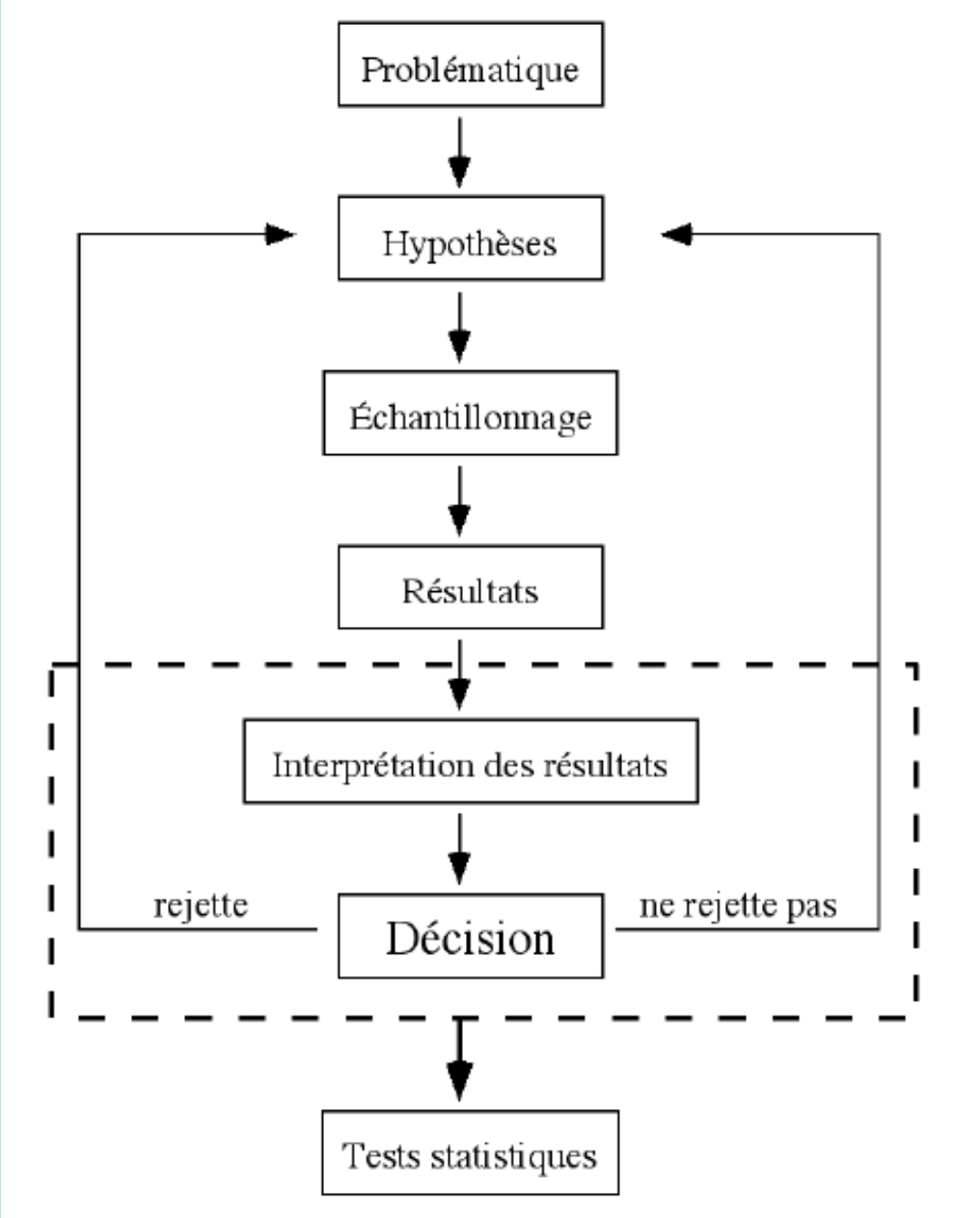
Introduction aux Statistiques

L1 STE



La démarche scientifique

But des statistiques

Permet de confirmer ou d'infirmer une hypothèse avec une marge d'erreur la plus petite possible et/ou prédire un événements à l'aide d'outils

Statistiques descriptives

Statistiques inférentielles

Statistiques descriptives

Méthodes statistiques utilisées pour construire des tables, des graphiques et des résumés numériques des données.

Statistiques inférentielles

Tirer une conclusion (inférence) ***objective*** à propos d'une population.

Basées sur l'information d'une population.

Population:

Ensemble des éléments qui forment le champ d'analyse d'une étude particulière. Attention à la connotation démographique!!

Taille notée : *N*

ex : Ensemble de toutes les voitures immatriculées en 21

Recensement:

Etude de tous les individus composant une population finie (pas toujours facile bien sûr).

Individu:

Elément composant la population.

ex: Un sol prélevé à Dijon, une voiture immatriculée en 21

Caractère:

Caractéristique propre à chacun des individus

ex : Teneur en Cd de ce sol, sa densité apparente..., couleur de la voiture, puissance.

Dans la plupart des cas, il est difficile d'obtenir l'information à partir de la **population** dans son ensemble. On utilise alors un **échantillon** pour tirer des conclusions sur la population.

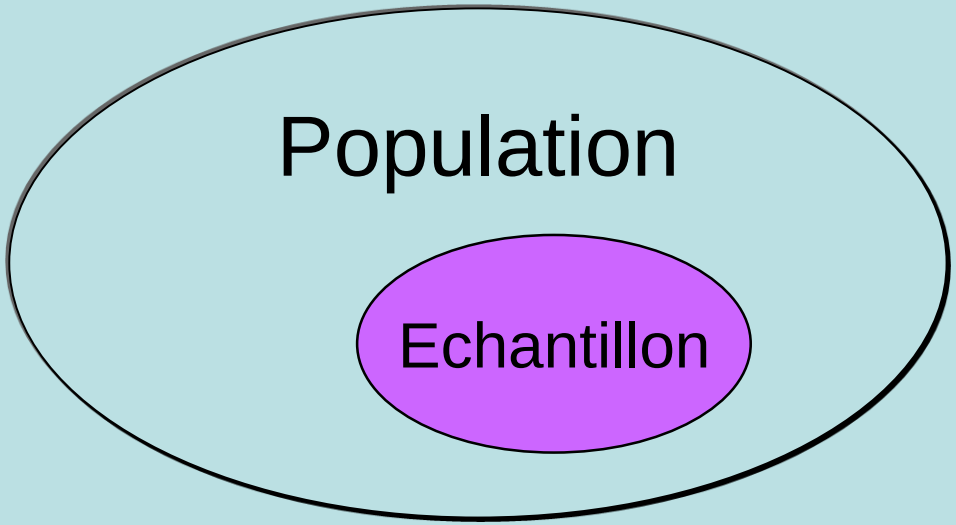
Echantillon (*sample*) :

Sous-groupe d'une population donnée.

Taille notée : *n*

ex : 20 sols viticoles prélevés autour de Beaune.

20 voitures passant devant la fac…



Caractéristiques qui doivent être prospectées lorsqu'on analyse des données:

- Type des variables
- Tables et méthodes graphiques
- Mesures numériques descriptives

Les éléments d'une population possèdent en commun le caractère d'être tous membres d'une population (!) mais ils varient selon d'autres critères…

Monnaies frappées sous l'Empire:

- Teneur en Ag
- Origine géographique
- Poids
- Usure
- Motif ….

CAPPADOCIA, Caesarea. *Tiberius, with Drusus Caesar.* AD 14-37. AR Drachm (3.57 g, 12h).

Le choix de la méthode statistique se fait suivant la **nature de la variable**.

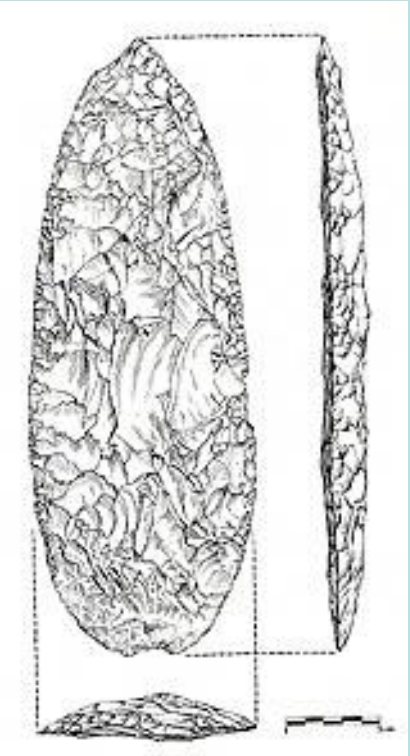
1. **Variables qualitatives**
   - Echelle nominale
   - Echelle ordinale

2. **Variables quantitatives**
   - Variables discrètes
   - Variables continues

Variable qualitative:

Modalité avec des mots ou des lettres (catégories).

ex :'homme', 'femme', de la variable ' sexe ', 'rouge', 'vert'… de la variable ' couleur '; 'non qualifié ', ' semi-qualifié ', ' qualifié ' de la variable ' qualification professionnelle ' type de silex…



Échelle nominale :

On dit d'une variable dont les catégories ne sont pas naturellement ordonnées, qu'elle est définie sur une échelle nominale.

ex : sexe, types de haches, types d'amphores...

