Data exploration

30h TD/TP

Évaluation (1ère et 2nde session) : Examen 2h sur papier/ Etude préalable d'un ou plusieurs jeux de données

Supports pédagogiques : Support de présentation et énoncé de TD/TP à chaque séance
Tutoriel sur l'ensemble du programme

L'objectif de ce module est d'introduire les méthodes permettant d'extraire de l'information pertinente de corpus de données.
A l'issue de cette formation, vous serez capables de :

- faire apparaitre des comportements particuliers des objets observés (détecter les individus ayant un comportement atypique, trouver des ensembles d'individus ayant un comportement similaire),
- trouver des liens entre les variables étudiées (ex: est-ce que le PIB d'un pays est lié à l'investissement en R&D, est-ce qu'un médicament a un impact sur le taux de guérison?),
- utiliser le langage dédié aux statistiques R (pas programmer mais utiliser).

Seules les méthodes descriptives seront abordées dans ce module. Il se complète avec les modules de 2ème année : IA (modèles de prévision/machine learning) et statistiques (gestion de l'incertitude)

Corpus/jeu de données

Un jeu de données est un tableau avec

- en ligne les individus/instances observés (unité statistique)
- en colonne les variables/attributs étudiées

Sur l'exemple ci-dessous les individus sont les incidents survenus dans la ville de Chicago depuis 2001 et les variables sont la date, le type d'incident, la position géographique, etc…

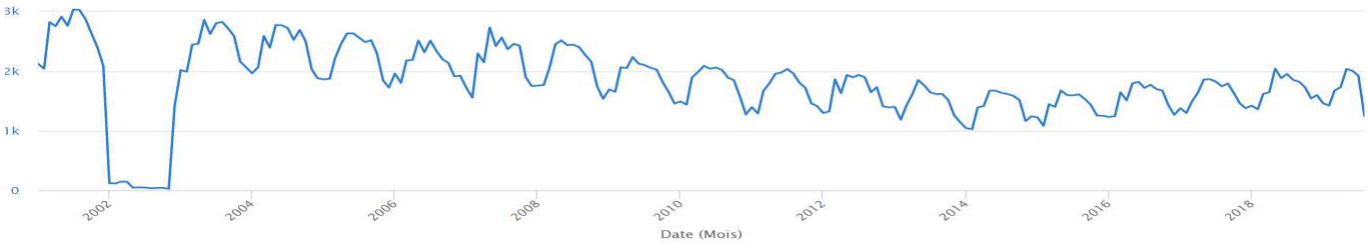
Incidents in Chicago from 2001 to Present

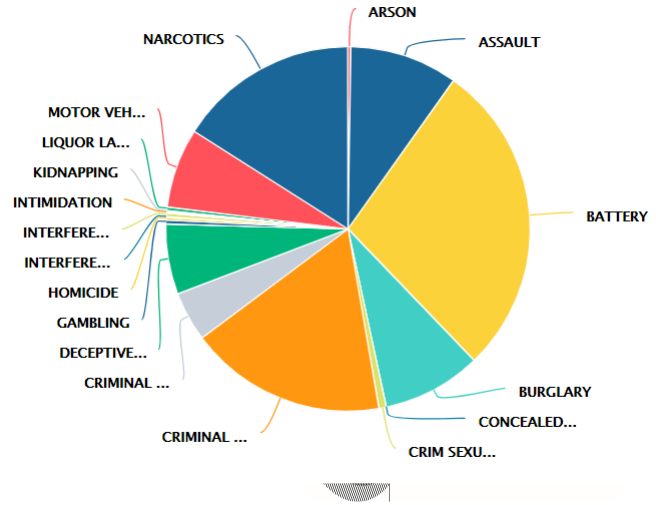
Informations | Tableau | Carte | Analyse | Export | API

