TECHNIQUES D'ENQUÊTE 1981, VOL. 7 NO. 2

L'INFÉRENCE STATISTIQUE BASÉE SUR DES PLANS D'ÉCHANTILLONNAGE COMPLEXES

Gad Nathan[1]

Les problèmes d'inférence statistique basée sur des plans d'échantillonnage complexes sont décrits. La définition du paramètre d'intérêt est de grande importance et on doit décider si on veut estimer un paramètre de la population à nombre fini ou un paramètre du modèle de superpopulation. Les méthodes générales basées sur la statistique généralisée de Wald et sa modification, aussi bien que la modification des statistiques classiques sont présentées. Les méthodes spécifiques pour la régression et les modèles linéaires et pour l'analyse des données qualitatives sont décrites en détail.

1. INTRODUCTION

Règle générale, l'application des méthodes classiques d'inférence, telles que la régression, l'analyse de variance ou les tests d'hypothèses, suppose que les observations proviennent d'un échantillon aléatoire simple d'une population infinie ayant une distribution de probabilité particulière à une famille hypothétique. La grande diffusion des programmes informatiques a rendu l'utilisation de ces méthodes particulièrement facile. Toutefois, il est presque impossible de les appliquer telles quelles à des données recueillies au moyen de plans d'échantillonnage complexes.

---

1 G. Nathan, Hebrew University, Jérusalem et Bureau de la statistique d'Israel.

Dans le présent article, nous tentons de donner quelques conseils pratiques sur ce qui peut et ne doit pas être fait dans ces cas. Cet examen est fondé sur des travaux récents (voir bibliographie) portant sur ces questions et comprend plusieurs exemples d'applications possibles.

Quiconque veut entreprendre une analyse statistique doit tout d'abord définir les paramètres qu'il faut estimer. Pour ce faire, on peut s'inspirer des travaux de Brewer et Mellor (1973) et de Smith (1976). De leur côté, Kish et Frankel (1974) soutiennent qu'une inférence valable doit porter sur des paramètres de la population finie, notamment le coefficient de régression d'une population:

$$B = \sum_{i=1}^{N} (X_i-\bar{X})(Y_i-\bar{Y}) \; / \; \sum_{i=1}^{N} (X_i-\bar{X})^2 \; ,$$

ou les coefficients de corrélation multiple ou partielle ou d'autres mesures qui sont définies par rapport à la population finie seulement, sans utiliser un modèle de superpopulation. Dans ce cas, l'inférence est basée sur un plan d'échantillonnage (Särndal, 1978), c'est-à-dire qu'elle est fondée uniquement sur les propriétés de la distribution d'échantillonnage. Cependant, on peut aussi utiliser l'inférence basée sur un modèle pour estimer un paramètre de la population finie (Hartley et Sielken, 1975).

Il existe une autre possibilité d'estimation dans ce domaine, celle que propose par exemple Fienberg (1980). Selon lui, l'inférence doit porter sur les paramètres d'une distribution de probabilité (superpopulation) dont la population finie constitue la réalisation. On peut trouver des exemples de ce type d'inférence dans les travaux de Konijn (1962), Fuller (1975), Thomsen (1978) et Pfeffermann et Nathan (1981). Si les paramètres se rapportent à un modèle de superpopulation, l'inférence ne peut être basée uniquement sur un plan d'échantillonnage; elle doit être fondée sur un modèle (Särndal, 1978) ou encore, sur un modèle et un plan d'échantillonnage en même temps. Si l'on suppose qu'il y a indépendance entre la distribution du modèle et la distribution d'échantillonnage, l'inférence classique (basée sur un modèle) est alors valide et le plan d'échantillonnage peut seul avoir un effet sur l'efficacité de l'inférence.

Ces deux démarches peuvent soulever de vives objections, car la méthode basée sur un modèle repose en grande partie sur des hypothèses appliquées à un modèle théorique qui ne peuvent généralement pas être garanties; par voie de conséquence, l'inférence ne sera pas robuste s'il y a déviation de ce modèle. Par ailleurs, dans le cas de l'inférence basée sur un plan d'échantillonnage, les paramètres d'une population finie sont habituellement des "copies" des paramètres d'un modèle théorique qui présentent une valeur descriptive très faible, à moins qu'il ne s'agisse d'un modèle de base. Par exemple, le coefficient de corrélation d'une population finie peut être une mesure utile de la relation qui existe entre deux variables, mais seulement si cette relation est approximativement linéaire.

Dans de nombreux cas, il est préférable de chercher un certain équilibre entre ces deux méthodes d'inférence. Il est possible d'y arriver si l'on utilise uniquement les paramètres d'une population finie qui constituent des éléments approximatifs des paramètres de la superpopulation d'un modèle pertinent auquel les données peuvent être appliquées. Par exemple, lorsque des équations de régression distinctes sont ajustées aux sous-ensembles appropriés, on obtient alors un meilleur ajustement linéaire que si l'on procède à une régression générale. Si les sous-ensembles sont de taille assez grande, les coefficients de régression de la population finie pourront mieux s'apparenter aux paramètres de la superpopulation et, ainsi, l'inférence portant sur les paramètres d'une population finie pourra être considérée comme caractérisant les paramètres de la superpopulation.

