Techniques d'enquête, décembre 1984
Vol. 10, no. 2, pp. 155–170
Statistique Canada

Analyse de données qualitatives d'enquêtes Complexes: Quelques expériences canadiennes[1]

D.A. Binder, M. Gratton, M.A. Hidiroglou, S. Kumar et J.N.K. Rao[2]

RÉSUMÉ

Les tests de la qualité de l'ajustement, les tests d'indépendance dans un tableau de contingence à deux dimensions, les modèles log-linéaires et les modèles de régression logistique sont examinés par rapport aux échantillons d'enquêtes basées sur un plan de sondage complexe. On relève quelques approximations des lois sous l'hypothèse nulle et on présente des exemples à partir de données de l'Enquête Santé Canada et de l'Enquête sur la Population Active du Canada. Il est aussi brièvement question des possibilités de mise en oeuvre d'un progiciel pour les méthodes étudiées.

KEYWORDS: χ^2 statistic; Wald Statistics; Goodness of fit; Independence in two-way tables; Log-linear and logistic regression model.

1. INTRODUCTION

Un résumé du développement des techniques modernes d'analyse des données qualitatives a été présenté dans un excellent document d'étude par Imrey, Koch et Stokes (1981). Comme ces techniques ont été conçues pour des échantillons aléatoires d'unités tirées indépendament et ayant une même fonction de répartition, elles ne s'appliquent pas directement aux échantillons d'enquêtes basées sur un plan de sondage complexe.

Koch *et coll.* (1975) et Shuster et Downing (1976) ont élaboré des méthodes asymptotiques qui reposent sur la variable de Wald et permettent de tenir compte du plan de sondage, mais elles exigent qu'on dispose du fichier complet de données ou, au moins, de la matrice des covariances estimées des estimations pour chaque case. Cohen (1976) et Altham (1976) ont proposé un modèle simple de classification automatique et démontré que la variable généralisée de Wald utilisée pour vérifier la qualité de l'ajustement est un multiple de la variable χ^2 lorsque ce modèle est valide. Brier (1978) a analysé un modèle semblable, mais s'est penché sur des hypothèses générales au sujet des probabilités dans chaque case et il a prouvé qu'un multiple de la variable correspondante de Pearson est asymptotiquement distribué comme une variable aléatoire χ^2 lorsque le modèle est valide. Fellegi (1980) a réduit la variable χ^2 à l'aide d'un facteur de correction calculé à partir de la moyenne des estimations des effets du plan de sondage. Fay (1985) a construit des variables χ^2 et G^2 pour une analyse fondée sur la méthode du jackknife; ces statistiques permettent également de tenir compte du plan de sondage, mais elles nous obligent de connaître les estimations obtenues dans chaque case au niveau des unités primaires d'échantillonnage. Rao et Scott (1981) ont décrit un moyen de rectifier la variable χ^2 ou (G^2) en utilisant l'approximation de Satterthwaite pour la distribution asymptotique de χ^2 et la matrice complète des covariances estimées.

---

[1] Cet article est une version augmentée et revisée de celui présenté lors du Séminaire sur les développements récents dans l'analyse des grands fichiers de données tenu à Luxembourg du 16 au 18 novembre 1983. Le séminaire était patronné par l'Office Statistique des Communautés Européennes.

[2] D.A. Binder, Division, des méthodes d'enquête-institutions et agriculture, M. Gratton, Planification et support informatique, M.A Hidiroglou, Division des méthodes d'enquêtes-entreprises, S. Kumar, Division des méthodes de recensement et d'enquête-ménages, Statistique Canada, Tunney's Pasture, Ottawa, Ontario, Canada, K1A 0T6, et J.N.K. Rao, Department of Mathematics and Statistics, Carleton University, Ottawa, Ontario, Canada.

On examine ici les problèmes relatifs à l'ajustement de modèles et aux tests d'hypothèses à partir de données qualitatives d'enquêtes complexes. Quand des données sont recueillies en fonction d'un plan de sondage complexe, il est nécessaire de modifier les méthodes classiques décrites par Imrey, Koch et Stokes pour faire des inférences valables. Si les tableaux publiés indiquent l'effet du plan de sondage sur les fréquences des cases et les fréquences marginales, il est possible de produire une approximation de la loi sous l'hypothèse nulle des statistiques des tests proposés sans disposer du fichier. Toutefois, si on dispose du fichier des données, on peut utiliser d'autres procédés qui seront décrits plus bas.

Pour donner une idée générale des modifications qu'il faut apporter aux méthodes habituelles, la section 2 aborde le problème classique de la qualité du test de la validité de l'ajustement. Ensuite, la section 3 examine les tests d'indépendance dans un tableau de contingence à deux dimensions. A la section 4, les modèles log-linéaires seront examinés d'une façon générale. Les modèles de régression logistique sont décrits à la section 5. La section 6 résume les travaux en cours pour l'élaboration d'un progiciel à Statistique Canada pour les méthodes étudiées ici et la section 7 porte sur la qualité de ces méthodes. Les exemples numériques sont tirés surtout de l'Enquête Santé Canada. Une application relative à l'Enquête sur la Population Active du Canada est présentée à la section 5.

2. QUALITÉ DE L'AJUSTEMENT

2.1 Échantillonnage multinomial

Supposons que nous choisissons n observations indépendantes, $Y_1, \ldots, Y_n$, qui ont une même distribution discrète à k catégories, où $\Pr(Y = i) = \pi_i$; $\Sigma_{i=1}^{k} \pi_i = 1$. Nous observons le vecteur aléatoire $\underset{\sim}{n} = (n_1, \ldots, n_{k-1})^T$, qui suit une loi multinomiale. Notre estimateur de $\pi = (\pi_1, \ldots, \pi_{k-1})^T$ est $\underset{\sim}{p} = \underset{\sim}{n}/n$. Cet estimateur est sans biais et sa matrice de variances-covariances est $\underset{\sim}{V}\{\underset{\sim}{p}\} = (\underset{\sim}{D}_\pi - \underset{\sim}{\pi}\underset{\sim}{\pi}^T)/n = \underset{\sim}{P}/n$, où $\underset{\sim}{D}_\pi = \text{diag}\ \{\pi_1, \ldots, \pi_{k-1}\}$. Notons que $\underset{\sim}{P}^{-1} = \underset{\sim}{D}_\pi^{-1} + (\underset{\sim}{1}\underset{\sim}{1}^T/\pi_k)$. Asymptotiquement, $n^{1/2}(\underset{\sim}{p} - \underset{\sim}{\pi}) \rightarrow N(\underset{\sim}{0}, \underset{\sim}{P})$. Pour une valeur donnée de $\underset{\sim}{\pi}_o$, le problème de la qualité de l'ajustement consiste à tester l'hypothèse:

$$H_o: \underset{\sim}{\pi} = \underset{\sim}{\pi}_o,$$

contre l'hypothèse

$$H_1: \underset{\sim}{\pi} \neq \underset{\sim}{\pi}_o. \tag{2.1}$$

Soit $\underset{\sim}{P}_o$ la valeur de $\underset{\sim}{P}$ associée à $\underset{\sim}{\pi}_o$, le statistique de Wald pour ce test est

$$\begin{aligned} W_1 &= n(\underset{\sim}{p} - \underset{\sim}{\pi}_o)^T \underset{\sim}{P}_o^{-1} (\underset{\sim}{p} - \underset{\sim}{\pi}_o) \\ &= n \sum_{i=1}^{K} \{(p_i - \pi_{io})^2/\pi_{io}\}, \end{aligned}$$

ce qui n'est pas autre chose que le test du Khi-carré de Pearson. Si H_o est vrai, cette statistique suit asymptotiquement une loi de type χ^2_{k-1}. Le test du rapport de vraisemblance pour ce problème repose sur l'expression

$$LR_1 = 2n \sum_{i=1}^{k} p_i \log(p_i/\pi_{io}).$$

Étant donné que $2p_i \log(p_i/\pi_{io})$ est asymptotiquement équivalent à $2(p_i - \pi_{io}) + (p_i - \pi_{io})^2/\pi_{io}$ si H_o, est vraie, on peut voir que le test du rapport de vraisemblance est asymptotiquement équivalent au test du khi-carré de Pearson sous l'hypothèse H_o.

Un autre test possible de l'hypothèse nulle est obtenu par le vecteur des logarithmes $\underset{\sim}{\mu}_o = \log \underset{\sim}{\pi}_o$ et $\underset{\sim}{\hat{\mu}} = \log \underset{\sim}{p}$. Si H_0 est vraie, $\underset{\sim}{\hat{\mu}} - \underset{\sim}{\mu}_o$ est asymptotiquement équivalent à $\underset{\sim}{D}_{\pi_o}^{-1}(\underset{\sim}{p} - \underset{\sim}{\pi}_o)$. Par conséquent, $n^{1/2}(\underset{\sim}{\hat{\mu}} - \underset{\sim}{\mu}_o) \rightarrow N(\underset{\sim}{0}, \underset{\sim}{D}_{\pi_o}^{-1} - \underset{\sim}{1}\underset{\sim}{1}^T)$ sous H_o et la statistique de Wald est

$$\begin{aligned} W_2 &= (\underset{\sim}{\hat{\mu}} - \underset{\sim}{\mu}_o)^T \left[\underset{\sim}{D}_{\pi_o} + (\underset{\sim}{\pi}_o \underset{\sim}{\pi}_o^T/\pi_{ko})\right](\underset{\sim}{\hat{\mu}} - \underset{\sim}{\mu}_o) \\ &\doteq \sum_{i=1}^{k} \pi_{io}(\hat{\mu}_i - \mu_{io})^2, \end{aligned}$$

où $\mu_{ko} = \log \pi_{ko}$ et $\hat{\mu}_k = \log p_k$.

