

DÉNOMBREMENT ET ÉCHANTILLONNAGE

Laurent DELSOL

Dans de nombreuses situations on cherche à dénombrer le nombre d'individus, d'animaux, de végétaux ou d'objets se trouvant dans une zone géographique donnée.

Il est parfois possible de faire un décompte exact des individus. Cependant, il est le plus souvent très difficile, voire impossible de dénombrer tous les individus en raison de la grande étendue que l'on doit considérer ou parce que la population est trop importante. Dans ces situations, on cherche plutôt à utiliser des méthodes, appelées méthodes d'échantillonnage, permettant d'avoir une approximation de la taille de la population étudiée en n'observant que certains individus ou certaines zones géographiques. L'objectif de cet atelier et de vous familiariser avec ces méthodes au travers de quelques exemples et de vous présenter les résultats mathématiques sur lesquels elles reposent.

1 Exercice introductif

Chaque participant lance 6 fois 5 dés et note les résultats obtenus. On supposera que les lancers sont indépendants et que les dés ne sont pas truqués. Chaque participant compte le nombre de faces 5 ou 6 qu'il a observé. Il fait également la moyenne des scores obtenus.

- Comparer les résultats obtenus par chacun d'entre vous. D'où proviennent ces différences ?
- Lorsqu'on lance un dé :
 - quelle est la probabilité que l'on tombe sur une certaine face ?
 - quelle est la probabilité de tomber sur 5 ou 6 ?
 - quel est le score moyen auquel on peut s'attendre ?
- Quelle est la proportion globale de 5 ou 6 observés ?
- Quelle est la moyenne globale des scores obtenus ?
- Que remarquez-vous ?

Ce phénomène, appelé loi des grands nombre est à la base de nombreuses méthodes statistiques. Supposons que l'on dispose de n observations indépendantes $(X_1, \dots, X_n)$ (appelées échantillon) d'un phénomène aléatoire X de valeur moyenne m . On peut montrer de manière rigoureuse que sous des conditions générales la moyenne des $(X_1, \dots, X_n)$ converge vers m lorsque n tend vers l'infini. En d'autres termes la valeur moyenne de X qui nous est inconnue peut être approchée par la moyenne de n observations de ce phénomène si n est assez grand. De même, la probabilité d'un événement peut être approchée par la proportion de fois où il a été observé au cours de n observations indépendantes. En pratique, $n > 30$ suffit assez souvent pour obtenir une bonne approximation.

Pour aller plus loin...

Comme nous venons de le voir, la loi des grands nombres fournit une méthode générale permettant d'approcher la moyenne inconnue m du phénomène étudié (respectivement la probabilité P qu'un événement A se produise) à partir de la moyenne $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ de quelques observations indépendantes de celui-ci (respectivement la proportion p_n de fois où il s'est produit parmi n observations indépendantes). Construites à partir d'observations de phénomènes aléatoires, nos approximations présentent une certaine variabilité qui semble diminuer lorsque le nombre d'observations augmente (voir les graphiques présentés).

Il est donc important de donner en plus d'une approximation de m ou de P une idée de la précision de la valeur obtenue. On appelle variance de nos observations l'écart carré moyen à la moyenne défini par

$$\sigma_{n-1}^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2.$$

On peut montrer sous des hypothèses générales que σ_{n-1}^2 converge vers la variance V (écart carré moyen à la moyenne m) du phénomène observé et que la loi de

$$T = \frac{\bar{X} - m}{\sqrt{\frac{\sigma_{n-1}^2}{n}}}$$

peut être approchée lorsque le nombre n d'observations est assez grand par une loi connue appelée loi normale centrée réduite (notée $\mathcal{N}(0, 1)$) [théorème de la limite centrale]. Le fait de connaître cette loi permet de donner des approximations de la probabilité d'événements tels que $\{|T| \leq c\}$ par la probabilité qu'une variable U de loi normale centrée réduite soit en valeur absolue plus petite que c . On sait par exemple que l'on a

$$P(|T| \leq 1.96) \approx 0.95.$$

Ce résultat est intéressant car il permet de donner ce que l'on appelle un intervalle de confiance pour m , c'est à dire un intervalle construit à partir de nos observations dans lequel on peut dire que m se trouve (avec un risque de 0.05 de se tromper). En effet, on a

