Chapitre 7| Statistiques descriptives

But : étude de jeux de données décrivant une ou plusieurs caractéristiques communes d'une population.

Ex : taille d'individus, nbre de voitures par individu, sport pratiqué ...

I Vocabulaire.

La population est l'ensemble des individus concernés par l'étude. Un échantillon est un sous-ensemble de cette population.

Une variable ou caractère est une propriété commune aux individus de la population. Elle peut être qualitative (oui/non, football/natation/gymnastique...) ou quantitative (à valeur dans $\mathbb{R}, \mathbb{R}^2, \dots$)

Un jeu de données

$$x = (x_1, \dots, x_n)$$

est produit par l'observation d'une variable sur un échantillon de taille n.

But : étudier ces jeux de données à l'aide de représentations graphiques et d'indicateurs (moyenne, dispersion...).

Lien avec les chapitres précédents :

On modélise souvent le jeu de données comme la réalisation de variables aléatoires $(X_1, \dots, X_n)$:

$$(x_1, \dots, x_n) = (X_1(\omega), \dots, X_n(\omega)) \quad \text{pour un } \omega \in \Omega.$$

Les X_i sont souvent supposées i.i.d. et suivant une certaine loi.

On peut alors essayer d'estimer des paramètres ou faire de test d'hypothèse (ces questions sortent du cadre de ce chapitre).

II Variables qualitatives.

Une variable est dite qualitative lorsque ses valeurs ne sont pas des quantités mesurées par des nombres mais des catégories (sexe, couleurs...)

Ces catégories peuvent dans certains cas être ordonnées (par exemple par des fréquences : rarement, souvent, très souvent...).

Pour chaque valeur possible de la variable on définit

- l'effectif : le nombre d'occurrence de cette valeur dans le jeu de données
- la fréquence : le quotient de l'effectif par la taille de l'échantillon.

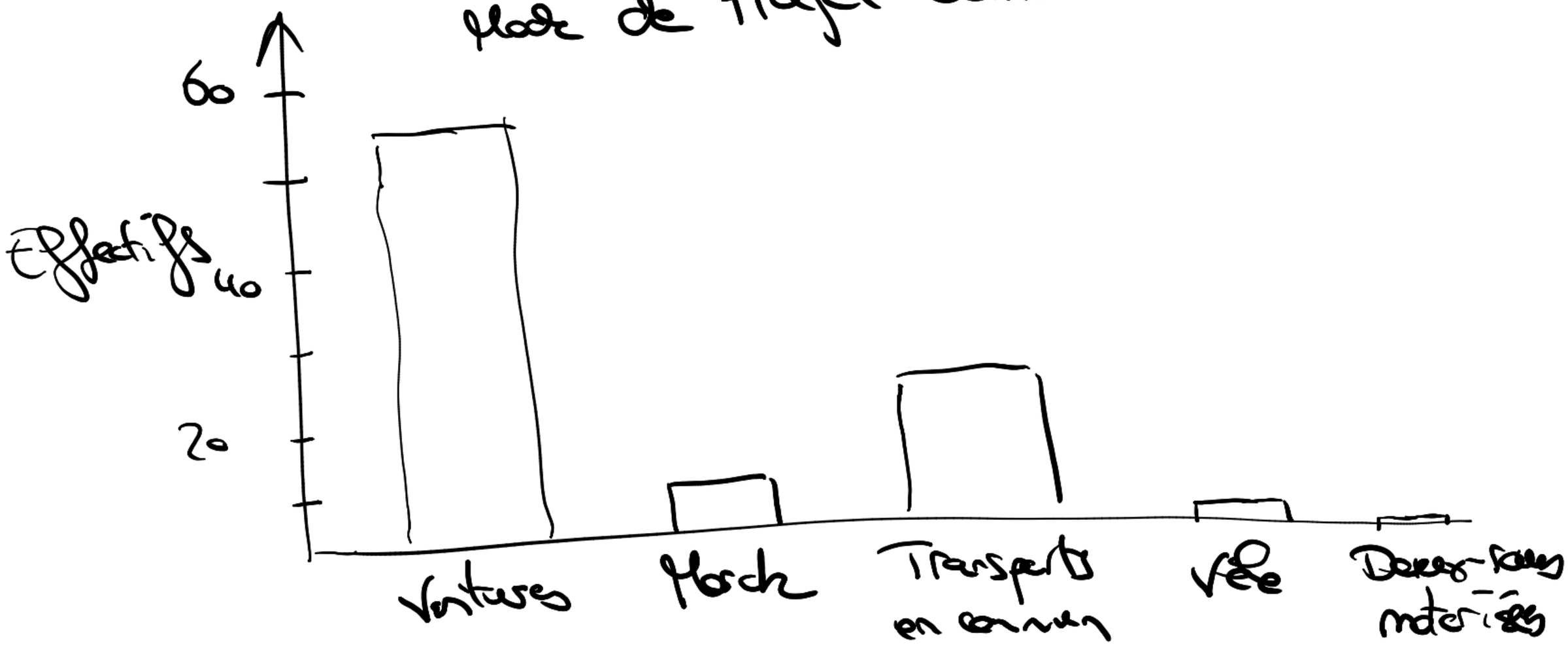
Le jeu de données peut être représenté par des tableaux ou graphiques.

Exemple. (source SDES - Insee, 2018-2019)

Questionnaire sur le mode de trajet domicile travail posé à 100 personnes vivant dans de communes désormais peuplées

