Chapitre # (PS) 3 Statistiques Descriptives

Il existe trois types de mensonges : les mensonges simples, les sacrés mensonges et les statistiques.

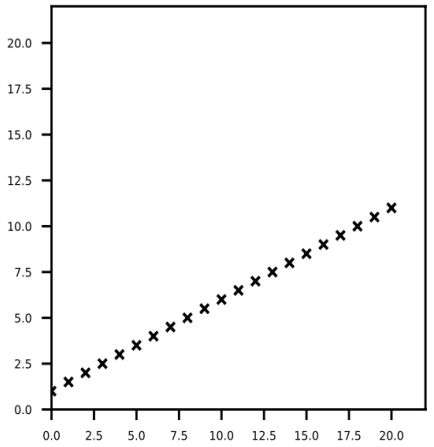
— Mark Twain

Résumé & Plan

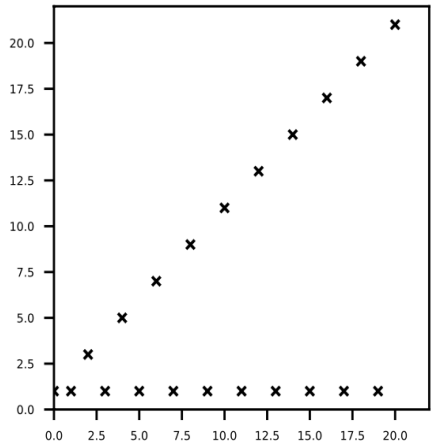
Les statistiques sont présentes dans beaucoup de domaines en Sciences, notamment dans l'exploitation de grosses quantités de données. Il existe plusieurs types de statistiques : vous en étudierez deux en BCPST. Les statistiques descriptives d'une part (en première année) dont le but est d'étudier des séries de données et d'en dégager des caractéristiques (objectif du présent chapitre), d'autre part les statistiques inférentielles dont l'objectif est de savoir si des données semblent provenir ou non de réalisations d'une certaine variable aléatoire (seront étudiées en 2ème année).

- Les énoncés importants (hors définitions) sont indiqués par un ♥.
- Les énoncés et faits à la limite du programme, mais très classiques parfois, seront indiqués par le logo [H.P] . Si vous souhaitez les utiliser à un concours, il faut donc en connaître la preuve ou la méthode mise en jeu. Ils doivent être considérés comme un exercice important.
- Les preuves déjà tapées sont généralement des démonstrations non exigibles en BCPST, qui peuvent être lues uniquement par les curieuses et curieux. Nous n'en parlerons pas en cours.

Jouons à un petit jeu pour commencer : nous avons tracé ci-dessous des points provenant de mesures expérimentales. Pouvez-vous deviner si les courbes en question sont des droites ou non (réponses page suivante) ?

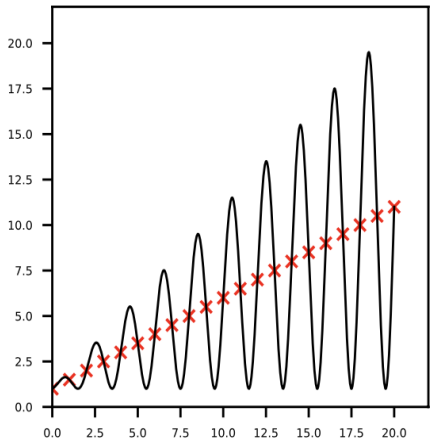


Est-ce une droite ?

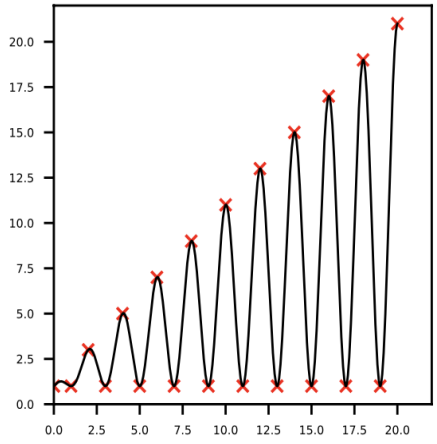


Est-ce une droite ?

Lycée Camille Guérin – Poitiers



Pas du tout.



Non plus!

Que nous apprend ce « jeu » ? Qu'il n'y a que deux façons de répondre à la question « la courbe est-elle une droite ? » à partir de l'observation de certains de ses points :

1. Si les points ne sont pas alignés, on peut être certain (si on a confiance en ses mesures...) qu'il ne s'agit pas d'une droite.
2. Si les points sont alignés, il est possible que la courbe soit une droite mais rien ne permet de l'affirmer.

C'est là tout le problème de l'expérimentation scientifique : si le résultat d'une expérience contredit le modèle, on peut en déduire que le modèle est faux/imparfait et qu'il faut le corriger. En revanche, si le résultat de l'expérience satisfait au modèle, on ne peut pas affirmer que le modèle est juste (un exemple n'est pas une preuve) mais seulement qu'il est cohérent avec ce qui a été observé jusqu'à présent...

1. INTRODUCTION

1.1. Statistique versus probabilité ?

Quelles sont les différences entre probabilités et statistiques ? Il est fréquent, même pour des mathématiciens confirmés, de faire la confusion entre les deux et ce pour deux raisons. Premièrement les outils mathématiques sont les mêmes et un problème concret ne relève jamais exclusivement d'un domaine ou de l'autre. Pour le traiter, on doit jongler en permanence entre probabilités et statistiques. Deuxièmement, si les deux domaines ont leur propre vocabulaire afin justement de permettre la distinction, on utilisera dans la pratique indifféremment un terme probabiliste ou statistique. On le fait par commodité car ceci permet d'alléger grandement le discours, et c'est le contexte qui déterminera s'il s'agit de probabilité ou de statistique.

La différence entre ces deux branches des mathématiques s'explique en terme d'objectif. Pour le comprendre étudions le petit jeu suivant : un joueur parie sur le résultat du lancé d'un dé à 6 faces. S'il annonce le bon résultat, il gagne la valeur de la face en euros, s'il perd il ne gagne rien du tout. Quelle stratégie le joueur doit-il adopter pour maximiser ses gains ?
Il s'agit de faire la remarque suivante : quelle que soit la face qu'il annonce, il a toujours une chance sur six de gagner. En revanche, les gains en cas de bonne réponse dépendent de la face : s'il annonce 1 il a une chance sur six de gagner 1 euros, s'il annonce 2 il a une chance sur six de gagner 2 euros etc...On devine ainsi que le joueur à tout intérêt à parier sur le 6 : il gagnera aussi souvent qu'avec une autre face mais gagnera plus. Mathématiquement cela s'écrit en terme **de gain moyen**, c'est-à-dire les chances de gagner multipliées par le gain, cela représente ce que le joueur gagne en moyenne par partie. On voit alors que le gain moyen en pariant sur le 6 est de $1/6 * 6 = 1$ euros tandis que les gains moyens sur les autres faces sont plus faibles : on vient de prouver que jouer le 6 est la meilleure stratégie. Le joueur vient de faire des **probabilités** .

Le raisonnement se tient mais on part tout de même d'une hypothèse : toutes les faces ont la même probabilité de sortir, à savoir $1/6$, mais est-ce bien le cas ? Pour le vérifier, le joueur lance avant de commencer à parier 100 fois le dé et remarque que le 6 ne sort pas une seule fois et que le 4 sort une fois sur deux... il se dit donc naturellement que le dé n'est pas équilibré et que le 4 a une chance sur deux de sortir tandis que le 6 n'a aucune chance de sortir, il vient de faire des **statistiques**. Il en déduit donc qu'il vaut mieux jouer le 4 (gain moyen de $1/2 * 4 = 2$ euros) que le 6 (gain moyen de $0 * 6 = 0$ euros), il vient de refaire des probabilités.

La connaissance des chances de sortie d'une face s'appelle la **loi de probabilité** du dé.

Faire des probabilités, c'est utiliser cette loi pour faire des calculs de chance, déterminer des stratégies etc... Faire des statistiques c'est essayer de déterminer cette loi à partir de quelques observations, c'est-à-dire à partir d'une information partielle. On devine alors que, comme pour le problème de la nature des courbes données en introduction, on ne pourra pas donner des réponses définitives : s'il est possible de rejeter certaines hypothèses, il est bien plus difficile, pour ne pas dire impossible, de garantir la validité de celles que nous conserverons.

Dans la pratique, on fait sans cesse des allers-retours entre statistiques et probabilités : on commence par observer un phénomène et en déduire quelques informations sur la loi (stat.), on en déduit quelques propriétés (proba.), on vérifie que ces propriété se réalisent bien (stat.) etc...

1.2. Vocabulaire

Une **population** (statistique) est un ensemble fini d'éléments. Les éléments de la population sont appelés ses **individus**.

Pour réaliser une étude statistique sur une population, on recueille auprès de chaque individu de la population la « valeur » d'une propriété appelée **caractère** (ou variable statistique), objet de l'étude.

