

R2.08 - Statistique descriptive Cours 1 - Description unidimensionnelle

A. Ridard

A propos de ce document

- Pour naviguer dans le document, vous pouvez utiliser :
 - le menu (en haut à gauche)
 - l'icône en dessous du logo IUT
 - les différents liens
- Pour signaler une erreur, vous pouvez envoyer un message à l'adresse suivante :
anthony.ridard@univ-ubs.fr

Plan du cours

- 1 Tableaux et graphiques
 - Variables qualitatives ou discrètes
 - Variables continues

- 2 Résumés numériques
 - Indicateurs de localisation
 - Indicateurs de dispersion
 - Quantiles
 - Indicateurs de forme

Dans l'exploration des données, une première étape consiste à étudier séparément chacune des variables, c'est la description unidimensionnelle.



Cette phase est indispensable mais évidemment insuffisante car elle ne tient pas compte des éventuelles relations entre les variables (cf. Cours 2).

On considère alors une variable X dont on possède n valeurs (données **brutes**) :

$$x_1, x_2, \dots, x_n$$

Sa description est faite sous forme de **tableaux**, **graphiques** et **résumés numériques**.



Cette phase exploratoire est facilitée par l'usage de l'informatique^a.

a. Nous utiliserons d'abord un tableur, puis nous exploiterons Python

1 Tableaux et graphiques

2 Résumés numériques

Leur présentation diffère selon la nature des variables.

- 1 Tableaux et graphiques
 - Variables qualitatives ou discrètes
 - Variables continues

- 2 Résumés numériques
 - Indicateurs de localisation
 - Indicateurs de dispersion
 - Quantiles
 - Indicateurs de forme

Pour chaque valeur ou modalité x_i de la variable, on note :

- n_i l'**effectif** de x_i c'est à dire son nombre d'occurrences dans l'échantillon
- $f_i = \frac{n_i}{n}$ la **fréquence** de x_i



Ne pas confondre

- Les données **brutes** (individu par individu) de la forme :

$$x_1, x_2, \dots, x_n$$

Dans ce cas, x_i peut coïncider avec x_j .

- Les données **regroupées par valeur** (qualitatives ou discrètes) de la forme :

x_1	x_2	...	x_p
n_1	n_2	...	n_p

Dans ce cas, x_i est différent de x_j dès que $i \neq j$!

Pour une variable **qualitative nominale**, on utilisera :

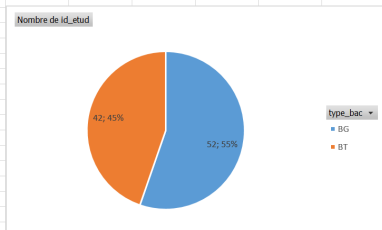
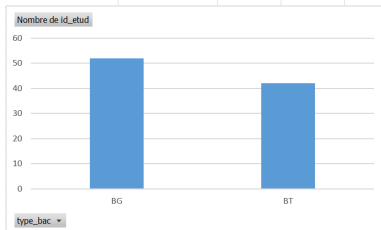
- un **diagramme en barres** (verticales ou horizontales) : chaque modalité est représentée par une barre de longueur proportionnelle à sa fréquence¹, l'épaisseur² étant sans importance
- un **diagramme circulaire**³ : chaque modalité est représentée par une portion de superficie proportionnelle à sa fréquence

Pour une variable **qualitative ordinale** ou **quantitative discrète**, on utilisera de préférence un **diagramme en barres verticales** en rangeant les modalités de la plus petite à la plus grande le long de l'axe horizontal.

