



Analyse univariée

(Rappels et compléments)

Source :
Cours Data Exploration (ING1)

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 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%

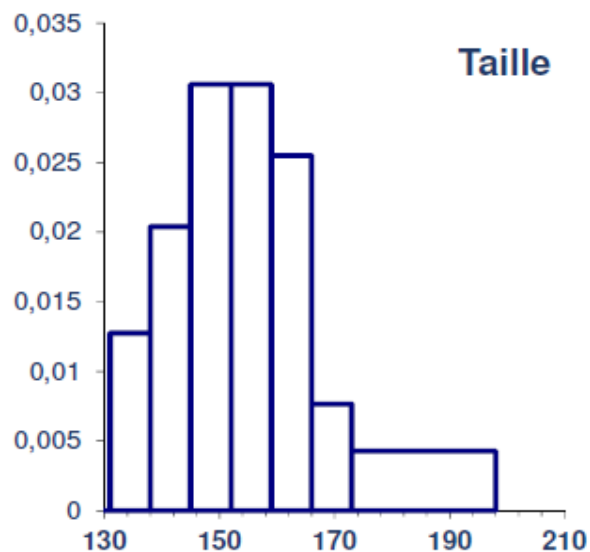
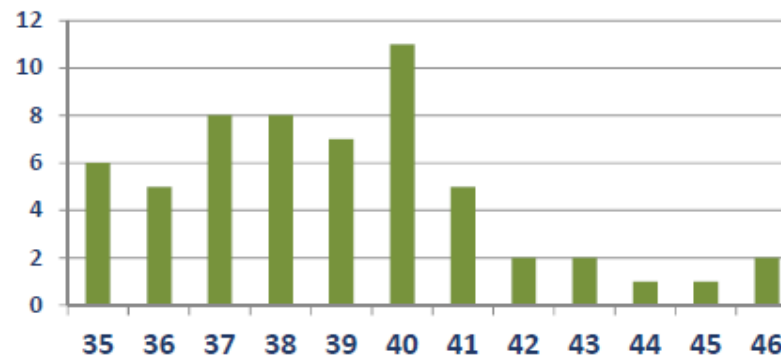
Variables quantitatives : Représentation graphique

Représentation graphique d'une variable discrète

Diagramme en bâtons

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

Pointures de chaussures



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 :

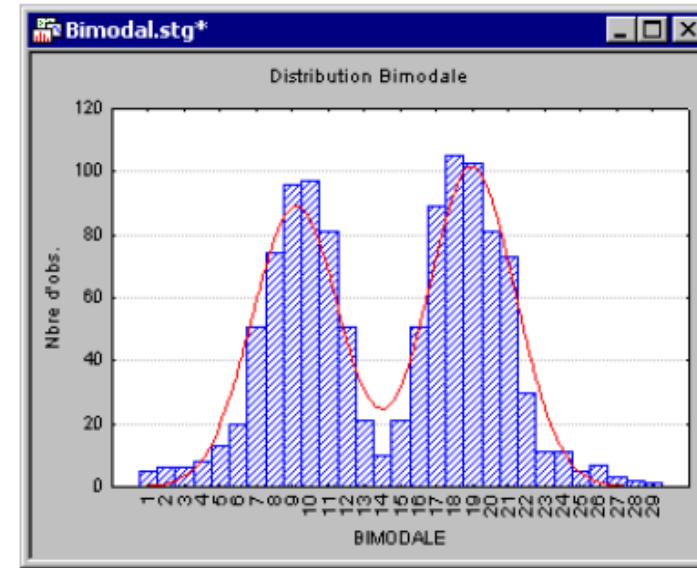
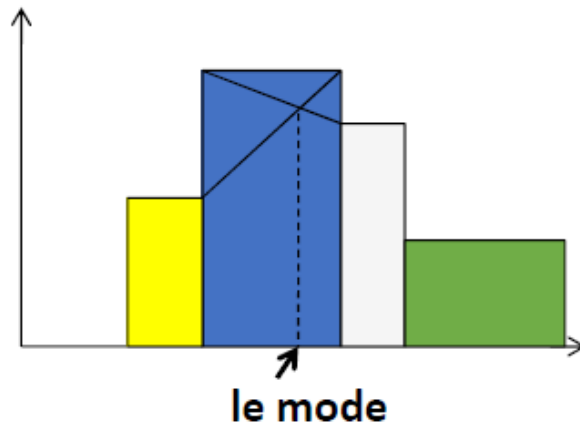
- Les **indicateurs de position** (moyenne, mode, médiane, quartiles). Ils positionnent la série des valeurs observées autour d'une tendance centrale.
- Les **indicateurs de dispersion** (variance, écart-type, étendue inter-quartile). Ils indiquent la fluctuation des valeurs de la série autour d'une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

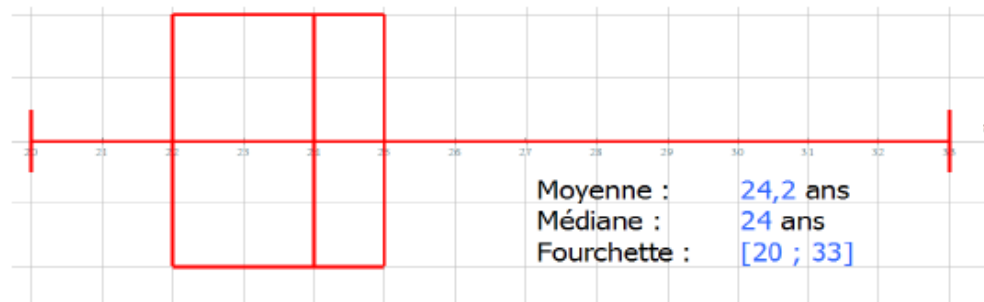
➤ Les quartiles :

