ORESTE NASI

Tests de rang sur les variances, asymptotiquement les plus puissants maximin

Statistique et analyse des données, tome 3, nº 1 (1978), p. 13-28
<http://www.numdam.org/item?id=SAD_1978__3_1_13_0>

Statistique et Analyse des Données
1 - 1978

TESTS DE RANG SUR LES VARIANCES, ASYMPTOTIQUEMENT LES PLUS PUISSANTS MAXIMIN.

NASI Oreste

(Laboratoire de Statistique, Université de Nancy I)

0 - INTRODUCTION

Cet article contient l'application aux doubles-échantillons de lois continues des résultats que nous donnions dans [4] et qui étaient eux-mêmes une extension de résultats de HAJEK et SIDAK ([1]).

Pus précisément nous donnons des tests de rang pour tester l'hypothèse d'égalité des lois de deux variables contre l'hypothèse de variances différentes, à moyennes égales.

Si on considère une variable X de densité et variance finies, ces tests sont asymptotiquement uniformément les plus puissants maximin pour l'alternative multiple où le deuxième échantillon est issu de la variable $Y = \sigma X + \beta$ et le premier de la variable $Z = e^{\Delta_n}(Y - E(Y)) + E(Y)$, quels que soient σ strictement positif et β réel.

Enfin les efficacités relatives sont calculées, ce qui permet de constater que plusieurs de ces tests ont des propriétés asymptotiques voisines.

Nous ne donnons pas la démonstration détaillée de tous les résultats, en particulier de ceux du chapitre 1 qui sont des applications directes aux doubles-échantillons des résultats obtenus dans [4] . Notre souci est de présenter, autant que possible, la construction des tests de manière applicable pour les utilisateurs.

1 - DONNEES ET RESULTATS DE BASE

1.1 - Données

On considère la suite de doubles-échantillons :

$$X^n = (X_{n,1}, X_{n,2}, \dots, X_{n,m_n}, X_{n,m_n+1}, \dots, X_{n,m_n+m'_n}), \quad n \in \mathbb{N}^*$$

constitués de variables mutuellement indépendantes. Nous supposerons que, pour tout n, les variables $X_{n,1}, \dots, X_{n,m_n}$ sont de même fonction de répartition F_n continue et les variables $X_{n,m_n+1}, \dots, X_{n,m_n+m'_n}$ de même fonction de répartition G_n continue ; nous nous intéressons à la suite de couples d'hypothèses :

$$(1.1) \qquad \begin{aligned} &H_n^o : F_n \equiv G_n \\ &H_n^1 : F_n \text{ est de densité } f(x, \Delta_n) \text{ et } G_n \text{ de densité } f(x,o), \end{aligned}$$

où $f(x, \theta)$ est une famille de densités ainsi définie :

$\theta \in I$, I intervalle ouvert de $\mathbb{R}$ contenant l'origine,

$\forall \theta \in I$, $f(x, \theta)$ est une densité par rapport à la mesure de Lebesgue sur $\mathbb{R}$, et $\{\Delta_n\}$ est une suite d'éléments de I, non nuls et de signe constant.

Désignons, quand elles existent, par $f'(x, \theta)$ ou $f'_\theta(x)$ la dérivée partielle de f en θ et par $I_f(\theta)$ la quantité d'information de la densité $f(x, \theta)$:

$$f'(x, \theta) = f'_\theta(x) = \lim_{u \to 0} \frac{f(x, \theta+u) - f(x, \theta)}{u} \qquad \theta \in I, \quad x \in \mathbb{R}$$

$$(1.2) \qquad I_f(\theta) = \int_{\mathbb{R}} \left(\frac{f'(x, \theta)}{f(x, \theta)}\right)^2 f(x, \theta)\, dx, \qquad \theta \in I.$$

Enfin rappelons la définition de la fonction Ψ_f de Hajek et Sidak relative à la famille $\{f(x, \theta)\}$:

$$(1.3) \qquad \Psi_f(u, \theta) = \frac{f'_\theta(F_\theta^{-1}(u))}{f_\theta(F_\theta^{-1}(u))} \qquad u \in]0,1[, \quad \theta \in I,$$

où $F_\theta(x)$, que nous noterons aussi $F(x, \theta)$, est la fonction de répartition de $f(x, \theta)$ et $F_\theta^{-1}(u)$, notée aussi $F^{-1}(u, \theta)$ est la fonction "inverse" de $F_\theta(x)$:

$$F_\theta^{-1}(u) = F^{-1}(u, \theta) = \inf \{x / F(x, \theta) \geqslant u\}, \quad u \in]0,1[.$$

Dans la suite nous négligerons certains indices n pour écrire par exemple m et m' au lieu de m_n et m'_n ou encore $X^n = (X_1, \dots, X_{m+m'})$.

1.2 - Résultats de base

Soit φ une application de $]0,1[$ dans $\mathbb{R}$ de carré intégrable et de moyenne $\bar{\varphi} = \int_0^1 \varphi(u)\,du$, et soit la suite de statistiques de rang des doubles-échantillons X^n :

$$(1.4)\qquad S_n^{\varphi} = \left\{\frac{m\,m'}{m+m'} \int_0^1 (\varphi(u) - \bar{\varphi})^2 du\right\}^{-1/2} \left(\sum_{i=1}^{m} \varphi\left(\frac{R_{ni}}{m+m'+1}\right) - \frac{m}{m+m'} \sum_{i=1}^{m+m'} \varphi\left(\frac{i}{m+m'+1}\right)\right)$$

où $R^n = (R_{n,1}, \dots, R_{n,m+m'})$ désigne le vecteur rang de Hajek de l'échantillon global X^n , défini dans [4] p. 18.

