Régression binaire par "capture-recapture"

Jean-Baptiste ANTENORD [1,2,3] & Etienne BILLETTE de VILLEMEUR [3,4]

[1] *Centre de Recherche en Gestion et en Economie du Développement (CREGED), Université Quisqueya, 218 avenue Jean Paul II, Haut de Turgeau, Port-au-Prince, Haïti.*

[2] *Centre de Techniques de Planification et d'Économie Appliquée (CTPEA), 19 Boulevard Harry Truman (Bicentenaire), Port-au-Prince, Haïti (W.I).*

[3] *Faculté des Sciences Économiques et Sociales, Université de Lille, Cité scientifique, 59655 Villeneuve d'Ascq Cedex, France.*
E-mail: jb.antenord@univ-lille.fr & etienne.de-villemeur@univ-lille.fr

[4] *Chaires Universitaires Toussaint Louverture, 2 bis rue Neptune, Delmas, Haïti.*

Résumé. Les outils théoriques développés dans le cadre des méthodes dites de "capture-recapture" (MCR) permettent d'élaborer des estimateurs statistiques corrigeant au moins partiellement les biais liés à la non-représentativité des échantillons. L'objectif de cette communication est de proposer, dans le cas où l'on dispose de deux échantillons indépendants mais pas nécessairement représentatifs, un estimateur de la probabilité (conditionnelle) qu'un individu de caractéristique x ait également la caractéristique y, où x et y sont deux variables binaires. Une formule pour évaluer la variance de cette estimateur est également offerte. L'estimateur naïf de cette même probabilité (celui obtenu quand les deux échantillons sont considérés n'en former qu'un seul, que l'on suppose représentatif) est généralement biaisé - y compris de manière asymptotique *i.e.* pour de grands échantillons. Cependant, ce n'est pas toujours le cas. Aussi un test statistique est proposé afin d'établir si l'on peut éviter d'opérer un redressement statistique à cause des problèmes de représentativité (et donc considérer l'estimateur naïf comme étant sans biais).

Mots-clés. Estimation par méthode de "capture-recapture", régression binaire, représentativité des échantillons.

Abstract. Theoretical tools developed in the context of the so-called "capture-recapture" methods (MCR) make it possible to develop statistical estimators that at least partially correct the biases related to sample non-representativeness. The objective of this communication is to offer an estimator for the (conditional) probability that an individual of characteristics x displays also characteristics y, where x and y are two binary variables, when two independent samples are available that may *not* be representative.

We also exhibit a formula as to compute the variance of this estimator. The naive estimator of this probability (as obtained when both samples are pooled and considered as single sample supposedly representative) is generally biased - including asymptotically *i.e.* for large samples. This is not always the case, though. We thus offer a statistical test to establish whether accounting for the representativeness issue may be dispensed with (so that it is possible to consider the naive estimator as unbiased).

Keywords. Estimation by capture-recapture methods, binary regression, sample representativeness.

1 Mise en contexte - et bref survol de la contribution

Dans de nombreux contextes, l'inexistence de bases de sondage empêche la réalisation d'enquêtes fiables. Cette situation s'observe tant pour des populations animales difficiles à joindre que pour certaines populations humaines, notamment dans les pays à appareillage statistique déficient. Les méthodes dites de "capture-recapture" dont la problématique est d'estimer la taille d'une population dans un contexte où il n'est pas possible de procéder à un recensement (Lavallée et Rivest, 2012), sont de plus en plus utilisées et permettent de corriger au moins partiellement les biais liés à la non-représentativité des échantillons issus des sondages. Ces méthodes fournissent des outils permettant d'estimer la taille des populations à partir de différentes listes de données (listes complètes ou incomplètes) jouant le rôle "d'échantillons". Différents cas d'application sont répertoriés dans la littérature (Chao et *al.*, 2001). De l'estimation du nombre de victimes dans une guerre (Manrique-Vallier et al., 2012) ou dans des conflits armés (Mitchell, 2014), à celle du nombre de personnes souffrant d'une pathologie donnée ou encore pour estimer le nombre d'unités non répertoriées lors d'un recensement (Zwane et Heijden, 2008), les méthodes de "capture-recapture" (MCR) ne cessent de prendre de l'ampleur.

L'objectif de cette communication est rendre compte du programme de recherche développé au Centre de Techniques de Planification et d'Économie Appliquée ($CTPEA$) visant à contribuer au développement théorique et à l'application des MCR pour effectuer des enquêtes en Haïti. Il s'agit de mettre au point des estimateurs statistiques permettant d'opérer dans un contexte où la représentativité des échantillons n'est pas établie. Plus précisément, nous montrons ici comment estimer la probabilité (conditionnelle) qu'un individu de caractéristique x ait également la caractéristique y - dans le cas où l'on dispose de deux échantillons *indépendants* mais *pas* nécessairement *représentatifs*.

