

Information chiffrée : statistique bivariée

Cours — 1re ens. sci.

Les filles choisissent-elles les mêmes langues que les garçons ? Les personnes les plus grandes ont-elles vraiment de plus grands pieds ? Le prix du tabac influence-t-il le nombre de fumeurs ? Pour répondre à ce genre de questions, une seule série de données ne suffit pas : il faut observer **deux caractères en même temps** sur les mêmes individus, puis les *croiser*. C'est ce que fait ce chapitre, avec les outils que l'on retrouve partout — dans les études de santé, les rapports sur le climat, les articles économiques ou les débats politiques. L'enjeu dépasse les mathématiques : savoir lire un tableau croisé ou un nuage de points, c'est aussi savoir repérer une conclusion trop rapide, un pourcentage trompeur ou une prévision qui n'a plus aucun sens. Bref, c'est un chapitre d'esprit critique autant que de calcul.

1 Croiser deux caractères qualitatifs

Avant de croiser quoi que ce soit, remettons en place le vocabulaire de base de la statistique, déjà rencontré en seconde. Il est simple, mais il faut l'employer avec précision.

Définition

Une étude statistique porte sur une **population** : l'ensemble des êtres ou des objets étudiés (les élèves d'un lycée, les habitants d'une ville, les logements d'un quartier...). Chaque élément de la population est un **individu**.

Un **caractère** est une propriété que l'on observe sur chaque individu. Il est :

- **qualitatif** lorsque ses valeurs, appelées *modalités*, ne sont pas des nombres : la langue étudiée, le moyen de transport, le genre... ;
- **quantitatif** lorsque ses valeurs sont des nombres sur lesquels les calculs ont un sens : la taille, la pointure, un prix, une température...

En seconde, vous avez déjà croisé deux caractères qualitatifs à l'aide d'un tableau. Reprenons proprement cet outil, car il est au cœur de la première partie du chapitre.

Définition

Lorsqu'on observe **deux caractères qualitatifs** sur les mêmes individus, on regroupe les résultats dans un **tableau croisé d'effectifs** : chaque case indique le nombre d'individus qui présentent *à la fois* la modalité de sa ligne et la modalité de sa colonne.

On complète le tableau par une ligne et une colonne « Total », appelées les **marges** du tableau : elles donnent les effectifs de chaque modalité prise seule, et l'effectif total de la population dans la case en bas à droite.

Exemple

Dans un lycée, on interroge les 200 élèves de première sur leur LV2 (espagnol, allemand ou italien), en distinguant filles et garçons. Voici le tableau croisé obtenu :

	Espagnol	Allemand	Italien	Total
Filles	60	25	15	100
Garçons	50	30	20	100
Total	110	55	35	200

Chaque case intérieure se lit avec un « et » : 60 élèves sont *des filles et* étudient l'espagnol ; 30 élèves sont *des garçons et* étudient l'allemand. Les marges se lisent avec un seul caractère : 110 élèves en tout étudient l'espagnol, et le lycée compte 100 filles pour 100 garçons. La case en bas à droite retrouve bien l'effectif total : $100 + 100 = 110 + 55 + 35 = 200$.

Les effectifs bruts ne permettent pas toujours de comparer deux groupes, surtout quand ils n'ont pas la même taille. On passe alors aux fréquences — et il y en a de deux sortes, qu'il ne faut surtout pas confondre.

Définition

Dans un tableau croisé d'effectifs :

- une **fréquence marginale** est la fréquence d'une modalité dans la population **entière** : on divise l'effectif total de cette modalité (lu dans la marge) par l'effectif total de la population ;
- une **fréquence conditionnelle** est la fréquence d'une modalité **à l'intérieur d'un groupe** : on divise l'effectif d'une case par le total du groupe considéré (le total de sa ligne ou de sa colonne). Elle répond aux questions du type « *parmi les filles*, quelle proportion étudie l'espagnol ? ».

Méthode

Calculer et interpréter une fréquence dans un tableau croisé.

