Techniques d'enquête, décembre 1994
Vol. 20, n° 2, pp. 121–128
Statistique Canada

Techniques de saisie-ressaisie pour l'estimation de la taille de la population et de totaux de population lorsqu'on ne dispose pas d'une base de sondage complète

K.H. POLLOCK, S.C. TURNER et C.A. BROWN[1]

RÉSUMÉ

Nous présentons une solution formelle, fondée sur un modèle, au problème de l'estimation de la taille de bases de sondage (listes), au moyen d'un échantillonnage avec saisie-ressaisie; cette solution a été largement utilisée pour estimer des populations d'animaux et pour le redressement des chiffres du recensement des États-Unis. Lorsqu'on dispose de deux listes incomplètes, il est facile d'estimer la taille totale de la base, en utilisant l'estimateur de Lincoln-Petersen. Il s'agit d'un estimateur fondé sur un modèle, qui repose sur l'hypothèse clé que les deux listes sont indépendantes. Dès qu'on dispose d'une estimation de la taille de la population (base de sondage), on peut estimer un total de population pour une caractéristique quelconque si cette caractéristique est mesurée auprès d'un échantillon de la population. Une analyse des propriétés de l'estimateur de Lincoln-Petersen est présentée. Nous examinons un exemple dans lequel les établissements sont des bateaux de pêche en activité au large de la côte Atlantique des États-Unis. L'estimation de la taille de la base de sondage, puis de totaux de population, à l'aide d'un modèle de saisie-ressaisie devrait avoir de vastes applications dans les enquêtes auprès des établissements, vu son côté pratique et les économies qu'elle procure. Toutefois, la possibilité que des biais soient introduits par suite du non-respect de certaines hypothèses doit être prise en considération.

MOTS CLÉS: Bases incomplètes; échantillonnage par saisie-ressaisie; enquêtes sur la pêche sportive; enquêtes téléphoniques; enquêtes aux points d'accès.

1. INTRODUCTION

Dans la théorie classique de l'échantillonnage, on suppose qu'il existe une base de sondage complète, c'est-à-dire qu'il existe, au moins théoriquement, une liste complète des unités de la population. On peut alors extraire un échantillon probabiliste de la population. Les estimateurs de paramètres de la population, par exemple une moyenne ou un total, ont alors des propriétés connues et peuvent être facilement étudiés théoriquement et numériquement. Les ouvrages sur la théorie de l'échantillonnage, comme celui de Cochran (1978), étudient cette situation et énoncent les propriétés des estimateurs pour des plans de sondage courants comme l'échantillonnage aléatoire simple, l'échantillonnage aléatoire stratifié et l'échantillonnage à plusieurs degrés (par grappes).

En pratique, dans les enquêtes auprès des établissements ou des entreprises, il se peut qu'on ne dispose pas d'une base complète. Ainsi, les listes d'établissements dont disposent les associations professionnelles ou les organismes gouvernementaux sont souvent incomplètes. Une façon d'aborder ce problème consiste à utiliser la méthode à bases multiples initialement élaborée par Hartley (1962, 1974). C'est la méthode employée, notamment, dans les enquêtes sur les établissements agricoles effectuées par le National Agricultural Statistics Service (USDA) (Vogel et Kott 1993). Ces enquêtes utilisent une liste incomplète d'établissements agricoles ainsi qu'une base aréolaire dans laquelle sont dénombrés tous les établissements appartenant à une unité d'échantillonnage. La liste, donc, est incomplète, tandis que la base aréolaire est théoriquement complète. (On dispose d'une liste de toutes les unités aréolaires et, dans chacune de ces unités, tous les établissements agricoles peuvent théoriquement être dénombrés.)

Il y a toutefois des situations où le recours à une base aréolaire peut, pour des raisons pratiques, n'être pas possible. Le chercheur peut n'avoir à sa disposition que des listes incomplètes d'établissements. En général, dans ce cas, on fusionne toutes les listes incomplètes et l'on ne tient pas compte du fait que l'ensemble puisse demeurer incomplet. Selon l'importance des carences de la liste globale, les estimations de la taille de la population et des totaux de population peuvent alors comporter un biais négatif prononcé.

Nous présentons plus loin une solution formelle à ce problème; il s'agit d'une solution fondée sur un modèle, qui fait appel à l'échantillonnage avec saisie-ressaisie. Les modèles d'échantillonnage avec saisie-ressaisie sont largement utilisés dans l'échantillonnage de populations animales (Seber 1982) et pour redresser les chiffres du recensement des États-Unis afin d'éliminer le sous-dénombrement (Feinberg 1992). Dans le cas le plus simple, celui où l'on possède deux listes incomplètes, nous considérons les unités "marquées" comme celles qui appartiennent aux deux listes et les unités non marquées comme celles qui n'appartiennent pas aux deux listes. Il est facile d'estimer la taille totale de la base au moyen de l'estimateur de Lincoln-Petersen (Seber 1982, p. 59). Il s'agit d'un estimateur

[1] K.H. Pollock, North Carolina State University, Raleigh, NC 27695; S.C. Turner et C.A. Brown, National Marine Fisheries Service, Miami, FL 33149, U.S.A.

fondé sur un modèle, qui repose sur l'hypothèse clé de l'indépendance des deux listes. Dès qu'on dispose d'une estimation de la taille de la population, on peut estimer un total de population pour une caractéristique quelconque si cette caractéristique est mesurée auprès d'un échantillon de la population.

