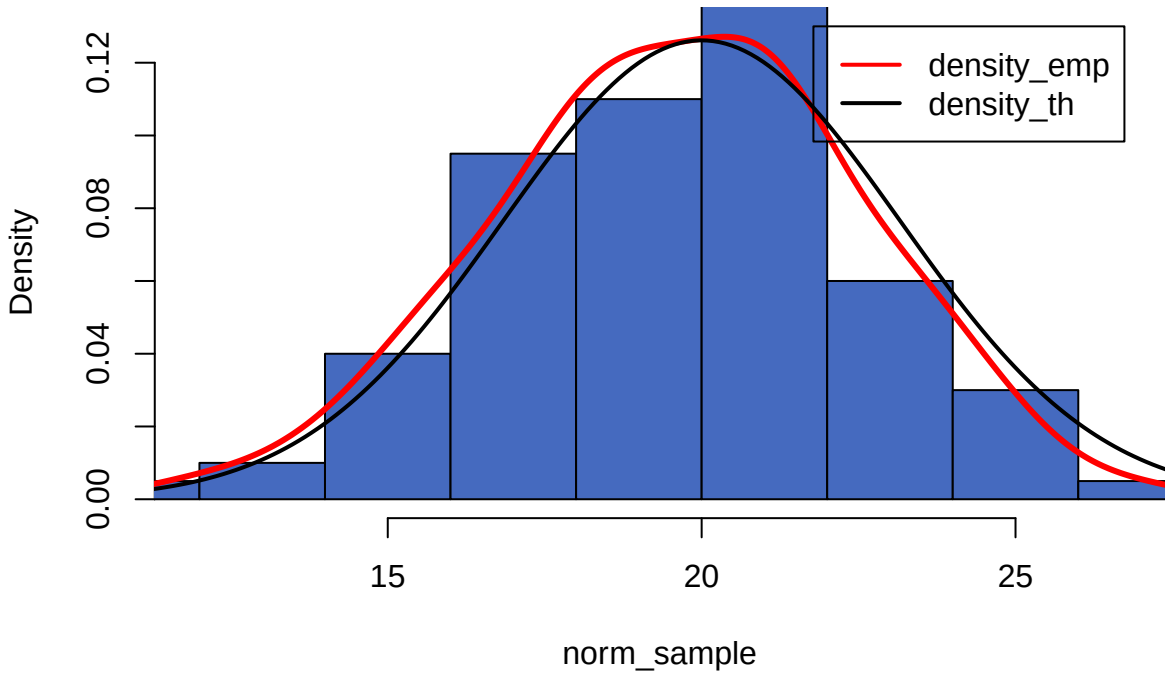
Simulation des données

On commence par simuler un échantillon gaussien à l'aide de la commande "rnorm" pour un échantillon de taille $n = 100$, d'espérance $\mu = 20$ et de variance $\sigma = 10$

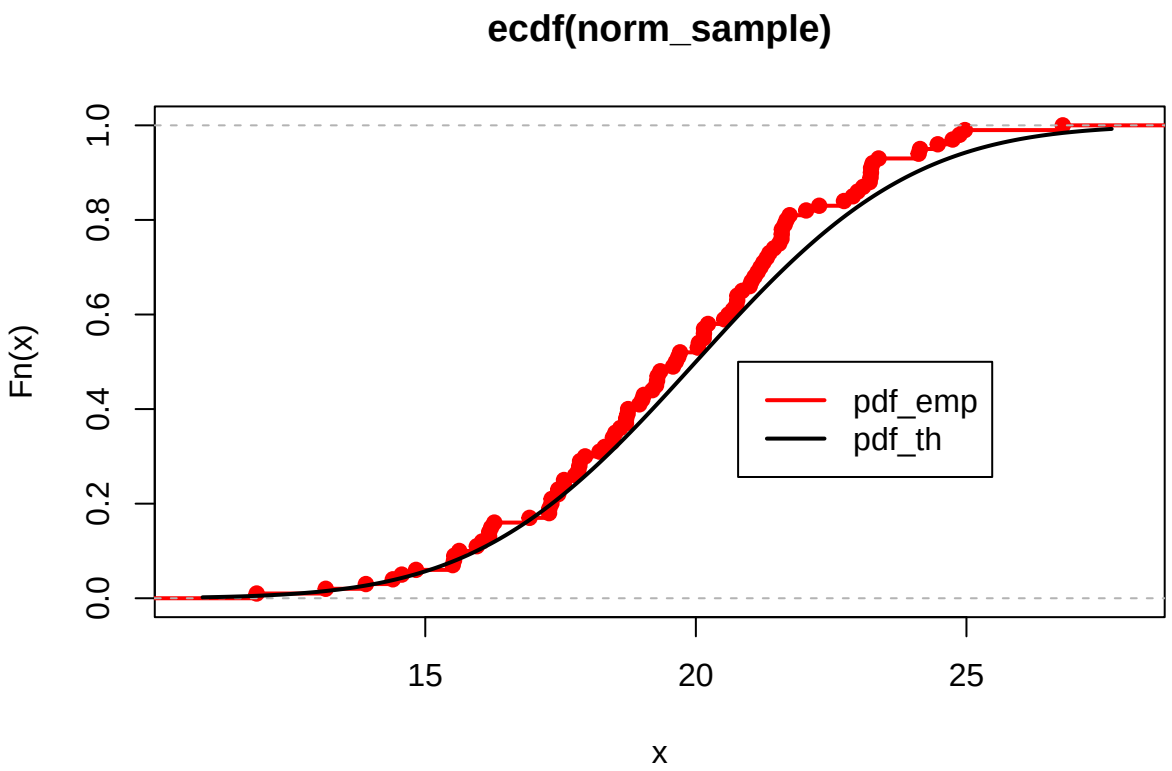
```
# Paramètres de l'échantillon

n = 100
mu = 20
ect = sqrt(10)

norm_sample <- rnorm(n, mu, ect)
```

