

REVUE DE STATISTIQUE APPLIQUÉE

W. KRYSICKI

Application de la méthode des moments à l'estimation des paramètres d'un mélange de deux distributions de Rayleigh

Revue de statistique appliquée, tome 11, n° 4 (1963), p. 25-34

http://www.numdam.org/item?id=RSA_1963__11_4_25_0

© Société française de statistique, 1963, tous droits réservés.

L'accès aux archives de la revue « *Revue de statistique appliquée* » (<http://www.sfds.asso.fr/publicat/rsa.htm>) implique l'accord avec les conditions générales d'utilisation (<http://www.numdam.org/conditions>). Toute utilisation commerciale ou impression systématique est constitutive d'une infraction pénale. Toute copie ou impression de ce fichier doit contenir la présente mention de copyright.

NUMDAM

Article numérisé dans le cadre du programme
Numérisation de documents anciens mathématiques
<http://www.numdam.org/>

APPLICATION DE LA MÉTHODE DES MOMENTS A L'ESTIMATION DES PARAMÈTRES D'UN MÉLANGE DE DEUX DISTRIBUTIONS DE RAYLEIGH

W. KRYSICKI

Professeur à l'École Polytechnique de LODZ (Pologne)

I - INTRODUCTION

Comme on le sait, ce n'est que rarement, dans les applications pratiques, que l'on rencontre une population générale homogène. Beaucoup plus souvent apparaissent des populations hétérogènes, formées par le mélange de deux, et parfois d'un plus grand nombre, de populations dans lesquelles la caractéristique étudiée est soumise à la même distribution d'un type de densité connu $f(x, a, b, \dots)$ mais avec diverses valeurs des mêmes paramètres a, b, \dots en nombre fini. En désignant par p le pourcentage de la première des populations partielles et nous limitant au cas d'une population générale composée de deux populations partielles, nous obtiendrons la formule de la densité d'un mélange de deux distributions :

$$p f(x, a_1, b_1, \dots) + (1 - p) f(x, a_2, b_2, \dots) \quad (1.1)$$

En 1894 K. Pearson [1]⁽¹⁾ a publié le premier ouvrage sur ce sujet, ouvrage concernant l'estimation des paramètres du mélange de deux distributions normales.

Dans son travail, il étudiait la fonction :

$$y = \frac{n_1}{\sqrt{2\pi} \sigma_1} e^{-\frac{(x-b_1)^2}{2\sigma_1^2}} + \frac{n_2}{\sqrt{2\pi} \sigma_2} e^{-\frac{(x-b_2)^2}{2\sigma_2^2}} \quad (1.2)$$

où $n_1 + n_2 = N$ désignait le nombre total des observations ; il est évident qu'il suffit de poser $n_1/N = p$ pour que la droite de la formule (1.2) prenne l'aspect (1.1).

Gumbel [2], Mendenhall et Hader [3], ainsi que Ridder [4] ont publié d'autres travaux dans ce domaine.

II - PRESENTATION DU PROBLEME

Dans le présent travail nous adoptons la distribution de Rayleigh comme distribution paraissant dans la formule (1.1) ; la densité est alors définie par la formule

(1) Ces numéros correspondent aux références bibliographiques.

$$f(x, k) = \frac{x}{k} \exp\left(-\frac{x^2}{2k}\right), \quad 0 < x < +\infty, \quad k > 0 \quad (2.1)$$

Il est facile de démontrer que cette densité atteint un maximum pour $x = \sqrt{k}$ dont la valeur est :

$$f(\sqrt{k}) = \frac{1}{\sqrt{k}} \exp\left(-\frac{1}{2}\right) \approx \frac{0,6065}{\sqrt{k}}$$

Ainsi la dominante de cette distribution est $x = \sqrt{k}$.