Dans ce qui suit, nous présentons d'abord le cadre d'analyse de manière générale, avant de nous concentrer sur le cas particulier qui nous occupe, où l'on suppose que l'on dispose de *deux* échantillons non-nécessairement représentatifs mais *indépendants*. On établit alors, en suivant les MCR, un estimateur de la probabilité (conditionnelle) qu'un individu de caractéristique x aient également la caractéristique y, où x et y sont deux variables binaires. Une formule pour le calcul de la variance de cette estimateur

est également proposée. Nous établissons ensuite que l'estimateur naïf (tel qu'obtenu quand les deux échantillons sont considérés n'en former qu'un seul, que l'on suppose représentatif) est généralement biaisé - y compris de manière asymptotique. Cependant, ce n'est pas toujours le cas. Nous proposons donc un test statistique pour établir s'il est possible de s'affranchir du redressement statistique sous-jacent à notre estimateur.

2 Estimateurs statistiques à partir d'échantillons non-représentatifs

2.1 Le cadre d'analyse des méthodes de "capture-recapture"

On considère une population $\mathcal{U}$ définie par $\mathcal{U} = \{1, 2, ..., N\}$, où N désigne la taille de la population. Chaque individu de la population possède un vecteur de caractéristiques $z \in \mathcal{Z}$, où $\mathcal{Z}$ dénote l'ensemble *fini* des caractéristiques possibles. On note n^z le nombre total d'individus dotés du vecteur de caractéristiques z.

On suppose qu'on fait K observations partielles de cette population. Pour chaque individu $i \in \mathcal{U}$, on note $\omega^i = (\omega_1^i, \omega_2^i, ..., \omega_K^i)$ l'historique des "captures" de cet individu ($\omega_k^i = 1$ si l'individu apparait dans la liste k et $\omega_k^i = 0$ sinon). On peut alors noter respectivement par n_ω^z et μ_ω^z les fréquences observées et théoriques des individus de caractéristiques z avec l'historique de capture $\omega \in \Omega$, où Ω désigne l'ensemble de tous les historiques de capture. Par définition,

$$n^z = \sum_{\omega \in \Omega} n_\omega^z \qquad \text{et} \qquad N = \sum_{z \in \mathcal{Z}} n^z.$$

Les méthodes de "capture-recapture" visent à extraire l'information contenue dans les K observations partielles, pour en inférer les caractéristiques de la partie de la population qui n'a pu être observée.

2.2 Le cas de deux listes - ou échantillons - supposés indépendants

On considère dans ce qui suit le cas particulier de deux listes: $K = 2$. L'ensemble des historiques possibles de capture-recapture s'écrit donc:

$$\Omega = \{11; 10; 01; 00\}.$$

On suppose également que, pour chaque individu de caractéristique z, la probabilité qu'il soit présent ou non dans la première liste est *indépendante* de sa probabilité d'être présent ou non dans la deuxième. En d'autre termes, on suppose que les deux listes peuvent être considérées comme des échantillons indépendants.

Si l'on note r_1^z est la probabilité de présence du type $z \in \mathcal{Z}$ dans la liste 1 et r_2^z est sa probabilité de présence dans la liste 2, on a alors les relations suivantes:

	Présent dans la liste 2	Absent de la liste 2
Présent dans la liste 1	$n_{11}^z = n^z\,(r_1^z)\,(r_2^z)$	$n_{10}^z = n^z\,(r_1^z)\,(1-r_2^z)$
Absent de la liste 1	$n_{01}^z = n^z\,(1-r_1^z)\,(r_2^z)$	$n_{00}^z = n^z\,(1-r_1^z)\,(1-r_2^z)$.

Seuls n_{11}^z, n_{10}^z et n_{01}^z sont observés. L'effectif n_{00}^z, par définition, ne peut l'être. Cependant, sous l'hypothèse d'indépendance des deux listes – ou échantillons, il est possible d'inférer des trois effectifs observés, l'effectif non-observé n_{00}^z, l'effectif total n^z ainsi que les deux probabilités r_1^z et r_2^z.
Plus précisément, les effectifs n_{00}^z et n^z peuvent être estimés comme suit:

$$\widehat{n_{00}^z} = \frac{n_{01}^z n_{10}^z}{n_{11}^z}, \tag{1}$$

$$\widehat{n^z} = n_{01}^z + n_{10}^z + n_{11}^z + \widehat{n_{00}^z} = \frac{1}{n_{11}^z}\left(n_{01}^z + n_{11}^z\right)\left(n_{10}^z + n_{11}^z\right) \tag{2}$$

(Lavallée et Rivest, 2012). Par suite, les estimateurs de probabilité de capture peuvent s'écrire:

$$\widehat{r_1^z} = \frac{n_{11}^z + n_{10}^z}{\widehat{n^z}} = \frac{n_{11}^z}{n_{01}^z + n_{11}^z}, . \tag{3}$$

$$\widehat{r_2^z} = \frac{n_{11}^z + n_{01}^z}{\widehat{n^z}} = \frac{n_{11}^z}{n_{10}^z + n_{11}^z}. \tag{4}$$

Dans ce qui suit, nous allons nous prévaloir de cette approche pour estimer sans biais une probabilité conditionnelle, alors même qu'on ne dispose pas, au moins *apriori,* d'échantillon représentatif.

