R2.08 Outils numérique pour la statistique descriptive

↳ 3 amphis
↳ 3x 2h de TP en salle machine en lien avec la SAÉ 2.4

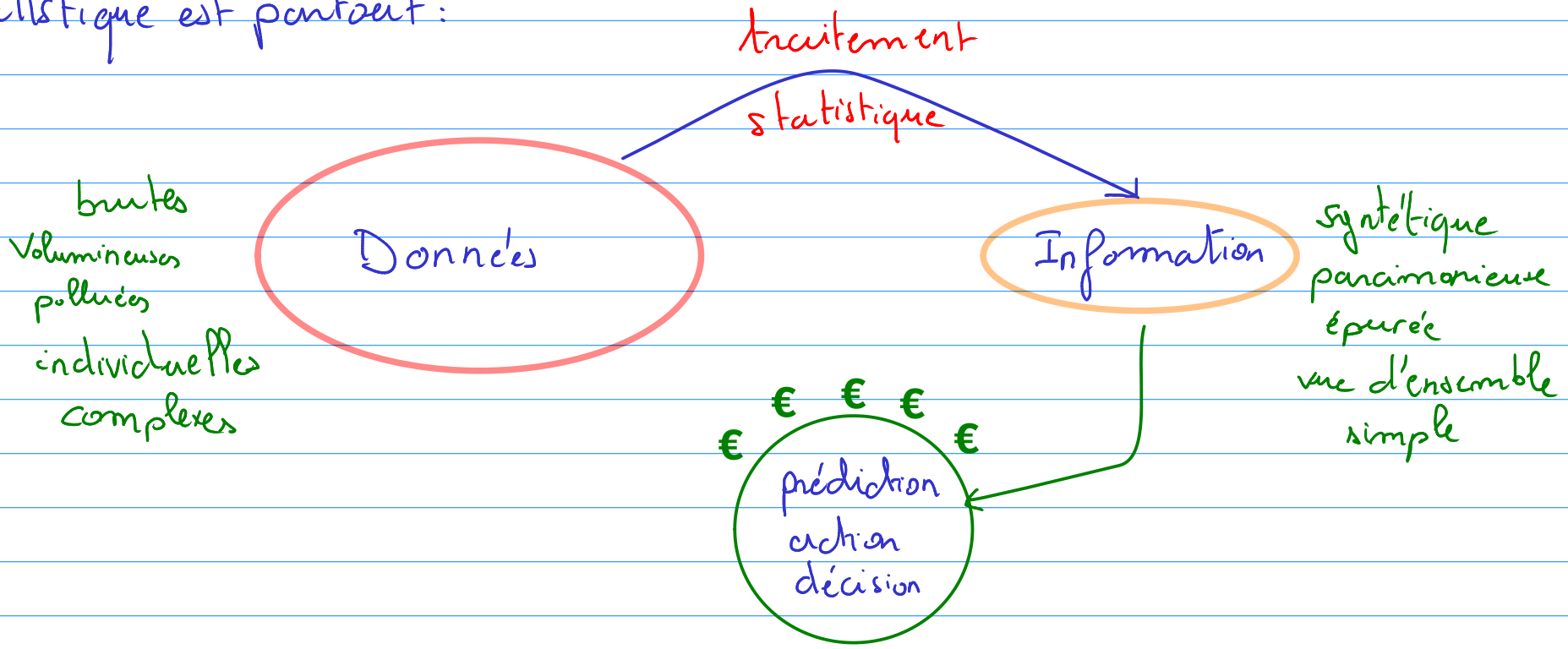
Évaluation :

↳ 1 qcm essentiellement centré sur le cours (28 mars ?)

↳ 1 dossier SAÉ (2 avril)

