Techniques d'enquête, juin 1988
Vol. 14, n° 1, pp. 63–73
Statistique Canada

Logiciel d'ordinateur personnel pour l'estimation de la variance dans des enquêtes complexes

DAN SCHNELL, WILLIAM J. KENNEDY, GARY SULLIVAN, HEON JIN PARK et WAYNE A. FULLER[1]

RÉSUMÉ

Le présent article contient la description d'un programme d'ordinateur personnel servant à l'estimation de la variance pour de grandes enquêtes. Ce programme, connu sous le nom de PC CARP, permet de calculer des estimations pour des agrégats, des rapports, des moyennes, des quantiles et des coefficients de régression et d'estimer les variances correspondantes.

MOTS CLÉS: Échantillon d'enquête; estimation de la variance; logiciel d'enquête.

1. INTRODUCTION

L'analyse de données d'enquête comporte habituellement un grand nombre d'observations et des calculs relativement complexes en ce qui a trait à la variance. Grâce aux derniers perfectionnements qui leur ont été appliqués, les ordinateurs personnels peuvent maintenant servir à traiter des données d'enquêtes complexes. Nous nous proposons dans cet article de décrire un programme d'ordinateur personnel qui est destiné à l'analyse de données d'enquête et qui a été mis au point à l'université Iowa State.

La création d'un logiciel d'ordinateur personnel pour l'estimation de la variance a été le résultat de la collaboration entre l'université Iowa State et le International Statistical Programs Center du U.S. Census Bureau. Par ce projet, le Census Bureau visait à offrir aux pays en voie de développement un logiciel qui pourrait être utilisé sur place pour traiter des données recueillies sur place. Le projet de l'université Iowa State s'inscrivait dans un programme du Census Bureau qui avait pour but de créer des logiciels pour la gestion des enquêtes et le contrôle et la totalisation des données.

Au début des années 1970, l'université Iowa State, se fondant sur les travaux d'Hidiroglou (1974) et Fuller (1975), élaborait un programme ayant pour but de calculer des coefficients de régression et la matrice estimée des covariances de ces coefficients pour des données d'enquête. Ce programme, connu sous le nom de SUPER CARP, devait plus tard être modifié de manière à inclure l'estimation d'agrégats et de rapports, les statistiques de sous-population, les tableaux à double entrée et les échantillons à deux degrés. La dernière révision de SUPER CARP remonte à 1980. Ce programme a ouvert la voie à la création de logiciels pour ordinateurs personnels. Pour cette raison, le programme d'ordinateur personnel a été baptisé PC CARP.

2. CARACTÉRISTIQUES DU PROGRAMME

PC CARP est conçu pour les modèles IBM PC, IBM PC/XT, IBM PC/AT et les ordinateurs compatibles. Il exige une mémoire d'au moins 410 kilo-octets ainsi qu'un coprocesseur "mathématique".

[1] Dan Schnell, Centers for Disease Control, 1600 Clifton Road, NE, Atlanta, (Georgie) 30333. William J. Kennedy, Gary Sullivan, Heon Jin Park et Wayne A. Fuller, Département de statistique, Iowa State University, Ames, Iowa 50011, États-Unis.

PC CARP traite avec autant de facilité et d'efficacité les longues et les petites séries de données. Il ne pose pas de limite pour le nombre de strates ou de grappes qui peuvent être incluses dans une série de données et peut accepter jusqu'à 50 paramètres à la fois. Il accepte, par ailleurs, les fichiers-disques à format de bloc fixe ou à structure libre.

Le programme peut servir à calculer les variances pour des échantillons à un ou à deux degrés avec correction pour population finie. Dans le cas des échantillons à plus de deux degrés, les termes de correction pour population finie n'existent qu'à deux niveaux. En ce qui concerne les échantillons à deux degrés, le programme calcule les taux de sondage à l'intérieur des grappes à partir des taux de sondage des strates et des poids des observations.

En règle générale, on connaîtra la strate et la grappe (unité primaire d'échantillonnage) auxquelles appartient chaque observation du fichier et on connaîtra aussi le poids de ces observations, qui correspond à l'inverse de la probabilité de sélection. L'utilisateur est libre d'introduire les taux de sondage du premier degré. Les plans de sondage simples, comme l'échantillonnage aléatoire simple, n'exigent pas tous ces renseignements. Dans de tels cas, il est possible d'avoir moins de paramètres.

Dans le cas d'un échantillonnage stratifié, toutes les observations qui sont contenues dans la même strate doivent être regroupées. S'il s'agit d'un échantillonnage par grappes, toutes les observations contenues dans la même grappe doivent être regroupées.

