REVUE DES SCIENCES DE L'EAU, Rev. Sci. Eau 17/1(2004) 23-47

Comparaison par simulation de Monte Carlo des propriétés de deux estimateurs du paramètre d'échelle de la loi exponentielle : méthode du maximum de vraisemblance (MV) et méthode des moindres carrés (MC)

Comparison of statistical properties of two estimators (the method of maximum of likelihood estimator and the method of least squares estimator) of the simple exponential distribution using a Monte Carlo simulation method

S. SAMBOU

Reçu le 1er octobre 2002, accepté le 15 mai 2003*.

SUMMARY

Exponential distributions are frequently applied in hydrology: drought frequency analysis of the duration and severity of water flow conditions MATHIEU L. *and al.* (1991); regional frequency of storm intensities ARNAUD P., LAVABRE J. (1999); partial duration of hydrological droughts KJELDSEN T. R. *and al.* (1999); and daily rainfall modelling CHAPMAN T.G. (1997); KABAILI Z. (1983). This method has only one parameter, and it is easy to use. Its parameter is mainly estimated using the maximum likelihood estimator (MLE) or the method of moments estimator (MOME), but the least square estimator (LSE) can also be applied. For the one-parameter exponential distribution, MOME and MLE give the same expression for the parameter:

$$\hat{x}_0 = \frac{\sum_{k=1}^{E_x} x_k}{E_x}$$

Using LSE requires a E_x size sample of exponential variables and involves the following steps:

Université Cheikh Anta DIOP, Faculté des Sciences et Techniques, Département de Physique, Dakar-Fann, Sénégal.

Correspondance : sousamb@ucad.refer.sn

* Les commentaires seront reçus jusqu'au 30 septembre 2004.

1. Sorting the E_x variables in the sample in ascending order

2. Associating to each quantile x_k whose rank is k in the sorted sample an empirical frequency $\hat{F}_k$ using Hazen's law such as : $\hat{F}_k = \dfrac{k - 0.5}{E_x}$

3. Plotting x_k against $\ln\left(1 - \hat{F}_k\right)$ and using LSE to calculate $\hat{x}_0$ by :

$$\hat{x}_0 = -\frac{\sum_{k=1}^{E_x} x_k \ln\left(1-\hat{F}_k\right)}{\sum_{k=1}^{E_x}\left\{\ln\left(1-\hat{F}_k\right)\right\}^2}$$

In this paper we compare the asymptotic behaviour of the statistical properties (mean and variance) of the MLE and the LSE. These comparisons must be made by using a great number of sample parameter estimations. In practice, only one historical sample of variables issued from a known exponential distribution was available, from which only one parameter can be calculated. To overcome this difficulty, samples of variables whose original theoretical exponential distribution is known are generated using the Monte Carlo numerical method. Samples of estimated parameters (using the MLE or the LSE) are then created from the above samples of random variables, and the statistical properties of the two estimators are then calculated. These different successive steps are summarised below:

1. Generate sample of finite size E_x for known exponential variables

2. Use this sample to estimate one parameter using MLE or LSE

3. Do steps 1 to 2 N_p times to collect a N_p size sample of parameter estimations

4. Use this sample to calculate statistical properties (mean and variance) for the two estimators.

According to this approach, sizes E_x and N_p should influence the statistical properties of the two estimators. We have verified this with a one-parameter exponential law, with a known theoretical parameter $x_0 = 1$. Samples of estimated parameters of size N_p have been generated from virtual samples of size E_x issued from a population following the above statistical distribution. During this operation, one of these sizes, E_x or N_p, has been held constant, while the other, N_p or E_x, changed with a constant step. Statistical properties of the estimators have then been calculated for each of the two cases.

Let $Var_{E_x}\left(\hat{x}_0\left(N_p\right)\right)$ and $E_{E_x}\left(\hat{x}_0\left(N_p\right)\right)$ be statistical properties (variance and mean) of the two estimators for fixed values of E_x, and $Var_{N_p}\left(\hat{x}_0\left(E_x\right)\right)$ and $E_{N_p}\left(\hat{x}_0\left(E_x\right)\right)$, the same statistical properties for fixed values of N_p.

Plotting $Var_{E_x}\left(\hat{x}_0\left(N_p\right)\right)$ for $E_x = 10$ and $E_x = 100$ shows that for large values of N_p (1000 to 5000) variance tends towards a constant value, close to 0.1 for E_x=10 (figure 1a) and to 0.01 for E_x=100 (figure 1b), both equal to $1/E_x$, when the MLE is used. When parameters are estimated with the LSE, variance tends towards a constant value, greater than the preceding ones (figure 1a and figure 1b). Plotting $Var_{N_p}\left(\hat{x}_0\left(E_x\right)\right)$ when N_p=1000 is constant, the variance decreases as E_x grows whatever the estimator, but for a given value of E_x, the variance is always greater when the LSE is used (figure 3). These two calculations show that asymptotic variance depends only on size E_x of samples of known exponential distribution.

Plotting $E_{E_x}\left(\hat{x}_0\left(N_p\right)\right)$ when E_x =10, for important values of N_p, the mean is close to the true parameter for the MLE, and greater than this true parameter for the LSE (figure 2a). When E_x=100, the mean is close to the true parameter for the two estimators (figure 2b). From these calculations we notice that the asymptotic mean depends only on the size of E_x for known exponential variables and on the used estimator. The MLE seems to present no bias for the mean, while the LSE presents a bias for small values of E_x, but this bias disappears as E_x increases. To quantify this degree of dependence, we have plotted $E_{N_p}\left(\hat{x}_0\left(E_x\right)\right)$ for Np = 1000 (figure 4). For the two estimators, the mean presents an initial bias, when E_x is low and the bias disappears when E_x becomes higher. The initial bias is more important with the LSE.

In summary, the asymptotic statistical properties of the two estimators (mean and variance) depend only on the size of E_x for known exponential distribution variables.

Empirical plots are unstable for low sample sizes, are sensitive to sampling, and are very difficult to explain. Analytical expressions for the asymptotic statistical properties of the two estimators are needed for realistic comparison. According to formulae (1) and (2), statistical properties depend on $E_\infty\left(x_k\right)$ and $E_\infty\left(x_k^2\right)$ respectively and the asymptotic mean of x_k and x_k^2. $E_\infty\left(x_k\right)$ and $E_\infty\left(x_k^2\right)$ have been derived using the density of probability of x_k through statistics of rank. Asymptotic statistical properties of the two estimators have then been evaluated using the expressions of $E_\infty\left(x_k\right)$ and $E_\infty\left(x_k^2\right)$.

We let $E_\infty\left[\hat{x}_0\left(E_x\right)\right]$ be the asymptotic mean of estimator, $\mathrm{Var}_\infty\left[\hat{x}_0\left(E_x\right)\right]$ be the asymptotic variance, and x_0 be the theoretical parameter of exponential distribution. By plotting $E_\infty\left[\hat{x}_0\left(E_x\right)\right]$ for x_0 =1, we note that this expression has a constant value, equal to unity when the MLE was applied, and that it decreases quickly to unity when the LSE was applied (figure 6). By plotting $Var_\infty\left[\hat{x}_0\left(E_x\right)\right]$ for x_0 = 1, we also note that the theoretical asymptotic variance diminishes as E_x grows, but is greater when the LSE was applied (figure 5). By comparing with empirical plots when x_0 = 1, we establish the same trends.

Theoretical derivations of asymptotic statistical proprieties have confirmed empirical experiences:

- The MLE for a one-parameter exponential presents no bias.
- The LSE for a one-parameter exponential is a consistent estimator of the simple exponential parameter.

Key words: *one-parameter exponential distribution, Monte Carlo numerical simulation, maximum of likelihood estimator, least squares estimator, mean, variance.*

RÉSUMÉ

La loi exponentielle est très répandue en hydrologie : elle est faiblement paramétrée, de mise en œuvre aisée. Deux méthodes sont fréquemment utilisées pour estimer son paramètre : la méthode du maximum de vraisemblance et la méthode des moments, qui fournissent la même estimation. À côté de ces deux méthodes, il y a celle des moindres carrés qui est très rarement utilisée

pour cette loi. Dans cet article, nous comparons le comportement asymptotique de l'estimateur de la méthode des moindres carrés avec celui de la méthode du maximum de vraisemblance en partant d'une loi exponentielle à un seul paramètre a connu, puis en généralisant les résultats obtenus à partir de la dérivation des expressions analytiques. L'échantillon historique disponible en pratique étant unique, et de longueur généralement courte par rapport à l'information que l'on désire en tirer, l'étude des propriétés statistiques des estimateurs ne pourra se faire qu'à partir d'échantillons de variables aléatoires représentant des réalisations virtuelles du phénomène hydrologique concerné obtenus par simulations de Monte Carlo. L'étude par simulation de Monte Carlo montre que pour de faibles échantillons, l'espérance mathématique des deux estimateurs tend vers le paramètre réel, et que la variance de l'estimateur des moindres carrés est supérieure à celle de l'estimateur du maximum de vraisemblance.

