STATISTIQUES

"Je ne crois jamais une statistique à moins de l'avoir moi-même falsifiée."
Winston Churchill

1 INTRODUCTION

Confronté à un grand nombre de données, le statisticien cherche à clarifier leur répartition pour en dégager la part d'aléa et permettre des prises de décision.

Exemple

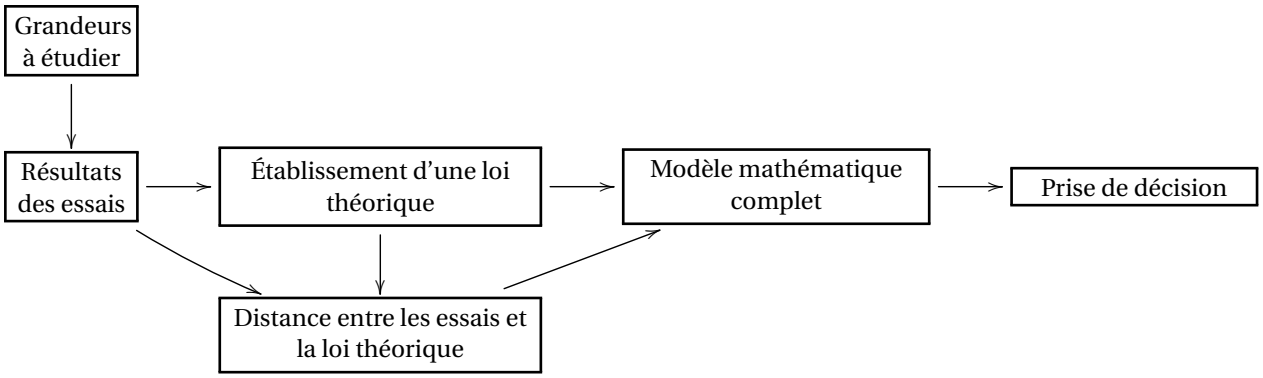
- Répartition des tailles des individus dans un groupe → *Remplir les rayonnages de magasins d'habillement*
- Pièces conformes et défectueuses en sortie d'usine → *Adapter le processus de production*
- Aliments préférés d'un groupe d'individus → *Commercialisation de nouvelles tablettes de chocolats*

Le statisticien sera en particulier tenté de chercher s'il n'existe pas une loi sous-jacente qui expliquerait les résultats des différentes expériences, ou plus simplement une règle mathématique qui permettrait de décrire et de quantifier au mieux les résultats pour faciliter leur exploitation.

Par exemple, lorsque vous construisez un bâtiment, vous utilisez pour cela des matériaux qui ont été normalisés. Des essais statistiques ont été réalisés et ont permis d'élaborer un modèle mathématique qui détermine leur comportement.
En effet, il est préférable avant de construire votre plancher, de savoir si vos poutres ne vont pas rompre sous le poids ou tout simplement trop se déformer. Ainsi, tous les essais que vous avez réalisés auparavant sur les matériaux vous donnent le comportement "*normal*" de vos poutres.
Par contre, malgré tous nos efforts, deux poutres ne se comporteront jamais de façon exactement identique.

Lorsque l'on a établi une loi "théorique", les essais ne donneront jamais un résultat parfaitement conforme, mais il y aura une certaine distance entre les essais réalisés et la loi théorique. La mesure de cette distance permettra en particulier de prévoir des marges de sécurité. Si les essais montrent que les comportements sont tous très proches les uns des autres, nous pourrons avoir des marges d'erreur réduites. En revanche, si les essais présentent une grande *dispersion*, il est préférable d'augmenter les marges d'erreur pour éviter un accident parce que l'on serait tombé sur une poutre un peu moins bonne que les autres.

Dans ce cours, nous resterons sur des cas extrêmement simples et basiques par rapport à cet objectif très ambitieux. Mais gardez en tête la logique complète présentée dans cette introduction car cela vous aidera à mieux comprendre la finalité des notions que nous révisons ici.



2 STATISTIQUES À UNE VARIABLE

A Classer les données

Définition 2.1 :
Une **population** (statistique) est un ensemble fini d'éléments sur lesquels portent une étude statistique.
Les éléments de la population sont appelés **individus**.
Le nombre total d'individus est appelé **effectif total** et souvent noté N.

Une partie de la population est appelée **échantillon** de la population.

Exemple

- les élèves de la classe forment une population.
- 5 élèves pris au hasard (ou non) dans la classe en forment un échantillon.
- des objets manufacturés en sortie d'usine forment une population. Chaque objet représente alors un individu.

Définition 2.2 (Caractère)
Un **caractère** est une donnée que l'on peut observer sur les individus de la population.
La caractère est dit **quantitatif** lorsqu'il est numérique (taille, poids, durée...).
Sinon, il est dit **qualitatif** (couleur des yeux...).

Pour un caractère noté x, on notera en général $x_1, x_2, \cdots, x_n$ les valeurs qu'il peut prendre.

Définition 2.3 (Série statistique)
Une **série statistique** est la donnée d'une population et d'un caractère.

L'objet d'une étude statistique est d'évaluer un caractère sur l'ensemble d'une population sans avoir à tester tous les individus. Pour cela, on choisit un échantillon **représentatif** de cette population sur lequel vont porter les tests.

Les résultats de ces tests seront ensuite extrapolés à la population tout entière.

Par exemple, pour des élections, les instituts de sondage ne sondent pas la totalité de la population (les électeurs), mais uniquement un échantillon choisi pour être le plus représentatif possible. Cela suppose que l'échantillon soit suffisamment grand, mais aussi qu'il recoupe au mieux les différentes situations possibles au regard de leur proportion dans la population (différences d'âges, positions géographiques...).

Cependant, les sondages ne remplacent jamais le vote et le comptage complet des votes exprimés qui seuls donnent un résultat fiable (?) et exact.

Exemple

- Dans un sondage politique, le caractère (nom de la personne ou du parti pour lequel on vote) n'est pas numérique. C'est un caractère qualitatif.
- Si vous testez la teneur en chocolat des tablettes en sortie de production, ce sera un caractère quantitatif.

Dans une **série statistique**, le statisticien cherche à regrouper les données qui se ressemblent pour obtenir une vision plus globale. On dit qu'il crée des **classes**.

Définition 2.4 (Classes)
Pour une série statistique dont le caractère peut prendre un ensemble de valeurs possibles,
les **classes** sont des parties de cet ensemble qui en forment une partition.

Explications :

Une classe est un regroupement de caractères suivant un critère prédéfini.
Pour pouvoir faire une étude statistique, les classes doivent former une **partition** de l'ensemble des caractères présents. C'est-à-dire :

- toutes les caractères doivent appartenir à une classe
- un caractère ne peut appartenir qu'à une seule classe à la fois.

Pour un caractère donné, il existe de nombreuses façons de créer des classes différentes.

Exemple

Pour répartir les tailles d'individus adultes (au cm près), on peut créer les classes :

$$\Big\{ < 150\,;\, [150\,;\, 159]\,;[160\,;\, 169]\,;\ldots;\, [190\,;\, 199]\,;\geq 200\Big\}$$

Les classes ainsi définies sont *disjointes*[1] et *recouvrent*[2] toutes les tailles possibles. Elles forment donc bien une partition de mes données.
On peut considérer les classes comme de nouveaux caractères. Ici, nous avions un caractère quantitatif (tailles en cm) et avons créé des classes qui donnent à présent un caractère qualitatif (des intervalles). Pour obtenir à nouveau un caractère quantitatif à partir de ces classes, nous pourrions remplacer l'intervalle par son milieu (ce qui donne donc un nombre). Cela demande de modifier un peu le traitement des extrémités "< 150" et "≥ 200" pour pouvoir leur assigner une valeur représentative.

Exemple

Pour étudier les couleurs préférées des français pour leur voiture, je peux créer les classes :

$$\Big\{\text{Rouge ; Vert ; Bleu ; Blanc ; Noir ; Gris ; Autre couleur}\Big\}$$

Ce sont des caractères qualitatifs.

