Statistique et analyse des données

Michel Maurin

Tests statistiques sur les distributions de Dirichlet

Statistique et analyse des données, tome 9, n° 3 (1984), p. 45-74

<http://www.numdam.org/item?id=SAD_1984__9_3_45_0>

Statistiques et Analyse de données
1984 - Vol. 9 n° 3 - pp. 45-74

TESTS STATISTIQUES SUR LES DISTRIBUTIONS DE DIRICHLET

MAURIN Michel

Centre d'Evaluation et de Recherches des Nuisances et de l'Energie, I.R.T.
109 Av. S. ALLENDE, BP 75 - 69672 BRON CEDEX

Résumé : La distribution de Dirichlet et la structure statistique associée sont définies sur le simplexe unité de l'espace vectoriel R^M ($M \geq 2$). Elles font intervenir M scalaires a_i réels positifs. Dans cette étude on examine les tests statistiques d'hypothèses (unilatéraux et bilatéraux) sur l'un d'entre eux, les autres étant connus, ou fantômes (paramètres importuns).

On examine aussi le cas où le paramétrage est le suivant :

$$A = \sum_1^M a_i, \quad \theta_i = a_i/A \quad i = 1,\ldots,M, \quad \sum_1^M \theta_i = 1.$$

Les problèmes distributionnels que l'on rencontre ne sont pas résolus de manière exacte. Les développements se terminent par des questions d'ajustement classiques, au vu des premiers moments (d'ordre 1 et 2 pour les tests conditionnels) et du champ de variation de chaque statistique scalaire considérée.

Abstract : Dirichlet's distribution and its associated statistical family are defined on unit simplex of the euclidian space R^M ($M \geq 2$). They depend on M positive scalar parameters a_i. We examine statistical tests (one and two sided) about one particular a_i, the others being known or not (nuisance parameters). We study equally another parametrization

$$A = \sum_{1}^{M} a_i, \quad \theta_i = a_i/A \quad i = 1,\ldots,M, \quad \sum_{1}^{M} \theta_i = 1$$

Distributionnal problems are not exactly solved. So we end tests' determination with classical adjustment practices, knowing the first moments (order 1 and 2 for conditionnal tests) and the statistic's fields.

Mots clés : distribution de Dirichlet, structure statistique exponentielle, tests d'hypothèse, paramètres fantômes, tests conditionnels, tests semblables.

Indices de classification : STMA : 05-020 ; 03-070; 03-140.

INTRODUCTION

La structure de Dirichlet est associée à la distribution de Dirichlet définie sur le simplexe unité de tout espace vectoriel R^M ($M \geq 2$) ; elle est paramétrée par un vecteur a de R_+^{*M}, à savoir M scalaires de R_+^*. On imagine donc aisément que les tests d'hypothèse sur un paramètre donné (mettons a_1) peuvent avoir lieu en présence de paramètres fantômes. Nous examinons ci-après quelques-uns de ces tests, le travail se déroule en trois temps :

- utilisation des propriétés de la distribution de Dirichlet pour réduire le nombre de paramètres ;
- tests sans paramètres fantômes ;
- en cas de paramètres fantômes, présentation d'une méthode approchée applicable aux paramètres naturels des distributions de structures exponentielles d'ordre 2, (résolution approchée du problème de Lehmann-Scheffe).

Tout au long de cette présentation nous rencontrerons des problèmes distributionnels, problèmes de nature Analytique qui sont permanents en Statistique. Ils sont ici résolus de manière approchée et classique en recherchant des ajustements des lois connaissant leurs premiers moments, et le domaine correspondant.

Sur le plan des notations les sommations $\sum_{i=1}^{M}$ et produits $\prod_{i=1}^{M}$ sont effectués sur l'ensemble des indices i = 1,..., M, sauf mention contraire explicite (par exemple $\Sigma_{\mathcal{J}}$ c'est-à-dire sur $j \in \mathcal{J}$).
En III.1.1 il y a aussi les sommations $\sum_{p=1}^{P}$ effectuées sur les statistiques échantillons d'indice p compris entre 1 et P.

Les applications $\Gamma(.)$, $\Psi(.)$, $\Psi'(.)$ sont les fonctions d'EULER habituelles.

I - LA STRUCTURE STATISTIQUE DE DIRICHLET

I.1 - La densité de Dirichlet

Pour tout M supérieur ou égal à 2, on note S_1^M le simplexe unité de l'espace vectoriel R^M, y un point du simplexe de coordonnées y_i i = 1,...,M, a un point de R_+^{*M} de coordonnées a_i i = 1,...,M.

On appelle distribution de Dirichlet la distribution définie sur S_1^M, à l'aide de la densité qui pour tout y de S_1^M et a de R_+^{*M} a pour valeur :

$$d_M^*(y;a) = \frac{\Gamma(\sum_1^M a_j)}{\Gamma(M) \prod_1^M \Gamma(a_i)} \prod_1^M y_k^{a_k-1}$$

(densité par rapport à la mesure uniforme de Lebesgue sur l'hyperplan qui supporte le simplexe unité). Si l'on distingue un indice $\bar{j}$ particulier, la densité par rapport à l'élément $\prod_{j\neq\bar{j}}^M dy_j$ a pour valeur :

$$d_M^{\bar{j}}(y;a) = \frac{\Gamma(\sum_1^M a_j)}{\prod_1^M \Gamma(a_i)} \prod_1^M y_k^{a_k-1}$$

avec $y_{\bar{j}} = 1 - \sum_{j\neq\bar{j}} y_j$.

C'est bien une densité, comme l'a montré Lejeune-Dirichlet en 1839 ; Wilks l'a baptisée du nom de Dirichlet dans les années 60, [20], mais on la rencontre dans la littérature depuis des temps plus anciens sans nom particulier ; Good [9] et Johnson [11] ont relevé sa présence à plusieurs reprises, dans des listes respectives de publications d'ailleurs disjointes. Elle est bien connue dans le contexte bayésien puisqu'elle est "conjuguée" par rapport à la distribution multinomiale [2,5]. On note $D_M(a)$ la distribution de Dirichlet définie par la densité $d_M(a)$, (pour mettons $\bar{j}$=M), paramétrée par le vecteur a de coordonnées positives. Le cas M=2 est celui de la distribution Béta classique sur [0,1].

Le paramétrage peut être mis sous une forme différente. Si on pose A la somme $\Sigma^M a_i$, et pour tout i=1,...,M, $\theta_i = a_i/A$, la distribution est paramétrée par un point θ.A du produit cartésien du simplexe ouvert $\overset{\circ}{S}_1^M$ et de R_+^*, (on note $D_M(A,\theta)$ la distribution ainsi paramétrée). C'est un paramétrage analogue à celui de la distribution de Fisher-Von-Mises [8] avec un paramètre θ de direction et un paramètre A de concentration [15]. La pratique des tests statistiques sur un paramètre a_i, A ou θ_i est facilitée par les lemmes préliminaires suivants, dus aux propriétés de la distribution de Dirichlet.

Lemme 1

Pour tout M supérieur à 2, et tout L compris entre 2 et M-1, soit $Q_1\ Q_2,\ldots,Q_L$ une partition de l'ensemble {1,...,M}. A tout point y de S_1^M on fait correspondre le point z de S_1^L de coordonnées $z_l = \Sigma y_j \quad j \in Q_l \quad l = 1,\ldots,L$. Si le point y suit sur S_1^M la distribution $D_M(a)$, le point z suit sur S_1^L la distribution $D_L(b)$, avec b le point de R_+^{*L} de coordonnées $b_l = \underset{Q_l}{\Sigma}\ a_i$.

Démonstration dans Wilks [20].

Lemme 2

Pour tout M supérieur à 2 et tout J compris entre 2 et M-1, soit $\mathcal{J}$ un sous-ensemble de J indices de {1,...,M}. A tout point y de S_1^M on fait correspondre le point t de S_1^J qui a pour coordonnées

$$t_j = \frac{y_j}{\underset{\mathcal{J}}{\Sigma} y_k}, \quad j \in \mathcal{J}.$$

Si le point y suit $D_M(a)$, le point t suit $D_J(c)$, avec c le point de R_+^{*J} de coordonnées $c_j = a_j \quad j \in \mathcal{J}$.

