INFORMATISATION DU CALCUL D'ESTIMATIONS POUR LES ENQUÊTES COMPLEXES[1]

M.A. Hidiroglou[2]

Très souvent, les organismes statistiques chargés de recueillir des données d'enquête s'acquittent aussi de toutes les étapes du traitement jusqu'à la mise en tableau des résultats. En outre, les programmes informatiques appliqués aux observations sont, dans la plupart des cas, adaptés au plan d'enquête. Les résultats de ces calculs peuvent varier de paramètres descriptifs simples, comme un total ou une moyenne, aux paramètres plus complexes nécessaires au études analytiques telles que la comparaison de domaines, l'analyse de régression et l'analyse de tableaux de contingence. Cet article décrit un programme informatique qui calcule ces statistiques et les erreurs d'échantillonnage correspondantes pour certains plan d'échantillonnage courants.

1. INTRODUCTION

Un grand nombre de statistiques sont calculées à partir de données qui, dans bien des cas, sont recueillies dans de grandes enquêtes complexes à l'échelle nationale ou régionale. Les résultats de ces calculs peuvent varier de paramètres descriptifs simples, comme un total ou une moyenne, aux paramètres plus complexes nécessaires aux études analytiques telles que la comparaison de domaines, l'analyse de régression et l'analyse de tableaux de contingence. L'estimation par domaine concerne la production de statistiques sur des sous-groupes de la population observée, qui ne sont pas explicitement prévus dans le plan de sondage. Yates (1960) offre une mine de renseignements sur l'estimation des moyennes et des écarts entre moyennes au niveau des domaines. Hartley (1959) et Rao (1975) donnent une excellente description des

---

[1] Exposé présenté à l'assemblée annuelle de l'American Statistical Association, Détroit, août 1981.

[2] M.A. Hidiroglou Division des méthodes d'enquêtes "Entreprises", Statistique Canada

méthodes utilisées pour l'estimation par domaine. Les estimateurs de la variance des estimations obtenues pour les domaines ne sont que des modifications directes des estimateurs de la variance appliqués aux variables statistiques simples. Cela n'est toutefois pas le cas des variables très complexes. L'estimation d'équations de régression basées sur des données d'enquêtes pose plusieurs problèmes. En effet, il faut formuler les équations de régression, définir la population sur laquelle les inférences porteront et estimer la variance des coefficients de l'équation (voir Konijn (1962), Kish et Frankel (1974) et Fuller (1975)). Les tests d'hypothèses appliqués aux tableaux de contingence en tenant compte du plan de sondage ont été analysés par Nathan (1969, 1972), Rao et Scott (1981), Fellegi (1980), Garza-Hernandez et McCarthy (1962) et par Koch, Freeman et Freeman (1975), pour ne nommer que ces auteurs-là.

Très souvent, les organismes statistiques chargés de recueillir des données d'enquête s'acquittent aussi de toutes les étapes du traitement jusqu'à la mise en tableau des résultats. En outre, les programmes informatiques appliqués aux observations sont, dans la plupart des cas, adaptés au plan d'enquête. Il est fort possible que chaque fois qu'un nouveau plan de sondage est appliqué, il faille mettre au point des programmes tout à fait nouveaux pour, par exemple, estimer des totaux et leur variance. Ce travail prend beaucoup de temps, coûte très cher et il est fastidieux, parce que, dans une certaine mesure, répétitif. Une autre possibilité consiste à utiliser un ensemble de logiciels statistiques comme le SPSS ou le SAS. Ces programmes informatiques permettent d'obtenir facilement des estimations pondérées. Toutefois, les variances calculées par ces procédés ne prennent pas en considération des facteurs liés au plan de sondage, comme la stratification ou la division de l'échantillon en grappes, à moins que les instructions nécessaires ne soient ajoutées aux programmes. L'utilisateur doit donc connaître assez bien le langage de ces méthodes s'il veut obtenir de bonnes estimations de la variance des paramètres d'un sondage.

Des travaux ont été effectués récemment dans le but d'élaborer des programmes qui calculent des variances pour une catégorie générale de plans d'enquêtes. Parmi ces programmes, on trouve, entre autres, STDEER de Shah (1974), SURREGR de Holt (1975), ainsi que SUPER CARP et MINI CARP de Hidiroglou, Fuller et Hickman (1980). Essentiellement, ces programmes exigent que l'utilisateur indique l'estimateur qu'il veut utiliser et les variables qu'il doit analyser. Nous supposerons ici qu'il y a vérification des ensembles de

données auxquels les programmes sont appliqués et imputation des données qui manquent. Dans les pages qui suivent, nous présentons une description de SUPER CARP et de MINI CARP. À l'aide de SUPER CARP, on peut calculer des estimations de totaux, des estimations par quotient, la différence entre des estimations obtenues par la méthode des quotients et construire des tests applicables à des tableaux de contingence pour des échantillons stratifiés à plusieurs degrés. SUPER CARP offre quelques techniques de régression qui conviennent à des observations touchées par des erreurs de réponse (de mesure). L'utilisateur peut estimer des matrices des covariances pour les moyennes et les totaux de sous-populations ou de strates. MINI CARP est un programme plus petit qui diffère de SUPER CARP en ce sens qu'il ne contient pas les techniques de régression de SUPER CARP. Une comparaison des applications de ces deux programmes est présentée au tableau 1.

