Chapitre # (PS) 2 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 : nous en étudierons deux en BCPST. Les statistiques descriptives d'une part 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.

La statistique descriptive est la branche des statistiques qui regroupe les nombreuses techniques utilisées pour décrire un ensemble relativement important de données. Cette description peut être réalisée au moyen de calculs de grandeurs (moyenne, variance, écart-type, mode, *etc.*) ou au moyen de descriptions visuelles (graphiques par exemple). Nous nous poserons également une seconde question : on peut représenter deux séries de données sous forme d'un nuage de points, on peut alors se demander s'il est possible d'évaluer la corrélation entre les deux séries, et plus précisément la faculté de l'une à dépendre de l'autre.

1. STATISTIQUES DESCRIPTIVES UNIVARIÉES

Cadre

Dans cette section, les notations n, p, i, j désigneront des entiers même si cela n'est pas précisé.

1.1. Série statistique

Définition 1 | Population & Échantillon

Une *population* est un ensemble fini dont les éléments sont appelés des *individus*. Le nombre d'individus d'une population est appelé sa *taille*. Un sous-ensemble d'une population est appelé un *échantillon* de cette population.

Remarque 1 Le mot « population » ne signifie pas que l'on considère des individus. Par exemple, si vous réalisez plusieurs titrages pour votre T.I.P.E., on parlera alors de population de concentrations.

Très souvent, on ne pourra pas mener notre étude sur la population entière mais sur une sous-partie que l'on espère représentative.

Définition 2 | Caractère, Série statistique

- Un *caractère* x de la population est une donnée *qualitative* ou *quantitative* attachée à chaque individu de la population. On notera x_i la valeur du caractère x pour un individu i.

- ◇ un caractère est dit *quantitatif* s'il prend des valeurs quantifiables, souvent des réels mais éventuellement des p-uplets ou des matrices.
- ◇ Un caractère est dit *qualitatif* s'il correspond à une propriété qui ne se quantifie pas.

- Une *série statistique* de taille n, est une famille $(x_1, \dots, x_n)$ à n éléments.

Par exemple, la couleur des cheveux, l'opinion politique sont des caractères qualitatifs.

Définition 3 | 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 $(x_1 y_1, x_2 y_2, \dots, x_n y_n)$.

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

Définition 4 | Modalités

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

Exemple 1 (Série des poids) Le tableau ci-dessous regroupe des données de poids d'individus numérotés entre 1 et n.

Individu	1	2	…	n
Masse	62	80	…	74

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 craactère étudié est le diamètre de la pièce en centimètre.

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

1.2. Modes de définition d'une série statistique

Une série statistique peut être donnée sous plusieurs formes.

- Soit on décide de fournir toutes les données de manière exhaustive,
- soit on décide de les regrouper en modalités et/ou classes, et on doit alors fournir un tableau d'effectifs associés.

DONNÉES REGROUPÉES PAR MODALITÉ. Si la taille de l'échantillon est trop grande on préfèrera donner les nombres d'individus associés à chaque modalité. On appelle cela l' *effectif* associé à ladite modalité.

Exemple 3 (Âges d'enfants) Dans cet exemple, une série d'âges sur des enfants entre 0 et 2 ans. On fournit alors les effectifs associés à chaque âge.

Âge	0	1	2	3
Effectif	20	31	3	10

DONNÉES REGROUPÉES EN CLASSES. 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*. Il peut également arriver que nos modalités comportent des classes et des modalités ponctuelles.

Exemple 4 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.

1.3. Effectif & Fréquence

1.3.1. Caractère discret

Définition 5 | Effectif

Soit une série statistique de taille n admettant un nombre fini de modalités ou à défaut de classes notées $a_1, \dots, a_p$.

- Pour $j \in [\![1\,,\,p]\!]$, on définit l'*effectif* n_j associé à la valeur a_j comme étant le nombre d'individus i pour lesquels $x_i = a_j$.
- Pour $j \in [\![1, p]\!]$, on définit la *fréquence de* f_j associée à la valeur a_j comme étant la proportion d'individus pour lesquels $x_i = a_j$, c'est-à-dire
$$f_j = \frac{n_j}{n}.$$

Proposition 1 | Formule des fréquences totales / effectifs totaux

Sous les mêmes hypothèses que précédemment, on a :
$$\sum_{j=1}^{p} n_j = n \quad \text{et} \quad \sum_{j=1}^{p} f_j = 1.$$

Dans le cas où les caractères étudiés sont des réels (ils peuvent donc être ordonnés), on va introduire les effectifs cumulés croissants et les fréquences cumulées croissantes.

Définition 6 | Effectifs cumulés croissants

On suppose ici les modalités $a_1, \dots, a_p$ rangées dans l'ordre croissant ($a_1 < a_2 < \dots < a_p$) (pour un regroupement en intervalles $]a, b]$ et $]c, d]$, cela correspond à $a < b \leqslant c < d$). Soit $j \in [\![1\,,\,p]\!]$.

- On définit *l'effectif cumulé croissant* associé à la modalité a_j comme le nombre d'observations $x_i \leqslant a_j$, c'est donc le nombre d'observations inférieures ou égales à a_j.
- On définit la *fréquence cumulée (croissante)* associée à la modalité a_j comme la proportion d'observations $x_i \leqslant a_j$, c'est-à-dire : $\mathrm{F}_j = \dfrac{\mathrm{N}_j}{n}$.

Exemple 5 Déterminer les effectifs/fréquences cumulé(e)s de l'Exemple 3.

