THÈME 3
SOUS-THÈME 3-1 : ESTIMATION D'UN EFFECTIF PAR ÉCHANTILLONNAGE

Mots-clés

Capture-marquage-recapture, échantillonnage, intervalle de confiance.

Références au programme

Appréhender la biodiversité et son évolution, ainsi que comprendre l'impact des actions humaines sur celles-ci nécessite de se doter de quelques outils mathématiques pour évaluer des effectifs et des variations de populations. La méthode de « capture-marquage-recapture », couramment répandue pour estimer des populations d'animaux, participe à cette approche scientifique en exploitant les notions de proportionnalité et d'intervalles de confiance.

Savoirs

La méthode de capture-marquage-recapture (CMR) repose sur des calculs effectués sur un échantillon d'une population. Si on fait l'hypothèse que la proportion d'individus marqués est identique dans l'échantillon de recapture et dans la population totale, l'effectif de la population totale s'obtient par le calcul d'une quatrième proportionnelle. Cependant, la proportion d'individus marqués calculée sur un échantillon dépend de cet échantillon. C'est la fluctuation d'échantillonnage. Pour en tenir compte, on assortit la proportion calculée sur un seul échantillon d'un niveau de confiance, toujours strictement inférieur à 100%.

Savoir-faire

Estimer une abondance par la méthode de CMR fondée sur le calcul d'une quatrième proportionnelle.

Simuler des échantillons de même effectif pour visualiser la fluctuation d'échantillonnage.

En utilisant une formule donnée pour un intervalle de confiance au niveau de confiance de 95%, estimer un paramètre inconnu dans une population de grande taille à partir d'un échantillon.

Notions mathématiques travaillées

- Calcul d'une quatrième proportionnelle
- Simulation d'échantillons d'une variable aléatoire de Bernoulli
- Fluctuation d'échantillonnage
- Intervalle de confiance

Histoire, enjeux, débats

La méthode de capture-marquage-recapture utilise des concepts et des outils mathématiques qui relèvent du domaine de la biostatistique. Cette partie des mathématiques permet notamment d'apporter des réponses quantitatives à des questions relatives à la biologie, la médecine, l'environnement et la biodiversité.

La biostatistique prend naissance au XVIIe siècle à la suite du traitement statistique des données démographiques. Son développement au cours de l'histoire a toujours été lié à celui des probabilités et des statistiques. Il bénéficie aujourd'hui des possibilités nouvelles liées au développement de l'informatique (Big Data, IA) : entraînés sur de vastes ensembles de données, des algorithmes dits d'apprentissage automatique (*machine learning*) permettent d'automatiser certaines prises de décisions.

La présentation de la méthode de CMR permet de sensibiliser les élèves à la fluctuation d'échantillonnage et à la notion d'intervalle de confiance dans le cadre particulier de l'estimation de l'effectif d'une population.

Il convient d'indiquer que la fluctuation d'échantillonnage et la notion d'intervalle de confiance interviennent dans d'autres domaines scientifiques (contrôle de qualité), mais également en sciences humaines (fourchettes de sondage).

Description de la méthode

La méthode de CMR permet d'estimer le nombre d'individus dans une population. Pour cela, on procède en deux temps.

Premier temps : capture et marquage

On capture M individus dans une population de taille N inconnue et qu'on souhaite déterminer. On marque (bagues…) ces individus et on les relâche au sein de la population. On note $p = \frac{M}{N}$ la proportion d'individus marqués dans la population ; M étant connu, l'estimation de N revient à celle de p.

Deuxième temps : recapture

On capture un nouvel échantillon de m individus dans la population. On compte le nombre, noté , d'individus marqués dans cet échantillon. On note $f = \frac{m}{n}$ la proportion (ou fréquence) d'individus marqués dans l'échantillon de recapture. Si on admet que p et f sont égaux, alors un calcul de quatrième proportionnelle donne : $N = M \times \frac{m}{n}$

Fluctuation d'échantillonnage

L'activité 1 permet de sensibiliser les élèves au fait que la valeur $f = \frac{m}{n}$ retenue comme estimation de p et permettant de calculer N dépend de l'échantillon de recapture.

Un théorème mathématique (le théorème de De Moivre-Laplace, voir la rubrique « complément pour l'enseignant ») permet de quantifier le niveau de confiance que l'on peut accorder à cette estimation en fonction de la taille n de l'échantillon. Pour de grandes valeurs de n, un intervalle de confiance I_c de p, au niveau de confiance de 95 %, peut être donné par la formule :

$$I_c = [f - \frac{1}{\sqrt{n}} ; f + \frac{1}{\sqrt{n}}]$$

Cela signifie que, si on procède un grand nombre de fois à la recapture d'échantillons de même taille n alors, dans au moins 95 % des cas, p appartient à l'intervalle $[f - \frac{1}{\sqrt{n}}\ ;\ f + \frac{1}{\sqrt{n}}]$, intervalle centré sur f et d'amplitude $\frac{2}{\sqrt{n}}$. . On remarque ici que plus l'échantillon est grand, plus l'intervalle de confiance permettant d'estimer p est resserré autour de f et donc plus l'estimation de p par f est précise (mais cela ne change rien au niveau de confiance, qui reste de 95%).

