No 11-522-XIF au catalogue

La série des symposiums internationaux de Statistique Canada - Recueil

Symposium 2005 : Défis méthodologiques reliés aux besoins futurs d'information

2005

Statistique Canada Statistics Canada

Canada

Recueil du symposium de 2005 de Statistique Canada
Défis méthodologiques reliés aux besoins futurs d'information

EFFET DE L'INFORMATIVITÉ DU PLAN DE SONDAGE SUR L'INFÉRENCE STATISTIQUE

David A. Binder, Milorad S. Kovacevic et Georgia Roberts[1]

RÉSUMÉ

Les plans typiques de sondage complexe donnent lieu à des échantillons informatifs, c'est-à-dire que la distribution d'une variable dans l'échantillon est différente de sa distribution dans la population. Afin de déterminer et de mesurer l'effet de l'informativité, on comparera les variances des paramètres estimés (de même que les paramètres estimés) fondées sur le plan à celles fondées sur le modèle, dans un modèle logistique d'après l'hypothèse que le modèle formulé est vrai. Pour évaluer l'adéquation de la modélisation des données provenant d'échantillons informatifs, nous considérerons deux façons de faire : utiliser une inférence fondée sur le plan pour les paramètres du modèle ou utiliser une inférence fondée sur le modèle. Nous proposons une nouvelle approche bivariée pour évaluer l'effet de l'informativité du plan de sondage qui tient compte des effets sur les estimations et sur leur écart-type. Une étude par simulation d'envergure, basée sur la génération d'une population sous un modèle postulé, utilisant des paramètres estimés dérivés de l'Enquête nationale sur la santé de la population (ENSP), nous permet de détecter la présence d'informativité, de la mesurer, et de comparer la robustesse des deux approches retenues.

MOTS CLÉS : selon le plan, selon le modèle, selon le plan-modèle, mise en grappes informative, puissance de test.

1. INTRODUCTION

L'informativité d'un échantillon est un concept de modélisation. Si la distribution de l'unité échantillonnée est différente de celle que l'on obtiendrait en faisant l'échantillonnage directement à partir du modèle, alors l'échantillonnage est dit **informatif**. Le plan est dit **ignorable** pour une analyse particulière s'il a la propriété que les résultats de l'analyse ne sont pas affectés par l'informativité du plan de sondage. Tous les plans de sondage non informatifs mènent à l'ignorabilité, mais l'inverse n'est pas vrai (Binder et Roberts, 2001).

Certains analystes ajustent le même modèle à des données d'enquête en utilisant une approche fondée sur le plan de sondage ainsi qu'une approche fondée sur le modèle et, si les estimations ponctuelles des coefficients du modèle sont les mêmes sous les deux approches, concluent que l'échantillonnage n'était pas informatif et poursuivent selon une approche fondée sur le modèle. Cependant, les estimations ponctuelles fondées sur le plan et celles fondées sur le modèle peuvent être semblables, même si les hypothèses au sujet de la loi du modèle sont incorrectes pour l'échantillon. Donc, lorsque les estimations ponctuelles sont semblables, le fait que les **estimations des variances par rapport au plan de sondage** ne soient pas proches des **estimations des variances par rapport au modèle** pourrait être une indication que l'échantillonnage est « informatif », particulièrement si l'échantillon est de grande taille. Si l'approche privilégiée dans ces conditions est encore celle fondée sur le modèle, ce dernier devrait être modifié afin de s'assurer que la distribution d'échantillon des unités échantillonnées soit valide sous le modèle. (Cette condition pourrait toutefois être difficile à satisfaire.)

L'objectif de cet exposé est de chercher des moyens de déterminer si le plan de sondage a une incidence sur les conclusions de fond tirées d'une analyse. À cet égard, nous comparons la méthode fondée sur le plan de sondage standard d'analyse à certaines autres méthodes fréquemment utilisées, c'est-à-dire les méthodes fondées sur un modèle standard, sur un modèle utilisant des poids normalisés et sur des modèles mixtes avec effets aléatoires pour la mise en grappes. Nous procédons à une étude par simulation dans laquelle nous générons une population finie satisfaisant entièrement notre modèle hypothétique. Par stratification et mise en grappes des résultats, nous tenons

[1] David A. Binder, Statistique Canada, 17J Immeuble R.-H.-Coats, Pré Tunney, Ottawa, ON, K1A0T6, dbinder49@hotmail.com, Milorad S. Kovacevic, Statistique Canada, 17J Immeuble R.-H.-Coats, Pré Tunney, Ottawa, ON, K1A0T6, kovamil@statcan.ca, Georgia Roberts, Statistique Canada, 17J Immeuble R.-H.-Coats, Pré Tunney, Ottawa, ON, K1A0T6, robertg@statcan.ca.

compte de l'effet de l'informativité des échantillons tirés à partir de la population finie. Puis, nous évaluons et comparons l'effet de l'informativité du plan d'échantillonnage sur les diverses méthodes.

À la section 2, nous décrivons l'étude par simulation. Nous présentons à la section 3 les mesures utilisées pour l'évaluation et les résultats obtenus pour l'évaluation de l'effet de l'informativité sur les estimations ponctuelles et les estimations de la variance. Nous donnons les détails d'une étude de l'effet de l'informativité sur la puissance et la grandeur des tests à la section 4. À la section 5, nous proposons une approche pour évaluer l'informativité des données d'échantillon. Enfin, à la section 6, nous présentons certaines conclusions.

2. ÉTUDE PAR SIMULATION

2.1 (Modèle de) superpopulation

Afin d'obtenir une évaluation empirique de l'effet de l'informativité du plan d'échantillonnage sur l'analyse des données d'enquête, nous avons réalisé une grande étude par simulation. Nous avons simulé un modèle de la relation entre ***la perte d'autonomie chez les personnes âgées*** (***LOSS***) et plusieurs facteurs associés à leur état de santé et à leurs habitudes, motivé par un modèle ajusté aux données provenant des deux premiers cycles (1994-1995 et 1996-1997) de l'Enquête nationale sur la santé de la population (ENSP) au Canada et présenté dans Martel, Bélanger et Berthelot (2002).

Le modèle que nous avons simulé exprime la probabilité qu'une personne âgée autonome perde son autonomie en fonction du sexe, de l'âge, de l'indice de masse corporelle, de la présence de maladies chroniques et des habitudes d'usage du tabac. Le modèle est de la forme :

$$\begin{aligned}\text{logit}(\boldsymbol{LOSS}) = {} & \beta_0 + \beta_1 * \boldsymbol{SEX} + \beta_2 * \boldsymbol{AGEGR} + \beta_3 * \boldsymbol{UNDERWGT} \\ & + \beta_4 * \boldsymbol{OVERWGT} + \beta_5 * \boldsymbol{CHRDS} + \beta_6 * \boldsymbol{SMOK}\end{aligned} \tag{2.1}$$

Toutes les variables du modèle sont binaires : *LOSS* (1, si la personne perd son autonomie au cours de la période de deux ans étudiée, 0 autrement), *SEX* (0 pour les femmes, 1 pour les hommes), *AGEGR* (0 pour l'âge compris dans [65,75), 1 pour l'âge égal à 75+), *UNDERWGT* (1 pour *BMI* (indice de masse corporelle) ≤18,5, 0 autrement), *OVERWGT* (1 pour BMI ≥25, 0 autrement), *CHRDS* (1 si au moins l'un des dix problèmes de santé chroniques est présent, 0 autrement), *SMOK* (1 si la personne fume quotidiennement ou si elle a arrêté récemment, 0 autrement). À part *LOSS*, toutes les variables sont mesurées au début de la période de deux ans.

