Composante du produit nº 12-001-X
Division des méthodes d'enquêtes auprès des entreprises

Article

# Estimation bayésienne linéaire en population finie en considérant en particulier les données catégoriques

par Kelly Cristina M. Gonçalves, Fernando A.S. Moura
et Helio S. Migon

Juin 2014

Statistique Canada Statistics Canada

Canada

## Comment obtenir d'autres renseignements

Pour toute demande de renseignements au sujet de ce produit ou sur l'ensemble des données et des services de Statistique Canada, visiter notre site Web à www.statcan.gc.ca.

Vous pouvez également communiquer avec nous par :

**Courriel** à infostats@statcan.gc.ca

**Téléphone** entre 8 h 30 et 16 h 30 du lundi au vendredi aux numéros sans frais suivants :

| | |
|---|---|
| • Service de renseignements statistiques | 1-800-263-1136 |
| • Service national d'appareils de télécommunications pour les malentendants | 1-800-363-7629 |
| • Télécopieur | 1-877-287-4369 |

**Programme des services de dépôt**

| | |
|---|---|
| Service de renseignements | 1-800-635-7943 |
| Télécopieur | 1-800-565-7757 |

## Comment accéder à ce produit

Le produit n° 12-001-X au catalogue est disponible gratuitement sous format électronique. Pour obtenir un exemplaire, il suffit de visiter notre site Web à www.statcan.gc.ca et de parcourir par « Ressource clé » > « Publications ».

## Normes de service à la clientèle

Statistique Canada s'engage à fournir à ses clients des services rapides, fiables et courtois. À cet égard, notre organisme s'est doté de normes de service à la clientèle que les employés observent. Pour obtenir une copie de ces normes de service, veuillez communiquer avec Statistique Canada au numéro sans frais 1-800-263-1136. Les normes de service sont aussi publiées sur le site www.statcan.gc.ca sous « À propos de nous » > « Notre organisme » > « Offrir des services aux Canadiens ».

Publication autorisée par le ministre responsable de Statistique Canada

© Ministre de l'Industrie, 2014

Tous droits réservés. L'utilisation de la présente publication est assujettie aux modalités de l'entente de licence ouverte de Statistique Canada (http://www.statcan.gc.ca/reference/licence-fra.html).

This publication is also available in English.

**Note de reconnaissance**

Le succès du système statistique du Canada repose sur un partenariat bien établi entre Statistique Canada et la population du Canada, ses entreprises, ses administrations et les autres établissements. Sans cette collaboration et cette bonne volonté, il serait impossible de produire des statistiques exactes et actuelles.

**Signes conventionnels**

Les signes conventionnels suivants sont employés dans les publications de Statistique Canada :

| | |
|---|---|
| . | indisponible pour toute période de référence |
| .. | indisponible pour une période de référence précise |
| ... | n'ayant pas lieu de figurer |
| 0 | zéro absolu ou valeur arrondie à zéro |
| $0^{s}$ | valeur arrondie à 0 (zéro) là où il y a une distinction importante entre le zéro absolu et la valeur arrondie |
| $^{p}$ | provisoire |
| $^{r}$ | révisé |
| x | confidentiel en vertu des dispositions de la *Loi sur la statistique* |
| $^{E}$ | à utiliser avec prudence |
| F | trop peu fiable pour être publié |
| * | valeur significativement différente de l'estimation pour la catégorie de référence (p<0,05) |

Techniques d'enquête, juin 2014
Vol. 40, N° 1, pp. 17-31
Statistique Canada, N° 12-001-X au catalogue


# Estimation bayésienne linéaire en population finie en considérant en particulier les données catégoriques

**Kelly Cristina M. Gonçalves, Fernando A.S. Moura et Helio S. Migon**[1]

Résumé

L'estimateur bayésien linéaire en population finie est obtenu en partant d'un modèle de régression à deux degrés spécifié uniquement par les moyennes et les variances de certains paramètres du modèle associés à chaque degré de la hiérarchie. Nombre d'estimateurs fondés sur le plan de sondage usuels décrits dans la littérature peuvent être obtenus en tant que cas particuliers. Un nouvel estimateur par le ratio est également proposé pour la situation pratique où de l'information auxiliaire est disponible. L'application de la même approche bayésienne linéaire est proposée pour estimer des proportions pour des données catégoriques multiples associées aux unités de la population finie, ce qui constitue la principale contribution des présents travaux, et est illustrée au moyen d'un exemple numérique.

**Mots-clés :** Échangeabilité; modèle linéaire; prédiction bayésienne linéaire.

# 1 Introduction

Les sondages sont depuis longtemps un moyen important d'obtenir de l'information exacte auprès d'une population finie. Par exemple, les gouvernements ont besoin de statistiques descriptives sur la population pour évaluer et mettre en œuvre leurs politiques. Au cours du premier tiers du XX$^{e}$ siècle, la principale préoccupation des spécialistes de la statistique officielle était d'établir une norme de pratique acceptable. Neyman (1934) a créé un cadre de ce type en introduisant des méthodes de randomisation dans le processus d'échantillonnage. Il préconisait l'utilisation de la distribution de randomisation, ou distribution aléatoire, induite par le plan de sondage pour évaluer les propriétés fréquentistes de diverses procédures. Il a également introduit la notion de stratification avec répartition optimale de la taille de l'échantillon et l'utilisation de probabilités de sélection inégales. Ses travaux sont considérés comme la pierre angulaire de la théorie des sondages fondée sur le plan de sondage et ont inspiré de nombreux auteurs. Par exemple, Horvitz et Thompson (1952) ont proposé une théorie générale de l'échantillonnage avec probabilités inégales et la méthode d'estimation pondérée par les probabilités, qui est connue sous le nom d'« estimateur de Horvitz-Thompson ».

Partout dans le monde, les organismes statistiques officiels ont trouvé la théorie des sondages fondée sur le plan de sondage très intéressante. Comme l'ont fait remarquer Skinner, Holt et Smith (1989), page 2, la raison principale de cet intérêt est qu'essentiellement, cette théorie ne dépend pas d'une loi de probabilité. En effet, depuis les travaux de Neyman, tous les progrès réalisés en théorie des sondages ont été fortement influencés par l'usage descriptif du sondage. Cette approche a eu pour conséquence un manque de développements théoriques concernant l'utilisation analytique des sondages, en particulier à des fins de prédiction. Dans certaines situations précises, l'approche fondée sur le plan de sondage s'est avérée inefficace, fournissant des prédicteurs inadéquats. Ainsi, l'estimation sur petits domaines et la présence de non-réponse ne peuvent pas être traitées par l'approche fondée sur le plan de sondage sans

1. Kelly Cristina M. Gonçalves, Departamento de Estatística, Universidade Federal do Rio de Janeiro (UFRJ), RJ, Brésil. Courriel : kelly@im.ufrj.br; Fernando A. S. Moura, Departamento de Estatística, Universidade Federal do Rio de Janeiro (UFRJ), RJ, Brésil. Courriel : fmoura@im.ufrj.br; Helio S. Migon, Departamento de Estatística, Universidade Federal do Rio de Janeiro (UFRJ), RJ, Brésil. Courriel : migon@im.ufrj.br.

faire appel à des hypothèses implicites, ce qui équivaut à émettre l'hypothèse d'un modèle. Les partisans de l'approche fondée sur le plan de sondage soutiennent que l'inférence fondée sur un modèle dépend en grande partie des hypothèses du modèle, lesquelles pourraient ne pas être vérifiées. Par ailleurs, l'inférence par intervalle sur les paramètres de la population cible (habituellement les totaux et les moyennes) s'appuie sur le théorème central limite, qui n'est pas applicable dans de nombreuses situations pratiques, où la taille de l'échantillon n'est pas suffisamment grande et/ou les hypothèses d'indépendance des variables aléatoires concernées ne sont pas raisonnables.

Basu (1971) rejetait les estimations des quantités de population dépendantes de la règle d'échantillonnage, comme les probabilités d'inclusion. Selon lui, cette méthode d'estimation ne satisfaisait pas le principe de vraisemblance dont il était un partisan. Basu (1971) a créé l'exemple de l'éléphant de cirque pour montrer que l'estimateur de Horvitz-Thompson pouvait produire des estimations inappropriées et a proposé un estimateur de rechange. La question qui se pose est celle de savoir s'il est possible de concilier les deux approches. Dans le contexte du modèle de superpopulation, Zacks (2002) a montré que l'on peut retrouver certains estimateurs fondés sur le plan de sondage en utilisant une approche basée sur un modèle de régression généralisée. Little (2003) affirme qu'une spécification prudente du modèle, sensible au plan de sondage, permet de répondre aux préoccupations concernant les spécifications du modèle, et que la statistique bayésienne offre un traitement cohérent et unifié de l'inférence par sondage descriptive et analytique. Il a donné certains exemples afin d'illustrer comment l'inférence fondée sur le plan de sondage classique peut découler de la perspective bayésienne en utilisant certains modèles avec priors non informatifs.