Dans la suite la famille $f(x, \theta)$ sera astreinte aux conditions suivantes :

(C1) : Pour presque tout x , $f(x, \theta)$ est absolument continue en θ sur I , c'est-à-dire il existe une application $h(x, \theta)$ telle que, pour presque tout x :

$$\forall\ a \text{ et } b \in I \ / \ -\infty < a \leq b < +\infty\ , \int_a^b h(x, \theta)\,d\theta = f(x, b) - f(x, a) \ .$$

On sait que dans ce cas :

$$h(x, \theta) = f'(x, \theta) \text{ pour presque tout } \theta \text{ de } I \ .$$

Nous confondrons, dans la suite, ces deux fonctions $h(x, \theta)$ et $f'(x, \theta)$.

(C2) : Pour presque tout x il existe un voisinage de 0 où $f'(x, \theta)$ est continue en θ .

(C3) : $I_f(\theta)$ est une fonction finie, non nulle et continue en θ sur un voisinage de 0 .

(C4) : $\int_{\mathbb{R}} |f'(x, \theta)|dx$ est pour presque tout x , continue en θ sur un voisinage de 0 .

(C5) : Pour toutes suites de réels v_n et t_n telles que :

$$(1.5)\qquad \lim_{n\to+\infty} v_n = +\infty \quad \text{et} \quad \lim_{n\to+\infty} t_n = 0\ ,$$

on a :

$$\lim_{n\to+\infty} \int_{\{x/f(x,\, t_n) > 0 \text{ et } \left|\frac{f'(x,\, t_n)}{f(x,\, t_n)}\right| > v_n\}} \left(\frac{f'(x,\, t_n)}{f(x,\, t_n)}\right)^2 dx = 0$$

Ces conditions diffèrent, en particulier (C1) et (C2) , des conditions que nous donnions dans [4] . Elles sont cependant plus adaptées aux densités que nous allons étudier et on vérifiera sans peine que les résultats de [4] restent vrais sous ces nouvelles conditions.

Résultat 1 : si

$$(a)\qquad \lim_{n\to+\infty} \min(m_n, m'_n) = +\infty$$

$$(b)\qquad \lim_{n\to+\infty} \frac{m_n m'_n}{m_n + m'_n} \Delta_n^2 = \mathcal{C}\ , \quad 0 < \mathcal{C} < +\infty$$

(c) φ est la somme non constante d'un nombre fini de fonctions monotones de carrés intégrables,

(d) la famille $\{f(x, \theta)\}$ satisfait aux conditions (C1) à (C5),

(e) $$\lim_{\theta \to 0} \int_0^1 \varphi(u)\, \Psi_f(u, \theta)\, du = \int_0^1 \varphi(u)\, \Psi_f(u, 0)\, du \quad ,$$

alors sous les hypothèses alternatives H_n^1 , la suite de statistiques S_n^φ définie par (1.4) converge en loi vers la loi normale $\mathcal{N}(\mu_1, 1)$ de moyenne μ_1 et d'écart-type 1 avec

$$\mu_1 = \sqrt{\mathcal{E}}\ \frac{\int_0^1 \varphi(u)\, \Psi_f(u, 0)\, du}{\left(\int_0^1 (\varphi(u) - \bar{\varphi})^2 du\right)^{\frac{1}{2}}}$$

<u>Démonstration</u> : Il s'agit de l'application du théorème du paragraphe 3 de [4] au cas où :

$$c_{ni} = t_{ni} = \begin{cases} \Delta_n & \text{si } 1 \leqslant i \leqslant m \\ 0 & \text{si } m+1 \leqslant i \leqslant m+m' \end{cases}$$

et $a_n(i) = \varphi(\frac{i}{m+m'+1})$.

Ce résultat est vrai pour les $\Delta_n > 0$. S'ils sont strictement négatifs il suffit de remplacer S_n^φ par $-S_n^\varphi$.

Notons maintenant :

$$\Psi_f^*(u) = \Psi_f(u, 0) \qquad \forall\, u \in \,]0,1[\quad , \text{et}$$

(1.6) $$S_n^*(f) = (\frac{m\,m'}{m+m'}\ I_f(0))^{-1/2} \left(\sum_{i=1}^{m} \Psi_f^*(\frac{R_{ni}}{m+m'+1}) - \frac{m}{m+m'} \sum_{i=1}^{m+m'} \Psi_f^*(\frac{i}{m+m'+1})\right) ,$$

et considérons les tests de H_n^o contre H_n^1 de fonctions de test :

(1.7) $$\begin{cases} \Phi_n = 1 & \text{si} \quad S_n^*(f) \geqslant k_{1-\alpha} \\ = 0 & \text{si} \quad S_n^*(f) < k_{1-\alpha} \end{cases}$$

où k_y est le quantile d'ordre y de la fonction de répartition Φ de $\mathcal{N}(0,1)$:

$$\Phi(k_y) = y \qquad \forall\, y \in \,]0,1[\quad .$$

(Autrement dit on rejette l'hypothèse H_n^o en faveur de H_n^1 si $S_n^*(f) \geqslant k_{1-\alpha}$ et on l'accepte dans le cas contraire).

On a le résultat suivant :

<u>Résultat 2</u> : si on a (a) et (b) et :

(c) Ψ_f^* est la somme non constante d'un nombre fini de fonctions monotones de carrés intégrables,

(d) la famille $\{f(x, \theta)\}$ satisfait aux conditions (C1) à (C5)

(e) $\lim_{\theta \to 0} \int_0^1 \Psi_f^*(u)\, \Psi_f(u, \theta)\, du = \int_0^1 (\Psi_f^*(u))^2\, du$,

alors la suite de fonctions de tests (1.7) détermine un test asymptotiquement uniformément le plus puissant de seuil α de H_n^o contre H_n^1 , de puissance asymptotique égale à

$1 - \Phi(k_{1-\alpha} - \sqrt{\mathcal{C} I_f(0)})$.

