Statistique descriptive

Théorie et applications

Jaouad Madkour
`jmadkour@uae.ac.ma`

Faculté des Sciences Juridiques, Économiques et Sociales de Tanger
Département des Sciences économiques et de Gestion

23 février 2026

jmadkour.org

Introduction

1. Statistique vs Statistiques
2. Statistique descriptive vs Statistique inférentielle
3. Plan du cours

Statistique *vs* Statistiques

Statistique

La statistique est la science dont l'objet est d'extraire, analyser et exploiter l'information contenue dans des données afin de prendre des décisions rationnelles dans des domaines par l'incertitude comme l'économie, la finance, l'assurance et le marketing.

Statistiques

Les statistiques désignent un ensemble de données numériques collectées de manière systématique dans un champs d'étude bien défini. Il s'agit par exemple de statistiques de naissances en démographie, de statistiques pluviométriques en météorologie ou de statistiques de production industrielle en économie.

Statistique descriptive *vs* Statistique inférentielle

Statistique descriptive

La statistique descriptive sert à faire une description graphique et numérique des phénomènes observés. Elle a pour objet de résumer un grand nombre de données en quelques indicateurs synthétiques faciles à manipuler et de présenter ces données de manière à en faciliter la visualisation.

Statistique inférentielle

La statistique inférentielle consiste à inférer et à extrapoler à la population toute entière les propriétés empiriques mises en évidence sur un échantillon représentatif et à valider ou invalider des hypothèses sur les propriétés théoriques de la population.

Plan du cours

Chapitre 1

Généralités sur la statistique

1. Terminologie statistique
2. Typologie des variables statistiques
3. Opérateurs arithmétiques

SECTION 1

TERMINOLOGIE STATISTIQUE

1. POPULATION, ÉCHANTILLON & INDIVIDU
2. VARIABLE STATISTIQUE, MODALITÉ & DONNÉE STATISTIQUE

Sous-section 1

Population, échantillon & individu

1. Population
2. Échantillon
3. Individu

Terminologie statistique

Population

Une population est un ensemble d'éléments regroupés et observés de manière agrégée selon un critère commun. Il peut s'agir d'êtres vivants ou d'objets matériels ou immatériels.

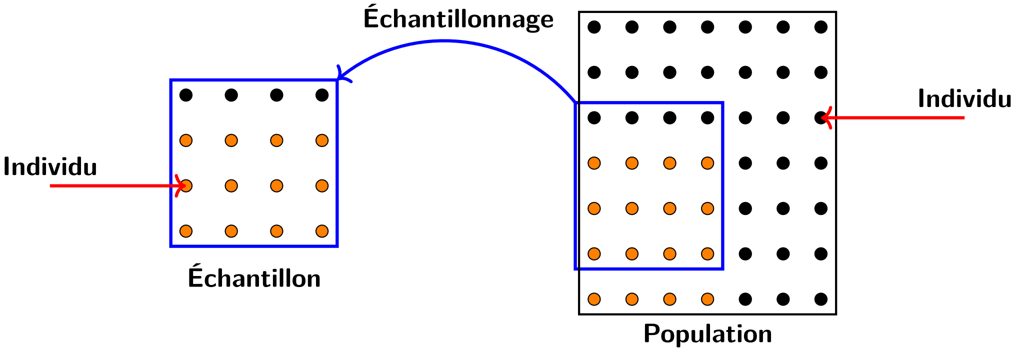
Échantillon

Un échantillon est un ensemble d'éléments sélectionnés dans une population selon une méthode d'échantillonnage bien définie.

Individu

Un individu est un élément d'un échantillon ou d'une population. Un individu est également appelé unité statistique.

Terminologie statistique



Terminologie statistique

Exemple

Une société de télécommunications mène une enquête auprès d'étudiants inscrits dans des universités et des grandes écoles afin d'évaluer leurs besoins en durées d'appels et en volumes de données. La société leur adresse à cet effet un questionnaire comportant des questions sur leur niveau d'études, sur leur type de logement, sur leur besoin mensuel en durées d'appels et sur leur besoin mensuel en volumes de données. La population ciblée par cette enquête est formée d'un ensemble d'étudiants et l'échantillon considéré est composé d'étudiants inscrits dans des universités ou dans des grandes écoles. Ainsi, tout étudiant inscrit dans une université ou dans une grande école constitue un individu de l'échantillon enquêté. Par contre, un étudiant poursuivant ses études supérieures dans d'autres types d'établissements est, certes, un individu de la population cible mais il ne fait pas partie de l'échantillon enquêté.

Terminologie statistique

Sous-section 2

Variable statistique, modalité & donnée statistique

1. Variable statistique
2. Modalité
3. Donnée statistique

Terminologie statistique

Variable statistique

Une variable statistique est le critère selon lequel les individus d'un échantillon ou d'une population sont observés dans une étude statistique. Une variable statistique est désignée par une lettre majuscule, X par exemple.

Modalité

Une modalité est une valeur prise par la variable X observée sur les individus d'un échantillon ou d'une population dans une étude statistique. Les modalités d'une variable statistique X sont indiquées par des lettres minuscules numérotées de 1 à $k \in \mathbb{N}^*$: $x_1, x_2, \cdots, x_k$.

Donnée statistique

Une donnée statistique est l'observation d'une variable statistique X sur un individu d'un échantillon ou d'une population dans une étude statistique. Une donnée statistique est également appelée observation.

Exemple

Le questionnaire adressé par la société de télécommunications aux étudiants inscrits dans des universités et des grandes écoles comporte une question sur leurs besoins mensuels en volumes de données. Les étudiants doivent cocher la case correspondant à leur réponse :

☐ 10 Go ☐ 20 Go ☐ 30 Go ☐ 40 Go ☐ 50 Go

La variable X étudiée dans cette enquête est « le besoin mensuel des étudiants en volumes de données », elle prend 5 modalités : $x_1 = 10$, $x_2 = 20$, $x_3 = 30$, $x_4 = 40$ et $x_5 = 50$. L'un des étudiants questionnés a coché la case 30 Go. Il a de ce fait fourni la donnée statistique suivante : le besoin mensuel de cet étudiant en volumes de données est de 30 Go.

Section 2

Typologie des variables statistiques

1. Variables quantitatives
2. Variables qualitatives

Sous-section 1

Variables quantitatives

1. Variable quantitative
2. Variable quantitative discrète
3. Variable quantitative continue

Typologie des variables statistiques

Variable quantitative

Une variable statistique est quantitative lorsque ses modalités sont des nombres représentant une mesure ou un dénombrement. Les modalités d'une variable quantitative sont appelées valeurs.

Variable quantitative discrète

Une variable quantitative est discrète si elle prend un ensemble fini et dénombrable de valeurs.

Variable quantitative continue

Une variable quantitative est continue si elle prend un ensemble infini de valeurs.

Classe

Les valeurs d'une variable continue sont rangées dans des classes. Une classe est un intervalle semi-fermé $[x_i^-, x_i^+[$ caractérisé par une amplitude a_i et par un centre c_i :

$$a_i = x_i^+ - x_i^- \qquad\qquad c_i = \frac{x_i^- + x_i^+}{2}$$

Typologie des variables statistiques

Exemple

Le questionnaire adressé aux étudiants contient une question sur leur besoin mensuel en volumes de données. Les étudiants doivent choisir parmi les réponses suivantes :

☐ 10 Go ☐ 20 Go ☐ 30 Go ☐ 40 Go ☐ 50 Go

Le besoin mensuel des étudiants en volumes de données est une variable quantitative puisqu'elle mesure le volume de données en gigaoctets dont les étudiants ont besoin chaque mois. Il s'agit d'une variable discrète parce qu'elle prend un ensemble fini et dénombrable de valeurs, en l'occurrence 10, 20, 30, 40 et 50 Go.

Typologie des variables statistiques

Exemple

Le questionnaire contient également une question sur leur besoin mensuel en durées d'appels. Les étudiants doivent cocher la case correspondant à leur réponse :

☐ [0 , 4h[☐ [4h , 8h[☐ [8h , 12h[☐ [12h , 16h[☐ [16h , 20h]

Le besoin mensuel des étudiants en durées d'appels est une variable quantitative puisqu'elle mesure la durée des appels en heures. Il s'agit d'une variable continue parce qu'elle prend un ensemble infini de valeurs allant de 0 à 20h, regroupées en 5 classes d'amplitudes égales à 4h. Le centre de la première classe est égal à 2h et le centre de la dernière classe est égal à 18h.

Sous-section 2

Variables qualitatives

1. Variable qualitative
2. Variable qualitative ordinale
3. Variable qualitative nominale

Typologie des variables statistiques

Variable qualitative

Une variable statistique qui n'est pas quantitative est dite qualitative. Les modalités d'une variable qualitative sont appelées catégories, une variable qualitative est de ce fait une variable catégorielle.

Variable qualitative ordinale

Une variable qualitative est ordinale s'il est possible de trier ses catégories.

Variable qualitative nominale

Une variable qualitative est nominale s'il n'est pas possible de trier ses catégories.

Exemple

Le questionnaire adressé aux étudiants contient une question sur leur niveau d'études. Les étudiants doivent choisir parmi les réponses suivantes :

☐ Bac+2 ☐ Bac+3 ☐ Bac+5 ☐ Bac+8

Le niveau d'études est une variable qualitative puisqu'un niveau d'études ne peut être quantifié. Il s'agit d'une variable ordinale parce que le niveau Bac+8 est supérieur au niveau Bac+5 lui-même supérieur au niveau Bac+3 qui est à son tour supérieur au niveau Bac+2.

Exemple

D'autre part, le questionnaire contient une question sur le type de logement des étudiants. Les étudiants doivent cocher la case correspondant à leur réponse :

☐ Domicile familial ☐ Cité universitaire ☐ Collocation ☐ Autres

Le type de logement est une variable qualitative puisqu'un type de logement ne peut être mesuré. Il s'agit d'une variable nominale parce qu'aucun tri de ses catégories n'est possible.

Section 3

Opérateurs arithmétiques

1. Opérateur d'addition
2. Opérateur de multiplication

Sous-section 1

Opérateur d'addition

1. Définition
2. Propriétés

Opérateurs arithmétiques

Définition

Opérateur d'addition Σ

L'opérateur d'addition Σ permet d'écrire la somme des valeurs $x_1, x_2, \cdots, x_k$ d'une variable quantitative X comme suit :

$$\sum_{i=1}^{k} x_i \equiv x_1 + x_2 + \cdots + x_k$$

où i est une variable muette représentant l'indice de l'addition. On lit : somme des valeurs x_i avec i allant de 1 à k.

Propriétés

Propriété 1

$$\sum_{i=1}^{k} a x_i = a \sum_{i=1}^{k} x_i \ , \ \forall a \in \mathbb{R}$$

Démonstration

$$\sum_{i=1}^{k} a x_i = a x_1 + a x_2 + \cdots + a x_k$$

$$= a \underbrace{\left(x_1 + x_2 + \cdots + x_k\right)}_{\sum_{i=1}^{k} x_i} = a \sum_{i=1}^{k} x_i$$

Opérateurs arithmétiques

Propriété 2

$$\sum_{i=1}^{k}(x_i + y_i) = \sum_{i=1}^{k} x_i + \sum_{i=1}^{k} y_i$$

Démonstration

$$\begin{aligned}\sum_{i=1}^{k}(x_i + y_i) &= (x_1 + y_1) + (x_2 + y_2) + \cdots + (x_k + y_k) \\ &= \underbrace{(x_1 + x_2 + \cdots + x_k)}_{\sum_{i=1}^{k} x_i} + \underbrace{(y_1 + y_2 + \cdots + y_k)}_{\sum_{i=1}^{k} y_i} \\ &= \sum_{i=1}^{k} x_i + \sum_{i=1}^{k} y_i\end{aligned}$$

Opérateurs arithmétiques

Propriété 3

$$\sum_{i=1}^{k} x_i = \sum_{i=1}^{h} x_i + \sum_{i=h+1}^{k} x_i \ , \ \exists h \in \mathbb{N}^* \mid 1 < h < k$$

Démonstration

$$\begin{aligned}
\sum_{i=1}^{k} x_i &= x_1 + x_2 + \cdots + x_h + x_{h+1} + x_{h+2} + \cdots + x_k \\
&= \underbrace{(x_1 + x_2 + \cdots + x_h)}_{\sum_{i=1}^{h} x_i} + \underbrace{(x_{h+1} + x_{h+2} + \cdots + x_k)}_{\sum_{i=h+1}^{k} x_i} \\
&= \sum_{i=1}^{h} x_i + \sum_{i=h+1}^{k} x_i
\end{aligned}$$

Propriété 4

$$\sum_{i=1}^{k} a = k.a \ , \ \forall a \in \mathbb{R}$$

Démonstration

$$\begin{aligned}\sum_{i=1}^{k} a &= \underbrace{a + a + \cdots + a}_{k \text{ fois}} \\ &= k.a\end{aligned}$$

Sous-section 2

Opérateur de multiplication

1. Définition
2. Propriétés

Définition

Opérateur de multiplication Π

L'opérateur de multiplication Π permet d'écrire le produit des valeurs $x_1, x_2, \cdots, x_k$ d'une variable quantitative X comme suit :

$$\prod_{i=1}^{k} x_i \equiv x_1 \times x_2 \times \cdots \times x_k$$

où i est une variable muette représentant l'indice de la multiplication. On lit : produit des valeurs x_i avec i allant de 1 à k.

Opérateurs arithmétiques

Propriétés

Propriété 1

$$\prod_{i=1}^{k} a = a^k \ , \ \forall a \in \mathbb{R}$$

Démonstration

$$\prod_{i=1}^{k} a = \underbrace{a \times a \times \cdots \times a}_{k \text{ facteurs}}$$
$$= a^k$$

Opérateurs arithmétiques

Propriété 2

$$\prod_{i=1}^{k}(x_i \times y_i) = \prod_{i=1}^{k} x_i \times \prod_{i=1}^{k} y_i$$

Démonstration

$$\begin{aligned}\prod_{i=1}^{k}(x_i \times y_i) &= (x_1 \times y_1) \times (x_2 \times y_2) \times \cdots \times (x_k \times y_k) \\ &= \underbrace{(x_1 \times x_2 \times \cdots \times x_k)}_{\prod_{i=1}^{k} x_i} \times \underbrace{(y_1 \times y_2 \times \cdots \times y_k)}_{\prod_{i=1}^{k} y_i} \\ &= \prod_{i=1}^{k} x_i \times \prod_{i=1}^{k} y_i\end{aligned}$$

Opérateurs arithmétiques

Propriété 3

$$\prod_{i=1}^{k} x_i = \prod_{i=1}^{h} x_i \times \prod_{i=h+1}^{k} x_i \ , \ \exists h \in \mathbb{N}^* \mid 1 < h < k$$

Démonstration

$$\begin{aligned}
\prod_{i=1}^{k} x_i &= x_1 \times x_2 \times \cdots \times x_h \times x_{h+1} \times x_{h+2} \times \cdots \times x_k \\
&= \underbrace{(x_1 \times x_2 \times \cdots \times x_h)}_{\prod_{i=1}^{h} x_i} \times \underbrace{(x_{h+1} \times x_{h+2} \times \cdots \times x_k)}_{\prod_{i=h+1}^{k} x_i} \\
&= \prod_{i=1}^{h} x_i \times \prod_{i=h+1}^{k} x_i
\end{aligned}$$

Opérateurs arithmétiques

Propriété 4

$$\prod_{i=1}^{k} a x_i = a^k \prod_{i=1}^{k} x_i \ , \ \forall a \in \mathbb{R}$$

Démonstration

$$\begin{aligned}
\prod_{i=1}^{k} a x_i &= a x_1 \times a x_2 \times \cdots \times a x_k \\
&= \underbrace{(a \times a \times \cdots \times a)}_{k \text{ facteurs}} \times \underbrace{(x_1 \times x_2 \times \cdots \times x_k)}_{\prod_{i=1}^{k} x_i} \\
&= a^k \prod_{i=1}^{k} x_i
\end{aligned}$$

Chapitre 2

Statistique descriptive univariée

1. Définitions
2. Représentations graphiques
3. Indicateurs statistiques

Section 1

Définitions

1. Effectif, effectif total & fréquence
2. Effectifs et fréquences cumulés
3. Séries et distributions statistiques
4. Tableau statistique

Sous-section 1

Effectif, effectif total & fréquence

1. Effectif
2. Effectif total
3. Fréquence

Définitions

Effectif, effectif total & fréquence

Effectif

L'effectif n_i d'une modalité x_i est le nombre d'individus d'un échantillon ou d'une population caractérisés par la modalité x_i. L'effectif n_i est également appelé fréquence absolue.

Effectif total

L'effectif total n d'un échantillon ou d'une population est le nombre d'individus qui le (la) composent. Il est égal à la somme des effectifs n_i des modalités x_i : $n_1 + n_2 + \cdots + n_k = n$.

Fréquence

La fréquence f_i d'une modalité x_i est la proportion d'individus d'un échantillon ou d'une population caractérisés par la modalité x_i. Elle est également appelée fréquence relative. La fréquence f_i d'une modalité x_i est donnée par le rapport entre l'effectif n_i de la modalité x_i et l'effectif total n de l'échantillon ou de la population selon la formule suivante : $f_i = \frac{n_i}{n}$.

Une fréquence f_i s'exprime sous forme d'une fraction, d'un nombre décimal ou d'un pourcentage. Elle est comprise entre 0 et 1 (ou entre 0 et 100 lorsqu'elle est exprimée en pourcentage) et la somme des fréquences est égale à 1 car :

$$f_1 + f_2 + \cdots + f_{k-1} + f_k = \frac{n_1}{n} + \frac{n_2}{n} + \cdots + \frac{n_{k-1}}{n} + \frac{n_k}{n} = \frac{n}{n} = 1$$

Dans la pratique, pour éviter que la somme des fréquences soit différente de 1 à cause de problèmes d'arrondis, on exprime la dernière fréquence f_k en fonction des autres fréquences $f_1, f_2, \cdots, f_{k-1}$:

$$f_1 + f_2 + \cdots + f_{k-1} + f_k = 1$$
$$f_k = 1 - f_1 - f_2 - \cdots - f_{k-1}$$

Exemple

Le dépouillement des questionnaires adressés aux étudiants d'une grande école a permis de relever leurs besoins mensuels en volumes de données :

10	40	30	50	10	40	20	30	20	10
20	50	10	40	30	10	30	40	30	10
40	50	30	20	40	50	10	40	30	50
10	50	20	50	30	50	30	20	40	10
30	50	50	10	40	40	50	40	20	50
30	20	40	10	50	50	50	20	50	30
40	20	50	20	50	30	50	40	30	40
10	20	10	20	30	30	40	40	20	20
20	30	10	50	20	50	50	40	50	10

Définitions

Effectif, effectif total & fréquence

La valeur $x_1 = 10$ est observée 15 fois dans cet échantillon, son effectif est donc $n_1 = 15$. De la même manière, les valeurs $x_2 = 20$ et $x_3 = 30$ sont observées 17 fois chacune, leurs effectifs respectifs sont $n_2 = n_3 = 17$, la valeur $x_4 = 40$ est observée 18 fois et la dernière valeur $x_5 = 50$ est observée 23 fois. L'effectif total n est donné par la somme des effectifs n_i de toutes les valeurs x_i soit un total de $n = 90$ individus dans cet échantillon.

Par ailleurs, la fréquence f_i de la modalité x_i est donnée par le rapport entre son effectif n_i et l'effectif total n, on obtient les fréquences suivantes : $f_1 = 15/90$, $f_2 = f_3 = 17/90$, $f_4 = 18/90$ et $f_5 = 23/90$.