Le tableau 1 contient une liste des paramètres statistiques qui peuvent être calculés à l'aide du PC CARP. Outre les fonctions énumérées dans le tableau 1, il existe des cartes supplémentaires

Tableau 1

Fonctions d'analyse disponibles dans PC CARP

Analyse	Coefficient de variation	Matrice des covariances	Effet du plan	Remarques
Analyses de populations				
Estimation d'agrégats	X	X	X	50 variables au maximum
Estimation de rapports	X	X	X	50 variables au maximum sans les covariances, 15 avec covariances
Différence de rapports			X	15 variables au maximum
Analyses de strates				
Agrégats	X		X	50 variables au maximum
Moyennes	X		X	50 variables au maximum
Proportions	X		X	50 variables au maximum
Analyses de sous-populations				
Agrégats	X		X	Classement recoupé Variables multiples
Moyennes	X		X	Classement recoupé Variables multiples
Proportions	X		X	Classement recoupé Variables multiples
Rapports	X		X	Classement recoupé Variables multiples
Autres analyses				
Tableau à double entrée		X		50 cases au maximum, test de proportionnalité
Régression		X		50 variables au maximum Tests à plusieurs degrés de liberté, valeurs prédites résidus
Analyse unidimensionnelle			X	Variables multiples, fonctions de distribution cumulatives empiriques, quantiles

pour l'estimation de la fonction logistique et la stratification a posteriori. Ces fonctions supplémentaires sont décrites dans les sections 4.4 et 4.5. Un "X" sous la rubrique "matrice des covariances" signifie que l'on peut calculer la matrice des covariances d'un vecteur d'estimations du genre de celles indiquées dans la colonne de gauche. L'erreur type peut être calculée pour tous les paramètres statistiques mais la matrice des covariances d'un vecteur ne peut l'être que pour un nombre limité de paramètres. En outre, le coefficient de variation peut être calculé pour un grand nombre de paramètres. L'effet du plan, désigné par DEFF, peut être calculé sur demande pour un grand nombre de paramètres. Voir Kish (1965) pour une description de l'effet du plan.

Les analyses de population (agrégat, rapport et différence de rapports) et les analyses de strate s'effectuent d'une manière simple. Pour ce qui est des autres analyses (analyse de sous-population, de tableau à double entrée, de régression et analyse unidimensionnelle), on trouvera les détails pertinents à la section V.

Les analyses de sous-population offrent à l'utilisateur la possibilité de croiser des variables de classification. Grâce à cela, l'utilisateur peut créer de nouvelles structures de classification à l'aide de deux variables de classification ou plus. Supposons, par exemple, que les données d'entrée comprennent les variables de classification "âgé" (6 niveaux), "sexe" (2 niveaux) et "niveau d'instruction" (5 niveaux). Si nous combinons ces trois variables, nous allons obtenir une nouvelle structure de classification à soixante niveaux. Suivant cette structure, l'utilisateur peut établir des estimations pour un nombre indéterminé de variables dépendantes.

L'analyse des tableaux à double entrée est définie par deux variables de classification et une variable dépendante. Il est possible de définir plus d'une variable dépendante pour une paire de variables de classification. Pour chaque variable dépendante, le programme calcule les tables de fréquences de cases ainsi que les tables de proportions fondées sur les totaux de ligne, les totaux de colonne et le grand total. Il calcule aussi des erreurs types pour tous les estimateurs et produit une fonction des observations visant à tester l'hypothèse de la proportionnalité. Cette fonction est fondée sur une approximation de Satterthwaite de la distribution du critère chi-carré de Pearson. Voir aussi Rao et Scott (1984).

L'analyse de régression par les moindres carrés pondérés permet de calculer des estimations de coefficients ainsi qu'une matrice estimée des variances-covariances qui tient compte du plan de sondage. Ces calculs sont présentés dans Fuller (1975) et repris en bref dans Hidiroglou et coll. (1980). Des tests F à plusieurs degrés de liberté pour des séries de coefficients et les variables t habituelles sont également prévus. L'utilisateur a aussi la possibilité de calculer des résidus et des valeurs estimatives.

L'analyse unidimensionnelle produit des statistiques qui décrivent la distribution d'une variable. L'utilisateur définit la variable d'intérêt et détermine une sous-population en choisissant une catégorie d'une variable de classification. Ainsi, l'utilisateur pourrait choisir de recueillir des données sur le revenu personnel des membres de la catégorie des professions libérales dans la classification des professions. Le programme produit en l'occurrence des estimations de la moyenne de sous-population, de la variance, de la fonction de distribution, de quantiles et de l'écart interquartile.

