Statistiques descriptives de mono variable

Benoit Gaüzère, Stéphane Canu
benoit.gauzere@insa-rouen.fr

INSA Rouen Normandie - ITI

30 janvier 2024

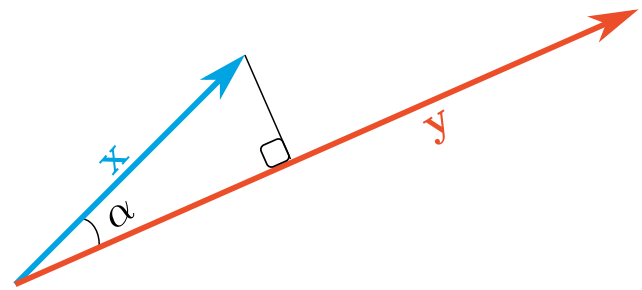
Plan

Rappels : espace vectoriel $\mathbb{R}^n$

vecteur $\mathbf{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$ transposé : $\mathbf{x}^\top = (x_1, \ldots, x_n)$

norme $\|\mathbf{x}\|^2 = \sum_{i=1}^{n} x_i^2 = \mathbf{x}^\top \mathbf{x}$

produit scalaire $\mathbf{x}^\top \mathbf{y} = \sum_{i=1}^{n} x_i y_i$



cosinus $\cos(\alpha) = \dfrac{\mathbf{x}^\top \mathbf{y}}{\|\mathbf{x}\| \|\mathbf{y}\|}$

projection $P_{\mathbf{y}}(\mathbf{x}) = \underbrace{\dfrac{\mathbf{x}^\top \mathbf{y}}{\|\mathbf{y}\|} \dfrac{\mathbf{y}}{\|\mathbf{y}\|}}_{\mathbf{y}^\top \mathbf{y}} = \|\mathbf{x}\| \cos\alpha \dfrac{\mathbf{y}}{\|\mathbf{y}\|}$

Quel est le meilleur résumé d'une variable quantitative ?

- $S_5 = \{1, 5, 2, 4, 3\}$ $\quad$ $\mathbf{x} = (1, 5, 2, 4, 3)^\top \in \mathbb{R}^5$

Quelle est la valeur c qui **résume** au mieux les observations ?

Nous pouvons poser un principe variationnel :

$$\min_{c \in \mathbb{R}} J(c) \quad \text{avec :}$$

$$J(c) \quad = \quad \sum_{i=1}^{n} (x_i - c)^2 = \|\mathbf{x} - c\,\mathbb{1}\|^2$$

$$\Leftrightarrow \min_{c \in \mathbb{R}} \underbrace{(1-c)^2 + (5-c)^2 + (2-c)^2 + (4-c)^2 + (3-c)^2}_{J_5(c) = 5c^2 - 30c + 55}$$

Quel est le meilleur résumé d'une variable quantitative ?

- $S_5 = \{1, 5, 2, 4, 3\}$ $\quad \mathbf{x} = (1, 5, 2, 4, 3)^\top \in \mathbb{R}^5$

Quelle est la valeur c qui **résume** au mieux les observations ?

Nous pouvons poser un principe variationnel :

$$\min_{c \in \mathbb{R}} J(c) \quad \text{avec :}$$

$$J(c) = \sum_{i=1}^{n} (x_i - c)^2 = \|\mathbf{x} - c\,\mathbb{1}\|^2$$

$$\Leftrightarrow \min_{c \in \mathbb{R}} \underbrace{(1-c)^2 + (5-c)^2 + (2-c)^2 + (4-c)^2 + (3-c)^2}_{J_5(c) = 5c^2 - 30c + 55}$$

- $S_6 = \{1, 5, 2, 4, 3, 100\}$ $\quad J_6(c) = 6c^2 - 230c + 10055$

Quel est le meilleur résumé ?

Résumer S_n par $c \in \mathbb{R}$

$$S_n = \{x_1, x_2, \ldots, x_i, \ldots, x_n\}$$

Solution

$$\min_{c \in \mathbb{R}} J(c) \Leftrightarrow \frac{dJ(c)}{dc} = 0 \Leftrightarrow 2\sum_{i=1}^{n}(x_i - c) = 0 \Leftrightarrow c = \frac{1}{n}\sum_{i=1}^{n} x_i$$

C'est la moyenne empirique : $c = \frac{1}{n}\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} \frac{1}{n} x_i = \sum_{i=1}^{n} \hat{f}_i x_i$

Remarque : pratiquement il est recommandé d'éliminer les valeurs extrêmes lorsque l'on calcule une moyenne empirique (typiquement 2 à chaque extrême).

La Moyenne I

Définition : Moyenne

On appelle moyenne d'un échantillon $x_1, x_2, ..., x_n$.

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

(Dans le cas d'une variable qualitative ça n'a pas de sens.)

La Moyenne II

Propriété : si x_i sont regroupées en modalités m_i on a aussi $\bar{x} = \sum_{i=1}^{r} m_i f_i$ où m_i :

- est la valeur de la modalité m_i si la variable est discrète comme un age,
- est la moyenne des valeurs de la modalité,

. $$m_i = [20 - 30] \quad \rightarrow \quad m_i = \frac{1}{n_i} \sum_{j \in m_i} x_j$$

Moyenne théorique

La moyenne théorique c'est l'espérance : $\mathbb{E}(X) = \lim_{n \to \infty} \bar{x}$

Propriété : $\mathbb{E}(X) = \int x \mathbb{P}(x) dx$.

L'exemple de pile ou face illustre bien la différence entre moyenne théorique (espérance) et moyenne empirique.

Existe t'il autre chose que la moyenne pour parler d'un ensemble de valeurs ?

Surtout quand la moyenne dit des bêtises !
Moyenne : $\{1, 5, 2, 4, 3, 100\} = 115/6 = 19,17$

Médiane (le point milieu)

Définition : Médiane

La médiane est la valeur $\widehat{M}$ telle que :

$$\widehat{F}_X(\widehat{M}) = \widehat{\mathbb{P}}(X \leq \widehat{M}) = \frac{1}{2}$$

(si on tire une valeur au hasard dans l'échantillon, on a autant de chance d'être au dessous que au dessus.)

x_i	n_i	f_i	F_i	$\widehat{F}_X$
1	1	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{12}$
2	1	$\frac{1}{6}$	$\frac{2}{6}$	$\frac{3}{12}$
3	1	$\frac{1}{6}$	$\frac{3}{6}$	$\frac{5}{12}$
4	1	$\frac{1}{6}$	$\frac{4}{6}$	$\frac{7}{12}$
5	1	$\frac{1}{6}$	$\frac{5}{6}$	$\frac{9}{12}$
100	1	$\frac{1}{6}$	1	$\frac{11}{12}$

$\widehat{F}_X(3) = 0,42$
$\widehat{F}_X(4) = 0,58$

$\widehat{M} = 3,5$

Médiane

Médiane **Théorique**

$$\min_{c \in \mathbb{R}} J_t(c) \quad \text{avec} \quad J_t(c) = \mathbb{E}\left(|X - c|\right)$$

Médiane **Empirique**

on remplace l'espérance par la moyenne :

$$\min_{c \in \mathbb{R}} J(c) \quad \text{avec} \quad J(c) = \sum_{i=1}^{n} |x_i - c|$$

$$\frac{\partial J(c)}{\partial c} = 0 \Leftrightarrow \sum_{i=1}^{n} \text{signe}(x_i - c) = 0$$

Remarque : La médiane est plus robuste aux valeurs extrêmes

Mode I

Définition : Mode

Le mode $\widehat{M_a}$ d'un échantillon est égal à :

$$\arg\max_{x \in \Omega}\{\Pr(x)\}$$

- Dans le cas discret, c'est la valeur la plus fréquente :

$$\widehat{M_a} = \arg\max_{i=1,\ldots,r} \hat{f}_i$$

- Dans le cas continu, c'est le **pic** de la distribution d'une variable continue.

