Introduction aux statistiques pour les archéologues

Josef Wilczek

Masarykova univerzita, Brno
L'univesité de Bourgogne, Dijon
josef.wilczek@hotmail.com

Planning

- Quelques définitions
- Types des variables
- Statistiques descriptives
  - Tables des fréquences
  - Méthodes graphiques
  - Les tendances centrales
  - Les tendances de la variabilité (dispersion)
  - Distributions

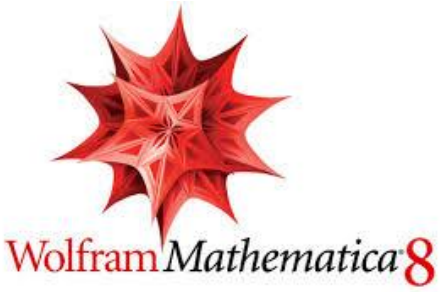
Statistique

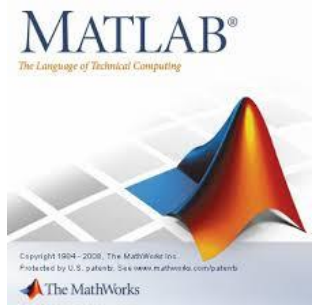
- La science qui a pour but de développer des connaissance sur le monde qui nous entours, en utilisant des données empiriques.
- Basé sur la mathématique et la théorie de probabilité

- Pourquoi ?
  - Pour la description des données, pour chercher des tendances, structurations, …
  - Pour tester les hypothèses
  - Pour la prédiction des valeurs

Software

- MS Excel
- StatSoft Statica
- IBM SPSS ver 12
- Mathworks MATLAB
- Wolfram Mathematica 8
- R – Project for Statistical Computing
- PAST

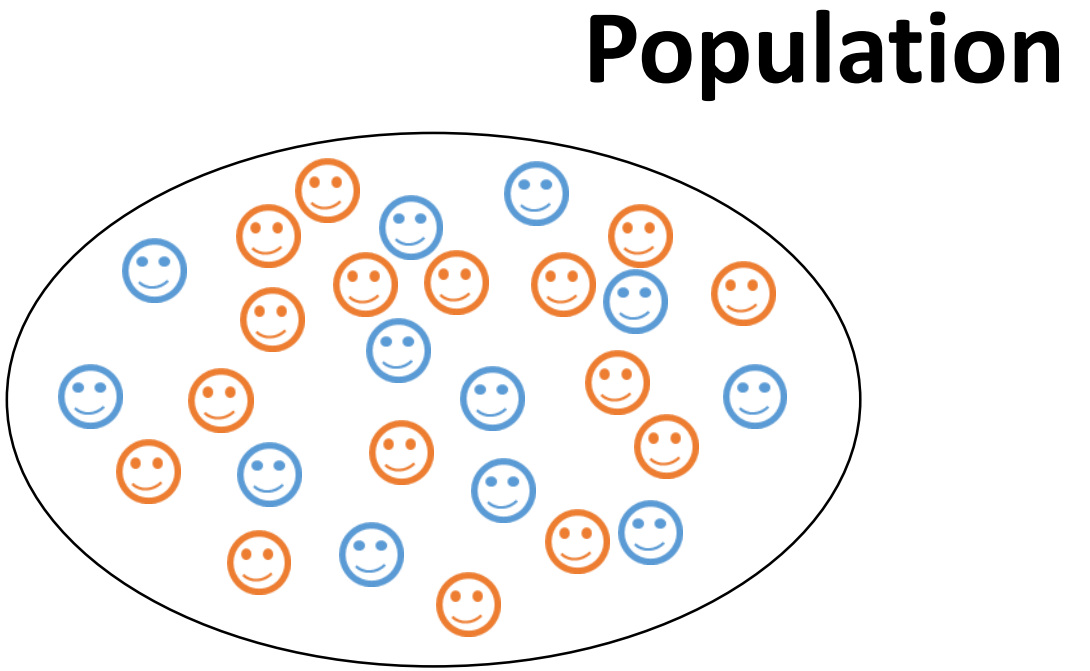




Quelques définitions

Quelques définitions

- **Population**
  - Ensemble des éléments qui forment le champ d'analyse d'une étude particulière
  - La taille est notée : N
  - Attention de pas mélanger avec la population démographique ! (la population peut être aussi bien un objet)

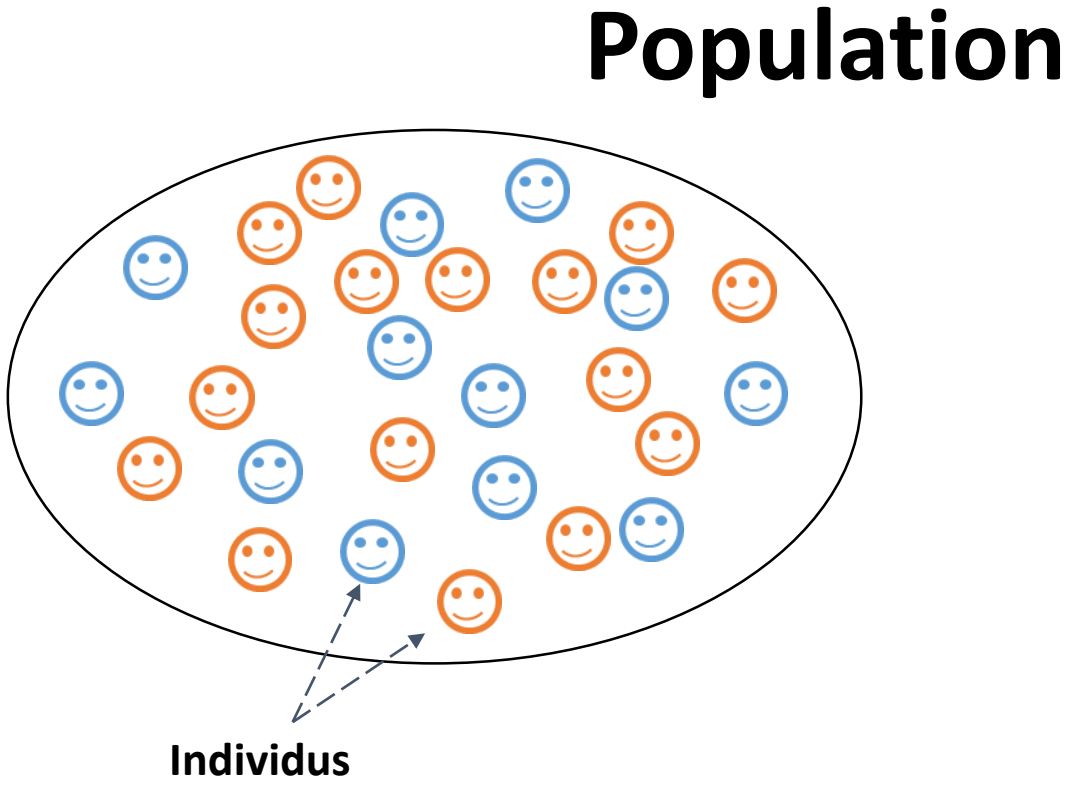


- *Les étudiants du labo ($N = 153$)*
- *Tous les vases néolithiques de Dijon ($N = 206505$)*

Quelques définitions

- **Individu**
  - = entité / objet statistique
  - Elément composant la population

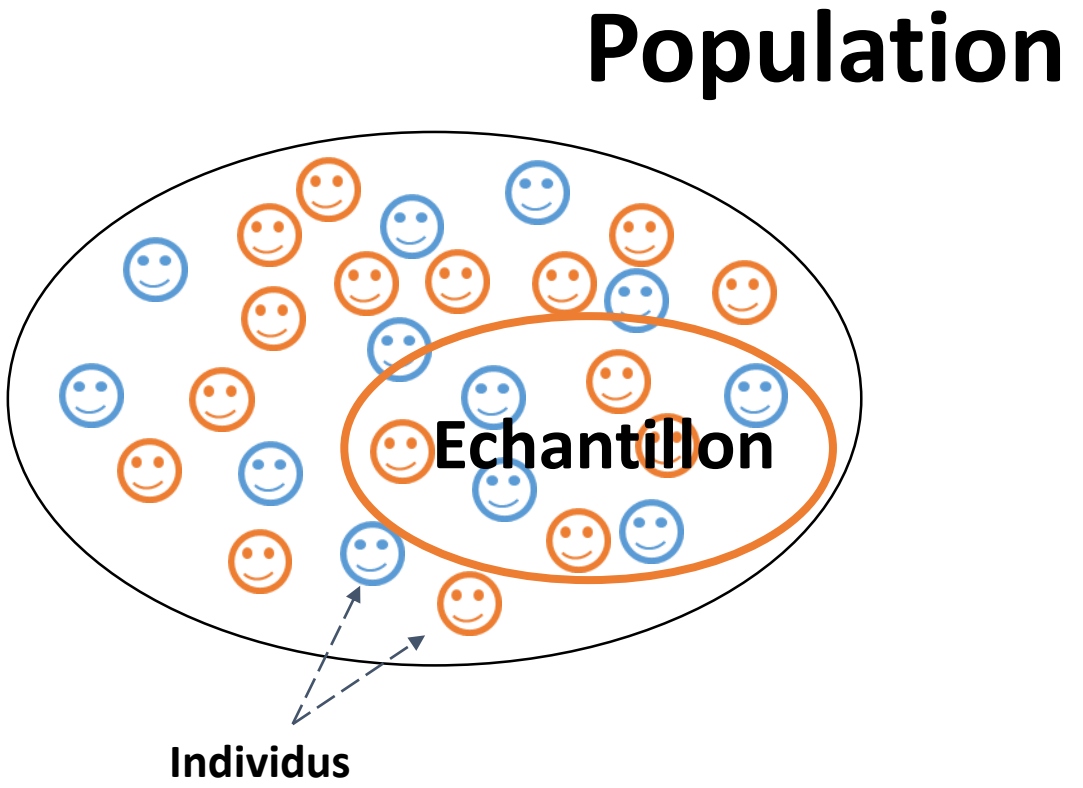
- *Etudiant du labo*
- *Une vase de la tombe*
- <u>Un individu statistique ne signifie pas « un pièce »</u>
  - *Dépôt des haches*
  - *Tous les vases d'une cimetière*