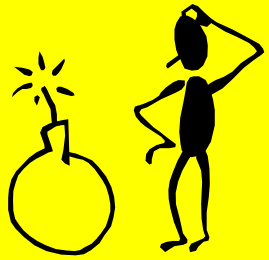
Three black-on-white mugs recovered from a single room at Pueblo Bonito, New Mexico. (Courtesy of the American Museum of Natural History, photograph by P. Hollembeak and J. Beckett.)

Échelle ordinale :

Si les catégories peuvent être ordonnées, on est en présence d'une échelle ordinale.

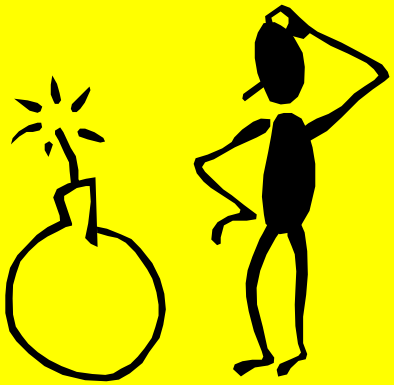
ex : qualification professionnelle (travail d'un potier) 'non qualifié', 'semi - qualifié', 'qualifié '

ATTENTION: Indique la position dans une série ordonnée mais pas l'importance de la différence. Pas de calculs algébriques!!



Une variable dichotomique est une variable qui ne comporte que 2 catégories.

'H' ou 'F', ' présence ' ou ' absence ', 'positif ' ou ' négatif ', ' marche ' ou ' arrêt ' ...



ATTENTION: On peut coder (0,1) des variables dichotomiques, cela ne signifie pas que les catégories ont un ordre logique. Ex. H/F!!

Variables quantitatives :
Modalités avec valeurs numériques.

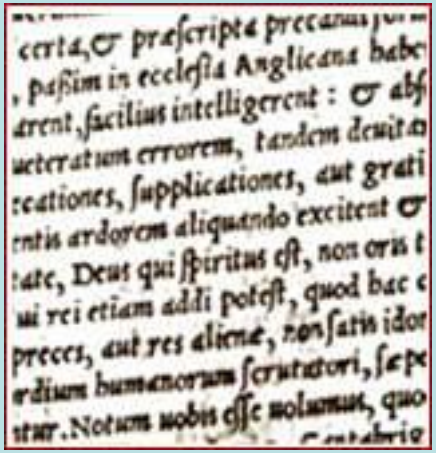
Ex: Teneur en Cd d'un sol, poids d'une pièce, nombre de sangliers sur une commune,…

Attention à l'unité!

Variables discrètes:

Une variable quantitative est dite **discrète** si l'étendue des valeurs possibles est dénombrable, c'est-à-dire si les valeurs peuvent être énumérées sous la forme d'une liste de chiffre (a_1, a_2,…) ou plus souvent d'entiers naturels (0, 1, 2, 3,…).

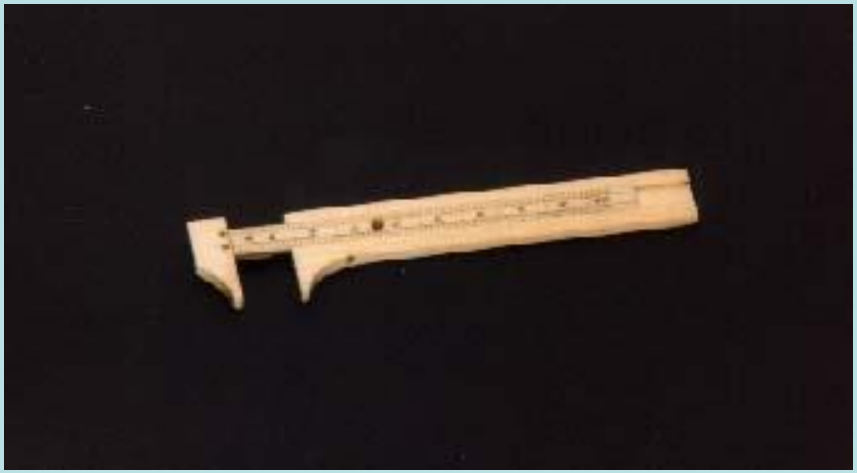
ex : Nombre d'objets dans un dépôt, nombre de mots dans une phrase, nombre de raisins sur une grappe, Nombre de mots dans un texte…


certa, & præscripta precan
, passim in ecclesia Anglicana habe
arent, facilius intelligerent : & abs
ueteratum errorem, tandem deuita
reationes, supplicationes, aut grati
ntis ardorem aliquando excitent &
tate, Deus qui spiritus est, non ori
ui rei etiam addi potest, quod hac
preces, aut res alienæ, non satis idon
rdium humanorum scrutatori, sæpe
tur. Notum nobis esse uolumus, quo

Variables continues:

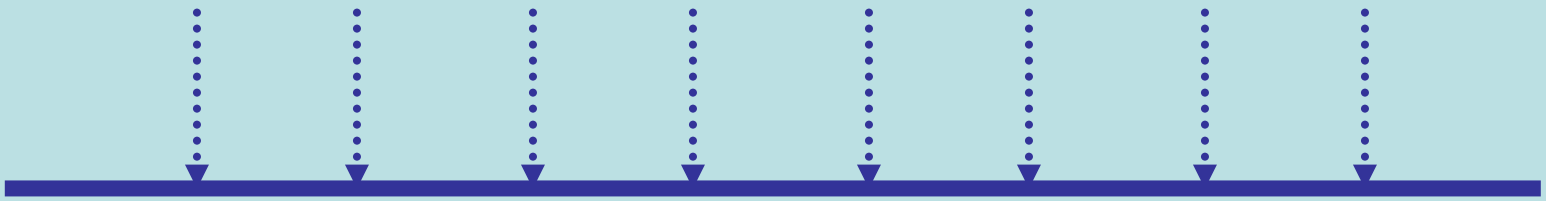
Une variable quantitative est dite **continue** si les valeurs possibles ne sont pas dénombrables.

Ex: poids d'un sanglier, concentration en Cd dans un sol,

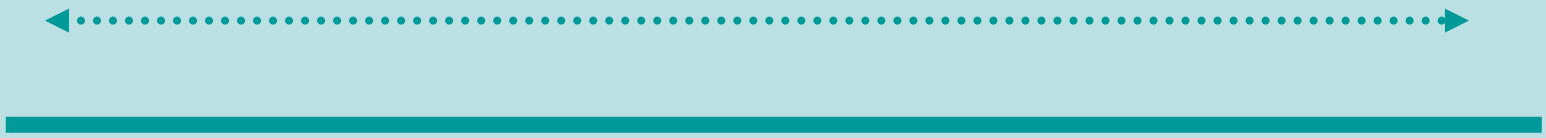


En fait, variable discrète du fait de la précision…

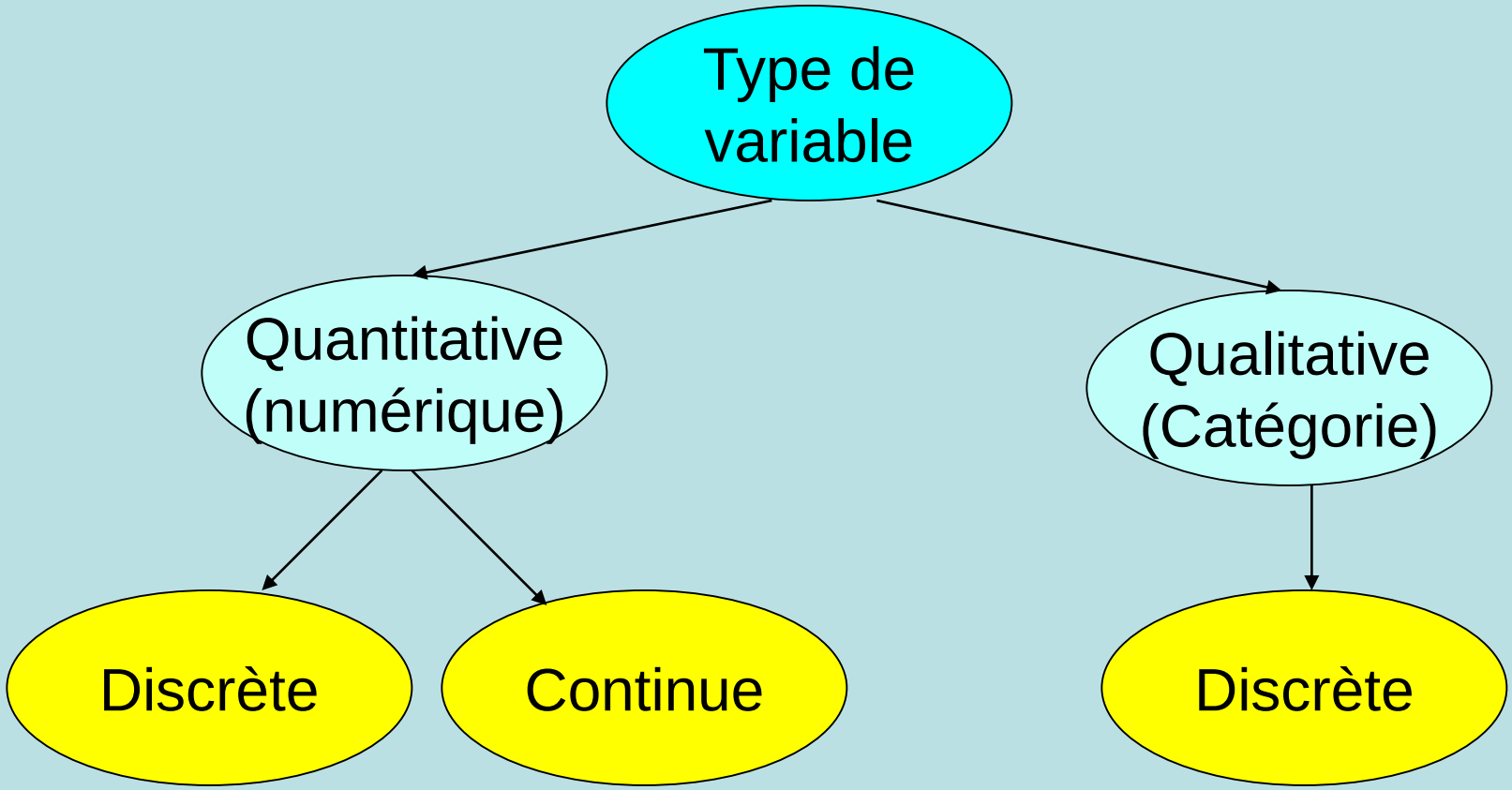
Variables discrètes -- Gaps entre les valeurs possibles



