

Estimation statistique

Semaine 4

Table des matières

Objectifs	1
Exercice 1. Estimateur de l'espérance et la variance.	1
Exercice 2. Maximum de vraisemblance	2

Objectifs

Dans cette séance, on étudiera quelques concepts clés de l'estimation d'une grandeur statistique, par exemple le paramètre d'une loi de probabilité. On présentera les critères d'évaluation de la qualité d'un estimateur ainsi que deux méthodes classiques pour construire des estimateurs : la méthode des moments et la méthode de maximum de vraisemblance.

Exercice 1. Estimateur de l'espérance et la variance.

On considère un échantillon $X = (X_1, \dots, X_n)$ constitué de n variables indépendantes, de même loi d'espérance m et de variance σ^2 finie.

Question 1

Justifier que la moyenne empirique \bar{X}_n est un estimateur sans biais et convergent du paramètre m . Est-il asymptotiquement gaussien ?

Solution. La solution peut être rédigée sur une feuille libre. On ne reportera que les justifications principales (théorèmes, hypothèses).

$$\text{biais}(\bar{X}_n) = \mathbb{E}(\bar{X}_n) - m = m - m = 0$$

Ainsi \bar{X}_n est un estimateur sans biais du paramètre m et d'après la loi faible des grands nombres :

Soit $\epsilon > 0$,

$$\lim_{n \rightarrow \infty} P(|\bar{X}_n - m| \geq \epsilon) = 0$$

\bar{X}_n converge vers m .

En posant $Z_n = \frac{\bar{X}_n - m}{\frac{\sigma}{\sqrt{n}}}$, d'après le Théorème central limite, Z_n converge vers la loi $N(0, 1)$ donc \bar{X}_n est asymptotiquement de loi $N(m, \frac{\sigma^2}{n})$.

Question 2

On définit

$$s_n^2(X) = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2$$

Montrer que $s_n^2(X)$ est un estimateur biaisé de la variance σ^2 . Proposer un estimateur sans biais $\hat{\sigma}_n^2$ de la variance.

Indication : Faire apparaître m dans le calcul de $s_n^2(X)$.

Solution. La solution peut être rédigée sur une feuille libre. On ne reportera que le résultat.

$$\text{biais}(s_n^2(X)) = \mathbb{E}(s_n^2(X)) - \sigma^2 = \frac{n-1}{n}\sigma^2 - \sigma^2 \neq 0$$

$$\hat{\sigma}_n^2 = \frac{n}{n-1}s_n^2(X)$$

Question 3

On considère l'échantillon simulé suivant

```
set.seed(12345)
n = 15
x = rnorm(n, m = 1, sd = 10)
```

Calculer l'estimation biaisée et l'estimation non-biaisée de la variance de cet échantillon en utilisant les formules de la question 2. Comparer le résultat à la commande `var`. Commenter ce résultat.

Solution.

```
## estimation biaisée
s_n <- mean((x-mean(x))^2)
cat("estimation biaisée : ", s_n, "\n")
```

```
## estimation biaisée : 69.36674
```

```
## estimation non-biaisée
s_nb <- mean((x-mean(x))^2)*(n/(n-1))
cat("estimation non-biaisée : ", s_nb, "\n")
```

```
## estimation non-biaisée : 74.32151
```

```
## Sachant que la commande var nous donne la variance de l'échantillon x, on en
## déduit que l'estimation non biaisée est très précise.
cat("variance de l'échantillon x : ", var(x), "\n")
```

```
## variance de l'échantillon x : 74.32151
```

Exercice 2. Maximum de vraisemblance

On considère un échantillon $X = (X_1, \dots, X_n)$ constitué de n variables indépendantes de loi exponentielle de paramètre $\theta > 0$

$$f(x; \theta) = \begin{cases} \theta e^{-\theta x} & \text{si } x > 0, \\ 0 & \text{sinon.} \end{cases}$$

