Techniques d'enquête, juin 1995
Vol. 21, n° 1, pp. 81–88
Statistique Canada

Estimation de la médiane à l'aide d'informations supplémentaires

GLEN MEEDEN[1]

RÉSUMÉ

On étudie le problème de l'estimation de la médiane d'une population finie quand une variable auxiliaire est présente. On propose des estimateurs ponctuels et des estimateurs par intervalle fondés sur une approche bayesienne non informative. L'estimateur ponctuel est comparé à d'autres estimateurs possibles et l'on constate qu'il donne de bons résultats dans diverses situations.

MOTS CLÉS: Enquête par sondage; estimation; médiane; variable auxiliaire; quantile; approche bayesienne non informative.

1. INTRODUCTION

On a beaucoup étudié le problème de l'estimation de la moyenne d'une population en présence d'une variable auxiliaire dans les ouvrages sur l'échantillonnage de populations finies. L'estimateur par le quotient a souvent été utilisé dans ce cas. Pour le problème de l'estimation de la médiane d'une population, la situation est très différente. Ce n'est que depuis peu que l'on étudie ce problème. Chambers et Dunstan (1986) ont proposé une méthode pour estimer la fonction de distribution d'une population ainsi que les quantiles associés. Ils ont supposé que la valeur de la variable auxiliaire était connue pour toutes les unités dans la population et leur estimateur provenait d'une méthode fondée sur un modèle. Rao et coll. (1990) ont proposé des estimateurs par le quotient et des estimateurs de différence pour la médiane à l'aide d'une méthode fondée sur un plan. Kuk et Mak (1989) ont proposé deux autres estimateurs de la médiane de la population. Pour utiliser les estimateurs de Kuk et Mak, il suffit de connaître les valeurs de la variable auxiliaire pour les unités de l'échantillon ainsi que sa médiane pour toute la population. L'efficacité de ces estimateurs dépend directement de la probabilité de "concordance" plutôt que de la validité d'une hypothèse de linéarité entre la variable étudiée et la variable auxiliaire.

Récemment, Meeden et Vardeman (1991) ont traité d'une approche bayesienne non informative de l'échantillonnage de populations finies. Cette nouvelle approche utilise la "distribution a posteriori de Polya" comme distribution prédictive pour les unités non observées de la population une fois l'échantillon observé. L'approche donne souvent des estimations ponctuelles et par intervalle qui ressemblent beaucoup à celle que l'on obtient dans le cadre de la théorie "fréquentiste" classique. De plus, l'approche est facile à appliquer dans le cas de problèmes difficiles à traiter avec la théorie classique. Dans le présent article, nous montrons comment cette méthode peut être utilisée dans le cas du problème de l'estimation de la médiane d'une population quand une variable auxiliaire est présente, et nous comparons cette méthode à certaines des autres méthodes proposées.

2. ESTIMATION DE LA MÉDIANE

Considérons une population finie contenant N unités. Pour l'unité i, représentons par y_i la caractéristique étudiée et par x_i la variable auxiliaire. Nous supposons que tant y_i que x_i sont des nombres réels et que cela est connu pour toutes les unités dans la population. Représentons par s un échantillon typique de taille n choisi par échantillonnage aléatoire simple sans remise. Pour des raisons d'ordre pratique, nous supposons un échantillonnage aléatoire simple puisque dans beaucoup de problèmes de ce genre, il arrivera souvent que l'échantillonnage ne sera pas aléatoire simple. Avant de considérer le problème de l'estimation de la médiane d'une population, nous examinons certains faits bien connus à propos du problème de l'estimation de la moyenne.

Considérons le modèle de superpopulation où l'on suppose que pour chaque i, $y_i = bx_i + u_i e_i$. Ici, b est un paramètre inconnu alors que les u_i sont des constantes connues et que les e_i sont des variables aléatoires indépendantes suivant la même distribution avec espérance mathématique nulle. Comme la moyenne de population peut être représentée par $N^{-1}(\sum_{i\in s} y_i + \sum_{j\notin s} y_j)$, nous pourrions penser que $N^{-1}(\sum_{i\in s} y_i + \hat{b}\sum_{j\notin s} x_j)$ serait une estimation raisonnable de la moyenne chaque fois que $\hat{b}$ est une estimation raisonnable de b. Un choix particulier de $\hat{b}$ est l'estimateur par les moindres carrés pondéré où les poids sont déterminés par les u_i. Par exemple, si, pour tout i,$u_i = \sqrt{x_i}$, l'estimateur résultant n'est que l'estimateur par le quotient habituel. Alors que si, pour tout i, $u_i = x_i$, alors $\hat{b} = n^{-1}\sum_{i\in s}(y_i/x_i)$ et l'estimateur résultant est celui qui a été étudié par Basu (1971). (Voir aussi Royall (1970).) À l'aide de ce modèle de superpopulation, il est facile de créer des populations où l'estimateur par le quotient a une erreur quadratique moyenne plus petite que celle de l'estimateur de Basu et inversement. Une étude de simulation assez limitée portant sur diverses populations a permis de constater que le rendement de l'estimateur de Basu est très semblable à celui de l'estimateur par le quotient, bien que dans la majorité des cas l'estimateur par le quotient donne un meilleur résultat que l'estimateur de Basu. Cela n'est pas inattendu, compte tenu de l'utilisation répandue de l'estimateur par le quotient.

[1] Glen Meeden, School of Statistics, University of Minnesota, Minneapolis, MN 55455, U.S.A.