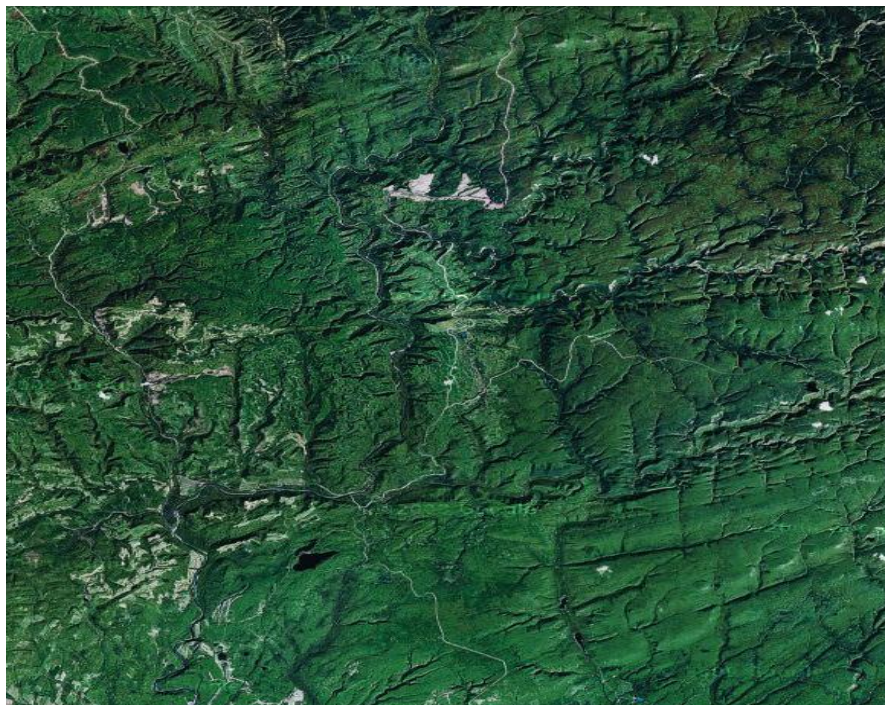
$$P\left(\bar{X} - \sqrt{\frac{\sigma_{n-1}^2}{n}} \times 1.96 \leq m \leq \bar{X} + \sqrt{\frac{\sigma_{n-1}^2}{n}} \times 1.96\right) = P(|T| \leq 1.96) \approx 0.95.$$

On peut voir que la précision augmente lorsque le nombre d'observations croît.

Ce résultat est également intéressant pour construire des tests statistiques portant sur la valeur de la moyenne m .

2 Echantillonnage aléatoire

Voici tout d'abord un premier exemple dans lequel on cherche à évaluer le nombre N d'érables se trouvant dans une forêt s'étendant sur une grande superficie ($32.5 \times 32.5 = 1056.25 \text{ km}^2$). La carte suivante représente les érables se trouvant dans la zone d'étude (image tirée de Google Maps).



Pensez-vous que cela soit envisageable de dénombrer

- tous les érables de cette forêt ?
- les érables se trouvant dans différentes zones d'assez petite taille ?

