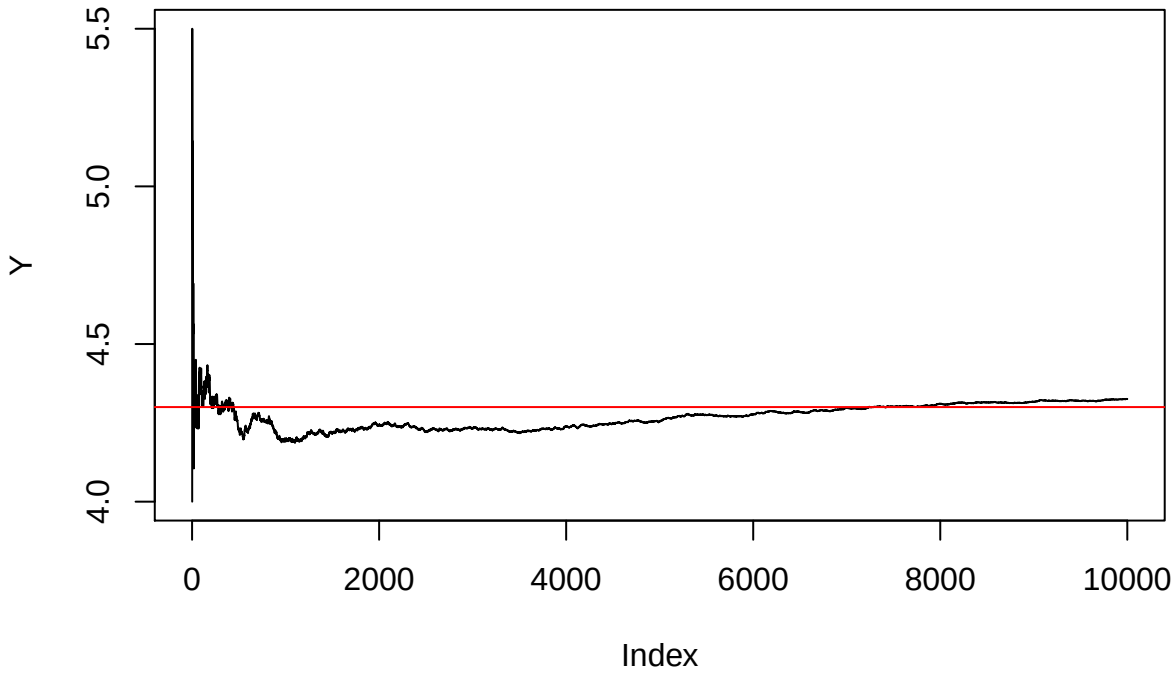
TP 3 : Estimateurs de la moyenne et de la variance

Apéritif Illustrons tout d'abord la loi (forte) des grands nombres pour des variables de Poisson de paramètre $\lambda = 4.3$ par exemple :

