Statistique univariée

- Diagramme en bâton
- Histogramme
- Caractéristiques de tendances centrales
- Caractéristiques de dispersion

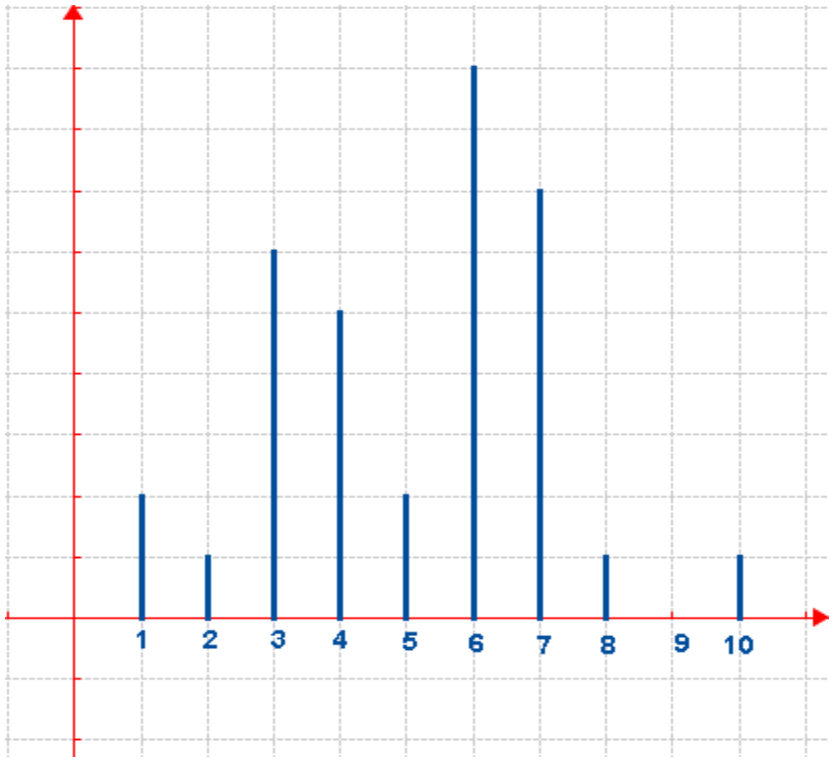
Effectifs et fréquences

- Un calcul peut être effectué sur tous les caractères : Compter le nombre d'unités statistiques qui ont une modalité donnée.
- On utilisera le terme effectif de la modalité i. On notera cet effectif n_i.
- On aura besoin de ramener les effectifs en pourcentage. On parlera alors de fréquences. On notera cette fréquence f_i.
- $f_i = n_i / n$ où n est l'effectif total.
- Remarque : n_i est aussi appelé fréquence absolue et f_i fréquence relative.

Diagramme en bâton

- Il concerne tous les types de caractères sauf le caractère quantitatif continu.
- C'est une représentation planaire.
  - Sur l'axe des x on a les différentes modalités du caractère
  - sur l'axe des y on a la fréquences relatives ou absolues des modalités