Dans le contexte bayésien, une autre proposition intéressante en vue de concilier l'approche fondée sur le plan de sondage et celle fondée sur un modèle a été avancée par Smouse (1984). La méthode incorpore l'information a priori dans les modèles d'inférence en population finie en s'appuyant sur des méthodes bayésiennes d'estimation par les moindres carrés et nécessite seulement la spécification des premier et deuxième moments de la distribution concernée, qui décrivent les connaissances a priori au sujet des structures présentes dans la population. L'approche peut remplacer les méthodes de randomisation et se situe à mi-chemin entre deux visions extrêmes, à savoir les procédures fondées sur le plan de sondage d'une part et celles fondées sur des modèles de superpopulation d'autre part. Dans un rapport inédit, O'Hagan (1985) a présenté des estimateurs bayésiens linéaires dans certains contextes de sondage particuliers et O'Hagan (1987) a également dérivé des estimateurs bayésiens linéaires pour certains modèles de réponse randomisés. O'Hagan (1985) a pris en considération plusieurs structures de population, dont la stratification et la mise en grappes, en émettant des hypothèses appropriées au sujet des premier et deuxième moments, et a montré comment obtenir certains estimateurs fondés sur le plan fréquemment utilisés en temps que cas particuliers de cette approche plus générale. Il a également fait remarquer que ses estimations ne tenaient pas compte de l'échantillonnage non informatif. Il a cité Scott (1977) et fait le commentaire que l'échantillonnage informatif devrait être exécuté par une analyse bayésienne complète. Une référence importante concernant le traitement de l'échantillonnage informatif avec des modèles hiérarchiques figure dans Pfeffermann, Moura et Silva (2006).

La présentation de l'article est la suivante. La section 2 décrit de manière générale l'approche d'estimation bayésienne linéaire appliquée à un modèle de régression linéaire général pour la prédiction en population finie et montre comment obtenir certains estimateurs fondés sur le plan de sondage à titre de cas particuliers. À la section 3, un nouvel estimateur est proposé pour la situation pratique où de

l'information auxiliaire est disponible. À la section 4, l'approche d'estimation bayésienne linéaire est étendue aux données catégoriques multiples. Enfin, la section 5 offre certaines conclusions et des suggestions de futures recherches.

# 2 Estimation bayésienne linéaire en population finie

L'approche bayésienne s'est avérée fructueuse dans de nombreuses applications, particulièrement lorsque l'analyse des données a été améliorée par des jugements d'expert. Cependant, si les modèles bayésiens possèdent de nombreuses caractéristiques intéressantes, leur application requiert souvent la spécification complète d'une loi a priori, ou prior, pour un grand nombre de paramètres. Goldstein et Wooff (2007), section 1.2, soutiennent que, à mesure que le problème se complexifie, notre aptitude réelle à spécifier complètement la loi a priori et/ou le modèle d'échantillonnage en détail diminue. Ils concluent que, dans de telles situations, il est nécessaire d'élaborer des méthodes qui reposent sur une spécification faisant appel à la croyance partielle.

Hartigan (1969) a proposé une méthode d'estimation, qu'il a nommée approche d'estimation bayésienne linéaire, qui ne nécessite que la spécification des premier et deuxième moments. Les estimateurs résultants ont la propriété de minimiser la perte quadratique a posteriori parmi tous les estimateurs qui sont linéaires en les données et peuvent être considérés comme des approximations des moyennes a posteriori. L'approche d'estimation bayésienne linéaire, qui est employée pleinement dans le présent article, est décrite brièvement ci-dessous.

## 2.1 Approche bayésienne linéaire

Soit $\mathbf{y}_s$ le vecteur des observations et $\boldsymbol{\theta}$, le paramètre à estimer. Pour chaque valeur de $\boldsymbol{\theta}$ et chaque estimation possible $\mathbf{d}$, appartenant à l'espace paramétrique $\boldsymbol{\Theta}$, nous associons une fonction de perte quadratique $L(\boldsymbol{\theta},\mathbf{d}) = (\boldsymbol{\theta}-\mathbf{d})'(\boldsymbol{\theta}-\mathbf{d}) = tr(\boldsymbol{\theta}-\mathbf{d})(\boldsymbol{\theta}-\mathbf{d})'$. Nous souhaitons avant tout trouver la valeur de $\mathbf{d}$ qui minimise $r(\mathbf{d}) = E[L(\boldsymbol{\theta},\mathbf{d})\,|\,\mathbf{y}_s]$, la valeur prévue conditionnelle de la fonction de perte quadratique fournie par les données.

Supposons que la distribution conjointe de $\boldsymbol{\theta}$ et $\mathbf{y}_s$ est spécifiée partiellement par leurs deux premiers moments seulement :

$$\begin{pmatrix}\boldsymbol{\theta}\\ \mathbf{y}_s\end{pmatrix} \sim \left[\begin{pmatrix}\mathbf{a}\\ \mathbf{f}\end{pmatrix}, \begin{pmatrix}\mathbf{R} & \mathbf{AQ}\\ \mathbf{QA}' & \mathbf{Q}\end{pmatrix}\right], \tag{2.1}$$

où $\mathbf{a}$ et $\mathbf{f}$, respectivement, désignent les vecteurs des moyennes et $\mathbf{R}$, $\mathbf{AQ}$ et $\mathbf{Q}$, les éléments de la matrice de covariance de $\boldsymbol{\theta}$ et $\mathbf{y}_s$.

L'estimateur bayésien linéaire (EBL) de $\boldsymbol{\theta}$ est la valeur de $\mathbf{d}$ qui minimise la valeur prévue de cette fonction de perte quadratique dans la classe de toutes les estimations linéaires de la forme $\mathbf{d} = \mathbf{d}(\mathbf{y}_s) = \mathbf{h} + \mathbf{H}\mathbf{y}_s$, pour un vecteur $\mathbf{h}$ et une matrice $\mathbf{H}$. Donc, l'EBL de $\boldsymbol{\theta}$, $\hat{\mathbf{d}}$, et sa variance associée, $\hat{V}(\hat{\mathbf{d}})$, sont donnés respectivement par

$$\hat{\mathbf{d}} = \mathbf{a} + \mathbf{A}(\mathbf{y}_s - \mathbf{f}) \text{ et } \hat{V}(\hat{\mathbf{d}}) = \mathbf{R} - \mathbf{AQA}'. \tag{2.2}$$

Il convient de souligner que l'EBL dépend de la spécification des premier et deuxième moments de la distribution conjointe partiellement spécifiée en (2.1). Le problème de l'obtention de ces quantités est traité aux sections 2.3.1 et 4.1 pour certains cas particuliers.

## 2.2 Approche bayésienne linéaire en population finie

Considérons $U = \{u_1, \ldots, u_N\}$ une population finie comprenant $N$ unités. Soit $\mathbf{y} = (y_1, \ldots, y_N)'$ le vecteur des valeurs d'intérêt des unités dans $U$. Le vecteur de réponses $\mathbf{y}$ est divisé en le vecteur des valeurs connues observées sur l'échantillon $\mathbf{y}_s$ de taille $n$, et en le vecteur des valeurs non observées $\mathbf{y}_{\bar{s}}$ de dimension $N - n$. Le problème général consiste à prédire une fonction du vecteur $\mathbf{y}$, telle que le total $T = \sum_{i=1}^{N} y_i = \mathbf{1}_s' \mathbf{y}_s + \mathbf{1}_{\bar{s}}' \mathbf{y}_{\bar{s}}$, où $\mathbf{1}_s$ et $\mathbf{1}_{\bar{s}}$ sont les vecteurs de 1 de dimensions $n$ et $N - n$, respectivement. Dans l'approche fondée sur un modèle, cela se fait habituellement en se servant d'un modèle paramétrique hypothétique pour les valeurs de population $y_i$, puis en obtenant le meilleur prédicteur linéaire sans biais empirique (EBLUP) pour le vecteur inconnu $\mathbf{y}_{\bar{s}}$ sous ce modèle. Habituellement, l'erreur quadratique moyenne de l'EBLUP de $T$ s'obtient par approximation d'ordre deux, de même qu'un estimateur sans biais. Voir Valliant, Dorfman et Royall (2000), chapitre 2, pour des renseignements détaillés.

