Une expérience ludique de capture-marquage-recapture pour l'initiation au raisonnement probabiliste indispensable au statisticien-modélisateur

Éric Parent, Jean-Jacques Boreux, Etienne Rivot, Sophie Ancelet

Une expérience ludique de capture-marquage-recapture pour l'initiation au raisonnement probabiliste indispensable au statisticien-modélisateur

Éric PARENT [1] & Jean-Jacques BOREUX [2] &
Étienne RIVOT [3] & Sophie ANCELET [4]

[1] *AgroParisTech, UMR 518, 16 rue Claude Bernard, 75005 Paris ; eric.parent@agroparistech.fr*
[2] *Dpt Sc. & G. Environnement, Université de Liège, site d'Arlon, Av de Longwy, 185, B6700 Belgique ; jj.boreux@ulg.ac.be*
[3] *UMR ESE, Ecology and Ecosystem Health, Institut Agro, INRAE, 65 rue de St Brieuc, 35042 Rennes ; etienne.rivot@agrocampus-ouest.fr*
[4] *Institut de Radioprotection et de Sûreté Nucléaire (IRSN), PSE-SANTE/SESANE/LEPID, 92262 Fontenay-Aux-Roses Cedex, France ; sophie.ancelet@irsn.fr*

Résumé. Les méthodes de capture-marquage-recapture sont des méthodes astucieuses d'échantillonnage peu invasives pour évaluer le nombre d'individus dans une population. Utilisées principalement en écologie, elles trouvent aussi des applications de portée bien plus large dans divers domaines tels que la sociologie et la psychologie expérimentales. Du point de vue de la pédagogie, elles permettent d'illustrer de façon simple, pratique et vivante de nombreux points clés du raisonnement probabiliste indispensables au statisticien-modélisateur. A l'aide d'une expérience ludique facile à effectuer en salle avec des gommettes, des haricots secs, une cuillère à soupe et un saladier, nous montrons comment aborder de façon simple et intéressante les points-clés suivants dans le cadre d'un problème d'estimation de la taille inconnue d'une population :

- les ingrédients de base du problème de statistique inférentielle considéré, en particulier, inconnues *versus* observables ;
- la construction d'un modèle probabiliste/stochastique possible, fondé sur l'assemblage de plusieurs briques binomiales élémentaires, ainsi que les différentes décompositions possibles de la vraisemblance associée ;
- la recherche d'estimateurs, leur étude théorique ainsi que la comparaison de leurs propriétés mathématiques par simulation numérique ;
- les différences opérationnelles majeures entre approches statistiques fréquentielle et bayésienne.

Cette expérience permet également d'illustrer en quoi le travail d'un statisticien-modélisateur ressemble bien souvent à celui d'un enquêteur de police....

Mots-clés. Capture-marquage-recapture, estimation, loi binomiale, raisonnement probabiliste, statistique bayésienne.

Abstract. Capture-mark-recapture techniques are smart non-invasive sampling methods to evaluate the number of individuals in a population. Primarily used in ecology, they also find applications with a much broader scope in various fields such as experimental sociology and psychology. From a pedagogical standpoint, they nicely illustrate, in a very simple and practical way, many key points of probabilistic reasoning as a rationale essential to statistician-modellers.

Starting from an affordable toy experiment that can be easily performed indoors with stickers, beans, a tablespoon and a saladbowl, we show how to deal with the following key points in a simple and interesting way, in the specific context of estimating the unknown size of a population :

— the basic ingredients of a problem of inferential statistics, unknowns *versus* observables ;
— the design of a possible probabilistic/stochastic model by assembling several binomial building blocks, as well as the many possible decompositions of the associated likelihood ;
— the search for estimators, their theoretical study as well as the comparative study of their mathematical properties using computer simulations ;
— the major operational differences between the frequentist and Bayesian statistical approaches,

This experiment also illustrates how the everyday work of a statistician-modeller often resembles the one of a police investigator ....

Keywords. Capture-mark-recapture, estimation, binomial distribution, probabilistic reasoning, Bayesian statistics.

1 Introduction

Les stats ? Un pensum ! L'enseignement de la statistique peut avoir laissé un souvenir tristement aride, et ce, même pour les anciens étudiants les plus passionnés par les mathématiques, d'ailleurs peut-être à leur tour devenus enseignants-chercheurs en statistique et/ou statisticiens dans le secteur public ou privé. Depuis le siècle dernier, l'approche pédagogique la plus souvent privilégiée en statistique inférentielle consiste à consciencieusement faire reconnaître par les étudiants *une* situation dans un catalogue de différentes situations typiques afin d'appliquer, par analogie, *la* technique statistique appropriée tirée d'une boîte à outils mathématiques. Ainsi, par exemple, l'estimation de l'effet de covariables sur une variable-réponse binaire appelle souvent systématiquement les étudiants à l'utilisation d'une régression logistique et celle d'une variable-réponse de type comptage à l'utilisation d'une régression de Poisson. À noter que ces outils sont, pour la plupart, déjà implémentés de manière générique et optimale dans les logiciels de statistique classiques (R, SAS, Stata, ...), ce qui facilite largement leur utilisation mais peut parfois nuire à la réflexion critique.

Aujourd'hui, le fraîchement émoulu *data scientist* reçoit par ailleurs un enseignement axé sur les défis informatiques qui se posent lorsqu'on essaie d'utiliser des données massives pour répondre à un questionnement. Pour autant, la compréhension des points-clés du raisonnement probabiliste –permettant la mise en équations explicite d'un problème concret en vue de faire parler des données– aura-t-elle réellement progressé ces dernières années ? Bien que les modèles stochastiques les plus complexes (décrits dans un cours de Master par exemple) ne reposent, le plus souvent, que sur l'assemblage de sous-modèles élémentaires beaucoup plus simples (Wikle et al., 1998), force est de constater que de nombreux maîtres de stage déplorent souvent par la suite la timidité excessive, voire le manque d'autonomie et d'envie créatrices des apprentis-chercheurs qu'ils accueillent, notamment en début de thèse. À ce niveau académique, les compétences de formalisation sont pourtant indispensables : faute de réaliser

l'éventail des modélisations possibles que permettent déjà les structures aléatoires de base, comment contribuer au développement de nouveaux modèles ?

La responsabilité de ce handicap amène à se questionner sur la pédagogie que nous-mêmes déployons lors de nos enseignements de niveau M2 ou inférieur. En ces temps où nombre de tenants de l'*Intelligence Artificielle* et du *Big Data* plaident volontiers pour un apprentissage sans modèle statistique, comment motiver notre audience à saisir l'intérêt de la construction explicite d'un raisonnement probabiliste pour répondre à un problème concret ?

La mise en place d'expériences ludiques et faciles à réaliser en salle par les étudiants peut permettre de rendre tangibles et intéressantes la construction explicite et les propriétés mathématiques de modèles probabilistes, qu'il soient élémentaires ou déjà élaborés (e.g., hiérarchiques). Ainsi, afin d'illustrer la construction du modèle de Bernoulli pour tests en duo/trio, Azaïs (2004) décrit une expérience ludique de dégustation pour distinguer deux produits semblables (sodas au cola) à partir de résultats binaires. De même, afin d'illustrer la construction de la loi normale multivariée, Tibshirani et al. (2011) modélisent les impacts au jeu de fléchettes, rappellent les propriétés de la transformée de Fourier et cherchent la stratégie judicieuse pour viser à maximiser son score. La dernière section du livre *Teaching Statistics* de Gelman and Nolan (2017) est consacrée à un atelier de construction d'hélicoptères en papier (Box, 1992; Annis, 2005), afin de mettre en place, ciseaux en main, une séance très appliquée de planification expérimentale.

Les manuels de statistique ont tendance à laisser croire que la loi de Bernoulli n'est à réserver qu'aux premières séances d'un cours élémentaire de statistique. En accord avec Collett (2002), nous réfutons ce parti pris. Cette loi de probabilité est d'intérêt dans de nombreux cursus comme les sciences de la vie ou les sciences humaines et sociales. Elle trouvera également son intérêt dans le cadre d'une initiation au raisonnement probabiliste, qui peut se faire, par exemple, dans un cours de statistique portant sur l'approche bayésienne et/ou la modélisation hiérarchique.

