REVUE DE STATISTIQUE APPLIQUÉE

R. SNEYERS

Sur la notion d'indépendance climatologique

Revue de statistique appliquée, tome 14, n° 2 (1966), p. 31-36

<http://www.numdam.org/item?id=RSA_1966__14_2_31_0>

SUR LA NOTION D'INDÉPENDANCE CLIMATOLOGIQUE [1]

R. SNEYERS
Institut Royal Météorologique de Belgique, Uccle

1 - INTRODUCTION

Il est bien connu que les erreurs d'échantillonnage qui affectent les estimations des paramètres lors de l'ajustement d'une loi de probabilité à un échantillon aléatoire et simple de valeurs observées d'une variable aléatoire donnée sont généralement d'autant plus grandes que la taille de l'échantillon est petite. C'est notamment le cas pour les estimateurs fournis par la méthode des moments ou dérivés du principe de la vraisemblance maximale et pour lesquels les variances des erreurs d'échantillonnage sont inversement proportionnelles à la taille de l'échantillon.

Il en résulte que lorsque les échantillons sont très petits, les ajustements qu'on en déduit sont pratiquement inutilisables en raison de l'ampleur prise par ces erreurs.

Tous les petits échantillons ne doivent cependant pas être traités comme tels et la constatation précédente perd de sa rigueur si la variable étudiée est en corrélation avec une autre variable dont la loi de distribution est bien connue et déterminée.

Le but de cette note est, en effet, de montrer que dans un cas assez général, dès que l'on dispose d'une série de valeurs correspondantes des deux variables et pour autant que le coefficient de corrélation entre ces variables dépasse un seuil minimal, un ajustement plus précis que celui fourni par la méthode ordinaire peut être obtenu.

2 - GENERALITES

Soient x et y deux variables aléatoires dont les lois de distributions se définissent au moyen de la même variable réduite grâce aux relations (2) :

---

(1) Communication présentée à la 35e Session de l'Institut International de Statistique. Beograd, 1965.

(2) On notera que les paramètres d'échelle σ_1 et σ_2 peuvent être de mêmes signes ou de signes contraires.

$$x = \sigma_1 t + \mu_1 \qquad \text{et} \qquad y = \sigma_2 t + \mu_2 \tag{1}$$

et soit

$$y = a_1 x + b_1 \tag{2}$$

la relation linéaire, déduite des relations (1) par élimination de t, qui détermine la loi de distribution de la variable y à partir de celle de la variable x.

On a donc les relations :

$$a_1 = \sigma_2 / \sigma_1 \qquad \text{et} \qquad b_1 = \mu_2 - a_1 \mu_1 \tag{3}$$

Supposons en outre que les paramètres σ_1 et μ_1 soient connus et que les coefficients a_1 et b_1 soient des inconnues que l'on cherche à estimer.

Il s'ensuit que si l'on dispose d'un échantillon aléatoire et simple de taille n de valeurs correspondantes de x et de y et si $\hat{\sigma}_i$ et $\hat{\mu}_i$ sont les estimations de σ_1 et de μ_i (i = 1, 2) qu'on en déduit, on peut considérer en toute généralité les fonctions :

$$\hat{a}_1 = \hat{\sigma}_2 / \sigma_1 \qquad \text{et} \qquad \hat{b}_1 = \hat{\mu}_2 - \hat{a}_1 \mu_1 \tag{4}$$

comme estimateurs des inconnues, tandis que dans le cas de l'existence d'une corrélation non nulle entre les variables x et y, on peut aussi envisager comme estimateurs de a_1 et de b_1 les fonctions :

$$a_1^* = \hat{\sigma}_2 / \hat{\sigma}_1 \qquad \text{et} \qquad b_1^* = \hat{\mu}_2 - a_1^* \hat{\mu}_1 \tag{5}$$

en raison de la réduction de leur variance que l'existence de la corrélation permet d'envisager.