Âge	…	…	…	…
Effectifs cumulés	…	…	…	…
Fréquences cumulées	…	…	…	…

Proposition 2 | Lien effectifs / effectifs cumulés

Sous les mêmes hypothèses que précédemment, on a :

1. $\forall j \in [\![1\,,\,p]\!], \quad \mathrm{N}_j = \sum_{k=1}^{j} n_k, \quad$ et $\quad \mathrm{F}_j = \sum_{k=1}^{j} f_k$.
2. Les effectifs et fréquences cumulés sont croissant(e)s :
$$0 \leqslant \mathrm{N}_1 \leqslant \mathrm{N}_2 \leqslant \dots \leqslant \mathrm{N}_p = n, \quad 0 \leqslant \mathrm{F}_1 \leqslant \mathrm{F}_2 \leqslant \dots \leqslant \mathrm{F}_p = 1.$$

Preuve

1. On écrit d'abord, avec la convention $a_0 = -\infty$:
$$\{i \in [\![1\,,\,n]\!] \mid x_i \leqslant a_j\} = \bigcup_{k=1}^{j} \{i \in [\![1\,,\,n]\!] \mid a_{k-1} < x_i \leqslant a_k\}.$$
En passant au cardinal dans cette réunion disjointe, il vient : $\mathrm{N}_j = \sum_{k=1}^{j} n_k$. Il suffit de diviser alors par n pour obtenir la version avec fréquences.
2. Il suffit de constater que pour $i, j \in [\![1\,,\,p]\!]$ tels que $1 \leqslant i < j \leqslant p$, nous avons l'inclusion :
$$\{k \in [\![1\,,\,n]\!] \mid x_k \leqslant a_i\} \subset \{k \in [\![1\,,\,n]\!] \mid x_k \leqslant a_j\}$$
puisque les classes sont ordonnées. Il suffit ensuite de passer au cardinal. On divise par n pour obtenir la version avec fréquences.

1.3.2. Caractère continu

Définition 7 | Effectif

Soit une série statistique regroupée en classes $\mathrm{I}_j = [a_{j-1}, a_j[, j \in [\![1\,,\,p]\!]$, (avec $(a_0, \dots, a_p)$ une suite croissante).

- Pour $j \in [\![1\,,\,p]\!]$, on définit l'*effectif* n_j associé à la valeur a_j comme étant le nombre d'individus i pour lesquels $x_i \in \mathrm{I}_j$.
- Pour $j \in [\![1, p]\!]$, on définit la *fréquence de* f_j associée à la valeur a_j comme étant la proportion d'individus pour lesquels $x_i \in \mathrm{I}_j$, c'est-à-dire : $f_j = \dfrac{n_j}{n}$.

Exemple 6 Par exemple, les notes sur 10 à une interrogation.

Données en classes	$[1,4[$	$[4,7[$	$[7,10[$
Effectifs	3	2	7

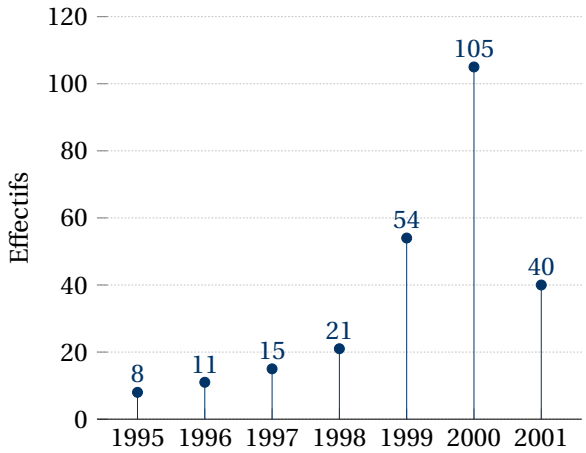
1.4. Représentation graphique

Les séries de données statistiques peuvent être représentés de plusieurs manières, en lieu et place de tableaux comme présentés *supra*.

DIAGRAMME EN BÂTONS & HISTOGRAMMES. Pour établir un diagramme en bâtons on va tracer pour chaque modalité un bâton (un rectangle long et fin) centré en a_j et de hauteur f_j ou n_j. Ce type de graphique est adapté aux données ponctuelles et aux données qualitatives. Pour les données continues regroupées en classes, on tracera plutôt ce que l'on appelle un *histogramme* (voir plus bas).

Exemple 7 (Diagramme en bâtons) Par exemple, ici nous avons représenté le tableau d'effectifs d'âges suivant :

Âge	1995	1996	1997	1998	1999	2000	2001
Effectif	8	11	15	21	54	105	40



Lorsque le nombre de modalités est trop grand, et même infini, nous avons dit que l'on regroupait généralement les données en classes. Pour chaque classe on va alors tracer un rectangle dont la largeur vaut l'amplitude de l'intervalle et dont *l'aire* est proportionnelle à la fréquence ou à l'effectif de la classe.

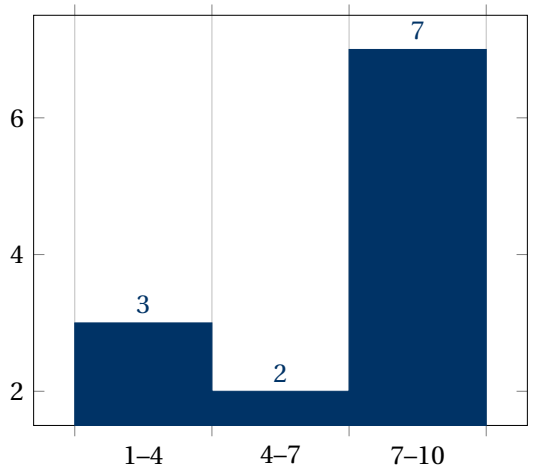
Attention

- C'est l'aire du rectangle qui est importante, pas sa hauteur même s'il y a bien entendu une relation entre les deux.
- La plupart du temps, les longueurs des bases des rectangles sont identiques, et les hauteurs sont égales aux effectifs.

