REVUE DE STATISTIQUE APPLIQUÉE

M. PINHAS

Valeur de l'information d'un échantillon ordonné : le cas gaussien

Revue de statistique appliquée, tome 28, nº 2 (1980), p. 87-90
<http://www.numdam.org/item?id=RSA_1980__28_2_87_0>

VALEUR DE L'INFORMATION D'UN ÉCHANTILLON ORDONNÉ : LE CAS GAUSSIEN

M. PINHAS

1. – UN THÉORÈME DE WATTERSON

Le point de départ de la présente étude se trouve dans le résultat suivant du à Watterson [1] :

Soit (X , Y) un vecteur gaussien de dimension 2, de paramètres μ_x , μ_y , σ_x^2, σ_y^2 et ρ. On considère un échantillon de taille n d'observations vectorielles (x_i , y_i) , i = 1 . . . n . Après avoir ordonné de manière croissante les réalisations de la première composante X : $x_{(1)} \leqslant \ldots \leqslant x_{(n)}$, on désigne les valeurs correspondantes de Y (qui ne sont pas nécessairement rangées en ordre croissant) par $y_{[1]} , \ldots y_{[n]}$.

Alors, à propos des variables $Y_{[r]}$, r = 1, . . . n, on peut affirmer, entre autres propriétés, que :

$$E\,Y_{[r]} = \mu_y + \rho\, \sigma_y\, \mu_{r:n}$$

$$\operatorname{Var} Y_{[r]} = (1 - \rho^2)\, \sigma_y^2 + \rho^2\, \sigma_y^2\, \sigma_{r:n}^2 \ ,$$

où $\mu_{r:n}$ et $\sigma_{r:n}^2$ sont respectivement l'espérance et la variance de la $r^{\text{ème}}$ variable aléatoire d'un échantillon ordonné de taille n, issu d'une loi parente gaussienne centrée réduite.

Il est utile de savoir d'une part qu'il existe des tables donnant la valeur de ces caractéristiques lorsque n est inférieur ou égal à 20 [2], et d'autre part que le caractère symétrique et normalisé de la dernière distribution mentionnée autorise l'écriture des relations :

$$\sum_{r=1}^{n} \mu_{r:n} = 0 \ , \ \mu_{r:n} = - \mu_{n-r+1:n}$$

$$\sum_{r=1}^{n} \sum_{s=1}^{n} \sigma_{rs:n} = n \quad (\sigma_{rs:n} \text{ covariance des v.a. de rang r et s}).$$

Mots-Clés : **Valeur de l'information.**
Information qualitative.

Deux applications de ce théorème proposons nous dans cette note. La première consiste en une évaluation du prix maximum que consentirait à payer un agent économique confronté à n projets indépendants et incompatibles, pour disposer d'une certaine information *qualitative* les concernant.

Dans le second développement nous décrivons une procédure de calcul des primes d'assurance, fondée sur le cumul des réclamations formulées par les assurés et, de ce fait, susceptible de créer entre eux une réelle émulation.

2. – PREMIÈRE APPLICATION : VALEUR MOYENNE DE L'INFORMATION PROVENANT D'UN ÉCHANTILLON ORDONNÉ

Pour simplifier l'exposé limitons nous au cas $n = 2$ et considérons un agent économique neutre à l'égard du risque(*) et devant effectuer un choix entre deux perspectives incompatibles et indépendantes, dont le gain aléatoire est représenté par la v.a. V_i, gaussienne, de paramètres μ_i et σ_i^2 $(i = 1, 2)$. Supposons chaque variable V_i statistiquement liée à une autre variable $U_i(\alpha_i, \gamma_i^2)$, de telle sorte que les deux couples (X_i, Y_i) soient indépendants, gaussiens centrés réduits, et de coefficient de corrélation linéaire ρ :

par définition $$X_i = \frac{U_i - \alpha_i}{\gamma_i} \quad \text{et} \quad Y_i = \frac{V_i - \mu_i}{\sigma_i} .$$

Il n'est évidemment pas restrictif d'admettre que ρ est positif.

Enfin nous émettons l'hypothèse qu'avant de se prononcer, le décideur peut prendre connaissance, non pas des réalisations des variables X_i, mais de leur classement.

Quelle valeur moyenne (maximale) Δg *attribuer à cette information* ?

Calculons d'abord le gain moyen $\widetilde{g}$ résultant de la procédure informative. Pour cela il est nécessaire d'examiner deux cas :

a) si l'évènement $(X_1 < X_2)$ se produit (probabilité 1/2), alors

$$(X_{(1)}, Y_{[1]}) = (X_1, Y_1) .$$

Or d'après le résultat du paragraphe I :

$$E\, Y_{[1]} = \rho\, \mu_{1:2} \quad \left(\text{quantité négative, car } \mu_{1:2} = -\pi^{-\frac{1}{2}}\right)$$

$$E\, Y_{[2]} = \rho\, \mu_{2:2} = -\rho\, \mu_{1:2}$$

D'où $\quad E\, V_{[i]} = \mu_i + \rho\, \sigma_i\, \mu_{i:2}$. Posons $\quad g_a = \max\limits_{i=1,2} \{\mu_i + \rho\, \sigma_i\, \mu_{i:2}\}$

(*) Le traitement du cas bernoullien général ne peut être conduit à son terme qu'au prix de lourdes approximations.

b) si $(X_1 > X_2)$ a lieu (probabilité 1/2), on obtient au contraire

$(X_{(1)}, Y_{[1]}) = (X_2, Y_2)$ et donc $E\,V_{[i]} = \mu_j + \rho\,\sigma_j\,\mu_{i:2}$ avec $j \neq i$.

