JOURNAL DE LA SOCIÉTÉ FRANÇAISE DE STATISTIQUE

OLIVIER NUÑEZ
DANIEL PEÑA

Discussion de l'article de Fréchet (1940) « Sur une limitation très générale de la dispersion de la médiane »

Journal de la société française de statistique, tome 147, n° 2 (2006), p. 67-72

<http://www.numdam.org/item?id=JSFS_2006__147_2_67_0>

© Société française de statistique, 2006, tous droits réservés.

L'accès aux archives de la revue « Journal de la société française de statistique » (http://publications-sfds.math.cnrs.fr/index.php/J-SFdS) 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/

DISCUSSION DE L'ARTICLE DE FRÉCHET (1940) SUR UNE LIMITATION TRÈS GÉNÉRALE DE LA DISPERSION DE LA MÉDIANE

Olivier NUÑEZ *, Daniel PEÑA*

1. Introduction

Dans cet intéressant travail daté de 1940, Fréchet discute quelques unes des idées fondamentales de la robustesse et démontre être un des précurseurs de l'approche minimax qui sera centrale dans les travaux postérieurs de Huber sur ce sujet. Pour développer ces idées, Fréchet réalise une étude rigoureuse de l'efficacité de la médiane qu'il compare à celle de la moyenne arithmétique. Il vise à montrer dans cette étude que cette efficacité dépend de manière cruciale du choix de la métrique utilisée pour mesurer les écarts. Ainsi, selon que l'on utilise l'écart quadratique moyen ou l'écart probable, on est amené à conclure ou bien à une légère supériorité de la moyenne ou à une supériorité « infinie » de la médiane. Ce résultat est basé sur des majorations de l'efficacité de ces deux estimateurs qu'obtient Fréchet pour différentes métriques. Ces majorations permettent de comparer le comportement de la médiane et de la moyenne dans des situations extrêmement défavorables pour l'une ou pour l'autre de ces mesures. Pour illustrer son propos, Fréchet conçoit ensuite une expérience de simulation dans laquelle les donné générées par le tirage successif de cartes à jouer sont issues d'une distribution dont la kurtosis est contrôlée par l'expérimentateur. À l'aide de cette expérience, Fréchet confirme de manière empirique que la médiane peut être beaucoup plus efficace que la moyenne lorsque la distribution a des queues lourdes.

Afin de replacer la contribution de Fréchet dans une perspective historique, nous consacrerons la section suivante a une brève histoire du débat entre médiane et moyenne. Dans la seconde et dernière section de cet discussion, nous mènerons une étude de l'efficacité relative de la médiane par rapport à la moyenne au sein de la famille puissance exponentielle dont le paramètre permet de contrôler la kurtosis. Cette étude est très similaire dans ses fins à l'étude empirique réalisée par Fréchet.

Notons enfin que, par souci de clarté, nous emprunterons la terminologie et les notations utilisées par Fréchet.

* Departamento de Estadística, Universidad Carlos III de Madrid, nunez@est-econ.uc3m.es, daniel.pena@uc3m.es

2. Historique du débat

La discution sur les avantages relatifs de la moyenne par rapport à la médiane comme estimateur de la valeur centrale d'une distribution symétrique, apparaît dès les origines de la statistique. Les livres de Barnett et Lewis (1998) et Hampel *et al.* (1986) présentent une bonne analyse historique de cette problèmatique. Ainsi, Poisson en 1824 soutenait déjà que la moyenne arithmétique n'a pas de bonnes propriétés lorsque l'on considère la distribution de probabilité dont la densité est

$$f(x) = \frac{1}{\pi(1+x^2)}$$

et que nous appelons aujourd'hui distribution de Cauchy, qui de fait l'utilisa quelques années plus tard à des fins similaires. Fisher (1922) dans sa critique des estimateurs des moments proposés par Pearson, démontre que la moyenne arithmétique et la variance sont très peu efficaces pour la majorité des distributions de la famille de Pearson et qu'il convient d'utiliser ces estimateurs en des proches voisinages de la distribution normale.

