

Mesures de la dispersion

L'un des objectifs de cette leçon est de vous montrer que toute mesure comporte un degré de variation, que l'on pense par exemple, à la moyenne de vos notes, à la taille des étudiants de votre promotion...

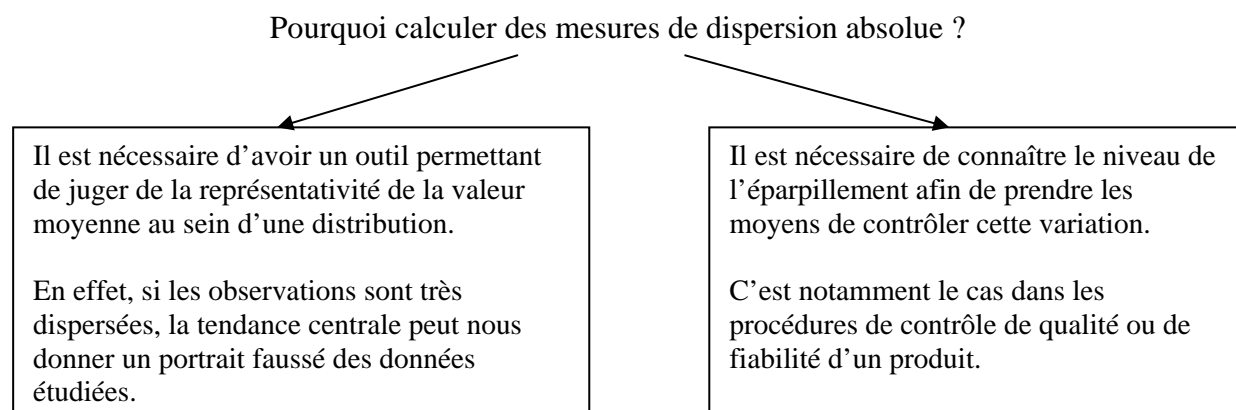
La variation fait partie de la réalité et vous devez savoir en rendre compte.

Les mesures de tendance centrale ne peuvent donc à elles seules résumer convenablement un ensemble de données.

Les mesures de dispersion nous indiquent à quel point les observations ont tendance à s'éloigner de la tendance centrale calculée.

1. Les mesures de dispersion absolue

Tout comme les mesures de tendance centrale, les mesures de dispersion absolue s'expriment dans les mêmes unités que les observations (par ex en € pour les salaires..)



Les 3 mesures de dispersion absolue le plus fréquemment utilisées sont l'étendue, l'écart moyen et l'écart type.

1.1 L'étendue

L'étendue est la plus simple et la plus sommaire des mesures de dispersion. Elle représente la différence entre la plus grande et la plus petite valeur d'une série d'observations, c'est-à-dire l'écart entre les valeurs extrêmes. L'étendue est donc très sensible aux valeurs extrêmes.

Cette caractéristique n'est pas définie pour des distributions groupées car les valeurs extrêmes ne sont pas toujours connues exactement après le regroupement en classes.

Cette mesure de dispersion s'avère utile pour exprimer les écarts de température, ou encore pour décrire les variations boursières sur une période donnée.

1.2 L'écart absolu moyen

Supposons qu'après son échec au bac, Léon décide de se changer les idées en partant en voyage organisé. Connaissant le directeur de l'agence, ce dernier indique à Léon que l'âge moyen des femmes célibataires inscrites au voyage A est de 19 ans, tandis que l'âge moyen de celles inscrites au voyage B est de 31 ans.

Sans hésiter, Léon s'inscrit au voyage A.

La distribution des âges des femmes célibataires inscrites à chaque voyage est donnée ci dessous.

Age des femmes allant au voyage A	n_i
2	3
4	1
5	1
7	1
10	1
11	2
34	1
35	2
50	1
58	1
Σ	14

L'âge moyen des femmes pour le voyage A est :
 $\bar{x}_A = 19$

Age des femmes allant au voyage B	n_i
18	1
19	5
20	2
45	2
46	1
47	1
48	1
50	1
Σ	14

L'âge moyen des femmes pour le voyage B est :
 $\bar{x}_B = 31$

Si Léon ne s'était pas uniquement fié à la moyenne, il aurait sûrement pris une autre décision. Ce qu'espérait Léon, en fait, c'était une moyenne d'âge de 19 ans et très peu de variation de l'âge des individus autour de cette valeur moyenne. Bref, Léon aurait préféré qu'une dispersion très petite soit associée à cette valeur moyenne de 19 ans.



L'écart moyen est une des mesures de dispersion qui aurait permis à Léon de mieux jauger la situation.