On représente maintenant les données sous la forme d'un histogramme ainsi que les fonction de répartition empirique à l'aide de la fonction (ecdf)

```
# Histogramme
hist(norm_sample, breaks=10, prob = TRUE,
     col = "#456ABF", border= "black", ylim = c(0,0.13), xlim =c(min(norm_sample), max(norm_sample)))
lines(density(norm_sample), col = "red", lwd = 3)
# Vous pouvez jouer sur le paramètres "breaks" pour faire varier le nombre
# de barres dans votre histogramme
# On peut également comparer cela à la densité théorique de la loi normale

x=seq(min(norm_sample)-1,max(norm_sample)+1,0.1)
lines(x,dnorm(x,mu,ect), col= "black", lwd = 2)
legend(max(norm_sample)-5, 0.13, legend = c("density_emp", "density_th"),
       col = c("red","black"), lwd = 2)
```




```
# Fonction de repartition empirique

plot(ecdf(norm_sample), col = "red", lwd =2)
x=seq(min(norm_sample)-1,max(norm_sample)+1,0.1)
lines(x,pnorm(x,mu,ect), col= "black", lwd = 2)
legend(max(norm_sample)-6,0.5, legend = c("pdf_emp", "pdf_th"), col = c("red","black"), lwd = 2)
```

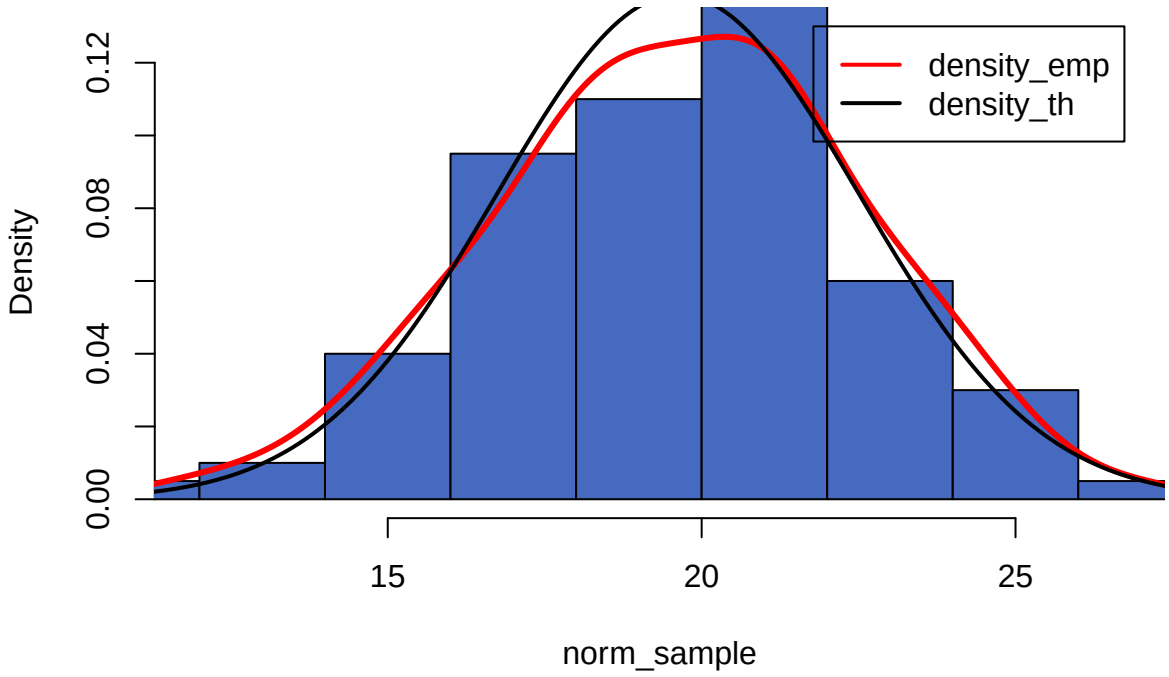


On peut ensuite donner une estimation de l'espérance et de la variance de nos données de la façon suivante :