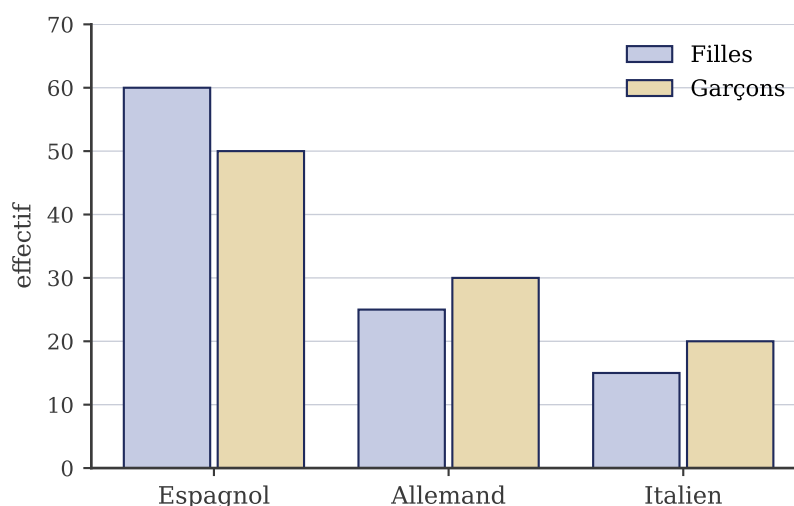
1. **Identifier la population de référence**, c'est-à-dire le groupe dans lequel on se place : la population entière (fréquence marginale) ou un sous-groupe comme « les filles » (fréquence conditionnelle). C'est cet effectif qui va au dénominateur — c'est l'étape décisive.
2. **Diviser** l'effectif concerné par l'effectif de référence, puis convertir en pourcentage si c'est plus parlant.
3. **Écrire une phrase d'interprétation** qui précise la population de référence : « parmi les... », « dans l'ensemble du lycée... ». Un pourcentage sans sa population de référence ne veut rien dire.

Exemple

Reprenons le tableau des LV2.

- *Fréquence marginale* : la fréquence des élèves étudiant l'espagnol est $\frac{110}{200} = 0,55$. Interprétation : dans l'ensemble du lycée, 55 % des élèves étudient l'espagnol.
- *Fréquence conditionnelle* : parmi les filles, la fréquence de l'espagnol est $\frac{60}{100} = 0,60$. Interprétation : 60 % **des filles** étudient l'espagnol.
- *Le piège classique* : « 60 % des filles font espagnol » ne signifie pas que « 60 % des hispanisants sont des filles » ! Cette seconde proportion se calcule dans un autre groupe de référence, celui des 110 élèves qui étudient l'espagnol : $\frac{60}{110} \approx 0,545$, soit environ 54,5 %. Les deux phrases utilisent les mêmes cases du tableau, mais pas le même dénominateur : elles ne racontent pas la même chose.

Pour rendre un tableau croisé lisible d'un coup d'œil, rien ne vaut un graphique. Le plus adapté pour comparer des groupes est le **diagramme en barres groupées** : pour chaque modalité du premier caractère, on place côte à côte une barre par groupe, et la hauteur de chaque barre est proportionnelle à l'effectif.



Le graphique rend la comparaison immédiate : l'espagnol domine largement dans les deux groupes, les filles y sont un peu plus nombreuses que les garçons, et c'est l'inverse pour l'allemand et l'italien. Aucun calcul n'a été nécessaire — c'est toute la force d'une bonne représentation.

Remarque

On peut aussi construire un **diagramme circulaire par groupe** (un « camembert » pour les filles, un autre pour les garçons) : chaque secteur a un angle proportionnel à l'effectif de la modalité dans le groupe. Attention cependant : un diagramme circulaire montre des **proportions à l'intérieur d'un groupe**, pas des effectifs. Si les deux groupes n'ont pas le même effectif total (imaginez 40 filles et 160 garçons), un secteur visuellement plus grand chez les filles peut correspondre à *moins* d'élèves en nombre que chez les garçons. Avant de comparer deux camemberts, toujours vérifier les effectifs totaux de chaque groupe.

2 Croiser deux caractères quantitatifs : le nuage de points

Passons au second cas de figure : les deux caractères observés sont maintenant **quantitatifs**, comme une taille et une pointure. Un tableau croisé n'est plus adapté — les valeurs sont trop nombreuses et

trop dispersées. L'outil naturel est graphique.

