Statistiques - Semestre 2

BUT informatique

Patrice Pérot

28 avril 2023

Chapitre 1

Vocabulaire des statistiques

La statistique est l'étude de la **collecte de données**, leur **analyse**, leur **traitement**, l'**interprétation des résultats et leur présentation** afin de rendre les données compréhensibles par tous. C'est à la fois une science, une méthode et un ensemble de techniques.

L'analyse des données est utilisée pour **décrire les phénomènes** étudiés, **faire des prévisions** et **prendre des décisions** à leur sujet. En cela, la statistique est un outil essentiel pour la compréhension et la gestion des phénomènes complexes.

Les données étudiées peuvent être de toute nature, ce qui rend la statistique utile dans tous les champs disciplinaires : de l'économie à la biologie en passant par la psychologie et l'informatique.

1.1 Vocabulaire

Les statistiques consistent en diverses méthodes de classement des données tels que les tableaux, les histogrammes et les graphiques, permettant d'organiser un grand nombre de données. Les statistiques se sont développées dans la deuxième moitié du XIXe siècle dans le domaine des sciences humaines (sociologie, économie, anthropologie, ...). Elles se sont dotées d'un vocabulaire particulier.

1.1.1 Épreuve statistique

Les **statistiques descriptives** visent à étudier les caractéristiques d'un ensemble d'observations comme les mesures obtenues lors d'une expérience. L'**expérience** est l'étape préliminaire à toute étude statistique. Il s'agit de prendre « contact » avec les observations. De manière générale, la méthode statistique est basée sur le concept suivant :

Définition 1. L'**épreuve statistique** est une expérience que l'on provoque.

Exemple 1. Le recensement permet d'établir la population officielle de chaque commune. Il fournit également des informations sur les caractéristiques de la population : âge, profession, moyens de transport utilisés, conditions de logement...

1.1.2 Population

Définition 2. On appelle **population** l'ensemble sur lequel porte une étude statistique. Cet ensemble est noté Ω.

Remarque 1. Ce terme vient du fait que la démographie, étude des populations humaines, a occupé une place centrale aux débuts de la statistique, notamment au travers des recensements de population.

Exemple 2. Lorsqu'on s'intéresse à la circulation automobile dans une ville, la population est constituée de l'ensemble des véhicules susceptibles de circuler dans cette ville à une date donnée. Dans ce cas Ω =ensemble des véhicules.

1.1.3 Individu (unité statistique)

Une population est composée d'individus. Les individus qui composent une population statistique sont appelés unités statistiques.

Définition 3. On appelle individu tout élément de la population Ω, il est noté ω (ω dans Ω).

Exemple 3. Si on étudie la production annuelle d'une usine de puces électroniques. La population est l'ensemble des puces produites durant l'année et une puce électronique constitue un individu.

1.1.4 Caractère (variable statistique)

La statistique « descriptive » cherche à décrire une population donnée. en étudiant ses différentes caractéristiques.

Définition 4. On appelle **caractère** (ou variable statistique, dénotée V.S) toute application $X : \Omega \to C$. L'ensemble C est dit : ensemble des valeurs du caractère X (c'est ce qui est mesuré ou observé sur les individus).

Exemple 4. Taille, température, nationalité, couleur des yeux, catégorie socioprofessionnelle ...

1.1.5 Modalités

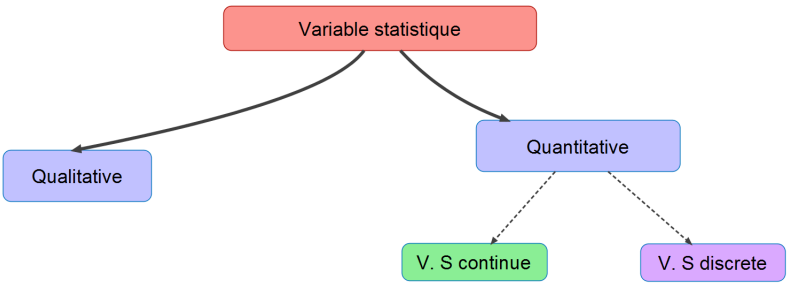
Définition 5. Les **modalités** d'une variable statistique sont les différentes valeurs que peut prendre celle-ci. Ce sont les différentes situations dans lesquelles les individus peuvent se trouver à l'égard du caractère considéré.

Exemple 5. Si la variable est la situation familiale, les modalités peuvent être « célibataire », « marié », « divorcé ».

Si la variable est la taille d'un individu, les modalités sont des nombres réels.

1.2 Types de caractères

Il existe deux catégories de caractères : les caractères qualitatifs et les caractères quantitatifs.



1.2.1 Caractère qualitatif

Les **caractères qualitatifs** sont ceux dont les modalités ne sont pas des nombres (on ne peut pas les quantifier). Il s'agit de qualités.

Exemple 6. Le caractère « état d'une maison » est un caractère qualitatif car on peut considérer les modalités suivantes : ancienne, dégradée, nouvelle, rénovée.

1.2.2 Caractère quantitatif

Les **caractères quantitatifs** sont des caractères dont les modalités sont des nombres. Il s'agit de quantités. Ainsi, l'âge, la taille ou le salaire d'un individu sont des caractères quantitatifs.

On distingue les variables quantitatives discrètes et les variables quantitatives continues :

— une **variable quantitative discrète** est une variable ne prenant que des valeurs isolées (souvent des valeurs entières). Par exemple, le nombre de maisons par quartier d'une ville.

— une **variable quantitative continue** et une variable dont les modalités peuvent être n'importe quel nombre d'un intervalle. Par exemple, le temps de réalisation d'une tâche ou la taille d'un individu.

Chapitre 2

Statistiques à une variable

2.1 Représentation graphique d'une série statistique

On distingue les méthodes de représentation d'une variable statistique en fonction de la nature de cette variable (qualitative ou quantitative). Le plus souvent, on représente des séries statistiques sous la forme d'un tableau ou sous la forme d'un graphique.

