Chapitre 3. Statistiques descriptives : Présentation de données

Mathématiques et statistiques appliquées
Département TC1-IUT de Sceaux

Damien THOMINE

Objectifs

- Reconnaître les différents types de variables et savoir choisir le type de graphique approprié.
- Lire et construire des histogrammes.
- Manier et représenter les effectifs et les fréquences cumulées.

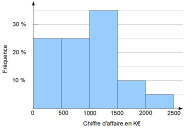
Plan du cours

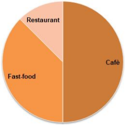
Définition

La **statistique descriptive** est un ensemble de techniques mathématiques permettant de présenter, décrire, résumer des observations faites sur une grande population.

Numéro	Departemer	Typologie	Nombre de salariés	CA en K€	alea
77001	94	Cafè	3	1 385,95	0,230362
77002	75	Fast-food	1	524,37	0,342594
77003	75	Cafè	1	483,50	0,048512
77004	75	Fast-food	1	365,33	0,972409
77005	92	Cafè	1	118,22	0,880434
77006	93	Fast-food	2	1 287,60	0,629347
77007	78	Cafè	1	1 380,08	0,938249
77008	75	Cafè	3	2 422,31	0,978339
77009	75	Restaurant	2	1 962,96	0,367627
77010	92	Cafè	2	1 087,58	0,454776
77011	92	Restaurant	2	237,59	0,015701
77012	92	Cafè	2	883,38	0,56861
77013	92	Fast-food	2	1 034,89	0,26755
77014	91	Cafè	3	724,52	0,345339
77015	75	Fast-food	1	615,06	0,530956
77016	75	Fast-food	2	1 410,98	0,112432
77017	78	Fast-food	4	1 375,41	0,965496
77018	75	Fast-food	3	1 707,93	0,03474
77019	92	Fast-food	2	738,52	0,870205
77020	94	Restaurant	0	2 388,67	0,597142
77021	75	Fast-food	4	1 640,03	0,513861
77022	78	Cafè	4	1 015,31	0,926572
77023	78	Cafè	0	1 295,19	0,439703
77024	78	Cafè	2	872,32	0,003737
77025	78	Cafè	1	164,87	0,977728
77026	75	Restaurant	2	930,14	0,909419
77027	75	Cafè	1	228,39	0,409777
77028	91	Cafè	1	106,45	0,195815
77029	92	Cafè	1	423,97	0,013714
77030	95	Fast-food	0	1 125,20	0,788084
77031	75	Cafè	3	1 411,69	0,678988
77032	75	Cafè	1	876,37	0,511315
77033	75	Cafè	2	1 097,61	0,017123
77034	75	Restaurant	2	962,68	0,66016
77035	92	Fast-food	3	1 349,85	0,559269
77036	75	Cafè	4	1 433,37	0,279256
77037	78	Fast-food	3	970,52	0,918249
77038	95	Fast-food	3	499,20	0,785718
77039	92	Fast-food	0	1 758,69	0,876999
77040	78	Cafè	1	298,76	0,114249

- **Présentation des données** : Tableaux et graphiques





$\hookrightarrow$ ce chapitre.

- **Paramètres statistiques :** médiane, moyenne, écart-type...

$\hookrightarrow$ prochains chapitres.

Section 1

Vocabulaire des Statistiques Descriptives

Populations et variables statistiques

- Une **population** statistique est l'ensemble sur lequel on effectue des observations. Les éléments de cet ensemble sont appelés les **individus**.
- Ce qui est observé ou mesuré sur les individus d'une population statistique est une **variable (ou caractère) statistique**. Les valeurs prises par une variable statistique sont appelées ses **modalités**

Exemple. Statistique à l'IUT de Sceaux.
Population : les étudiants inscrits en DUT TC

- Variable : Âge. Modalités : 17, 18, 19, 20,...
- Variable : Département de résidence. Modalités : $\boxed{}$
- Variable : $\boxed{}$ Modalité : STMG, ES, S...

