REVUE DE STATISTIQUE APPLIQUÉE

R. GIRAUD
P. THIONET

Simulation par aléatoires de l'indice de concordance de Theil

Revue de statistique appliquée, tome 16, nº 1 (1968), p. 59-75
<http://www.numdam.org/item?id=RSA_1968__16_1_59_0>

SIMULATION PAR ALÉATOIRES DE L'INDICE DE CONCORDANCE DE THEIL

R. GIRAUD et P. THIONET

L'objet de la présente note est de condenser le mémoire écrit par M. Giraud en vue de l'obtention d'un Diplôme d'Etudes Supérieures, sur un sujet que nous lui avions proposé. Il s'agissait de poursuivre l'étude de l'indice Q de Theil destiné à comparer des prévisions et leurs réalisations, indice présenté par nous-même dans le Journal de la Société de Statistique de Paris - 1964 (3ème trimestre, p. 181-9) [1, 2].

C'est le Professeur FRECHET qui nous donna l'impulsion au départ, en nous engageant à faire poursuivre cette étude par un de nos chercheurs. Toutefois M. FRECHET pensait plutôt à des calculs de Q effectués sur des séries numériques réelles. Nous avons pensé qu'il serait très difficile de nous procurer de telles données, où la valeur prévue, présentée sans fard à côté de la valeur finalement retenue pour la réalisation, tend à prouver qu'on fait (de temps à autre) des prévisions hasardeuses. Il n'était d'ailleurs pas question de faire rechercher de telles données par un étudiant.

En simulant les erreurs de prévisions $P_i - R_i$ *par des variables aléatoires, nous rentrions au contraire dans le domaine des recherches d'ordre universitaire et c'est ce qui fut fait. Pour essayer d'être plus clair, nous présentons dans une première partie les questions de statistique mathématique étudiées par M. Giraud en vue de la dite simulation. Dans la deuxième partie, on appliquera ces résultats au problème des erreurs de prévision.*

Bien entendu on n'a répondu ainsi que partiellement au but initialement proposé, l'assimilation des erreurs de prévision à des variables aléatoires étant d'autant plus discutable qu'on a été conduit à particulariser énormément le type de variables. C'est dire qu'il y a place pour d'autres recherches, apparemment plus difficiles.

Ajoutons qu'il a aussi été question de l'indice de Theil dans une communication au Congrès Européen d'Econométrie de 1967 (Bonn), pour comparer à des réalisations les résultats de certains modèles prévisionnels. Il semble donc que cet indice intéresse diverses classes de chercheurs [3].

P. T.

1ère PARTIE - ETUDE D'UNE VARIABLE ALEATOIRE DETERMINEE PAR UN χ^2 CENTRE ET PAR UN χ^2 DECENTRE

La variable Q qui va être étudiée mérite d'être rapprochée d'une autre variable Q de HOGBEN, PINKHAM et WILK, qui a fait l'objet de deux articles des Annals of Mathematical Statistics, 1964, [4]. Cette dernière variable Q est de la forme

$$W/\sqrt{W^2 + \chi^2}$$

le χ^2 ayant N degrés de liberté et W étant une variable de Laplace Gauss $\mathcal{N}(\theta, 1)$. $\chi^2 + W^2$ est donc un χ_c^2 décentré, de même que W^2

Notre variable Q est de la forme

$$Q = \chi/(a + \chi_c) \tag{1}$$

χ_c^2 étant un χ^2 décentré ; et comme pour la variable des Annals, les numérateur et dénominateur ne sont pas indépendants, admettant un "facteur commun" (au sens de l'analyse factorielle) :

$$\chi^2 = \xi_1^2 + \sum_2^N \xi_i^2 \qquad ; \qquad \chi_1^2 = (a + \xi_1)^2 + \sum_2^N \xi_i^2 \tag{2}$$

L'analogie s'arrête là, car nous ne donnons aucun moyen perfectionné de tabuler la variable Q (ce qui pourrait faire l'objet d'autres recherches).

Expressions diverses de Q

1/ La variable Q se présente comme fonction des N aléatoires X_i de Laplace Gauss normées

$$X_i : \mathcal{N}(0, 1) \qquad ; \qquad i = 1, 2, \ldots, N$$

$$Q = \frac{\sqrt{\sum X_i^2}}{a + \sqrt{\sum (a_i + X_i)^2}} \tag{3}$$

les a_i étant N paramètres, avec $\sum a_i^2 = a^2$.

Posant $\sum X_i^2 = \chi^2$, et substituant aux X_i les variables ξ_i, on peut donc ne faire dépendre Q que de 2 variables aléatoires liées χ, χ_c ou de 2 variables indépendantes

$$\xi_1 \qquad \text{et} \qquad \chi'^2 = \sum_2^N \xi_i^2$$

Cette dernière formulation peut céder la place à la suivante.

