Chapitre 2 :

1 : Moyenne, médian, mode

Définition 1 : La moyenne empirique de l'échantillon x_i i = 1, 2, …, N, notée $\overline{x}$, est la moyenne arithmétique de ces vals observées.

$$\overline{x} = \frac{1}{N} \sum_{i=1}^{N} x_i$$

Définition 2 : Le mode empirique de l'échantillon x_i i=1, …, N est la valeur qui correspond à la fréquence la plus élevé
Graphique sur feuille

Remarques :

- Le mode empirique est la valeur la plus fréquemment observée ds l'échantillon
- Le mode n'est pas nécessairement unique
- On utilise le mode principalement pr étudier les distributions des vars catégoriques ou des vars num discrètes
- Soit x une var num continue dont les vals sont regroupées ds les classes C_l l = 1, …, L. Le centre de la classe correspond à la fréquence la plus élevée est définie comme le mode de cette var

Exemple : x_i i = 1, 2, …, 7
PIB/hab en $

Allemagne	Canada	E-U	FR	Italie	Japon	R-U
41 259.18	42 294,23	58 559,67	35 765	29 296	34 813	43 020,20

$\overline{x} = \frac{1}{N} \sum_{i=1}^{N} x_i$ N = 7

$= \frac{1}{7} \sum_{i=1}^{N} x_i$ =1/7 (41 259,18, …, 43 020,20)

= 40 715,5

Exemple : Le tableau ci-dessous présente le nbr de paiements d'împots par les entreprises. (y_i, i = 1, …, 15) ds 15 pays

i	1	2	3	4	5	6	7	8
y_i	9	12	11	10	9	8	9	8

i	9	10	11	12	13	14	15
y_i	9	14	8	23	9	9	6

$\overline{x}$ = 1/15 (9+12+…+6) = 10,27

Propriété : (la moyenne empirique) Soit a et b des nbr réels et supposons que ns observons x_i et y_i i = 1, 2, …, N

1- Soit y, définie par yi = a + bx. Alors $\overline{y} = a + b\overline{x}$
2- x_i = ay + bx, alors $\overline{z} = a\overline{y} + b\overline{x}$

Proposition : La somme des écarts à la moyenne est nulle :

$$\sum_{i=1}^{N}(xi - \overline{x}) = 0$$

$\overline{x}$ ne dépend pas de i

Demonstration : $\sum_{i=1}^{N}(xi - \overline{x}) = \sum_{i=1}^{N} xi - \sum_{i=1}^{N} \overline{x}$

$= \sum_{i=1}^{N} xi - N\overline{x}$

$= N\overline{x} - N\overline{x} = 0$

Exemple :

1- Après avoir changé l'unité de la variable de $ en 1000$. Calculer la moyenne de cette var notée y_i i = 1, 2, …, 7

i	1	2	3	4	5	6	7
y_i	41,26	42,29	58,56	35,76	29,30	34,81	43,02

$y_i = 0 + 1/1000\ x_i$; $\overline{y} = 1/1000\ \overline{x} = 40{,}72$

$\tilde{y}$ = ecart à la moyenne

Les quantiles :

Définition : La quantile empirique d'ordre h de l'echantillon x_i i=1n…,N notée q_h est la val telle que 100 x h% des observations ds l'echantillon sont inférieures à cette val.

Remarques : Les cas spécifiques des quantiles.

