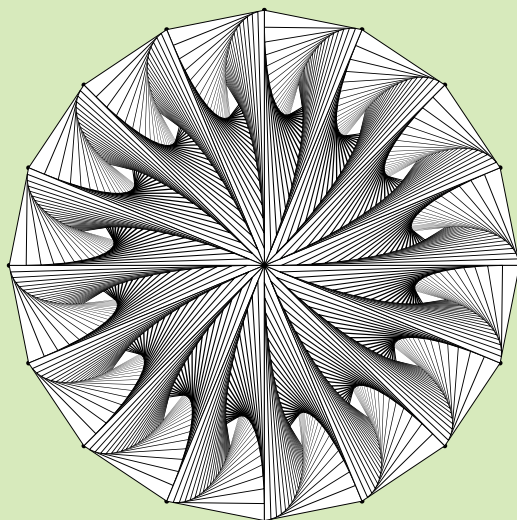


# Chapitre 11

# Statistiques



## Hors Sujet



**Titre :** « Gold »

**Auteur :** NEO RAUCH

**Présentation succincte de l'auteur :** Neo Rauch, né le 18 avril 1960 à Leipzig (RDA), est un artiste contemporain allemand, connu pour ses peintures monumentales influencées par les artistes surréalistes Giorgio de Chirico et René Magritte. Selon lui, ses tableaux sont dictés par des impératifs sur lesquels il n'a aucun contrôle, émergeant de l'« étrange zone d'obscurité située entre la raison et l'irrationalité dans laquelle l'artiste recherche une proie ».

Gold pourrait être considéré comme un fragment de rêve d'une société qui ne rêve plus en termes d'archétypes ou de symboles sexuels, mais sous forme d'images issues d'un dépotoir composé de débris d'icônes culturelles et des désirs de consommateurs.

Document réalisé à l'aide de  $\text{\LaTeX}$

Auteur : D.Zaccanaro

Site : [wicky-math.fr.nf](http://wicky-math.fr.nf)

Lycée : Jean Durand

## Table des matières

<b>I) Médiane et quartile</b>	<b>1</b>
I-1 Définitions et exemples . . . . .	1
I-2 Propriétés de la médiane et des quartiles . . . . .	3
I-3 Diagramme en boîte (ou boîte à moustache) . . . . .	4
I-4 Effet d'un changement affine . . . . .	5
<b>II) Moyenne, variance et écart type</b>	<b>6</b>
II-1 Définition . . . . .	7
II-2 Interprétation de l'écart-type . . . . .	7
II-3 Effet d'un changement affine . . . . .	8

## LEÇON 11

## Statistiques



## Résumé

En présence de données statistiques, on cherche à synthétiser les informations pour les rendre lisibles. Pour cela, on s'appuie sur quelques indicateurs numériques tel la moyenne et la médiane, ou encore les quartiles et l'écart-type. On commence à utiliser la moyenne en astronomie au XIV<sup>e</sup> siècle et seulement deux siècles plus tard pour l'écart type. Les propriétés des indicateurs numériques sont étudiées à partir de la deuxième moitié du siècle dernier, notamment par l'anglais Yule. L'usage des statistiques est très ancien (par exemple en Chine, 4000 ans avant JC) mais l'étude mathématique de cet outil est elle très récente.

## I) Médiane et quartile

## I-1 Définitions et exemples

**Définition 1 :**

On appelle médiane tout réel  $m_e$  tel que :

au moins 50% des termes de la série ont une valeur inférieure ou égal à  $m_e$

et

au moins 50% des termes de la série ont une valeur supérieure ou égal à  $m_e$

**Exemple :**

Un élève a obtenu les 10 notes suivantes :

$$x_1 = 4 \quad x_2 = 6 \quad x_3 = 7 \quad x_4 = 10 \quad x_5 = 11 \quad x_6 = 14 \quad x_7 = 14 \quad x_8 = 15 \quad x_9 = 16 \quad x_{10} = 19$$

