Séance 5 - Statistique bayésienne

Exercice 1. Chez les mammifères, les males possèdent un chromosome X et un chromosome Y, tandis que les femmes possèdent deux copies du chromosome X. Chez un individu, chacun des 2 chromosomes est hérité d'un seul des 2 parents.

L'hémophilie est une maladie génétique humaine liée à un allèle récessif localisé sur le chromosome X. Cela signifie qu'une femme portant l'allèle responsable de la maladie sur l'un de ses deux chromosomes n'est pas affectée par la maladie. La maladie est, en revanche, généralement fatale pour une femme qui porterait les deux copies mutées (un événement très rare de toutes façons).

1. Une femme sait que son frère est affecté, mais que son père ne l'est pas. A priori, quelle est la loi de son propre état, θ, que l'on considère comme un paramètre binaire (porteuse ou non de l'allèle récessif) ?
2. On prend maintenant en compte les données suivantes. Cette femme a deux fils (non-jumeaux) et aucun des deux n'est affecté. Quelle est la loi a posteriori du paramètre θ ?

Exercice 2. On considère la fonction de vraisemblance donnée par

$$p(y_1, \ldots, y_n|\theta) \propto \prod_{i=1}^{n} \exp\left(-\frac{1}{2\sigma^2}(y_i - \theta)^2\right), \quad y_i \in \mathbb{R}$$

où la moyenne, θ, est un paramètre inconnu. On suppose que la variance, σ^2, est connue, et que la loi a priori de θ est dégénérée (non-informative)

$$p(\theta) \propto 1\,.$$

On pose $\bar{y} = \sum_{i=1}^{n} y_i/n$.

1. Montrer que $p(\theta|y) = p(\theta|\bar{y})$.
2. Montrer que la loi a posteriori du paramètre θ est donnée par

$$\theta|y \sim N(\bar{y}, \sigma^2/n)\,.$$

Exercice 3. On considère le modèle gaussien dont la vraisemblance est donnée par

$$p(y_1, \ldots, y_n|\theta) \propto \frac{1}{\theta^{n/2}} \exp\left(-\frac{n}{2\theta}s_n^2\right)$$

où

$$s_n^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - m)^2\,.$$

Le paramètre θ représente la variance, qui est inconnue. On suppose que la moyenne m est connue et que la loi de θ est dégénérée (uniforme sur une échelle log)

$$p(\theta) \propto \frac{1}{\theta}\,.$$

On pose $\bar{y} = \sum_{i=1}^{n} y_i/n$.

1. On considère une variable aléatoire de loi χ^2_ν, dont la densité est donnée par

   $$p(x) \propto x^{\nu/2-1}e^{-x/2}\,, \quad x > 0, \nu > 0\,.$$

   Soit $s^2 > 0$. Montrer que la loi de la variable $\nu s^2/\chi^2_\nu$ admet pour densité

   $$p(x) \propto \frac{1}{x^{\nu/2+1}}e^{-\nu s^2/2x}\,, \quad x > 0\,.$$

   Cette loi est notée Inv-$\chi^2(\nu, s^2)$.
2. Montrer que la loi a posteriori du paramètre θ est donnée par

   $$\theta = \sigma^2|y \sim \text{Inv}\chi^2(n, s^2_n)$$

Exercice 4.

Désormais, m et σ^2 sont des paramètres inconnus et aléatoires. On considère donc $\theta = (m, \sigma^2)$, de loi a priori non-informative

$$p(m, \sigma^2) \propto \frac{1}{\sigma^2}\,.$$

1. Déterminer la vraisemblance du paramètre θ. En déduire que la loi a posteriori s'écrit

   $$p(m, \sigma^2|y) \propto \frac{1}{(\sigma^2)^{n/2+1}} \exp\left(-\frac{1}{2\sigma^2}((n-1)s^2_{n-1} + n(\bar{y} - m)^2)\right)$$

   où s^2_{n-1} est l'estimateur sans biais de la variance :

   $$s^2_{n-1} = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

2. Montrer que la loi marginale a posteriori de σ^2 est

   $$\sigma^2|y \sim \text{Inv}\chi^2(n-1, s^2_{n-1})\,.$$

3. Justifier la validité de l'algorithme de simulation de la loi a posteriori suivant

   ```
   sigma.2 = (n-1)*var(y)/rchisq(10000, n-1)
   m = rnorm(10000, mean(y), sd = sqrt(sigma.2/n))
   ```

4. Programmer cet algorithme en langage `R`, et l'utiliser pour estimer les paramètres de moyenne et variance des longueur des sépales des iris de Fisher. Comparer les résultats obtenus aux résultats théoriques.
5. Justifier la validité de l'algorithme de simulation de la loi a posteriori répétant les opérations suivantes (depuis une condition initiale fixée)

   ```
   sigma.2 = sum((y -m)^ 2)/rchisq(1, n)
   m = rnorm(1, mean(y), sd = sqrt(sigma.2/n))
   ```

6. Programmer cet algorithme en langage `R`, et comparer les résultats obtenus aux résultats théoriques pour le même jeu de données que précédemment.