Il reste toutefois à établir à quelle condition la variance des estimateurs (5) sera moindre que celle des estimateurs (4) afin de pouvoir faire le meilleur choix.

3 - LA CORRELATION MINIMALE

Si l'on pose $\hat{a}_1 = a_1 + d\hat{a}_1$ et $a_1^* = a_1 + d a_1^*$ et si l'on définit $d\hat{\sigma}_1$ et $d\hat{\sigma}_2$ d'une façon analogue, il vient avec (4), en négligeant les termes d'ordre supérieurs en $d\hat{\sigma}_1$ dans le développement de $d\alpha_1^*$:

$$d\hat{a}_1 = a_1 \frac{d\hat{\sigma}_2}{\sigma_2} \qquad \text{et} \qquad d a_1^* = a_1 \frac{d\hat{\sigma}_2}{\sigma_2} - a_1 \frac{d\hat{\sigma}_1}{\sigma_1} \tag{6}$$

c'est-à-dire :

$$d a_1^* = d\hat{a}_1 - a_1 \frac{d\hat{\sigma}_1}{\sigma_1}$$

En élevant les termes de cette dernière relation au carré et en prenant les moyennes, on en tire (1) :

---

(1) Il est clair que ces relations ne sont qu'approchées dans le cas où les estimateurs $\hat{\sigma}_1$ et $\hat{\sigma}_2$ présentent un biais.

$$\text{var } a_1^* = \text{var } \hat{a}_1 + (a_1/\sigma_1)^2 \text{ var } \hat{\sigma}_1 - 2\ (a_1/\sigma_1) \text{ cov } (\hat{a}_1, \hat{\sigma}_1)$$

ou encore

$$\text{var } \hat{a}_1 - \text{var } a_1^* = (a_1/\sigma_1)^2 \left[\frac{2}{a_1} \text{cov } (\hat{\sigma}_2, \hat{\sigma}_1) - \text{var } \hat{\sigma}_1 \right] \qquad (7)$$

De la même manière il vient :

$$d\, b_1^* = d\, \hat{b}_1 - a_1\, \mu_1 \left(\frac{d\, \hat{\mu}_1}{\mu_1} - \frac{d\, \hat{\sigma}_1}{\sigma_1} \right)$$

ce qui conduit à :

$$\text{var } b_1^* = \text{var } \hat{b}_1 + (a_1\, \mu_1)^2 \text{ var } \left(\frac{\hat{\mu}_1}{\mu_1} - \frac{\hat{\sigma}_1}{\sigma_1} \right) - 2\, a_1\, \mu_1 \text{ cov } \left(\hat{b}_1, \frac{\hat{\mu}_1}{\mu_1} - \frac{\hat{\sigma}_1}{\sigma_1} \right)$$

On en déduit :

$$\text{var } \hat{b}_1 - \text{var } b_1^* = (a_1\, \mu_1)^2 \left[\frac{2}{a_1} \text{cov } \left(\frac{\hat{\mu}_2}{\mu_1} - \frac{\hat{\sigma}_2}{\sigma_1}, \frac{\hat{\mu}_1}{\mu_1} - \frac{\hat{\sigma}_1}{\sigma_1} \right) - \text{var } \left(\frac{\hat{\mu}_1}{\mu_1} - \frac{\hat{\sigma}_1}{\sigma_1} \right) \right] \qquad (8)$$

Il en résulte que les estimateurs a_1^* et b_1^* devront être préférés aux estimateurs $\hat{a}_1$ et $\hat{b}_1$ dès que l'on a :

$$\text{var } \hat{a}_1 - \text{var } a_1^* > 0 \qquad \text{et} \qquad \text{var } \hat{b}_1 - \text{var } b_1^* > 0$$

c'est-à-dire, grâce aux relations (7) et (8)

$$\frac{2}{a_1} \text{cov } (\hat{\sigma}_2, \hat{\sigma}_1) - \text{var } \hat{\sigma}_1 > 0$$

et

$$\frac{2}{a_1} \text{ cov } \left(\frac{\hat{\mu}_2}{\mu_1} - \frac{\hat{\sigma}_2}{\sigma_1}, \frac{\hat{\mu}_1}{\mu_1} - \frac{\hat{\sigma}_1}{\sigma_1} \right) - \text{var } \left(\frac{\hat{\mu}_1}{\mu_1} - \frac{\hat{\sigma}_1}{\sigma_1} \right) > 0 \qquad (9)$$

