Statistiques

Activité 1 page 38

a) La moyenne vaut : $\frac{188}{15} \approx 12,53$.
Le mode est 13 ou 16 ; l'étendue est 9.

b) Premier tableau : par linéarité de la moyenne, on trouve : $\overline{x} \approx 13,53$.
Second tableau : $\overline{x} \approx 2 \times 12,53 \approx 25,06$

I Les indicateurs statistiques

1) Effectif

Définition :

L'effectif de la valeur d'un caractère est le nombre d'individus ayant cette valeur de caractère.

2) Fréquence

Définition :

La fréquence f d'une valeur d'un caractère est la proportion d'individus ayant cette valeur de caractère : $f = \frac{n}{N}$, où n est l'effectif de la valeur du caractère et N l'effectif total.

3) Étendue

Définition :

L'étendue d'une série statistique est la différence entre les valeurs extrêmes du caractère.

4) Mode

Définition :

Le mode d'une série statistique est la valeur du caractère ayant l'effectif le plus grand.

5) Moyenne (pondérée)

Définition :

Soit une série statistique dont les valeurs du caractère sont $x_1, x_2, \cdots, x_k$ et $n_1, n_2, \cdots, c_k$ effectifs associés.
La moyenne de la série statistique, notée $\bar{x}$, a pour valeur :

$$\bar{x} = \frac{n_1 x_1 + n_2 x_2 + \cdots + n_k x_k}{n_1 + n_2 + \cdots + n_k}$$

Conséquence : Lorsqu'on présente la série statistique en ne donnant que la liste des valeurs, alors la moyenne est

$$\frac{x_1 + x_2 + \cdots + x_k}{k}$$

Théorème : Si on appelle f_i la fréquence de la valeur x_i , alors :

$$\bar{x} = f_1 x_1 + f_2 x_2 + ... + f_k x_k.$$

La linéarité de la moyenne :

Théorème :
Soit k un nombre réel. Soit $x_1, x_2, \cdots, x_n$ les valeurs du caractère d'une série statistique et $\bar{x}$ leur moyenne.
Alors :
• la moyenne de la série $kx_1, kx_2, \cdots, kx_n$ est $k\bar{x}$;
• la moyenne de la série $x_1 + k, x_2 + k, \cdots, x_i + k, \cdots, x_n + k$ est $\bar{x} + k$.

Exemples :
• Si la moyenne au contrôle de biologie dans une classe est de 8 sur 20 et que le professeur décide d'augmenter toutes les notes de 10%, alors la nouvelle moyenne est de 8,8.
• Si la moyenne au contrôle d'histoire-géographie est de 8,7 sur 20 et que le professeur décide d'ajouter 1 point à tous les élèves, alors la nouvelle moyenne est de 9,7.

6) Médiane

Définition :

La médiane d'une série statistique est le nombre tel que :
50 % au moins des individus ont une valeur du caractère inférieure ou égale à ce nombre et 50 % au moins des individus ont une valeur supérieure ou égale à ce nombre.

Médiane d'un caractère quantitatif discret

On considère une série statistique dont les valeurs du caractère sont rangées par ordre croissant, chacune de ces valeurs figurant un nombre de fois égal à son effectif.
• Si le nombre de données est impair, donc de la forme $2n+1$, la médiane est le terme du milieu, c'est-à-dire le rang de terme $n+1$.
• Si le nombre de données est pair, donc de la forme $2n$, la médiane est la demi-somme des termes de rangs n et $n+1$.

7) Quartiles

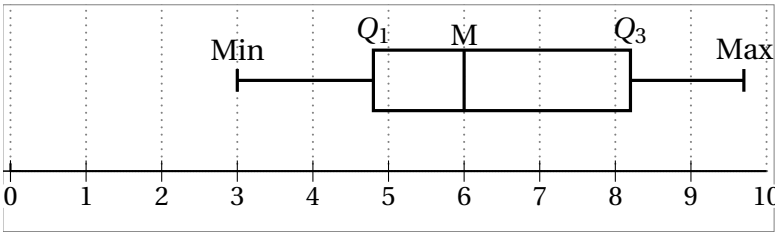
Définition :

Le premier quartile d'une série statistique, noté Q_1 est la première valeur de la série, rangée par ordre croissant, tel que 25 % des valeurs de la série soient inférieures ou égales à Q_1.
Le troisième quartile d'une série statistique, noté Q_3 est la première valeur de la série, rangée par ordre croissant, tel que 75 % des valeurs de la série soient inférieures ou égales à Q_3.
Remarque : Q_1 est la valeur x_i de la série dont l'indice i est le premier entier supérieur ou égal à $\frac{n}{4}$ (si n est l'effectif de la série).
Q_3 est la valeur x_i de la série dont l'indice i est le premier entier supérieur ou égal à $\frac{3n}{4}$ (si n est l'effectif de la série).