3. DÉTAILS SUR LE PROGRAMME

PC CARP est écrit presque entièrement en FORTRAN, le langage de programmation scientifique le plus connu, et c'est le compilateur Professional FORTRAN d'IBM qui a été choisi pour le projet. Une petite partie du code — quelques sections de l'interface avec l'utilisateur — est rédigé en langage assembleur d'IBM.

Durant l'élaboration du programme, on a surtout cherché à créer une interface conviviale et à réduire au minimum le nombre de fois que les données doivent être parcourues. On a réussi à créer une interface conviviale en mettant en place un système de réponse interactif à écran. Par ailleurs, un algorithme monopasse pour l'estimation de la variance de paramètres statistiques

élémentaires a permis de réduire au minimum la fréquence de lecture des fichiers. Une seule lecture des données suffit pour calculer la plupart des estimateurs et leur variance.

Le programme peut calculer des estimateurs pour la population totale, pour chaque strate ou pour des sous-populations déterminées. Ces estimateurs sont pour la plupart des fonctions de totaux d'échantillon pondérés. Si l'on veut calculer, par exemple, l'estimateur des rapports $R_1 = Y_1/X_1$ et $R_2 = Y_2/X_2$, on recueille les totaux pour Y_1, X_1 Y_2, et X_2. Si l'estimation concerne la population totale, on forme ces totaux par une seule lecture des données. Dans le cas des estimations de strate, les totaux peuvent être accumulés, et combinés si nécessaire, puis restitués strate par strate. Comme les données sont groupées par strate, une lecture suffit pour déterminer les totaux de strate pour n'importe quelle quantité de strates. Pour ce qui a trait aux estimations relatives à une sous-population, une seule lecture des données pourrait ne pas suffire si le nombre de catégories définies par la structure de classification est élevé. L'analyse de régression et l'analyse unidimensionnelle exigent deux lectures de données.

Tous les estimateurs, à l'exception des estimateurs d'agrégats, sont des fonctions non linéaires des moments pondérés de l'échantillon. Il est donc nécessaire d'utiliser une méthode conforme à une fonction non linéaire pour estimer la variance de la distribution approximative de ce genre d'estimateurs. Voir Wolter (1985) pour une analyse de l'estimation de la variance pour des enquêtes complexes. La méthode de Taylor (méthode des écarts statistiques) est la méthode d'estimation de la variance utilisée dans PC CARP. Il a été démontré que la méthode de Taylor était en règle générale aussi efficace sinon plus efficace que les autres méthodes d'estimation de la variance des paramètres statistiques à l'étude (p. ex.: les rapports). Voir, par exemple, Frankel (1971). La méthode de Taylor pour l'estimateur de rapports est décrite dans des ouvrages aussi courants que celui de Cochran (1977) et la même méthode pour l'estimateur de coefficients de régression est décrite dans Fuller (1975).

Dans la plupart des cas, l'estimateur et sa variance estimée peuvent être calculés par la même lecture de données parce que l'approximation de Taylor du premier ordre de la variance peut être exprimée en fonction des variances des totaux. Par exemple, l'approximation de Taylor du premier ordre de $\hat{R} = \hat{Y}/\hat{X}$ est

$$\hat{R} \doteq R + X^{-1}(\hat{Y} - R\hat{X}),$$

où $R = Y/X$ est le rapport des totaux réels. Par conséquent, la variance estimée d'un rapport $\hat{R} = \hat{Y}/\hat{X}$ peut être calculée à l'aide de la variance estimée des totaux de Y, X, et $(Y - X)$. De même, la matrice estimée des covariances pour $\hat{R}_1 = \hat{Y}_1/\hat{X}_1$ et $\hat{R}_2 = \hat{Y}_2/\hat{X}_2$, peut être calculée à l'aide des variances estimées des totaux de Y_1, X_1, $(Y_1 - X_1)$, Y_2, X_2, $(Y_2 - X_2)$, $(Y_1 - Y_2)$, $(Y_1 - X_2)$, $(Y_2 - X_1)$, et $(X_1 - X_2)$.

L'algorithme servant au calcul de la moyenne pondérée, des sommes des carrés pondérées et des matrices de produits croisés est décrit dans Herraman (1968). Pour les valeurs d'échantillon $\{X_i\}$ et les poids correspondants $\{W_i\}$, la suite des moyennes pondérées $\bar{X}_K$, et des sommes des carrés corrigées pondérées S_K est calculée à l'aide des formules suivantes

$$\bar{X}_K = \bar{X}_{K-1} + a_K d_K \text{ et } S_K = S_{K-1} + D_K - D_K a_K,$$

où $d_k = X_K - \bar{X}_{K-1}$, $a_K = W_K(\Sigma_{i=1}^{K} W_i)^{-1}$, et $D_K = d_k^2 W_K$.