2/ Posant $\sum_1^N a_i X_i = a\chi Y$, on peut vérifier que Y et χ sont indépendantes. On a

$$Q = \frac{\chi}{a + \sqrt{a^2 + 2a\chi Y + \chi^2}} \tag{4}$$

avec $Y = \cos \Omega$

en désignant par Ω l'angle des vecteurs (a_i) et (X_i). La distribution de Y est bien connue, c'est celle du coefficient de corrélation statistique R de n couples $(x_i y_i)$ d'observations gaussiennes de variables x, y non corrélées ; elle est liée à celle du t de Student-Fisher ; Y^2 suit une "loi Beta". Le calcul donne :

$$E(Y^2) = \frac{1}{N} \quad ; \qquad V(Y^2) = \frac{2(N-1)}{N^2(N+2)}$$

3/ On sait que $E\chi^2 = N$, $V\chi^2 = 2N$. Considérons donc

$$\sqrt{a^2 + 2a\chi Y + \chi^2}.$$

Dès que N est "un peu grand", le terme χY est négligeable vis à vis des termes a^2 et χ^2. Du moins est-ce vrai en probabilité, puisque χY et χ^2 sont deux aléatoires (si a^2 est une constante).

Ceci conduit à assimiler (sans doute un peu vite) la variable Q et la variable suivante :

$$Q_1 = \frac{\chi}{a + \sqrt{a^2 + \chi^2}} \tag{5}$$

lorsque N est grand, pratiquement pour $N > 20$. L'avantage de Q_1 est de ne plus dépendre que de l'aléatoire unique χ.

Tables de la variable Q_1

La graduation de Q_1, c'est-à-dire l'établissement de ses tables, est facile, du fait que les fonctions

$$q = \frac{x}{a + \sqrt{a^2 + x^2}} \Longleftrightarrow x = \frac{2\,aq}{1 - q^2} \tag{5'}$$

$$(x \neq 0)$$

qui sont réciproques l'une de l'autre, sont monotones (croissantes). On ne s'intéresse qu'aux valeurs positives de x (de χ), ce qui laisse varier q de 0 à 1 ; et on a

$$\text{Prob}\,(Q_1 < q) = \text{Prob}\,(\chi < x)$$

La table reproduite ici correspond à $N = 25$, le paramètre a variant de 30 à 500, les seuils de probabilité retenus pour les besoins du statisticien étant :

$$\text{Prob}\,(Q_1 > q) = 10\,\% \qquad 5\,\% \qquad 2\,\% \qquad 1\,\%$$

L'intérêt de Q_1 substituée à Q (dans des conditions qui seront désignées par H.S. initiales d'Hypothèse Simplificatrice) est qu'on rencontre des difficultés pratiques à calculer les tables de la distribution de la variable Q proprement dite.

Tableau 1 - Valeur que Q_1 a la probabilité 1 - P de dépasser. Pour N = 25

a	1 - P = 0, 10	0,05	0,02	0,01
30	0, 096	0, 101	0, 106	0, 109
35	0, 083	0, 086	0, 091	0, 094
40	0, 072	0, 076	0, 080	0, 082
45	0, 064	0, 067	0, 071	0, 073
50	0, 058	0, 061	0, 064	0, 066
55	0, 053	0, 055	0, 058	0, 060
60	0, 048	0, 051	0, 053	0, 055
65	0, 045	0, 047	0, 049	0, 051
70	0, 041	0, 043	0, 045	0, 047
75	0, 039	0, 040	0, 042	0, 044
80	0, 036	0, 038	0, 040	0, 041
85	0, 034	0, 036	0, 037	0, 039
90	0, 032	0, 034	0, 035	0, 036
95	0, 030	0, 032	0, 033	0, 034
100	0, 029	0, 030	0, 032	0, 033
105	0, 027	0, 029	0, 030	0, 031
110	0, 026	0, 027	0, 029	0, 030
115	0, 025	0, 026	0, 028	0, 029
120	0, 024	0, 025	0, 026	0, 027
125	0, 023	0, 024	0, 025	0, 026
130	0, 022	0, 023	0, 024	0, 025
135	0, 021	0, 022	0, 023	0, 024
140	0, 020	0, 021	0, 023	0, 023
145	0, 020	0, 021	0, 022	0, 022
150	0, 019	0, 020	0, 021	0, 022
160	0, 018	0, 019	0, 020	0, 020
170	0, 017	0, 018	0, 018	0, 019
180	0, 016	0, 017	0, 017	0, 018
190	0, 015	0, 016	0, 016	0, 017
200	0, 014	0, 015	0, 016	0, 016
210	0, 013	0, 014	0, 014	0, 015
220	0, 013	0, 013	0, 014	0, 015
230	0, 012	0, 013	0, 014	0, 014
240	0, 012	0, 012	0, 013	0, 013
250	0, 011	0, 012	0, 012	0, 013
275	0, 010	0, 011	0, 011	0, 012
300	0, 009	0, 010	0, 010	0, 011
350	0, 008	0, 008	0, 009	0, 009
400	0, 007	0, 007	0, 008	0, 008
450	0, 006	0, 006	0, 007	0, 007
500	0, 005	0, 006	0, 006	0, 006

Deux représentations géométriques de Q et techniques de calcul numérique

1/ Coordonnées bipolaires et polaires

L'expression (1) $Q = \chi/(a + \chi_c)$ invite à représenter en coordonnées bipolaires les courbes $Q = q$; il vient :

$$q = \rho/(a + \rho_1) \qquad (6)$$

ou $\qquad \rho_1 = p\rho - a \qquad$ avec $\qquad p = \frac{1}{q}$

C'est une famille d'ovales, de paramètre p (ou 1/q).

