ISSN 0248-3289

PUBLICATIONS
DU
LABORATOIRE
DE
STATISTIQUE
ET
PROBABILITES

N° 01 89

Juin 1989

PRESENTATION DES TESTS DU RAPPORT DE VRAISEMBLANCES, DE WALD ET DES MULTIPLICATEURS DE LAGRANGE

Alain BACCINI

Université Paul SABATIER
TOULOUSE

PRESENTATION DES TESTS
DU RAPPORT DE VRAISEMBLANCES, DE WALD
ET DES MULTIPLICATEURS DE LAGRANGE

Alain BACCINI

UNIVERSITE PAUL SABATIER
Laboratoire de Statistique et Probabilités
U.A. - C.N.R.S. 745
118, route de Narbonne
31062 TOULOUSE CEDEX

N° 01-89
Juin 1989

RESUME

L'objectif de cet article est essentiellement pédagogique ; aussi ne contient-il aucun résultat original, son but étant de présenter, de façon simple, un certain nombre de résultats concernant trois tests d'hypothèses aujourd'hui d'usage courant en statistique et, en particulier, en économétrie.

Nous mettons tout d'abord l'accent sur les hypothèses, relatives à la vraisemblance, nécessaires à la mise en oeuvre de ces tests. Ensuite, après avoir rappelé le principe du test du rapport de vraisemblances, nous présentons deux variantes pour le test de WALD comme pour le test des multiplicateurs de LAGRANGE, puisqu'on rencontre tantôt l'une et tantôt l'autre dans la littérature statistique ; les liens entre ces variantes sont notamment précisés. Un exemple commun sert d'illustration au calcul effectif des différentes statistiques de test. Dans un dernier point, nous montrons l'équivalence, dans le modèle multinomial, du test des multiplicateurs de LAGRANGE et du test du khi-deux, puis nous indiquons les relations entre les trois statistiques de test dans certains cas particuliers du modèle linéaire.

1. INTRODUCTION

1.1. Les données considérées.

On considère un échantillon de n variables aléatoires réelles (v.a.r.), supposées indépendantes et identiquement distribuées (i.i.d.) et notées $X_1, \dots, X_n$.

En fait, ce qui suit s'étend sans difficulté au cas de vecteurs aléatoires (à valeurs dans $\mathbb{R}^k$), comme au cas de v.a.r. indépendantes mais non identiquement distribuées (cas correspondant, notamment, au modèle linéaire) ; dans ce dernier cas, il est toutefois nécessaire de faire des hypothèses supplémentaires sur les X_i pour conserver les propriétés asymptotiques énoncées plus loin. Les restrictions indiquées ci-dessus ne sont donc faites que pour plus de commodité.

On suppose que la loi de probabilité commune des X_i ($i=1,\dots,n$) dépend d'un paramètre θ de dimension p ($p \in \mathbb{N}^*$) et l'on considère donc le modèle statistique $(\mathbb{R}, \mathcal{B}_{\mathbb{R}}, (P_\theta, \theta \in \Theta \subset \mathbb{R}^p))^n$, où Θ désigne l'espace paramétrique.

On suppose également que cette loi de probabilité admet une densité par rapport à une mesure σ-finie μ définie sur $\mathbb{R}$; nous noterons $f(x,\theta)$ cette densité et $L(x_1,\dots,x_n,\theta)$ la vraisemblance du modèle ci-dessus ; on a donc $L(x_1,\dots,x_n,\theta) = \prod_{i=1}^{n} f(x_i,\theta)$.

1.2. Les hypothèses du modèle.

Tout ce qui suit suppose vérifiées un certain nombre d'hypothèses de régularité sur le modèle envisagé, hypothèses classiques en statistique inférentielle et vérifiées par la plupart des lois de la famille exponentielle.

H_1 : Le support $\mathcal{X}$ de $f(x,\theta)$ est indépendant de θ.

L'importance de cette hypothèse dans les calculs portant sur la vraisemblance est évidente.

H_2 : L'espace paramétrique Θ est un ouvert de $\mathbb{R}^p$ et l'on a : $\forall x \in \mathcal{X}, \quad \forall \theta \in \Theta, \quad f(x,\theta) > 0$.

Cette hypothèse permet d'une part d'envisager la dérivation de f, selon θ, en tout point de Θ, d'autre part de définir le logarithme de f en tout point de $\mathcal{X}\times\Theta$.

H_3 : En tout point de $\mathcal{X}\times\Theta$, toutes les dérivées partielles d'ordre 1 et d'ordre 2 de $f(x,\theta)$, selon θ, sont définies.

On peut ainsi définir le score du modèle, $S(x,\theta) = \mathrm{grad}_\theta[\mathrm{Log}\ f(x,\theta)]$, et considérer son information, au sens de FISHER, $J(\theta) = E_\theta[S(X,\theta)^t S(X,\theta)]$.

$S(x,\theta)$ est un vecteur-colonne à p éléments et $J(\theta)$, si ses éléments sont finis, est une matrice carrée d'ordre p, symétrique et semi-définie-positive.

H_4 : En tout point θ de Θ, $f(x,\theta)$ est deux fois différentiable selon θ sous le signe $\int_{\mathbb{R}} d\mu(x)$.

Cette hypothèse entraîne que l'espérance de $S(X,\theta)$ est nulle et donc que $J(\theta)$ est la matrice des variances-covariances de $S(X,\theta)$.

H_5 : Pour tout θ de Θ, $J(\theta)$ est définie-positive.