Démonstration : On applique le théorème 1 du paragraphe 6 de [4] .

Rappelons que si l'on note, pour tout n , $H_n = \{P_n\}$ l'ensemble des lois de probabilité des échantillons X^n qui vérifient H_n^o et Q_n la loi de probabilité vérifiant H_n^1 , la conclusion du résultat signifie que

$$\lim_{n \to +\infty} \left(\sup_{P_n \in H_n} \int \phi_n\, dP_n \right) = \alpha \quad \text{et}$$

$$\lim_{n \to +\infty} \left(\beta(\alpha, H_n^o, Q_n) - \int \phi_n\, dQ_n \right) = 0$$

où $\beta(\alpha, H_n^o, Q_n)$ est la puissance d'un test uniformément le plus puissant de seuil α de H_n^o contre H_n^1 . Par ailleurs la puissance asymptotique est égale à :

$$\lim_{n \to +\infty} \int \phi_n\, dQ_n \ .$$

Remarque : ici aussi le résultat vaut pour les $\Delta_n > 0$. Sinon on remplace $S_n^*(f)$ par $-S_n^*(f)$. On notera que ces tests de rang sont asymptotiquement optimaux parmi tous les tests, y compris ceux qui ne sont pas de rang.

2 - UNE FAMILLE $\{f(x, \theta)\}$ PARTICULIERE

Soit, pour un réel a et une densité g sur $\mathbb{R}$, la famille de densités :

(2.1) $f(x, \theta) = e^{-\theta} g((x-a)e^{-\theta} + a) \quad \forall\, \theta \in \mathbb{R} \ , \ \forall\, x \in \mathbb{R}$.

Remarque 1 : si X désigne une variable de densité g , il est aisé de voir que $f(x, \theta)$ est la densité de la variable $e^{\theta}(X-a) + a$.
La condition (C1) sur la famille $f(x, \theta)$ implique que g doit être absolument continue et donc presque partout dérivable. Notons g' la dérivée de g . On a :

(2.2) $f'(x, \theta) = -e^{-\theta} g((x-a)e^{-\theta} + a) - (x-a)e^{-2\theta} g'((x-a)e^{-\theta} + a)$.

Il résulte des propriétés des fonctions absolument continues que si g est continue sur $\mathbb{R}$ et admet une dérivée continue sauf en un nombre fini de points où il existe une dérivée à droite et une dérivée à gauche finies, alors (C1) et (C2) sont vérifiées pour la famille (2.1).

Par ailleurs on a, d'après (2.2) :

$$\frac{f'(x,\theta)}{f(x,\theta)} = -1 - (x-a)\, e^{-\theta}\, \frac{g'((x-a)e^{-\theta}+a)}{g((x-a)e^{-\theta}+a)} ,$$

d'où d'après (1.2) :

$$I_f(\theta) = \int_{\mathbb{R}} \left(1 + (x-a)e^{-\theta}\, \frac{g'((x-a)e^{-\theta}+a)}{g((x-a)e^{-\theta}+a)}\right)^2 e^{-\theta}\, g((x-a)e^{-\theta}+a)\, dx$$

$$(2.3) \qquad = \int_{\mathbb{R}} \left(1 + (x-a)\, \frac{g'(x)}{g(x)}\right)^2 g(x)\, dx \; .$$

On constate que $I_f(\theta)$ ne dépend pas de θ . La condition (C3) sera donc vérifiée si l'intégrale (2.3) est finie et non nulle.

De plus :

$$\int_{\mathbb{R}} |f'(x,\theta)|\, dx = \int_{\mathbb{R}} |g(x) + (x-a)\, g'(x)|\, dx$$

$$\leqslant 1 + \int_{\mathbb{R}} |(x-a)\, g'(x)| dx \; .$$

Il suffit donc que cette dernière intégrale soit finie pour que la condition (C4) soit satisfaite.

Terminons l'examen des conditions sur $\{f(x,\theta)\}$: d'après la proposition V.3 de [3] la condition (C5) est vérifiée si pour toutes suites $\{v_n\}$ et $\{t_n\}$ qui vérifient (1.5) , il existe une suite $\{v'_n\}$ tendant vers $+\infty$ et un entier N tel que :

$$\forall\, n \geqslant N ,\ \forall\, x/g(xe^{-t_n}+a) \neq 0 \ \text{ et } \ \left|x\, \frac{g'(x\, e^{-t_n}+a)}{g(x\, e^{-t_n}+a)}\right| > v_n \Rightarrow |x| > v'_n \; .$$

Déterminons maintenant la fonction de Hajek et Sidak de la famille (2.1).
Notons G la fonction de répartition de g et G^{-1} la fonction "inverse" de G . D'après (1.3) on a :

$$F(x,\theta) = G((x-a)e^{-\theta}+a) \qquad \forall\, x \in \mathbb{R} , \quad \forall\, \theta \in \mathbb{R} ,$$

$$F^{-1}(u,\theta) = e^{\theta}(G^{-1}(u) - a) + a \qquad \forall\, u \in\,]0,1[\;, \quad \forall\, \theta \in \mathbb{R} ,$$

et :

$$(2.4) \qquad \Psi_f(u,\theta) = -1 - (G^{-1}(u) - a)\, \frac{g'(G^{-1}(u))}{g(G^{-1}(u))} \qquad \forall\, u \in\,]0,1[\;, \quad \forall\, \theta \in \mathbb{R} \; .$$

On remarque que Ψ_f ne dépend pas de θ .

Nous allons donner deux lemmes qui montrent l'intérêt de la famille de densités (2.1).

<u>Notation</u> : l'hypothèse simple H^1_n définie en (1.1) sera notée :

$$\{(f(x,\Delta_n) \; ; \; f(x,0))\} \; .$$

Cette hypothèse peut donc être notée, pour $f(x, \theta)$ donnée par (2.1) :

(2.5) $\{(e^{-\Delta_n} g((x-a)e^{-\Delta_n} + a) \; ; \; g(x))\}$.