La connaissance de l'intervalle I_c permet d'encadrer la valeur de p, puis d'encadrer celle de N toujours avec un niveau de confiance de 95%.

Dans la rubrique « complément pour l'enseignant », on explique d'où provient mathématiquement cet intervalle de confiance I_c.

L'objectif pédagogique de la séquence d'enseignement construite autour de ce sous-thème est de faire percevoir la fluctuation due à l'échantillonnage et la manière dont on peut quantifier cette fluctuation.

Conditions d'application de la méthode

Pour éviter l'introduction de biais, la méthode de capture-marquage-recapture s'applique à une population fermée n'évoluant pas entre les deux captures (marquage et recapture). Ainsi, il ne faut pas que des individus puissent quitter ou entrer dans la population étudiée, par exemple à l'occasion de flux migratoires. De même, la durée entre la capture de marquage et la recapture doit être suffisamment restreinte pour éviter naissances et décès, mais suffisamment important pour assurer un brassage uniforme des individus marqués parmi l'ensemble de la population.

Dans les faits, le modèle s'applique à certains types de populations : banc de poissons localisé dans une zone géographique, population d'amphibiens implantée aux abords d'un étang, implantation de campagnols dans une prairie, fourmilière isolée, zone de nidification du Fou de Bassan…

Proposition de mise en œuvre d'une séquence d'enseignement

Étape 1 : simulation de la méthode de CMR à l'aide de haricots secs

Un bocal contient un grand nombre de haricots secs. Ce nombre, noté N, n'est pas connu (ou uniquement de l'enseignant) et les élèves ont été sensibilisés à la difficulté de dénombrer la totalité des haricots. L'objectif est d'estimer ce nombre en effectuant plusieurs prélèvements.

1. Phase de capture et de marquage : on prélève M haricots que l'on marque à l'aide d'un feutre. On les remet dans le bocal et on mélange.



2. Phase de recapture : on prélève un échantillon de n haricots et on compte parmi eux le nombre m de ceux qui sont marqués.

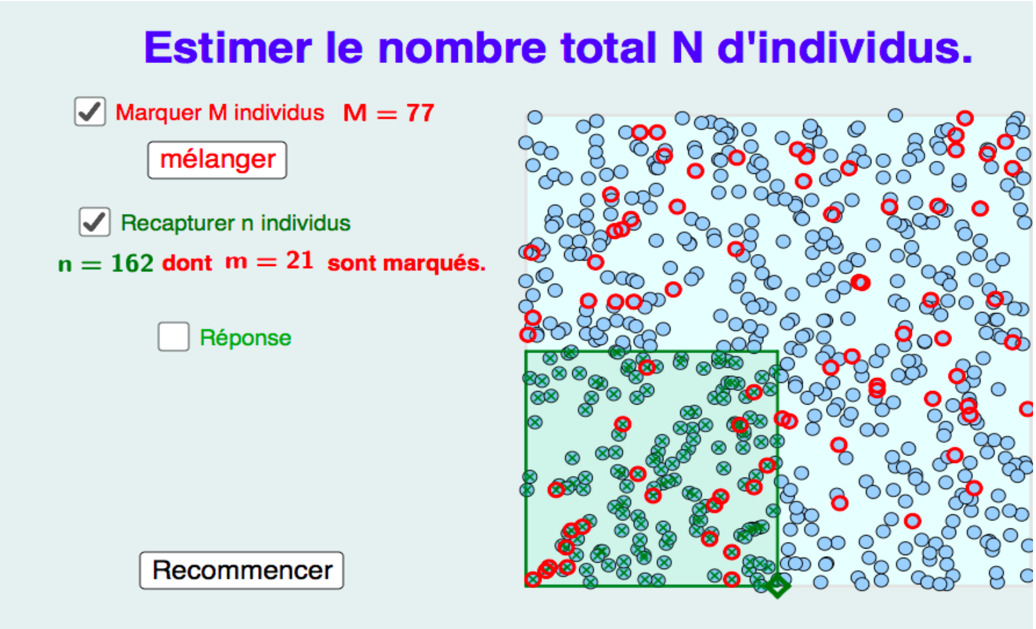
3. Exploitation des résultats : calcul de N en utilisant une quatrième proportionnelle.
4. Sensibilisation à la fluctuation des échantillons.