Si H_5 est vraie, $J(\theta)$ est inversible, $\forall\theta\in\Theta$.

On sait que les hypothèses H_1 à H_5 sont nécessaires pour établir l'inégalité de CRAMER-RAO.

H_6 : Le modèle est identifiable, c'est-à-dire : $\theta\neq\theta' \Rightarrow P_\theta \neq P_{\theta'}$.

Cette hypothèse est nécessaire pour obtenir la convergence de l'estimateur maximum de vraisemblance (E.M.V.) $\hat{\theta}_n$ de θ vers la vraie valeur du paramètre.

H_7 : $\forall(j,k)\in\{1,\ldots,p\}^2$, $\dfrac{\partial^2 \mathrm{Log}\ f(x,\theta)}{\partial\theta_j\,\partial\theta_k}$ est uniformément continue en x et continue en θ sur $\mathcal{X}\times\Theta$.

Les hypothèses H_1 à H_7 assurent la normalité et l'efficacité asymptotiques de l'E.M.V. ; autrement dit, on a la propriété suivante :

Propriété 1 : Sous les hypothèses H_1 à H_7,
$$\sqrt{n}[\hat{\theta}_n - \theta] \xrightarrow[n \to +\infty]{\text{loi}} N(0, J^{-1}(\theta)).$$

1.3. Les hypothèses à tester.

Considérons un entier q vérifiant $q \leqslant p$ et une application h de $\mathbb{R}^p$ dans $\mathbb{R}^q$ supposée deux fois continûment différentiable ; $h : \theta \mapsto h(\theta) = (h_1(\theta), \ldots, h_q(\theta))$.

On veut tester ici l'hypothèse nulle $H_0 : h(\theta) = 0$, contre son contraire $\bar{H}_0$ (un niveau de test α étant fixé).

Les trois tests que nous présentons dans la suite utilisent l'E.M.V. $\hat{\theta}_n$ de θ, ainsi que l'E.M.V. de θ sous la contrainte H_0 (autrement dit $\theta \in \Theta_0 = \Theta \cap \operatorname{Ker} h$), que nous noterons $\hat{\theta}_n^0$. On a, de façon évidente, $L(x, \hat{\theta}_n^0) \leqslant L(x, \hat{\theta}_n)$, $\forall x \in \mathcal{X}$. Des précisions sur la détermination de $\hat{\theta}_n^0$ sont données en 4.1.

2. TEST DU RAPPORT DE VRAISEMBLANCES

2.1. Principe.

Ce test, bien connu, consiste à considérer la statistique $\Lambda_n = \dfrac{L(X,\hat{\theta}_n^0)}{L(X,\hat{\theta}_n)}$, vérifiant $0 \leqslant \Lambda_n \leqslant 1$, et à prendre comme région critique $\Lambda_n < \lambda_\alpha$ (intuitivement, les petites valeurs de Λ_n correspondent à $\bar{H}_0$) ; λ_α est défini par $P_{H_0}[\Lambda_n < \lambda_\alpha] = \alpha$, α étant le niveau du test. Ce test fut proposé par NEYMAN and PEARSON (1928).

Sous cette forme, sa mise en oeuvre (à savoir le calcul de λ_α en fonction de α, ou l'inverse) ne peut se faire que lorsqu'on connait la loi de Λ_n sous H_0, c'est-à-dire seulement dans quelques cas particuliers.

2.2. Comportement asymptotique.

On pose $R_n = -2 \text{ Log } \Lambda_n$; on a $0 \leqslant R_n \leqslant +\infty$, et l'on obtient un test équivalent au précédent en prenant pour région critique $R_n > k_\alpha$. L'intérêt d'utiliser R_n est qu'on connaît sa distribution asymptotique sous H_0, donnée par la propriété suivante :

<u>Propriété 2</u> : Si H_0 est vraie, $R_n \xrightarrow[n\to+\infty]{\text{loi}} \chi^2_q$,

où χ^2_q désigne une loi de khi-deux à q degrés de liberté (q est la dimension de l'espace image de h).

Ce résultat fut initialement établi par WILKS (1938) et généralisé par WALD (1943) ; pour sa démonstration, on pourra se reporter à SERFLING (1980) ou à MONFORT (1982).

<u>Remarques</u> : 1) De façon plus générale, en ne supposant pas H_0, mais sous une suite d'hypothèses "convergeant" d'une certaine façon vers H_0, WALD (1943) a établi que R_n converge en loi vers une loi de khi-deux décentrée, à q d.d.l., le paramètre de décentrage s'annulant si et seulement si H_0 est vraie (sur ce point, voir aussi LEHMANN, 1986, chapitre 8).

2) Pour θ fixé dans $\bar{\Theta}_0$, on vérifie que $R_n \xrightarrow[n\to+\infty]{} +\infty$, ce qui entraîne que le test est convergeant.

2.3. Exemple.

Nous introduisons ici un exemple très simple qui nous servira d'illustration dans toute la suite.

Considérons la loi multinomiale à 4 catégories, notons p_1, p_2, p_3 et p_4 ses paramètres et X_i (i=1,...,n) le vecteur aléatoire des indicatrices des 4 catégories associé à la i-ième épreuve (les X_i sont supposés indépendants) ; on a $X_i = (X_i^1, X_i^2, X_i^3, X_i^4)$, chaque composante X_i^j (j=1,...,4) prenant la valeur 1 avec la probabilité p_j et la valeur 0 avec la probabilité $1-p_j$.