Lemme 1 : si la loi de densité g est de moyenne finie égale à a et de variance finie, et si l'échantillon X^n vérifie l'hypothèse (2.5) , alors :

$$E(X_{n,1}) = E(X_{n,m+1}) = a$$

$$\text{et} \quad Var(X_{n,1}) = e^{2\Delta_n} Var(X_{n,m+1}) \quad .$$

Ce lemme est facile à vérifier à partir de la remarque 1. Il montre qu'en testant H_n^o contre H_n^1 avec la famille $\{f(x, \theta)\}$ définie par (2.1) on fait un test d'égalité de deux lois contre une différence en variance, mais à moyennes égales.

Posons, pour tout σ strictement positif et tout β réel :

$$g_{\sigma,\beta}(x) = \frac{1}{\sigma} \; g\left(\frac{x-\beta}{\sigma}\right) \quad \forall x \in \mathbb{R} \quad .$$

Remarque 2 : si X désigne une variable de densité g , $g_{\sigma,\beta}$ est la densité de la variable $\sigma X + \beta$.

Lemme 2 : soit R^n le vecteur rang du double échantillon X^n . La loi de R^n sous l'hypothèse :

$$\{(e^{-\Delta_n} g_{\sigma,\beta}((x - a\sigma - \beta) \; e^{-\Delta_n} + a\sigma + \beta) \; ; \; g_{\sigma,\beta}(x))\}$$

ne dépend pas de σ et β .

Démonstration :

Soit X une variable de densité $g(x)$ et $Y^n = (Y_1, \ldots, Y_{m+m'})$ un échantillon satisfaisant à l'hypothèse (2.5). D'après la remarque 1 , $Y_1, Y_2, \ldots, Y_m$ ont même densité que la variable $e^{\Delta_n}(X - a) + a$ et $Y_{m+1}, \ldots, Y_{m+m'}$ ont même densité que X .

Pour σ strictement positif et β réel quelconque, considérons les variances :

$$Z_i = \sigma Y_i + \beta \qquad 1 \leq i \leq m+m' \; .$$

$Z_1, \ldots, Z_m$ ont donc même densité que $\sigma e^{\Delta_n}(X - a) + a\sigma + \beta$

et $Z_{m+1}, \ldots, Z_{m+m'}$ ont même densité que $\sigma X + \beta$.

De la remarque 2 il résulte que $g_{\sigma,\beta}$ est la densité de $\sigma X + \beta$, et en utilisant de plus la remarque 1 , que :

$$e^{-\Delta_n} g_{\sigma,\beta}((x - a\sigma - \beta) \; e^{-\Delta_n} + a\sigma + \beta) \quad \text{est la densité de la variable}$$

$$e^{\Delta_n} (\sigma X + \beta - a\sigma - \beta) + a\sigma + \beta = \sigma \, e^{\Delta_n}(X - a) + a\sigma + \beta \; .$$

Par conséquent $Z^n = (Z_1, \ldots, Z_{m+m'})$ est de même loi que l'échantillon X^n considéré dans le lemme. Comme la fonction $x \to \sigma x + \beta$ est strictement croissante, le rang de Z^n est le même que celui de Y^n qui ne dépend pas de σ et β. Ce qui prouve le lemme.

Ce lemme va nous permettre, à partir du résultat 3 ci-dessous, de construire des tests asymptotiquement uniformément les plus puissants (en abrégé : a.u.p.p.) maximin de H_n^o contre une hypothèse H_n^1 multiple.

Notation : Soit $\mathcal{D} = \{(h, k)\}$ un ensemble de couples de densités. Nous dirons que le double-échantillon X^n satisfait à l'hypothèse multiple $\{(h(x) ; k(x)) , (h, k) \in \mathcal{D}\}$ si $X_1, X_2, \dots, X_m$ sont de même loi de densité h et $X_{m+1}, \dots, X_{m+m'}$ de même loi de densité k pour un couple (h, k) de $\mathcal{D}$. Ainsi nous dirons que X_n vérifie l'hypothèse :

$$(2.6) \qquad \{(e^{-\Delta_n} g_{\sigma,\beta}((x - a\sigma - \beta) e^{-\Delta_n} + a\sigma + \beta) ; g_{\sigma,\beta}(x)) , \quad \forall \sigma > 0 , \quad \forall \beta \in \mathbb{R}\} ,$$

si $X_1, \dots, X_m$ sont de densité $e^{-\Delta_n} g_{\sigma,\beta}((x - a\sigma - \beta) e^{-\Delta_n} + a\sigma + \beta)$ et $X_{m+1}, \dots, X_{m+m'}$ de densité $g_{\sigma,\beta}$ pour un couple (σ, β) de $\mathbb{R}^{+*} \times \mathbb{R}$.