	ID	Case Number	Date	Block	IUCR	Primary Type	Description	Location Descrip
1	9788696	HX438215	22 septembre 2014 05:00	030XX S KOSTNER AVE	0460	BATTERY	SIMPLE	STREET
2	9779391	HX428579	15 septembre 2014 12:00	052XX S NEWCASTLE AVE	1310	CRIMINAL DAMAGE	TO PROPERTY	RESIDENCE
3	9752199	HX402207	25 août 2014 07:50	002XX W CHICAGO AVE	0460	BATTERY	SIMPLE	STREET
4	9831862	HX481538	25 octobre 2014 13:00	072XX S CAMPBELL AVE	041A	BATTERY	AGGRAVATED: HANDGUN	RESIDENTIAL YA
5	9825038	HX475187	18 octobre 2014 23:00	044XX N SPAULDING AVE	0910	MOTOR VEHICLE THEFT	AUTOMOBILE	STREET
6	9840868	HX489480	17 octobre 2014 14:00	020XX W BIRCHWOOD AVE	5002	OTHER OFFENSE	OTHER VEHICLE OFFENSE	STREET
7	5016770	HM620064	24 septembre 2006 12:00	059XX S CALUMET AVE	1310	CRIMINAL DAMAGE	TO PROPERTY	RESIDENCE

Pour tirer de l'information pertinente et compréhensible par tout le monde à partir de ce genre de tableau il faut produite :

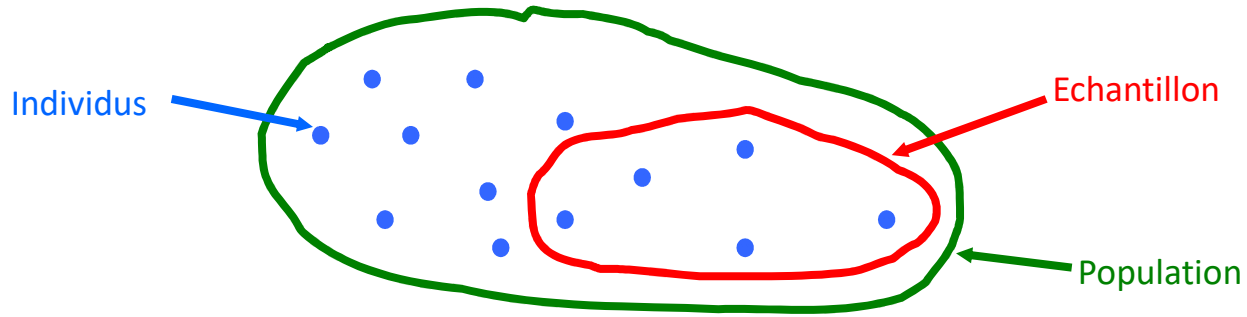
- une présentation la plus synthétique possible
- une représentation graphique appropriée
- un résumé numérique





Population et échantillon

- *Population* : Ensemble des individus
- *Echantillon* : Sous-ensemble de la population



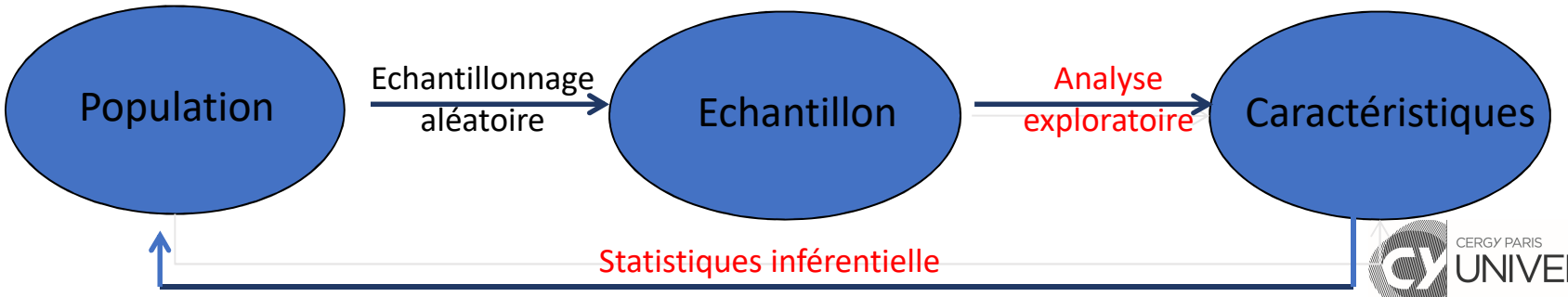
- *Recensement* : Etude de tous les individus d'une population
- *Sondage* : Etude d'une partie de la population

Remarque : la population peut être infinie et le recensement impossible.