II Diagramme en boîte (ou diagramme de Tukey ou boîte à moustaches)

Les deux quartiles Q_1, Q_3, la médiane M d'une série statistique, associés aux valeurs extrêmes (minimum et maximum) permettent d'appréhender certaines caractéristiques de la répartition des valeurs.

Exemple :
Voici la série des températures (en degré Celcius) relevées sous abri à différents moments de la journée. Elles sont classées par ordre croissant.
3 ; 3,8 ; 4,5 ; 4,8 ; 5 ; 5,5 ; 5,7 ; 5,8 ; 6,2 ; 7 ; 7,3 ; 8,2 ; 9 ; 9,2 ; 9,5 ; 9,7
Les valeurs extrêmes sont 3 et 9,7.
La médiane vaut 6 (moyenne entre 5,8 et 6,2).
Le premier quartile est $Q_1 = 4,8$; le troisième quartile est $Q_3 = 8,2$. Le diagramme en boîte est alors :



Les diagrammes en boîte servent à faire des comparaisons de deux séries statistiques.

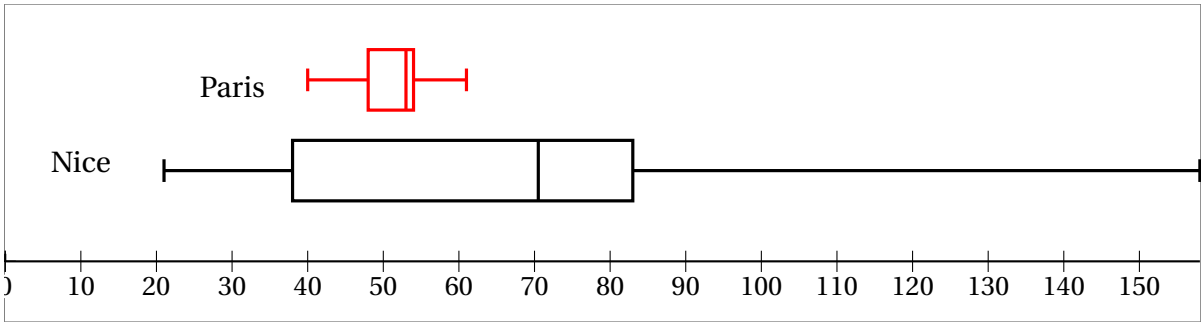
Exemple :
Les séries suivantes donnent les précipitations moyennes mensuelles en millimètres à Nice et à Paris :

	J	F	M	A	M	J	J	A	S	O	N	D
Nice	67	83	71	70	39	37	21	38	83	109	158	92
Paris	53	48	40	45	53	57	54	61	54	50	58	51

Pour effectuer la comparaison, on va ranger chaque série par ordre croissant :
Nice : 21 ; 37 ; 38 ; 39 ; 67 ; 70 ; 71 ; 83 ; 83 ; 92 ; 109 ; 158
Paris 40 ; 45 ; 48 ; 50 ; 51 ; 53 ; 53 ; 54 ; 54 ; 57 ; 58 ; 61

Pour Nice, on a : Min =21 ; Max = 158 ; $Q_1 = 38$; $M_1 = 70,5$ et $Q_3 = 83$
Pour Paris, on a : Min =40 ; Max = 61 ; $Q_1 = 48$; $M_1 = 53$ et $Q_3 = 54$

Diagrammes en boîtes :



Les précipitations sont plus régulières tout au long de l'année à Paris (série moins dispersée). LA totalité des valeurs de la série des précipitations à Paris est comprise entre le premier quartile et la médiane de la série des précipitations à Nice.
Pour la ville de Nice, plus de la moitié des mois ont des précipitations supérieures au maximum de Paris.