Remarque : Comme nous l'avons indiqué en 1.1, ce qui est fait dans cet article sur des v.a.r. i.i.d. se généralise sans difficulté à des vecteurs aléatoires i.i.d. ; c'est dans ce cadre plus général que rentre l'exemple traité, si on le présente sous la forme ci-dessus. Notons toutefois que, vue la simplicité de cet exemple, on peut se ramener à des v.a.r. X_i en posant, par exemple, $X_i = \ell$ si la i-ième observation appartient à la ℓ-ième catégorie (ℓ=1,...,4).

Posons $N_j = \sum_{i=1}^{n} X_i^j$ et notons n_j la valeur observée de N_j.

On suppose $0<p_j<1$, $\forall j=1,...,4$ et $\sum_{j=1}^{4} p_j = 1$; pour que les différentes composantes du paramètre θ soient linéairement indépendantes, on pose $\theta = (p_1, p_2, p_3)$; on a ainsi p=3 et Θ est l'intérieur du tétraèdre dont les sommets sont les points de coordonnées (0,0,0), (1,0,0), (0,1,0) et (0,0,1).

Le logarithme de la vraisemblance du modèle peut s'écrire

$$\text{Log } L(x,\theta) = C + \sum_{j=1}^{3} n_j \text{ Log } p_j + n_4 \text{ Log}(1 - \sum_{j=1}^{3} p_j),$$

où $x = (n_1, n_2, n_3, n_4)$ et où C est une constante relativement à θ.

On sait alors que l'E.M.V. de θ est $\hat{\theta}_n = \frac{1}{n}\begin{bmatrix} N_1 \\ N_2 \\ N_3 \end{bmatrix}$.

Supposons maintenant que l'on souhaite tester

$$H_0 : p_1 = p_2, \quad \text{contre} \quad \bar{H}_0 : p_1 \neq p_2 ;$$

cela est équivalent à introduire la fonction h de $\mathbb{R}^3$ dans $\mathbb{R}$ (q=1), définie par $h(\theta) = p_1 - p_2$, et à tester $h(\theta)=0$ contre $h(\theta)\neq 0$.

L'estimation maximum de vraisemblance de θ sous la contrainte $h(\theta)=0$ conduit à la solution unique (voir 4.5) $\hat{\theta}_n^0 = \frac{1}{n}\begin{bmatrix} \frac{N_1 + N_2}{2} \\ \frac{N_1 + N_2}{2} \\ N_3 \end{bmatrix}$;

on en déduit $\Lambda_n = \left(\frac{N_1 + N_2}{2}\right)^{N_1 + N_2} \frac{1}{N_1^{N_1} N_2^{N_2}}$,

et $R_n = 2[N_1 \mathrm{Log}\, N_1 + N_2 \mathrm{Log}\, N_2 - (N_1 + N_2)\mathrm{Log}\, \frac{N_1 + N_2}{2}]$.

<u>Application numérique</u> : Considérons les données suivantes :

$n = 210$, $n_1 = 47$, $n_2 = 55$, $n_3 = 83$, $n_4 = 25$.

La valeur de $\hat{\theta}_n$ est sensiblement (0,22 0,26 0,40) et celle de $\hat{\theta}_n^0$ sensiblement (0,24 0,24 0,40).

Il vient $r_n = 2(47\mathrm{Log}47 + 55\mathrm{Log}55 - 102\mathrm{Log}51) \simeq 0,628$.

La valeur limite d'une loi de khi-deux à 1 degré de liberté (q=1) au niveau 10% est 2,706, ce qui conduit à accepter l'hypothèse H_0.

3. TEST DE WALD

3.1. Notations.

Nous noterons ici $[h(\theta)]$ le vecteur-colonne $q\times 1$ associé à $h(\theta)$ et $D_h(\theta)$ la matrice jacobienne de h ;$D_h(\theta)$ est $q\times p$ et telle que $[D_h(\theta)]^j_\ell = \dfrac{\partial}{\partial\theta_j} h_\ell(\theta)$; nous la supposerons de rang q, ce qui nécessite que θ soit "bien conditionné".

Posons $\Sigma(\theta) = D_h(\theta)\, J^{-1}(\theta)\, {}^tD_h(\theta)$;
$\Sigma(\theta)$ est $q\times q$, symétrique et définie-positive (car $J(\theta)$ l'est, d'après H_5, et que $D_h(\theta)$ est de rang q) ; $\Sigma^{-1}(\theta)$ existe et est également symétrique et définie-positive.

Remarques : 1) On voit que les hypothèses faites sur le modèle statistique permettent de définir, pour toute valeur de θ, deux métriques sur $\mathbb{R}^p$ (ou une métrique sur $\mathbb{R}^p$ et une sur son dual), à partir de $J(\theta)$ et de $J^{-1}(\theta)$; celles faites sur h permettent, de plus, de définir pour tout θ deux métriques sur $\mathbb{R}^q$, à partir de $\Sigma(\theta)$ et de $\Sigma^{-1}(\theta)$.

2) Dans l'estimation de $h(\theta)$ par un estimateur sans biais, $\dfrac{1}{n}\,\Sigma(\theta)$ représente la borne inférieure de l'inégalité de CRAMER-RAO.

3.2. Principe du test.

Considérons $W_n = n\, {}^t[h(\hat\theta_n)]\, \Sigma^{-1}(\hat\theta_n)[h(\hat\theta_n)] = n\|h(\hat\theta_n)\|^2_{\Sigma^{-1}(\hat\theta_n)}$;
W_n est la statistique du test de WALD.