Définition

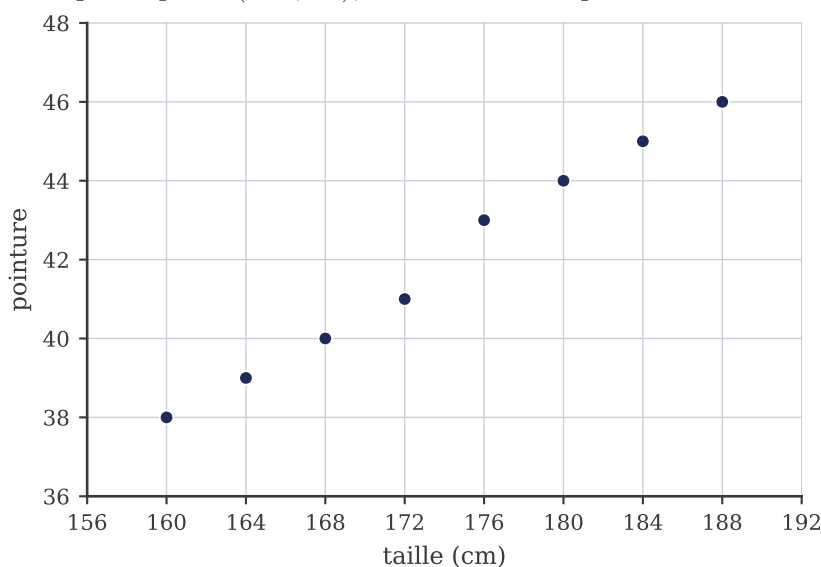
On observe deux caractères quantitatifs sur n individus : le premier individu donne le couple de valeurs $(x_1; y_1)$, le deuxième le couple $(x_2; y_2)$, et ainsi de suite. Le **nuage de points** de la série est l'ensemble des points de coordonnées $(x_i; y_i)$ placés dans un repère : **un point par individu**.

Exemple

On mesure la taille (en cm) et la pointure de 8 élèves d'une classe. Ce sont ces données qui nous serviront de fil rouge jusqu'à la fin du chapitre.

Taille (cm)	160	164	168	172	176	180	184	188
Pointure	38	39	40	41	43	44	45	46

Chaque colonne du tableau devient un point du nuage : l'élève qui mesure 160 cm et chausse du 38 est représenté par le point $(160; 38)$, et ainsi de suite pour les huit élèves.



Que lit-on sur ce nuage ? D'abord une **tendance** : quand la taille augmente, la pointure augmente aussi — le nuage « monte » de gauche à droite, la tendance est croissante. Ensuite une **forme** : les points ne sont pas exactement alignés, mais presque. C'est cette quasi-régularité qui va nous permettre, dans la section suivante, de résumer le nuage par une droite.

Remarque

Un nuage de points peut aussi avoir une tendance **décroissante** (quand une grandeur augmente, l'autre diminue : l'altitude et la température, par exemple), ou ne présenter **aucune tendance** visible (les points forment un paquet sans direction). Dans la vraie vie, des données ne sont jamais parfaitement alignées : on cherche seulement si les points sont *à peu près* alignés. C'est une appréciation graphique, pas un calcul.

Avant d'ajuster une droite, on repère un point particulier qui résume le nuage : son « centre de gravité ».

Définition

Le **point moyen** d'un nuage de n points est le point $G(\bar{x}; \bar{y})$, où $\bar{x}$ est la moyenne des abscisses et $\bar{y}$ la moyenne des ordonnées :

$$\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n} \quad \text{et} \quad \bar{y} = \frac{y_1 + y_2 + \dots + y_n}{n}.$$

Méthode

Calculer les coordonnées du point moyen.

1. Additionner **toutes les abscisses**, puis diviser par le nombre n d'individus : on obtient $\bar{x}$.
2. Additionner **toutes les ordonnées**, puis diviser par n : on obtient $\bar{y}$.
3. Conclure en donnant le point $G(\bar{x}; \bar{y})$, et le placer sur le nuage pour vérifier qu'il tombe bien « au milieu » des points — c'est un excellent contrôle visuel du calcul.

