.

Statistiques

Définition d'une variable statistique (Rappels)

- **Population :** Ensemble $\mathcal{P}$ des éléments étudiés dans le contexte considéré
- **Individu :** Elément p désigné de la population $\mathcal{P}$
- **Echantillon :** Une partie E des individus de la population $\mathcal{P}$
- **Caractère :** Aspect sur lequel porte l'étude. On en distingue deux catégories :
  a) *Les caractères qualitatifs*
  b) *Les caractères quantitatifs*
- **Modalités :** Valeur réalisée du caractère étudié dans la population considérée.

Une variable statistique est alors une variable aléatoire X définie sur $\mathcal{P}$ qui à tout individu p associe la valeur $X(p) = x$ du caractère étudié (via X ici donc). Toute valeur d'observation x du caractère est une réalisation $x \in X(\mathcal{P})$ nommée *modalité*.

Relevés Numériques de l'échantillon (Rappels)

Dans cette section, la population est notée $\mathcal{P}$ et X désigne un caractère étudié sur cette population.

- **Effectif total :** Valeur eff. total $= N$ du cardinal de la population $\mathcal{P}$
- **Effectif :** Valeur eff. $(E) = n_E$ du cardinal de l'échantillon $E \subset \mathcal{P}$ considéré.
- **Fréquence :** Valeur freq. $(E) = f_E = \dfrac{n_E}{N} = \dfrac{\text{eff.}(E)}{\text{eff. total}}$ relative à l'échantillon considéré.

Les effectifs ou fréquences d'une modalité x de caractère sont associés à l'échantillon E de la façon suivante :

$$E = \{p \in \mathcal{P} \mid X(p) = x\} = X^{-1}(x)$$

Ainsi, eff.$(x) = \text{card}\left(X^{-1}(x)\right)$. En cas d'ambiguïté, on écrit E_x ou $E(x)$.

Vocabulaire : On appelle *mode* du caractère X la valeur mode $= \text{argmax}\{\text{eff.}(x_i)\ ; i \leq k\}$.

Définition : [Relevé cumulé (dé)croissant]
Soit $(x_i)_{i\leq k}$ les modalités ordonnées associées à une population. On définit :

- L'effectif cumulé croissant (resp. décroissant) de $(x_i)_{i\leq k}$ comme la suite des sommes partielles :

$$\forall i \leq k \quad \text{eff. cumul. croiss}(x_i) = ecc_i = \sum_{s\leq i} \text{eff.}(\{p \in \mathcal{P} \mid \text{caractère}(p) = x_s\})$$

- La fréquence cumulée croissante (resp. décroissante) de $(x_i)_{i\leq k}$ comme la suite des sommes partielles :

$$\forall i \leq k \quad \text{freq. cumul. croiss}(x_i) = fcc_i = \sum_{s\leq i} \text{frq.}(\{p \in \mathcal{P} \mid \text{caractère}(p) = x_s\}) = \frac{\text{eff. cumul. croiss}(i)}{\text{eff. total}}$$

Les analogues décroissant(e)s étant défini(e)s en retirant les effectifs ou fréquences de façons successives à partir de N ou de 1 respectivement.

Exemple 1 : [Baccalauréat 2026] On considère les candidats présents à la session de juin, en France.

Voie	Générale	Technologique	Professionnelle
Effectif	380 168	148 331	216 076

Identifier le caractère, calculer les fréquences et déterminer le mode. Un cumul a-t-il ici un sens naturel ?
Source : DEPP, résultats provisoires du baccalauréat, 11 juillet 2026, France.

