Moyenne ou médiane?

G. Haesbroeck

Département de Mathématique – Université de Liège

Congrès de la SBPMef, août 2017

Définitions

Soient $x_1, x_2, \ldots, x_n$ n observations quantitatives.

Moyenne:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}.$$

Médiane: *observation du milieu* après avoir ordonné les valeurs de la plus petite à la plus grande:

$$x_{(1)} \leq \ldots\ldots \leq x_{(k+1)} \leq \ldots\ldots x_{(k)} \leq x_{(k+1)} \leq \ldots \leq x_{(n)}$$

Si n est impair $(n = 2k + 1)$,

$$\tilde{x} = x_{(k+1)}$$

Si n est pair $(n = 2k)$,

$$\tilde{x} = \frac{x_{(k)} + x_{(k+1)}}{2}$$

Moyenne ou médiane?

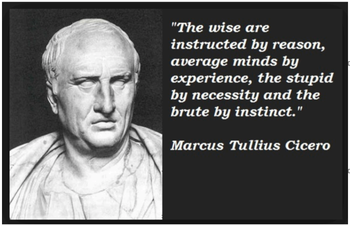
La notion de moyenne est souvent rencontrée dans notre vie de tous les jours et dans les médias.

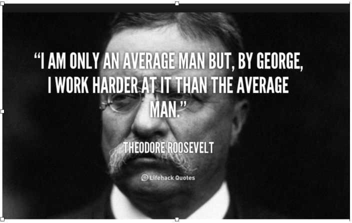
La notion de médiane est moins fréquente mais elle est tout de même rentrée dans les usages courants.

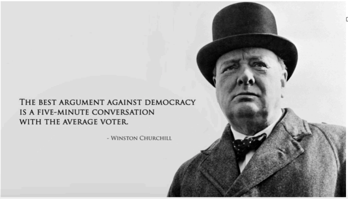
Vie de tous les jours: l'expression "être dans la moyenne"

"I SEE WE HAVE AN AVERAGE INCOME, WE'RE AN AVERAGE AGE, AND OUR HOUSE IS AN AVERAGE SIZE. NELSON, I WANT A DIVORCE."

Citations sur l'"homme moyen"







Le concept d'"homme moyen" était au centre des travaux de notre compatriote Adolphe Quetelet (1796-1874).

La moyenne dans la presse

RTL INFO — ACTU | SPORT | PEOPLE & BUZZ | VOUS | VIDÉOS | ALERTEZ-NOUS

> ACTU > BELGIQUE > SOCIÉTÉ

On en consomme en moyenne 4 kilos chaque année: mais qu'est-ce qui fait le succès des chips?

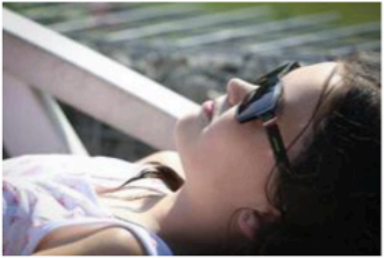
Publié le 03 août 2017 à 14h18 |

La moyenne dans la presse

RTL INFO ACTU SPORT PEOPLE & BUZZ VOUS VIDÉOS ALERTEZ-NOUS

La moyenne des températures maximales a battu un record en juin

Agence Belga , publié le 01 juillet 2017 à 09h55 |



(Belga) Le mois de juin a été particulièrement chaud cette année, la moyenne des températures maximales - 23,9° - ayant battu son record de 2003 (23,8°), indique samedi l'IRM.

En raison de la vague de chaleur qui a touché le pays du 18 au 22 juin, les températures moyennes du mois ont été très élevées à Uccle.

La médiane dans la presse

RTL INFO | ACTU | SPORT | PEOPLE & BUZZ | VOUS | VIDÉOS | ALERTEZ-NOUS

> ACTU > MONDE > INTERNATIONAL

Première augmentation du revenu médian aux Etats-Unis depuis la récession

AFP , publié le 13 septembre 2016 à 18h48 |

Pour la première fois depuis le début de la crise financière il y a huit ans, le revenu médian des ménages américains a augmenté en 2015 sur un an, même s'il n'a pas encore retrouvé son pic de la fin des années 1990.

Le revenu médian est défini comme le niveau situé entre la moitié des revenus de ceux qui gagnent le moins et l'autre moitié qui gagne le plus. Il reflète donc bien la situation de la classe moyenne.

La médiane dans la presse

rtbf.be — Info | Sport | Culture | Auvio | TV | Radio | Plus | S'inscrire | Se connecter

INFO

À la une | Fil Info | Belgique | Plus | Alertez-nous

Régions | Bruxelles | Brabant Wallon | Hainaut | Liège | Namur | Plus

Le prix médian des maisons à Bruxelles plutôt stable

RTBF - © RTBF

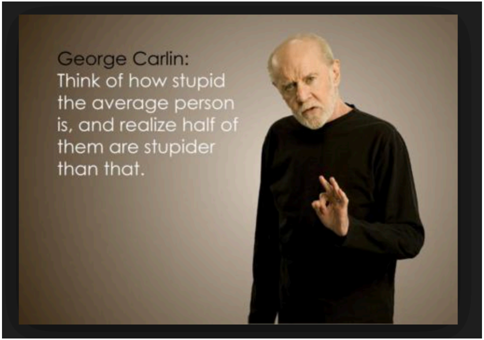
Rédaction RTBF

Le prix médian des maisons à Bruxelles-Ville a chuté de 26,4% en 2010, atteignant 277 000 euros. Celui des maisons à Woluwe-Saint-Pierre a diminué de 14,6% pour s'établir à 360 000 euros, d'après l'association des Ventes notariales de Bruxelles.