On rappelle que la variable aléatoire

$$S_n = \sum_{i=1}^n X_i$$

suit la loi Gamma(n, θ) dont la densité est donnée ci-dessous

$$g(x; n, \theta) = \begin{cases} \frac{\theta^n}{(n-1)!} x^{n-1} e^{-\theta x} & \text{si } x > 0, \\ 0 & \text{sinon.} \end{cases}$$

Question 1

On considère que les observations suivantes sont issues d'une loi exponentielle de paramètre inconnu.

```
x <- c(0.03 , 0.29 , 0.04 , 0.04 , 0.48 , 0.41, 0.05, 0.00 ,0.06,  
0.05 ,0.18 ,0.01 ,0.06, 0.01, 0.09 ,0.02 ,0.08, 0.01, 0.26, 0.12)
```

Calculer l'estimation du paramètre θ par la méthode du maximum de vraisemblance. Vérifier que l'estimateur EMV correspond à l'estimateur par la méthode des moments.

Solution. La solution peut être en partie rédigée sur une feuille libre. On reportera l'estimateur ainsi trouvé.

Calculons le log-vraisemblance :

$$l(\theta; x) = \ln(L_n(\theta; x)) = \sum_{i=1}^n f(x; \theta) = \sum_{i=1}^n (\ln(\theta) - \theta x_i)$$

En dérivant la fonction de log-vraisemblance par rapport à θ puis en cherchant la valeur annulant cette dérivée on trouve :

$$\frac{\partial l}{\partial \theta} = 0 \Rightarrow \sum_{i=1}^n \left(\frac{1}{\theta} - x_i\right) = 0$$

$$\text{Donc : } \hat{\theta}_{MV} = \frac{n}{\sum_{i=1}^n x_i} = \frac{1}{\bar{X}_n}$$

```
## EMV  
n <- length(x)  
hat_theta = 1/mean(x)  
hat_theta
```

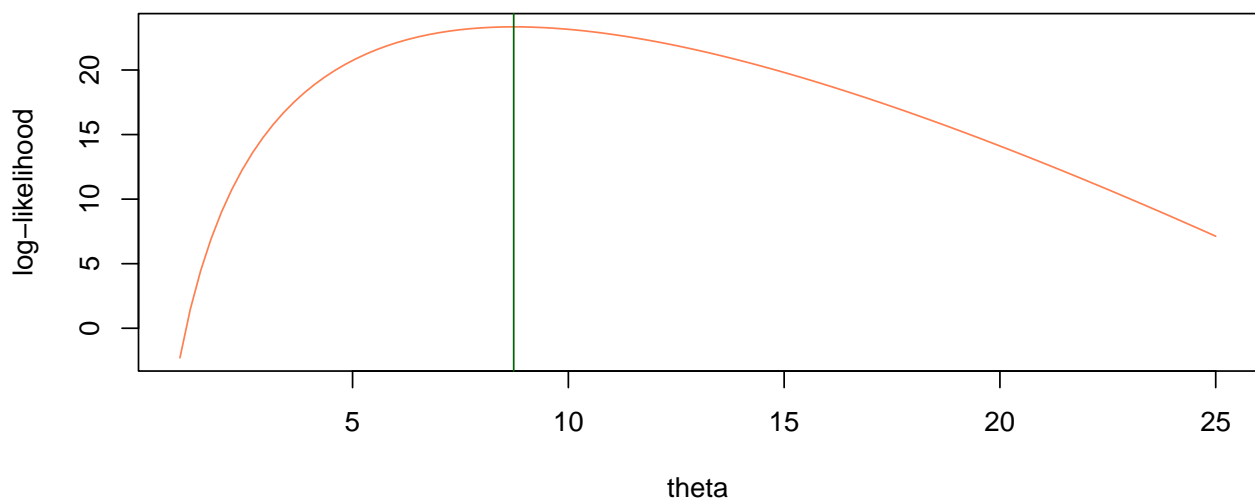
```
## [1] 8.733624
```

Question 2

Tracer la courbe de la log-vraisemblance pour θ appartenant à l'intervalle (0,25). L'optimum est-il très marqué ? Pensez vous que θ pourra être estimé précisément.