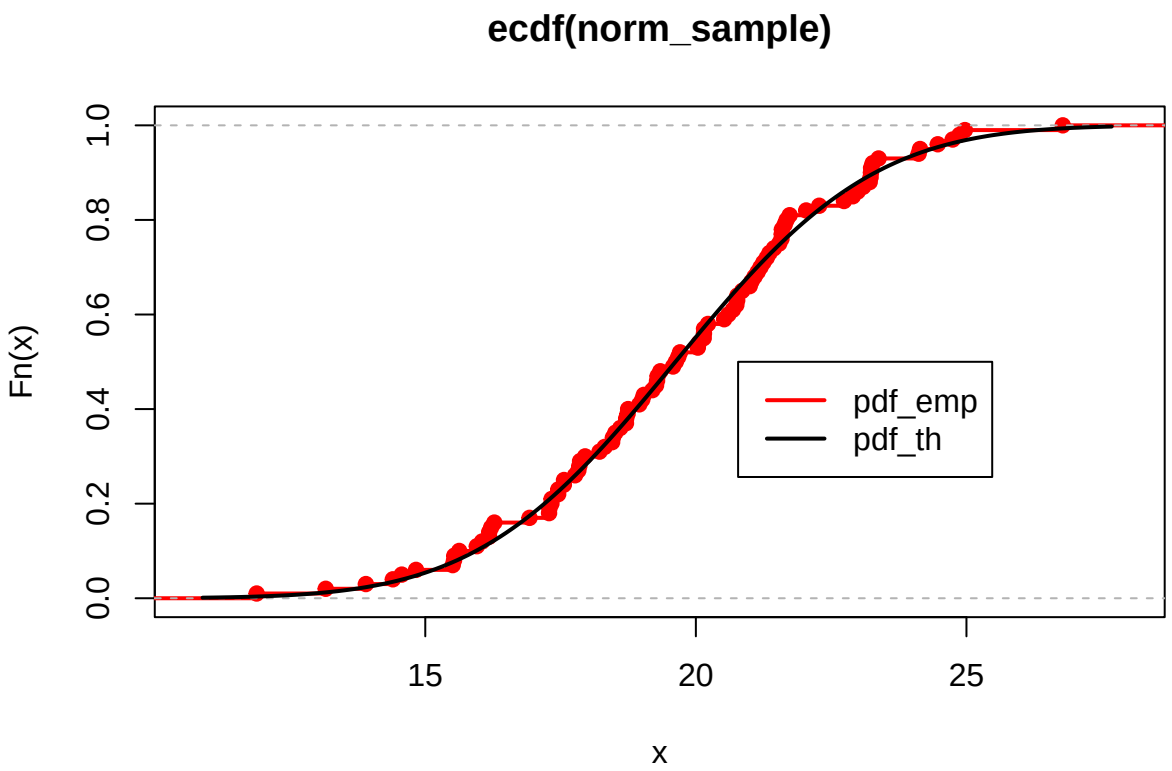
```
mean_emp <- mean(norm_sample)
var_emp <- var(norm_sample)
sd_emp <- sd(norm_sample)
```

On se propose ensuite de reprendre nos précédents graphes mais en y supperposant les densités/pdf avec les paramètres obtenus précédemment

```
# Histogramme
hist(norm_sample, breaks=10, prob = TRUE,
     col = "#456ABF", border= "black", ylim = c(0,0.13), xlim =c(min(norm_sample), max(norm_sample)))
lines(density(norm_sample), col = "red", lwd = 3)

x=seq(min(norm_sample)-1,max(norm_sample)+1,0.1)
lines(x,dnorm(x,mean_emp,sd_emp), col= "black", lwd = 2)
legend(max(norm_sample)-5, 0.13, legend = c("density_emp", "density_th"),
       col = c("red","black"), lwd = 2)
```




```
# Fonction de repartition empirique

plot(ecdf(norm_sample), col = "red", lwd =2)
x=seq(min(norm_sample)-1,max(norm_sample)+1,0.1)
lines(x,pnorm(x,mean_emp,sd_emp), col= "black", lwd = 2)
legend(max(norm_sample)-6,0.5, legend = c("pdf_emp", "pdf_th"), col = c("red","black"), lwd = 2)
```



On test maintenant la normalité de données à l'aide du test de Shapiro Wilk pour différents tailles de notre jeu de données

```
# Jeu de données actuel

shapiro.test(norm_sample)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  norm_sample
## W = 0.99466, p-value = 0.9654
```

```
# Lorsque n = 10

n=10
norm_sample_10 <- rnorm(n, mu, ect)
shapiro.test(norm_sample_10)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  norm_sample_10
## W = 0.95827, p-value = 0.766
```

```
# Lorsque n = 1000

n=1000
norm_sample_1000 <- rnorm(n, mu, ect)
shapiro.test(norm_sample_1000)
```

```
##
##  Shapiro-Wilk normality test
```

```
##
## data:  norm_sample_1000
## W = 0.99731, p-value = 0.09522
```

On peut exécuter plusieurs fois notre test de shapiro sur les différents échantillons, on remarque que le test conduira toujours au non rejet de l'hypothèse nulle, à savoir, au non rejet de la normalité des données.

