Thème 3–1 : Estimation d'un effectif par échantillonnage

1 Principe de la méthode CMR

On suppose qu'on doit trouver le nombre d'individus d'une certaine espèce et qu'on ne peut pas les compter directement. On va noter ce nombre N. Comme on ne peut pas compter les individus, on ne peut pas connaître **exactement** la valeur de N. Mais on va obtenir une **estimation** de cette valeur en utilisant la méthode **Capture-Marquage-Recapture (CMR)**.

1. On capture un nombre d'individus M que l'on marque (balise sur la nageoire d'un poisson, bague en métal à la patte d'un oiseau ou sur l'oreille d'un rongeur, etc.).
2. On relâche ces individus marqués dans leur milieu.
3. On recapture n individus dans le même milieu et on compte les individus marqués parmi ces n individus. On note m le nombre d'individus marqués après la recapture.

On va alors considérer que la proportion d'individus marqués dans l'ensemble du milieu est la même que la proportion d'individus marqués parmi les individus obtenus lors de la recapture. En d'autres terme, on considère

$$\frac{m}{n} = \frac{M}{N}.$$

Comme on connaît déjà M, m, n, on peut retrouver N par un calcul de proportionnalité :

$$N = \frac{M \times n}{m}.$$

Exemple 1

On veut compter le nombre de "Tarier de la Réunion" (une espèce d'oiseau) sur l'île de la Réunion. Il s'agit d'un oiseau qui vit uniquement sur cette île et qui est appelé "tec-tec" là bas. On capture ainsi 800 tec-tecs, que l'on marque avec une bague en métal à la patte, puis que l'on relâche ensuite dans la nature. On recapture ensuite 200 tec-tecs, et on dénombre 17 oiseaux marqués. Dans ce cas, on évalue le nombre total d'oiseaux sur l'île à

$$N = \frac{M \times n}{m} = \frac{800 \times 200}{17} \approx 9412$$

On peut ainsi estimer la population de tec-tecs à 9412 individus. Bien sûr il ne s'agit que d'une estimation, nous ne pouvons pas être sûr de ce nombre à moins de compter tous les oiseaux, mais on a déjà dit que c'était impossible.

Remarque

Attention, pour que la méthode CMR s'applique, il y a quelques conditions à respecter :

- après le marquage, il faut laisser le temps aux individus marqués de bien se mélanger dans toute la population.
- Le milieu doit être "fermé" : il ne doit pas y avoir d'arrivées ou de départs d'individus. Ainsi la recapture doit aller assez vite.
- Les individus marqués ne doivent pas être effectés et les marques ne doivent pas être perdues.

2 Fluctuation d'échantillonnage et intervalle de confiance

Dans la méthode CMR, on fait l'hypothèque que la proportion

$$f = \frac{m}{n}$$

d'animaux marqués dans l'échantillon obtenu après recapture est égale à (ou en tous cas proche de) la proportion

$$p = \frac{M}{N}$$

d'animaux marqués dans l'ensemble du milieu étudié. La proportion $f = \frac{m}{n}$ s'appelle la **fréquence observée** d'animaux marqués dans l'échantillon.

En pratique, cette fréquence dépend de l'échantillon.