On a alors, par exemple  $m_e = 12,5$

Supposons que son professeur, dans un jour de grande bonté, supprime sa plus faible note. L'élève n'a donc plus que 9 notes et la médiane devient alors :

$$m_e = 14$$

**Remarque :** On constate que la détermination de la médiane est différente suivant que l'effectif total  $N$  est pair ou impair :

- Lorsque l'effectif total  $N$  est impair, il n'y a pas de difficulté, la médiane  $m_e$  est le terme central, à savoir le terme de rang  $\frac{N+1}{2}$ . On a donc :  $m_e = x_{\frac{N+1}{2}}$ .
- Lorsque l'effectif total  $N$  est pair, l'usage veut que l'on choisisse pour médiane  $m_e$  la moyenne des deux termes centraux, à savoir : les termes de rang  $\frac{N}{2}$  et  $\frac{N}{2} + 1$ . On a donc :

$$m_e = \frac{x_{\frac{N}{2}} + x_{\frac{N}{2}+1}}{2}$$

Cependant, tout réel de l'intervalle  $\left[ x_{\frac{N}{2}}; x_{\frac{N}{2}+1} \right]$  conviendrait également. (En effet, dans certaines situations, la moyenne des deux termes centraux, qui n'est pas une valeur de la série, n'a pas de sens : par exemple, quel est le jour médian du mois de juin ? Le mois de juin comporte 30 jours. Les deux termes centraux sont 15 et 16 (15ème jour et 16ème jour). Dire que "le jour médian est le 15,5ème" n'a pas de sens. Mieux vaut dire (dans ce type de situation) : "le jour médian est le 15ème jour" ou "le jour médian est le 16ème jour" (au choix!) ...)

### Exercice 1 :

Quelle est la médiane de la série suivante :

$$x_1 = 1 \quad x_2 = 1 \quad x_3 = 1 \quad x_4 = 1$$



### Définition 2 :

On appelle premier quartile tout réel  $Q_1$  tel que :

au moins 25% des termes de la série ont une valeur inférieure ou égal à  $Q_1$

et

au moins 75% des termes de la série ont une valeur supérieure ou égal à  $Q_1$

On appelle troisième quartile tout réel  $Q_3$  tel que :

au moins 75% des termes de la série ont une valeur inférieure ou égal à  $Q_3$

et

au moins 25% des termes de la série ont une valeur supérieure ou égal à  $Q_3$

### Remarque :

- Le deuxième quartile  $Q_2 = m_e$
- Les trois quartiles partagent l'ensemble des valeurs en quatre sous ensembles de (presque) même effectif.
- On a toujours  $Q_1 \leq m_e \leq Q_3$



### Exemple :

En reprenant les 10 notes de l'élève fictif du premier exemple, on obtient  $Q_1 = 7$  et  $Q_3 = 15$