Les statistiques inférentielles (ING2) permettent d'extrapoler des résultats observés sur un échantillon à toute une population.

Analyse exploratoire et statistique inférentielle

- Analyse exploratoire:
  - 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, …
  - Recherche de sous-groupes homogènes (clustering)
- Statistique inférentielle :
  - Trouver des estimateurs non biaisés et efficients pour passer de l'échantillon à la population.
  - L'outil mathématique sous jacent est la théorie des probabilités



Nature des variables

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

Les variables sont classées suivant leur nature :

Différents types d'analyse

Vu au lycée

- 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 de 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.

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 variables étudiées 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

Analyse univariée

(Rappels et compléments)

Représentation synthétique

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 x_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 une 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.

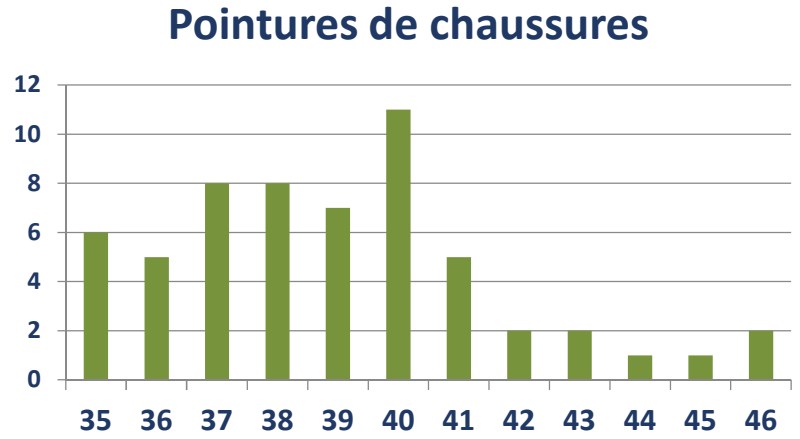
La fréquence permet de comparer des échantillons de tailles différentes

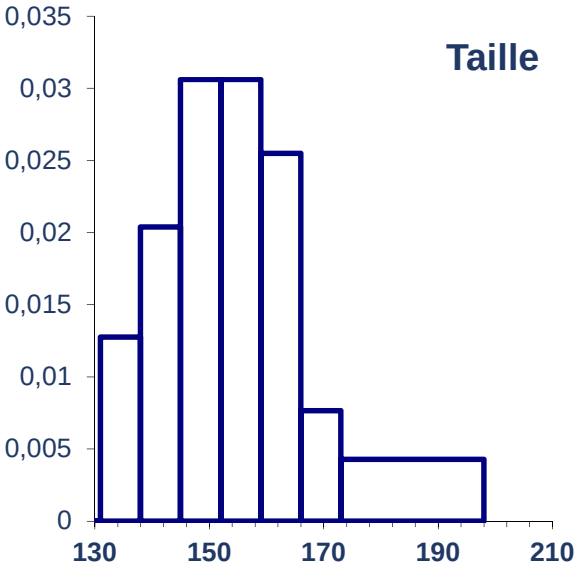
Zone Euro (avant 2010)	
Pas Euro	35%
Zone Euro	65%
Total	100%

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 $\approx E[1+10\times\log_{10}(n)/3]$
- base du rectangle proportionnelle à la longueur de l'intervalle
- hauteur du rectangle proportionnelle à l'effectif

Variables quantitatives : Résumés numériques

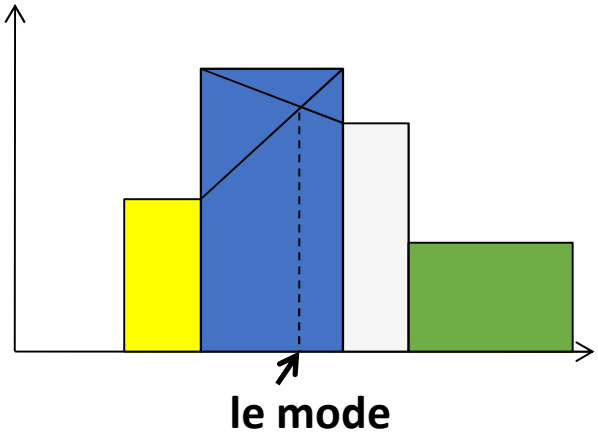
On distingue deux types de résumés numériques :