Démonstration

a) Pour tout $s \geq 0$ et $y \in S_1^M$, soit x le point de R_+^{*M} qui pour tout $i = 1,\ldots,M$ a pour coordonnées $x_i = s\, y_i$.
Avec toute distribution $D_M(a)$ sur S_1^M, on introduit la distribution gamma $\gamma(\Sigma a_i, 1)$, quand y suit $D_M(a)$ et s suit $\gamma(\Sigma a_i, 1)$, y et s indépendants, on vérifie que x suit dans R^M la distribution produit des M distributions gamma $\gamma(a_i, 1)$ puisque

$$\frac{D(x_1,\ldots,x_n)}{D(y_1,\ldots,y_{n-1},s)} = s^{M-1}$$

et

$$\frac{\Gamma(\Sigma a_i)}{\prod_1^M \Gamma(a_j)} \prod_1^M y_k^{a_k-1} \frac{s^{\Sigma a_i - 1} e^{-s}}{\Gamma(\Sigma^M a_i)} dy_1 \ldots dy_{M-1}\, ds = \prod_1^M \frac{x_i^{a_i-1} e^{-x_i}}{\Gamma(a_i)} dx_i.$$

b) $\mathcal{J}$ étant un sous-ensemble de J indices de $\{1,\ldots,M\}$, $2 \leq J \leq M-1$, on considère la distribution marginale en x_j, $j \in \mathcal{J}$ dans R^J, sa densité a pour valeur

$$\prod_{\mathcal{J}} \frac{x_j^{a_j-1} e^{-x_j}}{\Gamma(a_j)}$$

Il en résulte que le point t du simplexe S_1^J de coordonnées $t_j = \frac{x_j}{\sum_{\mathcal{J}} x_k} = \frac{y_j}{\sum_{\mathcal{J}} y_k}$ suit la distribution de Dirichlet $D_J(c)$ avec c le point de coordonnées $c_j = a_j \quad j \in \mathcal{J}$ [10, 15, 20].

□

Remarque : on peut démontrer le lemme 1 de cette façon.

Conséquence :

Ces deux lemmes permettent de diminuer le nombre des paramètres d'une distribution de Dirichlet. Le lemme 2 permet de supprimer K paramètres ($1 \leq K \leq M-2$), le lemme 1 de ramener leur nombre à 2, ils peuvent être combinés de plusieurs façons selon le cas.

I.2 - Structure et forme canonique

On constate immédiatement que la structure associée aux distributions de Dirichlet (S_1^M, $\mathcal{B}(S_1^M)$, $\{D_M(a), a \in \mathbb{R}_+^{*M}\}$) est exponentielle complète [2,3]. Pour plus de commodité on peut la mettre sous sa forme canonique. On utilise pour cela le changement de variable $\bar{L}$ qui à tout $j = 1,\ldots,M$ fait correspondre le point τ de coordonnées $\tau_j = -\log(y_j)$; les τ_j sont les statistiques canoniques de la distribution associées aux paramètres naturels a_j. L'image par $\bar{L}$ de S_1^M est une variété V_M de R_+^M, l'image de $D_M(a)$ est la distribution $\mathcal{F}_M(a_i;1)$ définie sur V_M dont la densité par rapport à la mesure uniforme "de surface" sur V_M a pour valeur [15]:

$$\frac{d\,\mathcal{F}_M(a;1)}{d\,v_M} = \frac{\Gamma(A)}{\prod_1^M \Gamma(a_i)}\; e^{-\sum_1^M a_i\tau_i}\left(\sum_1^M e^{-2\,\tau_k}\right)^{-1/2}$$

Pour tout $j = 1,\ldots,M$ et $z_j \in \mathbb{C}$ tel que $\mathrm{Re}(z_j) > -a_j$, on sait que la transformée de Laplace de cette distribution a pour valeur

$$\mathcal{L}_\tau(z) = \prod_1^M \frac{\Gamma(a_i+z_i)}{\Gamma(a_i)} \times \frac{\Gamma(A)}{\Gamma(A+\sum_1^M z_j)} \qquad [2,15].$$

par prolongement analytique c'est une fonction méromorphe définie sur $\mathbb{C}^M$ avec pour poles simples les points d'affixes respectifs $z_j = -a_j - n$, $n \in \mathbb{N}$, $j = 1,\ldots,M$, donc dérivable à l'origine.

On utilise également les résultats suivants :

Lemme 3

Les ensembles :

$$C_M = \{\tau \in \mathbb{R}_+^{*M} : \tau_i = u_i+\delta_i, \quad \delta_i \geq 0 \quad i=1,\ldots,M, \; u \in V_M \}$$

$$C_M^o = \{\tau \in \mathbb{R}_+^{*M} : \tau_i = u_i+\delta_i, \quad \delta_i > 0 \quad i=1,\ldots,M, \; u \in V_M\}$$

sont convexes; par conséquent toute droite passant par un point à distance finie de V_M avec des coefficients directeurs positifs ou nuls, non tous nuls, n'a pas de point commun avec V_M à distance finie.

Démonstration dans [15].

□

Lemme 4

La distribution $\mathcal{T}_M(a;1)$ a pour espérance mathématique le point $\bar{\Psi}(a)$ qui pour tout $i=1,\ldots,M$ a pour coordonnées $\bar{\Psi}_i(a) = \Psi(A)-\Psi(a_i)$, $\bar{\Psi}(a) \in C_M^o$, et pour variance-covariance la matrice $\Omega(a)$ qui a pour éléments, pour tout $i,j = 1,\ldots,M$:

$$\Omega_{ij}(a) = \Psi'(a_i)\, \delta_{ij} - \Psi'(A) \; ;$$

(Ψ et Ψ' fonctions eulériennes, $\Psi = \Gamma'/\Gamma$).

Démonstration dans [15].

□

Quand la distribution D_M est paramétrée par le paramètre naturel a, il est commode d'introduire le paramétrage de la valeur moyenne défini pour tout $i=1,\ldots,M$ par

$\mu_i = E(\tau_i) = \bar{\Psi}_i(a) = \frac{\partial}{\partial a_i} \text{Log} \frac{\Gamma^M(\Sigma a_j)}{\Pi^M \Gamma(a_k)}$, et aussi un paramétrage mixte avec des a_i et des μ_j [2].

L'utilité de ces nouveaux paramétrages est due à des propriétés des structures exponentielles.

II - TESTS STATISTIQUES SUR LES PARAMETRES

Nous envisageons des tests d'hypothèse sur chacun des paramètres a_i, θ_j ou A pris individuellement, selon que les autres sont connus ou non (fantômes).

II.1 - Paramétrage en a

On suppose que a_1 est le paramètre d'intérêt. Quand les paramètres restants sont connus, on peut éventuellement se ramener à la distribution $D_2(a_1,A_2)$ en posant $A_2 = \Sigma_2^M a_i$, ou toute autre distribution $D_L(a_1, A'_2,\ldots,A'_j,\ldots,A'_L)$ (Lemme 1). Quand il y a M-2 ou moins paramètres a_k fantômes le lemme 2 nous permet de les supprimer, on peut ensuite se ramener si on le désire à une distribution $D_2(a')$. Au contraire si tous les a_j j=2,...,M sont fantômes il est avantageux de se ramener à la distribution $D_2(a_1,A_2)$ dans laquelle il n'y a plus qu'un seul paramètre inconnu.

II.2 - Paramétrage en A, θ.

Ce paramétrage permet de tester des hypothèses soit sur le paramètre A, soit sur chacun des coefficients directeurs θ_i. Nous supposerons en l'utilisant que tous les paramètres θ_i sont simultanément connus ou fantômes. A l'aide du lemme 1 on se ramène à une distribution $D_2(A,\theta_*)$ avec $\theta_* = \Sigma\theta_i \quad i \in \mathcal{J}_*$ où $\mathcal{J}_*$ est un sous-ensemble non vide de {1,...,M}; de la sorte le paramètre vectoriel θ n'intervient que par un scalaire unique. Nous envisageons le test sur A avec θ (donc θ_*) connu ou non, et le test sur θ_1 (on prend $\mathcal{J}_* \equiv \{1\}$) avec A connu ou non. Dans toutes ces applications les distributions considérées ne contiennent qu'un paramètre fantôme au plus.