TABLEAU 1. Applications de SUPER CARP (S) et de MINI CARP (M)

Estimation à plusieurs variables de	Pour		
	Toute la population	Des strates individuelles	Des sous-populations
Paramètres simples			
. Moyennes	S,M	S,M	S,M
. Totaux	S,M	S,M	S,M
. Quotients	S,M	S,M	S,M
. Différences entre quotients	S,M		
. Proportions	S,M	S,M	S,M

Paramètres complexes		Tests	
. Moindres carrés pondérés	S	. Coefficients de régression	S
. Erreurs pondérées sur les variables (covariances connues et estimées des erreurs)	S	. Validité de l'ajustement	S,M
		. Indépendance, tableau à deux dimensions	S,M

2. DESCRIPTION GÉNÉRALE

2.1 Notation

En général, SUPER CARP et MINI CARP peuvent être utilisés avec des données obtenues à l'aide d'un plan stratifié à plusieurs degrés. Si le plan de sondage comporte s degrés d'échantillonnage, un vecteur de degrées à g dimensions est mis en mémoire pour chaque observation. On peut représenter ce vecteur de données de la façon suivante:

$$(Z_{hi_{\sim s}1}, Z_{hi_{\sim s}2}, \ldots, Z_{hi_{\sim s}g}),$$

où h = 1, 2, ..., L représente les strates, $i_s = (i_1, i_2, \ldots, i_s)$ indique le degré d'échantillonnage, i_1, = 1, 2, ..., n_h correspond aux unités du premier degré d'échantillonnage, i_2 = 1, 2, ..., n_{hi1} correspond aux unités du deuxième degré d'échantillonnage, ..., et i_s = 1, 2, ..., n_{hi} désigne les unités du dernier degré d'échantillonnage, s. $Z_{hi\ k}$ représente la hi_sième observation pour la kième variable d'intérêt. Un poids attribué à la hi_sième observation sera désigné par le terme w_{hi} . Ces poids sont inversement proportionnels à la probabilité que chaque unité finale d'échantillonnage soit sélectionnée. On énumère les variables qui serviront à l'analyse (que ce soit un total, une estimation par quotient ou par régression) au moyen d'un vecteur de sélection $v = (v_1, v_2, \ldots, v_{p+1})$, où $1 \leq v_k \leq g$ pour k = 1, 2, ..., p+1. Supposons que le type d'analyse et les variables requises ont été choisis par l'utilisateur, et disons que le vecteur de sélection de la hi_sième observation est le suivant:

$$(Y_{hi_{\sim s}}, X_{hi_{\sim s}1}, X_{hi_{\sim s}2}, \ldots, X_{hi_{\sim s}g}),$$

où Y représente la variable dépendante et X, les variables indépendantes s'il s'agit d'une analyse de régression. Il est à noter que v_1 correspond toujours à la variable dépendante dans le cas d'une régression. Quant aux autres types d'analyse, l'ordre à l'intérieur du vecteur de sélection n'a pas d'importance.

2.2 Types de calculs

Cette section présente un aperçu des variables statistiques simples définies dans les programmes et donne une liste partielle des options de régression offertes. Une description complète de toutes les options est donnée dans le guide de SUPER CARP ou de MINI CARP (1980).

i) Estimateur de totaux

L'estimateur de totaux se formule comme suit:

$$\hat{X}_{(k)} = \sum_{h} \sum_{i_1} \cdots \sum_{i_s} w_{hi_{\sim s}} X_{hi_{\sim s}(k)}, \quad k = 1, 1, 2, \ldots, p.$$

La matrice des covariances estimées de

$$\hat{\underset{\sim}{X}} = \{\hat{X}_{(1)}, \hat{X}_{(2)}, \ldots, \hat{X}_{(p)}\}$$

est

$$v_1(\hat{X}) = \sum_{h=1}^{L} (n_h - 1)^{-1} n_h(1-f_h) \sum_{i_1=1}^{n_h} (\hat{d}_{hi_{\sim 1}.} - \bar{\hat{d}}_{h..})^T (\hat{d}_{hi_{\sim 1}.} - \bar{\hat{d}}_{h..}) \quad (2.2.1)$$

où

$$\hat{d}_{hi_{\sim 1}.} = \{\hat{d}_{hi_1(1)}, \hat{d}_{hi_1(2)}, \ldots, \hat{d}_{hi_1(p)}\}$$

$$\hat{d}_{hi_1(k)} = \sum_{i_2=1}^{n_{hi_1}} \cdots \sum_{i_s=1}^{n_{hi_{s-1}}} w_{hi_s} X_{hi_s(k)}$$

$$\hat{\bar{d}}_{h..} = n_h^{-1} \sum_{i_1=1}^{n_h} \hat{d}_{hi_1}.$$

Il est à noter que la formule de variance ci-dessus peut être appliquée à des plans d'échantillonnage avec probabilité proportionnelle à la taille (ppt) avec ou sans remise. Dans le cas de l'échantillonnage avec remise, il faut calculer seulement la variance au premier degré (Des Raj, 1968, p. 120) et les facteurs de correction, f_h, sont fixés à zéro. Dans les enquêtes à grande échelle, il est souvent supposé que les grappes du premier degré d'échantillonnage ont été sélectionnées avec remise, même si, en réalité, l'échantillonnage était fait sans remise. Un corollaire de cette hypothèse et du fait que les pas de sondage sont petits est que la variance ainsi calculée se rapproche beaucoup de celle qu'on aurait obtenue en tenant compte de tous les degrés d'échantillonnage et de toutes les techniques de sélection. Si les pas de sondage sont assez importants à chaque degré d'échantillonnage et si l'échantillon a été prélevé par échantillonnage aléatoire simple sans remise à chaque degré, on peut tirer avantage de la règle de Des Raj (1966) pour calculer la composante de la variance à chaque degré d'échantillonnage. La matrice des covariances pour un échantillon formé de s degrés s'écrit:

$$v(\hat{\underset{\sim}{X}}) = \sum_{r=1}^{s} v_r(\hat{\underset{\sim}{X}})$$

où pour $r \geq 2$

$$v_r(\hat{\underset{\sim}{X}}) = \sum_{h=1}^{L} \sum_{i_1=1}^{n_h} \cdots \sum_{i_{r-1}=1}^{n_{hi_{r-2}}} \left[\prod_{j=0}^{r-2} \frac{n_{hi_j}}{N_{hi_j}} \right] \tag{2.2.2}$$

$$\times\ n_{hi_{r-1}} (n_{hi_{r-1}} - 1)^{-1} (1-f_{hi_{r-1}})$$

$$\times \sum_{i_r} (\hat{d}_{hi_r}(.) - \hat{\tilde{d}}_{hi_{r-1},.}(.))^T$$

$$\times\ \hat{d}_{hi_r}(.) - \hat{\tilde{d}}_{hi_{r-1},.}(.)$$

où

$$f_{hi_{r-1}} = n_{hi_{r-1}} N_{hi_{r-1}}^{-1},\ n_{hi_o} = n_h,\ N_{hi_o} = N_h,$$

$$\hat{d}_{hi_r}(.) = \hat{d}_{hi_r}(1), \hat{d}_{hi_r}(2), \ldots, \hat{d}_{hi_r}(p)$$

$$\hat{d}_{hi_r}(k) = \sum_{i_{r+1}} \cdots \sum_{i_s} w_{hi_s} x_{hi_s}(k)$$

$$\hat{\tilde{d}}_{hi_{r-1},.}(.) = n_{hi_r}^{-1} \sum_{i_r} \hat{d}_{hi_r}(.) .$$

On peut donc estimer la variance pour un plan de sondage à r degrés en calculant les composantes qui correspondent à chaque degré d'échantillonnage ($v_r(\hat{\underset{\sim}{X}})$) et en en faisant l'addition. On peut faire ce calcul en passant r fois en machine l'ensemble de données. La première fois, le programme repère les strates et les unités du premier degré d'échantillonnage pour obtenir $v_1(\hat{\underset{\sim}{X}})$. La deuxième fois, le programme enregistre les unités

primaires d'échantillonnage initiales comme des "strates" et les unités secondaires comme des grappes, ce qui donne $v_2(\hat{\underset{\sim}{X}})$. Enfin, lorsque les ($r \geq 2$) données sont lues la rième fois, les unités initialement du (r-1)ième degré d'échantillonnage sont considérées comme des "strates" et les unités du rième degré comme des grappes, pour calculer $v_r(\hat{\underset{\sim}{X}})$.

À chaque passage en machine des données, une fraction d'échantillonnage $g_{hi_{r-1}}$ est enregistrée pour l'unité hi_{r-1}, où

$$g_{hi_{r-1}} = 1 - \left[\prod_{j=0}^{r-2} \frac{n_{hi_j}}{N_{hi_j}}\right] \left[1 - \frac{n_{hi_{r-1}}}{N_{hi_{r-1}}}\right] .$$

Par cette méthode, le programme calcule $v_r(\hat{\underset{\sim}{X}})$ selon la structure donnée par $v_1(\hat{\underset{\sim}{X}})$.

Si les fractions de sondage ne sont pas négligeables à chaque degré et si l'enquête est basée sur un plan d'échantillonnage sans remise avec probabilité proportionnelle à la taille à tous les degrés, il faut prendre en considération la probabilité composée de sélection dans l'estimation de la variance à chaque degré. SUPER CARP et MINI CARP ne calculent pas les probabilités composées de sélection. Lorsque deux unités par strate ont été sélectionnées sans remise avec des probabilités inégales, on peut obtenir la variance de l'estimateur du total à l'aide de l'équation (2.2.1) en incluant, pour chaque strate, un facteur de correction qui contient la probabilité composée de sélection. Ce facteur de correction est le résultat de l'expression suivante:

$$f_h = \frac{2\pi_{h12} - \pi_{h1}\ \pi_{h2}}{\pi_{h12}} \quad , h=1, 2, \ldots, L$$

où π_{h12} représente la probabilité composée de sélection des unités choisies 1 et 2. Lorsque $n_h \geq 2$ et les probabilités composées de sélection ne sont pas connues, il est possible d'utiliser l'approximation de la variance pour l'échantillonnage sans remise formulée par Gray (1975). Gray démontre que les variances d'un échantillon constitué sans remise et avec des probabilités inégales de sélection peuvent être décomposées en une variance "avec remise" multipliée par un facteur de correction fini pour la population, ce facteur étant déterminé par les probabilités composées. Le facteur de correction a été évalué comme à peu près égal à 1 moins l'inverse de la fraction de sondage pour les populations composées de plus de 15 éléments à chaque degré. Au moyen de l'approximation de Gray, on peut calculer les variances pour les plans d'échantillonnage sans remise à plusieurs degrés avec des probabilités inégales de sélection.

