

Chapitre 7 | Statistiques descriptives |

But : étude de jeux de données.
décrivent 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) :

$$(a_1, \dots, a_n) = (X_1(\omega), \dots, X_n(\omega)) \quad \text{pour un } \omega \in \Omega.$$

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 des communes entièrement 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 barres:
mode de trajet domicile-travail

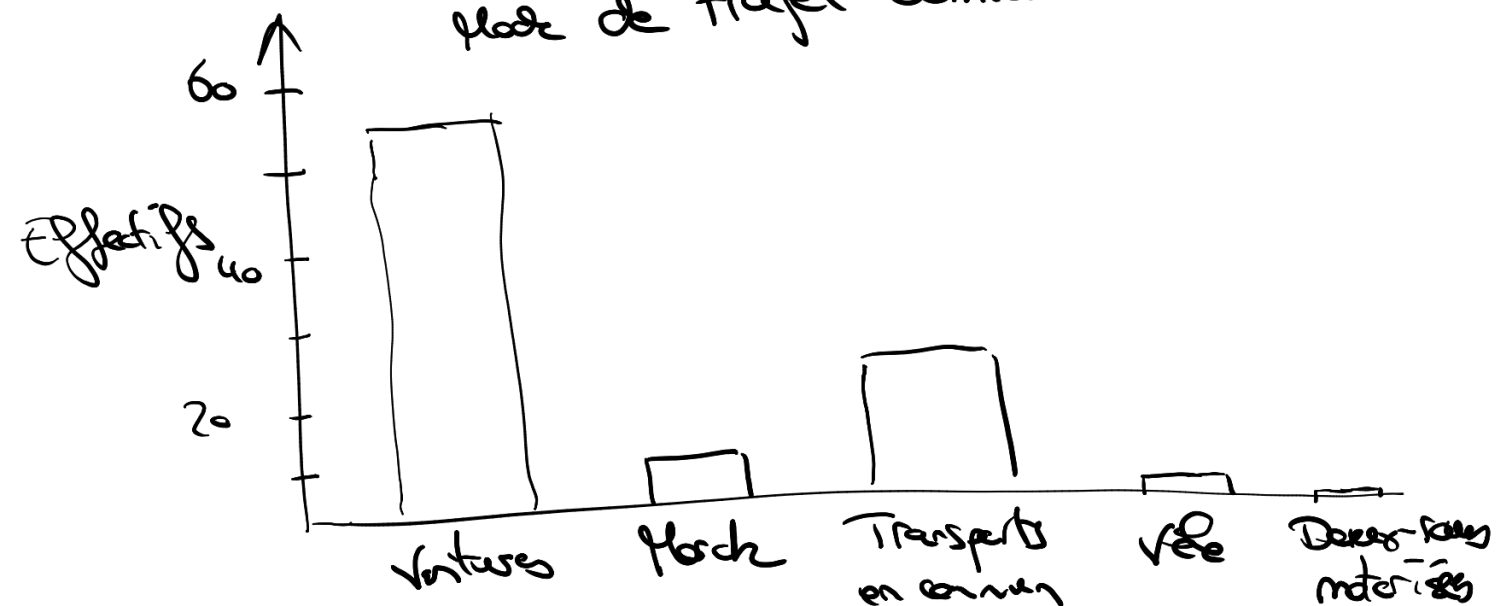
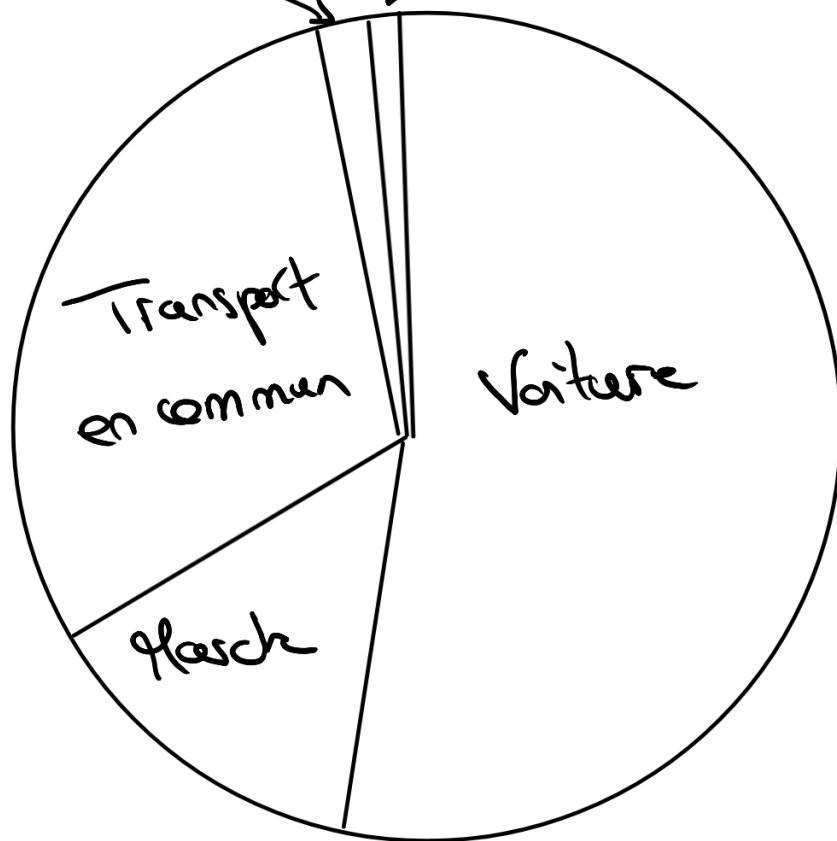


