Description mono variable

Stéphane Canu
stephane.canu@litislab.eu

M8 - Principes du traitement de l'information

February 7, 2018

Plan

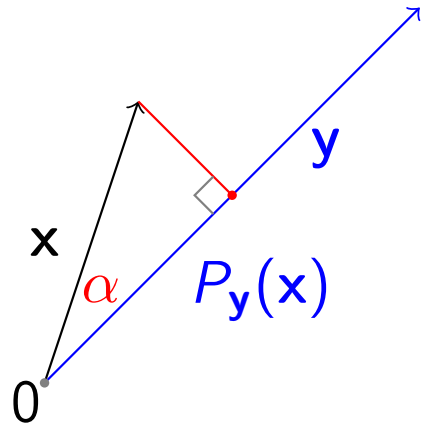
Rappels de M4 : espace vectoriel $\mathbb{R}^n$

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

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 :

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

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

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

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 :

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

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

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

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

Quel est le meilleur résumé ?

Comment synthétiser un échantillon $S_n = \{x_1, x_2, \ldots, x_i, \ldots, x_n\}$
. $\longrightarrow$ par une seule information c ?

Nous pouvons poser un principe variationnel :

$$\min_{c \in \mathbb{R}} J(c) \quad \text{avec} \quad J(c) = \sum_{i=1}^{n} (x_i - c)^2$$

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 : $c = \frac{1}{n} \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} \frac{1}{n} x_i = \sum_{i=1}^{n} \hat{f}_i x_i$

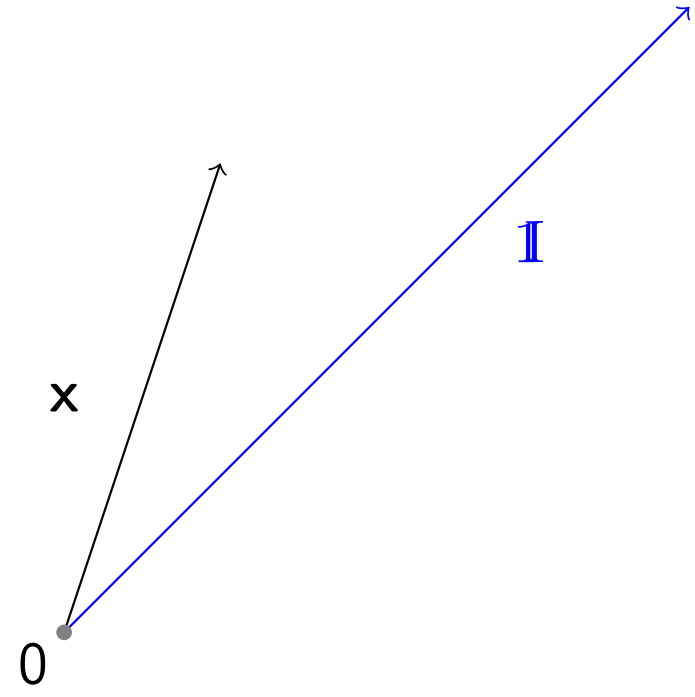
Remarque : pratiquement il est recommandé d'éliminer les valeurs extrêmes lorsque l'on calcule une moyenne empirique (typiquement 2 à chaque extrême).

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 $\mathbb{1}$

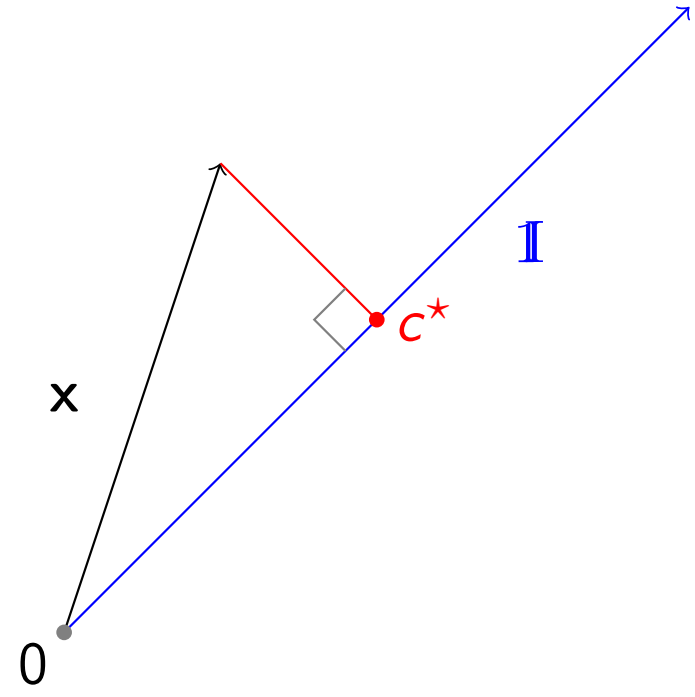
$$c^\star = \frac{\mathbf{x}^\top \mathbb{1}}{\mathbb{1}^\top \mathbb{1}} = \frac{\sum_{i=1}^{n} x_i}{n}$$

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 $\mathbb{1}$

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

Definition (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.

