N° d'ordre : 35/2010-M/MT

Republique Algerienne Democratique et Populaire

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université des Sciences et de la Technologie Houari Boumedienne

Faculté des Mathématiques

MEMOIRE

Présenté pour l'obtention du diplôme de Magister

En : Mathématiques

Spécialité : Recherche Opérationnelle, Méthodes Stochastiques

Par Malika HAMRAT

Sujet

Modèles de Capture-Recapture Multiples Estimation d'une Population Fermée

Soutenu publiquement le 25 Février 2010, devant le jury composé de :

Mr.	**BOUKHETALA**	Kamel	Professeur	U.S.T.H.B.	Président.
Mr.	**OUAFI**	Rachid	Maître de Conférences	U.S.T.H.B.	Directeur de thèse.
Mr.	**AKNOUCHE**	Abdelhakim	Maître de Conférences	U.S.T.H.B.	Examinateur
Mme.	**SADKI**	Ourida	Maître de Conférences	U.S.T.H.B.	Examinatrice

...وَقُل رَّبِّ زِدْنِي عِلْمًا ﴿١١٤﴾

من سورة: طه

Remerciements

En premier lieu, je remercie Dieu le tout puissant de m'avoir prêté main pour la réalisation de ce travail.

À Mr OUAFI Rachid, j'exprime toute ma gratitude et ma reconnaissance pour la patience et la générosité avec lesquelles il a su me guider durant les travaux de recherches.

Mes sincères remerciements vont aussi à Mr BOUKHETALA Kamel, professeur à l'U.S.T.H.B, qui m'a fait l'honneur de présider le jury.

Mes remerciements s'adressent également à Mr AKNOUCHE Abdelhakim, maître de conférences à l'U.S.T.H.B et Mme SADKI Ouardia, maître de conférences à l'U.S.T.H.B pour leur gratitude et leur dévouement qui ont permis d'honorer le jury.

Je remercie vivement *Mr BENTARZI* Mohamed, professeur à *l'USTHB,* pour son aide et soutien.

Je ne saurais oublier de remercier encore une fois, Mr AKNOUCHE et Mr HAMDI qui m'ont beaucoup aidé.

Dédicaces

À ceux qui m'ont donné la vie, je leurs témoigne mon amour et ma reconnaissance

À ma mère qui m'a appris à être patiente pour surmonter les difficultés

À mon très cher père

À mes très chers frères Abdelkader, Moussa, Abdelkarim

À mes amies Isma et Feriel

Malika

Modèles de Capture-Recapture Multiples : Estimation d'une Population Fermée

Résumé

Les modèles de capture-recapture sont des outils efficaces pour estimer les paramètres des populations, tels que la taille, le taux de survie et le taux de reproduction.

Dans notre travail, nous nous sommes intéressés aux modèles de capture-recapture uni-état à temps discret pour estimer la taille d'une population fermée, difficile à dénombrer. Les données de capture-recapture sont issues d'un procédé d'échantillonnage qui consiste à capturer, marquer puis recapturer des individus de cette population. L'opération est répétée en un nombre fini d'occasions de capture. Nous avons étudié les méthodes d'estimation bayésienne afin de traiter les données multiples issues d'un modèle de capture-recapture qui suppose l'indépendance entre les occasions de capture et l'homogénéité des individus, la loi conjuguée gamma est choisie comme loi a priori pour notre paramètre d'intérêt.

L'autre approche élaborée dans notre thèse est basée sur la modélisation log-linéaire dans le but de traiter l'hétérogénéité due aux comportements des individus après leur capture initiale, la dépendance entre occasions de capture ou la différence entre individus. Un estimateur log-linéaire de la taille de la population est obtenu en utilisant la méthode du maximum de vraisemblance.

L'évaluation des résultats obtenus est illustrée selon des données épidémiologiques en tant que des données de capture-recapture.

Table des matières

Liste des Figures

Liste des Tables

Introduction

L'estimation des paramètres des populations; tels que le taux de survie, le taux de reproduction et la taille de la population, est un problème universel, a interessé beaucoup plus les biologistes que les statisticiens. Une nouvelle stratégie d'échantillonnage a vu le jour dite *capture-recapture* pour répondre à ce problème.

Le procédé d'échantillonnage de capture-recapture consiste à capturer, marquer puis recapturer des individus d'une population. Cette dernière peut être fermée, c'est-à-dire il n'y a pas de naissance ou d'immigration, ni de mort ou émmigration.
Autrement dit; la taille de la population est invariante durant l'échantillonnage, à l'opposé la population est dite ouverte.

Les modèles de capture-recapture déstinés aux populations fermées ont un seul site, ils sont dit : modèles *uniétat.* Par contre ceux destinés aux populations ouvertes ont plusieurs sites, ils sont dit : modèles *multiétats.* Les deux types peuvent avoir des occasions de capture discètes ou continues.

Les modèles de capture-recapture à temps discret uniétat, ont leurs origines au $19^{ème}$ siècle, Peterson (1889)[27], Lincoln (1930)[27] ont initié leur développement mathématique à partir de la réalisation de deux applications remarquables en écologie animale, ils utilisaient le retour des bancs de poissons pour estimer la taille de la population nord américaine, ce modèle a deux occasions de capture. Schnabel [27] a amélioré le modèle classique de Peterson en multipliant les opérations (pêche, marquage, remise à l'eau, recapture ...). En les traitant globalement, le modèle obtenu est *le modèle de capture-recapture multiple.* Ce procédé a été traditionnellement utilisé pour estimer les populations animales, ce n'est que récemment que le modèle de capture-recapture a été considéré et développé pour l'estimation de la taille des

populations humaines. Les travaux de Wittes [30], El Khorazaty et Coll [14]en épidémiologie et de Wolter [31], Cowan et Malec [13] en démographie, montrent l'importance et l'intérêt. Des travaux remarquables ont été réalisés dans ce contexte, Seber [24] et Schwarz & Seber [24], autres références importantes (Otis, Burnham, White & Anderson [22]). Sans oublier la contribution de Darroch [8], Castledine [8], Philip Smith [28], et plusieurs d'autres à l'inférence bayésienne sur les modèles de capture-recapture.

Otis et al [22] ont suggéré trois sources de variation pour les probabilités de capture (l'occasion de capture (t), le comportement des individus après leur capture initiale (b) et l'hétérogénéité entre individus (h)). En tenant compte de la dernière source de variation des probabilités de capture, il résulte des modèles de capture-recapture homogènes notés M_0, M_t, M_b, M_{tb} et autres hétérogènes notés $M_h, M_{th}, M_{bh}, M_{tbh}$.

La présence d'hétérogénéité due à la différence entre individus, occasions de capture ou le comportement des individus après leur capture initiale peut biaiser l'estimateur de la taille de la population. Pollock et Otto [24] ont considéré l'estimateur de Jackknife, Norris et Pollock [24], Pledger [24] suggèrent la partition des individus en groupes homogènes, Fienberg [16], Cormack [12], Agresti [2] ont introduit l'approche log-linéaire sur les données de capture-recapture.

Notre objectif est d'estimer la taille d'une population fermée, difficile à dénombrer en utilisant les modèles de capture-recapture a temps discret uniétat.

Ce présent mémoire est constitué de cinq chapitres.

Dans le premier chapitre nous présentons des notions de base de la statistique inférentielle sur lesquelles repose notre travail.

Dans le deuxième chapitre, nous présentons les aspects et les notations utilisées concernant les expériences de capture-recapture ainsi que les modèles associés. Nous donnons les définitions des modèles de capture-recapture homogènes et les modèles hétérogènes ainsi que leurs paramètres, illustrons avec un exemple pour le modèle M_t.

Le troisième chapitre est consacré à l'inférence bayésienne des données de capture-recapture issues du modèle homogène M_t qui suppose l'indépendance entre les occasions de capture. Nous décrivons dabors une expérience de capture-recapture comme sa modèlisation statistique. L'estimation ponctuelle consiste à calculer l'estimateur de Bayes pour la

fonction de perte quadratique généralisé et son risque associé. Ajoutons l'estimation par intervalle pour notre paramètre d'intérêt. Terminons ce chapitre par une étude de simulation.

L'objectif à atteindre dans le quatrième chapitre est la modélisation log-linéaire des données de capture-recapture multiples. Au premier lieu, nous présentons des généralités sur les modèles linéaires généralisés qui représentent une méthodologie de régression unifiant une large variété de réponses, en se limitant au modèles log-linéaires. Ensuite, nous décrivons des méthodes iteratives pour calculer l'estimateur du maximum de vraisemblance, paramètre d'intérêt du modèle. Nous traitons en dérnier, les modèles de capture-recapture suivant : M_0, M_t, M_{th}, M_b où nous déduisons un estimateur log-linéaire de la taille de la population ainsi que leurs propriétés (biais, variance asymptotique,...).

Dans le cinquième chapitre nous présentons une étude épidémiologique dont le but est d'estimer le nombre de malades non enregistrés d'une pathologie bien déterminée et d'évaluer les systèmes de surveillance ainsi que les registres épidémiologiques.

Chapitre 1

Notions de base de la statistique inférentielle

1.1 Introduction

La statistique inférentielle regroupe des *théories* et des méthodes ayant pour but d'une part d'*induire* des *informations* concernant la *distribution de probabilité* (ou une de ses *caractéristiques*) d'un caractère concernant les individus d'une population, et ce sur la base de l'observation d'une partie restreinte (échantillon) de cette population, et d'autre part de mesurer le *degré* de *l'incertitude* correspondant à cette induction, au moyen d'outils probabilistes. Les informations induites peuvent être exploitées dans un processus de prise de décision dans des problèmes gouvernés par l'incertitude. Ces information peuvent être :

· Une *valeur ponctuelle* appelée à remplacer une caractéristique particulière inconnue de la distribution d'une population : paramètre d'une loi (moyenne, variance, valeur maximum, médiane, probabilité d'un évènement particulier,...) en se basant sur l'observation d'un échatillon tiré de la population.

· Un *intervalle aléatoire* couvrant (contenant), avec une certaine probabilité assez importante, la valeur inconnue d'une caractéristique de la distribution de la population, en se basant bien entendu sur l'observation d'un échantillon tiré de la population.

· *confirmation* ou *infirmation d'une hypothèse* consernant la distribution d'une population (ou une de ses caractéristiques), en se basant toujours sur l'observation d'un échantillon tiré de cette population.

Dans le premier cas, le processus d'induction (et de mesure d'incertitude inhérente) d'une valeur ponctuelle est connu sous le nom de : *théorie d'estimation ponctuelle.* Dans le second cas, il s'agit de la *théorie de l'estimation par intervalle de confiance.* Dans le troisième l'induction est nommée : *théorie des tests d'hypothèses.*

Dans ce chapitre, nous exposerons des notions de base concernant la théorie d'estimation ponctuelle, (en se limitant à l'estimation paramétrique sur laquelle repose notre travail) et à l'estimation par intervalle de confiance.

Cet exposé est inspiré du cours "*statistique inférentielle*" [6] et [3].

1.2 Estimation ponctuelle (paramétrique)

1.2.1 Principes de l'estimation ponctuelle

Introduction

Estimation ponctuelle *paramétrique* suppose que la loi de probabilité de la population est connue et que seulement les paramètres sont ignorés, par contre l'estimation *non-paramétrique* ne suppose aucune hypothèse sur la loi de probabilité de la population.

Considérons une population statistique et notons X la variable aléatoire désignant le caractère quantitatif faisant l'objet d'une étude statistique.

Dans les problèmes statistiques paramétriques, la loi de probabilité de la variable aléatoire parente X dépend, généralement, d'un certain paramètre scalaire (ou vectoriel) inconnu noté θ qui appartient à l'espace paramétrique (l'ensemble des valeurs possibles du paramètre) noté Θ. La population est donc caractérisée par une famille de lois de probabilité, indéxée par le paramètre θ, que l'on désigne par

$$\{f(x,\theta), x \in \mathfrak{X}, \theta \in \Theta\}$$

La vraie distribution, qui est un membre de cette famille, et qui correspond à une seule valeur de θ, notée θ_0 et est appelée : la vraie valeur du paramètre inconnu θ. Cette fonction

de distribution sera complètement spécifiée des que le paramètre (ou l'état de la nature) θ est connu.

Plus généralement, soit $g(\theta)$ une caractéristique inconnue liée à une population statistique qu'on voudrait estimer à partir de l'observation d'un échantillon aléatoire $\underline{X}$ de cette population. On peut estimer $g(\theta)$ à l'aide d'une fonction de l'échantillon (indépendante d'aucun paramètre inconnu) dont l'expression dépend de ce que représente θ pour la population.

· Si $g(\theta)$ représente la moyenne de la population $(g(\theta) = E(X))$, il est de coutume de l'estimer par la moyenne de l'échantillon (moyenne empirique)

$$\overline{X} = \frac{X_1 + X_2 + \ldots + X_n}{n}$$

· Si $g(\theta)$ représente la variance de la population $(g(\theta) = Var(X))$, il est de coutume de l'estimer par la variance de l'échantillon (variance empirique)

$$S^2 = \frac{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2}{n}$$

· Si $g(\theta)$ représente la plus grande valeur de la population (par exemple loi $U[0,\theta]$), il est de coutume de l'estimer par la plus grande valeur de l'échantillon

$$X_{(n)} = \max(X_1,\ X_2, \ldots,\ X_n)$$

Autrement dit chaque statistique peut être considérée comme un estimateur pour la fonction à estimer $g(\theta)$. Une question importante se pose alors : quelle statistique $T(X)$ doit-on choisir pour estimer une caractéristique $g(\theta)$?
Bien entendu on souhaite avoir un "bon" estimateur (selon un sens précis : sans biais, efficacité, ...) qui fournit, le mieux possible, des renseignements plus riches, plus précis et plus certains sur la caractéristique à estimer.

1.2.2 Estimation ponctuelle dans le cadre décisionnel

Fonction de décision

En se basant sur la quantité d'information contenue dans l'échantillon aléatoire $\underline{X}$ représentatif prélevé d'une population, on peut décider d'attribuer au paramètre inconnu θ (dont

dépend la loi de probabilité da la variable aléatoire parente X), la valeur numérique $\theta(x_1, x_2, \ldots, x_n)$. Par manque d'information complète sur la population (car la décision est prise seulement à la lumière des renseignements fournis par un échantillon de taille n extrait de cette population), il est évident que l'on soit trop optimiste si on exige que cette décision prise sur la valeur de θ est certainement correcte. La prise de décision dans une situation pareille, où régne l'incertitude est toujours susceptible d'erreur. Il est donc nécessaire de garder présent à l'esprit le coût de la perte qui peut être entraînée par une décision incorrecte, qui considérait $\theta(x_1, x_2, \ldots, x_n)$ comme étant la valeur du paramètre θ, alors que la vraie valeur est, en réalité.différente.

On suppose qu'on désire entreprendre une action et qu'il est à faire un choix parmi un certain nombre d'actions possibles. On suppose que l'action dépend d'un paramètre inconnu θ dont dépend la loi $f_x(.;\theta)$ de la population à étudier.
La méthode statistique consiste à choisir un échantillon (aléatoire simple) $X_1, X_2, \ldots, X_n$ issu de la population et de décider d'une action sur la base des valeurs de l'échantillon. On procède comme suit :

· Définir l'espace Θ des valeurs possibles du paramètre.

· Définir l'ensemble de toutes les actions possibles A.

Remarque 1.2.1 *Dans le problème d'estimation ponctuelle l'ensemble A est en bijection avec Θ*

· Choisir une régle (fonction) de décision qui pour chaque échantillon $X_1, X_2, \ldots, X_n$ associe une action dans A.

Définition 1.2.1 *Une régle de décision (dite aussi fonction ou stratégie de décision), $\widehat{\theta}(X_1, X_2, \ldots, X_n)$, est une application mesurable de l'espace de l'échantillon, $\underline{X}$, dans l'espace des actions A (et donc dans Θ)*

$$\begin{aligned} \widehat{\theta}(.) &: \mathfrak{X} \longrightarrow \Theta \\ \underline{x} &= (x_1, x_2, \ldots, x_n) \in \mathfrak{X} \rightsquigarrow \widehat{\theta}(x_1, x_2, \ldots, x_n) \in \Theta \end{aligned}$$

On appelle cette régle de décision estimateur (ponctuel) de θ et on le note simplement $\widehat{\theta}(\underline{X})$. Une action particulière, ou une réalisation particulière $\widehat{\theta}(x_1, x_2, \ldots, x_n)$ de l'esimateur $\widehat{\theta}(\underline{X})$, est dite estimation (ponctuelle) de θ, et elle est notée $\widehat{\theta}(\underline{x})$.

Plus généralement, on peut chercher à estimer une caractéristique quelconque, autre que la moyenne et la variance, liée à une population statistique. Ainsi, en considérant une caractéristique (fonction quelconque du paramètre scalaire ou vectoriel inconnu θ), $g(\theta)$, on note $\widehat{g}(\underline{x})$ l'estimation ponctuelle de cette caractéristique inconnue. La statistique ou, plus précisément, l'estimateur qui fournit cette estimation est notée $\widehat{g}(\underline{X})$

Fonction de perte

L'importance et la gravité des conséquences d'une décision erronée différent d'un problème de décision à un autre; ce qui nous suggère d'utiliser une certaine mesure, pour comparer les différentes décisions possibles, et pour juger de leurs qualités. Il est indispensables alors d'exprimer quantitativement les conséquences économiques de chaque régle de décision ou estimateur $\widehat{\theta}(\underline{X})$, pour chaque valeur du paramètre (ou de l'état de la nature) à l'aide d'une fonction , dite *fonction de perte*, notée $\mathfrak{L}(.,.)$.

Définition 1.2.2 *Une fonction de perte est une fonction mesurable de l'échantillon et le paramètre en question autrement dit; une variable aléatoire définie sur l'espace* $(\Theta, B_\Theta) \times (\Theta, B_\Theta)$ *et à valeur dans* $\mathbb{R}^+$, *c'est-à-dire*

$$\left(\widehat{\theta}, \theta\right) \in \Theta \times \Theta \rightsquigarrow \mathfrak{L}\left(\widehat{\theta}(X_1,\ X_2, \ldots,\ X_n), \theta\right) \in \mathbb{R}^+$$

L'étude du problème de l'estimation ponctuelle, dans le cadre décisionnel, se heurte, à la difficulté ou même très souvent à l'impossibilité d'évaluer et d'interpréter la perte économique par une fonction explicite $\mathfrak{L}\left(\widehat{\theta}(\underline{X}), \theta\right)$ pour chaque estimation $\widehat{\theta}(\underline{x})$, et pour chaque valeur du paramètre inconnu θ à estimer. Bien que l'approche décisionnelle du problème de l'estimation ponctuelle n'est pas évidente à cause de la difficulté relative à l'identification de la fonction de perte, pour chaque problème d'estimation, il reste toujours possible d'intégrer l'estimation ponctuelle dans le cadre de la théorie de la décision par un choix conventionnel de la fonction de perte.

Naturellement, la fonction de perte à choisir, pour surmonter ce type de difficulté, doit être une fonction valable pour une grande catégorie de problèmes d'estimation ponctuelle.

Choix conventionnel de la fonction de perte

Il est montré, sous certaines conditions souvent réalisables, que la fonction de perte qui satisfait les besoins demandés dans l'étude de l'estimation ponctuelle, peut être approximée par une quantité positive ou nulle de la forme :

$$C(\theta)\left(\widehat{\theta}(\underline{X})-\theta\right)^2$$

où le coefficient $C(\theta)$, dépendant seulement du paramètre θ, est strictement positif.
C'est la raison pour laquelle, la famille de fonctions de perte la plus courante dans les problèmes d'estimation est :

$$\begin{aligned}\mathfrak{L}\left(\widehat{\theta}(\underline{X}),\theta\right) &= C(\theta)\left(\widehat{\theta}(\underline{X})-\theta\right)^2, \forall\left(\widehat{\theta},\theta\right)\in\Theta\times\Theta \\ \text{avec } C(\theta) &> 0\ , \forall\theta\in\Theta\end{aligned}$$

On remarque que la perte exprimée par cette fonction s'annule lorsque $\widehat{\theta}(\underline{X})$ est égale à la vraie valeur du paramètre inconnu θ, et est positive et proportionnelle au carré de l'erreur commise, lorsque θ est estimé par une valeur $\widehat{\theta}(\underline{x})$ qui en est différente.
Le coefficient $C(\theta)$, strictement positif quelque soit la valeur de θ, n'est pas toujours facile à spéciffier pour chaque problème d'estimation, de plus il n'a pas d'influence sur l'efficacité relative de deux estimateurs, $\widehat{\theta}_1(\underline{X})$ et $\widehat{\theta}_2(\underline{X})$.
Ainsi, dans les problèmes d'estimation, $C(\theta)$ est fréquemment mis égal à l'unité et la fonction de perte se réduit à la fonction quadratique

$$\mathfrak{L}\left(\widehat{\theta}(\underline{X}),\theta\right)=\left(\widehat{\theta}(\underline{X})-\theta\right)^2, \forall\left(\widehat{\theta},\theta\right)\in\Theta\times\Theta$$

On note que cette dernière fonction de perte est une fonction convexe de $\widehat{\theta}$ et de θ individuellement. Elle est aussi favorisée par la simplicité relative de l'analyse mathématique des expressions qui interviennent dans la recherche de l'estimateur et lors de l'étude de ces propriétés. Il existe cependant d'autres fonctions de perte qui peuvent convenir à plusieurs problèmes d'estimation. On cite à titre d'exemple quelques unes d'entre-elles :

Fonction de perte absolue

$$\mathfrak{L}\left(\widehat{\theta}(\underline{X}),\theta\right)=\left|\widehat{\theta}(\underline{X})-\theta\right|, \forall\left(\widehat{\theta},\theta\right)\in\Theta\times\Theta$$

Fonction de perte quadrilatique

$$\mathfrak{L}\left(\widehat{\theta}\left(\underline{X}\right),\theta\right)=\left(\widehat{\theta}\left(\underline{X}\right)-\theta\right)^{4},\forall\left(\widehat{\theta},\theta\right)\in\Theta\times\Theta$$

Inconvénient : Les fonctions de perte données présentent un inconvénient majeur qui est la symetrie. Elles attribue les mêmes conséquences pour un écart positif ou négatif. Pourtant, nombre de domaines d'application tels que la finance et les assurances, sont caractérisés par l'asymétrie, d'où l'introduction d'autres fonction de perte (par exemple, la Value at Risque). Il n'existe aucune régle ou méthode générale qui prenne en charge le choix de la fonction de perte. Cette tâche est assez délicate ; elle demande une certaine expertise dans le domaine où le problème d'estimation en question est posé. Néanmoins, on souligne que ce choix de la fonction de perte dépend de plusieurs facteurs et en particulier de :

· la qualité de l'estimateur

· l'utilisation de cet estimateur

Fonction de risque

Même si l'on peut identifier de façon adéquate la fonction de perte $\mathfrak{L}\left(\widehat{\theta}\left(\underline{X}\right),\theta\right)$, on ne peut l'utiliser directement pour comparer deux estimateurs, puisque, dépendant de l'échantillon, pour certains échantillons $\underline{x}$ on peut avoir par exemple $\mathfrak{L}\left(\widehat{\theta}_1\left(\underline{X}\right),\theta\right)<\mathfrak{L}\left(\widehat{\theta}_2\left(\underline{X}\right),\theta\right)$ et pour d'autres $\underline{x}^{'}$ on peut avoir $\mathfrak{L}\left(\widehat{\theta}_1\left(\underline{X}\right),\theta\right)>\mathfrak{L}\left(\widehat{\theta}_2\left(\underline{X}\right),\theta\right)$. Pour résoudre ce problème, on compare plutôt les pertes moyennes pour chaque règle de décision, d'où l'introduction d'une nouvelle fonction comme critère de choix d'estimateur : *la fonction de risque.*

Définition 1.2.3 *La fonction de risque (notée $R(\widehat{\theta};\theta)$) est l'espérance mathématique de la fonction de perte :*

$$R(\widehat{\theta};\theta)=E\left(\mathfrak{L}\left(\widehat{\theta}\left(\underline{X}\right),\theta\right)\right),\forall\left(\widehat{\theta},\theta\right)\in\Theta\times\Theta$$

Cette moyenne dépend en effet seulement de la règle de l'estimation adoptée $\widehat{\theta}\left(\underline{x}\right)$ et de la caractéristique à estimer θ . Elle permet alors la comparaison des estimateurs, selon les valeurs possibles du paramètre inconnu θ, dont dépend la loi de la variable aléatoire X.

Critère de choix d'estimateurs

On a vu que toute fonction de décision (et donc toute statistique) $T(\underline{X})$ peut être considérée comme estimateur pour un paramètre inconnu $g(\theta)$. Une question importante se pose, alors : Quel estimateur $T(\underline{X})$ doit on choisir pour estimer une caractéristique $g(\theta)$? Bien entendu on souhaite avoir un "bon" estimateur (selon un sens à préciser) qui fournit, le mieux possible, des renseignements plus riches, plus précis sur la caractéristique à estimer.
Il est évident que le choix d'un estimateur doit être fait selon un certain critère de comparaison des estimateurs.

Nous abordons le principe de l'erreur quadratique moyenne, fréquemment, adoptées en statistique. Ce principe permet de comparer des estimateurs et par la suite de faire le choix de l'estimateur optimal.

Soient $g(\theta)$ une caractéristique inconnue à estimer et $T(\underline{X})$ un estimateur quelconque, à valeurs dans $g(\Theta)$, pour $g(\theta)$. Considérons la fonction de perte quadratique

$$\mathfrak{L}\left(T(\underline{X}), g(\theta)\right) = \left(T(\underline{X}) - g(\theta)\right)^2, \forall \left(T, g(\theta)\right) \in g(\Theta) \times g(\Theta)$$

Il parait donc raisonnable d'utiliser l'erreur quadratique moyenne (c'est la fonction de risque associée à une fonction de perte quadratique), qui est l'espérance de la fonction de perte, par rapport à l'échantillon X, quelque soit l'estimateur T et $\theta \in \Theta$

$$\mathfrak{R}(T, \theta) = E(\mathfrak{L}\left(T(\underline{X}), g(\theta)\right)) = E\left(\left(T(\underline{X}) - g(\theta)\right)^2\right), \forall \left(T, g(\theta)\right) \in g(\Theta) \times g(\Theta)$$

Définition 1.2.4 *On dit que l'estimateur $T_1(\underline{X})$ est préférable, au sens quadratique, à l'estimateur $T_2(\underline{X})$, et on note $T_1(\underline{X}) \geq T_2(\underline{X})$, si*

$$\mathfrak{R}\left(T_1, g(\theta)\right) \leq \mathfrak{R}\left(T_2, g(\theta)\right), \forall \theta \in \Theta$$

Si en outre, il existe une valeur θ_0 de θ telle que

$$\mathfrak{R}\left(T_1, g(\theta_0)\right) < \mathfrak{R}\left(T_2, g(\theta_0)\right)$$

alors l'estimateur $T_1(\underline{X})$ est dit strictement préférable, toujours au sens quadratique, à $T_2(\underline{X})$ et on note

$$T_1(\underline{X}) > T_2(\underline{X})$$

Définition 1.2.5 *Un estimateur $T(\underline{X})$ est dit admissible, s'il n'existe pas d'estimateur $T^*(\underline{X})$ qui lui est strictement préférable.*

L'ensemble de tous les estimateurs admissibles est appelé classe admissible.
Un estimateur inadmissible $T(\underline{X})$ peut être écarté de l'ensemble des estimateurs possibles de $g(\theta)$, car il existe au moins un estimateur $T_1(\underline{X})$ qui lui est préférable. Ainsi, lors de la recherche d'un estimateur approprié pour un paramètre inconnu on se restreint seulement à l'ensemble de tous les estimateurs admissibles (classe admissible).
En pratique, la situation n'est pas toujours aussi simple, car en général un estimateur $T_1(\underline{X})$ peut être meilleur (au sens de l'erreur moyenne quadratique) qu'un autre estimateur $T_2(\underline{X})$ sur un ensemble de valeurs de θ, mais ce n'est pas forcément le cas sur un autre ensemble.

Définition 1.2.6 *Un estimateur $T(\underline{X})$ de $g(\theta)$ est dit uniformément meilleur, au sens quadratique, si*

$$\mathfrak{R}\left(T\left(\underline{X}\right), g(\theta)\right) \leq \mathfrak{R}\left(S\left(\underline{X}\right), g(\theta)\right), \forall \theta \in \Theta$$

pour tout estimateur $S(\underline{X})$ de $g(\theta)$.

C'est-à-dire que l'erreur quadratique moyenne de l'estimateur $T(\underline{X})$ est inférieure ou égale à celle de tout autre estimateur $S(\underline{X})$ de $g(\theta)$.

Remarque 1.2.2 *Si l'inégalité large précédente est une inégalité stricte pour au moins une valeur de θ, alors l'estimateur $T(\underline{X})$ est dit estimateur strictement uniformément meilleur de $g(\theta)$, toujours au sens quadratique.*

Il est alors logique de se poser la question importante suivante : existe t-il un estimateur de $g(\theta)$ uniformément meilleur, au sens de l'erreur quadratique moyenne ?

La réponse n'est affirmative que dans quelques situations très simples et triviaux pour la raison suivante : supposons que la vraie valeur de θ est θ_0; alors la vraie valeur de $g(\theta)$ est $g(\theta_0)$. Comme $S(\underline{X})$ est un estimateur quelconque, alors pour $S(\underline{X}) = g(\theta_0)$; l'inégalité implique $R(T, g(\theta_0)) \leq 0$, ce qui entraîne $R(T, g(\theta_0)) = 0$.