Néanmoins, la moyenne arithmétique a eu un rôle central comme estimateur de la valeur centrale et ce en raison de deux résultats fondamentaux. Le premier est dû à Gauss qui proposa la méthode des moindres carrés et justifia ainsi l'usage de la moyenne arithmétique comme valeur typique d'un échantillon de mesures supposées normales. La grande simplicité du calcul dans la méthode des moindres carrés, comparée à la méthode basée sur les écarts absolus jadis utilisée entre autres par Laplace, a été un facteur déterminant de la popularité de la moyenne dans la pratique actuelle. Le second résultat est la méthode du maximum de vraisemblance introduite par Fisher, qui justifia son usage et son optimalité pour la distribution normale. Comme l'indique Fréchet, Fisher démontra en outre que l'efficacité relative de la médiane par rapport à la moyenne en terme d'écart quadratique est de 79.78% pour cette même distribution. Autrement dit, dans ce cas l'écart quadratique de la médiane est de l'ordre de 25% supérieur à celui de la moyenne.

Tukey (1960) compare ces deux estimateurs dans des voisinages de la distributions normale utilisant une contamination symétrique du type :

$$f(x) = (1-\varepsilon)\phi(x;\mu,\sigma) + \varepsilon\phi(x;\mu,k\sigma),$$

où $k > 0$ et $\phi(.;\mu,\sigma)$ désigne la densité de la distribution normale de moyenne μ et d'écart type σ. Il est facile de voir que, pour cette famille, la variance de la moyenne empirique est

$$\sigma^2_{V_n} = \frac{\sigma^2}{n}\left[1+\varepsilon\left(k^2-1\right)\right],$$

tandis que la variance de la médiane est, asymptotiquement

$$\sigma^2_{M_n} = \frac{1}{4nf^2(m)} = \frac{\sigma^2}{n}\frac{\pi}{2}\frac{k^2}{\left[\varepsilon+k(1-\varepsilon)\right]^2}.$$

Par conséquent, pour n suffisamment grand, on a :

$$\frac{\sigma_{V_n}}{\sigma_{M_n}} = \sqrt{\frac{2}{\pi}} \left[\varepsilon/k + (1-\varepsilon)\right] \sqrt{1+\varepsilon\left(k^2-1\right)}$$

Notons, que lorsque k croît la variance de la moyenne tend vers l'infini, alors que celle de la médiane reste bornée. En outre, lorsque ε est petit, la perte d'efficacité de la médiane est minime dans le cas de la distribution normale. De nouveau, nous constatons qu'en augmentant la masse des queues de la distribution, ce qui se traduit ici par une augmentation de la contamination, l'efficacité relative de la médiane par rapport à la moyenne tend vers l'infini. Plus récemment, Sherman (1997) réalisa une étude similaire au sein de la famille puissance exponentielle dont on s'inspirera dans l'étude de la section suivante.

Enfin, une comparaison plus générale fut réalisée postérieurement par Huber (1964) qui proposa des voisinages de contamination généraux d'une distribution F du type

$$(1-\varepsilon)F(x-\mu) + \varepsilon G(x-\mu)$$

où G est une distribution arbitraire.

Un des aspects importants de l'article de Fréchet, est de considérer l'efficacité pour des métriques différentes de l'écart quadratique moyen. Concrètement, il utilise l'écart absolu moyen et l'écart probable (ou intervalle interquartile) :

$$\frac{1}{n}\sum |x_i - M_n| \quad et \quad \frac{Q_3 - Q_1}{2}$$

Ces deux mesures de la dispersion sont plus robustes que la variance et sont en outre des estimateurs bien plus efficaces du paramètre d'échelle de distributions à forte kurtosis. Sur ce point, Fréchet anticipe un résultat désormais établi dans la litérature sur l'absence d'une mesure naturelle de la variabilité et du paramètre d'échelle. Une étude postérieure sur les mesures proposées par Fréchet peut être trouvée au chapitre 12 du livre de Hoaglin, Mosteller et Tukey (1983). Actuellement, la théorie de la robustesse donne un rôle prépondérant à l'écart absolu (en anglais, median of absolute deviations, MAD), même si l'écart interquartile continue à être aussi utilisé.