Notons que la valeur de référence est 0 pour toutes les variables incluses dans le modèle logistique (2.1). Les variables reliées à l'indice de masse corporelle *BMI* ont pour origine une variable *BMIGR* comprenant trois catégories (0 pour *BMI* ≤18,5, 1 pour 18,5 <*BMI* <25, et 2 pour *BMI* ≥25). Les dix problèmes de santé chroniques considérés sont l'asthme, l'arthrite, les maux de dos, la bronchite ou l'emphysème, le diabète, la maladie cardiaque, le cancer, les effets d'un accident vasculaire cérébral, l'incontinence urinaire et le glaucome ou la cataracte.

2.2 Population finie simulée

Nous avons simulé une population finie de 2,5 millions de personnes de telle sorte que ces dernières présentent certaines caractéristiques de la sous-population de personnes âgées de 65 ans et plus qui étaient autonomes au moment du premier cycle de l'ENSP réalisée au Canada. Les variables ont été générées sous forme de variables aléatoires de Bernoulli en utilisant les probabilités conjointes estimées d'après l'échantillon de l'ENSP au premier cycle.

Après avoir simulé les valeurs de *SEX*, *AGEGR*, etc., nous avons également créé la variable dépendante *LOSS* sous forme de variable de Bernoulli avec probabilité égale à

$$\boldsymbol{p_x} = \boldsymbol{p}(\boldsymbol{LOSS} = 1 \mid \boldsymbol{x}) = \left[1 + \exp(-\boldsymbol{x}'\hat{\theta})\right]^{-1},$$

où $\boldsymbol{x}$ et $\theta = (\beta_1, \beta_2, ..., \beta_7)$ sont définis par le modèle (2.1) et où θ a été estimé d'après l'échantillon de l'ENSP comme étant $\hat{\theta} = (\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \hat{\beta}_4, \hat{\beta}_5, \hat{\beta}_6)$ =(-3,799, 0,382, 1,388, 1,139, 0,406, 0,641, 0,484). La proportion de personnes faisant partie de la population finie simulée qui ont perdu leur autonomie est de 0,1009.

2.3 Mise en grappes et stratification simulées de la population finie

Nous avons mis la population finie en grappes de deux façons : l'une est une mise en grappes purement aléatoire (non informative) et l'autre, est, dans une certaine mesure, une « mise en grappes informative ». Dans les deux cas, les tailles des grappes variaient de 20 à 60 enregistrements individuels.

Dans le cas de la mise en grappes aléatoire, nous avons ordonné les enregistrements individuels aléatoirement, puis les avons affectés à des grappes dont les tailles ont été générées sous forme de nombres entiers répartis uniformément entre 20 et 60. Par exemple, si le premier nombre aléatoire était 29, les 29 premiers enregistrements individuels ont été affectés à la première grappe; si le deuxième nombre aléatoire était 43, les 43 enregistrements individuels suivants ont formé la deuxième grappe, et ainsi de suite. De cette façon, nous avons créé 62 600 grappes et avons calculé que la corrélation intragrappe était de 0,0001.

La mise en grappes informative a été réalisée de la façon suivante. À partir des premiers 1,875 millions d'enregistrements individuels, nous avons créé 62 600 grappes selon la même procédure de mise en grappes aléatoire que celle susmentionnée, excepté que les tailles des grappes ont été générées sous forme de nombres entiers aléatoires compris entre 10 et 50. À partir des 625 000 enregistrements restants, nous avons créé 62 500 groupes de 10 enregistrements chacun, de telle façon qu'environ 6 260 groupes contiennent uniquement des enregistrements pour lesquels LOSS = 1 et que les autres contiennent uniquement des enregistrements pour lesquels LOSS = 0. Alors, pour chacune des quelque 6 260 grappes ayant la proportion la plus grande d'enregistrements avec LOSS = 1, nous avons ajouté un groupe contenant 10 enregistrements « LOSS = 1 ». Toutes les autres grappes ont reçu un groupe de 10 enregistrements avec LOSS = 0. De cette façon, nous avons créé 62 600 grappes de tailles variant de 20 à 60, pour lesquelles la corrélation intragrappe était de 0,2637.

Pour chacune des mises en grappes de la population finie, nous avons stratifié les grappes de deux façons : (i) aucune stratification et (ii) deux strates, où la première strate contenait les 25 % de grappes ayant la proportion la plus forte d'enregistrements avec « LOSS = 1 » et la seconde, les 75 % de grappes restantes.

2.4 Plans d'échantillonnage et génération d'information sur le plan

Nous avons utilisé un plan d'échantillonnage différent selon que la population était ou non stratifiée. Dans les deux cas, le plan d'échantillonnage consistait en un échantillon de grappes sélectionnées sans remise avec probabilité de sélection proportionnelle à la taille de la grappe. La sélection a été faite par la méthode de Sampford, telle qu'elle est implémentée dans la procédure SURVEYSELECT de SAS. Dans le cas de la population non stratifiée, nous avons tiré un échantillon de 30 grappes. Dans le cas de la population stratifiée, nous avons sélectionné 15 grappes à partir de chacune des deux strates, ce qui signifie que nous avons clairement suréchantillonné la première strate, donc donné une plus grande probabilité de sélection aux grappes ayant une plus grande p_i (c.-à-d., une plus grande proportion d'enregistrements avec LOSS = 1).

Les poids d'échantillonnage originaux des enregistrements inclus dans un échantillon seront constants dans les grappes, puisqu'il n'y a pas eu de sous-échantillonnage dans les grappes. Nous avons ensuite stratifié à posteriori ces poids originaux en cinq stratifications à posteriori en nous fondant sur les dénombrements connus de (*AGEGR* X *SEX*) et d'*URBRUR*. Après la stratification à posteriori, les poids des enregistrements contenus dans une même grappe n'étaient plus nécessairement tous égaux.

Pour chaque échantillon de grappes, nous avons produit 500 rééchantillonnages bootstrap. Dans le cas du plan non stratifié, pour chaque rééchantillonnage bootstrap, nous avons tiré un échantillon aléatoire simple avec remise de taille n-1(= 29) grappes. Dans le cas du plan stratifié, nous avons sélectionné un échantillon aléatoire avec remise de taille n_h (= 14) pour la h^e strate, h=1,2. Nous avons alors obtenu les poids bootstrap dans le b^e *rééchantillonnage* bootstrap en commençant par ajuster les poids d'échantillonnage originaux afin de révéler l'inclusion (éventuellement plus d'une fois) de certains grappes et l'exclusion d'autres grappes selon la formule pour le b^e rééchantillonnage bootstrap :