L'estimateur habituel d'un total de population pour l'échantillonnage aléatoire simple sans remise est

$$\hat{Y} = N\bar{y}, \tag{1.1}$$

où N est connu et $\bar{y}$ est la moyenne de l'échantillon; voir par exemple Cochran (1978, p. 21). La variance de $\hat{Y}$ est donnée par

$$\text{Var}(\hat{Y}) = N^2\text{Var}(\bar{y}), \tag{1.2}$$

où

$$\text{Var}(\bar{y}) = \frac{S^2}{n}\left(\frac{N-n}{N}\right),$$

S^2 est la variance de la population et $(N - n/N)$ est appelé facteur de correction pour population finie. L'estimateur (1.1) est également un estimateur sans biais du total de population.

Dans le cas qui nous occupe, l'estimateur est

$$\hat{Y} = \hat{N}\bar{y}, \tag{1.3}$$

où $\hat{N}$ est obtenu par la méthode de saisie-ressaisie.

Il est clair que les propriétés de l'estimateur (1.3) sont plus difficiles à évaluer, car $\hat{N}$ et $\bar{y}$ sont deux variables aléatoires; ce n'est pas le cas de l'estimateur (1.1), où N est une valeur connue. Dans notre cas, on a une variance estimée de $\hat{Y}$ donnée par

$$\widehat{\text{Var}}(\hat{Y}) = (\hat{N})^2\,\widehat{\text{Var}}(\bar{y}) + (\bar{y})^2\,\widehat{\text{Var}}(\hat{N}) + \widehat{\text{Var}}(\bar{y})\,\widehat{\text{Var}}(\hat{N}), \tag{1.4}$$

en supposant que $\bar{y}$ et $\hat{N}$ sont indépendants et en utilisant un résultat exact dû à Goodman (1960). L'estimateur (1.3) est un estimateur sans biais seulement si $\hat{N}$ et $\bar{y}$ sont des estimateurs sans biais respectivement de la taille et de la moyenne de la population, ce qui n'est habituellement pas le cas en pratique. Nous examinons l'estimateur (1.3) dans l'exemple de l'enquête sur la grande pêche pélagique présenté à la section 3.

Le reste du présent article est structuré comme suit. À la section 2, nous passons en revue la littérature relative à la saisie-ressaisie, pour donner un aperçu des types de modèles disponibles. À la section 3, nous présentons un exemple d'enquête par sondage portant sur les bateaux de pêche. (Nous considérons un bateau comme un établissement commercial.) Bien que cet exemple comporte certaines caractéristiques uniques, nous croyons qu'il est représentatif, sous plusieurs aspects, d'autres enquêtes effectuées auprès des établissements. Dans la dernière section, nous faisons le bilan des forces et des faiblesses de la méthode de saisie-ressaisie pour l'estimation de la taille de la base de sondage dans les enquêtes auprès des établissements. Plusieurs des idées présentées ici exigeront d'être étayées par de futures recherches.

2. BREF SURVOL DES MODÈLES DE SAISIE-RESSAISIE

Il va de soi qu'une couverture complète des publications relatives à la saisie-ressaisie se situe hors du cadre du présent article. Pour un compte rendu plus détaillé, nous recommandons les travaux de Seber (1982), White et coll. (1982), Pollock et coll. (1990) et Pollock (1991). L'aperçu général donné par Pollock (1991) est une bonne introduction à la littérature du domaine; nous en suivons ici très étroitement la démarche. Les autres sources sont des ouvrages et des monographies qui s'adressent au lecteur qui veut pousser plus loin sa recherche et a plus de temps à sa disposition.

Nous allons examiner brièvement le modèle de Lincoln-Petersen pour deux échantillons, puis des modèles plus généraux à population fermée et à population ouverte pour plus de deux échantillons et, enfin, une méthode combinant des modèles à population fermée et à population ouverte dans un même plan d'échantillonnage. Pollock et coll. (1990, p. 9) présentent un organigramme donnant une vue d'ensemble des modèles et des liens qui existent entre eux.

2.1 Le modèle de Lincoln-Petersen

Il s'agit du modèle de saisie-ressaisie le plus ancien, le plus simple et le mieux connu. Le premier à l'utiliser fut Laplace, pour l'estimation de la population de la France. Petersen, au tournant du siècle, a été le premier à appliquer ce modèle au domaine de la pêche. Seber (1982, chapitre 3) en donne les détails dans une excellente analyse.

Dans son application originale au domaine de la pêche, la méthode peut être décrite de la façon suivante. Un échantillon de M poissons est capturé, et les poissons sont marqués et relâchés. Plus tard, un deuxième échantillon de n poissons est capturé, dont m sont marqués. On obtient de façon intuitive un estimateur en supposant égales les proportions de poissons marqués dans l'échantillon et dans la population,

$$m/n = M/N, \tag{2.1}$$

ce qui donne

$$\hat{N} = Mn/m. \tag{2.2}$$

Un estimateur modifié, atténuant le biais dans le cas des petits échantillons, est donné par Chapman (1951):

$$\hat{N}_c = \left[(M+1)(n+1)/(m+1)\right] - 1. \tag{2.3}$$

L'expression suivante donne une estimation de la variance de $\hat{N}_c$:

$$\widehat{\text{Var}}(\hat{N}_c) = \frac{(M+1)(n+1)(M-m)(n-m)}{(m+1)^2(m+2)}. \quad (2.4)$$

Voir par exemple Seber (1982, p. 60).

Les hypothèses fondamentales de ce modèle sont les suivantes:

(a) la population est complètement fermée aux ajouts et aux suppressions;

(b) tous les poissons ont une probabilité égale d'être capturés dans chaque échantillon;

(c) les marques restent en place et sont détectées.