Cas d'une série statistique (discrète ou continue) avec regroupement en classes

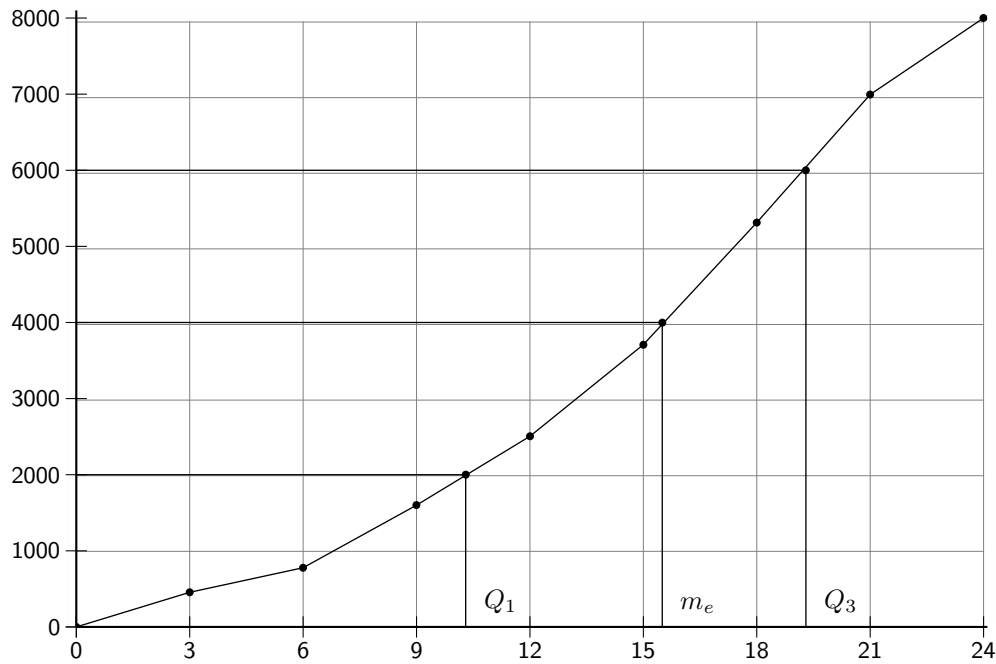


### Exemple :

La répartition des accidents corporels de la route selon les heures de la journée est décrite par le tableau suivant, pour l'année 1999. Répartitions des accidents corporels de la route

Tranche horaire	[0; 3[	[3; 6[	[6; 9[	[9; 12[	[12; 15[	[15; 18[	[18; 21[	[21; 0[	Total
Nombre d'accidents	4550	3230	8220	9050	12040	16040	16820	10050	80000
Effectifs cumulés croissants	4550	7780	16000	25050	37090	53130	69950	80000	

On trace ensuite le polygone des effectifs cumulés croissants :



**Remarque :**

- Une simple lecture graphique donne souvent une précision satisfaisante.
- Si on construit le polygone des fréquences cumulées croissantes alors  $Q_1$ ,  $m_e$  et  $Q_3$  sont les antécédents respectifs de 0,25 ; 0,5 et 0,75.
- Dans le cas d'un regroupement en classe, les statisticiens parlent rarement de valeur médiane mais plutôt de classe médiane.

**I-2 Propriétés de la médiane et des quartiles**

**Propriété 1 :**

Soient  $N \geq 5$  et  $(x_i)$  une famille de réels ordonnés dans l'ordre croissant.  
 Soient  $Q_1$ ,  $Q_3$  et  $m_e$  les quartiles et la médiane de la série  $(x_i)$ .  
 Soit  $m$  et  $M$  le minimum et le maximum de la série  $(x_i)$ .  
 Si l'on remplace  $m$  par un réel de  $]-\infty; Q_1[$  ou  $M$  par un réel de  $]Q_3; +\infty[$  alors les quartiles restent inchangés.  
 Si l'on remplace  $m$  par un réel de  $]-\infty; m_e[$  ou  $M$  par un réel de  $]m_e; +\infty[$  alors la médiane reste inchangée.

**Exemple :**

Considérons la série suivante :

$$x_1 = 1 \quad x_2 = 5 \quad x_3 = 8 \quad x_4 = 15 \quad x_5 = 29 \quad x_6 = 35$$

On a :  $Q_1 = x_2 = 5$  ;  $m_e = \frac{1}{2}(x_3 + x_4) = 11,5$  ;  $Q_3 = x_5 = 29$ .

Si l'on remplace  $m = x_1 = 1$  par un réel de  $]-\infty; 5[$ , cela ne changera pas les valeurs de  $Q_1$  ;  $m_e$  et  $Q_3$ . (Même si la série est à réordonner) Par contre, si l'on remplace  $m$  par un réel supérieur à  $Q_1$ , par exemple par 9. En réordonnant la série, on obtient :

$$y_1 = 5 \quad y_2 = 8 \quad y_3 = 9 \quad y_4 = 15 \quad y_5 = 29 \quad y_6 = 35$$

On constate que  $Q_1$  devient égal à  $y_2 = 8$  et  $m_e$  devient égal à  $\frac{1}{2}(y_3 + y_4) = 12$ .