On passe de là à une représentation polaire, avec

$$\rho_1^2 = \rho^2 + 2a\rho \sin \theta + a^2$$

$$\boxed{\rho = \frac{2a}{p^2 - 1} (p + \sin \theta)} \tag{7}$$

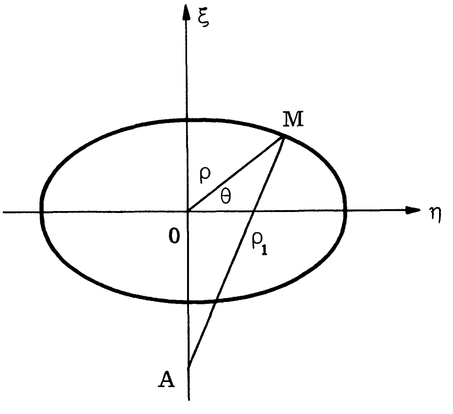

Fig. 1

<u>α) Cas de N = 2</u>

Le point $M(\rho, \theta)$ a une distribution de probabilité, dans le plan (ρ, θ) telle que :

$\rho^2 = OM^2$ est la variable χ^2, à $N = 2$ degrés de liberté

θ est uniformément répartie sur $(0, 2\pi)$ (isotropie)

de sorte que

$$OM^2 = \xi^2 + \eta^2 = \chi^2$$

$$AM^2 = (\xi + a)^2 + \eta^2 = \chi_c^2$$

Les variables ξ et η sont deux variables de Laplace Gauss normées indépendantes.

On voit ainsi que, <u>si N = 2</u>, la distribution de M dans son plan est schématisée par des angles égaux issus de 0 et des cercles concentriques de centre 0.

La distribution de Q est schématisée par une famille d'ovales. Le domaine situé à l'intérieur de l'ovale $(1/q = p)$ a pour mesure la probabilité que Q soit inférieur à q.

Chaque ovale est en fait inclus entre 2 cercles concentriques voisins l'un de l'autre, de rayons χ' et χ''.

En interpolant, il a été possible, graphiquement, d'en déduire (grossièrement), les valeurs suivantes de la probabilité.

Tableau 2

q =	0,08	0,07	0,06	0,05	0,04	0,03
Prob $(Q \leqslant q)$ =	99 %	98 %	93 %	85 %	70 %	50 %

β) Cas de $N > 2$: La modification apportée au cas précédent est apparemment minime.

$$\rho^2 = OM^2 \text{ est toujours un } \chi^2 \text{(à N degrés de liberté)}$$

$\rho^2 = \xi^2 + \eta^2$; ξ est encore une variable $\mathcal{N}(0, 1)$, indépendantes de η mais : η^2 est un χ^2 à $N - 1$ degrés de liberté.

Conséquence : La distribution est anisotrope : θ n'est plus du tout uniforme, si les cercles de rayon ρ délimitent encore des domaines dont la mesure est Prob$(\chi^2 < \rho^2)$.

De même il est toujours exact que le domaine situé à l'intérieur de l'ovale de paramètre $p = 1/q$ a pour mesure la probabilité de $(Q < q)$.

L'ovale d'équation (7) : $\rho = 2a(p + \sin\theta)/(p^2 + 1)$ a pour grand axe $2a/(p - 1)$ pour $\theta = \pi/2$; et pour petit axe $2a/(p + 1)$ pour $\theta = -\pi/2$.

La moyenne des axes est $2ap/(p^2 - 1)$, c'est-à-dire la valeur de ρ pour $\theta = 0$.

Le cercle de rayon $\rho = 2ap/(p^2 - 1)$, c'est-à-dire $p = \dfrac{a + \sqrt{a^2 + \rho^2}}{\rho}$ correspond justement à :

$$q = \frac{1}{p} = \frac{\rho}{a + \sqrt{a^2 + \rho^2}}$$

et la mesure de l'intérieur de ce cercle n'est pas autre que la probabilité de $(Q_1 < q)$.

Ainsi Q est très voisin de Q_1 parce que le cercle de rayon $2ap/(p^2 - 1)$ chevauche sur l'ovale, les rayons vecteurs plus grands que $2ap/(p^2 - 1)$ et ceux plus petits se compensant quel que soit N.

Exemple. - $a = 30$

				N = 10		N = 5		N = 2	
q	$\frac{2a}{(p^2+1)}$	$\frac{2ap}{(p-1)}$	$\frac{2a}{(p-1)}$	χ_{10}	Proba	χ_5	Proba	χ_2	Proba
0,08	4,44	4,83	5,19	4,81	0,99				
0,07	3,90	4,21	4,50	4,27	0,95	3,88	0,99		
0,06	3,39	3,60	3,81	3,66	0,80	3,65	0,98		
0,05	2,85	3,00	3,15	3,05	0,50	3,03	0,95	3,03	0,99
								2,88	0,98
0,04	2,31	2,40	2,50					2,49	0,95
								2,14	0,90

Conclusion : Aucune interpolation ne donnerait de résultats différents de Q_1, vu l'imprécision du procédé.

Restent possibles des calculs précis d'intégrales doubles, qu'on pourrait effectuer sur ordinateur.

