Analyse univariée (partie 2)

Vincent Audigier
vincent.audigier@lecnam.net

CNAM, Paris

STA101 2019-2020

Introduction

- Une série statistique peut être présentée par des tableaux ou des représentations graphiques
- La façon de présenter cette série dépend de la nature de la variable
- Ces représentations sont riches, mais peu pratiques quand le nombre de variables devient grand
- Besoin d'une description plus synthétique

Indicateurs

Pour une variables quantitative X, on complète la description de la série par des résumés numériques :

- indicateurs de tendance centrale (ou de position)
- indicateurs de dispersion
- indicateurs de forme

Plan

Indicateur de tendance centrale

- Pour une variable quantitative X, la description de la série de données $x_1, \ldots, x_n$ peut être complétée par des indicateurs de tendance centrale

- Pour définir un indicateur :
  - Choisir une mesure d'erreur locale

$$d(x_i, c)$$

  - Puis choisir un critère global

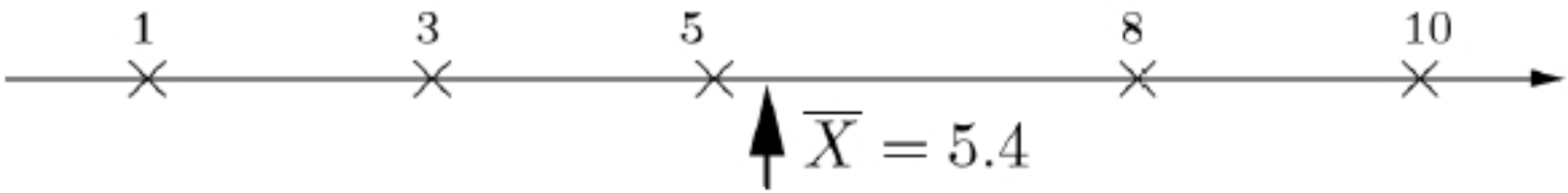
$$J(c) = \frac{1}{n} \sum_{i=1}^{n} d(x_i, c)$$

  - Minimiser ce critère

Moyenne empirique

$$d(x_i, c) = (x_i - c)^2 \qquad \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- Géométriquement : $\bar{x}$ est le centre de gravité des n points définis par la série, affecté du même poids $1/n$



- Sensibilité aux valeurs aberrantes[1]

[1] Une donnée sera considérée comme *aberrante* si elle n'est pas issue de la même distribution que celle qui tient pour la majorité des données

Moyenne empirique

On peut généraliser la moyenne empirique au cas où les individus ont des poids différents

Soit $p_1, p_2, ..., p_n$ le poids de chaque individu, alors la moyenne empirique s'écrit

$$\sum_{i=1}^{n} p_i x_i$$

propriété de linéarité de la moyenne empirique : pour $a, b, \in \mathbb{R}$

$$\overline{ax + b} = a\bar{x} + b$$

Médiane empirique

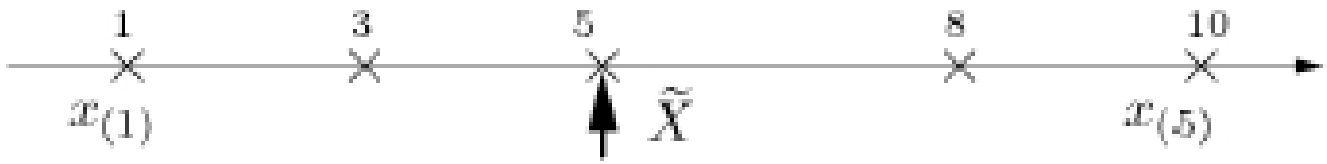
$$d(x_i, c) = |x_i - c|$$

La médiane empirique, notée $\tilde{x}$, est définie comme un réel partageant la série en deux groupes de même effectif

- Si n est impair, alors la médiane empirique est l'observation au centre de la série ordonnée