D'après (2.3) et (2.4) on peut noter :

$$(2.7) \qquad I_g = I_f(\theta) = \int_{\mathbb{R}} \left(1 + (x-a) \frac{g'(x)}{g(x)}\right)^2 g(x)\, dx \qquad \forall \theta \in \mathbb{R} ,$$

$$(2.8) \qquad \Psi_g(u) = \Psi_f(u, \theta) = -1 - (G^{-1}(u) - a) \frac{g'(G^{-1}(u))}{g(G^{-1}(u))} \qquad \forall \theta \in \mathbb{R} , \quad \forall u \in]0,1[.$$

Soient les statistiques :

$$(2.9) \qquad S_n^*(g) = \left(\frac{m\, m'}{m+m'} I_g\right)^{-\frac{1}{2}} \left(\sum_{i=1}^{m} \Psi_g\left(\frac{R_{ni}}{m+m'+1}\right) - \frac{m}{m+m'} \sum_{i=1}^{m+m'} \Psi_g\left(\frac{i}{m+m'+1}\right)\right) ,$$

et les fonctions de test :

$$(2.10) \qquad \begin{cases} \phi_n = 1 & \text{si} \quad S_n^*(g) \geq k_{1-\alpha} \\ = 0 & \text{si} \quad S_n^*(g) < k_{1-\alpha} , \end{cases}$$

on a le résultat suivant :

Résultat 3 : si les hypothèses (a) et (b) du résultat 1 sont vérifiées et si :

(c) Ψ_g est la somme non constante d'un nombre fini de fonctions monotones de carrés intégrables,

(d) $\{f(x, \theta)\}$ donnée en (2.1) satisfait aux conditions (C1) à (C5) ,

alors la suite de tests de fonctions de tests (2.10) détermine un test a.u.p.p. maximin de seuil α des hypothèses H_n^o contre les hypothèses (2.6), de puissance asymptotique égale à $1 - \Phi(k_{1-\alpha} - \sqrt{\mathcal{E} I_g})$.

Démonstration :

Cela découle du résultat 2 du paragraphe 1 et du lemme 2 ci-dessus. On notera que la condition (e) du résultat 2 est vérifiée d'après (2.8) . Rappelons la définition d'un test a.u.p.p. maximin de seuil α de H_n^o contre H_n^1 , H_n^o et H_n^1 désignant des hypothèses multiples :

H_n^o : la loi de probabilité de X^n appartient à $H_n = \{P_n\}$,

H_n^1 : la loi de probabilité de X^n appartient à $K_n = \{Q_n\}$.

Soit $M(\alpha, H_n^o)$ l'ensemble des fonctions de tests ϕ_n' telles que :

$$\sup_{P_n \in H_n} \int \phi_n' \, dP_n = \alpha ,$$

et $\beta(\alpha, H_n^o , H_n^1)$ la puissance d'un test uniformément le plus puissant maximin de H_n^o contre H_n^1 de seuil α :

$$\beta(\alpha, H_n^o , H_n^1) = \sup_{\phi_n' \in M(\alpha, H_n^o)} \left(\inf_{Q_n \in K_n} \int \phi_n' \, dQ_n \right) .$$

On dit que les fonctions de test ϕ_n déterminent un test a.u.p.p. maximin de seuil α de H_n^o contre H_n^1 si :

$$(2.11) \qquad \lim_{n \to +\infty} \sup_{P_n \in H_n} \int \phi_n \, dP_n = \alpha$$

$$(2.12) \qquad \lim_{n \to +\infty} \left(\beta(\alpha, H_n^o , H_n^1) - \inf_{Q_n \in K_n} \int \phi_n \, dQ_n\right) = 0 .$$

Ici comme les ϕ_n sont des fonctions du vecteur de rang de X^n , (2.11) et (2.12) deviennent respectivement :

$$\lim_{n \to +\infty} \int \phi_n \, dP_n = \alpha \quad \forall P_n \in H_n$$

$$\lim_{n \to +\infty} \left(\beta(\alpha, H_n^o , H_n^1) - \int \phi_n \, dQ_n\right) = 0 \qquad \forall Q_n \in K_n ,$$

cette dernière égalité découlant de (2.12) en vertu du lemme 2. De même la puissance asymptotique s'écrit :

$$\lim_{n \to +\infty} \inf_{Q_n \in K_n} \int \phi_n \, dQ_n = \lim_{n \to +\infty} \int \phi_n \, dQ_n \qquad \forall Q_n \in K_n .$$

3 - TESTS SUR LA VARIANCE

3.1 - Construction des tests

Dans la suite nous supposerons que g est une densité de moyenne et de variance finies :

$$\bar{g} = \int_{\mathbb{R}} x\, g(x)\, dx < +\infty$$

$$v^2(g) = \int_{\mathbb{R}} (x - \bar{g})^2\, g(x)\, dx < +\infty \quad .$$

Soit la famille de densités :

(3.1) $\qquad D = \{g_{\sigma,\beta} \quad , \forall\, \sigma > 0 \quad \forall\, \beta \in \mathbb{R}\}$.

On voit facilement, à partir de la remarque 2, que la moyenne et la variance de $g_{\sigma,\beta}$ vérifient :

(3.2)
$$\bar{g}_{\sigma,\beta} = \sigma\, \bar{g} + \beta$$
$$v^2(g_{\sigma,\beta}) = \sigma^2\, v^2(g) \qquad \forall\, \sigma > 0 \ , \quad \forall\, \beta \in \mathbb{R} \ .$$

Nous allons étudier des tests des hypothèses H_n^o contre les hypothèses multiples

(3.3) $\qquad \{ (e^{-\Delta_n}\, g_{\sigma,\beta}((x - \bar{g}_{\sigma,\beta})\, e^{-\Delta_n} + \bar{g}_{\sigma,\beta}) \ ; \ g_{\sigma,\beta}(x)) \ , \quad \forall\, \sigma > 0 \ , \quad \forall\, \beta \in \mathbb{R} \}$.

Un double-échantillon X^n qui vérifie cette hypothèse multiple est tel que $X_{m+1}, \ldots, X_{m+m'}$ est un échantillon aléatoire d'une variable Y de densité appartenant à D , et $X_1, \ldots, X_m$ un échantillon aléatoire de la variable $Z = e^{\Delta_n}\,(Y - E(Y)) + E(Y)$ (d'après (3.2) et la remarque 1 du paragraphe 2). De plus on a $E(Z) = E(Y)$ et $\mathrm{Var}(Z) = e^{2\Delta_n}\, \mathrm{Var}(Y)$.