2/ Autre représentation graphique.

Partons de :

$$Q = \frac{\chi}{a + \sqrt{a^2 + 2a\chi Y + \chi^2}} \tag{4}$$

Posant $Q < q$, $q = \frac{1}{p}$, l'équation

$$p = \frac{a + \sqrt{a^2 + 2axy + x^2}}{x} \tag{4'}$$

se réduit à

$$px - a = \sqrt{a^2 + 2axy + x^2}$$

ou

$$x(p^2x - 2ap - 2ay - x) = 0$$

Elle se décompose en une droite :

$$x = 0$$

et une famille Dp de droites :

$$p^2x - 2ap - (2ay + x) = 0 \tag{8}$$

Les Dp ont pour enveloppe l'hyperbole H, d'équation

$$a^2 + x(2ay + x) = 0 \tag{9}$$

On pourrait vérifier que deux droites quelconques de la famille Dp' et Dp'' se coupent bien hors de la demi-bande B, définie comme suit :

$$B : (x > 0 \quad ; \quad -1 < y < +1) \qquad \text{(cf fig. 2)}$$

Considérons alors sur B la distribution en (x, y), produit des distributions

| de χ_N sur le demi-axe positif Ox ;

| de Y, corrélation statistique de Fisher, sur l'intervalle (- 1, + 1) de Oy.

La droite Dp (de paramètre p) partage B en deux domaines, dont les mesures respectives sont :

Prob(Q < q) à gauche de Dp (région en forme de trapèze rectangle)

Prob(Q > q) à droite de Dp (région non bornée à droite).

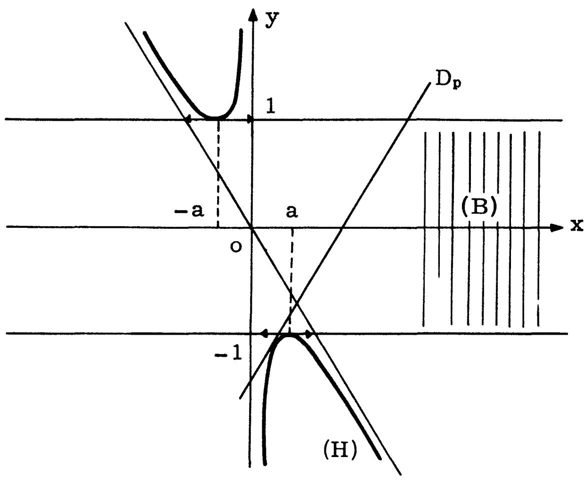

Fig. 2

En résumé

On pourra calculer Prob(Q < q) = F(q) en intégrant sur le trapèze rectangle l'élément différentiel, produit des probabilités élémentaires de X et de Y.

Ces calculs n'ayant pas été faits, il est difficile de dire s'ils sont plus simples que ceux relatifs au premier cas. A priori, il semble devoir être plus pratique d'intégrer sur une aire limitée par un segment de droite, que sur une aire limitée par un ovale.

Remarque 1. Cas de N = 1 : On obtient alors (figure 3)

$$q = \frac{|u|}{a + |a + u|}$$

$$\text{Prob}(Q < q) = \text{Prob}(u' < U < u'')$$

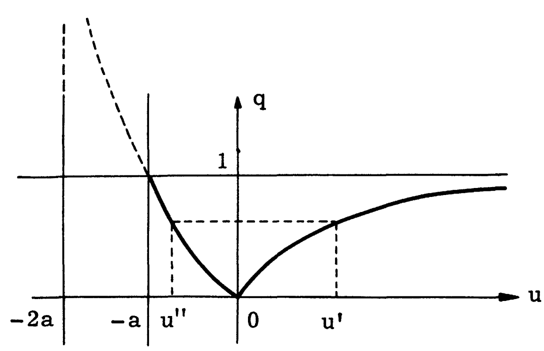

Fig. 3

Le cas est d'ailleurs sans aucun intérêt pratique.

Remarque 2. Cas d'une variable plus générale, du type

$$Q_2 = \frac{X}{a + \sqrt{b^2 + X^2}} \quad ; \qquad \text{avec : X définie sur} \qquad (0, +\infty).$$

où b n'est plus identique à a, et où X^2 n'est plus nécessairement un χ^2.

On notera d'abord que Q_2 est encore définie sur (0, 1), et que la fonction q(x)

$$q = x\,[a + \sqrt{b^2 + x^2}]^{-1}$$

reste monotone croissante ; d'où il suit qu'on a encore :

$$\text{Prob}(Q_2 < q) = \text{Prob}(X < x)$$

Ce fait présentera quelque intérêt dans la 2ème partie de cet exposé (modèle N° 4).

Remarque 3. Cas où N est très grand

Quand N est grand, les variables Q et Q_1 sont indiscernables. Mais l'une et l'autre sont asymptotiquement des variables de Laplace-Gauss.

$$Q \sim \mathcal{N}\left(\frac{\sqrt{N}}{2a} \ , \ \frac{1}{8a^2}\right)$$

Ceci peut s'établir par développement limité de

$$Q_1 = \frac{\chi}{a}\left[1 + \left(1 + \frac{\chi^2}{a^2}\right)^{1/2}\right]^{-1};$$

ce qui donne

$$Q \# \frac{\sqrt{N}}{2a}\left(1 - \frac{N}{4a^2}\right) + \frac{1}{2\sqrt{2a}}\left(1 - \frac{3N}{4a^2}\right)\xi \quad ; \quad \text{avec} \quad \xi \ : \ \mathcal{N}(0, 1).$$

On suppose enfin N/a^2 petit pour obtenir le résultat annoncé.

