REVUE DE STATISTIQUE APPLIQUÉE

B. ROY

Somme d'un nombre aléatoire de termes aléatoires. Application aux problèmes de stockage

Revue de statistique appliquée, tome 8, n° 1 (1960), p. 51-60

<http://www.numdam.org/item?id=RSA_1960__8_1_51_0>

SOMME D'UN NOMBRE ALÉATOIRE DE TERMES ALÉATOIRES APPLICATION AUX PROBLÈMES DE STOCKAGE

B. ROY

Attaché à la direction Scientifique de la Société d'Économie et de Mathématique Appliquées

En matière économique ou industrielle, les flux qui se rencontrent naturellement, ne présentent que très rarement un écoulement continu. Ils apparaissent au contraire généralement comme fractionnés en un nombre aléatoire de lots élémentaires ayant eux-mêmes une taille aléatoire, si bien que la quantité de flux écoulé durant une période se trouve être une variable aléatoire de la forme :

$$Z = \sum_{i=1}^{i=N} X_i$$

où N représente le nombre de lots écoulés durant la période, et où les X_i représentent la taille de ces lots. C'est l'étude, sous certaines hypothèses, de l'addition qu'implique de telles variables, dont il va être question ici.

Dans tout ce qui suit, les hypothèses suivantes seront implicitement supposées vérifiées :

1/ N est une variable aléatoire indépendante de la suite des X_i.

2/ Les X_i sont des variables aléatoires indépendantes, et de même loi,

et les notations définies ci-dessous seront sans cesse conservées :

loi de probabilité totale de Z	$H(z) = Pr\ [Z < z]$
loi de probabilité totale de X	$F(x) = Pr\ [X < x]$
loi de probabilité élémentaire de N	$p_n = Pr\ [N = n]$

On peut donner un support concret à ces notions générales en se référant au cas particulier d'un stock dont les sorties sont motivées par des commandes. Dans ce cas en effet, la quantité de produits écoulés durant un mois par exemple est bien une variable du type Z satisfaisant en général aux hypothèses précisées, et dans laquelle N et X symbolisent respectivement les variables aléatoires que sont les nombres mensuels de commandes et la taille de ces commandes. L'examen des bons de commandes fournit directement des échantillons des variables N et X à partir desquels il est possible de se faire une idée des lois F(x) et p_n.

Ainsi, les variables aléatoires de type Z se rencontrent-elles fréquemment dans les problèmes concrets de Recherche Opérationnelle. Si l'on sait faire l'addition d'un nombre aléatoire de termes aléatoires indépendants et de même loi, on pourra déduire la loi H des lois composantes F et p; si l'on ne sait pas faire une telle addition, on se verra contraint d'ajuster directement la loi H sur

un échantillon de la variable Z. Or, des raisons d'ordre à la fois statistique et structurel permettent de penser qu'en général, il est préférable de passer par l'intermédiaire des lois composantes.

Plaçons-nous tout d'abord d'un point de vue strictement statistique, et supposons que les renseignements dont on dispose couvrent q périodes. Il en résulte trois sortes d'échantillons :

Le premier de taille λ_q relatif à la variable X (λ désignant la valeur moyenne de N sur une période).

Le second de taille q relatif à la variable N.

Le troisième de taille q relatif à la variable Z.

Le premier de ces échantillons est en général très suffisant pour obtenir une estimation satisfaisante de la loi F(x). Il en est de même du second pour l'ajustement de la loi p_n. En effet, bien qu'il soit de taille notablement plus petite, il se rapporte à une variable discrète pour laquelle il est fréquemment possible de déterminer a priori un type de loi tel que, loi binomiale ou loi de Poisson. Les choses sont très différentes pour le troisième, car il se rapporte à une variable continue, en général très dispersée et à coup sûr (nous aurons l'occasion de le voir) beaucoup plus dispersée que les variables X ou N prises séparément. Ainsi, la méthode (que l'on peut qualifier de globale) qui consiste à déterminer directement H(z) à partir de ce dernier échantillon, s'avèrera souvent beaucoup moins précise que celle (que l'on peut qualifier d'analytique) qui repose sur une détermination préalable des lois composantes. C'est bien naturel , puisque le troisième échantillon ne contient qu'une partie seulement de l'information contenue dans les deux premiers.