Exemple 2

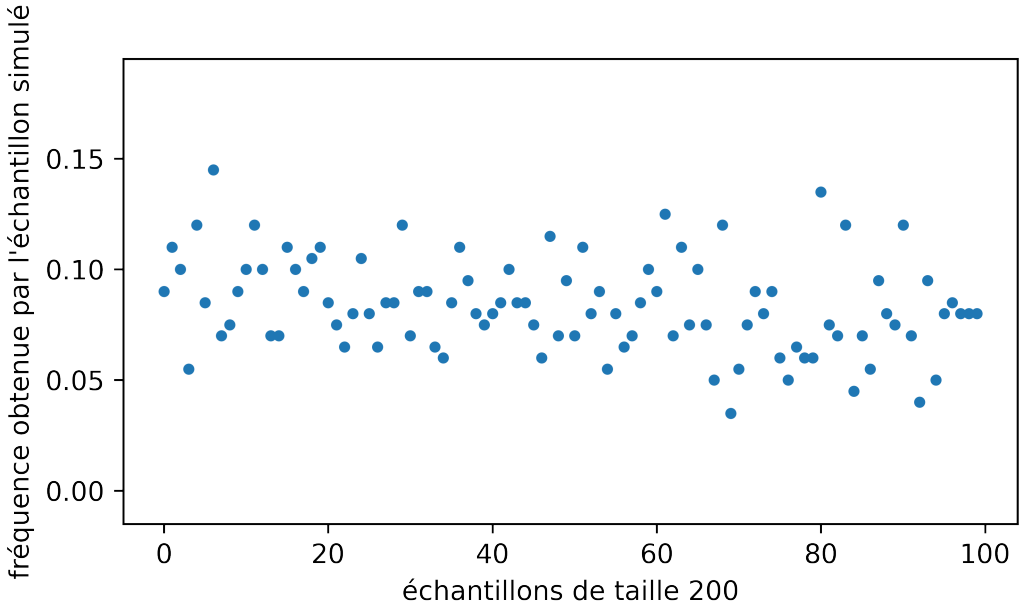
On continue l'exemple des tec-tecs, les oiseaux originaires de la Réunion. On recommence la capture de 200 tec-tecs à 10 reprises différentes et voici ce qu'on obtient.

Capture	1	2	3	4	5	6	7	8	9	10
Nombre de tec-tecs marqués	18	15	10	15	19	17	13	19	18	19
Proportion f de tec-tecs marqués	0.090	0.075	0.050	0.075	0.095	0.085	0.065	0.095	0.090	0.095
Estimation de la population totale N	8889	10667	16000	10667	8421	9411	12307	8421	8889	8421

On voit que l'estimation de la population N de tec-tecs dépend de chacun des échantillons, c'est ce qu'on appelle **la fluctuation d'échantillonnage**. On peut ainsi se demander si toutes les estimation $f = \frac{m}{n}$ obtenues sont de bonnes estimations de

$$p = \frac{M}{N}.$$

Par exemple, la troisième capture, qui ne contenait que 10 oiseaux marqués, semble s'écarter des autres résultats. En fait, on peut simuler plusieurs échantillons (à l'aide du langage Python par exemple), et on voit que les estimations de la proportion p ne sont pas toutes les mêmes. Ci-dessous, on a simulé 100 recaptures de 200 tec-tecs.



3 Intervalle de confiance

En fait, on peut montrer que la plupart des estimations que l'on obtient sont relativement bonnes, et que les estimations deviennent meilleures si l'on augmente la taille de l'échantillon, c'est-à-dire si on capture par exemple 600 ou 1000 tec-tecs au lieu de 200. On peut quantifier la notion de **marge d'erreur** de ces échantillons.

Propriété 1

Dans 95% des cas, l'écart entre la fréquence observée $f = \frac{m}{n}$ et la proportion $p = \frac{M}{N}$ est inférieur à

$$\frac{1}{\sqrt{n}}.$$

Autrement dit, dans 95% des cas, la proportion p appartient à l'intervalle

$$I = \left[\frac{m}{n} - \frac{1}{\sqrt{n}}; \frac{m}{n} + \frac{1}{\sqrt{n}}\right].$$

