Statistique bayésienne

Devoir maison obligatoire (Dauphine)

Paul Hardouin

February 26, 2020

Contents

Les enseignants des collèges et lycées français souhaitant obtenir une mutation professionnelle sont classés en fonction d'un nombre de points qui dépend de leur situation personnelle et de leur carrière. Le fichier **mutations2.csv** donne le nombre de points nécessaire pour obtenir une mutation dans les lycées de l'académie de Versailles en 2012, pour diverses disciplines enseignées ; c'est une mesure de l'attractivité de chaque établissement pour les enseignants. Par exemple, en mathématiques, il suffisait de 21 points pour pouvoir être nommé au lycée Georges Braque d'Argenteuil, mais il en fallait 464 pour être nommé au lycée Michelet de Vanves. Nous allons étudier ce nombre de points, dans un cadre bayésien.

Pour des couples (établissement, discipline), on dispose du nombre de points nécessaire (colonne **Barre**) pour obtenir une mutation, ainsi que de caractéristiques de l'établissement : nombre de candidats au baccalauréat par série, taux de réussite au baccalauréat par série, taux de réussite attendu (qui dépend notamment du tissu socioprofessionnel des parents d'élèves), taux d'accès des élèves de seconde et de première au baccalauréat. Par souci d'homogénéité des données, on considère uniquement les filières du lycée général, même si beaucoup des établissements concernés préparent aussi au baccalauréat technologique et parfois au baccalauréat professionnel.

1. Régression linéaire

On propose d'abord un modèle linéaire gaussien. On cherche à expliquer le nombre de points nécessaire à une mutation (colonne **Barre**) par les caractéristiques du lycée.

1.1. Régression linéaire bayésienne

==========
Effectuer une régression linéaire bayésienne et interpréter les coefficients obtenus.
==========

En premier lieu, on charge les librairies et les données.

On fait ensuite de l'inférence bayésienne avec la loi a priori g de Zellner. Comme nous n'avons pas de connaissance particulière de ce type de données, on choisit un a priori assez léger en prenant g = n (nombre d'individus).

```
# Mise en forme des données
  Y = data[, 6]
  X = cbind(1,as.matrix(data[, 7:23]))
  n = length(Y)
# Calcul de Beta Hat (maximum de vraisemblance)
  betaHat = solve(t(X)%*%X)%*%t(X)%*%Y
# Calcul de S2 (maximum de vraisemblance)
  s2 = t(Y-X%*%betaHat)%*%(Y-X%*%betaHat)
# A priori de Zellner
  g=n
# Espérance a posteriori de beta
  betaPost = betaHat * g / (g + 1)
# Espérance a posteriori de sigma^2 (inverse gamma)
  a = n/2
  b = s2/2 + 1/(2*g+2) * ((t(betaHat)%*%t(X))%*%(X%*%betaHat))
  sigma2Post = b / (a-1)
```

A partir des estimations obtenues, on peut estimer des intervalles de crédibilité. On effectue aussi une simulation pour pourvoir afficher les densités de distributions des coefficients significatifs.

```
# Simulation
  variance = as.integer(sigma2Post) * g / (g + 1) * solve(t(X)%*%X)
  simul = rmvnorm(1e4, mean = betaPost, sigma = variance)
# Intervalles
  semiRange = 1.96*sqrt(diag(variance))
  lowRange = betaPost - semiRange
  highRange = betaPost + semiRange
# Aggregation resultats
  D = data.frame(low_95 = lowRange, beta = betaPost, high_95 = highRange)
  setattr(D, "row.names", c("intercept",row.names(D)[2:dim(D)[1]]))
```