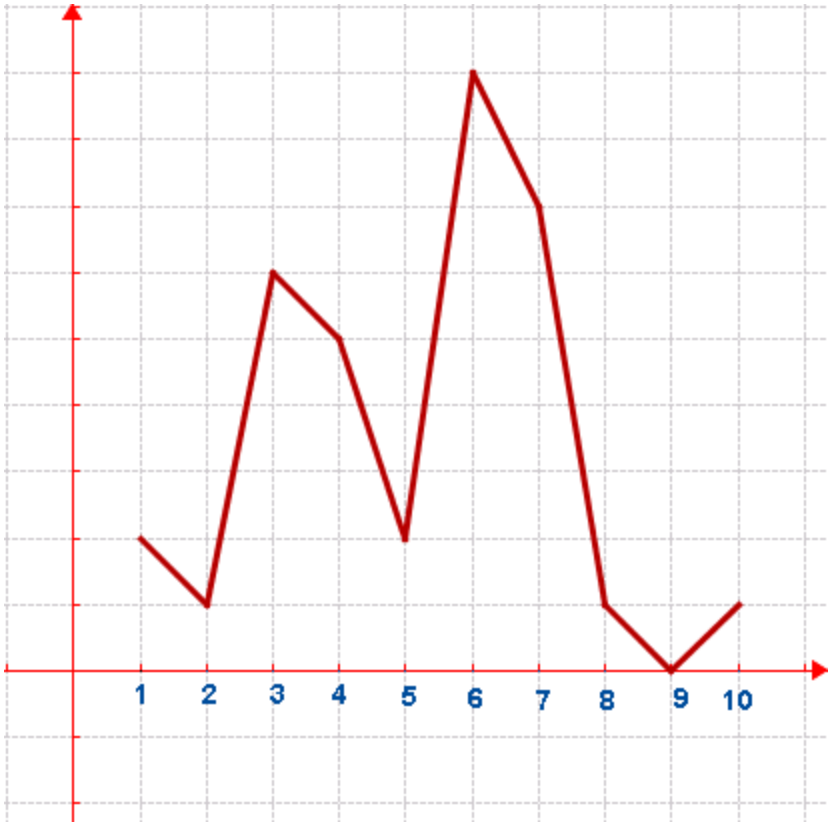
Note	1	2	3	4	5	6	7	8	9	10
n_i	2	1	6	5	2	9	7	1	0	1



Polygone de fréquences

- Il concerne les caractères de type qualitatif ordinal ou quantitatif discret.
- On représente dans le plan le polygone (mod_i, n_i)

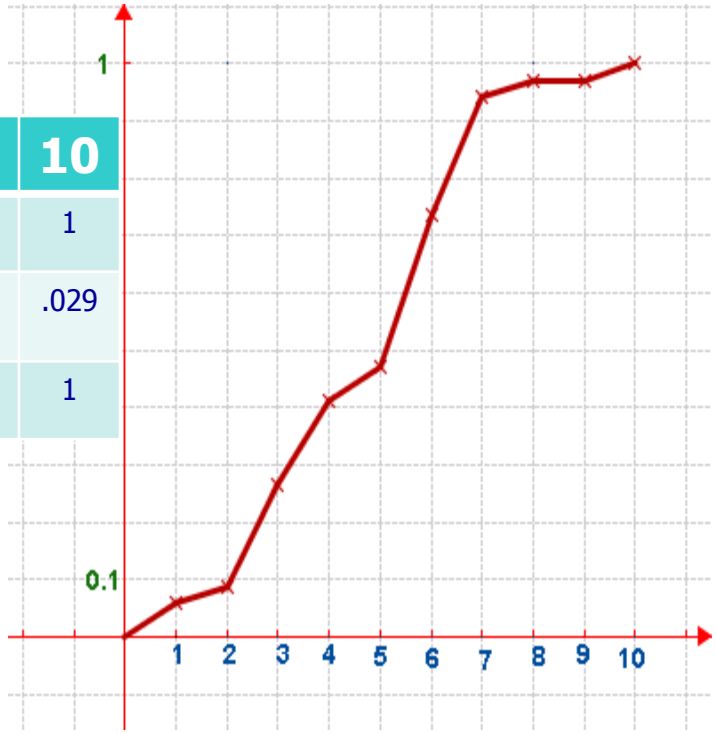
Note	1	2	3	4	5	6	7	8	9	10
n_i	2	1	6	5	2	9	7	1	0	1



Courbe de fréquences cumulées

- Il concerne les caractères de type qualitatif ordinal ou quantitatif discret.
- On représente dans le plan le polygone (mod_i,fc_i) où fc_i est la fréquence cumulée des modalités inférieures ou égales à mod_i.