Il est possible de calculer simultanément trois composantes de variance pour n'importe quel estimateur donné. Il s'agit de la composante de variance du premier degré, de la composante de variance optionnelle du second degré et de la variance optionnelle d'un échantillonnage aléatoire simple utilisée dans le calcul de l'effet du plan. Calculer toutes ces composantes par une seule

lecture de données exige beaucoup d'espace mémoire. Cependant, lorsqu'on travaille avec de grands échantillons, la suppression de lectures entières compense l'utilisation d'espace mémoire additionnel.

Le programme exécute régulièrement des vérifications visant à déceler des erreurs de calcul comme un diviseur nul. Si, par exemple, l'utilisateur introduit une série de données avec une strate à une seule grappe, le programme attribuera une variance nulle à la strate, complètera les calculs et fera imprimer un message d'erreur qui signalera la strate en question.

Le système de traitement d'erreurs a été conçu de manière à prévenir les arrêts de programme causés par des erreurs de l'utilisateur qui peuvent être corrigées facilement. Le programme renferme des fonctions de vérification pour l'absence de réponse, l'inexactitude de noms de fichiers et l'inexactitude de définitions de variables d'analyse. Si une erreur est décelée, le PC CARP permet à l'utilisateur de réintroduire les données ou de sortir du programme.

On a évalué la précision du programme en construisant des exemples d'application puis en comparant les résultats à ceux obtenus à l'aide du programme principal SUPER CARP. La série de données de Longley (1967) a servi à évaluer la précision du programme de régression. Des vérifications supplémentaires ont été faites à l'aide de PROC MATRIX du progiciel SAS. Voir Barr et coll. (1979). On a constaté que PC CARP avait un niveau de précision numérique équivalent à celui des progiciels principaux. On a aussi vérifié la cohérence interne de PC CARP en calculant des estimateurs équivalents à l'aide d'options différentes (p. ex., calcul d'une moyenne de sous-population au moyen de l'option "analyse de sous-population" et de l'option "analyse de rapports").

Lorsque PC CARP demande de l'information à l'utilisateur, une série de questions à réponse brève apparaît à l'écran avec des instructions détaillées. La première série d'affichages demande à l'utilisateur de fournir des renseignements sur l'organisation et l'emplacement des données. Des options "Help" et "Go Back" peuvent être utilisées à plusieurs endroits.

La seconde phase de l'exécution du programme a trait à la définition de l'analyse. Dans cette phase, l'utilisateur choisit le genre d'analyse, les options pertinentes et les variables d'analyse. Un nombre indéfini d'analyses peuvent être exécutées à l'aide des données définies dans la phase 1.

4. CHARACTÉRISTIQUES SPECIALES

4.1 Tableau à double entrée

Comme nous l'avons vu dans la Section 2, cette option fournit automatiquement à l'utilisateur quatre tableaux dont les éléments sont déterminés par le genre de liaisons qui ont été définies dans la construction du tableau à propos des fréquences par case.

Nous exposons ci-dessous la méthode qui a servi à construire le tableau des proportions de case et la matrice estimée des covariances des proportions. Supposons que le tableau est formé de R lignes et de C colonnes et posons $\hat{Y}_{rc}$ comme la fréquence estimée pour la case rc. Soit $\hat{Y}$ le vecteur colonne $RC \times 1$ des fréquences de case, créé en reproduisant l'une au-dessous de l'autre, dans l'ordre normal, les colonnes du tableau $R \times C$. Soit

$$\hat{Y}_{..} = \sum_{r=1}^{R} \sum_{c=1}^{C} \hat{Y}_{rc},$$

$$\hat{P}_{rc} = \hat{Y}_{..}^{-1} \hat{Y}_{rc}$$

le total estimé de la population et la proportion de case estimée pour la case rc respectivement.

Soit $\hat{P}$ le vecteur colonne $RC \times 1$, analogue à $\hat{Y}$, composé des RC valeurs $\hat{P}_{rc}$ disposées en colonnes. La matrice estimée des covariances pour $\hat{P}$ est

$$\hat{V}_{PP} = \hat{Y}_{..}^{-2}[I_{RC} - (\hat{P} \otimes J'_{RC})]\, \hat{V}_{YY}[I_{RC} - (\hat{P} \otimes J'_{RC})]',$$

où $\hat{V}_{YY}$ est la matrice estimée des covariances du vecteur des fréquences de case $\hat{Y}$, I_{RC} est la matrice unité de dimension $RC \times RC$ et J_{RC} est un vecteur colonne de dimension $RC \times 1$, constitué de uns.

