§ 2 Statistique descriptive pour une variable statistique continue: distribution empirique continue

■ Objectifs

Pour chaque notion étudiée (moyenne, médiane, écart-type, ...), le lecteur doit se préoccuper de savoir la calculer

1° à partir de données brutes, sans ordinateur;
2° à partir de données groupées, sans ordinateur;
3° à partir de données brutes, avec *Mathematica;*
4° à partir de données groupées, avec *Mathematica*.

■ Packages locaux

A partir de la commande **aide[]**, charger les *packages locaux*

```
<< Statistique`
```

```
<< Tableaux`
```

§ 2.1 *Distribution empirique continue et fonction de densité*

■ Données groupées en classes

D'un échantillon d'étudiants de sexe masculin, on a mesuré la masse de chacun. Les masses ont été arrondies à l'entier. Voici les données groupées en 7 classes:

Masses en kg	Nombre d'étudiants
45 - 54	5
55 - 59	14
60 - 64	33
65 - 69	47
70 - 74	26
75 - 79	13
80 - 89	2

La variable aléatoire X = *masse corporelle de l'étudiant* est continue.
Pour une variable continue, les effectifs sont associés non à une valeur mais à un intervalle.

L'intervalle	45 - 54	représente plus précisément l'intervalle	[44.5; 54.5[,
	55 - 59		[54.5; 59.5[,
	60 - 64		[59.5; 64.5[,
	65 - 69		[64.5; 69.5[,
	70 - 74		[69.5; 74.5[,
	75 - 79		[74.5; 79.5[,
	80 - 89		[79.5; 89.5[.

On définit ainsi une liste des <u>bornes des classes</u> :

$b_0, b_1, b_2, \ldots, b_k$

```
b = {44.5, 54.5, 59.5, 64.5, 69.5, 74.5, 79.5, 89.5};
```

Remarquez qu'il y a k classes mais (k+1) bornes.

On calcule ensuite les centres des classes

$$c_1 = \frac{b_0 + b_1}{2}, \quad c_2 = \frac{b_1 + b_2}{2}, \quad \ldots, \quad c_k = \frac{b_{k-1} + b_k}{2}$$

$$\boxed{c_j = \frac{b_{j-1} + b_j}{2}}$$

```
c = (Drop[b, 1] + Drop[b, -1])/2
```

$$c = \frac{\text{Drop}[b, 1] + \text{Drop}[b, -1]}{2}$$

```
{49.5, 57., 62., 67., 72., 77., 84.5}
```

```
effectifs = {5, 14, 33, 47, 26, 13, 2};
```

Nombre de classes

```
k = Length[effectifs]
```

```
7
```

Taille de l'échantillon

```
n = Apply[Plus, effectifs]
```

```
140
```

Fréquences

```
freq = effectifs/n
```

$$freq = \frac{effectifs}{n}$$

$$\left\{\frac{1}{28}, \frac{1}{10}, \frac{33}{140}, \frac{47}{140}, \frac{13}{70}, \frac{13}{140}, \frac{1}{70}\right\}$$

■ Répartition uniforme de la fréquence par classe

Considérer que les effectifs sont concentrés au centre des classes est déconseillé. La variable statistique étant continue, nous désirons avoir une *distribution continue*. Pour y parvenir, nous allons répartir les effectifs uniformément dans chaque classe.

Comment répartir les 5 étudiants dans la classe [44.5; 54.5[? Subdivisons l'intervalle en 5 intervalles partiels égaux
[44.5; 46.5[, [46.5; 48.5[, [48.5; 50.5[, [50.5; 52.5[, [52.5; 54.5[
puis attribuons un étudiant à chaque intervalle partiel. Nous dirons que

la fréquence de l'intervalle [44.5; 54.5[est de $\frac{5}{140}$

tandis que, pour chaque intervalle partiel,

la fréquence de l'intervalle [44.5; 46.5[est de $\frac{1}{140}$;

la fréquence de l'intervalle [46.5; 48.5[est de $\frac{1}{140}$;

la fréquence de l'intervalle [48.5; 50.5[est de $\frac{1}{140}$;

la fréquence de l'intervalle [50.5; 52.5[est de $\frac{1}{140}$;

la fréquence de l'intervalle [52.5; 54.5[est de $\frac{1}{140}$.

Pour rendre la distibution continue, nous acceptons de fractionner encore ces intervalles partiels

la fréquence de l'intervalle [44.5; 45.5[est de $\frac{1}{280}$,

ce que nous interprétons comme suit:

en choisissant une personne au hasard, il y a une chance sur 280 que son poids
se situe dans l'intervalle [44.5; 45.5[.

Plus généralement, à chaque classe est associée une fréquence

$$f_1 = \frac{n_1}{n} = f\left[b_0;\ b_1\right[= \text{fréquence de la classe numéro 1}$$

$$\ldots$$

$$f_j = \frac{n_j}{n} = f\left[b_{j-1};\ b_j\right[= \text{fréquence de la classe numéro } j$$

$$\ldots$$

$$f_k = \frac{n_k}{n} = f\left[b_{k-1};\ b_k\right[= \text{fréquence de la classe numéro } k$$

$$\sum_{j=1}^{k} f_j = 1$$

Dans notre exemple, la fréquence de la 4-ème classe - c'est-à-dire la fréquence de l'événement "la masse appartient à l'intervalle [64.5; 69.5[" - est de $\frac{47}{140}$, ce que l'on peut noter

$$f_4 = f\left[64.5;\ 69.5\right[= \frac{47}{140}$$

La signification de l'événement "*la masse est de 66 k*g" doit être précisée. Déterminons sa fréquence. S'il s'agit de f([65.5; 66.5[) on peut estimer sa valeur *en répartissant les effectifs uniformément sur toute la largeur de la classe*

$$f\left[65.5;\ 66.5\right[= \frac{66.5 - 65.5}{69.5 - 64.5} f\left[64.5;\ 69.5\right[= \frac{1}{5} f_4 = \frac{47}{700}$$