S'il faut faire de l'estimation par domaine pour quelques variables, on définit une nouvelle variable, ${}_dY_{hi_{\sim s}}(k)$, pour tous les éléments de la population, où

$$ {}_dY_{hi_{\sim s}}(k) = \begin{cases} Y_{hi_{\sim s}}(k) \text{ si le } hi_{\sim s}\text{ième élément appartient au domaine} \\ \qquad d \text{ (disons } D_d) \\ 0 \text{ autrement} \end{cases} $$

Une autre façon de définir ${}_dY_{hi_{\sim s}}(k)$ consiste à appliquer la formule ${}_dY_{hi_{\sim s}}(k) = {}_da_{hi_{\sim s}}\ Y_{hi_{\sim s}}(k)$, où

$$ {}_da_{hi} = \begin{cases} 1 \text{ si le } hi_s\text{ième élément appartient à } D_d \\ 0 \text{ autrement} \end{cases} $$

Notons que, si $\hat{Y}$ et $v(\hat{\underset{\sim}{Y}})$ sont des estimateurs sans biais pour Y et $V(\hat{\underset{\sim}{Y}})$ respectivement, les résultats de l'estimation par domaine, ${}_d\hat{\underset{\sim}{Y}}$ et $v({}_d\hat{\underset{\sim}{Y}})$, sont

également des estimations sans biais pour $_d\hat{\underset{\sim}{Y}}$ et $V(_d\hat{\underset{\sim}{Y}})$. Il devient alors possible d'appliquer les formules classiques pour $\hat{\underset{\sim}{Y}}$ et $v(\hat{\underset{\sim}{Y}})$ aux variables "synthétiques" $_dY_{hi\underset{\sim}{s}}$. On peut calculer des totaux pour chaque strate en considérant les strates comme des variables de classification.

ii) Estimation par quotient

Le vecteur $\{ Y_{hi\underset{\sim}{s}}(1), X_{hi\underset{\sim}{s}}(1), \ldots, Y_{hi\underset{\sim}{s}}(p), X_{hi\underset{\sim}{s}}(p) \}$ est utilisé dans l'analyse et les rapports estimés sont les suivants:

$$\hat{R}(t) = \hat{X}^{-1}_{(t)} \hat{Y}_{(t)}, \quad t = 1, 2, \ldots, p ;$$

où $Y_{(1)}$ et $X_{(t)}$ ont la forme présentée dans la section précédente. La matrice des covariances estimées de $R = R(1), R(2), \ldots, R(p)$ est la même que la matrice décrite dans la section précédente avec

$$\hat{d}_{hi\underset{\sim}{r}}(t) = \hat{X}^{-1}_{(t)} \sum_{i_{r+1}} \cdots \sum_{i_s} w_{hi\underset{\sim}{s}} \{Y_{hi\underset{\sim}{s}}(1) - \hat{R}(t)\, X_{hi\underset{\sim}{s}}(t)\}; \quad t = 1, \ldots, p.$$

On peut appliquer l'estimation par quotient au calcul de la moyenne de chaque variable d'intérêt en fixant toutes les variables X égales à 1. Il est possible de calculer la moyenne pour chaque domaine en substituant $_dY_{hi\underset{\sim}{s}}(t)$ à $Y_{hi\underset{\sim}{s}}(t)$ et $_da_{hi\underset{\sim}{s}}$ à $X_{hi\underset{\sim}{s}}(t)$. Lorsqu'il faut obtenir les proportions de Y pour une sous-population et pour un domaine D_d, le numérateur du quotient est la somme pondérée des $_dY_{hi\underset{\sim}{s}}(t)$ et le dénominateur est la somme pondérée des $Y_{hi\underset{\sim}{s}}(t)$. Le rapport estimé entre deux variables définies dans un domaine D_d peut être évalué de la même façon. On peut également calculer des proportions et des rapports pour des strates qui représentent des variables de classification.

iii) Estimation par régression

On a accordé récemment beaucoup d'attention aux notions de régression qui s'appliquent aux enquêtes par sondage. Cet intérêt est attribuable à plusieurs facteurs. D'abord, on met davantage l'accent sur les enquêtes analytiques, bien que certaines questions concernant les meilleures méthodes de pondération des observations ne soient pas encore résolues. Deuxièmement, la construction de modèles en général, surtout ceux fondés sur les méthodes de régression, a suscité beaucoup d'intérêt, ainsi que des critiques, quant aux possibilités qu'elle présente pour le calcul des estimations d'enquête. SUPER CARP attribue le poids approprié aux observations et calcule la variance des estimations des coefficients de régression selon une méthode conçue par Fuller (1975).

Les coefficients de régression estimés à partir d'un échantillon stratifié en grappes proviennent de l'équation:

$$\underset{\sim}{b} = (\underset{\sim}{X}_n' \, W \, \underset{\sim}{X}_n)^{-1} \quad \underset{\sim}{X}_n' \, W \, \underset{\sim}{Y}_n$$

où le (rs)ième élément de $(\underset{\sim}{X}_n' \underset{\sim}{\sim} W \underset{\sim}{\sim} \underset{\sim}{X}_n)$ est

$$\sum_{h=1}^{L} \sum_{i_1=1}^{n_h} \sum_{i_2=1}^{n_{hi_1}} X_{hi_1i_2r} \quad X_{hi_1i_2s} \quad W_{hi_1i_2}$$

et le (r)ième élément de $\underset{\sim}{X}_n' \underset{\sim}{\sim} W \underset{\sim}{\sim} \underset{\sim}{Y}_n$ est

$$\sum_{h=1}^{L} \sum_{i_1=1}^{n_h} \sum_{i_2=1}^{n_{hi_1}} X_{hi_1i_2r} \quad X_{hi_1i_2} \quad W_{hi_1i_2} \, .$$

