Chapitre n° 15
Statistiques descriptives

1 Préambule

1.1 Les statistiques

Les statistiques sont la science qui étudie la collecte, le traitement, l'analyse, la présentation et l'interprétation de données d'observation expérimentale. Son principal objectif est l'extraction d'informations pertinentes (représentations graphiques, indicateurs chiffrés représentatifs appelées «statistiques», liens de corrélation, etc.,) d'un tableau de données brutes. On parle aussi de la statistique (ou de la science des données) pour ne pas confondre le domaine scientifique et les caractéristiques calculées.

Bien que les statistiques, comme d'autres sciences, utilisent de nombreux outils mathématiques (en particulier issus de la théorie des probabilités), il ne s'agit pas à proprement parler d'une branche des mathématiques puisque son champ d'étude (les données observables) n'est pas abstrait. On parle plutôt de «mathématiques appliquées».

Les statistiques sont désormais utilisées avec efficacité dans presque tous les domaines scientifiques : sciences naturelles (agronomie, biologie, écologie, géologie, médecine, etc.), sciences humaines et sociales (économie, géographie, politique, sociologie, etc.), sciences expérimentales (chimie, ingénierie, physique appliquée, etc.) et sciences exactes (astronomie, informatique, mathématiques, physique théorique, etc.).

En BCPST, les deux branches principales des statistiques abordées en cours sont :

— les statistiques descriptives qui regroupent les méthodes pour décrire de manière exhaustive un ensemble complet d'observations disponibles (en maths-info BCPST1),

— les statistiques inférentielles qui regroupent les techniques permettant de prédire les caractéristiques d'un groupe général (la population) à partir d'un groupe particulier (l'échantillon) en fournissant une mesure de certitude de la prédiction (tests statistiques en info BCPST2).

Dans les deux cas, les notions présentées sont mises en pratique en TP (SVT, physique-chimie, informatique) et en TIPE. Ces concepts statistiques de base sont indispensables à toute carrière scientifique.

Il existe de nombreuses autres branches des statistiques que vous pourrez être amenés à étudier plus tard : l'analyse des données (ou «data science», qui utilise des outils mathématiques plus généraux que ceux issus des probabilités, comme la logique, les ensembles, les applications, l'algèbre linéaire, etc.), le «big data» (lorsque le nombre de données disponibles est colossal et impose des techniques particulières), l'apprentissage statistique (un des champs d'étude du «machine learning» en informatique et plus généralement de l'intelligence artificielle), la biostatistique (regroupant les méthodes spécifiques à l'agronomie, la biologie, l'écologie, la géologie, la médecine, etc.), la physique statistique (permettant de décrire le comportement d'un système à l'échelle macroscopique à partir des règles physiques régissant l'évolution de ses particules à l'échelle microscopique, comme par exemple la thermodynamique), la finance quantitative, etc.

1.2 Vocabulaire

On considère une expérience scientifique quelconque à l'issue de laquelle des données sont observées, mesurées et collectées.

Définition 1

Chaque unité statistique sur laquelle sont observées une ou plusieurs données est appelée un ***individu****. La* ***population*** *est l'ensemble des individus et sa* ***taille*** *est le nombre total d'individus.*

Définition 2

L'ensemble des données collectées est appelé une ***série statistique****. Les données d'un même type observées sur chaque individu de la population sont appelées un* ***caractère*** *qui peut être* ***qualitatif*** *(couleur, forme, nomenclature, etc.) ou* ***quantitatif*** *(longueur, masse, concentration, etc.). On dit que la série statistique est* ***univariée*** *lorsqu'un seul caractère est observé dans la population,* ***bivariée*** *pour deux caractères, ou plus généralement* ***multivariée****.*

Une manière de présenter tous les résultats de l'observation d'un caractère est de fournir la liste des données brutes. Plus précisément, si x est un caractère et n est la taille de la population, alors la liste des observations de x est de la forme $(x_1, x_2, \dots, x_n)$ où chaque x_i est une valeur de x (chiffrée, lorsque x est quantitatif, ou non).

Certaines de ces valeurs peuvent égales. Pour simplifier, on peut alors regrouper les valeurs identiques en précisant leur nombre.

Définition 3

Les valeurs différentes prises par un caractère sont appelées ses ***modalités****. Pour chaque modalité, le nombre d'individus partageant cette même valeur est appelé son* ***effectif****. En particulier, la taille de la population est égale à la somme des effectifs.*

Si x est un caractère et p est son nombre de modalités, alors on peut présenter les observations de x par un tableau de la forme :

x_1	x_2	$\dots$	x_p
n_1	n_2	$\dots$	n_p

où chaque n_i est l'effectif de la modalité x_i. Ainsi, la taille de la population est égale à $\boxed{n = \sum_{i=1}^{p} n_i}$.

Attention. Il ne faut pas confondre la taille n de la population (égale au nombre total d'observations) et le nombre p de modalités (egal au nombre d'observations différentes). Plus précisément, et même si la notation est ambigüe, il ne faut pas confondre la liste $(x_1, x_2, \dots, x_n)$ des n valeurs de x (dont certaines peuvent être égales) et l'ensemble $\{x_1, x_2, \dots, x_p\}$ des p modalité de x (qui sont toutes différentes).

Définition 4

Dans le cas d'un caractère quantitatif, si ses modalités sont trop nombreuses, elles peuvent être regroupées par ***classes****, c'est-à-dire par intervalles de valeurs continues. Pour chaque classe de modalités, sa longueur (en tant qu'intervalle) est appelée son* ***amplitude****, et le nombre d'individus partageant cette même plage de valeurs est son* ***effectif****.*

Ainsi, les observations d'un caractère quantitatif x dont les modalités sont regroupées par classes peuvent être présentées par un tableau de la forme :

$[a_1, b_1]$	$[a_2, b_2]$	$\dots$	$[a_p, b_p]$
n_1	n_2	$\dots$	n_p

où chaque n_i est l'effectif de la classe $[a_i, b_i]$ (et la taille de la population est égale à $n = \sum_{i=1}^{p} n_i$).

Attention. Les classes de modalités doivent être deux à deux disjointes. En particulier, si les classes de modalités sont consécutives, c'est-à-dire si $a_1 < b_1 = a_2 < b_2 = a_3 < \cdots < b_{p-1} = a_p < b_p$ alors elles doivent être de forme $[a_1, b_1[$, $[a_2, b_2[$, $\dots$, et $[a_p, b_p[$.

Conseil. En pratique, lorsqu'on regroupe les modalités par classes, on essaie d'utiliser des classes de même amplitude, c'est-à-dire telles que $b_1 - a_1 = b_2 - a_2 = \cdots = b_p - a_p$. De plus, pour simplifier les calculs, on suppose que les modalités sont uniformément réparties au sein de chaque classe. Ainsi, on pourra remplacer une classe de modalité $[a_i, b_i[$ par son **centre** $c_i = (a_i + b_i)/2$ dans les calculs.