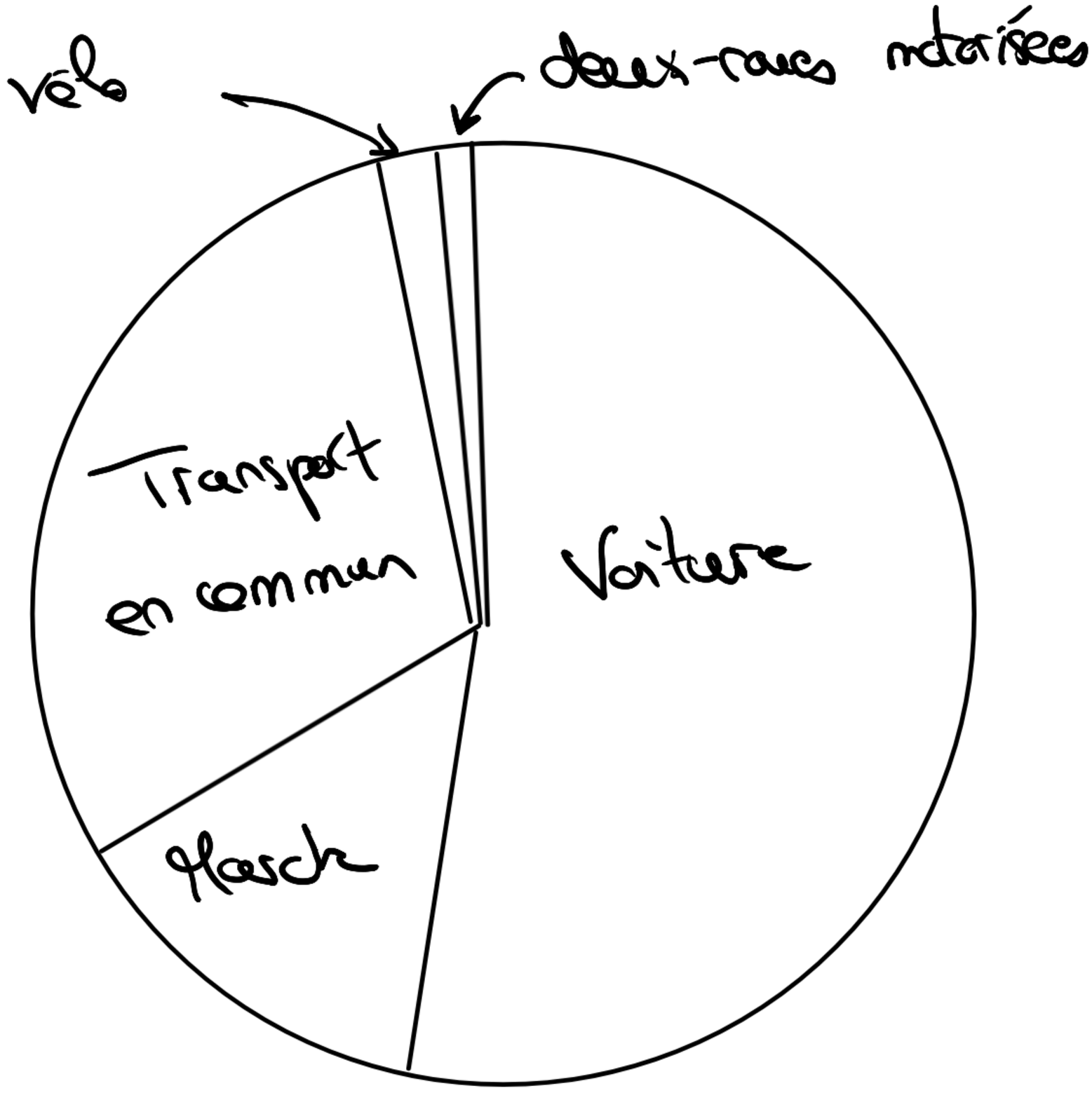
Mode de trajet	Voiture	Marche	Transports en commun	Vélo	Deux-roues motorisés
Effectif	54	10	28	5	3
Fréquence	0,54	0,1	0,28	0,05	0,03

Diagramme à bandes :



1 Diagramme circulaire

Mode de trajet domicile - travail



L'aire de la zone correspondant à chaque catégorie est proportionnelle à l'effectif.

III Variables quantitatives univariées.

On s'intéresse ici à des variable ayant pour valeurs possibles des réels.

1) Variable quantitative discrète.

On suppose ici que les valeurs possible de la variable sont $m_1, \dots, m_p$ avec $p \in \mathbb{N}^*$ et $m_1 < m_2 < \dots < m_p$.

Pour un jeu de données

$$x = (x_1, \dots, x_n)$$

on définit

- les effectifs n_i pour $i \in \{1, \dots, p\}$ par

$$n_i = \operatorname{card}\{k \in \{1, 2, \dots, n\} : x_k = m_i\}$$

- les fréquences f_i pour $i \in \{1, \dots, p\}$ par

$$f_i = \frac{n_i}{n}$$

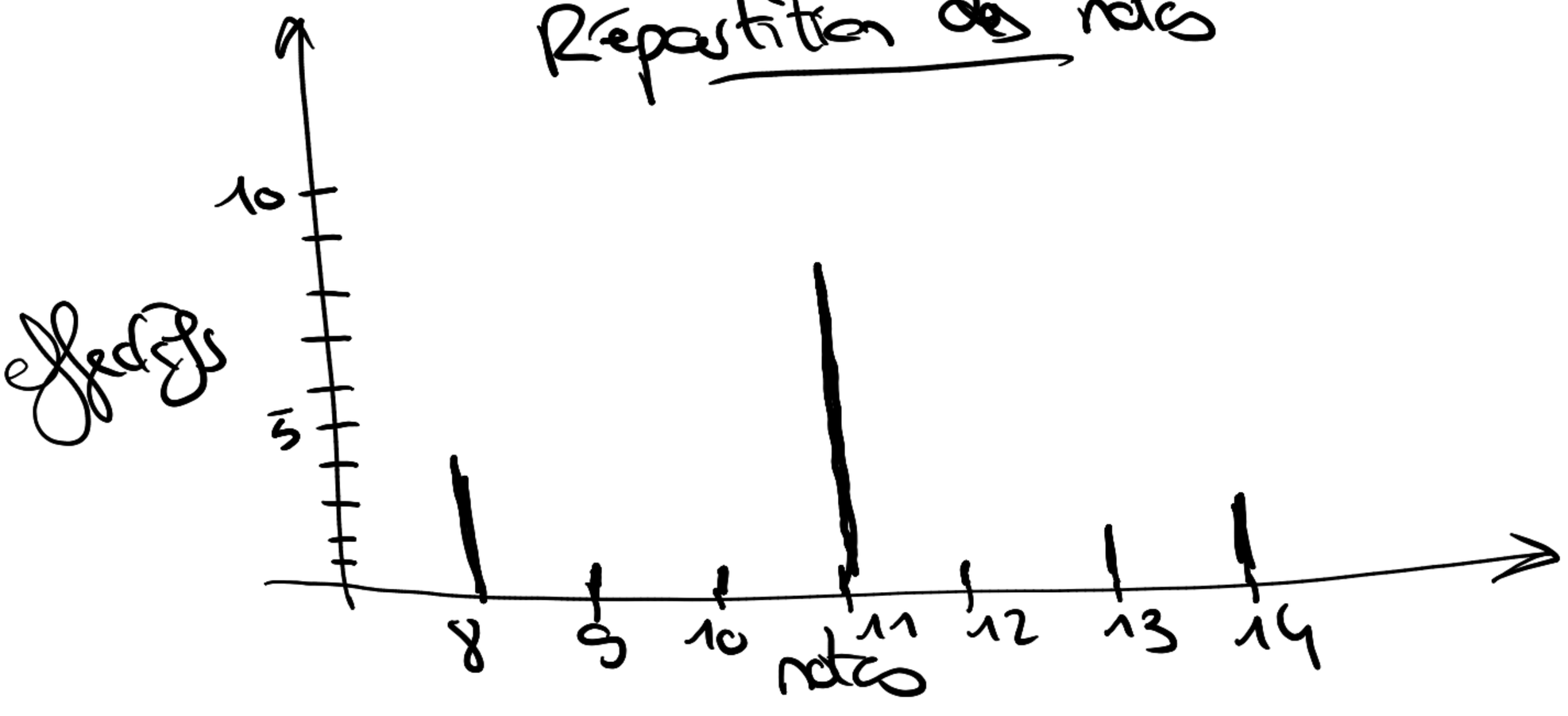
- les fréquences cumulées f_i, pour $i \in \{1, \dots, p\}$ par

$$F_i = \sum_{j=1}^{i} f_j .$$

Exemple : notes à un examen

Note	8	9	10	11	12	13	14
Effectif	4	1	1	8	1	2	3
Fréquence	0,20	0,05	0,05	0,40	0,05	0,1	0,15
Fréquence cumulée	0,20	0,25	0,30	0,70	0,75	0,85	1

Diagramme en barres (ou bâtons)