Pour pouvoir assurer la meilleure correspondance possible entre les paramètres d'un modèle et ceux d'une population finie, il faut procéder à une analyse exploratoire en profondeur du modèle avant d'aborder l'analyse théorique. Cette analyse exploratoire des divers modèles alternatifs peut, dans la plupart des cas, être fondée sur des mesures descriptives pour lesquelles le plan d'échantillonnage peut être pris en compte, ou encore sur des représentations graphiques. Toutefois, les résultats doivent être interprétés avec prudence en fonction du plan d'échantillonnage. Par exemple, un petit nombre de grands résidus ayant des coefficients de pondération de l'échantillon faibles peut être beaucoup moins important

qu'un grand nombre de petits résidus présentant des coefficients de pondération élevés. Un des outils d'analyse particulièrement utile dans le cas de la régression est la différence entre les coefficients de régression pondérée et non pondérée. Un écart important démontre souvent que le modèle ne convient pas.

Une fois les paramètres définis, il faut déterminer le genre d'inférence, soit l'estimation ponctuelle, l'estimation par intervalle ou les tests d'hypothèses. Il apparaît que l'estimation ponctuelle et les intervalles de confiance conviennent très bien aux paramètres d'une population finie, alors que les tests d'hypothèses, plus particulièrement d'hypothèses simples, sont valables uniquement lorsqu'ils sont appliqués à des paramètres d'une superpopulation d'un modèle très bien défini. Par exemple, l'hypothèse que les moyennes d'une paire d'ensembles sont égales peut être admise seulement s'il s'agit des moyennes d'une superpopulation et non des réalisations de leur population finie. Lorsqu'on veut éviter la formulation d'un modèle, il est recommandé de procéder à l'estimation ponctuelle ou de calculer les intervalles de confiance, plutôt que de faire des tests d'hypothèses, pour établir la différence entre les moyennes d'un ensemble. S'il faut appliquer ces tests aux paramètres d'une population finie, il est préférable de tester une hypothèse composée (par exemple tester si la différence entre les moyennes se situe à l'intérieur d'un ordre de valeurs) que de tester une hypothèse simple, c'est-à-dire vérifier si la différence est zéro. Il convient de souligner que, dans le cas d'échantillons assez grands, tout écart plus grand que zéro, peu importe s'il l'est à peine, sera considéré comme différent de zéro de façon significative.

Nous allons maintenant examiner certaines méthodes générales d'analyse des données provenant de plans d'échantillonnage complexes, de même que certaines autres méthodes particulières concernant les modèles linéaires et les tests de validité de l'ajustement et d'indépendance dans les tables de contingence. De façon générale, nous étudierons l'inférence se rapportant aux paramètres d'une population finie. Cependant, nous estimons que cette inférence est pertinente seulement si les paramètres s'apparentent aux paramètres du modèle de superpopulation, ce qui laisse à chacun la liberté

d'appliquer soit la méthode basée sur un plan d'échantillonnage soit celle basée sur un modèle, selon le degré de confiance que chacun a dans la validité d'un modèle sous-jacent.

2. MÉTHODES GÉNÉRALES DE BASE

2.1 Statistique généralisée de Wald

Si l'hypothèse qu'il faut tester est linéaire (ou peut être linéarisée) dans les valeurs prévues de statistiques asymptotiquement normales, pour lesquelles il existe un estimateur convergent de la matrice des variances, on peut alors utiliser la statistique généralisée de Wald (Grizzle, Starmer et Koch (1969), Koch, Freeman et Freeman (1976), Freeman, Freeman, Brock et Koch (1976), Shah, Holt et Folsom (1977) et Koch, Stokes et Brock (1980)).

Nous supposons que nous voulons tester l'hypothèse:

$$H_o: \quad X\beta = \theta_o, \tag{2.1.1}$$

où X est une matrice $r^X p$ connue de plein rang du plan d'échantillonnage. β est un vecteur inconnu $p^X 1$ des paramètres (soit des paramètres d'une population finie soit des paramètres d'une superpopulation) et θ_o est un vecteur connu $r^X 1$ des constantes. Si l'hypothèse n'est pas linéaire, on peut procéder à une approximation de premier ordre des séries de Taylor (Nathan, 1972, et Shuster et Downing, 1976).

Nous supposons qu'il y a un estimateur asymptotique normal convergent $\hat{\beta}$ de β ainsi qu'un estimateur convergent $(\hat{V})$ de la matrice des covariances de $\hat{\beta}$, dont la distribution est indépendante de celle de $\hat{\beta}$.

Alors, la statistique généralisée de Wald définie par:

$$X^2_w = (X\hat{\beta}-\theta_o)' \, (X\hat{V}X')^{-1} \, (X\hat{\beta}-\theta_o) \tag{2.1.2}$$

est distribuée asymptotiquement, selon l'hypothèse nulle, comme χ^2 ayant des degrés de liberté correspondant à la dimension de l'hypothèse (p-r).

La compatibilité de $\hat{\beta}$ et de $\hat{V}$ et les distributions asymptotiques de $\hat{\beta}$ et de X^2_W peuvent être étudiées en fonction de la distribution de la superpopulation.

La principale difficulté que présente cette démarche a trait au calcul de l'estimateur convergent $\hat{V}$ de la matrice des covariances, lorsque $\hat{\beta}$ n'est pas linéaire dans les données de l'échantillon (ce cas est très fréquent). Rao (1975) a étudié les diverses méthodes d'estimation de la variance qui peuvent être appliquées, notamment la linéarisation (Tepping, 1968), la répétition équilibrée (McCarthy, 1969) et la méthode dite "Jackknife" (Miller, 1974). Il existe également plusieurs programmes informatiques qui peuvent être utilisés dans ces cas, par exemple SUPERCARP (Hidiroglou, Fuller et Hickman, 1980), SUDAAN (Shah, 1978) pour la linéarisation et OSIRIS IV: PSALMS pour la répétition équilibrée. Une liste complète et une comparaison des programmes est présentée par Kaplan, Francis et Sedransk (1979).