Cette approximation est le résultat du fait que, si H_o est vraie,

$$\pi_{ko}(\hat{\mu}_k - \mu_{ko}) \doteq p_k - \pi_{ko}$$
$$= -(\underset{\sim}{p} - \underset{\sim}{\pi}_o)^T \underset{\sim}{1} - (\underset{\sim}{\mu} - \underset{\sim}{\mu}_o)^T \underset{\sim}{\pi}_o .$$

Notons que W_2 est aussi asymptotiquement équivalent à la statistique de test du Khi-carré de Pearson sous l'hypothèse nulle.

2.2 Autres méthodes d'échantillonnage

Ces résultats au sujet de W_1, W_2 et LR_1 sont biens connus. Le problème qui nous intéresse ici est celui de l'hypothèse plus générale selon laquelle $n^{1/2}(\underset{\sim}{p} - \underset{\sim}{\pi}) \rightarrow N(\underset{\sim}{0}, \underset{\sim}{V})$, où $\underset{\sim}{V}$ n'est pas nécessairement égale à $\underset{\sim}{P}$. Ici, $\underset{\sim}{p}$ est une estimation de $\underset{\sim}{\pi}$ qui est calculée à partir des données d'une enquête et qui peut dépendre des poids d'échantillonnage et d'autres facteurs de correction, ce qui est souvent le cas dans les enquêtes dont le plan de sondage est complexe. Nous supposons que $\hat{\underset{\sim}{V}}$ est un estimateur convergent de $\underset{\sim}{V}$. Nous examinerons maintenant deux procédés. Premièrement, on peut construire la statistique de Wald appropriée pour le plan de sondage en question. On obtient ainsi

$$W_3 = n(\underset{\sim}{p} - \underset{\sim}{\pi}_o)^T \hat{\underset{\sim}{V}}^{-1} (\underset{\sim}{p} - \underset{\sim}{\pi}_o),$$

où le rang de $\hat{\underset{\sim}{V}}$ est $k-1$, de sorte que W_3 suit asymptotiquement une loi de type χ^2_{k-1} si H_o est vraie.

Une deuxième possibilité consisterait à utiliser les fonctions W_1, W_2 ou LR_1 directement pour vérifier l'hypothèse nulle. On sait que, selon la théorie de la loi normale à plusieurs variables, la distribution de $n(\underset{\sim}{p} - \underset{\sim}{\pi}_o)^T \underset{\sim}{P}_o^{-1} (\underset{\sim}{p} - \underset{\sim}{\pi}_o)$ est celle de $\Sigma\, \delta_i Z_i^2$ où $\{Z_i^2\}$ est un ensemble de variables aléatoires indépendantes de type χ^2_1 et $\underset{\sim}{\delta} = (\delta_1, \ldots, \delta_{k-1})^T$ sont les valeurs propres de $\underset{\sim}{P}_o^{-1} \underset{\sim}{V}$; voir Johnson et Kotz (1970, pg. 150). Ce résultat a été démontré par Rao et Scott (1981), qui ont appelé les δ_i les effets généraux du plan de sondage. notons que, pour $k = 2$, $\delta = n\sigma_p^2 / \{\pi_o(1 - \pi_o)\}$, où $\sigma_p^2 = V\{p\}$, ce qui correspond à l'effet habituel du plan de sondage pour p si H_o est vraie.

2.3 Approximations

En général, la fonction de répartition de combinaisons linéaires de variables aléatoires de type χ^2_1 est plutôt compliquée, quoique leurs moments soient faciles à calculer. Rao et Scott (1981) ont proposé deux approximations pour le calcul des seuils des tests. Dans la première approximation, la distribution en cause est considérée comme étant proportionnelle à celle d'une variable aléatoire de type χ^2_{k-1}, et on obtient la constante de proportionnalité en égalant la moyenne de la distribution approximative et celle de la distribution théorique. On parvient ainsi au résultat suivant:

$$\sum_{i=1}^{k-1} \delta_i Z_i^2 \doteq \{\sum_{i=1}^{k-1} \delta_i / (k-1)\} \chi^2_{k-1} \tag{2.2}$$

Or,

$$\Sigma\, \delta_i = \mathrm{tr}(\underset{\sim}{P}_o^{-1} \underset{\sim}{V})$$
$$= \sum_{i=1}^{k} v_{ii} / \pi_{io}$$
$$= \sum_{i=1}^{k} d_i (1 - \pi_{io}),$$

La solution dépend donc seulement des effets du plan de sondage dans chaque case, $\{d_i\}$, où v_{ii} est le $i^{\text{ième}}$ élément de la diagonale de $\hat{\underset{\sim}{V}}$ et $d_i = v_{ii} / [\pi_{io}(1 - \pi_{io})]$. Cette approximation est particulièrement commode lorsqu'on ne dispose pas de la matrice de variances-covariances au complet, mais que les effets du plan de sondage dans chaque case sont connus, ce qui est souvent le cas dans les publications officielles.

Tableau 1

Répartition par âge des buveurs de 1 à 6 consommations par semaine.
Répartition par âge de la population canadienne projetée par le recensement (1978-1979)

	Âge							
Répartition de la population	15-19 .133	20-24 .127	25-34 .218	35-44 .152	45-54 .140	55-64 .115	65+ .115	Total 1.000
Répartition des buveurs de 1 à 6 cons./semaine	.117	.150	.264	.175	.148	.093	.053	1.000
Effet du plan de sondage	1.4	1.2	2.2	1.1	0.6	1.1	1.0	

Exemple 1

À partir des données de l'Enquête Santé Canada (1978-1979), qui est fondée sur un échantillon stratifié à plusieurs degrés, on a calculé la répartition par âge des buveurs de 1 à 6 consommations par semaine dans un échantillon de 5,204 personnes âgées de 15 ans et plus. Une description de cette enquête est présentée dans la publication "La santé des Canadiens" (n° 82-538F au catalogue de Statistique Canada).

Les résultats, que sont extraits d'une étude de Hidiroglou et Rao (1981), figurent au tableau 1. La valeur brute de W_1 est 298 et elle diminue à 248 après le calcul de l'approximation (2.2). Pour ces données, les rectifications par la stratification a posteriori des groupe d'âge et sexe produisent des effets de plan de sondage assez petits.

Une deuxième approximation de la distribution de $\Sigma\ \delta_i Z_i^2$ qui a été proposée par Rao et Scott (1981) est l'approximation de Satterthwaite (1946): $\Sigma\ \delta_i Z_i^2 \approx a\chi_\nu^2$. Pour calculer a et ν, il faut résoudre

$$\Sigma\ \delta_i^2 = \mathrm{tr}\{(\underset{\sim}{P}_{\cdot}^{-1}\hat{\underset{\sim}{V}})^2\}$$

$$= \sum_{i=1}^{k}\sum_{j=1}^{k} v_{ij}^2/(\pi_{io}\pi_{jo}).$$

Cependant, cette solution dépend de tous les éléments de la matrice $\hat{\underset{\sim}{V}}$. Le fait important ici à souligner est qu'il est nécessaire de corriger la statistique du test multinomial pour établir le seuil critique approprié.

Une autre approximation possible est celle proposée par Fellegi (1980) et qu'on calcule en divisant la statistique $n(p-\underset{\sim}{\pi}_o)^T\ \underset{\sim}{P}_o^{-1}(p-\underset{\sim}{\pi}_o)$ par la moyenne des effets du plan de sondage, $\bar{d}$, au lieu de la moyenne pondérée qui figure dans l'équation (2.2). L'effet de cette modificaiton sur les données du tableau 1 est que la valeur corrigée de la statistique du khi-carré devient 243, ce qui est comparable au résultat qu'on obtient avec l'approximation de Rao et Scott (1981).