3. Efficacité relative dans la famille puissance exponentielle

À partir des majorations données par Fréchet, on déduit directement la minoration suivante pour l'efficacité asymptotique de la médiane relative à celle de moyenne :

$$\lim_{n\to\infty} \frac{\sigma_{V_n}}{\sigma_{M_n}} \geqslant \frac{\sqrt{3}}{3} \simeq 58\%$$

Cette borne inférieure correspond au cas de données provenant d'une distribution uniforme. C'est le cas le plus défavorable pour la médiane. Cependant,

à mesure que la kurtosis de la distribution dont sont issues les données s'accroît, cette minoration devient de moins en moins pertinente. Afin d'illustrer ce fait, on peut étudier l'efficacité de M_n relative à V_n au sein de la famille puissance exponentielle de densité

$$f(x;a) = \frac{1}{2\sigma\Gamma(1+a)} \exp\left(-\left|\frac{x-\mu}{\sigma}\right|^{1/\alpha}\right), \quad \alpha \in]0,\infty)$$

symétrique autour de μ et de paramètre d'échelle σ. Cette famille contient en particulier les distributions de Laplace ($\alpha = 1$) et la gaussienne ($\alpha = 0.5$). Notons, que lorsque $\alpha \to 0$ cette densité converge vers la densité uniforme. Par contre, lorsque $\alpha \to \infty$ on obtient des distributions dont les queues sont de plus en plus massives. La Figure 1 montre les densités de différents membres de cette famille pour $\mu = 0$ et $\sigma = 1$.

Notons, que cette famille fut utilisée par Box et Tiao (1973) dans le but de comparer des estimateurs bayésiens. On peut montrer (Sherman, 1997) que la médiane et la moyenne ont la même efficacité lorsque $\alpha = .714$, ce qui correspond à une distribution proche de la normale. Cependant, à mesure que α augmente, la moyenne devient nettement moins efficace que la médiane. Ainsi, pour $\alpha = 1$ (distribution de Laplace), la moyenne a un écart quadratique moyen 41% supérieur à celui de la médiane, et pour $\alpha = 5$ le rapport atteint 500!

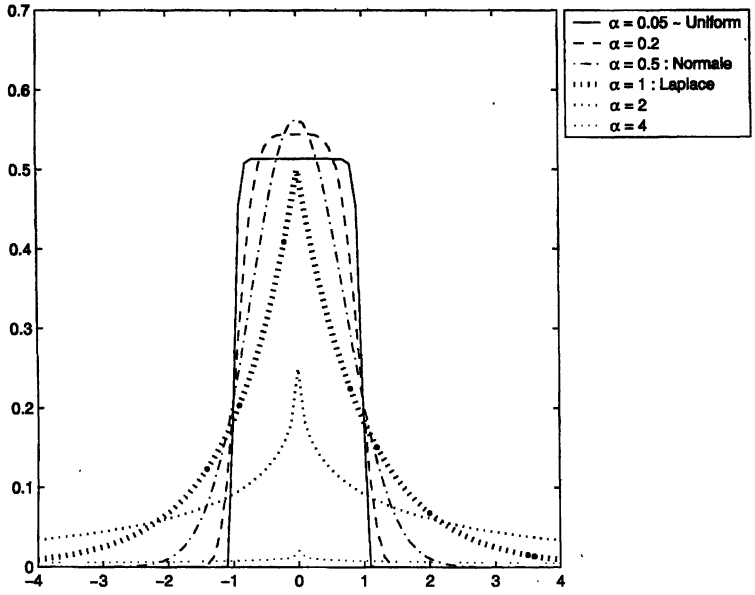

FIG 1. — Densités de la famille puissance exponentielle pour différentes valeurs de α.

Plus généralement, on montre que l'efficacité relative (asymptotique) de la médiane par rapport à la moyenne est selon la valeur de α égale à

$$\lim_{n\to\infty} \frac{\sigma_{V_n}}{\sigma_{M_n}} = \frac{\sqrt{3}}{3}\sqrt{\frac{\Gamma(3\alpha+1)}{\Gamma(\alpha+1)^3}}$$

