1 Notions de base

1.1 Contexte

Les statistiques descriptives ont pour but de décrire de manière synthétiques les caractéristiques d'un ensemble d'individus. Les individus peuvent être des personnes, mais aussi des objets ou des faits.

1.2 Vocabulaire

Population : ensemble de tous les individus que l'on souhaite étudier.
Chaque élément d'une population s'appelle une **unité statistique**.

Echantillon : ensemble des individus que l'on interroge, ou dont on connaît les informations.

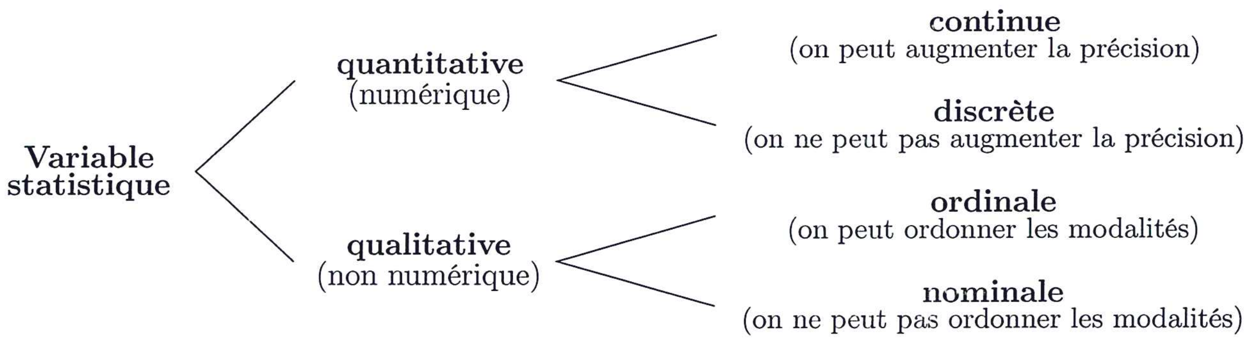
Si l'échantillon est égal à la population entière, on parle d'un **recensement**. Si l'échantillon n'est formé que d'une partie de la population, on parle d'un **sondage**.

Variable statistique : caractéristique que l'on souhaite étudier.

Une variable statistique est dite **quantitative** si les valeurs qu'elle peut prendre sont des nombres. Dans le cas contraire, elle est **qualitative**, et ses valeurs s'appellent des **modalités** ou des **catégories**.

Si les valeurs que peut prendre une variable statistique quantitative peuvent être listées (par exemple 0, 1, 2, 3, ...) et leur précision ne peut pas être augmentée, on dit que la variable est **discrète**. Si la précision des valeurs peut être modifiée (par exemple 2.3 ou 2.31 ou 2.314), on dit qu'elle est **continue**.

Si les valeurs que peut prendre une variable statistique qualitative peuvent être ordonnées (mise dans un certain ordre, par exemple *jamais*, *parfois*, *souvent*, *toujours*), la variable est dite **ordinale**. Dans le cas contraire, elle est **nominale**.



Exemples

a) On demande à 200 passagers pris au hasard dans un aéroport de donner leur nationalité.

Population : les passagers d'un aéroport
Echantillon : les 200 passagers
Unité statistique : 1 passager
S'agit-il d'un recensement ou d'un sondage ? sondage
Variable statistique : nationalité
Type de variable : qualitative nominale
Modalités ou valeurs : suisse, français, allemand, etc...

b) Durant tout le mois d'avril, on mesure la température maximale de la journée à la Tour-de-Peilz.

Population : Les jours du mois d'avril
Echantillon = population
Unité statistique : 1 jour
S'agit-il d'un recensement ou d'un sondage ? recensement
Variable statistique : température max à la Tour-de-Peilz
Type de variable : quantitative continue
Modalités ou valeurs : $[0°; 30°]$ par exemple

c) On demande à 5'000 familles résidant en Suisse le nombre d'enfants de leur foyer.

