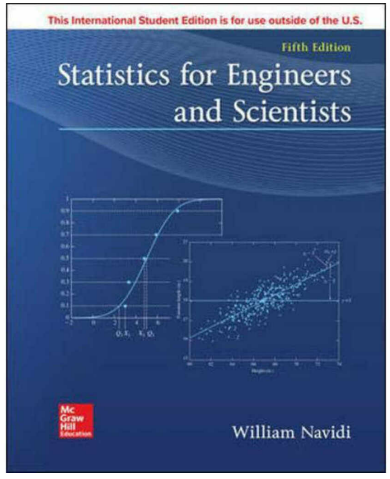
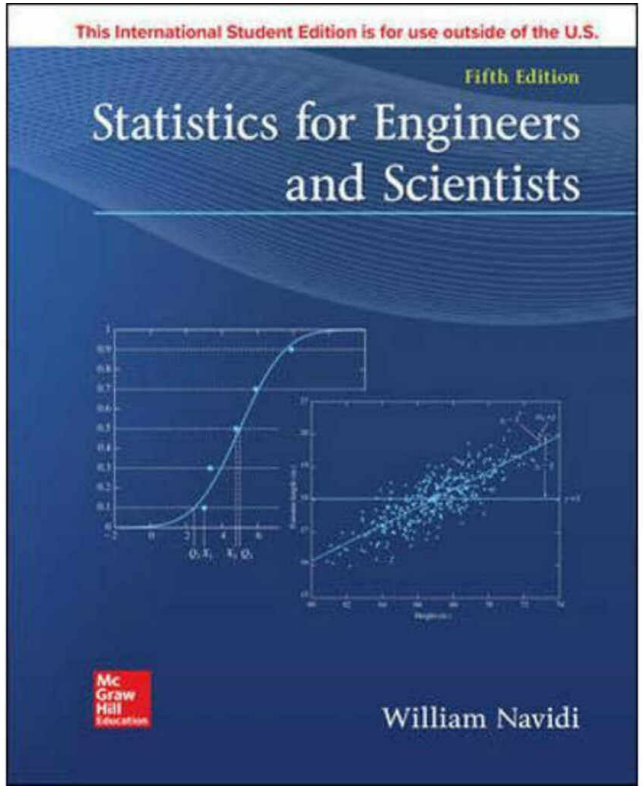
Statistiques pour l'Ingénieur
Description monovariable

Stéphane Canu
asi.insa-rouen.fr/enseignants/~scanu
scanu@insa-rouen.fr



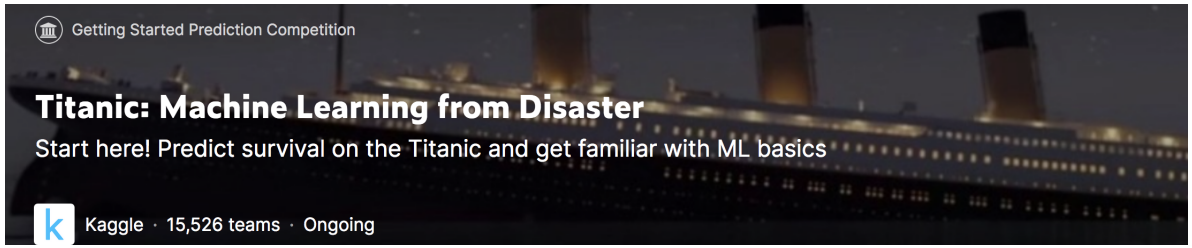
ITI 3, INSA Rouen Normandie, Janvier 2024

Lecture road map



https://moodle.insa-rouen.fr/course/view.php?id=93

L'exemple du Titanic



- Survived: Outcome of survival (0 = No; 1 = Yes)
- Pclass: Socio-economic class (1 = Upper; 2 = Middle; 3 = Lower)
- Name: Name of passenger
- Sex: Sex of the passenger
- Age: Age of the passenger (Some entries contain NaN)
- SibSp: Number of siblings and spouses of the passenger aboard
- Parch: Number of parents and children of the passenger aboard
- Ticket: Ticket number of the passenger
- Fare: Fare paid by the passenger
- Cabin Cabin number of the passenger (Some entries contain NaN)
- Embarked: Port of embarkation of the passenger (C = Cherbourg; Q = Queenstown; S = Southampton)

https://www.kaggle.com/c/titanic

Les trois distinctions

Definition (Domaine d'une variable)

Le **domaine d'une variable** est l'ensemble des valeurs que cette variable peut prendre

$\Omega_{genre} = \{M, F\}$

Definition (Variable discrète)

une **variable est discrète** si le cardinal de son domaine est dénombrable

genre, département, PCS, nombre de visites...