Des comparaisons empiriques des estimateurs de la variance ont été établies par Kish et Frankel (1974) et par Richards et Freeman (1980), de même que des comparaisons théoriques, par Krewski et Rao (1981).

Toutefois, il convient de porter une attention particulière à la stabilité de l'estimateur de la variance, surtout lorsqu'il y a un grand nombre de paramètres. De plus, il faut aussi tenir compte des conditions dans lesquelles la convergence et les propriétés asymptotiques se réalisent, dans le cas des plans complexes. Par exemple, avec un plan d'échantillonnage à deux degrés, il est possible que les résultats asymptotiques commandent un grand nombre d'unités primaires d'échantillonnage (UPE) et un grand nombre d'unités du dernier degré par UPE.

2.2 Approximation et élaboration d'un modèle de covariances

Les problèmes courants liés au calcul d'un estimateur convergent stable de la matrice des covariances ont poussé les spécialistes à essayer d'appliquer des approximations simplifiées à ces estimateurs. Le principe de base est que si l'on suppose une certaine structure de la matrice des covariances, on peut alors utiliser des estimateurs plus stables d'un petit nombre de paramètres.

L'approximation peut être calculée à l'aide de la méthode basée sur un plan d'échantillonnage, directement en fonction de la matrice des covariances. Si l'on peut faire des hypothèses sur l'égalité des effets du plan d'échantillonnage sur les variances et covariances à l'intérieur d'un sous-groupe donné de paramètres, les estimateurs d'ensemble de la covariance peuvent alors être utilisés. Cette démarche est préconisée par Nathan (1973), Fuller et Rao (1978), Fellegi (1980) et Lepkowski et Landis (1980).

Par ailleurs, l'élaboration d'un modèle de la structure de la population proprement dite peut favoriser la construction de matrices de covariances simplifiées qui peuvent être facilement estimées (voir Altham (1976), Fuller et Battese (1973), Tomberlin (1979), Holt, Richardson et Mitchell (1980), Imrey, Sobel et Francis (1980) et Pfeffermann et Nathan (1981).

2.3 Modification des tests classiques

L'utilisation très répandue de programmes informatiques types a favorisé l'élaboration de nouveaux tests qui tiennent compte de plans d'échantillonnage complexes. Ces travaux peuvent être évalués sous l'angle d'une extension naturelle de l'utilisation des effets du plan d'échantillonnage comme facteurs multiplicatifs des variances basées sur un échantillon aléatoire simple d'une même taille, en vue d'apporter les modifications appropriées lorsqu'il s'agit de plans d'échantillonnage complexes.

La rectification peut certes être basée sur les effets du plan de divers estimateurs ou sur la moyenne des effets du plan (voir Cowan et Binder (1978), Fay (1979), Fellegi (1980), Rao et Scott (1981) et Scott et Holt (1981).

L'autre solution possible est d'étudier le comportement des fonctions des observations utilisées dans un test selon un certain modèle de superpopulation et de modifier en conséquence les fonctions types (Cohen, 1979, et Campbell, 1977).

3. MÉTHODES PARTICULIÈRES

3.1 Régression et modèles linéraires

La définition préalable du modèle et des paramètres d'intérêt est extrêmement importante dans le cas de l'analyse de la régression et des modèles linéaires. Ainsi, lorsque des relations de régression différentes doivent être calculées pour diverses strates ou UPE dans un plan d'échantillonnage à deux degrés, le paramètre d'intérêt peut être une simple moyenne des coefficients de régression (Konijn, 1962), une moyenne pondérée des coefficients (Pfeffermann et Nathan, 1981) ou leur valeur prévue (selon une certaine distribution préalable - Porter, 1973).

Règle générale, le modèle et les paramètres d'intérêt doivent être définis en fonction de la structure présumée de la population globale et ne doivent pas se rapporter à la structure du plan d'échantillonnage. Toutefois, dans de nombreux cas, le plan d'échantillonnage reproduit la structure de la population de sorte que les variables du plan font partie du modèle. A titre d'exemple, considérons le modèle:

$$E(Y|X_1,X_2) = X_1 \beta_{1.2} + X_2 \beta_{2.1} \qquad (3.1.1)$$

où X_1 comprend uniquement les variables qui ne se rapportent pas au plan d'échantillonnage et X_2 représente toutes les variables incluses dans le

plan d'échantillonnage complexe, c'est-à-dire que la distribution de l'échantillonnage dépend seulement de X_2:

$$P(s|X_1,X_2) = P(s|X_2). \tag{3.1.2}$$

L'estimation de $\beta_{1.2}$ et de $\beta_{2.1}$ dans la définition (3.1.1) et l'inférence qui en est établie peuvent être calculées selon la méthode classique, comme si l'échantillonnage était aléatoire simple, pour autant que le terme (3.1.1) se vérifie.

Cependant, si les variables du plan, X_2, ne sont pas incluses dans l'équation de régression d'intérêt:

$$E(Y|X_1) = X_1\beta_1 \tag{3.1.3}$$

et que la variable du plan, X_2, est en corrélation avec Y (fonction conditionnelle de X_1), alors l'estimation type par la méthode des moindres carrés ordinaires de β_1 n'est pas convergente (voir Nathan et Holt (1980) et Holt et Smith (1979) qui préconisent l'utilisation d'estimations modifiées pondérées et non pondérées de β_1, qui sont convergentes). Holt, Smith et Winter (1980) donnent un exemple d'application de ces estimateurs.

