Actes du Symposium de 2018 de Statistique Canada
Combiner pour vaincre : innovations dans l'utilisation de multiples sources de données

Mesurer l'incertitude en présence de multiples sources de données

Sharon L. Lohr[1]

Résumé

Dans un échantillon probabiliste avec réponse complète, la marge d'erreur fournit une mesure fiable et théoriquement justifiée de l'incertitude. Toutefois, lorsque des estimations provenant de multiples échantillons ou sources de données administratives sont combinées, les marges d'erreur traditionnelles sous-estiment l'incertitude — les différences entre les statistiques de diverses sources dépassent souvent la variabilité d'échantillonnage estimée. J'examinerai des méthodes proposées pour mesurer l'incertitude découlant de sources de données combinées en appliquant le tout à l'estimation de la prévalence du tabagisme et des taux d'agressions sexuelles, et je décrirai quelques orientations possibles de la recherche.

Mots-clés : marge d'erreur; calage; enquêtes complexes; calcul de la moyenne des modèles bayésiens; agression sexuelle; prévalence du tabagisme.

1. Introduction

1.1 Combiner des données pour répondre aux besoins en matière de statistiques

On connaît bien les défis relatifs aux statistiques officielles. Les taux de réponse aux enquêtes diminuent. Parallèlement, on assiste à une demande croissante d'information plus détaillée, disponible plus rapidement et moins coûteuse. La conférence avait pour thème la façon de répondre aux demandes croissantes en combinant des données ou des statistiques provenant de plusieurs sources, soit des enquêtes, des données administratives, des données de capteurs ou des données recueillies sur Internet.

L'objectif est de permettre une estimation plus exacte (ou moins coûteuse) des quantités d'intérêt grâce aux statistiques provenant des données combinées. Il reste toutefois une difficulté essentielle, à savoir l'évaluation de l'exactitude des statistiques issues de données combinées. La plupart des estimations d'enquête sont accompagnées d'intervalles de confiance qui indiquent l'ampleur de l'incertitude due à la variabilité d'échantillonnage dans les estimations. En général, les statistiques issues de données administratives et de données de capteurs n'ont pas de variabilité d'échantillonnage et sont habituellement présentées sans mesure d'incertitude. Il existe cependant d'autres types d'erreurs, qu'on peut mieux connaître en étudiant de multiples sources de données.

Dans le présent article, j'aborderai certaines questions relatives à la mesure de l'incertitude pour trois méthodes couramment utilisées dans la combinaison de données — les méthodes utilisant des bases de sondage multiples, les modèles hiérarchiques et le calage — à partir de trois exemples. Je présenterai deux d'entre eux dans la présente section et le troisième à la section 5.

1.2 Tabagisme chez les adultes américains

La figure 1.2-1 présente des estimations ponctuelles et des intervalles de confiance pour le pourcentage d'adultes américains ayant fumé au moins 100 cigarettes dans leur vie. Siegfried et coll. (2017) ont calculé les statistiques indiquées à la figure 1.2-1 à partir de cinq enquêtes probabilistes représentatives à l'échelle nationale : la *Tobacco Use Supplement of the Current Population Survey* (TUS-CPS), l'étude *Population Assessment of Tobacco and Health*

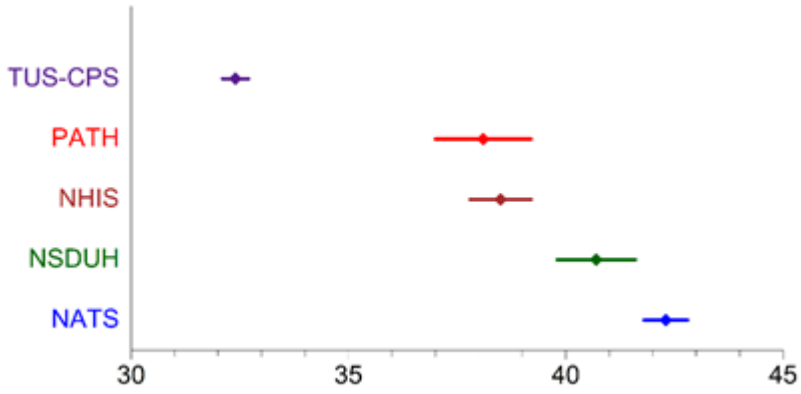
[1]Sharon L. Lohr, professeur émérite, Arizona State University, Tempe, AZ, 85287-1804, États-Unis, www.sharonlohr.com

(PATH), la *National Health Interview Survey* (NHIS), la *National Survey of Drug Use and Health* (NSDUH) et la *National Adult Tobacco Survey* (NATS).

Toutes les enquêtes ont été menées à peu près à la même période (de 2013 à 2015) et avaient des populations cibles comprenant la population civile des États-Unis ne résidant pas en établissement âgée de 18 ans ou plus, si bien qu'il était possible de calculer une estimation pour les adultes américains âgés de 18 ans ou plus à partir de chacune des enquêtes. Certaines enquêtes demandaient directement aux personnes interrogées si elles avaient fumé au moins 100 cigarettes au cours de leur vie; pour les autres, l'estimation pouvait être calculée à partir des réponses à plusieurs questions. Cependant, toutes les estimations ponctuelles de la figure 1.2-1 estiment, en théorie, environ le même paramètre de population.

