LES EFFETS DU PLAN D'ÉCHANTILLONNAGE SUR L'ANALYSE DES DONNÉES D'ENQUÊTE

Gad NATHAN

Les méthodes classiques d'inférence sont basées sur la supposition que les observations proviennent d'un échantillon aléatoire simple. Avant de les appliquer aux données recueillies au moyen d'un plan d'échantillonage complexe, il faut déterminer si les méthodes classiques sont valables. On montrera que cela dépend de la définition des paramètres d'intérêt et des modèles de base supposés. Après avoir déterminé qu'il y a des effets du plan d'échantillonage complexe sur l'analyse, il faut établir comment on peut prendre ces effets en considération dans l'analyse des données. Une possibilité est de construire des statistiques qui sont basées sur le plan d'échantillonage complexe. Une autre possibilité est de modifier des statistiques ou des tests d'hypothèses classiques afin qu'ils conviennent au plan d'échantillonage. Cette possibilité facilite l'emploi de programmes informatiques standardisés. Les deux méthodes sont illustrées par leur application aux modèles de régression et à l'analyse des données qualitatives.

Introduction

Quand on obtient des données d'un sondage basé sur un plan d'échantillonage complexe, il existe des méthodes standardisées pour l'estimation ponctuelle de fonctions des variables de la population, telles que des moyennes, des proportions, des totales ou des ratios. En plus, on peut estimer les variances de ces estimateurs et obtenir ainsi des intervalles de confiance, sur la base du théorême de la limite centrale. D'autre part, pour l'analyse des données qui proviennent d'un échantillon aléatoire simple, on peut employer des méthodes standardisées d'analyse, telles que la régression, l'analyse de variance et les modèles logarithmiques linéaires. La grande diffusion des programmes informatiques rend l'utilisation de ces méthodes très facile et, parfois, trop facile. On les emploie souvent pour analyser des données qui proviennent d'un plan

de sondage complexe, sans assurer que le plan de sondage n'a pas d'effet sur l'analyse.

Dans cet exposé on étudiera d'abord dans quelles conditions on peut vraiment employer les méthodes standardisées – des méthodes d'estimation classiques de la théorie de sondage ou des méthodes d'analyse employées pour les échantillons aléatoires simples – pour l'analyse des données d'un sondage à plan complexe. On verra que cela dépend, d'abord, de la définition qui convient du paramètre d'intérêt et des suppositions faites sur le modèle de base. Si on ne peut pas employer des méthodes standardisées, il y a deux approches possibles pour l'analyse. La première essaye de développer les méthodes spécialisées, qui tiennent compte du plan de sondage. Cette approche dépend surtout de la possibilité d'estimer les variances et les covariances des fonctions linéaires des observations, en employant la statistique généralisée de Wald. La deuxième approche est de modifier les méthodes d'analyse existantes, pour les adapter aux plans de sondage complexes. Pour cela on doit, souvent, ne connaître que les effets du plan (DEFF) des combinaisons linéaires des observations.

Les paramètres à estimer

Tout d'abord il faut décider quels sont les paramètres à estimer. Selon les chercheurs qui basent toute leur inférence sur le plan de sondage (par exemple Kish et Frankel, 1974), les seuls paramètres valables pour l'analyse des données sont les paramètres de la population finie. Par exemple, pour étudier les relations entre deux variables, X et Y, selon cette approche, il faut estimer le coefficient de régression de la population, defini par:

$$B = \frac{\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{N}(X_i - \bar{X})^2}. \tag{2.1}$$

Dans ce cas, on n'a pas besoin d'un modèle, et l'inférence est basée seulement sur le plan de sondage, c'est-à-dire sur les propriétés de la distribution d'échantillonage. Néanmoins, il faut préciser que les paramètres de la population finie ont de l'intérêt pour l'analyse seulement s'ils ont un rapport quelconque avec les paramètres d'un modèle. Par exemple, le paramètre B, (2.1), est une "copie" du paramètre β, de la régression linéaire simple. C'est-à-dire B est l'estimateur des moindres carrés pour β dans le modèle de régression simple:

$$\mathcal{E}(Y) = \alpha + \beta X. \tag{2.2}$$

D'autres chercheurs (par exemple Fienberg, 1980) constatent que l'inférence doit être basée sur un modèle de superpopulation qui lie les variables par une distribution de probabilité. La population finie constitue, selon cette approche, un échantillon aléatoire simple de la distribution supposée. Les seuls paramètres d'intérêt sont donc ceux du modèle – c'est-à-dire ceux de la distribution de probabilité. Par exemple, si le modèle est celui de la régression linéaire simple, donnée par (2.2), α et β doivent être les paramètres à estimer.

Si on accepte cette approche et, en plus, si on suppose qu'il y a indépendance entre la distribution du modèle et celle de l'échantillonage, l'inférence classique, basée sur le modèle, est valable et le plan de sondage peut avoir un effet seulement sur l'efficacité de l'inférence. Il faut remarquer que la validité de cette approche dépend de l'existence du modèle supposé et de son indépendance de la distribution de l'échantillonage – des suppositions qui doivent être verifiées. Il faut surtout verifier si l'inférence est robuste quand il y a une déviation de ce modèle.

