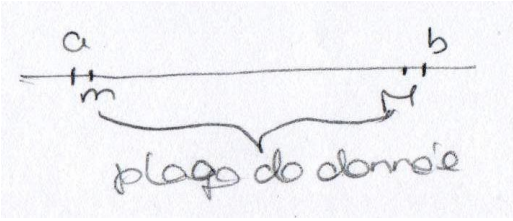
Statistique

Comment regrouper les données choisies ?

a. Combien de classes k ?
b. Une fois connu k, comment déterminer les limites de classe ?

a. -> règle de Sturge :
   Soit n le nombre de données
   $k \cong 1 + 3.32 * \log_{10} n$
   ⚠ k le nombre de classe est un entier

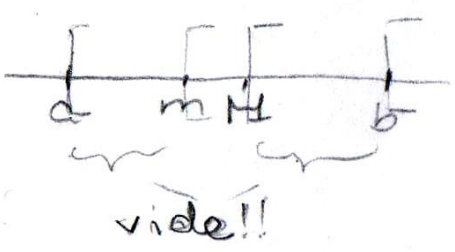
b. Soit $\varepsilon > 0$ “très petit”
   Soit m la plus petite donnée observée
   Soit M la plus grande donnée observée
   On pose a = m – ε
   b = M + ε



   c = $\frac{b-a}{k}$ amplitude de chacune des classes
   Les limites de classes : l_1, …, l_{K+1} sont définies par
   l_1 = a et $\forall\, p \in [2, k+1]$, $l_p = l_{p-1} + c = a + (p-1) * c$
   ➔ Les k classes adaptées au jeu de données sont $[l_1, l_2[$, $[l_2, l_3[$, …, $[l_k, l_{k+1}[$

Remarque :

- l_{k+1} = b
- $\varepsilon = \frac{M-m}{1000}$ idée pour ε
- Si ε très grand :



IV. Les indicateurs statistiques

3 types de données :

- Données brutes : pour toutes les natures de variable
- Données individuelles : pour variables qualitatives ou quantitatives discrètes
- Données regroupées : pour les quantitatives continues

- Données brutes : x_i avec i ∈ {1, …, n}
- Données individuelles $(x_k, n_k)_{k \in \{1,\dots,p\}}$
  $\{x_1, \dots, x_p\}$ les différentes réponses
  N_k effectif associée à x_k
- Données regroupées ; $([c_i, c_{i+1}[, n_i)_{i \in \{1,\dots,k\}}$
  Avec $[c_1, c_2[, \dots, [c_k, c_{k+1}[$ les k classes
  N_i l'effectif associée à la ieme classe

Ex : étude couleur des yeux en région PACA :

- Blanc, vert, bleu, marron, vert, noir, bleu, bleu, vert, marron

➔ Données brutes

Bleu	4
Marron	2
Vert	3
Noir	1

➔ Données individuelles

1° cas : si on a des variables qualitatives, il n'y a qu'un indicateur statistique possible -> mode

Dans ce cas :

Mode : la valeur la plus représentée dans le jeu de donnée : le plus grand effectif (grande fréquence)

Sur l'exemple précédent, mode = bleu

2° cas : variables quantitatives

- Moyenne (arithmétique) $\bar{x}$

Formules :

- Sur données brutes : $\bar{x} = \frac{1}{n} \sum_{1=1}^{n} x_i$
- Sur données individuelles : $\bar{x} = \frac{1}{n} \sum_{1=1}^{n} n_i * x_i = \sum_{1=1}^{n} f_i * x_i$
- Sur données regroupées : $\bar{x} = \frac{1}{n} \sum_{1=1}^{n} n_i * c_i$
  Avec c_i le centre de la ieme classe
  Si la ieme classe est $[l_i, l_{i+1}[$, $c_i = \frac{l_{i+1} - l_1}{2}$

Inconvénients :

- Sensibilité aux données « extrêmes »

Ex : nombre d'enfant pas famille :

Jeu 1 : 0 1 2 0 2 1 4 3 2 1 $\bar{x} = \frac{16}{10} = 1.6$

Jeu 2 : 0 1 2 0 2 1 4 3 2 10 $\bar{x} = \frac{25}{10} = 2.5$

	[0, 2.5[	[2.5, 5[	[5, 7.5[	[7.5, 10[
Jeu 1	8	2	0	1
Jeu 2	7	2	0	1

Jeu 1 : $\bar{x} = \frac{1}{10} * (8 * 1.25 + 2 * 3.75) = 1.75$

Jeu 2 : $\bar{x} = \frac{1}{10} * (7 * 1.25 + 2 * 3.75 + 1 * 8.75) = 2.5$

Médiane : (second quartile)

Def : grandeur observable qui sépare la population en 2 parties de même effectif

Pratique :

Cas discrète

Valeurs	Fréquences cumulées
x_1	F_1
…	
x_p	$F_p = 1$

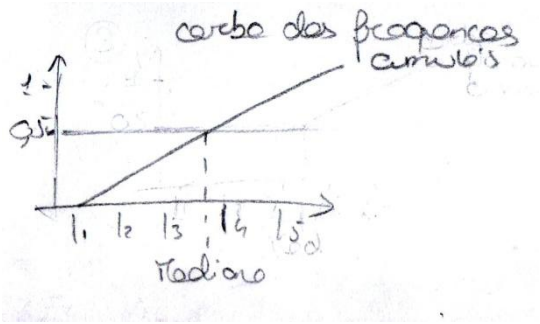
La première valeur pour laquelle la fréquence cumulée est ___ supérieure à 0.5

Cas continue

Classes	Fréquences cumulés
$[l_1, l_2[$	F_1
$[l_2, l_3[$	F_2
…	…
$[l_k, l_{k1}[$	$F_k = 1$

1° étape : on cherche la première classe pour laquelle la fréquence cumulée est > 0.5

➔ Classe $[l_i, l_{i+1}[$
Médiane ∈ $[l_i, l_{i+1}[$



Med = $l_1 + (l_{i+1} - l_i) * \frac{(0.5 - F(l_i))}{F(l_i) - F(l_i)}$

Remarque : $F(l_{i+1}) > 0.5$, et $F(l_i) <= 0.5$

Quartiles : Q_1, Q_2, Q_3

Méthode : idem que médiane avec

- pour Q_1 au lieu de considérer le seuil de 0.5, on considère le seuil 0.25
- pour Q_3 idem mais aussi 0.75 au lieu de 0.5

Boite à moustaches (box plot)