2ème PARTIE - APPLICATION A LA MISE A L'EPREUVE DE L'INDICE DE THEIL

1 - Rappels

Soit N grandeurs économiques G_i, i = 1, 2, ..., N, dont les valeurs à une date donnée font l'objet de prévisions

$$P_1, P_2, \ldots P_N$$

et qui prendront en fait les valeurs

$$R_1, R_2, \ldots R_N$$

Il se peut agir également d'une variable économique G, de ses valeurs prévues et ensuite observées, à des dates successives i = 1, 2, ... N.

L'indice étudié est :

$$Q = \frac{\sqrt{\sum (P_i - R_i)^2}}{\sqrt{\sum R_i^2 + \sum P_i^2}} \qquad \ldots \qquad (1)$$

Il est facile de voir que :

1/ $0 \leqslant Q \leqslant 1$

2/ $Q = 0$ si et seulement si $P_i = R_i$ quel que soit i ;

3/ $Q = 1$ dans les cas suivants :

3.1 - $P_i = 0$ pour tout i, quels que soient les R_i,

3.2 - $R_i = 0$ pour tout i, quels que soient les P_i,

3.3 - $P_i = - \lambda R_i$. $\lambda > 0$ pour tout i.

2 - Nature de l'étude

Dans [2] on avait considéré l'indice Q comme une fonction des 2n variables P_i, R_i. En fait les $(P_i - R_i)$ sont les erreurs commises sur les P_i par un service chargé de la prévision des R_i.

Nous assimilerons ces erreurs à certaines variables aléatoires, dont tout d'abord la loi n'est pas spécifiée mais dont on peut admettre qu'elles ont des moments d'ordre 1 et 2. Après quoi nous introduirons des hypothèses supplémentaires plus ou moins fortes.

Pour un corps d'hypothèses spécifié, l'indice Q étant devenu une variable aléatoire, nous nous proposons de calculer les valeurs q_α telles que

$$\text{Prob}[Q > q_\alpha] = \alpha$$

avec α = 10 %, 5 %, 2 %, 1 %, que les statisticiens appellent seuils de signification.

Cette étude nous donnera ainsi une idée de l'importance qu'il convient d'attribuer à telle ou telle valeur de l'indice Q, lorsqu'on le calcule sur des séries statistiques réelles.

Remarque - Nous considérerons les R_i comme des constantes ; c'est-à-dire que nous admettrons que les réalisations ne sont pas affectées par les prévisions qui ont été faites, ce qui est commode pour les calculs.

Le contraire est souvent vrai ; par exemple (phénomène de "feed back") on sait que des sondages, à la veille des élections, peuvent avoir une certaine influence sur la façon de voter.

Lorsque le calcul de Q porte, pour une grandeur économique, sur des prévisions

$$P_1, P_2, \dots, P_N$$

et des réalisations

$$R_1, R_2, \dots, R_N$$

à des dates

$$t = 1, 2, \dots, N \qquad \text{(disons également espacées)}$$

il pourrait même être nécessaire de tenir compte de modifications en chaîne :

Ainsi, un retard observé, à la date t, dans la réalisation R_t, peut avoir une influence sur R_{t+1} ou R_{t+2}.

Pour simplifier, nous excluons ces problèmes de la présente étude.

Hypothèse Simplificatrice (passage de Q à Q_1)

Nous aurons souvent recours à une hypothèse très commode (désignée par H. S) qui ne saurait d'ailleurs être rigoureusement vérifiée : nous supposerons négligeable le terme

$$\sum_i R_i (P_i - R_i)$$

à côté de $\sum R_i^2$ d'une part et $\sum (P_i - R_i)^2$ d'autre part, dans

$$\sum P_i^2 = \sum R_i^2 + 2 \sum R_i (P_i - R_i) + \sum (P_i - R_i)^2 .$$

Or, la plupart du temps nous devons aussi supposer les $(P_i - R_i)$ indépendantes, ce qui exclut l'existence d'une condition telle que :

$$\sum R_i (P_i - R_i) = 0$$

Mais quand les $(P_i - R_i)$ seront assimilées à des aléatoires d'espérance mathématique nulle, l'hypothèse H. S est admissible si les $(P_i - R_i)$ admettent un moment d'ordre 2, et d'autant mieux que N sera grand.