III Variance ; Écart type

1) Variance

Considérons deux groupes d'élèves, l'un de dix élèves et l'autre de huit élèves ; leurs notes de mathématiques à un contrôle sont :

Première série :

note x_i	1	2	3	17	20
effectif n_i	3	1	1	1	4

Deuxième série :

note x_i	8	10	11	12
effectif n_i	1	2	4	1

La moyenne de la première série est : $\dfrac{n_1x_1+\cdots+n_5x_5}{n_1+\cdots+n_5}=\dfrac{105}{10}=10,5$.

La moyenne de la deuxième série est : $\dfrac{84}{8}=10,5$.

Les deux moyennes sont égales ; pourtant, la répartition des notes n'est pas du tout la même.

Il faut donc trouver un moyen de mesurer la dispersion des nombres autour de la moyenne.

Un premier moyen est l'étendue, mais ce n'est pas très fiable.

Nous allons voir un deuxième moyen, qui est l'écart type.

Définition :

Soit une série statistique donnée par le tableau :

Valeur du caractère	x_1	x_2	$\cdots$	x_p	Total
Effectif	n_1	n_2	$\cdots$	n_p	N

La moyenne de cette série est : $\overline{x}=\dfrac{n_1x_1+n_2x_2+\cdots+n_Px_p}{N}$.

La **variance** est le nombre V défini par :

$$\mathbf{V=\frac{n_1(x_1-\overline{x})^2+n_2(x_2-\overline{x})^2+\cdots+n_p(x_p-\overline{x})^2}{N}}$$

V est donc la **moyenne des carrés des écarts entre chaque valeur x_i et la moyenne**.

Autre formulation de la variance :

Pour chaque indice i, on a : $(x_i-\overline{x})^2=x_i^2-2x_i\overline{x}+\overline{x}^2$.

En remplaçant dans le calcul de la variance chaque $(x_i-\overline{x})^2$ par ce que l'on vient de trouver, on obtient :

$$\begin{aligned}
V&=\frac{1}{N}\left[n_1\left(x_1^2-2x_1\overline{x}+\overline{x}^2\right)+n_2\left(x_2^2-2x_2\overline{x}+\overline{x}^2\right)+\cdots+n_p\left(x_p^2-2x_p\overline{x}+\overline{x}^2\right)\right]\\
&=\frac{1}{N}\left[n_1x_1^2+n_2x_2^2+\cdots+n_px_p^2-2n_1x_1\overline{x}-2n_2x_2\overline{x}-\cdots-n_px_p\overline{x}+n_1\overline{x}^2+n_2\overline{x}^2+\cdots+n_p\overline{x}^2\right]\\
&=\frac{1}{N}\left[n_1x_1^2+n_2x_2^2+\cdots+n_px_p^2-2\overline{x}(n_1x_1+n_2x_2+\cdots+n_px_p)+\overline{x}^2\left(n_1+n_2+\cdots+n_p\right)\right]\\
&=\frac{1}{N}\left[n_1x_1^2+n_2x_2^2+\cdots+n_px_p^2-2\overline{x}\times N\overline{x}+\overline{x}^2N\right]\\
&=\frac{1}{N}\left[n_1x_1^2+n_2x_2^2+\cdots+n_px_p^2-N\overline{x}^2\right]\\
&=\frac{n_1x_1^2+n_2x_2^2+\cdots+n_px_p^2}{N}-\overline{x}^2.
\end{aligned}$$

donc :

$$\boxed{\mathbf{V=\frac{n_1x_1^2+n_2x_2^2+\cdots+n_px_p^2}{N}-\overline{x}^2}}$$

Exemple : pour la deuxième série de notes :

note x_i	8	10	11	12
x_i^2	64	100	121	144
effectif n_i	1	2	4	1

$$V = \frac{(1\times 64)+(2\times 100)+(4\times 121)+(1\times 144)}{8} - 10,5^2 = \frac{892}{8} - 10,5^2 = 111,5 - 110,25 = 1,25.$$

2) Écart type

La variance est homogène aux carrés des valeurs de la série. Pour avoir une grandeur homogène aux valeurs de la série, on définit **l'écart type** de la série par : $\sigma = \sqrt{\mathbf{V}}$.
L'écart type est la racine carrée de la variance.