Si le modèle linéaire:

$$E(Y_i|x_i) = x_i'\beta \tag{3.1.4}$$

$$\text{cov}(Y_i,Y_j|x_i,x_j) = \begin{cases} \sigma^2 & i=j \\ 0 & i\neq j \end{cases} \tag{3.1.5}$$

est vrai pour toutes les unités (i, j=1, ..., N) d'une population finie et x_i, le vecteur colonne p^X1, comprend toutes les variables du plan d'échantillonnage, alors l'estimation non pondérée des moindres carrés ordinaires:

$$\hat{\beta} = (X_n' X_n)^{-1} X_n' Y_n \tag{3.1.6}$$

basée sur les valeurs échantillonnées

$$X'_{\underset{pxn}{n}} = (x_1, \ldots, x_n) \text{ et } Y'_n = (Y_1, \ldots, Y_1)$$

produit le "meilleur" estimateur de β sans biais d'un modèle linéaire, peu importe le plan d'échantillonnage. On le qualifie de meilleur parce qu'il définit une variance minimum basée sur le modèle. Toutefois, $\hat{\beta}$ n'est pas, de façon générale, un estimateur sans biais basé sur le plan d'échantillonnage, ni même un estimateur convergent (basé sur le plan) du paramètre de la population:

$$B = (X'_N X_N)^{-1} X'_N Y_N, \tag{3.1.7}$$

où $X'_{\underset{p^xN}{N}} = (x_1, \ldots, x_N)$ et $Y'_N = (Y_1, \ldots, Y_N)$.

L'estimateur convergent de B selon le plan d'échantillonnage est l'estimateur pondéré:

$$\hat{\beta}_W = (X'_n W_n X_n)^{-1} X'_n W_n Y_n, \tag{3.1.8}$$

où la matrice de pondération $W_n = \text{diag}\,(\pi_1^{-1}, \ldots, \pi_n^{-1})$, est la matrice diagonale n^xn des inverses des probabilités d'inclusion de l'échantillon $\pi_i = Pr\,(i \epsilon s)$

De toute évidence, la compatibilité de $\hat{\beta}_W$, comme estimateur de , ne dépend pas de ce que le modèle (3.1.4) se vérifie; en outre, la pertinence de l'estimation de B, lorsque le modèle ne se vérifie pas, peut être mise en doute. On peut démontrer que, si certaines conditions se réalisent dans un modèle non-linéaire, ce qui suppose que l'espérance mathématique conditionnelle de Y (selon X) est une fonction différenciable de X, l'espérance de B basée sur le modèle peut être exprimée approximativement comme une moyenne pondérée des pentes de cette fonction aux points X_i (les coefficients de pondération dépendant seulement de $X_i - \bar{X}$). Cependant, dans la pratique, cette interprétation a une valeur limitée.

De toute façon, $\hat{\beta}_W$ est un estimateur sans biais de β basé sur le modèle, chaque fois que le modèle (3.1.4) se vérifie. Règle générale, il ne sera pas un estimateur optimal de β selon (3.1.5) lorsqu'il y a échantillonnage à probabilité inégale, mais il le sera si la variance conditionnelle du modèle de Y_i est proportionnelle à Π_i, soit:

$$V(Y_i | x_i) = k\Pi_i \quad . \tag{3.1.9}$$

Comme l'estimateur pondéré $\hat{\beta}_W$ est plus robuste que l'estimateur non pondéré $\hat{\beta}$, en ce sens qu'il est à la fois un estimateur sans biais de β basé sur le modèle, si le modèle est vrai, et un estimateur convergent de B basé sur le plan d'échantillonnage, si le modèle ne se vérifie pas, il est recommandé d'utiliser l'estimateur pondéré $\hat{\beta}_W$ pour estimer B, chaque fois qu'on n'est pas sûr si le modèle (3.1.4) - (3.1.5) se vérifie. Il reste alors aux spécialistes à déterminer si B est un paramètre valable à estimer.

Il convient de souligner que, dans le cas de plans d'échantillonnage autopondérés, $\hat{\beta}$ et $\hat{\beta}_W$ correspondent. L'estimateur $\hat{\beta}_W$ (3.1.8) peut être calculé directement à partir de programmes informatiques types qui définissent la régression pondérée (par exemple BMDP) à l'aide des coefficients de pondération $1/\Pi_i$, ou encore à l'aide d'autres programmes (par exemple SPSS) qui affectent la régression non pondérée aux variables transformées $Y_i/\sqrt{\Pi_i}$ et $x_i/\sqrt{\Pi_i}$, mais pas aux variables pondérées Y_i/Π_i, x_i/Π_i. Cependant, il est bon de noter que, sous l'hypothèse alternative, les variances et covariances des estimateurs sont incorrectes et que les tests de signification usuels (par exemple, les tests F) ne sont pas valides et peuvent fausser grandement les conclusions.

Soit le modèle (3.1.4) - (3.1.5), la variance du modèle de $\hat{\beta}$ est:

$$V(\hat{\beta} | X_n) = \sigma^2 (X_n' X_n)^{-1} , \tag{3.1.10}$$

ce qui est le résultat fourni par les programmes de régresssion non pondérée. Cependant, la variance du modèle $\hat{\beta}_W$ est:

$$V(\hat{\beta}_W | X_n) = \sigma^2 (X_n' W_n X_n)^{-1} X_n' W_n' W_n X_n (X_n' W_n X_n)^{-1} \quad . \tag{3.1.11}$$

Le programme de régression pondérée, dont les coefficients sont $1/\Pi_i$, définit une valeur de $(X_n' W_n X_n)^{-1}$ de la variance du modèle de $\hat{\beta}_W$, ce qui correspond à (3.1.11) seulement si $W_n = I_n$. Par conséquent, aucun résultat concernant les erreurs types ou les tests d'hypothèses n'est exact.