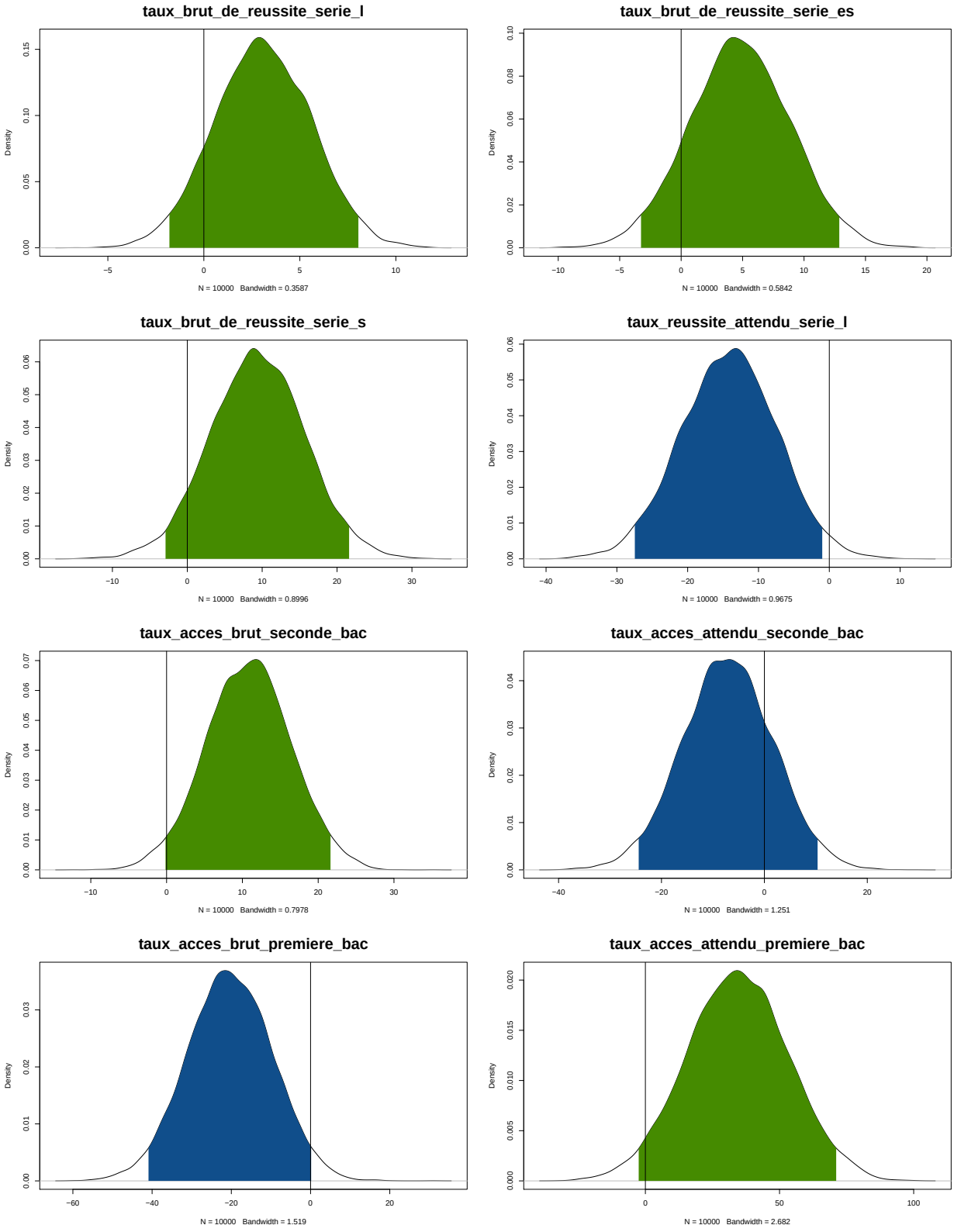
En observant les estimations des coefficients et leurs intervalles de crédibilité, on peut identifier des variables qui contribuent **positivement** à la **barre** cible et d'autres qui contribuent **négativement**. C'est-à-dire des coefficients estimés nettement différents de 0, avec une distribution crédible à 95% du même signe (ou quasiment).

	low_95	beta	high_95
intercept	**-1549.0**	**-471.6**	**605.8**
effectif_presents_serie_l	-2.4	0.8	3.9
effectif_presents_serie_es	-2.1	0.3	2.7
effectif_presents_serie_s	-2.0	0.0	2.0
taux_brut_de_reussite_serie_l	**-1.8**	**3.1**	**8.1**
taux_brut_de_reussite_serie_es	**-3.3**	**4.8**	**12.9**
taux_brut_de_reussite_serie_s	**-2.9**	**9.4**	**21.7**
taux_reussite_attendu_serie_l	**-27.5**	**-14.3**	**-1.0**
taux_reussite_attendu_serie_es	-12.1	3.8	19.7
taux_reussite_attendu_serie_s	-22.8	-4.3	14.2
effectif_de_seconde	-1.2	0.0	1.2
effectif_de_premiere	-1.7	-0.4	1.0
taux_acces_brut_seconde_bac	**-0.2**	**10.7**	**21.6**
taux_acces_attendu_seconde_bac	**-24.5**	**-7.1**	**10.4**
taux_acces_brut_premiere_bac	**-41.0**	**-20.3**	**0.3**
taux_acces_attendu_premiere_bac	**-2.6**	**34.4**	**71.3**
taux_brut_de_reussite_total_series	-30.2	-5.4	19.5
taux_reussite_attendu_total_series	-46.5	-4.1	38.4

On remarque en particulier la très faible contribution des variables suivantes :

- **effectif_presents_serie_l**
- **effectif_presents_serie_es**
- **effectif_presents_serie_s**
- **effectif_de_seconde**
- **effectif_de_première**.

Ci-dessous, quelques densités calculées à partir des tirages de la loi a posteriori, pour illustrer ces interprétations.



