Cours d'introduction à l'analyse statistique

4

Paramètres de dispersion d'une distribution

L3 LISS - Université Paris-Dauphine, Arnold Chassagnon, LEDa-SDFi, Octobre 2010

arnold.chassagnon@dauphine.fr

Les paramètres de dispersion évaluent le niveau d'étalement de la série autour de la valeur centrale. Ils complètent les paramètres de position en permettant de comparer des série dont les paramètres de position sont proches, mais où la forme de la dispersion est très différente. Ces notions n'ont de sens que pour des variables ordonnées.

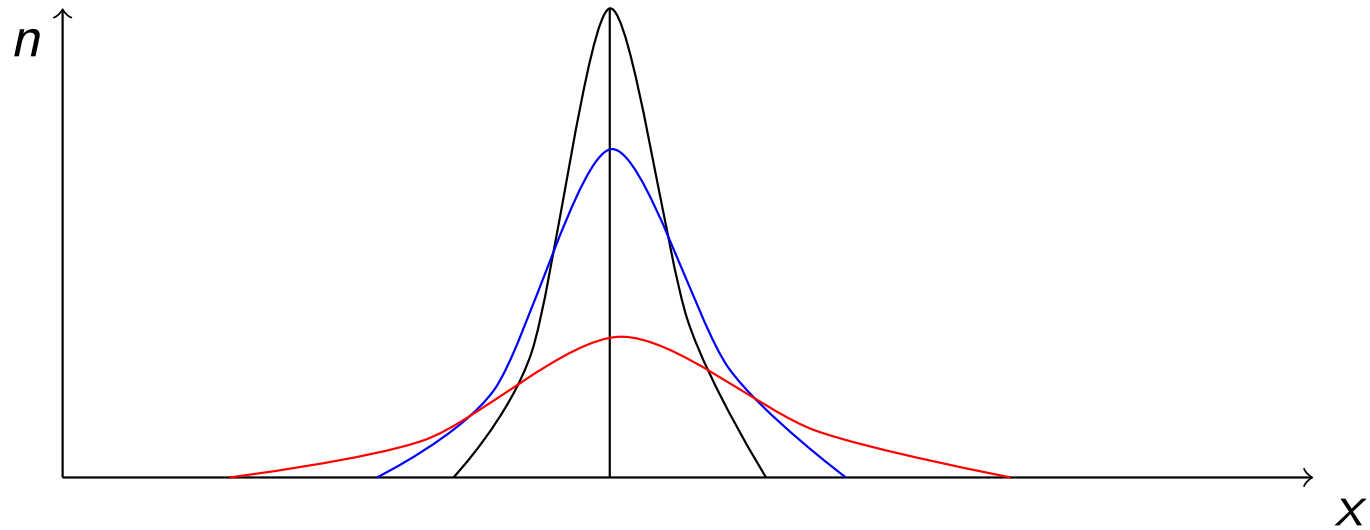


FIGURE: Les trois courbes se situent dans les mêmes gammes de valeurs, mais ont des étalements très différents. Les paramètres de dispersion, de symétrie et d'aplatissement le mettent clairement en évidence.

PLAN

- Ecart interquantile
- Variance, écart-type, coefficient de variation
- Coefficient d'asymétrie ou de skewness
- Coefficient d ?aplatissement ou Kurtosis

Ecart interquantile

L'écart interquartile est la taille de l'intervalle situé au centre de la série et incluant 50% des observations :

$$\text{écart} = Q_3 - Q_1.$$

plus cet écart est grand, plus la dispersion des observations est forte.

Variance

La variance mesure la distance des réalisations de la variable par rapport à la moyenne.

Définition

La variance est définie comme un moment d'ordre 2.

$$Var(X) = E\left[(X - E(X))^2\right]$$

Remarque

En réécrivant la variable $(X - E(X))^2$ sous la forme $(X - E(X))^2 = X^2 - 2E(X) * X + (E(X))^2$, on réécrit sa moyenne comme $E\left[(X - E(X))^2\right] = E(X^2) - 2E(X) * E(X) + (E(X))^2 = E(X^2) - E(X)^2$; d'où une autre formule de la variance

$$Var(X) = E(X^2) - E(X)^2$$

Propriétés de la Variance

Additivité

La variance d'une somme de variable X, Y est la somme des variance quand les deux variables sont indépendantes :

$$Var(X + Y) = Var(X) + Var(Y) \quad \underline{\text{si } X \text{ et } Y \text{ sont indépendantes}}$$

sinon, dans le cas général, la variance d'une somme égale :

$$Var(X + Y) = Var(X) + Var(Y) + 2Cov(X, Y)$$

- Notez que $Cov(X, \alpha) = 0$, $Var(X + \alpha) = Var(X)$ lorsque α est un scalaire. Par ailleurs, $Cov(X, X) = Var(X)$.