C'est-à-dire, qu'on peut estimer $g(\theta)$ sans aucune erreur, ce qui, en pratique, est impossible. On conclut, donc, qu'en général l'estimateur uniformément meilleur au sens de l'erreur

quadratique moyenne, par rapport à l'ensemble de tous les estimateurs admissibles possibles, n'existe pas. Par conséquent, on doit être moins exigeant dans la recherche d'un estimateur, d'où la nécessité d'imposer aux estimateurs certaines conditions qui, d'une part assurent certaines propriétés désirables et réalisables et, d'autre part restreignent la recherche du meilleur estimateur, au sens quadratique, à une classe restreinte des estimateurs satisfaisant ces conditions. Etant donné l'ensemble des estimateurs d'une certaine caractéristique $g(\theta)$, la fonction de risque quadratique permet d'isoler le sous-ensemble de tous les estimateurs admissibles. Cependant, l'utilisation de cette fonction de risque ne permet pas généralement sans autres informations supplémentaires sur le paramètre θ, de préférer un estimateur à un autre de cette classe pour toute valeur de θ. Reste à savoir maintenant comment et sur quelle base peut-on préférer ou choisir (ou encore privilégier) un des estimateurs de la classe admissible ?

Décomposition de l'erreur quadratique moyenne Soit $T(\underline{X})$ un estimateur quelconque de $g(\theta)$. L'erreur quadratique moyenne correspondante $\mathfrak{R}(T, g(\theta))$ peut s'écrire sous la forme suivante

$$\begin{aligned}
\mathfrak{R}(T, g(\theta)) &= E\left(T(\underline{X}) - g(\theta)\right)^2, \theta \in \Theta \\
&= E\left(T(\underline{X}) - E\left(T(\underline{X})\right) + E\left(T(\underline{X})\right) - g(\theta)\right)^2, \theta \in \Theta \\
&= E\left(T(\underline{X}) - E\left(T(\underline{X})\right)\right)^2 + \left(E\left(T(\underline{X})\right) - g(\theta)\right)^2, \theta \in \Theta
\end{aligned}$$

De cette dernière expression on constate que l'erreur quadratique moyenne est la somme de deux quantités : la variance de l'estimateur $T(\underline{X})$ qui est strictement positive (elle s'annule pour l'estimateur trivial constant presque partout quelque soit l'échantillon) et la quantité positive ou nulle $\left(E\left(T(\underline{X})\right) - g(\theta)\right)^2$. La racine carrée de cette quantité est dite biais.

1.2.3 Estimateur sans biais

Définition 1.2.7 *Un estimateur $T(\underline{X})$ de $g(\theta)$ est dit **sans biais (ESB)** de $g(\theta)$ si l'espérance de cet estimateur existe et est telle que*

$$E_\theta\left(T(\underline{X})\right) = g(\theta), \forall \theta \in \Theta$$

Dans le cas contraire, il est dit biaisé et son biais est $E_\theta\left(T(\underline{X})\right) - g(\theta)$

On considère le fait d'être sans biais comme une qualité de l'estimateur. L'une des raisons qui pousse qu'un estimateur soit sans biais est certainement que l'on a souvent affaire à des distributions dans lesquelles les fortes probabilités sont concentrées autour de l'espérance mathématique. S'il en est ainsi pour l'estimateur, on aura de belles probabilités d'obtenir l'estimation au voisinage de l'estimé. Sinon, on aurra de belles probabilités d'obtenir l'estimation au voisinage de l'espérance mathématique de l'estimateur, qui n'est pas l'estimé. Volentairement, on aura mis à coté, on aura créé, une erreur **systématique**. De plus, lorsque la taille de l'échantillon est grande, la plus part des estimateurs ont une distribution presque normale (théorème centrallimite) : les probabiltés sont bien, dès lors, concentrées autour de l'espérance mathématique.

Remarque 1.2.3 *Dans la recherche d'un estimateur sans biais, s'il existe, d'une fonction $g(\theta)$, il est souvent possible de le construire à partir d'un estimateur biaisé approprié en faisant quelques ajustement simples.*

Le choix de l'estimateur biaisé souhaitable ne suit aucune règle, toutefois la connaissance des espérances des premiers moments empiriques et la connaissance de l'espérance d'une statistique exhaustive, si elle existe, peuvent être utiles pour accomplir cette tâche.
Pour être plus clair supposons que $T(\underline{X})$ soit un estimateur biaisé de $g(\theta)$ tel que :

$$E\left(T(\underline{X})\right) = \alpha g(\theta) + \beta$$

où α et β sont des constantes réelles avec $\alpha \neq 0$.
La statistique $T'\left(\underline{X}\right) = \dfrac{T(\underline{X}) - \beta}{\alpha}$, $\alpha \neq 0$ est un estimateur sans biais de $g(\theta)$.

1.2.4 Estimateur Sans Biais de Variance Minimale

L'absence du biais est une propriété souhaitable d'un estimateur mais elle n'est pas, en soi, une propriété suffisante ; elle n'est même pas nécessaire pour définir un "bon"(selon un certain sens) estimateur, et cela même au sens du critère de l'erreur quadratique moyenne. Car un estimateur sans biais peut avoir une grande variance ce qui implique une erreur quadratique moyenne aussi grande que celle que peut avoir un estimateur biaisé avec un biais faible. Autrement dit, l'estimateur sans biais peut, même, être un estimateur inadmissible,

au sens de l'erreur quadratique moyenne, par rapport à l'ensemble de tous les estimateurs possibles.

Définition 1.2.8 *Un estimateur $T(\underline{X})$ de $g(\theta)$ est dit estimateur sans biais de variance minimale de $g(\theta)$ si :*

1. *Il est sans biais de $g(\theta)$,*
2. *$Var(T(\underline{X})) \leq Var(S(\underline{X}))$ pour tout estimateur sans biais $S(\underline{X})$ de $g(\theta)$.*

Remarque 1.2.4 *L'estimateur sans biais de variance minimale peut ne pas exister et s'il existe cela ne veut en aucune façon dire qu'il n'existe pas un estimateur biaisé dont la variance est strictement plus petite que celle de cet estimateur.*

Théorème 1.2.1 *Si l'estimateur sans biais de variance minimale existe alors il est unique.*

1.2.5 Famille complète et statistique complète

La notion de statistique complète est très importante, en particulier dans la recherche de l'estimateur sans biais de varaince minimale ainsi que dans le problème des tests d'hypothèses.

Famille complète de distribution

Définition 1.2.9 *La famille de distributions $f(x,\theta), x \in \mathfrak{X}$ et $\theta \in \Theta$, est dite famille complète, par rapport à θ, si et seulement si, il n'existe aucune fonction $H(\underline{X})$ Borel mesurable de la variable aléatoire X, dont l'espérance mathématique est nulle quelque soit θ, sauf si la fonction $H(\underline{X})$ elle même est nulle presque partout. C'est-à-dire $f(x,\theta)$ est complète, si et seulement si,*

$$E(H(\underline{X})) = 0, \forall\theta \in \Theta \Longrightarrow H(\underline{X}) = 0$$

presque partout.

Définition 1.2.10 *Une statistique $T(\underline{X})$ est dite complète si et seulement si, sa famille de distributions est complète.*

Théorème 1.2.2 *La famille de lois de Bernoulli $B(\theta)$ ainsi que la famille de lois Binomiales $B(n,\theta)$ sont complètes.*

Rôle de statistique exhaustive

La statistique exhaustive si elle existe, joue un rôle très important dans la recherche de l'estimateur sans biais de variance minimale, si ce dernier existe. Le théorème suivant dit de Rao-Blackwell montre que l'estimateur sans biais de variance minimale est toujours fonction d'une statistique exhaustive.

Théorème 1.2.3 *(Rao-Blackwell). Si T est un ESB pour $g(\theta)$ et $S(\underline{X})$ est une statistique exhaustive alors :*

- *La fonction $T^* = E(T/S)$ est une statistique;*
- *$T^* = E(T/S)$ est un ESB pour $g(\theta)$;*
- *$Var(T^*) \leq Var(T)$.*

Remarque 1.2.5 *$T^* = E(T/S)$ est un estimateur sans biais de $g(\theta)$ mais n'est pas nécessairement l'estimateur sans biais de variance minimale.*

Remarque 1.2.6 *On déduit du théorème précédent que dans la recherche de l'estimateur sans biais de variance minimale on doit se limiter, seulement, à la classe des estimateurs sans biais qui sont fonctions d'une statistique exhaustive si elle existe.*

Remarque 1.2.7 *Supposons qu'il existe un estimateur sans biais et un, uniquement, fonction de la statistique exhaustive $T(\underline{X})$; il est clair; dans ce cas, que $T^* = E(T/S)$ est lui même l'estimateur sans biais de variance minimale de $g(\theta)$.*

De la troisième remarque précédente on voit clairement l'importance de l'unicité de l'estimateur sans biais fonction d'une statistique exhaustive.

Le théorème suivant précise le rôle que joue une statistique exhaustive et complète si elle existe dans la recherche de l'estimation sans biais de variance minimale s'il existe.

Théorème 1.2.4 *(Lehmann-Scheffe). Si S est une statistique exhaustive et complète et si $T^* = T^*(S)$ est un ESB de $g(\theta)$ fonction de S alors T^* est l'unique ESB de variance minimale $(ESBVM)$ de $g(\theta)$*

Borne inférieur de Cramer-Rao

L'estimateur sans biais de variance minimale, d'une certaine fonction $g(\theta)$ du paramètre inconnu θ, est le "meilleur" estimateur sans biais, au sens de l'erreur quadratique moyenne, c'est-à-dire qu'il possède une petite varaince et donc la plus petite erreur quadratique moyenne dans la classe des estimateurs sans biais de $g(\theta)$.

On a vu précédemment que c'est seulement dans certains cas dont celui le cas où la statistique exhaustive et complète existe, que l'on peut établir l'existance de l'estimateur sans biais de variance minimale à partir d'un estimateur sans biais quelconque, s'il existe, et cela grâce aux théorèmes de Rao-Blackwell et Lehman-Sheffe.

Mais dans le cas général, même lorsque cet estimateur existe on n'est pas toujours en mesure de le trouver; il est alors important d'établir une certaine base de comparaison entre les variances des estimateurs sans biais, du paramètre inconnu θ (ou plus généralement d'une fonction $g(\theta)$).

Conditions de Régularité (pour l'existance de la borne de Cramer-Rao)

CR_1) Le support de la densité de probabilité de la population statistique ne dépend pas du paramètre inconnu θ.

CR_2) L'espace paramétrique, $\Theta =]a, b[$, $\forall x \in \mathfrak{X}$, où a et b sont des constantes réelles (finies ou infinies).

CR_3) La dérivée partielle $\dfrac{\partial L(\theta/\underline{x})}{\partial\theta}$ existe pour toute valeur θ de Θ et pour toute réalisation $\underline{x}$ de l'espace de l'échantillon (sauf peut être sur un ensemble de mesure nulle).

CR_4) $\int L(\theta/\underline{x})\, d\underline{x}$ peut être dérivée par rapport à θ sous le signe d'intégration.

CR_5) L'espérance $E\left(\dfrac{\partial L(\theta/\underline{x})}{\partial\theta}\right)^2 = \int \left(\dfrac{\partial L(\theta/\underline{x})}{\partial\theta}\right)^2 f(\underline{x},\theta)\, d\underline{x}$ existe et est strictement positive, $\forall\theta \in \Theta$ et $\forall\underline{x} \in \mathfrak{X}$.

CR_6) Il existe un estimateur sans bians $T(\underline{X})$ tel que l'intégrale $\int t(\underline{x})\, L(\theta, \underline{x})\, d\underline{x}$ peut être dérivée sous le signe d'intégration. Un estimateur qui possède cette dernière propriété est appellé estimateur régulier.

Théorème 1.2.5 ***Inégalité de Cramer-Rao.***

Soit une population statistique dont la famille de densités de probabilité est donnée par : $\{f(x,\theta)\,; x \in \mathfrak{X}, \theta \in \Theta\}$

Alors pour tout estimateur sans biais $T(\underline{X})$ de $g(\theta)$ on a, sous les conditions de régularité précédentes, l'inégalité suivante dite inégalité de Cramer-Rao :

$$Var(T(\underline{X})) \geq \frac{\left(g'(\theta)\right)^2}{E\left(\frac{\partial \log L(\theta/\underline{x})}{\partial\theta}\right)^2}$$

où $L(\theta/\underline{x})$ est la fonction de vraisemblance.

Définition 1.2.11 *La quantité strictement positive, calculée sur la base d'un échantillon de taille n, notée $I_n(\theta)$*

$$E\left(\frac{\partial \log L(\theta/\underline{x})}{\partial\theta}\right)^2,$$

est appelée ***Informtion de Fischer,*** *et la quantité strictement positive*

$$\frac{\left(g'(\theta)\right)^2}{E\left(\frac{\partial \log L(\theta/\underline{x})}{\partial\theta}\right)^2}$$

est appelée ***Borne (inférieure) de Cramer-Rao.***

Propriété

L'information de Fischer $I_n(\theta)$ peut être calculée par la formule suivante

$$I_n(\theta) = -E\left(\frac{\partial^2 \log L(\theta/\underline{x})}{\partial\theta^2}\right)$$

On souligne que pour plusieurs familles de lois, en particulier la famille exponentielle, le calcul de l'espérance de $-E\left(\frac{\partial^2 \log L(\theta/\underline{x})}{\partial\theta^2}\right)$ est nettement plus rapide que le calcul de l'espérance de la quantité aléatoire $E\left(\frac{\partial \log L(\theta/\underline{x})}{\partial\theta}\right)^2$; il est donc plus commode d'utiliser la formule de propriété.

1.2.6 Estimateur efficace

Le théorème de l'égalité de Cramer-Rao assure (sous les conditions de régularité) qu'on ne peut pas trouver un estimateur sans biais de $g(\theta)$ dont la variance est plus petite que

la borne inférieure de Cramer-Rao. Cependant, il ne nous indique rien sur la possibilité d'existance d'un estimateur sans biais de variance minimale, il serait souhaitable que sa varinace atteingne la borne inférieure de Cramer-Rao.

Définition 1.2.12 *On dit qu'un estimateur sans biais $T(\underline{X})$ de $g(\theta)$ est un estimateur efficace de $g(\theta)$ si et seulement si la variance de $T(\underline{X})$ atteint la borne inférieure de Cramer-Rao.*

Définition 1.2.13 *L'efficacité relative e_r de l'estimateur $T_1(\underline{X})$ de $g(\theta)$ par rapport à l 'estimateur $T_2(\underline{X})$ de $g(\theta)$ est définie par*

$$e_r = 100\frac{E\left(T_1 - g(\theta)\right)^2}{E\left(T_2 - g(\theta)\right)^2}\%$$

cas particulier

Si $T_1(\underline{X})$ et $T_2(\underline{X})$ sont deux estimateurs sans biais de $g(\theta)$ alors l'efficacité relative e_r; peut s'écrire sous la forme suivante

$$e_r = 100\frac{Var(T_1)}{Var(T_2)}\%$$

Propriété

L'estimateur efficace est un estimateur sans biais 100% efficace. Pour un estimateur sans biais donné, on calcule généralement sa variance et on la compare avec la borne inférieure de Cramer-Rao, afin de savoir si l'estimateur est efficace et d'avoir, dans le cas où il ne l'est pas, une idée sur l'écart entre sa variance et la borne de C.R, si elle existe.
Le théorème suivant fournit une condition nécessaire et suffisante pour qu'un estimateur sans biais soit efficace.

Théorème 1.2.6 *Soit $T(\underline{X})$ un estimateur sans biais de $g(\theta)$, sous les conditions de régularité CR_1 et CR_6, la variance de $T(\underline{X})$ atteint la borne inférieure de Cramer-Rao, autrement dit l'estimateur est efficace, si et seulement si, $\frac{\partial \log L(\theta/\underline{x})}{\partial\theta}$ peut être factorisée comme suit*

$$\frac{\partial \log L(\theta/\underline{x})}{\partial\theta} = A(\theta)\left(T(\underline{X}) - g(\theta)\right)$$

où $A(\theta)$ est une fonction qui ne dépend que de θ. Et dans ce cas l'information de Fischer $I_n(\theta)$ est donnée par

$$I_n(\theta) = A(\theta)\, g'(\theta)$$

Corollaire 1.2.1 *Si les conditions de régularité CR_1 à CR_6 sont satisfaites et si*

$$\frac{\partial \log L(\theta/\underline{x})}{\partial \theta} = A(\theta)\left(T(\underline{X}) - g(\theta)\right),$$

alors $T(\underline{X})$ est un estimateur efficace de θ dont la variance minimale est

$$Var\left(T(\underline{X})\right) = \frac{1}{A(\theta)} = \frac{1}{I_n(\theta)}$$

Remarque 1.2.8 *Si les conditions de régularité sont satisfaites, alors, la borne de Cramer-Rao existe; mais n'est pas nécessairement atteinte. Dans ce cas l'estimateur efficace n'existe pas. Néanmoins, il reste possible de trouver l'estimateur sans biais de variance minimale.*

Corollaire 1.2.2 *Si les conditions de régularité ne sont pas satisfaites alors l'efficacité absolue n'aura plus de sens. Cependant, il est toujours possible de chercher l'estimateur sans biais de variance minimale.*

1.2.7 Estimateur asymptotiquement efficace

Définition 1.2.14 *Sous les conditions de régularité un estimateur sans biais $T(\underline{X})$ de $g(\theta)$ dont l'efficacité absolue approche 100% quand n tend vers l'infini est appelé estimateur asymptotiquement efficace, c'est-à-dire;*

$$\lim_{n\to\infty} 100\frac{\left(g'(\theta)\right)^2}{I_n(\theta)\,Var(T)}\% = 100\%$$

1.3 Méthodes d'estimation ponctuelle

1.3.1 Introduction

Jusqu' à présent, on a introduit la notion d'estimation ponctuelle tout en citant quelques propriétés souhaitables des estimateurs, sans aborder la réponse à la question importante suivante :

Comment peut-on construire des estimateurs qui possèdent au moins quelques unes des propriétés désirables étudiées précédemment?

Il existe plusieurs méthodes pour la recherche de ces estimateurs. Le choix de la méthode à utiliser dépend quelque fois du problème (de l'estimation) posé, ainsi que tous les renseignements possibles que possède le statisticien.

Ainsi pour la recherche des estimateurs des paramètres inconnus $\underline{\theta} = (\theta_1, \theta_2, \ldots, \theta_m)$, en se basant seulement sur les informations contenues dans un $n-$échantillon, d'une famille de lois de probabilité donnée par la famille de densités dans le cas d'une variable aléatoire continue ou de lois de probabilité dans le cas où cette variable est discrète :

$$\{f(x; \theta_1, \theta_2, \ldots, \theta_m); x \in \mathfrak{X}, \underline{\theta} \in \Theta \subseteq \mathbb{R}^m\}$$

On peut utiliser l'une des méthodes bien connues suivantes :

· méthode *des moments*;

· méthode *du maximum de la fonction de vraisemblance*;

· méthode *Bayésienne*

Il existe par ailleurs une troisième méthode d'estimation dite Méthode des moindres carrés, cette méthode est fréquemment utilisée en statistique particulièrement dans l'analyse des modèles linéaires et des modèles de régression, dont l'application n'exige pas la connaissance de la famille de lois du variable aléatoire qui régit le phénomène étudié, mais uniquement l'existence de la moyenne et de la variance de cette variable.

On notera que ces méthodes d'estiamtion peuvent donner, comme on verra ultérieurement, des estimateurs différents pour les mêmes paramètres dans un même problème d'estimation.

1.3.2 Méthode du maximum de vraisemblance

Introduction

La méthode du maximum de vraisemblance est l'une des méthodes d'estimation qui convient au problème d'estimation des paramètres $\underline{\theta} = (\theta_1, \theta_2, \ldots, \theta_m)$ d'une population statistique dont la loi de probabilité dépend de ces paramètres.

Autrement dit lorsque la famille de lois est de la forme

$$\{f(x; \theta_1, \theta_2, \ldots, \theta_m); x \in \mathfrak{X}, \underline{\theta} \in \Theta \subseteq \mathbb{R}^m\}$$

Pour éclaircir l'idée, supposons pour le moment, que cette famille ne dépend que d'un seul paramètre scalaire $\theta \in \Theta \subseteq \mathbb{R}$.

En parcourant Θ, le paramètre θ engendre une famille de lois de probabilité pour la population statistique. La vraie loi, appartenant à cette famille, correspond à une seule valeur déterminée $\theta = \theta_0$, inconnue, appelée vraie valeur du paramètre θ.
Soit $\underline{X} = (X_1, X_2, \ldots, X_n)$ un échantillon aléatoire simple extrait de cette population. La famille de lois de cet échantillon est donné (sous l'indépendance et l'equidistributivité) par

$$f(\underline{x};\theta) = \prod_{j=1}^{n} f(x_j;\theta), \quad x \in \mathfrak{X}, \theta \in \Theta$$

On rappelle que $f(\underline{x};\theta)$, considérée comme fonction de la variable (déterministe) θ ($\underline{x}$ étant considérée fixé) est dite fonction de vraisemblance notée $L(\theta;\underline{x})$.
Si l'échantillon aléatoire simple X' a donné lieu à la réalisation particulière $\underline{x}' = \left(x_1', x_2', \ldots, x_n'\right)$, il est naturel de penser attribuer au paramètre inconnu θ la valeur θ_0 pour laquelle $\underline{x}'$ est la plus vraisemblable parmi toutes les autres réalisations $\underline{x}$.

Principe de la méthode

Le principe de la méthode du vraisemblance, est de chercher les valeurs des paramètres inconnus $\underline{\theta} = (\theta_1, \theta_2, \ldots, \theta_m)$ qui maximisent la fonction de vraisemblance $L(\theta; x')$ par rapport à θ pour tout x' fixé

Définition 1.3.1 *L'estimation par la méthode du maximum de vraisemblance, des paramètres $\underline{\theta} = (\theta_1, \theta_2, \ldots, \theta_m)$ de la population $f(x;\underline{\theta})$, est la valeur $\widehat{\theta}(\underline{x}) = \left(\widehat{\theta}_1(\underline{x}), \widehat{\theta}_2(\underline{x}), \ldots, \widehat{\theta}_n(\underline{x})\right)$ de $\underline{\theta}$, qui maximise la fonction de vraisemblance $L(\underline{\theta};\underline{x})$ par rapport à $\underline{\theta}$ pour $\underline{x}$ fixé.*

Autrement dit, la valeur $\widehat{\theta}(\underline{x})$ de $\underline{\theta}$ est une estimation du maximum de vraisemblance pour $\underline{\theta}$ si

$$L\left(\widehat{\theta}(\underline{x});\underline{x}\right) \geq L(\underline{\theta};\underline{x}), \quad \underline{\theta} \in \Theta \subseteq \mathbb{R}^m$$

Les statistiques $\widehat{\theta}(\underline{x}) = \left(\widehat{\theta}_1(\underline{x}), \widehat{\theta}_2(\underline{x}), \ldots, \widehat{\theta}_n(\underline{x})\right)$, sont appelées les estimateurs du maximum de vraisemblance des paramètres inconnus θ_1, θ_2 , ..., θ_m.

Remarque 1.3.1 *Si la fonction de vraisemblance est de la forme $L(\underline{\theta};\underline{x}) = C(\underline{x})g(\underline{\theta};\underline{x})$, où $C(\underline{x}) > 0$, $\forall \underline{x} \in \mathfrak{X}$; alors $C(\underline{x})$ n'intervient pas dans la recherche de la valeur de $\underline{\theta}$ qui maximise $L(C(\underline{x});\underline{x})$.*

En effet, la fonction de vraisemblance n'est pas nécessairement la fonction de l'échantillon $f(\underline{x};\underline{\theta})$ mais peut être une fonction positive qui lui est proportionnelle et dont le coefficient de proportionnalité est une fonction strictement positive de l'échantillon et ne dépend pas des paramètres inconnus en question.
Ce qui assure que les fonctions $L(\underline{\theta};\underline{x})$ et $g(\underline{\theta};\underline{x})$ atteignent leurs maximas pour la même abcisse $\widehat{\theta}(\underline{x})$, $\underline{x} \in \mathfrak{X}, \underline{\theta} \in \Theta \subseteq \mathbb{R}^m$. Ainsi on peut chercher l'estimation du maximum de vraisemblance en maximisant $g(\underline{\theta};\underline{x})$.

Remarque 1.3.2 *Les fonctions* $L(\underline{\theta};\underline{x})$ *et* $\log L(\underline{\theta};\underline{x})$, $(L > 0, \underline{x} \in \mathfrak{X}$ *et* $\underline{\theta} \in \Theta \subseteq \mathbb{R}^m)$ *atteignent leurs maximas pour la même valeur de* $\underline{\theta}$ *; alors il est très commode dans plusieurs cas (spécialement pour les familles de lois appartenant à la famille de lois exponnentielle) de maximiser plutôt* $\log L(\underline{\theta};\underline{x})$,

$$\max_{\underline{\theta}\in\Theta} L(\underline{\theta};\underline{x}) = \max_{\underline{\theta}\in\Theta} \log L(\underline{\theta};\underline{x})$$

Techiniques de différentiation

Soit une famille de lois de probabilités d'une population statistique telle que les conditions suivantes-dites conditions de régularité- soient satisfaites :

$\cdot \mathfrak{X} = \{x \in \mathbb{R}, f(x;\theta) > 0\}$ ne dépend pas des paramètres à estimer ;

$\cdot$ La fonction de vraisemblance $L(\underline{\theta};\underline{x})$ est continûment dérivable d'ordre deux.

Alors sous ces deux conditions (de régularité) l'estimation $\widehat{\theta}(\underline{x})$ par la méthode de vraisemblance, des paramètres $\underline{\theta} = (\theta_1, \theta_2, \ldots, \theta_m)$ est donnée par la résolution du système d'équations suivants :

A) Cas d'un paramètre scalaire

$$\frac{\partial L(\underline{\theta};\underline{x})}{\partial \theta} = 0; \quad \text{(équation normale)}$$
$$\left.\frac{\partial^2 L(\underline{\theta};\underline{x})}{\partial \theta^2}\right|_{\underline{\theta}=\widehat{\theta}(\underline{x})} < 0$$

ou encore par la résolution du système équivalent :

$$\frac{\partial \log L(\underline{\theta};\underline{x})}{\partial \theta} = 0; \quad \text{(équation normale)}$$

$$\left.\frac{\partial^2 \log L\left(\underline{\theta};\underline{x}\right)}{\partial\theta^2}\right\rfloor_{\underline{\theta}=\widehat{\theta}(\underline{x})} < 0$$

La statistique $\widehat{\theta}\left(\underline{x}\right)$ est l'estimateur du maximum de vraisemblance du paramètre inconnu θ à estimer.

B) Cas d'un paramètre vectoriel

L'estimation $\widehat{\theta}\left(\underline{x}\right) = \left(\widehat{\theta}_1\left(\underline{x}\right), \widehat{\theta}_2\left(\underline{x}\right), \ldots, \widehat{\theta}_n\left(\underline{x}\right)\right)$ du vecteur $\underline{\theta} = (\theta_1, \theta_2, \ldots, \theta_m)$ est donné sous les conditions de régularité par la résolution du système d'équations normales suivant :

$$\begin{cases} \frac{\partial L(\underline{\theta};\underline{x})}{\partial\theta_1} \\ \vdots \\ \frac{\partial L(\underline{\theta};\underline{x})}{\partial\theta_r} \\ \vdots \\ \frac{\partial L(\underline{\theta};\underline{x})}{\partial\theta_m} \end{cases}, \qquad \text{(équations normales)}$$

avec la condition que la matrice symétrique.

$$M = \left(..m_{ij}..\right)\rfloor_{\theta_i=\widehat{\theta}_i} \quad \text{où} \quad m_{ij} = -\frac{\partial^2 L\left(\underline{\theta};\underline{x}\right)}{\partial\theta_i\partial\theta_j}; \quad i,j = 1,\ldots,m$$

soit définie positive.

On peut, en tenant compte de la Remarque (1.3.2) , trouver l'estimation du maximum de vraisemblance par la résolution du système d'équations suivant

$$\begin{cases} \frac{\partial \log L(\underline{\theta};\underline{x})}{\partial\theta_1} \\ \vdots \\ \frac{\partial \log L(\underline{\theta};\underline{x})}{\partial\theta_r} \\ \vdots \\ \frac{\partial \log L(\underline{\theta};\underline{x})}{\partial\theta_m} \end{cases}, \qquad \text{(équations normales)}$$

avec la condition que la matrice symétrique

$$K = \left(..k_{ij}..\right)\rfloor_{\theta_i=\widehat{\theta}_i} \text{ où } k_{ij} = -\frac{\partial^2 \log L\left(\underline{\theta};\underline{x}\right)}{\partial\theta_i\partial\theta_j}; \quad i,j = 1,\ldots,m$$

soit définie positive.

La méthode du maximum de vraisemblance, comme on vient de voir, repose sur une idée intuitive. Néanmoins, la justification de sa popularité réside dans les diverses proptiétés intéressantes des estimateurs qu'elle fournit, en particulier les propriétés de convergence et la propriété d'invarance.

Propriétés des estimateurs du maximum de vraisemblance

Théorème 1.3.1 *(propriété d'invariance). Si* $\widehat{\theta}(\underline{X})$ *est un estimateur du maximum de vraisemblance de* θ *et* $g(.)$ *une fonction monotone, alors l'estimateur du maximum de vraisemblance de* $g(\theta)$ *est* $g\left(\widehat{\theta}(\underline{X})\right)$.