La meilleure solution à ce dilemme général entre l'inférence basée sur un modèle (de superpopulation) et l'inférence basée sur le plan de sondage est de chercher les modèles qui décrivent vraiment bien la population finie. Dans ce cas les paramètres de la population finie seront assez proches des paramètres correspondants du modèle, parce que la population est un très grand échantillon aléatoire simple de la superpopulation. Par exemple, si le modèle de régression simple (2.2) convient dans la superpopulation infinie, le paramètre B, défini par (2.1), est un estimateur efficace sans biais du paramètre β du modèle (2.2) et en sera assez proche si la population est assez grande.

L'emploi des méthodes standardisées

Pour les deux approches mentionnées ci-dessus – celle basée sur le plan de sondage et celle basée sur un modèle – l'emploi des méthodes standardisées pour l'analyse des données basées sur un plan de sondage complexe est valable seulement dans certaines conditions. Il faut d'abord que le plan de sondage ne dépende pas des valeurs des variables d'interêt. Dans ce cas, si on se base sur le plan de sondage, les méthodes d'estimation de la théorie de sondage classique sont valables pour l'estimation des paramètres de la population finie, même si le plan de sondage est complexe. D'autre part, les méthodes classiques d'analyse, qui sont basées sur la supposition d'un

échantillon aléatoire simple, ne sont pas nécessairement valables, même si le modèle sous-jacent convient.

Par exemple, si le plan de sondage est à probabilités de tirage inégales - disons π_i pour l'unité i - on emploie souvent l'estimateur de Horvitz-Thompson, donné par:

$$\hat{B}_W = \frac{\sum_{i=1}^{n}(x_i - \hat{\bar{X}}_W)(y_i - \hat{\bar{Y}}_W)/\pi_i}{\sum_{i=1}^{n}(x_i - \hat{\bar{X}}_W)^2/\pi_i}, \tag{3.1}$$

où $\hat{\bar{X}}_W$ et $\hat{\bar{Y}}_W$ sont les estimateurs pondérés de Horvitz-Thompson pour $\bar{X}$ et $\bar{Y}$. Cet estimateur est convergent, relativement au plan de sondage, pour B, donné par (2.1), même si le modèle (2.2) ne convient pas. D'autre part, l'estimateur classique des moindres carrées:

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \tag{3.2}$$

est biaisé, selon l'approche du plan de sondage.

Au contraire, selon le modèle (2.2), les deux estimateurs - $\hat{B}_W$, (3.1), et $\hat{\beta}$, (3.2), - sont sans biais. Si on veut employer un estimateur qui est convergent pour les deux approches, il est préférable d'employer $\hat{B}_W$ que $\hat{\beta}$. D'ailleurs, dans le cas homoscedastique, c'est-à-dire si Y a une variance constante, l'estimateur $\hat{\beta}$ est plus efficace que $\hat{B}_W$, selon l'approche du modèle.

Il faut encore préciser qu'on peut employer les logiciels statistiques standardisés, comme SAS, SPSS ou BMDP, pour calculer l'estimateur $\hat{B}_W$, par la régression ponderée. Mais ces logiciels ne calculeront pas correctement la variance de l'estimateur, selon aucune des deux approches.

Un problème plus grave peut se poser, si le plan de sondage dépend de la variable d'interêt. Pour examiner les effets possibles de la sélection sur l'analyse, prenons l'exemple de la régression simple. Si le modèle de régression (2.2) convient dans la population totale, comme par exemple pour les données représentées à la figure 1, chaque échantillon alléatoire simple pris de cette population donnera des estimateurs sans biais, A_1 et B_1 de α et de β, par la méthode des moindres carrées.

En plus, même si on sélectionne un échantillon de valeurs de x, selon un critère quelconque, l'estimateur des moindres carrées sera sans biais. Par

exemple si on sélectionne des unités qui ont des valeurs de x entre deux limites – x_L et x_U – comme représentées dans figure 2, ou même si on sélectionne seulement celles qui sont en dehors de ces limites, comme dans figure 3, les estimateurs des moindres carrées pour α et β, donnés par A_2 et B_2 et par A_3