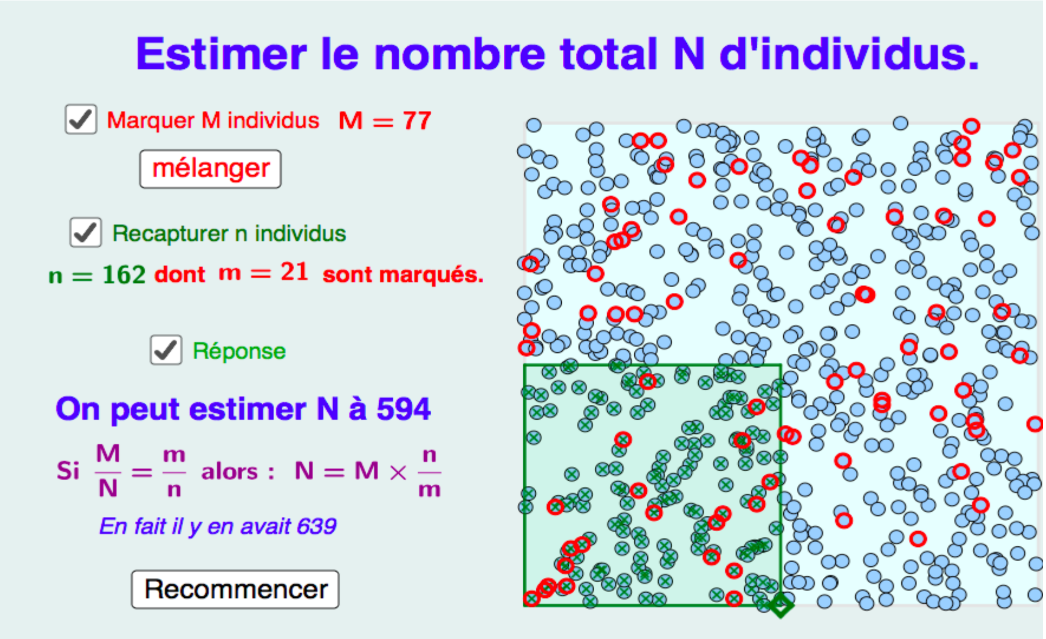
Le deuxième temps (phase de recapture) peut être répété plusieurs fois : après avoir remis dans le bocal les n haricots de la recapture, on réitère le prélèvement d'un échantillon de même taille n. On compare les résultats obtenus.

Étape 2 : simulation de la méthode de CMR avec une animation GeoGebra

Télécharger l'animation Geogebra© intitulée « CMR : explication de la méthode ».

Cette animation permet de présenter la méthode de CMR de manière visuelle. Une certaine population (on peut imaginer des poissons dans un étang) est représentée par des points bleus placés aléatoirement, et on cherche à estimer leur effectif inconnu N. L'appliquette permet de capturer et marquer un certain nombre M d'individus, puis d'en recapturer n et de compter le nombre m marqués dans cette recapture. En cochant la case *Réponse*, l'estimation ainsi que la vraie valeur de sont affichées. En cliquant sur le bouton *Recommencer*, une nouvelle population est générée aléatoirement.





Retrouvez éduscol sur

Étape 3 : simulations numériques d'échantillons

Principe

Pour appréhender la fluctuation d'échantillonnage, on simule k échantillons de même taille n d'une variable de Bernoulli de paramètre p. On rappelle qu'une variable aléatoire suivant une loi de Bernoulli de paramètre p prend la valeur 1 avec la probabilité p et la valeur 0 avec la probabilité 1- p

Un échantillon de taille n est un n-uplet $(X_1, X_2, \dots X_n)$ de n variables aléatoires suivant chacune la loi de Bernoulli de paramètre p.

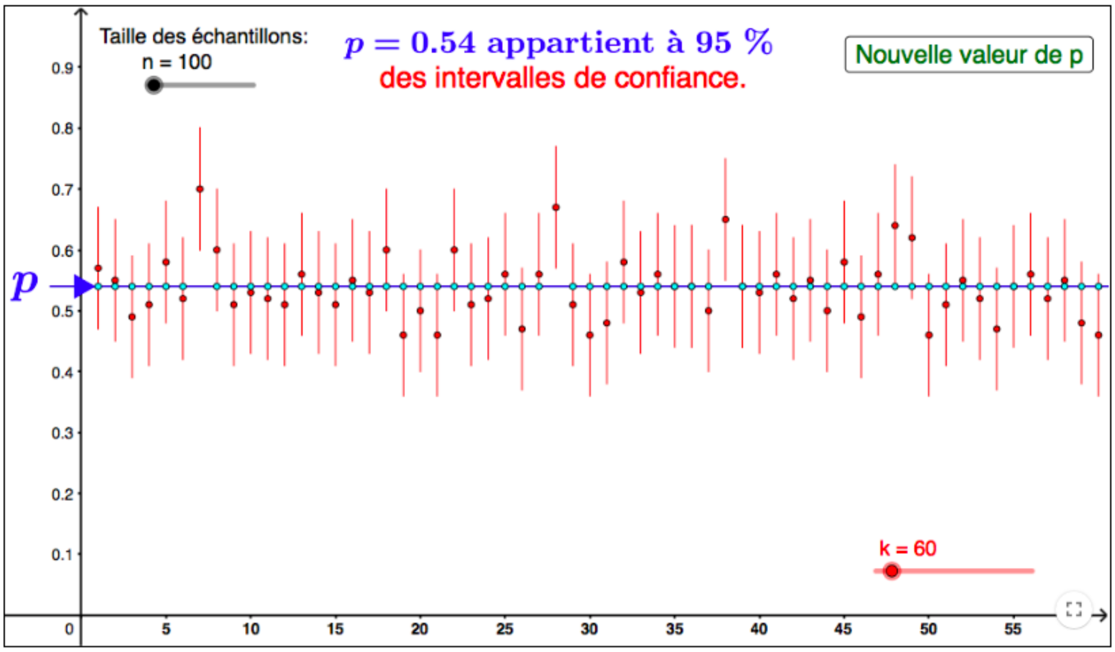
On calcule alors la fréquence d'apparition du 1 dans chacun des k échantillons de taille n, et on représente la liste des k fréquences observées. On observe qu'elles fluctuent autour de p En réalisant cette représentation pour différentes valeurs de n, on observe que l'amplitude de la fluctuation est d'autant plus limitée que la taille des échantillons est grande.