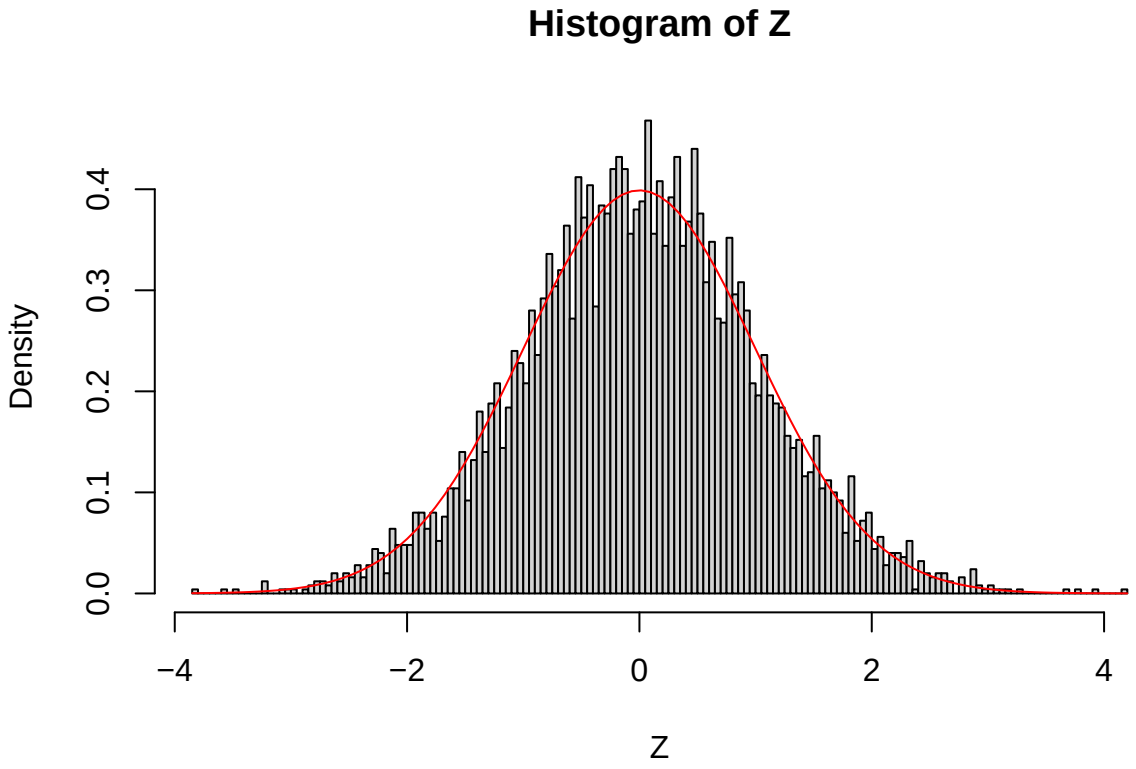
```
n = 10000
X = rpois(n,lambda=4.3)
# Y = cumsum(X)/1:n pour une version directe

## sinon :
Y = replicate(n,0) # j'initialise Y avec des zéros

for (k in 1:n){
  Y[k] = sum(X[1:k])/k
}

plot(Y,type="l") # affiche Y en fonction de k
abline(a=4.3,b=0,col="red") # superpose une droite d'équation y = 4.3 (horizontale)
```



Puis, le théorème Central limite :

```
n = 5000 # nb de variables X_j pour chaque valeur de i
N = 5000 # nb de valeurs de Z à simuler
Z = replicate(N,0)
for (i in 1:N){
  Z[i] = sqrt(n)*(sum(rpois(n,lambda=4.3))/n - 4.3)/sqrt(4.3)
}
```

```r
hist(Z,breaks = 190, freq=FALSE)
curve(dnorm(x, mean=0, sd=1),add=TRUE,col ="red")
```



Exercice 1 : Etude de données réelles, estimateurs de la moyenne et de la variance.

1. Voici le code pour la fonction h:

```r
h = function(x,alpha,tau){
  if (x < tau) {
    0
  }
  else {
    alpha*exp(-alpha*(x-tau))
  }
}
```

