E. J. Gumbel

La précision de la Moyenne Arithmétique et de la Médiane

Aktuárské vědy, Vol. 6 (1936), No. 4, 145–154

Persistent URL: http://dml.cz/dmlcz/144668

La précision de la Moyenne Arithmétique et de la Médiane.

E. J. Gumbel, Université de Lyon.

1) La distribution de la médiane.
2) Calcul des précisions pour quelques distributions initiales.
3) Conclusions.

Pour caractériser une répartition observée à l'aide d'une seule valeur, on choisit une des moyennes. Parmi elles, la moyenne arithmétique et la médiane jouent le premier rôle. Il s'agit de savoir, laquelle il faut choisir. Ce problème classique est résolu à l'aide de plusieurs critères — qui se contredisent. Nous n'allons pas reprendre tous ces arguments. Nous nous limitons plutôt à apporter quelques détails sur la question spéciale des précisions comparées de la moyenne arithmétique et de la médiane.

1. La distribution de la médiane.

Les moyennes, calculées d'après les observations, sont des variables statistiques, ayant par conséquent des dispersions dont l'inverse mesure leurs précisions. Nombreux sont les praticiens qui supposent que la moyenne arithmétique est toujours plus précise que la médiane. Cette opinion dérive du fait que cette proposition est juste pour la distribution de Gauss qui joue un rôle fondamental en statistique. Mais ce théorème étudié par M. Haag[1]) ne vaut pas pour une distribution quelconque.

Dans un récent travail, M. Fréchet[2]) a établi la distribution de la médiane pour $N = 3$ observations. Il en a tiré trois théorèmes intéressants: Pour une distribution initiale symétrique, la distribution de la médiane est symétrique (I). La médiane de la distribution de la médiane est égale à la médiane de la distribution initiale (II). Pour la distribu-

[1]) J. Haag: Sur la combinaison des résultats d'observations. Comptes Rendus 179, No. 24, p. 1388. Paris 1924.

[2]) M. Fréchet: Sur les précisions comparées de la moyenne et de la médiane. Aktuárské Vědy T. V., no. 1, Prague 1935.

tion initiale exponentielle symétrique, la médiane est plus précise que la moyenne arithmétique (III).

Le but de ces lignes est d'étendre ces théorèmes à un nombre impair quelconque d'observations. Soit $w(x)$ la distribution dite initiale ayant l'espérance mathématique $\bar{x}$ et l'écart type σ. Soit $W(x)$ la fonction des probabilités totales et c la médiane définie, d'après la méthode habituelle, par

$$W(c) = \tfrac{1}{2}. \tag{1}$$

Soit enfin $N = 2m - 1$ le nombre d'observations. Puisque une seule observation ne se prête pas au calcul d'une moyenne, il faut supposer

$$m \geqq 2.$$

Alors la médiane est simultanément la $m - 1^{\text{ième}}$ valeur d'en haut et d'en bas. Sa distribution $\mathfrak{w}_{m-1}(x)$ est donc un cas spécial de la distribution de la $m^{\text{ième}}$ valeur,[3]) ce qui mène à

$$\mathfrak{w}_{m-1}(x) = \frac{(2m-1)!}{(m-1)!^2} W^{m-1}(x)\,(1 - W(x))^{m-1}\, w(x). \tag{2}$$

La variable de cette distribution est la médiane. Pourtant nous l'écrivons x. Car à chaque valeur de la variable initiale appartient une certaine densité de probabilité pour qu'elle soit la médiane. Pour une distribution initiale symétrique, pour laquelle

$$w(-x) = w(x); \quad W(x) = 1 - W(x)$$

on obtient

$$\mathfrak{w}_{m-1}(-x) = \mathfrak{w}_{m-1}(x)$$

ce qui prouve (I) pour N arbitraire.

La médiane, variable statistique soumise à la distribution (2), possède une valeur dominante $\tilde{c}$ qu'on obtient d'après

$$\frac{\mathrm{d} \lg \mathfrak{w}_{m-1}(x)}{\mathrm{d}x} = 0.$$

La dominante de la médiane est donc la solution de

$$\frac{m-1}{W} w - \frac{m-1}{1-W} w + \frac{w'}{w} = 0 \tag{3}$$

ou

$$(m-1)\frac{\mathrm{d} \lg W(1-W)}{\mathrm{d}x} = -\frac{\mathrm{d} \lg w}{\mathrm{d}x}. \tag{3'}$$

Si la distribution initiale est symétrique, on obtient la solution habituelle (1). Dans le cas contraire, ces équations fournissent une solu-

[3]) E. J. Gumbel: Les valeurs extrèmes des distributions statistiques. Annales de l'Institut Henri Poincaré, T. IV, fasc. 2, Paris 1935.

tion

$$W(\tilde{c}) = \tfrac{1}{2} + f(N) \tag{1'}$$

dont on tire une correction pour le calcul de la médiane valable pour de petits nombres d'observations. Si la distribution initiale est telle qu'on peut négliger le terme $w'(\tilde{c})$ dans (3), de même, si le nombre d'observations augmente, la solution de (3) tend vers la valeur habituelle (1). La fonction $f(N)$ dependant de la distribution unitiale diminue pour un nombre croissant d'observations.