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,

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

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 : $\{1, 5, 2, 4, 3, 100\} = 115/6 = 19,17$

Mediane (le point milieu)

Definition (Mediane)

C'est le valeur $\widehat{M}$ telle que : $\widehat{F}_X(\widehat{M}) = \hat{\mathbb{P}}(X \leq \widehat{M}) = \frac{1}{2}$

où $\widehat{F}_X(x)$ est la fonction de répartition empirique de l'échantillon

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

Exemple

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	$\frac{11}{12}$

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

$\widehat{M} = 3,5$

Si la variable est continue ($n_i = 1$), il suffit de trier les observations et de prendre la valeur centrale (n impair) où la moyenne des deux valeurs centrales (n pair)

Mediane

La médiane théorique est : $\min_{c\in\mathbb{R}} J_t(c)$ avec $J_t(c) = \mathbb{E}(|X - c|)$

Pour la médiane empirique on remplace l'espérance par la moyenne,

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

à l'optimum on a : $$\frac{\partial J(c)}{\partial c} = \sum_{i=1}^{n} \text{signe}(x_i - c) = 0$$

Et pour avoir autant de signes $-$ que de signes $+$, il faut prendre c au milieu.

Remarque : La médiane est plus robuste aux valeurs extrêmes

Definition (Mode)

C'est $Argmax_{x \in \Omega}\{\mathbb{P}(x)\}$

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

$$\widehat{M}_a = \underset{i=1,\ldots,r}{Argmax}\, \hat{f}_i$$

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

La définition du mode n'exige pas de la variable qu'elle soit quantitative.
Remarque: A 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 :

$$moyenne \ = médiane \ = mode \ = \mu$$

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

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

- mode : les fréquences / les probabilités

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

Résumé central

Quand calculer quel type d'indicateur ?

$\longrightarrow$ ça dépend de la nature de la variable considérée :

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 ?

$\longrightarrow$ ça dépend de la nature de la variable considérée :

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	–

Résumé central

génère n = 200 points au hasard (suivant une loi exponentielle de paramètre 1)

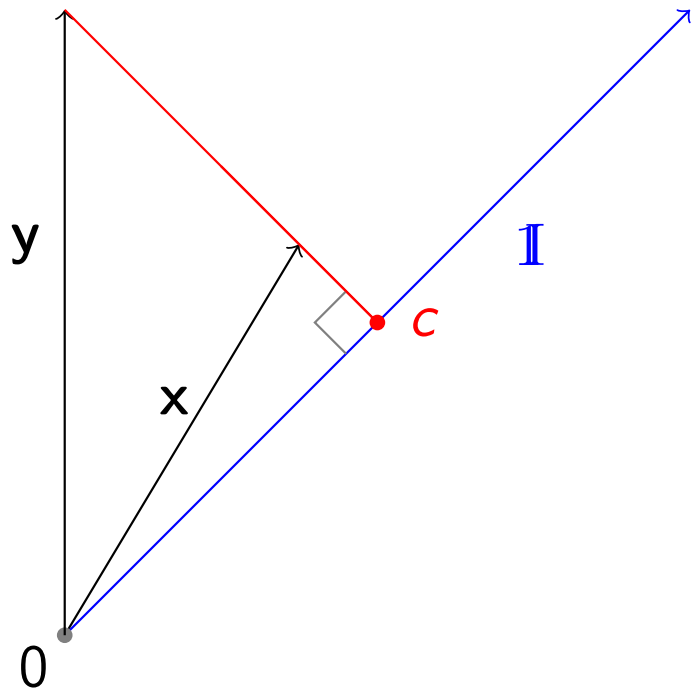
```
x = random('exp',1,200,1);
x2 = [x; 1000];                 % ajoutons une valeur aberrante
[mean(x) mean(x2)]              % calcul des moyennes des deux échantillons
    1.0072    5.9773
[median(x) median(x2)]          % calcul des médianes des deux échantillons
    0.7415    0.7418
[mode(x) mode(x2)]              % calcul des modes des deux échantillons
    0.0026    0.0026
nbin=7;                         % On discretise la variable continue
d=linspace(min(x),max(x),nbin+1);
for i=1:nbin
   H(i)=length(find(x>d(i)&x<=d(i+1)));
end
[v,i] = max(H);                 % on recherche la classe d'effectif maximal
le_mode = (d(i) + d(i+1))/2     % Le mode est le centre de l'interval associé
    0.4746
```

- la médiane est robuste
- le mode d'une variable continue ne se calcule pas directement

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 : $\{3.1, 3, 2.9, 2.8, 3.2\} = 3$

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

Calcul d'un paramètre de dispersion

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

Definition (Variance)

C'est la moyenne des carrées des écarts :

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

- Écart: $e_i = (x_i - c)$ Écart carré: $e_i^2 = (x_i - c)^2$
- 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 :

$$\begin{aligned}\widehat{\sigma}^2 &= \frac{1}{n}\sum_{i=1}^{n}(x_i - c)^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}\left(x_i^2 - 2x_i c + c^2\right) = \frac{1}{n}\sum_{i=1}^{n}(x_i)^2 - c^2\end{aligned}$$

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

$$\mathbb{P}\left(|X - c| < k\sigma\right) \geq \alpha$$

Pire des cas : Tchebychev

$$\mathbb{P}\big(|X - c| \geq \varepsilon\big) \leq \frac{\sigma^2}{\varepsilon^2} \qquad k = \frac{1}{\sqrt{\alpha}}$$

Cas Gaussien

pour un grand échantilon ($n \geq 100$)

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

Si on fixe $\alpha = 0,05 \quad (5\%) \quad \Longrightarrow \quad \frac{1}{\sqrt{\alpha}} = 4,47$

$\Longrightarrow$ plus de 95% des observations $\in [c - 4,47\widehat{\sigma}\ ,\ c + 4,47\widehat{\sigma}]$

$\Longrightarrow$ plus de 95% des observations $\in [c - 1,96\widehat{\sigma}\ ,\ c + 1,96\widehat{\sigma}]$

Si maintenant on fixe k : c'est l'approche « Six sigma » $\alpha = \frac{1}{k^2}$

$\Longrightarrow$ plus de 89% des observations $\in [c - 3\widehat{\sigma}\ ,\ c + 3\widehat{\sigma}]$

$\Longrightarrow$ plus de 99,7% des observations $\in [c - 3\widehat{\sigma}\ ,\ c + 3\widehat{\sigma}]$

Les fractiles

Definition (Fractiles)

On appelle 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,

$$\widehat{\Phi}_p \text{ telle que } \widehat{F}_X(\widehat{\Phi}_p) = p$$

Cas particuliers : 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 $\widehat{F}(\widehat{Q}_2) = \frac{3}{4}$.

