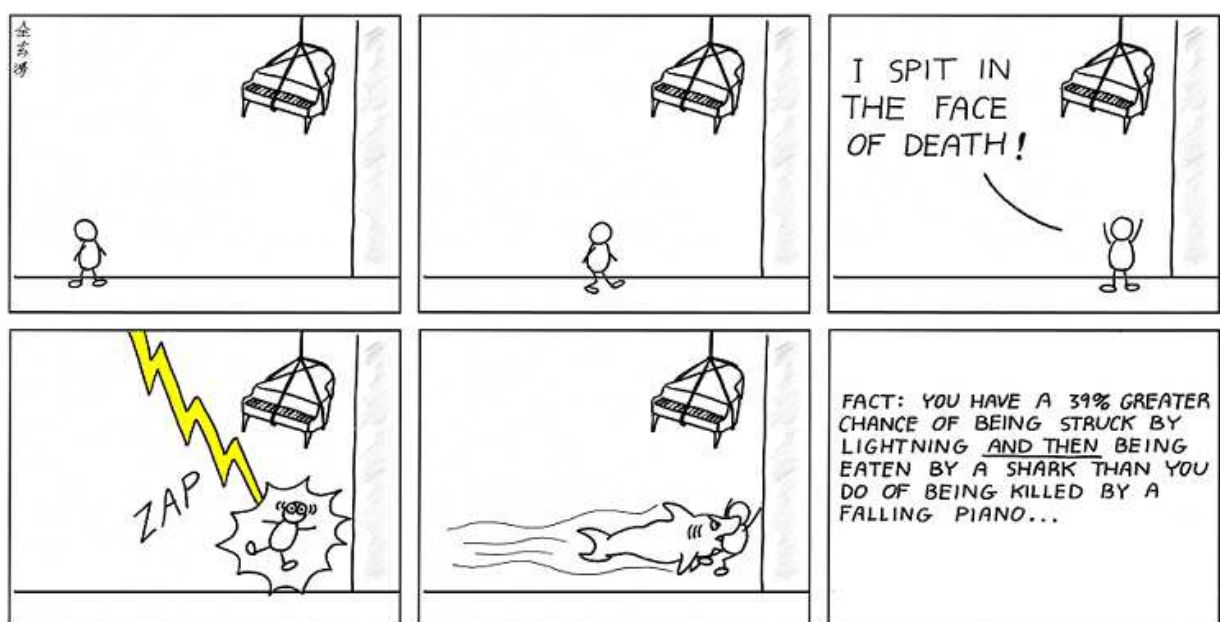


Chapitre 6

Statistiques



1. RAPPELS DE STATISTIQUE UNIVARIÉE

1.1. Vocabulaire

La **population** est l'ensemble des individus/objets sur lesquels portent l'étude statistique. (*Par exemple candidats au concours de la filière ECT, habitants de la France ...*) Les éléments d'une population sont appelés **individus**. On appelle **taille de l'échantillon**, le nombre de ses éléments, noté N par la suite.

Le **caractère** (ou **variable**) d'une série statistique est une propriété étudiée sur chaque individu :

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

Les plus gros "consommateurs" de statistiques sont les **assureurs** (risques d'accidents, de maladie des assurés), les **médecins** (épidémiologie), les **démographes** (populations et leur dynamique), les **économistes** (emploi, conjoncture économique), les **météorologues** ...

Définition 1 :

On considère une série statistique à caractère quantitatif, dont les p valeurs sont données par : x_1, x_2, \dots, x_p d'effectifs associés n_1, n_2, \dots, n_p avec $n_1 + n_2 + \dots + n_p = N$.

- A chaque valeur (ou classe) est associée une **fréquence** f_i : c'est la proportion d'individus associés à cette valeur.
- $f_i = \frac{n_i}{N}$ 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 : 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 – 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	11	12	13	14	15	16	17	18	19
Eff.	0	1	2	1	1	2	3	5	6	2	3	0	2	0	1	1	0	0	0
Fréq. en %	0	3	7	3	3	7	10	17	20	7	10	0	7	0	3	3	0	0	0

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

On peut aussi faire un regroupement par classe, ce qui rend l'étude moins précise, mais qui permet d'avoir une vision plus globale.