Pour établir la médiane M_e de la distribution il faut résoudre l'équation

$$\frac{1}{k} \int_0^{M_e} x \exp\left(-\frac{x^2}{2k}\right) dx = 0,5$$

d'où nous obtenons

$$M_e = \sqrt{2k \text{ Log } 2}$$

La courbe de la densité possède un point d'inflexion

$$x = \sqrt{3k}$$

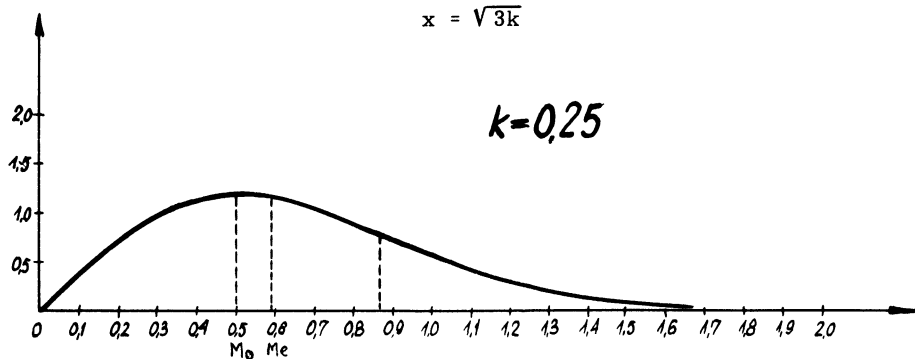


Fig. 1.

La figure (1) représente la courbe de densité de Rayleigh pour $k = 0,25$ où l'on a indiqué la mode M_0 , la médiane M_e , ainsi que l'abscisse du point d'inflexion. La figure (2) correspond à $k = 0,0625$.

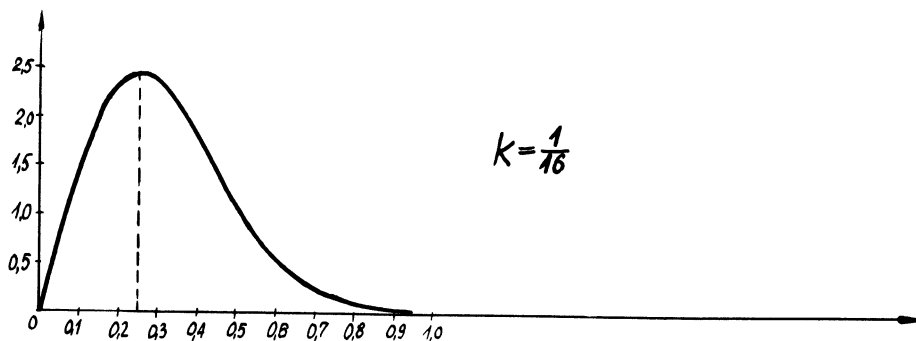


Fig. 2.

Conformément à (1.1), la densité du mélange de deux distributions de Rayleigh dans la proportion $p : (1 - p)$ est exprimée par la formule (2.2) :

$$\frac{p}{k_1} x \exp\left(-\frac{x^2}{2k_1}\right) + \frac{1-p}{k_2} x \exp\left(-\frac{x^2}{2k_2}\right) \quad (2.2)$$

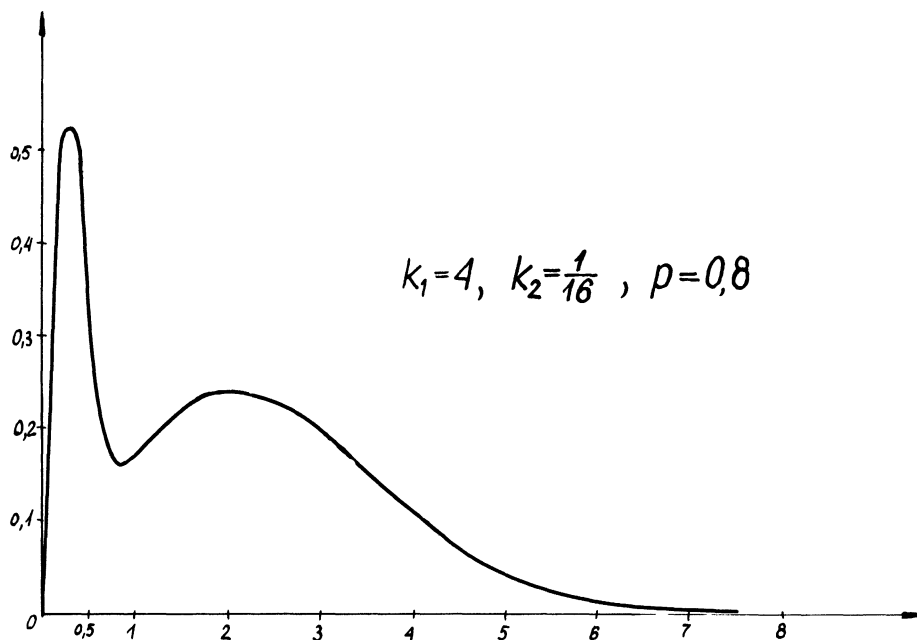


Fig. 3.

