Nº 12-001-X au catalogue
ISSN 1712-5685

Techniques d’enquête

Bayes, étayé par des idées fondées sur le plan, est le meilleur paradigme global pour l’inférence en enquête par échantillonnage

par Roderick J. Little

Date de diffusion : le 15 décembre 2022

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 suivants :

- Service de renseignements statistiques 1-800-263-1136
- Service national d'appareils de télécommunications pour les malentendants 1-800-363-7629
- Télécopieur 1-514-283-9350

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 « Contactez-nous » > « Normes de service à la clientèle ».

Note de reconnaissance

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

Une version HTML est aussi disponible.

This publication is also available in English.

Techniques d'enquête, décembre 2022
Vol. 48, nº 2, pp. 279-306
Statistique Canada, nº 12-001-X au catalogue

Bayes, étayé par des idées fondées sur le plan, est le meilleur paradigme global pour l'inférence en enquête par échantillonnage

Roderick J. Little[1]

Résumé

Des arguments conceptuels et des exemples sont présentés qui suggèrent que l'approche d'inférence bayésienne pour les enquêtes permet de répondre aux défis nombreux et variés de l'analyse d'une enquête. Les modèles bayésiens qui intègrent des caractéristiques du plan de sondage complexe peuvent donner lieu à des inférences pertinentes pour l'ensemble de données observé, tout en ayant de bonnes propriétés d'échantillonnage répété. Les exemples portent essentiellement sur le rôle des variables auxiliaires et des poids d'échantillonnage, et les méthodes utilisées pour gérer la non-réponse. Le présent article propose 10 raisons principales de favoriser l'approche d'inférence bayésienne pour les enquêtes.

Mots-clés : Inférence bayésienne calée; inférence fondée sur le plan de sondage; modèles de mélange de schémas d'observation; modèles de superpopulation; pondération d'enquête; post-stratification; échantillonnage avec probabilité proportionnelle à la taille; propension à répondre; splines pénalisés.

1. Introduction

À mon avis, les inférences bayésiennes constituent le meilleur paradigme inférentiel pour l'inférence statistique à partir d'enquêtes, qu'il s'agisse d'échantillonnages probabilistes ou non probabilistes. Consultez, à titre d'exemple, Ericson (1969), Binder (1982), Rubin (1987), Ghosh et Meeden (1997), Little (2003ab, 2004, 2012, 2015), Sedransk (2008) et Fienberg (2011). Cependant, les propriétés fondées sur le plan des inférences bayésiennes sont importantes parce que « tous les modèles sont incorrects » et qu'une vaste acceptation des résultats nécessite des inférences ayant de bonnes caractéristiques d'exploitation dans un échantillonnage répété. En particulier, les modèles bayésiens doivent intégrer des caractéristiques complexes du plan de sondage pour donner lieu à des inférences approximativement calées, c'est-à-dire des intervalles de crédibilité qui sont près des niveaux nominaux lorsqu'ils sont traités comme des intervalles de confiance dans un échantillonnage répété (Rubin, 1984, 2019; Little, 2006). Dans les grands échantillons, les modèles de travail flexibles permettent d'éviter les hypothèses fortes au sujet des paramètres qui mèneraient à des estimations potentiellement biaisées.

Pour orienter les propos, considérons le problème de dérivation d'une estimation ponctuelle q d'une quantité de population finie Q, et d'une estimation par intervalles à 95 % $I_{0,95} = (l, u)$ qui saisit une incertitude de q; l'intervalle pourrait avoir une interprétation fréquentiste comme un intervalle de confiance à 95 %, ou une interprétation bayésienne comme un intervalle de crédibilité *a posteriori* à 95 % pour Q. Je crois que les scientifiques qui ne sont pas statisticiens ont l'habitude d'interpréter l'intervalle I de manière bayésienne, en tant qu'intervalle fixe saisissant l'incertitude à propos de Q. Je ne me

1. Roderick J. Little, Département de biostatistique, University of Michigan. Courriel : rlittle@umich.edu.

concentre toutefois pas indûment sur les différentes interprétations de $I_{0.95}$ selon les deux paradigmes. La valeur nominale de 95 % est établie par convention, et d'autres niveaux peuvent être choisis.

L'échantillonnage de la population finie présente une caractéristique intéressante qui consiste à travailler avec des quantités réelles (bien qu'inconnues). En cas d'inférence par enquête « analytique » où l'accent est mis sur les paramètres des modèles idéalisés de la population, comme les coefficients de régression dans un modèle de régressions multiples, établissons la quantité de population finie Q en tant qu'estimation du paramètre d'intérêt si le modèle était ajusté aux données pour toute la population, selon une certaine méthode d'ajustement convenue comme les moindres carrés ou le maximum de vraisemblance. Une des caractéristiques utiles de cette construction est que Q est une quantité réelle au lieu d'être la caractéristique d'un modèle hypothétique simplifié de la population (par exemple Little, 2004).

Dans le paradigme bayésien, l'inférence pour Q repose sur sa distribution prédictive *a posteriori* en fonction des données, pour des choix judicieux de modèle et de distribution *a priori* concernant des paramètres inconnus. Par conséquent, q pourrait constituer la moyenne prédictive *a posteriori* de Q, et $I_{0.95}$, la valeur du 2,5^{e} au 97,5^{e} centile de la distribution prédictive *a posteriori*, ou les limites de la fourchette de valeurs de Q pour la densité *a posteriori* la plus élevée, en supposant que la distribution prédictive *a posteriori* est unimodale. Une des caractéristiques utiles de l'approche bayésienne est l'intégration automatique des « corrections de la population finie » dans la distribution prédictive *a posteriori* des quantités de population finie – à mesure que l'échantillon converge vers une population finie, la variance *a posteriori* tend vers zéro.

L'accent est mis sur la création de modèles appropriés et de distributions *a priori*. Le calcul était autrefois un défi majeur et demeure une considération pratique, même si c'est moins le cas aujourd'hui, avec l'avènement des méthodes de Monte-Carlo par chaîne de Markov et les progrès rapides accomplis en matière de calculs bayésiens. Par conséquent, il s'avère plus difficile qu'il y a, disons, 30 ans de soutenir l'idée que Bayes est intéressant sur le plan conceptuel, mais trop difficile à mettre en œuvre.

Voici la structure de la suite du présent article. Dans la section 2, je présente une notation et je décris officiellement les modèles et les distributions *a priori* qui sont nécessaires dans l'approche bayésienne relative aux données d'enquête, avec et sans non-réponse. Je décris à la section 3 les caractéristiques habituellement souhaitables d'une inférence à propos de Q et j'indique les raisons qui me font croire que le paradigme bayésien pour des modèles choisis de manière appropriée peut se révéler plus efficace que l'approche fondée sur le plan de sondage pour obtenir ces caractéristiques. Je présente dans la section 4 une variété d'exemples visant à illustrer les points de la section 3. À titre de conclusion, dans la section 5, je propose 10 raisons d'adopter Bayes dans le cadre d'échantillonnage.

2. Notation et article fondamental

Dans la présente section, je présente la notation et un article fondamental qui sous-tend en grande partie la réflexion à l'origine du présent article. Supposons que $Y=(y_1,\ldots,y_N)$ et $S=(S_1,\ldots,S_N)$, où $N<\infty$ est le nombre d'unités dans la population, y_i est l'ensemble de variables d'enquête et S_i, l'indicateur de sélection pour la i^{e} unité, dont la valeur est 1 si l'unité i^{e} est sélectionnée, et 0 autrement. Supposons que Z correspond à l'information du plan de sondage, comme une strate ou des indicateurs de grappe, et que z_i correspond à la valeur de Z pour l'unité i. Examinons l'inférence à propos de la quantité de population finie $Q(Y,Z)$, par exemple le total de la population $Q(Y,Z)=\sum_{i=1}^{N} y_i$, où $Y=(y_1,\ldots,y_N)$. Une approche générale fondée sur un modèle traite S et Y comme des variables aléatoires, dont la distribution conjointe en fonction de Z est :

$$f_{S,Y|Z}(S,Y|Z,\theta,\psi) = f_{Y|Z}(Y|Z,\theta)\, f_{S|Y,Z}(S|Z,Y,\psi), \tag{2.1}$$

où $f_{Y|Z}$ représente la densité des variables d'enquête Y indexée par des paramètres θ inconnus, et où $f_{S|Y,Z}$ représente le modèle pour l'inclusion indexée par des paramètres ψ inconnus. Dans le cas d'un échantillonnage probabiliste sans non-réponse, la distribution de l'échantillonnage est connue et ne dépend pas de Y, c'est-à-dire que :

$$f_{S|Y,Z}(S|Z,Y,\psi) = f_{S|Z}(S|Z); \tag{2.2}$$

les méthodes fondées sur le plan de sondage fondent les inférences sur la distribution des statistiques dans un échantillonnage répété tiré de cette distribution.

Pour une enquête comportant de la non-réponse totale, une inclusion se produit lorsqu'une unité est sélectionnée, et puis répond, si sélectionnée. Ainsi, supposons que $R_i=1$ si l'unité sélectionnée i répond, et que $R_i=0$ dans le cas contraire. L'approche fondée sur un modèle modélise la distribution conjointe de S, R et Y compte tenu de Z sous la forme suivante :

$$f_{S,R,Y|Z}(S,R,Y|Z,\theta,\psi) = f_{Y|Z}(Y|Z,\theta)\, f_{S|Y,Z}(S|Z,Y,\psi)\, f_{R|S,Y,Z}(R|Z,Y,S,\phi), \tag{2.3}$$

en ajoutant à l'équation (2.1) un modèle correspondant à une non-réponse totale ayant une densité $f_{R|S,Y,Z}$. La modélisation des indicateurs pour les schémas de non-réponse partielle traite également les données manquantes partielles (par exemple Little, 2003b).

Le traitement de S, R et Y comme des variables aléatoires est une des principales caractéristiques de Rubin (1978) qui est, selon moi, un des articles phares de l'histoire de la statistique. L'article présente des conditions selon lesquelles les données manquantes et les mécanismes de sélection sont ignorables, c'est-à-dire qu'ils n'ont pas à être modélisés pour une inférence fondée sur la vraisemblance, en élargissant les définitions de l'ignorabilité aux données manquantes, dans Rubin (1976), tout en procurant un cadre pour

l'inférence en cas de sélection ou de données manquantes non ignorables. L'importance de l'article pour l'échantillonnage passe facilement inaperçue, puisque son axe principal est le rôle du mécanisme d'attribution de traitement en cas d'inférence relative à des effets de causalité. Le mécanisme d'attribution est ignorable dans une attribution de traitement aléatoire, comme dans des essais cliniques aléatoires. L'article expose ensuite le cadre général des inférences causales en comparant les traitements, caractéristique qui a valu à l'article sa popularité. Toutefois, il fournit également une justification bayésienne de l'échantillonnage aléatoire, en tant que moyen pour éviter le recours à un modèle pour procéder à la sélection.