Définition 2.5 (Effectif)
L'**effectif** d'une classe, correspond au nombre de fois que l'on retrouve un élément de cette classe dans la série de données.
On peut ainsi créer un tableau classe/effectif.

Exemple

Je tire à pile ou face 10 fois de suite. J'écris P lorsque j'obtiens Pile et F lorsque j'obtiens Face. Voici mes résultats :

$$PFFPPFPPPPF$$

J'ai obtenu 7 fois pile et 4 fois face. J'ai donc le tableau des effectifs suivant :

Classe	Pile	Face
Effectif	7	4

Définition 2.6 :
Calculer la fréquence c'est établir la proportion des effectifs d'une classe par rapport à l'effectif total.
La fréquence peut s'exprimer sous forme de nombre décimal, de fraction ou de pourcentage.
La fréquence d'une classe s'obtient par la formule :

$$\text{fréquence} = \frac{\text{effectif de la classe}}{\text{effectif total}}$$

[1] Aucune mesure n'appartient à deux classes en même temps
[2] Chaque taille se trouve dans une classe

Propriété 2.7 :

La somme des fréquences de toutes les classes doit toujours être égale à 1.

Preuve :

Si on note n_i l'effectif de la classe i, alors f_i sa fréquence est donnée par

$$f_i = \frac{n_i}{\sum_j n_j}$$

Donc

$$\sum_i f_i = \sum_i \frac{n_i}{\sum_j n_j} = \frac{\sum_i n_i}{\sum_j n_j} = 1$$

■

Méthode :

Pour éviter les erreurs dans le calcul des effectifs ou des fréquences, rajoutez toujours une colonne "Total" à la fin du tableau. Cela permet de vérifier que vous n'avez oublié aucune donnée (ou compté plusieurs fois la même).

Exemple

Classe	Pile	Face	Total
Effectif	7	4	11
Fréquence (fraction)	$\frac{7}{11}$	$\frac{4}{11}$	$\frac{11}{11} = 1$
Fréquence (décimal)	$\approx 0,64$	$\approx 0,36$	$\approx 0,63 + 0,36 = 1$
Fréquence (pourcentage)	$\approx 63\%$	$\approx 36\%$	$\approx 63\%+36\% = 100\%$

Définition 2.8 (Effectifs cumulés croissants)

L'**effectif cumulé croissant** de la classe x_k est égal à

$$\sum_{i \leq k} n_i = n_1 + n_2 + \cdots + n_k$$

L'effectif cumulé croissant de la dernière classe est alors égal à l'effectif total.

On définit de la même manière, les effectifs cumulés décroissants, ou les fréquences cumulées croissantes et décroissantes.

Exemple

Voici les statistiques des notes obtenues lors d'un DS :

Classe	$[0,2]$	$]2,4]$	$]4,6]$	$]6,8]$	$]8,10]$	$]10,12]$	$]12,14]$	$]14,16]$	$]16,18]$	$]18,20]$	Total
Effectif	2	2	8	13	5	6	6	1	0	2	45
Effectif cumulé	2	4	12	25	30	36	42	43	43	45	–
Fréquence ($\approx$ %)	4,4	4,4	17,8	28,9	11,1	13,3	13,3	2,2	0	4,4	100
Fréquence cumulée ($\approx$ %)	4,4	8,9	26,7	55,6	66,7	80,0	93,3	95,6	95,6	100	–

On remarque que pour calculer les fréquences cumulées croissantes, il est préférable d'utiliser le calcul de l'effectif cumulé (exact) plutôt que d'utiliser le calcul de fréquences (approché). En effet dans ce dernier cas, on ajoute les erreurs successives de chaque calcul de fréquence : c'est à éviter dans la mesure du possible.

B Représenter les données

Pour faciliter la visualisation des statistiques, les données peuvent être représentées graphiquement sous forme de diagrammes. Les diagrammes usuels sont[3] :

- le diagramme en barres ou histogramme
- le diagramme en bande (effectifs cumulés)
- le diagramme circulaire
- le nuage de points

Exemple

Dans les exemples suivants, nous présentons les parts de marché des navigateurs web en France fin 2013[4]

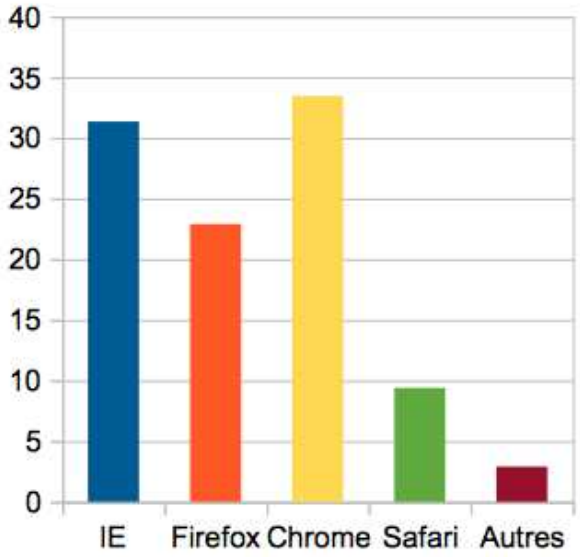
Navigateur	IE Microsoft	Mozilla Firefox	Chrome Google	Safari Apple	Autres	Total
Fréquence	31,35%	22,9%	33,46%	9,39%	2,9%	100%

Diagramme en barres : On construit un diagramme en barres en utilisant des hauteurs[5] de barres proportionnelles à l'effectif (ou à la fréquence).

Exemple

Navigateur	IE Microsoft	Mozilla Firefox	Chrome Google	Safari Apple	Autres	Total
Fréquence	31,35%	22,9%	33,46%	9,39%	2,9%	100%

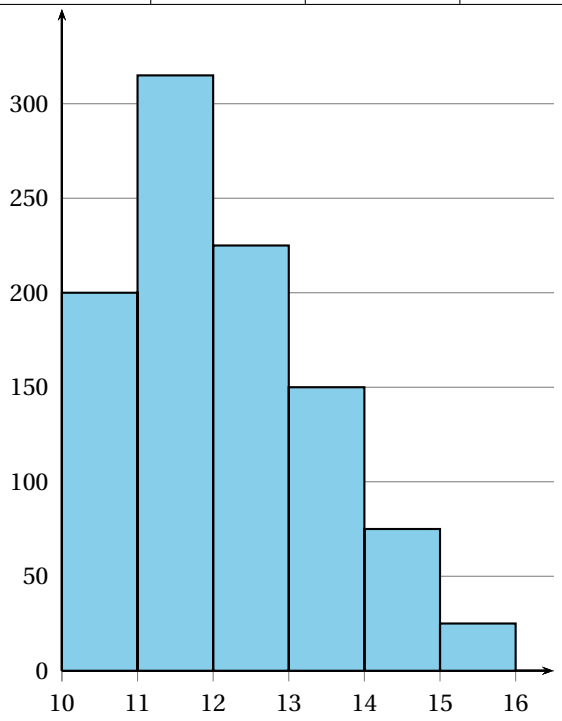
[3] Mais il en existe bien d'autres : diagrammes en boîtes, radars...
[4] France, novembre 2013, source Statcounter
[5] Il s'agit plutôt de l'aire de la barre qui doit être proportionnelle à l'effectif comme nous allons le voir plus loin.



Lorsqu'il s'agit de données numériques *continues* regroupées en classes (par exemple les tailles des individus en cm, la durée d'un trajet...), on parle alors d'**histogramme**.

Exemple

Horaire	$10 \leq t < 11$	$11 \leq t < 12$	$12 \leq t < 13$	$13 \leq t < 14$	$14 \leq t < 15$	$15 \leq t < 16$	Total
Nombre d'entrées	200	315	225	150	75	25	990



Dans l'exemple précédent, toutes les classes avaient la même *amplitude*. La hauteur des barres était donc proportionnelles à l'effectif (ou à la fréquence). Lorsque les classes n'ont pas la même amplitude on préférera autant que possible avoir des **surfaces** des barres proportionnelles à l'effectif. [6].

[6] Cela revient à prendre la moyenne pondérée par l'amplitude.