Exemple 8 (Histogramme) Par exemple, sur les notes précédentes.

Données en classes	$[1,4[$	$[4,7[$	$[7,10[$
Effectifs	3	2	7

Sur cet exemple les classes ont toutes la même longueur, mais ceci n'est aucunement obligatoire.



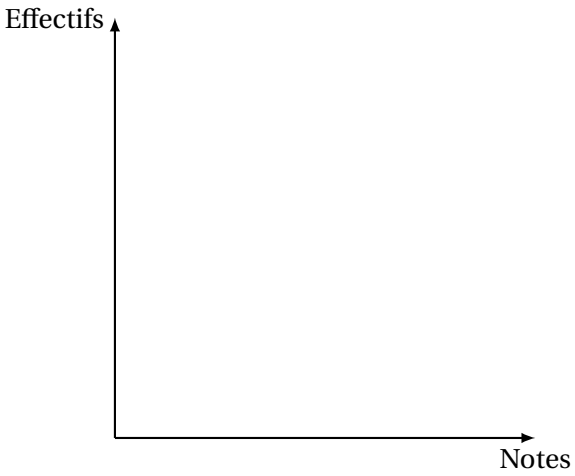
POLYGONE DES FRÉQUENCES CUMULÉES CROISSANTES. Ce type de graphique nous sera par exemple utile pour déterminer efficacement une médiane. Voyons comment il est construit.

- Pour des modalités ponctuelles, on place les points (a_i, F_i). On va alors relier les points (a_i, F_i) à (a_{i+1}, F_i) puis (a_{i+1}, F_i) à $(a_{i+1}, \mathrm{F}_{i+1})$. Schématiquement on trace un trait horizontal puis un trait vertical.
- Pour des modalités regroupées en classe $]\alpha_i, \alpha_{i+1}]$ (*resp.* $[\alpha_i, \alpha_{i+1}[$), on place les points A_i de coordonnées $(\alpha_{i+1}, \mathrm{F}_i)$ (*resp.* (α_i, F_i)), à la droite (*resp.* gauche) de l'intervalle donc. On relie ensuite simplement les points A_i par une ligne brisée.

Remarque 2 Cette représentation est pertinente si les individus sont répartis à peu près uniformément au sein de la classe. Si vous avez des raisons de penser que ce n'est pas le cas, il peut être judicieux de créer de nouvelles classes.

Exemple 9 (Notes sur 5) Considérons le tableau d'effectifs des notes d'un devoir noté sur 5 suivant :

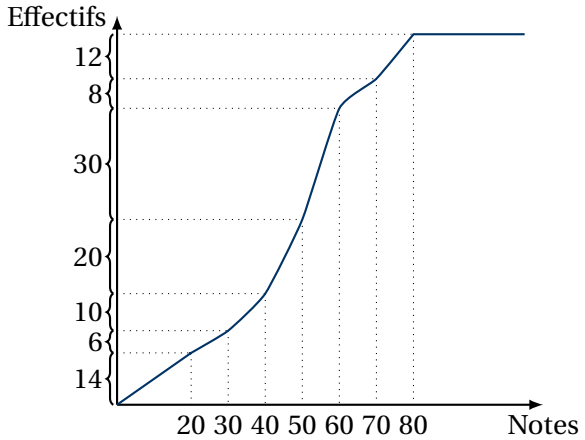
Notes en classe	0	1	2	3	4	5
Effectifs	2	4	3	3	6	2



Exemple 10 **(Notes sur 100)** 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.

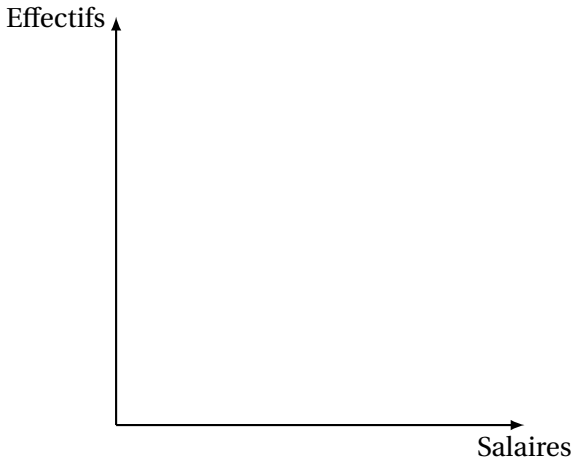


Lire graphiquement le nombre approximatif de notes inférieures à 30, et 56.

Exemple 11 **(Salaires)** Une étude portant sur les salaires mensuels des employés en CDI à temps complet d'une entreprise a permis d'établir le tableau ci-dessous.

x_i	$[1000,1200[$	$[1200,1300[$	$[1300,1400[$	$[1400,1500[$	$[1500,1900[$	$[1900,2200]$
n_i	28	24	20	8	24	6
N_i	28	52	72	80	104	110

Tracer le polygone des fréquences cumulées.



1.5. Caractéristiques de position

Les caractéristiques de position d'une série statistique sont des grandeurs dont la vocation est de mesurer la position des données qui constituent la série statistique.

Mode & Classe modale.