La méthode analytique d'autre part permettra dans de nombreux problèmes de fonder la loi H sur certaines grandeurs ayant un sens concret très précis. L'introduction de ces grandeurs comme paramètres dans H pourra rendre possible un réajustement périodique facile de cette loi ou une adaptation convenable à certaines instabilités telles que trend, fluctuations saisonnières, variation d'une clientèle ou d'un parc d'appareils. Il est bien clair que la méthode globale, en mélangeant tout, ne peut prétendre à de tels avantages, et cela d'autant plus que le fractionnement risque de rendre l'échantillon numéro 3 inutilisable.

Cette souplesse et cette sécurité de la méthode analytique constituent le point de vue structurel auquel il a été fait allusion précédemment. Comme il se trouve lié à chaque cas particulier il est difficile de le développer davantage. Il sera illustré par l'exemple concret présenté à la fin de l'article.

L'importance du rôle joué, dans les problèmes concrets, par l'addition d'un nombre aléatoire de termes aléatoires indépendants et de même loi, étant ainsi mise en évidence, abordons maintenant l'aspect théorique des problèmes soulevés par de telles additions. Supposons donc les lois F(x) et p_n données, et examinons ce que l'on peut dire de la loi H(z).

En premier lieu, il est possible d'exprimer H à l'aide d'un développement en série faisant intervenir les lois F et p. Pour cela introduisons la suite e_n d'évènements exclusifs : e_n ayant lieu lorsque N = n, il vient :

$$Pr\,[Z < z] = p_o Pr\,[Z < z/e_o] + \ldots + p_n Pr\,[Z < z/e_n] + \ldots$$

Posons :

$$Pr\,[X < z/e_n] = F_n(z)$$

cette fonction se déduit aisément de la fonction F(x) par convolution (puisqu'elle n'est autre que la loi de la somme de n variables aléatoires indépendantes de type X). Il vient alors :

$$H(z) = \sum_{n=0}^{n=\infty} p_n F_n(z) \qquad (1)$$

avec la convention suivante :

$$F_0(z) = Y(z) \text{ fonction de Heaviside.}$$

Cette formule permet théoriquement de calculer H(z) pour une valeur donnée de z. En fait, le calcul est assez pénible car on est obligé de prendre un très grand nombre de termes dans le développement (1), et de conserver pour chacun d'eux un très grand nombre de décimales (chaque terme étant très petit) si l'on veut obtenir une précision suffisante pour H(z).

Il est par contre quasi impossible de résoudre l'équation :

$$H(z) = 1 - \epsilon$$

équation très importante dans les problèmes de stockage en particulier. Ainsi cette formule est d'un intérêt très restreint.

On peut ensuite déduire la fonction caractéristique $\Phi_Z(t)$ des fonctions caractéristiques des variables X et N.

En effet,

$$\Phi_Z(t) = E(e^{itZ}) = E\,[E(e^{itZ/e}\ n)] = E\,[\varphi_X^N(t)]$$

$\varphi_X(t)$ désignant la fonction caractéristique de la variable X.

Si l'on désigne par G(u) la fonction génératrice de la variable N il vient :

$$\Phi_Z(t) = G\,[\varphi_X(t)] \qquad (2)$$

Cette seconde formule est beaucoup plus importante que la précédente parce qu'elle permet de déduire les moments de la variable Z de ceux des variables X et N.

Il est facile d'établir les formules suivantes relatives à la moyenne et à la variance :

$$E(Z) = E(N)\,E(X)$$

$$V(Z) = E(N)\,V(X) + V(N)\,E^2(X) \qquad (3)$$

La seconde formule (3) lie la variance de Z aux variances de X et N. Elle est particulièrement importante lorsqu'on est en droit de penser que la variable Z suit une loi de Gauss. Elle permet en effet dans ce cas d'estimer les paramètres de cette loi à partir des échantillons 1 et 2. Ce procédé semble très nettement préférable à celui qui consiste à estimer ces paramètres à partir du troisième échantillon.

On peut également calculer les moments d'ordre supérieur, lesquels peuvent être utiles pour ajuster la loi H à des lois plus compliquées que celle de Gauss, telles que loi de Pierson ou loi de Gram-Charlier.

Laissons maintenant le cas général, et intéressons-nous à celui plus particulier, mais très important, où N suit une loi de Poisson.