Il est clair que si H_0 est vraie, c'est-à-dire si $h(\theta)=0$, son estimation $h(\hat\theta_n)$ est voisine de 0 et donc W_n est une quantité petite ; la région critique du test est donc de la forme $W_n > k_\alpha$.

Remarque : Il arrive que l'on considère $S_n(x,\theta) = \text{grad}_\theta[\text{Log } L(x,\theta)]$, $J_n(\theta) = E_\theta[S_n(X,\theta)\,{}^tS_n(X,\theta)]$ et $\Sigma_n(\theta) = D_h(\theta)J_n^{-1}(\theta)\,{}^tD_h(\theta)$. Dans le cas où les v.a.r. X_i sont i.i.d., il vient alors :

$$S_n(x,\theta) = \sum_{i=1}^{n} S(x_i,\theta),\quad J_n(\theta) = n\, J(\theta) \quad\text{et}\quad \Sigma_n(\theta) = \frac{1}{n}\,\Sigma(\theta).$$

On peut par conséquent réécrire : $W_n = \|h(\hat\theta_n)\|^2_{\Sigma_n^{-1}(\hat\theta_n)}$.

3.3. Comportement asymptotique.

Il est donné par la propriété suivante :

Propriété 3 : Si H_0 est vraie, $W_n \xrightarrow[n\to+\infty]{\text{loi}} \chi^2_q$.

Cette propriété fut établie par WALD (1943), dans l'article où il étudiait également le comportement asymptotique de la statistique R_n du test du rapport de vraisemblances. On pourra encore se reporter à SERFLING (1980) ou à MONFORT (1982) pour la démonstration. Notons qu'en s'écartant de H_0 dans les mêmes conditions que celles précisées dans la remarque 1 du 2.2, on obtient le même résultat asymptotique que pour le test du rapport de vraisemblances.

3.4. Variante.

Rappelons la propriété 1 établie sous les hypothèses H_1 à H_7 :

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow[n\to+\infty]{\text{loi}} N(0,\ J^{-1}(\theta)).$$

L'application h ayant été supposée deux fois différentiable, on en déduit :

$$\sqrt{n}[h(\hat{\theta}_n) - h(\theta)] \xrightarrow[n\to+\infty]{\text{loi}} N(0, D_h(\theta) J^{-1}(\theta)\ {}^tD_h(\theta)) \qquad (3.1)$$

(voir, par exemple, LEHMANN, 1983).

Supposons définie la matrice des variances-covariances de $h(\hat{\theta}_n)$ et notons la $V_{h(\hat{\theta}_n)}(\theta)$; la statistique figurant dans le terme de gauche de (3.1) a donc pour matrice des variances-covariancesn $V_{h(\hat{\theta}_n)}(\theta)$.

En supposant vérifiées les hypothèses supplémentaires sur $f(x,\theta)$ pour assurer la convergence de tous les moments (voir, par exemple, BOROVKOV,1987), il vient :

$n\, V_{h(\hat{\theta}_n)}(\theta) \xrightarrow[n\to+\infty]{} D_h(\theta)\ J^{-1}(\theta)\ {}^tD_h(\theta) = \Sigma(\theta)$, et donc :

$$\frac{1}{n}\, V^{-1}_{h(\hat{\theta}_n)}(\theta) \xrightarrow[n\to+\infty]{} \Sigma^{-1}(\theta).$$

Si, de plus, cette convergence est uniforme, $V^{-1}_{h(\hat{\theta}_n)}(\hat{\theta}_n)$ et $n\,\Sigma^{-1}(\hat{\theta}_n)$ ont alors le même comportement asymptotique ; par conséquent, si H_0 est vraie, la statistique $W'_n = {}^t[h(\hat{\theta}_n)]V^{-1}_{h(\hat{\theta}_n)}(\hat{\theta}_n)[h(\hat{\theta}_n)]$ converge en loi vers une loi de khi-deux à q d.d.ℓ. et peut être utilisée comme statistique du test de WALD.

Remarques : 1) Si $h(\hat{\theta}_n)$ est un estimateur sans biais et efficace de $h(\theta)$, alors $V_{h(\hat{\theta}_n)}(\theta) = \frac{1}{n} D_h(\theta)J^{-1}(\theta)\ {}^tD_h(\theta) = \frac{1}{n}\Sigma(\theta)$ et l'on a $W'_n = W_n$, $\forall n\in\mathbb{N}^*$.

2) Si h est une fonction affine du paramètre ($h(\theta)=A\theta+B$, cas fréquent dans la pratique), alors $D_h(\theta)=A$ et il vient :

$$V_{h(\hat{\theta}_n)}(\theta) = A\ V_{\hat{\theta}_n}(\theta)\ {}^tA.$$

De plus, dans ce cas, il suffit que $\hat{\theta}_n$ soit sans biais et efficace pour θ pour avoir $W'_n = W_n$.

3.5. Exemple.

On reprend ici l'exemple introduit en 2.3.

$$h(\theta) = p_1 - p_2 \Rightarrow h(\hat{\theta}_n) = \frac{1}{n}(N_1 - N_2).$$

$D_h(\theta)$ est de dimension 1×3, vaut (+1 -1 0) et est donc égale à $D_h(\hat{\theta}_n)$.