Quand (modèle N° 4) nous supposerons $E(P_i - R_i) = \delta_i$ non nulle, notre hypothèse H. S impliquera au contraire $\sum R_i \, \delta_i \# 0$, c'est-à-dire une orthogonalité entre vecteurs (R_i) et (δ_i) qui appelle beaucoup de réserves. Toutefois il nous suffira (Remarque 2, fin de la 1ère partie) de supposer (ce qui est bien moins strict) :

$$\sum R_i (P_i - R_i - \delta_i) \# 0$$

et d'écrire

$$\sum P_i^2 = \sum (R_i^2 + 2 R_i \delta_i) + 2 \sum R_i (P_i - R_i - \delta_i) + \sum (P_i - R_i)^2$$

avec

$$\sum R_i^2 = a^2 \sigma^2 \quad ; \quad \sum R_i^2 + 2 R_i \delta_i = b^2 \sigma^2 \quad ; \quad \sum (P_i - R_i)^2 = \sigma^2 \chi_c^2$$

d'où

$$Q \# Q_2 = \frac{\chi_c}{a + \sqrt{b^2 + \chi_c^2}}$$

Le rôle de H. S est de limiter nos calculs, en assimilant Q à Q_1 aléatoire fonction d'une seule variable aléatoire (en pratique χ^2 ou χ_c^2) et non de deux aléatoires.

Il existe vraisemblablement d'autres approximations visant au même but et (le cas échéant) meilleures.

Les modèles d'erreurs

Modèle N° 1 : $P_i - R_i = X_i = \sigma_i u_i$

U_i serait une valeur de la variable aléatoire $\mathcal{N}(0, 1)$ de Laplace-Gauss réduite. Les U_i sont mutuellement indépendantes (en probabilité). $\sigma_i = \sigma$ est constante.

Alors, en posant

$$\sum_1 R_i^2/N = r^2$$

$$r\sqrt{N}/\sigma = a$$

l'indice Q de THEIL s'identifie à la variable aléatoire Q de la 1ère Partie.

Et si N est assez grand, la distribution de Q diffère peu de celle de Q_1 (Tableau 1 ci-dessus).

Discussion du modèle N° 1

Le modèle n° 1 d'erreurs n'est pas réaliste.

1/ Nous avons supposé que σ était le même pour tous les R_i, alors que les postes i (disons d'une comptabilité nationale) correspondant aux R_i sont très différents. On pourrait alors plutôt supposer σ_i proportionnel à R_i ou (mieux) à $\sqrt{R_i}$.

2/ L'indépendance mutuelle des U_i n'est pas réaliste : c'est ainsi que les postes (i) des comptes nationaux ne sont pas indépendants entre eux. Des relations linéaires, dites identités comptables, lient certains d'entre eux. Dès lors $\sum X_i^2 = \sigma^2\chi^2$ n'a plus N degrés de liberté, mais un nombre de degrés moindre. En moyenne $\sum X_i^2$ est moins grand que prévu. .

Exemple : Si $R_1 + R_2 + R_3 = R_4 + R_5$

et $P_1 + P_2 + P_3 = P_4 + P_5$,

alors $X_1 + X_2 + X_3 = X_4 + X_5$;

Posons

$$\sum_1^5 X_i^2 = \sigma^2\chi^2 \ :$$

χ^2 est alors une variable à 4 degrés de liberté (et non 5).

Conséquence : Raisonnons en faisant l'hypothèse H.S. Il revient alors au même de dire que χ^2 ou que Q est plus petit que prévu, parce qu'on oublie l'existence des identités comptables. L'indice Q risquerait donc de paraître non significatif, alors qu'en fait il serait déjà significatif, c'est-à-dire décelerait une baisse de qualité de la prospective.

3/ L'hypothèse $E(U_i) = 0$ traduit l'absence d'erreurs systématiques de la part du service chargé des prévisions. Or de telles erreurs existent en fait et sont longues à se laisser découvrir. On doit se demander quelle serait l'influence de ces erreurs s'ajoutant aux erreurs aléatoires.

Conclusion : Aucune de ces critiques n'est vraiment destructrice mais elles conduisent à certaines complications du modèle d'erreurs, qui vont être à présent formulées. En revanche on ne peut oublier une autre catégorie de critiques, celles concernant le bien-fondé de toute l'hypothèse gaussienne ; disons que d'autres lois de probabilités pourraient bien être plus près de la vérité mais présenteraient l'inconvénient majeur de conduire à des calculs inextricables ; dès lors il conviendrait de s'abandonner aux calculateurs électroniques, ce qui change totalement la nature des recherches (et des critiques à leur opposer). Tenons-nous en donc aux modèles gaussiens d'erreurs.

Modèle N° 2 : non indépendance des U_i

On peut introduire dans les calculs précédents les covariances à côté des variances. Posons $X^2 = \Sigma(P_i - R_i)^2$.

Soit ρ_{ij} la corrélation entre U_i et U_j. Si N est assez grand, Q est à peu près gaussien. On n'a pas altéré $E(X^2)$; mais le calcul montre que $V(X^2)$ est multiplié par

$$1 + (N - 1)\overline{\rho}$$

$\overline{\rho}$ étant la moyenne des ρ_{ij}. Il suffirait donc d'avoir $\overline{\rho} = 0,1$ pour que la variance se trouve multipliée par 3 (si N = 21 par exemple). Il est vrai que, dans le cas des identités comptables, il faudrait tenir les corrélations pour négatives (ce qui abaisse la variance).

