

# Statistique bayésienne avec R

Julien JACQUES

## Les données

On s'intéresse à la longueur de saumon juvénils le long d'une rivière de Bretagne (source [http://sirs.agrocampus-ouest.fr/bayes\\_V2/index.html](http://sirs.agrocampus-ouest.fr/bayes_V2/index.html)).

Sur cette rivière est installée une ferme d'élevage, et on se pose la question de l'impact de cette ferme sur les poissons dans le reste de la rivière.

```
ForkLengthData <- read.csv("~/Enseignement/Formations/Stat Bayesienne-v2/Rcode/ForkLengthData.csv",
                             sep=";")
ForkLengthData$Location=as.factor(ForkLengthData$Location)
levels(ForkLengthData$Location)=c("farm", "downstream", "other", "upstream")
boxplot(ForkLengthinmillimeters~Location, data=ForkLengthData)
```



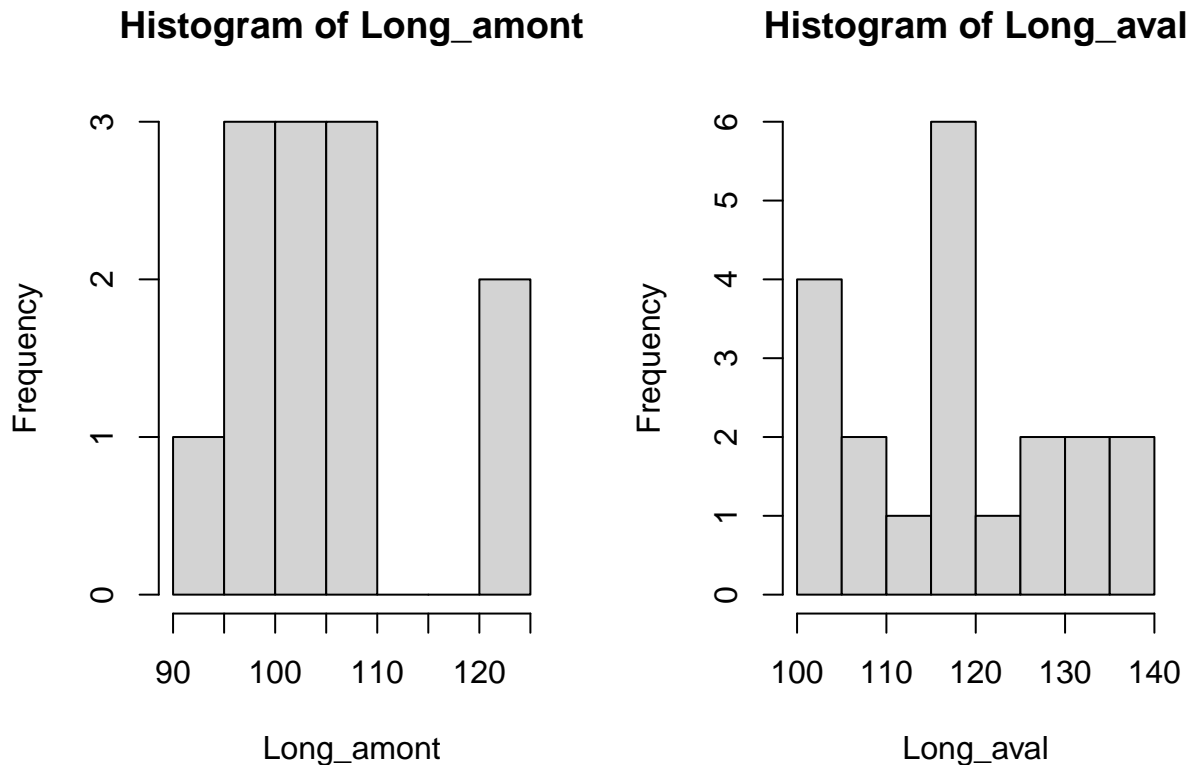
Visuellement, les poissons ont l'air plus grands en aval de la ferme qu'en amont.