Les différences entre les estimations ponctuelles sont plus grandes que ce qu'on pourrait attendre au vu de l'étroitesse des intervalles de confiance, qui indiquent uniquement l'erreur d'échantillonnage. L'estimation de la TUS-CPS, en particulier, est nettement plus basse que les autres, et l'estimation de la NATS est sensiblement plus élevée.

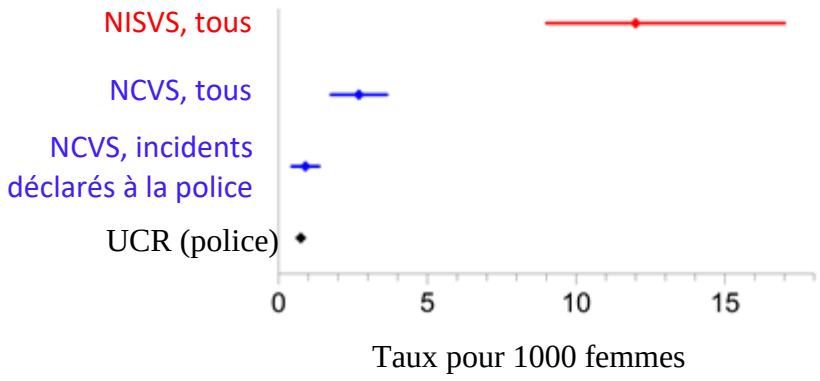
Figure 1.2.-1
Pourcentage d'adultes américains ayant fumé au moins 100 cigarettes, avec intervalles de confiance de 95 %



1.3 Agressions sexuelles en 2015

La figure 1.3-1 présente quatre estimations des viols et des agressions sexuelles aux États-Unis en 2015, avec des intervalles de confiance de 95 %. Lohr (2019) étudie les sources de données de ces estimations.

Figure 1.3.-1
Taux de viols et d'agressions sexuelles de femmes aux États-Unis en 2015, selon trois sources, avec intervalles de confiance de 95 %



Les statistiques de la ligne supérieures proviennent de la *National Intimate Partner and Sexual Violence Survey* (NISVS) de 2015, tirées du tableau 1 de Smith et coll. 2017. Les statistiques des deuxième et troisième lignes sont tirées de la *National Crime Victimization Survey* de 2015 (NCVS, Bureau of Justice Statistics , 2018a, 2018b). La deuxième ligne comprend tous les viols et agressions sexuelles déclarés par des femmes selon la NCVS de 2015; la troisième ligne comprend seulement la victimisation connue par la police d'après les répondantes à la NCVS.

Les deux enquêtes comportent un certain nombre de différences. La NISVS estime le nombre de femmes adultes des États-Unis, pour 1 000 femmes, ayant subi au moins un viol ou une tentative de viol au cours des 12 mois précédents. La statistique de la NCVS estime le nombre de femmes victimes de viol (viols perpétrés et tentatives) et d'agression sexuelle pour 1 000 femmes au cours des 12 mois précédents. Les âges pris en compte diffèrent également : la NISVS réalise des entrevues auprès des femmes âgées d'au moins 18 ans alors que la NCVS interroge aussi des personnes âgées de 12 à 17 ans (qui présentent généralement l'un des taux les plus élevés de victimisation par agression sexuelle). En effet, la NCVS comprend plus de types de crimes (autres types d'agression sexuelle et viol), comprend les victimisations multiples par femme et comprend un groupe d'âge supplémentaire qui enregistre habituellement des taux élevés d'agression sexuelle. Pour toutes ces raisons, on s'attendrait à ce que la statistique de la NCVS soit plus élevée que celle de la NISVS, mais elle est en fait plus basse. On peut en grande partie expliquer la différence par les questions différentes utilisées dans les deux enquêtes : bien que les définitions du viol soient semblables dans la NCVS et la NISVS, les questions mesurent en fait des paramètres de population différents. Ce n'est toutefois pas la seule explication (Lohr, 2019).

L'estimation par la NCVS des viols et des agressions sexuelles déclarées à la police se rapproche davantage de la quatrième estimation de la figure 1.3-1, tirée du système Uniform Crime Reporting (UCR) du *Federal Bureau of Investigation* (FBI). Le système UCR dénombre les crimes dont les organismes d'application de la loi ont eu connaissance. Le tableau 1 du FBI (2016) fait état de 126 134 viols en 2015, mais il ne présente pas de répartition selon le sexe. Pour obtenir le taux de victimisation estimé chez les femmes, j'ai réparti ces données entre hommes et femmes selon les proportions de victimes indiquées dans les données du *National Incident Based Reporting System* de 2015 (Puzzanchera et coll., 2017).

Les données du système UCR visent à recenser les crimes connus des organismes d'application de la loi. Le FBI n'indique pas de mesure de l'incertitude pour les estimations ni de mesure de la variabilité à partir de la procédure d'imputation utilisée pour les données manquantes. Les autres types d'erreurs susceptibles de se trouver dans une erreur quadratique moyenne, comme la variabilité de la classification de la criminalité et de son enregistrement par les organismes d'application de la loi, n'ont pas été suffisamment étudiés. Ainsi, l'estimation ponctuelle de la statistique du système UCR à la figure 1.3-1 n'est pas accompagnée d'un intervalle de confiance, mais cela ne signifie pas que l'estimation n'a pas d'erreur, mais seulement qu'on n'a pas de bonne mesure de l'incertitude.