La matrice $\hat{V}_{PP}$ sert à calculer la variable à tester pour l'hypothèse de la proportionnalité. L'hypothèse nulle, en l'occurrence, est que les éléments intérieurs du tableau de la population sont le produit des proportions marginales. Voir Rao et Scott (1984) pour une analyse des tests relatifs à ce genre d'hypothèses. Le test dans PC CARP est fondé sur une approximation de Satterthwaite de la distribution du critère chi-carré de Pearson, construite comme s'il s'agissait de proportions multinomiales. L'approximation est valide pour n'importe quelle variable d'analyse.

4.2 Estimation de quantiles

L'option "analyse unidimensionnelle" permet notamment de calculer des estimations de quantiles ainsi qu'un estimateur de l'erreur type des quantiles. Lorsqu'on veut calculer des quantiles, il faut tout d'abord construire un estimateur de la fonction de distribution cumulative. Une première lecture des données permet de déterminer l'intervalle des observations de même que la moyenne et l'écart type de l'échantillon. Par la même occasion, on relève les trois plus grandes et les trois plus petites observations.

La fonction de distribution cumulative estimée est définie par

$$\hat{F}_Y(x) = \left(\sum_{t=1}^{m} w_t Z_{St} \right)^{-1} \sum_{t=1}^{m} w_t Z_{St} I_Y(x),$$

où la sommation est étendue aux m éléments de l'échantillon, w_t est le poids d'échantillon, Z_{St} est une fonction indicatrice qui prend la valeur 1 si l'observation est comprise dans la sous-population d'intérêt et la valeur 0 dans le cas contraire, et $I_Y(x)$ est égale à 1 si $Y < x$ et est égale à zéro dans le cas contraire. Le domaine de la variable est divisé en 100 intervalles et la fonction de distribution cumulative est estimée aux 101 points définis par cette division.

Le programme estime ensuite la matrice des covariances pour la fonction de distribution estimée évaluée à 25 points, $j = 1, 5, \ldots, 96$. Les erreurs types estimées sont lissées à l'aide d'une moyenne mobile de trois points et une interpolation permet d'établir une erreur type estimée pour chacun des 101 points de la fonction de distribution estimée. Une interpolation linéaire permet de créer une fonction de distribution estimée qui est croissante au sens large. À l'aide des erreurs types lissées, le programme détermine une borne supérieure monotone croissante et une borne inférieure monotone croissante, qui délimitent un intervalle de confiance ponctuel de 95% pour la fonction de distribution. Ces bornes sont ensuite inversées pour former des intervalles de confiance de 95% pour les quantiles. L'écart interquartile et l'erreur type correspondante sont aussi estimés.

L'estimation de quantiles repose sur une théorie qui suppose l'existence d'une fonction de distribution de superpopulation sous-jacente avec une densité positive. Voir Francisco (1987) pour les aspects théoriques et Park (1987) pour les aspects relatifs aux calculs.

4.3 Estimation par régression

Les estimations des coefficients d'un modèle de régression linéaire sont calculées par la méthode des moindres carrés pondérés. Suivant la méthode décrite dans Fuller (1975), le programme calcule un estimateur de la matrice des covariances du vecteur des coefficients en tenant compte du plan de sondage.

Le vecteur des coefficients est estimé par

$$\hat{b} = (X' WX)^{-1} X' WY,$$

où X est la matrice $n \times p$ des valeurs de la variable indépendante, $\hat{Y}$ est le vecteur à n dimensions des valeurs de la variable dépendante, W est une matrice dont la diagonale est formée des poids des observations et le reste est formé de zéros, et n est le nombre total des observations. La variance de $\hat{b}$ est estimée par

$$\hat{V}(\hat{b}) = (X'WX)^{-1}\hat{G}_W(X'WX)^{-1}.$$

La matrice $\hat{G}_W$ est

$$\hat{G}_W = C \sum_{i=1}^{L} h_i \sum_{j=1}^{n_i} (\hat{d}_{ij} - \bar{d}_{i.})(\hat{d}_{ij} - \bar{d}_{i.})',$$

où

$$\hat{d}_{ij} = \sum_{k=1}^{m_{ij}} X_{ijk}\hat{v}_{ijk}W_{ijk},$$

$$\hat{v}_{ijk} = Y_{ijk} - \hat{b}'X_{ijk},$$

$$\bar{d}_{i.} = n_i^{-1}\sum_{j=1}^{n_i} \hat{d}_{ij},$$

$h_i = (n_i - 1)^{-1}n_i$, $C = (n - p)^{-1}(n-1)$, m_{ij} est le nombre d'éléments contenus dans la grappe j de la strate i, n_i est le nombre de grappes dans la strate i, n est le nombre total d'observations, L est le nombre de strates et p est le nombre de coefficients estimés. L'estimateur de la variance calculé par cette option diffère de l'estimateur habituel calculé par les moindres carrés pondérés en ce que la matrice $\hat{G}_W$ est substituée à $(X'WX)s^2$.