Exemple

Horaire	$10 \leq t < 11$	$11 \leq t < 12$	$12 \leq t < 13$	$13 \leq t < 14$	16	Total
Nombre d'entrées	200	315	225	150	100	990

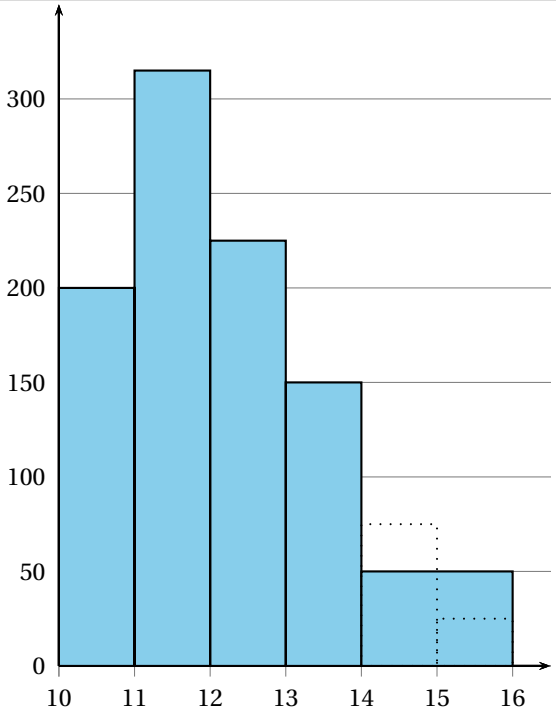


Diagramme en bandes : On construit un diagramme en bandes en collant bout à bout les barres du diagramme en barres.

Exemple

Navigateur	IE Microsoft	Mozilla Firefox	Chrome Google	Safari Apple	Autres	Total
Fréquence	31,35%	22,9%	33,46%	9,39%	2,9%	100%

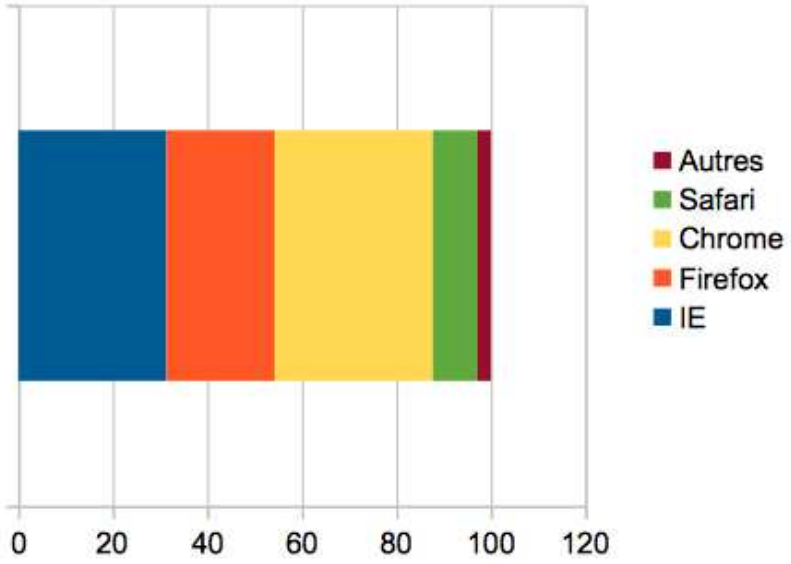
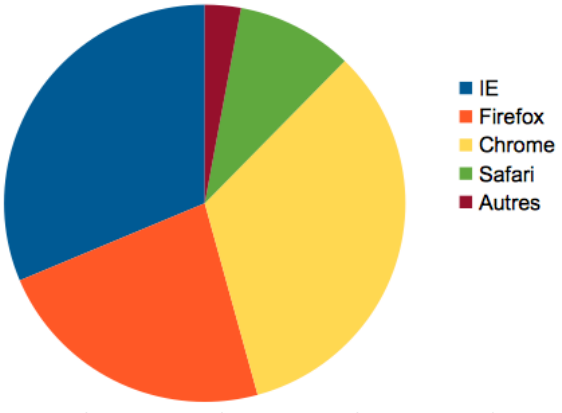


Diagramme circulaire : On construit un diagramme circulaire en découpant le cercle suivant la proportion de chaque classe.

Si je travaille en degrés, pour connaître le secteur angulaire associé à chaque classe, je multiplie la fréquence correspondante par 360 (pour que la totalité fasse 360^o, c'est à dire un cercle complet).
Lorsque la fréquence est en pourcentage, je multiplie le pourcentage par 360 puis divise par 100. C'est simplement une règle de trois.

Exemple

Navigateur	IE Microsoft	Mozilla Firefox	Chrome Google	Safari Apple	Autres	Total
Fréquence	31,35%	22,9%	33,46%	9,39%	2,9%	100%
Secteur angulaire	113^o	82^o	121^o	34^o	10^o	360^o

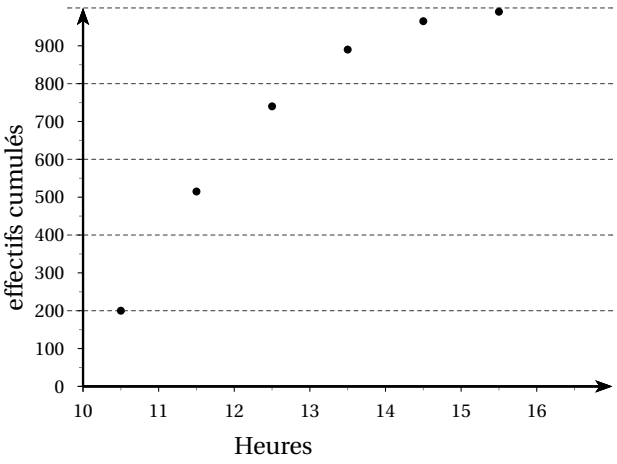


Remarque : Un diagramme circulaire est souvent appelé *diagramme camembert*[7].

Nuage de points : On peut aussi placer les données sur un repère (et éventuellement les relier entre elles). Ce type de diagramme sera surtout utile lors des statistiques bivariées (à deux variables). Ici, nous l'utilisons comme un histogramme en remplaçant les barres par un point à leur sommet.

Exemple

Reprenons l'exemple avec le nombre d'entrées en fonction de l'heure.
En abscisse, nous mettons les heures (nous prendrons le centre de la classe).
En ordonnée, nous plaçons les effectifs cumulés.



C Indicateurs de position

Les indicateurs de position ne sont définis que pour des données statistiques *numériques*. Si les données sont des couleurs, par exemple, ces indicateurs n'ont pas de sens.
Dans la suite, on suppose une série statistique de caractère x prenant un nombre fini de valeurs $x_1, x_2, \cdots, x_n$.

Chaque valeur x_i a pour effectif n_i dans la série statistique. L'effectif total est noté $N = \sum_{i=1}^{n} n_i$.

Mode :

Définition 2.9 (Mode et classe modale)
Le **mode** est la valeur du caractère avec le plus grand effectif.
La **classe modale** est la classe avec le plus grand ratio effectif/amplitude.

Exemple

Le mode permet par exemple de trouver les heures de pointes dans certains magasins.
Dans un histogramme, le mode correspond au plus haut bâton. Dans l'exemple précédent avec le nombre entrées selon les heures, la classe modale est la classe de 11h à midi.
Il serait donc très maladroit de fermer à cette heure-ci.

Moyenne :

[7] S'agit-il d'un hommage au sapeur du même nom connu pour son ingéniosité mathématique ?

Définition 2.10 (moyenne)

La moyenne est égale à :

$$\bar{x} = \frac{\sum_{i=1}^{n} n_i x_i}{N} = \frac{\text{somme des "valeur} \times \text{effectif"}}{\text{effectif total}}$$

Explications :

Soit une série statistique numérique dont la somme de toutes les valeurs est égale à $\sum_{i=1}^{n} n_i x_i = T$.

Si toutes les données de la série avaient eu la même valeur $\bar{x}$, quelle serait cette valeur pour que le total reste égal à T ? Cette valeur $\bar{x}$ serait exactement la moyenne de la série statistique.