Toutefois, l'estimateur du coefficient de corrélation multiple défini à l'aide de la régression pondérée:

$$\hat{R}^2 = \frac{(Y_n - X_n \hat{\beta}_W)' W_n (Y_n - X_n \hat{\beta}_W)}{(Y_n - \bar{y}_n \underset{-n}{1})' W_n (Y_n - \bar{y}_n \underset{-n}{1})} , \qquad (3.1.12)$$

où $\bar{y}_n = (\Sigma_s Y_i/\Pi_i) / (\Sigma_s 1/\Pi_i)$ est un estimateur convergent basé sur un plan d'échantillonnage du coefficient de corrélation multiple de la population:

$$R^2 = \frac{(\underset{N}{Y} - \underset{N}{X} B)' (\underset{N}{Y} - \underset{N}{X} B)}{(\underset{N}{Y} - \bar{Y}_N \underset{M}{1})' (\underset{N}{Y} - \bar{Y}_N \underset{-N}{1})} \qquad (3.1.13)$$

où $\bar{Y}_N = (1/N) \underset{-N}{1}' \underset{N}{Y}$.

La variance basée sur un plan de $\hat{\beta}_W$, lequel doit être considéré comme la mesure appropriée de la précision de $\hat{\beta}_W$ en tant qu'estimateur de B, ne peut pas en général être calculée uniquement à l'aide des probabilités d'inclusion de premier ordre Π_i. Pour la plupart des plans d'échantillonnage appliqués couramment, la variance de $\hat{\beta}_W$ basée sur un plan est estimée à l'aide d'une des méthodes d'estimation de la variance mentionnées plus haut, soit la linéarisation, la répétition équilibrée ou le "jackknife" (voir Jonrup et Rennermalm (1976) et Holt et Scott (1981)).

3.2 Analyse des données qualitatives

La méthode la plus simple d'analyse des données qualitatives est celle qui consiste à construire une classification unique de la population en k classes comportant des probabilités (fréquences relatives)

$\underline{p}' = (p_1, \ldots, p_{k-1})$. Lorsqu'on veut tester l'hypothèse nulle de la validité de l'ajustement d'une distribution connue, $\underline{p}_o = (p_{o1}, \ldots, p_{ok-1})$:

$$H_o: \underline{p} = \underline{p}_o \ , \tag{3.2.1}$$

on peut suivre les démarches expliquées au chapitre 2.

Nous supposons qu'il existe un estimateur d'enquête convergent $\hat{\underline{p}}' = (\hat{p}_1, \ldots, \hat{p}_{k-1})$ de $\underline{p}'$. S'il est asymptotiquement normal:

$$\sqrt{n}\ (\hat{\underline{p}} - \underline{p}) \rightsquigarrow N(\underline{0}, V) \tag{3.2.2}$$

et qu'il existe un estimateur convergent $\hat{V}$ de V, alors la statistique généralisée de Wald:

$$X_W = n(\hat{\underline{p}} - \underline{p}_o)'\ \hat{V}^{-1}(\hat{\underline{p}} - \underline{p}_o) \ , \tag{3.2.3}$$

qui est distribuée asymptotiquement par χ^2_{k-1} sous H_o, peut être utilisée pour tester H_o.

Pour bon nombre de plans d'échantillonnage simples, des estimateurs convergents de V sont obtenus directement alors que, pour les plans plus complexes, ils peuvent être calculés à l'aide des méthodes classiques. Cependant, si des tests de validité de l'ajustement doivent être effectués pour une variété de variables et de classifications, il est alors préférable d'utiliser la variable χ^2 type:

$$\chi^2 = n \sum_{i=1}^{k} (\hat{p}_i - p_{oi})^2/p_{oi} = n(\hat{\underline{p}}-\underline{p}_o)P_o^{-1}(\hat{\underline{p}}-\underline{p}_o), \tag{3.2.4}$$

où $P_o = \text{diag}\ (\underline{p}_o) - \underline{p}_o\underline{p}_o'$, tout en y apportant les modifications qui s'imposent. Rao et Scott (1981) ont démontré que la distribution asymptotique χ^2 sous H_o est celle de la somme pondérée de k-1 variables indépendantes χ^2 ayant chacune un degré de liberté.

$$\chi^2 \rightsquigarrow \sum_{i=1}^{k-1} \lambda_i Z_i^2 \ ; \quad Z_i^2 \sim N(0,1) \quad \text{indépendantes} \tag{3.2.5}$$

où $\lambda_1, \ldots, \lambda_{k-1}$ sont les valeurs propres de

$$D = P_o^{-1} V \quad (\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_{k-1} > 0). \tag{3.2.6}$$

On peut alors procéder à un test prudent de (3.2.1) en utilisant la variable X^2/λ_1 et une distribution χ^2_{k-1}. λ_1 peut représenter l'effet maximum du plan sur toutes les combinaisons linéaires des composantes de $\hat{\underline{p}}$. Par exemple, avec un échantillonnage stratifié proportionnel, $\lambda_1 \leq 1$ de sorte que X^2 peut être utilisé comme une fonction conservative des observations dans le test.

