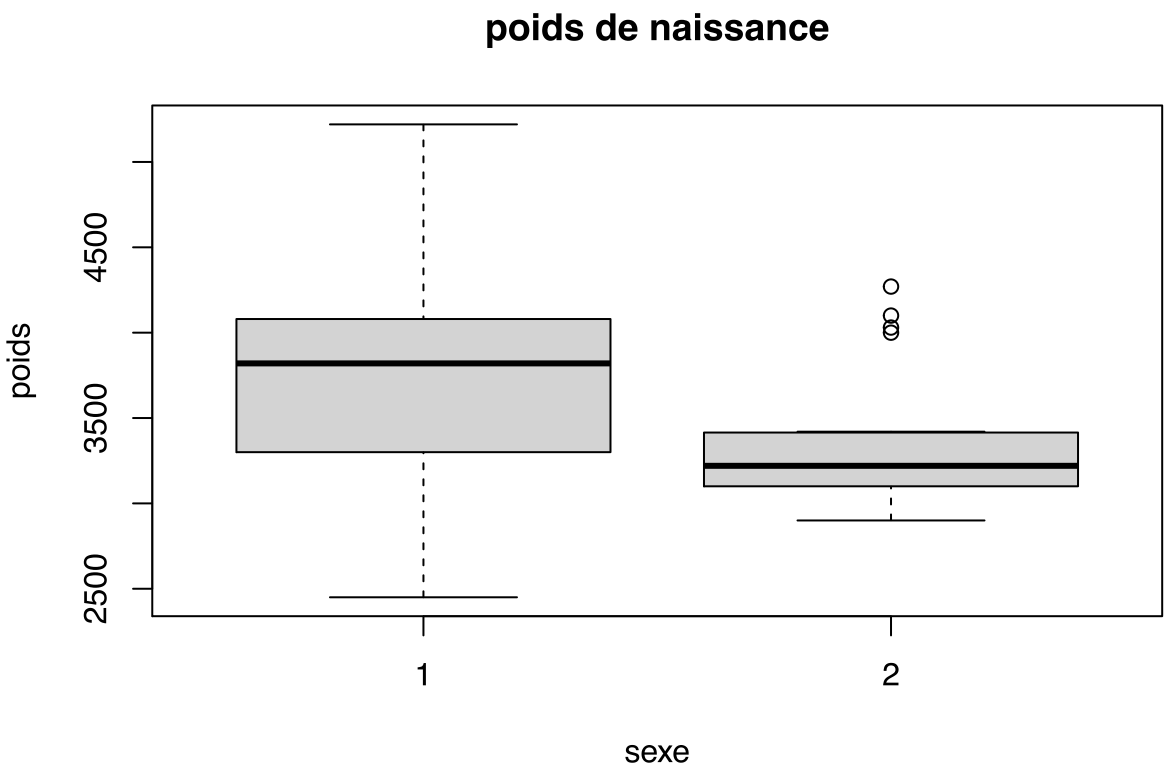
Statistique bayésienne avec R

Exercice sur les poids de naissance

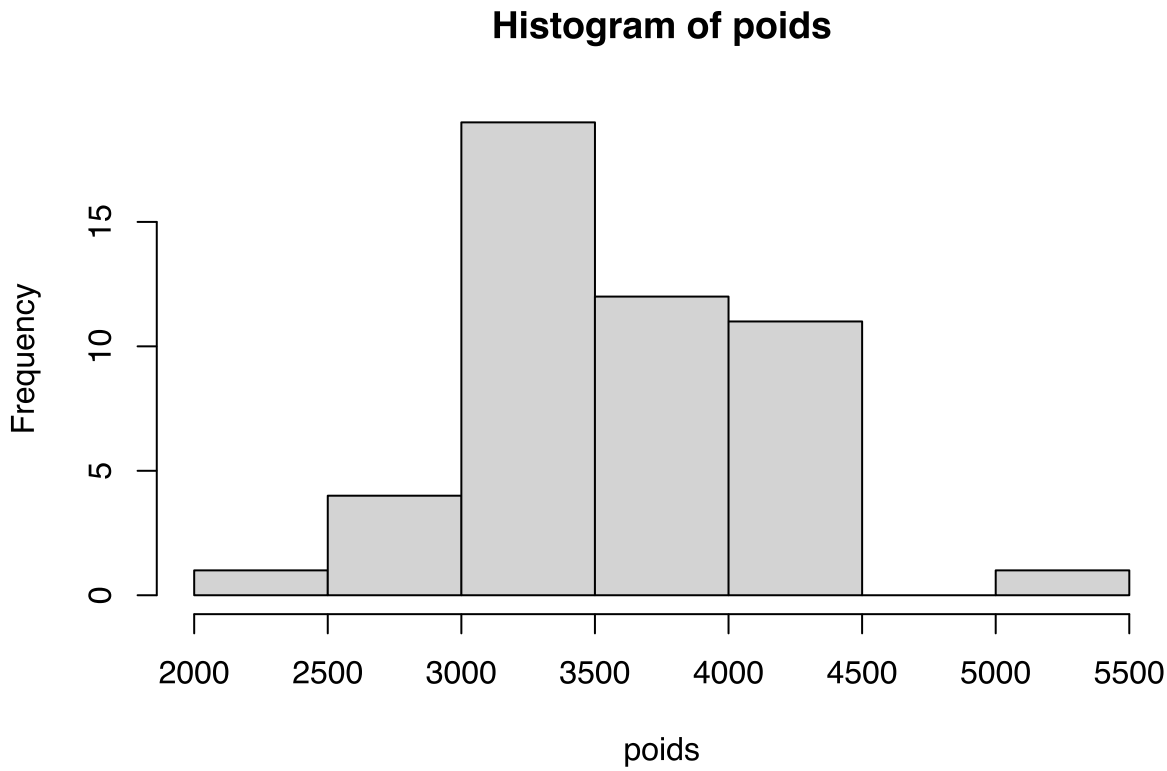
Julien JACQUES

```
data=read.table('Rcode/poidsnaissance.txt',header = T,sep=',',row.names = 1)
data$OBS=NULL
sexe=data$SEXE+1
poids=data$POIDNAIS
boxplot(poids~sexe,main="poids de naissance")
```



L'histogramme du poids de naissance ressemble à peu près à une loi gaussienne, ce qui est confirmé par le test de Shapiro

```
hist(poids)
```




```
shapiro.test(poids)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  poids
## W = 0.96218, p-value = 0.1237
```

Estimation bayésienne

Estimer le poids de naissance moyen

De façon fréquentiste :

```
mean(poids)
```

```
## [1] 3590
```

Les données étant distribuées suivant une loi gaussienne, nous allons choisir un a priori conjugué gaussien. Il nous reste à déterminer les moyennes et variances a priori.

L'histogramme nous donne l'idée d'un a priori gaussien centré en 3250g. Pour l'écart-type, il va traduire la confiance que l'on a dans notre a priori.

Nous avons vu que le MAP est alors donné par :

$$\hat{\theta} = E[\theta|\underline{x}] \quad = \quad \frac{\tau^2 \frac{\sigma^2}{n}}{\tau^2 + \frac{\sigma^2}{n}} \left(\frac{\bar{x}}{\frac{\sigma^2}{n}} + \frac{\mu}{\tau^2} \right)$$

dont on peut coder le calcul. En jouant sur la valeur de l'ecart-type a priori τ, on pourra jouer sur la confiance en notre a priori et examiner son influence sur le MAP

```
s=sd(poids)
tau=100
n=length(poids)
MAP=(tau^2*s^2/n)/(tau^2+s^2/n)*(mean(poids)/(s^2/n)+3250/(tau^2))
print(MAP)
```

```
## [1] 3461.82
```