- **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 **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 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

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

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

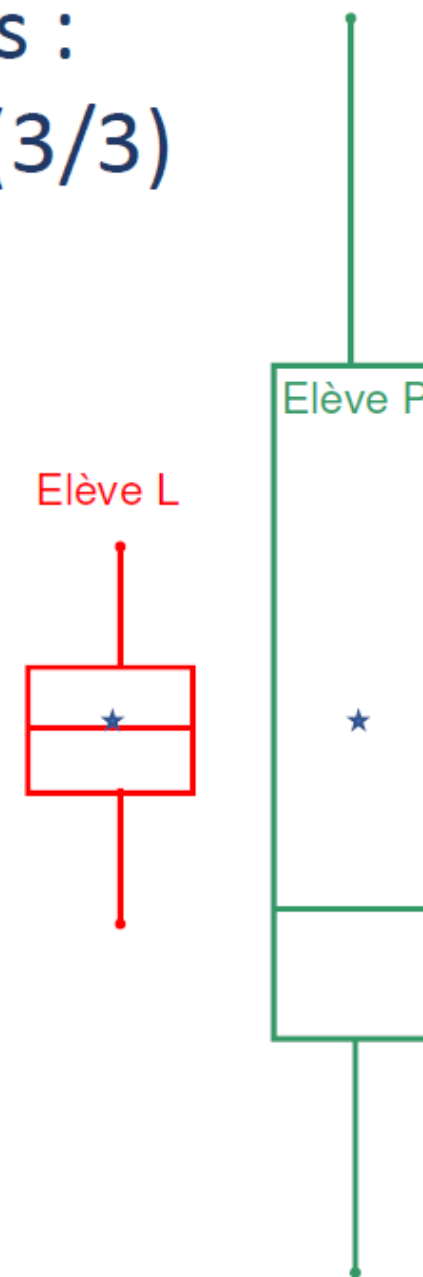
Variables quantitatives : Indicateurs de position (3/3)

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 \Rightarrow 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

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

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

$$\begin{aligned}\bar{\tilde{x}} &= \frac{1}{n} \sum_{k=1}^n \tilde{x}_k = \frac{1}{n} \sum_{k=1}^n \frac{x_k - \bar{x}}{s_x} = \frac{1}{s_x} \frac{1}{n} \sum_{k=1}^n (x_k - \bar{x}) = \frac{1}{s_x} \left(\frac{1}{n} \sum_{k=1}^n x_k - \bar{x} \right) = \frac{1}{s_x} (\bar{x} - \bar{x}) \\ s_{\tilde{x}}^2 &= \frac{1}{n} \sum_{k=1}^n (\tilde{x}_k - \bar{\tilde{x}})^2 = \frac{1}{n} \sum_{k=1}^n \tilde{x}_k^2 = \frac{1}{n} \sum_{k=1}^n \left(\frac{x_k - \bar{x}}{s_x} \right)^2 = \frac{1}{s_x^2} \left(\frac{1}{n} \sum_{k=1}^n (x_k - \bar{x})^2 \right) = \frac{1}{s_x^2} s_x^2 = 1\end{aligned}$$

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

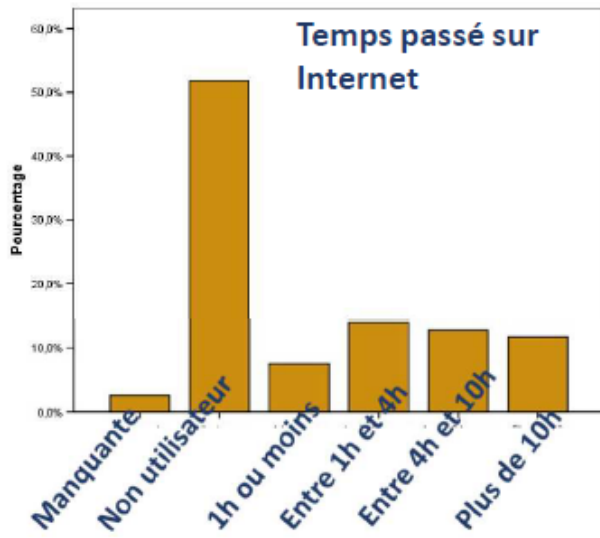


Pays	Taux de chômage	PIB
Allemagne	-0,88	0,18
Autriche	-1,08	0,37
Belgique	-0,52	0,20
Danemark	-0,54	0,38
Espagne	2,51	-0,21
Estonie	-0,09	-1,02
...
Moyenne	0	0
Ecart-type	1	1

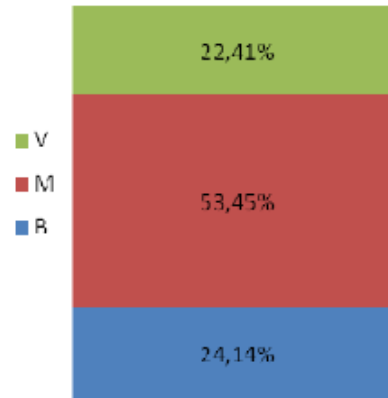
Variables qualitatives

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.

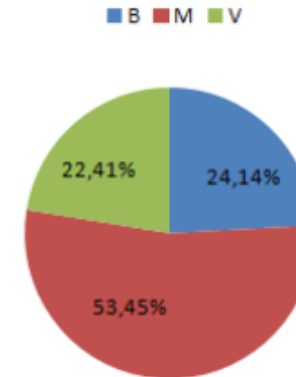
Représentation des variables nominales
diagramme en secteurs ou en barre



Couleur des yeux



Couleur des yeux



Représentation des variables ordinales
diagramme en bâtons

Résumé des représentations graphiques

