



Statistiques descriptives

Volume horaire : 25h de présentiel sous forme de cours/TPs de 2h30
5h en salle pour un mini projet en binôme

Evaluation : Un examen et un mini projet. La note du module est la moyenne des deux notes.

La présence aux cours/TP est obligatoire. Les absences non justifiées minoreront la note du module.

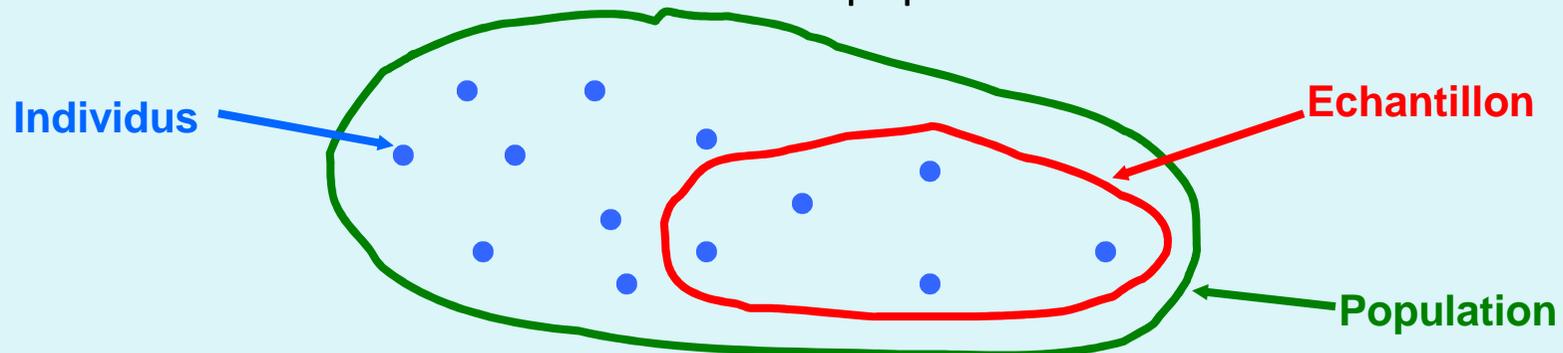
Mini projet : L'objectif du mini projet est de mettre en application les notions vues dans ce module sur votre propre jeu de données. Un document décrivant plus en détail le contenu du mini-projet se trouve sur Arel.

Vous devez dès à présent trouver votre jeu de données avec comme contraintes :

- Au moins 2 variables qualitatives (*) chacune avec moins de 5 modalités (*)
- Au moins 5 variables quantitatives (*)
- Un nombre d'observations supérieur à 30

Votre jeu de données doit être validé par votre enseignant référent avant le 03/11/2014
Vous pouvez vous aider des sites d'Europstat ou de l'Ined.

- Définition : Ensemble de méthodes scientifiques qui permettent de *décrire* un corpus de données observées au travers :
 - une présentation la plus synthétique possible
 - une représentation graphique appropriée
 - un résumé numérique
- Terminologie :
 - *Individus* : objets équivalents sur lesquels on observe les mêmes caractéristiques
 - *Population* : Ensemble des individus
 - *Echantillon* : Sous-ensemble de la population



- *Recensement* : Etude de tous les individus d'une population
- *Sondage* : Etude d'une partie de la population

- Statistiques descriptives :