Définition 5

Pour chaque modalité (ou chaque classe de modalités) d'un caractère, on définit sa ***fréquence*** *comme le rapport de son effectif par la taille de la population.*

Propriété 1

Les fréquences sont des réels de $[0,1]$ *et leur somme est égale à* 1*.*

Démonstration. Si on note n_i l'effectif d'une modalité x_i (ou d'une classe de modalités $[a_i, b_i[$), alors sa fréquence est égale à $\mathrm{f}_i = n_i/n$ où $n = \sum_{i=1}^{p} n_i$ est la taille de la population. Donc $\mathrm{f}_i \in [0,1]$, car $0 \leqslant n_i \leqslant n$, et $\sum_{i=1}^{p} \mathrm{f}_i = \sum_{i=1}^{p} n_i/n = n/n = 1$ par linéarité de la somme.

□

2 Statistiques univariées

On considère une série statistique univariée, c'est-à-dire constituée par les observations d'un seul caractère noté x qui peut être qualitatif ou quantitatif (et dont les modalités peuvent être regroupées par classes dans le deuxième cas).

2.1 Représentations graphiques

Définition 6

Un ***diagramme circulaire*** *d'un caractère à p modalités (ou classes de modalités) est un disque découpé en p secteurs angulaires dont l'angle de chaque secteur est proportionnel à l'effectif de la modalité correspondante.*

Conseil. Si on note n_i l'effectif d'une modalité x_i (ou d'une classe de modalités $[a_i, b_i[$) et $\mathrm{f}_i = n_i/n$ sa fréquence où $n = \sum_{i=1}^{p} n_i$ est la taille de la population, alors l'angle θ_i du secteur correspondant dans le diagramme circulaire s'obtient par la règle de proportionnalité suivante :

$$\begin{array}{|c|c|}\hline n & 2\pi \\ \hline n_i & \theta_i \\ \hline \end{array} \quad \text{donc} \quad \theta_i = \frac{2\pi n_i}{n} = 2\pi \mathrm{f}_i.$$

En particulier, on remarque que l'angle de chaque secteur est également proportionnel à la fréquence de la modalité correspondante.

Remarque. Puisque l'aire d'un secteur angulaire d'angle θ_i et de rayon R est égale à $\theta_i R$, l'aire de chaque secteur est également proportionnelle à l'effectif n_i (et donc à la fréquence f_i) de la modalité correspondante. Cette remarque justifie la pertinence graphique du diagramme circulaire : nos yeux humains observent plus naturellement les aires que les angles des secteurs.

Conseil. Le diagramme circulaire est particulièrement recommandé pour les caractères qualitatifs car ses modalités n'ont pas besoin d'être ordonnées selon un axe, contrairement au diagramme en bâtons (voir ci-dessous).

Définition 7

Un ***diagramme en bâtons*** *d'un caractère x quantitatif à p modalités notées* $x_1 < x_2 < \cdots < x_p$ *est un graphique représentant des barres rectangulaires de longueurs égales aux effectifs et basées sur les modalités disposées le long d'un axe de coordonnées.*

Remarque. Graphiquement, il revient au même de représenter des barres de longueurs égales aux fréquences (quitte à changer l'échelle du deuxième axe de coordonnées) puisque chaque effectif n_i est proportionnel à la fréquence correspondante $\mathrm{f}_i = n_i/n$ où $n = \sum_{i=1}^{p} n_i$ est la taille de la population.

Attention. On peut également utiliser un diagramme en bâtons pour un caractère qualitatif, mais dans ce cas l'ordre des modalités sur l'axe est choisi de manière complètement arbitraire. En pratique, il est préférable d'utiliser un diagramme circulaire pour un caractère qualitatif.

Conseil. Dans le cas où les modalités sont regroupées par classes, le graphique analogue au diagramme en bâtons est l'histogramme (voir ci-dessous) où les barres sont remplacées par des rectangles.

Définition 8

*Un **histogramme** d'un caractère quantitatif dont les modalités sont regroupées par classes est un graphique représentant des rectangles de hauteurs (ou bien d'aires) égales aux effectifs et basées sur les classes de modalités disposées le long d'un axe de coordonnées.*

Remarque. L'aire d'un rectangle basé sur une classe de modalité $[a_i, b_i[$ est de la forme $(b_i - a_i)h_i$ où h_i est sa hauteur. En particulier, si les classes ont même amplitude (ce qui est recommandé), alors l'aire de chaque rectangle est proportionnelle à sa hauteur (puisque le coefficient de proportionnalité $b_i - a_i$, égal à l'amplitude, est identique pour toutes les classes). Dans ce cas, il revient donc au même de choisir les hauteurs ou bien les aires égales aux effectifs (quitte à changer l'échelle du deuxième axe de coordonnées).

Attention. Si les classes n'ont pas la même amplitude (ce qui n'est pas recommandé), deux classes de même effectif peuvent avoir des aires différentes (si on fixe les hauteurs égales aux effectifs) ou bien des hauteurs différentes (si on fixe les aires égales aux effectifs). Le choix entre les hauteurs ou bien les aires égales aux effectifs dépend de l'effet visuel souhaité, mais il doit être clairement précisé en légende.

Remarque. Dans tous les cas, on peut remplacer les effectifs par les fréquences, cela revient graphiquement au même (quitte à changer l'échelle du deuxième axe de coordonnées).

Définition 9

*Soit x un caractère quantitatif. On définit l'**effectif cumulé** de tout réel t comme le nombre d'individus de modalité inférieure ou égale à t, c'est-à-dire la somme des effectifs n_i des modalités $x_i \leqslant t$. On définit de même la **fréquence cumulée** de tout $t \in \mathbb{R}$ qu'on note $F_x(t)$. La courbe représentative de la fonction F_x est appelée la **courbe des fréquences cumulées**.*

Conseil. Si on note $x_1 < x_2 < \cdots < x_p$ les modalités et $\mathrm{f}_1, \mathrm{f}_2, \ldots, \mathrm{f}_p$ leurs fréquences correspondantes, alors la fréquence cumulée est égale à : 0 sur $]-\infty, x_1[$, f_1 sur $[x_1, x_2[$, $\mathrm{f}_1 + \mathrm{f}_2$ sur $[x_2, x_3[$, $\mathrm{f}_1 + \mathrm{f}_2 + \mathrm{f}_3$ sur $[x_3, x_4[$, …, et $\sum_{i=1}^{p} \mathrm{f}_i = 1$ sur $[x_p, +\infty[$. Autrement dit, la fréquence cumulée de tout $t \in \mathbb{R}$ est égale à :

$$F_x(t) = \begin{cases} 0 & \text{si } t < x_1 \\ \sum_{i=1}^{k} \mathrm{f}_i & \text{si } x_k \leqslant t < x_{k+1} \text{ (où } k \in [\![1, p-1]\!]) \\ 1 & \text{si } x_p \leqslant t \end{cases}.$$

On reconnaît une fonction constante par morceaux (on dit aussi «en escalier»).