L'approche bayésienne de la prédiction en population finie repose souvent sur l'hypothèse d'un modèle paramétrique, mais elle vise à trouver la loi a posteriori de $T$ sachant $\mathbf{y}_s$. On peut obtenir les estimations ponctuelles en spécifiant une fonction de perte, quoique dans de nombreux problèmes pratiques, on considère souvent la moyenne a posteriori dont la variance associée est donnée par la variance a posteriori, c'est-à-dire :

$$E(T \mid \mathbf{y}_s) = \mathbf{1}_s' \mathbf{y}_s + \mathbf{1}_{\bar{s}}' E(\mathbf{y}_{\bar{s}} \mid \mathbf{y}_s) \text{ et } V(T \mid \mathbf{y}_s) = \mathbf{1}_{\bar{s}}' V(\mathbf{y}_{\bar{s}} \mid \mathbf{y}_s) \mathbf{1}_{\bar{s}}. \tag{2.3}$$

Il est possible d'obtenir une approximation des quantités dans (2.3) en utilisant une approche d'estimation bayésienne linéaire. Ici, nous obtiendrons en particulier les estimateurs en émettant l'hypothèse d'un modèle hiérarchique à deux degrés général en population finie, spécifié uniquement par sa moyenne et sa matrice de variance-covariance, présenté dans Bolfarine et Zacks (1992), page 76. Les cas particuliers décrivant les structures de population habituellement observées en pratique peuvent être dérivés facilement de (2.4). Le modèle général peut s'écrire :

$$\mathbf{y} \mid \boldsymbol{\beta} \sim [\mathbf{X}\boldsymbol{\beta}, \mathbf{V}] \text{ et } \boldsymbol{\beta} \sim [\mathbf{a}, \mathbf{R}], \tag{2.4}$$

où $\mathbf{X}$ est une matrice de covariables de dimensions $N \times p$, avec les lignes $\mathbf{X}_i = (x_{i1}, \ldots, x_{ip})$, $i = 1, \ldots, N$; $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ est un vecteur de dimension $p \times 1$ de paramètres inconnus, et $\mathbf{y}$, sachant $\boldsymbol{\beta}$, est un vecteur aléatoire de moyenne $\mathbf{X}\boldsymbol{\beta}$ et de matrice de covariance connue $\mathbf{V}$ de dimensions $N \times N$. De manière analogue, $\mathbf{a}$ et $\mathbf{R}$ sont, respectivement, le vecteur des moyennes a priori de dimension $p \times 1$ et la matrice de covariance a priori de dimensions $p \times p$ de $\boldsymbol{\beta}$.

Puisque le vecteur de réponses $\mathbf{y}$ est divisé en $\mathbf{y}_s$ et $\mathbf{y}_{\bar{s}}$, la matrice $\mathbf{X}$, qui est supposée connue, est divisée de manière analogue en $\mathbf{X}_s$ et $\mathbf{X}_{\bar{s}}$, et $\mathbf{V}$ est divisée en $\mathbf{V}_s$, $\mathbf{V}_{\bar{s}}$, $\mathbf{V}_{s\bar{s}}$ et $\mathbf{V}_{\bar{s}s}$. L'objectif premier est de prédire $\mathbf{y}_{\bar{s}}$ sachant l'échantillon observé $\mathbf{y}_s$, puis le total $T$. Voici les étapes que nous avons

suivies : premièrement, nous avons utilisé une loi a priori conjointe qui n'est spécifiée que partiellement en ce qui concerne les moments comme il suit :

$$\begin{pmatrix}\mathbf{y}_{\bar{s}}\\ \mathbf{y}_s\end{pmatrix}\Big|\boldsymbol{\beta} \sim \left[\begin{pmatrix}\mathbf{X}_{\bar{s}}\boldsymbol{\beta}\\ \mathbf{X}_s\boldsymbol{\beta}\end{pmatrix},\begin{pmatrix}\mathbf{V}_{\bar{s}} & \mathbf{V}_{s\bar{s}}\\ \mathbf{V}_{\bar{s}s} & \mathbf{V}_s\end{pmatrix}\right].$$

Donc, en appliquant le résultat général dans l'équation (2.2), l'EBL de $E(\mathbf{y}_{\bar{s}}|\mathbf{y}_s,\boldsymbol{\beta})$ et la perte quadratique prévue minimale (variance associée) sont donnés par :

$$\hat{E}(\mathbf{y}_{\bar{s}}|\mathbf{y}_s,\boldsymbol{\beta}) = \mathbf{X}_{\bar{s}}\boldsymbol{\beta} + \mathbf{V}_{s\bar{s}}\mathbf{V}_s^{-1}(\mathbf{y}_s - \mathbf{X}_s\boldsymbol{\beta}) \text{ et } \hat{V}(\mathbf{y}_{\bar{s}}|\mathbf{y}_s,\boldsymbol{\beta}) = \mathbf{V}_{\bar{s}} - \mathbf{V}_{s\bar{s}}\mathbf{V}_s^{-1}\mathbf{V}_{s\bar{s}}. \quad (2.5)$$

**Remarque 1** : Il convient de souligner que, sous l'hypothèse de normalité, $E(\mathbf{y}_{\bar{s}}|\mathbf{y}_s,\boldsymbol{\beta})$ et $V(\mathbf{y}_{\bar{s}}|\mathbf{y}_s,\boldsymbol{\beta})$ sont données, respectivement, par les membres de droite des équations de (2.5). L'EBL et sa variance associée donnés en (2.5) peuvent être considérés respectivement, comme des approximations de $E(\mathbf{y}_{\bar{s}}|\mathbf{y}_s,\boldsymbol{\beta})$ et $V(\mathbf{y}_{\bar{s}}|\mathbf{y}_s,\boldsymbol{\beta})$ pour les cas de non-normalité.

Maintenant, si nous revenons au modèle (2.4), nous devons adapter la structure (2.1) et utiliser les résultats de (2.2) pour obtenir l'EBL de $\boldsymbol{\beta}$ et sa variance associée, $\hat{V}(\hat{\boldsymbol{\beta}})$, donnés respectivement par :

$$\hat{\boldsymbol{\beta}} = \mathbf{a} + \mathbf{R}\mathbf{X}_s'(\mathbf{X}_s\mathbf{R}\mathbf{X}_s' + \mathbf{V}_s)^{-1}(\mathbf{y}_s - \mathbf{X}_s\mathbf{a}) \text{ et } \hat{V}(\hat{\boldsymbol{\beta}}) = \mathbf{C} = \mathbf{R} - \mathbf{R}\mathbf{X}_s'(\mathbf{X}_s\mathbf{R}\mathbf{X}_s' + \mathbf{V}_s)^{-1}\mathbf{X}_s\mathbf{R}. \quad (2.6)$$

Il est facile de voir que, dans (2.6), la première équation peut être réécrite sous la forme $\hat{\boldsymbol{\beta}} = \mathbf{C}(\mathbf{X}_s'\mathbf{V}_s^{-1}\mathbf{y}_s + \mathbf{R}^{-1}\mathbf{a})$, où $\mathbf{C}^{-1} = \mathbf{R}^{-1} + \mathbf{X}_s'\mathbf{V}_s^{-1}\mathbf{X}_s$. Il convient de souligner que, si nous plaçons une loi a priori vague sur $\boldsymbol{\beta}$, en prenant $\mathbf{R}^{-1} \to 0$, nous obtenons l'estimateur par les moindres carrés minimal de $\boldsymbol{\beta}$ : $\hat{\boldsymbol{\beta}}_{LS} = (\mathbf{X}_s'\mathbf{V}_s^{-1}\mathbf{X}_s)^{-1}\mathbf{X}_s'\mathbf{V}_s^{-1}\mathbf{y}_s$.