Dans la modélisation de superpopulation fréquentiste (par exemple Valliant, Dorfman et Royall, 2000), les paramètres des modèles sont traités comme étant fixes; dans une modélisation bayésienne pour les enquêtes, ces paramètres sont attribués dans le cadre d'une distribution *a priori*, et les inférences pour $Q(Y)$ reposent sur sa distribution prédictive *a posteriori* en fonction des données. Dans les grands échantillons, la distribution *a priori* joue un rôle mineur, et les deux méthodes procurent des réponses semblables pour des modèles comparables; plus particulièrement, l'estimation du maximum de vraisemblance d'un paramètre constitue essentiellement le mode de distribution *a posteriori* dans le cas d'une distribution *a priori* uniforme, et a donc une interprétation bayésienne. Dans les petits échantillons, l'incertitude entourant les paramètres du modèle se propage lorsqu'ils sont intégrés à la distribution *a posteriori*. Cette approche de la propagation de l'erreur dans les paramètres permet aux inférences bayésiennes des modèles et des distributions *a priori* judicieusement choisis d'être mieux étalonnées que les inférences issues de la modélisation de superpopulation, dans le sens où elles ont de meilleures propriétés fréquentistes dans un échantillonnage répété (Rubin, 1978). Ainsi, à mon avis, la « modélisation de superpopulation est super, mais Bayes la surpasse ».

3. Inférence fondée sur le plan de sondage et inférence fondée sur un modèle

Les ouvrages portant sur l'échantillonnage font l'objet de nombreuses controverses (par exemple Smith, 1976, 1994; Kish, 1995; Brewer, 2013; Little, 2014) entre l'inférence « fondée sur le plan de sondage » qui repose sur la distribution de l'échantillonnage (2.2) et l'inférence « fondée sur un modèle » qui repose sur la distribution du modèle $f_{Y|Z}\left(Y\,|\,Z,\theta\right)$ si la sélection est ignorable, ou sur la distribution complète du modèle (2.1) ou (2.3) si la sélection est non ignorable ou en présence de non-réponse. Je cherche des inférences fondées à la fois sur le plan et sur un modèle, par le fait qu'elles découlent de modèles bayésiens, mais qui ont de bonnes propriétés fondées sur le plan.

Rubin (1984) fait une distinction entre une *inférence statistique pour un ensemble de données particulier*, et les *propriétés* de cette inférence – convergence, couverture de confiance – dans un

échantillonnage répété. Pour être généralement crédible, l'inférence doit aussi afficher de bonnes propriétés d'échantillonnage répété. L'objectif de l'approche « fondée sur le plan de sondage » doit aussi en être un pour les modèles bayésiens pour les enquêtes – il faut choisir le modèle et la distribution *a priori* de façon à obtenir des inférences ayant de bonnes propriétés fondées sur le plan. Pour y parvenir, les caractéristiques du plan d'échantillonnage probabiliste complexe doivent faire partie du modèle – stratification et pondération intégrées à l'aide de covariables, échantillonnage à plusieurs degrés intégré à l'aide de modèles hiérarchiques. L'inclusion d'une distribution *a priori* dans la modélisation bayésienne, décriée par certains comme constituant une autre hypothèse à ajouter au modèle, me fournit un outil complémentaire à la modélisation de superpopulation. Elle procure une plus grande souplesse que la modélisation de superpopulation, ce qui restreint efficacement le nombre de distributions *a priori* uniformes possibles.

En plus d'avoir de bonnes propriétés fréquentistes, l'inférence qui repose sur q et I doit être appropriée – Rubin (2019) emploie le mot « pertinente » – à l'ensemble de données réalisé. Supposons que D désigne les données à la base de l'inférence et $\tilde{D}$, la réalisation particulière de D, les valeurs réellement obtenues de l'échantillon et du répondant. Que q et I découlent d'un modèle bayésien formel, d'une équation estimative ou d'une procédure algorithmique, ils doivent fournir une bonne inférence pour les données $\tilde{D}$, pas d'autres ensembles de données D qui seraient possiblement obtenus. Les méthodes bayésiennes ont tendance à présenter cette propriété en raison des conditions de distribution *a posteriori* appliquées à $\tilde{D}$; mais il faut également un intervalle de confiance approximativement valide s'il est considéré comme un intervalle crédible qui conditionne $\tilde{D}$, ne serait-ce que parce qu'il s'agit de la façon dont un non-statisticien aurait tendance à l'interpréter. Cette perspective manque parfois aux intervalles de confiance fondés sur le plan, comme le montrent les exemples 2 et 4 ci-dessous.

En somme, l'un des objectifs communs de l'inférence fondée sur le plan de sondage et de celle fondée sur un modèle consiste à obtenir une valeur de q qui utilise à bon escient les données, présente certaines propriétés comme la convergence par rapport au plan qui suppose qu'elle n'est pas trop éloignée de Q, et un intervalle I qui est aussi étroit que le permettent les renseignements contenus dans les données, tout en incluant Q selon une probabilité près de la valeur nominale de 95 %. Rubin (2019) associe ces propriétés à un statisticien « rempli de sagesse », dans son discours Waksberg.

Les méthodes fondées sur le plan sont souvent simplifiées pour éviter le recours à un modèle, parce que les propriétés comme la convergence par rapport au plan ne reposent pas sur un modèle, pour les données. Cependant, le rendement des méthodes fondées sur le plan dépend souvent d'un modèle implicite, et la modification de l'estimation reposant sur un modèle plus réaliste peut améliorer l'inférence, selon une perspective fondée sur le plan de sondage ou sur un modèle. Ce point est illustré dans les exemples 3 à 5 ci-dessous.

La question de l'ensemble de références approprié pour les propriétés d'un échantillonnage répété, comme les intervalles de confiance, est truffée de difficultés, surtout en ce qui concerne l'adéquation d'appliquer un conditionnement sur des statistiques ancillaires ou sur celles s'en rapprochant (par exemple Birnbaum, 1962; Berger et Wolpert, 1988; Ghosh, Reid et Fraser, 2010). L'évaluation des propriétés d'une inférence bayésienne pour un échantillonnage répété soulève également ces questions, mais ces dernières ne s'appliquent pas à l'inférence elle-même parce que la distribution *a posteriori* conditionne $\tilde{D}$.

L'approche fondée sur le plan de sondage, pour l'inférence pour les enquêtes par échantillonnage, comporte une portée trop limitée, ce qui empêche de traiter adéquatement bon nombre des problèmes liés à ce type d'enquêtes, dans la pratique. Elle comporte les limites suivantes :

1. L'inférence fondée sur le plan de sondage est asymptotique, et elle ne procure pas d'inférences valides dans de petits échantillons. Examinons l'exemple simple qui suit.

   Exemple 1. Inférence concernant la moyenne de la population par échantillonnage aléatoire simple. Considérons l'inférence concernant la moyenne de la population d'une variable Y tirée d'un échantillon aléatoire simple de taille n à partir d'une population de taille N. L'intervalle de confiance type à 95 %, fondé sur le plan, prend la forme suivante

   $$\bar{y} \pm 1{,}96\, s\sqrt{(1/n - 1/N)}, \tag{3.1}$$

   où $\bar{y}$ est la moyenne de l'échantillon et s, l'écart-type de l'échantillon. Cet intervalle est asymptotique et ne fournit pas d'inférences valides avec petit échantillon. Si, notamment, Y est continu, on obtient habituellement une meilleure inférence en remplaçant 1,96, le 97,5^{e} centile de la distribution normale, par le 97,5^{e} centile d'une distribution qui tient compte de l'incertitude à propos de l'estimation de la variance, comme la distribution t comportant $n-1$ degrés de liberté. La procédure présume toutefois une distribution normale pour Y et ainsi elle n'est pas fondée sur le plan. Si Y est binaire, avec les valeurs 0 et 1, alors $\bar{y}$ est la proportion de l'échantillon, $s = \sqrt{\bar{y}(1-\bar{y})}$ et (3.1) est l'intervalle de Wald asymptotique, qui a un très mauvais rendement avec les petits échantillons, surtout lorsque la proportion réelle se rapproche de 0 ou de 1. L'intervalle bayésien crédible pour l'*a priori* de Jeffreys ou uniforme présente de bien meilleures propriétés fréquentistes. Consultez les travaux de Dean et Pagano (2015) et de Franco, Little, Louis et Slud (2019) qui présentent des comparaisons des intervalles de Wald avec des solutions de rechange pour les plans de sondage complexes. L'inférence fondée sur le plan de sondage est souvent une mauvaise option pour les petits échantillons, plus particulièrement pour l'estimation sur petits domaines où on fait appel à un modèle pour Y afin d'« obtenir un renforcement par emprunt de données » aux différents domaines.

2. Une inférence fondée sur le plan de sondage ne gère pas d'unité d'enquête, de non-réponse partielle ou d'erreur de réponse, parce que ces problèmes nécessitent des modèles qui génèrent habituellement des résultats satisfaisants.
3. Une inférence fondée sur le plan de sondage n'est pas normative, d'une manière permettant de prescrire un choix approprié de méthode d'inférence pour les données à traiter. Le choix approprié d'un estimateur nécessite effectivement un modèle implicite, comme dans l'estimation « assistée par un modèle » (par exemple Särndal, Swensson and Wretman, 1992). L'estimateur de régression ou par quotient, par exemple, l'intégration d'information auxiliaire, ou l'estimateur de Horvitz-Thompson ou de Hájek pour l'insertion de poids d'enquête, reposent tous sur des modèles implicites, et si ce modèle est loin d'être réaliste, ces méthodes se révéleraient très sous-optimales – les éléphants de Basu (1971) constituant un exemple extrême et satirique. L'inférence bayésienne fondée sur des modèles plus souples tend à donner de meilleurs résultats, comme il sera établi dans l'exemple 5 ci-dessous.
4. Une inférence fondée sur le plan de sondage ne traite pas de la manière de générer des inférences pour des échantillonnages non probabilistes, qui sont de plus en plus utilisés en raison du coût et des difficultés liées à l'obtention de vrais échantillons aléatoires.