1- Si k est entre 0 et 100, le quantile d'ordre h = k/100 est appelé kième centile (ou percentile)
2- Les quantiles d'ordre 0,25 ; 0,5 ; 0,75 sont appelés le premier, le deuxième, et le troisième QUARTILES repectivement. **Q_1, Q_2, Q_3**
3- Le deuxième QUARTILE Q_2, noté également m_x, est appelé le médian

Notons $x_{(i)}$ i = 1, …, N les observations placées en ordre. C'est-à-dire

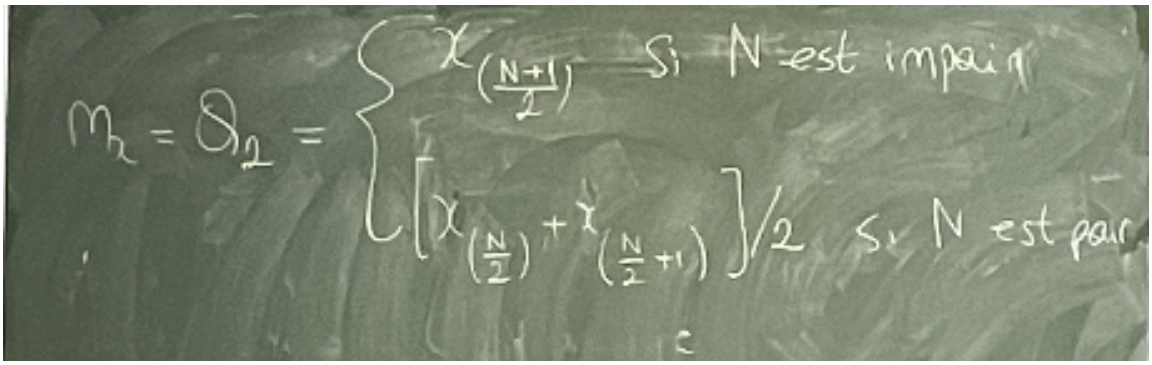
$x_{(1)} \leq x_{(2)} \leq x_{(3)}$

Nous trouvons le quantile d'ordre h en utilisant $q_h = x_{(1+(N-1)h)}$

$Q_1 = x_{(1 + (N-1)/4)}$

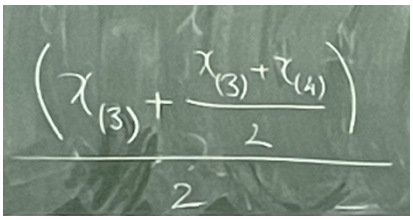
$m_x = Q_2 = x_{(1 + (N-1)/2)}$

$Q_3 = x_{(1 + (N-1)\ x\ ¾)}$



Ds ce cas où l'indice (i) n'est pas un entier on utilise la méthode d'interpelation.

Exemple : si $q_h = x_{(3,25)}$ ns prenons la val qui est située à 25% du chemin entre $x_{(3)}$ et $x_{(4)}$



$x_{(3)} + 0,25\ (x_{(4)} - x_{(3)})$

$q_h = x_{(5,6)}$
$q_h = x_{(5)} + 0,6\ (x_{(6)} - x_{(5)})$

Exercice :

i	1	2	3	4	5	6	7
y_i	41,26	42,29	58,56	35,76	29,30	34,81	43,02

i	1	2	3	4	5	6	7
y_i	29,3	34,81	35,76	41,26	42,29	43,02	58,56

Min = $y_{(1)}$ = 29,3 Max = $y_{(7)}$ = 58,56

Les quantiles d'ordre 0,2 et 0,8

a- h = 0,2 $q_h = Q_{(0,2)} = y_{(1 + (N-1) \times 0,2)} = y_{(1 + 6 \times 0,2)}$ $y_{(2,2)} = y_{(2)} + 0,2(y_{(3)} - y_{(2)})$
= 35,81 + 0,2 (35,76 – 34,81) = 35

b- $h = 0,8$. $Q_{(0,8)} = y_{(1 + 6 \times 0,8)} = y_{(5,8)} = y_{(5)} + 0,8\,(y_{(6)}-y_{(5)}) = 42,29 + 0,8\,(43,02 - 42,29) = 42,87$