**Remarque :** On dit parfois que la médiane et les quartiles sont insensibles aux termes extrêmes.

 **Preuve**

En remplaçant  $x_1$  par un réel de  $] -\infty; Q_1[$ , on ne change pas le nombre de termes de la série qui ont une valeur inférieure ou égale à  $Q_1$  (il y en aura donc toujours au moins 25%) ni le nombre de termes de la série qui ont une valeur supérieure ou égale à  $Q_1$  (il y en aura donc toujours au moins 75%). Donc  $Q_1$  reste une valeur convenable du premier quartile de la série. Même raisonnement pour le reste...

### I-3 Diagramme en boîte (ou boîte à moustache)


**Définition 3 :**

Soit  $(x_i)$  une famille de nombre réels ordonnés dans l'ordre croissant, notons  $x_N$  la valeur maximale de la famille. Soit  $m_e$ ,  $Q_1$  et  $Q_3$  la médiane, le premier quartile et le troisième quartile de cette série.

1. On appelle étendue la différence  $x_N - x_1$ . (Différence entre les termes extrêmes de la série)
2. On appelle interquartile la différence  $Q_3 - Q_1$ .
3. On appelle intervalle interquartile l'intervalle  $[Q_1; Q_3]$ .

**Remarque :** l'interquartile est un indicateur de dispersion (au même titre que l'étendue ou l'écart-type). Son avantage est qu'il ne tient compte que de 50% de la population, ce qui a pour effet d'ignorer les valeurs extrêmes souvent marginales. Il est donc assez utilisé car considéré comme "standard".

 **Exemple :**

En reprenant les notes de l'élève fictif initial, on obtient  $19 - 4 = 15$  comme étendue,  $15 - 7 = 8$  comme interquartile et  $[7; 15]$  comme intervalle interquartile.

On résume toutes ces données dans ce qu'on appelle une boîte à moustache, pour mieux comprendre observons l'exemple suivant :

Les séries suivantes donnent les précipitations moyennes mensuelles en millimètres à Nice et à Paris. Tableau des précipitations à Nice et Paris

Mois	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

On range les valeurs de chaque série par ordre croissant :

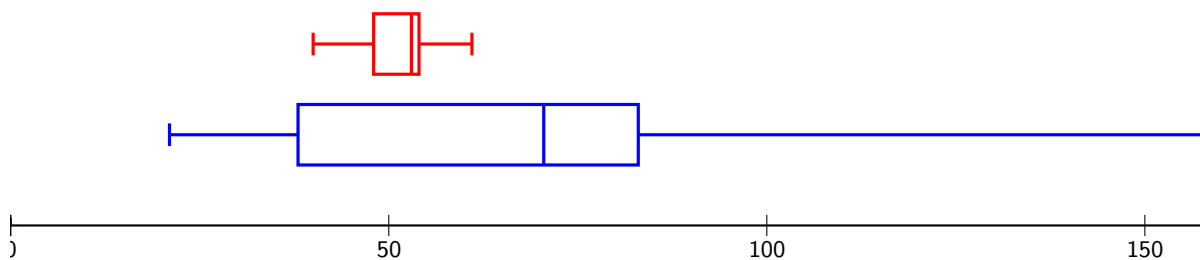
Nice : 21 - 37 - 38 - 39 - 67 - 70 - 71 - 83 - 83 - 92 - 109 - 158

Paris : 40 - 45 - 48 - 50 - 51 - 53 - 53 - 54 - 54 - 57 - 58 - 61

Par conséquent, pour Nice on obtient :  $Q_1 = 38$ ;  $m_e = 70,5$  et  $Q_3 = 83$