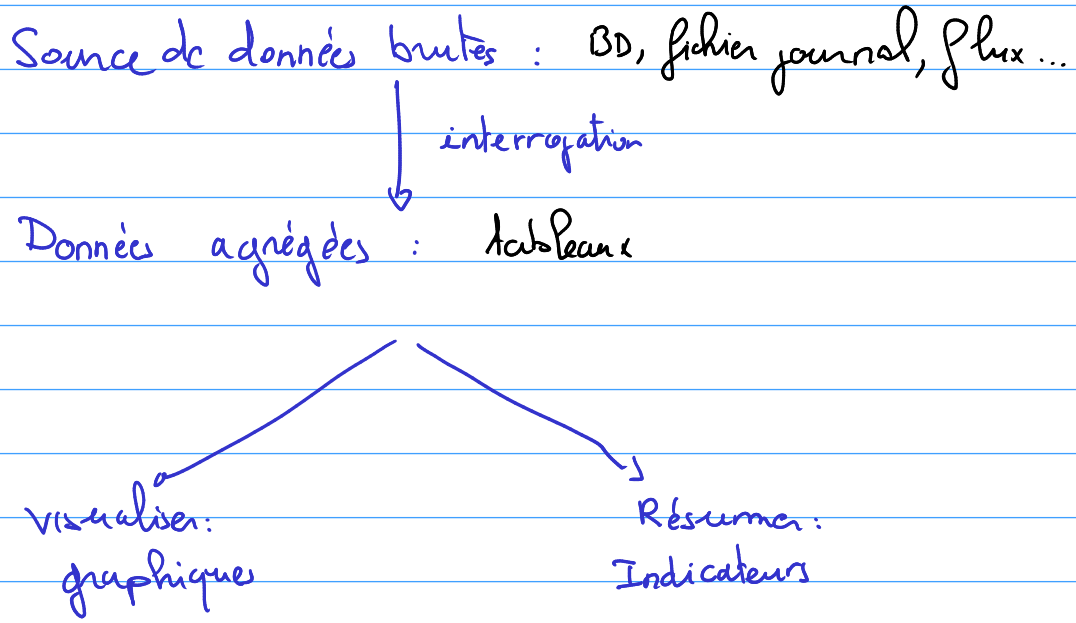
Statistique descriptive

La statistique est partout :



Santé, sport, finances, ingénierie, sciences de la vie, agronomie, écologie, sociologie, psychologie, réseaux, marketing, qualité de production, jeux vidéo, ...

⚠ Cas d'utilisation : révèle ou dissimule des vérités importantes



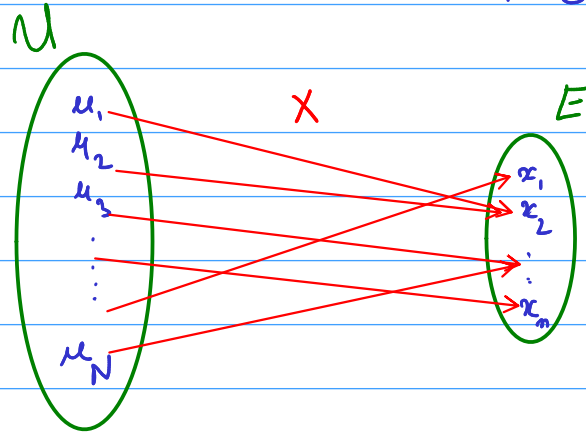
Ex : Frontmandbet vend des smartphones reconditionnés

Vue de la BD "Stocks" :

Id	Marque	Modèle	Année	OS	Cap	Couleur	Réseau	Vérouillage	Condition	Garantie	livraison offerte	Prix	...	taille écran	poids	...
x2298A	Apple	iPhone12	2020	iOS	64	Bleu	4G,5G	déblocqué	Etat correct - parfait état	6 mois	oui	602€		6.1	164	
b3125	Samsung	S10	2019	Android	128	Noir	4G	débloqué		12 mois	non	321€		6.1	157	
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮		⋮	⋮	

→ Visualer l'état du stock
→ Comparer l'offre par catégorie/ état général / os ... à la concurrence } ⇒ caractériser l'adéquation à la demande
→ Suivre des indicateurs clés résumant l'offre (prix, techno, âge etc..) } € € € €

Modèle et vocabulaire



- Population $\mathcal{U}$: ensemble des entités soumises à l'étude
- Individus u_i : entités élémentaires
- Caractère X : critère observable sur les individus
- Modalités x_i : valeurs prises par le critère

Rem : ↳ La taille de la population N est supposée **grande** pour que l'étude ait un intérêt
↳ On mesure souvent **plusieurs** caractères chez le **même** individu (cf ex)
↳ L'objet de l'étude est le caractère. Les individus sont insignifiants
↳ Formalisme mathématique : caractère ≡ fonction

$$X : \mathcal{U} \longrightarrow E$$
$$u \longmapsto X(u) \ (= x)$$

Ex de Frontmarket

↳ $\mathcal{U} = \{x229A, b328, \dots 432onc\}$ = {smartphones en stock}

↳ un individu ≡ un smartphone en stock

↳ plusieurs caractères observés : ↳ Marques : $\mathcal{U} \longrightarrow E = \{$"Apple", "Samsung", ...$\}$