Quelques définitions

- **Echantillon**
  - Souvent (jamais) on ne peut pas étudier toute la population => on fait donc une échantillonnage
  - Sous-groupe d'une population donnée
  - Noté : n

- *20 étudiants choisi de façon aléatoire*
- *20 bols*

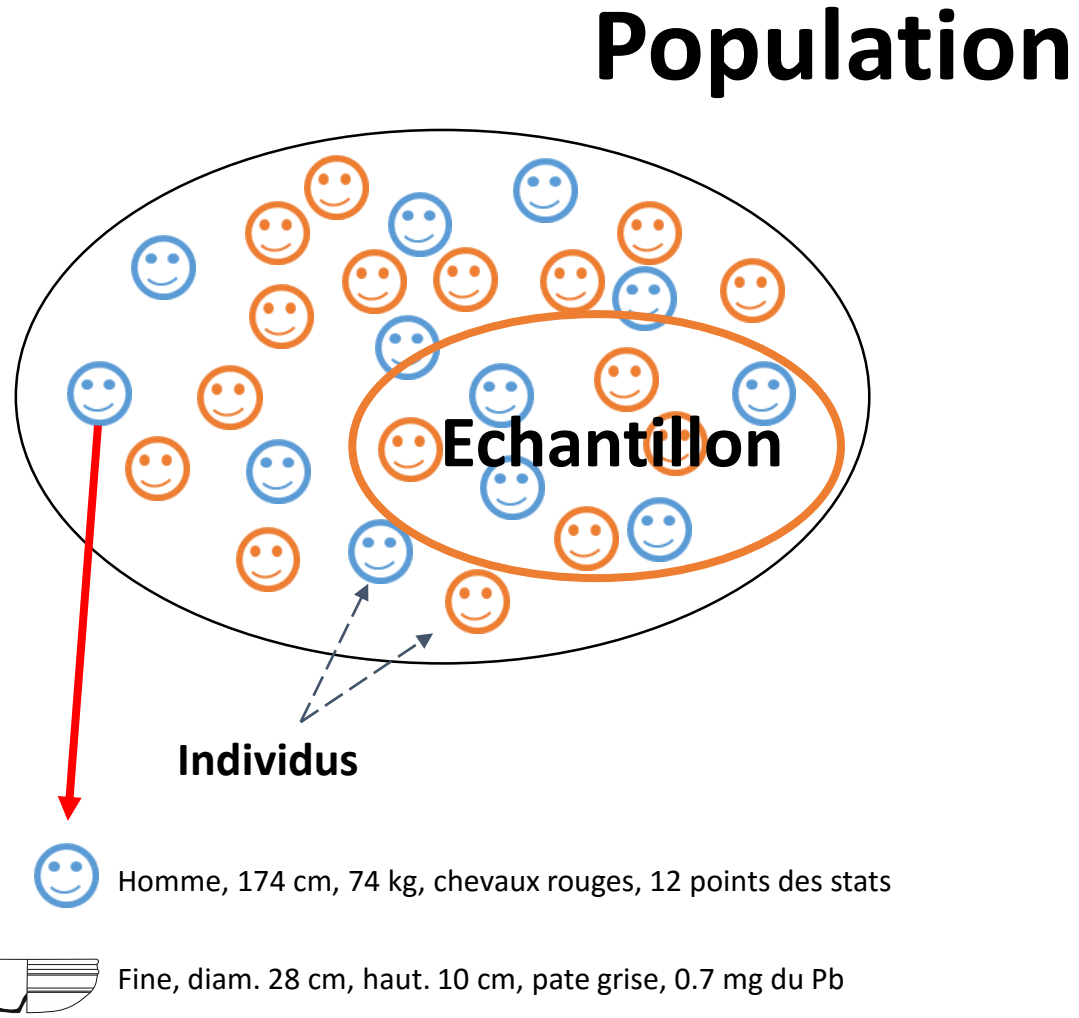


Quelques définitions

- **Caractère**
  - Caractéristique propre à chacun des individus

- *Un étudiant*
  - *gendre, sa taille, hauteur, largeur, poids, …, couleur des chevaux, des yeux, la note du test, …*

- *Une vase*
  - *hauteur, largeur, poids, …, couleur des la pate, teneur en Pb dans la pate, décoration, …*



Quelques définitions

Etudiants

Variables / Descripteurs

Individus

Individu	Sexe	Taille	Poids	Couleur du chevaux	La note du test
Etudiant 1	Homme	174	76	Gris	3
Etudiant 2	Homme	168	84	Blond	2
Etudiant 3	Femme	158	57	Rouge	3
…	…	…	…	…	…
Etudiant 152	Homme	179	80	Vert	1
Etudiant 153	Femme	173	76	Brun	2

Vases ($N = 294$)

Individu	Pate	Diam	Hauteur	Couleur de pate	Pb (%)
Vase 1	Fine	28	10	Brun	0.7
Vase 2	Fine	18	40	Jeune	12.1
Vase 3	Graphitique	17	23	Gris	0.8
…	…	…	…	…	…
Vase 293	Grossier	40	18	Noir	10.2
Vase 294	Jemná	28	20	Noir	9.8

Homme, 174 cm, 74 kg, chevaux rouges, 12 points des stats

Fine, diam. 28 cm, haut. 10 cm, pate grise, 0.7 mg du Pb

Types des variables

Variables qualitatives (mots)

- Nominales
- Dichotomiques
- Ordinales

Variables quantitatives (nombres)

- Discrètes
- Continues

- Le choix de la méthode statistique dépend du type de la variable.

Variables qualitatives (mots)

- Descripteurs qui expriment des qualités qu'on peut observer sur les individus
- Les « valeurs » s'appelent **modalités** *(gendre – 2 modalités)*
- Les modalités se excluent *(je suis un homme, je ne suis pas une femme)*

- **Nom**inale
  - Variable dont les catégories ne sont pas naturellement ordonnées
  - Règle : *« soit ca, soit ca ou ca »*
  - *Ex. couleur des yeux (vert/bleu/brun)*

- **Di**chotomique
  - A seulement deux catégories
  - Règle : *« soit ca ou ca »*
  - *Ex. Les cotés d'un monnaie (pile, face), sexe*

- **Ord**inale
  - Les modalités peuvent être ordonnées (de plus petit au plus grand)
  - Règle d'échelle
  - *Ex. Qualité d'un produit / d'un artisan (bon, mieux, meilleur), satisfaction (pas de tout, moyennement, très satisfé)*

- *étudiants*
  - *genre (homme, femme)*
  - *couleur des chevaux (blonde, brune, noir, rouge)*
  - *spécialisation (bronze, Hallstatt, La Tène, …)*
  - *qualification (pas qualifié, semi-qualifié, qualifié)*
  - *évaluation des élevés (excellent, bon, moyen, pas bon, désastre)*

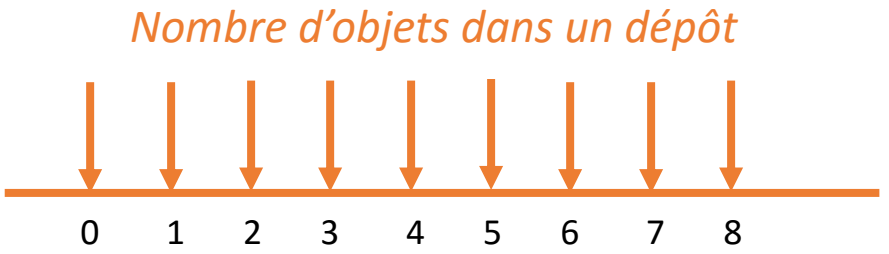
- *céramique*
  - *le type fonctionnel (bol, pot, couvercle, bouteille, …)*
  - *présence d'un pied (oui, non)*
  - *profilation du bord (évasé, étroit, rentrent)*

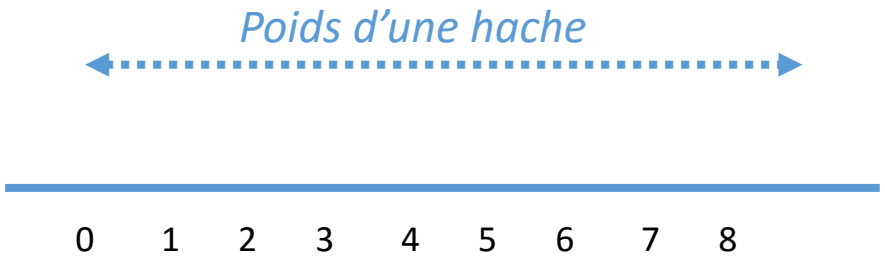
Variables quantitatives

- Modalités avec valeurs numériques

- Discrètes
  - Les valeur sont comptables, on peut les diviser par un nombre entier
  - Règle : On comte « des pièces »
  - Les valeurs sont souvent exprimés par les nombres entiers (0, 1, 2, 3, ….)
  - *Ex: nombre des objets dans une sépulture, des sites d'une époque…*

- Continues
  - Les valeurs possibles ne sont pas dénombrables
  - Les valeurs peuvent être mesuré de façon continu
  - Théoriquement, il n'y a pas de gap entre les valeurs possibles
  - *Ex. poids, hauteur, largeur, longueur, concentration en Cd dans un sol, …*





Variables quantitatives

- Modalités avec valeurs numériques

- Discrètes
  - Les valeur sont comptables, on peut les diviser par un nombre entier
  - Règle : On comte « des pièces »
  - Les valeurs sont souvent exprimés par les nombres entiers (0, 1, 2, 3, ….)
  - *Ex: nombre des objets dans une sépulture, des sites d'une époque…*

- Continues
  - Les valeurs possibles ne sont pas dénombrables
  - Les valeurs peuvent être mesuré de façon continu
  - Théoriquement, il n'y a pas de gap entre les valeurs possibles
  - *Ex. poids, hauteur, largeur, longueur, concentration en Cd dans un sol, …*