Attention : Même lorsqu'une valeur vaut 0, il ne faut pas oublier de la compter (malheureusement pour vos notes).

Exemple

Dans une classe, on fait une enquête sur le nombre de livres lus par chaque élève au mois de mars. Voici les résultats anonymes :

$$\{0;1;1;15;0;3;0;2;8;0;0;1;3;0;7;0;2;0;1;0;0;0;3\}$$

La moyenne vaut :

$$\bar{x} = \frac{0+1+1+15+0+3+0+2+8+0+0+1+3+0+7+0+2+0+1+0+0+0+3}{25} = \frac{49}{25} = 1,96$$

Pour mieux visualiser les réponses, on peut aussi compter les effectifs en regroupant ceux qui ont lu le même nombre de livres :

nombre de livres lus	0	1	2	3	7	8	15	Total
effectif	12	4	3	3	1	1	1	25
valeur × effectif	12 × 0	4 × 1	3 × 2	3 × 3	1 × 7	1 × 8	1 × 15	49

On trouve alors

$$\bar{x} = \frac{49}{25} = 1,96$$

Remarque : Dans une série statistique, les valeurs ne peuvent pas à la fois être toutes (strictement) plus petites ou toutes plus grandes que la moyenne. Il y en a forcément des plus petites et aussi des plus grandes que la moyenne. Par contre, en général, il n'y a pas le même nombre de valeurs plus petites que de valeurs plus grandes que la moyenne.

Moyenne pondérée : Dans certains cas, on peut vouloir donner plus d'importance à certaines valeurs qu'à d'autres. On va donc utiliser l'idée précédente pour "gonfler" artificiellement les effectifs de certaines valeurs.

Par exemple, dans un trimestre, on peut vouloir qu'un contrôle compte 6 fois plus qu'une interrogation de cours. Dans le calcul de la moyenne, chaque contrôle comptera comme s'il apparaissait 6 fois. C'est ce que l'on appelle le coefficient.

On retrouve alors la définition d'un barycentre :

Définition 2.11 :

La moyenne pondérée est égale à :

$$\bar{x} = \frac{\text{somme des "valeur} \times \text{coefficient"}}{\text{somme des coefficients}}$$

Exemple

	DS	Interro	DS	DS	Khôlles	Total
Note	12	20	9	10	13	-
Coefficient	6	1	6	6	0,5	19,5
Note × Coefficient	12 × 6	20 × 1	9 × 6	10 × 6	13 × 0,5	212,5

La moyenne vaut donc

$$\bar{x} = \frac{212,5}{19,5} \approx 10,90$$

Exemple **(Moyenne pondérée et centre de gravité)**

Graduation de la règle : abscisse (cm)	6	8	10	11	30	Total
Masse (g)	8	10	14	10	3	45
Masse×abscisse	48	80	140	110	90	468

La moyenne pondérée est donc

$$M = \frac{468}{45} = 10,4\text{cm}$$

C'est le centre de gravité associé au système. Nous voyons bien dans cet exemple qu'il n'y a pas la même masse de chaque côté du point d'équilibre. En fait, plus une masse est loin, plus elle agit fortement. Le terme "masse× abscisse" correspond au *moment* en mécanique[8].
Vous pouvez alors méditer cette phrase d'Archimède : "Donnez-moi un point d'appui et je soulèverais le monde". Reste à trouver ce point d'appui, mais aussi la barre assez rigide !

Les propriétés du barycentre se retrouvent donc avec la moyenne :

Propriété 2.12 **(Linéarité de la moyenne)**
La moyenne est linéaire.
Pour deux caractères numériques x, y et un réel λ,

$$\overline{x+\lambda y} = \overline{x} + \lambda \overline{y}$$

Preuve :

C'est une conséquence de la linéarité de la somme.

$$\begin{aligned}\overline{x+\lambda y} &= \frac{\sum_{i=1}^{n} x_i + \lambda y_i}{N} \\ &= \sum_{i=1}^{n} \frac{x_i}{N} + \lambda \sum_{i=1}^{n} \frac{y_i}{N} \\ &= \overline{x} + \lambda \overline{y}\end{aligned}$$

■

[8] Il faut aussi multiplier par la force de gravité pour obtenir une force, la formule du moment est donc "poids × distance" où la distance est celle entre le point mû et la droite qui passe par le point d'application de la force et dont un vecteur directeur est la force elle-même.

Médiane : Nous avons vu que la moyenne donne beaucoup d'importance aux valeurs extrêmes et ne renseigne pas sur le nombre de valeurs plus grandes ou plus petites.
Il existe un autre indicateur qui se contente de se placer au milieu, sans s'occuper de savoir si une valeur "pèse lourd" ou pas. C'est la médiane.

Définition 2.13 :
La **médiane** Me d'une série statistique est la valeur telle qu'il y a autant d'éléments plus grands que d'éléments plus petits.
Soit une série qui comprend n valeurs **ordonnées**,

- si n est impair, alors Me est la valeur "du milieu",
- si n est pair, alors Me est la demi-somme de la valeur juste au dessus et juste au dessous du milieu.

Remarque : Parfois la définition mathématique ne comprend pas la deuxième partie de cet énoncé. Dans ce cas, une série peut avoir plusieurs médianes.

L'inconvénient de la médiane est qu'il n'existe pas d'expression littérale simple pour la calculer, contrairement à la moyenne.

Exemple

Reprenons l'exemple précédent utilisé pour la moyenne.

nombre de livres lus	0	1	2	3	7	8	15	Total
effectif	12	4	3	3	1	1	1	25
effectif cumulé	12	16	19	22	23	24	15	-

Il y a au total 25 données, c'est un nombre impair. Je laisse donc la moitié au dessus : 12 valeurs et la moitié en dessous : 12 valeurs.
La médiane est alors la valeur "du milieu" : la $13^{\text{ème}}$. Il y a 12 zéro, la treizième est un "1".

$$Me = 1$$

Cette valeur est à comparer à la moyenne : $\bar{x} = 1,96$

Exemple

Des élèves doivent se placer pour la photo de classe et on mesure donc leur taille (hors chapeau pour le déguisement).

$$\Big\{172;162;190;190;169;164;177;181;189;161;164;182;185;188;169;190;193;$$
$$189;179;180;173;193;166;164;163;164;190;176;176;192;173;194\Big\}$$

Lorsque l'on ordonne les données :

$$\Big\{161;162;163;164;164;164;164;166;169;169;172;173;173;176;176;177;179;180;$$
$$181;182;185;188;189;189;190;190;190;190;192;193;193;194\Big\}$$

Il y a 32 données au total, la médiane est donc entre la $16^{\text{ème}}$ et la $17^{\text{ème}}$ donnée.
On prend la moyenne des deux : $Me = \dfrac{177+179}{2} = 178$

Exemple

La médiane est plus difficile à obtenir que la moyenne (puisqu'il faut déjà classer les données). Mais elle ne doit pas être négligée pour autant.
Pour illustration, voici un extrait d'un article de 2014 sur le site d'Europe 1.

Salaire moyen d'un salarié : 2.128 euros. "En 2011, dans le secteur privé et les entreprises publiques (...), le salaire moyen net de tous prélèvements sociaux s'est élevé à 2 128 euros", a indiqué lundi la Dares. Le tout pour les employés travaillant à temps plein.
Ces chiffres varient évidemment en fonction du statut : le salaire net moyen est plus bas pour les employés (1.649 euros) et pour les ouvriers (1.680 euros), mais plus élevé pour les professions intermédiaires (2.309 euros) et les cadres (4.302 euros).

Salaire médian d'un salarié : 1.712 euros.