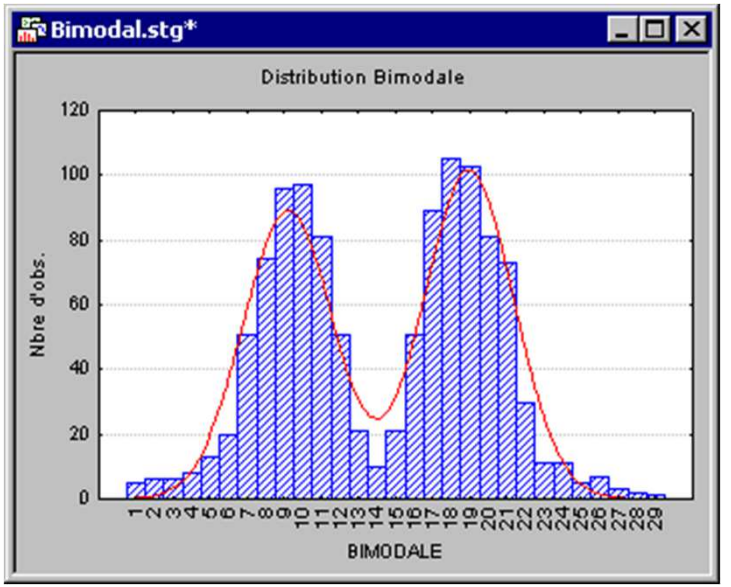
• Les ***indicateurs de position*** (moyenne, mode, médiane, quartiles). Ils positionnent la série des valeurs observées autour d'une tendance centrale.

• Les ***indicateurs de dispersion*** (variance, écart-type, étendue inter-quartile). Ils indiquent la fluctuation des valeurs de la série autour d'une tendance centrale en général.

Variables quantitatives : Indicateurs de position (1/3)

- ***Le mode*** est la valeur observée d'effectif maximum.
  Il sert notamment 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***

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

Garde les mêmes propriétés que l'espérance

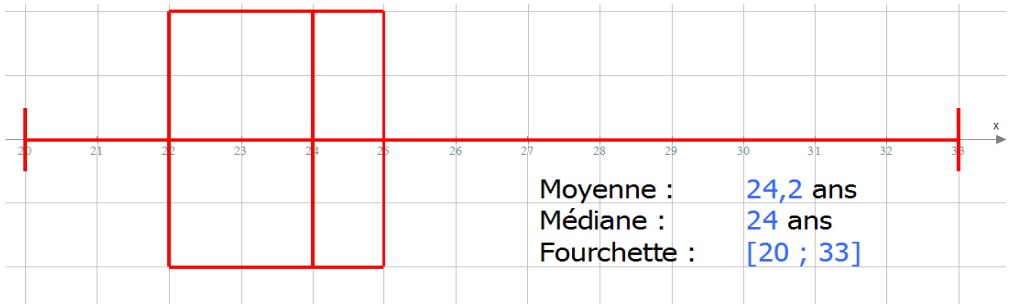
!!! Cet indicateur est très sensible aux valeurs extrêmes de la série !!!

Variables quantitatives : Indicateurs de position (2/3)

➢ ***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* ***1er quartile Q_1*** est la valeur qui sépare la série en ¼ inférieur et ¾ supérieur.
- *Le* ***3ème quartile Q_3*** est la valeur qui sépare la série en ¾ inférieur et ¼ supérieur.