Dans cet article, nous développons les idées en germe dans Dudley (1983) qui illustraient une expérience de capture-marquage-recapture (*CMR*) avec des friandises colorées du type *M&M's*. Dans la section 2, nous décrivons les détails d'une expérience ludique facile à effectuer en salle avec des gommettes, des haricots secs, une cuillère à soupe et un saladier. Nous l'avons réalisée de nombreuses fois avec des étudiants de niveau M1 d'un cursus de mathématique, de niveau M2 d'un cursus d'écologie, et de dernière année d'écoles d'ingénieurs en statistique, dans le cadre d'ateliers d'enseignement et d'écoles-chercheurs. Si on les livre à eux-même lors d'une première demi-heure, les étudiants, matheux ou écologues, se précipitent généralement sur l'obtention d'une valeur ponctuelle, sans se soucier d'indiquer la moindre variabilité. Tous sont enclins à confondre probabilité et fréquence empirique. Certains écologues sont peut-être un peu moins réticents à admettre la présence d'incertitudes, mais aucun étudiant n'a d'emblée recours à un vocabulaire probabiliste. En réaction, nous pensons que prendre le temps d'exploiter toutes les aspects de cette expérience permet de mettre en lumière de façon simple et intéressante les points-clés du raisonnement probabiliste, indispensables au statisticien-modélisateur, dans le cadre spécifique d'un problème d'estimation de la taille inconnue d'une population :

— les ingrédients de base du problème de statistique inférentielle considéré, en particulier en distinguant les grandeurs que l'on voit, les observables,

de celles qu'on ne voit pas mais qui sont nécessaires pour poser le problème, les inconnues (section 3) ;
— la construction d'un modèle probabiliste possible, basé sur l'assemblage de plusieurs briques binomiales élémentaires, ainsi que les différentes décompositions possibles de la vraisemblance associée (section 4) ;
— la recherche d'estimateurs, leur étude théorique ainsi que la comparaison de leurs propriétés mathématiques par simulation numérique, qui feront l'objet de la section 5 ;
— les différences opérationnelles majeures entre approche statistique fréquentielle et perspective bayésienne, discutées dans la section 6 .

2 Une expérience ludique de capture-marquage-recapture à réaliser en groupes

Le recensement est le dénombrement exhaustif des individus constituant une population statistique. Dans la pratique, un recensement nécessite généralement des moyens importants et la durée des enquêtes est un frein quand l'estimation de l'effectif doit être rendue rapidement. Dans ce contexte, travailler sur un modèle probabiliste de capture-marquage-recapture pour estimer la taille d'une population est un recours intéressant. Pour pouvoir s'inscrire dans les hypothèses des modèles standards de la famille *CMR*, il faut notamment, comme pour un recensement, que le nombre d'individus dans un milieu fermé n'évolue pas : les mouvements migratoires, les processus naturels de natalité et de mortalité sont tous nuls ou s'équilibrent durant la mesure. Il faut par ailleurs que la répartition des individus dans le milieu soit homogène et que la probabilité de capture ne varie pas durant la durée des opérations.

Historiquement, ce sont les écologues intéressés par l'estimation de la taille des populations animales dans le milieu naturel qui ont développé le modèle *CMR* (Seber et al., 1982; McCrea and Morgan, 2014; Royle et al., 2013). Nous prenons ici l'exemple de l'estimation d'une population de poissons vivant dans un milieu fermé, comme un lac ou un tronçon de rivière délimité en amont et en aval par des filets. Le client du statisticien pourrait être la gestionnaire d'une société de pêche intéressée par le nombre de saumons juvéniles quittant une rivière pour rejoindre la mer comme dans Rivot and Prévost (2002). Une première pêche (considérée comme une première opération de capture) donne C_1 poissons qui sont marqués et remis à l'eau. Une fois l'homogénéité du milieu restaurée, une seconde pêche (souvent avec le même dispositif) fournit C_2 poissons dont C_{21} poissons qui ont été marqués lors de la première pêche et donc $C_{20} = C_2 - C_{21}$ poissons qui sont non marqués. Le modèle *CMR* trouve également application dans d'autres domaines, en particulier au cours d'expérimentations pour les sciences humaines et sociales. Par exemple, Leyland et al. (1993) relatent comment la ville de Glasgow a mis en oeuvre une méthode *CMR* pour évaluer le nombre de prostituées afin de mieux appréhender les risques de propagation des maladies infectieuses.

Sauf rares exceptions, il n'est pas possible d'emmener toute une classe sur le terrain pour collecter des données de ce type. Aussi, nous proposons une expérience de *CMR* simple à réaliser en salle et en groupes, en vue d'estimer la taille inconnue d'une population. Afin de rendre la mise en situation la plus

concrète possible, nous utiliserons donc par la suite le vocabulaire emprunté à l'écologie halieutique : *pêche* pour *capture*, *poisson* pour *individu*, etc.

FIGURE 1 – Gommettes, haricots secs, cuillère à soupe et saladier lors d'une expérience de capture-marquage-recapture réalisée en salle.

Cette expérience de *CMR* nécessite que les étudiants soient répartis en groupes ; nous avons constaté qu'un effectif de trois étudiants par groupe permet un partage du travail efficace et une discussion inventive. L'expérience proposée a l'avantage de ne demander que peu de matériel à mettre à disposition de chaque groupe (voir Figure 1) :

- Une cuillère à soupe : c'est l'instrument de pêche ;
- Un saladier (diamètre 30 à 40 cm) ou un moule à cake ($30 \times 10 \times 10$ cm^3) : c'est le lac (ou le tronçon de rivière) ;
- Un kilogramme de riz pour simuler l'eau du lac ;
- Un paquet de haricots secs, 500 grammes de lingots blancs ou rouges feront l'affaire : ce sont les poissons ;
- Un paquet d'étiquettes auto–adhésives (e.g., gommettes) de couleurs variées pour le marquage des poissons.

L'expérience se déroule comme suit. Tout d'abord, chaque groupe d'étudiants prépare le saladier du groupe voisin : il y dispose une couche de riz ainsi qu'un certain nombre de haricots secs qu'il aura préalablement comptés. Requérant la discrétion des préparateurs, le professeur relèvera un à un les effectifs exacts [1] de la population contenue dans chacun des saladiers (*i.e.*, l'état de la nature). Après avoir récupéré son saladier auprès de son voisin (libre au professeur d'imaginer de plus savantes permutations), chaque groupe doit réaliser, au moins dans un premier temps, deux pêches successives, la taille de la population qu'il évalue lui étant donc inconnue. En pratique, il est conseillé de faire au moins six coups de cuillère par pêche afin d'obtenir des captures suffisamment nombreuses pour une bonne estimation des inconnues du problème (décrites dans la section suivante). Après chaque pêche, le groupe :

1. On supposera que ces effectifs auront été comptés séparément par chacun des 3 étudiants afin de limiter au maximum le risque d'erreur de mesure

a) marque les poissons pêchés à l'aide de gommettes d'une couleur fixée mais différente d'une pêche à l'autre (voir Figure 2) ;
b) compte et note le nombre de poissons pêchés.

Nous suggérons également de marquer la seconde pêche et d'effectuer une troisième pêche, bien que ce ne soit guère l'usage rencontré sur le terrain. Cette expérience permet ainsi à chaque groupe de récolter ses propres données de *CMR*.

FIGURE 2 – Étudiants en action lors d'une expérience de capture-marquage-recapture.

Pour chaque groupe d'étudiants, l'objectif est de proposer une estimation du nombre de haricots secs contenus dans son saladier **et** une fourchette de confiance, au vu des données qu'il a collectées. Outre observer le mode de fonctionnement en collectif des étudiants, il est intéressant de leur faire exprimer leur conception de ce qu'est une probabilité, et comment s'y prendre pour répondre à l'objectif fixé. Face à un problème réel, comment ont-ils recours au raisonnement probabiliste et aux outils statistiques ? Comment assimilent-ils les informations ? Quoique les questions posées soient très simples : *Combien de poissons ? Avec quelle (in)certitude ?*, le problème probabiliste est déjà suffisamment élaboré pour donner matière à un questionnement scientifique fructueux.

3 Les ingrédients de base d'un problème de statistique inférentielle

Il n'est certes pas facile de réfréner l'impatience amusée des étudiants pour pêcher des haricots et coller des gommettes, mais on peut profiter de leur intérêt pour les faire réfléchir aux résultats qu'ils vont obtenir au terme de leur expérience de *CMR*. Pour plus de clarté, on se focalisera plutôt, à ce stade, sur le cas simple de deux pêches successives. La maïeutique prescrit de demander aux étudiants de lister puis de classer l'ensemble des ingrédients qui vont jouer un rôle dans le problème de statistique inférentielle posé. Pour faire progresser

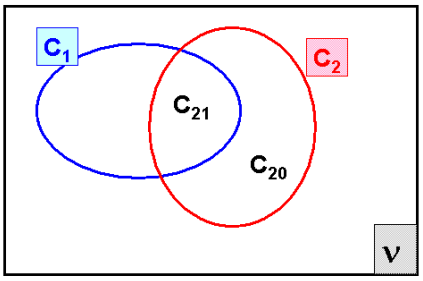


FIGURE 3 – Diagramme de Venn indiquant les différents résultats de comptage obtenus lors d'une méthode de capture-marquage-recapture basée sur deux pêches successives. C_1 est le nombre total de poissons pêchés et marqués la première fois, C_{21} désigne le nombre de poissons pêchés lors de la seconde pêche ayant déjà été marqués lors de la première pêche, C_{20} le nombre de poissons pêchés lors de la seconde pêche non marqués lors de la première pêche.

les idées, il faut oser utiliser des mots dont le sens se précisera au fur et à mesure de leur emploi dans divers contextes, car c'est en remettant régulièrement l'ouvrage sur le métier que l'on affine sa compréhension. La notion de probabilité en est un parfait exemple. Nous suggérons ici de commencer par rappeler aux étudiants qu'en statistique inférentielle, deux sortes d'objets se distinguent selon leur nature et qu'il faut les nommer pour avancer.