## Approche fréquentiste

On peut confirmer cela à l'aide d'un test statistique classique :

```
Long_aval=ForkLengthData[ForkLengthData$Location=="downstream",2]
Long_amont=ForkLengthData[ForkLengthData$Location=="upstream",2]
```

```
par(mfrow=c(1,2))
hist(Long_aval);hist(Long_aval)
```



```
t.test(Long_aval,Long_aval,alternative="greater")
```

```
##
## Welch Two Sample t-test
##
## data: Long_aval and Long_aval
## t = 3.5407, df = 27.447, p-value = 0.0007233
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
## 6.689227 Inf
## sample estimates:
## mean of x mean of y
## 118.0500 105.1667
```

```
wilcox.test(Long_aval,Long_aval,alternative="greater")
```

```
## Warning in wilcox.test.default(Long_aval, Long_aval, alternative = "greater"):
## cannot compute exact p-value with ties
```

```
##
## Wilcoxon rank sum test with continuity correction
##
## data: Long_aval and Long_aval
## W = 197.5, p-value = 0.001341
## alternative hypothesis: true location shift is greater than 0
```

En effet, les poissons sont plus grands en aval de la ferme !

L'estimation de l'espérance du poids en aval  $\mu_d$  et en amont  $\mu_u$  est de façon fréquentiste donné par les

moyennes empiriques

```
tmp=t.test(Long_aval);
cat('Estimation de l espérance en aval : ',tmp$estimate,'\n')

## Estimation de l espérance en aval : 118.05
cat('intervalle de confiance : [',tmp$conf.int[1],',',tmp$conf.int[2],'] ','\n')

## intervalle de confiance : [ 112.7388 , 123.3612 ]
tmp=t.test(Long_amont);
cat('Estimation de l espérance en amont : ',tmp$estimate,'\n')

## Estimation de l espérance en amont : 105.1667
cat('intervalle de confiance : [',tmp$conf.int[1],',',tmp$conf.int[2],'] ','\n')

## intervalle de confiance : [ 99.42699 , 110.9063 ]
```

## Approche bayésienne

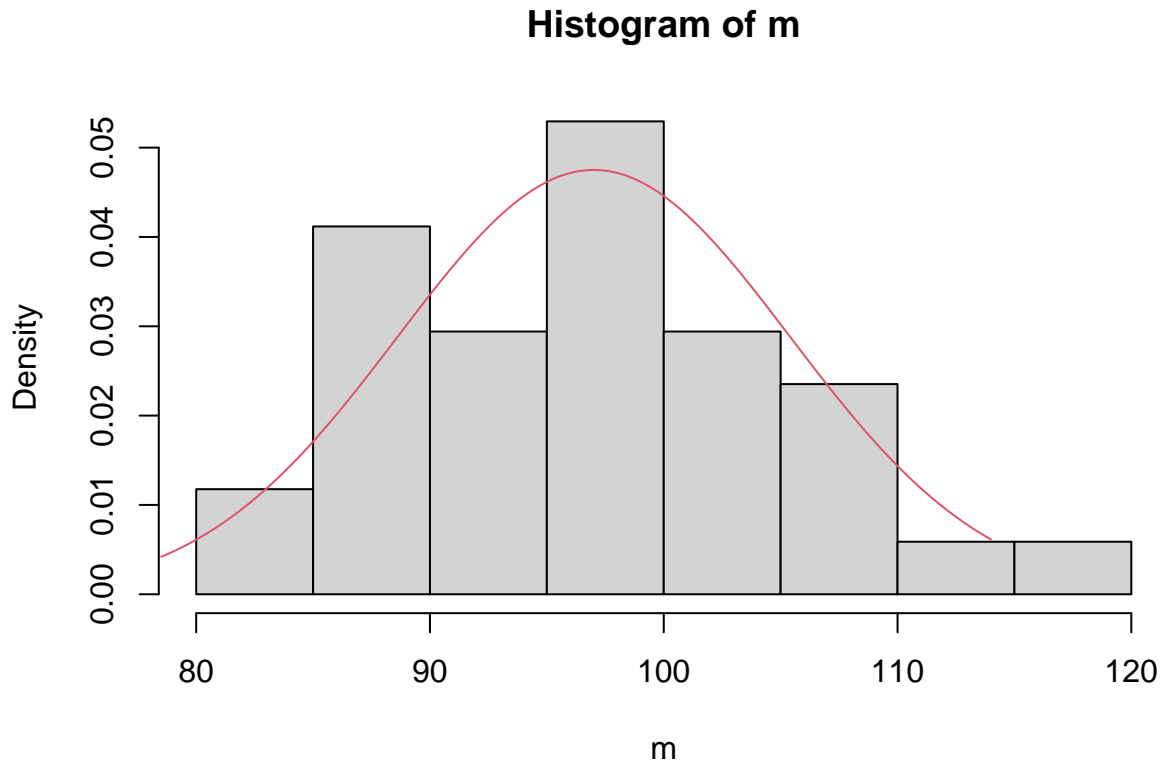
Intégrons de l'information supplémentaire à l'aide d'une estimation bayésienne; Nous supposons un a priori Gaussien sur  $\mu_d$  et  $\mu_u$  :  $\mathcal{N}(\mu, \sigma^2)$