Mots clés : *distribution exponentielle, simulation numérique, Monte Carlo, méthode des moindres carrés, méthode du maximum de vraisemblance, moyenne, variance.*

1 - INTRODUCTION

La loi exponentielle est très populaire en hydrologie dans le domaine de la gestion des ressources en eau où ses applications sont fort nombreuses et l'énumération que nous allons en faire est loin d'être exhaustive : sévérités des étiages (MATHIEU *et al.*, 1991), distribution des pluies non nulles en une station (LEBEL *et al.*, 1995), modélisation des débits caractéristiques (GALÉA et PRUDHOMME, 1997), modélisation des pluies journalières (CHAPMAN, 1997 ; PANTOGLOU et TZIAFETAS, 1989 ; KÉBAILI, 1983), modélisation des débits journaliers (HØYBYE et LÀSLZLÒ, 1997), génération de hyétogrammes horaires (ARNAUD et LAVABRE, 1999), modélisation des sécheresses hydrologiques (KJELDSEN *et al.*, 2000). Sous sa forme générale, elle comporte deux paramètres (un de position et un d'échelle). Cependant la version à un seul paramètre, qui se déduit aisément de la première (LANG *et al.*, 1999), est la plus usitée, et c'est sur celle-ci que nous allons porter notre attention ; nous la désignerons par la suite par L1. L'unique paramètre peut être estimé par la méthode du maximum de vraisemblance (MV) ou des moments (M), et par la méthode des moindres carrés (MC) qui est une autre méthode d'estimation. Si les deux premières qui fournissent une estimation triviale du paramètre sont très courantes pour cette loi, il n'en est pas de même de la troisième. Nous nous proposons donc dans cet article de comparer les propriétés statistiques (moyenne, variance) des méthodes (MV) et (MC). Deux approches ont été utilisées pour la comparaison :

- l'approche empirique qui est fondée sur la simulation numérique de Monte Carlo
- l'approche théorique qui repose sur la dérivation d'expressions analytiques des caractéristiques des estimateurs à partir de celles des observations de rang k dans un classement en ordre croissant de l'échantillon.

Dans l'approche empirique un nombre N_p de simulations d'échantillons de variables aléatoires de taille N est utilisé pour générer un échantillon de

paramètres d'effectif N_p estimés par la méthode MV ou par la méthode MC. Ces variables aléatoires ont été extraites d'une population suivant une loi L1 connue après tirages successifs indépendants dans une loi uniforme. La variance et l'espérance mathématique empiriques des deux estimateurs ont ensuite été calculées puis comparées entre elles. Le résultat obtenu montre en particulier que la variance est plus faible quand on utilise la méthode MV.

L'approche théorique repose sur le mode d'estimation du paramètre par les méthodes MV et MC, à partir d'échantillons de variables aléatoires classées par ordre croissant. L'expression du paramètre fait intervenir alors la réalisation de rang k dans un classement en ordre croissant, x_k. Les expressions analytiques des caractéristiques statistiques des deux estimateurs (moyenne $<x_k>$ et variance $Var(x_k)$) dépendent des caractéristiques statistiques des observations de rang k dans un classement par ordre croissant de l'échantillon, et diffèrent entre elles du coefficient de ces derniers. Une étude poussée de ces coefficients a été faite par simulation de Monte Carlo. Il en ressort qu'aux erreurs d'échantillons près, l'espérance mathématique des deux estimateurs tend vers la vraie valeur du paramètre, et que la variance est plus faible quand on utilise la méthode MV. Ce dernier résultat est du à l'influence des coefficients et à la prépondérance des termes en $Cov(x_k, x_l)$ dans l'expression de la variance du paramètre selon l'estimateur.

L'intérêt de ce travail est purement théorique, et la simplicité de la formulation de la loi exponentielle à un seul paramètre a beaucoup facilité le traitement par simulation de Monte Carlo.

2 – PRÉSENTATION DE LA LOI EXPONENTIELLE

2.1 Loi exponentielle à deux paramètres ou L2

Sous sa forme théorique, la fonction de répartition de la loi exponentielle à deux paramètres s'exprime par :

$$F(x) = \Pr ob(X < x) = 1 - e^{\left[-\left(\frac{x-x_0}{a}\right)\right]} \qquad (1)$$

où x_0 paramètre de position, a paramètre d'échelle.

Ses caractéristiques statistiques théoriques se déduisent de cette expression par :

- Espérance mathématique :

$$E(X) = x_0 + a \qquad (2)$$

- Variance

$$Var(X) = a^2 \qquad (3)$$

Ses paramètres peuvent être estimés par la méthode du maximum de vraisemblance ou par la méthode des moindres carrés.

Dans le premier cas (il s'agit du maximum de vraisemblance) on a :

$$\hat{x}_0 = Min\left(x_i, i = 1,..., N\right) \tag{4}$$

$$\hat{a} = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{5}$$

Dans le second cas on a :

$$\hat{x}_0 = Min\left(x_i, i = 1,..., N\right) \tag{6}$$

$$\hat{a} = -\frac{\sum_{k=1}^{N} x_k Ln\left(1 - \hat{F}_k\right)}{\sum_{k=1}^{N}\left\{Ln\left(1 - \hat{F}_k\right)\right\}^2} \tag{7}$$

où $\hat{F}_k = \dfrac{k - 0.5}{N}$ (formule de Hazen)

Dans le cas d'un processus de Poisson, la loi exponentielle à deux paramètres utilisée pour les valeurs supérieures à un seuil donné correspond à la loi de Gumbel utilisée pour les valeurs maximales annuelles (LANG *et al.*, 1999).

2.2 Loi exponentielle à un paramètre ou L1

Si X variable aléatoire supérieure à un seuil S suit une loi exponentielle à deux paramètres (x_0,a) alors ($x - x_0$) suit une loi exponentielle à un paramètre d'expression mathématique (LANG *et al.*, 1999) :

$$F\left(x\right) = 1 - e^{-\left[\frac{x}{a}\right]} \tag{8}$$

3 - MÉTHODOLOGIE

L'étude que nous voulons mener dans cet article est basée sur un nombre élevé d'échantillons de tailles différentes d'observations issues d'une population suivant une loi de type L1 dont le paramètre est connu. En pratique un seul échantillon historique de taille finie est disponible. Pour y remédier, on utilise en général la procédure de simulation de Monte Carlo pour générer les échantillons de réalisations virtuelles des variables aléatoires issues d'une population exponentielle connue selon la démarche ci-dessous :

1. tirer au hasard, dans une loi uniforme, une fréquence théorique F comprise entre 0 et 1,

2. calculer une variable aléatoire par $x = -aLn(1 - F)$ où a est connu,

3. répéter les étapes 1 et 2 N fois, ce qui donne un échantillon d'observations d'effectif N,

4. utiliser cet échantillon pour calculer une estimation $\hat{a}$ du paramètre théorique a, par la méthode MC ou par la méthode MV,

5. répéter les étapes 3 et 4 N_p fois, pour obtenir, suivant la méthode d'estimation, un échantillon d'effectif N_p d'estimations $\hat{a}$ de ce paramètre,

6. cet échantillon de paramètres virtuels joue un rôle important dans l'étude que nous allons entreprendre. Il fournit directement la variance et l'espérance mathématique de l'estimateur.

3.1 Effet de N et N_p sur les propriétés des estimateurs

La taille N de l'échantillon et le nombre N_p de simulations d'échantillons vont avoir un effet sur les caractéristiques statistiques des estimateurs utilisés, que nous étudions sur l'exemple d'une population représentée par une loi exponentielle de paramètre connu et égal à l'unité. Une variable aléatoire extraite de cette population s'exprime par :

$x = -Ln(1 - F)$ où F est une fréquence comprise entre 0 et 1.

En utilisant la procédure de Monte Carlo, nous avons généré N_p simulations d'échantillons de paramètres à partir d'échantillons de variables aléatoires de taille N issues d'une population suivant une loi exponentielle de paramètre connu $a=1$, en faisant varier respectivement N (ou N_p) avec un pas constant. Les courbes empiriques représentant l'évolution des caractéristiques statistiques des estimateurs des deux méthodes MV et MC, $<\hat{a}/N>$ et $<\hat{a}/N_p>$ pour la moyenne, $Var(\hat{a}/N)$ et $Var(\hat{a}/N_p)$ pour la variance ont été tracées. En les comparant entre elles, nous constatons que :

Pour N fixé,

• $Var(\hat{a}/N)$ se stabilise à partir de N_p = 1 000 ; elle tend vers 1/N quand on utilise la méthode MV, et vers une constante supérieure à cette valeur quand on utilise la méthode MC (figure 1a et 1b).

• $<\hat{a}/N>$ est stable dès que N_p dépasse 4 000 ; elle tend vers la vraie valeur pour les deux méthodes ; l'écart entre les deux méthodes diminue lorsque N_p augmente (figure 2a et 2b).