La matrice des covariances estimées de $\underset{\sim}{b}$ est calculée ainsi:

$$v(\underset{\sim}{b}) = (\underset{\sim}{X}_n' \underset{\sim}{W} \underset{\sim}{X}_n)^{-1} \hat{\underset{\sim}{G}}_n (\underset{\sim}{X}_n' \underset{\sim}{W} \underset{\sim}{X}_n)^{-1} ,$$

où le (rs)ième élément de $\hat{\underset{\sim}{G}}_n$ est

$$\hat{g}_n(r,s) = \frac{n-1}{n-p} \sum_{h=1}^{L} \frac{n_h(1-f_h)}{(n_h-1)} \sum_{i_1=1}^{n_h} (\hat{d}_{hi_1.r} - \bar{d}_{h..r}) \times (\hat{d}_{hi_1.s} - \bar{d}_{h..s})$$

où

$$\hat{d}_{hi_1i_2r} = x_{hi_1i_2r} \hat{v}_{hi_1i_2} w_{hi_1i_2} ,$$

$$\hat{v}_{hi_1i_2} = y_{hi_1i_2} - \sum_{r=1}^{P} \hat{b}(r) x_{hi_1i_2} ,$$

$$\hat{d}_{hi_1.r} = \sum_{i_2=1}^{n_{hi_1}} \hat{d}_{hi_1i_2r} ,$$

$$\bar{d}_{h..r} = n_h^{-1} \sum_{i_1=1}^{n_h} \hat{d}_{hi_1.r} ,$$

$$n = \sum_{h=1}^{L} \sum_{i_1=1}^{n_h} n_{hi_1} .$$

La méthode d'estimation de la variance est fondée sur une extension asymptotique de la série de Taylor appliquée au vecteur des coefficients de régression obtenus pour l'échantillon. Cette technique offre plusieurs avantages

par rapport à celle des échantillons superposés répétés et équilibrés (balanced repeated replication) et à la technique dite "jack-knife". Premièrement, elle est assez facile à programmer et on peut l'adapter aux plans d'échantillonnage à plusieurs degrés. Ensuite, aucune contrainte n'est imposée au plan d'enquête (par exemple, deux échantillons répétés par strate) et les hypothèses sur lesquelles repose la méthode exigent un comportement régulier dans certains moments de la distribution de la population d'intérêt. Troisièmement, cette méthode requiert moins de calculs que les autres techniques.

Les données contiennent assez souvent des erreurs de mesure. Des notions théoriques pour les modèles de régression qui prennent en considération ces erreurs ont été présentées par Fuller (1980a), Fuller (1980b) et par Fuller et Hidiroglou (1978). SUPER CARP est également assez souple pour effectuer des tests d'hypothèses concernant n'importe quel sous-ensemble de paramètres de régression.

iv) Tableaux de contingence

SUPER CARP et MINI CARP exécutent le test de validité de l'ajustement ainsi que le test d'indépendance avec des données d'enquêtes complexes. Ces deux tests tiennent compte de la division en strates et en grappes prévue par le plan de sondage. Comme l'ont fait remarquer Rao et Scott (1981), les utilisateurs de la méthode classique du khi-carré de Pearson peuvent être sérieusement induits en erreur si les effets du plan d'enquête sont importants.

Pour le test de validité de l'ajustement, SUPER CARP et MINI CARP utilisent la valeur corrigée de la fonction discriminante de Wald, calculée à l'aide de la formule suivante:

$$F_{WG} = [(k-1)d]^{-1} (d-k+2) (\hat{\underset{\sim}{p}}-\underset{\sim}{p}_0)^T \hat{\underset{\sim}{V}}^{-1} (\hat{\underset{\sim}{p}}-\underset{\sim}{p}_0)$$

où

$\hat{\underset{\sim}{p}} = (\hat{p}_1, \hat{p}_2, \ldots, \hat{p}_{k-1})^T$ est le vecteur des proportions estimées à partir de la configuration des strates et des grappes,

$\underset{\sim}{p}_o = (p_{o1}, p_{o2}, \ldots, p_{o,k-1})^T$ est le vecteur des proportions hypothétiques

$\hat{V}$ = la matrice des covariances de $\hat{\underset{\sim}{p}}$ calculée à partir de la configuration des strates et des grappes,

k = le nombre de catégories examinées,

$$d = \sum_{h=1}^{L} (n_h - 1),$$

L correspond au nombre de strates dans l'échantillon et n_i représente le nombre de grappes dans la iième strate. On obtient la matrice des covariances $\hat{\underset{\sim}{V}}$ par les méthodes décrites pour l'estimation par quotient. Dans les grands échantillons, F est distribuée approximativement comme une variable F centrée à k-1 et d-k+2 degrés de liberté lorsque l'hypothèse nulle est vraie.

Quant au test d'indépendance, Fuller (SUPER CARP, p. 65-69) a élaboré une technique qui tient compte du plan de sondage. Un tableau de contingence où la population est répartie selon deux critères contient R rangés et C colonnes, et l'hypothèse nulle qu'il faut vérifier se définit comme suit: H_o: $P_{ij} = P_{i+} P_{+j}$ ou $P_{+j} = P_{i+}^{-1} P_{ij}$ où P_{ij} est la proportion de la population qui figure dans la (ij)ième case, $P_{+j} = \sum_i P_{ij}$ et $P_{i+} = \sum_j P_{ij}$.