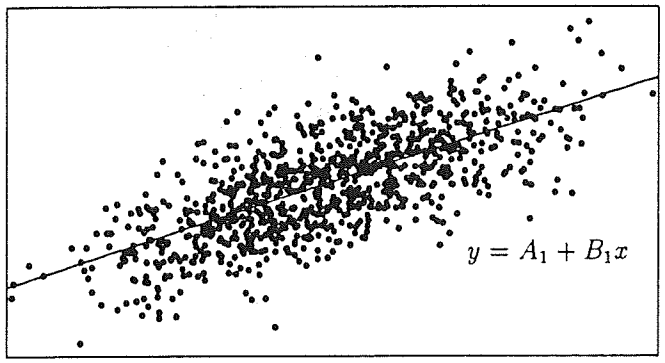

Figure 1: Population totale

et B_3, respectivement, sont sans biais. C'est–à–dire que:

$$\mathcal{E}(A_1) = \mathcal{E}(A_2) = \mathcal{E}(A_3) = \alpha;$$

$$\mathcal{E}(B_1) = \mathcal{E}(B_2) = \mathcal{E}(B_3) = \beta.$$

La même chose est vraie pour les estimateurs basés sur le plan de sondage, comme, par exemple, $\hat{B}_W$, donné par (3.1), si les probabilités de tirage ne dépendent pas de Y. D'ailleurs si on sélectionne les unités à base des valeurs de y, ces mêmes estimateurs seront, en général, biaisés. Par exemple la sélection d'unités qui ont des valeurs de y entre deux limites – y_L et y_U – comme représenté dans figure 4, en dehors de ces limites, comme dans figure 5, ou au dessus–de y_U, comme dans figure 6, rendra les estimateurs à moindres carrées – A_4–A_6 et B_4–B_6 – biaisés. Cest à dire que:

$$\mathcal{E}(A_j) \neq \alpha;\ \mathcal{E}(B_j) \neq \beta;\ (j = 4, 5, 6)$$

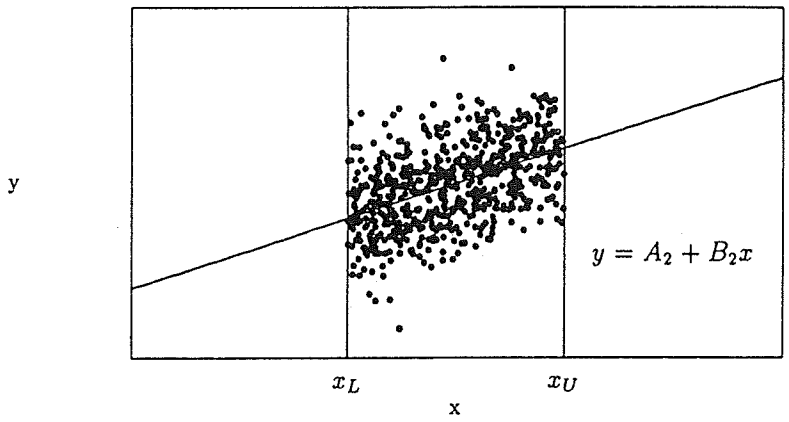

Figure 2: Selection sur $x : x_L < x < x_U$

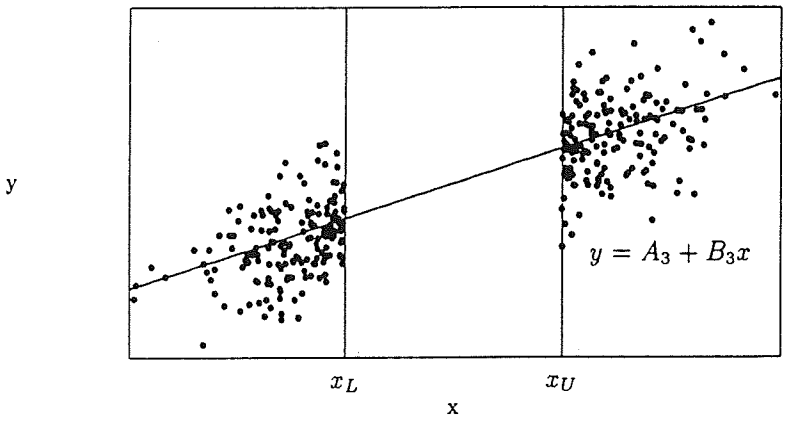

Figure 3: Selection sur $x : x < x_L, x > x_U$

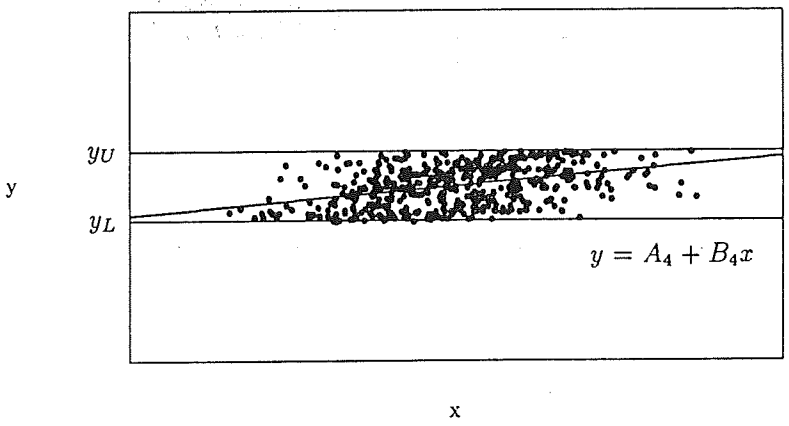

Figure 4: Selection sur $y : y_L < y < y_U$

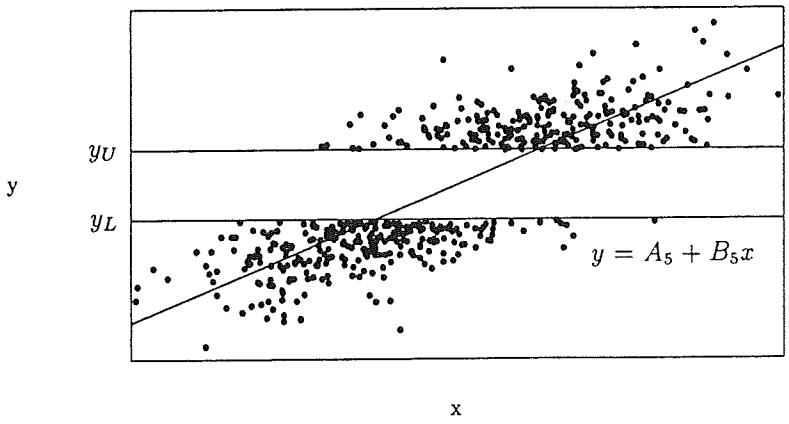

Figure 5: Selection sur $y : y < y_L, y > y_U$

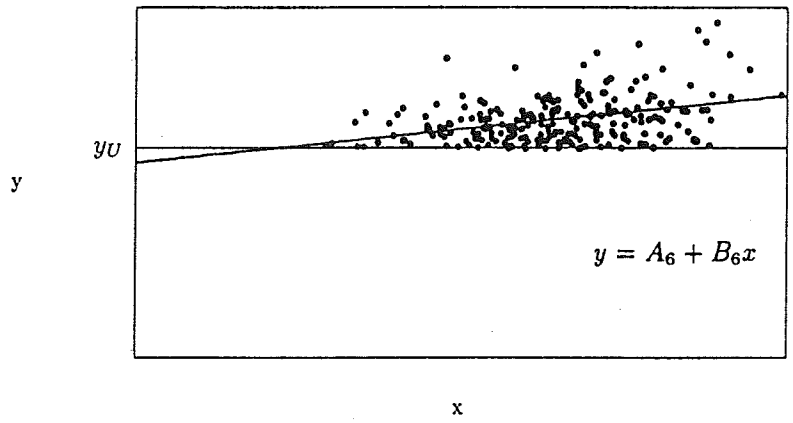

Figure 6: Selection sur y : $y > y_U$

On peut montrer que même si la sélection de l'échantillon est basée sur une autre variable, disons Z, qui n'est pas indépendante de Y, les estimateurs des moindres carrées seront biaisés. Sous certaines conditions pour les relations entre les variables X, Y et Z, Nathan et Holt (1980) ont montré que l'espérance mathématique conditionnelle de l'estimateur $\hat{\beta}$, étant données les valeurs de Z dans toute la population - $\boldsymbol{Z}_U$, est donnée, à l'ordre de $O(n^{-1})$, par:

$$\mathcal{E}(\hat{\beta}|\boldsymbol{Z}_U) = \beta + \frac{\sigma_y}{\sigma_x}\left\{\frac{\rho_{yz\cdot x}\rho_{xz}(1-\rho_{xy}^2)(1-\rho_{xz}^2)(Q-1)}{1+\rho_{xz}^2(Q-1)}\right\}, \qquad (3.3)$$

où σ_x et σ_y sont les écarts types de X et de Y, $\rho_{yz\cdot x}$, ρ_{xz} et ρ_{xy} sont des coefficients de corrélation (partielle ou marginale) du modèle et $Q = \mathcal{E}(s_z^2|\boldsymbol{Z}_U)/\sigma_z^2$ est le ratio entre l'espérance mathématique conditionnelle de la variance de Z dans l'échantillon et celle dans la population. Ainsi on trouve que $\hat{\beta}$ est un estimateur convergent de β si $Q = 1$, c'est-à-dire si ces deux variances sont égales en expectation. Nathan et Holt (1980) proposent des estimateurs modifiés qui sont convergents, même si cette condition ne convient pas.

L'utilisation des statistiques basées sur le plan de sondage

L'hypothèse d'interêt est souvent linéaire, (ou peut être linéarisée), dans les valeurs prévues de statistiques qui ont une distribution normale asymptotique. Si, en plus, on peut obtenir un estimateur convergent de la matrice

de covariances de ces statistiques, la statistique généralisée de Wald (Grizzle, Starmer et Koch, 1969) peut être utilisée pour prendre le plan de sondage en considération. Disons que l'hypothèse à vérifier peut être écrite sous la forme:

$$H_0 : \boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{\theta}_0, \tag{4.1}$$

où $\boldsymbol{X}$ est la matrice de plan connue d'ordre $r \times p$, $\boldsymbol{\beta}$ est un vecteur de paramètres inconnus d'ordre $p \times 1$, et $\boldsymbol{\theta}_0$ est un vecteur de constantes connues d'ordre $r \times 1$. Si l'hypothèse n'est pas linéaire, on peut souvent procéder à une approximation de premier ordre des series de Taylor ayant la forme (4.1).