$Q_2 = m_y = y_{(N+1/2)} = y_{(4)} = 41,26$

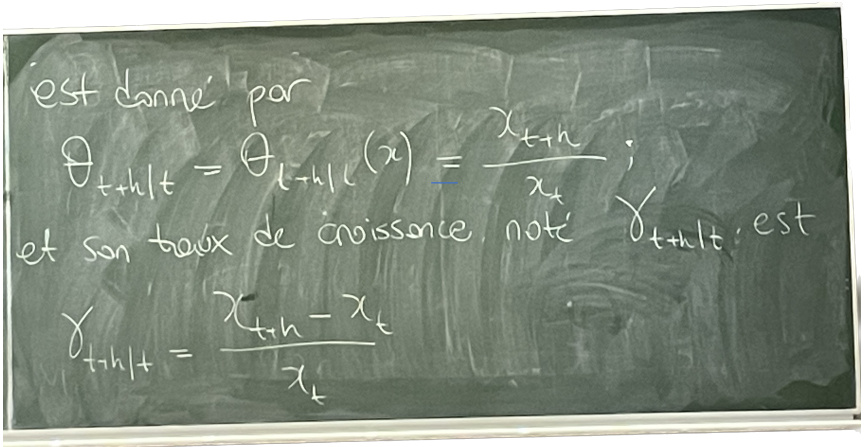
$Q_1 = Q_{(0,25)} = y_{(1 + (N-1)/4)} = y_{(2,5)} = (y_{(3)}+y_{(2)})/2 = 35,28$

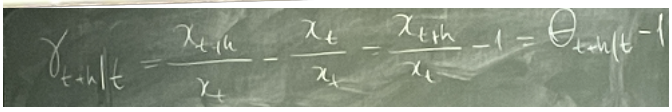
$Q_3 = 42,65$

Transformation des séries temporellles :

x_t $t = 1, 2, \ldots, T$

Définition : Soit une var x mesurée en t et t+h, h>0 Son coeff multiplicateur multiplicateur entre t et t+h, noté $\theta_{t+h\,t}$





Exercice : Déterminer le taux de croissance et coeff multiplicateur entre 2000 et 2001

(i) $x_{2001} = 3x_{2000}$. (ii) $x_{2001} = x_{2000/8}$

(i) $t = 2000$ $h = 1$. $t+h = 2001$

$\theta_{t+h\,t} = \theta_{2001\,2000} = x_{2001}/x_{2000} = 3x_{2000}/x_{2000} = 3$

$\delta_{2001\,2000} = 2$

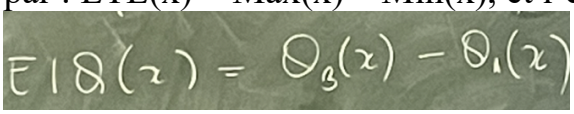
(ii) $\theta_{2001\,2000} = (x_{2000/8})/x_{2000} = 1/8$

$x_{2001\,2000} = 1/8 - 1 = -7/8$

Mesure de la dispersion :

Étendue et écart-interquartile :

Définition : l'étendue d'un échantillon x_i $i = 1, \ldots, N$, notée ETE(x) est définie par : ETE(x) = Max(x) – Min(x), et l'écart-interquartile de x est donné par



Propriété : Soit α et β deux nbr réels, x_i $i = 1, \ldots, N$ une var stat, et y_i $i = 1, \ldots, N$ définie par $y_i = \alpha + \beta x_i$ Alors $ETE(y) = \alpha + \beta ETE(x)$; $E\,Q(y) = \alpha + \beta EIQ(y)$

Exemple : y_i i = 1, 2, …, 7
Min (y) = 29,30
Max (y) = 58,56
$Q_1(y)$ = 35,28
ETE(y) = Max(y) – Min(y) = 58,56 – 29,30 = 29,26
EIQ(y) = $Q_3(y) – Q_1(y)$ = 42,25 – 35,28 = 7,7