Multiplication par un scalaire

La variance d'une variable X multipliée par un scalaire est la multiplication de la variance par le carré du scalaire :

$$Var(\lambda\ X) = \lambda^2\ Var(X)$$

- Entrainez vous à démontrer ce résultat

Variances, avec des échantillons d'une variable iid

Supposons qu'on ait N variables iid, $X_1, \ldots, X_N$, c'est-à-dire, *indépendantes identiquement distribués*

Attention : $Var(NX) \neq Var(X_1 + X_2 + \cdots + X_N)$

En effet $\quad Var(NX) = N^2 \; Var(X)$

alors que $\quad Var(X_1 + X_2 + \cdots + X_N) = N \; Var(X)$

Ceci traduit le fait que la variable $X_1 + X_2 + \cdots + X_N$, qui mélange les occurences (a priori différentes) de N variables est beaucoup moins dispersée que la variable NX, qui multiplie par N l'occurence de la seule variable X.

Variance d'une moyenne empirique

La moyenne empirique : $\quad \bar{X} = \dfrac{X_1 + X_2 + \cdots + X_N}{N}$

sa variance : $\quad Var(\bar{X}) = \dfrac{Var(X)}{N}$

Ecart-type

L'écart type est la racine de la variance. On s'intéresse à la racine du moment d'ordre deux, afin d'avoir une mesure qui est comparable à la variable et en particulier aux paramètres de position.

Définition

L'écart-type est la racine de la variance

$$\sigma = \sqrt{Var(X)}$$

- **calculé à partir des données individuelles**

L'écart-type vérifie

$$\sigma^2 = \frac{1}{N}\sum(x_i - \bar{x})^2 = \frac{1}{N}\sum x_i^2 - \bar{x}^2$$

Exemple

Calculer l'écart interquartile et l'écart type du tableau de données individuelles suivant :

j	1	2	3	4	5
Xj	6	4	3	2	3

Ecart-type

$\bar{X} = (1/5) * (6 + 4 + 3 + 2 + 3) = 3.6$

$Var(X) = (1/5) * (6^2 + 4^2 + 3^2 + 2^2 + 3^2) - (3.$

$\sigma = \sqrt{1.84} = 1.36$

Ecart interquartile

modalites	2	3	4	6
f_i cumulées	.2	.6	.8	1

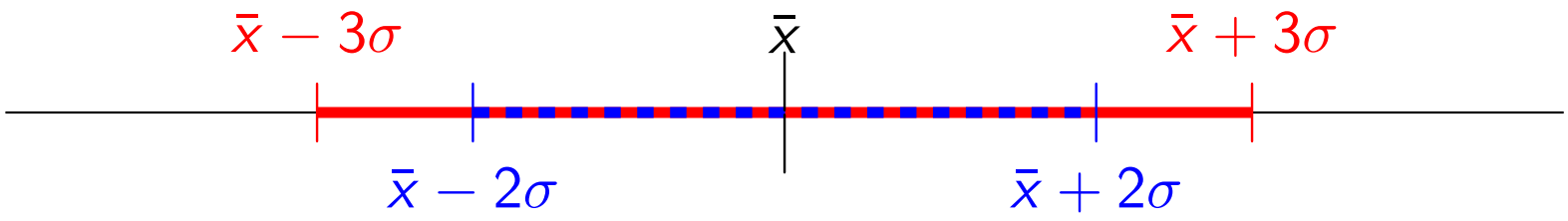
$Q_1 = 3,\ Q_2 = 3,\ Q_3 = 4$

écart interquartile=1

Intérêt pratique de l'écart type

Vrai quel que soit la loi

- au moins **75%** des valeurs se situent entre -2 et +2 écarts type de la moyenne
- au moins **89%** des valeurs se situent entre -3 et +3 écarts type de la moyenne (Chebychev' s inequality)



Seulement pour la loi normale

- **95%** des valeurs exactement se situent entre -2 et +2 écarts type de la moyenne
- **99%** des valeurs exactement se situent entre -3 et +3 écarts type de la moyenne.