Le graphique est un support visuel qui permet :
La synthèse : il permet de visualiser d'un seul coup d'œil les principales caractéristiques (mais on perd de l'information) ;
La découverte : il met en évidence les tendances ;
Le contrôle : il met en évidence les anomalies ;
La recherche des régularités : il met en évidence les régularités dans le mouvement et les répétition du phénomène.

2.1.1 Distribution à caractère qualitatif

Pour une variable quantitative, on utilise un diagramme en barres ou un diagramme circulaire.

Diagramme en barres : Les modalités sont portées sur l'axe des abscisses, de façon arbitraire. Pour chaque modalité, on trace un rectangle dont la hauteur est proportionnelle à l'effectif correspondant.

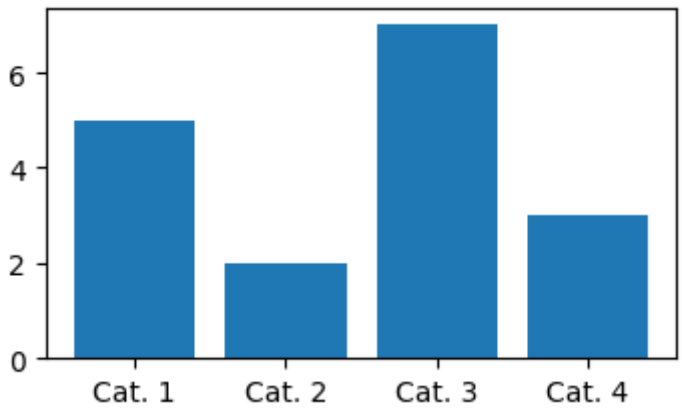


Diagramme en barres

Exercice 1. Le nombre d'habitants, en millions, de quelques pays européens est donné par le tableau suivant :

Pays	Allemagne	Belgique	Espagne	France	Italie	Pologne	Portugal
Population (en millions)	83,1	11,4	47,3	67,8	59,0	38,4	10,3

Construire le diagramme en barres associé à cette série statistique.

Diagramme circulaire : On partage un disque en secteurs angulaires, correspondant aux modalités, et dont l'angle est proportionnel à l'effectif de la modalité correspondante.

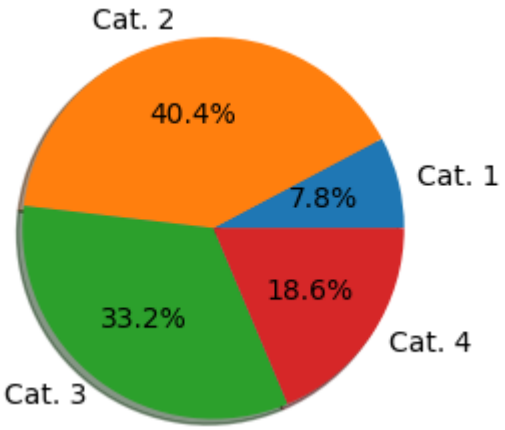


Diagramme circulaire

Exercice 2. En 2017, 74 % des actifs en emploi qui déclarent se déplacer pour rejoindre leur lieu de travail utilisent leur voiture, 16 % prennent des transports en commun, 6 % s'y rendent à pied et 2 % en vélo.

Dessiner un diagramme circulaire permettant d'illustrer cette affirmation.

2.1.2 Distribution à caractère quantitatif discret

On représente généralement une série statistique quantitative discrète par un diagramme en bâtons.

Diagramme en bâtons : On place les valeurs x_i sur l'axe des abscisses et pour chaque x_i, on trace un segment vertical dont la hauteur est proportionnelle à l'effectif correspondant.

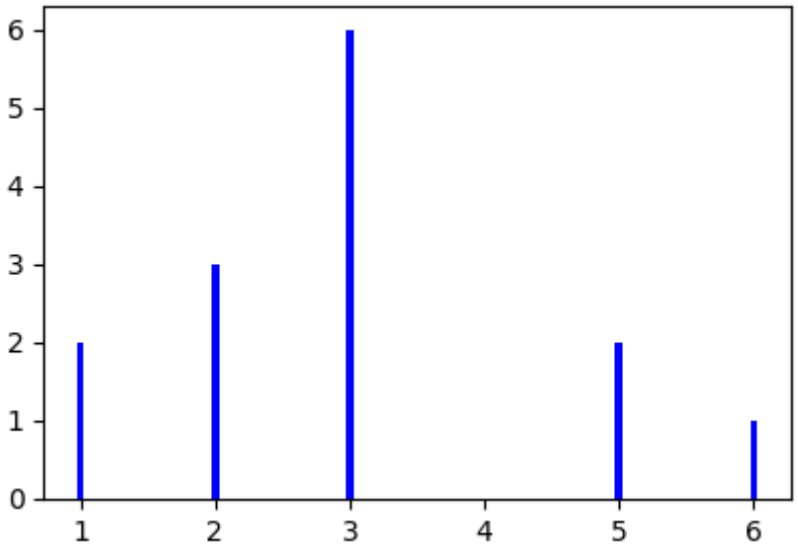


Diagramme en bâtons

Exercice 3. Les notes obtenues à un partiel d'un groupe d'étudiants sont les suivantes :

