TP Algorithme EM

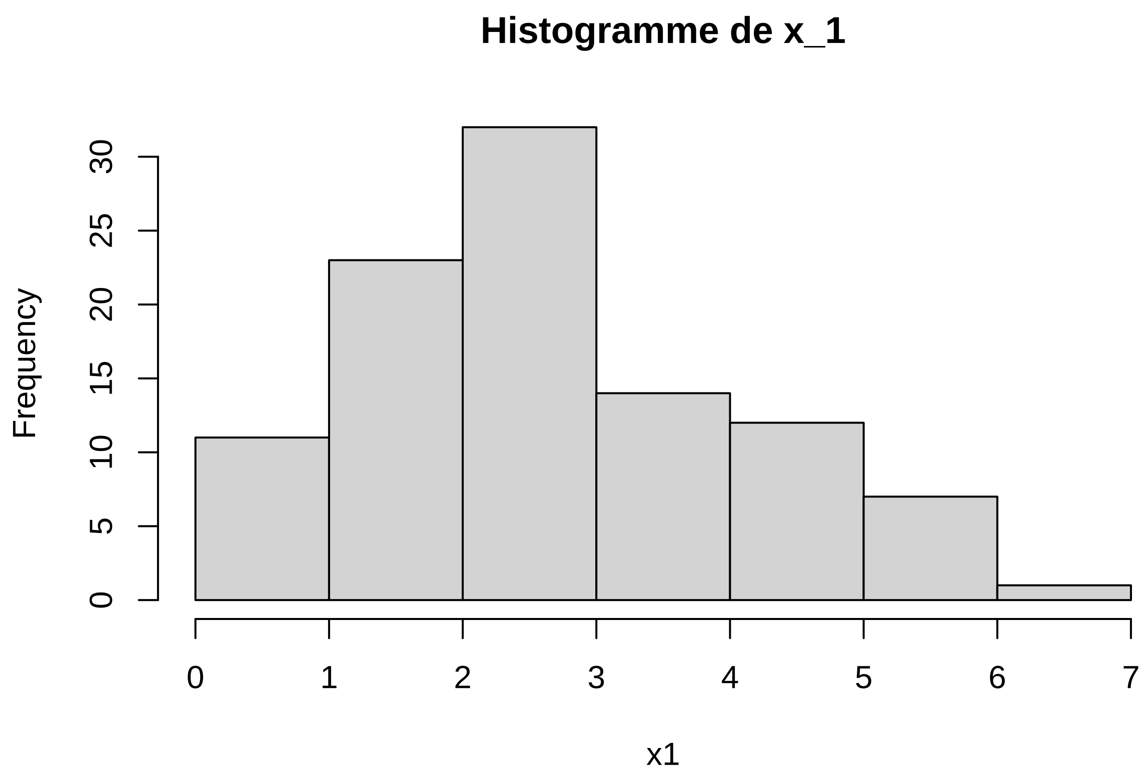
Buchon Valentin, Louan Ourvouai

05 November 2022

Simulation

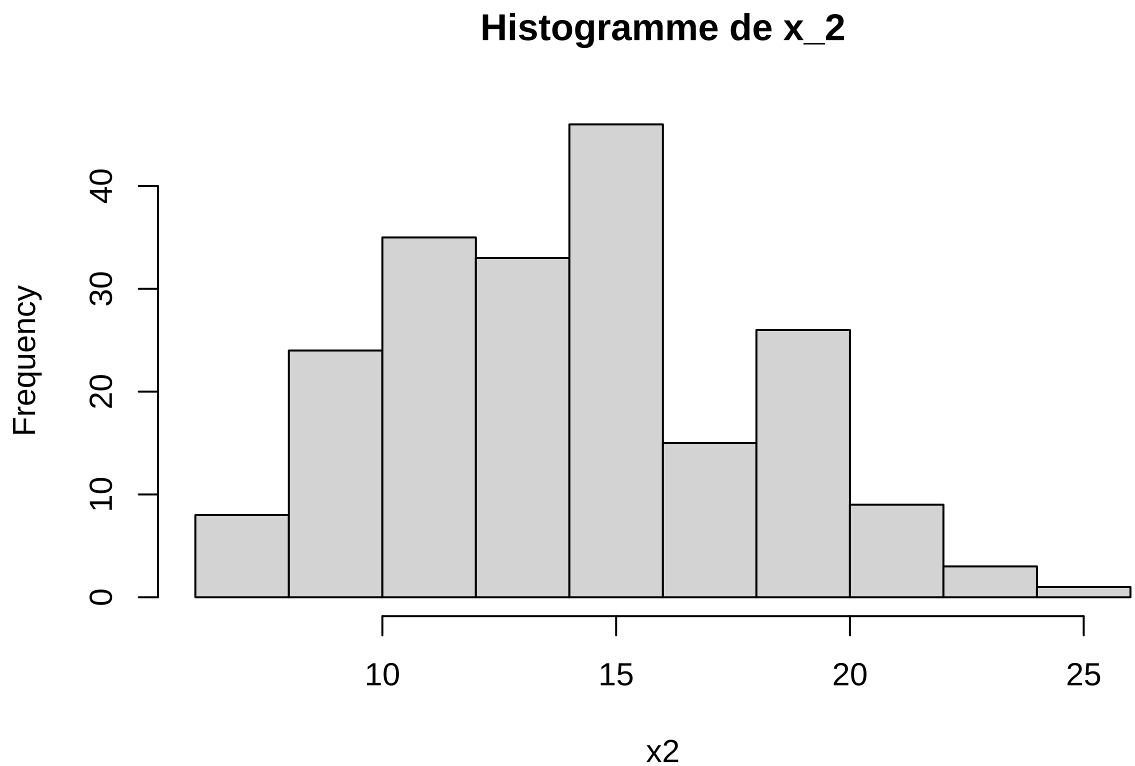
Question 1

On simule l'échantillon x_1 de taille $n = 100$ d'une loi de Poisson de paramètre $\lambda = 3$.