Diagramme circulaire

Mode de trajet domicile - travail

vélo deux-roues motorisés



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

III Variables quantitatives univariées.

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

1) Variables quantitatives discrètes.

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

Par un jeu de données
 $x = (x_1, \dots, x_n)$

on définit

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

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

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

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

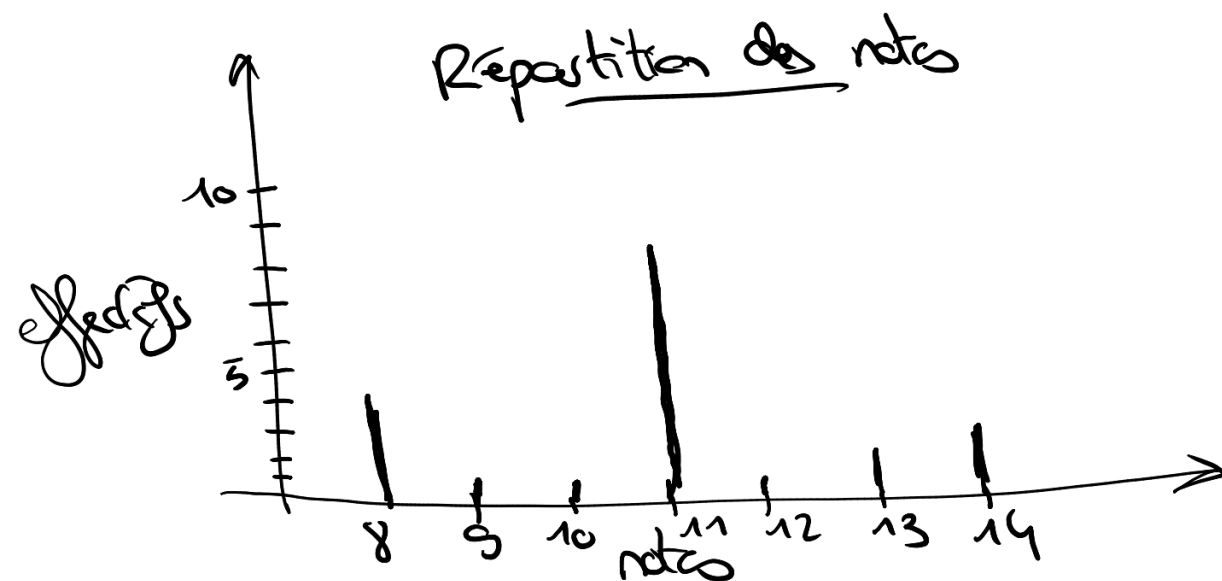
- les fréquences cumulées F_i , pour $i \in \{1, \dots, p\}$ par