Definition (Variable quantitative)

une variable discrète est **quantitative** si l'ensemble de ses modalités est comparable [1]. Toutes les variables continues sont quantitatives.

age, nombre de visites...

---

[1] $\forall x, y$ deux modalités quelconques, soit $x < y$ soit $x \geq y$

Variables qualitatives

Une variable qualitative peut être observée mais non mesurée

Definition (Modalités)

On appelle **modalités** l'ensemble des valeurs distinctes que peut prendre la variable observée. C'est le domaine d'une variable discrète. On note:

$$\Omega = \{m_1, m_2, ..., m_i, ..., m_r\}$$

avec $r = card(\Omega)$.

Exemples de modalités:

- oui / non (variable binaire)
- un ensemble de variétés de maïs (variable qualitative)
- un peu / moyen / beaucoup (qualitative ordinale)
- nombre des personnes à la caisse d'un supermarché (quantitative)
- nombre de mois de chômage entre deux empois (quantitative)

Definition (Effectif)

On appelle **effectif** d'une modalité m_i le nombre de fois n_i où cette modalité est apparue dans l'échantillon.

Definition (Fréquence)

On appelle **fréquence** d'une modalité le rapport dans un échantillon du nombre de fois où cette modalité est apparue (le rapport de l'effectif, n_i) divisé par la taille de l'échantillon, n. Pour la modalité m_i, la fréquence $f_i = \frac{n_i}{n}$.

Definition (Probabilité – OPM)

La **probabilité** $\mathbb{P}(m_i)$ d'une modalité m_i est la limite de la fréquence observée lorsque la taille de l'échantillon tend vers l'infini[2].

---

[2] Rigoureusement, il faudrait définir les probabilités à priori et considérer que l'on observe un échantillon d'une variable aléatoire tirée selon cette loi de manière indépendante et identiquement distribué (pour plus de détails voir par exemple `http://fr.wikipedia.org/wiki/Probabilité`).

Variables quantitatives discrètes

Exemples :

- nombre de produits achetés par un client
- l'appréciation donnée par un spectateur à un film (1:5)[3]
- intervalle (transformation d'une variable continue)
- nombre des personnes à la caisse d'un supermarché

[3] http://www.netflixprize.com/

Variables quantitatives discrètes

Modalités ordonnées : effectif, effectif cumulé, fréquence et probabilité

Modalité (m_i)	effectifs n_i	N_i	$\widehat{f_j} = \frac{n_i}{n}$	$100 \times \widehat{F_j}$ (%)
24	1	1	$\frac{1}{38}$	2,63 %
26	2	3	$\frac{2}{38}$	7,89 %
29	3	6	$\frac{3}{38}$	15,79 %
31	2	8	$\frac{2}{38}$	21,05 %
33	4	12	$\frac{4}{38}$	31,58 %
35	3	15	$\frac{3}{38}$	39,47 %
37	3	18	$\frac{3}{38}$	47,37 %
38	1	19	$\frac{1}{38}$	50,5 %
41	6	25	$\frac{6}{38}$	69,79 %
43	3	28	$\frac{3}{38}$	73,68 %
45	1	29	$\frac{1}{38}$	76,32 %
46	4	33	$\frac{4}{38}$	86,84 %
49	5	38	$\frac{5}{38}$	100 %
Total	38	-	1	-

La notion d'effectif cumulé n'a pas de sens pour les variables qualititives

Fréquence cumulée

Definition (Effectif cumulé)

On appelle **effectif cumulé** d'une modalité : $N_i = \sum_{j, m_j \leq m_i} n_j$

Definition (**Fréquence cumulée**)

On appelle **fréquence cumulée** : $\widehat{F}_i = \sum_{j, m_j \leq m_i} f_j$

Propiétés:

- On a $N_r = n$ et $\widehat{F}_r = 1$.
- Les $\widehat{F}_i$ définissent la loi cumulative empirique $\widehat{F}_i = \widehat{F}(m_i) = \hat{\mathbb{P}}(X \leq m_i)$
- On appelle $h_i = n_{i+1} - n_i$ la distance inter modalités.
- Ces quantités ont un sens lorsque la variable est quantitative.