Mode II

Remarques :

- La définition du mode n'exige pas de la variable qu'elle soit quantitative.
- À ces objets empiriques (moyenne, médiane, mode) on peut associer des objets théoriques.
- Dans le cas une loi normale : moyenne théorique $=$ médiane théorique $=$ mode théorique, *i.e.* si X suit une **loi normale**, $X \sim N(\mu, \sigma^2)$, alors, on a :

$$\textit{moyenne} = \textit{médiane} = \textit{mode} = \mu$$

Résumé central

On considère l'échantillon $S_n = \{x_1, x_2, \ldots, x_i, \ldots, x_n\}$.

- moyenne : la moyenne empirique / l'espérance

$$\min_{c\in\mathbb{R}} \sum_{i=1}^{n} (x_i - c)^2; \qquad c = \sum_{i=1}^{n} f_i x_i \qquad \mathbb{E}(X) = \int x\mathbb{P}(x)dx$$

- médiane : les fréquences cumulées / la fonction de répartition

$$\min_{M\in\mathbb{R}} \sum_{i=1}^{n} |x_i - M|; \qquad \widehat{\mathbb{P}}(X \leq M) = \frac{1}{2} \qquad \mathbb{P}(X \leq M) = \frac{1}{2}$$

- mode : les fréquences / les probabilités

$$\arg\max_{i\in\{1,\ldots,n\}} f_i \qquad \arg\max_{x\in\Omega} \mathbb{P}(x)$$

Résumé central

Quand calculer quel type d'indicateur ?

Type de variable	Exemple	Moyenne	Médiane	Mode
Qualitative multimodale	les départements...			
Qualitative bimodale	oui/non (0/1)			
Ordinale	j'aime...			
Quantitative discrète	nombre...			
Quantitative continue	temps...			

Résumé central

Quand calculer quel type d'indicateur ?

Type de variable	Exemple	Moyenne	Médiane	Mode
Qualitative multimodale	les départements...	–	–	OK
Qualitative bimodale	oui/non (0/1)	OK	–	ok
Ordinale	j'aime...	ok	OK	ok
Quantitative discrète	nombre...	OK	OK	OK
Quantitative continue	temps...	OK	OK	–

Et une fois que l'on dispose d'une valeur centrale ?

Surtout quand c'est la même !
Moyenne : $\{1, 5, 2, 4, 3\} = 15/5 = 3$
Moyenne : $\{3.1, 3, 2.9, 2.8, 3.2\} = 3$

Calcul d'un paramètre de dispersion

On considère l'échantillon $S_n = \{x_1, x_2, \ldots, x_i, \ldots, x_n\}$ de moyenne $\bar{x} = c$

Définition : Variance

C'est la moyenne des carrés des écarts à la moyenne :

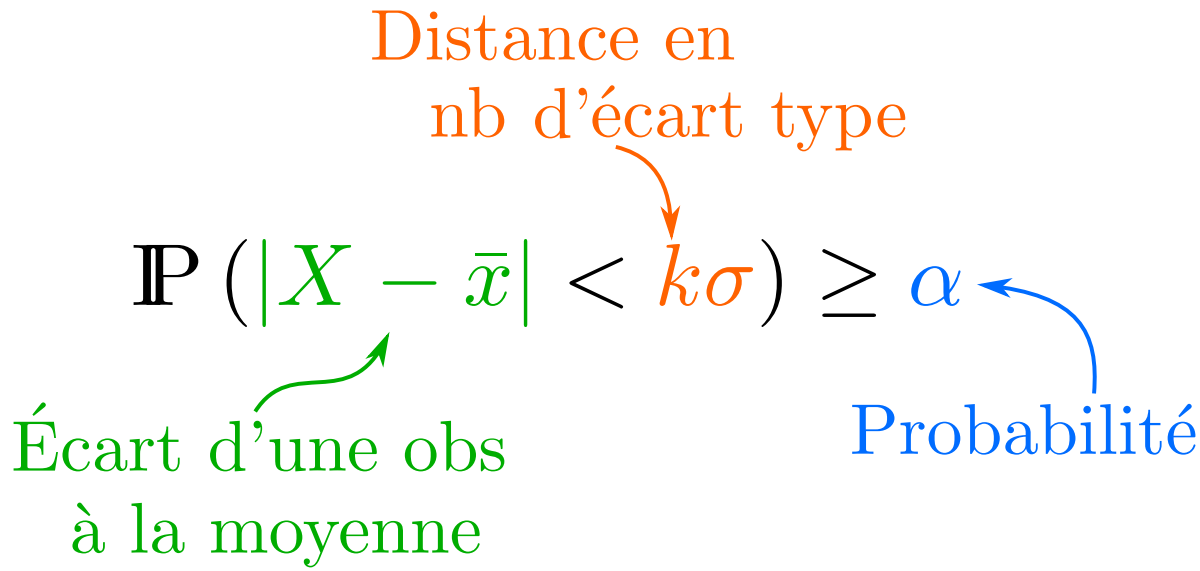
$$\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

- Écart : $e_i = (x_i - \overline{x}) \Rightarrow$ Écart carré : $e_i^2 = (x_i - \overline{x})^2$
- Variance : $\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} f_i e_i^2$

Note pour les calculs :

$$\begin{aligned}\widehat{\sigma}^2 &= \tfrac{1}{n}\textstyle\sum_{i=1}^{n}(x_i - \overline{x})^2 = \tfrac{1}{n}\sum_{i=1}^{n}\left(x_i^2 - 2x_i\overline{x} + \overline{x}^2\right) \\ &= \tfrac{1}{n}\textstyle\sum_{i=1}^{n}(x_i^2) - 2[\tfrac{1}{n}\sum_{i=1}^{n}(x_i)]\overline{x} + \overline{x}^2 = \tfrac{1}{n}\sum_{i=1}^{n}(x_i)^2 - \overline{x}^2\end{aligned}$$

Usage de la variance : donner une idée du domaine des observations I



Usage de la variance : donner une idée du domaine des observations II

Cas Gaussien

pour un grand échantillon $(n \geq 100)$:

$$\mathbb{P}\big(|X - \overline{x}| \geq u_{\alpha/2}\, \sigma\big) \leq \alpha$$

avec $u_{\alpha/2}$ pris dans la table de la loi normale.