1.2.1 Définition

Il s'agit d'une moyenne arithmétique d'écarts par rapport à une valeur centrale qui peut être la moyenne arithmétique ou encore la médiane.

Pour calculer l'écart moyen, il faut donc :

- 1) calculer la moyenne arithmétique ou la médiane des observations
- 2) déterminer l'écart absolu – c'est-à-dire l'écart abstraction faite du signe algébrique – entre chaque observation et la mesure de tendance centrale retenue (\bar{x} ou M_e)
- 3) calculer la moyenne de ces écarts absolus

1.2.2 Ecart absolu moyen par rapport à \bar{x}

. Pour les séries statistiques :
$$\bar{e}_{\bar{x}} = \frac{1}{n} \sum_{i=1}^n |x_i - \bar{x}|$$

. Pour les distributions d'effectifs :
$$\bar{e}_{\bar{x}} = \frac{1}{n} \sum_{i=1}^p n_i |x_i - \bar{x}|$$

. Pour les distributions de fréquences :
$$\bar{e}_{\bar{x}} = \sum_{i=1}^p f_i |x_i - \bar{x}|$$

Il est impératif de prendre les valeurs absolues des écarts $|x_i - \bar{x}|$ car, par construction, la moyenne des écarts à la moyenne arithmétique est nulle (propriété 1 de la moyenne arithmétique).

1.2.2.1 Dans une distribution non groupée

L'exemple précédent montre que l'écart moyen de l'âge des femmes célibataires optant pour le voyage B est de 13,57 ans.

1.2.2.2 Dans une distribution groupée

Reprenons l'exemple de la répartition des employés d'une PME selon le salaire mensuel en € dont nous avons déjà calculé le salaire moyen : $\bar{x} = 1232,14$ €

Les observations étant groupées par classes, il nous faut adopter la convention du centre de classe pour calculer l'écart absolu moyen :

classes	n_i	Centres de classes x_i	Ecarts absolus $ x_i - \bar{x} $	$n_i x_i - \bar{x} $
[1000, 1250[64	1125	107,14	6856,96
[1250, 1500[7	1375	142,86	1000,02
[1500, 1750[10	1625	392,86	3928,6
[1750, 2000[3	1875	642,86	1928,58
Σ	84			13 714,16

D'où $\bar{e}_{\bar{x}} = 163,26$

1.2.3 Ecart absolu moyen par rapport à M_e

. Pour les séries statistiques :
$$\bar{e}_{M_e} = \frac{1}{n} \sum_{i=1}^n |x_i - M_e|$$

. Pour les distributions d'effectifs :
$$\bar{e}_{M_e} = \frac{1}{n} \sum_{i=1}^p n_i |x_i - M_e|$$

. Pour les distributions de fréquences :
$$\bar{e}_{M_e} = \sum_{i=1}^p f_i |x_i - M_e|$$

1.2.4 Propriétés

Contrairement à l'étendue, l'écart moyen prend en considération chacune des observations et nous indique la distance moyenne séparant les observations de la moyenne arithmétique, ou médiane, de celles-ci.

Il est relativement simple à calculer et à interpréter.

Malheureusement le fait que des valeurs absolues apparaissent dans la formule de l'écart moyen en limitera l'utilisation comme mesure de dispersion.

1.3 La variance et l'écart type

1.3.1 Définition

La variance $V(x)$ est la moyenne arithmétique des carrés des écarts à la moyenne arithmétique. Elle mesure la dispersion des observations autour de la moyenne de celles-ci.

.Pour les séries statistiques :
$$V(x) = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$$

.Pour les distributions d'effectifs :
$$V(x) = \frac{1}{n} \sum_{i=1}^p n_i (x_i - \bar{x})^2$$

.Pour les distributions de fréquences :
$$V(x) = \sum_{i=1}^p f_i (x_i - \bar{x})^2$$

La variance est une mesure statistique importante, mais elle possède un inconvénient majeur : elle ne s'exprime pas dans la même unité de mesure que les données originales. Pour obtenir une mesure de dispersion qui s'exprime dans les mêmes unités que les valeurs originales, il faut passer à l'écart type.