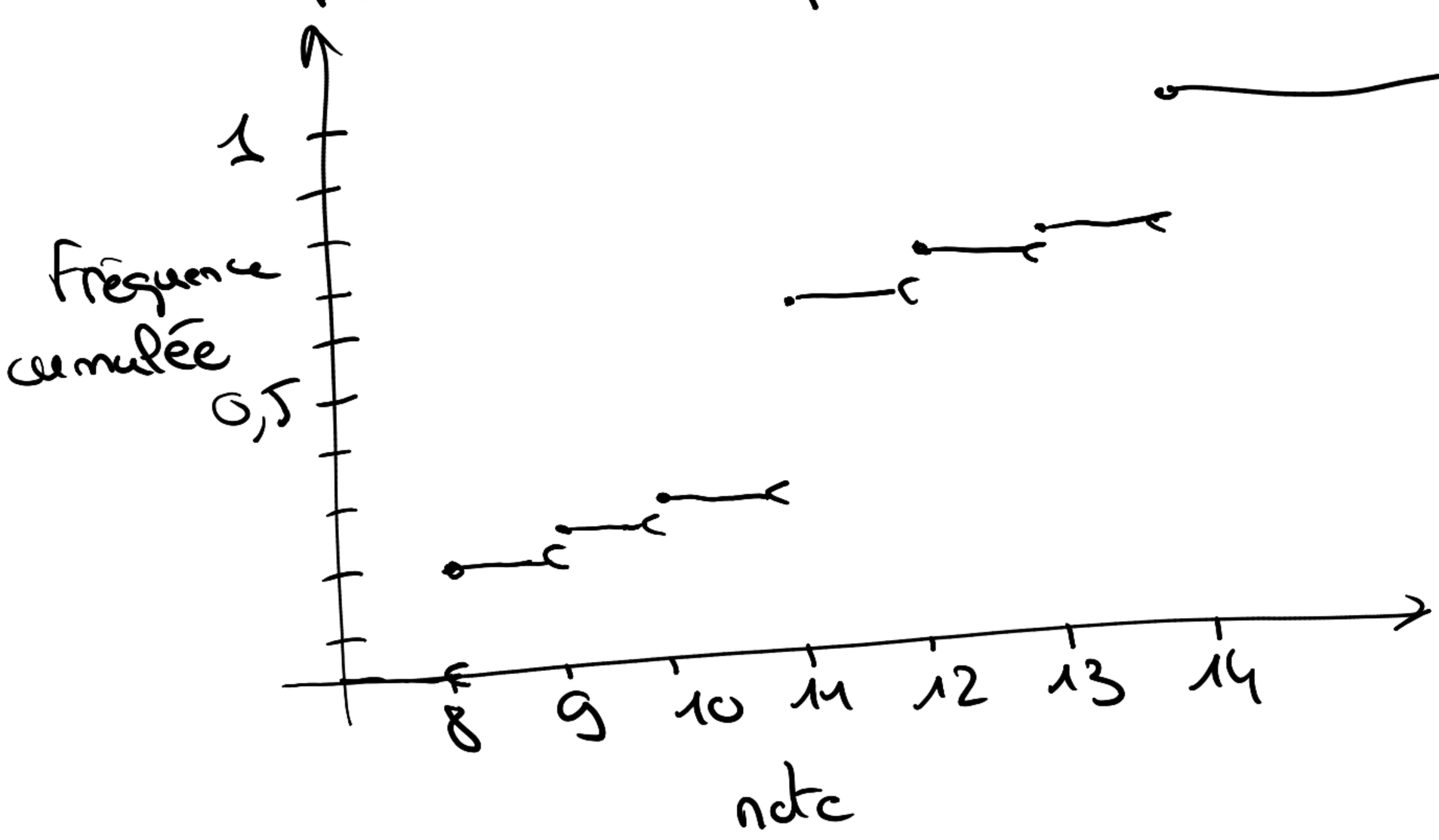
On peut définir la fonction de répartition empirique, pour tout $t \in \mathbb{R}$

$$F_n(t) = \begin{cases} 0 & \text{si } t \in]-\infty, m_1[\\ F_0 & \text{si } t \in [m_i, m_{i+1}[, \ i \in \{1, \dots, p-1\} \\ 1 & \text{si } t \in [m_p, +\infty[\end{cases}$$

Rq : on a, pour tout $t \in \mathbb{R}$,

$$F_n(t) = \frac{1}{n} \text{ coord} \{k \in \{1, \dots, n\} : x_k \leq t\}$$

En traçant le graphe de F_n on obtient un diagramme cumulatif.



2) Variables quantitatives continues.

On considère ici le cas où la variable prend ses valeurs dans un ensemble non discret (les variables ayant un très grand nbre de valeurs possibles peuvent être traitées comme des variables continues).

Dans ce cas on peut regroupe les données par classes ($a_0 < a_1 < \dots < a_p$)

$$\underbrace{[a_0, a_1[}_{=C_1} \quad \underbrace{[a_1, a_2[}_{=C_2} \quad \dots \quad \underbrace{[a_{p-1}, a_p]}_{=C_p}$$

et définir les effectifs et fréquences correspondants :

- effectifs $n_i = \text{card}\{k \in \{1, \dots, n\} : x_k \in C_i\}$
- fréquences $f_i = \frac{n_i}{n}$
- fréquences cumulées