Les figures 3 et 4 donnent les densités du mélange de deux distributions de Rayleigh pour diverses valeurs des paramètres.

Les moments simples de la distribution de Rayleigh (2.1) donnés par la formule

$$\alpha_m = \frac{1}{k} \int_0^{\infty} x^{m+1} \exp\left(-\frac{x^2}{2k}\right) dx ,$$

peuvent être calculés facilement en exprimant l'intégrale ci-dessus par la fonction gamma.

Ainsi dans le cas où m est un nombre pair nous obtenons, après quelques transformations :

$$\alpha_{2l} = (2k)^l \Gamma(l + 1) \quad l = 1, 2, 3, \dots \quad (2.3)$$

et dans le cas où m est un nombre impair :

$$\alpha_{2l-1} = (2k)^{\frac{l-1}{2}} \Gamma\left(l + \frac{1}{2}\right) \quad l = 1, 2, 3, \dots \quad (2.4)$$

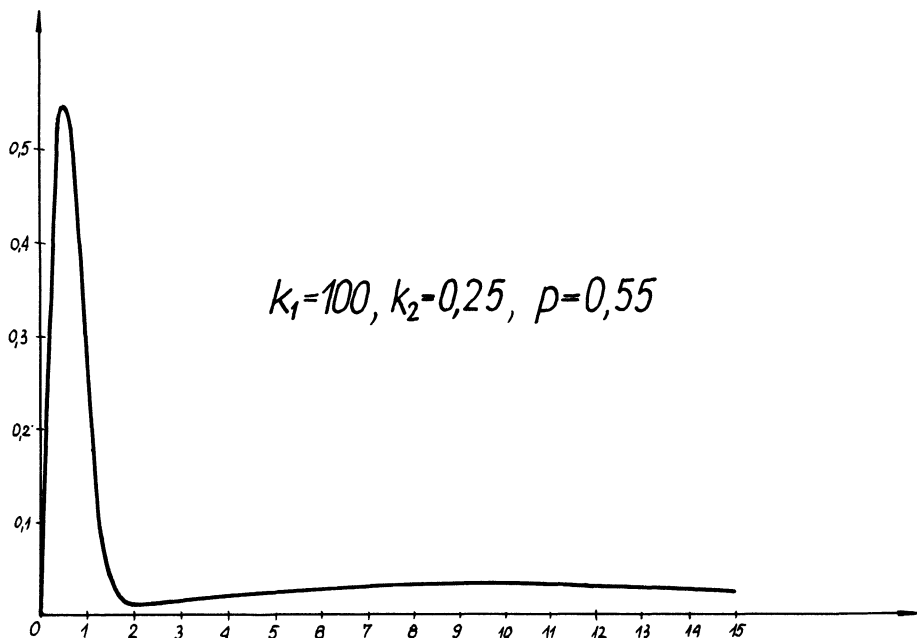


Fig. 4

Supposons maintenant que l'on a pris un échantillon aléatoire de n éléments de la population générale formée du mélange de deux distributions de Rayleigh dont la densité est donnée par la formule (2.2). Les résultats étant

$$x_1, x_2, \dots, x_n \quad (2.5)$$

A partir des résultats (2.5) nous calculons les moments simples de l'échantillon et il s'agit d'estimer sur cette base les paramètres inconnus du mélange (2.2).

Nous allons considérer deux cas selon que le rapport $p : (1 - p)$ est connu ou non.

III - 1er CAS : p ET LES PARAMETRES b_1 ET b_2 INCONNUS

Dans ce cas, pour estimer les trois paramètres p, b_1, b_2 nous utiliserons les trois premiers moments simples calculés à l'aide de l'échantillon aléatoire (2.5)

$$a_1 = \frac{1}{n} \sum_1^n x_i ; a_2 = \frac{1}{n} \sum_1^n x_i^2 ; a_3 = \frac{1}{n} \sum_1^n x_i^3 \quad (3.1)$$

