

REVUE DE STATISTIQUE APPLIQUÉE

BERNARD COLIN
NACÉRA MELLAL

Rapport des quantiles empiriques des composantes d'une loi bidimensionnelle

Revue de statistique appliquée, tome 50, n° 4 (2002), p. 55-80

http://www.numdam.org/item?id=RSA_2002__50_4_55_0

© Société française de statistique, 2002, 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/>

RAPPORT DES QUANTILES EMPIRIQUES DES COMPOSANTES D'UNE LOI BIDIMENSIONNELLE

Bernard COLIN⁽¹⁾, Nacéra MELLAL⁽²⁾

⁽¹⁾ *Département de Mathématiques et d'Informatique, Université de Sherbrooke,
Sherbrooke J1K-2R1, (Québec) Canada*

⁽²⁾ *Département de Mathématiques et de Statistique, Université d'Ottawa,
Ottawa K1N-6N5 (Ontario) Canada*

RÉSUMÉ

Notant la sensibilité particulière aux données aberrantes de l'estimateur usuel du rapport de deux moyennes, on considère comme alternative, afin de pallier cette situation, le quotient de deux médianes et plus généralement le quotient de deux quantiles. On montre que ce dernier est, pour le rapport des quantiles théoriques, un estimateur asymptotiquement normal, sans biais et convergent. Des simulations montrent que le rapport des médianes, en tant qu'estimateur du rapport des moyennes, se compare fort bien à des estimateurs de type Hodges-Lehmann ou à des estimateurs construits à l'aide du rapport de moyennes tronquées ou censurées et qu'il est, par ailleurs, moins sensible aux données aberrantes que le quotient des moyennes empiriques.

Mots-clés : Loi bidimensionnelle, rapport des quantiles, copules, loi asymptotique.

ABSTRACT

Outliers in data set may have some drastic effects on the performance of some estimators. In this case for example, the estimation of the ratio of two means may be greatly affected by such outliers. To deal with this problem, one can use, as a competitor of the means ratio, the medians ratio or more generally the quantiles ratio. We show, in this paper, that the empirical quantiles ratio is an asymptotically normal, unbiased and consistent estimator of the theoretical quantiles ratio. Some simulations show that medians ratio, used as an estimator of means ratio performs as well as some others like Hodges-Lehmann, truncated means ratio or winsorized means ratio estimators and moreover is less sensitive to outliers than the means ratio.

Keywords : Bidimensional probability measures, quantiles ratio, copula, asymptotic normality.

1. Introduction

Il est usuel dans certains domaines comme ceux de la finance, de l'économie, de la sociologie et bien d'autres encore, que l'on soit amené à effectuer dans le but de construire des indices, le quotient de deux variables aléatoires X et Y lequel se ramène fréquemment au quotient des deux moyennes empiriques \bar{X} et \bar{Y} . Il en est de même en théorie de l'échantillonnage, que ce soit dans le cas de l'estimation d'un

rapport, ou dans le cas où l'on désire améliorer les performances d'un estimateur d'un total par le biais d'une information auxiliaire.

Connaissant la sensibilité particulière des moyennes empiriques à l'égard des observations aberrantes et donc de l'impact que cela pourrait avoir sur la qualité de l'estimateur usuel du quotient de deux moyennes, on pourrait penser à utiliser, à cette fin, le rapport des moyennes tronquées ou le rapport des moyennes censurées ou encore le rapport de diverses statistiques déduites de la considération d'estimateurs de types M , L , ou R autrement dit, d'estimateurs qui sont respectivement : solution d'un problème de minimisation d'une fonction des observations et d'un paramètre, combinaison linéaire des statistiques d'ordre ou fonction des rangs des observations.

Toujours dans ce contexte, l'approche fonctionnelle mise à l'avant par les U et les V -statistiques pourrait s'avérer des plus intéressantes tant du point de vue asymptotique que du point de vue, plus empirique, du rééchantillonnage.

Cependant considérant le fait que la médiane d'une loi de probabilité constitue un paramètre de tendance centrale et que la médiane empirique est moins sensible aux observations aberrantes que ne l'est la moyenne empirique, nous proposons, en particulier dans le cas des lois symétriques, de considérer à titre d'estimateur concurrent au quotient de deux moyennes, le quotient de deux médianes, ou plus généralement le quotient de deux quantiles empiriques. Cette idée découle de l'exemple suivant emprunté à un problème d'estimation en théorie de l'échantillonnage et tiré de l'ouvrage de Cochran [2].

Exemple. – Le nombre total d'habitants (en milliers) dans 196 grandes villes des Etats-Unis (quatre grandes villes : New-York, Chicago, Philadelphie et Détroit, ont été omises) est de 22919 en 1920 et de 29351 en 1930. On prélève au hasard 9 villes parmi les 196, on note respectivement par (x_i) et par (y_i) le nombre d'habitants (en milliers) de ces villes en 1930 et en 1920. L'ensemble des observations est présentée dans le tableau suivant :

x_i :	80	115	106	64	104	57	85	570	93
y_i :	76	120	78	40	93	2	94	25	74

Les valeurs numériques $y_6 = 2$ et $x_8 = 570$ correspondent de fait à des valeurs aberrantes car elles devraient être respectivement de 25 et 70. Si l'on désire alors estimer la croissance relative moyenne de la population des 196 villes retenues de 1920 à 1930, à l'aide de l'échantillon observé, en utilisant le quotient des moyennes empiriques \bar{x}/\bar{y} , on trouve comme estimation 2,11 alors que le vrai rapport \bar{X}/\bar{Y} est de $29351/22919 = 1,28$. Cependant le rapport des médianes empiriques $med(x_i)/med(y_i) = x_{(5)}/y_{(5)}$ a pour valeur numérique $93/76 = 1,22$. Cette dernière valeur, beaucoup plus proche de la réalité que celle fournie par le quotient des moyennes empiriques, incite à penser que le rapport des médianes empiriques représente un estimateur du rapport des moyennes de la population plus "robuste" que l'estimateur habituellement utilisé.

Désignant par $\hat{\xi}_p$ et $\hat{\eta}_q$ les quantiles empiriques d'ordres p et q de deux variables aléatoires X et Y quelconques, nous nous proposons, dans les paragraphes qui suivent, d'étudier le rapport $\hat{\xi}_p/\hat{\eta}_q$ (loi à distance finie, espérance, variance, loi asymptotique) et d'envisager en particulier le rapport $\hat{\xi}_{0,5}/\hat{\eta}_{0,5}$ comme alternative à l'estimateur

$\overline{X}/\overline{Y}$ dans le cas où la loi du couple (X, Y) est symétrique. Nous illustrerons enfin le tout à l'aide de simulations.

Afin de préciser le cadre théorique de l'étude qui suit, nous considérerons une structure statistique paramétrique dominée de la forme :

$$(\mathbb{R}^2, \mathcal{B}_{\mathbb{R}^2}, \mathbb{P}_\theta; \theta \in \Theta \subseteq \mathbb{R}^k, \mu)$$

où μ désigne une mesure de référence σ -finie (en général la mesure de Lebesgue) telle que $\mathbb{P}_\theta \ll \mu \forall \theta \in \Theta$ et où Θ désigne l'espace paramètre de la mesure de probabilité considérée. Associé à cette structure statistique, nous noterons par $(\mathbb{R}^2, \mathcal{B}_{\mathbb{R}^2}, \mathbb{P}_\theta)^{\otimes n}$ le modèle empirique engendré par n réalisations indépendantes du modèle initial.

2. Loi du rapport des statistiques d'ordre des composantes d'un vecteur aléatoire de \mathbb{R}^2

On désignera par X et Y les composantes d'un vecteur aléatoire de \mathbb{R}^2 , par $(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$ les observations issues du modèle tel qu'introduit ci-dessus et l'on notera par $X_{(r)}$ et $Y_{(s)}$ les statistiques d'ordres r et s associées respectivement aux observations (X_i) et (Y_i) pour tout $i = 1, 2, \dots, n$.

2.1. Loi du couple $(X_{(r)}, Y_{(s)})$

Soit $F_{X_{(r)}, Y_{(s)}}(x, y)$ la fonction de répartition du vecteur aléatoire $(X_{(r)}, Y_{(s)})$. Pour des raisons de commodité d'écriture nous noterons désormais cette fonction de répartition par : $F_{(r,s)}(x, y)$. Posons alors :

$$\begin{aligned} p_1 &= p_1(x, y) = \mathbb{P}_\theta(X \leq x, Y \leq y); & p_2 &= p_2(x, y) = \mathbb{P}_\theta(X \leq x, Y > y) \\ p_3 &= p_3(x, y) = \mathbb{P}_\theta(X > x, Y \leq y); & p_4 &= p_4(x, y) = \mathbb{P}_\theta(X > x, Y > y) \end{aligned}$$

L'événement $(X_{(r)} \leq x, Y_{(s)} \leq y)$ ne peut se réaliser que s'il y a au moins r observations X_i telles que $X_i \leq x$ et au moins s observations Y_j telles que $Y_j \leq y$. Notons par A_{lm} l'événement défini par : "il y a exactement l observations X_i ($l \geq r$) telles que $X_i \leq x$ et il y a exactement m observations Y_j ($m \geq s$) telles que $Y_j \leq y$ ". Pour des valeurs de l et m fixées, notons également par $A_{lm,k}$ l'événement défini par : "il y a exactement k couples (X_i, Y_i) tels que $X_i \leq x$ et $Y_i \leq y$ " où $k \leq \min(l, m)$. Pour que ce dernier événement puisse se réaliser il faut donc qu'il y ait exactement : $l - k$ couples tels que $X_i \leq x$ et $Y_i > y$, $m - k$ couples tels que $X_i > x$ et $Y_i \leq y$ et enfin $n - l - m + k$ couples tels que $X_i > x$ et $Y_i > y$. La probabilité d'un tel événement est alors donnée par :

$$\mathbb{P}(A_{lm,k}) = \frac{n!}{k!(l-k)!(m-k)!(n-l-m+k)!} p_1^k p_2^{l-k} p_3^{m-k} p_4^{n-l-m+k}$$

Afin de déterminer $\mathbb{P}(A_{lm})$, il suffit alors d'effectuer la somme des probabilités $\mathbb{P}(A_{lm,k})$ sur toutes les valeurs de k compatibles avec les valeurs de l, m et n .

