Statistique et analyse des données

J-F. Ingenbleek

Test de rang et processus autorégressif d'ordre un

Statistique et analyse des données, tome 4, n° 1 (1979), p. 47-54

<http://www.numdam.org/item?id=SAD_1979__4_1_47_0>

Statistiques et Analyse des Données
1 - 1979 pp. 47, 54.

TEST DE RANG ET PROCESSUS AUTOREGRESSIF D'ORDRE UN

J-F. INGENBLEEK, Université Libre de Bruxelles.

Résumé. Considérons une réalisation finie d'un processus aléatoire. On désire tester l'hypothèse nulle que le processus générateur est un bruit blanc par rapport à l'alternative que ce processus est autorégressif d'ordre un. Suivant une méthode analogue à celle qui est développée dans [1], on construit une statistique de rang à puissance localement maximale. On développe une condition suffisante de normalité asymptotique sous l'hypothèse nulle. On montre que l'extension des scores de Wilcoxon, de Van der Waerden et médians définissent une statistique asymptotiquement normale.

0. RAPPELS ET INTRODUCTION.

0.1. Rappel

Soient $X_{N1}, X_{N2}, \ldots, X_{NN}$ N variables aléatoires telles que

$$X_{Ni} = \Delta . C_{Ni} + e_{Ni} \quad , \quad i = 1,2, \ldots N ,$$

où Δ est un paramètre inconnu, les C_{Ni} $(i = 1,2,\ldots N)$ N constantes données et les e_{Ni} $(i = 1,2, \ldots N)$ N variables aléatoires indépendantes identiquement distribuées suivant une fonction de répartition F(x) connue, de densité f(x).
Sous certaines conditions générales concernant F(x) [1], on peut montrer que la statistique de rang localement la plus puissante pour tester l'hypothèse nulle $H_o : \Delta = 0$ par rapport à l'alternative $H_1 : \Delta > 0$ est donnée par

$$S_N = \sum_{i=1}^{N} C_{Ni}\, a_N(R_{Ni}, F) \quad ,$$

où $R_{N1}, R_{N2}, \ldots, R_{NN}$ sont les rangs des variables $X_{N1}, \ldots, X_{NN}$ et les $a_N(i,f)$ $(i=1,2,\ldots N)$ N constantes, appelées scores, données par

$$a_N(i,f) = N \binom{N-1}{i-1} \int_0^1 - \frac{f'(F^{-1}(u))}{f(F^{-1}(u))}\, u^{i-1} (1-u)^{N-i}\, du \quad , \quad \text{où } F^{-1}(x) = \inf\{t : F(t) \geq x\}$$

Notons $Z_{N(i)}$ la i^e valeur ordonnée parmi N variables indépendantes et équidistribuées suivant la fonction de répartition F(x). Les scores $a_N(i,f)$ s'écrivent encore :

$$a_N(i,f) = \mathbb{E}\left[- \frac{f'(Z_{N(i)})}{f(Z_{N(i)})} \right] \quad ,$$

où l'espérance est prise dans la distribution de $Z_{N(i)}$.

Rappelons que, dans le cas particulier où $C_{Ni} = 0$ pour $i \leq N_1$ et $C_{Ni} = 1$ pour $i > N_1$, on retrouve notamment la statistique de Wilcoxon, Van der Waerden et de la médiane en prenant pour F(x), les fonctions de répartition des variables logistique, normale et double exponentielle.