Exemple

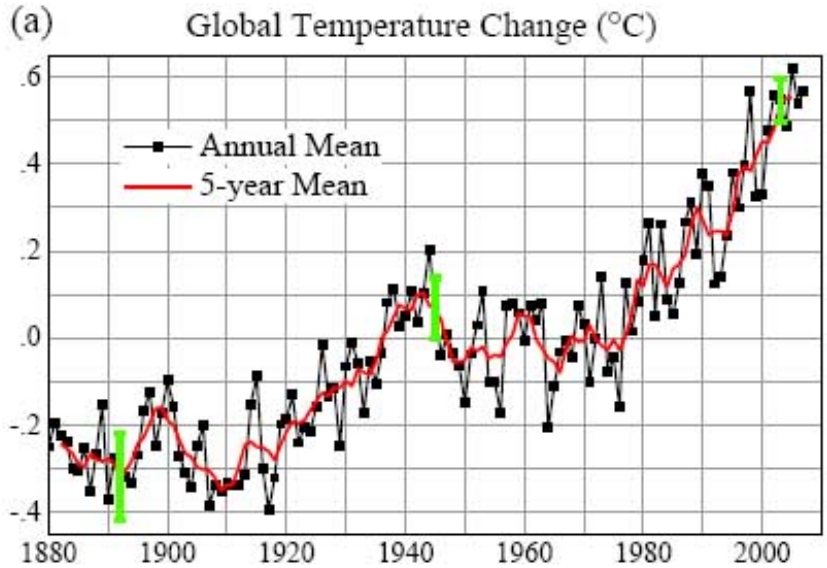
nom	*age*
louis	13
paul	35
joséphine	24
lucien	43
mike	56
pedro	27
..	...

$\Rightarrow$

modalité	*effectif*	*fréquence*	*eff. cumulé*	*fréq. cumulée*
13	381	$2,06$	381	2,05
14	576	$3,11$	957	5,17
15	441	$2,38$	1398	7,55
16	744	$4,02$	2142	11,57
17	553	$2,99$	2695	14,56
...	...	...	...	...
68	441	$2,38$	18509	100,00
total	18509	100		

Dans le cas continu on peut toujours définir la fonction de répartition empirique $\widehat{F}_i = \widehat{F}(m_i) = \hat{\mathbb{P}}(X \leq m_i)$

Variables quantitatives continues



Graphe des variations de températures annuelles du globe terrestre par rapport à la période de référence 1951-1980 (Air and ocean data from weather stations, ships and satellites)[4].

Exemples :

- la température ($\Omega = [-273.15, +\infty]$)
- concentration, intensité, prix, poids, taille, distance...
- rapport, proportion ($\Omega = [0, 1]$)

[4] `http://www.nasa.gov/topics/earth/features/earth_temp.html`

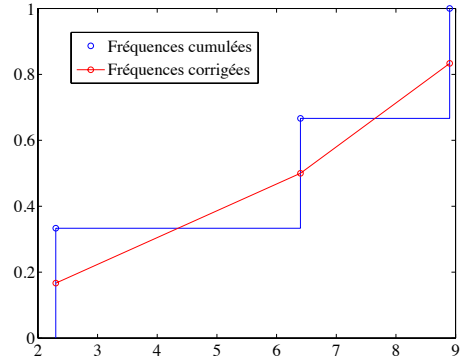
Fonction de répartition empirique

Définition : Fréquences corrigées : $\widehat{F}_i^c = \widehat{F}_i - \frac{1}{2}(\widehat{F}_i - \widehat{F}_{i-1})$:

$\rightarrow$ c'est le centre de l'intervalle $[\widehat{F}_{i-1}, \widehat{F}_i]$

Exemple : $n = 3$ observations $\rightarrow$ (6,4, ; 2,3 ; 8,9)

x_i	2,3	6,4	8,9
f_i	1/3	1/3	1/3
F_i	1/3	2/3	3/3 = 1
F_i^c	1/6	1/2	5/6



Definition (fonction de répartition empirique)