Il est clair que la résolution de ces inéquations dépend de la forme des estimateurs $\hat{\mu}_1$, $\hat{\mu}_2$, $\hat{\sigma}_1$ et $\hat{\sigma}_2$ et qu'en outre, on ne peut les résoudre si l'on ne fait aucune hypothèse sur ceux-ci.

A cet effet, nous avons considéré successivement le cas où les estimateurs sont linéaires et celui où ceux-ci sont fournis par la méthode des moments.

Ces hypothèses se justifient par le fait que dans le cas où nous nous sommes placés on peut toujours construire des estimateurs linéaires dont l'efficacité est satisfaisante [1], tandis que dans le cas de distributions normales, la méthode des moments donne les estimations les plus efficaces.

1er cas :

Les estimateurs $\hat{\sigma}_1$, $\hat{\sigma}_2$, $\hat{\mu}_1$ et $\hat{\mu}_2$ sont des estimateurs linéaires.

Soient λ_i et ν_i, $i = 1, 2, \ldots, n$ les constantes qui définissent ces estimateurs linéaires et soient x_i, y_i, $i = 1, 2, \ldots, n$ les valeurs correspondantes observées des variables x et y.

On a donc :

$$\hat{\sigma}_1 = \Sigma \lambda_i x_i, \quad \hat{\sigma}_2 = \Sigma \lambda_i y_i, \quad \hat{\mu}_1 = \Sigma \nu_i x_i \qquad \text{et} \qquad \hat{\mu}_2 = \Sigma \nu_i y_i \tag{10}$$

Si l'on note par ailleurs qu'en conséquence du caractère aléatoire et simple des séries x_i et y_i on a :

$$\text{cov}\ (x_i, y_j) = 0 \qquad \text{pour} \qquad i \neq j,$$

et que σ_1 et σ_2 peuvent être de mêmes signes ou de signes contraires, il vient avec (10) et (1) :

$$\text{cov}\ (\hat{\sigma}_1, \hat{\sigma}_2) = \Sigma \lambda_i^2 \text{ cov } (x_i, y_i) = (\Sigma \lambda_i^2) \text{ cov } (x, y) = (\Sigma \lambda_i^2) |\rho| \sigma_1 \sigma_2 \text{ var } t \tag{11}$$

sachant que ρ désigne le coefficient de corrélation entre x et y. De même, on trouve :

$$\text{var } \hat{\sigma}_1 = (\Sigma \lambda_i^2) \sigma_1^2 \text{ var } t \tag{12}$$

Il s'ensuit que la première des relations (9) se réduit à la condition :

$$|\rho| > \frac{1}{2} \tag{13}$$

Cette condition est également celle à laquelle se réduit la seconde relation (9). Pour le voir, il suffit de remarquer que dans l'hypothèse d'estimateurs linéaires, on passe de la première relation (9) à la seconde en remplaçant les coefficients λ_i par les coefficients $(\nu_i/\mu_1 - \lambda_i/\sigma_1)$.(1)

2e cas :

Les variables x et y sont des variables normales. Dans ces conditions, on a les relations :

$$\hat{\mu}_1 = \Sigma x_i/n, \quad \hat{\mu}_2 = \Sigma y_i/n, \quad \hat{\sigma}_1^2 = \Sigma (x_i - \hat{\mu}_1)^2/n, \quad \hat{\sigma}_2^2 = \Sigma (y_i - \hat{\mu}_2)^2/n.$$