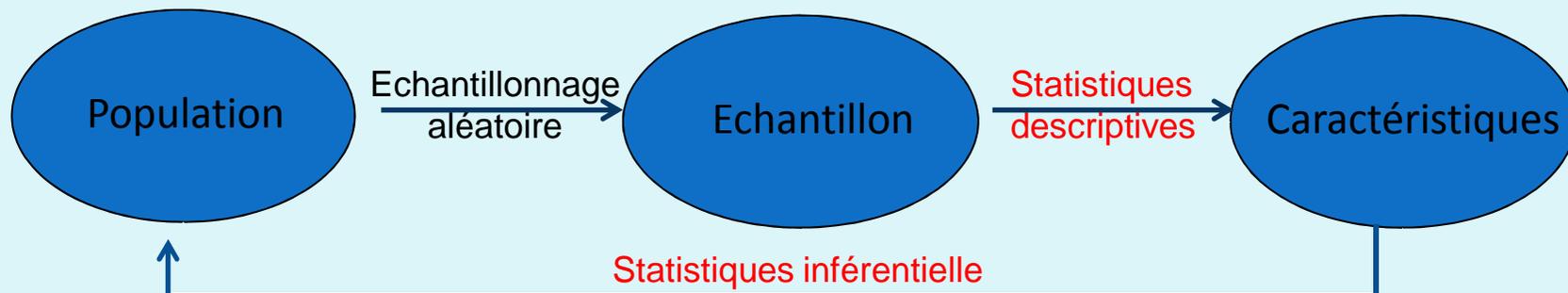
ING 1

- Présentation synthétique des données (tableau de contingence,...)
- Description à travers des résumés chiffrés : moyennes, médianes, écarts-types, corrélations, ...
- Description à travers des résumés graphiques : histogrammes, diagrammes en bâton ou circulaire, ...

- Statistique inférentielle :

ING 2

- Trouver des estimateurs non biaisés et efficaces pour passer de l'échantillon à la population.
- L'outil mathématique sous jacent est la théorie des probabilités



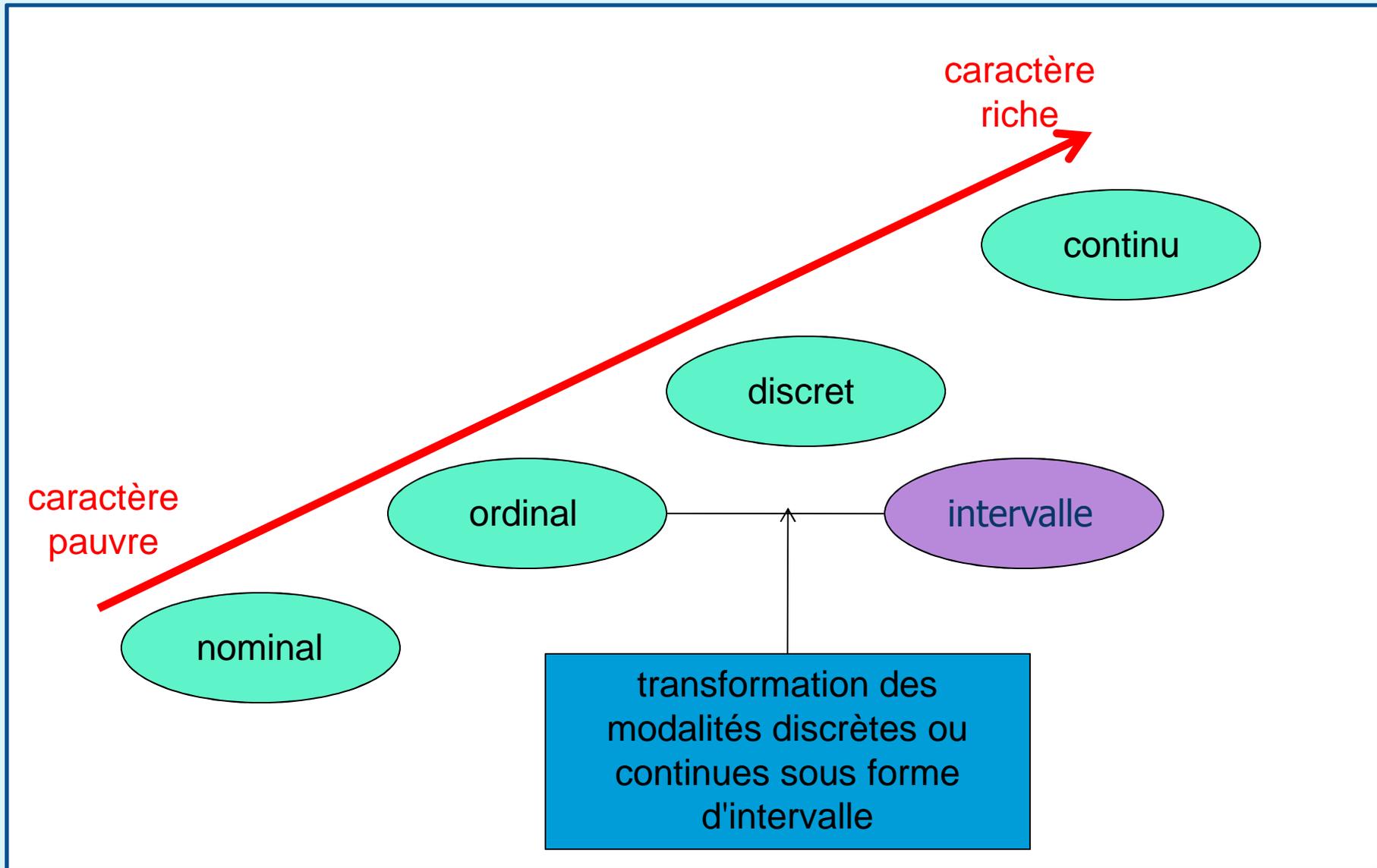
Chaque individu est décrit par un ensemble de caractéristiques appelées *variables (ou caractères)*

