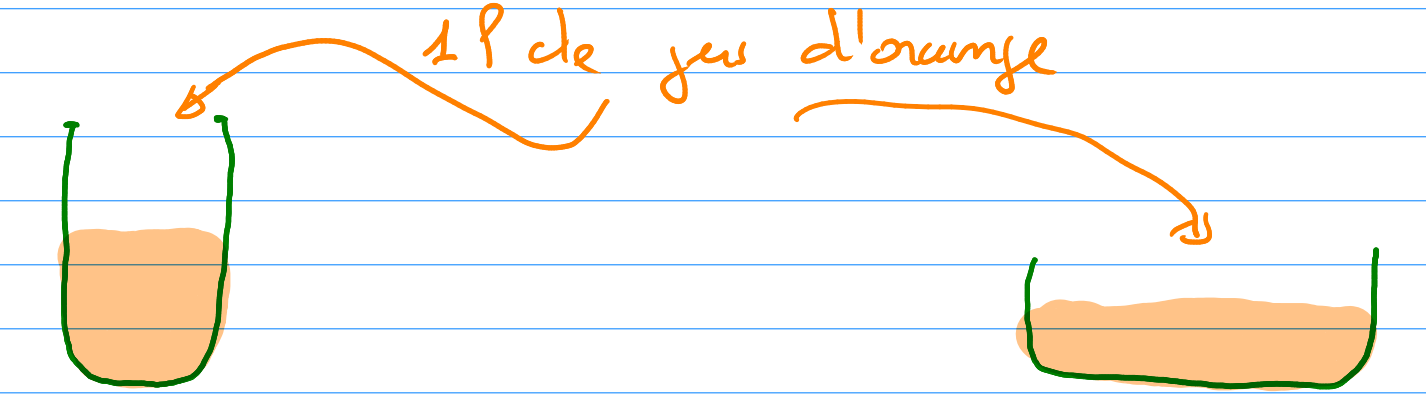
Représentation graphique des variables continues

↳ Regrouper les données en classes

classe i	n_i	f_i
$[a_0 ; a_1[$	n_1	f_1
$[a_1 ; a_2[$	n_2	f_2
⋮	⋮	⋮
$[a_{k-1} ; a_k[$	n_k	f_k
	N	1

N : taille pop

↳ idée :



verre + étroit
≡
niveau + haut

Verre + large
≡
niveau + bas

quantité identique

Donc :

fréquence de la classe ⟶ surface sur graphique

amplitude de la classe ⟶ largeur "

? ⟶ hauteur "

rem : hauteur × largeur = surface ⇒ hauteur = surface / largeur

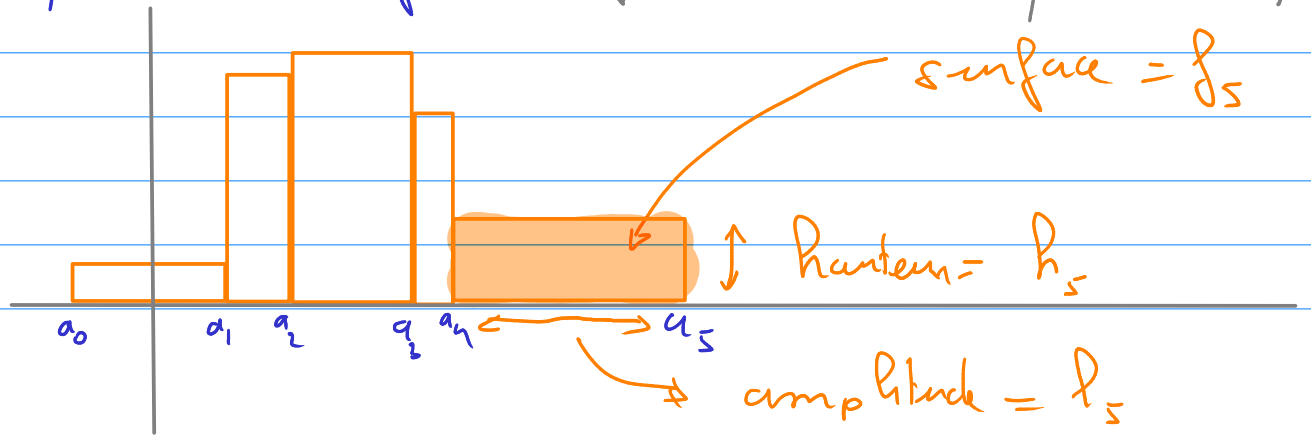
Def : Pour une classe $[a_{i-1}; a_i[$ de fréquence f_i et d'amplitude $l_i = (a_i - a_{i-1})$, la quantité

$$h_i = \frac{f_i}{l_i}$$

est la valeur de l'histogramme sur la classe $[a_{i-1}; a_i[$

Rem : ↳ la fonction histogramme peut être définie plus généralement pour tout réel