On suppose, en plus, qu'on a un estimateur convergent, $\widehat{\boldsymbol{\beta}}$, de $\boldsymbol{\beta}$, et un estimateur convergent, $\widehat{\boldsymbol{V}}$, de la matrice de covariances de $\widehat{\boldsymbol{\beta}}$, dont la distribution est indépendante de celle de $\widehat{\boldsymbol{\beta}}$.

On definit donc la statistique généralisée de Wald par:

$$X^2_W = (\boldsymbol{X}\widehat{\boldsymbol{\beta}} - \boldsymbol{\theta}_0)'(\boldsymbol{X}\widehat{\boldsymbol{V}}\boldsymbol{X}')^{-1}(\boldsymbol{X}\widehat{\boldsymbol{\beta}} - \boldsymbol{\theta}_0). \tag{4.2}$$

Sous certaines conditions, cette statistique a une distribution asymptotique de χ^2, sous l'hypothèse nulle, avec un nombre de degrés de liberté correspondant à la dimension de l'hypothèse, c'est-à-dire $p - r$.

Pour un exemple de cette méthode, prenons l'analyse des données qualitatives, d'abord pour la classification de la population en k classes comportant les probabilités, $\boldsymbol{p}' = (p_1, \ldots, p_{k-1})$. Si on veut tester l'hypothèse de la validité de l'ajustement d'une distribution connue: $\boldsymbol{p}'_0 = (p_{0\,1}, \ldots, p_{0\,k-1})$:

$$H_0 : \boldsymbol{p} = \boldsymbol{p}_0, \tag{4.3}$$

on peut suivre les démarches ci-dessus.

On suppose qu'un estimateur convergent $\widehat{\boldsymbol{p}}' = (\widehat{p}_1, \ldots, \widehat{p}_{k-1})$ de $\boldsymbol{p}'$ est obtenu de l'enquête et que cet estimateur est asymptotiquement normal:

$$\sqrt{n}(\widehat{\boldsymbol{p}} - \boldsymbol{p}_0) \longrightarrow N(\mathbf{0}, \boldsymbol{V}). \tag{4.4}$$

Si $\widehat{V}$ est un estimateur convergent de V, alors la statistique généralisée de Wald est:

$$X^2_W = n(\hat{\boldsymbol{p}} - \boldsymbol{p}_o)'\widehat{\boldsymbol{V}}^{-1}(\hat{\boldsymbol{p}} - \boldsymbol{p}_o). \tag{4.5}$$

Cette statistique est distribuée asymptotiquement, sous H_0, par χ^2 avec $k-1$ degrés de liberté et peut être employée pour tester H_0.

Pour tester l'indépendance dans une table de contingence, on peut exprimer l'hypothèse comme:

$$H_0 : h_{ij}(\boldsymbol{p}) = p_{ij} - p_{i+}p_{+j} = 0, (i = 1,\ldots,r-1; j = 1,\ldots,c-1;), \tag{4.6}$$

où p_{ij} est la probabilité de distribution de la population dans la case $(i\,j)$, p_{i+} et p_{+j} sont les probabilités marginales et $\boldsymbol{p}' = (p_{11},\ldots,p_{r-1\,c-1})$.

Si $\widehat{V}_h/n$ est un estimateur convergent de la matrice des covariances des estimateurs convergents $[\boldsymbol{h}(\hat{\boldsymbol{p}})]' = [h_{11}(\hat{\boldsymbol{p}}),\ldots,h_{r-1\,c-1}(\hat{\boldsymbol{p}})]$ de $\boldsymbol{h}(\boldsymbol{p})$, la statistique généralisée de Wald, appliquée au test de H_0, est donnée par:

$$X^2_{WI} = n[\boldsymbol{h}(\hat{\boldsymbol{p}})]'\widehat{\boldsymbol{V}}_h^{-1}[\boldsymbol{h}(\hat{\boldsymbol{p}})]. \tag{4.7}$$

Il faut remarquer que l'utilisation de la statistique généralisée de Wald peut être appliquée selon les deux approches à l'inférence mentionnées dans l'introduction. D'un côté on peut voir les paramètres $\boldsymbol{\beta}$ comme paramètres d'un modèle théorique et definir $\boldsymbol{V}$ comme la matrice des covariances de $\boldsymbol{\beta}$, selon ce modèle. Dans ce cas l'estimation de $\boldsymbol{V}$ peut se faire sans aucun rapport au plan de sondage et les résultats sont valables seulement si le modèle convient vraiment.