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

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)



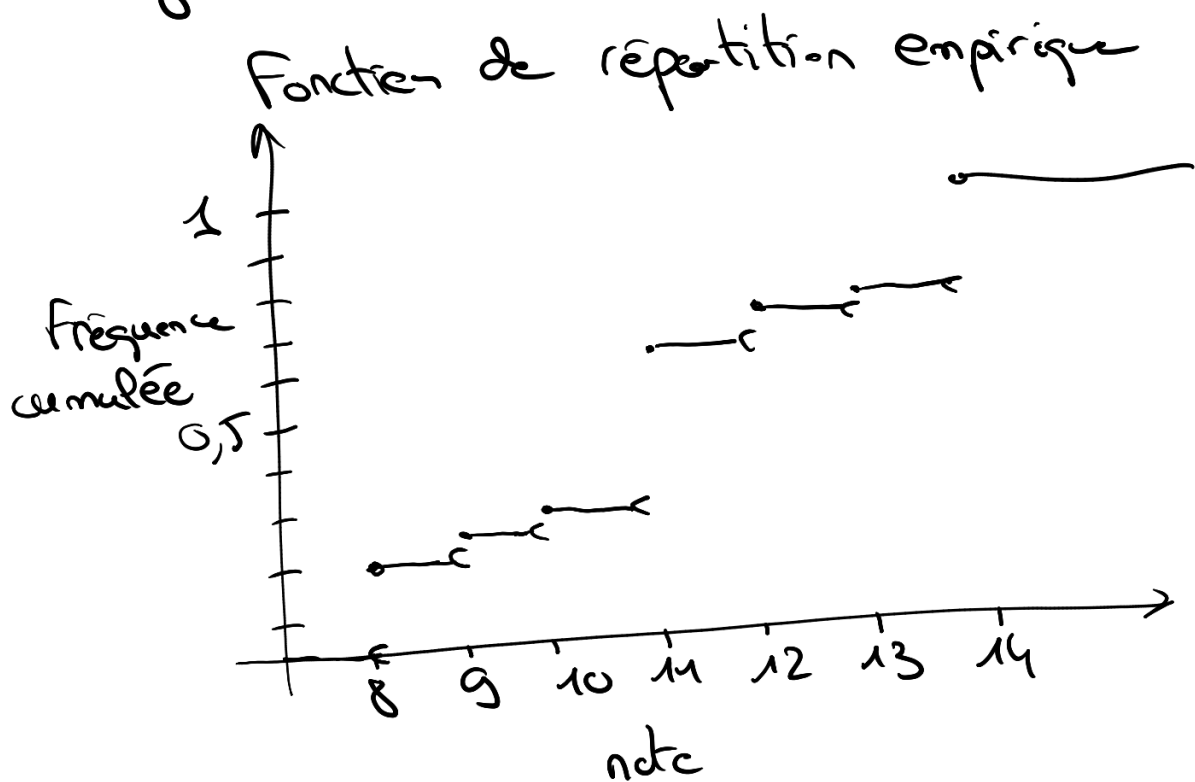
On peut définir la fonction de répartition empirique, pour tout $t \in \mathbb{R}$

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

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

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

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



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 regrouper les données par classes (où $a_0 < a_1 < \dots < a_p$)