La population étant généralement trop importante pour pouvoir recueillir les données de tous ses individus, on travaille souvent sur un **échantillon**, c'est-à-dire un sous-ensemble de la population. Une des difficultés de la statistique est, pour que l'étude soit pertinente, de s'assurer que l'échantillon choisi soit **représentatif** de la population.

On appelle alors **effectif total** la taille de l'échantillon étudié (son cardinal).

On appelle **série statistique** l'ensemble des mesures du caractère effectuées sur les individus de l'échantillon de population. Enfin, le caractère étudié est dit :

- **quantitatif** lorsqu'il est numérique (taille d'une personne, nombre d'enfant(s), durée d'une réunion, coordonnées d'un point d'une courbe, notes à l'épreuve de mathématiques du baccalauréat, concentration d'une solution, ...) ; Un caractère quantitatif est **discret** s'il prend un nombre fini de valeurs (valeurs isolées) ou **continu** s'il peut prendre toutes les valeurs réelles entre deux limites.
- **qualitatif** sinon (couleur des cheveux, sexe des nouveaux nés, nature d'un traitement médicamenteux, ...).

L'objet de ce chapitre est de donner un résumé (graphique ou numérique) le plus complet possible des informations apportées par une série statistique.

2. STATISTIQUE À UNE VARIABLE (UNIVARIÉE)

2.1. Distribution d'une série statistique quantitative

SÉRIE STATISTIQUE, MODALITÉ. L'observation d'un caractère quantitatif sur $n \in \mathbb{N}^\star$ individus se traduit par un relevé de mesures sous la forme d'une liste $x = (x_1, \dots, x_n)$ de données brutes. Cette liste est une **série statistique**.

Exemple 1 On relève les pointures des garçons d'une classe contenant 32 garçons : $x = (\underbrace{41, 43, 41, 37, \dots, 40}_{\text{32 valeurs}})$

Définition 1 | Opérations sur les séries statistiques

- Soit une série statistique $x = (x_1, x_2, \dots, x_n)$ et une fonction $f : \mathbb{R} \to \mathbb{R}$, on définit alors la série statistique $f(x)$ comme étant $(f(x_1), f(x_2), \dots, f(x_n))$.
- De même si $x = (x_1, x_2, \dots, x_n)$ et $y = (y_1, y_2, \dots, y_n)$ sont deux séries de même taille, on notera $x + y$ la série $(x_1 + y_1, x_2 + y_2, \dots, x_n + y_n)$, xy la série $(x_1 y_1, x_2 y_2, \dots, x_n y_n)$. *(Ce type de notations permettra de simplifier l'écriture de certaines proposition mathématiques dans ce chapitre.)*

Par exemple, si $x = (0, 1, 2, 3, 3, 4, 5, 5)$ alors $x^2 = (0, 1, 4, 9, 9, 16, 25, 25)$.

Définition 2 | Modalités

On appelle *modalités* d'un caractère les valeurs possibles qu'il peut prendre.

Exemple 2 **(Diamètres de pièces)** Le tableau ci-dessous regroupe les diamètres en cm de 48 pièces prélevées dans la production d'une machine.

19	26	23	20	22	24	20	24
22	20	21	19	21	22	19	20
21	21	22	21	23	22	21	24
25	23	22	19	20	26	24	25
23	26	25	25	21	22	25	24
23	22	24	24	25	23	25	22

Il s'agit donc d'un échantillon statistique de taille 48 dans la population des pièces fabriquées par la machine. Le caractère étudié est le diamètre de la pièce en centimètre.

Les modalités sont : 19, 20, 21, 22, 23, 24, 25 et 26.

Effectifs, fréquences (cas discret)

Définition 3 | Effectif (cumulé), fréquence (cumulée)

Les valeurs prises par une série statistique $x = (x_1, x_2, \dots, x_n)$ appartiennent à un ensemble fini $\{v_1, v_2, \dots, v_p\}$ où $v_1 < v_2 < \dots < v_p$ (avec $(p \leqslant n)$. On rappelle que les nombres réels distincts $v_1, v_2, \dots, v_p$ sont appelés les modalités de la série statistique.

Alors on note, pour tout $i \in [\![1; p]\!]$:

- n_i l'**effectif** de la modalité v_i : n_i est le nombre de x_j vérifiant $x_j = v_i$;
- l'effectif total de l'échantillon est alors $n = n_1 + \dots + n_p$;
- N_i l'**effectif cumulé croissant** de la modalité v_i : $\mathrm{N}_i = \sum_{k=1}^{i} n_k$ est le nombre de x_j vérifiant $x_j \leqslant v_i$;
- f_i la **fréquence** de la modalité v_i : $f_i = \dfrac{n_i}{n}$
- F_i la **fréquence cumulée croissante** de la modalité v_i : $\mathrm{F}_i = \sum_{\ell=1}^{i} f_k$ est la fréquence du nombre de x_j vérifiant $x_j \leqslant v_i$

Exemple 3

Sur une parcelle de soja, on a mesuré la hauteur en cm de 100 plantes à l'âge de 6 semaines.

Les résultats obtenus sont les suivants :

Hauteurs en cm	36	37	38	39	40	41
Effectifs	6	11	26	32	14	11
Fréquences						
Fréquences cumulées						

Déterminer l'effectif total de cette série et compléter le tableau précédent.

Exemple 4 On a mesuré la pointure de 32 garçons dans une classe. Compléter la colonne des effectifs cumulés.

Garçons	Pointure		
xi(valeur)	ni(effectif)	fi(fréquence)	Effectifs cumulés
40	2	0,06	
41	3	0,09	
42	6	0,19	
43	11	0,34	
44	3	0,09	
45	2	0,06	
46	3	0,09	
47	2	0,06	
total	32	1	

Effectifs, fréquences (cas continu). Parfois le nombre de modalités est trop grand, voire infini pour des modalités dites continues (c'est-à-dire à valeurs réelles). Il est alors nécessaire de regrouper les modalités en classes disjointes, le plus souvent en des intervalles qui ne sont pas forcément de tailles égales. Inversement les modalités non regroupées en classe sont dites *ponctuelles*.

Exemple 5 Par exemple, lorsque l'on étudie la démographie urbaine française on peut regrouper les communes en classes selon leur nombre d'habitants :

- Hameaux de 1 à 99 habitants,
- Village de 100 à 1999 habitants,
- Ville de 200 à 99999 habitants,
- Agglomération à partir de 100000 habitants.

Définition 4 | Effectif et fréquence : cas continu

Soit $x = (x_1, \dots, x_n)$ une série statistique.
On regroupe les données $x_1, x_2, \dots, x_n$ dans p intervalles
$I_1 = [v_0, v_1[, I_2 = [v_1, v_2[, \dots, I_p = [v_{p-1}, v_p[$ appelés **classes**.
Ces classes doivent former une partition de l'ensemble des mesures x_i. Plus précisément :

- toutes les mesures doivent appartenir à une classe
- une mesure ne peut appartenir qu'à une seule classe à la fois.

Alors on note, de la même façon que précédemment, pour tout $i \in [\![1; p]\!]$:

- n_i l'**effectif** de la classe I_i : n_i est le nombre de x_j vérifiant $v_{i-1} \leqslant x_j < v_i$;
- N_i l'**effectif cumulé croissant** de la classe I_i : $N_i = \sum_{k=1}^{i} n_k$ est le nombre de x_j vérifiant $v_0 \leqslant x_j < v_i$;
- f_i la **fréquence** de la classe I_i : $f_i = \dfrac{n_i}{n}$
- F_i **la fréquence cumulée croissante** de la classe I_i : $F_i = \sum_{k=1}^{i} f_k$ est la fréquence du nombre de x_j vérifiant $v_0 \leqslant x_j < v_i$

On résume alors la série statistique à l'aide d'un tableau identique au précédent (où les classes figurent en lieu et place des modalités).

Exemple 6 On considère une série statistique obtenue à partir d'un échantillon d'enfants de 7 ans.

Masses (kg)	[18,21[	[21,24[	[24,26[	[26,28[	[28,31[	[31,35[
Effectifs	6	12	47	30	12	3
Effectifs cumulés						
Fréquences						

Déterminer l'effectif total de cette série et compléter le tableau précédent.