Dans les figures 1.2-1 et 1.3-1, la variabilité entre les estimations provenant de différentes sources est supérieure à la variabilité au sein d'une source, qui est donnée par les intervalles de confiance. En ce qui concerne les statistiques de la NCVS et de la NISVS, cela s'explique en partie par le fait que les questions posées dans le cadre de la NISVS entraînent plus de déclarations d'agression sexuelle que les questions posées dans le cadre de la NCVS. Pour ce qui

est des statistiques sur le tabagisme, les différences pourraient être causées par des erreurs systématiques et par les méthodes employées par les enquêtes (Siegfried et coll., 2017).

2. Méthodes de combinaison de données

Lohr et Raghunathan (2017) et l'Académie nationale des sciences, de l'ingénierie et de la médecine (2017) décrivent les méthodes statistiques pouvant servir à la combinaison de données. Il s'agit (1) du couplage d'enregistrements, (2) de l'estimation pour les petits domaines, (3) de l'imputation, (4) des méthodes utilisant des bases de sondage multiples, (5) des modèles hiérarchiques et (6) du calage. Dans le présent article, je décris brièvement certaines questions liées à la mesure de l'incertitude dans les trois dernières méthodes.

Toutes les méthodes comportent des procédures établies de mesure de l'incertitude, mais les mesures obtenues sont parfois irréalistes et trop faibles. L'incertitude dans les statistiques issues de données combinées est causée par :

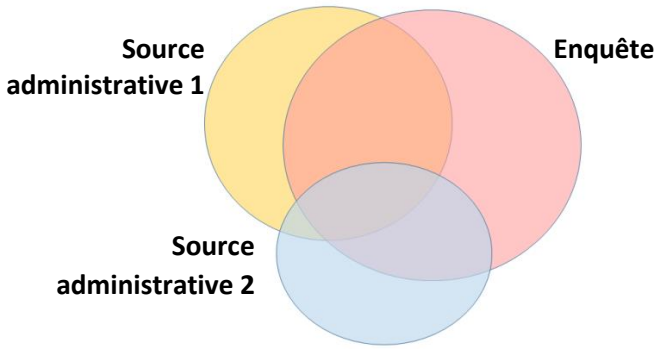
A. une erreur d'échantillonnage dans une des sources;
B. une erreur systématique dans une des sources;
C. les différences entre sources;
D. les modèles statistiques utilisés pour la combinaison de données.

La plupart des mesures d'incertitude publiées comprennent la variabilité d'échantillonnage, mais, comme nous l'avons vu dans les exemples de la section 1, les mesures ne tiennent pas compte des erreurs systématiques ni des autres différences susceptibles d'entraîner des variations entre les statistiques issues de différentes sources.

3. Méthodes utilisant une base de sondage multiple

La figure 3-1 présente la structure d'une base de sondage multiple avec trois sources de données potentielles. Chaque source couvre seulement une partie de la population. Les enquêtes avec des bases de sondage multiples sont souvent utilisées avec des enquêtes probabilistes indépendantes, mais dans la figure 3-1, deux des sources sont des données administratives, comme le système UCR ou les dossiers de santé informatisés, qui peuvent être des recensements.

Figure 3.-1
Trois sources de données (bases de sondage)



L'estimateur d'un total de population pour la figure 3-1 est la somme des sept totaux de domaine estimés, où trois domaines se trouvent dans une seule source de données, trois dans exactement deux sources, et le dernier dans les trois sources.

Et que dire de l'incertitude? Pour une enquête utilisant une base de sondage multiple, la variance du total de population estimé est une fonction des matrices de covariance du vecteur des totaux de domaine estimés pour chaque source (Lohr, 2011). Cependant, cette variance rend compte uniquement de l'incertitude découlant de l'erreur d'échantillonnage, ce qui n'a pas d'incidence sur les sources de données administratives. Il se peut que l'erreur quadratique moyenne du total estimé soit nettement plus grande que la variance si certaines estimations de domaine sont biaisées pour les totaux de population du domaine correspondant.

Les différences entre les estimations des enquêtes illustrées dans les figures 1.2-1 et 1.3-1 donnent des raisons de remettre en question l'hypothèse de la méthode à base de sondage multiple selon laquelle chaque source fournit des estimations sans biais des mêmes paramètres de domaines de population. Toutefois, l'estimation par la NCVS du nombre de viols déclarés à la police correspond à la statistique du FBI (2016), et il est possible d'utiliser des méthodes à base de sondage multiple avec ces deux sources de données. Pour évaluer l'incertitude des estimations combinées, il faut en savoir davantage sur les erreurs affectant les deux sources de données.

Quand toutes les sources ou certaines sources ont une couverture incomplète de la population, les méthodes à base de sondage multiple peuvent donner des estimations, qui s'appliquent à la population définie par l'union de toutes les bases de sondage. Il faut toutefois savoir à combien de bases appartient chaque point de données, de chaque source. L'incident déclaré par un répondant de la NCVS est-elle également dans les enregistrements du FBI (c'est-à-dire dans le chevauchement des deux bases de sondage), ou se situe-t-elle dans la partie de la population couverte seulement par la NCVS? À l'inverse, un viol déclaré aux organismes d'application de la loi est-il également dans la portée de la NCVS (ce qui ne sera pas le cas s'il s'agit d'un enfant, d'un résident de pays étranger ou d'une personne résidant en établissement)? Si les classifications de domaines sont inexactes, il se peut que les estimations des totaux de population soient biaisées et les estimations de l'incertitude trop faibles. Lohr (2011) et Lin (2013) ont présenté des méthodes d'ajustement des estimations et de leurs variances pour une possible classification erronée de domaine.