- Si n est pair, alors on peut choisir le milieu de l'intervalle $\left]x_{(n/2)}; x_{(n/2+1)}\right[$

$$\tilde{X} = \frac{x_{(n/2)} + x_{(n/2+1)}}{2}$$



- La moitié des observations sont supérieures (inférieures) à $\tilde{x}$

- Robuste aux valeurs aberrantes

Quantiles empiriques

On appelle **quantiles empiriques** les valeurs partageant la série ordonnée en un certain nombre de parties de même effectif

- en deux parties : médiane (ou Q(1/2))
- en 4 parties : les quartiles, notés Q(1/4), Q(1/2), Q(3/4)
- en dix parties : les déciles, notés Q(1/10), Q(2/10),..., Q(9/10)
- en cent parties : les percentiles, notés Q(1/100), Q(2/100),..., Q(9/100)

Pour $0 < q < 1$, le **quantile empirique** d'ordre q, noté $Q(q)$, est défini par

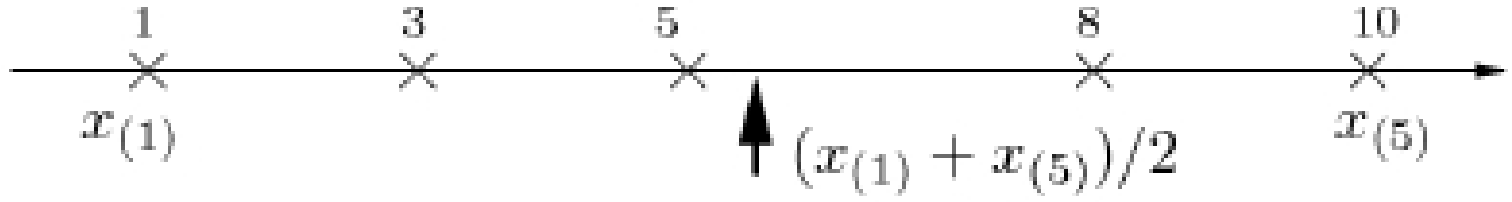
$$Q(q) = \begin{cases} \frac{x_{(nq)} + x_{(nq+1)}}{2} & \text{si } nq \text{ est un entier} \\ x_{([nq]+1)} & \text{sinon} \end{cases}$$

Moyenne des valeurs extrêmes

Les valeurs $x_{(1)}$ et $x_{(n)}$ sont respectivement l'observation extrême inférieure et supérieure de l'échantillon.

$$J(c) = sup_{1 \leq i \leq n} |x_i - c| \qquad \frac{x_{(1)} + x_{(n)}}{2}$$

- Géométriquement : point au milieu du domaine de variation des observations



- Sensible aux valeurs aberrantes

Plan

Variance empirique

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

- Propriétés : $s_X^2 = \overline{x^2} - \overline{x}^2 \qquad s_{aX+b}^2 = a^2 s_X^2$
- $s_X = \sqrt{s_X^2}$ est appelé écart-type (empirique)
- limite : la variance doit toujours se comparer à la valeur moyenne : une variance de 10 pour $\overline{x} = 12$ ou $\overline{x} = 200$ n'a pas la même signification.

Exemple

	maxO3	T9	T12	T15	Ne9	Ne12	Ne15	Vx9	Vx12	Vx15	maxO3v
écart-type	28.2	3.1	4.0	4.5	**2.6**	2.3	2.3	**2.6**	2.8	2.8	28.3
moyenne	90.3	18.4	21.5	22.6	**4.9**	5.0	4.8	**-1.2**	-1.6	-1.7	90.6

Coefficient de variation empirique / étendue

Le **coefficient de variation empirique**

$$CV = \frac{s_X}{\overline{x}}$$

- Il s'agit d'un indicateur sans dimension
- En pratique, on utilise le seuil de 0.15 pour établir une faible/forte variabilité

	maxO3	T9	T12	T15	Ne9	Ne12	Ne15	Vx9	Vx12	Vx15	maxO3v
écart-type	28.2	3.1	4.0	4.5	**2.6**	2.3	2.3	**2.6**	2.8	2.8	28.3
moyenne	90.3	18.4	21.5	22.6	**4.9**	5.0	4.8	**-1.2**	-1.6	-1.7	90.6
CV	0.3	0.2	0.2	0.2	**0.5**	0.5	0.5	**-2.2**	-1.7	-1.7	0.3