Section 2

Effectifs et fréquences

Tri à plat et effectifs

Numéro	Typologie
77001	Cafè
77002	Fast-food
77003	Cafè
77004	Fast-food
77005	Cafè
77006	Fast-food
77007	Cafè
77008	Cafè
77009	Restaurant
77010	Cafè
77011	Restaurant
77012	Cafè
77013	Fast-food
77014	Cafè
77015	Fast-food
77016	Fast-food
77017	Fast-food
77018	Fast-food
77019	Fast-food
77020	Restaurant
77021	Fast-food
77022	Cafè
77023	Cafè
77024	Cafè
77025	Cafè
77026	Restaurant
77027	Cafè
77028	Cafè
77029	Cafè
77030	Fast-food
77031	Cafè
77032	Cafè
77033	Cafè
77034	Restaurant
77035	Fast-food
77036	Cafè
77037	Fast-food
77038	Fast-food
77039	Fast-food
77040	Cafè

Le tableau ici à gauche présente les résultats d'un recensement des établissements de restauration d'une ville pour ce qui concerne la variable “Type”.

Il s'agit des données non traitées qu'on appelle la **série brute**.

Pour résumer ces données, il faut faire un **tri à plat** de la variable, c'est à dire :

- Faire l'inventaire des modalités de cette variable.
- Pour chaque modalité, compter le nombre d'individus (**l'effectif**) ayant cette modalité.

Type	Café	Fast-food	Restaurant	Total
Effectifs	20	15	5	40

Tableau des effectifs

Les effectifs peuvent donc être représentés dans des tableaux tels que :

Modalités	x_1	x_2	$\dots$	x_i	$\dots$	x_p	Total
Effectifs	n_1	n_2	$\dots$	n_i	$\dots$	n_p	n

Pour chaque modalité x_i, on a noté n_i l'effectif correspondant et n la **taille de la population**. On remarque alors que

$$n_1 + n_2 + \cdots + n_i + \cdots + n_{p-1} + n_p = n.$$

La **distribution** des effectifs d'une variable est l'ensemble des effectifs de toutes ses modalités.

Fréquences

La **fréquence** d'une modalité

est la proportion d'individus ayant cette modalité sur le total de la population.
Si n_i est l'effectif d'une modalité pour une population de taille n, sa fréquence sera

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

exprimée souvent en pourcentage .

Exemple. La fréquence de la modalité “Restaurant” parmi les établissements de la ville est

$$f_{\text{restaurant}} = \frac{n_{\text{restaurant}}}{n} = \frac{5}{40} = 0,125 = 12,5\%$$

autrement dit : 12,5% des établissements sont des restaurants.

Tableau de fréquences

Comme pour les effectifs, on peut présenter les fréquences dans un tableau :

Type	Café	Fast-food	Restaurant	Total
Fréquence	50%	37,5%	12,5%	100%

Ces pourcentages sont calculés à partir du tableau des effectifs :

Type	Café	Fast-food	Restaurant	Total
Effectif	20	15	5	40

$f_{\text{café}} = \frac{n_{\text{café}}}{n} = \frac{20}{40} = 0,5 = 50\%$
$f_{\text{fast-food}} = \frac{n_{\text{fast-food}}}{n} = \frac{15}{40} = 0,375 = 37,5\%$

Test

On considère la variable “Nombre de salariés”.

Compléter :

Tableau des fréquences

Nb. de salariés	0	1	2	3	4	Total
Fréquence			30%	20%	10%	100%

Tableau des effectifs

Nb. de salariés	0	1	2	3	4	Total
Effectif	4	12				40

Section 3

Taxonomie et représentations

Objectif

Les objectifs de cette section sont de :

- distinguer différents types de variables statistiques ("Taxonomie") ;
- discuter de différentes représentations de données, en fonction de leurs caractéristiques ("Représentation").

Variables quantitatives et qualitatives

Une variable quantitative

est une variable dont les modalités représentent une **quantité**.
On peut calculer la moyenne et les autres paramètres statistiques.

Exemples. Âge, taille, nombre d'enfants...

Une variable qualitative

est une variable dont les modalités ne représentent pas une quantité, pour laquelle la notion de moyenne n'a donc pas de sens.

Exemples. oui/non, sexe...

