1. Statistiques descriptives univariées et bivariées

Ne dis pas peu de choses en beaucoup de mots, mais dis beaucoup de choses en peu de mots.

Pythagore

Dans ce chapitre, nous commençons par introduire du vocabulaire et des grandeurs permettant de décrire le plus précisément possible une série statistique univariée.
Puis, dans un second temps, nous étudierons les couples de séries statistiques, plus précisément nous mettrons en oeuvre des méthodes permettant de savoir si une des deux variables a une influence sur l'autre.

1.1 Introduction et vocabulaire

Un **individu** est un élément de l'ensemble sur lequel porte l'étude statistique.
L'ensemble des individus constitue la **population** *(Par exemple, les étudiants du CPES, les habitants de la France, etc.)*
Si cette population est trop nombreuse, on peut être amené à ne travailler que sur un sous-ensemble de celle-ci que l'on appelle **échantillon** issu de la population. On appelle **taille de l'échantillon**, le nombre des éléments de l'échantillon, noté N par la suite.
Le **caractère** (ou **variable**) d'une série statistique est une propriété étudiée sur chaque individu de la population ou de l'échantillon.

1. Lorsque le caractère ne prend que des valeurs numériques *(taille en cm, un temps en secondes, une note sur 20, etc.)*, il est **quantitatif**.
2. Sinon, on dit qu'il est **qualitatif** *(couleur des yeux, sport pratiqué, ville de naissance, etc.)*: les variables ne sont pas des nombres.

Faire des **statistiques**, c'est recueillir, organiser, synthétiser, représenter et exploiter des données, numériques ou non, dans un but de comparaison, de prévision, de constats...
Les statistiques sont employées dans de nombreux domaines :

- dans les assurances (risques d'accidents, de maladie, ...)
- en médecine (évaluation de l'efficacité d'un traitement, lien entre maladie et mode de vie, épidémiologie...)
- en agronomie (sélection des variétés, études de descendance...)
- en économie (emploi, conjoncture économique...)
- en sociologie (enquêtes, sondages...)
- et la liste est loin d'être exhaustive !

Cette année, nous étudierons deux "types" de statistiques descriptives :

- **les statistiques univariées** : étude d'un seul caractère sur une population ou échantillon
- **les statistiques bivariées** : étude de deux caractères sur une population (notamment la corrélation ou non entre ces deux caractères)

En L3, vous étudierez les **statistiques inférentielles** que l'on pourrait résumer en : *comment déduire des informations sur la population connaissant ces informations seulement sur un échantillon de cette population ?*

1.2 Statistiques univariées

1.2.1 Premières définitions

Définition 1.1 (Fréquence) On considère une série statistique à caractère quantitatif, dont les p valeurs sont données par :

$$x_1, x_2, \ldots, x_p, \text{ d'effectifs associés } n_1, n_2, \ldots, n_p, \text{ avec } n_1 + n_2 + \ldots + n_p = N.$$

- À chaque valeur (ou classe) est associée une **fréquence** f_i : c'est la proportion d'individus associés à cette valeur.
- $f_i = \dfrac{n_i}{N}$ est un nombre compris entre 0 et 1, que l'on peut écrire sous forme de pourcentage.
- L'ensemble des fréquences de toutes les valeurs du caractère s'appelle la **distribution des fréquences** de la série statistique.

Exemple 1.1 Voici les notes obtenues à un contrôle dans une classe de 30 élèves : (**Série A**)

2–3–3–4–5–6–6–7–7–7–8–8–8–8–8–9–9–9–9–9–9–10–10–11–11–11–13–13–15–16

On peut représenter cette série par un tableau d'effectifs et le compléter par la distribution des fréquences.

Notes	1	2	3	4	5	6	7	8	9	10
Eff.										
Fréq. en %										

Notes	11	12	13	14	15	16	17	18	19
Eff.									
Fréq. en %									

Remarque : On peut vérifier que la somme des fréquences est égale à 1 *(ou à* 100 *si on les exprime en pourcentages).*

Dans le cas d'une série statistique où le caractère étudié est continu, on peut regrouper les valeurs du caractères en **classes.** On parle alors de **regroupement en classes**.