Théorème 1.3.2 *(Estimateur du MV et statistique exhaustive)*

S'il existe une statistique exhaustive $T(\underline{X})$ *pour* θ *alors l'estimateur du maximum de vraisemblance de* θ *est une fonction de cette statistique exhaustive.*

Théorème 1.3.3 *(Estimateur de MV et efficacité)*

Si l'estimateur efficace de θ *existe alors c'est l'estimateur du maximum de vraisemblance de* θ.

Théorème 1.3.4 *(propriété asymptotique)*

L'estimateur du maximum de vraisemblance est asymptotiquement normal. De plus sa matrice de covariance asymptotique est l'inverse de l'information de Fisher.

$$\sqrt{n}\left(\widehat{\theta}_n - \theta\right) \to N\left(0, \left[I(\theta)\right]^{-1}\right)$$

1.3.3 Méthode bayésienne

Si l'expérimentateur ou le statisticien disposant d'un échantillon aléatoire remarque ou a été informé que les paramètres $\theta_1, \theta_2, \ldots, \theta_m$ se comportent eux même comme des variables aléatoires dont la loi de probabiltité, appelée loi *a priori*, notée

$$P(\theta_1, \theta_2, \ldots, \theta_m), \underline{\theta} \in \Theta \subseteq \mathbb{R}^m,$$

alors, dans ce cas, il est évident, qu'il doit faire rentrer ces informations supplémentaires dans l'estimation de ces paramètres. La méthode d'estimation qui convient à ce genre de paramètre est connue sous le nom méthode bayésienne.

Formule de Bayes

On dispose d'un échantillon aléatoire simple $\underline{X} = (X_1, X_2, \ldots, X_n)$ extrait d'une population ayant une famille de lois $\{P(\underline{x}, \theta), \underline{x} \in \mathfrak{X}, \theta \in \Theta \subseteq \mathbb{R}\}$ dont la fonction de vraisemblance

s'écrit

$$L(\underline{\theta};\underline{x}) = \prod_{i=1}^{n} P(x_i,\theta)$$

La distribution conjointe de $(\underline{x},\theta)$ s'obtient par

$$P(\underline{x},\theta) = P(\underline{x}/\theta)\,P(\theta) \tag{1.3.1}$$

La formule de Bayes est basée sur la décomposition inverse de (1.3.1)

$$P(\underline{x},\theta) = P(\theta/\underline{x})\,P(\underline{x})$$

La densité de θ conditionnelle à l'observation $\underline{x}$ de l'échantillon $\underline{X}$ est dite densité *a postériori* de θ

$$P(\theta/\underline{x}) = \frac{P(\underline{x}/\theta)\,P(\theta)}{P(\underline{x})} \tag{1.3.2}$$

Le dénominateur ne dépond pas de θ.

$P(\underline{x})$ est appelée la densité *prédictive* de $\underline{X}$, (la loi marginale de $\underline{X}$). On peut écrire

$$P(\underline{x}) = C^{-1} = \begin{cases} \displaystyle\int_{\Theta} P(\underline{x}/\theta)\,P(\theta)\,d\theta, & \theta \text{ continu} \\ \displaystyle\sum_{\Theta} P(\underline{x}/\theta)\,P(\theta), & \theta \text{ discrèt} \end{cases}$$

(1.3.2) peut s'écrire sous la forme

$$P(\theta/\underline{x}) = CP(\underline{x}/\theta)\,P(\theta)$$

La quantité C, appelée constante de normalisation, nous permet d'écrire

$$P(\theta/\underline{x}) \propto P(\underline{x}/\theta)\,P(\theta)$$

où " $\propto$ " est le symbole de proportionnalité.

$P(\underline{x}/\theta)$, considérée comme fonction de θ, est appelée fonction de vraisemblance de θ sachant $\underline{x}$ qu'on notera $L(\theta/\underline{x})$.

La distribution a posteriori de θ peut être réécrite alors sous la forme suivante

$$P(\theta/\underline{x}) \propto L(\theta/\underline{x})\,P(\theta)$$

Etant donné la loi de probabilité a priori de θ et après que $\underline{x}$ soit observée, la densité a posteriori de θ, $P(\theta/\underline{x})$ calculée via le théorème de Bayes, est la base de toute analyse bayésienne, car toute inférense bayésienne concernant le paramètre θ dépend de cette loi.

Recherche des stratégies de Bayes

Considérons d'abord le cas où on a pas de données. Soit $\mathfrak{L}\left(\widehat{\theta},\theta\right)$ la perte associée à $\widehat{\theta}$, $\theta \in \Theta \subset \mathbb{R}^m$.

Définition 1.3.2 *On appelle risque moyen de Bayes associé à une loi a priori P l'espérance de* $R(T,\theta)$ *noté* $r_{_P}(\theta)$*:*

$$r_{_P}(\theta) = \begin{cases} \int\limits_{\Theta} R(T,\theta)P(\theta)\,d\theta, & \theta \text{ continu} \\ \sum\limits_{\Theta} R(T,\theta)P(\theta), & \theta \text{ discrèt} \end{cases}$$

Théorème 1.3.5 *La procédure de Bayes consiste à trouver* $\widehat{\theta}_0$ *qui minimise* $\int\limits_{\Theta} \mathfrak{L}\left(\widehat{\theta},\theta\right) P(\theta)\,d\theta$ *par rapport à* θ*.*

Si $\underline{X} = (X_1, X_2, \ldots, X_n)$ est observé, l'estimateur de Bayes $\widehat{\theta}_0(\underline{X})$ minise $\int\limits_{\Theta} \mathfrak{L}\left(\widehat{\theta}(\underline{x}),\theta\right) P(\theta/\underline{x})\,d\theta$

Remarque 1.3.3 $\int\limits_{\Theta} \mathfrak{L}\left(\widehat{\theta}(\underline{x}),\theta\right) P(\theta/\underline{x})\,d\theta$ *est appelé risque à postériori*

Corollaire 1.3.1 *La régle de Bayes (estimateur) est la moyenne de la loi à postériori si* $\mathfrak{L}\left(\widehat{\theta}(\underline{X}),\theta\right) = \left(\widehat{\theta}(\underline{X}) - \theta\right)^2, \left(\widehat{\theta},\theta\right) \in \Theta \times \Theta$ *est utilisée.*

Corollaire 1.3.2 *Si* $\mathfrak{L}\left(\widehat{\theta}(\underline{X}),\theta\right) = \left|\widehat{\theta}(\underline{X}) - \theta\right|$ *(fonction de perte écart absolu), alors la régle de Bayes est la médiane à postériori.*

Pour représenter nos croyances a priori sur les paramètres, on peut utiliser une classe de densité conjuguée [Reiffa et Scheaifer (1961)] . Lorsqu'aucune information sur les paramètres n'est connue, il est suggéré d'utiliser les lois a priori non informative, qui exprime notre état d'ignorance sur les paramètres. Nous présentons dans ce qui suit la loi a priori de Jeffreys (1961) et la loi a priori de référence.

Lois a priori conjuguées

Nous avons vu que l'essentiel de la méthode bayésienne consistait en l'introduction sur l'espace des paramètres Θ d'une distribution a priori que l'on supposait muni d'une densité $P(\underline{\theta})$. Cette distribution doit traduire la connaissance qu'a le statisticien des états de la nature. Il est alors commander de choisir la densité a priori $P(\underline{\theta})$ parmi une famille F de densités satisfaisant aux conditions suivantes :

· La famille F doit se prêter suffisamment au calcul analytique pour permettre le calcul de la densité a posteriori et du risque de Bayes associé.

· F doit être fermée en ce sens que la distribution a posteriori associée à un élément de F doit encore appartenir à F.

· F doit être assez riche, en d'autres termes doit dépendre de suffisamment de paramètres pour exprimer convenablement les idées a priori du statisticien.

Lois a priori non informative

Le choix d'une distribution a priori non informative conduit souvent à la spécification d'une mesure et non d'une probabilité.
La procédure de spécification d'une mesure a priori non informative revient à définir une mesure sur l'espace paramétrique Θ à partir du mécanisme d'échantillonnage décrit par l'échantillon $\underline{X} \in \mathfrak{X}$ et la probabilité d'échantillonnage.

Mesures à priori Si $P(\underline{x}/\theta)$, $\theta = (\theta_1, \theta_2, \ldots, \theta_m) \in \Theta \subseteq \mathbb{R}^m$ est la distribution de l'échantillon $\underline{X}$, une mesure a priori sera caractérisée par sa densité $P(\theta)$ par rapport à la mesure de Lebesgue. Cette densité est une fonction réelle positive ou nulle mais d'integrale non nécessairement finie.

Mesure a priori de Jeffreys Le mode de spécification d'une mesure a priori non informative connu sous le non de mesure a priori de Jeffreys consiste à assigner à un modèle d'échantillonnage caractérisé par sa vraisemblance $P(\underline{x}/\theta)$ la mesure a priori de densité

$$P(\theta) = [\det I(\theta)]^{1/2}$$

par rapport à la mesure de Lebesgue.

La matrice $I(\theta)$ est la matrice d'information de Fisher d'élèment $I_{i,j}(\theta), i, j = 1, 2, \ldots, m$ définit comme étant

$$I_{i,j}(\theta) = -E\left(\frac{\partial^2}{\partial\theta_i \partial\theta_i} \ln P(\underline{x}/\theta)\right)$$

Les conditions de régularité sont supposées vérifier.

Mesure a priori de référence L'analyse de référence développé par Bernardo (1979) est exposée de manière détaillée dans Bernardo et Smith (1994) est un mode général de spécification d'une loi a priori contenant aussi peu d'information que possible.
Soit un modèle d'échantillonnage décrit par $P(\underline{x}/\underline{\theta})$ et une densité à priori $P(\underline{\theta})$. La densité marginale de $\underline{X}$ est noté $P(\underline{x})$ et $P(\underline{\theta}/\underline{x})$ désigne la densité a postériori. On a deux distributions de probabilité sur $\Theta \times \mathfrak{X}$: La loi jointe $P(\underline{\theta}) P(\underline{x}/\underline{\theta})$ et le produit de deux marginales $P(\underline{\theta}) P(\underline{x})$ (paramètre et échantillon indépendants). L'information apportée par un modèle statistique est naturellement fonction de la dépendance entre $\underline{X}$ et $\underline{\theta}$, il est naturel de la mesurer en comparant $P(\underline{\theta}) P(\underline{x}/\underline{\theta})$ et $P(\underline{\theta}) P(\underline{x})$. La divergence de Kullback est utilisée comme mode de comparaison :

$$\begin{aligned} K_p &= \int \ln\left[\frac{p(\underline{\theta}) P(\underline{x}/\underline{\theta})}{p(\underline{\theta}) P(\underline{x})}\right] P(\underline{\theta}) P(\underline{x}/\underline{\theta}) \, d\underline{x} d\theta \\ &= \int \left[\ln \frac{P(\underline{x}/\underline{\theta})}{P(\underline{x})}\right] P(\underline{\theta}) P(\underline{x}/\underline{\theta}) \, d\underline{x} d\theta \end{aligned}$$

Pour une taille d'échantillon finie, K_p est en général un nombre positif et on appellera a priori de référence la probabilté a priori qui minimise K_p.

1.4 Estimation par région de confiance

Introduction

Supposons que l'on observe un échantillon aléatoire simple $\underline{X}$, de taille n, pélevé dans une population statistique dont la famille de densités ou, plus généralement, la famille de lois de probabilité est donnée sous la forme :

$$\{f(\underline{x}; \theta_1, \theta_2, \ldots, \theta_m); \underline{x} \in \mathfrak{X}, \underline{\theta} \in \underline{\Theta} \subseteq \mathbb{R}^m\}$$

Le problème de l'estimation ponctuelle consiste, comme on l'a vu, en la recherche et l'étude des propriétés d'un estimateur ponctuel

$$\widehat{\theta}\left(\underline{X}\right)=\left(\widehat{\theta}_1\left(\underline{x}\right),\widehat{\theta}_2\left(\underline{x}\right),\ldots,\widehat{\theta}_n\left(\underline{x}\right)\right),$$

du paramètre vectoriel

$$\underline{\theta}=\left(\theta_1,\theta_2,\ldots,\theta_m\right)$$

Au lieu de chercher un estimateur ponctuel $\widehat{\theta}\left(\underline{X}\right)$ de $\underline{\theta}$, on peut s'intéresser à la recherche d'un sous ensemble aléatoire $\underline{\Theta}^*\left(\underline{X}\right)$ de $\underline{\Theta}$ dont le nombre moyen de réalisations $\underline{\Theta}^*\left(\underline{x}\right)$, en répétant l'observation de l'échantillon plusieurs fois, couvre la vraie valeur de $\underline{\theta}$, $100\left(1-\alpha\right)\%$ de fois où α est une probabilité non nulle fixée à priori et est dite probabilité de risque. La probabilité $1-\alpha$ est appelée *niveau de confiance* ou coefficient de confiance lié à la région aléatoire $\underline{\Theta}^*\left(\underline{x}\right)$.
Le problème statistique qui consiste à chercher une région aléatoire $\underline{\Theta}^*\left(\underline{x}\right)$ qui contienne ou couvre la vraie valeur de $\underline{\theta}$ avec une probabilité donnée $\left(1-\alpha\right)$, $0<\alpha<1$, c'est à dire

$$P\left(\underline{\Theta}^*\left(\underline{x}\right)\ \text{contient}\ \underline{\theta}\right)=1-\alpha$$

est connu académiquement sous le nom : Estimation par Région de Confiance. La région aléatoire $\underline{\Theta}^*\left(\underline{X}\right)$ est dite région de confiance de niveau de confiance $1-\alpha$.
Si $\underline{\theta}$ est un paramètre scalaire, c'est-à-dire que $\theta\in\underline{\Theta}=\Theta\subseteq\mathbb{R}$, alors la région aléatoire de confiance se réduit à un intervalle aléatoire $I\left(\underline{X}\right)$ de la forme

$$\underline{\Theta}^*\left(\underline{X}\right)=I\left(\underline{X}\right)=\left]\theta_2\left(\underline{X}\right),\theta_1\left(\underline{X}\right)\right[\subseteq\Theta\subseteq\mathbb{R};$$

telle que

$$P\left(\theta\in\left]\theta_2\left(\underline{X}\right),\theta_1\left(\underline{X}\right)\right[\right)=1-\alpha\ ,\ \text{avec}\ \alpha\in\left]0,1\right[$$

cet intervalle aléatoire est dit intervalle (aléatoire) de confiance de niveau $100\left(1-\alpha\right)\%$.

L'étendue d'un intervalle aléatoire de confiance de niveau $100\left(1-\alpha\right)\%$ est définie comme suit

$$e=\theta_1\left(\underline{X}\right)-\theta_2\left(\underline{X}\right)$$

Intervalle de confiance bilatéral

Supposons qu'on désire avoir un intervalle aléatoire de confiance de niveau $1-\alpha$ où $\alpha \in]0,1[$. Soient, à cet effet, α_1 et α_2 deux probabilité telles que $\alpha_1+\alpha_2=\alpha$, et supposons que l'on puisse trouver deux statistiques $\theta_1(\underline{X})$ et $\theta_2(\underline{X})$ telles que l'intervalle aléatoire $[\theta_1(\underline{X}),\ +\infty[$ (resp $]-\infty,\ \theta_2(\underline{X})]$) couvre la vraie valeur de θ avec la probabilité α_1 (resp α_2).

Autrement dit soient $\theta_1(\underline{X})$ et $\theta_2(\underline{X})$ qui vérifient le système

$$\begin{cases} P(\theta \in [\theta_1(\underline{X}),\ +\infty[) = \alpha_1 \\ P(\theta \in]-\infty,\ \theta_2(\underline{X})]) = \alpha_2 \end{cases}$$

Alors en décomposnt l'espace d'épreuve Ω de la manière suivante :

$$\Omega = \{\omega : \theta \in]\theta_2(\underline{X}), \theta_1(\underline{X})[\} \cup \{\omega : \theta \in [\theta_1(\underline{X}),\ +\infty[\} \cup \{\omega : \theta \in]-\infty,\ \theta_2(\underline{X})]\}$$

Du fait que les événements aléatoires, du seconde membre de l'égalité précédente sont disjointe, en prenant la probabilité des deux membres de l'expression précédentes, on obtient

$$P(\Omega) = P(\theta \in]\theta_2(\underline{X}), \theta_1(\underline{X})[) + P(\theta \in [\theta_1(\underline{X}),\ +\infty[) + P(\theta \in]-\infty,\ \theta_2(\underline{X})])$$

d'où il découle

$$P(\theta \in]\theta_2(\underline{X}), \theta_1(\underline{X})[)\, 1-(\alpha_1+\alpha_2) = 1-\alpha$$

Cette expression s'écrit, encore, sous la forme équivalente

$$P(\theta_2(\underline{X}) < \theta < \theta_1(\underline{X})) = 1-\alpha;$$

ainsi, on déduit l'intervalle aléatoire de confiance de niveau $1-\alpha$ pour θ :

$$I(\underline{X}) =]\theta_2(\underline{X}), \theta_1(\underline{X})[$$

qui, en répétant l'échantillonnage plusieurs fois, couvre la vraie valeur du paramètre inconnu θ, $100(1-\alpha)\,\%$.

Définition 1.4.1 *Un intervalle aléatoire de confiance* ***bilatéral*** *est un intervalle dont les deux bornes, inférieure et supérieure $\theta_2(\underline{X})$ et $\theta_1(\underline{X})$ respectivement, sont des statistiques (fonctions mesurables de l'échantillon ne dépendant d'aucun paramètre inconnu) vérifiant le système d'équation suivant*

$$\begin{cases} P(\theta \in [\theta_1(\underline{X}),\ +\infty[) = \alpha_1 \\ P(\theta \in]-\infty,\ \theta_2(\underline{X})]) = \alpha_2 \end{cases}$$

Construction des intervalles de confiance

L'objet de ce paragraphe est la construction des intervalles de confiance, à l'aide des méthodes suivantes :

· Méthode de la fonction pivotale

· Méthode générale

Méthode de la fonction pivotale

Soit $\underline{X}$ un échantillon aléatoire de taille n issu d'une population dont la famille de densités, dépend d'un paramètre réel inconnu θ

$$\{f(\underline{x};\theta);\underline{x}\in\mathfrak{X},\ \theta\in\underline{\Theta}\subseteq\mathbb{R}\}$$

Supposons que l'on dispose :

1. d'une statistique exhaustive $T(\underline{X})$, si elle existe, ou dans le cas échéant, de l'estimateur du maximum de vraisemblance de θ.

2. d'une variable aléatoire $U(T,\theta)$, fonction de la statistique exhaustive ou de l'estimateur du maximum de vraisemblance T, du paramètre θ en question, vérifiant les deux conditions suivantes :

a. la loi de probabilité de la variable aléatoire $U(T,\theta)$ est complètement spécifiée (elle ne dépend d'aucun paramètre inconnu)

b. la variable aléatoire $U(T,\theta)$ est une fonction monotone en θ.

Définition 1.4.2 *la fonction aléatoire $U(T,\theta)$ qui satisfait les deux conditions précédentes a et b est dite fonction pivotale pour le paramètre θ.*

Si l'on possède une fonction pivotale $U(T,\theta)$ pour θ alors, pour une probabilité de risque $\alpha\in]0.1[$ donné a priori, on peut construire un intervalle de confiance bilatéral pour θ à coefficient de confiance $1-\alpha$ de la manière suivante :

Soient α_1 et α_2 deux probabilités telles que $\alpha_1+\alpha_2=\alpha$. Comme la loi de probabilité du variable aléatoire $U(T,\theta)$ est connue et ne dépend d'aucun paramètre inconnu, on peut trouver, alors, les nombres réels u_1 et u_2 qui ne dépendent pas de θ tels que :

$$\begin{cases} P(U(T(\underline{X}),\theta)>u_1)=\alpha_1 \\ P(U(T(\underline{X}),\theta)\leq u_2)=1-\alpha_1 \end{cases}$$

ou encore

$$\begin{cases} P\left(U\left(T\left(\underline{X}\right),\theta\right) > u_1\right) = 1-\alpha_1 \\ P\left(U\left(T\left(\underline{X}\right),\theta\right) \leq u_2\right) = \alpha_2 \end{cases}$$

ce qui peut s'écrire sous la forme

$$\left(u_1 < P\left(U\left(T\left(\underline{X}\right),\theta\right) < u_2\right) = 1-\alpha \right. \tag{1.4.1}$$

La solution en θ, de cette double inégalité est un intervalle de confiance de niveau $100\left(1-\alpha\right)\%$. Si $U\left(T\left(\underline{X}\right),\theta\right)$ est monotone croissante en θ, et si on peut résoudre explicitement en θ la double inégalité intervenante dans la formule (1.4.1), alors cette dernière peut s'écrire sous la forme équivalente

$$P\left(U^{-1}\left(u_1,T\right) < \theta < U^{-1}\left(u_2,T\right)\right) = 1-\alpha,$$

ou encore

$$P\left(\theta \in \left]U^{-1}\left(u_1,T\right),U^{-1}\left(u_2,T\right)\right[\right) = 1-\alpha, \quad \text{avec } \alpha \in \left]0.1\right[$$

ce qui montre que l'intervalle aléatoire $\left]U^{-1}\left(u_1,T\right),U^{-1}\left(u_2,T\right)\right[$ est un intervalle aléatoire de confiance de niveau $1-\alpha$ pour θ.

Les intervalles crédibles

Nous nous plaçons dans le cadre bayésien.

Définition 1.4.3 *Un ensemble C_θ dans l'espace des paramètres Θ est dit $(1-\alpha)-$ intervalle crédible si et seulement si*

$$p(\theta \in C_\theta/\underline{x}) = 1-\alpha$$

Les intervalles H.P.D crédibles

Définition 1.4.4 *Soit $P(\underline{\theta}/\underline{x})$ la densité a posteriori de $\underline{\theta}$. La région $\mathcal{C} \subset \Theta$ est dite HPD (Highest Posterior Density) crédible de niveau de confiance $1-\alpha$ si*

$\cdot P(\underline{\theta} \in \mathcal{C}/\underline{x}) = 1-\alpha$

$\cdot \forall \underline{\theta} \in \mathcal{C}, \underline{\theta}' \notin \mathcal{C} : P(\underline{\theta}/\underline{x}) \geq P(\underline{\theta}'/\underline{x})$

Chapitre 2

Généralités sur les modèles de capture-recapture

2.1 Introduction

Avant toute mise en oeuvre d'une procédure d'échantillonnage de capture-recapture, il convient de bien définir l'objectif recherché, notre intérêt dans ce travail est d'estimer la taille d'une population fermée.

De manière générale, dans la procédure de capture-recapture, les individus d'une population sont capturés, marqués grâce à un code d'identification individuel puis relâchés dans la population. Cette population est échantillonnée séquentiellement à des occasions discrètres. A chaque occasion, de nouveaux individus sont donnés marques puis sont soumis à nouveau à l'échantillonnage, les individus déja identifiés sont répertoriés puis remis aussi dans la population. Il peut s'agir indifféremment d'identification visuelle ou bien de capture physique.

Dans ce chapitre on présente des notions générales consernant les modèles de capture-recapture (à temps discret, uni-état), dont'on définit les modèles homogénes et hétérogènes. Ces différents modèles de capture-recapture ont été suggérés par Otis et al (1978).

2.2 Expérience de capture-recapture

2.2.1 Zone et période d'étude

Avant de commencer une expérience de capture-recapture on délimite la zone d'étude (plus la zone d'étude est étendue, plus exacts seront les résultats). Dans le cas d'une population animale, le plan d'échantillonnage nécessite l'installation de pièges uniformement dans toute la zone d'étude pour éviter les situations où il sera impossible de piéger certains individus. La nature des pièges utilisés depend de la population étudiée (trappes de capture, filets, pièges photographiques...). L'éxpérience de capture-recapture est effectuée sur une période bien précise.

2.2.2 Population à estimer

Population animale

Les expériences de capture-recapture sont souvent et généralement menées sur les populations d'animaux, des petits mammifères peu visibles le jour et qui circulent surtout la nuit (poissons, oiseaux, ...).

Population humaine

Ces expériences de capture-recapture peuvent être aussi effectuées sur des populations humaines dans des circonstances bien définies (période, zone, cause, conséquences...).

Si la population n'est affectée par aucun changement sur les plans des naissances, des décès, d'immigration ou d'émigration pendant le processus d'échantillonnage, elle est dite population fermée, si non elle est ouverte.

2.2.3 Echantillonnage de la population

Soit une population fermée de taille N inconnue (à estimer). On tire un échantillon initial de taille n_0, tous les individus capturés sont marqués avant d'être remis dans la population. A la deuxième occasion de capture, on tire un deuxième échantillon indépendamment du premier de taille n_1, on compte le nombre d'individus marqués m_1 et on marque les non marqués, on

refait la procédure à t, $(t \geq 2)$ occasions de capture. La probabilité d'observer un individu est appellée probabilité de capture (efficacité du piège dans le cas d'une population animale). Les expériences de capture-recapture les plus simple possèdent deux occasions de capture, l'estimateur qui découle est connu sous le nom d'estimateur de Peterson, l'hypothèse de base est que la proportion des individus marqués observés dans l'échantillon de recapture est égale à leur proportion dans la population échantillonnée :

$$\frac{m_1}{n_1} = \frac{n_0}{N} \text{ d'où : } N = \frac{n_1 n_0}{m_1}$$

Vu que : la probabilité de $m_1 = 0$ est finie c'est-à-dire il n'y a pas d'individus marqués à la seconde occasion de capture, alors l'estimateur de Peterson peut avoir un biais infini, il a été corrigé par Chapman et Lincoln (1951) , [27] comme suit :

$$\widehat{N} = \frac{(n_0 + 1)(n_1 + 1)}{(m_1 + 1)} - 1 \tag{2.1.1}$$

2.3 Les modèles de capture-recapture fondamentaux

Otis, Burnham, White et Anderson [22] ont developpé les méthodes d'estimation de ces expériences : le modèle de capture-recapture peut incorporer trois sources de variations pour les probabilités de capture des individus :

1. Effet temporel ($time : t$) : cause les variations de probabilité de capture entre les occasions de capture (dans le cas d'une population animale, une nuit froide peut réduire la mobilité des animaux ce qui rend la probabilité de capture plus faible)

2. Hétérogénéité ($heterogeneity : h$) : la probabilité de capture varie d'un individu à un autre (cette différence est due aux facteurs physiques tels que : l'âge et le sexe qui ne sont généralement pas observés).

3. Comportement des individus ($behavioral : b$) : la première capture d'un individu change son comportement ce qui mène à la variation de sa probabilité de capture, (un animal est capturé la première fois peut devenir plus prudent à la suite de sa capture).

Huit modèles fondamentaux de capture-recapture pour une population fermée ont été élaborés pour tenir compte des sources de variations des probabilités de capture : M_t, M_b, M_h, ajoutant les modèles qui incorporent plus d'une source de variation : M_{tb}, M_{th}, M_{bh},

M_{tbh}. Le modèle le plus simple n'admet aucune source de variation des probabilités de capture, noté M_0.

2.3.1 Définitions et notation

On considère l'expérience de capture-recapture definie antérieurement. Pour tous couple (i, j) tel que : $(i = 1, \ldots, N,\ j = 1, \ldots, t)$, on définit :

p_{ij} = la probabilité de capture de l'individu i à l'occasion de capture j.

c_{ij} = la probabilité de recapture de l'individu i à l'occasion de capture j après sa capture initiale.

On a vu que : les sources de variations des p_{ij} sont :

1. l'occasion de capture (t)
2. la réponse des individus aux captures (b)
3. la différence entre individus (h)

En tenant compte de la troisième source de variation, il resulte : les modèles homogènes et les modèles hétérogènes.

2.3.2 Modèles homogènes

Ces modèles supposent l'homogénéité des individus, ils possédent que deux sources de variation des probabilités de capture p_{ij} : l'occasion de capture (t) et la réponse des individus à leur capture initiale (b), ce qui donne les modèles suivants : M_0, M_t, M_b, M_{tb}.

Modèle M_0

Le modèle M_0 est le plus simple des huit modèles présentés, il ne considère aucune source de variation des probabilités de capture.

Hypothèse : tous les individus ont la même probabilité de capture à chaque occasion, et cette probabilité demeure constante durant l'expérience :

$$\forall i = 1, \ldots, N,\ \forall j = 1, \ldots, t \text{ on a : } p_{ij} = p,$$

Ce modèle possède deux paramètres : $N \in \mathbb{N}^*$ et $p \in (0.1)$.

Modèle M_t

Le modèle M_t admet une seule source de variation de probabilités de capture : les occasions de capture (t) . Il suppose que les t occasions de capture sont indépendantes. Cette procédure a été proposée pour la première fois par Darroch [8]. Cependant, ce modèle ne serait adéquat que si les pièges de capture étaient différents à chaque occasion, même si la méthode de capture est unique pour toutes les occasions de capture. Ce modèle est appliqué dans plusieurs domaines : écologie, démographie, épidémiologie,...

Hypothèse : A chaque occasion de capture, tous les individus ont la même probabilité d'être capturés c'est-à-dire :

$$p_{ij} = p_j, \forall i = 1, \ldots, N, \ \forall j = 1, \ldots, t \quad ,$$

c'est-à-dire p_j est la probabilité de capture d'un individu quelconque à l'occasion j, ainsi le modèle de capture-recapture M_t admet $t+1$ paramètres : $p_1, \ldots, p_t, N$.

Exemple: *Estimation du nombre d'erreurs dans un produit software [32]:*

Deux lecteurs ont vérifié deux copies identiques d'un produit software. Le premier a détecté 30 erreurs, le deuxième en a détecté 24, en comparant les deux résultats finaux, ils ont constaté que le nombre d'erreurs détectées par chacun est de 20.
Quel est le nombre d'erreurs non détecté par l'un et non par l'autre?