S'il s'agit de l'événenement "la masse vaut exactement un 66 kg", commençons par de petits intervalles autour de 66 :

$$f\left[65.95;\ 66.05\right[= \frac{66.05 - 65.95}{69.5 - 64.5} f\left[64.5;\ 69.5\right[= \frac{0.1}{5} f_4 = \frac{47}{7000}$$

$$f\left[65.995;\ 66.005\right[= \frac{66.005 - 65.995}{69.5 - 64.5} f\left[64.5;\ 69.5\right[= \frac{0.01}{5} f_4 = \frac{47}{70\,000}$$

En prenant une suite d'intervalles emboîtés dont les largeurs tendent vers 0, on peut conclure que *la fréquence d'un événement réduit à un point est nulle:*

$$f\ (\{66\}) = 0$$

On peut intuitivement interpréter ce dernier résultat comme suit : "Il n'y a quasiment aucune chance pour qu'une personne pèse <u>exactement</u> 66.000 000 000 ... kg.

Généralisons. La fréquence d'un événement A peut être écrite sous la forme

$$\boxed{f\ (A) = \frac{\text{effectif de } A}{\text{effectif total}} = \frac{n_A}{n}}$$

A l'intérieur d'une classe, lorsqu'on distribue la fréquence uniformémement sur toute la largeur de la classe, on a

$$\boxed{b_{j-1} \le x < b_j \quad \Longrightarrow \quad f\left[b_{j-1};\ x\right[= \frac{x - b_{j-1}}{b_j - b_{j-1}} f_j}$$

■ Propriétés de la fréquence d'un événement

Les événements sont des <u>intervalles</u> de nombres réels, ou des réunions et intersections d'intervalles.

En particulier, on a

$$0 \le f\ (A) \le 1$$

$$f\ (\emptyset) = \frac{0}{n} = 0 \qquad \text{fréquence de l'événement impossible;}$$

$$f\ (\mathbb{R}) = \frac{n}{n} = 1 \qquad \text{fréquence de l'événement certain.}$$

Deux événements disjoints (c'est-à-dire tels que $A \cap B = \varnothing$) sont appelés incompatibles.
La fonction fréquence est additive. Par exemple, pour deux événements incompatibles

$$f\left(\left[59.5;\ 64.5\right[\bigcup \left[64.5;\ 69.5\right[\right) = f\left[59.5;\ 69.5\right[= \frac{33+47}{140} = \frac{33}{140} + \frac{47}{140} = f\left[59.5;\ 64.5\right[+ f\left[64.5;\ 69.5\right[$$

Plus généralement,

$$\boxed{A \cap B = \varnothing \implies f(A \cup B) = f(A) + f(B)}$$

En conséquence, lorsqu'on calcule la fréquence d'un intervalle, il importe peu que l'intervalle soit ouvert ou fermé

$$f[a;\ b] = f\{a\} + f]a;\ b[+ f\{b\} = 0 + f]a;\ b[+ 0 = f]a;\ b[$$

■ La fréquence cumulée comme fonction de distribution empirique continue

Calculons d'abord la fréquence cumulée aux bornes des classes

```
freqCum = Accumulate[freq]
```

$$\left\{\frac{1}{28}, \frac{19}{140}, \frac{13}{35}, \frac{99}{140}, \frac{25}{28}, \frac{69}{70}, 1\right\}$$

en insérant la fréquence 0 au début de la liste

```
afficheTableau[{"Bornes des classes", "Fréquences cumulées"},
 None, {b, Prepend[freqCum, 0]}]
```

Bornes des classes	44.5	54.5	59.5	64.5	69.5	74.5	79.5	89.5
Fréquences cumulées	0	$\frac{1}{28}$	$\frac{19}{140}$	$\frac{13}{35}$	$\frac{99}{140}$	$\frac{25}{28}$	$\frac{69}{70}$	1

La fréquence cumulée est

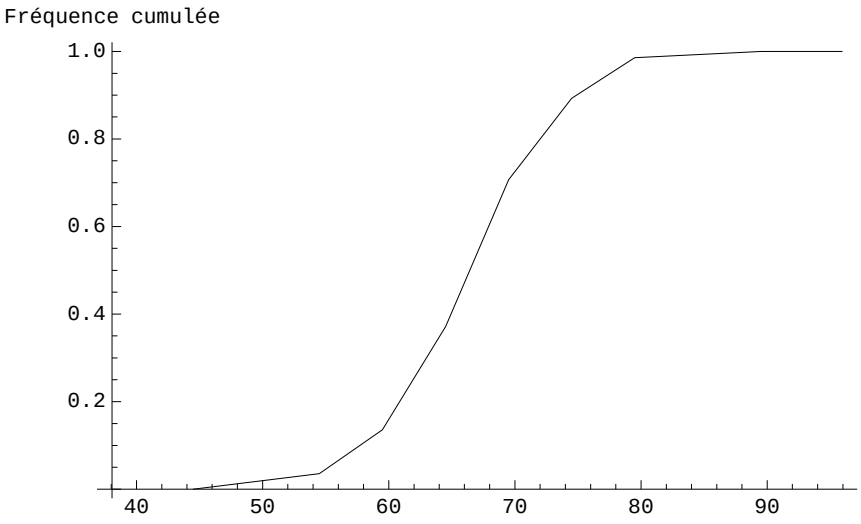
$$F(x) = f]-\infty,\ x] = \text{fréquence des modalités} \leq x$$

```
afficheTableau[{"x", "F(x)"}, None, {b, Prepend[freqCum, 0]}]
```

x	44.5	54.5	59.5	64.5	69.5	74.5	79.5	89.5
F(x)	0	$\frac{1}{28}$	$\frac{19}{140}$	$\frac{13}{35}$	$\frac{99}{140}$	$\frac{25}{28}$	$\frac{69}{70}$	1

Pour satisfaire l'hypothèse que les effectifs sont uniformément distribués dans chaque classe, on interpole linéairement entre ces points. On obtient ainsi une fonction F qui est *continue* et *affine par morceaux*.

```
frequenceCumuleeContinue[b, freq, AxesLabel → {None, "Fréquence cumulée"}]
```