1.2. Covariables significatives + comparaison à la méthode fréquentiste

==========
Choisir les covariables significatives. Comparer au résultat obtenu par une analyse fréquentiste.
Afin de réduire le coût computationnel, il peut être intéressant d'effectuer une présélection des covariables considérées.
==========

Pour effectuer ce choix, on va mettre en oeuvre un échantillonneur de Gibbs. Pour cela, on reprend les fonctions vues pendant la formation.

```
# fonction pour calculer la log-vraisemblance marginale
##########################################################
marglkd = function(gamma, X, Y, n, g){
  q=sum(gamma)
  X1=X[ ,c(T,gamma)]
  if(q==0)
    {m = -n/2 * log(t(Y)%*%Y)}
  else
    {m = -q/2*log(g+1) -n/2*log(t(Y)%*%Y
         -g/(g+1)*t(Y)%*%X1%*%solve(t(X1)%*%X1)%*%t(X1)%*%Y)}
  return(m)
}
```

```
# Echantilloneur de gibbs
##########################################################
echGibbs = function(niter, X, Y, n, g){
  # initialisation
    nC = dim(X)[2]
    gamma = matrix(F, nrow = niter, ncol = nC-1)
    gamma0 = sample(c(T, F), size = nC-1, replace = TRUE) # valeur initiale aléatoire
    lkd = rep(0, niter)
    modelnumber = rep(0, niter)
  # boucle
    oldgamma = gamma0
    for(i in 1:niter){
      newgamma = oldgamma
      for(j in 1:nC-1){
        g1 = newgamma; g1[j]=TRUE
        g2 = newgamma; g2[j]=FALSE
        ml1 = marglkd(g1, X, Y, n, g)
        ml2 = marglkd(g2, X, Y, n, g)
        p = c(ml1,ml2)-min(ml1,ml2)
        newgamma[j] = sample(c(T,F), size=1, prob=exp(p))
      }
      gamma[i,] = newgamma
      lkd[i] = marglkd(newgamma, X, Y, n, g)
      modelnumber[i] = sum(newgamma*2^(0:(nC-2)))
      oldgamma = newgamma
    }
  # output
    return(list(gamma=gamma, lkd=lkd, modelnumber=modelnumber))
}
```

On met en place l'experience avec 10.000 iterations.

```
# echantillonnage
  niter = 10000
  resGibbs = echGibbs(niter, X, Y, n, g)
  gamma = resGibbs$gamma
  modelnumber = resGibbs$modelnumber
```

Pour vérifier la convergence, on lisse pour chaque covariable les valeurs booléennes obtenues, afin d'obtenir une estimation visuelle de la probabilité de la covariable. On obtient des courbes qui se stabilisent assez vite, avant 500 itérations, valeur que l'on prend pour le burn-in. Même on constate d'importantes oscillations autour de la valeur cible. Les courbes ne sont pas affichées dans ce rapport pour une meilleure lisibilité.

```
# controle de convergence
  library(zoo)
  for(i in 1:(dim(X)[2]-1)) plot(rollapply(gamma[,i], width=500, FUN=mean), type="l")
```

On affiche ci-dessous les modèles les plus probables pour notre étude. On s'aperçoit que les modèles les plus probables sont très parcimonieux. Ainsi, le modèle retenu n'a qu'un seul coefficient non nul, **taux_acces_attendu_premiere_bac**.

	M1:12.7%	M2: 8.8%	M3: 4.5%	M4: 4.4%	M5: 3.4%	Fréquence
effectif_presents_serie_l	.	.	.	.	.	0.04
effectif_presents_serie_es	.	.	.	.	.	0.05
effectif_presents_serie_s	.	.	.	.	.	0.05
taux_brut_de_reussite_serie_l	.	.	.	.	.	0.05
taux_brut_de_reussite_serie_es	.	.	.	.	.	0.09
taux_brut_de_reussite_serie_s	.	.	.	.	.	0.08
taux_reussite_attendu_serie_l	.	.	.	.	.	0.12
taux_reussite_attendu_serie_es	.	.	**TRUE**	.	.	**0.12**
taux_reussite_attendu_serie_s	.	.	.	.	**TRUE**	**0.10**
effectif_de_seconde	.	.	.	.	.	0.04
effectif_de_premiere	.	.	.	.	.	0.05
taux_acces_brut_seconde_bac	.	.	.	.	.	0.09
taux_acces_attendu_seconde_bac	.	**TRUE**	.	.	.	**0.20**
taux_acces_brut_premiere_bac	.	.	.	.	.	0.09
taux_acces_attendu_premiere_bac	**TRUE**	.	.	.	.	**0.33**
taux_brut_de_reussite_total_series	.	.	.	**TRUE**	.	**0.11**
taux_reussite_attendu_total_series	.	.	.	.	.	0.11

On compare ce résultat avec une approche fréquentiste. On fait une recherche un modèle avec un critère BIC, et on retrouve la même conclusion, à savoir un modèle avec un seul coefficient non nul, **taux_acces_attendu_premiere_bac**.

```
##
## Call:
## lm(formula = Barre ~ taux_acces_attendu_premiere_bac, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -380.36 -199.53 -129.26  -25.62 1686.87
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    -503.790    261.114  -1.929  0.05423 .
## taux_acces_attendu_premiere_bac    9.808      3.094   3.170  0.00161 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 420.4 on 514 degrees of freedom
## Multiple R-squared:  0.01918,    Adjusted R-squared:  0.01727
## F-statistic: 10.05 on 1 and 514 DF,  p-value: 0.001614
```

1.3. Mutations en mathématiques et en anglais

==========

On se concentre maintenant uniquement sur les mutations en mathématiques et en anglais. Répéter l'analyse pour chacune de ces deux catégories. Que penser de l'hypothèse que les covariables agissent de la même manière dans ces deux disciplines ?