On note:
n_1 le nombre d'erreurs détectés par le premier lecteur
n_2 le nombre d'erreurs détectés par le deuxième lecteur
n_{12} le nombre d'erreurs détectés par les deux lecteurs.

1^{er} *lecteur*	$2^{ème}$ *lecteur*		
	erreurs détectés	*erreurs non détectés*	
erreurs détectés	n_{12}	$n_1 - n_{12}$	n_1
erreurs non détectés	$n_2 - n_{12}$	$N - n_1 - n_2 + n_{12}$	$N - n_1$
	n_2	$N - n_2$	N

En appliquant l'estimateur de Lincoln-Peterson (2.1.1) sur les données de ce problème, on trouve $\widehat{N} = 36$.

La méthodologie des modèles de capture-recapture peut être généralisée au cas où il y aurait plusieurs lecteurs.

On suppose que :

• il y a m lecteurs, chaqu'un travaille indépendamment de l'autre, sur les mêmes copies de produit.

• la probabilité que le lecteur i détecte une erreur donnée est p_i indépendamment des autres erreurs détectées par le lecteur i ou autres lecteurs .

Soit:

· n_i le nombre total d'erreurs détectées par le lecteur i $(i = 1, 2, \ldots, m)$.

· u_i le nombre total des erreurs non détectées par le lecteur i mais détectées par d'autres lecteurs.

· n_{ij} le nombre d'erreurs détectées par les deux lecteurs i et j à la fois.

· u_{ij} le nombre d'erreurs non détectées ni par le lecteur i ni par le lecteur j mais par d'autres.

Si $n_{_T}$ est le nombre d'erreurs découvertes par les m lecteurs et n_0 le nombre d'erreurs non détectées par aucun lecteur, alors le nombre d'erreurs dans le document software est :

$$N = n_{_T} + n_0$$

On considère les deux lecteurs i et j parmis m lecteurs. Les données enregistées sont réparties dans le tableau suivant

lecteur i	*lecteur j*		*total*
	erreurs détectés	*erreurs non détectés*	
erreurs détectés	n_{ij}	$n_i - n_{ij}$	n_i
erreurs non détectés	$n_j - n_{ij}$	$u_{ij} + n_0$	$u_i + n_0$
	n_j	$u_j + n_0$	$N = n_T + n_0$

Trouvons l'estimateur du maximum de vraisemblance $(EMV\ )$ $\widehat{N}$ qui maximise la fonction de vraisemblance suivante :

$$L(N) = \begin{pmatrix} N \\ n_{_T} \end{pmatrix} \prod_{i=1}^{m} p_i^{n_i} (1 - p_i)^{N - n_i},$$

La fonction log-vraisemblance pour estimer les paramètres $(N,\ p_1, p_2, \ldots, p_m)$ est donnée par

$$\ln L(N) = \ln \begin{pmatrix} N \\ n_{_T} \end{pmatrix} + \sum_{i=1}^{m} n_i \ln p_i + \sum_{i=1}^{m} (N - n_i) \ln(1 - p_i) \tag{2.3.2}$$

Pour N donné, l'EMV de p_i est obtenu par

$$\widehat{p_i} = \frac{n_i}{N}, \quad i = 1, 2, \ldots, m. \tag{2.3.3}$$

En remplacant p_i dans **2.3.2** par $\widehat{p_i}$ dans **2.3.3** on trouve la fonction log-vraisemblance suivante

$$\begin{aligned} \ln L(N) &= \ln \begin{pmatrix} N \\ n_{_T} \end{pmatrix} + \sum_{i=1}^{m} n_i \ln \left(\frac{n_i}{N}\right) + \sum_{i=1}^{m} (N - n_i) \ln \left(\frac{N - n_i}{N}\right) \qquad (2.3.4) \\ &= \ln N! - \ln(N - n_T)! + \sum_{i=1}^{m} (N - n_i) \ln(N - n_i) - Nm \ln N + cste. \end{aligned}$$

Un estimateur de N est l'EMV approximé de Darroch, est la solution unique de l'équation suivante (Darroch, 1958), [8] :

$$1 - \frac{n_T}{N} = \prod_{i=1}^{m} (1 - \frac{n_i}{N}), \quad \text{avec } N \geqslant n_T. \tag{2.3.5}$$

Modèle M_b

Le modèle M_b suppose que la seule source de variation de probabilités de capture est la réponse de l'individu à sa capture initiale.

Hypothèse: au début de l'expérience, tous les individus admettent la même probabilité de capture p. Dès qu'un individu non marqué est capturé, sa probabilité de capture varie de p à c, et il garde cette nouvelle probabilité jusqu'à la fin de l'expérience, même s'il est recapturé dans les prochaines occasions de capture.

Ce modèle suppose que tous les individus réagissent de la même manière à la première capture, c'est-à-dire que tous les individus marqués admettent la même probabilité de capture.

Remarque : il est plus logique que les probabilités de capture varient suite aux captures successives (après la $2, 3, \ldots$). Otis et al,[22] ont montré que cette généralisation n'a pas d'effet sur l'estimation de N.

Ce modèle admet donc trois paramètres :

-la taille de la population N

-la probabilité de capture d'un individu non marqué p

-la probabilité de capture d'un individu marqué c

Modèle M_{tb}

Le modèle M_{tb} possède deux sources de variation de probabilités de capture : les occasions de capture (t) et la réponse des individus à la première capture (b).

Hypothèse: A chaque occasion de capture j, tous les individus non marqués ont la même probabilité de capture p_j, et tous les individus marqués ont la même probabilité de capture c_j, autrement dit

$$\begin{aligned} p_{ij} &= p_j, \forall i = 1, \ldots, N, \ \forall j = 1, \ldots, t \\ c_{ij} &= c_j, \forall i = 1, \ldots, N, \ \forall j = 2, \ldots, t \quad , \end{aligned}$$

donc, ce modèle possède $2t$ paramètres, $p_1, \ldots, p_t, c_2, \ldots, c_t, N$.

2.3.3 Modèles hétérogènes

On a vu que les probabilités de capture varient selon trois sources : l'effet des occasions de capture ($t : time$), la réaction des individus à la première capture ($b : behavioral$) et les différences entre les individus ($h : heterogeneity$) .

les modèles hétérogènes supposent que les individus sont hétérogènes entre eux À chaque modèle homogène correspond un modèle hétérogène :

Tableau 2.1: Correspondance entre les modèles homogènes et hétérogènes.

Modèles homogènes	Modèles hétérogènes
M_0	M_h
M_t	M_{th}
M_b	M_{bh}
M_{tb}	M_{tbh}

Modèle M_h

Hypothèse : chaque individu de la population a sa propre probabilité de capture indépendamment des autres, il l'a conserve durant l'expérience.
Ce modèle admet $N+1$ paramètres : $p_1, \ldots, p_N$, N.
Dans ce cas, p_i sont des paramètres de nuisance, qui sont inestimables.

Différentes approches ont été élaborées pour traiter ce problème : Norris et Pollock , Pledger[24] ont suggéré la partition de la population en groupes homogènes selon des critères physiques (âge, sexe,...).

Modèle M_{th}

Le modèle M_{th} possède deux sources de variation de probabilités p_{ij} : l'effet des occasions de capture (t) , et l'hétérogénèité (h) .
Hypothèse : à chaque occasion de capture, chaque individu possède sa propre probabilité de capture indépendamment des autres.

Les travaux de Agresti [2] ont traité ce modèle. Ce dernier possède $Nt+1$ paramètres.

Modèle M_{bh}

Le modèle suppose deux sources de variation de probabilités de capture : la réponse de l'individu à la première capture et la différence de comportement entre les individus.
Hypothèse : chaque individu i $(i = 1, \ldots, N)$ a deux probabilités de capture : une avant qu'il ne soit capturé pour la première fois p_i, et une après c_i.
Le modèle admet $2N+1$ paramètres : $p_1, \ldots, p_N, c_1, \ldots, c_N$ et N .

Modèle M_{tbh}

Ce modèle considère les trois sources de variation de probabilités de capture : l'occasion de capture (t), la réponse des individus au premier capture (b) , la différence entre le comportement des individus (h).
Tous les modèles précèdents sont des cas particuliers du modèle M_{tbh}.
Les paramètres du modèle sont :
$\cdot Nt$ probabilités de capture : $p_{ij}, i = 1, \ldots, N,\ j = 1, \ldots, t$

· $(t-1)\,N$ probabilités de recapture (probabilités de capturer les individus déja capturés au moin une fois) $c_{ij} : i = 1, \ldots, N,\ j = 2, \ldots, t$

·la taille de la population N.

Donc ce modèle possède $Nt + (t-1)\,N + 1$ paramètres, il est le plus réaliste mais sa complexité ne permet pas d'estimer la taille de la population.

Chapitre 3

Analyse bayésienne des données de capture-recapture

3.1 Introduction

Les statistiques bayésiennes forment des travaux mathématiques intéréssants : Etant donné un modèle statistique, on construit une mesure de probabilité conjointe sur les paramètres et les observations, à partir de la distribution d'échantillonnage des observations et d'une distribution a priori sur les paramètres, alors on déduit une distribution prédictive sur les observations et une distribution à posteriori sur les paramètres, cette dernière nous facilite la réponse à tout problème d'inférence statistique sur les paramètres.

Plusieurs chercheurs ont appliqué l'approche bayésienne aux modèles de capture-recapture multiples afin d'estimer la taille d'une population fermée. Des travaux remarquables ont été réalisés par : Smith [28], Castledine [8]. Gasy & Staley,[24] ont constitué une bibliographie exhaustive des approches bayésiennes pour obtenir des inférences sur la taille d'une population fermée.

Dans un premier temps, nous décrivons une expérience de capture-recapture effectuée sur une population fermée de taille inconnue, sous l'hypothèse d'indépendance entre occasions de capture et l'homogénéité des individus, ensuite une analyse bayésienne est présentée dans le but d'estimer la taille de cette population. Terminons ce chapitre par une étude de simulation.

3.2 Inférence bayésienne sur les données de capture-recapture

Soit une population fermée de taille N inconnue. Nous effectuons une expérience de capture-recapture de t $(t \geqslant 2)$ occasions de capture où nous supposons qu'à chaque occasion de capture i tous les individus ont la même probabilité de capture p_i et l'indépendance des occasions de capture. A la complétion de la $i^{ème}$occasion de capture, les données enregistrées sont

· n_i : la taille de l'échantillon.

· m_i : le nombre d'individus marqués dans l'échantillon.

· M_i : le nombre d'individus marqués dans la population juste avant le $i^{ème}$ échantillonnage alors avant le $i^{ème}$ échantillonnage nous avons

$$M_1 = 0,\ M_i = \sum_{j=1}^{i-1} (n_j - m_j), \qquad i = 2, \ldots, t+1.$$

Remarque 3.2.1 $M_i = M_{i-1} + (n_{i-1} - m_{i-1}),\ i = 2, \ldots, t+1$

Les $(n_i - m_i)$ individus non marqués sont à leur tour marqués puis les n_i individus sont remis dans la population.

A l'occasion d'échantillonnage i, les individus de la population ont deux caractères : soit ils sont marqués c'est-à-dire ils ont été capturés au moins une fois pendant les $i-1$ occasions de capture, soit non marqués. Alors nous avons M_i individus marqués dans la population avant le $ième$ tirage et $N - M_i$ individus non marqués. Nous considérons m_i le nombre d' individus marqués dans le $i^{ème}$ échantillon.

m_i est une variable aléatoire qui suit une loi hypergéométrique ayant la densité

$$f(m_i/M_i, N) = \frac{\binom{M_i}{m_i}\binom{N-M_i}{n_i-m_i}}{\binom{N}{n_i}}, m_i = 0, 1, \ldots, \inf\{M_i, n_i\}$$

d'où la fonction de vraisemblance de l'échantillon $\underline{m} = (m_1, m_2, \ldots, m_i)$ est

$$L(m_1, m_2, \ldots, m_i; N, M_i) = \prod_{j=1}^{i} \frac{\binom{M_j}{m_j}\binom{N-M_j}{n_j-m_j}}{\binom{N}{n_j}}, \forall i = 1 \ldots t.$$

A l'occasion de capture i, Pour n_i assez grand et $\frac{M_i}{N}$ petit de telle façon que $n_i\frac{M_i}{N}$ tend vers une constante λ_i , $f\left(m_i/M_i, N\right)$ est approximativement une densité de la loi de Poison de paramètre λ_i

$$f\left(m_i/\lambda_i\right) = \exp\left(-\lambda_i\right)\frac{\lambda_i^{m_i}}{m_i!}, m_i = 0, 1, \ldots$$

avec

$$\lambda_i \ = n_i\frac{M_i}{N}$$

En posant $\omega = \frac{1}{N}$, la fonction de vraisemblance s'écrit

$$L\left(m_1, n_1, \ldots, m_i, n_i; \omega\right) = \prod_{j=1}^{i} f\left(m_j/\omega\right)$$

on obtient alors

$$\begin{aligned} L\left(m_1, n_1, \ldots, m_i, n_i; \omega\right) &= \prod_{j=1}^{i} \exp\left(-\lambda_j\right)\frac{\lambda_j^{m_j}}{m_j!} \\ &= \prod_{j=1}^{i} \exp\left(n_j\frac{M_j}{N}\right)\frac{\left(n_j\frac{M_j}{N}\right)^{m_j}}{m_j!} \\ &= \omega^{\sum_{j=1}^{i} m_j}\frac{\prod_{j=1}^{i}\left(n_jM_j\right)^{m_j}}{\prod_{j=1}^{i} m_j!}\exp\left(\omega\sum_{j=1}^{i} n_jM_j\right) \end{aligned}$$

Alors

$$L\left(m_1, n_1, \ldots, m_i, n_i; \omega\right) = \omega^{S_i}\exp\left(-T_i\omega\right)\frac{\prod_{j=1}^{i}\left(n_jM_j\right)^{m_j}}{\prod_{j=1}^{i} m_j!}$$

avec

$$S_i = \sum_{j=1}^{i} m_j \ \text{ et } \ T_i = \sum_{j=1}^{i} n_jM_j$$

Remarque 3.2.2 *D'après le théoème de factorisation nous avons S_i, T_i sont deux statistiques exhaustives pour ω.*

Calculons la loi à posteriori de ω sachant les statistiques exhaustives T_i, S_i, après la completion du $i^{ème}$occasion de capture.

Puisque la densité de l'échantillon est celle de Poisson alors la loi a priori conjuguée du paramètres $\omega = \dfrac{1}{N}$ est une gamma $\Gamma(a, b)$ d'où

$$f(\omega) = \frac{b^a}{\Gamma(a)}\omega^{a-1}\exp(-b\omega), \omega > 0.a > 0, b > 0. \tag{3.3.1}$$

Calculons la loi a posteriori de ω sachant les statistiques exhaustives S_i et T_i, après la complétion du $i^{ème}$ occasion de capture :

D'après la formule de Bayes, la densité à posteriori de ω est donnée par

$$f(\omega/T_i, S_i) = \frac{L(\underline{n}, \underline{m}; \omega) f(\omega)}{\displaystyle\int_{\Theta} L(\underline{n}, \underline{m}; \omega) f(\omega) d\omega}, \ \Theta =]0, +\infty[\tag{3.3.2}$$

alors en utilisant les formules (3.3.1) et (3.3.2), on obtient

$$f(\omega/T_i, S_i) \ \alpha \ L(\underline{n}, \underline{m}; \omega) f(\omega)$$

$$\begin{aligned} f(\omega/T_i, S_i) \quad \alpha \quad & \frac{b^a}{\Gamma(a)}\omega^{S_i+a-1}\exp[-(T_i+b)\omega] \\ \alpha \quad & \omega^{S_i+a-1}\exp[-(T_i+b)\omega] \end{aligned} \tag{3.3.3}$$

la densité à posteriori de ω est une $\Gamma(S_i + a, T_i + b)$. Le dénominateur de (3.3.2) est la densité prédictive des données ne dépend pas du paramètre ω, noté C_i^{-1}.
Vu que $N \geqslant M_{i+1}$ après l'occasion de capture i alors

$$C_i^{-1} = \int_0^{\frac{1}{M_{i+1}}} L(\underline{n}, \underline{m}; \omega) f(\omega) d\omega$$

Calculons la densité prédictive de $(\underline{n}, \underline{m})$, (densité marginale)

$$
\begin{aligned}
C_i^{-1} &= \int_0^{\frac{1}{M_{i+1}}} \frac{b^a}{\Gamma(a)} \frac{\prod_{j=1}^{i} (n_j M_j)^{m_j}}{\prod_{j=1}^{i} m_j!} \omega^{S_i+a-1} \exp\left[-(T_i+b)\,\omega\right] d\omega \\
&= \frac{b^a}{\Gamma(a)} \frac{\prod_{j=1}^{i} (n_j M_j)^{m_j}}{\prod_{j=1}^{i} m_j!} \int_0^{\frac{1}{M_{i+1}}} \omega^{S_i+a-1} \exp\left[-(T_i+b)\,\omega\right] d\omega
\end{aligned}
$$

on pose $\dot{I} = \int_0^{\frac{1}{M_{i+1}}} \omega^{S_i+a-1} e^{-(T_i+b)\omega} d\omega$.

En utilisans le changement de variable $z = 2(T_i+b)\omega$, l'integrale $\dot{I}$ s'écrit

$$
\begin{aligned}
\dot{I} &= \left(\frac{1}{T_i+b}\right)^{S_i+a-1} \frac{\Gamma(S_i+a)}{2(T_i+b)} \int_0^{\frac{2(T_i+b)}{M_{i+1}}} \frac{1}{\Gamma(S_i+a)} \left(\frac{z}{2}\right)^{\frac{2(S_i+a)}{2}-1} \exp\left(-\frac{z}{2}\right) dz \\
&= \frac{\Gamma(S_i+a)}{2^{S_i+a}(T_i+b)^{S_i+a}} \chi^2_{2(S_i+a)}\left[2(T_i+b)/M_{i+1}\right]
\end{aligned}
$$

où $\chi^2_\delta(.)$ la fonctionde répartition de chi-deux à δ degré de liberté dite loi de *Pearson*

Alors la loi prédictive de $(\underline{n}, \underline{m})$ est donnée par

$$
\begin{aligned}
f(m_1, n_1, \ldots, m_i, n_i) &= \frac{b^a}{\Gamma(a)} \frac{\prod_{j=1}^{i} (n_j M_j)^{m_j}}{\prod_{j=1}^{i} m_j!} \frac{\Gamma(S_i+a)}{(T_i+b)^{S_i+a}} \chi^2_{2(S_i+a)}\left[2(T_i+b)/M_{i+1}\right] \\
, \quad m_j &= 0, 1, \ldots, M_j \text{ et } n_j = 0, 1, \ldots, N.
\end{aligned}
$$

La densité a posteriori de $\omega = \frac{1}{N}$ est celle d'une $\Gamma(S_i+a, T_i+b)$ alors la denstité a posteriori de N est d'une gamma inverse de paramètres $(S_i+a, \frac{1}{T_i+b})$.

3.2.1 Estimation ponctuelle

L'estimateur traditionnel de Schnabel est donné par [27]

$$N_i^{''} = \frac{T_i}{S_i} = \sum_{j=1}^{i} K_j N_j^{'},$$

avec

$$K_j = \frac{m_j}{\sum_{t=1}^{j} m_t}$$

et

$$N_j^{'} = n_j \sum_{t=0}^{j-1} \frac{n_t - m_t}{m_j}$$

$N_j^{'}$ est l'estimateur traditionnel de Peterson de N après l'occasion de capture j.

Fonction de perte quadratique généralisée

Nous considèrons la fonction de perte $\mathfrak{L}\left(\widehat{N}, N\right)$ qui caractérise la perte encourue par l'estimation incertaine de N par un estimateur $\widehat{N}$. L'estimateur de Bayes de N minimise l'espérance a postériori de $\mathfrak{L}\left(\widehat{N}, N\right)$, autrement dit : minimise $E_{\omega/T_i,S_i}\left[\mathfrak{L}\left(\widehat{N}, N\right)\right]$.

La fonction de perte adoptée dans notre analyse est la fonction de perte quadratique généralisée, définie comme suit

$$\mathfrak{L}\left(\widehat{N}, N\right) = \frac{\left(N - \widehat{N}\right)^2}{N^x}\ , x = 0, 1, 2. \tag{3.3.4}$$

$x = 0, 1, 2$ correspond respectivement à la fonction de perte quadratique, du chi-deux et proportionel au carré.

A la fin de la $i^{ème}$occasion de capture, l'estimateur de Bayes $\widehat{N_i}$ défini antérieurement est obtenu en minimisant $E_{N/T_i,S_i}\left[\mathfrak{L}\left(\widehat{N_i}, N\right)\right]$:

$$\begin{aligned} E_{N/T_i,S_i}\left[\mathfrak{L}\left(\widehat{N_i}, N\right)\right] &= \int_0^{\frac{1}{M_{i+1}}} \frac{\left(N - \widehat{N_i}\right)^2}{N^x} f_{\omega/T_i,S_i}(\omega)\, d\omega \\ \frac{\partial E_{NT_i,S_i}\left[\mathfrak{L}\left(\widehat{N_i}, N\right)\right]}{\partial \widehat{N_i}} &= -2 \int_0^{\frac{1}{M_{i+1}}} \frac{1}{N^{x-1}} f_{\omega/T_i,S_i}(\omega)\, d\omega + 2\widehat{N_i} \int_0^{\frac{1}{M_{i+1}}} \frac{1}{N^x} f_{\omega/T_i,S_i}(\omega)\, d\omega \end{aligned}$$

alors

$$\frac{\partial E_{NT_i,S_i}\left[\mathfrak{L}\left(\widehat{N_i},N\right)\right]}{\partial \widehat{N_i}}=0\Leftrightarrow \widehat{N_i}=\frac{E_{N/T_i,S_i}\left(\frac{1}{N^{x-1}}\right)}{E_{N/T_i,S_i}\left(\frac{1}{N^{x}}\right)}$$

d'où

$$E_{N/T_i,S_i}\left(\frac{1}{N^x}\right)=\int_0^{\frac{1}{M_{i+1}}}\frac{1}{N^x}f_{\omega/T_i,S_i}\left(\omega\right)d\omega$$

donc

$$\widehat{N_i}=\frac{T_i+b}{S_i+a+x-1}\frac{\chi^2_{2(S_i+a+x-1)}\left[2\left(T_i+b\right)/M_{i+1}\right]}{\chi^2_{2(S_i+a+x)}\left[2\left(T_i+b\right)/M_{i+1}\right]}$$

avec $\chi^2_\delta\left(.\right)$ la fonction de répartition de chi-deux à δ degré de liberté dite loi de *Pearson.* Le risque associé est l'espérence a posteriori de la fonction de perte $E_{N/T_i,S_i}\left[\mathfrak{L}\left(\widehat{N_i},N\right)\right]$ est donné par

$$\begin{aligned}
E_{N/T_i,S_i}\left[\mathfrak{L}\left(\widehat{N_i},N\right)\right] &= \int_0^{\frac{1}{M_{i+1}}}\frac{\left(N-\widehat{N_i}\right)^2}{N^x}f_{\omega/T_i,S_i}\left(\omega\right)d\omega \\
&= \int_0^{\frac{1}{M_{i+1}}}\frac{1}{N^{x-2}}f_{\omega/T_i,S_i}\left(\omega\right)d\omega\ -2\widehat{N_i}\int_0^{\frac{1}{M_{i+1}}}\frac{1}{N^{x-1}}f_{\omega/T_i,S_i}\left(\omega\right)d\omega \\
&\quad +\widehat{N_i}^2\int_0^{\frac{1}{M_{i+1}}}\frac{1}{N^x}f_{\omega/T_i,S_i}\left(\omega\right)d\omega
\end{aligned}$$

vu que

$$\widehat{N_i}=\frac{E_{N/T_i,S_i}\left(\frac{1}{N^{x-1}}\right)}{E_{N/T_i,S_i}\left(\frac{1}{N^{x}}\right)}$$

alors

$$\begin{aligned}
E_{N/T_i,S_i}\left[\mathfrak{L}\left(\widehat{N_i},N\right)\right] &= E_{N/T_i,S_i}\left(\frac{1}{N^{x-2}}\right)-2\frac{E_{N/T_i,S_i}\left(\frac{1}{N^{x-1}}\right)}{E_{N/T_i,S_i}\left(\frac{1}{N^{x}}\right)}E_{N/T_i,S_i}\left(\frac{1}{N^{x-1}}\right) \\
&\quad +\left(\frac{E_{N/T_i,S_i}\left(\frac{1}{N^{x-1}}\right)}{E_{N/T_i,S_i}\left(\frac{1}{N^{x}}\right)}\right)^2E_{N/T_i,S_i}\left(\frac{1}{N^{x}}\right)
\end{aligned}$$

finalement

$$E_{N/T_i,S_i}\left[\mathfrak{L}\left(N,\widehat{N_i}\right)\right]=E_{N/T_i,S_i}\left(\frac{1}{N^{x-2}}\right)-\frac{E^2_{N/T_i,S_i}\left(\frac{1}{N^{x-1}}\right)}{E_{N/T_i,S_i}\left(\frac{1}{N^{x}}\right)}$$

avec

$$E_{N/T_i,S_i}\left(\frac{1}{N^{\upsilon}}\right)=\frac{\Gamma\left(a+S_i+\upsilon\right)}{\left(b+T_i\right)^{\upsilon}\Gamma\left(a+S_i\right)}\frac{\chi^2_{2(S_i+a+\upsilon)}\left[2\left(T_i+b\right)/M_{i+1}\right]}{\chi^2_{2(S_i+a)}\left[2\left(T_i+b\right)/M_{i+1}\right]}$$

3.2.2 Estimation ensembliste

En utilisant l'approche fréquentiste, on construit un intervalle de confiance pour N en se basant sur la propriété de normalité asymptotique des estimateurs.
Si $\widehat{N}$ est un estimateur de N et $\widehat{v}$ est un estimateur de sa variance alors un intervalle de niveau de confiance $1-\alpha$ $(0<\alpha<1)$ est tel que

$$p\left(N\in\left(0,\widehat{N}+z_\alpha\sqrt{\widehat{v}}\right)\right)=1-\alpha$$

où z_α est le $\alpha-$quantile de la loi normale centrée et réduite.

Soit $p_{N/S_i,T_i}$ est la densité à postériori de N. Un intervalle $(1-\alpha)\,100\%$ crédible est un sous ensemble C de Θ $(\Theta=]0,\infty[)$ tel que

$$p_{N/S_i,T_i}\left(C\right)=1-\alpha$$

3.3 Etude de simulation

Modèle de capture-recaptue M_0

Pour vérifier les résultats théoriques calculés, nous proposons une étude de simulation de capture-recapture (sous les hypothèses du modèle M_0).

Le modèle de capture-recapture homogène M_0 suppose que les occasions de capture sont indépendantes et tous les individus ont la même probabilité de capture p $(p\in]0.1[)$.

$1^{ère}$ partie : Simulation

On suppose que notre population a une taille constante N $(N\in\mathbb{N}^*)$ à estimer. Le nombre d'occasions de capture est $t,(t=2,3,\ldots)$, n_i $(i=1,2,\ldots,t)$ est la taille du *ième* échantillon tiré. Le vecteur $(n_1,\ n_2,\ \ldots,\ n_t)$ est choisi aléatoirement. En utilisant la commande *"randsample"* du logiciel Matlab, nous aurons n_i $(i=1,2,\ldots,t)$ nombres aléatoires qui

représentent les individus tirés de la population au *ième* tirage. À chaque occasion de capture i, on compare les nombres aléatoires tirés avec ceux tirés aux occasions de capture précédentes (les $(i-1)$ occasions de capture), et on compte le nombre de répétition m_i, elles représentent le nombre des individus marqués dans le *ième* échantillon. Les données disponibles sont : n_i, m_i, M_i (le nombre cumulé de répétitions just avavt le *ième* tirage).

2ème partie : Estimation

On suppose qu'aucune information est disponible sur le paramètre N, pour cela on choisi les paramètres de la loi a priori gamma comme suit : $a = b = 0$, ceci correspond a l'utilisation de la loi a priori non informative de Jeffres pour N.

L'estimateur de Bayes de la taille de la population est calculé à chaque occasion de capture i, noté $\widehat{N_i}$ $(i = 1, 2, \ldots, t)$, ainsi que son risque R_i associés à la fonction de perte quadratique $(x = 0)$, chi-deux $(x = 1)$ et proportionnelle au carré $(x = 2)$.

• Les entrées de simulation contiennent les valeurs suivantes : $[N,\ t,\ a,\ b,\ x,\ (n_1,\ \ldots,\ n_t)]$ L'expérience sera itérée 1000 fois pour les mêmes paramètres.

• Les sorties sont les valeurs moyennes estimées de paramètre N_i (estimateur de Bayes) à chaque occasion de capture i et la moyenne de son risque associé R_i notés respectivement $\widehat{N}_{imoy}$, R_{imoy}.

Dans ce qui suit nous présentons quelques exemples où les résultats sont répartis dans les tableaux qui suivent. L'occasion de capture i, le nombre d'individus capturés à cette occasion de capture n_i, le nombre d'individus marqués m_i ainsi que apparaissent dans les premières colonnes de ces tableaux.