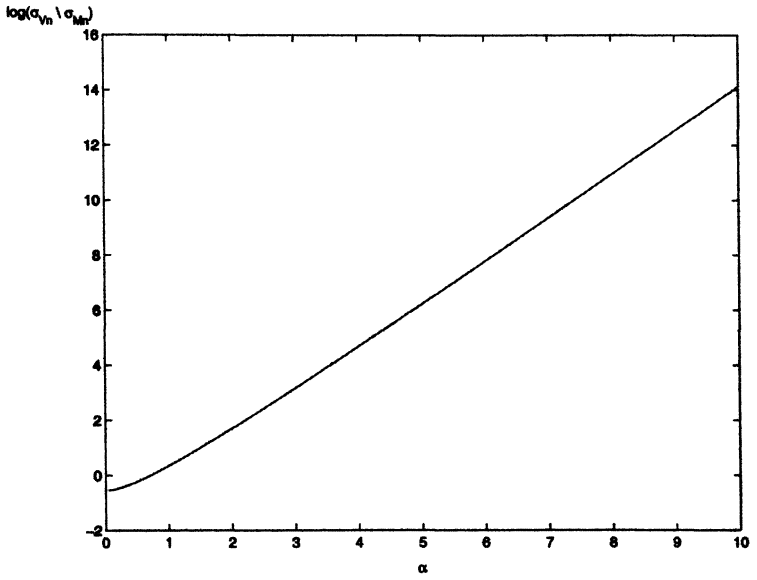

FIG 2. — Logarithme de l'efficacité relative de la médiane par rapport à la moyenne en fonction de α.

Notons que lorsque $\alpha \to 0$ (c.à.d, la distribution devient uniforme), ce rapport décroît vers $\sqrt{3}/3$ qui correspond à la minoration donnée plus haut. En utilisant la formule de Binet pour la fonction log-gamma (voir Whittaker and Watson, 1990, p. 251), il est possible de montrer que le logarithme de cette efficacité relative croît à la même vitesse que $\frac{3}{2}\log(3)\alpha$ lorsque $\alpha \to \infty$. La figure 2 montre les variations du logarithme de cette efficacité relative en fonction de α. Ce graphique confirme que le logarithme du rapport d'efficacité est approximativement linéaire, ce qui démontre que la médiane devient rapidement supérieure à la moyenne lorsque l'on augmente α.

Enfin, notons que l'efficacité relative asymptotique est la même pour les trois métriques considérées par Fréchet :

$$\lim_{n\to\infty} \frac{E_{V_n}}{E_{M_n}} = \lim_{n\to\infty} \frac{\theta_{V_n}}{\theta_{M_n}} = \lim_{n\to\infty} \frac{\sigma_{V_n}}{\sigma_{M_n}}.$$

Par conséquent, le comportement de l'efficacité relative en function de α décrit précédemment est aussi valide pour l'écart absolu moyen et l'écart probable.

4. Conclusion

Les outils d'analyse que propose Fréchet dans son article annoncent les concepts fondateurs de la théorie actuelle de la robustesse. Ainsi, dans la majoration générale de l'efficacité que donne Fréchet, est déjà en germe l'approche minimax que proposa ensuite Huber pour étudier la robustesse des estimateurs. Par ailleurs, l'expérience de simulation conçue par Fréchet contient déjà les ingrédients principaux développés dans la vaste étude de la robustesse réalisée à Princeton en 1970/71 par Bickel, Huber et Tukey entre autres chercheurs, dont l'objet était les mesures robustes de localisation pour des distributions à queues lourdes (voir Gross et Tukey, 1973) et qui jeta les bases de la théorie actuelle de la robustesse.

Références

BOX G.E. P. and TIAO, G. C. (1973) *Bayesian Inference in Statistical Analysis*, Wiley

BARNETT V. and LEWIS T. (1998) *Outliers in Statistical Data*, Wiley

FISHER R. A. (1922) On the mathematical foundations of theoretical statistics, *Philos.Trans. Roy. Soc. London*, A 222, 309-368.

GROSS A. M. and TUKEY J. W. (1973). The estimators of the Princeton robustness study, Tech. Rep. 38, Ser. 2, Dept. of Statistics, Princeton University, Princeton, N.J.

HAMPEL F. R. *et al.* (1986) *Robust Statistics*, Wiley.

HOAGLIN D.C., MOSTELLER F. y TUKEY J.W.(1983) *Understanding Robust ad Exploratory Data Analyses*. Wiley

HUBER P. J. (1964) Robust estimation of a location parameter, *Ann Math. Statist.*, 35, 73-101.

SHERMAN M. (1997) Comparing the Sample mean and the Sample Median : An Exploration in the Exponential Power Family, *The American Statistician*, 51,1,52-54.

TUKEY J.W. (1960) A survey of sampling from contaminated distributions, en *contributions to Probability and Statistics*, I. Olkin (ed), Stanford University Press.

WHITTAKER E. T. and WATSON G. N. (1990) *A Course in Modern Analysis*, 4th ed. Cambridge, England : Cambridge University Press.