JOURNAL DE LA SOCIÉTÉ FRANÇAISE DE STATISTIQUE

ANNE RUIZ-GAZEN

Introduction à la discussion de l'article de Maurice Fréchet

Journal de la société française de statistique, tome 147, n° 2 (2006), p. 17-21

<http://www.numdam.org/item?id=JSFS_2006__147_2_17_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/

INTRODUCTION À LA DISCUSSION DE L'ARTICLE DE MAURICE FRÉCHET (Journal de la Société Statistique de Paris, 1940)

Anne RUIZ-GAZEN *

Je tiens tout d'abord à remercier Henri Caussinus pour cette réédition de l'article de Maurice Fréchet « Sur une limitation très générale de la dispersion de la moyenne » et pour m'avoir associée à ce projet. L'objectif de ma contribution est essentiellement de présenter les neuf textes de discussion écrits pour l'occcasion par douze chercheurs statisticiens en écho à cet article de 1940. Plus de soixante années séparent l'article de référence de cette discussion. Mais il n'est qu'à voir la variété et la richesse des textes des intervenants pour comprendre le caractère précurseur de l'article en question et l'intérêt de sa réédition commentée. Les intervenants ont tour à tour qualifié l'article d'intéressant (O. Nuñez et D. Peña), de pertinent (J. Avérous), d'original (E. Ronchetti, P. Rousseeuw) voire de remarquable (C. Croux, S. Morgenthaler) et ils sont surtout frappés par son caractère moderne, très en avance sur son temps (P. Rousseeuw), précurseur (M. Armatte, O. Nuñez et D. Peña) et visionnaire (J. Avérous). Mais force est de constater que la plupart des auteurs ne connaissaient pas l'existence de ce papier. S. Morgenthaler précise même qu'il n'a trouvé aucune référence à cet article dans la littérature statistique usuelle, ce qui explique sans doute que certains des résultats obtenus par M. Fréchet aient été redécouverts plus de quarante ans après (*cf.* P. Rousseeuw). Ainsi, il semble que les idées novatrices développées par M. Fréchet aient été « étouffées » pendant plusieurs décénies avant d'être à nouveau considérées et développées de façon très active (*cf.* J. Avérous).

Ma présentation s'articule autour de quatre axes. Après avoir synthétisé les commentaires concernant le débat entre moyenne et médiane et les apports de l'article sur ce sujet, j'insisterai sur les aspects précurseurs du papier sur le plan de la statistique appliquée, de la statistique robuste mais aussi des simulations. Puis, suivant les commentaires de certains auteurs, je replacerai l'article de M. Fréchet dans un ensemble de travaux plus généraux et mentionnerai certains autres de ses travaux présentant un intérêt pour la statistique. Enfin, je terminerai en indiquant divers prolongements de l'article envisagés par les intervenants.

* Institut de Mathématiques, LSP, Université Toulouse 3 et GREMAQ, U.M.R. C.N.R.S. C5604, Université Toulouse I, 21, allée de Brienne, 31000 Toulouse.
ruiz@cict.fr

1. Médiane contre moyenne

L'objet de l'article de M. Fréchet est essentiellement la mise en évidence de l'avantage de l'usage de la médiane comparativement à l'usage de la moyenne en étudiant notamment l'efficacité asymptotique relative. Pour l'époque, l'article est original dans la mesure où la moyenne est l'estimateur habituellement utilisé, mais s'inscrit aussi dans la continuité d'un débat où l'utilisation de la moyenne (et de l'écart-type) est contestée. La contribution de M. Armatte décrit de façon détaillée l'historique de ce débat entre moyenne et médiane de ses origines à 1940. En particulier, il nous rappelle que la théorie des erreurs a mobilisé plusieurs grands mathématiciens du XVIII$^{\text{ème}}$ siècle (D. Bernoulli, Gauss, Lagrange, Laplace,...) et qu'à l'époque, le « triplet gagnant » pour le choix d'un centre, d'une loi des erreurs et d'un critère d'ajustement est moyenne, loi de Laplace-Gauss et méthode des moindres carrés. M. Armatte précise aussi que ce triplet domine le siècle suivant essentiellement parce qu'« il est consolidé par plusieurs théorèmes qui lient les trois ingrédients et fournissent les bases de la statistique mathématique (...) » mais aussi parce qu'il aboutit à des calculs algébriques simples. Ce choix est toutefois contesté au XIX$^{\text{ème}}$ siècle car les fondements de la moyenne ne sont pas « considérés comme corrects et suffisants », que « les expérimentateurs ont des raisons sérieuses de ne pas accepter une seule loi des erreurs pour toutes les situations expérimentales » et que le principe des moindres carrés donne « un poids déraisonnable aux grandes erreurs en valeur absolue ». L'article de M. Fréchet se situe justement « dans la continuité de cette tentative de fonder une théorie alternative des erreurs ».