Question 2

On simule l'échantillon de taille $n = 200$ d'une loi de Poisson de paramètre $\lambda = 15$.



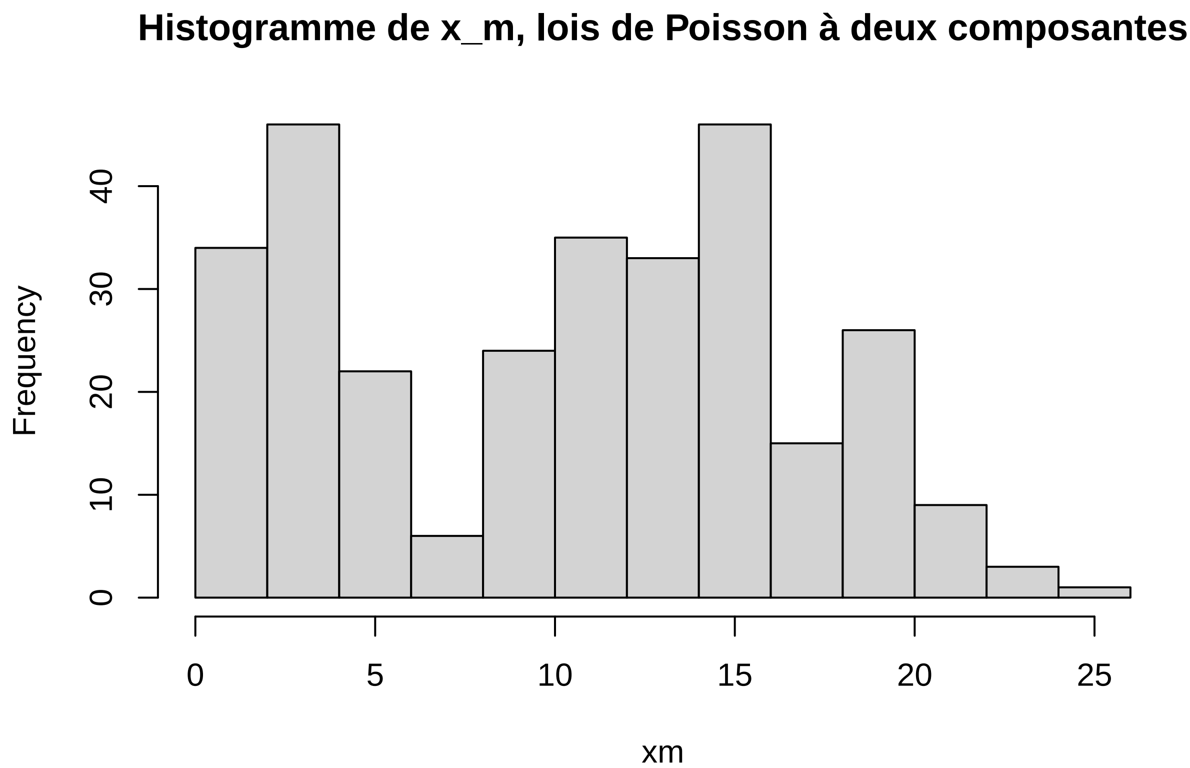
Question 3

On crée alors un vecteur contenant 100 valeurs égales à 1, et 200 valeurs égales à 2.