Pour N_p fixé,

• $Var(\hat{a}/N_p)$ décroît pour les deux méthodes quand N augmente ; elle est plus faible quand on utilise la méthode MV (figure 3).

• $<\hat{a}/N_p>$ présente un biais initial correspondant aux erreurs d'échantillonnage qui se stabilise très vite et tend vers la vraie valeur lorsque N croît quelle que soit la méthode utilisée (figure 4).

Il ressort de cette comparaison que :

• la variance des deux estimateurs ne dépend pas du nombre N_p de simulations, pourvu qu'il soit suffisamment élevé, mais plutôt de l'effectif N des échantillons des variables aléatoires obtenues par tirages dans une population suivant une loi exponentielle connue ;

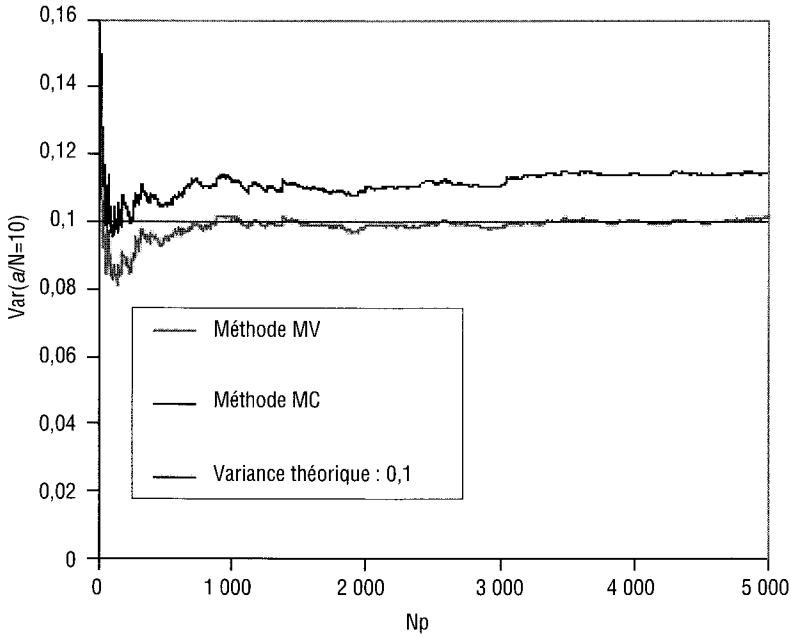

Figure 1a *Var* $(\hat{a}/N = 10)$ en fonction du nombre N_p de simulations.
Var $(\hat{a}/N = 10)$ *as a function of the number of simulations,* N_p.

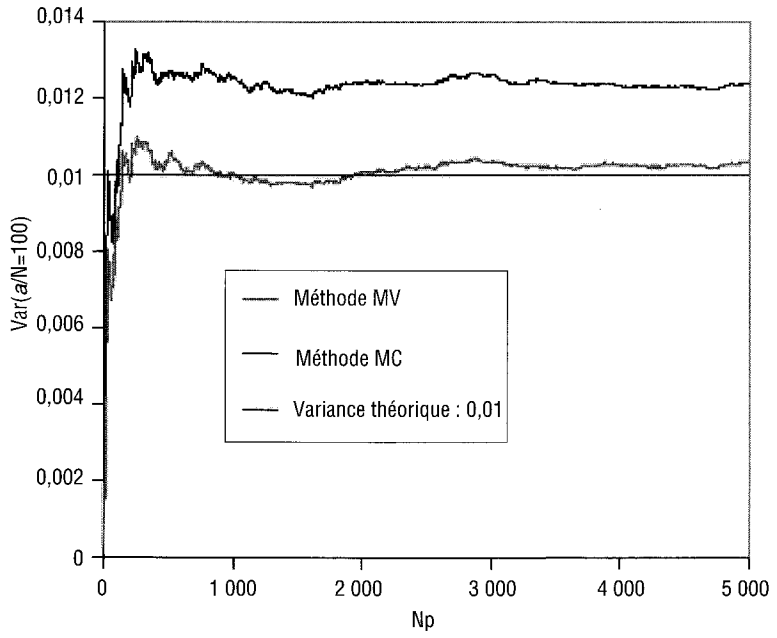

Figure 1b *Var* $(\hat{a}/N = 100)$ en fonction du nombre N_p de simulations.
Var $(\hat{a}/N = 100)$ *as a function of the number of simulations,* N_p.

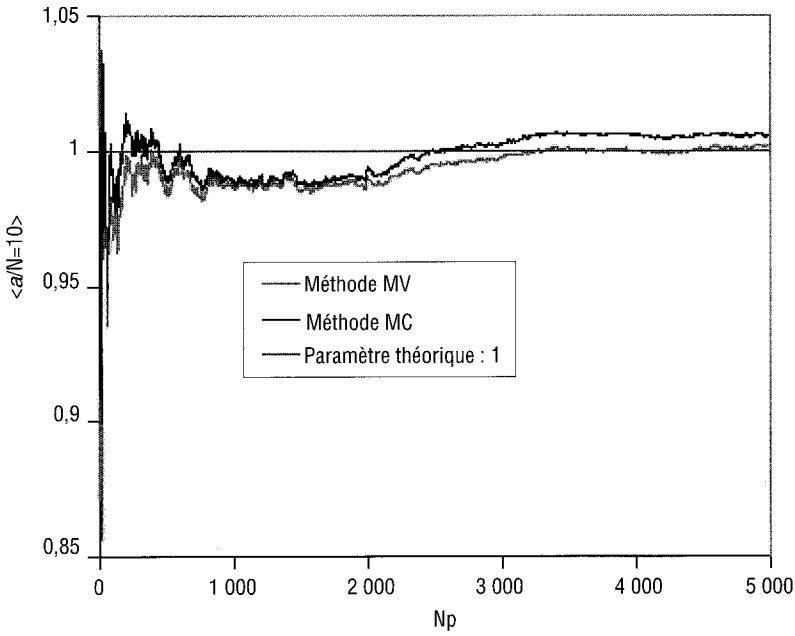

Figure 2a $\langle \hat{a}/N = 10 \rangle$ en fonction du nombre N_p de simulations.
$\langle \hat{a}/N = 10 \rangle$ *as a function of the number of simulations,* N_p.

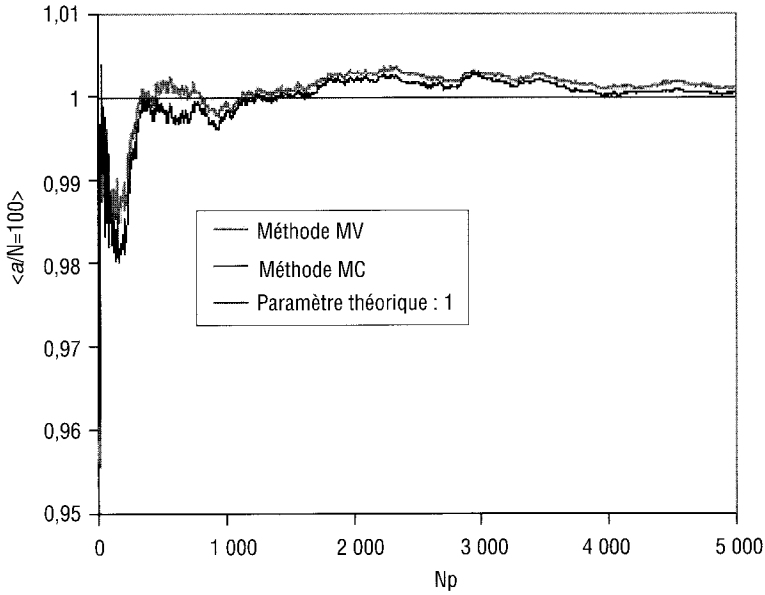

Figure 2b $\langle \hat{a}/N = 100 \rangle$ en fonction du nombre N_p de simulations.
$\langle \hat{a}/N = 100 \rangle$ *as a function of the number of simulations,* N_p.