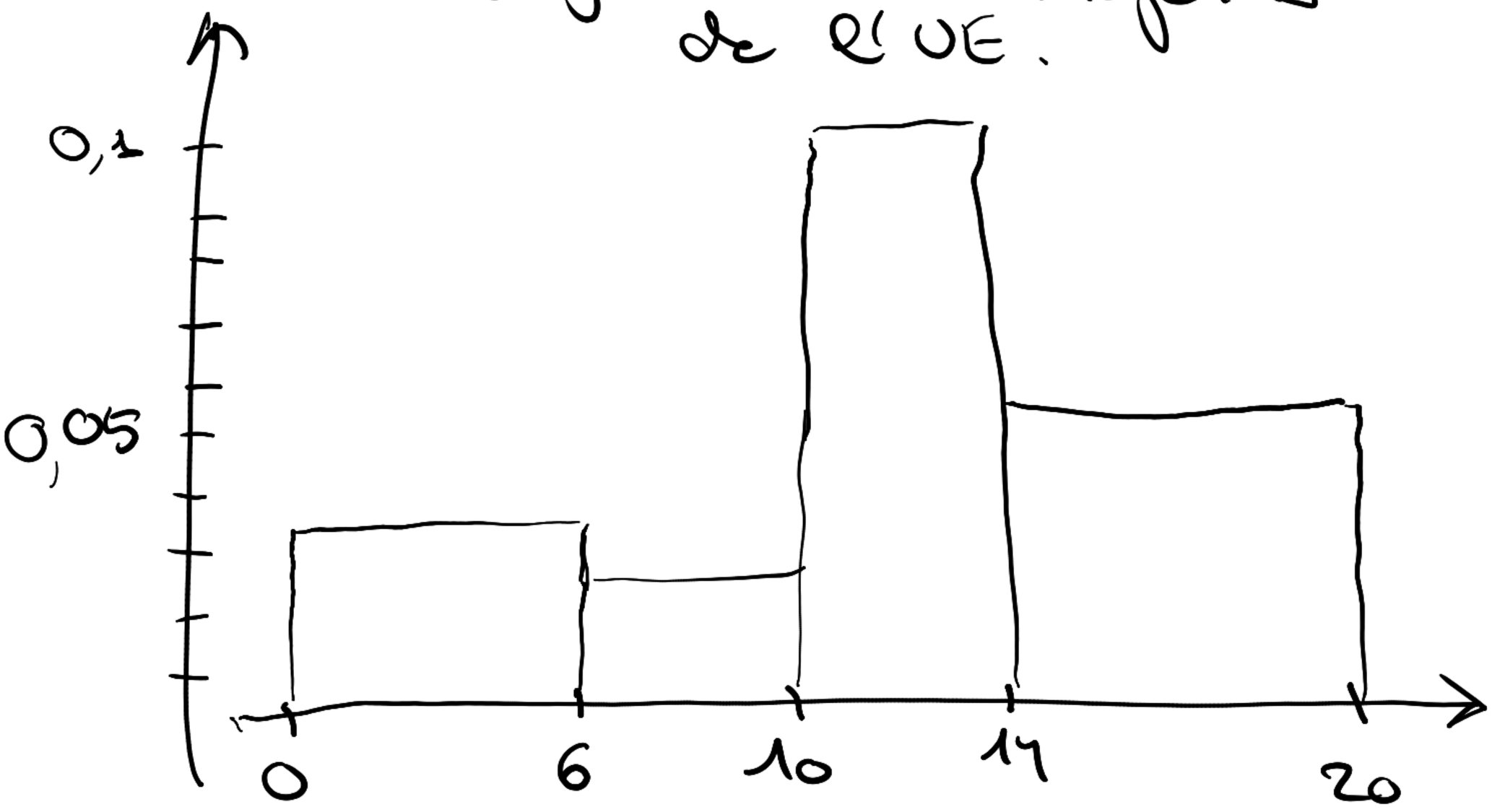
Exemple. moyennes d'une UE

moyenne	$[0,6[$	$[6,10[$	$[10,14[$	$[14,20]$
effectifs	4	2	8	6
fréquence	0,20	0,10	0,40	0,30
fréquences cumulées	0,20	0,30	0,70	1

Histogramme

Principe: diagramme composé de rectangles, chacun associé à une classe, dont la base est donnée par l'intervalle $[a_{i-1}, a_i[$ et l'aire est égale à la fréquence f_i.

Histogramme des moyennes de l'UE.



Hauteurs des rectangles :

$\frac{0,2}{6} = 0,0333\ldots$; $\frac{0,1}{4} = 0,025$; $\frac{0,4}{4} = 0,1$; $\frac{0,3}{6} = 0,05$

La somme des aires des rectangles est égale à 1.

Sans regrouper par classes, on peut s'intéresser à la fonction de répartition empirique. On suppose ici que l'on a rangé les données par ordre croissant :

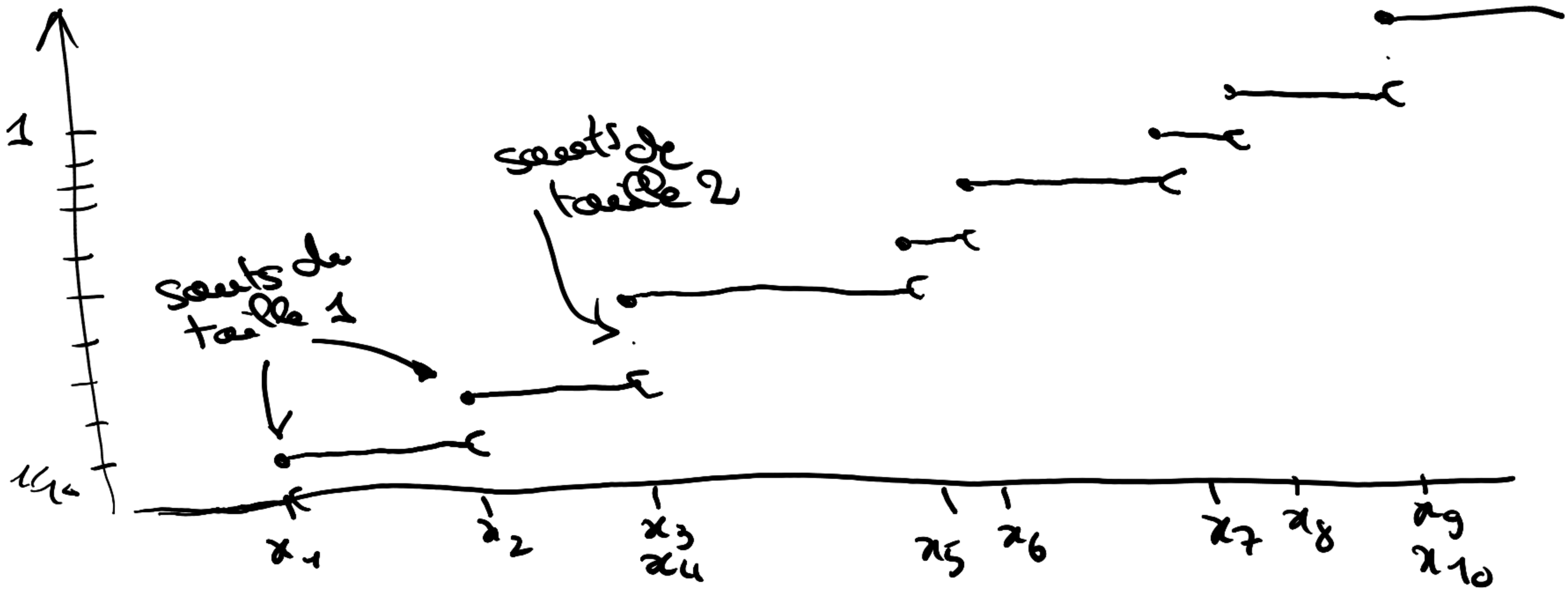
$$x_1 \leq x_2 \leq x_3 \leq \dots \leq x_n .$$

On définit, pour tout $t \in \mathbb{R}$

$$F_n(t) = \frac{1}{n} \mathrm{card}\left\{ k \in \{1, \dots, n\} : x_k \leq t \right\}$$

$$= \sum_{\substack{k=1 \\ x_k \leq t}}^{n} \frac{1}{n}$$

Tracé de F_n avec $n = 10$



3) Moyenne et variance empiriques

Définition. La moyenne empirique d'un jeu de données $(x_1, \dots, x_n)$ est le réel $\bar{x}$ défini par

$$\bar{x} = \frac{1}{n} \sum_{k=1}^{n} x_k .$$

Rq: si $x = (x_1, \dots, x_n)$ est un jeu de données correspondant à une variable discrète, de valeurs possibles $m_1, \dots, m_p$ effectifs $n_1, \dots, n_p$ et fréquences $f_1, \dots, f_p$,

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{p} n_i \, m_i$$

$$= \sum_{i=1}^{p} f_i \, m_i .$$

Exemple: si $x = (8.5, 10, 13.5, 4, 19, 11)$

$$\bar{x} = 11$$

Définition : La variance empirique d'un jeu de données $x = (x_1, x_2, \dots, x_n)$ est le réel positif $s_n^2(x)$ défini par

$$s_n^2(x) = \frac{1}{n} \sum_{k=1}^{n} (x_k - \bar{x})^2$$

$$= \frac{1}{n} \sum_{k=1}^{n} x_k^2 - (\bar{x})^2$$

L'écart-type associé est $s_n(x) = \sqrt{s_n^2(x)}$.