Les résultats précédents sont synthétisés dans le tableau suivant :

x_i	n_i	$f_i = n_i/n$	Approximations	Corrections	Pourcentages
10	15	15/90	0.17	0.17	17%
20	17	17/90	0.19	0.19	19%
30	17	17/90	0.19	0.19	19%
40	18	18/90	0.20	0.20	20%
50	23	23/90	0.26	0.25	25%
Σ	$n = 90$	1	1.01	1	100%

Sous-section 2

Effectifs et fréquences cumulés

1. Effectifs cumulés
2. Fréquences cumulées

Effectifs cumulés

Effectif cumulé croissant

L'effectif cumulé croissant N_i^+ d'une modalité x_i est le nombre d'individus caractérisés par des modalités au plus égales à la modalité x_i. Il est calculé selon la formule suivante :

$$N_i^+ = n_1 + n_2 + \cdots + n_i$$

Effectif cumulé décroissant

L'effectif cumulé décroissant N_i^- d'une modalité x_i est le nombre d'individus caractérisés par des modalités au moins égales à la modalité x_i. Il est calculé selon la formule suivante :

$$N_i^- = n_k + n_{k-1} + \cdots + n_i$$

Exemple

Les effectifs cumulés croissants N_i^+ et les effectifs cumulés décroissants N_i^- des besoins mensuels des étudiants en volumes de données x_i sont calculés dans le tableau suivant :

x_i	n_i		N_i^+	n_i		N_i^-
10 $\downarrow$	15 $\downarrow$	$\longrightarrow$	15	15 $\uparrow$	$\longrightarrow$	90
	$+$	$\swarrow$		$+$	$\nwarrow$	
20 $\downarrow$	17 $\downarrow$	$\longrightarrow$	32	17 $\uparrow$	$\longrightarrow$	75
	$+$	$\swarrow$		$+$	$\nwarrow$	
30 $\downarrow$	17 $\downarrow$	$\longrightarrow$	49	17 $\uparrow$	$\longrightarrow$	58
	$+$	$\swarrow$		$+$	$\nwarrow$	
40 $\downarrow$	18 $\downarrow$	$\longrightarrow$	67	18 $\uparrow$	$\longrightarrow$	41
	$+$	$\swarrow$		$+$	$\nwarrow$	
50 $\downarrow$	23 $\downarrow$	$\longrightarrow$	90	23 $\uparrow$	$\longrightarrow$	23
Σ	$n = 90$		–	$n = 90$		–

L'effectif cumulé croissant N_i^+ de la modalité x_i est obtenu en cumulant les effectifs n_i de manière croissante en partant de l'effectif $n_1 = 15$ de la première modalité $x_1 = 10$ jusqu'à arriver à l'effectif n_i de la modalité x_i dont on veut calculer l'effectif cumulé croissant N_i^+. En revanche, l'effectif cumulé décroissant N_i^- de la modalité x_i est obtenu en cumulant les effectifs n_i de manière décroissante en partant de l'effectif $n_5 = 23$ de la dernière modalité $x_5 = 50$ jusqu'à arriver à l'effectif n_i de la modalité x_i dont on veut calculer l'effectif cumulé décroissant N_i^-.

Par ailleurs, notons que, d'une part, l'effectif cumulé croissant $N_1^+ = 15$ de la première modalité $x_1 = 10$ est égal à l'effectif $n_1 = 15$ de cette première modalité et que l'effectif cumulé croissant $N_5^+ = 90$ de la dernière modalité $x_5 = 50$ est égal à l'effectif total $n = 90$. Notons, d'autre part, que l'effectif cumulé décroissant $N_1^- = 90$ de la première modalité $x_1 = 10$ est égal à l'effectif total $n = 90$ et que l'effectif cumulé décroissant $N_5^- = 23$ de la dernière modalité $x_5 = 50$ est égal à l'effectif $n_5 = 23$ de cette dernière modalité.

Fréquences cumulées

Fréquence cumulée croissante

La fréquence cumulée croissante F_i^+ d'une modalité x_i est la proportion d'individus caractérisés par des modalités au plus égales à la modalité x_i. Elle est calculée selon la formule suivante :

$$F_i^+ = f_1 + f_2 + \cdots + f_i$$

Fréquence cumulée décroissante

La fréquence cumulée décroissante F_i^- d'une modalité x_i est la proportion d'individus caractérisés par des modalités au moins égales à la modalité x_i. Elle est calculée selon la formule suivante :

$$F_i^- = f_k + f_{k-1} + \cdots + f_i$$

Exemple

Les fréquences cumulées croissantes F_i^+ et les fréquences cumulées décroissantes F_i^- des besoins mensuels des étudiants en volumes de données x_i sont calculées dans le tableau suivant :

x_i	f_i		F_i^+	f_i		F_i^-
10 $\downarrow$	0.17 $\downarrow$	$\longrightarrow$	0.17	0.17 $\uparrow$	$\longrightarrow$	1
	$+$	$\swarrow$		$+$	$\nwarrow$	
20 $\downarrow$	0.19 $\downarrow$	$\longrightarrow$	0.36	0.19 $\uparrow$	$\longrightarrow$	0.83
	$+$	$\swarrow$		$+$	$\nwarrow$	
30 $\downarrow$	0.19 $\downarrow$	$\longrightarrow$	0.55	0.19 $\uparrow$	$\longrightarrow$	0.64
	$+$	$\swarrow$		$+$	$\nwarrow$	
40 $\downarrow$	0.20 $\downarrow$	$\longrightarrow$	0.75	0.20 $\uparrow$	$\longrightarrow$	0.45
	$+$	$\swarrow$		$+$	$\nwarrow$	
50 $\downarrow$	0.25 $\downarrow$	$\longrightarrow$	1	0.25 $\uparrow$	$\longrightarrow$	0.25
Σ	1		–	1		–

La fréquence cumulée croissante F_i^+ de la modalité x_i est obtenue en cumulant les fréquences f_i de manière croissante en partant de la fréquence $f_1 = 0.17$ de la première modalité $x_1 = 10$ jusqu'à arriver à la fréquence f_i de la modalité x_i dont on veut calculer la fréquence cumulée croissante F_i^+. En revanche, la fréquence cumulée décroissante F_i^- de la modalité x_i est obtenue en cumulant les fréquences f_i de manière décroissante en partant de la fréquence $f_5 = 0.25$ de la dernière modalité $x_5 = 50$ jusqu'à arriver à la fréquence f_i de la modalité x_i dont on veut calculer la fréquence cumulée décroissante F_i^-.

Par ailleurs, notons que, d'une part, la fréquence cumulée croissante $F_1^+ = 0.17$ de la première modalité $x_1 = 10$ est égal à la fréquence $f_1 = 0.17$ de cette première modalité et que la fréquence cumulée croissante $F_5^+ = 1$ de la dernière modalité $x_5 = 50$ est égale au total des fréquences 1. Notons, d'autre part, que la fréquence cumulée décroissante $F_1^- = 1$ de la première modalité $x_1 = 10$ est égale au total des fréquences 1 et que la fréquence cumulée décroissante $F_5^- = 0.25$ de la dernière modalité $x_5 = 50$ est égale à la fréquence $f_5 = 0.25$ de cette dernière modalité.

Remarque

L'expression « au plus » signifie « inférieur ou égal » et l'expression « au moins » signifie « supérieur ou égal », cela sous-entend que les modalités sont triées. Par conséquent, les effectifs et les fréquences cumulés ne peuvent être calculés que dans le cas d'une variable quantitative (discrète ou continue), dans le cas d'une variable qualitative ordinale mais pas dans le cas d'une variable qualitative nominale.

Sous-section 3

Séries et distributions statistiques

1. Séries statistiques
2. Distributions statistiques

Séries statistiques

Série statistique

Une série statistique est une liste de modalités correspondant à l'observation d'une variable X sur les individus d'un échantillon ou d'une population dans une étude statistique.

Série statistique brute

Une série statistique est dite brute si les modalités ne sont pas triées.

Série statistique ordonnée

Une série statistique est dite ordonnée si les modalités sont triées.

Exemple

Le dépouillement des questionnaires adressés aux étudiants d'une grande école a permis de relever la série statistique brute de leurs besoins mensuels en volumes de données :

10	40	30	50	10	40	20	30	20	10
20	50	10	40	30	10	30	40	30	10
40	50	30	20	40	50	10	40	30	50
10	50	20	50	30	50	30	20	40	10
30	50	50	10	40	40	50	40	20	50
30	20	40	10	50	50	50	20	50	30
40	20	50	20	50	30	50	40	30	40
10	20	10	20	30	30	40	40	20	20
20	30	10	50	20	50	50	40	50	10

Exemple

Présentée sous une forme brute, nulle information n'est à déceler sans un traitement préalable des données. Le tri croissant de la série statistique brute conduit à la série statistique ordonnée suivante :

10	10	10	10	10	10	10	10	10	10
10	10	10	10	10	20	20	20	20	20
20	20	20	20	20	20	20	20	20	20
20	20	30	30	30	30	30	30	30	30
30	30	30	30	30	30	30	30	30	40
40	40	40	40	40	40	40	40	40	40
40	40	40	40	40	40	40	50	50	50
50	50	50	50	50	50	50	50	50	50
50	50	50	50	50	50	50	50	50	50

Le passage d'une série brute à une série ordonnée fournit au moins trois informations, à savoir : le minimum $x_{min} = 10$Go, le maximum $x_{max} = 50$Go et l'étendue de la série $x_{max} - x_{min} = 40$Go.

Distributions statistiques

Distribution statistique des effectifs

Une distribution statistique des effectifs est une association (x_i, n_i) entre les modalités x_i d'une variable X et les effectifs n_i correspondants.

Distribution statistique des fréquences

Une distribution statistique des fréquences est une association (x_i, f_i) entre les modalités x_i d'une variable X et les fréquences f_i correspondantes.

Remarque

La distribution des effectifs et la distribution des fréquences donnent, à un facteur près, la même information. Néanmoins, la distribution des fréquences est à privilégier si l'objectif de l'étude statistique est de comparer les observations d'une variable sur deux ou plusieurs échantillons.

Exemple

Distribution statistique des effectifs des besoins mensuels en volumes de données des étudiants d'une grande école :

x_i	10	20	30	40	50
n_i	15	17	17	18	23

Exemple

Distribution statistique des fréquences des besoins mensuels en volumes de données des étudiants d'une grande école :

x_i	10	20	30	40	50
f_i	0.17	0.19	0.19	0.20	0.25

Sous-section 4

Tableau statistique

1. Définition
2. Interprétation

Définition

Tableau statistique

Le *tableau statistique* est une représentation synthétique d'une distribution statistique. Il regroupe les modalités x_i, leurs effectifs n_i, leurs fréquences f_i, leurs effectifs cumulés N_i et leurs fréquences cumulées F_i.

x_i	n_i	f_i	N_i^-	N_i^+	F_i^-	F_i^+
x_1	n_1	f_1	$N_1^- = n$	$N_1^+ = n_1$	$F_1^- = 1$	$F_1^+ = f_1$
x_2	n_2	f_2	N_2^-	N_2^+	F_2^-	F_2^+
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_k	n_k	f_k	$N_k^- = n_k$	$N_k^+ = n$	$F_k^- = f_k$	$F_k^+ = 1$
Σ	n	1	–	–	–	–

Interprétation

- Le nombre d'individus caractérisés par x_1 est n_1 ;
- La proportion d'individus caractérisés par x_2 est f_2 ;
- Le nombre d'individus caractérisés par au moins x_1 est N_1^- ;
- Le nombre d'individus caractérisés par au plus x_2 est N_2^+ ;
- La proportion d'individus caractérisés par au moins x_1 est F_1^- ;
- La proportion d'individus caractérisés par au plus x_2 est F_2^+.

Remarque

Dans le cas d'une variable continue, les effectifs cumulés croissants et les fréquences cumulées croissantes sont associés aux bornes supérieures des classes. En revanche, les effectifs cumulés décroissants et les fréquences cumulées décroissantes sont associés à leurs bornes inférieures.

Exemple : Besoin mensuel des étudiants en volumes de données

Le dépouillement des questionnaires adressés aux étudiants d'une grande école a permis de relever leurs besoins mensuels en volumes de données :

Volume de données (Go)	n_i	f_i	N_i^-	N_i^+	F_i^-	F_i^+
10	15	0.17	90	15	1	0.17
20	17	0.19	75	32	0.83	0.36
30	17	0.19	58	49	0.64	0.55
40	18	0.20	41	67	0.45	0.75
50	23	0.25	23	90	0.25	1
Σ	90	1	–	–	–	–

Exemple : Besoin mensuel des étudiants en volumes de données

- Le nombre d'étudiants ayant besoin de 20 Go de données par mois est $n_2 = 17$;
- La proportion d'étudiants ayant besoin de 30 Go de données par mois est $f_3 = 0.19$;
- Le nombre d'étudiants ayant besoin d'au moins 20 Go de données par mois est $N_2^- = 75$;
- Le nombre d'étudiants ayant besoin d'au plus 30 Go de données par mois est $N_3^+ = 49$;
- La proportion d'étudiants ayant besoin d'au moins 10 Go de données par mois est $F_1^- = 1$;
- La proportion d'étudiants ayant besoin d'au plus 30 Go de données par mois est $F_3^+ = 0.55$.

Exemple : Besoin mensuel des étudiants en durées d'appels

Le dépouillement des questionnaires adressés aux étudiants d'une grande école a permis de relever leurs besoins mensuels en durées d'appels :

Durée d'appels (h)	n_i	f_i	N_i^-	N_i^+	F_i^-	F_i^+
$[0;4[$	20	0.22	90	20	1	0.22
$[4;8[$	24	0.27	70	44	0.78	0.49
$[8;12[$	34	0.38	46	78	0.51	0.87
$[12;16[$	7	0.07	12	85	0.13	0.94
$[16;20]$	5	0.06	5	90	0.06	1
Σ	90	1	–	–	–	–

Exemple : Besoin mensuel des étudiants en durées d'appels

- Le nombre d'étudiants ayant besoin de 4 à 8 heures d'appels par mois est $n_2 = 24$;
- La proportion d'étudiants ayant besoin de 8 à 12 heures d'appels par mois est $f_3 = 0.38$;
- Le nombre d'étudiants ayant besoin d'au moins 8 heures d'appels par mois est $N_3^- = 46$;
- Le nombre d'étudiants ayant besoin d'au plus 8 heures d'appels par mois est $N_2^+ = 44$;
- La proportion d'étudiants ayant besoin d'au moins 0 heures d'appels par mois est $F_1^- = 1$;
- La proportion d'étudiants ayant besoin d'au plus 12 heures d'appels par mois est $F_3^+ = 0.87$.

Exemple : Niveau d'études des étudiants

Le dépouillement des questionnaires adressés aux étudiants d'une grande école a permis de relever leurs niveaux d'études :

Niveau d'études	n_i	f_i	N_i^-	N_i^+	F_i^-	F_i^+
Bac+2	40	0.44	90	40	1	0.44
Bac+3	25	0.28	50	65	0.56	0.72
Bac+5	20	0.22	25	85	0.28	0.94
Bac+8	05	0.06	5	90	0.06	1
Σ	90	1	–	–	–	–

Exemple : Niveau d'études des étudiants

- Le nombre d'étudiants ayant un niveau d'études équivalent à Bac+3 est $n_2 = 25$;
- La proportion d'étudiants ayant un niveau d'études équivalent à Bac+5 est $f_3 = 0.22$;
- Le nombre d'étudiants ayant un niveau d'études au moins égal à Bac+3 est $N_2^- = 50$;
- Le nombre d'étudiants ayant un niveau d'études au plus égal à Bac+5 est $N_3^+ = 85$;
- La proportion d'étudiants ayant un niveau d'études au moins égal à Bac+3 est $F_2^- = 0.56$;
- La proportion d'étudiants ayant un niveau d'études au plus égal à Bac+5 est $F_3^+ = 0.94$.

Exemple : Type de logement des étudiants

Le dépouillement des questionnaires adressés aux étudiants d'une grande école a permis de relever le type de logement qu'ils occupent :

Type de logement	n_i	f_i
Domicile familial	50	0.55
Cité universitaire	20	0.22
Collocation	15	0.17
Autres	05	0.06
Σ	90	1

Exemple : Type de logement des étudiants

- Le nombre d'étudiants résidant en cité universitaire est $n_2 = 20$;
- La proportion d'étudiants résidant au domicile familial est $f_1 = 0.55$.

Remarque

Le type de logement étant une variable qualitative nominale, seuls sont calculés les effectifs et les fréquences des modalités. Les effectifs et les fréquences cumulés ne sont pas calculés parce qu'il n'est pas possible d'ordonner les catégories d'une variable qualitative nominale.

Section 2

Représentations graphiques

1. Diagramme en barres
2. Diagramme en bâtons
3. Histogramme
4. Courbes cumulatives

Diagramme en barres

Diagramme en barres

Le diagramme en barres est un graphique qui associe à chaque catégorie d'une variable qualitative un rectangle de hauteur égale à son effectif ou à sa fréquence.

Utilisation

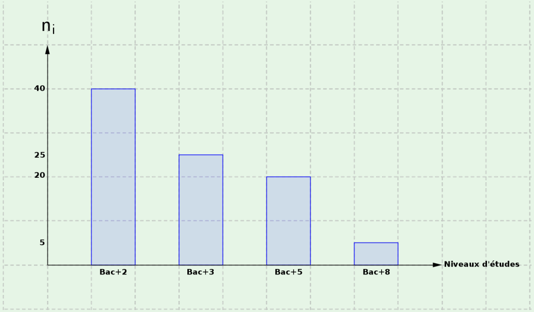
Représentation graphique des effectifs et des fréquences d'une variable qualitative.

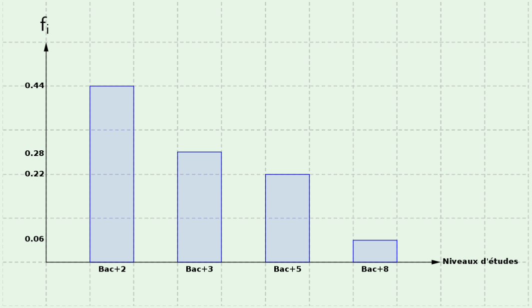
Remarques

- L'ordre des catégories d'une variable qualitative ordinale doit être respecté lors de la construction du diagramme en barres ;
- Les rectangles d'un diagramme en barres sont disjoints.

Exemple : Niveau d'études d'étudiants inscrits dans une grande école

Niveau d'études	Bac+2	Bac+3	Bac+5	Bac+8
n_i	40	25	20	5
f_i	0.44	0.28	0.22	0.06





Exemple : Type de logement d'étudiants inscrits dans une grande école

Type de logement	Domicile familial	Cité universitaire	Collocation	Autres
n_i	50	20	15	5
f_i	0.55	0.22	0.17	0.06

Diagramme en bâtons

Diagramme en bâtons

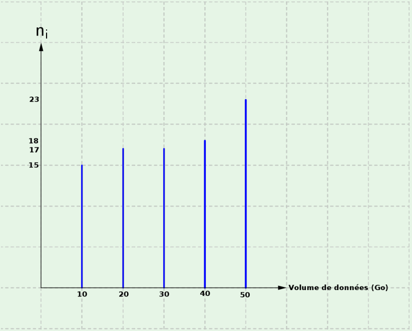
Le diagramme en bâtons est un graphique qui associe à chaque valeur d'une variable quantitative discrète un segment de longueur égale à son effectif ou à sa fréquence.

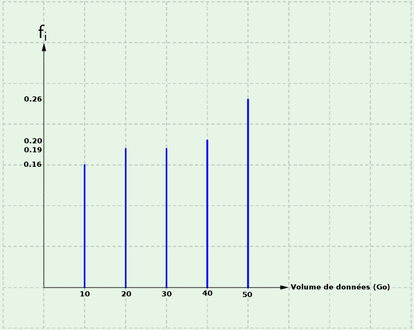
Utilisation

Représentation graphique des effectifs et des fréquences d'une variable quantitative discrète.