Conseil. Dans le cas où les modalités sont regroupées par classes consécutives de la forme $[a_1, b_1[$, $[a_2, b_2[$, …, $[a_p, b_p[$ avec $a_1 < b_1 = a_2 < b_2 = a_3 < \cdots < b_{p-1} = a_p < b_p$, et si on note f_i la fréquence d'une classe $[a_i, b_i[$, alors la fréquence cumulée est égale à : 0 sur $]-\infty, a_1]$, f_1 en $b_1 = a_2$, $\mathrm{f}_1 + \mathrm{f}_2$ en $b_2 = a_3$, $\mathrm{f}_1 + \mathrm{f}_2 + \mathrm{f}_3$ en $b_3 = a_4$, …, et $\sum_{i=1}^{p} \mathrm{f}_i = 1$ sur $[b_p, +\infty[$. Pour calculer la fréquence cumulée dans une classe $[a_k, b_k[$, on suppose que les modalités sont uniformément réparties au sein de la classe, et donc que la courbe représentative de F_x est le segment de droite reliant les extrémités $(a_k, F_x(a_k) = \sum_{i=1}^{k-1} \mathrm{f}_i)$ et $(b_k, F_x(b_k) = \sum_{i=1}^{k} \mathrm{f}_i)$ qui admet pour équation :

$$y = \left(\frac{F_x(b_k) - F_x(a_k)}{b_k - a_k}\right)(x - a_k) + F_x(a_k) = \frac{\mathrm{f}_k}{b_k - a_k}(x - a_k) + \sum_{i=1}^{k-1} \mathrm{f}_i.$$

On reconnaît une fonction affine par morceaux (on dit aussi «polygonale»).

Propriété 2

La fréquence cumulée F_x est une fonction réelle définie sur $\mathbb{R}$ qui vérifie les propriétés suivantes :

- *F_x est croissante ;*
- *F_x est constante égale à 0 sur l'intervalle des valeurs inférieures à la plus petite modalité ;*
- *F_x est constante égale à 1 sur l'intervalle des valeurs supérieures à la plus grande modalité ;*
- *si les modalités ne sont pas regroupées par classes, alors F_x est en escalier, c'est-à-dire constante entre deux modalités consécutives et discontinue en chaque modalité ;*
- *si les modalités sont regroupées par classes (et uniformément réparties au sein de chaque classe), alors F_x est polygonale, c'est-à-dire affine sur chaque classe de modalités et continue sur $\mathbb{R}$.*

Démonstration. Par définition de la fréquence cumulée et en supposant, dans le cas où les modalités sont regroupées par classes, que les modalités sont uniformément réparties au sein de chaque classe.

□

Remarque. Graphiquement, la courbe représentative des effectifs cumulés a exactement la même allure que celle des fréquences cumulées puisque ça revient à multiplier l'échelle des ordonnées par un facteur égal à la taille de la population.

2.2 Caractéristiques de position

Une caractéristique de position permet de résumer une liste d'observations à une seule valeur. Autrement dit, c'est un indicateur représentant la tendance générale des données d'une série statistique univariée.

Les caractéristiques de position présentées ici sont le mode, la médiane et la moyenne (arithmétique), mais il en existe d'autres.

Les caractéristiques de position sont à distinguer des caractéristiques de dispersion (voir la section suivante) qui permettent de mesurer la variabilité des données.

Définition 10

*Le **mode** d'un caractère est sa modalité d'effectif maximal. Dans le cas d'un caractère quantitatif dont les modalités sont regroupées par classes, la classe d'effectif maximal est appelée la **classe modale**.*

Remarque. Il revient bien sûr au même de considérer la modalité (ou la classe) de fréquence maximale.

Attention. Le mode (ou la classe modale) n'est pas toujours unique. On dit que le caractère est **unimodal** lorsqu'une seule modalité (ou une seule classe) apparaît plus fréquemment que les autres, **bimodal** lorsque deux modalités (ou deux classes) ont le même effectif strictement supérieur aux autres, etc.

Attention. Il n'est pas pertinent de parler de classe modale lorsque les classes de modalités n'ont pas la même amplitude. En effet, plus une classe est de grande amplitude, plus son effectif est important sans que cela soit significatif pour la position de la série statistique.

Remarque. Le mode est la seule caractéristique qu'on peut calculer pour les caractères qualitatifs. Les autres caractéristiques de position comme la médiane et la moyenne (voir ci-dessous) ainsi que les caractéristiques de dispersion (voir la section suivante) ne sont pas définies si le caractère n'est pas quantitatif.

Définition 11

*Soit x un caractère quantitatif. On définit la **médiane**, notée Q_2, comme une valeur qui partage la population en deux : la moitié des individus ont des modalités inférieures à Q_2 alors que l'autre moitié ont des modalités supérieures.*

Attention. La médiane n'est pas unique en général ! Plus précisément, si la taille de la population est un nombre pair de la forme $n = 2m$ et qu'on ordonne la liste des observations de x :

$$\underbrace{x_1 \leqslant x_2 \leqslant \cdots \leqslant x_m}_{\text{moitié inférieure}} \leqslant \underbrace{x_{m+1} \leqslant \cdots \leqslant x_{2m}}_{\text{moitié supérieure}}$$

alors Q_2 peut prendre n'importe quelle valeur entre x_m et x_{m+1}. Par contre, si la taille de la population est un nombre impair de la forme $n = 2m+1$ et qu'on ordonne la liste des observations alors $Q_2 = x_{m+1}$.

Propriété 3

On a $F_x(Q_2) = 1/2$. En particulier, la médiane est égale à l'abscisse du point d'intersection entre la courbe des fréquences cumulées et la droite horizontale d'équation $y = 1/2$.

Démonstration. Par définitions, l'effectif cumulé de la médiane, c'est-à-dire le nombre d'individus de modalité inférieure ou égale à Q_2, est égal à la moitié de la population. On en déduit que $F_x(Q_2) = 1/2$.

□

Remarque. Puisque F_x prend des valeurs croissantes de 0 à 1, il suffit, d'après le théorème des valeurs intermédiaires, qu'elle soit continue pour que l'équation $F_x(Q_2) = 1/2$ admette une solution.

- Si les modalités sont regroupées par classes, alors F_x est polygonale (donc continue) donc la courbe des fréquences cumulées intersecte la droite horizontale d'équation $y = 1/2$ et la médiane est bien définie graphiquement (et est unique). De plus, si les modalités sont uniformément réparties au sein de chaque classe, l'équation $F_x(Q_2) = 1/2$ à résoudre pour calculer la médiane est de degré 1.
- Si les modalités ne sont par regroupées par classes, alors F_x est en escalier donc la courbe des fréquences cumulées et la droite horizontale d'équation $y = 1/2$ s'intersectent aucune fois (la hauteur $1/2$ passe entre deux «marches» de l'escalier) ou bien une infinité de fois (une «marche» de l'escalier a pour hauteur $1/2$). Dans le premier cas, la médiane n'est pas bien définie graphiquement : Q_2 est égale à l'abscisse du point de discontinuité entre les deux «marches» où passe la hauteur $1/2$. Dans le deuxième cas, la médiane n'est pas unique : Q_2 peut prendre n'importe quelle valeur d'abscisse d'un point de la «marche» qui a pour hauteur $1/2$.