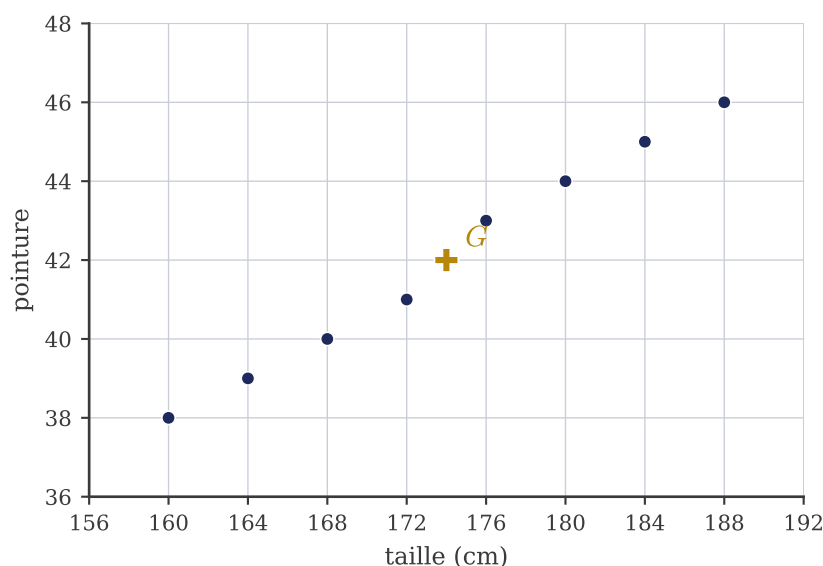
Exemple

Calculons le point moyen du nuage taille–pointure.

$$\bar{x} = \frac{160 + 164 + 168 + 172 + 176 + 180 + 184 + 188}{8} = \frac{1\,392}{8} = 174.$$

$$\bar{y} = \frac{38 + 39 + 40 + 41 + 43 + 44 + 45 + 46}{8} = \frac{336}{8} = 42.$$

Le point moyen du nuage est donc $G(174; 42)$. Interprétation : l'« élève moyen » de ce groupe mesure 174 cm et chausse du 42 — même si, en réalité, aucun des huit élèves n'a exactement ces mesures.



3 L'ajustement affine

Notre nuage taille–pointure a des points presque alignés. L'idée clé du chapitre : remplacer ce nuage par **une droite** qui le résume. On perd un peu de précision, mais on gagne un modèle simple qui permet de faire des prévisions.

Définition

Lorsque les points d'un nuage sont à peu près alignés, on peut modéliser le lien entre les deux caractères par une droite d'équation $y = ax + b$ qui passe « au plus près » des points. Cette droite est appelée une **droite d'ajustement** du nuage, et on dit qu'on réalise un **ajustement affine**.

Son équation permet d'**estimer** la valeur d'un caractère à partir de l'autre : c'est tout l'intérêt de la démarche.

Il existe plusieurs façons d'obtenir une droite d'ajustement. Aucune n'est « la vraie » : ce sont des méthodes différentes qui donnent des droites voisines.

Méthode

Ajuster au jugé.

On trace **à la règle**, à main levée, une droite qui passe au plus près de l'ensemble des points : à peu près autant de points au-dessus qu'en dessous, sans privilégier une extrémité du nuage. Bonne pratique : faire passer la droite par le **point moyen** G , qui joue le rôle de centre d'équilibre du nuage. On lit ensuite graphiquement son coefficient directeur et son ordonnée à l'origine pour obtenir une équation approchée.

Méthode

Construire la droite de Mayer.

1. **Ranger** les points du nuage par abscisses croissantes.
2. **Partager** le nuage en deux moitiés : les points « de gauche » d'un côté, les points « de droite » de l'autre (si l'effectif est impair, on met un point de plus dans l'une des deux moitiés).
3. **Calculer les points moyens partiels** : G_1 , point moyen de la première moitié, et G_2 , point moyen de la seconde, exactement comme à la section précédente.
4. La **droite** (G_1G_2) est la droite de Mayer. Pour obtenir son équation $y = ax + b$: le coefficient directeur vaut $a = \frac{y_{G_2} - y_{G_1}}{x_{G_2} - x_{G_1}}$, puis on trouve b en écrivant que la droite passe par G_1 (on remplace x et y par les coordonnées de G_1 et on résout).