Définition 8 | 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 3 Il est possible qu'une série statistique admette plusieurs modes ou classes modales.

MOYENNE.

Définition 9 | Moyenne

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

Proposition 3 | Cas d'un regroupement en classes : lien avec les effectifs

- Soit $x = (x_1, \dots, x_n)$ une série statistique de modalités $(a_1, \dots, a_p)$ ponctuelles, alors : $\overline{x} = \dfrac{1}{n}\sum_{j=1}^{p} n_j a_j = \sum_{j=1}^{p} f_j a_j$.
- On dit alors que $\overline{x}$ est la *moyenne pondérée* des a_j par les fréquences f_j.

Preuve Constater simplement que dans $\sum_{i=1}^{n} x_i$, chaque x_i apparaît un nombre n_j de fois où n_j désigne l'effectif de x_i. On a donc : $\sum_{j=1}^{p} n_j a_j = \sum_{i=1}^{n} x_i$, d'où l'on tire ensuite les formules.

Quand on travaille avec des données regroupées par classes cette définition n'est pas utilisable. Dans cette situation on va alors considérer que les valeurs sont uniformément réparties dans les intervalles et prendre pour moyenne de la série la moyenne des milieux des intervalles pondérés par les effectifs.

Définition 10 | Moyenne pour des données regroupées en classes

Soit $x = (x_1, \dots, x_n)$ une série statistique quantitative. Supposons que les modalités $(a_j)_{j\in[\![1,p]\!]}$ de cette série correspondent à des intervalles $[b_j, c_j[$. On définit alors la *moyenne* par : $\overline{x} = \dfrac{1}{n}\sum_{j=1}^{p} n_j \times \left(\dfrac{b_j + c_j}{2}\right) = \sum_{j=1}^{p} f_j \times \left(\dfrac{b_j + c_j}{2}\right)$.

Proposition 4 | Propriétés de la moyenne

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 : $\overline{x + y} = \overline{x} + \overline{y}$.
- **[Mélange]** Soit z la série statistique obtenue en « concaténant » les séries x et y, *i.e.* $z = (x_1, \dots, x_n, y_1, \dots, y_m)$, alors : $\overline{z} = \dfrac{n\overline{x} + m\overline{y}}{n + m}$.

Preuve

- 🖌
- 🖌
- 🖌

Exemple 12 Dans les classes de 1BC1 (classe de 43), la moyenne des notes à un concours blanc est de 9.38, en 1BC2 (classe de 41) elle est 8.81. Calculer la moyenne sur l'ensemble des deux classes à ce concours blanc.

🖌

MÉDIANE. On suppose ici que nos modalités sont des réels. La moyenne est fortement influencée par les valeurs extrêmes, donc dans ce cas la donnée de la moyenne est assez peu instructive pour rendre compte de l'allure de la série statistique, et on privilégie une autre quantité appelée la médiane.

Définition 11 | Médiane pour des données ponctuelles

Soit x une série statistique de taille n dont les modalités sont données dans l'ordre croissant $a_1 < a_2 < \ldots < a_n$.

- On appelle *médiane de x* le réel $Q_2(x)$ défini par :
$$Q_2(x) = \begin{cases} a_{\frac{n+1}{2}} & \text{si } n \text{ est impair,} \\ \frac{1}{2}\left(a_{\frac{n}{2}} + a_{\frac{n}{2}+1}\right) & \text{si } n \text{ est pair.} \end{cases}$$
- Un individu dont le caractère correspond à la médiane est dit être un *individu médian*.

Exemple 13 (Médiane & Valeurs extrêmes) La médiane a, par rapport à la moyenne, l'avantage d'être peu influencée par les valeurs extrêmes. Elle est alors plus représentative que la moyenne lorsque la série comporte des valeurs très grandes ou très petites.
Par exemple, en France en 2014 le salaire moyen mensuel était de 1934 euros pour les femmes et 2389 euros pour les hommes tandis que le salaire médian mensuel était de 1619 euros pour les femmes et 1882 euros pour les hommes. La différence s'explique par le fait que les très hauts salaires, même s'ils sont peu nombreux, tirent la moyenne vers la haut.

Lorsque les données sont regroupées en classe, la définition de la médiane est purement graphique.

Définition 12 | Médiane pour des données continues

Soit une série statistique regroupée en classes $I_j = [a_{j-1}, a_j[, j \in [\![1\,,\,p]\!]$, avec $(a_0, \ldots, a_p)$ une suite croissante. On définit la *médiane* de x comme l'abscisse du point de la courbe des fréquences cumulées croissantes d'ordonnée $\frac{1}{2}$.

Remarque 4 (Version plus formelle)

- On définit la *classe médiane*, comme étant l'intervalle $I_j = [a_{j-1}, a_j[$ de sorte que $F_{j-1} < \frac{1}{2}$ et $F_j \geqslant \frac{1}{2}$ (c'est donc la classe associée à la première fréquence cumulée dépassant $\frac{1}{2}$).
- On définit alors : $Q_2(x) = a_{j-1} + \frac{a_j - a_{j-1}}{F_j - F_{j-1}} \times \left(\frac{1}{2} - F_{j-1}\right)$.
- D'où vient l'expression précédente ? On remarque simplement que
$$y = \tfrac{F_j - F_{j-1}}{a_j - a_{j-1}}(x - a_{j-1}) + F_{j-1}$$
est l'équation de la droite reliant (a_{j-1}, F_{j-1}) et (a_j, F_j) (c'est une partie du polygone des fréquences cumulées). D'après notre définition, pour avoir la médiane, on remplace simplement y par $\frac{1}{2}$ dans cette équation.