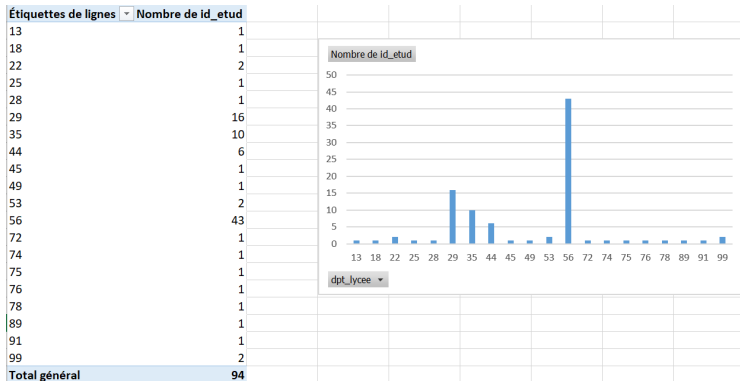
-
1. Au sens large : les n_i sont aussi appelés fréquences absolues, et les f_i fréquences relatives
 2. Lorsque les barres sont des bâtons, on parle de **diagramme en bâtons**
 3. Ou encore camembert (en anglais **pie-chart**)

Une variable qualitative nominale :

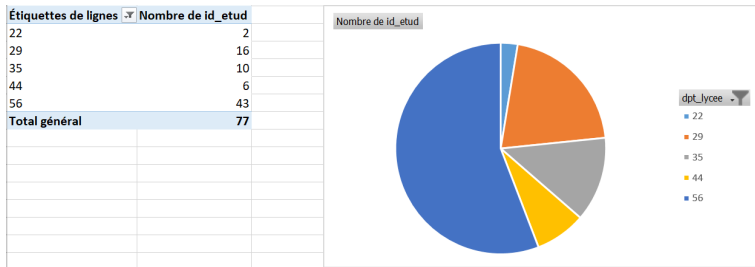
Étiquettes de lignes	Nombre de id_etud
BG	52
BT	42
Total général	94



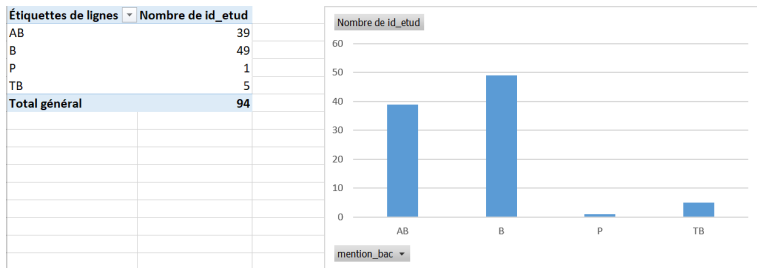
Une variable qualitative nominale "qui se prend pour" une variable quantitative :



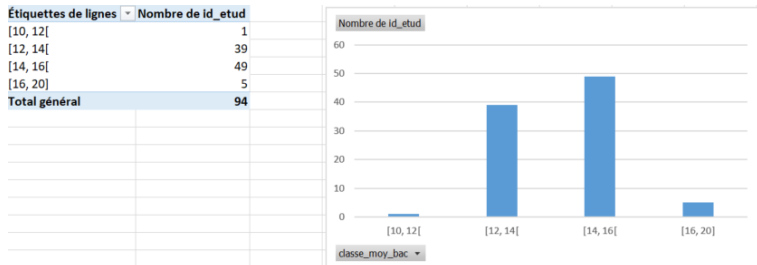
Pour cette variable qualitative nominale (filtrée), on préférera :



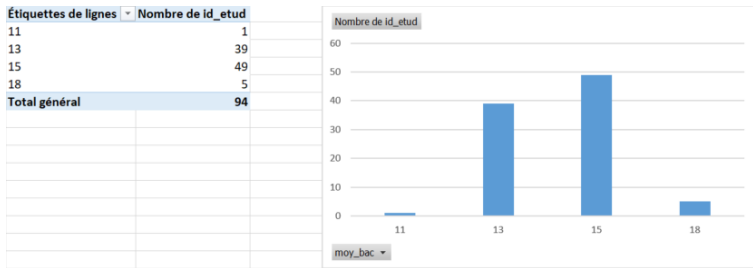
Une variable qualitative ordinale dont l'ordre n'est pas bien géré par l'outil :



Pour cette information, on "calculera" une nouvelle variable qualitative ordinale :

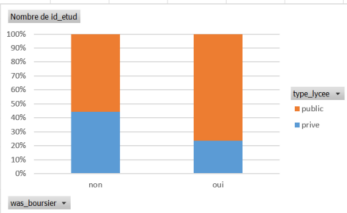
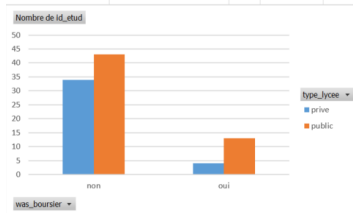