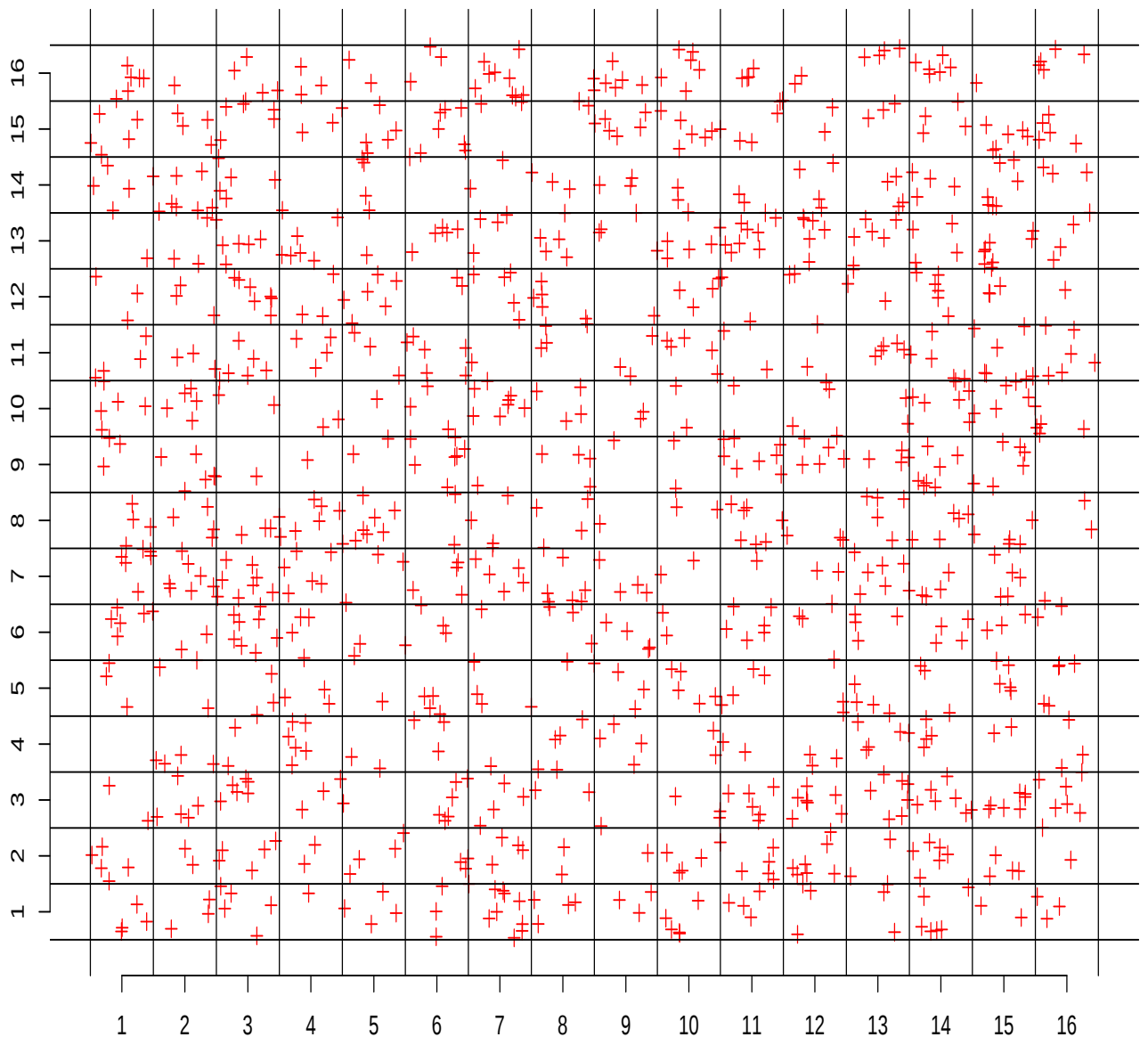
On suppose que l'on peut découper la forêt en 256 zones plus petites et de même forme sur lesquelles on peut compter de manière exacte le nombre d'érables.

- Quel est, en fonction de N , le nombre moyen m d'érables par zone ?
- Comment pensez-vous que l'on puisse donner une approximation de m en observant seulement un petit nombre de ces zones ?
- Comment choisir les zones à étudier ? Combien faut-il en prendre ?