Variables continues -- *Théoriquement*,
pas de gap entre les valeurs possibles



Types de variables



Caractéristiques qui doivent être prospectées lorsqu'on analyse des données:

- Type des variables
- Tables et méthodes graphiques
- Mesures numériques descriptives
- Allures des distributions
- Détection des points éloignés (*ouliers*)

Distribution de fréquences (absolues ou relatives)

- Un simple moyen et efficace pour organiser et présenter les données tel qu'on peut avoir une image globale de l'endroit ou les mesures sont concentrées et dans quelle mesure elles sont dispersées.

- Convient aux données **qualitatives et quantitatives**.

Poids des fragments de poteries trouvés sur un site néolithique (g)

11.8	**3.6**	**16.6**	**13.5**	**4.8**	**8.3**
.	.	.	.	.	.
.	.	.	.	.	.
6.2	**11.2**	**10.4**	**7.2**	**5.5**	**14.5**

Distribution de fréquence

lower limit		upper limit	freq	rel freq	%freq
2	up to	5	3	0.10	10.00
5	up to	8	6	0.20	20.00
8	up to	11	8	0.27	26.67
11	up to	14	7	0.23	23.33
14	up to	17	4	0.13	13.33
17	up to	20	2	0.07	6.67

Distribution de fréquences cumulées

Convient aux données quantitatives seulement.

lower limit		upper limit	freq	cum freq	cum rel freq	cum % freq
2	up to	5	3	3	0.10	10.00
5	up to	8	6	9	0.30	30.00
8	up to	11	8	17	0.57	56.67
11	up to	14	7	24	0.80	80.00
14	up to	17	4	28	0.93	93.33
17	up to	20	2	30	1.00	100.00

Diagrammes en bâtons

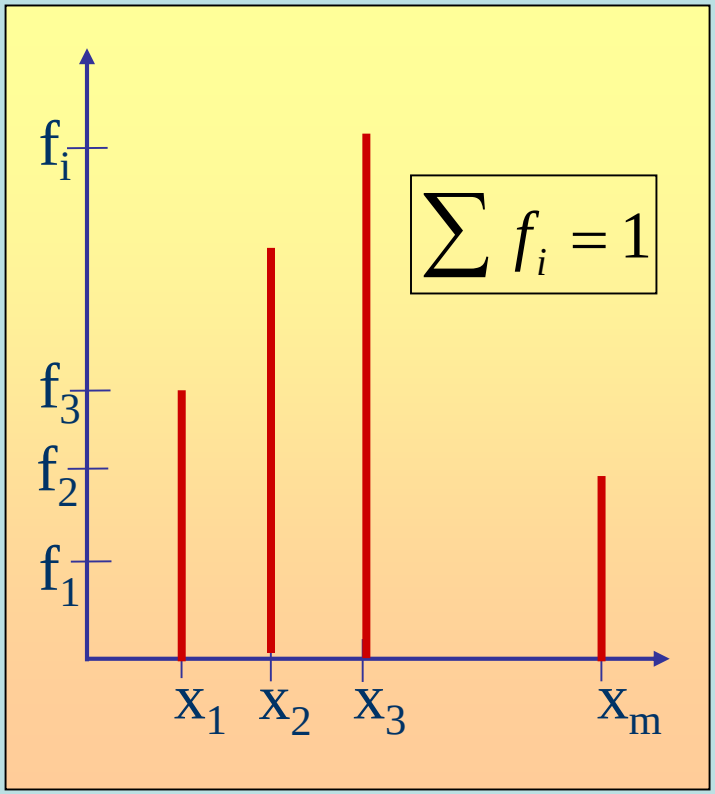
Diagrammes circulaires (*pie-chart*)

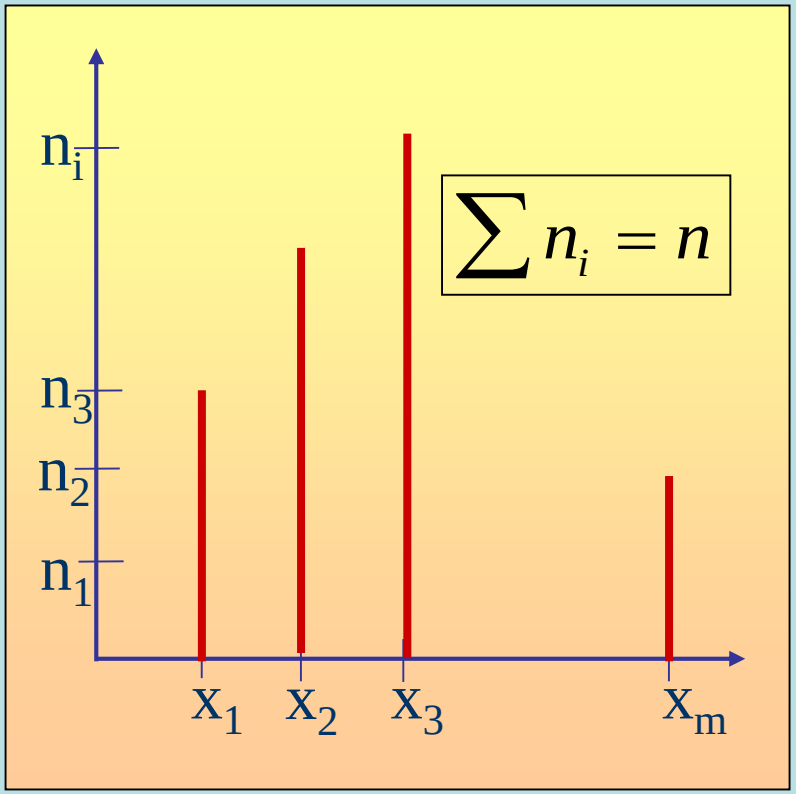
Histogrammes

Polygones de fréquences cumulées

Diagrammes en bâtons (*bar chart*)

Variables qualitatives sur une échelle nominale ou ordinale.