$$w_{hij}^{(b)} = w_{hij}\, k_{hi}^{(b)} \frac{n_h}{n_h - 1},$$

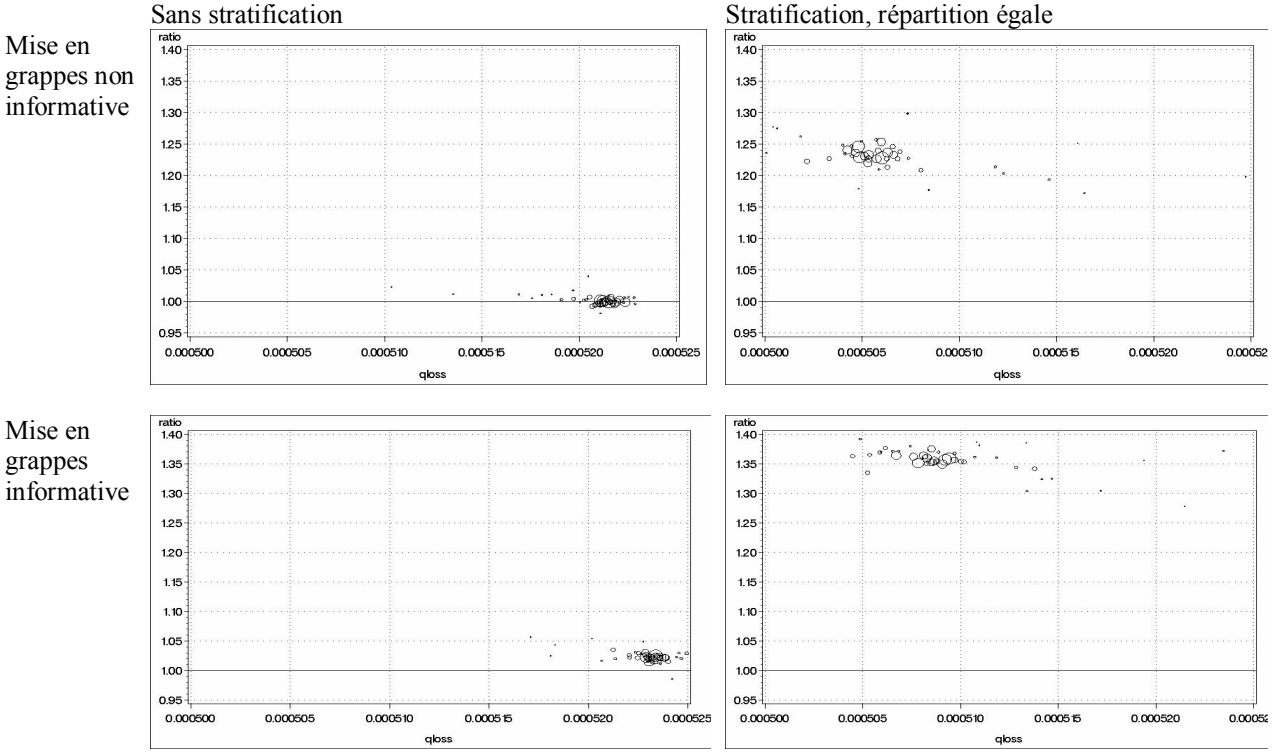
où w_{hij} est le poids d'échantillonnage original de la j^e personne provenant de la i^e grappe dans la h^e strate, $h = 1,2$, et $k_{hi}^{(b)}$ est le nombre de répétitions de la i^e grappe dans le b^e rééchantillonnage bootstrap. Soulignons que $\sum_i k_{hi}^{(b)} = n_h - 1$. Nous avons ensuite stratifié à posteriori ces poids par rapport aux dénombrements connus des stratifications à posteriori, de la même façon qu'avaient été calibrés les poids originaux pour l'échantillon complet.

2.5 Configuration de la simulation de Monte Carlo

Compte tenu des deux types de mise en grappes (non informative et informative) et des deux options de stratification (sans et avec), nous avions quatre **définitions de population** ayant divers niveaux d'informativité. Nous avons tiré 500 échantillons de Monte Carlo à partir de chacune des quatre populations, c'est-à-dire i) mise en grappes « non informative », pas de stratification, 30 grappes (**faible inf.**); ii) mise en grappes « informative », pas de stratification, 30 grappes; iii) mise en grappes « non informative », stratification, 15 grappes par strate; et iv) mise en grappes « informative », stratification, 15 grappes par strate (**forte inf.**).

Afin de vérifier le degré d'informativité atteint, nous avons calculé les probabilités conditionnelles d'inclusion, sachant x, de la variable LOSS, c'est-à-dire, $\boldsymbol{ploss}(\boldsymbol{x}) = \Pr\{\boldsymbol{i} \in \boldsymbol{sample} | \boldsymbol{LOSS_i} = 1, \boldsymbol{x}\}$ et $\boldsymbol{qloss}(\boldsymbol{x}) = \Pr\{\boldsymbol{i} \in \boldsymbol{sample} | \boldsymbol{LOSS_i} = 0, \boldsymbol{x}\}$. Si le plan d'échantillonnage est non informatif, ces probabilités ne dépendront pas de LOSS, et, pour le même $\boldsymbol{x}$, le ratio $\boldsymbol{ploss}(\boldsymbol{x}) / \boldsymbol{qloss}(\boldsymbol{x}) \approx 1$. Notons qu'il existe seulement 48 valeurs différentes pour le vecteur $\boldsymbol{x}$. Au graphique 1, nous présentons, pour chacune des quatre populations définies, les ratios sur l'axe vertical et les probabilités d'inclusion des enregistrements pour lesquels LOSS = 0 sur l'axe horizontal. Les diamètres des cercles représentent la taille du groupe ***x dans*** la population. L'examen du graphique 1 montre que la mise en grappes informative conjuguée à la stratification augmente la valeur du ratio, qui passe à 1,35 (en moyenne). Autrement dit, une personne pour laquelle LOSS = 1 a 1,35 fois plus de chance d'être incluse dans l'échantillon qu'une personne ayant les mêmes caractéristiques **x** pour laquelle LOSS = 0.

Graphique 1 : Ratio des probabilités conditionnelles d'inclusion pour les personnes pour lesquelles LOSS = 0 et LOSS=1, sachant $\boldsymbol{x}$



Dans la suite de l'article, nous présentons les résultats pour les deux configurations extrêmes, c'est-à-dire faible informativité (faible inf.) et forte informativité (forte inf.).

2.6 Approches inférencielles pour le modèle logistique

Nous considérons les cinq approches inférencielles qui suivent pour ajuster le modèle logistique aux données provenant des échantillons sélectionnés :

i) [PLAN] Approche entièrement fondée sur le plan de sondage où les variances sont estimées par la méthode du bootstrap avec équation d'estimation linéarisée (voir Binder, Kovacevic, Roberts, 2004). Les estimations résultantes des paramètres et les estimations de leur variance sont indiquées par $\hat{\theta}_{\boldsymbol{p}}$ et $\hat{\boldsymbol{V}}_{\boldsymbol{p}}(\hat{\theta}_{\boldsymbol{p}})$. La programmation a été faite en SAS en utilisant la procédure SAS IML.

ii) [PLAN-MODÈLE] Combinaison d'une estimation ponctuelle pondérée des paramètres, $\hat{\theta}_{\boldsymbol{p}}$, et d'une estimation fondée sur un modèle des variances en utilisant l'estimateur intercalé robuste de la variance non corrigé pour la mise en grappes, $\hat{\boldsymbol{V}}_{\xi}(\hat{\theta}_{\boldsymbol{p}})$. Cette approche a été mise en œuvre en utilisant la procédure LOGISTIC dans SUDAAN et en fixant DESIGN = SRS, SEMETHOD = model et en spécifiant la variable de poids de sondage dans un énoncé WEIGHT.

iii) [MODÈLE] Approche fondée sur un modèle (non pondéré) où les estimations ponctuelles $\hat{\theta}_{\xi}$ et leurs variances $\hat{\boldsymbol{V}}_{\xi}(\hat{\theta}_{\xi})$ sont les unes et les autres calculées comme si l'on suivait un plan d'échantillonnage aléatoire simple d'unités avec remise. Cette approche a été appliquée en utilisant la procédure LOGISTIC dans SUDAAN et en spécifiant DESIGN=WR, SEMETHOD=model et sans utiliser l'énoncé WEIGHT.