3 Régression binaire avec deux échantillons indépendants

3.1 Le modèle sous-jacent

On suppose que le vecteur de caractéristiques $z \in \mathcal{Z}$ se réduit à une paire de variables binaires. Autrement dit $z = yx$ avec $y \in Y = \{0,1\}$ et $x \in X = \{0,1\}$. Par suite $\mathcal{Z} = \{11, 10, 01, 00\}$.

Tout se passe comme si la population était générée par une loi multinomiale:
$(n^{00}, n^{01}, n^{10}, n^{11}) \sim \mu\left(N, q^{00}, q^{01}, q^{10}, q^{11}\right)$ avec

$$q^{11} = \frac{n^{11}}{N} = \bar{q}\mathring{q}, \qquad q^{00} = \frac{n^{00}}{N} = \left(1-\mathring{q}\right)\left(1-\underline{q}\right), \tag{5}$$

$$q^{10} = \frac{n^{10}}{N} = \underline{q}\left(1-\mathring{q}\right), \qquad q^{01} = \frac{n^{01}}{N} = \left(1-\overline{q}\right)\mathring{q}, \tag{6}$$

où $\mathring{q} = (n^{11} + n^{01})/N$ est la proportion d'individus avec la caractéristique $x = 1$ dans la population, et $\bar{q}$, $\underline{q}$ sont les deux probabilités conditionnelles suivantes:

$$\bar{q} = \text{Prob}\{y = 1 \mid x = 1\} = \frac{n^{11}}{n^{11} + n^{01}},$$
$$\underline{q} = \text{Prob}\{y = 1 \mid x = 0\} = \frac{n^{10}}{n^{10} + n^{00}}.$$

Premier pas vers une approche générale permettant de faire de l'inférence statistique en absence d'échantillons représentatifs, cet article se propose d'établir un estimateur de la probabilité conditionnelle[1] $\bar{q} = Prob\{y = 1 \mid x = 1\}$ quand on dispose de deux échantillons indépendants.

3.2 Un estimateur de probabilité conditionnelle par la MCR

3.2.1 Estimateur de $\bar{q}$

Par définition, l'estimateur de la probabilité conditionnelle $\bar{q}$ s'écrit:

$$\widehat{\bar{q}} = \frac{\widehat{n^{11}}}{\widehat{n^{11}} + \widehat{n^{01}}}. \tag{7}$$

L'estimateur $\widehat{n^{yx}}$, $yx \in \{00, 01, 10, 11\}$, fourni par l'expression (2) directement dérivée de la MCR, permet de ré-écrire $\widehat{\bar{q}}$ comme suit:

$$\widehat{\bar{q}} = \frac{n_{11}^{01}\left(n_{01}^{11} + n_{11}^{11}\right)\left(n_{10}^{11} + n_{11}^{11}\right)}{\left[n_{11}^{01}\left(n_{01}^{11} + n_{11}^{11}\right)\left(n_{10}^{11} + n_{11}^{11}\right) + n_{11}^{11}\left(n_{01}^{01} + n_{11}^{01}\right)\left(n_{10}^{01} + n_{11}^{01}\right)\right]}. \tag{8}$$

3.2.2 Variance de l'estimateur $\widehat{\bar{q}}$

Le calcul de la variance de l'estimateur $\widehat{\bar{q}}$ repose sur celui de la variance estimée de $\widehat{n_{00}^{yx}}$. Les recherches se basant sur les MCR pour estimer la taille N d'une population difficile à joindre, adoptent en général la formule de la variance asymptotique de l'estimateur du nombre total des individus qui n'ont pas été capturés (Sekar et Deming, 1949; Manly, 1969). Cette variance asymptotique – calculée pour N grand – sur-estime la variance de la population dans le cas de populations de taille peu élevée (Gerber et *al.*, 2014).

[1] On se restreint sans perte de généralité à l'estimation de la probabilité conditionnelle $\bar{q} = Prob\{y = 1 \mid x = 1\}$. Les estimateurs des trois autres probabilités conditionnelles que l'on peut définir dans ce contexte sont aisément déduits de celui que nous proposons.

C'est pourquoi, nous établissons dans ce qui suit l'expression de la variance exacte de la distribution binomiale[2] pour approximer la variance de l'estimateur $\widehat{\bar{q}}$.