La *fréquence cumulée* (on dit aussi <u>fonction de répartition</u> empirique) est une fonction

$$F : \mathbb{R} \longrightarrow \mathbb{R}$$

qui vérifie

$$\text{pour les } x \in \{b_0,\ b_1,\ \ldots,\ b_k\},\ \text{on a}$$

$$F(x) = \frac{\text{nombre d'éléments inférieurs ou égaux à } x}{\text{effectif total}} = f\left]-\infty;\ x\right]$$

c'est-à-dire

$$\begin{array}{lcccc}
F(b_0) & = & 0 & = & f]-\infty;\ b_0] \\
F(b_1) & = & f_1 & = & f]-\infty;\ b_1] \\
F(b_2) & = & f_1 + f_2 & = & f]-\infty;\ b_2] \\
\ldots & . & \ldots & . & \ldots \\
F(b_{k-1}) & = & f_1 + f_2 + \ldots + f_{k-1} & = & f]-\infty;\ b_{k-1}] \\
F(b_k) & = & 1 & = & f]-\infty;\ \infty[
\end{array}$$

$$\begin{array}{lcccccc}
\text{Pour } x \le b_0, & F(x) & = & f]-\infty;\ b_0] & = & 0 \\
\text{Pour } b_0 \le x \le b_1, & F(x) & = & f]-\infty;\ b_0] + f]b_0;\ x] & = & F(b_0) + \frac{x-b_0}{b_1-b_0} f_1 \\
\text{Pour } b_1 \le x \le b_2, & F(x) & = & f]-\infty;\ b_1] + f]b_1;\ x] & = & F(b_1) + \frac{x-b_1}{b_2-b_1} f_2 \\
\ldots & .. & . & \ldots & . & \ldots \\
\text{Pour } b_{k-1} \le x \le b_k, & F(x) & = & f]-\infty;\ b_{k-1}] + f]b_{k-1};\ x] & = & F(b_{k-1}) + \frac{x-b_{k-1}}{b_k-b_{k-1}} f_k \\
\text{Pour } b_k \le x & F(x) & = & f]-\infty;\ \infty[& = & 1
\end{array}$$

■ Relation entre distribution empirique et fréquence

La valeur de la fonction de distribution empirique en x est égale à la fréquence de l'intervalle $]-\infty;\ x]$

$$\boxed{\text{Pour tout } x \in \mathbb{R},\ \text{on a} \qquad F(x) = f]-\infty;\ x]}$$

En d'autres termes, $F(x)$ représente la fréquence de l'événenent "*être inférieur ou égal à x*".
F est aussi appelé fonction *fréquence cumulée continue.*

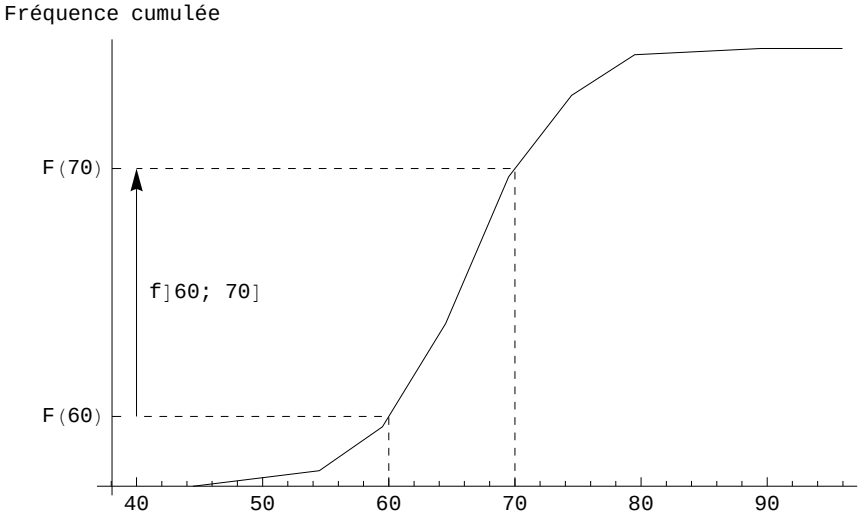
On remarquera que la distribution F est une fonction d'une variable réelle dont la représentation graphique est immédiate.

Il n'en est pas de même de la fonction fréquence f dont l'argument n'est pas un nombre réel mais un événement. La fréquence d'un intervalle peut se déduire de la distribution

$$\boxed{f(]a,\ b]) = F(b) - F(a)}$$

Plus simplement - mais abusivement - on note aussi $f]a,\ b]$. Dans notre exemple numérique,

$$f]\,60,\ 70] = F(70) - F(60)$$



■ La densité de fréquence

La fréquence cumulée étant une fonction affine par morceaux, intéressons-nous à la *pente* de chacun des morceaux. Situons-nous à l'intérieur de la classe numéro j et désignons la pente par h_j.

$$h_j = \frac{F(b_j) - F(b_{j-1})}{b_j - b_{j-1}} = \frac{f_j}{b_j - b_{j-1}} = \frac{f]\,b_{j-1};\ b_j]}{b_j - b_{j-1}}$$

Le nombre h_j a une double interprétation

* d'une part,

$$\boxed{h_j = \frac{f_j}{b_j - b_{j-1}} = \frac{f]\,b_{j-1};\ b_j]}{b_j - b_{j-1}}}$$

représente *la fréquence divisée par l'amplitude de la classe* d'où le nom de *densité de fréquence;*

en mots, $\text{Densité de la classe j} = \frac{\text{fréquence de la classe j}}{\text{amplitude de la classe j}}$

* d'autre part,

$$\boxed{h_j = \frac{F(b_j) - F(b_{j-1})}{b_j - b_{j-1}}}$$

représente la *pente* moyenne de la fonction F sur la classe numéro j.

```
h = densites[b, freq]
```

```
{0.00357143, 0.02, 0.0471429, 0.0671429, 0.0371429, 0.0185714, 0.00142857}
```