Posons, pour toute densité h de la famille D définie en (3.1) :

(3.4) $\qquad f_h(x, \theta) = e^{-\theta}\, h((x - \bar{h})\, e^{-\theta} + \bar{h}) \qquad \forall\, x \in \mathbb{R} \ , \quad \forall\, \theta \in \mathbb{R}$,

$\bar{h}$ désignant la moyenne de la densité h .

Appelons Ψ_h et I_h la fonction de Hajek et Sidak et la quantité d'information correspondantes. On a :

<u>Lemme 3</u> :

(3.5) $\qquad \Psi_h(u) = -1 - (H^{-1}(u) - \bar{h})\, \dfrac{h'(H^{-1}(u))}{h(H^{-1}(u))}$

(3.6) $\qquad I_h = \displaystyle\int_{\mathbb{R}} \left(1 + (x - \bar{h})\, \frac{h'(x)}{h(x)}\right)^2 h(x)\, dx$,

où H^{-1} désigne la fonction "inverse" de la fonction de répartition H de la densité h .

<u>Démonstration</u> :

On a : $\quad f_h(x, \theta) = e^{-\theta}\, g((x - a)\, e^{-\theta} + a)$

avec $g = h$ et $a = \bar{h}$. C'est donc un cas particulier de la famille de densités $\{f(x, \theta)\}$

définies par (2.1). Les relations (3.5) et (3.6) découlent immédiatement de (2.8) et (2.7). Soient encore les statistiques de rang des échantillons X^n :

(3.7) $$S_n(h) = \left(\frac{m\,m'}{m+m'} I_h\right)^{-1/2} \left(\sum_{i=1}^{m} \Psi_h\left(\frac{R_{ni}}{m+m'+1}\right) - \frac{m}{m+m'} \sum_{i=1}^{m+m'} \Psi_h\left(\frac{i}{m+m'+1}\right)\right) .$$

On a :

Lemme 4 : $\forall\, h \in D$,

(3.8) $$\Psi_h = \Psi_g$$

(3.9) $$I_h = I_g$$

(3.10) $$S_n(h) = S_n(g) .$$

Démonstration :

$$\forall\, h \in D \;,\quad \exists\, \sigma > 0 \text{ et } \beta \in \mathbb{R} \;/\; h = g_{\sigma,\beta} .$$

$$\Rightarrow H(x) = G\left(\frac{x-\beta}{\sigma}\right)$$

$$\Rightarrow H^{-1}(u) = \inf\{x \;/\; H(x) \geq u\} = \sigma\, G^{-1}(u) + \beta .$$

D'autre part, $\bar{h} = \sigma\,\bar{g} + \beta$ d'après (3.2) et

$$\frac{h'(x)}{h(x)} = \frac{1}{\sigma} \frac{g'\left(\frac{x-\beta}{\sigma}\right)}{g\left(\frac{x-\beta}{\sigma}\right)} .$$

D'où en remplaçant dans (3.5) ,

$$\Psi_h(u) = -1 - (G^{-1}(u) - \bar{g}) \frac{g'(G^{-1}(u))}{g(G^{-1}(u))}$$

$$= \Psi_g(u) \qquad \forall\, u \in \,]0,1[.$$

Par ailleurs (3.6) donne :

$$I_h = \int_{\mathbb{R}} \left(1 + (x - \sigma\,\bar{g} - \beta)\, \frac{1}{\sigma} \frac{g'\left(\frac{x-\beta}{\sigma}\right)}{g\left(\frac{x-\beta}{\sigma}\right)}\right)^2 \frac{1}{\sigma}\, g\left(\frac{x-\beta}{\sigma}\right) dx$$

$$= \int_{\mathbb{R}} \left(1 + (u - \bar{g}) \frac{g'(u)}{g(u)}\right)^2 g(u)\, du = I_g ,$$

ce qui prouve (3.9). L'égalité (3.10) découle immédiatement de (3.8) et (3.9).

Remarque : si g est telle que la famille $\{f_h(x, \theta)\}$ satisfait aux conditions (C1) , (C2) et (C3), l'égalité (3.9) sur les quantités d'informations se déduit de l'égalité (3.8) sur les fonctions de Hajek et Sidak à partir de la proposition 1 de [4] où il est établi que :

$$I_h = \int_0^1 \Psi_h^2(u)\, du \quad .$$

Le lemme 4 permet de construire à partir de n'importe quelle densité de la famille D un test a.u.p.p. maximin pour une alternative concernant la famille D de densités.

En effet, d'après les égalités du lemme, on peut poser :

$$(3.11) \qquad \Psi_D(u) = \Psi_h(u) = -1 - (H^{-1}(u) - \bar{h})\, \frac{h'(H^{-1}(u))}{h(H^{-1}(u))} \qquad \forall\, h \in D$$

$$(3.12) \qquad I_D = I_h = \int_0^1 \Psi_h^2(u)\, du \qquad \forall\, h \in D$$

$$(3.13) \qquad S_n(D) = \left(\frac{m\, m'}{m+m'}\, I_h\right)^{-1/2} \left(\sum_{i=1}^{m} \Psi_h\left(\frac{R_{ni}}{m+m'+1}\right) - \frac{m}{m+m'} \sum_{i=1}^{m+m'} \Psi_h\left(\frac{i}{m+m'+1}\right) \right) \forall\, h \in D$$

et si on considère les fonctions de test :

$$(3.14) \qquad \begin{cases} \phi_n = 1 & \text{si} \quad S_n(D) \geqslant k_{1-\alpha} \\ = 0 & \text{si} \quad S_n(D) < k_{1-\alpha} \quad , \end{cases}$$

on a le résultat suivant :

Résultat 4 :

Si les hypothèses (a) et (b) du résultat 1 sont vérifiées et si :

(c) Ψ_D est la somme non constante d'un nombre fini de fonctions monotones de carrés intégrables,

(d) $\{f_h(x, \theta)\}$ satisfait aux conditions (C1) à (C5) pour une densité h de D ,

alors les fonctions de test (3.14) déterminent un test a.u.p.p. maximin de seuil α de H_n^o contre :

$$(3.15) \qquad \{(e^{-\Delta_n} h((x - \bar{h})\, e^{-\Delta_n} + \bar{h})\ ;\ h(x)) \ , \qquad \forall\, h \in D \ \} \ ,$$

de puissance asymptotique égale à $1 - \Phi(k_{1-\alpha} - \sqrt{\mathscr{C} I_D}\ )$.