Pour commencer nous présentons les tests en présence de paramètres connus, (chapitre III); dans ce cas le nombre des autres paramètres importe peu en fait. En revanche les lemmes 1 et 2 sont mis à profit dans le cas contraire (chapitre IV). Nous supposons que l'on dispose d'un échantillon de P points y^p $p = 1,\ldots,P$ du simplexe S_1^M, qui résultent de P tirages indépendants de la distribution $D_M(a)$; dans l'espace des statistiques exhaustives τ_i, le point τ suit la loi $\mathcal{F}_M(a;1)$.

III - TESTS SANS PARAMETRES FANTOMES

III.1 - Tests sur a_1.

1.1 - La vraisemblance de l'échantillon des points τ^p , $p = 1,\ldots,P$ est proportionnelle à la quantité

$$\left(\frac{\Gamma(A)}{\prod \Gamma(a_i)}\right)^P e^{-P \sum_1^M a_i \bar{\tau}_i}$$

où l'on pose $j = 1,\ldots,M$, $\bar{\tau}_j = \frac{1}{P} \sum_{p=1}^{P} \tau_j^p = -\frac{1}{P} \sum_{p=1}^{P} \text{Log } y_j^p$.

On note $\mathcal{F}_M(a;P)$ la distribution définie sur C_M du point $\bar{\tau}$ de coordonnées $\bar{\tau}_i$, $i = 1,\ldots,M$, (on a vu que C_M est convexe). C'est une distribution de structure exponentielle canonique, son espérance mathématique est évidemment égale à $\bar{\Psi}(a)$, et sa matrice de variance-covariance égale à $\Omega(a)/P$. Lorsque les paramètres a_j , $j = 2,\ldots,M$ sont connus, la statistique $\bar{\tau}_i$ est exhaustive pour a_1, la distribution marginale en $\bar{\tau}_1$ est une distribution scalaire de structure exponentielle canonique. L'estimateur du maximum de vraisemblance de a_1 est la solution de l'équation

$$\Psi(A_2+\hat{a}_1) - \Psi(\hat{a}_1) = \bar{\tau}_1$$

en posant $A_2 = \Sigma_2^M a_i$.

Avec une probabilité égale à 1 cette équation admet une racine réelle positive unique $\hat{a}_1$. D'après l'expression de la densité sous la forme canonique

$e^{-Pa_1 \bar{\tau}_1} \alpha(a)\, \beta(\bar{\tau}_1)$ il existe les tests UMP et UMPB classiques des distributions scalaires de structure exponentielle. Si l'on teste l'hypothèse a_1^0 contre a_1^1, $a_1^1 > a_1^0$, la région de rejet est de la forme : $\bar{\tau}_1 < \tau_1^*$. Pour tout $i = 1,\ldots,M$ et pour tout $z_j \in \mathbb{C}$ dont la partie réelle est positive ou nulle, la transformée de Laplace de $\mathcal{F}_{\mathbb{M}}(a;P)$ a pour valeur

$$\mathcal{L}_{\tau}(\tfrac{z}{P})]^P = \left[\frac{\Gamma(A)}{\prod_M \Gamma(a_i)} \; \frac{\prod_M \Gamma(a_j+z_j/P)}{\Gamma(A+\sum_M z_k/P)}\right]^P ,$$

c'est une fonction méromorphe dans $\mathbb{C}^M$ avec pour seuls poles simples les points $z_i = -a_i P - nP$, $n \in \mathbb{N}$, donc dérivable à l'origine. La transformée de Laplace de la distribution marginale en $\bar{\tau}_1$ a pour valeur

$$\mathcal{L}_{\tau_1}(\tfrac{z_1}{P}))^P = \left[\frac{\Gamma(A)}{\Gamma(a_1)} \; \frac{\Gamma(a_1+z_1/P)}{\Gamma(A+z_1/P)}\right]^P .$$

Le calcul de l'inverse, et de la densité de la loi de $\bar{\tau}_1$ paraît difficile, en revanche par dérivation à l'origine on peut facilement calculer les cumulants de cette distribution. On a :

moyenne	$K_1 = \Psi(A) - \Psi(a_1)$
variance	$K_2 = (\Psi'(a_1)-\Psi'(A))/P$
cum. d'ordre 3	$K_3 = (\Psi''(A) - \Psi''(a_1))/P^2$.

1.2 - Pour la mise en oeuvre des tests sur a_1, on peut alors substituer à la loi proprement dite une loi approchée ajustée sur les premiers moments (ou cumulants, [5]). Un tel ajustement n'est évidemment pas unique, pour le choisir on peut notamment tenir compte du domaine parcouru par la statistique dont on cherche la loi. En l'occurence, compte tenu du domaine des valeurs prises par $\bar{\tau}_1$, il paraît judicieux de prendre

comme loi approchée la loi log-normale définie sur R^+, ou une loi dont la densité est complétée par les polynômes orthogonaux correspondants (Figure 1), [6]. Les polynômes associés à la loi log-normale ne sont pas classiques, les premiers d'entre eux peuvent être établis par la méthode des déterminants [4], les polynômes de degré 3 et 4 sont calculés par exemple en [16]. L'ajustement d'une loi log-normale décalée sur les trois premiers moments peut être faite également avec la méthode de Wicksell [13], mais alors on ne maîtrise plus de la même façon le domaine de la loi approchée.

1.3 - On sait que pour tout α compris entre 0 et 1 il existe des tests UMP de l'hypothèse a_1^0 contre a_1^1, $a_1^1 > a_1^0$, (resp. $a_1^1 < a_1^0$), et un test UMPB de l'hypothèse a_1^1 contre a_1^0, $a_1^1 \neq a_1^0$, de niveau α. L'ajustement selon une loi approchée est un moyen d'obtenir des tests et des régions de rejet approchés. Quand on utilise l'ajustement log-normal la détermination des régions critiques est immédiate à l'aide des tables de la loi normale.

III.2 - Tests sur A.

Tous les θ_i $i = 1,\ldots,M$ étant connus, la vraisemblance de l'échantillon des τ^p indépendants est proportionnelle à la quantité :

$$\left[\frac{\Gamma(A)}{\prod_{1}^{M}\Gamma(A\theta_i)}\, e^{-A\sum_{1}^{M}\theta_j \bar{\tau}_j}\right]^p$$

La statistique $\bar{s}_\theta = \sum_{1}^{M}\theta_j \bar{\tau}_j$ est exhaustive pour A ; la distribution de $\bar{s}_\theta$ est une distribution scalaire de structure exponentielle canonique paramétrée par A. Si l'on teste l'hypothèse A^0 contre A^1, $A^1 > A^0$, la région critique est de la forme $\bar{s}_\theta < s_\theta^*$.

Pour tout $z \in \mathbb{C}$ tel que $Re(z) > -A$ la transformée de Laplace de la distribution en $\bar{s}_\theta$ a pour valeur :

$$\left[\frac{\Gamma(A)}{\prod_{1}^{M} \Gamma(A\theta_i)} \; \frac{\prod_{1}^{M} \Gamma(\theta_i(A+z/p))}{\Gamma(A+z/p)} \right]^P,$$

c'est une fonction méromorphe dérivable à l'origine. Le calcul de l'inverse (densité de $\bar{s}_\theta$), paraît difficile, en revanche on a facilement les premiers cumulants :

moyenne $\quad K_1 = \Psi(A) - \Sigma\, \theta_i\, \Psi(A\theta_i)$

variance $\quad K_2 = [\Sigma\theta_i^2\, \Psi'(A\theta_i) - \Psi'(A)]/P$

cum. d'ordre 3: $K_3 = [\Psi''(A) - \Sigma\, \theta_i^3\, \Psi''(A\theta_i)]/P^2$.

Pour tout α compris entre 0 et 1 il existe des tests UMP de l'hypothèse A^0 contre A^1, $A^1 > A^0$, (resp. $A^1 < A^0$), et un test UMPB de l'hypothèse A^0 contre A^1, $A^1 \neq A^0$, de niveau α . Compte tenu des valeurs prises par $\bar{s}_\theta$ il paraît indiqué de faire un ajustement basé comme précédemment sur la loi log-normale, ce qui permet d'avoir une approximation des régions de rejet des tests ci-dessus. Il paraît ici convenable de prendre une loi log-normale calée sur la valeur $\inf(\bar{s}_\theta) = -\Sigma\ \theta_i \operatorname{Log} \theta_i$(*) ; figure 2.