2.2. Représentations graphiques

Diagramme en bâtons : caractère quantitatif discret. Dans le cas discret, la façon la plus simple de montrer la répartition par effectifs d'une série selon les modalités du caractère (qualitatif ou quantitatif) consiste à bâtir un diagramme en bâtons, suite de segments parallèles à l'axe des ordonnées dont les hauteurs sont proportionnelles aux effectifs (et donc aux fréquences). Par exemple, pour l'exemple 3 précédent, on obtient (pour la hauteur des plantes) :

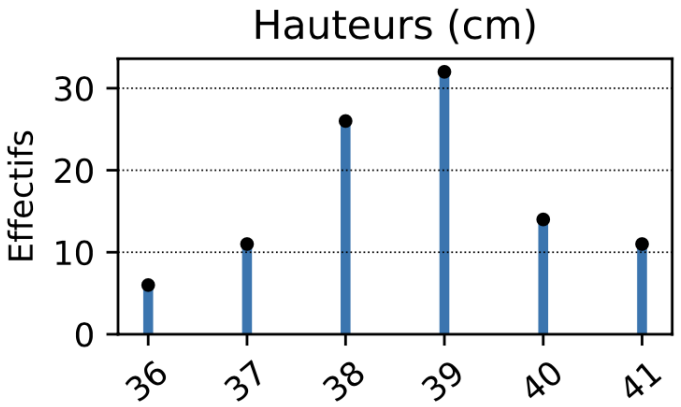
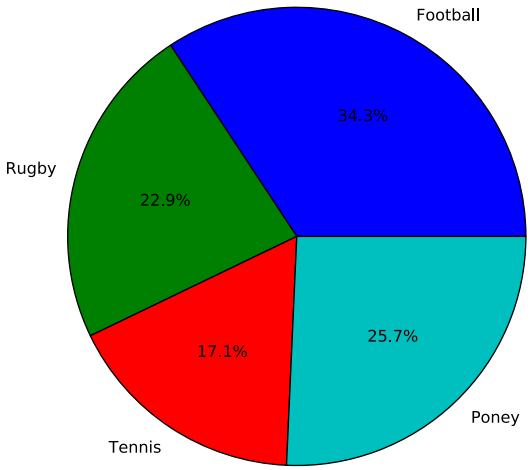
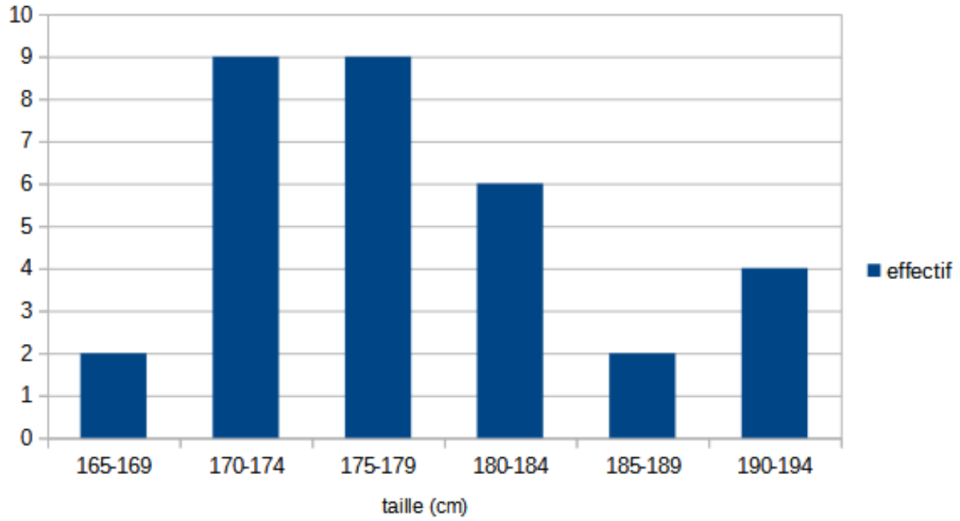


Diagramme circulaire : caractère qualitatif. Pour un caractère qualitatif, il est préférable de représenter les résultats sous la forme d'un **diagramme circulaire**.

Exemple 7 Dans une classe, chaque élève pratique un unique sport. 12 élèves pratiquent le football, 9 l'équitation, 8 le rugby et 6 le tennis. Le diagramme circulaire correspondant est :



HISTOGRAMME. Lorsqu'un caractère présente beaucoup de valeurs x_i, chaque valeur étant peu représentée, on effectue des regroupements en classes statistiques (intervalles). On peut alors représenter la distribution par un histogramme, c'est-à-dire des rectangles ayant pour base les intervalles choisis et d'aire proportionnelle à l'effectif n_i. Si on choisit des intervalles de même longueur, on obtient alors des rectangles dont la hauteur représente l'effectif. Ci-après un exemple.



POLYGÔNE DES EFFECTIFS CUMULÉS.

Définition 5

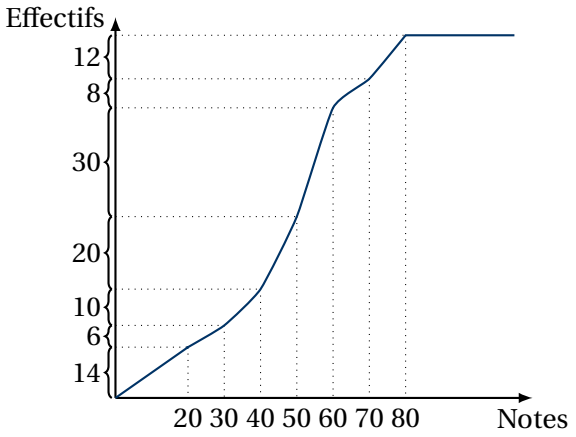
Lorsque les valeurs peuvent être ordonnées, on peut définir les fréquences cumulées en ajoutant à chaque fréquence les fréquences des valeurs précédentes.

- Dans le cas discret, le **polygône des effectifs cumulés** est la ligne brisée qui relie les points ayant pour abscisse la valeur du caractère, et en ordonnée son effectif.
- Dans le cas d'un regroupement par classes, on réalise une interpolation linéaire.

Exemple 8 **(Cas continu)** Considérons le tableau d'effectifs des notes, regroupées en classe, d'un devoir noté sur 100 suivant :

Notes en classe	[0,20[	[20,30[	[30,40[	[40,50[	[50,60[	[60,70[	[70,80[
Effectifs	14	6	10	20	30	8	12

On peut le représenter sous forme de polygone des fréquences cumulées.



1. Quel est l'effectif total ?

2. Lire graphiquement le nombre approximatif de notes inférieures à 30.

2.3. Paramètres de position

Les paramètres de position sont des outils numériques permettant d'avoir des informations sur le comportement global d'une série statistique x. Les principaux paramètres sont :

LE MODE. C'est, moralement, la valeur de la série statistique de plus grand effectif

Définition 6 | Mode & Classe modale

- On appelle *mode* d'une série statistique x toute modalité de x dont l'effectif est maximal parmi les effectifs de toutes les modalités.
- Lorsque les modes correspondent à des classes, on appelle alors *classe modale* la classe dont l'effectif est maximal.

Attention

Si vos classes sont de tailles différentes alors la classe modale n'est pas forcément la classe qui a le « plus haut » rectangle dans l'histogramme, mais plutôt celui qui a la plus grande aire.

Remarque 1 Il est possible qu'une série statistique admette plusieurs modes ou classes modales.

Exemple 9 Dans l'exemple des pointures des garçons, déterminer le mode de la série statistique.

La moyenne.

Définition 7 | Moyenne

Soit $x = (x_1, \dots, x_n)$ une série statistique. La *moyenne* de la série, notée $\overline{x}$, est définie par : $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$.

Remarque 2 Dans le cas continu, on remplace les classes du caractère par leur valeur milieu.

Proposition 1 | Propriétés de la moyenne (Linéarité)

Soient $x = (x_1, \dots, x_n)$ et $y = (y_1, \dots, y_m)$ deux séries statistiques quantitatives.

- **[Affine]** Soit $(a, b) \in \mathbb{R}^2$, alors : $\overline{ax + b} = a\overline{x} + b$.
- **[Somme]** Supposons ici que les deux séries statistiques sont de même longueur, alors : $\overline{x + y} = \overline{x} + \overline{y}$.

Preuve Immédiat par la linéarité de la somme.

Exemple 10 Dans l'exemple des pointures des garçons, déterminer la moyenne de la série statistique.

La médiane.

La médiane d'une série statistique permet de la « couper en deux ».

Définition 8 | Médiane

La **médiane** d'une série statistique permet de couper la population étudiée en deux groupes de même taille de la façon suivante : 50% de la population a un caractère inférieur à la médiane, et 50% de la population a un caractère supérieur à la médiane.

Selon le type de données, la détermination de la médiane sera différente.

1er cas : cas de données brutes. Lorsque la série est sous forme brute, il faut commencer par ordonner les valeurs.

- Si l'effectif total est impair, par exemple pour la série $(0, 1, 3, 3, 4, 7, 10)$, il n'y a pas d'ambiguité dans la définition de la médiane, celle-ci vaut :

- En revanche, si l'effectif total est pair, par exemple pour la série $(0, 1, 3, 3, 4, 5, 7, 10)$, on parle d'intervalle médian :

La médiane est alors définie comme la moyenne des bornes de l'intervalle médian.

Exemple 11 Dans l'exemple des pointures des garçons, déterminer la médiane de la série statistique.

2nd cas : données regroupés par classes. On peut alors utiliser judicieusement le polygone des effectifs cumulés.

Exemple 12 Déterminer une estimation de la médiane dans l'Exemple 8 (notes sur 100) :

LES QUARTILES. Les **quart**iles permettent de partager en **quatre** une série statistique.

