Méthode statistique de Capture-Recapture : Du dénombrement de cure-dents au dénombrement de poissons dans un étang

Maxime BOUCHER*

Niveau : Tout public
Durée : Adaptable - Au minimum 15 minutes.

Abstract

L'objectif de cet atelier est d'appréhender la notion d'estimateur en statistique et de la méthode statistique de capture-recapture pour dénombrer une population

Matériel

Un récipient opaque avec une ouverture (typiquement une bouteille de lait),
un lot de cure-dents,
un marqueur/feutre.

Quelle problématique ?

Qui ne s'est jamais retrouvé autour d'une étendue d'eau, un week-end, dans l'espoir de pêcher quelques poissons ? Et surtout, qui n'est jamais rentré chez soi sans aucuns poissons dans ses filets ? On est alors en droit de se demander : "Mais est-ce qu'il y a vraiment des poissons ici ?". La question est légitime mais la méthode pour y répondre est à décrire. La population à estimer est mouvante, on ne peut donc pas créer un découpage de l'espace pour ensuite estimer le nombre moyen de poissons par parcelles. On retire la possibilité de vider l'étendue d'eau pour compter à vue les poissons. Comment peut-on faire dans ce cas ?

*Université Libre de Bruxelles

La méthodologie

On peut supposer que l'on dispose : d'une méthode pour marquer les poissons (sans les blesser), d'une canne pour pêcher un poisson, et de temps. On suppose que les poissons sont pêchés un à un et remis à l'eau. Le nombre $N \in \mathbb{N}^*$ de poissons au total dans l'étendue d'eau est inconnu.

La **méthode de capture-recapture** repose sur la technique suivante :

1. Une première session (Capture) : on pêche, un à un, m poissons que l'on marque ($m \in \mathbb{N}^*$).
2. Une phase d'attente : une pause pour que les poissons reprennent leur répartition "habituelle".
3. Une deuxième session (Re-Capture) : on pêche, un à un, $n \in \mathbb{N}^*$ poissons. Sur ces n poissons, on note le nombre de poissons marqués observés.

Présentation au public

L'avantage de la méthode de Capture-Recapture est son accessibilité par le public. Pour un public jeune (collégiens, lycéens), on peut modéliser le dénombrement d'une population de poissons avec des cure-dents. Pour cela, on a besoin du matériel suivant :

- Un **récipient opaque** avec une ouverture (typiquement une bouteille de lait).
- Un **lot de cure-dents**.
- Un **marqueur/feutre**.

Les poissons sont modélisés par les cure-dents que le présentateur place dans le récipient. A noter que le nombre de cure-dents n'est connu que de lui. On crée une ouverture dans le récipient de la taille d'un cure-dents. En effet, dans la méthode on pêche un poisson à la fois, il est donc important de ne pouvoir sortir qu'un cure-dent à la fois.

On donne le récipient à un expérimentateur (un élève par exemple) : il peut effectuer l'expérience de la première session. Il tire un cure-dent, le marque, le redépose dans le récipient, mélange et recommence m fois. Le temps de pause est modélisé par un mélange appuyé du récipient. Puis on effectue la deuxième session où on observe notre n-échantillon.

Remarque 1 *L'avantage de cette expérience, c'est que l'on peut faire appréhender le rôle du m. Plus il est grand, mieux ce sera. En revanche, on constate qu'avoir un m grand devient rapidement fastidieux, il y a donc un compromis à faire entre "avoir un m grand" et "ne pas perdre trop de temps" (on peut faire un parallèle avec le coût de marquage d'une population dans la vie réelle).*

Remarque 2 *A noter que cette méthode peut servir à dénombrer des poissons dans une étendue d'eau, mais aussi d'autres populations animales.*

La modélisation statistique

Au cours de la première session, on fixe le paramètre m qui peut être vu comme une variable ajustée par l'expérimentateur. A la fin de la deuxième session, on dispose d'un n-échantillon que l'on note $(X_1, \ldots, X_n)$ où $X_i = 1$ si le i^{ieme} poisson est marqué et $X_i = 0$ sinon. On a :

$$\forall i \in [\![1, n]\!],\ X_i \overset{i.i.d}{\sim} \mathcal{B}(p) \tag{1}$$

La probabilité $p \in]0, 1[$ du modèle de Bernoulli est la même pour tous les X_i puisque l'on remet à l'eau tous les poissons. Par définition, au cours de la deuxième session, la probabilité de pêcher un poisson marqué vaut :

$$p = \frac{\text{nbr poissons marqués}}{\text{nbr total de poissons}} = \frac{m}{N} \tag{2}$$

