Distance entre la moyenne et la médiane d'une variable aléatoire :

une application d'une version latéralisée de l'inégalité de Bienaymé-Tchebycheff[1]

À propos de médiane et de moyenne

La médiane et la moyenne sont des outils statistiques couramment utilisés par exemple, pour les comparaisons de salaires. Ces deux outils, qui ont chacun leurs avantages et leurs inconvénients, devraient être employés de manière complémentaire selon le cadre de l'étude.

La médiane est généralement plus représentative pour l'analyse des petits groupes, car la présence de quelques cas extrêmes peut tirer fortement la moyenne vers le haut ou vers le bas et donner ainsi une image qui n'est pas représentative de la grande majorité des données. Prenons un exemple simple pour illustrer le phénomène : un groupe de 10 salaires, dont 8 de 2000 euros et 2 de 5000 euros. La moyenne est 2600 et la médiane de 2000. La moyenne est tirée vers le haut, dans une zone où il n'y a justement pas de salaires. La médiane est par contre représentative de la majorité des salaires analysés.

Dans les grands groupes, de plusieurs dizaines ou centaines de données analysées, la médiane et la moyenne tendent à se confondre. Les dispersions vers le haut et vers le bas tendent à se compenser et d'autre part les cas extrêmes isolés ont peu d'influence sur la moyenne.

Dans les grands groupes, c'est toutefois la médiane qui peut donner une image peu représentative de la réalité. C'est le cas lorsque les groupes analysés sont formés de sous-groupes eux-mêmes assez homogènes, mais bien distincts les uns des autres. Pour illustrer ce phénomène, prenons à nouveau un cas simple avec deux groupes de 100 salaires. Dans chacun des deux groupes, il y a un sous-groupe de bas salaires de 2000 euros et un sous-groupe de hauts salaires de 4000 euros. La proportion des deux sous-groupes est différente entre les deux groupes. Dans le premier, il y a 60% de bas salaires et 40% de hauts salaires. Dans le second groupe, la proportion est inversée. Dans cet exemple, la moyenne est de 2800 pour le premier groupe et de 3200 pour le second, soit un écart de 400, correspondant à 13,3% de la moyenne globale des deux groupes. Les médianes donnent par contre une image très différente, avec une valeur de 2000 pour le premier groupe et de 4000 pour le second. L'écart est alors de 2000, soit 66% de la moyenne globale. Dans cet exemple, l'écart calculé sur la base des médianes donne manifestement une image fortement biaisée de la réalité, alors que les moyennes permettent une bonne représentation de cette réalité.

Ceci est évidemment simpliste, mais correspond toutefois à certaines analyses statistiques basées sur les médianes, publiées par des organismes officiels.

Proposons-nous maintenant d'examiner sur un plan théorique, comment une information sur la dispersion de la distribution autour de sa moyenne donnée par le paramètre écart-type, σ, permet de situer une valeur médiane. Nous nous plaçons dans le cadre des variables aléatoires définies sur un espace probabilisé. Une définition de la médiane souvent utilisée est donnée par la formule $m_e = \min\left\{x \ /F(x) \geqslant \frac{1}{2}\right\}$ où F est la fonction de répartition de la variable aléatoire X. Elle peut s'appliquer indifféremment aux cas des variables discrètes ou autres. Cette définition de la médiane n'est pas toujours celle qui est implantée dans les logiciels calculant des paramètres statistiques[2]. Plus précisément en ce qui nous concerne, nous allons prouver que si $m = \mathbb{E}(X)$,

$$\mathbf{P}(X \leqslant m - \sigma) \leqslant \frac{1}{2} \text{ et } \mathbf{P}(X \leqslant m + \sigma) \geqslant \frac{1}{2}$$

ce qui permet de conclure que la médiane est dans l'intervalle $[m - \sigma, m + \sigma]$ au sens précédent lorsque $\frac{1}{2}$ admet au plus un antécédent par la fonction F (c'est le cas si X discrète finie, prend n valeurs $\{x_i / i \in [\![1, n]\!]\}$, rangées en ordre croissant, telles que pour tout $j \in [\![1, n]\!]$ $\sum_{i=1}^{j} \mathbf{P}(X = x_i) \neq \frac{1}{2}$).