↳ Années : $\mathcal{U} \longrightarrow E = \{2014; 2015; \dots, 2021\}$

↳ Condition : $\mathcal{U} \longrightarrow E = \{$"correct", "bon", "excellent", "neuf"$\}$

↳ Prix : $\mathcal{U} \longrightarrow E = [25; 1256]$ (€)

↳ Poids : $\mathcal{U} \longrightarrow E = [115; 234]$ g

⋮

Voc : ↳ 1 seul caractère = étude statistique **univariée** (cette ressource)

2 caractères " " **bivariée**

+de 2 " " " **multivariée**

Type d'une variable

↳ traitement différent pour { le caractère OS → 2 modalités : iOS, Android
                            { le caractère prix → beaucoup de prix possibles

Déf
↳ Un caractère est . quantitatif si les modalités sont des **nombres**
                   . qualitatif sinon

↳ Un caractère quantitatif est . discret si l'ensemble des modalités possibles est fini ou dénombrable
                               . continu sinon

Rem :
↳ on parle aussi de caractère qualitatif . nominal si les modalités ne sont pas ordonnées
                                         . ordinal sinon

↳ Un caractère quantitatif est aussi appelé variable statistique
↳ la distinction discret/continu est parfois difficile si $|E|$ est fini mais très grand

Ex de Frontmarket : Os, Marque, Modèle, : qualitatif (nominal)
Condition : qualitatif (ordinal)
Année, capacité : quantitatif discret
Taille écran, poids, prix : quantitatif continu

choix discutable

Rem : Le prix (par ex) est discret si on se limite au centime d'€. On le traitera comme une variable continue car les modalités sont très nombreuses

Étape 0 : obtenir une série statistique

Déf : une série statistique est la liste des valeurs observées pour chaque individu, pour un caractère donné

Ex de Frontmarket : [iOS, Android, Android, Android, iOS, ... , iOS,] caractère OS
[134, 122, 212, ... 194] caractère poids

Rem : Liste ⇒ valeurs non ordonnées et non uniques

Étape 1 : obtenir des tableaux
(Résumer sans perte)

↳ Dorénavant : on supposera toujours donnée une série $[\dots]$ pour un caractère donné X

↳ Rappel : les individus sont insignifiants sans ordre

Série $[x_3, x_2, x_3, x_1, x_2, x_1 \dots x_3] \equiv [\underbrace{x_1, x_1, \dots, x_1}_{\text{groupes}}, \underbrace{x_2, x_2, \dots, x_2}_{\text{groupes}}, \underbrace{x_3 \dots x_3}_{\text{groupes}}]$

Déf : On appelle **effectif** de la modalité x_i le **nombre d'occurences** de x_i
" " **fréquence** " " " **la proportion d'occurences** de x_i

↳ par extension, si $A \subset E$ on appellera **fréquence de A** le nombre d'occurences de toutes les modalités de A.

Tableau statistique

Modalités	x_1	x_2	$\dots$	x_i	$\dots$	x_m	E
Effectif	n_1	n_2	$\dots$	n_i	$\dots$	n_m	N
fréquence	f_1	f_2	$\dots$	f_i	$\dots$	f_m	1

N : taille de la pop°

Rem :
- Les modalités sont ordonnées par ordre croissant si la variable est quantitative ou qualitative ordinale
- Les effectifs somment à N
  Les fréquences somment à 1
- Pas de perte d'information sur le caractère, mais seulement sur les individus (insignifiants)
- Le tableaux résume toute l'information utile ⇒ pas d'autre interrogation de la source de données

Ex de Frontmarket : Caractère OS :

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

↳ affichage en % ou pas (0,099)
↳ attention aux arrondis : il faut sommer à 1

regroupement de modalités peu nombreuses

Caractère Année

An	≤ 2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	
n_i	56	105	178	273	339	494	607	585	416	2	3055
f_i	0,018	0,034	0,058	0,089	0,111	0,162	0,199	0,191	0,136	0,001	

Tableaux de variables continues (regroupement en classe avec pertes)

↳ nb de modalités > 20 $\Rightarrow$ Tableaux illisibles

↳ On regroupe les modalités dans des classes = Intervalles $[a_i ; a_{i+1}[$

Rem : le choix des classes
- est non unique
- est arbitraire
- peut { cacher la structure du caractère
  mettre en évidence " "

ex de Frontmarchet

- trop peu de classes
- mal choisies ?