Observant que la condition $l + m \leq n$ entraîne que $k \geq 0$ et que la condition $l + m > n$ entraîne que $k \geq l + m - n$, il vient, en posant $u = \text{Sup}(0, l + m - n)$ et $v = \text{min}(l, m)$:

$$\mathbb{P}(A_{lm}) = \sum_{k=u}^v \frac{n!}{k! (l-k)! (m-k)! (n-l-m+k)!} p_1^k p_2^{l-k} p_3^{m-k} p_4^{n-l-m+k}$$

Désignant par $M_n(k, l, m)$ le coefficient $n! / k! (l-k)! (m-k)! (n-l-m+k)!$ et tenant compte des contraintes sur les indices l et m , il s'ensuit que :

$$\begin{aligned} F_{(r,s)}(x, y) &= \mathbb{P}(X_{(r)} \leq x, Y_{(s)} \leq y) \\ &= \sum_{l=r}^n \sum_{m=s}^n \sum_{k=u}^v M_n(k, l, m) p_1^k p_2^{l-k} p_3^{m-k} p_4^{n-l-m+k} \end{aligned}$$

On notera que ce résultat est indépendant de la nature discrète ou non de la fonction de répartition du vecteur aléatoire considéré. Par ailleurs si $y \rightarrow \infty$ on vérifie sans peine que :

$$\lim_{y \rightarrow \infty} p_2(x, y) = \lim_{y \rightarrow \infty} p_4(x, y) = 0$$

que :

$$\lim_{y \rightarrow \infty} p_1(x, y) = \mathbb{P}_\theta(X \leq x); \quad \lim_{y \rightarrow \infty} p_3(x, y) = \mathbb{P}_\theta(X > x)$$

et que :

$$k \rightarrow l \text{ avec probabilité } 1; \quad m - k \rightarrow n - l \text{ avec probabilité } 1$$

Ainsi en posant $\mathbb{P}_\theta(X \leq x) = F_X(x)$ on a :

$$\lim_{y \rightarrow \infty} F_{(r,s)}(x, y) = \mathbb{P}(X_{(r)} \leq x) = \sum_{l=r}^n \binom{n}{l} (F_X(x))^l (1 - F_X(x))^{n-l}$$

et également, avec des notations analogues :

$$\lim_{x \rightarrow \infty} F_{(r,s)}(x, y) = \mathbb{P}(Y_{(s)} \leq y) = \sum_{m=s}^n \binom{n}{m} (F_Y(y))^m (1 - F_Y(y))^{n-m}$$

Ces deux dernières expressions ne sont autres, bien entendu, que les lois des statistiques d'ordre $X_{(r)}$ et $Y_{(s)}$.

Il se peut dans la pratique que la détermination de $F_{(r,s)}(x, y)$ entraîne des calculs fastidieux; on pourra alors, si l'on ne dispose pas d'autres moyens pour

l'évaluer, utiliser des approximations numériques telles que celles décrites dans Galambos [6].

Si l'on désigne par $F_{X,Y}(x, y)$ la loi du vecteur aléatoire (X, Y) on a :

$$p_1(x, y) = F_{X,Y}(x, y); p_2(x, y) = F_X(x) - F_{X,Y}(x, y)$$

$$p_3(x, y) = F_Y(y) - F_{X,Y}(x, y); p_4(x, y) = 1 - F_X(x) - F_Y(y) + F_{X,Y}(x, y)$$

et la densité du couple $(X_{(r)}, Y_{(s)})$, notée $f_{(r,s)}(x, y)$, a pour expression :

$$\sum_{l=r}^n \sum_{m=s}^n \sum_{k=u}^v M_n(k, l, m)$$

$$\times \frac{\partial^2}{\partial x \partial y} \left[(F_{X,Y}(x, y))^k (F_X(x) - F_{X,Y}(x, y))^{l-k} (F_Y(y) - F_{X,Y}(x, y))^{m-k} \right]$$

$$\times (1 - F_X(x) - F_Y(y) + F_{X,Y}(x, y))^{n-l-m+k}$$

2.2. Loi du rapport $X_{(r)} \diagdown Y_{(s)}$

Nous nous contenterons de donner ici l'expression générale de la densité de la variable aléatoire $X_{(r)} \diagdown Y_{(s)}$ (cette dernière étant, par ailleurs, définie presque sûrement en vertu du fait que, dans le cadre retenu, $\mathbb{P}(Y_{(s)} \neq 0) = 1$)

Considérons le changement de variable suivant :

$$U = X_{(r)} \text{ et } V = X_{(r)} \diagdown Y_{(s)}$$

La valeur absolue du jacobien de la transformation ayant pour expression $|U| \diagdown V^2$, il s'ensuit que la densité de probabilité du couple (U, V) est donnée par :

$$f_{U,V}(u, v) = f_{(r,s)}\left(u, \frac{u}{v}\right) \frac{|u|}{v^2} \text{ avec } v \in \mathbb{R} \setminus \{0\}$$

Par conséquent :

$$f_V(v) = \int f_{(r,s)}\left(u, \frac{u}{v}\right) \frac{|u|}{v^2} du$$

ou encore en posant $u \diagdown v = t$, il vient :

$$f_V(v) = \int f_{(r,s)}(vt, t) |t| dt \text{ où } v \in \mathbb{R} \setminus \{0\}$$

De la même manière si l'on pose $W = Y_{(s)} \diagdown X_{(r)} = 1 \diagdown V$ on a :

$$f_W(w) = \int f_{(r,s)}(t, wt) |t| dt \text{ où } w \in \mathbb{R} \setminus \{0\}$$

Par ailleurs si les variables aléatoires X et Y sont échangeables, les statistiques d'ordre $X_{(r)}$ et $Y_{(r)}$ sont régies par la même loi pour tout $r = 1, 2, \dots, n$, de sorte que les variables $X_{(r)}/Y_{(s)}$ et $Y_{(r)}/X_{(s)}$ sont identiquement distribuées. De plus si $r = s$ les variables $Z = X_{(r)}/Y_{(r)}$ et $1/Z = Y_{(r)}/X_{(r)}$ ont même loi.

Mentionnons enfin que le cadre théorique considéré permet, à l'aide du changement de variables habituel $U = F_X(X)$ et $V = F_Y(Y)$ (à ne pas confondre avec le précédent), de se ramener systématiquement au cas où les variables aléatoires suivent une loi uniforme $\mathcal{U}_{[0,1]}$ sur l'intervalle $[0, 1]$. On peut alors vérifier sans peine que la fonction de répartition $F_{U,V}(u, v)$ du couple (U, V) , est donnée par :

$$F_{U,V}(u, v) = F_{X,Y}(F_X^{-1}(u), F_Y^{-1}(v))$$

Cette dernière expression, dite "représentation uniforme" de $F_{X,Y}$ (voir Sklar [14], Fréchet [5]), porte le nom de "copule" (voir également Kimeldorf et Sampson [8]).

Notant que pour tout $r, s = 1, 2, \dots, n$ les ensembles $\{F_X(X_{(r)})\}$ et $\{F_Y(Y_{(s)})\}$ ne sont autres que les ensembles $\{U_{(r)}\}$ et $\{V_{(s)}\}$ des statistiques d'ordres r et s des variables aléatoires U et V , il vient :

$$\frac{X_{(r)}}{Y_{(s)}} = \frac{F_X^{-1}(U_{(r)})}{F_Y^{-1}(V_{(s)})}$$

Outre le fait que cette représentation uniforme constitue un cadre commun pour l'ensemble des lois considérées, elle permet certaines simplifications dans les calculs, et s'avère particulièrement intéressante dans le cas où l'on désire effectuer des simulations si, de surcroît, la copule est archimédienne (voir Ling [9], Genest et MacKay [7]). Dans ce cas en effet, la simulation s'obtient en générant deux variables aléatoires indépendantes de lois uniformes sur l'intervalle $[0, 1]$.

3. Loi asymptotique du rapport de deux quantiles empiriques des composantes d'un vecteur aléatoire de \mathbb{R}^2

Dans ce paragraphe de nature plus technique, ne faisant appel cependant qu'à des outils et des résultats classiques, nous montrerons la convergence en loi du rapport des quantiles empiriques $\hat{\xi}_p/\hat{\eta}_q$. Dans le cas des lois $F_{X,Y}(x, y)$ symétriques, nous présenterons les conséquences pratiques de ce résultat en considérant le rapport $\hat{\xi}_{0,5}/\hat{\eta}_{0,5}$ en tant qu'estimateur du quotient $\mathbb{E}(X)/\mathbb{E}(Y)$. Des simulations illustreront l'aspect asymptotique ainsi que les comportements respectifs, en présence de données aberrantes, des estimateurs $\hat{\xi}_{0,5}/\hat{\eta}_{0,5}$ et \bar{X}/\bar{Y} .