- *Quantité des pépins dans un orange*
- *Nombre des haches dans un dépôt*
- *Poids d'un orange*
- *Poids des haches dans une tombe*
- *Nombre des mots dans une phrase*
- *Quantité de Pb dans une hache*

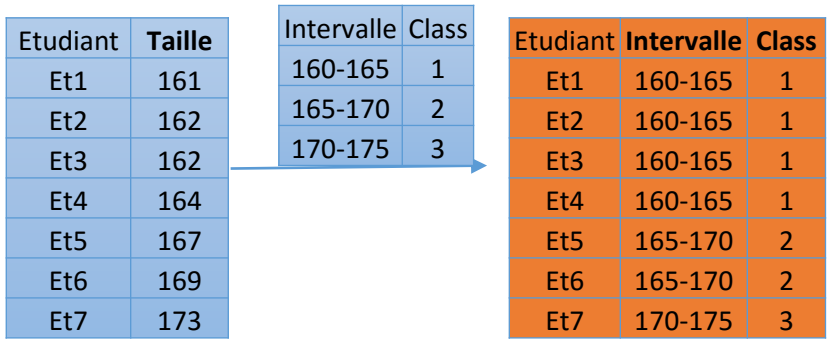
Variables quantitatives

- Modalités avec valeurs numériques

- Discrètes
  - Les valeur sont comptables, on peut les diviser par un nombre entier
  - Règle : On comte « des pièces »
  - Les valeurs sont souvent exprimés par les nombres entiers (0, 1, 2, 3, ….)
  - *Ex: nombre des objets dans une sépulture, des sites d'une époque…*

- Continues
  - Les valeurs possibles ne sont pas dénombrables
  - Les valeurs peuvent être mesuré de façon continu
  - Théoriquement, il n'y a pas de gap entre les valeurs possibles
  - *Ex. poids, hauteur, largeur, longueur, concentration en Cd dans un sol, …*

- Les variables continues peuvent être « discrétisées » - ils peuvent être regroupées dans les clases (intervalles)
  - Ex. La taille des étudiants qui sont entre 160 et 175 peut être exprimé par trois intervalles (de 5 cm).



Etudiant	**Taille**
Et1	161
Et2	162
Et3	162
Et4	164
Et5	167
Et6	169
Et7	173

Intervalle	Class
160-165	1
165-170	2
170-175	3

Etudiant	**Intervalle**	**Class**
Et1	160-165	1
Et2	160-165	1
Et3	160-165	1
Et4	160-165	1
Et5	165-170	2
Et6	165-170	2
Et7	170-175	3

Variable continue

Variable discrète

Types des variables

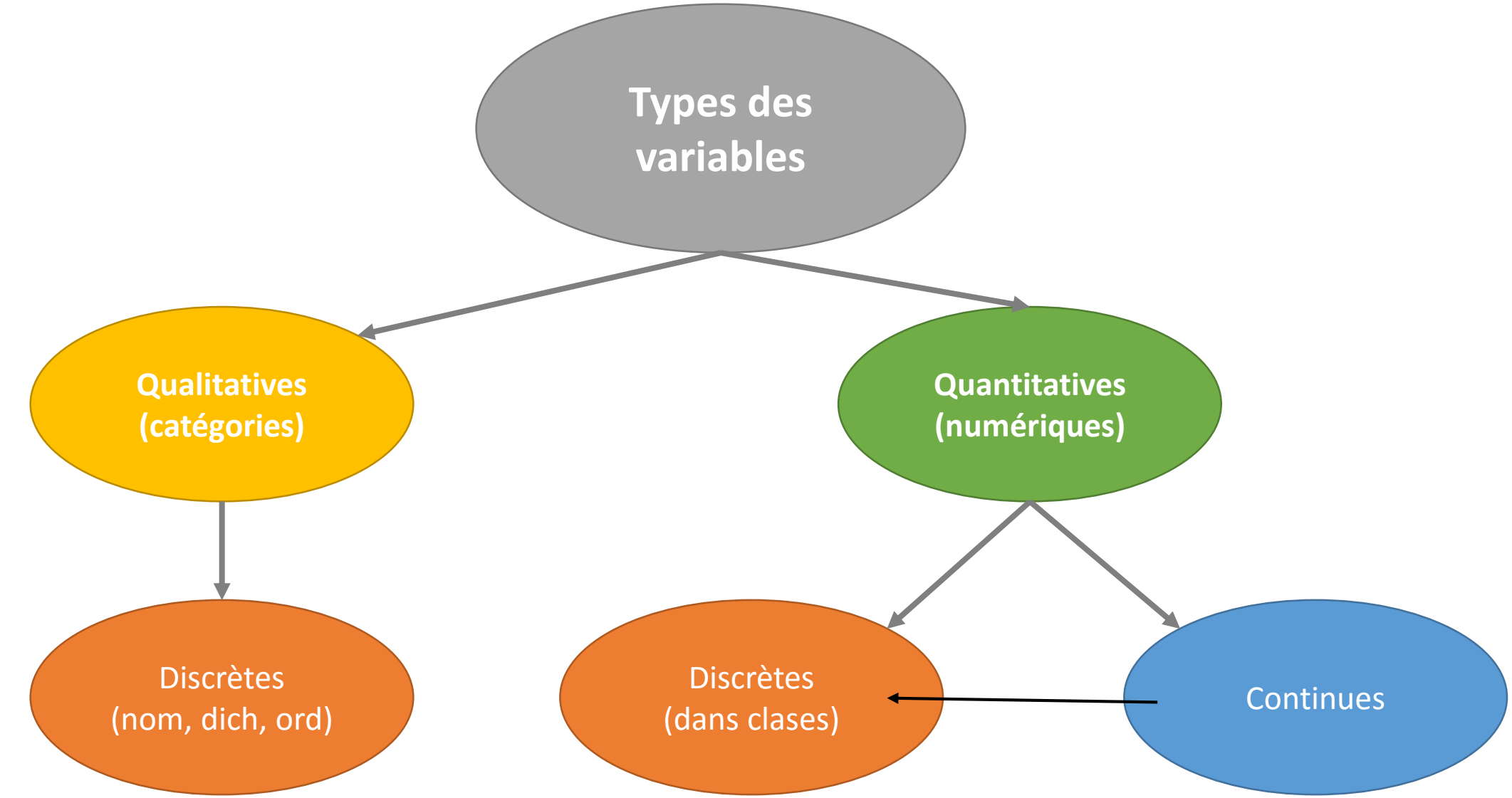
Etudiants ($N = 153$)

Individu	Sexe	Taille	Poids	Couleur du chevaux	La note du test
Etudiant 1	Homme	174	76	Gris	3
Etudiant 2	Homme	168	84	Blond	2
Etudiant 3	Femme	158	57	Rouge	3
…	…	…	…	…	…
Etudiant 152	Homme	179	80	Vert	1
Etudiant 153	Femme	173	76	Brun	2

Vases ($N = 294$)

Individu	Pate	Diam	Hauteur	Couleur de pate	Pb (%)
Vase 1	Fine	28	10	Brun	0.7
Vase 2	Fine	18	40	Jeune	12.1
Vase 3	Graphitique	17	23	Gris	0.8
…	…	…	…	…	…
Vase 293	Grossier	40	18	Noir	10.2
Vase 294	Jemná	28	20	Noir	9.8

Quantitative ou qualitative ?



Mots qu'on a dans les classes.

Les valeurs numériques dans les clases.

Les nombres continues.

Statistiques

1) Statistiques descriptives
2) Statistiques inférentielles

1) Statistiques descriptives

- Pourquoi ?
  - Pour présenter notre données
  - Pour décrire les caractéristiques des bases (toujours avant des analyse des donnes)
- Comment ?
  - Par tables
  - Par des graphiques
  - Par des mesures numériques descriptives (moyens, écarts, médians, ….)
  - Par la caractérisation des distributions

- Avantages
  - Pas trop complique
  - Au premier coup d'oeil on peut voir
    - les caractéristiques des donnés et tendances
    - et parfois quelque chose intéressante…

- Désavantages
  - des résultats (les graphiques, moyens, écarts, etc.) sont des descriptions et ne permettent pas de dire quelque chose avec une certitude
  - *Ex : « Les longueurs des tombes féminins sont plus longues de celles des hommes. "*
  - Mais est-ce que ca peut être juste l'effet du hasard ?

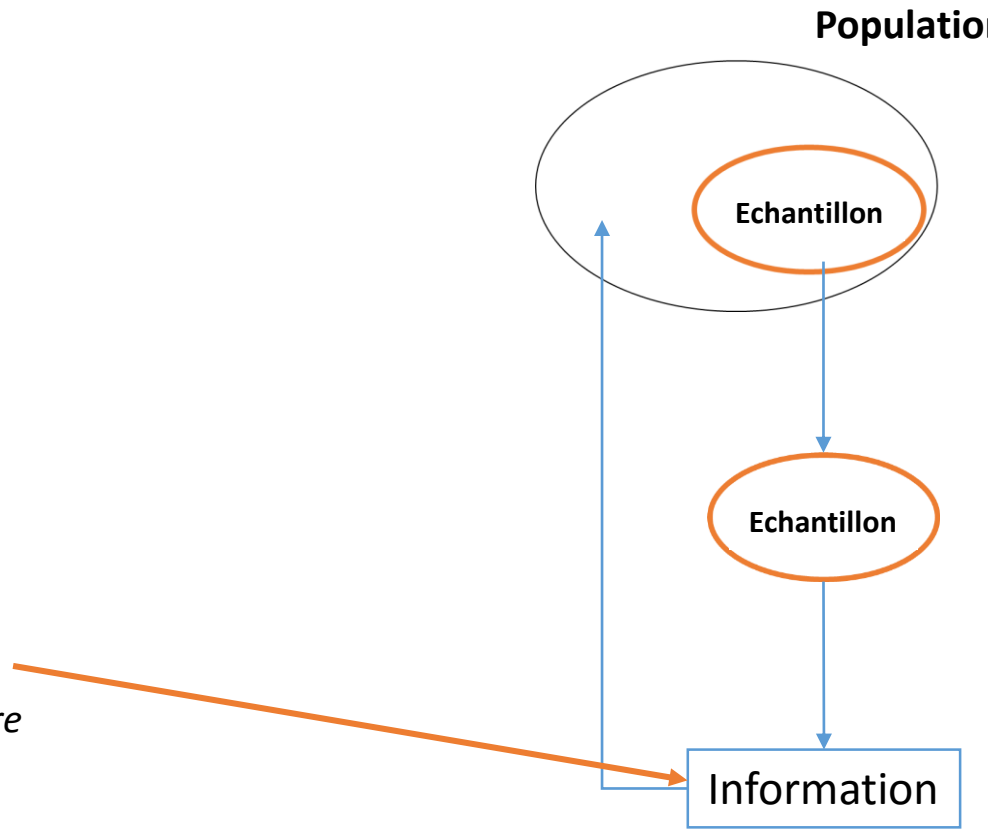
2) Statistiques inférentielles

- = statistique inductive
- D'après un échantillon choisi de la population on trouve quelques informations, qui sont ensuite généralisées sur la population