Exemple 1.2 On étudie la superficie du logement en m^2 sur un échantillon de 1000 foyers, on obtient la **série B** suivante.

Superficie	$[20;40[$	$[40;60[$	$[60;80[$	$[80;100[$	$[100,140[$	$[140,200]$
Effectif	240	208	160	212	129	51
Fréquence en %						

Pour des caractères isolés, on peut aussi faire un **regroupement en classes**, ce qui rend l'étude moins précise, mais qui permet d'avoir une vision plus globale.

Exemple 1.3 Pour la **série A**, si on regroupe les données par classes d'amplitude 5 points, on obtient le tableau suivant.

Notes	$[0;5[$	$[5;10[$	$[10;15[$	$[15;20[$	Total
Effectif					
Fréquence					

Définition 1.2 (Effectifs cumulées, fréquences cumulées) On considère une série statistique à caractère quantitatif, dont les p valeurs sont données par :

$$x_1, x_2, \ldots, x_p, \text{ d'effectifs associés } n_1, n_2, \ldots, n_p, \text{ avec } n_1 + n_2 + \ldots + n_p = N.$$

- **L'effectif cumulé croissant** (resp. **décroissant**) de la valeur x_i est la somme des effectifs de toutes les valeurs **inférieures** (resp. **supérieures**) **ou égales** à x_i.
- **La fréquence cumulée croissante** (resp. **décroissante**) de la valeur x_i est la somme des fréquences de toutes les valeurs **inférieures** (resp. **supérieures**) **ou égales** à x_i.

Exemple 1.4 Pour la **Série A**, on obtient :

Notes	1	2	3	4	5	6	7	8	9	10
Eff. cum. crois.										
Fréq. cum. crois.										

Notes	11	12	13	14	15	16	17	18	19
Eff. cum. crois.									
Fréq. cum. crois.									

Exemple 1.5 Pour la **Série B** :

Superficie	[20;40[	[40;60[	[60;80[	[80;100[	[100,140[	[140,200]
Eff. cumulé croissant						
Fréq. cum. croiss.						

1.2.2 Représentation graphique des séries statistiques

Diagramme en bâtons et fonction de répartition

Pour représenter graphiquement une série statistique où le caractère X prend des valeurs discrètes, on utilise un **diagramme en bâtons** où les « hauteurs » des bâtons sont proportionnelles aux effectifs (ou aux fréquences).

Exemple 1.6 Traçons le diagramme en bâtons associé à la **Série A**.

On peut aussi tracer le diagramme des fréquences cumulées. On obtient alors une courbe en escalier, appelée **fonction de répartition** de X.

Exemple 1.7 Traçons la fonction de répartition associée à la **Série A**.

Propriété 1.1 La fonction de répartition est définie sur $\mathbb{R}$ et croissante de 0 à 1.

Histogramme

Pour représenter graphiquement une série statistique où le caractère est regroupé en classes, on utilise un **histogramme** où les « aires » des rectangles sont proportionnelles aux effectifs (ou aux fréquences).

Exemple 1.8 Traçons l'histogramme associé à la **Série B**.

On peut aussi tracer **la fonction de répartition**, celle-ci se calcule aux extrémités de classes à partir des fréquences cumulées F_i.
Si l'on suppose, comme dans l'histogramme, que la distribution des individus est uniforme à l'intérieur de chaque classe $[e_i, e_{i+1}[$, on peut joindre les points de coordonnées (e_i, F_i) et (e_{i+1}, F_{i+1}) par un segment de droite, la courbe obtenue est alors continue et croissante de 0 à 1.

Exemple 1.9 Traçons la fonction de répartition associée à la **Série B**.

1.2.3 Paramètres d'une série statistique univariée

Le but de cette étude est de substituer à l'ensemble des valeurs de la série statistique à étudier, quelques paramètres dont les valeurs numériques résumeront aussi fidèlement que possible les données initiales. Il faut cependant être conscient que vouloir représenter une série statistique par quelques paramètres, même judicieusement choisis, conduit à une perte non négligeable d'information.