==========

On répéte l'analyse pour chacune de ces deux catégories, comme détaillé ci-après. Dans les 2 cas, les approches fréquentistes et bayésiennes sont cohérentes. En revanche, les modèles résultants sont nettement différents. Cela tend à réfuter l'hypothèse que les covariables agissent de la même manière dans ces deux disciplines MATHS et ANGLAIS.

1.3.1. MATHS

On travaille dans cette partie sur les infdividus **data$Matiere=="MATHS"**. L'approche bayésienne et l'échantillonneur de Gibbs nous mène à nouveau vers un modèle très réduit avec pour seul coefficient non nul celui de **taux_brut_de_reussite_serie_es**. Ce résultat est à prendre avec prudence, car le modèle retenu a somme toute une probabilité relativement faible de **3.1%**. En revanche, si l'on regarde les fréquences d'apparition, 2 covariables sortent du lot: **taux_brut_de_reussite_serie_l** et **taux_brut_de_reussite_serie_es**.

	M1: 2.6%	M2: 1.9%	M3: 1.5%	M4: 0.8%	M5: 0.8%	Fréquence
effectif_presents_serie_l	.	.	.	.	.	0.13
effectif_presents_serie_es	.	.	.	.	.	0.13
effectif_presents_serie_s	.	.	.	.	.	0.22
taux_brut_de_reussite_serie_l	.	**TRUE**	**TRUE**	**TRUE**	.	**0.52**
taux_brut_de_reussite_serie_es	**TRUE**	**TRUE**	.	**TRUE**	.	**0.52**
taux_brut_de_reussite_serie_s	.	.	.	.	.	0.13
taux_reussite_attendu_serie_l	.	.	.	.	.	0.13
taux_reussite_attendu_serie_es	.	.	.	.	.	0.15
taux_reussite_attendu_serie_s	.	.	.	.	.	0.15
effectif_de_seconde	.	.	.	.	.	0.17
effectif_de_premiere	.	.	.	.	.	0.23
taux_acces_brut_seconde_bac	.	.	.	.	.	0.16
taux_acces_attendu_seconde_bac	.	.	.	.	.	0.14
taux_acces_brut_premiere_bac	.	.	**TRUE**	.	.	**0.23**
taux_acces_attendu_premiere_bac	.	.	.	**TRUE**	**TRUE**	**0.26**
taux_brut_de_reussite_total_series	.	.	.	.	.	0.19
taux_reussite_attendu_total_series	.	.	.	.	.	0.16

L'approche fréquentiste, avec un choix de modèle selon le critère BIC, confirme le modèle trouvé avec l'approche bayésienne.

```
##
## Call:
## lm(formula = Barre ~ taux_brut_de_reussite_serie_es, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -145.97  -80.44  -34.60   60.41  375.63
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   -286.357    148.068  -1.934  0.05809 .
## taux_brut_de_reussite_serie_es    5.433      1.688   3.220  0.00212 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 115.6 on 57 degrees of freedom
## Multiple R-squared:  0.1539, Adjusted R-squared:  0.139
## F-statistic: 10.37 on 1 and 57 DF,  p-value: 0.00212
```

1.3.2. ANGLAIS

On travaille dans cette partie sur les infdividus **data$Matiere=="ANGLAIS"**. L'approche bayésienne et l'échantillonneur de Gibbs nous mène à nouveau vers un modèle très réduit avec deux coefficients non nuls, **taux_brut_de_reussite_serie_es** et **taux_reussite_attendu_serie_l**. Ce résultat est à prendre avec prudence, car le modèle retenu a somme toute une probabilité relativement faible de **1.1%**. On peut même se demander si un choix de modèle pertinent se dégage réellement de cette étude, tellement les probabilités sont faibles pour les modèles mis en avant par l'échantilloneur de Gibbs.

	M1: 1.1%	M2: 0.9%	M3: 0.8%	M4: 0.7%	M5: 0.6%	Fréquence
effectif_presents_serie_l	**.**	**.**	**.**	**.**	**TRUE**	**0.18**
effectif_presents_serie_es	.	.	.	.	.	0.15
effectif_presents_serie_s	.	.	.	.	.	0.17
taux_brut_de_reussite_serie_l	.	.	.	.	.	0.20
taux_brut_de_reussite_serie_es	**TRUE**	**TRUE**	**TRUE**	**.**	**.**	**0.57**
taux_brut_de_reussite_serie_s	.	.	.	.	.	0.16
taux_reussite_attendu_serie_l	**TRUE**	**.**	**.**	**TRUE**	**.**	**0.25**
taux_reussite_attendu_serie_es	.	.	.	.	.	0.17
taux_reussite_attendu_serie_s	**.**	**.**	**TRUE**	**.**	**.**	**0.31**
effectif_de_seconde	.	.	.	.	.	0.16
effectif_de_premiere	.	.	.	.	.	0.21
taux_acces_brut_seconde_bac	.	.	.	.	.	0.18
taux_acces_attendu_seconde_bac	.	.	.	.	.	0.17
taux_acces_brut_premiere_bac	.	.	.	.	.	0.17
taux_acces_attendu_premiere_bac	.	.	.	.	.	0.31
taux_brut_de_reussite_total_series	.	.	.	.	.	0.16
taux_reussite_attendu_total_series	.	.	.	.	.	0.21