Justification des inégalités

Soit X une variable aléatoire ayant des moments au moins d'ordre deux.
Rappelons l'inégalité de Bienaymé-Tchebycheff

$$\forall \varepsilon > 0 \quad \mathbf{P}(|X - \mathbb{E}(X)| \geqslant \varepsilon) \leqslant \frac{\sigma^2}{\varepsilon^2}$$

dont une application importante est d'établir la loi faible des grands nombres, i.e. la convergence (*en probabilité*) vers m de $Z_n = \frac{1}{n} \sum_{i=1}^{n} X_i$ où (X_n) est une suite de variables aléatoires indépendantes deux à deux, admettant une espérance m et une variance σ^2. Cet emploi ne sera pas utilisé dans notre étude. L'inégalité de Bienaymé-Tchebycheff peut s'établir par déduction[3] d'une inégalité plus simple à formuler connue sous le nom d'inégalité de Markov qui s'énonce ainsi : si X est une variable aléatoire discrète ou à densité dont l'univers

[1]Cette courte présentattion m'a été inspirée par la consultation du site http://www.se16.info/hgb/cheb.htm et de pages web d'HENRY BOTTOMLEY où vous trouverez des développements complémentaires. Je remercie JEAN LEJEUNE pour sa relecture et ses judicieuses remarques.

[2]On pourra consulter à ce sujet l'article bien documenté de JEAN LEJEUNE sur les différentes définitions des quantiles couramment usitées, publié dans le *Miroir des Maths n°1* et téléchargeable sur notre site web.

[3]Appliquer l'inégalité de Markov à la V.A.R. Y telle que$Y = |X - \mathbb{E}(X)|^2$ avec $\lambda = \varepsilon^2$ et remarquer que $\mathbb{E}(Y) = V(X) = \sigma^2$.

image $X(\Omega)$ est inclus dans $\mathbb{R}_+$ et qui possède une espérance, alors pour tout $\lambda > 0$, $\quad \mathbf{P}(X \geqslant \lambda) \leqslant \dfrac{\mathbb{E}(X)}{\lambda}$.

Nous allons maintenant déduire de cette inégalité, la proposition suivante :

Soit X une variable aléatoire ayant des moments au moins d'ordre deux, et $t > 0$. Posons $Z = X - \mathbb{E}(X)$ et $\sigma(X) = \sigma$. Nous avons $\mathbb{E}(Z) = 0$ et $\mathbb{E}(Z^2) = \sigma^2$.

Pour tout $c \geqslant 0$, $\quad P(Z \geqslant t) \leqslant \dfrac{\sigma^2 + c^2}{(t+c)^2}$ (1).

Preuve : Posons $Y = Z + c$ et $t + c = b$. Nous avons $(Z \geqslant t) = (Y \geqslant b) \subset (Y^2 \geqslant b^2)$. Appliquons l'inégalité de Markov avec Y^2 et $\lambda = b^2$:

Nous avons $\mathbf{P}(Y^2 \geqslant b^2) \leqslant \dfrac{\mathbb{E}(Y^2)}{b^2}$. Par linéarité de l'espérance, nous obtenons $\mathbb{E}(Y^2) = \mathbb{E}((Z+c)^2) = \mathbb{E}(Z^2) + 2c\mathbb{E}(Z) + c^2 = \sigma^2 + c^2$. En substituant les expressions de $\mathbb{E}(Y^2)$ et de b nous obtenons l'inégalité (1) annoncée au dessus.

Nous allons maintenant en déduire une version latéralisée de l'inégalité de Bienaymé-Tchebycheff :

$$\text{Pour tout } t > 0, \quad P(X - \mathbb{E}(X) \geqslant t) \leqslant \frac{\sigma^2}{\sigma^2 + t^2} \qquad (2)$$

Pour ce faire, nous montrons[4] que la fonction $c \mapsto \dfrac{\sigma^2 + c^2}{(t+c)^2}$ admet un minimum global sur $\mathbb{R}_+$.