L'association NVN a choisi de ne retenir que les prix médians, qui ne tiennent pas compte des prix extrêmes qui faussent les moyennes.

Les trois communes les plus chères sont désormais Etterbeek (392 500 euros), Woluwe-Saint-Lambert (387 500 euros) et Watermael-Boitsfort (375 000 euros).

"Mélange" entre moyenne et médiane



> Quand la grande majorité d'une population est très au-dessous de la moyenne, il y a de quoi s'inquiéter.

Les euphorismes de gregoire - **Grégoire Lacroix**

Moyenne ou médiane: quelle mesure exploiter?

La moyenne et la médiane sont deux *milieux* classiques, correspondant à différentes façons de définir une valeur centrale.

Mais il y a de nombreuses autres propositions dans la littérature:

- Le mid-range: $\frac{x_{(1)}+x_{(n)}}{2}$
- Les moyennes tronquées au seuil α: $\bar{x}_\alpha = \frac{1}{n-2[\alpha n]} \sum_{i=[\alpha n]+1}^{n-[\alpha n]} x_{(i)}$
- Les moyennes "Winsorisées":

$$\bar{x}_{l,r} = \frac{1}{n}\left(l x_{(l+1)} + x_{(l+1)} + x_{(l+2)} + \ldots + x_{(n-r-1)} + x_{(n-r)} + r x_{(n-r)}\right)$$

- Le "shorth" au seuil α: point milieu du plus petit intervalle contenant au moins $(1-\alpha)n$ observations.
- ...

La comparaison pourrait être plus large mais nous nous concentrons sur la moyenne et la médiane.

Plan de l'exposé

1. Comparaison de la moyenne et la médiane sur deux exemples
2. Ecart, dispersion et concentration (et inégalité de Tchebychev)
3. Premier pas vers la statistique inférencielle et efficacité
4. Robustesse

Illustration concrète 1: tailles de 107 étudiants de 1er BAC Ing Gestion

156	159	160	160	160	160	160	163	163	163
165	165	165	165	165	166	167	168	168	168
169	169	169	170	170	170	170	170	170	170
171	171	172	173	173	173	173	174	175	175
175	175	175	175	175	175	175	176	176	176
176	177	177	177	177	177	177	177	178	178
178	178	179	180	180	180	180	180	180	180
180	180	180	180	180	182	182	182	182	183
183	183	184	185	185	185	185	185	185	185
186	186	186	187	187	187	187	188	189	190
190	190	190	192	192	194	195			

Taille moyenne: 176.5cm
Taille médiane: 177cm

Moyenne ≈ Médiane dans cet exemple "symétrique"

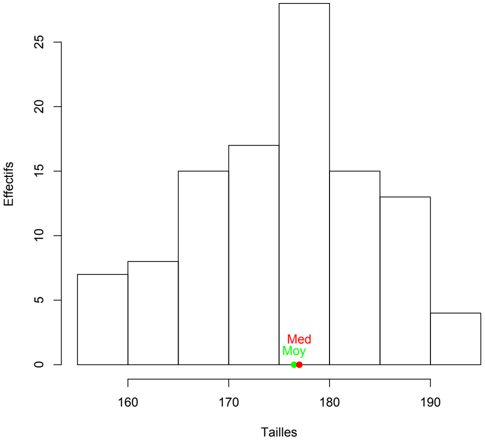


Illustration concrète 2: salaires (Source: Blog d'Horacio Gonzalez)

Dix diplômés ingénieurs civils se retrouvent pour fêter le dixième anniversaire de leur année de promotion.