La probabilité $\mathfrak{W}_{m-1}(x)$ pour que la médiane soit inférieure à x, est d'après (2)

$$\mathfrak{W}_{m-1}(x) = \frac{(2m-1)!}{(m-1)!^2} \int\limits_0^{W(x)} z^{m-1}(1-z)^{m-1}\, dz \tag{4}$$

grandeur qu'on peut évaluer à l'aide des tables de la fonction Gamma incomplète. Pour la médiane c de la distribution initiale on obtient, en posant $z = 1 - u$

$$\int\limits_0^{\frac{1}{2}} z^{m-1}\,(1-z)^{m-1}\, dz = \int\limits_{\frac{1}{2}}^{1} u^{m-1}\,(1-u)^{m-1}\, du.$$

D'après la condition de l'aire, cela mène à

$$\mathfrak{W}_{m-1}(c) = \tfrac{1}{2}$$

ce qui prouve (II) pour N arbitraire.

Pour calculer l'espérance mathématique et la précision de la médiane, il faudrait évaluer

$$\int\limits_{-\infty}^{+\infty} \mathfrak{w}_{m-1}(x)\, x\, dx \quad \text{et} \quad \int\limits_{-\infty}^{+\infty} \mathfrak{w}_{m-1}(x)\, x^2\, dx$$

ce qui n'est pas possible sans introduire la distribution initiale que l'on traite. Au lieu de ce procédé spécial, calculons la forme finale $\mathfrak{w}(x)$ vers laquelle tend la distribution (2) pour des valeurs croissantes d'observations. Différents auteurs, Fogelson[4]), Kolmogoroff[5]), Smirnoff[6]) et Eyraud,[7]) ont prouvé par différentes méthodes que cette forme finale est Gaussienne.

---

[4]) S. Fogelson: Sur la détermination de la médiane. Revue trimestrielle de Statistique, T. VII, fasc. 2, Varsovie 1930.

[5]) A. N. Kolmogoroff: La méthode de la médiane dans la théorie des erreurs. Recueil mathématique, T. 38, p. 47, Moscou 1931.

[6]) N. Smirnoff: Über das allgemeine Glied der Variationsreihe. Metron T. XII, no. 2, Rome 1935.

[7]) H. Eyraud: Sur la valeur la plus précise d'une distribution. C. R. de l'Ac. des Sc. T. 199, p. 817, Paris 1934.

Une preuve très simple est la suivante[8]): la densité de probabilité $\mathfrak{w}_{m-1}(c)$ de la dominante de la distribution (2) de la médiane

$$\mathfrak{w}_{m-1}(c) = \frac{(2m-1)!}{(m-1)!^2}\, 2^{-2m-2}\, w(c)$$

tend, d'après la formule de Stirling pour des valeurs croissantes du nombre d'observations, vers

$$\mathfrak{w}(c) = \frac{2}{\sqrt{\pi}}\,\frac{m}{m-1}\sqrt{m-\tfrac{1}{2}}\,w(c),$$

$$\mathfrak{w}(c) = 2\sqrt{\frac{m}{\pi}}\,w(c). \tag{5}$$

Si le nombre d'observations est suffisamment grand, on peut développer $W(x)$ autour de c en se limitant au premier membre. Car la probabilité de la dominante augmente comme $\sqrt{m}$. Posons donc dans (2)

$$W(x) = \tfrac{1}{2}(1 + 2\,(x-c)\,w(c)),$$
$$1 - W(x) = \tfrac{1}{2}\,(1 - 2\,(x-c)\,w(c))$$

et

$$w(x) = W'(x) = w(c).$$

En traitant d'abord les facteurs qui existent dans la densité de probabilité de la dominante, on aura

$$\mathfrak{w}(x) = \mathfrak{w}(c)\,[1 + 2\,(x-c)\,w(c)]^{m-1}\,[1 - 2\,(x-c)\,w(c)]^{m-1}.$$