Exemple : Besoin mensuel en volumes de données d'étudiants inscrits dans une grande école

Volumes de données (Go)	10	20	30	40	50
n_i	15	17	17	18	23
f_i	0.16	0.19	0.19	0.20	0.26





Histogramme

Histogramme

L'histogramme est un graphique qui associe à chaque classe d'une variable quantitative continue un rectangle d'aire égale à son effectif (ou à sa fréquence). La largeur du rectangle est égale à l'amplitude de la classe et sa hauteur est égale à sa densité d'effectif (ou de fréquence).

Densité d'effectif et densité de fréquence

La densité h_i associée à une classe $[x_i^-; x_i^+[$ est donnée par le rapport entre son effectif n_i (ou sa fréquence f_i) et son amplitude a_i, c'est à dire $h_i = n_i/a_i$ (ou $h_i = f_i/a_i$).

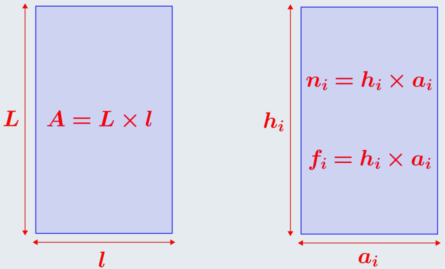
Utilisation

Représentation graphique des effectifs et des fréquences d'une variable quantitative continue.

Représentations graphiques

Histogramme

Rappel : Aire d'un rectangle

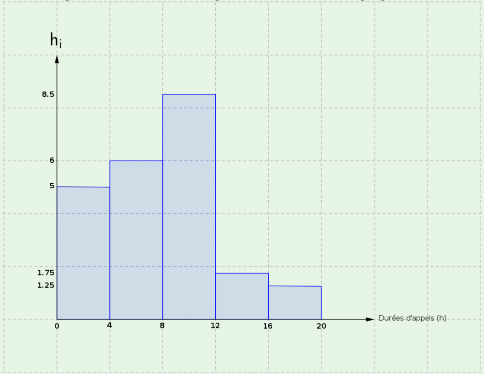


Expression d'une densité

$$h_i = \frac{n_i}{a_i} \qquad h_i = \frac{f_i}{a_i}$$

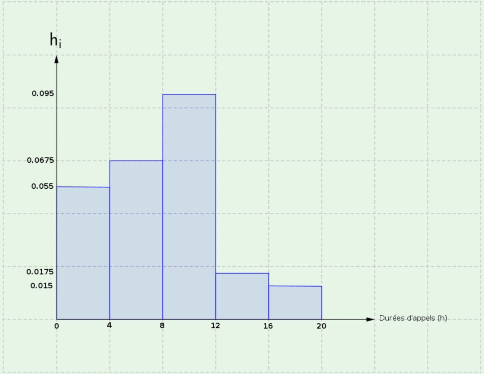
Exemple : Besoin mensuel en durées d'appels d'étudiants inscrits dans une grande école

Durées d'appels (h)	[0 ;4[	[4 ;8[	[8 ;12[	[12 ;16[	[16 ;20]
n_i	20	24	34	7	5
$h_i = n_i/a_i$	5	6	8.5	1.75	1.25



Exemple : Besoin mensuel en durées d'appels d'étudiants inscrits dans une grande école

Durées d'appels (h)	[0 ;4[	[4 ;8[	[8 ;12[	[12 ;16[	[16 ;20]
f_i	0.22	0.27	0.38	0.07	0.06
$h_i = f_i/a_i$	0.055	0.0675	0.095	0.0175	0.015



Courbes cumulatives

Courbes cumulatives

La courbe cumulative est un graphique qui associe à chaque modalité d'une variable quantitative son effectif cumulé ou sa fréquence cumulée. Elle est dite ascendante dans le cas des effectifs et des fréquences cumulés croissants, et descendante dans le cas contraire.

Utilisation

Représentation graphique des effectifs et des fréquences cumulés d'une variable quantitative.

Logique de construction

x_i	n_i	f_i	N_i^-	N_i^+	F_i^-	F_i^+
10	15	0.16	90	15	1	0.16
20	17	0.19	75	32	0.84	0.35
30	17	0.19	58	49	0.65	0.54
40	18	0.20	41	67	0.46	0.74
50	23	0.26	23	90	0.26	1

Logique de construction

x_i	n_i	f_i	N_i^-	N_i^+	F_i^-	F_i^+
$-\infty$	0	0	90	0	1	0
$\uparrow$	0	0	90	0	1	0
10	15	0.16	90	15	1	0.16
$\cdots$	0	0	75	15	0.84	0.16
$\cdots$	0	0	75	15	0.84	0.16
20	17	0.19	75	32	0.84	0.35
$\cdots$	0	0	58	32	0.65	0.35
$\cdots$	0	0	58	32	0.65	0.35
30	17	0.19	58	49	0.65	0.54
$\cdots$	0	0	41	49	0.46	0.54
$\cdots$	0	0	41	49	0.46	0.54
40	18	0.20	41	67	0.46	0.74
$\cdots$	0	0	23	67	0.26	0.74
$\cdots$	0	0	23	67	0.26	0.74
50	23	0.26	23	90	0.26	1
$\downarrow$	0	0	0	90	0	1
$+\infty$	0	0	0	90	0	1

Logique de construction

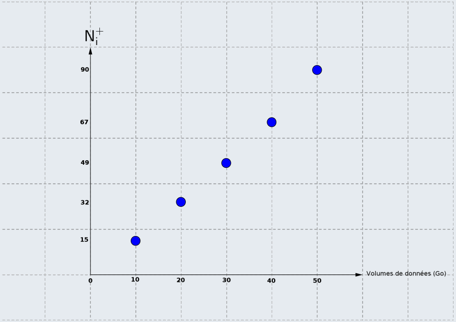
1. Imaginer toutes les valeurs $x_i \in \mathbb{R}$ pour la variable X, y compris celles démunies de sens ;
2. Attribuer des effectifs nuls aux valeurs ne faisant pas partie des modalités de X ;
3. Compléter le tableau statistique.

Constatations

- Les effectifs et les fréquences cumulés sont constants entre deux modalités consécutives ;
- Avant la première modalité de X, les effectifs et les fréquences cumulés croissants sont nuls ;
- Après la dernière modalité de X, les effectifs et les fréquences cumulés décroissants sont nuls ;
- Avant la première modalité de X, les effectifs cumulés décroissants sont égaux à n ;
- Avant la première modalité de X, les fréquences cumulées décroissantes sont égales à 1 ;
- Après la dernière modalité de X, les effectifs cumulés croissants sont égaux à n ;
- Après la dernière modalité de X, les fréquences cumulées croissantes sont égales à 1.

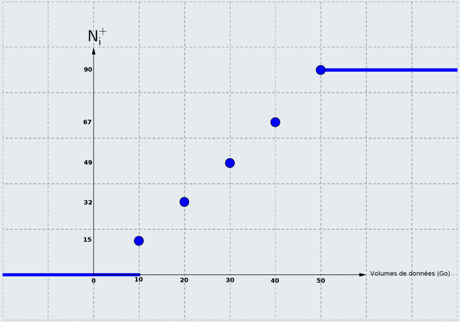
Étape 1 : Positionner les points de coordonnées (x_i, N_i^+)

x_i	10	20	30	40	50
N_i^+	15	32	49	67	90



Étape 2 : Tracer les demi-droites extrêmes

x_i	10	20	30	40	50
N_i^+	15	32	49	67	90

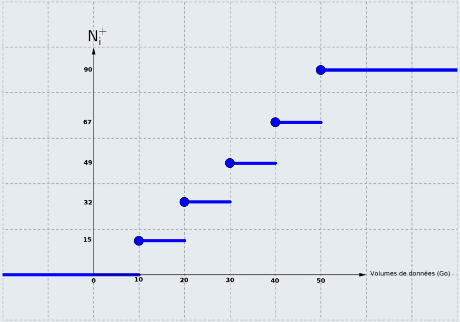


Représentations graphiques

Courbes cumulatives d'une variable discrète : Courbe cumulative ascendante des effectifs

Étape 3 : Tracer les segments intermédiaires

x_i	10	20	30	40	50
N_i^+	15	32	49	67	90



N_i^+ est le **nombre** d'individus qui ont **au plus** la modalité x_i.

Remarque

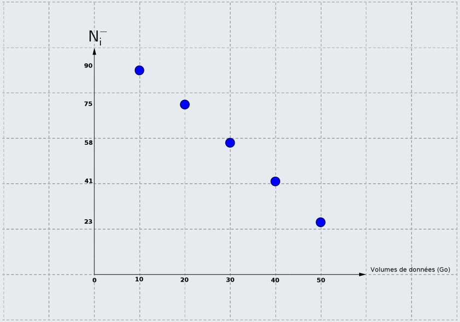
La courbe cumulative ascendante des effectifs d'une variable discrète est la représentation graphique d'une fonction càdlàg (continue à droite, limite à gauche) définie sur $\mathbb{R}$, non-décroissante et à valeurs dans $[0; n]$.

Représentations graphiques

Courbes cumulatives d'une variable discrète : Courbe cumulative descendante des effectifs

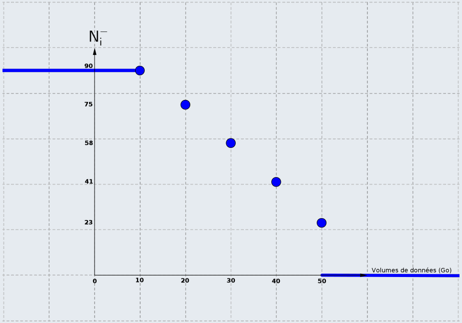
Étape 1 : Positionner les points de coordonnées (x_i, N_i^-)

x_i	10	20	30	40	50
N_i^-	90	75	58	41	23



Étape 2 : Tracer les demi-droites extrêmes

x_i	10	20	30	40	50
N_i^-	90	75	58	41	23

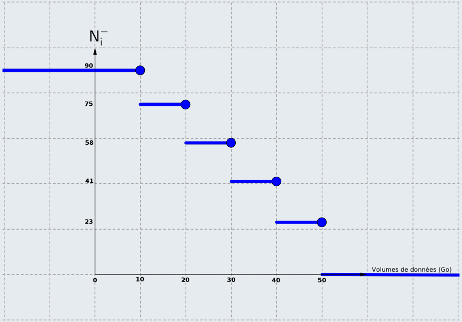


Représentations graphiques

Courbes cumulatives d'une variable discrète : Courbe cumulative descendante des effectifs

Étape 3 : Tracer les segments intermédiaires

x_i	10	20	30	40	50
N_i^-	90	75	58	41	23



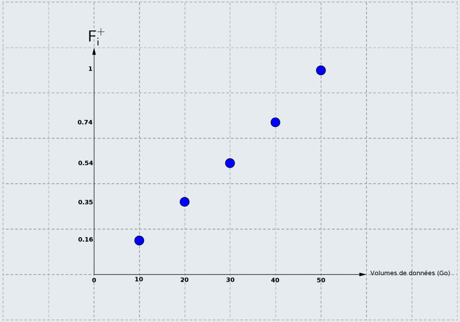
N_i^- est le **nombre** d'individus qui ont **au moins** la modalité x_i.

Remarque

La courbe cumulative descendante des effectifs d'une variable discrète est la représentation graphique d'une fonction càglàd (continue à gauche, limite à droite) définie sur $\mathbb{R}$, non-croissante et à valeurs dans $[0; n]$.

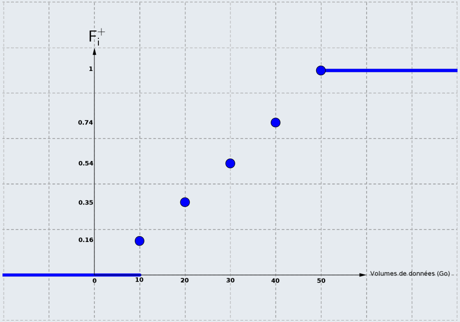
Étape 1 : Positionner les points de coordonnées (x_i, F_i^+)

x_i	10	20	30	40	50
F_i^+	0.16	0.35	0.54	0.74	1



Étape 2 : Tracer les demi-droites extrêmes

x_i	10	20	30	40	50
F_i^+	0.16	0.35	0.54	0.74	1

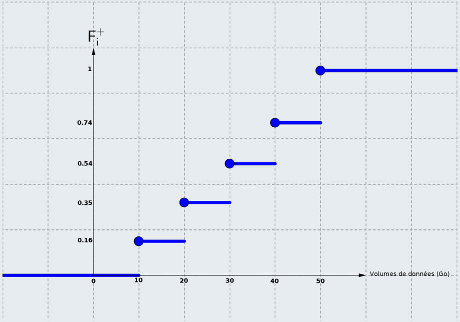


Représentations graphiques

Courbes cumulatives d'une variable discrète : Courbe cumulative ascendante des fréquences

Étape 3 : Tracer les segments intermédiaires

x_i	10	20	30	40	50
F_i^+	0.16	0.35	0.54	0.74	1



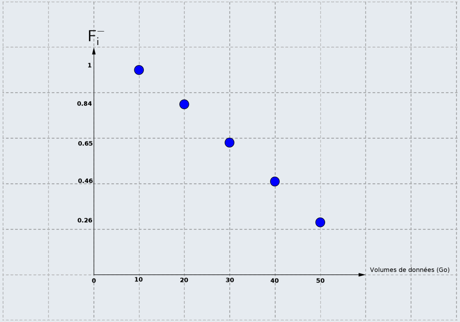
F_i^+ est la **proportion** d'individus qui ont **au plus** la modalité x_i.

Remarque

La courbe cumulative ascendante des fréquences d'une variable discrète est la représentation graphique d'une fonction càdlàg (continue à droite, limite à gauche) définie sur $\mathbb{R}$, non-décroissante et à valeurs dans $[0; 1]$.

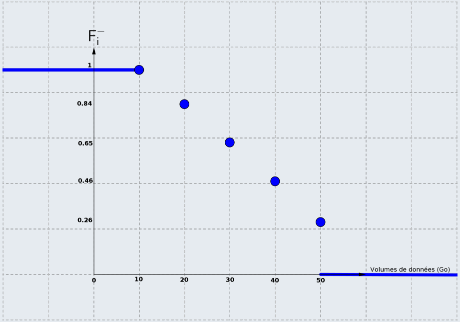
Étape 1 : Positionner les points de coordonnées (x_i, F_i^-)

x_i	10	20	30	40	50
F_i^-	1	0.84	0.65	0.46	0.26



Étape 2 : Tracer les demi-droites extrêmes

x_i	10	20	30	40	50
F_i^-	1	0.84	0.65	0.46	0.26

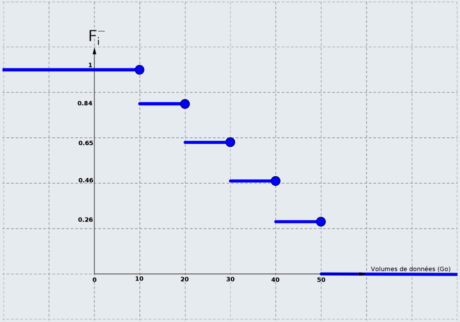


Représentations graphiques

Courbes cumulatives d'une variable discrète : Courbe cumulative descendante des fréquences

Étape 3 : Tracer les segments intermédiaires

x_i	10	20	30	40	50
F_i^-	1	0.84	0.65	0.46	0.26



F_i^- est la **proportion** d'individus qui ont **au moins** la modalité x_i.

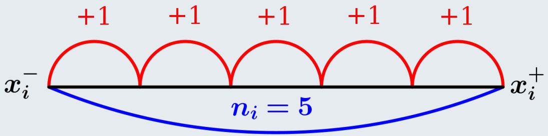
Remarque

La courbe cumulative descendante des fréquences d'une variable discrète est la représentation graphique d'une fonction càglàd (continue à gauche, limite à droite) définie sur $\mathbb{R}$, non-croissante et à valeurs dans $[0; 1]$.

Logique de construction

$[x_i^-;x_i^+[$	n_i	f_i	N_i^-	N_i^+	F_i^-	F_i^+
[0 ;4[	20	0.22	90	20	1	0.22
[4 ;8[	24	0.27	70	44	0.78	0.49
[8 ;12[	34	0.38	46	78	0.51	0.87
[12 ;16[	7	0.07	12	85	0.13	0.94
[16 ;20]	5	0.06	5	90	0.06	1

Hypothèse d'uniformité



Représentations graphiques

Courbes cumulatives d'une variable continue

$[x_i^-, x_i^+[$	x_i	n_i	f_i	N_i^-	N_i^+	F_i^-	F_i^+
	$-\infty$	0	0	90	0	1	0
$]-\infty; 0[$	$\uparrow$	0	0	90	0	1	0
		0	0	90	0	1	0
	0	4	0.04	90	4	1	0.04
	1	4	0.04	86	8	0.96	0.08
$[0; 4[$	2	4	0.04	82	12	0.92	0.12
	3	4	0.04	78	16	0.88	0.16
	3.99...	4	0.04	74	20	0.84	0.20
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	16	1	0.011	5	86	0.056	0.56
	17	1	0.011	4	87	0.044	0.67
$[16; 20]$	18	1	0.011	3	88	0.033	0.78
	19	1	0.011	2	89	0.022	0.89
	20	1	0.011	1	90	0.011	1
		0	0	0	90	0	1
$]20; +\infty[$	$\downarrow$	0	0	0	90	0	1
	$+\infty$	0	0	0	90	0	1

Logique de construction

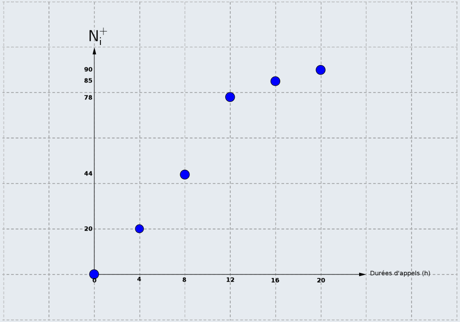
1. Imaginer toutes les valeurs $x_i \in \mathbb{R}$ pour la variable X, y compris celles démunies de sens ;
2. Attribuer des effectifs nuls aux valeurs ne faisant pas partie des modalités de X ;
3. Compléter le tableau statistique.

Constatations

- Les effectifs et les fréquences cumulés sont croissants ou décroissants à l'intérieur des classes ;
- Avant la première classe de X, les effectifs et les fréquences cumulés croissants sont nuls ;
- Après la dernière classe de X, les effectifs et les fréquences cumulés décroissants sont nuls ;
- Avant la première classe de X, les effectifs cumulés décroissants sont égaux à n ;
- Avant la première classe de X, les fréquences cumulées décroissantes sont égales à 1 ;
- Après la dernière classe de X, les effectifs cumulés croissants sont égaux à n ;
- Après la dernière classe de X, les fréquences cumulées croissantes sont égales à 1.