La variable aléatoire n_{00}^{yx} suit une distribution binomiale $B\left(n^{yx}, \pi_{00}^{yx}\right)$, où n^{yx} désigne l'effectif de la population avec les caractéristiques yx et $\pi_{00}^{yx} = \left(1 - r_1^{yx}\right)\left(1 - r_2^{yx}\right)$ la probabilité que ces individus échappent à toute observation. En utilisant les formules (1), (2) ainsi que (3) et (4) qui découlent de la MCR, il est aisé d'établir que[3]:

$$\widehat{V}\left(\widehat{n_{00}^{yx}}\right) = \widehat{n_{00}^{yx}}\left(1 - \frac{\widehat{n_{00}^{yx}}}{\widehat{n^{yx}}}\right) = \left(\frac{n_{01}^{yx}}{n_{01}^{yx} + n_{11}^{yx}}\right)\left(\frac{n_{10}^{yx}}{n_{10}^{yx} + n_{11}^{yx}}\right)\left(n_{01}^{yx} + n_{10}^{yx} + n_{11}^{yx}\right). \tag{9}$$

En accord avec l'hypothèse *d'indépendance* des deux échantillons, $Cov\left(\widehat{n_{00}^{01}}, \widehat{n_{00}^{11}}\right) = 0$.

Comme l'estimateur $\widehat{\bar{q}}$ tel que défini en (8) est non linéaire, la "méthode Delta" est utilisée pour approximer sa variance (Cassela et Berger, 2001). On obtient:

$$\widehat{V}\left(\widehat{\bar{q}}\right) \simeq \frac{\left(\widehat{n^{01}}\right)^2}{\left(\widehat{n^{01}} + \widehat{n^{11}}\right)^4}\widehat{V}\left(\widehat{n_{00}^{11}}\right) + \frac{\left(\widehat{n^{11}}\right)^2}{\left(\widehat{n^{01}} + \widehat{n^{11}}\right)^4}\widehat{V}\left(\widehat{n_{00}^{01}}\right), \tag{10}$$

où $\widehat{V}\left(\widehat{n_{00}^{yx}}\right)$ est calculée d'après (9). Par suite,

$$\widehat{V}\left(\widehat{\bar{q}}\right) \simeq \left(\frac{\left[n_{11}^{11}\left(n_{01}^{01} + n_{11}^{01}\right)\left(n_{10}^{01} + n_{11}^{01}\right)\right]^2\left(n_{11}^{11}n_{11}^{01}\right)^2 n_{01}^{11}n_{10}^{11}}{\left[n_{11}^{11}\left(n_{01}^{01} + n_{11}^{01}\right)\left(n_{10}^{01} + n_{11}^{01}\right) + n_{11}^{01}\left(n_{01}^{11} + n_{11}^{11}\right)\left(n_{10}^{11} + n_{11}^{11}\right)\right]^4}\right)\frac{n_{01}^{11} + n_{10}^{11} + n_{11}^{11}}{\left(n_{01}^{11} + n_{11}^{11}\right)\left(n_{10}^{11} + n_{11}^{11}\right)}$$
$$+ \left(\frac{\left[n_{11}^{01}\left(n_{01}^{11} + n_{11}^{11}\right)\left(n_{10}^{11} + n_{11}^{11}\right)\right]^2\left(n_{11}^{11}n_{11}^{01}\right)^2 n_{01}^{01}n_{10}^{01}}{\left[n_{11}^{11}\left(n_{01}^{01} + n_{11}^{01}\right)\left(n_{10}^{01} + n_{11}^{01}\right) + n_{11}^{01}\left(n_{01}^{11} + n_{11}^{11}\right)\left(n_{10}^{11} + n_{11}^{11}\right)\right]^4}\right)\frac{n_{01}^{01} + n_{10}^{01} + n_{11}^{01}}{\left(n_{01}^{01} + n_{11}^{01}\right)\left(n_{10}^{01} + n_{11}^{01}\right)}.$$

[2]La distribution multinomiale – dont les distributions marginales sont des binomiales, constitue l'une des distributions donnant les meilleures estimations de la taille de la population. C'est en tout cas ainsi si on la compare à la distribution hypergéométrique multivariée, quand on se place dans le cas de populations fermées – c'est à dire quand on suppose que la taille de la population, et quand le nombre total d'échantillons est maintenu constant (Garroch, 1958; Sandland et Cormach, 1984).