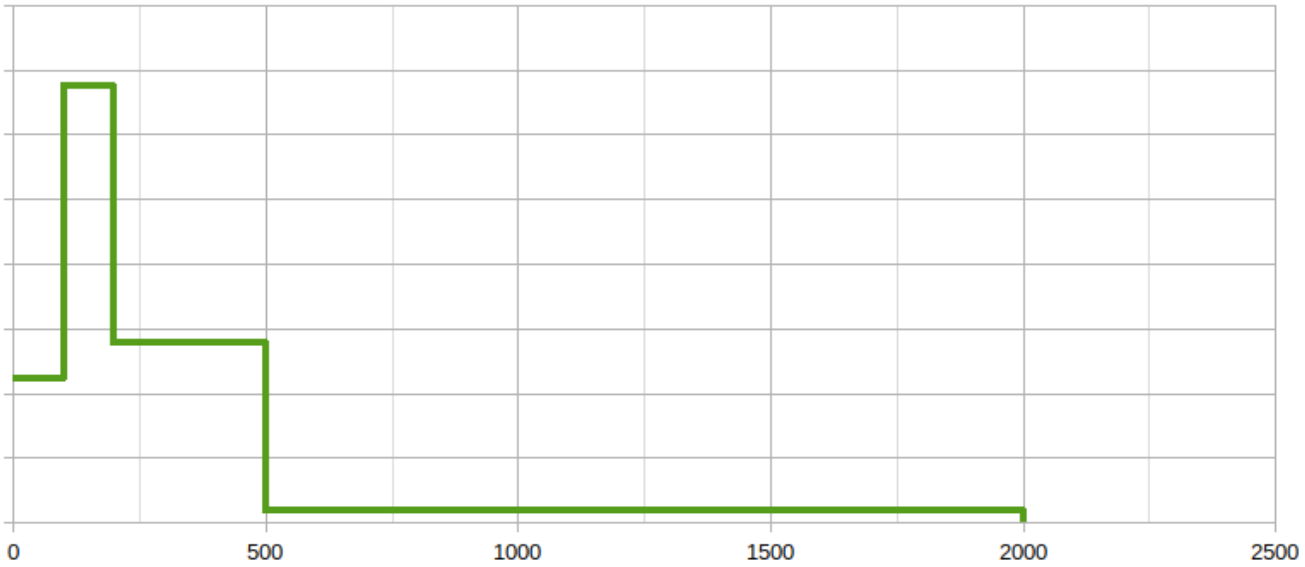
↳ par abus de langage, on appelle aussi histogramme la représentation graphique de cette fonction (constante sur chaque classe) ci-dessous :



ex de Frontmarchet

Prix :	$[0;100[$	$[100;200[$	$[200;500[$	$[500;2000]$	
n_i	342	1031	1276	406	3055
f_i	11%	34%	42%	13%	1
l_i	100	100	300	1500	
h_i	1,12	3,37	1,39	0,09	