Dans Meeden et Vardeman (1991), on a élaboré une approche bayesienne non informative à l'échantillonnage de populations finies, fondée sur la "distribution a posteriori de Polya". Pour le cas simple où aucune variable auxiliaire n'est présente, compte tenu des valeurs observées dans l'échantillon, cette approche introduit une distribution de Polya comme une pseudo-distribution a posteriori pour les unités non observées de la population. Cette pseudo-distribution a posteriori peut être utilisée pour obtenir des estimations ponctuelles et des estimations par intervalle de diverses quantités relatives à la population. Elle est liée à la méthode bootstrap bayesienne de Rubin (1981) et à la distribution *a priori* du processus Dirichlet de Ferguson (1973). Quand on estime la médiane, cette méthode donne des résultats semblables à ceux obtenus par Binder (1982). Un argument de Bayes pas à pas qui donne l'admissibilité des estimateurs résultants constitue une justification théorique de cette méthode. Voyez, par exemple, Meeden et Ghosh (1983). C'est dans cet article que l'on a démontré l'admissibilité de l'estimateur de Basu. Dans ce cas, on a montré que l'estimateur de Basu provenait d'une "distribution a posteriori" qui traite les ratios connus et inconnus, $r_i = y_i/x_i$ comme interchangeables. Il faut remarquer que cela ressemble beaucoup, en esprit, à la justification à l'aide du modèle de superpopulation présenté plus haut, où les ratios $r_i = y_i/x_i$ étaient indépendants et suivaient une distribution identique. Nous verrons que l'argument de Bayes pas à pas, qui sous-tend l'estimateur de Basu pour la moyenne, peut s'appliquer de façon simple aux estimateurs ponctuels et aux estimateurs par intervalle pour la médiane. Malheureusement, cela n'est pas le cas pour certains des autres estimateurs. Le ratio de la médiane des valeurs y dans l'échantillon sur la médiane des valeurs x dans l'échantillon multiplié par la médiane des valeurs x dans la population constitue un estimateur naturel, mais peut-être élémentaire, qui imite, dans un certain sens, l'estimateur par le quotient de la moyenne. Il n'y a pas de théorie connue fondée sur un modèle qui sous-tende cet estimateur, comme c'est le cas pour l'estimateur par le quotient de la moyenne.

Dans l'approche bayesienne de l'échantillonnage de populations finies, il faut préciser une distribution *a priori*. Puis, étant donné un échantillon, des inférences sont fondées sur la distribution a posteriori, qui est la distribution prédictive des unités non observées de la population compte tenu des unités dans l'échantillon. Dans l'approche bayesienne pas à pas, étant donné l'échantillon, on a toujours une distribution "a posteriori", mais elle ne découle pas d'une distribution *a priori* unique. Toutefois, on peut utiliser cette distribution "a posteriori" selon la méthode bayesienne habituelle pour trouver les estimateurs ponctuels et les estimateurs par intervalle des paramètres étudiés. Nous allons maintenant montrer comment le modèle bayesien pas à pas qui donne l'estimateur de Basu pour la moyenne peut aussi être utilisé quand on estime la médiane. Dans ce modèle, étant donné l'échantillon, la distribution prédictive des ratios non observés traite les ratios observés et les ratios non observés comme "interchangeables".

Pour plus de précision, supposons que notre échantillon renferme les n premières unités de la population. Nous construisons une urne qui contient n boules où l'on donne à la boule i la valeur du i-ième ratio observé, disons r_i. Nous commençons par choisir une boule au hasard dans l'urne, et la valeur observée est attribuée à l'unité non observée $n + 1$. Cette boule ainsi qu'une boule additionnelle ayant la même valeur sont remises dans l'urne. Une autre boule est choisie dans l'urne, et sa valeur est attribuée à l'unité non observée $n + 2$. Cette boule ainsi qu'une autre boule ayant la même valeur sont remises dans l'urne. On poursuit ce processus jusqu'à ce qu'un ratio ait été attribué à toutes les unités non observées. Une fois que nous avons attribué une valeur à toutes ces unités, nous avons observé une réalisation provenant de notre distribution "a posteriori" pour les ratios non observés, étant donné l'échantillon de ratios observés. Si, dans ce processus, on a attribué à l'unité non observée j le ratio avec valeur r, nous disons alors que sa valeur y_j est rx_j. Par conséquent, à l'aide d'un échantillonnage de Polya simple, étant donné l'échantillon, nous avons créé une distribution prédictive pour les unités non observées. Nous appelons cette distribution prédictive la "distribution a posteriori de Polya". Il est facile de vérifier que cette distribution prédictive donne l'estimateur de Basu lorsque nous estimons la moyenne de la population quand nous avons une fonction quadratique de perte.

Étant donné l'échantillon, la "distribution a posteriori de Polya" donne une distribution prédictive pour les unités non observées de la population et, par conséquent, elle donne aussi une distribution prédictive pour la médiane. Du point de vue de la théorie de la décision, la fonction de perte habituelle est l'erreur absolue quand nous estimons une médiane. Pour cette fonction de perte, l'estimation bayesienne n'est que la médiane de la distribution a posteriori ou prédictive pour la médiane de la population. Si l'on utilisait une fonction quadratique de perte pour estimer la médiane, alors l'estimation bayesienne ne serait que la moyenne de la distribution prédictive pour la médiane de la population. L'admissibilité de ces estimateurs quand nous utilisons une fonction de perte appropriée découle d'un argument bayesien pas à pas de la même façon que la preuve de l'admissibilité, pour l'estimateur de Basu, de la moyenne de la population. Dans Meeden et Vardeman (1991) et Meeden (1993), on a remarqué le fait un peu surprenant décrit ci-après. Pour beaucoup de distributions courantes, la moyenne de la distribution prédictive de la médiane de la population a donné de meilleurs résultats que la médiane de la distribution prédictive de la médiane de la population selon les deux fonctions de perte. Des résultats semblables valent pour ce problème. Par conséquent, notre estimateur sera la moyenne de la distribution prédictive de la médiane de la population, bien que nous suivrons l'usage courant et utiliserons l'erreur absolue comme fonction de perte. Nous désignerons cet estimateur par *estpp*. Cet estimateur ne peut être trouvé de façon explicite. Toutefois, nous le trouverons de façon approximative en simulant des observations à partir de la distribution a posteriori ou prédictive de la médiane de la