Solution. Compléter le code suivant

```
curve(n*log(theta)-theta*sum(x),  
      xname = "theta",  
      from = 1, to = 25, col = "coral",  
      ylab = "log-likelihood")  
  
abline(v = hat_theta, col = "darkgreen")
```



La log-vraisemblance étant assez plate autour de son maximum, les valeurs 8 ou 10 sont pratiquement aussi vraisemblables que l'estimation de maximum de vraisemblance. On s'attend donc que l'estimation de θ soit peu précise.

Question 3

Montrer que l'EMV est un estimateur biaisé. Montrer que l'estimateur défini par $\hat{\theta}_{SB} = (n-1)/S_n$ est sans biais. Calculer l'estimation sans biais du paramètre θ pour les données observées.

Solution.

$$\mathbb{E}[\hat{\theta}_{MV}] = \mathbb{E}\left[\frac{1}{\bar{X}_n}\right]$$

Où \bar{X}_n suit une loi gamma $(n, n\theta)$

Donc $\mathbb{E}[\hat{\theta}_{MV}] = \frac{n}{n-1}\theta$.

$\hat{\theta}_{MV}$ est donc un estimateur biaisé.

Ainsi, l'estimateur

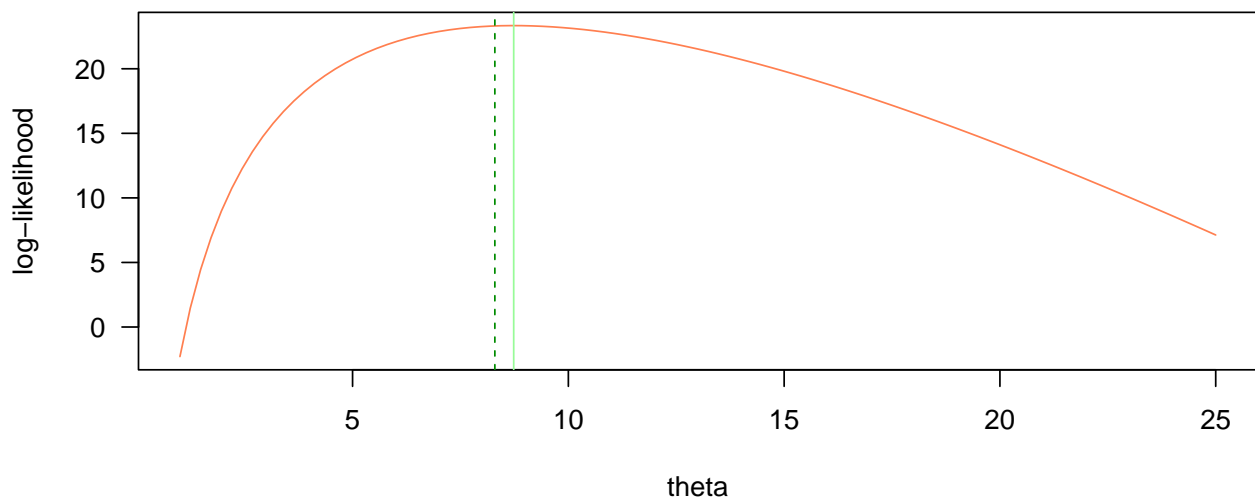
$$\hat{\theta}_{SB} = \frac{n-1}{n} \hat{\theta}_{MV}$$

est un estimateur sans biais de θ . On indique l'estimation de l'estimateur sans biais en pointillé sur le graphe précédent

```
curve(n*log(theta)-theta*sum(x),
      xname = "theta",
      from = 1, to = 25, col = "coral",
      ylab = "log-likelihood", las = 1)

# formule de l'estimateur sans biais
hat_theta_SB = (n-1)/sum(x)

abline(v = hat_theta, col = "palegreen")
abline(v = hat_theta_SB, col = "green4", lty = 2)
```



Question 4

Calculer la variance des estimateurs $\hat{\theta}_{SB}$ et $\hat{\theta}_{MV}$. En déduire que les estimateurs sont convergents.