[3]Lorsque cette même variance est calculée à partir de l'approximation asymptotique, on obtient

$$\widehat{V_a}\left(\widehat{n_{00}^{yx}}\right) = \widehat{n^{yx}}\frac{\left(1 - r_1^{yx}\right)\left(1 - r_2^{yx}\right)}{r_1^{yx}r_2^{yx}} = \left(\frac{n_{01}^{yx} + n_{11}^{yx}}{n_{11}^{yx}}\right)\left(\frac{n_{10}^{yx} + n_{11}^{yx}}{n_{11}^{yx}}\right)\left(\frac{n_{01}^{yx}n_{10}^{yx}}{n_{11}^{yx}}\right).$$

Comme indiqué plus haut, cette dernière expression surestime la variance:

$$\frac{\widehat{V}\left(\widehat{n_{00}^{yx}}\right)}{\widehat{V_a}\left(\widehat{n_{00}^{yx}}\right)} = \frac{\left(n_{01}^{yx} + n_{10}^{yx} + n_{11}^{yx}\right)\left(n_{11}^{yx}\right)^3}{\left(n_{01}^{yx} + n_{11}^{yx}\right)^2\left(n_{10}^{yx} + n_{11}^{yx}\right)^2} < \frac{\left(n_{11}^{yx}\right)^2}{\left(n_{01}^{yx} + n_{11}^{yx}\right)\left(n_{10}^{yx} + n_{11}^{yx}\right)} = \frac{n_{11}^{yx}}{\widehat{n^{yx}}}.$$

Cette sur-estimation est potentiellement d'autant plus importante que les listes sont incomplètes.

4 Estimateur naïf

Si l'on ne tient pas compte des problèmes d'échantillonnage, la probabilité conditionnelle $\overline{q}$ est obtenue par l'estimateur *"naïf*[4]" suivant:

$$\widetilde{\overline{q}} = \frac{o^{11}}{o^{11} + o^{01}},$$

où l'on note $o^{yx} = n_{01}^{yx} + n_{10}^{yx} + n_{11}^{yx}$ l'effectif observé avec les caractéristiques yx.

4.1 Biais de l'estimateur naïf

Puisque $\widehat{\overline{q}} = \left(o^{11} + \widehat{n_{00}^{11}}\right) / \left(o^{11} + \widehat{n_{00}^{11}} + o^{01} + \widehat{n_{00}^{01}}\right)$, l'estimateur naïf $\widetilde{\overline{q}}$ est généralement à l'origine d'un biais qui est donné par:

$$b_{\overline{q}} \equiv \widetilde{\overline{q}} - \widehat{\overline{q}} = \left(\frac{o^{11}o^{01}}{o^{11} + o^{01}}\right) \left[\frac{\left(\widehat{n_{00}^{01}}/o^{01}\right) - \left(\widehat{n_{00}^{11}}/o^{11}\right)}{o^{11} + \widehat{n_{00}^{11}} + o^{01} + \widehat{n_{00}^{01}}}\right]. \tag{11}$$

En utilisant le fait que $\widehat{n^{yx}} = \widehat{o^{yx}} + \widehat{n_{00}^{yx}} = \left(n_{01}^{yx} + n_{11}^{yx}\right)\left(n_{10}^{yx} + n_{11}^{yx}\right)/n_{11}^{yx}$ et puisque

$$\frac{\widehat{n_{00}^{yx}}}{o^{yx}} = \frac{n_{01}^{yx} n_{10}^{yx}}{n_{11}^{yx}\left(n_{01}^{yx} + n_{10}^{yx} + n_{11}^{yx}\right)},$$

l'expression du biais peut se ré-écrire:

$$\begin{aligned} b_{\overline{q}} = {} & \left[\left(n_{01}^{11} + n_{10}^{11} + n_{11}^{11}\right) + \left(n_{01}^{01} + n_{10}^{01} + n_{11}^{01}\right)\right]^{-1} \\ & \times \left[\frac{\left[n_{11}^{11}\left(n_{01}^{11} + n_{10}^{11} + n_{11}^{11}\right)\right] n_{01}^{01} n_{10}^{01} - n_{01}^{11} n_{10}^{11}\left[n_{11}^{01}\left(n_{01}^{01} + n_{10}^{01} + n_{11}^{01}\right)\right]}{\left(n_{11}^{11} + n_{01}^{11}\right)\left(n_{10}^{11} + n_{11}^{11}\right) n_{11}^{01} + n_{11}^{11}\left(n_{11}^{01} + n_{01}^{01}\right)\left(n_{10}^{01} + n_{11}^{01}\right)}\right]. \end{aligned} \tag{12}$$

4.1.1 Condition pour l'absence de biais

A partir de la formule (11), il est aisé d'établir que le biais de l'estimateur naïf $\widetilde{\overline{q}}$ est nul si et seulement si:

$$\frac{\widehat{n_{00}^{11}}}{\widehat{n^{11}}} = \frac{n_{01}^{11} n_{10}^{11}}{\left(n_{01}^{11} + n_{11}^{11}\right)\left(n_{10}^{11} + n_{11}^{11}\right)} = \frac{n_{01}^{01} n_{10}^{01}}{\left(n_{01}^{01} + n_{11}^{01}\right)\left(n_{10}^{01} + n_{11}^{01}\right)} = \frac{\widehat{n_{00}^{01}}}{\widehat{n^{01}}}. \tag{13}$$