0.2. Introduction.

Considérons à présent le processus autorégressif d'ordre 1 défini par

$$X_{Ni} = \Delta . X_{Ni-1} + e_{Ni} \quad , \quad i = 2,3, \ldots, N \quad ,$$

où Δ est un paramètre inconnu, les e_{Ni} $(i = 2,3, \ldots, N)$ N-1 variables aléatoires indépendantes et identiquement distribuées suivant une fonction de répartition F(x) connue ($\frac{d^2F}{dx^2} = \frac{df}{dx} = f'(x)$), et X_{N1} une variable aléatoire indépendante des e_{Ni} , distribuée suivant une fonction de répartition $F(x;\Delta)$ ($\frac{dF(x;\Delta)}{dx} = f(x;\Delta)$) dépendant de Δ. On suppose également que la condition $F(x;0) \equiv F(x)$ est satisfaite.
On désire tester l'hypothèse nulle H_o : $\Delta = 0$ par rapport à l'alternative H_1 : $\Delta > 0$ ($\Delta < 0$ ou $\Delta \neq 0$) à l'aide d'une statistique de rang $S_N = g_N(R_{N1}, \ldots, R_{NN})$, où g_N est une certaine fonction mesurable à déterminer.

1. STATISTIQUE DE RANG LOCALEMENT LA PLUS PUISSANTE.

1.1. On remarque que, sous l'hypothèse nulle H_o, chaque configuration de rang (R_{N1} , $R_{N2}, \ldots, R_{NN}$) porte la même probabilité de 1/N!. Il n'en est pas ainsi sous l'hypothèse alternative. Aussi pour tester l'hypothèse nulle H_o : $\Delta = 0$ par rapport à l'alternative

H_1' : $\Delta = \delta$ (hypothèse alternative simple), un test de rang au niveau α et à puissance maximum aura-t-il pour zone critique les $\alpha.N!$ configurations de rang $(R_{N1}, \ldots, R_{NN})$ les plus probables sous H_1' .

Supposons à présent que l'on puisse écrire, au voisinage de $\Delta = 0$,

$$P(R_{N1} = r_{N1}, \ldots, R_{NN} = r_{NN}|\Delta) = p_\Delta(r_{N1}, \ldots, r_{NN}) =$$

$$\frac{1}{N!} + \Delta . \partial_\Delta p_\Delta|_{\Delta=0} + \alpha(\Delta) \quad ;$$

Pour un δ suffisamment petit, les configurations de rang les plus probables sont donc celles pour lesquelles le terme du premier ordre

$$g(R_{N1}, \ldots, R_{NN}) = \partial_\Delta P_\Delta(R_{N1}, \ldots R_{NN})|_{\Delta=0} \quad \text{est le plus grand,}$$

$g_N(R_{N1}, R_{N2}, \ldots, R_{NN})$ fournit une statistique de rang à puissance localement maximum pour l'alternative H_1 : $\Delta > 0$. Si les densités $f(x_1, s)$ et $f(x)$ vérifient des conditions de dérivation sous le signe, on obtient

$$\partial_\Delta \int f(x_1, \Delta) \prod_{i=1}^{n-1} f(x_i - \Delta x_{i-1}) dx_1 \ldots dx_n =$$

$$g_N(R_{N1}, \ldots, R_{NN}) = S_N = a^{(1)}(R_{N1}) - \sum_{i=2}^{N} a_N(R_{Ni-1}, R_{Ni}) \quad ,$$

où

$$\begin{cases} a_N^{(1)}(i) = \mathbb{E}\, \dfrac{\partial_\Delta f(Z_{N(i)}, 0)}{f(Z_{N(i)})} & i = 1,2, \ldots, N \\ \\ a_N(i,j) = \mathbb{E}\left[Z_{N(i)} \cdot \dfrac{f'(Z_{N(j)})}{f(Z_{N(j)})} \right] & \begin{matrix} i = 1,2, \ldots N \\ j = 1,2, \ldots N \\ i \neq j \end{matrix} \end{cases}$$

$Z_{N(1)}, \ldots, Z_{N(N)}$ étant les N variables auxiliaires définies en 0.1.

2. DISTRIBUTION EXACTE SOUS H_o .

2.1. Soient des constantes arbitraires $a_N^{(1)}(i)$, $(i = 1,2, \ldots N)$ et $a_N(i,j)$ $(i \neq j = 1,2, \ldots N)$. Elles définissent une statistique du type 1.2. Celle-ci ne sera localement la plus puissante si les scores $a_N^{(1)}(i)$ et $a_N(i,j)$ vérifient les conditions imposées en 1.2. Signalons quelques propriétés de S_N pour des scores arbitraires.