iv) [MIXTENL] Approche fondée sur un modèle où les effets de la mise en grappes $\boldsymbol{u}_{\boldsymbol{i}}$ sont modélisés sous forme d'effets aléatoires additifs :

$$\xi_1 : \quad \boldsymbol{logit}(\boldsymbol{y}_{\boldsymbol{ij}}) = \boldsymbol{x}'_{\boldsymbol{ij}}\ \theta + \boldsymbol{u}_{\boldsymbol{i}} + \varepsilon_{\boldsymbol{ij}}\,.$$

Les paramètres du modèle θ sont estimés au moyen d'un estimateur fondé sur un modèle (non pondéré) $\hat{\theta}_{\xi 1}$ et la matrice de variance correspondante est estimée à l'aide de l'estimateur de la variance par rapport au modèle $\hat{\boldsymbol{V}}_{\xi 1}(\hat{\theta}_{\xi 1})$. Nous avons utilisé la procédure SAS PROC NLMIXED pour l'estimation. Il convient de souligner que le modèle ne tenait pas compte de la stratification dans le plan de sondage.

v) [MIXTENL-POND] Ajustement du modèle ξ_1 comme en (iv), mais en utilisant les poids de sondage, obtention des estimations ponctuelles $\hat{\theta}_{\boldsymbol{w}}$ et de leur matrice de variance par rapport au modèle en utilisant l'estimateur « intercalé » évalué aux valeurs estimées finales, $\hat{\boldsymbol{V}}_{\xi 1}(\hat{\theta}_{\boldsymbol{w}})$. Notons qu'en (iv) et (v), nous ne cherchons pas à estimer les composantes de la variance. L'estimation a été faite en utilisant la procédure SAS PROC NLMIXED avec la variable de pondération spécifiée dans l'énoncé REPLICATE. (Puisque la variable spécifiée dans l'énoncé REPLICATE doit avoir des valeurs entières et que les poids de sondage contenaient quatre chiffres après la virgule, la variable spécifiée dans l'énoncé REPLICATE était, en fait, 10 000*poids.)

3. EFFET DE L'INFORMATIVITÉ SUR LES ESTIMATIONS PONCTUELLES ET LES ESTIMATIONS DE LA VARIANCE

Nous considérons une gamme de mesures afin de comparer les diverses approches. Dans la description des mesures qui suit, l'indice inférieur *M* indique n'importe quelle approche et θ représente n'importe quel coefficient de notre modèle.

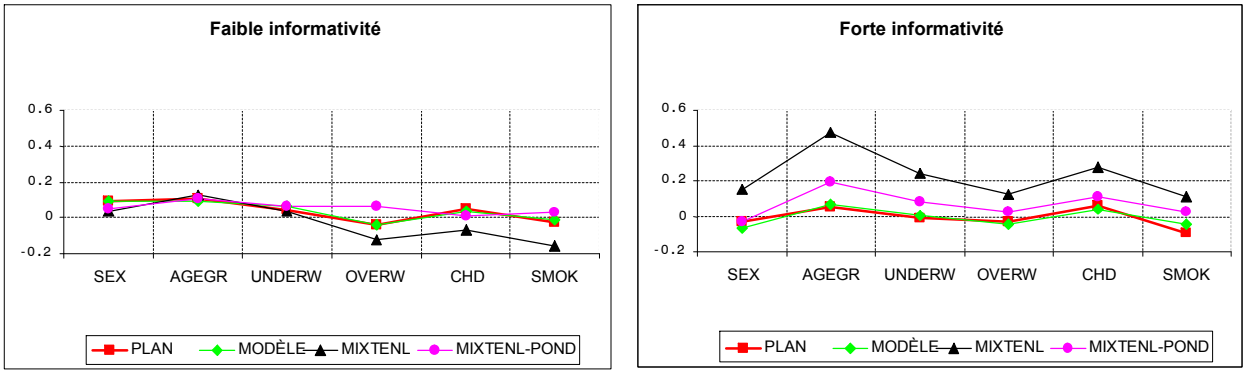
La première mesure examinée, qui a trait à l'effet de l'informativité sur les estimations ponctuelles, est **l'écart standardisé de l'estimation d'un paramètre par rapport à la valeur réelle :**

$$\frac{E_{sim}(\hat{\theta}_M)-\theta}{\sqrt{MSE_{sim}(\hat{\theta}_M)}}, \text{ où} \qquad (3.1)$$

$E_{sim}(\hat{\theta}_M)=\frac{1}{500}\sum_k \hat{\theta}_{M,k}$, $MSE_{sim}(\hat{\theta}_M)=\frac{1}{500}\sum_k(\hat{\theta}_{M,k}-\theta)^2$ et $\hat{\theta}_{M,k}$ est l'estimation de θ en utilisant l'approche M^e avec l'échantillon k^e de Monte Carlo. L'estimation du paramètre est d'autant meilleure que cette mesure s'approche de 0. Nous présentons nos résultats au graphique 2 pour les six coefficients du modèle, sauf l'ordonnée à l'origine. Le graphique de gauche montre les résultats pour le cas de faible informativité, et celui de droite, pour celui de forte informativité. Les variables du modèle sont portées sur l'axe horizontal et la grandeur de la mesure est représentée sur l'axe vertical. Les valeurs de la mesure pour les diverses variables pour une approche particulière sont reliées par des segments de droite d'une couleur particulière.

Dans les conditions de faible informativité, toutes les approches donnent des résultats semblables, la mesure demeurant proche de zéro pour toutes les variables, avec une légère exception pour MIXTENL. Sous les conditions de forte informativité, les tracés pour les différentes approches sont plus dispersés; cependant, PLAN et MODÈLE produisent des résultats fort semblables qui sont ceux qui s'approchent le plus de la ligne zéro. L'approche PLAN-MODÈLE donne exactement les mêmes résultats que l'approche PLAN, et n'est donc pas représentée sur ces graphiques. Dans une situation d'enquête réelle, puisque nous ne connaissons pas la vraie valeur de θ, nous ne pouvons calculer cette statistique.

Graphique 2 : Écart standardisé de l'estimation d'un paramètre par rapport à la valeur réelle

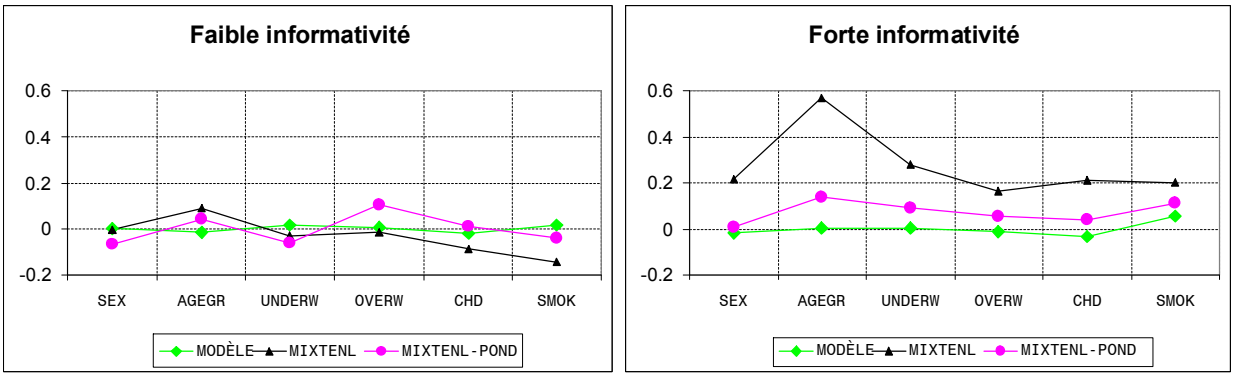


Une mesure que l'on peut calculer d'après des données d'enquête est **l'écart normalisé entre une estimation de rechange du paramètre $\hat{\theta}_M$ et celle fondée sur le plan $\hat{\theta}_p$.** Cette mesure, qui est une moyenne de l'écart standardisé sur les 500 échantillons de Monte Carlo, est définie de la façon suivante :

$$E_{sim}\left\{\frac{\hat{\theta}_M-\hat{\theta}_p}{\sqrt{\hat{V}_p(\hat{\theta}_p)}}\right\}=\frac{1}{500}\sum_k\frac{\hat{\theta}_{M,k}-\hat{\theta}_{p,k}}{\hat{V}_p(\hat{\theta}_{p,k})}. \qquad (3.2)$$

Les résultats sont présentés au graphique 3.