Des aspects du paradigme à base de sondage multiple s'appliquent aussi à de nombreuses autres méthodes de combinaison de données. En effet, dans de nombreuses méthodes, le problème fondamental consiste à déterminer la couverture et le chevauchement des différentes sources. Dans le cas des modèles hiérarchiques, de l'estimation de petits domaines, de l'imputation et du calage, on suppose souvent implicitement que les sources couvrent la même population. On peut évaluer cette hypothèse à l'aide de méthodes à base de sondage multiple et les utiliser en association avec d'autres modèles statistiques en cas de couverture incomplète par les sources.

4. Modèles hiérarchiques

Les modèles hiérarchiques, souvent utilisés en biostatistique à des fins de méta-analyse, tiennent compte expressément de l'hétérogénéité entre sources dans les estimations de l'incertitude. Prenons comme exemple le modèle de Manzi et coll. (2011) pour la moyenne $\bar{y}_{dj}$ du domaine d et de la source de données j :

$$\bar{y}_{dj} = \theta_d + \delta_{dj} + e_{dj}, \tag{4.1}$$

où θ_d est la moyenne globale dans le domaine d; δ_{dj} est un effet aléatoire pour l'écart de la source j par rapport à la moyenne de domaine globale, qu'on suppose suivre une distribution $N(\Delta_j, \tau_j^2)$; et e_{dj} est l'erreur d'échantillonnage de $\bar{y}_{dj}$. Dans l'équation (4.1), on suppose que les moyennes de domaine de la source j ont un biais moyen Δ_j. Ce modèle et les modèles connexes permettent la modélisation explicite du biais de chaque source.

Pour que les paramètres du modèle de l'équation (4.1) soient identifiables, il faut fournir plus d'information sur les paramètres θ_d, en définissant une source ou une combinaison de sources comme étant sans biais pour une fonction des paramètres. Si l'on combine les estimations sur le tabagisme de la figure 1-2.1, quelle source, le cas échéant, serait sans biais? Les modèles hiérarchiques ont des hypothèses fortes sur le biais, la forme du modèle et la couverture de la population (les problèmes touchant les enquêtes à base de sondage multiple ont également une incidence sur les

modèles hiérarchiques). Cependant, ces hypothèses sont explicites et certaines peuvent être mises à l'épreuve empiriquement.

Les modèles hiérarchiques présentent un avantage considérable : ils rendent compte de l'hétérogénéité entre les sources de données dans la distribution a posteriori des paramètres. La variance a posteriori peut être supérieure à la variance d'échantillonnage de chaque source individuelle.

5. Calage

5.1 Variance des estimateurs calés

Observons maintenant le calage, qui est probablement la méthode la plus courante de combinaison de données d'enquête avec les données d'une autre source. La variable d'intérêt dans l'enquête est notée y, et le vecteur de l'information auxiliaire, mesuré dans l'enquête et dans une source externe de totaux de contrôle, est notée $\boldsymbol{x}$. Le calage ajuste les poids d'enquête de façon à ce que $\widehat{\boldsymbol{X}}$ (le total estimé de la population $\boldsymbol{x}$ de l'enquête) soit égal à $\boldsymbol{X}$ (le total de la population de $\boldsymbol{x}$ de la source externe).

Dans le cas particulier d'une stratification a posteriori, la variance de l'estimateur post-stratifié $\hat{Y}_{ps} = \boldsymbol{X}'\widehat{\overline{\boldsymbol{Y}}}$, où $\boldsymbol{X}$ est le vecteur des totaux de contrôle des post-strates G et $\widehat{\overline{\boldsymbol{Y}}} = (\hat{Y}_1/\hat{X}_1, \ldots, \hat{Y}_G/\hat{X}_G)'$ est le vecteur des moyennes des post-strates estimé à partir de l'enquête. La variance de $\hat{Y}_{ps}$ est

$$V(\hat{Y}_{ps}) \approx \boldsymbol{X}'V\left(\widehat{\overline{\boldsymbol{Y}}}\right)\boldsymbol{X}. \tag{5.1}$$

La stratification a posteriori réduit presque toujours la variance de l'estimateur du total de population. Cependant, le fait qu'elle réduise l'erreur quadratique moyenne en cas de non-réponse dépend du mécanisme de non-réponse et des propriétés des totaux de contrôle $\boldsymbol{X}$.

Dever et Valliant (2010, 2016) ont montré que si les totaux de contrôle provenaient d'une enquête auxiliaire (l'*American Community Survey* est par exemple souvent utilisée comme source de totaux de contrôle) et non d'un recensement, et qu'ils présentaient donc une variabilité d'échantillonnage, la variance de l'équation (5.1) peut considérablement sous-estimer la variance de l'estimateur post-stratifié. Dans ce cas, l'estimateur de calage avec totaux de contrôle estimés est $\hat{Y}_{EC} = \widehat{\boldsymbol{X}}_{aux}{}'\widehat{\overline{\boldsymbol{Y}}}$, où $\widehat{\boldsymbol{X}}_{aux}$ est un estimateur du vecteur des totaux de contrôle de l'enquête auxiliaire, qui a une variance

$$V(\hat{Y}_{EC}) \approx \boldsymbol{X}'V\left(\widehat{\overline{\boldsymbol{Y}}}\right)\boldsymbol{X} + \overline{\boldsymbol{Y}}'V\left(\widehat{\boldsymbol{X}}_{aux}\right)\overline{\boldsymbol{Y}}. \tag{5.2}$$

Ici, $\overline{\boldsymbol{Y}}$ est le vecteur de population des moyennes de post-strate. Le deuxième terme de l'équation (5.2) peut avoir le même ordre de grandeur que le premier terme (et même être plus grand si l'enquête auxiliaire présente une grande variabilité).