On sait d'autre par que, dans le modèle multinomial, on a $J^{-1}(\theta) = n\ V_{\hat{\theta}_n}(\theta)$ (voir, par exemple, TASSI, 1989) ; il vient donc :

$$J^{-1}(\theta) = \begin{bmatrix} p_1(1-p_1) & -p_1p_2 & -p_1p_3 \\ -p_1p_2 & p_2(1-p_2) & -p_2p_3 \\ -p_1p_3 & -p_2p_3 & p_3(1-p_3) \end{bmatrix}.$$

On en déduit $\Sigma(\theta) = D_h(\theta) J^{-1}(\theta)^t D_h(\theta) = (p_1+p_2) - (p_1-p_2)^2$

et donc $$\Sigma(\hat{\theta}_n) = \frac{1}{n^2}[n(N_1+N_2) - (N_1-N_2)^2].$$

Finalement, $$W_n = n\ {}^t[h(\hat{\theta}_n)]\ \Sigma^{-1}(\hat{\theta}_n)[h(\hat{\theta}_n] = \frac{n(N_1-N_2)^2}{n(N_1+N_2)-(N_1-N_2)^2}$$

$$= \frac{(N_1-N_2)^2}{(N_1+N_2) - \dfrac{(N_1-N_2)^2}{n}}.$$

Variante : On sait que, dans le modèle multinomial, l'E.M.V. est sans biais et efficace pour θ ; h étant ici affine, on doit donc trouver $W_n^{\cdot} = W_n$.

q valant 1, $V_{h(\hat{\theta}_n)}(\theta)$ est un réel ; comme $h(\hat{\theta}_n) = \frac{1}{n}(N_1-N_2)$,

il vient : $$V_{h(\hat{\theta}_n)}(\theta) = \frac{1}{n^2}[Var(N_1) + Var(N_2) - 2\ Cov(N_1,N_2)]$$

$$= \frac{1}{n^2}[np_1(1-p_1) + np_2(1-p_2) + 2n\ p_1 p_2]$$

$$= \frac{1}{n}[(p_1+p_2) - (p_1-p_2)^2] = \frac{1}{n}\Sigma(\theta)\ ;$$

d'où le résultat.

Application numérique : Reprenant les mêmes données qu'en 2.3, on obtient :

$$w_n = \frac{64}{102 - \dfrac{64}{210}} \simeq 0{,}629.$$

Dans cet exemple, la différence entre les valeurs des deux statistiques R_n et W_n n'est que de un millième ; les conclusions sont donc, bien sûr, identiques.

4. TEST DES MULTIPLICATEURS DE LAGRANGE

Ce test porte aussi les noms de test de RAO ou de test du score.

4.1. Maximisation de la vraisemblance sous contraintes.

Pour maximiser, selon θ, la vraisemblance $L(X,\theta)$ sous la contrainte $h(\theta)=0$, on utilise la méthode des multiplicateurs de LAGRANGE, c'est-à-dire qu'on maximise le lagrangien de $\operatorname{Log} L(X,\theta)$, qui s'écrit :

$$\operatorname{Log} L(X,\theta) + {}^t[h(\theta)][\mu] \quad , \quad \text{où} \quad [\mu] = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_q \end{bmatrix} \text{ est le}$$

vecteur des multiplicateurs de LAGRANGE.

On résoud donc le système de p+q équations en θ et μ :

$$\left.\begin{aligned} S_n(X,\theta) + {}^tD_h(\theta)[\mu] &= 0 \\ h(\theta) &= 0 \end{aligned}\right\} \qquad (4.1).$$

On notera $\hat{\theta}_n^0$ la solution en θ et $\hat{\mu}_n$ la solution en μ.

<u>Remarque</u> : En général, pour ne pas compliquer le problème, on néglige les conditions d'ordre 2 ; dans la mesure où n est grand, cela n'est pas vraiment gênant, puisque, sous les hypothèses de régularité supposées ici, les équations de vraisemblance n'admettent asymptotiquement qu'une seule solution et qu'elle correspond à un maximum.

4.2. Principe du test.

On considère $M_n = \dfrac{1}{n}\, {}^t[\hat{\mu}_n]\Sigma(\hat{\theta}_n^0)[\hat{\mu}_n] = \dfrac{1}{n}\|\hat{\mu}_n\|^2_{\Sigma(\hat{\theta}_n^0)}$,

avec toujours $\Sigma(\theta) = D_h(\theta)J^{-1}(\theta)\, {}^tD_h(\theta)$.

M_n est la statistique du test des multiplicateurs de LAGRANGE.

Si H_0 est vraie, $\hat{\theta}_n^0 \simeq \hat{\theta}_n$ et $S_n(X,\hat{\theta}_n^0) \simeq S_n(X,\hat{\theta}_n) = 0$; on a donc dans ce cas ${}^tD_h(\hat{\theta}_n^0)[\hat{\mu}_n] \simeq 0$ et donc $\|\hat{\mu}_n\|$ doit être petit (on peut en particulier vérifier que ${}^tD_h$ est une application linéaire et continue, de noyau nul).

On prend donc pour région critique du test $M_n > k_\alpha$.

<u>Remarque</u> : En reprenant les notations de la remarque du 3.2, on peut réécrire :

$$M_n = {}^t[\hat{\mu}_n]\Sigma_n(\hat{\theta}_n^0)[\hat{\mu}_n] = \|\hat{\mu}_n\|^2_{\Sigma_n(\hat{\theta}_n^0)} .$$

4.3. <u>Comportement asymptotique.</u>

On retrouve le même résultat que pour les tests précédents, que H_0 soit vraie ou non ; on a donc, en particulier :

<u>Propriété 4</u> : Si H_0 est vraie, $M_n \xrightarrow[n \to +\infty]{\text{loi}} \chi^2_q$.