Exemple : Toujours pour la **série A**, si on regroupe les données par classes d'amplitude 5 points, on obtient :

Notes	[0 ; 5 [[5 ; 10 [[10 ; 15 [[15 ; 20 [total
Effectif	4	17	7	2	30
Fréquence	0,13	0,57	0,23	0,07	1

1.2. Caractéristiques de position

1.2.1 Moyenne

Définition 2 : Moyenne

Soit une série statistique à caractère quantitatif, dont les p valeurs sont données par x_1, x_2, \dots, x_p d'effectifs associés n_1, n_2, \dots, n_p avec $n_1 + n_2 + \dots + n_p = N$.

La **moyenne pondérée** de cette série est le nombre noté \bar{x} qui vaut

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

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 :

- Dans la **série A**, la moyenne du contrôle est égale à

$$\bar{m} = \frac{2 \times 1 + 3 \times 2 + \dots + 16 \times 1}{30} = \frac{254}{30} \approx 8,47$$

- si on regroupe par classe d'amplitude 5 points, une estimation de la moyenne est :

$$\bar{m} = \frac{2,5 \times 4 + 7,5 \times 17 + \dots + 17,5 \times 2}{30} = \frac{260}{30} \approx 8,67$$

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

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

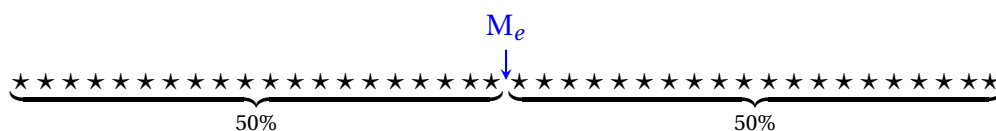
1.2.2 Médiane

Définition 3 : Médiane

Soit une série statistique ordonnée dont les n valeurs sont $x_1 \leq x_2 \leq x_3 \leq \dots \leq x_n$.

La **médiane** est un nombre M qui permet de diviser cette série en deux sous-groupes de même effectif.

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



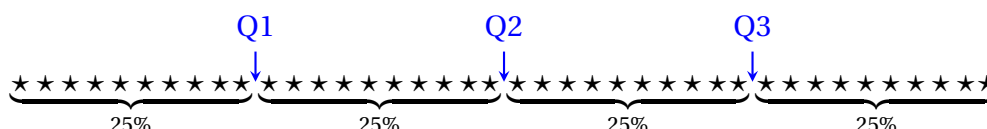
Exemple :

- La médiane de la série « 2 – 5 – 6 – 8 – 9 – 9 – 10 » est 8.
- La médiane de la série « 2 – 5 – 6 – 8 – 9 – 9 » est 7.
- La médiane de la série « 2 – 5 – 6 – 6 – 9 – 10 » est 6.

1.2.3 Quartiles

Définition 4 : 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, $Q_2 = Med(X)$.

Exemple : Pour la **série A**, la calculatrice nous donne $Q_1 = 7$, $Med = 8,5$ et $Q_3 = 10$.

1.3. Caractéristiques de dispersion

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

Exemple : Considérons deux élèves A et B ayant obtenu les notes suivantes :

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

Ces deux élèves ont tout deux une moyenne et une médiane de 10, mais l'élève 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.

1.3.1 Étendue, variance, écart-type

Définition 5 : Étendue

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 : L'étendue de la **série A** est de $E(A) = 16 - 2 = 14$.

Définition 6 : Variance et écart-type

Soit une série statistique à caractère quantitatif, dont les p valeurs sont données par x_1, x_2, \dots, x_p d'effectifs associés n_1, n_2, \dots, n_p avec $n_1 + n_2 + \dots + n_p = N$.

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

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

- On appelle **écart-type** de cette série statistique le nombre noté σ_X définie par $\sigma_X = \sqrt{V(X)}$

Remarque :

- Pour la variance d'une série statistique groupée, 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.