Étape 1 : Positionner les points de coordonnées (x_i^+, N_i^+)

$[x_i^-; x_i^+[$	$[0;4[$	$[4;8[$	$[8;12[$	$[12;16[$	$[16;20]$
N_i^+	20	44	78	85	90

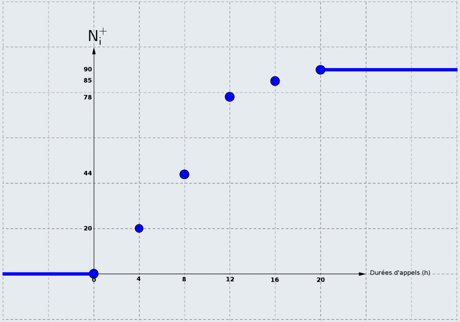


Représentations graphiques

Courbes cumulatives d'une variable continue : Courbe cumulative ascendante des effectifs

Étape 2 : Tracer les demi-droites extrêmes

$[x_i^-; x_i^+[$	$[0;4[$	$[4;8[$	$[8;12[$	$[12;16[$	$[16;20]$
N_i^+	20	44	78	85	90

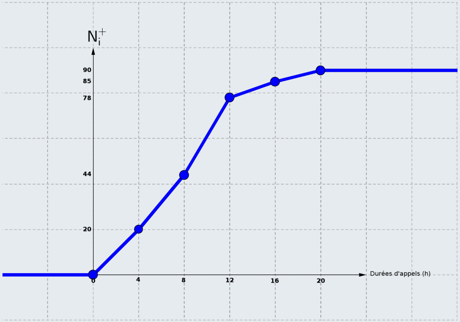


Représentations graphiques

Courbes cumulatives d'une variable continue : Courbe cumulative ascendante des effectifs

Étape 3 : Tracer les segments intermédiaires

$[x_i^-; x_i^+[$	$[0; 4[$	$[4; 8[$	$[8; 12[$	$[12; 16[$	$[16; 20]$
N_i^+	20	44	78	85	90



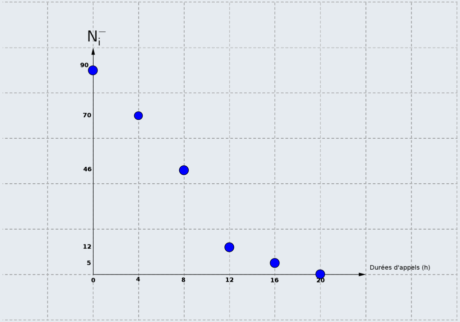
N_i^+ est le **nombre** d'individus qui ont **au plus** la modalité x_i^+.

Remarque

La courbe cumulative ascendante des effectifs d'une variable continue est la représentation graphique d'une fonction continue, définie sur $\mathbb{R}$, non-décroissante et à valeurs dans $[0; n]$.

Étape 1 : Positionner les points de coordonnées (x_i^-, N_i^-)

$[x_i^-; x_i^+[$	$[0;4[$	$[4;8[$	$[8;12[$	$[12;16[$	$[16;20]$
N_i^-	90	70	46	12	5

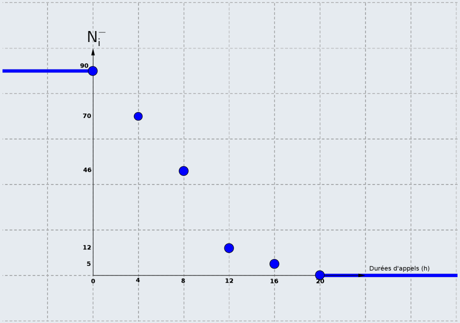


Représentations graphiques

Courbes cumulatives d'une variable continue : Courbe cumulative descendante des effectifs

Étape 2 : Tracer les demi-droites extrêmes

$[x_i^-; x_i^+[$	$[0;4[$	$[4;8[$	$[8;12[$	$[12;16[$	$[16;20]$
N_i^-	90	70	46	12	5

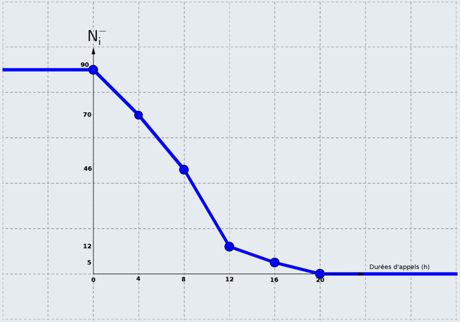


Représentations graphiques

Courbes cumulatives d'une variable continue : Courbe cumulative descendante des effectifs

Étape 3 : Tracer les segments intermédiaires

$[x_i^-; x_i^+[$	$[0;4[$	$[4;8[$	$[8;12[$	$[12;16[$	$[16;20]$
N_i^-	90	70	46	12	5



N_i^- est le **nombre** d'individus qui ont **au moins** la modalité x_i^-.

Remarque

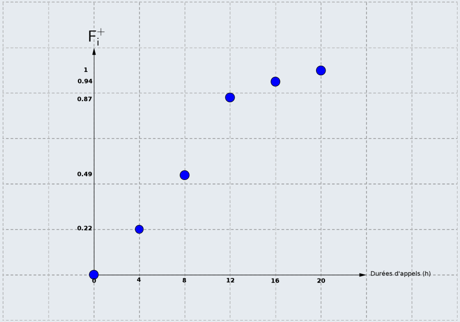
La courbe cumulative descendante des effectifs d'une variable continue est la représentation graphique d'une fonction continue, définie sur $\mathbb{R}$, non-croissante et à valeurs dans $[0; n]$.

Représentations graphiques

Courbes cumulatives d'une variable continue : Courbe cumulative ascendante des fréquences

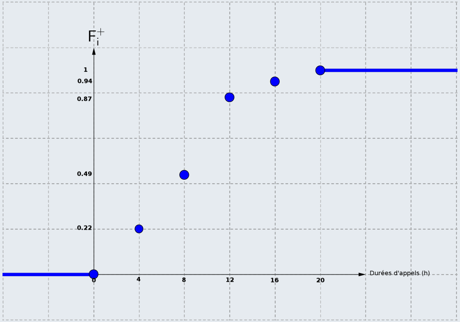
Étape 1 : Positionner les points de coordonnées (x_i^+, F_i^+)

$[x_i^-; x_i^+[$	$[0;4[$	$[4;8[$	$[8;12[$	$[12;16[$	$[16;20]$
F_i^+	0.22	0.49	0.87	0.94	1



Étape 2 : Tracer les demi-droites extrêmes

$[x_i^-; x_i^+[$	$[0;4[$	$[4;8[$	$[8;12[$	$[12;16[$	$[16;20]$
F_i^+	0.22	0.49	0.87	0.94	1

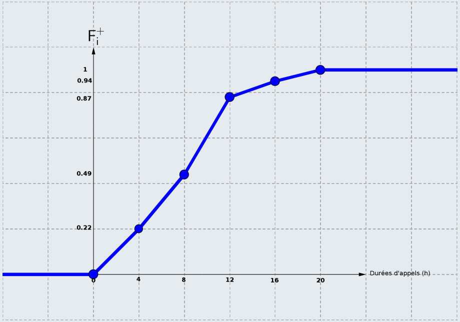


Représentations graphiques

Courbes cumulatives d'une variable continue : Courbe cumulative ascendante des fréquences

Étape 3 : Tracer les segments intermédiaires

$[x_i^-; x_i^+[$	$[0; 4[$	$[4; 8[$	$[8; 12[$	$[12; 16[$	$[16; 20]$
F_i^+	0.22	0.49	0.87	0.94	1



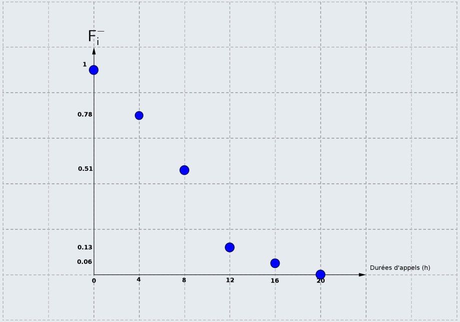
F_i^+ est la **proportion** d'individus qui ont **au plus** la modalité x_i^+.

Remarque

La courbe cumulative ascendante des fréquences d'une variable continue est la représentation graphique d'une fonction continue, définie sur $\mathbb{R}$, non-décroissante et à valeurs dans $[0; 1]$.

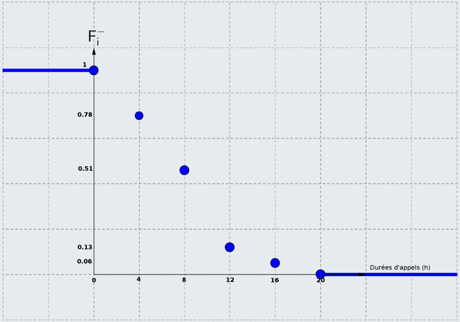
Étape 1 : Positionner les points de coordonnées (x_i^-, F_i^-)

$[x_i^-; x_i^+[$	$[0;4[$	$[4;8[$	$[8;12[$	$[12;16[$	$[16;20]$
F_i^-	1	0.78	0.51	0.13	0.06



Étape 2 : Tracer les demi-droites extrêmes

$[x_i^-; x_i^+[$	$[0;4[$	$[4;8[$	$[8;12[$	$[12;16[$	$[16;20]$
F_i^-	1	0.78	0.51	0.13	0.06

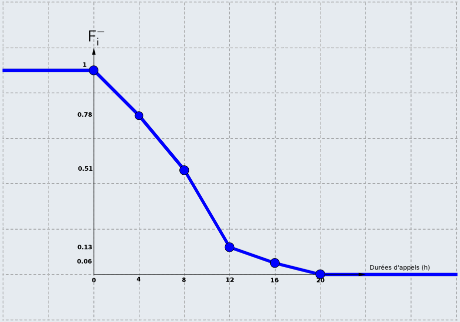


Représentations graphiques

Courbes cumulatives d'une variable continue : Courbe cumulative descendante des fréquences

Étape 3 : Tracer les segments intermédiaires

$[x_i^-; x_i^+[$	$[0;4[$	$[4;8[$	$[8;12[$	$[12;16[$	$[16;20]$
F_i^-	1	0.78	0.51	0.13	0.06



F_i^- est la **proportion** d'individus qui ont **au moins** la modalité x_i^-.

Remarque

La courbe cumulative descendante des fréquences d'une variable continue est la représentation graphique d'une fonction continue, définie sur $\mathbb{R}$, non-croissante et à valeurs dans $[0; 1]$.

SECTION 3

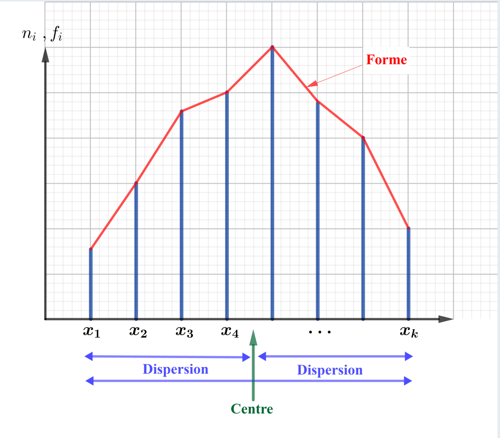
INDICATEURS STATISTIQUES

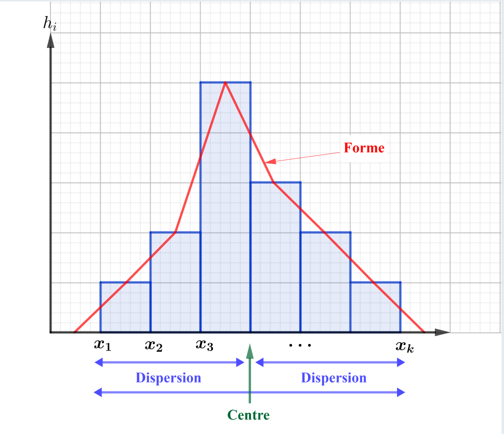
1. MESURES DE POSITION
2. MESURES DE DISPERSION
3. MESURES DE FORME
4. MESURES DE CONCENTRATION

Remarque

Dans cette section, seules seront considérées des variables quantitatives étant donné que les variables qualitatives ne se prêtent pas aux calculs mathématiques qui seront présentés.

Objectif


n_i , f_i
Forme
x_1
x_2
x_3
x_4
...
x_k
Dispersion
Dispersion
Centre


h_i
Forme
x_1
x_2
x_3
...
x_k
Dispersion
Dispersion
Centre

Sous-section 1

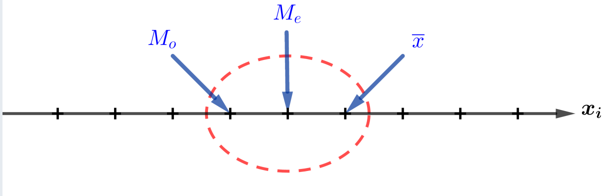
Mesures de position

1. Mesures de tendance centrale
   - 1.1 Mode
   - 1.2 Médiane
   - 1.3 Moyenne arithmétique
2. Mesures de positions relatives
   - 2.1 Quartiles
   - 2.2 Déciles
   - 2.3 Centiles
   - 2.4 Quantiles d'ordre α

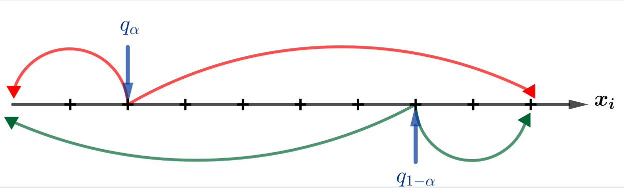
Mesure de position

Une mesure de position est un indicateur d'une position particulière dans une distribution statistique. Cette position particulière peut être le centre autour duquel les données sont réparties de façon équilibrée (mesure de tendance centrale), ou au contraire une position telle que les données se répartissent dans des proportions inégales de part et d'autre (mesure de position relative).

Positions centrales : Mode, médiane et moyenne arithmétique



Positions relatives : Quantiles



Mesures de tendance centrale

- Mode
- Médiane
- Moyenne arithmétique

Mode

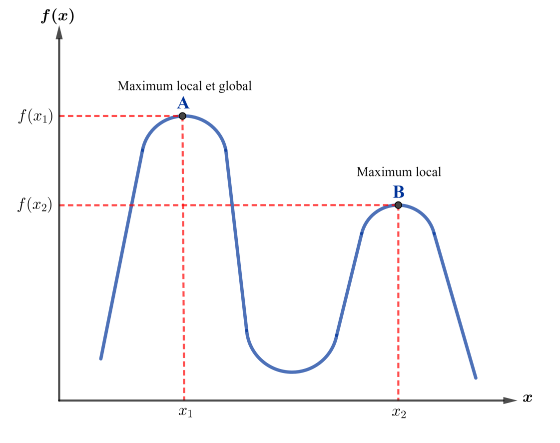
Mode

Le mode d'une distribution statistique, noté M_o, est la modalité du caractère la plus fréquemment rencontrée localement. C'est l'équivalent du maximum d'une fonction en analyse mathématique.

Remarque

Une distribution statistique peut avoir un, deux ou plusieurs modes, on parle alors respectivement de distributions unimodale, bimodale et multimodale.

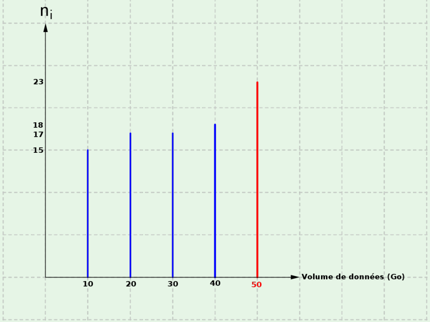
Maximum local - Maximum global

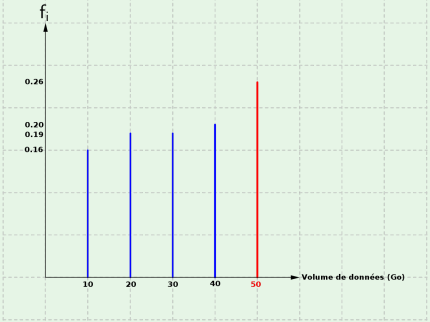


Règle

Le mode est la modalité correspondant au bâton le plus élevé localement.

Exemple : Besoin mensuel des étudiants en volumes de données





Le besoin mensuel des étudiants en volumes de données le plus fréquent est $M_o = 50$ Go.

Indicateurs statistiques

Mesures de position : Mesures de tendance centrale (Mode d'une variable discrète à partir du tableau statistique)

Règle

Le mode est la modalité correspondant à l'effectif (ou à la fréquence) maximal(e) localement.

Exemple : Besoin mensuel des étudiants en volumes de données

x_i	n_i	f_i
10	15	0.16
20	17	0.19
30	17	0.19
40	18	0.20
50	23	0.26

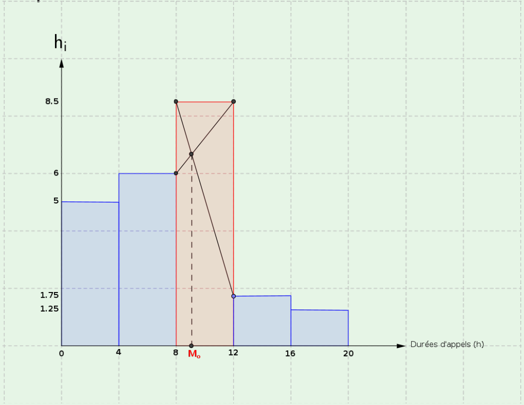
Le besoin mensuel des étudiants en volumes de données le plus fréquent est $M_o = 50$ Go.

Règle

1. Identifier la classe modale $\left[x_i^- ; x_i^+\right[$, celle qui a le rectangle le plus élevé localement.
2. Déterminer la valeur modale M_o graphiquement.

Exemple : Besoin mensuel des étudiants en durées d'appels

La classe modale est $[8; 12[$ et le mode est $M_o \approx 9$ heures. La durée d'appels la plus fréquente dont les étudiants ont besoin chaque mois est d'environ 9 heures.



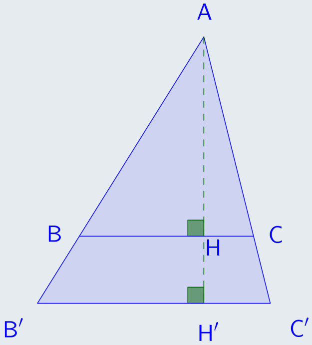
Règle

1. Identifier la classe modale $[x_i^-; x_i^+[$, celle qui a la densité h_i la plus élevée localement.
2. Calculer la valeur modale M_o analytiquement selon la formule suivante :

$$M_0 = x_i^- + \frac{d_1}{d_1 + d_2}\left(x_i^+ - x_i^-\right) \text{ avec } d_1 = h_i - h_{i-1} \text{ et } d_2 = h_i - h_{i+1}$$

Rappels : Triangles semblables & Théorème de Thalès

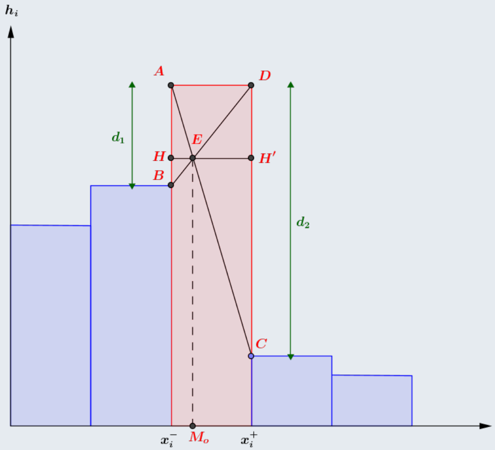
Deux triangles sont semblables si leurs angles sont égaux deux à deux. Par conséquent, ils ont la même forme mais pas nécessairement la même taille.



Théorème de Thalès :

$$\frac{AB}{AB'} = \frac{AC}{AC'} = \frac{BC}{B'C'} = \frac{AH}{AH'}$$

Démonstration : Les triangles ABE et CDE sont semblables



$$\frac{EH}{EH'} = \frac{AB}{DC} \Longleftrightarrow \frac{M_o - x_i^-}{x_i^+ - M_o} = \frac{d_1}{d_2}$$

$$M_o - x_i^- = \frac{d_1}{d_2}(x_i^+ - M_o)$$

$$M_o - x_i^- = \frac{d_1}{d_2}\left[(x_i^+ - x_i^-) - (M_o - x_i^-)\right]$$

$$M_o - x_i^- = \frac{d_1}{d_2}(x_i^+ - x_i^-) - \frac{d_1}{d_2}(M_o - x_i^-)$$

$$(M_o - x_i^-)\left(\frac{d_1 + d_2}{d_2}\right) = \frac{d_1}{d_2}(x_i^+ - x_i^-)$$

$$M_o = x_i^- + \frac{d_1}{d_1 + d_2}(x_i^+ - x_i^-)$$

Exemple : Besoin mensuel des étudiants en durées d'appels

La classe modale est celle qui à la densité h_i la plus élevée localement : $M_o \in [8; 12[$

$[x_i^-; x_i^+[$	n_i	a_i	h_i
$[0; 4[$	20	4	5
$[4; 8[$	24	4	6
$[8; 12[$	34	4	8.5
$[12; 16]$	7	4	1.75
$[16; 20]$	5	4	1.25

La valeur modale M_o est donnée par :

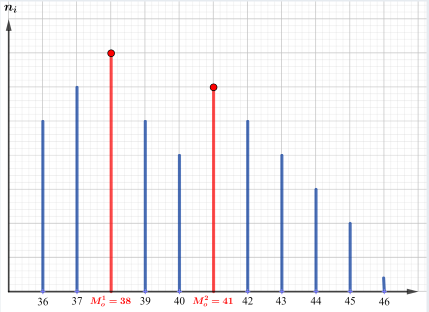
$$M_o = 8 + \frac{(8.5 - 6)}{(8.5 - 6) + (8.5 - 1.75)} \times (12 - 8) \approx 9.08 \text{ heures}$$

La durée d'appels la plus fréquente dont les étudiants ont besoin chaque mois est 9.08 heures.