Ce test, sous ce nom et sous cette forme, fut proposé par AITCHISON and SILVEY (1958) et par SILVEY (1959), qui établirent la propriété ci-dessus. On pourra également trouver la démonstration de cette propriété dans SILVEY (1975) ou dans MONFORT (1982).

Toutefois, ce test est équivalent à un test plus ancien, proposé par RAO (voir ci-dessous).

4.4. <u>Variante.</u>

$\hat{\theta}_n^0$ et $\hat{\mu}_n$ étant solutions de (4.1), on a ${}^tD_h(\hat{\theta}_n^0)[\hat{\mu}_n] = - S_n(X,\hat{\theta}_n^0)$, d'où l'on déduit :

$$M_n = \frac{1}{n}\, {}^t[\hat{\mu}_n]\, D_h(\hat{\theta}_n^0)\ J^{-1}(\hat{\theta}_n^0)\ {}^tD_h(\hat{\theta}_n^0)[\hat{\mu}_n]$$

$$= \frac{1}{n}\, {}^tS_n(X,\hat{\theta}_n^0)J^{-1}(\hat{\theta}_n^0)S_n(X,\hat{\theta}_n^0) = \frac{1}{n}\|S_n(X,\hat{\theta}_n^0)\|^2_{J^{-1}(\hat{\theta}_n^0)}$$

$$= \|S_n(X,\hat{\theta}_n^0)\|^2_{J_n^{-1}(\hat{\theta}_n^0)} \qquad (4.3).$$

C'est sous cette forme que RAO (1947) proposa, le premier, ce test, repris dans RAO (1965). Cela explique qu'on le rencontre parfois sous le nom de <u>test de RAO</u>. D'autre part, vue l'expression (4.3) de la statistique du test, ce test porte également le nom de <u>test du score.</u>

4.5. Exemple.

Reprenons encore une fois le même exemple.

Les équations (4.1) s'écrivent :

$$\begin{cases} \dfrac{N_1}{p_1} - \dfrac{N_4}{p_4} + \mu = 0, \\ \dfrac{N_2}{p_2} - \dfrac{N_4}{p_4} - \mu = 0, \\ \dfrac{N_3}{p_3} - \dfrac{N_4}{p_4} = 0, \\ p_1 = p_2 . \end{cases}$$

On vérifie sans difficulté qu'elles ont pour solution unique :

$$\hat{\theta}^0_n = \frac{1}{n} \begin{bmatrix} \dfrac{N_1 + N_2}{2} \\ \dfrac{N_1 + N_2}{2} \\ N_3 \end{bmatrix} , \quad \hat{\mu}_n = n \, \frac{N_1 - N_2}{N_1 + N_2} .$$

On a déjà déterminé en 3.5 $\Sigma(\theta) = (p_1 + p_2) - (p_1 - p_2)^2$;

on en déduit $\Sigma(\hat{\theta}^0_n) = \dfrac{N_1 + N_2}{n}$.

Il vient donc : $M_n = \dfrac{(N_1 - N_2)^2}{N_1 + N_2}$.

<u>Variante</u> :

Si l'on détermine maintenant M_n à partir du score, on écrit :

$$S_n(X,\theta) = \begin{bmatrix} \dfrac{N_1}{p_1} - \dfrac{N_4}{p_4} \\ \dfrac{N_2}{p_2} - \dfrac{N_4}{p_4} \\ \dfrac{N_3}{p_3} - \dfrac{N_4}{p_4} \end{bmatrix} \Rightarrow S_n(X,\hat{\theta}^0_n) = n \begin{bmatrix} \dfrac{N_1 - N_2}{N_1 + N_2} \\ -\dfrac{N_1 - N_2}{N_1 + N_2} \\ 0 \end{bmatrix} .$$

Ensuite, de l'expression de $J^{-1}(\theta)$ donnée en 3.5, on déduit :

$$J^{-1}(\hat{\theta}_n^0) = \begin{bmatrix} \frac{N_1+N_2}{2n}\left(1-\frac{N_1+N_2}{2n}\right) & -\left(\frac{N_1+N_2}{2n}\right)^2 & -\frac{(N_1+N_2)N_3}{2n^2} \\ -\left(\frac{N_1+N_2}{2n}\right)^2 & \frac{N_1+N_2}{2n}\left(1-\frac{N_1+N_2}{2n}\right) & -\frac{(N_1+N_2)N_3}{2n^2} \\ -\frac{(N_1+N_2)N_3}{2n^2} & -\frac{(N_1+N_2)N_3}{2n^2} & \frac{N_3}{n}\left(1-\frac{N_3}{n}\right) \end{bmatrix}.$$

Enfin, on vérifie bien :

$$\frac{1}{n}\,{}^tS_n(X,\hat{\theta}_n^0)\;J^{-1}(\hat{\theta}_n^0)\;S_n(X,\hat{\theta}_n^0) = \frac{(N_1-N_2)^2}{N_1+N_2} = M_n .$$

<u>Application numérique</u> : Avec les données déjà utilisées, on obtient :

$$m_n = \frac{(n_1-n_2)^2}{n_1+n_2} = \frac{64}{102} \simeq 0{,}627.$$

On a donc obtenu $r_n = 0{,}628$, $w_n = 0{,}629$ et $m_n = 0{,}627$; ces valeurs ne diffèrent qu'au millième et conduisent à l'acceptation de H_0.