Maintenant, en appliquant les propriétés bien connues des espérances et des variances conditionnelles, nous obtenons :

$$E[\mathbf{y}_{\bar{s}}|\mathbf{y}_s] = E(E(\mathbf{y}_{\bar{s}}|\mathbf{y}_s,\boldsymbol{\beta})|\mathbf{y}_s) \text{ et } V[\mathbf{y}_{\bar{s}}|\mathbf{y}_s] = E(V(\mathbf{y}_{\bar{s}}|\mathbf{y}_s,\boldsymbol{\beta})|\mathbf{y}_s) + V(E(\mathbf{y}_{\bar{s}}|\mathbf{y}_s,\boldsymbol{\beta})|\mathbf{y}_s). \quad (2.7)$$

En remplaçant $E(\mathbf{y}_{\bar{s}}|\mathbf{y}_s,\boldsymbol{\beta})$ et $V(\mathbf{y}_{\bar{s}}|\mathbf{y}_s,\boldsymbol{\beta})$ dans (2.7) par leur EBL respectif donné en (2.5) et puis, en remplaçant $E(\boldsymbol{\beta}|\mathbf{y}_s)$ et $V(\boldsymbol{\beta}|\mathbf{y}_s)$ par $\hat{\boldsymbol{\beta}}$ et $\hat{V}(\hat{\boldsymbol{\beta}})$ donnés en (2.6), nous obtenons l'EBL de $E[\mathbf{y}_{\bar{s}}|\mathbf{y}_s]$ et sa variance associée sous la forme :

$$\begin{aligned}&\hat{E}[\mathbf{y}_{\bar{s}}|\mathbf{y}_s] = \mathbf{X}_{\bar{s}}\hat{\boldsymbol{\beta}} + \mathbf{V}_{s\bar{s}}\mathbf{V}_s^{-1}(\mathbf{y}_s - \mathbf{X}_s\hat{\boldsymbol{\beta}}) \text{ et}\\ &\hat{V}[\mathbf{y}_{\bar{s}}|\mathbf{y}_s] = \mathbf{V}_{\bar{s}} - \mathbf{V}_{s\bar{s}}\mathbf{V}_s^{-1}\mathbf{V}_{s\bar{s}} + (\mathbf{X}_{\bar{s}} - \mathbf{V}_{s\bar{s}}\mathbf{V}_s^{-1}\mathbf{X}_s)\mathbf{C}(\mathbf{X}_{\bar{s}} - \mathbf{V}_{s\bar{s}}\mathbf{V}_s^{-1}\mathbf{X}_s)'.\end{aligned} \quad (2.8)$$

**Remarque 2** : De manière analogue à la remarque 1, sous l'hypothèse de normalité, nous avons que les membres de droite des équations (2.8) sont, respectivement, les valeurs de $E[\mathbf{y}_{\bar{s}}|\mathbf{y}_s]$ et $V[\mathbf{y}_{\bar{s}}|\mathbf{y}_s]$.

L'expression générale de l'EBL du total $T$ et de sa variance associée est obtenue en remplaçant $E(\mathbf{y}_{\bar{s}}|\mathbf{y}_s)$ et $V(\mathbf{y}_{\bar{s}}|\mathbf{y}_s)$ dans les équations (2.3) par leurs équivalents respectifs $\hat{E}[\mathbf{y}_{\bar{s}}|\mathbf{y}_s]$ et $\hat{V}[\mathbf{y}_{\bar{s}}|\mathbf{y}_s]$ :

$$\hat{T} = \mathbf{1}_s'\mathbf{y}_s + \mathbf{1}_{\bar{s}}'\hat{E}[\mathbf{y}_{\bar{s}}|\mathbf{y}_s] \text{ et } \hat{V}(\hat{T}) = \mathbf{1}_{\bar{s}}'\hat{V}[\mathbf{y}_{\bar{s}}|\mathbf{y}_s]\mathbf{1}_{\bar{s}}. \quad (2.9)$$

Il convient de souligner que, dans de nombreuses applications de (2.9), la matrice $\mathbf{V}$ est supposée être diagonale, ce qui implique que $\mathbf{V}_{\bar{s}\bar{s}} = \mathbf{0}$ et nous avons alors :

$$\hat{T} = \mathbf{1}_s' \mathbf{y}_s + \mathbf{1}_{\bar{s}}' \mathbf{X}_{\bar{s}} \hat{\boldsymbol{\beta}} \text{ et } \hat{V}\left(\hat{T}\right) = \mathbf{1}_{\bar{s}}' \left[\mathbf{V}_{\bar{s}} + \mathbf{X}_{\bar{s}} \mathbf{C} \mathbf{X}_{\bar{s}}'\right] \mathbf{1}_{\bar{s}}. \tag{2.10}$$

En guise d'illustration, nous considérons certains exemples discutés par O'Hagan (1985) et proposons un nouvel estimateur par le ratio, qui est l'une des contributions des présents travaux. Tous les exemples peuvent être traités comme des cas particuliers du modèle (2.4).

## 2.3 Retour sur certains plans de sondage fréquents

### 2.3.1 Échantillonnage aléatoire simple sans remise : échangeabilité

O'Hagan (1985) a examiné le cas simple où la population ne présente aucune structure pertinente, ce qui peut se faire en spécifiant :

$$E(y_i) = m, V(y_i) = v \text{ et } \operatorname{Cov}\left(y_i, y_j\right) = c, i, j = 1, \ldots, N, \forall i \neq j. \tag{2.11}$$

**Remarque 3** : On peut justifier la corrélation introduite dans le modèle (2.11) en invoquant l'imitation de l'échantillonnage aléatoire simple sans remise.

En appliquant le résultat général établi dans (2.10) à (2.11) avec $\boldsymbol{\beta}$ de dimension 1, $\mathbf{X} = \mathbf{1}_N$, $\mathbf{a} = m$, $\mathbf{R} = c$ et $\mathbf{V} = \sigma^2 \mathbf{I}$, où $\sigma^2 = v - c$, nous obtenons l'EBL de $T$ et sa variance associée :

$$\hat{T}_{srs} = n\bar{y}_s + (N - n)\hat{\mu} \text{ et } \hat{V}\left(\hat{T}_{srs}\right) = (N - n)\sigma^2 + (N - n)^2 c\sigma^2 \left(\sigma^2 + nc\right)^{-1}, \tag{2.12}$$

où

$\bar{y}_s = n^{-1} \mathbf{1}_s' \mathbf{y}_s$ est la moyenne d'échantillon,

$\hat{\mu} = \omega \bar{y}_s + (1 - \omega) m$ est la valeur prévue des valeurs non observées de $\mathbf{y}$ et

$\omega = \dfrac{n\sigma^{-2}}{c^{-1} + n\sigma^{-2}}$, où $\sigma^2 = v - c$.

Il convient de souligner que $\hat{\mu}$ est une moyenne pondérée de la moyenne a priori $m$ et de la moyenne d'échantillon $\bar{y}_s$, où $\omega$ est le ratio entre les deux quantités de population. La moyenne $m$ peut être considérée comme le prior du chercheur pour la moyenne de population réelle $\bar{y}$. L'incertitude au sujet de $y_i$ est divisée en deux composantes : l'incertitude au sujet du niveau global des $y_i$ (inter-variation) et l'incertitude quant à la part de chaque $y_i$ qui peut différer de ce niveau global (intra-variation). Une mesure utile de la variabilité des unités dans la population est donnée par

$$S^2 = \frac{1}{N - 1} \sum_{i=1}^{N} (y_i - \bar{y})^2.$$

Il n'est pas difficile de montrer que $E\left(S^2\right) = v - c = \sigma^2$. Par conséquent, $\sigma^2$ peut être interprétée comme une estimation a priori de la variabilité à l'intérieur de la population. Nous obtenons aussi

$V(\bar{y}) = c + N^{-1}\sigma^2$. Dans de nombreuses applications, $N$ est grand et la constante $c$ peut donc être considérée comme l'inter-variation.

En posant que $v \to \infty$ et en maintenant $\sigma^2$ fixe, c'est-à-dire en supposant que l'on ne connaît pas les priors, les estimations dans (2.12) donnent :

$$\hat{T}_{srs} = N\,\bar{y}_s \text{ et } \hat{V}\left(\hat{T}_{srs}\right) = N^2\left(1 - \frac{n}{N}\right)\frac{\sigma^2}{n}.$$

Ces expressions sont fort semblables à l'estimation bien connue du total et à sa variance dans le contexte fondé sur le plan de sondage pour le cas de l'échantillonnage aléatoire simple. O'Hagan (1985) a discuté de certains moyens possibles d'éviter la tâche difficile d'attribuer une valeur à $\sigma^2$. Le moyen le plus naturel de le faire consiste à trouver son EBL, mais linéaire en les carrés et les termes de variance des produits croisés. Cependant, il est nécessaire de spécifier les moments d'ordre quatre des $y_i$. Goldstein (1979) a proposé un EBL de la variance qui n'utilise que des fonctions linéaires des données. Néanmoins, on obtient une expression compliquée de la variance associée de son EBL modifié. O'Hagan (1985) a soutenu que, si l'information a priori au sujet des composantes de la variance est faible, toute estimation a posteriori s'approche des estimations non bayésiennes classiques obtenues en utilisant uniquement les données, lorsque ce gendre d'estimations est disponible. Par conséquent, il a proposé, en guise de procédure bayésienne approximative, d'introduire ces estimations de variance classiques par substitution dans l'EBL et dans sa variance associée lorsque cela est approprié. Dans le cas qui nous occupe, nous pouvons remplacer $\sigma^2$ par $s^2 = (n-1)^{-1}\sum_{i=1}^{n}(y_i - \bar{y}_s)^2$, qui est sans biais sous le plan pour $S^2$.