Le salaire moyen vaut 37.5 k €

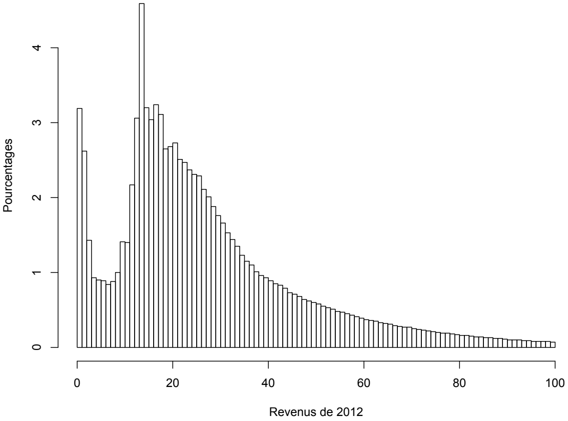
28–29–31–32–32–32–33–34–34–90

Le salaire médian vaut 32 k €

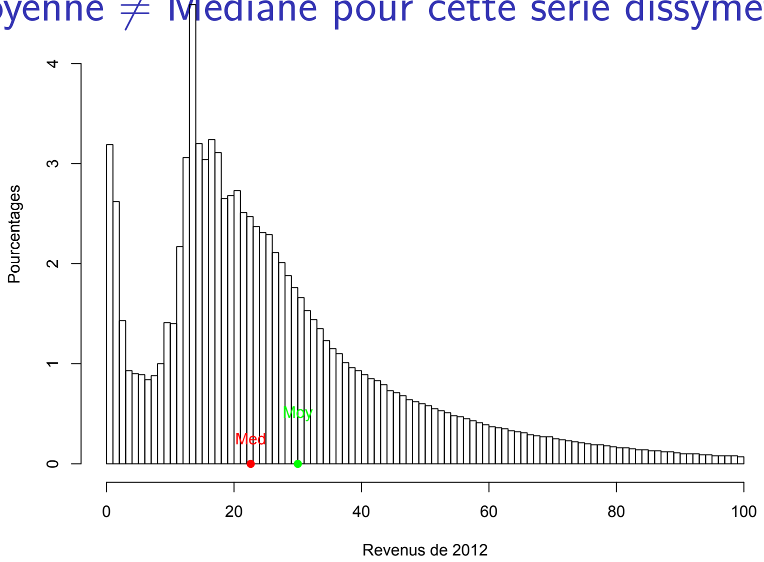
Exemple réel: répartition des revenus en Belgique en 2012

Classes (unité: 1000 Euro)		Nombres de déclarations	Pourcentages
1	[0, 1[	196536	3.19
2	[1, 2[	161248	2.62
3	[2, 3[	87849	1.43
4	[3, 4[	57573	0.93
5	[4, 5[	55228	0.90
6	[5, 6[	54656	0.89
...			
94	[93, 94[	5584	0.09
95	[94, 95[	5539	0.09
96	[95, 96[	5184	0.08
97	[96, 97[	4980	0.08
98	[97, 98[	4841	0.08
99	[98, 99[	4650	0.08
100	[99, 100[	4497	0.07
101	[100, $+\infty$[	143678	2.33
		6.157.995	100

Histogramme des revenus de 2012



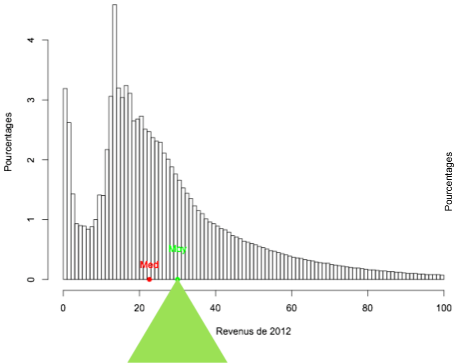
Moyenne $\neq$ Médiane pour cette série dissymétrique

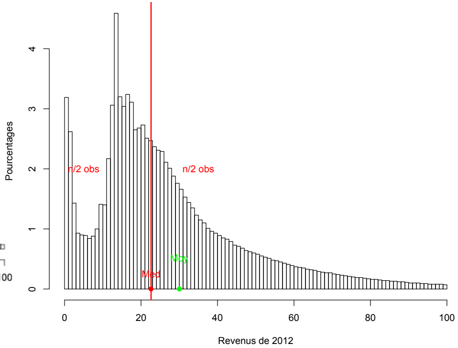


$\bar{x} = 30012$ et $\tilde{x} = 22623$

Comment expliquer cette différence?

Equilibre des masses $\longleftrightarrow$ Partage des masses





La "masse" des revenus de la classe $[100, +\infty[$ comptabilise 12.6 % de la masse totale des revenus!

Critère $L_2 \longleftrightarrow$ Critère L_1

Un centre "optimal" devrait être proche de l'ensemble des observations... ou, du moins, rendre aussi petits que possible les écarts entre les observations et lui-même.

La moyenne et la médiane vérifient chacune une telle propriété d'optimalité.

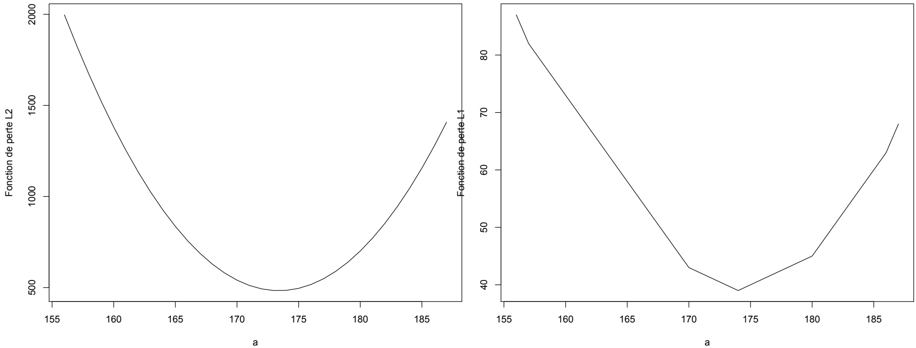
Critère (L_2) des moindres carrés pour la moyenne:

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 \leq \sum_{i=1}^{n}(x_i - a)^2 \ \forall\ a \in \boldsymbol{R}.$$

Critère (L_1) des moindres valeurs absolues pour la médiane:

$$\sum_{i=1}^{n}|x_i - \tilde{x}| \leq \sum_{i=1}^{n}|x_i - a| \ \forall\ a \in \boldsymbol{R}.$$

Critère $L_2 \longleftrightarrow$ Critère L_1



Ecart entre moyenne et médiane

Vu les définitions et caractéristiques différentes des deux paramètres, il est normal d'observer un écart entre $\bar{x}$ et $\tilde{x}$. Que peut-on dire à propos de celui-ci?