Il importe de faire comprendre aux élèves que la simulation, qui se fait à l'aide d'un générateur de nombres aléatoires compris entre 0 et 1, nécessite la connaissance a priori de la valeur de p (qui représente ici la proportion d'individus marqués dans la population totale). Chaque échantillon de taille n simule un échantillon de recapture. La fréquence d'apparition du 1 dans un échantillon simulé correspond à la fréquence d'individus marqués dans l'échantillon de recapture qu'il simule.

Avec GeoGebra

Télécharger l'animation GeoGebra© intitulée « [Intervalle de confiance et fluctuation d'échantillonnage]() ».

Pour visualiser la fluctuation d'échantillonnage, on fait varier le nombre d'échantillons ainsi que la taille n des échantillons de recapture. La droite horizontale bleue représente la proportion p utilisée pour la simulation. On constate qu'au moins 95 % des intervalles contiennent la valeur de p.



Avec un tableur

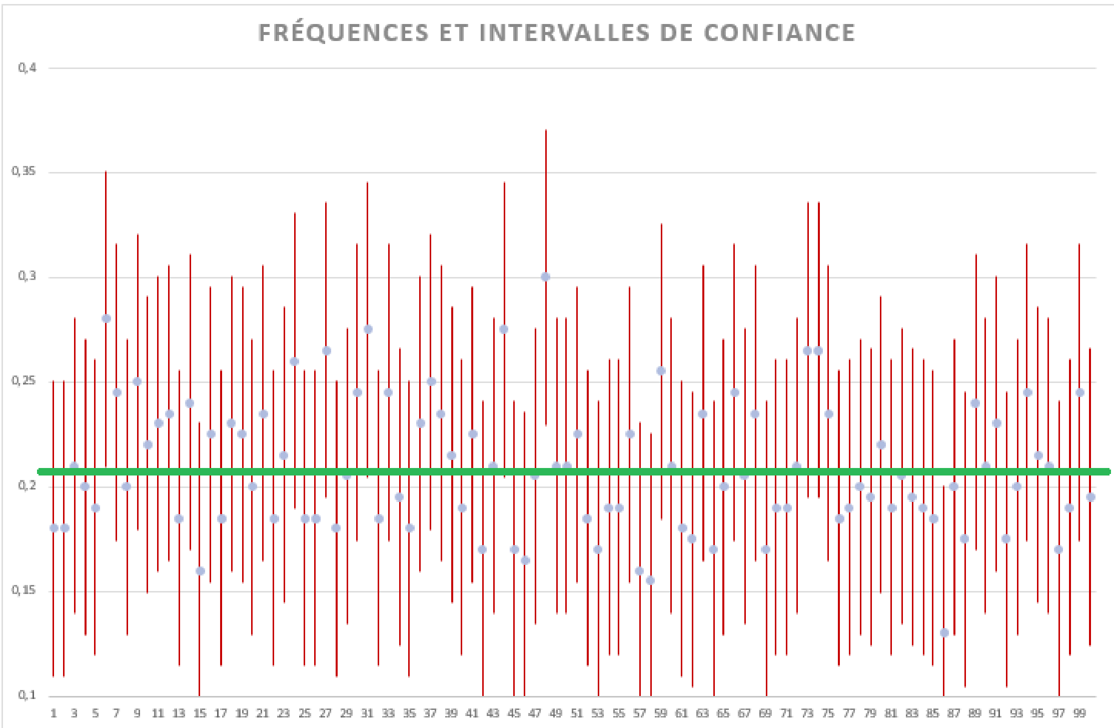
Pour réaliser la feuille de calcul fournie, les principales fonctions utiles sont :

- ALEA.ENTRE.BORNES(min ; max) : fonction qui renvoie un nombre entier entre min et max.
- ENT(ALEA()+p) : fonction qui renvoie 1 avec la probabilité *p* et 0 avec la probabilité 1 - p.
- ET(valeur logique 1 ; valeur logique 2 ;…) : fonction qui renvoie 1 (ou VRAI) lorsque toutes les conditions (mentionnées par le tableur comme des valeurs logiques) sont vraies et 0 sinon.
- NB.SI(plage ; critère) : fonction qui compte le nombre de cellules vérifiant le critère.