3. TESTS D'INDEPENDANCE DANS UN TABLEAU A DEUX DIMENSIONS

3.1 Echantillonnage multinomial

Nous supposons maintenant que les catégories de la distribution multinomiale peuvent être recoupées pour former un tableau de dimension $r\times c$ dans lequel, pour l'observation du couple de variables (Y_1, Y_2), nous avons $\Pr(Y_1 = i,\ Y_2 = j) = \pi_{ij}$; $\Sigma_{i=1}^{r}\Sigma_{j=1}^{c}\ \pi_{ij} = 1$. On définit $\pi_{i+} = \Sigma_{j=1}^{c}\ \pi_{ij}$ et $\pi_{+j} = \Sigma_{i=1}^{r}\ \pi_{ij}$ et $\underset{\sim}{\pi} = (\pi_{11}, \ldots, \pi_{1c}, \ldots, \pi_{r1}, \ldots, \pi_{rc})^T$, $\underset{\sim}{\pi}_R = (\pi_{1+}, \ldots, \pi_{(r-1)+})^T$, $\underset{\sim}{\pi}_C = (\pi_{+1}, \ldots, \pi_{+(c-1)})^T$, $\underset{\sim}{P}_R = \underset{\sim}{D}_{\pi_R} - \underset{\sim}{\pi}_R\underset{\sim}{\pi}_R^T$. $\underset{\sim}{P}_C = \underset{\sim}{D}_{\pi C} - \underset{\sim}{\pi}_C\underset{\sim}{\pi}_C^T$. Nous observons le vecteur aléatoire $\underset{\sim}{n}$ de la distribution multinomiale, où $E\{\underset{\sim}{n}\} = n\underset{\sim}{\pi}$. On définit $\underset{\sim}{p} = \underset{\sim}{n}/n$, $p_{i+} = \Sigma_j p_{ij}$, et $p_{+j} = \Sigma_i\ p_{ij}$.

Nous voulons vérifier l'hypothèse d'indépendance

$$H_o\colon \pi_{ij} - \pi_{i+}\pi_{+j} = 0 \text{ for } 1 \leq i \leq r-1;\ 1 \leq j \leq c-1,$$

contre

$$H_1\colon \pi_{ij} - \pi_{i+}\pi_{+j} \neq 0 \text{ pour un couple quelconque } (i, j).$$

Si nous construisons la variable $h_{ij} = p_{ij} - p_{i+}p_{+j}$, pour $1 \leq i \leq r-1$ et $1 \leq j \leq c-1$, on peut, par un échantillonnage multinominal sous H_o, exprimer la matrice de variances-convariances asymptotique de $\underset{\sim}{h} = (h_{11}, \ldots, h_{1,c-1}, \ldots, h_{r-1,c-1})^T$ sous la forme $\underset{\sim}{P}_R \otimes \underset{\sim}{P}_C$, où $\otimes$ est le produit direct de deux matrices. La statistique de Wald devient donc, si H_o est vraie,

$$\begin{aligned} W_4 &= \underset{\sim}{h}^T(\hat{\underset{\sim}{P}}_C^{-1} \otimes \hat{\underset{\sim}{P}}_R^{-1})\underset{\sim}{h} \\ &= \sum_{i=1}^{r}\sum_{j=1}^{c} (p_{ij} - p_{i+}p_{+j})^2/(p_{i+}p_{+j}), \end{aligned}$$

qui correspond au test habituel du khi-carré à $(r-1)\,(c-1)$ degrés de liberté.

Un autre test, qui est asymptotiquement équivalent à W_4 si H_o est vraie, est le test du rapport de vraisemblance exprimé par

$$LR_2 = 2n\left[\sum_{i=1}^{r}\sum_{j=1}^{c} p_{ij}\log p_{ij} - \sum_{i=1}^{r} p_{i+}\log p_{i+} - \sum_{j=1}^{c} p_{+j}\log p_{+j}\right].$$

On peut aussi résoudre ce problème en prenant un cas spécial des procédés décrits par Grizzle, Starner et Koch (1969). On calcule la statistique du test de Wald avec

$$\{f_{ij} = \log p_{ij} - \log p_{i+} - \log p_{+j};\ \text{for } 1 \leq i \leq r-1, \text{ and } 1 \leq j \leq c-1\}.$$

La matrice de variances-covariances asymptotique de $\hat{\underset{\sim}{f}} = (f_{11}, \ldots, f_{1,c-1}, \ldots, f_{r-1,c-1})^T$ est $(\underset{\sim}{D}_{\pi_R}^{-1} - \underset{\sim}{1}\,\underset{\sim}{1}^T) \otimes (\underset{\sim}{D}_{\pi_C}^{-1} - \underset{\sim}{1}\underset{\sim}{1}^T)$. On obtient ainsi une variable de Wald qui a la forme suivante:

$$W_5 = \underset{\sim}{f}^T\left[\left(\hat{\underset{\sim}{D}}_{\pi_R} + \frac{\underset{\sim}{\pi}_R\,\underset{\sim}{\pi}_R^T}{p_{r+}}\right) \otimes \left(\hat{\underset{\sim}{D}}_{\pi_C} + \frac{\underset{\sim}{\pi}_C\,\underset{\sim}{\pi}_C^T}{p_{+c}}\right)\right]\underset{\sim}{f}$$

On note que, si H_o est vraie, f_{ij} est asymptotiquement équivalent à

$$\frac{p_{ij}}{\pi_{i+}\pi_{+j}} - \frac{p_{i+}}{\pi_{i+}} - \frac{p_{+j}}{\pi_{+j}} + 1,$$

de sorte que $\sum_{i=1}^{r} \pi_{i+} f_{ij} \doteq \sum_{j=1}^{c} \pi_{+j} f_{ij} \doteq 0$. Cette approximation permet de remplacer W_5 par

$$W_5' = \sum_{i=1}^{r}\sum_{j=1}^{c} p_{i+}p_{+j} f_{ij}^2 .$$

Il convient de noter que si H_o est vraie, les fonctions W_4, LR_2 et W_5 sont toutes asymptotiquement équivalentes à

$$\sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(p_{ij} - \pi_{i+}\pi_{+j})^2}{\pi_{i+}\pi_{+j}} - \sum_{i=1}^{r} \frac{(p_{i+} - \pi_{i+})^2}{\pi_{i+}} - \sum_{j=1}^{c} \frac{(p_{+j} - \pi_{+j})^2}{\pi_{+j}} \tag{3.1}$$

L'utilité de ce résultat sera montrée à la section (3.3).

3.2 Autres méthodes d'échantillonnage

Au lieu de supposer que $\underset{\sim}{n}$ suit une loi multinomiale, on suppose maintenant que $n^{1/2}(\underset{\sim}{p} - \underset{\sim}{\pi}) \rightarrow N(\underset{\sim}{0}, \underset{\sim}{V})$ où $\underset{\sim}{p}$ est une estimation calculée à partir de données d'enquête et peut dépendre des poids d'échantillonnage et d'autres facteurs de correction. Dans ce cas, Shuster et Downing (1976) et Fellegi (1980) proposent une statistique de Wald construite à partir de $\{h_{ij} = p_{ij} - p_{i+}p_{+j}\}$. Soit $\underset{\sim}{J}_a$ la matrice de dimension $(a-1) \times a$ qui a la forme suivante

$$\underset{\sim}{J}_a = [\underset{\sim}{I} \mid \underset{\sim}{0}] \tag{3.2}$$

et définissons $\hat{\underset{\sim}{H}} = (\underset{\sim}{J}_r - \underset{\sim}{\pi}_R \underset{\sim}{1}^T) \otimes (\underset{\sim}{J}_c - \underset{\sim}{\pi}_C \underset{\sim}{1}^T) - (\underset{\sim}{\pi}_R \underset{\sim}{1}^T \otimes \underset{\sim}{\pi}_C 1^T)$

On calcule ensuite la statistique de Wald qui s'écrit

$$W_6 = \underset{\sim}{h}^T(\hat{\underset{\sim}{H}} \hat{\underset{\sim}{V}} \hat{\underset{\sim}{H}}^T)^{-1}\underset{\sim}{h},$$

qui, si H_0 est vraie, suit asymptotiquement une loi de type $\chi^2_{(r-1)(c-1)}$.

Nous pouvons aussi construire une statistique de Wald dans laquelle on définit $\{f_{ij} = \log p_{ij} - \log p_{i+} - \log p_{+j}\}$. C'est là un cas spécial du modèle log-linéaire qui est présenté à la section 4. On définit deux matrices qui sont respectivement de dimension $(r-1) \times r$ et $(c-1) \times c$:

$$\hat{\underset{\sim}{E}}_R = \left[\hat{\underset{\sim}{D}}^{-1}_{\pi_R} \mid \underset{\sim}{0}\right], \quad \hat{\underset{\sim}{E}}_C = \left[\hat{\underset{\sim}{D}}^{-1}_{\pi_C} \mid \underset{\sim}{0}\right].$$

Soit $\hat{\underset{\sim}{F}} = (\hat{\underset{\sim}{E}}_R - \hat{\underset{\sim}{E}}_R \underset{\sim}{1}\underset{\sim}{1}^T) \otimes (\hat{\underset{\sim}{E}}_C - \hat{\underset{\sim}{E}}_C \underset{\sim}{1}\underset{\sim}{1}^T) - (\hat{\underset{\sim}{E}}_R \underset{\sim}{1}\underset{\sim}{1}^T \otimes \hat{\underset{\sim}{E}}_C \underset{\sim}{1}\underset{\sim}{1}^T)$.