Les variables sont classées suivant leur nature :



Statistiques descriptives

Classification des type de variable





Statistiques descriptives

Les traitements

- Les statistiques **univariées** : On ne s'intéresse qu'à une seule variable.
Ex : Les salaires en Europe, nombre d'enfants par ménage, Temps de visite sur un site, Etude de l'âge des habitants d'un pays par classe d'âge.
- Les statistiques **bivariées** : On s'intéresse à l'étude simultanée de deux variables pour mesurer leur dépendance.
Ex : le vote est-il différent d'une CSP à l'autre ?
- Les statistiques **multivariées** : On s'intéresse à l'étude simultanée de p variables. Même problématique que pour le bidimensionnel mais avec p assez grand.
Ex : Etude du bien-être par département à travers des critères géo-socio-économiques (ensoleillement, nombre de théâtres, taux de suicides, infrastructure routière, ...)

A chaque type de variable (qualitatif nominal, quantitatif continu,...) correspond un traitement spécifique.



Statistiques descriptives

Méthodologie d'étude statistique

Etape 1 : Définir précisément le problème étudié :

- 1) Quels sont les objectifs de l'étude ?
 - ✓ recensement des différentes questions posées
 - ✓ déduction des différentes études statistiques à opérer
 - ✓ définition des caractères étudiés avec leur type
- 2) Quelle est la population étudiée ?
 - ✓ définition précise de l'unité statistique
 - ✓ définition du périmètre spatio-temporel
- 3) Comment récupérer et stocker l'information ?
 - ✓ Enquête ou données existantes ou un mixte
 - ✓ Choix de la technique d'échantillonnage
 - ✓ Récupération des données et validation des données récupérées

Etape 2 : Exécution des études statistiques avec les logiciels appropriés

Etape 3 : Rédaction du document de synthèse

- 1) Rappel du contexte : objectifs de l'étude, périmètre de l'étude, définitions des études statistiques
- 2) Insertion par étude des résumés chiffrés et graphiques
- 3) Interprétation des résultats en cohérence avec le périmètre et les résumés

Une présentation synthétique des données commence par un *tableau de contingence*.

Effectif : Pour chaque variable, il s'agit de compter le nombre d'individus ayant la même valeur/modalité i . On utilisera le terme effectif de la valeur/modalité i . On notera cet effectif n_i .