Definition (Distance inter quartile (DIQ))

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

MAD

la médiane des écarts absolus

$$MAD = \text{médiane}(|x_i - \widehat{M}|)$$

`http://en.wikipedia.org/wiki/Median_absolute_deviation`

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

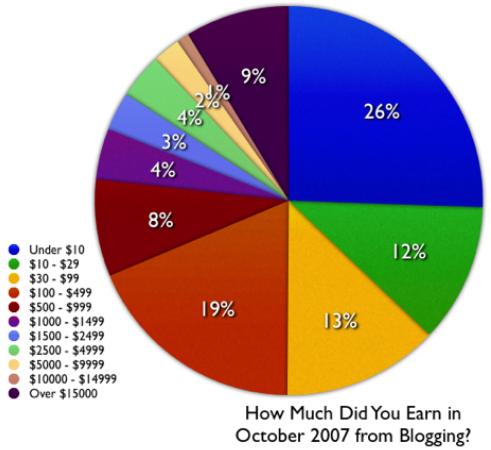
en admettant une hypothèse gaussiène, 95 % des observations (soit dix neuf sur vingt) sont dans l'intervalle

$$[3,9 - \sqrt{4.54} \times 1.96 \ , \ 4 + \sqrt{4.54} \times 1.96] = [-0,28 \ , \ 9,08]$$

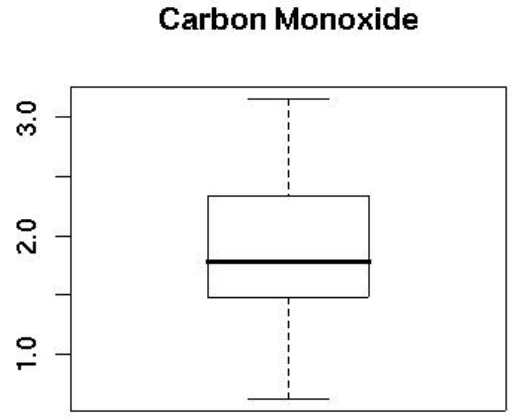
Plan

Résumé graphique des données

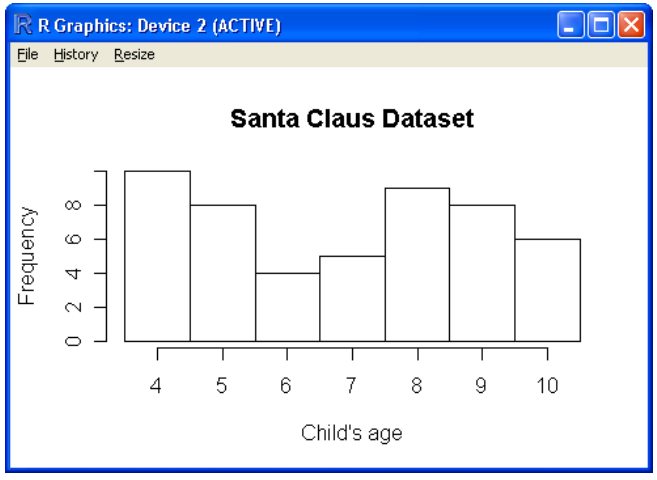
- Qualitatives : camemberts



- Quantitatives
  - Continues : boite à moustache

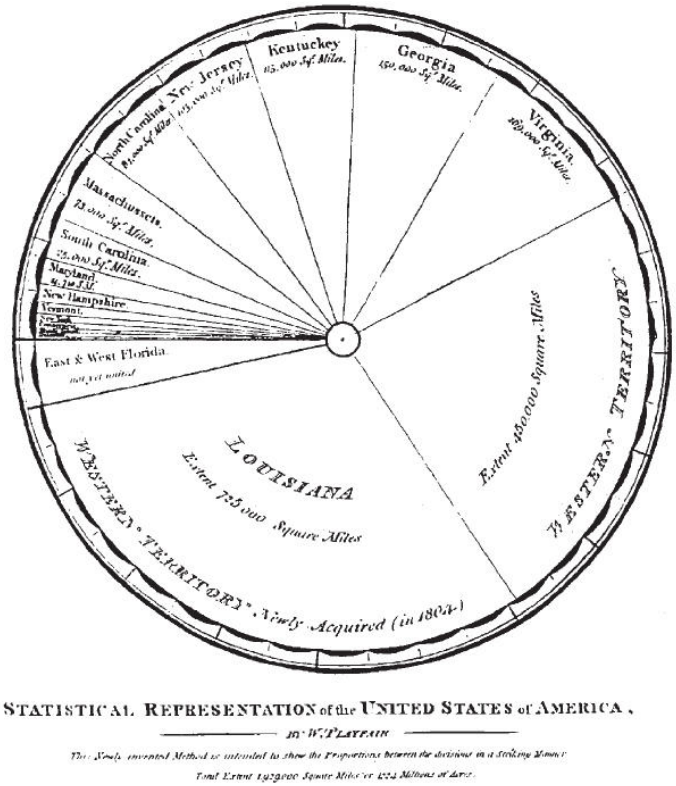


  - Discrètes : histogramme



Cas des variables qualitatives

Camemberts (diagramme en secteurs)



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

Variables qualitatives : boite à moustache

Un point hors de l'intervalle extrême est un point aberrant et on le représente par un plus où une double étoile. Un exemple de boite à moustache est donné figure 1.