Cela implique que la probabilité du modèle est inconnue (puisque N l'est). On utilise notre échantillon pour construire un estimateur de p, puis en déduire un estimateur de N. D'après la Loi des Grands Nombres (LGN), pour un échantillon provenant d'un modèle de Bernoulli :

$$\overline{X}_n := \frac{1}{n}\sum_{i=1}^{n} X_i \underset{n\to+\infty}{\overset{\mathbb{P}}{\longrightarrow}} \mathbb{E}[X_1] = p \tag{3}$$

On pose alors $\overline{X}_n$ notre estimateur (convergent) de p. Autrement dit, on sait que pour n grand, la valeur observée de $\overline{X}_n$ approxime la valeur inconnue de p. Ainsi, on peut facilement revenir à la quantité d'intérêt N :

$$\overline{X}_n \approx p = \frac{m}{N} \Leftrightarrow N \approx \frac{m}{\overline{X}_n} \tag{4}$$

D'après (4), on pose de manière naturelle $\hat{N}_n = \frac{m}{\overline{X}_n}$ l'estimateur de N. Cependant, il est possible que $\overline{X}_n = 0$ sur certaines expériences. On obtiendrait alors une estimation $N = +\infty$ qui n'a pas de sens. On doit donc modifier notre estimateur. Une manière de procéder est la suivante, on considère l'estimateur :

$$\hat{N}_{n,2} = \begin{cases} \frac{m}{\overline{X}_n} & \text{si } \overline{X}_n \neq 0 \\ m+1 & \text{si } \overline{X}_n = 0 \end{cases} \tag{5}$$

Cet estimateur traduit simplement le fait que si on ne pêche aucun poisson marqué, on ne peut obtenir d'estimation précise. On pose alors comme estimation $m+1$ que l'on sait volontairement fausse. Cependant, on peut montrer par le calcul que la probabilité $\mathbb{P}(\overline{X}_n = 0) \underset{n\to+\infty}{\longrightarrow} 0$. Si on considère un n suffisamment grand, alors l'éventualité de devoir donner comme estimation $m+1$ diminue. Si

on étudie cet estimateur, on peut montrer que c'est un estimateur **biaisé** pour n fixé (avec un biais positif, ce qui signifie que notre estimation sera supérieure à la vraie valeur de N), mais il est **asymptotiquement sans biais**. Autrement dit, pour des n (très) grands, la différence entre l'estimation et la vraie valeur diminue. De plus, en se basant sur la construction d'un intervalle de confiance au niveau $1-\alpha$ (où $\alpha \in]0,1[$ est le risque, traditionnellement $\alpha = 0.05, 0.01, 0.005$) de la proportion p, on peut construire un intervalle de confiance au niveau $1-\alpha$ pour N :

$$\left[\frac{\hat{N}_{n,2}\sqrt{n}}{\sqrt{n} + q_{1-\frac{\alpha}{2}}\sqrt{Q_n}}; \frac{\hat{N}_{n,2}\sqrt{n}}{\sqrt{n} - q_{1-\frac{\alpha}{2}}\sqrt{Q_n}} \right] \tag{6}$$

où $Q_n = \frac{\hat{N}_{n,2}}{m}\left(1 - \frac{m}{\hat{N}_{n,2}}\right)$ et où $q_{1-\frac{\alpha}{2}}$ est le quantile de la loi normale centrée réduite $\mathcal{N}(0,1)$ à l'ordre $1-\frac{\alpha}{2}$.

Remarque 3 *On peut présenter différents estimateurs pour étudier ce problème. Par exemple, on peut considérer l'estimateur* $\tilde{N} = \frac{m(n+1)}{n\overline{X}_n+1}$. *Cet estimateur n'a plus le problème de division par zéro et possède des propriétés (de biais notamment) plus intéressantes que* $\hat{N}_{n,2}$ *qui est l'estimateur intuitif.*

Enfin, afin d'illustrer au mieux les estimateurs $\hat{N}_n$, $\hat{N}_{n,2}$ ou même $\tilde{N}$, on peut facilement construire un petit code R ou Python qui prend comme entrée la taille de l'échantillon désirée n, le paramètre m ainsi que la valeur N à estimer. Cela permet d'illustrer la notion de "convergence" d'un estimateur (et donc d'illustrer la LGN) mais aussi d'appuyer le lien entre statistique et algorithme qui est inévitable de nos jours.