1.3.2 Écart inter-quartiles

Définition 7 :

On appelle **intervalle inter-quartiles** l'intervalle $[Q_1; Q_3]$.
L'amplitude de cet intervalle est appelée **écart inter-quartiles**.

Exemple :

- Dans la **série A**, l'intervalle inter-quartile est l'intervalle $[7; 10]$ dont l'écart vaut

$$10 - 7 = 3$$

- Cet intervalle comprend la moitié des notes de la série située au centre de celle-ci.

2. STATISTIQUES BIVARIÉES

Les statistiques à une variable s'intéressaient pour une population donnée, à **un** caractère donnée : 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**, en étudiant des **séries statistiques doubles**.

2.1. Analyse de caractères qualitatifs

2.1.1 Définitions et exemples

Définition 8 :

On appelle **série statistique à deux variables** la donnée de deux caractères X et Y définis sur une même population.

Une série statistique à deux variable est souvent présentée à l'aide d'un tableau d'effectifs à deux entrées. Un tableau d'effectifs à deux entrées présente l'un des caractères en ligne et l'autre caractère en colonne. À l'intersection d'une ligne et d'une colonne, on indique le nombre d'individus possédant simultanément ces deux valeurs de caractères.

Exemple : On considère le nombre d'admis dans 5 écoles de commerce en fonction de la filière d'origine :

École Filière	ESC Rennes	ESC Troyes	EM Normandie	EM Lyon	HEC Paris	Total
Scientifique	1633	401	414	2919	145	5512
Économique	1663	373	407	2375	46	4864
Technologique	832	602	613	482	3	2532
Littéraire	336	105	122	651	14	1228
Total	4464	1481	1556	6427	208	14136

2.1.2 Fréquences marginales

Les effectifs figurant dans la ligne et la colonne « Total » sont appelés **effectifs marginaux**. Les effectifs marginaux correspondent à l'étude d'un seul des deux caractères.

Dans l'exemple précédent, la colonne « Total » donne la répartition des élèves par filière et la ligne « Total » donne la répartition des élèves par école.

Définition 9 : Tableau des fréquences par rapport à l'effectif global

Un tableau des fréquences à deux entrées se présente comme un tableau d'effectifs à deux entrées dans lequel les effectifs sont remplacés par des fréquences.

Exemple : Pour obtenir le tableau des fréquences à partir du tableau précédent, il suffit de diviser chaque nombre du tableau par 14136. On obtient (arrondi au millième) :

École Filière	ESC Rennes	ESC Troyes	EM Normandie	EM Lyon	HEC Paris	Total
Scientifique	0.116	0.028	0.029	0.206	0.010	0.389
Économique	0.118	0.026	0.029	0.168	0.003	0.344
Technologique	0.059	0.042	0.043	0.034	0.001	0.179
Littéraire	0.025	0.007	0.008	0.046	0.001	0.088
Total	0.315	0.105	0.110	0.455	0.015	1

Les fréquences figurant dans la ligne et la colonne « Total » sont appelées fréquences marginales.

2.1.3 Fréquences conditionnelles

Définition 10 :

Si l'on fixe une valeur d'un des deux caractères; on obtient une série statistique à une variable appelée **série conditionnelle**. Le tableau d'effectifs de cette série correspond à une *ligne* ou à une *colonne* du tableau à double entrée.

Exemple : On cherche à étudier la répartition par filière des élèves de l'EM Lyon. Les effectifs de cette série sont donnés par la colonne « EM Lyon » du tableau d'effectifs à double entrée.

Filière	Scientifique	Économique	Technologique	Littéraire	Total
EM Lyon	2919	2375	482	651	6427

Pour obtenir le tableau des *fréquences* de cette série à une variable, il faut diviser chaque effectif par l'effectif total de l'ensemble des élèves admis à l'EM Lyon, soit 6427. On obtient :

Filière	Scientifique	Économique	Technologique	Littéraire	Total
EM Lyon	0.454	0.369	0.075	0.101	1

Ce tableau indique que 7,5% des élèves admis à l'EM Lyon sont issus de la filière ECT. On parle de la *fréquence conditionnelle* ou *fréquences des élèves de la filière ECT sachant qu'ils sont admis à l'EM Lyon*.

