

Cours 1 : Statistique descriptive

1. Terminologie

Population : ensemble des individus observables (Ex. patients hôpital, étudiants licence 1)

Echantillon : ens. individus observés. (Ex. groupe classe de 30 étudiants)

Taille : nombre d'individu de l'échantillon.

Caractère : grandeur ou attribut attaché à un individu

caractère

- qualitatif : groupe sanguin
- quantitatif : poids
- discret : groupe sanguin
- continu : poids, âge

Modalités d'un caractère : ensemble des valeurs possibles :

ex. gre sanguin : $\{O, A, AB, B\}$.

2. Représentation des données

Série brute : en liste suivant l'ordre d'acquisition.

Ex. 1

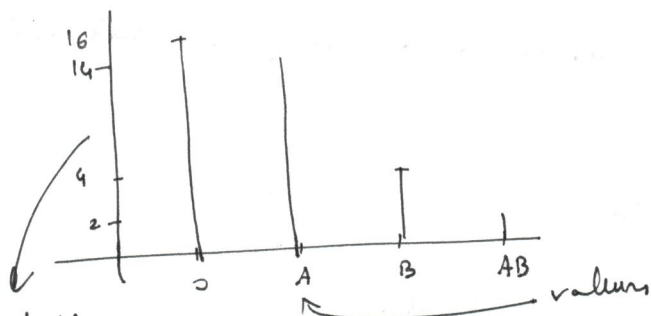
Tableau valeurs - effectifs (le caractère prend un nombre fini et petit de valeurs)

valeurs x_i	0	A	B	AB
effectifs n_i	16	14	4	2
fréquences f_i	$\frac{16}{36}$	$\frac{14}{36}$	$\frac{4}{36}$	$\frac{2}{36}$

$$\text{fréquence } f_i = \frac{n_i}{n_{\text{tot}}} = \frac{n_i}{n_1 + n_2 + \dots + n_r}$$

taille de l'échantillon

Diagramme en bâtons : pour un caract. qualitatif



effectifs
ou fréquences

Tableau Classes - effectifs quand le caractère prend un grand nombre de val., éventuellement infini.

Ex. 2

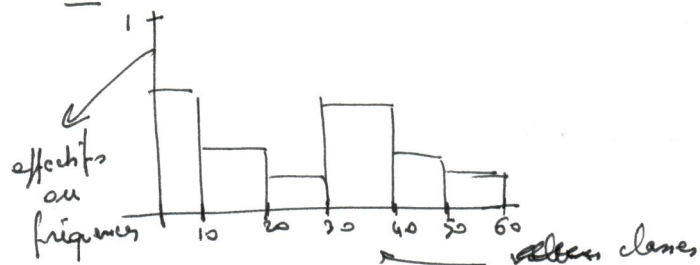
population	$[0, 10[$	$[10, 20[$	$[20, 30[$	$[30, 40[$	etc.
effectifs	10	6	2	2	
fréquences	$\frac{10}{24}$	$\frac{6}{24}$	$\frac{2}{24}$	$\frac{2}{24}$	

Rem.
classes
d'amplitude
constante

Classe: $[x_{i-1}, x_i[$ intervalle
amplitude: $x_i - x_{i-1}$, milieu: $\frac{x_i + x_{i-1}}{2}$

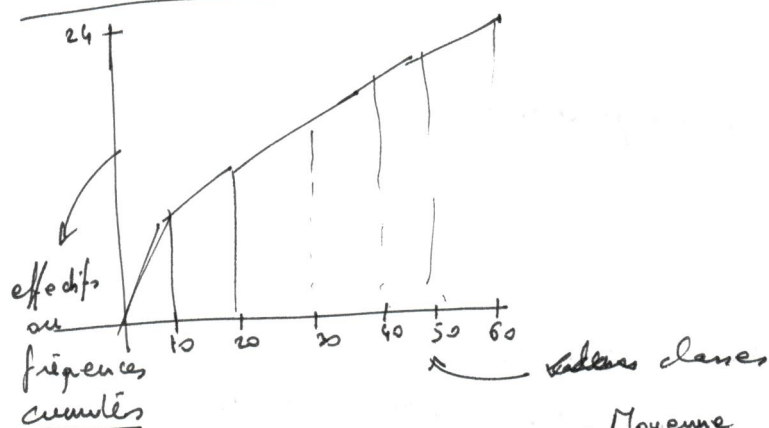
histogramme: pour des ~~cases~~ données par classe.

l'aire du rectangle doit être proportionnelle à l'effectif (ou à la fréquence) de la classe.



Attention au cas d'amplitude non constante.