Graphique 3 : Écart standardisé entre une estimation $\hat{\theta}_M$ et l'estimation fondée sur le plan $\hat{\theta}_p$
(moyenne sur 500 échantillons)



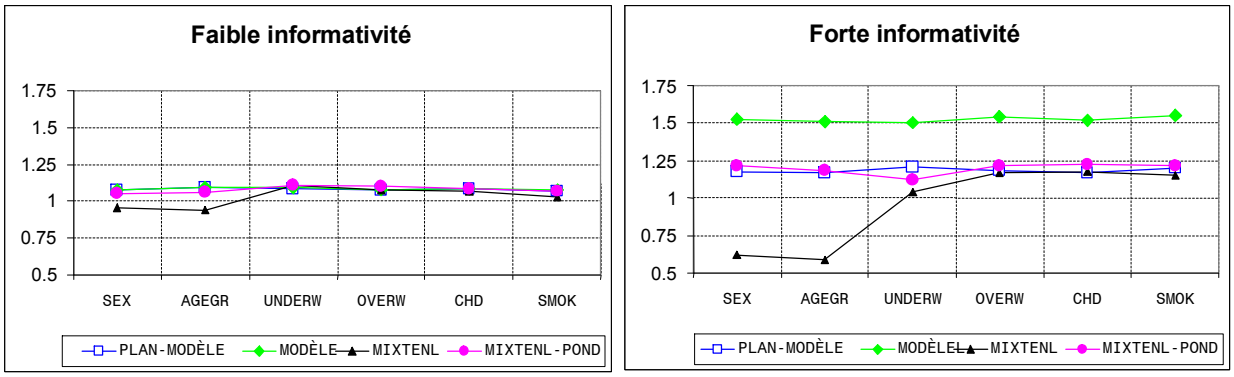
Dans le cas de faible informativité, toutes les approches produisent des courbes qui s'approchent de 0, celle de l'approche MODÈLE demeurant la plus proche. Dans le cas de forte informativité, la dispersion des tracés est plus grande, quoique la courbe de l'approche fondée sur un modèle soit encore proche de zéro. La courbe MIXTENL est celle qui exprime le comportement le plus extrême. Notons que, de nouveau, les graphiques ne contiennent pas de courbe pour l'approche PLAN-MODÈLE, car les estimations des paramètres en fonction de cette approche sont les mêmes que sous l'approche PLAN. Puisque cette mesure est calculable pour un échantillon unique, nous voulons la considérer comme une possibilité pour le dépistage de l'informativité.

La mesure de l'écart standardisé définie plus haut et illustrée au graphique 3 est une moyenne d'un écart standardisé sur 500 échantillons de Monte Carlo. Nous avons également étudié la variabilité sur les échantillons de Monte Carlo de l'écart standardisé obtenu pour diverses approches d'estimation. Nous constatons que les conditions de forte informativité produisent généralement une plus grande dispersion des écarts que les conditions de faible informativité. En outre, nous notons que, même si les approches MIXTENL-POND et MODÈLE ont des valeurs moyennes semblables, la première est nettement plus variable que la seconde, aussi bien d'après les conditions de faible que de forte informativité.

Les deux mesures décrites plus haut servent à comparer les estimations ponctuelles. La suivante nous permet de comparer les variances en calculant le **ratio de la variance $\hat{V}_p(\hat{\theta}_p)$ selon l'approche fondée sur le plan à la variance $\hat{V}_M(\hat{\theta}_M)$ selon une autre approche**. Au graphique 4, nous présentons les moyennes de ces ratios sur les 500 échantillons de Monte Carlo :

$$E_{sim}\left\{\frac{\hat{V}_p(\hat{\theta}_p)}{\hat{V}_M(\hat{\theta}_M)}\right\}. \qquad (3.3)$$

Graphique 4 : Ratio de la variance $\hat{V}_p(\hat{\theta}_p)$ selon l'approche fondée sur le plan à la variance $\hat{V}_M(\hat{\theta}_M)$ selon une autre approche (moyenne sur 500 échantillons)



Dans ce graphique, le degré de proximité de l'approche de rechange par rapport à l'approche fondée sur le plan en ce qui concerne l'estimation de la variance est indiqué par la proximité de la mesure par rapport à la valeur 1. Dans les conditions de faible informativité, toutes les approches ont produit des résultats très semblables et une mesure proche de 1. Dans les conditions de forte informativité, nous observons une plus grande dispersion des courbes et aucune approche ne donne une valeur de 1. L'approche MODÈLE est celle qui s'écarte le plus de celle fondée sur le plan pour cette mesure particulière, alors que pour les deux mesures examinées antérieurement, nous ne pouvions faire la distinction entre les deux. Ces résultats indiquent qu'une comparaison des variances est nécessaire lors de l'évaluation de l'informativité. Nous aborderons la question du choix de la mesure de façon plus approfondie à la section 5.

4. EFFET DE L'INFORMATIVITÉ SUR LA PUISSANCE (ET LA TAILLE) DES TESTS

Nous examinons maintenant l'effet de l'informativité sur les tests d'hypothèse au sujet des coefficients du modèle. Nous évaluons son effet sur la puissance (et la taille) des tests à l'aide de deux exemples. Dans le premier, l'hypothèse nulle testée est effectivement vérifiée dans la population par construction et, dans le deuxième, l'hypothèse nulle est fausse dans la population par construction.

Exemple 1 : Le modèle initial augmenté d'un terme SEX×AGEGR est ajusté à des ensembles de données d'échantillon. Puis, l'hypothèse nulle suivante est formulée : « **H_0 :** il n'existe pas d'interaction entre SEX et AGEGR ». La statistique de test généralement utilisée pour vérifier cette hypothèse est celle de Wald $\hat{X}^2_W = \hat{\theta}^2_{Sex\times Agegr} / \hat{V}(\hat{\theta}_{Sex\times Agegr})$ qui est alors comparée au 95^e centile d'une loi χ^2_1 et ($\chi^2_1(0,95) = 3,841$).