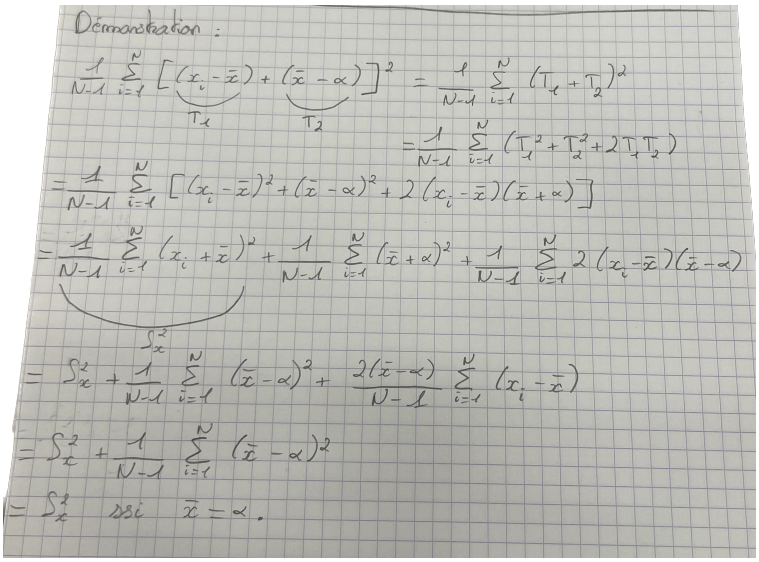
$y_{(8)}$ = 116,36
Min(y) = 29,30
Max(y) = 116,36
ETE(y) = 116,36 – 29,30 = 87,06
EIQ(y) = 11,38
EIQ est plus robuste aux vals extrêmes

Variance empirique et écart-type empirique :
Définition : la variance empirique de l'echantillon x_i i = 1, …, N notée S_x^2 est définie par : $S_x^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \overline{x})^2$

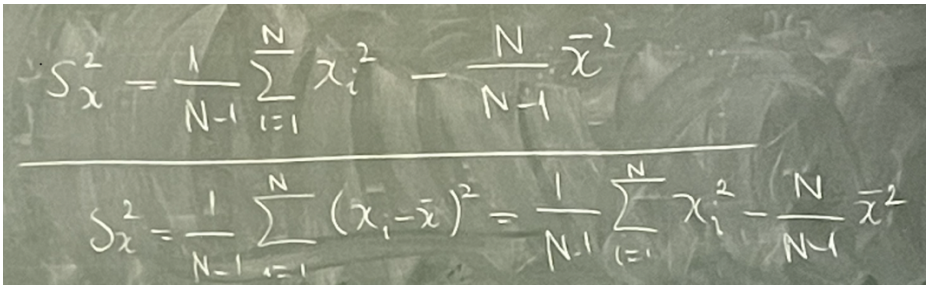
Définition : L'écart type, noté S_x est égale à $S_x = \sqrt{S_x^2}$

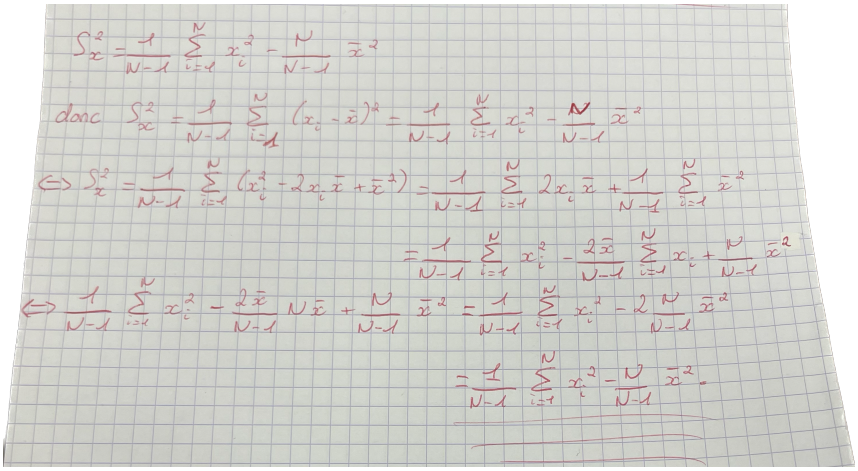
Propriété : la variance empirique de x_i i = 1, …, N est tjrs positive : $S_x^2 >= 0$