Observons ce que Dever et Valliant (2010) ont fait dans l'équation (5.2). Les totaux de contrôle d'une enquête auxiliaire ne sont pas égaux aux valeurs de la population en raison de la variabilité de l'échantillonnage. Elles peuvent être trop élevées pour certaines post-strates et trop basses pour d'autres. Cela produit un biais pour $\hat{Y}_{EC}$, mais on ne connaît ni la direction ni la taille du biais, et la variance de l'équation (5.1), qui est conditionnelle aux totaux de contrôle, n'en rend pas compte. L'équation (5.2) transforme l'incertitude sur le biais en variance de façon à le traduire dans l'intervalle de confiance de l'estimateur.

Les variances des équations (5.1) et (5.2) supposent que les variables $\boldsymbol{x}$ des données administratives ou de l'enquête auxiliaire sont identiques aux variables $\boldsymbol{x}$ de l'enquête principale. Cela n'est pas nécessairement vrai. Si, par exemple, les catégories de race et d'origine ethnique sont utilisées dans la post-stratification, il se peut que l'enquête principale ne définisse pas ou ne mesure pas ces catégories de la même façon que les données administratives ou l'enquête

auxiliaire : les définitions peuvent différer, une source pourrait utiliser des catégories de race multiples alors que d'autres ne le font pas, ou encore les questions ou le contexte sont susceptibles d'entraîner des réponses différentes.

Si la réponse à l'enquête principale est complète et que les variables $\boldsymbol{x}$ sont cohérentes d'une source à l'autre, alors la variance de l'équation (5.2) coïncide habituellement avec l'erreur quadratique moyenne de l'estimateur, quel que soit le choix de variables auxiliaires et de modèle de calage.

Toutefois, en cas de non-réponse, les différents choix de modèle de calage peuvent donner des réponses différentes. La variance de l'estimation dans l'équation (5.2) est conditionnelle au choix du modèle et n'inclut pas le biais pouvant résulter de ce choix.

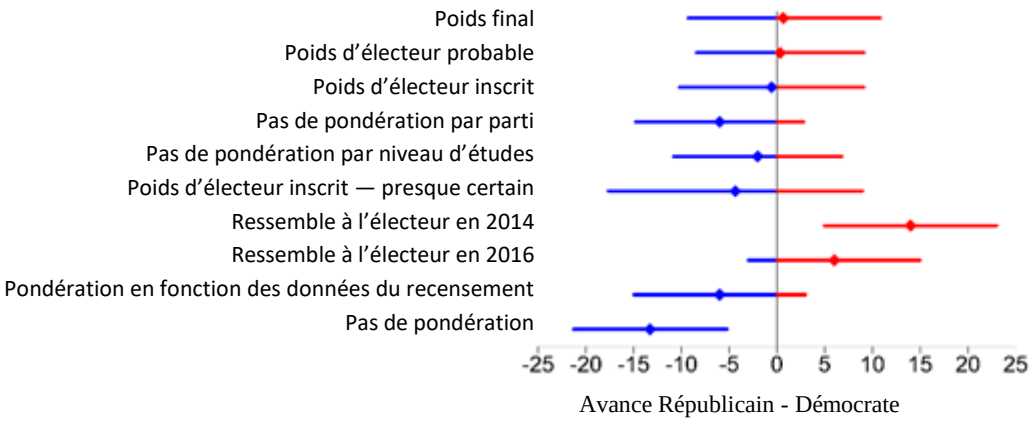
5.2 Sondages électoraux de mi-mandat aux États-Unis de 2018

Pour étudier la différence pouvant naître du choix du modèle de calage, observons les données de sondage publiées dans le *New York Times* (Andre et coll., 2018) à propos de l'élection de mi-mandat de 2018 dans la sixième circonscription de l'Illinois. Le sondage téléphonique a été mené du 4 au 6 septembre 2018; Cohn (2018) en a décrit la méthodologie. Pour cette circonscription , 36 455 appels ont été passés à de probables électeurs, 512 personnes ont répondu, ce qui donne un taux de réponse de 1,4 %.

Le républicain Peter Roskam devait être crédité de 45 % des intentions de vote et le démocrate Sean Casten de 44 %. Chacune de ces statistiques présentait une marge d'erreur de 4,7 points de pourcentage (les 11 % restants des personnes interrogées étaient indécises). Ainsi, en septembre 2018, le sondage montrait le candidat républicain en tête d'un point, soit à l'intérieur de la marge d'erreur de neuf points de pourcentage.

Or, le taux de réponse était faible et le sondage a été réalisé deux mois avant le scrutin. Des hypothèses fortes sous-tendent les variables utilisées dans la pondération, les personnes qui voteront selon les projections et ce qu'on prévoit que les personnes indécises feront.