Définition 12

*Soit x un caractère quantitatif observé sur une population de taille n. On note $(x_1, x_2, \dots, x_n) \in \mathbb{R}^n$ la liste des valeurs de x. La **moyenne** de x, notée $\overline{x}$, est définie par :*

$$\boxed{\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i}.$$

Conseil. Pour simplifier les calculs, on peut utiliser les valeurs différentes de x. Plus précisément, si on note $\{x_1, x_2, \dots, x_p\}$ l'ensemble des modalités de x, n_i l'effectif de chaque modalité x_i (donc $n = \sum_{i=1}^{p} n_i$) et $\mathrm{f}_i = n_i/n$ la fréquence de chaque modalité x_i, alors la moyenne s'obtient par la formule suivante :

$$\boxed{\overline{x} = \frac{1}{n}\sum_{i=1}^{p} n_i x_i = \sum_{i=1}^{p} \mathrm{f}_i x_i}.$$

Conseil. Dans le cas où les modalités sont regroupées par classes de la forme $[a_1, b_1[$, $[a_2, b_2[$, …, $[a_p, b_p[$, on peut calculer la moyenne en supposant que les modalités sont uniformément réparties au sein de chaque classe. Il suffit alors d'utiliser le centre $c_i = (a_i + b_i)/2$ de chaque classe $[a_i, b_i[$ et on obtient simplement la formule suivante :

$$\boxed{\overline{x} = \frac{1}{n}\sum_{i=1}^{p} n_i c_i = \sum_{i=1}^{p} \mathrm{f}_i c_i}.$$

Exercice 1. Écrire en Python une fonction `moyenne` qui prend en argument une liste de réels et qui renvoie la moyenne de ses éléments.

Propriété 4

*Soient x et y deux caractères quantitatifs en **relation affine**, c'est-à-dire tels que $y = \alpha x + \beta$ où α et β sont deux constantes réelles. Alors :*

$$\overline{y} = \alpha\overline{x} + \beta.$$

Démonstration. Par linéarité de la somme :

$$\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{n}\sum_{i=1}^{n}(\alpha x_i + \beta) = \alpha \underbrace{\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)}_{=\overline{x}} + \frac{\beta}{n}\underbrace{\left(\sum_{i=1}^{n} 1\right)}_{=n} = \alpha\overline{x} + \beta.$$

□

Remarque. On dit que la moyenne est linéaire. Ainsi, pour changer d'unité une moyenne (par exemple pour passer une vitesse moyenne de m/s à km/h, ou une température moyenne de °C à °F), il est inutile de convertir toutes les données, il suffit de convertir le résultat.

Attention. La moyenne est sensible aux valeurs extrêmes, contrairement à la médiane. Par exemple, si on augmente seulement la valeur de la plus grande modalité alors la moyenne augmente aussi alors que la médiane ne change pas.

Remarque. Ainsi, comparer la moyenne et la médiane donne une indication sur la dispersion des données :

- Si $Q_2 < \overline{x}$ alors les écarts entre les plus grandes modalités et la médiane sont plus importants que ceux entre la médiane et les plus petites modalités. Par conséquent, les modalités supérieures à la médiane (qui correspondent à une moitié de la population) sont plus dispersées que les modalités inférieures à la médiane (qui correspondent à l'autre moitié). Dans ce cas, la série statistique est dite **désaxée vers la droite** (par exemple la série $(1, 2, 4, 7, 11)$).
- Inversement, si $\overline{x} < Q_2$ (par exemple la série $(3, 6, 8, 9, 9)$), la série statistique est dite **désaxée vers la gauche** : les modalités supérieures à la médiane sont moins dispersées que les modalités inférieures à la médiane.
- Si $\overline{x} = Q_2$ (par exemple la série $(2, 5, 6, 8, 9)$), on dit que la série statistique est **symétrique**.

Dans le cas où la série statistique est fortement dissymétrique, la médiane est donc un meilleur indicateur de la position que la moyenne à cause de la sensibilité de cette caractéristique aux valeurs extrêmes.

2.3 Caractéristiques de dispersion

Dans cette section, on suppose que le caractère x observé est quantitatif.

Une caractéristique de dispersion permet de mesurer la variabilité des données d'une série statistique univariée autour des caractéristiques de position. Par exemple, comparer la moyenne et la médiane afin d'étudier la symétrie de la série statistique fournit une indication sur la dispersion. Mais il existe d'autres indicateurs plus simples.

Les caractéristiques de dispersion présentées ici sont la variance, l'écart-type, les quartiles et les déciles.

Définition 13

*Soit x un caractère quantitatif observé sur une population de taille n. On note $(x_1, x_2, \ldots, x_n) \in \mathbb{R}^n$ la liste des valeurs de x. La **variance** de x, notée s_x^2, est définie comme la moyenne des carrés des écarts à la moyenne, c'est-à-dire :*

$$\boxed{s_x^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2}.$$

*La racine carrée de la variance est appelée l'**écart-type** et on le note :*

$$\boxed{s_x = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2}}.$$

Remarque. La variance est plus simple à calculer mais l'intérêt de l'écart-type est qu'il s'exprime dans les mêmes unités que les données observées, comme la moyenne.

Conseil. Comme pour la moyenne, on peut simplifier les calculs à l'aide des valeurs différentes de x. Ainsi, si on note $\{x_1, x_2, \dots, x_p\}$ l'ensemble des modalités de x, n_i l'effectif de chaque modalité x_i (donc $n = \sum_{i=1}^{p} n_i$) et $\mathrm{f}_i = n_i/n$ la fréquence de chaque modalité x_i, alors l'écart-type peut s'écrire :

$$\boxed{s_x = \sqrt{\frac{1}{n}\sum_{i=1}^{p} n_i\Big(x_i - \overline{x}\Big)^2} = \sqrt{\sum_{i=1}^{p} \mathrm{f}_i\Big(x_i - \overline{x}\Big)^2}}.$$

De plus, dans le cas où les modalités sont regroupées par classes de centres c_1, c_2, …, c_p et si on suppose que les modalités sont uniformément réparties au sein de chaque classe, alors l'écart-type peut s'écrire :

$$\boxed{s_x = \sqrt{\frac{1}{n}\sum_{i=1}^{p} n_i\Big(c_i - \overline{x}\Big)^2} = \sqrt{\sum_{i=1}^{p} \mathrm{f}_i\Big(c_i - \overline{x}\Big)^2}}.$$

Attention. En statistiques inférentielles, la définition de la variance et l'écart-type, souvent notés σ_x^2 et σ_x, est différente : le n au dénominateur est remplacé par $n-1$. Il faut donc faire attention aux valeurs renvoyées par une calculatrice, un logiciel ou une application : en fonction des marques et des modèles, on peut obtenir des valeurs différentes (et également des notations s_x et σ_x inversées). Pour vérifier la définition utilisée, il suffit par exemple de calculer l'écart-type de la série $(1, 2, 3)$: on obtient $s_x = \sqrt{2/3} \approx 0,816$ en statistiques descriptives et $\sigma_x = \sqrt{2/2} = 1$ en statistiques inférentielles.