Un résultat plutôt attendu étant donnés que nos échantillons sont tirés selon une loi normale.

On finit cette série d'expériences en considérant un échantillon tiré selon une exponentielle. On prendra également un échantillon de taille 10, de taille 100 et de taille 1000.

On rappelle qu'une loi exponentielle de paramètre λ admet comme moyenne la valeur $\frac{1}{\lambda}$

```
# Simualtion d'échantillons suivants une loi exponentielles
moyenne = 20
lambda = 1/moyenne

exp_sample_10 <- rexp(10, lambda)
exp_sample_100 <- rexp(100, lambda)
exp_sample_1000 <- rexp(1000, lambda)

# Tests de normalité

shapiro.test(exp_sample_10)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  exp_sample_10
## W = 0.70688, p-value = 0.001067
```

```
shapiro.test(exp_sample_100)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  exp_sample_100
## W = 0.85903, p-value = 2.595e-08
```

```
shapiro.test(exp_sample_1000)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  exp_sample_1000
## W = 0.82558, p-value < 2.2e-16
```

Sans surprise, les tests de shapiro conduisent bien au rejet de l'hypothèse nulle, i.e. au rejet de l'hypothèse de normalité des données. Le rejet est d'autant plus important ou sûr que la taille de l'échantillon augmente.

Parenthèse sur la transformation de Boxcox

Présentation

Il s'agit d'une transformation proposée par Box et Cox dans les années 60 et qui consiste à transformer un échantillon prenant des valeurs positives de façon à ce que l'échantillon se rapproche d'une loi normale.

La transformation employée est une application $\varphi_\lambda : \mathbb{R}_+^\star \to \mathbb{R}$, pour tout $\lambda \in \mathbb{R}$ définie par

$$\varphi_\lambda(x) = \begin{cases} \dfrac{x^\lambda - 1}{\lambda} & \text{si} \quad \lambda \neq 0 \\ \log(x) & \text{si} \quad \lambda = 0 \end{cases}$$

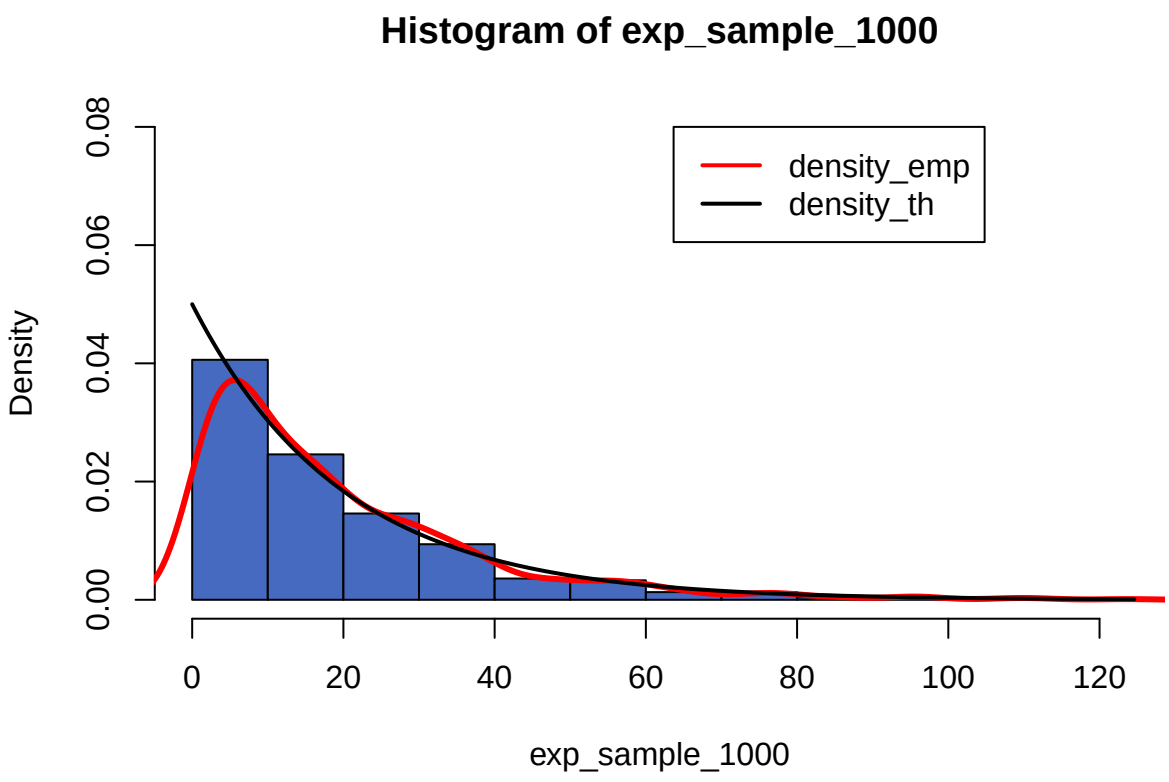
Dans le cas où $\lambda = 0$, il s'agit d'une transformation logarithmique des données. Lorsque $\lambda = 1$, on ne transforme pas réellement les données mais on se contente d'effectuer une translation de 1 de ces dernières. Notons que comme la fonction est strictement monotone et croissante, alors l'ordre des valeurs de x est conservé après transformation des données.