Prix :	$[0;100[$	$[100;200[$	$[200;500[$	$[500;2000]$	$305]$
n_i	342	1031	1276	406	
f_i	11%	34%	42%	13%	1

↳ choix possibles :
- même amplitude ($a_{i+1} - a_i = P$)
- même fréquence avec la classes ($f_i = \frac{N}{k}$, $i = 1$ à k)

Rem : ↳ Regroupement en classes ⇒ perte (modérée) d'information
↳ Les données brutes individuelles sont parfois déjà en classes

Tableaux de cumuls pour variables quanti ou ordinales

↳ Questions du type :
- combien de smartphones ont plus de 3 ans ?
- quelle proportion de smartphones ont au moins 64 Go ?

} à partir des tableaux sans nouvelle requête

Déf : On appelle effectif cumulé croissant de la modalité x_i, noté $n_i^\uparrow$, le nombre d'occurrences de valeurs inférieures ou égales à x_i

. fréquence cumulée croissante de la modalité x_i, notée $f_i^\uparrow$, la proportion d'occurrences de valeurs inférieures ou égales à x_i

Rem ↳
$$n_i^\uparrow = n_1 + n_2 + \dots + n_i$$
$$f_i^\uparrow = f_1 + f_2 + \dots + f_i$$

Ex de Frontmanbet (modalités en colonne pour changer)

Cap en Go

x_i	n_i	$n_i^\uparrow$	f_i	$f_i^\uparrow$
4	22	22	0,007	0,007
8	147	169	0,048	0,055
16	416	585	0,136	0,191
32	451	1036	0,148	0,339
64	581	1617	0,190	0,529
128	957	2574	0,313	0,843
256	363	2937	0,119	0,961
512	107	3044	0,035	0,996
1024	11	3055	0,004	1,000
	3 055		1	

Rem : On définit de la même manière les effectifs et fréquences cumulés décroissants $n_i^\downarrow$ et $f_i^\downarrow$

Ex : pourcentage de smartphones de capacité au moins 128 Go
$= f_i^\downarrow = 47\%$ (tableau à faire en ex)

Cas de données en classes

borne à droite de la classe

↳ définition analogue : $n_i^{\uparrow} = n_1 + \dots + n_i$

↳ lecture différente : pour la classe $[a_i ; a_{i+1}[$, $n_i^{\uparrow}$ = nb de valeurs $< a_{i+1}$

↳ à adapter si classe $]a_i ; a_{i+1}]$ ---- $\leq a_{i+1}$

ex de Frontmarket

Prix :	$[0;100[$	$[100;200[$	$[200;500[$	$[500;2000]$	
n_i	342	1031	1276	406	3055
$n_i^{\uparrow}$	342	1373	2649	3055	
f_i	11%	34%	42%	13%	1
$f_i^{\uparrow}$	11%	45%	87%	100%	

45% des smartphones coûtent moins de 200 €

Rem : pour des valeurs dans une classe, on pourrait interpoler (plus fond)

ex : % de smartphones coûtant moins de 250 €

Représentations graphiques

↳ doit être adapté au type de variable : quali
quanti discret
quanti continu

Variable qualitative

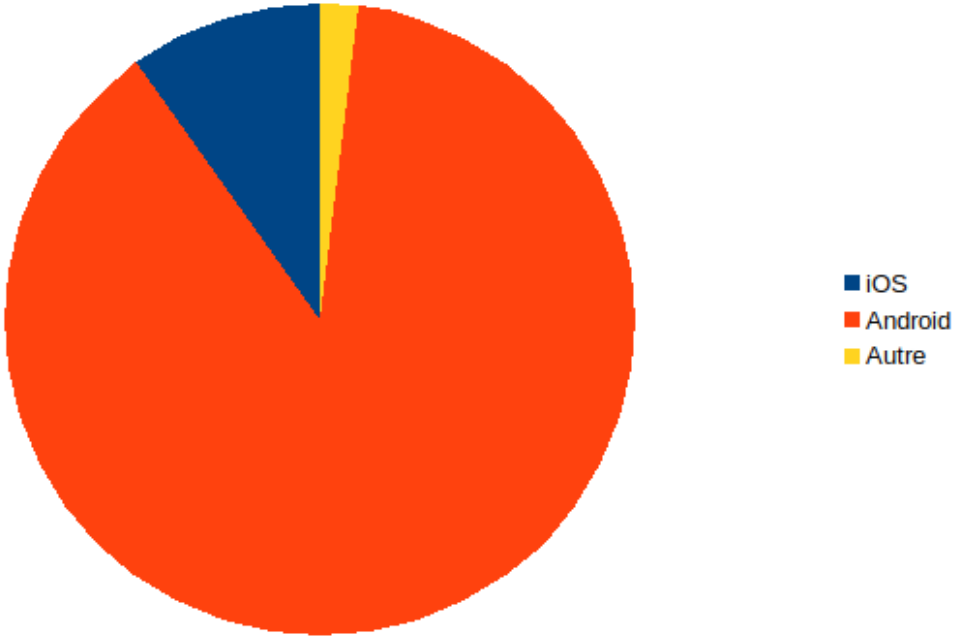
→ diagramme en secteurs (circulaire, camembert, pie chart)