Théorème 1 (*Formule de König-Huygens*)

La variance est égale à la différence entre la moyenne des carrés et le carré de la moyenne, c'est-à-dire :

$$\boxed{s_x^2 = \overline{x^2} - \overline{x}^2}.$$

Démonstration. On a :

$$\begin{aligned}
s_x^2 &= \frac{1}{n}\sum_{i=1}^{n}\Big(x_i - \overline{x}\Big)^2 \quad \text{par définition de la variance} \\
&= \frac{1}{n}\sum_{i=1}^{n}\Big(x_i^2 - 2\overline{x}x_i + \overline{x}^2\Big) \\
&= \frac{1}{n}\sum_{i=1}^{n} x_i^2 - 2\overline{x} \times \underbrace{\frac{1}{n}\sum_{i=1}^{n} x_i}_{=\overline{x}} + \overline{x}^2 \times \underbrace{\frac{1}{n}\sum_{i=1}^{n} 1}_{=1} \quad \text{par linéarité de la somme} \\
&= \overline{x^2} - 2\overline{x}^2 + \overline{x}^2 = \overline{x^2} - \overline{x}^2.
\end{aligned}$$

□

Remarque. En particulier, on obtient que la moyenne des carrés est supérieure au carré de la moyenne (car $s_x^2 \geqslant 0$), ce qui est un résultat non trivial. On peut aussi le retrouver graphiquement pour deux valeurs : le milieu d'une corde reliant deux points de la parabole d'équation $y = x^2$ a une ordonnée supérieure à l'image de la moyenne des abscisses de ces deux points (la fonction $x \mapsto x^2$ est dite convexe).

Conseil. En pratique, la formule de König-Huygens permet de simplifier les calculs de la variance, et donc aussi de l'écart-type.

Exercice 2. Écrire en Python deux fonctions `ecarttype1` et `ecarttype2` qui prennent en argument une liste de réels et qui renvoient l'écart-type de ses éléments à l'aide de la définition pour `ecarttype1` et de la formule de König-Huygens pour `ecarttype2`. Dans les deux cas, on pourra utiliser la fonction `moyenne` écrite dans l'exercice 1.

Propriété 5

*Soient x et y deux caractères quantitatifs en **relation affine**, c'est-à-dire tels que $y = \alpha x + \beta$ où α et β sont deux constantes réelles. Alors :*

$$s_y = |\alpha| s_x.$$

Démonstration. On a :

$$\begin{aligned}
s_y^2 &= \overline{y^2} - \overline{y}^2 \quad \text{d'après la formule de König-Huygens} \\
&= \overline{(\alpha x + \beta)^2} - \Big(\underbrace{\overline{\alpha x + \beta}}_{=\alpha\overline{x}+\beta}\Big)^2 \\
&= \underbrace{\overline{\alpha^2 x^2 + 2\alpha\beta x + \beta^2}}_{=\alpha^2\overline{x^2}+2\alpha\beta\overline{x}+\beta^2} - (\alpha\overline{x} + \beta)^2 \quad \text{car la moyenne est linéaire} \\
&= \alpha^2\overline{x^2} + 2\alpha\beta\overline{x} + \beta^2 - \left(\alpha^2\overline{x}^2 + 2\alpha\beta\overline{x} + \beta^2\right) \\
&= \alpha^2\left(\overline{x^2} - \overline{x}^2\right) \quad \text{après simplifications} \\
&= \alpha^2 s_x^2 \quad \text{d'après la formule de König-Huygens.}
\end{aligned}$$

Donc :

$$s_y = \sqrt{s_y^2} = \sqrt{\alpha^2 s_x^2} = |\alpha| s_x.$$

□

Remarque. Ainsi, l'écart-type n'est pas linéaire. On dit qu'il est invariant par translation ($s_{x+c} = s_x$ pour toute constante c) et homogène par dilatation ($s_{cx} = c s_x$ pour toute constante positive c).

Propriété 6

L'écart-type est positif. De plus, il s'annule si et seulement si toutes les observations sont égales entre elles.

Démonstration. On a $s_x \geqslant 0$ par définition. De plus :

$$s_x = 0 \iff \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2} = 0 \iff \sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2 = 0 \iff \left(\forall i \in [\![1, n]\!],\ (x_i - \overline{x})^2 = 0\right)$$

en reconnaissant une somme de réels positifs. Par conséquent :

$$s_x = 0 \iff \forall i \in [\![1, n]\!],\ x_i = \overline{x}.$$

Ainsi, l'écart-type s'annule si et seulement si toutes les valeurs x_i sont égales à une même constante (qui est nécessairement égale à $\overline{x}$).

□

Remarque. Cette propriété justifie que l'écart-type est une caractéristique de dispersion : si s_x est proche de 0 alors les modalités sont proches de $\overline{x}$ et donc peu dispersées. Réciproquement, plus l'écart-type est grand et plus les modalités sont dispersées atour de la moyenne.

Définition 14

*On définit les premier et troisième **quartiles**, notés Q_1 et Q_3, comme les abscisses des points d'intersection entre la courbe des fréquences cumulées et les droites horizontales d'équation $y = 1/4$ et $y = 3/4$. Ainsi, un quart des individus ont des modalités inférieures à Q_1 et un quart des individus ont des modalités supérieures à Q_3. On définit de même les premier et neuvième **déciles**, notés D_1 et D_9, avec les droites horizontales d'équation $y = 1/10$ et $y = 9/10$: un dixième des individus ont des modalités inférieures à D_1 et un dixième des individus ont des modalités supérieures à D_9.*

Attention. Comme pour la médiane, les quartiles et les déciles ne sont pas uniques en général et ne sont parfois pas bien définis graphiquement.

Remarque. Le deuxième quartile, qui correspond à l'abscisse du point d'intersection entre la courbe des fréquence cumulées et la droite horizontale d'équation $y = 2/4 = 1/2$, est donc égal à la médiane. Ce qui justifie la notation Q_2.

Remarque. La quantité $Q_3 - Q_1$ est appelée l'**écart interquartile**. Puisque la moitié des individus ont des modalités comprises entre Q_1 et Q_3, plus l'écart interquartile est grand et plus les modalités sont dispersées. Cette remarque justifie que les quartiles sont des indicateurs de la dispersion. On peut faire un raisonnement similaire avec les déciles.

3 Statistiques bivariées

On considère désormais une série statistique bivariée, c'est-à-dire constituée par les observations de deux caractères notés x et y qu'on suppose quantitatifs.

3.1 Généralités

La liste des observations de x et de y est de la forme :

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$$

où n est la taille de la population. On peut également présenter ces données en regroupant les couples d'observations identiques par un tableau de la forme :

(x_1, y_1)	(x_2, y_2)	$\ldots$	$(x_p, y_{p)}$
n_1	n_2	$\ldots$	n_p

où p est le nombre de modalités, c'est-à-dire le nombre de couples d'observations différents, et chaque n_i est l'effectif de la modalité (x_i, y_i), donc $n = \sum_{i=1}^{p} n_i$.

Pour simplifier, on ne regroupera pas les modalités par classes pour les statistiques bivariées.