Test

- La variable "Chiffre d'affaire" est Quantitative ☐ ou Qualitative ☐
- La variable "Couleur" est Quantitative ☐ ou Qualitative ☐
- La variable "Département" est Quantitative ☐ ou Qualitative ☐

Variables ordonnées et non ordonnées

Une variable est **ordonnée**

si ses modalités ont un ordre naturel.
Exemples. Taille de vêtement (XS, S, M, L, XL), les jugements, toutes les variables quantitatives.
$\Rightarrow$ on mettra toujours **les modalités dans l'ordre**

Une variable est **non ordonnée**

si ses modalités n'ont pas un ordre naturel. **Exemples.** Sexe, région...

Test Dans un sondage

- La variable "Genre préféré de film" avec modalités :
  "Action", "Comédie", "Drame", "Horreur"
  est ordonnée ☐ ou non-ordonnée ☐
- La variable "Fréquence de sortie au cinéma" avec modalités :
  "jamais", "quelques fois par an", "une fois par mois", "une fois par semaine"
  est ordonnée ☐ ou non-ordonnée ☐

Variables discrètes et continues

Une variable **quantitative** est dite :

discrète

si elle ne peut prendre que des **valeurs isolées**, généralement entières.

Exemples. Nombre d'enfants, la plupart des variables "Nombre de...", toutes les variables qualitatives sont discrètes.
On ne peut pas avoir 2,13 enfants.

continue

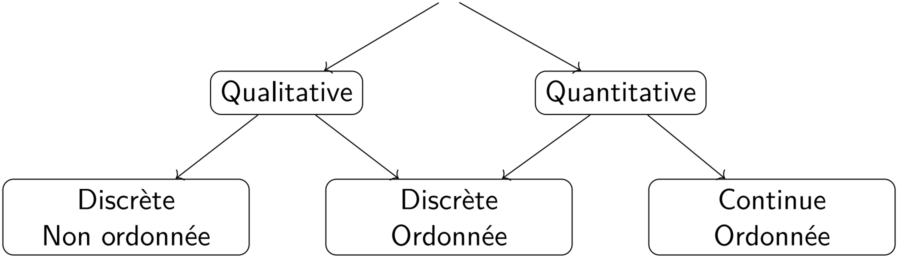
si elle peut prendre n'importe quelle valeur dans intervalle. Les modalités peuvent être des **nombres décimaux**.

Exemples. Taille, durée, poids...
On peut peser 63,51 kg.

Test :

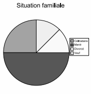
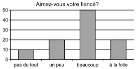
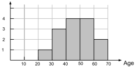
- La variable "Surface d'une exploitation agricole" est continue ☐ ou discrète ☐
- La variable "Nombre d'employés" est continue ☐ ou discrète ☐

Résumé


Qualitative
Quantitative
Discrète
Non ordonnée
Discrète
Ordonnée
Continue
Ordonnée

Graphiques

Pour communiquer efficacement il faut toujours choisir le graphique adapté à chaque type de variable.
Voici un tableau récapitulatif des graphiques qu'on présentera.

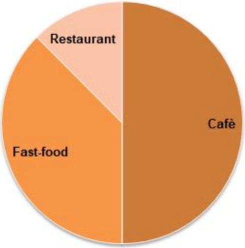
<table>
<tr><th>Graphique</th><th>Types de variable</th></tr>
<tr><td>Diagramme circulaire

</td><td>variable qualitative non-ordonnée avec peu de modalités</td></tr>
<tr><td>Diagramme en bâton

</td><td>variable qualitative ordonnée
variable qualitative avec beaucoup de modalités
variable quantitative discrète qui n'a pas été triée par classes (voir section 3)</td></tr>
<tr><td>Histogramme

</td><td>variable quantitative continue ou qui a été triée par classes</td></tr>
</table>

Diagrammes circulaires ou “camemberts”

Chaque modalité est représentée par un secteur circulaire dont l'angle (et donc la surface) est proportionnel à son effectif, et donc à sa fréquence.

Exemple.

Type	Café	Fast-food	Restaurant
Fréquence	50%	37,5%	12,5%



Quand l'utiliser

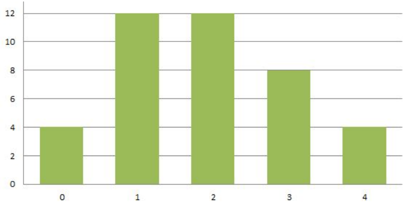
- pour les **variables discrètes non ordonnées**
- si on veut mettre en évidence les **fréquences relatives**

Diagramme en barres ou bâtons

Chaque modalité est représentée par une barre. Chaque barre a une base constante et une hauteur proportionnelle à l'effectif n_i ou à la fréquence f_i.