Définition : [Regroupement par classes]
On appelle *regroupement par classes* une partition $(C_i)_{i\leq k}$ de l'ensemble $X(\mathcal{P})$ des modalités de X par la population étudiée. Formellement, si la classe C_i est l'ensemble de valeurs du caractère $C_i = \{x_1 \ ; \ \dots \ ; \ x_m\}$ est de cardinal m alors les notions d'effectifs, fréquences, cumulées ou non, sont transmises aux regroupements de façon naturelle en sommant les valeurs associées à chaque x_j de C_i. Par exemple :

$$\text{eff. } (C_i) = \sum_{j\leq m} \text{eff.}(x_j) = \text{card } \left(X^{-1}(C_i)\right)$$

On peut aussi utiliser une partition de $\mathbb{R}$ en intervalles $(I_i)_{i\leq k}$ et considérer implicitement les classes $C_i = I_i \cap X(\mathcal{P})$. On conservera eff. $(I_i) = \text{card } \left(X^{-1}(C_i)\right)$ qui reste fini.

Exemple 2 : [Salaires du secteur privé] On se restreint aux postes rémunérés de 1 300 à moins de 2 100 euros nets mensuels en équivalent temps plein (EQTP), en France en 2025. Les effectifs, exprimés en milliers d'EQTP, sont arrondis ; les classes ont été regroupées.

Classe (euros)	[1300; 1500[	[1500; 1700[	[1700; 1800[	[1800; 2100[
Effectif	903	1987	1168	3195

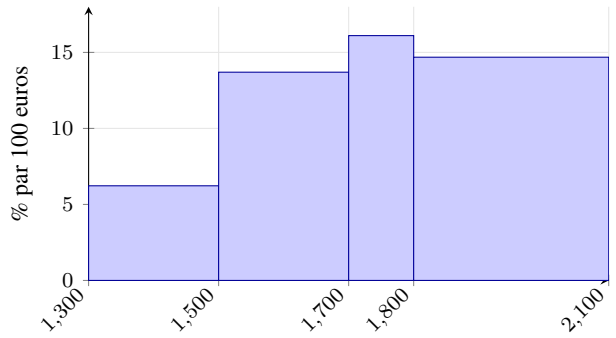
Calculer les fréquences et les fréquences cumulées sur ce sous-ensemble. Comparer la classe de plus grand effectif et celle de plus grande densité.
Source : Insee Première n° 2132, salaires du privé en 2025, publication d'octobre 2026.

Représentations Graphiques associées :

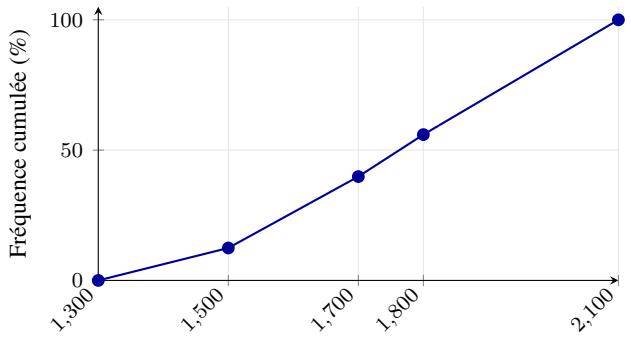
Effectuées en repère orthogonal :

- **Diagramme en bâtons** : modalités en abscisses et effectifs ou fréquences en ordonnées
- **Histogrammes** : classes de modalités en abscisses et densités en ordonnées - c'est la surface du rectangle qui a valeur de fréquence !
- **Courbes affines par morceaux** pour les relevés cumulés : modalités (ou classes) en abscisses, cumuls en ordonnées - souvent reliés par des segments.

Histogramme



Interpolation affine



Remarque : la courbe interpole les cumuls entre les bornes ; elle ne donne pas les fréquences exactes à l'intérieur des classes.

Cas d'étude avec plusieurs caractères

Supposons que, sur une même population $\mathcal{P}$, on étudie via Y un autre caractère statistique. Ainsi, pour chaque individu p, l'information statistique est un couple $(X(p); Y(p))$ d'observations.
Dans ce contexte, la seule donnée de valeurs de X observées sans considérations de celles de Y est dite *marginale*.

