Statistique et analyse des données

Guy Cirier

Statistiques exhaustives complètes sur des échantillons corrélés

Statistique et analyse des données, tome 13, n° 1 (1988), p. 1-7

<http://www.numdam.org/item?id=SAD_1988__13_1_1_0>

Statistique et Analyse des Données
1988 - Vol. 13 n° 1 - p. 1-7

STATISTIQUES EXHAUSTIVES COMPLETES SUR DES ECHANTILLONS CORRELES

Guy CIRIER

74, rue Dunois
75646 PARIS CEDEX 13

Résumé: *L'étude des statistiques exhaustives sur des echantillons correlés conduit à considérer les familles complètes: lorsque la statistique exhaustive en n+1 est fonction de la variable aléatoire en n+1 et de p statistiques définies sur l'échantillon de taille n, l'évolution du rapport de vraisemblance ne dépend que de l'aléa en n+1 et des p statistiques.*

Abstract *We consider complete families for sufficient statistics : suppose that the sufficient statistic at time n+1 is a function of the random variable and of p fixed statistics at time n : it is prouved that the likelihood ratio L_{n+1}/L_n depends only on the random variable at time n+1 and on the p statistics.*

Mots-clés: *Exhaustivité, Completion, Echantillon corrélé.*
Indices de classification STMA: *01-080, 02-070, 04-090.*

1-INTRODUCTION

La notion de statistique exhaustive fut introduite vers 1920 par R.A. Fischer (5). En 1935, les célèbres travaux de Darmois (4) ont établi la relation profonde entre l'exhaustivité et la famille exponentielle dans le cas d'échantillons à tirages indépendants. Vers la même époque, Neyman (7) remarquait la factorisation de la densite dans le cas d'échantillon corrélé. Bien plus tard, en 1950, Lehmann et Scheffé (6) ont montré l'unicité des statistiques complètes. En 1954, Bahadur (1) a introduit la notion de statistique exhaustive transitive pour étudier des décisions séquentielles. Depuis, à l'exception de quelques progrès sur la loi exponentielle, l'intérêt pour ces questions semble se ralentir. On peut trouver une synthèse sur ces travaux dans J.R. Barra (3) ou dans O.Barndorff Nielsen(2).

Manuscrit reçu le 12.8.86, révisé le 10.5.88

Cet article precise le role de la completion dans l'etude de l'exhaustivité sur des échantillons correlés: si l'alea admet un resume exhaustif f_{n+1} (resp.f_n) à l'epoque n+1 (resp.n), on peut toujours écrire que f_{n+1} est fonction de x_{n+1} et de p statistiques $(g_1,g_2,\cdots,g_p)=\mathbf{g}_p$, définies sur l'échantillon en n ; si la famille est complete, on montre que la densité de la transition entre n et n+1 (le rapport des vraisemblances L_{n+1}/L_n) ne depend que de x_{n+1} et des p statistiques de $\mathbf{g}_p$. Si, de plus, f_{n+1} est relié à f_n par une relation de récurrence, les transitions sont markoviennes.

2-NOTATIONS ET STATISTIQUES EXHAUSTIVES

X_n est un alea à valeur dans $D \subseteq R^k$ dont la loi depend d'un paramètre θ. Notons $\mathbf{x}_n=(x_1,x_2,\ldots x_n)$ un échantillon corrélé de taille n ; soient $L_n=p(\mathbf{x}_n \mid \theta)$ la densité de $\mathbf{x}_n$, sachant θ dans $\Theta=R^d$, et $f_n=f_n(\mathbf{x}_n)$, un résumé exhaustif de $\mathbf{x}_n$.

En n+1,on considère l'échantillon correlé $\mathbf{x}_{n+1}$, la statistique exhaustive f_{n+1} et la densité de $\mathbf{x}_{n+1}$,$L_{n+1}=p(\mathbf{x}_{n+1} \mid \theta)$. On suppose que f_{n+1} dépend de x_{n+1} et de p statistiques réelles sur $\mathbf{x}_n$, composantes de $\mathbf{g}_p=(g_1,g_2,\cdots,g_p)$:

$$f_{n+1}=f_{n+1}(x_{n+1},\mathbf{g}_p) \qquad (1)$$

Il va de soi que l'on peut toujours écrire f_{n+1} sous cette forme : il suffit, pour le vérifier; de choisir p=nk et de poser $\mathbf{x}_n=\mathbf{g}_p(\mathbf{x}_n)$, mais les cas intéressants seront ceux où p est petit devant n, ou encore, p est fixé quand n augmente.