La question est maintenant de savoir comment déterminer la valeur de λ. On passe les aspects théoriques (vraisemblance) et on va directement se concentrer sur la pratique en utilisant la fonction "boxcox" de R.

Mise en application

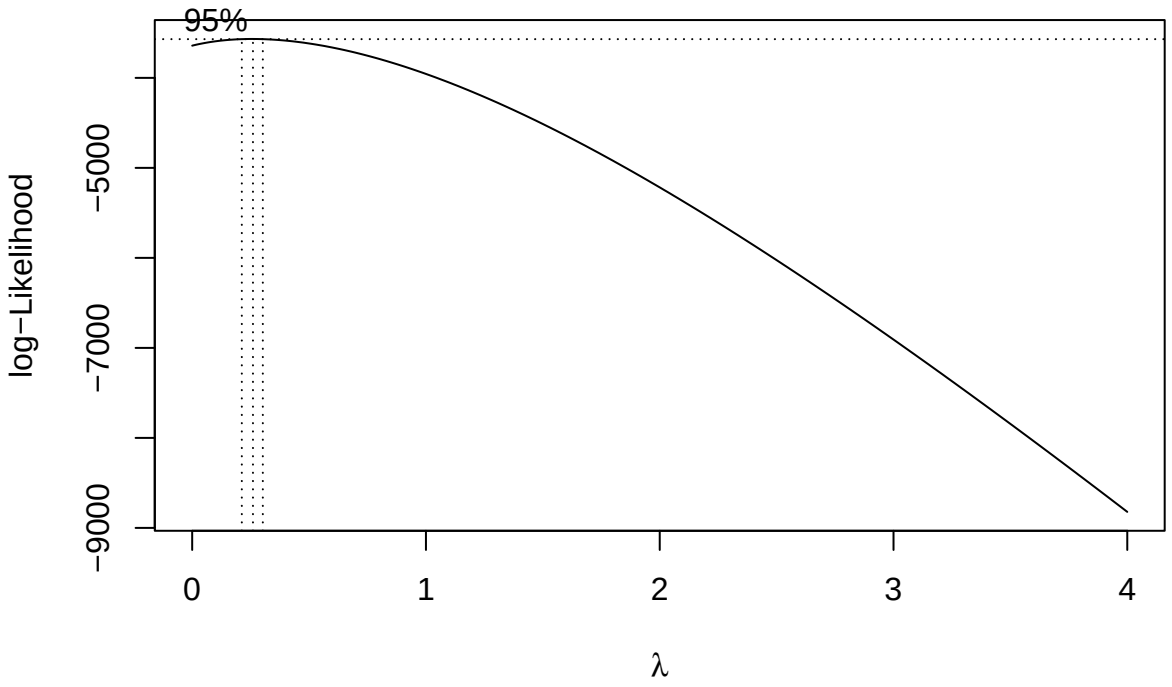
On va commencer par reprendre notre échantillon tiré selon la loi exponentielle de taille $n = 1000$ et on va regarder qu'elle est la bonne transformation à appliquer pour obtenir une loi normale

```
# Histogramme
hist(exp_sample_1000, breaks=10, prob = TRUE,
     col = "#456ABF", border= "black", ylim = c(0,0.08), xlim =c(0, max(exp_sample_1000)))
lines(density(exp_sample_1000), col = "red", lwd = 3)

x=seq(0,max(exp_sample_1000)+1,0.1)
lines(x,dexp(x,lambda), col= "black", lwd = 2)
legend(max(exp_sample_1000)-60, 0.08, legend = c("density_emp", "density_th"),
       col = c("red","black"), lwd = 2)
```