La statistique de Wald qui découle de ces définitions est

$$W_7 = \underset{\sim}{f}^T(\hat{\underset{\sim}{F}} \hat{\underset{\sim}{V}} \hat{\underset{\sim}{F}}^T)^{-1}\underset{\sim}{f}.$$

Comme il a été fait à la section 2 pour le problème de la qualité de l'ajustement, Rao et Scott (1981) ont montré que les distributions sous H_o des variables W_4, LR_2 et W_5, sont toutes asymptotiquement équivalentes à la distribution sous H_o de (3.1). Par conséquent, cette distribution sous H_o est identique à celle de $\Sigma_{i=1}^{(r-1)(c-1)} \delta_i Z_i^2$, où $\{Z_i^2\}$ est un ensemble de variables indépendantes qui suivent une loi de type χ_1^2 et les δ_i les valeurs propres de

$$(\underset{\sim}{P}_R^{-1} \otimes \underset{\sim}{P}_C^{-1})(\underset{\sim}{H}\underset{\sim}{V}\underset{\sim}{H}^T).$$

Cowan et Binder (1978) ont étudié les propriétés des valeurs propres issues d'un échantillon à deux degrés prélevé selon un plan autopondéré pour un tableau de dimension de 2×2. Ces auteurs ont constaté que la valeur propre augmente quand le degré d'indépendance des proportions dans chaque case diminu à l'intérieur des unités primaires d'échantillonnage.

3.3 Approximations

Une approximation de la distribution de $\Sigma\ \delta_i Z_i^2$ est

$$\Sigma\ \delta_i Z_i^2 \approx \frac{\Sigma\ \delta_i}{(r-1)\ (c-1)} \chi^2_{(r-1)\ (c-1)},$$

comme dans l'équation (2.2). Étant donné que cette statistique est asymptotiquement équivalente à l'expression (3.1), si H_o est vraie, on peut calculer la moyenne de (3.1) et obtenir

$$\Sigma\ \delta_i = \sum_{i=1}^{r}\sum_{j=1}^{c} d_{ij}\ (1 - \pi_{i+}\pi_{+j}) - \sum_{i=1}^{r} d_i^{(r)}\ (1 - \pi_{i+}) - \sum_{j=1}^{c} d_j^{(c)}(1-\pi_{\pm j}),$$

où d_{ij} est l'effet du plan de sondage dans chaque case; $d_i^{(r)}$ et $d_j^{(c)}$ sont l'effet du plan de sondage sur les fréquences marginales dans chaque rangée et chaque colonne respectivement. Ce sont Rao et Scott (1983) qui ont obtenu cette expression particulièrement simple. Enfin, Fellegi (1980) a proposé une autre approximation:

$$(\sum_{i=1}^{r}\sum_{j=1}^{c} d_{ij}/rc)\ \chi^2_{(r-1)\ (c-1)}$$

Exemple 2

Le tableau 2 comprend un tableau de dimension 4×2 qui est basé sur des données de l'Enquête Santé Canada. Ce tableau présente un classement recoupé de l'utilisation de médicaments en fonction de la consommation (quatre catégories: 0, 1, 2, 3+ classes de médicaments sur une période de deux jours) et selon le sexe (hommes, femmes). Ici, $n = 31{,}668$.

La valeur brute de W_4 is 774. La correction de Rao et Scott (1981) réduit cette valeur à 437, tandis que celle de Fellegi (1980) la fait chuter à 327. La statistique de Wald, W_6, est égale à 538. Hidiroglou et Rao (1981) ont constaté que l'approximation de Rao et Scott (1981) produit d'assez bons résultats en comparaison de ceux de l'approximation de Satterthwaite (1946) qui exige une connaissance complète de la matrice de variances-covariances.

4. MODÈLES LOG-LINÉAIRES

4.1 Échantillonnage multinomial

Nous faisons maintenant une extension des résultats de la section précédente à des classements recoupés plus généraux qui suivent une loi multinomiale. Les formules classiques pour ce genre de modèle figurent dans les ouvrages de Bishop, Fienberg et Holland (1975) et Fienberg (1980).

Tableau 2

Classes de médicaments consommés selon le sexe, Canada (1978-79)

		Nombre de classes de médicaments				
Sexe		0	1	2	3±	Total
Hommes	Proportion	0.293	0.134	0.048	0.021	0.496
	Effet du plan	1.56	3.37	1.15	1.38	0.00*
Femmes	Proportion	0.228	0.159	0.072	0.045	0.504
	Effet du plan	3.59	3.13	2.85	1.96	0.00*
Total	Proportion	0.521	0.293	0.120	0.066	1.000
	Effet du plan	6.03	6.46	1.65	2.57	

* A cause de la stratification a posteriori selon l'âge et le sexe, ces effets du plan sont nuls.

Le vecteur $\underset{\sim}{\pi} = (\pi_1,\ldots,\pi_k)^T$ contient les proportions dans chaque case et $\Sigma_{i=1}^{k} \pi_k = 1$. Nous observons $\underset{\sim}{n} = (n_1,\ldots,n_k)^T$, le vecteur des totaux dans chaque case, et à partir d'un échantillon aléatoire, de sorte que $\underset{\sim}{n}$ suit une loi multinomiale ($\Sigma\ n_i = n$). Définissons $\underset{\sim}{p} = \underset{\sim}{n}/n$ et

$$\underset{\sim}{\mu} = \log \underset{\sim}{\pi}.$$

Dans le modèle log-linéaire, le vecteur des paramètres $\underset{\sim}{\theta} = (\theta_1, \ldots, \theta_t)^T$, est tel que

$$\underset{\sim}{\mu}(\underset{\sim}{\theta}) = u(\underset{\sim}{\theta})\underset{\sim}{1} + \underset{\sim}{X}\underset{\sim}{\theta},$$

où $\underset{\sim}{X}$ est une matrice connue de dimension $k \times t$ et de rang complet et $\underset{\sim}{X}^T\underset{\sim}{1} = \underset{\sim}{0}$. Notons que $t \leq k-1$. Si $t = k-1$, le modèle est saturé.

Pour obtenir l'estimateur du maximum de vraisemblance pour $\underset{\sim}{\theta}$, on doit résoudre l'équation (4.1):

$$\underset{\sim}{X}^T(\underset{\sim}{p} - \hat{\underset{\sim}{\pi}}) = \underset{\sim}{0}, \tag{4.1}$$

où $\hat{\underset{\sim}{\pi}} = \underset{\sim}{\pi}(\hat{\underset{\sim}{\theta}})$. Asymptotiquement,

$$\hat{\underset{\sim}{\pi}} - \underset{\sim}{\pi} \doteq \underset{\sim}{P}\underset{\sim}{X}(\hat{\underset{\sim}{\theta}} - \underset{\sim}{\theta}),$$

où $\underset{\sim}{P} = \underset{\sim}{D}_{\pi} - \underset{\sim}{\pi}\underset{\sim}{\pi}^T$. La formule (4.1) permet d'écrire

$$\hat{\underset{\sim}{\theta}} - \underset{\sim}{\theta} \doteq (\underset{\sim}{X}^T\underset{\sim}{P}\underset{\sim}{X})^{-1}\underset{\sim}{X}^T(\underset{\sim}{p} - \underset{\sim}{\pi})$$

et

$$\hat{\underset{\sim}{\pi}} - \underset{\sim}{\pi} \doteq \underset{\sim}{P}\underset{\sim}{X}(\underset{\sim}{X}^T\underset{\sim}{P}\underset{\sim}{X})^{-1}\underset{\sim}{X}^T(\underset{\sim}{p} - \underset{\sim}{\pi}).$$

Étant donné que $n^{1/2}(\underset{\sim}{p} - \underset{\sim}{\pi}) \to N(\underset{\sim}{0}, \underset{\sim}{P})$ on peut écrire

$$n^{1/2}(\hat{\underset{\sim}{\theta}} - \underset{\sim}{\theta}) \to N[\underset{\sim}{0}, (\underset{\sim}{X}^T\underset{\sim}{P}\underset{\sim}{X})^{-1}]$$

$$n^{1/2}(\underset{\sim}{\pi} - \underset{\sim}{\pi}) \to N[\underset{\sim}{0}, \underset{\sim}{P}\underset{\sim}{X}(\underset{\sim}{X}^T\underset{\sim}{P}\underset{\sim}{X})^{-1}\underset{\sim}{X}^T\underset{\sim}{P}].$$

Supposons maintenant que l'expression linéaire $\underset{\sim}{X}\underset{\sim}{\theta}$ admet une décomposition de la forme $\underset{\sim}{X}_1\underset{\sim}{\theta}_1 + \underset{\sim}{X}_2\underset{\sim}{\theta}_2$ où $\underset{\sim}{X}_1$ et $\underset{\sim}{X}_2$ sont de rang complet, $\underset{\sim}{X}_1$ est de dimension $k \times r$, $\underset{\sim}{X}_2$ de dimension $k \times s$, $\underset{\sim}{\theta}_1$ de dimension $r \times 1$ et $\underset{\sim}{\theta}_2$ de dimension $s \times 1$, où $r + s = t$.