En introduisant la valeur $\mathfrak{w}(c)$, d'après (5), on obtient

$$\lg \mathfrak{w}(x) = \lg 2\sqrt{\frac{m}{\pi}}\,w(c) + (m-1)\lg\,(1 - 4\,(x-c)^2\,w^2(c))$$

ou en première approximation

$$\mathfrak{w}(x) = 2\sqrt{\frac{m}{\pi}}\,w(c)\,e^{-4(m-1)\,(x-c)^2 w^2(c)}$$

Pour de grandes valeurs de m, la distribution (2) tend vers la distribution de Gauss

$$\mathfrak{w}(x) = 2\sqrt{\frac{m}{\pi}}\,w(c)\,e^{-4m(x-c)^2 w^2(c)}. \tag{6}$$

L'espérance mathématique, la médiane et la dominante de la distribution de la médiane coïncident dans ces conditions avec la médiane de la distribution initiale définie par (1). On tire de (6) l'écart type de la

[8]) E. J. Gumbel: La distribution finale des valeurs voisines de la médiane. C. R. de l'Ac. des Sc. T. 199, p. 1173, Paris 1934.

médiane σ_c étudié par M. R. A. Fisher[9])

$$\sigma_c = \frac{1}{2\sqrt{N}\, w(c)} \tag{7}$$

expression valable pour des grands nombres d'observations.

Puisque l'écart type de la moyenne arithmétique $\bar{x}$ est

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{N}} \tag{8}$$

on conclut: Pour une distribution telle que la densité de probabilité de la médiane $w(c)$ est inférieure (égale, supérieure) à la valeur inverse du double de l'écart type σ, la précision de la moyenne arithmétique est supérieure (égale, inférieure) à la précision de la médiane. Ainsi le théorème (III) est étendu à une distribution quelconque et à un grand nombre d'observations. Ce critère permet de juger la précision des deux moyennes pour toute distribution qui permet le calcul de ces deux valeurs.

En formant la relation $R(c, \bar{x})$ de la précision $1 : \sigma_c$ de la médiane par rapport à celle de la moyenne arithmétique $1 : \sigma_{\bar{x}}$

$$R(c, \bar{x}) = 2\sigma\, w(c) \tag{9}$$

ou l'inverse

$$R(\bar{x}, c) = \frac{1}{2\sigma\, w(c)} \tag{9'}$$

on peut calculer le pourcentage de précision qu'on obtient pour l'une des moyennes par rapport à l'autre.[10]) Il est évident que ce critère ne s'applique qu'aux distributions qui permettent une expression simple pour l'écart type et pour la densité de probabilité de la médiane.

2. Calcul des précisions pour quelques distributions initiales.

Appliquons ce critère à quelques distributions. Pour l'équipartition

$$w(x) = \frac{1}{k}; \qquad 0 \leqq x \leqq k$$

l'écart type est

$$\sigma = \frac{k}{2\sqrt{3}}$$

donc

$$w(c) < \frac{1}{2\sigma}.$$

---

[9]) R. A. Fischer: Theory of statistical estimation. Cambridge Phil. Soc. T. 22, p. 700, 1925.

[10]) Voire: G. Darmois: Méthodes d'estimation. Actualités scientifiques et industrielles, No. 356. Hermann & Cie., Paris 1936.

Puisque

$$\frac{1}{k} < \frac{\sqrt{3}}{k}.$$

D'après

$$R(c, \bar{x}) = \frac{1}{\sqrt{3}}, \tag{9a}$$

la médiane n'atteint que 58% de la précision de la moyenne arithmétique.

Pour la distribution de Gauss, la densité de probabilité de la médiane est

$$w(c) = \frac{1}{\sigma\sqrt{2\pi}} < \frac{1}{2\sigma}.$$

La moyenne arithmétique est plus précise que la médiane, règle bien connue. D'après

$$R(c, \bar{x}) = \sqrt{\frac{2}{\pi}} \tag{9b}$$

la précision de la médiane n'atteint que 80% de la précision de la moyenne arithmétique.