Exemple : pour la première série de notes, on a : $V = \frac{1905}{10} - 10,5^2 = 80,25$.
L'écart type de la première série est $\sigma = \sqrt{V} = \sqrt{80,25} \approx 8,96$.
Celui de la deuxième série est $\sigma = \sqrt{1,25} \approx 1,118$.

L'écart type de la première série est plus grand que celui de la deuxième série : les notes sont plus dispersées dans le premier cas que dans le second.

Exercices n^o 7 ; 8 et 9 page 60

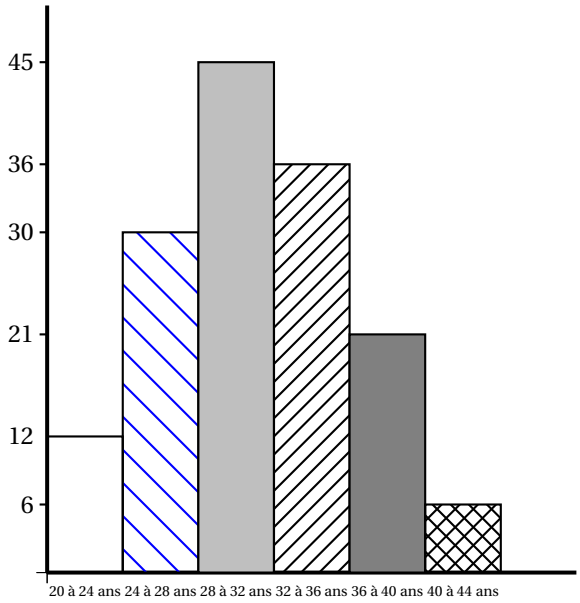
IV Histogrammes

Rappel : un histogramme est une représentation graphique sous la forme de rectangles, dont **l'aire** est proportionnelle aux effectifs.

Il y a deux cas possibles :

- Si le pas est constant (largeur des rectangles identiques), les aires des rectangles sont proportionnelles aux hauteurs des rectangles.

  Exemple : Brevet Polynésie juin 2007



(a) Compléter le tableau ci-dessous

Âge	$20 \leqslant$ âge < 24	$24 \leqslant$ âge < 28	$28 \leqslant$ âge < 32	$32 \leqslant$ âge < 36	$36 \leqslant$ âge < 40	$40 \leqslant$ âge < 44	Total
Centre de la classe	22						
Effectifs							
Fréquences en %							

(b) Quel est le pourcentage des employés qui ont strictement moins de 36 ans ?

(c) Calculer l'âge moyen d'un employé de cette entreprise.

- Si le pas n'est pas constant, les hauteurs des rectangles ne sont plus proportionnelles aux aires de ceux-ci. C'est le cas quant les données sont réparties par classes, avec des largeurs d'intervalles différentes.

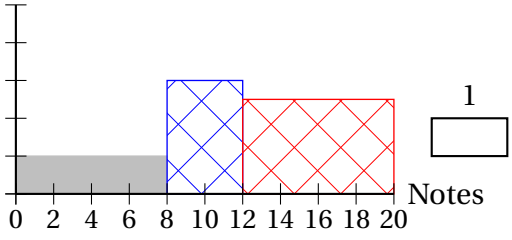
  Exemple :

  Considérons la répartiion des notes de 10 élèves.

Classe	[0 ; 8[	[8 ; 12[	[12 ; 20[
Effectif	2	3	5

On commence par choisir une unité sur l'axe des abscisses et une unité d'aire. Par exemple : 2 cm^2 pour un effectif de 1.