L'hypothèse relative à la population fermée peut être adoucie, mais même pour une population complètement ouverte, à laquelle cet estimateur ne s'applique pas, une modification de l'estimateur de Lincoln-Petersen est utilisée. L'hypothèse d'une capturabilité égale pose des problèmes dans la plupart des applications. Il pourrait bien y avoir une variabilité inhérente (hétérogénéité) dans les probabilités de capture des animaux individuels en raison de l'âge, du sexe ou d'autres facteurs. Il se peut aussi qu'une réaction à la capture initiale se manifeste. Dans la section qui suit, nous examinons des modèles à population fermée comportant plus de deux échantillons, dans lesquels les probabilités de capture des animaux peuvent être l'objet d'une variation dans le temps ainsi que d'une hétérogénéité et de réactions à la capture. La perte et la non-détection des marques peuvent être importantes. Une façon d'estimer la perte des marques consiste à utiliser deux marques (Seber 1982, p. 94).

2.2 Modèles à population fermée

Les modèles à population fermée reposent sur l'hypothèse qu'aucun mouvement de la population attribuable à des naissances, à des décès ou à une migration ne survient entre les périodes d'échantillonnage. Ces modèles, par conséquent, sont généralement appliqués à des études portant sur de courtes périodes (p. ex. on pose des filets chaque jour pendant cinq jours consécutifs). Les historiques de capture des animaux individuels sont les données qui servent à produire les estimations selon ces modèles. Parmi les premiers travaux dans ce domaine, mentionnons ceux de Schnabel (1938) et de Darroch (1958), qui ont examiné des modèles fondés sur une capturabilité égale des animaux dans chaque échantillon.

On possède aujourd'hui un ensemble de modèles dans lesquels les probabilités de capture diffèrent à cause de l'hétérogénéité (h), d'une réaction à la capture (b), d'une variation dans le temps (t) (c.-à-d. que la probabilité de capture au temps i diffère de la probabilité de capture au temps j) et de toutes les combinaisons de deux ou de trois de ces facteurs. Les huit modèles [$M(o)$, $M(h)$, $M(b)$, $M(bh)$, $M(t)$, $M(th)$, $M(tb)$, $M(thb)$] ont d'abord été examinés comme un tout par Pollock (1974), tandis que Otis et coll. (1978), White et coll. (1982) et Pollock et Otto (1983) les ont développés plus à fond. Otis et coll. (1978) ont fourni un programme informatique détaillé, appelé CAPTURE, pouvant être utilisé de concert avec leur monographie. Une version mise à jour permet d'obtenir des estimations pour sept des huit modèles et propose aux biologistes une méthode de sélection facilitant le choix d'un modèle. Cette méthode de sélection du modèle se fonde sur divers tests de qualité de l'ajustement. Récemment, Menkins et Anderson (1988) ont fait valoir que la méthode de sélection du modèle n'est pas très bonne pour les petites populations, à moins que les probabilités de capture soient démesurément élevées.

2.3 Modèles à population ouverte

Dans beaucoup d'études de saisie-ressaisie, on ne peut supposer que la population est fermée, c.-à-d. à l'abri d'ajouts ou de suppressions permanentes. Le modèle de base à population ouverte qui convient à cette situation est celui de Jolly-Seber (Jolly 1965; Seber 1965; Seber 1982, p. 196). Ce modèle permet l'estimation de la taille de la population à chaque échantillonnage, ainsi que l'estimation des taux de survie et du nombre de naissances entre les échantillonnages. La migration ne peut être séparée des naissances et des décès sans information additionnelle.

Le modèle de Jolly-Seber repose sur les hypothèses suivantes:

(a) pour tous les animaux présents dans la population au moment d'un échantillonnage particulier, la probabilité d'être capturés est la même;

(b) pour tous les animaux marqués qui sont présents dans la population immédiatement après un échantillonnage, la probabilité de survivre jusqu'à l'échantillonnage suivant est la même;

(c) les marques restent en place et sont détectées;

(d) toute l'émigration est permanente;

(e) tous les échantillons sont instantanés, et chaque animal est relâché immédiatement après avoir été prélevé.

Selon le modèle de base de Lincoln-Petersen, décrit à la section 2.1, les hypothèses (a), (c) et (e) devaient être faites. Seuls les animaux marqués servent à estimer les taux de survie; il n'est donc pas nécessaire, à strictement parler, de supposer l'égalité des taux de survie des animaux marqués et des animaux non marqués. En pratique toutefois, le biologiste voudra que les estimations des taux de survie portent sur l'ensemble de la population. Le modèle de Jolly-Seber permet que certains animaux soient perdus au moment de la capture et ne soient donc pas retournés dans la population. Le modèle exige par ailleurs que toute émigration soit permanente. Si des animaux émigrent, puis reviennent dans la population, il en résulte ce qu'on appelle une émigration temporaire, qui contredit gravement les hypothèses et engendre un biais important dans les estimations de la taille de la population.

2.4 Combinaison de modèles à population fermée et à population ouverte

Pollock (1982), Pollock et coll. (1990) et Kendall (1992) examinent des méthodes d'échantillonnage qui permettent de réunir dans un même plan des modèles à population fermée et des modèles à population ouverte. Ces méthodes présentent un avantage du fait qu'on peut considérer que les probabilités de capture sont inégales, possibilité que n'offre pas le modèle classique de Jolly-Seber. Autre avantage, ces modèles permettent de tenir compte d'une émigration temporaire des animaux.

