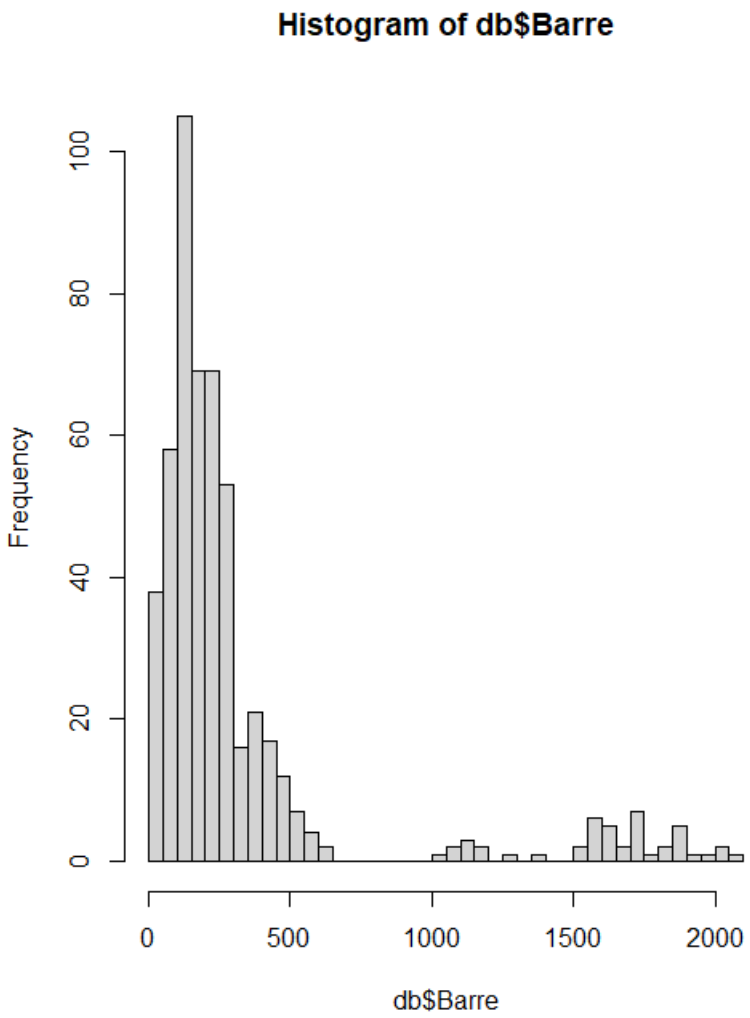
Nom et Prénom : GUZMAN Jorge

Projet : Cours d'inférence bayésienne

Objet : L'objet de ce document est d'appliquer les bases de la statistique bayésienne et donner réponses aux différents questions formulées dans le cadre de ce TP.

Données : Le dataset mise à disposition contient 516 observations et 23 variables dont 4 variables labelisées. L'objectif de ce projet est d'appliquer la statistique bayésienne afin d'inférer le nombre de points nécessaire pour une mutation. Dans le histogramme ci-dessous il est constaté une concentration importante entre 21 à 300 points, ce qui est confirmé dans le summary de la variable avec un min et un max à 21 et à 2056 points respectivement. Le quantile à 50% et moyenne sont à 196 et 321 respectivement ce qui permet de retenir la dissymétrie de la distribution.



Chapitre 1 : Régression linéaire

Pour la suite on va conserver uniquement la variable "Matiere" comme variable catégorielle car on gardant toutes les autres variables catégorielles on aura des petites échantillons par modalités ce qui empêchera la modélisation (exp. LYCEE RAFA - ANGLAIS, 2 observations). Il est nécessaire de conserver "barre" en première position pour pouvoir utilisée la librairie pour l'inférence bayésienne.

Pour information la librairie utilisée et « rstanarm » et non « BMS » car cette dernière ne reçoit pas des variables catégorielles.

La librairie « rstanarm » reçoit des variables catégorielles sans besoin d'une transformation numérique mais les chaînes de markof ne convergent pas même en augmentant le nombre d'interactions, les modèles réalisées font partie de la documentation de ce projet (avec toutes les variables, uniquement avec les variables « matière » et « établissement ».

Il est possible que la non convergence du modèle soit liée au set des données que compte avec un nombre faible des observations en considérant la combinaison « établissement » et « matière » donc la variable « établissement » a été exclue de l'analyse, cette exclusion a permis la convergence du modèle mais l'estimation de la variable « barre » a été mauvaise.