- Si on fixe $\alpha = 0,05 \quad (5\%) \Longrightarrow k = 1,96$
  $\Longrightarrow \quad 95\%$ des observations $\in [\overline{x} - 1,96\widehat{\sigma}\ ,\ \overline{x} + 1,96\widehat{\sigma}]$
- Si maintenant on fixe $k = 3$:
  - $\mathbb{P}\left(|X - \overline{x}| > 3\sigma\right) = 0.99865$
  - $\Rightarrow \mathbb{P}\left(|X - \overline{x}| < -3\sigma\right) = 1 - 0.99865 = 0.00135$
  - $\Rightarrow \quad 99,73\%$ des observations $\in [\overline{x} - 3\widehat{\sigma}\ ,\ \overline{x} + 3\widehat{\sigma}]$

Les quantiles I

Définition : Quantiles

On appelle quantiles (fractiles) à l'ordre p, $\forall p \in [0,1], \widehat{\Phi}_p$

$$\widehat{\mathbb{P}}(X \leq \widehat{\Phi}_p) = p$$

ou de manière équivalente : $\widehat{\Phi}_p$ telle que $\widehat{F}_X(\widehat{\Phi}_p) = p$

Les Quartiles

- $\widehat{\Phi}_{\frac{1}{4}} = \widehat{Q}_1$, telle que $\widehat{F}(\widehat{Q}_1) = \frac{1}{4}$,
- $\widehat{\Phi}_{\frac{1}{2}} = \widehat{Q}_2 = \widehat{M}$, telle que $\widehat{F}(\widehat{M}) = \frac{1}{2}$,
- $\widehat{\Phi}_{\frac{3}{4}} = \widehat{Q}_3$, telle que $\widehat{F}(\widehat{Q}_2) = \frac{3}{4}$.

Les quantiles II

Définition : Distance inter quartile (DIQ)

$$DIQ = \widehat{Q}_3 - \widehat{Q}_1$$

Résumé robuste

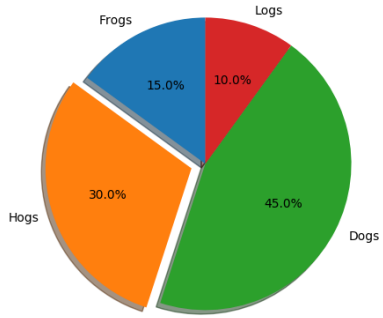
Considérons l'échantillon suivant : $4,\ 1,\ 6,\ 2,\ 72,\ 4,\ 6,\ 5,\ 1,\ 3,\ 7$

Statistique	Échantillon complet	Sans la valeur 72
Moyenne	10.09	3.9
Médiane	4	4
Variance	425.69	4.54
Distance interquartile	3.37	4

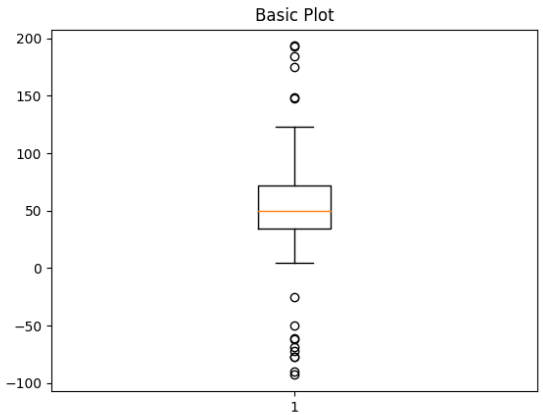
Plan

Résumé graphique des données

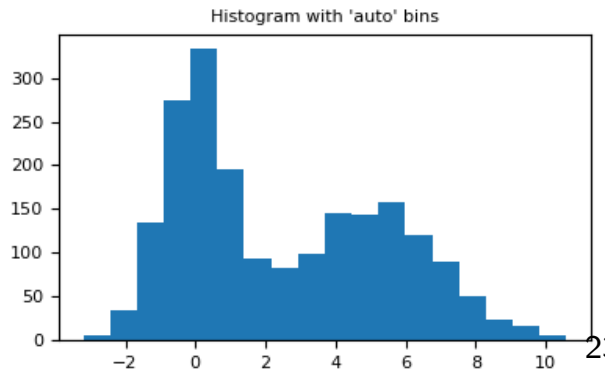
- Qualitatives :
  - Camemberts (pie charts)