Proposition 5 | La médiane partage en deux la série

Soit x une série statistique de taille n dont les modalités sont données dans l'ordre croissant $a_1 < a_2 < \ldots < a_n$.
$$\text{Card}\{i \in [\![1\,,\,n]\!] \mid x_i \leqslant Q_2(x)\} \geqslant \frac{n}{2} \quad \text{et} \quad \text{Card}\{i \in [\![1\,,\,n]\!] \mid x_i \geqslant Q_2(x)\} \geqslant \frac{n}{2}.$$

Remarque 5

- Cela signifie qu'il y a au moins autant d'éléments plus grands que d'éléments plus petits.
- Les symboles $\leqslant, \geqslant$ dans la proposition précédente sont justifiés par le fait que $\frac{n}{2}$ n'est pas toujours un entier.

Exemple 14 Calculer la médiane des séries statistiques des exemples mentionnés.

- Exemple 9 :
- Exemple 10 :
- Exemple 11 :

GÉNÉRALISATION : QUARTILES, DÉCILES, QUANTILES. Plutôt que d'introduire une quantité qui découpe en deux une série statistique (la médiane), on peut également partager en trois quatre *etc.*. De la même façon, on définit donc pour tout x série statistique :

- *Le premier quartile* comme étant la plus petite valeur de la série (ou éventuellement la moyenne de deux valeurs), notée Q_1, telle qu'au moins $\frac{1}{4}$ des effec-

tifs aient une valeur inférieure ou égale à Q_1. Pour un regroupement en classe, si $[a_{j-1}, a_j[$ est la classe de valeurs telle que $F_{j-1} < \frac{1}{4}$ et $F_j \geqslant \frac{1}{4}$ alors :

$$Q_1(x) = v_{j-1} + \frac{a_j - a_{j-1}}{f_j} \times \left(\frac{1}{4} - F_{j-1}\right).$$

- Le deuxième quartile est la médiane $Q_2(x)$.
- Le *troisième quartile* comme étant la plus petite valeur de la série (ou éventuellement la moyenne de deux valeurs), notée $Q_3(x)$, telle qu'au moins $\frac{3}{4}$ des effectifs aient une valeur inférieure ou égale à $Q_3(x)$. Pour un regroupement en classe, si $[a_{j-1}, a_j[$ est la classe de valeurs telle que $F_{j-1} < \frac{3}{4}$ et $F_j \geq \frac{3}{4}$, alors :

$$Q_3(x) = v_{j-1} + \frac{v_j - v_{j-1}}{f_j} \times \left(\frac{3}{4} - F_{j-1}\right).$$

- On a :

$$\text{Card}\{i \in [\![1\,,\,n]\!] \mid x_i \leqslant Q_1(x)\} \geqslant \frac{n}{4} \quad \text{et} \quad \text{Card}\{i \in [\![1\,,\,n]\!] \mid x_i \geqslant Q_2(x)\} \geqslant \frac{3n}{4}.$$

On définit de manière analogue la notion de décile.

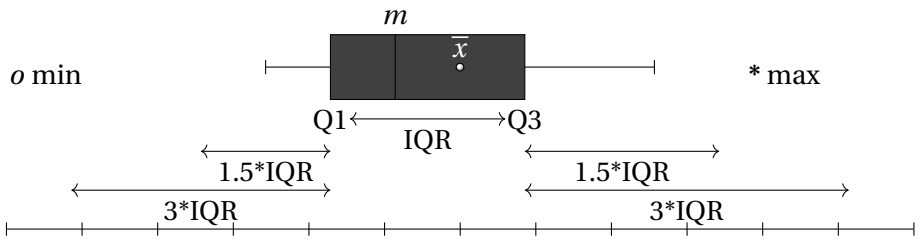
Définition 13 | Écarts

Soit x une série statistique de taille n. On appelle :
- *écart interquartile* la différence $Q_3(x) - Q_1(x)$ noté parfois « IQR ».
- *intervalle interquartile* l'intervalle $[Q_1(x), Q_3(x)]$.

La moitié au moins de la population se trouve donc dans l'intervalle interquartile.

VISUALISATION GRAPHIQUE DES QUANTILES : LE DIAGRAMME DE TUKEY (OU « BOÎTE À MOUSTACHE »). On peut représenter de manière graphique l'étendue, les quartiles et la médiane en dessinant un diagramme dit *diagramme de TUKEY* conçu de la manière suivante :

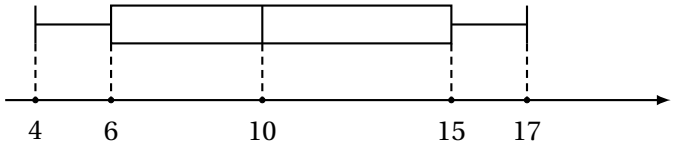
- au centre une boite allant du premier au troisième quartile, séparée en deux par la médiane ;
- de chaque côté une moustache allant du minimum au premier quartile pour l'une, et du troisième quartile au maximum pour l'autre.



REPRÉSENTATION EN BOÎTE À MOUSTACHE D'UNE SÉRIE STATISTIQUE

Exemple 15 Considérons par exemple la liste des des notes d'un devoir noté sur 20 suivant :

4	6	10	15	17



1.6. Caractéristiques de dispersion

L'idée des caractéristiques de dispersion est de donner une idée de la répartition de la série autour de sa moyenne ou de sa médiane. Les valeurs sont-elles relativement proches de la moyenne ou existe-t-il des valeurs très grandes et très petites ?