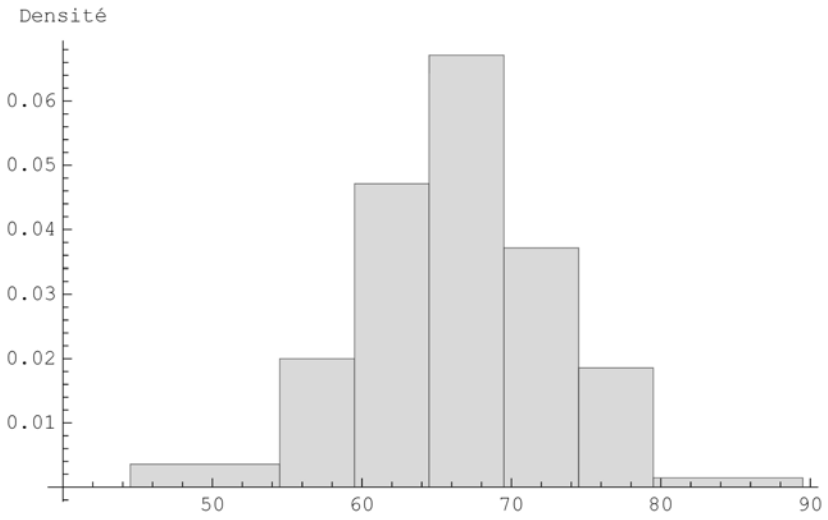
Ces k nombres permettent de définir une fonction $h : \mathbb{R} \longrightarrow \mathbb{R}$ constante par morceaux

$$\begin{array}{lllllll}
\text{Pour } x \le b_0, & h(x) & = & 0 & = & 0 \\
\text{Pour } b_0 \le x \le b_1, & h(x) & = & h_1 & = & \frac{f_1}{b_1 - b_0} \\
\text{Pour } b_1 \le x \le b_2, & h(x) & = & h_2 & = & \frac{f_2}{b_2 - b_1} \\
\ldots & \ldots & . & \ldots & . & \ldots \\
\text{Pour } b_{k-1} \le x \le b_k, & h(x) & = & h_k & = & \frac{f_k}{b_k - b_{k-1}} \\
\text{Pour } b_k \le x & h(x) & = & 0 & = & 0
\end{array}$$

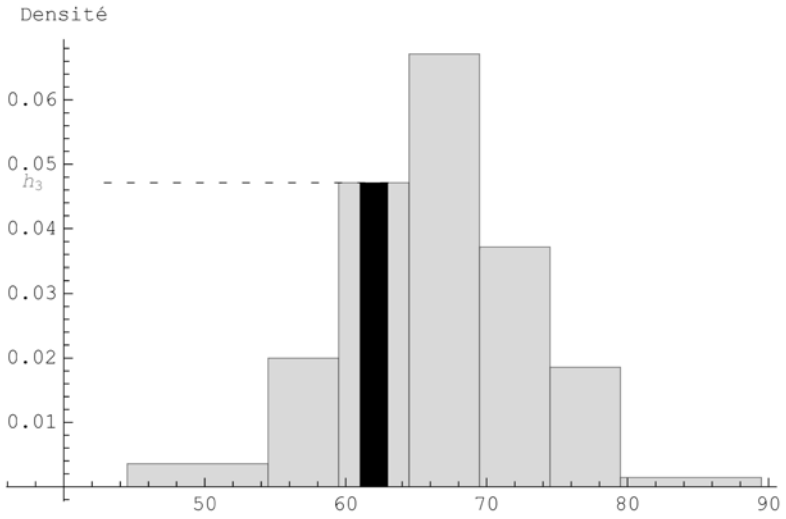
Histogramme

Nous verrons ci-après que la surface comprise entre l'axe des abscisses et le graphique de la fonction $h(x)$ = *densité de fréquence* a une signification statistique importante. C'est la raison pour laquelle on représente le graphique de la fonction *h* avec des rectangles et que l'on grise ces rectangles. Cette représentation est appelée *histogramme*:

```
histogramme[b, freq, AxesOrigin → {40, 0}, AxesLabel → {None, "Densité"}]
```



Dans le but d'interpréter le graphique précédent, répondons à la question : "Quelle est la fréquence de l'événement *la masse appartient à l'intervalle [61; 63[* ?".



L'axe vertical est l'axe des densités; il est gradué en *fréquence par kg*

$$h_3 = \frac{f_3}{5\ \text{kg}} = \frac{33}{700}\ \frac{1}{\text{kg}}$$

Conformément à la relation *fréquence* = *densité* * *amplitude*, on a

$$f\,]\,61;\ 63\,] = h_3 * (63\ \text{kg} - 61\ \text{kg}) = \frac{33}{700\ \text{kg}}\ 2\ \text{kg} = \frac{66}{700}$$

Du point de vue géométrique, pour le rectangle marqué en noir,

aire = *hauteur* * *largeur*

L'aire en noir représente donc la fréquence de l'événement *la masse appartient à l'intervalle [61; 63[*.

Retenons le résultat suivant. Dans un histogramme,

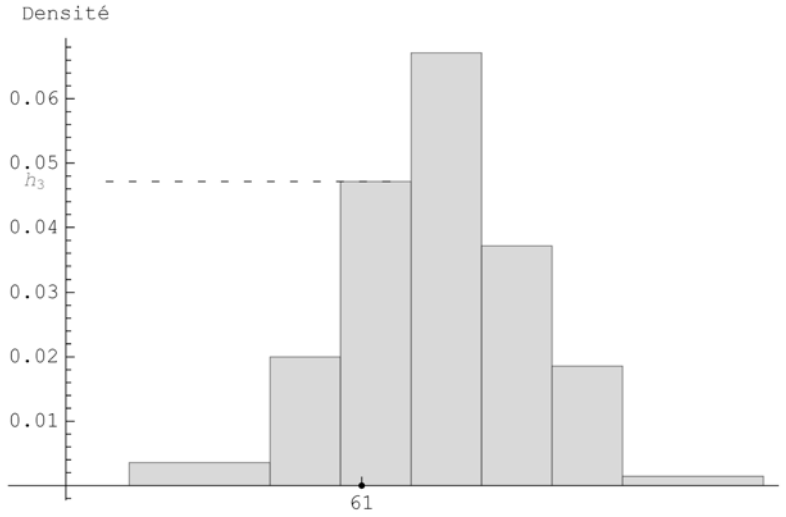
*** les *hauteurs* des rectangles représentent des *densités*;**

*** les *aires* des rectangles représentent des *fréquences*.**

En particulier, la somme des aires d'un histogramme est égale à 1.