- Quantitatives
  - Continues : boite à moustache

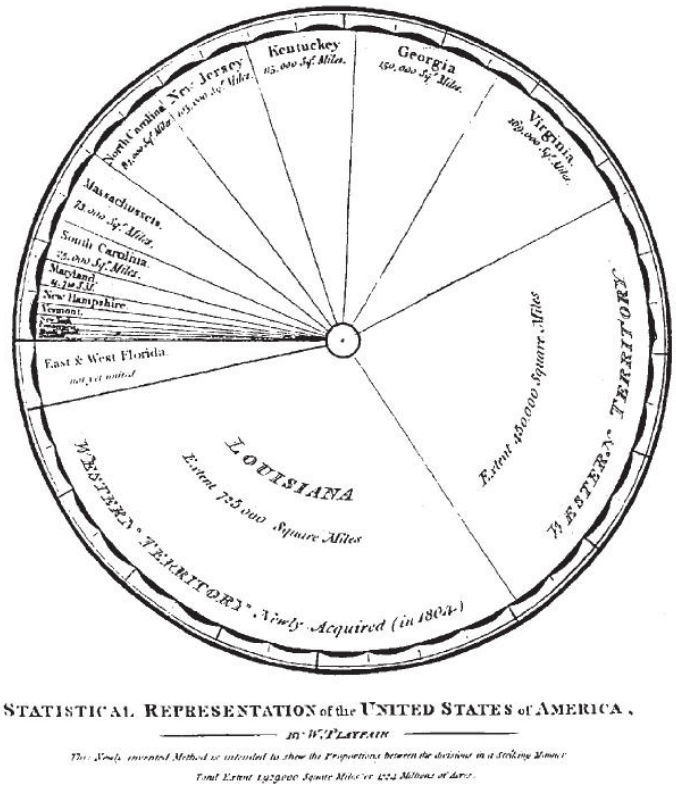


  - Discrètes : histogramme



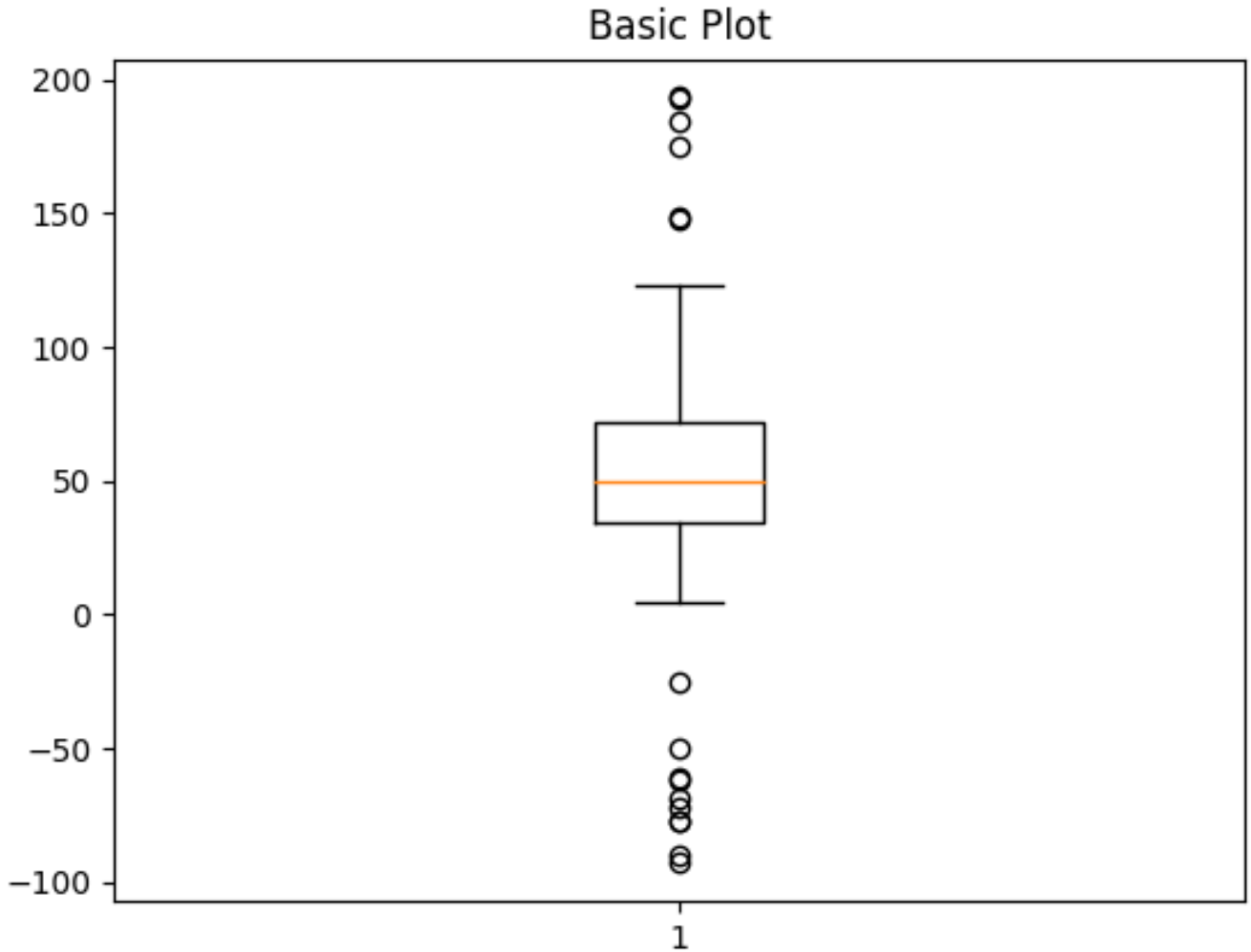
Cas des variables qualitatives

Camemberts (diagramme en secteurs,diagramme circulaire)
Angle du camembert : fréquence f_i multipliée par 360.



http://euclid.psych.yorku.ca/SCS/Gallery/images/playfair1805-pie2.jpg

Variables qualitatives : boite à moustache



Boite à moustache

Boite

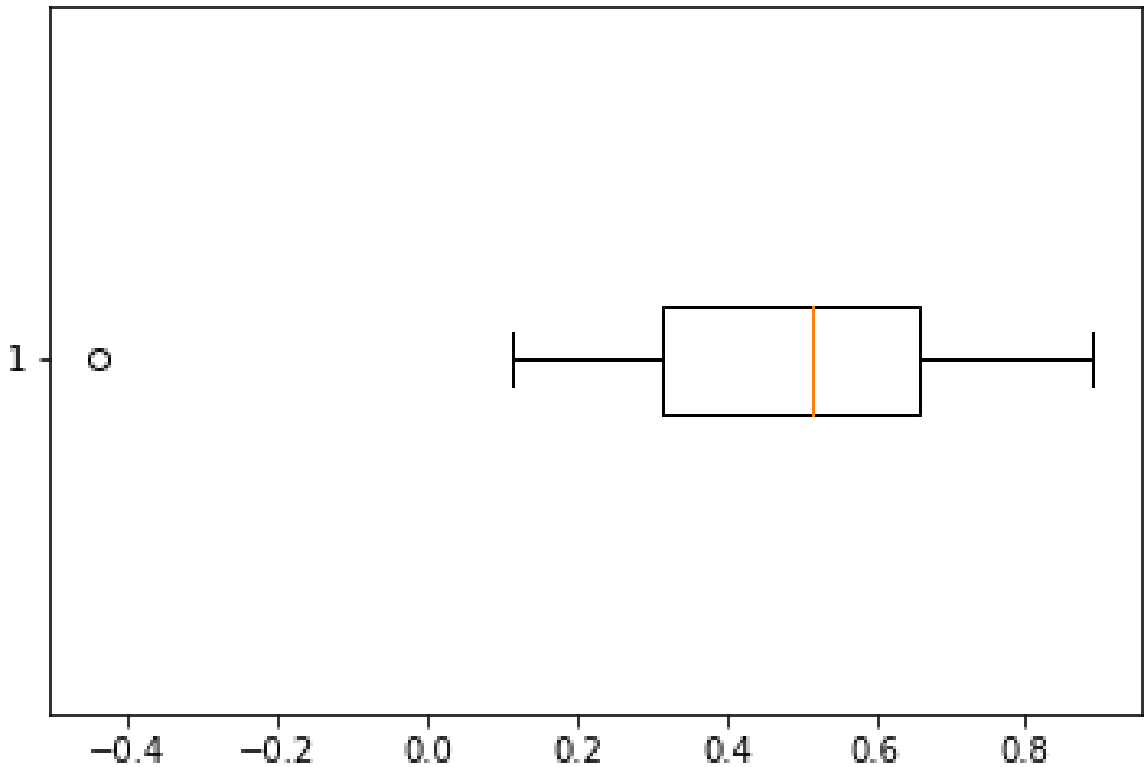
- Ligne : médiane
- Extrémités de la boîte : Quartiles

Épures (whiskers)

- Par défaut :
$$[\widehat{Q}_1 - \frac{3}{2}DIQ; \widehat{Q}_3 + \frac{3}{2}DIQ].$$
- Réglable : RTFM : `https://matplotlib.org/3.1.1/api/_as_gen/matplotlib.pyplot.boxplot.html`
- Les points hors épure sont représentés (outliers)