Proposition 1 :

La fréquence de A sachant B, notée $f_B(A)$ est égale à :

$$f_B(A) = \frac{\text{effectif de } A \cap B}{\text{effectif de B}}$$

Exemple : Dans l'exemple précédent, la fréquence des élèves d'ECT parmi les élèves admis à HEC Paris est égale à :

$$f = \frac{3}{208} \simeq 0,014$$

(ici A = « élèves issus de la filière ECT » et B = « élèves admis à HEC Paris »). Autrement dit, 1,4% des élèves admis à HEC Paris sont issus de la filière ECT.

2.2. Analyse de caractères quantitatifs

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.

2.2.1 Définitions et exemples

Tout au long de cette partie, nous illustrerons les différentes notions sur les deux exemples ci-dessus :

Exemple :

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

2. Le tableau suivant donne l'évolution du nombre d'adhérents d'un club de rugby de 2001 à 2006.

Année	2001	2002	2003	2004	2005	2006
Nombre d'adhérents	70	90	115	140	170	220

- Caractère n° 1 : l'année;
- Caractère n° 2 : le nombre d'adhérents.

Définition 11 : Moyenne empirique

Soit une série statistique à deux variables X et Y, dont les valeurs sont des couples $(x_i, y_i)_{1 \leq i \leq n}$

- On appelle **moyenne empirique** de X, noté \bar{X} , le nombre défini par :

$$\bar{X} = \frac{x_1 + x_2 + \dots + x_n}{n} = \frac{1}{n} \sum_{i=1}^n x_i$$

- On appelle **moyenne empirique** de Y, noté \bar{Y} , le nombre défini par :

$$\bar{Y} = \frac{y_1 + y_2 + \dots + y_n}{n} = \frac{1}{n} \sum_{i=1}^n y_i$$

Exemple : On reprend l'exemple 1 ci-dessus. Si X désigne le nombre d'heures d'ensoleillement et Y la température, alors :

$$\bar{X} = \frac{2790 + 2072 + 2763 + 1729 + 1574 + 1833 + 1685}{7} \simeq 2064$$

$$\bar{Y} = \frac{14,7 + 11,4 + 14,2 + 10,8 + 9,7 + 11,2 + 9,7}{7} \simeq 11,7$$

Définition 12 : Point moyen

On appelle point moyen, le point de coordonnées (\bar{X}, \bar{Y}) .

Exemple :

1. Le point moyen de l'exemple 1 a pour coordonnées (2064, 11,7).
2. Le point moyen de l'exemple 2 a pour coordonnées (2003,5, 134,2).

Définition 13 : Variance empirique

Soit une série statistique à deux variables X et Y, dont les valeurs sont des couples $(x_i, y_i)_{1 \leq i \leq n}$

- On appelle **variance empirique** de X, noté σ_X , le nombre défini par :

$$\mathbb{V}(X) = \frac{(x_1 - \bar{X})^2 + (x_2 - \bar{X})^2 + \dots + (x_n - \bar{X})^2}{n} = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{X})^2$$

- On appelle **variance empirique** de Y, noté σ_Y , le nombre défini par :

$$\mathbb{V}(Y) = \frac{(y_1 - \bar{Y})^2 + (y_2 - \bar{Y})^2 + \dots + (y_n - \bar{Y})^2}{n} = \frac{1}{n} \sum_{i=1}^n (y_i - \bar{Y})^2$$

On peut alors définir les **écarts-types empiriques** de X et Y, notés σ_X et σ_Y par :

$$\sigma_X = \sqrt{\mathbb{V}(X)} \quad \text{et} \quad \sigma_Y = \sqrt{\mathbb{V}(Y)}$$

