# Compter les marmottes

*Julien MOUTARD, Vincent ROZAN,*
*Jean MUTILLOD et Thomas DUEZ*

élèves de terminale S du Lycée d'Altitude de Briançon
Enseignant : Hubert PROAL
Chercheur : Camille PETIT (Institut Fourier de Grenoble)

## Sujet

On s'intéresse à une population de $N$ individus où $N$ est inconnu. On souhaite justement connaître $N$ mais il est impossible de compter les individus dans leur ensemble, et donc d'avoir une réponse exacte. En revanche, on est capable de capturer un individu au hasard dans la population. On peut alors décider de marquer ou de ne pas marquer cet individu, et de le relâcher ou non. On peut le faire autant de fois que l'on veut.

Comment peut-on estimer le nombre d'individus dans la population ? Proposer des méthodes, les valider par simulation et essayer de les comparer.

## Note du professeur

L'article squi suit a considérablement évolué au cours de l'année suite aux discussions avec les chercheurs. Il ne présente pas l'évolution des recherches des élèves, mais seulement l'aboutissement de leurs travaux.

## Mots-clés

MARMOTTE, ÉCHANTILLON, EFFECTIF, POPULATION, INFÉRENCE STATISTIQUE, ESTIMATION

## Méthode étudiée

Après plusieurs idées, nous avons proposé une méthode : nous avons une population de $N$ marmottes que l'on souhaite estimer. Nous capturons $M$ marmottes que nous marquons et relâchons.

Nous répétons ensuite $n$ fois la démarche suivante :
– Capturer $C$ marmottes (avec relâche)
– Noter $X_i$ le nombre de marmottes marquées parmi celles-ci, où $i$ est le numéro de notre capture.

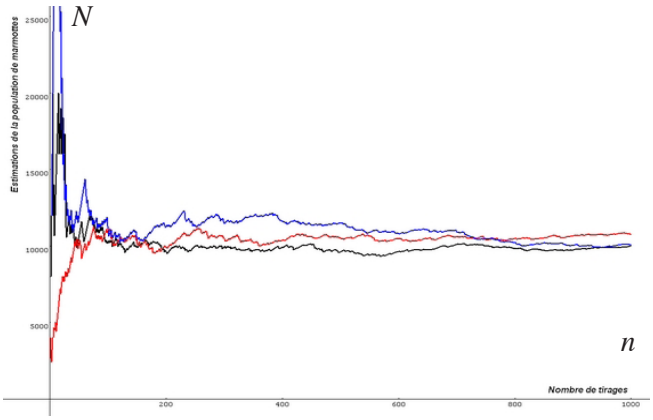
Ainsi nous sommes en mesure d'estimer $N$. Le rapport $X_i / C$ doit être proche de M/N, ce qui donne $\frac{X(i)}{C} \simeq \frac{M}{N}$, soit $N \simeq \frac{C}{X(i)} \times M$. Cette estimation est meilleure si on remplace $X_i$ par la moyenne $\overline{X}$ des $X_i$ sur les $n$ tirages. Voici notre formule :

$$N \simeq \frac{C}{\overline{X}} \times M \qquad \textbf{(1)}$$

Nous remarquons que cette formule dépend de nombreux paramètres ($M$, $C$ et $n$) et nous avons voulu comprendre comment les choisir pour avoir la meilleure estimation possible.

## Étude expérimentale

? Nous avons fixé $M$ = 400 et $C$ = 10 et nous avons fait des simulations [notre graphique montre 3 séries]. Nous avons remarqué qu'à partir d'un certain nombre de tirages $n$, les estimations varient « peu ».