Les erreurs quadratique bayésiennes et fréquentiste sont élèves (288 et 282 respectivement) donc le data set a été divisé selon le histogramme de distribution de la variable "Barre", tous les points supérieurs à 1000 ont été exclu de la modélisation ce qui réduit la base des données à 471 observations Vs les 516 observations du dataset original.

Avec les nouveau dataset les erreurs quadratiques sont divisée par deux (102) pour les deux approches, fréquentiste et bayésienne, sur cette base les deux modèles sont égaux.

Le retour de l'inférence bayésienne est la distribution des valeurs qui pourra prendre les coefficients. En revanche, la régression linéaire retour qu'un seule valeur pour chaque coefficient ainsi qu'une valeur de probabilité associée qui permet d'établir un niveau de significativité. Il est observé que si l'intervalle du coefficient de la statistique bayésienne touche le zéro ce coefficient n'est pas significatif dans la régression linéaire.

Pour finaliser cette première partie le même analyse sera effectué uniquement sur les mutations « mathématiques » et « anglais ».

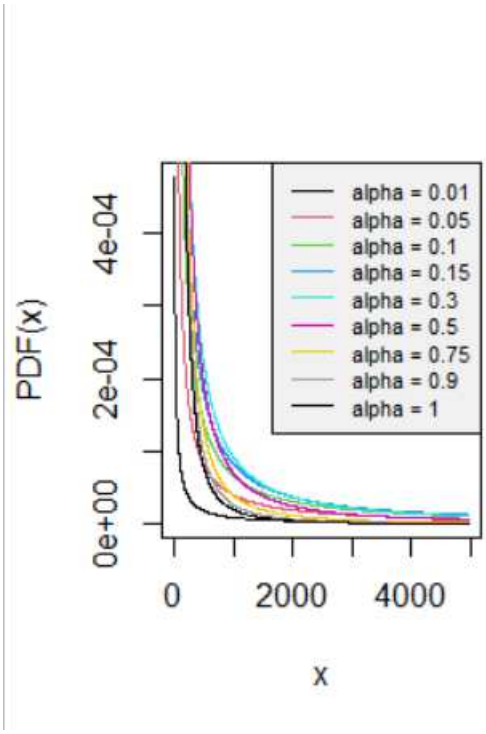
Sur la base de l'erreur quadratique le modèle fait que sur les mutation « anglais » sont très proche, 100 (modèle linéaire) et 99 (modèle bayésienne). En revanche, sur les modèles basé sur les mutations « mathématiques » l'écart est plus important, 64 (modèle linéaire) et 44 (modèle bayésienne).

Il est observé que dans le modèle « mathématique » seulement une covariable est significative (taux_reussite_attendu_total_series) pour la régression linéaire et toutes les covariables de la régression bayésienne touchent le zéro.

Pour le modèle « anglais » 4 covariables sont significatives (effectif_presents_serie_es, effectif_de_seconde, taux_acces_attendu_seconde_bac, taux_acces_attendu_premiere_bac). En revanche, que deux covariables de la régression bayésienne ne touchent pas le zéro (effectif_presents_serie_es, taux_acces_attendu_premiere_bac).

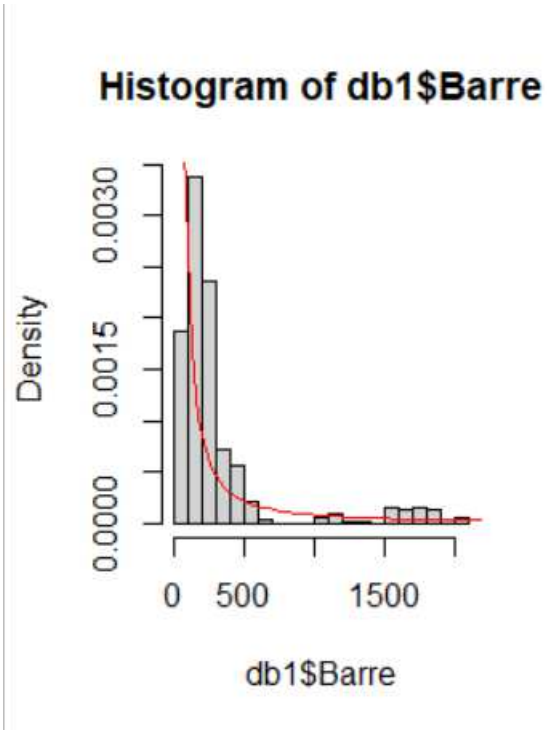
Ces deux constat permet d'affirmer que les covariables n'agissent de la même manière dans ces deux disciplines.

Chapitre 2 : Loi de Pareto



On observe sur le graphique ci-avant (voir code R)que si les valeurs d'alpha oscillant entre [0 ; 0,5] les coudes des courbes est plus éloigné des axes (x ; y). En revanche, avec un alpha [0,5 ;1] la courbe est plus proche des axes.