L'écart type (*ou écart quadratique moyen*) (*ou encore déviation standard*) est la racine carrée de la variance $V(x)$. C'est la mesure de dispersion la plus utilisée : il mesure aussi la dispersion des observations autour de la moyenne arithmétique et on le calcule en se basant sur les écarts existant entre chacune des observations et la valeur moyenne de celles-ci. Plus l'éparpillement des observations autour de la moyenne est important, plus l'écart est grand

.Pour les séries statistiques :

$$\sigma^2 = V(x) = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 \quad \text{et} \quad \sigma = \sqrt{V(x)} = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2}$$

.Pour les distributions d'effectifs :

$$\sigma^2 = V(x) = \frac{1}{n} \sum_{i=1}^p n_i (x_i - \bar{x})^2 \quad \text{et} \quad \sigma = \sqrt{V(x)} = \sqrt{\frac{1}{n} \sum_{i=1}^p n_i (x_i - \bar{x})^2}$$

.Pour les distributions de fréquences :

$$\sigma^2 = V(x) = \sum_{i=1}^p f_i (x_i - \bar{x})^2 \quad \text{et} \quad \sigma = \sqrt{V(x)} = \sqrt{\sum_{i=1}^p f_i (x_i - \bar{x})^2}$$

L'écart type est donc la moyenne quadratique des écarts à la moyenne arithmétique.

1.3.1.1 Dans une distribution non groupée

Reprenons le « cas » Léon pour lequel nous avons déjà calculé que l'âge moyen des femmes pour le voyage B est $\bar{x}_B = 31$, et construisons le tableau suivant :

Age des femmes allant au voyage B	n_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	$n_i (x_i - \bar{x})^2$
18	1	-13	169	169
19	5	-12	144	720
20	2	-11	121	242
45	2	14	196	392
46	1	15	225	225
47	1	16	256	256

48	1	17	289	289
50	1	19	361	361
Σ	14			2654

$$V(x) = \frac{1}{n} \sum_{i=1}^k n_i (x_i - \bar{x})^2 = 189,57 \quad \text{et} \quad \sigma(x) = \sqrt{189,57} = 13,8 \text{ ans}$$

1.3.1.2 Dans une distribution groupée

Dans ce cas, les formules de définition, comme celles de la moyenne ne peuvent pas être directement appliquées.

Par convention, on suppose que toutes les observations à l'intérieur d'une classe sont groupées en son centre. On retient donc comme variable statistique le centre de chaque classe

Classes	n_i	Centres de classe	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	$n_i (x_i - \bar{x})^2$
---------	-------	-------------------	-----------------	---------------------	-------------------------

1.3.2 Propriétés algébriques

P1 La variance et l'écart type sont nuls si et seulement si tous les écarts $(x_i - \bar{x})$ sont nuls, c'est-à-dire si toutes les valeurs observées sont égales entre elles et donc égales à leur moyenne : $x_1 = x_2 = \dots = \bar{x}$

P2 Formule développée

Le carré de l'écart type est égal à la moyenne des carrés moins le carré de la moyenne :

.Pour les séries statistiques : $\sigma^2 = V(x) = \frac{1}{n} \sum_{i=1}^n x_i^2 - \bar{x}^2$

.Pour les distributions d'effectifs : $\sigma^2 = V(x) = \frac{1}{n} \sum_{i=1}^p n_i x_i^2 - \bar{x}^2$

.Pour les distributions de fréquences : $\sigma^2 = V(x) = \sum_{i=1}^p f_i x_i^2 - \bar{x}^2$

Il s'agit en fait de la formule développée de la variance. Cette formule permet de simplifier les calculs lorsque le nombre d'observations est grand ;

Pour démontrer cette propriété, il suffit de développer la formule de la variance :

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

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

Sachant que

$$\frac{1}{n} \sum_i n_i x_i = \bar{x} \quad \text{et que} \quad \frac{1}{n} \sum_i n_i = 1,$$

on obtient :

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

P3 Variances intra et inter populations

Soit une population P d'effectif n composée de 2 sous populations



$$P_1, n_1, \bar{x}_1, v(x_1)$$

$$\text{avec } n_1 + n_2 = n$$

$$P_2, n_2, \bar{x}_2, v(x_2)$$

La variance de la population totale est égale à la moyenne pondérée des variances des sous populations augmentée de la variance des moyennes des différentes sous populations.

$$V(x) = \frac{1}{n} [n_1 v(x_1) + n_2 v(x_2)] + \frac{1}{n} [n_1 (\bar{x}_1 - \bar{x})^2 + n_2 (\bar{x}_2 - \bar{x})^2]$$

<div style="text-align: center;">  </div> <div style="border: 1px solid black; padding: 5px; text-align: center;"> Moyenne des variances Ou Variance intra population </div>	<div style="text-align: center;">  </div> <div style="border: 1px solid black; padding: 5px; text-align: center;"> Variance des moyennes Ou Variance inter population </div>
---	---

La variance intra population : $\frac{1}{n} [n_1 V(x_1) + n_2 V(x_2)]$ est la variance que l'on obtient si toutes les sous-populations ont la même moyenne.