Comment avoir une idée des hyperparamètres : nous pouvons utiliser les données des autres sites de mesure sur la rivière. Attention, on veut une idée de l'espérance de la longueur des poissons, non pas de la longueur elle-même.

```
extra=ForkLengthData[ForkLengthData$Location=="other",]
m=NULL
for (site in unique(extra$StationNumber)){
  m=c(m,mean(extra[extra$StationNumber==site,2]))
}
hist(m,probability = T)
mu=mean(m)
tau2=var(m)
cat('Moyenne des moyennes des tailles sur les autres sites :',mu,'\n')

## Moyenne des moyennes des tailles sur les autres sites : 97.0097
cat('Variance des moyennes des tailles sur les autres sites :',tau2,'\n')

## Variance des moyennes des tailles sur les autres sites : 70.53082
x=seq(min(m)-5,max(m)+5,0.01)
lines(x,dnorm(x,mu,sqrt(tau2)),col=2)
```



Comme la loi gaussienne est conjuguée pour elle même, si :

- $\underline{x} = (x_1, \dots, x_n)$  avec  $x_i \sim \mathcal{N}(\mu_d, \sigma^2)$
- $\mu_d \sim \mathcal{N}(\mu, \tau^2)$

alors la loi a posteriori de l'espérance en aval  $\mu_d$  est :

$$\mu_d | \underline{x} \sim \mathcal{N}\left(\frac{\tau^2 \bar{x} + \mu \frac{\sigma^2}{n}}{\tau^2 + \frac{\sigma^2}{n}}, \frac{\tau^2 \frac{\sigma^2}{n}}{\tau^2 + \frac{\sigma^2}{n}}\right)$$

A noter que l'on a supposé la variance identique en aval et en amont, ce qui signifie que les conditions aléatoire sont les mêmes, la présence de la ferme influençant éventuellement uniquement sur l'espérance.

L'espérance et la variance de la loi gaussienne a posteriori est donc :

```
nd=length(Long_aval)
mu_dpost=(mean(Long_aval)*tau2 + mu*var(Long_aval)/nd)/(tau2 + var(Long_aval)/nd)
sigma_dpost=(tau2 * var(Long_aval)/nd)/(tau2 + var(Long_aval)/nd)
cat('Espérance a posteriori, aval : ', mu_dpost, '\n')
```

```
## Espérance a posteriori, aval : 116.2898
```

```
cat('Variance a posteriori, aval : ', sigma_dpost, '\n')
```

```
## Variance a posteriori, aval : 5.900625
```