La représentation graphique des ces indicateurs est la boite de Tukey. Elle permet d'avoir une 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

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

Toute valeur de la série en dehors des moustaches est considérée comme ***atypique***

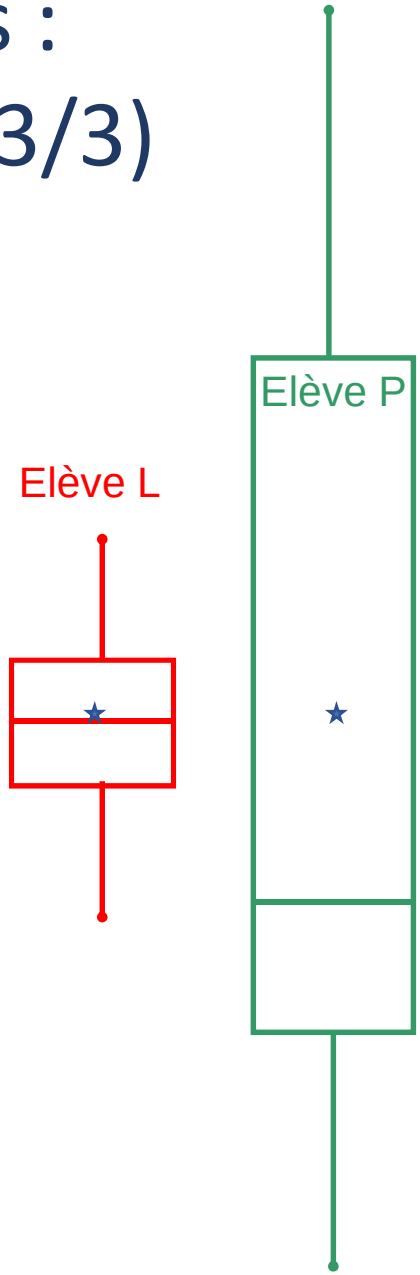
Variables quantitatives : Indicateurs de position (3/3)

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

	Moyenne	Médiane	Q1	Q3	m	M
Elève L	9,1	9	8	10	5	13
Elève P	9,1	6	5	14	0	27,5



Différence de fluctuation des notes ⇒ indicateurs de dispersion

Variables quantitatives : Indicateurs de dispersion

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

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

Garde les mêmes propriétés que la variance théorique

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

$$em = \frac{1}{n}\sum_i n_i \left| x_i - med \right|$$

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

	Variance	Ecart-type	em	Q3-Q1
Elève L	1,81	1,34	1	8
Elève P	35,48	5,96	4,85	5

Variables quantitatives : Variables centrées-réduites

On définit la série ***centrée-réduite*** de la façon suivante *:*

$$\tilde{x}_i = \left(\frac{x_i - \bar{x}}{s_x} \right)$$

La série est dite :

- centrée car de moyenne nulle
- réduite car de variance égale à 1

Démonstration

$$\bar{\tilde{x}} = \frac{1}{n}\sum_{k=1}^{n} \tilde{x}_k = \frac{1}{n}\sum_{k=1}^{n} \frac{x_k - \bar{x}}{s_x} = \frac{1}{s_x}\frac{1}{n}\sum_{k=1}^{n}(x_k - \bar{x}) = \frac{1}{s_x}\left(\frac{1}{n}\sum_{k=1}^{n} x_k - \bar{x}\right) = \frac{1}{s_x}(\bar{x} - \bar{x})$$

$$s^2_{\tilde{x}} = \frac{1}{n}\sum_{k=1}^{n}(\tilde{x}_k - \bar{\tilde{x}})^2 = \frac{1}{n}\sum_{k=1}^{n}\tilde{x}^2 = \frac{1}{n}\sum_{k=1}^{n}\left(\frac{x_k - \bar{x}}{s_x}\right)^2 = \frac{1}{s_x^2}\left(\frac{1}{n}\sum_{k=1}^{n}(x_k - \bar{x})^2\right) = \frac{1}{s_x^2}s_x^2 = 1$$

Pays	Taux de chômage	PIB
Allemagne	5,5	37430,1
Autriche	4,4	40064,8
Belgique	7,6	37727,8
Danemark	7,5	40189,9
Espagne	25,1	31903,8
Estonie	10,1	20393,3
…	…	…
Moyenne	**10,6**	**34851,6**
Ecart-type	**5,77**	**14203,93**