On appelle fonction de répartition empirique d'un échantillon pour la variable X la fonction $\widehat{F}_X$ définie sur l'intervalle $[x_1, x_n]$ par :

$$\widehat{F}_X(x) = \begin{cases} 0 & \text{si } x < x_1 \\ \widehat{F}_{i-1}^c + \dfrac{\widehat{F}_i^c - \widehat{F}_{i-1}^c}{x_i - x_{i-1}}(x - x_{i-1}) & \text{si } x_{i-1} \leq x < x_i, \qquad i = 2, n \\ 1 & \text{si } x \geq x_n \end{cases}$$

Exemples de fonctions de répartition empirique

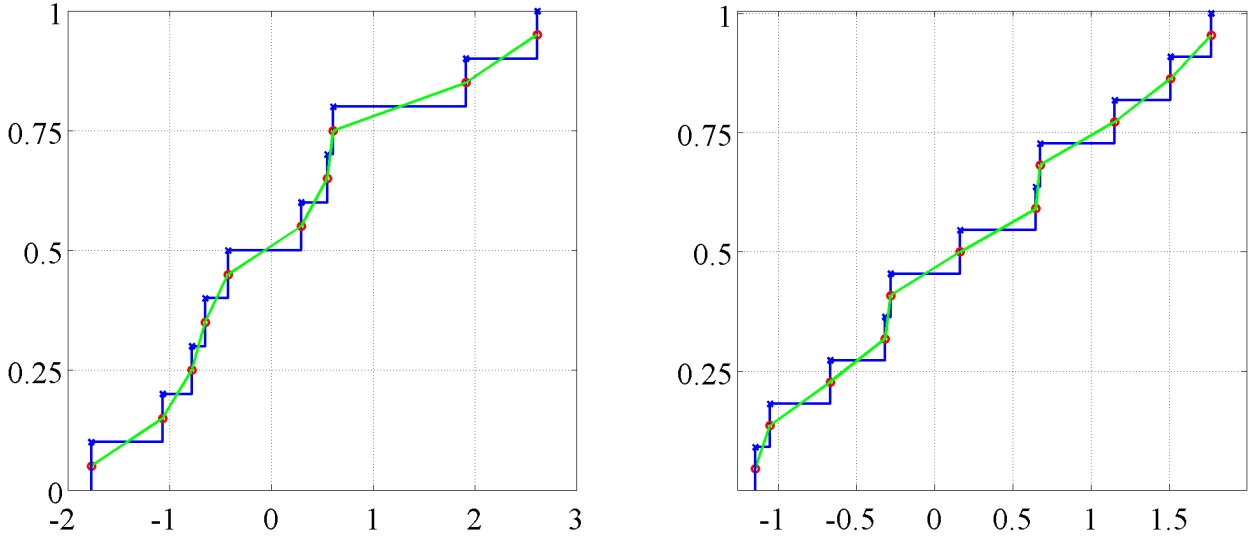


Figure: Fréquences cumulées (en bleue) et de fonction de répartition empirique (courbe verte) pour 10 observations (à gauche) et 11 observations (à droite).

```
x = np.array([-1.14, -1.05, -0.66, -0.31, -0.28, 0.16, 0.64, 0.67, 1.15, 1.51, 1.76])
n = x.shape[0]
F = np.arange(1, n+1, 1, dtype=int)/n  # frequences cumulees
Fc = F - 1/2*(F - np.append(0, F[1:n])) # fonct. de repartition empirique
plt.plot(x,Fc)
```

Fonctions de répartition : discret vs. continue

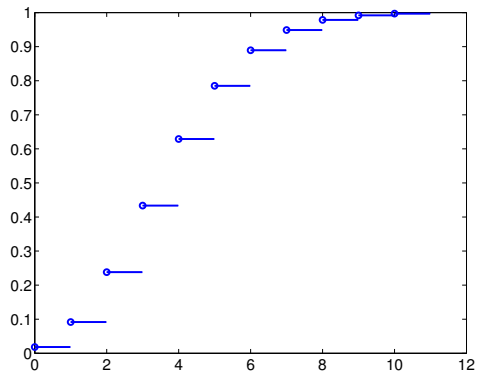
modalité	*eff.*	$\widehat{f}$	*e. c.*	$\widehat{F}$
]0, 18]	1405	7,59	1405	7,59
]18, 29]	1875	10,13	3280	17,72
]30, 39]	1172	6,33	4452	24,05
]40, 49]	3047	16,46	7499	40,51
]50, 59]	4920	26,58	12419	67,09
60 $\leq$	4920	32,91	18510	1000
total	18509	100		