Boite à moustache

0.5300
0.2841
0.4361
0.1869
0.8340
0.1406
0.6683
0.8088
0.3240
0.3618
0.3096
0.5754
0.2800
-0.720
0.6135
0.5610
0.4228
0.1138
0.8681
0.8665
0.5023
0.6334
0.8877

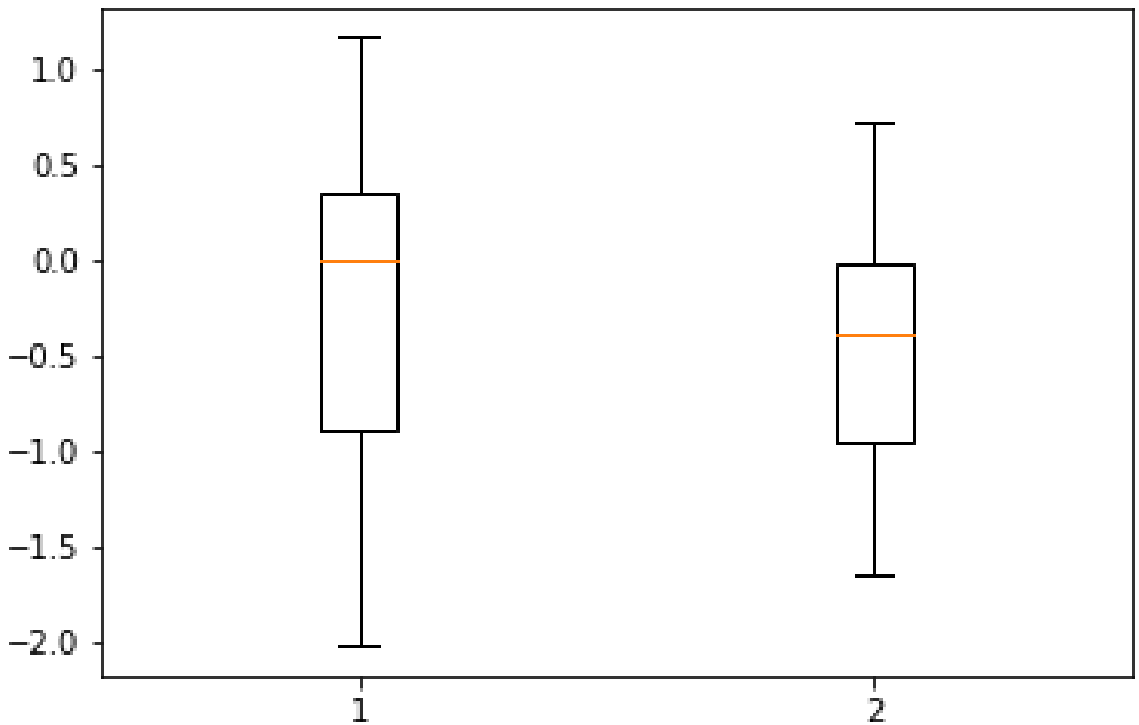


Exemple d'utilisation des boites à moustache

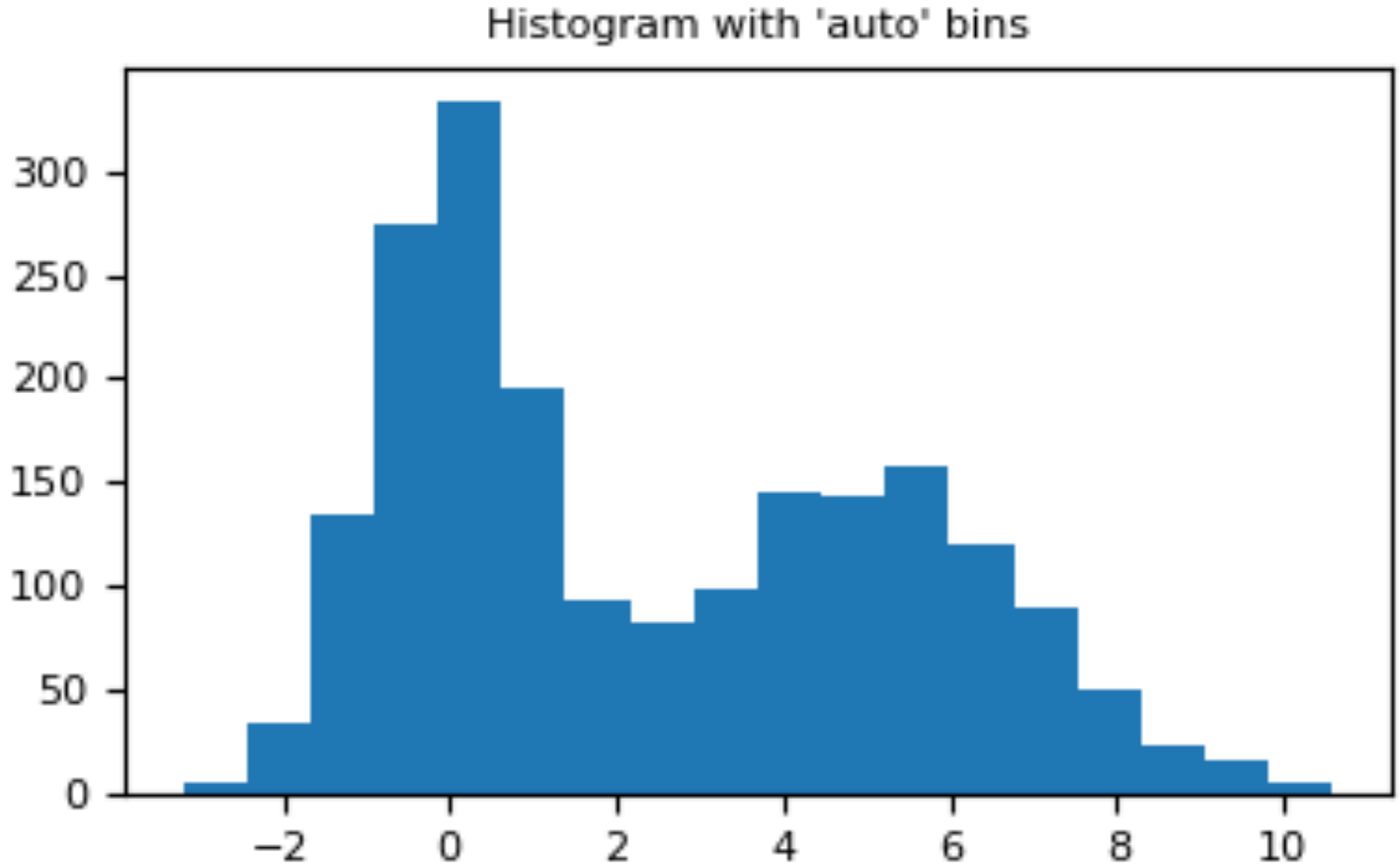
-1.1754	-0.6376
-0.3977	-0.3052
-1.9512	0.5161
0.9047	0.3681
0.6822	0.0715
-2.0183	-0.3123
-0.4543	0.7366
0.0421	-0.3635
-0.4582	-0.9618
0.2641	-0.2175
0.2920	0.0235
0.3868	-0.7281
0.5540	-1.3258
0.9437	-0.1650
0.0598	-0.1500
0.5747	0.5769
-1.4419	0.2333
-0.2355	-0.9544
-1.0214	-0.4148
0.1398	-1.0734
0.0816	-1.0024
-1.2582	-0.7818
-1.4685	-1.0750
-0.0329	-1.6426
-1.6925	-0.9326
-0.4757	-0.8195