Note	1	2	3	4	5	6	7	8	9	10
n_i	2	1	6	5	2	9	7	1	0	1
f_i	.059	.029	.176	.147	.059	. 264	.206	.029	.0	.029
fc_i	.059	.088	.264	0.411	.470	.734	.940	.969	.969	1



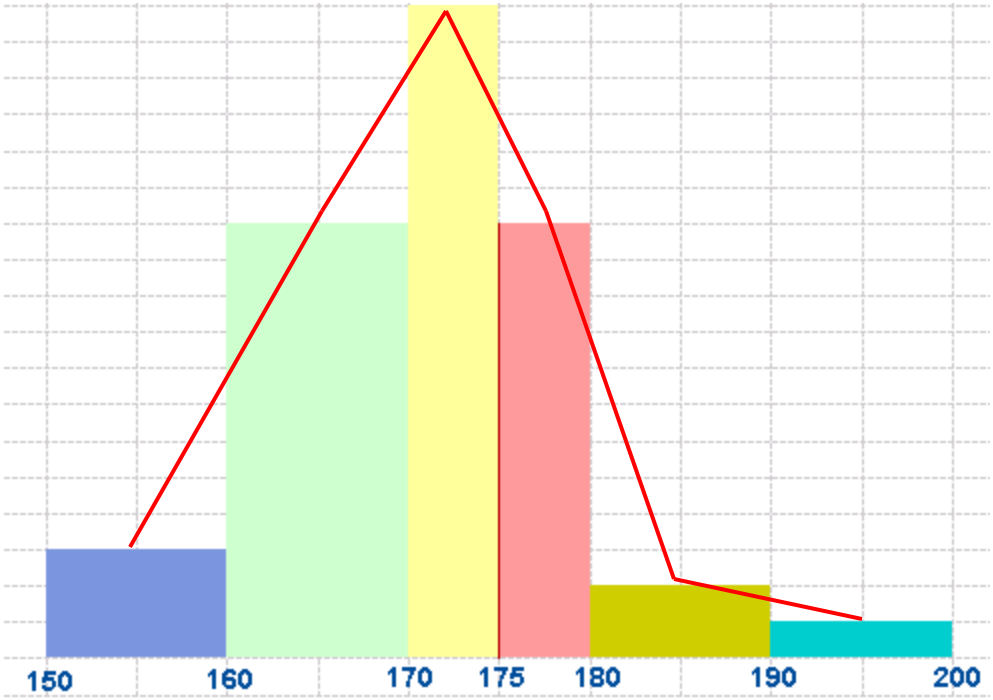
Histogramme

- Il concerne les caractères de type quantitatif continu et accessoirement de type quantitatif discret.
- On partitionne l'ensemble des valeurs du caractères en une famille finie d'intervalles contigus. Les valeurs ponctuelles n'ont pas de sens dans ce graphique.

Taille	[150, 160[	[160, 170[	[170, 175[	[175, 180]	[180, 190]	[190, 200]
Eff.	3	12	9	6	2	1

Dans un plan on place sur l'axe des x les différentes bornes des intervalles$[a_i, a_{i+1}[$. Pour chaque intervalle, on trace un rectangle dont la base est $[a_i, a_{i+1}[$ et **la surface est proportionnelle à n_i** où n_i est le nombre d'individus dont la modalité sur le caractère appartient à $[a_i, a_{i+1}[$.

Comme pour le diagramme en bâton, on peut tracer un polygone de fréquences. La base en x est le milieu des intervalles. Voir la ligne polygonale en rouge sur le graphique ci-contre.