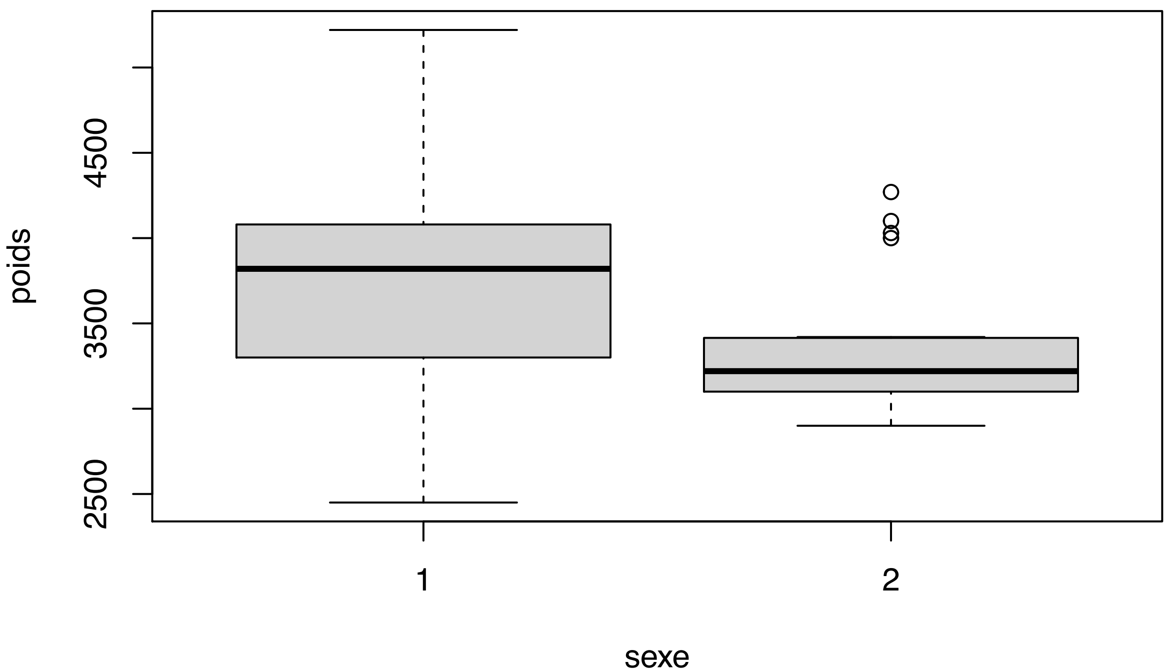
Les calculs sont faits ici *à la main*, nous verrons plus tard comment les faire sous R.

Test d'hypothèse et comparaison bayésienne de modèles

les garçons sont-ils plus lourds que les filles à la naissance ?

De façon fréquentiste c'est significatif,

```
boxplot(poids~sexe)
```




```
t.test(poids~sexe,alternative='greater')
```

```
##
##  Welch Two Sample t-test
##
## data:  poids by sexe
## t = 2.4522, df = 45.553, p-value = 0.009047
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  110.0071      Inf
## sample estimates:
## mean in group 1 mean in group 2
##        3728.103        3379.211
```

Pour la version bayesienne, on indiquera l'unilatéralité de l'hypothèse alternative en specifiant `nullInterval=c(0, Inf)`

```r
library(BayesFactor)
ttestBF(poids[sexe==1],poids[sexe==2],nullInterval=c(0, Inf))
```

```
## Bayes factor analysis
## --------------
## [1] Alt., r=0.707 0<d<Inf    : 4.525689  ±0%
## [2] Alt., r=0.707 !(0<d<Inf) : 0.1017375 ±0%
##
## Against denominator:
##   Null, mu1-mu2 = 0
## ---
## Bayes factor type: BFindepSample, JZS
```