x_i	eff. n_i	N_i	$\widehat{f}_j = \frac{n_i}{n}$	$100 \times \widehat{F}_j$
24	1	1	$\frac{1}{6}$	16,67 %
26	1	2	$\frac{1}{6}$	33,33%
29	1	3	$\frac{1}{6}$	50 %
31	1	4	$\frac{1}{6}$	66,67%
33	1	5	$\frac{1}{6}$	83,33%
35	1	6	$\frac{1}{6}$	100 %
Total	6	-	1	-

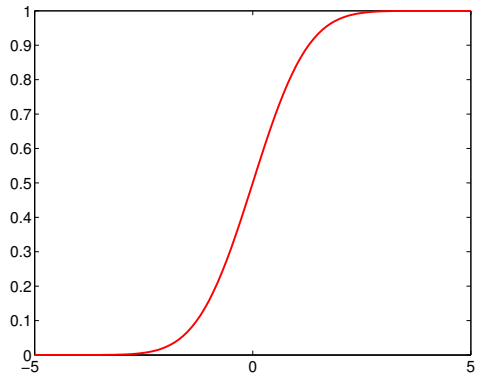
Definition (Fonction de répartition (l'OPM))

La fonction de répartition de la variable aléatoire réelle X est la fonction F_X qui à tout réel x associe

$$F_X(x) = \mathbb{P}(X \leq x)$$



$F_X(x)$ discret



$F_X(x)$ continu

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 :

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

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

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

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 :

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

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

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

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

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

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

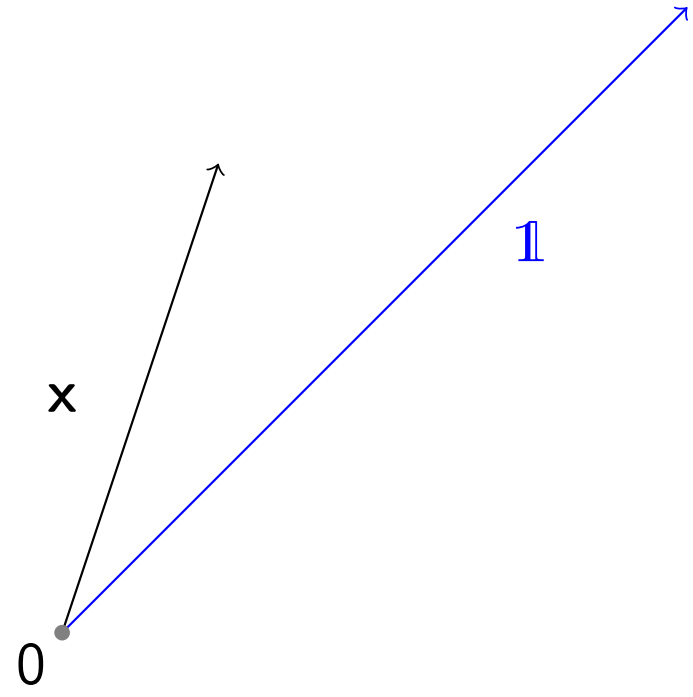
C'est la moyenne empirique : $c = \frac{1}{n} \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} \frac{1}{n} x_i = \sum_{i=1}^{n} \hat{f}_i x_i$