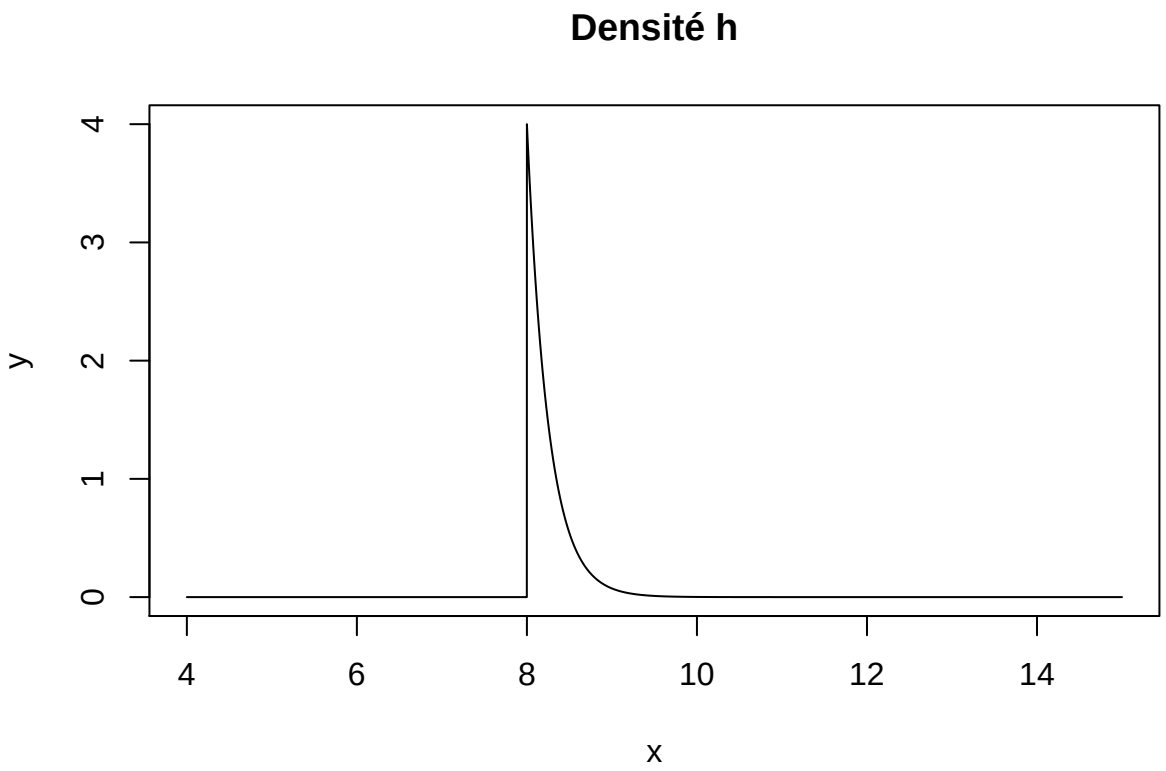
ou, avec des indicatrices

```r
hbis = function(x,alpha,tau){
  alpha*exp(-alpha*(x-tau))*(x >= tau)
  # on incorpore la condition "x >= tau" sous forme d'indicatrice
}
```

On veut appliquer h à un vecteur, mais h a plusieurs arguments qui ne sont pas tous des vecteurs. On fait donc une boucle for.

```r
x = seq(4,15,by = 0.001) # fonction seq, qui découpe un intervalle en petits "pas" (paramètre by)
n = length(x)
y = replicate(n,0) # j'initialise y avec des 0

for (i in 1:n){
```

```
  y[i] = h(x[i],alpha=4,tau=8)
}

plot(x,y,type="l",main="Densité h")
```



2. Ce modéle est paramètrique, ces paramètres sont α (qu'on peut appeler *taux*) et τ (qu'on peut appeler *décalage* ou *shift*)

(non demandé). On peut vérifier que h est bien une fonction de densité.

$$\int_{\mathbb{R}} h(x,\alpha,\tau)dx = \int_{\tau}^{+\infty} \alpha \exp(-\alpha(x-\tau))dx = [-\exp(-\alpha(x-\tau))]_{\tau}^{+\infty} = 1.$$

3. On sait d'après le cours que $m_n(X_1,\dots,X_n) = \frac{1}{n}\sum_{i=1}^n X_i$, et $v_n(X_1,\dots,X_n) = \frac{n}{n-1}\left(\frac{1}{n}\sum_{i=1}^n X_i^2 - \left(\frac{1}{n}\sum_{i=1}^n X_i\right)^2\right)$.

4. D'après le cours (ou la LFGN), on a les convergence p.s. suivantes :

$$m_n \to \mathbb{E}[X] = \int_{\tau}^{+\infty} \alpha x \exp(-\alpha(x-\tau))dx = [-x\exp(-\alpha(x-\tau))]_{\tau}^{+\infty} + \int_{\tau}^{+\infty} \exp(-\alpha(x-\tau))dx = \tau + 1/\alpha,$$