- Donc,
  - d'une population, on choisit un échantillon
  - de cet échantillon on obtient quelque informations
  - qu'on généralise pour la population

- Il s'agit surtout du « test » des hypothèses :
  - *H1 : Il y a des différences entre la taille de tombes entre les hommes et femmes.*
  - *H1 : Il y a une relation entre longueur et largeur des tombes.*

  - Est-ce que ces différences/relations sont significatifs?
  - On peut conclure la différence/relation avec certain niveau de la certitude ?



Statistiques descriptives

Statistiques descriptives

- But
  - décrire de la meilleure manière les données qu'on veut analyser
  - le premier pas avant l'analyse des données

- Comment ?
  1) Tables des fréquences
  2) Méthodes graphiques
  3) Les tendances centrales
  4) Les tendances de la variabilité (dispersion)
  5) Distributions

Statistiques descriptives

1) **Tables des fréquences**
2) Méthodes graphiques
3) Les tendances centrales
4) Les tendances de la variabilité (dispersion)
5) Distributions

Table de fréquences (absolues ou relatives)

- Utilisé pour des variables quantitatives et qualitatives
- On peut obtenir un images du fait
  - où les mesures se concentrent
  - et combien ils sont dispersées

- **La fréquence absolue**
  - Exprime combien des individués avec certain modalité/classe on a dans l'échantillon (n_i)
  - Pour les variables qualitatives
  - Pour les variables quantitative discrètes
  - Pas pour les variables quantitative continuelles (!) – un peu nonsense…

- **La fréquence relative**
  - $f_i = \frac{n_i}{N}$
    - n_i - la fréquence absolue du i-th modalité
    - N – somme des individus
  - Ratio entre la fréquence absolue d'un modalité/classe et la somme des individués.

Gendre	Fréquence absolue	Fréquence relative	%
Homme	73	0,48	48
Femme	80	0,52	52
Somme	153		

Classes des tailles	Fréquence absolue	Fréquence relative	%
160-165	22	0.14	14.38
165-170	48	0.31	31.37
170-175	15	0.10	9.80
175-180	34	0.22	22.22
180-185	30	0.20	19.61
185-190	4	0.03	2.61
	153		

Distribution des fréquences (absolues ou relatives) cumulées

- Seulement pour les donnés qualitatives
- C'est la même chose que les tables des fréquences, sauf que les valeurs consécutives se cumulent

Taille	Fréquence absolue	Fréquence absolue cumulée	Fréquence relative	Fréquence relative cumulée
160-165	22	22	0.14	0.14
165-170	48	70	0.31	0.45
170-175	15	85	0.10	0.55
175-180	34	119	0.22	0.77
180-185	30	149	0.20	0.97
185-190	4	153	0.03	1.00

Pourquoi ? 55% des individus ont la taille jusque 175 cm.

Statistiques descriptives

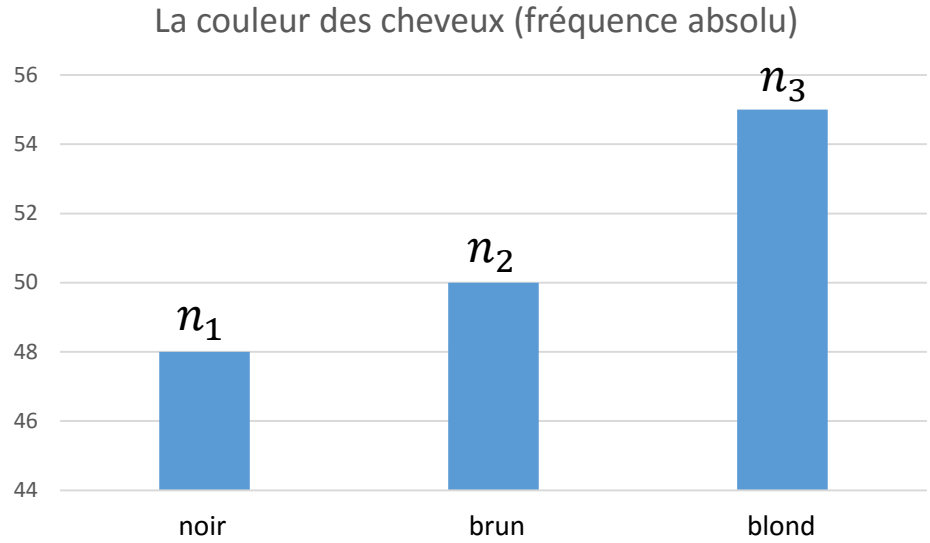
1) Tables des fréquences
2) **Méthodes graphiques**
3) Les tendances centrales
4) Les tendances de la variabilité (dispersion)
5) Distributions

Variables qualitatives discrètes

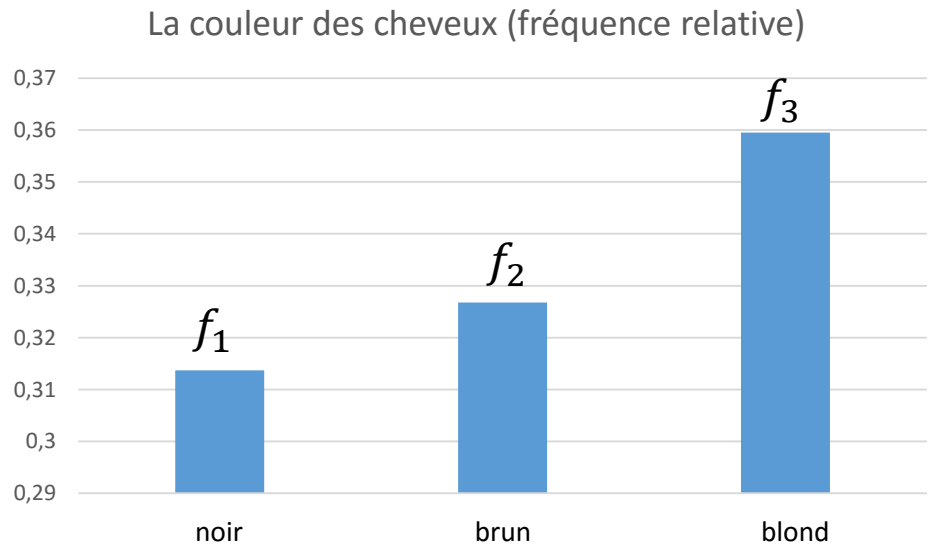
Diagramme en bâtons (bar chart/bar plot)

- Les bâtons correspondent aux modalités
- La taille d'un bâton exprime la fréquence absolue/relative d'une modalité

- **Que est-ce qu'il faut savoir ?**
- Les bâtons ne se touchent pas, car les variables ne sont pas ordinales (on ne peut pas les ordonner de plus petits aux plus grandes)
- C'est a nous de choisir l'ordre des bâtons dans le graphique
- Néanmoins on essaye de les ordonner
  - De plus petits aux plus grandes
  - Par l'ordre alphabétique



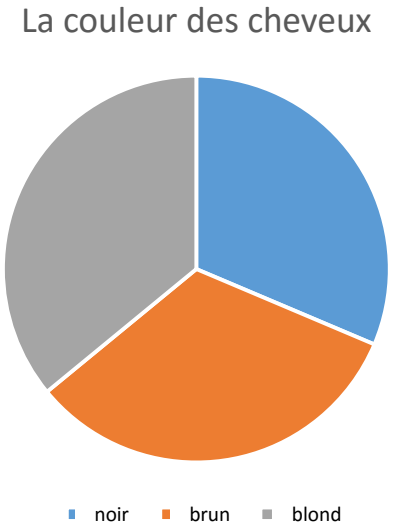
$$\sum n_i = n$$

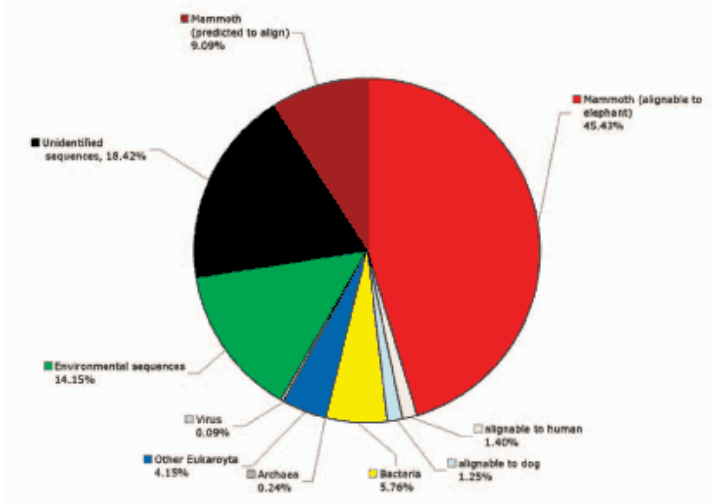


$$\sum f_i = 1$$

Variables qualitatives discrètes

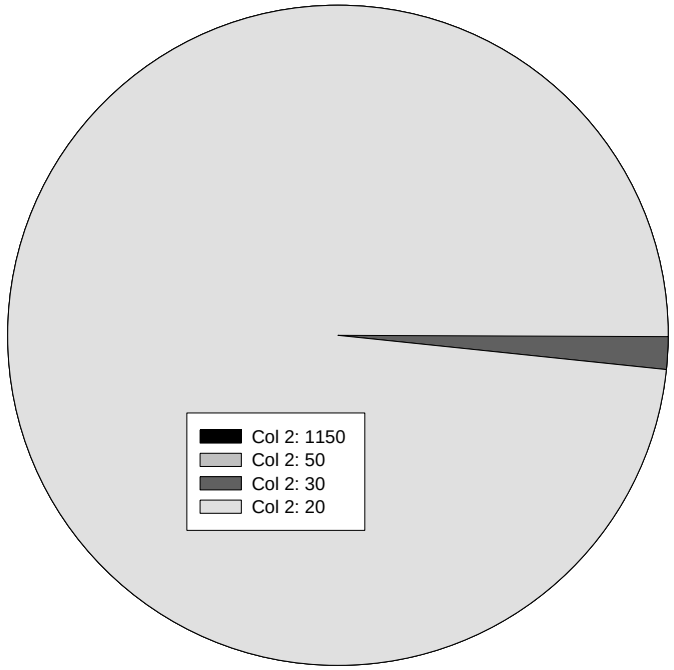
Diagrammes circulaires (pie chart / « camembert »)

