JOURNAL DE LA SOCIÉTÉ FRANÇAISE DE STATISTIQUE

CHRISTOPHE CROUX

Discussion. Sur une limitation très générale de la dispersion médiane by M. Fréchet

Journal de la société française de statistique, tome 147, n° 2 (2006), p. 45-49

<http://www.numdam.org/item?id=JSFS_2006__147_2_45_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 : «SUR UNE LIMITATION TRÈS GÉNÉRALE DE LA DISPERSION MÉDIANE» BY M. FRÉCHET

Christophe CROUX *

1. Introduction

Let me start by thanking the Editor for having "discovered" this remarkable paper of Maurice Fréchet, and for having invited me to discuss it. In this paper, published in 1940, the sample median is promoted as an easy-to-compute estimator, and expressions for the dispersion of the sample median are computed. This dispersion can be measured by the standard deviation, but also by the interquartile range, the latter being called "l'écart probable" in this paper. One of the mathematical contributions of this paper is that a lower bound for the relative efficiency of the sample median with respect to the sample mean is presented. This lower bound is valid for all unimodal distributions, having the median as modus.

Fréchet already knew that the statistical efficiency of the sample median may well be superior to the efficiency of the sample mean. He notices that the dispersion of the sample mean may tend to infinity for heavy tailed distributions, and considers this as a serious drawback, of importance in applications. Although he is nowhere mentioning the robustness of the median with respect to outliers, he clearly understood that the median has good sampling properties under much milder conditions than the sample average. Several decades before the pioneering work of Peter Huber on robust statistics, this paper already contains several important robustness ideas.

In my discussion I will first rewrite some of the results of Maurice Fréchet in more modern notation, and then illustrate them by a modest simulation study.

2. Main Results

Let $X \sim F$ and denote M_n the sample median computed from a random sample from the distribution F, while $\bar{X}_n$ stands as usual for the sample

* Faculty of Economics and Applied Economics, K.U.Leuven, Naamsestraat 69, B-3000 Leuven, Belgium. E-mail : christophe.croux@econ.kuleuven.ac.be

mean. The sample mean has a wonderful property :

$$\frac{\mathrm{SD}(\bar{X}_n)}{\mathrm{SD}(X)} = \frac{1}{\sqrt{n}} \tag{1}$$

where SD stands for the Standard Deviation of any random variable. Hence, the reduction in SD by replacing an observation by a sample average, is constant and this for all distributions F. While the equality (1) does not hold for the sample median, a lower bound ("limitation") for the reduction in sampling variance has been derived in this paper. Fréchet correctly notices that if other measures of dispersion than SD are used, the right hand side of (1) will become dependent on the underlying distribution F.

It is common to compare the sampling variability of two estimators by computing their asymptotic relative efficiency (ARE). The ARE of the sample median with respect to the sample mean is

$$\mathrm{ARE} = \lim_{n\to\infty} \frac{\mathrm{SD}^2(\bar{X}_n)}{\mathrm{SD}^2(M_n)} = 4f(\mathrm{med}(X))^2\mathrm{SD}^2(X) \tag{2}$$

with f the density of the distribution F. Since, under regularity conditions, both estimators are asymptotically normal, replacing the SD by other measures of dispersion will not alter the value of the ARE. It is well known that the sample median is asymptotically normal if the density at the population median is strictly positive, and the sample mean if the second moment of F exists. If the expression in (2) is larger than 1, then the sample median is most efficient, otherwise the sample mean.

While the ARE is not bounded above, it follows form Fréchet's paper that, if the density f is unimodal with modus equal to its median, then

$$\mathrm{ARE} \geqslant \frac{1}{3}. \tag{3}$$

It is not mentioned explicitly in Fréchet's paper, but a distribution F attaining the lower bound in (3) is the rectangular distribution. It is worth mentioning that the same lower bound has been obtained in the seminal paper of Hodges and Lehmann (1956), for the relative efficiency of the sign test with respect to the classical t-test. From their result, it also follows that the ARE of the Hodges-Lehman estimator, being defined as

$$\underset{1\leqslant i<j\leqslant n}{\mathrm{med}} \frac{X_i + X_j}{2},$$

is larger than 0.864 (over the class of symmetric distributions). Hence, using the Hodges-Lehman estimator instead of the sample mean yields at most 14% loss in asymptotic efficiency. This result can even be improved on, using R-estimators with normal scores (*e.g.* Jurečková and Sen, 1996).

3. A Simulation Experiment

To illustrate the results of Maurice Fréchet, we conduct a modest simulation experiment. We generated $m = 10000$ samples of size $n = 10, 20, 50, 100$ and 1000 from the following distributions :

Nor A standard normal distributions $N(0, 1)$, where it is known that the sample mean is the maximum likelihood estimator. One has ARE $= 2/\pi = 0.64$.