3.1 Les observables

Les observables désignent les grandeurs qu'on peut voir, toucher ou mesurer. Par convention, une lettre majuscule latine est utilisée pour les nommer lorsqu'il s'agit de variables aléatoires. A ce stade, un rappel informel simple de ce que désigne une variable aléatoire en théorie des probabilités peut être nécessaire, comme par exemple : *Il s'agit d'une quantité inconnue, avant d'avoir réalisé l'expérience de CMR, et qui peut prendre une collection imaginée de valeurs, munies de pondérations.*

Dans le cas d'une méthode de *CMR* fondée sur deux pêches successives, les observables C_1, C_{20}, C_{21}, C_2 ont déjà été définies à la section 2. Un petit dessin valant mieux qu'un long discours, nous pensons pertinent de présenter en complément le diagramme de Venn de la figure 3 sur lequel sont notamment représentés ces différents résultats de comptage.

Une observable est bien sûr à distinguer de son observation, sa réalisation qui, elle, est associée à une valeur numérique unique –stockable dans un ordinateur (par exemple, après avoir réalisé une expérience de *CMR*). Par convention, une observation est nommée avec une lettre minuscule latine.

3.2 Les inconnues

Les inconnues désignent les grandeurs qu'on ne voit pas. A ce stade, il est important de modérer une discussion générale pendant laquelle il faut insister constamment sur le fait que nous faisons *exister* les inconnues parce qu'on nomme ces concepts qui sortent de notre imagination. Il faut également faire astucieusement miroiter une facilité d'emploi à venir pour justifier l'usage de choisir des lettres grecques [2] pour les désigner. Après un temps de discussion apparaissent infailliblement les paramètres des lois d'échantillonnage de notre problème de statistique inférentielle :

— ν : la taille de la population ;
— π_1 : la probabilité de capture au cours de la première pêche ;
— π_2 : la probabilité de capture au cours de la seconde pêche, éventuellement supposée égale à π_1.

Dans notre problème, c'est ν le véritable paramètre d'*intérêt* car c'est l'inconnue sur laquelle focalisera principalement l'écologue. L'introduction de π_1 et π_2 dans le raisonnement probabiliste formel est néanmoins indispensable pour estimer ν.

4 Construction d'un modèle probabiliste par assemblage de briques élémentaires binomiales

4.1 Passer des inconnues aux observables

L'expérience est toujours en attente mais inconnues et observables ayant été identifiées, on peut désormais enquérir les étudiants du "moyen" pour passer des premières aux secondes, comme sur la figure 4 sur laquelle on a pris soin de faire figurer les inconnues en haut (le monde éthéré des abstractions) et les observables en bas, afin de désigner le niveau du terrain de la réalité expérimentale. À

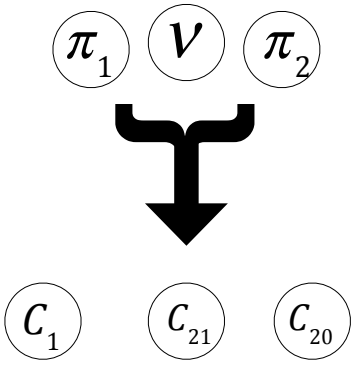


FIGURE 4 – Faire un modèle ? Passer des inconnues aux observables !

un certain moment apparaît un consensus pour construire un modèle probabi-

2. A cette occasion, nous nous sommes aperçus que, face à un public de plus en plus mondialisé, nous ne pouvions souvent plus tenir comme allant de soi la connaissance partagée de la culture gréco-latine.

liste. Il est alors profitable d'insister sur l'idée d'un assemblage de plusieurs lois binomiales élémentaires (appelées briques par la suite, par analogie avec les jeux de Lego) qui miment les résultats de capture et de recapture d'une expérience basée sur deux séquences successives de pêche. Le résultat C_1 de la première pêche correspond au nombre de poissons pêchés dans un lac ou un tronçon de rivière (délimité en amont en aval par des filets) contenant ν poissons. Sous les hypothèses décrites dans la section 2, une loi binomiale de paramètres ν et π_1 semble être un choix pertinent possible : chaque poisson est capturé de manière indépendante et avec la même probabilité π_1. Suivant la même logique, les résultats C_{21} et C_{20} de la deuxième pêche peuvent aussi être modélisés à l'aide de lois binomiales mais dont le paramètre de probabilité est π_2 et le nombre total de tirages possibles est C_1 et $\nu - C_1$ respectivement (Figure 3). Un point intéressant à souligner est l'assemblage des lois binomiales d'une pêche à l'autre : le résultat de la première pêche conditionne le paramétrage de la loi binomiale de la seconde pêche.

Est-ce que cette étape de construction explicite du modèle est un passage obligé ? Étonnamment, cette formalisation mathématique est également perçue comme un moyen de rechercher un agrément du collectif, voire une garantie de sérieux scientifique.

A ce stade, il nous semble utile de veiller à insister sur les trois points suivants, qui constituent des préalables indispensables pour répondre à tout problème concret de statistique inférentielle.

Pouvoir lister et justifier ses hypothèses de modélisation : il faut en faire pour avancer et limiter la gamme des incertitudes qu'on accepte de représenter. Mais lesquelles ? Par exemple, il faut rendre parlantes les conditions selon lesquelles les pêches sont indépendantes et suivent la même distribution de probabilité : *Les poissons ne se parlent pas, ils se font capturer de la même façon quelle que soit leur taille, il n'y pas d'effet du marquage sur le comportement des poissons, etc.* Ces hypothèses sont-elles réalistes et objectives ? Pourquoi les fait-on ? Il est facile de partager l'avis selon lequel $\pi_1 = \pi_2$. Cette probabilité commune sera notée π par la suite. Par contre, une hypothèse qu'on justifie par sa commodité, comme, par exemple : *"Techniquement, c'est uniquement ce qu'on saura traiter et il nous faut bien modestement commencer par quelque chose."* est une potion qui reste encore trop amère à avaler pour certains étudiants, nourris de déterminisme avec l'idée d'une science qui ne saurait être qu'exacte. N'ont-ils pas compris qu'un résultat scientifique est toujours né d'une simplification du problème posé, ce qui revient à considérer un cadre formel duquel on ne se donne pas la permission de sortir avant de finir le raisonnement et, alors seulement, éventuellement le remettre en question ?

Pouvoir simuler des données grâce à un programme informatique En tant que professeur, nous voudrions poursuivre l'idée d'un assemblage de plusieurs briques binomiales élémentaires du type :

$$\begin{aligned} C_1|\nu,\pi_1 &\sim dbin(\pi_1,\nu) \\ C_{20}|C_1,\nu,\pi_2 &\sim dbin(\pi_2,\nu-C_1) \\ C_{21}|C_1,\pi_2 &\sim dbin(\pi_2,C_1) \end{aligned} \tag{1}$$

en utilisant, par exemple, la syntaxe *dbin* de type BUGS (Lunn et al.,

2000) pour référer à la loi binomiale. Sauf rares exceptions, l'étudiant ne réagit pas immédiatement de cette façon qui lui semblera trop éthérée. En revanche, il ne renâclera pas à écrire et à exécuter sur son ordinateur un programme informatique de simulation de l'expérience. Reste à faire comprendre qu'appeler des fonctions de la famille *random* dans un algorithme, c'est se trouver *de facto* en présence d'un modèle probabiliste. Ceci est le message important à faire passer (et heureusement facile compte-tenu de l'engouement pour l'ordinateur). Pour notre cas d'étude, une routine de simulations de résultats de *CMR* pourrait par exemple s'écrire en R comme suit :

```
nu<-1000;  pi1<-0.75  ;
pi2=pi1
C1=rbinom(1,nu,pi1);
C21=rbinom(1,C1,pi2);
C20=rbinom(1,nu-C1,pi2);
C2=C21+C20;
```

A noter que, pour simuler des réalisations des observables (i.e. produire des données), on doit se placer dans la situation où les inconnues sont connues (ou supposées connues), ce qui est le cas de ν , π_1 et π_2 dans l'exemple ci-dessus.

Construire un graphe acyclique orienté Une bonne idée pour aider à poser les hypothèses de modélisation avant les conclusions est d'introduire un graphe acyclique orienté (voir par exemple Spiegelhalter et al. (1993)). Un tel graphe est souvent désigné par l'acronyme anglais *DAG*, pour *Directed Acyclic Graph*. Un DAG s'appuie sur des règles graphiques simples qui permettent d'utiliser des images intuitives pour aider à la conception de modèles et faciliter leur présentation . Les cercles (ou noeuds) représentent des variables aléatoires (i.e., inconnues ou observables), les flèches en traits pleins traduisent des lois de probabilité conditionnelles et les flèches en pointillés des opérations arithmétiques intermédiaires, comme sur la représentation de la figure 5 : $\nu - C_1$ désigne le nombre de poissons non marqués après la première pêche mais aussi l'un des paramètres de la loi binomiale suivie par la variable aléatoire C_{20}.