Par définition, $\bar{x}$ et $\tilde{x}$ sont comprises entre $x_{(1)}$ et $x_{(n)}$. D'où

$$|\bar{x} - \tilde{x}| \leq E$$

où E est l'étendue de la série, $x_{(n)} - x_{(1)}$.

On peut améliorer la borne comme suit:

$$|\bar{x} - \tilde{x}| \leq \frac{E}{2}$$

Démonstration (en supposant n pair, $n = 2k$)

Notons S_g et S_d les sommes $\sum_{i=1}^{k}(x_{(i)} - \tilde{x})$ et $\sum_{i=k+1}^{n}(x_{(i)} - \tilde{x})$.

On a $\bar{x} - \tilde{x} = \frac{1}{n}\sum_{i=1}^{n}\left(x_{(i)} - \tilde{x}\right) = \frac{1}{n}\left(S_g + S_d\right)$.

De là, $S_g \leq n(\bar{x} - \tilde{x}) \leq S_d$.

Or,

$$0 \leq x_{(i)} - \tilde{x} \leq E \;\text{ pour } i = k+1, \ldots, n$$

$$\text{et } \; -E \leq x_{(i)} - \tilde{x} \leq 0 \;\text{ pour } i = 1, \ldots, k$$

D'où

$$-\frac{k}{n}E \leq \bar{x} - \tilde{x} \leq \frac{k}{n}E,$$

ou encore $|\bar{x} - \tilde{x}| \leq \frac{k}{n}E \leq \frac{E}{2}$

Hotteling et Solomons, 1932

THE LIMITS OF A MEASURE OF SKEWNESS

By Harold Hotelling and Leonard M. Solomons, *Columbia University*

The measure of skewness

$$\underline{s} = \frac{\text{mean -- median}}{\text{standard deviation}}$$

is sometimes recommended because of its simplicity. Obviously neither this nor any other statistic can be of much value until something at least is known of its distribution in samples from populations of some plausible form. For populations near the normal form the inefficiency of the median as a statistic of location suggests that the standard error of $\underline{s}$ may be considerably greater than that of μ_3/σ^3. We know of no investigation of the sampling distribution of $\underline{s}$. Apparently even the range is unknown. The object of the present note is to show that $\underline{s}$ necessarily lies between -1 and 1.

ce qui donne $|\bar{x} - \tilde{x}| \leq s$ où s est l'écart-type.

Choix du centre et mesure de la répartition de la masse autour du centre: inégalité de Tchebychev

Pafnouti Tchebychev
(1821–1894)
Mathématicien russe (probabilité – statistique – théorie des nombres)

Tchebychev? Chebyshev?

Il n'existe aucune transcription officielle des noms russes. Les traductions se basent habituellement sur la phonétique:

- Les Anglais écrivent *Chebyshev*;
- Les Allemands écrivent *Tschebyschow*;
- Les Hollandais écrivent *Tsjebysjev*.

Sa contribution scientifique la plus connue dans le contexte de la théorie des probabilités est l'inégalité de *Bienaymé-Tchebychev*: soit X une variable aléatoire de moyenne μ et d'écart-type σ. Pour tout $t > 0$,

$$\boldsymbol{P}(|X - \mu| \geq t\,\sigma) \leq \frac{1}{t^2}.$$

C'est la version *empirique* de cette propriété qui a été insérée dans le référentiel.

- Tchebychev, gloire nationale, B. Hauchecorne, *Tangente*, 132, Janvier-février 2010.
- P.L. Chebyshev (1821-1894), P. Butzer et F. Jongmans, *Journal of Approximation theory*, 1999.

Inégalité de Tchebychev

Proposition

Soit $S = \{x_1, \ldots, x_n\}$ *une série de moyenne* $\bar{x}$ *et d'écart-type* s*.*
La proportion d'observations s'écartant d'au moins t *écarts-types de la moyenne est inférieure ou égale à* $\frac{1}{t^2}$*.*

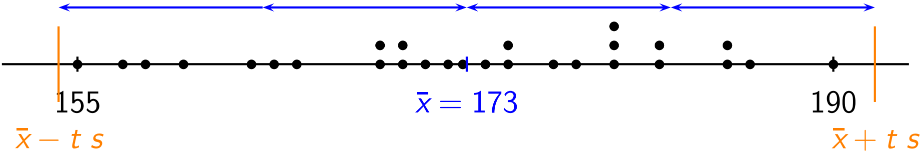
La démonstration de cette propriété ne fait appel qu'aux outils suivants:

- Définition de s^2;
- Exploitation des valeurs absolues;
- Majoration de sommes.