Le point 4 ne permet pas d'exclure le recours à des méthodes fondées sur le plan pour obtenir q et I, parce que nous pouvons toujours prétendre que nous disposons d'un échantillon probabiliste en adoptant un modèle pour les indicateurs de sélection S qui définit l'inclusion dans l'échantillon, et en estimant les paramètres inconnus dans ce modèle (par exemple Elliott et Valliant, 2017). Les statisticiens qui emploient des méthodes fondées sur le plan pour une inférence tirée d'échantillons aléatoires ont tendance à privilégier cette méthode de « quasi-randomisation ». Cette méthode comporte toutefois les mêmes limites que l'approche fondée sur le plan de sondage concernant les échantillons probabilistes, à savoir l'incapacité à gérer de petits échantillons, les données manquantes ou les erreurs de réponse; la boîte à outils bayésienne est beaucoup mieux garnie pour traiter ces problèmes.

4. Exemples

Une variété d'exemples sont présentés, certes plutôt axés sur mes propres travaux réalisés avec des collègues. J'évite les sujets comme l'estimation de petits domaines ou de séries chronologiques, parce que le besoin de modélisation est bien établi, dans ces cas.

Exemple 1 (suite). Échantillon aléatoire normal. Une des critiques souvent entendues à propos des méthodes fondées sur un modèle soutient que « si mon inférence repose sur un modèle et que ce modèle est incorrect, l'inférence le sera également. Puisque (pour paraphraser George Box) tous les modèles sont

incorrects, toutes les inférences fondées sur un modèle sont par conséquent incorrectes. Je préfère donc une inférence fondée sur le plan de sondage, qui ne demande pas d'hypothèses de modélisation ». Le raisonnement, quoique plausible, n'est pas si simple. La validité des méthodes fondées sur un modèle dépend du plan, de même que du degré et de la nature de l'erreur de spécification du modèle; le fait que les méthodes fondées sur le plan ne dépendent pas nécessairement des modèles ne constitue pas un signe incontestable de leur supériorité par rapport à ceux qui en dépendent.

Supposons qu'un échantillon aléatoire simple de taille n est prélevé à partir d'une distribution continue ayant une moyenne μ inconnue et un écart-type σ. Examinons trois estimations d'intervalle pour la moyenne de la population :

(i) L'intervalle A est l'intervalle de confiance type à 95 %, fondé sur le plan, équation (3.1).

(ii) L'intervalle B remplace le 97,5^{e} centile normal dans l'équation (3.1), à savoir 1,96, par le 97,5^{e} centile de la distribution t avec $n-1$ df.

(iii) L'intervalle C est l'intervalle de crédibilité *a posteriori* à 95 % fondé sur un modèle normal, avec une distribution $p(\mu,\sigma) \propto 1/\sigma$ selon l'*a priori* de Jeffreys.

D'un point de vue fréquentiste, lequel de ces intervalles est le plus juste ? L'intervalle A ne s'appuie sur aucune hypothèse de distribution pour Y, et B s'appuie sur une hypothèse normale – faut-il en déduire que A est supérieur ? Si n est grand, les intervalles seront presque les mêmes, et si n est petit, alors B est sans doute mieux que A, même si les données ne sont pas normales, puisqu'il tient compte de l'incertitude à propos de la variance.

Dans une enquête informelle, les deux tiers d'une classe récente de nos étudiants de troisième cycle bien formés ont préféré B à C, faisant valoir qu'il évite de choisir une distribution *a priori* et, par conséquent, fait moins d'hypothèses. Mais B et C sont également bons ou mauvais, puisqu'il s'agit de la même procédure! Que l'intervalle de confiance à 95 % soit précis selon la normalité et qu'il soit un intervalle de crédibilité à 95 % pour le choix d'*a priori* mentionné constitue simplement deux propriétés de la procédure. Juger une méthode par ses hypothèses apparentes est un excès de simplification.

Exemple 2. Échantillon aléatoire simple avec borne inférieure sur la variance (Little, 2006). Supposons dans l'exemple précédent que $n=7,\ \bar{y}=1, s=1.$ L'intervalle t normal (en ignorant les corrections d'une population finie) serait

$$\bar{y} \pm t_{6\ 975}\, s/\sqrt{n} \ = 1 \pm 2{,}447/\sqrt{7} = 1 \pm 0{,}925 \tag{4.1}$$

si, en fait, nous savons que $\sigma =$ 1,5, nous aurions un meilleur intervalle en

$$\bar{y} \pm z_{0{,}975}\, \sigma/\sqrt{n} \ = 1 \pm 1{,}96 \times 1{,}5/\sqrt{7} = 1 \pm 1{,}11, \tag{4.2}$$

l'intervalle le plus vaste tenant compte du fait que σ est plus grand que la valeur particulière de s pour cet échantillon. L'intervalle t (4.1) a une couverture de confiance exacte, mais compte tenu de ce que nous savons à propos de σ, il s'agit de la mauvaise inférence pour cet ensemble de données en particulier : nous ne devrions pas le retenir au détriment de (4.2) parce qu'il est plus étroit!

Supposons maintenant que nous savons que $\sigma > 1{,}5$, en raison d'une certaine source de variation additionnelle dont on n'a pas tenu compte. L'approche bayésienne intègre ensuite ce renseignement à la distribution *a priori* pour σ, ce qui génère un intervalle crédible plus large que dans (4.2). Quelle est la réponse fréquentiste ? L'intervalle de confiance t a toujours une couverture exactement nominale dans un échantillonnage répété, mais il est nettement trop étroit en tant qu'inférence pour l'ensemble de données observé parce qu'il ne tient pas compte de ce qui est connu à propos de σ. L'intervalle (4.2) est anticonservateur, malgré le fait qu'il soit plus grand que (4.1) pour l'ensemble de données réalisé – une propriété qu'il est possible d'observer pour des intervalles de confiance, mais pas pour des intervalles de crédibilité selon un modèle particulier. Des méthodes fréquentistes asymptotiques ne sont d'aucune aide dans le cas présent, alors quelle est la solution de rechange à Bayes dans cet exemple ?

Un exemple connexe apparaît dans une analyse de variance à effets aléatoires au sein d'un groupe, lorsque l'estimation des moindres carrés de la variance entre les groupes est négative – une analyse bayésienne aborde ce problème à l'aide d'une distribution *a priori* pour la variance entre les groupes qui ne permet pas d'obtenir des valeurs négatives. Les modèles à effets aléatoires et à effets mixtes sont importants pour traiter la mise en grappe dans les enquêtes, milieu dans lequel les méthodes bayésiennes fonctionnent mieux que le maximum de vraisemblance.

Exemple 3. Poststratification sur une covariable catégorique. La prévision (dans une modélisation) constitue une approche générale plus fiable de l'inférence que la pondération (dans une inférence fondée sur le plan de sondage). J'illustre cet énoncé général à l'aide d'un simple exemple de poststratification sur une seule variable Z. Vous trouverez ci-dessus un exemple plus complexe (exemple 7).

Examinons un échantillonnage probabiliste à probabilité égale comportant une seule variable catégorique de poststratification Z, pour laquelle le chiffre de population connu N_h est disponible pour chaque strate *a posteriori* h, $h = 1, 2, \ldots, H$. Supposons que $\bar{y}_h$ est la moyenne de l'échantillon dans une strate *a posteriori* h, d'après une taille d'échantillon de n_h, et $n = \sum_{h=1}^{H} n_h$, $N = \sum_{h=1}^{H} N_h$. L'estimation normale de la moyenne de la population est la moyenne poststratifiée :

$$\bar{y}_{\mathrm{PS}} = \sum_{h=1}^{H} P_h \bar{y}_h, \tag{4.3}$$

où $P_h = N_h / N$ et N correspond à la taille de la population. Cela peut être vu comme une moyenne pondérée

$$\bar{y}_{\mathrm{PS}} = \sum_{h=1}^{H} \sum_{i=1}^{n_h} w_i \, y_i,$$

où $w_i = N_h / (N n_h)$ est le poids de poststratification pour les unités échantillonnées dans une strate *a posteriori* h. Ces poids peuvent être très importants dans les strates « *a posteriori* » avec des petites tailles d'échantillon n_h, qui, contrairement à un échantillonnage stratifié qui repose sur Z, ne relèvent pas du contrôle de l'échantillonneur. Ces poids imposants peuvent entraîner une variabilité excessive dans $\bar{y}_{\mathrm{PS}}$. En fait, à proprement dit, $\bar{y}_{\mathrm{PS}}$ n'a pas de distribution dans un échantillonnage répété, car avec une probabilité positive, les tailles d'échantillon dans certaines strates *a posteriori* peuvent être nulles. Cela demeure vrai en cas de modification des strates *a posteriori* visant à assurer une valeur positive aux dénombrements d'échantillon des strates *a posteriori* de tout l'échantillon observé, par exemple par la combinaison de strates adjacentes.

L'approche type fondée sur le plan de sondage pour une variabilité excessive de $\bar{y}_{\mathrm{PS}}$ consiste à modifier les poids $\{w_i\}$, par exemple en éliminant les plus grands. Cependant, d'un point de vue prévisionnel, cette stratégie n'est pas une bonne idée. Le problème ne se situe pas au niveau des poids – les proportions de la population $\{P_h\}$ de chaque strate *a posteriori* sont connues, après tout – mais bien dans le manque de données de l'échantillon dans certaines strates *a posteriori* qui rend les estimations $\{\bar{y}_h\}$ peu fiables. Ce sont les estimations des strates *a posteriori* ayant peu de données qu'il faut modifier, non les poids $\{w_i\}$ rattachés aux unités échantillonnées. Selon les principes, le moyen de modifier $\{\bar{y}_h\}$ consiste à adopter un modèle qui relie Y et Z. L'approche fondée sur le plan de sondage, en évitant d'adopter ce modèle, mène au mauvais principe – modifier les poids au lieu des prévisions de valeurs non échantillonnées.

Point connexe, une approximation courante fondée sur le plan (Kish, 1992) mesure l'augmentation proportionnelle de la variance par rapport à la pondération en tant que $1 + \mathrm{cv}(w_i)$, où $\mathrm{cv}(w_i)$ est le coefficient de variation des poids; un élagage des poids réduit $\mathrm{cv}(w_i)$ ce qui entraîne par conséquent une augmentation proportionnelle. Cependant, cette règle empirique n'est valide que lorsque Y et Z ne sont pas liés, auquel cas une poststratification est inutile. Si Y et Z sont liés et que la taille de l'échantillon n'est pas trop petite, la variance de la moyenne poststratifiée est *plus petite*, non *plus grande*, que la variance de la moyenne non pondérée de l'échantillon (Holt et Smith, 1979; Little et Vartivarian, 2005). La règle empirique ne fonctionne pas parce que la relation entre Y et Z n'est pas modélisée.