Solution.

$$\text{Var}(\hat{\theta}_{MV}) = \mathbb{E}[\hat{\theta}_{MV}^2] - \mathbb{E}[\hat{\theta}_{MV}]^2 = \theta^2 \frac{1}{(n-1)(n-2)} - \theta^2 \left(\frac{n}{n-1}\right)^2 = \frac{(n, \theta)^2}{(n-2)(n-1)^2}$$

$$\text{Var}(\hat{\theta}_{SB}) = \mathbb{E}\left[\left(\frac{n-1}{n}\hat{\theta}_{MV}\right)^2\right] - \mathbb{E}\left[\frac{n-1}{n}\hat{\theta}_{MV}\right]^2 = \left(\frac{n-1}{n}\right)^2 \text{Var}(\hat{\theta}_{MV}) = \frac{\theta^2}{n-2}$$

On en déduit que lorsque $n \rightarrow \infty$, $\text{Var}(\hat{\theta}_{MV}) \rightarrow 0$ et $\text{Var}(\hat{\theta}_{SB}) \rightarrow 0$ donc $\hat{\theta}_{MV}$ et $\hat{\theta}_{SB}$ sont des estimateurs convergents.

Question 5

Calculer les valeurs théoriques de l'erreur quadratique moyenne de l'estimateur de maximum de vraisemblance et de l'estimateur sans biais.

$$\text{EQM}(\hat{\theta}_{MV}) = \left(\frac{n}{n-1}\right)^2 \left(\frac{1}{n-2} + \frac{1}{n^2}\right) \times \theta^2$$

et

$$\text{EQM}(\hat{\theta}_{SB}) = \frac{\theta^2}{n-2}$$

Solution. La solution peut être rédigée sur une feuille libre.

On sait que $\text{EQM}(\hat{\theta}_{MV}) = (\mathbb{E}[\hat{\theta}_{MV}] - \theta)^2 + \text{Var}(\hat{\theta}_{MV})$

Donc :

$$\text{EQM}(\hat{\theta}_{MV}) = \left(\frac{\theta}{n-1}\right)^2 + \frac{(n\theta)^2}{(n-2)(n-1)^2} = \theta^2 \left(\frac{(n-2) + n^2}{(n-1)^2(n-2)}\right) = \theta^2 \left(\frac{n}{n-1}\right)^2 \left(\frac{1}{n-2} + \frac{1}{n^2}\right)$$

De même, comme l'estimateur $\hat{\theta}_{SB}$ est un estimateur sans biais on obtient :

$$\text{EQM}(\hat{\theta}_{SB}) = \text{Var}(\hat{\theta}_{SB}) = \frac{\theta^2}{n-2}$$

Question 6

Comparer graphiquement les EQMs des deux estimateurs pour $\theta = 10$ et n compris entre 5 et 50.

Solution. Compléter ou commenter les codes suivants.

```
## erreur quadratique moyenne de l'estimateur de maximum de vraisemblance
eqm_mv = function(theta = 10, n){
  (n/(n-1))^2 * (1/(n-2) + 1/n^2)*theta^2
}

## erreur quadratique moyenne de l'estimateur sans biais
eqm_sb = function(theta = 10, n){
  theta^2*(1/(n-2))
}
```