Une variable quantitative discrète porteuse de la même information, mais approchée :



Pour comparer deux sous-populations⁴ selon une variable qualitative nominale :

Nombre de id_etud		Étiquettes de colonnes	
Étiquettes de lignes	privé	public	Total général
non	34	43	77
oui	4	13	17
Total général	38	56	94



- 1 Tableaux et graphiques
 - Variables qualitatives ou discrètes
 - Variables continues

- 2 Résumés numériques
 - Indicateurs de localisation
 - Indicateurs de dispersion
 - Quantiles
 - Indicateurs de forme

Les var. continues peuvent être **discrétisées** en **regroupant les valeurs par classe** :

$[e_1, e_2[$	$[e_2, e_3[$...	$[e_p, e_{p+1}[$
n_1	n_2	...	n_p

Pour une variable **continue discrétisée**, on utilisera un **histogramme** : chaque classe est représentée par une barre de superficie proportionnelle à sa fréquence.



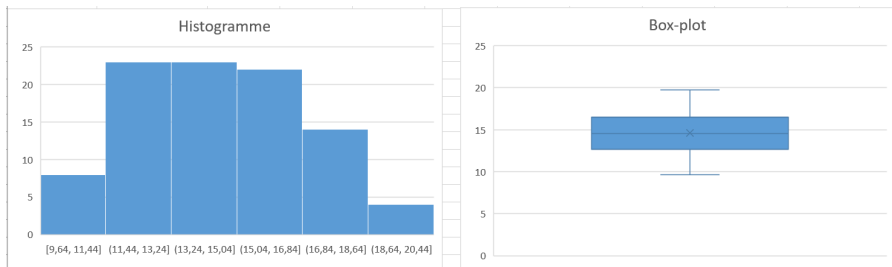
- Si toutes les classes ont la même amplitude^a, chaque classe est alors représentée par une barre de hauteur proportionnelle à sa fréquence.
- Ne pas confondre histogramme et diagramme en barres.

a. C'est le plus simple à mettre en œuvre, mais pas toujours le plus pertinent.



- La détermination du nombre de classes d'un histogramme est délicate, il existe des formules mais pas de règle absolue.
- Un trop faible nombre de classes fait perdre de l'information et aboutit à gommer les différences pouvant exister. En revanche, un trop grand nombre de classes "brouille" l'information.

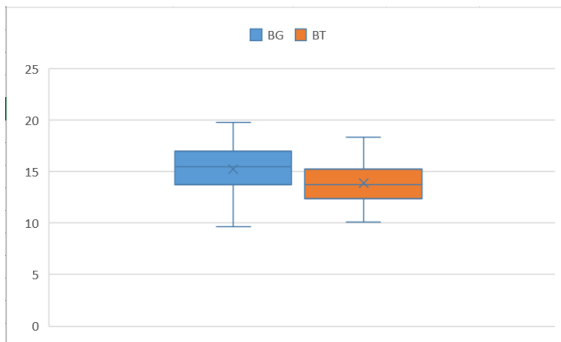
Une variable quantitative continue (lv_1) :



Un box-plot^a représente graphiquement des indicateurs numériques que l'on (re)verra plus tard comme les quartiles par exemple.

a. Boîte à moustaches

Pour comparer deux sous-populations⁵ selon une var. quantitative continue (lv_1) :



- 1 Tableaux et graphiques
- 2 Résumés numériques

Dans cette section, la variable X est **quantitative**.

- 1 Tableaux et graphiques
 - Variables qualitatives ou discrètes
 - Variables continues

- 2 Résumés numériques
 - Indicateurs de localisation
 - Indicateurs de dispersion
 - Quantiles
 - Indicateurs de forme

On cherche à résumer **au mieux** les observations x_1, \dots, x_n avec un seul nombre c .

Pour quantifier le « au mieux », on va utiliser l'erreur moyenne :

$$e = \frac{1}{n} \sum_{i=1}^n d(x_i, c)$$

où d mesure l'écart entre deux valeurs.