Le facteur de Bayes faut 4.525689

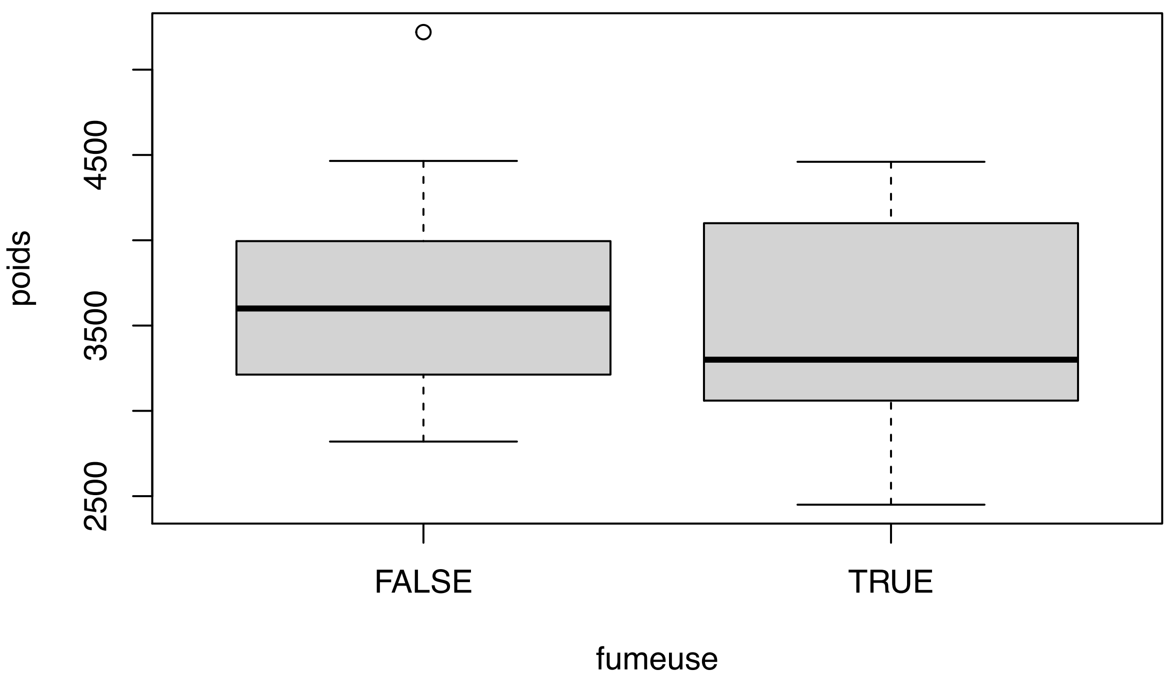
```r
log10(4.525689)
```

```
## [1] 0.6556847
```

De façon bayesienne, c'est *substantielle* (!), il est difficile de conclure à la supériorité de poids des garçons par rapport aux filles.

le poids de naissance dépend-il du fait que la mère soit fumeuse ?

```r
fumeuse=data$CIGJOUR>0
boxplot(poids~fumeuse)
```




```r
t.test(poids~fumeuse,alternative='greater')
```

```
##
##  Welch Two Sample t-test
##
## data:  poids by fumeuse
## t = 0.53947, df = 10.414, p-value = 0.3005
```

```
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  -295.1788       Inf
## sample estimates:
## mean in group FALSE  mean in group TRUE
##            3613.590            3487.778
```