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

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

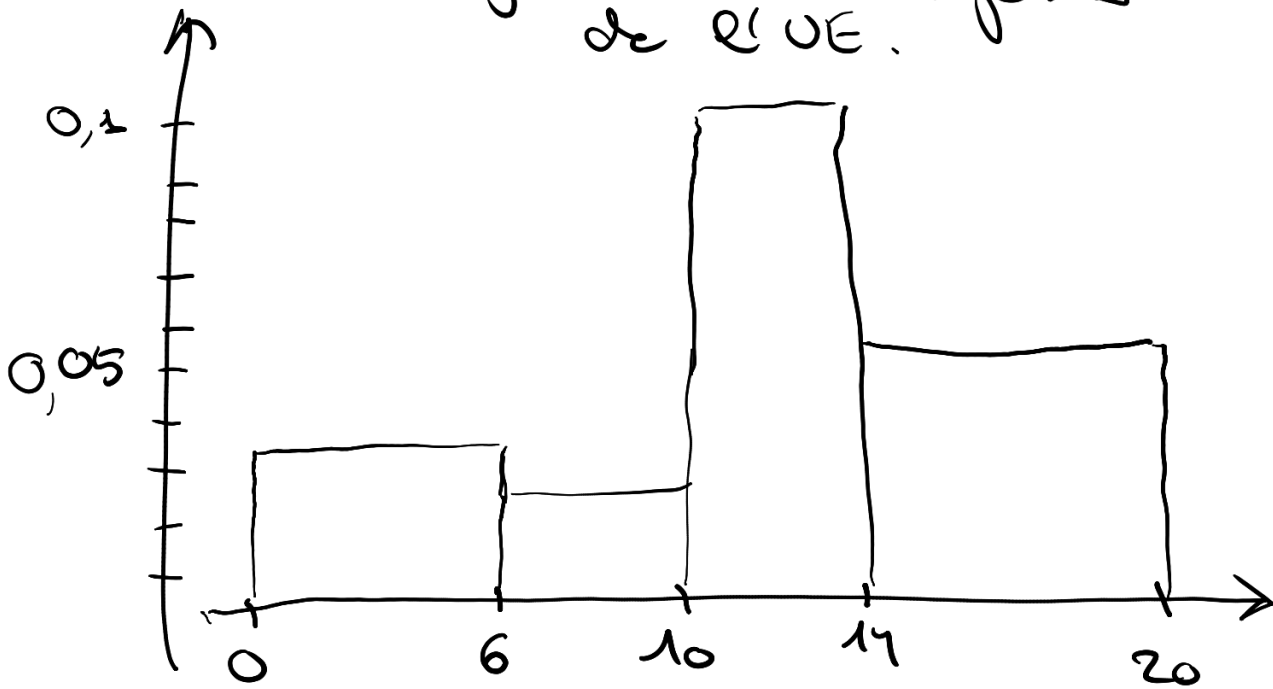
Exemple. moyennes d'une UE

moyennes	$[0,6[$	$[6,10[$	$[10,14[$	$[14,20]$
effectifs	4	2	8	6
fréquence	0,20	0,10	0,40	0,30
fréquences cumulees	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 :

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

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:

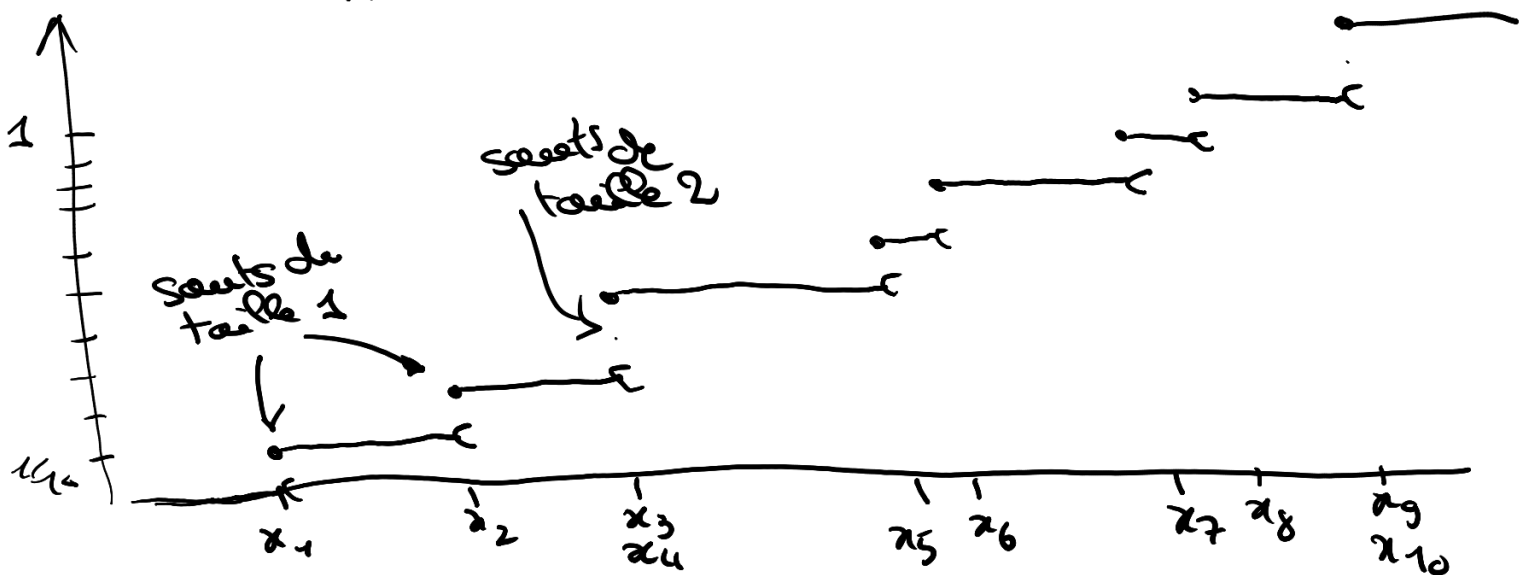
$$x_1 \leq x_2 \leq x_3 \leq \dots \leq x_n.$$

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

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

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

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 $x = (x_1, x_2, \dots, x_n)$ et le réel positif $s_n^2(x)$ définie par

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

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,

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

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 $x = (8,5, 10, 13, 4, 19, 11)$

$\bar{x} = 11$, $s^2_6(x) \approx 21,08$, $s_6(x) \approx 4,59$

1) Médiane, quantiles.

Définition: on appelle médiane d'un jeu 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

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

$\underbrace{\hspace{15em}}_{l+1 \text{ données}}$

$\underbrace{\hspace{15em}}_{l+1 \text{ données}}$

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

$$n = 2l,$$

on a

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

$\underbrace{\hspace{15em}}$
 $\underbrace{\hspace{15em}}$

l données
 l données

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

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

$$m = \begin{cases} x_{l+1} & \text{si } n = 2l + 1 \\ \frac{x_l + x_{l+1}}{2} & \text{si } n = 2l \end{cases}$$

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 $x = (1, 3, 4, 5, 0, 2, 6, 8, 100000)$

$$\bar{x} \approx 11114,33$$

$$m = 4$$

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

Définition: par $r \in]0, 1[$ 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 $Q(r)$.

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

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

$l+1$ données, avec $\frac{l+1}{n} = \frac{l+1}{\frac{l}{r} + 1} = r \frac{l+1}{l+1/r} \geq r$