Sur notre fil rouge : la première moitié du nuage (les quatre élèves les plus petits) a pour point moyen $G_1(166; 39,5)$ et la seconde moitié $G_2(182; 44,5)$. Le coefficient directeur de la droite de Mayer vaut donc $\frac{44,5 - 39,5}{182 - 166} = \frac{5}{16} \approx 0,31$.

Propriété

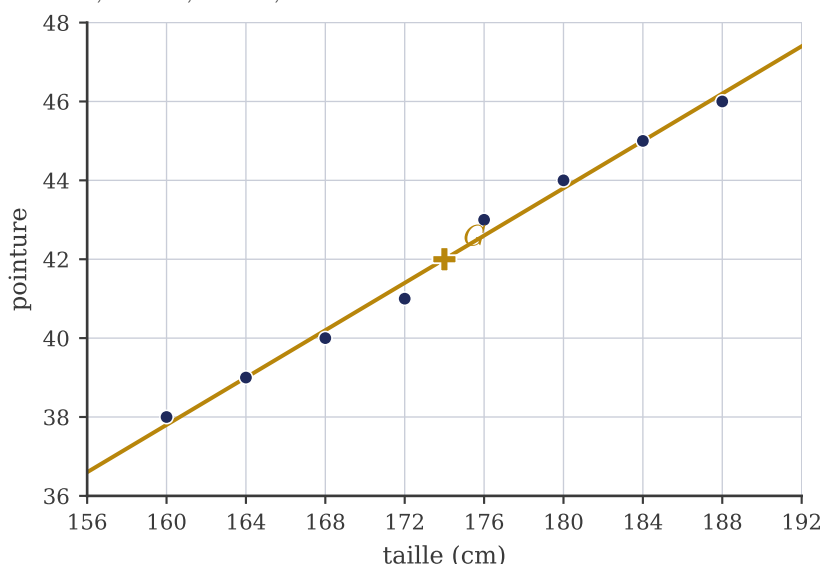
Lorsque le nuage est partagé en deux moitiés de même effectif, le point moyen G du nuage est le milieu du segment $[G_1G_2]$: la droite de Mayer **passer par le point moyen** G . C'est une bonne vérification après calcul — sur le fil rouge, le milieu de $[G_1G_2]$ a bien pour coordonnées $\left(\frac{166 + 182}{2}; \frac{39,5 + 44,5}{2}\right) = (174; 42)$.

Remarque

Une troisième méthode, dite des **moindres carrés**, est celle qu'utilisent le tableur et la calculatrice : la machine calcule automatiquement la droite « la plus proche » de l'ensemble des points, en un sens précis qu'il n'est pas nécessaire de connaître. Dans un tableur, on l'obtient en ajoutant une **courbe de tendance linéaire** au nuage (voir la dernière section). Aucune formule n'est à connaître : on demande la droite à la machine et on utilise son équation.

Exemple

Ajustons notre nuage taille–pointure. Entre les deux points extrêmes du nuage, la pointure augmente de $46 - 38 = 8$ quand la taille augmente de $188 - 160 = 28$ cm : le coefficient directeur d'une droite d'ajustement est donc proche de $\frac{8}{28} \approx 0,29$ — et la méthode de Mayer donnait $\approx 0,31$. Retenons la valeur simple $a = 0,3$, et faisons passer la droite par le point moyen $G(174; 42)$: l'ordonnée à l'origine vaut alors $b = 42 - 0,3 \times 174 = 42 - 52,2 = -10,2$. La droite d'ajustement retenue a donc pour équation $y = 0,3x - 10,2$. Vérification qu'elle passe par G : $0,3 \times 174 - 10,2 = 52,2 - 10,2 = 42$. C'est bien l'ordonnée de G .



Graphiquement, la droite longe tous les points de près : l'ajustement est bon, le modèle « $\text{pointure} = 0,3 \times \text{taille} - 10,2$ » résume fidèlement nos huit élèves.