Définition 15

*Le **nuage de points** de la série statistique bivariée des caractères quantitatifs x et y est la représentation graphique des n points (x_i, y_i) dans un plan muni d'un repère orthogonal.*

Remarque. L'axe des abscisses correspond donc aux modalités de x alors que l'axe des ordonnées correspond aux modalités de y. Puisque les observations de x sont de type différent de celles de y (et qui s'expriment en général avec des unités différentes), le repère utilisé n'a pas besoin d'être orthonormé.

Conseil. Afin d'obtenir un nuage de points clair et significatif, il convient de choisir sur chaque axe de coordonnées une échelle adaptée. D'autre part, des couples d'observations identiques (donc de même modalité) correspondent graphiquement à des points confondus. Pour distinguer graphiquement ces points superposés (dont leur nombre est égal à l'effectif de la modalité correspondante), on peut utiliser un **nuage de disques** au lieu d'un nuage de points, c'est-à-dire la représentation graphique de p disques de centre (x_i, y_i) et de rayon proportionnel à l'effectif n_i. Dans ce cas, il faut choisir convenablement le coefficient de proportionnalité utilisé pour représenter des disques ni trop grands ni trop petits et obtenir un graphique lisible.

Définition 16

*Le **point moyen** est le point du plan de coordonnées $(\overline{x}, \overline{y})$.*

3.2 Corrélation et ajustement affine

Une question centrale dans l'étude des séries statistiques bivariées est de mesurer la corrélation entre deux caractères, c'est-à-dire la façon dont ils s'influencent (ou non) entre eux. Pour cela, on peut utiliser la covariance et le coefficient de corrélation affine.

Définition 17

*La **covariance** de x et y, notée $s_{x,y}$, est définie comme la moyenne des produits des écarts à la moyenne, c'est-à-dire :*

$$\boxed{s_{x,y} = \frac{1}{n}\sum_{i=1}^{n}\Big(x_i - \overline{x}\Big)\Big(y_i - \overline{y}\Big).}$$

Remarque. La covariance s'exprime dans le produit des unités de x et de y.

Conseil. Si on utilise les modalités de (x, y), alors la covariance peut s'écrire :

$$\boxed{s_{x,y} = \frac{1}{n}\sum_{i=1}^{p} n_i\Big(x_i - \overline{x}\Big)\Big(y_i - \overline{y}\Big) = \sum_{i=1}^{p} \mathrm{f}_i\Big(x_i - \overline{x}\Big)\Big(y_i - \overline{y}\Big).}$$

Exercice 3. Écrire en Python une fonction `covariance` qui prend en arguments deux listes de réels (qu'on suppose de même taille) et qui renvoie la covariance de leurs éléments. On pourra utiliser la fonction `moyenne` écrite dans l'exercice 1.

Attention. Comme pour la variance et l'écart-type, il existe une définition différente de la covariance en statistiques inférentielles (le n au dénominateur est remplacé par $n-1$). Cette covariance est souvent notée $\sigma_{x,y}$ mais cela dépend des marques et des modèles de calculatrices, logiciels ou applications utilisés pour la calculer.

Remarque. Si x et y sont **positivement corrélés**, c'est-à-dire que les modalités de x sont grandes (donc supérieures à $\overline{x}$) lorsque les modalités de y sont grandes (donc supérieures à $\overline{y}$) et réciproquement que les modalités de x sont petites (donc inférieures à $\overline{x}$) lorsque les modalités de y sont petites (donc inférieures à $\overline{y}$), alors les termes de la somme dans la définition de la covariance sont tous positifs (comme produits de deux réels de même signe) et par conséquent la valeur de $s_{x,y}$ est positive. Graphiquement, le nuage de point semble suivre une direction croissante lorsque lorsque x et y sont positivement corrélés. Inversement, si x et y sont **négativement corrélées**, c'est-à-dire que les modalités de x sont grandes lorsque les modalités de y sont petites et réciproquement, alors la valeur de $s_{x,y}$ est négative (comme somme de produits de deux réels de signes opposés) et le nuage de points semble suivre une direction décroissante.

Théorème 2 (*Formule de König-Huygens*)

La covariance est égale à la différence entre la moyenne des produits et le produit de la moyenne, c'est-à-dire :

$$\boxed{s_{x,y} = \overline{(xy)} - (\overline{x})(\overline{y})}.$$

Démonstration. On a :

$$\begin{aligned}
s_{x,y} &= \frac{1}{n}\sum_{i=1}^{n}\Big(x_i - \overline{x}\Big)\Big(y_i - \overline{y}\Big) \quad \text{par définition de la covariance} \\
&= \frac{1}{n}\sum_{i=1}^{n}\Big(x_i y_i - \overline{y}x_i - \overline{x}y_i + \overline{xy}\Big) \\
&= \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \overline{y} \times \underbrace{\frac{1}{n}\sum_{i=1}^{n} x_i}_{=\overline{x}} - \overline{x} \times \underbrace{\frac{1}{n}\sum_{i=1}^{n} y_i}_{=\overline{y}} + \overline{xy} \times \underbrace{\frac{1}{n}\sum_{i=1}^{n} 1}_{=1} \quad \text{par linéarité de la somme} \\
&= \overline{xy} - \overline{y} \times \overline{x} - \overline{x} \times \overline{y} + \overline{x} \times \overline{y} = \overline{(xy)} - (\overline{x})(\overline{y}).
\end{aligned}$$

□

Propriété 7

La covariance vérifie les propriétés suivantes :

(i) $s_{x,x} = s_x^2$,

(ii) $s_{x,y} = s_{y,x}$ *(on dit que la covariance est* ***symétrique****)*,

(iii) $s_{cx,y} = cs_{x,y} = s_{x,cy}$ *pour toute constante* $c \in \mathbb{R}$,

(iv) $\boxed{s_{x+y}^2 = s_x^2 + 2s_{x,y} + s_y^2}$ *et donc* $s_{x,y} = \frac{1}{2}\left(s_{x+y}^2 - s_x^2 - s_y^2\right)$,

(v) si toutes les observations de x *ou de* y *sont égales entre elles alors* $s_{x,y} = 0$.

Démonstration. *(i)* Par définitions de la covariance et de la variance. *(ii)* Par commutativité du produit des nombres réels. *(iii)* Par linéarité de la moyenne, on a d'après la formule de König-Huygens :

$$s_{cx,y} = \overline{(cx)y} - \underbrace{(\overline{cx})}_{=c(\overline{x})}(\overline{y}) = \underbrace{\overline{c(xy)}}_{=c(\overline{xy})} - c(\overline{x})(\overline{y}) = c\Big(\overline{xy} - (\overline{x})(\overline{y})\Big) = cs_{x,y}.$$

Et de même pour $s_{x,cy} = cs_{x,y}$. *(iv)* On a :

$$\begin{aligned}
s_{x+y}^2 &= \overline{(x+y)^2} - \Big(\overline{x+y}\Big)^2 \quad \text{d'après la formule de König-Huygens} \\
&= \underbrace{\overline{x^2 + 2xy + y^2}}_{=\overline{x^2}+2\overline{xy}+\overline{y^2}} - \Big(\overline{x} + \overline{y}\Big)^2 \quad \text{car la moyenne est linéaire} \\
&= \overline{x^2} + 2\overline{xy} + \overline{y^2} - \Big(\overline{x}^2 + 2(\overline{x})(\overline{y}) + \overline{y}^2\Big) \\
&= \Big(\overline{x^2} - \overline{x}^2\Big) + 2\Big(\overline{xy} - (\overline{x})(\overline{y})\Big) + \Big(\overline{y^2} - \overline{y}^2\Big) \\
&= s_x^2 + 2s_{x,y} + s_y^2 \quad \text{d'après les formules de König-Huygens.}
\end{aligned}$$