Désignons les estimateurs des paramètres inconnus p, k_1, k_2 , calculés à partir des trois premiers moments simples (3.1) par $\hat{p}, \hat{k}_1, \hat{k}_2$, alors des formules (2.3), (2.4), (3.1) nous obtiendrons :

$$\begin{aligned}
\hat{\rho} \sqrt{\hat{k}_1} \sqrt{\frac{\pi}{2}} + (1 - \hat{\rho}) \sqrt{\hat{k}_2} \sqrt{\frac{\pi}{2}} &= a_1 \\
2\hat{\rho} \hat{k}_1 + 2(1 - \hat{\rho}) \hat{k}_2 &= a_2 \\
3\hat{\rho} \hat{k}_1^{3/2} \sqrt{\frac{\pi}{2}} + 3(1 - \hat{\rho}) \hat{k}_2^{3/2} \sqrt{\frac{\pi}{2}} &= a_3
\end{aligned} \tag{3.2}$$

En introduisant des inconnues auxiliaires définies par la formule

$$\sqrt{\hat{k}_i} = u_i, \quad i = 1, 2 \tag{3.3}$$

nous allons ramener le système d'équations (3.2) à un système de trois équations rationnelles

$$\begin{aligned}
\hat{\rho} u_1 + (1 - \hat{\rho}) u_2 &= \sqrt{\frac{2}{\pi}} a_1 \\
\hat{\rho} u_1^2 + (1 - \hat{\rho}) u_2^2 &= \frac{1}{2} a_2 \\
\hat{\rho} u_1^3 + (1 - \hat{\rho}) u_2^3 &= \frac{1}{3} \sqrt{\frac{2}{\pi}} a_3
\end{aligned} \tag{3.4}$$

Tirons $\hat{\rho}$ de la première des équations (3.4) et portons cette valeur dans la seconde et dans la troisième de ces mêmes équations ; dans ce cas on peut présenter ce système sous la forme :

$$\begin{aligned}
a_1 \sqrt{\frac{2}{\pi}} (u_1 + u_2) - u_1 u_2 &= \frac{1}{2} a_2 \\
a_1 \sqrt{\frac{2}{\pi}} (u_1^2 + u_1 u_2 + u_2^2) - u_1 u_2 (u_1 + u_2) &= \frac{1}{3} \sqrt{\frac{2}{\pi}} a_3
\end{aligned} \tag{3.5}$$

Introduisons encore une fois de nouvelles inconnues

$$\begin{aligned}
u_1 + u_2 &= v \\
u_1 u_2 &= w
\end{aligned} \tag{3.6}$$

Alors le système (3.5) prend la forme

$$\begin{aligned}
c_1 v - w &= c_2 \\
c_1 (v^2 - w) - vw &= c_3
\end{aligned} \tag{3.7}$$

où l'on a substitué

$$a_1 \sqrt{\frac{2}{\pi}} = c_1, \quad \frac{1}{2} a_2 = c_2, \quad \frac{1}{3} \sqrt{\frac{2}{\pi}} a_3 = c_3 \tag{3.8}$$

En résolvant le système (3.7) nous obtenons

$$v = \frac{c_3 - c_1 c_2}{c_2 - c_1^2}, \quad w = \frac{c_1 c_3 - c_2^2}{c_2 - c_1^2} \tag{3.9}$$

Pour plus de commodité introduisons les notations

$$c_3 - c_1 c_2 = A, \quad c_2 - c_1^2 = M, \quad c_1 c_3 - c_2^2 = B \quad (3.10)$$

On aura donc

$$v = \frac{A}{M}, \quad w = \frac{B}{M} \quad (3.11)$$

Pour obtenir les estimateurs recherchés $\hat{k}_i = u_i^2$, $i = 1, 2$, à partir des formules (3.6) et (3.11), nous calculons

$$\hat{k}_1 + \hat{k}_2 = \frac{A^2 - 2BM}{M^2} \quad (3.12)$$

$$\hat{k}_1 \hat{k}_2 = \frac{B^2}{M^2}$$

et nous obtenons l'équation

$$M^2 k^2 - (A^2 - 2BM) k + B^2 = 0 \quad (3.13)$$

Nous basant sur (3.10) et (3.8), nous obtenons finalement une équation dont les racines constituent les estimateurs recherchés \hat{k}_i , $i = 1, 2$, des paramètres inconnus k_i :

$$\left(a_2 - \frac{4}{\pi} a_1^2 \right)^2 k^2 - \left[\frac{8}{\pi} \left(\frac{1}{3} a_3 - \frac{1}{2} a_1 a_2 \right)^2 - 4 \left(a_2 - \frac{4}{\pi} a_1^2 \right) \left(\frac{2}{3\pi} a_1 a_3 - \frac{1}{4} a_2^2 \right) \right] k + 4 \left(\frac{2}{3\pi} a_1 a_3 - \frac{1}{4} a_2^2 \right)^2 = 0 \quad (3.14)$$