Nous voulons vérifier l'hypothèse

$$H_o: \underset{\sim}{\theta}_2 = \underset{\sim}{0},$$

contre l'hypothèse

$$H_1: \underset{\sim}{\theta}_2 \neq \underset{\sim}{0}.$$

Nous utilisons $\underset{\sim}{\theta}_1$, $\underset{\sim}{\theta}_2$, $\underset{\sim}{\pi}$, etc. pour représenter les estimations calculées à partir du modèle, si H_1 est vraie, et $\hat{\underset{\sim}{\theta}}_1$, $\hat{\underset{\sim}{\pi}}$, etc. pour représenter les estimations calculées, si H_o est vraie.

Or,

$$n^{1/2}(\underset{\sim}{\theta}_2 - \underset{\sim}{\theta}_2) \to N[\underset{\sim}{0}, (\tilde{\underset{\sim}{X}}_2^T\underset{\sim}{P}\tilde{\underset{\sim}{X}}_2)^{-1}]$$

où

$$\tilde{\underset{\sim}{X}}_2 = [\underset{\sim}{I} - \underset{\sim}{X}_1(\underset{\sim}{X}_1^T\underset{\sim}{P}\underset{\sim}{X}_1)^{-1}\underset{\sim}{X}_1^T\underset{\sim}{P}]\underset{\sim}{X}_2 \tag{4.2}$$

de sorte que la statistique de Wald est

$$W_8 = n\hat{\underset{\sim}{\theta}}_2^T \tilde{\underset{\sim}{X}}_2^T \hat{\underset{\sim}{P}} \tilde{\underset{\sim}{X}}_2 \underset{\sim}{\theta}_2.$$

Si H_o est vraie, cette statistique est asymptotiquement équivalente à la statistique du khi-carré de Pearson

$$n\,(\hat{\underset{\sim}{\pi}} - \hat{\underset{\sim}{\pi}})^T \hat{\underset{\sim}{D}}_\pi^{-1} (\hat{\underset{\sim}{\pi}} - \hat{\underset{\sim}{\pi}}),$$

où à la statistique du test du rapport de vraisemblance

$$LR_3 = 2n \sum_{i=1}^{k} p_i \log(\hat{\pi}_i / \hat{\pi}_i).$$

Si H_o est vraie, ces variables suivent asymsptotiquement une loi de type χ_s^2.

4.2 Autres méthodes d'échantillonnage

Nous supposons de nouveau que les proportions dans chaque case, $\underset{\sim}{\pi}$, satisfont à la condition $\underset{\sim}{\mu} = \log \underset{\sim}{\pi} = u(\underset{\sim}{\theta}_1, \underset{\sim}{\theta}_2)\underset{\sim}{1} + \underset{\sim}{X}_1 \underset{\sim}{\theta}_1 + \underset{\sim}{X}_2 \underset{\sim}{\theta}_2$ mais on a maintenant que $n^{1/2}(\underset{\sim}{p} - \underset{\sim}{\pi}) \to N(\underset{\sim}{0}, \underset{\sim}{V})$, où $\underset{\sim}{p}$ est une estimation calculée à partir des données d'une enquête.

Rao et Scott (1983) ont proposé une statistique de Wald pour tester $\underset{\sim}{\theta}_2 = \underset{\sim}{0}$. Soit $\underset{\sim}{C}$ une matrice de dimension $k \times s$ pour laquelle $\underset{\sim}{C}^T \underset{\sim}{X}_1 = \underset{\sim}{0}$, $\underset{\sim}{C}^T \underset{\sim}{1} = \underset{\sim}{0}$ and $\underset{\sim}{C}^T \underset{\sim}{X}_2$ est non singulière. Par exemple, si $\underset{\sim}{X}_1^T \underset{\sim}{X}_2 = \underset{\sim}{0}$, $\underset{\sim}{C} = \underset{\sim}{X}_2$ est une définition convenable. Cela revient à supposer que $\underset{\sim}{C}^T \underset{\sim}{\mu} = \underset{\sim}{0}$. Ainsi

$$\begin{aligned} \underset{\sim}{C}^T(\hat{\underset{\sim}{\mu}} - \underset{\sim}{\mu}) &\doteq \underset{\sim}{C}^T \underset{\sim}{D}_\pi^{-1} (\hat{\underset{\sim}{\pi}} - \underset{\sim}{\pi}) \\ &\doteq \underset{\sim}{C}^T \underset{\sim}{X} (\underset{\sim}{X}^T \underset{\sim}{P} \underset{\sim}{X})^{-1} \underset{\sim}{X}^T (\underset{\sim}{p} - \underset{\sim}{\pi}), \end{aligned}$$

où $\underset{\sim}{\pi}$ est reporté de l'équation (4.1), dans laquelle $\underset{\sim}{p}$ est une estimation calculée à partir des données d'une enquête.

On a donc la statistique de Wald suivante:

$$W_9 = n\hat{\underset{\sim}{\mu}}^T \underset{\sim}{C} [\underset{\sim}{C}^T \underset{\sim}{X} (\underset{\sim}{X}^T \hat{\underset{\sim}{P}} \underset{\sim}{X})^{-1} (\underset{\sim}{X}^T \hat{\underset{\sim}{V}} \underset{\sim}{X}) (\underset{\sim}{X}^T \hat{\underset{\sim}{P}} \underset{\sim}{X})^{-1} \underset{\sim}{X}^T \underset{\sim}{C}]^{-1} \underset{\sim}{C}^T \hat{\underset{\sim}{\mu}}.$$

Binder (1983) a obtenu des résultats semblables. Si, sous H_1, le modèle est saturé ($r+s = k-1$), alors $\underset{\sim}{p} = \underset{\sim}{\pi}$ et on peut écrire

$$W_9 = n\hat{\underset{\sim}{\mu}}^T \underset{\sim}{C} [\underset{\sim}{C}^T \hat{\underset{\sim}{D}}_\pi^{-1} \hat{\underset{\sim}{V}} \hat{\underset{\sim}{D}}_\pi^{-1} \underset{\sim}{C}]^{-1} \underset{\sim}{C}^T \hat{\underset{\sim}{\mu}}.$$

Rao et Scott (1984) ont démontré que, si on utilise $\hat{\underset{\sim}{P}}$ au lieu de $\hat{\underset{\sim}{V}}$ pour calculer W_9 et W_9', ces statistiques sont asymptotiquement équivalentes à la statistique du rapport de vraisemblance ou à la statistique du khi-carré de Pearson. Ces auteurs ont aussi montré que la statistique du test du rapport de vraisemblance a une distribution comme celle de $\sum_{i=1}^{s} \delta_i Z_i^2$ sous H_o, où $\{Z_i^2\}$ est un ensemble de variables indépendantes de type χ_1^2 et les $\{\delta_i\}$ sont les valeurs propres de

$$(\tilde{\underset{\sim}{X}}_2^T \underset{\sim}{P} \tilde{\underset{\sim}{X}}_2)^{-1} (\tilde{\underset{\sim}{X}}_2^T \underset{\sim}{V} \tilde{\underset{\sim}{X}}_2), \tag{4.3}$$

où $\tilde{\underset{\sim}{X}}_2$ est définie dans l'équation (4.2).

4.3 Approximations

Comme dans les sections précédentes, nous utilisons l'approximation suivante de la distribution sous l'hypothèse nulle

$$\sum_{l=1}^{s} \delta_i Z_i^2 \approx \left(\frac{\sum \delta_i}{s}\right) \chi_s^2.$$

Pour évaluer cette expression, il faut calculer la trace de la matrice (4.3). Rao et Scott (1984) ont démontré que, si le modèle admet des solutions explicites de $\hat{\underset{\sim}{\pi}}$ and $\hat{\underset{\sim}{\pi}}$, cette approximation dépend de la matrice $\underset{\sim}{V}$ seulement en fonction de l'effet du plan de sondage sur les fréquences des cases et les fréquences marginales. Cette observation est particulièrement utile quand on dispose seulement d'estimations de ces effets, ce qui est souvent le cas dans les tableaux publiés.

Exemple 3

Hidiroglou et Rao (1983) ont examiné toutes les estimations directes dans un tableau à trois dimensions de données de l'Enquête Santé Canada recoupées selon la consommation de médicaments (cinq catégories: 0, 1, 2, 3, 4+ classes de médicaments sur une période de deux jours) × l'âge (quatre catégories; 0-14, 15-44, 45-64, 65+) × et le sexe (hommes, femmes). Nous résumons ici le résultat d'un test d'indépendance de l'âge et du sexe dans chaque classe de médicament (n = 31,668). L'hypothèse nulle est donc

$$H_0\colon \pi_{ijk} = \pi_{ij+}\frac{\pi_{i+k}}{\pi_{i++}}.$$

Si on utilise la notation de Bishop, Fienberg et Holland (1975), dans laquelle $\log \pi_{ijk} = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{12(ij)} + u_{13(ik)} + u_{23(jk)} + u_{123(ijk)}$, l'hypothèse nulle peut s'écrire

$$H_0\colon u_{23(jk)} = u_{123(ijk)} = 0 \quad \text{pour tous les } (i, j, k).$$

La valeur brute de la statistique khi-carré est 23 pour 15 degrés de liberté. La valeur propre moyenne est de 1.39, de sorte que l'approximation réduit la valeur de khi-carré à 16. La valeur non corrigée du khi-carré conduirait donc l'analyste à rejeter l'hypothèse nulle au seuil de 10%, alors que l'approximation indique que h_o ne peut pas être rejetée, même au seuil de 30%.