et

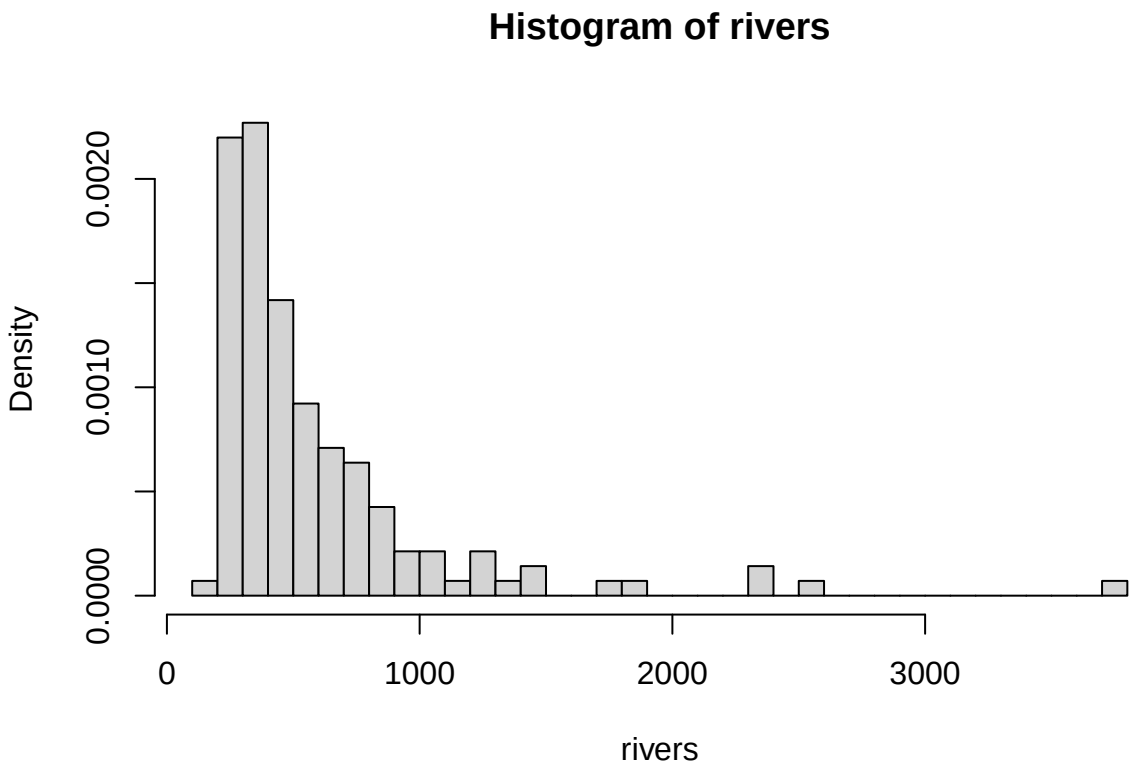
$$v_n \to \mathrm{Var}[X] = \int_{\tau}^{+\infty} \alpha x^2 \exp(-\alpha(x-\tau))dx - \mathbb{E}[X]^2$$

$$= \left[-x^2\exp(-\alpha(x-\tau))\right]_{\tau}^{+\infty} + \int_{\tau}^{+\infty} 2x\exp(-\alpha(x-\tau))dx - \mathbb{E}[X]^2 = \tau^2 + 2\tau/\alpha + 2/\alpha^2 - (\tau+1/\alpha)^2 = 2/\alpha^2 - 1/\alpha^2 = 1/\alpha^2.$$

Les estimateurs de $\hat{\alpha}$ et $\hat{\tau}$ des paramètres α et τ associés sont donc $\hat{\alpha} = \frac{1}{\sqrt{v_n}}$, et $\hat{\tau} = m_n - \sqrt{v_n}$.