Rq : si $x = (x_1, \dots, x_n)$ est un jeu de données correspondant à une variable discrète,

$$s_n^2(x) = \frac{1}{n} \sum_{i=1}^{p} n_i (m_i - \bar{x})^2$$

$$= \sum_{i=1}^{p} f_i (m_i - \bar{x})^2$$

Rq : Souvent on utilise la variance empirique modifiée

$$s_{n-1}^2(x) = \frac{1}{n-1} \sum_{k=1}^{n} (x_k - \bar{x})^2 .$$

Ex: si $x = (8{,}5\ ,\ 10\ ,\ 13\ ,\ 4\ ,\ 19\ ,\ 11)$

$\bar{x} = 11$, $s_6^2(x) \approx 21{,}08$, $s_6(x) \approx 4{,}59$

4) Médiane, quantiles.

Définition: On appelle médiane d'une série de données $x = (x_1, \dots, x_n)$ un réel m tel qu'au moins la moitié des données soient inférieures ou égales à m et au moins la moitié des données soit supérieures ou égales à m.

Supposons que l'on a rangé $x = (x_1, \dots, x_n)$ par ordre croissant:

$$x_1 \leq x_2 \leq x_3 \dots \leq x_n$$

- s'il existe un entier l tel que

$$n = 2l+1 = \frac{1}{1/2}\,l + 1,$$

alors on a

$$x_1 \leq x_2 \leq \dots \leq x_l \leq x_{l+1} \leq x_{l+2} \leq \dots \leq x_{n-1} \leq x_n$$

($l+1$ données) ($l+1$ données)

Il y a donc $l+1 \geq \frac{n}{2}$ données inférieures ou égales à x_{l+1} $(x_1, x_2, \dots, x_{l+1})$ et $l+1 \geq \frac{n}{2}$ données supérieures ou égales à x_{l+1} $(x_{l+1}, x_{l+2}, \dots, x_{2l+1} = x_n)$

Donc $m = x_{l+1}$ convient.

- s'il existe un entier ℓ tel que

$$n = 2\ell,$$

on a

$$x_1 \leq x_2 \leq \dots \leq \underbrace{x_\ell}_{\ell \text{ données}} \leq \underbrace{x_{\ell+1} \leq \dots \leq x_n}_{\ell \text{ données}}$$

donc n'importe quel nombre de l'intervalle $[x_\ell, x_{\ell+1}]$ convient pour la médiane m. Plusieurs définitions sont possibles. En voici deux exemples :

* $m = \begin{cases} x_{\ell+1} & \text{si } n = 2\ell+1 \\ x_\ell & \text{si } n = 2\ell \end{cases}$

$$m = \begin{cases} x_{\ell+1} & \text{si } n = 2\ell+1 \\ \dfrac{x_\ell + x_{\ell+1}}{2} & \text{si } n = 2\ell \end{cases}$$

Cette deuxième définition est souvent utilisée par les logiciels de calcul scientifique.

Rq : comparée à la moyenne empirique, la médiane est moins sensible aux valeurs extrêmes.

Ex : si $x = (1, 3, 4, 5, 0, 2, 6, 8, 1000)$

$\bar{x} \approx 111{,}33$

$m = 4$

On peut généraliser la notion de médiane avec la notion de quantile d'ordre r.

1 Définition: pour $r \in]0;1[$ on appelle quantile d'ordre r d'un jeu de données $x=(x_1,\dots,x_n)$ un réel $Q(r)$ tel qu'au moins une proportion r des données est inférieure ou égale à $Q(r)$ et au moins une proportion 1-r est supérieure ou égale à $Q(r)$.

Rq: s'il existe $\ell \in \mathbb{N}$ tel que $n = \frac{\ell}{r} + 1$, on a (en supposant les données triées par ordre croissant)

$$\underbrace{x_1 \leq \dots \leq x_\ell}_{\ell+1 \text{ données, avec } \frac{\ell+1}{n} = \frac{\ell+1}{\ell/r+1} = r \ \frac{\ell+1}{\ell+1/r} \geq r} \leq x_{\ell+1} \leq x_{\ell+2} \leq \dots \leq x_n$$

et

$$x_1 \leq \dots \leq x_\ell \leq \underbrace{x_{\ell+1} \leq \dots \quad \dots \leq x_n}_{n-\ell \text{ données avec}}$$

$$\frac{n-\ell}{n} = 1 - \frac{\ell}{\frac{\ell}{r}+1} = 1 - r\,\frac{\ell}{\ell+r} \geq 1-r$$

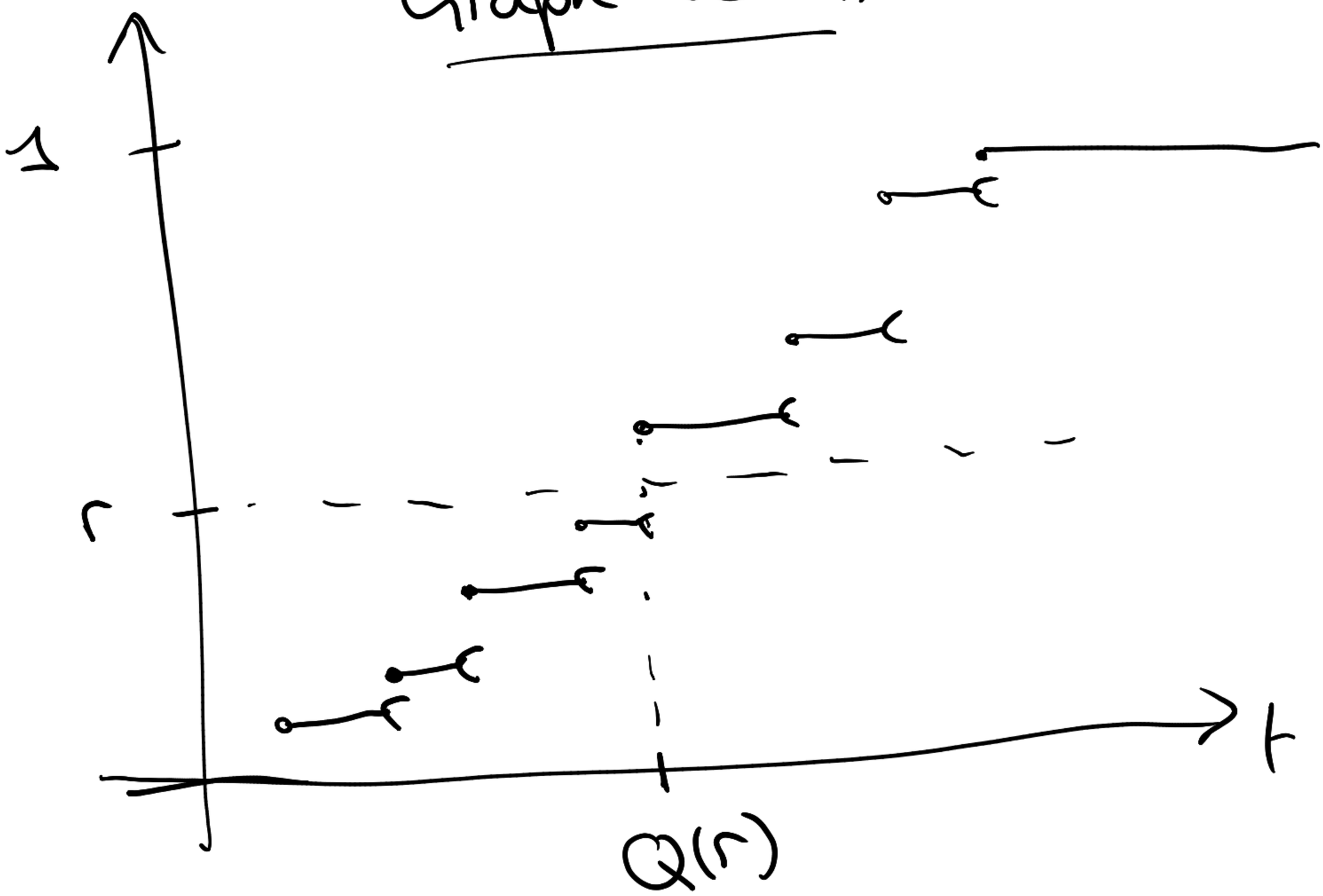
Donc $Q(r) = x_{\ell+1}$.