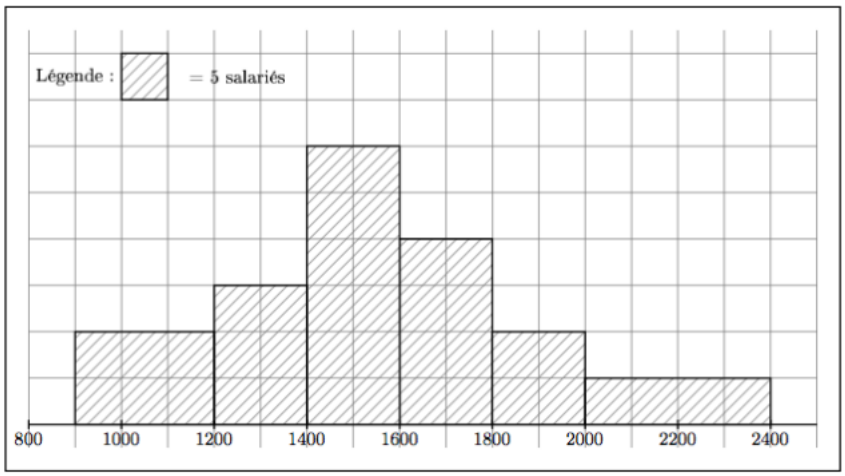
6 ; 6 ; 8 ; 10 ; 10 ; 10 ; 11 ; 11 ; 12 ; 12 ; 12 ; 12 ; 13 ; 13 ; 15 ; 15 ; 16

Construire le diagramme en bâtons associé à cette série statistique.

2.1.3 Distribution à caractère quantitatif continu

Lorsque la variable est quantitative continue, on utilise un histogramme.

Les valeurs de la série statistique sont rangées en classes.
Après avoir choisi une unité d'aire, pour chaque classe, on trace un rectangle dont la largeur correspond à la largeur de la classe et dont l'aire est proportionnelle à l'effectif correspondant.



Histogramme

Exercice 4. On considère l'histogramme précédent qui représente le nombre de salariés d'une entreprise en fonction du salaire.
À partir de cet histogramme, construire un tableau dont la première ligne contient les salaires (sous forme d'intervalles) et la seconde ligne contient le nombre de salariés dont le salaire appartient à l'intervalle indiqué dans la première ligne.

2.2 Indicateurs de position d'une série statistique

Les indicateurs de position ne concernent que les séries quantitatives.

2.2.1 Le mode

Définition 6. Le mode d'une série statistique est la valeur qui a le plus grand effectif.

Remarque 2. Il peut y avoir une ou plusieurs modes.

2.2.2 La médiane

Définition 7. On appelle **médiane**, notée M, toute valeur telle qu'au moins 50 % des valeurs de la série sont inférieures ou égales à M et au moins 50 % des valeurs de la série sont supérieures ou égales à M.

Remarque 3. Pour une série quantitative discrète, on prend généralement pour la médiane M la valeur « du milieu » si l'effectif total est impair, et la moyenne des deux valeurs « du milieu » si l'effectif total est impair.

Exercice 5. Déterminer la médiane de la série statistique de l'exercice 3 et interpréter ce résultat.

2.2.3 La moyenne

Définition 8. Pour une série quantitative discrète, on appelle **moyenne** de X, la quantité :

$$\overline{x} = \frac{1}{N}\sum_{i=1}^{n} n_i x_i = \sum_{i=1}^{n} f_i x_i$$

avec $N = \text{Card}(\Omega) = \sum_{i=1}^{n} n_i$.

Propriété 1. La moyenne est **linéaire** : quand on ajoute (resp. multiplie) une même quantité à toutes les valeurs d'une série, la nouvelle moyenne s'obtient en ajoutant (resp. multipliant) cette quantité à l'ancienne moyenne.

Exercice 6. On reprend les données de l'exercice 4.

1. En prenant comme valeurs les centres des intervalles, calculer le salaire moyen de cette entreprise.
2. Le chef d'entreprise souhaite augmenter le salaire de tous ses employés. Il hésite entre augmenter tous les salaires de 3 %, ou augmenter tous les salaires de 50 €.

   Calculer le salaire moyen après application de l'une et l'autre de ces solutions.

2.2.4 Les quartiles

Définition 9. On appelle **premier quartile**, noté Q_1, la première valeur x_i de la V.S X telle que $F(x_i) \geqslant 0,25$.
On appelle **troisième quartile**, noté Q_3, la première valeur x_i de la V.S X telle que $F(x_i) \geqslant 0,75$.

Remarque 4. On représente graphiquement les indicateurs médiane, quartiles, minimum et maximum par un diagramme en boîte.

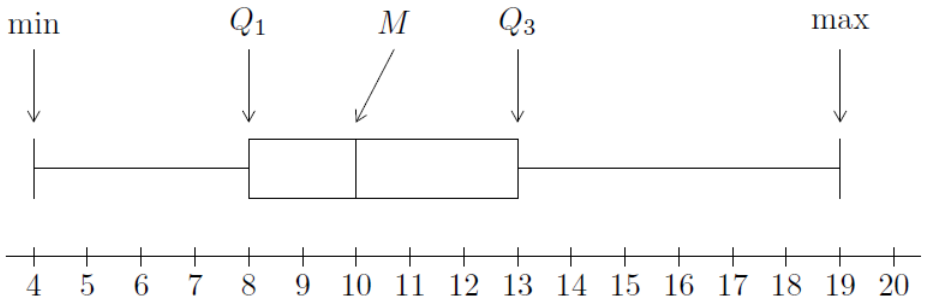


Diagramme en boîte

Cela permet de comparer facilement deux séries statistiques.

Exercice 7. Déterminer les premier et troisième quartile de la série statistique de l'exercice 3 et représenter cette série par un diagramme en boîte.

2.3 Indicateurs de dispersion d'une série statistique

Les indicateurs de dispersion ne concernent que les séries quantitatives. Ils permettent de mesurer la dispersion des valeurs les unes par rapport aux autres (proches ou éloignées les unes des autres).

2.3.1 L'étendue

Définition 10. La différence entre la plus grande valeur et la plus petite valeur du caractère, donnée par la quantité $e = x_{max} - x_{min}$, s'appelle l'**étendue** de la variable statistique X.

Remarque 5. Le calcul de l'étendue est très simple et permet de donner rapidement une première idée de la dispersion des observations.