Selon l'autre approche - celle qui est basée sur le plan de sondage, on regarde les paramètres $\boldsymbol{\beta}$ comme des paramètres de la population finie dont l'échantillon est tiré. Dans ce cas, la matrice des covariances $\boldsymbol{V}$ doit être celle du plan de sondage et non celle du modèle. Si le plan de sondage est complexe, il y a parfois des difficultés à trouver une méthode convenable pour estimer $\boldsymbol{V}$. C'est surtout le cas si l'estimateur $\hat{\boldsymbol{\beta}}$ n'est pas linéaire dans les données de l'échantillon. Même s'il existe maintenant, dans ce cas,

des méthodes variées pour estimer des variances comme la linéarisation, la répétition équilibrée et les méthode de *Jacknife* et de *Bootstrap*, il faut se rendre compte qu'on a, en général, un grande nombre de paramétres à estimer et qu'on doit assurer la stabilité de l'estimateur $\widehat{V}$, afin que l'inversion de $\widehat{V}$ puisse se faire sans difficultés.

La modification des statistiques standardisées

Pour éviter les problèmes que pose l'estimation instable des variances, mentionnés ci-dessus, il est souvent préférable de chercher les possibilités de modifier les statistiques standardisées, afin qu'elles conviennent au plan de sondage complexe. Cette possibilité a, en outre, l'avantage de faciliter l'emploi des logiciels statistiques. Celles-ci ne donnent que des statistiques standardisées, qui ne conviennent pas au plan de sondage complexe. Même si cette modification doit se baser aussi sur l'estimation des variances, la sensitivité de la modification à la stabilité des estimateurs est beaucoup plus basse que dans le cas où on doit inverser la matrice des covariances.

Nous suivons cette possibilité par son application aux problèmes d'analyse des données qualitatives de la section précédente. Par exemple, pour tester l'hypothèse de la validité de l'ajustement (4.3), dans le cas de sondage aléatoire simple, la statistique:

$$X_0^2 = n\sum_{i=1}^{k}(\hat{p}_i - p_{0i})^2/p_{0i} = n(\hat{\boldsymbol{p}} - \boldsymbol{p}_0)'\boldsymbol{P}_0^{-1}(\hat{\boldsymbol{p}} - \boldsymbol{p}_0), \tag{5.1}$$

où $\boldsymbol{P}_0 = \mathrm{diag}(\boldsymbol{p}_0) - \boldsymbol{p}_0\boldsymbol{p}_0'$, est la statistique standardisée généralement employée. Rao et Scott (1981) ont montré que pour un plan de sondage complexe cette statistique a une distribution qui est obtenue comme une somme ponderée de $k-1$ variables indépendantes, dont chacune a une distribution de χ^2 avec un degré de liberté:

$$X_0^2 \approx \sum_{i=1}^{k-1}\lambda_i Z_i^2, \tag{5.2}$$

où $Z_i \sim N(0,1)$ sont des variables normales standardisées indépendantes et $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_{k-1}$ sont les valeurs propres de la matrice $\boldsymbol{D} = \boldsymbol{P}_0^{-1}\boldsymbol{V}$.

On peut démontrer que la valeur propre la plus grande, λ_1, égale l'effet du plan (DEFF) maximal des combinaisons linéaires des valeurs de p_i:

$$\lambda_1 = \sup_{\boldsymbol{l}}[V_P(\boldsymbol{l}'\hat{\boldsymbol{p}})/V_{srs}(\boldsymbol{l}'\hat{\boldsymbol{p}})], \tag{5.3}$$

où $\boldsymbol{l}' = (l_1, \ldots, l_{k-1})$ est un vecteur quelconque de coefficients, V_P dénote la variance selon le plan de sondage complexe et V_{srs} celle pour l'échantillonage aléatoire simple.

Ainsi, pour le sondage stratifié représentatif, étant donnée que $\lambda_1 \leq 1$, l'emploi de X_0^2 donne un test prudent de l'hypothèse (4.3). Mais pour le sondage par grappes à plusieurs degrés, Holt, Scott et Ewings (1980) ont démontré, par des études empiriques, que l'emploi de X_0^2 peut produire un test qui a, en réalité, un niveau de signification beaucoup plus élevé que le niveau nominal. En général, pour utiliser le résultat (5.2) il faut évaluer les valeurs de λ_i. Ces valeurs peuvent être estimées par des estimateurs convergents $\hat{\lambda}_i$, qui sont les valeurs propres de la matrice $\widehat{\boldsymbol{D}} = \widehat{\boldsymbol{P}}^{-1}\widehat{\boldsymbol{V}}$, où $\widehat{\boldsymbol{P}} = \mathrm{diag}(\hat{\boldsymbol{p}}) - \hat{\boldsymbol{p}}\hat{\boldsymbol{p}}'$ et $\widehat{\boldsymbol{V}}$ est un estimateur convergent de $\boldsymbol{V}$. Mais si on avait un estimateur convergent de $\boldsymbol{V}$ convenable, on pourrait l'employer directement pour évaluer la statistique X_W^2. Dans d'autres cas, où l'estimation de $\boldsymbol{V}$ n'est pas stable, on peut souvent obtenir des estimateurs des effets du plan (DEFF) pour les cases individuelles: $\hat{d}_i = \hat{V}_{ii}/[p_i(1-p_i)]$. Rao et Scott (1981) ont démontré que la moyenne pondérée de ces effets du plan:

$$\hat{\lambda} = tr(\widehat{\boldsymbol{D}})/(k-1) = \sum_{i=1}^{k-1}(1-p_i)\hat{d}_i/(k-1) \tag{5.4}$$

est un estimateur convergent de $\mathcal{E}(X_0^2)/(k-1)$. Sur cette base ils ont proposé d'utiliser la statistique modifiée $X_0^2/\hat{\lambda}$, qui est approximativement distribuée comme χ^2 avec $k-1$ degrés de liberté, si l'hypothèse (4.3) convient. Des résultats empiriques de Hidiroglou et Rao (1987) pour des enquêtes de santé canadiennes et de Holt, Scott et Ewings (1980) pour des enquêtes britanniques démontrent qu'une approximation satisfaisante pour les niveaux de signification est obtenue par l'emploi de la statistique modifiée $X_0^2/\hat{\lambda}$.

Des méthodes modifiées peuvent être utilisées pour l'analyse des données à plusieurs variables qualitatives. Par exemple pour tester l'indépendance dans une table de contingence - c'est-à-dire l'hypothèse (4.6). Pour le sondage aléatoire simple, la statistique standardisée est:

$$X_I^2 = n\sum_{i=1}^{r}\sum_{j=1}^{c}(\hat{p}_{ij} - \hat{p}_{i+}\hat{p}_{+j})^2/(\hat{p}_{i+}\hat{p}_{+j})), \tag{5.5}$$

qui a une distribution asymptotique de χ^2 avec $(r-1)(c-1)$ degrés de liberté, sous l'hypothese (4.6). Quand le plan de sondage est complexe, Rao et Scott (1981) ont démontré que X_I^2 a encore une fois une distribution asymptotique qui est la somme ponderée de $(r-1)(c-1)$ variables, dont chacune a une distribution de χ^2 avec un degré de liberté:

$$X_I^2 \approx \sum_{i=1}^{r-1}\sum_{i=1}^{c-1} \delta_{ij} Z_{ij}^2, \tag{5.6}$$

où $Z_{ij} \sim N(0,1)$ sont des variables normales standardisées indépendantes et δ_{ij} sont les valeurs propres de la matrice $\boldsymbol{D}_h = (\boldsymbol{P}_r^{-1} \otimes \boldsymbol{P}_c^{-1})\boldsymbol{V}_h$, $\boldsymbol{P}_r = \text{diag } (\boldsymbol{p}_r) - \boldsymbol{p}_r\boldsymbol{p}_r'$, $\boldsymbol{P}_c = \text{diag } (\boldsymbol{p}_c) - \boldsymbol{p}_c\boldsymbol{p}_c'$, $\boldsymbol{p}_r = (p_{1+}, \ldots, p_{r-1+})$, $\boldsymbol{p}_c = (p_{+1}, \ldots, p_{+c-1})$ et $\boldsymbol{V}_h$ est la matrice des covariances des fonctions $h_{ij}(\widehat{\boldsymbol{p}})$. Les valeurs de δ_{ij} peuvent être considerées comme des effets du plan généralisés des statistiques $h_{ij}(\widehat{\boldsymbol{p}})$. L'approximation pour une distribution de χ^2 avec $(r-1)(c-1)$ degrés de liberté est obtenue pour la statistique modifiée $X_I^2/\hat{\delta}$, où $\hat{\delta}$ est la moyenne ponderée des effets du plan estimés pour $h_{ij}(\widehat{\boldsymbol{p}})$:

$$\hat{\delta} = \sum_{i=1}^{r}\sum_{j=1}^{c}(1-\hat{p}_{i+})(1-\hat{p}_{+j})\hat{\delta}_{ij}/[(r-1)(c-1)], \tag{5.7}$$

$$\hat{\delta}_{ij} = \hat{V}_{ij}/[\hat{p}_{i+}(1-\hat{p}_{i+})\hat{p}_{+j}(1-\hat{p}_{+j})], \tag{5.8}$$

où $\hat{V}_{ij}$ est un estimateur convergent de la variance de $h_{ij}(\widehat{\boldsymbol{p}})$. Encore une fois, pour evaluer $\hat{\delta}_{ij}$ directement selon (5.8) il faut estimer V_{ij}, qui est parfois difficile. Rao et Scott (1984) ont démontré qu'on peut évaluer $\hat{\delta}_{ij}$ en employant les valeurs estimées des effets du plan, $\hat{d}_{ij}$, des probabilités estimées des cases, $\hat{p}_{ij}$, et les effets de plan estimés, $\hat{d}_i(r)$ et $\hat{d}_j(c)$, des probabilités marginales estimées, $\hat{p}_{i+}$ et $\hat{p}_{+j}$:

$$\hat{\delta} = \frac{\sum_{i,j}(1-\hat{p}_{i+})(1-\hat{p}_{+j})\hat{d}_{ij} - \sum_i(1-\hat{p}_{i+})\hat{d}_i(r) - \sum_j(1-\hat{p}_{+j})\hat{d}_j(c)}{(r-1)(c-1)}. \tag{5.9}$$

Ces modifications ont été généralisées pour les tables de contingence à classification multiple pour tester des modèles log-linéaires directs par Rao et Scott (1984). Dans ce cas ils démontrent qu'on peut employer un estimateur

d'effet du plan moyen, $\hat{\delta}$, qui est une fonction des effets du plan estimés des probabilités des cases et de celles des probabilités marginales.