Figure 5.2.-1
Estimations pour la sixième circonscription de l'Illinois selon différents modèles de pondération et de participation



Andre et coll. (2018) ont entrepris ce que font rarement les autres sondeurs : montrer quels seraient les résultats selon différents modèles de pondération et hypothèses de participation électorale et présenter d'autres poids sur l'ensemble de données. La figure 5.2-1 montre des intervalles de confiance de 95 % pour le candidat républicain en tête d'un point de pourcentage selon des pondérations et des modèles de participation électorale différents. La partie rouge de chaque ligne indique que le candidat républicain est en tête selon les prévisions, et la partie bleue indique que le candidat démocrate est en tête selon les prévisions.

Les estimations ponctuelles dépendent fortement du modèle de pondération utilisé, et la variabilité entre les estimations des différents modèles de pondération rivalise avec celle provenant de l'erreur d'échantillonnage.

6. Tenir compte de l'incertitude des modèles de pondération

6.1 Moyenne de modèles par une approche bayésienne

Les différences entre les estimations de la figure 5.2-1 donnent à penser que les intervalles de confiance fondés uniquement sur l'erreur d'échantillonnage sous-estiment l'incertitude de l'estimateur. Le calcul de la moyenne de plusieurs modèles par une approche bayésienne peut servir à inclure l'incertitude du modèle. Hoeting et coll. (1999) en donnent un aperçu; Lohr et Brick (2017) appliquent la méthode au sondage *Literary Digest* de 1936.

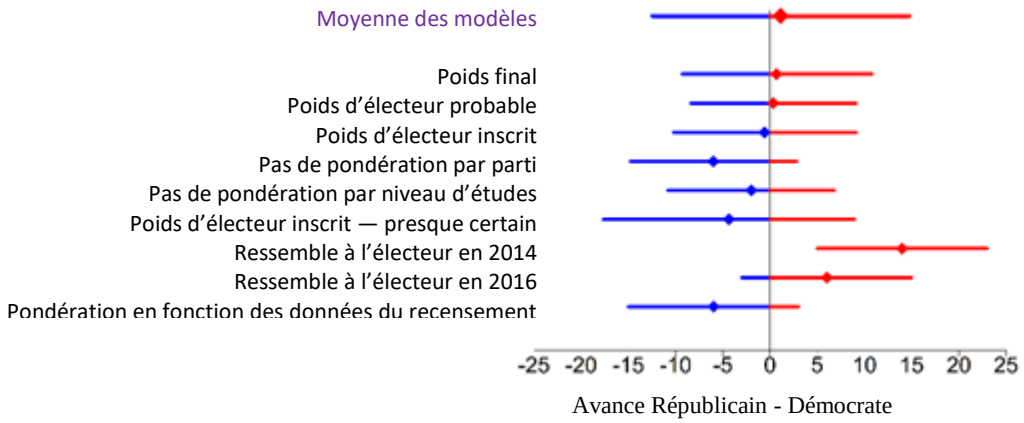
Soit une application avec les modèles de pondération $M_1, M_2, \ldots, M_K$ et les données D. La distribution à posteriori pour le modèle M_k, compte tenu des données, est $pr(M_k|D)$. La distribution à posteriori pour un paramètre θ est alors, compte tenu des données,

$$pr(\theta\,|D) = \sum_{k=1}^{K} pr(\theta\,|M_k,\ D)\ pr(M_k|\ D). \tag{6.1}$$

Ainsi, la moyenne à posteriori pour θ est la moyenne pondérée des estimations, pondérée par les probabilités à posteriori des modèles. La variance à posteriori comprend la variabilité d'échantillonnage de chaque estimation ainsi que la variabilité entre les estimations provenant de différents schémas de pondération.

La figure 6-1.1 montre un intervalle de prévision à posteriori fondé sur la moyenne bayésienne des modèles pour les données du sondage. L'information fournie par les données étant insuffisante, je n'ai pas pu obtenir de distributions à posteriori basées sur les données pour les modèles (en effet, il s'agit de modèles concernant les hypothèses des mécanismes de non-réponse et je ne dispose pas de données sur les non-répondants). C'est pourquoi j'ai choisi une distribution à priori pour les modèles fondée sur les descriptions subjectives d'Andre et coll. (2018). J'ai donné au modèle non pondéré une probabilité à priori de zéro puisqu'il semble qu'une certaine pondération est nécessaire.

Figure 6.1.-1
Estimations pour le sixième district du Congrès de l'Illinois, comprenant une estimation fondée sur la moyenne bayésienne de plusieurs modèles



L'intervalle fondé sur la moyenne des modèles de la figure 6-1.1 est plus grand que tous les intervalles des schémas de pondération individuels, ce qui exprime mon incertitude a priori au sujet des modèles de pondération. L'estimation ponctuelle est proche de celle du modèle avec les poids finaux, mais l'estimation de l'intervalle est plus grande, ce qui exprime l'incertitude supplémentaire sur la pondération. Tout comme la méthode de Dever et Valliant (2010), le

calcul de la moyenne des modèles bayésiens transforme l'incertitude sur le modèle de pondération d'un biais possible à une variance.

J'ai réalisé cette analyse deux mois avant les élections, en septembre 2018. Or, que s'est-il passé lors des élections de novembre? Selon *l'Illinois State Board of Elections* (2018), 146 445 personnes ont voté pour Roskam (républicain), et 169 001 personnes ont voté pour Casten. La différence (entre le candidat républicain et le candidat démocrate) était d'environ 7 points de pourcentage.