Dans les autres cas, l'utilisation de $X^2/\bar{\lambda}$

soit $$\bar{\lambda} = \frac{1}{k-1} \sum_{i=1}^{k-1} \lambda_i = \frac{1}{k-1} \sum_{i=1}^{k} d_i(1-p_i),$$

où $d_i = V[\hat{p}_i]/[p_i(1-p_i)]$ est l'effet du plan pour $\hat{p}_i$, s'est révélée un bon test d'approximation (Hidiroglou et Rao (1981), pour l'Enquête Santé Canada, et Holt, Scott et Ewings (1980), pour des enquêtes à grande échelle au Royaume-Uni). Une

approximation alternative , $X^2/\bar{d}$, où $\bar{d} = k^{-1} \sum_{i=1}^{k} d_i$, a été

proposée par Fellegi (1980).

La formulation directe de modèles pour $\underline{p}$ a été présentée par Altham (1976) et par Cohen (1976), mais leurs modèles comportent de sérieuses limitations puisqu'ils supposent que
$\lambda_1 = \lambda_2 = \ldots = \lambda_{k-1} = \bar{\lambda}$, ce qui équivaut à un effet constant du plan sur les catégories. Cela ne constitue pas généralement une hypothèse réaliste et a pour résultat que $X^2/\bar{\lambda}$ contient une distribution asymptotique χ^2_{k-1}. Pour effectuer un test d'indépendance dans une table de contingence, les hypothèses peuvent être exprimées:

$$H_o: h_{ij}(\underline{p}) = p_{ij} - p_{i+} p_{+j} = 0$$
$$(i=1, \ldots, r-1; \; j=1, \ldots, c-1), \tag{3.2.7}$$

où p_{ij} est la probabilité de distribution de la population dans la case (i, j), p_{i+}, p_{+j} sont les probabilités marginales et $\underline{p}' = (p_{11}, \ldots, p_{rc-1})$. La statistique généralisée de Wald appliquée au test de H_o est notée:

$$X^2_{WI} = n[\underline{h}(\hat{\underline{p}})]' \hat{V}_h^{-1} \underline{h}(\hat{\underline{p}}), \tag{3.2.8}$$

où $[\underline{h}(\hat{\underline{p}})]' = [h_{11}(\hat{\underline{p}}), \ldots, h_{r-1\ c-1}(\hat{\underline{p}})]$ et $\hat{V}_h/n$ est un estimateur convergent de la matrice des covariances de $\underline{h}(\hat{\underline{p}})$. Des versions de (3.2.8) ont été appliquées à des plans d'échantillonnage particuliers de même que diverses méthodes d'estimation de $\hat{V}_{h/n}$ ont été utilisées par Garza-Hernandez et McCarthy (1962), Nathan (1969, 1975) Shuster et Downing (1976) et Fellegi (1980).

Une variable modifiée s'apparentant à $X^2/\bar{\lambda}$ a été introduite par Rao et Scott (1981):

$$X^2_{CI} = (n/\hat{\delta}) \sum_{i=1}^{r} \sum_{j=1}^{c} (\hat{p}_{ij} - \hat{p}_{i+}\ \hat{p}_{+j})^2 / (\hat{p}_{i+}\ \hat{p}_{+j}), \tag{3.2.9}$$

où $\hat{\delta} = \frac{1}{(r-1)(c-1)} \sum_{i=1}^{r} \sum_{j=1}^{c} \hat{v}_{ij}(\underline{h})/(\hat{p}_{i+}\ \hat{p}_{+j})$ et

$\hat{v}_{ij}(\underline{h})/n$ est un estimateur convergent de la variance de $h_{ij}(\hat{\underline{p}})$. $\hat{\delta}$ peut être exprimé en fonction des effets du plan estimés de $h_{ij}(\hat{\underline{p}})$:

$$\hat{\delta} = \frac{1}{(r-1)(c-1)} \sum_{i=1}^{r} \sum_{j=1}^{c} (1-\hat{p}_{i+})(1-\hat{p}_{+j})\ \hat{\delta}_{ij}, \tag{3.2.10}$$

où $\hat{\delta}_{ij}$ est un estimateur de l'effet du plan d'échantillonnage, δ_{ij}, de $h_{ij}(\hat{\underline{p}})$:

$$\delta_{ij} = nV[h_{ij}(\hat{\underline{p}})]/[p_{i+}\ p_{+j}\ (1-p_{i+})(1-p_{+j})]\ . \tag{3.2.11}$$

Il peut être plus facile d'estimer les effets du plan d'échantillonnage que d'estimer les variances.

Des travaux de recherche empiriques menés par Holt, Scott et Ewings (1980) et par Hidiroglou et Rao (1981) démontrent que la distribution de X^2_{CI} s'apparente à $\chi^2(r-1)(c-1)$.

3.3 Autres genres d'analyses

Les modèles linéaires, de même que les tests de validité de l'ajustement et d'indépendance se prêtent à de nombreuses applications d'analyses, alors que d'autres genres d'analyses telles que l'analyse en composantes principales, l'analyse factorielle et discriminante, l'analyse de corrélation, la régression logistique, les modèles logarithmiques linéaires, les méthodes non paramétriques, etc. ne peuvent être appliquées aussi facilement. Bien que les méthodes générales décrites au chapitre deux puissent être utilisées, leur application pose certaines difficultés et, de fait, on n'a signalé que très peu de cas possibles.

Étant donné que les coefficients de corrélation sont un élément de base dans la plupart des analyses à plusieurs variables, un certain nombre d'études empiriques sur l'effet du plan d'échantillonnage sur l'estimation de ces coefficients ont été menées par Kish et Frankel (1974), Bebbington et Smith (1977) et Holt, Richardson et Mitchell (1980). Quoiqu'on ne puisse tirer de conclusions générales de leurs travaux, il apparaît que les effets du plan d'échantillonnage sont loin d'être négligeables. Bebbington et Smith (1977) ont également étudié la variabilité échantillonnale des estimateurs des composantes principales.