2.2. Il est clair que la distribution de S_N sous H_o est indépendante de la forme analytique de $F(x)$, puisque chaque configuration de rang $(R_{N1}, \ldots, R_{NN})$ porte la même probabilité de $1/N!$.

On peut aisément déterminer les moments de S_N. On a, par exemple,

$$\mathbb{E}\, S_N = \bar{a}_N^{(1)} + (N-1)\,\bar{a}_N$$

$$= \frac{1}{N} \sum_{i=1}^{N} a_N^{(1)}(i) + \frac{1}{N} \sum_{i=1}^{N} \sum_{\substack{j=1 \\ \neq i}}^{N} a_N(i,j) \quad .$$

Dans le cas particulier où

$$\sum_{\substack{j=1 \\ \neq i}}^{n} a_N(i,j) = \sum_{\substack{i=1 \\ \neq j}}^{N} a_N(i,j) = \frac{1}{N} \sum_{i=1}^{N} a_N^{(1)}(i) = 0 \quad \forall\, i,\, j = 1,2, \ldots N \quad ,$$

$$D^2 S_N = \sigma^2_{a(1)} + N\,\sigma^2_a$$

$$= \frac{1}{N} \sum_{i}^{n} a_N^{(1)2}(i) + \frac{1}{N-1} \sum_{i=1}^{n} \sum_{\substack{j \neq i \\ =1}}^{n} a_N^2(i,j) \quad .$$

2.3. La distribution de S_N peut s'obtenir par énumération des N! configurations de rang. Pour N grand, cette énumération devient laborieuse. Il est nécessaire de connaître la distribution asymptotique de S_N sous H_o .

3. DISTRIBUTION ASYMPTOTIQUE SOUS H_o .

3.1. Posons

$$S'_N = a_N(R_{N1}, R_{N2}) + a_N(R_{N2}, R_{N3}) + \ldots + a_N(R_{NN-1}, R_{NN}) + a_N(R_{NN}, R_{N1})$$

$$= S_N - a^{(1)}(R_{N1}) - a_N(R_{NN}, R_{N1}) \quad .$$

S'il existe des constantes c_n et b_n (n = 1,2, ...) tel que $\dfrac{S'_n - c_n}{b_n}$ converge en loi, vers une distribution N(0,1) alors S_n sera asymptotiquement normale (c_n, b_n) pour autant que

$$\begin{cases} \dfrac{a^{(1)}(R_{N1})}{b_n} \xrightarrow{P} 0 \\[2ex] \dfrac{a(R_{NN}, R_{N1})}{b_n} \xrightarrow{P} 0 \end{cases}$$

Dans les paragraphes qui suivent nous allons supposer que ces deux conditions sont satisfaites; nous étudions donc la normalité asymptotique de S'_N. Pour alléger l'écriture, nous noterons de la même manière S'_N et S_N .

3.2. A toute configuration de rang $(R_{N1}, \ldots, R_{NN})$, on peut associer une permutation σ de l'ensemble (1,2, ... n) en posant

$$\sigma(i) = R_{Ni} \qquad i = 1,2, \ldots N$$

Définissons la permutation s par

$$\begin{cases} s(i) = i+1 & , \quad i = 1,2, \ldots N-1 \\ s(N) = 1 \end{cases}$$

On peut alors trivialement voir que

<u>lemme 1.</u> $\quad S_N = \sum_{i=1}^{N} a_N(i, \sigma.s.\sigma^{-1}(i))$, où $\sigma(i) = R_{Ni}$.

Désignons par $\mathcal{S}_N$ l'ensemble des permutations de (1,2,... n). Soit R la relation d'équivalence:

$$\sigma \tilde{R} \tau \text{ ssi } \quad \sigma\, s\, \sigma^{-1} = \tau\, s\, \tau^{-1} .$$

Désignons par $\tilde{\sigma}$ les éléments de $\tilde{\mathcal{S}}_N$, quotient de $\mathcal{S}_N$ par $\tilde{R}$.