2.5 Applications des modèles de saisie-ressaisie

Naturellement, les modèles de saisie-ressaisie ont beaucoup été utilisés pour l'étude de la faune terrestre et des poissons. Toutefois, on peut maintenant observer plusieurs applications nouvelles, en dehors du domaine biologique. De nombreux auteurs ont appliqué les méthodes de saisie-ressaisie à l'estimation du sous-dénombrement du recensement. (Voir Feinberg (1992) pour une bibliographie complète.) Cowan, Breakey et Fischer (1986) s'en sont servi pour estimer le nombre de sans-abri dans une ville. Greene (1983) a utilisé la méthode pour estimer des paramètres démographiques propres à des populations criminelles. Wittes (1974) et Wittes, Colton et Sidel (1974) ont utilisé la saisie-ressaisie pour estimer le nombre de personnes atteintes de maladies à partir de listes d'hôpital et d'autres listes. L'échantillonnage de populations humaines difficiles à circonscrire, effectué à l'aide de l'échantillonnage par grappes, d'échantillons superposés et de l'échantillonnage par saisie-ressaisie, a été examiné par Sudman, Sirken et Cowan (1988).

3. MODÈLES DE SAISIE-RESSAISIE APPLIQUÉS À LA "LARGE PELAGIC SURVEY"

La "Large Pelagic Survey" est une enquête sur la pêche sportive réalisée par le National Marine Fisheries Service au moyen d'un plan d'enquête combinant une enquête téléphonique et une enquête aux points d'accès. Un échantillon de propriétaires de bateaux de pêche est constitué à partir d'une liste et l'on demande à ces propriétaires, par téléphone, de fournir des renseignements sur leur effort de pêche (nombre de sorties de pêche au cours d'une période donnée). Les données sur les prises par unité d'effort (prises par sortie) sont tirées d'un deuxième échantillon de propriétaires de bateaux interviewés aux points d'accès des bateaux, au retour des sorties de pêche. On combine les données des deux enquêtes pour estimer l'effort total et la prise totale d'espèces importantes comme le thon rouge.

Le fait que la liste des propriétaires de bateaux qui sert à l'enquête téléphonique soit très incomplète pose un problème sérieux dans cette enquête. Par conséquent, la théorie classique de l'échantillonnage, qui suppose une base complète de taille connue (N), ne convient pas et doit être modifiée. La méthode actuelle d'estimation de la taille de la base de sondage sous-jacente à la liste consiste à combiner deux listes (une liste téléphonique et une liste de bateaux interceptés à quai) et à utiliser le modèle de Lincoln-Petersen. On peut se demander si cette façon de procéder est la meilleure. Il serait peut-être possible, par exemple, de combiner plus de deux listes et ensuite d'utiliser les modèles à population fermée ou ouverte dont il a été question aux sections 2.2 et 2.3. Toutefois, nous laissons de côté ces questions pour le moment et nous commençons par examiner et évaluer la méthode actuelle, afin de montrer comment cette approche pourrait être employée utilement dans d'autres enquêtes auprès des établissements.

3.1 Le modèle de Lincoln-Petersen

3.1.1 Estimation de la taille de la base de sondage *(N)*

Selon la méthode actuelle, les bateaux "marqués" (M) sont ceux de la liste principale, établie en majeure partie d'après des interviews téléphoniques antérieures. L'échantillon de ressaisie est prélevé à quai, aux pompes à carburant, et l'on vérifie la liste complète des bateaux interceptés (n) afin de déterminer ceux qui sont "marqués" (m), (c.-à-d. ceux qui appartiennent à la liste principale). On obtient ensuite un estimateur de la taille de la base de sondage (N) en utilisant l'équation 2.3. Examinons maintenant les hypothèses du modèle et l'effet que la violation de ces dernières pourrait avoir sur le biais de l'estimateur de N.

Population fermée

Cette hypothèse n'est vraisemblablement pas respectée. Il se peut que des bateaux de pêche aient été inscrits sur la liste principale puis aient cessé leur activité de pêche (pertes), tandis que de nouveaux bateaux peuvent être entrés en activité une fois la période commencée (gains). Idéalement, il faudrait produire une estimation distincte de la taille de la base de sondage pour chaque période de deux semaines. L'avantage de l'estimateur fondé sur un modèle à population fermée de Lincoln-Petersen est qu'il est simple et pratique à utiliser. Le fait que la population ne soit pas fermée peut entraîner un biais positif ou négatif de l'estimateur.

On ne connaît pas pour le moment l'évolution probable de la taille de la flotte pendant la saison de pêche. Un plan de sondage à saisies-ressaisies multiples permettrait d'utiliser le modèle de Jolly-Seber pour estimer la taille de la flotte à chaque période. L'examen de ces estimateurs, ainsi que des estimateurs du taux de survie et du recrutement, nous permettra d'évaluer la validité de l'hypothèse d'une population fermée. Pour l'instant, nous devons nous limiter à des conjectures.

Capturabilité égale

Le non-respect de l'hypothèse de la capturabilité égale des individus peut être dû soit à l'hétérogénéité intrinsèque des probabilités de capture entre les individus, soit à une "réaction à la capture" qui ferait que pour les individus