Vocabulaire : Le tableau formé des effectifs de la population $\mathcal{P}$ associés aux couples de modalités $(x_i; y_j)$ est appelé tableau de tri croisé. Il peut tout aussi bien être organisé par regroupements de classes (voir exemple plus bas).

Définition : [Relevé conditionnel]
Soit $(X; Y)$ un couple de caractères statistique de la population $\mathcal{P}$ étudiée. Si $C \subset Y(\mathcal{P})$ est non vide (potentiellement un singleton) alors les effectifs ou fréquences conditionnels de X sachant $Y \in C$ sont ceux associés à la restriction de X à l'échantillon de population $Y^{-1}(C)$.
Exemple : [Immobilier parisien] Une agence dispose d'un relevé de 400 appartements vendus à Paris en 2026 (données fictives). On étudie la surface habitable X (en m^2) et le prix de vente Y (en milliers d'euros).

Surface / Prix	$[0; 250[$	$[250; 400[$	$[400; 600[$	$[600; 900[$	$[900; +\infty[$	Total
$[0; 25[$	30	17	3	0	0	
$[25; 40[$	18	42	15	5	0	
$[40; 60[$	2	28	65	21	4	
$[60; 80[$	0	4	26	45	5	
$[80; 100[$	0	0	6	29	15	
$[100; +\infty[$	0	0	1	4	15	
Total						400

On lit par exemple que 65 appartements de surface appartenant à $[40; 60[$ ont été vendus à un prix appartenant à $[400; 600[$.

1. Déterminer les effectifs marginaux des deux caractères en complétant les marges du tableau.
2. En déduire les classes modales marginales de surface et de prix (au sens du plus grand effectif).
3. Déterminer les effectifs de prix conditionnellement à la classe de surface $[40; 60[$.
4. Déterminer les fréquences de surface conditionnellement à la classe de prix $[600; 900[$.
5. Comparer la fréquence marginale des appartements d'au moins 80 m^2 avec leur fréquence conditionnellement à un prix d'au moins 900 milliers d'euros.

Remarque : les effectifs marginaux se lisent dans les marges du tableau. Pour calculer une fréquence marginale, on divise par l'effectif total ; pour une fréquence conditionnelle, on divise par l'effectif de la sous-population considérée.

Caractéristiques de position

On se place dans le cadre d'un caractère X quantitatif, de modalités $(x_i)_{i\leq k}$ rangés dans l'ordre (strictement) croissant, et défini sur une population $\mathcal{P}$.
On peut alors définir les caractéristiques, dites de positions de X, suivantes :

- **Médiane :** Toute valeur m vérifiant :

$$\text{freq.}\left(X^{-1}(]-\infty;m])\right) \geq \frac{1}{2} \quad \text{et} \quad \text{freq.}\left(X^{-1}([m;+\infty[)\right) \geq \frac{1}{2}$$

  Remarque : Lorsque l'effectif total est une valeur N paire explicitement connue, on choisit le centre des valeurs médiane comme étant la médiane conventionnelle.
- **Moyenne :** La valeur μ associée à X définie comme :

$$\mu_X = \bar{X} = \frac{1}{N}\sum_{j\leq k} x_j \times \text{eff.}(x_j) = \sum_{j\leq k} x_j \times \text{freq.}\,(x_j)$$

- **Quantiles :** On nomme quantile d'ordre $q \in]0;1[$ ou plus facilement $q-$tile toute valeur t vérifiant :

$$\text{freq.}\left(X^{-1}(]-\infty;t])\right) \geq q \quad \text{et} \quad \text{freq.}\left(X^{-1}([t;+\infty[)\right) \geq 1-q$$

  Remarque : Si $q = \frac{1}{2}$ alors on retrouve la notion de médiane.