Exemple 3.3.1 $N = 1500$, $t = 10$, $n = (75\ 82\ 65\ 91\ 101\ 87\ 73\ 112\ 60\ 99)$

Tableau 3.1 : Résultats de simulation du modèle M_0 pour $N = 1500$, $t = 6$

i	n_i	m_i	$\widehat{N}_{imoy}\ (x=0)$	R_{imoy}	$\widehat{N}_{imoy}\ (x=1)$	R_{imoy}	$\widehat{N}_{imoy}\ (x=2)$	R_{imoy}
1	75	0	-	-	-	-	-	-
2	82	4	-	-	-	-	-	-
3	65	7	1821.1	859880	1669.5	227.6346	1480.4	0.0929
4	91	13	1602.7	142080	1564.1	75.5922	1497.8	0.0423
5	101	19	1552.2	63200	1531.6	37.9003	1496.5	0.0231
6	87	21	1534.7	39390	1517.3	24.4106	1494.0	0.0154
7	73	21	1526.1	28670	1512.9	18.1321	1494.3	0.0116
8	112	36	1517.7	19550	1510.8	12.6126	1495.1	0.0081
9	60	22	1516.1	16380	1506.5	-	1495.0	-
10	99	40	1512.0	-	1503.2	-	1495.1	-

Exemple 3.3.2 $N = 1500$, $t = 8$, $n = (82\ 75\ 101\ 96\ 61\ 122\ 75\ 92)$,

Tableau 3.2 : Résultats de simulation du modèle M_0 pour $N = 1500$, $t = 8$

i	n_i	m_i	$\widehat{N}_{imoy}\ (x=0)$	R_{imoy}	$\widehat{N}_{imoy}\ (x=1)$	R_{imoy}	$\widehat{N}_{imoy}\ (x=2)$	R_{imoy}
1	82	0	-	-	-	-	-	-
2	75	4	-	-	-	-	-	-
3	101	10	1704.7	332700	1580.8	136.5009	1479.9	0.0685
4	96	16	1572.2	98300	1527.8	55.4037	1487.5	0.0330
5	61	13	1554.8	63190	1517.9	37.1394	1488.5	0.0230
6	122	30	1534.1	33410	1513.3	21.2942	1493.4	0.0136
7	75	23	1528.4	25500	1510.2	16.0110	1494.4	0.0103
8	92	32	1518.9	18670	1505.1	11.9278	1494.8	0.0078

Exemple 3.3.3 $N = 2000$, $t = 10$, $n = (92\ 85\ 101\ 96\ 71\ 122\ 75\ 92\ 101\ 87)$

Tableau 3.3 : Résultats de simulation du modèle M_0 pour $N = 2000$, $t = 9$

i	n_i	m_i	$\widehat{N}_{imoy}\ (x=0)$	R_{imoy}	$\widehat{N}_{imoy}\ (x=1)$	R_{imoy}	$\widehat{N}_{imoy}\ (x=2)$	R_{imoy}
1	92	0	-	-	-	-	-	-
2	85	4	-	-	-	-	-	-
3	101	9	2349.1	1077200	2179.1	240.9113	2409.9	-
4	96	13	2164.1	243500	2071.0	91.8357	2156.6	2.3828
5	71	12	2102.8	137600	2046.6	58.3831	2106.4	1.3771
6	122	25	2054.8	73600	2023.4	33.6432	2057.5	0.7358
7	75	19	2047.4	54900	2016.2	25.4577	2043.4	0.5447
8	92	26	2032.7	40400	2015.0	19.2113	2033.1	0.4026
9	101	31	2027.4	30700	2012.3	14.7397	2025.6	0.3055
10	87	30	2020.6	-	2007.4	-	2019.8	-

Exemple 3.3.4 $N = 2000$, $t = 13$, $n = (73\ 87\ 96\ 31\ 29\ 75\ 102\ 67\ 198\ 92\ 78\ 63\ 109\ 56\ 52)$

Tableau 3.4 : Résultats de simulation du modèle M_0 pour $N = 2000$, $t = 13$

i	n_i	m_i	$\widehat{N}_{imoy}\ (x=0)$	R_{imoy}	$\widehat{N}_{imoy}\ (x=1)$	R_{imoy}	$\widehat{N}_{imoy}\ (x=2)$	R_{imoy}
1	72	0	-	-	-	-	-	-
2	85	3	1827.6	-	-	-	-	-
3	77	6	1978.1	10990	2276.5	434.0530	-	-
4	96	11	1988.4	5020	2105.5	126.1548	2242.5	3.8078
5	82	13	1975.2	3040	2055.0	69.2623	2128.7	1.7038
6	122	23	1988.8	1790	2013.0	37.6097	2066.6	0.8478
7	75	18	1996.8	1360	2014.5	28.2506	2050.4	0.0615
8	92	25	1999.7	1020	2008.5	20.8903	20332.2	0.4415
9	101	31	1995.7	780	2007.2	15.8571	2049.8	0.3280
10	87	33	1996.7	-	2005.4	-	2015.8	-
11	75	38	2000.1	-	2007.1	-	2012.5	-
12	82	33	2000.1	-	2005.5	-	2009.0	-
13	99	42	2000.5	-	2005.1	-	2008.6	-

Représentation graphique

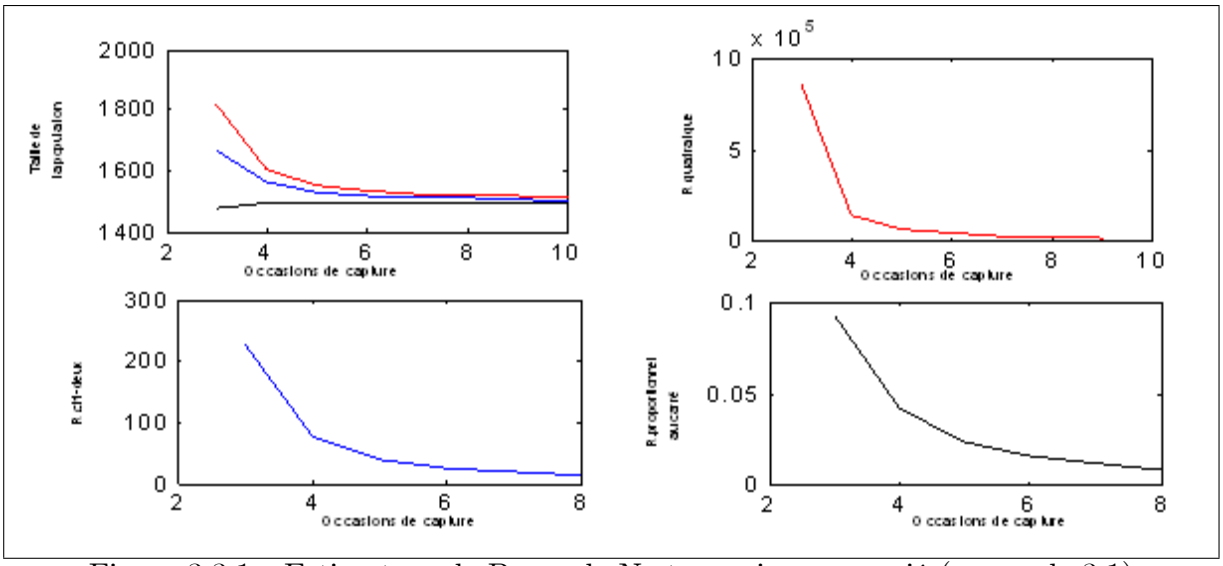

Figure 3.3.1 : Estimateur de Bayes de N et son risque associé (exemple 3.1)

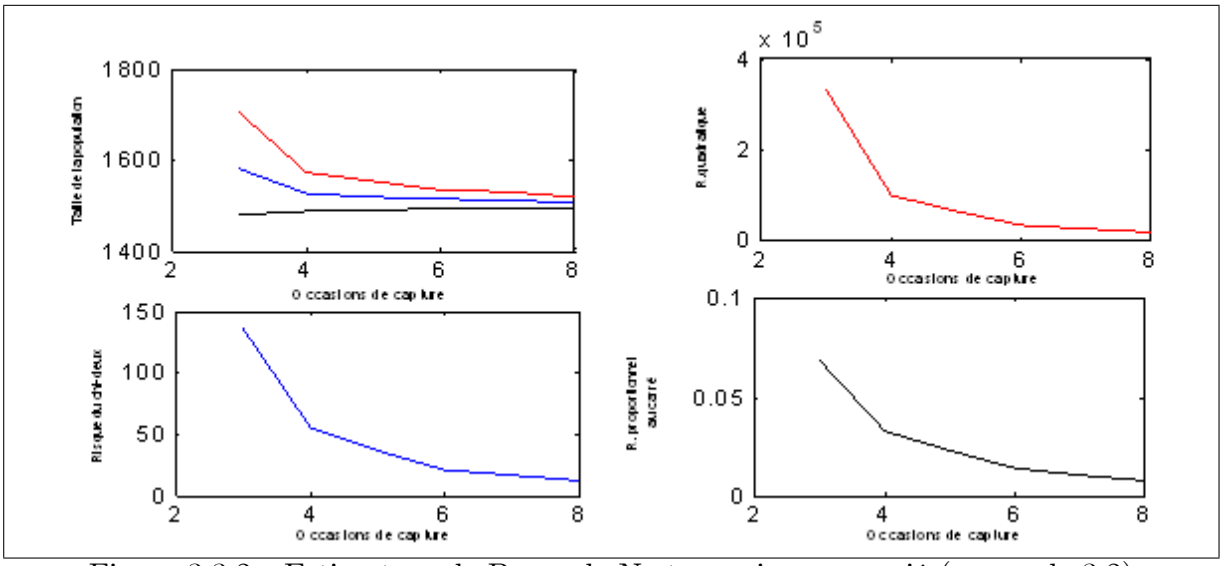

Figure 3.3.2 : Estimateur de Bayes de N et son risque associé (exemple 3.2)

.

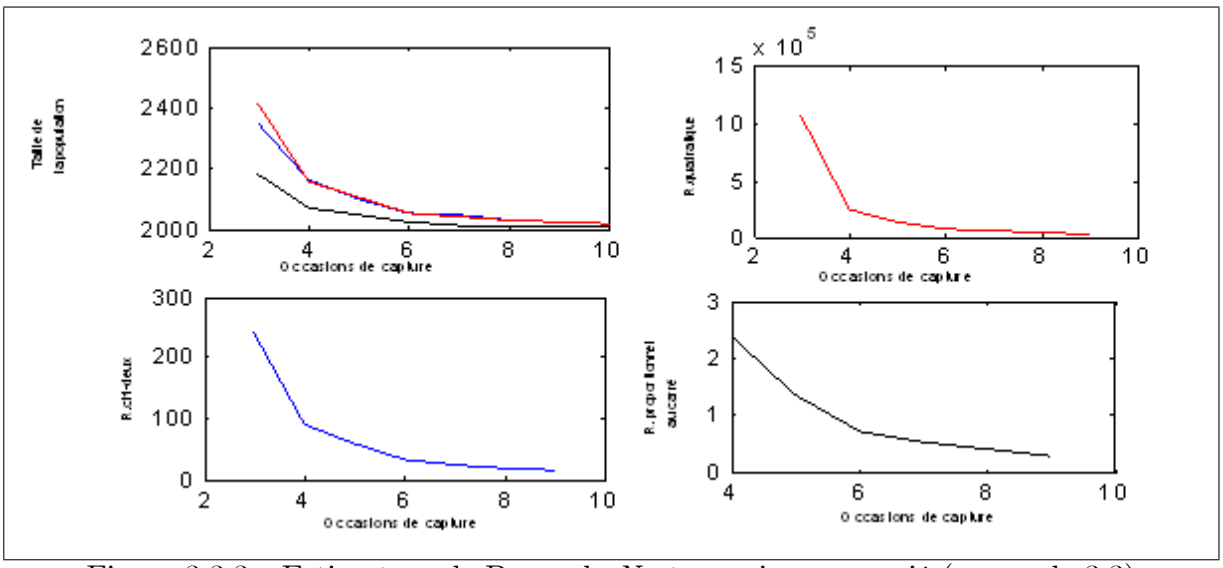

Figure 3.3.3 : Estimateur de Bayes de N et son risque associé (exemple 3.3)

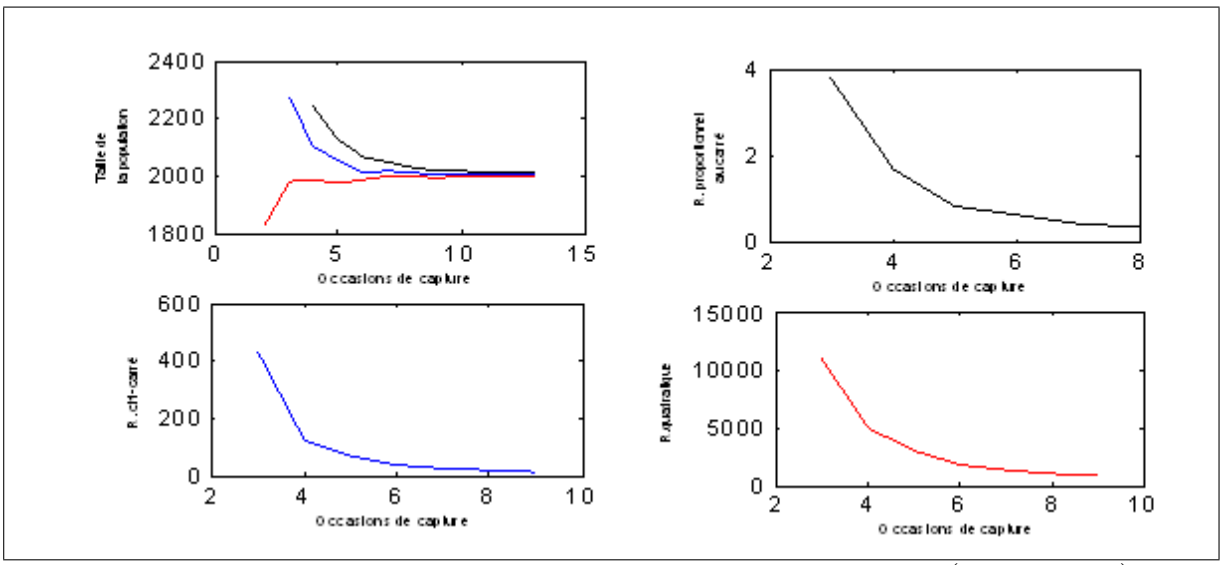

Figure 3.3.4 : Estimateur de Bayes de N et son risque associé (exemple 3.4)

Modèle de capture-recapture M_t

Le modèle de capture-recapture M_t est un modèle homogène, il suppose que les occasions de capture sont indépendantes et à chaque occasion de capture i, tous les individus ont la même probabilité de capture p_i $(i = 1, 2, \ldots, t)$.

· n_i $(i = 1, 2, \ldots, t)$ le nombre d'individus capturés à l'occasion de capture i, est une variable aléatoire suit une loi binomiale de paramètre N, p_i où p_i est la probabilité de capture d'un individu à l'occasion de capture i.
En utilisant la commande "*binornd*" du logiciel Matlab, on peut générer des valeurs de la variable aléatoire n_i $(i = 1, 2, \ldots, t)$.

· Les entrées de simulation sont : $(N,\ t,\ a,\ b,\ x,\ P)$ tel que $P = (p_1, \ldots, p_t)$.

· Les sorties sont les valeurs moyennes estimées de paramètre N_i (estimateur de Bayes) à chaque occasion de capture i et la moyenne de son risque associé R_i notés respectivement $\widehat{N}_{imoy}$, R_{imoy}.

Dans ce qui suit nous présentons quelques exemples où nous calculons l'estimateur de Bayes de N pour les trois fonctions de perte (quadratique, chi-carré et proportionnelle au carré):

Exemple 3.3.5 $N = 1500$, $t = 8$, $P = (0.12\ 0.13\ 0.11\ 0.128\ 0.14\ 0.18\ 0.19\ 0.2)$, $a = b = 0$,

Tableau 3.5 : Résultats de simulation du modèle M_t pour $N = 1500$, $t = 8$

i	p_i	m_i	$\widehat{N}_{imoy}$ $(x = 0)$	R_{imoy}	$\widehat{N}_{imoy}$ $(x = 1)$	R_{imoy}	$\widehat{N}_{imoy}$ $(x = 2)$	R_{imoy}
1	0.12	0	-	-	-	-	-	-
2	0.13	23	1618.4	146900	1561.3	76.8069	1490.3	0.0430
3	0.11	39	1537.2	41480	1519.5	25.6748	1499.7	0.0162
4	0.128	61	1520.4	19570	1508.8	12.5257	1495.9	0.0081
5	0.14	85	1510.1	15120	1502.0	9.0123	1494.0	0.0060
6	0.18	131	1506.5	9652	1500.7	-	1497.3	-
7	0.19	166	1504.5	-	1500.7	-	1498.0	-
8	0.2	199	1503.5	-	1500.4	-	1498.1	-

Exemple 3.3.6 $N = 1500$, $t = 10, P = (0.02\ 0.01\ 0.05\ 0.03\ 0.02\ 0.04\ 0.012)$, $a = b = 0$,

Tableau 3.6 : Résultats de simulation du modèle M_t pour $N = 1500$, $t = 10$

i	p_i	m_i	$\widehat{N}_{imoy}\ (x=0)$	R_{imoy}	$\widehat{N}_{imoy}\ (x=1)$	R_{imoy}	$\widehat{N}_{imoy}\ (x=2)$	R_{imoy}
1	0.020	0	-	-	-	-	-	-
2	0.030	5	-	-	-	-	1422.4	-
3	0.010	24	1583.9	104170	1534.1	57.0754	1488.8	0.0338
4	0.028	10	1558.3	70450	1528.3	41.5134	1494.2	0.0253
5	0.040	55	1523.8	25910	1510.9	16.4794	1493.1	0.0106
6	0.080	44	1514.0	12450	1509.1	6.0231	1493.1	0.0035
7	0.090	57	1509.0	-	1504.8	-	1494.7	-
8	0.020	14	1508.0	-	1504.3	-	1495.3	-
9	0.019	151	1502.7	-	1502.1	-	1496.7	-
10	0.0123	17	1502.6	-	1502.6	-	1496.6	-

Représentation graphique

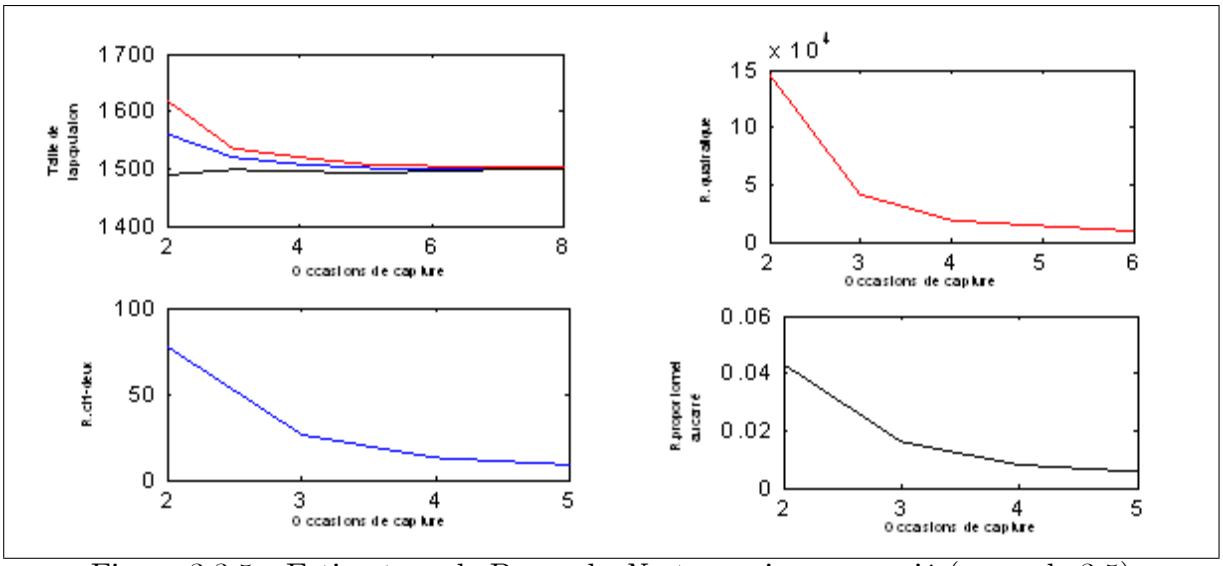

Figure 3.3.5 : Estimateur de Bayes de N et son risque associé (exemple 3.5)

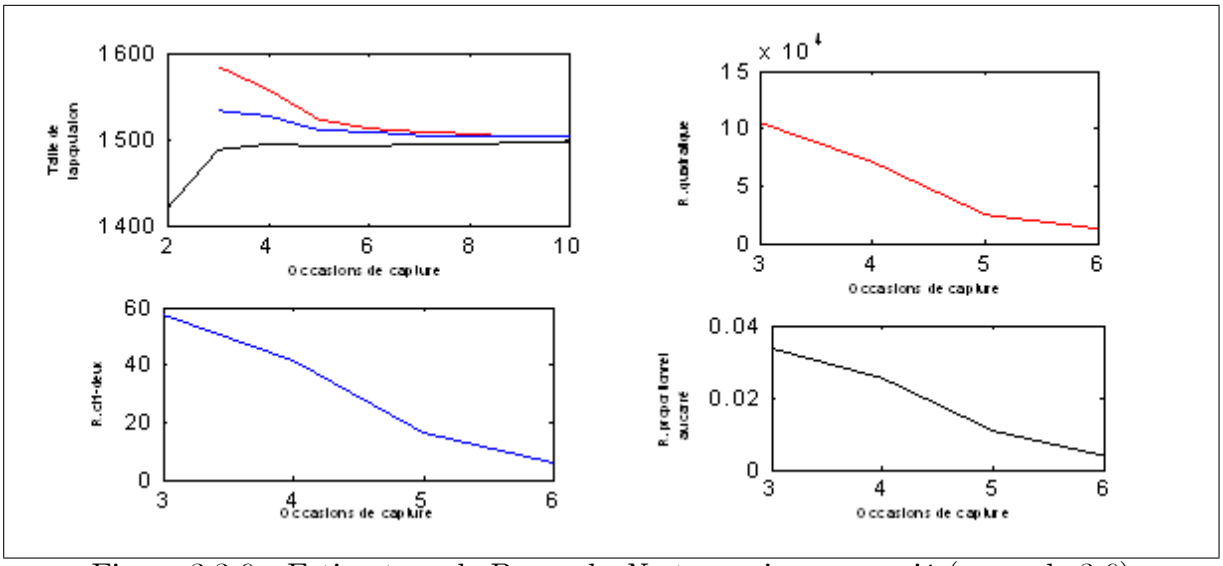

Figure 3.3.6 : Estimateur de Bayes de N et son risque associé (exemple 3.6)

variation des paramètres de la loi à priori

Exemple 3.3.7 $N = 2000$, $t = 10$, $n = (72\ 85\ 77\ 96\ 82\ 122\ 75\ 92\ 101\ 87)$

Tableau 3.7 : Résultats de simulation du modèle M_0 pour $N = 1500$, $t = 10$, $a = 0.027$ et $b = 0.108$

i	n_i	m_i	$\widehat{N}_{imoy}\ (x=0)$	R_{imoy}	$\widehat{N}_{imoy}\ (x=1)$	R_{imoy}	$\widehat{N}_{imoy}\ (x=2)$	R_{imoy}
1	72	0	-	-	-	-	142.1	0.9481
2	85	5	-	-	2741	-	1917.3	0.3133
3	77	24	3445.3	-	2245	412.9748	1987.3	0.1104
4	96	10	2224.7	376200	2076	121.4692	1999.2	0.0504
5	82	55	2119.7	170190	2028	67.2609	1994.2	0.0306
6	122	44	2061.7	84900	2026	38.1667	1995.8	0.0179
7	75	57	2041.8	60930	2019	28.3728	1997.5	0.0136
8	92	14	2030.0	44270	2016	21.3728	1994.2	0.0101
9	101	151	2021.4	3300	2015	21.0567	1992.8	0.0077
10	87	17	2016.6	-	2011	-	1993.0	-

Exemple 3.3.8 $N = 2000$, $t = 10$, $n = (72\ 85\ 77\ 96\ 82\ 122\ 75\ 92\ 101\ 87)$,

Tableau 3.8 : Résultats de simulation du modèle M_0 pour $N = 1500$, $t = 10$, $a = 2,7$ et $b = 108$

i	n_i	m_i	$\widehat{N}_{imoy}\ (x=0)$	R_{imoy}	$\widehat{N}_{imoy}\ (x=1)$	R_{imoy}	$\widehat{N}_{imoy}\ (x=2)$	R_{imoy}
1	72	0	136	15160	109.2	26.8106	97.7	0.1055
2	85	4	1188	643330	1000.5	223.7543	856.9	0.1376
3	77	8	1392.7	202100	1306.0	109.8342	1195.2	0.0663
4	96	14	1443.5	86480	1405.9	53.5846	1335.9	0.0339
5	82	17	1468.6	53040	1436.6	33.3910	1393.9	0.0217
6	122	30	1488.6	31460	1459.7	19.9460	1431.7	0.0131
7	75	23	1488.0	23550	1467.4	15.2185	1449.5	0.0101
8	92	31	1490.3	17680	1474.9	11.5250	1459.9	0.0076
9	101	39	1491.7	-	1478.8	-	1469.4	-
10	87	37	1493.7	-	1481.3	-	1475	-

Représentation graphique

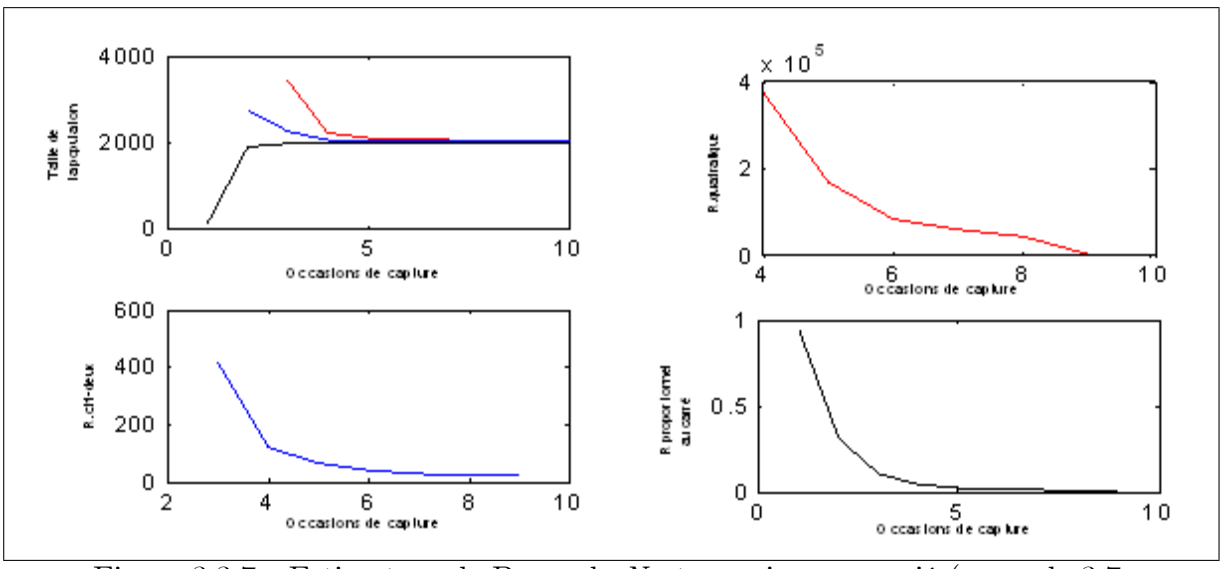

Figure 3.3.7 : Estimateur de Bayes de N et son risque associé (exemple 3.7, $a = 0.027, b = 0.108$)

2

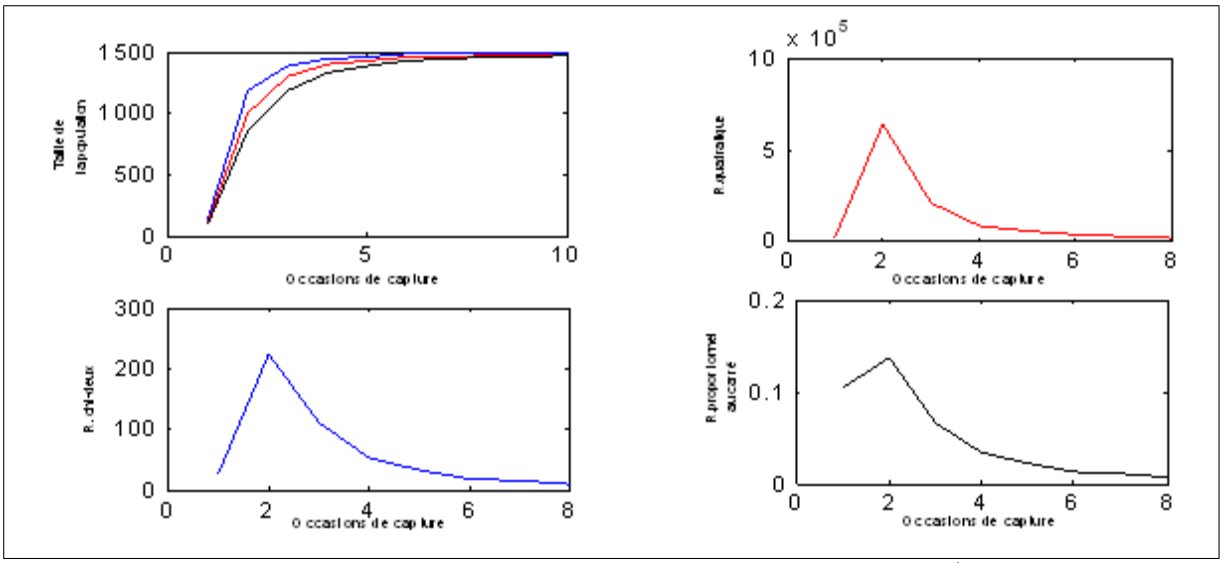

Figure 3.3.8 : Estimateur de Bayes de N et son risque associé (exemple 3.8, $a = 2.7, b = 108$)

3.3.1 Discussion

1.$a = 0, b = 0$

Modèle M_0

· Dans l'exemple (3.1), nous avons choisi les paramètres de simulation comme suit : $N = 1500$, $t = 10$ et le nombre d'individus capturés initialement est de 75.