Définition 9 | Quartiles

- Le premier quartile d'une série, noté Q_1, est la plus petite valeur pour laquelle au moins 25% des données sont inférieures ou égales à Q_1.
- Le troisième quartile d'une série, noté Q_3, est la plus petite valeur pour laquelle au moins 75% des données sont inférieures ou égales à Q_3.

Remarque 3 Le *deuxième quartile* Q_2 est défini comme étant la médiane.

Exemple 13 Pour la pointure des garçons, déterminer Q_1 et Q_3.

2.4. Paramètres de dispersion

Les paramètres de position ne sont généralement pas suffisants lorsque l'on veut analyser plus finement une série statistique. Si on prend par exemple les séries statistiques :

1. des notes d'un premier DS de math où tous les élèves ont eu 10 ;
2. des notes d'un deuxième DS de math où 20% élèves ont eu 0, 60% 10 et 20% 20.

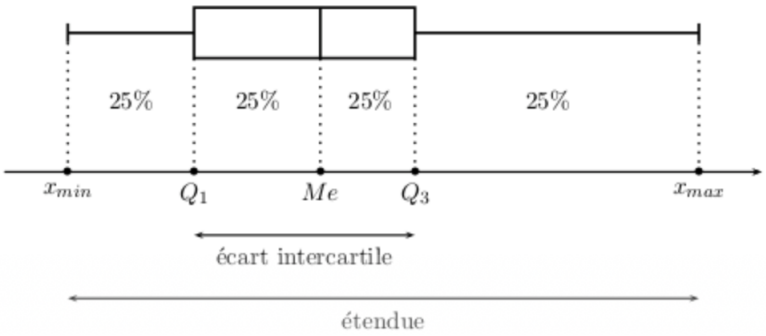
Alors tous les paramètres de position de ces deux séries coïncident. Pour les distinguer, il faut considérer les paramètres de dispersion, i.e. des outils qui permettent de quantifier l'écartement de la série autour de ses paramètres de position.

ETENDUE. Si $x_{\min}$ et $x_{\max}$ sont les valeurs extrêmes de la série, alors l'étendue est $E = x_{\max} - x_{\min}$.

Exemple 14 Déterminer l'étendue pour l'exemple de la série statistique des pointures des garçons.

L'ÉCART INTERQUARTILE. L'**écart interquartile** est la différence $Q_3 - Q_1$. *(L'intervalle interquartille* $[Q_1, Q_3]$ *contient donc* 50% *de l'effectif total.)*

Une « boîte à moustache » (ou diagramme de Tuckey) permet de représenter graphiquement la série statistique.



Exemple 15 Construire la boîte à moustache pour la série statistique des pointures des garçons.

DÉCILES. Les **déciles** sont les valeurs qui permettent de partager la série en 10 groupes de même effectif. On note généralement ces valeurs $D_1, \dots, D_9$. *(On remarque que D_5 est la médiane de la série)*

L'**écart terdécile** est la différence $D_9 - D_1$. *(L'intervalle interdécile $[D_1; D_9]$ contient donc 80% de l'effectif total.)*

VARIANCE. La variance permet de mesurer la dispersion d'une série statistique par rapport à sa moyenne

Définition 10 | Variance & Écart-Type

- La *variance* d'une série statistique quantitative à valeurs réelles $x = (x_1, x_2, \dots, x_n)$, est le nombre $\mathbb{V}_x$ défini par :
$$\mathbb{V}_x = \overline{x - \overline{x}^2} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2.$$
- L'*écart-type* d'une telle série, noté σ_x, est défini par :
$$\sigma_x = \sqrt{\mathbb{V}_x}.$$

$\mathbb{V}_x$ est la moyenne des carrés des écarts à la moyenne donc est toujours positive, d'où la bonne définition de l'écart-type.

Proposition 2 | Variance nulle

Soit une série statistique quantitative à valeurs réelles $x = (x_1, x_2, \dots, x_n)$ alors :
$$\mathbb{V}_x = 0 \iff \forall i \in [\![1\,,\,n]\!], \quad x_i = x_0.$$

Preuve Se démontre rapidement :

$$\begin{aligned}\mathbb{V}_x = 0 &\iff \sum_{i=1}^{n}(x_i - \overline{x})^2 = 0 \qquad \text{somme de termes positifs}\\ &\iff \forall i \in [\![1\,,\,n]\!], \quad x_i = \overline{x}\\ &\iff \forall i \in [\![1\,,\,n]\!], \quad x_i = x_0.\end{aligned}$$

C'est ce qu'on voulait.

Proposition 3 | KÖNIG-HUYGENS

Soit x une série statistique quantitative à valeurs réelles. Alors :
$$\mathbb{V}_x = \overline{x^2} - \overline{x}^2.$$

Preuve

Remarque 4 (Interprétation)

- Plus la variance est grande, plus la série s'éloigne de sa moyenne, et plus la série est donc « étalée ». Inversement, plus la variance est proche de zéro et plus la série est concentrée autour de sa moyenne.
- La variance ne donne pas d'informations sur une éventuelle asymétrie de la série.

Remarque 5 (Homogénéité) L'intérêt de l'écart-type par rapport à la variance est que l'écart-type s'exprime dans les mêmes unités que les modalités de la série.

Proposition 4 | Propriétés de la variance

Soit x une série statistique quantitative réelle, $(a, b) \in \mathbb{R}^2$ et y la série statistique $y = ax + b$. Alors : $\mathbb{V}_y = a^2\mathbb{V}_x$, $\sigma_y = |a|\,\sigma_x$.

Preuve En utilisant KÖNIG-HUYGENS et la linéarité de la moyenne :

$$\begin{aligned}\mathbb{V}_{ax+b} &= \overline{(ax+b)^2} - (\overline{ax+b})^2\\ &= \overline{a^2x^2 + 2abx + b^2} - (a\bar{x} + b)^2\\ &= a^2\overline{x^2} + 2ab\bar{x} + b^2 - a^2(\bar{x})^2 - 2ab\bar{x} - b^2\\ &= a^2\left(\overline{x^2} - (\bar{x})^2\right)\\ &= a^2\,\mathrm{V}(x)\end{aligned}$$

Exemple 16 Pour x la série statistique donnant la pointure des garçons, on a $V_x \simeq 3,3$ (ce qui est une « plutôt grande » dispersion.

On suppose dans la suite que toutes les données d'une série à nombre de modalités fini sont contenues dans une liste `L` donnée en paramètre. Voici la liste des principales fonctions à comprendre pour les statistiques univariées.

Pour avoir la liste des modalités, il suffit de créer une nouvelle liste sans doublon.

(Modalités)

```
def sans_doublon(L):
    """
    renvoie la liste des éléments de L, chaque élément \
    ↪ apparaissant une unique fois
    """
    M = []
    for x in L:
        if x not in M:
            M.append(x)
    return M
```

On peut également transformer la série statistique de départ en dictionnaire de clefs les modalités, et valeurs l'effectif associé à chaque modalité. C'est la fonction `dico_occur` déjà rencontrée.

(Dictionnaire des effectifs)

```
def dico_occur(L):
    D = {}
    for x in L:
        if x not in D:
            D[x] = 1
        else :
            D[x] += 1
    return D
```

On peut également revenir à une liste d'observations si on le souhaite.

(Dictionnaire des effectifs vers liste)

```
def dico_occur_vers_liste(D):
    L = []
    for x in D:
        # x est une modalité, que l'on duplique autant de \
        ↪ fois que nécessaire
        eff_x = D[x]
        for _ in range(eff_x):
            L.append(x)
    return L
```

La fonction de calcul de moyenne s'appuie notamment sur celle qui calcule la somme.

(Moyenne)

```
def moyenne(L):
    """
    Renvoie la moyenne des éléments d'une liste
    """
    S = 0
    for x in L:
        S += x
    return S/len(L)
```

Si l'on préfère, on peut aussi calculer directement la moyenne à l'aide du dictionnaire des effectifs : dans ce cas, on pondère par l'effectif associé.

(Moyenne avec effectifs)

```
def moyenne_avec_eff(D):
    """
    Renvoie la moyenne d'une série associée au dictionnaire \
    ↪ des effectifs
    D
    """
    S = 0
    N = 0 # nombre d'éléments de la série
    for x in D:
        eff_x = D[x]
        S += x*eff_x
        N += eff_x
    return S/N
```

Pour la variance, on utilise généralement la version König-Huygens de la formule : $\mathbb{V}_x = \overline{x^2} - \overline{x}^2$ si x désigne une série statistique.

(Variance)

```
def variance(L):
    """
    Renvoie la variance, version KH
    """
    S2 = 0
    for x in L:
        S2 += x**2
    return S2/len(L) - moyenne(L)**2
```

Voyons quelques exemples d'exécutions.