6.2 Défis

Bien entendu, les distributions à posteriori sont essentielles dans les estimations et elles dépendent des hypothèses. Dans d'autres applications de moyenne bayésienne de modèles, par exemple quand on examine différents modèles de régression, les données contiennent des renseignements auxquels les modèles correspondent bien. Toutefois, les modèles de non-réponse prévoient des données qui ne sont pas observées. J'ai utilisé des probabilités à priori subjectives pour les modèles de pondération, mais les probabilités à posteriori pouvaient être calculées pour les modèles si l'information était disponible dans la base de sondage ou une source externe. Une autre possibilité consisterait à inclure des renseignements provenant de données antérieures.

La théorie de l'inférence fondée sur le plan de sondage a été conçue afin d'éviter les problèmes de subjectivité dans les échantillons d'enquête. L'inférence fondée sur le plan évite les hypothèses de modèle, n'a pas besoin de loi à priori subjective, et a une théorie mathématique d'une magnifique élégance. Neyman (1934, p. 592) a soutenu que la « méthode représentative » (échantillonnage probabiliste) est supérieure, car « la construction de la bande de confiance est relativement indépendante de toute hypothèse arbitraire concernant les valeurs de θ. » La méthode représentative « rend superflu tout recours au théorème de Bayes » (Neyman, 1934, p. 562).

Toutefois, en cas de non-réponse, il *faut* faire des hypothèses de modèle. Dès qu'on adopte un modèle de non-réponse, on devient bayésien. On est cependant un bayésien qui adhère complètement au modèle de non-réponse qu'on a choisi pour la pondération. On a attribué une distribution à priori certaine basée sur un seil modèle de non-réponse.

Il faut encore prendre de nombreuses mesures et résoudre de nombreuses difficultés avant que ces méthodes soient prêtes à être utilisées dans des statistiques officielles. Il est facile d'obtenir une variabilité moindre en ne tenant compte que de modèles de pondération produisant des réponses semblables, et il est nécessaire d'établir des normes (comme cela a été le cas pour la déclaration des taux de réponse). Une des possibilités consisterait à enregistrer les modèles de pondération et les distributions à priori avant la collecte de données, comme pour l'enregistrement des protocoles d'essais cliniques et des analyses proposées.

Une approche basée sur la moyenne bayésienne de modèles présente un avantage : comme dans les modèles hiérarchiques, les hypothèses au sujet des modèles de non-réponse sont explicites. Il faut toujours faire des hypothèses au sujet des données manquantes, mais dans la plupart des enquêtes, ces hypothèses sont cachées dans les notes techniques ou, parfois, les détails des décisions relatives à la pondération ne sont pas publiés. En cas de modèle bayésien, les hypothèses sont ouvertement établies et chacun est libre de les évaluer.

7. Discussion

7.1 Examiner les propriétés des erreurs au moyen de sources de données multiples

Les estimations d'enquête illustrées dans les figures 1.2-1 et 1.3-1 diffèrent pour plusieurs raisons. Pour ce qui est de la NISVS et de la NCVS, la plupart des différences constatées dans les estimations sont probablement attribuables aux questions d'enquête. Dans la NCVS, on demande explicitement si le répondant a été violé ou agressé sexuellement alors que la NISVS pose des questions sur le comportement et les événements qui se sont produits. Lohr (2019) discute d'autres facteurs possibles de la différence entre la NISVS et la NCVS, et Siegfried et coll. (2017) examinent des raisons expliquant les différences entre les estimations sur le tabagisme.

L'adoption de modèles de pondération différents pour la non-réponse peut aussi expliquer certaines des différences dans les estimations d'enquête. La NCVS et la NISVS utilisent toutes deux une pondération, mais les modèles de pondération finaux ne sont pas décrits précisément. De plus, le taux de réponse de la NISVS est inférieur de 40 à 50 points de pourcentage environ au taux de réponse de la NCVS. Il se peut que les estimations de la NISVS soient plus sensibles à la repondération que les estimations de la NCVS parce que la non-réponse est plus importante dans la NISVS.

Il serait possible d'évaluer les enquêtes quand le FBI aura terminé sa transition vers le *National Incident-Based Reporting System*, qui recueille des données sur les caractéristiques des victimes et des délinquants pour chaque infraction. À l'heure actuelle, seuls 40 % des organismes participent au système, mais l'information détaillée pourrait servir à étudier le biais de non-réponse possible dans la NCVS (ou les erreurs possibles de sous-déclaration dans les statistiques sur l'application de la loi).

De même, pour ce qui est des différentes estimations du tabagisme au cours de la vie, on peut utiliser les mêmes sources pour explorer les propriétés des erreurs dans les autres sources. Par exemple, le CPS permet des entretiens par procuration. Cela explique-t-il en partie pourquoi ses estimations sont inférieures à celles des autres enquêtes? Willis et coll. (2017, p. 3) écrivent que les estimations plus basses de la TUS-CPS s'expliquent « probablement par diverses raisons, notamment des différences entre les enquêtes en ce qui concerne les caractéristiques techniques et l'administration des enquêtes ».

7.2 Le problème zéro

Colin Mallows (1998) juge que la participation des statisticiens intervient souvent trop tard dans les problèmes de recherche. Ils doivent être impliqués dans ce qu'il a appelé le problème zéro, le problème consistant à déterminer les sources de données pertinentes selon le problème. Selon lui : « Souvent, les arguments statistiques peinent à convaincre, car le fondement de leurs hypothèses n'est pas énoncé précisément. »

Quand on utilise de multiples sources à des fins d'inférence, le problème zéro consiste à évaluer la qualité et les propriétés des sources de données individuelles et à déterminer celles pertinentes dans la réponse aux questions de recherche.