## 2.3.2 Échantillonnage aléatoire simple stratifié sans remise

Soit $y_{hi}$ la $i^{\text{e}}$ unité, $i = 1,...,N_h$ appartenant à la strate $h$, $h = 1,..,H$. Nous supposons que la taille de strate, $N_h$, est connue pour toutes les strates. L'échangeabilité d'ordre deux dans chaque strate est énoncée dans O'Hagan (1985) sous la forme :

$$E(y_{hi}) = m_h, V(y_{hi}) = v_h, \text{Cov}(y_{hi}, y_{hj}) = c_h, i \neq j \text{ et } \text{Cov}(y_{hi}, y_{lj}) = d_{hl}, h \neq l.$$

**Remarque 4 :** Il est raisonnable de supposer que l'information obtenue au sujet d'une strate pourrait modifier les croyances au sujet des autres strates dans certaines applications spéciales. Cependant, si nous voulons imiter l'échantillonnage aléatoire simple stratifié, nous devons supposer que les observations dans les diverses strates ne sont pas corrélées, en posant que $d_{hl} = 0$.

Le modèle général (2.4) peut être appliqué à ce cas en prenant $\mathbf{X} = \text{diag}(\mathbf{X}_1, \ldots, \mathbf{X}_H)$ et $\mathbf{V} = \text{diag}(\mathbf{V}_1, \ldots, \mathbf{V}_H)$, avec $\mathbf{X}_h = \mathbf{1}_{N_h}$ et $\mathbf{V}_h = \sigma_h^2 \mathbf{I}_{N_h}$, où $\sigma_h^2 = v_h - c_h, \forall\, h = 1, \ldots, H$, $\mathbf{a} = (m_1, \ldots, m_H)'$, $\mathbf{R}$ est une matrice de dimensions $H \times H$ avec $R_{hl} = c_h$, si $h = l$ et $R_{hl} = d_{hl}$ autrement. L'EBL de $T$ et sa variance associée sont obtenus au moyen de (2.10) et figurent dans O'Hagan (1985). Les modèles pour l'échantillonnage en grappes sont donnés dans Bolfarine et Zacks (1992), page 11. L'EBL des modèles avec grappes figurent dans O'Hagan (1985).

# 3 Information auxiliaire : estimateur par le ratio

Dans de nombreuses situations pratiques, il est possible d'obtenir de l'information au sujet d'une variable auxiliaire $x_i$ (corrélée à $y_i$) pour toutes les unités de la population, ou du moins pour chaque unité échantillonnée, ainsi que la moyenne de population $\bar{X}$. En pratique, $x_i$ est souvent la valeur de $y_i$ à une période antérieure où a été effectué un recensement complet. Cette approche est utilisée dans des situations où la valeur prévue et la variance de $y_i$ sont proportionnelles à $x_i$, de sorte que dans la spécification de l'EBL, nous remplaçons certaines hypothèses au sujet des $y$ par des hypothèses au sujet des deux premiers moments du taux $y_i / x_i$. Autant que nous sachions, le nouvel estimateur par le ratio proposé ci-après est une nouvelle contribution à la théorie des sondages.

Le nouvel estimateur par le ratio est obtenu en tant que cas particulier du modèle (2.4) et en se servant de l'hypothèse d'échangeabilité, utilisée dans l'approche bayésienne linéaire, appliquée au taux $y_i / x_i$ pour tout $i = 1, \ldots, N,$ comme il est décrit ci-dessous :

$$E\left(\frac{y_i}{x_i}\right) = m, V\left(\frac{y_i}{x_i}\right) = v \text{ et } \operatorname{Cov}\left(\frac{y_i}{x_i}, \frac{y_j}{x_j}\right) = c, i, j = 1, \ldots, N, \forall i \neq j. \tag{3.1}$$

En appliquant le résultat général établi dans (2.10) à (3.1) avec $\mathbf{X} = (x_1, \ldots, x_N)'$, le vecteur $N \times 1$ de dimension $\mathbf{a} = m,$ de variables auxiliaires $\mathbf{R} = c$ et $\mathbf{V} = \sigma^2 \operatorname{diag}(x_1, \ldots, x_N)$, où $\sigma^2 = v - c,$ nous obtenons l'EBL de T et sa variance associée comme il suit :

$$\begin{aligned}
\hat{T}_{ra} &= n\bar{y}_s + (N - n)\hat{\mu}\bar{x}_{\bar{s}} \text{ et} \\
\hat{V}\left(\hat{T}_{ra}\right) &= (N - n)\bar{x}_{\bar{s}}\sigma^2 + (N - n)^2 \bar{x}_{\bar{s}}{}^2 \left(c^{-1} + \sigma^{-2} n\bar{x}_s\right)^{-1}, \text{ où} \\
\hat{\mu} &= \omega \frac{\bar{y}_s}{\bar{x}_s} + (1 - \omega) m \text{ et } \omega = \frac{\sigma^{-2} n\bar{x}_s}{\left(c^{-1} + \sigma^{-2} n\bar{x}_s\right)},
\end{aligned}$$

où $\bar{x}_{\bar{s}} = (N\bar{X} - n\bar{x}_s)/(N - n)$ est la moyenne des $x$ pour les unités non échantillonnées. En posant que $v \to \infty$ et $n \to \infty,$ mais en maintenant $\sigma^2$ fixe, nous réobtenons l'estimateur de type ratio de l'approche fondée sur le plan de sondage : $\hat{T}_{ra} = N\bar{X}(\bar{y}_s / \bar{x}_s)$.

# 4 Méthode bayésienne linéaire pour données catégoriques

Il arrive souvent que l'on s'intéresse à des cas où la caractéristique observée est celle de savoir si l'unité de population possède ou non un certain attribut d'intérêt. Nous pouvons définir une variable dichotomique $y_i = 1,$ si la $i^{\text{e}}$ unité possède cet attribut, ce qui est désigné comme une réussite, et $y_i = 0$ autrement. Pour le cas binaire lorsque la taille de l'échantillon n'est pas suffisamment grande pour appliquer le théorème central limite, l'approche fondée sur le plan de sondage pourrait faire appel à la randomisation introduite par le plan de sondage pour justifier la distribution des quantités aléatoires binaires. Par exemple, Cochran (1977), sections 3.4 et 3.5, montre comment appliquer les lois hypergéométrique et binomiale pour obtenir les intervalles de confiance pour les proportions de population quand on se sert de plans d'échantillonnages aléatoires simples avec et sans remise, respectivement. Par ailleurs, des approches dépendantes d'un modèle ont également été avancées et

appliquées pour prédire les totaux ou les moyennes dans les catégories d'intérêt. Malec, Sedransk, Moriarity et LeClere (1997) ont considéré un modèle hiérarchique logistique à deux niveaux, où les grappes forment le deuxième niveau. Ils ont également comparé les estimations bayésiennes entièrement hiérarchiques aux estimations bayésiennes empiriques et aux méthodes classiques. Moura et Migon (2002) ont présenté une approche basée sur un modèle hiérarchique logistique pour la prédiction de proportions sur petits domaines, en tenant compte des effets spatiaux ainsi que des effets d'hétérogénéité non structurée possibles. Nandram et Choi (2008) ont proposé un modèle multinomial-Dirichlet dépendant du temps pour prédire les résultats d'une élection sous non-réponse ignorable et non ignorable. Ils ont également utilisé une approche bayésienne pour répartir les électeurs indécis entre les candidats.

De nouveau, ici, nous n'avons pas besoin d'utiliser des hypothèses au sujet du modèle complet ni une approche de randomisation, mais nous devons émettre certaines hypothèses au sujet des premier et deuxième moments des quantités aléatoires concernées. L'EBL pour les données binaires a été introduit brièvement par O'Hagan (1985), mais ici, nous le développons d'une manière plus générale pour le cas où nous nous intéressons à l'analyse de plus d'un attribut dans une population. L'objectif est de décrire l'estimation de la proportion de réussites avec des données catégoriques. Soit $y_{ij}$ la variable qui indique que l'unité $i,\ i = 1,\ldots,N$ se trouve dans la catégorie $j,\ j = 1,\ldots,k$ donnée par

$$y_{ij} = \begin{cases} 1, \text{ si la } i^{\text{e}} \text{ unité possède le } j^{\text{e}} \text{ attribut;} \\ 0, \text{ autrement.} \end{cases}$$