Choix de la loi a priori d'alpha :



On regardant le histogramme de Barre, on peut partir sur l'hypothèse que cette variable suit une distribution de Pareto, donc on pourrais supposer que la fonction de max de vraisemblance prend cette forme. Sur la base de la littérature l'a priori aura une distribution Gamma.

When likelihood function is a continuous distribution [edit]

Likelihood	Model parameters	Conjugate prior distribution	Prior hyperparameters	Posterior hyperparameters[note 1]	Interpretation of hyperparameters
Pareto with known minimum x_m	k (shape)	Gamma	$\alpha,\ \beta$	$\alpha + n,\ \beta + \sum_{i=1}^{n} \ln \frac{x_i}{x_m}$	α observations with sum β of the order of magnitude of each observation (i.e. the logarithm of the ratio of each observation to the minimum x_m)

https://en.wikipedia.org/wiki/Conjugate_prior

Cette distribution Gamma compte avec deux paramètres alpha et beta qui seront estimés via le max de vraisemblance (voir code R) https://en.wikipedia.org/wiki/Gamma_distribution

Closed-form estimators [edit]

Consistent closed-form estimators of k and θ exists that are derived from the likelihood of the generalized gamma distribution.[19]

The estimate for the shape k is

$$\hat{k} = \frac{N\sum_{i=1}^{N} x_i}{N\sum_{i=1}^{N} x_i \ln(x_i) - \sum_{i=1}^{N} \ln(x_i) \sum_{i=1}^{N} x_i}$$

and the estimate for the scale θ is

$$\hat{\theta} = \frac{1}{N^2}\left(N\sum_{i=1}^{N} x_i \ln(x_i) - \sum_{i=1}^{N} \ln(x_i) \sum_{i=1}^{N} x_i\right)$$

If the rate parameterization is used, the estimate of $\hat{\beta} = \frac{1}{\hat{\theta}}$.

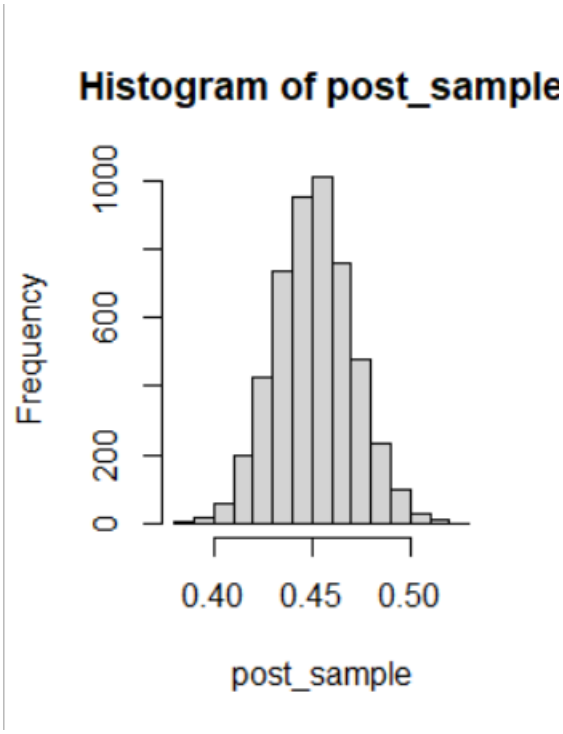
Choix de la loi a posteriori d'alpha :

Les paramètres seront calculés conforme à la littérature (voir tableau ci-avant), la fonction utilisée est fait sur R. https://en.wikipedia.org/wiki/Conjugate_prior

Posterior hyperparameters[note 1]	Interpretation of hyperparameters
$\alpha + n,\ \beta + \sum_{i=1}^{n} \ln \frac{x_i}{x_m}$	α observations with sum β of the order of magnitude of each observation (i.e. the logarithm of the ratio of each observation to the minimum x_m)

Intervalle de crédibilité à 95% :

Avec un échantillon de taille N=5000 le limite inférieur se trouve à 0.4145 et le supérieur à 0.4905, ce calcul est détaillé dans le code R mais on pourrais l'observé dans l'histogramme ci-dessous.



Analyse pour anglais et mathématiques, est-ce que les alpha sont égaux ?

Dans R l'exercice a été refait sur anglais et mathématiques, il est observé que même si les médianes sont proches les alphas sont différents.

Alpha mathématiques : [0,40 ; 0,66] avec une médiane à 0,52

Alpha anglais : [0,37 ; 0,63] avec une médiane à 0,49