Exemple 17 On code la série

19	26	23	20	22	24	20	24
22	20	21	19	21	22	19	20
21	21	22	21	23	22	21	24
25	23	22	19	20	26	24	25
23	26	25	25	21	22	25	24
23	22	24	24	25	23	25	22

avec la liste :

```
>>> L = [19, 26, 23, 20, 22, 24, 20, 24, 22, 20, 21, 19, 21, \
↪ 22, 19, 20, 21, 21, 22, 21, 23, 22, 21, 24, 25, 23, 22, 19, \
↪ 20, 26, 24, 25, 23, 26, 25, 25, 21, 22, 25, 24, 23, 22, 24, \
↪ 24, 25, 23, 25, 22]
>>> modalites(L)
[19, 26, 23, 20, 22, 24, 21, 25]
>>> D = dico_occur(L)
>>> D
{19: 4, 26: 3, 23: 6, 20: 5, 22: 9, 24: 7, 21: 7, 25: 7}
>>> dico_occur_vers_liste(D)
[19, 19, 19, 19, 26, 26, 26, 23, 23, 23, 23, 23, 23, 20, 20, \
↪ 20, 20, 20, 22, 22, 22, 22, 22, 22, 22, 22, 22, 24, 24, 24, \
↪ 24, 24, 24, 24, 21, 21, 21, 21, 21, 21, 21, 25, 25, 25, 25, \
↪ 25, 25, 25]
>>> moyenne(L)
22.5
>>> variance(L)
4.083333333333314
>>> moyenne_avec_eff(D) # on retrouve bien le même résultat
22.5
```

On peut également, après recherche du minimum et du maximum, renvoyer l'étendue de la série.

(Étendue d'une série)

```
def etendue(L):
    """
    Renvoie l'étendue de la série statistique des éléments de L
    """
    mini = L[0]
    maxi = L[0]
    for x in L[1:]:
        if x < mini:
            mini = x
        elif x > maxi:
            maxi = x
    return maxi - mini
```

Pour calculer la médiane, il faut au préalable trier la liste.

(Médiane)

```
def mediane(L):
    """
    Cherche la médiane d'une liste, après tri rapide des \
    ↪ observations
    """
    L_tri = tri_rapide_rec(L)
    n = len(L)
    if n % 2 == 1:
        # Nombre impair d'observations
        return L_tri[n//2]
    else:
        # Nombre pair d'observations
        return (L_tri[n//2-1] + L_tri[n//2])/2
```

Plus généralement, voici comment calculer les 3 quartiles.

(Quartiles)

```python
def quartiles(L):
    """
    renvoie Q1, Q2, Q3, après tri rapide des observations
    """
    L_tri = tri_rapide_rec(L)
    n = len(L)
    if n % 2 != 0:
        # Nombre impair d'observations
        Q2 = L_tri[n//2]
    else:
        # Nombre pair d'observations
        Q2 = (L_tri[n//2-1] + L_tri[n//2])/2
    if n % 4 != 0:
        # Nombre non multiple de 4 d'observations
        Q1 = L_tri[n//4]
        Q3 = L_tri[(3*n)//4]
    else:
        # Nombre multiple de 4 d'observations
        Q1 = L_tri[n//4-1]
        Q3 = L_tri[(3*n)//4-1]
    return Q1, Q2, Q3
```

3. STATISTIQUES DESCRIPTIVES BIVARIÉES

Dans une population donnée, on étudie simultanément deux caractères quantitatifs discrets X et Y : à partir d'un échantillon de n individus de cette population, on recueille les deux séries statistiques $x = (x_1, x_2, \dots, x_n)$ et $y = (y_1, y_2, \dots, y_n)$. On forme alors la série double $((x_1, y_1), (x_2, y_2), \dots, (x_n, y_n))$ qu'on suppose composée de n couples deux à deux distincts.

On veut savoir s'il existe une *corrélation* entre les séries x et y afin, éventuellement, d'en déduire un lien entre les caractères X et Y qui contredirait leur *indépendance*.

Attention : Les statistiques vont nous permettre de constater une plus ou moins forte corrélation entre deux caractères mais elles n'ont pas de légitimité pour discerner les causes de cette corrélation : un caractère est-il la cause de l'autre ; sont-ils tous deux conséquences d'une cause tierce ?

Par exemple, on observe une corrélation entre le nombre de nids de cigognes et le nombre de naissances humaines (le taux de natalité humaine et la population de cigognes suivent des trajectoires comparables). De là à affirmer un lien de causalité entre l'arrivée des cigognes et la naissances des bébés humains, ce n'est pas à la statistique de se prononcer...

Ainsi, il ne sera pas possible de conclure qu'il existe une relation de cause à effet entre deux caractères seulement du fait qu'ils sont corrélés : leur causalité nécessitera d'autres investigations que ce cours n'abordera pas.

3.1. Nuage de points

La série double présentée précédemment peut être représentée par un nuage de n points de $\mathbb{R}^2$: chaque point a pour coordonnées (x_k, y_k).

La représentation graphique du nuage de points est la première étape essentielle pour déterminer s'il existe ou non une relation entre x et y.

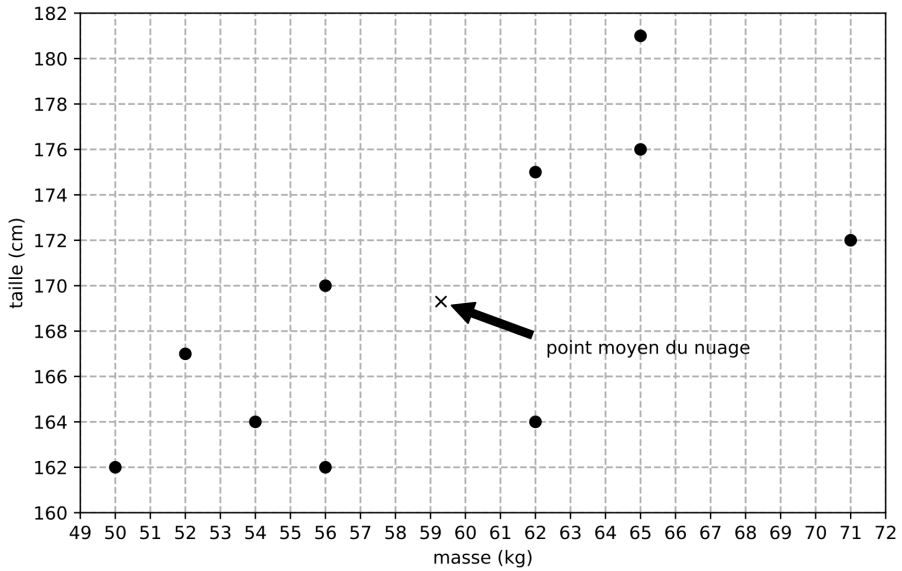
Définition 11 | Point moyen

Le point moyen d'une série statistique pour les caractères (x, y) est le point de coordonnées $G(\overline{x}, \overline{y})$.

Exemple 18 **(Relation entre le poids et la taille)** On relève la taille et le poids de 10 personnes. Les résultats sont résumés dans le tableau suivant :

poids (kg)	56	56	50	62	71	54	52	62	65	65
taille (cm)	162	170	162	164	172	164	167	175	176	181

On calcule alors : $\overline{x} = 59,3$ et $\overline{y} = 169,3$.
Le point moyen est donc le point G(59,3; 169,3).
Le nuage de points correspondant est donné par le graphique ci-après



3.2. Caractéristiques de position & dispersion

On cherche dans un premier temps à construire un test nous permettant de vérifier si les points du nuage sont alignés ou non. On introduit pour cela deux nouvelles définitions (covariance, coefficient de corrélation) et un théorème : l'inégalité de Cauchy-Schwarz.

Définition 12 | Covariance
Soit $(x,y) = ((x_1,y_1),\dots,(x_n,y_n))$ une série statistique. On appelle *covariance de x et de y*, notée $\mathbb{C}_{x,y}$, la quantité : $\mathbb{C}_{x,y} = \dfrac{1}{n}\sum_{k=1}^{n}(x_k-\overline{x})(y_k-\overline{y})$.

Remarque 6 **(Interprétation)** Si $\mathbb{C}_{x,y} < 0$ alors x et y ont tendance à varier dans des sens opposés (quand l'un augmente l'autre diminue), si $\mathbb{C}_{x,y} > 0$ alors ils ont tendance à varier dans le même sens.

Proposition 5 | Propriétés de la covariance
Soit $(x,y) = ((x_1,y_1),\dots,(x_n,y_n))$ une série statistique bivariée.
- **[Lien covariance/variance]** $\mathbb{C}_{x,x} = \mathbb{V}_x$.
- **[Formule de König-Huygens]** $\mathbb{C}_{x,y} = \overline{xy} - \overline{x}\cdot\overline{y}$.
- **[Symétrie]** $\mathbb{C}_{x,y} = \mathbb{C}_{y,x}$.
- **[Constante]** $\mathbb{C}_{x,c} = \mathbb{C}_{c,x} = 0$ pour toute constante $c \in \mathbb{R}$.
- **[Linéarité]** $\mathbb{C}_{\lambda x+\mu y,z} = \lambda\mathbb{C}_{x,z} + \mu\mathbb{C}_{y,z}$, $\mathbb{C}_{z,\lambda x+\mu y} = \lambda\mathbb{C}_{z,x} + \mu\mathbb{C}_{z,y}$.
- **[Variance d'une somme]** $\mathbb{V}_{x+y} = \mathbb{V}_x + \mathbb{V}_y + 2\mathbb{C}_{x,y}$.