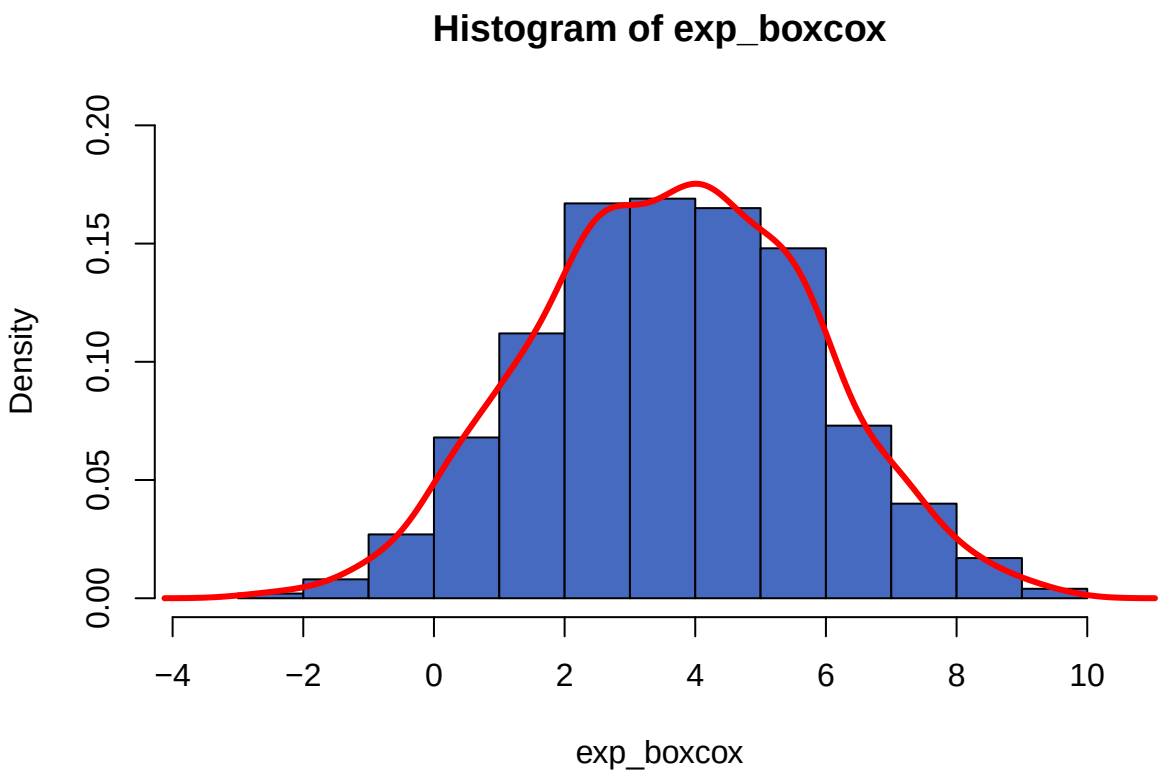


On détermine la bonne valeur de γ pour la transformation de boxcox de nos données. C'est celui qui va maximiser notre log-vraisemblance

```r
# Calcul du gamma optimal
library(MASS)

res_box <- boxcox(exp_sample_1000~1, lambda = seq(0,4,0.01))
```




```r
gamma_opt <- res_box$x[which.max(res_box$y)]

# Transformation des données et représentation des données

exp_boxcox <- (exp_sample_1000^(gamma_opt)-1)/gamma_opt

hist(exp_boxcox, breaks=10, prob = TRUE,
     col = "#456ABF", border= "black", ylim = c(0,0.20), xlim =c(min(exp_boxcox)-1, max(exp_boxcox)+1))
lines(density(exp_boxcox), col = "red", lwd = 3)
```




```
# On peut ensuite tester la normalité

shapiro.test(exp_boxcox)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  exp_boxcox
## W = 0.99821, p-value = 0.3796
```

Données réelles

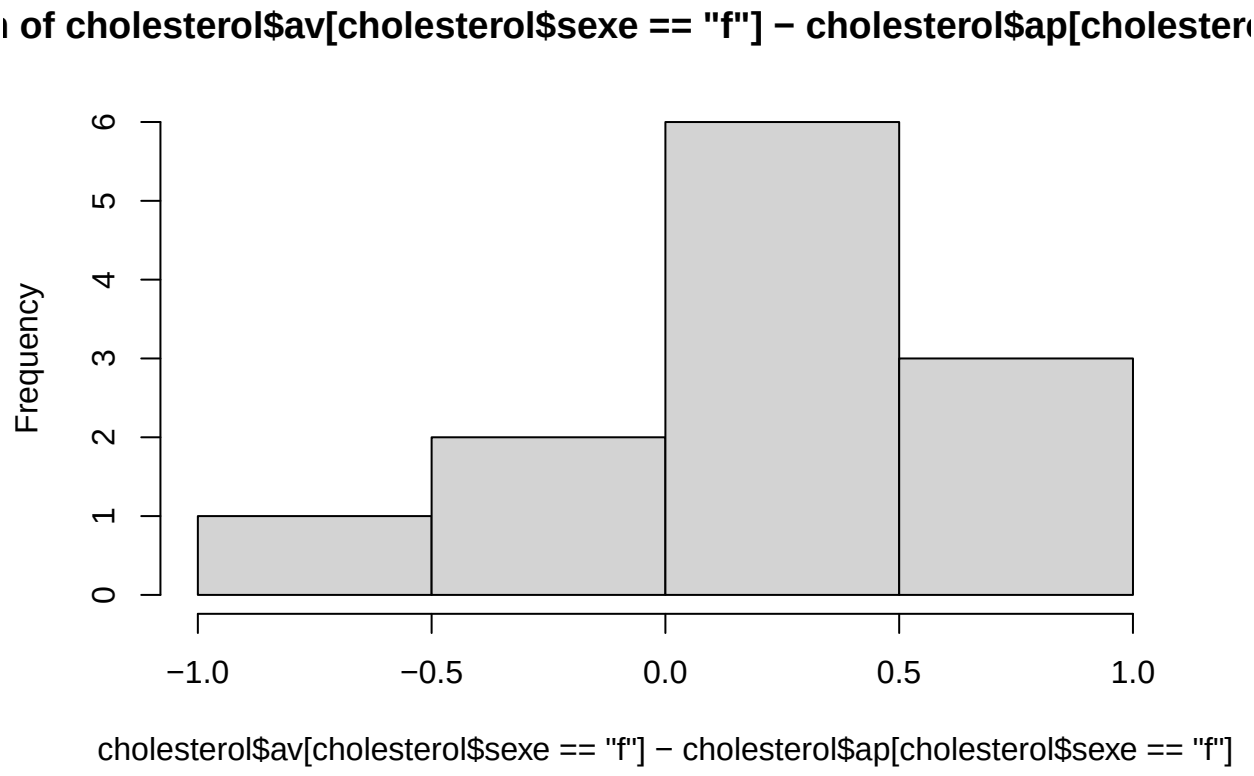
On se propose ici de reprendre les données "Cholesterol" étudiées lors du précédent TD afin de regarder si les données sont issues d'une distribution normale ou non pour les deux sexes.

```
cholesterol <- read.csv2("~/Desktop/A exporter/Cours/Lyon2/Statistiques_Master/cholesterol.csv")
```