Le theorème de Neyman(1935) permet de factoriser la densite L lorsqu'il existe un résumé exhaustif:

$$L_n=A_n(\mathbf{x}_n)\, B_n(f_n,\theta) \qquad \text{à l'époque n}$$

$$L_{n+1}=A_{n+1}(\mathbf{x}_{n+1})\, B_{n+1}(f_{n+1},\theta) \qquad \text{à l'époque n+1}$$

où les fonctions A ne dependent pas de θ et les B ne dépendent que du résumé exhaustif et de θ.

La loi de X_{n+1}, sachant $\mathbf{x}_n$, aura une densité égale à L_{n+1}/L_n, notée:

$$L_{n+1}/L_n(\mathbf{x}_{n+1})=a(\mathbf{x}_{n+1})\, b(f_{n+1},f_n,\theta) \qquad (2)$$

avec:

$$a(\mathbf{x}_{n+1})=A_{n+1}(\mathbf{x}_{n+1})/A_n(\mathbf{x}_n)$$

et: $$b(f_{n+1},f_n,\theta)= B_{n+1}(f_{n+1},\theta) / B_n(f_n,\theta)$$

On admet que toutes les conditions mathématiques de dérivation ou d'intégration sont remplies pour justifier les calculs.

3-STATISTIQUES COMPLETES.

La definition d'une famille complète s'écrit ici (voir Barra):

Définition

La famille de lois de X_{n+1}, sachant $\mathbf{x}_n$, de densité L_{n+1}/L_n, est complète si, pour toute statistique $W=w(\mathbf{X}_{n+1})$, vérifiant : $\int_D w(\mathbf{x}_{n+1})\{L_{n+1}/L_n\}(\mathbf{x}_{n+1})\,dx_{n+1}=0$; pour tout θ dans Θ et pour tout $\mathbf{x}_n$ fixe, on a : $W=0$.

4-STATISTIQUES EXHAUSTIVES COMPLETES

Rappelons que, si la loi est complète, il y a unicité de la statistique ayant une espérance fixée d'après le résultat de Lehmann-Scheffé (7)

Proposition 1

Soient une statistique exhaustive f_{n+1}, à l'époque $n+1$, de la forme $f_{n+1}=f_{n+1}(x_{n+1}, \mathbf{g}_p)$, et f_n, à l'époque n; si la famille de lois de X_{n+1}, sachant $\mathbf{x}_n$, de densité L_{n+1}/L_n, est complète et a un support D indépendant de $\mathbf{x}_n$, alors, cette densite s'écrit :

$$L_{n+1}/L_n(\mathbf{x}_{n+1})=a(x_{n+1}, \mathbf{g}_p, f_n)\,b(f_{n+1}, f_n, \theta) \qquad (3)$$

où a ne dépend pas de θ et où b ne depend que de f_{n+1}, f_n et θ .En outre, f_n s'exprime en fonction des seules statistiques $\mathbf{g}_p$.

Démonstration

*Il suffit de démontrer que le terme $a(\mathbf{x}_{n+1})$ dans L_{n+1}/L_n, dans la formule (2), ne dépend que de x_{n+1}, $\mathbf{g}_p$, f_n . En ce cas, la relation (3) est vérifiée et, en tenant compte de la relation (1), l'équation :

$$\phi(\mathbf{g}_p, f_n)=\int_D \{L_{n+1}/L_n\}(\mathbf{x}_{n+1})\,dx_{n+1}-1=0$$

détermine une relation entre $\mathbf{g}_p$ et f_n vérifiée quel que soit θ.