Vecteur : 1 2

Question 4

Pour simuler un mélange de Poisson de paramètres $\lambda_1 = 3$ et $\lambda_2 = 15$, on peut utiliser les simulations précédentes (en les pondérant par les coefficients $\pi_1 = 0.33$ et $\pi_2 = 1 - \pi_1 = 0.67$).



Algorithme EM pour un mélange de loi de Poisson à K composantes

Question 1

La création de l'initialisation est faite dans le code qui suit. On initialise les proportions π_k toutes égales à $\frac{1}{K}$, et les paramètres λ_k choisis aléatoirement parmi les observations.

```
# Paramètres du problèmes
# Ici K = 2 et n = 300
K<-2
n<-300

init<-function(X, K) {
  theta<-c()
  len = length(X)
  for (i in 1:K) {
    theta[i] = X[runif(1, min = 1, max = len)]
  }
  return(c(rep(1/K, K), theta))
}
```

Question 2

Le code qui suit permet de créer l'étape **E**. Le but de cette fonction est de retourner la matrice $T = \left(t_{i,k}^{(q)}\right)_{1\leq i\leq n, 1\leq k\leq K}$ à l'étape q.

Cette matrice est contituée des coefficients $t_{i,k}^{(q)}$ défini comme suivant : $t_{i,k}^{(q)} = \mathbb{P}\left(z_i = k | x_i, \theta^{(q)}\right)$
Les variables aléatoires $(z_i)_{1\leq i\leq n}$ sont les variables dont chacunes des observations $(x_i)_{1\leq i\leq n}$ proviennent. D'après la formule de Bayes, on peut calculer cette valeur. On obtient alors la formule suivante.

$$\begin{aligned} t_{i,k}^{(q)} &= \mathbb{P}\left(z_i = k | x_i, \theta_k^{(q)}\right) \\ &= \frac{\mathbb{P}\left(x = x_i | z_i = k, \theta_k^{(q)}\right)\mathbb{P}\left(z_i = k, \theta_k^{(q)}\right)}{\mathbb{P}\left(x = x_i, \theta_k^{(q)}\right)} \\ &= \frac{\pi_k f(x_i, \theta_k^{(q)})}{F(x_i, \Theta^{(q)})} \end{aligned}$$

(Avec F la densité totale telle que $F(x, \Theta) = \sum_{k=1}^{K} \pi_k f(x, \theta_k)$)

```r
# Implémentation de la densité totale de poisson, utile dans notre cas
density_pois_tot<-function(x, theta) {
  somme<-0
  K_<- as.integer(length(theta)/2)
  for (k in 1:K_) {
    somme<- somme + theta[k] * dpois(x, theta[k + K_])
  }
  return(somme)
}
```

```r
E_step<-function(dens, dens_tot, theta_q, X, K) {
  T<-matrix(nrow = length(X), ncol = K)
  for (i in 1:length(X)) {
    for (k in 1:K) {
      T[i,k]<-theta_q[k] * dens(X[i],theta_q[k + K]) / dens_tot(X[i], theta_q)
    }
  }
  return(T)
}
```

Question 4

On cherche alors à maximiser $Q(\theta, \theta^{(q)})$, voici son expression.

$$\begin{aligned} Q(\theta, \theta^{(q)}) &= \sum_{i=1}^{n}\sum_{k=1}^{K} t_{i,k} \log(\pi_k f_k(x_i, \theta_k)) \\ &= \sum_{i=1}^{n}\sum_{k=1}^{K} t_{i,k} \left(\log \pi_k + \log\left(\frac{e^{-\lambda_k}}{x_i!}\lambda_k^{x_i}\right)\right) \\ &= \sum_{i=1}^{n}\sum_{k=1}^{K} t_{i,k} \left(\log \pi_k - \lambda_k - \log x_i! + x_i \log(\lambda_k)\right) \end{aligned}$$

On cherche à maximiser cette quantité, on annule donc la dérivée par rapport à un paramètre λ_{k_0}.