Lap A Laplace distribution, with $f(x) = 0.5 \exp(-|x|)$, where it is known that the sample median is the maximum likelihood estimator. One has ARE $= 2$.

Uni A uniform distribution on $[-0.5; 0.5]$. This is the worst case distribution for the sample median. We have ARE $= 1/3$.

F0 A distribution with density $f(x) = |x|$, for x in $[-1, 1]$ and zero elsewhere. Note that the density at the median equals zero. This is an example of a distribution not covered by the theorem; the sample variability of the median is not "limited." We expect here ARE=0.

t2 A Student distribution with 2 degrees of freedom. The population mean is well defined, but this distribution has heavy tails and no second moment. We have ARE $= \infty$.

All of the above distributions are symmetric, hence the population median and mean coincide. Then we simulate finite sample relative efficiencies

$$\mathrm{ARE}_n = \frac{\mathrm{SD}_j^2 \bar{X}_n^j}{\mathrm{SD}_j^2 M_n^j} \tag{4}$$

with $\bar{X}_n^j$ and M_n^j, the average and median of the jth generated sample, for $j = 1, \ldots, m$. The results are reported in Table 1.

TABLE 1. — Simulated relative efficiencies of the sample median with respect to the sample mean for several sample sizes n and at different sampling distributions.

ARE_n	NOR	Lap	Uni	F0	t_2
$n = 10$	0.714	1.387	0.442	0.246	4.542
$n = 20$	0.684	1.489	0.390	0.157	6.166
$n = 50$	0.655	1.653	0.356	0.093	5.308
$n = 100$	0.641	1.756	0.346	0.065	6.540
$n = 1000$	0.637	1.931	0.338	0.020	7.812

One sees that the finite sample relative efficiencies converge to the expected values. Depending on the true distribution F, the sample median or the sample mean may be more efficient. At the distribution F0, the convergence rate of the sample median is slower than $n^{-1/2}$, as we infer from the convergence to zero

of ARE_n. On the other hand, for the Student distribution with 2 degrees of freedom, we see that the relative performance of the sample median becomes more and more superior with increasing sampling sizes.

To measure the dispersion of the estimator, we used the standard deviation in (4). Fréchet suggests in his paper to use also other measures of dispersion. Therefore we repeated the simulation exercise, now using the Interquartile Range (IQR) instead of SD in (4). Those results are then reported in Table 2. We see that, when both the sample mean and median are asymptotically normal, there is hardly any difference between Tables 1 and 2 (the convergence to the asymptotic value being a bit faster for the Laplace distribution when using IQR). But for the sampling distributions F0 and t_2 the numerical values of ARE_n are now clearly different.

TABLE 2. — As Table 1, but now with the sampling variability measured by the Interquartile Range.

ARE	NOR	Lap	Uni	F0	t_2
$n = 10$	0.717	1.682	0.418	0.135	2.110
$n = 20$	0.685	1.768	0.357	0.086	2.506
$n = 50$	0.654	1.874	0.352	0.050	3.016
$n = 100$	0.631	1.804	0.330	0.034	3.203
$n = 1000$	0.641	1.918	0.339	0.011	4.545

4. Conclusions

Focus in Fréchet's paper and in my discussion is on the statistical efficiency of the sample median. In the literature on robust statistics the median is often advocated as the optimal robust estimator. As such, the median has minimal sensitivity with respect to gross-errors (see also Croux, 1998), and it has the min-max bias property (*e.g.* Hampel *et al.*, 1986). In 1940, when Fréchet wrote his paper, the computational simplicity of the sample median was an important advantage. There are existing other estimators being robust, quite efficient over larger classes of distributions, and very easy to compute, such as trimmed means (see Croux and Haesbroeck, 2001, for a robustness comparison between several of these simple univariate location estimators). Other competitors for the median are M-estimators, which need to be computed iteratively, but have good efficiency and robustness properties. Moreover, M-estimators are also easy to construct in regression and multivariate models (*e.g.* Hampel *et al.*, 1986).

References

CROUX C. (1998), "Limit Behavior of the Empirical Influence Function of the Median", Statistics and Probability Letters, 37, 331-340.

CROUX C. and HAESBROECK G. (2001), "Maxbias Curves of Robust Scale Estimators Based on Subranges", Metrika, 53, 101-122.

HAMPEL F.R., RONCHETTI E.M., ROUSSEEUW P.J., and Stahel W.A. (1986), Robust Statistics : The Approach based on Influence Functions, New York : Wiley.

HODGES J.L., and LEHMANN E.L. (1956), "The Efficiency of Some Nonparametric Competitors of the t-test," *The Annals of Mathematical Statistics*, 27(2), 324-335.

JUREČKOVÁ J., and SEN P.K. (1996), Robust Statistical Procedures : Asymptotics and Interrelations, New York : Wiley.