*Soit la relation d'équivalence entre deux échantillons fixés de taille n :

$$\mathbf{x}_n \sim \mathbf{y}_n$$

si et seulement si :

$$f_n(\mathbf{x}_n)=f_n(\mathbf{y}_n)$$

$$g_i(\mathbf{x}_n)=g_i(\mathbf{y}_n) \qquad i=1,\ldots..p$$

Cette relation induit les relations suivantes :

-sur les resumés exhaustifs, en vertu de la relation (1) :

$$f_{n+1}(x_{n+1}, \mathbf{x}_n)=f_{n+1}(x_{n+1}, \mathbf{y}_n) \qquad \text{pour tout } x_{n+1} \text{ de D.}$$

-et, sur les densités :

$$\{L_{n+1}/L_n\}(x_{n+1}, \mathbf{y}_n)=a(x_{n+1}, \mathbf{y}_n)b(f_{n+1}, f_n, \theta)$$
$$=a(x_{n+1}, \mathbf{y}_n)/a(x_{n+1}, \mathbf{x}_n)\,\{L_{n+1}/L_n\}(x_{n+1}, \mathbf{x}_n).$$

*Or, on a, pour tout $\mathbf{x}_n$ fixe :

$$\int_D \{L_{n+1}/L_n\}(x_{n+1}, \mathbf{x}_n).\,dx_{n+1}=1$$

De même, pour tout $\mathbf{y}_n \sim \mathbf{x}_n$:

$$\int_D \{L_{n+1}/L_n\}(x_{n+1}, y_n)\, dx_{n+1}=1$$

mais, la relation sur les densités permet d'écrire :

$$\int_D a(x_{n+1}, y_n)/a(x_{n+1}, \mathbf{x}_n)\{L_{n+1}/L_n\}(x_{n+1}, \mathbf{x}_n)dx_{n+1}=1$$

dès que $\mathbf{x}_n \sim y_n$.

En retranchant la dernière équation de la première, on obtient :

$$\int_D (1-a(x_{n+1}, y_n)/a(x_{n+1}, \mathbf{x}_n))\{L_{n+1}/L_n\}(x_{n+1}, \mathbf{x}_n)dx_{n+1}=0$$

Comme la statistique est complète, on en déduit que : $1-a(x_{n+1}, y_n)/a(x_{n+1}, \mathbf{x}_n)=0$ dès que $\mathbf{x}_n \sim y_n$.

Remarque:

1-En fait, la relation d'équivalence ne porte que sur $\mathbf{g}_p$, puisque f_n s'exprime uniquement en fonction de $\mathbf{g}_p$.

2-La proposition conduit a rechercher les statistiques $\mathbf{g}_p$ sur $\mathbf{x}_n$ les plus globales possibles, au sens où le nombre p de composantes de $\mathbf{g}_p$ distinctes doit être le plus petit possible: $\mathbf{g}$ sera plus global que $\mathbf{h}$ si:

$$f_{n+1}(x_{n+1}, \mathbf{g}_p)=f_{n+1}(x_{n+1}, \mathbf{h}_q) \qquad (4)$$

pour tout x_{n+1}, avec $p<q$.

Par exemple, on peut eventuellement trouver une fonction g qui soit fonction de plusieurs fonctions h, et ainsi de suite. Si p est minimum, on dira alors que $\mathbf{g}_p$ est la memoire de X_{n+1} sur $\mathbf{x}_n$.

3-Dans le cas où $f_{n+1}=f_{n+1}(x_{n+1}, f_n)$, la densité de X_{n+1}, sachant $\mathbf{x}_n$, ne dépendra que de x_{n+1} et de f_n.

5-INDEPENDANCE DE X_{n+1} ET DE $\mathbf{X}_n$ CONDITIONNELLEMENT A' $\mathbf{g}_p$.

Ceci résulte du fait que D ne dépend pas de $\mathbf{x}_n$ et de (3).

Proposition 2.

Sous les conditions de la proposition 1, X_{n+1} et $\mathbf{X}_n$ sont independants conditionnellement à $\mathbf{g}_p$.