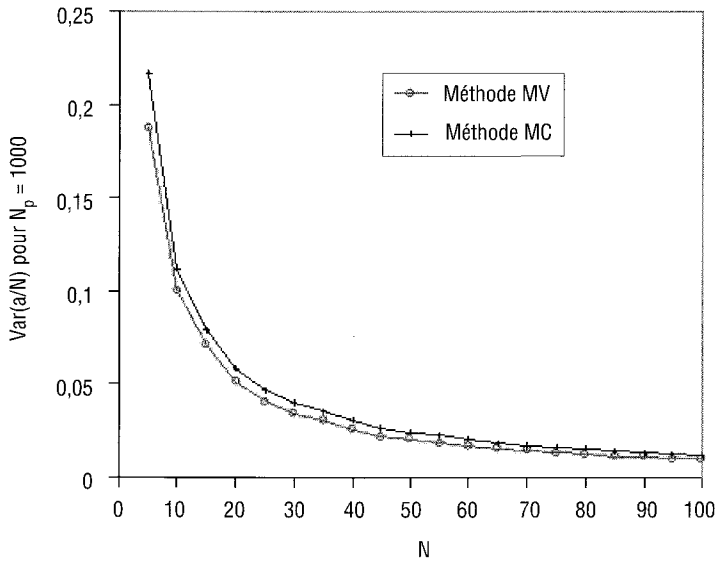

Figure 3 *Var(â/N)* pour N_p= 1000 en fonction de la taille N de l'échantillon.
Var(â/N) *for* N_p *= 1000 as a function of sample size N.*

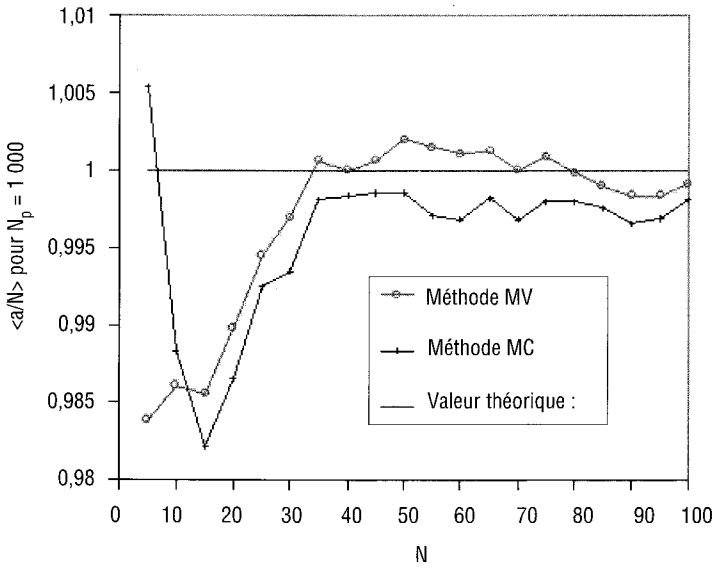

Figure 4 *<â/N>* pour N_p = 1000 en fonction de la taille N de l'échantillon.
<â/N> *for* N_p *= 1000 against sample size N.*

• la variance est sensible à l'estimateur : elle est plus faible quand on utilise la méthode MV ;

• l'espérance mathématique est moins sensible à l'estimateur ; pour N_p suffisamment élevé, il tend asymptotiquement vers la valeur du vrai paramètre dans les deux cas, avec un biais plus ou moins important.

3.2 Détermination des estimateurs du paramètre de la loi L1 par simulation de Monte Carlo

Pour calculer les caractéristiques statistiques des estimateurs, nous réalisons Np simulations selon la procédure ci-dessous :

• la population parente de la variable aléatoire est connue : elle est représentée par une loi exponentielle de paramètre *a*,

• de cette population parente est extrait un échantillon de réalisations virtuelles *x* de la variable aléatoire d'effectif *N* par tirages successifs et indépendants dans une loi uniforme,

• cet échantillon est classé par ordre croissant,

• à chaque réalisation *x* de rang *k* de la variable aléatoire nous associons une fréquence empirique au non-dépassement $\hat{F}_k$ estimée par une formule de probabilité empirique générale (CUNNANE C., 1978) :

$$\hat{F}_k = \frac{k - d}{N + 1 - 2d}$$

Dans cette étude nous avons utilisé la loi de Hazen qui correspond à $d = 0.5$. L'estimation $\hat{a}$ obtenue à partir de la *ième* simulation de *N* variables aléatoires est alors calculée par l'une des deux méthodes suivantes :

• méthode MC :

$$(\hat{a}/N)_{i,MC} = -\frac{\sum_{k=1}^{N} x_{k,i} Ln\left(1 - \hat{F}_k\right)}{\sum_{k=1}^{N} \left[Ln\left(1 - \hat{F}_k\right)\right]^2} \tag{9}$$

• méthode MV :

$$(\hat{a}/N)_{i,MV} = \frac{1}{N} \sum_{k=1}^{N} x_{k,i} \tag{10}$$

où $x_{k,i}$ est la variable aléatoire de rang *k* dans un classement en ordre croissant de l'échantillon de N observations obtenues lors de la simulation d'ordre *i*.

3.3 Caractéristiques statistiques de l'estimateur des Moindres Carrés

3.3.1 Estimation de l'espérance mathématique $E(\hat{a}/N)_{MC}$

À partir de l'équation 9, la moyenne des paramètres obtenus après N_p simulations s'exprime par :

$$\left\langle \hat{a} / N \right\rangle_{MC} = -\frac{1}{N_p} \sum_{i=1}^{N_p} \left\{ \frac{\sum_{k=1}^{N} x_{k,i} Ln\left(1 - \hat{F}_k\right)}{\sum_{k=1}^{N} \left[Ln\left(1 - \hat{F}_k\right)\right]^2} \right\} \tag{11}$$

Après arrangement, cette expression peut être mise sous la forme :

$$\left\langle \hat{a} / N \right\rangle_{MC} = -\frac{\sum_{k=1}^{N} \left\langle x_k \right\rangle Ln\left(1 - \hat{F}_k\right)}{\sum_{k=1}^{N} \left[Ln\left(1 - \hat{F}_k\right)\right]^2} \tag{12}$$

avec

$$\left\langle x_k \right\rangle = \frac{1}{N_p} \sum_{i=1}^{N_p} x_{k,i} \tag{13}$$

D'après l'équation (13), $\langle x_k \rangle$ est la moyenne des observations de rang k de l'ensemble des N_p simulations.

3.3.2 Expression analytique de la variance $Var(\hat{a}_{MC}/N)$

Par définition, la variance de l'estimateur la méthode MC s'écrit sous la forme :

$$Var(\hat{a}/N)_{MC} = \left\langle \hat{a}^2/N \right\rangle_{MC} - \left\langle \hat{a} / N \right\rangle_{MC}^2 \tag{14}$$

Pour calculer le premier terme, nous élevons (11) au carré :

$$\left\langle a^2/N \right\rangle_{i,MC} = \left\{ \frac{\sum_{k=1}^{N} x_{k,i} Ln\left(1 - \hat{F}_k\right)}{\sum_{k=1}^{N} \left[Ln\left(1 - \hat{F}_k\right)\right]^2} \right\}^2$$

Le carré moyen obtenu après N_p simulations s'exprime alors par :

$$\left\langle \hat{a}^2/N \right\rangle_{MC} = \frac{1}{N_p} \sum_{k=1}^{N_p} \left\{ \frac{\sum_{k=1}^{N} x_{k,i} \, Ln\left(1 - \hat{F}_k\right)}{\sum_{k=1}^{N} \left[Ln\left(1 - \hat{F}_k\right) \right]^2} \right\}^2$$

Après développement et arrangement, nous obtenons :

$$\left\langle \hat{a}^2/N \right\rangle_{MC} = \frac{\sum_{k=1}^{N} \left\langle x_k^2 \right\rangle \left[Ln\left(1 - \hat{F}_k\right) \right]^2 + 2 \sum_{k=1}^{N-1} \sum_{l=k+1}^{N} Ln\left(1 - \hat{F}_l\right) Ln\left(1 - \hat{F}_k\right) \left\langle x_k \, x_l \right\rangle}{\left\{ \sum_{k=1}^{N} \left[Ln\left(1 - \hat{F}_k\right) \right]^2 \right\}^2} \tag{15}$$

avec $$\left\langle x_k^2 \right\rangle = \frac{1}{N_p} \sum_{i=1}^{N_p} x_{k,i}^2 \text{ et } \left\langle x_k \, x_l \right\rangle = \frac{1}{N_p} \sum_{i=1}^{N_p} x_{k,i} \, x_{l,i}$$

Pour calculer le second terme de (14), nous élevons (12) au carré. Après développement et arrangement des termes, nous obtenons :

$$\left\langle \hat{a}/N \right\rangle_{MC}^2 = \frac{\sum_{k=1}^{N} \left\langle x_k \right\rangle^2 \left[Ln\left(1 - \hat{F}_k\right) \right]^2 + 2 \sum_{k=1}^{N-1} \sum_{l=k+1}^{N} Ln\left(1 - \hat{F}_k\right) Ln\left(1 - \hat{F}_l\right) \left\langle x_l \right\rangle \left\langle x_k \right\rangle}{\left\{ \sum_{k=1}^{N} \left[Ln\left(1 - \hat{F}_k\right) \right]^2 \right\}^2} \tag{16}$$

La soustraction membre à membre des équations (15) et (16), nous conduit à :

$$Var\left(\hat{a}/N\right)_{MC} = \frac{\sum_{k=1}^{N} \left[Ln\left(1 - \hat{F}_k\right) \right]^2 Var\left(x_k\right) + \sum_{k=1}^{N-1} \sum_{l=k+1}^{N} Ln\left(1 - F_k\right) Ln\left(1 - F_l\right) Cov\left(x_k, x_l\right)}{\left\{ \sum_{k=1}^{N} \left[Ln\left(1 - \hat{F}_k\right) \right]^2 \right\}^2} \tag{17}$$

où $$Var\left(x_k\right) = \left\langle x_k^2 \right\rangle - \left\langle x_k \right\rangle^2 \tag{18}$$

et $$Cov(x_k, x_l) = \langle x_k x_l \rangle - \langle x_k \rangle \langle x_l \rangle \tag{19}$$