Exprimons $\dfrac{\sigma^2 + c^2}{(t+c)^2}$ à l'aide de b, t :

$\dfrac{\sigma^2 + c^2}{(t+c)^2} = \dfrac{\sigma^2 + (b-t)^2}{b^2} = 1 + \dfrac{\sigma^2 + t^2 - 2bt}{b^2} = (\sigma^2 + t^2)\dfrac{1}{b^2} - 2t\dfrac{1}{b} + 1$. En posant $u = \dfrac{1}{b}$, nous avons affaire à un trinôme de variable u, que nous pouvons écrire $P(u) = (\sigma^2 + t^2)\left(u - \dfrac{t}{\sigma^2 + t^2}\right)^2 + 1 - \dfrac{t^2}{\sigma^2 + t^2}$. Il s'ensuit que $P(u)$ est minimum pour $u = u_0 = \dfrac{t}{\sigma^2 + t^2}$. Le minimum est $1 - \dfrac{t^2}{\sigma^2 + t^2} = \dfrac{\sigma^2}{\sigma^2 + t^2}$. La valeur c_0 de c donnant le minimum global de la fonction considérée plus haut est $\dfrac{\sigma^2}{t}$. L'inégalité annoncée est bien établie.

Nous pouvons en déduire une nouvelle inégalité en substituant $-X$ à X :

$$\text{Pour tout } t > 0, \quad P(-X + \mathbb{E}(X) \geqslant t) \leqslant \frac{\sigma^2}{\sigma^2 + t^2} \quad \text{ou encore} \quad P(X - \mathbb{E}(X) \leqslant -t) \leqslant \frac{\sigma^2}{\sigma^2 + t^2} \qquad (3)$$

En choisissant $t = \sigma$ dans (2), puis (3) nous obtenons[5] : $\boxed{\mathbf{P}(X \leqslant m + \sigma) \geqslant \frac{1}{2} \quad \text{et} \quad \mathbf{P}(X \leqslant m - \sigma) \leqslant \frac{1}{2}}$

Pour finir, nous comparons les majorants des probabilités obtenus par les deux inégalités de B.T. avec les probabilités calculées en supposant que la loi de X est "normale". Nous posons $A_k = (|X - \mathbb{E}(X)| \geqslant k\sigma)$ et $B_k = (X - \mathbb{E}(X) \geqslant k\sigma)$. Les inégalités de B.T. mentionnées dans ce qui précède peuvent s'écrire :

$$\text{Pour tout } k > 0, \quad \mathbf{P}\left(|X - \mathbb{E}(X)| \geqslant k\sigma\right) \leqslant \frac{1}{k^2} \quad \text{et} \quad \mathbf{P}\left(X - \mathbb{E}(X) \geqslant k\sigma\right) \leqslant \frac{1}{1+k^2}$$

k	$\dfrac{1}{k^2}$	$\dfrac{1}{1+k^2}$	$\mathbf{P}(A_k)$	$\mathbf{P}(B_k)$
0,1	10000,00%	99,00%	92,00%	46,00%
0,5	400,00%	80,00%	61,70%	30,90%
0,9	123,00%	55,00%	36,80%	18,40%
1	100,00%	50,00%	31,70%	15,90%
1,64	37,00%	27,00%	10,00%	5,00%
1,96	26,00%	20,70%	5,00%	2,50%
2	25,00%	20,00%	4,55%	2,28%
3	11,11%	10,00%	0,27%	0,13%
4	6,25%	5,88%	0,01%	0,00%
4,36	5,26%	5,00%	0,00%	0,00%
4,47	5,00%	4,76%	0,00%	0,00%
9,95	1,01%	1,00%	0	0
10	1,00%	0,99%	0	0

Nous remarquons bien sûr que contrairement à sa version latéralisée, l'inégalité de B.T. classique fournit des majorants de probabilités sans intérêt pour $k \leqslant 1$!

La médiocre qualité des majorants obtenus dans les deux cas s'explique par le fait que l'inégalité de Markov dont elles découlent, ne prend en compte que la valeur moyenne de la distribution de X sans tenir autrement compte de la loi de probabilité de X. Malgré cela, cette inégalité conserve cependant un bon rapport qualité/prix.

[4] par accès de « trinomite » aiguë, qui m'a laissé des séquelles. Une dérivation /c conduit aussi facilement au même résultat.

[5] l'inégalité (2) donne $\mathbf{P}(X < m + \sigma) \geqslant \frac{1}{2}$ qui implique $\mathbf{P}(X \leqslant m + \sigma) \geqslant \frac{1}{2}$ par croissance de la probabilité.