VALEURS EXTRÊMES ET ÉTENDUE.

Définition 14 | Valeurs extrêmes, étendue

Soit x une série statistique de taille n à valeurs réelles.
- Si les modalités sont en nombre fini, on appelle *valeurs extrêmes* de la série x les nombres : $\min_{i \in [\![1\,,\,n]\!]} x_i$, et $\max_{i \in [\![1\,,\,n]\!]} x_i$.
  Il s'agit donc des modalités maximales et minimales.
- Si les modalités sont regroupées en classes, on appelle *valeurs extrêmes* de la série la borne supérieure de la classe maximale et la borne inférieure de la classe minimale.
- On appelle *étendue* de la série statistique la différence entre la valeur maximale et la valeur minimale.

Remarque 6 L'étendue est facile à déterminer mais ne délivre que très peu d'informations car elle est très fortement affectée par les valeurs extrêmes. Par exemple en France le revenu annuel se situe entre 0 euros et environ 7 millions, ce qui ne nous donne pas vraiment une idée de la répartition des salaires dans la population.

VARIANCE, ÉCART-TYPE.

Définition 15 | Variance & Écart-Type

- La *variance* d'une série statistique quantitative à valeurs réelles $x = (x_1, x_2, \dots, x_n)$ de nombre de modalités finies $a_1, \dots, a_p$, est le nombre $\mathbb{V}_x$ défini par :
$$\mathbb{V}_x = \overline{x - \overline{x^2}} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{1}{n}\sum_{j=1}^{p} n_j(a_j - \overline{x})^2 = \sum_{j=1}^{p} f_j(a_j - \overline{x})^2.$$
- L'*écart-type* d'une telle série, noté σ_x, est défini par :
$$\sigma_x = \sqrt{\mathbb{V}_x}.$$

Dans le cas d'une série regroupée en classes on prendra pour valeurs a_j les centres des classes.

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

Proposition 6 | Variance nulle

Soit une série statistique quantitative à valeurs réelles $x = (x_1, x_2, \dots, x_n)$ de nombre de modalités finies $a_1, \dots, a_p$, alors :
$$\mathbb{V}_x = 0 \iff \forall i \in [\![1\,,\,n]\!], \quad x_i = x_0.$$

Preuve On raisonne par exemple avec l'expression en fréquences de la variance.

$$\begin{aligned}
&\sum_{j=1}^{p} f_j(a_j - \overline{x})^2 = 0 \\
&\iff \forall i \in [\![1\,,\,p]\!], \quad f_j(a_j - \overline{x})^2 = 0, && \text{somme de termes positifs} \\
&\iff \forall i \in [\![1\,,\,p]\!], \quad (a_j - \overline{x})^2 = 0, && f_j \neq 0 \\
&\iff \forall i \in [\![1\,,\,p]\!], \quad a_j = \overline{x}.
\end{aligned}$$

C'est ce qu'on voulait.

Proposition 7 | KÖNIG-HUYGENS

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

Preuve

$$\begin{aligned}
\mathbb{V}_x &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i^2 + \overline{x}^2 - 2x_i\overline{x}\right) = \frac{1}{n}\sum_{i=1}^{n}x_i^2 + \frac{1}{n}\sum_{i=1}^{n}\overline{x}^2 - 2\frac{\overline{x}}{n}\sum_{i=1}^{n}x_i \\
&= \overline{x^2} + \overline{x}^2 - 2\overline{x}^2 = \overline{x^2} - \overline{x}^2.
\end{aligned}$$

Remarque 7 (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 8 (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. On pourra alors faire des calculs faisant intervenir modalités, moyenne et écart-type (par exemple dans des situations d'estimation de paramètres ou de test statistique d'hypothèses).

Proposition 8 | 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]**

- **[En utilisant la définition]**
$$\begin{aligned}
\mathbb{V}_y &= \frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})^2 = \frac{1}{n}\sum_{i=1}^{n}(ax_i + b - (a\overline{x} + b))^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}a^2(x_i - \overline{x})^2 \\
&= a^2\mathbb{V}_x.
\end{aligned}$$

Pour la version écart-type, prendre simplement la racine dans l'expression.

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 connaitre pour les statistiques univariées.

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

(Modalités)

```
def sans_doublon(L):
    """
    Retourne 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 16 **(Diamètres de pièces)** 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
```

On peut ensuite tester si la quantité retournée est bien une médiane, en contrôlant la définition : au moins la moitié des observations sont supérieures ou égales à la médiane, et au moins la moitié des observations sont inférieures ou égales à la médiane.

```
def mediane_verif(L, m):
```

```
    """
    Renvoie True si m est bien une médiane de L
    """
    nb_inf = 0
    nb_sup = 0
    for x in L:
        if x >= m:
            nb_sup += 1
        if x <= m:
            nb_inf += 1
    return nb_sup >= len(L)/2 and nb_inf >= len(L)/2
```

Exemple 17 **(Diamètres de pièces)**

```
>>> etendue(L)
7
>>> m = mediane(L)
>>> mediane_verif(L, m)
True
```

Plus généralement, voici comment calculer les 3 quartiles.

(Quartiles)

```
def quartiles(L):
    """
    Retourne 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
```

2. STATISTIQUES DESCRIPTIVES BIVARIÉES

On a vu dans les sous-sections précédentes diverses manières d'extraire de l'information d'un échantillon statistique. Lorsque l'on ne dispose plus d'un seul mais de plusieurs échantillons statistiques, on peut, au delà de la simple étude des échantillons, étudier les éventuels liens entre eux, c'est l'objet de cette dernière sous-section. Pour des raisons de simplicité on se limitera à deux échantillons.