Il vient alors :

$$p_n = e^{-\lambda} \frac{\lambda^n}{n!} \qquad \text{avec } E(N) = \lambda$$

La fonction génératrice de la loi de Poisson a pour expression :

$$G(u) = \sum_{n=o}^{n=\infty} e^{-\lambda} \frac{\lambda^n}{n!} u^n = e^{\lambda(u-1)}$$

et la formule (2) s'écrit dans ce cas particulier :

$$\Phi_Z(t) = e^{\lambda [\varphi_X(t) - 1]} \quad (1)$$

D'où l'on déduit :

$$\psi_Z(t) = \lambda [\varphi_X(t) - 1],$$

ψ étant la seconde fonction caractéristique de la variable Z.

Si l'on développe maintenant en série la fonction φ :

$$\varphi_X(t) = 1 + m_1 it + \ldots + m_q \frac{(it)^q}{q!} + \ldots$$

Il vient :

$$\psi_Z(t) = \lambda m_1 it + \ldots + \lambda m_q \frac{(it)^q}{q!} + \ldots$$

D'où l'on déduit :

$$k_q = \lambda m_q \qquad (4)$$

formule liant le cumulant d'ordre q de la variable Z au moment de même ordre de la variable X.

Cette formule permet d'écrire le développement en série de Gram-Charlier de la fonction H(z) sous une forme particulièrement simple. Avant de le faire, disons quelques mots de ces séries.

Elles font intervenir la suite orthogonale que constituent les polynômes d'Hermite. Ceux-ci sont définis à partir des dérivées successives de la densité de Gauss :

$$y(t) = \frac{1}{\sqrt{2\pi}} e^{-t^2/2}$$

Il en résulte que le développement en série de Gram-Charlier de la densité de probabilité h(z) peut s'écrire, en ne faisant intervenir que ces dérivées :

$$h(z)dz = dz \sum_{r=o}^{r=\infty} c_r y^{(r)}(z)$$

---

(1) On remarquera que la loi H(z) est indéfiniment divisible.

les c_r étant des coefficients liés au moment d'ordre inférieur ou égal à r de la variable Z, par une formule assez compliquée que l'on n'utilisera pas.

Les premiers coefficients de ce développement prennent une forme plus simple lorsque l'on introduit la variable centrée réduite :

$$t = \frac{z - E(Z)}{\sqrt{V(Z)}}$$

La série de Gram-Charlier prend alors la forme :

$$h(z)dz = [y(t) - c_3 y^{(3)}(t) + c_4 y^{(4)}(t) - \ldots + (-1)^r c_r y^{(r)}(t) + \ldots] dt$$

qui ne contient pas de termes en y', y''.

Les coefficients c_r s'expriment assez simplement en fonction des cumulants de la variable Z d'ordre inférieur ou égal à r :

$$\left.\begin{aligned} c_3 &= \frac{k_3}{3!\, k_2^{3/2}} \\ c_4 &= \frac{k_4}{4!\, k_2^2} \\ c_5 &= \frac{k_5}{5!\, k_2^{5/2}} \end{aligned}\right\} \qquad (5)$$

La formule se complique quelque peu à partir de l'ordre 6.

Intégrons maintenant ce développement de manière à obtenir une expression de la loi de probabilité totale H(z). Il vient :

$$H(z) = \Pi(t) + \sum_{r=3}^{r=\infty} (-1)^r c_r y^{(r-1)}(t) \qquad \Pi(t) = \int_0^t y(t)dt \qquad (6)$$

Fermons ici la parenthèse ouverte sur les séries de Gram-Charlier.

La formule (4) permet d'écrire le développement de H(z) avec autant de termes que l'on voudra, en ne faisant intervenir que les moments de la distribution de X et le paramètre λ de la loi de Poisson. Une telle série n'aura d'intérêt pratique que si elle converge suffisamment vite. Or, les séries de Gram-Charlier sont particulièrement connues pour l'irrégularité de leur convergence . On sait néanmoins que dans l'addition d'un petit nombre de variables aléatoires satisfaisant à certaines conditions très générales, les premiers termes du développement de Gram-Charlier fournissent une approximation très satisfaisante de la loi de la somme. Or ici, la variable Z apparaît dans une certaine mesure comme la somme de λ variables Y, où Y est une variable obtenue en ajoutant un nombre aléatoire poissonnien de moyenne 1, de variable X.