La comparaison graphique se fait ainsi

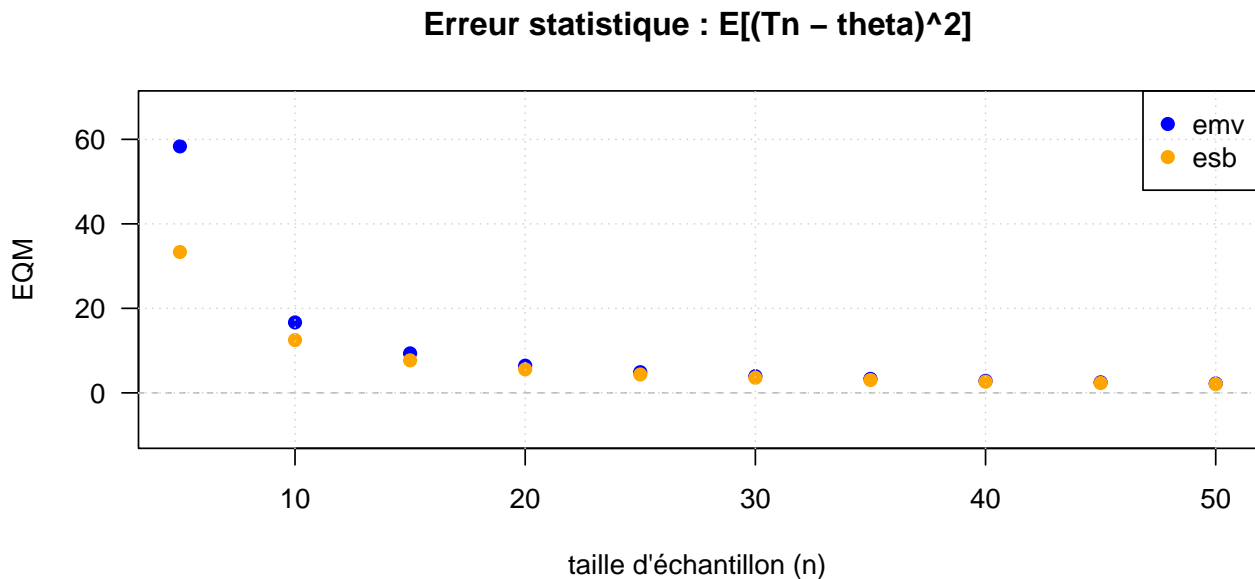
```
## On prend n compris entre 5 et 50
n <- seq(5, 50, by = 5)

## valeur du paramètre de la loi étudiée
theta = 10

## on affiche l'EQM avec un estimateur de MV
plot(n, eqm_mv(theta, n),
     col = "blue",
     xlab = "taille d'échantillon (n)",
     ylab = "EQM",
     main = "Erreur statistique : E[(Tn - theta)^2]",
     pch = 19, ylim = c(-10, max(eqm_mv(theta, n)+ 10)), las = 1)

## On affiche cette fois ci l'EQM pour un estimateur sans biais
points(n, eqm_sb(theta, n), col = "orange", pch = 19)
abline(h = 0, lty = 2, col = "grey")

grid()
legend("topright",
     col = c("blue", "orange"),
     pch = 19,
     legend = c("emv", "esb"))
```



On observe bien que pour des petites valeurs d'échantillons, l'*EMV* est un estimateur biaisé comparé à celui sans-biais. Or pour des valeurs plus importantes d'échantillons, les deux sont semblables en terme d'erreur statistique.

Question 7

Vérifier les résultats théoriques de la question 5 pour $\theta = 10$ et pour plusieurs valeurs de n entre 5 et 50 par des simulations de Monte Carlo utilisant $m = 10000$ répétitions.

Solution. Commenter le code suivant.

```
# vraie valeur
theta = 10
```

```

# tailles d'échantillon
tailles_echantillon <- seq(5, 50, by = 5)

# nombre de répétition MC
m = 10000

eqm_mv_mc = NULL
eqm_sb_mc = NULL

for (nn in tailles_echantillon){
  # Simulation de m échantillons de taille n suivant une loi exponentielle
  X <- matrix(rexp(nn*m, rate = theta), ncol = m)

  # estimateur de maximum de vraisemblance
  emv <- nn/apply(X, 2, FUN = sum)
  eqm_mv_mc <- c(eqm_mv_mc, mean((emv-theta)^2))

  # estimateur sans biais
  esb <- (nn-1)/apply(X, 2, FUN = sum)
  eqm_sb_mc <- c(eqm_sb_mc, mean((esb-theta)^2))
}

```

On désire afficher quoi déjà. Rappelle le ici.

