Projet – Analyse bayésienne

Ensembles de données
ab_data.csv, mimic3d.csv

Description des données:
Les données proviennent de:

- https://www.kaggle.com/zhangluyuan/ab-testing
- https://www.kaggle.com/drscarlat/predict-hospital-length-of-stay-classification/data

Les fichiers .csv sont disponible dans la filière "Data".

Questions:

1. Supposons que des responsables du marketing testent une nouvelle page web dans l'espoir d'augmenter le taux de conversion (la proportion de visiteurs qui deviennent membre). Lorsqu'un visiteur se présente sur le site web, on lui donne accès soit à la page originale ou à la nouvelle page, au hasard. Les résultats sont disponible dans le fichier *ab_data.csv*, qui répertorie les visites des utilisateurs en précisant si on leur donnait accès à la nouvelle page ou à l'ancienne, et s'il y a eu conversion. Résumez les données et préparez quelques graphiques.

2. Effectuer un test A/B bayésien pour mesurer l'effet de la page sur le taux de conversion. Vous pouvez définir et mettre à jour des distributions a priori indépendantes sur l'ancien et le nouveau taux de conversion et obtenir les distributions a posteriori correspondantes. Utilisez une loi a priori Beta(alpha=2, beta=20) pour l'ancien taux, qui représente ce qui a été observé par le passé. Qu'elle loi a priori utiliseriez-vous pour le nouveau taux de conversion?

3. Commencez avec un sous-ensemble de 100 observations, choises aléatoirement, et effectuez une analyse d'inférence. Quelle est la probabilité a posteriori que la nouvelle page ait un taux de conversion plus élevé? Conseil: utilisez des échantillons aléatoires indépendants provenant de la distribution a posteriori afin d'estimer la probabilité.

4. Ré-actualiser les distributions postérieures avec 100 observations supplémentaires. À partir de quelle taille d'échantillon est-ce que la forme spécifique des distributions a priori n'est plus pertinente?

5. Nous ne cherchons pas toujours seulement à estimer une réponse; parfois nous désirons aussi obtenir une distribution de probabilité pour celle-ci. Par exemple, si votre espérance de vie est de 80 ans, vous voudrez peut-être savoir si elle est uniforme sur l'intervalle 40-120 ans, ou si c'est plutôt une distribution dont la masse se retrouve près de 80 ans. Explorer les données qui se retrouvent dans le fichier *mimic3d.csv*, qui énumère les durées de séjour à l'hôpital ("LOSdays"), ainsi qu'un certain nombre d'autres variables.

6. Ajoutez une observation avec les valeurs les plus probables pour votre prochain séjour à l'hôpital (ou celui de quelqu'un d'autre). Effectuez une analyse de régression linéaire bayésienne. Quelle est la probabilité de rester plus de 2 jours à l'hôpital (et donc de s'absenter définitivement du travail)? Pour plus de simplicité, vous pouvez utiliser des distributions a priori normales.