De plus, si l'on désigne encore par ρ le coefficient de corrélation entre x et y, on a aussi (cf. [2] pp. 208, 211 et 80) :

$$\text{cov}\ (\hat{\sigma}_1, \hat{\sigma}_2) = \rho^2 \sigma_1 \sigma_2/2n, \quad \text{cov}\ (\hat{\sigma}_2, \hat{\mu}_1) = \text{cov}\ (\hat{\sigma}_1, \hat{\mu}_2) = \text{cov}\ (\hat{\sigma}_1, \hat{\mu}_1) = 0$$

$$\text{var } \hat{\sigma}_1 = \sigma_1^2/2n, \quad \text{var } \hat{\mu}_1 = \sigma_1^2/n \quad \text{et} \quad \text{cov}\ (\hat{\mu}_1, \hat{\mu}_2) = |\rho| \sigma_1 \sigma_2/n \tag{14}$$

On en déduit que cette fois la première des relations (9) conduit à :

$$2\rho^2 - 1 > 0 \qquad \text{c'est-à-dire :} \qquad |\rho| > \sqrt{\frac{1}{2}} = 0,707\ldots \tag{15}$$

tandis que la seconde se ramène à :

$$\frac{2}{a_1}\left[\frac{1}{\mu_1^2} \text{cov}\ (\hat{\mu}_2, \hat{\mu}_1) + \frac{1}{\sigma_1^2} \text{cov}\ (\hat{\sigma}_2, \hat{\sigma}_1)\right] - \left[\frac{1}{\mu_1^2} \text{var } \hat{\mu}_1 + \frac{1}{\sigma_1^2} \text{var } \hat{\sigma}_1\right] > 0$$

---

(1) En toute généralité on devrait écrire $\hat{\sigma}_2 = \Sigma \lambda_i' y_i$ et $\hat{\mu}_2 = \Sigma \nu_i' y_i$ où les λ_i' et les ν_i' forment une même permutation des λ_i et des ν_i. Comme de plus, $\Sigma(\lambda_i - \lambda_i')^2 \geqslant 0$ entraîne $\Sigma \lambda_i^2 \geqslant \Sigma \lambda_i \lambda_i'$ et que les coefficients $(\nu_i/\mu_1 - \lambda_i/\sigma_1)$ vérifient une relation analogue, il s'ensuit que la condition (13) donne, en réalité, la borne inférieure des valeurs minimales de $|\rho|$.

En posant $k = \sigma_1^2/\mu_1^2$, avec (14), cette relation devient :

$$\rho^2 + 2k|\rho| - \left(k + \frac{1}{2}\right) > 0$$

c'est-à-dire :

$$|\rho| > \rho_o \qquad \text{avec} \qquad \rho_o = -k + \sqrt{k^2 + k + \frac{1}{2}} \tag{16}$$

Il apparaît ainsi que pour des variables normales, l'estimateur a_1^* doit être préféré à l'estimateur $\hat{a}_1$ dès que $|\rho| > \sqrt{\frac{1}{2}}$ tandis que le choix peut se porter sur l'estimateur b_1^* dès que $|\rho|$ dépasse la valeur critique ρ_o qui décroît régulièrement de $\sqrt{\frac{1}{2}}$ à $\frac{1}{2}$ lorsque le rapport k varie de zéro à l'infini.

4 - APPLICATION

Le problème particulier d'ajustement qui vient d'être considéré trouve de fréquentes applications en climatologie, notamment à l'occasion d'études à l'échelle régionale.

On sait, en effet, que les réseaux climatologiques au moyen desquels on recueille les informations nécessaires à la description du climat ont généralement des durées de fonctionnement très inégales. Plus précisément, si la création de stations centrales de météorologie remonte communément à un passé assez lointain, celle des stations régionales n'excède le plus souvent pas dix ou vingt ans et rarement trente ans.