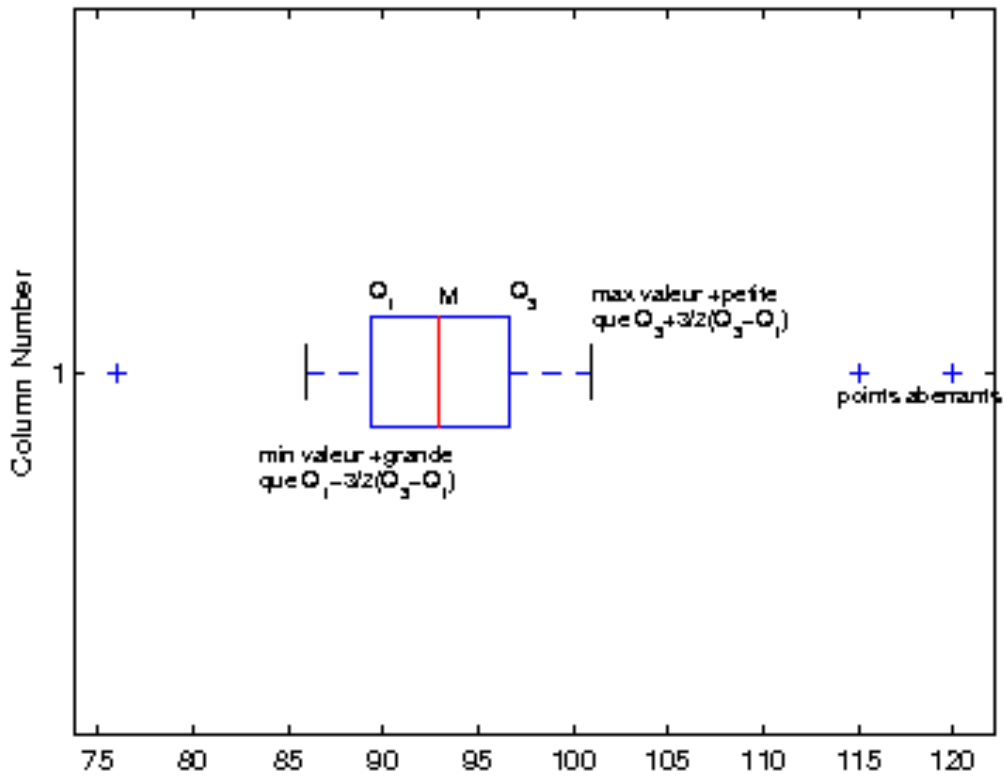


Figure : Exemple de boite à moustache

Boite à moustache

Definition (DIQ)

La Distance InterQuartile est définie de la manière suivante :

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

L'**épure** d'un échantillon est l'intervalle :

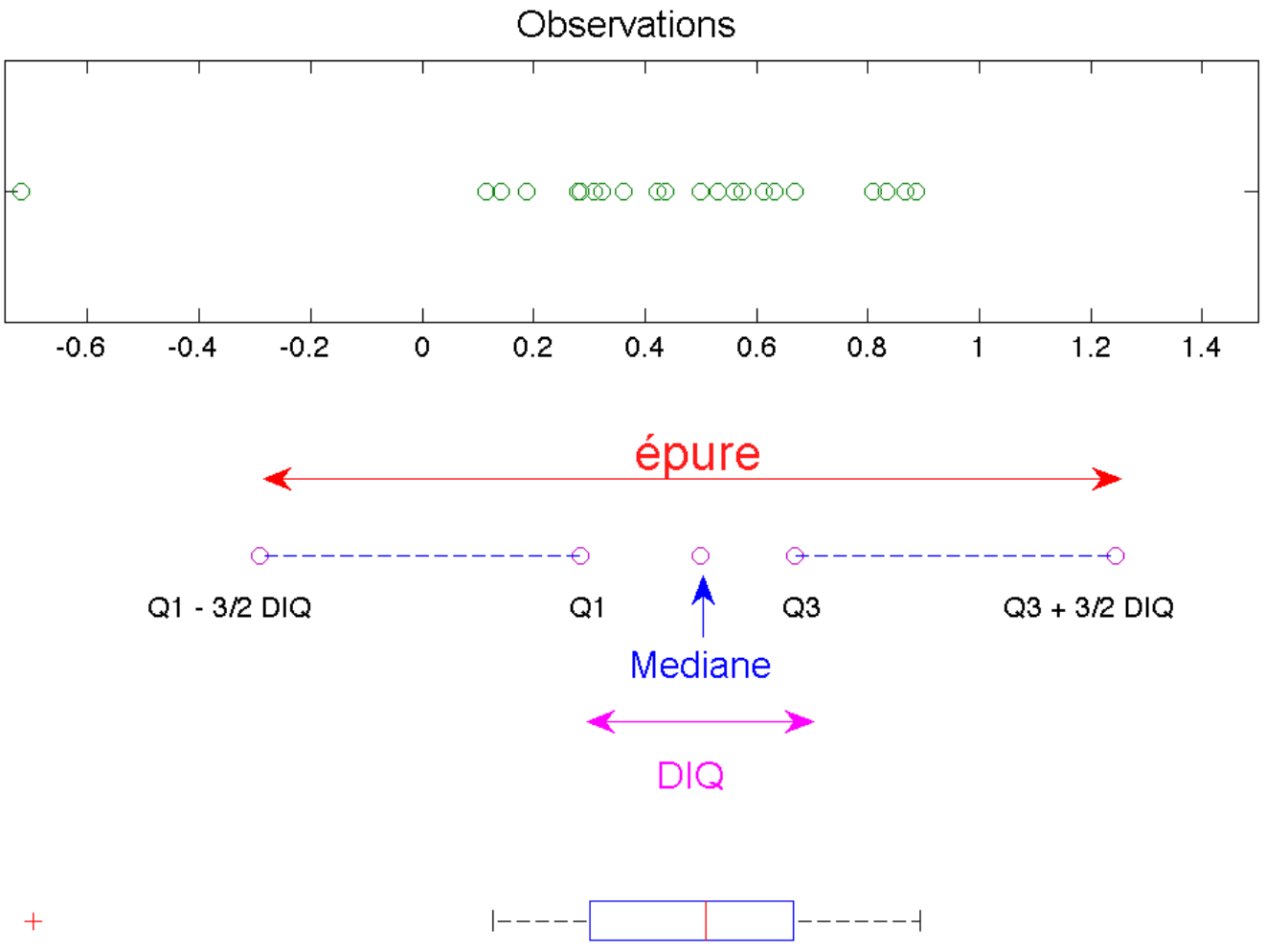
$$[\widehat{Q}_1 - \frac{3}{2}DIQ; \widehat{Q}_3 + \frac{3}{2}DIQ].$$