3.1. Loi du rapport de deux quantiles empiriques des composantes d'un vecteur aléatoire de \mathbb{R}^2

Soit $(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$ un échantillon de taille n issu d'un modèle empirique quelconque $(\mathbb{R}^2, \mathcal{B}_{\mathbb{R}^2}, \mathbb{P}_\theta)^{\otimes n}$. Désignons par $\hat{\xi}_p$ et $\hat{\eta}_q$ les quantiles empiriques d'ordres p et q obtenus à partir des observations (X_1, X_2, \dots, X_n) et (Y_1, Y_2, \dots, Y_n) des variables aléatoires X et Y . Il vient en utilisant les définitions et notations habituelles :

$$\hat{\xi}_p = X_{(r)} \text{ avec } r = \begin{cases} np & \text{si } np \text{ est un entier} \\ [np + 1] & \text{sinon} \end{cases}$$

et

$$\hat{\eta}_q = Y_{(s)} \text{ avec } s = \begin{cases} nq & \text{si } nq \text{ est un entier} \\ [nq + 1] & \text{sinon} \end{cases}$$

Il s'ensuit que la loi du couple $(\hat{\xi}_p, \hat{\eta}_q)$ se déduit aisément des résultats précédents et qu'il en est de même pour la loi du rapport $\hat{\xi}_p / \hat{\eta}_q = \hat{R}_{p,q}$.

3.2. Moments d'ordres 1 et 2 du rapport des quantiles empiriques

De façon générale, il ne sera pas possible de donner, pour tout n , une expression simple et explicite de $\mathbb{E}(\hat{R}_{p,q})$ et de $\mathbb{E}(\hat{R}_{p,q}^2)$ en raison principalement de la complexité de la loi de $\hat{R}_{p,q}$. On peut cependant en obtenir des approximations à l'aide des techniques usuelles.

3.2.1. Moment d'ordre 1

Le rapport $\hat{\xi}_p / \hat{\eta}_q$ s'exprimant sous la forme du quotient de deux statistiques d'ordre, par exemple $X_{(r)} / Y_{(s)} = F_X^{-1}(U_{(r)}) / F_Y^{-1}(V_{(s)})$, désignons par

$$\mu = {}^t(\mu_{(r)}, \nu_{(s)}) = {}^t(\mathbb{E}(U_{(r)}), \mathbb{E}(V_{(s)}))$$

le vecteur moyen du vecteur aléatoire $Z = {}^t(U_{(r)}, V_{(s)})$ et par Σ sa matrice de variance-covariance.

Posons $g(z) = g(u_{(r)}, v_{(s)}) = F_X^{-1}(u_{(r)}) / F_Y^{-1}(v_{(s)}) = x_{(r)} / y_{(s)}$ et notons par $g(Z)$ la variable aléatoire associée. Utilisant la version multidimensionnelle du théorème de Taylor, on a, en supposant que les conditions de différentiabilité

requis portant sur g soient remplies :

$$g(z) - g(\mu) = \sum_{k=1}^{n-1} \frac{1}{k!} \sum_{i_1=1}^2 \sum_{i_2=1}^2 \cdots \sum_{i_k=1}^2 \frac{\partial^k g(u_1, u_2)}{\partial u_{i_1} \partial u_{i_2} \cdots \partial u_{i_k}} \Big|_{z=\mu} \prod_{j=1}^k (z_{i_j} - \mu_{i_j}) \\ + \frac{1}{n!} \sum_{i_1=1}^2 \sum_{i_2=1}^2 \cdots \sum_{i_n=1}^2 \frac{\partial^n g(u_1, u_2)}{\partial u_{i_1} \partial u_{i_2} \cdots \partial u_{i_n}} \Big|_{z=z^*} \prod_{j=1}^n (z_{i_j} - \mu_{i_j})$$

où $z^* = \mu + t(z - \mu)$ pour un certain $t \in]0, 1[$

Le développement limité à l'ordre 2 s'exprime alors sous la forme :

$$g(z) = g(\mu) + {}^t \text{Grad}g \Big|_{\mu} (z - \mu) + \frac{1}{2!} \left({}^t (z - \mu) D^{(2)}g \Big|_{\mu} (z - \mu) \right) \\ + \frac{1}{3!} \sum_{i_1=1}^2 \sum_{i_2=1}^2 \sum_{i_3=1}^2 \frac{\partial^3 g(u_1, u_2)}{\partial u_{i_1} \partial u_{i_2} \partial u_{i_3}} \Big|_{z^*} \prod_{j=1}^3 (z_{i_j} - \mu_{i_j})$$

où $\text{Grad}g \Big|_{\mu}$ et $D^{(2)}g \Big|_{\mu}$ désignent respectivement le gradient et la matrice hessienne de la fonction g , évalués en $z = \mu$.

Notant, pour des raisons de commodité d'écriture, par $Q(z - \mu)$ la forme quadratique ${}^t (z - \mu) D^{(2)}g \Big|_{\mu} (z - \mu)$

$$= \frac{\partial^2 g}{\partial u_1^2} (u_{(r)} - \mu_{(r)})^2 + \frac{\partial^2 g}{\partial u_2^2} (v_{(s)} - \nu_{(s)})^2 + 2 \frac{\partial^2 g}{\partial u_1 \partial u_2} (u_{(r)} - \mu_{(r)}) (v_{(s)} - \nu_{(s)})$$

il vient :

$$\left| g(z) - g(\mu) - {}^t \text{Grad}g \Big|_{\mu} (z - \mu) - \frac{1}{2!} Q(z - \mu) \right| = o(\|z - \mu\|^3)$$

L'espérance de la variable aléatoire $g(Z) = X_{(r)} / Y_{(s)}$ peut donc s'exprimer sous la forme :

$$\mathbb{E} \left(\frac{X_{(r)}}{Y_{(s)}} \right) = \frac{F_X^{-1}(\mu_{(r)})}{F_Y^{-1}(\nu_{(s)})} + \frac{1}{2!} \mathbb{E}(Q(Z - \mu)) \\ + \frac{1}{3!} \sum_{i_1=1}^2 \sum_{i_2=1}^2 \sum_{i_3=1}^2 \frac{\partial^3 g(u_1, u_2)}{\partial u_{i_1} \partial u_{i_2} \partial u_{i_3}} \Big|_{z^*} \mathbb{E} \left(\prod_{j=1}^3 (Z_{i_j} - \mu_{i_j}) \right)$$

et par conséquent on a :

$$\left| \mathbb{E} \left(\frac{X_{(r)}}{Y_{(s)}} \right) - \frac{F_X^{-1}(\mu_{(r)})}{F_Y^{-1}(\nu_{(s)})} - \frac{1}{2!} \mathbb{E}(Q(Z - \mu)) \right| = o(\|Z - \mu\|^3)$$

Ainsi, les approximations d'ordres 1 et 2 de $\mathbb{E}(X_{(r)}/Y_{(s)})$ sont données respectivement par :

$$\frac{F_X^{-1}(\mu_{(r)})}{F_Y^{-1}(\nu_{(s)})} \text{ et par } \frac{F_X^{-1}(\mu_{(r)})}{F_Y^{-1}(\nu_{(s)})} + \frac{1}{2!} \mathbb{E}(Q(Z - \mu))$$

avec $\mu = {}^t(r/(n+1), s/(n+1))$ puisque les statistiques $U_{(r)}$ et $V_{(s)}$ sont les statistiques d'ordre r et d'ordre s de deux lois uniformes $\mathcal{U}_{[0,1]}$

Désignons par $\hat{\lambda}_p$ et $\hat{\omega}_q$ les quantiles empiriques d'ordres p et q déduits des statistiques d'ordres $F_X(X_{(i)})$ et $F_Y(Y_{(i)})$ $i = 1, 2, \dots, n$ issues de lois uniformes $\mathcal{U}_{[0,1]}$. Supposons que n, r et s tendent vers l'infini de sorte que $n^{-1}r \rightarrow p$ et que $n^{-1}s \rightarrow q$. On sait dans ce cas que sous des conditions très fréquemment vérifiées dans la pratique (voir Serfling [12], David [3]), les quantiles empiriques $\hat{\lambda}_p$ et $\hat{\omega}_q$ convergent, respectivement, presque sûrement vers p et q . Ainsi nous aurons en première approximation :

$$\begin{aligned} \mathbb{E} \left(\frac{\hat{\xi}_p}{\hat{\eta}_q} \right) &= \mathbb{E} \left(\hat{R}_{p,q} \right) = \frac{\xi_p}{\eta_q} + \frac{1}{2!} \mathbb{E}(Q(Z - \mu)) + o(\|Z - \mu\|^3) \\ &\simeq \frac{\xi_p}{\eta_q} + \frac{1}{2!} \mathbb{E}(Q(Z - \mu)) \end{aligned}$$

3.2.2. Moment d'ordre 2

Afin de trouver une approximation de la variance de $X_{(r)}/Y_{(s)}$, considérons la fonction $g(z) - \mathbb{E}(g(Z)) = x_{(r)}/y_{(s)} - \mathbb{E}(X_{(r)}/Y_{(s)})$. Cette dernière a pour expression :

$$\begin{aligned} &{}^t(z - \mu) \text{ Grad}g|_{\mu} + \frac{1}{2!} (Q(z - \mu) - \mathbb{E}(Q(Z - \mu))) \\ &+ \frac{1}{3!} \sum_{i_1=1}^2 \sum_{i_2=1}^2 \sum_{i_3=1}^2 \frac{\partial^3 g(u_1, u_2)}{\partial u_{i_1} \partial u_{i_2} \partial u_{i_3}} \Big|_{z^*} \left[\prod_{j=1}^3 (z_{i_j} - \mu_{i_j}) - \mathbb{E} \left(\prod_{j=1}^3 (Z_{i_j} - \mu_{i_j}) \right) \right] \end{aligned}$$

d'où :

$$\left| \begin{aligned} &g(z) - \mathbb{E}(g(Z)) - {}^t(z - \mu) \text{ Grad}g|_{\mu} \\ &- \frac{1}{2!} (Q(z - \mu) - \mathbb{E}(Q(Z - \mu))) \end{aligned} \right| = o(\|z - \mu\|^3)$$