Dans d'autres domaines également, l'effet du plan d'échantillonnage sur les logits a été examiné par Lepkowski et Landis (1980) et les intervalles de confiance des quantiles, par Woodruff (1952) et par Sedransk et Meyer (1978).

REMERCIEMENTS

L'auteur tient à remercier MM. D. Binder, N. Chinnappa, S.E. Fienberg, M. Hidiroglou, C.J.C. Hole, J.N.K. Rao et A. Scott de lui avoir consacré du temps et fait part de leurs commentaires.

BIBLIOGRAPHIE

[1] Altham, P.M.E. (1976), "Discrete Variable Analysis for Individuals Grouped into Families", Biometrika, 63, 263-269.

[2] Brewer, K.R. and Mellor, R.W. (1973), "The Effect of Sample Structure on Analytical Surveys", Aust. J. Statist., 15, 145-152.

[3] Bebbington, A.C. and Smith, T.M.F. (1977), "The Effect of Survey Design on Multivariate Analysis", The Analysis of Survey Data (C.A. O'Muircheartaich and C. Payne, Editors), Vol. 2. Model Fitting", New York: Wiley, 175-192.

[4] Campbell, C. (1977), "Properties of Ordinary and Weighted Least Squares Estimators for Two Stage Samples", Proc. Soc. Statist. Sect., Amer. Statist. Assoc., 800-805.

[5] Cohen, J.E. (1976), "The Distribution of the Chi-Squared Statistic Under Clustered Sampling", J. Amer. Statist. Assoc. 71, 665-670.

[6] Cowan, J. and Binder, D.A. (1978), "L'effet d'un plan d'échantillonnage à deux degrés sur les tests d'indépendance", Techniques d'enquête 4, n° 1, 16-29.

[7] Fay, R.E. (1979), "On Adjusting the Pearson Chi-Square Statistic for Clustered Sampling", Proc. Soc. Statist. Sect., Amer. Statist. Assoc. 402-405.

[8] Fellegi, I.P. (1980), "Approximate Tests of Independence and Goodness of Fit Based on Stratified Multistage Samples", J. Amer. Statist. Assoc. 75, 261-268.

[9] Fienberg, S.E. (1980), "The Measurement of Crime Victimization: Prospects for Panel Analysis of a Panel Survey", The Statistician, 29, 313-350.

[10] Fuller, W.A. (1975), "Regression Analysis for Sample Survey", Sanklya, 22, 117-132.

[11] Fuller, W.A. and Battese, C.E. (1973), "Transformations for Estimation of Linear Models with Nested-Error Structure", J. Amer. Statist. Assoc. 68, 626-632.

[12] Fuller, W.A. and Rao, J.N.K. (1978), "Estimation for a Linear Regression Model with Unknown Diagonal Covariance Matrix", Ann. Statist. 6., 1149-1158.

[13] Freeman, D.H. Jr., Freeman, J., Brock, D.B. and Koch, G.G., "Strategies in the Multivariate Analysis of Data from Complex Surveys II: An Application to the United States National Health Interview Survey", Inter. Statist. Rev. 44, 317-330.

[14] Garza-Hernandez, T. and McCarthy, P.J. (1962), "A Test of Homogeneity for a Stratified Sample", Proc. Soc. Statist. Sect., Amer. Statist. Assoc., 200-202.

[15] Grizzle, J.E., Starmer, C.F. and Koch, G.G. (1969), "Analysis for Categorical Data by Linear Models", Biometrics, 25, 489-504.

[16] Hartley, H.O. and Sielken, R.L. (1975), "A Superpopulation Viewpoint for Finite Population Sampling", Biometrics, 31, 411-422.

[17] Hidiroglou, M.A., Fuller, W.A. and Hickman, R.D. (1980), "Super Carp: Sixth Edition", Statistical Laboratory Survey Section, Iowa State University, Ames, Iowa.

[18] Hidiroglou, M.A. and Rao, J.N.K. (1981), "Chi-Square Tests for the Analysis of Categorical Data from the Canada Health Survey", Invited Paper for 43rd Session of I.S.I., Buenos-Aires.

[19] Holt, D. Richardson, S.C. and Mitchell, P.W. (1980), "The Analysis of Correlations In Complex Survey Data", (non publié).

[20] Holt, D. and Scott, A.J. (1981), "Regression Analysis Using Survey Data", The Statistician, 30. (à paraître).

[21] Holt, D. Scott, A.J., and Ewings, P.O. (1980), "Chi-Squared Test with Survey Data", J. Roy. Statist. Soc. A, 143, 302-330.

[22] Holt, D. and Smith, T.M.F. (1979), "Regression Analysis of Data from Complex Surveys", Roy. Statist. Soc. Conf. Oxford.

[23] Holt, D., Smith, T.M.F. and Winter, P.O. (1980), "Regression Analysis of Data from Complex Surveys", Jour. Roy. Statist. Soc. A, 143, 474-483.

[24] Imvrey, P., Sobel, E. and Francis, M. (1980), "Modeling Contingency Tables from Complex Surveys", Proc. Sect. Survey Meth., Amer. Statist. Assoc., 213-217.

[25] Jonrup, H. and Rennermalm, B. (1976), "Regression Analysis in Samples from Finite Population", Scand. Jour. Statist., 3, 33-37.