Vocabulaire : On distingue très souvent les quantiles particuliers suivants :

- Si $q = \frac{1}{4}$, on parle de *premier quartile*
- Si $q = \frac{3}{4}$, on parle de *troisième quartile*
- Si $q = \frac{1}{10}$, on parle de *premier décile*
- Si $q = \frac{9}{10}$, on parle de *dernier décile*
- Si $q = n\%$, on parle de *n-ième percentile*

On utilise souvent le **Diagramme de Tukey** -ou boîte à moustache- pour représenter synthétiquement ces données. On y place des indicateurs de positions type *quantiles*. Le plus souvent, minimum, maximum, quartiles (et médiane)
Exemple : [Distribution des salaires nets mensuels en EQTP, 2025]

Indicateur (euros)	Femmes	Hommes	Ensemble
D_1	1506	1534	1519
Médiane	2128	2328	2238
D_9	4028	4737	4441
P_{99}	8776	11743	10552
Moyenne	2578	2956	2797

Interpréter les quantiles, comparer moyenne et médiane, calculer les rapports interdéciles D_9/D_1. Peut-on déduire exactement la variance de ce tableau ?

Source : Insee Première n° 2132, salaires du privé en 2025, publication d'octobre 2026.

Caractéristiques de dispersion

Conformément au programme officiel, on se limite au cas de modalités réelles. Nous restons donc dans le cadre d'un caractère quantitatif X d'une population $\mathcal{P}$ donnée avec $X(\mathcal{P}) \subset \mathbb{R}$. Les modalités formeront la suite $(x_i)_{i\leq k}$ où k est le cardinal de $X(\mathcal{P})$ supposé fini.

- **Etendue :** Valeur $\max(X(\mathcal{P})) - \min(X(\mathcal{P}))$
- **Intervalle Interquartiles :** Si $Q_1(X)$ et $Q_3(X)$ désignent des premier et troisième quartiles respectivement, alors l'intervalle $[Q_1(X); Q_3(X)]$ est un intervalle interquartile.
  On peut aussi, par abus de langage, désigner la valeur d'amplitude $Q_3(X) - Q_1(X)$ de cet intervalle.
- **Variance :** Analogue statistique de la variance en probabilités :

$$V_X = Var(X) = \frac{1}{N}\sum_{i=1}^{k}(x_i - \bar{x})^2 \text{eff.}(x_i)$$

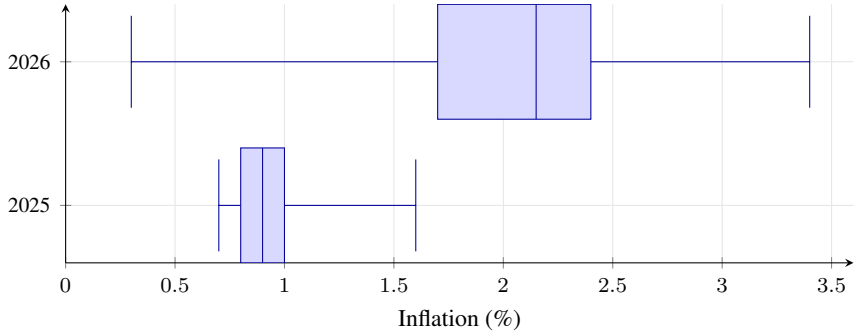
  si l'on note $\bar{x}$ la moyenne des valeurs du caractère X
- **Ecart-type :** Analogue statistique de la variance en probabilités : $\sigma_X = \sqrt{V_X}$
- **Coefficient de variation :** Valeur $\dfrac{\sigma_X}{\mu_X}$ (pour une moyenne strictement positive et un caractère à zéro significatif)

Exemple : [Inflation en France] On compare les glissements annuels de l'IPC (en pourcentage) de janvier à octobre 2025 et 2026.

Mois	J	F	M	A	M	J	J	A	S	O
2025	1,6	0,8	0,8	0,8	0,7	1,0	1,0	0,9	1,2	0,9
2026	0,3	0,9	1,7	2,2	2,4	1,8	2,1	2,4	3,1	3,4

Source : Insee, Informations rapides n° 271, 13 novembre 2026, glissements annuels de l'IPC et de l'ISJ.