Pour comparer visuellement 2 variables :

$$c_1 = -0.13 \quad c_2 = -0.55 \qquad \widehat{\sigma}_1 = 1.02 \quad \widehat{\sigma}_2 = 0.67$$



Variables discrètes : Histogramme



- Abscisse : intervalle de valeurs
- Ordonnée : hauteur
- Surface : nombre d'individu ou fréquence

Variables discrètes : Histogramme

- Variable discrète : on compte le nombre de fois ou la modalité est apparue.
- Variable Continue : on discrétise le domaine.

Discrétiser un domaine se donner une suite ordonnée de valeurs $\{b_i\}_{i=0,\dots,p}$ qui couvrent le domaine.

À chaque intervalle $[b_{i-1}, b_i[$ on associe une hauteur h_i telle que la surface du rectangle ainsi créé soit proportionnelle au nombre d'observations incluses dans l'intervalle $s_i = h_i(b_i - b_{i-1})$. On résume la situation en rappelant que l'on dispose :

- des bornes : $p+1$ bornes pour p classes $\{b_i\}_{i=0,p}$
- des intervalles $b_i - b_{i-1}$
- des effectifs et des surfaces s_i
- des hauteurs $s_i = h_i(b_i - b_{i-1}) \quad \Leftrightarrow \quad h_i = \frac{s_i}{b_i - b_{i-1}}$

Détail de la construction d'histogrammes

Comment choisir p le nombre d'intervalles ?

- la règle de Sturges (si on a n observations) : $p \geq 1 + \log n$
- la règle de Scott : $p = \frac{3{,}5\widehat{\sigma}}{n^{1/3}}$
- la règle de Freedman Diaconis : $p = 2\frac{DIQ}{n^{1/3}}$
- trouver p par équi-répartition tel que $s_i \geq 5$.
- `https://numpy.org/doc/stable/reference/generated/numpy.histogram_bin_edges.html#numpy.histogram_bin_edges`

Comment choisir les $\{b_i\}_{i=0,\ldots,p}$? On peut le faire par équirépartition des individus ou des intervalles :

- équirépartition des individus par classe : $b_0 = min, b_i$ est calculé tel qu'il y ait θ observations entre et b_{i+1}
- équirépartition des intervalles : $largeur = \frac{max-min}{p}$; $b_0 = min$; $b_i = b_0 + i\left(\frac{max-min}{p}\right)$

Comment construire un histogramme (variable continue)

Construction d'un histogramme équiréparti :

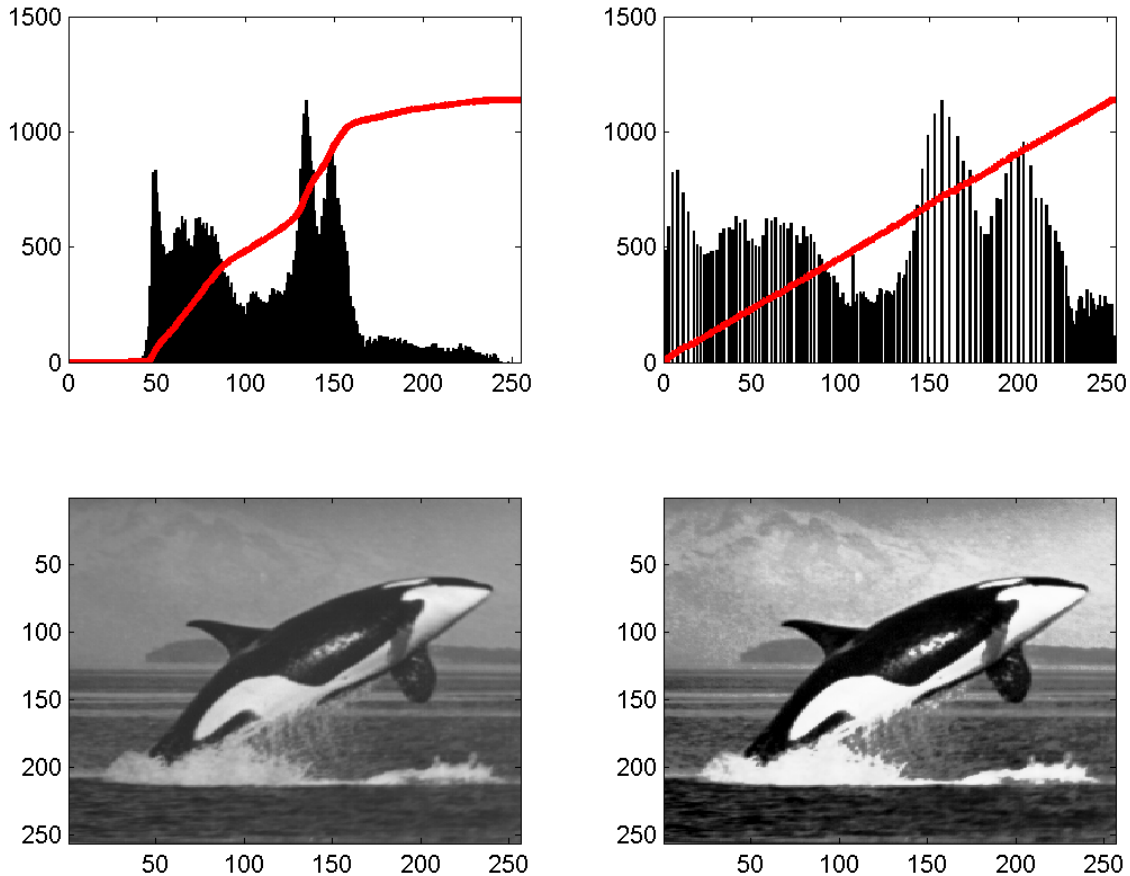
1. choisir p le nombre de classes,
2. calculer les $\{b_i\}_{i=0,\dots,p}$, les bornes des intervalles,
3. en déduire les s_i le nombre d'individus par classe. Si ce nombre est inférieur à cinq, on fusionne des classes.
4. calculer enfin les $h_i = \frac{s_i}{b_i - b_{i-1}}$ les hauteurs.