Le but principal est d'estimer un vecteur $\mathbf{p} = (p_1,\ldots,p_k)'$, où $p_j = N^{-1}\sum_{i=1}^{N} y_{ij},\ j = 1,\ldots,k,$ est la proportion d'unités dans la catégorie $j,$ sachant $\mathbf{y}_s,$ un vecteur de dimension $nk,$ défini comme étant $\mathbf{y}_s = (y_{11}, y_{21},\ldots, y_{n1},\ldots, y_{1k}, y_{2k},\ldots, y_{nk})'.$ Comme nous avons affaire à des situations dans lesquelles il n'est possible d'associer qu'un seul attribut à chaque unité, nous avons $\sum_{j=1}^{k} p_j = 1.$ Donc, nous ne devons estimer que $k-1$ paramètres, puisqu'il s'ensuit que $\hat{p}_k = 1 - \sum_{j=1}^{k-1}\hat{p}_j$ et que l'estimation de la variance est également obtenue de manière analogue par $\hat{V}(\hat{p}_k) = \sum_{j=1}^{k-1}\hat{V}(\hat{p}_j) + \sum_{l\neq j=1}^{k-1}\hat{\text{Cov}}(\hat{p}_j, \hat{p}_l).$

En l'absence de toute autre information structurelle, nous supposons que les unités dans une catégorie donnée sont échangeables d'ordre deux, mais nous ne supposons aucune échangeabilité entre les unités de différentes catégories. Nos croyances a priori sont exprimées pour $i = 1,\ldots,N,\ j = 1,\ldots,k-1,$ comme il suit :

$$\begin{aligned} m_j &= E(y_{ij}) = P(y_{ij} = 1), v_j = V(y_{ij}) = m_j(1-m_j) \text{ et} \\ \text{cov}(y_{ij}, y_{i'j}) &= P(y_{i'j} = 1 \,|\, y_{ij} = 1)P(y_{ij} = 1) - P(y_{ij} = 1)P(y_{i'j} = 1) \\ &= m_j(m_{jj} - m_j) = c_j,\ \forall i \neq i' \text{ et } \sigma_j^2 = v_j - c_j = m_j(1-m_{jj}), \end{aligned}$$

où $m_{jj} = P(y_{i'j} = 1 \,|\, y_{ij} = 1),$ pour tout $i \neq i'.$

Pour $j \neq j',$ nous obtenons de manière analogue la covariance entre ces catégories sous la forme

$$\text{cov}(y_{ij}, y_{i'j'}) = \begin{cases} m_j(m_{j'j} - m_{j'}), & \text{si } i \neq i', \\ -m_j m_{j'}, & \text{si } i = i'. \end{cases}$$

Souvent, nous ne possédons pas toutes les données $\mathbf{y}_s$, mais seulement une statistique exhaustive, comme la proportion dans l'échantillon pour chaque catégorie, $\bar{\mathbf{y}}_s$. Soit $\bar{\mathbf{y}}_s$ le vecteur de dimension $k-1$- dont la $j^{\text{e}}$ position est donnée par la moyenne d'échantillon pour la catégorie $j$. En utilisant le modèle général donné par (2.4), nous obtenons :

$$E(\bar{\mathbf{y}}_s) = E(E(\bar{\mathbf{y}}_s \mid \boldsymbol{\beta})) = \mathbf{a} \text{ et } \operatorname{Var}(\bar{\mathbf{y}}_s) = E(V(\bar{\mathbf{y}}_s \mid \boldsymbol{\beta})) + V(E(\bar{\mathbf{y}}_s \mid \boldsymbol{\beta})) = \mathbf{V}_s + \mathbf{R}.$$

En appliquant le modèle général donné dans (2.4), où la variable de réponse est donnée par $\bar{\mathbf{y}}_s$, le vecteur $\boldsymbol{\beta}$ est de dimension $k-1$, $\mathbf{X}_s = \mathbf{I}_s$ et $\mathbf{V} = \operatorname{diag}(\mathbf{V}_{\bar{s}}, \mathbf{V}_s)$, nous obtenons à partir de (2.10) :

$$\hat{\mathbf{p}} = \frac{n\bar{\mathbf{y}}_s + (N-n)\hat{\boldsymbol{\beta}}}{N} \text{ et } \hat{V}(\hat{\mathbf{p}}) = \frac{(N-n)^2[\mathbf{V}_{\bar{s}} + \mathbf{C}]}{N^2}, \tag{4.1}$$

où $\mathbf{C}^{-1} = \mathbf{R}^{-1} + \mathbf{V}_s$ et $\hat{\boldsymbol{\beta}} = \mathbf{C}\left(\mathbf{V}_s^{-1}\bar{\mathbf{y}}_s + \mathbf{R}^{-1}\mathbf{a}\right)$, comme il est énoncé en (2.6).

Soit $\mathbf{Q} = \mathbf{V}_s + \mathbf{R}$. L'EBL de $\mathbf{p}$ et sa variance associée donnés par (4.1) peuvent s'écrire en fonction des quantités a priori $m_j$, $m_{jj'}$ et $j = 1, \ldots, k-1$, en notant que $\mathbf{a} = (m_1, \ldots, m_{k-1})'$, $Q_{jj} = c_j + \sigma_j^2/n$ et $Q_{jj'} = m_j(m_{j'j} - m_{j'}) - m_j m_{j'j}/n$. Par conséquent, la matrice $\mathbf{R} = \{r_{jj'}\}, j, j' = 1, .., k-1$ avec $r_{jj} = c_j$ et $r_{jj'} = m_j(m_{j'j} - m_{j'})$ et $\mathbf{V}_s = 1/n\{v_{jj'}\}$, $j, j' = 1, .., k-1$ avec $v_{jj} = \sigma_j^2$ et $v_{jj'} = -m_j m_{j'j}$. De manière analogue, nous obtenons $\mathbf{V}_{\bar{s}} = n/(N-n)\mathbf{V}_s$.

## 4.1 Obtention des priors

L'obtention des priors est le processus consistant à formuler les connaissances et les croyances d'une personne au sujet d'une ou de plusieurs quantités incertaines sous forme d'une loi de probabilité pour ces quantités. Selon Garthwaite, Kadane et O'Hagan (2005), il est commode de concevoir la tâche d'obtention des priors comme faisant intervenir un facilitateur qui aide l'expert à formuler ses connaissances spécialisées sous forme probabiliste. Dans le contexte de l'obtention d'une loi a priori pour une analyse bayésienne, ce sont les connaissances a priori de l'expert qui sont tirées au clair, mais en général, l'objectif est d'exprimer les connaissances courantes de l'expert sous forme probabiliste. Si l'expert est un statisticien ou s'il connaît très bien les concepts statistiques, l'intervention d'un facilitateur pourrait ne pas être formellement nécessaire, mais cela est rare en pratique. O'Hagan (1998) a illustré au moyen d'un exemple pratique comment obtenir les premier et deuxième moments. En particulier, il a adopté l'approche bayésienne linéaire parce qu'elle permet aux ingénieurs d'appliquer facilement une procédure d'obtention des priors.

À la présente section, nous présentons certaines contraintes concernant les quantités a priori et une solution de rechange pour faciliter le processus d'obtention des priors en vue d'obtenir l'EBL pour des données catégoriques. Comme $m_j$ et $m_{jj'}$ sont des probabilités et que $\mathbf{R}$ et $\mathbf{V}_s$ sont les matrices de covariance dans le modèle (2.4), les contraintes qui suivent doivent être satisfaites :

1. $0 < m_j < 1$ et $0 \le m_{jj'} \le 1$, $j, j' = 1, \ldots, k-1$;
2. $\mathbf{R}$ et $\mathbf{V}_s$ sont des matrices symétriques définies positives.

Afin de vérifier si la condition (2.2) est satisfaite, on peut exécuter les étapes suivantes :

i. vérifier si $\mathbf{R}$ et $\mathbf{V}_s$ sont symétriques en vérifiant que $m_j m_{jj'} = m_{j'} m_{j'j}$;

ii. vérifier si $\mathbf{R}$ et $\mathbf{V}_s$ sont des matrices définies positives en trouvant les valeurs propres de $\mathbf{R}$ et $\mathbf{V}_s$. Si les valeurs propres sont positives, alors les matrices sont définies positives.

Il convient de mentionner que les valeurs propres sont les racines du polynôme caractéristique et que si ce polynôme est de degré $n, n \leq 4,$ il est possible d'obtenir analytiquement ses racines en appliquant Bhaskara, Cardan ou Ferrari; voir Jacobson (2009), chapitre 4, pour les formules. Cependant, si $n \geq 5,$ il est habituellement nécessaire d'appliquer une méthode itérative pour les obtenir. Néanmoins, pour les matrices de dimensions supérieures à $2 \times 2,$ il n'est pas simple d'obtenir analytiquement ces contraintes en se basant sur les valeurs propres. La proposition qui suit présente les conditions que $m_j$ et $m_{jj'}$, $j = 1, \ldots, k-1,$ doivent satisfaire afin d'obtenir un prior convenable pour un modèle multinomial comprenant trois catégories en utilisant l'approche d'estimation bayésienne linéaire.