.Cas particulier des distributions de Dirichlet invariantes par le groupe Σ_M des permutations.

Les distributions invariantes par Σ_M sont paramétrées par un vecteur a dont tous les coefficients directeurs θ_i sont égaux à 1/M. Les tests précédents sur A sont applicables, la statistique $\bar{s}_\theta = \sum_{1}^{M} \bar{\tau}_i/M$ est exhaustive pour A, les premiers cumulants sont ici :

moyenne $\quad K_1 = \Psi(A) - \Psi(A/M)$

variance $\quad K_2 = [\frac{1}{M}\, \Psi'(\frac{A}{M}) - \Psi'(A)]/P$

cum d'ordre 3: $K_3 = [\Psi''(A) - \frac{1}{M^2}\ \Psi''(\frac{A}{M})]/P^2$

(*) La loi "log-normale calée sur x_0" est la translatée de la loi ordinaire par la translation $x \to x+x_0$.

Ceci rectifie la présentation incorrecte de [15], chapitre VI.4.4.

. Remarque - Quand le paramètre de direction θ d'une distribution de Fisher-Von-Mises est connu, on connaît une statistique exhaustive pour A, et la fonction de répartition de sa distribution ([8] chapitre IV.1). C'est là un renseignement analytique dont on n'a pas l'équivalent ici ; mais il semble qu'en pratique ce soit un avantage "à la Phyrrus" compte tenu des difficultés numériques qu'entraîne l'expression exacte de la fonction de répartition, comparées à la facilité de la mise en oeuvre d'un ajustement sur la loi log-normale correctement calée.

III.3 - Tests sur θ_1.

Quand on connaît A on utilise le lemme 1 pour se ramener aux distributions $D_2(A,\theta_1)$ et $\mathcal{F}_2(A,\theta_1;P)$. La vraisemblance de l'échantillon des τ^P indépendants est proportionnelle à la quantité

$$\left[\frac{\Gamma(A)}{\Gamma(A\theta_1)\,\Gamma(A\theta_2)}\, e^{-A(\theta_1\bar{\tau}_1 + \theta_2\,\bar{\tau}_2)}\right]^P, \quad \theta_1+\theta_2 = 1$$

et

$$e^{AP\theta_1(\bar{\tau}_2-\bar{\tau}_1)}.$$

La statistique $\bar{\omega}_2 = \bar{\tau}_2-\bar{\tau}_1$ est exhaustive pour θ_1. Le test de l'hypothèse θ_1^o contre θ_1^1, $\theta_1^1 > \theta_1^o$ est un test UMP avec une région critique de la forme $\bar{\omega}_2 > \omega_2^*$. Pour tous z_1 et z_2 complexes tels que $Re(z_1)$ et $Re(z_2)$ sont positifs ou nuls, la transformée de Laplace de la distribution $\mathcal{F}_2(A,\theta_1;P)$ a pour valeur :

$$\tilde{\mathcal{L}}_{\tau_1\tau_2}(z_1,z_2) = \left[\frac{\Gamma(A)}{\Gamma(A\theta_1)\,\Gamma(A\theta_2)}\,\frac{\Gamma(A\theta_1+z_1/P)\,\Gamma(A\theta_2+z_2/P)}{\Gamma(A+(z_1+z_2)/P}\right]^P,$$

c'est une fonction méromorphe dont les poles simples sont $z_1 = -PA\theta_1-Pn_1$, $z_2 = -PA\theta_2-Pn_2$, n_1, $n_2 \in \mathbb{N}$. Si l'on fait le changement de variables dans $\mathbb{C}^2$

$$\begin{cases} \omega_1 = \tau_1 + \tau_2 \\ \omega_2 = \tau_2 - \tau_1 \end{cases}$$

la distribution image en $\omega_1\ \omega_2$ a pour fonction génératrice des moments l'image de $\tilde{\mathcal{L}}_{\tau_1,\tau_2}$ telle que $\bar{\tau}_1 z_1 + \bar{\tau}_2 z_2 = \bar{\omega}_2 Z_1 + \bar{\omega}_2 Z_2$, elle a donc pour valeur

$$\tilde{\mathcal{L}}_{\omega_1,\omega_2}(Z_1, Z_2) = \check{\mathcal{L}}_{\tau_1\tau_2}(Z_1 - Z_2,\ Z_1 + Z_2)\ ;$$

on en déduit que la fonction génératrice des moments de la distribution marginale en $\bar{\omega}_2$ a pour valeur

$$\check{\mathcal{L}}_{\omega_1\omega_2}(0, Z_2) = \tilde{\mathcal{L}}_{\tau_1\tau_2}(-Z_2, Z_2)$$

$$= \left[\frac{\Gamma(A\theta_1 - Z_2/P)\ \Gamma(A\theta_2 + Z_2/P)}{\Gamma(A\theta_1)\ \Gamma(A\theta_2)}\right]^P .$$

C'est une fonction méromorphe dérivable à l'origine.

De même le calcul de l'inverse (densité de $\bar{\omega}_2$) paraît difficile, par contre on a facilement les premiers cumulants :

moyenne $\quad K_1 = \Psi(A\theta_1) - \Psi(A\theta_2)$

variance $\quad K_2 = (\Psi'(A\theta_1) + \Psi'(A\theta_2))/P$

cum.d'ordre 3 : $K_3 = (\Psi''(A\theta_1) - \Psi''(A\theta_2))/P^2$

cum.d'ordre 4 : $K_4 = (\Psi'''(A\theta_1) + \Psi'''(A\theta_2))/P^3$

Comme dans les cas précédents, on sait que pour tout α compris entre 0 et 1 il existe un test UMP de l'hypothèse θ_1^o contre θ_1^1, $\theta_1^1 > \theta_1^o$, (resp. $\theta_1^1 < \theta_1^o$) et un test UMPB de l'hypothèse θ_1^o, $\theta_1^1 \neq \theta_1^o$, de niveau α .

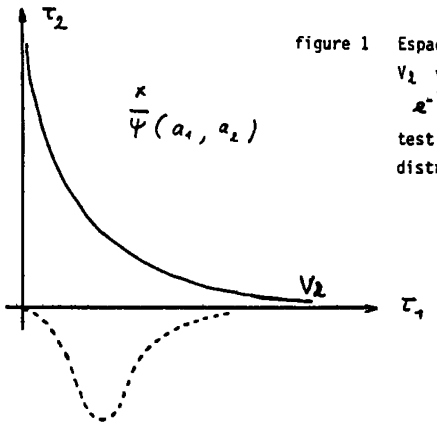

figure 1 Espace des statistiques canoniques, M=2 , V_2 variété définie par l'équation $e^{-\tau_1} + e^{-\tau_2} = +1$; $\tau_1 > 0$, $\tau_2 > 0$; test sur a_1 , (III-1) distribution marginale de $\overline{\overline{\tau}}_1$.

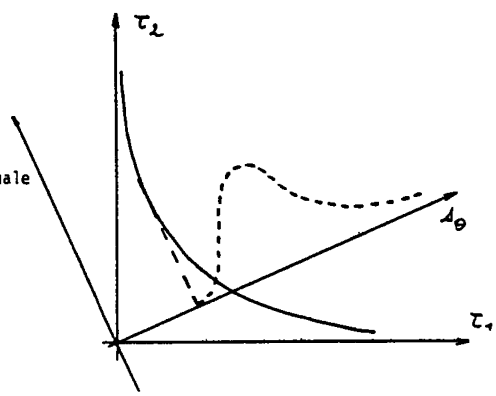

figure 2 test sur A , (III-2) distribution marginale de $\overline{s}_\theta$.