Caractéristiques de position

Définition 1.3 (Mode) Le **mode** d'une série statistique, noté M_0, est défini comme étant la valeur du caractère ayant le plus grand effectif.

Remarque : Il se lit simplement, dans le tableau statistique ou sur le diagramme en bâtons. Mais il ne peut ne pas être unique, plusieurs valeurs du caractère peuvent avoir le même effectif. Si la série a un seul mode, on parle de série ***unimodale****, deux modes : de série* ***bimodale****.*
Pour une série statistique regroupée en classes, on définit la ***classe modale*** *comme la classe de densité la plus élevée.*

Exemple 1.10

- Pour la série A, le mode vaut
- Pour la série B,

Définition 1.4 (Moyenne) Soit une série statistique à caractère quantitatif, dont les p valeurs sont données par $x_1, x_2, \ldots, x_p$, d'effectifs associés $n_1, n_2, \ldots, n_p$, avec $n_1 + n_2 + \ldots + n_p = N$.
La **moyenne pondérée** de cette série est le nombre noté $\overline{x}$ qui vaut

$$\overline{x} = \frac{n_1 x_1 + n_2 x_2 + \cdots + n_p x_p}{n_1 + n_2 + \cdots + n_p} = \frac{1}{N} \sum_{i=1}^{p} n_i x_i.$$

Remarque : Lorsque la série est regroupée en classes, on calcule la moyenne en prenant pour valeurs x_i le ***centre de chaque classe****. Ce centre est obtenu en faisant la moyenne des deux extrémités de la classe.*

Exemple 1.11

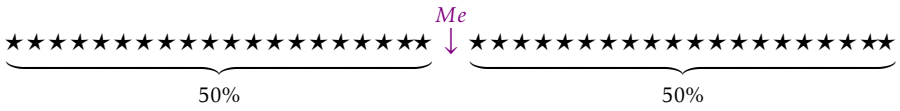
- Dans la **série A**, la moyenne du contrôle est égale à
- Dans la série B, la superficie moyenne des logements est égale à

Remarque : On peut aussi calculer une moyenne à partir de la distribution de fréquences.

$$\overline{x} = f_1 x_1 + f_2 x_2 + \cdots + f_p x_p = \sum_{i=1}^{p} f_i x_i.$$

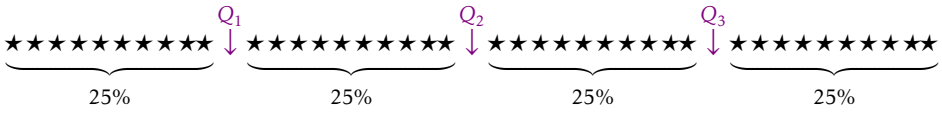
Définition 1.5 (Médiane) Soit une série statistique ordonnée dont les n valeurs sont $x_1 \leqslant x_2 \leqslant x_3 \leqslant \ldots \leqslant x_n$. La **médiane** est le nombre noté M_e qui permet de diviser cette série en deux sous-groupes de même effectif.

- Si n est **impair**, Me est la valeur de cette série qui est située au milieu, à savoir la valeur dont le rang est $\dfrac{n+1}{2}$ i.e. $x_{\frac{n+1}{2}}$.
- Si n est **pair**, Me est le centre de l'intervalle médian, qui est l'intervalle formé par les deux nombres situés "au milieu" de la série i.e. $x_{\frac{n}{2}}$ et $x_{\frac{n}{2}+1}$.



Exemple 1.12
- La médiane de la série "2 – 5 – 6 – 8 – 9 – 9 – 10" est
- La médiane de la série "2 – 5 – 6 – 8 – 9 – 9" est
- La médiane de la série "2 – 5 – 6 – 6 – 9 – 10" est

Définition 1.6 (Quartiles) Soit une série statistique, on appelle **quartiles** de la série un triplet de réels $(Q_1; Q_2; Q_3)$ qui sépare la série en quatre groupes de même effectif.



Remarque : Par définition, si X est une série statistique, $Q_2 = M_e(X)$.