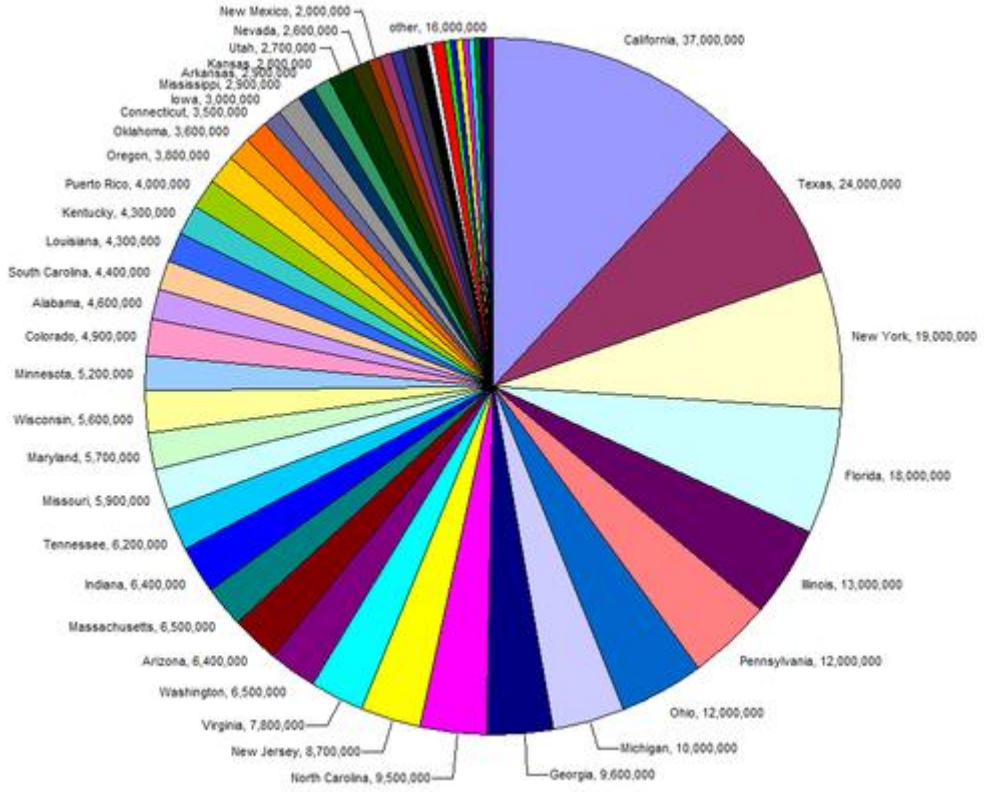




Variables qualitatives discrètes

Nombre de périodes de chômage en une année



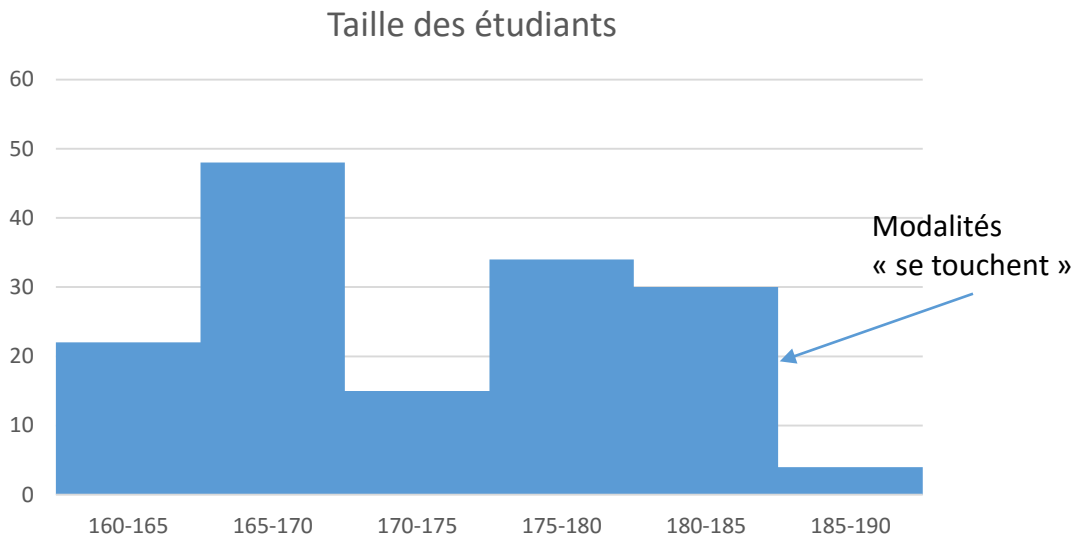


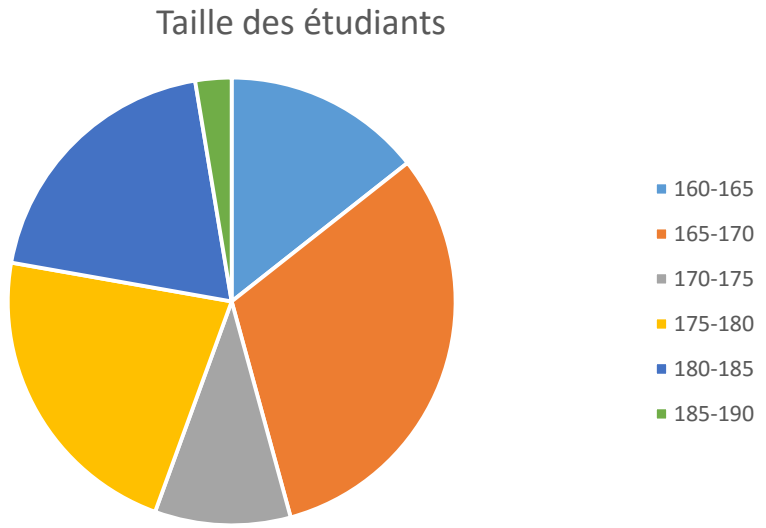
Variables quantitatives discrètes

- *Ex: La taille des étudiants en classes*

- On utilise
  - Diagramme des bâtons
  - Diagramme circulaire
  - Diagramme des courbes des fréquences cumulées

- Dans le diagramme des bâtons, les bâtons « se touchent » (car les valeurs sont quantitatives!)



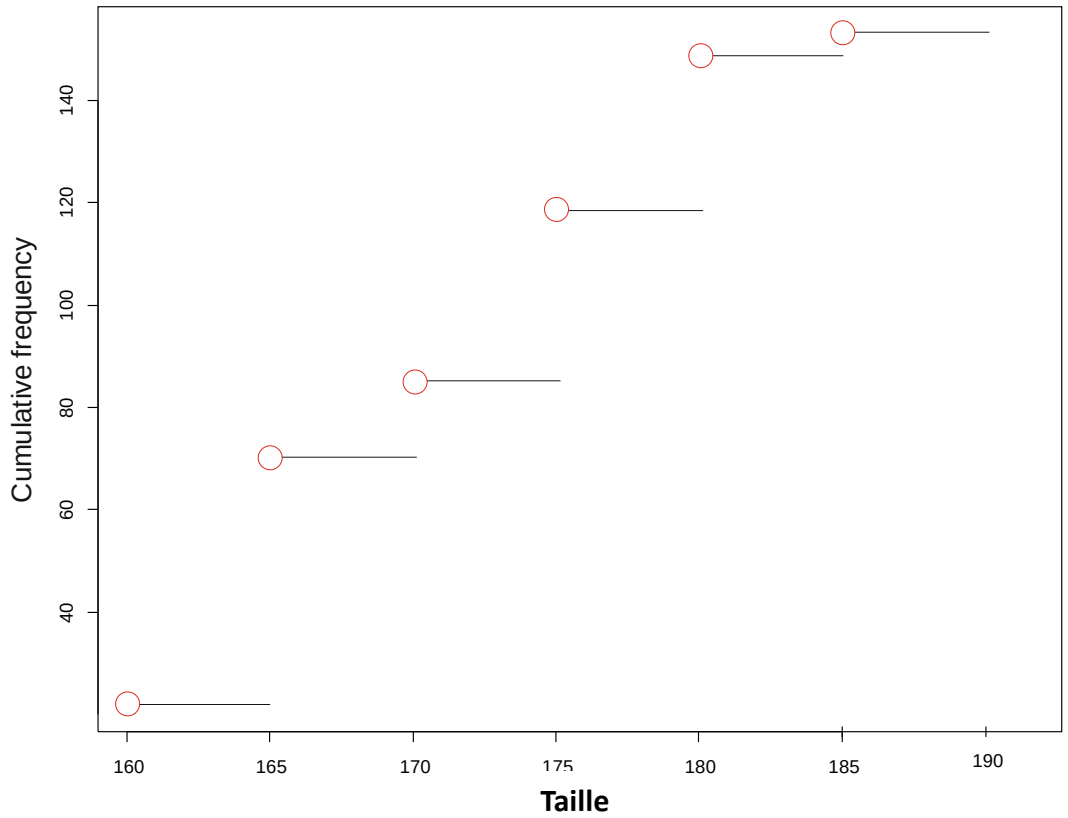


Variables quantitatives discrètes

Courbes des fréquences cumulées

- Il s'agit de courbes en escalier, c.-à-d. constantes sur chaque intervalle défini par deux modalités successives