(peu utilisé en sciences, très utilisé en eco, psycho, socio, ...)

Secteurs d'un disque de surfaces proportionnelles aux fréquences

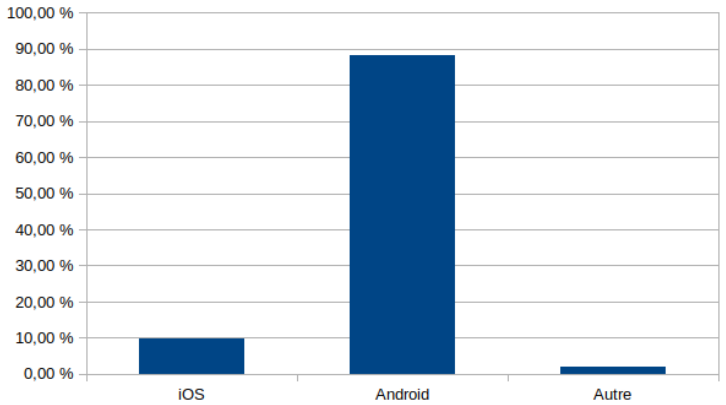
Ex de Frontmarket

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



Rem : ↳ lisible si peu de modalité (< 5 ou 6)
↳ on peut compléter en indiquant effectifs ou fréquences
→ Diagramme en bâtons (en barre, barplot)
x modalités sur un axe
y fréquence (ou effectif) sur l'autre

Ex de Frontmarket



Rem : ↳ Cela ne s'appelle pas un histogramme (confusion dans les media et bureautique)
↳ On ordonne les modalités si la variable est ordinale (état correct, bon état, ...)
↳ éventuellement horizontal (modalités en ordonnées, fréquences en abscisses)
↳ des variantes possibles si +sieurs jeux de données

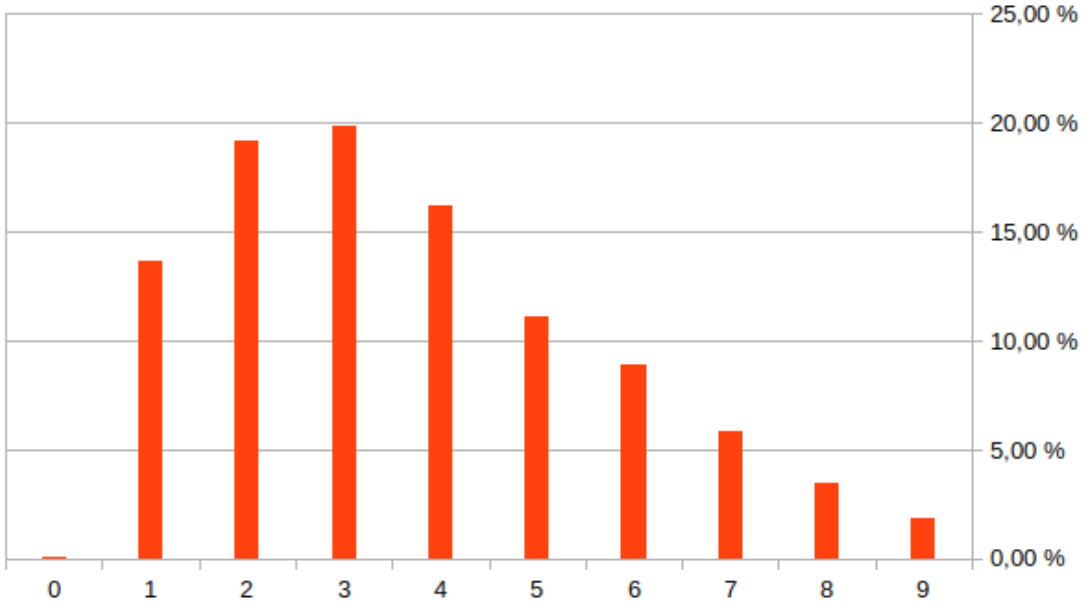
Variables quantitatives discrètes

- Diagramme en bâtons si nb de modalités pas trop grand (< 20)
- Modalités ordonnées et échelle respectées