Population : les familles résidant en Suisse
Echantillon : les 5'000 familles interrogées
Unité statistique : 1 famille
S'agit-il d'un recensement ou d'un sondage ? sondage
Variable statistique : nombre d'enfants
Type de variable : quantitative discrète
Modalités ou valeurs : 0, 1, 2, 3, ...

d) A la fin d'une période de révision, un enseignant demande à ses élèves de répondre à la question suivante :
Cette période de révision a-t-elle été utile ? ☐ *pas du tout,* ☐ *un peu* ☐ *beaucoup*

Population : les élèves de la classe
Echantillon = population
Unité statistique : 1 élève
S'agit-il d'un recensement ou d'un sondage ? recensement
Variable statistique : l'utilité de la période
Type de variable : qualitative ordinale
Modalités ou valeurs : pas du tout, un peu, beaucoup

1.3 Echelles de mesure

On distingue quatre types d'échelles de mesure.

Echelle nominale

Exemple : 0 (homme) 1 (femme)

- s'utilise pour une variable qualitative nominale
- sert uniquement à différencier les catégories d'une variable ($=$, $\neq$)
- ne permet pas d'établir une relation d'ordre ($<$, $>$)
- ne permet pas d'effectuer d'opération arithmétique ($+$, $-$, $\cdot$, $:$)

Echelle ordinale

Exemple : 0 (pas du tout) 1 (un peu) 2 (beaucoup)

- s'utilise pour une variable qualitative ordinale ou quantitative
- sert à différencier les catégories d'une variable ($=$, $\neq$)
- permet d'ordonner les catégories ($<$, $>$)
- ne permet pas d'effectuer d'opération arithmétique ($+$, $-$, $\cdot$, $:$)

Echelle d'intervalle

Exemple : les degrés Celsius

- s'utilise pour une variable quantitative
- sert à différencier les valeurs d'une variable ($=$, $\neq$)
- permet d'ordonner les valeurs ($<$, $>$)
- permet d'effectuer des additions / soustractions, mais pas des multiplications / divisions
- la valeur 0 ne signifie pas l'absence de la caractéristique

Echelle de rapport

Exemple : nombre d'enfants dans une famille : 0, 1, 2, ...

- s'utilise pour une variable quantitative
- sert à différencier les valeurs d'une variable ($=$, $\neq$)
- permet d'ordonner les valeurs($<$,$>$)
- permet d'effectuer les quatre opérations arithmétiques ($+$, $-$, $\cdot$, $:$)
- la valeur 0 signifie une absence de la caractéristique mesurée

Exemple

Voici quatre manières différentes de mesurer la consommation de cigarettes dans un questionnaire :

1. | Fumez-vous des cigarettes ? |
   |---|
   | 1. Oui 2. Non |

   Echelle : nominale

2. | A quelle fréquence fumez-vous des cigarettes ? |
   |---|
   | 0. Jamais 1. Rarement 2. Occasionnellement 3. Régulièrement |

   Echelle : ordinale

   (1 < 2 => rarement < occasionnellement)

3. | En moyenne, combien de cigarettes fumez-vous par jour ? |
   |---|
   | 1. 0 2. De 1 à 5 3. de 6 à 10 4. de 11 à 20 5. Plus de 20 |

   Echelle : ordinale

   6 à 10 < 11 à 20 ✓

   2 + 2 ≠ 4 : (1 à 5 + 1 à 5 ≠ 11 à 20)

4. | Combien de cigarettes fumez-vous par jour ? |
   |---|
   | .... |

   Echelle : de rapport

   0 : ne fume pas

   6 cigarettes/jour est bien le double de 3 cigarettes/jour

1.4 Tableau de distribution

Une fois les données récoltées, on les regroupe par modalité dans un tableau de distribution.

Exemples

a) On a demandé à tous les élèves d'une classe quelle était leur matière préférée parmi les matières suivantes : français, anglais, maths et musique.

musique	français	français	anglais	français
anglais	musique	maths	musique	musique
musique	musique	musique	français	français
français	anglais	musique	anglais	maths

Tableau de distribution :

Répartition des élèves selon leur matière préférée

matière	nombre d'élèves	fréquence
français	6	6/20 = 30%
anglais	4	20%
maths	2	10%
musique	8	40%
Total	20	100%

b) On a demandé à ces mêmes élèves leur dernière note d'anglais.