On compte le nombre de fois où la proportion appartient à l'intervalle I_c.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	Population totale :		2999		Population marquée :			900	Proportion théorique p :			0,3001		Nombre de VRAI:		97
2																
3	borne sup	0,320711	0,360711	0,275711	0,340711	0,385711	0,365711	0,380711	0,310711	0,390711	0,380711	0,390711	0,400711	0,395711	0,335711	0,390711
4	borne inf	0,179289	0,219289	0,134289	0,199289	0,244289	0,224289	0,239289	0,169289	0,249289	0,239289	0,249289	0,259289	0,254289	0,194289	0,249289
5	fréq.	0,25	0,29	0,205	0,27	0,315	0,295	0,31	0,24	0,32	0,31	0,32	0,33	0,325	0,265	0,32
6	p dans intervalle	VRAI	VRAI	FAUX	VRAI	VRAI	VRAI	VRAI	VRAI	VRAI	VRAI	VRAI	VRAI	VRAI	VRAI	VRAI
7																
8		Ech. 1	Ech. 2	Ech. 3	Ech. 4	Ech. 5	Ech. 6	Ech. 7	Ech. 8	Ech. 9	Ech. 10	Ech. 11	Ech. 12	Ech. 13	Ech. 14	Ech. 15
9	Indiv. 1	0	0	0	0	0	1	0	0	1	0	0	0	0	0	1
10	Indiv. 2	0	1	0	0	0	0	0	1	0	1	1	0	0	0	0
11	Indiv. 3	0	1	0	0	0	1	1	1	1	0	0	0	1	0	0
12	Indiv. 4	0	1	0	0	0	1	0	1	0	0	0	0	0	0	1
13	Indiv. 5	0	1	0	1	1	1	0	0	0	0	1	1	0	0	0
14	Indiv. 6	0	0	1	0	1	0	0	0	0	1	0	1	1	0	0
15	Indiv. 7	0	0	0	0	1	0	0	1	0	0	1	0	0	0	0
16	Indiv. 8	0	0	0	1	0	0	0	0	1	0	1	0	0	0	1
17	Indiv. 9	0	0	0	1	0	1	1	0	0	0	0	0	0	1	0

Représentation graphique du nuage des 100 fréquences observées complétées par leurs intervalles de confiance (en utilisant le diagramme boursier) :



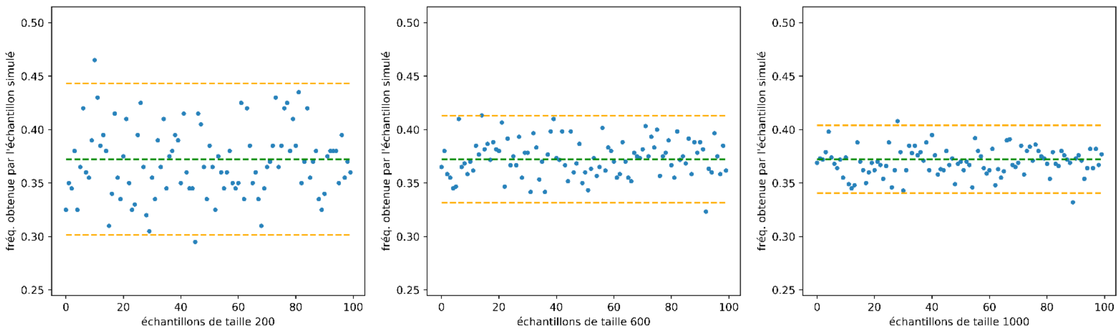
La droite horizontale verte représente la proportion p utilisée pour la simulation. On constate qu'au moins 95 % des intervalles contiennent la valeur de p.

Retrouvez éduscol sur

Il est important d'adapter le choix des outils numériques à la taille des nombres mis en jeu. Ainsi, l'utilisation du tableur peut s'avérer inopérante pour étudier un grand nombre d'échantillons.