La valeur moyenne estimée de la taille de la population se stabilise à la vraie valeur ($N = 1500$) à partir de la sixième occasion de capture, pour l'estimateur de Bayes associé à la fonction de perte quadratique, du chi-deux et proportionnelle au carré respectivement. Le risque moyen associé à cette valeur diminu en multipliant les occasions de capture (pour les trois fonctions de perte).

· Dans l'exemple (3.2) les paramètres choisis sont : $N = 1500$, $t = 8$ et le nombre d'individus initialement capturés est de 82.

La valeur estimée moyenne $\widehat{N}_{imoy}$ $(i = 1, \ldots, 8)$ tend vers la vraie valeur quand le nombre d'occasions de capture augmente, elle l'atteint à partir de $i = 6$ (pour la fonction de perte quadratique) et à partir de $i = 7$ (pour les fonctions de perte chi-carré et proportionnelle au carré). Le risque moyen R_{imoy} associé à $\widehat{N}_{imoy}$ est décroissant en fonction du nombre d'occasions de capture dans les trois cas.

· Nous avons choisi $N = 2000$, $t = 10$ et le nombre d'individus capturés initialement est de 92.

La valeur moyenne estimée de la taille de la population se stabilise à la vraie valeur $(N = 2000)$ à partir de $(i = 7)$, pour l'estimateur de Bayes associé à la fonction de perte quadratique, et $(i = 8)$ pour celui associé à la fonction de perte du chi-deux et proportionnelle au carré. Le risque moyen est décroissant pour les trois cas.

· Le dernier exemple du modèle de capture-recapture M_0 suppose que la vraie valeur de N est 2000, $t = 13$ et le nombre d'individus initialement capturés est 72.

La valeur moyenne estimée de la taille de la population se stabilise à la vraie valeur $(N = 2000)$ à partir de $(i = 11)$ pour la fonction de perte quadratique à partir de $(i = 13)$ pour les deux autres cas. Le risque moyen R_{imoy} associé à $\widehat{N}_{imoy}$ est décroissant en fonction du nombre d'occasions de capture dans les trois cas.

Une question importante qui se pose : Après quelle occasion de capture on doit cesser d'échantillonner ?

Le rique associé R_{imoy} est représenté graphiquement en fonction des occasions de capture. Prenons l'exemple 3.1, remarquons que le risque quadratique est décroissant, d'un taux de 5% entre la troisième et la quatrième occasion de capture et de 1% entre la cinquième et la sixième occasion de capture. Pour une valeur spécifiée de α, $(0 < \alpha < 1)$ si le risque associé d'un estimateur diminu d'une occasion à celle qui suive d'un pourcentage moins que α alors l'échantillonnage doit s'arrêter.

Modèle M_t

· Les entrées de simulation dans l'exemple (3.5) contiennent les valeurs suivantes : $N = 1500$, $t = 8$, $P = (0.12\ 0.13\ 0.11\ 0.128\ 0.14\ 0.18\ 0.19\ 0.2)$.

La valeur estimée moyenne $\widehat{N}_{imoy}$ $(i = 1, \ldots, 8)$ tend vers la vraie valeur quand le nombre d'occasions de capture augmente, elle l'atteint à partir de $i = 6$ (pour la fonction de perte quadratique) et à partir de $i = 7$ (pour les fonctions de perte chi-carré et proportionnelle au carré). Le risque moyen R_{imoy} associé à $\widehat{N}_{imoy}$ est décroissant en fonction du nombre d'occasions de capture dans les trois cas.

· Les entrées de simulation dans l'exemple (3.5) contiennent les valeurs suivantes : $N = 1500$, $t = 10, P = (0.02\ 0.01\ 0.05\ 0.03\ 0.02\ 0.04\ 0.012)$.

La valeur moyenne estimée de la taille de la population se stabilise à la vraie valeur ($N = 1500$) à partir de ($i = 6$) pour la fonction de perte proportionnelle au carré et à partir de ($i = 8$) pour les deux autres cas. Le risque moyen R_{imoy} associé à $\widehat{N}_{imoy}$ est décroissant en fonction du nombre d'occasions de capture dans les trois cas.

2. Variation des paramètres de la loi à priori

· Dans les exemples (3.1) , (3.7) et (3.8) , les paramètres de la loi à priori sont choisis comme suit : $a = b = 0$, $a = 0.027$ et $b = 0.108$, $a = 2.7$ et $b = 108$. Nous remarquons que la tendance de la valeur estimée moyenne $\widehat{N}_{imoy}$ dans le cas informatif vers la vraie valeur de la taille de la population est plus rapide que dans le cas noninformatif

Chapitre 4

Modélisation log-linéaire des données de capture-recapture

4.1 Introduction

Les modèles de capture-recapture hétérogènes considèrent les trois sources de variations des probabilités de capture : effet temporel, le comportement des individus après leur capture initiale, la différence entre les individus. Fienberg [16], Cormack [12] ont introduit l'approche basée sur les modèles log-linéaires afin de traiter cette hétérogénéité. Des travaux remarquables ont été réalisés dans ce contexte : Rivest et Levesque [24], Rivest et Daigle [25].

Les modèles log-linéaires appartiennent à la classe des modèles linéaires généralisés qui a été introduite par Nelder et Wedderburn en (1972) , puis approfondie par McCullagh et Nelder en (1989) [1].

L'objectif à atteindre en préparant ce chapitre est la modélisation log-linéaire des données de capture-recapture. Pour cela nous présentons une théorie générale sur les modèles linéaires généralisés. La forme log-linéaire est déduite pour chacun des modèles de capture-recapture suivants : M_0, M_t, M_{th}, M_b d'où un estimateur log-linéaire de la taille de la population est obtenu, terminons par le calcul de son biais et sa variance.

4.2 Les modèles linéaires généralisés

La classe des modèles linéaires généralisés notés LG, est une généralisation des modèles de régression linéaires en termes de loi de probabilité d'une part mais aussi en termes de lien à la linéarité. Leur objectif est d'étudier la relation entre une variable réponse et d'autre variables explicatives.

Soit Y une variable aléatoire, sa densité appartient à la famille exponentielle, $y = (y_1,\ y_2, \ldots,\ y_n)^{'}$ un vecteur $n \times 1$ d'observations indépendantes de Y tel que

$$\begin{cases} y_i = \mu_i + \varepsilon_i \\ E(Y_i) = \mu_i \qquad\qquad\qquad\qquad , \ i = 1, \ \ldots, \ n \\ var(Y_i) = var(\varepsilon_i) = \sigma_i^2 \end{cases}$$

ε un vecteur d'erreurs non observé d'espérance nulle et ε_i, ε_j sont non corrélées pour $i \neq j$. Un modèle linéaire généralisé pour Y est

$$\eta = g(\mu) = X\beta$$

où g est une fonction monotone dérivable, X est une matrice $(n \times d)$, d'éléments déterministes (réalisations des variables explicatives $X_1, \ldots, X_d$), β est un vecteur de d paramètres fixes à estimer.
Dans ce qui suit, nous présentons une théorie générale sur ces modèles en plus de détails.

4.2.1 Composantes du modèle linéaire-généralisé

Les modèles catalogués dans la classe des modèles linéaires généralisés sont caractérisés par trois composantes [1].

1. Composante aléatoire

La composante aléatoire identifie la distribution de probabilités de la variable à expliquer Y. On suppose que $(y_1, \ldots, y_n)$ n observations indépendantes de la variable aléatoire Y. Sa densité doit avoir une structure exponentielle.

Définition 4.2.1

appelle famille de loi exponentielle sur $\mathbb{R}^n$*, l'ensemble des lois de probabilités admettant une densité de la forme*

$$f(y_i, \theta_i, \phi) = \exp\left\{\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi)\right\} \tag{4.2.1}$$

.

La famille exponentielle inclut la plupart des lois usuelles comportant un ou deux paramètres : gaussienne, gamma, poisson, binomiale,....

Le paramètre θ_i est appelé paramètre *naturel* de la famille exponentielle, il varie de $i = 1, \ldots, n$ dépond des variables explicatives, ϕ est appelé paramètre de *dispersion.*

Quand ϕ est connu, la formule (4.2.1) est simplifiée sous la forme

$$f(y_i, \theta_i) = A(\theta_i) B(y_i) \exp[y_i Q(\theta_i)], \tag{4.2.2}$$

avec $A(\theta_i) = \exp[-b(\theta_i)/a(\phi)]$ et $Q(\theta_i) = \theta_i/a(\phi)$, $B(y_i) = \exp[c(y_i, \phi)]$.

2. Composante systématique (détérministe)

Les observations planifiées des variables explicatives sont organisées dans la matrice X dite *matrice du modèle (design matrix).*

La composante systématique d'un modèle LG relie les paramètres β_j $(j = 1, \ldots, d)$ aux variables explicatives avec un prédicteur linéaire

$$\eta_i = \sum_{j=1}^{d} \beta_j x_{ij} \quad , i = 1, \ldots, n.$$

où $x_{i1}, x_{i2}, \ldots, x_{id}$ représentent les valeurs de d variables explicatives pour la $i^{ème}$ observation.

La forme matricielle est donnée par

$$\eta = X\underline{\beta}$$

où $\underline{\beta} = (\beta_1, \ldots, \beta_d)'$, et $\eta = (\eta_1, \eta_2, \ldots, \eta_n)'$

3. Une fonction lien

La fonction lien exprime une relation fonctionnelle entre la composante aléatoire et la composante systématique.

Soit $\{\mu_i = E(Y_i), i = 1, \ldots, n\}$. Le modèle relie μ_i à η_i par

$$\eta_i = g(\mu_i), i = 1, \ldots, n \quad ,$$

g appelée *fonction lien*, est une fonction monotone et dérivable.

Remarque 4.2.1 *: g relie $E(Y_i)$ aux variables explicatives comme étant*

$$g(\mu_i) = \sum_{j=1}^{d} \beta_j x_{ij} \ , \ i = 1, \ldots, n.$$

▷ *La fonction lien $g(\mu) = \mu$, appelée fonction lien identité, $\eta_i = \mu_i$, elle spécifie le modèle linéaire de la moyenne elle même. C'est la fonction lien pour un modèle de regréssion avec une variable réponse Y de distribution normale.*

▷ *La fonction lien qui transforme la moyenne en paramètre naturel est dite fonction lien canonique :*

$$g(\mu_i) = Q(\theta_i) \ \ or\ Q(\theta_i) = \sum_{j=1}^{d} \beta_j x_{ij}$$

Dans le cas d'un échantillon gaussien, les densités d'une famille de lois $N(\mu; \sigma^2)$ s'écrivent

$$\begin{aligned} f(y, \mu) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(y-\mu)^2}{2\sigma^2}\right\} \qquad ,\sigma^2 \text{ connue.} \\ &= \exp\left\{-\frac{\mu^2}{2\sigma^2}\right\} \exp\left\{-\frac{y^2}{2\sigma^2} - \frac{1}{2}\ln(2\pi\sigma^2)\right\} \exp\left\{y\frac{\mu}{\sigma^2}\right\} \end{aligned}$$

on posant

$$\begin{aligned} Q(\theta) &= \frac{\theta}{\phi} = \frac{\mu}{\sigma^2} \\ A(\theta) &= \exp\left\{-\frac{\mu^2}{2\sigma^2}\right\} \\ B(y) &= \exp\left\{-\frac{y^2}{2\sigma^2} - \frac{1}{2}\ln(2\pi\sigma^2)\right\} \end{aligned}$$

La famille gaussienne se met sous la forme canonique qui n'est en fait qu'une famille exponentielle de paramètre de dispersion $\phi = \sigma^2$ et de paramètre naturel: $\theta = E(Y) = \mu$, donc sa fonction lien est la fonction identité.

Exemple 4.2.1 *: Le modèle logistique des données binaires*

Les variables réponses qualitatives qui représentes (succé, échec) sont des variables binaires à deux modalités (1, 0). On suppose que : $p\,(Y=1)=\pi$ *et* $p(Y=0)=1-\pi$ *alors :* $E(Y)=\pi$. *La variable aléatoire* Y *a une distribution de Bernouli, sa densité est donnée par*

$$\begin{aligned} f\,(y,\pi) &= \pi^{y}\,(1-\pi)^{1-y} \\ &= (1-\pi)\left(\frac{\pi}{1-\pi}\right)^{y} \\ &= (1-\pi)\exp\left(y\log\frac{\pi}{1-\pi}\right) \end{aligned}$$

pour $y=0$ *et* 1. *Cette densité appartient à la famille exponentielle, où* $A(\pi)=1-\pi$ *,* $B(y)=1$, $Q(\pi)=\log\frac{\pi}{1-\pi}$. *Le paramétre naturel est* $\log\frac{\pi}{1-\pi}$, *logit* π. *c'est le lien canonique, le modèle LG qui utilise le lien logit est dit modèle logistique.*

Tableau 3.1: Différents types de modèles linéaires généralisés.

composante aléatoire	*lien*	*composante déterministe*	*modèle*
normale	identité	continue	régression
normale	identité	catégorielle	analyse de la variance
normale	identité	mixte	analyse de covariance
binomiale	logit	mixte	régression logistique
poisson	log	mixte	log-linéaire

4.2.2 Maximum de vraisemblance pour un modèle linéaire généralisé

Dans ce qui suit nous calculons l'estimateur du maximum de vraisemblance pour β; le vecteur de paramètres du modèle linéaire généralisé. Cet estimateur existe et il est unique (Wedderburn 1996), [1].

Généralement, les équations de vraisemblance d'un modèle linéaire généralisé ne sont pas linéaires en β, nous citerons des méthodes itératives (méthode de Newton-Raphson, fisher scoring) pour les résoudre.

Moments et fonction de vraisemblance du modèle LG

Soit Y la composante aléatoire d'un modèle linéaire génerlisé, alors pour n observations indépendantes $y_1, \ldots, y_n$ de celle ci , nous notons

$$L_i = \log f\left(y_i, \theta_i, \phi\right),$$

donc la fonction log-vraisemblance du modèle est donnée par:

$$L = \sum_{i=1}^{n} L_i \quad ,$$

d'après la formule (4.2.1) nous avons

$$L_i = \left[y_i\theta_i - b\left(\theta_i\right)\right]/a\left(\phi\right) + c\left(y_i, \phi\right), \tag{4.2.3}$$

Proposition 4.2.1 $E(Y_i) = b^{'}\left(\theta_i\right)$, $Var\left(Y_i\right) = b^{''}\left(\theta_i\right) a\left(\phi\right)$

où $b^{'}\left(\theta_i\right)$ *et* $b^{''}\left(\theta_i\right)$ *représentent successivement la dérivé première et seconde de* $b\left(.\right)$ *par rapport à* θ_i

Preuve. D'après la formule $(4.2.3)$, il résulte

$$\begin{aligned} \partial L_i/\partial\theta_i &= \left[y_i - b^{'}\left(\theta_i\right)/a\left(\phi\right)\right] \text{ et} \\ \partial^2 L_i/\partial\theta_i^2 &= -b^{''}\left(\theta_i\right)/a\left(\phi\right) \quad , \end{aligned}$$

Sous les conditions de régularité satisfaites par la famille exponentielle, nous avons :

$$\begin{aligned} E\left(\frac{\partial L}{\partial\theta}\right) &= 0, \text{ et} \\ -E\left(\frac{\partial^2 L}{\partial\theta^2}\right) &= E\left(\frac{\partial L}{\partial\theta}\right)^2, \end{aligned}$$

donc pour une seule observation, il découle de la première formule:

$$E\left[Y_i - b^{'}\left(\theta_i\right)\right]/a\left(\phi\right) = 0$$

or :

$$\mu_i = E\left(Y_i\right) = b^{'}\left(\theta_i\right), \tag{4.2.4}$$

de la seconde formule il découle:

$$\begin{aligned} b^{''}(\theta_i)/a(\phi) &= E\left[Y_i - b^{'}(\theta_i)/a(\phi)\right]^2 \\ &= Var(Y_i)/\left[a(\phi)\right]^2. \end{aligned}$$

alors:

$$Var(Y_i) = b^{''}(\theta_i)\, a(\phi). \tag{4.2.5}$$

■

Proposition 4.2.2 *Les équations de vraisemblance d'un modèle LG sont données par*

$$\sum_i^n \frac{(y_i - \mu_i)\, x_{ij}}{var(Y_i)} \frac{\partial \mu_i}{\partial \eta_i} = 0 \ , \ j = 1, \ldots, d \tag{4.2.6}$$

Preuve. D'après la formule (4.2.3), la fonction log-vraisemblance est donnée par:

$$\begin{aligned} L(\beta) &= \sum_{i=1}^n L_i = \sum_i \log f(y_i, \theta_i, \phi) \\ &= \sum_{i=1}^n \frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + \sum_{i=1}^n c(y_i, \phi) \end{aligned} \tag{4.2.7}$$

Les équations de vraisemblance sont

$$\frac{\partial L(\beta)}{\partial \beta_j} = \sum_i \frac{\partial L_i}{\partial \beta_j} = 0 \ , \ j = 1, \ldots, d$$

pour dériver la log-vraisemblance (4.2.7) , nous utilisons la chaine suivante

$$\frac{\partial L_i}{\partial \beta_j} = \frac{\partial L_i}{\partial \theta_i} \frac{\partial \theta_i}{\partial \mu_i} \frac{\partial \mu_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta_j} \tag{4.2.8}$$

puisque $\partial L_i/\partial\theta_i = \left[y_i - b^{'}(\theta_i)\right]/a(\phi)$ et $\mu_i = b'(\theta_i)$, $var(Y) = b''(\theta_i)\, a(\phi)$ (d'après (4.2.3), (4.2.4) et (4.2.5) respectivement) alors

$$\frac{\partial L_i}{\partial \theta_i} = \frac{y_i - b^{'}(\theta_i)}{a(\phi)}$$

et

$$\frac{\partial \mu_i}{\partial \theta_i} = b''(\theta_i) = \frac{var(Y_i)}{a(\phi)}$$

aussi puisque $\eta_i = \sum_j \beta_j x_{ij}$ alors

$$\frac{\partial \eta_i}{\partial \beta_j} = x_{ij}$$

finalement puisque $\eta_i = g(\mu_i)$, $\partial \mu_i / \partial \eta_i$ dépend de la fonction lien du modèle.

En résumé, la formule (4.2.8) s'écrit

$$\begin{aligned} \frac{\partial L_i}{\partial \beta_j} &= \frac{y_i - b'(\theta_i)}{a(\phi)} \frac{a(\phi)}{var(Y_i)} \frac{\partial \mu_i}{\partial \eta_i} x_{ij} \\ &= \frac{(y_i - \mu_i) x_{ij}}{var(Y_i)} \frac{\partial \mu_i}{\partial \eta_i} \end{aligned} \tag{4.2.8}$$

les équations de vraisemblance sont alors

$$\sum_i^n \frac{(y_i - \mu_i) x_{ij}}{var(Y_i)} \frac{\partial \mu_i}{\partial \eta_i} = 0 \ , \ j = 1, \ldots, d \tag{4.2.9}$$

■

Remarque 4.2.2 · *Les équations de vraisemblance dépendent de β à travers μ_i car $\mu_i = g^{-1}\left(\sum_{j=1}^d \beta_j x_{ij}\right)$*

Les équations de vraisemblances (4.2.9) ne sont pas linéaires en β, pour cela on décrit des méthodes itératives pour résoudre une équation non linéaire ultérieurement.

Variance asymptotique d'estimateur du maximum de vraisemblance

La fonction de vraisemblance du modèle LG détermine la matrice de covariance asymptotique de $\widehat{\beta}$ estimateur du maximum de vraisemblance de β. Cette matrice est l'inverse de la matrice d'information de Fisher $I(\beta)$, avec éléments $E\left[-\partial^2 L(\beta) / \partial \beta_h \partial \beta_j\right]$.
En utilisant le résultat suivant sous les conditions de régularité vérifiées par la famille exponentielle,

$$E\left(\frac{\partial^2 L_i}{\partial \beta_h \partial \beta_j}\right) = -E\left(\frac{\partial L_i}{\partial \beta_h}\right)\left(\frac{\partial L_i}{\partial \beta_j}\right)$$

on trouve

$$\begin{aligned} E\left(\frac{\partial^2 L_i}{\partial \beta_h \partial \beta_j}\right) &= -E\left[\frac{(Y_i - \mu_i) x_{ih}}{var(Y_i)} \frac{\partial \mu_i}{\partial \eta_i} \frac{(Y_i - \mu_i) x_{ij}}{var(Y_i)} \frac{\partial \mu_i}{\partial \eta_i}\right] \\ &= \frac{-x_{ih} x_{ij}}{var(Y_i)} \left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2 \qquad , 1 \leq j, h \leq d \end{aligned}$$

puisque $L(\beta) = \sum_i L_i$, alors

$$E\left(-\frac{\partial^2 L}{\partial\beta_h \partial\beta_j}\right) = \sum_{i=1}^{n} \frac{x_{ih}x_{ij}}{var(Y_i)} \left(\frac{\partial\mu_i}{\partial\eta_i}\right)^2 \tag{4.2.10}$$

L'expression (4.2.10) représente l'élément (h, j) de la matrice $I(\beta)$, sa forme est donnée par

$$I(\beta) = X'WX \tag{4.2.11}$$

W est une matrice diagonale d'éléments

$$w_i = \frac{(\partial\mu_i/\partial\eta_i)^2}{var(Y_i)}, \quad i = 1, \ldots, n \tag{4.2.12}$$

La matrice de covariances de $\widehat{\beta}$ est estimée par

$$\widehat{cov}\left(\widehat{\beta}\right) = \widehat{I(\beta)}^{-1} = \left(X'\widehat{W}X\right)^{-1}$$

où $\widehat{W}$ est la matrice W évaluée à $\widehat{\beta}$. De (4.2.12) on remarque que W dépend de la fonction lien du modèle LG.

Estimation par les méthodes de Newton-Raphson et Fisher scoring

Méthode de Newton-Raphson C'est une méthode itérative pour résoudre une équation non linéaire, notre but est de détérminer la valeur $\widehat{\beta}$ qui maximise la fonction log-vraisemblance $L(\beta)$.

Soit

$$u' = (\partial L(\beta)/\partial\beta_1, \partial L(\beta)/\partial\beta_2, \ldots, \partial L(\beta)/\partial\beta_d)$$

et H la matrice d'éléments

$$h_{a,b} = \partial^2 L(\beta)/\partial\beta_a \partial\beta_b \quad , 1 \leq a, b \leq d$$

appelée la *matrice Hessiènne*. $u^{(k)}$et $H^{(k)}$ sont u et H évalués en $\beta^{(k)}$, l'itération k de $\widehat{\beta}$. L'étape k du processus itératif $(k = 0, 1, 2, \ldots)$ donne l'approximation de $L(\beta)$ au voisinage de $\beta^{(k)}$ par le developpement de Taylor comme suit

$$L(\beta) \approx L\left(\beta^{(k)}\right) + u^{(k)'}\left(\beta - \beta^{(k)}\right) + \frac{1}{2}\left(\beta - \beta^{(k)}\right)' H^{(k)}\left(\beta - \beta^{(k)}\right)$$

en résolvant l'équation

$$\frac{\partial L(\beta)}{\partial \beta} \approx u^{(k)} + H^{(k)}\left(\beta - \beta^{(k)}\right) = 0$$

pour β valeur résultante d'itération suivante $(k+1)$. Cette itération peut être exprimée comme suit

$$\beta^{(k+1)} = \beta^{(k)} - \left[H^{(k)}\right]^{-1} u^{(k)}, \tag{4.2.13}$$

en supposant que $H^{(k)}$ est une matrice nonsingulière. L'estimateur de maximum de vraisemblance est la limite de $\beta^{(k)}$ quand $k \to \infty$.

La convergence de $\beta^{(k)}$ vers $\widehat{\beta}$ de la méthode de Newton Raphson est souvent rapide.

Méthode de Fisher scoring Quand il s'agit des modèles linéaires généralisés, la méthode préconisée pour estimer le paramètre d'intérêt, est celle de Fisher scoring qui n'est rien d'autre que celle de Newton Raphson, sauf que dans le schéma itératif (4.2.13), H est remplacée par son espérance, a comme éléments $E\left[\partial^2 L(\beta)/\partial\beta_h \partial\beta_j\right]$ évalués à l'itération k alors,

$$E\left(H^{(k)}\right) = -I^{(k)}$$

avec $I^{(k)}$ est la matrice d'information de Fisher évaluée à l'itération k. Le schéma itérative à ce moment est donné par:

$$\beta^{(k+1)} = \beta^{(k)} + \left[I^{(k)}\right]^{-1} u^{(k)}$$

ou

$$I^{(k)}\beta^{(k+1)} = I^{(k)}\beta^{(k)} + u^{(k)} \tag{4.2.14}$$

4.2.3 Méthode des moindres carrés généralisée itérative

Une relation existe entre l'estimation par maximum de vraisemblance en utilisant Fisher scoring et l'estimation par les moindres carrés généralisés.
Dans le cas d'un modèle de régression linéaire ayant la forme générale suivante :

$$z = X\beta + \varepsilon$$

où la matrice de covaraince de ε est V, l'estimateur des moindres carrées généralisés de β est

$$\widehat{\beta} = \left(X'V^{-1}X\right)^{-1} X'V^{-1}z$$

Le terme à droite de l'expression (4.2.14) est le vecteur $I^{(k)}\beta^{(k)} + u^{(k)}$ et d'après l'expression (4.2.11) alors

$$I^{(k)}\beta^{(k+1)} = X'W^{(k)}X\beta^{(k)} + u^{(k)}$$

ayant pour éléments

$$\sum_{j=1}^{d}\sum_{i=1}^{n} x_{ih}x_{ij}\frac{\left(\frac{\partial\mu_i}{\partial\eta_i}\right)^2}{var\left(Y_i\right)}\beta_j^{(k)} + \sum_{i=1}^{n}\frac{1}{var\left(Y_i\right)}\left(y_i - \mu_i^{(k)}\right)x_{ih}\frac{\partial\mu_i^{(k)}}{\partial\eta_i^{(k)}}$$

simplifié comme suit

$$\sum_{i=1}^{n} x_{ih}\frac{\left(\frac{\partial\mu_i}{\partial\eta_i}\right)^2}{var\left(Y_i\right)}\left(\sum_{j=1}^{d} x_{ij}\beta_j^{(k)} + \left(y_i - \mu_i^{(k)}\right)\frac{\partial\mu_i^{(k)}}{\partial\eta_i^{(k)}}\right)$$

où $\mu_i^{(k)}$ et $\frac{\partial\mu_i^{(k)}}{\partial\eta_i^{(k)}}$ sont μ_i et $\frac{\partial\mu_i}{\partial\eta_i}$ évalués en $\beta = \beta^{(k)}$.
En posant

$$z_i^{(k)} = \sum_{j=1}^{d} x_{ij}\beta_j^{(k)} + \left(y_i - \mu_i^{(k)}\right)\frac{\partial\mu_i^{(k)}}{\partial\eta_i^{(k)}}$$

les composantes du vecteur $X'W^{(k)}X\beta^{(k)} + u^{(k)}$ ont la forme suivante

$$\sum_{i=1}^{n} x_{ih}\frac{\left(\frac{\partial\mu_i}{\partial\eta_i}\right)^2}{var\left(Y_i\right)}z_i^{(k)} = X'W^{(k)}z_i^{(k)}$$

alors les équations (4.2.13) de Fisher s'écrivent

$$\left(X'W^{(k)}X\right)\beta^{(k+1)} = X'W^{(k)}z^{(k)}$$

se sont les équations normales utilisées dans la méthode des moindres carrés généralisés qui ajuste le modèle linéaire avec $z^{(k)}$ variable dépendante, X matrice du plan d'expérience du modèle et $W^{(k)}$ est l'inverse de la matrice de covariance du modèle linéaire. La solution ainsi est donnée par

$$\beta^{(k+1)} = \left(X'W^{(k)}X\right)^{-1}X'W^{(k)}z^{(k)}$$

L'estimateur du maximum de vraisemblance est obtenu par la méthode des moindres carrés généralisée où la matrice poids (inverse de la covariance) change à chaque itération. Ce procédé est dit *méthode des moindres carrés généralisée itérative.*

4.3 Modèles log-linéaires

Une classe spécifique des modèles linéaires généralisés est la classe des modèles log-linéaire, alors la fonction lien est la fonction *log*, autrement dit le logarithme de l'espérance de la variable réponse sera une combinaison linéaire des variables explicatives.