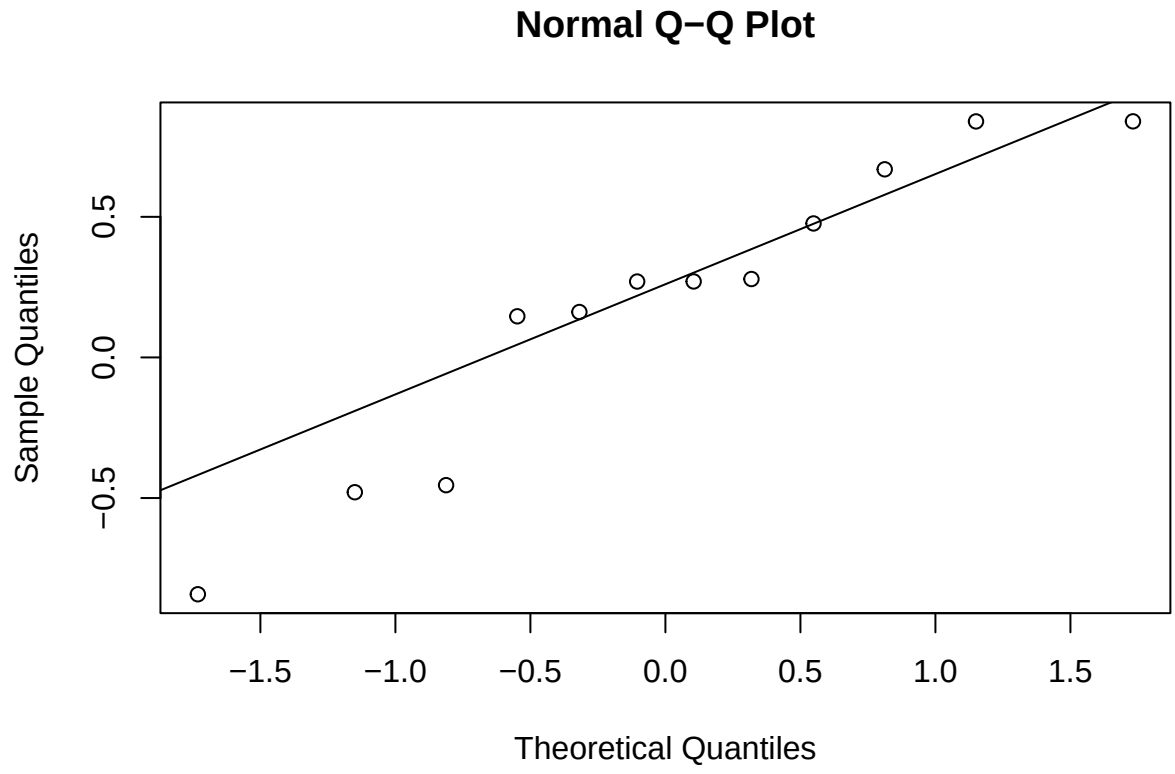
On peut faire cela sur les échantillons évaluant le taux de cholestérol avant la prise du médicament, après la prise du médicament ou encore sur l'échantillon différence. La démarche étant la même à chaque fois, on le fera uniquement pour l'échantillon différence.