Nous avons élaboré une stratégie pour dériver les courbes de puissance $\gamma(\theta_a)$ en vue de tester l'hypothèse qu'un coefficient est nul. Dans cette stratégie, nous supposons que la statistique de Wald pour l'approche M, dénotée par $\hat{X}^2_{W(M)}$, est proportionnelle à une variable χ^2 non centrée avec un degré de liberté, de sorte que la courbe de puissance pour l'approche M est

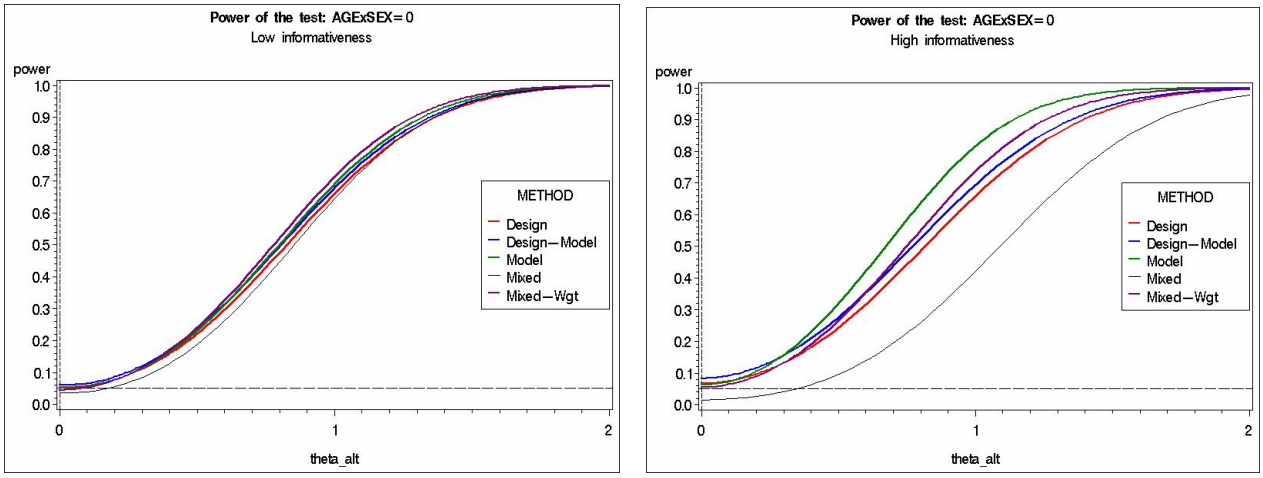
$$\gamma_M(\theta_a) = \text{Prob}\left\{\hat{X}^2_{W(M)} > 3,841 \mid \theta = \theta_a, \hat{X}^2_{W(M)} \sim c_M \chi^2_1(b_{a(M)})\right\},$$

où le paramètre de non-centralité est $b_{a(M)} = d'_{a(M)} V^{-1}_{(M)} d_{a(M)}$ avec $d_{a(M)} = \theta_a + biais_p(\hat{\theta}_M) = E_p(\hat{\theta}_M \mid \theta = \theta_a)$, et le paramètre de proportionnalité est $c_M = V^{-1}_M V_{(M)}$, où $V_M = E_p(\hat{V}_M(\hat{\theta}_M))$ et $V_{(M)} = V_p(\hat{\theta}_M)$. Pour les courbes de puissance simulées, ces deux variances sont l'une et l'autre estimées par calcul des moyennes des estimations appropriées sur 500 échantillons comme le montre le tableau 1. Une nouvelle approximation a été établie pour le cas non centré pour plus d'un degré de liberté et sera publiée ailleurs.

Tableau 1 : Approximations de la variance pour les calculs de puissance

	Approche ***M***				
	PLAN	PLAN-MODÈLE	MODÈLE	MIXTENL	MIXTENL-POND
V_M	$E_{sim}(\hat{V}_p(\hat{\theta}_p))$	$E_{sim}(\hat{V}_\xi(\hat{\theta}_p))$	$E_{sim}(\hat{V}_\xi(\hat{\theta}_\xi))$	$E_{sim}(\hat{V}_{\xi_1}(\hat{\theta}_{\xi_1}))$	$E_{sim}(\hat{V}_{\xi_1}(\hat{\theta}_w))$
$V_{(M)}$	$V_{sim}(\hat{\theta}_p)$	$V_{sim}(\hat{\theta}_p)$	$V_{sim}(\hat{\theta}_\xi)$	$V_{sim}(\hat{\theta}_{\xi_1})$	$V_{sim}(\hat{\theta}_w)$

Graphique 5 : Courbes de puissance pour tester l'hypothèse qu'il n'existe pas d'interaction entre SEX et AGEGR



Au graphique 5, nous ne présentons que la moitié de chaque courbe de puissance, puisque celles-ci sont symétriques par rapport à 0. Les courbes de faible informativité sont fort semblables pour les diverses méthodes. Les courbes de forte informativité divergent davantage, la courbe MIXTENL étant celle qui s'écarte le plus des autres.

Le tableau 2 donne la valeur du coefficient de proportionnalité ***c,*** la valeur de la courbe de puissance théorique au point correspondant à l'alternative vraie (qui est 0) et le taux de rejet empirique lorsque l'hypothèse a été testée en utilisant 500 échantillons de Monte Carlo. Les valeurs de la puissance et du taux de rejet sont assez proches, quelle que soit l'approche suivie, ce qui indique qu'au point 0, la courbe de puissance est bien approximée par le taux de rejet empirique. La puissance et le taux de rejet sont plus proches du niveau nominal de 5 % dans les conditions de faible informativité que dans celle de forte informativité. Le fait que les valeurs obtenues pour l'approche PLAN soient supérieures au niveau nominal de 5 % pourrait être causé par la petite taille des échantillons. L'approche MIXTENL semble donner des résultats différents de ceux des autres approches.

Tableau 2 : Coefficient de proportionnalité, puissance théorique à $\theta_a = 0$ et taux empirique de rejet

Informativité	Faible			Forte		
Approche *M*	$c = V_M^{-1}V_{(M)}$	Puissance théorique à θ_a=0 (taille)	Taux empirique de rejet	$c = V_M^{-1}V_{(M)}$	Puissance théorique à θ_a=0 (taille)	Taux empirique de rejet
PLAN	**1,029**	**0,053**	**0,058**	**1,143**	**0,067**	**0,080**
PLAN-MODÈLE	**1,084**	**0,060**	**0,051**	**1,277**	**0,083**	**0,082**
MODÈLE	**0,999**	**0,050**	**0,053**	**1,102**	**0,063**	**0,058**
MIXTENL	**0,870**	**0,036**	**0,014**	**0,657**	**0,016**	**0,014**
MIXTENL-POND	**0,948**	**0,044**	**0,049**	**1,038**	**0,054**	**0,060**

Exemple 2 : Considérons un cas où l'hypothèse nulle n'est pas vraie dans la population. En particulier, considérons l'hypothèse « **H_0:** SEX n'a pas d'influence sur la probabilité de perdre son autonomie. » Le coefficient de la variable

SEX est effectivement égal à 0,384 et cette valeur est indiquée par une ligne verticale sur le graphique 6. L'ordre et la dispersion des courbes de puissance demeurent assez semblables à ceux observés dans l'exemple précédent; cependant, cela signifie que, sous les conditions de forte informativité, la fourchette de valeur des courbes à la valeur vraie est nettement plus grande. Néanmoins, comme le montre le tableau 3, la puissance théorique et le taux de rejet empirique sont assez proches, aussi bien sous forte que sous faible informativité.