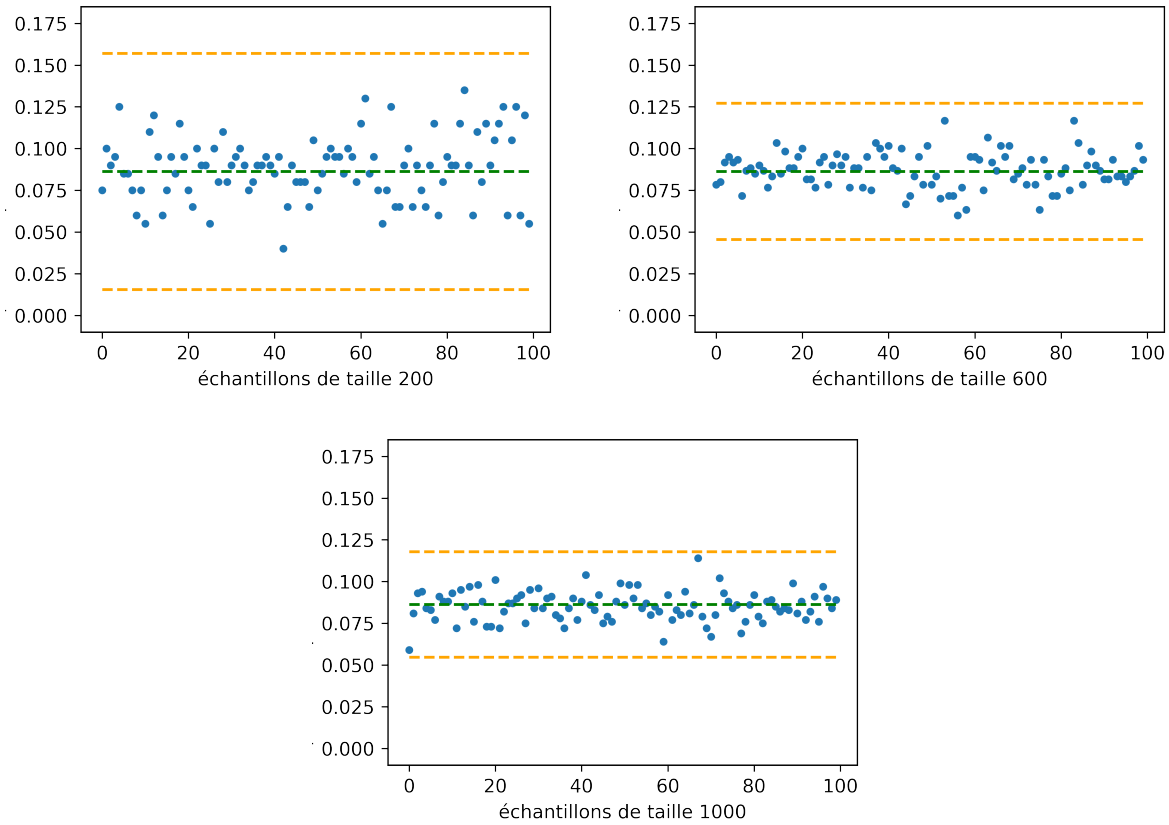
Définition 1 (Intervalle de confiance)

On dit que l'intervalle

$$I = \left[\frac{m}{n} - \frac{1}{\sqrt{n}}; \frac{m}{n} + \frac{1}{\sqrt{n}}\right].$$

est un **intervalle de confiance** au niveau de confiance 95% de la proportion $p = \frac{M}{N}$.

Ainsi, on retrouve bien que la marge d'erreur de $\frac{1}{\sqrt{n}}$ diminue lorsque la taille n de l'échantillon diminue. Cependant, le niveau de confiance de 95% ne change pas lorsque n augmente. Ci-dessous on a représenté les fréquences obtenues en utilisant des échantillons de tailles 200, 600, et 1000. En pointillé vert est représentée la proportion p que l'on cherche à estimer, et en orange on a les valeurs $p \pm \frac{1}{\sqrt{n}}$.



On peut aussi tracer, autour de chaque estimation obtenue, l'intervalle de confiance associé. On remarque, comme prévu, que la valeur p que l'on cherche à estimer est presque tout le temps dans l'intervalle de confiance. On remarque aussi que la taille des intervalles diminue lorsque la taille des échantillon augmente.