Exemple 1.13 Pour la **série A**, la calculatrice nous donne $Q_1 = 7$, $Me = 8.5$ et $Q_3 = 10$.

Exemple 1.14 Pour la **série B**, la calculatrice nous donne $Q_1 = 30$, $Me = 70$ et $Q_3 = 90$.

Caractéristiques de dispersion

L'inconvénient majeur des caractéristiques de position est qu'ils ne rendent pas compte de la répartition des données.

Exemple 1.15 Considérons deux étudiants A et B ayant obtenu les notes suivantes.

- Étudiant A : 0 ; 20 ; 5 ; 15 ; 17 ; 3.
- Étudiant B : 10 ; 8 ; 12 ; 10 ; 13 ; 7.

Ces deux étudiants ont tous deux une moyenne et une médiane de 10, mais l'étudiant B a été beaucoup plus "régulier" que l'élève A.

Les grandeurs définies dans cette section visent à mesurer la dispersion des données d'une série statistique.

Définition 1.7 (Etendue) On appelle **étendue** d'une série discrète X le réel défini par

$$E(X) = \max(X) - \min(X).$$

Il s'agit de la première mesure de la dispersion d'une série statistique. Son principal mérite a longtemps été d'exister et de fournir une information sur la dispersion très simple à obtenir.

Exemple 1.16

- L'étendue de la **série A** est de
- L'étendue de la **série B** est de

Définition 1.8 (Variance et écart-type) Soit une série statistique à caractère quantitatif, dont les p valeurs sont données par

$$x_1, x_2, \ldots, x_p, \text{ d'effectifs associés } n_1, n_2, \ldots, n_p, \text{ avec } n_1 + n_2 + \ldots + n_p = N.$$

- On appelle **variance** de cette série le nombre noté $V(X)$ qui vaut

$$V(X) = \frac{n_1(x_1 - \overline{x})^2 + n_2(x_2 - \overline{x})^2 + \cdots + n_p(x_p - \overline{x})^2}{N} = \frac{1}{N}\sum_{i=1}^{p} n_i(x_i - \overline{x})^2.$$

- On appelle **écart-type** de cette série statistique le nombre noté σ_X défini par

$$\sigma_X = \sqrt{V(X)}.$$

Remarque :

- *Pour la variance d'une série statistique regroupée en classes, à l'instar de la moyenne, on remplace les x_i par les centres c_i des classes $[x_i; x_{i+1}[$.*
- *La variance et l'écart-type mesurent la dispersion des valeurs prises par X autour de sa moyenne. Plus précisément, plus la variance/l'écart-type est grand(e), plus les valeurs sont dispersées.*

Exemple 1.17 Pour la **série A**, la calculatrice nous donne $\sigma_A \approx 3.3$.

Exemple 1.18 Pour la **série B**, la calculatrice nous donne $\sigma_B \approx 37.2$.

Définition 1.9 (Écart inter-quartile) On appelle **intervalle inter-quartile** l'intervalle $[Q_1; Q_3]$. L'amplitude de cet intervalle est appelée **écart inter-quartile**.

Exemple 1.19

- Dans la **série A**, l'intervalle inter-quartile

- Dans la **série B**, l'intervalle inter-quartile

1.3 Statistiques bivariées

Les statistiques à une variable s'intéressaient pour une population donnée, à **un** caractère donné : les notes à un devoir surveillé d'une classe, les salaires dans une entreprise, etc.
Lorsque l'on s'intéresse à l'étude simultanée de **deux** caractères d'une même population, on fait ce que l'on appelle des **statistiques à deux variables** (ou bivariées), en étudiant des **séries statistiques doubles.**

Étant données deux grandeurs statistiques quantitatives X et Y, il est naturel de chercher s'il existe une relation entre X et Y, *i.e* si l'une des deux grandeurs influence l'autre et de quelle manière. C'est l'objet de cette section. Nous étudierons ici uniquement le cas des couples de variables quantitatives. On considère deux variables quantitatives X et Y définies sur la même population de taille N. On dispose donc de N couples (x_i, y_i) avec $i \in [\![1, n]\!]$.