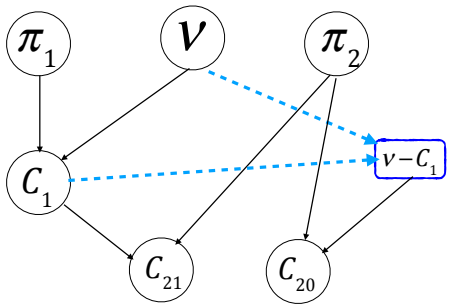


FIGURE 5 – Une bonne idée pour aider à la conception d'un modèle probabiliste et faciliter sa présentation : un graphe acyclique orienté.

4.2 Tous les chemins mènent à Rome

Construire un modèle probabiliste paramétrique pour décrire une expérience de *CMR* revient à imaginer de multiples tirages aléatoires en cascade dans des urnes (voir section 4.1), avec pour objectif de reproduire *in fine* des nombres comparables aux observations réellement faites. Comme on va le mettre en évidence ci-après, l'exemple *CMR* a pour avantage pédagogique de permettre de montrer qu'on peut parfois atteindre le même but par des chemins distincts, mais qui conduisent à la même vraisemblance et donc au même modèle probabiliste paramétrique.

Bien sûr, construire un DAG ne semble pas à proprement parler équivalent à écrire une vraisemblance, un mot-clé du cours de statistique que les étudiants ont généralement tous retenu. Hélas, ces derniers ne maîtrisent pas toujours les notions apportées par cette hydre à deux têtes : s'agit-il d'une fonction des données, ou bien des paramètres ? Une façon de remettre les notions en place est d'adopter la notation crochets de Gelfand and Smith (1990). Nous suggérons de définir d'abord la loi d'échantillonnage de C_1, C_{20}, C_{21} comme $[C_1, C_{20}, C_{21}|\nu, \pi_1, \pi_2]$ pour souligner qu'il faut se mettre en situation d'une puissance créatrice qui connaît les inconnues ν, π_1 et π_2 pour écrire le modèle [3]. Mais quand on voit cette même expression mathématique comme une fonction des inconnues (son second argument), il s'agit alors de la vraisemblance du modèle. En effet, le terme rebattu *maximum de vraisemblance* n'indique-t-il pas que c'est alors une fonction des inconnues ?

À ce stade, les propositions des étudiants mettront en évidence plusieurs façons de décrire un même modèle *CMR*. Ces propositions sont à réorganiser après avoir aidé à écrire chacune d'entre elles de façon un peu plus formelle.

La première possibilité est de suivre un raisonnement probabiliste conditionnel séquentiel. En d'autres termes, il s'agit d'opter pour la vision constructive séquentielle rencontrée dans la section 4.1 qui, fondée sur des hypothèses d'indépendance conditionnelle, permet d'écrire la loi de probabilité jointe des observables C_1, C_{20} et C_{21} sachant ν, π_1 et π_2 comme la décomposition en produits suivante :

$$[C_1, C_{20}, C_{21}|\nu, \pi_1, \pi_2] = [C_1|\nu, \pi_1] \times [C_{20}|C_1, \nu, \pi_2] \times [C_{21}|C_1, \pi_2]$$

À noter que le DAG de la figure 5, qui laisse apparaître clairement les relations d'indépendance conditionnelle entre observables, peut considérablement aider dans l'écriture de cette décomposition. On suit alors quasi-mécaniquement les deux étapes de l'expérience de *CMR*. La décomposition ci-dessus ainsi que les briques binomiales décrites dans la section 4.1 justifie le modèle décrit par le jeu d'équations 1.

Comme dans le chapitre 5 de Marin and Robert (2007), il peut être intéressant d'opposer à cette vision séquentielle la vision globale qui établit directement le bilan d'une expérience de *CMR* à deux pêches successives, à l'aide d'une loi multinomiale à 4 catégories [4] et dont les probabilités et effectifs associés sont indiqués dans le tableau 1.

3. Reconnaissons-là un tropisme bayésien partagé par les auteurs : la sigma-algèbre qui permettrait de définir une distribution conjointe sur C_1, ν, π_1, C_{21}, C_{20} et π_2 est ici enfouie à cent lieux sous le tapis.

4. Il faut ici avoir soin de ne pas postuler trop rapidement l'élimination de l'indice du π afin d'aider à distinguer facilement les deux phases de pêche

	Pêches 1 et 2	1^{ere} pêche seulement	2^{eme} pêche seulement	Jamais
Probabilité	$\pi_1\pi_2$	$\pi_1(1-\pi_2)$	$(1-\pi_1)\pi_2$	$(1-\pi_1)(1-\pi_2)$
Effectif	C_{21}	$C_1 - C_{21}$	C_{20}	$\nu - (C_1 + C_2 - C_{21})$

TABLE 1 – Probabilités de capturer un poisson et effectif associé à chacune des 4 catégories possibles (listées en colonne) d'une loi multinomiale décrivant le bilan global d'une expérience de *CMR* fondée sur deux pêches successives.

Enfin, certains souhaiteront profiter de la séance pour faire un rappel de calcul de probabilités en demandant tout d'abord aux étudiants de vérifier que la loi de $C_1, C_{20}, C_{21}|\nu, \pi_1, \pi_2$ est la même sous le modèle binomial séquentiel et le modèle multinomial de bilan. Dans le texte du matériau supplémentaire, on montre cette première équivalence puis, comment on retombe, par un simple changement de variable, sur la modélisation traditionnelle alternative d'une expérience de *CMR*, qui suppose un tirage sans remise des poissons recapturés lors de la 2ème pêche via une loi hypergéométrique [5] :

$$\begin{aligned} C_1|\nu, \pi_1 &\sim dbin(\pi_1, \nu) \\ C_2|\nu, \pi_2 &\sim dbin(\pi_2, \nu) \\ C_{21}|C_1, C_2, \nu &\sim hypergeometrique(\nu, C_1, C_2) \end{aligned} \tag{2}$$

L'expérience d'enseignement d'ateliers de ce type confirme que cette troisième façon de voir les choses (appelée *modèle avec composante hypergéométrique* par la suite) est immanquablement avancée par des étudiants de culture biologique, issus d'un Master d'écologie par exemple, car c'est sous cette forme que le modèle standard de capture-marquage-recapture y est généralement présenté. Dans cette troisième vision, on a complètement symétrisé le rôle de C_1 et C_2 (On pourra vérifier l'invariance de la génération de C_{21} si on permute le rôle de C_1 et C_2 dans le tirage hypergéométrique). Pour un statisticien, elle est aussi l'occasion de souligner les problèmes délicats de dépendance conditionnelle et d'en discuter. En effet, il serait totalement déraisonnable d'utiliser de nouveau un modèle binomial $C_{21}|C_1, \pi_2 \sim dbin(\pi_2, C_1)$ pour la dernière opération ci-dessus, car une fois la seconde pêche C_2 réalisée, on ne peut plus *générer* C_{21} indépendamment de C_2, ne serait-ce que parce que $C_{21} \leq C_2$!

5. Rappelons l'interprétation classique d'une variable aléatoire Y de loi hypergéométrique $(N+B, N, K)$: après un tirage sans remise de K boules dans une urne contenant N boules noires et B boules blanches, on observe le nombre y de boules noires obtenues. Sa loi de probabilité s'écrit :

$$[Y = y|N+B, N, K] = \frac{C_N^y C_B^{K-y}}{C_{N+B}^K}$$

5 Estimation fréquentiste d'une taille de population inconnue, avec un coup de main de R

Comme illustré par la photo de la figure 2, le grand moment arrive : on procède enfin à l'expérience de *CMR* ! Le tableau 2 donne un exemple de résultats obtenus après deux pêches successives et lors d'une séance pendant laquelle les étudiants étaient repartis en trois groupes. Jusqu'à la fin de l'exercice, les étudiants de chaque groupe ignoreront l'effectif véritable de leur saladier. Mais, pour information, les voici : Groupe 1 : 210 ; Groupe 2 : 376 ; Groupe 3 : 244.

	Groupe 1	Groupe 2	Groupe 3
C_1	99	116	82
C_{20}	56	56	74
C_{21}	62	25	39

TABLE 2 – Résultats d'une expérience de *CMR* à deux pêches avec trois groupes d'étudiants.

5.1 L'estimateur de Lincoln-Petersen

À ce stade, il convient d'interroger les étudiants sur leur meilleure estimée de l'effectif de leur population de poissons, non sans en avoir éventuellement profité pour rappeler brièvement la distinction entre *estimateur* et *estimation*. Il n'est alors pas rare que plusieurs d'entre eux supposent l'égalité des probabilités de capture π_1=π_2 et suggèrent ainsi que la proportion *connue* de poissons marqués lors de la seconde pêche soit une bonne estimation de la proportion *inconnue* de poissons marqués dans l'ensemble de la population (i.e., pris lors de la première pêche) : $\frac{C_{21}}{C_2} \approx \frac{C_1}{\nu}$. Poser l'égalité stricte des deux rapports permet de construire un estimateur ponctuel de l'effectif recherché : $\hat{\nu_p} = \frac{C_1 \times C_2}{C_{21}}$, appelé estimateur de Lincoln-Petersen (Chao et al., 2008). Cet estimateur est certes obtenu avec une approximation simple mais confondant probabilités et proportions. Hélas, cette confusion conceptuelle n'est par ailleurs pas nécessairement ressentie comme un raisonnement intuitif incomplet. Et hop, 188 poissons pour le premier groupe qui a bien envie de déclarer mission accomplie et de ranger ses cahiers ! Comme beaucoup, sa bonne volonté s'arrête généralement lorsqu'on requiert une fourchette de crédibilité, par exemple les quartiles 25% et 75% autour de ce meilleur pari qu'il a calculé à partir de ses résultats expérimentaux, même en acceptant une réponse ni formalisée, ni justifiée, mais simplement issue d'une intuition ou d'une discussion au sein du groupe.