3.4 Caractéristiques statistiques de l'estimateur du Maximum de Vraisemblance

3.4.1 Estimation de l'espérance mathématique

D'après l'équation (10), l'estimation de l'espérance mathématique du paramètre après N_p simulations est :

$$\left\langle \hat{a}/N \right\rangle_{MV} = \frac{1}{N_p} \sum_{i=1}^{N_p} \left\{ \sum_{k=1}^{N} x_{k,i} \right\}$$

Cette expression peut être réarrangée sous la forme :

$$\left\langle \hat{a}/N \right\rangle_{MV} = \frac{1}{N} \sum_{k=1}^{N} \left\langle x_k \right\rangle \tag{20}$$

3.4.2 Estimation de la variance

Comme précédemment, la variance est définie par :

$$Var\left(\hat{a}/N\right)_{MV} = \left\langle \hat{a}^2/N \right\rangle_{MV} - \left\langle \hat{a}/N \right\rangle_{MV}^2 \tag{21}$$

Pour évaluer le premier terme, nous élevons l'équation (10) au carré, nous obtenons, pour une simulation d'ordre *i* :

$$\left\langle \hat{a}^2/N \right\rangle_{i,MV} = \left\{ \frac{1}{N} \sum_{k=1}^{N} x_{k,i} \right\}^2$$

et le carré moyen après N_p simulations devient :

$$\left\langle a^2/N \right\rangle_{MV} = \frac{1}{N_p} \sum_{i=1}^{N_p} \left\{ \frac{1}{N} \sum_{k=1}^{N_p} x_{k,i} \right\}^2$$

ce qui conduit, après développement et arrangement à :

$$\left\langle \hat{a}^2/N \right\rangle_{MV} = \frac{1}{N^2} \left\{ \sum_{k=1}^{N} \left\langle x_k^2 \right\rangle + 2 \sum_{k=1}^{N-1} \sum_{l=k+1}^{N} \left\langle x_k x_l \right\rangle \right\} \tag{22}$$

Le second terme est obtenu en élevant au carré la relation (20). Il vient après développement et arrangement :

$$\left\langle \hat{a}/N \right\rangle^2 = \frac{1}{N^2} \left\{ \sum_{k=1}^{N} \left\langle x_k \right\rangle^2 + 2 \sum_{k=1}^{N-1} \sum_{l=k+1}^{N} \left\langle x_k \right\rangle \left\langle x_l \right\rangle \right\} \tag{23}$$

En soustrayant les équations (22) et (23), nous obtenons l'expression ci-dessous de la variance de l'estimateur de la méthode MV :

$$Var\left(\hat{a}/N\right)_{MV} = \frac{1}{N^2}\left\{\sum_{k=1}^{N} var\left(x_k\right) + 2\sum_{k=1}^{N-1}\sum_{l=1}^{N} cov\left(x_k, x_l\right)\right\} \qquad (24)$$

où $Var(x_k)$ et $Cov(x_k,x_l)$ ont été respectivement définis par les équations (18) et (19).

4 - COMPARAISON DES EXPRESSIONS ANALYTIQUES DES ESTIMATEURS

D'après les équations (12), (17), (20) et (24), les caractéristiques statistiques des estimateurs dépendent des caractéristiques statistiques moyenne, variance - covariance des observations de rang k (dans un classement en ordre croissant) obtenus au bout de N_p simulations.

Pour N_p = 1000, N = 100 et a = 1, nous représentons aux figures 5a) et 5b) l'évolution de la moyenne $\langle x_k \rangle$, et de la variance $var\{x_k\}$ de x_k. Dans les deux cas, la moyenne et la variance augmentent avec *k*.

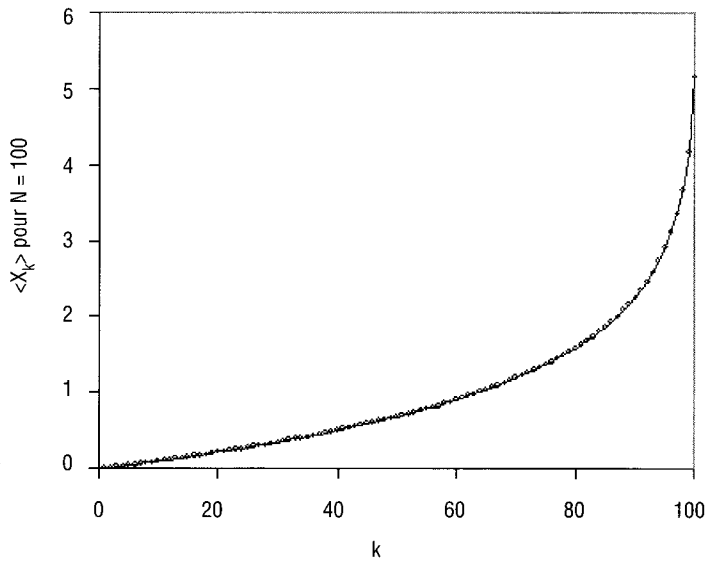

Figure 5a $\langle x_k \rangle$ en fonction de *k* pour N = 100.
$\langle x_k \rangle$ *as a function of* k *for N = 100.*

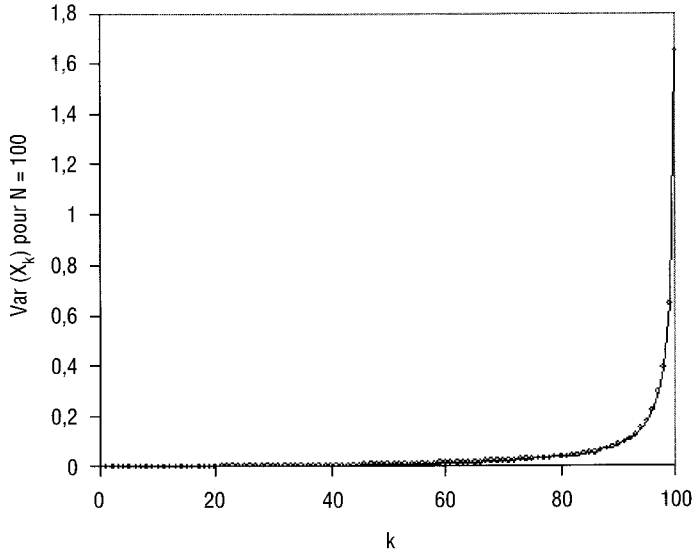

Figure 5b *Var*(x_k) en fonction de *k* pour N = 100.
Var(x_k) *as a function of* k *for N = 100.*

4.1 Comparaison de la moyenne des estimateurs

Nous écrivons l'espérance mathématique des deux estimateurs respectivement sous la forme ci-dessous :

4.1.1 Méthode MC

$$\left\langle \hat{a}/N \right\rangle_{MC} = \sum_{k=1}^{N} FMC(k)\left\langle x_k \right\rangle \tag{25}$$

où

$$FMC(k) = \left[-\frac{Ln\left(1 - \hat{F}_k\right)}{\sum_{k=1}^{N}\left[Ln\left(1 - \hat{F}_k\right)\right]^2} \right] \tag{26}$$

4.1.2 Méthode MV

$$\left\langle \hat{a}/N \right\rangle_{MV} = \sum_{k=1}^{N} FMV(k)\left\langle x_k \right\rangle \tag{27}$$

où

$$FMV(k) = \left[\frac{1}{N} \right] \tag{28}$$

D'après les équations (25) et (27) les moyennes des deux estimateurs diffèrent du coefficient de $\langle x_k \rangle$, qui est selon l'estimateur utilisé *FMC*(*k*) pour la méthode MC et *FMV*(*k*) pour la méthode MV. Pour N fixé, ce coefficient dépend du rang *k* pour la méthode MC ; il est constant pour la méthode MV. Comparer la moyenne des deux estimateurs revient donc à comparer ces coefficients. Nous avons posé N_p = 1000, N = 100, et a = 1, et simulé N_p échantillons de variables aléatoires d'effectif N issues d'une loi exponentielle de paramètre a. Ces échantillons ont été classés par ordre croissant. La moyenne $\langle x_k \rangle$ des réalisations x_k de rang *k*, les coefficients *FMV*(*k*), *FMC*(*k*) et les produits $FMV(k)^* \langle x_k \rangle$ et $FMC(k)^* \langle x_k \rangle$ ont été calculés et représentés en fonction de *k,* figure (6a) et figure (6b). Nous notons sur la figure (6a) que *FMV*(*k*) est constant alors que *FMC*(*k*) varie avec *k* d'une part, et que d'autre part *FMV*(*k*) > *FMC*(*k*) sauf pour les dernières valeurs ($k \geq 87$), pour lesquelles l'inégalité change de sens. C'est ce que nous observons sur la figure (6b) où $FMV(k)^* \langle x_k \rangle > FMC(k)^* \langle x_k \rangle$ pour $k \leq 87$. Ce changement de signe de l'écart nous a amené à représenter l'évolution en fonction du rang *k* du cumul des produits $\sum_{j=1}^{k} FMC(j)^* \langle x_j \rangle$ et $\sum_{j=1}^{k} FMV(j)^* \langle x_j \rangle$ pour *k* = 1,...,100 sur la figure (6c). Au fur et à mesure que *k* croît, les deux courbes tendent vers la même valeur qui est le paramètre réel *a* = 1 ; la valeur finale est cependant plus proche du vrai paramètre pour la méthode MV. C'est ce que nous observons sur la figure 4 représentant l'évolution de $\langle \hat{a}/N \rangle_{MC}$ et $\langle \hat{a}/N \rangle_{MV}$ avec N_p=1000, pour quelques valeurs de N allant de 5 à 100.