X	x_1	x_2	$\dots$	x_i	$\dots$	x_N
Y	y_1	y_2	$\dots$	y_i	$\dots$	y_N

1.3.1 Nuage de points

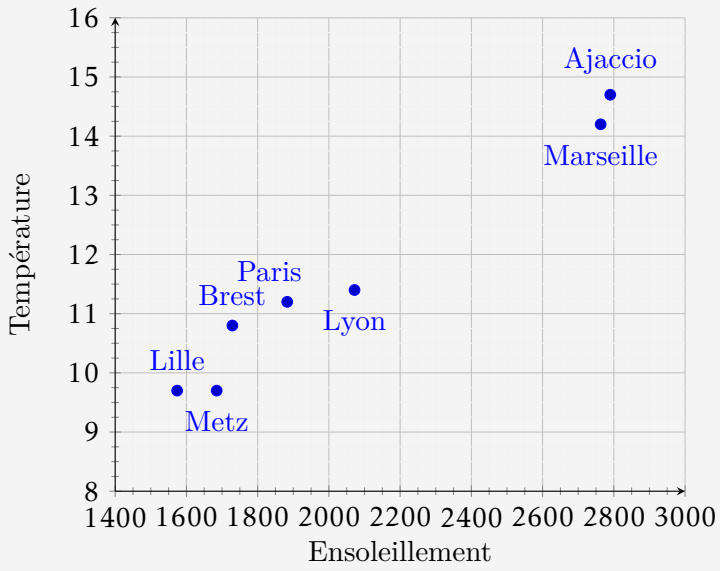
La représentation graphique la plus naturelle consiste à considérer chaque couple (x_i, y_i) comme les coordonnées d'un point M_i dans un plan muni d'un repère. L'ensemble des points M_i forme un **nuage de points**.

Exemple 1.20 Le tableau ci-dessous donne, pour chaque ville, le nombre moyen d'heures d'ensoleillement dans l'année, ainsi que la température moyenne :

Ville	Ajaccio	Lyon	Marseille	Brest	Lille	Paris	Metz
Ensoleillement	2790	2072	2763	1729	1574	1833	1685
Température	14,7	11,4	14,2	10,8	9,7	11,2	9,7

- Caractère numéro 1 : nombre d'heures d'ensoleillement dans la ville ;
- Caractère numéro 2 : température moyenne dans la ville.

Traçons le nuage de points de cette série statistique bivariée. Si on place l'ensoleillement en abscisses et la température en ordonnées :



Exercice 1.1 Soit X le nombre d'accidents et Y le nombre d'accidentés (dans ces accidents).

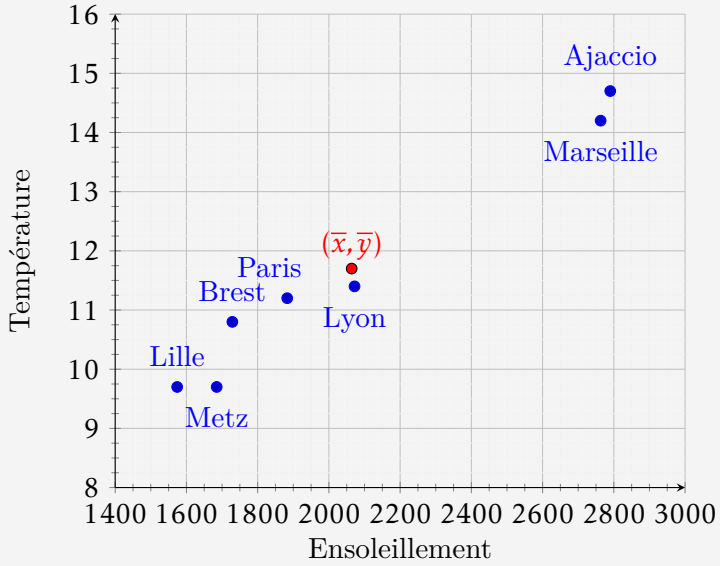
X	235	258	248	231	203	185	173
Y	325	354	340	299	308	264	238

Tracer le nuage de points représentant le nombre d'accidentés en fonction du nombre d'accidents.