Et pour Paris on obtient :  $Q_1 = 48$ ;  $m_e = 53$  et  $Q_3 = 54$

Voici la boîte à moustache résumant ces valeurs pour Nice (en bleu) et pour Paris (en rouge) :



#### I-4 Effet d'un changement affine



##### **Théorème 1 :**

Soit  $N \in \mathbb{N}^*$  Soit  $(x_i)_{1 \leq i \leq N}$  une famille de réels ordonnés dans l'ordre croissant de médiane  $m_e$  et de quartiles  $Q_1$  et  $Q_3$ . Soient  $a \in \mathbb{R}^*$  et  $b \in \mathbb{R}$ . Soit  $(y_i)_{1 \leq i \leq N}$  la famille de réels définis par :  $y_i = ax_i + b$  pour tout  $i$  compris entre 1 et  $N$ .

Si  $a > 0$  alors la famille  $(y_i)_{1 \leq i \leq N}$  est ordonnée dans l'ordre croissant. Les réels suivants :

$$m'_e = am_e + b \quad Q'_1 = aQ_1 + b \quad Q'_3 = aQ_3 + b$$

sont des valeurs convenables de la médiane et des quartiles de la famille  $(y_i)_{1 \leq i \leq N}$ . Si  $a < 0$  alors la famille  $(y_i)_{1 \leq i \leq N}$  est ordonnée dans l'ordre décroissant Les réels suivants :

$$m'_e = am_e + b \quad Q'_1 = aQ_1 + b \quad Q'_3 = aQ_3 + b$$

sont des valeurs convenables de la médiane et des quartiles de la famille  $(y_i)_{1 \leq i \leq N}$ .



##### **Preuve**

Notons  $f$  la fonction définie par  $f(x) = ax + b$ , on a dans ce cas :

$$f(x_i) = ax_i + b = y_i$$

Si  $a > 0$  alors  $f$  est croissante et donc  $x_i \leq x_{i+1} \implies y_i \leq y_{i+1}$ , par conséquent dans la liste des termes de la série  $S_2$ ,  $y_i$  occupe le même rang que  $x_i$  dans la liste des termes de  $S_1$ . Le résultat découle alors directement de la définition de la médiane et des premier et troisième quartiles.

Lorsque  $a < 0$ , la fonction affine  $f : x \mapsto ax + b$  est décroissante. On a alors :

$$x_i \leq Q_3 \iff f(x_i) \geq f(Q_3) \iff ax_i + b \geq aQ_3 + b \iff y_i \geq aQ_3 + b$$


Donc :

$$\{i \in \llbracket 1, N \rrbracket \text{ tels que } x_i \leq Q_3\} = \{i \in \llbracket 1, N \rrbracket \text{ tels que } y_i \geq aQ_3 + b\}$$

Comme il y a 75% au moins de termes de la série  $S_1$  qui sont inférieurs à  $Q_3$ , on en déduit que :

$$Q'_1 = aQ_3 + b$$

On procède de manière analogue pour  $m'_e$  et  $Q'_3$

 **Exemple :**

Dans une entreprise les salaires sont résumés par :

	Moyenne	Médiane	$Q_1$	$Q_3$	Minimum	Maximum
Salaire (en €)	1450	1400	1200	1800	1020	3800

Le conseil d'administration décide d'une augmentation des salaires de 2% auquel s'ajoute encore une indemnité de 10€.

Cela se traduit par la transformation affine  $f$  définie par :  $f(x) = 1,02x + 10$ . ( $I_{cia} > 0$ )

Cela donne :  $f(m) = 1050,4$ ;  $f(M) = 3886$  pour le minimum et le maximum.