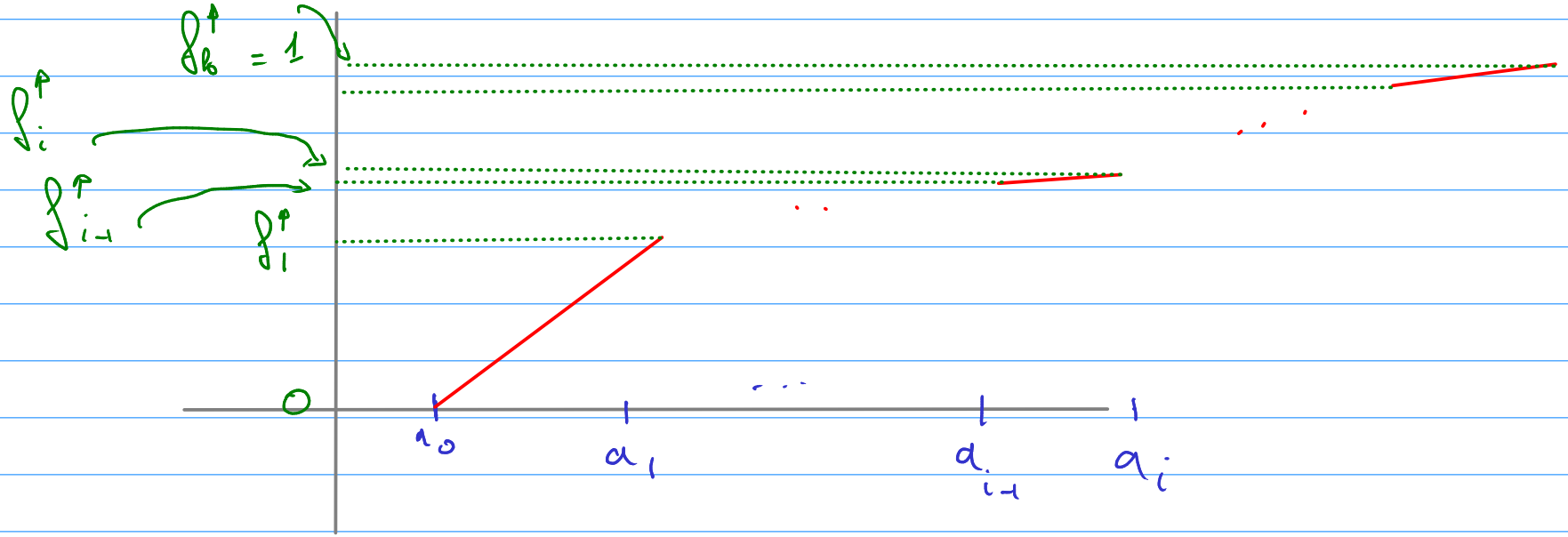
x 100 pour lisibilité



Rem :
- ↳ visualisation la plus simple d'un jeu de données continues
- ↳ choix des classes primordial
- ↳ en général, 1° beaucoup de classes (20;30) de même amplitude
  2° diminuer ensuite le nombre de classes
- ↳ adapté aux variables discrètes avec nombreuses modalités
- ↳ $\sum f_i = 1$ mais $\sum h_i$ n'a pas de sens
- ↳ L'unité de l'axe des ordonnées n'a pas d'interprétation particulière

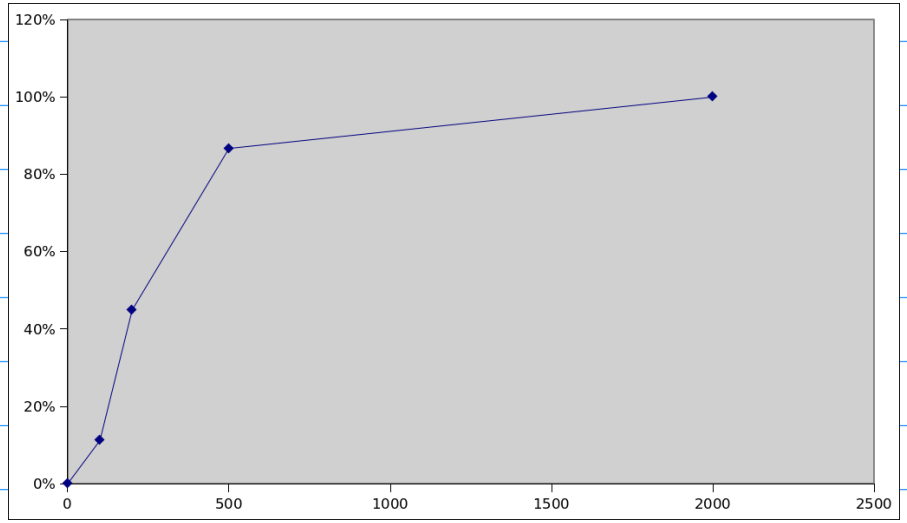
Courbes cumulatives

↳ pas sur le même graphique que l'histogramme



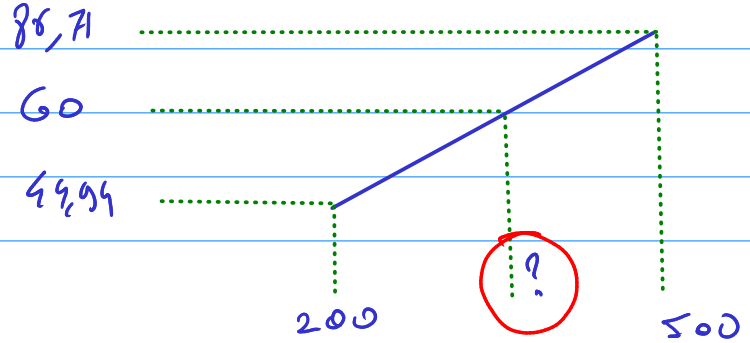
Ex de Frontmarket

$[a_{i-1}\ a_i[$	n_i	f_i	p_i ↑
[0 ; 100[	342	11,19%	11,19%
[100 ; 200[	1031	33,75%	44,94%
[200 ; 500[	1276	41,77%	86,71%
[500 ; 2000[	406	13,29%	100,00%
	3055	100%	