De même pour l'amont

```
nu=length(Long_amont)
mu_upost=(mean(Long_amont)*tau2 + mu*var(Long_amont)/nu)/(tau2 + var(Long_amont)/nu)
sigma_upost=(tau2 * var(Long_amont)/nd)/(tau2 + var(Long_amont)/nu)
cat('Espérance a posteriori, amont : ', mu_upost, '\n')
```

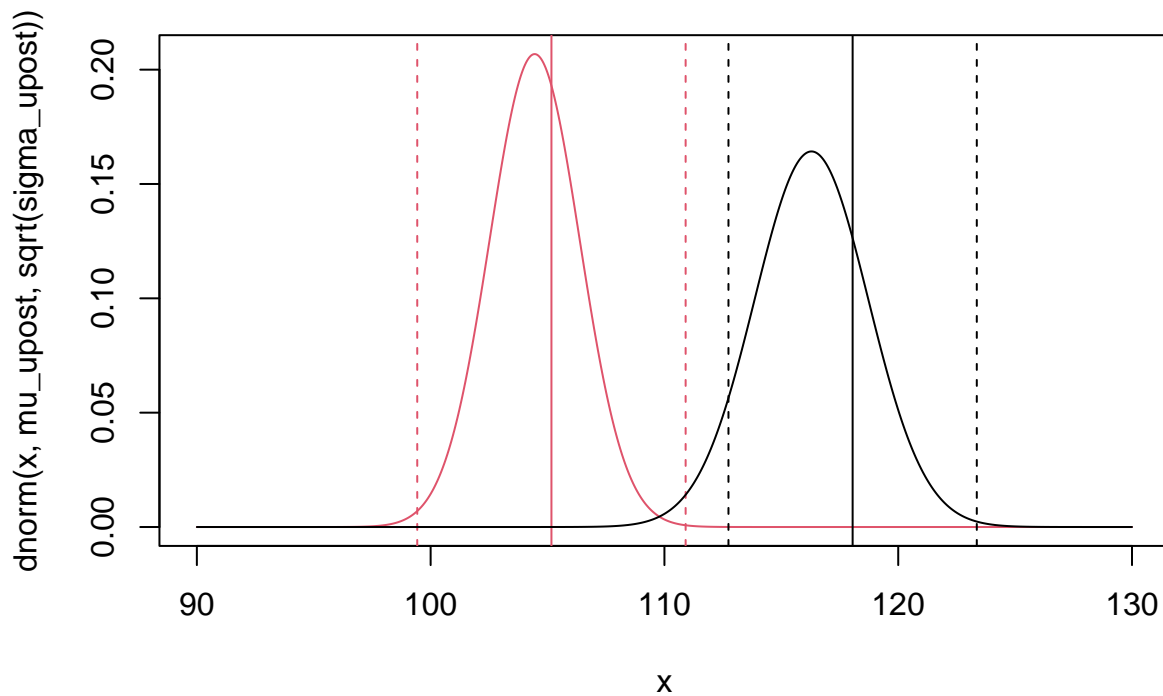
```
## Espérance a posteriori, amont : 104.4493
```

```
cat('Variance a posteriori, montant :', sigma_upost, '\n')
```

```
## Variance a posteriori, montant : 3.721482
```