Ces outils étaient déjà exploités par Tchebychev dans sa preuve originale.

Interprétation de la propriété de Tchebychev

Considérons 28 tailles d'étudiants de BLOC 1 Math

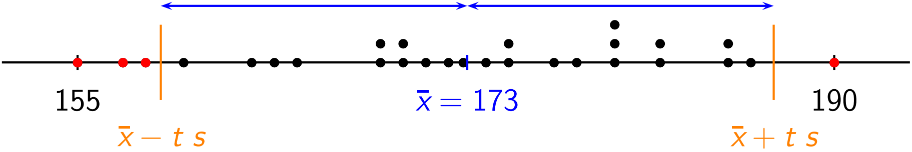


$s = 9.4$. Prenons $t = 2$

Aucune observation ne s'écarte d'au moins 2 écarts-types de la moyenne; L'inégalité est vérifiée!

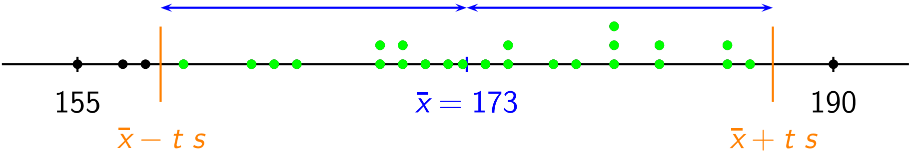
Interprétation de la propriété de Tchebychev

Prenons maintenant $t = 1.5$



Il y a 4 observations s'écartant d'au moins 1.5 écarts-types de la moyenne; La proportion observée est donc $\frac{4}{28} = 0.14 \leq 0.44 = \frac{1}{t^2}$

De façon équivalente: la proportion d'observations situées dans l'intervalle $]\bar{x} - t\ s, \bar{x} + t\ s[$ vaut au moins $1 - \frac{1}{t^2}$.



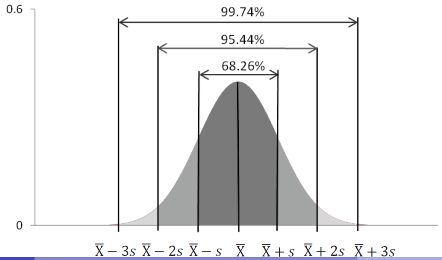
Il y a 24 observations dans l'intervalle; ce qui donne une proportion observée égale à 0.86$>$ 0.55.

Intérêt de la propriété de Tchebychev

- Discussion sur la répartition de la masse;
- Aucune hypothèse n'est imposée à la série de données;
- L'inégalité est aussi valide dans un espace probabilisé, cadre dans lequel elle est utile pour démontrer des résultats fondamentaux de convergence probabiliste (par ex: loi des grands nombres)

En pratique, la borne est très peu contraignante... et peu informative surtout lorsqu'une hypothèse paramétrique est plausible (y revenir lorsque l'on discute des masses sous la loi normale!)

Pour une distribution « en cloche », on a approximativement :



Et si on préfère travailler avec la médiane?

Une propriété de Tchebychev est également disponible!

Dans ce cas, un paramètre de dispersion plus naturel serait l'écart-absolu médian, défini par

$$E_{\tilde{x}} = \frac{1}{n} \sum_{i=1}^{n} |x_i - \tilde{x}|$$

Proposition

Soit $S = \{x_1, \ldots, x_n\}$ une série de médiane $\tilde{x}$ et d'écart-absolu médian $E_{\tilde{x}}$.
La proportion d'observations s'écartant d'au moins t écart-absolu médian de la médiane est inférieure ou égale à $\frac{1}{t}$.

La démonstration suit exactement la même logique que celle basée sur la moyenne et l'écart-type.

Un premier pas vers la statistique inférentielle

Dans la plupart des analyses statistiques, il n'est pas possible de mesurer la variable statistique sur l'ensemble de la population.

Seuls certains individus de la population sont considérés. Ils constituent un *échantillon* que l'on espère représentatif de la population visée.

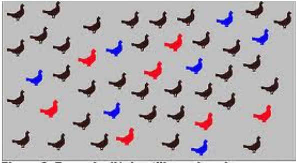
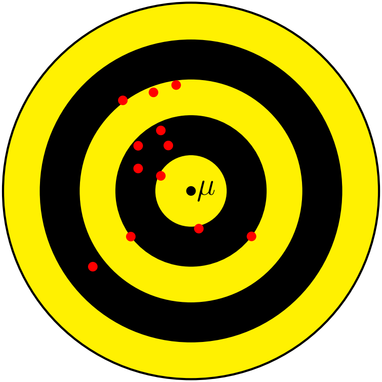


Figure 2. Exemple d'échantillons : les oiseaux en bleu constituent un échantillon, ceux en rouge, un autre

Le calcul d'une moyenne ou d'une médiane sur l'échantillon vise à estimer le *vrai* centre de la population.