L'objectif est donc de trouver le nombre c qui minimise l'erreur moyenne e .

Si on mesure l'écart entre deux valeurs de la manière suivante :

$$d(x_i, c) = (x_i - c)^2$$

$e = \frac{1}{n} \sum_{i=1}^n (x_i - c)^2$ est alors appelée l'**erreur quadratique moyenne**.

Dans ce cas, le minimum est réalisé⁶ pour :

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

On reconnaît la moyenne (arithmétique) des observations.
Elle est appelée **moyenne empirique** de l'échantillon et notée \bar{x} .

6. cf. exercice suivant pour une démonstration



Considérons la fonction f définie par :

$$\begin{aligned} f: \mathbb{R} &\longrightarrow \mathbb{R} \\ c &\longmapsto \frac{1}{n} \sum_{i=1}^n (x_i - c)^2 \end{aligned}$$

- 1 Calculer $f'(c)$.
- 2 En déduire le tableau de variations de f .
- 3 Conclure.



Cet indicateur est largement utilisé, il possède de bonnes propriétés algébriques, mais il est très sensible aux valeurs extrêmes et donc assez peu robuste.

Si on mesure l'écart entre deux valeurs de la manière suivante :

$$d(x_i, c) = |x_i - c|$$

$e = \frac{1}{n} \sum_{i=1}^n |x_i - c|$ est alors appelée l'**erreur absolue moyenne**.

Dans ce cas, le minimum est réalisé pour c vérifiant :

$$\text{Card}(\{i \in \llbracket 1, n \rrbracket \mid x_i < c\}) = \text{Card}(\{i \in \llbracket 1, n \rrbracket \mid x_i > c\})$$

Il s'agit de la **médiane** des observations notée M ou encore⁷ Q_2 .

7. Q_2 désigne le deuxième quartile qui sera défini plus tard

En triant dans l'ordre croissant l'échantillon formé de x_1, \dots, x_n , on obtient :

$$x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$$

On appelle $x_{(i)}$ la i -ème **statistique d'ordre**.

Selon la parité de n , on peut alors définir la médiane de la manière suivante :

$$M = \begin{cases} x_{(k+1)} & \text{si } n = 2k + 1 \\ \frac{x_{(k+1)} - x_{(k)}}{2} & \text{si } n = 2k \end{cases}$$

En notant j la partie entière de $\frac{n-1}{2}$ et d sa partie décimale, on a aussi⁸ :

$$M = x_{(j+1)} + d \times (x_{(j+2)} - x_{(j+1)})$$



Il existe d'autres définitions selon le choix de la valeur entre $x_{(k)}$ et $x_{(k+1)}$

8. Cette formule unique est implémentée dans Excel et Python (par défaut) et facilement généralisable



Calculer la médiane de :

① 1, 1, 2, 2, 2

② 1, 2, 5, 100



Contrairement à la moyenne, la médiane est un indicateur de localisation insensible aux valeurs extrêmes et donc robuste, mais sans propriété algébrique.

- 1 Tableaux et graphiques
 - Variables qualitatives ou discrètes
 - Variables continues

- 2 Résumés numériques
 - Indicateurs de localisation
 - **Indicateurs de dispersion**
 - Quantiles
 - Indicateurs de forme



Un indicateur de localisation doit toujours être accompagné d'un indicateur de dispersion mesurant la **variabilité** des données.

Si on utilise la moyenne, on complètera avec l'erreur quadratique moyenne en \bar{x} :

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

Elle est appelée la **variance empirique** de l'échantillon et notée s^2 .



- Pour faciliter l'interprétation de cette mesure, on préfère utiliser sa racine carrée qui s'exprime dans la même dimension que les données. Elle est appelée **écart-type empirique** de l'échantillon et noté s .
- Le rapport $\frac{s}{\bar{x}}$ renseigne^a sur la significativité de la moyenne.
- Pour éviter son « biais^b », on utilise la **variance empirique corrigée** :

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

Ou encore l'**écart-type empirique corrigé** $s^* = \sqrt{s^{*2}}$

-
- a. S'il est inférieur à 10%, la moyenne est significative compte tenu de la très faible variabilité
b. Cette notion dépasse le cadre de ce cours, mais sera abordée en Statistique inférentielle

Si on utilise la médiane, on complètera avec l'**écart inter-quartile** $Q_3 - Q_1$ contenant la « moitié centrale » des observations.