population. Selon le plan de sondage de Polya, pour les ratios décrits ci-dessus, nous pouvons simuler une réalisation possible de toute la population. Pour cette copie simulée, nous pouvons alors trouver sa médiane. Si nous répétons ce processus R fois, nous aurons alors simulé la distribution prédictive de la médiane de la population selon la "distribution a posteriori de Polya". Quand R est grand, la moyenne de ces R médianes simulées de la population donne, approximativement, l'estimateur *estpp*.

Nous allons maintenant comparer l'estimateur *estpp* à plusieurs autres estimateurs, dont un est tout simplement la médiane de l'échantillon des y_i. Pour cet estimateur, représenté par *estsm*, que nous employons comme étalon, nous ne tenons pas compte des renseignements contenus dans la variable auxiliaire. Un autre estimateur est l'analogue naturel de l'estimateur par le quotient de la moyenne de la population. Cet estimateur est étudié dans Kuk et Mak (1989), et nous le représentons par *estrm*. Il ne s'agit que du ratio de la médiane des valeurs y sur la médiane des valeurs x de l'échantillon, multiplié par la médiane de toutes les valeurs x dans la population. Les auteurs ont proposé deux autres estimateurs pour la médiane. Nous n'étudierons que le premier d'entre eux et le représentons par *estkm*. Cet estimateur a une justification intuitive plausible qui est présentée dans leur article. Rao, Kovar et Mantel (1990) ont étudié un estimateur fondé sur un plan de la médiane. Nous représenterons cet estimateur par *estrkm*. Comme le calcul de cet estimateur peut prendre beaucoup de temps, nous en obtiendrons une approximation à l'aide d'une méthode présentée dans Mak et Kuk (1993). Enfin, nous allons considérer l'estimateur proposé dans Chambers et Dunstan (1986) et le représenter par *estcd*. En fait, Chambers et Dunstan proposent toute une famille d'estimateurs, et nous n'en étudierons qu'un cas spécial qui convient lorsque $u_i = \sqrt{x_i}$ dans le modèle de superpopulation décrit au début de la présente section. Nous allons maintenant présenter brièvement le raisonnement qui les amène à leur estimateur de la médiane. Représentons par F la fonction de distribution cumulative associée aux valeurs y de la population. C'est-à-dire que F attribue un poids $1/N$ à chaque y_i de la population totale. La première étape consiste à obtenir un estimateur de $F(t)$ pour un nombre réel arbitraire t. Si s représente notre échantillon de taille n, alors, compte tenu de l'échantillon, nous pouvons poser

$$F(t) = N^{-1}\left\{ \sum_{i\in s} \Delta(t - y_i) + \sum_{j\notin s} \Delta(t - y_j)\right\}$$

où $\Delta(z)$ est la fonction en escalier qui a la valeur un lorsque $z \geq 0$ et zéro ailleurs. Puisque la première somme dans l'expression ci-dessus est connue une fois que nous avons observé l'échantillon, pour obtenir une estimation de $F(t)$, il suffit de trouver une estimation de la seconde somme. Or, d'après notre modèle supposé de superpopulation, les ratios de population $(y_i - bx_i)/\sqrt{x_i}$ sont des variables aléatoires indépendantes qui suivent la même distribution. Puisque, une fois l'échantillon s observé, $\hat{b} = \sum_{i\in s} y_i / \sum_{i\in s} x_i$ constitue une estimation naturelle de b, on pourrait agir comme si les n ratios connus $(y_i - \hat{b}x_i)/\sqrt{x_i}$ pour $i \in s$ sont des observations réelles tirées de cette distribution inconnue. Selon cette hypothèse, pour un t fixe et une unité j fixe qui ne fait pas partie de l'échantillon s, une estimation de $\Delta(t - y_j)$ est donnée tout simplement par le nombre des n ratios connus qui incorporent $\hat{b}$ inférieurs ou égaux à $(t - \hat{b}x_j)/\sqrt{x_j}$ divisé par n. Finalement, si nous calculons la somme, pour toutes les unités j non observées, des estimations de $\Delta(t - y_j)$, nous avons donc une estimation de la seconde somme dans l'expression ci-dessus pour $F(t)$ qui donne alors une estimation de $F(t)$. Une fois que nous pouvons estimer $F(t)$ pour tout t par, disons, $\hat{F}(t)$, l'estimation de la médiane de la population est $\inf\{t:\hat{F}(t) \geq 0.5\}$.