Quelle est l'approche bayésienne face à une variabilité excessive de $\bar{y}_{\mathrm{PS}}$? Cette dernière représente la moyenne *a posteriori* du modèle normal stratifié

$$\left(y_{hi} \,|\, \mu_h, \sigma^2\right) \overset{\text{iid}}{\sim} G\left(\mu_h, \sigma^2\right), \tag{4.4}$$

$$p\left(\mu_h, \sigma\right) \propto 1/\sigma, \tag{4.5}$$

où $G\left(a, b^2\right)$ désigne la distribution normale (gaussienne) ayant une moyenne a, une variance b^2 et l'équation (4.5) est la distribution *a priori* de Jeffreys sur la moyenne et l'écart-type pour chaque strate *a posteriori*. Puisque l'échantillon compte quelques strates *a posteriori* ayant peu de données, la distribution *a priori* doit être modifiée afin de permettre un renforcement par emprunt d'information tirée d'autres strates. Une des méthodes consiste à adopter un modèle normal à effets aléatoires

$$p\left(\mu_h \,|\, \mu, \tau, \sigma\right) \overset{\text{iid}}{\sim} G\left(\mu, \tau^2\right),\ p\left(\mu, \sigma, \tau\right) \propto \sigma^{-1}, \tag{4.6}$$

qui traite les moyennes de strate comme des effets aléatoires. Il serait aussi possible de traiter les variances dans chaque strate *a posteriori* comme des effets aléatoires distincts et les soumettre à une distribution *a priori*, au lieu de les regrouper. La moyenne *a posteriori* de $\bar{Y}$ pour la distribution *a priori* (4.6) déplace le poids w_i des unités échantillonnées dans la strate *a posteriori* h vers la valeur 1, avec un degré de rétrécissement qui dépend de la taille relative des estimations de σ et de τ (Lazzeroni et Little, 1998).

La distribution *a priori* (4.6) donne lieu à l'hypothèse non négligeable que les moyennes des strates *a posteriori* sont échangeables. Il est possible de la modérer en limitant le modèle à effets aléatoires à un sous-ensemble de strates *a posteriori* ayant peu d'unités échantillonnées; ou en remplaçant la moyenne constante μ dans (4.6) par une régression des caractéristiques C_h connues des strates *a posteriori*, comme dans :

$$p\left(\mu_h \,|\, \beta_0, \beta_1, \tau, \sigma, c_h\right) \overset{\text{ind}}{\sim} G\left(\beta_0 + \beta_1 c_h, \tau^2\right),\ p\left(\beta_0, \beta_1, \sigma, \tau\right) \propto \sigma^{-1}, \tag{4.7}$$

qui limite l'hypothèse de l'échangeabilité aux erreurs de la régression de μ_h sur c_h. Pour des généralisations complètes de cet exemple de base, consultez les travaux de Gelman et Little (1997), d'Elliott et Little (2000), d'Elliott (2007), de Gelman (2007) et de Si, Trangucci, Gabry et Gelman (2020).

Exemple 4. Estimateur par la régression de la moyenne, compte tenu d'une variable auxiliaire de la population. Si la variable auxiliaire Z de l'exemple précédent est continue, une façon habituelle de l'intégrer à l'inférence consiste à utiliser une estimation par régression de la moyenne :

$$q = \bar{y}_{\text{REG}} = \bar{y} + \hat{\beta}_1\left(\bar{Z} - \bar{z}\right),$$

où $\hat{\beta}_1$ est l'estimation des moindres carrés de la pente de Y sur Z dans l'échantillon, alors que $\bar{z}$ et $\bar{Z}$ correspondent respectivement à la moyenne de l'échantillon et de la population de Z. Dans une étude de simulation de cinq populations réelles, Royall et Cumberland (1981, 1985) évaluent les inférences axées sur q, avec (a) l'erreur type fondée sur le plan reposant sur un échantillonnage aléatoire simple, à savoir :

$$I_{0{,}95D} = \bar{y}_{\text{reg}} \pm 1{,}96\hat{\text{se}}_D\left(\hat{\beta}\right),\ \hat{\text{se}}_D\left(\hat{\beta}\right) = \sqrt{\left(1-f\right)s_{Y.Z}/n},$$

où $s^2_{Y.Z}$ est la variance résiduelle de l'échantillon et $f = n/N$, la fraction d'échantillonnage; et (b) l'erreur type de la prévision, reposant sur le modèle de régression linéaire normal ayant une variance constante, à savoir :

$$I_{0,95M} = \bar{y}_{\text{reg}} \pm t_{0,975,\,n-2} 1{,}96 \hat{\text{se}}_M\left(\hat{\beta}\right),$$

$$\hat{\text{se}}_M\left(\hat{\beta}\right) = \hat{\text{se}}_D\left(\hat{\beta}\right)\sqrt{\left(1+\left(\bar{z}-\bar{Z}\right)^2\right)\Big/\left(1-f\right)\left(\sum_{i=1}^{n}\left(z_i-\bar{z}\right)^2/n\right)}.$$

Les intervalles de confiance fondés sur le plan $I_{0,95D}$ affichent une très faible couverture de confiance conditionnelle lorsque le $\bar{z}$ observé dévie nettement de $\bar{Z}$. L'intervalle de confiance fondé sur un modèle $I_{0,95M}$ tient compte du manque d'équilibre concernant $\bar{z}$, mais est sensible à une erreur de spécification du modèle, surtout à un manque de linéarité dans la relation entre Y et Z ou à une variance résiduelle non constante. De solides estimations de l'erreur type, reposant sur l'estimateur sandwich ou l'estimateur jackknife, procurent des intervalles ayant de meilleures propriétés de couverture conditionnelle, même si on observe encore parfois une déviation par rapport aux taux de couverture nominaux. Une démarche de rechange serait l'inférence bayésienne, reposant sur un modèle plus souple reliant Y à Z, comme le modèle de spline pénalisée :

$$\left(y_i \,|\, z_i, \beta, \sigma^2\right) \overset{\text{ind}}{\sim} G\left(\text{spline}\left(z_i, \beta\right), \sigma^2 z_i^{\alpha}\right),$$

$$\text{spline}\left(z_i, \beta\right) = \beta_0 + \sum_{j=1}^{p}\beta_j z_i^j + \sum_{\ell=1}^{m}\beta_{\ell+p}\left(z_i-\kappa_\ell\right)_+^p,$$

$$\left(\beta_{l+p}\,|\,\tau\right) \overset{\text{iid}}{\sim} N\left(0,\tau^2\right), l=1,\ldots,m;\ p\left(\beta_0,\ldots,\beta_p,\alpha,\sigma,\tau\right)\propto 1/\sigma, 0<\alpha<2, \qquad (4.8)$$

où les constantes $\kappa_1 < \ldots < \kappa_m$ sont des nœuds fixes sélectionnés, et $(u)_+^p = u^p$ si $u > 0$ et 0, dans les autres cas (consulter, à titre d'exemple, Ruppert, Wand et Carroll, 2003). Le paramètre α permet d'avoir une variété de formes courantes d'hétéroscédasticité. Les erreurs types bayésiennes tiennent alors compte du déséquilibre de la distribution de Z dans l'échantillon et la population, alors que la souplesse du modèle limite le biais causé par une erreur de spécification du modèle.

Supposons que la quantité ciblée n'est pas la moyenne de la population de Y, mais la pente de la droite des moindres carrés de Y sur Z dans la population. Une démarche judicieuse consiste à imputer des valeurs non échantillonnées de Y à l'aide du modèle (4.8), puis d'estimer la pente de la droite de Y sur Z en tant que pente de la droite des moindres carrés pour les données sur la population insérées au préalable. Il est possible de propager l'incertitude grâce à une imputation multiple (Rubin, 1987), une méthode qui repose sur des principes bayésiens. Dans ce contexte, Little (2004) fait une distinction entre le « modèle ciblé » qui détermine la quantité de la population cible présentant un intérêt, la régression linéaire de Y sur X dans le cas présent, et le « modèle de travail » (4.8) qui constitue le fondement d'une inférence et sert à prévoir les variables d'enquête pour les unités non échantillonnées et non répondantes de la population. La distinction entre ces deux modèles procure une solide forme d'inférence bayésienne pour les enquêtes.

Szpiro, Rice et Lumley (2010) appliquent un principe semblable dans le réglage d'une régression de la superpopulation, et Little (2019) avance que cela se révèle plus simple que de modifier l'interprétation du paramètre, la démarche adoptée par Buja, Berk, Brown, George, Pitkin, Zhan et Zhang (2019). Szpiro

et coll. (2010) montrent que la démarche mène à une interprétation bayésienne de l'estimateur sandwich de la variance dans une régression, qui équivaut sur le plan asymptotique aux estimations de la variance par réutilisation de l'échantillon, comme le bootstrap ou le jackknife, qui sont souvent appliquées dans le cadre d'une enquête par échantillonnage.

Exemple 5. Inférence pour des échantillons ayant des probabilités inégales de sélection. En ce qui concerne les plans ayant des probabilités inégales de sélection, les articles classiques critiquant la méthode de modélisation des enquêtes (Kish et Frankel, 1974; Hansen, Madow et Tepping, 1983) n'incluent pas les probabilités de sélection dans le modèle, ce qui donne lieu à des inférences sensibles à une erreur de spécification du modèle. Les probabilités de sélection jouent un rôle de premier plan dans la solidité d'une inférence fondée sur un modèle, mais en tant que covariables du modèle plutôt que poids d'échantillonnage.

Examinons, par exemple, une inférence relative à la moyenne de la population $\bar{Y}$. Si y_i est la valeur de la variable d'enquête Y et π_i est la probabilité de sélection de l'unité i, l'estimateur habituel de la moyenne de Y, fondé sur le plan, pondère les unités échantillonnées (disons les unités $i=1,\ldots,n$) par l'inverse de π_i qui nous donne l'estimation de Horvitz-Thompson (Horvitz et Thompson, 1952)

$$\bar{y}_{\mathrm{HT}} = N^{-1}\sum_{i=1}^{n} y_i/\pi_i, \tag{4.9}$$

si la taille de la population N est connue, ou l'estimation de Hájek (1971)

$$\bar{y}_{\mathrm{HK}} = \left(\sum_{i=1}^{n} y_i/\pi_i\right)\Big/\left(\sum_{i=1}^{n} 1/\pi_i\right), \tag{4.10}$$

si N est estimé. La pondération est une démarche universelle où les unités échantillonnées obtiennent le même poids, sans tenir compte de la relation entre π_i et y_i. Cela se révélera possiblement inefficace – si la relation est faible, la pondération aura simplement pour effet de réduire la précision de l'estimation en l'absence d'une réduction compensatoire du biais.