Comme par la médiane, s'il n'existe pas $\ell \in \mathbb{N}$ tel que $n = \frac{\ell}{r} + 1$, il existe plusieurs définitions de $Q(r)$. Voici deux exemples.

* A partir de la fonction de répartition empirique :

$$Q(r) = \inf\{t \in \mathbb{R} : F_n(t) \geq r\}$$

Graph de F_n



Avantage : avec cette définition $Q(r)$ est une des valeurs du jeu de données.

Rq :- si q est le plus petit entier tel que $\frac{q}{n} \geq r$, autrement dit (avec $\lfloor y \rfloor$ la partie entière de y),

$$q = \begin{cases} \lceil rn \rceil & \text{si} \quad rn = \lfloor rn \rfloor \\ \lfloor rn \rfloor + 1 & \text{si} \quad rn > \lfloor rn \rfloor \end{cases}$$

alors avec cette définition $Q(r) = x_q$.

En effet, on a

$$f_n(x_q) = \sum_{\substack{k=1 \\ x_k \leq x_q}}^{n} \frac{1}{n} \geq \sum_{k=1}^{q} \frac{1}{n} = \frac{q}{n} \geq r$$

et pour tout $t < x_q$

$$f_n(t) = \sum_{\substack{k=1 \\ x_k < t}}^{n} \frac{1}{n} \leq \sum_{k=1}^{q-1} \frac{1}{n} \leq \frac{q-1}{n} < r$$

- si $n = \frac{\ell}{r} + 1$, $rn = \ell + r$, $\lfloor rn \rfloor = \ell < rn$,

$q = \lfloor rn \rfloor + 1 = \ell + 1$ et on retrouve $Q(r) = x_q = x_{\ell+1}$.

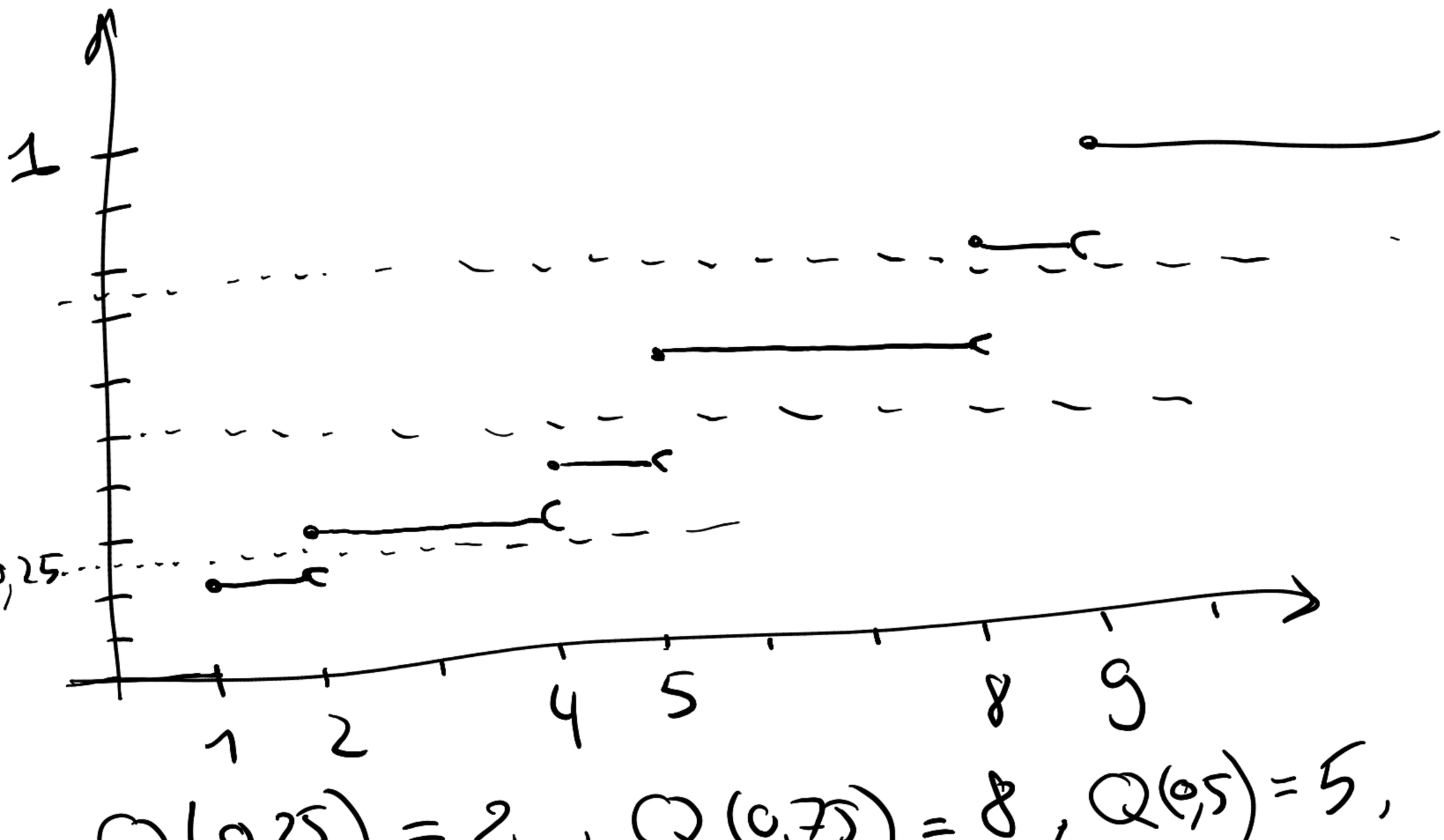
- si $r = \frac{1}{2}$ alors $q = \begin{cases} \ell+1 & \text{si } n = 2\ell+1 \\ \ell & \text{si } n = 2\ell \end{cases}$

on obtient la définition de la médiane

$$Q\left(\frac{1}{2}\right) = \begin{cases} x_{\ell+1} & \text{si } n = 2\ell+1 \\ x_\ell & \text{si } n = 2\ell \end{cases}$$

Exemple :

$a = (1, 1, 2, 4, 5, 5, 8, 8, 9, 9)$



$Q(0,25) = 2$, $Q(0,75) = 8$, $Q(0,5) = 5$,

* Autre définition possible de $Q(r)$:
interpolation linéaire (utilisée par défault par les bibliothèques numpy et pandas de python) :

$$Q(r) = a_{\lfloor r(n-1) \rfloor + 1} + \Big(r(n-1) \cdot \lfloor r(n-1) \rfloor \Big) \Big(a_{\lfloor r(n-1) \rfloor + 2} - a_{\lfloor r(n-1) \rfloor + 1} \Big)$$