Etendue

L'**étendue** est la différence des valeurs extrêmes $E = x_{(n)} - x_{(1)}$

- c'est un indicateur très sensible aux valeurs aberrantes
- utilisé en contrôle qualité pour détecter les valeurs aberrantes

Distance inter-quartile

$$Q(3/4) - Q(1/4)$$

- Elle est robuste vis-a-vis des valeurs aberrantes
- Exemple :

min	$Q(1/4)$	$Q(1/2)$	$\bar{x}$	$Q(3/4)$	max	$Q(3/4) - Q(1/4)$
42.00	70.75	81.50	90.30	106.00	166.00	35.25

Table: Indicateurs pour la variable maxO3

Représentation graphique

La boîte à moustaches ou “boxplot”

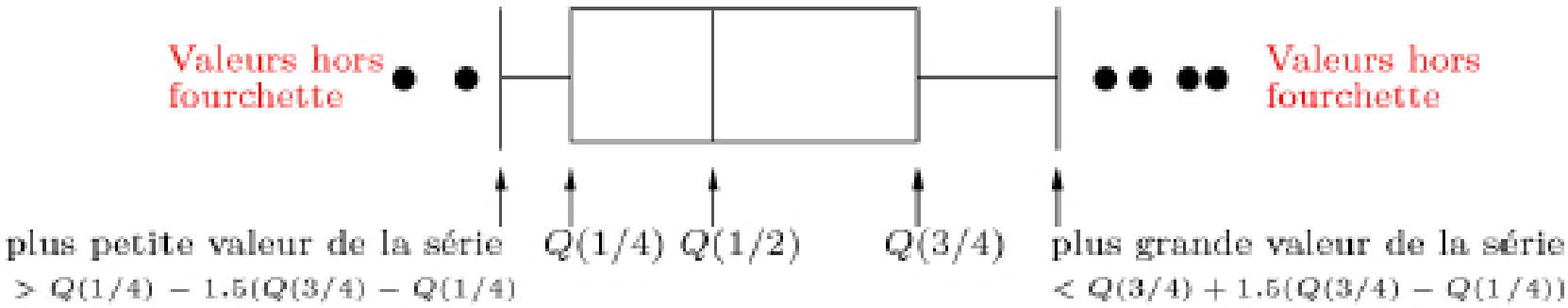
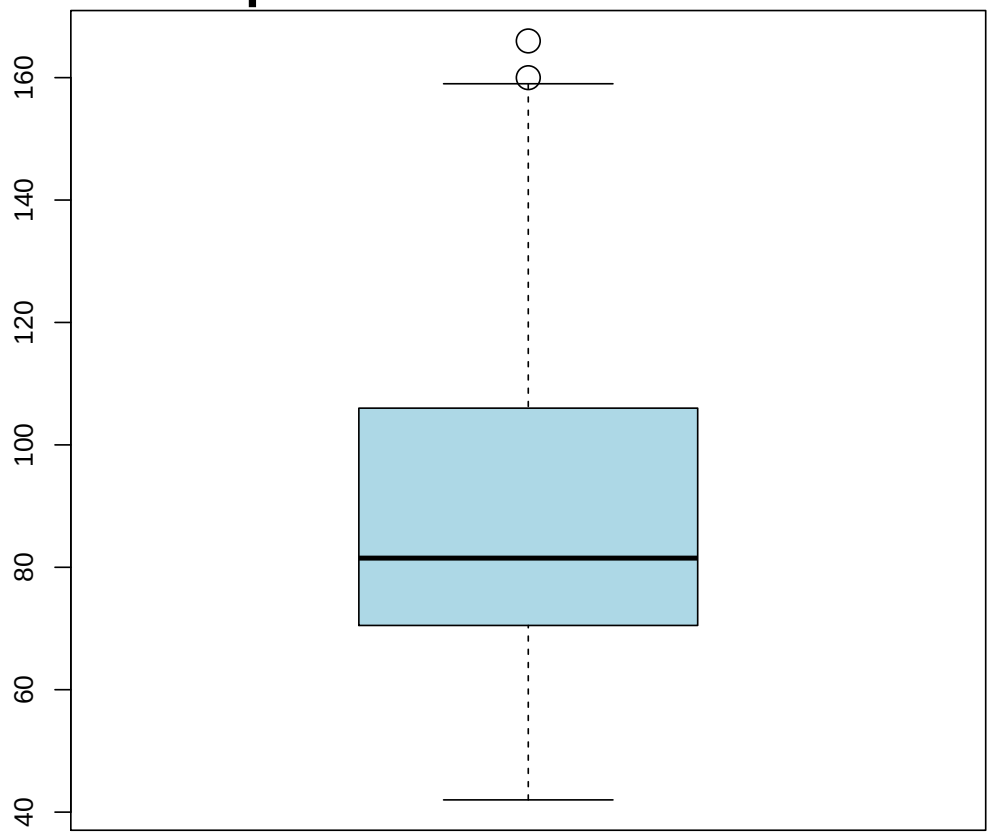