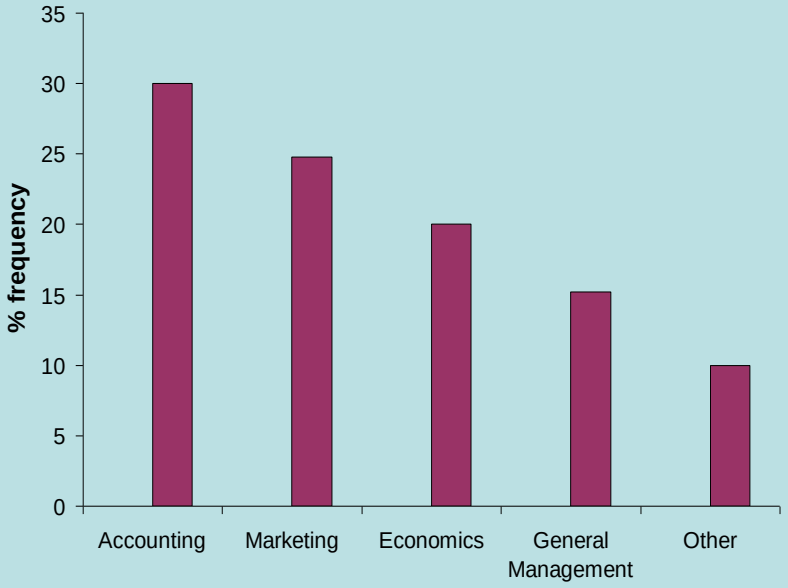
Diagrammes en bâtons (*bar chart*)

REMARQUES

•**Aucun ordre** n'est supposé

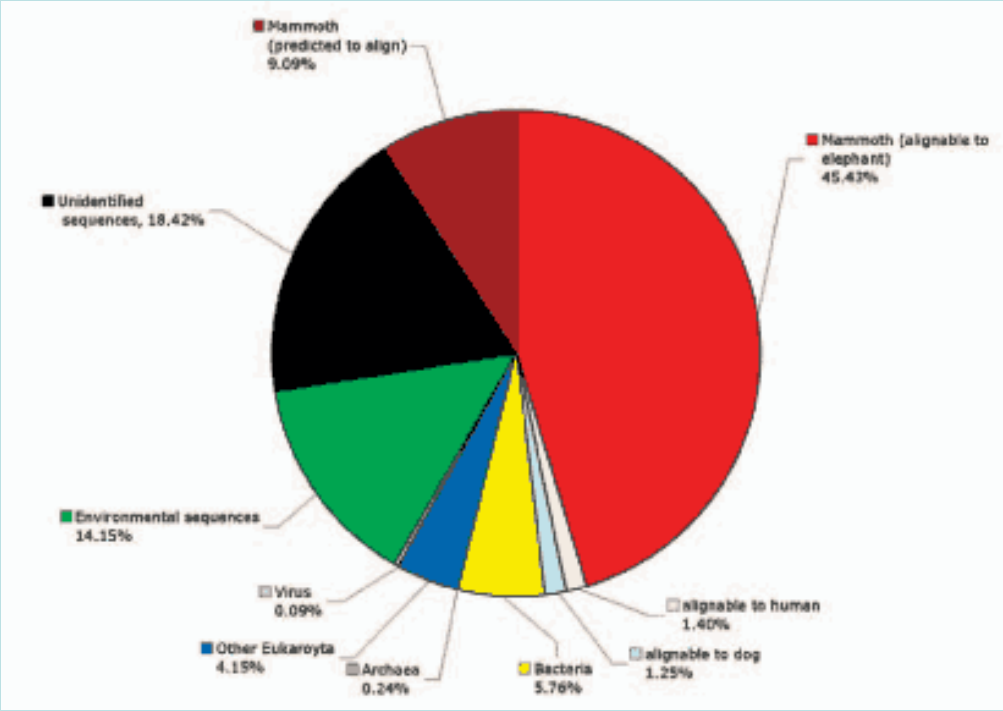
• Souvent les modalités ordonnées dans le sens des **fréquences croissantes ou par ordre alphabétique**

• Sur une échelle ordinale les données sont rangées suivant leur **ordre naturel**.



Diagrammes circulaires

Convient (surtout et éventuellement) aux données QUALITATIVES



Proportion of DNA sequence from different sources in the mammoth sample of Poinar et al. (2006).

Variables quantitatives discrètes

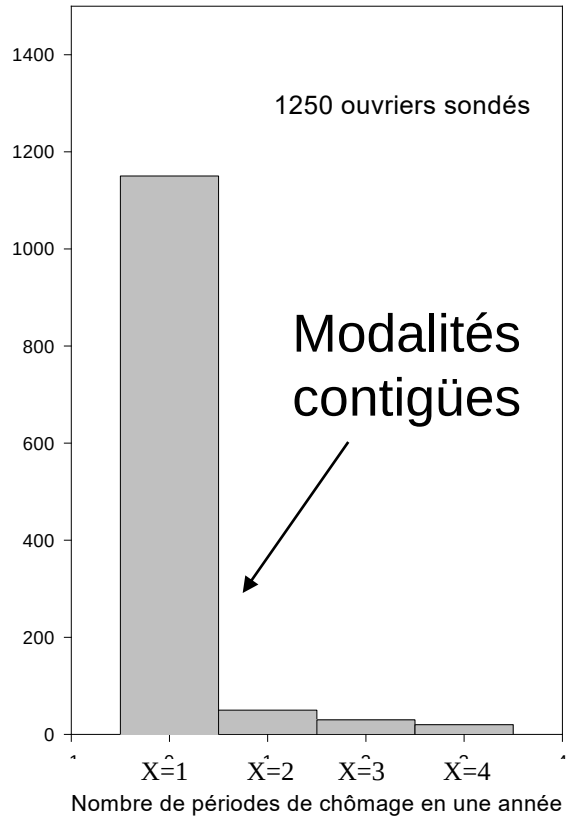
Modalités discontinues mais suivant un ordre naturel.

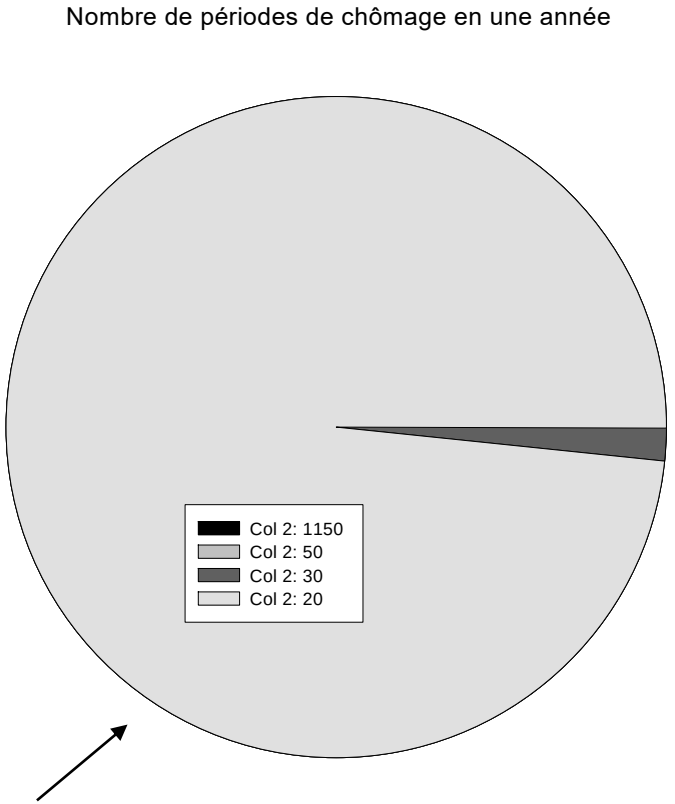
Même règles que pour des variables qualitatives d'échelle ordinale.

Tableau statistique, diagramme en bâtons, diagramme circulaire.

Dans le diagramme à bâtons, modalités successives contiguës.

Variables quantitatives discrètes



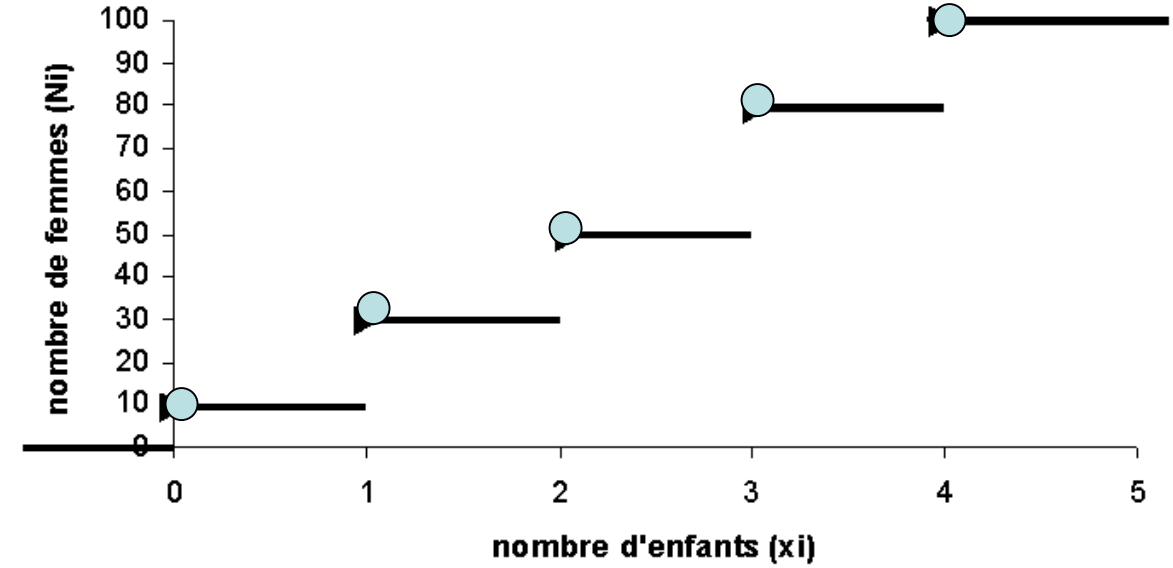


Possible mais peu révélateur car une modalité domine

Variables quantitatives discrètes

Courbes des fréquences cumulées. Il s'agit de courbes en escalier, c'est-à-dire constantes sur chaque intervalle défini par deux modalités successives,