Il en résulte que si, dans la plupart des cas, la distribution statistique des éléments climatologiques est bien connue pour la station centrale, par contre, cette connaissance est, apparemment du moins, limitée par la brièveté des séries d'observations pour les autres stations.

Cet état de choses n'est toutefois pas aussi grave qu'il semble à première vue et l'intérêt de l'étude théorique qui précède réside justement dans le fait qu'elle donne les conditions auxquelles des estimations plus précises pourront être obtenues.

L'existence même de ces conditions fournit d'autre part un moyen objectif d'appréciation de la qualité d'un réseau. Pour le montrer, nous avons considéré le cas des pointes maximales du vent au cours de chacun des mois de l'année en Belgique.

Leur étude montre, en effet, que le coefficient de corrélation entre la pointe maximale observée au cours d'un mois dans une station régionale donnée et la pointe maximale correspondante enregistrée à la station centrale d'Uccle varie d'une manière linéaire avec la distance qui sépare la station régionale de la station centrale.

De plus, si l'on exprime les coefficients ρ de corrélation en millièmes et les distances d en km, on trouve, en particulier, pour le mois de janvier :

$$\rho = 1\,000 - 1{,}42\,d \tag{17}$$

et pour le mois de juillet

$$\rho = 1\,000 - 5{,}63\,d \tag{18}$$

Par ailleurs, comme la distribution des pointes maximales du vent obéit à une loi doublement exponentielle, on sait (cf. [3] et [4]) que, lors de l'ajustement, il convient d'utiliser des estimateurs linéaires des paramètres de la loi de probabilité.

Il s'ensuit que les estimateurs a_1^* et b_1^* pourront être utilisés dès que $\rho > \frac{1}{2}$, c'est-à-dire, en vertu des relations (17) et (18), pour les stations régionales situées à des distances de la station centrale inférieures à 352 km pour le mois de janvier et à 89 km pour le mois de juillet.

Si l'on note alors que le réseau anémométrique du pays se décompose en triangles dont les côtés ont des longueurs variant de 30 km à 140 km, et qu'en outre, la distance maximale des stations régionales à la station centrale atteint 161 km, on en déduit que pour l'hiver la connaissance de la distribution des pointes maximales du vent à la station centrale est effectivement utile à l'étude statistique des pointes maximales du vent à l'échelle de tout le pays, mais que pour l'été, cette utilité n'intéresse qu'une région beaucoup plus restreinte.

5 - CONCLUSIONS

En résumé, on a vu que l'existence d'une corrélation entre deux séries d'observations ne permet pas nécessairement d'obtenir des déterminations plus efficaces de la loi de probabilité à laquelle l'une d'elle est soumise lorsqu'on connaît l'autre ; encore faut-il que cette corrélation atteigne un seuil minimal qui dépend de la forme des estimateurs des paramètres de la loi de probabilité en cause.

Du point de vue pratique, nous croyons que cette propriété justifie l'introduction de la notion d'indépendance climatologique que l'on réservera aux séries d'observations pour lesquelles le coefficient de corrélation est inférieur au seuil critique tel qu'il vient d'être défini.

Pour le reste, l'exemple choisi montre toute l'importance que cette notion revêt lors de l'exploitation des séries d'observations fournies par un réseau climatologique donné.

BIBLIOGRAPHIE

[1] BLOM G. (1958) - Statistical estimates and transformed beta-variables, John Wiley and Sons, Inc., New-York.

[2] KENDALL M. G. (1947) - The Advanced Theory of Statistics, Vol. I, London, Charles Griffin et C°.

[3] LIEBLEIN J. (1954) - A New Method of Analyzing Extreme-value Data, N. A. C. A., Technical note 3053, Washington, D. C.

[4] KIMBALL, F. B. (1956) - The bias in certain estimates of the parameters of the extreme- value distribution, Ann. Math. Statis., 27, p. 758-767.