Conclusion

Les quelques exemples donnés ci-dessus montrent qu'il y a une grande diversité dans les méthodes et les possibilités de traiter l'analyse des données recueillies au moyen d'un plan de sondage complexe. En réalité la diversité dans ce domaine est encore plus grande - comme on peut le voir en lisant l'ouvrage rédigé par Skinner, Holt et Smith (1989), qui résume des dizaines de travaux de recherches théoriques et empiriques concernant ce sujet. On doit encore préciser que ce domaine est relativement neuf dans l'étude des méthodes de sondage - pas plus de vingt ans ont passé depuis la publication des premiers travaux. L'emploi des modèles dans le domaine de la théorie de sondage est aussi relativement récent, et a sans doute contribué aux progrès réalisés en étudiant les effets de plan de sondage sur l'analyse. Ces progrès ont produit un rapprochement bénéfique entre la théorie classique de statistique, basée tout à fait sur les suppositions de modèles, et la théorie de sondage classique, qui ignorait, jusqu'à récemment, les modèles.

Malgré ce progrès, il faut noter que les statisticiens officiels n'ont pas encore pénétré serieusement ce domaine et que, pour des raisons diverses, ils produisent surtout des statistiques énumératives et laissent l'analyse aux autres agences et aux chercheurs substantifs. Ces derniers n'ont pas, en général, la formation statistique nécessaire et ils employent le plus souvent des logiciels statistiques standardisés, sans toujours comprendre leurs limitations et nuances d'emploi ou savoir s'ils conviennent vraiment aux conditions d'analyse. C'est surtout le cas concernant l'analyse de données qui sont recueillies au moyen d'un plan de sondage complexe. Comme on l'a vu ci-dessus, il est dangereux d'analyser ces données sans prendre en considération que le plan de sondage peut influencer l'analyse .

Même si la solution la meilleure est de développer des statistiques et des tests d'hypothèses adaptés au plan de sondage, comme mentionné dans la section 4, cette approche est difficile à réaliser en pratique. C'est surtout vrai si l'analyse secondaire doit être faite par des chercheurs en dehors de l'organisation qui est responsable de la collecte des données. Dans ce cas, ces chercheurs n'ont pas toujours accès aux détails du plan de sondage, qui sont nécesssaires pour employer les mèthodes de la section 4.

Pour ces raisons et à cause de l'emploi général des logiciels standardisés, il apparaît que la voie décrite dans la section 5 — celle de la modification

des statistiques standardisées – sera la plus appropriée dans la plupart des cas. D'ailleurs ces modifications, ont aussi besoin d'informations certaines sur les effets du plan de sondage. Les statisticiens d'enquêtes - officiels et autres - doivent s'organiser pour assurer que ces données seront disponibles aux chercheurs. D'autre part les statisticiens doivent développer des méthodes générales pour la modification des statistiques standardisées. Ces modifications doivent être introduites dans les logiciels statistiques standardisés.

Bibliographie

Fienberg, S. E. (1980). The measurement of crime victimization: prospects for panel analysis of a panel survey. *The Statistician* **29**, 313-350.

Grizzle, J. E., Starmer, C. F. et Koch, G. G. (1969). Analysis of categorical data by linear models. *Biometrics* **25**, 489-504.

Hidiroglou, M. A. et Rao , J. N. K. (1987). Chi-squared tests with categorical data from complex surveys: Part I – simple goodness–of–fit, homogeneity and independence in a two–way table with applications to the Canada Health Survey (1978–1979). *Journal of Official Statistics* **3**, 117-132.

Holt, D., Scott, A. J. et Ewings, P. D. (1980). Chi–squared tests with survey data. *Journal of the Royal Statistical Society A* **143**, 303-320.

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

Nathan, G. et Holt, D. (1980). The effect of survey design on regression analysis. *Journal of the Royal Statistical Society B* **42**, 377-386.

Rao, J. N. K. et Scott, A. J. (1981). The analysis of categorical data from complex surveys: chi-squared tests for goodness of fit and independence in two-way tables. *Journal of the American Statistical Association* **76**, 221-230.

Rao, J. N. K. et Scott, A. J. (1984). On chi-square tests for multiway contingency tables with cell proportions estimated from survey data. *Annals of Statistics* **12**, 46-60.

Skinner, C. J., Holt, D. et Smith, T. M. F. (1989). (éds.) *Analysis of complex surveys.* Chichester: Wiley.