et

$$a_1 \leq \dots \leq a_l \leq a_{l+1} \leq \dots \leq a_n$$

$n-l$ données avec

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

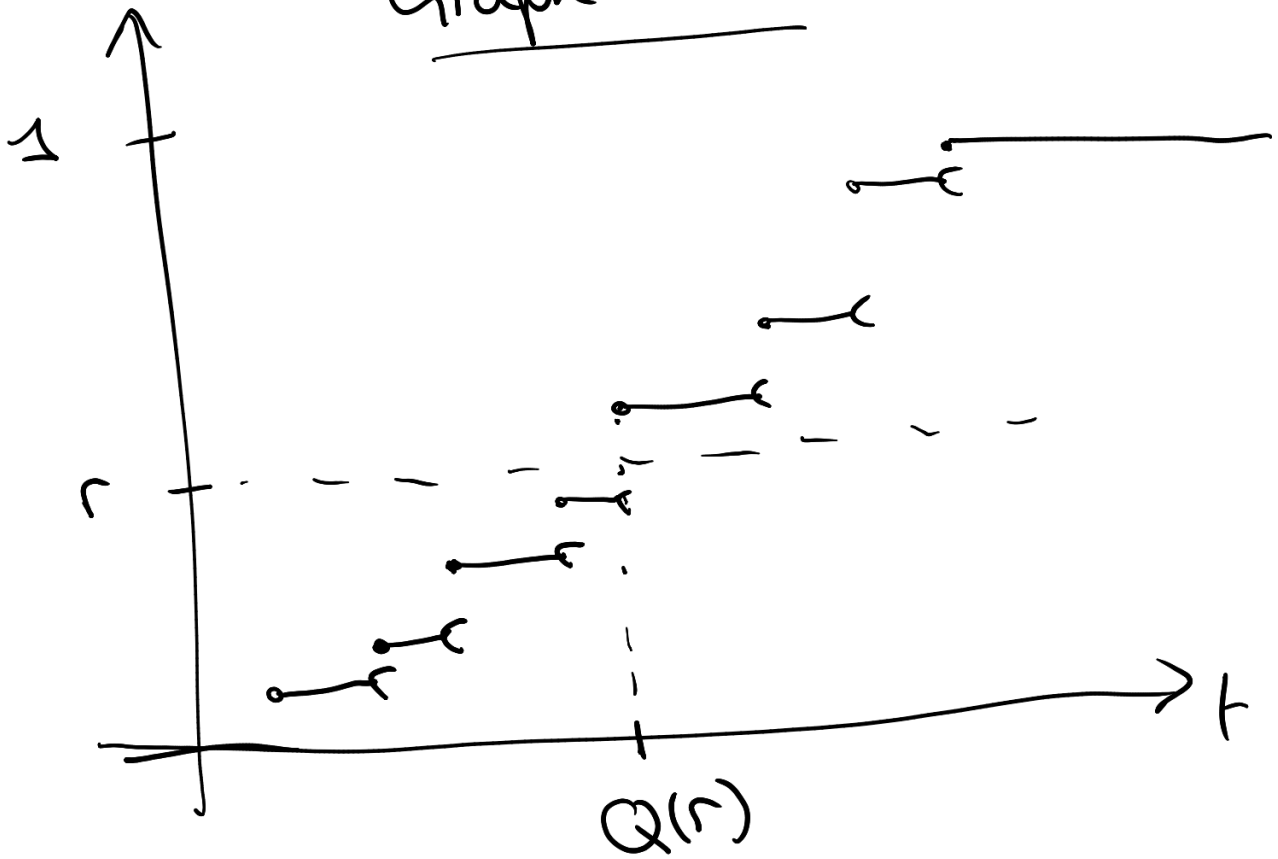
Donc $Q(r) = a_{l+1}$.

Comme par la médiane, s'il n'existe pas $l \in \mathbb{N}$ tel que $n = \frac{l}{r} + 1$, 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 of F_n



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

Req :- 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),

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

alors avec cette définition $Q(r) = \alpha_q$.

En effet, on a

$$F_n(\alpha_q) = \sum_{\substack{k=1 \\ \alpha_k \leq \alpha_q}}^n \frac{1}{n} \geq \sum_{k=1}^q \frac{1}{n} = \frac{q}{n} \geq r$$

et pour tout $t < \alpha_q$

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

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

$q = \lfloor rn \rfloor + 1 = l+1$ et on retrouve $Q(r) = \alpha_q = \alpha_{l+1}$.