L'expression (13) indique que l'estimateur naïf $\widetilde{\overline{q}}$ sera non biaisé si et seulement si la proportion des individus non-observés est identique pour les individus de caractéristiques $yx = 11$ et ceux de caractéristiques $yx = 01$; autrement dit, tous les individus de caractéristique $x = 1$, ont la même probabilité d'être observés.[5]

[4]Nous définissons un estimateur comme étant naïf quand il est basé sur les seules observations, sans prendre en compte les problèmes de représentativité des échantillons.

[5]Observons que la condition (13) n'implique pas forcément que les probabilités de capture r_1^{yx} et r_2^{yx} sont identiques pour tous les individus de caractéristique $x = 1$. Par contre, évidemment, si les probabilités de capture sont identiques, la condition (13) est nécessairement vérifiée.

4.1.2 Biais asymptotique

Même asymptotiquement, il n'y a aucune raison que l'estimateur naïf $\widetilde{\overline{q}}$ soit sans biais. En effet, en substituant dans l'expression de $b_{\overline{q}}$ donnée en (11) l'expression théorique des fréquences qui découle du modèle statistique sous-jacent (Section 3.1 et en particulier équations (5)-(6)), on obtient facilement

$$\lim_{N\to\infty} b_{\overline{q}} = \frac{\bar{q}\mathring{q}\underline{q}\,(1-\mathring{q})\left[(1-r_2^{01})\,(1-r_1^{01})-(1-r_1^{11})\,(1-r_2^{11})\right]}{\left[\bar{q}\mathring{q}+\underline{q}\,(1-\mathring{q})\right]\left\{\bar{q}\mathring{q}\left[1+(1-r_2^{11})\,(1-r_1^{11})\right]+\underline{q}\,(1-\mathring{q})\left[1+(1-r_2^{01})\,(1-r_1^{01})\right]\right\}}.$$

Ceci signifie que $b_{\overline{q}}$ est asymptotiquement nul si et seulement si

$$\left(1-r_1^{11}\right)\left(1-r_2^{11}\right)=\left(1-r_1^{01}\right)\left(1-r_2^{01}\right). \tag{14}$$

En d'autres termes, on retrouve le fait que, pour que l'estimateur naïf de $\overline{q}$ soit (asymptotiquement) sans biais, il faut que la probabilité des individus de ne faire partie d'aucune des deux listes soit identique pour les individus de caractéristiques $yx = 01$ et ceux de caractéristiques $yx = 11$ – c'est à dire pour tous les individus de caractéristique $x = 1$.

Bien-sûr, il n'y a aucune raison de supposer *apriori* que la condition (14) soit remplie.

4.2 Test statistique de la condition de biais nul.

La condition de biais nul, $b_{\overline{q}} = 0$, telle que donnée en (13) ou en (14) peut s'écrire sous la forme de l'hypothèse statistique suivante:

$$H_0: \quad p_{00}^{11} - p_{00}^{01} = 0 \quad (b_{\overline{q}} = 0)$$

où $p_{00}^{11} = (n_{00}^{11}/n^{11})$ est la proportion d'individus de caractéristiques $yx = 11$ absente des deux listes, et $p_{00}^{01} = (n_{00}^{01}/n^{01})$ est cette même proportion pour $yx = 01$.

On a déjà mentionné que n_{00}^{yx} obéit à une loi binomiale. Lorsque n^{yx} suffisamment élevé[6], la distribution de $p_{00}^{yx} = n_{00}^{yx}/n^{yx}$ tend en loi vers une distribution normale:

$$\widehat{p_{00}^{yx}} \xrightarrow{\mathcal{L}} \mathcal{N}\left(\left(1-r_1^{yx}\right)\left(1-r_2^{yx}\right), \frac{1}{n^{yx}}\left(1-r_1^{yx}\right)\left(1-r_2^{yx}\right)\left(r_1^{yx}+r_2^{yx}-r_1^{yx}r_2^{yx}\right)\right).$$