marqués la probabilité de capture est plus élevée ou plus faible que pour les individus non marqués. Dans les deux cas, lorsque les unités des listes sont des bateaux de pêche, nous croyons qu'il peut exister une hétérogénéité des probabilités de saisie entre les bateaux. Si l'hétérogénéité vaut pour les deux échantillons, les individus saisis sur la première liste auront tendance à être ceux pour lesquels la probabilité de saisie est élevée et donc ceux qui le plus probablement figureront sur la deuxième liste. La proportion d'individus marqués dans le deuxième échantillon (liste) sera donc trop élevée et l'estimateur de N comportera un biais négatif. Notons que cet argument intuitif montre clairement que ce n'est pas l'hétérogénéité en soi qui constitue un problème, mais la corrélation positive des probabilités de saisie entre les deux échantillons. Une autre façon d'énoncer l'hypothèse des probabilités égales de saisie consiste à dire que les probabilités de saisie sont indépendantes entre les deux échantillons. Une méthode par laquelle on peut tenter de réaliser l'indépendance des probabilités de saisie entre les deux échantillons consiste à utiliser des plans de sondage complètement différents pour les deux échantillons. C'est pourquoi nous avons recommandé plus tôt qu'une liste soit établie à partir d'interviews téléphoniques et l'autre, à partir d'interviews réalisées à quai. Toutefois, nous croyons qu'il reste une hétérogénéité et une dépendance dans les probabilités de saisie: nous pensons que pour les bateaux qui participent très activement à la pêche la probabilité de figurer sur l'une ou l'autre liste est plus grande (liste d'interviews téléphoniques et liste de bateaux à quai). Cette hétérogénéité causera un biais négatif dans l'estimation de la taille de la base de sondage, mais nous n'avons aucune idée de l'importance de ce biais négatif. Seber (1982, p. 86) fait un examen plus complet de l'hétérogénéité et de l'indépendance des échantillons.

Perte ou non-détection des marques

La situation est ici un peu confuse. On pourrait croire, au premier abord, que la perte ou la non-détection de marques est impossible dans cette application. Cela suppose toutefois que les bateaux ont tous des noms distincts ou que, si certains ont des noms identiques, d'autres renseignements comme le nom du capitaine permettent de distinguer toutes les unités figurant sur les listes. S'il n'est pas certain que l'on puisse identifier tous les bateaux, il n'est pas sûr non plus qu'on puisse confirmer si un bateau marqué a été ressaisi ou non. Il se peut aussi que les agents introduisent des erreurs, rendant ainsi difficile l'appariement d'une ressaisie avec son identification dans la liste initiale. Un mode d'emploi normalisé est actuellement mis au point et documenté, en vue de réduire au minimum les erreurs de ce genre dans l'avenir.

3.1.2 Estimation de l'effort total et des prises totales

L'effort total (E) (c.-à-d. le nombre total de sorties de pêche au cours d'une période définie) est ainsi estimé:

$$\hat{E} = \hat{N}\bar{e}, \qquad (3.1)$$

où $\hat{N}$ est l'estimation de la taille de la base (taille de la flotte) et $\bar{e}$, l'effort de pêche moyen (c.-à-d. le nombre moyen de sorties de pêche) obtenu au moyen des interviews téléphoniques. L'évaluation des propriétés de cet estimateur est plus difficile que quand N est connu, car $\hat{N}$ et $\bar{e}$ sont deux variables aléatoires. Nous croyons que $\bar{e}$ est biaisé vers le haut, car les bateaux de pêche peu actifs sont moins susceptibles de figurer sur la liste. Malheureusement, nous ne pouvons dire que $\hat{N}$ sera toujours biaisé vers le haut ou vers le bas. Les trois violations des hypothèses examinées en 3.1.1 peuvent être importantes (population fermée, hétérogénéité, perte de marques), et l'on ne sait trop dans quel sens le biais global de $\hat{N}$ se manifesterait. La seule méthode possible est de faire une simulation avec plusieurs scénarios différents quant aux violations des hypothèses. En utilisant l'équation (1.4), on peut exprimer ainsi la variance estimée de $\hat{E}$:

$$\widehat{\mathrm{Var}}(\hat{E}) = (\hat{N})^2\,\widehat{\mathrm{Var}}(\bar{e}) + (\bar{e})^2\,\widehat{\mathrm{Var}}(\hat{N}) + \widehat{\mathrm{Var}}(\bar{e})\,\widehat{\mathrm{Var}}(\hat{N}). \qquad (3.2)$$

Les prises totales (C) sont estimées par $\hat{C} = \hat{E}\bar{c}$, où $\hat{E}$ est l'effort de pêche total estimé et $\bar{c}$, les prises moyennes par unité d'effort calculées d'après les interviews à quai. Les mêmes préoccupations que celles touchant l'équation (3.1) s'appliquent sans doute à cette équation et, encore une fois, une simulation pourrait être très utile.

3.1.3 Illustration de la méthode

Dans cette section, nous présentons les estimations de la taille de la base de sondage et les estimations de l'effort total pour la pêche au thon rouge en Virginie pendant une partie de l'année 1992. Ces estimations proviennent d'une enquête plus vaste couvrant la côte est des États-Unis, de la Caroline du Nord au Massachusetts. Les estimations sont distinctes pour les bateaux affrétés et les bateaux privés.

Estimations de la taille de la base de sondage

Des listes de bateaux privés et de bateaux affrétés ont été établies, principalement d'après les interviews téléphoniques des saisons précédentes. Au cours de la saison 1992, des bateaux "marqués" et "non marqués" ont été échantillonnés aux pompes à carburant avant ou après les sorties de pêche.

Pour les bateaux privés, la taille de la liste était de $M = 335$ bateaux avant la saison. Un échantillon de $n = 374$ bateaux ont été interceptés aux pompes à carburant et, de ces derniers, $m = 49$ étaient marqués. L'estimateur de Chapman est $\hat{N}_c = 2{,}519$, $\widehat{ET}(\hat{N}_c) = 303.08$ et $\widehat{ET}$ relative $= 0.12$.

Pour les bateaux affrétés, la taille de la liste était de $M = 47$ bateaux avant la saison. Un échantillon de $n = 31$ bateaux ont été interceptés aux pompes à carburant et, de ces derniers, $m = 13$ étaient marqués. L'estimateur de Chapman est $\hat{N}_c = 109$, avec $\widehat{ET}(\hat{N}_c) = 17.88$ et $\widehat{ET}$ relative $= 0.16$.