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

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

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



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

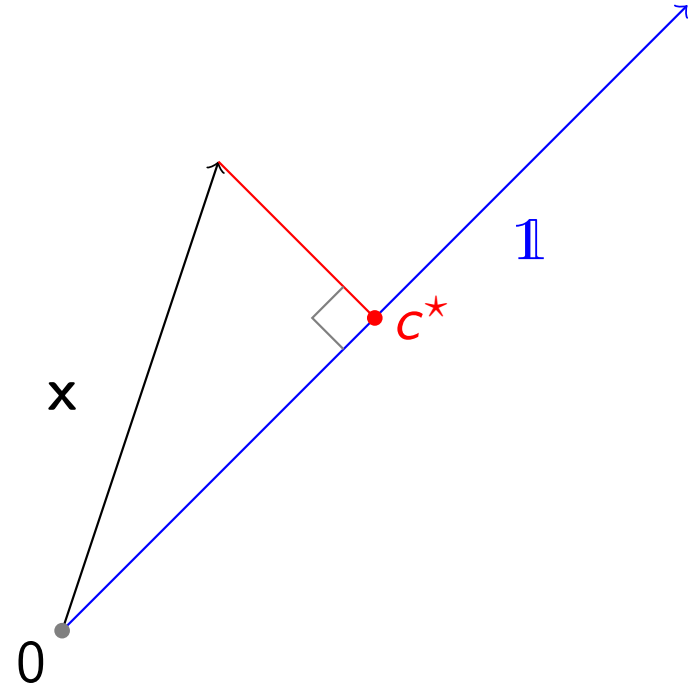
$c^\star$, la moyenne est aussi la projection orthogonale du vecteur des observations sur le vecteur $\mathbb{1}$

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

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 $\mathbb{1}$

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

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

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 ca 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,

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

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	$\frac{11}{12}$

$\widehat{F}_X(3) = 0,42$ $\quad$ $\widehat{M} = 3,5$

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

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 (OPM). 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

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

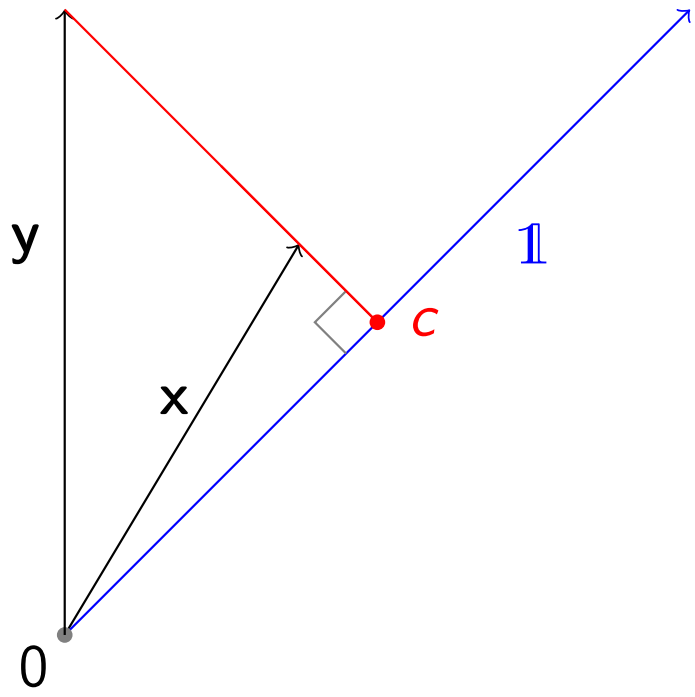
- mode : les fréquences / les probabilités

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

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 :

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

- Écart: $e_i = (x_i - \overline{x})$ 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 :

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

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

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

Pire des cas : Tchebychev

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

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

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

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

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

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

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

Cas Gaussien

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

MAD

la médiane des écarts absolus

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

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

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.