Selon cette interprétation,

$$f\{61\} = f]61;\ 61] = \text{aire du rectangle de hauteur } h_3 \text{ et de largeur } 0$$



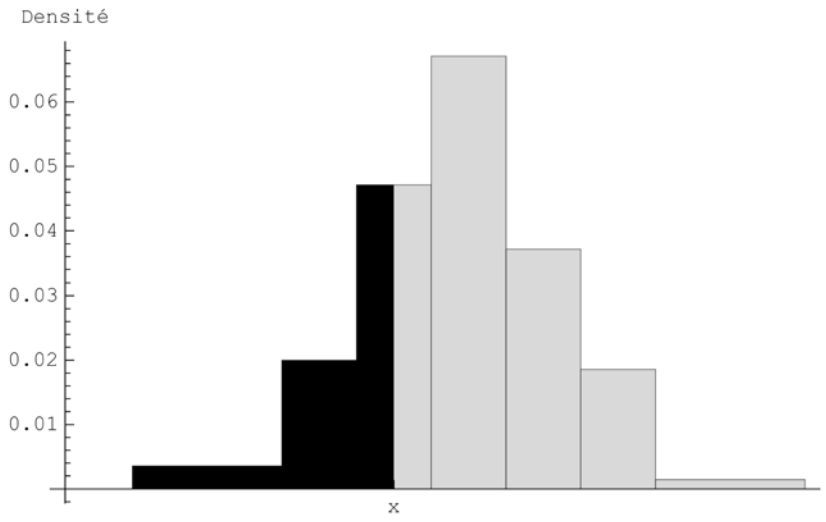
Il s'ensuit que

$$f\{61\} = 0$$
$$f[61;\ 63] = f]61;\ 63[= f[61;\ 63[$$

■ Relation entre distribution et densité

La valeur de la fonction de distribution en x est égale à l'aire de la surface délimitée par la fonction densité de fréquence au-dessus de l'intervalle $]-\infty,\ x]$.

Dans la figure ci-dessous, l'aire de la surface noire est égale à $F(x)$.

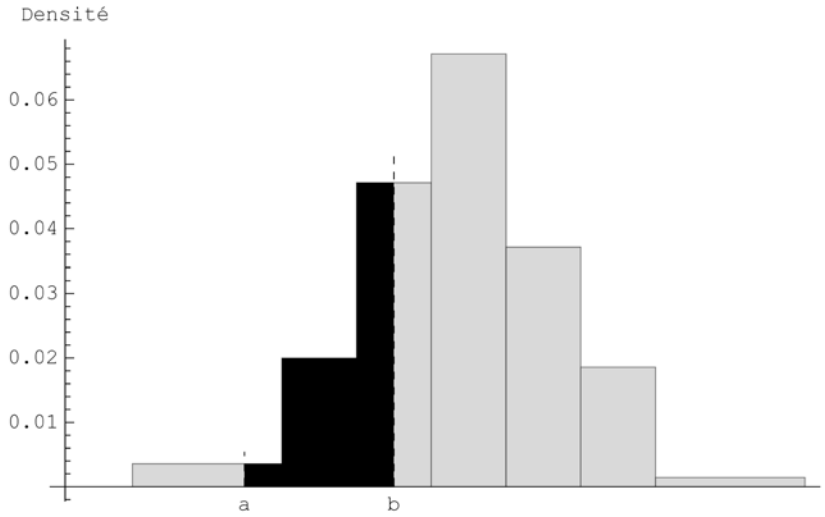


■ Relation entre fréquence et distribution

La fréquence d'un intervalle est égale à l'accroissement de la fonction de distribution sur cet intervalle

$$\boxed{f(]a,\ b]) = F(b) - F(a)}$$

Graphiquement, la fréquence de l'intervalle $]a,\ b]$ est représentée par l'aire de la surface délimitée par la fonction densité de fréquence au-dessus de l'intervalle $]a,\ b]$ (voir figure ci-dessous):



La fonction de distribution empirique contient toutes les informations sur la manière dont on se représente les données.

■ Relation entre densité et fréquence

A l'intérieur d'une classe (disons dans la classe numéro j), c'est-à-dire pour

$$]a,\ b] \subset]b_{j-1},\ b_j]$$

on a

$$f]a,\ b] = h_j\ (b - a)$$

En remplaçant "densité constante" par "densité moyenne", on peut généraliser la formule précédente à n'importe quels nombres a, b vérifiant $a \leqslant b$

$$f]a,\ b] = \bar{h}\ (b - a)$$

où $\bar{h}$ = <u>densité moyenne</u> sur l'intervalle $[a,\ b]$. Donc,

$$\boxed{\bar{h}_{[a,b]} = \frac{f]a,\ b]}{b - a}}$$

En mots:

$$(\text{densité moyenne sur l'intervalle}) = \frac{\text{fréquence de l'intervalle}}{\text{amplitude de l'intervalle}}$$

■ Relation entre densité et fonction de distribution

A l'intérieur d'une classe (disons dans la classe numéro j), c'est-à-dire pour

$$]a,\ b] \subset]b_{j-1},\ b_j]$$

on a

$$F(b) - F(a) = h_j\ (b - a)$$

En remplaçant "densité constante" par "densité moyenne", on peut généraliser la formule précédente à n'importe quels

nombres a, b vérifiant $a \leqslant b$

$$F(b) - F(a) = \bar{h}\,(b - a)$$

où $\bar{h}$ = densité moyenne sur l'intervalle $[a,\ b]$. Donc,

$$\boxed{\bar{h}_{[a,b]} = \frac{F(b) - F(a)}{b - a}}$$

En mots:
la densité moyenne sur l'intervalle $[a,\ b]$ est égale à la pente moyenne de la fonction F sur l'intervalle
(on dit aussi le taux d'accroissement de F sur l'intervalle $[a,\ b]$).