Variance d'une variable regroupée par modalité

On suppose qu'il y a n_i occurences de la valeur x_i, pour $i = 1, \ldots, n$.

$$\sigma^2 = \frac{1}{n} \sum_i (x_i - \bar{x})^2 = \frac{\sum_i n_i x_i}{n} - (\bar{x})^2$$

ou encore

$$\sigma^2 = \sum_i f_i \, (x_i - \bar{x})^2 = \sum_i f_i x_i - (\bar{x})^2$$

Exemple

Calculer la variance et l'écart-type du tableau de fréquences :

x_i	0	1	2	3	4
f_i	.2	.1	.4	.2	.1

Ecart-type

$$\bar{E}(X) = 0,2 * 0 + 0,1 * 1 + 0,4 * 2 + 0,2 * 3 + 0,1 * 4) = 1.9$$

$$\begin{aligned} Var(X) &= 0,2 * 0^2 + 0,1 * 1^2 + 0,4 * 2^2 + 0,2 * 3^2 + 0,1 * 4^2 - (1.9)^2 \\ &= 5,1 - 3.61^2 = 1.49 \end{aligned}$$

$$\sigma = \sqrt{1.49} = 1.22$$

Variance d'une variable regroupée par classes

On suppose qu'il y a n_i occurences de la classe *de centre* c_i, pour $i = 1, \ldots, n$. La moyenne que l'on calcule est la moyenne des centres de classes.

$$\sigma^2 = \frac{1}{n} \sum_i (c_i - \bar{x})^2 = \frac{\sum_i c_i x_i}{n} - (\bar{x})^2$$

ou encore

$$\sigma^2 = \sum_i f_i (c_i - \bar{x})^2 = \sum_i f_i c_i - (\bar{x})^2$$

Exemple

Calculer la variance et l'écart-type du tableau de fréquences :

x_i	$]0, 100]$	$]100, 200]$	$]200, 300]$
f_i	.3	.5	.2

Ecart-type

$$\bar{E}(X) = 0,3 * 50 + 5 * 150 + 0,2 * 250) = 140$$

$$\begin{aligned} Var(X) &= 0,3 * 50^2 + 0,5 * 150^2 + 0,2 * 250^2 - (140)^2 \\ &= 24500?19600 = 4900 \end{aligned}$$

$$\sigma = \sqrt{4900} = 70$$

Coefficient de variation

Le coefficient de variation est une mesure relative de l'écart type qui permet de prendre en compte l'ordre de grandeur de la moyenne.

Définition

Le coefficient de variation est l'écart-type rapporté à la moyenne

$$C = \frac{\sigma}{E(X)}$$

- **Exemples**

$$\mu = 140,\ \sigma = 70. \quad C = 70/140 = 0,5$$

$$\mu = 1,9,\ \sigma = 1,22. \quad C = 1,22/1,9 = 0,64$$

Coefficient d'asymétrie ou Skewness

C'est un moment d'ordre 3.