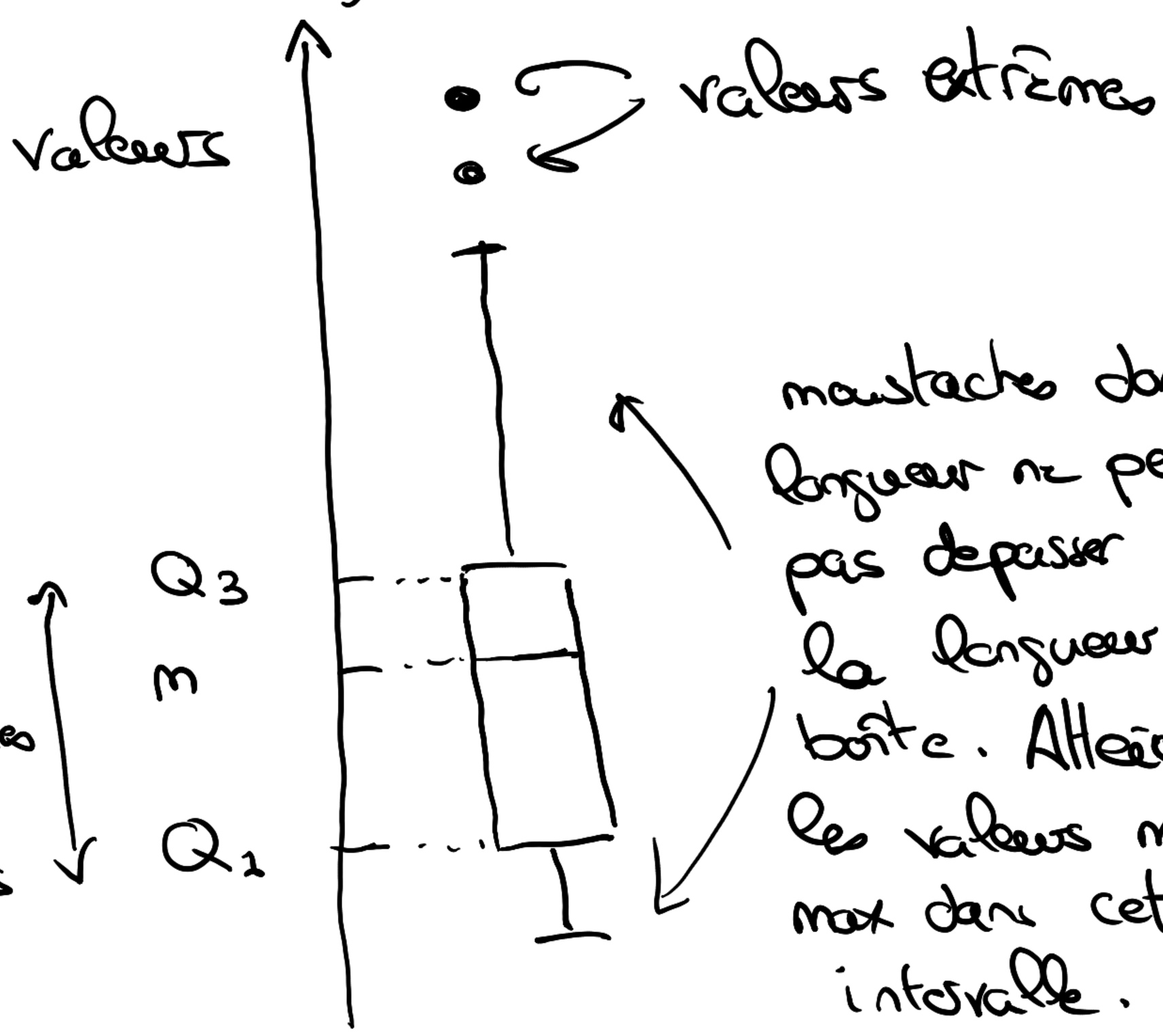
Rq : avec $r = \frac{1}{2}$ avec cette autre définition on obtient la médiane

$$Q\left(\frac{1}{2}\right) = \begin{cases} x_{\ell+1} & \text{si} \quad n = 2\ell+1 \\ \dfrac{x_\ell + x_{\ell+1}}{2} & \text{si} \quad n = 2\ell \end{cases}$$

5) Boîte à moustaches (boxplot)

Diagramme permettant de représenter certains indicateurs.

On définit $m = Q(0,5)$, les quartiles $Q_1 = Q(0,25)$, $Q_3 = Q(0,75)$



II Variables quantitatives bi-variées.

1) Corrélations.

Cette fois le jeu de données est de la forme $((x_1,y_1),(x_2,y_2),\dots,(x_n,y_n))$.

En plus de $\bar{x}$, $\bar{y}$, $s_n^2(x)$, $s_n^2(y)$ ont peut définir la covariance empirique

$$c_n = c_n(x,y) = \frac{1}{n}\sum_{k=1}^{n}(x_k-\bar{x})(y_k-\bar{y})$$

$$= \frac{1}{n}\sum_{k=1}^{n} x_k y_k - \bar{x}\bar{y}$$

et le coefficient de corrélation (lorsque $s_n(x)>0$ et $s_n(y_k)>0$)

$$\rho_n(x,y) = \frac{c_n(x,y)}{s_n(x)\,s_n(y)}$$

Puisque l'inégalité de Cauchy Schwarz implique

$$|c_n(x,y)| = \left| \frac{1}{n} \sum_{k=1}^{n} (x_k - \bar{x})(y_k - \bar{y}) \right|$$

$$\leq \underbrace{\sqrt{\frac{1}{n} \sum_{k=1}^{n} (x_k - \bar{x})^2}}_{= s_n(x)} \underbrace{\sqrt{\frac{1}{n} \sum_{k=1}^{n} (y_k - \bar{y})^2}}_{= s_n(y)},$$

on a

$$-1 \leq \rho_n(x,y) \leq 1$$

avec $\rho_n(x,y) = 1$ ou $\rho_n(x,y) = -1$

si et seulement si

$$\begin{pmatrix} x_1 - \bar{x} \\ \vdots \\ x_n - \bar{x} \end{pmatrix} \text{ et } \begin{pmatrix} y_1 - \bar{y} \\ \vdots \\ y_n - \bar{y} \end{pmatrix}$$