La méthode de modélisation intègre les probabilités de sélection en procédant à une régression de y_i sur π_i. La force de la relation entre y_i et π_i modère ensuite l'effet qu'aura la probabilité de sélection sur l'estimateur – si la relation est faible, le coefficient de régression de π_i est modeste et le poids d'échantillonnage a peu d'incidence. Cela procure des estimations plus efficaces.

Une régression linéaire de y_i sur π_i est sensible au biais en cas d'erreur de spécification de la linéarité, mais on peut réduire l'incidence de cette erreur en choisissant un modèle aboutissant à une estimation convergente par rapport au plan. De nombreux modèles satisfont à cette exigence. À titre d'exemple, voir Firth et Bennett (1998).

En cas d'échantillonnage stratifié à l'aide d'une variable de stratification Z, le modèle de régression naturel comporte des covariables qui consistent en variables indicatrices pour la strate. L'estimation résultante de la moyenne de la population d'une variable d'enquête continue Y est la moyenne stratifiée,

qui prend la même forme que dans l'équation (4.3), mais pour laquelle Z forme des strates au lieu de strates *a posteriori*. Dans cette situation, la pondération par l'inverse de la probabilité de sélection de chaque strate et la régression par variable indicatrice aboutiront toutes deux à l'estimateur (4.3). Les méthodes bayésiennes de réduction de la variance décrites dans l'exemple 3 tendent à être moins rentables dans le cas d'un échantillonnage stratifié qu'en poststratification, parce que l'échantillonneur a le contrôle sur la taille des échantillons de chaque strate.

Dans un échantillonnage de probabilité proportionnelle à la taille (PPT), la covariable Z correspond à la taille de l'unité et de $\pi_i = \min(cz_i, 1)$. Maintenant, puisque Z est une variable continue, la pondération et la régression pourraient donner des réponses différentes. On pourrait unifier les méthodes en prenant en considération les modèles qui procurent des estimations pondérées selon le plan de sondage lorsqu'ils sont utilisés pour prévoir les unités non échantillonnées. Plus particulièrement, en ignorant les corrections de la population finie, l'estimation de Horvitz-Thompson (4.9) est la moyenne *a posteriori* pour le « modèle Horvitz-Thompson ».

$$\left(y_i \,|\, z_i, \beta, \sigma^2\right) \overset{\text{ind}}{\sim} G\left(\beta z_i, \sigma^2 z_i^2\right),\ p(\beta, \sigma) \propto 1/\sigma, \tag{4.11}$$

et l'estimation Hájek (4.10), la moyenne *a posteriori* pour le « modèle Hájek » :

$$\left(y_i \,|\, z_i, \beta, \sigma^2\right) \overset{\text{ind}}{\sim} G\left(\beta, \sigma^2 z_i\right),\ p(\beta, \sigma) \propto 1/\sigma. \tag{4.12}$$

Ces modèles sous-jacents décrivent des situations dans lesquelles les estimations fondées sur le plan correspondantes sont optimales. Ils comportent toutefois de fortes hypothèses paramétriques. Une bonne méthode de modélisation bayésienne intègre ces modèles dans un modèle plus vaste, comme le modèle de spline pénalisée (4.8), comme l'ont proposé Zheng et Little (2003). Zheng et Little (2005) et Chen, Elliott, Haziza, Yang, Ghosh, Little, Sedransk et Thompson (2017) présentent des études de simulation indiquant que cette méthode de modélisation peut procurer des gains substantiels par rapport à une estimation de Horvitz-Thompson ou de Hájek, à la fois sur le plan de l'efficacité et du rapprochement de la couverture de confiance nominale (fréquentiste) dans des échantillons modérés – rappelons-nous que les résultats fondés sur le plan sont asymptotiques. Le modèle (4.8) s'agrandit facilement pour permettre l'insertion d'autres variables auxiliaires mesurées pour toutes les unités de population, et la souplesse des inférences de petits échantillons s'accroît grâce à l'inclusion de distributions *a priori* qui conviennent pour les paramètres du modèle.

Leon-Novelo et Savitsky (2019) examinent des modèles bayésiens pour la distribution conjointe de y_i et π_i, en appelant les modèles qui corrigent π_i des méthodes « d'extension ». Cependant, l'attribution d'une distribution à π_i me semble à la fois artificielle et inutile dans ce réglage; si on enregistre π_i pour toutes les unités de population, il peut être conditionné dans le modèle, comme dans des méthodes de régression classiques qui traitent les covariables comme des valeurs fixes. Même si les valeurs de π_i pour les unités non échantillonnées ne sont pas fournies aux analystes, il est possible de prévoir leurs valeurs à l'aide d'un théorème bayésien, comme dans Zangeneh et Little (2015).

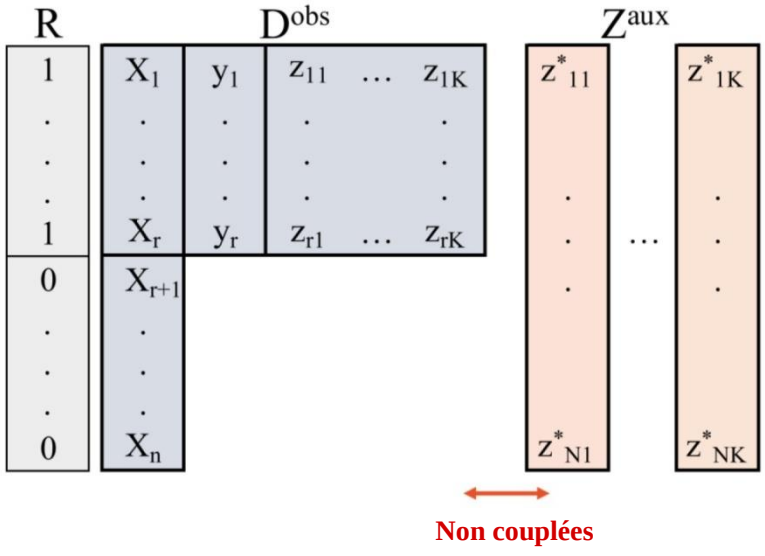
Exemple 6. Spline pénalisée de modèles de propension de réponse. Examinons la non-réponse totale pour une variable d'enquête Y, dans le cadre de l'observation d'un ensemble de variables, disons

$X_1, \ldots, X_p$ concernant les répondants et les non-répondants de l'échantillon. La propension de réponse pour l'unité i, $\theta_i = \Pr\left(R_i = 1 \middle| x_{i1}, \ldots, x_{ip}, \psi\right)$, où ψ représente des paramètres inconnus, joue un rôle important dans les méthodes de pondération et de prévision relatives à la non-réponse à l'enquête. Dans le cas de la pondération de la non-réponse, le poids d'échantillonnage est multiplié par le poids de non-réponse, de la façon suivante

$$\begin{aligned} w_i &= 1/\Pr(\text{l'unité } i \text{ est sélectionnée et répond}) \\ &= (1/\Pr(i \text{ sélectionnée})) \times (1/\Pr(i \text{ répond} \mid \text{sélectionnée})) \\ &\quad \text{poids d'échantillonnage} \times \text{poids de réponse.} \end{aligned}$$

Contrairement au poids d'échantillonnage, le poids de réponse est inconnu, et il faut indiquer clairement ce sur quoi la probabilité dépend. Little (2022) avance qu'elle doit conditionner les variables auxiliaires et d'enquête, mais pas d'autres variables qui pourraient avoir un effet sur elle. Comme pour les poids d'échantillonnage, une pondération par l'inverse de la propension de réponse estimée est une méthode universelle qui ne tient pas compte de la force de la relation entre R et Y. Une spline pénalisée de prévision de la propension (Zhang et Little, 2009) fait régresser Y sur la spline pénalisée de la propension de réponse estimée, alors que d'autres variables observées pour les répondants et les non-répondants apparaissent sur le plan paramétrique. La spline modélise une relation souple entre la propension de réponse et Y, ce qui donne de la robustesse à une erreur de spécification du modèle. La propriété équilibrante du score de propension procure également une double propriété de robustesse, en cela que la forme paramétrique de la régression des autres prédicteurs peuvent comporter une erreur de spécification sans entraîner de biais, à condition que la spline pénalisée saisisse la relation entre la propension et Y. Cette méthode fonctionne bien avec les méthodes de pondération, dans les simulations (Zhang et Little, 2009; Yang et Little, 2015), et la version exclusivement bayésienne avec distributions *a priori* sur les paramètres propage l'erreur dans l'estimation des propensions.

Figure 4.1 Structure des données dans l'exemple 7.



Exemple 7. Non-réponse totale avec poststratification. Un autre exemple dans lequel les approches bayésiennes et les approches fondées sur le plan de sondage qui ont trait à l'inférence diffèrent est celui de la non-réponse totale avec poststratification. Examinons le scénario de la figure 4.1, où Y est une variable d'enquête assujettie à une non-réponse, X est une variable observée pour toutes les unités de l'échantillon et où $Z^{\text{aux}} = (Z_1, \ldots, Z_K)$ est constitué d'un ensemble de variables auxiliaires catégoriques. La distribution conjointe de Z^{aux} est observée pour les répondants de l'enquête, et les distributions marginales de Z_k, $k = 1, \ldots, K$, sont aussi observées pour la population ou l'échantillon provenant de sources externes. Les unités des données auxiliaires ont une caractéristique distinctive, celle de ne pas être couplées aux unités de l'enquête. Ce scénario apparaît souvent dans les réglages où une poststratification sert à ajuster la non-réponse. Définissons R_i en tant qu'indicateur de réponse pour (y_i, z_i) dans une unité d'échantillonnage i, et supposons que

$$\Pr(R_i = 1 | x_i, z_i, y_i, \psi) = \Pr(R_i = 1 | x_i, z_i, \psi), \tag{4.13}$$

faisant en sorte que cette probabilité ne dépende pas de y_i. Il convient de souligner que si cette probabilité dépend de z_i, le mécanisme de réponse n'est pas manquant au hasard dans le sens de Rubin (1976), parce que les valeurs de z_i sont manquantes pour les non-répondants de l'enquête. Zangeneh et Little (2022) prennent en considération l'estimation bayésienne et l'estimation par maximum de vraisemblance pour les données présentant ce modèle. En considérant, pour simplifier, des modèles de type i.i.d. pour les unités i, la distribution conjointe de X, Z, Y et R est prise en compte de la façon suivante

$$\begin{aligned} f_{X,Z,Y,R}(x_i, z_i, y_i, r_i | \theta, \phi) &= f_{Y|X,Z,R}(y_i | x_i, z_i, \theta, R_i = r_i)\, f_{X,Z,R}(x_i, z_i, r_i | \phi) \\ &\overset{\text{selon l'éq. (4.13)}}{=} f_{Y|X,Z,R}(y_i | x_i, z_i, \theta)\, f_{X,Z,R}(x_i, z_i, r_i | \phi), \end{aligned}$$

où θ et ϕ sont des paramètres distincts (Little et Rubin, 2019, chapitre 6). Il est ensuite possible d'estimer les paramètres θ de cette factorisation à partir des données d'enquête des répondants $R_i = 1$, de même que les paramètres de la distribution conjointe de X et Z par la combinaison des données des répondants et des données auxiliaires pour ces variables.