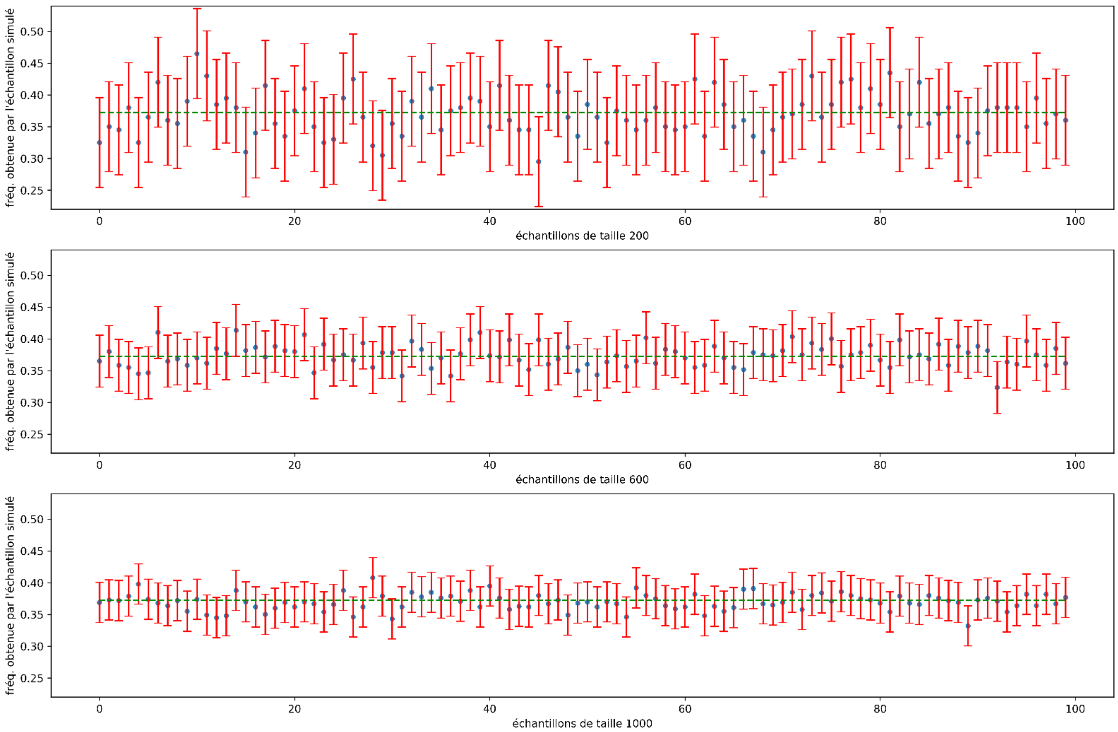
Avec le logiciel Python

On a représenté les nuages des fréquences observées correspondant à des échantillons de recapture de tailles différentes (200, 600, 1000) pour illustrer l'influence de la taille des échantillons sur l'amplitude de la fluctuation.



On observe que la fluctuation est d'autant plus limitée que la taille de l'échantillon est élevée et qu'au moins 95 % des fréquences observées ont un écart à p inférieur ou égal à $\frac{1}{\sqrt{n}}$. .

Comme avec le tableur, on peut programmer avec Python des intervalles de confiance au niveau de 95 %.



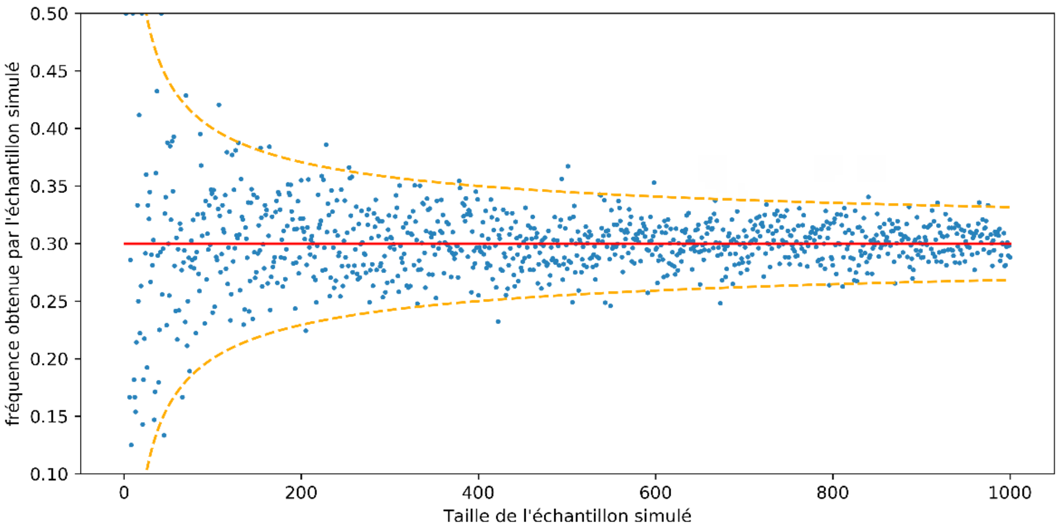
Les simulations précédentes ont permis d'illustrer les deux faits suivants :

- la proportion théorique p se situe, dans au moins 95 % des cas, dans l'intervalle centré sur la fréquence observée et d'amplitude $\frac{2}{\sqrt{n}}$;
- la fluctuation est d'autant plus faible que la taille des échantillons est grande.

Les fichiers sources CMR_multi.ipynb et CMR_multi.html permettent de récupérer les codes des programmes, ils sont disponibles dans le dossier compressé Annexes_CMR_multi.zip.

Évolution de la fluctuation en fonction de la taille des échantillons

Ci-dessous, on a simulé 1000 échantillons, dont la taille n varie de 1 à 1000, d'une expérience aléatoire suivant une loi de Bernoulli de paramètre $p = 0{,}3$. Les courbes oranges en pointillé représentent les fonctions $x \mapsto p - \frac{1}{\sqrt{x}}$ et $x \mapsto p + \frac{1}{\sqrt{x}}$ entre lesquelles se situent une majorité de fréquences observées.



On pourra faire le lien avec la ressource d'accompagnement (n°4) du programme de seconde pour l'algorithmique et la programmation.

Propositions d'activités

Il convient de distinguer les activités issues de données réelles (comptage d'animaux) dans laquelle la taille N de la population est inconnue des activités de simulations basées sur la génération de nombres aléatoires à partir d'une proportion p ou d'un effectif total N connus.

Activité 1 : estimation d'une population par la méthode de CMR.

On souhaite estimer la population de mouettes rieuses (*Chroicocephalus ridibundus*) en Camargue (Gard et Bouches-du-Rhône).

Pour cela, lors d'une première campagne, on capture au hasard sur ce territoire 1 000 mouettes rieuses qui sont baguées puis relâchées.