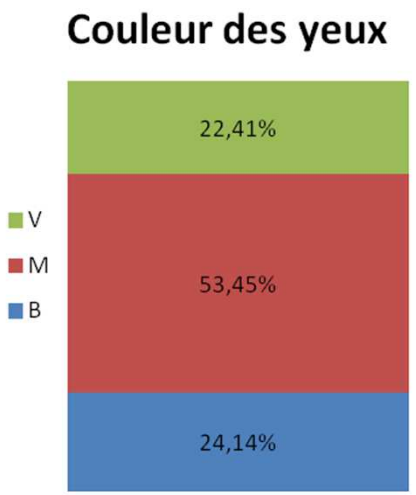
normalisation

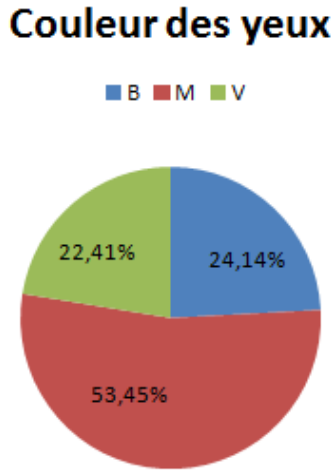
Pays	Taux de chômage	PIB
Allemagne	-0,88	0,18
Autriche	-1,08	0,37
Belgique	-0,52	0,20
Danemark	-0,54	0,38
Espagne	2,51	-0,21
Estonie	-0,09	-1,02
…	…	…
Moyenne	**0**	**0**
Ecart-type	**1**	**1**

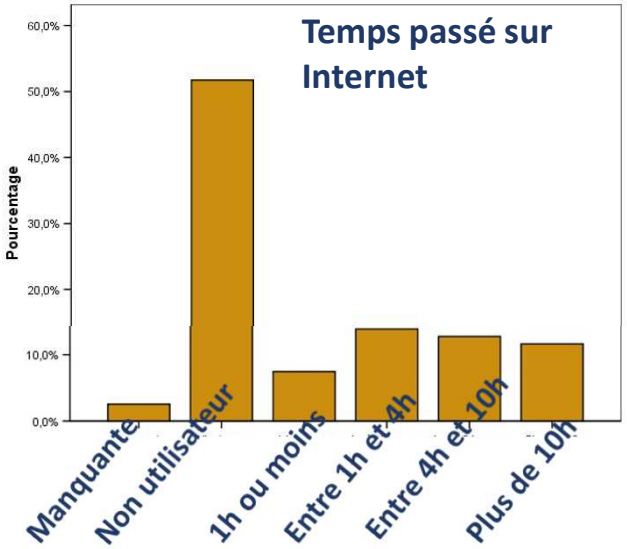
Variables qualitatives

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.