Si nous définissons $P_{ij|i}$ comme étant égal à $P_{i+}^{-1} P_{ij}$ et désignons les estimations obtenues pour l'échantillon par la notation $\hat{P}_{ij|i} = \hat{P}_{i+}^{-1} \hat{P}_{ij}$, on peut estimer $(P_{+1}, P_{+2}, \ldots, P_{+,C-1})$ en effectuant la régression de $\hat{P}_{ij|i}$ (i = 1, 2, ..., R; j = 1, 2, ..., C-1) par rapport à des vecteurs-lignes à (C-1) dimensions qui contiennent 1 à la jième entrée correspondant à $\hat{P}_{ij|i}$ et des zéros ailleurs. Cette régression relève de la méthode des moindres carrés généralisée parce que la structure des erreurs des $\hat{P}_{ij|i}$ n'est pas uniforme. On obtient une estimation de la matrice des covariances des $P_{ij|i}$, qui tient compte du plan d'enquête, au moyen des formules de covariances élaborées pour la méthode des quotients. La variable utilisée dans le test de H_0 est ensuite calculée à partir des sommes des carrés des résidus produits par la régression.

3. ENTRÉES

Dans une enquête, les données relatives à chaque unité choisie sont normalement caractérisées par la strate, le degré d'échantillonnage (premier, deuxième, ..., s ième) et un facteur de pondération. Les données doivent être classées selon cet ordre afin de produire des estimations de la variance qui correspondent à la structure des strates et des grappes.

SUPER CARP et MINI CARP utilisent un langage de commande composé de codes numériques placés à des positions fixes sur les cartes informatiques. Dans ces deux programmes, il ya six cartes de contrôle obligatoires qui doivent toujours faire partie des entrées. Un certain nombre de cartes de contrôle facultatives peuvent aussi être incluses lorsque des renseignements supplémentaires sont requis pour des options spécifiées sur les cartes obligatoires. Les cartes obligatoires comprennent la carte des paramètres, la carte des noms des variables, la carte du format, la carte de sélection, la carte d'analyse et la carte d'identification des variables. La carte des paramètres fournit

de l'information préliminaire d'ordre général, comme la définition du problème, le nombre d'observations à stocker, le support sur lequel se trouve les entrées (bande, disque ou cartes), la structure des données, de même que des renseignements sur la sortie des données et des contrôles pour la combinaison des strates. La carte du format indique la composition des données d'entrée ainsi que leur nature et le poids correspondant. La carte des noms des variables attribue des noms choisis aux zones des données d'entrée, suivant l'ordre dans lequel les données sont introduites. La carte de sélection contient les limites valables pour certaines variables lorsqu'une telle opération est nécessaire. La carte d'analyse énumère les analyses à effectuer (voir le tableau 1). Enfin, la carte d'identification des variables désigne les variables qui seront utilisées dans les analyses demandées. Parmi les cartes facultatives, on retrouve les cartes des fractions de sondage (on peut indiquer les pas de sondage de chaque strate), les cartes des erreurs sur les variables, qui donnent au programme la matrice des covariances pour les variables mesurées avec une erreur, et la carte des tests d'hypothèses qui permet de spécifier des coefficients de régression et de vérifier s'ils sont égaux à zéro.

4. CALCULS

4.1 Moyennes, sommes des carrés corrigées et produits vectoriels

Les moyennes, les sommes des carrés corrigées et les produits vectoriels sont des fonctions normalement traitées dans un programme d'enquête. Pour choisir les algorithmes nécessaires au calcul de ces fonctions, il faut prendre en considération le degré d'exactitude visé, la vitesse d'exécution et les contraintes liées au stockage des données. Beaton, Rubin et Barone (1976) ont admis que la recherche de méthodes de calcul très exactes doit être subordon-

née à la préoccupation de s'assurer que les données sont assez précises pour que les résultats soient significatifs. Divers modèles d'algorithmes à un et à deux passages ont été étudiés par Ling (1974), qui est arrivé à la conclusion qu'il n'y a pas d'algorithme supérieur aux autres dans tous les cas. Le meilleur algorithme pour un ensemble de données en particulier dépend des chiffres contenus dans cet ensemble. Une des suggestions formulées par Ling est d'effectuer des calculs en double précision pour obtenir des résultats plus précis que ceux des calculs en simple précision. Les algorithmes récurrents à un passage sont préférable aux méthodes habituelles à un passage du type "machine de bureau", parce qu'ils tendent à produire moins d'erreurs de calcul. Cela se note surtout dans les sous-programmes avec simple précision. Dans SUPER CARP et MINI CARP, on a choisi des algorithmes récurrents à un passage programmés avec double précision.

4.2 Inversion de matrices

L'inversion de matrices est nécessaire à la régression et à l'analyse de tableaux de contingence. Le choix de l'algorithme d'inversion dans une méthode informatique est très important, comme le démontre l'étude de Longley (1967), où l'auteur examine la précision de divers algorithmes d'inversion et découvre de sérieuses imperfections dans les calculs. Longley affirme avoir obtenu les résultats les plus précis en utilisant la technique d'orthonormalisation. Kopitze, Boardman et Graybill (1975) préconisent l'application de la décomposition de Cholesky comme algorithme d'inversion. Ces auteurs mentionnent que, contrairement à la méthode d'élimination de Gauss, la technique de Cholesky ne requiert pas de pivot pour stabiliser les matrices définies positives symétriques. L'inversion prend donc moins de temps. La décomposition ne demande pas beaucoup de place en mémoire et elle est plus facile à programmer que la méthode d'élimination de Gauss. Un autre avantage de la décomposition de Cholesky est qu'elle est assez précise, comme le démontre

l'analyse de Wilkinson (1965). De plus, cette technique permet de trouver les valeurs propres de systèmes d'équations ayant la forme A x = B x, où A est une matrice positive et B, une matrice semi-définie positive. Le calcul de valeurs propres est nécessaire dans SUPER CARP pour quelques-unes des analyses de régression avec erreurs sur les variables. Pour cette raison et compte tenu des critères de l'exactitude établis, on a adopté la décomposition de Cholesky comme méthode d'inversion dans SUPER CARP.