La variance inter population est la variance que l'on obtient si les sous populations sont homogènes .

Généralisation : Soit une population de taille n qui se divise en K sous populations k de taille n_k , ayant chacune une moyenne \bar{x}_k et une variance σ_k^2 . On peut montrer que :

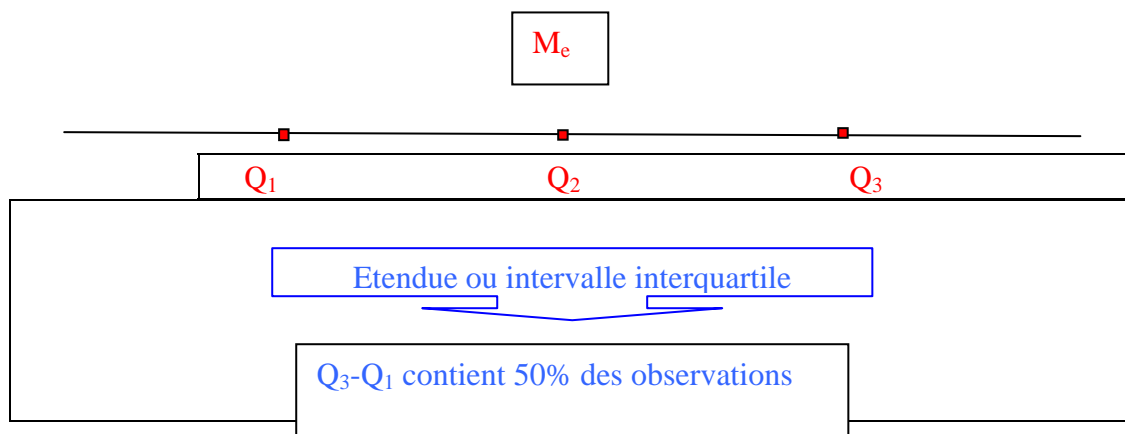
$$\sigma^2 = \frac{1}{n} \left[\sum_{k=1}^K n_k \sigma_k^2 \right] + \frac{1}{n} \left[\sum_{k=1}^K n_k (\bar{x}_k - \bar{x})^2 \right]$$

On peut donc décomposer une dispersion globale en calculant la part imputable aux dispersions internes (intra) et celle imputable à la dispersion des moyennes (inter)

1.4 L'écart interquartile

A l'instar de l'étendue, l'intervalle interquartile est une mesure de dispersion basée sur la distance entre 2 points déterminés. Pour l'étendue, ces 2 points étaient tout simplement la plus grande et la plus petite des valeurs observées. Dans le cas de l'intervalle interquartile, il nous faut calculer l'étendue interquartile, laquelle englobe approximativement 50 % des valeurs centrales de la distribution.

Ainsi nos points se situeront à la position occupée entre le premier (Q_1) et le troisième (Q_3) quartile.



Il est possible de calculer un intervalle semi-interquartile Q qui correspond simplement à la moitié de l'intervalle interquartile.

$$Q = \frac{Q_3 - Q_1}{2}$$

Plus la valeur de Q est petite, plus le degré de concentration de la moitié centrale des observations est élevé.

2. Les mesures de dispersion relative

2.1 Dispersion relative et dispersion absolue

L'écart type et les autres mesures de dispersion considérées jusqu'à présent sont des mesures de dispersion absolues. Cela signifie qu'elles s'expriment dans les mêmes unités que les observations originales.

Ainsi, la distribution des salaires mensuels des salariés d'une PME peut avoir un écart type de 208,76 € tandis que l'écart type de la distribution des salaires mensuels d'une petite entreprise anglaise est de 200 £.

Ou encore une distribution aura un écart type de 10 kg tandis qu'une autre aura un écart type de 10 ans.

S'il nous venait à l'esprit de comparer les dispersions de ces distributions,

pourrions nous conclure que la distribution ayant un écart type de 300 € est plus dispersée que celle ayant un écart type de 200 £ ?

de même, pourrions nous conclure que la dispersion ayant un écart type de 10kg a la même variabilité que celle ayant un écart type de 10 ans ?

Pouvons nous logiquement comparer des kg et des années ?

La réponse à ces questions est évidemment non : nous ne pouvons rien conclure en comparant simplement les mesures de dispersion absolue.

Afin d'effectuer des comparaisons, nous avons besoin d'une mesure du degré de dispersion relative au sein de la distribution étudiée.