Résumé des représentation graphiques

Variables qualitatives

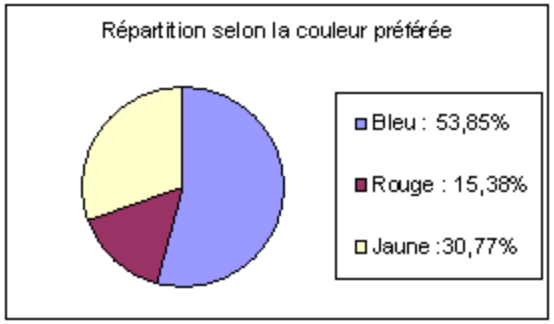
Variables quantitatives

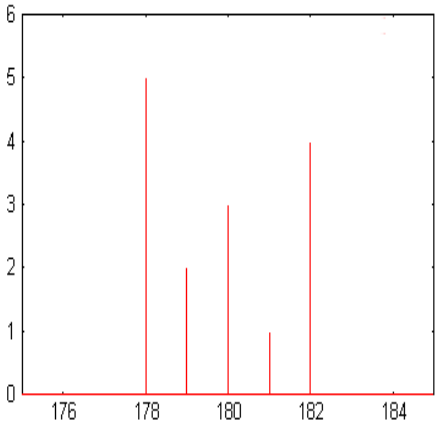
nominales

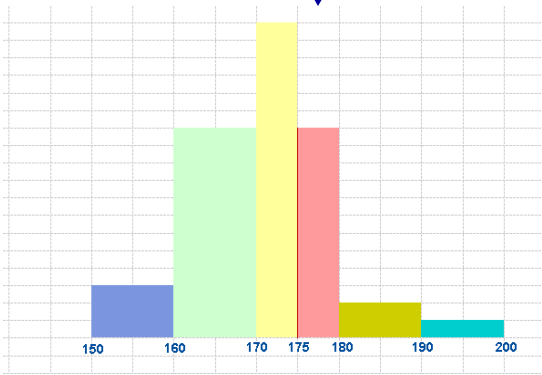
ordinales

discrètes

continues





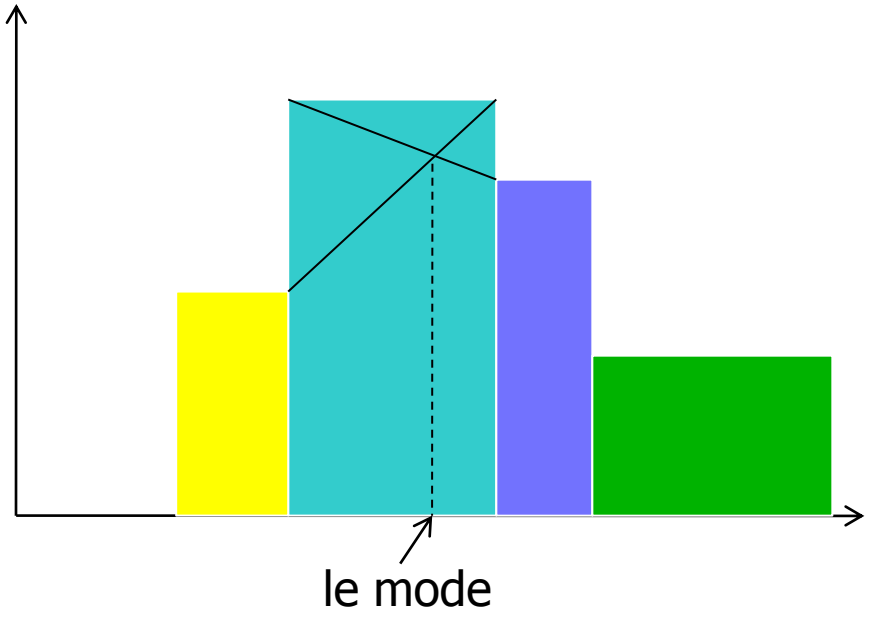


Indicateurs numériques

- La statistique descriptive synthétise les données pour les rendre intelligibles.
- Les tableaux et graphiques vus précédemment sont les premières synthèses.
- La suite consiste à calculer des indicateurs numérique.
- Les plus importantes sont :
  - Caractéristiques de tendances centrales pour expliquer ce qui explique principalement le caractère.
  - Caractéristiques de dispersion pour expliquer comment varie le caractère autour des tendances centrales.