Exemple : On reprend l'exemple 1 ci-dessus. Si X désigne le nombre d'heures d'ensoleillement et Y la température, alors :

$$\sigma_X = \sqrt{\frac{(2790 - 2064)^2 + (2072 - 2064)^2 + \dots + (1685 - 2064)^2}{7}} \simeq 472,8$$

$$\sigma_Y = \sqrt{\frac{(14,7 - 11,7)^2 + (11,4 - 11,7)^2 + \dots + (9,7 - 11,7)^2}{7}} \simeq 1,9$$

Définition 14 : Covariance empirique

Soit une série statistique à deux variables X et Y, dont les valeurs sont des couples $(x_i, y_i)_{1 \leq i \leq n}$. On appelle **covariance empirique** de (X, Y), noté $Cov(X, Y)$, le nombre défini par :

$$Cov(X, Y) = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{X})(y_i - \bar{Y}) = \frac{1}{n} \sum_{i=1}^n x_i y_i - \bar{X}\bar{Y}$$

Exemple : On reprend l'exemple 1 ci-dessus. Si X désigne le nombre d'heures d'ensoleillement et Y la température, alors :

$$Cov(X, Y) = \frac{1}{7} \times (2790 \times 14,7 + 2072 \times 11,4 + \dots + 1685 \times 9,7) - 2064 \times 11,7 \simeq 806$$

Définition 15 : Coefficient de corrélation linéaire

Soit une série statistique à deux variables X et Y, dont les valeurs sont des couples $(x_i, y_i)_{1 \leq i \leq n}$. On appelle **coefficient de corrélation linéaire** de (X, Y), noté $\rho(X, Y)$, le nombre défini par :

$$\rho(X, Y) = \frac{Cov(X, Y)}{\sigma_X \sigma_Y}$$

Exemple : On reprend l'exemple 1 ci-dessus. Si X désigne le nombre d'heures d'ensoleillement et Y la température, alors :

$$\rho(X, Y) = \frac{806}{472,8 \times 1,9} \simeq 0,91$$

2.2.2 Représentation graphique

Une première étape peut être de réaliser un graphique qui traduise les deux séries statistiques ci-dessus.

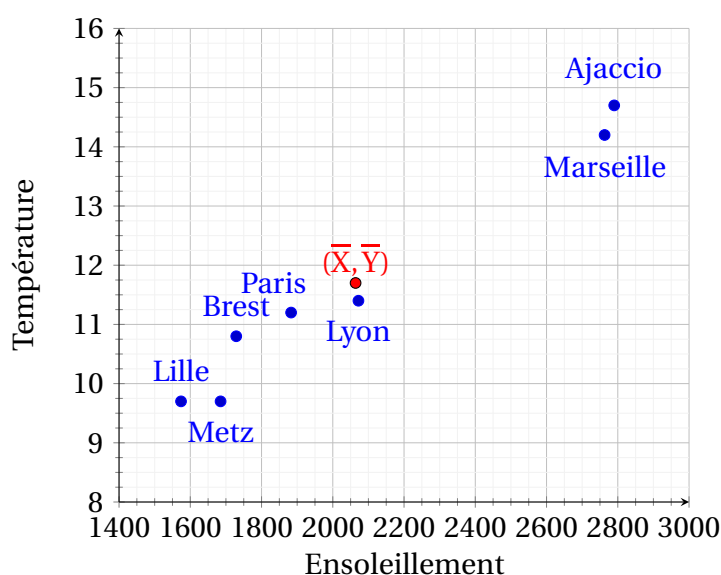
Définition 16 : Nuage de points

Soient X et Y deux variables statistiques numériques observées sur n individus.