Propriété : Soit α et β deux nbr réels x_i i = 1, …, N une var stat et $y_i = \alpha + \beta x_i$
Alors $S_y^2 = \beta^2 S_x^2$; βS_x
Proposition : La moyenne des carrés des écarts par rapport à la moyenne est min. Autrement dit l'expression : $\frac{1}{N-1}\sum_{i=1}^{N}(x_i - a)^2$ est min pr $\alpha = \overline{x}$
Démonstration :



Proposition : La variance empirique de x_i i = 1, …, N peut être calculée en utilisant la formule suivante :



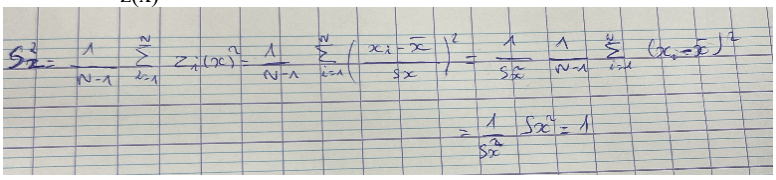


Définition : le coeff de var de x_i i = 1, …, N. $CV_x = \frac{Sx}{\overline{x}}$

Définition : Le z-score de la var x_i i = 1, …, N notée $z_i(x)$ est donné par $z_i(x) = x_i - \overline{x} \ / S_x$

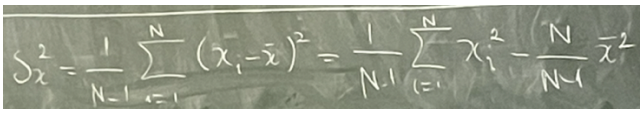
Proposition : $z_i\overline{(x)} = \frac{1}{N}\sum_{i=1}^{N} z_i(x) = \frac{1}{N}\sum_{i=1}^{N} xi - \overline{x} \ /Sx = 0$

1- $\overline{z}_{(x)} = 0$
2- $S_{z(x)} = 1$



Mesures de la relation existe entre 2 variables

Définition : Pr un échantillon de taille N des observation (x_i ; y_i) i = 1,…, N la covariance empirique est définie par : $\mathbf{S_{xy} = \frac{1}{N-1}\sum_{i=1}^{N}(xi - \overline{x})(yi - \overline{y})}$



$S_{xy} = \frac{1}{N-1}\sum_{i=1}^{N} xiyi - \frac{N}{N-1}\overline{xy}$

Définition : Le coeff de corrélation entre x_i et y_i : $r_{xy} = S_{xy} / S_xS_y$

Propriété : $-1 <= r_{xy} <= 1$

Propriété : Soit α et β deux nbrs réels, y_i et x_i deux variables stats i = 1, …, N et $g_i = \alpha + x_i$; $d_i = \beta + y_i$; $z_i = \alpha x_i$; $h_i = \beta y_i$; $w_i = x_i + y_i$

1) Sxx = S_x^2
2) Syx = Sxy
3) Sgd = Sxy
4) Szh = αβSxy
5) Sαx = 0
6) $S_w^2 = S_x^2 + S_y^2 + 2Sxy$

Calcul de la moyenne et la variance à partir d'un tableau de fréquence :

C_l	n_l
C_1	n_1
C_2	n_2
C_l	n_l
Total	N

Pr une var discrète où C_l sont les vals prises par la var :