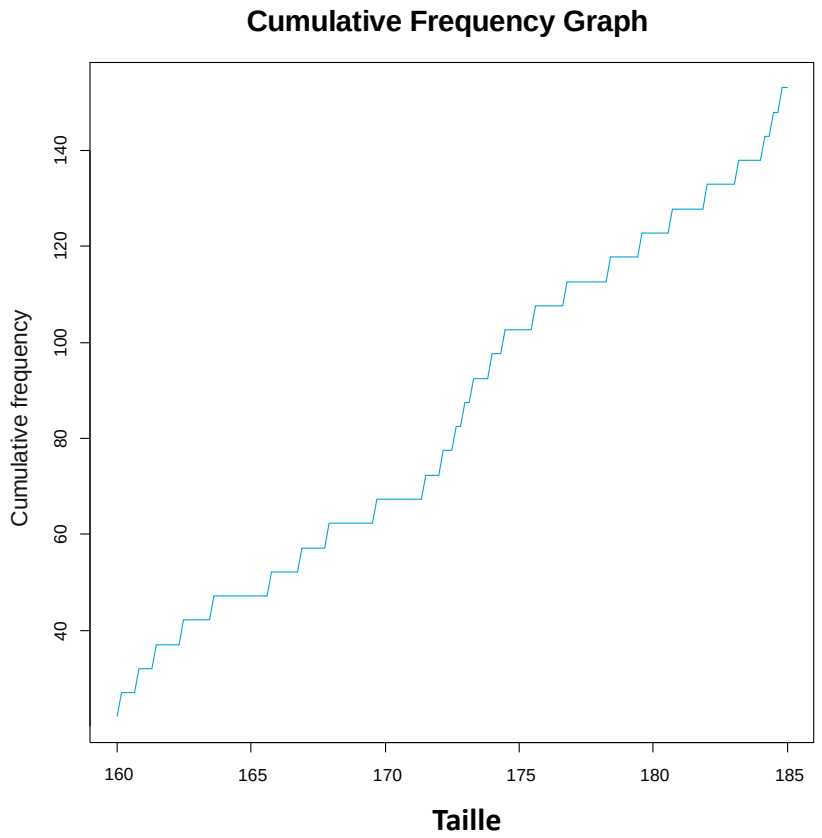
Taille	Fréquence absolue	Fréquence absolue cumulée
160-165	22	22
165-170	48	70
170-175	15	85
175-180	34	119
180-185	30	149
185-190	4	153



Variables quantitatives continues

Courbes des fréquences cumulées

Taille	Fréquence absolue	Fréquence relative cumulée
161	1	1
162	4	5
163	4	9
164	6	15
…	…	…
187	2	153

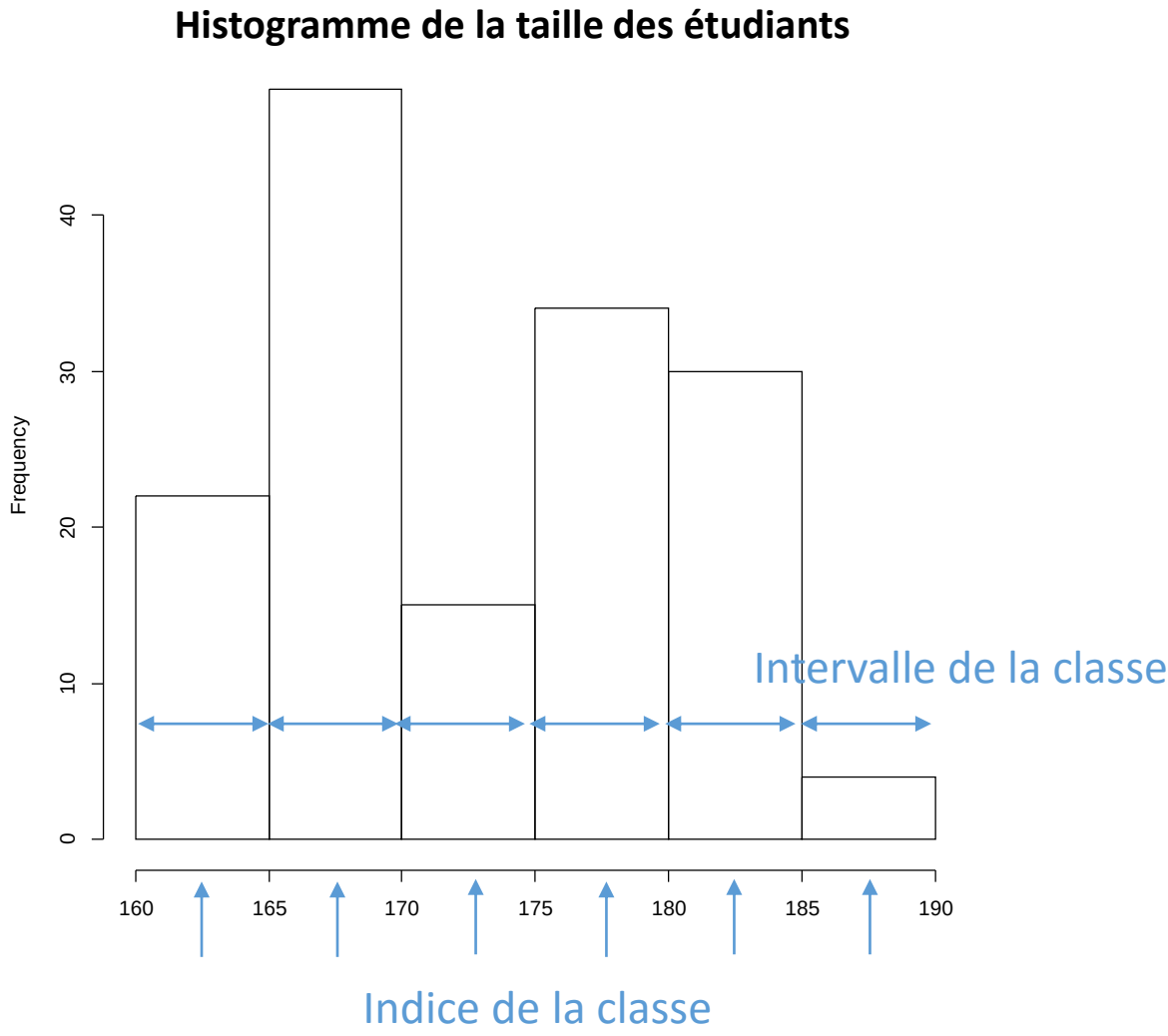


Variables quantitatives continues

Histogrammes

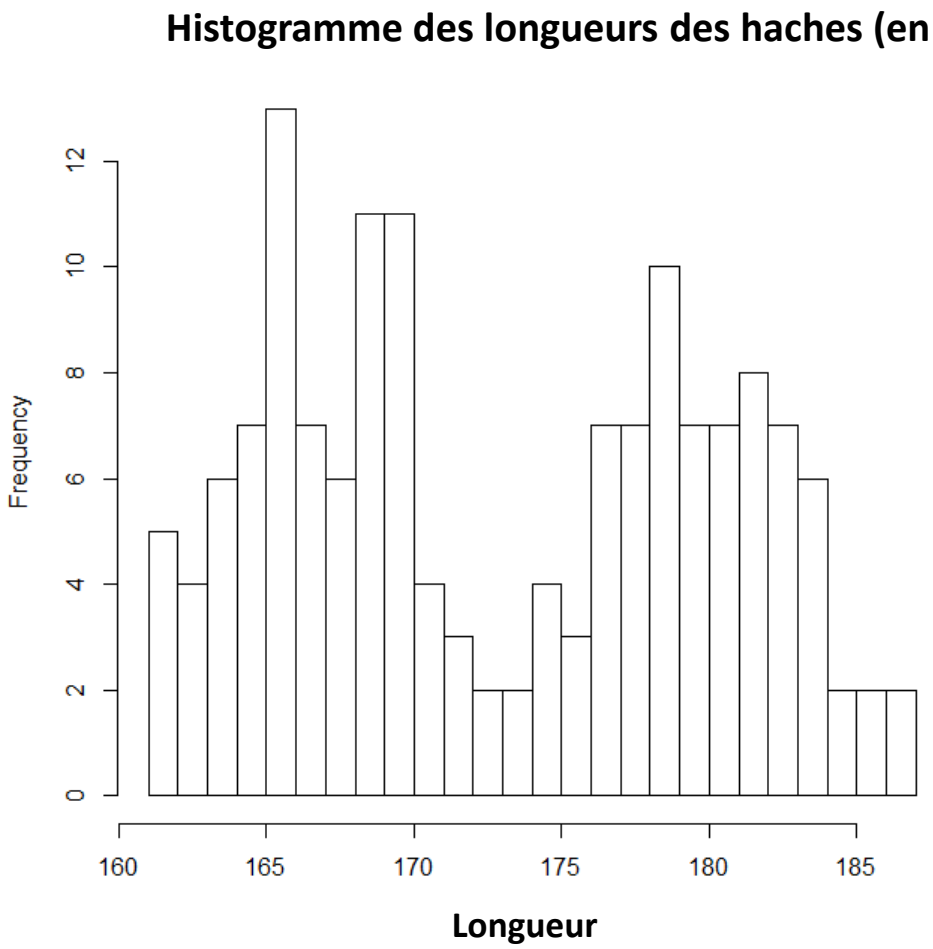
- Pas confondre avec le diagramme de bâton !
- les bâtons « se touchent » (car les valeurs sont quantitatives!)