D'après le théorème, cela donne :  $f(Q_1) = 1234$ ;  $f(m_e) = 1438$  et  $f(Q_3) = 1846$ . Enfin, la nouvelle moyenne est donnée par  $f(\text{moyenne})$ . En effet :

Notons  $(x_i)_{1 \leq i \leq N}$  la série des salaires initiaux et posons  $y_i = f(x_i)$ , pour  $i \in \llbracket 1, N \rrbracket$ . La série  $(y_i)_{1 \leq i \leq N}$  correspond aux nouveaux salaires. La moyenne  $y$  des nouveaux salaires est :

$$y = \frac{1}{N} \sum_{i=1}^N y_i = \frac{1}{N} \sum_{i=1}^N (ax_i + b) = a \text{moyenne} + b = f(\text{moyenne})$$

Cela donne  $y = 1489$ . D'où le nouveau tableau :

	Moyenne	Médiane	$Q_1$	$Q_3$	Minimum	Maximum
Salaire (en €)	1489	1438	1234	1846	1050,4	3886

 **Exercice 2 :**

Une enquête portant sur l'argent de poche mensuel des adolescents en France et au Japon a donné les résultats suivants :

	Moyenne	Médiane	$Q_1$	$Q_3$	Minimum	Maximum
France (en €)	30	27,5	22,5	36,5	15	75
Japon (en yens)	3000	2570	2100	3500	1500	8000

Sur un même graphique, dresser les diagrammes en boîte de ces deux séries. (A l'époque de l'enquête il fallait 97 yens pour 1 euro).

## II) Moyenne, variance et écart type

Dans ce paragraphe, nous utiliserons une nouvelle notation. Soit  $(z_i)_{1 \leq i \leq N}$  une série statistique. Certains de ces réels peuvent être confondus. Notons  $p$  le nombre de valeurs de la série ( $1 \leq p \leq N$ ) et, pour tout  $i \in \llbracket 1, p \rrbracket$  notons  $x_i$  ces valeurs et  $n_i$  l'effectif de  $x_i$ . On notera  $(x_i, n_i)_{1 \leq i \leq p}$  la série statistique ainsi obtenue où les  $x_i$  sont distincts deux à deux.



## II-1 Définition



### Définition 4 :

La moyenne d'une série statistique  $(x_i, n_i)_{1 \leq i \leq p}$  est le nombre  $\bar{x}$  défini par :

$$\bar{x} = \frac{1}{N} \sum_{i=1}^{i=p} n_i x_i \quad \text{où } N \text{ désigne l'effectif total}$$

La variance d'une série statistique  $(x_i, n_i)_{1 \leq i \leq p}$  est le nombre noté  $V$  et défini par :

$$\frac{1}{N} \sum_{i=1}^{i=p} n_i (x_i - \bar{x})^2$$

la variance est la moyenne des carrés des écarts à la moyenne

L'écart-type d'une série statistique  $(x_i, n_i)_{1 \leq i \leq p}$  est le nombre noté  $\sigma$  et défini par :

$$\sigma = \sqrt{V}$$

### Remarque :

- La variance est une somme de carrés. C'est donc une quantité positive. L'écart-type est donc bien défini. Et il s'exprime dans la même unité que la caractère étudié.
- Si on note  $f_i = \frac{x_i}{n_i}$  les fréquences des  $x_i$  les formules deviennent  $\bar{x} = \sum_{i=1}^{i=p} f_i x_i$  et  $V = \sum_{i=1}^{i=p} f_i (x_i - \bar{x})^2$
- Dans le cas d'un regroupement en classe, les calculs sont effectués en choisissant  $x_i$  au centre de chaque classe (c'est l'hypothèse de répartition uniforme de chaque classe)

Pour calculer la variance, on dispose d'une formule un peu plus pratique :