Qualité de l'estimation

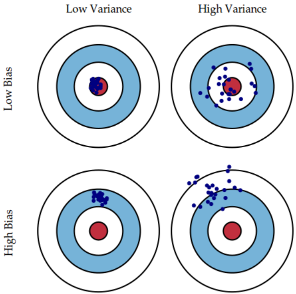
Soit μ le *vrai* centre de la population.



Le calcul de la moyenne ou de la médiane sur chaque échantillon donne une estimation du centre! Les estimations *fluctuent* autour du centre de la cible!

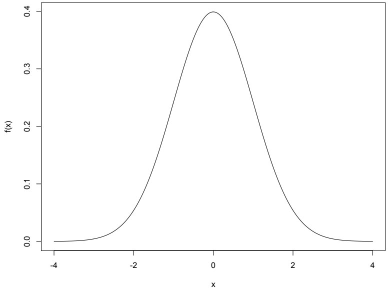
Qualité de l'estimation/l'estimateur

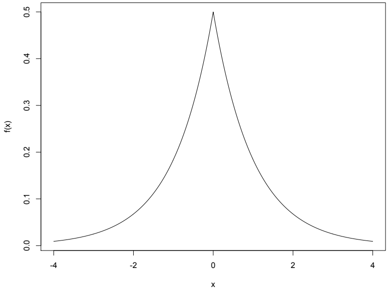
Plusieurs comportements sont possibles de la part d'un *estimateur*:



Et la performance de l'estimateur dépend de la distribution.

Performance de la moyenne et de la médiane



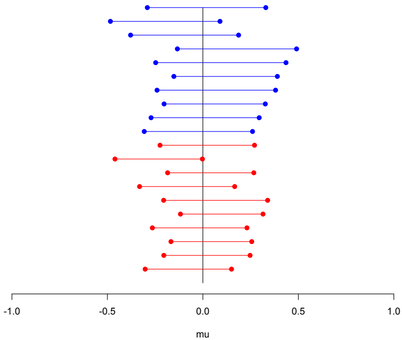


Distribution normale
Variance de $\tilde{x}$
$\approx \pi/2$ Variance de $\bar{x}$

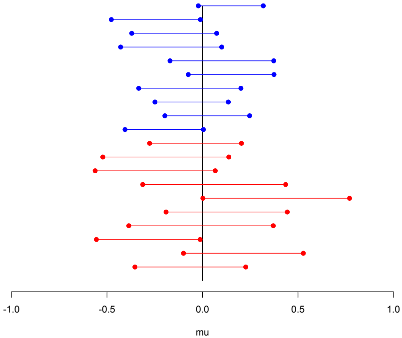
Distribution Laplace
Variance de $\tilde{x}$
$\approx 1/2$ Variance de $\bar{x}$

Interprétation des efficacités relatives

En rouge (resp. bleu): intervalles de confiance à 95% construits sur la moyenne (resp. médiane), $n = 50$.



Sous la normalité



Sous la distribution Laplace

Efficacité $\longleftrightarrow$ Robustesse

La performance des estimateurs se mesure habituellement sous un modèle spécifique (distribution F).

Dans la pratique, cependant, une fraction des observations peut être "contaminée" (erreurs d'encodage, faits exceptionnels...).

Le modèle théorique préconisé par les statisticiens robustes est plutôt le mélange suivant:

$$(1-\varepsilon)F + \varepsilon G$$

Les techniques qualifiées de *robustes* devraient être "insensibles" à la présence d'observations contaminées (minoritaires) et toujours estimer les caractéristiques de la distribution F.

Illustration

Soit la distribution $F = (1 - \varepsilon)N(0,1) + \varepsilon N(7, 0.25^2)$.

Si on part du principe que la cible est la distribution $N(0,1)$, on souhaite que les estimations de tendance centrale soient proches de 0.

Prélevons 10 échantillons de taille 100 de cette distribution et calculons sur chacun la moyenne, la médiane et la moyenne tronquée de seuil 0.1.

Illustration pour $\varepsilon = 0$

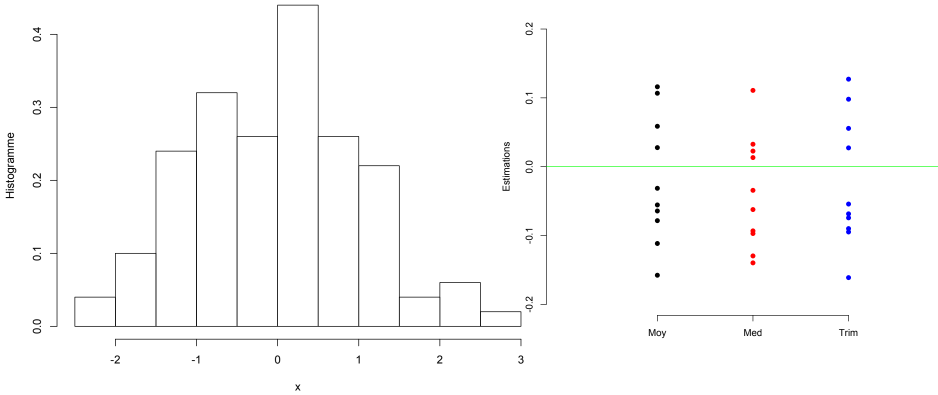
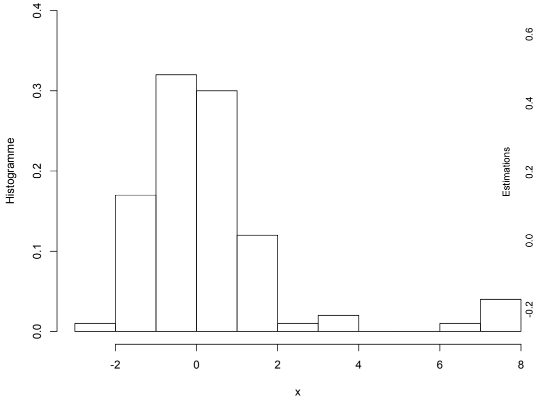