On peut encore imaginer que ρ_{ij} est en général nulle, sauf si $j = i \pm 1$, c'est-à-dire pour les termes consécutifs d'une série (chronologique surtout). La variance serait alors multipliée par

$$1 + 2\left(\frac{N - 1}{N}\right)\hat{\rho}$$

$\hat{\rho}$ désignant la moyenne des corrélations non nulles. Avec $\hat{\rho} = 0,25$, la variance sera accrue de 50 % par rapport au modèle N° 1 qui suppose l'indépendance.

Modèle N° 3

Si l'on ne suppose plus les σ_i égaux entre eux, alors

$$\sum (P_i - R_i)^2 = \sum \sigma_i^2 U_i^2 = X^2$$

n'est pas une variable χ^2 de Pearson. On trouve facilement

$$EX^2 = \sum \sigma_i^2 = N\mu_1$$

$$VX^2 = 2\sum \sigma_i^4 = 2N\mu_2$$

en appelant μ_1 μ_2 les deux premiers moments des σ_i^2 :

$$\mu_2 = \mu_1^2 + V(\sigma_i^2)$$

Supposant que X^2 ait la même moyenne qu'avec le modèle N° 1, la dispersion des σ_i^2 a pour conséquence un accroissement de VX^2.

Raisonnons en faisant l'hypothèse H. S : l'indice Q aura même valeur centrale que dans le modèle N° 1 ; mais sa dispersion autour de cette valeur centrale sera d'autant plus accrue que les σ_i seront plus différents les uns des autres.

Modèle N° 4

Introduisons dans le modèle N° 1 des erreurs systématiques δ_i, c'est-à-dire imaginons qu'on ait :

$$P_i = R_i + \delta_i + \delta U_i \quad , \quad \delta_i \neq 0$$

L'indice Q pourra se mettre sous la forme

$$Q = \frac{\chi_c}{a + \chi'_c}$$

où χ_c et χ'_c sont deux χ décentrés non indépendants, à paramètres de décentrages différents. Raisonnant encore dans l'hypothèse H. S, on aurait

$$Q \# Q_1 = \frac{\chi_c}{a + \sqrt{a^2 + \chi_c^2}}$$

le χ_c décentré se substituant au χ de Karl Pearson dans la formule (5).

Mais nous devons faire des réserves sur l'hypothèse H. S qui en pareil cas suppose :

$$\sum R_i \delta_i = 0$$

Si l'on se borne à tenir pour négligeable $\sum (R_i + \delta_i)U_i$, ce qui paraît raisonnable, il vient

$$Q \# Q_2 = \frac{\chi_c}{a + \sqrt{b^2 + \chi_c^2}}$$

où b^2 diffère de a^2 : $a^2 = \sum R_i^2/N\sigma^2$; $b^2 = \sum (R_i + \delta_i)^2/N\sigma^2$

D'après la remarque 2 (fin de la 1ère partie), on n'en aura pas moins

$$\text{Prob}(Q_2 < q) = \text{Prob}(\chi_c < x)$$

Pour passer aux applications, encore faudrait-il connaître le paramètre de décentrage de χ_c^2, ce qui suppose qu'on soit déjà très bien informé sur les erreurs systématiques. Ainsi on ne voit guère le moyen d'utiliser la table χ_c^2. Du point de vue qualitatif, le décentrage du χ^2 a pour effet d'augmenter (très notablement) les valeurs significatives de la variable, donc de Q. C'est dire que Q peut paraître significativement trop grand, alors qu'il ne le serait pas.

Application à l'étude d'une série numérique double (P_i , R_i) $i = 1, 2, \ldots, N$

On a pu remarquer que tous les résultats obtenus (le tableau 1 notamment) dépendaient non seulement de N (taille de la série) mais en plus encore du paramètre a.

Dans le cas du Modèle 1, on a par hypothèse

$$a^2 = \frac{\sum R_i^2}{\sigma^2}$$

1/ Il est possible qu'on veuille tester avec Q une hypothèse où la valeur de a est spécifiée, auquel cas il n'y a pas de problème.

2/ Il est aussi possible d'estimer a sur une série numérique, l'hypothèse (disons celle du modèle 1) étant supposée correcte. Alors l'estimation de σ^2 :

$$\hat{\sigma}^2 = \sum (P_i - R_i)^2/N$$

est suffisante et sans biais. L'estimation de $a^2 = \left(\sum R_i^2\right)/\sigma^2$, où $\left(\sum R_i^2\right)$ n'est jamais qu'une constante (d'après nos hypothèses initiales), pourrait donc être :

$$\hat{a}^2 = \left(\sum R_i^2\right)/\hat{\sigma}^2$$

$$= N\left(\sum R_i^2\right)/\sum (P_i - R_i)^2$$

Cette estimation est suffisante, elle est malheureusement biaisée (pour a^2 et a).

En pratique, a sera estimé sur un nombre N de couples relativement médiocre et l'imprécision de cette estimation restera grande.

3/ Enfin on serait satisfait s'il était possible tout à la fois de tester l'hypothèse et d'estimer a, c'est-à-dire de juger à première vue si un indice Q est fort ou est faible. Or ceci n'a pas paru de l'ordre des choses possibles :

Si l'on n'a aucune idée, à l'avance, sur la valeur de a, on ne peut utiliser le tableau 1.

Plus exactement :

Ayant estimé a par $\hat{a}$, la valeur de Q, lue sur la ligne $\hat{a}$ du tableau 1, ne devrait pas être (en général) une valeur significative.