Par exemple pour 32 observations entre 0 et 15. On en déduit $p = 6$ et les valeurs suivantes :

$$b_0 = 0 \quad b_1 = 2 \quad b_2 = 3 \quad b_3 = 4 \quad b_4 = 8 \quad b_5 = 10 \quad b_6 = 15$$
$$s_1 = 6 \quad s_2 = 5 \quad s_3 = 6 \quad s_4 = 5 \quad s_4 = 5 \quad s_5 = 5$$
$$h_1 = \frac{6}{2} \quad h_2 = 5 \quad h_3 = 6 \quad h_4 = \frac{5}{4} \quad h_5 = \frac{5}{2} \quad h_6 = \frac{5}{5}$$

Histogramme et traitement d'images

Redressement (où égalisation) d'histogramme [1] : $b_n(i) = \widehat{F}^{-1}\left(\frac{i}{p}\right)$



1. http://en.wikipedia.org/wiki/Histogram_equalization

Histogramme et traitement d'images

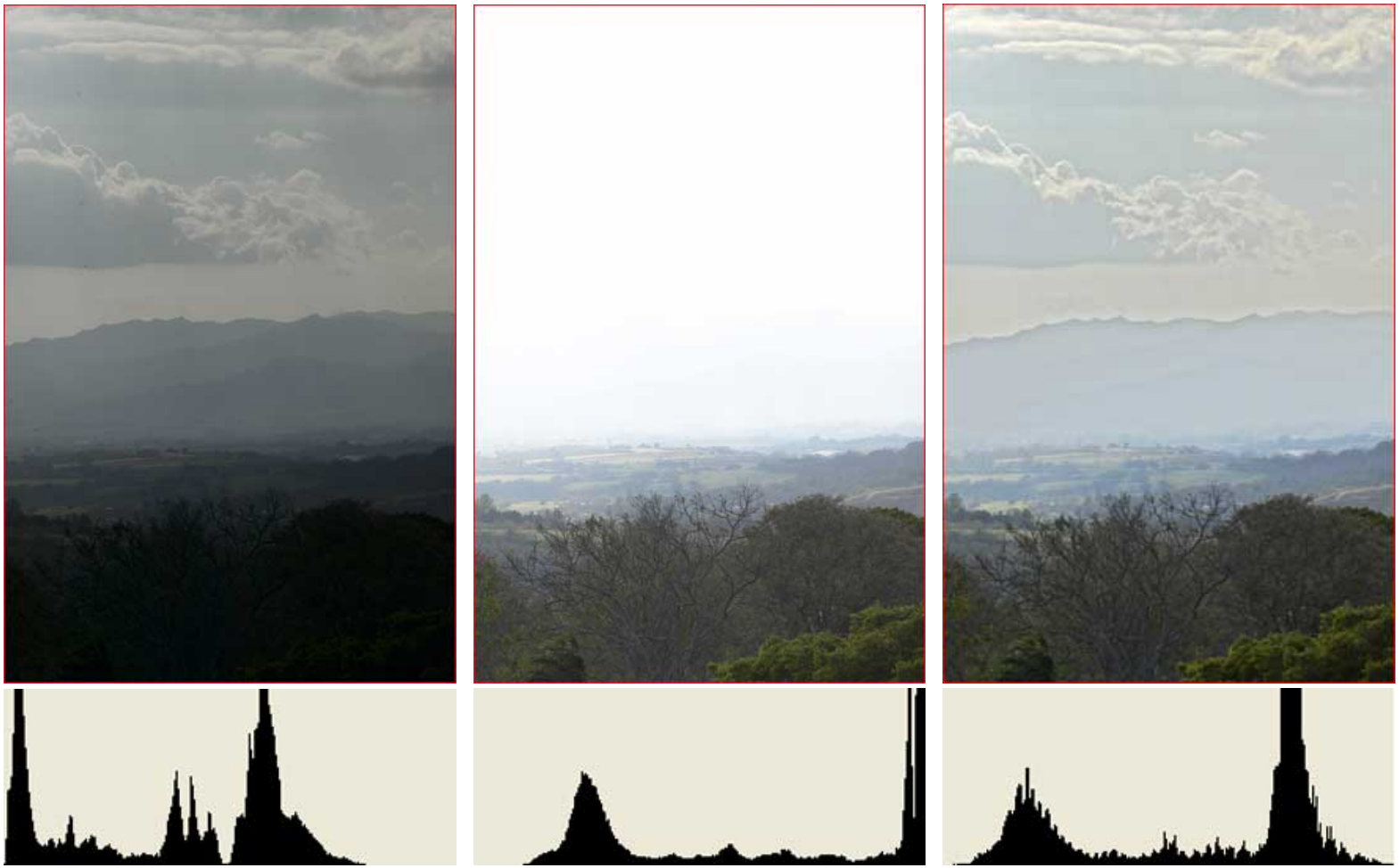


Figure – Exemple d'histogrammes `http://www.llvj.com/tutorials/understanding-series/understanding-histograms.shtml`.

Plan

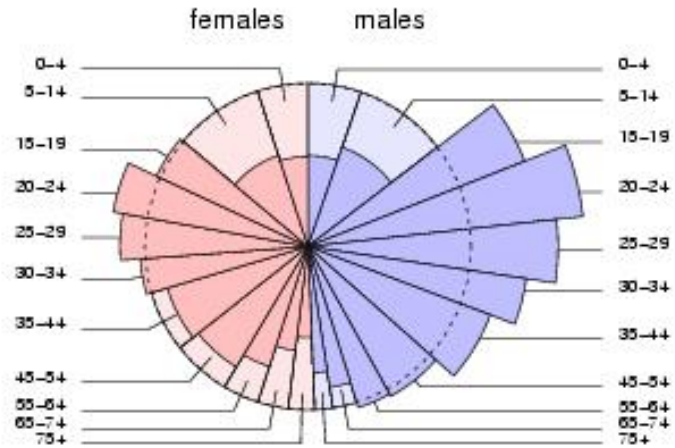
Empirique vs. Théorique

Cas d'une variable X discrète :

Nom Empirique	Empirique	Théorique	Nom Théorique
Fréquence	$\hat{f}_i = \frac{n_i}{n}$	$\mathbb{P}(X = m_i)$	Probabilité
Fonction de répartition	$\hat{F}_i = \frac{N_i}{N}$	$\mathbb{P}(X < a_i) = F(m_i)$	Fonction de répartition
Moyenne	$\bar{x} = \sum_{i=1}^{n} f_i x_i$	$\mathbb{E}(X) = \sum_{i=1}^{n} \mathbb{P}(m_i) m_i$	Espérance
Médiane	$\hat{F}(\hat{M}) = \frac{1}{2}$	$\mathbb{P}(X < M) = \frac{1}{2}$	Médiane
Variance	$\frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2$	$\sum_{i=1}^{n} \mathbb{P}(m_i)(m_i - \mathbb{E}[X])^2$	Variance

Cas d'une variable X continue : l'empirique ne change pas (avec $n_i = 1$)

