JOURNAL DE LA SOCIÉTÉ FRANÇAISE DE STATISTIQUE

JEAN AVÉROUS

Commentaires à propos de l'article de Maurice Fréchet : « Sur une limitation très générale de la dispersion de la médiane »

Journal de la société française de statistique, tome 147, n° 2 (2006), p. 39-43

<http://www.numdam.org/item?id=JSFS_2006__147_2_39_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/

COMMENTAIRES À PROPOS DE L'ARTICLE DE MAURICE FRÉCHET : «SUR UNE LIMITATION TRÈS GÉNÉRALE DE LA DISPERSION DE LA MÉDIANE»

Jean AVÉROUS *

On pourrait sous-titrer l'article de Maurice Fréchet :

Appel pour la poursuite de la lutte contre l'emploi exclusif des méthodes imposées par les « gaussistes »

Bien sûr, si le lecteur s'attache seulement aux résultats annoncés par Maurice Fréchet et s'en tient au premier degré dans la lecture des commentaires qui les accompagnent, ce deuxième titre peut paraître provocateur et sans fondement. C'est pourtant M. Fréchet lui-même qui qualifie sa communication de « cri d'alarme » dont il donne les raisons, dans un style plus feutré que dans ses précédentes attaques contre les abus de l'emploi du coefficient de corrélation linéaire. Il présente les résultats sur la limitation de la dispersion de la médiane comme une preuve supplémentaire de la supériorité de celle-ci sur la moyenne dans de très nombreuses situations, justifiant sa réhabilitation par des considérations plus générales dans la ligne de ses combats antérieurs et dont on peut, aujourd'hui encore, mesurer la pertinence et le caractère visionnaire.

Il suffit de lire les manuels édités durant les quarante années qui ont suivi cet appel, et certains autres plus récents, pour constater que ce cri n'a pas été entendu, ou a été étouffé, par la communauté statistique fascinée par les facilités de l'algèbre linéaire et bilinéaire plus que par le terrorisme des « algébristes » ou des « théoriciens » que sous-entend M. Fréchet dans le dernier paragraphe précédant les « Confrontations avec l'expérience », (on remarquera l'emploi de l'article défini, singulier dans tous les sens du terme, pour qui ne donne qu'un seul exemple, déjà remarquable cependant, compte tenu des moyens de calcul de l'époque). Cette dénonciation de l'attitude de certains est aussi, malheureusement, annonciatrice des querelles stériles et dommageables pour la statistique, entre statistique mathématique et statistique appliquée, qui perdurent de nos jours à l'état endémique, principalement en France.

Il est intéressant d'analyser en détail l'argumentation de M. Fréchet pour établir la comparaison des avantages de la médiane avec ceux de la moyenne selon la mesure de dispersion utilisée. La méthode est à l'opposé des « preuves

* Laboratoire de statistique et probabilités, Bâtiment 1R1, Université Paul-Sabatier, 118, route de Narbonne, 31062 Toulouse, France, *adresse courriel* : averous@cict.fr

par intimidation », comme Roger Koenker (1997) qualifie les preuves données par Gauss pour imposer les moindres carrés.

M. Fréchet se borne à trois mesures de dispersion étudiées plus en détail dans un article publié aussi en 1940 :

- l'écart quadratique moyen de la médiane empirique M_n de n observations indépendantes distribuées comme X, $\sigma_{M_n} = (\mathbb{E}(M_n - \mathbb{E}(M_n))^2)^{\frac{1}{2}}$,
- l'écart moyen $\theta_{M_n} = \mathbb{E}|M_n - med(M_n)|$ dont l'avantage invoqué ici, pour mesurer la dispersion de M_n, est l'égalité entre les médianes théoriques de M_n et de X (appelées « valeurs probables » par M. Fréchet). Il va sans dire que l'écart quadratique moyen de la moyenne $\sigma_{V_n} = (\mathbb{E}(V_n - \mathbb{E}(V_n))^2)^{\frac{1}{2}}$ présente un avantage analogue pour mesurer la dispersion de la moyenne, et qu'il faut donc se placer en « terrain neutre » pour comparer « équitablement » dispersion de V_n et dispersion de M_n ; d'où l'introduction de la troisième mesure,
- l'écart probable, moitié de la longueur de l'intervalle interquartiles, que M. Fréchet affirme « n'être pas plus lié à la position de la valeur moyenne de X qu'à sa valeur probable ».