5	4.5	3.5	5	6	3.5	4	2.5	4	4.5
4	4.5	4.5	4.5	3	4	4.5	5	3.5	4

Tableau de distribution :

Répartition des élèves selon leur note d'anglais

note	effectif	fréquence
1	0	0%
1,5	0	0%
2	0	0%
2,5	1	5%
3	1	5%
3,5	3	15%
4	5	25%
4,5	6	30%
5	3	15%
5,5	0	0%
6	1	5%
Total	20	100%

c) Enfin, on a demandé à ces élèves leur taille en centimètre.

plus petite

172	157	162	156	167	179	173	173	178	160
168	171	165	166	184	170	165	164	160	175

plus grande

Il s'agit d'une variable continue, donc il est très rare d'obtenir plusieurs fois la même valeur

=> on regroupe les données par classes

Tableau de distribution :

Répartition des élèves selon leur taille

155 est dans cette classe

160 non

	taille	nbre d'élèves	fréquence
II	[155 ; 160[	2	10%
IIII	[160 ; 165[	4	20%
~~IIII~~	[165 ; 170[	5	25%
~~IIII~~	[170 ; 175[	5	25%
III	[175 ; 180[	3	15%
I	[180 ; 185[	1	5%
	Total	20	100%

Aide pour le choix du nombre de classes : Table de Sturges

Nombre de données	Nombre **approximatif** de classes
Entre 10 et 22	5
Entre 23 et 44	6
Entre 45 et 90	7
Entre 91 et 180	8
Entre 181 et 360	9
Entre 361 et 720	10

on a pris 6 classes pour avoir des classes pratiques (4 irait aussi...)

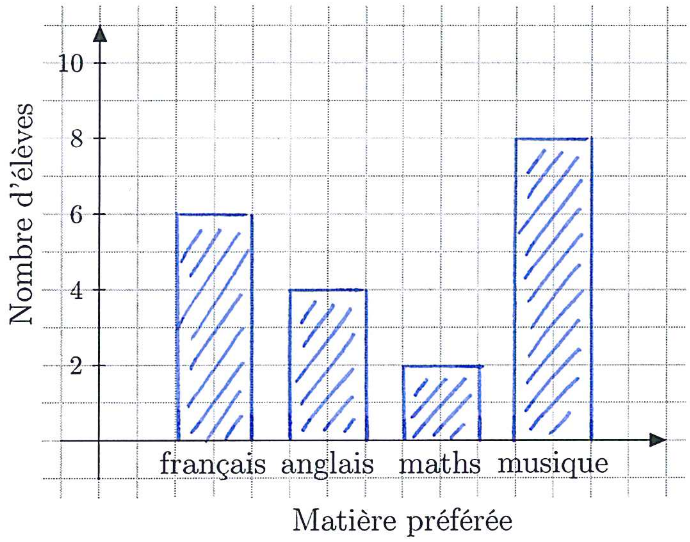
(Formule : $1 + \log_2(n)$)

2 Représentations graphiques

2.1 Représentation graphique d'une variable qualitative

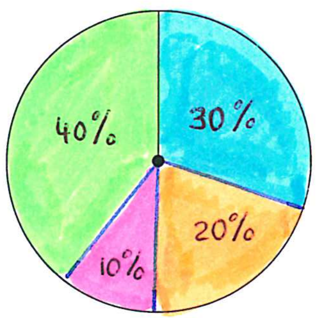
Répartition des 20 élèves d'une classe selon leur matières préférées (exemple p.5).

Diagramme à rectangles



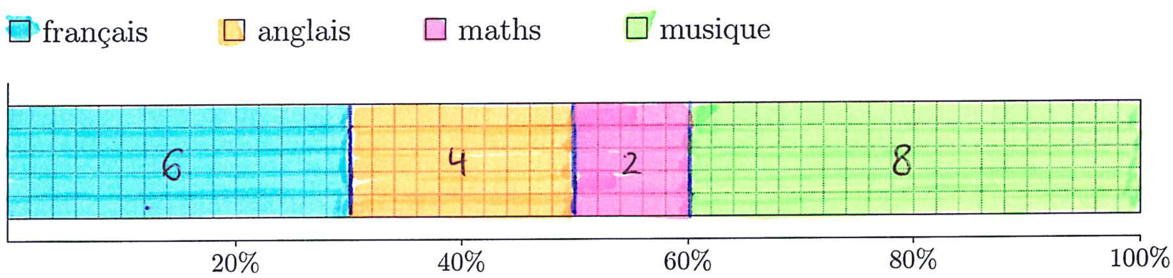
⚠ toujours laisser un espace entre les rectangles !