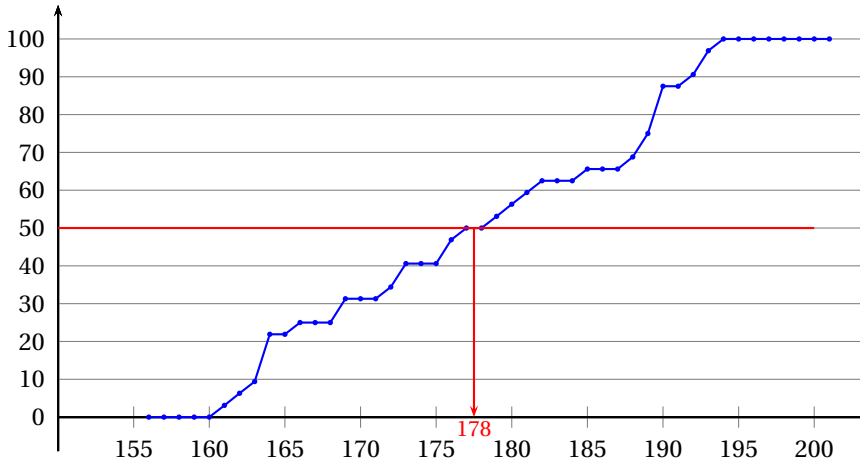
Le salaire moyen a le mérite de la simplicité mais il n'est pas assez précis. Un seul exemple pour tout comprendre : si vous faites la moyenne des rémunérations dans un groupe comprenant 999 smicards et un millionnaire, vous obtenez un chiffre qui n'est pas représentatif.
Les économistes préfèrent donc parler de salaire médian, c'est-à-dire le "salaire tel que la moitié des salariés de la population considérée gagne moins et l'autre moitié gagne plus", selon l'Insee. Le salaire médian est en moyenne 20% inférieur au salaire moyen et s'établissait en 2011 à 1.712 euros net par mois. En clair la moitié des Français gagnait plus, l'autre moitié moins.

Exemple

99,9% de la population française a plus de jambes que la moyenne !
mais en a autant que la médiane.

Exemple (Lecture de la médiane sur les fréquences cumulées)

On peut représenter les tailles des élèves sur un graphique avec les fréquences cumulées croissantes (ou les effectifs cumulés croissants). La médiane se lit très simplement :

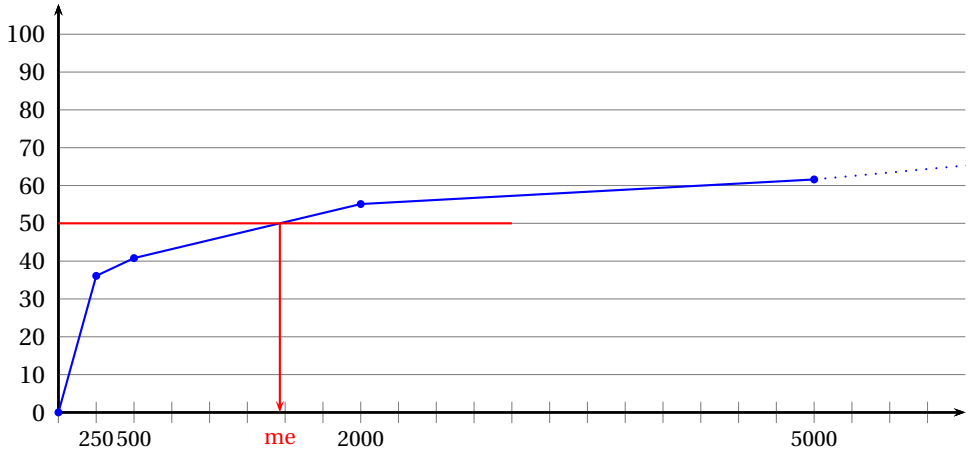


Exemple (Approximation de la médiane par interpolation linéaire)

Le rapport Uniagro 2015 indique la taille des entreprises selon l'âge des ingénieurs issus des écoles UNIAGRO. Voici le tableau pour les ingénieurs de moins de 30 ans :

Taille de l'entreprise	%
Moins de 250 salariés	36,1
250 à 499 salariés	4,7
500 à 1999 salariés	14,3
2000 à 4999 salariés	6,5
5000 salariés et plus	38,4
Total	100

On peut tracer le graphique des fréquences cumulées croissantes pour avoir une *approximation* de la médiane.



Le graphique permet de savoir que la taille médiane de l'entreprise est comprise entre entre 500 et 2000 salariés. Par contre, le tableau ne permet pas d'avoir de valeur plus précise. On peut donc supposer en première approximation que la répartition est uniforme entre 500 et 2000.
Les trois points $A(500,40,8)$, $M(\text{me},50)$ et $B(2000,55,1)$ sont alors alignés, et les vecteurs $\overrightarrow{AM}$, $\overrightarrow{BM}$ sont colinéaires. Donc

$$\frac{\text{me}-500}{50-40,8}=\frac{2000-500}{55,1-40,8}\quad\Rightarrow\quad \text{me}=500+9,2\frac{1500}{14,3}\approx 1465$$

La médiane approximative (que l'on pouvait avoir en lecture graphique) est donc de 1465 employés dans l'entreprise. En première approximation, la moitié des moins de 30 ans, travaillent dans une entreprise de moins de 1465 employés.
Cependant nous voyons ici qu'une approximation linéaire n'est pas bien adaptée à la forme de la courbe et on pourrait par exemple essayer d'utiliser plutôt une approximation logarithmique par exemple. Mais nous verrons cela plus tard avec les statistiques à deux variables.

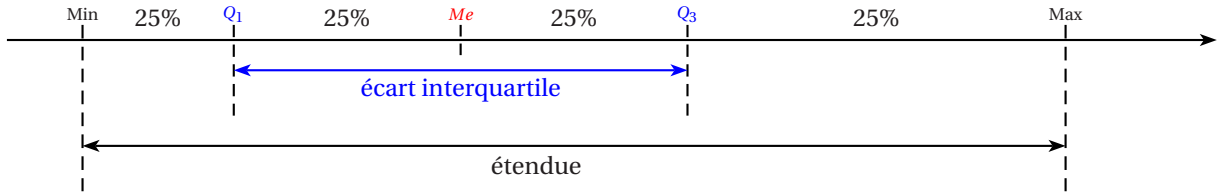
D Indicateurs de dispersion

Les indicateurs de dispersion indiquent si les valeurs sont très resserrées ou au contraire s'étendent largement. À nouveau, ces indicateurs n'existent que pour des caractères quantitatifs.

Déciles et quartiles

Définition 2.14 :

L'**étendue** d'une série statistique est la différence entre la plus grande valeur (max) et la plus petite valeur (min) de la série.
Le **premier quartile** Q_1 est la plus petite valeur de la série telle que *au moins* 25% des valeurs lui soient inférieures ou égales.
Le **troisième quartile** Q_3 est la plus petite valeur de la série telle que *au moins* 75% des valeurs lui soient inférieures ou égales.
L'**écart interquartile** est la différence $Q_3 - Q_1$



Le $k^{\text{ième}}$ **décile** est la plus petite valeur telle que *au moins* $k \times 10\%$ des valeurs lui soient inférieures ou égales.

Variance et écart-type

Définition 2.15 (Variance)

La variance d'une série statistique par rapport au caractère x est donnée par

$$\mathbb{V} = \sum_{i=1}^{n} \frac{n_i (x_i - \bar{x})^2}{N} = \sum_{i=1}^{n} f_i (x_i - \bar{x})^2$$

où f_i désigne la fréquence de la valeur x_i.

Explications :

Le principe de la variance est de mesurer une dispersion, c'est à dire un étalement des valeurs.
Elle mesure donc un étalement moyen autour de la moyenne. L'écart moyen entre x et $\overline{x}$ est toujours nul par linéarité de la moyenne, on étudie donc ici "la moyenne du carré de l'écart à la moyenne".

Attention : La variance n'est pas *homogène* avec x : lorsque x désigne une longueur, V désigne une longueur au carré. Il s'agit d'un rapport quadratique.