Remarque

Les modes des distributions statistiques multimodales ne sont pas des mesures de centre mais plutôt une indication sur l'hétérogénéité de la population.

Illustration : Paires de chaussures vendues



Remarque

Le mode est sensible au découpage en classes d'une variable continue.

Illustration

$[x_i^-; x_i^+[$	n_i	a_i	h_i
$[0; 4[$	20	4	5
$[4; 8[$	24	4	6
$[8; 12[$	34	4	8.5
$[12; 16[$	7	4	1.75
$[16; 20]$	5	4	1.25

$M_o \approx 9.08$ h

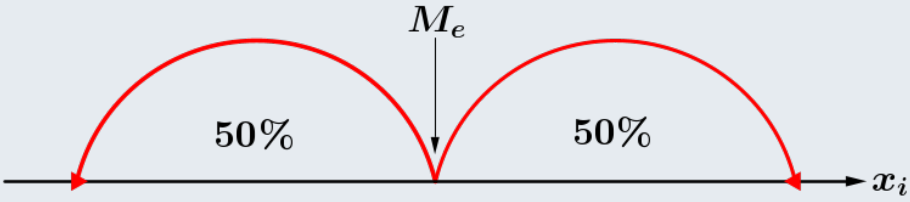
$[x_i^-; x_i^+[$	n_i	a_i	h_i
$[0; 8[$	44	8	5.5
$[8; 16[$	41	8	5.125
$[16; 20]$	5	4	1.25

$M_o \approx 7.49$ h

Médiane

Médiane

La médiane d'une distribution statistique, notée M_e, est la valeur du caractère qui partage la série statistique ordonnée en deux parties contenant chacune la moitié des observations environ.



Exemple 1 : La médiane est égale à 5

$$\begin{array}{ccccccccc} & & & & M_e & & & & \\ 2 & 3 & 3 & 4 & 5 & 6 & 6 & 7 & 8 \end{array}$$

	x_i	n_i	f_i	F_i^+	
	2	1	0.11	0.11	
	3	2	0.22	0.33	
	4	1	0.11	0.44	
$\rightarrow$	5	1	0.11	0.55	$\leftarrow$
	6	2	0.22	0.77	
	7	1	0.11	0.88	
	8	1	0.12	1	

5 est la médiane de cette distribution statistique. On remarque que sa fréquence cumulée croissante est immédiatement supérieure à 0.5.

Exemple 2 : La médiane est égale à 5 (elle se situe entre 5 et 5 !)

2	3	3	4	5	M_e ↓	5	6	7	8	9

	x_i	n_i	f_i	F_i^+	
	2	1	0.1	0.1	
	3	2	0.2	0.3	
	4	1	0.1	0.4	
→	5	2	0.2	0.6	←
	6	1	0.1	0.7	
	7	1	0.1	0.8	
	8	1	0.1	0.9	
	9	1	0.1	1	

5 est la médiane de cette distribution statistique. On remarque que sa fréquence cumulée croissante est immédiatement supérieure à 0.5.

Exemple 3 : La médiane se situe entre 5 et 6, on a un intervalle médian [5 ;6]

					M_e					
2	3	3	4	5	$\downarrow$	6	6	7	8	9

	x_i	n_i	f_i	F_i^+	
	2	1	0.1	0.1	
	3	2	0.2	0.3	
	4	1	0.1	0.4	
$\rightarrow$	5	1	0.1	0.5	$\leftarrow$
$\rightarrow$	6	2	0.2	0.7	
	7	1	0.1	0.8	
	8	1	0.1	0.9	
	9	1	0.1	1	

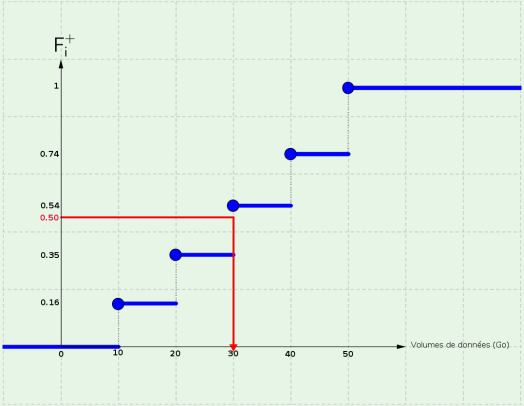
On a un intervalle médian [5 ;6]. On remarque que la fréquence cumulée croissante de sa borne inférieure est égale à 0.5. La médiane est le centre de l'intervalle médian $M_e = 5.5$.

Règle : (à partir du tableau statistique)

- Si aucune fréquence cumulée croissante n'est égale à 0.5, alors la médiane est la modalité dont la fréquence cumulée croissante est immédiatement supérieure à 0.5.
- Si la fréquence cumulée croissante d'une modalité x_i est égale à 0.5, alors on a un intervalle médian $[x_i; x_{i+1}]$. La médiane est le centre de cet intervalle.

Exemple 1 : Besoin mensuel des étudiants en volumes de données

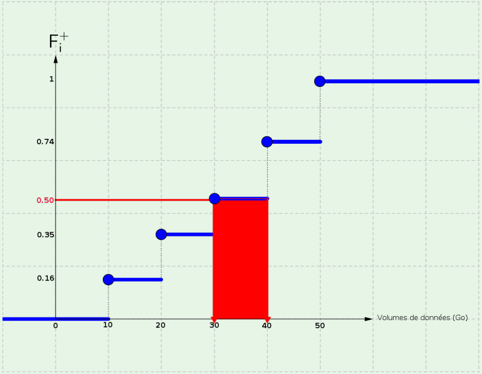
x_i	F_i^+
10	0.16
20	0.35
30	0.54
40	0.74
50	1



La médiane est $M_e = 30$ Go. La moitié des étudiants questionnés a besoin de moins de 30 Go de données par mois et la moitié a besoin de plus de 30 Go de données par mois.

Exemple 2 : Besoin mensuel des étudiants en volumes de données

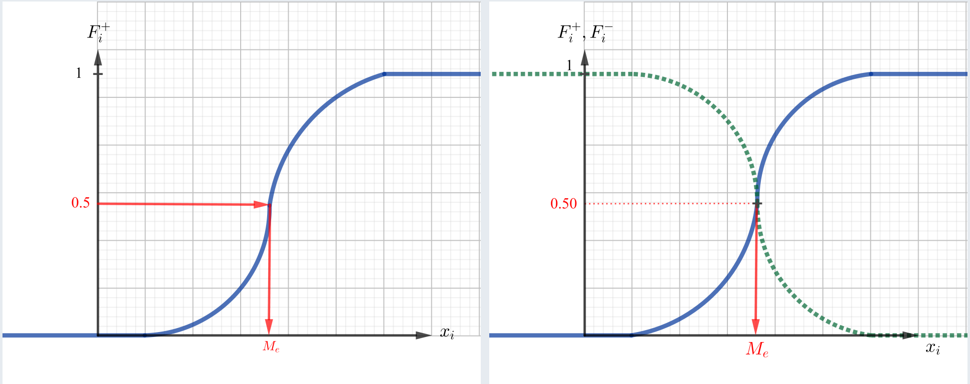
x_i	F_i^+
10	0.16
20	0.35
30	0.50
40	0.74
50	1



On a un intervalle médian $[30; 40]$. La médiane est le centre de cet intervalle $M_e = 35$ Go. La moitié des étudiants questionnés a besoin de moins de 35 Go de données par mois et la moitié a besoin de plus de 35 Go de données par mois.

Règle : (à partir d'une courbe cumulative)

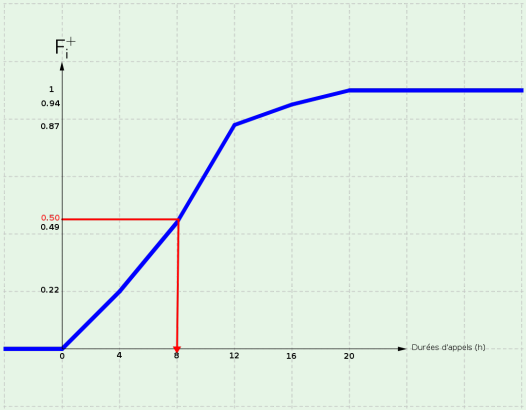
La médiane est l'abscisse du point associé à la fréquence cumulée croissante 0.5. La médiane est également l'abscisse du point d'intersection des courbes cumulatives ascendante et descendante.

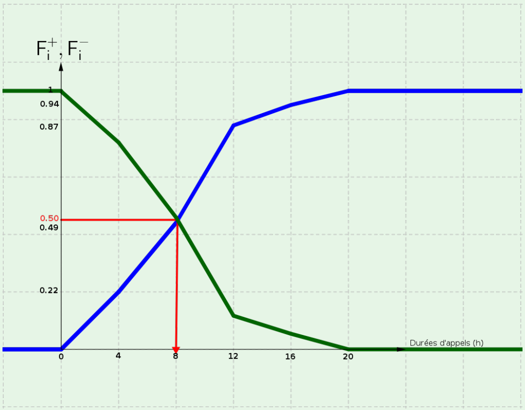


Indicateurs statistiques

Mesures de position : Mesures de tendance centrale (Médiane d'une variable continue)

Exemple : Besoin mensuel des étudiants en durées d'appels



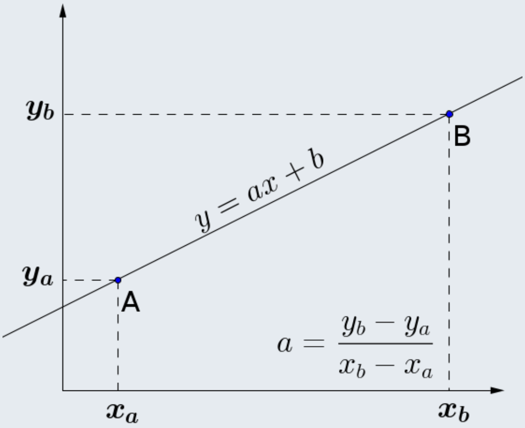


Règle : (à partir du tableau statistique)

- S'il y a une fréquence cumulée croissante égale à 0.5, alors la médiane est égale à la borne supérieure de la classe correspondant à la fréquence cumulée croissante égale à 0.5.
- Si aucune fréquence cumulée croissante n'est égale à 0.5, alors la *classe médiane* $[x_i^-; x_i^+[$ est la classe dont la fréquence cumulée croissante est immédiatement supérieure à 0.5. Dans ce cas, la valeur médiane M_e est calculée par interpolation linéaire comme suit :

$$M_e = x_i^- + \frac{0.5 - F_{i-1}^+}{F_i^+ - F_{i-1}^+}\left(x_i^+ - x_i^-\right)$$

Démonstration : Coefficient directeur d'une droite



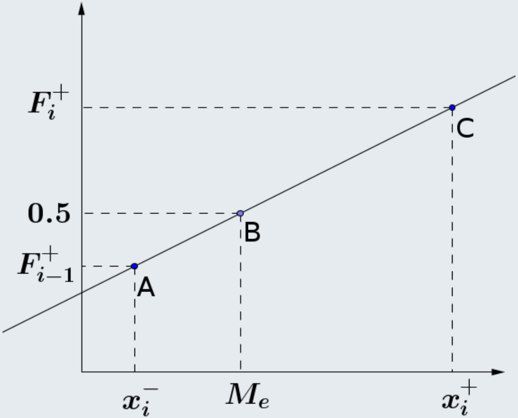
Démonstration (suite)

$$a = \frac{y_b - y_a}{x_b - x_a} = \frac{0.5 - F_{i-1}^{+}}{M_e - x_i^{-}}$$

$$a = \frac{y_C - y_A}{x_C - x_A} = \frac{F_i^{+} - F_{i-1}^{+}}{x_i^{+} - x_i^{-}}$$

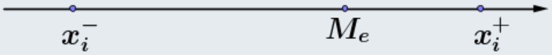
$$\Rightarrow \frac{0.5 - F_{i-1}^{+}}{M_e - x_i^{-}} = \frac{F_i^{+} - F_{i-1}^{+}}{x_i^{+} - x_i^{-}}$$

$$M_e = x_i^{-} + \frac{0.5 - F_{i-1}^{+}}{F_i^{+} - F_{i-1}^{+}} \times (x_i^{+} - x_i^{-})$$



En pratique

1. Identifier la classe médiane $[x_i^- ; x_i^+[$ et situer la valeur médiane M_e à l'intérieur de cette classe.

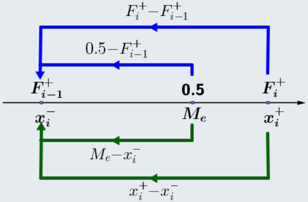


2. Associer la fréquence cumulée croissante 0.5 à la valeur médiane M_e.

En pratique

3. Trouver dans la colonne des fréquences cumulées croissantes les valeurs qui encadrent 0.5 et les associer aux bornes de la classe médiane.

4. Calculer la médiane par interpolation linéaire comme suit :

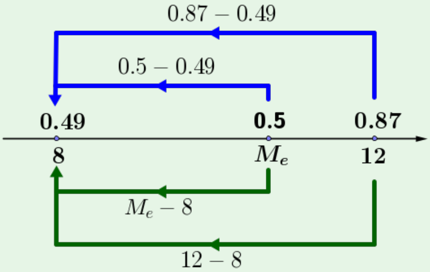


$$\frac{M_e - x_i^{-}}{x_i^{+} - x_i^{-}} = \frac{0.5 - F_{i-1}^{+}}{F_i^{+} - F_{i-1}^{+}}$$

$$M_e = x_i^{-} + \frac{0.5 - F_{i-1}^{+}}{F_i^{+} - F_{i-1}^{+}} \times \left(x_i^{+} - x_i^{-}\right)$$

Exemple 1 : Besoin mensuel des étudiants en durées d'appels

Durées (h)	F_i^+
$[0;4[$	0.22
$[4;8[$	0.49
$[8;12[$	0.87
$[12;16[$	0.94
$[16;20]$	1



$$M_e = 8 + \frac{0.5-0.49}{0.87-0.49} \times (12-8) \approx 8.11 \text{ heures}$$

La moitié des étudiants a besoin de moins de 8.11 heures d'appels par mois et la moitié a besoin de plus de 8.11 heures d'appels par mois.

Exemple 2 : Besoin mensuel des étudiants en durées d'appels

Durées (h)	F_i^+
$[0;4[$	0.22
$[4;8[$	0.50
$[8;12[$	0.87
$[12;16[$	0.94
$[16;20]$	1

La médiane est $M_e = 8$ heures. La moitié des étudiants a besoin de moins de 8 heures d'appels par mois et la moitié a besoin de plus de 8 heures d'appels par mois.

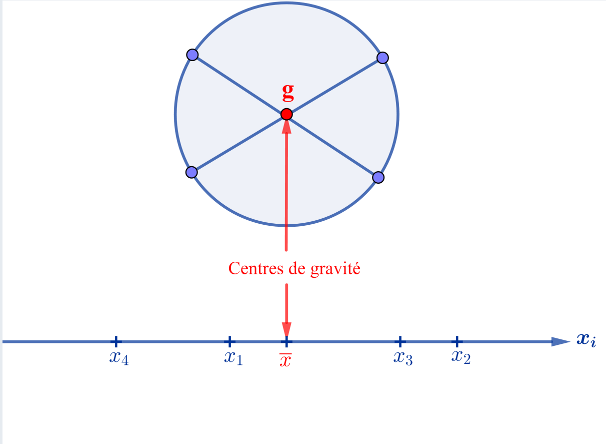
Moyenne arithmétique

Moyenne arithmétique

La *moyenne arithmétique* d'une variable statistique, notée $\overline{x}$, est la valeur qui concentre toute l'information contenue dans la distribution statistique. C'est l'équivalent d'un centre de gravité en géométrie. Elle est donnée par la formule suivante :

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{k} n_i x_i = \sum_{i=1}^{k} f_i x_i$$

Centres de gravité



Remarque

La moyenne arithmétique d'une variable continue X s'écrit :

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{k} n_i c_i = \sum_{i=1}^{k} f_i c_i$$

où c_i désigne le centre de la classe $[x_i^- ; x_i^+[$.

Exemple 1 : Besoin mensuel des étudiants en volumes de données

x_i	n_i	$n_i x_i$
10	15	150
20	17	340
30	17	510
40	18	720
50	23	1150
Σ	90	2870

$$\begin{aligned}\overline{x} &= \frac{1}{n}\sum_{i=1}^{5} n_i x_i \\ &= \frac{2870}{90} \\ &\approx 31.89 \text{ Go}\end{aligned}$$

Les étudiants questionnés ont besoin, en moyenne, de 31.89 Go de données par mois.

Exemple 2 : Besoin mensuel des étudiants en durées d'appels

$[x_i^-; x_i^+[$	c_i	n_i	$n_i c_i$
$[0;4[$	2	20	40
$[4;8[$	6	24	144
$[8;12[$	10	34	340
$[12;16[$	14	7	98
$[16;20]$	18	5	90
Σ	–	90	712

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{5} n_i c_i = \frac{712}{90} \approx 7.91 \text{ h}$$

Les étudiants questionnés ont besoin, en moyenne, de 7.91 h d'appels par mois.

Propriétés

Soient X et Y deux variables statistiques quantitatives prenant respectivement les valeurs $x_1, x_2, \cdots, x_k$ et $y_1, y_2, \cdots, y_k$ avec $k \in \mathbb{N}^*$ et a et b deux constantes réelles.

La moyenne arithmétique vérifie les propriétés suivantes :

$$\overline{ax + by} = a\overline{x} + b\overline{y}$$

$$\sum_{i=1}^{k} n_i(x_i - \overline{x}) = 0$$

Démonstrations

$$\begin{aligned}
\overline{ax+by} &= \frac{1}{n}\sum_{i=1}^{k} n_i(ax_i+by_i) \\
&= \frac{1}{n}\sum_{i=1}^{k}(an_ix_i+bn_iy_i) \\
&= \frac{1}{n}\sum_{i=1}^{k} an_ix_i + \frac{1}{n}\sum_{i=1}^{k} bn_iy_i \\
&= a\left(\frac{1}{n}\sum_{i=1}^{k} n_ix_i\right) + b\left(\frac{1}{n}\sum_{i=1}^{k} n_iy_i\right) \\
&= a\overline{x} + b\overline{y}
\end{aligned}$$

Exemples

$$\overline{x+y} = \overline{x} + \overline{y}$$

$$\overline{2x+y} = 2\overline{x} + \overline{y}$$

$$\overline{3x-y} = 3\overline{x} - \overline{y}$$

$$\overline{x-3y} = \overline{x} - 3\overline{y}$$

$$\overline{x+2y} = \overline{x} + 2\overline{y}$$

$$\overline{3x+2} = 3\overline{x} + 2$$

$$\overline{3-2y} = 3 - 2\overline{y}$$

Démonstrations (suite)

$$\begin{aligned}
\sum_{i=1}^{k} n_i(x_i - \overline{x}) &= \sum_{i=1}^{k} (n_i x_i - n_i \overline{x}) \\
&= \sum_{i=1}^{k} n_i x_i - \sum_{i=1}^{k} n_i \overline{x} \\
&= n \times \frac{1}{n} \sum_{i=1}^{k} n_i x_i - \overline{x} \sum_{i=1}^{k} n_i \\
&= n\overline{x} - \overline{x} n \\
&= 0
\end{aligned}$$

Remarque

La moyenne arithmétique est sensible aux valeurs extrêmes contrairement à la médiane qui n'est pas influencée par de telles observations. La médiane est, de ce fait, une mesure *robuste* du centre de la distribution statistique.