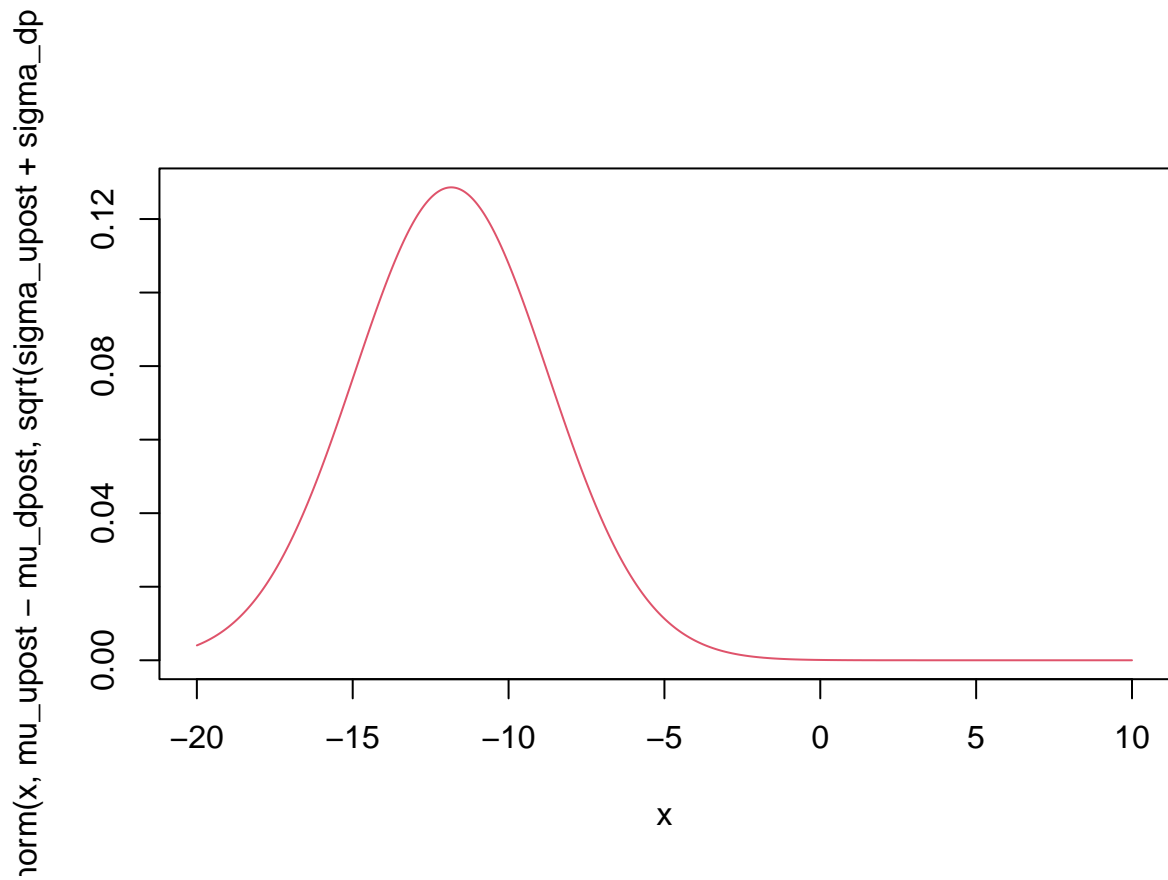
On peut alors représenter l'estimation fréquentiste (moyenne et intervalle de confiance) et bayésienne (loi a posteriori)

```
x=seq(90,130,0.01)
plot(x,dnorm(x,mu_upost,sqrt(sigma_upost)),type='l',col=2)
abline(v=mean(Long_amont),col=2)
tmp=t.test(Long_amont);
abline(v=tmp$conf.int[1],col=2,lty=2)
abline(v=tmp$conf.int[2],col=2,lty=2)
lines(x,dnorm(x,mu_dpost,sqrt(sigma_dpost)),type='l')
abline(v=mean(Long_aval))
tmp=t.test(Long_aval);
abline(v=tmp$conf.int[1],lty=2)
abline(v=tmp$conf.int[2],lty=2)
```



Enfin, on peut également analyser la loi a posteriori de la différence de longueur amont/aval  $\mu_u - \mu_d$ . Sous l'hypothèse d'indépendance, l'espérance est la différence et la variance la somme :

```
x=seq(-20,10,0.01)
plot(x,dnorm(x,mu_upost-mu_dpost,sqrt(sigma_upost+sigma_dpost)),type='l',col=2)
```



La probabilité que la taille des poissons en aval soit plus grande d'au moins 1mm qu'en amont est :

```
pnorm(-1,mu_upost-mu_dpost,sqrt(sigma_upost+sigma_dpost))
```

```
## [1] 0.9997627
```

Rq : on avait supposé que les variances étaient les mêmes, mais on n'a pas utilisé cette information....