Rem Lecture : { 44,94% de smartphones à moins de 200 €
88,71% " " 500 €

Si l'interpolation aurait un sens : 60% des smartphones à moins de ? €



Indicateurs statistiques

↳ But : Résumer la série par quelques nombres
- → position (localisation, tendance centrale)
- → dispersion (variabilité, inhomogénéité)
- → asymétrie, aplatissement etc... (pas dans cette ressource)

↳ Avec perte d'information

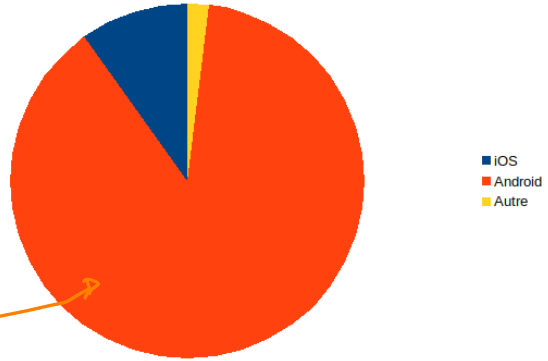
Indicateurs de positions

Def Le **mode** noté M_0 d'une série statistique est la modalité de plus grande **fréquence**, si elle existe et est unique

Rem :
- ↳ principalement pour variables qualitatives ou quantitatives discrètes
- ↳ classe modale pour variables continues (cf cours suivant)
- ↳ On parle de distribution bimodale ou multimodale si plusieurs valeurs les plus fréquentes → indication de plusieurs sous-populations distinctes
- ↳ Le mode est une valeur "privilégiée" en un certain sens

Ex de Frontmarket pour OS

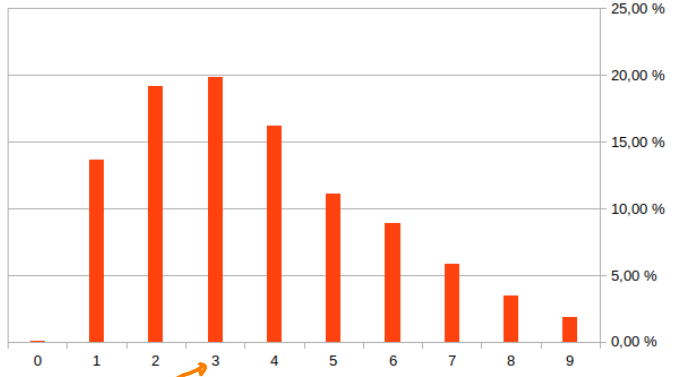
OS	iOS	Android	Autres	
n_i	301	2696	58	3055
f_i	9,9%	88,2%	1,9%	1



M_o = Android L'OS le plus proposé est Android

Ex de Frontmarket
Âge des smartphones

x_i	n_i	f_i
0	2	0,07 %
1	416	13,62 %
2	585	19,15 %
3	607	19,87 %
4	494	16,17 %
5	339	11,10 %
6	273	8,94 %
7	178	5,83 %
8	105	3,44 %
9	56	1,83 %
	3055	1



$M_o = 3$

L'âge le plus courant pour un smartphone est 3 ans

Def : On appelle médiane d'une série statistique quantitative la plus petite des modalités pour laquelle la moitié au moins des observations lui soient inférieures ou égales

En pratique : ↳ si données individuelles

1°/ on range les observations par ordre croissant

2°/ la médiane est l'observation de rang

$$\begin{cases} \frac{N}{2} & \text{si } N \text{ est pair} \\ \frac{N+1}{2} & \text{si } N \text{ est impair} \end{cases}$$

N : taille pop

↳ à partir du tableau des effectifs

1°/ on construit les fréquences cumulées croissantes $f_i^{\uparrow}$

2°/ la médiane est la plus petite valeur x_i telle que $f_i^{\uparrow} \geq 0,5$

Rem :
↳ La médiane est une valeur observée
↳ C'est un indicateur de tendance centrale : le "milieu du tableau"
↳ en grande population : autant de valeurs plus grandes que plus petites
↳ insensible aux valeurs extrêmes