5. MODÈLES DE RÉGRESSION LOGISTIQUE

5.1 Échantillonnage multinomial

Nous examinons maintenant un modèle de régression logistique pour la distribution conditionnelle d'une variable binaire y en fonction d'un vecteur donné $\underset{\sim}{x}$ de variables indépendantes. Plus précisément, il s'agit de la distribution conditionnelle suivante:

$$\Pr(y_i \mid x_i) = \pi(\underset{\sim}{x}_i)^{y_i}[1 - \pi(\underset{\sim}{x}_i)]^{1-y_i},$$

où $y_i \in \{0, 1\}$.

Dans le modèle de régression logistique

$$\log\left\{\frac{\pi(\underset{\sim}{x}_i)}{1 - \pi(\underset{\sim}{x}_i)}\right\} = \underset{\sim}{x}_i^T \underset{\sim}{\theta},$$

où $\underset{\sim}{\theta}$ est un vecteur inconnu de paramètres.

Notons que si $\underset{\sim}{x}_i$ est un vecteur qualitatif dont les éléments prennent la valeur 0 ou 1, ce modèle est un cas spécial du modèle log-linéaire décrit à la section 4. On permet ici à $\underset{\sim}{x}_i$ d'être arbitaire. L'extension de cette analyse au cas où la variable y comprend k catégories peut se faire directement. Il est également possible d'écrire le modèle de régression logistique sous une forme générale:

$$\log\left\{\frac{\pi(\underset{\sim}{x}_i)}{1 - \pi(\underset{\sim}{x}_i)}\right\} = f(\underset{\sim}{x}_i^T \underset{\sim}{\theta}),$$

pour une fonction connue, $f(\cdot)$, mais cet aspect n'est pas abordé ici.

L'estimateur du maximum de vraisemblance de $\underset{\sim}{\theta}$ se calcule à l'aide de l'équation

$$\underset{\sim}{X}^T(\underset{\sim}{y} - \hat{\underset{\sim}{\pi}}) = \underset{\sim}{0}$$

où $\underset{\sim}{y} = (y_1, \ldots, y_n)^T$, $\hat{\underset{\sim}{\pi}} = [\hat{\pi}(\underset{\sim}{x}_1), \ldots, \hat{\pi}(\underset{\sim}{x}_n)]^T$ et $\underset{\sim}{X} = [\underset{\sim}{x}_1 \ldots \underset{\sim}{x}_n]^T$.

Dans certaines conditions de régularité appropriées,

$$n^{1/2}(\hat{\underset{\sim}{\theta}} - \underset{\sim}{\theta}) \to N[\underset{\sim}{0}, n(\underset{\sim}{X}^T\underset{\sim}{\Delta}\underset{\sim}{X})^{-1}], \text{ où } \underset{\sim}{\Delta} = \underset{\sim}{D}_\pi(\underset{\sim}{I} - \underset{\sim}{D}_\pi).$$

Si $\underset{\sim}{X}\underset{\sim}{\theta} = \underset{\sim}{X}_1\underset{\sim}{\theta}_1 + \underset{\sim}{X}_2\underset{\sim}{\theta}_2$ et que nous voulons vérifier les hypothèses suivantes:

$$H_0\colon \underset{\sim}{\theta}_2 = \underset{\sim}{0}$$
$$H_1\colon \underset{\sim}{\theta}_2 \neq \underset{\sim}{0},$$

on obtient une statistique de Wald qui se définit ainsi:

$$W_{10} = n\,\hat{\underset{\sim}{\theta}}_2^T(\tilde{\underset{\sim}{X}}_2^T\,\underset{\sim}{\Delta}\,\underset{\sim}{X}_2)\,\hat{\underset{\sim}{\theta}}_2$$

où

$$\tilde{\underset{\sim}{X}}_2 = [\underset{\sim}{I} - \underset{\sim}{X}_1(\underset{\sim}{X}_1^T\underset{\sim}{\Delta}\underset{\sim}{X}_1)^{-1}\underset{\sim}{X}_1^T\hat{\underset{\sim}{\Delta}}]\underset{\sim}{X}_2.$$

Le test du rapport de vraisemblance repose alors sur la statistique

$$LR_4 = 2\sum_{i=1}^{n}\left[y_i \log\left(\frac{\hat{\pi}_i}{\hat{\bar{\pi}}_i}\right) + (1 - y_i)\log\left\{\frac{(1 - \hat{\pi}_i)}{(1 - \hat{\bar{\pi}}_i)}\right\}\right]$$

qui est asymptotiquement équivalente à W_{10} si H_0 est vraie.

5.2 Autres méthodes d'échantillonnage

Supposons maintenant que $n^{-1/2}\underset{\sim}{X}^T(\underset{\sim}{y} - \underset{\sim}{\pi}) \to N(\underset{\sim}{0}, \underset{\sim}{V})$ et que $\hat{\underset{\sim}{V}}$ est un estimateur convergent de $\underset{\sim}{V}$. Ici, $\underset{\sim}{y}$ n'est pas nécessairement un vecteur composé de 0 et de 1,mais il peut en réalité dépendre des poids d'échantillonnage et d'autres facteurs de correction. On peut normalement estimer que $\underset{\sim}{V}$ est une somme d'observations aléatoires et la plupart des plans d'échantillonnage admettent un estimateur convergent d'une somme d'observations (qui ne sont pas nécessairement indépendantes). Pour estimer $\underset{\sim}{V}$, nous utilisons $\hat{\underset{\sim}{\pi}}$ au lieu de $\underset{\sim}{\pi}$ dans l'estimateur. Étant donné que, asymptotiquement,

$$(\hat{\underset{\sim}{\theta}} - \underset{\sim}{\theta}) \doteq (\underset{\sim}{X}^T\underset{\sim}{\Delta}\underset{\sim}{X})^{-1}\underset{\sim}{X}^T(\underset{\sim}{y} - \underset{\sim}{\pi}),$$

il est possible de déduire que

$$n^{1/2}(\hat{\underset{\sim}{\theta}} - \underset{\sim}{\theta}) \to N[\underset{\sim}{0}, n^2(\underset{\sim}{X}^T\underset{\sim}{\Delta}\underset{\sim}{X})^{-1}\underset{\sim}{V}(\underset{\sim}{X}^T\underset{\sim}{\Delta}\underset{\sim}{X})^T];$$

voir Binder (1983) pour une explication détaillée de ce résultat. On peut alors construire une statistique de Wald à partir de la matrice de variances-covariances estimée de $\underset{\sim}{\theta}_2$.

Tableau 3

Modèle de régression logistique pour expliquer la consultation d'un médecin

Variable	Type de données	d.l.	Statistique de Wald
Âge	Qualitatif	4	19.232
Sexe	Qualitatif	1	12.494
Interaction âge-sexe	Qualitatif	4	36.001
Revenu de la famille	Qualitatif	5	14.642
Profession	Qualitatif	3	8.614
Interaction profession-sexe	Qualitatif	3	11.501
État matrimonial	Qualitatif	3	45.752
Antécédents médicaux	Qualitatif	2	36.700
Nombre de problèmes de santé	Quantitatif	1	81.554
Consommation de médicaments	Qualitatif	2	272.175
Nombre d'accidents	Quantitatif	2	106.372
Nombre de jours de maladie	Quantitatif	2	29.052
Taille de l'agglomération	Qualitatif	2	11.751
Rapport du nombre de médecins et de la population dans la province	Quantitatif	1	0.540

Exemple 4

On a ajusté un modèle de régression logistique à des données de l'Enquête Santé Canada sur 20,726 répondants pour expliquer le fait qu'ils ont ou n'ont pas consulté un médecin sur une période de douze mois. En tout, on a estimé que 77% de la population avait consulté un médecin au moins une fois. Les résultats sont résumés au tableau 3 (pour une description détaillée, voir binder (1983)). Le modèle de régression logistique a semblé très bien s'ajuster aux données.