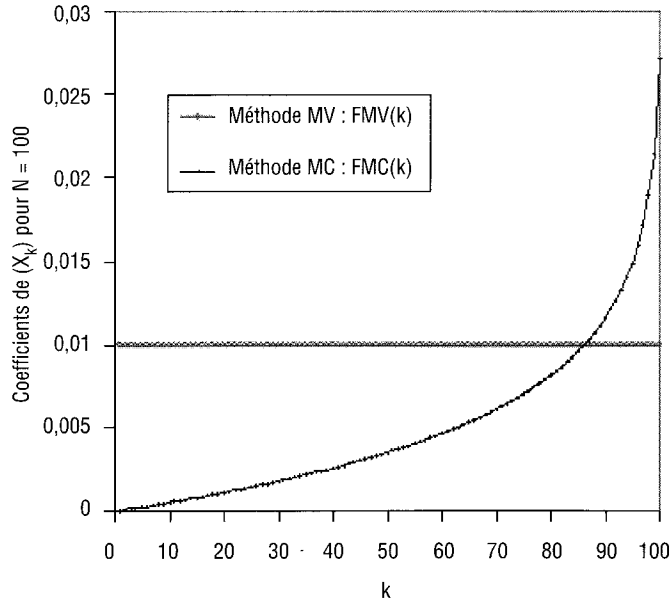

Figure 6a Coefficients de $\langle x_k \rangle$ en fonction de *k* pour N = 100.
Coefficients of $\langle x_k \rangle$ *against* k *for N = 100.*

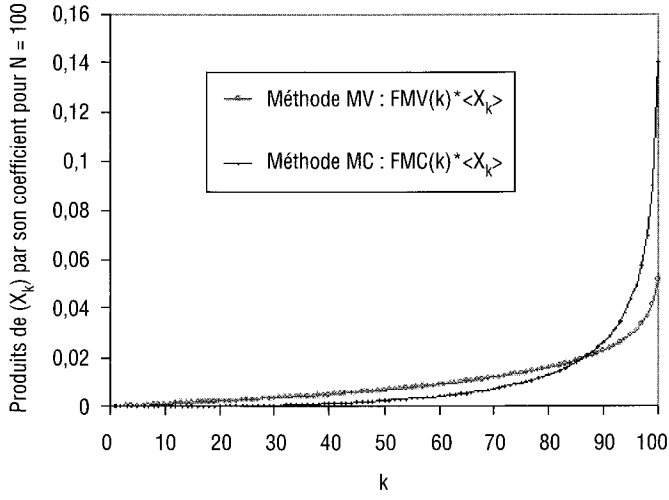

Figure 6b Produit de <x_k> par son coefficient en fonction de *k* pour N = 100.
Product <x_k> *and its coefficient as a function of* k *for N = 100.*

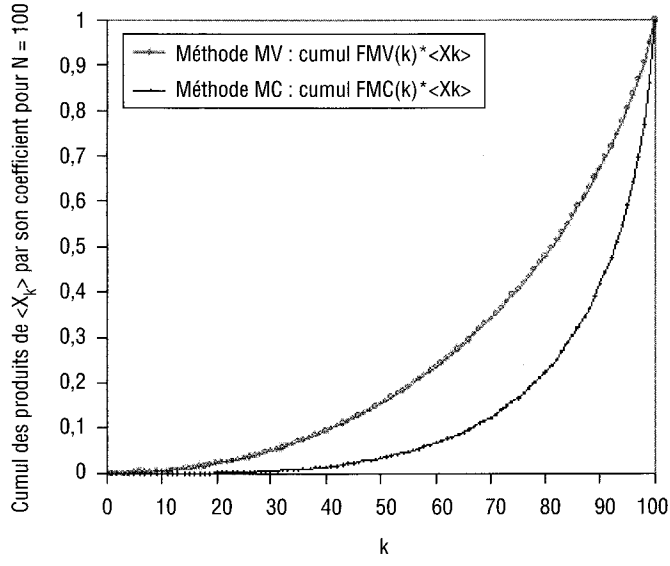

Figure 6c Cumul des produits de <x_k> par son coefficient en fonction de *k* pour N = 100.
Sum of products of <x_k> *and its coefficient as a function of* k *for N = 100.*

4.2 Comparaison des expressions analytiques de la variance selon l'estimateur

Nous mettons également la variance des deux estimateurs sous la forme :

4.2.1 Méthode MC

$$Var(\hat{a}/N)_{MC}=\sum_{k=1}^{N} FMC^{2}(k)^{*}Var(x_k)+2\sum_{k=1}^{N-1}\sum_{l=k+1}^{N} FMC(k)^{*}FMC(l)^{*}Cov(x_k,x_l) \quad (29)$$

4.2.2 Méthode MV

$$Var(\hat{a}/N)_{MV}=\sum_{k=1}^{N} FMV^{2}(k)Var(x_k)+2\sum_{k=1}^{N-1}\sum_{l=k+1}^{N} FMV(k)^{*}FMV(l)^{*}Cov(x_k,x_l) \quad (30)$$

De l'examen des équations (29) et (30), il ressort que les variance des deux estimateurs dépendent de la variance de l'observation de rang x_k, $Var(x_k)$, et de la covariance des observations de rangs k et l, x_k et x_l, $Cov(x_k,x_l)$. Elles diffèrent des coefficients de ces quantités, qui sont :

- $FMV^2(k)$ ou $FMC^2(k)$ pour la variance, selon que selon que l'on utilise la méthode MV ou la méthode MC ;
- $FMC(k)^*FMC(l)$ ou $FMV(k)^*FMV(l)$ pour la covariance, selon que l'on utilise la méthode MC ou MV.

Ce coefficient est dans tous les cas constant quand on utilise la méthode MV ; il varie en fonction de k quand on utilise la méthode MC.

Pour comparer les variances des deux estimateurs, nous générons comme précédemment, par simulation de Monte Carlo Np = 1000 échantillons de N = 100 réalisations virtuelles de variables aléatoires issues d'une loi exponentielle à un seul paramètre connu a = 1 ; pour chaque simulation, les observations sont classées en ordre croissant.

Nous calculons et représentons l'évolution en fonction de k :

- des coefficients de $Var(x_k)$, $FMC^2(k)$ pour la méthode MV et $FMV^2(k)$ pour la méthode M, figure (7a) ;
- les produits $FMV^2(k)^*Var(x_k)$ pour la méthode MV et $FMC^2(k)^*Var(x_k)$, pour la méthode MC, figure (7b) ;
- la covariance entre les observations de rang k et l, $Cov(x_k,x_l)$ pour k = 1, et l = 2, ...,100, figure (7c) ;
- les produits $FMC(k)^*FMC(l)^*Cov(x_k,x_l)$ et $FMV(k)^*FMV(l)^*Cov(x_k,x_l)$ pour k = 1 et l = 2, ...,100, figure (7c) et figure (7d).

Nous notons que :

- pour $k < 87$, nous avons toujours $FMV^2(k) > FMC^2(k)$ Figure (7a) et $FMV^2(k)^*Var(x_k) > FMC^2(k)^*Var(x_k)$ figure (7b) ; l'inégalité change de sens dans le cas contraire où $k \geq 87$ et l'amplitude des écarts devient alors beaucoup plus important ; l'effet de ces derniers termes doit être prépondérant sur la valeur numérique de la variance du paramètre selon l'estimateur.

• pour $k = 1$, et $l = 2, \ldots,100$, $FMC(k)*FMC(l)*Cov(x_k,x_l)$ est en valeur absolue plus élevé que $FMV(k)*FMV(l)*Cov(x_k,x_l)$, figure (7d).