Exemple.

Nb. de salariés	0	1	2	3	4
Effectif	4	12	12	8	4



Quand l'utiliser

- pour des **variables discrètes ordonnées**.
- si on veut mettre en évidence la valeur de chaque effectif.

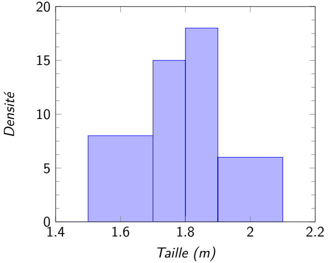
Histogramme

Chaque classe est représentée par une barre. L'**aire** de la barre est proportionnelle à l'effectif n_i ou à la fréquence f_i.

Exemple.

Taille (m)	[1.5, 1.7[	[1.7, 1.8[
Effectif	16	15

Taille (m)	[1.8, 1.9[	[1.9, 2.1[
Effectif	18	12



Quand l'utiliser

Ppour des **variables continues** ou des variables quantitatives discrètes triées par classe.

Résumé

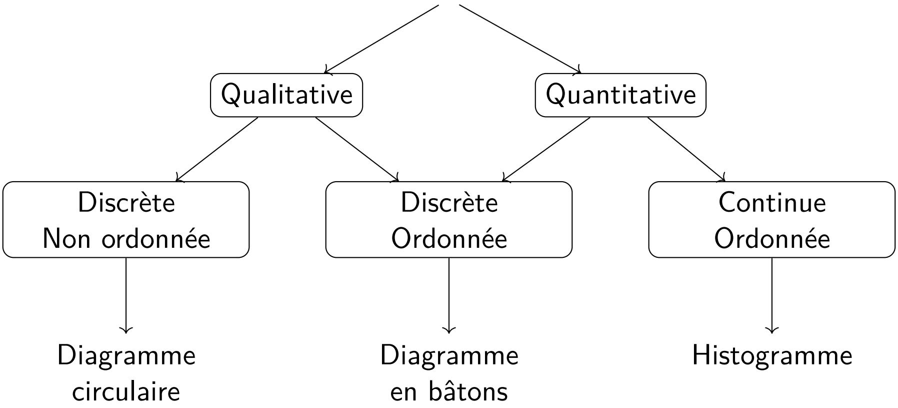

Qualitative
Quantitative
Discrète
Non ordonnée
Discrète
Ordonnée
Continue
Ordonnée
Diagramme
circulaire
Diagramme
en bâtons
Histogramme

Diagramme circulaire : quand ne pas l'utiliser 1

Le diagramme circulaire n'est pas adapté pour les **variables ordonnées**. On préfère le diagramme en bâton car on peut montrer l'ordre des modalités.

Non

Oui

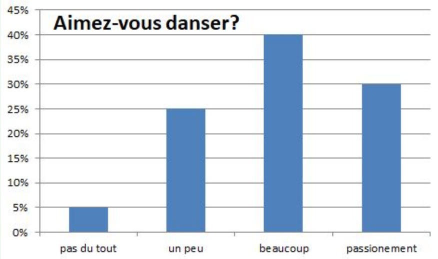


Diagramme circulaire : quand ne pas l'utiliser 1

C'est pour cette raison que l'on trouvera des diagrammes en bâtons pour résumer les notes des utilisateurs sur de nombreux sites :

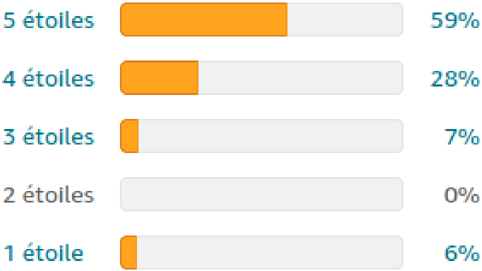
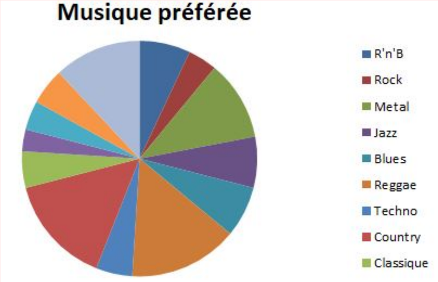