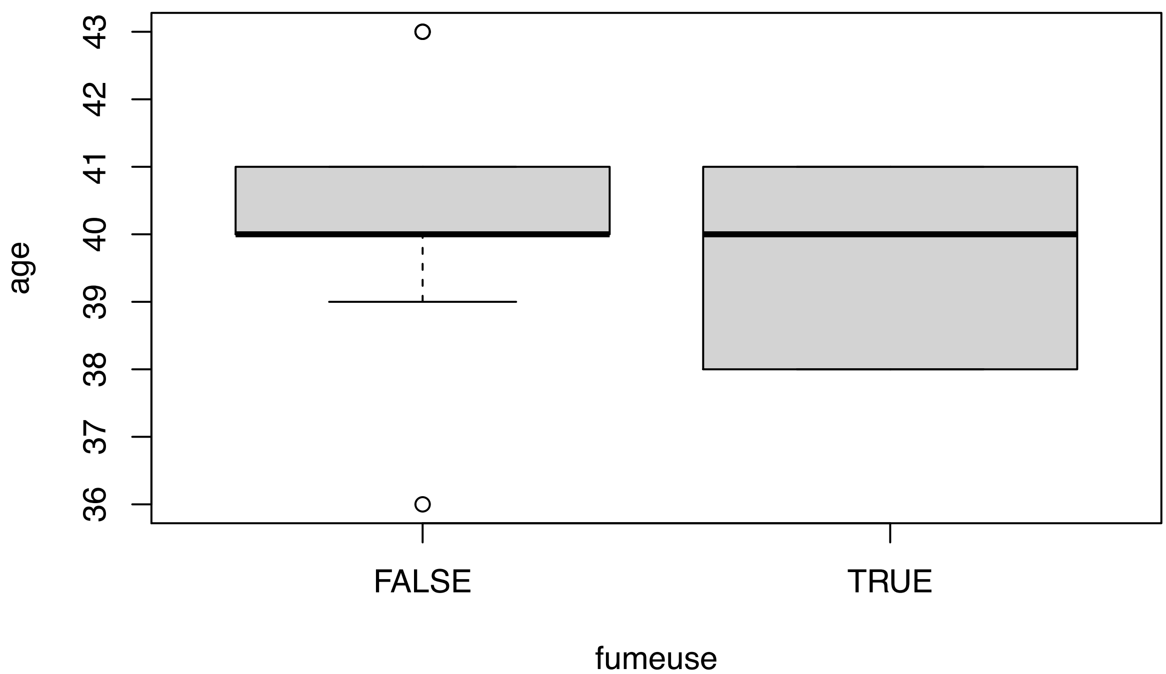
```
ttestBF(poids[fumeuse],poids[!fumeuse],nullInterval=c(-Inf,0))
```

```
## Bayes factor analysis
## --------------
## [1] Alt., r=0.707 -Inf<d<0    : 0.5668229 ±0.02%
## [2] Alt., r=0.707 !(-Inf<d<0) : 0.2385683 ±0%
##
## Against denominator:
##   Null, mu1-mu2 = 0
## ---
## Bayes factor type: BFindepSample, JZS
```

Non significatif.

l'âge gestationnel dépend-il du fait que la mère soit fumeuse ?

```
age=data$AGEGEST
boxplot(age~fumeuse)
```




```
t.test(age~fumeuse,alternative='greater')
```

```
##
##  Welch Two Sample t-test
##
## data:  age by fumeuse
## t = 0.96285, df = 11.094, p-value = 0.1781
## alternative hypothesis: true difference in means is greater than 0
```

```
## 95 percent confidence interval:
##  -0.3986471        Inf
## sample estimates:
## mean in group FALSE  mean in group TRUE
##            40.12821            39.66667
```

```r
ttestBF(age[fumeuse],age[!fumeuse],nullInterval=c(-Inf,0))
```

```
## Bayes factor analysis
## --------------
## [1] Alt., r=0.707 -Inf<d<0    : 0.84529   ±0%
## [2] Alt., r=0.707 !(-Inf<d<0) : 0.1953459 ±0%
##
## Against denominator:
##   Null, mu1-mu2 = 0
## ---
## Bayes factor type: BFindepSample, JZS
```

Non significatif.

Régression linéaire

Effectuer une régression du poids de naissance en fonction des autres variables disponibles

De façon fréquentiste

```r
m1=lm(POIDNAIS~.,data=data)
summary(m1)
```

```
##
## Call:
## lm(formula = POIDNAIS ~ ., data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1039.65  -192.17   -33.47   187.49  1277.90
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  621.107   3050.952   0.204  0.83969
## AGEGEST      164.131     57.489   2.855  0.00672 **
## SEXE        -422.069    145.419  -2.902  0.00593 **
## CIGJOUR       -2.552      5.493  -0.465  0.64472
## TAILMERE     -28.560     14.524  -1.966  0.05605 .
## POIDAVG      -17.229     17.021  -1.012  0.31737
## POIDFING      30.330     13.433   2.258  0.02934 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 459.7 on 41 degrees of freedom
## Multiple R-squared:  0.3654, Adjusted R-squared:  0.2725
## F-statistic: 3.935 on 6 and 41 DF,  p-value: 0.003388
```

```r
m2=step(m1)
```