4.3 Combinaison de strates

Lorsqu'une population échantillonnée est très hétérogène et qu'on applique plusieurs critères de stratification, il est fort possible que certaines strates contiennent seulement une grappe. Dans ce cas, il est impossible d'estimer la variabilité, et l'utilisateur peut demander que les strates composées d'une seule grappe soient combinées avec des strates voisines. Si cette demande n'est pas faite, SUPER CARP et MINI CARP excluent ces strates des calculs de la variance, mais les incluent pour les besoins d'estimation. Le programme produit une liste des strates à une seule grappe, ce qui peut aider l'utilisateur à combiner ce genre de strates quand il présente un programme par la suite. Pour cette combinaison, les strates à une grappe doivent avoir les mêmes caractéristiques que des strates voisines. On peut suggérer la méthode suivante qui se prête bien à la programmation. Lorsque le programme découvre une strate qui ne contient qu'une grappe, cette strate est fondue avec la strate suivante classée dans le fichier. Si la dernière strate est composée d'une seule grappe, la dernière strate est combinée avec l'avant-dernière. Une strate dont la fraction d'échantillonnage est égale à 1 n'est pas combinée parce qu'une telle strate n'influe pas sur la variance observée entre les unités primaires de l'échantillon. Les strates dont la fraction d'échantillonnage est égale à 1 ne doivent jamais figurer après une strate

formée d'une seule grappe. Pour s'assurer que cette exigence est respectée, il suffit d'enregistrer tous les groupes d'observations ayant une fraction d'échantillonnage égale à 1 au début du fichier. Quand deux strates sont regroupées, la fraction d'échantillonnage de la nouvelle strate est calculée en fonction de la fraction d'échantillonnage de chacune des deux strates combinées et du nombre d'éléments qu'elles contiennent.

4.4 Grappes formées d'un seul élément

Si des grappes qui ne contiennent qu'un élément se trouvent dans une strate au premier degré d'échantillonnage, la combinaison de strates voisines permet de calculer des estimations de la variance. Dans un plan de sondage à plusieurs degrés, il peut arriver que certains degrés d'échantillonnage produisent des grappes formées d'un seul élément. Pour ce genre de grappes, il est impossible de calculer la variation à l'intérieur de l'ensemble des grappes pour lesquelles cette variation peut être calculée. L'incidence de ces degrés d'échantillonnage sur la variance, lorsque certaines grappes n'ont qu'un élément, peut alors être représentée par cette approximation.

5. QUELQUES CARACTÉRISTIQUES SOUHAITABLES DANS UN PROGRAMME D'ESTIMATION DE LA VARIANCE

Francis, Heiberger et Velleman (1975) ont dressé une liste de critères utiles d'évaluation des programmes statistiques en général. Dans la présente section, nous énumérons les caractéristiques qu'un programme informatique doit posséder pour l'estimation de la variance dans les enquêtes complexes. Parmi ces facteurs nécessaires, mentionnons la documentation des utilisateurs, les contrôles des données à l'entrée, les listes imprimées et l'efficacité statistique. Nous examinons ici dans quelle mesure les caractéristiques de SUPER CARP et MINI CARP répondent à ces besoins.

Essentiellement, la documentation des utilisateurs est un guide qui explique à l'utilisateur la façon de se servir du programme. SUPER CARP et MINI CARP comprennent tous les deux un guide de ce genre qui se présente sous la forme suivante. D'abord, une introduction résume les diverses options statistiques offertes par la méthode. Ensuite, les données d'entrée et les commandes relatives aux analyses, aux variables et aux options sont expliquées, avec exemples à l'appui. Comme les données doivent être introduites et les commandes placées selon un ordre particulier, un organigramme d'analyse est inclus. Les techniques offertes par les programmes sont décrites en fonction des formules et des méthodes numériques utilisées, et avec quelques références à divers ouvrages.

Comme il a été mentionné précédemment, le langage de commande de SUPER CARP et de MINI CARP est constitué de codes numériques ou alphanumériques enregistrés dans des colonnes fixes sur des cartes. Francis, Heiberger et Velleman (1975) signalent que les langages de commande qui peuvent exécuter les calculs les plus efficaces sont fondés sur l'utilisation de codes numériques dans des colonnes fixes. L'inconvénient de cette méthode est que les utilisateurs doivent se référer trop souvent au manuel pour trouver des commandes. Il serait peut-être possible de permettre aux utilisateurs de demander des analyses et des options par des commandes semblables avec des mots anglais, en ajoutant un programme de traduction des commandes. L'avantage de cette méthode est qu'elle serait assez facile à apprendre, par contre le temps et le travail nécessaires pour programmer le traducteur en rendraient le coût prohibitif.