**Proposition 1** Supposons que nous obtenons $m_j,$ tel que $0 < m_j < 1,\ j = 1, 2.$ Alors, sachant $\rho_{11}, \rho_{12}$ et $\rho_{22},$ nous obtenons $m_{11}, m_{12}, m_{21}$ et $m_{22}$ au moyen de (4.2). Les quantités a priori $m_j$ et $m_{jj'},$ pour $j, j' = 1, 2,$ doivent satisfaire les contraintes qui suivent pour que les matrices $\mathbf{R}$ et $\mathbf{V}_s$ soient définies positives :

$$m_{11} > m_1 \ et \ m_{22} > m_2, m_{11}m_{22} - m_{11} - m_{22} + 1 > m_{12}m_{21} \text{ et}$$
$$m_{11}m_{22} - m_{11}m_2 - m_1 m_{22} > m_{12}m_{21} - 2m_2 m_{12}.$$

La vérification de la proposition 1 nécessite certaines opérations algébriques. Nous vérifions que les matrices $\mathbf{R}$ et $\mathbf{V}_s$ sont définies positives en utilisant (i) et (ii) susmentionnés. Nous faisons appel au fait que les valeurs propres d'une matrice de dimensions $2 \times 2$ sont positives si et seulement si son déterminant est positif et nous obtenons alors $m_{jj'},\ j, j' = 1, 2$ qui satisfait cette contrainte pour les deux matrices. Pour les cas comprenant plus de trois catégories, nous devons vérifier numériquement si les matrices $\mathbf{R}$ et $\mathbf{V}_s$ sont définies positives en remplaçant $m_j$ et $m_{jj'},\ j = 1, \ldots, k-1$ par leur valeur numérique.

Par ailleurs, si un expert a de la difficulté à spécifier certaines de ces probabilités conditionnelles $m_{jj'},$ il pourrait être plus simple d'attribuer un prior au coefficient de corrélation. Définissons $\rho_{jj'}$ comme étant le prior du coefficient de corrélation entre deux unités différentes dans les catégories $j$ et $j',$ c'est-à-dire :

$$\rho_{jj'} = \text{corr}\left(y_{ij}, y_{i'j'}\right) = \begin{cases} \dfrac{m_{jj} - m_j}{1 - m_j}, & j = j', \\ \dfrac{m_j\left(m_{j'j} - m_{j'}\right)}{\sqrt{m_j\left(1 - m_j\right)m_{j'}\left(1 - m_{j'}\right)}}, & j \neq j', \end{cases}$$

pour $i, i' = 1, \ldots, n,\ i \neq i',\ j, j' = 1, \ldots, k-1.$

Par conséquent, sachant $\rho_{jj'},\ j, j' = 1,\ldots,k-1,$ nous obtenons

$$m_{jj'} = \begin{cases} m_j + \rho_{jj}\left(1 - m_j\right) & j = j', \\ \dfrac{m_j m_{j'} + \rho_{j'j}\sqrt{m_j\left(1-m_j\right)m_{j'}\left(1-m_{j'}\right)}}{m_{j'}}, & j \neq j'. \end{cases} \tag{4.2}$$

Il convient de mentionner que, si l'on dispose de données provenant d'une enquête antérieure, il est possible qu'un expert utilise cette information. Par exemple, $m_j$ peut être obtenu en estimant la proportion d'unités dans la catégorie $j,\ j = 1,\ldots,k-1$ à partir de l'enquête antérieure. De façon analogue, $\rho_{jj'}$ peut être obtenu en utilisant les données d'une enquête antérieure. Comme l'indique la contrainte (2.1), $m_j$ ne peut pas prendre les valeurs 0 et 1, sinon les corrélations ne seraient pas définies.

## 4.2 Analyse de la sensibilité aux priors

Il est utile de vérifier si l'estimateur et sa variance associée dépendent des priors attribués. Nous traitons le cas simple ne comprenant que deux catégories. Soulignons que, dans le cas où il y a plus de deux catégories, le nombre de quantités a priori qu'il faut obtenir augmente rapidement, mais que l'on peut étendre les conclusions obtenues. Par ailleurs, en l'absence d'information a priori, nous pouvons utiliser des priors non informatifs et, comme il est décrit à la section 2.2, on retrouve alors les estimateurs de l'approche fondée sur le plan de sondage.

L'EBL pour la proportion en cas de données binaires peut être obtenu en tant que cas particulier de l'estimateur (4.1),

$$\hat{p}_1 = \frac{n\bar{y}_1 + (N-n)\hat{\mu}}{N},$$

où

$\hat{\mu} = \omega\bar{y}_1 + (1-\omega)m_1$ est la valeur prévue des valeurs non observées dans la catégorie 1,

$\omega = \dfrac{n\sigma_1^{-2}}{n\sigma_1^{-2} + c_1^{-1}}$,

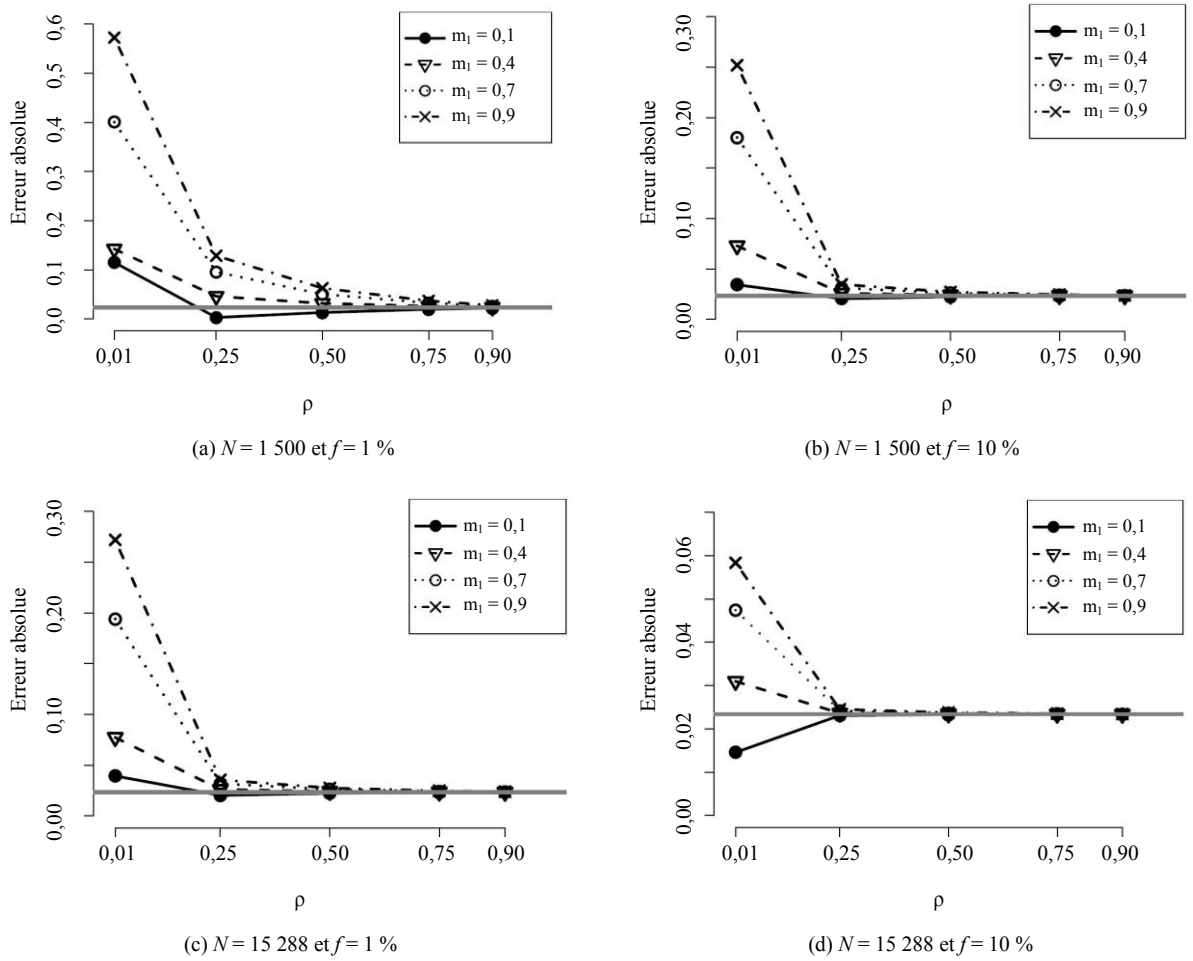
et $\hat{p}_2 = 1 - \hat{p}_1$. Notons que $\sigma_1^2$ et $c_1$ dépendent de $m_{11} = m_1 + \rho_{11}(1-m_1)$, voir page 13. Nous analysons comment les estimations sont affectées par $\rho_{11}$.

