

TP_2 stats bayésiennes (*Durée : 1h*)

Avertissement : L'objectif de ces consignes est seulement d'aider à comprendre l'exercice demandé. Vous pouvez faire d'autres choix ou aller beaucoup plus rapidement sans programmer toutes les fonctions intermédiaires demandées ici. Les fonctions demandées ici seront très courtes et souvent évidentes.

1. On travaillera sur les données du projet, et pour commencer sur une seule variable (vous pouvez reprendre ce TP avec plusieurs variables si vous avez fini).
 - Quelle loi choisir pour $\pi(X_i | \theta)$. Pour l'exercice, on travaillera avec une loi admettant un prior conjugué.
 - Donner la forme des priors conjugués pour cette loi. Si cela n'a pas été vu en cours, faire la preuve de la propriété de conjugaison.
 - Créer une fonction `update` qui en fonction des hyperparamètres du prior, et des données, calcule les hyperparamètres du posterior.
2. On veut maintenant construire une région de crédibilité bayésienne. On suppose pour cette question seulement que cette région est connexe (donc un intervalle si θ est unidimensionnel). Pour construire la région de crédibilité, on se propose donc de parcourir, à l'aide des quantiles de la loi a posteriori, tous les intervalles possibles de probabilité $1 - \alpha$ a posteriori.
 - Créer une fonction `region_asym` qui prend en argument les hyperparamètres du posterior, et $\alpha_1, \alpha_2 > 0$, et renvoie q_1, q_2 tels que

$$\mathbb{P}(\theta \leq q_1 | X) = \alpha_1$$
$$\mathbb{P}(\theta > q_2 | X) = \alpha_2$$

- Créer une fonction `region_vol_min`, qui prend en argument les hyperparamètres du posterior, et $\alpha \in [0, 1]$, et qui renvoie la région $(1 - \alpha)$ -crédible de volume minimal (i.e. $q_2 - q_1$ minimal) obtenue en parcourant plusieurs couples α_1, α_2 tels que $\alpha_1 + \alpha_2 = \alpha$.

On pourra utiliser cette méthode de construction de région de crédibilité lorsque le postérieur est unimodal et continue, et que l'on peut obtenir facilement les quantiles de ce dernier.

3. Dans cette question, on souhaite faire la même construction de région $(1 - \alpha)$ -crédible, mais cette fois ci en parcourant différents seuils possibles pour la densité de la loi a posteriori.
 - Créer un vecteur `theta_vec` maillant l'ensemble des paramètres, et calculer la densité a posteriori en chaque point de ce vecteur que l'on stockera dans un vecteur `y`.
 - Créer une fonction `region_hpd` qui à un seuil `s`, et le couple `(theta_vec, y)` renvoie la région *HPD* associé à ce seuil, ainsi que la probabilité a posteriori de cette région calculée par intégration numérique sur la même grille `theta_vec`.

- Créer une fonction `region_HPD_alpha` qui prend en argument les hyperparamètres du posterior, et $\alpha \in [0, 1]$, et qui renvoie la région HPD $(1 - \alpha)$ -crédible obtenue en parcourant plusieurs seuils s sur la densité a posteriori.

On pourra utiliser cette méthode de construction de région de crédibilité lorsque le nombre de paramètres est assez réduit pour pouvoir utiliser de l'intégration numérique sur une grille.

4. Dans cette question, on considère une situation où l'on disposerait seulement d'un échantillon simulé sous la loi a posteriori (situation qui arrivera le plus souvent comme on le verra plus tard dans le cours). On veut donc construire une région de crédibilité seulement avec cet échantillon.
 - Simuler de façon *had-oc* (en utilisant par exemple la propriété de conjugaison) un échantillon `theta_sim` de taille $n = 10000$ sous la loi a posteriori. Créer aussi un vecteur `z` correspondant à l'évaluation de la densité du posterior en `theta_sim`.¹
 - À partir de cette question, on se donne comme objectif de construire la région de crédibilité seulement avec `theta_sim`, et `z`. Comme précédemment, on va construire une fonction `region_HPD_sim` qui prend en argument un niveau α et le couple `(theta_sim, z)` renvoie la région HPD $(1 - \alpha)$ -crédible. On pourra pour cela classer les `theta_sim` par valeur croissante de `z`.

On pourra utiliser cette méthode de construction de région de crédibilité lorsque le nombre de paramètres est trop grand pour pouvoir utiliser de l'intégration numérique sur une grille, mais que l'on sait aisément simuler la loi a posteriori.

5. Comparer les résultats obtenus par les différentes méthodes vues au dessus. Par soucis de comparaison, que vaudrait l'intervalle de confiance fréquentiste de niveau α ? Faire une belle représentation de cette région de crédibilité. On pourra aussi reprendre des points de ce TP en l'adaptant avec plusieurs paramètres, des lois non explicites...

¹Évidemment, en pratique, on ne fera jamais cette étape, mais l'on disposera d'autres méthodes de simulation pour les cas où l'on ne dispose pas de propriété de conjugaison, et le vecteur `z` correspondra alors à la densité a posteriori calculée à une constante près.