Ce n'est qu'après cette évaluation qu'on peut combiner les données. Toutes les méthodes d'intégration reposent sur des modèles, et les mesures de l'incertitude des sources de données combinées dépendent des mesures des sources de données individuelles, qui sont susceptibles d'être sous-estimées. Ces variances individuelles sous-estimées sont héritées par l'estimation combinée.

Le fait de disposer de sources diverses, particulièrement si les sources ont différents types d'erreurs systématiques, peut aider à comprendre cette sous-estimation et à étudier les propriétés des erreurs de chaque source. Elles peuvent servir à étudier la qualité selon une approche systémique. La mesure de l'incertitude dans les études individuelles et dans les données combinées est un problème systémique. C'est pourquoi il faut une solution systémique, qui comprenne les erreurs de mesure et de non-réponse ainsi que la variabilité résultant des différents modèles de pondération.

Bibliographie

Andre, M., Buchanan, L., Bloch, M., Bowers, J., Cohn, N., Coote, A., Daniel, A., Harris, R., Katz, J., Rebecca Lieberman, R., Migliozzi, B., Murray, P., Pearce, A., Quealy, K., Weingart, E., and White, I. (2018), "We Polled Voters in Illinois's 6th Congressional District", *The New York Times*, https://www.nytimes.com/interactive/2018/upshot/elections-poll-il06-1.html, accessed September 10, 2018.

Bureau of Justice Statistics (2018a), "Rates of Rape/Sexual Assaults by Sex and Reporting to the Police, 1993-2016", Generated using the NCVS Victimization Analysis Tool at www.bjs.gov on September 11, 2018.

Bureau of Justice Statistics (2018b), "Standard Errors for Rates of Rape/Sexual Assaults by Sex and Reporting to the Police, 1993-2016", Generated using the NCVS Victimization Analysis Tool at www.bjs.gov on September 11, 2018.

Cohn, N. (2018), "Our Polling Methodology", *The New York Times* (September 6), https://www.nytimes.com/2018/09/06/upshot/live-poll-method.html, accessed September 10, 2018.

Dever, J., and Valliant, R. (2010), "A Comparison of Variance Estimators for Poststratification to Estimated Control Totals", *Survey Methodology*, 36, pp. 45-56.

Dever, J., and Valliant, R. (2016), "General Regression Estimation Adjusted for Undercoverage and Estimated Control Totals", *Journal of Survey Statistics and Methodology*, 4, pp. 289-318.

Federal Bureau of Investigation (2016), *Crime in the United States, 2015*. Washington, D.C.: Federal Bureau of Investigation

Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. (1999), "Bayesian Model Averaging: A Tutorial", *Statistical Science,* 14, pp. 382-401.

Illinois State Board of Elections (2018), *Official Canvass, General Election, November 6, 2018,* Springfield, IL: Illinois State Board of Elections.

Lin, D. (2013), *Measurement Error in Dual Frame Estimation,* Ph.D. Dissertation, Southern Methodist University.

Lohr, S. L. (2011), "Alternative Survey Sample Designs: Sampling with Multiple Overlapping Frames", *Survey Methodology,* 37, pp. 197-213.

Lohr, S. L. (2019), *Measuring Crime: Behind the Statistics, Boca Raton,* FL: CRC Press.

Lohr, S. L., and Brick, J. M. (2017), "Roosevelt Predicted to Win: Revisiting the 1936 Literary Digest Poll", *Statistics, Politics and Policy,* 8, pp. 65-84.

Lohr, S. L., and Raghunathan, T. E. (2017), "Combining Survey Data with Other Data Sources", Statistical Science, 32, pp. 293-312.

Mallows, C. (1998), "The Zeroth Problem", *The American Statistician,* 52, pp. 1-9.

Manzi, G., Spiegelhalter, D. J., Turner, R. M., Flowers, J., and Thompson, S. G. (2011), "Modelling Bias in Combining Small Area Prevalence Estimates from Multiple Surveys", *Journal of the Royal Statistical Society: Series A,* 174, pp. 31-50.

National Academies of Sciences, Engineering, and Medicine (2017). *Federal Statistics, Multiple Data Sources, and Privacy Protection: Next Steps,* Washington, D.C.: The National Academies Press.

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, pp. 558-625.

Puzzanchera, C., Smith, J., and Kang, W. (2017), "Easy Access to NIBRS Victims, 2015: Victims of Violence", https://www.ojjdp.gov/ojstatbb/ezanibrsdv/, accessed September 11, 2018.

Siegfried, Y., Morganstein, D., Piesse, A., and Lohr, S. (2017), "Why Independent Surveys with the Same Objective Yield Different Estimates", *Proceedings of the Survey Research Methods Section, American Statistical Association,* pp. 935-952.

Smith, S. G., Zhang, X., Basile, K. C., Merrick, M. T., Wang, J., Kresnow, M.-J., and Chen, J. (2018). *The National Intimate Partner and Sexual Violence Survey (NISVS): 2015 Data Brief.* Atlanta, GA: National Center for Injury Prevention and Control, Centers for Disease Control and Prevention.

Willis, G., Hartman, A., Reyes-Guzman, C., Seaman, E .L., Gibson, J. T., Goettsche, E., Chomenko, D., Mangold, K., and Block, M. (2017), *The 2014-2015 Tobacco Use Supplement to the Current Population Survey,* Rockville, MD: National Cancer Institute.