 Hicham Janati - hjanati@insea.ac.ma

TD/TP5: Modèle de capture-recapture

Le contexte typique dans lequel le modèle de capture-recapture est utilisé est celui d'une population animale dans la nature. Imaginons un biologiste qui souhaite estimer le nombre de poissons dans un étang. Pour ce faire, il va capturer un échantillon de poissons, les marquer (par exemple, en leur attachant une petite étiquette) puis les relâcher dans l'étang. Après un certain laps de temps, il va revenir et capturer à nouveau un échantillon de poissons. En examinant combien de poissons marqués sont présents dans le deuxième échantillon, le biologiste peut estimer la taille totale de la population de poissons dans l'étang.

Soit N la taille inconnue de la population totale que nous cherchons à estimer.

1. Lors de la première capture, un échantillon de n_1 poissons est prélevé dans la population, qui sont tous marqués par le biologiste.

2. Lors de la deuxième capture, un nouvel échantillon de n_2 poissons est prélevé dans la population, le biologiste à présent ne fait que compter le nombre de poissons n_{12} qui ont été déjà marqués à la première capture.

On observe donc n_1, n_2 et $n_{12} \leq n_2$. On cherche à estimer N. Les valeurs numériques sont: $n_1 = 125$, $n_2 = 110$ et $n_{12} = 15$.

1 Modèle à un paramètre

On ignore la probabilité de capture (le biais de séléction) et on s'intéresse uniquement à la modélisation de la recapture des poissons marqués à l'étape 2. On suppose donc que n_1 et n_2 sont des constantes données.

1. La loi hypergéométrique de paramètres (M, K, n) modélise le nombre de succès parmi n tirages sans remplacement dans une population de taille M où le nombre total de succès possible est K. Sa densité est donnée par:
$$\mathbb{P}(X = k) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}.$$
Comment peut-on modéliser $n_{12}|N$?

2. Déterminer une borne inférieure $N_0 \leq N$ en fonction de n_1, n_2 et n_{12}.

3. Déterminer l'estimateur de maximum de vraisemblance en analysant la monotonie d'une suite réelle indexée par N.

4. On se place désormais dans le cadre Bayésien. Quels sont les avantages et inconvénients de prendre une loi apriori uniforme sur $[|N_0, +\infty|]$?

5. Prenons à présent une loi a priori uniforme donnée par $\mathbb{1}_{[N_0, N_{\max}]}$ où $N_{\max}$ est un nombre supposé assez grand. Comment peut-on obtenir des statistiques a posteriori sur N numériquement ?

6. Implémentez le modèle avec `pymc`, faites le diagnostic MCMC et interprétez les résultats.

7. Comparez avec l'approche fréquentiste

2 Modèle à deux paramètres

On modélise à présent la probabilité de capture d'un poisson par un paramètre p. n_1 et n_2 et n_{12} sont désormais des variables aléatoires. Les paramètres d'intérêt sont donc (p, N). On pose ${n_2}' \stackrel{\text{def}}{=} n_2 - n_{12}$ le nombre de poissons "nouveaux" c-à-d capturés à l'étape 2 et non capturées à l'étape 1. On suppose que chaque poisson a une probabilité p d'être capturé en toutes circonstances. On modélise les nombres de capture avec la loi binomiale:

- marqués à l'étape 1: $n_1 \sim \mathcal{B}(N, p)$.
- marqués à létape 2: n_{12} parmi les poissons marqués n_1: $n_{12}|n_1 \sim \mathcal{B}(n_1, p)$.
- non marqués à l'étape 2: ${n_2}'$ parmi les poissons non marqués $N - n_1$: ${n_2}'|N, n_1 \sim \mathcal{B}(N - n_1, p)$.
- Définir $n_2 = n_{12} + {n_2}'$.

1. On suppose des loi a priori uniformes sur $p \sim \mathcal{U}([0, 1])$, et $N \sim \mathcal{U}([|N_0, +\infty|])$ supposés indépendants. Tracez le graphe probabiliste de ce modèle.
2. Implémentez ce modèle avec `pymc`, faites le diagnostic MCMC et interprétez les résultats.
3. Déterminer la distribution a posteriori de $p|N$
4. De même pour $N|p$. Pouvez-vous la reconnaitre ?
5. La loi binomiale négative est une distribution de probabilité discrète définie par deux paramètres : le nombre de succès r et la probabilité de succès p. Elle modélise le nombre d'échecs indépendants nécessaires pour obtenir exactement r succès, sachant que la probabilité de succès dans chaque essai est p. Sa densité est donnée par:

$$P(X = k) = \binom{k + r - 1}{k} \times p^r \times (1 - p)^k$$

   Notons $n_+ = n_1 + n_2 - n_{12}$. Montrez que $N - n_+|n_+, p$ suit une Bionomiale négative et précisez ses paramètres.
6. Proposez une meilleure procédure que `pymc` pour simuler la loi a posteriori.
7. Comment se comparent les deux implémentations ?