

Statistiques

Définition d'une variable statistique (Rappels)

- **Population** : Ensemble \mathcal{P} des éléments étudiés dans le contexte considéré
- **Individu** : Élément 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) = $f_E = \frac{n_E}{N} = \frac{\text{eff.}(E)}{\text{eff. total}}$ 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, $\text{eff.}(x) = \text{card}(X^{-1}(x))$. 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 associées à une population. On définit :

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

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

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

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

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 : on peut reprendre le cours de première année.

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 :

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

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 $\text{eff.}(I_i) = \text{card}(X^{-1}(C_i))$ qui reste fini.

Exemple 2 : on peut reprendre le cours de première année.

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.

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 $C \subset \mathbb{R}$ 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 : On donne un tableau des effectifs du couple (prix ; masse) d'un panier de biens d'une boutique :

prix / masse :	20 g	100 g	250g	500g	1000g
5 (euros)	23	15	2	0	0
7 (euros)	19	26	12	7	1
8.50 (euros)	13	21	35	22	14
9 (euros)	6	15	24	30	34
9.99 (euros)	1	3	9	24	5

Il s'agit d'un tableau de tri croisé. On peut, par exemple, lire dans ce tableau que la boutique possède 22 articles vendus à 8,50 euros qui pèsent 500 grammes.

1. Déterminer les effectifs marginaux de chaque modalité du caractère *masse*
2. Déterminer les effectifs de masse conditionnellement à 8,50 euros puis les fréquences de masse conditionnelles associées.

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 :

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

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 :

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

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

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

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)

Exemples : On peut reprendre le cours de première année

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 :

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

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 $\frac{\sigma_X}{\mu_X}$

Exemple : Déterminer les caractéristiques de dispersion définies plus haut à partir du tableau de statistiques suivant, puis réaliser le diagramme de Tukey associé :

Valeur	2	3	4	5
Effectif	3	14	7	8

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 :

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

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 $\rho(X; Y) = \frac{Cov(X; Y)}{\sigma_X \sigma_Y}$.

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]

Tracer le nuage de points associé à la série statistique dite *chronologique* extraite du tableau de CA d'une entreprise qui suit :

Année :	2015	2016	2017	2018	2019	2020	2021
Rang $x_k =$	0	1	2	3	4	5	6
CA (millions euros) $y_k =$	20,3	24,3	33,9	48,8	53,8	59,1	56,3

Exemple 2 :

Tracer le nuage de points associé à la série statistique de dates de naissances suivante :

individu p	Pierre	Simon	Julie	Alex	Sandra	Karima	Alicia
jour $x_k =$	12	7	8	30	29	4	7
mois $y_k =$	3	12	6	4	12	2	8

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 \mathcal{D} , une droite d'équation $y = ax + b$ la valeur : $r_i^2 = (y_i - (ax_i + b))^2$. 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 :

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

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 $R^2 : (a; b) \mapsto \sum_{i=1}^N (y_i - (ax_i + b))^2$ de deux variables où $(x_i)_{i \leq N}$ 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 :

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