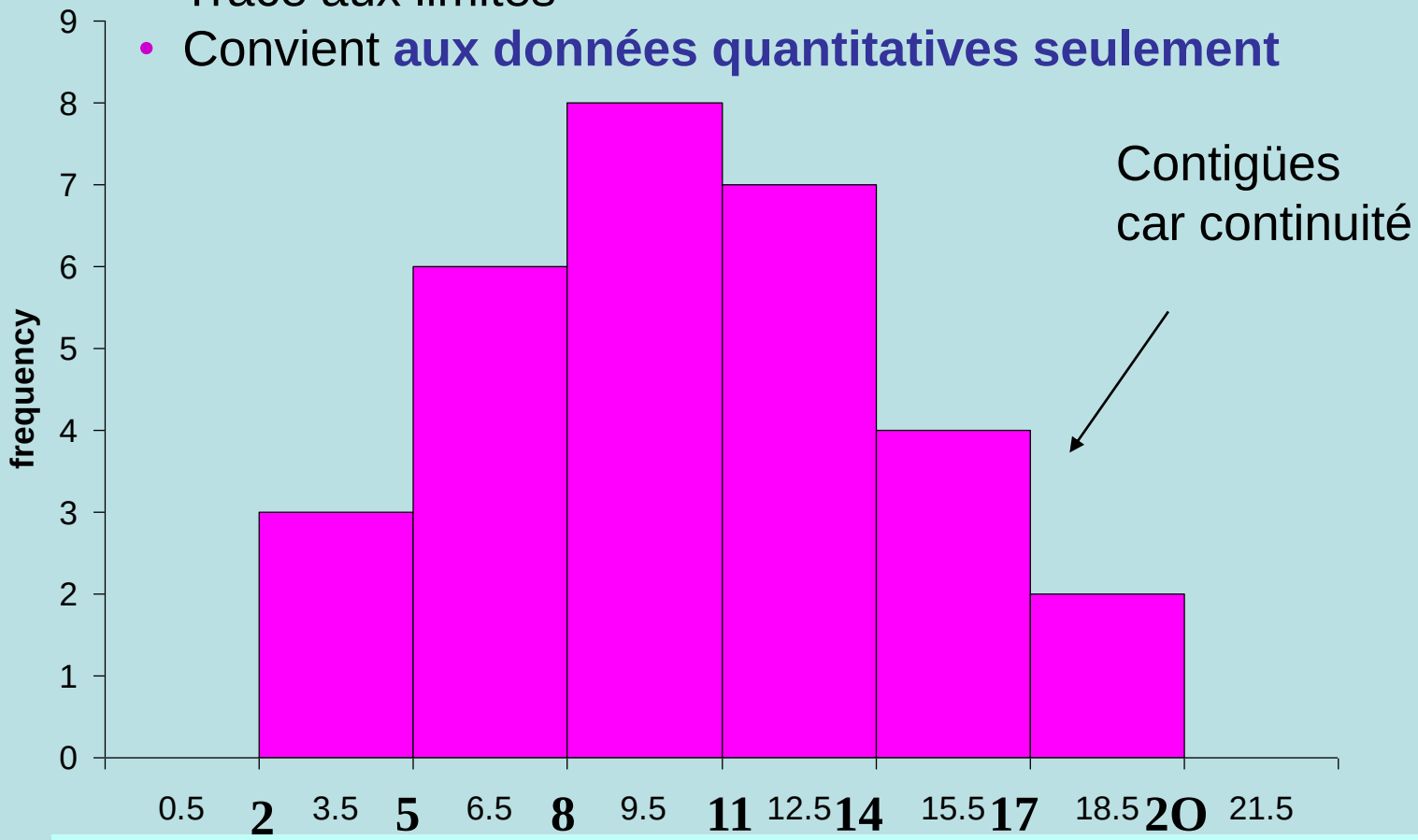
$$\text{Sur } [x_i, x_{i+1}[\text{ la fonction v aut } F_i^{\%} = \sum_{j=1}^{i} f_j^{\%}$$



Histogrammes

A ne pas confondre avec le diagramme baton!

- Tracé aux limites
- Convient **aux données quantitatives seulement**



Poids des fragments de poteries trouvés sur un site néolithique (g)

Histogrammes

Organisation par classe.

Soient ($[x_i,x_{i+1}[,f_i^{\%}$) et i de 0 à p-1, la distribution des fréquences.

On appellera **histogramme des fréquences** le diagramme formé des rectangles ($[x_i,x_{i+1}[$x[0,h_i]) où h est tel que **l'aire ainsi définie soit proportionnelle à $f_i^{\%}$**

Histogrammes

Dans la majorité des cas, une classe se rapporte à **plusieurs valeurs de la variable**. 15 g < œuf <= 16 g

Intervalle de classe : gamme des valeurs admissibles : de 15 g à 16 g, soit 1 g.

Indice de classe : valeur centrale de la classe. (15.5 g)

Perte d'information : répartition des valeurs à l'intérieur des classes.

Nombre de classe: combien??

Règle de Sturge:

$$\text{nombre de classes} = 1 + (3.3 \log_{10} n)$$

Règle de Yule :

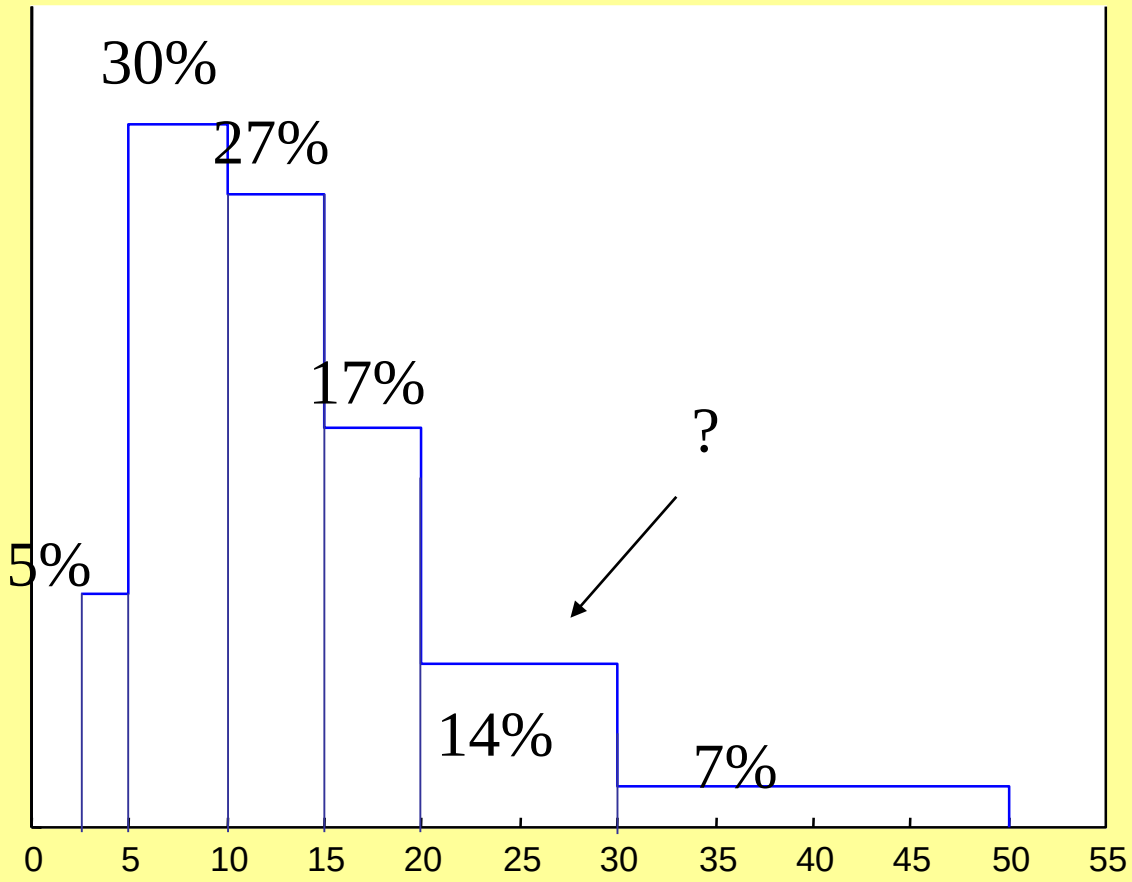
$$\text{nombre de classes} = 2.5\sqrt[4]{n}$$

Histogrammes

En divisant l'étendue de la variation par le nombre de classes
on obtient un ordre de grandeur de l'intervalle de classe.

$$\text{Int. de classe} = \frac{\text{Val max - val min}}{\text{Nbre de classes}}$$