5. On affiche les données sous forme d'histogramme :

```
hist(rivers,breaks=30,freq=FALSE)
```



Puis, on calcule les valeurs des estimateurs.

```
n = length(rivers) # taille de l'échantillon
m = mean(rivers) # estimateur de la moyenne
v = var(rivers) # estimateur de la variance sans biais (var est déjà sans biais sous R)
alpha_est = 1/sqrt(v) # calcul de l'estimateur de alpha
tau_est = m-sqrt(v) # calcul de l'estimateur de tau
alpha_est
```

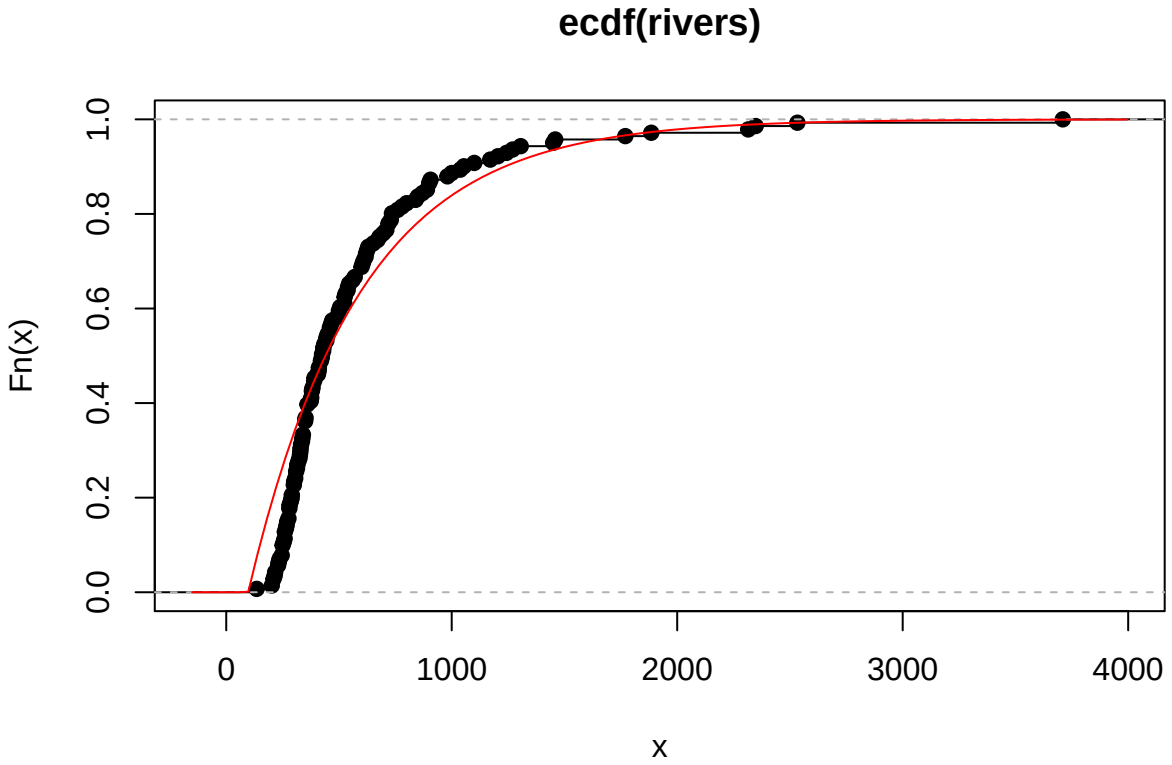
```
## [1] 0.002024821
```

```
tau_est
```

```
## [1] 97.31356
```

6. On superpose la fonction de répartition des données et la fonction de répartition théorique de la loi pour les paramètres estimés précédemment. On remarque que pour la fonction de répartition théorique, il suffit de prendre la fonction de répartition d'une loi exponentielle de paramètre alpha et et de la translater de tau. On utilisera `pmaloi` pour une fonction de repartition empirique d'une loi connue `maloi`.

```
plot(ecdf(rivers)) # fonction de répatition EMPIRIQUE de l'échantillon
curve(pexp(x-tau_est,rate=alpha_est),add=TRUE,col="red") # courbe de la fonction de répartition THEORIQ
```



On observe que les deux courbes sont proches, sans toutefois être confondues : le modèle est correct mais pourrait être raffiné pour mieux expliquer les données.