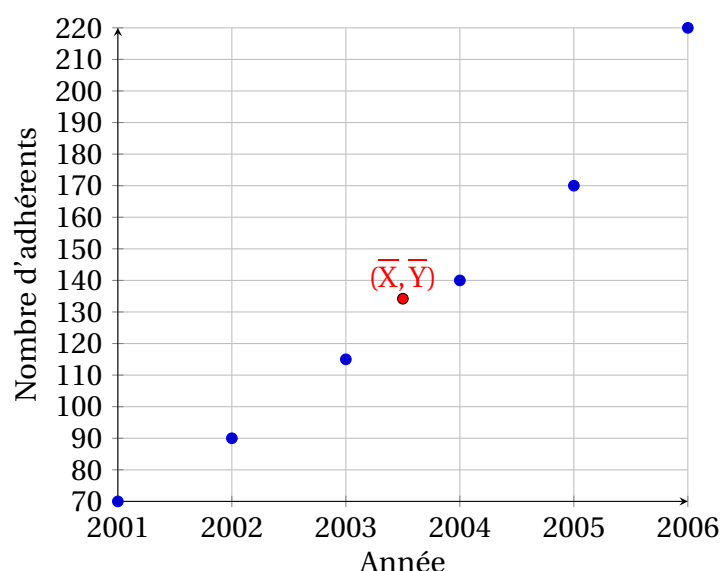
Dans un repère orthogonal, l'ensemble des n points de coordonnées (x_i, y_i) forme le **nuage de points** associé à cette série statistique.

Exemple :

1. Si on place l'ensoleillement en abscisses et la température en ordonnées :



2. Si on place l'année en abscisses, et le nombre d'adhérents en ordonnées :



2.2.3 Ajustement affine par la méthode des moindres carrés

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. Il existe de nombreuses manières d'obtenir un ajustement affine satisfaisant. L'une d'entre elle, présentée ici, est la méthode des moindres carrés.

L'idée est de chercher une droite qui minimise la somme des carrés des distances des points du nuage à cette droite. Ceci revient à chercher les réels a et b minimisant la quantité :

$$\sum_{i=1}^n (y_i - ax_i - b)^2.$$

Théorème 1 : Droite de régression linéaire de Y en X

Il existe une unique droite d'équation $y = ax + b$, rendant minimale $\sum_{i=1}^n (y_i - ax_i - b)^2$.

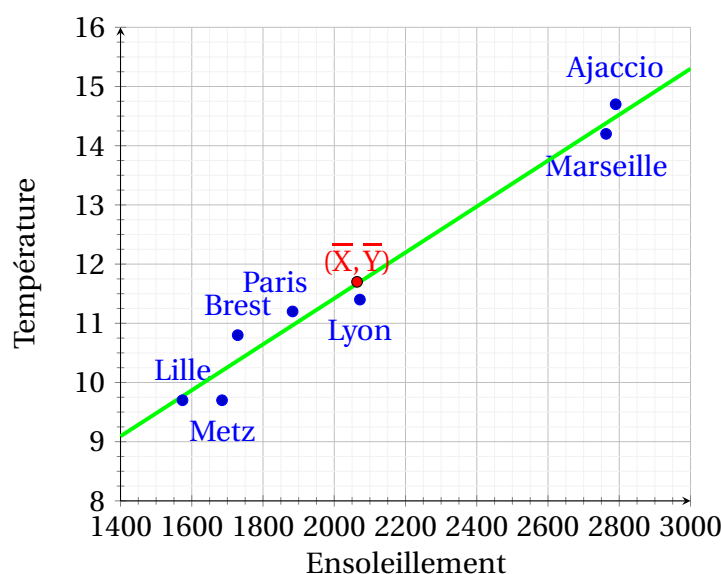
C'est la droite d'équation

$$y = \frac{\text{Cov}(X, Y)}{\sigma_X^2} (x - \bar{X}) + \bar{Y}$$

Cette droite est appelée **droite de régression linéaire de Y en X**.

Exemple : 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 :

$$y = 0,00388x + 3,66$$



Remarque :

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

Proposition 2 :

La droite de régression linéaire passe par le point moyen.

Exercice : Calculer la droite de régression linéaire pour l'exemple 2.

3. EXERCICES

Statistique univariée

6.1 Un commerçant observe, durant les sept premiers mois de l'ouverture de son officine, le chiffre d'affaire mensuel en milliers d'euros. Le résultat de l'observation est résumé dans le tableau suivant où X désigne le chiffre d'affaire correspondant.

X	12	13	15	19	21	22	24
-----	----	----	----	----	----	----	----

Calculer la moyenne \bar{X} , la variance $V(X)$, ainsi que la médiane de cette série.