Illustration pour $\varepsilon = 0.05$



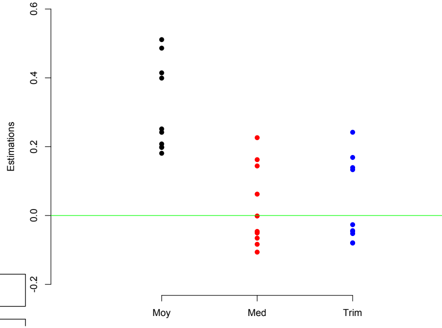
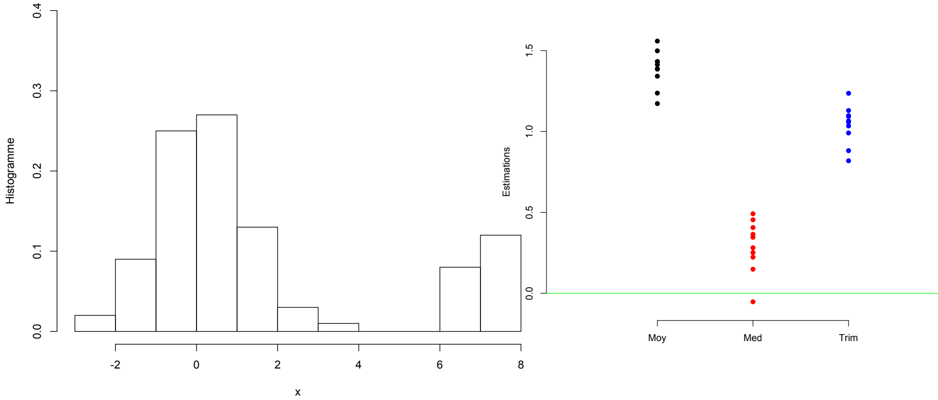


Illustration pour $\varepsilon = 0.20$



Mesures empiriques de robustesse

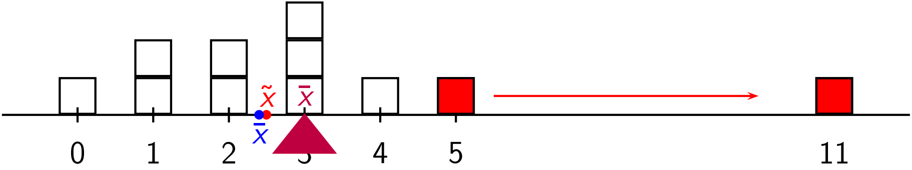
Des simulations permettent de déterminer si certaines techniques d'estimation sont robustes ou pas.

Il existe également la possibilité de mesurer la robustesse à partir d'outils disponibles dans la littérature. Certains de ceux-ci sont exploitables dans le contexte empirique:

- Mesure globale de robustesse: le point de rupture
- Mesure locale de robustesse: la fonction d'influence

Le point de rupture

Il s'agit de la fraction minimale d'observations à contaminer pour emmener l'estimation au delà de toute borne finie.
Prenons un exemple concret:

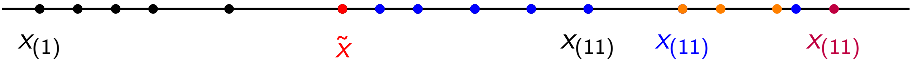


La médiane n'est pas perturbée par cette "contamination"!

Par contre, pour conserver l'équilibre, la moyenne est obligée de se décaler. Et si on accentue la perturbation, la moyenne sautera au delà de toute borne préalablement fixée.

Le point de rupture de la moyenne est égal à $1/n$!

et le point de rupture de la médiane?



La médiane peut résister jusqu'à 50% de contamination dans les données! Son point de rupture vaut $1/2$. Il s'agit de la technique d'estimation la plus robuste pour une tendance centrale.

Et les autres paramètres de tendance centrale?
Mid-range: $1/n$
Moyenne tronquée: α

La fonction d'influence

Elle mesure l'effet de l'ajout d'une observation quelconque sur le calcul d'un paramètre statistique.

Soient $x_1, \ldots, x_n$ les n observations de départ et x une nouvelle observation.

La fonction d'influence empirique est définie par

$$EIF(x) = \frac{\text{Estimation perturbée} - \text{Estimation initiale}}{1/n}$$

Ce qui importe, c'est le caractère borné de la fonction d'influence, illustrant ainsi la résistance de l'estimateur par rapport à une contamination très partielle.