[26] Kaplan, B., Francis, I., and Sedransk, J. (1979), "A Comparison of Methods and Programs for Computing Variances of Estimators from Complex Sample Surveys", Proc. Sect. Survey Meth., Amer. Statist. Assoc., 97-100.

[27] Kish, Leslie and Frankel, M.R. (1970), "Balanced Repeated Replication for Standard Errors", J. Amer. Statist. Assoc., 65, 1071-1094.

[28] Kish, L. and Frankel, M.R. (1974), "Inference from Complex Samples", (avec commentaires), J. Roy. Statist. Soc. B, 36, 1-37.

[29] Koch, G.G., Freeman, D.J., Jr., and Freeman, J.L. (1975), "Strategies in The Multivariate Analysis of Data from Complex Surveys", Inter. Statist. Rev. 43, 59-78.

[30] Koch, G.G., Stokes, M.E. and Brock, D. (1980), "Applications of Weighted Least Squares Methods for Fitting Variational Models to Health Survey Data", Proc. Sect. Survey Meth., Amer. Statist. Assoc., 218-223.

[31] Konijn, H.S. (1962), "Regression Analysis for Sample Surveys", J. Amer. Statist. Assoc. 57, 590-606.

[32] Krewski, D., and Rao, J.N.K. (1981), "Inference from Stratified Samples: Properties of the Linearization, Jackknife and Balanced Repeated Replication Methods", Ann. Statist., 9 (5/ 1010-1019).

[33] Lepkowski, J.M. and Landis, J.R. (1980), "Design Effects for Linear Contrasts of Proportions and Logits", Proc. Sect. Survey Meth., Amer. Statist. Assoc., 224-229.

[34] McCarthy, P.J. (1969), "Pseudo-Replication: Half-Samples", Inter. Statist. Rev. 37, 239-264.

[35] Miller, R.G. (1974), "The Jackknife--A Review", Biometrika 61, 1-15.

[36] Nathan, G. (1969), "Tests of Independence in Contingency Tables from Stratified Samples", New Developments in Survey Sampling (N.L. Johnson and H. Smith, eds.), New York: Wiley, 578-600.

[37] Nathan, G. (1972), "On the Asymptotic Power of Tests for Independence in Contingency Tables from Stratified Samples", J. Amer. Statist. Assoc., 67, 917-920.

[38] Nathan, G. (1973), "Approximate Tests of Independence in Contingency Tables from Complex Stratified Samples", National Center for Health Statistics, Vital and Health Statistics Series 2, No. 53, Washington, D.C.

[39] Nathan, G. (1975), "Tests of Independence in Contingency Tables from Stratified Proportional Samples", Sankhya C, 37, 77-87. [erratum: Sankhya C, 40, (1978), 190].

[40] Nathan, G. and Holt, D. (1980), "The Effect of Survey Design on Regression Analysis", J. Roy. Statist. Soc. B, 42, 377-386.

[41] Pfeffermann, D., and Nathan, G. (1981), "Regression Analysis of Data from Complex Samples", J. Amer. Statist. Assoc. 76, 681-689.

[42] Porter, R.M. (1973), "On the Use of Survey Sample Weights in the Linear Model", Annals of Economic and Social Measurement, 2, 141-158.

[43] Rao, J.N.K. (1975), "Analytic Studies of Sample Survey Data", Survey Methodology, Vol. 1, Supplementary Issue.

[44] Rao, J.N.K. and 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", J. Amer. Statist. Assoc. 76, 221-230.

[45] Richards, V. and Freeman, D.H. Jr. (1980), "A Comparison of Replicated and Pseudo-Replicated Covariance Matrix Estimators for the Analysis of Contingency Tables", Proc. Sect. Survey Meth., Amer. Statist. Assoc., 209-211.

[46] Särndal, C.E. (1978), "Design-Based and Model-Based Inference in Survey Sampling", Scand. J. Statist., 5, 27-52.

[47] Sedransk, S. and Meyer, J. (1978), "Confidence Intervals for the Quantiles of a Finite Population: Simple Random and Stratified Simple Random Sampling", J. Roy. Statist. Soc. B., 40, 239-252.

[48] Scott, A. and Holt, D. (1981), "The effect of Two-Stage Sampling on Ordinary Least Squares Methods", (non publié).

[49] Shah, B.V. (1978), "SUDAAN: Survey Data Analysis Software", Proc. Statist. Comp. Sect., Amer. Statist. Assoc., 146-151.

[50] Shah, B.V., Holt, M.M. and Folsom, R.E. (1977), "Inference About Regression Models from Sample Survey Data", Bull. Inter. Statist. Inst. 47, Bk. 3, 43-57.

[51] Shuster, J.J. and Downing, D.J. (1976), "Two-Way Contingency Tables for Complex Sampling Schemes", Biometrika 63, 271-278.

[52] Smith, T.M.F. (1976), "The Foundations of Survey Sampling: A Review (Avec Commentaires), J. Roy. Statist. Soc. A., 139, 183-195.

[53] Tepping, B.J. (1968), "The Estimation of Variance in Complex Surveys", Proc. Soc. Statist. Sect., Amer. Statist. Assoc., 66, 411-414.

[54] Thomsen, I. (1978), "Design and Estimation Problems When Estimating a Regression Coefficient From Survey Data", Metrika 25, 27-35.

[55] Tomberlin, T.J. (1979), "The Analysis of Contigency Tables of Data from Complex Samples", Proc. Sect. Survey Meth., Amer. Statist. Assoc., 152-157.

[56] Woodruff, Ralph S. (1952), "Confidence Intervals for Medians and Other Position Measures", J. Amer. Statist. Assoc. 47, 635-646.