Pour juger de l'effet des derniers termes, nous avons poussé l'analyse en étudiant l'évolution des cumuls :

• $\sum_{j=1}^{k} FMC^2(j)*Var(x_j)$ et $\sum_{j=1}^{k} FMV^2(j)*Var(x_j)$ pour $k = 1, \ldots,N$ (figure 7(e))

• $\sum_{j=1}^{k}\sum_{l=k+1}^{N} FMC(j)*FMC(l)*Cov(x_j,x_l)$ et $\sum_{j=1}^{k}\sum_{l=k+1}^{N} FMV(j)*FMV(l)*Cov(x_j,x_l)$

pour $k = 1, \ldots, N - 1$. Figure 7(f)).

Les figures 7 e) et 7 f) mettent en évidence le poids des derniers termes sur le résultat final : les cumuls relatifs à la méthode MC sont supérieurs à ceux relatifs à la méthode MV jusqu'aux dernières valeurs de k pour lesquelles la tendance s'inverse. Ce sont donc les derniers termes du cumul qui sont déterminent le résultat final, et qui expliquent que la variance soit plus élevée pour la méthode MC que pour la méthode MV : ces derniers sont plus élevés quand on utilise la méthode MC. Nous avons comparé sur la figure 7g) les cumuls des termes en $Var(x_k)$ et des termes en $Cov(x_k,x_l)$. Il ressort de la comparaison que ce sont les derniers qui imposent le résultat final.

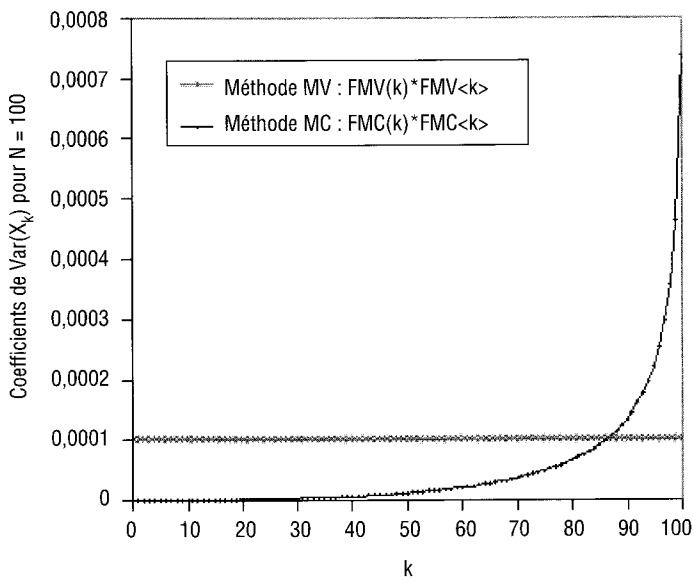

Figure 7a Coefficients de $Var<x_k>$ en fonction de k pour N = 100.
Coefficients of $Var<x_k>$ as a function of k *for N = 100.*

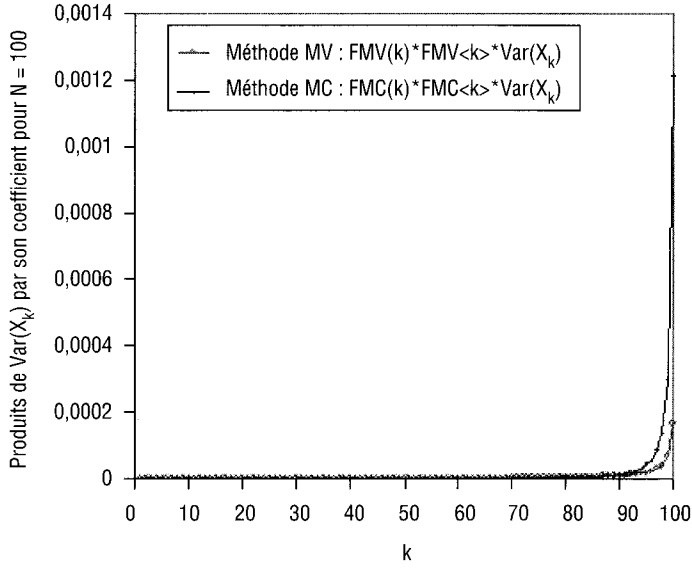

Figure 7b Produits *Var<x_k>* par son coefficient en fonction de *k* pour N = 100.
Product of Var<x_k> and its coefficient as a function of k *for N = 100.*

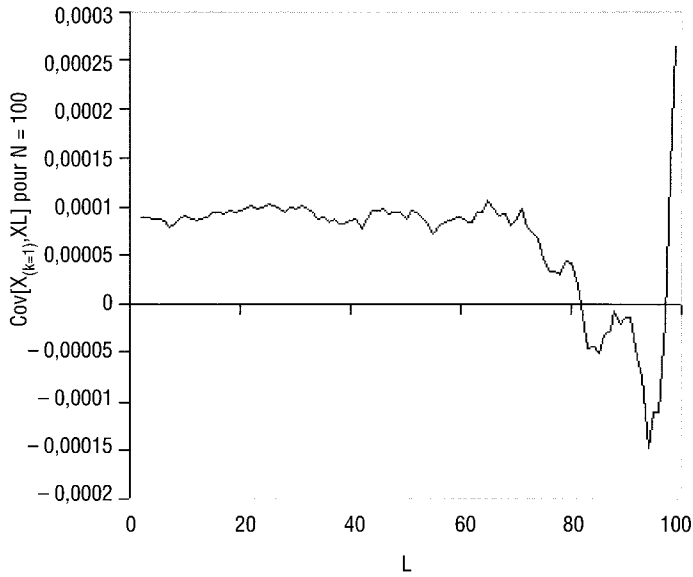

Figure 7c *Cov<$x_{k=1}$,x_l>* pour N = 100.
Cov<$x_{k=1}$,x_l> *pour N = 100.*

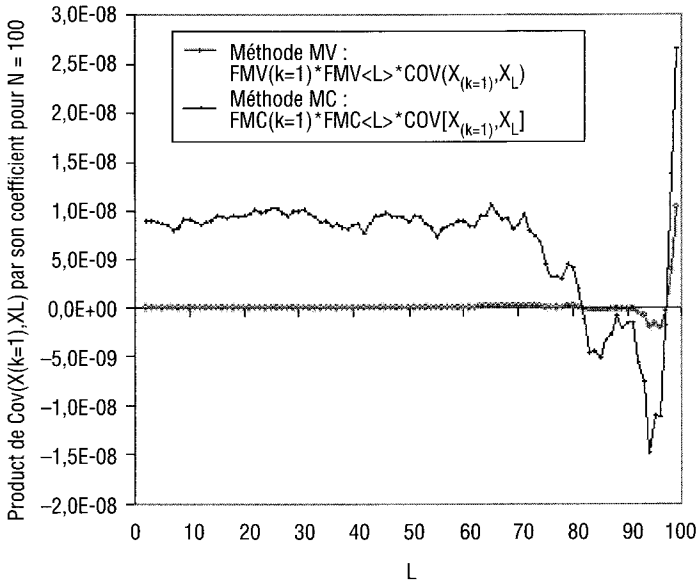

Figure 7d Cumul du produit de $Var<x_k>$ par son coefficient pour N = 100.
Sum of product of $Var<x_k>$ and its coefficient for N = 100.

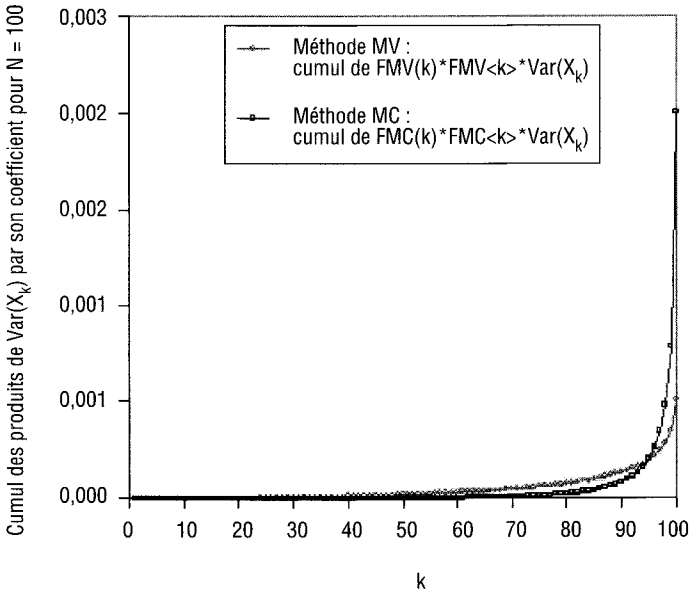

Figure 7e Produit de $Var<x_k>$ par son coefficient pour N = 100.
Product of $Var<x_k>$ and its coefficient for N = 100.