Définition

Le coefficient d'asymétrie ou Skewness est le moment d'ordre 3 centré

$$\mu_3 = E\left[(X - E(X))^3\right]$$

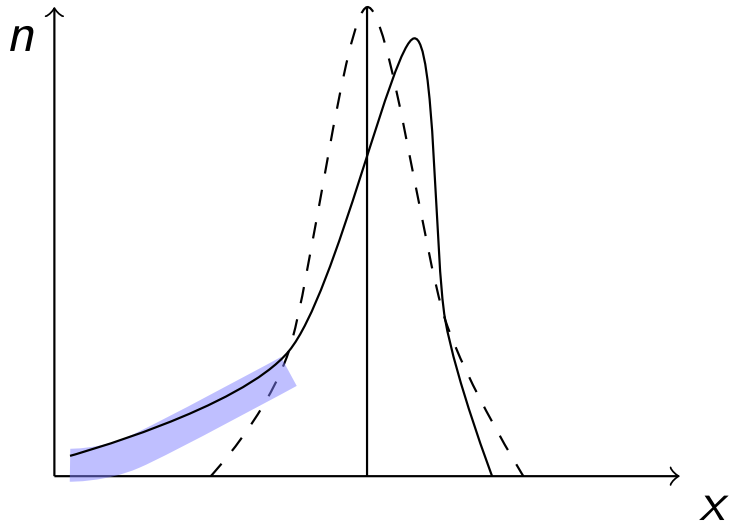
Définition

Le coefficient d'asymétrie ou Skewness de Fisher est relatif

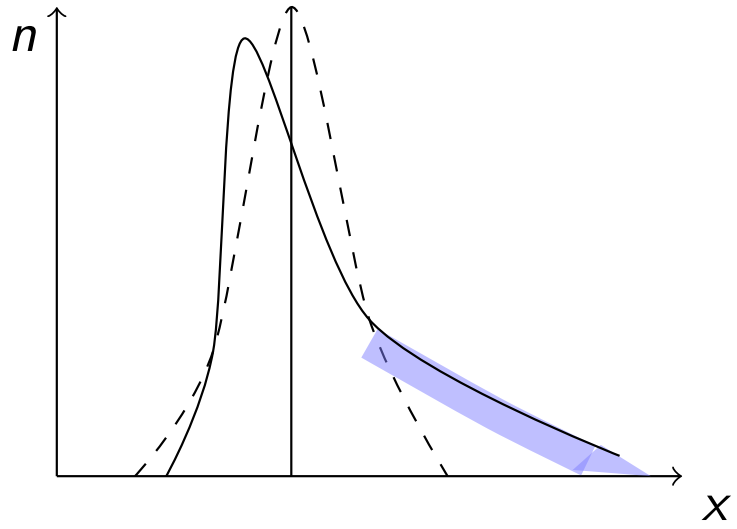
$$S = \frac{\mu_3}{\sigma^3}$$

Coefficient de Skewness et asymétrie

- Lorsque la distribution est symétrique, le coefficient de Skewness est nul.
- Lorsque la distribution possède une forte queue vers la droite, le coefficient de Skewness est positif (les + l'emportent).
- Lorsque la distribution possède une forte queue vers la gauche, le coefficient de Skewness est négatif (les - l'emportent).



Skewness à gauche, $S < 0$
[& mode à droite de la moyenne]



Skewness à droite, $S > 0$
[& mode à gauche de la moyenne]

Exemple

Montrer que la série suivante présente une queue de distribution vers la droite

Revenus	n_i	f_i
]0,100]	3	0.3
]100,200]	5	0.5
]200,300]	2	0.2
Total	10	1

- $\mu_3 = 0,3 * (50?140)^3 + 0,5 * (150?140)^3 + 0,2 * (250?140)^3 = 48000$
- $\sigma^3 = 70^3 = 343000$
- $S = 48000/343000 = 0,13$

$\Rightarrow$ S est positif, la série présente une queue de distribution vers la droite.

Coefficient d'aplatissement ou Kurtosis

C'est un moment d'ordre 4.

Définition

Le coefficient d'aplatissement ou Kurtosis est le moment centré d'ordre 4

$$\mu_4 = E\left[(X - E(X))^4\right]$$

Définition

Pearson a défini le coefficient d'aplatissement (Kurtosis) qui permet d'étudier la forme plus ou moins pointue ou aplatie :

$$K = \frac{\mu_4}{\sigma^4}$$

Fisher propose d'étudier $K' = K - 3$ ce qui permet de faire référence à une distribution particulière qui est la loi normale pour laquelle K vaut 3. Les logiciels statistiques vous donnent la valeur de K'.

Kurtosis et aplatissement

Le kurtosis donne une information sur les *QUEUES* de distribution. En effet, ce coefficient est grand quand il y a beaucoup de valeurs éloignées de la moyenne.