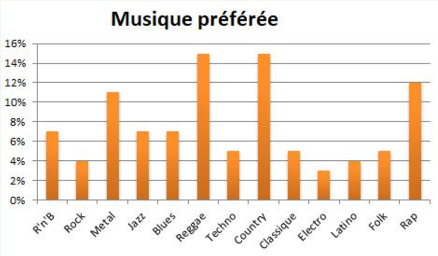
Diagramme circulaire : quand ne pas l'utiliser 2

Le diagramme circulaire n'est pas adapté si il y a **beaucoup de modalités**.

Non

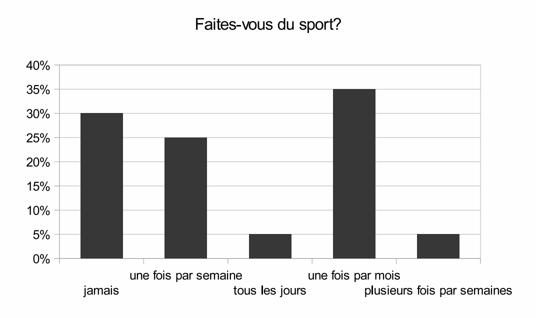


Oui



Test

Qu'est ce que ne va pas ?



Mots de fin

Dans tous les cas, l'important est de trouver une représentation qui **mette en valeur les aspects importants des données**.
Pour cela, la **sobriété** est fortement recommandée (pas de 3D, ne pas sur-utiliser les camemberts...).
Enfin, il ne s'agit que de quelques modes de représentations. Il en existe beaucoup d'autres, à choisir en fonction des données et de ce que l'on veut souligner :

- Cartes (par exemple, comment représenter une variable "Département" ?) ;
- Nuages de points (pour visualiser la dépendance entre deux variables quantitatives) ;
- Graphes...

Section 4

Tri à plat par classes

Tri à plat par classes

Numéro	CA en K€
77001	1 385,95
77002	524,37
77003	483,50
77004	365,33
77005	118,22
77006	1 287,60
77007	1 380,08
77008	2 422,31
77009	1 962,96
77010	1 087,58
77011	237,59
77012	883,38
77013	1 034,89
77014	724,52
77015	615,06
77016	1 410,98
77017	1 375,41
77018	1 707,93
77019	738,52
77020	2 388,67
77021	1 640,03
77022	1 015,31
77023	1 295,19
77024	872,32
77025	164,87
77026	930,14
77027	228,39
77028	106,45
77029	423,97
77030	1 125,20
77031	1 411,69
77032	876,37
77033	1 097,61
77034	962,68
77035	1 349,85
77036	1 433,37
77037	970,52
77038	499,20
77039	1 758,69
77040	298,76

A gauche, la série brute de la variable “Chiffre d’affaire” des établissements de la ville. Si one regroupe pas de modalités, on obtient un grand nombre de petits effectifs :

CA en K€	106,45	118,22	164,87	228,39	237,59	298,76	365,33	....
Effectif	1	1	1	1	1	1	1	...

Pour les variables qui ont **beaucoup de modalités**, il faut **grouper les modalités**. Dans le cas des variables quantitatives, on les groupe par **classes** (ou intervalles).

CA en K€	[0,500[	[500,1000[	[1000,1500[	[1500,2000[	[2000,2500[	Total
Effectif	10	10	14	4	2	40
Fréquences	25%	25%	35%	10%	5%	100%

L’**amplitude** d’une classe est la taille de l’intervalle.
Dans l’exemple, toutes les classes ont une amplitude de 500.

Le choix des classes

Le choix des classes est parfois délicat.

Pour que l'analyse soit bien lisible, on essayera d'avoir :

- des **classes d'amplitudes égales**

A éviter :

CA en K€	[0,500[	[500,1500[	[1500,2400[	[2400,2500[	Total
Effectif	...	...	...	...	40

- un **nombre de classes équilibré** : suffisamment de classes pour ne pas trop schématiser trop, mais pas trop nombreuses pour ne pas avoir beaucoup de tous petits effectifs.