Cadre

Dans toute la suite, on travaillera uniquement avec des séries statistiques discrètes.

2.1. Série

On va s'intéresser ici à deux caractères quantitatifs d'une même population. On notera n la taille de l'échantillon étudié et (x, y) les deux caractères étudiés.

Définition 16 | Série statistique bivariée

- Soient x et y deux séries statistiques de taille n. Alors on appelle *série statistique bivariée* une famille de $\mathbb{R}^2$ du type :
  $$((x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)).$$
- Un *caractère* (x, y) de la population est une donnée *qualitative* ou *quantitative* attachée à chaque individu de la population. On notera (x_i, y_i) la valeur du caractère (x, y) pour un individu i.
- On appelle *nuage de points* associé à l'échantillon (x, y) le tracé de tous les points de coordonnées (x_i, y_i) pour $i \in [\![1, n]\!]$.
- On appelle *point moyen* du nuage le point $(\overline{x}, \overline{y})$.

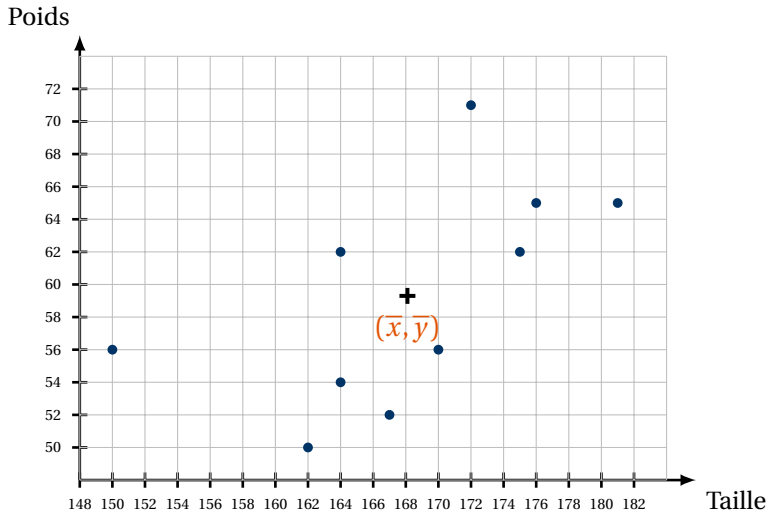
Définition 17 | Modalités

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

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 :

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

Voici alors le nuage de points associés.



2.2. Caractéristiques de position & dispersion

On définit les moyennes et variances de manière similaire au cas univarié, pour chacune des séries x et y. On définit aussi la covariance, qui nous sera très utile plus tard dans le problème de la droite de régression linéaire.

Définition 18 | 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 9 (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 9 | 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
- $\mathbb{C}_{x,x}=\dfrac{1}{n}\sum_{k=1}^{n}(x_k-\overline{x})^2=\mathbb{V}_x$
- $\begin{aligned}\mathbb{C}_{x,y}&=\frac{1}{n}\sum_{k=1}^{n}(x_k-\overline{x})(y_k-\overline{y})\\&=\frac{1}{n}\sum_{k=1}^{n}(x_ky_k-\overline{x}y_k-\overline{y}x_k+\overline{x}\cdot\overline{y}) &&\text{développement du produit}\\&=\frac{1}{n}\sum_{k=1}^{n}x_ky_k-\frac{1}{n}\sum_{k=1}^{n}\overline{x}y_k-\frac{1}{n}\sum_{k=1}^{n}\overline{y}x_k+\frac{1}{n}\sum_{k=1}^{n}\overline{x}\cdot\overline{y} &&\text{linéarité de la somme}\\&=\overline{xy}-\overline{x}\cdot\overline{y}-\overline{x}\cdot\overline{y}+\overline{x}\cdot\overline{y}\\&=\overline{xy}-\overline{x}\cdot\overline{y}\end{aligned}$
- 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.
- $\begin{aligned}\mathbb{V}_{x+y}&=\overline{(x+y)^2}-\overline{x+y}^2\\&=\overline{x^2+2xy+y^2}-(\overline{x}+\overline{y})^2\\&=\overline{x^2}+2\overline{xy}+\overline{y^2}-\overline{x}^2-\overline{y}^2-2\overline{x}\cdot\overline{y} &&\text{linéarité de l'espérance}\\&=\overline{x^2}-\overline{x}^2+\overline{y^2}-\overline{y}^2+2\left(\overline{xy}-\overline{x}\overline{y}\right) &&\text{formule de König-Huygens}\\&=\mathbb{V}_x+2\mathbb{C}_{x,y}+\mathbb{V}_y.\end{aligned}$

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

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

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

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

2.3. 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. Vous pourriez avoir ce besoin lors de vos TIPE. Notez qu'il est aussi possible d'étudier les dépendances polynomiales entre deux paramètres pour un degré quelconque, nous n'aborderons pas ce point ici.

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.