2.3.2 La variance

Définition 11. On appelle **variance** de cette série statistique X, le nombre :

$$\text{Var}(X) = \frac{1}{N}\sum_{i=1}^{n} n_i(x_i - \overline{x})^2$$

.

Remarque 6. La variance est la moyenne des carrés des écarts à la moyenne $\overline{x}$.

Exercice 8. On considère la série statistique suivante :

p_i	-8	-4	3	6
Effectif	3	7	5	2

Calculer la variance et l'écart-type de cette série statistique.

Le théorème suivant (Théorème de König-Huygens) donne une identité remarquable reliant la variance et la moyenne, parfois plus pratique dans le calcule de la variance.

Théorème 1. Soit (x_i, n_i) une série statistique de moyenne $\overline{x}$ et de variance Var(X). Alors :

$$\text{Var}(X) = \frac{1}{N}\sum_{i=1}^{n} n_i x_i^2 - \overline{x}^2$$

2.3.3 L'écart-type

Définition 12. La quantité $\sigma(X) = \sqrt{\text{Var}(X)}$ s'appelle l'**écart-type** de la variable statistique X.

Remarque 7. Il sert à mesurer la dispersion d'une série statistique autour de sa moyenne :
- Plus il est petit, plus les valeurs sont concentrées autour de la moyenne (on dit que la série est homogène).
- Plus il est grand, plus les valeurs sont dispersées autour de la moyenne (on dit que la série est hétérogène).

2.4 Courbe des effectifs cumulés croissants ou des fréquences cumulées croissantes

Définition 13. Quand les valeurs d'un caractère quantitatif sont rangées dans l'ordre croissant, on définit :
- L'**effectif cumulé croissant** d'une valeur est la somme des effectifs des valeurs inférieures ou égales à cette valeur ;
- la **fréquence cumulée croissante** d'une valeur est la somme des fréquences des valeurs inférieures ou égales à cette valeur.

Remarque 8. On définit de même les effectifs cumulés décroissants et les fréquences cumulées décroissantes.

Définition 14. La courbe des effectifs cumulés croissants (ou des fréquences cumulées croissantes), appelée aussi polygone des effectifs cumulés croissants (ou des fréquences cumulées croissantes), notamment dans le cas d'une série statistique quantitative continue, permet de retrouver la valeur de la médiane et des premier et troisième quartiles.

Exercice 9. On reprend les données de l'exercice 4.

1. Construire, en prenant comme valeurs les centres des intervalles, la courbe des effectifs cumulés croissants.
2. Déterminer graphiquement la médiane ainsi que les premier et troisième quartiles.
3. Pour parler des salaires de cette entreprise, vaut-il mieux utiliser la médiane ou la moyenne ?

2.5 Exercices

Exercice 10. Un magasin de sport a fait l'inventaire des paires de skis qu'il propose en location :

Taille des skis (en cm)	Nombre de paires
130	2
135	5
140	7
145	15
150	35
155	46
160	63
165	32
170	10
175	10

1. Construire la représentation graphique la plus adaptée à cette série statistique.
2. Calculer la taille moyenne des skis proposés par ce magasin.
3. Représenter la courbe des effectifs cumulés croissants et estimer graphiquement la médiane et les premier et troisième quartiles.
4. Déterminer par le calcul la médiane, les premier et troisième quartiles et vérifier la cohérence avec la question précédente.
5. Représenter cette série statistique par un diagramme en boîte.

Exercice 11. En Python, construire un module `statistiques` contenant les fonctions permettant de calculer :

- la moyenne ;
- la médiane ;
- les premier et troisième quartiles ;
- la variance ;
- l'écart-type.

Ces fonctions doivent prendre un paramètre obligatoire de type `list` correspondant aux valeurs d'une série statistique, et un paramètre facultatif de type `list` correspondant aux effectifs ou aux fréquences des valeurs de la série statistique.
Par exemple :

```
>>> moyenne([1, 5, 9, 4, 2])
4.2
>>> mediane([1, 5, 9, 4, 2], [3, 1, 2, 2, 1])
4
```

Exercice 12. Le tableau ci-dessous donne le nombre de naissances (en milliers) par an en France métropolitaine entre 1901 et 1920.

Année	1901	1902	1903	1904	1905	1906	1907	1908	1909	1910
Nombre de naissances en milliers	917,1	904,4	884,5	877,1	865,6	864,7	829,6	849,0	824,7	828,1
Année	1911	1912	1913	1914	1915	1916	1917	1918	1919	1920
Nombre de naissances en milliers	793,5	801,6	795,9	757,9	483,0	384,7	412,7	472,8	507,0	838,1

1. Calculer le nombre moyen de naissances par an en France métropolitaine entre 1901 et 1920. Arrondir la réponse à la centaine.
2. Calculer la médiane, les premier et troisième quartiles de cette série statistique.
3. Construire le diagramme en boîte de cette série statistique.

Exercice 13. Le fichier `parc-informatique-2020.csv` recense le parc informatique d'une entreprise en 2020.
Répondre aux questions suivantes en utilisant exclusivement Python.

1. Déterminer l'âge moyen des photocopieurs.
2. Dessiner le diagramme circulaire donnant la répartition des différents types de matériels informatiques.
3. Dessiner le diagramme en bâtons de l'année de livraison des PC fixes.

Exercice 14. Le fichier `controle-RGPD-2021.csv` recense les contrôles qui ont été effectués dans le cadre du RGPD.
Répondre aux questions en utilisant Python :

1. Déterminer le département dans lequel il y a eu le plus de contrôles concernant la santé.
2. Dessiner le diagramme en barres représentant le nombre de contrôles en fonction du type d'activité.
3. Dessiner le diagramme circulaire représentant le nombre de contrôles en région Île de France en fonction du département.

Chapitre 3

Statistiques à deux variables

Ce chapitre s'intéresse au lien qu'il peut y avoir entre deux caractères quantitatifs d'un même individu.

Par exemple, concernant les salariés d'une entreprise, on peut s'interroger sur le lien éventuel entre leurs salaires et leurs ancienneté.