On pourrait choisir de manière déterministe les zones à étudier précisément. Cependant, un choix déterministe de celles-ci peut présenter des défauts. En effet, se restreindre à étudier des zones définies a priori par l'observateur ou suivant une grille régulière peut amener un biais si ces zones ne sont pas représentatives de l'ensemble de la forêt. Afin

de parer à ce problème, on propose souvent de réaliser un échantillonnage aléatoire qui consiste à choisir de manière aléatoire (et avec remise) les zones à étudier. Dans le cas le plus simple où l'on suppose que les arbres sont répartis de manière homogène, toutes les zones ont la même probabilité d'être observées.

Decoupage de la foret en parcelles plus petites



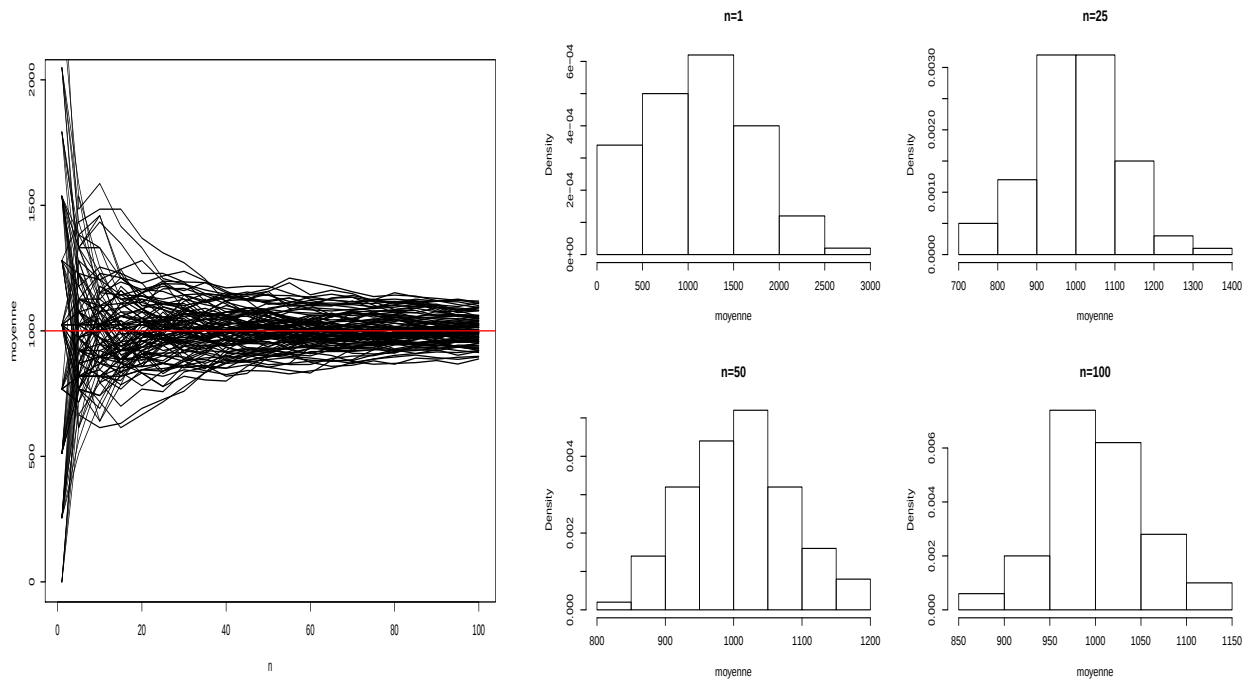
On note X_i le nombre d'érables se trouvant dans la i ème parcelle observée. Comme les parcelles sont choisies au hasard et de manière indépendantes, ces X_i forment un échantillon de variables indépendantes dont la moyenne commune est m . Grâce au phénomène de la loi des grands nombres que nous avons observé au cours de l'exercice introductif, on sait que la moyenne de ces variables donnera une bonne approximation de m .