Exemple

Avec l'exemple des livres lus :

nombre de livres lus	0	1	2	3	7	8	15	Total
effectif	12	4	3	3	1	1	1	25

La moyenne est $\overline{x} \approx 1,96$.
La variance est égale à

$$\mathbb{V} \approx \frac{12\,(0-1,96)^2 + 4\,(1-1,96)^2 + 3\,(2-1,96)^2 + 3\,(3-1,96)^2 + 1\,(7-1,96) + 1\,(8-1,96)^2 + 1\,(15-1,96)^2}{25} = 11,40$$

Théorème 2.16 (Formule de König-Huygens)

$$\mathbb{V} = \overline{x^2} - \overline{x}^2$$

Preuve :

$$\begin{aligned}
\mathbb{V} &= \sum_{i=1}^{n} f_i (x_i - \overline{x})^2 \\
&= \sum_{i=1}^{n} f_i \left(x_i^2 - 2x\overline{x} + \overline{x}^2\right) \\
&= \sum_{i=1}^{n} f_i x_i^2 - 2\sum_{i=1}^{n} f_i x\overline{x} + \sum_{i=1}^{n} f_i \overline{x}^2 \\
&= \overline{x^2} - 2\overline{x}\sum_{i=1}^{n} f_i x + \overline{x}^2 \sum_{i=1}^{n} f_i \\
&= \overline{x^2} - 2\overline{x}\,\overline{x} + \overline{x}^2 && \text{car } \sum_i f_i = 1 \\
&= \overline{x^2} - \overline{x}^2
\end{aligned}$$

■

Exemple

On peut reprendre l'exemple précédent :

nombre de livres lus (x_i)	0	1	2	3	7	8	15	Total	
effectif (n_i)	12	4	3	3	1	1	1	25	
$n_i \times x_i$	12×0 0	4×1 4	3×2 6	3×3 9	1×7 7	1×8 8	1×15 15	49	$\overline{x} = \frac{49}{25} \approx 1{,}96$
$n_i \times x_i^2$	12×0^2 0	4×1^2 4	3×2^2 12	3×3^2 27	1×7^2 49	1×8^2 64	1×15^2 225	381	$\overline{x^2} = \frac{381}{25} \approx 15{,}24$

Donc $\mathbb{V} = \overline{x^2} - \overline{x}^2 \approx 15{,}24 - 1{,}96^2 = 11{,}40$

Propriété 2.17 :

Pour un caractère x et deux constantes réelles quelconques $a, b \in \mathbb{R}^2$,

$$\mathbb{V}(ax+b) = a^2\mathbb{V}(x)$$

Explications :

Il est intuitif que $\mathbb{V}$ soit insensible à l'ajout d'une constante : la translation des valeurs ne modifie pas leur dispersion.
Par contre, l'homothétie de rapport a influence la dispersion, c'est comme un changement d'échelle. Comme la variance est quadratique par rapport au caractère, elle est multipliée par a^2. L'écart type que l'on définit un peu plus loin permet de résoudre cet écueil.

Preuve :

$$\begin{aligned}
\mathbb{V}(ax+b) &= \overline{(ax+b)^2} - \overline{(ax+b)}^2 \\
&= \overline{a^2x^2 + 2abx + b^2} - \left(a\overline{x} + b\right)^2 && \text{par linéarité de la moyenne} \\
&= a^2\overline{x^2} + 2ab\overline{x} + b^2 - \left(a^2\overline{x}^2 + 2ab\overline{x} + b^2\right) && \text{par linéarité de la moyenne} \\
&= a^2\left(\overline{x^2} - \overline{x}^2\right) \\
&= a^2\mathbb{V}(x)
\end{aligned}$$

■

Propriété 2.18 (Positivité de la variance)

Pour tout caractère x, $\mathbb{V}(x) \geq 0$.
La variance est nulle si et seulement si x ne prend qu'une seule valeur.

Preuve :

La définition de la variance l'exprime comme une somme de termes positifs (des effectifs fois des carrés).
C'est donc un nombre positif.
Il est nul si et seulement si tous ses termes sont nuls, c'est-à-dire si x ne prend que la valeur de sa moyenne. ■

On a bien $\mathbb{V}(x) \geq 0$ même si x prend des valeurs négatives.

Ceci permet de définir sa racine carrée.

Définition 2.19 (Écart-type)

L'écart-type d'une série statistique pour le caractère x est la racine carrée de sa variance.

$$s_x = \sigma_x = \sqrt{\mathbb{V}}$$

C'est la raison pour laquelle on note souvent la variance $\mathbb{V} = \sigma_x^2 = s_x^2$
L'intérêt de l'écart type est qu'il est *homogène* avec le caractère x : il est dans la même unité que x.

Propriété 2.20 :

Pour toutes constantes $a, b \in \mathbb{R}^2$,

$$\sigma(ax+b) = |a|\sigma(x)$$

Attention : N'oubliez pas la valeur absolue !

Exemple

Dans l'exemple des livres lus, $\sigma \approx \sqrt{11,40} \approx 3,38$.

3 STATISTIQUES À DEUX VARIABLES

Les statistiques bivariées (ou plus généralement multivariées) permettent d'établir des liens entre différents caractères sur une population. Nous n'étudierons que les caractères quantitatifs.

Attention : Les statistiques permettent de constater une *corrélation* entre plusieurs caractères, mais elles n'ont pas de légitimité pour discerner les causes de cette corrélation. Est-ce qu'un caractère est cause de l'autre, est-ce qu'ils sont tous deux conséquences d'une cause tierce ?

Exemple

On a observé une corrélation entre le nombre de cigognes en Alsace et le taux de natalité ces dernières années (le taux de natalité et la population de cigogne ont chuté suivant des trajectoires comparables). De là, à affirmer un lien de cause à effet entre l'arrivée des cigognes et la naissances des enfants, ce n'est pas à la statistique de se prononcer.

A Indicateurs

Définition 3.1 (Modalité et effectifs)

Soient deux caractères x, y sur une population, tels que
x puisse prendre les valeurs $(x_1, x_2, \cdots, x_p)$
y puisse prendre les valeurs $(y_1, y_2, \cdots, y_q)$.

a) Un couple (x_i, y_j) s'appelle **modalité** (conjointe) des caractères x, y.

b) L'**effectif conjoint** de la modalité (x_i, y_j) est le nombre d'individus vérifiant simultanément les caractères $x = x_i$ et $y = y_j$.

c) L'**effectif marginal** de x_i est le nombre d'individus vérifiant le caractère $x = x_i$ (indépendamment de la valeur du caractère y).

Explications :

Les effectifs conjoints peuvent être décrits par un tableau à deux entrées appelé **tableau de contingence**.
Chaque case du tableau correspond à l'effectif conjoint.
Les effectifs marginaux s'obtiennent par sommation sur les lignes ou les colonnes.

$x \backslash y$	y_1	y_2	$\cdots$	y_q	
x_1	$n_{1,1}$	$n_{1,2}$	$\cdots$	$n_{1,q}$	$n_{1,\bullet} = \sum_{j=0}^{q} n_{1,j}$ ← modalité marginale de x_1
x_2	$n_{2,1}$	$n_{2,2}$	$\cdots$	$n_{2,q}$	$n_{2,\bullet} = \sum_{j=0}^{q} n_{2,j}$ ← modalité marginale de x_2
$\vdots$				$\vdots$	
x_i	$n_{i,1}$	$n_{i,2}$	$\cdots$	$n_{i,q}$	$n_{i,\bullet} = \sum_{j=0}^{q} n_{i,j}$ ← modalité marginale de x_i
$\vdots$				$\vdots$	
x_p	$n_{p,1}$	$n_{p,2}$	$\cdots$	$n_{p,q}$	$n_{n,\bullet} = \sum_{j=0}^{q} n_{p,j}$ ← modalité marginale de x_p
	$n_{\bullet,1} = \sum_{i=1}^{p} n_{i,1}$ mod. marg. de y_1	$n_{\bullet,2} = \sum_{i=1}^{p} n_{i,2}$ mod. marg. de y_2		$n_{\bullet,q} = \sum_{i=1}^{p} n_{i,q}$ mod. marg. de y_q	

Définition 3.2 (Fréquences conjointes et marginales)

La fréquence conjointe de la modalité (x_i, y_j) est donnée par

$$f_{i,j} = \frac{\text{effectif conjoint de la modalité } (x_i, y_j)}{\text{effectif total}}$$

La fréquence marginale de la modalité (x_i) est donnée par

$$f_{i,\bullet} = \frac{\text{effectif marginal de la modalité } x_i}{\text{effectif total}}$$