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



L'échelle des abscisses doit être respectée

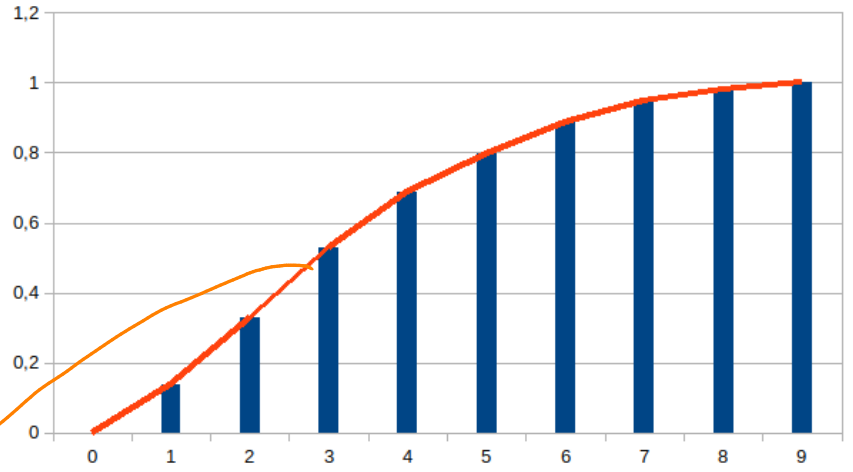
Courbes cumulatives

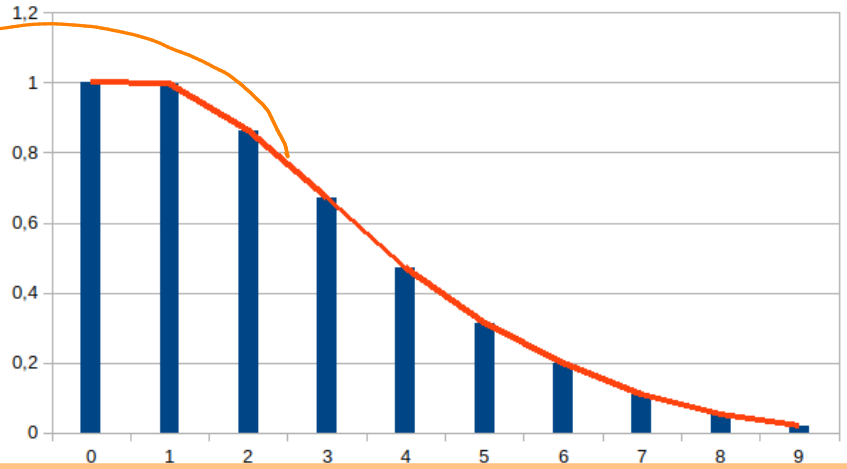
pour variables quantitatives discrètes

- ↳ diagramme en bâtons pour $n_i^{\uparrow}$, $n_i^{\downarrow}$, $f_i^{\uparrow}$ et $f_i^{\downarrow}$
- ↳ ne met pas les hautes et basses fréquences en évidence
- ↳ lecture facile des modalités "seuils" (ex: au moins 80% des individus)

Ex de Frontmarbet pour l'âge

x_i	f_i	$f_i^{\uparrow}$	$f_i^{\downarrow}$
0	0,07 %	0,07 %	100,00 %
1	13,62 %	13,68 %	99,93 %
2	19,15 %	32,83 %	86,32 %
3	19,87 %	52,70 %	67,17 %
4	16,17 %	68,87 %	47,30 %
5	11,10 %	79,97 %	31,13 %
6	8,94 %	88,90 %	20,03 %
7	5,83 %	94,73 %	11,10 %
8	3,44 %	98,17 %	5,27 %
9	1,83 %	100,00 %	1,83 %





L'interpolation n'a pas de sens (confort visuel seulement)