5.3 Variables explicatives qualitatives

La théorie décrite dans cette section a été présentée par G. Roberts dans un document non publié (Université Carleton). Dans le cas qui nous occupe, toutes les variables explicatives sont qualitatives. Les domaines sont notés $\{1, \ldots, I\}$. Nous définissons p_i l'estimation calculée à partir des données de l'enquête de la proportion du $i^{\text{ème}}$ domaine et $\hat{N}_i$ l'estimation de la taille du $i^{\text{ème}}$ domaine, N_i. Suivant le modèle, la proportion espérée dans le $i^{\text{ème}}$ domaine est f_i où

$$\log \{f_i/(1-f_i)\} = \underset{\sim}{a}_i^T \theta,$$

pour $\underset{\sim}{a}_i$ connu et $\underset{\sim}{\theta}$ au paramètre inconnu. Nous définissons $\underset{\sim}{A} = [\underset{\sim}{a}_1, \ldots, \underset{\sim}{a}_I]^T$ et $\underset{\sim}{D}_{\hat{N}} = \text{diag}\{\hat{N}_1, \ldots, \hat{N}_I\}$. Selon le modèle, l'estimateur calculé à partir des données de l'enquête de $\underset{\sim}{f} = (f_1, \ldots, f_I)^T$ est $\hat{\underset{\sim}{f}}$, la solution de l'équation

$$\underset{\sim}{A}^T \underset{\sim}{D}_{\hat{N}} (\underset{\sim}{p} - \hat{\underset{\sim}{f}}) = \underset{\sim}{0}. \tag{5.1}$$

Puisque asymptotiquement

$$\hat{\underset{\sim}{\theta}} - \underset{\sim}{\theta} \doteq (\underset{\sim}{A}^T \underset{\sim}{\Delta} \underset{\sim}{A})^{-1} \underset{\sim}{A}^T \underset{\sim}{D}_N (\underset{\sim}{p} - \underset{\sim}{f}),$$

où $\underset{\sim}{\Delta} = \text{diag}\{N_1 f_1(1-f_1), \ldots, N_I f_I(1-f_I)\}$, il s'en suit que

$$n^{1/2}(\underset{\sim}{\theta} - \underset{\sim}{\theta}) \to N[\underset{\sim}{0}, (\underset{\sim}{A}^T \underset{\sim}{\Delta} \underset{\sim}{A})^{-1} \underset{\sim}{A}^T \underset{\sim}{D}_N \underset{\sim}{V}_P \underset{\sim}{D}_N \underset{\sim}{A} (\underset{\sim}{A}^T \underset{\sim}{\Delta} \underset{\sim}{A})^{-1}]$$

lorsque $n^{1/2}(\underset{\sim}{p} - \underset{\sim}{f}) \to N(\underset{\sim}{0}, \underset{\sim}{V}_P)$.

Suivant un échantillonnage binomial indépendant, la matrice de variances-covariances est réduite à $(N/n)(\underset{\sim}{A}^T \underset{\sim}{\Delta} \underset{\sim}{A})^{-1}$, où n est la taille de l'échantillon.

Le test du rapport de vraisemblance pour tester la qualité de l'ajustement est

$$LR_5 = 2(n/\hat{N}) \sum_{i=1}^{I} \hat{N}_i [p_i \log(p_i/\hat{f}_i) + (1-p_i) \log\{(1-p_i)/(1-\hat{f}_i)\}],$$

où n est la taille d'échantillon et $\hat{N} = \Sigma \hat{N}_i$. Si H_o est vraie, cette statistique est asymptotiquement équivalente à

$$W_{11} = (n/\hat{N}) \sum_{i=1}^{I} \hat{N}_i (p_i - \hat{f}_i)^2 / [f_i (1 - f_i)].$$

En général, la distribution de LR_5 sera celle de $\Sigma \delta_i Z_i^2$, où $\{Z_i\}$ est un ensemble de variables aléatoires indépendantes distribuées selon une loi χ_1^2, et $\{\delta_i\}$ sont les valeurs propres de $\hat{N}^{-1} \underset{\sim}{D}_{\hat{N}} [\underset{\sim}{\Delta}^{-1} - \underset{\sim}{A} (\underset{\sim}{A}^T \underset{\sim}{\Delta} \underset{\sim}{A})^{-1} \underset{\sim}{A}^T] \underset{\sim}{D}_{\hat{N}} \underset{\sim}{V}_P \underset{\sim}{D}_{\hat{N}} [\underset{\sim}{\Delta}^{-1} - \underset{\sim}{A} (\underset{\sim}{A}^T \underset{\sim}{\Delta} \underset{\sim}{A})^{-1} \underset{\sim}{A}^T] \underset{\sim}{\Delta} \underset{\sim}{D}_{\hat{N}}^{-1}$. En prenant l'espérance mathématique de W_{11} et l'approximation

$$W_{11} \approx \frac{\Sigma \delta_i}{I - s} \chi^2_{I-s}$$

où $s = \text{rang}(\underset{\sim}{A})$, il s'en suit que

$$\Sigma \delta_i = (n/\hat{N}) \sum_{i=1}^{I} \hat{N}_i v_{ii}^{(r)} / \{f_i (1 - f_i)\}$$

où $v_{ii}^{(r)} = V\{p_i - \hat{f}_i\}$. On peut calculer les $\{v_{ii}^{(r)}\}$ en utilisant la relation $\underset{\sim}{p} - \hat{\underset{\sim}{f}} \doteq [\underset{\sim}{I} - \text{diag}\{f_i(1-f_i)\}\underset{\sim}{A}(\underset{\sim}{A}^T\underset{\sim}{\Delta}\underset{\sim}{A})^{-1}\underset{\sim}{A}^T\underset{\sim}{D}_N](\underset{\sim}{p} - \underset{\sim}{f})$.

Exemple 5

Les données de l'enquête sur la population active d'octobre 1980 ont été utilisées pour ajuster des modèles logistiques (logit) de la probabilité d'avoir un travail. L'échantillon est composé d'hommes âgés de 15 à 64 ans qui font partie de la population active et n'étudient pas à plein temps. Un modèle logit, quadratique pour les variables âge et instruction, a été ajusté à ces données. On a défini des groupes d'âge en divisant l'intervalle [15, 64] en dix parties et le $j^{\text{ième}}$ groupe d'âge est représenté par l'intervalle $[10 + 5j, 14 + 5j]$, $j = 1, 2, \ldots, 10$. Le milieu de chaque groupe d'âge a été utilisé comme la valeur de l'âge de toutes les personnes appartenant au groupe correspondant. On a établi six niveaux d'instruction en attribuant à chaque personne une valeur basée sur la médiane du nombre d'années d'instruction. Ainsi, la création de catégories d'âge et d'instruction produit un tableau comprenant soixante cases.

Soit $\pi_i = \Pr\{\text{une personne classée dans la } i^{\text{ième}} \text{ case a un emploi}\}$, $i = 1, 2, \ldots, 60$. Nous supposons que $0 < \pi_i < 1$. Par conséquent, $1 - \pi_i$ représente la probabilité qu'une personne dans la $i^{\text{ième}}$ cas soit chômeur. Le modèle ajusté a la forme suivante:

$$\ln \frac{\pi_i}{1-\pi_i} = \beta_o + \beta_1 a_i + \beta_2 a_i^2 + \beta_3 d_i + \beta_4 d_i^2, \qquad (1)$$

$$i = 1, 2, \ldots, 60$$

où les a_i et les d_i sont les valeurs des variables âge et instruction des personnes classées dans la $i^{\text{ième}}$ case.

À partir des estimations p_i obtenues pour les π_i dans l'échantillon, les valeurs de la statistique de Pearson, W_{11}, et du rapport de vraisemblance LR_5 ont été calculées et les résultats sont $W_{11} = 98.94$ et $LR_5 = 101.20$. La valeur critique de la distribution khi-carré à 55 degrés de liberté qui délimite une région de 5% à droite est 73.31. Les valeurs calculées de W_{11} et de LR_5 permettent de rejeter le modèle (1). Toutefois, les résultats obtenus pour W_{11} ou LR_5 sont appropriés seulement si l'échantillon est aléatoire.

L'estimation de la valeur propre moyenne, $\Sigma\, \delta_i/55$, pour tester la qualité de l'ajustement est égale à 1.88. Ce calcul a pour effet de diminuer W_{11} à 52.63 et LR_5 à 53.83. Cette correction permet donc de constater que les données concordent avec le modèle (1).

On a également songé à employer la statistique de Wald, $(\underset{\sim}{p} - \hat{\underset{\sim}{f}})^T[\hat{V}^{(r)}]^-(\underset{\sim}{p} - \hat{\underset{\sim}{f}})$, pour tester la qualité de l'ajustement. (Notons que l'inverse généralisé de $\hat{V}^{(r)}$ est utitisé pour ce cas puisque cette matrice est singulière.) Quand $p_i = 1$, il faut que la valeur de p_i soit un peu perturbée pour qu'on puisse calculer la statistique de Wald. La statistique de Wald s'avère instable pour notre problème. De faibles perturbations dans les estimations de p entraînent des changements considérables dans la valeur de la statistique de Wald.
En outre, cette valeur est très élevée ici à cause de l'instabilité de la matrice de variances-covariance estimée qui entre dans le calcul de la statistique de Wald. Cette statistique est au moins trente fois plus élevée que nos valeurs corrigées du khi-carré.