Voici deux cas particuliers examinés par Zangeneh et Little (2022) :

(1) Aucune covariable X et une seule variable de stratification *a posteriori* Z. L'hypothèse relative aux données manquantes de l'équation (4.13) se réduit ensuite à $\Pr(R_i = 1 | Z_i, Y_i, \psi) = \Pr(R_i = 1 | Z_i, \psi)$. L'estimation des paramètres de la distribution conditionnelle de Y compte tenu de Z est réalisée à partir des répondants de l'enquête, et celle des paramètres de la distribution marginale de Z, à partir des données auxiliaires sur Z. Si notamment Z est multinomial avec $\Pr(z_i = j | \phi) = \phi_j$, l'estimation à l'aide du maximum de vraisemblance de ϕ_j est tout simplement la proportion des chiffres auxiliaires dans la strate *a posteriori* j. Si, en outre, Y compte tenu de $Z = j$ est présumé normal, avec une moyenne μ_j et une variance σ_j^2, l'estimation par maximum de vraisemblance de la moyenne de la population Y qui en découle

est simplement la moyenne stratifiée *a posteriori* obtenue dans l'équation (4.3). Cet exemple simple est intéressant sur le plan théorique parce que le mécanisme de réponse n'est pas manquant au hasard, mais ignorable en ce qui a trait à l'inférence de vraisemblance; le fait d'être manquant au hasard est une condition suffisante, mais pas toujours nécessaire en matière d'ignorabilité.

(2) X est une seule variable catégorique et Z, un stratificateur catégorique *a posteriori*. Maintenant, la distribution conjointe de X, Z et R n'est pas mentionnée dans un modèle saturé et nécessite d'autres hypothèses afin d'établir le modèle. Un modèle « RAKE » restreint, avec données manquantes non au hasard, part de l'hypothèse que les distributions marginales de Z_1 et Z_2 sont différentes pour les répondants et les non-répondants, mais les rapports de cotes de Z_1 et Z_2 sont les mêmes pour les uns et les autres. Il en résulte un modèle exactement identifié. Le ratissage du tableau de fréquences des répondants de Z_1 et Z_2 jusqu'aux marges auxiliaires de Z_1 et Z_2 procure des estimations selon le maximum de vraisemblance de ϕ, selon ce modèle RAKE (Little et Wu, 1991; Little, 1993). L'estimateur poststratifié de la moyenne de Y est alors

$$\bar{Y}_{\text{rake}} = \sum_{j=1}^{J}\sum_{k=1}^{K} P_{jk}\left(\hat{\phi}\right)\bar{Y}_{jk}\left(\hat{\theta}\right),$$

où $P_{jk}\left(\hat{\phi}\right)$ est la proportion de la population pour $X = j, Z = k$ tirée du ratissage des fréquences de répondants jusqu'aux marges, et $\bar{Y}_{jk}\left(\hat{\theta}\right)$ est la moyenne estimée de Y compte tenu de $X = j, Z = k$ tirée du modèle pour Y compte tenu de X et Z.

Voici quelques commentaires concernant cette démarche :

(1) Il convient de souligner ici aussi qu'il s'agit d'un modèle avec des données manquantes non au hasard, et que ce modèle est moins fiable qu'un modèle ayant des données manquantes au hasard qui présume que la réponse dépend de X, mais pas de Z. Ce faisant, la méthode a tendance à afficher moins de biais que d'autres méthodes alternatives qui reposent sur une hypothèse de données manquantes au hasard pour obtenir des estimations convergentes.

(2) Comme dans le cas de l'exemple 4, la méthode du maximum de vraisemblance qu'emploie ce modèle ne comporte pas de modification des poids appliqués aux données dans une cellule $X = j, Z = k$ à l'aide de fonctions de distance arbitraires – les estimations selon le maximum de vraisemblance sont pleinement efficaces dans le modèle théorique.

(3) Il est possible de pallier la rareté de données dans les cellules par l'ajout de distributions *a priori* appropriées pour les paramètres et par l'application de méthodes bayésiennes. À titre d'exemple, pour l'inférence ayant trait aux paramètres θ de la distribution de Y compte tenu de $X = j, Z = k$, il serait possible de présumer d'une distribution *a priori* uniforme sur les

principaux effets de X et Z, mais d'une distribution *a priori* normale sur les interactions, en les faisant rétrécir jusqu'à zéro. Cela permet d'obtenir un rétrécissement de la moyenne de la cellule $\bar{y}_{jk}$ jusqu'aux moyennes ajustées tirées du modèle additif. Dans la terminologie fréquentiste, on parlera d'un modèle mixte d'analyse de la variance à effets principaux fixes et interactions aléatoires.

(4) Le principe de ratissage des marges X et Z est bien connu, mais il convient de souligner que si X est un stratificateur et Z, un stratificateur *a posteriori*, la démarche habituelle consiste à effectuer une itération du ratissage, en couplant d'abord la marge X, puis la marge Z : dans le modèle théorique, un ratissage itératif constitue la bonne procédure. Le ratissage correspond dans le cas présent à un maximum de vraisemblance, mais les formes bayésiennes de ratissage peuvent aussi servir à propager l'incertitude du paramètre.

Exemple 8. Analyse indirecte par modèles de mélange de schémas d'observation. Une analyse indirecte par modèles de mélange de schémas d'observation est une méthode d'évaluation du biais de non-réponse pour la moyenne d'une variable d'enquête Y assujettie à une non-réponse, en présence d'un ensemble de covariables observées pour les non-répondants et les répondants. Dans le passé, la quantité de données manquantes, mesurées en tant que taux de réponse, constituait la mesure la plus souvent utilisée pour évaluer la qualité d'une enquête. Mais les taux de réponse font abstraction des renseignements que renferment des covariables auxiliaires observées pour les non-répondants. Des méthodes reposant sur la probabilité estimée de non-réponse, comme l'indicateur R (Schouten, Cobben et Bethlehem, 2009) et la mesure q2 (Särndal et Lundström, 2010), ne tiennent pas compte de la solidité de l'association de la variable d'intérêt de l'enquête et de la probabilité de réponse, qui devrait sans doute être prise en compte dans l'évaluation du biais. Ces mesures supposent la présence de données manquantes au hasard, mais si des variables auxiliaires sont disponibles pour l'ajustement de la non-réponse, ce sont les déviations par rapport aux données manquantes au hasard qui entraînent un biais.

Andridge et Little (2011) proposent une analyse indirecte par modèles de mélange de schémas d'observation, méthode qui repose sur un modèle de mélange de schémas d'observation pour les non-réponses dans lequel sont combinées de manière simple et intuitive les principales caractéristiques de l'ajustement des non-réponses. Supposons que Y_i désigne la valeur d'un résultat d'enquête continu et $Z_i = \left(Z_{i1}, Z_{i2}, \ldots, Z_{ip}\right)$, la valeur des covariables p pour l'unité i dans l'échantillon. Seulement r des n unités échantillonnées répondent, de sorte que les données observées correspondent à $\left(Y_i, Z_i\right)$ pour $i = 1, \ldots, r$ et à Z_i pour $i = r+1, \ldots, n$. Supposons que R désigne l'indicateur de réponse, de sorte que pour l'unité i $R_i = 1$ si Y_i est observé et $R_i = 0$ si Y_i est manquant. Afin de réduire la dimensionnalité, nous remplaçons Z par une seule variable substitut X ayant la plus forte corrélation avec Y dans l'échantillon de répondants. Il est possible d'estimer cette variable substitut en effectuant une régression

de Y sur Z à l'aide des données des répondants, en incluant d'importants prédicteurs de Y, de même que des interactions et des termes non linéaires, si cela convient. Nous adoptons plus précisément le modèle de régression $E\left(Y \mid Z, R=1\right)=\alpha_0+\alpha Z$, et supposons $X=\alpha Z$. La distribution conjointe de X, Y et R à l'aide du modèle de mélange de schémas d'observation suivant, qui prend une forme comparable à celle présentée dans Little (1994) :

$$\begin{aligned}
&\left(X, Y \mid R=r\right) \sim G_2\left((\mu_x^{(r)}, \mu_y^{(r)}), \Sigma^{(r)}\right) \\
&R \sim \text{Bernoulli}(\pi) \\
&\Sigma^{(r)}=\begin{pmatrix} \sigma_{xx}^{(r)} & \rho^{(r)}\sqrt{\sigma_{xx}^{(r)}\sigma_{yy}^{(r)}} \\ \rho^{(r)}\sqrt{\sigma_{xx}^{(r)}\sigma_{yy}^{(r)}} & \sigma_{yy}^{(r)} \end{pmatrix}
\end{aligned}$$

où N_2 désigne la distribution bivariée normale. Nous nous intéressons au biais apparaissant dans la moyenne marginale de Y, que nous pouvons inscrire de cette façon : $\mu_y=\pi\mu_y^{(1)}+\left(1-\pi\right)\mu_y^{(0)}$. Andridge et Little (2011) présument que