Après avoir dit que l'usage exclusif de la moyenne et de l'écart-type était regrettable, M. Fréchet rappelle que, lorsque la dispersion est mesurée par l'écart-type, il existe des lois de probabilités pour lesquelles la médiane est moins dispersée que la moyenne mais il s'agit de lois particulières. Un des objectifs de son papier est de montrer qu'en changeant de mesure de dispersion, la médiane est moins dispersée que la moyenne pour une classe générale de lois de probabilités. Plus précisemment, M. Fréchet considère comme mesures de dispersion l'écart moyen (moyenne des écarts en valeur absolue) et la demi-longueur de l'intervalle interquartile et obtient des majorations des dispersions de la moyenne et de la médiane sous des hypothèses générales.

Les contributions de R. Koenker, d'O. Nuñez et D. Peña et de J. Avérous apportent aussi un éclairage historique. R. Koenker donne notamment un exemple d'analyse de données datant de la fin du XVIII$^{\text{ème}}$ siècle et du début du XIX$^{\text{ème}}$, révélateur de la controverse au sujet de la loi normale. O. Nuñez et D. Peña donnent un historique du débat entre moyenne et médiane en citant notamment les apports de Poisson, Cauchy et Fisher. Enfin, J. Avérous rappelle quels sont les chercheurs qui ont découvert ou redécouvert le caractère optimal de la médiane pour l'écart absolu.

2. Un article moderne

2.1. Des considérations de statistique appliquée

Plusieurs intervenants (dont J. Avérous, R. Koenker, E. Ronchetti et P. Rousseeuw) insistent sur le fait que, tout en étant résolumment tourné vers les mathématiques pures, M. Fréchet était aussi intéressé par la statistique appliquée au sens moderne du terme. Ainsi, R. Koenker trouve assez remarquable que M. Fréchet mette l'accent sur « l'importance d'une pratique statistique solide dans les sciences et la politique publique ». Quant au caractère visionnaire de la fin de la partie A, il a en particulier été souligné par J. Avérous, E. Ronchetti et P. Rousseeuw qui voit dans l'expression « méthodes empiriques de découverte » de M. Fréchet les prémisses de l'analyse exploratoire de données « à la Tukey ». Relevons aussi que M. Fréchet est très sensible à une simplification sur le plan numérique. Il précise même (page 73) que « cette simplification, indifférente aux théoriciens, est capitale pour les statisticiens ».

2.2. Un article précurseur de statistique robuste

La statistique robuste conduit généralement à la remise en cause des méthodes communément utilisées en statistique en raison de leur faible fiabilité en présence de données atypiques. Ainsi, concernant l'estimation d'un paramètre de centrage, la moyenne qui est optimale pour la loi normale en terme de variance est très sensible à la présence de données atypiques. Ce manque de robustesse est mis en évidence en considérant d'autres distributions que la loi normale (distributions à queues lourdes ou mélanges) et en utilisant des mesures modernes de sensibilité des méthodes statistiques considérées. Cette démarche est comparable à celle des tenants de la médiane dans les siècles passés (*cf.* M. Armatte), et à celle de M. Fréchet dans cet article, qui considèrent d'autres distributions que la loi normale (sans se focaliser exclusivement sur le problème de valeurs atypiques) pour montrer l'avantage de la médiane sur la moyenne. Suivant la contribution de J. Avérous, on peut toutefois remarquer que les lois de probabilités traditionnellement envisagées comme alternatives à la loi normale en statistique robuste (de même que celles envisagées par M. Fréchet dans son article) ne sont pas complètement générales et se situent souvent en quelque sorte « au voisinage » de la loi normale. Par ailleurs, en statistique robuste, la médiane n'est qu'un estimateur parmi les nombreux estimateurs envisagés en remplacement de la moyenne (voir notamment la célèbre étude de robustesse de Princeton citée par O. Nuñez et D. Peña).

C. Croux et P. Rousseeuw font remarquer que la comparaison asymptotique de la dispersion d'un estimateur robuste à la dispersion d'un estimateur non robuste (dans l'article, médiane contre moyenne) est une notion qui nous est familière puisqu'il s'agit de l'efficacité asymptotique relative (ARE) très utilisée en particulier en statistique robuste. Concernant la médiane, C. Croux, S. Morgenthaler et P. Rousseeuw précisent que la borne inférieure sur l'efficacité relative asymptotique de la médiane par rapport à la moyenne

obtenue par M. Fréchet a été redécouverte par Hodges et Lehmann quelques années après.

Une autre idée originale de M. Fréchet est de proposer des alternatives robustes à l'écart-type comme mesures de dispersion. Les alternatives qu'il envisage sont l'écart moyen à la médiane qu'il dénomme « écart-moyen » et la demi-longueur de l'intervalle interquartile qu'il dénomme « écart probable ». La comparaison de ces mesures de dispersion par le biais d'inégalités pour des variables aléatoires quelconques fait l'objet d'une autre publication de M. Fréchet [1] (1940). Cet article fournit les démonstrations des inégalités à la base des arguments développés dans l'article dont il est question ici.