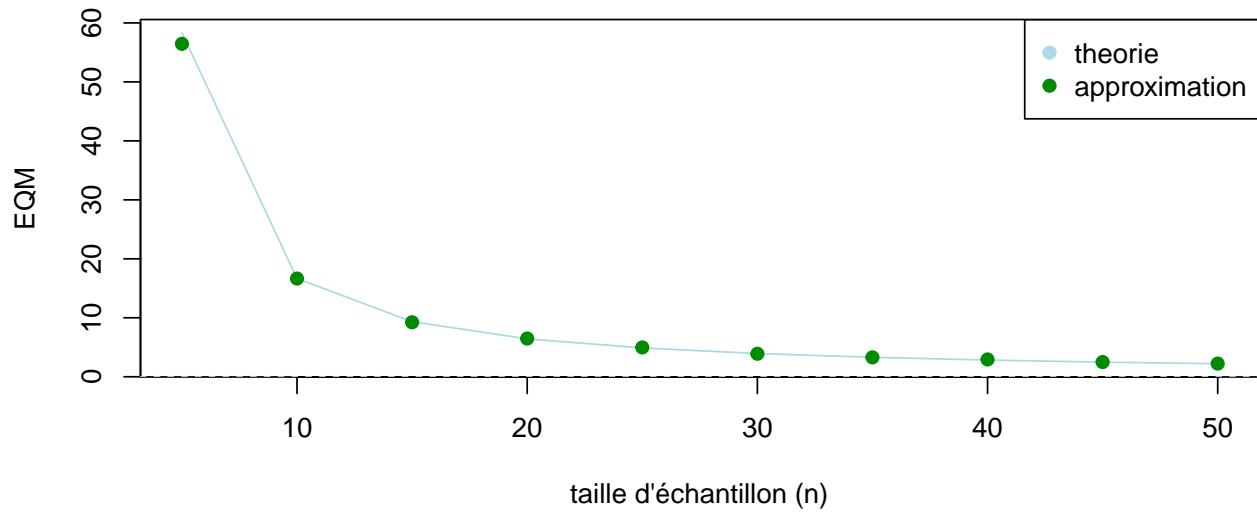
```

# On affiche les valeurs théoriques et expérimentales de l'EQM de MV
plot(n, eqm_mv(10, n),
     col = "lightblue",
     xlab = "taille d'échantillon (n)",
     ylab = "EQM",
     main = "Approximation MV",
     type = "l")
points(n, eqm_mv_mc, col = "green4", pch = 19)
abline(h = 0, lty = 2, col = "grey")

legend("topright",
      col = c("lightblue", "green4"),
      pch = 19,
      legend = c("theorie", "approximation"))

```

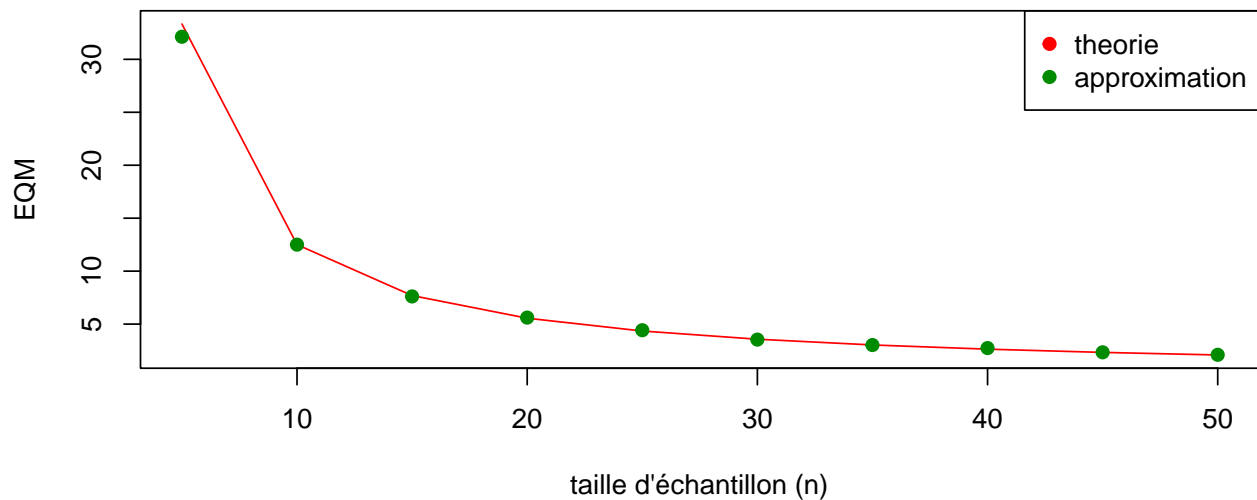
Approximation MV



```
# On affiche les valeurs théoriques et expérimentales de l'EQM sans-biais
plot(n, eqm_sb(10, n),
     col = "red",
     xlab = "taille d'échantillon (n)",
     ylab = "EQM",
     main = "Approximation SB",
     type = "l")
points(n, eqm_sb_mc, col = "green4", pch = 19)
abline(h = 0, lty = 2, col = "grey")

legend("topright",
      col = c("red", "green4"),
      pch = 19,
      legend = c("theorie", "approximation"))
```