$$
\begin{aligned}
&\frac{\partial Q(\theta,\theta^{(q)})}{\partial \lambda_{k_0}} = 0 \\
\Longleftrightarrow\ &\frac{\partial}{\partial \lambda_{k_0}} \left(\sum_{i=1}^{n} \sum_{k=1}^{K} t_{i,k} \left(\log \pi_k - \lambda_k - \log x_i! + x_i \log(\lambda_k)\right) \right) = 0 \\
\Longleftrightarrow\ &\sum_{i=1}^{n} t_{i,k_0} \left(-1 + \frac{x_i}{\lambda_{k_0}} \right) = 0 \\
\Longleftrightarrow\ &\frac{1}{\lambda_{k_0}} \sum_{i=1}^{n} t_{i,k_0} x_i = \sum_{i=1}^{n} t_{i,k_0} \\
\Longleftrightarrow\ &\lambda_{k_0} = \frac{\sum_{i=1}^{n} t_{i,k_0} x_i}{\sum_{i=1}^{n} t_{i,k_0}}
\end{aligned}
$$

On obtient donc une formule pour $\lambda_{k_0}^{(q+1)}$.
Pour obtenir les proportions $(\pi_k)_{1\le k\le K}$, il faut résoudre un problème d'optimisation sous contraintes (car la somme des proportions doit valoir 1, soit $\sum_{k=1}^{K} \pi_k = 1$).
On définit alors le Lagrangien du problème comme suivant.
$\mathcal{L}(\theta,\lambda) = Q(\theta,\theta^{(q)}) + \lambda \left(\sum_{k=1}^{K} \pi_k - 1 \right)$
On cherche alors à résoudre ce système.

$$
\begin{cases} \frac{\partial \mathcal{L}(\theta,\lambda)}{\partial \theta} = 0 \\ \frac{\partial \mathcal{L}(\theta,\lambda)}{\partial \lambda} = 0 \end{cases}
\Longleftrightarrow
\begin{cases} \frac{\partial}{\partial \pi_{k_0}} \left(\sum_{i=1}^{n} t_{i,k_0} \log \pi_{k_0} \right) + \frac{\partial}{\partial \pi_{k_0}} \left(\lambda \sum_{k=1}^{K} \pi_k - 1 \right) = 0 \\ \sum_{k=1}^{K} \pi_k - 1 = 0 \end{cases}
$$

$$
\Longleftrightarrow
\begin{cases} \frac{1}{\pi_{k_0}} \sum_{i=1}^{n} t_{i,k_0} - \lambda = 0 \\ \sum_{k=1}^{K} \pi_k = 1 \end{cases}
\Longleftrightarrow
\begin{cases} \pi_{k_0} = \frac{1}{\lambda} \sum_{i=1}^{n} t_{i,k_0} \\ \sum_{k=1}^{K} \pi_k = 1 \end{cases}
$$

Comme on a que $\lambda = \sum_{i=1}^{n} \sum_{k=1}^{K} t_{i,k} = \sum_{i=1}^{n} 1 = n$, on obtient alors la formule suivante pour la proportion k_0 à l'itération $(q+1)$.
$\pi_{k_0}^{(q+1)} = \frac{1}{n} \sum_{i=1}^{n} t_{i,k_0}$
On peut alors implémenter l'étape M calculant les paramètre $\theta^{(q+1)}$.

```
M_step<-function(T_, X) {
  theta_q<-c()
  n<-length(X)
  K<-dim(T_)[2]
  for (k in 1:K) {
    sum_ti_xi<- 0
    sum_ti<- 0
    for (i in 1:n) {
      sum_ti_xi<-sum_ti_xi + T_[i,k]*X[i]
      sum_ti<-sum_ti + T_[i,k]
    }
    theta_q[k]<- sum_ti / n
    theta_q[k + K]<- sum_ti_xi / sum_ti
  }
```

```
  return(theta_q)
}
```

Question 5

Pour appliquer l'algorithme EM, on applique les étapes **E** et **M** jusqu'à convergence. Cela donne l'implémentation suivante.

```
algorithme_EM<-function(dens, dens_tot, X, K, eps) {
  theta_q<-init(X, K)
  T_<-E_step(dens,dens_tot, theta_q, X, K)
  theta_q1<-M_step(T_, X)
  while (sum((theta_q - theta_q1)^2) / sum((theta_q)^2) > eps) {
    theta_q<-theta_q1
    T_<-E_step(dens,dens_tot, theta_q1, X, K)
    theta_q1<-M_step(T_, X)
  }
  return(theta_q1)
}
```

Résultat : 0.3458449 0.6541551 3.300331 14.62741 Ce qui sont bien les résultats attendu avec $\lambda_1 = 3$, $\lambda_2 = 15$, $\pi_1 = \frac{1}{3}$ et $\pi_2 = \frac{2}{3}$