5. RESULTATS COMPLEMENTAIRES

5.1. Lien avec le test du khi-deux.

Si le modèle statistique considéré est un modèle multinomial à k catégories, on peut poser $\theta = (p_1, \ldots, p_{k-1})$, et l'hypothèse nulle $h(\theta)=0$ peut être testée au moyen du test du khi-deux.

La statistique de test est $K_n = \sum_{j=1}^{k} \frac{(N_j - n\hat{p}_j^0)^2}{n\hat{p}_j^0}$, où $\hat{p}_j^0$ désigne l'E.M.V. de p_j sous H_0.

Compte-tenu des propriétés asymptotiques bien connues de K_n, on doit avoir, pour n grand et si H_0 est vraie, K_n voisin de R_n, W_n et M_n.

En réalité, on a la propriété suivante :

Propriété 5 : Dans un modèle multinomial, $K_n = M_n$, $\forall n \in \mathbb{N}^*$.

En effet, dans le modèle multinomial, les éléments diagonaux de $J^{-1}(\theta)$ sont les quantités $p_j(1-p_j)$ $(j=1,\ldots,k-1)$, et les éléments non diagonaux les quantités $-p_j p_\ell$ (voir 3.5) ; par ailleurs, l'élément générique de $S_n(X,\theta)$ est $\frac{N_j}{p_j} - \frac{N_k}{p_k}$, $(j=1,\ldots,k=1)$, de sorte que l'on peut vérifier :

$$Z(\theta) = \frac{1}{n}\, {}^t S_n(X,\theta) J^{-1}(\theta) S_n(X,\theta) = \frac{1}{n} \sum_{j=1}^{k} \frac{N_j^2}{p_j} - n.$$

Il s'ensuit que $M_n = Z(\hat{\theta}_n^0) = \frac{1}{n} \sum_{j=1}^{k} \frac{N_j^2}{\hat{p}_j^0} - n = K_n$.

D'autre part, le nombre de d.d.ℓ. de la statistique K_n est bien égal à q.

Exemple : Dans l'exemple utilisé jusqu'ici, on a $\hat{p}_1^0 = \hat{p}_2^0 = \frac{N_1 + N_2}{2n}$, $p_3^0 = \frac{N_3}{n}$ et $\hat{p}_4^0 = \frac{N_4}{n}$.

On a donc : $$K_n = \frac{\left(N_1 - \frac{N_1 + N_2}{2}\right)^2}{\frac{N_1 + N_2}{2}} + \frac{\left(N_2 - \frac{N_1 + N_2}{2}\right)^2}{\frac{N_1 + N_2}{2}}$$

$$= \frac{(N_1 - N_2)^2}{N_1 + N_2} = M_n .$$

5.2. <u>Relations entre les trois statistiques de test.</u>

<u>Propriété 6</u> : Si le logarithme de la vraisemblance du modèle considéré est une fonction quadratique en θ et si $h(\theta)$ est une application affine, on a alors :

$$R_n = W_n = M_n \quad , \quad \forall n \in \mathbb{N}^* .$$

La première condition énoncée ci-desus s'écrit

$$\text{Log } L(x,\theta) = C - \frac{1}{2} {}^t\theta Q \theta , \tag{5.1}$$

où Q est carrée d'ordre p, symétrique et définie-positive ; c'est en particulier le cas d'une vraisemblance multinormale, si l'on suppose connue la matrice des variances - covariances.

La seconde condition s'écrit $h(\theta) = A\theta + B$, (5.2)
où A est $q \times p$ et B est $q \times 1$.

BUSE (1982) a établi cette propriété en supposant (5.2) et dans le cadre du modèle linéaire (supposant toujours connue la matrice des variances-covariances) ; ENGLE (1984) l'a montrée pour toute vraisemblance vérifiant (5.1), mais pour une hypothèse nulle plus restrictive que (5.2). Il est clair que le résultat reste vrai en supposant seulement (5.1) et (5.2).

<u>Propriété 7</u> : Dans le modèle linéaire, lorsque la matrice des variances-covariances des erreurs est à estimer et que $h(\theta)$ est une application affine, on a :

$$M_n \leqslant R_n \leqslant W_n \quad , \quad \forall n \in \mathbb{N}^* .$$

D'abord établi dans divers cas particuliers (voir, notamment, BERNDT and SAVIN, 1977) ce résultat fut donné dans sa forme générale par BREUSCH (1979). Il est clair que, pour n fini, cela peut entraîner des conclusions contradictoires selon la statistique de test choisie.

5.3. Approche basée sur la géométrie.

L'idée sous-jacente à ces trois tests est que, si H_0 est vraie, $\hat{\theta}_n$ et $\hat{\theta}_n^0$ sont proches ; en fait, chacune des trois statistiques R_n, W_n et M_n constitue une mesure particulière de l'écart soit entre $\hat{\theta}_n$ et $\hat{\theta}_n^0$, soit entre $h(\hat{\theta}_n)$ et $h(\hat{\theta}_n^0)$, cette mesure étant liée à la fonction Log $L(x,\theta)$. On trouvera une très intéressante illustration de ceci dans BUSE (1982), basée sur l'étude de la courbure de la fonction $L(x,\theta)$.