Il est évident qu'en calculant les racines de l'équation (3.14) en fonction des moments a_1 , a_2 , a_3 , calculés à partir de l'échantillon, nous pouvons obtenir soit des racines complexes conjuguées, soit des racines réelles qui ne seront pas toutes les deux positives. Cela peut prouver que notre hypothèse que la population étudiée a été formée du mélange de deux distributions de Rayleigh n'est pas vraie, mais cela peut aussi arriver lorsque l'échantillon ne contient pas de nombreux éléments bien que notre hypothèse soit vraie.

Pour des échantillons de grande taille avec une probabilité égale à 1, en admettant que notre hypothèse est vraie, nous aurons les inégalités (3.15), ci-après.

Dans le cas où l'équation admettra deux racines positives, il reste la question de savoir laquelle des deux racines de l'équation (3.14) est l'estimateur \hat{k}_1 et laquelle l'estimateur \hat{k}_2 ?

Nous allons montrer que c'est indifférent : il résulte de la forme des équations du système (3.2) que l'échange simultané qu'on y fait de \hat{k}_1 et \hat{k}_2 d'une part et de p et $1 - p$ d'autre part ne change aucune des équations du système (3.2).

Prenons donc l'une des racines de l'équation (3.14) comme \hat{k}_1 , par exemple, la plus petite, que nous désignerons par k' , par là même nous considérons la seconde racine k'' comme \hat{k}_2 et calculons \hat{p} par exemple de la première des équations (3.2).

$$\hat{p} = \frac{a_1 \sqrt{\frac{2}{\pi}} - \sqrt{\hat{k}_2}}{\sqrt{\hat{k}_1} - \sqrt{\hat{k}_2}}$$

Si cependant au début nous avons adopté le contraire, c'est-à-dire $k' = \hat{k}_2$, $k'' = \hat{k}_1$, le second membre de l'égalité précédente représenterait alors $1 - \hat{p}$.

Les estimateurs \hat{k}_1 , \hat{k}_2 ainsi obtenus des paramètres inconnus k_1 , k_2 sont fonction des moments (a_1 , a_2 , a_3) de l'échantillon. Comme la valeur moyenne du moment a_i de l'échantillon est égale au moment α_i de la population générale, il résulte donc du théorème de Khintchine que le moment a_i dans l'échantillon, pour $n \rightarrow \infty$, converge en probabilité vers le moment α_i dans la population, si ce dernier existe, évidemment.

Si donc $k_1 \neq k_2$, de la continuité des estimateurs \hat{k}_1 et \hat{k}_2 au point $(\alpha_1, \alpha_2, \alpha_3)$ considérés comme fonctions des (a_1, a_2, a_3) , il résulte qu'ils sont des estimateurs convergents pour des paramètres inconnus k_1 , k_2 car si en même temps les a_i , ($i = 1, 2, 3$) sont suffisamment proches des α_i , on a alors :

$$\hat{k}_1 > 0, \hat{k}_2 > 0 \text{ et } 0 < \hat{p} < 1 \quad (3.15)$$

Considérons maintenant le cas $k_1 = k_2 = k$

Alors
$$\alpha_1 = \sqrt{\frac{\pi k}{2}}, \alpha_2 = 2k, \alpha_3 = 3k \sqrt{\frac{\pi k}{2}}$$

Calculons ensuite les limites en probabilité (stochastiques) des coefficients de l'équation (3.14) : pour $n \rightarrow \infty$, on a :

$$a_2 - \frac{4}{\pi} a_1^2 \rightarrow \alpha_2 - \frac{4}{\pi} \alpha_1^2 = 0$$

$$\frac{1}{3} a_3 - \frac{1}{2} a_1 a_2 \rightarrow \frac{1}{3} \alpha_3 - \frac{1}{2} \alpha_1 \alpha_2 = 0$$

$$\frac{2}{3\pi} a_1 a_3 - \frac{1}{4} a_2^2 \rightarrow \frac{1}{3\pi} \alpha_1 \alpha_3 - \frac{1}{4} \alpha_2^2 = 0$$

Ces limites stochastiques sont - comme on le sait - égales aux valeurs moyennes des distributions asymptotiquement normales que suivent, pour les grandes valeurs de n , les variables aléatoires figurant à la gauche des dernières expressions. Les variances de ces variables ont la forme [5, 27.7]

$$Cn^{-1} + O(n^{-3/2})$$

où C est une constante déterminée.