Exemple : Notes sur 20 de deux groupes d'étudiants

Groupe A	Groupe B
15 15 15 15 15	1 15 15 15 15
$\overline{x} = 15$	$\overline{x} = 12.2$
$M_e = 15$	$M_e = 15$

Mesures de position relative ou quantiles

- Quartiles
- Déciles
- Centiles
- Quantiles d'ordre α
- Mode de calcul

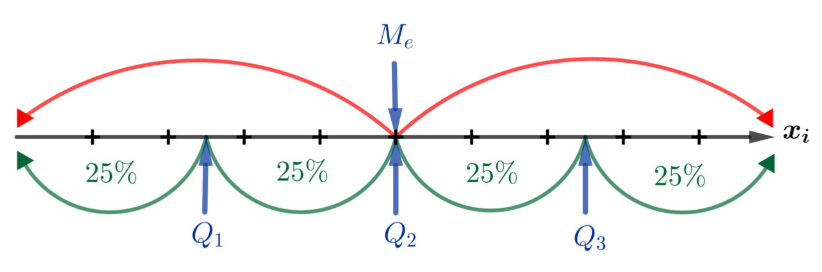
Quantiles

Les quantiles ou les fractiles sont des valeurs qui partagent la série statistique ordonnée en un certain nombre de parties contenant chacune une même proportion d'observations. Ils constituent, de ce fait, une généralisation de la notion de médiane qui partage la série statistique en deux parties contenant chacune la moitié des observations.
Les quantiles les plus connus sont les quartiles, les déciles et les centiles qui divisent la série statistique en respectivement 4, 10 et 100 parties de même effectif. Par conséquent, les quantiles ne se situent pas nécessairement au centre de la distribution statistique mais dans des positions dites relatives.

Quartiles

Quartiles

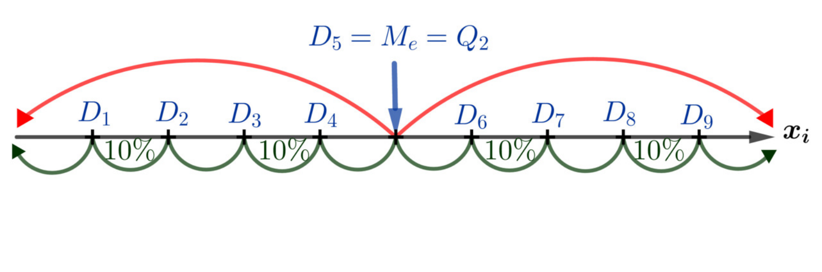
Les *quartiles* sont des valeurs qui partagent la série statistique ordonnée en 4 parties contenant chacune 25% des observations. Ils sont au nombre de 3 et sont notés Q_1, Q_2 et Q_3.



Déciles

Déciles

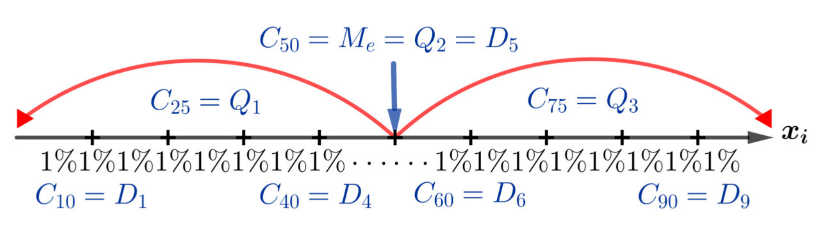
Les déciles sont des valeurs qui partagent la série statistique ordonnée en 10 parties contenant chacune 10% des observations. Ils sont au nombre de 9 et sont notés D_1, D_2... et D_9.



Centiles

Centiles

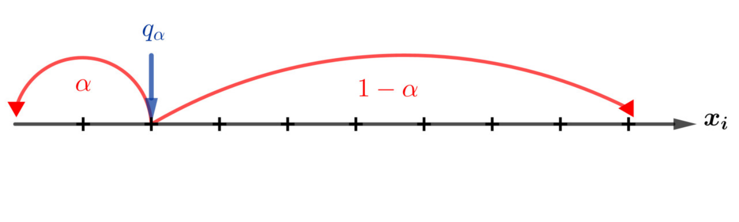
Les *centiles* sont des valeurs qui partagent la série statistique ordonnée en 100 parties contenant chacune 1% des observations. Ils sont au nombre de 99 et sont notés C_1, C_2... et C_{99}.



Quantiles d'ordre α

Quantiles d'ordre α

De façon générale, on définit les *quantiles d'ordre* α, avec $\alpha \in]0;1[$, comme les valeurs qui partagent la série statistique ordonnée en parties contenant chacune une proportion α des observations et on les note q_α. Ainsi, une proportion α des observations est inférieure à q_α et une proportion $1-\alpha$ est supérieure à cette valeur.



Remarque

- Quartiles : $Q_1 = q_{0.25}$, $Q_2 = q_{0.50}$ et $Q_3 = q_{0.75}$;
- Déciles : $D_1 = q_{0.10}$, $D_2 = q_{0.20}$,... et $D_9 = q_{0.90}$;
- Centiles : $C_1 = q_{0.01}$, $C_2 = q_{0.02}$,... et $C_{99} = q_{0.99}$.

Règle : (à partir du tableau statistique)

- S'il y a une fréquence cumulée croissante égale à α alors on a un *intervalle quantile* délimité par la modalité associée à la valeur α et par la modalité suivante.
- Si aucune fréquence cumulée croissante n'est égale à α alors le quantile recherché est la modalité ayant une fréquence cumulée croissante immédiatement supérieure à α.

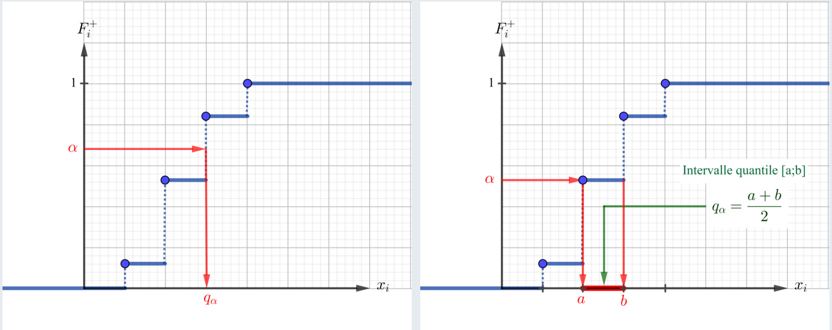
Exemple : Besoin mensuel des étudiants en volumes de données

x_i	F_i^+
10	0.16
20	0.35
30	0.54
40	0.74
50	1

- $Q_1 = 20$ Go, $Q_2 = M_e = 30$ Go et $Q_3 = 50$ Go
- $D_1 = 10$ Go, $D_5 = M_e = 30$ Go et $D_9 = 50$ Go
- $C_1 = 10$ Go, $C_{50} = M_e = 30$ Go et $C_{99} = 50$ Go
- $q_{0.18} = 20$ Go, $q_{0.50} = M_e = 30$ Go et $q_{0.70} = 40$ Go

Règle : (À partir d'une courbe cumulative)

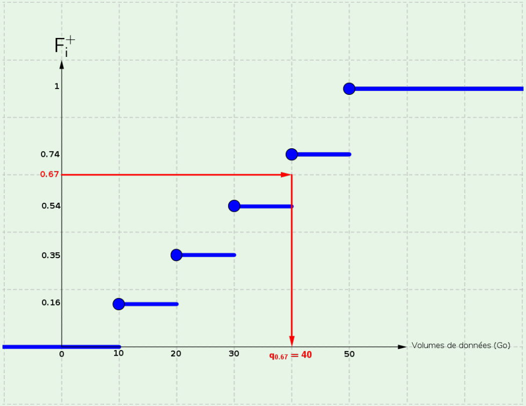
- Si la valeur α se situe entre deux paliers de la courbe cumulative ascendante, alors le quantile q_α est l'abscisse commune aux deux paliers.
- Si la valeur α est en face d'un palier de la courbe cumulative ascendante, alors on a un *intervalle quantile* délimité par les abscisses des extrémités du palier.

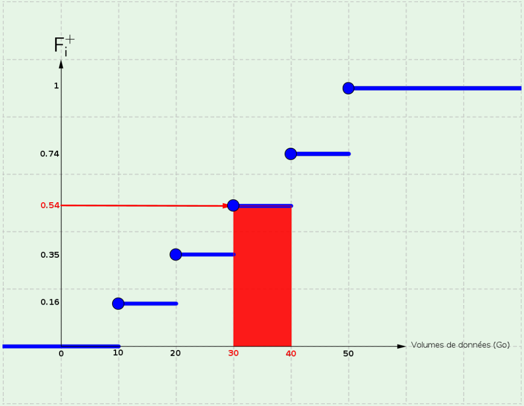


Indicateurs statistiques

Mesures de position : Mesures de position relative (Mode de calcul pour une variable discrète)

Exemple : Besoin mensuel des étudiants en volumes de données



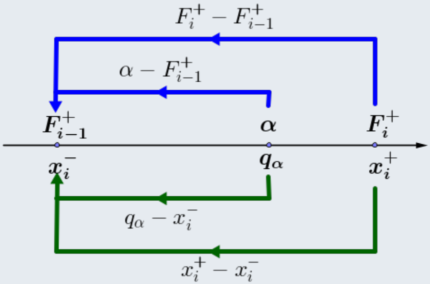


Règle : (à partir du tableau statistique)

- S'il y a une fréquence cumulée croissante égale à α, alors le quantile d'ordre α est égal à la borne supérieure de la classe correspondant à la fréquence cumulée croissante α.
- Si aucune fréquence cumulée croissante n'est égale à α, alors la classe $[x_i^- ; x_i^+[$ contenant le quantile d'ordre α est celle qui a la fréquence cumulée croissante immédiatement supérieure à α. Dans ce cas, la valeur du quantile q_α est calculée par interpolation linéaire comme suit :

$$q_\alpha = x_i^- + \frac{\alpha - F_{i-1}^+}{F_i^+ - F_{i-1}^+} \times (x_i^+ - x_i^-)$$

En pratique



$$\frac{q_\alpha - x_i^-}{x_i^+ - x_i^-} = \frac{\alpha - F_{i-1}^+}{F_i^+ - F_{i-1}^+}$$

$$q_\alpha = x_i^- + \frac{\alpha - F_{i-1}^+}{F_i^+ - F_{i-1}^+} \times (x_i^+ - x_i^-)$$

Exemple : Besoin mensuel des étudiants en durées d'appels

Durées (h)	F_i^+
$[0;4[$	0.22
$[4;8[$	0.49
$[8;12[$	0.87
$[12;16[$	0.94
$[16;20]$	1

- $M_e \in [8;12[$;
- $Q_1 \in [4;8[$, $Q_2 \in [8;12[$ et $Q_3 \in [8;12[$;
- $D_1 \in [0;4[$, $D_5 \in [8;12[$ et $D_9 \in [12;16[$;
- $C_1 \in [0;4[$, $C_{55} \in [8;12[$ et $C_{99} \in [16;20]$.

Exemple (suite): Besoin mensuel des étudiants en durées d'appels

$$M_e = 8 + \frac{0.50 - 0.49}{0.87 - 0.49} \times 4 \approx 8.11 \text{ h}$$

$$Q_1 = 4 + \frac{0.25 - 0.22}{0.49 - 0.22} \times 4 \approx 4.44 \text{ h}$$

$$Q_2 = 8 + \frac{0.50 - 0.49}{0.87 - 0.49} \times 4 \approx 8.11 \text{ h}$$

$$Q_3 = 8 + \frac{0.75 - 0.49}{0.87 - 0.49} \times 4 \approx 10.74 \text{ h}$$

$$D_1 = 0 + \frac{0.10 - 0}{0.22 - 0} \times 4 \approx 1.82 \text{ h}$$

$$D_5 = 8 + \frac{0.50 - 0.49}{0.87 - 0.49} \times 4 \approx 8.11 \text{ h}$$

$$D_9 = 12 + \frac{0.90 - 0.87}{0.94 - 0.87} \times 4 \approx 13.71 \text{ h}$$

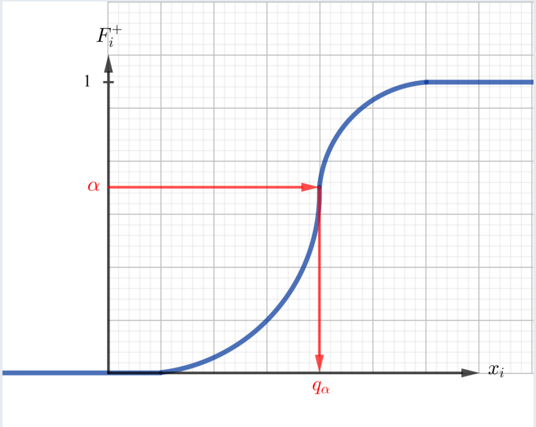
$$C_1 = 0 + \frac{0.01 - 0}{0.22 - 0} \times 4 \approx 0.18 \text{ h}$$

$$C_{55} = 8 + \frac{0.55 - 0.49}{0.87 - 0.49} \times 4 \approx 8.63 \text{ h}$$

$$C_{99} = 16 + \frac{0.99 - 0.94}{1 - 0.94} \times 4 \approx 19.33 \text{ h}$$

Règle : (à partir d'une courbe cumulative)

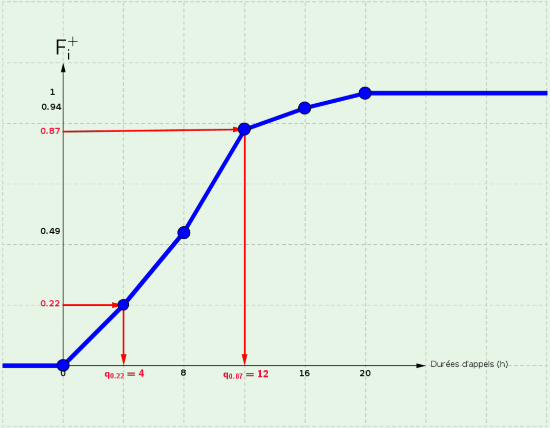
Le quantile q_α est l'abscisse du point correspondant à la fréquence cumulée croissante α.



Indicateurs statistiques

Mesures de position : Mesures de position relative (Mode de calcul pour une variable continue)

Exemple : Besoin mensuel des étudiants en durées d'appels



Sous-section 2

Mesures de dispersion

Exemple introductif : Taux de rentabilité d'actifs financiers

Actif A	Actif B	Actif C
0% 0% 0% 0% 0%	-2% -1% 0% 1% 2%	-4% -3% 0% 3% 4%
$\overline{x} = 0\%$	$\overline{x} = 0\%$	$\overline{x} = 0\%$
$M_e = 0\%$	$M_e = 0\%$	$M_e = 0\%$

Les mesures de tendance centrale ne donnent aucune information sur la *dispersion* des données !

Mesure de dispersion

Une *mesure de dispersion* permet de décrire la variabilité des données dans une série statistique. Elle vient ainsi compléter les mesures de position qui passent sous silence cet aspect.

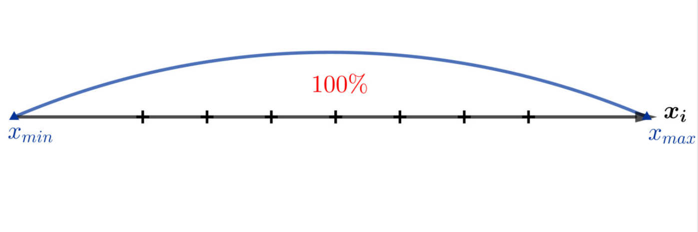
Dispersion dans un intervalle

- Intervalle de variation & Étendue d'une série statistique
- Intervalle interquartile & Écart interquartile
- Intervalle interdécile & Écart interdécile
- Intervalle intercentile & Écart intercentile

Intervalle de variation & Étendue d'une série statistique

L'*intervalle de variation* d'une série statistique est l'intervalle borné par la valeur minimale x_{min} et la valeur maximale x_{max} de variable statistique X, il est noté $IV = [x_{min}; x_{max}]$. Cet intervalle contient la totalité des données.
L'amplitude de l'intervalle de variation IV, notée E, est appelée *étendue de la série*. On a :

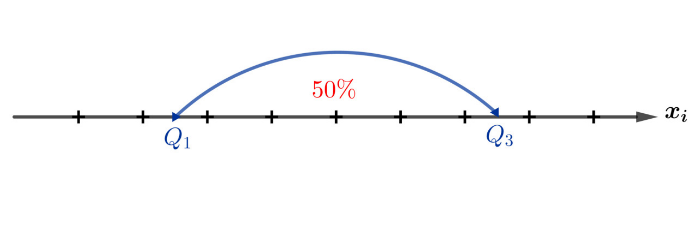
$$E = x_{max} - x_{min}$$



Intervalle interquartile & Écart interquartile

L'*intervalle interquartile* d'une série statistique est l'intervalle borné par le premier et le dernier quartiles de la variable X, il est noté $IQ = [Q_1; Q_3]$. Cet intervalle contient la moitié des données. L'amplitude de l'intervalle interquartile IQ, notée EIQ, est appelée *écart interquartile ou étendue interquartile*. On a :

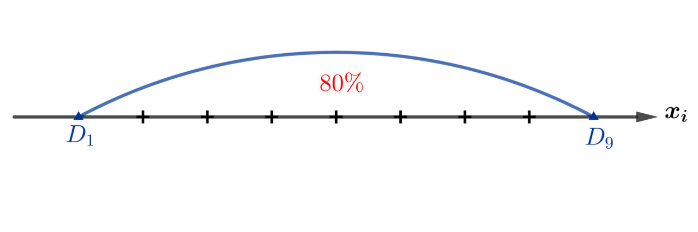
$$EIQ = Q_3 - Q_1$$



Intervalle interdécile & Écart interdécile

L'*intervalle interdécile* d'une série statistique est l'intervalle borné par le premier et le dernier déciles de la variable X, il est noté $ID = [D_1; D_9]$. Cet intervalle contient 80% des données.
L'amplitude de l'intervalle interdécile ID, notée EID, est appelée *écart interdécile ou étendue interdécile*. On a :

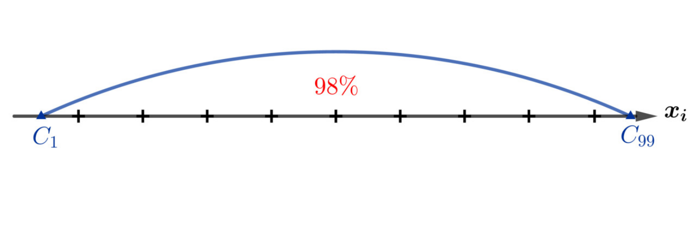
$$EID = D_9 - D_1$$



Intervalle intercentile & Écart intercentile

L'*intervalle intercentile* d'une série statistique est l'intervalle borné par le premier et le dernier centiles de la variable X, il est noté $IC = [C_1; C_{99}]$. Cet intervalle contient 98% des données.
L'amplitude de l'intervalle intercentile IC, notée EIC, est appelée *écart intercentile ou étendue intercentile*. On a :

$$EIC = C_{99} - C_1$$



Indicateurs statistiques

Mesures de dispersion : Dispersion dans un intervalle

Exemple 1 : Besoin mensuel des étudiants en volumes de données

$$\begin{aligned}
IV &= [x_{min}; x_{max}] = [10; 50] & \rightarrow \quad & E = 50 - 10 = 40 \text{ Go} \\
IQ &= [Q_1; Q_3] = [20; 50] & \rightarrow \quad & EIQ = 50 - 20 = 30 \text{ Go} \\
ID &= [D_1; D_9] = [10; 50] & \rightarrow \quad & EID = 50 - 10 = 40 \text{ Go} \\
IC &= [C_1; C_{99}] = [10; 50] & \rightarrow \quad & EIC = 50 - 10 = 40 \text{ Go}
\end{aligned}$$

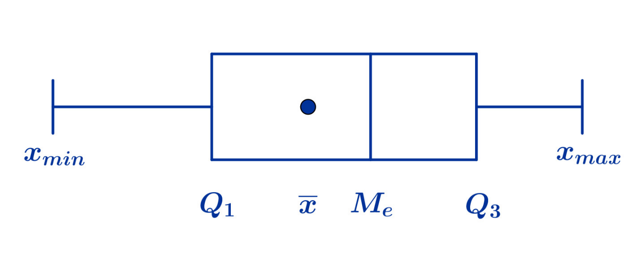
Exemple 2 : Besoin mensuel des étudiants en durées d'appels

$$\begin{aligned}
IV &= [x_{min}; x_{max}] = [0; 20] & \rightarrow \quad & E = 20 - 0 = 20 \text{ h} \\
IQ &= [Q_1; Q_3] = [4.44; 10.74] & \rightarrow \quad & EIQ = 10.74 - 4.44 = 6.3 \text{ h} \\
ID &= [D_1; D_9] = [1.82; 13.71] & \rightarrow \quad & EID = 13.71 - 1.82 = 11.89 \text{ h} \\
IC &= [C_1; C_{99}] = [0.18; 19.33] & \rightarrow \quad & EIC = 19.33 - 0.18 = 19.15 \text{ h}
\end{aligned}$$

Remarque

La dispersion d'une distribution statistique peut être résumée à l'aide d'une *boîte de Tukey*. Il s'agit d'un diagramme représentant simultanément l'intervalle de variation, l'intervalle interquartile, la moyenne arithmétique et la médiane de la distribution statistique.