3. LES POPULATIONS

Nous allons comparer ces estimateurs à l'aide de plusieurs populations différentes. Nous commençons avec trois populations réelles. La première est un groupe de 125 villes américaines. La variable x est leur population en 1960, en millions d'habitants, alors que leur variable y est la population correspondante en 1970, à nouveau, en millions d'habitants. La deuxième est un groupe de 304 comtés américains. La variable x est le nombre de familles dans les comtés en 1960, alors que la variable y est la population totale du comté en 1960. Les deux variables sont présentées en milliers d'habitants. La troisième population est un groupe de 331 grosses sociétés. La variable x est leurs ventes totales en 1974 et la variable y, leurs ventes totales en 1975. Les ventes sont présentées en milliards de dollars. Nous représentons ces trois populations par *ppcities*, *ppcounties* et *ppsales*. Pour les trois populations, les coefficients de corrélation sont .947, .998 et .997, respectivement. Ces populations ont été étudiées dans Royall et Cumberland (1981). Notre *ppcounties* ressemble à leur population Counties60 sauf que, pour nous, la variable x représente le nombre de familles plutôt que le nombre de ménages.

Nous avons aussi étudié six populations artificielles. Dans chaque cas, nous avons d'abord choisi la variable auxiliaire x, puis nous avons produit la variable y à partir de la variable auxiliaire. Dans certains cas, nous avons suivi le modèle de superpopulation décrit au début de la section précédente pour un certain choix des u_i. Dans d'autres cas, nous n'avons pas respecté l'hypothèse selon laquelle la moyenne de y_i est bx_i, étant donné la valeur x_i. Dans tous les cas, les erreurs, les e_i étaient des variables aléatoires indépendantes suivant une distribution normale identique de moyenne nulle et de variance un.

Dans la première population, *ppgamma20*, les x_i étaient un échantillon aléatoire tiré d'une distribution gamma avec vingt comme paramètre de forme et un comme paramètre d'échelle. Alors, étant donné x_i, la distribution conditionnelle de y_i était normale avec moyenne $1.2x_i$ et variance x_i, c.-à-d. $u_i = \sqrt{x_i}$.

Dans la deuxième population, *ppgamma5a*, la valeur des x_i était de dix plus une valeur d'un échantillon aléatoire provenant d'une distribution gamma avec cinq comme

paramètre de forme et un comme paramètre d'échelle. Alors, étant donné x_i, la distribution conditionnelle de y_i était normale avec moyenne $3x_i$ et variance x_i.

Dans *ppgamma5b*, la variable auxiliaire était la même que dans *ppgamma5a*. Alors, étant donné x_i, la distribution conditionnelle de y_i était normale avec moyenne $3x_i$ et variance x_i^2.

Dans *ppstskew*, la variable auxiliaire était fortement désaxée vers la droite avec moyenne 42.63, médiane 39.29 et variance 204.59. Alors, étant donné x_i, la distribution conditionnelle de y_i était normale avec moyenne $x_i + 5$ et variance $9x_i$.

Dans *ppln*, la variable auxiliaire était un échantillon aléatoire provenant d'une population lognormale avec moyenne et écart-type (du logarithme) de 4.9 et de .586 respectivement. Alors, étant donné x_i, la distribution conditionnelle de y_i était normale avec moyenne $x_i + 2\log x_i$ et variance x_i^2.

Dans *ppexp*, la valeur de la variable auxiliaire était cinquante plus une valeur d'un échantillon aléatoire provenant de la distribution exponentielle traditionnelle. Alors, étant donné x_i, la distribution conditionnelle de y_i était normale avec moyenne $80 - x_i$ et variance $(.6 \log x_i)^2$.

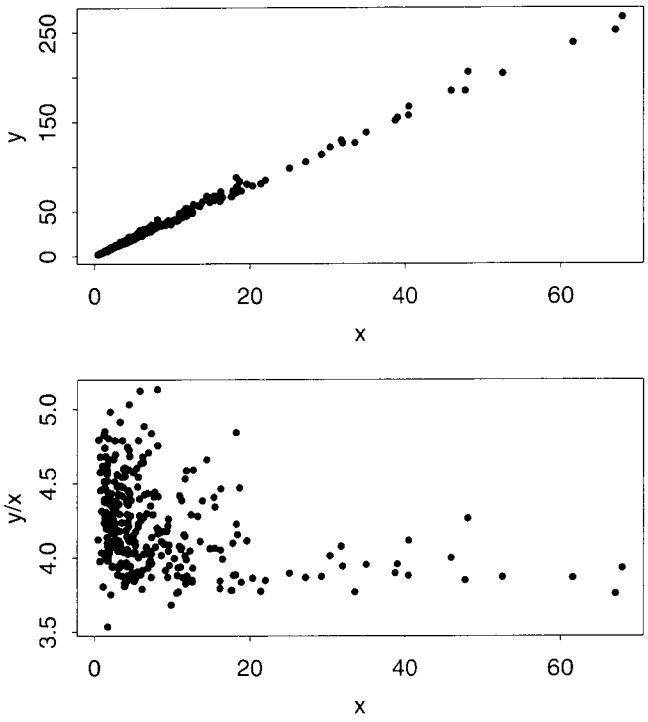

Figure 1. Pour *ppcounties*, tracé de y par rapport à x et de y/x par rapport à x où x représente le nombre de familles (en milliers) vivant dans un comté et y, la population totale (en milliers) du comté pour 304 comtés.

Toutes les populations renferment 500 unités sauf *ppstskew*, qui en a 1,000. Les coefficients de corrélation entre les deux variables pour ces six dernières populations sont de .76, .87, .41, .61, .58 et $-.28$, respectivement.