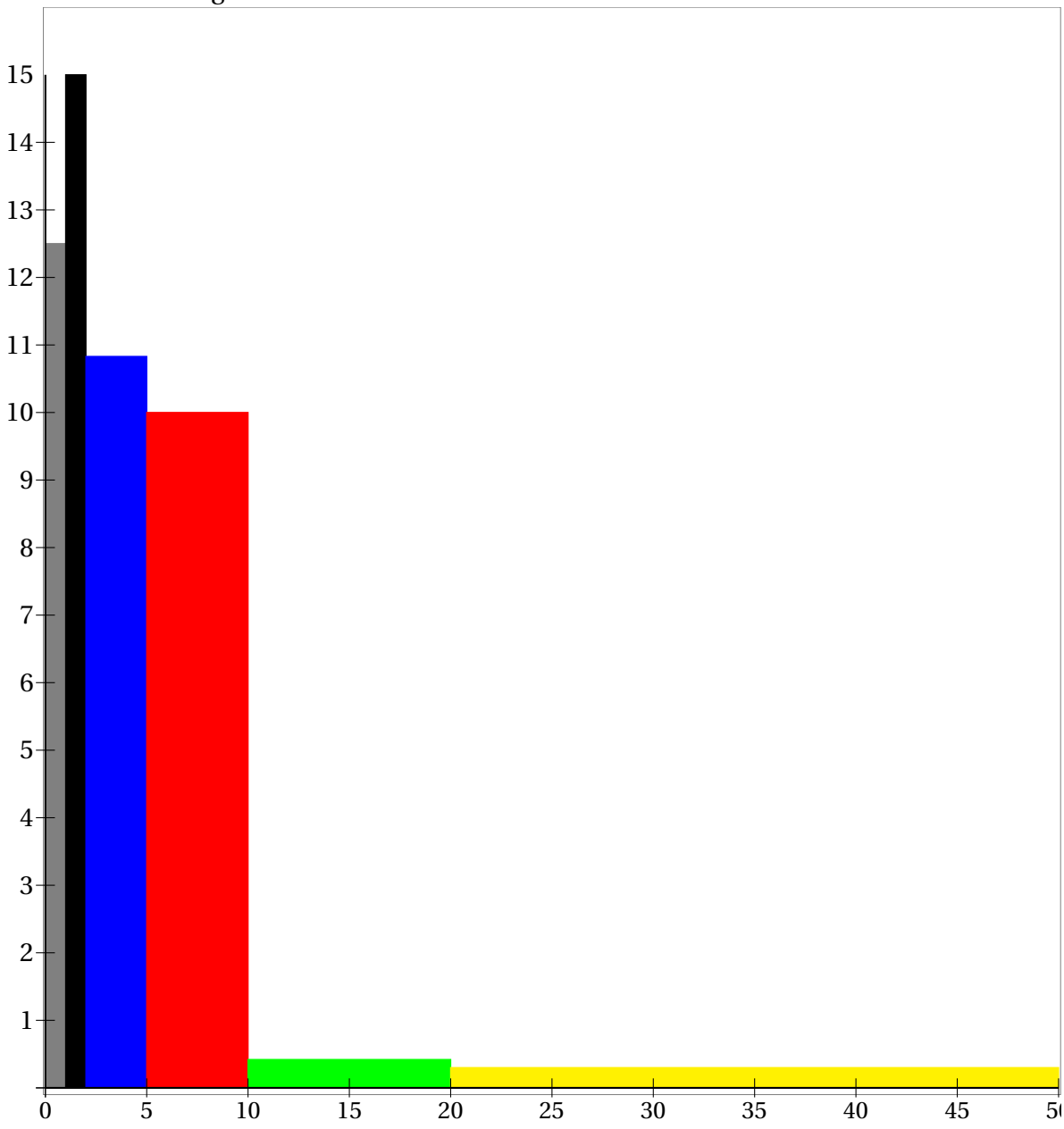
Classe	[0 ; 8[	[8 ; 12[	[12 ; 20[
Effectif	2	3	5
Aire en cm^2	4	6	10
Largeur en cm	4	2	4
Hauteur du rectangle	1	3	$\frac{5}{2} = 2,5$



Autre exemple : On a relevé les distances domicile-lieu de travail pour les salariés d'une entreprise.

Distance (en km)	[0 ; 1[	[1 ; 2[	[2 ; 5[	[5 ; 10[	[10 ; 20[	[20 ; 50[
Effectifs	30	36	78	120	10	24

Construire l'histogramme relatif à cette série.



Exercice :
Le tableau ci-dessous représente la répartition des durées de 70 films (rn min).

Durée en minutes	[100 ; 120[	[120 ; 160[	[160 ; 180[	[180 ; 260[
Effectifs	20	30	10	10

Représenter la situation par un histogramme.
On commencera à 100 sur l'axe des abscisses ; échelle : 1 cm pour 20 minutes en abscisses ; aire : 1 cm^2 pour 5 unités.