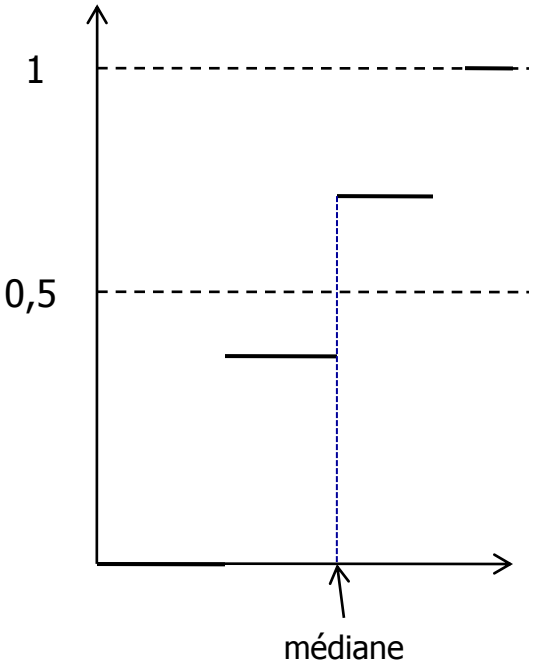
Tendance centrale : le mode

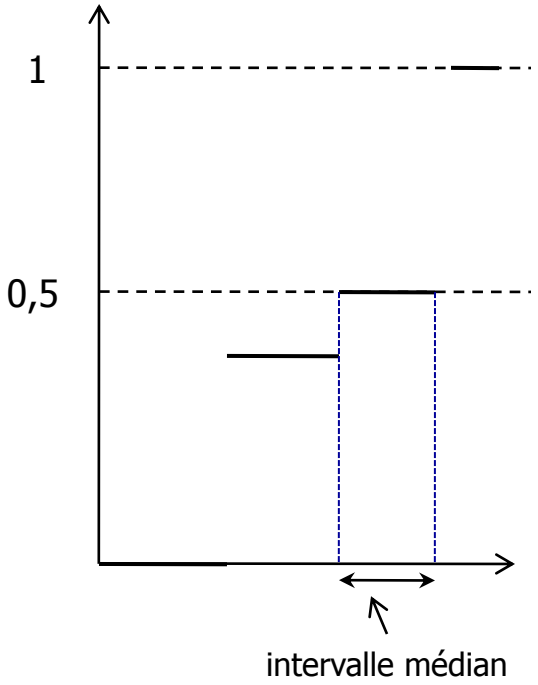
- C'est la modalité observé d'effectif maximum. Elle concerne tous les types de caractères.
- Il sert essentiellement à détecter si la population est homogène ou éventuellement constituée de deux ou plusieurs populations hétérogènes
- Dans le cas du type quantitatif discret, si on a deux effectifs maximum côte à côte on parle d'intervalle modal $[c_i, c_{i+1}]$
- Dans le cas du type quantitatif continu il faut tenir compte des classes adjacentes. Voir graphique ci-contre