Figure: Principe de construction de la boîte à moustaches

La boîte correspond à l’intervalle de valeurs $[Q(1/4), Q(3/4)]$. On y adjoint la valeur de la médiane Q(1/2) (et parfois aussi celle de la moyenne)

Boîte à moustaches pour la variable maxO3



Centrage-réduction

Le centrage réduction d'une série est très pratique pour situer une valeur parmi les autres.

$$z_i = \frac{x_i - \bar{x}}{s_x}$$

- le centrage permet de voir directement si une valeur est supérieure ou inférieure à la moyenne
- la réduction permet de voir rapidement les valeurs au-delà de deux écart-type et de comparer les valeurs sur des variables différentes

	maxO3	T9	T12	T15	maxO3	T9	T12	T15
20010601	87.00	15.60	18.50	18.40	-0.12	-0.88	-0.75	-0.93
20010602	82.00	17.00	18.40	17.70	-0.29	-0.44	-0.77	-1.09
20010603	92.00	15.30	17.60	19.50	0.06	-0.98	-0.97	-0.69
20010604	114.00	16.20	19.70	22.50	0.84	-0.69	-0.45	-0.03
20010605	94.00	17.40	20.50	20.40	0.13	-0.31	-0.25	-0.49
20010606	80.00	17.70	19.80	18.30	-0.37	-0.21	-0.43	-0.96
moyenne	90.30	18.36	21.53	22.63	0	0	0	0
ecart-type	28.19	3.12	4.04	4.53	1	1	1	1

Table: extrait de ozone : données brutes (à gauche) et centrées réduites (à droite)

Eléments d'interprétation

- Inégalité de Bienaymé-Tchebychev

$$P(|z| \geq 2) \leq 25\%$$

- Pour une loi normale centrée-réduite

$$P(|z| > 1.96) = 5\%$$

- Les valeurs centrées-réduites supérieures à 2 sont remarquables

Plan

Coefficient d'asymétrie

- Le **coefficient d'asymétrie** de Fisher

$$\gamma_1 = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{x_i - \overline{x}}{s_X} \right)^3$$

  - 0 pour une série dont la répartition est symétrique (e.g. loi normale)
  - > 0 si la queue de la distribution est à droite (e.g. loi exponentielle)
  - < 0 si la queue de la distribution est à gauche