Calculer les caractéristiques de position et de dispersion des deux séries. Pour les quartiles, choisir les valeurs aux rangs $\lceil N/4 \rceil$ et $\lceil 3N/4 \rceil$. Comparer les distributions.



Ici les moustaches correspondent aux valeurs extrêmes.

Cas d'étude avec plusieurs caractères

Les caractères X et Y sont supposés quantitatifs réels dans cette partie. On notera en conséquence $(x_i)_{i\leq k}$ et $(y_j)_{j\leq l}$ les modalités respectives des caractères X et Y (en nombre fini chacun donc).
On peut alors définir :

- **Covariance :** Analogue statistique de la covariance en probabilités :

$$Cov(X;Y) = \frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{l}(x_i-\bar{x})(y_j-\bar{y})\text{eff.}((x_i;y_j))$$

  si l'on note $\bar{x}$ la moyenne des valeurs du caractère X et $\bar{y}$ la moyenne des valeurs du caractère Y.
- **Coefficient de corrélation linéaire :** Valeur $\rho(X;Y) = \dfrac{Cov(X;Y)}{\sigma_X\sigma_Y}$, si les deux écarts-types sont non nuls.

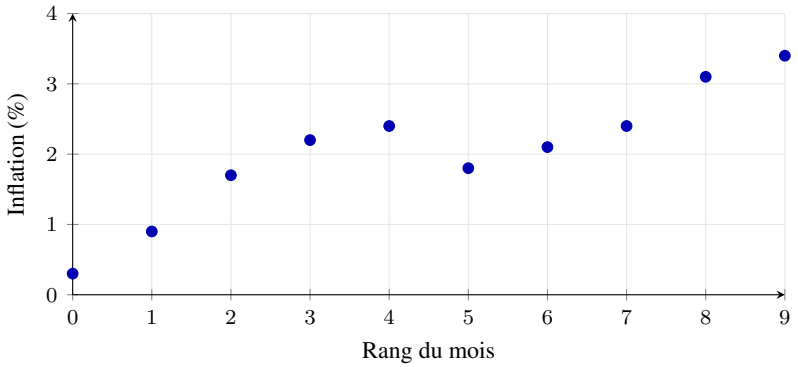
Nous illustrerons donc le coefficient de corrélation linéaire à l'aide de la méthode des moindres carrés.

Représentations Graphiques associées :

On appelle *Nuage de points* la données des points d'un repère orthogonal définis par :

$$\mathcal{N} = \{M(x;y) \mid \exists p \in \mathcal{P} \ \ (x;y) = (X(p);Y(p))\}$$

Exemple 1 : [Série chronologique] Reprenons l'inflation de janvier à octobre 2026. On associe à chaque mois son rang x_k (de 0 à 9) et le glissement annuel y_k de l'IPC. Tracer le nuage, puis calculer la covariance et la corrélation.



Exemple 2 : [Inflation globale et sous-jacente] L'ISJ cherche à dégager la tendance de fond des prix, notamment en écartant certains prix volatils. On considère les mêmes dix mois.

Mois	J	F	M	A	M	J	J	A	S	O
IPC	0,3	0,9	1,7	2,2	2,4	1,8	2,1	2,4	3,1	3,4
ISJ	0,7	0,9	1,1	1,2	1,5	1,0	1,3	1,1	1,5	1,7

Source : Insee, Informations rapides n° 271, 13 novembre 2026, glissements annuels de l'IPC et de l'ISJ.
Tracer le nuage $(IPC;ISJ)$, calculer ρ et discuter le lien entre corrélation et causalité : les indicateurs partagent une partie de leurs données de prix.