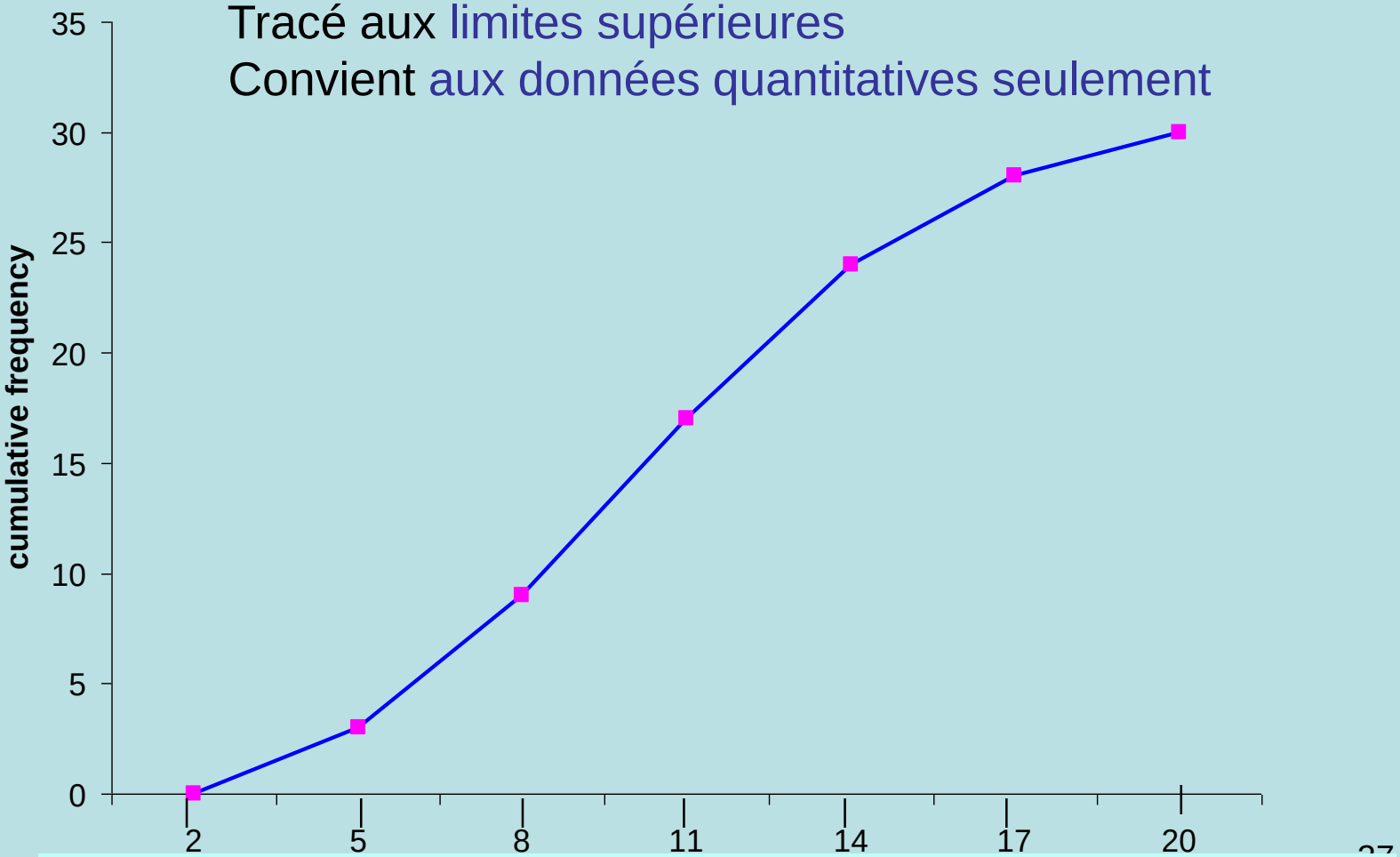
Histogrammes… pas toujours intuitifs!



Polygones de fréquences cumulées

Tracé aux limites supérieures

Convient aux données quantitatives seulement



Poids des fragments de poteries trouvés sur un site néolithique (g)

Caractéristiques qui doivent être prospectées lorsqu'on analyse des données:

- Type des variables
- Tables et méthodes graphiques
- Mesures numériques descriptives

Moyenne

convient aux données quantitatives.

Médiane

convient aux données quantitatives et aux données qualitatives sur une échelle ordinale.

Mode

convient aux données quantitatives et aux données qualitatives.

Par individus

Moyenne arithmétique:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

x barre

Moyenne arithmétique pondérée:

$$\bar{x} = \frac{\sum_{i=1}^{d} w_i x_i}{\sum_{i=1}^{d} w_i}$$

Population :

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$$

Moyenne dans le cas continu (données catégorisées, groupées)

On définit une subdivision de l'ensemble des valeurs donnant la distribution continue; soit ([x_i,x_{i+1}[,n_i) avec i de 0 à p-1 cette subdivision. Soit m_i le centre des classes,

$$m_i = \frac{x_i + x_{i+1}}{2}$$

On prend comme moyenne de x, la moyenne de la distribution discrète (m_i,n_i), avec i de 0 à p-1

La tendance centrale (moyenne)

Classes $[x_i,x_{i+1}[$	Centres	Effectifs	Effectifs pondérés
$[x_0,x_1[$	m_1	n_1	$n_1.m_1$
$[x_1,x_2[$	m_2	n_2	$n_2.m_2$
…	…		
$[x_p,x_{p-1}[$	m_p	n_p	$n_p.m_p$
		$\sum_{i=1}^{p} n_i$	$\sum_{i=1}^{p} n_i m_i$

$$\overline{x} = \frac{\sum_{i=1}^{p} n_i m_i}{\sum_{i=1}^{p} n_i}$$

En fait :

$$\overline{x} = \frac{\sum_{i=1}^{p} n_i \hat{m}_i}{\sum_{i=1}^{p} n_i}$$

m chapeau c'est une estimation!

Médiane

Les données sont classées par ordre de magnitude.

Valeur pour laquelle la fréquence cumulée est égale à 0.50 ou point qui partage la distribution en 2 parties égales.

$$med = x_{\left(\frac{n+1}{2}\right)}$$

Pour *n* impair

$$med = \frac{x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}}{2}$$

Pour *n* pair

$$med = L + \frac{n/2 - \sum n_i(\text{inf})}{n_i(med)} \cdot c$$

L: limite inférieure de la classe médiane
n: nombre total d'observations
$\sum n_i(\text{inf})$: somme des fréquences absolues des classes se situant avant la classe médiane.
ni(med) : fréquence de la classe médiane
c: largeur de la classe médiane

Exercice

Médiane :propriétés

Souvent utilisée pour les données démographiques.

Particulièrement adaptée pour décrire la tendance centrale des **échelles ordinales** et des distributions **très étalées** pour lesquelles la moyenne pondère exagérément les valeurs extrêmes.

La médiane est plus **conservatrice**. Donne **l'individu type**.

Se prête mal aux calculs algébriques

Le mode (mod) d'une variable qualitative (ou quantitative discrète) est la valeur qui possède la fréquence la plus élevée.

Quelques propriétés…

Le mode n'est pas toujours la valeur centrale de la distribution.

Une distribution peut avoir **un ou plusieurs modes**.

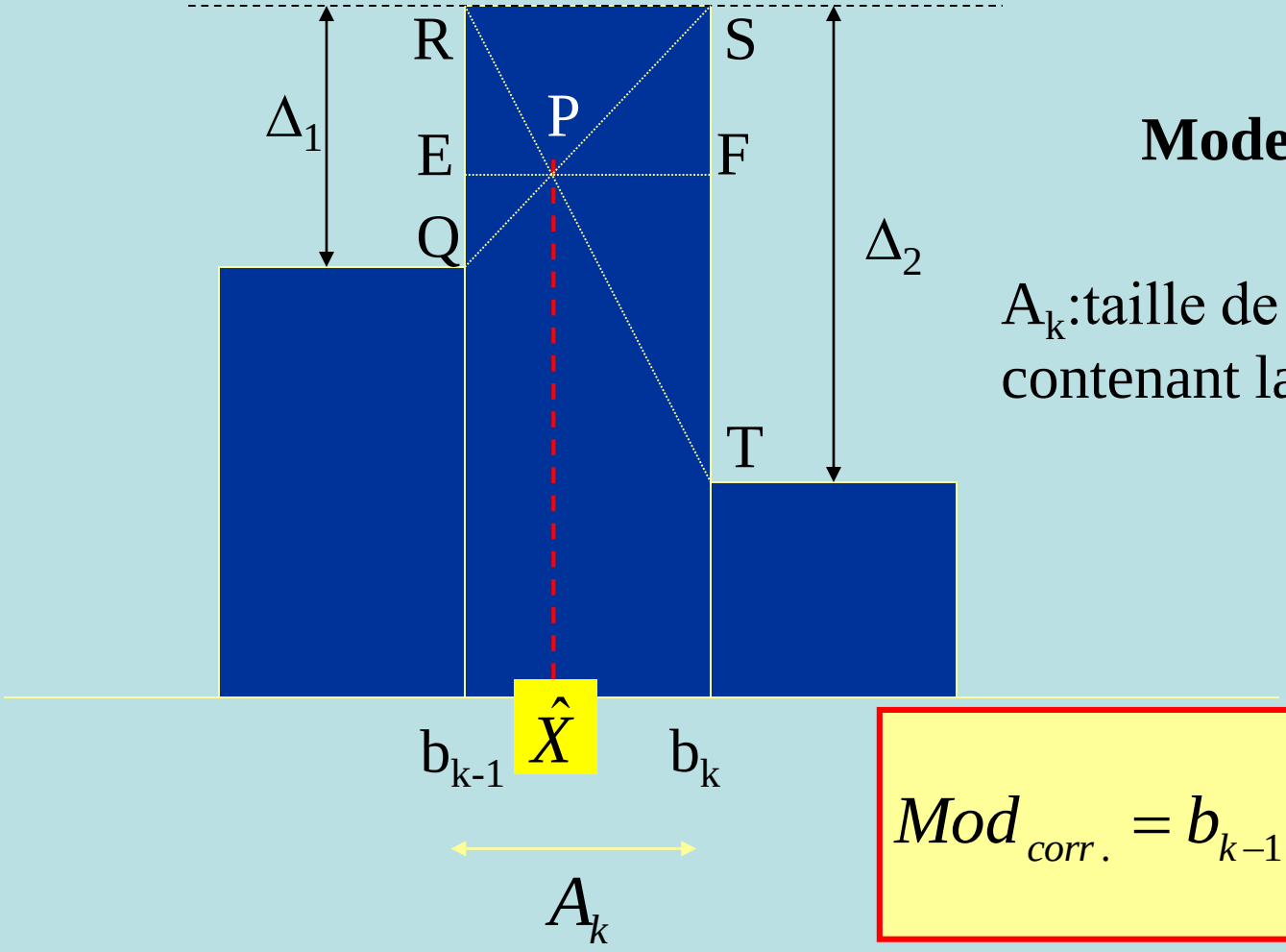
N'est pas affecté par les **valeurs exceptionnelles**.