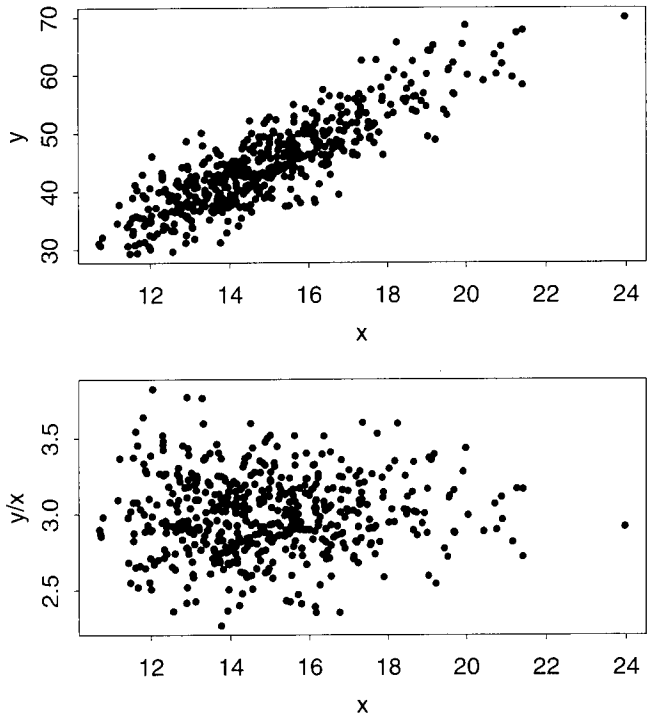

Figure 2. Pour *ppgamma5a*, tracé de y par rapport à x et de y/x par rapport à x.

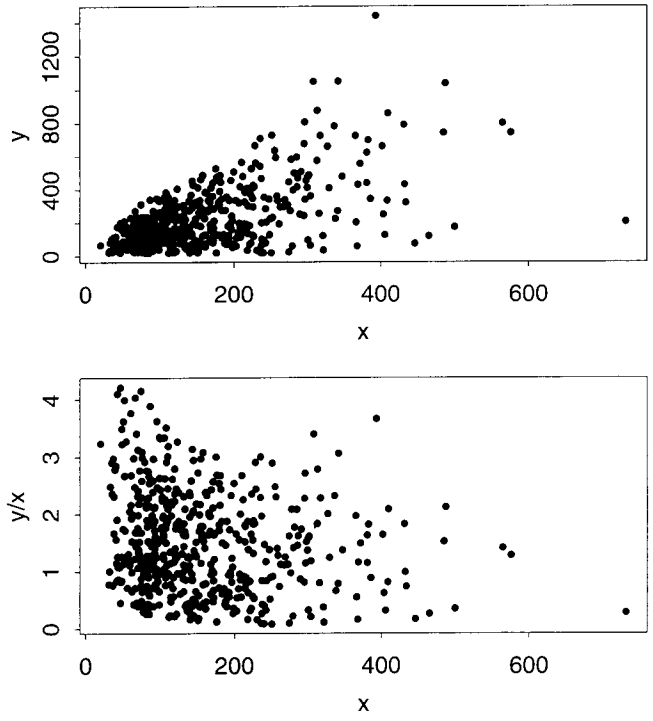

Figure 3. Pour *ppln*, tracé de y par rapport à x et de y/x par rapport à x.

Dans la majorité des exemples, où l'on utilise des estimateurs par quotient, tant les y_i que les x_i sont habituellement strictement positifs. Dans la population *ppstskew*, 13 des 1,000 unités ont une valeur y négative. Dans la construction originale de la population *ppln*, un nombre beaucoup plus important de valeurs étaient négatives. La population a été modifiée pour que toutes les valeurs soient supérieures à zéro.

On remarquera que ces populations ont été construites selon divers scénarios pour le lien entre les variables x et y. *Ppgamma20* et *ppgamma5a* respectent les hypothèses du modèle de superpopulation qui mènent à *estcd*, alors que *ppgamma5b* est compatible avec les hypothèses qui sous-tendent *estpp*. Dans *ppstskew*, la variance conditionnelle de y_i étant donné x_i, est compatible avec *estcd*, alors que pour la population *ppln* non modifiée elle était compatible avec *estpp*. Dans ces deux cas, l'hypothèse pour l'espérance conditionnelle n'est pas respectée. Dans le cas des populations *ppcounties*, *ppgamma5a* et *ppln*, nous avons représenté graphiquement y par rapport à x et y/x par rapport à x. Les résultats sont présentés dans les figures 1 à 3.

L'estimateur *estpp* est fondé sur l'hypothèse qu'étant donné l'échantillon s, notre opinion à propos des ratios observés, c.-à-d. les ratios y_i/x_i pour $i \in s$, et notre opinion sur les ratios non observés, c.-à-d. les ratios y_j/x_j pour $j \notin s$ sont à peu près interchangeables. En particulier, cela signifie que l'opinion qu'on peut avoir à propos d'un ratio y_j/x_j ne devrait pas dépendre de la taille de x_j. En fait, *ppgamma5b* a été construite de façon que cela soit effectivement vrai. Par contre, dans le cas du modèle de superpopulation qui mène à l'estimateur *estcd*, nous nous attendrions à ce que la variabilité des ratios diminue à mesure qu'augmente la taille de la variable x tandis que la valeur moyenne des ratios dans toute bande verticale étroite demeure à peu près constante lors du déplacement de la bande vers la droite. Cela se voit clairement dans le tracé des ratios pour la population *ppgamma5a*. Pour les autres populations, sauf pour *ppgamma20*, la valeur des ratios dépend en fait de la taille de x. Cela se voit clairement dans les tracés pour *ppcounties* et *ppln*. Par conséquent, ces populations devraient constituer des cas types intéressants pour l'estimateur *estpp*. *Ppexp* a été incluse comme cas type pour qu'on puisse voir ce qui se produirait si l'on respectait peu les hypothèses sur lesquelles sont fondés *estpp* et *estcd*.