Ajustement linéaire par méthode des moindres carrés

1. **Contexte :** Données statistiques à deux variables $(X;Y)$ quantitatives réelles d'une population $\mathcal{P}$ de cardinale $N \geq 2$. Le nuage de points semble former une (presque) droite.
2. **Objectif :** Déterminer l'équation de la droite $\mathcal{D} : y = mx + p$ qui minimise les écarts avec les points du nuage $\mathcal{N}$ associé à $(X;Y)$ par projection le long de l'axe des ordonnées.
   On appelle alors *résidu de* $M_i(x_i;y_i)$ du nuage $\mathcal{N}$ sur D, une droite d'équation $y = ax + b$ la valeur : $r_i = y_i - (ax_i + b)$. Nous cherchons la droite $\mathcal{D}$ pour laquelle la quantité :
   $$R^2(a;b) = \sum_{i=1}^{N} r_i^2$$
   est minimale parmi toutes les droites possibles. Nous observerons que cela revient à chercher le couple $(m;p) \in \mathbb{R}^2$ qui minimise une certaine fonction.
3. **Moyens :** La donnée des couples $(x_i;y_i)$ pour $i \leq N = \text{card}(\mathcal{P})$ associés à chaque individu de la population étudiée.

Principe de la méthode : Si $X(\mathcal{P})$ est de cardinal au moins 2, alors il existe un couple $(m;p)$ unique dans $\mathbb{R}^2$ vérifiant :

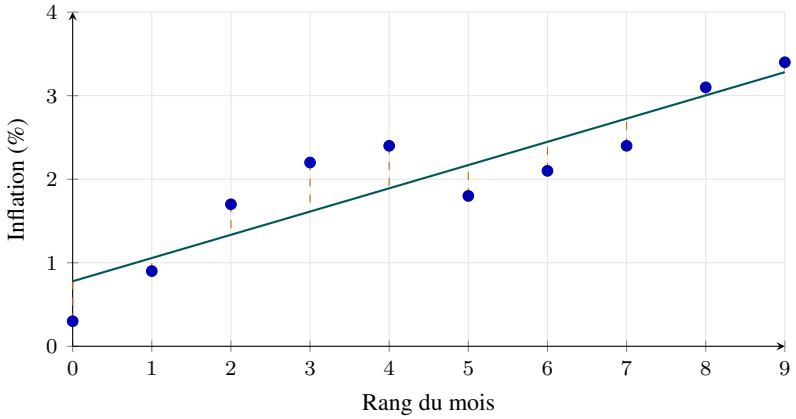
$$\sum_{i=1}^{N} \left(y_i - (mx_i + p)\right)^2 = \min_{(a;b)\in\mathbb{R}^2} R(a;b)^2$$

On dit alors que $\mathcal{D} : y = mx + p$ est la droite d'ajustement affine par méthode des moindres carrés du nuage $\mathcal{N}$.
Démonstration : [à compléter en classe]
On va étudier la fonction $R^2 : (a;b) \mapsto \sum_{i=1}^{N} \left(y_i - (ax_i + b)\right)^2$ de deux variables où $(x_i)_{\leq N}$ et $(y_i)_{i\leq N}$ sont des constantes dans l'étude. On en cherche le minimum global sur $\mathbb{R}^2$ et en quel(s) point(s) il est réalisé.

Théorème : Le problème d'optimisation sans contrainte énoncé admet pour solution le couple $(m;p)$ défini comme :

$$m = \frac{Cov(X;Y)}{V_X} \quad \text{et} \quad p = \bar{y} - m\bar{x} = \frac{1}{N}\left(\sum_{i=1}^{N}\left(y_i - \frac{Cov(X;Y)}{V_X}x_i\right)\right)$$

Application : Reprendre l'exemple chronologique. Calculer m et p, puis représenter la droite et les résidus. Discuter la pertinence d'une extrapolation à 2027.



Remarque : les segments verticaux figurent les résidus. La droite passe par le point moyen $G(\bar{x};\bar{y})$.