L'approche fréquentiste, avec un choix de modèle selon le critère BIC, conclue sur des coefficients **tous nuls**, ce qui confirme les doutes quant-aux résultats de l'approche bayésienne.

```
##
## Call:
## lm(formula = Barre ~ 1, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -205.48 -115.48  -68.48    3.52 1491.52
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   226.48      40.94   5.532  1.1e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 295.2 on 51 degrees of freedom
```

2. Loi de Pareto

On ignore maintenant les covariables, et on s'intéresse uniquement à la loi du nombre de points nécessaire (colonne **Barre**). La loi gaussienne peut paraitre peu pertinente pour ces données : on va plutôt proposer une loi de Pareto. Pour $m > 0$ et $\alpha > 0$, on dit que $Z \sim Pareto(m, \alpha)$ si Z est à valeur dans $[m, +\infty[$ de densité

$$f_Z(z; m, \alpha) = \alpha \frac{m^\alpha}{z^{\alpha+1}} \mathbb{I}_{\{z \geq m\}}$$

On impose $m = 21$ au vue des données. En effet, la valeur minimale de **Barre** étant 21, il nous faut prendre une valeur de $m \geq 21$ pour pouvoir manipuler des vraisemblances non nulles. Aussi, pour toute la suite, on omettra de considérer $\mathbb{I}_{\{z \geq m\}}$ dans nos raisonnement, sachant que cette valeur sera toujours égale à 1.

2.4. Générer des réalisations de Pareto

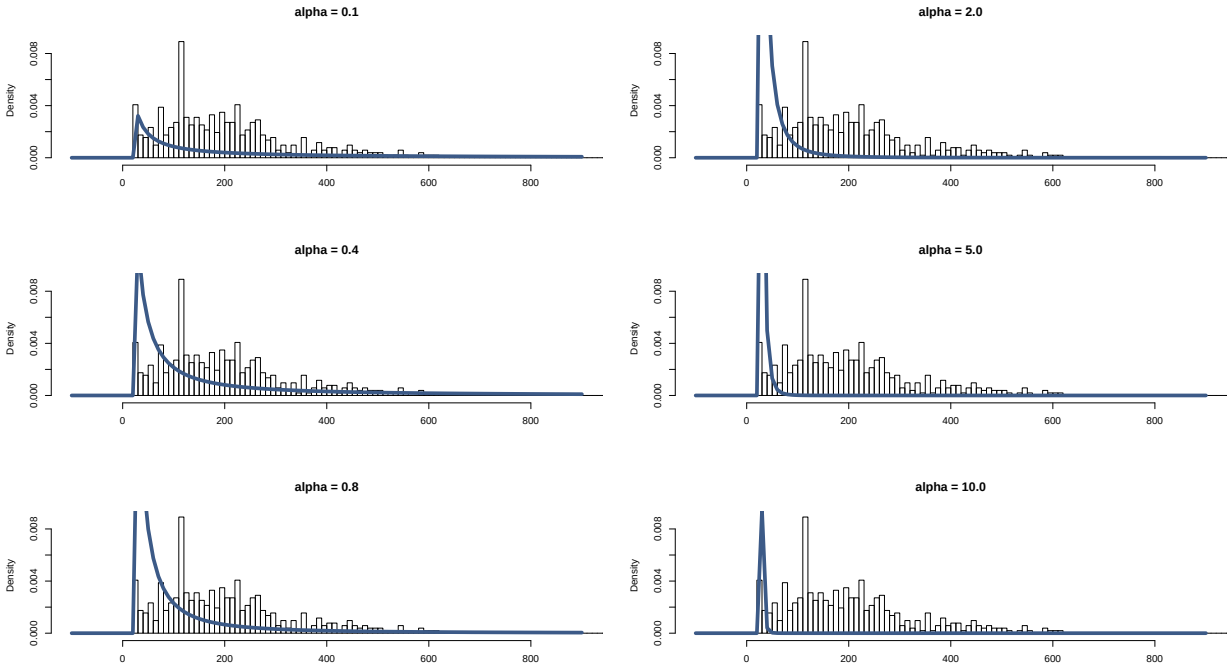
==========

Chercher un package R permettant de générer des réalisations d'une loi de Pareto. Visualiser l'impact du paramètre α.

==========

Pour générer des réalisations d'une loi de Pareto, on choisit d'utiliser le package **actuar**, avec la fonction **dpareto1** pour obtenir la densité, et la fonction **rpareto1** pour générer les réalisations.

Comme on peu facilement le montrer à partir de la fonction de densité, le maximum est atteint en $z = m$ avec une valeur $f_Z(m; m, \alpha) = \frac{\alpha}{m}$. Ainsi, le maximum augmente avec α. En revanche, plus α est grand, plus la descente est abrupte. Ci-dessous, on affiche les densités pour 6 valeurs différentes de α, et on les superpose à l'histogramme de **Barre**. Des quelques valeurs testées pour le paramètre α, ce sont les valeurs $\alpha = 0.4$ et $\alpha = 0.8$ qui semblent la plus en adéquation avec l'histogramme: **on s'attend donc à une estimation de la valeur de α entre 0.4 et 0.8**.