4 Interpolation et extrapolation

Une droite d'ajustement ne sert pas qu'à décorer le nuage : son équation permet d'**estimer des valeurs que l'on n'a pas mesurées**. Le vocabulaire distingue deux situations, et la distinction est essentielle.

Définition

On dispose d'une droite d'ajustement construite sur des données observées.

- **Interpoler**, c'est estimer une valeur située **à l'intérieur** de la plage des données observées (ici, pour une taille comprise entre 160 et 188 cm).
- **Extrapoler**, c'est estimer une valeur située **à l'extérieur** de cette plage (pour une taille inférieure à 160 cm ou supérieure à 188 cm).

Méthode

Estimer une valeur avec la droite d'ajustement.

1. Écrire l'équation de la droite d'ajustement : $y = ax + b$.
2. **Remplacer** x par la valeur connue et calculer y (ou, si c'est y qui est connu, résoudre l'équation pour trouver x).
3. **Conclure par une phrase**, en arrondissant de façon cohérente avec le contexte, et en précisant s'il s'agit d'une interpolation ou d'une extrapolation — la fiabilité n'est pas la même.

Exemple

Utilisons notre droite d'équation $y = 0,3x - 10,2$.

- *Interpolation*. Quelle pointure prévoir pour un élève de 170 cm ? On calcule $0,3 \times 170 - 10,2 = 51 - 10,2 = 40,8$. Le modèle prévoit une pointure d'environ 41. La valeur 170 est à l'intérieur de la plage observée (160 à 188 cm) : cette estimation est raisonnable.
- *Extrapolation*. Et pour une personne de 210 cm ? Le calcul donne $0,3 \times 210 - 10,2 = 63 - 10,2 = 52,8$... soit du 53 ! Une telle pointure n'existe pratiquement pas : à 22 cm au-delà de notre plus grand élève, la relation observée sur la classe n'a plus aucune raison d'être valable. Le calcul est juste, mais la prévision est absurde.

Remarque

Esprit critique. Trois réflexes à garder devant tout ajustement :

- plus on s'éloigne du nuage de points, moins la prévision est fiable — une extrapolation lointaine peut donner des résultats absurdes, comme notre pointure 53 ;
- un modèle affine n'est valable que **sur la plage étudiée** : rien ne garantit que le phénomène reste linéaire au-delà ;
- un **lien apparent** entre deux grandeurs ne prouve pas que l'une cause l'autre. Les ventes de glaces et le nombre de coups de soleil augmentent ensemble chaque été : aucune des deux ne provoque l'autre, c'est l'ensoleillement qui fait grimper les deux. Avant de conclure « x agit sur y », toujours se demander si une troisième grandeur ne pilote pas les deux à la fois.

5 Avec un tableur

Toutes les notions du chapitre se retrouvent au tableur, qui est l'outil attendu en pratique. On saisit les données en colonnes : les tailles en colonne A (cellules A2 à A9), les pointures en colonne B (cellules B2 à B9), avec les titres en ligne 1.

- **Obtenir le nuage de points** : sélectionner les deux colonnes de données (plage A1:B9), puis menu *Insertion* → *Graphique* → type « **nuage de points (XY)** ». La première colonne fournit les abscisses, la seconde les ordonnées.
- **Calculer le point moyen** : la fonction MOYENNE fait tout le travail.

Cellule	Formule	Résultat
D2	=MOYENNE(A2:A9)	174 (c'est $\bar{x}$)
E2	=MOYENNE(B2:B9)	42 (c'est $\bar{y}$)

- **Obtenir la droite des moindres carrés** : clic droit sur un point du nuage → *Ajouter une courbe de tendance* → choisir le type **linéaire** → cocher « **afficher l'équation sur le graphique** »

». Le tableur trace la droite et affiche son équation $y = ax + b$, prête à être utilisée pour interpoler ou extrapoler.

Le tableur calcule vite et sans erreur — mais il ne réfléchit pas à votre place. C'est à vous de vérifier que les points sont à peu près alignés avant d'ajuster, de distinguer interpolation et extrapolation, et de garder l'œil critique sur ce que le modèle a le droit de dire.