Pour les femmes

```
# Echantillon différence du taux de cholesterol

# Histogramme
hist(cholesterol$av[cholesterol$sexe=="f"]-cholesterol$ap[cholesterol$sexe=="f"], breaks=4)
```




```
# qqplot : il est d'usage de représenter aussi la droite en Henry sur
# ce type de graphe pour un jugement qualtitatif sur la normalité des
# données
qqnorm(cholesterol$av[cholesterol$sexe=="f"]-cholesterol$ap[cholesterol$sexe=="f"])
qqline(cholesterol$av[cholesterol$sexe=="f"]-cholesterol$ap[cholesterol$sexe=="f"])
```



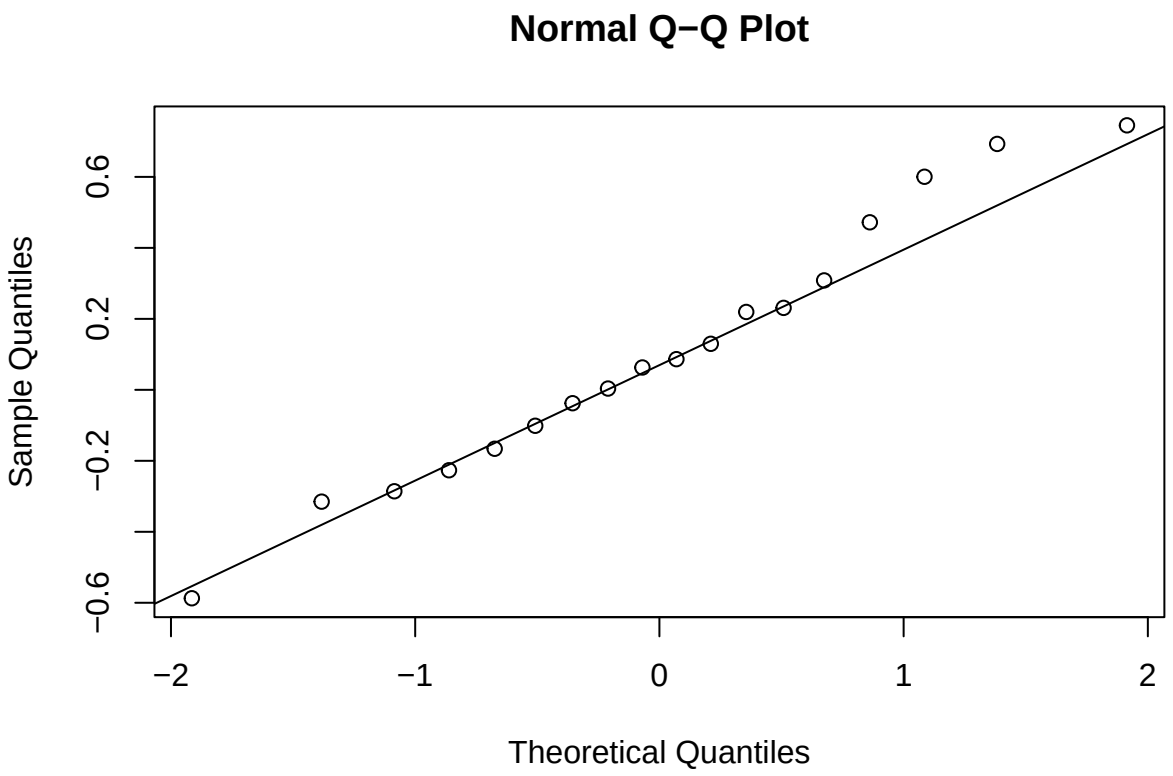
```r
# test de shapiro
shapiro.test(cholesterol$av[cholesterol$sexe=="f"]-cholesterol$ap[cholesterol$sexe=="f"])
```

```
##
##  Shapiro-Wilk normality test
##
## data:  cholesterol$av[cholesterol$sexe == "f"] - cholesterol$ap[cholesterol$sexe == "f"]
## W = 0.91513, p-value = 0.2481
```

Pour les hommes

```r
# Echantillon différence du taux de cholesterol

# Histogramme
hist(cholesterol$av[cholesterol$sexe=="h"]-cholesterol$ap[cholesterol$sexe=="h"], breaks=4)
```


```r
# qqplot : il est d'usage de représenter aussi la droite en Henry sur
# ce type de graphe pour un jugement qualtitatif sur la normalité des
# données
qqnorm(cholesterol$av[cholesterol$sexe=="h"]-cholesterol$ap[cholesterol$sexe=="h"])
qqline(cholesterol$av[cholesterol$sexe=="h"]-cholesterol$ap[cholesterol$sexe=="h"])
```




```
# test de shapiro
shapiro.test(cholesterol$av[cholesterol$sexe=="h"]-cholesterol$ap[cholesterol$sexe=="h"])
```

```
##
##  Shapiro-Wilk normality test
##
## data:  cholesterol$av[cholesterol$sexe == "h"] - cholesterol$ap[cholesterol$sexe == "h"]
## W = 0.97629, p-value = 0.9039
```

Il est également possible de tester la normalité des données à l'aide d'un test de Kolmogorov-Smirnov (ks.test).