Le programme calcule une fonction des observations R^2 multiple pour les modèles qui ont une ordonnée ou une abscisse à l'origine. De plus, il effectue toujours un test F pour la régression générale et offre à l'utilisateur la possibilité de tester des sous-ensembles de coefficients.

4.4 Régression logistique

Cette option permet d'obtenir des estimations du modèle logistique à plusieurs variables. Les algorithmes destinés à la régression logistique ont été élaborés après que la version initiale de PC CARP a vu le jour. Comme la fonction de moyenne pour le modèle logistique est non linéaire dans les paramètres, les estimateurs sont calculés à l'aide d'un algorithme d'itération fondé sur les moindres carrés pondérés. Les variances des estimateurs sont calculées au moyen des méthodes décrites dans Fuller (1975) et appliquées à l'estimation non linéaire. Voir aussi Binder (1983). Au point de vue des modalités d'application, l'option "régression logistique" est en tout point semblable à l'option "régression". Par exemple, les variables dépendantes et indépendantes y sont définies de la même façon.

4.5 Stratification a posteriori

Après que l'on a élaboré la version originale du programme PC CARP, une option "stratification a posteriori" a été mise au point pour un bon nombre des estimateurs. La stratification a posteriori est censée être une opération par laquelle on corrige les poids de manière à obtenir des estimations pour certaines catégories, qui concordent avec des totaux de population connus. C'est ce que Fuller et Sullivan (1987) appellent la stratification a posteriori gamma.

Le programme calcule la variance de l'estimateur de stratification a posteriori à partir d'une formule dans laquelle l'estimateur est exprimé comme la somme d'estimateurs de rapports.

4.6 Regroupement de strates

Aux fins du calcul de la variance, l'utilisateur peut recourir à cette option afin d'éliminer les strates à une grappe. Lorsque cette option est appliquée, chaque strate à une grappe est assimilée à la strate suivante dans la série de données. On modifie en conséquence le numéro de strate et le numéro de grappe des enregistrements touchés par cette fusion. S'il existe des taux de sondage de strates, les nouveaux taux sont définis par

$$f_i^* = (n_i f_i^{-1} + n_{i+1} f_{i+1}^{-1})^{-1}(n_i + n_{i+1}),$$

où la strate i, avec $n_i = 1$, a été combinée avec la strate $i + 1$. Ces nouveaux taux sont enregistrés dans un fichier annexe pour usage ultérieur. L'ordre de classement des strates influera sur le mode de regroupement des séries de données et les taux de sondage de strates. Le programme exige une lecture de données additionnelle lorsque l'utilisateur recourt à l'option "regroupement de strates" ou à l'option "à deux degrés".

4.7 Imputation par la méthode "hot deck"

PC CARP exige des séries de données complètes pour l'analyse. Pour le contrôle et l'imputation des données, beaucoup d'utilisateurs choisiront d'écrire un programme spécial ou encore utiliseront l'un des programmes offerts à cette fin dans la gamme PC.

Un programme d'imputation par "hot deck", appelé PRE CARP, peut être fourni avec le PC CARP pour ceux qui en font la demande. La méthode "hot deck" consiste à remplacer une valeur manquante par la valeur de l'article correspondant de l'enregistrement qui vient immédiatement avant dans le fichier de données. PRE CARP permet à l'utilisateur de définir une variable de classification qui peut renfermer jusqu'à 10 catégories de telle sorte que la valeur manquante est remplacée par la valeur de l'enregistrement précédent dans la même catégorie. PRE CARP créera aussi une variable indicatrice pour chaque variable pour laquelle il y a des valeurs manquantes. Cette variable indicatrice peut être ensuite utilisée avec l'option "analyse de sous-population" pour calculer des moyennes en fonction des observations originales.

5. EXEMPLES

Dans cette section, nous effectuons plusieurs analyses à l'aide d'une série de données composée et présentons les durées d'exécution. Le but des passages d'essai n'est pas d'étudier toutes les combinaisons possibles de facteurs qui peuvent influer sur la durée de l'exécution mais d'estimer le temps requis pour exécuter quelques-unes des analyses offertes dans le programme.

Les données d'essai ont été constituées à partir d'un sous-ensemble de la seconde édition de la National Health and Nutrition Examination Survey (NHANES II). La série ainsi constituée comprend 2400 observations réparties en 32 strates. Chaque strate compte deux unités primaires d'échantillonnage et ces unités d'échantillonnage sont de taille variée. De plus, chaque observation est affectée d'un poids de sondage non nul.