- Sensible aux valeurs aberrantes

Coefficient d'aplatissement

- Le **coefficient d'aplatissement** de Fisher

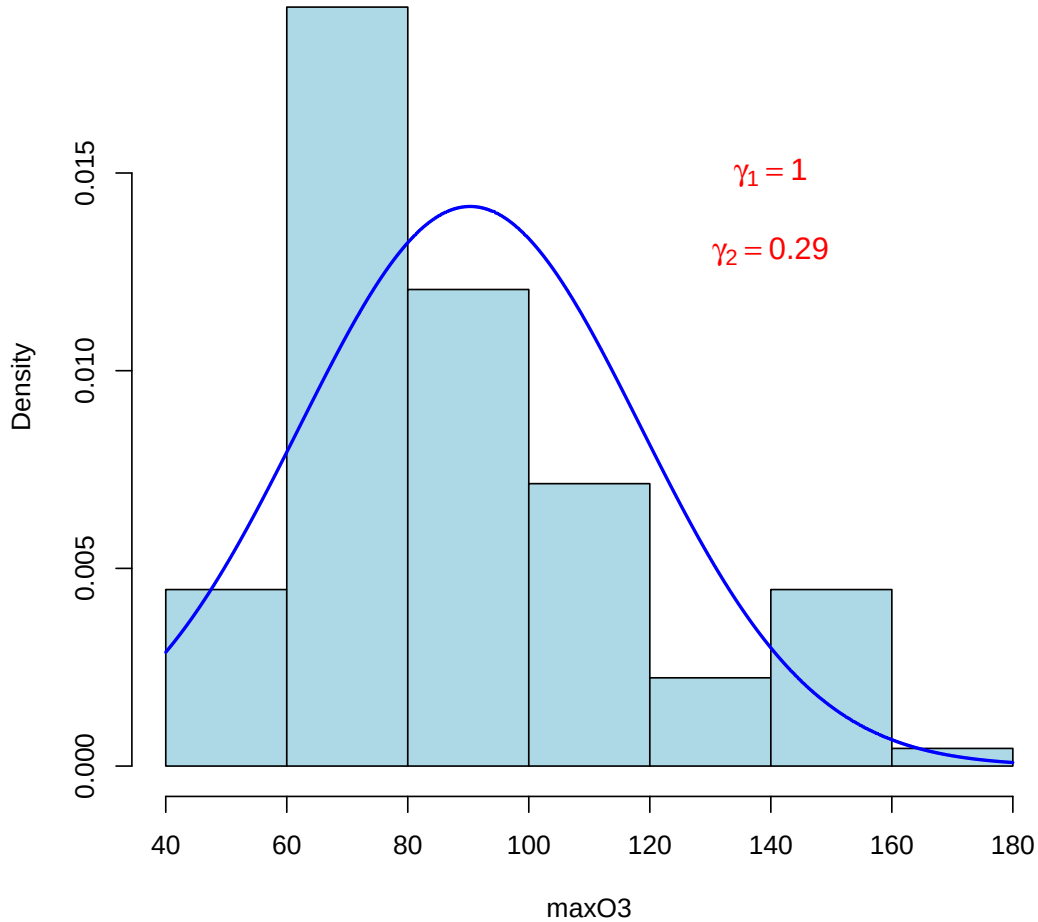
$$\gamma_2 = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \overline{x}}{s_X}\right)^4 - 3$$

  - Sert à comparer la concentration des valeurs à celle d'une loi normale centrée pour laquelle ce coefficient vaut 3
  - > 0 signifie distribution plus concentrée ou plus pointue que la Gaussienne
  - < 0 signifie distribution plus aplatie que la Gaussienne

- Sensible aux valeurs aberrantes

Exemple

	γ_1	γ_2
maxO3	1.00	0.29
T9	0.62	0.18
T12	0.78	0.21
T15	0.59	-0.26
Ne9	-0.50	-1.10
Ne12	-0.68	-0.48
Ne15	-0.49	-0.76
Vx9	-0.17	-0.54
Vx12	0.48	0.18
Vx15	0.08	-0.22
maxO3v	0.97	0.23



Conclusion

- L'analyse univariée permet d'apprécier les caractéristiques de chacune des variables en les résumant
- Elle s'effectue différemment selon la nature des variables
- Cette analyse peut être effectuée sous forme de tableaux, graphiques ou indicateurs
- Elle est indispensable pour pouvoir repérer d'éventuelles anomalies dans les données ou identifier des valeurs particulières