Exemples de choix aléatoires :

x	6	9	3	16	10	10	6	9	9	2	11	5	4	3	9	11	15	9	2	2
y	8	13	13	11	14	15	10	7	4	7	6	1	10	4	1	11	9	16	10	4
x	12	14	5	6	15	14	11	15	14	5	10	11	8	12	8	6	13	5	6	3
y	13	14	9	12	13	9	1	12	4	16	6	16	10	8	15	1	1	7	10	4

- Compter le nombre d'érables se trouvant dans les zones sélectionnées par cet échantillonnage aléatoire.
- Donner une approximation de m à partir des données recueillies.
- En déduire une approximation de N .

ATTENTION : Il est important de comprendre qu'il y a de la variabilité dans les résultats obtenus en fonction de l'échantillonnage. On peut obtenir de moins bons résultats. Cependant, on peut voir que la variabilité des résultats diminue lorsque la taille de l'échantillon augmente.



On peut aller plus loin :

- calculer la variance de nos observations.
- donner la valeur de l'intervalle de confiance associé à m .
- en déduire un intervalle de confiance pour N .

Remarque : Si la répartition des arbres est vraiment différente dans certaines portions de la forêt (changements de reliefs, de ressources naturelles, présence d'un cours d'eau), il est pertinent de réaliser un échantillonnage stratifié tenant compte de ces variations.

3 Méthode de capture re-capture

On s'intéresse à présent à une autre méthode de dénombrement, appelée capture re-capture, qui n'est pas basée sur un choix aléatoire de parcelles. Elle peut être utilisée dans des situations pour lesquelles la répartition des individus n'est pas uniforme mais nécessite que les individus présents dans une portion de la zone à un instant t_0 donné se soient réparti de manière uniforme sur l'ensemble de la zone étudiée au bout d'un certain temps (à l'instant $t_0 + T$).

Cette méthode a l'avantage de pouvoir être utilisée dans des situations où l'on n'est pas en mesure de compter de manière exacte les individus présents dans une zone donnée (même si sa superficie est petite). C'est notamment le cas lorsque l'on souhaite donner un ordre de grandeur du nombre de poissons se trouvant dans une grande étendue d'eau. Il peut être difficile voire impossible de compter le nombre exact de poissons se trouvant dans une petite portion de cette étendue d'eau en raison de la profondeur, du fait que l'eau est trouble, de la fragilité des poissons, ... Supposons pour simplifier les choses que l'on souhaite faire subir le moins de stress possible aux poissons et que l'on décide de relâcher immédiatement toute prise. On ne peut donc observer qu'un poisson à la fois (cette restriction pourrait être enlevée mais cela compliquerait les calculs et les résultats).

Expérimentation et mise en pratique :

Vous disposez d'un récipient contenant un certain nombre N inconnu d'objets identiques. Vous ne pouvez sortir du récipient qu'un objet à la fois et devez mélanger avant chaque tirage afin de que l'objet sorti la fois précédente ne se trouve pas forcément sur le dessus. Vous avez à votre disposition un feutre indélébile.

- Comment pensez-vous que l'on puisse avoir une idée du nombre total d'objets présents dans la boîte ?
- A quoi peut servir le feutre ?

Imaginons que les objets contenus dans la boîte ne soient pas identiques. Supposons qu'il y ait un nombre connu N_M d'objets portant un signe distinctif.