6. PROBLÈMES ACTUELS DE L'UTILISATION DE PROGICIELS

Grâce aux progrès réalisés dans la technologie informatique, les opérations de collecte, de stockage et d'accès aux données sont devenues faciles et efficaces. Des systèmes puissants d'application générale, comme TPL, STATPAK et ESTIMATION SYSTEM, permettent à des utilisateurs et à des analystes d'estimer des totaux et leur variance avec assez peu de difficulté. En outre, un certain nombre de progiciels analytiques offerts tels BMDP, SPSS et SAS se prêtent extrêment bien à certains contextes. Mais la capacité de ces progiciels à exécuter des analyses comme celles décrites ici est limitée. Par exemple, dans les tests d'hypothèses ou l'inférence statistique, ces progiciels supposent que les données analysées proviennent d'enquêtes basées sur des échantillons aléatoires simples.

A l'heure actuelle, il n'existe pas de progiciel intégré comme ceux mentionnés plus haut qui soit conçu pour des analyses des types de données dont il est question dans les sections précédentes. Par conséquent, quiconque cherche une solution rapide à un problème de ce genre doit habituellement avoir recours aux progiciels existants, lesquels risquent de ne pas être appropriés.

Les possibilités qui s'offrent sont les suivantes:

- utiliser les progiciels existants, mais en les modifiant;
- utiliser les programmes spéciaux qui existent déjà;
- écrire des programmes sur mesure;
- adopter une combinaison des choix ci-dessus.

Pour les analyses présentées dans ce document, c'est en modifiant le programme MINI CARP (Hidiroglou, Fuller et Hickman, 1980) qu'on a obtenu les résultats des exemples 1, 2 et 3. Pour l'exemple 4, on a utilisé une combinaison de programmes écrits en PL/1 et en SAS. L'analyse des données de l'Enquête sur la Population Active (exemple 5) a été faite à l'aide de SAS et de programmes sur mesure.

Chacune des possibilités énumérées plus haut présente des inconvénients sur le plan pratique, notamment:

a) s'il est nécessaire de modifier un progiciel qui existe déjà, il faut posséder une connaissance détaillée de ses mécanismes;

b) il peut être nécessaire de reproduire des renseignements identiques sur différents fichiers de données, étant donné que les solutions susmentionnées ne peuvent pas être intégrées comme les systèmes généraux;

c) en comparaison des progiciels intégrés qui sont conçus de manière à être facilement utilisables par différents utilisateurs, les autres programmes manquent souvent d'élégance et d'efficacité opérationnelle;

d) il est souvent impossible d'obtenir une documentation complète sur les programmes spéciaux ou faits sur mesure, ce qui limite l'accessibilité de ces logiciels.

Des travaux sont actuellement en cours pour mettre au point des programmes écrits en SAS qui permettent d'effectuer un grand nombre des analyses décrites ici. Notre objectif final est semblable à celui formulé par Shah (1981), c'est-à-dire élaborer un progiciel complet pour l'analyse des données d'enquête. Tous les efforts visant à atteindre cet objectif sont justifiés si nous voulons éviter les problèmes qu'éprouvent actuellement les spécialistes qui doivent soit construire leurs propres programmes, soit utiliser les progiciels qui existent déjà, ce qui peut produire des résultats et des conclusions erronés.

7. CONCLUSION

Nous avons examiné quelques-uns des problèmes que pose l'ajustement de modèles à des données qualitatives recueillies en fonction de plans d'échantillonnage complexes. La méthode fondamentale qu'on utilise ici est de calculer la statistique de Wald qui convient au modèle ajusté ou d'employer le test approprié sur les plans fondés sur l'échantillonnage multinomial et ensuite trouver une bonne approximation de la distribution sous l'hypothèse nulle.

Nous n'avons pas abordé la question du choix entre les méthodes basées sur le modèle et celles basées sur le plan de sondage. On a plutôt mis l'accent sur le processus d'inférence basé sur le plan de sondage.

Pour résumer cette dichotomie, prenons de nouveau le test d'indépendance dans un tableau de contingence à deux dimensions. L'indépendance nous intéresse quand nous voulons savoir si la valeur de la variable Y_1 a un effet sur notre connaissance de la variable Y_2. Si la réponse est non pour toutes les unités de la population, on dit alors que ces variables sont indépendantes. Mais si nous connaissons la valeur de Y_3, il est encore possible que Y_1 et Y_2 ne soient

pas des variables indépendantes. Cette possibilité est particulièrement importante quand Y_3 est une variable (comme la strate géographique) qui est utilisée dans la structure du plan d'échantillonnage. Étant donné que ce genre de variable est habituellement connue pour toutes les unités échantillonnées, on peut soit a) ne plus accorder d'importance à la question de l'indépendance, soit b) mettre Y_3 en marge et s'intéresser seulement à Y_1 et Y_2, non à leur distribution conditionnelle. Si on choisit b), les résultats présentés dans les sections précédentes semblent appropriés. Dans certains cas, il peut être possible de vérifier si Y_1 et Y_2 sont conditionnellement indépendants pour Y_3 donné.

Toutefois, il existe une autre difficulté. Supposons que nous voulons connaître les proportions π_{ij} dans chaque case d'un tableau pour une population finie de taille N. Dans un recensement de cette population, il est peu probable qu'on obtienne la relation $\pi_{ij} = \pi_{i+}\,\pi_{+j}$ exactement. Au mieux, on peut souhaiter qu'une mesure quelconque d'association comme, par exemple $N\Sigma\Sigma(\pi_{ij} - \pi_{i+}\pi_{+j})^2\pi_{i+}\pi_{+j}$, soit faible. Notons que même dans un modèle de superpopulation à indépendance exacte, on ne s'attendrait pas à ce que ce coefficient d'association soit nul. Il faut peut-être en définitive tester des hypothèses telles que

$$H_o\text{: coefficient d'association} < C$$

$$H_1\text{: coefficient d'association} > C.$$

D'autres recherches doivent être menées sur ce sujet. Toutefois, dans des cas pratiques où la fraction de sondage n'est pas grande, les méthodes décrites ici sont applicables.

BIBLIOGRAPHIE

ALTHAM, P.A.E. (1976). "Discrete Variable Analysis for Individuals Grouped Into Families", *Biometrika*, 63, 263-269.

BINDER, D.A. (1983). Les variances d'estimateurs asymptotiquement normaux basés sur des enquêtes complexes. Version française d'un article qui sera publié dans *International Statistical Review*, 51.

BISHOP, Y.M.M., FIENBERG, S.E. et HOLLAND, P.W. (1975). *Discrete Multivariate Analysis: Theory and Practice*. Cambridge, Mass: MIT Press.

BRIER, S.S. (1978). "Discrete Data Models With Random Effects", Technical Report, University of Minnesota, School of Statistics.

COHEN, J.E. (1976). "The Distribution of the Chi-Squared Statistic Under Cluster Sampling Form Contingency Tables", *Journal of the American Statistical Association*, 71, 665-670.

COWAN, J. et BINDER, D.A. (1978). The effect of a two-stage sample design on tests of independence in a 2 by 2 table. *Techniques d'enquête*, 4, 16-28.

FAY, R.E. (1985). A jackknifed chi-squared test for complex samples. *Journal of American Statistical Association,* 80, 148-157.

FELLEGI, I.P. (1980). Approximate tests of independence and goodness of fit based on stratified multistage samples. *Journal of the American Statistical Association*, 71, 665-670.

FIENBERG, S.E. (1980). *The Analysis of Cross Classified Data,* (2^e ed.). Cambridge, Mass: MIT Press.

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

HIDIROGLOU, M.A., FULLER, W.A. et HICKMAN, R.D. (1980). MINICARP: A program for estimating simple descriptive statistics and their variances for multi-stage stratified designs. Iowa State University: Ames, Iowa.

HIDIROGLOU, M.A. et RAO, J.N.K. (1981). Tests khi-carré pour l'analyse d'observations classées de l'Enquête Santé Canada. Version française d'un document présenté aux réunions de l'Institut international de statistique, Buenos Aires, 1981.

HIDIROGLOU, M.A. et RAO, J.N.K. (1983). Chi-square tests for the analysis of three-way contingency tables from the Canada Health Survey. Technical Report. Statistics Canada.

IMREY, P.B., KOCH, G.G. et STOKES (1981). Categorical data analysis: Some reflections of the log linear model and logistic regression. Part I: Historical Methodological Overview. *International Statistical Review*, 49, 265-283.

JOHNSON, N.L. et KOTZ, S. (1970). Continuous Univariate Distributions. Boston: Houghton Mifflin.

KOCH, G.G., FREEMAN, D.H. JR., and FREEMAN, J.L. (1975). "Strategies in the Multivariate Analysis of Data From Complex Surveys". *International Statistical Review*, 43, 59-78.

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.

RAO, J.N.K. et SCOTT, A.J. (1983). On Chi-square tests for multiway contingency tables with cell proportions estimates from survey data. Carleton Mathematical Series No. 199, Carleton University, Ottawa.

SATTERTHWAITE, F.E. (1946). An approximate distribution of estimates of variance components. *Biometrics*, 2, 110-114.

SHAH, B.V. (1981). Development of survey data analysis software. Research Triangle Institute, Research Triangle Park, North Carolina.

SHUSTER, J.J. et DOWNING, D.J. (1976). Two-way contingency tables for complex sampling schemes. *Biometrika*, 63, 272-276.