Ici, il faut noter que nous avons rencontré le cas (expérimentalement) une fois : il se trouvait alors que le modèle d'erreurs (modèle N° 1) était totalement irréaliste, en ce sens que toutes les erreurs $P_i - R_i$ étaient grandes en module, ce qui ne peut guère se produire pour une variable de Laplace Gauss U_i. On peut penser que Q ne sera significatif que dans des cas comparables à celui-ci.

4/ Il convient donc de poser le problème d'une autre façon :

A la suite de nombreux essais de prévision, et des confrontations faites plus tard sur les réalisations quand on les connaît, un service de prospective (ou de prévision, ou de conjoncture) est supposé avoir acquis une technique telle que ses prévisions soient homogènes, - autrement dit que l'écart-type σ - et finalement le paramètre "a" - relatifs à tel ou tel groupe de rubriques, sont censés rester stables au cours des années.

Bien entendu, "a" sera "meilleur" pour les groupes de rubriques où la prévision est la plus aisée (disons le groupe des salaires, en comptabilité nationale) ; "a" sera "mauvais" pour les groupes de rubriques dont la valeur, voire même le signe, sont difficiles à prévoir, c'est-à-dire les soldes (solde de la balance des paiements, de la balance commerciale, épargne, etc..). A noter que "a" mauvais signifie "a" petit.

La procédure raisonnable semble devoir être finalement d'estimer "a" au moyen des données concernant l'année T et de juger du caractère de Q au moyen des données concernant l'année T + 1.

5/ D'autre part ; si l'on veut comparer les valeurs de 2 indices Q_1 et Q_2 concernant des données distinctes, il conviendrait qu'il s'agisse bien de données également nombreuses ($N_1 \# N_2$) et de tailles comparables ($r_1 \# r_2$). Finalement si les écarts-types sont comparables ($\sigma_1 \# \sigma_2$), le fait que Q_1 diffère notablement de Q_2 peut encore s'expliquer en partie par l'effet du hasard (distribution de la variable Q).

6/ Remarque

On peut encore se demander s'il convient par exemple d'accepter le modèle N° 1 contre le modèle N° 4 : absence ou existence d'erreurs systématiques. Il ne semble pas que la statistique Q soit choisie à priori pour fournir ici le test le plus puissant. Dans le cas où le modèle N° 4 supposerait $\delta_i = \delta$ (la même erreur pour tous les i), le test de Student-Fisher s'imposerait à ce point de vue.

Plus généralement la statistique Q n'a aucune position privilégiée pour supplanter les tests connus. Il faut voir avant tout dans Q un bon indice de concordance.

CONCLUSION

On vient d'étudier un problème concret où la valeur donnée à N est couramment de l'ordre de 10, et parfois moins (quand il s'agit d'une série chronologique). Pourtant nous n'avons trouvé la solution numérique exacte du problème que d'une part pour N = 1 ou 2, - d'autre part pour N grand (en fait si N est de l'ordre de 25, la solution approchée Q_1 et la solution asymptotique Laplace-Gaussienne ne donnent guère de résultats différents). En revanche, en appliquant la solution Q_1 avec N = 10, on est assuré de n'avoir que des résultats assez médiocres - disons des ordres de grandeur -

Il serait souhaitable que les recherches se poursuivent, en vue d'une meilleure approximation - ou d'un calcul pratique des intégrales doubles donnant la solution exacte du problème.

Toutefois, le phénomène est assez général. Par exemple la théorie des files d'attente aux guichets (disons de la SNCF) est fort simple s'il n'y a qu'un guichet, - est encore praticable à la rigueur pour 2 guichets, après quoi il convient de traiter le cas d'un nombre très élevé de guichets. Les vraies difficultés commencent avec le cas d'un nombre réaliste de guichets.

Dans certains problèmes combinatoires (comme celui du nombre d'isolés dans un groupe) on retrouve la même difficulté : entre les cas qu'on peut traiter directement et le cas asymptotique, le problème posé dans le cas général possède une solution théorique qui ne se prête pas au calcul, - quand on la connaît.

On constate même assez souvent que la solution asymptotique convient très mal dans les cas où l'infini est loin ; ceci n'exclut pas la possibilité de trouver de bonnes approximations ou d'essayer une méthode de Monte-Carlo si l'on en a les moyens. Présentement, on peut trouver trop mince le point de contact entre le problème traité (1ère Partie) et les besoins des utilisateurs, pour justifier l'emploi de grands moyens de calculs.

BIBLIOGRAPHIE

[1] THEIL M. - Economic Forecast and Policy (Amsterdam 2° Ed. 1961).

[2] THIONET P. - Un indice statistique destiné à la comparaison des prévisions et des réalisations, Journal de la Société de Statistique de Paris, 105-3 (1964) 181-89.

[3] JORGENSON Dale W. - The predictive perforlance of quaterly econometric models of the United States, The Econometric Society, Bonn 1967.

[4] HOGBEN D., PINKHAM R.S., WILK M.B. -

1/ The moments of a variate related to the non-central t.

1/ An approximation to the distribution of Q (a variate related to the non-central t).

Annals of Mathematical Statistics (March 1964) p. 298-14 et 315-18.