Densité ponctuelle (prolongement pour lecteurs avertis)

Sur le site http://www.collegedusud.ch/app/applmaths
suivez les liens **Documents *Mathematica* / Annexes: Statistiques I**
et téléchargez le cahier **2-1_densite_ponctuelle.nb**

■ Moyenne

Pour calculer la moyenne de données groupées en classes, on peut utiliser la formule suivante dans laquelle les c_j désignent les centres des classes et les f_j les fréquences des classes correspondantes

$$\boxed{m = \bar{x} = \sum_{j=1}^{k} c_j\, f_j}$$

Cette formule est exacte alors même que l'on suppose que la densité est uniforme dans chaque classe.

```
m = c.freq
```

```
66.3036
```

■ Ecart-type

Pratiquement, on peut utiliser la formule approximative suivante

$$s \approx \sqrt{\sum_{j=1}^{k} (c_j - m)^2\, f_j}$$

Cette formule n'est pas tout à fait exacte car elle suppose que les effectifs sont concentrés aux centres des classes alors que les densités sont uniformes dans chaque classe.

```
s = √((c - m)^2.freq)
```

```
6.68198
```

■ Classe modale

On appelle *classe modale* la classe dont la densité est maximale. (Attention : il ne s'agit pas nécessairement de la classe dont la fréquence est maximale !)

Dans notre exemple, la classe modale est l'intervalle [64.5; 69.5[.
Dans le cas où plusieurs classes sont de densité maximale, on dit que la distribution est *multimodale*.

Avec *Mathematica*,

```
h = densites[b, freq]
```

{0.00357143, 0.02, 0.0471429, 0.0671429, 0.0371429, 0.0185714, 0.00142857}

```
clMod = Flatten[Position[h, Max[h]]]
```

{4}

Usuellement, le milieu de la classe modale est appelé mode:

```
mo = c[[4]]
```

67.

■ Médiane (ou quantile interpolé $\frac{1}{2}$)

La médiane est le nombre me = $Q_{\frac{1}{2}}$ tel que $F(\text{me}) = \frac{1}{2}$. Le quantile $\frac{1}{2}$ auquel il correspond est dit *interpolé* car la fonction F est construite par interpolation linéaire.

Calcul sans ordinateur

```
b
```

{44.5, 54.5, 59.5, 64.5, 69.5, 74.5, 79.5, 89.5}

```
N[freqCum]
```

{0.0357143, 0.135714, 0.371429, 0.707143, 0.892857, 0.985714, 1.}

x	64.5	me	69.5
$F(x)$	0.371429	0.5	0.707143

$$\frac{0.5 - 0.371429}{\text{me} - 64.5} = \frac{0.707143 - 0.371429}{69.5 - 64.5}$$

$$\frac{0.128571}{\text{me} - 64.5} = \frac{0.335714}{5}$$

$$(\text{me} - 64.5)\ 0.335714 = 0.642855$$

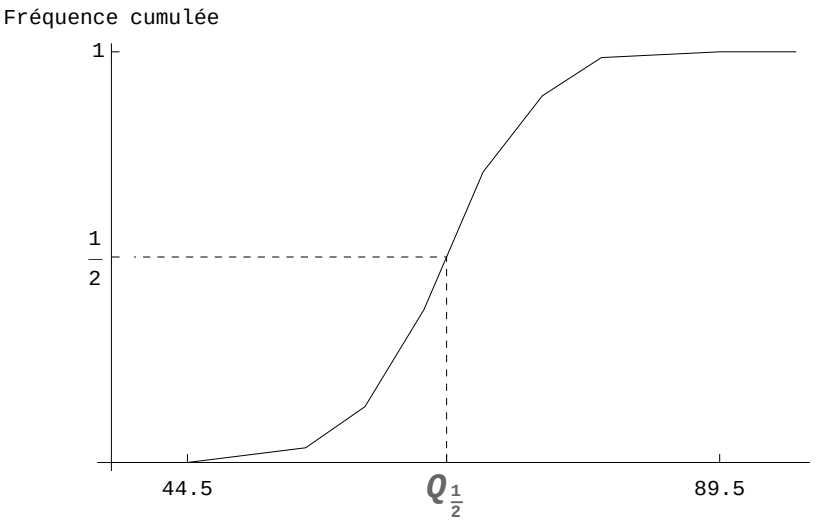
$$\text{me} - 64.5 = \frac{0.642855}{0.335714}$$

$$\text{me} = 64.5 + \frac{0.642855}{0.335714} = 66.4149$$

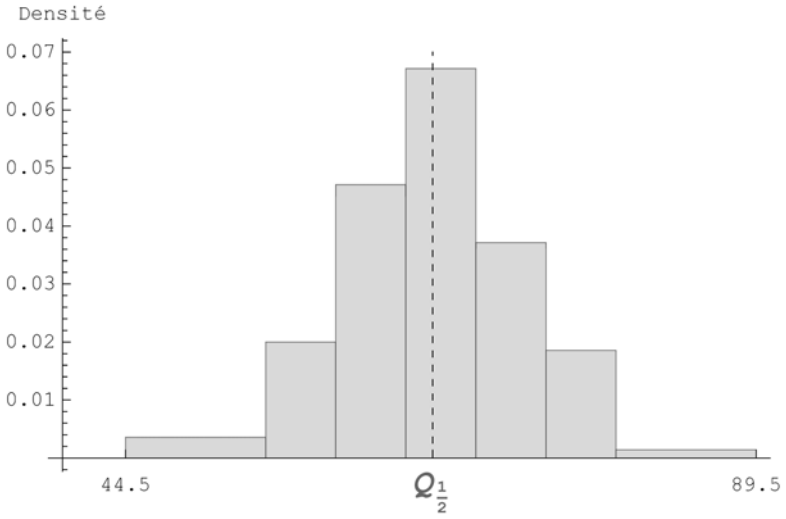
Calcul avec Mathematica

```
me = quantileC[b, freq, 1/2]
```