Les moustaches de la boite, seront données par la plus petite et la plus grande observation dans l'épure. Les points hors épure vont être représentes par une étoile ou un $\circ$. Un point est dit « suspect » s'il est hors épure mais dans l'intervalle (représenté par une étoile)

$$[\widehat{Q}_1 - 3 \cdot DIQ; \widehat{Q}_3 + 3 \cdot DIQ].$$

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

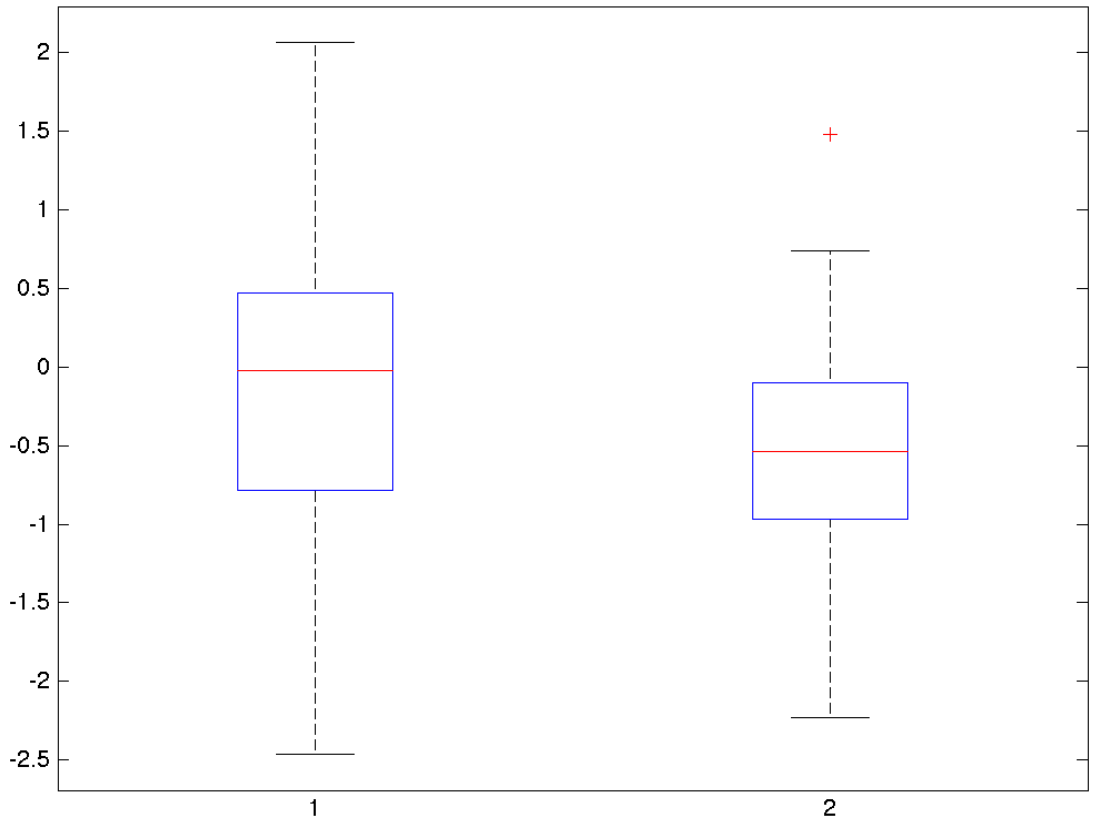


Exemple d'utilisation des boites à moustache

-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

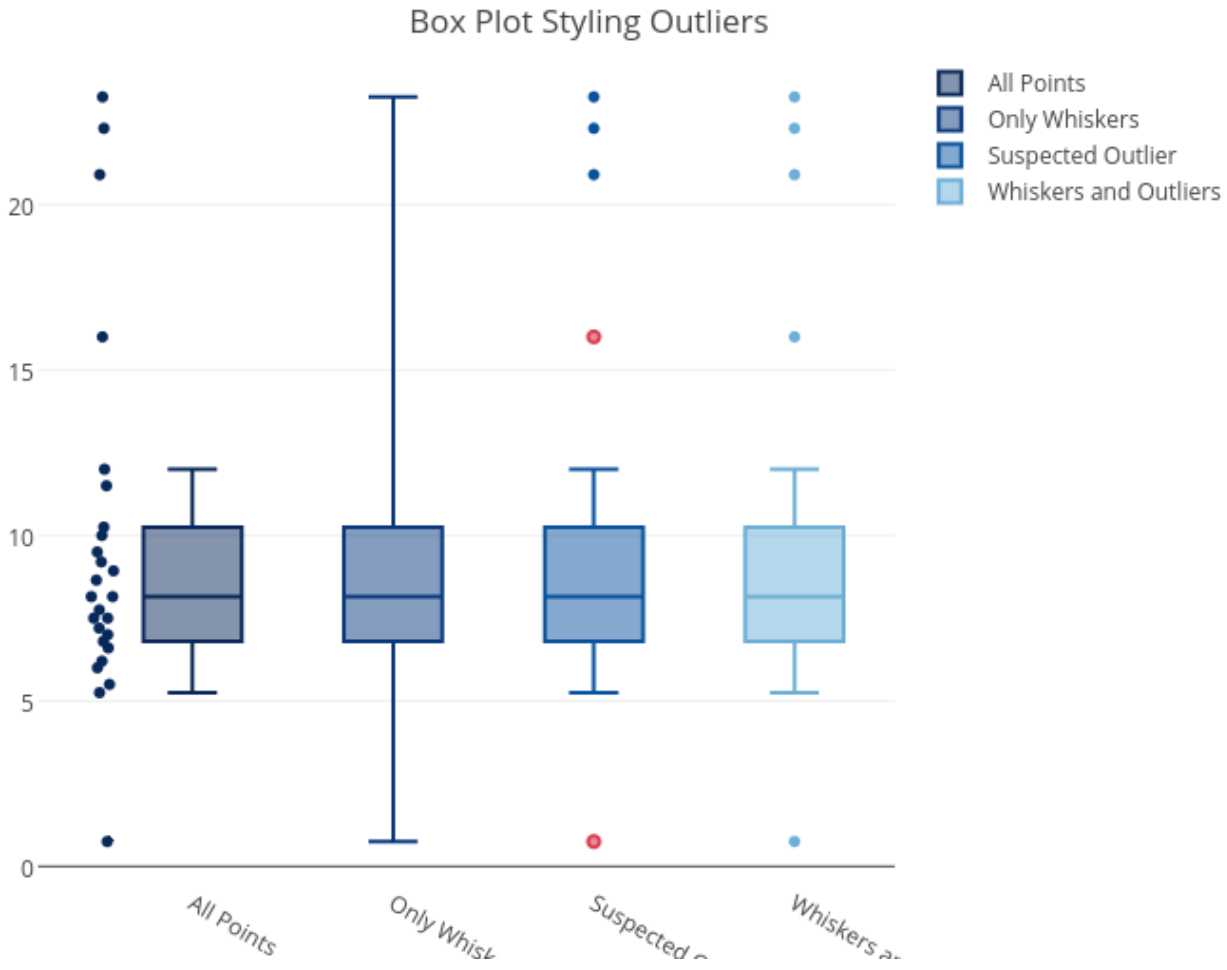