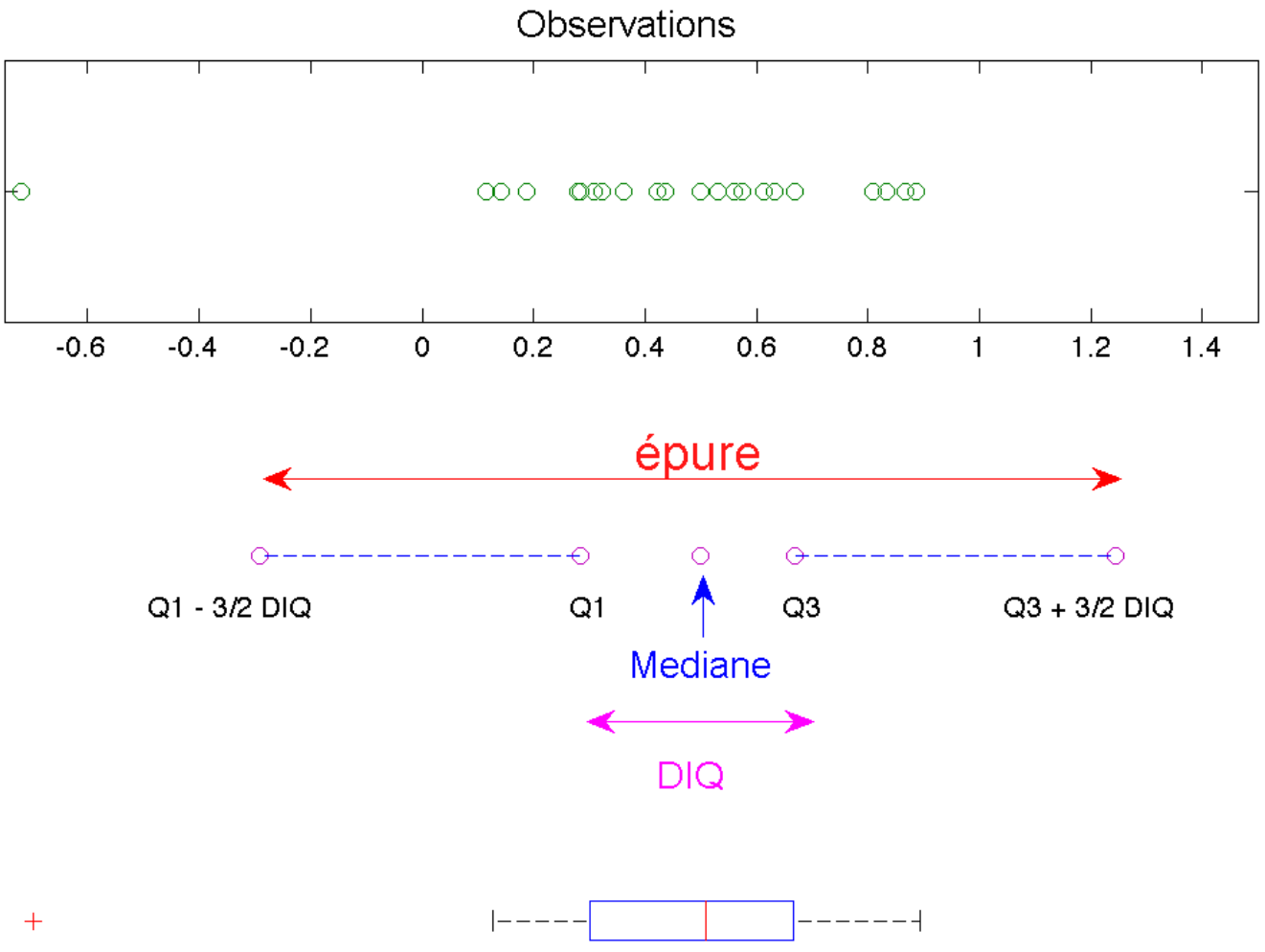
- $\widehat{\Phi}_{\frac{1}{4}} = \widehat{Q}_1$, telle que $\widehat{F}(\widehat{Q}_1) = \frac{1}{4}$,
- $\widehat{\Phi}_{\frac{1}{2}} = \widehat{Q}_2 = \widehat{M}$, telle que $\widehat{F}(\widehat{M}) = \frac{1}{2}$,
- $\widehat{\Phi}_{\frac{3}{4}} = \widehat{Q}_3$, telle que $\widehat{F}(\widehat{Q}_2)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

Definition (Distance inter quartile (DIQ))

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

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



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 $\circ$.