Approximation SB



On observe que les deux estimateurs sont convergents, comme montré en question 5.

Question 8

Simuler $m = 10000$ échantillons de taille $n = 20$ d'une loi exponentielle de paramètre $\theta = 10$. Décrire l'histogramme de l'estimateur de maximum de vraisemblance. Quelle densité est-elle théoriquement attendue ?

Indiquer la moyenne de la loi par une barre verticale rouge et la valeur de θ par une barre verticale bleue.

Solution. Pour trouver la densité théoriquement attendue, on peut effectuer un changement de variable. Reporter la densité attendue.

Densité : Si on prend $Y = \sum X_i$, on a $\hat{\theta}_{MV} = \frac{n}{Y}$. Comme les X_i suivent une loi exponentielle $\exp(\theta)$, alors Y suit une loi gamma $\Gamma(n, \theta)$.

En posant $Z = \frac{Y}{n}$ et en calculant sa fonction de répartition et sa densité de probabilité, on démontre que Z suit une loi gamma: $\Gamma(n, n\theta)$.

Comme $\hat{\theta}_{MV} = \frac{n}{Y} = \frac{1}{Z}$ on arrive à démontrer que $\hat{\theta}_{MV}$ suit une loi inverse gamma : $IG(n, n\theta)$.

Pour la réponse numérique, créer les codes de calculs dans un bout de code (note: on peut simuler directement la loi de l'EMV, plutôt que de calculer l'estimateur pour les échantillons de loi exponentielle).

```
# Valeur theta
theta <- 10
# Taille échantillon
n <- 20
# Nombre échantillons
m <- 10000

# Estimateur de maximum de vraisemblance avec loi inverse gamma

# Simulation de m échantillons de taille n suivant une loi exponentielle
X <- matrix(rexp(n*m, rate = theta), ncol = m)

# estimateur de maximum de vraisemblance
emv <- n/apply(X, 2, FUN = sum)

# Histogramme empirique
hist(emv, breaks = 40, freq = FALSE,
     main = "Histogramme de l'estimateur MV de theta",
     xlab = "estimateur de MV",
     ylab = "densité",
     col = "lightblue", border = "white")

# paramètres de la loi inverse gamma
beta <- n * theta
alpha <- n

# densité théorique de loi inverse gamma(n,n*theta)
curve((beta^alpha / gamma(alpha)) * x^(-(alpha + 1)) * exp(-beta / x),
      col = "darkgreen", lwd = 2, add = TRUE)

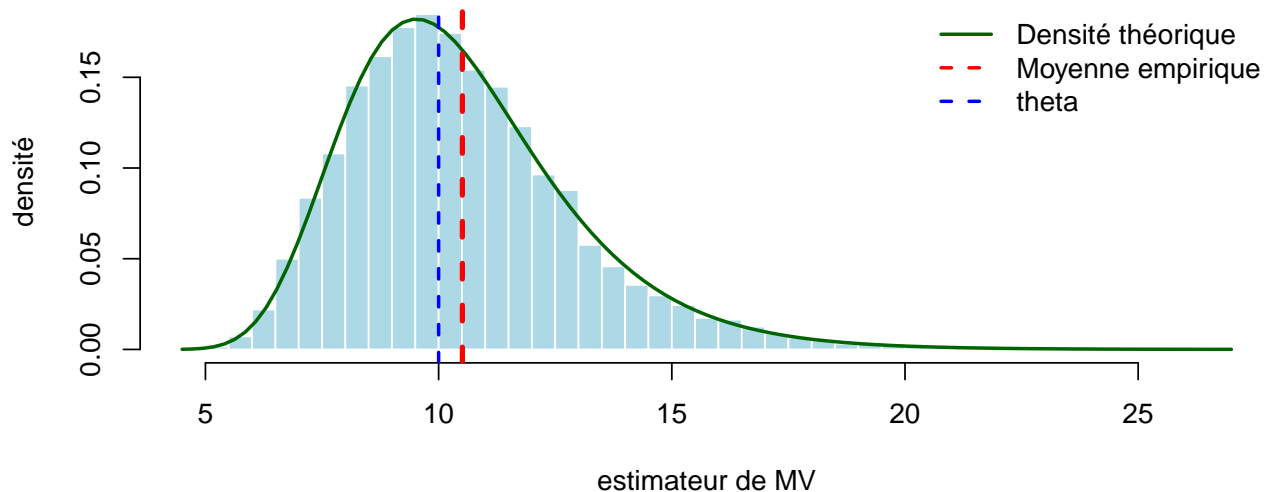
# Moyenne empirique
abline(v = mean(emv), col = "red", lwd = 3, lty = 2)
# Valeur de theta
abline(v = theta, col = "blue", lwd = 2, lty = 2)
```

```

legend("topright",
      legend = c("Densité théorique", "Moyenne empirique", "theta"),
      col = c("darkgreen", "red", "blue"),
      lty = c(1, 2, 2), lwd = 2, bty = "n")