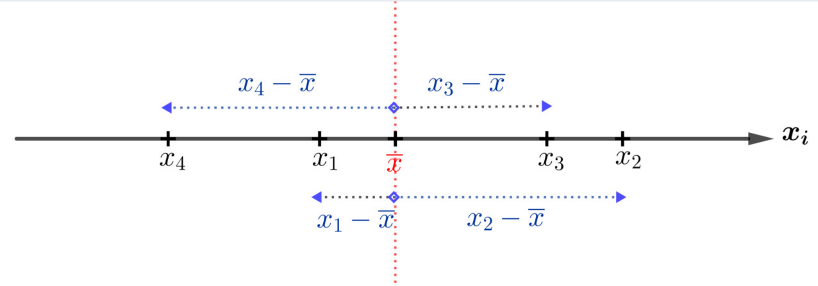
Boîte de Tukey



Dispersion autour du centre

- Écart absolu moyen
- Variance
- Écart-type
- Coefficient de variation

Écarts par rapport à la moyenne arithmétique



Écart moyen

L'*écart moyen* d'une variable statistique, notée EM, est la moyenne arithmétique des écarts des observations de la variable X par rapport au centre généralement mesuré par la moyenne arithmétique $\overline{x}$:

$$EM = \frac{1}{n}\sum_{i=1}^{k} n_i(x_i - \overline{x}) = 0$$

Écart absolu moyen

L'*écart absolu moyen* d'une variable statistique, notée EAM, est la moyenne arithmétique des valeurs absolues des écarts des observations de la variable X par rapport au centre généralement mesuré par la moyenne arithmétique $\overline{x}$:

$$\text{EAM} = \frac{1}{n}\sum_{i=1}^{k} n_i\,|x_i - \overline{x}| \geq 0$$

Exemple 1 : Besoin mensuel des étudiants en volumes de données

x_i	n_i	$\lvert x_i - \overline{x} \rvert$	$n_i \lvert x_i - \overline{x} \rvert$
10	15	21.89	328.35
20	17	11.89	202.13
30	17	1.89	32.13
40	18	8.11	145.98
50	23	18.11	416.53
Σ	90	–	1125.12

$$\text{EAM} = \frac{1}{n} \sum_{i=1}^{5} n_i \lvert x_i - \overline{x} \rvert = \frac{1125.12}{90} \approx 12.50 \text{ Go}$$

Les besoins mensuels des étudiants en volumes de donnée s'éloignent de la moyenne $\overline{x} = 31.89$ Go de 12.50 Go en moyenne.

Exemple 2 : Besoin mensuel des étudiants en durées d'appels

$[x_i^-; x_i^+[$	c_i	n_i	$\lvert c_i - \overline{x} \rvert$	$n_i \lvert c_i - \overline{x} \rvert$
$[0; 4[$	2	20	5.91	118.2
$[4; 8[$	6	24	1.91	45.84
$[8; 12[$	10	34	2.09	71.06
$[12; 16[$	14	7	6.09	42.63
$[16; 20]$	18	5	10.09	50.45
Σ	–	90	–	328.18

$$\text{EAM} = \frac{1}{n}\sum_{i=1}^{5} n_i \lvert c_i - \overline{x} \rvert = \frac{328.18}{90} \approx 3.65 \text{ h}$$

Les besoins mensuels des étudiants en durées d'appels s'éloignent de la moyenne $\overline{x} = 7.91$ heures de 3.65 heures en moyenne.

Variance & écart-type

La *variance* d'une variable statistique X, notée σ_x^2 ou $V(X)$, est la moyenne des carrés des écarts à la moyenne $\overline{x}$:

$$\sigma_x^2 = \frac{1}{n}\sum_{i=1}^{k} n_i \left(x_i - \overline{x}\right)^2$$

L'*écart-type* d'une variable statistique X, noté σ_x, est la racine carrée de la variance : $\sigma_x = \sqrt{\sigma_x^2}$.

Théorème de König-Huygens : Formule réduite de la variance

La variance peut se réécrire $\sigma_x^2 = \overline{x^2} - \overline{x}^2$ avec

$$\overline{x^2} = \frac{1}{n}\sum_{i=1}^{k} n_i x_i^2$$

Démonstration

$$\begin{aligned}
\sigma_x^2 &= \frac{1}{n}\sum_{i=1}^{k} n_i\left(x_i-\overline{x}\right)^2 = \frac{1}{n}\sum_{i=1}^{k} n_i\left(x_i^2-2x_i\overline{x}+\overline{x}^2\right)\\
&= \frac{1}{n}\sum_{i=1}^{k}\left(n_i x_i^2 - n_i 2x_i\overline{x}+n_i\overline{x}^2\right) = \frac{1}{n}\sum_{i=1}^{k} n_i x_i^2 - \frac{1}{n}\sum_{i=1}^{k} n_i 2x_i\overline{x} + \frac{1}{n}\sum_{i=1}^{k} n_i\overline{x}^2\\
&= \frac{1}{n}\sum_{i=1}^{k} n_i x_i^2 - 2\overline{x}\left(\frac{1}{n}\sum_{i=1}^{k} n_i x_i\right) + \overline{x}^2\left(\frac{1}{n}\sum_{i=1}^{k} n_i\right) = \overline{x^2} - 2\overline{x}.\overline{x} + \overline{x}^2 = \overline{x^2} - \overline{x}^2
\end{aligned}$$

Indicateurs statistiques

Mesures de dispersion : Dispersion autour du centre (Variance & écart-type)

Exemple 1 : Besoin mensuel des étudiants en volumes de données

x_i	n_i	$n_i x_i^2$	$n_i\left(x_i-\overline{x}\right)^2$
10	15	1500	7187.5815
20	17	6800	2403.3257
30	17	15300	60.7257
40	18	28800	1183.8978
50	23	57500	7543.3583
Σ	90	109900	18378.889

$$\sigma_x^2=\frac{1}{n}\sum_{i=1}^{5} n_i\left(x_i-\overline{x}\right)^2=\frac{18378.889}{90}\approx 204.21 \Rightarrow \sigma=\sqrt{\sigma_x^2}\approx 14.29 \text{ Go}$$

$$\sigma_x^2=\overline{x^2}-\overline{x}^2=\frac{109900}{90}-(31.89)^2\approx 204.14 \Rightarrow \sigma=\sqrt{\sigma_x^2}\approx 14.29 \text{ Go}$$

Les besoins mensuels des étudiants en volumes de donnée s'éloignent de la moyenne $\overline{x}=31.89$ Go de 14.29 Go en moyenne.

Exemple 2 : Besoin mensuel des étudiants en durées d'appels

$[x_i^-; x_i^-[$	c_i	n_i	$n_i c_i^2$	$n_i(c_i - \overline{x})^2$
$[0;4[$	2	20	80	698.562
$[4;8[$	6	24	864	87.5544
$[8;12[$	10	34	3400	148.5154
$[12;16[$	14	7	1372	259.6167
$[16;20]$	18	5	1620	509.0405
Σ	–	90	7336	1703.289

$$\sigma_x^2 = \frac{1}{n}\sum_{i=1}^{5} n_i\left(c_i - \overline{x}\right)^2 = \frac{1703.289}{90} \approx 18.93 \Rightarrow \sigma = \sqrt{\sigma_x^2} \approx 4.35 \text{ h}$$

$$\sigma_x^2 = \overline{c^2} - \overline{x}^2 = \frac{7336}{90} - (7.91)^2 \approx 18.94 \Rightarrow \sigma = \sqrt{\sigma_x^2} \approx 4.35 \text{ h}$$

Les besoins mensuels des étudiants en durées d'appels s'éloignent de la moyenne $\overline{x} = 7.91$ heures de 4.35 heures en moyenne.

Remarque

La variance n'a pas de signification concrète, il s'agit simplement d'une étape intermédiaire dans le calcul de l'écart-type. Néanmoins, elle peut être utilisée comme mesure de dispersion tout en prenant garde à bien interpréter l'unité de mesure transformée à la puissance 2 car ceci peut induire en erreur. A titre d'exemples, le mètre qui est une unité de mesure des longueurs se transforme en mètre carré qui est une unité de mesure des surfaces et le litre se transforme en litre carré qui n'a pas de signification concrète.

Propriétés

Soit X une variable statistique quantitative et a et b deux constantes réelles. La variance vérifie les propriétés suivantes :

$$V(aX) = a^2 V(X)$$

$$V(X + b) = V(X)$$

$$V(aX + b) = a^2 V(X)$$

Indicateurs statistiques

Mesures de dispersion : Dispersion autour du centre (Variance & écart-type)

Démonstrations

$$
\begin{aligned}
V(aX) &= \frac{1}{n}\sum_{i=1}^{k} n_i\left(ax_i - \overline{ax}\right)^2 = \frac{1}{n}\sum_{i=1}^{k} n_i\left(ax_i - a\overline{x}\right)^2 \\
&= \frac{1}{n}\sum_{i=1}^{k} n_i\left[a(x_i - \overline{x})\right]^2 = \frac{1}{n}\sum_{i=1}^{k} n_i a^2\left(x_i - \overline{x}\right)^2 \\
&= a^2\frac{1}{n}\sum_{i=1}^{k} n_i\left(x_i - \overline{x}\right)^2 = a^2 V(X)
\end{aligned}
$$

X	aX	n_i
x_1	ax_1	n_1
x_2	ax_2	n_2
$\vdots$	$\vdots$	$\vdots$
x_i	ax_i	n_i
$\vdots$	$\vdots$	$\vdots$
x_k	ax_k	n_k
Σ		n

Démonstrations (suite)

$$
\begin{aligned}
V(X+b) &= \frac{1}{n}\sum_{i=1}^{k} n_i\left(x_i + b - \overline{x+b}\right)^2 \\
&= \frac{1}{n}\sum_{i=1}^{k} n_i\left[x_i + b - (\overline{x} + b)\right]^2 \\
&= \frac{1}{n}\sum_{i=1}^{k} n_i\left(x_i + b - \overline{x} - b\right)^2 \\
&= \frac{1}{n}\sum_{i=1}^{k} n_i\left(x_i - \overline{x}\right)^2 = V(X)
\end{aligned}
$$

X	$X+b$	n_i
x_1	x_1+b	n_1
x_2	x_2+b	n_2
$\vdots$	$\vdots$	$\vdots$
x_i	x_i+b	n_i
$\vdots$	$\vdots$	$\vdots$
x_k	x_k+b	n_k
	Σ	n

Démonstrations (suite)

$$V(aX+b) = V(aX)$$
$$= a^2 V(X)$$

X	$aX+b$	n_i
x_1	ax_1+b	n_1
x_2	ax_2+b	n_2
$\vdots$	$\vdots$	$\vdots$
x_i	ax_i+b	n_i
$\vdots$	$\vdots$	$\vdots$
x_k	ax_k+b	n_k
	Σ	n

Exemples

$$V(2X) = 4V(X)$$

$$V(X+3) = V(X)$$

$$V(2X+3) = 4V(X)$$

Coefficient de variation

Le coefficient de variation, noté CV, est une mesure de dispersion relative, il est égal au rapport entre l'écart-type et la moyenne arithmétique, c'est un nombre sans dimension :

$$CV = \frac{\sigma}{\overline{x}}$$

Le coefficient de variation permet de comparer les dispersions de variables quantitatives n'ayant pas la même unité de mesure.

Exemples : Besoin mensuel des étudiants en volumes de données et en durées d'appels

$$\overline{x} = 31.39 \text{ Go}$$

$$\sigma_x = 14.29 \text{ Go}$$

$$CV_x = \frac{\sigma}{\overline{x}} = \frac{14.29 \cancel{\text{Go}}}{31.39 \cancel{\text{Go}}} \approx 0.4552$$

$$\overline{y} = 7.91 \text{ h}$$

$$\sigma_y = 4.35 \text{ h}$$

$$CV_y = \frac{\sigma}{\overline{y}} = \frac{4.35 \cancel{\text{h}}}{7.91 \cancel{\text{h}}} \approx 0.5499$$

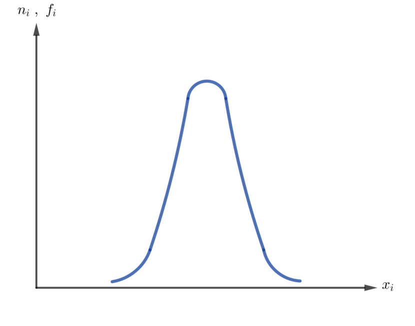
La distribution statistique des besoins mensuels des étudiants en durées d'appels est plus dispersée relativement à celle de leurs besoins mensuels en volumes de données.

Sous-section 3

Mesures de forme

Mesures de forme

Une *mesure de forme* indique le niveau de déviation d'une distribution statistique par rapport à une situation dite « normale » représentée par une courbe en cloche symétrique, ni pointue ni aplatie. Le degré de déviation est donné par une mesure d'*asymétrie* et par une mesure d'*aplatissement*.

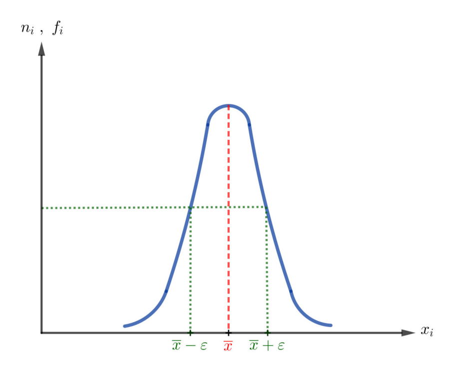


Coefficients d'asymétrie

- Coefficient de Fisher
- Coefficient de Pearson

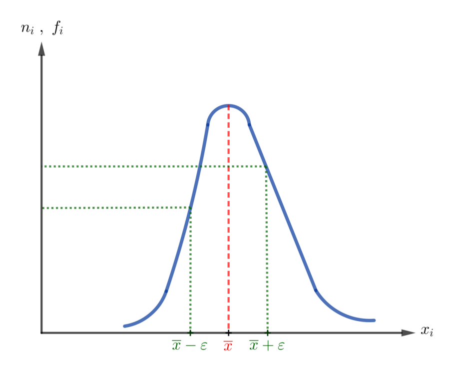
Distribution symétrique

Les valeurs symétriques par rapport au centre de la distribution statistique ont les mêmes effectifs et les mêmes fréquences.



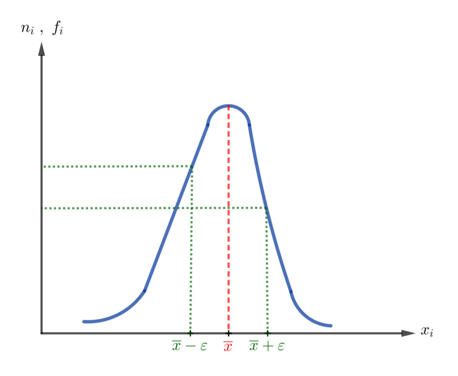
Distribution asymétrique à droite

Les valeurs supérieures au centre de la distribution statistique sont plus fréquentes relativement à leurs symétriques.



Distribution asymétrique à gauche

Les valeurs inférieures au centre de la distribution statistique sont plus fréquentes relativement à leurs symétriques.



Coefficient de Fisher

Coefficient de *skewness*

$$\gamma_1 = \frac{1}{n}\sum_{i=1}^{k} n_i \left(\frac{x_i - \overline{x}}{\sigma}\right)^3$$

Interprétation

- $\gamma_1 = 0$: la distribution est *symétrique* ;
- $\gamma_1 > 0$: la distribution est *asymétrique à droite* ;
- $\gamma_1 < 0$: la distribution est *asymétrique à gauche*.

Coefficient de Pearson

Coefficient de *skewness normalisé*

$$\beta_1 = \gamma_1^2 = \left[\frac{1}{n}\sum_{i=1}^{k} n_i \left(\frac{x_i - \overline{x}}{\sigma}\right)^3\right]^2$$

Interprétation :

- $\beta_1 = 0$: la distribution est *symétrique* ;
- $\beta_1 > 0$: la distribution est *asymétrique à droite* (si $\mu_3 > 0$) ou *à gauche* (si $\mu_3 < 0$) avec :

$$\mu_3 = \frac{1}{n}\sum_{i=1}^{k} n_i \left(x_i - \overline{x}\right)^3$$

Exemple 1 : Besoin mensuel des étudiants en volumes de données

x_i	n_i	$n_i\left(\frac{x_i-\overline{x}}{\sigma_x}\right)^3$
10	15	-33.75
20	17	-8.704
30	17	-0.017
40	18	3.888
50	23	50.531
Σ	90	11.948

$$\gamma_1 = \frac{1}{n}\sum_{i=1}^{5} n_i\left(\frac{x_i-\overline{x}}{\sigma_x}\right)^3 = \frac{11.948}{90} \approx 0.133$$

$$\beta_1 = \left[\frac{1}{n}\sum_{i=1}^{5} n_i\left(\frac{x_i-\overline{x}}{\sigma_x}\right)^3\right]^2 = (0.133)^2 \approx 0.018$$

La distribution des besoins mensuels des étudiants en volumes de données est asymétrique à droite. Cela signifie que les étudiants qui ont des besoins mensuels en volumes de données supérieurs à la moyenne sont plus fréquents que ceux qui ont des besoins inférieurs à la moyenne.

