TD 1: Modèles bayésiens conjugués

Objectifs:

1. Découvrir des exemples d'application concrets des statistiques Bayésiennes.
2. Se familiariser avec les lois de probabilités classiques et leur utilisation pratique en modélisation.
3. Pratiquer, s'exercer et tester ses connaissances et sa compréhension de la méthode Bayésienne.

Nom	**Domaine**	**Densité**	**Espérance**	**Variance**
Bernoulli($p,\ 0 \leq p \leq 1$)	$\{0,1\}$	$p^x(1-p)^{1-x}$	p	$p(1-p)$
Binomiale($n \in \mathbb{N},\ 0 \leq p \leq 1$)	$\{0,\dots,n\}$	$\binom{n}{x}p^x(1-p)^{n-x}$	np	$np(1-p)$
Poisson($\lambda > 0$)	$\mathbb{N}$	$\frac{\lambda^x e^{-\lambda}}{x!}$	λ	λ
Exponentielle($\lambda > 0$)	$\mathbb{R}^+$	$\lambda e^{-\lambda x}$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$
Normale($\mu \in \mathbb{R},\ \sigma^2 > 0$)	$\mathbb{R}$	$\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$	μ	σ^2
Beta($\alpha, \beta > 0$)	$[0,1]$	$\frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)}$	$\frac{\alpha}{\alpha+\beta}$	$\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$
Gamma($\alpha, \beta > 0$)	$\mathbb{R}^+$	$\frac{\beta^\alpha}{\Gamma(\alpha)}x^{\alpha-1}e^{-\beta x}$	$\frac{\alpha}{\beta}$	$\frac{\alpha}{\beta^2}$
Inverse-Gamma($\alpha > 1, \beta > 0$)	$\mathbb{R}^+$	$\frac{\beta^\alpha}{\Gamma(\alpha)}x^{-\alpha-1}e^{-\beta/x}$	$\frac{\beta}{\alpha-1}$	$\frac{\beta^2}{(\alpha-1)^2(\alpha-2)}$, si $\alpha > 2$

Table 1: Tableau des distributions de probabilité classiques

Exercice 1 (Modèlisation des sinistres en assurance) – Une compagnie d'assurance automobile souhaite mettre à jour son estimation de la fréquence des sinistres d'un client A après n années. Le nombre de sinistres par an est souvent modélisé par une loi de Poisson car les sinistres sont des événements rares et indépendants. On note ce nombre par la variable aléatoire N qui suit une loi de Poisson $\mathcal{P}(\lambda)$. On rappelle sa densité:

$$\mathbb{P}(N = k) = \frac{\lambda^k e^{-\lambda}}{k!}, \quad k \in \mathbb{N}$$

Les observations des n dernières années du client A notées par $N_1, \dots, N_n$ sont données par 1, 0, 3, 2, 0 (ici $n = 5$). Comme la moyenne de $\mathcal{P}(\lambda)$ est égale à λ, λ peut être interprétée comme le nombre moyen de sinistre par an. L'assurance a également des données historiques du nombre moyen de sinistres par an pour chacun de ses m clients: $\mu_1, \dots, \mu_m$ dont la moyenne et l'écart type sont donnés par $\bar{\mu} = \frac{1}{m}\sum_{i=1}^m \mu_i = 3.4$ et $\hat{\sigma_\mu} = \sqrt{\frac{1}{m}\sum_{i=1}^m (\mu_i - \bar{\mu})^2} = 2$.

Dans toutes les questions suivantes, donnez d'abord les formules en fonction des variables de l'énoncé avant d'utiliser les valeurs empiriques mentionnées.

1. Estimez λ en adoptant une approche fréquentiste. Cet estimateur est noté $\widehat{\lambda_F}$.

2. On souhaite à présent adopter une approche Bayésienne en prenant une loi a priori $\lambda \sim \text{Gamma}(\alpha, \beta)$. On note sa moyenne λ_{prior}. Comment peut-on choisir ses paramètres α, β ?

3. Déterminez la loi a posteriori $\lambda | N_1, \dots, N_n$.

4. À quoi correspond l'estimateur de Bayes dans ce cas ?

5. Écrivez la formule de l'estimateur de Bayes $\widehat{\lambda_B}$ comme combinaison convexe de la moyenne fréquentiste et de la moyenne a priori, c-à-d, trouvez $Z \in [0, 1]$ tel que:

$$\widehat{\lambda_B} = Z\widehat{\lambda_F} + (1 - Z)\lambda_{\text{prior}}$$

6. En théorie de crédibilité, Z est un score associé au client A. Il est appelé "facteur de crédibilité". Pourquoi à votre avis ?