Pour définir le **premier** et le **troisième quartile**, on généralise la formule de la médiane de manière à partager l'échantillon, non plus en deux, mais en quatre.

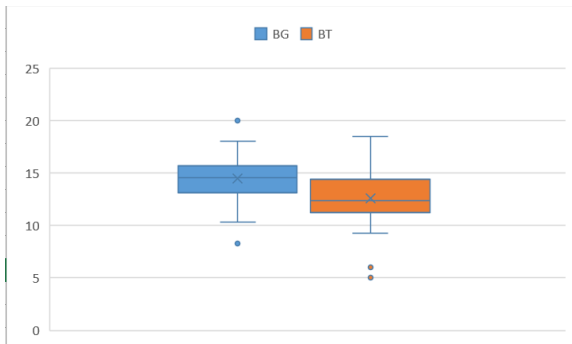
En notant j la partie entière de $\frac{1}{4}(n-1)$ et d sa partie décimale, on définit :

$$Q_1 = x_{(j+1)} + d \times (x_{(j+2)} - x_{(j+1)})$$

En notant j la partie entière de $\frac{3}{4}(n-1)$ et d sa partie décimale, on définit :

$$Q_3 = x_{(j+1)} + d \times (x_{(j+2)} - x_{(j+1)})$$

La médiane et l'écart inter-quartile se représentent graphiquement à l'aide d'une boîte à moustaches (ou box-plot) :





- La boîte est limitée par Q_1 et Q_3 , sa longueur est l'écart inter-quartile
- Elle est coupée en deux par la médiane Q_2
- L'extrémité inférieure de la moustache est définie par :

$$\max\left\{x_{(1)}, Q_1 - \frac{3}{2}(Q_3 - Q_1)\right\}$$

- L'extrémité supérieure de la moustache est définie par :

$$\min\left\{x_{(n)}, Q_3 + \frac{3}{2}(Q_3 - Q_1)\right\}$$

- Les valeurs (points) qui « s'écartent trop » sont dites **aberrantes**

- 1 Tableaux et graphiques
 - Variables qualitatives ou discrètes
 - Variables continues

- 2 Résumés numériques
 - Indicateurs de localisation
 - Indicateurs de dispersion
 - **Quantiles**
 - Indicateurs de forme

Les quartiles se généralisent par les quantiles d'ordre p avec $0 < p < 1$.

En notant j la partie entière de $p(n-1)$ et d sa partie décimale, on définit :

$$\tilde{Q}_p = x_{(j+1)} + d \times (x_{(j+2)} - x_{(j+1)})$$



- $M = \tilde{Q}_{0.5}$
- $Q_1 = \tilde{Q}_{0.25}$ et $Q_3 = \tilde{Q}_{0.75}$
- En partageant l'échantillon en 10, on obtient les **déciles** $\tilde{Q}_{0.1}, \dots, \tilde{Q}_{0.9}$
- En partageant l'échantillon en 100, on obtient les **centiles** $\tilde{Q}_{0.01}, \dots, \tilde{Q}_{0.99}$
- Plus le découpage est fin, plus le résumé est riche !

- 1 Tableaux et graphiques
 - Variables qualitatives ou discrètes
 - Variables continues
- 2 Résumés numériques
 - Indicateurs de localisation
 - Indicateurs de dispersion
 - Quantiles
 - Indicateurs de forme

On peut enfin résumer l'asymétrie et l'applatissement avec les indicateurs suivants :

- Le **coefficient d'asymétrie** : $\gamma_3 = \frac{m_3}{s^3}$
- Le **coefficient d'applatissement** : $\gamma_4 = \frac{m_4}{s^4}$

$$\text{où } m_3 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^3 \text{ et } m_4 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^4$$



- m_3 et m_4 sont appelés **moments centrés** d'ordre 3 et 4
- γ_3 et γ_4 peuvent servir à « tester » le caractère gaussien d'un échantillon