Conclusion

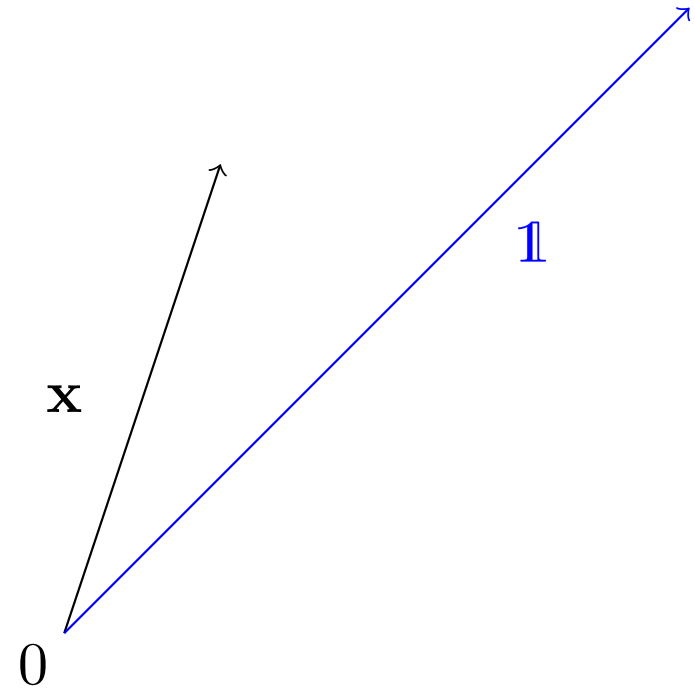


- Résumé quantitatif
  - ordre 1 : centrage
  - ordre 2 : dispersion
  - ordres supérieurs : asymétrie, aplatissement...
- Résumé qualitatif (graphique)
  - qualitative : camembert
  - quantitative (continue) : boite à moustache.
  - quantitative discrète : histogramme
- Méthode exploratrice
  - pour une exploration interactive des données

Interprétation géométrique de la moyenne

$$J(c) = \sum_{i=1}^{n}(x_i - c)^2$$

$$= \|\mathbf{x} - c\,\mathbb{1}\|^2$$



- $\mathbf{x} = (x_1, x_2, ..., x_n)^\top \in \mathbb{R}^n$
- $\mathbb{1} = (1, 1, ..., 1)^\top \quad \in \mathbb{R}^n$

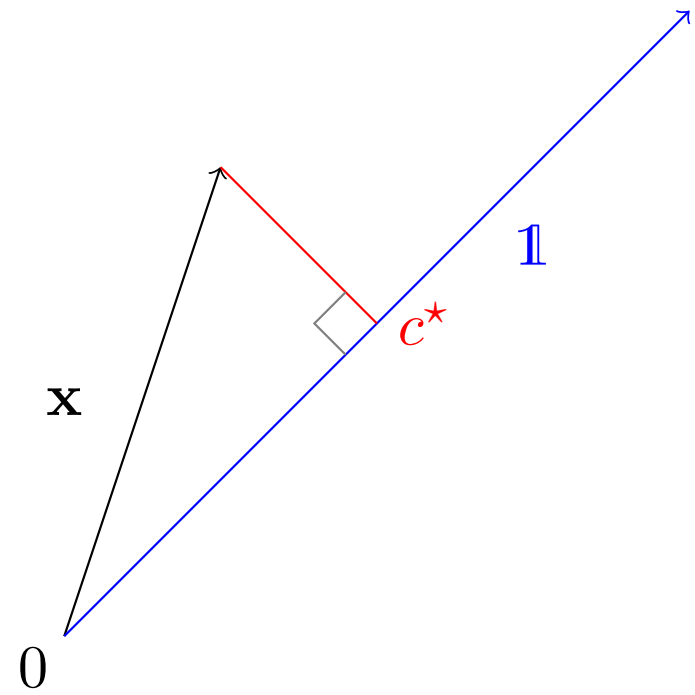
$c^\star$, la moyenne est aussi la projection orthogonale du vecteur des observations sur le vecteur $\mathbb{1}$

$$c^\star = \frac{\mathbf{x}^\top \mathbb{1}}{\mathbb{1}^\top \mathbb{1}} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Interprétation géométrique de la moyenne

$$J(c) = \sum_{i=1}^{n}(x_i - c)^2$$

$$= \|\mathbf{x} - c\,\mathbb{1}\|^2$$

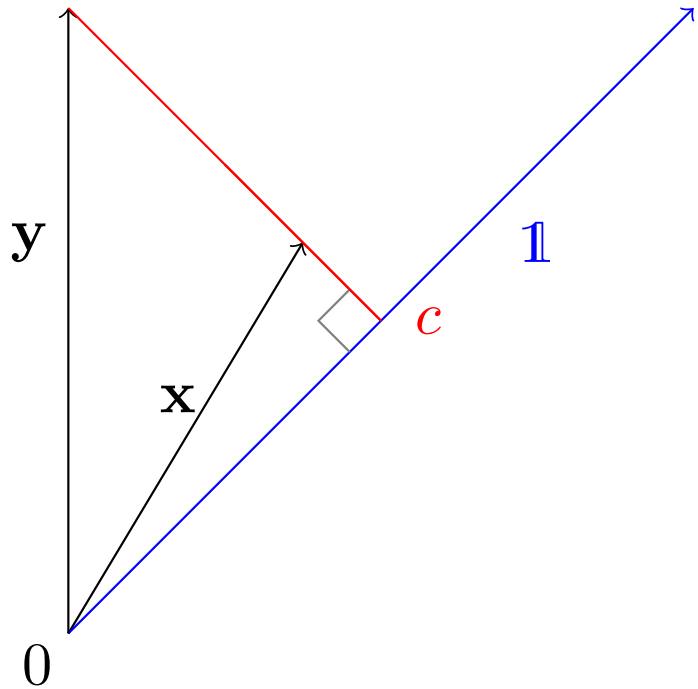


- $\mathbf{x} = (x_1, x_2, ..., x_n)^\top \in \mathbb{R}^n$
- $\mathbb{1} = (1, 1, ..., 1)^\top \quad \in \mathbb{R}^n$

$c^\star$, la moyenne est aussi la projection orthogonale du vecteur des observations sur le vecteur $\mathbb{1}$

$$c^\star = \frac{\mathbf{x}^\top \mathbb{1}}{\mathbb{1}^\top \mathbb{1}} = \frac{\sum_{i=1}^{n} x_i}{n}$$

La moyenne des écarts à la moyenne : la variance



La variance $\widehat{\sigma}^2$

$$\widehat{\sigma}^2 = \frac{1}{n} \|\mathbf{x} - c\,\mathbb{1}\|^2$$

Pour deux échantillons $\mathbf{x}$ et $\mathbf{y}$ de même moyenne, la variance traduit leur proximité avec le vecteur $\mathbb{1}$.

L'Écart type : $\widehat{\sigma} = \sqrt{\widehat{\sigma}^2}$

C'est la norme du vecteur $\mathbf{x} - c\,\mathbb{1}$ (divisée par $\sqrt{n}$)

Et Pythagore nous dit que : $\|\mathbf{x}\|^2 = nc^2 + \|\mathbf{x} - c\,\mathbb{1}\|^2$