$$\Pr\left(R=1 \mid Y, X\right) = f\left(X^*+\lambda Y\right),$$

pour une certaine fonction non spécifiée f et une constante connue λ. Dans le cas présent, $X^*=X\sqrt{\sigma_{yy}^{(1)}/\sigma_{xx}^{(1)}}$ est le substitut X proportionné de manière a obtenir la même variance que pour Y, ce qui facilite l'interprétation de λ grâce au positionnement de X et de Y à la même échelle. Ce mécanisme fait appel à des données manquantes au hasard lorsque $\lambda=0$, et dévie progressivement de celles-ci à mesure que λ augmente. Dans cette hypothèse, les paramètres sont exactement identifiés et l'estimation selon le maximum de vraisemblance de la moyenne de Y, en faisant la moyenne des schémas d'observation, correspond à

$$\hat{\mu}_y = \bar{y}_1+\left(\frac{n-r}{n}\right)\left(\frac{\lambda+\hat{\rho}}{\lambda\hat{\rho}+1}\right)\left(\bar{x}_0-\bar{x}_1\right),$$

où $\bar{x}_1, \bar{y}_1$ sont les moyennes de répondants de X et Y et $\bar{x}_0$, la moyenne de X pour les non-répondants. L'indice de biais correspond alors à l'ajustement de la moyenne de l'échantillon $\bar{y}_1$ considérée dans cette estimation, à savoir

$$\left(\frac{n-r}{n}\right)\left(\frac{\lambda+\hat{\rho}}{\lambda\hat{\rho}+1}\right)\left(\bar{x}_0-\bar{x}_1\right), \tag{4.14}$$

où $\hat{\rho}$ est la corrélation de l'échantillon de répondants. Cet ajustement intègre de manière simple et intuitive trois facteurs clés qui influent sur le possible biais – le taux de non-réponse $\left(n-r\right)/n$, la corrélation $\hat{\rho}$ entre X et Y, et la déviation de $\bar{x}_0$ par rapport à $\bar{x}_1$. L'indice croît notamment en fonction du taux de non-réponse et de l'écart de moyenne de X pour les répondants et les non-répondants.

Les données ne fournissent aucun renseignement à propos du paramètre λ – ce qui est habituellement le cas dans l'emploi de méthodes qui modélisent des déviations du mécanisme de données manquantes au hasard. À la suite de Little (1994), Andridge et Little (2011) proposent une analyse de la sensibilité dans laquelle on génère des estimations pour une fourchette de valeurs de λ entre 0 et l'infini, particulièrement 0, 1 et l'infini; le choix de 0 correspond à des données manquantes au hasard, celui de 1 présuppose que le biais de Y est le même que le biais de la variable substitut X^*, alors que le choix de l'infini correspond à la déviation la plus marquée par rapport aux données manquantes au hasard; dans ce cas, les estimations présentent la variance la plus élevée. Puisque λ varie de 0 à l'infini, le facteur mitoyen $(\lambda+\hat{\rho})/(\lambda\hat{\rho}+1)$ varie de $\hat{\rho}$ (quand $\hat{\mu}_y$ correspond à l'estimateur par la régression classique de la moyenne) à $1/\hat{\rho}$ (quand $\hat{\mu}_y$ correspond à l'estimateur par la régression inverse proposée par Brown [1990]). La sensibilité de l'estimation pour le choix de λ est faible lorsque $\hat{\rho}$ se situe près de 1, ce qui indique la présence d'une variable substitut robuste, et forte lorsque $\hat{\rho}$ se situe près de 0, auquel cas, la variable substitut est faible. Il s'avère donc essentiel de disposer de variables auxiliaires constituant de bons prédicteurs des résultats de l'enquête.

Des versions bayésiennes du modèle de mélange de schémas d'observation sont déjà au point, ce qui permet la propagation de l'erreur dans les paramètres du modèle. Il est en outre possible d'appliquer le modèle de manière à créer des imputations multiples de valeurs de non-répondants, ce qui permet d'intégrer à l'indice les éléments d'un plan d'échantillonnage complexe.

Plus récemment, le modèle a été utilisé pour développer des indices d'écart par rapport à l'échantillonnage aléatoire, en l'appliquant à la sélection du modèle plutôt qu'à la non-réponse (Little, West, Boonstra et Hu, 2020; Boonstra, Little, West, Andridge et Alvaredo-Leiton, 2021). Le perfectionnement de ces travaux vise (a) à remplacer le paramètre λ par $\phi=\lambda/(1+\lambda)$, qui constitue une meilleure paramétrisation parce que ϕ va de 0 (données manquantes au hasard) à 1, et (b) (avec l'ajout d'autres hypothèses) à permettre aux données manquantes de dépendre également de variables auxiliaires orthogonales à X. L'expansion de la méthode a aussi permis de gérer des résultats binaires (Andridge, West, Little, Boonstra et Alvarado-Leiton, 2019) et des indices de biais potentiels dans les coefficients de régression (West, Little, Andridge, Boonstra, Ware, Pandit et Alvarado-Leiton, 2021).

5. Conclusion : 10 raisons d'adopter Bayes pour les inférences d'enquête

Mes exemples visent à donner une idée de l'ampleur des possibilités de la modélisation bayésienne pour les données d'enquête, sans pour autant se révéler un tant soit peu exhaustifs. Bayes est aussi utile dans des domaines que je n'ai pas abordés, notamment l'échantillonnage à plusieurs degrés, les séries chronologiques, modèles d'analyse de classes latentes et de facteurs, les erreurs de mesures, la

combinaison de données issues de différentes sources, la création de données synthétiques pour prévenir la divulgation, et ainsi de suite. Je termine par un résumé des raisons qui m'amènent à favoriser l'approche d'inférence bayésienne pour les enquêtes :

1. L'approche fondée sur le plan de sondage est asymptotique et trop limitée pour gérer les divers problèmes relatifs aux inférences tirées d'enquêtes, qu'elles soient de nature probabiliste ou non probabiliste.
2. L'approche bayésienne est à la fois unifiée et suffisamment souple pour gérer les divers problèmes rencontrés dans les enquêtes, en plus de comprendre une modélisation de la superpopulation sous forme d'inférence tirée d'un vaste échantillon.
3. Des modèles bayésiens choisis avec soin peuvent fournir des intervalles crédibles qui possèdent de bonnes propriétés fondées sur le plan de sondage dans un échantillonnage répété. Il est notamment possible de modéliser la pondération, la stratification et la stratification *a posteriori* à l'aide de covariables et d'intégrer la mise en grappes à partir de modèles bayésiens hiérarchiques. Des modèles souples qui incorporent des caractéristiques du plan rendent inutiles des démarches hybrides comme l'estimation assistée par un modèle (par exemple Särndal, Swensson et Wretman, 1992).
4. Les premières critiques de la méthode de modélisation portent sur des modèles n'intégrant aucune caractéristique du plan et, par conséquent, qui sont sensibles à l'erreur de spécification du modèle. Il est possible, voire souhaitable d'éviter le recours à ces modèles.
5. Les calculs bayésiens d'intégration de paramètres de nuisance aboutissent à des inférences ayant de bonnes propriétés fréquentistes à partir de petits comme de grands échantillons.
6. L'approche fondée sur le plan de sondage pour traiter certains problèmes et fondée sur un modèle dans d'autres situations donne lieu à des incohérences logiques (voir, par exemple, Little, 2012, section 4.3); l'approche bayésienne fournit des inférences unifiées et logiquement cohérentes.
7. Dans l'approche bayésienne, la spécification des distributions *a priori* constitue un point fort, non une faiblesse, parce qu'elle donne une plus grande souplesse à la modélisation. Pour certains problèmes, des *a priori* « objectives » faibles aboutissent à des résultats équivalant à ceux obtenus de solutions fréquentistes conventionnelles. Pour d'autres problèmes, des *a priori* « subjectives » plus robustes donnent des réponses utiles dans le cas de modèles non identifiés, comme la non-réponse des données manquantes non au hasard.
8. Les récentes avancées méthodologiques et l'augmentation de la puissance informatique ont grandement réduit les difficultés de calcul dans l'approche bayésienne.
9. La modélisation inscrit la recherche par enquête dans le processus principal de la modélisation statistique pour d'autres types de données. Le paradigme bayésien gère aisément les caractéristiques particulières de la recherche par enquête – complexité du plan de sondage et accent sur des quantités de population finie.

10. L'approche bayésienne ne nie pas l'utilité de l'échantillonnage probabiliste pour le plan, elle qui se révèle extrêmement précieuse pour l'obtention des inférences robustes limitant le recours à des hypothèses contestables concernant la représentativité de l'échantillon.

Remerciements

Je tiens à remercier le Comité du prix Waksberg de m'avoir donné la chance de présenter ces travaux, de même que Yajuan Si et un examinateur pour leurs commentaires constructifs.

Bibliographie

Andridge, R.H., et Little, R.J. (2011). Proxy pattern-mixture analysis for survey nonresponse. *Journal of Official Statistics*, 27, 2, 153-180.

Andridge, R.R., West, B.T., Little, R.J.A., Boonstra, P.S. et Alvarado-Leiton, F. (2019). Indices of non-ignorable selection bias for proportions estimated from non-probability samples. *Journal of the Royal Statistical Society, Series C, Applied Statistics*, 68, 5, 1465-1483.

Basu, D. (1971). An essay on the logical foundations of survey sampling, Part I. Dans *Foundations of Statistical Inference*, (Éds., V.P. Godambe et D.A. Sprott), Toronto: Holt, Rinehart and Winston, 203-242.

Berger, J.O., et Wolpert, R.L. (1988). The likelihood principle. *Institute of Mathematical Statistics Lecture Notes-Monograph Series*, 6, 1-199.

Binder, D.A. (1982). Non-parametric Bayesian models for samples from finite populations. *Journal of the Royal Statistical Society, Series B,* 44,3, 388-393.

Birnbaum, A. (1962). On the foundations of statistical inference (avec discussion). *Journal of the American Statistical Association*, 57, 269-326.

Boonstra, P.S., Little, R.J., West, B.T., Andridge, R.R. et Alvaredo-Leiton, F. (2021). A simulation study of diagnostics for bias in non-probability samples. *Journal of Official Statistics*, 37, 3, 751-769.

Brewer, K. (2013). Trois controverses dans l'histoire de l'échantillonnage. *Techniques d'enquête*, 39, 2, 275-289. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2013002/article/11883-fra.pdf.

Brown, C.H. (1990). Protecting against nonrandomly missing data in longitudinal studies. *Biometrics*, 46, 143-155.

Buja, A., Berk, R., Brown, L., George, E., Pitkin, E., Zhan, L. et Zhang, K. (2019). Models as approximations 1: Consequences illustrated with linear regression. *Statistical Science*, 34, 4, 580-583.