Diagramme circulaire



- français 30/100 = 108°/360°
- anglais 20/100 = 72°/360°
- maths 10/100 = 36°/360°
- musique 40/100 = 144°/360°

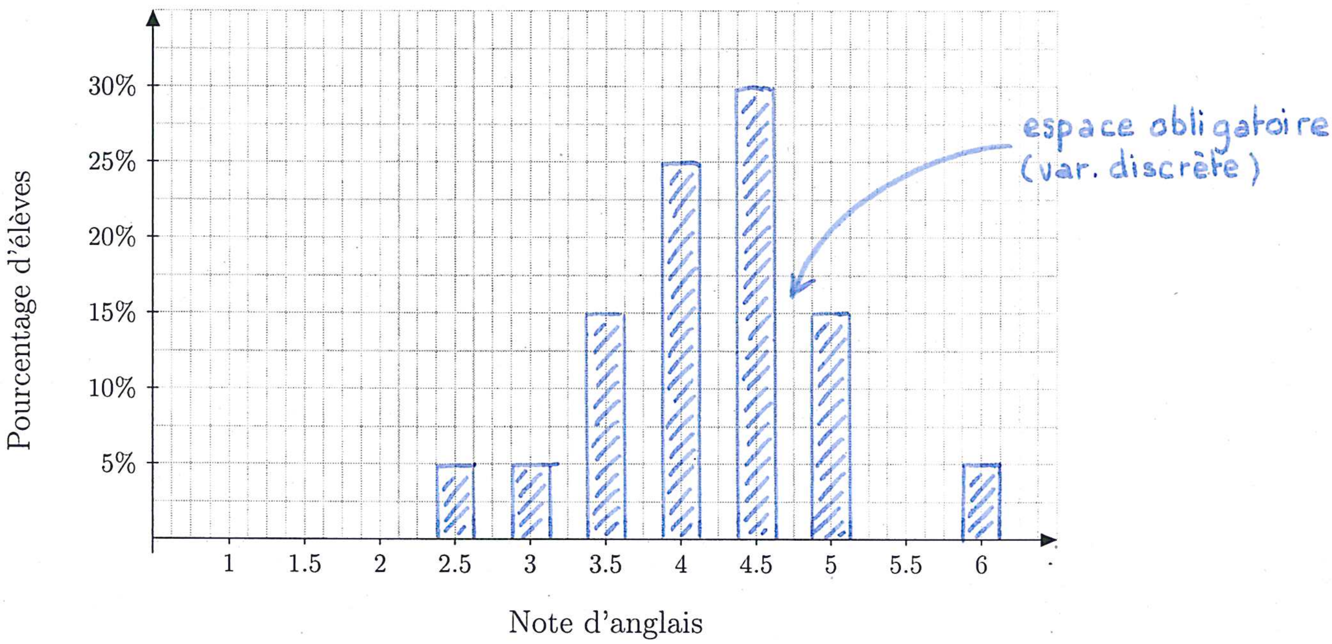
Diagramme linéaire



2.2 Représentation graphique d'une variable quantitative discrète

Répartition des 20 élèves d'une classe selon leur note d'anglais (exemple p.5).

Diagramme en bâtons (ou rectangles)