2.5. Choix de la loi a priori

==========
Choisir une loi a priori pour α.
==========

Sans connaissance particulière, nous choisissons de déterminer la prior de Jeffrey.
Dans un premier temps, on dérive 2 fois la log-vraisemblance.

$$l(\alpha|z) = log(\alpha) + \alpha.log(m) - (\alpha+1).log(z)$$

$$\frac{\partial l}{\partial \alpha} = \frac{1}{\alpha} + log(m) - log(z)$$

$$\frac{\partial^2 l}{\partial \alpha^2} = -\frac{1}{\alpha^2}$$

Cette dernière valeur ne dépend plus de z, donc est égale à son espérance.
On en déduit donc l'information de Fisher pour la loi de Pareto.

$$I(\alpha) = -E\left[\frac{\partial^2 l}{\partial \alpha^2}\right] = \frac{1}{\alpha^2}$$

On en déduit la prior de Jeffrey.

$$\pi(\alpha) \propto \sqrt{I(\alpha)} = \frac{1}{\alpha}$$

Cette loi est impropre, mais comme nous allons le voir ensuite, la posterior associée sera elle bien définie.

2.6. Identification de la loi a posteriori

==========
Donner la loi a posteriori pour α.
==========

Soit n le nombre d'individus.

$$\pi(\alpha|Z) \propto \pi(\alpha).L(\alpha|Z) \propto \frac{1}{\alpha} \cdot \prod_i \left(\alpha.\frac{m^\alpha}{z_i^{\alpha+1}}\right) \propto \alpha^{n-1} \cdot e^{-\alpha \cdot \sum_i log(\frac{z_i}{m})}$$

On reconnait une loi Gamma.

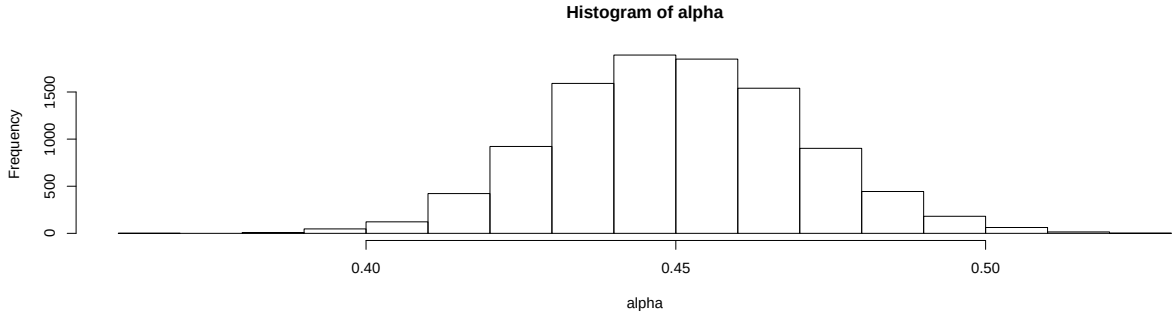
$$\alpha_{posterior} \sim \Gamma(n, \sum_i log(\tfrac{z_i}{m}))$$

2.7. Tirage d'un échantillon de la loi a posteriori

==========
Par la méthode de votre choix, tirer un échantillon de la loi a posteriori de α. Donner un intervalle de crédibilité à 95%.
==========

A l'aide de la fonction **rgamma**, on tire un échantillon de la loi a posteriori.

```
# parametres
  Z = data$Barre
  n = length(Z)
  m = 21
  S = sum(log(Z/m))
  niter = 10000
# tirage
  alpha = rgamma(niter, n, S)
  hist(alpha)
```



L'histogramme confirme l'intuition que nous avions eu en testant les différentes valeurs de pour le paramètre α, à savoir une estimation entre 0.4 et 0.8. Ci-dessous, les intervalles de crédibilités à 95%. On constate que les valeurs expérimentales sont tres proches des valeurs théoriques.

```
# Intervalles de crédibilité experimental
  quantile(alpha, c(0.025, 0.975))
```

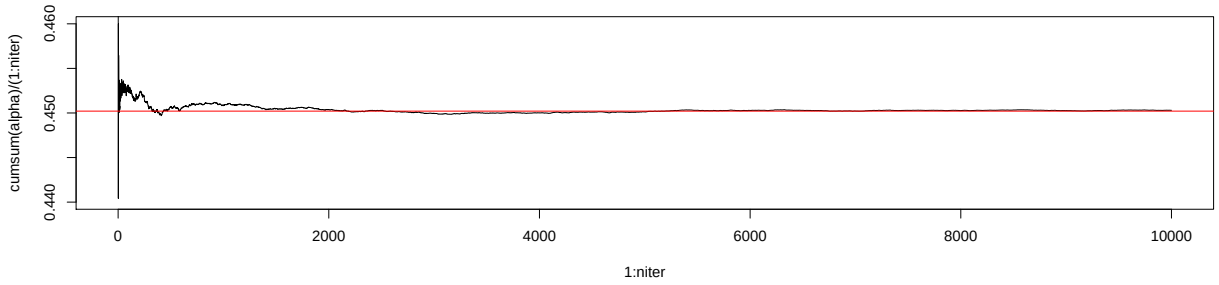
```
##      2.5%     97.5%
## 0.4124107 0.4905134
```

```
# Intervalles de crédibilité theorique
  qgamma(c(.025, .975), n, S)
```

```
## [1] 0.4121942 0.4898709
```