Graphique 6 : Courbes de puissance pour tester l'hypothèse que la variable SEX n'a pas d'effet sur la probabilité de perdre son autonomie

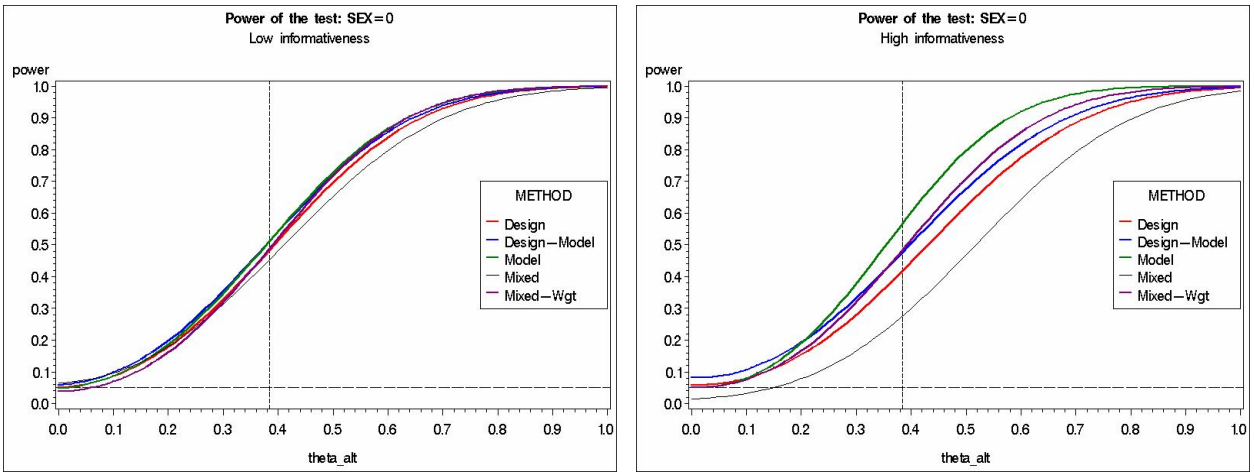


Tableau 3 : Coefficient de proportionnalité, puissance théorique à θ_a= 0,384 et taux empiriques de rejet

Informativité	Faible			Forte		
Approche *M*	$\boldsymbol{c} = \boldsymbol{V}_{\boldsymbol{M}}^{-1}\boldsymbol{V}_{(\boldsymbol{M})}$	Puissance théorique à θ_a=0,384	Taux empirique de rejet	$\boldsymbol{c} = \boldsymbol{V}_{\boldsymbol{M}}^{-1}\boldsymbol{V}_{(\boldsymbol{M})}$	Puissance théorique à θ_a=0,384	Taux empirique de rejet
PLAN	**0,989**	**0,483**	**0,505**	**1,070**	**0,416**	**0,450**
PLAN-MODÈLE	**1,061**	**0,511**	**0,510**	**1,261**	**0,475**	**0,484**
MODÈLE	**0,971**	**0,510**	**0,527**	**1,006**	**0,566**	**0,584**
MIXTENL	**1,131**	**0,453**	**0,448**	**0,636**	**0,275**	**0,246**
MIXTENL-POND.	**0,889**	**0,488**	**0,480**	**0,997**	**0,483**	**0,485**

5. MESURE PROPOSÉE DE L'INFORMATIVITÉ

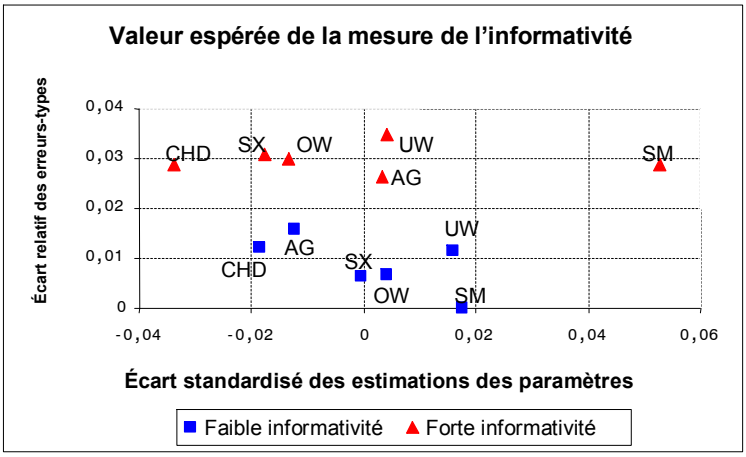
Nous supposons que le chercheur souhaite fonder son inférence statistique sur la distribution des observations résultant de la génération d'une population finie sous un modèle, suivie de la sélection de l'échantillon conformément au plan d'échantillonnage. Étant donné que le modèle réel générant la population finie est celui qui a été utilisé dans l'approche MODÈLE, à la présente section, nous nous limiterons à l'examen de ce modèle. Si le plan d'échantillonnage n'est pas informatif, nous savons que $\boldsymbol{E}_{\xi\boldsymbol{p}}(\hat{\theta}_{\boldsymbol{p}}) = \boldsymbol{E}_{\xi\boldsymbol{p}}(\hat{\theta}_{\xi})$ et $\boldsymbol{E}_{\xi\boldsymbol{p}}(\hat{V}_{\xi}(\hat{\theta}_{\boldsymbol{p}})) = \boldsymbol{E}_{\xi\boldsymbol{p}}(\hat{V}_{\boldsymbol{p}}(\hat{\theta}_{\boldsymbol{p}}))$ (Binder et Roberts, 2003). Partant de cela et de ce que nous avons observé aux sections précédentes, nous proposons une mesure bivariée pour évaluer l'informativité : une composante devrait permettre la comparaison des estimations ponctuelles et l'autre devrait se rapporter à l'estimation de la variance. Nous proposons donc la mesure suivante, qui peut être calculée à partir d'un seul échantillon, pour comparer les approches PLAN et MODÈLE :

$$\left\{ \frac{\hat{\theta}_{\xi} - \hat{\theta}_{\boldsymbol{p}}}{\sqrt{\hat{V}_{\boldsymbol{p}}(\hat{\theta}_{\boldsymbol{p}})}}, 1 - \sqrt{\frac{\hat{V}_{\xi}(\hat{\theta}_{\boldsymbol{p}})}{\hat{V}_{\boldsymbol{p}}(\hat{\theta}_{\boldsymbol{p}})}} \right\}.$$

La transformation particulière des estimations de la variance montrée ici a été choisie afin d'assurer qu'elle soit du même ordre de grandeur que la première composante. Les deux composantes seraient proches de 0 dans la situation de non-informativité.

Nous avons créé un diagramme de dispersion des versions moyennes de simulation de cette mesure au graphique 7 :

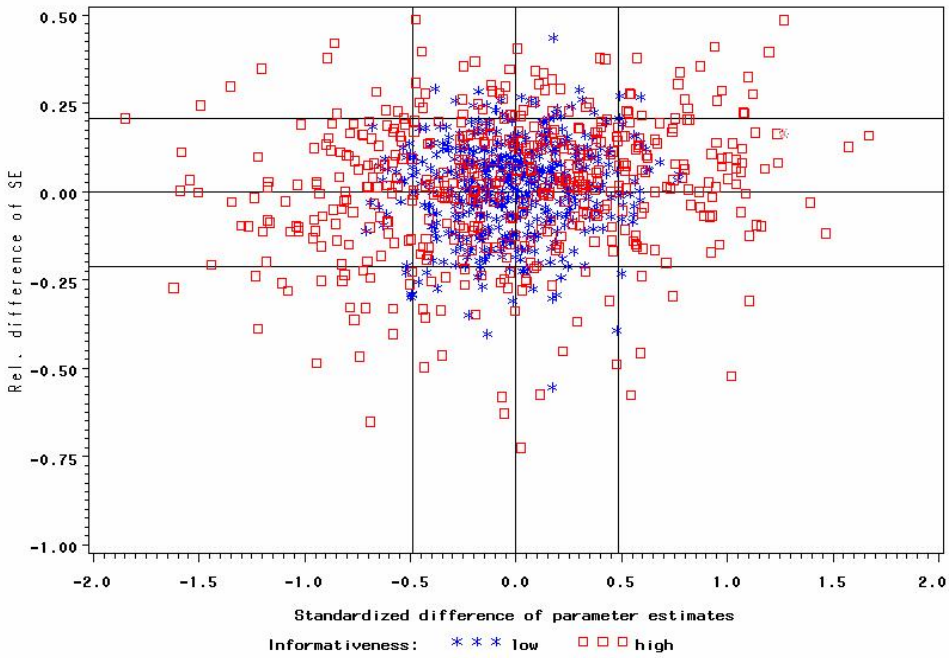
Graphique 7 : Valeur moyenne de la mesure proposée de l'informativité sur 500 échantillons



Il existe une séparation nette entre les cas de faible et de forte informativité pour les valeurs moyennes de variance de la mesure bivariée qui sont portées sur l'axe vertical, tandis que toute distinction concernant les valeurs moyennes des estimations ponctuelles est moins évidente, à part la plus grande dispersion pour le cas de forte informativité.