Estimations de l'effort total

L'effort total et les prises totales ont été estimés à une fréquence hebdomadaire. À titre d'illustration, nous présentons ici les calculs pour la semaine du 8 au 14 juin 1992 en ce qui a trait à l'effort total.

Effort total - bateaux privés

$\hat{N}_c = 2{,}519$ bateaux, $\widehat{\text{Var}}(\hat{N}_c) = 91{,}856.4706$, $\bar{e} = 0.15108$ sorties par interview, $\widehat{\text{Var}}(\bar{e}) = 0.001242$ et $\widehat{ET}(\bar{e}) = 0.0352$. En utilisant ces estimations, on obtient

$$\hat{E} = \hat{N}_c \times \bar{e} = 2{,}519 \times 0.15108 = 380.57 \text{ sorties},$$

$$\widehat{\text{Var}}(\hat{E}) = \widehat{\text{Var}}(\bar{e})(\hat{N}_c^2) + \widehat{\text{Var}}(\hat{N}_c)(\bar{e})^2 + \widehat{\text{Var}}(\hat{N}_c)\widehat{\text{Var}}(\bar{e}) = 10{,}091.6633, \text{ et}$$

$$\widehat{ET}(\hat{E}) = 100.45.$$

Il est utile de calculer également la variance de l'effort total, en supposant connue la taille de la base de sondage. Dans ce cas, on a $\widehat{\text{Var}}(\hat{E}) = 7{,}780.9384$ avec $\widehat{ET}(\hat{E}) = 88.77$; autrement dit, 89% de l'erreur-type de l'estimation de l'effort total est attribuable à la variation de l'effort moyen, tandis que seulement 11% en est attribuable à l'estimation de la taille de la base.

Effort total - bateaux affrétés

Pour les bateaux affrétés, $\hat{E} = 59.95$ sorties avec $\widehat{\text{Var}}(\hat{E}) = 512.5100$ et $\widehat{ET}(\hat{E}) = 22.64$.

La variance de l'estimation de l'effort total, si l'on suppose connue la taille de la base de sondage, est $\widehat{\text{Var}}(\hat{E}) = 404.8926$ avec $\widehat{ET}(\hat{E}) = 20.12$. Encore ici, 89% de l'erreur-type de l'estimation de l'effort total est attribuable à la variation de l'effort moyen, tandis que seulement 11 % en est attribuable à l'estimation de la taille de la base.

3.2 Utilisation de plus de deux listes

À la section 2, nous avons indiqué qu'il existe beaucoup plus de possibilités de modélisation si l'on dispose de plusieurs listes (plus de deux). Nous allons examiner ici des modèles à population fermée et à population ouverte, dans le contexte le plus général. Nous entrevoyons le plan d'échantillonnage de la façon suivante. Avant le début de la saison de pêche, un échantillon préliminaire serait prélevé pour établir une liste (interviews téléphoniques ou bateaux interceptés à quai). Au cours de chaque période (par exemple de deux semaines), on établirait une liste additionnelle au moyen d'une enquête téléphonique ou d'interviews réalisées à quai. Chaque bateau aurait alors un historique de saisie qui indiquerait sur quelles listes il figure. (Par exemple, pour cinq périodes, un historique de saisie de 1 1 1 0 1 indiquerait que le bateau figure sur toutes les listes sauf celle de la quatrième période.)

La structure de l'échantillon et de la population serait alors celle indiquée au tableau 1. La première question à examiner est de savoir si nous devons utiliser des modèles à population fermée ou à population ouverte. La façon de procéder la plus logique consiste à ajuster le modèle à population ouverte de Jolly-Seber d'abord, puis à l'utiliser pour évaluer l'hypothèse d'une population fermée.

Tableau 1

Structure de la population selon un modèle à population ouverte*

Période	Liste préalable	Listes établies pendant la saison (p. ex. toutes les deux semaines)					
	0	1	2	3	. . .		k
Tailles de la population marquée	M_0	M_1	M_2	M_3	. . .		M_k
Tailles de la population totale	N_0	N_1	N_2	N_3	. . .		N_k

* Les tailles de la population marquée et de la population totale sont indiquées pour l'ensemble de l'étude.

3.2.1 Modèles à population ouverte

Dans le modèle de Jolly-Seber déjà examiné à la section 2.3, un certain nombre de paramètres sont identifiables (tableau 2). Notons qu'il est possible d'estimer le nombre de bateaux de pêche de la flotte à toutes les périodes de la saison, sauf la dernière (c.-à-d. que $\hat{N}_k$ ne peut être estimé). Cette application du modèle, qui comporte l'utilisation d'une liste préalable, est avantageuse du fait qu'elle écarte toute inquiétude quant à la possibilité que la liste préalable soit périmée, car le modèle permet les ajouts et les suppressions avant le début de la saison. Le modèle de Jolly-Seber est toutefois désavantageux en raison de sa complexité plus grande. Pour chaque période, une taille de la base de sondage est établie et les paramètres du taux de survie et du recrutement doivent être estimés. Parfois, les estimations de ces paramètres ont une faible précision, à moins que les échantillons ne soient de grande taille. Le modèle de Jolly-Seber a aussi l'inconvénient d'exiger l'hypothèse d'une probabilité égale de saisie.