5.2 L'estimateur de Schnabel-Chapman

Soulignant le problème majeur posé par l'estimateur de Lincoln-Petersen quand $C_{21} = 0$, on présente alors un estimateur alternatif, appelé estimateur de Schnabel-Chapman et défini par : $\hat{\nu_s} = \frac{(C_1+1)\times(C_2+1)}{C_{21}+1} - 1$ (Amstrup et al., 2010). Viennent alors assez spontanément les deux questions suivantes : *Quel est le meilleur des deux estimateurs proposés ? Pourriez-vous proposer d'autres estimateurs ?* Le plus souvent, elles ne manquent pas de déstabiliser l'audience...

Pour soulager la tension qui commence à s'installer quand les étudiants réalisent qu'on les chatouille désagréablement sur leurs conceptions de l'incertitude, un travail empirique par simulations procure un secours appréciable ; par exemple, il est facile[6] d'écrire un programme en R qui simule 100000 répétitions de notre expérience CMR pour $\nu = 200, \pi_1 = \pi_2 = 0.1$, calcule les deux estimateurs, en trace la répartition empirique et évalue leurs caractéristiques (biais, variance et risque quadratique). La comparaison des deux estimateurs apparaît en figure 6 et, bien sûr, l'estimateur de Schnabel-Chapman se révèle non biaisé et de risque quadratique plus faible que celui de Lincoln-Petersen qui, de plus, peut-être non défini quand la probabilité de capture est suffisamment faible pour obtenir $C_{21} = 0$ avec une probabilité non négligeable. Encore faut-il bien faire comprendre que cette simulation s'effectue sous la loi d'échantillonnage, avec des paramètres qui ne sont plus inconnus mais fixés à une valeur que le modélisateur a donnée. Il est également possible d'entraîner les étudiants

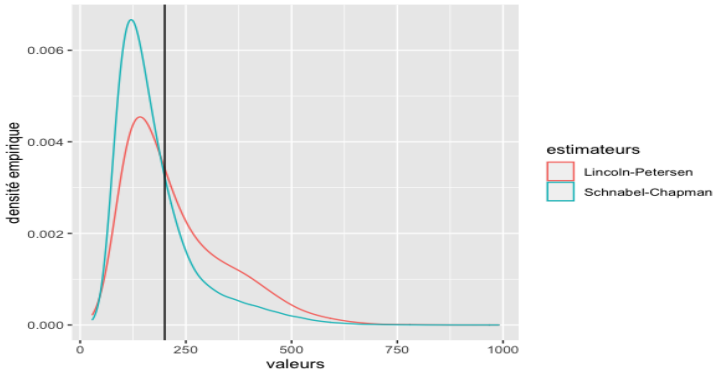


FIGURE 6 – Comparaison de la répartition des estimateurs de Petersen-Lincoln et de Schnabel-Chapman pour une population $\nu = 200$ et une probabilité de capture fixée à $\pi_1 = \pi_2 = 0.1$. L'estimateur de Petersen-Lincoln n'est pas défini ($C_{21} = 0$) dans environ 13% des cas. L'espérance de l'estimateur de Petersen-Lincoln est ici évaluée à 225 quand il est défini, celle de Schnabel-Chapman à 178, avec pour risques quadratiques $\mathbb{E}((\hat{\nu} - \nu)^2)$ respectifs 15090 et 10060.

les plus férus de programmation informatique dans une étude par simulations un peu plus poussée dont l'objectif est de comparer l'évolution du biais relatif des estimateurs de Petersen-Lincoln et de Schnabel-Chapman en fonction de la taille de la population ν (allant par exemple de 100 à 1000 par pas de 50). On pourra, par exemple pour chaque valeur de ν, générer 20000 jeux de données de CMR en fixant la probabilité de pêche π à 0.30. La figure 7 illustre encore plus clairement que, contrairement à l'estimateur de Schnabel-Chapman, l'estimateur de Petersen-Lincoln est biaisé : il sur-estime la taille de la population ν et cette sur-estimation est d'autant plus marquée que la taille de la population est petite.

6. Voir le code des figures 6 et 7 en matériau supplémentaire de l'article

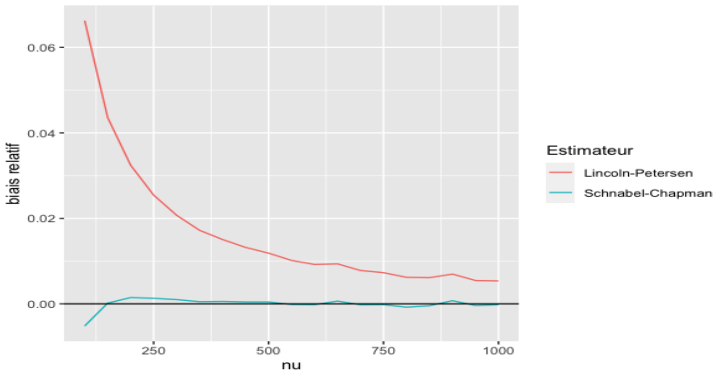


FIGURE 7 – Évolution du biais (empirique) relatif des estimateurs de Petersen-Lincoln et de Schnabel-Chapman en fonction de la taille ν de la population, pour une probabilité de capture fixée à $\pi_1 = \pi_2 = 0.3$.

5.3 Simuler pour mieux comprendre ce que signifie la loi d'un estimateur

Continuons de procéder par simulation afin de comparer les propriétés des estimateurs proposés. Pour obtenir une évaluation empirique de l'écart-type des estimateurs de Lincoln-Petersen (quand celui-ci est défini) et Schnabel-Chapman, on peut effectuer de nombreux tirages dans la loi hypergéométrique (pour simuler des réalisation de C_{21}) dont on aura réglé les paramètres sur une première approximation de ν et π. De notre expérience, il ressort que nombre d'étudiants, même au niveau Master, n'ont pas perçu l'intérêt de la répétition simulée précédente pour évaluer le comportement d'un estimateur et, par conséquent, n'ont pas compris ce que représentait la loi d'échantillonnage et la cohabitation d'estimateurs multiples. Par ailleurs, force est de constater que les notions d'intervalle de confiance et de risque restent encore trop souvent non assimilées... Conseil au pédagogue : ne pas sortir de ses gonds, reprendre tranquillement les bases de la démarche statistique pour tenter de faire comprendre à son auditoire que se contenter d'une estimation ponctuelle, c'est être pressé d'avoir raison mais certainement pas une attitude scientifique responsable.

5.4 Calcul des moments de l'estimateur de Schnabel-Chapman

R possédant un générateur hypergéométrique *rhyper*, on pourrait bien sûr d'après l'équation 2, simuler C_{21} sachant C_1 et C_2 pour faire une étude empirique des caractéristiques de l'estimateur de Schnabel-Chapman, mais pour les étudiants les plus férus de résultats théoriques (notamment ceux dans un parcours mathématique de niveau au moins M1), on peut les aider à entreprendre vaillamment une analyse de portée plus générale. Pour en évaluer les propriétés théoriques, il faut alors s'appuyer sur les propriétés de la loi hypergéométrique.

Moyennes arithmetique et inverse Pour une variable aléatoire Y de loi hypergéométrique $(N+B, N, K)$, interprétée comme le nombre de boules noires

obtenues après un tirage sans remise de K boules dans une urne contenant N boules noires et B boules blanches, on peut d'abord montrer les espérances mathématiques suivantes :

$$\begin{aligned} \mathbb{E}(Y) &= \frac{N}{N+B}K \\ \mathbb{E}\left(\frac{1}{Y+1}\right) &= \frac{N+B+1}{(N+1)(K+1)} \end{aligned} \tag{3}$$

En posant $K = C_2, N = C_1, B = \nu - C_1, Y = C_{21}$, on démontre le caractère sans biais de l'estimateur de Schnabel-Chapman $\hat{\nu_s}$ dans les circonstances où C_1 et C_2 sont suffisamment grands pour que $C_1 + C_2 \geq \nu$. On trouve $\mathbb{E}(\hat{\nu_s}) = \nu$ par déconditionnement sur les variables aléatoires indépendantes C_1 et C_2.

Moyenne inverse d'ordre 2 Poursuivant l'étude des propriétés de l'hypergéométrique, on trouve dans Wittes (1972) [7] la preuve que la quantité

$$\begin{aligned} s^2 &= \frac{(C_1+1)(C_2+1)}{(C_{21}+1)^2}\frac{(C_1-C_{21})(C_2-C_{21})}{C_{21}+2} \\ &= \hat{\nu}_s^2 + 3\hat{\nu}_s + 2 - (C_1+1)(C_1+2)(C_2+1)(C_2+2)/(C_{21}+1)(C_{21}+2) \end{aligned}$$

est telle que :

$$\mathbb{E}(s^2) = \mathbb{E}(\hat{\nu}_s^2) - \nu^2 = \mathbb{V}(\hat{\nu_s})$$

c'est à dire que s^2 est un estimateur (sans biais) de la variance (conditionnelle à C_1 et C_2) de l'estimateur de Schnabel-Chapman.