$c_1 = -0.13 \quad c_2 = -0.55 \qquad \widehat{\sigma}_1 = 1.02 \quad \widehat{\sigma}_2 = 0.67$



Taille des moustache et taille de l'échantillon

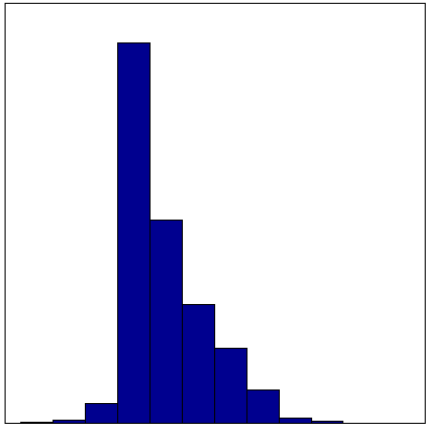
n petit
- min–max
- Tukey

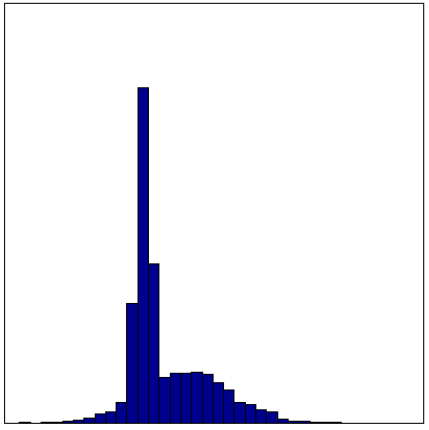
n grand
- 9% – 91 %
- 2% – 98 %

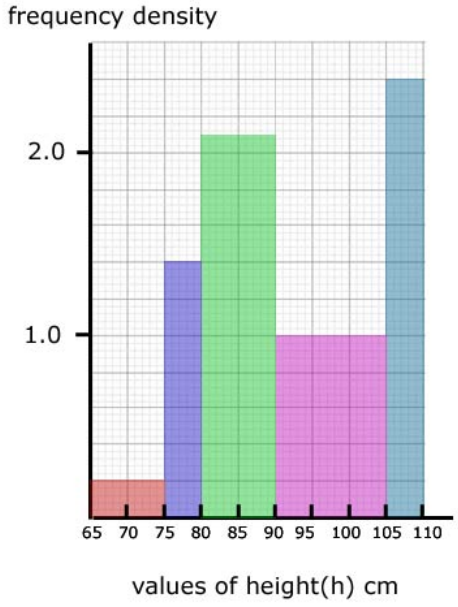


https://plot.ly/r/box-plots/

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.
- Dans le cas d'une variables continues : 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)

$$p \geq 1 + \log n \qquad \text{par exemple,} \qquad p = 1 + \frac{10}{3} \log_{10} n$$