On considère donc une population Ω avec $\text{Card}(\Omega) = N$ et on pose :

$$\begin{array}{rccc} Z & : & \Omega & \to & \mathbb{R}^2 \\ & & \omega & \mapsto & Z(\omega) = (X(\omega), Y(\omega)) \end{array}$$

Dans ce cas, Z est appelée variable statistique à deux variables. Le couple (X, Y) est appelé le couple de la variable statistique.

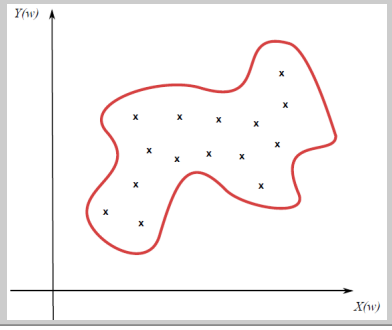
3.1 Représentation des séries statistiques à deux variables

On rassemble les données statistiques dans un tableau.
A chaque ω_i, on associe (x_i, y_i). On rassemblera les données comme dans le tableau suivant :

ω_i	ω_1	ω_2	...	ω_N
Variable X	$X(\omega_1)$	$X(\omega_2)$	...	$X(\omega_N)$
Variable Y	$Y(\omega_1)$	$Y(\omega_2)$	...	$Y(\omega_N)$

On peut alors représenter ces données statistiques par un nuage de points :

Définition 15. On appelle **nuage de points** d'une série statistique à deux variables l'ensemble des points du plan $(X(\omega_i); Y(\omega_i))$, pour i allant de 1 à N.



Définition 16. On appelle **point moyen** d'une série statistique à deux variables le point $\overline{M}$ de coordonnées $(\overline{x}; \overline{y})$ où $\overline{x}$ et $\overline{y}$ sont les moyennes des séries X et Y.

Exercice 15. La série statistique à deux variables suivante décrit la superficie certifiée de production biologique exprimée en hectares (ha) en France de 2004 à 2009 : y_i est la superficie pour l'année $2003+x_i$.

Année	2004	2006	2007	2008	2009
x_i	1	3	4	5	6
y_i	468	500	497	502	526

Source des données : Eurostat

1. Tracer dans un repère adapté le nuage de points associé à cette série statistique.
2. Calculer les coordonnées du point moyen et le placer dans le repère.

3.2 Indicateurs statistiques

Les séries X et Y peuvent être vues comme des séries à une variable. On peut donc calculer les moyennes $\overline{x}$ et $\overline{y}$, les variances Var(X) et Var(Y), et enfin les écart-types σ_X et σ_Y des séries X et Y.

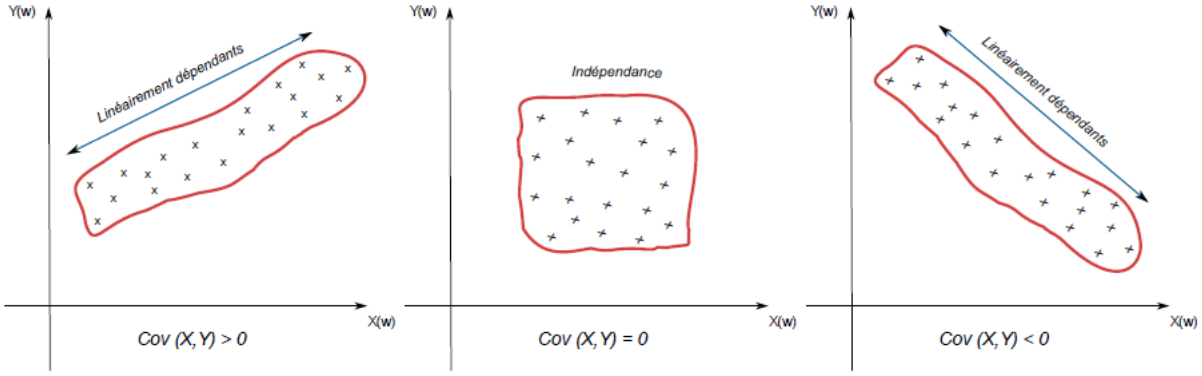
3.2.1 Notion de covariance

Définition 17. On appelle **covariance** des variables X et Y le nombre défini par :

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

Remarque 9.

- La covariance est un paramètre qui donne la variabilité de X par rapport à Y :



- $\text{Cov}(X,X) = \text{Var}(X)$ et $\text{Cov}(Y,Y) = \text{Var}(Y)$.
- On a également : $\text{Cov}(X,Y) = \frac{1}{N}\sum_{i=1}^{N} x_i y_i - \overline{xy}$

Exercice 16. On reprend les données de l'exercice 15. Calculer la covariance de cette série statistique.

3.3 Ajustement linéaire

Dans le cas où on peut mettre en évidence l'existence d'une relation linéaire significative entre deux caractères quantitatifs continus X et Y (le nuage de points est a plus ou moins la forme d'une droite), on peut chercher à formaliser la relation qui unit ces deux variables à l'aide d'une équation de droite et qui résume cette relation. Cette démarche s'appelle l'ajustement linéaire.