Le tableau 3 récapitule moyenne et écart-type des estimateurs de Lincoln-Petersen (par simulation) et de Schnabel-Chapman (par la théorie) pour les données du tableau 2.

Groupe	Petersen-Lincoln	Ecart-type	Schnabel-Chapman	Ecart-type
G1	188	10.6	188	9.9
G2	376	57.4	368	51.8
G3	238	23.2	236	21.4

TABLE 3 – Estimations et leurs écart-types des estimateurs (Petersen-Lincoln et Schnabel-Chapman) de l'effectif ν de poissons pour les 3 groupes d'étudiants (Voir le code en matériau supplémentaire de l'article).

5.5 L'estimateur du maximum de vraisemblance

Montons la barre ! Peu encourageant à la première lecture des équations de ce modèle, trouver le maximum de vraisemblance ne pose pourtant aucune difficulté numérique particulière pour les étudiants ayant assimilé leur cours d'optimisation. Supposons $\pi_1 = \pi_2$. En matériau supplémentaire de l'article, on montre que l'expression de $[C_1, C_{20}, C_{21}|\nu, \pi]$, vue comme une fonction de ν et π, est proportionnelle à :

$$\frac{\nu!}{(\nu - C_1 - C_{20})!}\pi^{(C_1+C_2)}(1-\pi)^{2\nu-(C_1+C_2)}$$

7. Remarque : Une erreur typographique s'est glissée dans Wittes (1972) : la quantité *inverse factorial moment* d'ordre k doit s'écrire $\mathbb{E}(\prod_{i=1}^{k}(n_{12}+i)^{-1})$ et non $\mathbb{E}(\prod_{i=1}^{k}(n_{12}+i)^{-2})$

Maximiser cette fonction de vraisemblance revient à maximiser son logarithme, d'où : $\hat{\pi}_{mv} = \frac{C_1+C_2}{2\hat{\nu}_{mv}}$. Pour trouver $\hat{\nu}_{mv}$, on cherchera le maximum en ν de la fonction ci-dessous -dite de *vraisemblance profilée*- on a substitué π par sa valeur optimale en fonction de ν :

$$\frac{\nu!}{(\nu - C_1 - C_{20})!} \times \left(\frac{C_1 + C_2}{2\nu}\right)^{(C_1+C_2)} \times \left(1 - \left(\frac{C_1 + C_2}{2\nu}\right)\right)^{2\nu-(C_1+C_2)}$$

Pour maximiser cette expression, on pourra simplement parcourir toutes les valeurs (discrètes) possibles pour l'effectif de poissons ν.

Un intervalle de confiance pour cet estimateur peut s'obtenir de multiples manières, par exemple en ayant recours à l'approximation par un chi-deux de la déviance profilée (Cox and Hinkley, 1974; Casella and Berger, 2001). Mais rares furent les étudiants que nous avons rencontrés qui possédaient le niveau de culture suffisant en statistique mathématique asymptotique pour y parvenir.

Finalement, le tableau 4 résume les estimations de l'effectif ν de poissons et les intervalles de confiance asymptotiques à 95% obtenus pour chacun des 3 groupes d'étudiants pour les 3 estimateurs étudiés jusqu'à présent.

Gpe	$\hat{\nu}_{025.p}$	$\hat{\nu}_p$	$\hat{\nu}_{975.p}$	$\hat{\nu}_{025.s}$	$\hat{\nu}_s$	$\hat{\nu}_{975.s}$	$\hat{\nu}_{025.mv}$	$\hat{\nu}_{mv}$	$\hat{\nu}_{975.mv}$
G1	168	188	208	169	188	207	172	189	213
G2	258	376	494	267	368	469	297	385	534
G3	192	238	283	194	236	278	205	242	299

TABLE 4 – Estimations et intervalles de confiance à 95% obtenus pour 3 estimateurs de l'effectif ν de poissons pour 3 groupes d'étudiants. Les suffixes utilisés sont "$_p$" pour Petersen-Lincoln, "$_s$" pour Schnabel-Chapman, "$_{mv}$" pour maximum de vraisemblance. Les intervalles de confiance à 95% calculés pour les estimateurs de Lincoln-Petersen et de Schnabel-Chapman supposent la normalité asymptotique de ces estimateurs. Celui du maximum de vraisemblance, lui dissymétrique, est obtenu grâce à l'approximation par un chi-deux de la vraisemblance profilée. Le code R figure en matériau supplémentaire de l'article.

6 La piste bayésienne

Si l'on dispose de plus d'une journée pour mener cet atelier *Gommettes, haricots et saladier*, consacrer une séance à l'initiation au raisonnement bayésien (Robert, 2005; Parent and Bernier, 2007; Boreux et al., 2010; McElreath, 2020) représente un défi intéressant compte-tenu de deux de ses aspects de caractère typiquement bayésien : après tout, chaque groupe d'étudiants a préparé le matériel de ses voisins, et donc possède de l'information *a priori* sur l'ordre de grandeur du nombre possible de haricots dans un saladier et de l'efficacité de la cuillère à soupe comme instrument de pêche ! Et prendre un pari subjectif quant aux valeurs des variables du problème fait ici au moins autant sens qu'imaginer d'évaluer une fréquence d'occurrence lors d' une hypothétique répétition expérimentale.

6.1 Commodité de l'approche bayésienne

L'objectif d'un atelier *CMR* traité en bayésien sera essentiellement de faire saisir l'idée que de nombreuses difficultés techniques relatives à la mise en pratique de l'approche bayésienne sont aujourd'hui levées (Brooks, 2003). Pour une première initiation opérationnelle à l'inférence bayésienne, nous suggérons de s'appuyer sur les logiciels *clic-boutons* de la famille BUGS (Gilks et al., 1994; Lunn et al., 2000) et le langage déclaratif associé très proche de R. Pour un premier contact avec l'inférence bayésienne à partir d'algorithmes Monte-Carlo par Chaînes de Markov (MCMC), le package R baptisé "rjags" [8] qui permet d'appeler le logiciel Jags de Plummer (2015) a reçu notre faveur, car il fonctionne efficacement sous tous les systèmes d'exploitation de l'ordinateur : Windows, MacOs ou Linux. Sur nos données, pour le goupe 1, la figure 8 permet ainsi de visualiser [9] la loi *a posteriori* jointe de ν et π (pour le groupe d'étudiants 1) et de constater l'existence d'une corrélation *a posteriori* entre ces deux quantités. Les caractéristiques des lois marginales *a posteriori*, présentées dans la

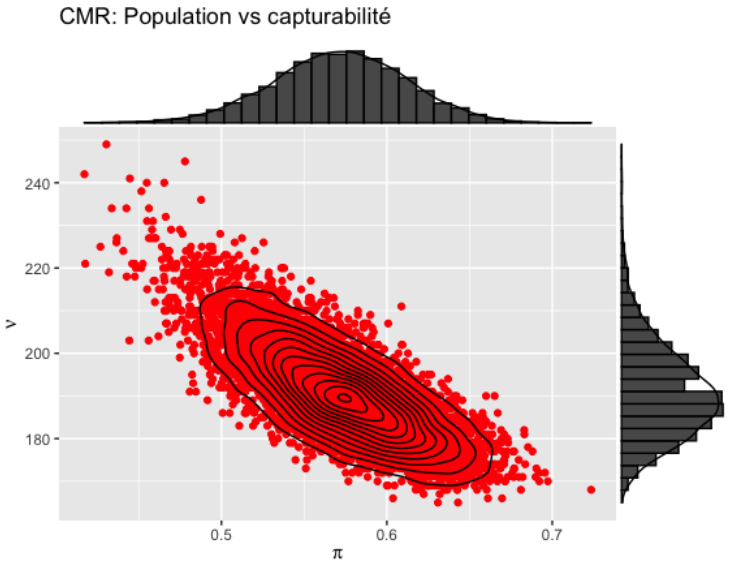


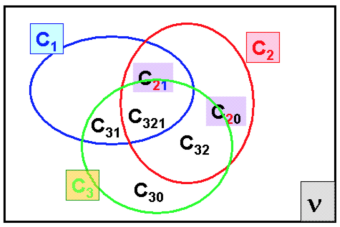
FIGURE 8 – Loi *a posteriori* de la capturabilité π et de la taille de la population de poissons ν connaissant les données obtenues par le groupe 1 pour une loi a priori bêta(2,1) sur π et uniforme sur ν.

table 5 s'interprètent naturellement en terme de pari intuitif : pour le premier groupe, il y 95 chances sur 100 que la taille de la population se situe entre 173 et 215 poissons (*c-à-d, je suis prêt à parier à* 95 *contre* 5 *sur cette assertion*). Pour les groupes qui ont continué leur expérience de CMR en passant de deux à trois pêches successives avec remise (voir figure 9), le problème inférentiel serait nettement plus compliqué à résoudre selon l'approche fréquentiste. Mais en

8. https ://cran.r-project.org/web/packages/rjags/index.html
9. L' obtention de cette figure et du tableau suivant est détaillée en matériau supplémentaire

param	mean	sdt	q.025	q25	q50	q75	q.975
ν_prior	250	144	11	125	249	373	487
ν_post	**191**	11	**173**	183	**190**	197	**215**
π_prior	0.67	0.24	0.16	0.50	0.71	0.87	0.98
π_post	0.57	0.04	0.49	0.54	0.57	0.60	0.65

TABLE 5 – Statistiques *a priori* et *a posteriori* relatives à l'effectif de poissons ν et à la capturabilité π pour le groupe 1. En gras figurent deux estimateurs bayésiens classiques (i.e., moyenne et médiane *a posteriori*) ainsi que les bornes de l'intervalle de crédibilité à 95% de ν.