On est donc en droit de penser que les premiers termes du développement constitueront une approximation très souvent acceptable de la loi H, plutôt que d'entrer dans des considérations théoriques compliquées relatives à cette convergence, nous avons préféré exécuter quelques calculs numériques. Les principaux résultats obtenus sont présentés dans le tableau ci-contre. Il a été calculé pour différents types de lois F et différentes valeurs de t, la valeur exacte :

$$\Pr\left[\frac{Z - E(Z)}{\sqrt{V(Z)}} > t\right] \quad \text{(à l'aide de la formule 1),}$$

SOMME DE TERMES ALEATOIRES X EN NOMBRE ALEATOIRE DE POISSON (A = 4)

LOI de X		t	$Pr\left[\frac{Z - E(z)}{\sqrt{V(z)}} > t\right]$		
nature	paramètre		vraie valeur	Gram-Charlier(1)	Gauss
GAUSS	m = 0	1	0,145080	0,143530	0,158655
	σ quelconque	2	0,026245	0,026136	0,022750
	$m = \frac{\sigma}{2}$	1	0,153967	0,143970	0,158655
	quelconque	2	0,036618	0,041742	0,022750
	$m = \sigma$ σ quelconque	3	0,006970	0,007612	0,001350
EXPONENTIELLE	λ quelconque	2	0,0437456	0,058138 (0,052172)	0,022750
		3	0,0110077	0,012613 (0,015552)	0,001350

et la valeur approchée obtenue en n'utilisant que les deux premiers termes de la série de Gram-Charlier (ordre 3, ordre 4), ainsi que la valeur obtenue en limitant cette série à son terme principal (loi de Gauss).

Le paramètre λ a été choisi égal à 4 : plus petit, il aurait été vraiment trop petit, plus grand, il aurait rendu trop pénible l'application de la formule (1), laquelle doit déjà être écrite avec 17 termes dans le cas retenu. Cette formule limitait également le choix de la fonction F.

On peut constater sur le tableau que l'approximation de Gram-Charlier limitée à ses deux premiers termes, est en général très suffisante pour les applications concrètes, et cela même dans les très petites probabilités (t = 3). Par contre, l'approximation gaussienne s'avère en général très mauvaise.

Ainsi, il sera fréquemment possible d'écrire :