Calcul de la fonction d'influence de la moyenne

La moyenne de départ vaut $\frac{x_1+x_2+\ldots+x_n}{n}$
Après avoir ajouté x, la moyenne devient $\frac{x_1+x_2+\ldots+x_n+x}{n+1}$
L'écart entre la moyenne perturbée et la moyenne initiale vaut donc

$$\frac{x_1+x_2+\ldots+x_n+x}{n+1} - \frac{x_1+x_2+\ldots+x_n}{n}$$
$$= \frac{n(x_1+x_2+\ldots+x_n+x)-(n+1)(x_1+x_2+\ldots+x_n)}{n\,(n+1)}$$

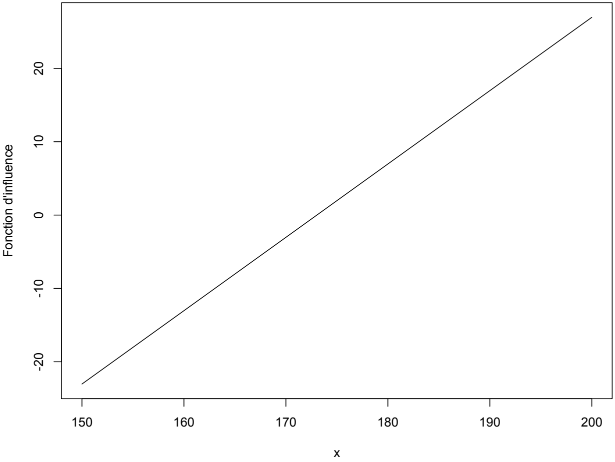
Il reste

$$\frac{n\,x-(x_1+x_2+\ldots+x_n)}{n\,(n+1)} = \frac{x-\bar{x}}{n+1}$$

En tenant compte de l'importance de la perturbation, il vient

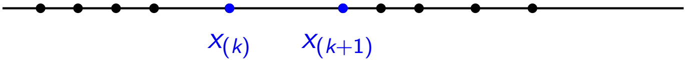
$$EIF(x) = x - \bar{x}$$

Sur les 28 tailles de BLOC 1 math

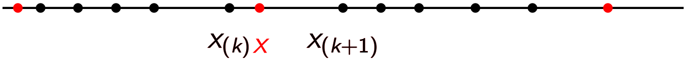


Calcul de la fonction d'influence de la médiane

Supposons n pair ($n = 2k$).
La médiane de départ vaut $\frac{x_{(k)}+x_{(k+1)}}{2}$



Après avoir ajouté x, la médiane devient l'observation du milieu d'un ensemble de $2k+1$ valeurs, et plus précisément



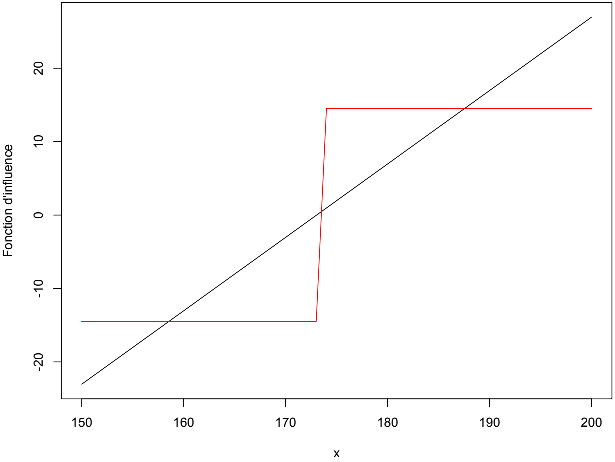
En d'autres termes, $\begin{cases} x_{(k)} & \text{si } x \leq x_{(k)} \\ x & \text{si } x_{(k)} < x \leq x_{(k+1)} \\ x_{(k+1)} & \text{si } x > x_{(k+1)} \end{cases}$

Calcul de la fonction d'influence de la médiane (suite)

Au final,

$$EIF(x) = \begin{cases} (n+1)(x_{(k)} - x_{(k+1)})/2 & \text{si } x \leq x_{(k)} \\ (n+1)\left(x - (x_{(k)} + x_{(k+1)})/2\right) & \text{si } x_{(k)} < x \leq x_{(k+1)} \\ (n+1)(x_{(k+1)} - x_{(k)})/2 & \text{si } x > x_{(k+1)} \end{cases}$$

Sur les 28 tailles de BLOC 1 math



Conclusion: moyenne ou médiane?

- Les deux paramètres sont utiles en analyse exploratoire.
  - Lorsque les deux paramètres donnent une estimation similaire du centre, “tout va bien”.
  - Lorsque les deux paramètres ne donnent pas une estimation similaire, il faut approfondir l’analyse.
- Les notions de fonction d’influence et de point de rupture sont simples à introduire, exploitent des notions élémentaires de mathématique et permettent de montrer que la statistique ne se résume pas à la bête application de formules.
- Une première introduction à la statistique inférentielle (fluctuations des estimations...) rendrait les élèves plus critiques par rapport à certains résultats de sondage présentés dans la presse.