- 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 des valeurs telle que le min $s_i \geq 5$.

Comment choisir les $\{b_i\}_{i=0,\ldots,p}$?

- équirépartition des individus par classe :

$b_0 = min$ $\qquad$ b_i calculé tel qu'il y ai θ observations entre b_i et b_{i+1}

- équirépartition des intervalles :

$$largeur = \frac{max - min}{p}; \qquad b_0 = min; \qquad b_i = b_0 + i\left(\frac{max - min}{p}\right)$$

Comment construire une histogramme (variable continue)

Construction d'un histogramme équiréparti :

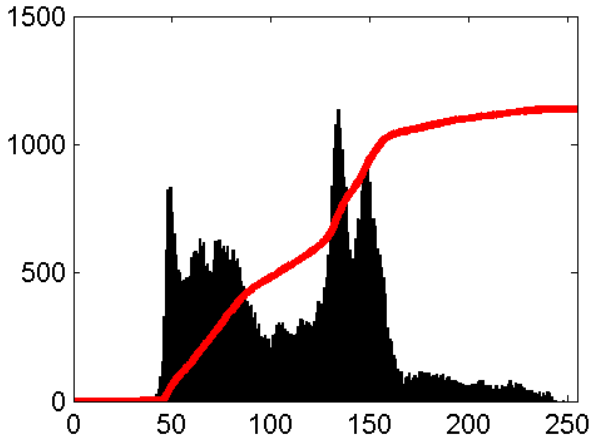
1. choisir p le nombre de classes,
2. calculer les $\{b_i\}_{i=0,\dots,p}$, les bornes des intervalles,
3. en déduire les s_i le nombre d'individus par classe. Si ce nombre est inférieur à cinq, on fusionne des classes.
4. calculer enfin les $h_i = \frac{s_i}{b_i - b_{i-1}}$ les hauteurs.

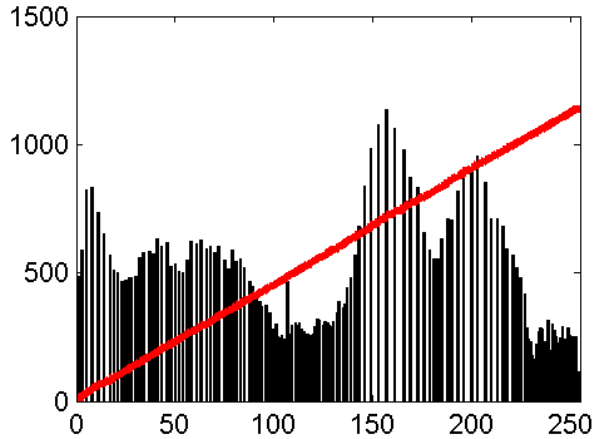
Par exemple pour 32 observations entre 0 et 15. On en déduit $p = 6$ et les valeurs suivantes :

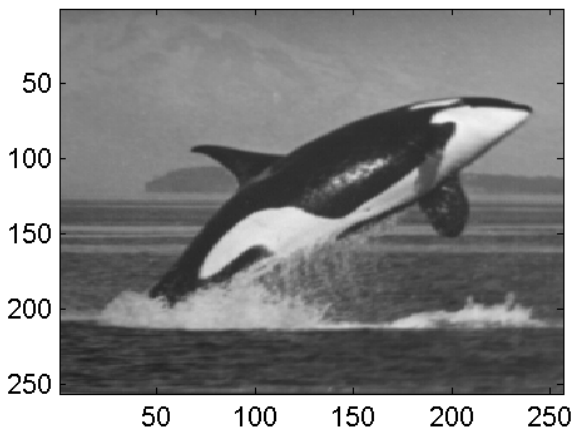
$$b_0 = 0 \quad b_1 = 2 \quad b_2 = 3 \quad b_3 = 4 \quad b_4 = 8 \quad b_5 = 10 \quad b_6 = 15$$
$$s_1 = 6 \quad s_2 = 5 \quad s_3 = 6 \quad s_4 = 5 \quad s_4 = 5 \quad s_5 = 5$$
$$h_1 = \frac{6}{2} \quad h_2 = 5 \quad h_3 = 6 \quad h_4 = \frac{5}{4} \quad h_5 = \frac{5}{2} \quad h_6 = \frac{5}{5}$$