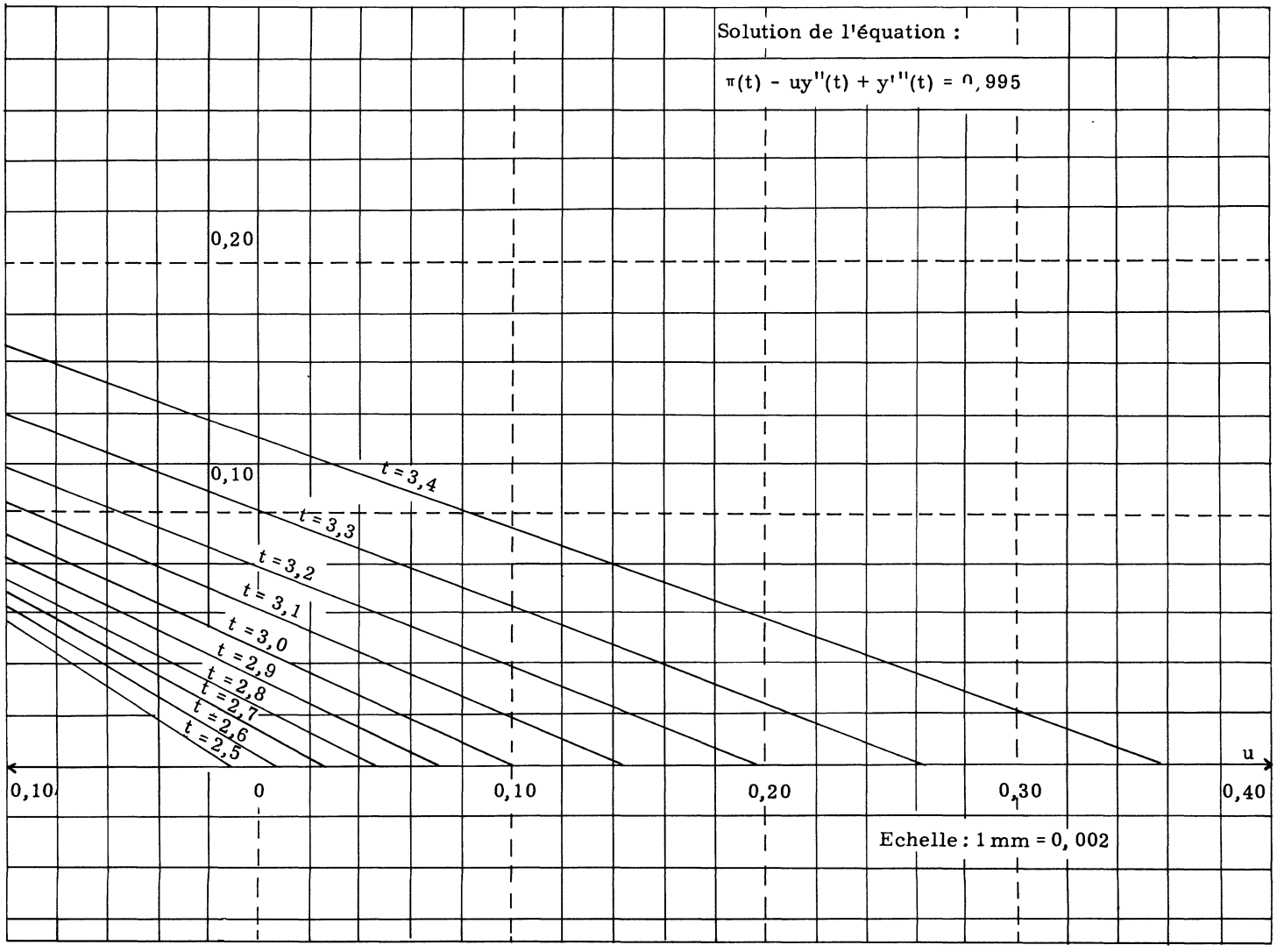
$$H(z) \# \Pi(t) - uy''(t) + vy'''(t). \qquad (7)$$

où

$$t = \frac{z - \lambda m_1}{\sqrt{\lambda m_2}} \qquad u = \frac{m_3}{6\sqrt{\lambda m_2^3}} \qquad v = \frac{m_4}{24\ \lambda m_2^2} \qquad (7)$$

---

(1) La série de Gram-Charlier a été limitée à ses 2 premiers termes; les valeurs entre parenthèses correspondent à un terme de plus.


Solution de l'équation :
π(t) - uy''(t) + y'''(t) = 0,995
0,20
0,10
t = 3,4
t = 3,3
t = 3,2
t = 3,1
t = 3,0
t = 2,9
t = 2,8
t = 2,7
t = 2,6
t = 2,5
u
0,10
0
0,10
0,20
0,30
0,40
Echelle : 1 mm = 0, 002

Cette expression permet de calculer facilement H(z) lorsqu'on se donne z car les fonctions Π, y'' et y''' sont tabulées.

Le calcul inverse qui consiste à résoudre l'équation :

$$H(z) = 1 - \varepsilon,$$

est un peu moins immédiat. On peut le faire par approximations successives. On peut aussi remarquer que :

$$\Pi(t) - uy''(t) + vy'''(t) = 1 - \varepsilon$$

est une formule linéaire en u et v pour t et ε fixés.

Il est donc très facile de tracer l'abaque donnant la valeur de t en fonction de u et v pour ε donné. La résolution graphique de l'équation considérée est alors immédiate. On pourra constater sur l'abaque ci-joint la grande précision de cette méthode graphique.

Afin d'illustrer ces formules, présentons pour terminer le problème concret qui fut à l'origine de cette étude.

L'objet de ce problème était l'élaboration d'une règle d'approvisionnement pour les tubes de casing. (Ce sont des tubes destinés à la recherche pétrolière). L'élaboration d'une telle règle nécessitait un examen préalable des lois de la consommation de ces tubes. Il importe tout d'abord de préciser que ceux-ci se distinguent par leur diamètre et par l'épaisseur et la qualité de l'acier qui les compose. La donnée d'un diamètre, d'une épaisseur et d'une qualité détermine ce que l'on appelle une spécification. Les tubes de même diamètre sont assemblés en colonnes comprenant en général plusieurs spécifications. Ces colonnes constituent l'unité élémentaire de consommation. Elles sont descendues en terre par des appareils que l'on sépare en trois catégories : appareils lourds, moyens et légers, qui pour nous seront repérés par les indices A, B, C.