[6] C'est l'hypothèse généralement adoptée. En pratique, on a une bonne approximation quand $n^{yx} \geq 30$ et $min\left(n^{yx}\pi_{00}^{yx},\, n^{yx}\pi_{00}^{yx}\left(1-\pi_{00}^{yx}\right)\right) \geq 5$. Dans le cas de petits échantillons, le théorème central limite n'est pas applicable. Différents tests doivent alors être envisagés. Deux des tests les plus utilisés sont le test *khi-deux* de Pearson et le test exact de Fisher. En réalité, ces deux tests sont utilisés pour tester l'indépendance entre une variable X (en lignes) et une variable Y (en colonnes) rangées dans un tableau de contingence de r lignes et c colonnes. Dénommés aussi modèle à essais comparatifs, de double dichotomie ou modèle d'homogénéité, basés sur une double binomiale, ils sont aussi utilisés pour comparer deux proportions (Kroll, 1989 ; Yates, 1984). Dans le cas du test *khi-deux* de Pearson, une correction de continuité est réalisée pour l'adapter au cas de petits échantillons quand au moins une des cellules du tableau de contingence a un effectif inférieur à 5 sous la contrainte que les effectifs partiels théoriques soient tous supérieurs à 3 (Yates, 1984). C'est pourquoi le test exact de Fischer est le plus souvent conseillé (Yates, 1984; Kroll, 1989 ; Yu, 2017). Ce test dont la démarche est basée sur une distribution hypergéométrique, permet de calculer de manière exacte la probabilité critique p_c.

L'estimateur de la proportion p_{00}^{yx} est donné par:

$$\widehat{p_{00}^{yx}} = \left(1 - \widehat{r_1^{yx}}\right)\left(1 - \widehat{r_2^{yx}}\right) = \frac{n_{01}^{yx} n_{10}^{yx}}{\left(n_{01}^{yx} + n_{11}^{yx}\right)\left(n_{10}^{yx} + n_{11}^{yx}\right)}. \tag{15}$$

La différence $\widehat{p_{00}^{11}} - \widehat{p_{00}^{01}}$ s'écrit donc:

$$\widehat{p_{00}^{11} - p_{00}^{01}} = \frac{n_{01}^{11} n_{10}^{11}\left[n_{11}^{01}\left(n_{01}^{01} + n_{10}^{01} + n_{11}^{01}\right)\right] - n_{01}^{01} n_{10}^{01}\left[n_{11}^{11}\left(n_{01}^{11} + n_{10}^{11} + n_{11}^{11}\right)\right]}{\left(n_{01}^{11} + n_{11}^{11}\right)\left(n_{10}^{11} + n_{11}^{11}\right)\left(n_{01}^{01} + n_{11}^{01}\right)\left(n_{10}^{01} + n_{11}^{01}\right)}. \tag{16}$$

Puisque les deux variables aléatoires p_{00}^{11} et p_{00}^{01} sont indépendantes, la variance de leur différence est donnée par:

$$\widehat{V}\left(\widehat{p_{00}^{11} - p_{00}^{01}}\right) = \widehat{V}\left(\widehat{p_{00}^{11}}\right) + \widehat{V}\left(\widehat{p_{00}^{01}}\right) \tag{17}$$

et, sous l'hypothèse H_0:

$$\begin{aligned}\widehat{V}\left(\widehat{p_{00}^{11} - p_{00}^{01}}\right) = &\left(\frac{n_{11}^{01}\left[n_{11}^{11}\left(n_{01}^{11} + n_{10}^{11} + n_{11}^{11}\right)\right] + n_{11}^{11}\left[n_{11}^{01}\left(n_{01}^{01} + n_{10}^{01} + n_{11}^{01}\right)\right]}{n_{11}^{01}\left(n_{01}^{11} + n_{11}^{11}\right)\left(n_{10}^{11} + n_{11}^{11}\right) + n_{11}^{11}\left(n_{01}^{01} + n_{11}^{01}\right)\left(n_{10}^{01} + n_{11}^{01}\right)}\right) \\ &\times \left(\frac{n_{11}^{01}\left(n_{01}^{11} n_{10}^{11}\right) + n_{11}^{11}\left(n_{01}^{01} n_{10}^{01}\right)}{\left(n_{01}^{11} + n_{11}^{11}\right)\left(n_{10}^{11} + n_{11}^{11}\right)\left(n_{01}^{01} + n_{11}^{01}\right)\left(n_{10}^{01} + n_{11}^{01}\right)}\right).\end{aligned} \tag{18}$$

L'hypothèse d'égalité des probabilités p_{00}^{11} et p_{00}^{01} ne peut être rejetée si , au risque α donné[7]:

$$P\left(\left|Z_{\widehat{\overline{p}}} = \frac{\left(\widehat{p_{00}^{11} - p_{00}^{01}}\right)}{\sqrt{\widehat{V}\left(\widehat{p_{00}^{11} - p_{00}^{01}}\right)}} \xrightarrow{\mathcal{L}} N(0,1)\right| \leq z_{\frac{\alpha}{2}}\right) = 1 - \alpha.$$