<u>lemme 2.</u> S_N est distribuée comme la variable aléatoire $\sum_{i=1}^{N} a_N(i, \tilde{\sigma}(i))$, où $\tilde{\sigma}$ est une permutation aléatoire de distribution uniforme sur $\mathcal{D}$.

Ce dernier lemme permet de rattacher l'étude de la statistique S_N à celle des statistiques bilinéaires de rang. Rappelons qu'une statistique bilinéaire de rang se définit de la manière suivante : soient $d_N(i,j)$ (i = 1,2, ... N, j = 1,2, ... N) N^2 coefficients et σ une permutation aléatoire de distribution uniforme sur $\mathcal{S}_N$; alors la variable aléatoire
$T_N = \sum_{i=1}^{N} d_N(i, \sigma(i))$ est appelée statistique bilinéaire de rang.

W. Hoeffding et M. Motoo ([2], [3]) ont étudié séparément la normalité asymptotique de T_N. M. Motoo, en s'appuyant sur un théorème central limite pour des variables non indépendantes a trouvé une condition suffisante de normalité asymptotique plus générale que celle de W. Hoeffding qui lui, étudie la convergence des moments de T_N.
Avant d'examiner la normalité asymptotique de S_N, signalons le lemme suivant, dont la vérification est immédiate :

<u>lemme 3.</u> Soient $a_N(i,.) = \sum_{\substack{j=1 \\ \neq i}}^{n} a_N(i,j)$, $a_N(.,j) = \sum_{\substack{i=1 \\ \neq j}}^{n} a_N(i,j)$ et

$$a'_N(i,j) = a_N(i,j) - \frac{a_N(.,i) + (N-1)a_N(i,.)}{N(N-2)} - \frac{a_N(j,.) + (N-1)a_N(.,j)}{N(N-2)}$$

Si $\bar{a}_N = 0$ alors $a'_N(i,.) = a'_N(.,j) = 0 \quad i,j = 1,2, \ldots N$, de plus les statistiques basées sur les scores $a_N(i,j)$ et $a'_N(i,j)$ sont égales .

3.3. Il est alors aisé de montrer que

Proposition. Soient $a'_N(i,j)$ les scores déduits de $a_N(i,j)$ conformément au lemme 3, (avec $\bar{a}_N = 0$)

Si $\lim_N \dfrac{\max_{i,j=1,2\ldots n} a'^2_N(i,j)}{\frac{1}{n}\sum_{i=1}^{n}\sum_{j\neq i}^{n} a'^2_N(i,j)} = 0$ (condition dite de Hoeffding), S_N est asymptotiquement normale $(0, \mathbb{D}(S_n))$.

Esquissons la démonstration. On définit à partir des scores $a'_N(i,j)$ une statistique bilinéaire T_N en posant

$$d_N(i,j) = \begin{cases} a_N(i,j) & i \neq j \\ 0 & i = j \end{cases}$$

T_N est définie sur $\mathcal{S}_N$ et S_N sur $\mathcal{D}$; cependant T_N et S_N possèdent asymptotiquement les mêmes moments. La condition de Hoeffding assurant la normalité de T_N, on en déduit celle de S_N.

4. QUELQUES CAS OU LA PROPOSITION EST VERIFIEE.

4.1. Les scores optimaux vérifient-ils la condition de Hoeffding ? Dans le paragraphe qui suit nous considérons (par analogie avec [1]) la classe des scores $a_N(i,j)$ vérifiant les trois conditions suivantes :

(i) $\bar{a}_N = \dfrac{1}{N(N-1)} \sum_{i=1}^{n}\sum_{j\neq i}^{n} a_N(i,j) = 0$;

(ii) $a_N(i,j) = E\,\varphi(Z_{N(i)}, Z_{N(j)})$ où $Z_{N(i)}$ est la variable auxiliaire définie précédemment, mais correspondant à une variable uniforme sur $[0,1]$, et où $\varphi(x,y)$ est une fonction réelle définie sur $[0,1] \times [0,1]$, de carré sommable.