(v) Supposons par exemple que toutes les observations de y sont égales à une même constante, donc que $s_y^2 = 0$, et notons c cette constante. On sait que $s_{x+y}^2 = s_{x+c}^2 = s_x^2$ car la variance est invariante par translation. De plus, on d'après le point précédent :

$$\underbrace{s_{x+y}^2}_{=s_x^2} = s_x^2 + 2s_{x,y} + \underbrace{s_y^2}_{=0} \quad \text{donc} \quad s_x^2 = s_x^2 + 2s_{x,y}.$$

Après simplifications, on en déduit bien que $s_{x,y} = 0$.

□

Remarque. Le dernier point est cohérent avec l'interprétation de la covariance pour la corrélation. Par exemple, si les modalités de y sont peu dispersées (donc si s_y est proche de 0) alors les variations des valeurs des modalités de x ont peu d'influence sur les valeurs des modalités de y qui varient peu. Il y a donc peu de corrélation entre les deux caractères (et $s_{x,y}$ est proche de 0).

Théorème 3 (*Inégalité de Cauchy-Schwarz*)

La covariance est inférieure en valeur absolue aux produits des écarts-types, c'est-à-dire :

$$\boxed{\Big|s_{x,y}\Big| \leqslant s_x s_y}.$$

De plus, il y a égalité si et seulement si x *et* y *sont en relation affine.*

Démonstration. Si $s_x = 0$ ou $s_y = 0$ alors l'inégalité de Cauchy-Schwarz est évidente car dans ce cas $s_{x,y} = 0$ (puisque toutes les observations de x ou de y sont égales entre elles). Il y a même égalité. De plus, x et y sont bien en relation affine dans ce cas (par exemple $y = 0x + \overline{y}$ si $s_y = 0$).
On peut donc supposer que $s_x \neq 0$ et $s_y \neq 0$, et on considère la fonction suivante :

$$P : t \mapsto \underbrace{s_{tx+y}^2}_{\geqslant 0} = \underbrace{s_{tx}^2}_{=t^2 s_x^2} + 2\underbrace{s_{tx,y}}_{=ts_{x,y}} + s_y^2 = \underbrace{s_x^2}_{=a}\, t^2 + \underbrace{2s_{x,y}}_{b}\, t + \underbrace{s_y^2}_{=c}.$$

On reconnaît une fonction polynomiale de degré 2 qui est toujours positive. Le polynôme ne peut pas avoir deux racines (sinon il serait négatif entre les racines, ce qui est absurde). On en déduit que son discriminant est nul ou strictement négatif. Or :

$$\Delta = b^2 - 4ac = \Big(2s_{x,y}\Big)^2 - 4s_x^2 s_y^2 = 4\left(s_{x,y}^2 - s_x^2 s_y^2\right) \leqslant 0 \quad \text{donc} \quad s_{x,y}^2 \leqslant s_x^2 s_y^2.$$

En passant à la fonction $t \mapsto \sqrt{t}$ qui est strictement croissante, on obtient bien l'inégalité de Cauchy-Schwarz :

$$\Big|s_{x,y}\Big| \leqslant s_x s_y \quad \text{car} \quad s_x \geqslant 0 \quad \text{et} \quad s_y \geqslant 0.$$

De plus, il y a égalité si et seulement si $\Delta = 0$. Dans ce cas, il existe une racine $t_0 \in \mathbb{R}$ de P, c'est-à-dire $P(t_0) = s_{t_0x+y}^2 = 0$. On en déduit que toutes les observations du caractère $t_0x + y$ sont égales à une même constante. Si on note c cette constante, on a $t_0x + y = c$ donc $y = -t_0x + c$. Par conséquent, x et y sont bien en relation affine.

□

Remarque. On retrouve un résultat similaire à celui de l'inégalité de Cauchy-Schwarz en géométrie (voir le chapitre n° 14) avec une démonstration analogue (toujours aussi originale et «jolie»).

Définition 18

Si ni les observations de x ni celles de y ne sont toutes égales entre elles (donc si $s_x \neq 0$ et $s_y \neq 0$), on définit le ***coefficient de corrélation affine****, noté $r_{x,y}$, par :*

$$\boxed{r_{x,y} = \frac{s_{x,y}}{s_x s_y}}.$$

Remarque. Le coefficient de corrélation affine n'a pas d'unités. C'est une version «normalisée» de la covariance.

Théorème 4

Le coefficient de corrélation affine est un réel de $[-1, +1]$. De plus, il est égal à -1 ou à $+1$ si et seulement si les deux caractères sont en relation affine.

Démonstration. D'après l'inégalité de Cauchy-Schwarz.

□

Remarque. Puisque l'écart-type est positif, le signe du coefficient de corrélation affine s'interprète de la même manière que celui de la covariance pour la corrélation des deux caractères. D'autre part, plus $|r_{x,y}|$ est proche de 1 et plus x et y sont en corrélation affine, c'est-à-dire proche d'être en relation affine. Graphiquement, le nuage de points est concentré aux environs d'une droite strictement croissante lorsque $r_{x,y}$ est proche de $+1$ et aux environs d'une droite strictement décroissante lorsque $r_{x,y}$ est proche de -1. Dans le cas extrême où $|r_{x,y}| = 1$, tous les points du nuage sont alignés sur une droite dont on peut déterminer l'équation en reprenant la démonstration du cas d'égalité de l'inégalité de Cauchy-Schwarz : dans ce cas, on a vu que $\Delta = 0$ et donc que le polynôme P admet pour unique racine :

$$t_0 = \frac{-b}{2a} = \frac{-2s_{x,y}}{2s_x^2} = \frac{-s_{x,y}}{s_x^2}.$$

De plus, $s_{t_0x+y}^2 = P(t_0) = 0$ donc toutes les observations du caractère $t_0x + y$ sont égales à une même constante qui est nécessairement égale à $c = \overline{t_0x + y} = t_0\overline{x} + \overline{y}$. Par conséquent, la droite admet pour équation :

$$y = -t_0x + c = \frac{s_{x,y}}{s_x^2}x + \frac{-s_{x,y}}{s_x^2}\overline{x} + \overline{y} = \frac{s_{x,y}}{s_x^2}\Big(x - \overline{x}\Big) + \overline{y}.$$

Définition 19

*La **droite de régression affine** de la série statistique bivariée des caractères quantitatifs x et y est la droite du plan euclien qui admet pour équation :*

$$\boxed{y = \frac{s_{x,y}}{s_x^2}\Big(x - \overline{x}\Big) + \overline{y}}.$$

Conseil. Une analyse dimensionnelle à l'aide des unités de x et y permet de vérifier l'homogénéité de l'équation de la droite de régression affine.