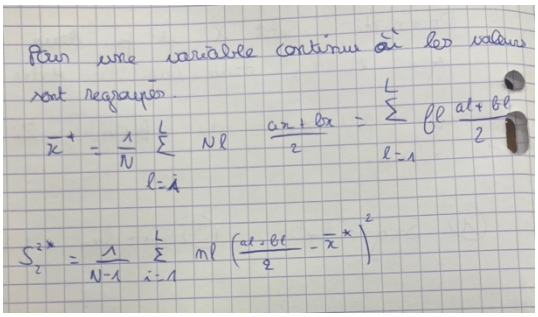
. $\overline{x} = \frac{1}{N}\sum_{i=1}^{N} nl\ Cl = \sum_{i=1}^{N} fl\ Cl$

. $\frac{1}{N-1}\sum_{i=1}^{N} nl\ (Cl - \overline{x}\ )^2$

Exemple :

i	1	2	3	4	5
x_i	2	2	1	1	2

Pour une var continue où les vals sont regroupés :



Calcul pour la position :

$\overline{x} = \frac{1}{N} \sum_{i=1}^{N} x_i$

$Q_h = x_{(1+(N+1)/h)} + Q_1 + Q_3$

Calcul pour la dispersion :

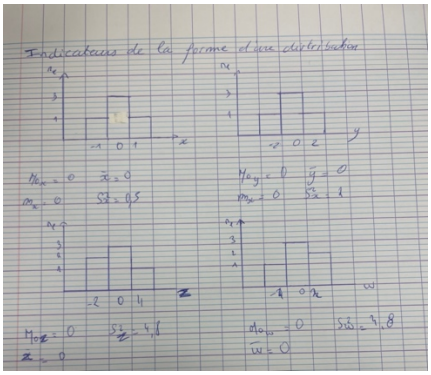
$S_x^2 = \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \overline{x})^2$

$S_x = \sqrt{S_x^2}$

$CV_x = \frac{Sx}{\overline{x}}$

ETE(y) = Max(y) – Min(y)

EIQ(y) = $Q_3(y) - Q_1(y)$



Définition : Le moment centré d'ordre r de la var x_i i=1, …, N notée $m_r(x)$ est définie par $m_r(x)$: $\frac{1}{N} \sum_{i=1}^{N} (x_i - \overline{x})^r$

Remarque :

$m_1(x) = 0$

$m_2(x) = \frac{N-1}{N} S_x^2$

Propriété : Soit α et β deux nbrs réels x_i i=1, …, N une var stat et $y_i = \alpha + \beta x_i$

Alors : $m_r(x) = \beta^r m_r(x)$

Propriété : Soit x_i i =1, …, N une var stat. Si la dist de x est symétrique

1) Mo(x) = m_x

2) Mo(x) = $\overline{x}$
3) $M_3(x) = 0$

i	1	2	3	4	5
x_i	-1	0	0	0	1
y_i	-2	0	0	0	2
z_i	-2	-2	0	0	4
w_i	-4	0	0	2	2

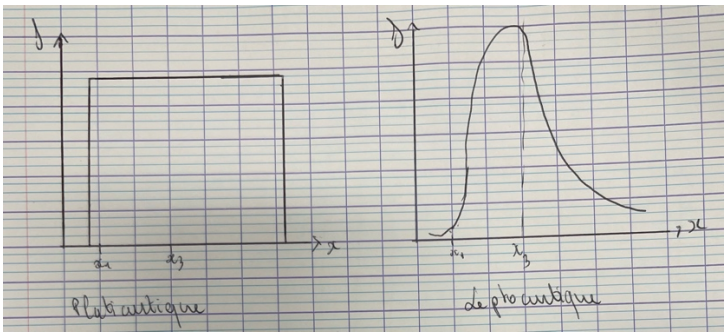
. $\overline{x} = \overline{y} = \overline{z} = \overline{w}$
. $S_x^2 = 0{,}5$. $S_y^2 = 2$. $S_z^2 = S_w^2 = 4{,}8$
$m_3(z) = \frac{1}{N}\sum_{i=1}^{N}(z_i - \overline{z})^3 = \frac{1}{5}((-2)^3 + (-2)^3 + 0^3 + 0^3 + 4^3) = 9{,}6$
$m_3(w) = \frac{1}{N}\sum_{i=1}^{N}(w_i - \overline{w})^3 = \frac{1}{5}((-4)^3 + 0^3 + 0^3 + 2^3 + 2^3) = -9{,}6$
Les distributions sont asymétriques