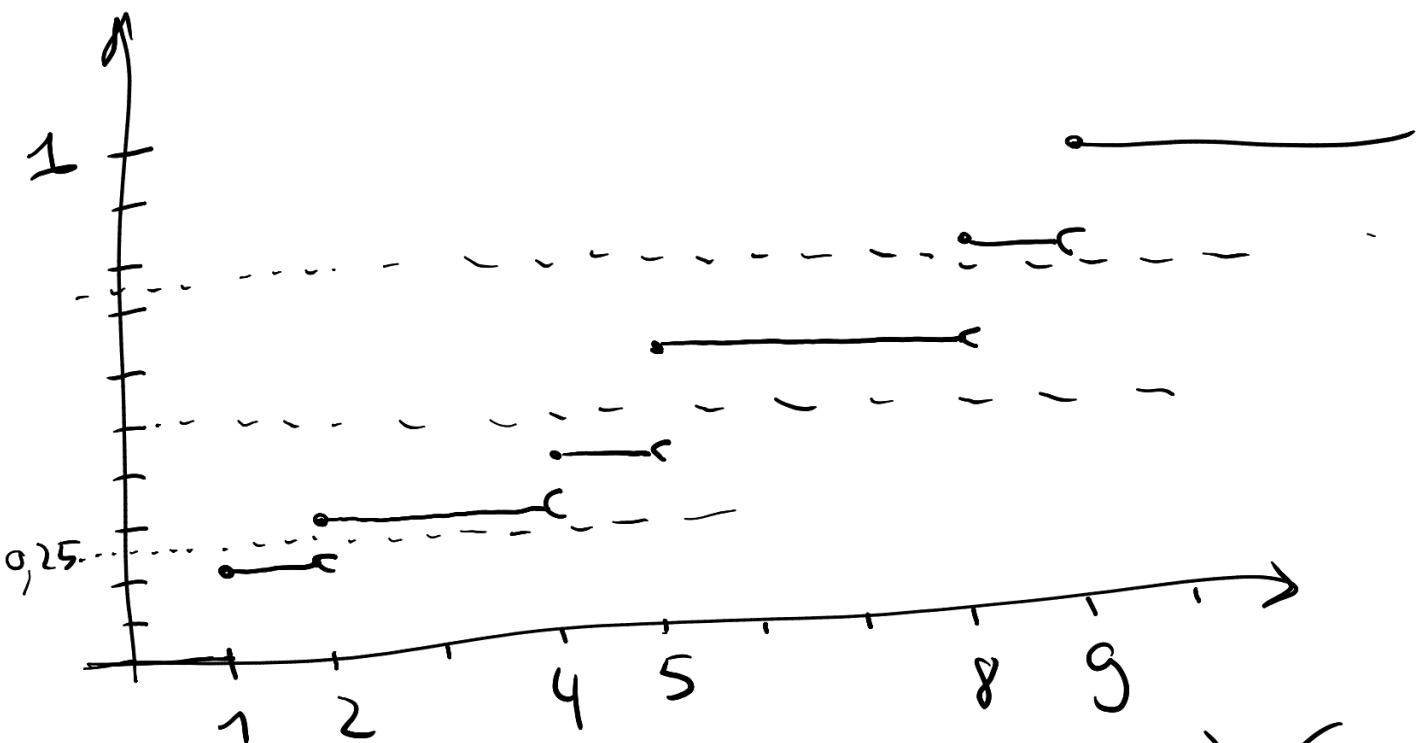
- si $r = \frac{1}{2}$ alors $q = \begin{cases} 2l+1 & \text{si } n = 2l+1 \\ 2l & \text{si } n = 2l \end{cases}$

on obtient la définition de la médiane

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

Exemple:

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



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

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

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

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

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

5) Boîte à moustaches (boxplot)

Diagramme permettant de représenter certains indicateurs.

On définit $m = Q(0,5)$, les quartiles

$Q_1 = Q(0,25)$, $Q_3 = Q(0,75)$