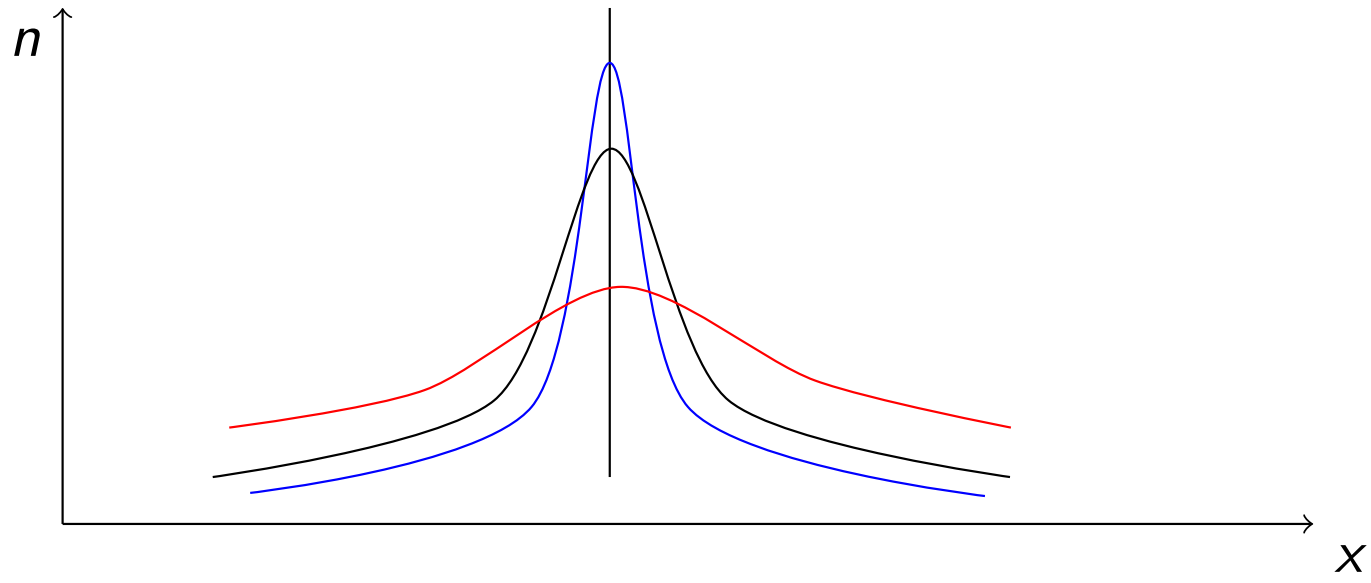


FIGURE: Un kurtosis positif ($K' > 0$) indique que les *queues* comptent *plus d'observations* que dans une distribution gaussienne. Un kurtosis négatif ($K' < 0$) indique que les *queues* comptent *moins d'observations* que dans une distribution gaussienne. Un kurtosis nul est celui d'une loi gaussienne

Exemple

Montrer que la série suivante est moins aplatie qu'une distribution normale, c'est à dire $K' < 0$.

Revenus	n_i	f_i
]0,100]	3	0.3
]100,200]	5	0.5
]200,300]	2	0.2
Total	10	1

- $\mu_4 = 0,3 * (50?140)^4 + 0,5 * (150?140)^4 + 0,2 * (250?140)^4 = 48970000$
- $\sigma^4 = 70^4 = 24010000$
- $K' = (48970000/24010000) - 3 = 2,039?3 = -0,96$

$\Rightarrow K'$ est négatif, les queues de distribution sont moins épaisses que les queues de la loi normale.

Petits Entraînements

Vous caractériserez les séries suivantes en utilisant les paramètres statistiques étudiés : écart interquartile, écart-type, coefficient de variation, S et K.

Nombre d'enfants dans un échantillon de familles.

Nb d'enfants	0	1	2	3	4	5	6 et +
Nb de familles	12	15	20	30	13	6	4

Entrées aux urgences d'un hôpital selon l'âge.

Classes	0-2	2-5	5-10	10-20	20-30	30-50
Effectifs	17	21	57	55	35	15

CA en milliers d'un produit selon le nom dudit produit.

Nom	MP-1	XP-2	ZP-3	RP-4
CA	170	210	300	150

Distribution des hôtels de Rouen selon le nombre d ?étoiles.

Nb *	*	**	***	****	*****
Effectifs	15	10	5	3	1

Entraînements sur les bases de données

1. Sur FichierExemple2, analyser la symétrie et l'aplatissement de la variable indemnisation totale. Commenter abondamment.