IV - 2ème CAS : p CONNU, LES PARAMETRES b_1 , b_2 INCONNUS

Ce cas se présente en pratique assez souvent, et notamment lorsque dans la production en série effectuée par exemple par deux machines (ou ensembles de machines) le nombre des objets fabriqués en série par chaque machine (ou chaque ensemble de machines) dans un même temps est

connu. Dans ce cas p est connu et il ne reste à estimer par la méthode des moments que les paramètres b_1 et b_2 du mélange des distributions de Rayleigh. Il suffit donc de prendre en considération les deux premières équations du système (3.2) et après l'introduction des inconnues auxiliaires selon la formule (3.3) - les deux premières équations de système (3.4). Après élimination de u_2 nous obtenons de ce système l'équation suivante par rapport à l'inconnue u_1 :

$$pu_1^2 - 2\sqrt{\frac{2}{\pi}} pa_1u_1 + \left[\frac{2}{\pi} a_1^2 - \frac{a_2}{2} (1 - p) \right] = 0 \quad (4.1)$$

De cette équation, nous déduisons u_1 , et ensuite, nous basant sur la formule (3.3), nous obtenons

$$\hat{k}_1 = \frac{q}{2p} - a_2 + \frac{2(p - q)}{\pi p} a_1^2 \pm 2a_1 \sqrt{\frac{q}{\pi p} \left(a_2 - \frac{4}{\pi} a_1^2 \right)} \quad (4.2)$$

où $q = 1 - p$.

Remplaçant p par q dans l'équation (4.1), nous obtiendrons une équation concernant u_2 , il en résulte qu'en échangeant p et q , le second membre de l'égalité (4.2) va déterminer les deux valeurs \hat{k}_2 . La première des équations (3.4) montre que, si dans la formule (4.2) nous posons le signe plus, nous obtiendrons la valeur correspondante de l'estimateur \hat{k}_2 en échangeant réciproquement dans cette formule p et q et en posant le signe moins devant la racine carrée.

Au signe moins devant la racine carrée dans la formule (4.2) pour \hat{k}_1 , correspond, outre le changement entre p et q , le signe plus devant la même racine carrée, le second membre de l'équation (4.2) représentera alors \hat{k}_2 .

Si donc $p = \frac{1}{2}$, les équations concernant u_1 et u_2 deviennent identiques, alors l'une des valeurs (n'importe laquelle) déterminée par la formule (4.2) est considérée comme \hat{k}_1 et l'autre comme \hat{k}_2 .

Si $p \neq \frac{1}{2}$, alors lorsque $k_1 > k_2$ il faut adopter dans la formule (4.2) le signe supérieur, si au contraire $k_1 < k_2$ on adopte le signe inférieur et, dans la formule pour \hat{k}_2 les signes contraires à ceux qui sont donnés.

V - VARIANCE ASYMPTOTIQUE DES ESTIMATEURS

Dans le cas général, c'est-à-dire lorsqu'il s'agit de l'estimation des trois paramètres inconnus p , b_1 , b_2 le calcul des variances asymptotiques des estimateurs serait assez pénible et sans grand intérêt, aussi nous bornerons-nous au cas où p est connu.