⚠ ↳ Les tableurs usuels (Excel, calc, google sheet..) ont une définition FAUSSE de la médiane si N est pair

↳ de plus, calcul sur données individuelles uniquement pour tableurs

Ex de Frontmarket : capacité en Go

x_i	4	8	16	32	64	128	256	512	1024
n_i	22	147	416	451	581	957	363	107	11
n_i+	22	169	585	1036	1617	2574	2937	3044	3055
f_i+	0,72	5,53	19,15	33,91	52,93	84,26	96,14	99,64	100,00

La capacité médiane est 64 Go : 53% des smartphones ont au plus 64 Go
47% " ; au moins 64 Go

Dans l'idéal, 50 - 50.

Def : On appelle **moyenne** (arithmétique) d'une série statistique quantitative le nombre

$$\bar{x} = \frac{\text{somme des observations}}{\text{taille population}}$$

↳ c'est la moyenne usuelle

↳ à partir des données individuelles $X(u_1), \dots, X(u_N)$ $\quad \bar{x} = \dfrac{\sum_{i=1}^{N} X(u_i)}{N}$ (taille pop)

↳ à partir du tableau des effectifs :

(nb de modalités) (taille pop)

$$\bar{x} = \frac{1}{N} \sum_{i=1}^{m} n_i x_i = \sum_{i=1}^{m} f_i x_i$$

Ex de Frontmarket : capacité en Go

x_i	4	8	16	32	64	128	256	512	1024
n_i	22	147	416	451	581	957	363	107	11
f_i	0,72 %	4,81 %	13,62 %	14,76 %	19,02 %	31,33 %	11,88 %	3,50 %	0,36 %

$$= \sum_{i=1}^{9} x_i f_i$$

$$\bar{x} = 4 \times 0,0072 + 8 \times 0,0481 + \dots + 512 \times 0,035 + 1024 \times 0,0036$$

$$= 111 \text{ Go}$$

La capacité moyenne des smartphones est 111 Go

Rem : La médiane est un indicateur de tendance centrale plus adapté dans ce cas. Rappel : $M_e = 64$ Go

↳ Si l'on transforme une série statistique par une application affine

$$x_i \longrightarrow y_i = ax_i + b$$

alors $\bar{y} = a\bar{x} + b$

↳ ce n'est plus forcément vrai si la transformation n'est pas affine

ex : carré $x_i \longrightarrow ax_i^2 + b$

logarithme $x_i \longrightarrow \ln(x_i)$

Déf : Les extrêmes d'une série statistique sont

↳ le **min** : plus petite valeur observée

↳ le **max** : plus grande valeur observée

Rem : attention à l'interprétation abusive et potentiellement trompeuse de ces indicateurs

ex : 1 smartphone à 1990 €

↳ très exceptionnel

Déf : Le premier quartile Q_1 est la plus petite des valeurs observées telle que 25% au moins des observations lui soient inférieure ou égales

Le troisième quartile Q_3 est la plus petite des valeurs observées telle que 75% au moins des observations lui soient inférieure ou égales

↳ Définition similaire à la Médiane : 25% → Q_1, 50% → Q_2 = médiane, 75% → Q_3

↳ Coupent les observations en 4 parties d'effectifs (à peu près) égaux

↳ Q_1 : observation de rang $\begin{cases} \frac{N}{4} & \text{si } N \text{ multiple de } 4 \\ \lfloor \frac{N}{4} \rfloor + 1 & \text{sinon} \end{cases}$

$\lfloor x \rfloor$ = partie entière

↳ Q_3 : $\begin{cases} \frac{3N}{4} & \text{si } N \text{ multiple de } 4 \\ \lfloor \frac{3N}{4} \rfloor + 1 & \text{sinon} \end{cases}$

↳ À partir du tableau : $Q_1 \equiv$ première modalité x_i telle que $f_i^{\uparrow} \geq 0{,}25$

$Q_3 \equiv$ " " $\geq 0{,}75$

↳ Plus généralement, pour $p \in [0;1]$

le quantile d'ordre p est la première modalité x_i telle que $f_i^{\uparrow} \geq p$

Fréquents : déciles $d_1; \ldots; d_9$ $(p = 0{,}1; \ldots 0{,}9)$

centiles $c_{11}; \ldots; c_{99}$ $(p = 0{,}01; \ldots; p = 0{,}99)$

coupent en $\frac{1}{p}$ classes d'effectifs (à peu près) égaux