	Gp1	Gp3
C1	99	82
C20	56	74
C21	62	39
C30	41	40
C321	43	12
C31	15	14
C32	38	27

FIGURE 9 – Diagramme de Venn (à gauche) et données recueillies par les groupes d'étudiants 1 et 3 (à droite) lors d'une expérience de CMR basée sur trois pêches successives.

bayésien, les résultats *a posteriori* obtenus avec seulement quelques lignes de programmation (voir le code BUGS en matériau supplémentaire) mettent facilement en évidence le gain de précision apporté par une troisième pêche sur la connaissance des inconnues : pour le groupe 1 par exemple, sachant les quatre informations supplémentaires apportées par la troisième pêche, la meilleure estimée du nombre d'individus se décale en effet de 191 à 215 tandis que l'écart-type *a posteriori* se réduit de 11 à 7.

6.2 Efficacité : l' intérêt pour un statisticien classique d'emprunter la piste bayésienne

Convenons pourtant que comparer des algorithmes (non des méthodes scientifiques) selon les résultats des calculs et leurs facilités de mise en oeuvre ne veut rien dire en soi. Une raison plus subtile peut motiver un statisticien classique à emprunter *incognito* la piste bayésienne.

Pour les étudiants matheux de M1 ou M2, cette expérience CMR est aussi l'occasion de voir (ou de revoir) quelques éléments de la théorie de l'estimation. Une espérance *a posteriori*, comme celle évaluée pour ν au tableau 5 est une fonction (éventuellement non explicite) des données, il s'agit donc d'un estimateur au sens classique. Il est même d'ailleurs extrêmement intéressant pour un

statisticien *classique* car c'est un estimateur non dominé (encore appelé efficace dans certains manuels de théorie de l'estimation). Par définition, un estimateur est efficace si on ne peut pas trouver un autre estimateur dont le risque quadratique soit uniformément meilleur sur tout le domaine de définition de ν. Or, le théorème de la classe complète de Wald (1947) établit une passerelle essentielle entre statistiques classique et bayésienne : il précise que tous les estimateurs efficaces (les seuls dignes d'intérêt pour le statisticien !) sont engendrés par la classe des estimateurs bayésiens et leurs limites.

6.3 Des propriétés structurelles pour modéliser ? de Finetti et l'échangeabilité.

Du point de vue plus général de l'initiation à la modélisation probabiliste à partir d'une expérience de CMR, même la simple écriture binomiale séquentielle est déjà fort riche d'enseignements : les étudiants ont tendance à calquer un modèle d'urne avec tirages sans remise sur le mécanisme de pêche. Ce stéréotype n'est pas approprié ici : il s'agit plutôt d'une somme de comportements individuels Bernoulli *pris/pas pris* indépendants. Les conditions *iid* peuvent être ici postulées très raisonnablement : un poisson capturé n'alerte pas ses congénères, l'espace disponible ne modifie pas la capturabilité résiduelle. Par exemple, lorsqu'un dispositif par pêche électrique est utilisé, la technique de pêche fait que, dans le disque d'influence du dispositif électrique, tous les poissons, gros ou petits, sont attirés irrésistiblement vers l'anode tenue par le technicien. A un niveau plus avancé de l'enseignement, des raisons structurelles de symétrie plaident également pour un modèle binomial : il semblerait en effet déraisonnable de ne pas postuler l'invariance de la représentation probabiliste prédictive du système par toute permutation de l'ordre des individus capturés et, assumant cette propriété quelque soit la taille de la population, le théorème de représentation de de Finetti (1937) impose alors une vraisemblance binomiale pour ces données individuelles binaires ! Ce théorème d'un grand auteur bayésien du siècle passé démontre également l'existence mathématique d'une quantité sur laquelle conditionner pour retourner à l'indépendance –nous avons ici proposé le couple (ν, π)– et de la loi qui doit lui être associée. Voilà que réapparaît la loi *a priori*, et une autre histoire de modélisation...

6.4 Interpréter la probabilité ?

À ce stade, hasarder un oecuménisme optimiste entre les postures fréquentiste et bayésienne s'avèrera toxique (Lecoutre, 1997) car les interprétations de la probabilité ne se recoupent pas du tout de façon *cohérente* au sens de Lindley (2013). Une p-value n'a rien d'une crédibilité bayésienne. La philosophie cognitive doit servir : expliquer pourquoi probabilité fréquentiste et crédibilité bayésienne ne sont pas la même chose participe à la rigueur intellectuelle et c'est un fait crucial pour la bonne formation d'un statisticien. Aussi se doit-on de prendre du temps pour répondre aux questions des étudiants concernant l'approche statistique bayésienne et en rappeler formellement les bases (Savage, 1954, 1971; Bernardo and Smith, 2009; Kadane, 2011; Lindley, 2013). D'ailleurs, comme ce paradigme semble d'emblée pour certains d'interprétation intuitive plus naturelle et immédiate (Collectif Biobayes, 2015; Lambert, 2018),la pilule de la cohérence mathématique que garantit le calcul des probabilités sera d'autant

moins difficile à faire passer. Leur questionnement le plus inquiet portera sans doute sur la loi *a priori* : dans notre exemple, l'incorporation d'expertise probabilisée sur les valeurs des inconnues ν, π_1 et π_2. C'est simplement une question de modélisation (voir à ce propos le point de vue provocateur de Spiegelhalter et al. (2004) à la page 73 du chapitre 3 de leur ouvrage) mais la modélisation, ce n'est pas si facile ! Il faut dire qu'on peut discuter sans état d'âme du choix des lois *a priori* et réaliser une analyse de sensibilité, afin d'être conscient de l'impact potentiel de ces choix, au vue des données disponibles. Dans l'exemple de *CMR*, comme on ne fait pas de recensement, on est obligé de reconnaître que notre état de connaissance concernant la cible ν est incertain. Il y a donc un sens à représenter cette incertitude par une distribution de probabilité et ce même quand cet état de connaissance est très réduit. Par exemple, changer la loi uniforme sur la gamme des effectifs pour une loi uniforme sur le *logarithme* de ν, c'est considérer que la même chance est donnée à tous les ordres de grandeurs possibles. Le choix d'une loi *a priori* bêta sur π de paramètres a=2 et b=1 serait un choix acceptable pour encoder, par exemple, le jugement d'un expert de la pêche électrique annonçant que son évaluation moyenne personnelle de l'efficacité π est de l'ordre de 2/3 mais, sans grande confiance, puisqu'il n'est prêt à parier qu'une chance sur deux environ pour l'intervalle $[0.4, 0.8]$.

7 Épilogue

7.1 Perspectives

Plusieurs extensions plus spécifiques peuvent être apportées à cette expérience ludique et facile à réaliser en salle avec des gommettes, des haricots secs, une cuillère à soupe et un saladier. Citons, par exemple :

- Il est possible de poursuivre l'initiation aux algorithmes MCMC (Robert and Casella, 2013) : la faible dimension du problème permet de se lancer dans l'implémentation d'un algorithme de Gibbs simple et de comparer les résultats à ceux obtenus avec le logiciel Jags. Remarquons que les calculs peuvent se faire à la main en s'appuyant sur la conjugaison partielle (loi *a priori* bêta), mais il faut expliquer l'algorithme de Gibbs et éventuellement les techniques de Raoblackwellisation (Casella and Robert, 1996), notamment en ce qui concerne l'inconnue π. Les lois conditionnelles complètes sont ici explicites : loi bêta pour π et loi discrète pour ν (dérivées en matériau supplémentaire). La litérature sur les échantillonneurs par algorithmes markoviens est en constante évolution (Lunn et al., 2009) ; d'autres outils d'échantillonnage, par exemple, du type STAN (Gelman et al., 2015) ou NIMBLE (de Valpine et al., 2017) peuvent également être testés à partir de notre expérience jouet.
- Certaines perspectives intéresseront davantage les écologues que les probabilistes. Il n'est guère difficile d'adapter le matériel de notre expérience de CMR pour faire une introduction aux techniques de capture avec enlèvements successifs, un autre moyen d'usage courant en écologie pour évaluer la taille d'une population (Rivot et al., 2008) ou pour s'initier à la représentation d'un système dynamique à état discret en adoptant des règles de mortalité et de naissance sur la population (King et al., 2009). Enfin les méthodes de *CMR* font l'objet de constructions hiérarchiques

fructueuses (Rivot and Prévost, 2002) pour représenter les ressemblances entre années, sites, etc. et en tirer profit pour une inférence plus riche d'informations.