En élevant l'expression de $g(z) - \mathbb{E}(g(Z))$ au carré il vient :

$$\begin{aligned} (g(z) - \mathbb{E}(g(Z)))^2 &= \left({}^t(z - \mu) \text{Grad}g|_{\mu} \right)^2 \\ &\quad + \frac{1}{4} (Q(z - \mu) - \mathbb{E}(Q(Z - \mu)))^2 \\ &\quad + {}^t(z - \mu) \text{Grad}g|_{\mu} (Q(z - \mu) - \mathbb{E}(Q(Z - \mu))) + h(z) \end{aligned}$$

où $h(z)$ est une fonction ne comportant que des puissances de z supérieures ou égales à 3. En prenant l'espérance des deux membres de l'égalité ci-dessus, en isolant les termes de puissances inférieures ou égales à 2 des autres et en tenant compte du fait que $\mathbb{E}(Z - \mu) = 0$, on a :

$$\begin{aligned} \text{Var}(g(Z)) &= \mathbb{E} \left({}^t(Z - \mu) \text{Grad}g|_{\mu} {}^t \text{Grad}g|_{\mu} (Z - \mu) \right) \\ &\quad - \frac{1}{4} (\mathbb{E}(Q(Z - \mu)))^2 + \mathbb{E}(h'(Z)) \end{aligned}$$

où $\mathbb{E}(h'(Z))$ ne fait intervenir que des moments centrés d'ordres supérieurs ou égaux à 3.

Ainsi l'approximation d'ordre 1 de $\text{Var} \left(\frac{X_{(r)}}{Y_{(s)}} \right) = \text{Var} \left(\frac{F_X^{-1}(U_{(r)})}{F_Y^{-1}(V_{(s)})} \right)$ est donnée par :

$$\begin{aligned} &\mathbb{E} \left({}^t(Z - \mu) \text{Grad}g|_{\mu} {}^t \text{Grad}g|_{\mu} (Z - \mu) \right) \\ &= {}^t \text{Grad}g|_{\mu} \Sigma \text{Grad}g|_{\mu} \text{ avec } \Sigma = \mathbb{E} \left((Z - \mu) {}^t (Z - \mu) \right) \end{aligned}$$

Quant à l'approximation d'ordre 2 elle est donnée par :

$${}^t \text{Grad}g|_{\mu} \Sigma \text{Grad}g|_{\mu} - \frac{1}{4} \left[\begin{aligned} &\frac{\partial^2 g}{\partial u_1^2} \text{Var}(U_{(r)}) + \frac{\partial^2 g}{\partial u_2^2} \text{Var}(V_{(s)}) \\ &+ 2 \frac{\partial^2 g}{\partial u_1 \partial u_2} \text{Cov}(U_{(r)}, V_{(s)}) \end{aligned} \right]^2$$

Rappelant que $X_{(r)} = F_X^{-1}(U_{(r)})$ et que $Y_{(s)} = F_Y^{-1}(V_{(s)})$, on vérifie aisément que :

$$\begin{aligned} \frac{dx_{(r)}}{du_{(r)}} \Big|_{u_{(r)}=u} &= \left(\left(\frac{du_{(r)}}{dx_{(r)}} \right)^{-1} \Big|_{x_{(r)}=F_X^{-1}(u)} \right) = \frac{1}{f_X[F_X^{-1}(u)]} \text{ et que :} \\ \frac{dy_{(s)}}{dv_{(s)}} \Big|_{v_{(s)}=v} &= \frac{1}{f_Y[F_Y^{-1}(v)]} \end{aligned}$$

De même :

$$\begin{aligned} \left. \frac{d^2 x_{(r)}}{du_{(r)}^2} \right|_{u_{(r)}=u} &= \frac{d}{du_{(r)}} [f_X(x_{(r)})]^{-1} \Big|_{x_{(r)}=F_X^{-1}(u)} \\ &= -f'_X(F_X^{-1}(u)) (f_X(F_X^{-1}(u)))^{-3} \quad \text{et} \\ \left. \frac{d^2 y_{(s)}}{dv_{(s)}^2} \right|_{v_{(s)}=v} &= -f'_Y(F_Y^{-1}(v)) (f_Y(F_Y^{-1}(v)))^{-3} \end{aligned}$$

Ces dérivées étant connues, il est alors possible d'évaluer les dérivées partielles de la fonction g et d'en déduire les approximations de $\mathbb{E}(X_{(r)}/Y_{(s)})$ et de $Var(X_{(r)}/Y_{(s)})$ compte tenu du fait que :

$$Var(U_{(r)}) = \frac{r(n-r+1)}{(n+1)^2(n+2)} \quad \text{et que} \quad Var(V_{(s)}) = \frac{s(n-s+1)}{(n+1)^2(n+2)}$$

Quant à la covariance de $U_{(r)}$ et de $V_{(s)}$ il est nécessaire de connaître la loi conjointe du couple $(U_{(r)}, V_{(s)})$ pour la déterminer. Cependant, si les variables sont non corrélées ou si le couple $(U_{(r)}, V_{(s)})$ est "asymptotiquement indépendant" (voir Srivastava [15]) les calculs précédents s'en trouvent simplifiés.

3.3. Loi asymptotique du rapport $\hat{\xi}_p/\hat{\eta}_q$

Considérant la représentation de $F_{X,Y}(x,y)$ sous la forme de la copule $F_{U,V}(u,v) = F_{X,Y}(F_X^{-1}(u), F_Y^{-1}(v))$, désignons respectivement par $\hat{\xi}_p, \hat{\eta}_q, \hat{\lambda}_p$ et $\hat{\omega}_q$ les quantiles empiriques d'ordre p et d'ordre q des variables aléatoires X et Y et des variables aléatoires $U = F_X(X)$ et $V = F_Y(Y)$, déduits de l'échantillon

$$(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$$

et de l'échantillon

$$(F_X(X_1), F_Y(Y_1)), (F_X(X_2), F_Y(Y_2)), \dots, (F_X(X_n), F_Y(Y_n)))$$

De même, désignons respectivement par ξ_p, η_q, λ_p et ω_q les quantiles théoriques d'ordre p et d'ordre q des variables aléatoires X et Y et des variables aléatoires U et V .

Appliquant le théorème de représentation des quantiles de Bahadur [1] ainsi que le théorème de Slutsky (voir Serfling [12]) aux "pseudo-observations"

$$(U_1, V_1), (U_2, V_2), \dots, (U_n, V_n)$$

il vient, en notant que $\lambda_p = p$ et que $\omega_q = q$:

$$\begin{pmatrix} \hat{\lambda}_p - p \\ \hat{\omega}_q - q \end{pmatrix} = \begin{pmatrix} p - F_U^{(n)}(p) \\ q - F_V^{(n)}(q) \end{pmatrix} + \mathbf{R}_n \mathbf{1}$$

où $F_U^{(n)}$ et $F_V^{(n)}$ désignent respectivement les fonctions de répartition marginales empiriques des variables U et V , où $\mathbf{R}_n = O\left(n^{-3/4} (\text{Log} n)^{3/4}\right)$ avec probabilité 1 et où $\mathbf{1}$ désigne le vecteur de \mathbb{R}^2 dont les composantes sont égales à 1. Or :

$$\left(F_U^{(n)}(p), F_V^{(n)}(q)\right) = \frac{1}{n} \sum_{i=1}^n (\mathbb{I}_{]-\infty, p]}(U_i), \mathbb{I}_{]-\infty, q]}(V_i))$$

donc :

$$\mathbb{E}\left(F_U^{(n)}(p), F_V^{(n)}(q)\right) = \frac{1}{n} \sum_{i=1}^n (\mathbb{E}(\mathbb{I}_{]-\infty, p]}(U_i)), \mathbb{E}(\mathbb{I}_{]-\infty, q]}(V_i))) = (p, q)$$

De plus pour tout $i = 1, 2, \dots, n$ on a :

$$\text{Var}(\mathbb{I}_{]-\infty, p]}(U_i)) = p(1-p); \text{Var}(\mathbb{I}_{]-\infty, q]}(V_i)) = q(1-q)$$

$$\text{et } \text{Cov}(\mathbb{I}_{]-\infty, p]}(U_i), \mathbb{I}_{]-\infty, q]}(V_i)) = F_{U,V}(p, q) - pq = F_{X,Y}(\xi_p, \eta_q) - pq$$

et par conséquent les vecteurs aléatoires $Z_i = {}^t(\mathbb{I}_{]-\infty, p]}(U_i), \mathbb{I}_{]-\infty, q]}(V_i))$ sont indépendants et identiquement distribués de vecteur moyen $\mu = {}^t(p, q)$ et de matrice de variance-covariance Σ donnée par :

$$\Sigma = \begin{bmatrix} p(1-p) & F_{U,V}(p, q) - pq \\ F_{U,V}(p, q) - pq & q(1-q) \end{bmatrix}$$

La version multidimensionnelle du théorème central limite appliquée à la quantité

$$\frac{1}{n} \sum_{i=1}^n Z_i = {}^t(F_U^{(n)}(p), F_V^{(n)}(q))$$

nous permet de conclure que $\sqrt{n} \left(\frac{1}{n} \sum_{i=1}^n Z_i - \mu\right) \xrightarrow{L} N(0, \Sigma)$ et donc que le vecteur aléatoire ${}^t(F_U^{(n)}(p), F_V^{(n)}(q))$ est "asymptotiquement normal" de vecteur moyen $\mu = {}^t(p, q)$ et de matrice de variance-covariance $\frac{1}{n}\Sigma$ (propriété que l'on notera par le symbole : $\rightsquigarrow AN\left(\mu, \frac{1}{n}\Sigma\right)$). Il s'ensuit alors que :

$${}^t(\hat{\lambda}_p, \hat{\omega}_q) \rightsquigarrow AN\left({}^t(p, q), \frac{1}{n}\Sigma\right)$$