5.4. Aspects logiciels.

Sur le plan du calcul, les différences entre les statistiques de test sont les suivantes :

. la statistique R_n du test du rapport de vraisemblances nécessite deux estimations maximum de vraisemblance ($\hat{\theta}_n$ et $\hat{\theta}_n^0$), mais pas d'inversion de matrice ;

. la statistique W_n du test de WALD n'utilise qu'une seule estimation ($\hat{\theta}_n$, sans contrainte), mais nécessite deux inversions de matrices (d'abord $J(\hat{\theta}_n)$, puis $\Sigma(\hat{\theta}_n)$) ;

. la statistique M_n du test des multiplicateurs de LAGRANGE utilise également une seule estimation ($\hat{\theta}_n^0$, sous la contrainte H_0) et requiert une inversion de matrice (celle de $J(\hat{\theta}_n^0)$).

Pour la mise en oeuvre de ces tests, un logiciel doit donc disposer, entre autres, de procédures de calcul matriciel (produit et inversion), ainsi que de procédures d'optimisation de fonctions, notamment sous contraintes. Notons qu'en général, l'optimisation d'une fonction est réalisée, sur les logiciels les plus courants, par la méthode de NEWTON-RAPHSON qui utilise à chaque itération l'inverse d'une matrice de même nature que $J(\theta)$ (voir, par exemple, MONFORT, 1982).

Signalons que le logiciel GLIM (Generalized Linear Interactive Modelling) fournit directement deux des trois statistiques : la statistique R_n du test du rapport de vraisemblances est égale à la <u>déviance</u> fournie par GLIM (voir Mc CULLAGH and NELDER, 1983), à condition de prendre comme modèle de référence le modèle saturé (modèle implicite comprenant autant de paramètres que d'observations) et comme modèle ajusté celui correspondant à l'hypothèse nulle H_0 ; d'autre part, la statistique M_n du test des multiplicateurs de LAGRANGE est égale au khi-deux donné par GLIM, en prenant toujours les mêmes modèles, de référence et ajusté (voir PREGIBON, 1982).

Par contre, la statistique W_n du test de WALD n'est pas fournie directement ; toutefois, GLIM fournit $V_{\hat{\theta}_n}(\hat{\theta}_n)$, matrice des variances-covariances de $\hat{\theta}_n$ calculée en $\hat{\theta}_n$; ainsi, lorsque h est une application affine de θ, on obtient W_n' par calcul matriciel à partir de $V_{\hat{\theta}_n}(\hat{\theta}_n)$ (voir la remarque 2 du 3.4).

REFERENCES

Aitchison, J., and Silvey, S.D. (1958), "Maximum Likelihood Estimation of Parameters Subject to Restraints", *The Annals of Mathematical Statistics,* 29,813-828.

Berndt, E.R., and Savin, N.E. (1977), "Conflict Among Criteria for Testing Hypotheses in the Multivariate Linear Regression Model", *Econometrica,* 45,1263-1277.

Borovkov, A. (1987), "Statistique Mathématique", MIR, Moscou.

Breusch, T.S. (1979), "Conflict Among Criteria for Testing Hypotheses : Extensions and Comments", *Econometrica,* 47,203-207.

Buse, A. (1982), "The Likelihood Ratio, Wald, and Lagrange Multiplier Tests : An Expository Note", *The American Statistician,* 36,153-157.

Chow, G.C. (1983), "Econometrics", Mc Graw-Hill, International Student Edition, Singapore.

Engle, R.F. (1984), "Wald, Likelihood Ratio and Lagrange Multiplier Tests in Econometrics", in *Handbook of Econometrics,* vol.2, Elsevier Science Publisher, Amsterdam, 775-826.

Kendall, M.G., and Stuart, A. (1967), "The Advanced Theory of Statistics", vol.2, Griffin, London.

Lehmann, E.L. (1983), "Theory of Point Estimation", Wiley, New York.

Lehmann, E.L. (1986), "Testing Statistical Hypotheses", Wiley, New York.

Mc Cullagh, P. and Nelder, J.A. (1983), "Generalized Linear Models", Chapman and Hall, London.

Monfort, A. (1982), "Cours de Statistique Mathématique", Economica, Paris.

Neyman, J., and Pearson, E.S. (1928), "On the Use and Interpretation of Certain Test Criteria for the Purposes of Statistical Inference", *Biometrika,* 20A,175-240, 263-294.

Pregibon, D. (1982), "Score Tests in GLIM with Applications", in *Proceedings of the International Conference on Generalised Linear Models,* Lecture Notes in Statistics, 14, Springer-Verlag, New York, 87-97.

Rao, C.R. (1947), "Large Sample Tests of Statistical Hypotheses Concerning Several Parameters with Applications to Problems of Estimation, *Proceedings of Cambridge Philosophical Society,* 44,50-57.

Rao, C.R. (1965), "Linear Statistical Inference and Its Applications", Wiley, New York.

Serfling, R.J. (1980), "Approximation Theorems of Mathematical Statistics", Wiley, New York.

Silvey, S.D. (1959), "The Lagrangian Multiplier Test", *The Annals of Mathematical Statistics,* 30,389-407.

Silvey, S.D. (1975), "Statistical Inference", Chapman and Hall, London.

Tassi, P. (1989), "Méthodes Statistiques", Economica, Paris.

Wald, A. (1943), "Tests of Statistical Hypotheses Concerning Several Parameters when the Number of Observations is Large", *Transactions of American Mathematical Society,* 54,426-482.

Wilks, S.S. (1938), "The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses", *The Annals of Mathematical Statistics,* 9,60-62.