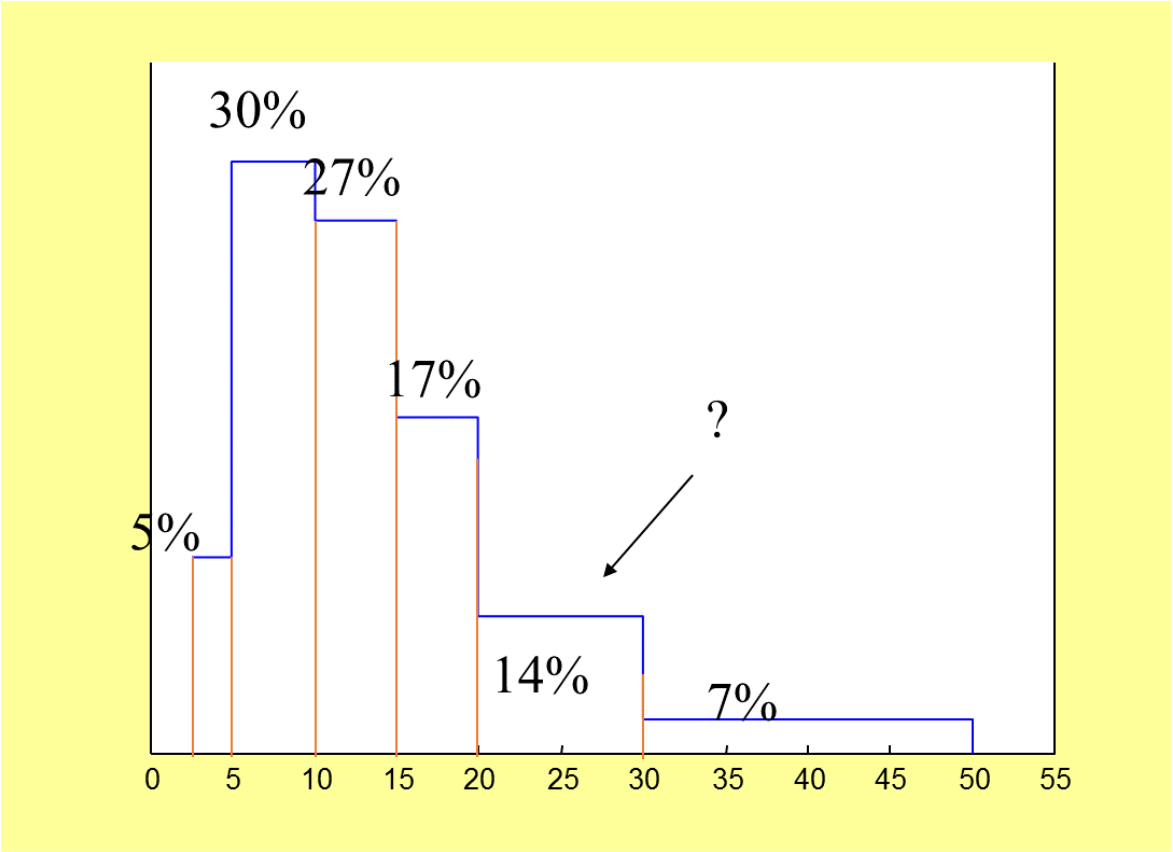
- Combien des classes ?
  - Règle de Sturge: $nombre\ de\ classes = 1 + (3.3log_{10}n)$
  - Règle de Yule: $nombre\ de\ classes = 2.5\sqrt[4]{n}$
  - *n* est le nombre d'individus



Sturge: $1 + (3.3log_{10}153) = 1 + (3.3 * 2.18) = 8$
Yule: $2.5\sqrt[4]{153} = 2.5 * 3.5 = 9$

Variables quantitatives continues



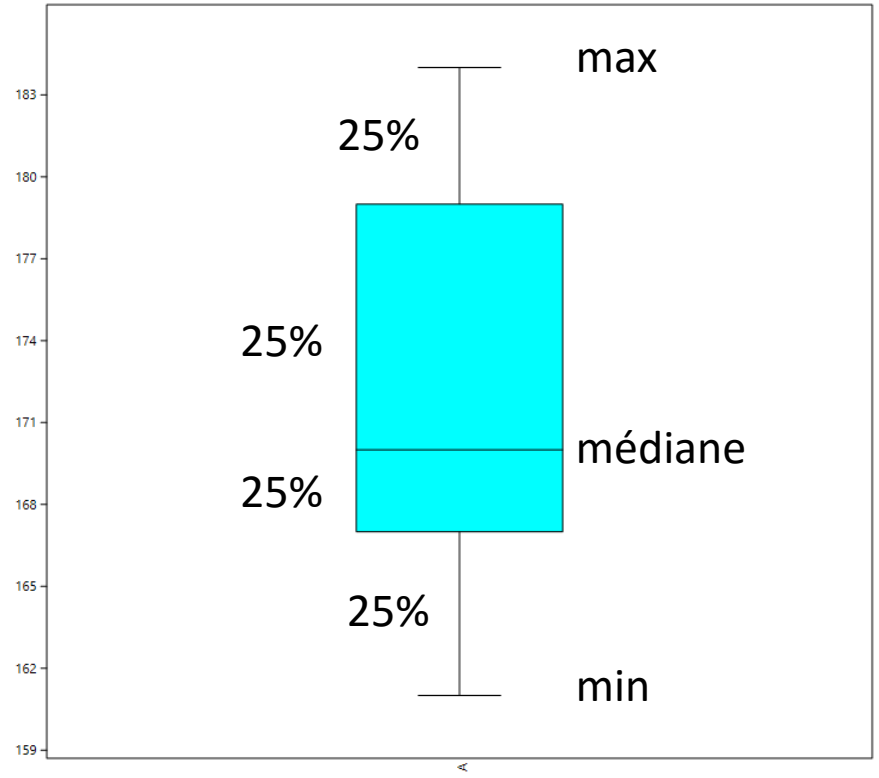


Variables quantitatives continues

Boîtes à moustaches (boxplots)

- Graphique, qui sépare les donnés sur 4 parties qui sont plus ou moins égales
- Pour variables quantitatives continus et discrètes
- Pourquoi ?
  - On peut voir où les données se concentrent
  - Pour trouver intervalle comprenant 50 % des observations les plus au centre de la distribution

Boxplot de la taille des étudiants



Comment faire les tableaux des fréquences et des graphiques ?

Dataset

- 150 haches d'Age du Bronze
- Tombe (homme / femme)
- Taille (variable continue)
- Poids (variable continue)
- Type (Neyruz / Salez / Altantique)

Excercice (Excel)

1) Faire un tableau des fréquences et barplot pour les variables Tombe et Type
2) Faire un camembert pour les variables Tombe et Type

Excercice (PAST)

1) Faire un barplot pour Tombe et Type
2) Faire un camembert pour Tombe et Type
3) Faire un histogramme de la Taille et Poids
4) Faire un boxplot de la Taille et Poids
5) Comparer la Taille des H et F par boxplots

1

	FEMME	HOMME	Atlantique	Neyruz	Salez
Type	-	-	-	-	-
Name	FEMME	HOMME	Atlantique	Neyruz	Salez
1 •	80	73	48	50	55

2

Color	Symbol	Name	Tombe
		-	-
		3	Tombe
Black	Dot	F	80
Black	Dot	M	73

		-	-
		3	Type
Black	Dot	Atlantique	48
Black	Dot	Neuruz	50
Black	Dot	Salez	55

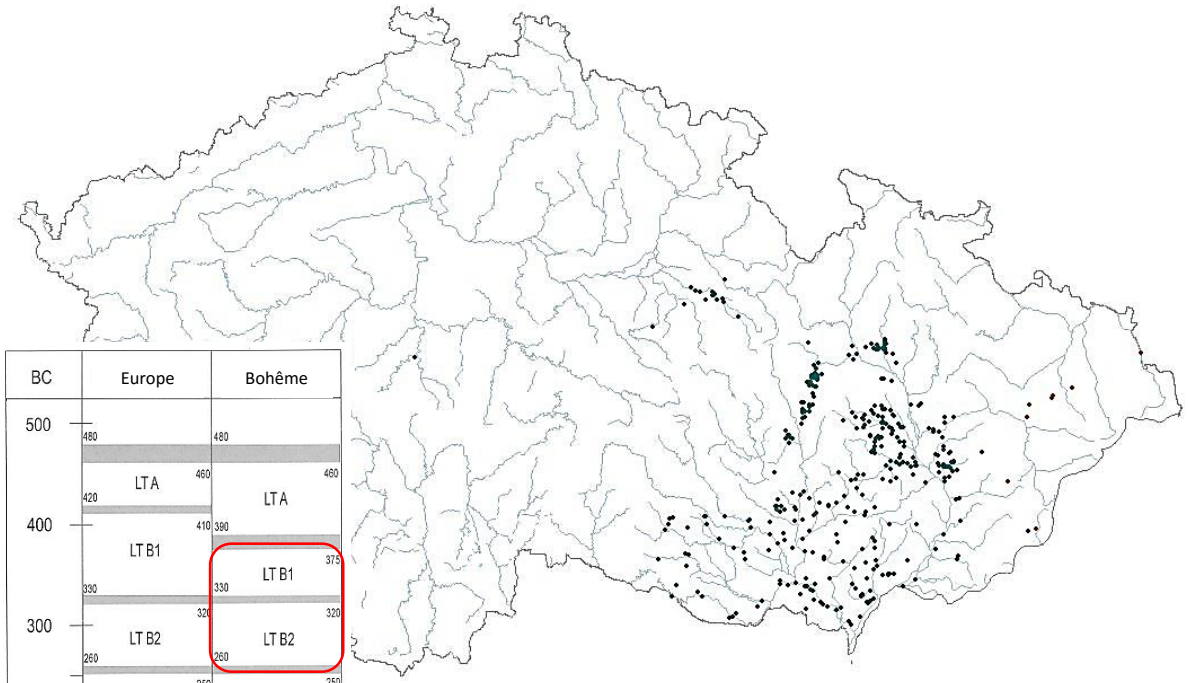
3-4

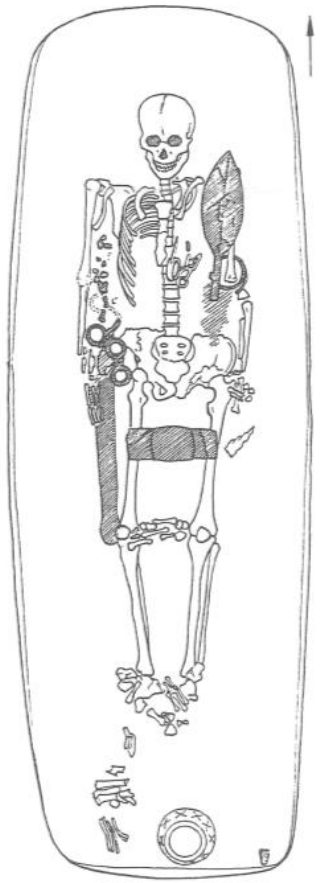
	Taille	Poids
Type	-	-
Name	Taille	Poids
1 •	164	64
2 •	169	64
3 •	168	65
4 •	164	62
5 •	167	66
6 •	169	66
7 •	165	69
8 •	167	62