1.3.2 Tendance centrale

Définition 1.10 Pour chaque série, on peut calculer comme paramètre de tendance centrale, les moyennes $\overline{x}$ et $\overline{y}$. On obtient ainsi un nouveau point G de coordonnées $(\overline{x}, \overline{y})$ appelé **point moyen.**

Exemple 1.21 On reprend l'exemple des villes.



Remarque : On peut l'interpréter comme la généralisation de moyenne à deux dimensions.

Exercice 1.2 Calculer le point moyen pour la série statistique des accidents et accidentés.

1.3.3 Dispersion

Pour chaque série, on peut aussi calculer sa variance pour mesurer la dispersion des individus autour de leur moyenne. La généralisation de cette mesure de dispersion dans le cas d'un couple de variables est la mesure de la dispersion des points du nuage autour de leur point moyen G, c'est ce que l'on appelle la **covariance** de la série double, notée $\text{cov}(X, Y)$.

Définition 1.11 On définit la **covariance** d'un couple de variables quantitatives (X, Y) ainsi :

$$\text{cov}(X, Y) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})(y_i - \overline{y}) = \frac{1}{N}\sum_{i=1}^{N} x_i y_i - \overline{xy}.$$

Remarque : Si X et Y sont identiques, $\text{cov}(X, Y) = V(X)$. La covariance est donc bien une mesure de dispersion.

Exemple 1.22 On reprend l'exemple des villes. Si X désigne le nombre d'heures d'ensoleillement et Y la température, alors :

Exercice 1.3 Calculer la covariance pour la série statistique des accidents et accidentés.

La covariance a toutefois un inconvénient majeur : elle possède des unités. On lui préfère donc le **coefficient de corrélation linéaire** qui est un nombre sans dimension. En effet, il s'agit de la covariance normée par les écarts types de chaque série.

Définition 1.12 On définit le coefficient de corrélation linéaire d'un couple de variables quantitatives ainsi :

$$\text{r}(X, Y) = \frac{\text{cov}(X, Y)}{\sqrt{V(X)V(Y)}} = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y}.$$

Exemple 1.23 On reprend l'exemple des villes. Si X désigne le nombre d'heures d'ensoleillement et Y la température, alors :

Exercice 1.4 Calculer le coefficient de corrélation linéaire pour la série statistique des accidents et accidentés.

Remarque : On peut montrer que ce nombre est compris entre -1 et 1 que si les points sont alignés $r = 1$ ou $r = -1$, les variables X et Y sont liées linéairement.
Le coefficient de corrélation linéaire mesure donc l'intensité d'une relation linéaire entre les deux séries statistiques quantitatives.
Si $r = 0$ ou proche de 0, X et Y sont linéairement indépendantes, cela peut signifier qu'il n'existe pas de relation entre ces deux séries ou que, si il existe une relation, elle n'est pas linéaire.

1.3.4 Ajustement affine

Lorsque les points du nuage paraissent presque alignés, on peut avoir l'idée de chercher quelle droite approcherait le mieux les points de ce nuage. Une telle droite permettrait notamment de faire des prévisions.
Nous cherchons donc une droite d'équation $y = ax+b$ « la plus proche possible » de l'ensemble des points du nuage. Plusieurs méthodes sont possibles.

Méthode empirique

On peut tracer la droite **à main levée**. On voit, sans peine, que l'on peut ainsi obtenir différentes solutions qui peuvent même être très différentes pour certains nuages de points, ce qui n'est guère satisfaisant.

Méthode des moindres carrés

Dans l'équation, nous avons exprimé Y en fonction de X, les deux variables ne jouent pas le même rôle : Y est la variable à expliquer (ou variable endogène) et X la variable explicative (ou variable exogène).
On veut obtenir une valeur pour Y lorsque X prend une valeur donnée : ainsi, à tout point d'abscisse x_i, on associe le point de la droite d'ajustement d'ordonnée $\widehat{y_i} = ax_i + b$ qui peut être différent de la vraie valeur y_i.
La droite d'ajustement devant passer le plus près possible de l'ensemble des points du nuage, il faut minimiser les écarts entre les y_i et les $\widehat{y_i}$ correspondants. Les valeurs absolues étant de manipulation délicate dans les calculs, on préfère minimiser la somme des carrés de ces écarts d'où le nom de **méthode des moindres carrés.**

Le problème revient alors à déterminer a et b tels que $\sum_{i=1}^{N}(y_i - ax_i - b)^2$ soit minimum.