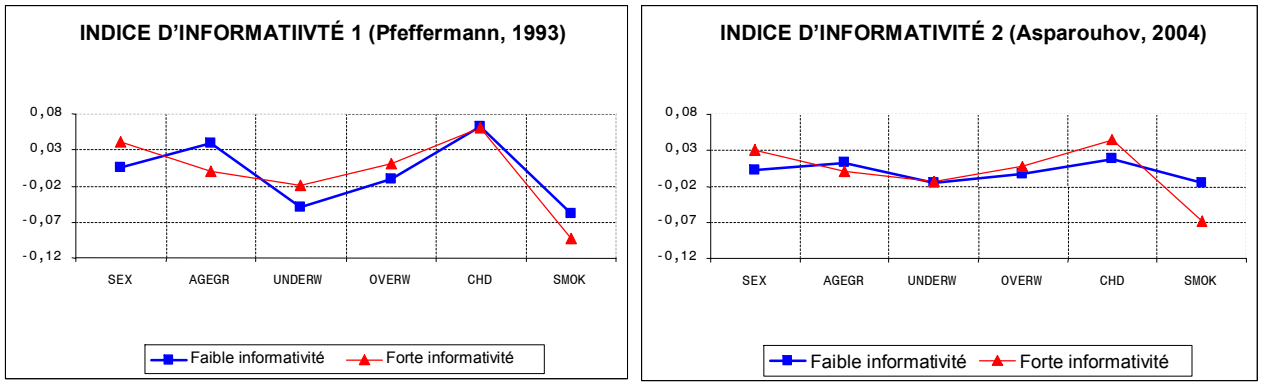
Au graphique 8, nous utilisons les résultats individuels provenant des 500 échantillons de Monte Carlo selon les conditions de faible et de forte informativité pour présenter un diagramme de dispersion de la mesure proposée pour le coefficient de la variable CHD (problème de santé chronique), qui est celle pour laquelle l'écart entre les estimations ponctuelles produites par les approches MODÈLE et PLAN est le plus faible. Sur chacun des axes, nous montrons les 5^{e} et 95^{e} centiles des points obtenus sous les conditions de faible informativité. Alors qu'un pourcentage très élevé des points faible inf. se situent à l'intérieur du carré, 50 % seulement des points forte informativité y sont.

Graphique 8 : Diagramme de dispersion de la mesure proposée obtenue sur 500 échantillons pour la variable CHD pour deux niveaux d'informativité



Il convient de souligner que d'autres mesures de l'informativité sont proposées dans les travaux antérieurs (p. ex., Dumouchel et Duncan (1983), Pfeffermann (1993) et Asparouhov (2004)). Cependant, elles sont entièrement univariées et ne comparent que des estimations ponctuelles. Dans le graphique qui suit, nous présentons les moyennes sur 500 échantillons de l'indice de Pfeffermann et de l'indice d'Asparouhov. De toute évidence, ces deux mesures ne permettent pas de faire la distinction entre les cas de faible et de forte informativité quand les estimations ponctuelles fondées sur le modèle et celles fondées sur le plan sont très semblables.

Graphique 9 : Indice d'informativité de Pfeffermann (1993) et indice d'informativité d'Asparouhov (2004)
Moyenne sur 500 échantillons



6. CONCLUSION

Nous avons étudié dans le présent article l'effet de l'informativité sur certaines conclusions de fond tirées d'une analyse de données d'enquête. Nous avons généralisé une population finie et élaboré un outil permettant de faire varier le niveau d'informativité du plan d'échantillonnage afin de comparer les propriétés de plusieurs approches analytiques d'après divers niveaux d'informativité. Nous constatons que, pour chaque approche considérée, les

résultats obtenus dans des conditions de faible informativité diffèrent de ceux enregistrés sous des conditions de forte informativité et que des différences existent entre les approches. Donc, l'informativité est importante ainsi que et la façon dont elle est traitée.

Nous avons inclus dans l'étude deux approches fondées sur un modèle mixte simple comportant une composante aléatoire destinée à tenir compte de la mise en grappes dans le plan de sondage. En général, le modèle mixte n'a pas donné de bons résultats, indiquant qu'il n'est pas suffisant de tenir compte du plan d'échantillonnage en se limitant à l'inclusion d'un effet de grappe aléatoire dans le modèle, comme il est souvent recommandé dans les articles analytiques. Nous constatons que l'utilisation d'une estimation pondérée dans le modèle mixte améliore légèrement les résultats. Nous notons aussi que l'approche fondée sur le plan n'est pas catégoriquement la meilleure pour toutes les mesures utilisées dans la présente étude par simulation. Il pourrait en être ainsi à cause des petites tailles d'échantillon utilisées dans la simulation. Nous nous attendons à ce que les résultats produits par l'approche fondée sur le plan s'améliore considérablement si l'on augmente les tailles d'échantillon ou que l'on accroît la complexité du plan d'échantillonnage. Nous aimerions étudier ces aspects de façon plus approfondie. Pour évaluer l'informativité, nous soutenons qu'il est nécessaire d'adopter une approche bivariée, mais nous devons poursuivre les travaux avant de pouvoir proposer un test bivarié formel pour déterminer si l'informativité du plan d'échantillonnage est significative.

RÉFÉRENCES

Asparouhov, T. (2004), "Weighting for Unequal Probability of Selection in Multilevel Modeling", *Mplus Web Notes*: No.8.

Binder, D.A., Kovacevic, M.S. et Roberts, G. (2004), "Design-Based Methods For Survey Data: Alternative Uses Of Estimating Functions", *Proceedings Of The Section On Survey Research Methods*, 3301-3312, JSM, Toronto

Binder, D.A. et Roberts, G.R. (2001), "Can informative designs be ignorable?", *Newsletter of the Survey Research Methods Section, Issue* 12, American Statistical Association.

Binder, D. A. et Roberts, G.R. (2003), "Design-based and Model-based Methods for Estimating Model Parameters", Dans *Analysis of Survey Data*, (eds. R.L. Chambers and Chris Skinner) Wiley, Chichester, 29-48.

DuMouchel, W.H. et Duncan, G.J. (1983), "Using Sample Survey Weights in Multiple Regression Analysis of Stratified Samples", *Journal of American Statitical Association*, 78, 535-543.

Martel, L., Bélanger, A. et Berthelot, J.-M. (2002), "Perte et regain de l'autonomie chez les personnes âgées", *Rapports sur la santé*, 13, 35-48.

Pfeffermann, D. (1993), "The Role of Sampling Weights When Modeling Survey Data", *International Statistical Review*, 317-338.