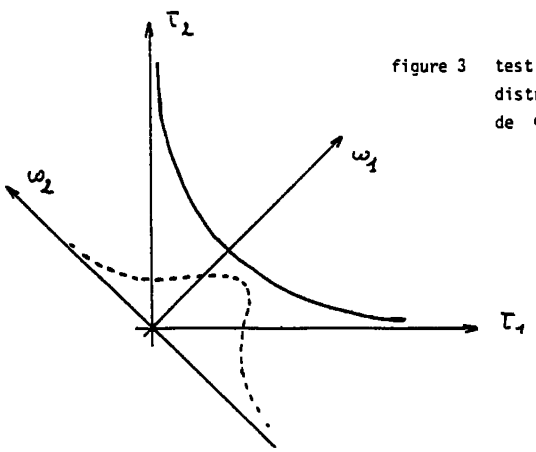

figure 3 test sur Θ_1 , (III-3) distribution marginale de $\overline{\overline{\omega}}_2$

De la même façon on a une approximation des régions de rejet à l'aide d'une loi approchée ajustée sur les premiers cumulants. Dans le cas présent la statistique $\bar{\omega}_2$ prend ses valeurs dans R, il paraît convenable de prendre comme loi d'ajustement une loi normale complétée au besoin par les polynômes d'Hermite associés, [6], figure 3.

IV - TESTS EN PRESENCE DE PARAMETRES FANTOMES

IV.1 - Généralités

La présence de paramètres fantômes complique l'élaboration des tests, en revanche de tels problèmes se présentent davantage en pratique. La méthodologie générale consiste à donner la valeur de l'hypothèse nulle au paramètre d'intérêt et à conditionner la distribution par rapport à des statistiques exhaustives pour les paramètres inconnus [5], ce qui revient en gros à remplacer des paramètres sur lesquels on ne sait rien par des statistiques "astucieuses". Ci-après nous nous ramenons pour simplifier à un paramètre fantôme unique. La méthode des tests conditionnels est par excellence adaptée aux distributions des structures exponentielles mises sous forme canonique. Celles-ci possèdent des statistiques exhaustives pour les paramètres naturels inconnus quand les paramètres naturels d'intérêt sont fixés par hypothèse [3,5,14], et avec elles il existe des tests qui ont une structure de Neyman relativement aux statistiques exhaustives pour les paramètres fantômes [5, 14, 17, 19] ; il s'agit du problème de Lehmann-Scheffe.

Si l'on prend par exemple le paramètre d'intérêt a_1, les autres a_j $j = 2,\ldots,M$ et $A_2 = \Sigma_2^M a_j$ étant inconnus, on a vu que la statistique $\bar{\tau}_2$ est exhaustive pour A_2, sous une hypothèse $a_1 = a_1^o$. Il en résulte que pour tout α compris entre 0 et 1 il existe des tests de l'hypothèse a_1^1 contre a_1^o, $a_1^1 < a_1^o$ (resp. $a_1^1 > a_1^o$, $a_1^1 \neq a_1^o$) qui sont semblables, donc UMPB [19], de niveau α. On envisage ensuite les paramètres d'intérêt A puis θ_1 ; les statistiques $\bar{\omega}_2$ et $\bar{s}_1 = \theta_1 \bar{\tau}_1 + \theta_2 \bar{\tau}_2$ sont respectivement

exhaustives pour θ_1 et A (avec les hypothèses respectives A^o et θ_1^o). Par conséquent pour chacun de ces tests on peut considérer les distributions respectives des variables $(\bar{\tau}_1, \bar{\tau}_2)$ (à savoir $\mathcal{T}_2(a;P)$),

$$\begin{pmatrix} \bar{\omega}_1 \\ \bar{\omega}_2 \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix} \begin{pmatrix} \bar{\tau}_1 \\ \bar{\tau}_2 \end{pmatrix} \quad \text{et} \quad \begin{pmatrix} \bar{s}_1 \\ \bar{s}_2 \end{pmatrix} = \begin{pmatrix} \theta_1 & \theta_2 \\ -\theta_2 & \theta_1 \end{pmatrix} \begin{pmatrix} \bar{\tau}_1 \\ \bar{\tau}_2 \end{pmatrix}$$

qui sont liées aux précédentes de façon linéaire, et les conditionner respectivement par rapport à $\bar{\tau}_2$, $\bar{\omega}_2$, $\bar{s}_1$.

IV.2 - Difficultés analytiques et méthode approchée.

En pratique, à notre connaissance, [10], on ne connait l'expression analytique de la densité d'aucune de ces distributions, mais on connait l'expression de leur transformée de Laplace, méromorphe et dérivable à l'origine. A partir de ces dernières on a vu qu'il est simple de déduire les transformées de Laplace des distributions marginales ; il n'en est pas de même pour les transformées des distributions conditionnelles [12], et le calcul de chacun des moments de ces lois conditionnelles nécessite au moins le calcul d'une inverse de transformée par moment.

On est confronté à la situation analytique suivante :

a) la distribution de Dirichlet, comme toute distribution de structure exponentielle complète permet d'utiliser les tests conditionnels en présence de paramètres fantômes, (structure de Neyman) ;

b) on ne connait pas les distributions conditionnelles nécessaires pour cette mise en oeuvre ;

c) on dispose de la transformée de Laplace de la distribution elle-même.

La méthode approchée suivante est destinée à résoudre cet obstacle analytique, et à éviter le calcul d'un trop grand nombre d'inverses de transformées, elles-mêmes approchées. Elle s'articule selon les 3 points suivants à propos des paramètres naturels :

i) il revient au même de conditionner par rapport à une statistique fonction monotone de la statistique exhaustive initiale ;

ii) le paramétrage de la valeur moyenne [2] sert de guide pour le choix des fonctions monotones ci-dessus ;

iii) on remplace les lois $\mathcal{F}_2(a;P)$ par les distributions normales $\mathcal{N}(\bar{\Psi}(a), \Omega(a)/P)$ qui ont les mêmes moments d'ordre 1 et 2.

Les deux premiers points sont rigoureux, c'est dans le second qu'interviennent les paramétrages de la valeur moyenne et mixtes, dus à des propriétés de structures exponentielles canoniques. C'est dans le dernier point qu'a lieu l'approximation, elle s'appuie sur l'identité des moments d'ordre 1 et 2, et sur le théorème de la limite centrale à plusieurs dimensions [1]; cette approximation est d'autant plus précise que l'effectif P est plus élevé. On verra ci-après comment cette approximation normale iii) s'associe de manière simple avec le paramétrage de la valeur moyenne ii) pour les paramètres naturels. Sur le plan numérique l'on verra que les problèmes rencontrés sont du même ordre de difficulté que ceux d'une (seule) inversion de transformée de Laplace [7]. Les tests suivants sont une illustration de cette méthode.

IV.3 - Tests sur a_1.

3.1 - Le paramétrage initial est le paramétrage naturel a_1, A_2. Comme indiqué ci-dessus on remplace la distribution $\mathcal{F}_2(a;P)$ par la loi normale bivariée de moyenne $\bar{\Psi}(a)$ et de variance-covariance $\Omega(a)/P$. Par définition le paramétrage de la valeur moyenne est donné par les coordonnées de $\bar{\Psi}(a)$; ici on conserve le paramétrage mixte a_1 et

$$\mu_2 = \bar{\Psi}(a)_2 = E(\bar{\tau}_2) = \frac{\partial}{\partial A_2} \text{Log} \frac{\Gamma(A)}{\Gamma(a_1)\,\Gamma(A_2)} = \Psi(a_1+A_2)-\Psi(A_2).$$

On vérifie que l'équation $\Psi(a_1+A_2) - \Psi(A_2) = \bar{\tau}_2$ possède une racine réelle $\hat{A}_2(a_1,\bar{\tau}_2)$ positive unique et que l'application $\hat{A}_2$ est croissante en $\bar{\tau}_2$. On est ainsi amené à considérer la statistique de conditionnement égale à $\hat{A}_2(a_1,\bar{\tau}_2)$. Son intérêt réside dans le fait qu'à l'occasion du conditionnement par rapport à $\hat{A}_2(a_1,\bar{\tau}_2)$, donc $\bar{\tau}_2$, la distribution normale de remplacement prend une forme simplifiée ; puisque le paramètre de la valeur moyenne est précisément le vecteur espérance mathématique des statistiques exhaustives $\bar{\Psi}(a)$, [2], le même que celui de la loi normale de substitution. En particulier la distribution de $\bar{\tau}_1$ conditionnelle par rapport à $\hat{A}_2(a_1,\bar{\tau}_2)$ est la loi normale scalaire de moyenne $\Psi(a_1+\hat{A}_2)-\Psi(a_1)$ et de variance

$$\text{var}(\bar{\tau}_1/\bar{\tau}_2) = [\Psi'(a_1)-\Psi'(a_1+\hat{A}_2) - \frac{\Psi'(a_1+\hat{A}_2)^2}{\Psi'(\hat{A}_2)-\Psi'(a_1+\hat{A}_2)}] \times \frac{1}{p}.$$