Propriété 3.3 :

La modalité marginale s'obtient comme somme des fréquences marginales :

$$f_{i,\bullet} = \sum_j f_{i,j}$$

B Représentation graphique

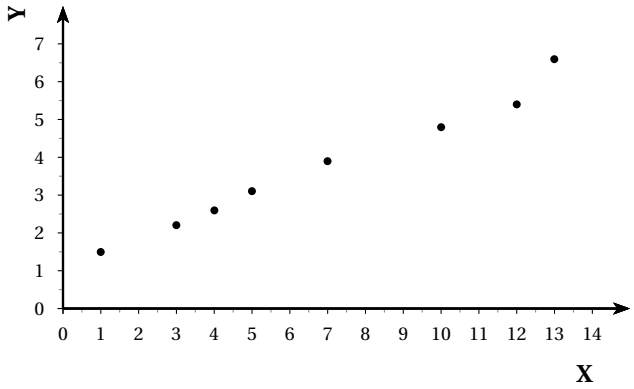
Les fréquences conjointes (ou effectifs conjoints) des caractères (x, y) peuvent être représentés par des nuages de points. La modalité (x_i, y_j) est représentée par un disque[9]<+++> dont le centre a les coordonnées (x_i, y_j) et dont la surface est proportionnelle à la fréquence.
Souvent, la notion de fréquence ou même d'effectif n'est pas pertinente, on représente alors les données par des points comme dans l'exemple qui suit.

Exemple

Voici un tableau représentant les valeurs prises par deux caractères.

X	1	3	4	5	7	10	12	13
Y	1,5	2,2	2,6	3,1	3,9	4,8	5,4	6,6

On peut représenter ces données par un nuage de points :



On observe que les points sont presque alignés. L'objet de la section suivante sera de trouver droite qui décrit au mieux cet alignement.

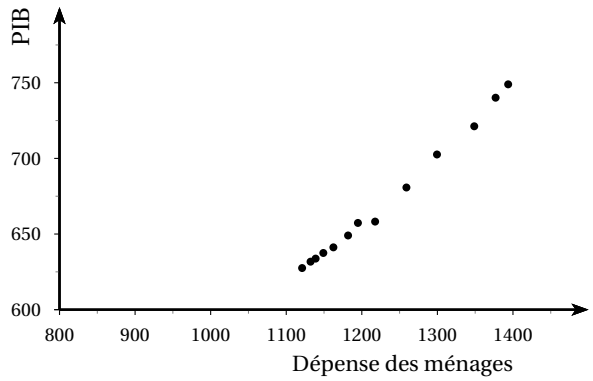
Exemple

Voici un tableau indiquant le PIB et les dépenses de consommation des ménages en France de 1990 à 2002 en euros constants (base 1995).

Année	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002
Produit intérieur brut	1 121,0	1 132,2	1 149,1	1 138,9	1 162,4	1 181,8	1 194,9	1 217,6	1 259,1	1 299,5	1 348,8	1 377,1	1 393,7
Dépenses de consommation finale des ménages	627,5	631,7	637,5	633,7	641,2	649,0	657,3	658,2	680,7	702,6	721,2	740,1	748,9

Milliards d'euros 1995, Source: INSEE

[9] On peut remplacer le disque par toute autre forme plus ou moins exotique.



Bien que moins flagrant que pour le dernier exemple, on observe ici un relatif alignement entre les points.

C Régression linéaire

Définition 3.4 (Point moyen)

Le **point moyen** d'une série statistique pour les caractères (x, y) est le point de coordonnées

$$G(\overline{x}, \overline{y})$$

où $\overline{x}$ et $\overline{y}$ désignent respectivement les moyennes marginales de x et de y.

Explications :

Dans le plan, le point G est le barycentre des points de coordonnées (x_i, y_j) affectés de leur fréquence (pour la masse du point).

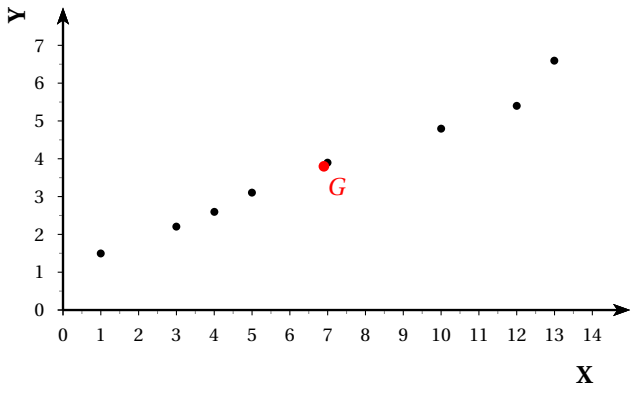
Exemple

Avec l'exemple précédent,

$$\overline{x} = \frac{1+3+4+5+7+10+12+13}{8} \approx 6,9$$

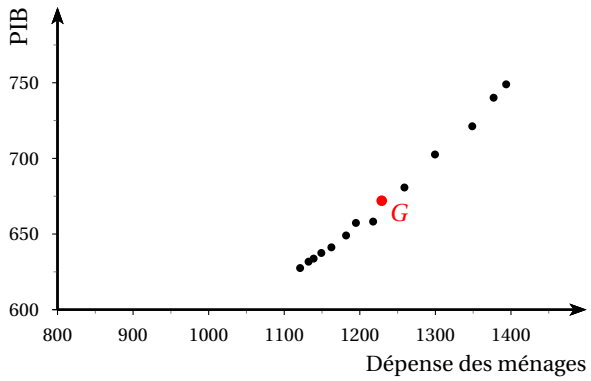
$$\overline{y} = \frac{1.5+2.2+2.6+3.1+3.9+4.8+5.4+6.6}{8} \approx 3,8$$

Le point moyen est donc $G(6,9;3,8)$



Exemple

Avec l'exemple PIB/Dépense des ménages, le point moyen est $G(1229, 672)$.



Définition 3.5 (Covariance)

La covariance de deux caractères (x, y) d'une série statistique est définie par

$$s_{x,y} = \frac{\sum\limits_{\substack{1\leq i\leq p\\1\leq j\leq q}} n_{i,j}\left(x_i-\overline{x}\right)\left(y_j-\overline{y}\right)}{N}$$

Avec $\overline{x}$ et $\overline{y}$ les moyennes de x et y avec leurs effectifs marginaux,
$n_{i,j}$ l'effectif conjoint de la modalité (x_i, y_j)
$N = \sum\limits_{\substack{1\leq i\leq p\\1\leq j\leq q}} n_{i,j}$ l'effectif total.

Propriété 3.6 (Formule de Kœnig-Huygens)

$$s_{x,y} = \overline{xy} - \overline{x}\cdot\overline{y}$$

Preuve :

$$\begin{aligned}
s_{x,y} &= \frac{\sum\limits_{i=1}^{p}\sum\limits_{j=1}^{q} n_{i,j}\left(x_i-\overline{x}\right)\left(y_j-\overline{y}\right)}{N}\\
&= \frac{\sum\limits_{i=1}^{p}\sum\limits_{j=1}^{q} n_{i,j}\left(x_iy_j-\overline{x}y_j-\overline{y}x_j+\overline{x}\,\overline{y}\right)}{N}\\
&= \frac{\sum\limits_{i=1}^{p}\sum\limits_{j=1}^{q} n_{i,j}x_iy_j}{N} - \overline{x}\,\overline{y}-\overline{y}\,\overline{x}+\overline{x}\,\overline{y}\\
&= \overline{xy}-\overline{x}\,\overline{y}
\end{aligned}$$

■

Propriété 3.7 :

Pour un caractère x, la covariance $s_{x,x}$ est égale à la variance V.

$$s_{x,x} = s_x^2$$

Explications :

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

La covariance permet de mesurer la dispersion du nuage de points autour du point moyen. Sa dimension (en physique) est égal au produit des dimensions des caractères.
Ici, il n'y a pas d'écart-type. La covariance peut être négative.