Posons de nouveau :

$$g(\hat{\lambda}_p, \hat{\omega}_q) = F_X^{-1}(\hat{\lambda}_p) / F_Y^{-1}(\hat{\omega}_q) = \hat{\xi}_p / \hat{\eta}_q$$

Il résulte de la dernière propriété que (voir Serfling [12]) :

$$\frac{\hat{\xi}_p}{\hat{\eta}_q} \rightsquigarrow AN \left(g(p, q), \frac{1}{n} {}^t \text{Grad}g|_{(p,q)} \Sigma \text{Grad}g|_{(p,q)} \right)$$

Or $g(p, q) = \xi_p / \eta_q$ et :

$$\text{Grad}g|_{(p,q)} = \left(\frac{\frac{1}{f_X(F_X^{-1}(p)) F_Y^{-1}(q)} - F_X^{-1}(p)}{f_Y(F_Y^{-1}(q)) (F_Y^{-1}(q))^2} \right) = \left(\frac{\frac{1}{f_X(\xi_p) \eta_q} - \xi_p}{f_Y(\eta_q) \eta_q^2} \right)$$

Par conséquent : ${}^t \text{Grad}g|_{(p,q)} \Sigma \text{Grad}g|_{(p,q)} =$

$$\frac{p(1-p)}{\eta_q^2 f_X^2(\xi_p)} + \frac{\xi_p^2 q(1-q)}{\eta_q^4 f_Y^2(\eta_q)} - \frac{2\xi_p(F_{X,Y}(\xi_p, \eta_q) - pq)}{\eta_q^3 f_X(\xi_p) f_Y(\eta_q)}$$

Nous obtenons donc le résultat suivant concernant la convergence en loi de l'estimateur $\hat{\xi}_p / \hat{\eta}_q$ du rapport ξ_p / η_q des quantiles théoriques ξ_p et η_q :

$$\sqrt{n} \left(\frac{\hat{\xi}_p}{\hat{\eta}_q} - \frac{\xi_p}{\eta_q} \right) \xrightarrow{L} N \left(0, \frac{p(1-p)}{\eta_q^2 f_X^2(\xi_p)} - \frac{2\xi_p(F_{X,Y}(\xi_p, \eta_q) - pq)}{\eta_q^3 f_X(\xi_p) f_Y(\eta_q)} + \frac{\xi_p^2 q(1-q)}{\eta_q^4 f_Y^2(\eta_q)} \right)$$

Dans le cas où $p = q = 0,5$ et que de plus la loi du vecteur aléatoire (X, Y) est symétrique, le résultat ci-dessus montre que $\hat{\xi}_{0,5} / \hat{\eta}_{0,5}$ est un estimateur asymptotiquement sans biais et convergent du rapport des moyennes théoriques $\mathbb{E}(X) / \mathbb{E}(Y) = \xi_{0,5} / \eta_{0,5}$. Il pourrait donc, en présence de données aberrantes, constituer une alternative intéressante à l'estimateur usuel \bar{X} / \bar{Y} et se substituer, dans ce cas, à ce dernier.

3.4. Remarque

On peut également appliquer le théorème de représentation des quantiles de Bahadur et le théorème de Slutsky aux variables aléatoires $\hat{\xi}_p$ et $\hat{\eta}_q$ et considérer dans ce cas l'expression :

$${}^t (\hat{\xi}_p - \xi_p, \hat{\eta}_q - \eta_q) = {}^t \left(\frac{p - F_X^{(n)}(\xi_p)}{f_X(\xi_p)}, \frac{q - F_Y^{(n)}(\eta_q)}{f_Y(\eta_q)} \right) + \mathbf{R}_n {}^t \mathbf{1}$$

où $F_X^{(n)}$ et $F_Y^{(n)}$ désignent, dans ce cas, les fonctions de répartition marginales empiriques des composantes X et Y . Posons :

$$\left(F_X^{(n)}(\xi_p), F_Y^{(n)}(\eta_q) \right) = \frac{1}{n} \sum_{i=1}^n (\mathbb{I}_{]-\infty, \xi_p]}(X_i), \mathbb{I}_{]-\infty, \eta_q]}(Y_i))$$

et notons que les vecteurs $Z_i = {}^t (\mathbb{I}_{]-\infty, \xi_p]}(X_i), \mathbb{I}_{]-\infty, \eta_q]}(Y_i))$ $i = 1, 2, \dots, n$ sont indépendants et identiquement distribués de vecteur moyen $\mu = {}^t(p, q)$ et de matrice de variance-covariance Σ donnée par :

$$\Sigma = \begin{bmatrix} p(1-p) & F_{X,Y}(\xi_p, \eta_q) - pq \\ F_{X,Y}(\xi_p, \eta_q) - pq & q(1-q) \end{bmatrix}$$

Il en découle que :

$$\sqrt{n} \left(\frac{1}{n} \sum_{i=1}^n Z_i - \mu \right) \xrightarrow{L} N(0, \Sigma)$$

et donc ${}^t (F_X^{(n)}(\xi_p), F_Y^{(n)}(\eta_q))$ est un vecteur aléatoire “asymptotiquement normal” de vecteur moyen $\mu = {}^t(p, q)$ et de matrice de variance-covariance $\frac{1}{n}\Sigma$. On peut alors conclure que :

$${}^t \left(\frac{p - F_X^{(n)}(\xi_p)}{f_X(\xi_p)}, \frac{q - F_Y^{(n)}(\eta_q)}{f_Y(\eta_q)} \right) \rightsquigarrow AN \left(0, \frac{1}{n} \Sigma' \right)$$

où :

$$\Sigma' = \begin{bmatrix} \frac{p(1-p)}{f_X^2(\xi_p)} & \frac{F_{X,Y}(\xi_p, \eta_q) - pq}{f_X(\xi_p) f_Y(\eta_q)} \\ \frac{F_{X,Y}(\xi_p, \eta_q) - pq}{f_X(\xi_p) f_Y(\eta_q)} & \frac{q(1-q)}{f_Y^2(\eta_q)} \end{bmatrix}$$

Il vient donc :

$$\begin{aligned} \sqrt{n} {}^t (\hat{\xi}_p - \xi_p, \hat{\eta}_q - \eta_q) &= \sqrt{n} {}^t \left(\frac{p - F_X^{(n)}(\xi_p)}{f_X(\xi_p)}, \frac{q - F_Y^{(n)}(\eta_q)}{f_Y(\eta_q)} \right) \\ &\quad + O \left(n^{-\frac{1}{4}} (\text{Log} n)^{\frac{3}{4}} \right) {}^t \mathbf{1} \end{aligned}$$

et tenant compte du fait que $O \left(n^{-\frac{1}{4}} (\text{Log} n)^{\frac{3}{4}} \right) {}^t \mathbf{1} \xrightarrow{p.s} (0, 0)$, on a :

$$\sqrt{n} {}^t (\hat{\xi}_p - \xi_p, \hat{\eta}_q - \eta_q) \xrightarrow{L} N(0, \Sigma')$$

Ce résultat n'est autre que celui obtenu par Siddiqui [13], mais l'approche s'appuyant sur le théorème de représentation des quantiles de Bahadur, en rend la démonstration plus simple. De plus dans le cas où $p = q = 0,5$ on retrouve le résultat de Maritz [10].

4. Simulations et résultats

Les résultats des simulations présentés ci-dessous, issus en partie des travaux de Mellal [11], se composent de deux parties. Dans la première partie, et bien que cela ne soit pas directement l'objet du présent travail, nous comparons, dans le cas d'une population finie (situation la plus fréquente en échantillonnage) et uniquement de manière empirique, le quotient $\hat{\xi}_{0,5} / \hat{\eta}_{0,5}$ à d'autres estimateurs concurrents du paramètre $\mathbb{E}(X) / \mathbb{E}(Y)$. Dans la seconde partie, nous présentons, dans le cas de deux lois bidimensionnelles continues, quelques résultats concernant la normalité asymptotique du quotient de deux quantiles ainsi que le comportement des estimateurs $\hat{\xi}_{0,5} / \hat{\eta}_{0,5}$ et \bar{X} / \bar{Y} en présence de données aberrantes.