On ne s'intéressera désormais qu'à une seule spécification, d'ailleurs quelconque, pour laquelle on définira les variables aléatoires suivantes :

Z = consommation mensuelle;

Z_A, Z_B, Z_C = consommation mensuelle des appareils lourds, moyens et légers;

X_A, X_B, X_C = longueurs de la spécification considérée, entrant dans les colonnes descendues par les appareils lourds, moyens et légers

N_A, N_B, N_C = nombres mensuels de colonnes descendues par les appareils lourds, moyens et légers.

Il est bien clair que l'on a :

$$Z = Z_A + Z_B + Z_C$$

Z_A = somme de N_A variables aléatoires X_A; Z_B = somme de N_B variables aléatoires X_B; Z_C = somme de N_C variables aléatoires X_C.

L'indépendance des différentes variables X entrant dans chacune de ces sommes était très certainement bien réalisée. L'indépendance entre les variables X et N, moins évidente, a pu être testée avec succès.

Des raisons théoriques permettaient de supposer que les variables N suivaient des lois de Poisson dont le paramètre était proportionnel au nombre d'appareils lourds, moyens et légers (a, b, c) existant en parc, durant le mois considéré. Soit :

$$E(N_A) = a\,\lambda_A$$

$$E(N_B) = b\,\lambda_B$$

$$E(N_C) = c\,\lambda_C$$

(λ_A, λ_B, λ_C étant des caractéristiques de chaque type d'appareil).

Ces hypothèses ont pu être vérifiées.

La formule (5) donnait donc une expression simple des cumulants des variables Z_A, Z_B, Z_C. Les cumulants de Z se déduisent de ces derniers par une simple addition (les variables Z_A, Z_B, Z_C étant indépendantes). D'où :

$$k_q = a\lambda_A m_{A,q} + b\lambda_B m_{B,q} + c\lambda_C n_{C,q} \qquad (8)$$

formule donnant les cumulants de Z en fonction de l'état de parc a, b, c des paramètres λ et des moments m des variables X.

Les moments m étaient en fait les seules caractéristiques des lois F_A, F_B, F_C des variables X_A, X_B, X_C qu'il était matériellement possible d'obtenir. En effet, ces variables étaient liées à une autre variable aléatoire, longueur des colonnes (dont la loi de probabilité était connue) par un mécanisme probabiliste fort compliqué.

Dans ces conditions, il était difficile d'obtenir un ajustement analytique de cette loi. Par contre, il était possible de simuler l'ensemble du mécanisme sur ordinateurs et d'obtenir ainsi de très bons échantillons des variables X à partir desquels on pouvait estimer les moments m.

La formule (8) jointe aux formules (5) et (6) fournissait donc une approximation de la loi de probabilité de la consommation mensuelle de chaque spécification, en fonction de l'état du parc qui restait paramétrique, ce qui était indispensable. La maniabilité de l'expression obtenue fut grandement accrue par la quasi insensibilité des coefficients u et v aux variations de parc. Ainsi la solution de l'équation :

$$\Pi(t) - uy''(t) + vy'''(t) = 1 - \varepsilon$$

était-elle indépendante des paramètres a, b, c, ce qui simplifiait considérablement le calcul des stocks de sécurité.

Une des difficultés du problème que nous avions à traiter résidait dans l'instabilité du processus de consommation. Il était donc indispensable de fonder la loi de consommation H(z) sur des grandeurs que l'on pourrait estimer régulièrement de manière à surveiller la validité de H. La formule obtenue répond précisément à ces exigences : le calcul périodique des λ étant une chose facile, celui des moments m pouvant être répété sans trop de difficultés chaque fois qu'une modification dans la décomposition des colonnes est signalée. Aucun ajustement analytique direct des lois H n'aurait pu présenter la souplesse de la méthode de Monte-Carlo, pour la détermination des moments m.

L'examen rapide du problème des tubes de casing, montre que, les résultats qui précèdent, malgré leur aspect théorique et leur caractère limité, peuvent rendre des services non négligeables dans certaines applications concrètes.