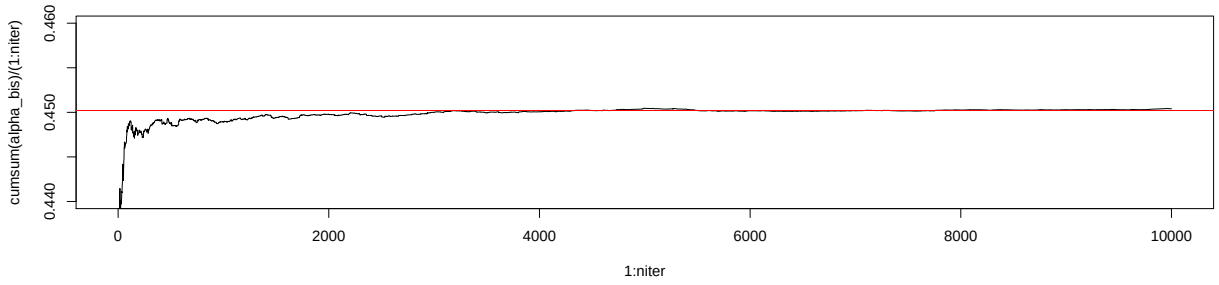
Enfin, la figure ci-dessous nous montre, la bonne convergence de notre estimateur.

```
# convergence du monte-carlo classique
  plot(1:niter, cumsum(alpha)/(1:niter), type="l", ylim=c(0.440,0.460))
  abline(h=n/S, col=2) # esperance analytique
```



Pour le fun, on fait un échantillonnage d'importance avec comme distribution instrumentale la loi gaussienne, mais cela n'améliore pas notre estimateur qui avait déjà convergé rapidement avec le Monte-Carlo classique.

```
# convergence de l'échantillonnage d'importance
  tirage = rnorm(niter, n/S, sqrt(n)/S)
  alpha_bis = tirage * dgamma(tirage, n, S) / dnorm(tirage, n/S, sqrt(n)/S)
  plot(1:niter, cumsum(alpha_bis)/(1:niter), type="l", ylim=c(0.440,0.460))
  abline(h=n/S, col=2) # esperance analytique
```



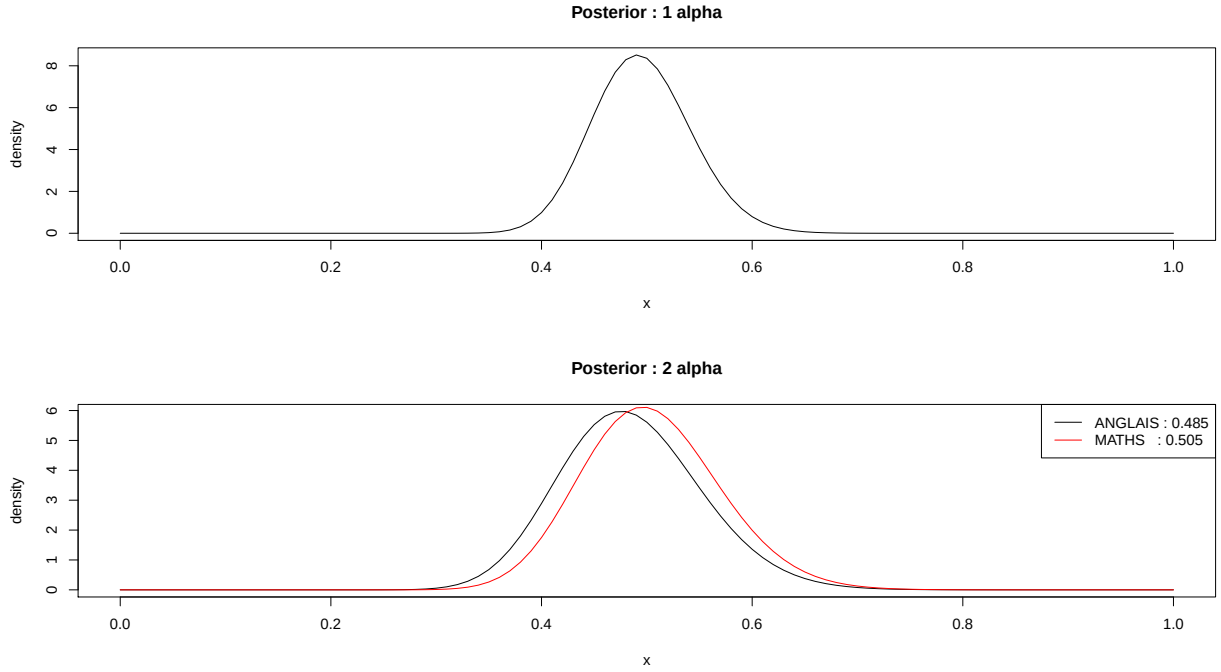
2.8. Mutations en mathématiques et en anglais