Tableau 2

Structure du modèle à population ouverte de Jolly-Seber*

Période	Valeurs préalables	Valeurs établies pendant la saison						
	0	1	2	3	.	.	k-1	k
Population marquée	$(M_0 \equiv 0)$	$\hat{M}_1$	$\hat{M}_2$	$\hat{M}_3$	.	.	$\hat{M}_{k-1}$	-
Population totale	-	$\hat{N}_1$	$\hat{N}_2$	$\hat{N}_3$	.	.	$\hat{N}_{k-1}$	-
Taux de survie	$\hat{p}_0$	$\hat{p}_1$	$\hat{p}_2$		.	$\hat{p}_{k-2}$		-
Recrutement		$\hat{B}_1$	$\hat{B}_2$		.	$\hat{B}_{k-2}$		

* Les estimateurs des paramètres identifiables sont indiqués pour la taille de la population marquée, la taille de la population totale, le taux de survie et le recrutement.

Une autre question importante qui se pose au sujet du modèle de Jolly-Seber a trait à ce qu'on appelle l'"émigration temporaire". Un bateau de pêche peut devenir inactif

dans certaines périodes, puis reprendre son activité. Le modèle de Jolly-Seber fait l'hypothèse que les bateaux de pêche qui cessent leur activité ne reviennent pas. Cette question exige un examen plus approfondi. L'utilisation du plan robuste (c.-à-d. d'une combinaison de modèles à population fermée et de modèles à population ouverte) permet de tenir compte de l'émigration temporaire. Un tel plan exigerait l'établissement de deux listes dans un court intervalle à chaque période.

3.2.2 Modèles à population fermée

Si les estimations du modèle de Jolly-Seber des "taux de survie" et du "recrutement" laissaient croire que la population est fermée (c.-à-d. que N demeure constant), les modèles généraux à population fermée présentés à la section 2.2 pourraient s'appliquer. Les avantages sont une précision accrue de $\hat{N}$ attribuable à l'utilisation d'un plus grand nombre de listes, ainsi qu'une robustesse accrue de $\hat{N}$ à l'égard des probabilités inégales de saisie. L'inconvénient principal est un accroissement de la complexité.

4. ANALYSE

4.1 Méthodes à employer quand les listes sont incomplètes

(i) Compléter la liste

L'avantage est que le chercheur dispose d'une base complète et n'a pas besoin de généraliser les résultats en fonction d'une taille estimée de la base de sondage. L'inconvénient est le coût et, parfois, l'impossibilité de compléter la liste.

(ii) Utiliser une base aréolaire

Cette solution est avantageuse du fait qu'on doit seulement dénombrer les établissements des régions qui servent à former l'échantillon. Elle pourrait en revanche se révéler inefficace si les entreprises étaient dispersées à l'intérieur de chaque grande région.

(iii) Utiliser une liste et une base aréolaire (méthode à bases multiples)

Les avantages sont évidemment une précision accrue et une couverture de tous les établissements. Cette méthode peut par contre se révéler coûteuse ou encore être impossible à mettre en oeuvre.

(iv) Utiliser la saisie-ressaisie pour l'estimation de la taille de la base de sondage

Cette méthode pratique a l'avantage d'être moins coûteuse que les trois précédentes. Son inconvénient est l'introduction possible d'un biais si les hypothèses de la méthode de saisie-ressaisie ne sont pas respectées et s'il faut inclure une variation attribuable à l'estimation de la taille de la base de sondage dans les estimations de la variance des estimateurs de totaux de population.

4.2 Estimation de la taille de la base par la méthode de saisie-ressaisie

Dans la présente section, nous examinons les hypothèses des modèles, la précision des estimations, l'estimation de totaux de population et les problèmes spéciaux propres aux plans d'échantillonnage plus complexes dans les cas où la méthode de saisie-ressaisie est appliquée à l'estimation de la taille de la base de sondage.

Hypothèses des modèles

(i) Population fermée

La taille de la base de sondage peut-elle être considérée comme constante, de telle sorte qu'on puisse utiliser les modèles à population fermée? La réponse variera selon que l'enquête est seulement un cliché pris à un moment précis ou une série d'observations s'étalant sur une certaine période. Elle dépendra aussi de la rapidité avec laquelle les établissements ferment leurs portes ou de la rapidité d'apparition de nouveaux établissements. Nous croyons que c'est le type d'établissement étudié qui déterminera s'il y a lieu d'utiliser des modèles à population fermée ou des modèles à population ouverte.

Il y a aussi la question de l'émigration temporaire, c.-à-d. de la sortie de la base de certains établissements, qui y reviennent ensuite. Il s'agissait d'un problème possible dans l'exemple des bateaux de pêche, car certains bateaux cessent leur activité, puis redeviennent actifs. Le problème pourrait se poser dans d'autres enquêtes auprès des établissements, si les entreprises arrêtent et reprennent leur activité fréquemment et gardent le même nom quand elles redeviennent actives.

(ii) "Capturabilité inégale" et indépendance des listes

Comme nous l'avons dit plus haut, les listes utilisées doivent idéalement être indépendantes, de façon que les estimations de la taille de la base de sondage ne soient pas biaisées. En pratique, il peut se révéler difficile de trouver deux listes ou plus qui soient indépendantes.

(iii) Perte de marques - Identification unique des établissements

Les noms des établissements doivent être uniques et ne prêter à aucune confusion, sinon certains appariements entre les listes pourraient être omis ou erronés. Ce problème se posait dans l'exemple des bateaux de pêche, les premières années. À notre avis, il ne devrait pas s'agir d'un problème majeur dans la plupart des enquêtes auprès des établissements.

Précision des estimations

Les listes utilisées doivent être de taille suffisante pour permettre une précision adéquate de l'estimation ($\hat{N}$) de la taille de la base de sondage. Seber (1982, p. 96) examine l'estimation de Lincoln-Petersen en détail et présente des graphiques des tailles d'échantillon requises pour différents niveaux de précision. Pollock et coll. (1990) présentent des données sur la taille des échantillons dans le cas des modèles à population ouverte.