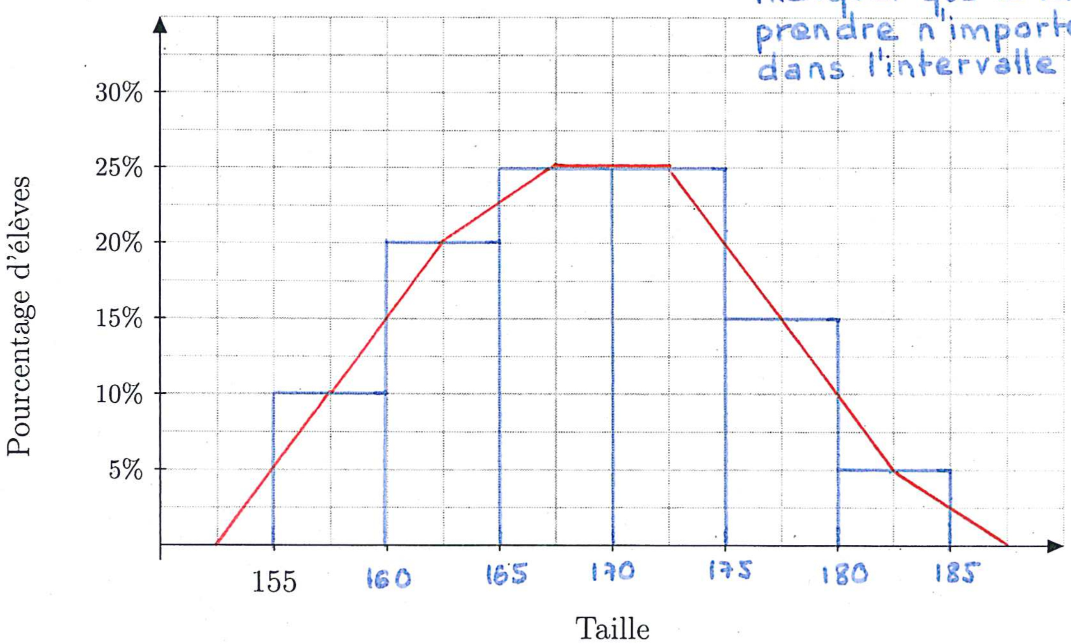


2.3 Représentation graphique d'une variable quantitative continue

Répartition des 20 élèves d'une classe selon leur taille (exemple p.6).

Histogramme et polygone de fréquences

les rectangles sont collés pour indiquer que la variable peut prendre n'importe quelle valeur dans l'intervalle