==========
On se concentre uniquement sur les mutations en mathématiques et en anglais. Répéter l'analyse pour chacune de ces deux catégories. Que pensez-vous de l'hypothèse que $\alpha_{maths} = \alpha_{anglais}$?
==========

2.8.1. Densités des lois a posteriori

On trace les densités des lois a posteriori, pour chacune des 2 catégories.

```
# parametres
  Z1 = data$Barre[data$Matiere=="ANGLAIS"]
  Z2 = data$Barre[data$Matiere=="MATHS"]
  n1 = length(Z1)
  n2 = length(Z2)
  S1 = sum(log(Z1/m))
  S2 = sum(log(Z2/m))
  niter = 10000
# densites
  par(mfrow=c(2, 1))
  curve(dgamma(x, n1+n2, S1+S2), xlim=c(0, 1), main="Posterior : 1 alpha", ylab="density")
  curve(dgamma(x, n1, S1),       xlim=c(0, 1), main="Posterior : 2 alpha", ylab="density")
  curve(dgamma(x, n2, S2), col=2, add=T)
  str1 = sprintf("ANGLAIS : %.3f",n1/S1)
  str2 = sprintf("MATHS   : %.3f",n2/S2)
  legend("topright", c(str1,str2), col=1:2, lty=1)
```



Les 2 distributions semblent légèrement translatées, avec $\hat{\alpha}_{ANGLAIS} = 0.485$ et $\hat{\alpha}_{MATHS} = 0.505$. Intuitivement, on veut donc rejeter l'hypothèse $\alpha_{maths} = \alpha_{anglais}$.

2.8.2. Facteur de Bayes

Pour confirmer cette intuition, on décide de calculer le facteur de Bayes pour comparer les 2 modèles suivants.

$m_0(z)$: un seul alpha commun au 2 catégories.
$m_1(z)$: un alpha par catégorie.

Soient n_1 le nombre d'individus, $P_1 = \prod_i z_i$ et $S_1 = \sum_i log(\frac{z_i}{m}))$ pour la catégories **ANGLAIS**.
Soient n_2 le nombre d'individus, $P_2 = \prod_i z_i$ et $S_2 = \sum_i log(\frac{z_i}{m}))$ pour la catégories **MATHS**.

Le facteur de Bayes s'écrit comme suit

$$B = \frac{m_0(z)}{m_1(z)}$$

Modèle $m_0(z)$

$$m_0(z) = \int L(\alpha|Z) \,.\, \pi(\alpha)\, d\alpha = \frac{1}{P_1.P_2} \int \alpha^{n_1+n_2-1} \,.\, e^{-\alpha \,.\, (S_1+S_2)}\, d\alpha$$

$$m_0(z) = \frac{\Gamma(n_1+n_2-1)}{P_1.P_2.(S_1+S_2)^{n_1+n_2-1}} \int \alpha \,.\, \left[\frac{(S_1+S_2)^{n_1+n_2-1}}{\Gamma(n_1+n_2-1)} \,.\, \alpha^{n_1+n_2-1-1} \,.\, e^{-\alpha \,.\, (S_1+S_2)} \right] d\alpha$$

On reconnait l'espérance d'une loi $\Gamma(n_1+n_2-1, S_1+S_2)$, donc

$$m_0(z) = \frac{\Gamma(n_1+n_2-1)}{P_1.P_2.(S_1+S_2)^{n_1+n_2-1}} \cdot \frac{n_1+n_2-1}{S_1+S_2} = \frac{\Gamma(n_1+n_2)}{P_1.P_2.(S_1+S_2)^{n_1+n_2}}$$

Modèle $m_1(z)$

Par un raisonnement identique, on en déduit

$$m_1(z) = \frac{\Gamma(n_1)}{P_1.(S_1)^{n_1}} \cdot \frac{\Gamma(n_2)}{P_2.(S_2)^{n_2}}$$

On peut donc calculer le facteur de Bayes. Pour le calcul, on va utiliser log-gamma pour éviter d'atteindre des valeurs non-manipulables par l'ordinateur.

```
# Facteur de Bayes [on ne calcule pas P1 et P2, qui se neutralisent dans B]
  log_m0 = lgamma(n1+n2)-(n1+n2)*log(S1+S2)
  log_m1 = lgamma(n1)+lgamma(n2)-n1*log(S1)-n2*log(S2)
  B = exp(log_m0-log_m1)
  sprintf("log10[Bayes Factor] = %.2f",log10(B))
```

```
## [1] "log10[Bayes Factor] = 0.31"
```

Le facteur de Bayes est en faveur du modèle $m_0(z)$, ce qui tend à faire **accepter** l'hypothèse $\alpha_{maths} = \alpha_{anglais}$, contrairement à notre intuition précédente. En revanche, ce résultat est à prendre avec des pincettes, car **l'évidence est faible (<0.5)**.