Lors d'une seconde campagne, quelques temps plus tard, on capture au hasard sur le même territoire 1 200 oiseaux. On constate que sur cet échantillon 239 oiseaux sont bagués.

On suppose que toutes les captures sont indépendantes les unes des autres et que le milieu est clos (population identique lors des deux campagnes de captures).

1. Donner un intervalle de confiance de p au niveau de confiance de 95 % (arrondir les bornes à 10^{-3}).
2. En déduire un encadrement de N au niveau de confiance de 95 %.

Activité 2 : effet de la taille de l'échantillon de recapture

On considère une population de taille N à estimer. Une première capture a permis de marquer $M = 315$ individus, que l'on relâche ensuite dans la population.

1. Lors d'une deuxième capture, on récupère n = 1 600 individus dont m = 112 sont marqués. Donner un encadrement de N au niveau de confiance de 95 %.
2. Répondre à la même question si on avait seulement recapturé 400 individus dont 30 marqués.
3. Commenter ces résultats

Activité 3 : simulations à l'aide de Python

Partie A : simulations de lois

1. Écrire en langage Python une fonction bernoulli(p) simulant une variable aléatoire suivant une loi de Bernoulli de paramètre p, c'est-à-dire une fonction qui renvoie 1 avec une probabilité p et qui renvoie 0 avec une probabilité 1-p.

```
bernoulli(0.4)
```

```
1
```

2. Ecrire une fonction freq_rep_bernoulli(n, p) qui renvoie la fréquence d'apparition de la valeur 1 dans un échantillon de taille n de variables aléatoires suivant une loi de Bernoulli de paramètre p.

```
freq_rep_bernoulli(1000, 0.3)
```

```
0.276
```

3. La fréquence calculée par la fonction freq_rep_bernoulli(n, p) est une estimation de la proportion p à partir d'un échantillon de taille n.
Un intervalle de confiance pour cette proportion, au niveau de 95 %, est :

$$\left[f-\frac{1}{\sqrt{n}}\,;f+\frac{1}{\sqrt{n}}\right]$$

Ecrire une fonction freq_rep_bernoulli_confiance(n, p) identique à la fonction précédente mais qui renvoie en plus les bornes d'un intervalle de confiance au niveau de 95 %.

```
freq_rep_bernoulli_confiance(1000, 0.3)
```

```
(0.284, 0.2523772233983162, 0.31562277660168375)
```

4. Cette question nécessite de connaître la notion de liste en Python.
Écrire une fonction liste_freq_rep_bernoulli(e, n, p) qui renvoie une liste de e éléments.
Chaque élément de la liste contient un résultat de la fonction freq_rep_bernoulli(n, p).

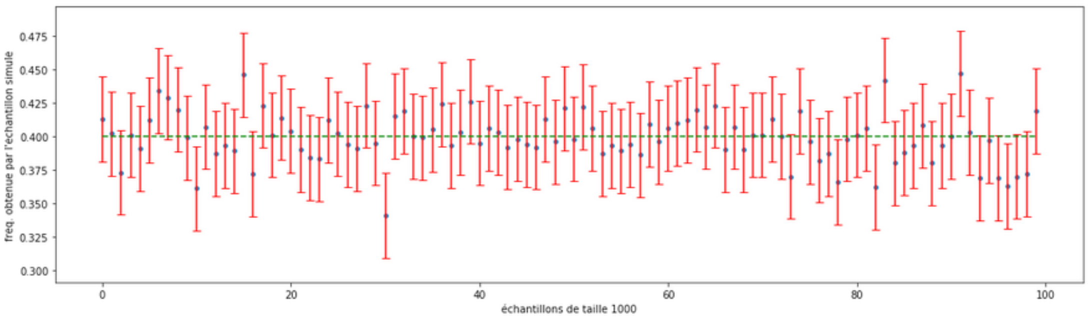
```
liste_freq_rep_bernoulli(5, 1000, 0.4)
```

```
[0.414, 0.39, 0.385, 0.394, 0.416]
```

5. Cette question nécessite de connaître le module matplotlib.pyplot permettant de représenter graphiquement des données.
Il s'agit de réaliser une représentation graphique des fréquences, et de leurs intervalles de confiance, obtenus pour un ensemble d'échantillons.
Pour une partie des élèves, on donnera l'ensemble des éléments et on ne demandera qu'une observation de la figure obtenue.

```
x = list(range(nb_echantillon))
ymin, ymax = min(liste_freq), max(liste_freq)

plt.figure(figsize = (18, 5))
plt.ylim(ymin-0.05, ymax+0.05)
plt.xlabel("échantillons de taille {}".format(taille_echantillon))
plt.ylabel("fréq. obtenue par l'échantillon simulé")
plt.scatter(x, liste_freq, s=14)
plt.errorbar(x, liste_freq, yerr = 1/sqrt(taille_echantillon), fmt = 'none', capsize = 4, errorevery = 1, ecc
plt.plot(x, [p_bernoulli]*nb_echantillon, 'g--')
plt.show()
# plt.savefig('repr_ech_freq_confiance.png', dpi = 600)
```