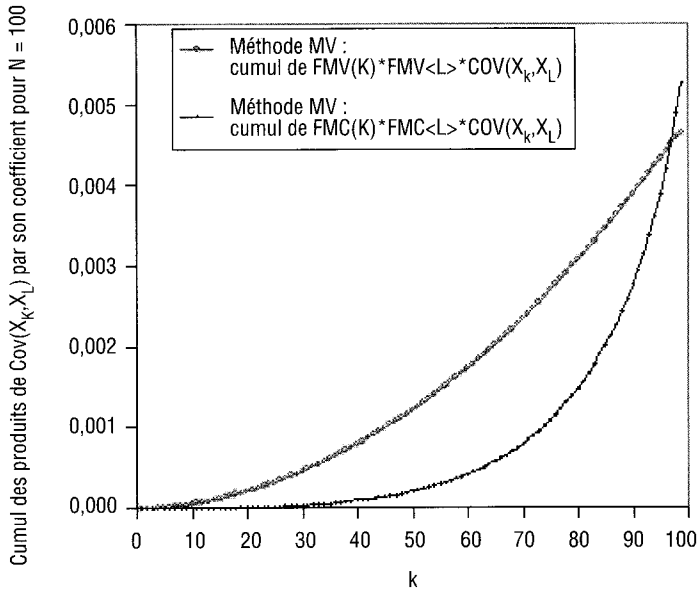

Figure 7f Cumul du produit de $Cov(x_k,x_l)$ par son coefficient en fonction de k pour N = 100.

Sum of product of coefficient of $Cov(x_k,x_l)$ *and its coefficient as a function of k for N = 100.*

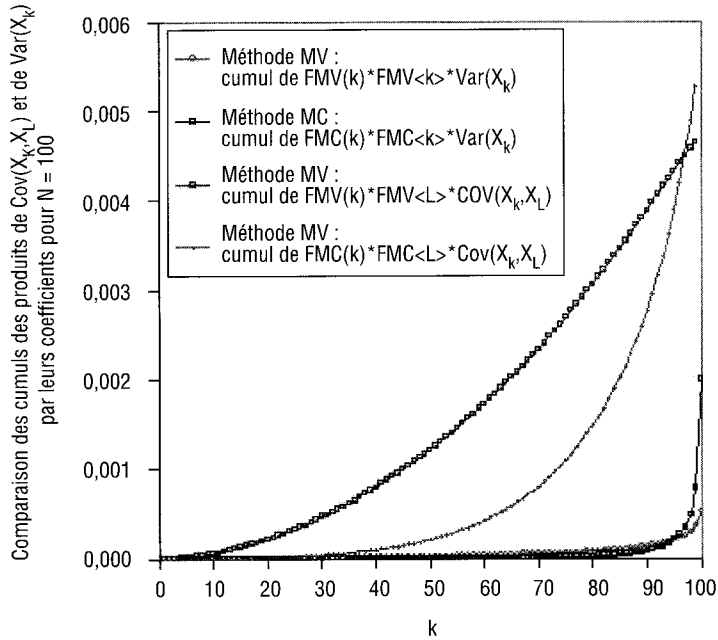

Figure 7g Cumul du produit de $Var(x_k)$ et de $Cov(x_k,x_l)$ par leurs coefficients pour N = 100.

Sum of product of $Var(x_k)$ and $Cov(x_k,x_l)$ *and their coefficient for N = 100.*

5 - SYNTHÈSE DES ESSAIS

La simulation numérique de Monte Carlo a d'abord été utilisée pour comparer les caractéristiques statistiques de deux estimateurs du paramètre *a* de la loi exponentielle à un paramètre : la méthode MV et la méthode MC. L'application a été faite pour $a = 1$. L'analyse des graphes obtenus montre que :

- quelle que soit la méthode utilisée, pour un nombre suffisamment élevé de simulation *Np*, l'espérance mathématique de l'estimateur tend vers la vraie valeur du paramètre avec un biais plus ou moins important provenant des erreurs d'échantillonnage ;

- la variance des estimateurs ne dépend pas du nombre Np des simulations, mais de l'effectif N des échantillons de variables aléatoires. Elle décroît en fonction de l'effectif *N* des échantillons d'observations virtuelles pour les deux méthodes. Elle est plus faible quand on utilise la méthode MV.

L'expression analytique des estimateurs a été dérivée pour des échantillons de réalisations virtuelles d'observations classés en ordre croissant. L'examen des expressions obtenues montre que ces caractéristiques statistiques dépendent de celles des observations de rang *k*, x_k, moyenne $<x_k>$, variance $Var(x_k)$ et covariance $Cov(x_k,x_l)$ et diffèrent des coefficients de ces quantités.

Pour *N* fixé, ce coefficient est constant quand on utilise la méthode MV et varie avec le rang *k* de x_k quand on utilise la méthode MC. Ce coefficient influence les valeurs des caractéristiques statistiques des estimateurs, particulièrement la variance. Une application faite pour $a = 1$ montre que :

- l'espérance mathématique présente pour les deux méthodes (MV et MC) un biais initial qui décroît très vite quand N augmente ; elle tend vers la vraie valeur du paramètre dans les deux cas, avec un biais du aux erreurs d'échantillonnage ;

- la variance des deux estimateurs décroît quand N augmente et reste plus faible quand on utilise la méthode MV. Ce dernier provient de la prépondérance des termes en $Cov(x_k,x_l)$ et de l'influence des derniers termes des cumuls des produits. Nous retrouvons ainsi un résultat déjà connu pour la loi de Gumbel (LOWERY et NASH, 1970).

6 - CONCLUSION

Deux méthodes d'estimation sont généralement utilisées pour calculer le paramètre de la loi exponentielle à un seul paramètre : la méthode du maximum de vraisemblance (méthode MV) et la méthode des moments (méthode M). Ces méthodes fournissent une estimation triviale du paramètre. La méthode des moindres carrés (méthode MC) est très peu utilisée. Dans cet article nous avons comparé les caractéristiques statistiques (moyenne et variance) de deux estimateurs : méthode MV et méthode MC, empiriquement par simulation de Monte Carlo, théoriquement après dérivation des expres-

sions analytiques de ces caractéristiques statistiques à partir de la statistique des rangs. L'étude par simulation de Monte Carlo a montré que pour un nombre suffisamment élevé de simulations, l'espérance mathématique des deux estimateurs tend vers le vrai paramètre, alors que la variance est plus faible quand on utilise la méthode MV. Ce dernier résultat provient de l'effet des coefficients de $Var(x_k)$ et $Cov(x_k,x_l)$ du poids relative des termes en $Cov(x_k,x_l)$ dans l'expression analytique de la variance selon l'estimateur. LOWERY et NASH, (1970) ont abouti à une conclusion similaire pour la loi de Gumbel.

REMERCIEMENTS

L'auteur remercie les réviseurs anonymes pour leurs remarques et suggestions qui ont largement contribué à faire de l'article ce qu'il est.

RÉFÉRENCES BIBLIOGRAPHIQUES

ARNAUD P., LAVABRE J. (1999) : Using a stochastic model for generating hourly hyetographs to study extreme rainfall. *Hydrol. Sci. J.* 44 (3) : 433-446.

CUNNANE C.,1978. Unbiaised plotting positions. *J. of Hydrology*, 37, 205-222.

CHAPMAN T.G. (1997) Stochastic models for daily rainfall in the Western Pacific. *Mathematic and computer simulation*. 43 (3-6) : 351-358.

GALEA G., PRUDHOMME C. (1997) Notions de base et concepts utiles pour la compréhension de la modélisation synthétique des régimes de crue des bassins versants au sens des modèles QdF. *Rev. Sci. Eau* 10 (1) : 64-101.

HØYBYE J., LÀSLZLÒ I. (1997) Analysis of extreme hydrological events in a moonsoon climate catchment : the Hongru River, China. *Hydrol. Sci. J.* 42 (3) : 343-356.

KABAILI Z. (1983) Contribution à l'étude statistique de la pluie dans la région de Tunis. Thèse présentée à l'INPT en vue d'obtenir le grade de Docteur-Ingénieur. 223 pages.

KJELDSEN T.R., LUNDORF A., ROSBERG A.(2000) Use of a two-component exponentiel distribution in partial duration modelling of hydrological droughts in Zimbabwean rivers. *Hydrol. Sci. J.* 45 (2) : 285-298.

LANG M., OUARDA T., BOBEE B., 1999. Towards operational guidelines for overthreshold modeling. *J. Hydrol.* 225,103-117.

LEBEL T., AMANI A., CAZENAVE F., LECOCQ J., TAUPIN J.D., ELGUERO E., GERARD M., LOWERY M.D., NASH J.E., 1970. A comparison of methods of fitting the double exponential distribution. *J. Hydrol.*, 10, 259-275.

MATHIEU L., PERREAULT L., BOBEE B., ASHKAR F. (1991) : Frequence analysis of water : Duration and severity. Computer Methods in Water Resources II. Vol. 2. *Computational hydraulics and hydrology proceedings*. (Proceeding of the 2e international Conference held in Marrakesh, Morocco, 20-22 February 1991). Computational Mechanics Publications. Springer Verlag 141-156.

PANTOGLOU G., TZIAFETAS G. (1989) : Stochastic models for rainfall in the greater Athens Area. *Zeitschrift für Meteorologie* 39 (35) : 273-277.