3.2 - On a donc facilement obtenu une loi conditionnelle approchée, (elle ne dépend plus de A_2). Plus précisément on a les deux premiers moments approchés de cette loi ; en effet il n'est sans doute pas indiqué de conserver cette loi normale telle quelle car les valeurs prises par $\bar{\tau}_1/\bar{\tau}_2$ sont supérieures ou égales à $-\log(1-e^{-\bar{\tau}2})$, avec $0 < -\log(1-e^{-\bar{\tau}2}) < E(\bar{\tau}_1/\bar{\tau}_2)$ puisque le point $\bar{\Psi}(a_1,\hat{A}_2)$ appartient à C_2^o (lemme A); figure 4.

On est de nouveau ramené à un problème d'ajustement avec une loi scalaire ; pour une résolution rapide on peut prendre la loi log-normale correctement calée sur $-\log(1-e^{-\bar{\tau}2})$, dont on connait la moyenne $E(\bar{\tau}_1/\bar{\tau}_2) + \text{Log}(1-e^{-\bar{\tau}2})$ et la variance $\text{var}(\bar{\tau}_1/\bar{\tau}_2)$.

Pour tout α compris entre 0 et 1, on peut donc obtenir des régions critiques (approchées) des tests UMPB de l'hypothèse a_1^o contre a_1^1 avec $a_1^1 > a_1^o$, $a_1^1 < a_1^o$ $a_1^1 \neq a_1^o$, de niveau α .

Conclusions

Les tests ci-dessus utilisent les trois points mentionnés en IV.2, le codage de la valeur moyenne donne facilement une nouvelle statistique de conditionnement pour laquelle la loi normale bivariée de remplacement prend une forme simple. La mise en oeuvre dépend de la résolution numérique d'une équation dont on sait qu'elle admet une racine réelle unique, c'est tout à fait le même genre de calculs que ceux que propose Daniels pour inverser une transformée de Laplace [7]. Ce calcul unique donne les deux premiers moments de la loi conditionnelle recherchée, il en faudrait au moins trois pour en avoir autant en calculant directement les moments [12].

IV.4 - Tests sur A

Les autres tests ne portent pas sur des paramètres naturels, mais les statistiques de conditionnement sont liées de façon linéaire aux statistiques exhaustives précédentes. On prend à présent les paramètres A et θ_1, avec $\theta_2 = 1-\theta_1$. Les nouvelles variables :

$$\begin{pmatrix} \bar{\omega}_1 \\ \bar{\omega}_2 \end{pmatrix} = \mathcal{H} \begin{pmatrix} \bar{\tau}_1 \\ \bar{\tau}_2 \end{pmatrix} \quad , \mathcal{H} = \begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix}$$

sont commodes, on l'a vu, pour conditionner par rapport à $\bar{\omega}_2$. La loi de remplacement de ces variables est à présent la loi normale bivariée image [1], $\mathcal{N}(\mathcal{H}\,\bar{\Psi}(a), \mathcal{H}\,\Omega(a)\,\mathcal{H}'/P)$. D'après l'expression de $\mathcal{H}\,\bar{\Psi}(a)$ il est indiqué de prendre comme nouvelle variable de conditionnement la solution $\hat{\theta}_1(A,\bar{\omega}_2)$ de l'équation $\Psi(A\theta_1)-\Psi(A\theta_2) = \bar{\omega}_2$. On vérifie que pour tout ω_2 réel au deuxième membre, cette équation possède une racine réelle unique $\hat{\theta}_1(A,\bar{\omega}_2)$ monotone en $\bar{\omega}_2$. En conditionnant la loi normale par rapport à θ_1, donc $\bar{\omega}_2$, la loi conditionnelle approchée de $\bar{\omega}_1$ par rapport à $\hat{\theta}_1$ a pour moyenne $E(\bar{\omega}_1/\bar{\omega}_2) = 2\Psi(A)-\Psi(A\hat{\theta}_1)-\Psi(A-A\hat{\theta}_1)$ et pour variance

$$\mathrm{var}(\bar{\omega}_1/\bar{\omega}_2) = [\Psi'(A\hat{\theta}_1)+\Psi'(A\hat{\theta}_2)-4\Psi'(A) - \frac{[\Psi'(A\hat{\theta}_1)-\Psi'(A\hat{\theta}_2)]^2}{\Psi'(A\hat{\theta}_1)+\Psi'(A\hat{\theta}_2)}] \times \frac{1}{P}$$

$$\hat{\theta}_1 + \hat{\theta}_2 = 1,$$

(cette loi ne dépend plus de θ_1). On remarque par ailleurs que les valeurs prises par $\bar{\omega}_1/\bar{\omega}_2$ sont supérieures ou égales à 2 Log (2 ch $\bar{\omega}_2/2$), avec $E(\bar{\omega}_1/\bar{\omega}_2) > 2$ Log (2ch $\bar{\omega}_2/2$) puisque $\bar{\Psi}(A,\hat{\theta}_1) \in C_2^o$, figure 5.

On a de nouveau les deux premiers moments (approchés) de la loi conditionnelle cherchée, et on est ramené à un problème d'ajustement de loi scalaire, connaissant les deux premiers moments. Il paraît judicieux dans ce cas de prendre la loi log-normale calée sur 2 Log (2 ch $\bar{\omega}_2/2$) de moyenne $E(\bar{\omega}_1/\bar{\omega}_2)$-2 Log (2 ch $\bar{\omega}_2/2$) et de variance var $(\bar{\omega}_1/\bar{\omega}_2)$.

Pour tout α compris entre 0 et 1 on en déduit les régions de rejet approchées des tests d'hypothèse UMPB sur une valeur de A^o, de niveau α. Sur le plan numérique cette solution dépend de la résolution en θ_1 de l'équation

$$\Psi(A\theta_1) - \Psi(A\theta_2) = \bar{\omega}_2.$$

IV.5 - Tests sur θ_1.

5.1 - On reprend les paramètres A et θ_1. Les nouvelles variables

$$\begin{pmatrix} \bar{s}_1 \\ \bar{s}_2 \end{pmatrix} = \mathscr{U}_\theta \begin{pmatrix} \bar{\tau}_1 \\ \bar{\tau}_2 \end{pmatrix}, \quad \mathscr{U}_\theta = \begin{pmatrix} \theta_1 & \theta_2 \\ -\theta_2 & \theta_1 \end{pmatrix}$$

sont commodes pour conditionner par rapport à $\bar{s}_1$, (changement de variables linéaire). La loi de remplacement de ces variables est ici la loi normale bivariée image $\mathscr{N}(\mathscr{U}_\theta\ \bar{\Psi}(a), \mathscr{U}_\theta\ \Omega(a)\mathscr{U}_\theta' /P)$. D'après

l'expression de $\mathcal{L}_\theta \bar{\Psi}(a)$ il est indiqué de prendre comme nouvelle statistique de conditionnement la solution $\hat{A}(\theta_1, \bar{s}_1)$ de l'équation $\Psi(A)-\theta_1 \Psi(A\theta_1)-\theta_2 \Psi(A\theta_2) = \bar{s}_1$. On vérifie que pour tout $\bar{s}_1$ supérieur ou égal à $s_1^* = -\theta_1 \text{ Log } \theta_1 - \theta_2 \text{ Log } \theta_2$ au deuxième membre, cette équation possède une racine réelle unique $\hat{A}(\theta_1, \bar{s}_1)$ croissante en $\bar{s}_1$ [15]. Quand on conditionne la loi normale ci-dessus par rapport à $\hat{A}(\theta_1, \bar{s}_1)$, donc $\bar{s}_1$, la loi conditionnelle approchée de $\bar{s}_2$ par rapport à $\bar{s}_1$ à pour moyenne $E(\bar{s}_2/\bar{s}_1) = (\theta_1-\theta_2)\Psi(\hat{A})-\theta_1\Psi(\hat{A}\theta_2)+\theta_2\Psi(\hat{A}\theta_1)$ et pour variance (+)

$$\text{var}(\bar{s}_2/\bar{s}_1) = [\theta_2^2\Psi'(\hat{A}\theta_1)+\theta_1^2\Psi'(\hat{A}\theta_2)-(\theta_1-\theta_2)^2\Psi'(\hat{A})$$

$$- \frac{(\theta_1\theta_2(\Psi_2'-\Psi_1')+(\theta_2-\theta_1)\Psi'(\hat{A}))^2}{\theta_1^2\Psi_1' + \theta_2^2\Psi_2' - \Psi'(\hat{A})}] \frac{1}{p}$$