Définition 3.8 (Coefficient de correlation)
Pour deux caractères (x, y) d'écart-type $s_x \neq 0$ et $s_y \neq 0$, on définit le **coefficient de corrélation** par

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

Le coefficient de corrélation est une grandeur sans dimension.
Remarque : $r_{x,y} = r_{y,x}$

Propriété 3.9 :
Le coefficient de corrélation est toujours compris dans l'intervalle $[-1, 1]$.

$$r_{x,y} = \pm 1 \iff \exists (a,b) \in \mathbb{R}^2, \text{ tel que } y = ax + b$$

Explications :

Lorsque le coefficient de corrélation vaut ± 1, cela veut dire que x et y sont liés de façon affine. Dans le nuage de point, les points sont alignés sur la droite d'équation $y = ax + b$.
Par extension, lorsque le coefficient de corrélation se rapproche de ± 1 ($|r_{x,y}|$ proche de 1), les points ont tendance à s'aligner.
On dit alors que les caractères x et y sont **corrélés**.
A contrario, lorsque le coefficient de corrélation se rapproche de 0, cela indique que les caractères ne sont pas corrélés. Le nuage de point ne prend pas la forme d'une droite.
Le fait que $r_{x,y}$ soit dans $[-1, 1]$ provient de l'inégalité de Cauchy-Schwarz que l'on retrouve dans de nombreux domaines des mathématiques. Nous n'approfondissons pas ce point ici.
L'équivalence est simplement le cas d'égalité pour l'inégalité de Cauchy-Schwarz.
La droite de régression nous permettra d'interpréter différemment cette équivalence.

Attention : La corrélation de deux caractères n'indique pas un lien de causalité entre eux. Ils peuvent dépendre d'un tierce paramètre. L'étude du coefficient de corrélation permet de conjecturer l'existence d'un lien de causalité et d'argumenter sa leur faveur a postériori, mais elle ne permet jamais de le *démontrer*.
Le calcul du coefficient de corrélation ne remplace pas un "examen visuel" du nuage de point (souvent beaucoup plus efficace pour déceler si l'alignement est pertinent).

Exemple

Avec l'exemple précédent en X, Y :

$$\overline{xy} = \frac{1 \cdot 1,5 + 3 \cdot 2,2 + 4 \cdot 2,6 + 5 \cdot 3,1 + 7 \cdot 3,9 + 10 \cdot 4,8 + 12 \cdot 5,4 + 13 \cdot 6,6}{8} \approx 32,5$$

Donc

$$s_{x,y} \approx 32,5 - 6.9 \cdot 3.8 \approx 6,3$$

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

Donc $s_x = \sqrt{s_{x,x}} = \sqrt{\overline{x^2} - \overline{x}^2} \approx 4,11$ $\overline{y^2} = \dfrac{1.5^2 + 2.2^2 + 2.6^2 + 3.1^2 + 3.9^2 + 4.8^2 + 5.4^2 + 6.6^2}{8} \approx 17$

Donc $s_y = \sqrt{s_{y,y}} = \sqrt{\overline{y^2} - \overline{y}^2} \approx 1,61$
On en déduit

$$r_{x,y} = \frac{s_{x,y}}{s_x\ s_y} \approx 0,99$$

Le coefficient de corrélation est très proche de 1, cela rend compte du très bon alignement des points.

Définition 3.10 (Droite de régression)

La droite de régression d'un nuage de points pour les caractères (x, y) est la droite passant par le point moyen $G(\overline{x}, \overline{y})$ et de coefficient directeur $a = \frac{s_{x,y}}{s_x^2}$.

Propriété 3.11 (Équation de la droite de regression linéaire)

La droite de régression a pour équation

$$Y = aX + b$$

avec $a = \frac{s_{x,y}}{s_x^2}$ et $b = \overline{y} - a\overline{x}$

Théorème 3.12 :

La droite de régression est l'unique droite d'équation $y = ax + b$ qui minimise la somme des carrés des distances verticales entre la droite et les points du nuage.
a et b minimisent la grandeur :

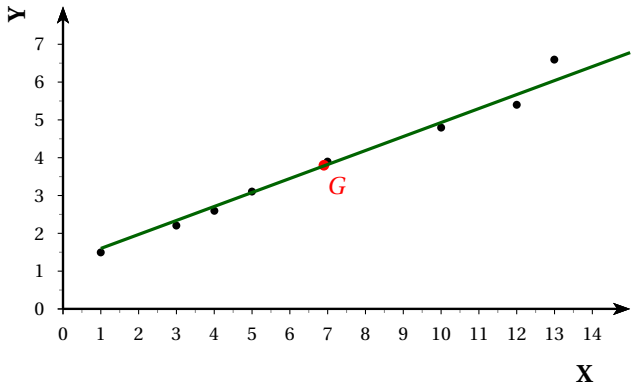
$$\sum_{k=1}^{p+q} \left(y_k - (ax_k + b)\right)^2$$

Preuve :

Admis. ■

Exemple

Avec l'exemple X, Y, on trouve $a = \frac{s_{x,y}}{s_{x,x}} \approx \frac{6,3}{16,9} = 0,37$ et $b = 1,23$



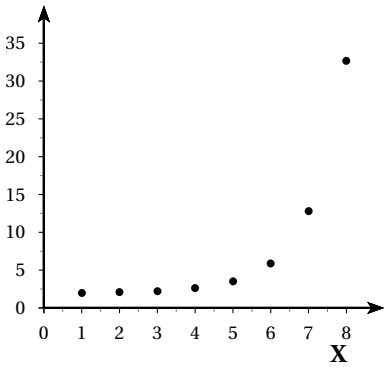
D Se ramener à un ajustement affine

Dans de nombreux cas, on observe un alignement, non suivant une droite, mais d'après une forme puissance, exponentielle ou logarithmique. Il faut alors faire un changement de variable pour se ramener au cas linéaire.

Exemple

X	1.4	4	5	6	9	12	14	15
Y	27	44	54	68	119	214	334	406

Si on cherche un ajustement affine, on trouve $r_{x,y} = 0,95$. La valeur proche de 1 du coefficient de corrélation pourrait faire croire que l'ajustement affine est satisfaisant. Pourtant, une simple observation de la courbe disqualifie directement cette hypothèse : la courbe ressemble à une exponentielle.

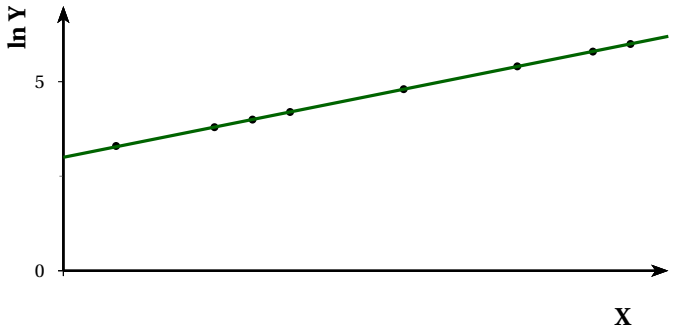


On essaie donc un ajustement du type : $Y = \lambda e^{aX} = e^{ax+b}$, avec $b = \ln \lambda$ (ici $\lambda > 0$ d'après la forme de la courbe).
On cherche une relation

$$\ln Y = ax + b$$

On établit le tableau correspondant :

X	1.4	4	5	6	9	12	14	15
$\ln(Y)$	3.3	3.8	4	4.2	4.8	5.4	5.8	6



On obtient alors $a \approx 0,2$ et $b \approx 3$ avec un coefficient de corrélation $r_{x,y} > 0,999$

Exemple

Il est souvent difficile de discerner à l'œil nu une courbe puissance d'une courbe exponentielle.
Dans le cas, où l'on conjecture une relation de type puissance, on a alors la relation

$$Y = \lambda X^a = e^{a \ln X + b}$$

avec $b = \ln \lambda$ si $\lambda > 0$, c'est à dire que la courbe est croissante.
Alors on peut écrire

$$\ln Y = a \ln X + b$$

On fait alors les tableaux statistiques avec $\ln X$ et $\ln Y$.
Dans le cas précédent, on obtient alors un coefficient de corrélation plus faible $r_{x,y} = 0,94$, mais surtout l'allure du nuage de points n'est clairement pas affine :