Actuellement, comme le rappellent O. Nuñez et D. Peña, le MAD (médiane des écarts absolus à la médiane) est l'estimateur le plus couramment utilisé en statistique robuste. Mais l'intervalle interquartile que propose M. Fréchet est aussi reconnu en tant qu'estimateur robuste. C. Croux ainsi qu'O. Nuñez et D. Peña remarquent que, sous certaines conditions de régularité, utiliser d'autres mesures de dispersion que la variance pour calculer le ARE ne change pas sa valeur. Toutefois, P. Rousseeuw note que plusieurs auteurs contemporains ont redécouvert l'idée d'utiliser des mesures robustes de dispersion pour mesurer l'efficacité à taille d'échantillon finie.

Remarquons pour finir ce paragraphe que la contribution de P. Rousseeuw détaille tout particulièrement les liens entre l'article de M. Fréchet et la littérature de la statistique robuste. Certaines déductions de M. Fréchet (établissement de (11ter)) sont même apparentées à des déductions de type « point de rupture ».

2.3. Un article précurseur dans l'utilisation des simulations

E. Ronchetti note que M. Fréchet adopte une démarche moderne de chercheur en méthodologie statistique avec l'étude d'approximations asymptotiques validées par des simulations.

Cette méthode de « simulations » que propose M. Fréchet est astucieuse puisqu'elle peut être mise en œuvre sans l'aide des ordinateurs (inexistants à l'époque !). C. Croux ainsi qu'O. Nuñez et D. Peña ont repris le même type de simulations mais avec des moyens modernes et une grande variété de lois de probabilités.

Ce même exemple a aussi été analysé très en détail par M. Genton, Y. Ma et E. Parzen. Ces intervenants remarquent que les données traitées par M. Fréchet suivent une distribution discrète avec des ex-æquos et que la médiane empirique usuelle « ne traite pas correctement ce type de données ». En suivant pas à pas le principe de simulations proposé par M. Fréchet et en le généralisant à des tailles d'échantillons plus grandes, M. Genton, Y. Ma et E. Parzen montrent en particulier que la médiane calculée par M. Fréchet à partir des données qu'il propose ne suit pas une distribution gaussienne. Ainsi, à partir de cette expérience, M. Genton, Y. Ma et E. Parzen nous convainquent

1. Je remercie J. Avérous de me l'avoir indiquée.

qu'il faut utiliser une définition des quantiles adaptée à la présence d'ex-æquos. Une telle notion a été récemment proposée et étudiée par E. Parzen.

3. Au sujet d'autres travaux de Maurice Fréchet

M. Armatte, J. Avérous et R. Koenker précisent que cet article fait partie d'une « plus vaste campagne » de M. Fréchet, statisticien, dans laquelle on trouve aussi une critique de la mauvaise utilisation du coefficient de corrélation linéaire.

Comme il le dit lui même, M. Fréchet est avant tout un mathématicien spécialiste d'analyse. M. Armatte nous rappelle que ses travaux sur les espaces fonctionnels lui valent sa réputation internationale et qu'ils ont « une résonance directe dans le domaine de la statistique mathématique ».

E. Ronchetti mentionne ainsi la notion de différentielle au sens de Fréchet qui est couramment utilisée en statistique robuste puisqu'elle permet de travailler sur des espaces fonctionnels.

4. Prolongements

Les intervenants indiquent plusieurs prolongements de l'article de M. Fréchet dans la littérature moderne. Les extensions vers la statistique robuste sont les prolongements les plus évidents. Ils ont fait quasiment l'unanimité parmi les intervenants et nous avons déjà consacré un paragraphe à ce sujet. Mais d'autres prolongements ont été envisagés par certains auteurs.

Ainsi, J. Avérous rappelle les extensions multivariés de la médiane et de l'intervalle interquartile et cite aussi la notion de profondeur. Selon O. Nuñez et D. Peña, l'article de M. Fréchet contient déjà « en germe » l'approche minimax. S. Morgenthaler cite l'utilisation d'ensembles de lois non paramétriques et de mélanges.

Des prolongements récents ou restant à développer sont aussi précisés. M. Genton, Y. Ma et E. Parzen montrent que la distribution exacte de la médiane ainsi que celle du demi-intervalle interquartile (cas particuliers de L-estimateurs) peuvent être calculées pour des échantillons dépendants et des distributions à queues lourdes. Enfin, R. Koenker propose de définir une médiane et plus généralement des quantiles de Fréchet dans le domaine de l'analyse de données fonctionnelles.

Référence complémentaire

FRÉCHET M. (1940), « Comparaison des diverses mesures de la dispersion », *Revue de l'Institut International de Statistique*, 8 (1), 1-12.