4.1. Cas d'une population finie

À l'aide des techniques de simulation usuelles, telles que décrites par exemple dans Devroye [4], on génère des échantillons $(x_i, y_i) \ i = 1, 2, \dots, n$, provenant d'une population finie de taille 1000 issue d'une loi normale $N(\mu, \Sigma)$ où :

$$\mu = {}^t(3, 7) \text{ et } \Sigma = \begin{bmatrix} 0,5 & 0,7 \\ 0,7 & 2 \end{bmatrix}$$

Afin d'estimer le rapport des moyennes $\mathbb{E}(X) / \mathbb{E}(Y) = 3 / 7 = 0,42857$, on considère les estimateurs suivants :

- Rapport des moyennes empiriques : \bar{x} / \bar{y}
- Rapport des médianes empiriques : $\hat{\xi}_{0,5} / \hat{\eta}_{0,5}$
- Rapport des moyennes tronquées empiriques : \bar{x}_t / \bar{y}_t , où :

$$(\bar{x}_t, \bar{y}_t) = \left(\frac{1}{n - 2[n\alpha]} \sum_{i=[n\alpha]+1}^{n-[n\alpha]} x_{(i)}, \frac{1}{n - 2[n\alpha]} \sum_{i=[n\alpha]+1}^{n-[n\alpha]} y_{(i)} \right)$$

avec $\alpha = 0,1 ; 0,2 ; 0,3$

– Rapport des moyennes censurées empiriques : $\overline{x_w} / \overline{y_w}$, où :

$$\overline{x_w} = \frac{1}{n} \left([n\alpha] x_{([n\alpha])} + \sum_{i=[n\alpha]+1}^{n-[n\alpha]} x_{(i)} + [n\alpha] x_{(n-[n\alpha]+1)} \right),$$

$$\overline{y_w} = \frac{1}{n} \left([n\alpha] y_{([n\alpha])} + \sum_{i=[n\alpha]+1}^{n-[n\alpha]} y_{(i)} + [n\alpha] y_{(n-[n\alpha]+1)} \right)$$

avec $\alpha = 0, 1 ; 0, 2 ; 0, 3$

– Rapport des estimateurs de Hodges-Lehmann : med_{z^1} / med_{z^2} , où :

$$z^1 = \left\{ \frac{1}{2} (x_{(i)} + x_{(j)}) \quad i \leq j = 1, 2, \dots, n \right\},$$

$$z^2 = \left\{ \frac{1}{2} (y_{(i)} + y_{(j)}) \quad i \leq j = 1, 2, \dots, n \right\}$$

À l'aide de 500 échantillons de tailles n , ($n = 10, 20, 30, 40, 50$) issus de la population considérée ci-dessus et pour chacun des cinq estimateurs précédents, on détermine la moyenne, la variance et l'intervalle de confiance symétrique interquantile de niveau 0, 95 des estimations obtenues. On répète alors 200 fois cette opération afin d'estimer la longueur théorique des intervalles de confiance et l'on détermine de plus la moyenne générale des estimations ainsi que la moyenne des variances obtenues à chaque répétition.

Plus précisément, soit T un estimateur de $\theta = \mathbb{E}(X) / \mathbb{E}(Y)$. Pour chaque répétition r ($r = 1, 2, \dots, 200$), on dispose de 500 estimations $t_{1,r}, t_{2,r}, \dots, t_{500,r}$ de θ à partir desquelles l'on détermine les valeurs empiriques suivantes :

$\bar{t}_r =$ Moyenne des estimations ($t_{i,r}$) ; $s_{t_r}^2 =$ Variance des estimations ($t_{i,r}$)

$L_r =$ Longueur de l'intervalle de confiance symétrique, pour θ , de niveau 0, 95

Les valeurs numériques qui apparaissent dans les tableaux ci-après sous les noms de Moyenne, Variance et "Long^r.intervalle" désignent respectivement les quantités suivantes :

$$\frac{1}{200} \sum_{r=1}^{r=200} \bar{t}_r ; \quad \frac{1}{200} \sum_{r=1}^{r=200} s_{t_r}^2 \quad \text{et} \quad \frac{1}{200} \sum_{r=1}^{r=200} L_r$$

Les résultats des simulations montrent que le quotient des médianes fournit une estimation du quotient des moyennes tout à fait comparable à celle que donnent les autres estimateurs et en particulier à ceux de Hodges-Lehmann, en étant, de surcroît,

d'un usage plus aisé.

Légende

- RMOYT : Rapport des moyennes tronquées
- RMOYW : Rapport des moyennes censurées
- REHOLE : Rapport des estimateurs de Hodges-Lehmann
- RMOYE : Rapport des moyennes
- RMEDE : Rapport des médianes

Estimations du rapport $\mathbb{E}(X) / \mathbb{E}(Y) = 0,42857$

n=10 :

Estimateur	Moyenne	Variance	“Long ^r .intervalle”
RMOYT ($\alpha = 0,1$)	0,4294	23×10^{-4}	0,1357
RMOYT ($\alpha = 0,2$)	0,4290	30×10^{-4}	0,1535
RMOYT ($\alpha = 0,3$)	0,4328	47×10^{-4}	0,1957
RMOYW ($\alpha = 0,1$)	0,4302	26×10^{-4}	0,1463
RMOYW ($\alpha = 0,2$)	0,4303	43×10^{-4}	0,1864
RMOYW ($\alpha = 0,3$)	0,4345	66×10^{-4}	0,2291
REHOLE	0,4291	19×10^{-4}	0,1702
RMOYE	0,4293	18×10^{-4}	0,1661
RMEDE	0,4291	25×10^{-4}	0,1858

n=20 :

Estimateur	Moyenne	Variance	“Long ^r .intervalle”
RMOYT ($\alpha = 0,1$)	0,4273	12×10^{-4}	0,0977
RMOYT ($\alpha = 0,2$)	0,4286	16×10^{-4}	0,1129
RMOYT ($\alpha = 0,3$)	0,4292	23×10^{-4}	0,1338
RMOYW ($\alpha = 0,1$)	0,4297	14×10^{-4}	0,1062
RMOYW ($\alpha = 0,2$)	0,4291	27×10^{-4}	0,1475
RMOYW ($\alpha = 0,3$)	0,4318	50×10^{-4}	0,2029
REHOLE	0,4284	9×10^{-4}	0,1194
RMOYE	0,4281	9×10^{-4}	0,1124
RMEDE	0,4271	13×10^{-4}	0,1333

n=30 :			
Estimateur	Moyenne	Variance	“Long ^r .intervalle”
RMOYT ($\alpha = 0,1$)	0,4280	7×10^{-4}	0,0764
RMOYT ($\alpha = 0,2$)	0,4281	10×10^{-4}	0,0911
RMOYT ($\alpha = 0,3$)	0,4284	15×10^{-4}	0,1108
RMOYW ($\alpha = 0,1$)	0,4285	11×10^{-4}	0,0927
RMOYW ($\alpha = 0,2$)	0,4282	23×10^{-4}	0,1351
RMOYW ($\alpha = 0,3$)	0,4299	43×10^{-4}	0,1873
REHOLE	0,4279	6×10^{-4}	0,0963
RMOYE	0,4281	6×10^{-4}	0,0921
RMEDE	0,4275	9×10^{-4}	0,1102
n=40 :			
Estimateur	Moyenne	Variance	“Long ^r .intervalle”
RMOYT ($\alpha = 0,1$)	0,4279	6×10^{-4}	0,0675
RMOYT ($\alpha = 0,2$)	0,4280	7×10^{-4}	0,0769
RMOYT ($\alpha = 0,3$)	0,4288	12×10^{-4}	0,0962
RMOYW ($\alpha = 0,1$)	0,4285	9×10^{-4}	0,0841
RMOYW ($\alpha = 0,2$)	0,4300	22×10^{-4}	0,1326
RMOYW ($\alpha = 0,3$)	0,4285	41×10^{-4}	0,1813
REHOLE	0,4279	5×10^{-4}	0,0827
RMOYE	0,4279	4×10^{-4}	0,0790
RMEDE	0,4275	3×10^{-4}	0,0962
n=50 :			
Estimateur	Moyenne	Variance	“Long ^r .intervalle”
RMOYT ($\alpha = 0,1$)	0,4283	4×10^{-4}	0,0587
RMOYT ($\alpha = 0,2$)	0,4279	6×10^{-4}	0,0694
RMOYT ($\alpha = 0,3$)	0,4282	8×10^{-4}	0,0811
RMOYW ($\alpha = 0,1$)	0,4284	8×10^{-4}	0,0794
RMOYW ($\alpha = 0,2$)	0,4294	20×10^{-4}	0,1255
RMOYW ($\alpha = 0,3$)	0,4305	39×10^{-4}	0,1763
REHOLE	0,4278	4×10^{-4}	0,0709
RMOYE	0,4279	3×10^{-4}	0,0692
RMEDE	0,4275	5×10^{-4}	0,0851

4.2. Cas de lois continues

Nous considérons, dans cette seconde partie, la loi normale $N(\mu, \Sigma)$ introduite ci-dessus où :

$$\mu = {}^t(3, 7) \text{ et } \Sigma = \begin{bmatrix} 0,5 & 0,7 \\ 0,7 & 2 \end{bmatrix}$$

et la loi de Gumbel $\mathcal{G}m(\alpha)$ de paramètre $\alpha = 0,5$ dont la fonction de répartition $F_{X,Y}(x, y)$ a pour expression :

$$F_{X,Y}(x, y) = 1 - e^{-x} - e^{-y} + e^{-x-y-0,5xy} \text{ où } x > 0 \text{ et } y > 0$$

et pour laquelle les rapports des moyennes et des médianes ont pour valeur 1. De plus, pour chaque loi, nous considérons également les rapports des quantiles empiriques $\hat{\xi}_p/\hat{\eta}_q$ pour les valeurs de (p, q) suivantes : $(0,7, 0,4)$ et $(0,25, 0,75)$. Les valeurs théoriques de ces rapports sont données par :

$$\begin{aligned} \mathcal{G}m(0,5) : \frac{\xi_{0,7}}{\eta_{0,4}} &= 2,35691; \frac{\xi_{0,25}}{\eta_{0,75}} = 0,20752 \\ N(\mu, \Sigma) : \frac{\xi_{0,7}}{\eta_{0,4}} &= 0,50752; \frac{\xi_{0,25}}{\eta_{0,75}} = 0,31722 \end{aligned}$$

4.2.1. Comparaison des estimateurs \bar{X}/\bar{Y} et $\hat{\xi}_{0,5}/\hat{\eta}_{0,5}$ de $\mathbb{E}(X)/\mathbb{E}(Y)$

Utilisant à nouveau les techniques de simulation décrites par Devroye [4], nous générons, pour chaque loi, 500 échantillons (ne contenant aucune valeur aberrante) de taille n ($n = 30, 50, 100$). Les résultats des simulations sont présentés ci-dessous :