Pr une var x_i i = 1, …, N
1) La distribution de x est symétrique si $m_3(x) = 0$
2) La distribution est étalée vers la gauche si $m_3(x)<0$
3) La distribution est étalée vers la droite si $m_3(x)>0$

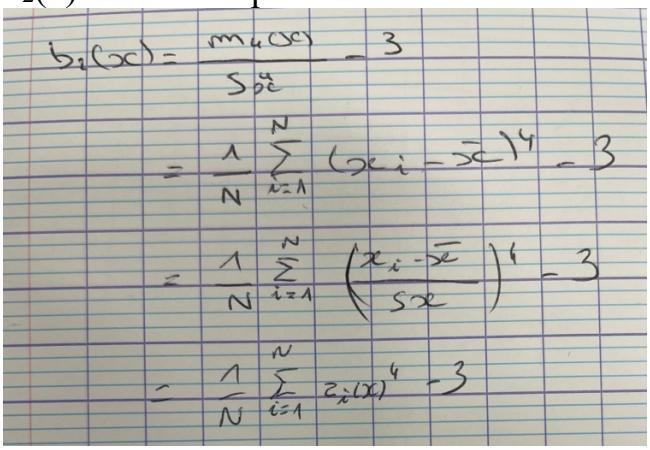
Si $m_3 < 0$ alors $b_1 < 0$
Si $m_3 > 0$ alors $b_1 > 0$
Si $m_3 = 0$ alors $b_1 = 0$

Définition : Le coeff d'asymétrie de Fisher-Pearson de x_i i=1, …,N noté $b_1(x)$ est donnée : $b_1(x) = \frac{m3(x)}{Sx\text{^}3}$

Définition :
- Une distribution est dite applatie ou platicurtique si pr une variance donnée, une forte variation de la var entraine une faible variation de sa freq relative.
- Une distribution est dite pointue ou leptocurtique si pr une variance donnée une faible variation de la var entraîne une forte var de sa freq relative



Définition : Le coeff d'applatissement de Fisher-Pearson de x_i i = 1, …, N noté $b_2(x)$ est définie par :



$$b_2(x) = \frac{m_4(x)}{s_x^4} - 3$$

$$= \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^4 - 3$$

$$= \frac{1}{N}\sum_{i=1}^{N}\left(\frac{x_i - \bar{x}}{s_x}\right)^4 - 3$$

$$= \frac{1}{N}\sum_{i=1}^{N} z_i(x)^4 - 3$$

Si $b_2(x) < 0$ la dist est platicurtique
Si $b_2(x) > 0$ la dist est leptocurtique

Exemple :

i	1	2	3	4	5	6	7	8
y_i	29,3	34,81	35,76	41,26	42,29	43,02	58,56	116,36

$z_i(y) = -0,743$ 2,3564
$z_i(y)^3 = -0,4101$ 13,0898
$z_i(y)^4 = 0,3047$ 30,8302

$b_1(y) = \frac{1}{N}\sum_{i=1}^{N} z_i(y)^2 = \frac{1}{8}((-0,743)^3 + \ldots + (2,3564)^3) = 1,5414$
Etalée vers la droite
$b_2(y) = \frac{1}{N}\sum_{i=1}^{N} z_i(y)^4 - 3 = \frac{1}{8}(0,3047 + \ldots + 30,8302) - 3 = 0,9153$
Leptocurtique / à la normal