Attention. Les coefficients $\overline{x}$, $\overline{y}$, s_x^2 et $s_{x,y}$ sont des constantes qu'on calcule à l'aide des observations. Mais les variables x et y dans l'équation de la droite de régression affine sont muettes : elles correspondent aux coordonnées d'un point quelconque de la droite et ne doivent pas être confondues avec les caractères x et y même si la notation est ambigüe. Pour éviter les confusions, on peut utiliser des majuscules pour les coordonnées :

$$Y = \frac{s_{x,y}}{s_x^2}\Big(X - \overline{x}\Big) + \overline{y}.$$

Propriété 8

La droite de régression affine passe par le point moyen du nuage de points.

Démonstration. En effet, si $X = \overline{x}$ on obtient bien que $Y = \overline{y}$.

□

Remarque. La droite de régression affine est toujours définie, même si les points du nuage ne sont pas alignés. Lorsque $|r_{x,y}|$ est proche de 1, le nuage de points est concentré aux environs de la droite de régression affine (qui est strictement croissante si $s_{x,y} > 0$ et strictement décroissante si $s_{x,y} < 0$). Lorsque $r_{x,y}$ est proche de 0, le nuage de points est dispersé autour de la droite de régression affine.

Lorsque la corrélation affine est importante (c'est-à-dire que $|r_{x,y}|$ est proche de 1 et que le nuage de points est concentré aux environs d'une droite), il existe de nombreuses façons de trouver une droite qui soit «proche» du nuage de points (il faudrait avant tout définir précisément ce que signifie être «proche»). Réaliser un **ajustement affine** consiste à trouver l'équation d'une telle droite. La droite de régression affine est un exemple d'ajustement affine (l'exemple le plus simple). Elle est aussi appelée la **droite d'ajustement affine selon la méthode des moindres carrés** car elle minimise la somme des carrés des distances verticales. Plus précisément, si $y = ax + b$ est l'équation d'une droite quelconque du plan (où $(a, b) \in \mathbb{R}^2$), alors la distance verticale d'un point (x_i, y_i) du nuage à la droite est égale à $|y_i - (ax_i + b)|$ et donc la somme des carrés des distances verticales est égale à :

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

Par conséquent, trouver la droite qui minimise cette quantité revient à chercher les valeurs de a et b pour lesquelles la fonction F (de deux variables) atteint son minimum. Nous verrons dans le chapitre n° 28 sur les fonctions à deux variables que (par analogie avec les fonctions à une variable) les deux dérivées partielles de F s'annulent en ces valeurs. Or :

$$\begin{aligned}
\frac{\partial F}{\partial a}(a, b) &= \sum_{i=1}^{n} -2x_i\Big(y_i - ax_i - b\Big) \\
&= -2\sum_{i=1}^{n} x_i y_i + 2a\sum_{i=1}^{n} x_i^2 + 2b\sum_{i=1}^{n} x_i \quad \text{par linéarité de la somme} \\
&= -2n\Big(\overline{xy}\Big) + 2an\Big(\overline{x^2}\Big) + 2bn\Big(\overline{x}\Big) \quad \text{d'après la définition de la moyenne}
\end{aligned}$$

$$\text{et} \quad \frac{\partial F}{\partial b}(a,b) = \sum_{i=1}^{n} -2\Big(y_i - ax_i - b\Big)$$
$$= -2\sum_{i=1}^{n} y_i + 2a\sum_{i=1}^{n} x_i + 2b\sum_{i=1}^{n} 1 \quad \text{par linéarité de la somme}$$
$$= -2n\Big(\overline{y}\Big) + 2an\Big(\overline{x}\Big) + 2bn \quad \text{d'après la définition de la moyenne.}$$

Par conséquent :

$$\begin{cases} \dfrac{\partial F}{\partial a}(a,b) = 0 \\ \dfrac{\partial F}{\partial b}(a,b) = 0 \end{cases} \iff \begin{cases} -2n\left(\overline{xy}\right) + 2an\left(\overline{x^2}\right) + 2bn\left(\overline{x}\right) = 0 \\ -2n\left(\overline{y}\right) + 2an\left(\overline{x}\right) + 2bn = 0 \end{cases}$$
$$\iff \begin{cases} \left(\overline{x^2}\right)a + \left(\overline{x}\right)b = \overline{xy} \\ \left(\overline{x}\right)a + b = \overline{y} \end{cases}$$
$$\iff \begin{pmatrix} \overline{x^2} & \overline{x} \\ \overline{x} & 1 \end{pmatrix}\begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} \overline{xy} \\ \overline{y} \end{pmatrix}.$$

On reconnaît une système linéaire de deux équations à deux inconnues dont la matrice des coefficients a un déterminant égal à :

$$\det\begin{pmatrix} \overline{x^2} & \overline{x} \\ \overline{x} & 1 \end{pmatrix} = \overline{x^2} - \overline{x}^2 = s_x^2 \quad \text{d'après la formule de König-Huygens.}$$

On peut supposer que $s_x^2 \neq 0$ sinon toutes les observations de x seraient égales entre elles et l'étude de la corrélation serait inintéressante (tous les points du nuage seraient alignés sur une droite verticale). Par conséquent, le système linéaire admet une unique solution égale à :

$$\begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} \overline{x^2} & \overline{x} \\ \overline{x} & 1 \end{pmatrix}^{-1}\begin{pmatrix} \overline{xy} \\ \overline{y} \end{pmatrix} = \frac{1}{s_x^2}\begin{pmatrix} 1 & -\overline{x} \\ -\overline{x} & \overline{x^2} \end{pmatrix}\begin{pmatrix} \overline{xy} \\ \overline{y} \end{pmatrix} = \frac{1}{s_x^2}\begin{pmatrix} \overline{xy} - \left(\overline{x}\right)\left(\overline{y}\right) \\ -\left(\overline{xy}\right)\left(\overline{x}\right) + \left(\overline{x^2}\right)\left(\overline{y}\right) \end{pmatrix}.$$

On peut simplifier ces résultats à l'aide de la formule de König-Huygens :

$$a = \frac{s_{x,y}}{s_x^2} \quad \text{et} \quad b = \frac{-\Big(s_{x,y} + \left(\overline{x}\right)\left(\overline{y}\right)\Big)\left(\overline{x}\right) + \Big(s_x^2 + \left(\overline{x}\right)^2\Big)\left(\overline{y}\right)}{s_x^2} = \frac{-s_{x,y}}{s_x^2}\overline{x} + \overline{y}.$$

Finalement, on en déduit que la droite d'équation $y = ax + b$ qui minimise la somme des carrés des distances verticales est celle obtenue en remplaçant a et b par les valeurs trouvées, c'est-à-dire :

$$y = \frac{s_{x,y}}{s_x^2}x - \frac{s_{x,y}}{s_x^2}\overline{x} + \overline{y} = \frac{s_{x,y}}{s_x^2}\Big(x - \overline{x}\Big) + \overline{y}.$$

On retrouve l'équation de la droite de régression affine qui correspond donc aussi à la droite d'ajustement affine selon la méthode des moindres carrés.