Démonstration :

Cela résulte du lemme 4 et du résultat 3 où on prend $a = \bar{g}$.

En effet, pour tout h appartenant à D ,

$$\exists\, \sigma > 0 \quad \text{et} \quad \beta \in \mathbb{R}\ /\ h = g_{\sigma,\beta}$$

d'où : $a\, \sigma + \beta = \sigma\, \bar{g} + \beta = \bar{h}$

d'après (3.2). L'hypothèse alternative (2.6) du test considéré au résultat 3 coïncide donc avec celle de l'énoncé. Par ailleurs le lemme 4 permet de remplacer Ψ_g et $S_n^*(g)$ donnés par (2.8) et (2.9) par Ψ_D et $S_n(D)$ déterminés à partir d'une densité h quelconque de D .

Remarque : En pratique on a intérêt à déterminer Ψ_D et $S_n(D)$ à partir de la densité h de D qui amène les calculs les plus simples. Il s'agit souvent de celle dont la moyenne $\bar{h}$ vaut 0 .

3.2 - Exemples

Soit g la densité de la loi normale centrée et réduite et $D = \{g_{\sigma,\beta} , \sigma > 0 , \beta \in \mathbb{R}\}$.
D est donc l'ensemble des densités des variables $\sigma X + \beta$ où X est de loi normale centrée et réduite. Si on note $\mathcal{N}(\beta, \sigma)$ la densité de la loi normale de moyenne β et d'écart-type σ , D est donc l'ensemble des densités normales :

$$D = \{\mathcal{N}(\beta, \sigma) , \forall \sigma > 0 , \forall \beta\in\mathbb{R}\} .$$

L'alternative multiple (3.15) du résultat 4 devient :

$$(3.16) \qquad \{(\mathcal{N}(\beta, e^{\Delta_n} \sigma) ; \mathcal{N}(\beta, \sigma)) , \forall \sigma > 0 , \forall \beta \in \mathbb{R} \} .$$

On trouve, à partir de (3.11), (3.12) et (3.13) :

$$\Psi_D (u) = -1 - [\Phi^{-1}(u)]^2 \quad \forall u \in]0,1[,$$

$$I_D = 2$$

$$S_n(D) = \sqrt{\frac{m + m'}{2m m'}} \left(\sum_{i=1}^{m} [\Phi^{-1} (\frac{R_{ni}}{m+m'+1})]^2 - \frac{m}{m+m'} \sum_{i=1}^{m+m'} [\Phi^{-1} (\frac{i}{m+m'+1})]^2 \right) .$$

Cette dernière statistique est la statistique de Klotz ([2]) . Si les conditions (a) et (b) du résultat 1 sont satisfaites, le test de région de rejet $S_n(D) \geqslant k_{1-\alpha}$ est a.u.p.p. maximin de H_n^o contre l'alternative (3.16).
La puissance asymptotique est égale à $1 - \Phi(k_{1-\alpha} - \sqrt{2\mathcal{C}})$.
Nous avons étudié quelques autres exemples, concernant les familles D engendrées par les densités suivantes :

double-exponentielle : $g(x) = \frac{1}{2} e^{-|x|} \quad \forall x \in \mathbb{R}$

logistique : $g(x) = \dfrac{e^{-x}}{(1 + e^{-x})^2} \quad \forall x \in \mathbb{R}$

log-normale : $g(x) = \begin{cases} \dfrac{1}{x\sqrt{2\pi}} e^{-\frac{\mathrm{Log}^2 x}{2}} & \text{si } x > 0 \\ 0 & \text{si } x \leqslant 0 \end{cases}$

Nous résumons dans le tableau ci-dessous les données qui permettent de construire des tests a.u.p.p. maximin à l'aide du résultat 4 , de (3.13) et (3.14).

Table 1 :

nom de la densité g	$\Psi_D(u)$	I_D
normale	$-1 - [\Phi^{-1}(u)]^2$	2
double-exponentielle	$-1 - \mathrm{Log}\,(1 - \|1-2u\|)$	1
logistique	$-1 + (1-2u)\,\mathrm{Log}\,(\frac{1}{u} - 1)$	$\frac{1}{3} + \frac{\pi^2}{9}$
log-normale	$\Phi^{-1}(u) - \sqrt{e}\, e^{-\Phi^{-1}(u)}[\Phi^{-1}(u)+1]$	$1+2e^3 - 2e$

4 - EFFICACITE RELATIVE ASYMPTOTIQUE

Il est rare en pratique de connaître les densités des échantillons sur lesquels on effectue des tests ; en général g est donc inconnue et on ne peut déterminer la statistique optimale $S_n(D)$. Le calcul de l'efficacité relative asymptotique de Pittmann donne une idée de la robustesse des tests bâtis sur $S_n(D)$ suivant les différentes densités que peut suivre l'échantillon.