(elle ne dépend plus de A). On dispose à nouveau des deux premiers moments de la loi conditionnelle approchée. La statistique de conditionnement $\bar{s}_1$ est nécessairement supérieure ou égale à s_1^*, et la statistique conditionnée $\bar{s}_2/\bar{s}_1$ prend ses valeurs sur un intervalle borné $[s_2' \; s_2'']$, figure 6.

5.2 - Par rapport aux deux exemples précédents cet intervalle de variation n'est pas simple à déterminer ; quand à la loi d'ajustement on peut penser à une loi Béta, à condition que l'on ait (condition nécessaire et suffisante pour un ajustement Béta à partir des deux premiers moments) :

$$[s_2''-E(\bar{s}_2/\bar{s}_1)] \; [E(\bar{s}_2/\bar{s}_1)-s_2'] < \text{var } (\bar{s}_2/\bar{s}_1).$$

On est amené à introduire une approximation supplémentaire à cause de l'intervalle de variation de $\bar{s}_2/\bar{s}_1$; pour cela on remplace la variété V_2 dans R^2 par l'hyperbole d'équation $\tau_1\tau_2 = s_1^{*2}/4\theta_1\theta_2$ qui a un point

---

(+) On a posé $\Psi_1' = \Psi'(\hat{A}\theta_1)$; $\Psi_2' = \Psi'(\hat{A}\theta_2)$

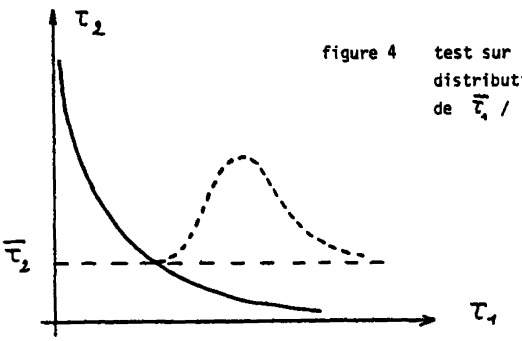

figure 4 test sur a_1 , (IV-3) distribution conditionnelle de $\overline{\tau_1}$ / $\overline{\tau_2}$.

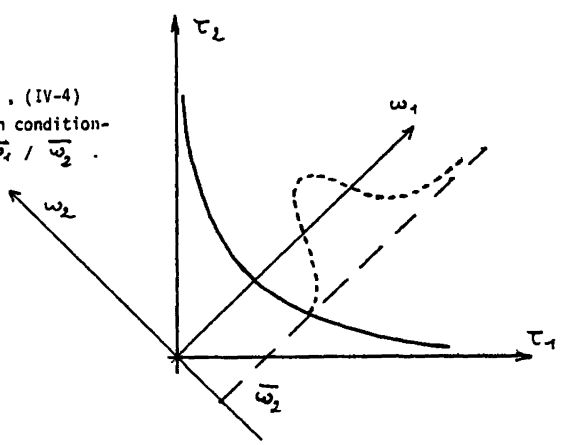

figure 5 test sur A , (IV-4) distribution conditionnelle de $\overline{\omega_1}$ / $\overline{\omega_2}$.

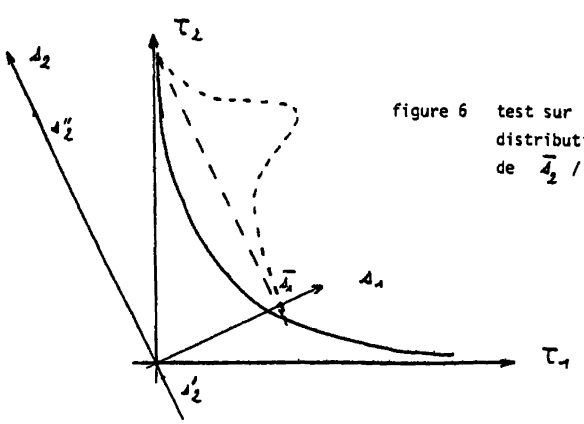

figure 6 test sur θ_1 , (IV-5) distribution conditionnelle de $\overline{\Delta_2}$ / $\overline{\Delta_1}$

double avec V_2 au point de coordonnées

$$\begin{cases} \tau_1^* = s_1^*/2\theta_1 \\ \tau_2^* = s_1^*/2\theta_2 \end{cases} , \quad \begin{cases} s_1^* \\ s_2^* = \frac{\theta_1-\theta_2}{2\theta_1\theta_2}\ s_1^* \end{cases}$$

Les bornes du nouvel intervalle de variation sont à présent les deux racines réelles d'une équation du second degré, elles sont égales à :

$$\tilde{s}_2' \ , \ \tilde{s}_2'' = \frac{\theta_1-\theta_2}{2\theta_1\theta_2}\ \bar{s}_1 \ \pm \ \frac{\theta_1^2+\theta_2^2}{2\theta_1\theta_2}\ \sqrt{\bar{s}_1^2 - s_1^{*2}} \ .$$

Sur ce nouvel intervalle on peut faire un ajustement Béta si la condition suivante est vérifiée

$$(E(\bar{s}_2/\bar{s}_1) - \frac{\theta_1-\theta_2}{2\theta_1\theta_2}\ \bar{s}_1)^2 - (\frac{\theta_1^2+\theta_2^2}{2\theta_1\theta_2})^2\ (\bar{s}_1^2 - s_1^{*2}) + \mathrm{var}(\bar{s}_2/\bar{s}_1) < 0 \ ;$$

quand aux régions de rejet des tests d'hypothèse sur θ_1, on peut par exemple les déterminer à l'aide de tables de la loi Béta [18]. Cette troisième application est donc un peu plus complexe que les précédentes, mais ce sont des questions de "fin de parcours", l'application elle-même repose sur la résolution numérique de l'équation

$$\Psi(A)-\theta_1\Psi(A\theta_1) - \theta_2\Psi(A\theta_2) = \bar{s}_1.$$

5.3 - On peut également, faut de mieux, conserver l'approximation normale de $\bar{s}_2/\bar{s}_1$.