Démonstration:

Soient A un évènement ne dépendant que de X_{n+1} et B un évènement ne dépendant que de $\mathbf{X}_n$. On se fixe $\mathbf{g}_p$ et on complète la base formée par cet ensemble de composantes par $(g_{p+1}\cdots g_{nk})=\mathbf{g}_{nk/p}$, de sorte que $(x_{n+1}, \mathbf{x}_n)$ soit en bijection avec $(x_{n+1}, \mathbf{g}_{nk})$. Soit J le jacobien de cette transformation : il ne dépend que de $\mathbf{x}_n$. La probabilite de A, B sachant $\mathbf{g}_p$ s'écrit :

$$\text{Prob}\{A, B/\mathbf{g}_p\}=\int_A \int_{\mathbf{g}_{nk/p}{}^{-1}(B)/\mathbf{g}_p} J\{L_{n+1}/L_n\}L_n\, d\mathbf{g}_{nk/p}\, dx_{n+1}$$

Or, $J L_n$ ne dépend que de $\mathbf{x}_n$ et $\{L_{n+1}/L_n\}$, que de X_{n+1} et de g_p, et est ainsi independant de $\mathbf{g}_{nk/p}$:

$$\text{Prob}\{A, B/\mathbf{g}_p\}=\int_{A/\mathbf{g}_p}\int_{\mathbf{g}_{nk/p}{}^{-1}(B)/\mathbf{g}_p} (x_{n+1}, \mathbf{x}_n)J L_n\, d\mathbf{g}_{nk/p}\, dx_{n+1}$$
$$=\int_{A/\mathbf{g}_p}\{L_{n+1}/L_n\}(x_{n+1}, \mathbf{x}_n)\}\{\int_{\mathbf{g}_{nk/p}{}^{-1}(B)/\mathbf{g}_p} J L_n d\mathbf{g}_{nk/p}\}dx_{n+1}$$

$$= \int_{A/g_p} \{L_{n+1}/L_n\}(x_{n+1}, \mathbf{x}_n)\} \, dx_{n+1} \int_{g_{nk}/p^{-1}(B)/g_p} J L_n \, dg_{nk/p}$$

$$= \text{Prob}\{A/g_p\} \text{Prob}\{B/g_p\}$$

d'ou, le resultat.

6-GENERALISATION ET INTERPRETATION

Dans ce qui precede, le domaine D est indépendant de $\mathbf{x}_n$; l'examen des démonstrations montre que l'on peut assouplir cette condition restrictive et la remplacer par: D est fonction de g_p, sans changer les conclusions obtenues.

En résumé, l'échantillon $\mathbf{x}_n$ se décompose en deux sous échantillons: le premier, g_p, conditionne X_{n+1}, et le second, engendre par g_{nk}, est indépendant de X_{n+1}, conditionnellement au premier.

On peut aussi constater, sur la structure de L_{n+1}/L_n, que les seules statistiques exhaustives possibles sont :

-à l'époque n, des fonctions de g_p,

-à l'époque n+1, des fonctions de g_p et de x_{n+1},

mais , on ne peut ,actuellement , preciser le role de g_p dans f_n car les deux relations entre g_p et f_n :

$$f_{n+1} = f_{n+1}(x_{n+1}, g_p)$$

et :

$$\phi(g_p, f_n) = 0$$

ne permettent pas, sans hypothèses supplémentaires, d'exprimer g_p en fonction de f_n Par exemple, notons $\mathbf{f}_n = (f_1, \ldots f_n)$; si $f_{n+1} = f_{n+1}(x_{n+1}, \mathbf{f}_n)$, on a: $\mathbf{f}_n = g_p$, mais la relation $\phi(\mathbf{f}_n, f_n) = 0$ peut être vérifiée, a priori , par des statistiques très variées sur $\mathbf{f}_n$.

7-RECURRENCE DES RESUMES EXHAUSTIFS

On suppose maintenant que f_n et f_{n+1} sont liés par une relation de récurrence du type:

$$f_{n+1} = f_{n+1}(x_{n+1}, f_n) \qquad (5)$$

et, que cette relation réalise une bijection entre f_{n+1} et x_{n+1}, pour f_n fixé.