Taille	n= 30	n= 50	n= 100
<i>Gm(0,5)</i>			
Moyenne des rapports \bar{x}/\bar{y}	1,059343	0,998716	1,021885
Moyenne des rapports $\hat{\xi}_{0,5}/\hat{\eta}_{0,5}$	1,135308	1,036030	1,038880
Variance des rapports \bar{x}/\bar{y}	0,133604	0,04639	0,035563
Variance des rapports $\hat{\xi}_{0,5}/\hat{\eta}_{0,5}$	0,288202	0,110809	0,055526
<i>N(μ, Σ)</i>			
Moyenne des rapports \bar{x}/\bar{y}	0,432612	0,430611	0,429557
Moyenne des rapports $\hat{\xi}_{0,5}/\hat{\eta}_{0,5}$	0,431503	0,427840	0,428736
Variance des rapports \bar{x}/\bar{y}	0,000157	0,000100	0,000050
Variance des rapports $\hat{\xi}_{0,5}/\hat{\eta}_{0,5}$	0,000434	0,000253	0,000142

On remarque ainsi expérimentalement qu'en l'absence de données aberrantes, les estimateurs \bar{X}/\bar{Y} et $\hat{\xi}_{0,5}/\hat{\eta}_{0,5}$ du rapport $\mathbb{E}(X)/\mathbb{E}(Y)$ se comparent fort bien l'un à l'autre et ce d'autant mieux, que la taille de l'échantillon est grande ce qui,

dans la pratique, se révèle très fréquent. On peut donc, dans ce cas, substituer sans crainte le rapport des médianes empiriques au rapport des moyennes empiriques afin d'estimer le rapport des moyennes théoriques.

4.2.2. Influence de données aberrantes sur les estimateurs \bar{X}/\bar{Y} et $\hat{\xi}_{0,5}/\hat{\eta}_{0,5}$

Afin d'évaluer l'impact de la présence de données aberrantes sur le comportement des estimateurs \bar{X}/\bar{Y} et $\hat{\xi}_{0,5}/\hat{\eta}_{0,5}$, nous avons procédé comme suit :

Nous avons simulé deux séries de 500 échantillons indépendants, de tailles respectives $n = 20, 30$, et 50 , du vecteur aléatoire (X, Y) de loi normale $N(\mu, \Sigma)$ où :

$$\mu = {}^t(3, 7) \text{ et } \Sigma = \begin{bmatrix} 0,5 & 0,7 \\ 0,7 & 2 \end{bmatrix}$$

Dans la première série (Série 1), nous avons remplacé au hasard, dans le cas des échantillons de tailles $n = 20$ et 30 , l'une des observations X_i par $\mathbb{E}(X) + 7\sigma_X$ et l'une des observations Y_i par $\mathbb{E}(Y) + 7\sigma_Y$, et dans le cas des échantillons de taille $n = 50$, nous avons remplacé au hasard, deux observations X_i et $X_{i'}$ par $\mathbb{E}(X) + 7\sigma_X$ et par $\mathbb{E}(X) + 10\sigma_X$ et deux observations Y_i et $Y_{i'}$ par $\mathbb{E}(Y) + 7\sigma_Y$ et par $\mathbb{E}(Y) + 10\sigma_Y$. Dans la seconde série (Série 2), nous avons conservé la même procédure de substitution pour les observations X_i ($n = 20, 30$) et pour les observations X_i et $X_{i'}$ ($n = 50$), alors que pour $n = 20$ et 30 , nous avons remplacé au hasard Y_i par $\mathbb{E}(Y) - 7\sigma_Y$ et que pour $n = 50$, nous avons remplacé au hasard Y_i et $Y_{i'}$ par $\mathbb{E}(Y) - 7\sigma_Y$ et par $\mathbb{E}(Y) - 10\sigma_Y$. On trouvera ci-après les résultats des deux simulations.

Série 1	n= 20	n= 30	n= 50
Moyenne des rapports \bar{x}/\bar{y}	0.497219	0.473850	0.514991
Moyenne des rapports $\hat{\xi}_{0,5}/\hat{\eta}_{0,5}$	0.439165	0.435861	0.438509
Variance des rapports \bar{x}/\bar{y}	0.000470	0.000255	0.000259
Variance des rapports $\hat{\xi}_{0,5}/\hat{\eta}_{0,5}$	0.000606	0.000494	0.000298

Série 2	n= 20	n= 30	n= 50
Moyenne des rapports \bar{x}/\bar{y}	0.499850	0.473688	0.513601
Moyenne des rapports $\hat{\xi}_{0,5}/\hat{\eta}_{0,5}$	0.443592	0.435305	0.438422
Variance des rapports \bar{x}/\bar{y}	0.000473	0.000240	0.000226
Variance des rapports $\hat{\xi}_{0,5}/\hat{\eta}_{0,5}$	0.000691	0.000474	0.000289

On constate ainsi expérimentalement que l'estimateur \bar{X}/\bar{Y} est plus sensible aux données aberrantes que l'estimateur $\hat{\xi}_{0,5}/\hat{\eta}_{0,5}$ et ce dernier devrait donc, dans la pratique, être préféré au rapport des moyennes empiriques, dans tous les cas où l'on redoute la présence de telles données dans l'échantillon considéré.

4.2.3. Comportement des estimateurs $\hat{\xi}_{0,7}/\hat{\eta}_{0,4}$ et $\hat{\xi}_{0,25}/\hat{\eta}_{0,75}$

En ce qui concerne les estimateurs de $\xi_{0,7}/\eta_{0,4}$ et de $\xi_{0,25}/\eta_{0,75}$, dont les valeurs théoriques sont respectivement 2,35691 et 0,20752 dans la cas de la loi $\mathcal{G}m(0, 5)$ et 0,50752 et 0,31722 dans le cas de la loi $N(\mu, \Sigma)$, nous obtenons, toujours à l'aide des mêmes techniques, les résultats suivants :

Taille	n= 30	n= 50	n= 100
$\mathcal{G}m(0, 5)$			
Moyenne des rapports $\hat{\xi}_{0,7}/\hat{\eta}_{0,4}$	2,620186	2,396945	2,383535
Moyenne des rapports $\hat{\xi}_{0,25}/\hat{\eta}_{0,75}$	0,254326	0,212714	0,220094
Variance des rapports $\hat{\xi}_{0,7}/\hat{\eta}_{0,4}$	1,635145	0,544573	0,316988
Variance des rapports $\hat{\xi}_{0,25}/\hat{\eta}_{0,75}$	0,013761	0,003908	0,002997
$N(\mu, \Sigma)$			
Moyenne des rapports $\hat{\xi}_{0,7}/\hat{\eta}_{0,4}$	0,511239	0,508195	0,508062
Moyenne des rapports $\hat{\xi}_{0,25}/\hat{\eta}_{0,75}$	0,318685	0,320113	0,318160
Variance des rapports $\hat{\xi}_{0,7}/\hat{\eta}_{0,4}$	0,000667	0,000389	0,000217
Variance des rapports $\hat{\xi}_{0,25}/\hat{\eta}_{0,75}$	0,000416	0,000283	0,000140

4.2.4. Normalité asymptotique

À l'aide des 500 simulations d'échantillons de taille 100 précédentes, nous illustrons dans les figures 1 à 6 ci-après, pour les deux lois considérées, les densités empiriques des estimateurs $\hat{\xi}_{0,5}/\hat{\eta}_{0,5}$, $\hat{\xi}_{0,7}/\hat{\eta}_{0,4}$ et $\hat{\xi}_{0,25}/\hat{\eta}_{0,75}$ et nous indiquons, entre parenthèses, les valeurs des variances empiriques et des variances théoriques (en caractères gras dans les figures 1 à 6), déduites des lois asymptotiques correspondantes pour $n = 100$.

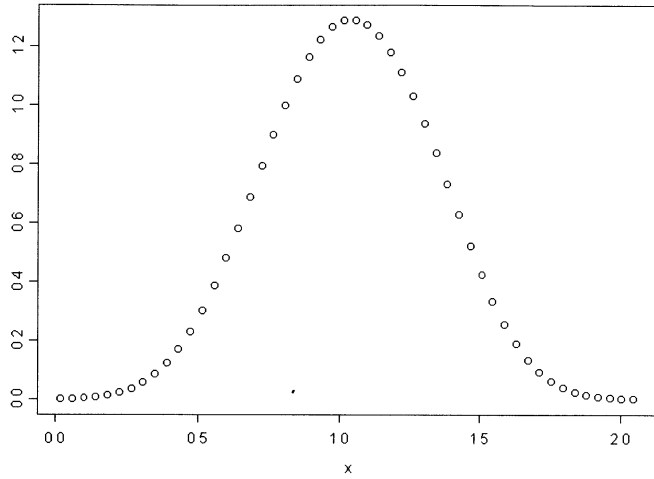


FIGURE 1
 $\mathcal{G}m(0,5) : \hat{\xi}_{0,5} / \hat{\eta}_{0,5}$
 (0,055526 ; 0,050517)