Estimation de totaux de population

Une fois obtenue l'estimation de la taille de la base de sondage, il arrive souvent que cette estimation soit combinée à une moyenne d'échantillon pour donner une estimation d'un total de population ($\hat{Y} = \hat{N}\bar{y}$). L'estimation du total de population peut comporter un biais et une variance additionnelle du fait que $\hat{N}$ est estimé. L'estimation peut aussi être biaisée parce que $\bar{y}$ n'est pas fondé sur un échantillon aléatoire de la base complète.

Plans de sondage plus complexes

Dans le présent article, nous avons examiné l'estimation de la taille de la base de sondage dans un contexte d'échantillonnage aléatoire simple, au moyen de la méthode de saisie-ressaisie. D'autres questions se posent si des plans de sondage plus complexes sont utilisés. Par exemple, dans le cas des plans stratifiés, on peut se demander s'il y a lieu d'estimer la taille de la base dans chaque strate séparément, ou d'estimer la taille de la base totale pour ensuite répartir celle-ci entre les strates en supposant des probabilités égales de présence des différentes strates dans les listes incomplètes. Une question encore plus complexe a trait à la façon d'estimer la taille de la base dans les plans de sondage à plusieurs degrés. Il est clair que ces domaines devront faire l'objet de recherches additionnelles.

REMERCIEMENTS

Les auteurs tiennent à remercier le rédacteur en chef, le rédacteur associé et deux arbitres anonymes pour leurs commentaires utiles qui ont permis d'améliorer sensiblement le contenu de cet article.

BIBLIOGRAPHIE

CHAPMAN, D.G. (1951). Some properties of the hypergeometric distribution with application to zoological census. *University of California Publication in Statistics*, 1, 131-160.

COCHRAN, W.G. (1978). *Sampling Techniques* (3ième édition). New York: John Wiley and Sons.

COWAN, C.D., BREAKEY, W.R., et FISCHER, P.J. (1986). The methodology of counting the homeless. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 170-175.

DARROCH, J.N. (1958). The multiple-recapture census I: estimation of a closed population. *Biometrika*, 45, 343-359.

FIENBERG, S.E. (1992). Bibliography on capture-recapture modelling with application to census undercount adjustment. *Survey Methodology*, 18, 143-154.

GOODMAN, L.A. (1960). On the exact variance of products. *Journal of the American Statistical Association*, 55, 708-713.

GREENE, M.A. (1983). Estimating the size of the criminal population using an open population approach. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 8-13.

HARTLEY, H.O. (1962). Multiple Frame Surveys. *Proceedings of the Social Statistics Section, American Statistical Association*, 203-206.

HARTLEY, H.O. (1974). Multiple frame methodology and selected applications. *Sankhyā*, C, 36, 99-118.

JOLLY, G.M. (1965). Explicit estimates from capture-recapture data with both death and immigration-stochastic model. *Biometrika*, 52, 225-247.

KENDALL, W.L. (1992). Robust Design in Capture-Recapture Sampling: Modelling Approaches and Estimation Methods. Unpublished PhD. dissertation, North Carolina State University, Biomathematics Program.

MENKINS, G.E., et ANDERSON, S.H. (1988). Estimation of small mammal population size. *Ecology*, 69, 1952-1959.

OTIS, D.L., BURNHAM, K.P., WHITE, G.C., et ANDERSON, D.R. (1978). Statistical inference for capture data on closed animal populations. *Wildlife Monographs*, 62, 1-125.

POLLOCK, K.H. (1974). The Assumption of Equal Catchability of Animals in Tag-Recapture Experiments. Unpublished Ph.D. dissertation, Cornell University, Biometrics Unit.

POLLOCK, K.H. (1982). A capture-recapture design robust to unequal probability of capture. *Journal of Wildlife Management*, 46, 752-757.

POLLOCK, K.H., NICHOLS, J.D., HINES, J.E., et BROWNIE, C. (1990). Statistical inference for capture-recapture experiments. *Wildlife Monographs*, 107, 1-97.

POLLOCK, K.H., et OTTO, M.C. (1983). Robust estimation of population size in closed animal populations from capture-recapture experiments. *Biometrics*, 39, 1035-1049.

POLLOCK, K.H. (1991). Modelling, capture, recapture, and removal statistics for estimation of demographic parameters for fish and wildlife populations: past, present and future. *Journal of the American Statistical Association*, 86, 225-238.

SCHNABEL, Z.E. (1938). The estimation of the total fish population of a lake. *American Mathematical Monthly*, 45, 348-352.

SEBER, G.A.F. (1965). A note on the multiple recapture census. *Biometrika*, 52, 249-259.

SEBER, G.A.F. (1982). *The Estimation of Animal Abundance and Related Parameters* (2ième édition). New York: MacMillan.

SUDMAN, S., SIRKEN, M.G., et COWAN, C.D. (1988). Sampling rare and elusive populations. *Science*, 240, 991-995.

VOGEL, F.A., et KOTT, P. (1993). Multiple frame establishment surveys. *Proceedings, International Conference on Establishment Surveys.*

WHITE, G.C., ANDERSON, D.R., BURNHAM, K.P., et OTIS, D.L. (1982). *Capture-Recapture and Removal Methods for Sampling Closed Populations.* Los Alamos, NM: Los Alamos Laboratory.

WITTES, J.T. (1974). Applications of a multinomial capture-recapture model to epidemiological data. *Journal of the American Statistical Association*, 69, 79-93.

WITTES, J.T., COLTON, T., et SIDEL, V.W. (1974). Capture-recapture method for assessing the completeness of case ascertainment when using multiple information sources. *Journal of Chronic Diseases*, 27, 25-36.