7. Expliquez l'influence du nombre d'années d'observation n sur l'estimation.

Exercice 2 (Bayesian A/B testing) – Google souhaite comparer l'efficacité de deux publicités (A et B) en menant un test A/B. Chaque publicité est affichée à un certain nombre d'utilisateurs n_A, n_B. On mesure le taux de conversion θ_A (resp. θ_B), c'est-à-dire la probabilité qu'un utilisateur clique sur la publicité A (resp. B) après l'avoir vue. On note X_A et X_B le nombre de clics respectivement obtenus pour les publicités A et B. Le but de l'étude est de comparer θ_A et θ_B en utilisant d'abord une approche fréquentiste, puis une approche bayésienne. On observe les chiffres suivants: $X_A = 40$, $X_B = 65$, $n_A = 1100$, $n_B = 1300$.

1. Approche Fréquentiste

1. Quel est le modèle approprié ?

2. Proposez un estimateur pour chaque paramètre θ_A et θ_B avec l'approche fréquentiste. On note ces estimateurs $\widehat{\theta_{Af}}$ et $\widehat{\theta_{Bf}}$.

3. Montrez que ces estimateurs sont asymptotiquement Gaussiens et déterminer les paramètres de leur limite Gaussienne.

4. On suppose que les données des deux publicités sont indépendantes. En déduire la distribution asymptotique de leur différence.

5. On considère l'hypothèse $H_0 : \theta_A = \theta_B = \theta$, proposez un estimateur de θ. Déduire de la question précédente une statistique de la forme $W \overset{\text{def}}{=} \frac{\widehat{\theta_{Af}} - \widehat{\theta_{Bf}}}{Z}$ avec Z à déterminer telle que:

$$W \overset{n_A, n_B \to +\infty}{\sim} \mathcal{N}(0, 1).$$

6. En déduire un moyen de tester si l'hypothèse $\theta_A < \theta_B$ est vraie.

2. Approche Bayésienne

Google souhaite maintenant adopter une approche bayésienne en utilisant une loi Beta comme a priori :

$$\theta_A \sim \text{Beta}(\alpha_A, \beta_A), \quad \theta_B \sim \text{Beta}(\alpha_B, \beta_B)$$

7. Pour quelles valeurs de α, β obtiendrait-on des lois a priori non-informatives ?
8. Déterminez la loi a posteriori de θ_A et θ_B après observation des données.
9. Comment peut-on définir un estimateur bayésien de $\theta_A - \theta_B$?
10. Proposez une méthode de simulation empirique pour évaluer $\mathbb{P}(\theta_A < \theta_B)$.
11. Implémentez cette méthode et comparez le résultat à celui de l'approche fréquentiste.
12. Les entreprises en tech qui ont recours à la procédure du test A/B pour décider la meilleure version d'un produit, site-web, système de recommandation etc.. ont tendance à adopter l'approche Bayésienne. Comment pouvez-vous l'expliquer ?

Exercice 3 (Rendements de portefeuille) – Un analyste financier veut estimer la rentabilité moyenne μ d'un portefeuille d'actions. On suppose que les rendements passés $X_1, \dots, X_n$ suivent une loi normale:

$$X_i | \mu \sim \mathcal{N}(\mu, \sigma^2)$$

.

Partie 1. σ^2 connue.

On suppose σ^2 connue. L'analyste a une croyance a priori sur μ et modélise cette incertitude par une loi normale :

$$\mu \sim \mathcal{N}(\mu_0, \tau_0^2)$$

1. Donnez l'estimateur de μ selon une approche fréquentiste.
2. Déterminez la loi a posteriori de μ après observation des rendements passés.
3. Déduisez l'estimateur bayésien de μ.
4. Expliquez comment cet estimateur prend en compte l'information a priori et les données observées.

Partie 2. σ^2 inconnue.

On suppose maintenant que la variance σ^2 est inconnue et qu'on la modélise avec une loi a priori Inverse-Gamma :

$$\sigma^2 \sim \mathrm{IG}(\alpha_0, \beta_0)$$

5. Déterminez la loi jointe a posteriori de (μ, σ^2) après observation des rendements.
6. Identifiez la distribution marginale a posteriori de σ^2.
7. Donnez l'estimateur bayésien de μ en intégrant l'incertitude sur σ^2.
8. Comparez cet estimateur avec celui obtenu lorsque σ^2 était supposé connue.

Exercice 4 (Modélisation de durée) – Un hôpital souhaite modéliser le temps d'attente T des patients avant une consultation médicale. On suppose que T suit une loi exponentielle :

$$T|\lambda \sim \text{Exp}(\lambda)$$

où λ représente le taux d'arrivée des patients. L'hôpital utilise une approche bayésienne et modélise λ avec une loi Gamma :

$$\lambda \sim \text{Gamma}(\alpha_0, \beta_0)$$

1. Donnez l'estimateur fréquentiste de λ basé sur les durées d'attente observées.
2. Déterminez la loi a posteriori de λ après observation des temps d'attente.
3. Trouvez l'estimateur bayésien de λ.
4. Comparez l'estimateur bayésien et l'estimateur fréquentiste.