```

Histogramme de l'estimateur MV de theta



L'estimateur de maximum de vraisemblance suit donc bien une loi inverse gamma de paramètres $(n, n\theta)$.

Question 9

Pour les données de l'énoncé,

```

x = c(0.03 , 0.29 , 0.04 , 0.04 , 0.48 , 0.41, 0.05,
      0.00 , 0.06, 0.05 , 0.18 , 0.01 , 0.06, 0.01, 0.09 , 0.02 , 0.08, 0.01, 0.26, 0.12)

```

a-t-on suffisamment de preuves pour affirmer que le paramètre inconnu θ est inférieur ou égal à $\theta_0 = 10$? Et pour affirmer qu'il est inférieur ou égal à $\theta_0 = 13$ (rappel : $\hat{\theta} \approx 8.3$) ?

Solution. Calculer la probabilité pour que la somme $S_n = \sum_{i=1}^n X_n$ soit supérieure à la valeur observée ($s_n = 2.29$) sous l'hypothèse que la vraie valeur est $\theta_0 = 10$, puis $\theta_0 = 13$. Examiner chacune des deux hypothèses : semblent-elles raisonnable ou déraisonnable ?

```

# nombre de données
n = length(x)
s_n <- sum(x)

# hypothèse testée theta0 = 10 ou 13
theta0 = 10

# pourcentage d'échantillons compatible avec des données observées
# pour l'hypothèse testée

cat("Probabilité que S_n > valeur observée avec theta = 10 est : ",
    ,round(pgamma(s_n, shape = n, rate = theta0, lower.tail = FALSE)*100),"% \n")

## Probabilité que S_n > valeur observée avec theta = 10 est : 24 %

cat("Probabilité que S_n > valeur observée avec theta = 13 est : ",
    ,round(pgamma(s_n, shape = n, rate = 13, lower.tail = FALSE),2)*100,"% \n")

```

Probabilité que $S_n >$ valeur observée avec $\theta = 13$ est : 2 %

La probabilité que S_n soit supérieure à la valeur observée est assez forte (24%) quand $\theta_0 = 10$ et faible (2%) quand $\theta_0 = 13$. Donc l'hypothèse que $\theta_0 = 10$ est plausible alors que celle que $\theta_0 = 13$ est peu plausible.