Retrouvez éduscol sur

Partie B : mise en œuvre de la méthode de CMR

On veut simuler la méthode de capture-marquage-recapture sur une population de taille N inconnue dont 800 individus ont été marqués au cours de la première phase (capture). Pour cela, on choisit au hasard, à l'aide de Python, une valeur pour N qu'on stocke dans une variable effectif_total. Le rapport $p = \frac{800}{N}$ est la proportion d'individus marqués dans cette population.

```
effectif_total = randint(4000, 6000)
marques = 800
proportion = marques/effectif_total
```

1. À quelle étape de la méthode de Capture-Marquage-Recapture correspond la fonction freq_rep_bernoulli(n,p) ?
2. Écrire une fonction eval_pop_totale(n, m, p) qui renvoie une estimation ponctuelle de l'effectif de la population totale à partir de la recapture d'un échantillon de taille n contenant m individus marqués.

```
: eval_pop_totale(200, 800, proportion)
4571
```

3. Compléter la fonction eval_pop_totale(n, m, p) de sorte qu'elle renvoie également les bornes d'un encadrement de l'effectif de la population totale obtenus à partir de l'intervalle de confiance à 95% de la fréquence calculée.

```
: eval_pop_totale(200, 800, proportion)
(3624, 5333, 10090)
```

Les fichiers fournis CMR.ipynb et CMR.html permettent de récupérer les codes sources. Ces documents sont disponibles dans le dossier compressé Annexes_CMR.zip.

Pour aller plus loin

Les statistiques et la médecine

Daniel Schwartz (1917 – 2009) est l'auteur d'un livre intitulé « Le jeu de la science et du hasard » avec comme sous-titre « La statistique et le vivant ». La lecture de ce livre, ou de quelques extraits, permet d'interroger la place des statistiques dans notre société et leur introduction relativement récente dans le domaine des sciences du vivant et notamment de la médecine.

Le livre *Les principes généraux de statistique médicale* (1840) de Jules Gavarret permet d'aborder également la place des statistiques en médecine en adoptant un point de vue historique.

Traitement de données

Les élèves souhaitant développer leurs compétences en informatique et mathématiques peuvent découvrir des packages dédiés au traitement de données en Python et notamment la bibliothèque Panda. Le logiciel R peut également est une piste possible pour les plus férus.

Compléments pour le professeur

Cette rubrique vise à expliquer pourquoi l'intervalle $I_c= \left[f-\frac{1}{\sqrt{n}} ; f+\frac{1}{\sqrt{n}}\right]$ peut être qualifié d'intervalle de confiance au niveau de confiance de 95%.

Le théorème de de Moivre-Laplace (cas particulier du théorème limite central) stipule que, si $X_1, X_2, \ldots, X_n$ sontvariables aléatoires indépendantes et de même loi de Bernoulli de paramètre p alors si on note $F_n = \frac{X_1+X_2+\cdots+X_n}{n}$ la fréquence d'obtention du, alors, pour tout nombre réel strictement positif a,

$$\lim_{n\to+\infty} P\left(\sqrt{\frac{n}{p(1-p)}}\,(F_n-p)\epsilon[-a,a\,]\right) = \frac{1}{\sqrt{2\pi}}\int_{-a}^{a} \mathrm{e}^{-\frac{t^2}{2}}\,\mathrm{d}t$$

Or, les valeurs de la fonction : $g : a \mapsto \frac{1}{\sqrt{2\pi}}\int_{-a}^{a} \mathrm{e}^{-\frac{t^2}{2}}$ sont tabulées et $g(1{,}96) \approx 0{,}95$.

Si, pour n suffisamment grand, on assimile $P\left(\sqrt{\frac{n}{p(1-p)}}\,(F_n-p\,)\epsilon\,[-1{,}96\,;1{,}96]\right)$ et sa limite, on obtient :

$$P\left(|F_n-p| \leq 1{,}96\sqrt{\frac{p(1-p)}{n}}\right) = 0{,}95.$$

Or $1{,}96 < 2$ et $p(1-p) < \frac{1}{4}$.. Donc $1{,}96\sqrt{\frac{p(1-p)}{n}} < 1$ et, comme l'événement $\left(|F_n-p| \leq 1{,}96\sqrt{\frac{p(1-p)}{n}}\right)$ est inclus dans l'événement $\left(|F_n-p| \leq \frac{1}{\sqrt{n}}\right)$, on en déduit que $P\left(|F_n-p| \leq \frac{1}{\sqrt{n}}\right) \geq 0{,}95$.

En pratique, pour un entier fixé, on a observé une valeur numérique explicite f qui est une réalisation de la variable F_n c'est-à-dire qu'il existe une issue ω_0 telles que $f=F_n(\omega_0)$.
Proposer d'estimer la valeur inconnue par l'intervalle de confiance $\left[f-\frac{1}{\sqrt{n}}, f+\frac{1}{\sqrt{n}}\right]$, c'est faire le pari que ω_0 est l'un des éléments de l'évènement $\left\{\omega \in \Omega\,;\ |F_n(\omega)-p| \leq \frac{1}{\sqrt{n}}\right\}$.
Le théorème de de Moivre-Laplace permet d'affirmer que la probabilité de gagner ce pari est d'au moins 0,95.