Les variables contenues dans la série de données sont les suivantes:

1. Sexe	1 = homme, 2 = femme
2. Origine raciale	1 = blanc, 2 = noir, 3 = autre
3. Caractère agricole	1 = ménage non agricole, 2 = ménage agricole
4. Revenu	revenu du ménage (en milliers de dollars)
5. Âge	âge (en années).

Figure 1. Résultats de l'analyse C; âge moyen selon le sexe et l'origine raciale.

MOYENNES DE SOUS-POPULATION

Variable dépendante: âge

Catégorie	Estimation	E.T.	C.V.	DEFF (effet du plan)
Sexe = 1.0000	Origine raciale = 1.0000			
	3.06811D+01	6.19678D-01	2.0197D-02	1.3967D+00
Sexe = 1.0000	Origine raciale = 2.0000			
	3.13016D+01	7.88580D-01	2.5193D-02	9.5384D-01
Sexe = 1.0000	Origine raciale = 3.0000			
	3.41579D+01	2.24111D+00	6.5610D-02	2.0965D+00
Sexe = 2.0000	Origine raciale = 1.0000			
	1.33742D+01	3.18904D-01	2.3845D-02	1.2588D+00
Sexe = 2.0000	Origine raciale = 2.0000			
	*********			
Sexe = 2.0000	Origine raciale = 3.0000			
	1.71957D+01	9.53816D-01	5.5468D-02	1.1389D+00

Une variable qui prend la valeur 1 pour chaque observation (variable constante) a été créée par le programme. Les analyses exécutées sont les suivantes:

A. Revenu moyen de la population échantillonnée
B. Revenu moyen par strate
C. Âge moyen pour la classification selon le sexe et l'origine raciale
D. Fonctions de distribution de l'échantillon selon l'âge pour les groupes agricoles et non agricoles.

L'analyse A (estimation du revenu moyen) a été exécutée au moyen de l'option "analyse de rapports", où le revenu servait de numérateur et la variable constante de dénominateur. Les estimations du revenu moyen par strate (analyse B) ont été calculées directement au moyen de l'option "moyenne de strate". L'analyse C a été exécutée au moyen de l'option "moyenne de sous-population" en croisant les variables de classification "sexe" et "origine raciale" et en définissant l'"âge" comme la variable dépendante. Les résultats de l'analyse C sont reproduits dans la figure 1. Les symboles "*****" qui paraissent sous la classification "sexe = 2 origine raciale = 2" indiquent qu'il n'y avait pas d'observations relatives à cette catégorie. Les valeurs de l'effet du plan mettent en relief l'importance de tenir compte du plan de sondage dans l'estimation de la variance. Par exemple, l'effet du plan pour l'estimation correspondant à la catégorie "sexe = 1 et origine raciale = 3" est environ 2. Cela signifie que la variance estimée de la moyenne de l'échantillon pour un échantillonnage aléatoire simple équivaut à la moitié de la variance estimée pour un échantillonnage en grappes stratifié. Les caractéristiques de la distribution de l'âge pour chacun des deux états de la variable "caractère agricole" ont été estimées au moyen de l'option "analyse à une variable". La figure 2 donne les résultats de l'analyse à une variable pour les ménages non agricoles. Toutes les variances et les estimations d'erreurs types indiquées dans cette figure tiennent compte du plan de sondage.

La durée d'exécution des analyses A, B, C et D pour les 2,400 observations a été de 70, 135, 120 et 360 secondes respectivement. L'exécution s'est faite sur IBM PC AT, les données étant stockées sur le disque dur et lues en structure non imposée. Les taux de sondage des strates n'ont pas été introduits dans le programme. Les résultats ont été acheminés au récepteur de contrôle et à un fichier-disque. Les effets du plan pour les estimations ont dû être considérés dans les quatre analyses. Les trois premières analyses ont exigé chacune une seule lecture des données. Un plus

Figure 2. Résultats de l'analyse unidimensionnelle pour les ménages non agricoles (Exemple D).