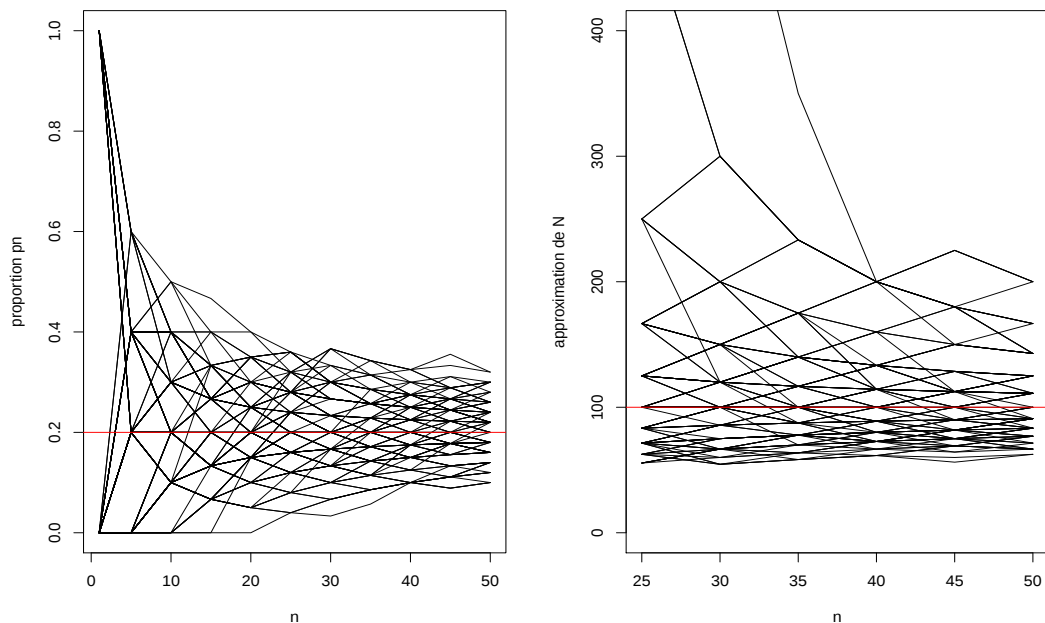
- Quelle serait la probabilité P de choisir au hasard un objet portant un signe distinctif en fonction de N ?

- Supposons que l'on fasse n tirages avec remise dans le récipient en notant si l'objet porte une marque ou non. Comment peut-on donner une approximation p_n de P ?
- Dédire de p_n une approximation de N .

Revenons à notre problème concret. Tous nos objets sont identiques.

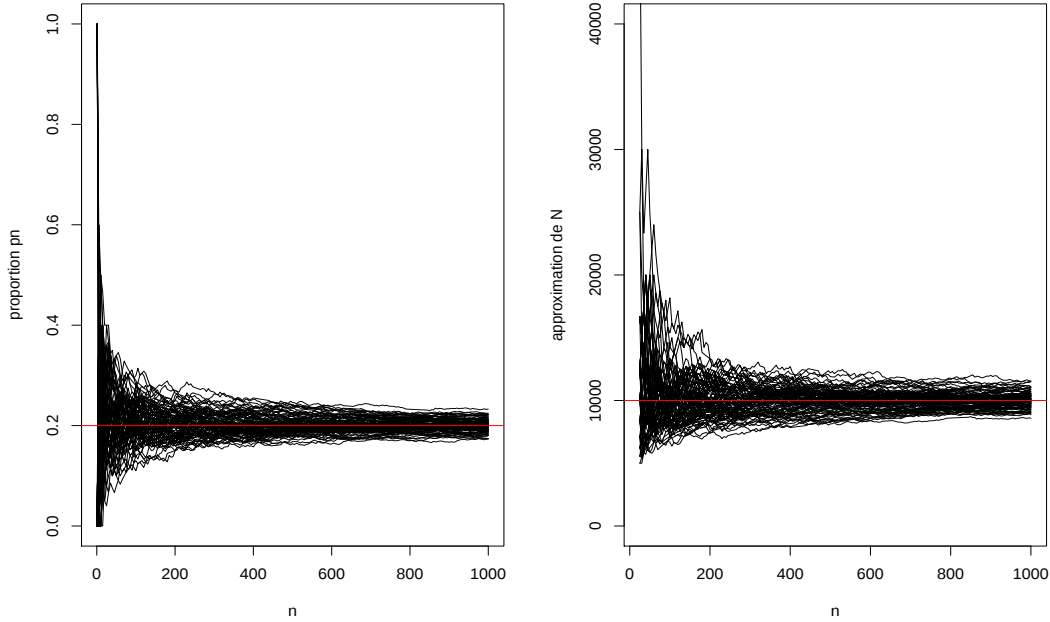
- Proposer une méthode permettant de se ramener à l'étude d'un récipient contenant 20 objets marqués. [ETAPE DE CAPTURE]
- Mélanger les objets pour modéliser la répartition uniforme des individus marqués dans la zone d'étude. [DELAI D'ATTENTE]
- Effectuer ensuite $n = 40$ tirages avec remise. En déduire une approximation de N . [ETAPE DE RECAPTURE]