Les listes imprimées par SUPER CARP et MINI CARP indiquent la technique statistique appliquée et les variables utilisées dans l'analyse. Une partie de chaque liste imprimée montre le numéro de la version de SUPER CARP ou de MINI CARP et la date de la dernière mise à jour. Ces renseignements peuvent

servir à repérer des erreurs dans la version indiquée et à les corriger. De plus, divers messages sont également imprimés. Par exemple, certains messages concernent les contrôles des données à l'entrée, comme lorsqu'on tente d'enregistrer plus de variables que le programme ne le permet, quand on tente d'introduire trop de grappes ou quand le format des données d'entrée ne répond pas aux normes établies. Si certaines strates contiennent une seule grappe, une liste de ces strates sera dressée. Dans le cas où l'utilisateur demande la combinaison de strates formées d'une grappe seulement, les strates ainsi produites sont imprimées.

SUPER CARP et MINI CARP sont écrits en FORTRAN avec double précision. Ils peuvent être utilisés sur les ordinateurs dotés d'un compilateur FORTRAN moyennant quelques petites modifications du langage de contrôle des travaux. On peut aussi ajouter de nouvelles techniques statistiques à SUPER CARP ou à MINI CARP. Ces techniques peuvent s'intégrer au programme sous la forme de sous-programmes qui peuvent être reliés aux autres éléments du logiciel.

6. BIBLIOGRAPHIE

[1] Beaton, A.E., Rubin, D.B., et Barone, J.L. (1976), The Acceptability of Regression Solutions: Another Look at Computational Accuracy. Journal of the American Statistical Association, 71, 158-168.

[2] Raj, Des (1968), Sampling Theory, McGraw-Hill Inc.

[3] Raj, D. (1966), Some Remarks on a Simple Procedure of Sampling Without Replacement, Journal of the American Statistical Association, 61, 391-397.

[4] Francis, I., Heiberger, R.M., et Velleman, P.F. (1975), Criteria and Considerations in the Evaluation of Statistical Program Packages. The American Statistician, 29, 52-56.

[5] Fuller, W.A. (1975), Regression Analysis for Sample Surveys, Sankhya, 37, 117-132.

[6] Fuller, W.A. et Hidiroglou, MA. (1968), Regression Estimation After Correcting for Attenuation, Journal of the American Statistical Association, 73, 99-104.

[7] Fuller, W.A. (1980a), Properties of Some Estimators for the Errors-in-Variables Model, Annals of Statistics, 8, 407-422.

[8] Fuller, W.A. (1980b), Estimation of Measurement Error Models from Cluster Samples, Communication présentée aux réunions de l'American Educational Research Association. Boston, Massachusetts.

[9] Garza-Hernandez, T. et McCarthy, P.J. (1962), A Test of Homegeneity for a Stratified Sample. Proceedings of the Social Statistics Section of the American Statistical Association, 200-202.

[10] Gray, G.B. (1975), Components of Variance Model in Multistage Stratified Samples, Techniques d'enquête, Statistique Canada, 1, 27-43.

[11] Hartley, H.O. (1959), Analytic Studies of Survey Data. Instituto de Statistica, Rome, Volume in onore di Corrado Gini.

[12] Hidiroglou, M.A., Fuller, W.A. et Hickman, R.D., (1980), SUPER CARP, Survey Section, Iowa State University, Ames, Iowa.

[13] Hidiroglou, M.A., Fuller, W.A., et Hickman, R.D. (1980), MINI CARP, Survey Section, Iowa State University, Ames, Iowa.

[14] Holt, Mary M. (1977), SURREGR: Standard Errors of Regression Coefficients from Sample Survey Data, rapport non publié, Research Triangle Institute, Research Triangle Park, North Carolina.

[15] Kish, L., et Frankel, M.R. (1974), Inference from Complex Samples. Journal of the Royal Statistical Society B. 36, 1-37.

[16] Konijn, H.S. (1962), Regression Analysis in Sample Surveys. Journal of the American Statistical Association 5 7, 590-606.

[17] Kopitze, R., Boardman, T.J. et Graybill, F.A. (1975), Least Square Programs - A Look at the Square Root Procedure. The American Statistician, 29, 64-66.

[18] Ling, R.G. (1974), Comparaison of Several Algorithms for Computing Sample Means and Variances. Journal of the American Statistical Association, 69, 859-866.

[19] Longley, James (1967), An Appraisal of Least Squares Programs for the Electronic Computer from the Point of View of the User. Journal of the American Statistical Association, 62, 819-841.

[20] Nathan, G. (1972), On the Asymptotic Power for Tests for Independence in Contingency Tables from Stratified Samples. Journal of the American Statistical Association, 67, 917-920.

[21] Rao, J.N.K. (1975), Analytic Studies of Sample Survey Data. Techniques d'enquête, Statistique Canada.

[22] Rao, J.N.K. (1975), Unbiased Variance Estimation for Multistage Designs, Sankhya C, 37, 133-139.

[23] Rao, J.N.K. et Scott, A.J. (1981), The Analysis of Categorical Data from Complex Sample Surveys: Chi-square Tests for Goodness-of-Fit and Independence in Two-Way Tables. Journal of the American Statistical Association, 76, 221-230.

[24] Shah, B.V. (1974), STDERR: Standard Errors Program for Sample Survey Data, Research Triangle Institute, Research Triangle Park, North Carolina.

[25] Wilkinson, J.H. (1975), The Algebraic Eigenvalue Problem. Oxford: Clarendon Press, 229-233.

[26] Yates, F. (1960), Sample Methods for Censuses and Surveys. Charles Griffin and Sons, London, Third Edition.