1. Si $\rho_{11} \to 0$, alors $\omega \to 0$ et $\hat{\mu} \to m_1$. Donc, l'estimateur pour les valeurs non observées dépend en grande partie de la valeur du prior.
2. Si $\rho_{11} \to 1$, alors $\omega \to 1$ et $\hat{\mu} \to \bar{y}_1$. Donc, l'estimateur pour les valeurs non observées ne dépend pas de la valeur du prior.

En outre, il est facile de voir que $n/N \to 1,\ \hat{p}_1 \to \bar{y}_1$. Pour illustrer ces résultats, nous avons créé un jeu de données artificielles en fixant la proportion réelle à $\mathbf{p} = (0{,}2380; 0{,}7620)'$ et la moyenne d'échantillon à $\bar{\mathbf{y}}_s = (0{,}2614; 0{,}7386)'$. Ces valeurs ont été tirées de Moura et Migon (2002). Puis, nous avons déterminé comment les valeurs de $m_1,\ N,\ f = n/N$ et $\rho_{11}$ affectent l'estimateur $\hat{p}_1$. La

figure 4.1 donne la représentation graphique en deux dimensions de l'erreur absolue de $\hat{p}_1$ en fonction de $\rho_{11}$ pour certains cas particuliers. La courbe grise représente l'erreur absolue entre la proportion d'échantillon $\bar{y}_1$ et la proportion réelle $p_1$.

Il faut souligner que, à mesure que $f$ ou $N$ augmente, l'erreur absolue diminue pour toute valeur du prior. De surcroît, quand $\rho_{11} \to 0$, l'erreur absolue augmente quand $m_1$ diffère considérablement de la proportion réelle $p_1$, mais elle diminue à mesure que la taille de l'échantillon augmente. Enfin, quand $\rho_{11} \to 1$, nous observons que l'erreur absolue de $\hat{p}_1$ tend vers l'erreur absolue de la proportion d'échantillon $\bar{y}_1$. Donc, si nous avons une bonne information a priori, en ce qui concerne $m_1$, l'estimateur proposé donne de bons résultats pour toutes les valeurs de $\rho_{11}$. Cependant, si aucune information a priori n'est disponible, des priors non informatifs caractérisés par $\rho_{11} \to 1$ peuvent être utilisés et nous obtenons des résultats similaires à ceux de l'approche fondée sur le plan de sondage.




(a) $N$ = 1 500 et $f$ = 1 %

(b) $N$ = 1 500 et $f$ = 10 %

(c) $N$ = 15 288 et $f$ = 1 %

(d) $N$ = 15 288 et $f$ = 10 %

**Figure 4.1 Erreur absolue pour $m_1 \in \{0{,}1; 0{,}4; 0{,}7; 0{,}9\}$, $N \in \{1\,500, 15\,288\}$ et $f \in \{1\,\%, 10\,\%\}$ fixes et $\rho_{11} \in \{0{,}01; 0{,}25; 0{,}5; 0{,}75; 0{,}9\}$ variable. La courbe grise représente l'erreur absolue de la proportion d'échantillon $\bar{y}_1$**

# 5 Conclusion

Obtenir une loi a priori conjointe complète pour un grand nombre de dimensions représenterait une tâche énorme. La méthode bayésienne linéaire nécessite seulement l'obtention des moyennes, des variances et des covariances a priori des paramètres. Elle est particulièrement utile lorsqu'un statisticien n'est pas disponible pour procéder à l'obtention complète des priors. Un exemple d'obtention réussie des priors en utilisant cet estimateur figure dans O'Hagan (1998).

Nous avons dérivé les estimateurs fondés sur le plan de sondage bien connus en appliquant la structure de l'EBL à une approche faisant appel à un modèle de régression généralisée. Nous avons étendu l'estimateur aux données catégoriques et conclu que même si cet estimateur nécessite l'obtention des priors de nombreuses quantités, il est possible de les reparamétriser ou de travailler avec des priors non informatifs. L'exemple numérique illustre le comportement des estimations sous forme d'une fonction de la taille de l'échantillon et des spécifications des paramètres a priori. Cependant, nous sommes conscients que l'obtention des priors pour un grand nombre des paramètres n'est pas une tâche facile si l'on ne dispose pas d'information provenant d'enquêtes antérieures. Néanmoins, les exemples dont il est question dans l'article montrent que même si on ne dispose pas d'information a priori, il est possible d'obtenir les estimateurs fondés sur le plan de sondage équivalent en fixant une variance suffisamment grande pour les priors. En outre, les praticiens des sondages qui doivent obtenir des estimations pour un grand nombre de variables se rendraient également compte qu'ils ne seraient pas capables de produire des estimations d'une précision satisfaisante pour toutes les variables, quelle que soit l'approche utilisée. Enfin, nous avons montré comment l'approche de l'EBL et l'approche fondée sur le plan de sondage peuvent être conciliées.

# Remerciements

Les présents travaux ont été réalisés dans le cadre de la thèse de maîtrise de Kelly C.M. Gonçalves sous la supervision de Fernando Moura et de Helio Migon, dans le programme d'études supérieures de l'UFRG. Kelly est titulaire d'une bourse décernée par le Coordenação de Aperfeiçoamento de Pessoal do Ensino Superior (CAPES). Fernando Moura et Helio Migon ont reçu l'appui financier du Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq-Brazil, BPPesq).

Les auteurs remercient le rédacteur en chef, le rédacteur associé et deux examinateurs de leurs commentaires très constructifs.

# Bibliographie

Basu, D. (1971). An essay on the logical foundations of survey sampling, Part 1 (avec discussion). Dans *Foundations of Statistical Inference*, (Éds., Godambe et Sprott), 203-242. Holt, Reinhart and Wilnston, Toronto.

Bolfarine, H., et Zacks, S. (1992). *Prediction Theory for Finite Populations*. New York : Springer-Verlag.

Cochran, W. (1977). *Sampling Techniques*. New York : John Wiley & Sons, Inc.

Garthwaite, P., Kadane, J. et O'Hagan, A. (2005). Statistical methods for eliciting probability distributions. *Journal of the American Statistical Association*, 100, 680-701.

Goldstein, M., et Wooff, D. (2007). *Bayes Linear Statistics: Theory and Methods*. Durham University, UK : Wiley series in probability and statistics.

Goldstein, M. (1979). The variance modified linear Bayes estimator. *Journal of the Royal Statistical Society*, 41, 96-100.

Hartigan, J. (1969). Linear bayesian methods. *Journal of the Royal Statistical Society*, Séries B (Méthodologique), 446-454.

Horvitz, D., et Thompson, D. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47, 663-685.

Jacobson, N. (2009). *Basic Algebra,* Vol 1. Dover Books on Mathematics.

Little, R.J. (2003). The Bayesian approch to sample survey inference. Dans *Analysis of Survey Data*, (Éds., R.L. Chambers et C.J. Skinner), chapitre 4, 49-52. New York : John Wiley & Sons Inc.

Malec, D., Sedransk, J., Moriarity, C.L. et LeClere, F.B. (1997). Small area inference for binary variables in National Health Interview Survey. *Journal of the American Statistical Association*, 92, 815-826.

Moura, F., et Migon, H. (2002). Bayesian spatial models for small area estimation of proportions. *Statistical Modelling*, 2, 183-201.

Nandram, B., et Choi, J. (2008). Une répartition bayésienne des électeurs indécis. *Techniques d'enquête*, 34, 1, 41-54.

Neyman, J. (1934). On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, 97, 558-625.

O'Hagan, A. (1985). Bayes linear estimators for finite populations. *Rapport technique 58*, Department of Statistics - University of Warwick.

O'Hagan, A. (1987). Bayes linear estimators for randomized response models. *Journal of the American Statistical Association*, 82, 580-585.

O'Hagan, A. (1998). Eliciting expert beliefs in substantial practical applications. *The Statistician*, 47, 21-35.

Pfeffermann, D., Moura, F.A.S. et Silva, P.L.N. (2006). Multi-level modelling under informative sampling. *Biometrika*, 93, 943.

Scott, A.J. (1977). Large-sample posterior distributions for finite populations. *Annals of Mathematical Statistics*, 42, 1113-1117.

Skinner, C., Holt, D. et Smith, T. (1989). *Analysis of complex surveys*. New York : John Wiley & Sons, Inc.

Smouse, E. (1984). A note on bayesian least squares inference for finite population models. *Journal of the American Statistical Association*, 79, 390-392.

Valliant, R., Dorfman, A. et Royall, R. (2000). *Finite Population Sampling and Inference: A Prediction Approach*. New York : John Wiley & Sons, Inc.

Zacks, S. (2002). In the footsteps of Basu: The predictive modelling approach to sampling from finite population. *Sankhyā: The Indian Journal of Statistics*, Séries A, 64, 532-544.