Il faut bien comprendre ici encore que la nature aléatoire des observations produit de la variabilité au niveau des approximations obtenues. On voit au travers du graphique ci-dessous que de petites erreurs d'approximation concernant la proportion p peuvent conduire à des erreurs assez importantes dans l'approximation de N . Cela vient du fait qu'en général p est petit et que on obtient une approximation de N en divisant N_M par p . On peut cependant montrer que cette variabilité décroît lorsque le nombre d'observations augmente.



Cette méthode est en général utilisée lorsque les nombres N , N_M et n sont plus importants. Dans ce contexte on peut obtenir de meilleurs résultats en termes d'approximation. Voici par exemple ce que l'on peut obtenir sur 100 échantillons en prenant $N = 10000$ et

$N_M = 2000$.



Lorsque le nombre d'observation et la proportion de poissons marqués sont assez grand (c'est le cas dans ce dernier exemple) on peut également utiliser le théorème de la limite centrale pour montrer que

$$P\left(p_n - \sqrt{\frac{p_n(1-p_n)}{n}} \times 1.96 \leq P \leq p_n + \sqrt{\frac{p_n(1-p_n)}{n}} \times 1.96\right) \approx 0.95.$$

On peut voir que la précision de l'intervalle de confiance augmente lorsque le nombre d'observations croît.

Supposons que sur 1000 observations on ait une proportion de 0.199 poissons marqués.

- Donner un intervalle de confiance pour P .
- En déduire un intervalle de confiance pour N .

4 Méthode du maximum de vraisemblance

Une autre méthode appelée maximum de vraisemblance peut être utilisée pour donner une approximation du nombre N d'individus. L'idée principale est de choisir la valeur de N avec laquelle on a le plus de chance d'obtenir les observations qui ont été faites. Prenons

un exemple concret avec peu de données pour illustrer la méthode.