IV.6 - Exemple

A titre d'application formelle de ces développements, on considère le partage du corps électoral français au cours du deuxième tour des élections présidentielles de 1981. Il s'agit d'un partage en trois

parties, on prend les indices i=2 et 3 pour MM. les candidats Mitterand et Giscard d'Estaing et i=1 pour les suffrages non exprimés. Comme échantillon on prend les P=24 partages des résultats de chaque région, de Paris intra-muros, et de l'ensemble DOM-TOM. Si l'on considère que ces partages indépendants suivent la même loi de Dirichlet sur le simplexe S_1^3, les estimateurs du maximum de vraisemblance ont les valeurs suivantes [15] :

$$\hat{a}_1 = 11,01 \qquad \hat{\theta}_1 = 0,173$$
$$\hat{a}_2 = 26,64 \qquad \hat{\theta}_2 = 0,419$$
$$\hat{a}_3 = 25,95 \qquad \hat{\theta}_3 = 0,408$$
$$\hat{A} = 63,6$$

a) On envisage des tests sur a_1 ignorant les valeurs a_2 et a_3. On se limite aux statistiques $\bar{\tau}_1$ et $\bar{\tau}_2$ qui correspondent aux voix non exprimées et exprimées. On sait que $\bar{\tau}_2$ est exhaustive pour $A_2 = a_2+a_3$; les données de l'échantillon donnent

$$\bar{\tau}_1 = 1,78997 \qquad \bar{\tau}_2 = 0,19194, \quad (\frac{\bar{\tau}_1}{\bar{\tau}_2}) \in C_2^o,$$

Pour chaque hypothèse a_1^o, on doit rechercher la valeur $\hat{A}_2(a_1^o,\bar{\tau}_2)$; à titre d'exemple on a le tableau de calcul suivant, avec les régions de rejet de tests bilatéraux $a_1^1 \neq a_1^o$ issues de l'approximation log-normale calée sur $-\log(1-e^{-\tau 2})$ de la loi de $\bar{\tau}_1/\bar{\tau}_2$, pour le niveau .05.

a_1^o	6	7	10	15	18	19
$\hat{A}_2$	28,85228	33,57850	47,75666	71,38633	85,563999	90,28986
$E(\bar{\tau}_1/\bar{\tau}_2)$	1,830586	1,81808	1,795804	1,778684	1,773017	1,771529
$\sigma^2(\bar{\tau}_1/\bar{\tau}_2)$	$6,1008\ 10^{-4}$	$4,4496\ 10^{-4}$	$2,1513\ 10^{-4}$	$9,461\ 10^{-5}$	$6,5468\ 10^{-5}$	$5,8703\ 10^{-5}$
intervalle de confiance 0,95 (approx. log-nor	1,79223 1,88915	1,785329 1,86724	1,77303 1,82999	1,763582 1,801353	1,760455 1,79187	1,759334 1,789386

On peut donc accepter les hypothèses $a_1^0 = 7, 10, 15, 18$.

b) On envisage des tests sur A ignorant θ_1, les statistiques de l'échantillon sont égales à

$$\bar{\omega}_1 = 1{,}98191, \quad \bar{\omega}_2 = -1{,}59803$$

Pour chaque hypothèse A^0 on doit rechercher la valeur $\hat{\theta}_1(A^0, \bar{\omega}_2)$.
Par exemple on a le tableau de calcul suivant, avec les régions de rejet de tests bilatéraux $A_1^1 \neq A_1^0$ issues de l'approximation log-normale calée sur 2 Log 2 ch($\bar{\omega}_2/2$) de la loi de $\bar{\omega}_1/\bar{\omega}_2$, pour le niveau .05.

A^0	30	40	70	110
$\hat{\theta}_1$	0,179089	0,17642	0,17293	0,17126
$\bar{E}(\bar{\omega}_1/\bar{\omega}_2)$	0,03358	0,02514	0,01431	0,09110
$\sigma^2_{\bar{\omega}_1/\bar{\omega}_2}$	$9{,}39\ 10^{-5}$	$5{,}26\ 10^{-5}$	$1{,}71\ 10^{-5}$	$6{,}92\ 10^{-6}$
intervalle à 0,95	1,985	1,980	1,973	1,971
(log-nor)	2,023	2,008	1,989	1,9817

On peut donc accepter les hypothèses $A^0 = 40, 70$.

c) On envisage des tests sur θ_1 ignorant A. Pour chaque hypothèse θ_1^0 on doit rechercher la valeur $\hat{A}(\theta_1^0, \bar{s}_1)$. Dans ce cas les statistiques $\bar{s}_1$, $\bar{s}_2$ de l'échantillon dépendent de θ_1^0. On a par exemple, le tableau de calcul suivant, avec les régions de rejet de tests bilatéraux $\theta_1^1 \neq \theta_1^0$ issues de l'approximation normale de la loi de $\bar{s}_2/\bar{s}_1$, pour le niveau .05.

θ_1^o	0,15	0,17	0,20
$\hat{A}$	57,048	65,740	45,710
$E(\bar{s}_2/\bar{s}_1)$	-1,6311	-1,4701	-1,2783
$\sigma^2_{\bar{s}_2/\bar{s}_1}$	$3,24\ 10^{-3}$	$2,35\ 10^{-3}$	$2,69\ 10^{-3}$
$\bar{s}_2$	-1,493	-1,453	-1,393
intervalle	-1,742	-1,565	-1,380
à 0,95 (normal)	-1,520	-1,375	-1,176

On peut donc accepter l'hypothèse $\theta_1^o = 0.17$.

IV.7 - Conclusions

Il est immédiat d'en faire autant avec des tests unilatéraux.

La distribution de Dirichlet est une distribution qui s'applique aux partages entre un nombre connu M d'acteurs, $M \geq 2$. Il y a autant de paramètres scalaires qu'il y a de participants. On a montré ci-dessus comment on peut facilement faire un test d'hypothèse sur l'un d'eux au cours d'une procédure en deux étapes, en présence de paramètres fantômes ou non :

- une étape où l'on met en oeuvre des propriétés de la distribution de Dirichlet ;
- une étape où l'on utilise des propriétés communes aux distributions de structures exponentielles complètes d'ordre 2. Les transformées de Laplace et/ou les fonctions génératrices de moments quand il n'y a pas de paramètres fantômes ; le paramétrage mixte, l'approximation normale de la distribution des statistiques exhaustives et au besoin un changement de variables linéaires sont les outils analytiques mis en oeuvre. Les questions d'ordre numérique se réduisent à peu de choses quand on dispose d'un minimum d'informatique, (les applications ont été faites avec un micro de 4 K mots).

Techniquement on doit terminer dans tous les cas par une question d'ajustement. On peut la résoudre avec des méthodes classiques plus ou moins sophistiquées, disposant dans le cas des tests conditionnels des moments d'ordre 1 et 2, et du champ de variation de la statistique conditionnelle.
L'édition de tables statistiques pour la mise en oeuvre des tests est inutile quand on se ramène, comme ici, à un ajustement normal ou log-normal.

BIBLIOGRAPHIE

[1] ANDERSON T.W.
An introduction to multivariate analysis.J. Wiley 1958.

[2] BARNDORFF-NIELSEN O. -
Information and exponential families. J. Wiley, 1977.

[3] BARRA R.
Notions fondamentales de statistique mathématique. Dunod 1971

[4] BUCHWALTER H.
Polynômes orthogonaux et fonction hypergéométrique, Lyon 1.

[5] COX D.R., HINCKLEY D.V.
Theoretical statistics. Chapmann and Hall, 1974.

[6] CRAMER H.
Mathematical methods of Statistics. Princeton Un. Press, 1946

[7] DANIELS H.E.
Saddlepoint approximations in statistics
Ann. of Math. Stat. Vol. 25, pages 631-650, 1954.

[8] DEGERINE S.
Etude des structures statistiques associées aux lois de Von-Mises, Thèse de 3ème Cycle, USMGrenoble, 1975.

[9] GOOD I.J.
The estimation of probabilities. An essay of modern Bayesian methods. Research monograph n° 30, MIT Press, 1965.

[10] HLADIK J.
La transformation de LAPLACE à plusieurs variables, Masson, 1969.

[11] JOHNSON N.L.
An approximation to the multinomial distribution, some properties and applications, Biometrika Vol. 47, pages 93-102, 1960.

[12] KAUFMANN A.
Cours moderne de calcul de probabilités. A. Michel, 1965.

[13] KENDALL M.G., STUART A.
The advanced theory of statistics, Tome 1, C. Griffin 1963.

[14] LEHMANN E.L.
Testing statistical hypothesis, J. Wiley, 1959.

[15] MAURIN M.
Contribution à l'étude des partages, probabilités et statistiques sur le simplexe unité, Thèse de Docteur-Ingénieur, USMG, 1983.

[16] MAURIN M., CHALONS J.M., BEGUE D., ISARD M.
Les niveaux de bruit élevés produits par la circulation automobile, 11ème congrès international d'Acoustique, Paris, 1983, IRT-CERNE 1983

[17] MONTFORT A.
Cours de statistique mathématique. Economica, 1980.

[18] PEARSON K.S.
Tables of incomplete Beta function, Cambridge Un. Press 1956.

[19] ULMO J., BERNIER J.
Eléments de décision statistique, PUF, 1973.

[20] WILKS S.S.
Mathematical statistics, John Wiley 1963.