Traitons comme exemple d'une distribution asymétrique celle de la plus grande valeur[11])

$$w(x) = \alpha e^{-y-e^{-y}}$$

où

$$y = \alpha(x - u).$$

La constante u est la dominante. L'autre est reliée à l'écart type par

$$\sigma = \frac{\pi}{\sqrt{6}\,\alpha}.$$

La probabilité d'une valeur inférieure à x est

$$W(x) = e^{-e^{-y}}.$$

Puisque la distribution est asymétrique l'équation (3) peut servir pour calculer la dominante de la médiane pour un nombre restreint d'observations, faites sur la plus grande valeur. D'après

$$\frac{w}{W} = e^{-y}; \qquad -\frac{w'}{w} = 1 - \varepsilon^{-y}$$

on peut écrire (3) dans la forme

$$(m-1)\, e^{-y}\, (1 - 2W) = (1 - W)\,(1 - e^{-y}). \tag{3c}$$

Posons

---

[11]) E. J. Gumbel: La plus grande valeur. Aktuárské Vědy, T. V., nos. 2, 3, 4. Prague 1935-36.

$$W = \tfrac{1}{2}e^{-\varepsilon},$$
$$e^{-y} = \lg 2 + \varepsilon,$$

où ε, fonction de m, sera la correction nécessaire pour de petits nombres d'observations. En développant, d'après des puissances croissantes de ε, et en négligeant ε^2 par rapport à 1, on obtient

$$2\varepsilon\,(m - 1)\,(\lg 2 + \varepsilon) = (1 + \varepsilon)\,(1 - \varepsilon - \lg 2),$$

ou

$$2\varepsilon\,(m - 1) \lg 2 + 2\varepsilon^2\,(m - 1) = 1 - \lg 2 - \varepsilon \lg 2.$$

En négligeant même $m\varepsilon^2$ par rapport à 1, la première approximation devient

$$\varepsilon = \frac{1 - \lg 2}{(2m - 1) \lg 2}.$$

La dominante de la médiane calculée à l'aide de

$$W(\tilde{c}) = \tfrac{1}{2}e^{-\frac{0{,}44270}{N}} \tag{1'c}$$

sera un peu inférieure à la valeur habituelle.

Pour un nombre suffisamment grand d'observations, on obtient la densité de probabilité de la médiane

$$w(c) = \frac{\alpha \lg 2}{2}.$$

On aura donc

$$w(c) < \frac{1}{2\sigma}$$

car

$$\lg 2 = 0{,}69315 < \frac{\sqrt{6}}{\pi} = 0{,}77970.$$

D'après

$$R(c, \bar{x}) = \frac{\pi \lg 2}{\sqrt{6}}. \tag{9c}$$

La précision de la médiane est 90% de la précision de la moyenne arithmétique. On vérifie aisément que ce resultat vaut aussi pour la distribution de la première valeur.[11])

Pour la distribution asymétrique exponentielle réduite

$$w(x) = e^{-x};\ \ W(x) = 1 - e^{-x}$$

qui régit, par exemple, les distances entre les émissions radioactives l'espérance mathématique, égale à l'écart type, est 1. Pour de petits nombres d'observations, on obtient la dominante de la médiane d'après (3) par

$$(m-1)\left(\frac{1}{W}-\frac{1}{1-W}\right)=\frac{1}{1-W} \tag{3d}$$

ou

$$\frac{m-1}{W}=\frac{m}{1-W}$$

ce qui mène à

$$W(\tilde{c})=\frac{m-1}{2m-1}. \tag{1'd}$$

La dominante de la médiane est donc, pour de petits nombres d'observations, un peu inférieure à la valeur (1) qu'on choisit d'ordinaire. Mais pour des valeurs croissantes du nombre d'observations, cette différence disparait. La densité de probabilité de la dominante est alors

$$w(\tilde{c})=\tfrac{1}{2}.$$

Mais de même

$$\frac{1}{2\sigma}=\frac{1}{2}.$$

D'après

$$R(c,\bar{x})=1 \tag{9d}$$

la précision de la moyenne arithmétique est égale à la précision de la médiane.

Traitons enfin la distribution exponentielle symétrique. Cette première loi de Laplace étudiée par M. Fréchet[2]) se compose de deux branches

$$w(x)=\begin{cases}\dfrac{e^{x}}{2} & \text{pour } x<0\\[2mm] \dfrac{e^{-x}}{2} & \text{pour } x>0\end{cases}$$

La probabilité est

$$W(x)=\begin{cases}\dfrac{e^{x}}{2} & \text{pour } x<0\\[2mm] 1-\dfrac{e^{-x}}{2} & \text{pour } x>0.\end{cases}$$

On obtient donc, pour de petits nombres d'observations, deux valeurs dominantes de la médiane. En effet, l'équation (3') mène,

$$\text{pour } x<0 \text{ à } (m-1)\left(1-\frac{W}{1-W}\right)=-1;$$

et (3'c)

$$\text{pour } x>0 \text{ à } (m-1)\left(\frac{1-W}{W}-1\right)=1$$

dont on tire

$$(m-1)(1-2W)=W-1;\ (m-1)(1-2W)=W$$

ou

$$m=2Wm-W;\ m-1=2mW-W$$

c'est-à-dire

$$W(\tilde{c})=\tfrac{1}{2}\left(1+\frac{1}{N}\right);\quad W(\tilde{c})=\tfrac{1}{2}\left(1-\frac{1}{N}\right). \tag{1'c}$$

La première solution est un peu supérieure, l'autre un peu inférieure à la valeur habituelle zéro. En introduisant la fonction des probabilités totales, on obtient, en désignant les deux solutions par $'\tilde{c}$ et $\tilde{c}'$ en première approximation

$$'\tilde{c}=\frac{1}{N};\qquad \tilde{c}'=-\frac{1}{N}.$$

Les deux valeurs de la médiane ont la même densité de probabilité.