Pour celleux qui veulent aller plus loin : une autre loi universelle

1. Si A est de taille $n \times n$ avec des entrées i.i.d. Gaussiennes centrées réduites, alors M définie par

$$M = \frac{A + A^T}{\sqrt{2}}$$

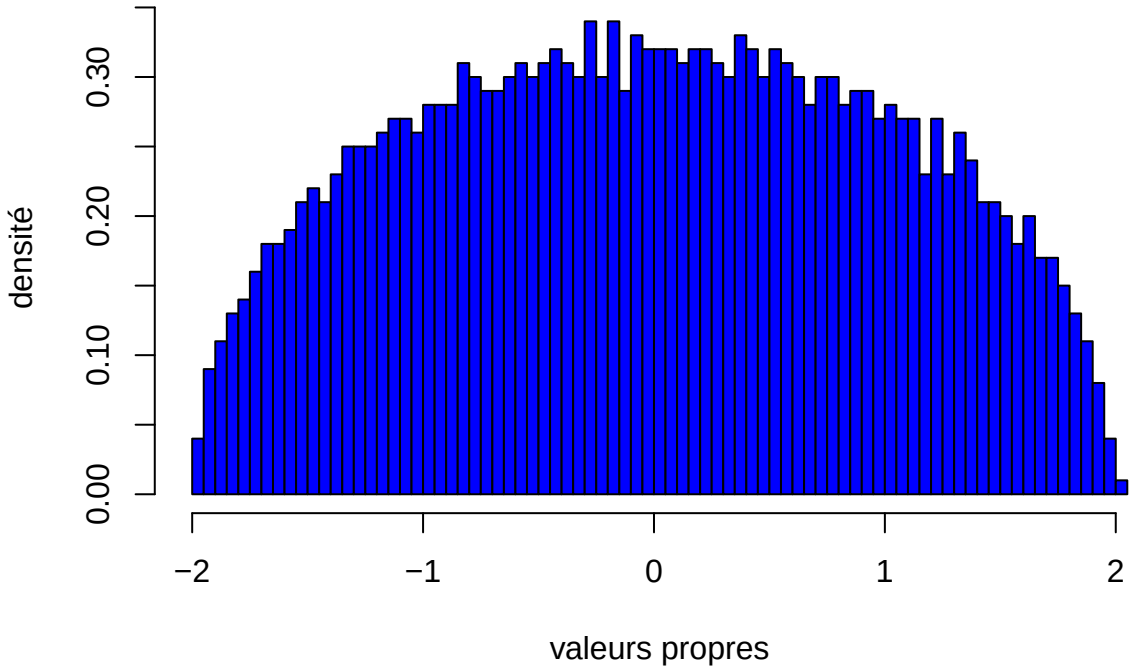
est bien symétrique. De plus, pour $i < j$ et $k < l$ avec $(i, j) \neq (k, l)$, M_{ij} et M_{kl} sont bien indépendantes, de loi normale d'espérance $\frac{0+0}{\sqrt{2}} = 0$ et de variance $\frac{1+1}{(\sqrt{2})^2} = 1$. Enfin, les éléments diagonaux $M_{ii} = \frac{2A_{ii}}{\sqrt{2}} = \sqrt{2}A_{ii}$ suivent des lois $\mathcal{N}(0, 2)$.

2. Voici le code.

```
n=2000 #la matrice sera de taille n fois n
X = rnorm(n*n,mean=0,sd=1) #on simule n^2 gaussiennes standard i.i.d.
A = matrix(X, ncol=n) #on range X en une matrice de n colonnes (et donc n lignes)
M = (A + t(A))/sqrt(2) #on calcule M à partir de A comme indiqué dans l'énoncé
```

3. Voici le code pour calculer le spectre de $M/\sqrt{n}$ et afficher l'histogramme des valeurs propres.

```
spectre = eigen(M/sqrt(n),symmetric=TRUE, only.values=TRUE)
hist(spectre$values,breaks=120,freq=FALSE,main="Histogramme des valeurs propres de M",xlab="valeurs pro
```



De façon assez jolie, la loi du spectre de la matrice M semble avoir une densité qui à la forme… d'un demi-cercle. *On visualise un autre théorème universel ici, s'appliquant à des objets plus particuliers (les matrices aléatoires) : il s'agit de la loi du demi-cercle de Wigner.*