A éviter :

CA en K€	[0,2500[
Effectif	40

et

CA en K€	[100,110[	[110,120[	[120,130[	[130,140[	...
Effectif	1	1	1	1	...

- **des chiffres ronds aux extrémités** ou des chiffres ayant une signification particulière.

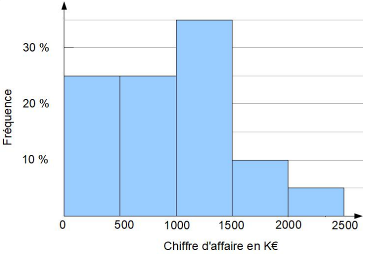
A éviter :

CA en K€	[106,735[	[735,1364[	[1364,1993[	[1993,2622[	Total
Effectifs	...	...	...	...	40

L'Histogramme

est le graphique utilisé pour la représentation des effectifs et des fréquences des variables analysées par classes.

- à chaque classe on associe un rectangle dont **la base est l'intervalle**
- **la surface** de chaque rectangle **est proportionnelle à la fréquence** de la classe
- **les rectangles sont collés** les uns aux autres, pour montrer qu'il n'y a pas de trou entre les classes,

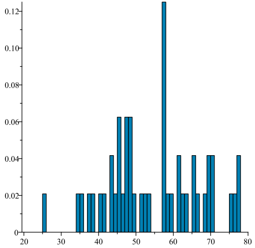


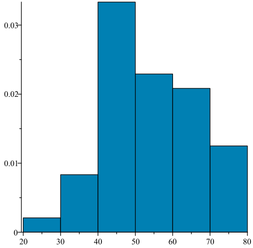
Exemple.

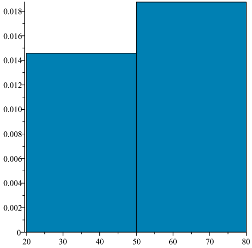
CA en K€	[0,500[	[500,1000[	[1000,1500[	[1500,2000[	[2000,2500[
Fréquences	25%	25%	35%	10%	5%

Test : Choix de l'amplitude des classes

Les histogrammes ci-dessous représentent les fréquences de l'âge des 49 inscrits à un cour de Yoga (en abscisse, les années).

1. 

2. 

3. 

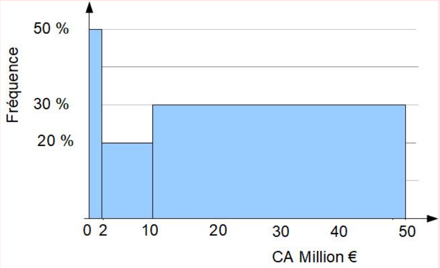
L'amplitude des classes de **2.** est ☐ et de **3.** est ☐.
Le graphique ☐ donne le moins d'information et ☐ donne le plus d'information.
Le graphique ☐ représente mieux la répartition globale des âges.

Histogramme : erreur à éviter

Exemple. Parmi les PME d'une région, on a

CA annuel en Millions d'euros	Micros Entreprises [0,2[	Petites Entreprises [2,10[	Moyennes Entreprises [10,50[
Fréquences	50%	20%	30%

Non

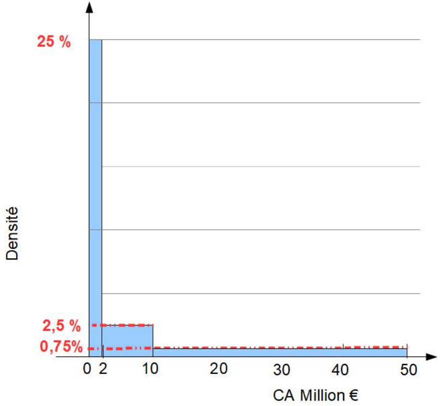


Ce graphique donne une idée fausse : les "Moyennes entreprises" y apparaissent trop nombreuses

Histogramme pour classes d'amplitudes différentes

Si l'amplitude des classes varie, on construit un histogramme où le rectangle, associé à une classe , a

- comme **base** l'intervalle qui défini la classe
- une **aire** proportionnelle à la **fréquence**.
- la **hauteur** proportionnelle à $\frac{\text{fréquence}}{\text{amplitude}}$, qu'on appelle **densité**.