Pays	Taux de chômage	PIB	Zone Euro (avant 2010)
Allemagne	5,5	37430,1	Zone Euro
Autriche	4,4	40064,8	Zone Euro
Belgique	7,6	37727,8	Zone Euro
Danemark	7,5	40189,9	Pas Euro
...



Zone Euro (avant 2010)	
Pas Euro	7
Zone Euro	13
Total	20

Taux de chômage	
[4,4;9,6[9
[9,6;14,8[7
[14,8;19,9[1
[19,9;25,1[2
Total	20

N.B. Le tableau de contingence des variables continues nécessite un regroupement des valeurs (cf. histogramme)

Fréquence : Quand 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.

La fréquence permet de comparer des échantillons de tailles différentes

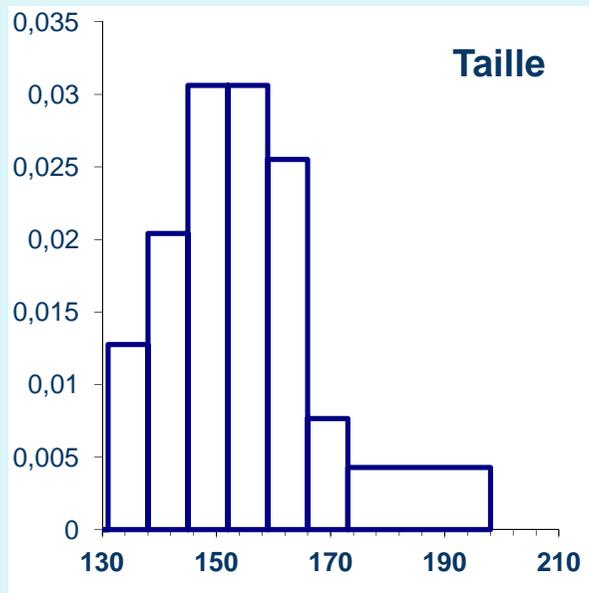
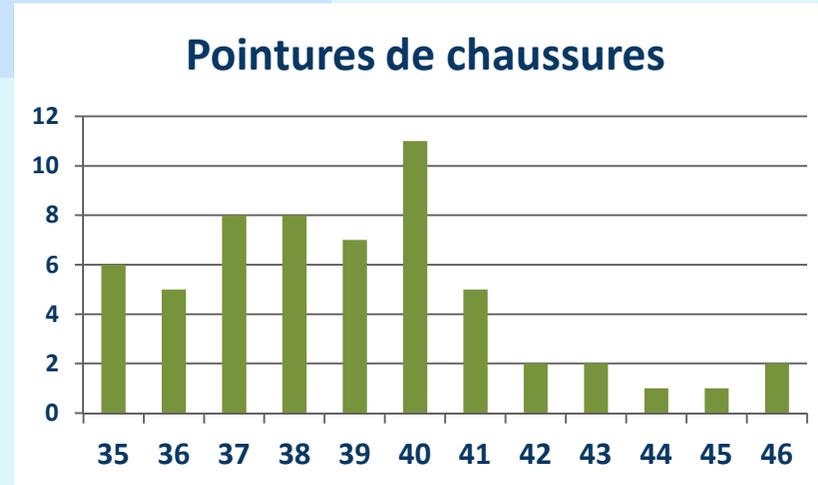
Zone Euro (avant 2010)	
Pas Euro	35%
Zone Euro	65%
Total	100%

Variables quantitatives – Représentation graphique

Représentation graphique d'une variable discrète

Diagramme en bâtons

- bâton par valeur discrète
- hauteur du bâton proportionnelle à l'effectif de la valeur



Représentation graphique d'une variable continue

Histogramme

- regroupement des valeurs par intervalle (classe)
- nombre de classes $\approx E[1+10 \times \log(n)/3]$
- base du rectangle proportionnelle à la longueur de l'intervalle
- hauteur du rectangle proportionnelle à l'effectif