Courbe des effectifs (ou fréquences) cumulés:



Indicateurs numériques:

- Moyenne
- médiane et quantiles
- écart-type et variance

Moyenne: notée $\bar{x}$

pour série brute $x_1, \dots, x_n$ c'est $\bar{x} = \frac{1}{n} (x_1 + \dots + x_n)$
 pour valeurs-effectifs (x_i, n_i) c'est $\bar{x} = \frac{1}{n} (x_1 n_1 + x_2 n_2 + \dots + x_p n_p)$

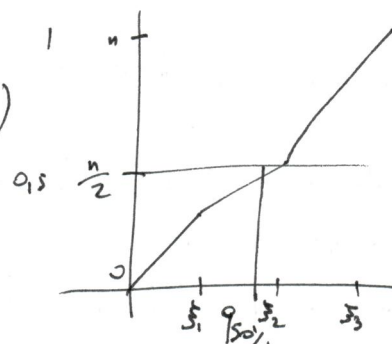
avec $n = n_1 + \dots + n_p$

Médiane: notée $q_{50\%}$: c'est la valeur qui partage l'échantillon en deux parties, une partie moitié qui a valeurs $\leq q_{50\%}$, l'autre qui a valeurs $\geq q_{50\%}$.

i) On réordonne la série brute: $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$

$$q_{50\%} = \begin{cases} x_{(k+1)} & \text{si } n = 2k+1 \text{ impair} \\ \frac{x_{(k)} + x_{(k+1)}}{2} & \text{si } n = 2k \text{ pair} \end{cases}$$

Attention, c'est



ii) graphiquement, sur la courbe des effectifs (ou fréquences) cumulés

iii) Si classes - effectifs (ou fréq.)
 n_i f_i

On ~~cherche~~ d'abord la classe où se trouve $q_{50\%}$. $[f_{i-1}, f_i]$

ensuite on a une formule d'interpolation

$$\frac{q_{50\%} - f_{i-1}}{f_i - f_{i-1}} = \frac{\frac{n}{2} - n_{i-1}}{n_i - n_{i-1}}$$

Quantiles ($q_{25\%}$, $q_{50\%}$ ← médiane, $q_{75\%}$) ils partagent les valeurs observées en quatre parties de même effectif.

Méthodes de calcul :

i) Série brute. On réordonne la S.B. $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$.

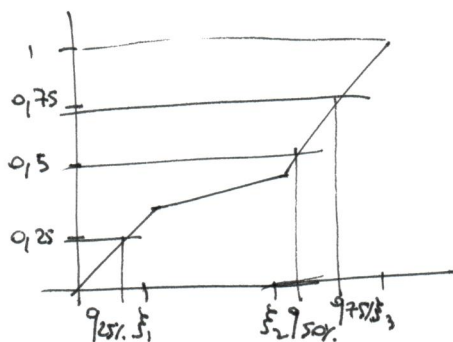
$$q_{25\%} = x_{\left(\frac{n}{4}\right)}$$

$$q_{75\%} = x_{\left(\frac{3n}{4}\right)}$$

Attention. 1) si $\frac{n}{4}$, $\frac{3n}{4}$ pas entiers, on prend l'entier supérieur.

2) Cette-ci est une convention qui n'est pas universelle.

ii) graphiquement

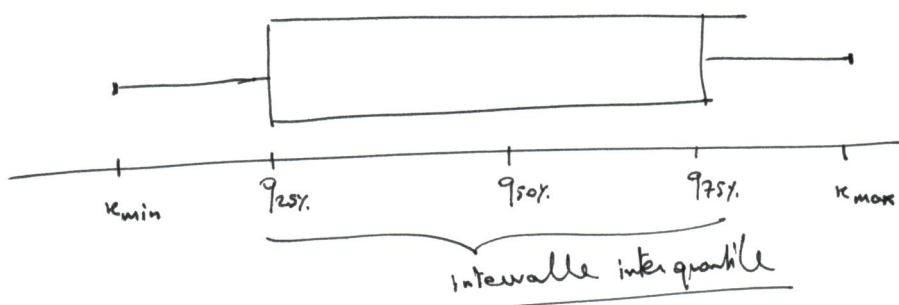


iii) Si classes - effectifs (ou fréq.) par interpolation, avec les formules

$$\frac{q_{25\%} - f_{i-1}}{f_i - f_{i-1}} = \frac{0,25 - f_{i-1}}{f_i - f_{i-1}} = \frac{\frac{n}{4} - n_{i-1}}{n_i - n_{i-1}}$$

et sim. avec $\left(q_{75\%}, 0,75, \frac{3n}{4}\right)$

Boxplot
Boxplot ou boîte à moustaches



Variance s_{n-1}^2 (Attention, convention non universelle)

mesure de la dispersion autour de la moyenne.

~~et par~~ série brute: $s_{n-1}^2 = \frac{1}{n-1} \left((x_1 - \bar{x})^2 + \dots + (x_n - \bar{x})^2 \right)$

écart-type (standard deviation) la racine de la variance $s_{n-1} = \sqrt{\frac{1}{n-1} (\dots)}$