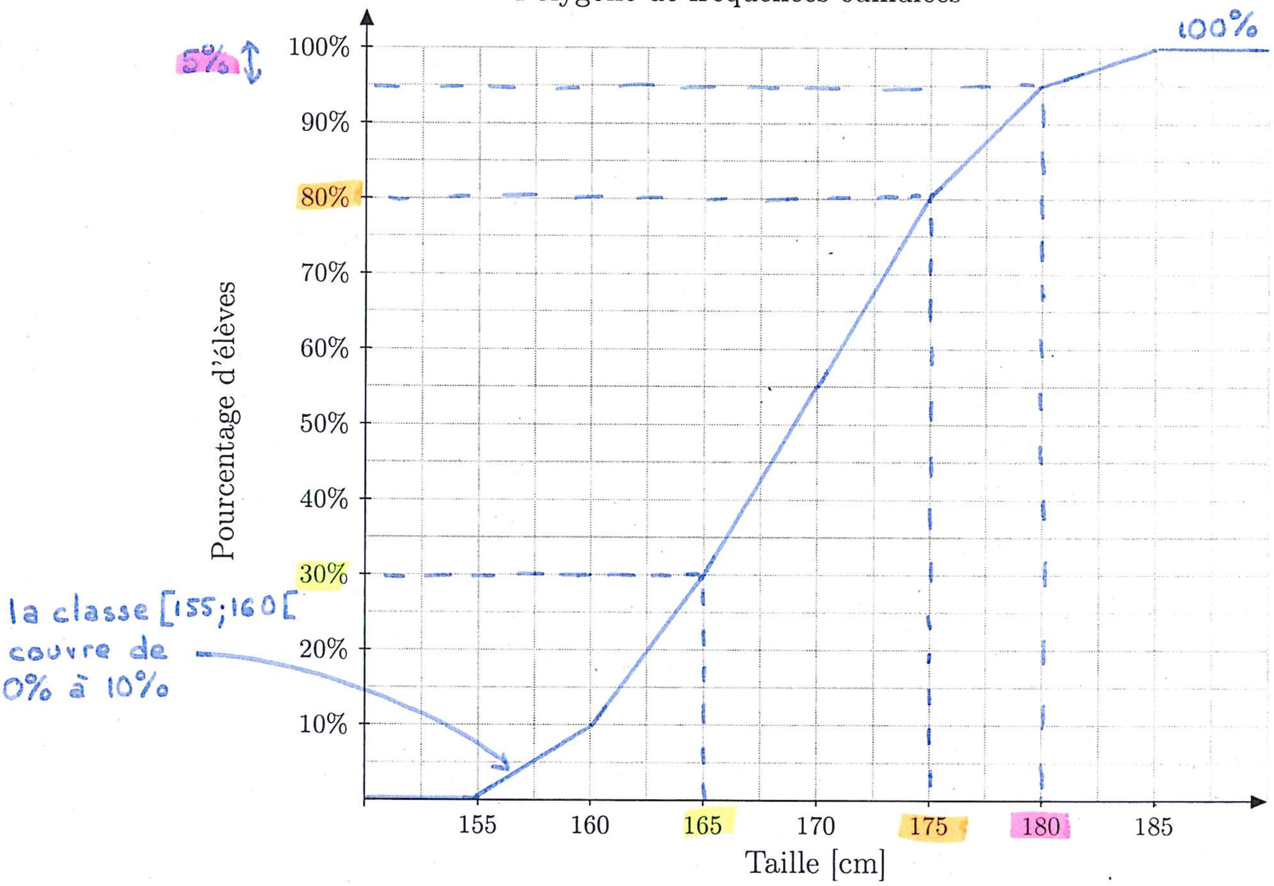


Polygone de fréquences cumulées (ou courbe de fréquences cumulées)

Répartition des élèves selon leur taille

taille	effectif	fréquence	fréquence cumulée
[155 ; 160[	2	10%	10%
[160 ; 165[	4	20%	10% + 20% = 30%
[165 ; 170[	5	25%	30% + 25% = 55%
[170 ; 175[	5	25%	80%
[175 ; 180[	3	15%	95%
[180 ; 185[	1	5%	100%
Total	20	100%	

Polygone de fréquences cumulées



30% des élèves mesurent moins de 165 cm, ou 1,65 mètre

80 % des élèves mesurent moins de 1.75 mètre.

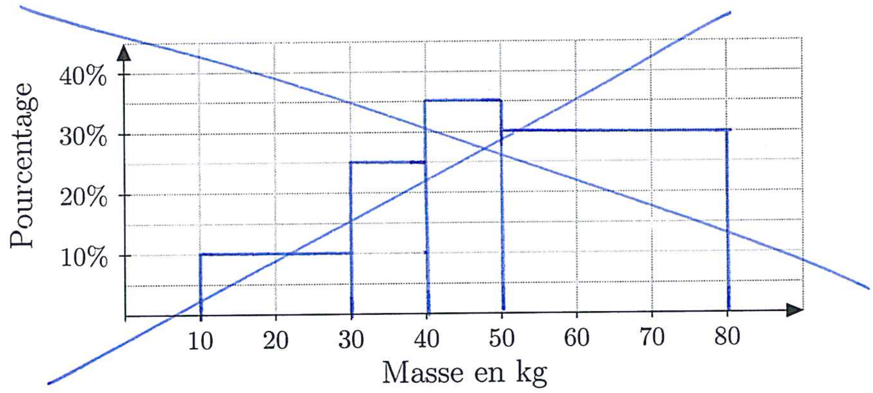
5% des élèves mesurent au moins 180 cm, ou 1,80 mètre

Histogramme dans le cas de classes inégales

On veut représenter par un histogramme la répartition des roches dans une rocaille de fleurs selon leur masse. Cette répartition est donnée par le tableau suivant :

masse en kg	pourcentage de roches
[10 ; 30[	10%
[30 ; 40[	25%
[40 ; 50[	35%
[50 ; 80[	30%
Total	100%

Histogramme obtenu en utilisant directement les fréquences données dans le tableau :

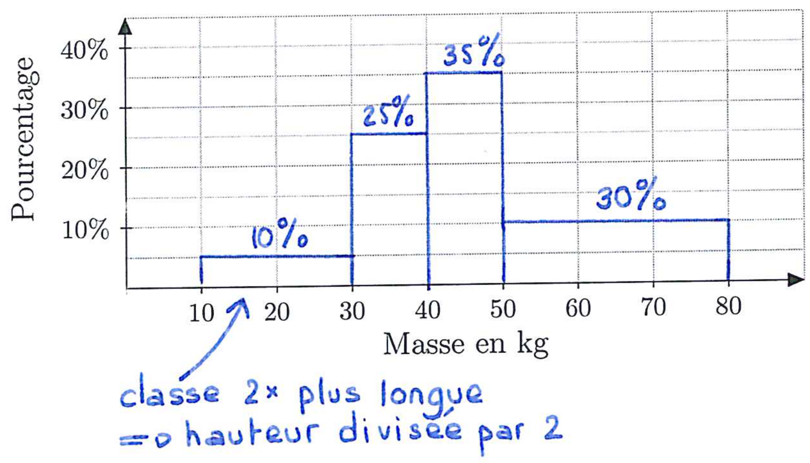


Cet histogramme donne-t-il une information correcte ou faussée ? faussée

Principe de proportionnalité des aires :

L'aire d'un rectangle doit être proportionnelle au pourcentage.
(et non la hauteur)

Histogramme correct (respectant le principe de proportionnalité des aires) :



3 Mesures de tendance centrale

But : résumer une série statistique par une seule valeur.