66.4149

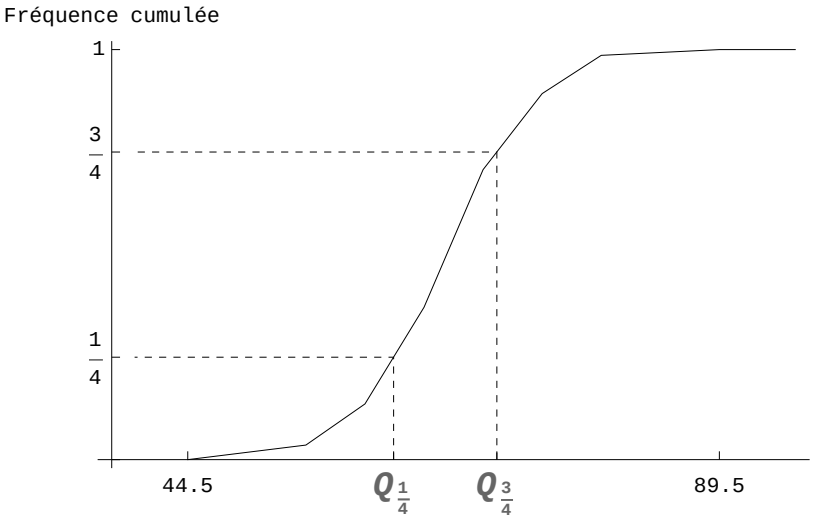


La médiane partage l'histogramme en deux parties d'aires égales.



■ Intervalle interquartile

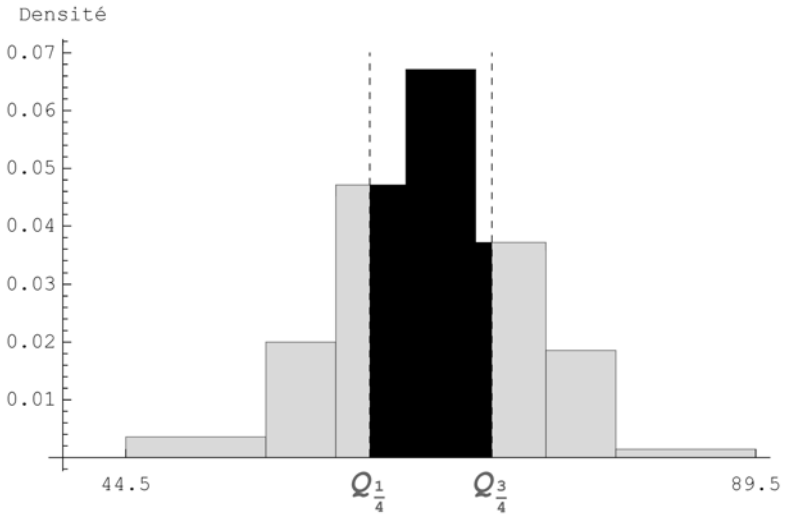
L'intervalle interquartile représente la différence entre les quantiles $\frac{3}{4}$ et $\frac{1}{4}$ (voir la figure qui suit):



```
interQuartile = quantileC[b, freq, 3/4] - quantileC[b, freq, 1/4]
```

```
8.7296
```

Il s'ensuit que, dans l'intervalle interquartile se situe exactement 50 % de l'effectif total.



■ Exercice 2 - 1 [Sans ordinateur]

Dans une ferme, à une date déterminée, on a pesé les oeufs qui ont été produits:

Masse de l'oeuf [g]	*Nombre d'oeufs*
28 – 37	3
38 – 47	51
48 – 52	74
53 – 57	112
58 – 62	92
63 – 72	62
73 – 82	6

a) Calculez les fréquences et représentez graphiquement la distribution empirique.
Calculez la médiane et l'intervalle interquartile.

b) Calculez les densités de fréquence et représentez graphiquement l'histogramme.
Déterminez la classe modale.

c) Calculez ou déterminez
la moyenne arithmétique;
la variance;
l'écart-type.

■ Exercice 2 - 2 [Avec Mathematica]

Mêmes questions que dans l'exercice précédent.

■ Exercice 2 - 3 [Sans ordinateur]

Avec les données de l'exercice 2-1,

a) calculez la fréquence des intervalles suivants

```
f ([49.5; 50.5[)
f ([50; 52[)
f ([50; 52])
f ([60; 80[)
f ([-∞; 60[)
```

b) calculez la valeur de la fonction de distribution empirique aux abscisses suivantes

```
F (20)
F (40)
F (60)
F (80)
F (100)
```

c) Vérifiez les relations

```
F (40) = f (] -∞; 40[)
F (60) = f (] -∞; 60[)
f ([60; 80[) = F (80) - F (60)
```

d) A partir des résultats de la partie b), calculez la fréquence des intervalles suivants

```
f ([40; 60[)
f ([40; 80[)
```

§ 2.2 *Erreurs de groupement*

■ Données brutes

Dans l'exemple "*Masses corporelles d'étudiants*" présenté dans le § 2.1, les données étaient groupées. Voici maintenant les observations originelles, appelées données brutes, à partir desquelles le groupement a été établi

```
x = {69, 59, 70, 72, 67, 49, 69, 67, 62, 65, 60, 68, 71, 75, 62, 77, 74, 59, 65, 62, 67,
   74, 54, 63, 54, 69, 61, 72, 65, 54, 62, 71, 71, 74, 80, 61, 80, 67, 65, 65, 69,
   69, 77, 62, 73, 61, 58, 77, 59, 73, 73, 66, 62, 57, 55, 61, 62, 67, 57, 55, 61,
   67, 79, 69, 64, 70, 68, 68, 59, 67, 67, 64, 77, 73, 67, 57, 66, 68, 72, 78, 75,
   62, 55, 64, 62, 71, 66, 67, 70, 68, 52, 77, 60, 65, 61, 57, 58, 70, 69, 66, 76,
   69, 67, 63, 77, 77, 69, 72, 66, 60, 65, 62, 65, 61, 66, 72, 73, 76, 61, 63, 66,
   64, 65, 61, 64, 61, 70, 66, 60, 65, 67, 56, 67, 66, 73, 70, 73, 73, 68, 64};
```