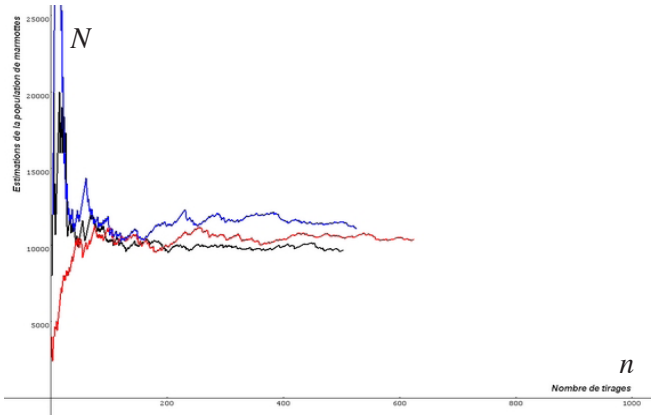



Pour avoir une définition de « varier peu », nous avons proposé le protocole suivant : si pour 100 estimations consécutives, l'écart entre l'estimation maximale et l'estimation minimale est inférieur à 400, alors on arrête la simulation et on estime la population par la dernière valeur obtenue. [Les estimations utilisent toujours la formule (1)]

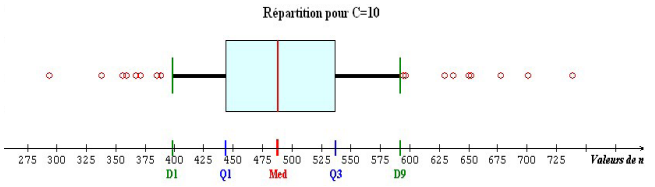
Deux nouveaux paramètres sont apparus, *le nombre d'estimations consécutives* et *l'écart* qui déterminent l'arrêt des simulations. Nous les avons fixés de manière arbitraire, tout comme la taille de la population totale que nous avons fixée à $N$=10 000 pour pouvoir réaliser des simulations. Suite à des discussions avec notre chercheur, nous nous sommes aperçu que le choix d'un écart

d'arrêt de 400 était pertinent dans le cas d'une population de 10000 individus, mais peut-être pas pour une population beaucoup plus grande. Il aurait été plus judicieux de faire une première estimation et de prendre comme critère 4% de cette première estimation.

Avec le critère d'arrêt ci-dessus, voici ce que donnent les graphiques précédents (et surtout la valeur d'arrêt $n$): nous obtenons pour $n$ respectivement les valeurs 503, 624 et 526.




Si nous répétons de nombreuses fois (100 fois) les [séries de] simulations, nous obtenons la répartition des valeurs de $n$ suivante :




Ce qui nous donne comme *neuvième décile* 593,5. Le neuvième décile nous intéresse car il signifie qu'avec ce protocole, pour estimer une population de 10000 individus, nous avons dû compter, dans 90% des cas, moins de 594 paquets de 10 marmottes. Autrement dit, nous avons dû prélever moins de 400+10×594=6340 marmottes dans la population.

Maintenant que le protocole expérimental est déterminé, nous avons voulu tester l'influence des différents paramètres fixés. Nous avons pris d'autres valeurs de $C$ et réalisé différentes simulations avec $M$=400. Voici les résultats d'une [série de] simulation pour chaque valeur de $C$ :

| Valeur de C | Valeur du 9° décile | Valeur de $n$ | Nombre de marmottes comptées | Estimation de la valeur de $N$ |
|---|---|---|---|---|
| 10 | 594 | 387 | 4270 | 9382 |
| 20 | 479 | 455 | 9500 | 10898 |
| 30 | 397 | 280 | 8800 | 8528 |
| 15 | 526 | 425 | 6775 | 11333 |
| 5 | 777 | 685 | 3825 | 9786 |
| 1 | 1499 | 982 | 1382 | 9580 |

? Nous pouvons remarquer que le nombre de marmottes comptées augmente avec C.