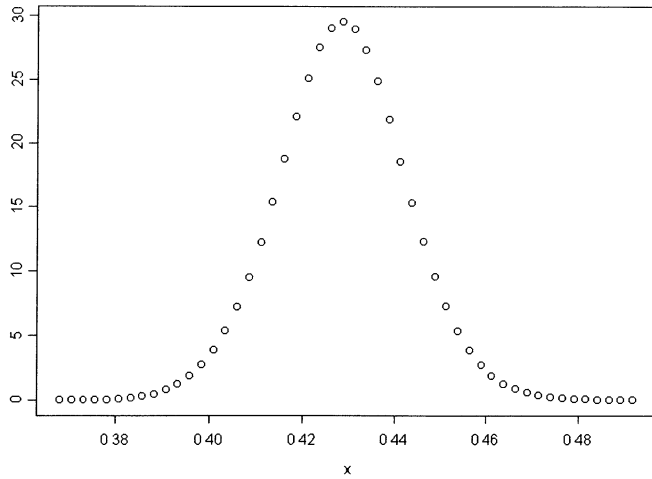


FIGURE 2
 $N(\mu, \Sigma) : \hat{\xi}_{0,5} / \hat{\eta}_{0,5}$
 (0,000142 ; 0,000142)

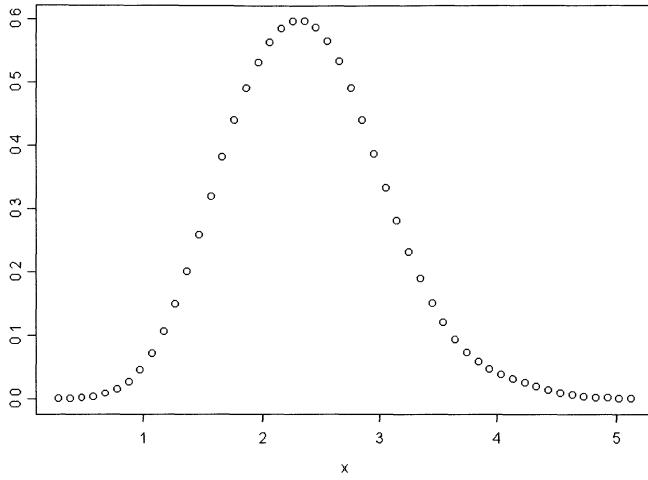


FIGURE 3
 $Gm(0, 5) : \hat{\xi}_{0,7} / \hat{\eta}_{0,4}$
 (0,316988 ; 0,174270)

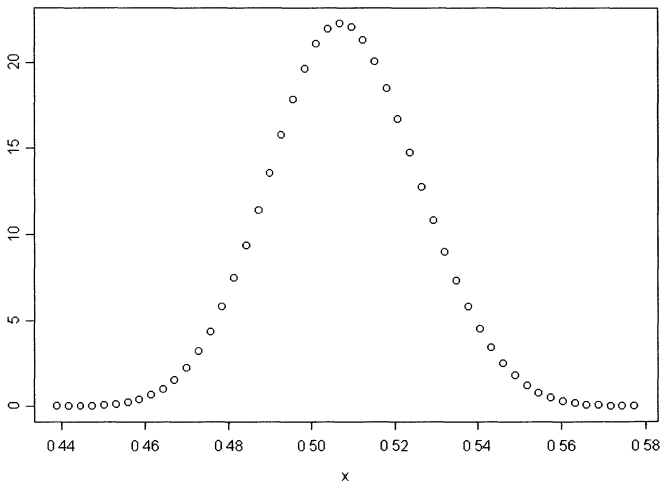


FIGURE 4
 $N(\mu, \Sigma) : \hat{\xi}_{0,7} / \hat{\eta}_{0,4}$
 (0,000217 ; 0,000162)

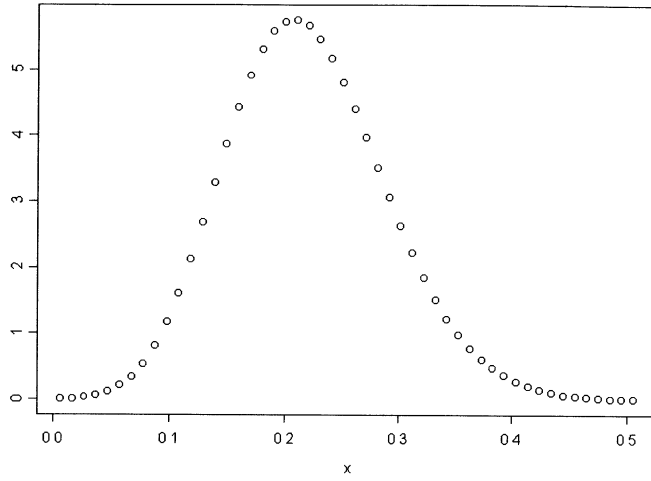


FIGURE 5
 $Gm(0,5) : \hat{\xi}_{0,25} / \hat{\eta}_{0,75}$
 (0,002997 ; 0,0034168)

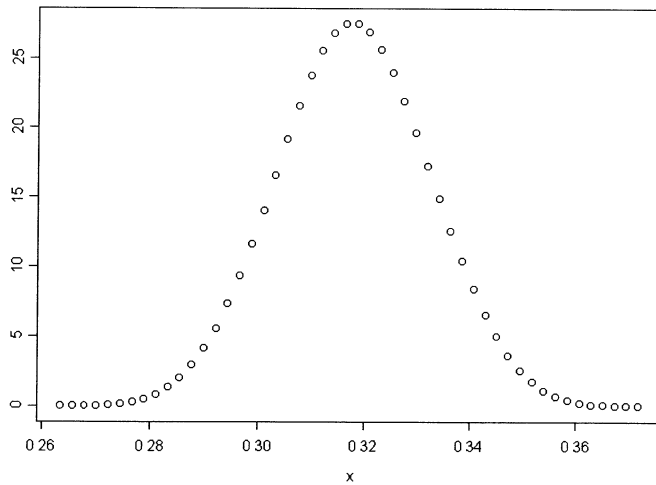


FIGURE 6
 $N(\mu, \Sigma) : \hat{\xi}_{0,25} / \hat{\eta}_{0,75}$
 (0,000140 ; 0,000090)

5. Conclusion

Notant la sensibilité aux données aberrantes de l'estimateur \bar{X}/\bar{Y} du rapport des moyennes théoriques $\mathbb{E}(X)/\mathbb{E}(Y)$, nous nous sommes intéressés, afin de pallier cette situation, au rapport de deux médianes et, plus généralement, au rapport de deux quantiles empiriques. Nous avons montré, dans le cadre des lois continues, que le rapport des quantiles empiriques $\hat{\xi}_p/\hat{\eta}_q$ est un estimateur asymptotiquement normal, sans biais et convergent du rapport des quantiles théoriques correspondant ξ_p/η_q , ce qui, par exemple dans le cas des lois symétriques, confère au rapport des médianes empiriques $\hat{\xi}_{0,5}/\hat{\eta}_{0,5}$ le statut d'estimateur concurrent à \bar{X}/\bar{Y} , pour le paramètre $\mathbb{E}(X)/\mathbb{E}(Y) = \xi_{0,5}/\eta_{0,5}$ de la loi du vecteur aléatoire (X, Y) . Par ailleurs, nous avons constaté expérimentalement que, le rapport des médianes empiriques $\hat{\xi}_{0,5}/\hat{\eta}_{0,5}$ se compare fort bien, dans le cadre précédent, à certains estimateurs du quotient des moyennes théoriques $\mathbb{E}(X)/\mathbb{E}(Y)$, qu'il est asymptotiquement normal et que, de surcroît, il est moins sensible aux données aberrantes que ne l'est, par exemple, l'estimateur usuel \bar{X}/\bar{Y} . Enfin, pour donner suite à ce travail, nous étudierons dans un cadre plus formel, l'efficacité relative asymptotique, la fonction d'influence et la robustesse des estimateurs de la forme $\hat{\xi}_p/\hat{\eta}_q$.

Références

- [1] BAHADUR R.R., *A note on quantiles in large samples*, Ann. Math. Statist., 37 (1966), 577-580.
- [2] COCHRAN W.G., *Sampling Techniques*, Third Edition, John Wiley & Sons, (1977).
- [3] DAVID H.A., *Order statistics*, John Wiley & Sons, (1981).
- [4] DEVROYE L., *Non-Uniform Random Variate Generation*, Springer-Verlag, (1986).
- [5] FRÉCHET M., *Sur les tableaux de corrélation dont les marges sont données*, Ann. Univ. Lyon, Sér.3, 14 (1951), 53-77.
- [6] GALAMBOS J., *Order statistics of samples from multivariate distribution*, J. Amer. Stat. Assoc., 70, no.351 (1975), 674-680.
- [7] GENEST C. et MACKAY R.J., *Copules archimédiennes et familles de lois bidimensionnelles dont les marges sont données*, La Revue Canadienne de Statistique, vol 14, 2 (1986), 145-159.
- [8] KIMELDORF G. and SAMPSON A.R., *Uniform representations of bivariate distributions*, Com. Statist. 4 (1975), 617-627.
- [9] LING C.H., *Representation of associative functions*, Publ. Math. Debrecen, 12 (1965), 189-212.
- [10] MARITZ J.S., *Estimating the covariance matrix of bivariate medians*, Statist. Prob. Lett. 12, no.4 (1991) 305-309.

- [11] MELLAL N., *Rapport des quantiles des composantes d'une loi bidimensionnelle*, Thèse de doctorat, Université de Sherbrooke, Canada, (2000).
- [12] SERFLING R.J., *Approximation theorems of mathematical statistics*, John Wiley & Sons, (1980).
- [13] SIDDIQUI M.M., *Distribution of quantiles in samples from a bivariate population*, Journal of Research of the National Standards-B Mathematics and Mathematical Physics, vol.64B, no3, (1960), 145-150.
- [14] SKLAR A., *Fonctions de répartition à n dimensions et leurs marges*, Publ. Inst. Statist. Univ. Paris, 8 (1959), 229-231.
- [15] SRIVASTAVA M.S., *Asymptotic independence of certain statistics connected with the extreme order statistics in a bivariate population*, Sankya : the Indian journal of Statistics : Series A, vol.29,2 (1967), 175-182.