Par suite, on ne peut rejeter l'hypothèse d'égalité des probabilités p_{00}^{11} et p_{00}^{01} dès lors que

$$\begin{aligned}&\frac{\left\{\left(n_{01}^{11} n_{10}^{11}\right)\left[n_{11}^{01}\left(n_{01}^{01} + n_{10}^{01} + n_{11}^{01}\right)\right] - \left(n_{01}^{01} n_{10}^{01}\right)\left[n_{11}^{11}\left(n_{01}^{11} + n_{10}^{11} + n_{11}^{11}\right)\right]\right\}^2}{\left[\left(n_{01}^{11} n_{10}^{11}\right) n_{11}^{01} + \left(n_{01}^{01} n_{10}^{01}\right) n_{11}^{11}\right]\left(n_{01}^{11} + n_{11}^{11}\right)\left(n_{10}^{11} + n_{11}^{11}\right)\left(n_{01}^{01} + n_{11}^{01}\right)\left(n_{10}^{01} + n_{11}^{01}\right)} \leq \\ &\qquad \frac{n_{11}^{01}\left[n_{11}^{11}\left(n_{01}^{11} + n_{10}^{11} + n_{11}^{11}\right)\right] + n_{11}^{11}\left[n_{11}^{01}\left(n_{01}^{01} + n_{10}^{01} + n_{11}^{01}\right)\right]}{n_{11}^{01}\left(n_{01}^{11} + n_{11}^{11}\right)\left(n_{10}^{11} + n_{11}^{11}\right) + n_{11}^{11}\left(n_{01}^{01} + n_{11}^{01}\right)\left(n_{10}^{01} + n_{11}^{01}\right)}\left(z_{\alpha/2}\right)^2.\end{aligned} \tag{19}$$

[7]Pour de meilleures performances, une correction de continuité peut être réalisée. Si $\widehat{p_{00}^{11}} < \widehat{p_{00}^{01}}$, on substitue $\widehat{p_{00}^{11}} - \widehat{p_{00}^{01}}$ au numérateur par $\left\{\widehat{p_{00}^{11}} - \widehat{p_{00}^{01}} + \frac{1}{2}\left[\left(\widehat{n_{00}^{01}}\right)^{-1} + \left(\widehat{n_{00}^{11}}\right)^{-1}\right]\right\}$. Dans le cas contraire, on considère au numérateur $\left\{\widehat{p_{00}^{11}} - \widehat{p_{00}^{01}} - \frac{1}{2}\left[\left(\widehat{n_{00}^{01}}\right)^{-1} + \left(\widehat{n_{00}^{11}}\right)^{-1}\right]\right\}$.

References

[1] CASSELA, G. et BERGER, R. L. (2001). *Statistical Inference*. Second Edition, Thomson Learning.

[2] CHAO, A. et *al.* (2001). *The applications of capture-recapture models to epidemiological data*. Statistics in medicine, vol. 20, no 20, p. 3123-3157.

[3] GERBER, B. D. et *al.* (2014). *Estimating the abundance of rare and elusive carnivores from photographic-sampling data when the population size is very small*. Population eco-logy, vol. 56, no 3, p. 463-470.

[4] KROLL, N. E. A. (1989). Testing Independence in 2 x 2 Contingency Tables. Journal of Educational Statistics Spring, Vol. 14, No. I, pp. 47-79.

[5] LAVALLÉE, P. et RIVEST, L-P. (2012). *Capture-recapture sampling and indirect sampling*. Journal of Official Statistics, vol. 28, no 1, p. 1.

[6] MANLY, B. F. J. (1969). *Some Properties of a Method of Estimating the Size of Mobile Animal Populations*. Biometrika, Vol. 56, No. 2, pp. 407-410.

[7] MANRIQUE-VALLIER, D. et *al.* (2013). *Multiple systems estimation techniques for estimating casualties in armed conflicts*. Counting Civilian Casualties: An Introduction to Recording and Estimating Nonmilitary Deaths in Conflict, vol. 165.

[8] MITCHELL, S. A. (2014). *Capture-recapture Estimation for Conflict Data and Hierarchical Models for Program Impact Evaluation*. Doctoral dissertation, Harvard University.

[9] SEKAR,C. C. et DEMING, W. E. (1949). *On a Method of Estimating Birth and Death Rates and the Extent of Registration*. Journal of the American Statistical Association, Vol. 44, No. 245, pp. 101-115.

[10] YATES, F. (1984). Tests of Significance for 2 $\times$ 2 Contingency Tables. Journal of the Royal Statistical Society. Series A (General), Vol. 147, No. 3, pp. 426-463.

[11] YU, Y. et *al.* (2017).Tests of Independence for a 2 2 Contingency Table with Random Margins. International Journal of Statistics and Probability; Vol. 6, No. 2.

[12] ZWANE, E. N. et VAN DER HEIJDEN, P. G. M.(2008). *Capture-recapture studies with incomplete mixed categorical and continous covariates*. Journal of data science, vol. 6, p. 557-572.