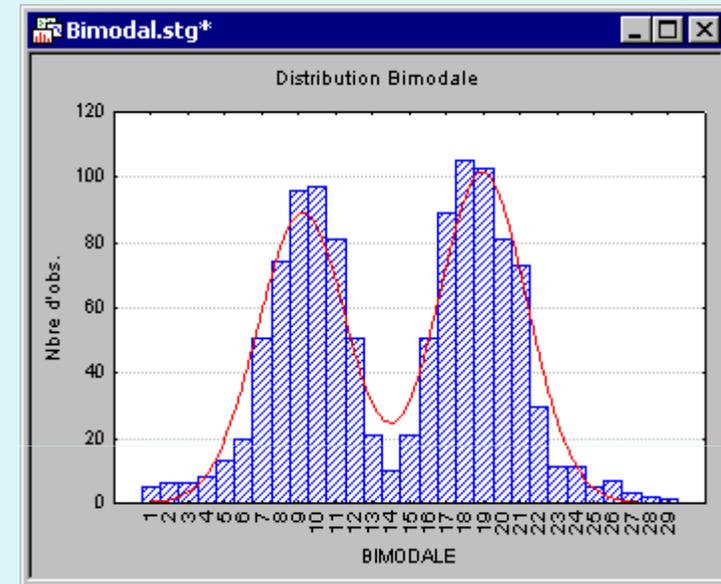
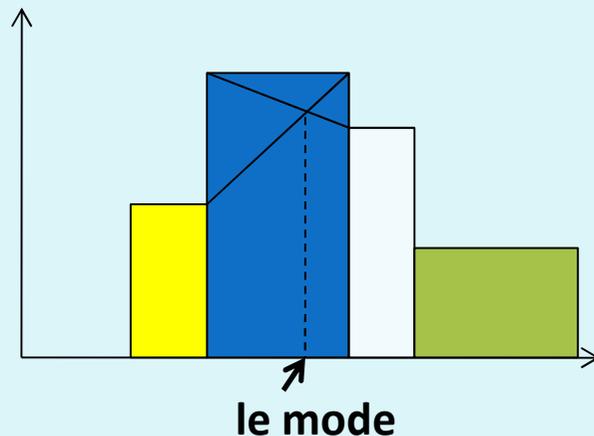
Statistiques univariées

Variables quantitatives - Résumés numériques

On distingue deux types de résumés numériques :

- Les *indicateurs de position* (moyenne, mode, médiane, quartiles). Ils positionnent la série des valeurs observées autour d'une tendance centrale.
- Les *indicateurs de dispersion* (variance, écart-type, étendue inter-quartile). Ils indiquent la fluctuation des valeurs de la série autour en général d'une tendance centrale.

- **Le mode** est la valeur observée d'effectif maximum. Il sert essentiellement à détecter si la population est homogène ou éventuellement constituée de deux ou plusieurs sous-populations. Dans le cas du type quantitatif continu il faut tenir compte des classes adjacentes.

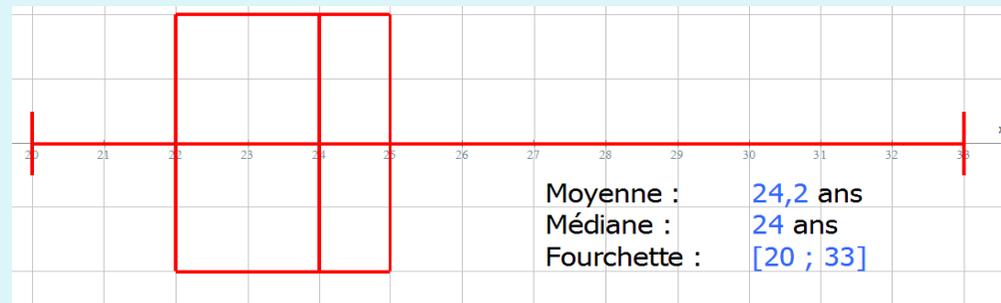


- **La moyenne** est calculée en faisant la somme des valeurs de la série pondérées par leur effectifs. Cette somme est ensuite divisée par l'effectif total. Cet indicateur est très sensible aux valeurs extrêmes de la série.