3.3.1 Coefficient de corrélation

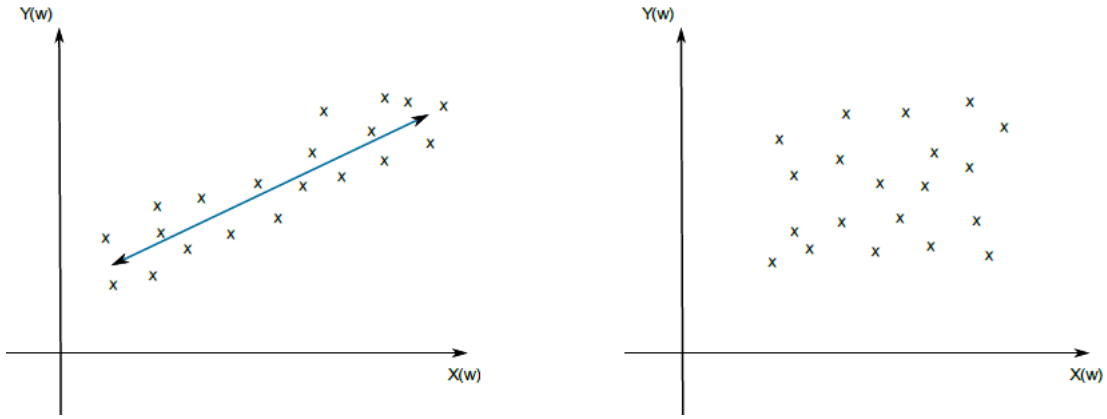
Définition 18. On appelle **coefficient de corrélation** le nombre défini par :

$$\rho_{XY} = \frac{\text{Cov}(X,Y)}{\sigma_X \sigma_Y}$$

Propriété 2. Le coefficient ρ_{XY} est un nombre compris entre -1 et 1.

Le coefficient ρ_{XY} mesure le degré de liaison linéaire entre X et Y (voir figures suivantes) :

- plus la valeur approchée de ρ_{XY} est proche de 1 ou de -1, plus X et Y sont liées linéairement.
- plus la valeur approchée de ρ_{XY} est proche de 0, moins il y a de liaison linéaire entre X et Y .



La série statistique représentée par le nuage de points de gauche a un coefficient de corrélation proche de 1 alors que celle représentée par le nuage de points de droite a un coefficient de corrélation proche de 0.

Exercice 17. On reprend les données de l'exercice 15. Calculer le coefficient de corrélation. Interpréter le résultat.

3.3.2 Droite de régression

Lorsque le coefficient de corrélation est proche de 1 ou de -1, le nuage de points a la forme d'une droite. On cherche l'équation de la droite qui passe « au plus près » de chacun des points.

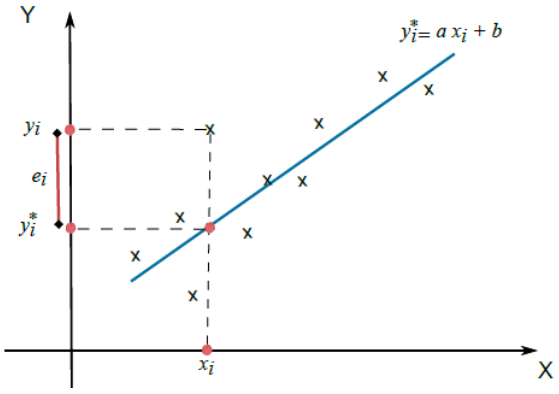
Plusieurs méthodes permettent d'obtenir cette droite :

- au jugé : on trace la droite qui nous semble ajuster au mieux le nuage de points (généralement, on la fait passer par le point moyen), et on détermine graphiquement son équation ;
- la droite de Mayer : on partage le nuage de points rangés dans l'ordre croissant de leurs abscisses en deux sous-groupes de même effectif. La droite passant par les points moyens des deux sous-groupes de points est appelée droite de Mayer et fournit un ajustement satisfaisant ;
- la méthode des moindres carrés (voir après).

Exercice 18. On reprend les données de l'exercice 15.

1. Tracer au jugé une droite qui ajuste au mieux le nuage en passant par le point moyen. Estimer alors à quelle superficie certifiée de production biologique on aurait pu s'attendre en 2015.
2. Déterminer l'équation de la droite de Mayer associée à cette série statistique. Estimer alors par le calcul à quelle superficie certifiée de production biologique on aurait pu s'attendre en 2015.

Méthode des moindres carrés :
Soit d la droite d'équation $y = ax + b$. Pour chaque i allant de 1 à N, on définit l'écart entre le point $M_i(x_i ; y_i)$ et la droite d par le nombre $e_i = y_i - (ax_i + b)$.



Définition 19. La **méthode des moindres carrés** est la méthode qui permet de minimiser la somme des carrés des écarts entre chaque point du nuage de points et la droite, c'est-à-dire à trouver a et b tels que la somme $\sum_{i=1}^{N} e_i^2$ soit minimale

Soit U la fonction à deux variables définie par $U(a,b) = \sum_{i=1}^{N} e_i^2 = \sum_{i=1}^{N} (y_i - (ax_i + b))^2$

Puisqu'on cherche un minimum, on $\dfrac{\partial U}{\partial a} = \dfrac{\partial U}{\partial b} = 0$.

La condition $\dfrac{\partial U}{\partial b} = 0$ implique $\overline{y} - a\overline{x} - b = 0$ soit $b = \overline{y} - a\overline{x}$.

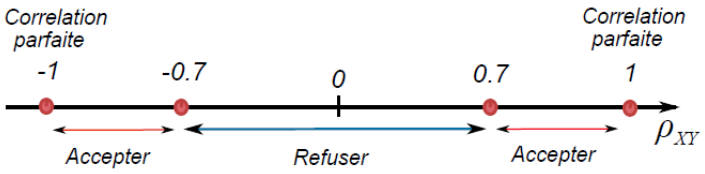
Dans ce cas, $U(a,b) = \sum_{i=1}^{N} (y_i - \overline{y} - a(x_i - \overline{x}))^2$

La condition $\dfrac{\partial U}{\partial a} = 0$ implique $\sum_{i=1}^{N} -2(x_i - \overline{x})(y_i - \overline{y} - a(x_i - \overline{x})) = 0$ soit $a = \dfrac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(X)}$.

Propriété 3. La droite d'ajustement obtenue par la méthode des moindres carrés a pour équation $y = ax + b$, avec $a = \dfrac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(X)}$ et $b = \overline{y} - a\overline{x}$.