4.3.1 Modèle log-linéaire de Poisson

Soit $n = (n_1, \ldots, n_s)$ un vecteur de s variables indépendantes représentant un dénombrement, $\mu = (\mu_1, \ldots, \mu_s)$ est le vecteur moyen $E(n)$.
Le modèle de dénombrement le plus simple est le modèle de *poisson*.

n_i est une variable à expliquer (réponse) représentant un dénombrement a comme espérance

$$\mu_i = E(n_i),$$

et

$$f(n_i; \mu_i) = \exp(-\mu_i) \frac{\mu_i^{n_i}}{n_i!}, \quad n_i = 0, 1, 2, \ldots. \tag{4.3.1}$$

la densité de n_i a la forme exponentielle simplifiée avec : $\theta_i = \mu_i$, $A(\mu_i) = \exp(-\mu_i)$, $B(n_i) = \frac{1}{n_i!}$ et $Q(\mu_i) = \log \mu_i$.
Le paramètre naturel est : $\log \mu_i$, alors la fonction lien canonique est $\eta_i = \log \mu_i$, d'où le modèle qui utilise ce lien

$$\log \mu_i = \sum_{j=1}^{d} \beta_j x_{ij} \ , \ i = 1, \ldots, s.$$

avec x_{ij} réalisation de la variable explicative X_{ij}, $\beta = (\beta_1, \ldots, \beta_d)$ le paramètre du modèle à estimer, alors

$$\begin{cases} n = \exp X\beta + \varepsilon \\ E(n) = \exp X\beta \\ \quad var(n) = var\varepsilon = diag(\mu_1, \ldots, \mu_s) \end{cases}$$

avec $E(\varepsilon_i) = 0$, $var(\varepsilon_i) = \mu_i$ et $cov(\varepsilon_i, \varepsilon_j) = 0$ pour $i \neq j$.
Ce modèle est dit: modèle *log-linéaire de poisson.*

4.3.2 Maximum de vraisemblance d'un modèle log-linéaire de Poisson

Etant donné un modèle log-linéaire, nous estimons β, son vecteur de paramètres en utilisant la méthode du maximum de vraisemblance. Vu que le modèle log-linéaire est un modèle linéaire généralisé, alors ces équations de vraisemblance ne sont pas linéaire en β, pour les résoudre nous proposons d'utiliser la méthode de Newton-Raphson qui est rien d'autre que la méthode de Fisher scoring dans ce cas.

Les équations de vraisemblance

Pour un vecteur de dénombrement n avec $E(n) = \mu$, le modèle log-linéaire est

$$\log \mu = X\beta$$

c'est-à-dire

$$\log \mu_i = \sum_j \beta_j x_{ij} \ , \ i = 1, \ldots, s \tag{4.3.2}$$

d'après la formule (4.3.1) et (4.2.7) le noyau de la fonction log-vraisemblace est

$$L(\beta) = \sum_i (n_i \log \mu_i - \mu_i)$$

d'aprés la formule (4.3.2), il résulte

$$L(\beta) = \sum_i \left[n_i \sum_j \beta_j x_{ij} - \exp\left(\sum_j \beta_j x_{ij} \right) \right] \tag{4.3.3}$$

alors

$$\frac{\partial L(\beta)}{\partial \beta_j} = \sum_i (n_i - \mu_i) x_{ij} \ , \ j = 1, \ldots, d \tag{4.3.4}$$

d'après la formule (4.2.9) les équations de vraisemblance d'un modèle log-linéaire de poisson sont données par

$$\frac{\partial L(\beta)}{\partial \beta_j} = 0 \Leftrightarrow \sum_i (n_i - \mu_i) x_{ij} = 0, \quad j = 1, \ldots, d \tag{4.3.5}$$

autrement écrit

$$\begin{cases} \sum_{i=1}^{s} x_{i1}\left[n_i - \exp\left(\sum_j \beta_j x_{ij}\right)\right] = 0 \\ \sum_{i=1}^{s} x_{i2}\left[n_i - \exp\left(\sum_j \beta_j x_{ij}\right)\right] = 0 \\ \vdots \qquad\qquad\qquad\qquad\qquad \vdots \\ \sum_{i=1}^{s} x_{id}\left[n_i - \exp\left(\sum_j \beta_j x_{ij}\right)\right] = 0 \end{cases} \Longleftrightarrow X'(n-\mu) = 0$$

Les équations de vraisemblance ne sont pas linéaires en β, alors le modèle log-linéaire n'a pas d'estimateur direct.

Autrement dit; l'estimateur du maximum de vraisemblance n'a pas une forme explicite.

Estimation par les méthodes de Newton-Raphson et de Fisher scoring

Le chemain itératif de la méthode de Newton-Raphson est donné par

$$\beta^{(k+1)} = \beta^{(k)} - \left[H^{(k)}\right]^{-1} u^{(k)}, \quad k = 0, 1, \ldots \tag{4.3.6}$$

avec H est la matrice Hessiènne du modèle et u est le vecteur gradient

d'après la formule (4.3.4), il résulte

$$u_j = \frac{\partial L(\beta)}{\partial \beta_j} = \sum_i (n_i - \mu_i) x_{ij}, \quad j = 1, \ldots, d$$

alors le vecteur gradient u est donné par

$$u = X'(n - \mu)$$

où n est le vecteur des observations et $\mu = E(n)$, X est la matrice du modèle log-linéaire.

De l'expréssion (4.2.10) nous avons

$$h_{j,h} = \frac{\partial^2 L_i}{\partial \beta_j \partial \beta_h} = -\sum_i x_{ij}\mu_i x_{ih}, \quad 1 \le j, h \le d$$

alors la matrice Hessiènne est donnée par

$$H = \begin{pmatrix} -\sum_i x_{i1}^2\mu_i & -\sum_i x_{i1}\mu_i x_{i2} & \ldots & -\sum_i x_{i1}\mu_i x_{id} \\ -\sum_i x_{i2}\mu_i x_{i1} & -\sum_i x_{i2}^2\mu_i & \ldots & -\sum_i x_{i2}\mu_i x_{id} \\ \vdots & \vdots & \ddots & \vdots \\ -\sum_i x_{id}\mu_i x_{i1} & -\sum_i x_{id}\mu_i x_{i2} & \ldots & -\sum_i x_{id}^2\mu_i \end{pmatrix}$$

d'où

$$H = -X^{'}WX$$

avec $W = diag\left(\mu_1, \ldots, \mu_s\right)$.
La $k^{ème}$ approximation $\mu^{(k)}$ de $\widehat{\mu}$ calculée en $\beta^{(k)}$ telle que $\mu^{(k)} = \exp\left(X\beta^{(k)}\right)$, donne $\beta^{(k+1)}$.
Le chemin iteratif (4.3.6), simplifié pour les modèles log-linéaires de poisson comme suit

$$\beta^{(k+1)} = \beta^{(k)} + \left[X^{'}W^{(k)}X\right]^{-1} X^{'}\left(n - \mu^{k}\right)$$

Remarque 4.3.1 *La matrice Hessiènne H est indépendante des variables aléatoires n_i, alors $E\left(H\right) = H$, ce qui fait que la méthode de Newton-Raphson et de Fisher scoring coïncident.*

4.3.3 Propriétés asymptotiques de $\widehat{\beta}$

Soit $\widehat{\beta}$ l'estimateur du maximum de vraisemblance du modèle log-linéaire, $\widehat{\beta}$ est asymptotiquement, sans biais, de variance minimale, donc $cov\widehat{\beta}$ atteint la borne de Cramer-Rao c'est-à-dire

$$cov\widehat{\beta} = \left[I\left(\beta\right)\right]^{-1}$$

avec $I\left(\beta\right)$ est la matrice d'information de Fisher, l'élément $I_{j,h}\left(\beta\right)$ de cette matrice est donné par

$$I_{j,h}\left(\beta\right) = E\left(-\frac{\partial^2 L}{\partial\beta_j\partial\beta_h}\right) = \sum_i x_{ij}\mu_i x_{ih}$$

alors

$$I\left(\beta\right) = X^{'}WX$$

ce qui donne

$$cov\widehat{\beta} = \left[X^{'}\widehat{W}X\right]^{-1}$$

4.3.4 Validation et critères de sélection d'un modèle log-linéaire

Sélection

Déviance Aprés avoir ajusté un modèle linéaire généralisé à un ensemble de données, il est nécessaire par la suite, de comparer les valeurs ajustées $\widehat{y}$, sous ce modèle, aux valeurs obsevées y.

Nous notons

$$D(y, \widehat{\mu}) = -2\left[l\left(\widehat{\mu}, \phi, y\right) - l\left(y, \phi, y\right)\right]$$

avec $l\left(\widehat{\mu}, \phi, y\right)$ est la fonction log-vraisemblanse maximisée en β, $l\left(y, \phi, y\right)$ est le maximum sous le modèle saturé.

$D(y, \widehat{\mu})$ est appelée *déviance* pour le modèle courant.

Remarque 4.3.2 *pour la distribution de poisson, la déviance n'est rien d'autre que la statistique de test du rapport du maximum de vraisemblance G^2 donnée par*

$$G^2 = 2\sum_{i=1}^{n}\left[y_i \log\left(\frac{y_i}{\widehat{\mu_i}}\right) - (y_i - \widehat{\mu_i})\right]$$

Critères AIC et BIC Les critères AIC (*Akaike Information Criterion*) et BIC (*Bayesian Information Criterion*) sont basés sur la philosophie suivante : plus la vraisemblance est grande, plus grande est la log-vraisemblance et meilleur est le modèle.

Pour choisir des modèles plus parcimonieux, une stratégie consiste à pénaliser la vraisemblance par une fonction du nombre de paramètres.

·par définition l'AIC pour un modèle à p paramètres est

$$AIC = -2l\left(\widehat{\beta}\right) + 2p$$

·Le critère de choix de modèle le BIC pour un modèle à p paramètres estimer sur n observations est défini par

$$BIC = -2l\left(\widehat{\beta}\right) + p\log(n)$$

nous choisissons ainsi le modèle qui possède le plus petit AIC ou BIC.

4.4 Modèles log-linéaires pour les tables de contingence

Les données se présentent généralement sous la forme d'une table de contingence obtenue par le croisement de plusieurs variables qualitatives dont chaque cellule contient un effectif ou une fréquence à modéliser.

4.4.1 Table de contingence à deux entrées

Nous considérons une table de contingence $I \times J$ c'est-à-dire avec deux variables réponses catégorielles X, Y ayant I, J modalités respéctivement. La probabilité d'une cellule (i, j) est p_{ij}, sa fréqence observée est notée n_{ij}, sa fréqence moyenne (théorique) est $\mu_{ij} = Np_{ij}$ avec N est le nombre total des individus observés.

Le modèle log-linéaire utilise la loi de poisson de paramètre μ_{ij} $\left(E\left(n_{ij}\right) = \mu_{ij}\right)$ pour les observations n_{ij} ou la loi multinomiale pour les $I \times J$ cellules.

Modèle log-linéaire d'indépendance

Nous supposons que les deux variables réponses X, Y sont indépendantes, ce qui implique

$$\mu_{ij} = Np_{i+}p_{+j},\ 1 < i \leq I \text{ et } 1 < j \leq J \tag{4.4.1}$$

avec

p_{i+} est la probabilté marginale de la $i^{ème}$ ligne

p_{+j} est la probabilité marginale de la $j^{ème}$ colonne.

La forme log-linéaire pour l'expréssion (**4.4.1**) est la suivante

$$\log \mu_{ij} = \beta_0 + \beta_i^X + \beta_j^Y, \quad 0 < i \leq I \text{ et } 0 < j \leq J \tag{4.4.2}$$

avec: $\beta_0 = \ln N$, $\beta_i^X = \ln\, p_{i+}$, $\beta_j^Y = \ln p_{+j}$ sous la condition

$$\sum_i \beta_i^X = \sum_j \beta_j^Y = 0$$

β_i^X, β_j^Y représentent effet ligne, effet colonne respectivement.

La formule (4.4.2) est la forme générale d'un modèle *log-linéaire d'indépendance.*

Modèle log-linéaire saturé

Le modèle log-linéaire dans le cas de dépendance a la forme suivante

$$\log \mu_{ij} = \beta_0 + \beta_i^X + \beta_j^Y + \beta_{ij}^{XY}, \quad 1 < i \leq I \text{ et } 1 < j \leq J \tag{4.4.3}$$

β_{ij}^{XY} est un paramètre qui représente l'effet d'interaction entre X et Y telque

$$\sum_i \beta_{ij}^{XY} = \sum_j \beta_{ij}^{XY} = 0$$

Le modèle $(4, 4, 3)$ est dit *modèle saturé.* C'est un modèle *hiérarchique, a*utrement dit le modèle inclu tous les termes d'ordre inférieur composés des variables qui apparaissent dans le terme d'ordre maximum. Si le modèle contient β_{ij}^{XY}, alors il contient β_i^X et β_j^Y forcement. un exemple d'un modèle *nonhiérarchique* est:

$$\log \mu_{ij} = \beta_0 + \beta_i^X + \beta_{ij}^{XY}$$

Types d'indépendance et modèles log-linéaires associés

une table de contingence à trois dimentions croise trois variables réponses X, Y, Z ont différents types d'indépendance.

On suppose que : p_{ijk} est la probabilité de la cellule (i, j, k).

·Les trois variables sont mutuellement indépendantes si et seulement si :

$$p_{ijk} = p_{i++}p_{+j+}p_{++k} \quad \text{pour } i = 1, \ldots, I,\ j = 1, \ldots, J \text{ et } k = 1, \ldots, K\ ,$$

avec: p_{i++} est la probabilité marginale de la variable X.

p_{+j+} est la probabilité marginale de la variable Y.

p_{++k} est la probabilité marginale de la variable Z.

dans ce cas la forme du modèle log-linéaire est celle d'un modèle d'indépendance :

$$\log \mu_{ijk} = \beta_0 + \beta_i^X + \beta_j^Y + \beta_k^Z$$

·La variable Y est conjointement indépendante de X et Z si et seulement si:

$$p_{ijk} = p_{i+k}p_{+j+} \quad \text{pour tous } i,\ j \text{ et } k\ ,$$

Le modèle log-linéaire est:

$$\log \mu_{ijk} = \beta_0 + \beta_i^X + \beta_j^Y + \beta_k^Z + \beta_{ik}^{XZ}$$

de la même façon, X peut être conjointement indépendante de Y et Z ou , Z peut être conjointement indépendante de X et Y.

$\cdot X$ et Y sont conditionnellement indépendantes, sachant Z si et seulement si, l'indépendance est vérifiée pour chaque table partielle en fixant Z, autrement dit:

$$p_{ij/k} = p_{i+/k} p_{+j/k}, \quad \text{pour tous } i,\ j \text{ et } k\ ,$$

avec:

$$p_{ij/k} = p(X = i, Y = j/Z = k),$$

l'indépendance conditionnelle de X et Y sachant Z, résulte le modèle log-linéaire suivant:

$$\log \mu_{ijk} = \beta_0 + \beta_i^X + \beta_j^Y + \beta_k^Z + \beta_{ik}^{XZ} + \beta_{jk}^{YZ}.$$

Le modèle log-linéaire saturé d'une table de contingence à trois dimensions est le suivant:

$$\log \mu_{ijk} = \beta_0 + \beta_i^X + \beta_j^Y + \beta_k^Z + \beta_{ik}^{XY} + \beta_{ik}^{XZ} + \beta_{ik}^{YZ} + \beta_{ijk}^{XYZ},$$

avec: β_{ijk}^{XYZ} est le coefficient d'interaction entre les trois variables.

Table 4.2 : les modèles log-linéaires pour les tables de contingence à trois dimensions

$modèle\ log-linéaire$	$notation$
$\log \mu_{ijk} = \beta_0 + \beta_i^X + \beta_j^Y + \beta_k^Z$	(X, Y, Z)
$\log \mu_{ijk} = \beta_0 + \beta_i^X + \beta_j^Y + \beta_k^Z + \beta_{ik}^{XY}$	(XY, Z)
$\log \mu_{ijk} = \beta_0 + \beta_i^X + \beta_j^Y + \beta_k^Z + \beta_{ik}^{XY} + \beta_{ik}^{XZ}$	(XY, YZ)
$\log \mu_{ijk} = \beta_0 + \beta_i^X + \beta_j^Y + \beta_k^Z + \beta_{ik}^{XY} + \beta_{ik}^{XZ} + \beta_{jk}^{YZ}$	(XY, YZ, XZ)
$\log \mu_{ijk} = \beta_0 + \beta_i^X + \beta_j^Y + \beta_k^Z + \beta_{ik}^{XY} + \beta_{ik}^{XZ} + \beta_{ik}^{YZ} + \beta_{ijk}^{XYZ}$	(XYZ)

4.5 Modèle log-linéaire dans le cas des données de capture-recapture

Nous considèrons une expérience de capture-recapture effectuée sur une population fermée de taille inconnue N. Soit t $(t \geq 2)$ le nombre d'occasions de capture de cette expérience.

Un historique de capture individuel peut être exprimé comme vecteur ligne :
$W = (w_1, \ldots, w_t)$ où

$$w_j = \begin{cases} 1 & \text{si l'individu est capturé à l'occasion } j \\ 0 & \text{si non} \end{cases}$$

avec $j = 1, 2, \ldots, t$.
Le nombre d'historiques de capture des individus capturés exactement j fois est C_t^j alors le nombre total des historiques de capture est $C_t^1 + C_t^2 + \ldots + C_t^t$

$$2^t = \sum_{j=0}^{t} C_t^j$$

A la fin de l'expérience, le nombre d'historiques observables possibles est $s = 2^t - 1$.

Remarque 4.5.1 *Si $t = 3$, les historiques de capture observables possibles sont*

$$(1,0,0)\,,\ (0,1,0)\,,\ (0,0,1)\,,\ (1,1,0)\,,\ (1,0,1)\,,\ (0,1,1)\,,\ (1,1,1)\,.$$

Remarque 4.5.2 *L'historique non observable celui des individus qui n'ont jamais été capturés durant l'expérience est défini comme suit :* $(0, \ldots, 0)$.

Nous notons, pour le *ième* historique de capture observable $W^{(i)}$, $i = 1, \ldots, s$
$n_i =$ le nombre d'individus dans la population ayant l'historique de capture $W^{(i)} = (w_1^{(i)}, \ldots, w_t^{(i)})$
$P_i =$ la probabilité d'appartenance à l'historique $W^{(i)}$
$\mu_i =$ la fréquence prédite associée a n_i, $\mu_i = NP_i$, (fréquence attendue)
$n =$ le nombre total d'individus capturés jusqu'à la dernière occasion de capture, autrement dit le nombre total d'individus capturés au moins une fois durant l'expérience, d'où

$$n = \sum_{i=1}^{2^t-1} n_i.$$

4.5.1 Modèle multinomial

Nous considèrons les variables aléatoires $N-n, n_1, \ldots, n_s$.

$N-n$ suit une loi binomiale de paramètres (N, P_0) et n_i suit une loi binomiale de paramètres (N, P_i), ce qui donne $(N-n, n_1, \ldots, n_s)$ suit une loi multinomiale de paramètre $(N, P_0, P_1, \ldots, P_s)$.

Ce modèle a été introduit par Darroch (1958). La fonction de vraisemblance est donnée par

$$L(N, P_0, P_1, \ldots, P_s) = \frac{N!}{(N-n)! \prod_{i=1}^{s} n_i!} P_0^{N-n} \prod_{i=1}^{s} P_i^{n_i}$$

4.5.2 Modèle de Poisson

Le modèle statistique de capture-recapture a une *fréquence prédite* μ_i associée à n_i. Le but de notre analyse statistique est d'estimer μ_0, le nombre d'individus oubliés (non observés) durant l'expérience, alors l'estimation est de N

$$\widehat{N} = n + \widehat{\mu_0}$$

Nous supposons que les variables aléatoires $n_1, \ldots, n_s$ sont indépendantes suivent une loi de Poisson de paramètres $\mu_1, \ldots, \mu_s$, respectivement, alors

$$f(n_i, \mu_i) = \frac{e^{-\mu_i} \mu_i^{n_i}}{n_i!}, \ i = 1, \ldots, s$$

D'aprés la formule (4.3.1), la densité de n_i a la forme exponentielle où $\log \mu_i$ est le paramètre naturel.

Montrons que pour les modèles de capture-recapture multiples M_0, M_t, M_{th} et M_b la fréquence prédite μ_i est log-linéaire, peut être écrite en fonction d'une matrice $s \times d$ (design matrix) X où d est la longueur du vecteur des paramètres du modèle log-linéaire β, nous déduisont les variables explicatives pour chaque modèle

4.5.3 Modèle M_0

Nous considérons le modèle de capture-recapture définit au premier chapitre. Rappelons qu'il est le plus simple des modèles de capture-recapture, d'où les occasions de capture sont indépendantes avec probabilité de capture commune p.

Nous avons

$$E(n_i) = \mu_i$$

donc

$$p_i = p^{\sum_{j=1}^{t} w_j^{(i)}} (1-p)^{t-\sum_{j=1}^{t} w_j^{(i)}}, i = 1, \ldots, s.$$

vu que

$$\mu_{w^{(i)}} = N p_{w^{(i)}}$$

alors

$$\mu_i = N p^{\sum_{j=1}^{t} w_j^{(i)}} (1-p)^{t-\sum_{j=1}^{t} w_j^{(i)}}, i = 1, \ldots, s. \tag{4.4.1}$$

Le modèle (4.4.1) est un modèle multiplicatif, linéarisant ce dérnier par la transformation logarithmique comme suit

$$\ln \mu_i = \ln N + \sum_{j=1}^{t} w_j^{(i)} \ln p + \left(t - \sum_{j=1}^{t} w_j^{(i)} \right) \ln (1-p),$$

donc

$$\ln \mu_i = \ln \left[N (1-p)^t \right] + \sum_{j=1}^{t} w_j^{(i)} \ln \frac{p}{1-p}, \tag{4.4.2}$$

nous posons

$$\begin{aligned} \beta_0 &= \ln \left[N (1-p)^t \right] \\ \beta_1 &= \ln \frac{p}{1-p}, \end{aligned}$$

et

$$\begin{aligned} X_{i0} &= 1, \\ X_{i1} &= \sum_{j=1}^{t} w_j^{(i)}, i = 1, \ldots, s, \end{aligned}$$

le modèle (4.4.2) peut être écrit sous la forme

$$\ln \mu_i = \beta_0 + \beta_1 x_{i1}, i = 1, \ldots, s \tag{4.4.3}$$

où x_{i1} est une réalisation de la variable explicative X_{i1}.

Remarque 4.5.3 *: X_1 = le nombre de fois où l'individu a été capturé, alors X_1 une variable aléatoire suit une loi binomiale $\beta(t, p)$.*

Le modèle (4.4.3) est un modèle *log-linéaire*, sa forme matricielle est donnée par

$$\underline{\eta} = X\underline{\beta}$$

où X est la matrice du modèle (design matrix) $s \times 2$ écrite comme suit

$$X = \begin{pmatrix} 1 & 1 \\ 1 & 2 \\ \vdots & \vdots \\ 1 & t \end{pmatrix}.$$

et

$$\underline{\beta} = (\beta_0, \beta_1)'$$

vecteur des paramètres du modèle log-linéaire, aussi

$$\underline{\eta} = (\ln \mu_1, \ldots, \ln \mu_s)', \quad \underline{\eta} \in \mathbb{R}^s$$

Pour ce modèle la fréquence prédite μ_0 des individus non enregistrés, est

$$\mu_0 = e^{\beta_0},$$

c'est-à-dire $\mu_0 = \left[N\,(1-p)^t\right]$.

4.5.4 Modèles M_t, M_h et M_{th}

Le modèle M_t suppose que les occasions de capture sont indépendantes et chaque occasion a sa propre probabilité de capture.

Soit : p_j la probabilité de capture d'un individu à l'occasion j, $j = 1, \ldots, t$.

Dans ce cas

$$p_i = \prod_{j=1}^{t} p_j^{w_j^{(i)}} \left(1 - p_j\right)^{1 - w_j^{(i)}},$$

or

$$\mu_i = N \prod_{j=1}^{t} p_j^{w_j^{(i)}} \left(1 - p_j\right)^{1 - w_j^{(i)}}, i = 1, \ldots, s.$$

donc nous avons un modèle multiplicatif, passant au logarithme

$$\begin{aligned} \ln \mu_i &= \ln N + \sum_{j=1}^{t} w_j^{(i)} \ln p_j + \sum_{j=1}^{t} \left(1 - w_j^{(i)}\right) \ln (1 - p_j) \\ &= \ln \left[N \prod_{j=1}^{t} (1 - p_j) \right] + \sum_{j=1}^{t} w_j^{(i)} \ln \frac{p_j}{1 - p_j} \end{aligned} \tag{4.4.4}$$

en posant

$$\begin{aligned} \beta_0 &= \ln \left[N \prod_{j=1}^{t} (1 - p_j) \right] \\ \beta_j &= \ln \frac{p_j}{1 - p_j} \end{aligned}$$

et

$$X_{i0} = 1, \quad X_{ij} = w_j^{(i)}, i = 1, \ldots, s$$

le modèle (4.4.4) s'écrit sous la forme

$$\ln \mu_i = \beta_0 + \sum_{j=1}^{t} \beta_j x_{ij}, i = 1, \ldots, s \tag{4.4.5}$$

avec x_{ij} est une réalisation de la variable explicative X_{ij}.

La formule (4.4.5) décrit le modèle *log-linéaire*, sa forme matricielle est donnée par

$$\underline{\eta} = X\underline{\beta}$$

avec $\underline{\beta} = (\beta_0, \beta_1, \ldots, \beta_t)'$ est le vecteur des paramètres du modèle, X la matrice du modèle a $t + 1$ colonnes et $\underline{\eta} = (\ln \mu_1, \ldots, \ln \mu_s)'$.

Sous ce modèle

$$\begin{aligned} \mu_0 &= e^{\beta_0}, \text{ c'est-à-dire} \\ \mu_0 &= N \prod_{j=1}^{t} (1 - p_j). \end{aligned}$$

Exemple 4.4.3 : Sous l'hypothèse d'un modèle M_t, nous supposons que *t=2.*

D'après la formule du modèle log -linéaire (4.4.5) nous avons

$$\begin{cases} \log \mu_{00} = \beta_0 \\ \ln \mu_{10} = \beta_0 + \beta_1 \\ \ln \mu_{01} = \beta_0 + \beta_2 \\ \ln \mu_{11} = \beta_0 + \beta_1 + \beta_2 \end{cases} \tag{4.4.6}$$

la forme matricielle s'écrit alors

$$\begin{pmatrix} \ln \mu_{00} \\ \ln \mu_{10} \\ \ln \mu_{01} \\ \ln \mu_{11} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix}$$

Le modèle M_t suppose que les t occasions de capture sont indépendantes. Bishop, Fienberg et Holland (1975) ont montré que cette hypothèse n'est pas toujours vérifiée; dans ce cas des interactions entres les occasions de capture s'imposent pour une bonne modélisation des données. Une colonne de la matrice du plan X correspondante a une interaction entre la $jème$ et la $kème$ occasion de capture est : $w_j \times w_k$. Avec cette nouvelle matrice du plan, l'intercept β_0 reste $\log \mu_0$.

Agresti (1994) et Darroch, Fienberg, Glonek et Junker (1993) ont suggéré de modéliser l'hétérogénité dans les probabilités de capture des N individus en ajoutant l'effet d'interaction dans la formule (4.4.5), pour tout couple d'occasions de capture. Il résulte le modèle log- linéaire pour M_{th} a $d = t + 2$ paramètres. La colonne d'hétérogéniété est : $\sum_{j>k} w_j \times w_k$, [24].

Le modèle log-linéaire pour M_{th} est donné par la formule suivante

$$\ln \mu_i = \beta_0 + \sum_{j=1}^{t} \beta_j x_{ij} + x_{i(t+1)} \beta_{t+1}, i = 1, \ldots, s \tag{4.4.7}$$

où $x_{i(t+1)}$ est une réalisation de la variable explicative $X_{i(t+1)} = w_j^{(i)} \times w_k^{(i)}$, $i = 1, \ldots, s$,[19].

4.5.5 Modèle M_b

Une généralisation de M_0 est que le comportement d'individu change aprés sa première capture de p à c.

Sous le modèle de capture-recapture M_b, la fréquence prédite est modélisée comme suit

$$\mu_{_i} = Np_i,$$

avec

$$p_i = (1-p)^{\min\left\{j;w_j^{(i)}=1\right\}-1} pc^{\sum_{j=1}^{t} w_{j-1}^{(i)}} (1-c)^{t-\left(\min\left\{j,w_j^{(i)}=1\right\}-1\right)-\left(\sum_{j=1}^{t} w_{j-1}^{(i)}\right)-1},$$

posons

$$\begin{aligned} X_{i1} &= \min\left\{j, w_j^{(i)} = 1\right\}, \text{et} \\ X_{i2} &= \sum_{j=1}^{t} w_j^{(i)} \end{aligned}$$

alors

$$\ln \mu_i = \ln\left[Np\,(1-c)^{t-1}\right] + X_1 \ln\frac{1-p}{1-c} + X_2 \ln\frac{c}{1-c}.$$

en notant

$$\begin{aligned} \beta_0 &= \ln\left[Np\,(1-c)^{t-1}\right], \text{ et} \\ \beta_1 &= \ln\frac{1-p}{1-c}, \text{ et} \\ \beta_2 &= \ln\frac{c}{1-c}. \end{aligned}$$

donc

$$\ln \mu_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2}\ ,\ i = 1, \ldots, s$$

d'où la modélisation *log-linéaire*.

4.5.6 Estimation de N

Nous considèrons le modèle log-linéaire de poisson précédent, en utilisant les fréquence observées n_i $(i = 1, \ldots, s)$ nous estimons le vecteur de paramètre de ce modèle β. Cette opération se fait à l'aide de la fonction $glmfit$ du $MATLAB$.