5

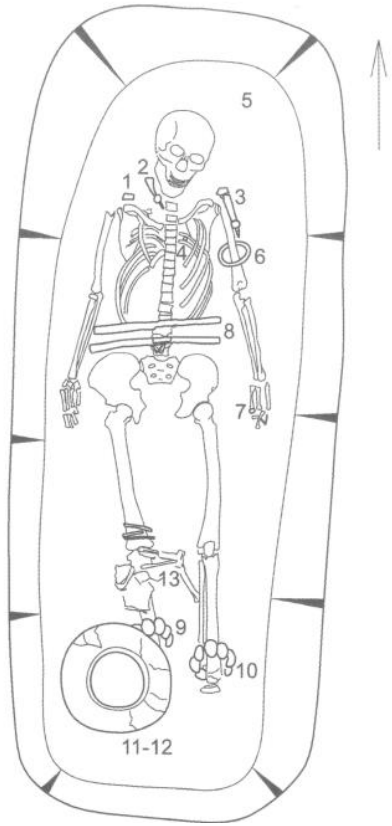
Type	-	-
Name	M	F
1 •	164	168
2 •	169	166
3 •	168	168
4 •	164	169
5 •	167	168
6 •	169	169
7 •	165	170
8 •	167	179
9 •	166	181
10 •	169	177

Problématique

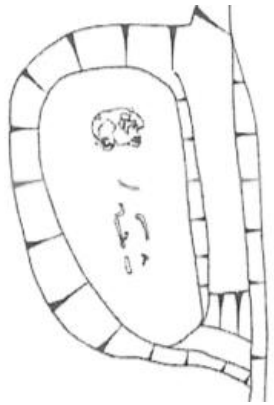




Homme

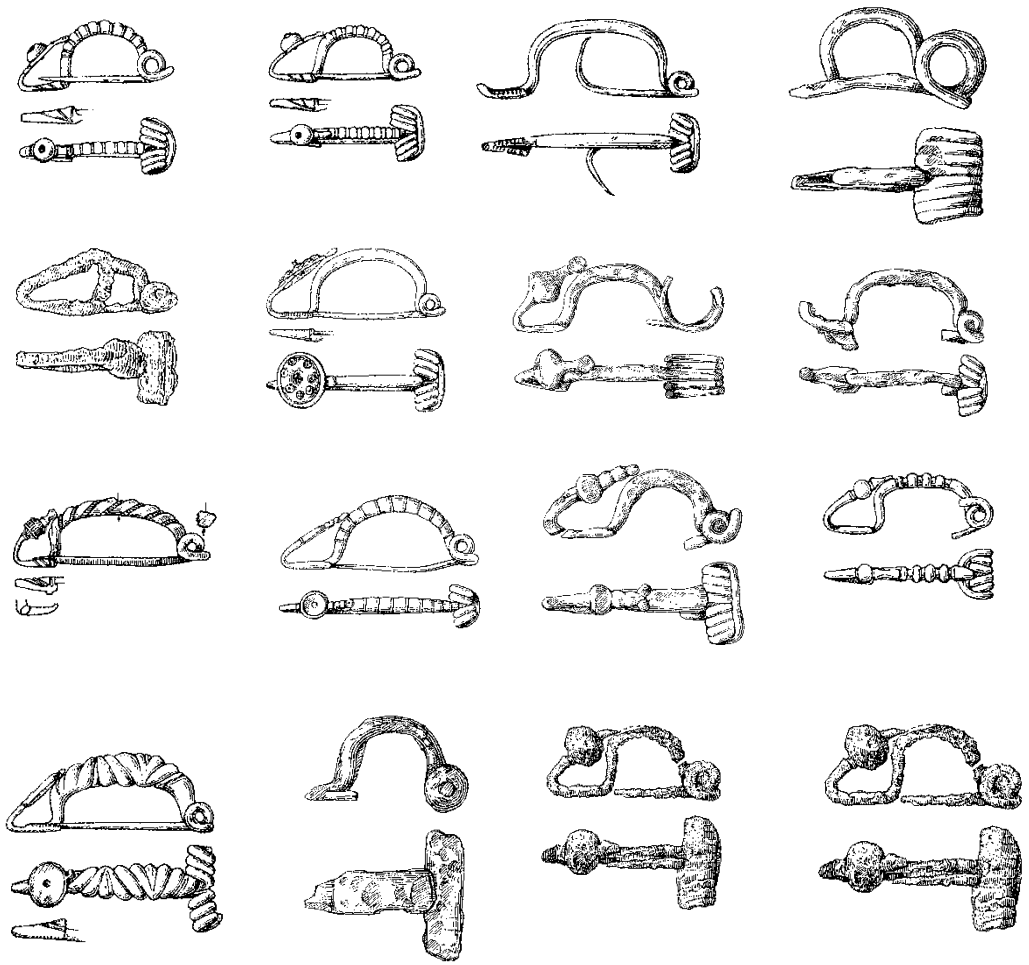


Femme



Enfant

Problématique



Homme

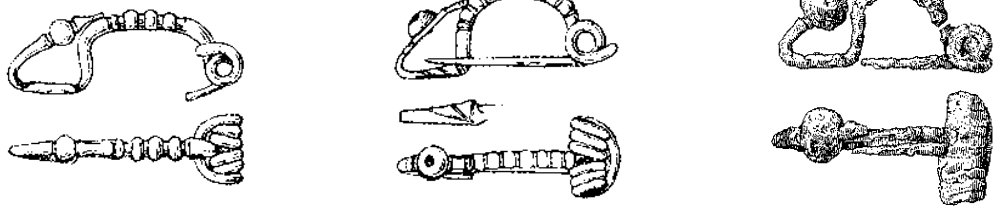
Femme

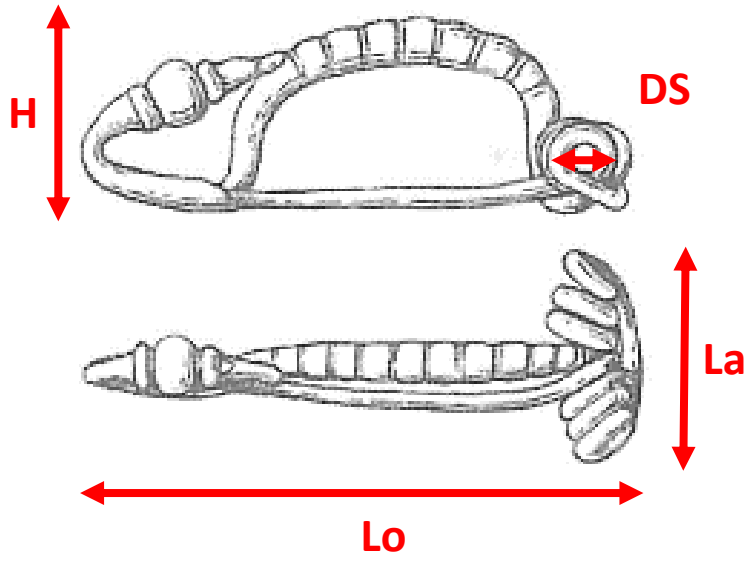
Enfant

Problématique

- 120 fibules ont été analysés

- 5 paramètres pour chaque fibule
  - Lo – longueur
  - La – largeur
  - H – hauteur
  - DS – diamètre de la spirale
  - Type (Duchcov, Münsingen, Bulle)





Fibule	Gendre	Lo	La	H	DS	Type
Fib1	Homme	34,33	30,08	26,34	11,53	Duchcov
Fib2	Homme	49,97	22,08	23,49	9,31	Munsingen
Fib3	Homme	57,25	12,78	17,56	5,82	Duchcov
Fib4	Homme	46,2	24,94	11,49	10,74	Bulle
Fib5	Homme	31,51	28,93	9,49	8,7	Duchcov
…	…	…	…	…	…	…
Fib120	Enfant	41,17	16,04	14,83	13,88	Munsingen

Problématique

Projet :

1) Comparez les nombre des fibules des trois genres (H / F / E) :
    1) par diagramme en bâtons (barplot)
    2) par diagrammes circulaires (pie chart / « camembert »)

2) Comparez les nombre des trois types des fibules (Duchcov / Munsingen / Bulle)
    1) par diagramme en bâtons (barplot)
    2) par diagrammes circulaires (pie chart / « camembert »)

3) Construisez des histogrammes pour chaque des variables appropriés, en déterminant le bon nombre des intervalles.

4) Comparez les longueurs des spirales (Lo) des trois types des fibules par boxplot. Sont ils différents ?

5) Comparez les largeurs des spirales (La) des trois genres par boxplot. Sont ils différents ?

6) Divisez la variable hauteurs des spirales (H) dans les intervalles (par 5 cm).
    1) Construisez un table de fréquences absolues et relatives (en Excel)
    2) Construisez un table de fréquences absolues et relatives cumulés (en Excel)