Chen, Q., Elliott, M.R., Haziza, D., Yang, Y., Ghosh, M., Little, R.J., Sedransk, J. et Thompson, M. (2017). Approaches to improving survey-weighted estimates. *Statistical Science*, 32, 2, 227-248.

Dean, N., et Pagano, M. (2015). Evaluating confidence interval methods for binomial proportions in clustered surveys. *Journal of Survey Statistics and Methodology*, 3, 484-503.

Elliott, M.R. (2007). Réduction bayésienne des poids pour les modèles de régression linéaire généralisée. *Techniques d'enquête*, 33, 1, 27-40. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2007001/article/9849-fra.pdf.

Elliott, M.R., et Little, R.J. (2000). Model-based approaches to weight trimming. *Journal of Official Statistics*, 16, 191-210.

Elliott, M.R., et Valliant, R. (2017). Inference for nonprobability samples. *Statistical Science*, 32, 2, 249-264.

Ericson, W.A. (1969). Subjective Bayesian models in sampling finite populations (avec discussion). *Journal of the Royal Statistical Society, Series B*, 31, 195-233.

Fienberg, S.E. (2011). Bayesian models and methods in public policy and government settings. *Statistical Science*, 26, 2, 212-226.

Firth, D., et Bennett, K.E. (1998). Robust models in probability sampling. *Journal of the Royal Statistical Society, Series B*, 60, 3-21.

Franco, C., Little, R.J., Louis, T.A. et Slud, E.V. (2019). Comparative study of confidence intervals for proportions in complex sample surveys. *Journal of Survey Statistics and Methodology*, 7, 3, 334-364.

Gelman, A. (2007). Struggles with survey weighting and regression modeling (avec discussion). *Statistical Science*, 22, 2, 153-188.

Gelman, A., et Little, T.C. (1997). Stratification a posteriori en un grand nombre de catégories par régression logistique hiérarchique. *Techniques d'enquête*, 23, 2, 135-145. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/1997002/article/3616-fra.pdf.

Ghosh, M., et Meeden, G. (1997). *Bayesian Methods for Finite Population Sampling*. Londres: Chapman and Hall.

Ghosh, M., Reid, N. et Fraser, D.A.S. (2010). Ancillary statistics: A review. *Statistica Sinica*, 20, 1309-1332.

Hájek, J. (1971). Comment on a paper by D. Basu. Dans *Foundations of Statistical Inference*, (Éds., V.P. Godambe et D.A. Sprott), Toronto: Holt, Rinehart and Winston, 236.

Hansen, M.H., Madow, W.G. et Tepping, B.J. (1983). An evaluation of model-dependent and probability-sampling inferences in sample surveys (avec discussion). *Journal of the American Statistical Association*, 78, 776-793.

Holt, D., et Smith, T.M.F. (1979). Post stratification. *Journal of the Royal Statistical Society, Series A*, 142, 1, 33-46.

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

Kish, L. (1992). Weighting for unequal Pi. *Journal of Official Statistics*, 8, 183-200.

Kish, L. (1995). The hundred years' wars of survey sampling. *Statistics in Transition*, 2, 813-830. Reproduit comme chapitre 1 de *Leslie Kish: Selected Papers*, (2003, Éds., G. Kalton et S. Heeringa), New York: John Wiley & Sons, Inc.

Kish, L., et Frankel, M.R. (1974). Inferences from complex samples (avec discussion). *Journal of the Royal Statistical Society, Series B*, 36, 1-37.

Lazzeroni, L.C., et Little, R.J. (1998). Random-effects models for smoothing post-stratification weights. *Journal of Official Statistics*, 14, 61-78.

Leon-Novelo, L.G., et Savitsky, T.D. (2019). Fully Bayesian estimation under informative sampling. *Electronic Journal of Statistics*, 13, 1608-1645.

Little, R.J. (1993). Post-stratification: A modeler's perspective. *Journal of the American Statistical Association*, 88, 1001-1012.

Little, R.J. (1994). A class of pattern-mixture models for normal missing data. *Biometrika*, 81, 3, 471-483.

Little, R.J. (2003a). The Bayesian approach to sample survey inference. Dans *Analysis of Survey Data*, (Éds., R.L. Chambers et C.J. Skinner), New York: John Wiley & Sons, Inc., 49-57.

Little, R.J. (2003b). Bayesian methods for unit and item nonresponse. Dans *Analysis of Survey Data*, (Éds., R.L. Chambers et C.J. Skinner), New York: John Wiley & Sons, Inc., 289-306.

Little, R.J. (2004). To model or not to model? Competing modes of inference for finite population sampling. *Journal of the American Statistical Association*, 99, 546-556.

Little, R.J. (2006). Calibrated Bayes: A Bayes/frequentist roadmap. *The American Statistician*, 60, 3, 213-223.

Little, R.J. (2012). Calibrated Bayes: An alternative inferential paradigm for official statistics (avec discussion). *Journal of Official Statistics*, 28, 3, 309-372.

Little, R.J. (2014). Survey sampling: Past controversies, current orthodoxies, and future paradigms. Dans *Past, Present and Future of Statistical Science*, COPSS 50th Anniversary Volume, (Éds., X. Lin, D.L. Banks, C. Genest, G. Molenberghs, D.W. Scott et J.-L. Wang), CRC Press.

Little, R.J. (2015). Calibrated Bayes, an inferential paradigm for official statistics in the era of big data. *Statistical Journal of the International Association of Survey Statisticians*, 31, 4, 555-563.

Little, R.J. (2019). Comment on "Models as approximations 1: Consequences illustrated with linear regression" by A. Buja et al. *Statistical Science*, 34, 4, 580-583.

Little, R.J. (2022). A note about the definition of propensity weights. À paraître dans le *Journal of Survey Statistics and Methodology*.

Little, R.J., et Rubin, D.B. (2019). *Statistical Analysis with Missing Data*, 3rd edition. New York: John Wiley & Sons, Inc.

Little, R.J., et Vartivarian, S. (2005). La pondération pour la non-réponse augmente-t-elle la variance des moyennes de sondage ? *Techniques d'enquête*, 31, 2, 175-183. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2005002/article/9046-fra.pdf.

Little, R.J., et Wu, M.M. (1991). Models for contingency tables with known margins when target and sampled populations differ. *Journal of the American Statistical Association*, 86, 87-95.

Little, R.J., West, B.T., Boonstra, P.S. et Hu, J. (2020). Measures of the degree of departure from ignorable sample selection. *Journal of Survey Statistics and Methodology*, 8, 5, 932-964.

Royall, R.M., et Cumberland, W.G. (1981). The finite population linear regression estimator and estimator of its variance-an empirical study. *Journal of the American Statistical Association*, 76, 924-930.

Royall, R.M., et Cumberland, W.G. (1985). Conditional coverage properties of finite population confidence intervals. *Journal of the American Statistical Association*, 80, 355-359.

Rubin, D.B. (1976). Inference and missing data. *Biometrika*, 53, 581-592.

Rubin, D.B. (1978). Bayesian inference for causal effects: The role of randomization. *Annals of Statistics*, 6, 1, 34-58.

Rubin, D.B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Annals of Statistics*, 12, 1151-1172.

Rubin, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley & Sons, Inc.

Rubin, D.B. (2019). Le calage conditionnel et le sage statisticien. *Techniques d'enquête*, 45, 2, 199-210. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/pub/12-001-x/2019002/article/00010-fra.pdf.

Ruppert, D., Wand, M.P et Carroll, R.J. (2003). *Semiparametric Regression*, Cambridge University Press.

Särndal, C.-E., et Lundström, S. (2010). Plan d'estimation : détermination de vecteurs auxiliaires en vue de réduire le biais de non-réponse. *Techniques d'enquête*, 36, 2, 141-156. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2010002/article/11376-fra.pdf.

Särndal, C.-E., Swensson, B. et Wretman, J.H. (1992). *Model Assisted Survey Sampling*. New York: Springer Verlag.

Schouten, B., Cobben, F. et Bethlehem, J. (2009). Indicateurs de la représentativité de la réponse aux enquêtes. *Techniques d'enquête*, 35, 1, 107-121. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2009001/article/10887-fra.pdf.

Sedransk, J. (2008). Assessing the value of Bayesian methods for inference about finite population quantities. *Journal of Official Statistics*, 24, 495-506.

Si, Y, Trangucci, R., Gabry, J.S. et Gelman, A. (2020). Ajustement de pondération hiérarchique bayésienne et inférence d'enquête. *Techniques d'enquête*, 46, 2, 193-228. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2020002/article/00003-fra.pdf.

Smith, T.M.F. (1976). The foundations of survey sampling: A review (avec discussion). *Journal of the Royal Statistical Society, Series A*, 139, 183-204.

Smith, T.M.F. (1994). Sample surveys 1975-1990: An age of reconciliation? (avec discussion). *Revue Internationale de Statistique*, 62, 5-34.

Szpiro, A.A., Rice, K.M. et Lumley, T. (2010). Model-robust regression and a Bayesian "sandwich" estimator. *Annals of Applied Statistics*, 4, 4, 2099-2113.

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

West, B.T., Little, R.J., Andridge, R.R., Boonstra, P.S., Ware, E.B., Pandit, A. et Alvarado-Leiton, F. (2021). Measures of selection bias in regression coefficients estimated from non-probability samples. *The Annals of Applied Statistics*, 15(3), 1556-1581.

Yang, Y., et Little, R.J. (2015). A comparison of doubly robust estimators of the mean with missing data. *Journal of Statistical Computation and Simulation*, 85, 16, 3383-3403.

Zangeneh, S.Z., et Little, R.J. (2015). Bayesian inference for the finite population total in heteroscedastic probability proportional to size samples. *Journal of Survey Statistics and Methodology*, 3, 162-192.

Zangeneh, S.Z., et Little, R.J. (2022). Likelihood based estimation of the finite population mean with post-stratification information under nonignorable nonresponse. À paraître dans la *Revue Internationale de Statistique*.

Zhang, G., et Little, R.J. (2009). Extensions of the penalized spline of propensity prediction method of imputation. *Biometrics*, 65, 3, 911-918.

Zheng, H., et Little, R.J. (2003). Penalized spline model-based estimation of the finite population total from probability-proportional-to-size samples. *Journal of Official Statistics*, 19, 2, 99-117.

Zheng, H., et Little, R.J. (2005). Inference for the population total from probability-proportional-to-size samples based on predictions from apenalized spline nonparametric model. *Journal of Official Statistics*, 21, 1-20.