### Théorème 2 :

La variance d'une série statistique  $(x_i, n_i)_{1 \leq i \leq p}$  peut se calculer avec la relation suivante :

$$V = \frac{1}{N} \sum_{i=1}^{i=p} n_i x_i^2 - \bar{x}^2$$



### Preuve

$$V = \sum_{i=1}^{i=p} f_i (x_i - \bar{x})^2 = \sum_{i=1}^{i=p} f_i (x_i^2 - 2x_i \bar{x} + \bar{x}^2) = \sum_{i=1}^{i=p} f_i x_i^2 - 2\bar{x} \sum_{i=1}^{i=p} f_i x_i + \bar{x}^2 = \sum_{i=1}^{i=p} f_i x_i^2 - 2\bar{x}^2 + \bar{x}^2 = \sum_{i=1}^{i=p} f_i x_i^2 + \bar{x}^2$$

## II-2 Interprétation de l'écart-type

La variance est la moyenne des carrés des écarts à la moyenne. Elle mesure donc la dispersion des valeurs autour de la moyenne. Elle n'est pas très parlante car elle s'exprime dans le carré de l'unité du caractère.

L'écart-type a l'avantage de s'exprimer dans la même unité que le caractère.

L'écart-type permet de comparer la dispersion de deux séries. Contrairement à l'interquartile, il tient compte de l'ensemble de la population.

 **Exercice 3 :**

L'élève A a obtenu les dix notes suivantes :


10    15    16    13    8    11    12    12    13    15

L'élève B a obtenu les dix notes suivantes :

11    9    9    10    15    7    12    12    14    13

Calculer les moyennes de A et B. Quel est l'élève qui a les résultats les plus homogènes ?

**II-3 Effet d'un changement affine**

 **Théorème 3 :**  
 Soit  $(x_i, n_i)_{1 \leq i \leq p}$  une série statistique de variance  $V$  et d'écart-type  $\sigma$ .  
 Soient  $a \in \mathbb{R}^*$  et  $b \in \mathbb{R}$ .  
 Soit  $(y_i, n_i)_{1 \leq i \leq p}$  la série statistique définie par  $y_i = ax_i + b$ , pour tout  $i \in \llbracket 1, p \rrbracket$ .  
 Notons  $V'$  sa variance et  $\sigma'$  son écart-type.  
 Alors :  $V' = a^2V$  et  $\sigma' = |a|\sigma$

 **Preuve**

On rappelle que

$$\bar{y} = a\bar{x} + b$$

Par conséquent :

$$V' = \frac{1}{N} \sum_{i=1}^{i=p} n_i (y_i - \bar{y})^2 = \frac{1}{N} \sum_{i=1}^{i=p} n_i (ax_i + b - a\bar{x} - b)^2 = \frac{1}{N} \sum_{i=1}^{i=p} n_i (ax_i - a\bar{x})^2 = a^2 \times \frac{1}{N} \sum_{i=1}^{i=p} n_i (x_i - \bar{x})^2 = a^2V$$

De plus

$$\sigma' = \sqrt{V'} = \sqrt{a^2V} = |a| \sqrt{V} = |a| \sigma$$

 **Exercice 4 :**

1. Soit  $(x_i, n_i)_{1 \leq i \leq p}$  une série statistique de moyenne  $\bar{x}$  et d'écart-type  $\sigma$ . Trouver deux réels  $a$  positif et  $b$ , de telle sorte, qu'après une transformation affine  $y = ax + b$ , la nouvelle série  $(y_i, n_i)_{1 \leq i \leq p}$  vérifie  $\bar{y} = 0$  et  $\sigma' = 1$ . On dit que la nouvelle série obtenue est centrée et réduite.
2. On donne les deux séries A et B suivantes :

	Moyenne	Ecart-type	$Q_1$	$Q_3$	Médiane	Minimum	Maximum
Série A	0,314	0,015	0,303	0,326	0,314	0,27	0,352
Série B	0,21	0,014	0,2	0,22	0,21	0,171	0,247

Trouver les nouveaux paramètres des séries A' et B' centrées et réduites correspondant à A et B