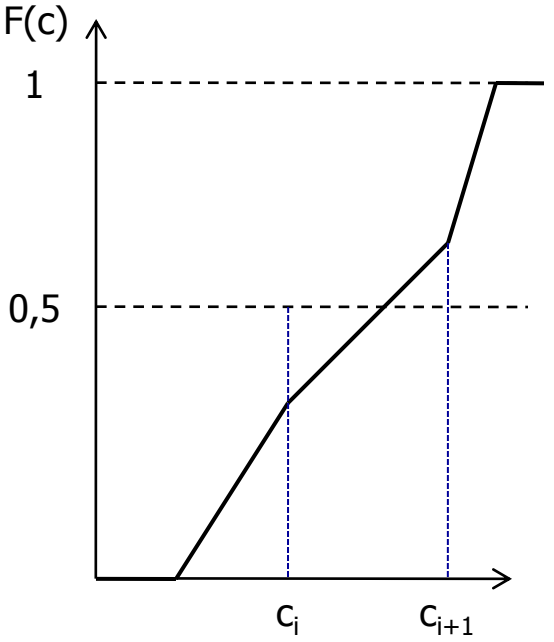
Tendance centrale : Médiane

- La médiane est la modalité qui sépare la population en deux groupes d'effectifs égaux. Elle n'a de sens qu'avec le type qualitatif ordinal et les types quantitatifs.





Type quantitatif discret



$mé = c_i + (0,5 - F(c_i)/(F(c_{i+1})-F(c_i))$

Les valeurs c_i sont les milieux des intervalles

Type quantitatif continu

Tendance centrale : Moyenne

- La moyenne est une valeur calculée qui est la moyenne des modalités c_i pondérée par les effectifs n_i. Elle n'a de sens qu'avec les types quantitatifs.
- Contrairement à la médiane, ce n'est pas nécessairement une modalité du caractère.
- C'est la tendance la plus utilisée pour des raisons algébriques plus que statistique. Elle est facile à calculer et c'est une fonction dérivable (voir analyse multivariée)
- Contrairement à la médiane, elle est très sensible à l'intensité des valeurs du caractère.
- Notation : Si c est le caractère alors la moyenne est notée $\overline{c}$
- En résumé la moyenne et la médiane sont complémentaires.

Moyenne et Médiane : perte d'information

- Quand on passe de la série statistique (c_i) à la moyenne que perd-on comme information ?

Considérons le problème suivant :

$$\min_{x\in R} \mathrm{err}_2(x) = \frac{1}{n}\sum_i n_i (c_i - x)^2$$

$$\mathrm{err}_2'(x) = 0 \Leftrightarrow -\frac{2}{n}\sum n_i (c_i - x) = 0 \Leftrightarrow \sum n_i \cdot c_i = \sum n_i . x = n.x \Leftrightarrow x = \frac{1}{n}\sum n_i \cdot c_i = \bar{c}$$

Choisir la moyenne comme résumé consiste à minimiser l'erreur quadratique.

- Quand on passe de la série statistique (c_i) à la médiane que perd-on comme information ?

De même on peut définir

$$\min_{x\in R} err_1(x) = \frac{1}{n}\sum_i n_i |c_i - x|$$

On peut démontrer que dans ce cas la valeur optimale est la médiane.

Choisir la médiane comme résumé consiste à minimiser l'erreur en valeur absolue. On parle dans ce cas d'écart-médian.

dispersion : écart-type, écart-médian

- Quand on utilise la moyenne comme tendance centrale, il faut utiliser l'écart-type du caractère comme indicateur de dispersion noté $\sigma(C)$

$$\text{Variance}(C) = \frac{1}{n}\sum_i n_i (c_i - \bar{c})^2 \text{ et } \sigma(C) = \sqrt{\frac{1}{n}\sum_i n_i (c_i - \bar{c})^2}$$

- Quand on utilise la médiane comme tendance centrale, il faut utiliser l'écart-médian du caractère comme indicateur de dispersion