Preuve
- $\mathbb{C}_{x,x} = \dfrac{1}{n}\sum_{k=1}^{n}(x_k-\overline{x})^2 = \mathbb{V}_x$
- $\mathbb{C}_{x,y} = \dfrac{1}{n}\sum_{k=1}^{n}(x_k-\overline{x})(y_k-\overline{y})$
  $= \dfrac{1}{n}\sum_{k=1}^{n}(x_ky_k - \overline{x}y_k - \overline{y}x_k + \overline{x}\cdot\overline{y})$ *développement du produit*
  $= \dfrac{1}{n}\sum_{k=1}^{n}x_ky_k - \dfrac{1}{n}\sum_{k=1}^{n}\overline{x}y_k - \dfrac{1}{n}\sum_{k=1}^{n}\overline{y}x_k + \dfrac{1}{n}\sum_{k=1}^{n}\overline{x}\cdot\overline{y}$ *linéarité de la somme*
  $= \overline{xy} - \overline{x}\cdot\overline{y} - \overline{x}\cdot\overline{y} + \overline{x}\cdot\overline{y}$
  $= \overline{xy} - \overline{x}\cdot\overline{y}$
- Immédiat puisque le produit de deux réels est commutatif.
- Constatons que $\overline{c} = c$, où $c = (c,\dots,c)$ désigne par abus de notation la série constante. Donc $\mathbb{C}_{x,c} = \overline{x}\cdot c - \overline{x}\cdot c = 0 = \mathbb{C}_{c,x}$ par symétrie.
- Conséquence de la linéarité de la somme.
- $\mathbb{V}_{x+y} = \overline{(x+y)^2} - \overline{x+y}^2$
  $= \overline{x^2+2xy+y^2} - (\overline{x}+\overline{y})^2$ *linéarité de la moyenne*
  $= \overline{x^2} + 2\overline{xy} + \overline{y^2} - \overline{x}^2 - \overline{y}^2 - 2\overline{x}\cdot\overline{y}$
  $= \overline{x^2} - \overline{x}^2 + \overline{y^2} - \overline{y}^2 + 2(\overline{xy} - \overline{x}\overline{y})$ *formule de König-Huygens*
  $= \mathbb{V}_x + 2\mathbb{C}_{x,y} + \mathbb{V}_y$.

Remarque 7 La covariance est une généralisation de la variance pour plusieurs caractères. Lorsqu'elle est appliquée à un seul caractère, on retrouve la formule de la variance.

Avant de poursuivre, commençons par une inégalité relative aux sommes, importante pour la suite.

Lemme 1 | Inégalité de CAUCHY-SCHWARZ

Soient $x = (x_1, \dots, x_n)$ et $y = (y_1, \dots, y_n)$ deux vecteurs de $\mathbb{R}^n$.

- **[Inégalité]** $\left|\sum_{i=1}^n x_i y_i\right| \leqslant \sqrt{\sum_{i=1}^n x_i^2} \times \sqrt{\sum_{i=1}^n y_i^2}$.
- **[Cas d'égalité]** $\left|\sum_{i=1}^n x_i y_i\right| = \sqrt{\sum_{i=1}^n x_i^2} \times \sqrt{\sum_{i=1}^n y_i^2} \iff \exists \lambda \in \mathbb{R}, \quad x = \lambda y.$

Preuve *(Point clef — **Introduire la fonction** $P : \lambda \in \mathbb{R} \longmapsto \sum_{i=1}^n (x_i + \lambda y_i)^2$**, c'est un polynôme en** λ*)*

Soit $\lambda \in \mathbb{R}$, alors par linéarité de la somme : $P(\lambda) = \sum_{i=1}^n x_i^2 + 2\lambda \sum_{i=1}^n x_i y_i + \lambda^2 \sum_{i=1}^n y_i^2$. C'est un polynôme en λ de degré 1 ou 2.

- **[1er cas]** Si $\sum_{i=1}^n y_i^2 = 0$, alors $y = 0_{\mathbb{R}^n}$ et l'inégalité est évidente (elle devient $0 \leqslant 0$).
- **[2ème cas]** Si $\sum_{i=1}^n y_i^2 > 0$, alors P est un trinôme, positif, donc de discriminant négatif :

$$\Delta = 4\left(\sum_{i=1}^n x_i y_i\right)^2 - 4\left(\sum_{i=1}^n x_i^2\right)\left(\sum_{i=1}^n y_i^2\right) \leqslant 0 \iff \left(\sum_{i=1}^n x_i y_i\right)^2 \leqslant \left(\sum_{i=1}^n x_i^2\right)^2 \left(\sum_{i=1}^n y_i^2\right)^2$$

passage à la racine

$$\iff \left|\sum_{i=1}^n x_i y_i\right| \leqslant \sqrt{\sum_{i=1}^n x_i^2} \times \sqrt{\sum_{i=1}^n y_i^2}.$$

Enfin, le cas d'égalité est obtenu lorsque :

$\Delta = 0 \iff$ P possède une racine double, $\iff$ P s'annule sur $\mathbb{R}$ (car positif),

$\iff \exists \lambda \in \mathbb{R}, \quad P(\lambda) = 0 = \sum_{i=1}^n (x_i + \lambda y_i)^2$ *somme de termes positifs*

$\iff \exists \lambda \in \mathbb{R}, \quad \forall i \in [\![1, n]\!], \quad x_i = \lambda y_i$

$\iff \exists \lambda \in \mathbb{R}, \quad x = \lambda y.$

Définition/Proposition 1 | Coefficient de corrélation

Soit $(x, y) = ((x_1, y_1), \dots, (x_n, y_n))$ une série statistique bivariée.

- **[Inégalité de CAUCHY-SCHWARZ]**
$$\left|\mathbb{C}_{x,y}\right| \leqslant \sqrt{\mathbb{V}_x}\sqrt{\mathbb{V}_y} \quad \text{ou encore} \quad |\mathbb{C}_{x,y}| \leqslant \sigma_x \sigma_y.$$
- Si x, y sont d'écart-type non nul, on appelle *coefficient de corrélation entre x et y* la quantité : $\rho_{x,y} = \dfrac{\mathbb{C}_{x,y}}{\sigma_X . \sigma_Y} \in [-1, 1]$.
- **[Cas d'égalité]** $\rho_{x,y} = \pm 1 \iff \exists a, b \in \mathbb{R}, y = ax + b$. *(la série y dépend de x et de manière affine)*

On voit toute de suite l'intérêt de la seconde partie de la proposition afin de mesurer la dépendance affine d'une série statistique par rapport à une autre.

Preuve Commençons par appliquer l'inégalité de CAUCHY-SCHWARZ (lemme précédent aux vecteurs) $(x_1 - \overline{x}, \dots, x_n - \overline{x})$, et $(y_1 - \overline{y}, \dots, y_n - \overline{y})$. On obtient alors :

$$\left|\sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y})\right| \leqslant \sqrt{\sum_{i=1}^n (x_i - \overline{x})^2} \times \sqrt{\sum_{i=1}^n (y_i - \overline{y})^2}$$

division par n de chaque côté

$$\iff \left|\frac{1}{n}\sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y})\right| \leqslant \sqrt{\frac{1}{n}\sum_{i=1}^n (x_i - \overline{x})^2} \times \sqrt{\frac{1}{n}\sum_{i=1}^n (y_i - \overline{y})^2}$$

$$\iff |\mathbb{C}_{x,y}| \leqslant \sqrt{\mathbb{V}_x}\sqrt{\mathbb{V}_y} \iff \rho_{x,y} \leqslant 1.$$

On en déduit alors le cas d'égalité

$\rho_{x,y} = \pm 1 \iff |\mathbb{C}_{x,y}| = \sigma_x \sigma_y$

$\iff (x_1 - \overline{x}, \dots, x_n - \overline{x}), (y_1 - \overline{y}, \dots, y_n - \overline{y})$ réalisent le cas d'égalité dans CAUCHY-SCHWARZ

$\iff \exists \lambda \in \mathbb{R}, \quad \forall i \in [\![1, n]\!], \quad y_i - \overline{y} = \lambda(x_i - \overline{x})$

$\iff \exists a, b \in \mathbb{R}, \quad y = ax + b$ en notant $a = \lambda, b = \overline{y} - \lambda \overline{x}$.