Soient g et g^* deux densités sur $\mathbb{R}$ et

$$D = \{g_{\sigma,\beta}\ ,\ \sigma > 0\ ,\ \beta \in \mathbb{R}\}\ ,\quad D^* = \{g^*_{\sigma,\beta}\ ,\ \sigma > 0\ ,\ \beta \in \mathbb{R}\}\ .$$

Considérons le test des hypothèses H_n^o contre les hypothèses :

$$H_n^1 = \{(e^{-\Delta_n} h^*((x - \bar{h}^*)\, e^{-\Delta_n} + \bar{h}^*)\ ;\ h^*(x))\ ,\ \forall\, h^* \in D^*\}\ . \tag{4.1}$$

Définition : Nous dirons que le test est bâti sur les statistiques $S_n(D)$ s'il est déterminé par les fonctions de test :

$$\begin{cases} \phi_n = 1 & \text{si}\quad S_n(D) \geqslant k_{1-\alpha} \\ = 0 & \text{si}\quad S_n(D) < k_{1-\alpha}\ . \end{cases} \tag{4.2}$$

Si le double-échantillon satisfait à l'hypothèse (4.1), la suite $S_n(D)$ converge en loi, d'après le résultat 1 et le lemme 2 , vers une loi normale $\mathcal{N}(\mu_1, 1)$ avec

$$\mu_1 = \sqrt{\mathscr{C}}\ \frac{\int_0^1 \Psi_D(u)\, \Psi_{D^*}(u)\, du}{\left(\int_0^1 (\Psi_D(u) - \bar{\Psi}_D)^2\, du\right)^{1/2}}\ ,$$

où $\bar{\Psi}_D = \int_0^1 \Psi_D(u)\, du$. Si la densité g est telle que la famille $\{f_g(x, \theta)\}$ satisfait aux conditions (C1), (C2) et (C3) , on a d'après les propositions 1 et 2 de [4] :

$$\bar{\Psi}_D = 0 \quad \text{et} \quad \int_0^1 \Psi_D^2(u)\, du = I_D$$

d'où :

$$(4.3) \qquad \mu_1 = \left(\frac{\mathscr{C}}{I_D}\right)^{1/2} \int_0^1 \Psi_D(u) \quad \Psi_{D^*}(u)\, du \ .$$

Sous les conditions habituelles, la puissance asymptotique du test de H_n^o contre H_n^1 bâti sur $S_n(D)$ est donc donnée par :

$$\lim_{n\to+\infty} Q_n(S_n(D) \geqslant k_{1-\alpha}) = \lim_{n\to+\infty} Q_n(S_n(D) - \mu_1 \geqslant k_{1-\alpha} - \mu_1) = 1 - \Phi(k_{1-\alpha} - \mu_1) \quad ,$$

où Q_n désigne une loi de probabilité satisfaisant à H_n^1 .

Si on considère le test de H_n^o contre H_n^1 bâti cette fois-ci sur $S_n(D^*)$, on obtient, d'après le résultat 4, une puissance asymptotique égale à $1 - \Phi(k_{1-\alpha} - \mu_2)$ avec :

$$(4.4) \qquad \mu_2 = \sqrt{\mathscr{C} I_D} \quad .$$

<u>Définition</u> : On appelle efficacité asymptotique relative du test bâti sur $S_n(D)$ contre le test bâti sur $S_n(D^*)$, le rapport :

$$(4.5) \qquad e = \left(\frac{\mu_1}{\mu_2}\right)^2 \quad .$$

<u>Remarque</u> : On vérifie facilement que pour toutes densités g et g^* , e est inférieur à 1 , qu'il est égal à 1 si g égale g^* et que plus e est grand plus la puissance asymptotique du test bâti sur $S_n(D)$ approche celle, optimale, qui serait obtenue par le test bâti sur $S_n(D^*)$.

A partir de (4.3), (4.4) et (4.5) on obtient le résultat suivant :

<u>Résultat 5</u> :

Si les conditions (a) et (b) du résultat 1 sont vérifiés et si

(c) Ψ_D et Ψ_{D^*} sont sommes non constantes d'un nombre fini de fonctions monotones de carrés intégrables,

(d) les familles $\{f_h(x, \theta)\}$ et $\{f_{h^*}(x, \theta)\}$ satisfont aux conditions (C1) à (C5) pour un élément h et un élément h^* de D et D^* respectivement, alors l'efficacité relative asymptotique du test bâti sur $S_n(D)$ contre celui bâti sur $S_n(D^*)$, pour le test des hypothèses H_n^o contre les hypothèses H_n^1 définies par (4.1), est égale à :

$$e = \frac{\left(\int_0^1 \Psi_D(u) \ \Psi_{D^*}(u) \ du\right)^2}{I_D \ I_{D^*}} .$$

Nous avons calculé e pour les différentes densités considérées à la fin du chapitre 3 . On obtient les résultats contenus dans le tableau suivant :

Table 2 :

D engendré par la densité : \ D^* engendré par la densité :	Normale	double-exponentielle	logistique	log-normale
normale	1	0,98	0,98	0,41
double-exponentielle	0,98	1	0,996	0,36
logistique	0,98	0,996	1	0,35
log-normale	0,41	0,36	0,35	1

On remarquera le comportement similaire des tests bâtis sur $S_n(D)$ quand D est engendrée par les densités normale, double-exponentielle ou logistique.

Si g est inconnue, on a tout intérêt à utiliser le test bâti à partir de la densité logistique qui exige les calculs les moins difficiles (voir table 1). En particulier, contrairement au test de Klotz, il ne nécessite pas une table (ou un programme de calcul) des valeurs de Φ^{-1} .

REFERENCES

[1] Hajek J. et Sidak Z. - Theory of rank tests.
Academia Publishing House of the Czechoslovak Academy of Sciences, Prague (1967).

[2] Klotz J. - Nonparametric tests for scale.
Annals of Mathematical Statistics, volume 33, 498-512 (1962).

[3] NASI O. - Thèse de 3ème Cycle, Université de Nancy I (1975).

[4] NASI O. - Lois asymptotiques de statistiques de rang sous des alternatives continues ou discrètes.
Statistiques et Analyse des données, 01-02 (1976).