2.2 Le coefficient de variation

Soit une entreprise qui possède 2 établissements :

l'établissement 1 se trouve en France : $\bar{x}_1 = 1500$ € et $\sigma_1 = 120$ €

l'établissement 2 se trouve aux Etats-Unis : $\bar{x}_2 = 800$ \$ et $\sigma_2 = 70$ \$

On vous demande de calculer la dispersion relative des salaires dans ces 2 établissements.

Comme vous l'avez constaté l'écart type et la moyenne s'expriment dans la même unité que la variable statistique ce qui pose un problème lorsqu'il s'agit de comparer les dispersions de distributions qui ne sont pas exprimées dans la même unité.
Dans ce cas, il faut utiliser le coefficient de variation.

2.2.1 Définition

Le coefficient de variation (*ou coefficient de variabilité*) est une caractéristique de dispersion relative définie comme le rapport de l'écart type à la moyenne.

$$CV = \frac{\sigma_x}{\bar{x}} \quad \text{ou} \quad c_v = \frac{\sigma_x}{\bar{x}}$$

C'est un nombre sans dimension, souvent exprimé en %.

Compte tenu de la propriété P1 de la variance, le coefficient de variation est nul ssi tous les écarts $(x_i - \bar{x})$ sont nuls, c'est-à-dire si toutes les valeurs observées sont égales entre elles et donc égales à leur moyenne : $x_1 = x_2 = \dots = \bar{x}$

2.2.2 Utilisation

Dans la mesure où il est indépendant des unités choisies, le coefficient de variation permet d'effectuer des comparaisons de séries très différentes : soit parce qu'elles ne sont pas exprimées dans la même unité, soit parce que leur moyennes sont très différentes.

3. Les moments

La moyenne arithmétique et la variance ne sont que 2 cas particuliers de valeurs qu'on appelle « moments de la distribution ».

3.1 Définition

On appelle moment d'ordre r par rapport à a :

a désigne l'origine du moment

r fait référence à l'ordre du moment

.Pour les séries statistiques : $aM_r = \frac{1}{n} \sum_{i=1}^n (x_i - a)^r$

.Pour les distributions d'effectifs : $aM_r = \frac{1}{n} \sum_{i=1}^p n_i (x_i - a)^r$

.Pour les distributions de fréquences : $aM_r = \frac{1}{n} \sum_{i=1}^p f_i (x_i - a)^r$

Les moments s'avèrent particulièrement utiles pour l'étude des caractéristiques de forme d'une distribution à 1 variable ainsi que pour l'étude des distributions à 2 variables.

Deux cas peuvent être distingués :

- les moments par rapports à l'origine (*ou moments simples*) (*ou encore moments non centrés*)
- les moments par rapport à la moyenne (*ou moments centrés*)

3.2 Les moments simples : m_r

Les moments simples, ou moments non centrés, sont des moments d'ordre r pour lesquels l'origine est égale à 0 ($a=0$) :

.Pour les séries statistiques : $m_r = \frac{1}{n} \sum_{i=1}^n x_i^r$

.Pour les distributions d'effectifs : $m_r = \frac{1}{n} \sum_{i=1}^p n_i x_i^r$

.Pour les distributions de fréquences : $m_r = \frac{1}{n} \sum_{i=1}^p f_i x_i^r$

Propriété:

P1 Le moment simple d'ordre 1 se confond avec la moyenne arithmétique.

$$\text{si } r=1 \quad m_1 = \frac{1}{n} \sum_{i=1}^p n_i x_i = \bar{x}$$

3.3 Les moments centrés : μ_r

Les moments centrés sont des moments d'ordre r pour lesquels l'origine est la moyenne arithmétique ($a = \bar{x}$) :

.Pour les séries statistiques : $\mu_r = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^r$

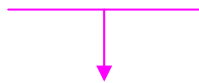
.Pour les distributions d'effectifs : $\mu_r = \frac{1}{n} \sum_{i=1}^p n_i (x_i - \bar{x})^r$

.Pour les distributions de fréquences : $\mu_r = \frac{1}{n} \sum_{i=1}^p f_i (x_i - \bar{x})^r$

Propriétés:

P1 : le moment centré d'ordre 1 est toujours nul

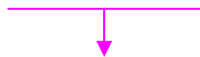
si $r=1$ $\mu_1 = \frac{1}{n} \sum_i n_i (x_i - \bar{x}) = 0$



On retrouve ici la première propriété de la moyenne arithmétique :
la somme des écarts entre valeurs observées et \bar{x} est nulle

P2 : le moment centré d'ordre 2 se confond avec la variance

si $r=2$ $\mu_2 = \frac{1}{n} \sum_i n_i (x_i - \bar{x})^2 = V(x)$



On retrouve ici la formule de définition de la variance