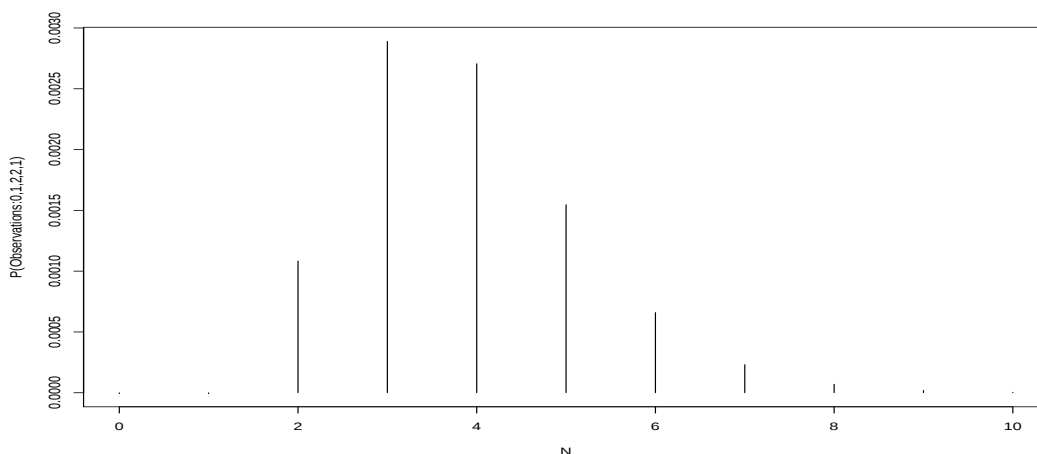
On suppose que l'on cherche à donner une approximation du nombre de rhinocéros se trouvant dans une zone géographique assez vaste. On suppose qu'une étude préalable a montré que les rhinocéros ont chaque jour une probabilité de $\frac{1}{3}$ d'aller se désaltérer à l'un des rares points d'eau en fin d'après-midi. On propose de se rendre à plusieurs reprises autour de ce point d'eau et de compter le nombre de rhinocéros que l'on observe. Pour simplifier nos calculs, on supposera que les rhinocéros ont des comportements indépendants les uns des autres (ce qui peut s'expliquer par le fait qu'ils vivent en solitaire). On suppose également que les jours d'étude sont choisis de manière à avoir des observations indépendantes. On fait les décomptes suivants :

Nombre de rhinocéros	0	1	2	2	1
----------------------	---	---	---	---	---

On note b_i le nombre de rhinocéros observés le jour i . On peut montrer que le nombre de rhinocéros venant s'abreuver en fin d'après-midi un jour donnée peut être modélisé par une variable aléatoire B de loi $Bin(N, \frac{1}{3})$. Par conséquent on peut donner une expression explicite de la probabilité d'avoir observé $b_1, \dots, b_5$ rhinocéros en fonction de N :

$$P_N(\text{Observations} : 0, 1, 2, 2, 1) = \left(\frac{2}{3}\right)^{5N} \frac{N^4(N-1)^2}{2^8}.$$

- Est-ce que N peut valoir 0 ou 1 ?
- A partir du graphique ci-dessous dire pour quelle valeur de N la probabilité est la plus forte que se produise ce que l'on a observé.



- En déduire la valeur de N la plus probable à partir des données.

Cette méthode présente également de la variabilité à cause de l'aspect aléatoire des observations. L'ampleur de cette variabilité décroît lorsque le nombre d'observations aug-

mente. Nous n'en parlerons pas en détail ici.

On peut également s'intéresser à appliquer cette méthode dans des contextes différents. Imaginons par exemple qu'un nouveau participant arrive en fin de séance et qu'il ne sache pas combien de dés vous avez lancé au début. S'il dispose uniquement du nombre de 5 ou 6 observés pour chaque lancé de 5 dés, il peut utiliser la méthode du maximum de vraisemblance pour déterminer à partir des observations la valeur la plus vraisemblable de dés que vous avez jeté. En effet, ici aussi le nombre de 5 ou 6 obtenus à chacun de vos lancers peut être modélisé par une variable de loi $\text{Bin}(N, \frac{1}{3})$, où N représente le nombre inconnu de dés que vous avez lancé.

5 En guise de conclusion

Nous avons vu différentes méthodes statistiques permettant de donner une approximation du nombre d'individus mais aussi de donner un ordre de grandeur de la précision de notre estimation. Comme nous l'avons vu, le caractère aléatoire de ces méthodes amène une certaine variabilité de nos résultats, qui diminue cependant lorsque le nombre d'observations augmente. Il est donc important de prendre en compte la qualité des données (taille de l'échantillon, représentativité) lorsque l'on cherche à interpréter les résultats donnés par une méthode d'estimation (par exemple un sondage). Les méthodes basiques que nous avons vues sont en général modifiées et complexifiées pour fournir de meilleurs résultats.

Prise de notes et calculs intermédiaires