Exemple.

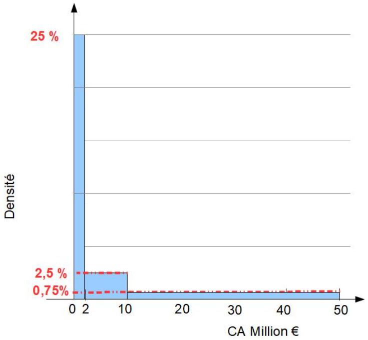
CA annuel en Millions d'euros	Micros entreprises [0,2[	Petites entreprises [2,10[	Moyennes entreprises [10,50[
Fréquences=**aire**	50%	20%	30%
Amplitude=**base**	$2-0=2$	$10-2=8$	$50-10=40$
Densité=**hauteur**	$\frac{50\%}{2}=25\%$	$\frac{20\%}{8}=2,5\%$	$\frac{30\%}{40}=0,75\%$

Densité

La densité d'une classe

$$\text{densité} = \frac{\text{fréquence}}{\text{amplitude}}$$

représente la fréquence moyenne d'une classe d'amplitude égale à 1.



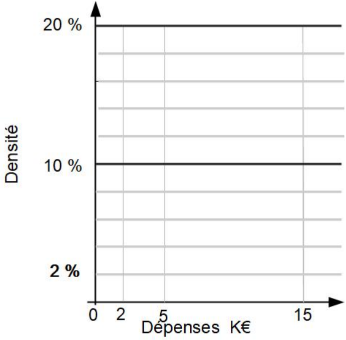
Exemple on pourrait dire que :
Environ 25 % des entreprises ont un CA compris entre 1 et 2 M€
Environ 2,5 % des entreprises ont un CA compris entre 5 et 6 M€

CA annuel en Millions d'euros	Micros Entreprises [0,2[	Petites Entreprises [2,10[	Moyennes Entreprises [10,50[
Fréquences	50%	20%	30%
Densité	$\frac{50\%}{2-0} = 25\%$	$\frac{20\%}{10-2} = 2,5\%$	$\frac{30\%}{50-10} = 0,75\%$

Test : Construire un histogramme

Exemple. Parmi les détenteurs d'une carte de fidélité d'un magasin d'électroménager, on a recensé :

Dépense annuelle en K€	Petits clients [0,2[	Moyens clients [2,5[	Gros clients [5,15[
Fréquence	20%	60%	20%
Amplitude			
Densité			



Section 5

Effectifs et Fréquences cumulés

Effectifs cumulés

Pour une variable ordonnée , l'**effectif cumulé** de la valeur x est

$$\begin{aligned} N_x &= \text{nombre d'individus assosiés à une modalité } \leq x \\ &= \text{somme de tous les effectifs } n_i \text{ pour } i \leq x \end{aligned}$$

Exemple. Le tableau des effectifs pour une étude sur une population de 18 familles concernant la variable "nombre d'enfants".

Nombre d'enfants i	0	1	2	3	4	Total
Effectif n_i	6	4	5	2	1	18

L'effectif cumulé de 2 est :

$$N_2 = n_0 + n_1 + n_2 = 6 + 4 + 5 = 15$$

c'est-à-dire : Il y a 15 familles qui ont 2 enfants ou moins.

Fréquences cumulées

Pour une variable ordonnée , la **fréquence cumulée** de la valeur x est

$$\begin{aligned} F_x &= \text{proportion d'individus assosiés à une modalité } \leq x \\ &= \text{somme de toutes les fréquences } f_i \text{ pour } i \leq x \end{aligned}$$

Exemple. Le tableau des fréquences pour une étude sur une population de 18 familles concernant la variable “Nombre d’enfants”.

Nombre d’enfants i	0	1	2	3	4	Total
Fréquences f_i	33%	22%	28%	11%	6%	100%

La fréquence cumulée de 2 est :

$$F_2 = f_0 + f_1 + f_2 = 33\% + 22\% + 28\% = 83\%$$

c’est-à-dire : 83% familles qui ont 2 enfants ou moins.