Représentation des variables nominales
diagramme en secteurs ou en barre

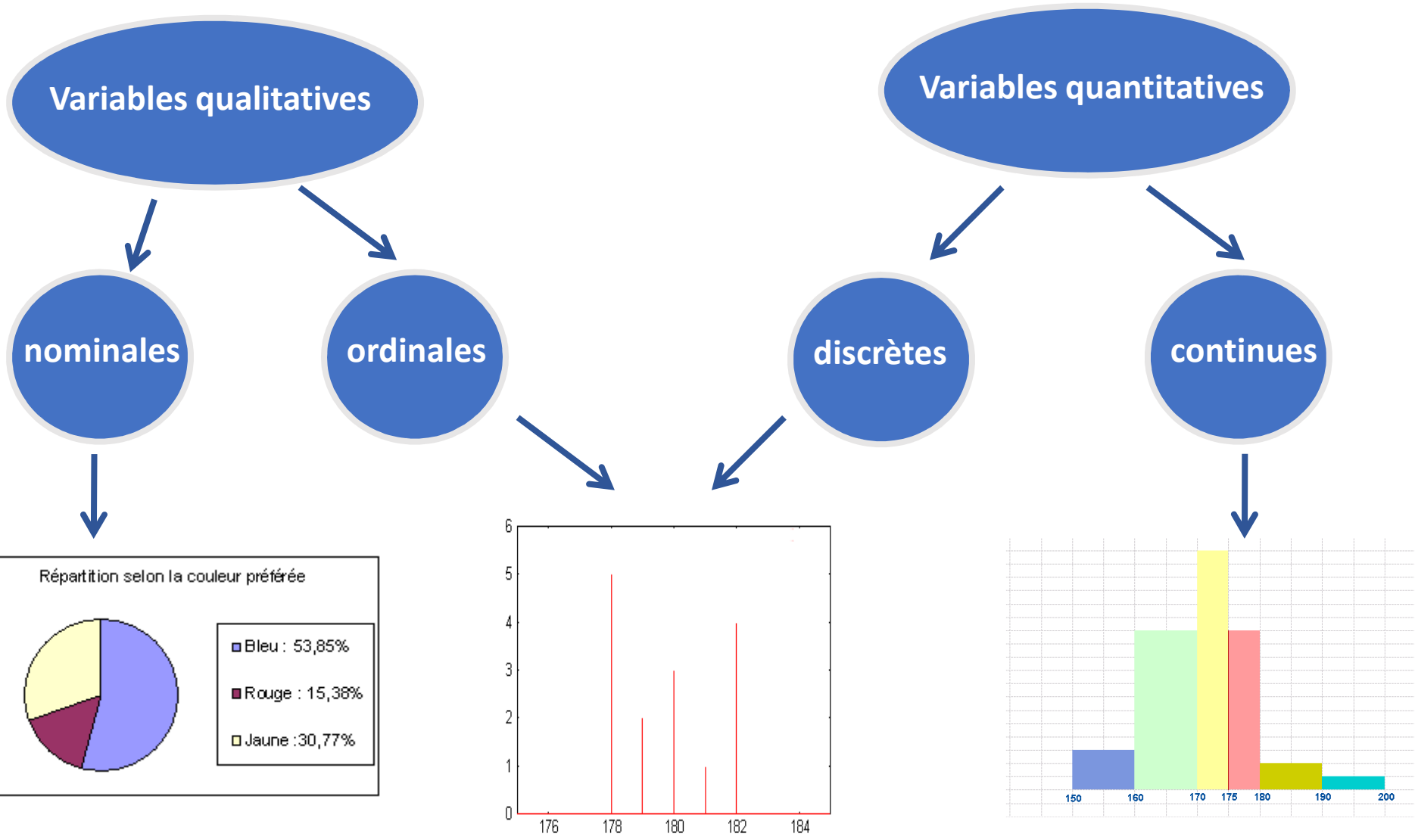






Représentation des variables ordinales
diagramme en bâtons

Résumé des représentations graphiques



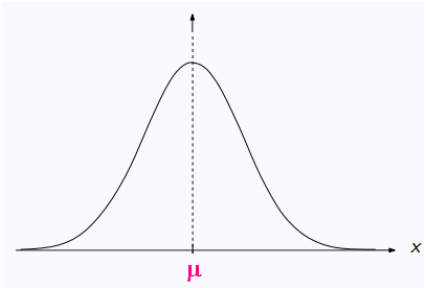
Quelques rappels

Probabilités		Statistiques	
Variable aléatoire	X	Série observée :	$\{x_i\}_{i=1,\dots,n}$
Espérance	$E(X)$	Moyenne	$\bar{x}$
Variance théorique	$V(X) = E[(X - E(X)^2] = E(X^2) - E(X)^2$	Variance empirique	$s_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i)^2 - \bar{x}^2$
Covariance théorique	$cov(X,Y) = E[(X - E(X)(Y - E(Y)] = E(XY) - E(X)E(Y)$	Covariance empirique	$c_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\bar{y}$

Soient X, Y et Z trois variables aléatoires et a et b deux constantes.

$E(aX + b) = aE(X) + b \qquad E(a) = a$
$V(aX + b) = a^2V(X) \qquad V(a) = 0$
$V(X + Y) = V(X) + V(Y) - 2cov(X,Y)$
$cov(X,Y) = cov(Y,X)$
$cov(aX,Y) = acov(X,Y)$
$cov(X + Z,Y) = cov(X,Y) + cov(Z,Y)$

Si X suit une loi normale $N(\mu, \sigma^2)$ alors $E(X) = \mu$, $V(x) = \sigma^2$ et



Une loi normale modélise un phénomène qui fluctue autour de sa moyenne avec des probabilités qui deviennent plus faibles au fur et à mesure qu'on s'en éloigne.