4. CERTAINS RÉSULTATS DES SIMULATIONS

Pour comparer les six estimateurs, 500 échantillons aléatoires simples de diverses tailles ont été tirés parmi les neuf populations. Pour chaque échantillon, la valeur des six estimateurs a été calculée. Pour l'estimateur *estpp*, cela signifie qu'il a fallu trouver cet estimateur de façon approximative en simulant $R = 500$ réalisations de la distribution prédictive de la médiane de la population induite par la "distribution a posteriori de Polya". Dans chaque cas, on a calculé la valeur moyenne et l'erreur absolue moyenne de l'estimateur. Dans le tableau 1, la valeur moyenne de tous les estimateurs sauf *estsm* est présentée. Tous les estimateurs sont approximativement sans biais sauf dans un cas, *estcd* pour la population *ppln*. Nous n'avons pas inclus les résultats pour *estsm*, puisqu'il est bien établi que cet estimateur est sans biais. Dans le tableau 2, l'erreur absolue moyenne pour les six estimateurs est présentée. Nous voyons, dans ce tableau, que *estcd* et *estpp* sont de loin les meilleurs estimateurs. Ils donnent tous les deux de meilleurs résultats que les quatre autres estimateurs dans tous les cas sauf un. Dans *ppexp*, *estsm* a donné de meilleurs résultats que ces deux estimateurs, mais il s'agit d'un cas où l'on ne pouvait s'attendre à ce que l'un ou l'autre de ces deux estimateurs donne de bons résultats. Pour les sept premières populations, le rendement était presque identique alors que, pour la population *ppln*, l'estimateur *estpp* est préféré et que, pour la population *ppstskew*, c'est le contraire.

Tableau 1

Valeur moyenne de cinq estimateurs de la médiane pour 500 échantillons aléatoires simples

Population (médiane)	Taille de l'échantillon	Valeur moyenne de l'estimateur				
		estrm	*estkm*	*estrkm*	*estcd*	*estpp*
ppcities (1.90)	25	.197	.196	.193	.195	.195
ppsales (1.24)	30	1.21	1.25	1.23	1.25	1.24
ppcounties (18.33)	30	18.21	18.60	18.66	18.26	18.39
ppexp (29.02)	30	29.03	29.05	29.00	29.03	29.05
ppgamma5a (43.90)	30	43.82	43.88	43.91	43.99	43.89
	50	43.90	43.91	43.85	44.06	43.90
ppgamma5b (44.17)	30	43.84	43.96	44.19	44.15	43.61
	50	44.28	44.37	44.18	44.18	43.98
ppgamma20 (23.15)	30	23.47	23.28	23.14	23.46	23.77
	50	23.34	23.18	23.17	23.43	23.18
ppln (170.25)	30	171.15	169.38	168.12	185.01	170.61
	50	169.15	167.54	167.65	185.03	169.61
ppstskew (46.12)	30	43.66	40.27	45.88	45.50	45.11
	50	44.04	40.70	46.01	45.43	45.37

Tableau 2

Erreur absolue moyenne de six estimateurs de la médiane pour 500 échantillons aléatoires simples

Population	Taille de l'échantillon	Erreur absolue moyenne de l'estimateur					
		estsm	*estrm*	*estkm*	*estrkm*	*estcd*	*estpp*
ppcities	25	.0326	.0161	.0162	.0155	.0075	.0072
ppsales	30	.1797	.0770	.0797	.0870	.0244	.0245
ppcounties	30	3.12	.586	.964	1.34	.215	.214
ppexp	30	.43	.49	.48	.47	.48	.46
ppgamma5a	30	1.36	.96	1.03	.89	.54	.53
	50	.95	.74	.78	.65	.44	.43
ppgamma5b	30	2.84	2.74	2.71	2.58	2.37	2.38
	50	2.08	2.04	2.01	1.89	1.80	1.85
ppgamma20	30	1.08	1.06	1.05	.88	.67	.64
	50	.94	.77	.78	.73	.51	.49
ppln	30	25.9	25.8	24.2	21.62	21.4	17.0
	50	18.0	20.1	17.9	16.46	17.7	12.7
ppstskew	30	3.86	4.26	6.69	3.21	2.72	3.14
	50	2.92	3.63	5.82	2.55	2.20	2.51

En pratique, il arrive souvent qu'on veuille obtenir et des estimations par intervalle et des estimations ponctuelles pour les paramètres étudiés. Dans Kuk et Mak (1989) et dans Chambers et Dunstan (1986), on suggère des méthodes qui permettent de trouver, à l'aide de la théorie asymptotique, des estimations par intervalle fondées sur les estimateurs proposés par ces auteurs. Mais dans aucun cas les auteurs n'ont trouvé d'estimateurs par intervalle. Meeden et Vardeman (1991) ont fait remarquer comment on peut trouver de façon approximative des régions de crédible à 95% approximatives fondées sur la "distribution a posteriori de Polya". Si nous posons que $q(.025)$ et $q(.975)$ représentent respectivement les .025 ième et .975 ième quantiles de la collection des médianes pour 500 populations simulées sous une "distribution a posteriori de Polya", alors $(q(.025), q(.975))$ est un intervalle de crédible à 95% approximatif. (Voir Berger 1985 pour la définition de tels intervalles.) Le tableau 3 présente l'étendue moyenne et la fréquence relative de couverture de ces intervalles. Nous constatons que, pour ces populations, les intervalles ont des propriétés "fréquentistes" raisonnables. Cela n'est peut-être pas inattendu étant donné la discussion présentée dans Meeden et Vardeman (1991). Mais, par contre, une seule des populations a été construite de façon que les ratios y_i/x_i soient interchangeables. Ces résultats laissent supposer que les estimateurs ponctuels et les estimateurs par intervalle de la médiane fondés sur la "distribution a posteriori de Polya" pour les ratios sont assez robustes relativement à l'hypothèse d'interchangeabilité et qu'ils devraient donner de bons résultats dans diverses situations. Nous parlerons plus longuement de cette question dans la section 5.