Effectifs et fréquences cumulés sont liés par la relation $F_x = \frac{N_x}{n}$

Tableau des effectifs et fréquences cumulés

Nombre d'enfants i	0	1	2	3	4	Total
Effectifs n_i	6	4	5	2	1	18
Effectifs cumulés N_i	6	10	15	17	18	

$N_0 = n_0 = 6$
$N_1 = N_0 + n_1 = 6 + 4 = 10$
$N_2 = N_1 + n_2 = 10 + 5 = 15$
$N_3 = N_2 + n_3 = 15 + 2 = 17$
$N_4 = N_3 + n_4 = 17 + 1 = 18 = n =$effectif total

Nombre d'enfants x_i	0	1	2	3	4
Fréquences f_i	33%	22%	28%	11%	6%
Fréquences cumulées F_i	33%	55%	83%	94%	100%

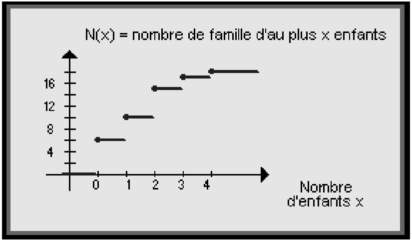
Le même principe s'applique aux fréquences.

Graphique pour les variables discrètes

Nombre d'enfants i	0	1	2	3	4	Total
Effectifs cumulés N_i	6	10	15	17	18	

Exemple. Le nombre de famille qui ont moins de 2,3 est le nombre de famille qui ont 2 enfants ou moins, c'est-à-dire $N_{2,3} = N_2$ et $F_{2,3} = F_2$

Si la variable est discrète alors $N_x = N_i$ et $F_x = F_i$ si $i \leq x < i + 1$. Donc le graphique des effectifs et des fréquences cumulés (i.e des courbes $y = N_x$ et $y = F_x$) est une courbe en **escalier**.



Effectifs et Fréquences cumulés : classes

Exemple. Le tableau de fréquences parmi les exploitations agricoles pour la variable “Surface” :

Surfaces (ha)	]0 ; 3]	]3 ; 5]	]5 ;10]	]10 ; 20]	]20 ; 30]	]30 ; 50]
Fréquence	39%	29%	24%	4%	3%	1%

On peut déduire que, les exploitations qui ont une surface :
de moins 3 ha sont 39%
$\hookrightarrow F_3$=39%
de moins 5 ha sont 68%=39%+29%
$\hookrightarrow F_5$=68%
de moins 10 ha sont 92%=39%+29%+24%
$\hookrightarrow F_{10}$=92%

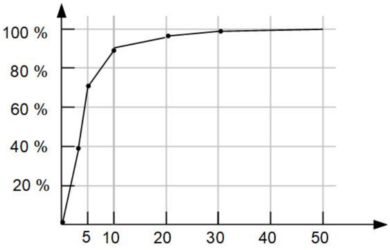
Surfaces (ha) i	0	3	5	10	20	30	50
Fréquences cumulées F_i	0%	39%	68%	92%	96%	99%	100%

Graphique pour les variables par classes

Surfaces (ha) i	3	5	10	20	30	50
Fréquences cumulées F_i	39%	68%	92%	96%	99%	100%

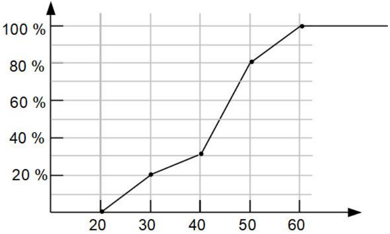
On s'attend que le nombre de exploitations d'une surface $\leq 7,3$ soit plus grand de 68% et plus petit de 92%, c'est-à-dire $F_5 < F_{7,3} < F_{10}$.

Si la variable est traitée par classe alors N_x et F_x sont des **fonctions croissantes**.
Pour tracer le graphique des effectifs et des fréquences cumulés on utilisera une courbe qui **croit linéairement** entre les valeurs connues.



Test

Le graphique suivant représente les fréquences cumulées de l'âge des employés d'une entreprise



20% des employés ont moins de $\boxed{}$ ans.

$\boxed{}$% des employés ont moins de 50 ans.

$\boxed{}$% des employés ont entre 30 et 50 ans.

Environ 50% des employés ont moins de $\boxed{}$ ans.