C'est un problème d'optimisation d'une fonction de deux variables a et b.
On commence par chercher les points critiques. Pour cela, on calcule la dérivée partielle par rapport à a et la dérivée partielle par rapport à b, on obtient le système d'équations suivant :

$$\begin{cases} \sum_{i=1}^{N} x_i(y_i - ax_i - b) = 0 \quad (1) \\ \sum_{i=1}^{N} (y_i - ax_i - b) = 0 \quad (2) \end{cases}$$

En développant l'équation (2) et en divisant chaque terme par N, on obtient : $\overline{y} = a\overline{x} + b$ autrement dit, **la droite d'ajustement des moindres carrés de Y en X passe par le centre de gravité G du nuage de points.**

Le système possède une unique solution et on peut vérifier (calculs que je vous épargne) qu'il s'agit d'un **minimum.** On obtient :

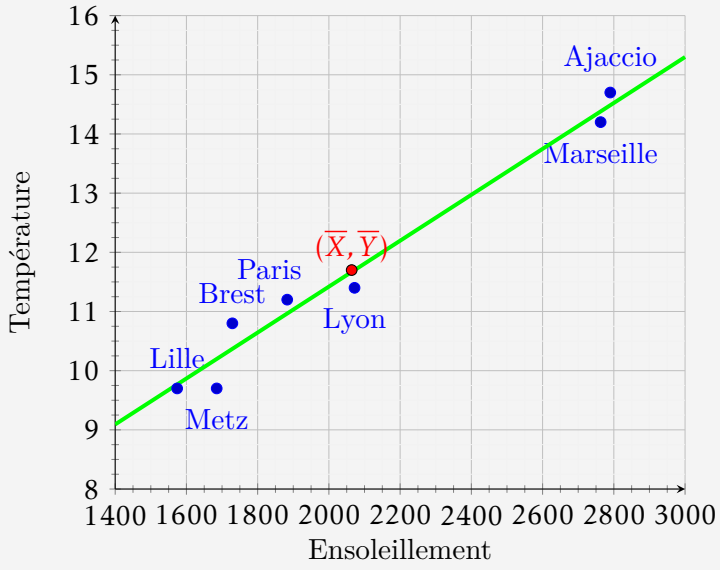
$$a = \frac{\text{cov}(X,Y)}{V(X)} \quad \text{et} \quad b = \overline{y} - a\overline{x}.$$

Méthode 1.1 (Déterminer une droite d'ajustement par la méthode des moindres carrés)

- On précise ce qui logiquement apparaît comme variable à expliquer (par exemple Y) et comme variable explicative (par exemple X).
- On calcule alors a et b à l'aide de la calculatrice ou d'un tableur.

Exemple 1.24 On reprend l'exemple **1** ci-dessus. Si X désigne le nombre d'heures d'ensoleillement et Y la température, alors on trouve pour l'équation de la droite de régression linéaire de Y en X :

$$y =$$



Remarque :

- *Plus* $|r(X,Y)|$ *est proche de 1, plus les points du nuage sont proches de l'alignement.* $|r(X,Y)|$ *ne valant 1 que lorsqu'ils sont alignés.*
- *Si* $r(X,Y) > 0$*, alors la droite est de pente positive :* X *et* Y *varient dans le même sens (lorsque l'une croît, l'autre croît, lorsque l'une décroît, l'autre décroît aussi).*
- *Si* $r(X,Y) < 0$*, alors la droite est de pente négative :* X *et* Y *varient dans des sens opposés (lorsque l'une croît, l'autre décroît).*

Exercice 1.5 Revenons sur la série statistique des accidents et accidentés :