$$em(C) = \frac{1}{n}\sum_i n_i |c_i - me|$$

- Dans tous les cas de figures, on peut utiliser l'étendue

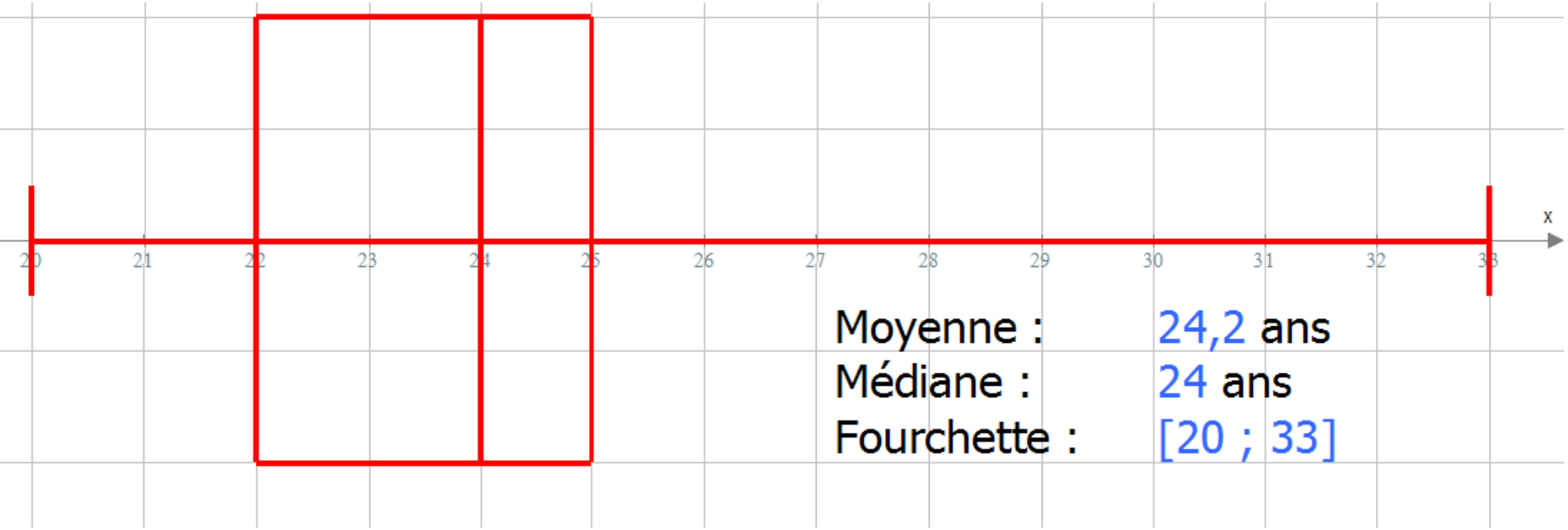
$$et(C) = \text{Max}_i\ (c_i) - \text{Min}_i\ (c_i)$$

Dispersion : Etendue et quantiles

- Dans tous les cas de figures, on peut utiliser l'étendue du caractère $et(C) = \mathrm{Max}_i\ c_i - \mathrm{Min}_i\ c_i$
- Les quantiles : On ordonne les valeurs du caractères et on divise cette série ordonnée en paquets égaux. Chaque modalité extrême de chaque paquet est un quantile.
- Quand on divise en quatre paquets, on obtient les quartiles. Le 2ème quartile est la médiane. Le 1er quartile découpe la population en deux sous populations : la première sous population est formée du ¼ des individus, la deuxième est formée des ¾ des individus.
- On utilise souvent le 1er et le 99ème centiles pour éliminer des individus dits aberrants.

Boite à moustache

- C'est un résumé inventé par John Tukey
- Il consiste à afficher sur un graphique cinq indicateurs révélateurs d'un profil : la plus petite valeur, le 1er quartile, la médiane, le 3ème quartile et la plus grande valeur.



Résumé des indicateurs numériques

Série observée

Note	1	2	3	4	5	6	7
Elève L	9	10	8	7	10	9	11
Elève P	14	2	16	5	6	5	16

Série ordonnée

Elève L	7	8	9	9	10	10	11
Elève P	2	5	5	6	14	16	16

Tendance centrale

	Moyenne	Médiane
Elève L	9,1	9
Elève P	9,1	6

Dispersion

	Variance	Ecart-type	Q1	Q3	M_sup	M_inf
Elève L	1,81	1,34	8,5	10	12,25	6,25
Elève P	35,48	5,96	5	16	32,5	-11,5