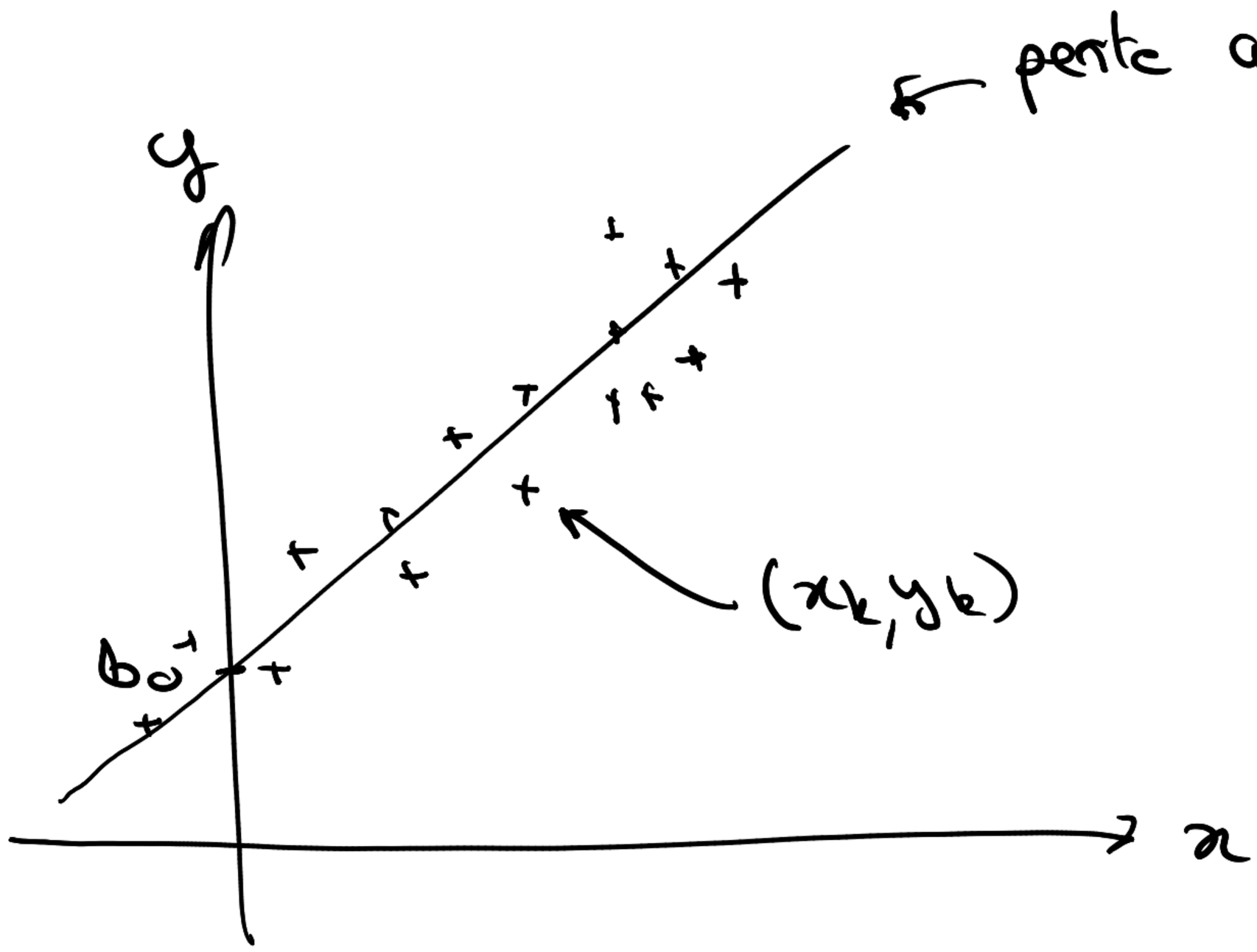
sont colinéaires, c'est à dire il existe $\lambda_1, \lambda_2 \in \mathbb{R}$ tels que, pour tout $k \in \{1, \dots, n\}$

$$\lambda_1 x_k + \lambda_2 y_k = \lambda_1 \bar{x} + \lambda_2 \bar{y}$$

↳ les (x_k, y_k) sont sur la même droite.

2) Régression linéaire

Principe : lorsque $\rho_n(x,y)$ est proche de -1 ou 1 on veut trouver une droite donnant une bonne approximation de la répartition des données.



Pour $a, b \in \mathbb{R}$ on pose

$$J(a,b) = \sum_{k=1}^{n} \left(y_k - (a x_k + b) \right)^2$$

Minimiser J revient à chercher la meilleure droite dans le sens des moindres carrés.

Proposition : si $s_n(x) > 0$, l'application J admet un minimum global sur $\mathbb{R}^2$, atteint en $(a_0, b_0) \in \mathbb{R}^2$ satisfaisant

$$\begin{cases} a_0 = \dfrac{c_n(x,y)}{s_n^2(x)} \\ b_0 = \bar{y} - a_0 \bar{x} \end{cases}$$

Idée de preuve

J est C^∞ sur $\mathbb{R}^2$ car polynomiale, et

$$\begin{cases} \dfrac{\partial J}{\partial a}(a,b) = -2 \displaystyle\sum_{k=1}^{n} x_k \left(y_k - (a x_k + b)\right) \\ \dfrac{\partial J}{\partial b}(a,b) = -2 \displaystyle\sum_{k=1}^{n} \left(y_k - (a x_k + b)\right) \end{cases}$$

Donc tout point critique (a_0, b_0) vérifie

$$\begin{cases} \dfrac{\partial J}{\partial a}(a_0, b_0) = 0 \\ \dfrac{\partial J}{\partial b}(a_0, b_0) = 0 \end{cases}$$

$$\Leftrightarrow \begin{cases} \sum_{k=1}^{n} x_k y_k - a_0 \sum_{k=1}^{n} x_k^2 - b_0 \sum_{k=1}^{n} x_k = 0 \\ \sum_{k=1}^{n} y_k - a_0 \sum_{k=1}^{n} x_k - b_0 n = 0 \end{cases}$$

$$\Leftrightarrow \begin{cases} a_0 \left(S_n^2(x) + \bar{x}^2 \right) + b_0 \bar{x} = c_n(x,y) + \bar{x}\bar{y} \\ a_0 \bar{x} + b_0 = \bar{y} \end{cases}$$

$$\Leftrightarrow \begin{cases} a_0 \left(S_n^2(x) + \cancel{\bar{x}^2} \right) + \left(\cancel{\bar{y}} - \cancel{a_0 \bar{x}} \right) \bar{x} = c_n(x,y) + \cancel{\bar{x}\bar{y}} \\ b_0 = \bar{y} - a_0 \bar{x} \end{cases}$$

$$\Leftrightarrow \begin{cases} a_0 = \dfrac{c_n(x,y)}{S_n^2(x)} \\ b_0 = \bar{y} - a_0 \bar{x} \end{cases}$$

Pour justifier que ce point critique est le minimum global de J on peut montrer que J est strictement convexe sur $\mathbb{R}^2$ en montrant que la matrice Hessienne $H(x,y)$ de J est définie positive en tout $(x,y)\in\mathbb{R}^2$.

On a, pour tout $(x,y)\in\mathbb{R}^2$,

$$H(x,y) = \begin{pmatrix} 2\sum_{k=1}^{n} x_k^2 & 2\sum_{k=1}^{n} x_k \\ 2\sum_{k=1}^{n} x_k & 2\sum_{k=1}^{n} 1 \end{pmatrix}$$

$$= -2n \begin{pmatrix} s_n^2(x) + \bar{x}^2 & \bar{x} \\ \bar{x} & 1 \end{pmatrix}$$

Comme, pour tout $(a,y) \in \mathbb{R}^2$,

$$\det\left(H(a)\right) = \frac{1}{n^2}\left(s_n^2(a) + \bar{a}^2 - \bar{a}^2\right) = \frac{s_n^2(a)}{n^2} > 0$$

et $\mathrm{Tr}\left(H(a)\right) = s_n^2(a) + \bar{a}^2 + 1 > 0$,

J est bien strictement convexe.