Bon indicateur des **populations hétérogènes** qui présentent une ou plusieurs valeurs dominantes

Se prête mal aux calculs stat. et algébriques

Attention, varie si l'on modifie l'intervalle de classe.

La tendance centrale (le mode)

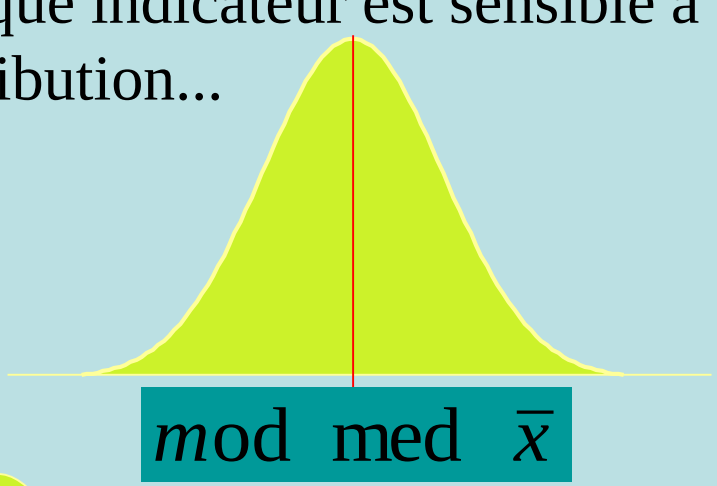


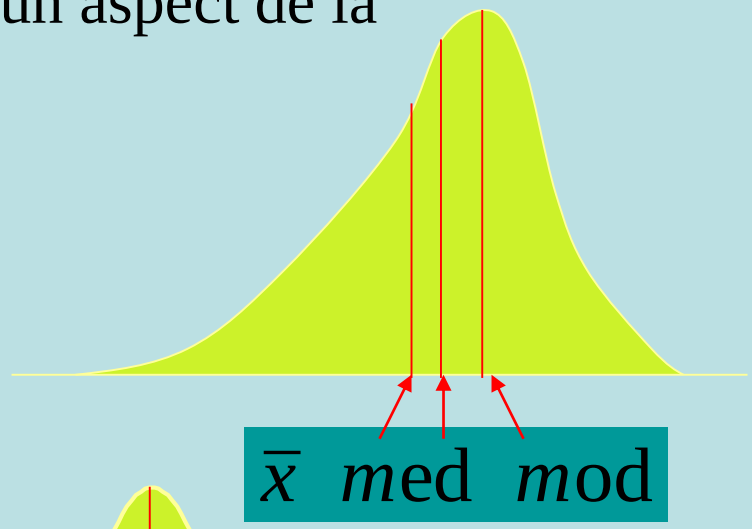
Mode corrigé

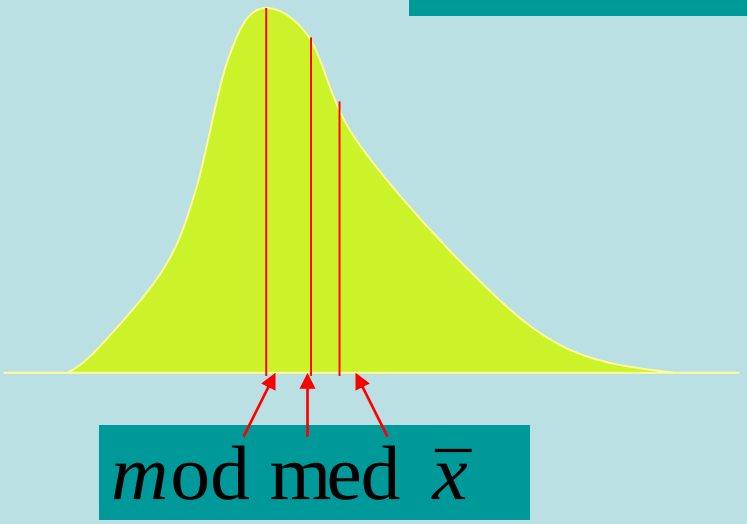
A_k:taille de l'intervalle contenant la classes modale

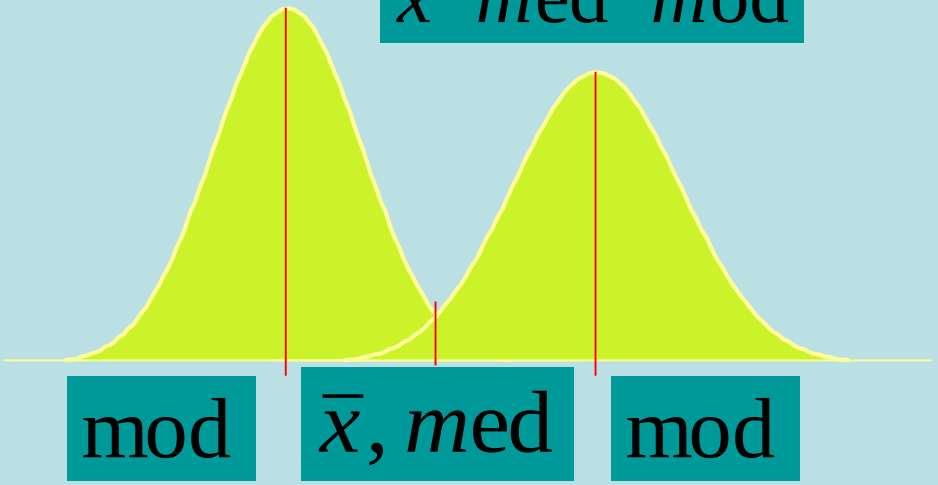
$$Mod_{corr.} = b_{k-1} + \left(\frac{\Delta_1}{\Delta_1 + \Delta_2} \right) A_k$$

Chaque indicateur est sensible à un aspect de la distribution...





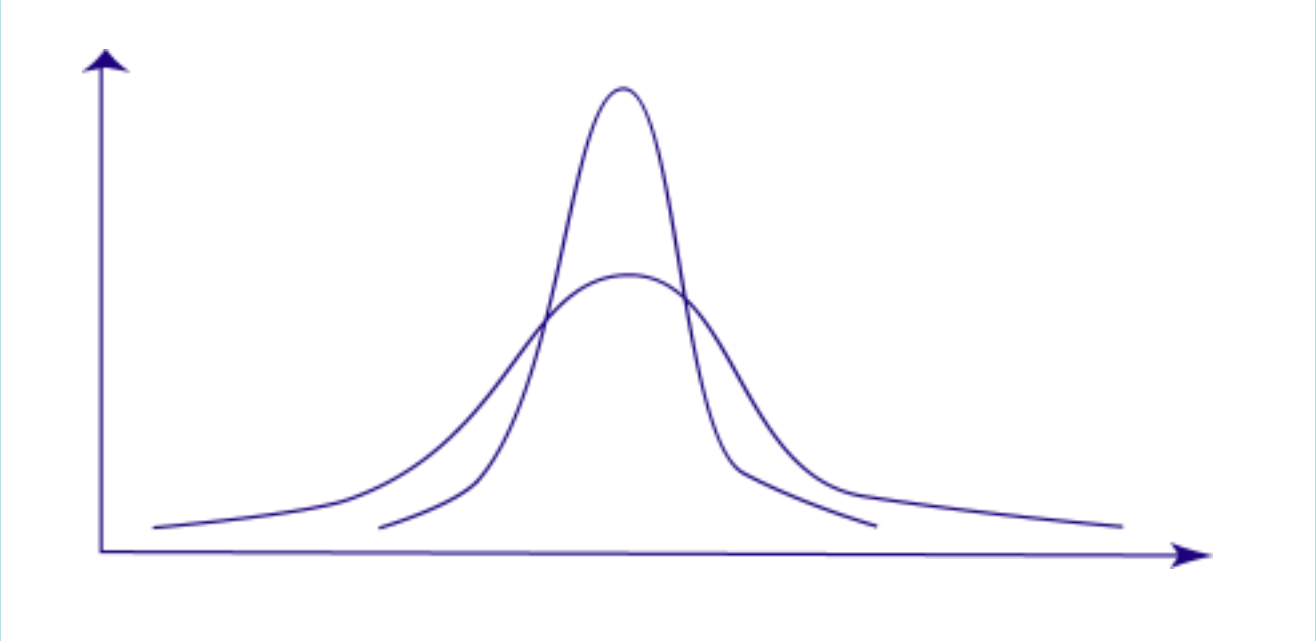




Caractéristiques qui doivent être prospectées lorsqu'on analyse des données:

- Type des variables
- Tables et méthodes graphiques
- Mesures numériques descriptives
- Allures des distributions
- Détection des points éloignés (*ouliers*)

Deux distributions de fréquence peuvent avoir la même moyenne, la même médiane et le même mode et présenter des formes très différentes:



Etendue de la variation (*range*) ou *empan* ou *marge de variation*

C'est la différence entre la plus grande valeur et la plus petite valeur de la variable.

Etendue = maximum - minimum

Exemple

Valeur maximum x = 174 mm

Valeur minimum x = 140 mm

Etendue de la variation = 174 - 140 = 34 mm

Mesure de la dispersion (écarts moyen, écart médian)

Ecart moyen :

$$E.M. = \frac{\sum_{i=1}^{n} \left| x_i - \bar{x} \right|}{n}$$

Ecart médian :

$$E.med. = \frac{\sum_{i=1}^{n} \left| x_i - med \right|}{n}$$

Au niveau de la **population statistique**, la variance est la moyenne arithmétique des carrés des écarts des valeurs à leur moyenne:

Moyenne : $$\mu = \frac{\sum x}{N}$$

Variance de la population: $$\sigma^2 = \frac{\sum_{i=1}^{n} (x_i - \mu)^2}{N}$$

Dans la cas d'un **échantillonnage aléatoire**, la meilleure estimation de la variance de la population est:

$$s_x^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$$

Estimateur sans biais

Les formules précédentes se rapportent à des données brutes. Pour une **distribution de fréquence**, il faut employer:

$$s_x^{\,2} = \frac{\sum_{i=1}^{k} f_i \left(x_i - \bar{x}\right)^2}{n-1}$$

k : nombre de classes
f_i : la fréquence de la classe i
x_i : la valeur centrale de la classe i

Propriétés:

• La variance est toujours > ou = à 0

• La variance est égale à 0 si toutes les valeurs sont identiques

• En ajoutant une constante aux données, la variance ne change pas.

• En multipliant par une constante, on modifie la variance par un facteur multiplicatif égal au carré de la constante d ’origine

• Si $Y=aX+b$, $s^2(Y)=a^2.s^2(X)$ et $s(Y)=a.s(X)$

Mesure de la dispersion (la variance)

xi	fi	xi-moy	(xi-moy)^2	fi*(xi-moy)^2
142.5	1	-16.7	278.89	278.89
147.5	1	-11.7	136.89	136.89
152.5	9	-6.7	44.89	404.01
157.5	17	-1.7	2.89	49.13
162.5	16	3.3	10.89	174.24
167.5	3	8.3	68.89	206.67
172.5	3	13.3	176.89	530.67

Moyenne =	n		somme	1780.5
159.2	50			
s2	**36.3367347**			

$$s_x^2 = \frac{1780.5}{49} = 36.33 mm^2$$

L’écart type d’une distribution est égale à la racine de la variance

$$\sigma = \sqrt{\sigma^2}$$

population

$$s_x = \sqrt{s_x^2}$$

échantillon

Même unité que la moyenne!!

Mesure de la dispersion (le coefficient de variation))

Écart type de 3 m n'a pas la même signification si l'on se rapporte à 50 m ou 1000 m!

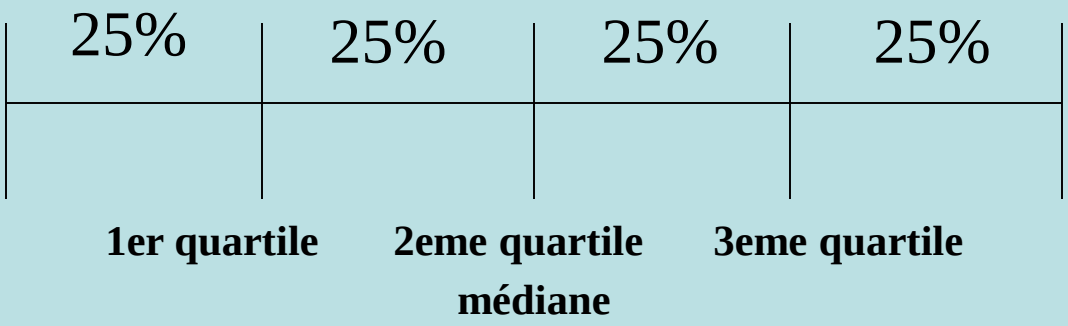
$$C.V. = \frac{100 s_x}{\bar{x}}$$

échantillon

L'intervalle interquartile est une mesure de dispersion correspondant à l'intervalle comprenant 50% des observations les plus au centre de la distribution.

Quantiles:

- Quartiles : 4 parties égales
- Déciles : 10 parties égales
- Centiles : 100 parties égales



Organiser les n observations en distribution de fréquence

Quartiles = observations pour lesquelles la fréquence relative cumulée dépasse respectivement 25%, 50% et 75%

Autre méthode: Calcul du j^e quartile

Soit i la partie entière de $j.(n+1)/4$ et k la partie fractionnelle de $j.(n+1)/4$. Soit $x_{(i)}$ et $x_{(i+1)}$ les valeurs des observations classées respectivement en i^e et $(i+1)^e$ position (après classement par ordre croissant.

Le j^e quartile est égale à:

$$Q_j = x_{(i)} + (k.(x_{(i+1)}-x_{(i)})$$

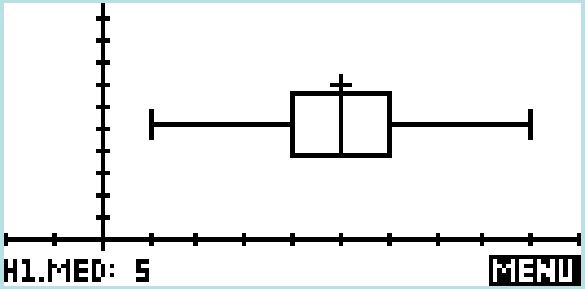
Exemple: 1 2 4 4 5 5 5 6 7 9

Q_1 à la position (n+1)/4 = 2.75 soit entre 2e et 3e observation

$Q_1 = x_{(2)} + 0.75\,(x_{(3)} - x_{(2)}) =$ **3.5**

$Q_2 = x_{(5)} + 0.5\,(x_{(6)} - x_{(5)}) =$ **5**

$Q_3 = x_{(8)} + 0.25\,(x_{(9)} - x_{(8)}) =$ **6.25**



50% data dans cet intervalle

Intervalle interquartile:

$IQ = Q_3 - Q_1 = 6.25 - 3.5 =$ **2.75**

Groupement en classes (variable continue):
1er quartile : classe pour laquelle la freq. Rel. Cum. > 25%
2eme quartile : classe pour laquelle la freq. Rel. Cum. > 50%
•3eme quartile : classe pour laquelle la freq. Rel. Cum. > 75%

$$Q = L + \left[\frac{(nq) - \sum n_i(\inf)}{n_i(quartile)} \right] \cdot c$$

L: borne inf de la classe du quartile
n: nombre total d'observations
q :1/4 pour 1er quartile, 1/2 pour Q_2, 3/4 pour Q_3
$\sum n_i(\inf)$: Somme des freq abs. des classes se situant avant la classe du quartile.
n_i(quartile): fréquence absolue de la classe du quartile.
c: largeur de la classe du quartile.

Mesure de la dispersion (intervalle interquartile)

Conc	Freq abs	Freq abs cum	Freq rel cum
100-200	10	10	0.1
200-300	20	30	0.3
300-400	40	70	0.7
400-500	30	100	1
Total	100		

Classe 1er quartile : 200-300

$$Q_1 = 200 + \left[\frac{(100 \cdot 1/4) - 10}{20}\right] \cdot 100 = 275$$

$$Q_2 = 350$$

$$Q_3 = 416.66$$

$$IQ = 416.66 - 275 = 141.66$$