V Moyennes mobiles

Objectif : On s'intéresse à des séries chronologiques (dépendant du temps) et de mettre en évidence des tendances. Elle ssont plus ou moins faciles à mettre en œuvre, d'autant qu'il peut y avoir des fluctuations importantes (comme le cours de la bourse ...)

Pour adoucir les variations, on effectur un lissage par de smoyennes mobiles.

Exemple :
On a un tableau de valeurs :

Date	t_1	t_2	$\cdots$	t_i	t_{n-1}	t_n	
Valeur	x_1	x_2	$\cdots$	x_i	$\cdots$	x_{n+1}	x_n

Lisser une série chronologique par les moyennes mobiles d'ordre 3 consiste à remplacer les valeurs x_i $(2 \leqslant i \leqslant n-1)$ du premier tableau par les moyennes $y_i = \dfrac{x_{i-1} + x_i + x_{i+1}}{3}$.
On obtient ainsi le tableau :

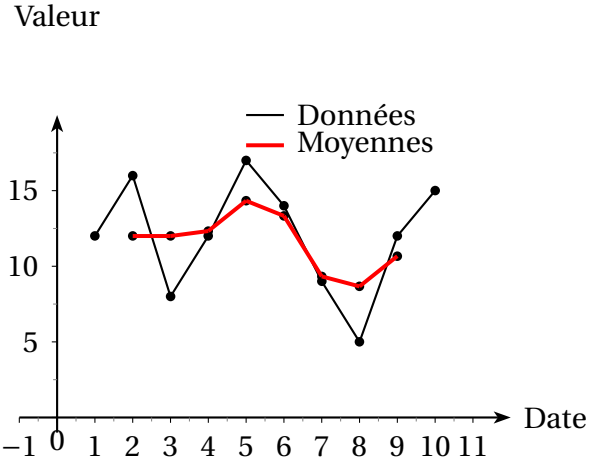
Date	t_1	t_2	$\cdots$	t_i	t_{n-1}	t_n	
Moyenne		y_2	$\cdots$	y_i	$\cdots$	y_{n+1}	

On perd les première et dernière valeurs.

Exemple ;
Le tableau suivant donne les relevés de températures effectués 10 jours consécutifs, en un même lieu.

Date	1	2	3	4	5	6	7	8	9	10
Température	12	16	8	12	17	14	9	5	12	15
Moyennes mobiles d'ordre 3		12	12	12,33	14,33	13,33	9,33	8,67	10,67	

On peut la représenter dans le même repère que la série non lissée :



Voir aussi l'exemple du livre pages 44-45

VI Effet de structure

En 2006, le personnel d'une petite entreprise était constitué par 3 cadres et par 7 ouvriers.
En 2007, la structure de l'entreprise change : il y a maintenant 2 cadres et 8 oucriers. Le salaire mensuel d'un cadre était alors de 10 000 € et celui d'un ouvrier de 1 000 € .
En 2007, le salaire mensuel d'un cadre est de 10 100 € et celui d'un ouvrier de 1 100 € .

De 2006 à 2007, le salaire moyen des cadres a augmenté car il est passé de 10000 à 10100 € . De même, le salaire moyen des ouvriers a augmenté car il est passé de 1 000 à 1 100 € .

Calculons alors les salaires moyens dans l'entreprise en 2000 et en 2001 :
En 2006, le salaire moyen valait : $\dfrac{3\times 10\ 000+7\times 1\ 000}{10}=3\ 700\,€$.
En 2007, le salaire moyen valait : $\dfrac{2\times 10\ 100+8\times 1\ 100}{10}=2\ 900\,€$.
Alors qu'on aurait pu s'attendre à une augmentation du salaire moyen, celui-ci diminue. Ce phénomène s'appelle **effet de structure**.

Autre exemple :
Comparons deux entreprises qui emploient toutes deux 1 000 ingénieurs qui se répartissent en quatre catégories.
Entreprise 1 :

Catégorie	1	2	3	4	Total
Effectif	100	200	300	400	
Salaire en euros	180	240	260	320	

Entreprise 2 :

Catégorie	1	2	3	4	Total
Effectif	80	190	350	380	
Salaire en euros	200	260	280	340	

Comparer alors les deux salaires moyens.