➤ Les quartiles :

- **La médiane** est la valeur qui sépare la population en deux groupes d'effectifs égaux. Elle n'a de sens que sur une série rangée par ordre croissant.
- **Le 1^{er} quartile Q_1** est la valeur qui sépare la série en $\frac{1}{4}$ inférieur et $\frac{3}{4}$ supérieur.
- **Le 3^{ème} quartile Q_3** est la valeur qui sépare la série en $\frac{3}{4}$ inférieur et $\frac{1}{4}$ supérieur.

La représentation graphique de ces indicateurs est la **boîte de Tukey**. Elle permet d'avoir un aperçu graphique rapide de la distribution des valeurs de la série et permet beaucoup d'interprétation.



Les valeurs extrêmes de cette représentation sont les « moustaches » définies en général par

$$m = Q_1 - 1,5 \times (Q_3 - Q_1) \quad \text{et} \quad M = Q_3 + 1,5 \times (Q_3 - Q_1)$$

Toutes valeurs de la série en dehors des moustaches est considérée comme **atypique**

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

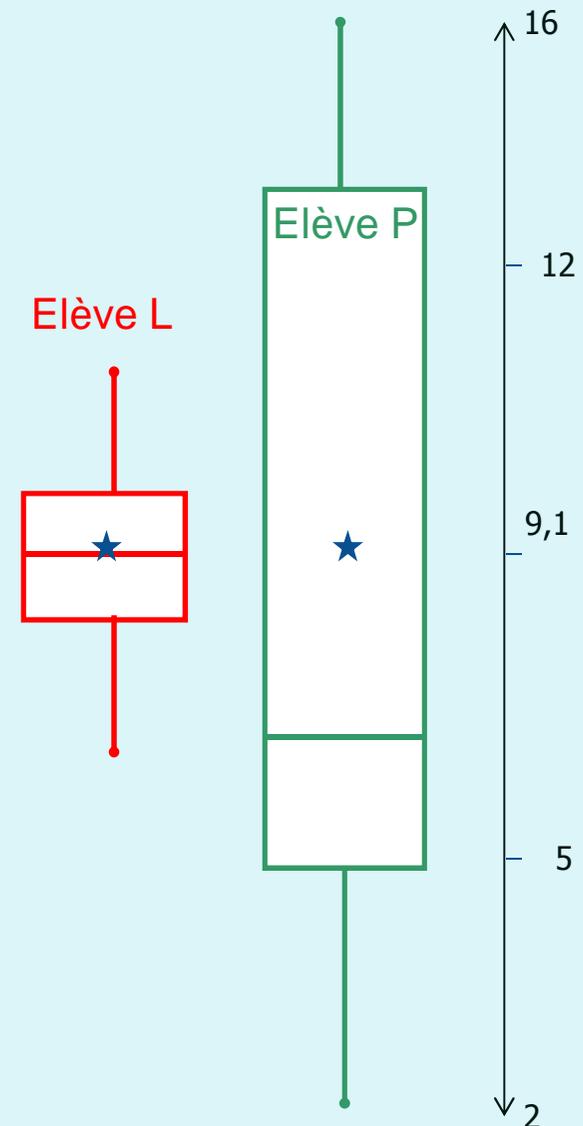
Elève L	7	8	9	9	10	10	11
Elève P	2	5	5	6	14	16	16

Tendance centrale

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

Dispersion

	Variance	Ecart-type	Q1	Q3
Elève L	1,81	1,34	8	10
Elève P	35,48	5,96	5	14



- *La variance* mesure l'écart au carré entre les valeurs de la série et leur moyenne

$$s^2 = \frac{1}{n} \sum_i n_i (x_i - \bar{x})^2$$

Afin de garder la même unité que la variable, on utilise *l'écart-type* $s = \sqrt{s^2}$

Tout comme la moyenne ces deux indicateurs sont sensibles aux valeurs extrêmes de la série.