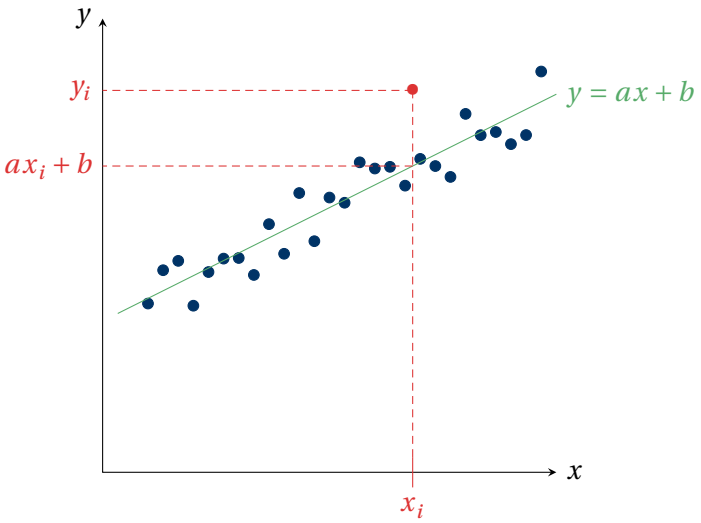
[Objectif] On veut alors chercher la droite d'équation $y = ax + b$ qui passe « le mieux » par notre nuage de points. Parfois on sait que la relation existe et on veut déterminer a et b.

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 : $y_i - ax_i - b$.

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) :

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

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



Problème de régression linéaire

ce problème de deux manières. Il s'agira donc de minimiser la fonction de deux variables :

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

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 l'espérance déjà établie, nous avons :

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

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

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 :

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

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

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

identité remarquable

Pour tout $a\in\mathbb{R}$, notons $f(b) = \mathrm{F}(a,b) = \sum_{i=1}^{n}(y_i-ax_i)^2 - 2b\sum_{i=1}^{n}(y_i-ax_i) + nb^2$. 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 :

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

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

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

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

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

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 mini-

male là où g' s'annule, *i.e.* en $a = \frac{\mathbb{C}_{x,y}}{\sigma_x^2}$. En résumé, nous avons montré :

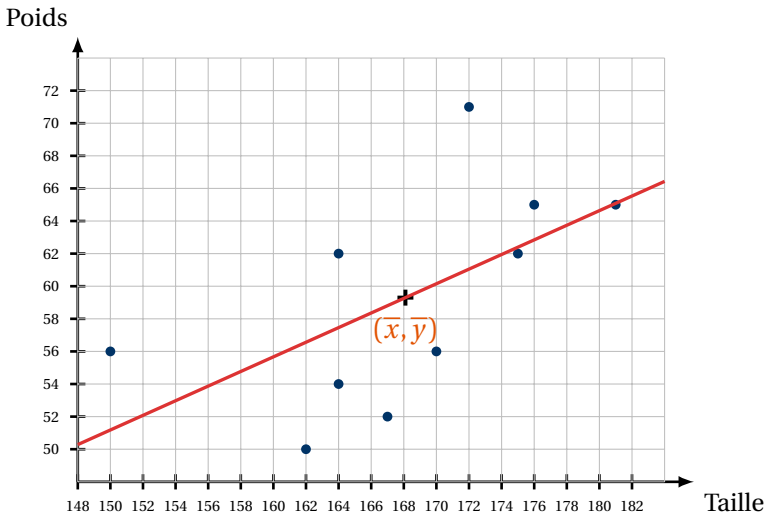
$$\forall a, b \in \mathbb{R}, \quad F(a,b) \geqslant F\left(a, \overline{y} - a\overline{x}\right) \geqslant F\left(\frac{\mathbb{C}_{x,y}}{\sigma_x^2}, \overline{y} - \frac{\mathbb{C}_{x,y}}{\sigma_x^2}\overline{x}\right).$$

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

Remarque 10 **(À propos du vocabulaire)** Pourquoi parle-t-on de « régression linéaire » ? La réponse est une erreur de traduction. Le mathématicien anglais Sir Galton étudiait les tailles des fils (y_j) en fonction de la taille de leur père (x_j) ; et a constaté un « retour à la moyenne ». En effet, les grands individus ont en moyenne des enfants plus petits qu'eux et les petits individus ont des enfants plus grand qu'eux. En Anglais le terme « retour à la moyenne » est « régression to the mean », ce terme a ensuite été mal transposé au Français.

Regardons ce que donne cette droite sur l'Exemple 18.

Exemple 19 **(Relation entre le poids et la taille)**
Dans cet exemple, les calculs fournissent : $a^\star \approx 0.448553748740826$, $b^\star \approx -16.10188516333285$ (attention à l'axe des ordonnées qui commence ici à 50 environ, il n'est donc pas surprenant d'avoir une ordonnée à l'origine négative).



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 : $\boxed{r^2(x,y) = \rho_{x,y}^2} = \frac{\mathbb{C}_{x,y}^2}{\mathbb{V}_x \mathbb{V}_y} \in [0,1]$.

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 11 **(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 : $r^2(x,y) \geqslant 0.9$.

Autres ajustements se ramenant à une régression linéaire. On peut penser à beaucoup d'ajustements. Par exemple :

1. si l'on souhaite tester la relation $y = \lambda e^{\alpha x}$, avec $(\alpha,\lambda) \in \mathbb{R} \times \mathbb{R}^{+\star}$ avec y série statistique strictement positive, on peut constater qu'elle est équivalente à $(\ln y = \ln\lambda + \alpha x$: on fait alors la régression sur $(x, \ln y)$. Comment retrouver α, λ à partir de $a^\star, b^\star$?

2. Si l'on souhaite tester la relation $y = a\ln x + b$, avec $(a,b) \in \mathbb{R}$, on peut faire alors la régression sur $(\ln x, y)$.

2.4. 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)))
```

(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)
```

(Coefficients de régression)

```
def regression_lin(X, Y):
    """
    Retourne 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 es séries statistiques ci-après, la série X correspondant à des relevés de poids, et Y de taille.

```
>>> 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.