Tableau 3

Étendue moyenne et fréquence relative de couverture pour un intervalle de crédible à .95 pour la médiane fondée sur la "distribution a posteriori de Polya" pour 500 échantillons aléatoires simples

Population	Taille de l'échantillon	Étendue moyenne	Fréquence de couverture
ppcities	25	.041	.968
ppsales	30	.141	.964
ppcounties	30	1.44	.994
ppexp	30	2.26	.944
ppgamma5a	30	2.70	.950
	50	2.15	.956
ppgamma5b	30	11.67	.932
	50	8.86	.942
ppgamma20	30	3.24	.960
	50	2.51	.964
ppln	30	84.8	.934
	50	65.4	.956
ppstskew	30	15.52	.936
	50	12.00	.938

5. DISCUSSION

Le calcul de l'estimateur *estpp* est fondé sur l'hypothèse que les ratios de population y_i/x_i sont interchangeables. Cette hypothèse peut être décrite mathématiquement de deux manières distinctes mais liées. La première est le modèle de superpopulation déjà présenté; la seconde découle de la "distribution a posteriori de Polya", qui est fondée sur un argument bayesien pas à pas et qui donne une interprétation bayesienne non informative de l'estimateur. Cette deuxième méthode peut être utilisée quel que soit le paramètre estimé. Quand on estime la moyenne, cette méthode mène à l'estimateur de Basu, qui donne à peu près le même résultat que l'estimateur par le quotient, bien que ce dernier donne habituellement un résultat légèrement supérieur. Quand on estime la médiane, cette méthode mène à l'estimateur dont on traite dans le présent article. Ici, nous avons soutenu que la "distribution a posteriori de Polya" pour les ratios mène à de bons estimateurs ponctuels et à de bons estimateurs par intervalle de la médiane quand une variable auxiliaire est présente et qu'elle semble assez robuste par rapport à l'hypothèse de l'interchangeabilité des ratios y_i/x_i.

Royall et Cumberland (1981) ont présenté une étude empirique de l'estimateur par le quotient et d'estimateurs de sa variance. Ils ont soutenu qu'étant donné un échantillon il arrive souvent qu'une estimation de la variance, fondée sur le modèle de superpopulation, qui mène à l'estimateur par le quotient, soit plus raisonnable qu'une estimation fondée sur un plan basée sur une distribution d'échantillonnage probabiliste. Dans Royall et Cumberland (1985), ils ont démontré que, sous réserve de la moyenne de l'échantillon de la variable auxiliaire, les propriétés de couverture conditionnelle de l'intervalle de crédible habituel fondé sur un plan, pour la moyenne de la population, n'étaient absolument pas fiables.

Nous allons maintenant aborder la question du comportement conditionnel des intervalles pour la médiane fondée sur la "distribution a posteriori de Polya" qui ont été élaborés dans cet article. Dans les études de simulation présentées plus tôt, nous avons utilisé un échantillonnage aléatoire simple pour des raisons d'ordre pratique. Pour obtenir une idée du comportement conditionnel de la "distribution a posteriori de Polya", nous avons étudié cinq de nos populations. Dans chaque cas, nous avons classé la population à l'aide des valeurs de la variable auxiliaire x. Nous avons alors tiré 500 échantillons aléatoires de la première (ou plus petite) moitié de la population, puis 500 échantillons aléatoires additionnels de la deuxième (ou plus grande) moitié de la population et finalement 500 échantillons aléatoires de plus dans le tiers médian de la population. Puis, nous avons calculé l'intervalle de crédible à .95 pour la médiane fondé sur la "distribution a posteriori de Polya" qui suppose l'interchangeabilité des ratios y_i/x_i. Dans le tableau 4, nous présentons les résultats pour les estimateurs fondés sur la "distribution a posteriori de Polya" de la médiane. (Nous avons aussi calculé la valeur moyenne et l'erreur absolue moyenne de l'estimateur *estcd* pour ces exemples. Nous n'avons pas inclus

ces résultats ici puisqu'ils correspondent étroitement aux résultats de la "distribution a posteriori de Polya".) Nous voyons que leur comportement conditionnel, du moins dans ces cas, ressemble beaucoup à leur comportement inconditionnel. En bref, les estimations par intervalle de la médiane fondées sur la "distribution a posteriori de Polya" devraient avoir des propriétés "fréquentistes" raisonnables, quelle que soit la façon dont l'échantillon a été choisi, pourvu que la population respecte, de façon approximative, notre opinion que les ratios sont à peu près interchangeables.