Remarque 10. Le coefficient de corrélation ρ_{XY} permet de justifier l'ajustement linéaire. On adopte les critères suivants :

- Si $|\rho_{XY}| < 0,7$ alors l'ajustement linéaire est refusé (droite refusée).
- Si $|\rho_{XY}| \geqslant 0,7$ alors l'ajustement linéaire est accepté (droite acceptée).



Exercice 19. On reprend les données de l'exercice 15.

1. Déterminer l'équation de la droite d'ajustement obtenue par la méthode des moindres carrées.
2. Estimer par le calcul la superficie certifiée de production biologique à laquelle on aurait pu s'attendre en 2015 et comparer avec les résultats obtenus précédemment.

Exercice 20. Compléter le module `statistiques` en ajoutant deux fonctions permettant de calculer la covariance et le coefficient de corrélation.

3.4 Exercices

Exercice 21.

Ajouter au module `statistiques` une fonction Python qui prend en arguments deux listes `X` et `Y`, et qui trace sur un graphique :

- le nuage de points ;
- le point moyen ;
- la droite d'ajustement obtenue par la méthode des moindres carrés.

En titre du graphique, on fera apparaître l'équation de la droite d'ajustement ainsi que le coefficient de corrélation.

Tester avec les données de l'exercice 15.

Exercice 22. Un couple envisage de louer son appartement dans quelques années. Il s'intéresse au montant mensuel des loyers dans sa région pour un appartement équivalent au sien.
Ces montants sont donnés dans le tableau ci-dessous :

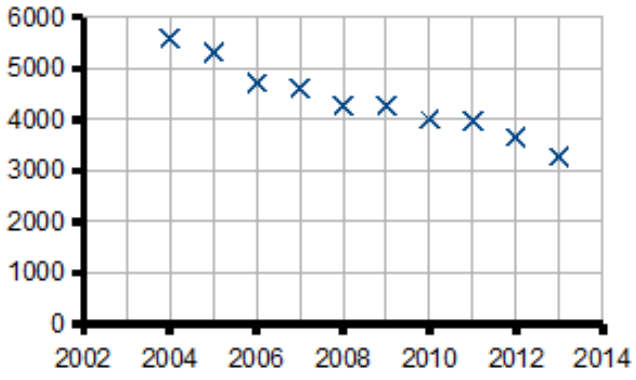
Année	2012	2013	2014	2015	2016	2017
Rang : x_i	0	1	2	3	4	5
Loyer mensuel (en euros) : y_i	610	612	619	628	634	640

où x_i désigne le rang de l'année mesuré à partir de l'année 2012 et y_i le montant mensuel moyen du loyer (en euros) des appartements entre 2012 et 2017.

1. Donner le coefficient de corrélation linéaire de la série statistique $(x_i\ ;\ y_i)$ arrondi au millième et expliquer pourquoi ce résultat permet d'envisager un ajustement affine.
2. Donner l'équation de la droite de régression de y en x sous la forme $y = ax + b$, où a et b sont à arrondir au centième.
3. On décide d'ajuster le nuage de points de cette série statistique $(x_i\ ;\ y_i)$ par la droite d'équation $y = 6,4x + 608$.
   (a) Déterminer le montant du loyer mensuel que peut espérer ce couple en 2020.
   (b) En quelle année le couple peut-il espérer louer son appartement plus de 700 € ?

Exercice 23. On s'intéresse à l'évolution du nombre de tués dans des accidents de la circulation depuis 2004 en France. On considère le tableau et le graphique suivants qui représentent le nombre d'accidents corporels en fonction des années :

Année	Nombre de tués
2004	5 593
2005	5 318
2006	4 709
2007	4 620
2008	4 275
2009	4 273
2010	3 992
2011	3 963
2012	3 653
2013	3 268