ANALYSE UNIDIMENSIONNELLE 1

Variable de classification: caractère agricole état: 1

Nombre d'éléments de l'échantillon dans la sous-population = 111

Variable dépendante: âge

Variance de la sous-population = 4.25645D + 02

C.V. de la sous-population = 7.14101D-01

Moyenne de la sous-population

Estimation	E.T.	C.V.	DEFF
2.8891089D + 01	2.2543875D + 00	7.80305D-02	9.86336D-01

Valeurs extrêmes des éléments de l'échantillon dans la sous-population

Valeurs les plus faibles	Nombre de valeurs observées	Identification de la première observation: strate	grappe	poids
1.000D + 00	1	32	1	3.000D + 00
2.000D + 00	1	15	1	2.000D + 00
3.000D + 00	3	10	2	3.000D + 00

Valeurs les plus élevées	Nombre de valeurs observées	Identification de la première observation: strate	grappe	poids
7.400D + 01	2	29	1	3.000D + 00
7.100D + 01	2	7	1	2.000D + 00
7.000D + 01	4	7	1	2.000D + 00

Quantiles

	Estimation	E.T.	Intervalle de confiance de 95%
0.01	2.2690811D + 00	7.3167585D − 01	(8.05729D − 01, 3.73243D + 00)
0.05	4.2364814D + 00	1.1977759D + 00	(3.34942D + 00, 8.14053D + 00)
0.10	7.7750203D + 00	1.3563759D + 00	(5.36691D + 00, 1.07924D + 01)
0.25	1.3652930D + 01	1.4225576D + 00	(9.99238D + 00, 1.56826D + 01)
0.50	1.9449315D + 01	2.2740912D + 00	(1.57192D + 01, 2.48156D + 01)
0.75	4.5698071D + 01	4.7577709D + 00	(3.58936D + 01, 5.49247D + 01)
0.90	6.2787426D + 01	2.3472775D + 00	(5.51417D + 01, 6.45308D + 01)
0.95	6.5923423D + 01	1.2228344D + 00	(6.43837D + 01, 6.92750D + 01)
0.99	7.1714993D + 01	1.1425033D + 00	(7.05401D + 01, 7.40000D + 01)

Écart interquartile

Estimation	E.T.
3.2045141D + 01	4.2434890D + 00

grand nombre de paramètres statistiques ont été calculés pour les analyses B et C que pour l'analyse A. Quant à l'analyse D, elle a nécessité quatre lectures de données, soit deux pour chaque analyse à une variable.

REMERCIEMENTS

Outre le U.S. Census Bureau, le Soil Conservation Service du Département de l'agriculture de É.-U. a apporté son soutien au projet PC CARP. Nous tenons à remercier les arbitres et les éditeurs pour leurs commentaires utiles. Les personnes dans les pays en voie de développement qui s'intéressent au programme PC CARP sont priées de s'adresser à: PC CARP, International Statistical Programs Center, U.S. Bureau of the Census, Washington (D.C.), É.-U. 20233. Les autres personnes sont priées de s'adresser à: PC CARP, Statistical Laboratory, Iowa State University, Ames (Iowa), É.-U. 50011. Des exemplaires du programme ont été offerts à des organismes répartis dans plus de 20 pays.

BIBLIOGRAPHIE

BARR, A.J., GOODNIGHT, J.H., SALL, J.P., BLAIR, W.H., et CHILKO, D.M. (1979). SAS User's Guide. SAS Institute, Raleigh, North Carolina.

BINDER, D.A. (1983). On the variances of asymtoticaly normal estimators from complex surveys. *Revue Internationale de Statistique*, 51, 279-292.

COCHRAN, W.G. (1977). *Sampling Techniques*. New York: John Wiley.

DISKIN, B.A. (1985). Microcomputers in developing country statistical offices: current use and a look to the future. Actes de la 45^{e} assemblée de l'Institut international de statistique, Amsterdam.

FRANCISCO, C.A. (1987). Estimation of quantiles and the interquartile range in complex surveys. Thèse de doctorat non publiée, Iowa State University, Ames, Iowa.

FRANKEL, M.R. (1971). *Inference from Survey Samples*. Institute of Social Research, University of Michigan, Ann Arbor.

FULLER, W.A. (1975). Regression analysis for sample survey. Sankhyā C 37, 117-132.

FULLER, W.A., et SULLIVAN, G. (1987). Gamma Post Stratification. Report to the U.S. Bureau of the Census. Département de statistique, Iowa State University, Ames, (Iowa).

HERRAMAN, C. (1968). Sums of squares and products matrix. *Applied Statistics*, 17, 289-292.

HIDIROGLOU, M.A. (1974). Estimation of regression parameters for finite populations. Thèse de doctorat non publiée, Iowa State University.

HIDIROGLOU, M.A., FULLER, W.A., et HICKMAN, R.D. (1980). *SUPER CARP*, Département de statistique, Iowa State University, Ames, Iowa.

KISH, L. (1965). *Survey Sampling*. New York: John Wiley.

LONGLEY, J.W. (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.

PARK, H.J. (1987). Univariate Analysis in PC CARP. Creative Component pour le M.S. non publiée, Iowa State University, Ames, Iowa.

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

WOLTER, K.M. (1985). *Introduction to Variance Estimation*. New York: Springer-Verlag.