(iii) $0 < \int_0^1 \int_0^1 \varphi^2(x,y)\,dx\,dy \;(= \mathbb{E}\,\varphi^2) < \infty$.

Si $F(x)$ est strictement croissante, les scores optimaux vérifient la condition (ii) en prenant

$$\varphi(x,y) = F^{-1}(x) \quad \frac{f'(F^{-1}(y))}{f(F^{-1}(y))}$$

Par extension du théorème a V.1.4 de [1] on a

Propriété : $\lim_{N\to\infty} \dfrac{1}{N(N-1)} \sum_{i=1}^{n}\sum_{j\neq i}^{n} a_N^2(i,j) = \mathbb{E}\,\varphi^2$

On peut également montrer sans difficulté que

Propriété : $\lim_{N\to\infty} \frac{1}{N(N-1)} \sum_{i=1}^{n} \sum_{j\neq i}^{n} a'^2_N(i,j) = c^2 > 0$

Pour la classe de scores considérée, la condition de Hoeffding prend ainsi la forme plus simple :

$$\lim_{N\to\infty} \frac{\max_{i,j=1,2\ldots n} a'^2_N(i,j)}{N} = 0 \quad .$$

4.2. Lemme 1 : Soient $Z_{N(N)}$ et $Z_{N(1)}$ les variables auxiliaires définies en 0.1 avec une fonction de répartition F(x) admettant des moments par rapport à l'origine jusqu'à l'ordre k; alors

$$\lim_{N\to\infty} \frac{\mathbb{E}\, Z^k_{N(N)}}{N} = \lim_{N\to\infty} \frac{\mathbb{E}\, Z^k_{N(1)}}{N} = 0 \quad .$$

Proposition : Soient $a_N(i,j)$ des scores appartenant à la classe considérée, et optimaux pour une certaine fonction de répartition F(x). Alors, si $\frac{f'(x)}{f(x)}$ est bornée, les $a_N(i,j)$ vérifient la condition de Hoeffding.

Démonstration : Il est facile de voir que l'on a

$$a^2_N(i,j) \leq K\, \mathbb{E}\, Z^2_{N(1)} + K\, \mathbb{E}\, Z^2_{N(N)} \quad .$$

Par ailleurs, si $|a_N(i,j)|$ est bornée pour tout i,j par B_N, alors $|a'_N(i,j)|$ est bornée par $7\, B_N$. Il suffit ensuite d'appliquer le lemme 4.2.
Il découle de cette propriété que les scores optimaux correspondent à une distribution logistique (extension des scores de Wilcoxon) et double exponentielle (extension des scores médians) vérifient la condition de Hoeffding.

4.3. Propriété. Les scores optimaux correspondant à une distribution normale (extension des scores de Van der Waerden) vérifient la condition de Hoeffding.

Démonstration : Les scores normaux s'écrivent

$$a_N(i,j) = \mathbb{E}\, Z_{N(i)} \cdot Z_{N(j)}$$

où $Z_{N(i)}$ est la variable auxiliaire définie en 0.1 avec une fonction de répartition normale. Le théorème découle alors du lemme 4.2 en remarquant que, d'une manière générale,

$$\mathbb{E}^2\, Z_{N(i)} \cdot Z_{N(j)} \leq 8\, \mathbb{E}\, Z^4_{N(1)} + 8\, \mathbb{E}\, Z^4_{N(N)}$$

Références.

[1] HAJEK, J. et SIDAK, Z. : Theory of rank test, Academic Press, New York (1967).

[2] HOEFFDING, W. (1951) : A combinatorial central limit Theorem, AMS 22, 558-566.

[3] MOTOO, M. (1957) : On the Hoeffding's combinatorial central limit Theorem, AISM 8, 145-154.

REMERCIEMENTS.

Je tiens à remercier le Referee pour ses utiles remarques.