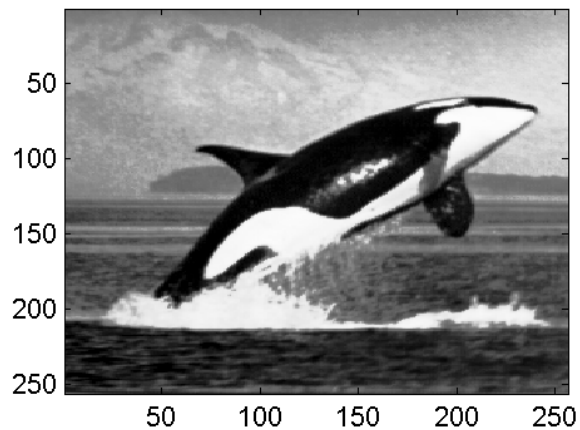
Histogramme et traitement d'images

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

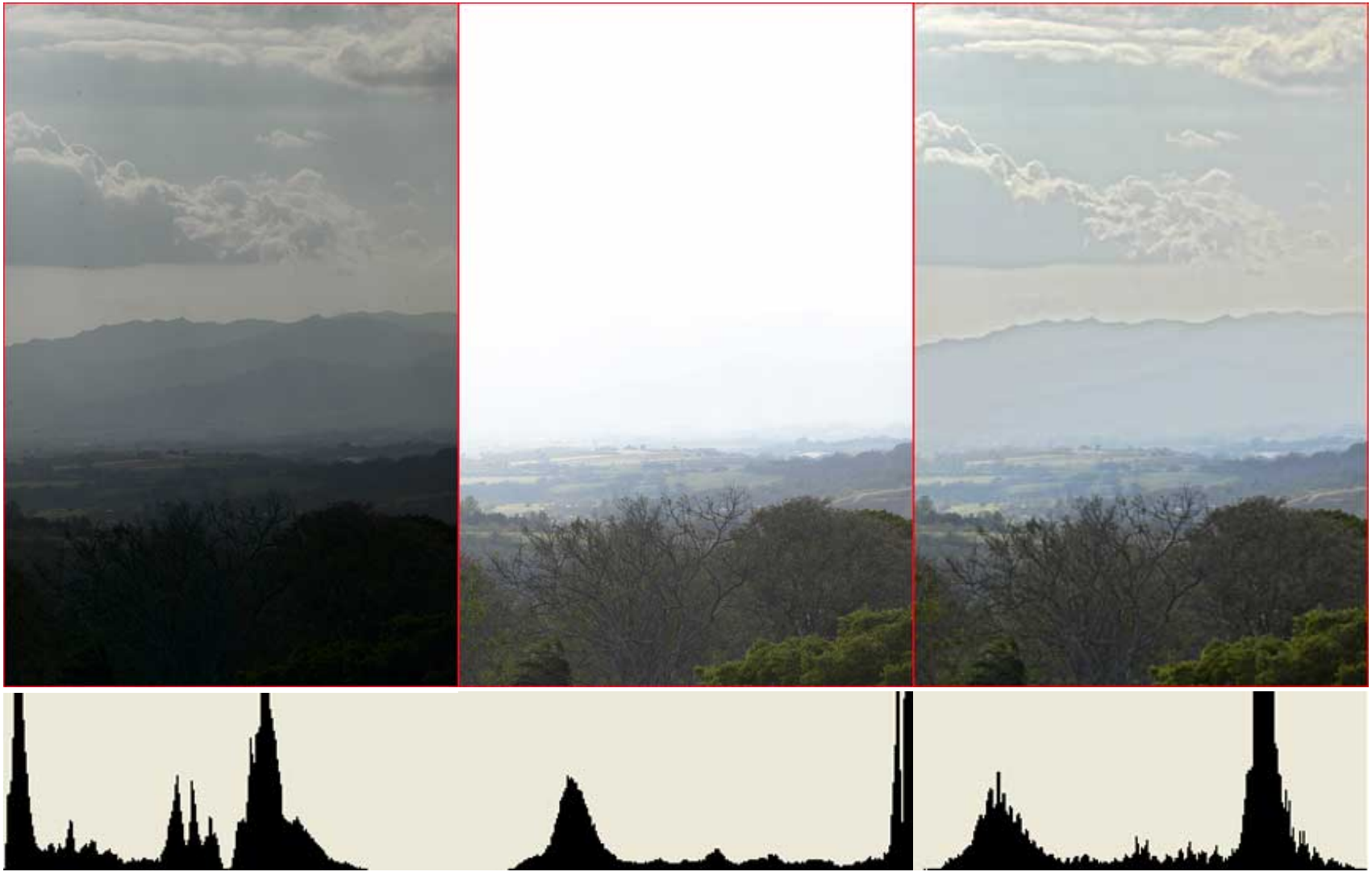


Figure : Exemple d'histogrammes `http://www.llvj.com/tutorials/understanding-series/understanding-histograms.shtml`.

Plan

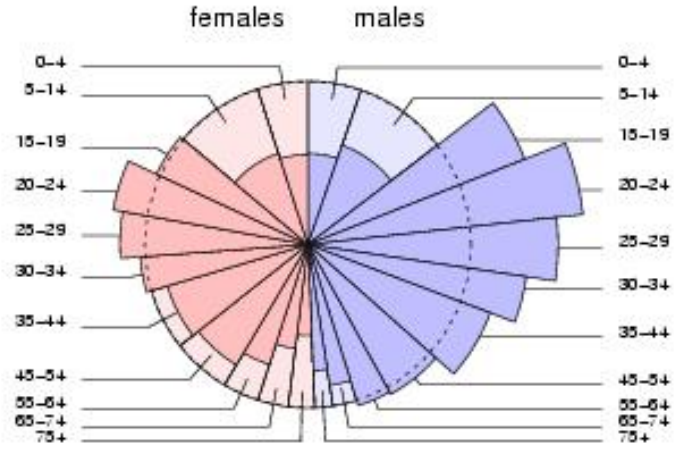
Empirique vs. OPM

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