3.1 Moyenne

$$\boxed{\bar{x} = \frac{x_1 + x_2 + x_3 + \cdots + x_n}{n}}$$

Exemple 1

On reprend les dernières notes d'anglais des 20 élèves d'une classe (exemple p.5).

5	4.5	3.5	5	6	3.5	4	2.5	4	4.5
4	4.5	4.5	4.5	3	4	4.5	5	3.5	4

Calcul de la moyenne : $\frac{5+4+4,5+4,5+3,5+\ldots+4,5+4}{20} = 4,2$

Calculatrice : 5 $\boxed{\Sigma+}$ 4 $\boxed{\Sigma+}$ 4,5 $\boxed{\Sigma+}$ ... 4 $\boxed{\Sigma+}$ (affiche n=20)

$\boxed{2nd}$ $\overset{\bar{x}}{\boxed{x^2}}$ => donne bien 4,2

Calcul de la moyenne à partir du tableau de distribution (formule plus rapide) :

Répartition des 20 élèves d'une classe selon leur note d'anglais

note	effectif	fréquence
1 / 1.5 / 2	0	0%
2.5	1	5%
3	1	5%
3.5	3	15%
4	5	25%
4.5	6	30%
5	3	15%
5.5	0	0%
6	1	5%
Total	20	100%

n : effectif
c : catégorie/valeur/modalité

$$\boxed{\bar{x} = \frac{n_1 \cdot c_1 + n_2 \cdot c_2 \cdots + n_k \cdot c_k}{n} = f_1 \cdot c_1 + f_2 \cdot c_2 \cdots + f_k \cdot c_k}$$

$\bar{x} = \frac{1 \cdot 2,5 + 1 \cdot 3 + 3 \cdot 3,5 + \ldots + 1 \cdot 6}{20} = 4,2$ ou $\bar{x} = 0,05 \cdot 2,5 + 0,05 \cdot 3 + 0,15 \cdot 3,5 + \ldots + 0,05 \cdot 6 = 4,2$

Calculatrice : 2,5 $\boxed{2nd}$ $\underbrace{\overset{FRQ}{\boxed{1/x}}\ 1}_{\text{fréqu.}}$ $\boxed{\Sigma+}$ ...

Exemple 2

On reprend les tailles des 20 élèves d'une classe (exemple p.6).

172	157	162	156	167	179	173	173	178	160
168	171	165	166	184	170	165	164	160	175

Calcul direct de la moyenne (à partir des données brutes) :

$$\bar{x} = \frac{172 + 168 + 157 + \ldots + 160 + 175}{20} = 168,25 \quad (\cong 168 \text{ cm})$$

Calcul de la moyenne à partir du tableau de distribution :

Répartition des 20 élèves d'une classe selon leur taille

taille	valeur centrale	effectif	fréquence	fréqu. cum.
[155 ; 160[	157,5	2	10%	10%
[160 ; 165[	162,5	4	20%	30%
[165 ; 170[	167,5	5	25%	55%
[170 ; 175[	172,5	5	25%	80%
[175 ; 180[	177,5	3	15%	95%
[180 ; 185[	182,5	1	5%	100%
Total		20	100%	

n : effectif
c : valeur centrale

$$\boxed{\bar{x} = \frac{n_1 \cdot c_1 + n_2 \cdot c_2 \cdots + n_k \cdot c_k}{n} = f_1 \cdot c_1 + f_2 \cdot c_2 \cdots + f_k \cdot c_k}$$

$$\bar{x} = \frac{2 \cdot 157,5 + 4 \cdot 162,5 + 5 \cdot 167,5 + \ldots + 1 \cdot 182,5}{20} = 169$$

ou alors

$$\bar{x} = 0,1 \cdot 157,5 + 0,2 \cdot 162,5 + 0,25 \cdot 167,5 + \ldots + 0,05 \cdot 182,5 = 169$$

Remarque

La valeur de la moyenne n'est pas la même selon la méthode utilisée. Dans le deuxième cas, on ne dispose plus des données brutes, et on doit donc estimer la valeur de $\bar{x}$ avec l'information disponible.