La fréquence manquante $\mu_{_0}$ est exprimée à l'aide du vecteur β

$$\mu_{_0} = NP_{_0} = h(\beta)$$

d'où

$$\widehat{\mu_{_0}} = h\left(\widehat{\beta}\right)$$

finalement, nous estimons la taille totale de la population N par

$$\widehat{N} = n + h\left(\widehat{\beta}\right) \tag{4.4.9}$$

pour les modéles de capture-recapture précédents, nous avons $h\left(\widehat{\beta}\right) = \exp\left(\widehat{\beta}_0\right)$alors

$$\widehat{N} = n + \exp\left(\widehat{\beta}_0\right) \tag{4.4.10}$$

Variance de $\widehat{N}$

Calculons la variance de $\widehat{N}$ sous le modèle de poisson

D'aprés l'expréssion (4.4.9)

$$\widehat{N} = n + h\left(\widehat{\beta}\right) \tag{4.4.11}$$

alors

$$\begin{aligned} var\left(\widehat{N}\right) &= var\left(n + h\left(\widehat{\beta}\right)\right) \\ &= var(n) + var\left(h\left(\widehat{\beta}\right)\right) + 2cov\left(n, h\left(\widehat{\beta}\right)\right) \end{aligned} \tag{4.4.12}$$

Nous avons montré que la fréquence manquante μ_0 a la forme suivante

$$\mu_0 = \exp\beta_0$$

sous les hypothèse des modèles M_0, M_t, M_{th}, M_b, alors:

$$\mu_0 = h(\beta) \text{ d'où } \widehat{\mu}_0 = h\left(\widehat{\beta}\right)$$

donc

$$h(\beta) = NP_0 \text{ où } P_0 = 1 - \sum_{i=1}^{s} P_i$$

alors

$$h(\beta) = NP_0\frac{1-P_0}{1-P_0} = \mu\frac{P_0}{1-P_0} \quad \text{où} \quad \mu = \sum_{i=1}^{s}\mu_i$$

en posant $\dfrac{P_0}{1-P_0} = c$, nous aurons $h(\beta) = \mu c$. nous pouvons déduire que $\dfrac{h\left(\widehat{\beta}\right)}{n} \simeq \widehat{c}$.

Calculons $cov\left(n, h\left(\widehat{\beta}\right)\right)$:

$$\begin{aligned} cov\left(n, h\left(\widehat{\beta}\right)\right) &= E\left(nh\left(\widehat{\beta}\right)\right) - EnE\left(h\left(\widehat{\beta}\right)\right) \\ &= E\left(n^2\frac{h\left(\widehat{\beta}\right)}{n}\right) - EnE\left(n\frac{h\left(\widehat{\beta}\right)}{n}\right) \\ &\simeq \widehat{c}\left[En^2 - (En)^2\right] \\ &\simeq \widehat{c}varn \end{aligned}$$

puisque $n = \sum_{i=1}^{s} n_i$ et $n_1, \ldots, n_s$ sont des variables indépendantes, et n_i suit la loi de poisson de paramètre μ_i $(i = 1, \ldots, s)$ alors lavariable aléatoire n suit la loi de poisson de paramètre $\mu = \sum_{i=1}^{s}\mu_i$, donc

$$cov\left(n, h\left(\widehat{\beta}\right)\right) \simeq \widehat{c}\mu$$

l'expression (4.4.10) est simplifiée comme suit

$$var\left(\widehat{N}\right) \simeq (1 + 2\widehat{c})\,\mu + var\left(h\left(\widehat{\beta}\right)\right)$$

Biais de $\widehat{N}$

D'aprés l'expression (4.4.9)on a $\widehat{N} = n + h\left(\widehat{\beta}\right)$,
Le développement de $h\left(\widehat{\beta}\right)$ au voisinage de β est donné par

$$h\left(\widehat{\beta}\right) = h(\beta) + \left(\widehat{\beta} - \beta\right)\nabla h\left(\widehat{\beta}\right) + \frac{1}{2}\left(\widehat{\beta} - \beta\right)'\nabla^2 h\left(\widehat{\beta}\right)\left(\widehat{\beta} - \beta\right)$$

alors

$$\widehat{N} = n + h(\beta) + \left(\widehat{\beta} - \beta\right)\nabla h\left(\widehat{\beta}\right) + \frac{1}{2}\left(\widehat{\beta} - \beta\right)'\nabla hf\left(\widehat{\beta}\right)\left(\widehat{\beta} - \beta\right) \tag{4.4.13}$$

McCullagh (1987) a donné une expréssion du biais de l'estimateur $\widehat{\beta}$ dans le cadre d'un modèle linéaire généralisé (Peter McCullagh. (1991)),[17] simplifiée pour les modèles log-linéaire comme suit

$$\begin{aligned} b &= E\left(\widehat{\beta}\right) - \beta \\ &= var\left(\widehat{\beta}\right) X'W\xi \end{aligned}$$

où $\xi = -\frac{1}{2} diag\left(X\left(X'WX\right)^{-1}X'\right)$,
le biais de N est déduit de l'expréssion (4.4.10)

$$E\left(\widehat{N}\right) - N = E\left[\left(\widehat{\beta}-\beta\right)\nabla h\left(\widehat{\beta}\right) + \frac{1}{2}\left(\widehat{\beta}-\beta\right)'\nabla^2 h\left(\widehat{\beta}\right)\left(\widehat{\beta}-\beta\right)\right] \tag{4.4.14}$$

Théorème 4.5.1 *[4]soit X une variable multidimentionelle, d'espérance μ et de matrice de variances covariances V. Si A est une matrice carrée alors*

$$E\left(X'AX\right) = \mu'A\mu + tr\left(AV\right)$$

il découle alors

$$E\left[\left(\widehat{\beta}-\beta\right)'\nabla^2 h\left(\widehat{\beta}\right)\left(\widehat{\beta}-\beta\right)\right] = b'\left(\nabla^2 h\left(\widehat{\beta}\right)\right)b + tr\left(\nabla^2 h\left(\widehat{\beta}\right) var\left(\widehat{\beta}\right)\right)$$

le biais (4.4.14) de $\widehat{N}$ est ainsi

$$E\left(\widehat{N}\right) - N = b'\left(\nabla h\left(\widehat{\beta}\right)\right) + \frac{1}{2}\left[b'\left(\nabla^2 h\left(\widehat{\beta}\right)\right)b + tr\left(\nabla^2 h\left(\widehat{\beta}\right) var\left(\widehat{\beta}\right)\right)\right]$$

Chapitre 5

Analyse des données épidémiologiques à l'aide des modèles de capture-recapture

5.1 Introduction

L'épidémiologie

"C'est la discipline qui étudie la dynamique des phénomènes de santé dans les populations, dans le but de mettre en évidence les facteurs qui les déterminent, ainsi que le rôle de ces facteurs, et de mettre en œuvre les mesures de correction appropriées", Daniel Schwartz.

"L'épidémiologie est l'étude de la distribution des maladies chez l'homme et des facteurs qui les influencent" B.MacMahon, 1960.

L'épidémiologie est l'étude de la distribution et des déterminants des états de santé ou des évènements de santé dans une population définie et l'application de cette étude au contrôle des problèmes de santé" Last, 1988.

L'épidémiologie est l'étude des facteurs influant sur la santé et les maladies des populations humaines.(médecine qui se rapporte à la répartition, à la fréquence et à la gravité des états pathologiques).

Pour élaborer et mettre en oeuvre des stratégies efficaces d'amélioration de la santé publique, axées sur la prévention et la lutte, il faut posséder suffisamment de données sur la maladie: répartition géographique et temporelle, modalités de survenue et sujets atteints.

Dans ce chapitre, nous présentons une étude en épidémiologie pour donner un aspect pratique aux résultats décrits sur le plan théorique. En premier temps, nous définissons des concepts nécessaires pour l'étude, ensuite nous présentons la structure des données épidémiologiques et leurs transformations autant que des données de capture-recapture. Términons par une application numérique sur un exemple réel dans le but d'estimer le nombre de cas de la pathologie visée ainsi que l'évaluation de l'exhaustivité des systèmes de surveillance et des registres.

5.2 Concepts généraux

Système de surveillance:

Il fait parti du système d'information sanitaire, c'est un outil de travail indispensable, dit source ou liste dans la términologie d'un statisticien.

Modalités de surveillance:

▷ Système de déclarations obligatoires (25 maladies transmissibles doivent être déclarées, sur la base de critères cliniques ou biologiques).

▷ Réseaux de laboratoires d'analyses biologiques et médicales.

▷ Réseau de pharmaciens pour la surveillance des ventes en médicaments.

▷ Enquêtes d'environnement.

▷ Analyse des certificats de décès.

5.2.1 Conditions d'application de la méthodologie de capture-recapture en épidémiologie

Conditions statistiques

• La population étudiée est fermée.

• Les sources sont indépendantes, c'est à dire que la probabilité qu'un individu soit recensé dans une source ne dépend pas de la probabilité qu'il soit recensé par une autre source. Avec plus de deux sources, la dépendance entre les sources peut être évaluée et prise en compte dans l'estimation grâce à l'application des modèles log-linéaires au données de capture-recapture.

• L'homogénéité de capture des cas: pour une source donnée, tous les individus de la population étudiée ont la même probabilité d'identification, pour cela, la notification des cas dans une source ne doit pas être liée à des variables caractérisant les cas (âge, sexe, lieu de résidence, gravité de la maladie...).

Conditions implicites

• Tous les cas identifiés sont des vrais cas.

• Les cas identifiés appartiennent à la zone géographique étudiée et à la même période d'étude.

• Tous les vrais cas communs et seulement les vrais cas communs sont identifiés, ce qui impose l'application d'un identifiant commun unique entre les sources.

5.2.2 Structure des données

Nous considèrons chaque liste enregitrée (source) comme un échantillon capturé, le nombre d'identifications, et les noms sont utilisés comme marques. Etre capturé dans un échantillon correspond à être identifié dans une liste.
Nous avons noté trois différences entre l'échantillonnage de capture-recapture d'une population animale et d'une population humaine

▷ Dans une étude sur une population animale, des méthodes de capture identiques peuvent être utilisées durant l'échantillonnage, mais à l'épidémiologie, différents types de sources d'identification sont utilisées.

▷ Il y a toujours des échantillons d'animaux capturés, mais à l'épidémiologie les listes sont limitées.

▷ Il y a un temps d'ordre dans l'expérience de capture-recapture d'animaux, mais en général les listes sont obtenues en même temps.

Supposons que N la taille de la population à étudier soit inconnue, ça sera notre paramètre d'intérêt, et t le nombre de listes ou sources disponibles, la présence et l'absence dans une liste sont notés 1, 0 respectivement. Si le nombre de listes est 3 alors, l'individu identifié par la première source, non par la deuxième et ni par la troisième appartient à l'historique de capture $(1\ 0\ 0)$, un individu identifié dans les trois sources appartient à l'historique de capture $(1\ 1\ 1)$. Chaque individu de la population appartient à l'un des historiques suivants : $(1\ 0\ 0)$, $(0\ 1\ 0)$, $(0\ 0\ 1)$, $(1\ 1\ 0)$, $(1\ 0\ 1)$, $(0\ 1\ 1)$, $(1\ 1\ 1)$, $(0\ 0\ 0)$ avec $(0\ 0\ 0)$ est l'historique de capture des individus non identifiés par aucune source.
Les individus identifiés dans les trois sources appartiennent à l'historique de capture $(1\ 1\ 1)$.

individus	*liste*1	*liste*2	*liste*3
1	X_{11}	X_{12}	X_{13}
2	X_{21}	X_{22}	X_{23}
...	...	...	...
M	X_{M1}	X_{M2}	X_{M3}
M+1	0	0	0
...	...	...	...
N	0	0	0

Supposons qu'il y a M individus identifiés et $N - M$ individus non comptés, donc $N - M$ appartiennent à l'historique $(0\ 0\ 0)$. Alors les données d'identification de tous les individus peuvent être exprimées par une matrice $N \times t$, notée X d'éléments $X_{ij} = 1$ si l'individu i est identifié dans la liste j, 0 si non. Les données d'identification des individus peuvent être colléctées sous forme de données catégorielles comme le montre l'exemple suivant :

*liste*1	*liste*2	*liste*3	*données*
0	0	0	n_{000}
0	0	1	n_{001}
0	1	0	n_{010}
0	1	1	n_{011}
1	0	0	n_{100}
1	0	1	n_{101}
1	1	0	n_{110}
1	1	1	n_{111}

en considérant que $n_{w_1,\ldots,w_t}$ le nombre d'individus ayant l'historique de capture $W = (w_1, w_2, \ldots w_t)$ où $w_j = 0$ indique une absence dans la liste j et $w_j = 1$ indique une présence dans la liste j.

Pour $t = 3$, il y a sept cellules ovbservées, n_{100} est le nombre d'individus ayant l'historique de capture (1 0 0) c'est-à-dire que le nombre d'individus identifié par la première liste et non identifiés par la deuxième et à la troisième liste. n_{000} est le nombre d'individus oubliés durant le recueil des données.

Structure des données pour l'analyse bayésienne

nous supposons avoir trois sources, 1, 2, 3 , notons

$n_1 = n_{1++}$: le nombre d'individus identifiés par la première source (1) alors

$$n_{1++} = n_{100} + n_{110} + n_{101} + n_{111}$$

$n_2 = n_{+1+}$: le nombre d'individus identifiés par la deuxième source (2) alors

$$n_{+1+} = n_{110} + n_{010} + n_{011} + n_{111}$$

$n_3 = n_{++1}$: le nombre d'individus identifiés par la troisième source (3) alors

$$n_{++1} = n_{101} + n_{011} + n_{001} + n_{111}$$

Le nombre d'individus recapturés (réidentifiés) m_2 dans la deuxième source est donné par

$$m_2 = n_{110} + n_{111}$$

Le nombre d'individus réidentifiés dans la troisième source m_3 est donnée par

$$m_3 = n_{101} + n_{011} + n_{111}$$

Les données enregistrées ainsi seront affichées dans le tablau suivant

i	n_i	m_i	M_i
1	$n_1 = n_{1++}$	0	0
2	$n_2 = n_{+1+}$	$m_2 = n_{110} + n_{111}$	n_1
3	$n_3 = n_{++1}$	$m_3 = n_{101} + n_{011} + n_{111}$	$n_1 + n_2$

5.3 Estimation du nombre total de cas à l'aide des modèles log-linéaires

Estimation du nombre total de cas avec plus de deux sources

Le modèle log-linéaire permet d'analyser des données de comptage croisées et d'estimer le nombre de cas attendu dans chaque cellule du tableau de contingence y compris la cellule vide, la valeur estimer pour le nombre de cas identifiés par aucune source permet d'obtenir une estimation de N et de sa variance. d'étudier les relations entre des variables qualitatives croisées dans un tableau de contingence. Il représente le logarithme népérien de la fréquence attendue d'une cellule du tableau comme une combinaison linéaire d'effets principaux et d'interactions. Dans la situation particulière de la capture-recapture, le tableau de contingence a une cellule structurellement vide correspondante à l'absence de notifications de cas dans l'ensemble des sources. Pour estimer les effectifs attendus le modèle utilise toutes les cellules du tableau sauf celle définie comme étant structurellement vide, et pour laquelle on attend une estimation. La présence d'une cellule structurellement vide rend impossible l'ajustement d'un modèle prenant en compte l'interaction d'ordre maximum entre toutes les sources. Les modèles log-linéaires permettent de calculer des estimations prenant en compte les dépendances entre les sources ainsi que les variables d'hétérogénéité de capture.

5.4 Application numérique

Modèles log-linéaires

Exemple réel :

Exemple 5.4.1 *Evaluation de la surveillance des infections à méningocoques en France en* 1996 *par la méthode de CR :*[*23*].
Le recueil des données:

▷*Le système de déclaration obligatoire (faite par les medecins) : DO.*

▷*Le centre national de référence de méningocoques : CNR.*

▷*Le réseau de surveillance des méningites bactériennes : EPIBAC.*

Objectifs de l'étude :

▷Déterminer le nombre total d'infections à méningocoque prises en charge en milieu hospitalier(confirmées par isolement de neisseria meningitidis dans le sang ou le liquide céphalorachidien LCR) en France 1996.

▷Evaluer l'exhaustivité des différents systèmes de surveillance : DO, CNR, EPIBAC.
La population à étudiée : la population résidant en France métropolitaine en .La période d'étude était comprise entre le premier Janvier 1996 et le 31 Décembre 1996.

Description des données

Table 5.1 : Répartition des cas de méningocoque selon les historiques de capture

DO	CNR	$EPIBAC$	*nombre d'individus observés*
1	0	0	47
0	1	0	44
0	0	1	21
1	1	0	93
0	1	1	65
1	0	1	18
1	1	1	127

Les modèles log-linéaires résultats de différentes hypothèses concernant les dépendance entres sources sont décrits, en déduisant leurs matrices du plan ainsi que leurs paramètres.

Hypothèse 5.1.1

Les trois sources, sont dépendantes.

D'après l'expréssion **4.4.7** le modèle log-linéaire a la forme suivante

$$\begin{cases} \ln \mu_{100} = \beta_0 + \beta_1 \\ \ln \mu_{010} = \beta_0 + \beta_2 \\ \ln \mu_{001} = \beta_0 + \beta_3 \\ \ln \mu_{110} = \beta_0 + \beta_1 + \beta_2 + \beta_4 \\ \ln \mu_{011} = \beta_0 + \beta_2 + \beta_3 + \beta_5 \\ \ln \mu_{101} = \beta_0 + \beta_1 + \beta_3 + \beta_6 \\ \ln \mu_{111} = \beta_0 + \beta_1 + \beta_2 + \beta_3 + \beta_4 + \beta_5 + \beta_6 \end{cases}$$

sa forme matricielle est donnée par

$$\eta = X\beta$$

où $\eta = (\ln \mu_{100}, \ldots, \ln \mu_{111})$ et X une matrice (7×7), matrice du modèle, $\beta = (\beta_0, \ldots, \beta_6)'$ autrement écrit

$$\begin{pmatrix} \ln \mu_{100} \\ \ln \mu_{010} \\ \ln \mu_{001} \\ \ln \mu_{110} \\ \ln \mu_{011} \\ \ln \mu_{101} \\ \ln \mu_{111} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \\ \beta_6 \end{pmatrix}$$

Ce modèle sera noté DC, CE, DE à la suite. C'est le modèle saturé

Remarque 5.4.1 *ce modèle contient autant de paramètres que d'observations, il est dit modèle saturé.*

Hypothèse 5.1.2 les sources DO et $EPIBAC$ sont dépendantes, CNR et $EPIBAC$ sont dépendantes or que DO et CNR sont indépendantes.

D'après l'expréssion **4.3.7** ce modèle log-linéaire a la forme suivante

$$\begin{pmatrix} \ln \mu_{100} \\ \ln \mu_{010} \\ \ln \mu_{001} \\ \ln \mu_{110} \\ \ln \mu_{011} \\ \ln \mu_{101} \\ \ln \mu_{111} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 & 0 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \end{pmatrix}$$

Ce modèle sera noté DE, CE.

Hypothèse 5.1.3 une dépendance existe entre DO et CNR, DO et $EPIBAC$, mais CNR et $EPIBAC$ sont indépendantes.

La forme matricielle de ce modèle log-linéaire est donnée par

$$\begin{pmatrix} \ln \mu_{100} \\ \ln \mu_{010} \\ \ln \mu_{001} \\ \ln \mu_{110} \\ \ln \mu_{011} \\ \ln \mu_{101} \\ \ln \mu_{111} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \end{pmatrix}$$

Ce modèle sera noté DC, DE.

Hypothèse 5.1.4 DO et CNR sont dépendantes, CNR et $EPIBAC$ sont dépendantes, DO et $EPIBAC$ sont indépendantes.

Sa forme matricielle est donnée par

$$\begin{pmatrix} \ln \mu_{100} \\ \ln \mu_{010} \\ \ln \mu_{001} \\ \ln \mu_{110} \\ \ln \mu_{011} \\ \ln \mu_{101} \\ \ln \mu_{111} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 & 0 & 1 \\ 1 & 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \end{pmatrix}$$

Ce modèle sera noté DC, CE.

Hypothèse 5.1.5 une dépendance existe entre DO et CNR.

Le modèle log-linéaire s'écrit alors comme suit

$$\begin{pmatrix} \ln \mu_{100} \\ \ln \mu_{010} \\ \ln \mu_{001} \\ \ln \mu_{110} \\ \ln \mu_{011} \\ \ln \mu_{101} \\ \ln \mu_{111} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{pmatrix}$$

Le modèle sera noté DC, E.

Hypothèse 5.1.6 une dépendance existe entre DO et $EPIBAC$.

Il résulte le modèle log-linéaire suivant

$$\begin{pmatrix} \ln \mu_{100} \\ \ln \mu_{010} \\ \ln \mu_{001} \\ \ln \mu_{110} \\ \ln \mu_{011} \\ \ln \mu_{101} \\ \ln \mu_{111} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{pmatrix}$$

Le modèle sera noté DE, C.

Hypothèse 5.1.7 une dépendance existe entre CNR et $EPIBAC$.

Le modèle log-linéaire qui résulte

$$\begin{pmatrix} \ln \mu_{100} \\ \ln \mu_{010} \\ \ln \mu_{001} \\ \ln \mu_{110} \\ \ln \mu_{011} \\ \ln \mu_{101} \\ \ln \mu_{111} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 1 \\ 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{pmatrix}$$

Ce modèle sera noté D, CE.

Hypothèse 5.1.8 les trois sources sont indépendantes

Le modèle log-linéaire d'indépendance est le suivant

$$\begin{pmatrix} \ln \mu_{100} \\ \ln \mu_{010} \\ \ln \mu_{001} \\ \ln \mu_{110} \\ \ln \mu_{011} \\ \ln \mu_{101} \\ \ln \mu_{111} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix}$$

Le modèle sera noté D, C, E.

Les résultats obtenus sont données dans le tableau suivant

Tableau 5.2 : Analyse log-linéaire et modèles, estimation du nombre de cas d'infection à méningocoque

numéro du modèle	*modèles*	*l'intercept* $\widehat{\beta}_0$	$\widehat{\mu}_{000} = \exp\left(\widehat{\beta}_0\right)$	N	G^2	*ddl*
1	DC, CE, DE	3.9257	51	466	0	0
2	DE, CE	3.1017	22	437	5.38	1
3	DC, DE	2.6543	14	429	18.6	1
4	**DC, CE**	4.0043	**55**	**470**	**0.11**	**1**
5	DC, E	2.9124	18	433	21.03	2
6	DE, C	2.5860	13	428	18.70	2
7	D, CE	3.2454	26	441	6.37	2
8	D, C, E	2.7755	16	431	21.50	3

N : nombre de cas total estimés, G^2 : statistique du test du rapport de vraisemblance, *ddl* : degré de liberté

L'analyse pas à pas descendante à partir du modèle saturé prenant en compte toutes les interactions d'ordre 2 entre les 3 sources montrait qu'en dehors du modèle saturé, un seul modèle présentait une bonne adéquation avec les données. Ce modèle incluait les deux dépendances observées précédemment : entre les sources DO et CNR et entre CNR et EPIBAC (modèle numéro 4) (**tableau 5.2**). L'estimation du nombre total de cas d'IM confirmés par isolement de NM dans le sang ou le LCR en France, déclarés par au moins une des trois sources ou non déclarés était de 470 cas avec ce modèle.

Remarque 5.4.2 *: La modélisation log-linéaire à été réalisée à l'aide du logiciel MATLAB 7.0 avec la commande glmfit dédiée à l'analyse des modèles linéaires généralisés*

Conclusion et perspéctives

Dans ce travail nous avons étudie les modèles de capture-recapture uni-état à temps discret, dont le but est d'estimer la taille d'une population fermée, difficile à dénombrer. Pour ce faire, nous avons adopté deux approches : la première est l'approche bayésienne, à travers laquelle nous avons considéré un modèle de capture-recapture homogène où les occasions de capture sont indépendantes et la probabilité de capture change d'une occasion à une autre. D'après une étude de simulation, nou avons constaté que l'estimateur de Bayes pour la taille de la population obtenu à la dernière occasion de capture est préférable à ceux obtenus aux occasions de capture précédentes.
Dans une deuxième approche basée sur les modèles log-linéaires, nous avons traité l'hétérogénéité due à la dépendance entre occasions de capture, ou le comportement des individus après leur capture initiale. Nous avons constaté que la dépendance entre occasions de capture peut être incorporée en ajoutant les termes d'intéractions. Nous avons déduit que : Si le nombre d'occasions de capture augmente alors le nombre de modèles log-linéaires augmente rapidement, cela engendre un problème pour la séléction du meilleur modèle.

En outre les modèles de capture-recapture uni-état, à temps discret ou continu sont des outils efficaces servant à traiter les populations fermée non observées dans leur totalité.

Un autre axe de recherche important s'agissant des modèles de capture-recapture multi-état à temps discret ou continu qui sont destinés aux populations ouvertes dont les paramètres d'intérêt sont le taux de survie, le taux de reproduction et la probabilité de transition entre les sites.

Bibliographie

[1] Agresti, A. (2007). *Categorical Data Analysis, Second Edition.* John Willey and Sons, New York.

[2] Agresti, A. (1994) Simple capture-recapture models permitting unequal catchability and variable sampling effort. *Biomertics* **50**, 494-500.

[3] Aknouche (2009). Introduction à la Statistique Inférentielle, *cours non publié.*

[4] Atkinson, R. (2000) . *Robust Diagnostic Regression Analysis.* Springer-Verlag New York Berlin Heidelberg.

[5] Baker.S.G. (1990) . A simple EM algorithm for capture-recapture data with categorical covariates. *Biometrics,* **46**, 1193- 1200.

[6] Bentarzi. (2009). Statistique Inférentielle, *cours non publié.*

[7] Besse P. (2003). Pratique de la modélisation statistique. Publication du laboratoire de statistique et probabilités, *UMR CNRS C 5583. Université Paul Sabatier- 31062 - Toulouse cedex 4.*

[8] Castledine, B.J. (1981). A bayesian analysis of multiple recapture sampling for a closed population. *Biometrika* **67**.1, 197-210.

[9] Chaieb, ML. (1999). Utilisation des modèles linéaires généralisés pour estimer la taille d'une population animale fermée.Thèse de maître ès sciences. Faculté des Sciences et de Génie, U.Laval.Québec.

[10] Chao, A, Tsay, P.K, Sheng, H, Siang, L, Wen, Y et Chao, D. (2001). The applications of capture-recapture models to epidemiological data.*Statist-Med*, 3123-3157.

[11] Chevallier E. (2001). Estimations locales de la prévalence de l'usage d'opiacés et cocaïne en France. Observatoire francais des drogues et des toxicomanies, 105, rue La Fayette 75010 Paris *(OFDT).*

[12] Cormack ,R.M .(1989). Log-linear models for capture-recapture. *Biometrics*, **45**, 395-413.

[13] Cowan et Malek. (1986) Capture-racapture models when both sources have clustered observations. *J. Am. Statist. Assoc.* 81, 347-353

[14] El Khorazaty and Coll (1977). Estimating the total number of events with data from multiple record systems. *Int Statist. Rev.* 48, 129-57.

[15] Fang, K. Zhang,Y. (1980) . *Generalized Multivariate Analysis.* Springer Verlag, New york.

[16] fienberg.A. (1972) . The multiple recapture censusfor closed populaions and incomplete 2^k contingency tables. *Biometrika*, **80**, 27-38.

[17] Gauss, M. Codeiro, Peter McCullagh. (1991) . Bias Correction in Generalised Linear Models. *Journal of the Royal Statistical Society. Series B (Methodological),* Volume **53**, Issue **3**, 629-643.

[18] Hamrat. M & Ouafi. R.(2007) . Modèles de capture-recapture multiples en épidémiologie. *Journés de statistique théorique et Appliquée, Biskra, 21 et 22 Avril.*

[19] Hamrat. M & Ouafi. R. (2008) . Modélisation log-linéaire des données de capture-recapture. *Journées de statistique, Modélisation et Application, USTHB, Alger.*

[20] Kenneth, H. Pollock and Mark C, Otto. (1983). Robust estimation of population size in closed animal populations from capture- recapture experiments. *Biometrics* ,**39**, 1035-1049.

[21] McCullagh.P & Nelder J.A.(1989) . *Generalized Linear Models, Second edition.* Chapman & Hall. London.

[22] Otis, D. L, Burnham, K. P, White, G. C and Anderson, D. R. (1978). Statistical inference from capture data on closed animal populations. *Wildlife Monographs* **62.**

[23] Perrocheau, A. (2001).Evaluation de la surveillance des infections à méningocoques en France 1996 par la méthode de capture-recapture.

[24] Rivest L.P, Lévesque T. (2001). Improved log-linear model estimators of abundance in capture-recapture experiments. *The Canadian Journal of Statistics*, **29**, 555-572.

[25] Rivest L.P, Daigle. G. (2004) . Log-linear Models for the Robust Design in Mark-Recapture Experiments. *Biometrics* **60**, 100-107.

[26] Robert, C. (1992). L'analyse statistique bayésienne. *Economica,* **49***, rue Héricart, 75015 Paris.*

[27] Schnabel Zoe, E. (1938). The estimation of the total fish population of a lake. *Amer Math Mon*, **45**, 348-352.

[28] Smith P.J (1988). Bayesian methods for multiple capture-recapture surveys. *Biometrics*, **44**, 1177-1189.

[29] Wang X, He CZ, Sun D. (2007). Bayesian population estimation for small sample capture-recapture data using noninformative priors. *Journal of Statistical Planing and Inference*, **137**, 1099-1118.

[30] Wittes, J.T (1974) . Application of a multinomial capture-recapture model of epidemiological data. J.Am. Statist. Assoc. 69, 93-7.

[31] Wolter, K.M (1986) . Some coverage error models for census data. J.Am. Statist. Assoc. 81, 338-46.

[32] Young H.C (2006). Estimating the number of undetected software errors via the correlated capture-recapture model. *European Journal of operational Research,* **175**, 1180-1192.