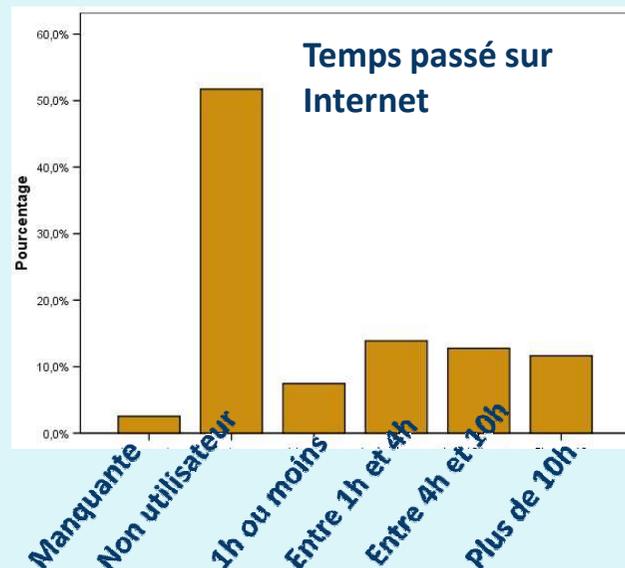
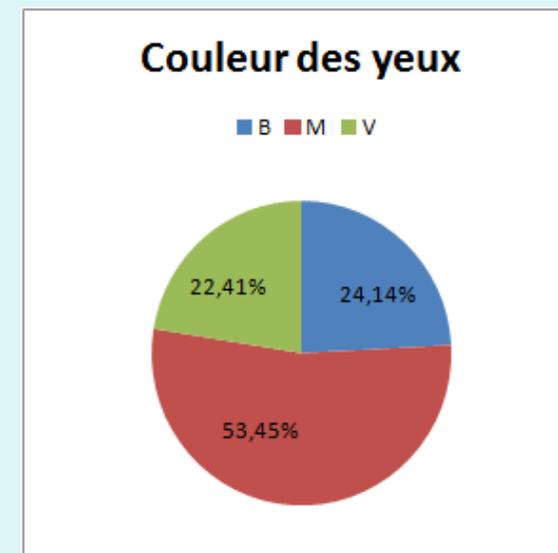
- *L'écart-médian* mesure l'écart entre les valeurs de la série et leur médiane

$$em = \frac{1}{n} \sum_i n_i |x_i - \bar{x}|$$

On peut aussi utiliser *l'étendue* de la série $\max \{x_i\} - \min \{x_i\}$ ou *l'écart inter-quartiles* $Q_3 - Q_1$

Les observations d'une variable qualitative sont des modalités et ne sont pas numériques. Les traitements précédents n'ont donc pas lieu d'être (moyenne, variance,...) sauf le mode. On se contente de faire des tableaux de contingence et des représentations graphiques. Pour les représentations graphiques, on distingue

- le cas de variables nominales avec un diagramme en secteurs



- le cas des variables ordinales avec un diagramme en bâtons

Variables qualitatives

nominales

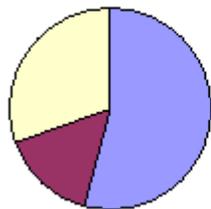
ordinales

Variables quantitatives

discrètes

continues

Répartition selon la couleur préférée



■ Bleu : 53,85%
 ■ Rouge : 15,38%
 □ Jaune : 30,77%