6.2 Le tableau de l'espérance de vie X à la naissance pour les hommes, dans un certain pays, durant une période de 8 ans, est le suivant :

Année	1	2	3	4	5	6	7	8
Hommes	75,3	75,5	75,8	75,9	76,8	77,2	77,6	

Déterminer la moyenne, la variance, le premier quartile, la médiane, le troisième quartile, ainsi que l'écart inter-quartile, de l'espérance de vie des hommes en France pendant la période considérée.

6.3 Le répartition des notes aux tests logiques de l'ESSEC pour les admissibles en 2008 est donnée par le tableau suivant :

Notes	[0; 4[[4; 6[[6; 8[[8; 10[[10; 12[[12; 14[[14; 16[[16; 20[
Fréquences (%)	1,08	4,81	21,03	32,45	28,00	9,50	2,88	0,24

Calculer la moyenne et la médiane de cette série. L'ESSEC annonce une moyenne de 9,40. Expliquer la différence avec la valeur trouvée.

Statistique bivariée

6.4 Le tableau de l'espérance de vie à la naissance pour les hommes et les femmes en France, entre 2000 et 2007, est le suivant :

Année	2000	2001	2002	2003	2004	2005	2006	2007
Hommes	75,3	75,5	75,8	75,9	76,8	76,8	77,2	77,6
Femmes	82,8	82,9	83	82,9	83,9	83,8	84,2	84,5

On note X l'espérance de vie des hommes et Y l'espérance de vie des femmes.

1. Calculer les moyennes empiriques \bar{X} et \bar{Y} .
2. Calculer les variances et écart-types empiriques $V(X)$, $V(Y)$, σ_X et σ_Y .
3. Calculer la covariance empirique $Cov(X, Y)$.
4. En déduire le coefficient de corrélation linéaire $\rho(X, Y)$. Un ajustement linéaire entre les variables mesurant les espérances de vie pour les hommes et pour les femmes est-il pertinent?
5. Tracer sur un même graphique le nuage de points associée à la série statistique double (X, Y) et la droite de régression linéaire.

6.5 L'évolution du chiffre d'affaire (en millions d'euros) d'une entreprise depuis sa création en 2002 est donnée par le tableau suivant :

Année	2002	2003	2004	2005	2006	2007	2008	2009
Chiffre d'affaire	0,7	1,6	2	2,4	2,5	2,8	3	3

On note X la variable donnant l'année et Y celle donnant le chiffre d'affaire.

1. Calculer les moyennes empiriques \bar{X} et \bar{Y} .
2. Calculer les variances et écart-types empiriques $\mathbb{V}(X)$, $\mathbb{V}(Y)$, σ_X et σ_Y .
3. Calculer la covariance empirique $Cov(X, Y)$.
4. En déduire le coefficient de corrélation linéaire $\rho(X, Y)$. Un ajustement linéaire entre X et Y est-il pertinent?
5. Calculer le coefficient de corrélation de X et $\exp(Y)$ et le comparer à celui de X et Y . En déduire ce que l'on appelle un ajustement logarithmique de Y en X .

4. TABLE DES MATIÈRES

1	Rappels de statistique univariée	2
1.1	Vocabulaire	2
1.2	Caractéristiques de position	3
1.2.1	Moyenne	3
1.2.2	Médiane	3
1.2.3	Quartiles	4
1.3	Caractéristiques de dispersion	4
1.3.1	Étendue, variance, écart-type	4
1.3.2	Écart inter-quartiles	5
2	Statistiques bivariées	5
2.1	Analyse de caractères qualitatifs	5
2.1.1	Définitions et exemples	5
2.1.2	Fréquences marginales	6
2.1.3	Fréquences conditionnelles	6
2.2	Analyse de caractères quantitatifs	7
2.2.1	Définitions et exemples	7
2.2.2	Représentation graphique	9
2.2.3	Ajustement affine par la méthode des moindres carrés	10
3	Exercices	12
3.1	Statistique univariée	12
3.2	Statistique bivariée	12