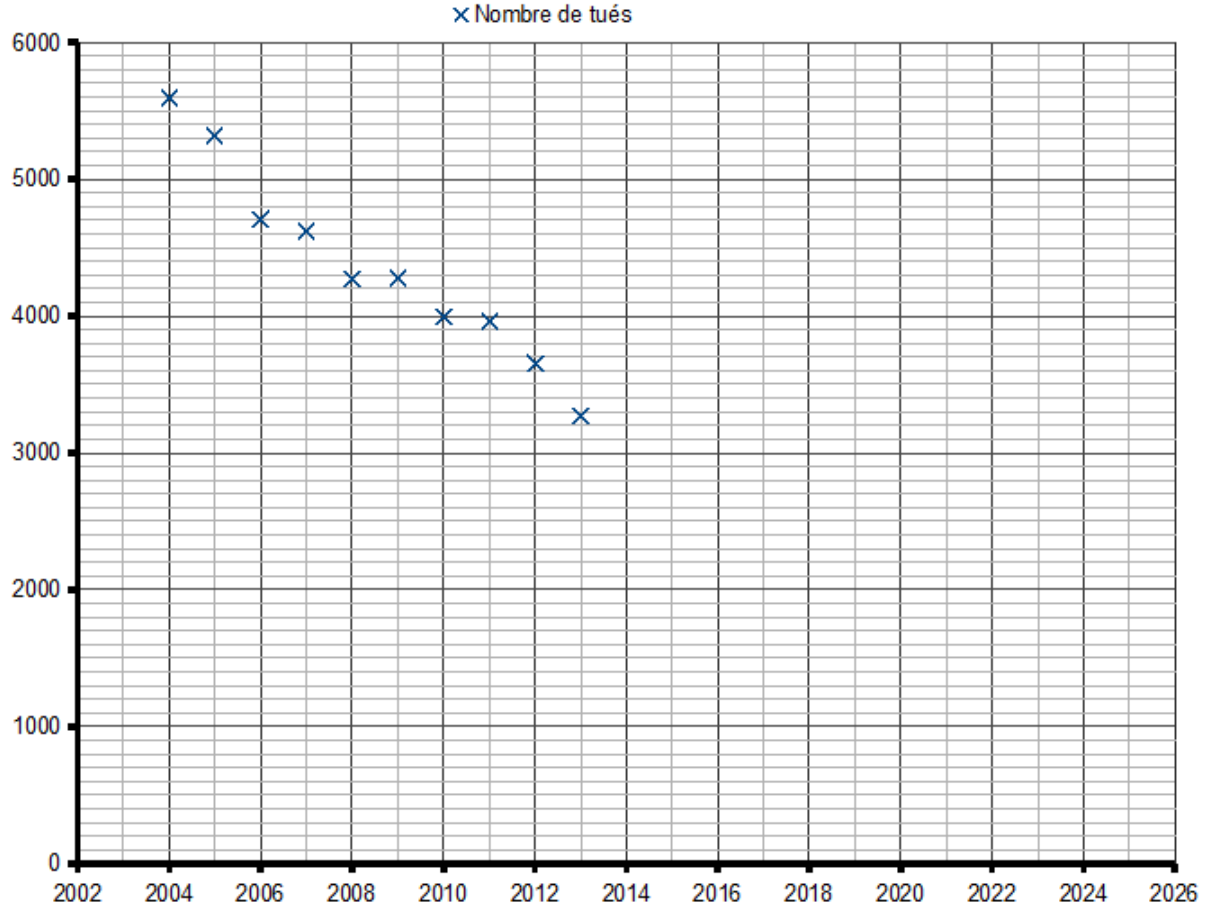


Source : https://www.onisr.securite-routiere.gouv.fr/

1. En 2013, on a décidé de se fixer comme objectif à l'horizon 2025 une poursuite de la baisse de la mortalité dans les mêmes conditions.
   (a) Déterminer une équation de la droite d'ajustement du nuage de points selon la méthode des moindres carrés.
   (b) À l'aide de ce modèle, déterminer graphiquement quelle prévision on peut faire sur le nombre de tués en 2025.
2. En réalité, entre 2014 et 2021, le nombre de tués est donné par le tableau suivant :

Année	2014	2015	2016	2017	2018	2019	2020	2021
Nombre de tués	3 384	3 461	3 477	3 448	3 248	3 244	2 541	2 944

(a) Compléter le graphique suivant avec les données de ce tableau.



(b) Comment peut-on expliquer la baisse exceptionnelle du nombre de tués en 2020 ?

(c) Que peut-on penser du modèle choisi dans la question 1. ?

(d) À l'aide d'un nouveau modèle de son choix, déterminer une nouvelle prévision du nombre de tués en 2025.

Exercice 24. Le tableau suivant indique la teneur de l'air en dioxyde de carbone (CO_2), observée depuis le début de l'ère industrielle.
Dans le tableau ci-dessous, x_i représente le rang de l'année et y_i la teneur en CO_2 exprimée en parties par million (ppm).

Année	1850	1900	1950	1990
Rang de l'année x_i	0	50	100	140
Teneur en CO_2 y_i	275	290	315	350

On veut modéliser cette évolution par une fonction dont la courbe est voisine du nuage de points. Plusieurs types de fonctions semblent utilisables.

1. Représenter dans un repère le nuage de points associé à la série statistique $(x_i\ ;\ y_i)$.
2. Modélisation par une fonction affine
   (a) À l'aide d'une calculatrice, donner le coefficient de corrélation linéaire, arrondi au centième, de la série $(x_i\ ;\ y_i)$.
   (b) À l'aide d'une calculatrice, donner une équation de la droite de régression de y en x par la méthode des moindres carrés, sous la forme $y = ax + b$, avec a arrondi au centième et b à l'unité. Représenter cette droite dans le repère ci-dessus.
   (c) Selon ce modèle, quelle teneur en CO_2 peut-on prévoir en 2030 ?

3. Modélisation par une fonction f définie par $f(x) = 250 + B\mathrm{e}^{Ax}$.
   On pose $z_i = \ln\left(y_i - 250\right)$. On admet que la série $(x_i\ ;\ z_i)$ a pour coefficient de corrélation linéaire 0,999 et qu'une équation de la droite de régression de z en x par la méthode des moindres carrés est : $z = 0,01x + 3,2$.
   (a) Selon ce modèle, quelle teneur en $C0_2$ peut-on prévoir en 2030 ?
   (b) Donner une équation de la courbe d'ajustement de y en x, sous la forme $y = f(x) = 250 + B\mathrm{e}^{Ax}$, avec A arrondi au centième et B à l'unité.
   (c) En déduire des valeurs approchées décimales arrondies à l'unité près de $f(0)$, $f(50)$, $f(100)$, $f(140)$.
4. Laquelle des deux prévisions de la teneur en CO_2 pour 2010 vous semble la plus plausible ? Pourquoi ?

Exercice 25. Le tableau ci-dessous donne l'évolution des ventes de vélos à assistance électrique en France entre 2007 et 2017.

Année	2007	2009	2011	2013	2015	2017
Rang de l'année : x_i	0	2	4	6	8	10
Nombre de vélos à assistance électrique vendus (en milliers) : n_i	10	23	37	57	102	278

Données : Observatoire du Cycle

1. (a) Représenter dans un repère le nuage de points associé à la série statistique $(x_i\ ;\ n_i)$.
   (b) Expliquer pourquoi un ajustement affine ne semble pas envisageable.
2. On pose $y_i = \ln(x_i)$. Compléter au centième près le tableau suivant.

Année	2007	2009	2011	2013	2015	2017
Rang de l'année : x_i	0	2	4	6	8	10
Nombre de VAE vendus (en milliers) : n_i	10	23	37	57	102	278
$y_i = \ln(n_i)$	2,3					

3. Donner une équation de la droite de régression de y en x (on arrondira les coefficients au dixième).
4. En déduire une équation de la courbe d'ajustement de n en x.
5. Si l'évolution se poursuit de la même façon, quel devrait être, en milliers, le nombre de vélos à assistance électrique vendus en France en 2023 ?