7.2 Conclusions

Dans la France des années 1650, le Chevalier de Méré avait un entêtant problème de jeu :

— Lancer un dé équilibré au maximum quatre fois, et gagner si vous obteniez un six ;
— Lancez deux dés au maximum vingt-quatre fois pour obtenir un double-six.

Quel était le meilleur pari ?

Reprenant la solution erronée que le Chevalier de Méré en avait présentée, Blaise Pascal et Pierre de Fermat planchèrent sur le problème. L'histoire des sciences retient que c'est ainsi qu'ensemble, ils développèrent les premiers éléments de la théorie des probabilités...

Finalement, l'expérience que nous proposons ici –ludique et facile à effectuer en salle avec des gommettes, des haricots secs, une cuillère à soupe et un saladier– n'est peut-être qu'un simple retour aux sources, avec le jeu, ses paris sur les résultats possibles et l'observation répétée de données expérimentales. A partir de données réelles, collectées par les étudiants eux–mêmes, elle permet de développer, dès les premières séances d'un cours de statistique inférentielle, de nombreux points-clés du raisonnement probabiliste –qu'il soit fréquentiste ou bayésien– indispensables au statisticien-modélisateur. Gageons qu'en l'abordant de façon délibérément simple et empirique, elle suscitera néanmoins la réflexion et mobilisera la faculté d'abstraction des étudiants face aux questions liées à la quantification des incertitudes.

Matériau supplémentaire

Les auteurs sont convaincus que la reproductibilité totale est la norme minimale pour juger des travaux scientifiques. Un fichier *html* est disponible pour vérifier et reproduire tous les chiffres et résultats de cet article à la page web de la revue.

Remerciements

Les auteurs remercient Jacques Bernier pour les nombreuses discussions, souvent vigoureuses mais toujours constructives, concernant la cohérence et la rationalité du discours statistique.

Références

Amstrup, S., McDonald, T., and Manly, B. (2010). *Handbook of Capture-Recapture Analysis*. Princeton University Press.

Annis, D. H. (2005). Rethinking the paper helicopter : Combining statistical and engineering knowledge. *The American Statistician*, 59(4) :320–326.

Azaïs, J.-M. (2004). Illustration de la méthode des plans d'expériences sur la comparaison de boissons au cola. *Journal de la société française de statistique*, 145(4) :69–78.

Bernardo, J. M. and Smith, A. F. (2009). *Bayesian theory*, volume 405. John Wiley & Sons.

Boreux, J.-J., Parent, E., Bernier, J., and Bernier, J. (2010). *Pratique du calcul bayésien*, volume 118. Springer.

Box, G. E. (1992). Teaching engineers experimental design with a paper helicopter. *Quality Engineering*, 4(3).

Brooks, S. P. (2003). Bayesian computation : a statistical revolution. *Philosophical Transactions of the Royal Society of London. Series A : Mathematical, Physical and Engineering Sciences*, 361(1813) :2681–2697.

Casella, G. and Berger, R. L. (2001). *Statistical inference*. Duxbury/Thomson Learning.

Casella, G. and Robert, C. P. (1996). Rao-blackwellisation of sampling schemes. *Biometrika*, 83(1) :81–94.

Chao, A., Pan, H.-Y., and Chiang, S.-C. (2008). The petersen–lincoln estimator and its extension to estimate the size of a shared population. *Biometrical Journal : Journal of Mathematical Methods in Biosciences*, 50(6) :957–970.

Collectif Biobayes (2015). *Initiation à la statistique bayésienne : bases théoriques et applications en alimentation, environnement, épidémiologie et génétique*. Ellipses.

Collett, D. (2002). *Modelling binary data*. CRC press.

Cox, D. R. and Hinkley, D. V. (1974). *Theoretical Statistics*. Chapman and Hall, London.

de Finetti, B. (1937). La prévision : ses lois logiques, ses sources subjectives. In *Annales de l'institut Henri Poincaré*, volume 7, pages 1–68.

de Valpine, P., Turek, D., Paciorek, C. J., Anderson-Bergman, C., Lang, D. T., and Bodik, R. (2017). Programming with models : writing statistical algorithms for general model structures with nimble. *Journal of Computational and Graphical Statistics*, 26(2) :403–413.

Dudley, B. (1983). A practical study of the capture/recapture method of estimating population size. *Teaching Statistics*, 5(3) :66–70.

Gelfand, A. E. and Smith, A. F. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85(410) :398–409.

Gelman, A., Lee, D., and Guo, J. (2015). Stan : A probabilistic programming language for Bayesian inference and optimization. *Journal of Educational and Behavioral Statistics*, 40(5) :530–543.

Gelman, A. and Nolan, D. (2017). *Teaching statistics : A bag of tricks*. Oxford University Press.

Gilks, W. R., Thomas, A., and Spiegelhalter, D. J. (1994). A language and program for complex Bayesian modelling. *Journal of the Royal Statistical Society : Series D (The Statistician)*, 43(1) :169–177.

Kadane, J. B. (2011). *Principles of uncertainty*. CRC Press.

King, R., Morgan, B., Gimenez, O., and Brooks, S. (2009). *Bayesian analysis for population ecology*. CRC press.

Lambert, B. (2018). *A Student's Guide to Bayesian Statistics*. Sage.

Lecoutre, B. (1997). C'est bon à savoir. *Et si vous étiez un bayésien qui s' ignore. Modulad*, 18 :81–87.

Leyland, A., Barnard, M., and McKeganey, N. (1993). The use of capture-recapture methodology to estimate and describe covert populations : An application to female street-working prostitution in Glasgow. *Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique*, 38(1) :52–73.

Lindley, D. V. (2013). *Understanding uncertainty*. John Wiley & Sons.

Lunn, D., Spiegelhalter, D., Thomas, A., and Best, N. (2009). The bugs project : Evolution, critique and future directions. *Statistics in medicine*, 28(25) :3049–3067.

Lunn, D. J., Thomas, A., Best, N., and Spiegelhalter, D. (2000). Winbugs-a Bayesian modelling framework : concepts, structure, and extensibility. *Statistics and computing*, 10(4) :325–337.

Marin, J.-M. and Robert, C. (2007). *Bayesian core : a practical approach to computational Bayesian statistics*. Springer Science & Business Media.

McCrea, R. S. and Morgan, B. J. (2014). *Analysis of capture-recapture data*. CRC Press.

McElreath, R. (2020). *Statistical rethinking : A Bayesian course with examples in R and Stan*. CRC press.

Parent, E. and Bernier, J. (2007). *Le raisonnement bayésien : modélisation et inférence*. Springer Science & Business Media.

Plummer, M. (2015). Jags version 4.0. 0 user manual. *Lyon. Available online at : http ://sourceforge. net/projects/mcmc-jags*.

Rivot, E. and Prévost, E. (2002). Hierarchical Bayesian analysis of capture mark recapture data. *Canadian Journal of Fisheries and Aquatic Sciences*, 59(11) :1768–1784.

Rivot, E., Prévost, E., Cuzol, A., Baglinière, J.-L., and Parent, E. (2008). Hierarchical Bayesian modelling with habitat and time covariates for estimating riverine fish population size by successive removal method. *Canadian Journal of Fisheries and Aquatic Sciences*, 65(1) :117–133.

Robert, C. (2005). *Le choix bayésien : Principes et pratique*. Springer Science & Business Media.

Robert, C. and Casella, G. (2013). *Monte Carlo statistical methods*. Springer Science & Business Media.

Royle, J. A., Chandler, R. B., Sollmann, R., and Gardner, B. (2013). *Spatial capture-recapture*. Academic Press.

Savage, L. J. (1954). *The foundations of statistics*. Courier Corporation.

Savage, L. J. (1971). Elicitation of personal probabilities and expectations. *Journal of the American Statistical Association*, 66(336) :783–801.

Seber, G. A. F. et al. (1982). *The estimation of animal abundance and related parameters*, volume 8. Blackburn press Caldwell, New Jersey.

Spiegelhalter, D. J., Abrams, K. R., and Myles, J. P. (2004). *Bayesian approaches to clinical trials and health-care evaluation*, volume 13. John Wiley & Sons.

Spiegelhalter, D. J., Dawid, A. P., Lauritzen, S. L., and Cowell, R. G. (1993). Bayesian analysis in expert systems. *Statistical science*, pages 219–247.

Tibshirani, R. J., Price, A., and Taylor, J. (2011). A statistician plays darts. *Journal of the Royal Statistical Society : Series A (Statistics in Society)*, 174(1) :213–226.

Wald, A. (1947). An essentially complete class of admissible decision functions. *The Annals of Mathematical Statistics*, pages 549–555.

Wikle, C. K., Berliner, L. M., and Cressie, N. (1998). Hierarchical Bayesian space-time models. *Environmental and Ecological Statistics*, 5(2) :117–154.

Wittes, J. T. (1972). Note : On the bias and estimated variance of Chapman's two-sample capture-recapture population estimate. *Biometrics*, pages 592–597.