Définissons de même

$$g_b = \max_{\substack{i=1,2 \\ j\neq i}} \{\mu_j + \rho\,\sigma_j\,\mu_{i:2}\}\ .$$

Et donc $\tilde{g} = \frac{1}{2}(g_a + g_b)$. Par ailleurs il est clair que, sans information, le gain moyen s'élèverait à :

$$\hat{g} = \max(\mu_1, \mu_2)\ .$$

Nous allons montrer que la valeur de l'information $\Delta g = \tilde{g} - \hat{g}$ s'exprime de façon simple à l'aide des paramètres du modèle.

Première hypothèse : $|\mu_2 - \mu_1| \geqslant -\mu_{1:2}\,\rho(\sigma_1 + \sigma_2)$.

- si $\mu_2 \geqslant \mu_1$:

$\hat{g} = \mu_2$

$g_b = \mu_2 + \rho\,\sigma_2\,\mu_{1:2}$

$g_a = \mu_2 - \rho\,\sigma_2\,\mu_{1:2}$ (à cause du signe négatif de $\rho\,\mu_{1:2}$)

D'où $\tilde{g} = \mu_2$ et $\Delta g = 0$.

- si $\mu_2 \leqslant \mu_1$:

$\hat{g} = \mu_1$

$g_a = \mu_1 + \rho\,\sigma_1\mu_{1:2}$

$g_b = \mu_1 - \rho\,\sigma_1\,\mu_{1:2}$

D'où $\tilde{g} = \mu_1$ et $\Delta g = 0$.

Deuxième hypothèse : $|\mu_2 - \mu_1| < -\mu_{1:2}\,\rho(\sigma_1 + \sigma_2)$.

Alors $g_a = \mu_2 - \rho\,\sigma_2\,\mu_{1:2}$ et $g_b = \mu_1 - \rho\,\sigma_1\,\mu_{1:2}$,

et

$$\tilde{g} = \frac{1}{2}(\mu_1 + \mu_2) - \frac{1}{2}\,\rho\,\mu_{1:2}\,(\sigma_1 + \sigma_2)\ .$$

Finalement, quel que soit le signe de $\mu_2 - \mu_1$, Δg s'écrit :

$$\Delta g = -\frac{1}{2}\,|\mu_2 - \mu_1| - \frac{1}{2}\rho\,\mu_{1:2}\,(\sigma_1 + \sigma_2)\ .$$

Nous constatons ainsi que, pour μ_1 et μ_2 fixés, la valeur de l'information est une fonction non décroissante de ρ, σ_1 et σ_2.

3. – UNE AUTRE APPLICATION : LE CALCUL DE CERTAINES PRIMES D'ASSURANCE

Dans ce paragraphe nous envisageons d'utiliser le théorème de Watterson en vue d'établir une procédure de tarification actuarielle destinée à une population de risques a priori homogène. A cet effet nous supposons que le risque afférent à l'assuré i est représentable par un vecteur gaussien (X_i , Y_i), de caractéristiques $\mu_x , \sigma_x^2 , \mu , \sigma^2$ et ρ. Ces n vecteurs, mutuellement indépendants ont pour première composante la somme des réclamations formulées au cours de la période $[-t , 0]$, et pour Y_i le montant des dommages relatifs à l'année $[0 , 1]$.

Le calcul des primes individuelles pour la période $[0 , 1]$, qu'il est raisonnable de fonder sur le classement des réalisations (x_i), pourrait selon nous être effectué conformément au schéma suivant :

1) Calcul de la prime globale p correspondant au risque $\sum_1^n Y_i$, par exemple, au moyen du principe de l'indice de sécurité [3]. Ce qui donne présentement : $p = n\mu + \frac{\tau}{2} n\sigma^2$ ($\tau > 0$, valeur de l'indice de sécurité retenue).

2) En plus du montant μ, chaque assuré [i] doit verser à la compagnie un correctif $C_{[i]}$.

3) Les quantités $C_{[i]}$ sont proportionnelles aux espérances $E\,Y_{[i]}$. Par conséquent, en égalant à p la somme des primes individuelles, il vient :

$$\frac{C_{[i]}}{\mu + \rho\sigma\mu_{i:n}} = \frac{\tau\,\sigma^2}{2\,\mu}$$

et donc :

$$C_{[i]} = \frac{\tau}{2}\,\sigma^2 \left(1 + \rho\,\mu_{i:n}\frac{\sigma}{\mu}\right).$$

Notons enfin que le regroupement des résultats x_i en k classes s'avère pratiquement indispensable lorsque n est grand (non-existence de tables pour les $\mu_{i:n}$).

BIBLIOGRAPHIE

[1] WATTERSON. – Linear estimation in censored samples from multivariate normal populations. *Ann. Math. Statisti. 30,* 814-24.

[2] TEICHROEW. – Tables of expected values of order statistics. *Ann. Math. Statist. 27,* 410-26.

[3] PINHAS. – Assurance à la carte et prévention (Colloque Structures économiques et Econométrie, Lyon 1980).