Viennent ensuite le calcul de bornes pour ces dispersions, que l'on ne peut améliorer si X est « seulement » assujetti à ce que sa valeur probable soit aussi sa valeur « dominante », condition qui, en pratique, pour un statisticien, restreint la loi de X aux lois unimodales symétriques... pour lesquelles moyenne (quand elle existe) et valeur probable sont confondues ; on est bien loin de la généralité souhaitée.

Enfin dans ce décor adroitement planté, la comparaison est effectuée en trois actes :

- Acte I : « On connaît la belle relation $9ter$ » ...(chanson gaussiste)
- Acte II : Mais,
  - pour des variables aléatoires particulières ce n'est pas si beau que cela et on veut des résultats valables pour une classe « très générale »,
  - de plus, « il n'est pas équitable d'employer, pour estimer cette précision, l'écart quadratique moyen, notion intimement liée à la moyenne » de même, cela va sans dire, il ne faut pas employer l'écart moyen intimement lié à la médiane.
- Acte III : Le verdict de l'écart probable : infinie supériorité de la médiane.

Prendre l'écart probable pour arbitre sous-entend qu'il n'est « intimement lié » ni à la moyenne ni à la médiane. Il semble bien pourtant que l'intimité (dont la mesure n'est pas précisée) est plus forte avec la médiane : l'intervalle interquartiles est l'intervalle de masse $\frac{1}{2}$ qui a même médiane que la distribution, ce qui permet d'obtenir, dans la classe des distributions de moyenne, médiane et écart probable fixés, des distributions d'écart quadratique moyen aussi grand que l'on veut. M. Fréchet a beau jeu alors de montrer que, côté médiane, $\sqrt{n}\frac{E_{M_n}}{E_X}$ est borné tandis que, côté moyenne, $\sqrt{n}\frac{E_{V_n}}{V_X}$ ne peut être borné

pour la (large...) classe de variables aléatoires considérée. M. Fréchet conclut donc, dans un style très Pascalien, que l'avantage de la médiane fourni par la formule 11*bis* comparée à 11*ter* est infiniment plus considérable que celui, qualifié modestement de sensible, que la formule 9*ter* concède à la moyenne par comparaison avec 9*bis*. Cette infinie supériorité, si l'on veut bien faire l'impasse sur la définition d'une mesure de supériorité relative à une mesure de dispersion, est donc obtenue avec un arbitre un peu partial, et un titre plus convenable serait : Sur la limitation d'*une* dispersion de la médiane.

Mais l'essentiel, dans cette communication, outre son caractère précurseur pour la notion de robustesse, est l'énoncé de quelques principes qu'il ne faut pas se lasser de répéter (Michel Armatte (2002) a relevé cette phrase de M. Fréchet dans un article de 1935 : « il y a des morts qu'il faut tuer plusieurs fois ») : avoir le souci du sens des méthodes utilisées pour traiter un problème donné et travailler à expliciter le sens des résultats mathématiquement démontrés pour mieux identifier les problèmes qu'ils peuvent résoudre. C'est ce que fait M. Fréchet dans l'introduction et dans ses réponses à M. Roy.

Pour terminer ces commentaires, notons que cette préoccupation du sens est à la base de travaux qui se sont développés quarante ans après cette communication. Le sens de la médiane comme paramètre de centrage optimum pour l'écart absolu était connu depuis longtemps et avait été étendu au cas multidimensionnel dans le cadre de la recherche opérationnelle (appelée alors application industrielle) par Weber (1909) ; redécouvert dans le cadre statistique par Gini et Galvani (1929) il était semble-t-il oublié au moment où M. Fréchet présentait cette communication et il fallut attendre encore huit ans avant sa redécouverte par Haldane (1948) et presque quarante de plus pour que soit étudiée en détail, par Kempermann (1987), la médiane dans un espace de Banach. La généralisation des intervalles interquantiles, qui peuvent être considérés en un certain sens comme une alternative aux « nombres peu déterminés » auxquels M. Fréchet fait allusion dans la réponse à M. Roy à propos du choix du paramètre de centrage lorsqu'on dispose de peu de données, viendra plus tard via les quantiles multivariés (Breckling et Chambers (1988), Chaudhuri (1996)) ou directement par extension des propriétés de Kempermann à des boules optimales (Avérous et Meste (1997)). Les idées lancées par Tukey en 1975 relèvent de cette même préoccupation et sont à l'origine de la notion de profondeur qui est depuis une quinzaine d'années un domaine de recherche très actif (*cf.* Liu *et al.* (1999), Mosler (2002) ainsi que Serfling (2002) pour de récentes bibliographies sur ce domaine où la notion de partie centrale optimale, généralisation des intervalles interquantiles, est la notion de base). Ce n'est que très récemment que les étroits liens entre quantiles multivariés, profondeur et dispersion ont été abordés (Avérous (2002)).