Pour obtenir les données numériques précédentes,
accédez au site http://www.collegedusud.ch/app/applmaths
suivez les liens **Documents *Mathematica* / Supports de cours: Statistiques I**
et téléchargez le cahier **2-stat_I.nb**

Taille de l'échantillon

```
n = Length[x]
```

```
140
```

Valeurs extrêmes

```
{Min[x], Max[x]}
```

```
{49, 80}
```

■ Groupement

Nous expliquons ici comment on peut passer des données brutes aux données groupées du § 2.1. On divise l'intervalle des modalités en classes d'égale entendue, sauf peut-être aux deux extrémités. Pour extraire l'essentiel de l'information sans qu'elle soit brouillée par des complications inutiles, le nombre de classes est choisi ni trop grand (rarement au-delà de 15 classes), ni trop petit (rarement au-dessous de 5 classes).

Comme bornes des classes, nous avons choisi

```
b = {44.5, 54.5, 59.5, 64.5, 69.5, 74.5, 79.5, 89.5};
```

Les milieux des classes sont

```
c = (Drop[b, -1] + Drop[b, 1])/2
```

```
{49.5, 57., 62., 67., 72., 77., 84.5}
```

Calculons les effectifs de chaque classe

```
effectifs = BinCounts[x, {b}]
```

```
{5, 14, 33, 47, 26, 13, 2}
```

Le premier élément est l'effectif des individus dont la masse m vérifie $44.5 \leq m < 54.5$
et ainsi de suite.

```
freq = effectifs/n
```

$$\left\{\frac{1}{28}, \frac{1}{10}, \frac{33}{140}, \frac{47}{140}, \frac{13}{70}, \frac{13}{140}, \frac{1}{70}\right\}$$

■ Paramètres empiriques

Calculons la moyenne des données brutes, puis la moyenne des données groupées

```
mB = Mean[x]; N[mB]
```

```
66.2786
```

```
mG = c.freq; N[mG]
```

```
66.3036
```

Pour comparer les deux moyennes, on peut calculer l'erreur relative due au groupement :

```
(mG - mB)/mB
```

```
0.000377196
```

qui vaut ici environ 0.04 %.

Calculons l'écart-type des données brutes, puis l'écart-type des données groupées (valeurs numériques approchées)

```
sB = StandardDeviationMLE[x]; N[sB]
```

```
6.33817
```

$$sG = \sqrt{(c - mG)^2.freq}\ ;\ N[sG]$$

```
6.68198
```

Calculons la médiane des données brutes, puis la médiane des données groupées

```
meBrut = InterpolatedQuantile[x, 1/2]
```

```
66
```

```
meGr = quantileC[b, freq, 1/2]
```

```
66.4149
```

Calculons l'intervalle interquartile des données brutes, puis l'intervalle interquartile des données groupées

```
interQuartBrut = InterpolatedQuantile[x, 3/4] - InterpolatedQuantile[x, 1/4];
N[interQuartBrut]
```

```
8.5
```

```
interQuartGr = quantileC[b, freq, 3/4] - quantileC[b, freq, 1/4]
```

```
8.7296
```

■ Erreur de groupement

On peut observer ci-dessus comment le groupement des données peut modifier la moyenne, l'écart-type, la médiane et l'écart interquartile. On appelle *erreur de groupement* l'erreur induite par le groupement des données. Une autre manière de grouper les données donnerait probablement des résultats encore différents.

L'erreur de groupement est due au fait que l'hypothèse

les données brutes sont réparties uniformémement dans chaque classe

n'est vérifiée que d'une manière approximative.

■ Exercice 2 - 4 [Avec *Mathematica*]

Partons des données brutes à partir desquelles on a effectué le groupement donné dans l'exercice 2-1:

accédez au site http://www.collegedusud.ch/app/applmaths
suivez les liens **Documents *Mathematica* / Annexes: Statistiques I**
et téléchargez le cahier **2-4_donnees_exercice.nb**

a) Comparez la moyenne arithmétique des données brutes et la moyenne arithmétique des données groupées. Plus précisément, calculez l'erreur relative due au groupement.

b) Question analogue pour l'écart-type.

c) Question analogue pour la médiane.

d) Question analogue pour l'intervalle interquartile.

■ Exercice 2 - R [Révision]

D'une table de mortalité (Suisse 1988-1993), on a extrait les données suivantes pour 100000 personnes de sexe masculin:

Âge	Nombre de décès cumulés
0	0
40	4743
50	7341
60	13 688
70	28 960
80	57 691
90	89 909
110	100 000

a) [Méthode libre] Pour chaque classe d'âge, calculez la fréquence de décès et la densité de décès.

b) [Avec *Mathematica*] Dessinez l'histogramme.

c) [Sans ordinateur] Calculez la fréquence de décès entre 77 et 84 ans.

d) [Avec *Mathematica*] Calculez l'âge moyen de décès et l'écart-type.

e) [Sans ordinateur] Ecrivez les formules pour calculer l'âge moyen de décès et l'écart-type. Remplacez les symboles par les valeurs numériques pour montrer comment elles s'appliquent ici.

f) [Sans ordinateur] A quel âge $\frac{1}{4}$ des hommes sont-ils décédés ?

g) [Avec *Mathematica*] Calculez l'âge médian de décès et l'intervalle interquartile.

■ Polygone des fréquences (Supplément facultatif)

Sur le site http://www.collegedusud.ch/app/applmaths/
suivez les liens **Documents *Mathematica* / Annexes: Statistiques I**
téléchargez et consultez le cahier **Supplement_2-1.nb**