Exemple 2 : Besoin mensuel des étudiants en durées d'appels

$[x_i^-; x_i^+[$	c_i	n_i	$n_i\left(\frac{c_i-\overline{y}}{\sigma_y}\right)^3$
$[0;4[$	2	20	-50.4
$[4;8[$	6	24	-2.16
$[8;12[$	10	34	3.74
$[12;16[$	14	7	19.18
$[16;20]$	18	5	62.45
Σ	–	90	32.81

$$\gamma_1 = \frac{1}{n}\sum_{i=1}^{5} n_i\left(\frac{c_i-\overline{y}}{\sigma_y}\right)^3 = \frac{32.81}{90} \approx 0.36$$

$$\beta_1 = \left[\frac{1}{n}\sum_{i=1}^{5} n_i\left(\frac{c_i-\overline{y}}{\sigma_y}\right)^3\right]^2 = (0.36)^2 \approx 0.13$$

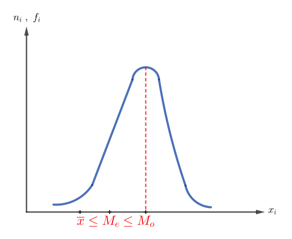
La distribution des besoins mensuels des étudiants en durées d'appels est asymétrique à droite. Cela signifie que les étudiants qui ont des besoins mensuels en durées d'appels supérieurs à la moyenne sont plus fréquents que ceux qui ont des besoins mensuels en durées d'appels inférieurs à la moyenne.

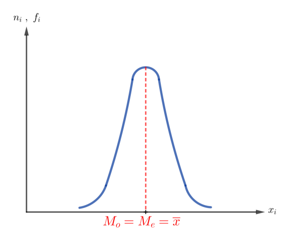
Indicateurs statistiques

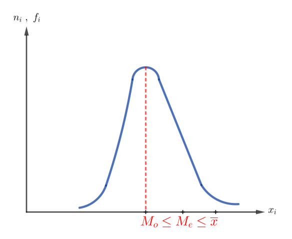
Mesures de forme : Mesures d'asymétrie

Remarque

Il y a un lien entre l'asymétrie d'une distribution statistique et l'ordre de ses mesures de tendance centrale. Celles-ci sont égales entre elles dans le cas d'une distribution symétrique et différentes dans le cas d'une distribution asymétrique et respectent un ordre bien précis suivant la nature de l'asymétrie :





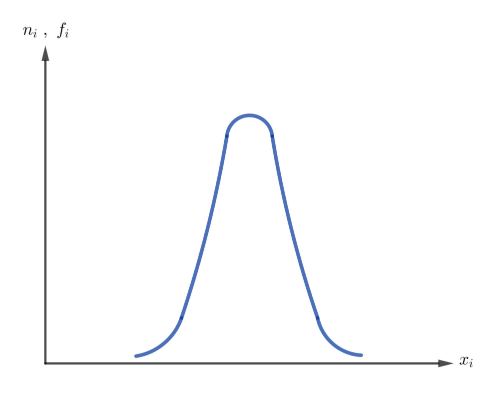


Coefficients d'aplatissement

- Coefficient de Pearson
- Coefficient de Fisher

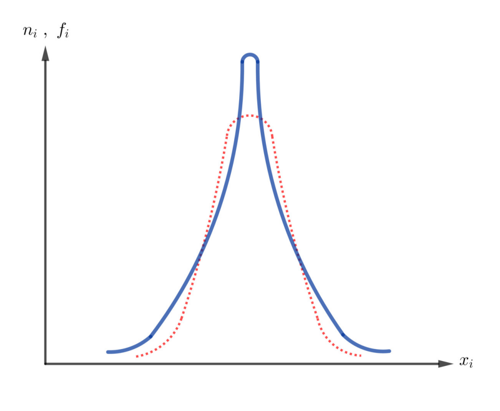
Distribution mésokurtique

Les valeurs extrêmes et les valeurs centrales ont des fréquences modérées.



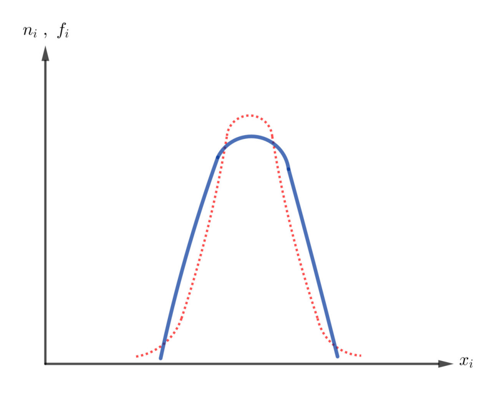
Distribution leptokurtique

Les valeurs extrêmes et les valeurs centrales sont plus fréquentes relativement à une distribution normale (en pointillés).



Distribution platykurtique

Les valeurs extrêmes et les valeurs centrales sont moins fréquentes relativement à une distribution normale (en pointillés).



Coefficient de Pearson

Coefficient de *kurtosis*

$$\beta_2 = \frac{1}{n}\sum_{i=1}^{k} n_i \left(\frac{x_i - \overline{x}}{\sigma}\right)^4$$

Interprétation

- $\beta_2 = 3$: distribution *mésokurtique* (courbe normale);
- $\beta_2 > 3$: distribution *leptokurtique* (moins aplatie que la normale);
- $\beta_2 < 3$: distribution *platykurtique* (plus aplatie que la normale).

Coefficient de Fisher

Coefficient d'*excès de kurtosis*

$$\gamma_2 = \beta_2 - 3 = \frac{1}{n}\sum_{i=1}^{k} n_i \left(\frac{x_i - \overline{x}}{\sigma}\right)^4 - 3$$

Interprétation

- $\gamma_2 = 0$: distribution *mésokurtique* (courbe normale);
- $\gamma_2 > 0$: distribution *leptokurtique* (moins aplatie que la normale);
- $\gamma_2 < 0$: distribution *platykurtique* (plus aplatie que la normale).

Exemple 1 : Besoin mensuel des étudiants en volumes de données

x_i	n_i	$n_i\left(\frac{x_i-\overline{x}}{\sigma_x}\right)^4$
10	15	50.625
20	17	6.96
30	17	0.0017
40	18	2.333
50	23	65.69
Σ	90	125.6097

$$\beta_2 = \frac{1}{n}\sum_{i=1}^{5} n_i\left(\frac{x_i-\overline{x}}{\sigma_x}\right)^4 = \frac{125.6097}{90} \approx 1.40$$

$$\gamma_2 = \frac{1}{n}\sum_{i=1}^{5} n_i\left(\frac{x_i-\overline{x}}{\sigma_x}\right)^4 - 3 = 1.40 - 3 = -1.6$$

La distribution des besoins mensuels des étudiants en volumes de données est platykurtique. Cela signifie que les valeurs extrêmes sont moins fréquentes relativement à une distribution normale.

Exemple 2 : Besoin mensuel des étudiants en durées d'appels

$[x_i^-; x_i^+[$	c_i	n_i	$n_i\left(\frac{c_i-\overline{x}}{\sigma_x}\right)^4$
$[0; 4[$	2	20	68.54
$[4; 8[$	6	24	0.95
$[8; 12[$	10	34	1.80
$[12; 16[$	14	7	26.85
$[16; 20]$	18	5	144.88
Σ	–	90	243.02

$$\beta_2 = \frac{1}{n}\sum_{i=1}^{5} n_i\left(\frac{c_i-\overline{y}}{\sigma_y}\right)^4 = \frac{243.02}{90} \approx 2.70$$

$$\gamma_2 = \frac{1}{n}\sum_{i=1}^{5} n_i\left(\frac{c_i-\overline{y}}{\sigma_y}\right)^4 - 3 = 2.70 - 3 = -0.30$$

La distribution des besoins mensuels des étudiants en durées d'appels est platykurtique. Cela signifie que les valeurs extrêmes sont moins fréquentes relativement à une distribution normale.

Sous-section 4

Mesures de concentration

1. Masse, masse cumulée et masse cumulée relative
2. Médiale et écart médiale-médiane
3. Courbe de Lorenz et Indice de Gini

Mesure de concentration

Une *mesure de concentration* permet de décrire (essentiellement) la distribution de la masse salariale entre des individus et de mesurer son niveau de concentration. Une faible concentration indique une répartition égalitaire des salaires : un certain pourcentage d'individus reçoit le même pourcentage de la masse salariale. En revanche, une forte concentration indique une répartition inégalitaire des salaires : un petit pourcentage d'individus reçoit un grand pourcentage de la masse salariale.

Masse, masse cumulée et masse cumulée relative

Masse & Masse totale

La *masse* m_i d'une valeur x_i d'une variable statistique positive X est le produit de la valeur x_i et de son effectif n_i, c'est à dire $m_i = n_i x_i$. La masse totale M est donnée par la somme des masses :

$$M = \sum_{i=1}^{k} m_i$$

Masse cumulée & Masse cumulée relative

La *masse cumulée* M_i d'une valeur x_i est donnée par : $M_i = m_1 + m_2 + \cdots + m_i$. Sa *masse cumulée relative* G_i est donnée par le rapport entre sa masse cumulée M_i et la masse totale M :

$$G_i = \frac{M_i}{M}$$

Exemple 1 : Besoin mensuel des étudiants en volumes de données

x_i	n_i	m_i	M_i	G_i
10	15	150	150	0.0523
20	17	340	490	0.1707
30	17	510	1000	0.3484
40	18	720	1720	0.5993
50	23	1150	2870	1
Σ	90	2870	–	–

Exemple 2 : Besoin mensuel des étudiants en durées d'appels

$[x_i^-; x_i^+[$	c_i	n_i	m_i	M_i	G_i
$[0; 4[$	2	20	40	40	0.0562
$[4; 8[$	6	24	144	184	0.2584
$[8; 12[$	10	34	340	524	0.7360
$[12; 16[$	14	7	98	622	0.8736
$[16; 20]$	18	5	90	712	1
Σ	–	90	712	–	–

Médiale et écart médiale-médiane

Médiale

La *médiale* M_l est la valeur du caractère positif X qui partage la série statistique ordonnée des masses m_i en deux parties contenant chacune la moitié de la masse totale M. La médiale est de ce fait une médiane calculée sur la série des masses m_i.

Écart médiale-médiane

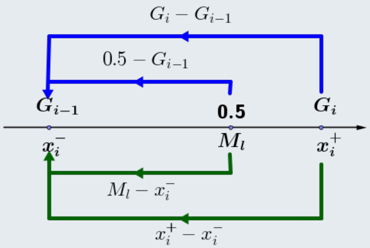
L'*écart médiale-médiane* ΔM est la différence entre la médiale et la médiane : $\Delta M = M_l - M_e$.

Interprétation

- Si $\Delta M = 0$, alors la concentration est nulle et la distribution est parfaitement égalitaire ;
- Si $\Delta M > E > 0$, alors la concentration est forte et la distribution est fortement inégalitaire ;
- Si $E > \Delta M > 0$, alors la concentration est faible et la distribution est faiblement inégalitaire.

Règle de calcul de la médiale

- Dans le cas d'une variable discrète, la médiale est la valeur qui a la masse cumulée relative immédiatement supérieure à 0.5 (sinon, on a un intervalle médial et la médiale en est le centre).
- Dans le cas d'une variable continue, identifier la classe médiale $[x_i^-; x_i^+[$, celle qui a une masse cumulée relative immédiatement supérieure à 0.5 puis calculer la médiale par interpolation linéaire (sinon, la médiale est la borne supérieure de la classe ayant une masse cumulée relative égale à 0.5):



$$\frac{M_l - x_i^-}{x_i^+ - x_i^-} = \frac{0.5 - G_{i-1}}{G_i - G_{i-1}}$$

$$M_l = x_i^- + \frac{0.5 - G_{i-1}}{G_i - G_{i-1}} \times (x_i^+ - x_i^-)$$

Exemple 1 : Besoin mensuel des étudiants en volumes de données

x_i	F_i^+	G_i
10	0.16	0.0523
20	0.35	0.1707
30	0.54	0.3484
40	0.74	0.5993
50	1	1

$M_e = 30$ Go

$M_l = 40$ Go

$\Delta M = 40 - 30 = 10$ Go

$E = 50 - 10 = 40$ Go

$\Delta M < E$: la concentration de la distribution est faible

Exemple 2 : Besoin mensuel des étudiants en durées d'appels

$[x_i^-; x_i^+[$	F_i^+	G_i
$[0; 4[$	0.22	0.0562
$[4; 8[$	0.49	0.2584
$[8; 12[$	0.87	0.7360
$[12; 16[$	0.94	0.8736
$[16; 20]$	1	1

$$M_e \in [8; 12[\Rightarrow M_e = 8 + \frac{0.5 - 0.49}{0.87 - 0.49} \times (12 - 8) \approx 8.11 \text{ h}$$

$$M_l \in [8; 12[\Rightarrow M_l = 8 + \frac{0.5 - 0.2584}{0.7360 - 0.2584} \times (12 - 8) \approx 10.02 \text{ h}$$

$$\Delta M = 10.02 - 8.11 = 1.91 \text{ h}$$

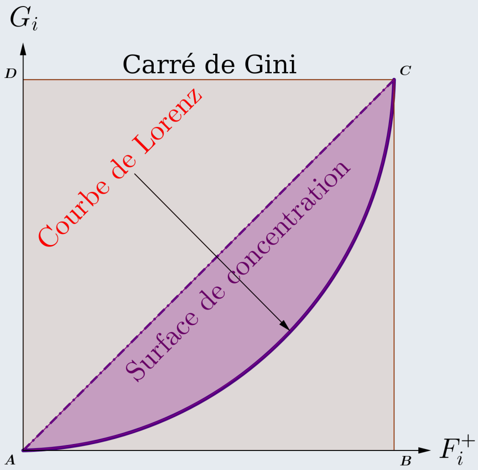
$$E = 20 - 0 = 20 \text{ h}$$

$\Delta M < E$: la concentration de la distribution est faible

Courbe de Lorenz et indice de Gini

Courbe de Lorenz

La *courbe de Lorenz* (ou *courbe de concentration*) met en relation graphiquement les fréquences cumulées croissantes F_i^+ d'une distribution statistique (en abscisses) et ses masses cumulées relatives G_i (en ordonnées). La courbe de Lorenz s'inscrit dans le *carré de Gini* ABCD. La surface entre la courbe de Lorenz et la diagonale $[AC]$ est appelée *surface de concentration*.



G_i
Carré de Gini
D
C
Courbe de Lorenz
Surface de concentration
A
B
F_i^+

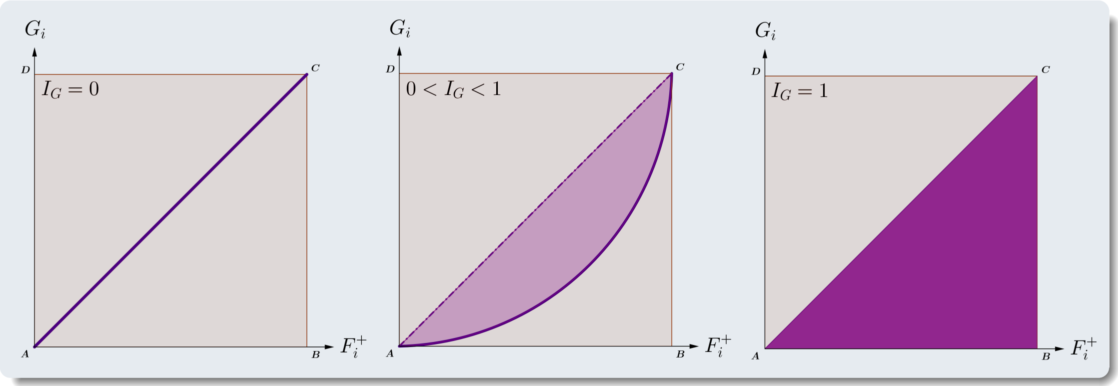
Indice de Gini

L'*indice de Gini* I_G est une mesure de la concentration d'une distribution statistique. Il est égal au double de l'aire de concentration et il est compris entre 0 (concentration nulle) et 1 (concentration complète):

$$I_G = \frac{\text{aire de concentration}}{\text{aire du triangle ABC}} = 2 \times \text{aire de concentration}$$

Interprétation

- Si I_G est nul, alors la concentration de la distribution est nulle ;
- Si I_G est proche de 0, alors la concentration de la distribution est faible ;
- Si I_G est proche de 1, alors la concentration de la distribution est forte ;
- Si I_G est égal à 1, alors la concentration de la distribution est maximale.



G_i
D
C
$I_G = 0$
A
B
F_i^+
G_i
D
C
$0 < I_G < 1$
A
B
F_i^+
G_i
D
C
$I_G = 1$
A
B
F_i^+

Application : Inégalités des salaires (en 1000 DH) dans une entreprise

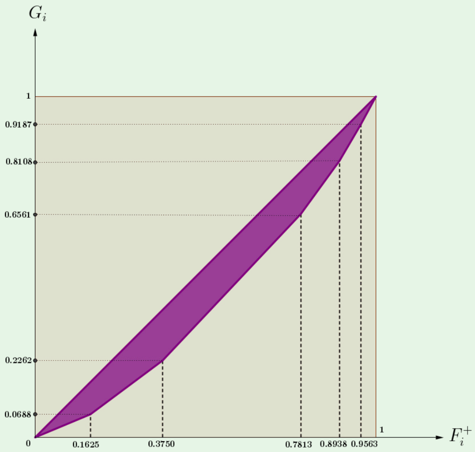
Salaires	c_i	n_i	f_i	F_i^+	$n_i c_i$	M_i	G_i
$[3;5[$	4	26	0.1625	0.1625	104	104	0.0688
$[5;9[$	7	34	0.2125	0.3750	238	342	0.2262
$[9;11[$	10	65	0.4063	0.7813	650	992	0.6561
$[11;15[$	13	18	0.1125	0.8938	234	1226	0.8108
$[15;17[$	16	10	0.0625	0.9563	160	1386	0.9187
$[17;19]$	18	7	0.0437	1	126	1512	1
Σ	–	160	1	–	1512	–	–

$$M_e \in [9;11[\Rightarrow M_e = 9 + \frac{0.5 - 0.3750}{0.7813 - 0.3750} \times (11 - 9) \approx 9615.3 \text{ DH}$$

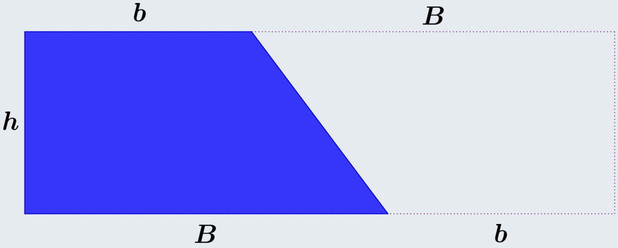
$$M_l \in [9;11[\Rightarrow M_l = 9 + \frac{0.5 - 0.2262}{0.6561 - 0.2262} \times (11 - 9) \approx 10273.8 \text{ DH}$$

$$\Delta M = 10273.8 - 9615.3 = 658.5 \text{ DH} \qquad E = 19 - 3 = 16000 \text{ DH}$$

Application (suite): Inégalités des salaires (en 1000 DH) dans une entreprise



Rappel : Aire d'un trapèze



- Aire du rectangle : $h(B+b)$
- Aire du trapèze : $A = h(B+b)/2$

Application (suite): Inégalités des salaires (en 1000 DH) dans une entreprise

Trapèze i	h_i	b_i	B_i	A_i
Trapèze 1	0.1625	0	0.0688	0.0056
Trapèze 2	0.2125	0.0688	0.2262	0.0313
Trapèze 3	0.4063	0.2262	0.6561	0.1792
Trapèze 4	0.1125	0.6561	0.8108	0.0825
Trapèze 5	0.0625	0.8108	0.9187	0.0540
Trapèze 6	0.0437	0.9187	1	0.0419
Σ	–	–	–	0.3945

- Aire du trapèze i : $A_i = h_i(B_i + b_i)/2$
- Aire de concentration : $0.5 - \Sigma_{i=1}^{6} A_i = 0.5 - 0.3945 = 0.1055$
- Indice de Gini : $I_G = 2 \times$ Aire de concentration $= 0.211$

La concentration des salaires dans cette entreprise est faible. La distribution des salaires dans cette entreprise est en conséquence faiblement inégalitaire.