Plus généralement, on reconnait actuellement que le sens des valeurs typiques de centrage, dispersion, dissymétrie (skewness), aplatissement (kurtosis) est précisé par les relations d'ordre pour lesquelles ces valeurs sont monotones, relations d'ordre qui portent le sens des concepts considérés. Les articles

fondateurs de cette approche ont été écrits par Bickel et Lehmann (1975-76-79), suivis par Oja (1983) puis Dabrowska (1985).

Le « cri d'espoir », comme M. Fréchet appelle aussi sa communication, n'a donc pas été vain ; on trouvera dans l'article de Koenker (1997) la parfaite justification de la recommandation de M. Fréchet pour la facilité de calcul de la médiane et de ses extensions multivariées. De plus, la multitude d'articles et de communications sur le thème « $\mathbb{L}_1$ data analysis » ou sur la comparaison des méthodes $\mathbb{L}_1$ et $\mathbb{L}_2$ démontre le caractère visionaire du dernier paragraphe de la partie A qui justifierait à lui seul le rappel de cette communication à la communauté statistique.

Références

ARMATTE M. (2002) Maurice Fréchet statisticien, enquêteur et agitateur public. *Revue d'histoire des Mathématiques* **7**, 7-65.

AVÉROUS J. (2002) Quantile Functions and Spread for Multivariate Distributions. In *Statistical Data Analysis Based on the* $\mathbb{L}_1$*-Norm and related methods*, (Y. Dodge ed.) 3-14, Birkhauser.

AVÉROUS J. and MESTE M. (1997) Median balls : an extension of the interquatile intervals to multivariate distributions. *Journal of Multivariate Analysis* **63**, 222-241.

BICKEL P.J. and LEHMANN E.L. (1975) Descriptive statistics for nonparametric models (I. Introduction, II. Location). *Ann. Statist.* **3**, 1039-1069.

BICKEL P.J. and LEHMANN E.L. (1976) Descriptive statistics for nonparametric models (III. Dispersion). *Ann. Statist.* **4**, 1139-1158.

BICKEL P.J. and LEHMANN E.L. (1979) Descriptive statistics for nonparametric models (IV. Spread). In *Contributions to Statistics*, (J. Jurečková ed.) 33-40, Academia, Prague.

BRECKLING J. and CHAMBERS R. (1988) M-quantiles. *Biometrika* **75**, 751-771.

CHAUDHURI P. (1996) On a geometric notion of quantiles for multivariate data. *Journal of American Statistical Association* **91**, 862-872.

DABROWSKA D. (1985) Descriptive parameters of location, dispersion and stochastic dependence. Mathematische Operationforshung und Statistik, Series statistics **16**, 63-88.

GINI C. e GALVANI L. (1929) Di taluni estensioni dei concetti di media ai caratteri qualitativi. *Metron* **8**.

KEMPERMAN J.H.B. (1987) The median of a finite measure on a Banach space. In *Statistical Data Analysis Based on the* $\mathbb{L}_1$*-Norm and related methods*, (Y. Dodge ed.) 217-230, North-Holland.

KOENKER R. (1997) $\mathbb{L}_1$ computation, an interior monologue. $\mathbb{L}_1$-Statistical Procedures and Related Topics. IMS Lecture Notes - Monograph Series Volume **31**, 15-32.

KOLTCHINSKII V. (1997) M-estimation convexity and quantiles. *Annals of Statistics* **25**, 435-477.

LIU R., PARELIUS J.M. and SINGH K. (1999) Multivariate analysis with data depth : Descriptive statistics, graphics and inference. *Annals of Statistics* **27**, 783-858. With Discussion.

MOSLER (2002) Multivariate Dispersion and Depth. Lecture Notes in Statistics, Springer.

OJA H.(1983) Descriptive statistics for multivariate distributions. *Statistics and Probability Letters* **1**, 327-332.

SERFLING R. (2002) Quantile functions for multivariate analysis : approaches and applications. *Statistica Neerlandica* **56**, 214-232. (see also other papers by Serfling and/or Zuo, Y.).

TUKEY J.W. (1975) Mathematics and picturing data. *Proc. Intern. Congr. Math. Vancouver 1974* **2**, 523-531.

WEBER A. (1909) Uber den Standort der Industrien, Tubingen. English translation by Freidrich, C.J. (1929) Alfred Weber's Theory of location of Industries, University of Chicago Press.