Nous aurons alors [4, 27.7.3]

$$D^2(\hat{k}_1) = D^2(a_1) \left(\frac{\partial \hat{k}_1}{\partial a_1} \right)^2 + 2 \text{Cov}(a_1, a_2) \frac{\partial \hat{k}_1}{\partial a_1} \frac{\partial \hat{k}_1}{\partial a_2} + D^2(a_2) \left(\frac{\partial \hat{k}_1}{\partial a_2} \right)^2 \quad (5.1)$$

où l'on doit calculer les valeurs des dérivées partielles au point

$$a_1 = \alpha_1 = \sqrt{\frac{\pi}{2}} (p \sqrt{k_1} + q \sqrt{k_2}) \quad (5.2)$$

$$a_2 = \alpha_2 = 2 (pk_1 + qk_2), \quad q = 1 - p$$

Nous calculons donc les variances et la covariance des moments (par exemple selon [6, (10.5), (10.6)]) et nous obtenons :

$$D^2 (a_1) = \frac{1}{2n} [p (4 - \pi p) k_1 + q (4 - \pi q) k_2 - 2 pq \sqrt{k_1 k_2}] \quad (5.3)$$

$$\begin{aligned} \text{Cov} (a_1, a_2) = \frac{1}{n} \sqrt{\frac{\pi}{2}} [p (1 + 2q) k_1^{3/2} + q (1 + 2p) k_2^{3/2} - \\ - 2pq (k_1 k_2^{1/2} + k_1^{1/2} k_2)] \end{aligned} \quad (5.4)$$

La variance du moment de l'échantillon a_2 dépend également du quatrième moment α_4 de la population générale que nous déterminons à partir des formules (2.3) et (3.2).

$$\alpha_4 = 8 pk_1^2 + 8 qk_2^2$$

Ceci étant posé, nous aurons :

$$D^2 (a_2) = \frac{4}{n} [p (2 - p) k_1^2 + q (2 - q) k_2^2 - 2 pq k_1 k_2] \quad (5.5)$$

Nous calculons ensuite les dérivées partielles supposant $k_1 > k_2$, c'est-à-dire admettant dans la formule (4.2) le signe supérieur nous obtenons

$$\frac{\partial \hat{k}_1}{\partial a_1} = \frac{4 (p - q)}{\pi p} a_1 + \frac{2 \left(a_2 - \frac{8}{\pi} a_1^2 \right)}{\sqrt{\frac{\pi p}{q} \left(a_2 - \frac{4}{\pi} a_1^2 \right)}} \quad (5.6)$$

$$\frac{\partial \hat{k}_1}{\partial a_2} = \frac{q}{2p} + \frac{a_1 q}{\sqrt{\pi p q \left(a_2 - \frac{4}{\pi} a_1^2 \right)}} \quad (5.7)$$

Les valeurs de ces dérivées au point (a_1, a_2) déterminé par la formule (5.2) seront respectivement

$$\begin{aligned} \left(\frac{\partial \hat{k}_1}{\partial a_1} \right)_{(a_1, a_2)} &= \frac{-4 \sqrt{k_1 k_2}}{p \sqrt{2\pi} (\sqrt{k_1} - \sqrt{k_2})} \\ \left(\frac{\partial \hat{k}_1}{\partial a_2} \right)_{(a_1, a_2)} &= \frac{\sqrt{k_1}}{2p (\sqrt{k_1} - \sqrt{k_2})} \end{aligned} \quad (5.8)$$

Pour finir, nous portons (5.3), (5.4), (5.5) et (5.8) dans (5.1) et nous obtenons :

$$\begin{aligned} D^2 (\hat{k}_1) = \frac{k_1}{np^2 (\sqrt{k_1} - \sqrt{k_2})^2} \left[p (1 + q) k_1^2 - p (1 + 2q) k_1^{3/2} k_2^{1/2} + \right. \\ \left. + 4 p \left(\frac{4}{\pi} - p \right) k_1 k_2 - 6 pq k_1^{1/2} k_2^{3/2} + q \left(\frac{16}{\pi} - 3 q - 1 \right) k_2^2 \right] \end{aligned} \quad (5.9)$$

REFERENCES

- [1] PEARSON K. - Contributions to the mathematical theory of evolution. Philos Trans. Roy. Soc. London, Vol. 185 A, 1894, 71-110.
- [2] GUMBEL E. J. - La dissection d'une répartition. Annales de l'Université de Lyon, Série 3, Section A, Fascicule 2, 1940, 39-51.
- [3] MENDENHALL W., HADER R. J. - Estimation of parameters of mixed exponentially distributed failure time distributions from censored life test data. Biometrika, Vol. 45, 1958, 504-520.
- [4] RIDDER P. - The method of moments applied to a mixture of two exponential distributions. Annals of math. Stat., 1960.
- [5] CRAMER H. - Mathematical Methods of statistics, 1946.
- [6] KENDALL M. G., STUART A. - The advanced theory of statistics, vol. 1, 1958.