On retrouve en ce cas un résultat indiqué par Bahadur, sans démonstration, en 1954, et sans supposer la complétion

Définition

Soit f_n, une suite de statistiques exhaustives ; f_n est transitive si, pour tout évènement A, ne dépendant que de X_{n+1}, la probabilité conditionnelle de A, sachant $\mathbf{x}_n$, ne dépend que de f_n.

En ce cas :

Proposition 3 :

Soit une suite de statistiques exhaustives completes, reliées par la relation (5); f_n est transitive.

Demonstration:
Il suffit de remplacer f_{n+1} par (5) dans la formule (3).

Exemple:
Pour que la moyenne soit exhaustive et complète, elle doit, au moins, avoir des transitions markoviennes.

8-EXEMPLES

Les exemples qui suivent, ne font qu'adapter au cas qui nous interésse, des résultats connus sur la famille exponentielle et permettent de caracteriser completement la loi d'une transition en fonction de x_{n+1}, f_n et f_{n+1}.

Exemple1
Determinons $a(x_{n+1}, g_p, f_n)$ quand $b(f_{n+1}, f_n, \theta)$ est définie par:

$$b(f_{n+1}, f_n, \theta)=h(\theta)\ \exp\{-<C_{n+1}(\theta), f_{n+1}>+<C_n(\theta), f_n>\}$$

l'équation $\phi(g_p, f_n)=0$ s'écrit ici:

$$\int_D h(\theta)\ \exp\{-<C_{n+1}(\theta), f_{n+1}>+<C_n(\theta), f_n>\}\ dx_{n+1}=1$$

Si f_{n+1} est en bijection avec x_{n+1} et si on suppose $C_{n+1}(\theta)=\theta$, on a une transformation de Laplace (la statistique est donc complète) et on peut resoudre aisement cette équation en cherchant:
$-u(x_{n+1}, f_{n+1})$ solution de:

$$\int_D u\ \exp\{-<\theta, f_{n+1}>\}\ dx_{n+1}=C_n(\theta)$$

$-v(x_{n+1}, f_{n+1})$ solution de:

$$\int_D v\ \exp\{-<\theta, f_{n+1}>\}\ dx_{n+1}=h(\theta)$$

On en déduit l'expression de a:

$$a=\Sigma\ (-1)^p/p!\ v^*(<u, f_n>)^{*p}$$

Exemple2
Si :

$$f_{n+1}=l_{n+1}(x_{n+1})+f_n$$

on pose : $d_n=\theta-C_n(\theta)$
Si u et v sont solutions de :

$$\int_D u\ \exp\{-<\theta, l_{n+1}>\}\ dx_{n+1}=d_n$$

$$\int_D v\ \exp\{-<\theta, l_{n+1}>\}\ dx_{n+1}=h(\theta)$$

on obtient :

$$a=\Sigma\ (1/p!)\ v^*(<u, f_n>)^{*p}$$

mais, ici , u et v ne dependent que de x_{n+1}.
Tous ces résultats sont bien conformes à ceux prévus par la proposition 1.

Remerciements: l'auteur remercie les rapporteurs de cet article qui ont contribué à en améliorer le texte.

Bibliographie:

(1)Bahadur , R., Sufficiency and statistical decision , A.M.S. , 1954 , 25 ,pp. 423-462.
(2)Barndoff Nielsen , O., Information and exponential families , Wiley , 1978.
(3)Barra , J .R., Notions fondamentales de statistique mathématique , Dunod , 1971.
(4)Darmois , G., Sur les lois de probabilité à estimation exhaustive ",C.R.A.S. ,1935 , 200 , pp. 12-65.
(5)Fischer , R.A., "A mathematical examination of the method of determining the accuracy of an observation by the mean error and by the square error." , M.N.R.Astron.Soc., 1920, 80 , pp. 758-780.
(6)Lehmann, E.L. and Scheffé, H., Completness similar regions and unbiaised estimation" , Sankhya, 1950 , 10 ,pp. 305-340.
(7)Neyman, J., Su un theorema concernente le cosiddette statistiche sufficiente. , Inst. Ital. Atti. Giorn., 1935, pp .320-344.