Remarque 8 Dans les faits, le coefficient de corrélation linéaire est, en valeur absolue, très rarement exactement égal à 1 . Néanmoins, si sa valeur absolue est « proche » de 1 alors on peut considérer qu'il existe une relation linéaire *(c'est un abus de langage, il faudrait plutôt dire une relation affine)* entre les valeurs de la série y et celles de la série x. On dit alors que les séries x et y sont **linéairement corrélées**. Cependant, le coefficient de corrélation linéaire ne quantifie que l'alignement des points de la série double, c'est-à-dire d'un **échantillon** de la population totale. Aussi, si l'effectif de l'échantillon est "trop petit" rapporté à l'effectif total de la population alors, même avec un coefficient de corrélation égal à ± 1, on ne pourra pas en déduire que les caractères X et Y sont linéairement corrélés... En revanche, s'il est éloigné de 1 (en valeur absolue), on en déduira que les caractères X et Y ne sont pas linéairement corrélés : une courbe passant par des points alignés est peut-être une droite ; une courbe passant par des points qui ne sont pas alignés n'est pas une droite (cf. l'introduction du chapitre). Aussi, même lorsque $\rho_{x,y} = 0$, on ne saurait dire si les caractères sont **indépendants** (il peut exister une corrélation autre que linéaire) alors que la réciproque est vraie : si deux caractères sont indépendants alors ils ne sont pas linéairement corrélés ! Enfin, on rappelle encore une fois que la corrélation de deux caractères n'indique pas un lien de causalité entre eux : ils peuvent dépendre d'un tierce paramètre. L'étude du coefficient de corrélation permet de conjecturer l'existence d'un lien de causalité et d'argumenter en sa faveur a postériori, mais elle ne permet jamais de le *démontrer*.

3.3. Ajustement affine

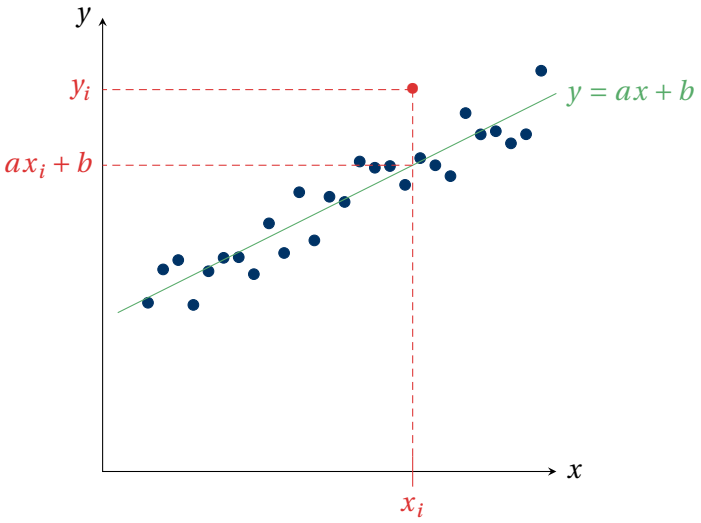
Il est courant, en physique-chimie, en sciences industrielles, ou plus généralement dans toute discipline expérimentale comme la biologie, la chimie, l'économie d'avoir à comparer des données expérimentales et de conjecturer une éventuelle dépendance linéaire entre deux paramètres donnés.

Le problème. L'idée de l'ajustement affine est la suivante : on dispose de deux séries statistiques (souvent expérimentales) x et y et on soupçonne qu'il existe une relation les liant de la forme $y = ax + b$. Ce soupçon peut provenir :

- du tracé du nuage de points (x, y),
- et/ou du calcul de $\rho_{x,y}$, que l'on observe proche de ± 1.

[Objectif] On veut alors chercher la droite d'équation $y = ax + b$ qui passe « le mieux » par notre nuage de points. Parfois on sait que la relation existe et on veut déterminer a et b.

Plus précisément, soit $(x_i, y_i)_{1\leqslant i\leqslant n}$ avec $n \geqslant 1$ est un nuage de n points provenant de séries statistiques x, y. En regardant un dessin, nous voyons que si l'on approche le nuage par la droite $y = ax + b$ avec $(a, b) \in \mathbb{R}^2$, alors l'écart entre cette droite et le nuage, au point x_i, $i \in [\![1\,,\, n]\!]$, est donné par : $y_i - ax_i - b$.



Problème de régression linéaire

Sauf que l'on veut que tous les écarts soit minimum. Pour cela, on peut chercher à trouver le minimum des fonctions ci-après (en a, b) :

$$\max_{1\leqslant i\leqslant n} |y_i - ax_i - b|\,, \quad \sum_{i=1}^{n} |y_i - ax_i - b|\,, \quad \mathrm{F}(a,b) = \sum_{i=1}^{n} (y_i - ax_i - b)^2.$$

Plus ces quantités sont petites, plus tous les écarts à la droite seront également petits. Dans le dernier cas, on parle de *minimisation au sens des moindres carrés* *(à cause de la présence des carrés)* et c'est cette minimisation que nous allons essayer de réaliser car c'est pour celle-ci que les calculs sont les plus simples. Nous pouvons résoudre ce problème de deux manières. Il s'agira donc de minimiser la fonction de deux variables :

$$\boxed{\begin{array}{c|ccc} \mathrm{F} & \mathbb{R}^2 & \longrightarrow & \mathbb{R} \\ & (a,b) \in \mathbb{R}^2 & \longmapsto & \displaystyle\sum_{i=1}^{n} (y_i - ax_i - b)^2\,. \end{array}}$$

Le problème est qu'il s'agit la d'une fonction de deux variables pour laquelle nous n'avons pas de méthode : la méthode classique du tableau de variations ne fonctionne plus ici.

Cas réellement affine $y = ax + b$. Si y est réellement affine en x, *i.e.* de la forme $y = ax + b$ avec $a, b \in \mathbb{R}$, alors d'après les propriétés de la covariance et de la moyenne déjà établies, nous avons :

$$\overline{y} = \overline{ax + b} = a\overline{x} + b \implies \text{la droite passe par le point moyen,} \quad \boxed{b = \overline{y} - a\overline{x}}$$

$$\mathbb{C}_{y,x} = \mathbb{C}_{ax+b,x} = a\mathbb{C}_{x,x} + 0 \implies \boxed{a = \frac{\mathbb{C}_{y,x}}{\sigma_x^2}}.$$

Il s'avère que le couple (a, b) obtenu dans le cas très particulier où $y = ax + b$, noté $(a^\star, b^\star)$ dans la suite, est également la solution du cas général. C'est ce que nous montrons dès à présent.

Théorème 1 | Existence de la droite des moindres carrés

Soit (x, y) une série statistique double constituée d'une suite de couples $((x_k, y_k))_{1\leqslant k\leqslant n}$. Alors $(a^\star, b^\star)$ défini par :

$$a^\star = \frac{\mathbb{C}_{x,y}}{\sigma_x^2}, \qquad b^\star = \overline{y} - a^\star\overline{x} = \overline{y} - \frac{\mathbb{C}_{x,y}}{\sigma_x^2}\overline{x},$$

est l'unique point de $\mathbb{R}^2$ qui minimise F, c'est-à-dire :

$$\forall (a,b) \in \mathbb{R}^2, \quad \mathrm{F}(a^\star, b^\star) \leqslant \mathrm{F}(a,b).$$

La droite de régression par la méthode des moindres carrés de y en x a donc pour équation : $\boxed{y = \dfrac{\mathbb{C}_{x,y}}{\sigma_x^2}(x - \overline{x}) + \overline{y}.}$

Preuve Nous admettrons l'unicité, on justifie simplement que :

$$\forall (a,b)\in\mathbb{R}^2,\quad F(a^\star,b^\star)\leqslant F(a,b).$$

Soit $(a,b)\in\mathbb{R}^2$. Alors par linéarité de la somme :

$$\begin{aligned}F(a,b)&=\sum_{i=1}^{n}(y_i-ax_i-b)^2=\sum_{i=1}^{n}((y_i-ax_i)-b)^2\\&=\sum_{i=1}^{n}(y_i-ax_i)^2-2b\sum_{i=1}^{n}(y_i-ax_i)+\sum_{i=1}^{n}b^2\\&=\sum_{i=1}^{n}(y_i-ax_i)^2-2b\sum_{i=1}^{n}(y_i-ax_i)+nb^2.\end{aligned}$$

identité remarquable

Pour tout $a\in\mathbb{R}$, notons $f(b)=F(a,b)=\sum_{i=1}^{n}(y_i-ax_i)^2-2b\sum_{i=1}^{n}(y_i-ax_i)+nb^2$. Alors par le calcul précédent, on voit que f est un trinôme (en b), et de courbe une parabole orientée vers le haut puisque $n>0$. Ainsi, f est minimale là où sa dérivée s'annulle. Mais :

$$f'(b)=-2\sum_{i=1}^{n}(y_i-ax_i)+2nb=0\iff b=\frac{1}{n}\left(\sum_{i=1}^{n}y_i-a\sum_{i=1}^{n}x_i\right)=\overline{y}-a\overline{x}.$$