Suite au congrès de Gap et aux discussions avec notre chercheur, nous avons décidé de fixer à 4600 le nombre de marmottes comptées ($C$×$n$=4600) et de voir dans ce cas-là l'influence des valeurs de $C$ sur l'estimation de la population.

| Valeur de $C$ | Valeur de $n$ | Nombre total de marmottes prélevées $C$×$n$+$M$ | Estimation de la valeur de $N$ |
|---|---|---|---|
| 1 | 4600 | 5000 | 10575 |
| 5 | 920 | 5000 | 8889 |
| 10 | 460 | 5000 | 9388 |
| 15 | 307 | 5005 | 10407 |
| 20 | 230 | 5000 | 12349 |
| 30 | 153 | 4990 | 8827 |

Les estimations semblent être du même ordre dans chaque cas. Ceci nous fait penser qu'il est inutile de faire des « paquets » de comptage de marmottes, c'est-à-dire que l'on peut prendre $C$=1.

Nous en sommes restés là pour l'étude expérimentale, bien conscients que l'influence de beaucoup de paramètres reste à étudier.

## Essai d'étude théorique

Dans notre protocole, nous avons considéré que si lors de 100 simulations consécutives, nous avions un écart d'estimation inférieur ou égal à 400, alors nous arrêtions le processus.

Lors de nos séances d'atelier, nous avons essayé de comprendre si nous ne pouvions pas estimer de manière théorique la valeur d'arrêt $n$ qui correspond à notre protocole.

On réalise $k$ simulations de tirages de $C$ marmottes. Nous obtenons à chaque nouvelle simulation une [la] moyenne du nombre de marmottes [déjà] marquées, que l'on note $m_k$. Ainsi quand on réalise une simulation supplémentaire, on a :

$$m_{k+1}=\frac{k}{k+1}m_k+\frac{x_{k+1}}{k+1}$$

où $x_{k+1}$ est le résultat de la dernière simulation.

Or $x_{k+1}$ est entre 0 et $C$ (il y a entre 0 et $C$ marmottes marquées dans un paquet de $C$ marmottes), donc

$$\frac{k}{k+1}m_k \leqslant m_{k+1} \leqslant \frac{k}{k+1}m_k+\frac{C}{k+1}$$

Nous avons aussi voulu que l'écart entre la plus grande estimation et la plus petite sur 100 estimations successives soit inférieur à 400. [D'après la formule (1), cela signifie que] si au lieu de 100 nous avions pris 2 nous aurions aussi l'inéquation suivante :

$$\frac{m_k-m_{k+1}}{m_k\times m_{k+1}} \leqslant \frac{400}{M\times C}$$

Nous avons essayé de comprendre comment nous pour-

rions manipuler ces inéquations pour obtenir un résultat théorique mais nous n'avons abouti à aucun résultat.

Lors d'une rencontre avec notre chercheur, ce dernier nous a montré que si l'on fixe le nombre de marmottes que l'on veut compter, par exemple à 4 600, *le choix de C n'intervient pas dans l'estimation*.

## Démonstration

Nous réalisons $n$ expériences identiques qui sont de choisir $C$ marmottes et de compter les marmottes marquées dans le tirage.

On fixe le nombre de marmottes que l'on veut compter, par exemple $n \times C = 4\,600$.

Comme précédemment, notons $X_i$ le nombre de marmottes marquées dans le i-ème tirage de C marmottes.

Nous avons

$X_i = x_i(1) + x_i(2) + ... + x_i(C)$,

où $x_i(j)$ prend la valeur 1 si la j-ème marmotte du paquet numéro i est marquée et 0 sinon.

Ainsi par exemple $X_1 = x_1(1) + x_1(2) + ... + x_1(C)$.

$\overline{X} = \dfrac{X(1) + X(2) + \ldots + X(n)}{n}$, ce qui donne

$\overline{X} = \dfrac{x_1(1) + \ldots + x_1(C) + \ldots + x_n(C)}{n}$ et

$\dfrac{\overline{X}}{C} = \dfrac{x_1(1) + \ldots + x_1(C) + \ldots + x_n(C)}{n\,C}$

où $nC$ vaut 4600.

*Ainsi, le quotient* $\dfrac{\overline{X}}{C}$ *est la moyenne des 4600 valeurs* $x_i(j)$ *et ne dépend donc pas de C. Finalement, notre estimation ne dépend pas de C, il revient au même de compter les marmottes marquées une par une ou par paquets de taille C.*

***