Tableau 4

Valeur moyenne et erreur absolue moyenne pour l'estimateur ponctuel et étendue moyenne et fréquence de couverture pour un intervalle de crédible à .95 de la médiane fondée sur la "distribution a posteriori de Polya" pour 500 échantillons aléatoires simples tirés de l'ensemble de la population, de la moitié "la plus petite", de la moitié "la plus grande" et du tiers "médian"

Population	Taille de l'échantillon	Partie de la population d'où l'échantillon est tiré	Valeur moyenne	Erreur moyenne	Étendue moyenne	Fréquence de couverture
ppcities	25	population totale	.195	.0072	.041	.968
		½ la plus petite	.192	.0047	.033	.994
		½ la plus élevée	.196	.0078	.048	.988
		⅓ médian	.201	.0114	.055	.922
ppcounties	30	population totale	19.4	.220	1.46	.990
		½ la plus petite	18.6	.305	1.34	.942
		½ la plus élevée	18.1	.283	1.59	.954
		⅓ médian	18.5	.252	1.35	.964
ppsales	30	population totale	1.24	.0072	.141	.964
		½ la plus petite	1.24	.027	.153	.966
		½ la plus élevée	1.23	.020	.125	.982
		⅓ médian	1.23	.027	.139	.944
ppgamma5a	30	population totale	43.9	.53	2.70	.950
		½ la plus petite	43.8	.55	2.82	.948
		½ la plus élevée	44.0	.53	2.55	.940
		⅓ médian	43.9	.47	2.63	.974
ppgamma5b	30	population totale	43.6	2.38	11.7	.932
		½ la plus petite	42.2	2.69	11.6	.890
		½ la plus élevée	45.1	2.25	11.2	.950
		⅓ médian	45.2	2.27	11.3	.936

Comme on peut le voir en étudiant les tracés de y_i/x_i par rapport à x_i et les résultats de nos simulations, le fait que les ratios y_i/x_i diminuent à mesure que x_i augmente ne semble pas avoir beaucoup d'importance. Toutefois, il est crucial que la valeur moyenne des ratios dans la bande étroite au-dessus d'un petit intervalle de valeurs x possibles reste assez constante à mesure que nous déplaçons le petit intervalle vers la droite. Dans la figure 2, le tracé des ratios pour *ppgamma5a* est un exemple d'une telle représentation graphique. En fait, c'est de cette façon que la population a été construite, puisqu'elle est conforme aux hypothèses qui sous-tendent *estcd*. Dans les figures 1 et 3, nous voyons, pour *ppcounties* et pour *ppln*, que la valeur moyenne des ratios dans une bande étroite tend à diminuer à mesure que nous nous déplaçons vers la droite, et cela aide à expliquer le rendement relativement moins bon des estimateurs de la "distribution a posteriori de Polya" dans ce cas. Globalement toutefois, le rendement des procédures fondées sur la "distribution a posteriori de Polya" semble être assez robuste par rapport à l'hypothèse d'interchangeabilité.

Comme autre possibilité, nous pourrions considérer un plan de sondage plus équilibré fondé sur la stratification de la population en fonction de la variable auxiliaire. Par exemple, considérons à nouveau la population *ppgamma5b*, classée selon ses valeurs x_i. Nous avons construit dix strates où la première strate est composée des unités comptant les cinquante valeurs les plus faibles de x_i, la deuxième strate, des unités groupant les cinquante et unième à centième plus petites valeurs de x_i, et ainsi de suite. Nous avons alors tiré 500 échantillons aléatoires stratifiés de taille cinquante où cinq unités ont été choisies au hasard dans chaque strate. Pour ces exemples, la valeur moyenne de *estpp* était de 43.94 et l'erreur absolue moyenne, de 1.81. L'étendue moyenne de l'estimateur par intervalle correspondant était de 8.95 et la fréquence relative de couverture, de .938. Il faut remarquer que ces chiffres sont très semblables à ceux des tableaux 1 et 2, où nous avons eu recours à un échantillonnage aléatoire simple.

REMERCIEMENTS

Cette recherche a été partiellement supporté par la subvention SES 9201718 de la NSF.

BIBLIOGRAPHIE

BASU, D. (1971). An essay on the logical foundations of survey sampling, part one. Dans *Foundations of Statistical Inference*. Toronto: Holt, Reinhart and Winston, 203-242.

BERGER, J.O. (1985). *Statistical Decision and Bayesian Analysis*. New York: Springer-Verlag.

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

CHAMBERS, R.L., et DUNSTAN, R. (1986). Estimating distribution functions from survey data. *Biometrika*, 73, 597-604.

FERGUSON, T.S. (1973). A Bayesian analysis of some non-parametric problems. *Annals of Statistics*, 1, 209-230.

KUK, A.Y.C., et MAK, T.K. (1989). Median estimation in the presence of auxiliary information. *Journal of the Royal Statistical Society, Series B*, 51, 261-269.

MAK, T.K., et KUK, A. (1993). A new method for estimating finite-population quantiles using auxiliary information. *The Canadian Journal of Statistics*, 21, 29-38.

MEEDEN, G., et GHOSH, M. (1983). Choosing between experiments: applications to finite population sampling. *Annals of Statistics*, 11, 296-305.

MEEDEN, G., et VARDEMAN, S. (1991). A noninformative Bayesian approach to interval estimation in finite population sampling. *Journal of the American Statistical Association*, 86, 972-980.

MEEDEN, G. (1993). Noninformative nonparametric Bayesian estimation of quantiles. *Statistics and Probability Letters*, 16, 103-109.

RAO, J.N.K., KOVAR, J.G., et MANTEL, H.J. (1990). On estimating distribution functions and quantiles from survey data using auxiliary information. *Biometrika*, 77, 365-375.

ROYALL, R.M. (1970). On finite population sampling theory under certain linear regression models. *Biometrika*, 57, 377-387.

ROYALL, R.M., et CUMBERLAND, W.D. (1981). An empirical study of the ratio estimator and estimators of its variance. *Journal of the American Statistical Association*, 76, 66-88.

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

RUBIN, D.B. (1981). The Bayesian bootstrap. *Annals of Statistics*, 9, 130-134.