On vient alors de montrer que :

$$\forall a,\quad(\forall b\in\mathbb{R},\quad f(b)\geqslant f(\overline{y}-a\overline{x}))\iff\forall a,b\in\mathbb{R},\quad F(a,b)\geqslant F(a,\overline{y}-a\overline{x}).$$

On souhaite encore trouver le minimum en a du minorant. On considère donc ensuite :

$$g:a\in\mathbb{R}\longmapsto F(a,\overline{y}-a\overline{x}).$$

Justifions de-même que g est un trinôme en a.

$$\begin{aligned}g(a)&=\sum_{i=1}^{n}[a(x_k-\overline{x})-(y_k-\overline{y})]^2\\&=a^2\sum_{i=1}^{n}(x_k-\overline{x})^2-2a\sum_{i=1}^{n}(x_k-\overline{x})(y_k-\overline{y})\\&\quad+\sum_{i=1}^{n}(y_k-\overline{y})^2\\&=a^2n\sigma_x^2-a\left(2n\mathbb{C}_{x,y}\right)+n\sigma_y^2,\end{aligned}$$

identité remarquable

polynôme de degré 2 en a

$$g'(a)=2an\sigma_x^2-2n\mathbb{C}_{x,y}.$$

Comme g est encore un trinôme de graphe une parabole orientée vers le haut, elle est minimale là où g' s'annule, *i.e.* en $a=\dfrac{\mathbb{C}_{x,y}}{\sigma_x^2}$. En résumé, nous avons montré :

$$\forall a,b\in\mathbb{R},\quad F(a,b)\geqslant F(a,\overline{y}-a\overline{x})\geqslant F\left(\frac{\mathbb{C}_{x,y}}{\sigma_x^2},\overline{y}-\frac{\mathbb{C}_{x,y}}{\sigma_x^2}\overline{x}\right).$$

Cette inégalité prouve que $\left(\dfrac{\mathbb{C}_{x,y}}{\sigma_x^2},\overline{y}-\dfrac{\mathbb{C}_{x,y}}{\sigma_x^2}\overline{x}\right)$ est un minimum global de F.

Exemple 19 Voici un tableau représentant les valeurs prises par deux caractères.

x	1	3	4	5	7	10	12	13
y	1,5	2,2	2,6	3,1	3,9	4,8	5,4	6,6

Calculons leur coefficient de corrélation à l'aide de la formule de KÖNIG-HUYGENS :

Calculs préliminaires :

$$\bar{x}=\frac{1+3+4+5+7+10+12+13}{8}=6,875$$

$$\bar{y}=\frac{1,5+2,2+2,6+3,1+3,9+4,8+5,4+6,6}{8}=3,7625$$

$$\overline{x^2}=\frac{1^2+3^2+4^2+5^2+7^2+10^2+12^2+13^2}{8}=64,125$$

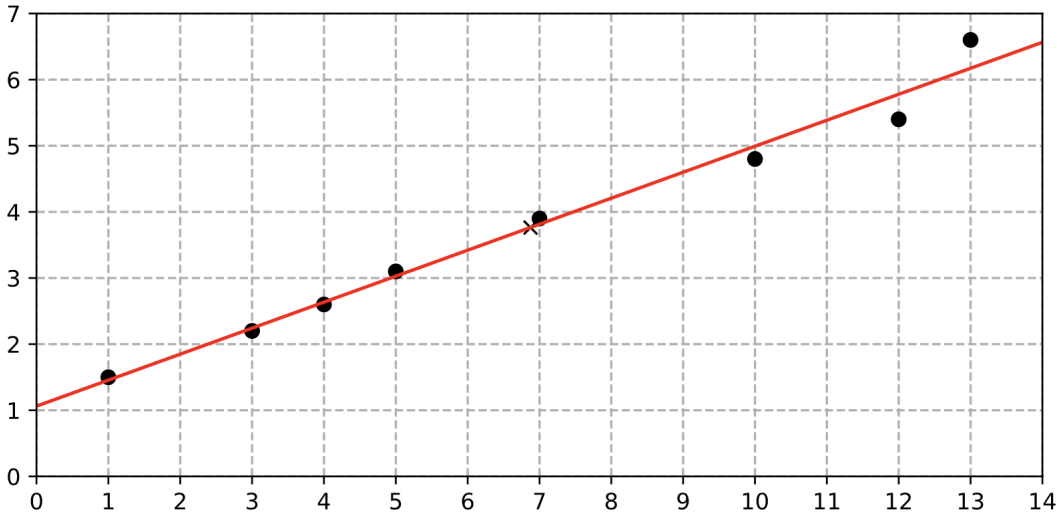
$$\overline{y^2}=\frac{1,5^2+2,2^2+2,6^2+3,1^2+3,9^2+4,8^2+5,4^2+6,6^2}{8}=16,80375$$

$$\overline{xy}=\frac{1\times1,5+3\times2,2+4\times2,8+5\times3,1+7\times3,9+10\times4,8+12\times5,4+13\times6,6}{8}$$

$$=32,4875$$

On en déduit que :

- la covariance de x et y est $\sigma_{xy}=\overline{xy}-\bar{x}\bar{y}\approx6,62$
- leurs écarts-types sont $\sigma_x=\sqrt{\overline{x^2}-(\bar{x})^2}\approx4,11$ et $\sigma_y=\sqrt{\overline{y^2}-(\bar{y})^2}\approx1,63$
- leur coefficient de corrélation est $\rho_{xy}=\dfrac{\sigma_{xy}}{\sigma_x\sigma_y}\approx0,99$ il est proche de 1 : on réalise un ajustement affine ...
- la droite des moindres carrés associée est (approximativement) celle d'équation $y=0,393x+1,063$



QUALITÉ D'UNE RÉGRESSION LINÉAIRE. Comment évaluer la « justesse » d'un ajustement ? Pour y répondre on définit un nouvel indicateur statistique : le coefficient de détermination, plutôt que le seul coefficient de corrélation.

Définition/Proposition 2 | Coefficient de détermination d'une régression

Soit (x,y) une série statistique double constituée d'une suite de couples $((x_k,y_k))_{1\leqslant k\leqslant n}$. On appelle *coefficient de détermination de x et y*, noté $r^2(x,y)$, la quantité définie par : $\boxed{r^2(x,y)=\rho_{x,y}^2}=\dfrac{\mathbb{C}_{x,y}^2}{\mathbb{V}_x\mathbb{V}_y}\in[0,1]$.

Note | *Puisque $\rho_{x,y}\in[-1,1]$, son carré est bien dans $[0,1]$.*

Attention

- Ce n'est donc pas le coefficient de corrélation, mais son carré.
- Nous n'avons pas défini $r(x,y)$, $r^2(x,y)$ est une notation mais ne désigne pas un carré.

Remarque 9 (Interprétation) Ainsi $r^2(x,y)=1$ correspond à une adéquation parfaite tandis que $r^2(x,y)$ proche de 0, équivalent à $\rho_{x,y}$ proche de 0, indique une faible liaison linéaire ce qui peut signifier qu'il n'y a pas de lien entre x et y ou bien que x et y sont liés par une relation non-affine. En général, on considère une régression linéaire comme « satisfaisante » lorsque : $r^2(x,y)\geqslant 0.9$.

3.4. Informatique

En utilisant directement les définitions de chaque quantité, on en déduit les fonctions associées ci-dessous.

(Covariance)

```
def covariance(L, M):
    """
    Renvoie la covariance des deux séries
    """
    Prod = [L[i]*M[i] for i in range(len(M))]
    return moyenne(Prod) - moyenne(L)*moyenne(M)
```

(Coefficient de corrélation)

```
def coeff_cor(X, Y):
    """
    Renvoie le coefficient de corrélation des deux séries
    """
    return covariance(X, \
    ↪ Y)/(ma.sqrt(variance(X))*ma.sqrt(variance(Y)))
```

(Coefficients de régression)

```
def regression_lin(X, Y):
    """
    renvoie les coefficients a, b de régression  linéaire \
    ↪ associée au
    nuage de points (X, Y)
    """
    a = covariance(X, Y)/variance(X)
    b = moyenne(Y)-a*moyenne(X)
    return a, b
```

Exemple 20 (Relation entre le poids et la taille) Regardons ce que donne cette fonction sur les séries statistiques ci-après, la série X correspondant à des relevés de poids, et Y de taille (Exemple 18).

```
>>> X = [150, 170, 162, 164, 172, 164, 167, 175, 176, 181]
>>> Y = [56, 56, 50, 62, 71, 54, 52, 62, 65, 65]
>>> a, b = regression_lin(X, Y)
>>> a
0.4485537487408167
>>> b # on retrouve bien les bonnes valeurs
-16.10188516333129
>>> coeff_cor(X, Y)**2
0.3442851600160366
```

Le coefficient de détermination est très inférieur à 0.9, la régression est donc très mauvaise (ce que l'on pouvait constater graphiquement).

Exceptionnellement, le TD de ce chapitre sera fait au travers d'un TP d'Informatique.