Pour de grands nombres d'observations, on obtient

$$w(\tilde{c})=\tfrac{1}{2}$$

puisque l'écart type est

$$\sigma=\sqrt{2}.$$

On aura

$$w(\tilde{c})>\frac{1}{2\sigma}.$$

D'après

$$R(\bar{x},c)=\frac{1}{\sqrt{2}} \tag{9'c}$$

la moyenne arithmétique n'atteint que 70% de la précision de la médiane.[12])

Enfin, traitons un exemple où la relation des précisions dépend des constantes. Pour la distribution de Galton,[13]) où l'on prend la médiane comme unité des mesures,

$$w(x)=\frac{1}{s\sqrt{2\pi}\,x}\,e^{-\frac{1}{2s^2}\lg^2 x}$$

l'écart type est

$$\sigma=e^{\frac{1}{2}s^2}\sqrt{e^{s^2}-1}$$

tandis que

$$w(\tilde{c})=\frac{1}{s\sqrt{2\pi}}.$$

[12]) Voire: E. B. Wilson: First and second laws of error. Quaterly Publ. Amer. Stat. Association, September 1923. — E. B. Wilson and M. M. Hilferty: Note on C. S. Peirce's Experimental Discussion of the laws of errors. Proc. Nat. Ac. of Sc., T. 15, no. 2, pp. 120—125, Washington 1929.

[11]) E. J. Gumbel, Über eine Verteilungsgesetz, Zeitschrift für Physik, Bd. 37 No 6. Berlin 1926.

La précision de la moyenne est donc supérieure (égale, inférieure) à celle de la médiane suivant que

$$\frac{e^{s^2}(e^{s^2}-1)}{s^2} \gtreqless \frac{\pi}{2}.$$

On obtient l'égalité pour

$$s^2 = 0{,}29859; \quad e^{s^2} = 1{,}34795.$$

On peut interprèter ce resultat de deux manières suivant les significations qu'on attribue à e^{s^2}. Car $e^{s^2} = 1 + v^2$ où $v = \sigma/\bar{x}$ est le coéfficient de variation, dont on tire la condition

$$\sigma \gtreqless 0{,}5899\bar{x}.$$

D'autre part

$$e^{\frac{1}{2}s^2} = \bar{x}$$

dont on tire la condition

$$\bar{x} \gtreqless 1{,}161.$$

Four de petites valeurs de l'écart type par rapport à l'espérance mathématique ou pour de petites valeurs de l'espérance mathématique par rapport à la médiane, ce sera la médiane, pour de grandes valeurs ce sera l'espérance mathématique qui est plus précise. Dans le cas spécial

$$\sigma = 0{,}5899\bar{x}$$

les précisions des deux moyennes coïncident, tandis que leurs espérances mathématiques diffèrent, cas opposé à la distribution de Gauss, où les deux moyennes coïncident tandis que leurs précisions diffèrent.

3. Conclusions.

La médiane de la distribution de la médiane est égale à la médiane de la distribution initiale. La dominante de la distribution de la médiane, solution de (3), tend vers la même valeur. Enfin, la distribution de la médiane tend vers une distribution de Gauss.

Les distributions considérées prouvent que les relations entre les deux moyennes et leurs précisions sont multiformes. Il se peut que les espérances mathématiques des deux moyennes coïncident et leurs précisions diffèrent. Pour une autre distribution, les deux moyennes diffèrent et leurs précisions coincident. Enfin — et ce sera le cas général — les deux moyennes et leurs précisions diffèrent.

Nous avons vérifié que la moyenne arithmétique est plus précise que la médiane pour l'équipartition, pour la distribution de Gauss et pour la distribution de la plus grande valeur; qu'elle est moins précise pour la distribution exponentielle symétrique, tandis que la relation des précisions dépend de la grandeur de l'écart type pour la distribution de Galton. Dans un cas spécial de cette distribution et pour la distribution exponentielle asymétrique les deux précisions sont égales.