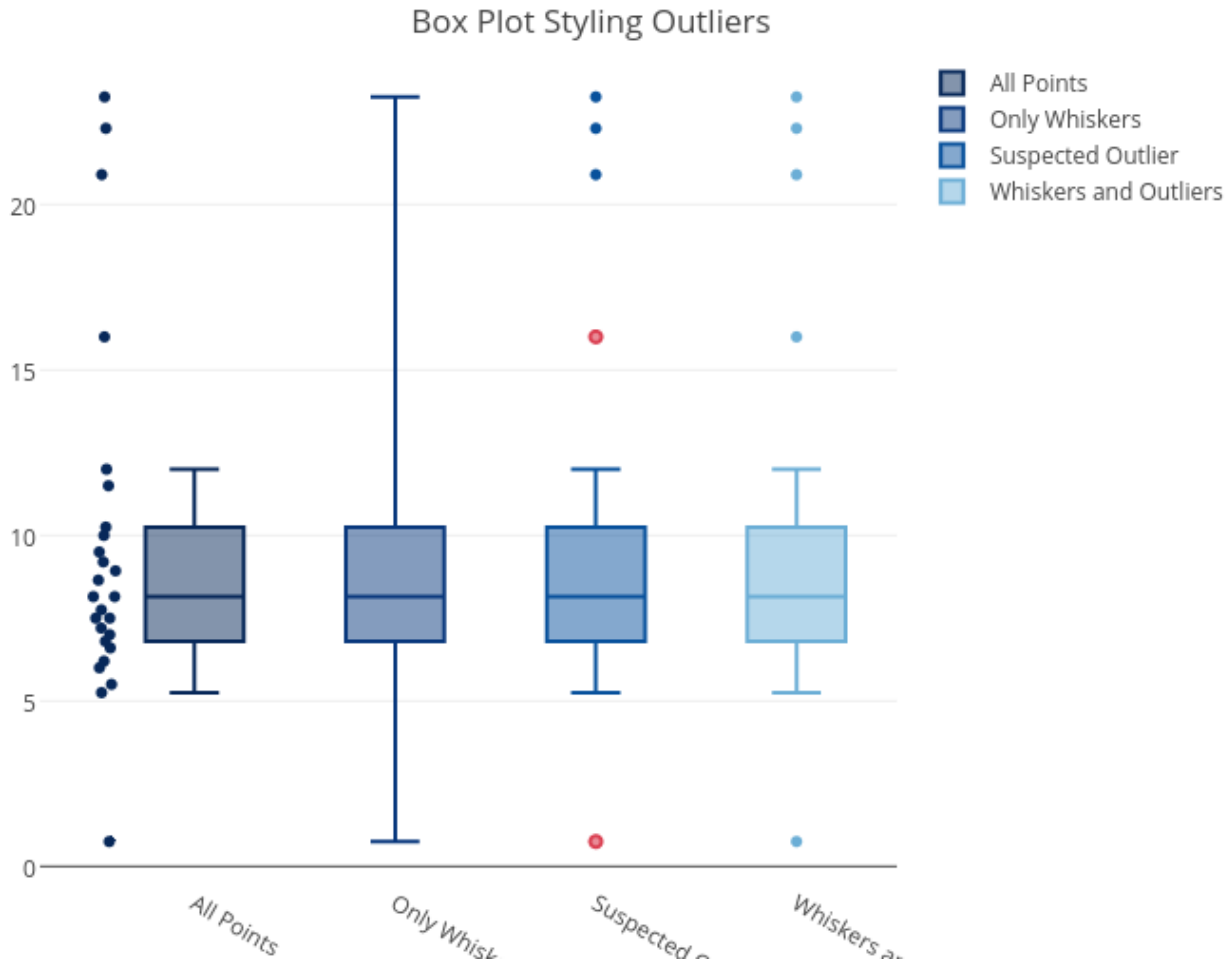
Taille des moustache et taille de l'échantillon

n petit

- min–max
- Tukey(1969)

n grand

- 9% – 91 %
- 2% – 98 %



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

Boxplot in python

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

url="https://fxjollois.github.io/cours-2016-2017/donnees/tips.csv"
dataframe=pd.read_csv(url)
```

avec Pandas

```
boxplot = dataframe.boxplot()
```

avec matplotlib

```
x = dataframe.tip
fig1, ax1 = plt.subplots()
ax1.boxplot(x)
```

avec seaborn

```
ax = sns.boxplot(x=x)
```

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

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

Les moments d'ordre supérieur

Moments : $m_k = \frac{1}{n}\sum_{i=1}^{n} x_i^k, \qquad k = 1, \ldots, \infty$

Moments centrées : $\mu_k = \frac{1}{n}\sum_{i=1}^{n} (x_i - \overline{x})^k, \qquad k = 1, \ldots, \infty$

Le coefficient d'asymétrie (skewness en anglais) correspond à une mesure de l'asymétrie de la distribution

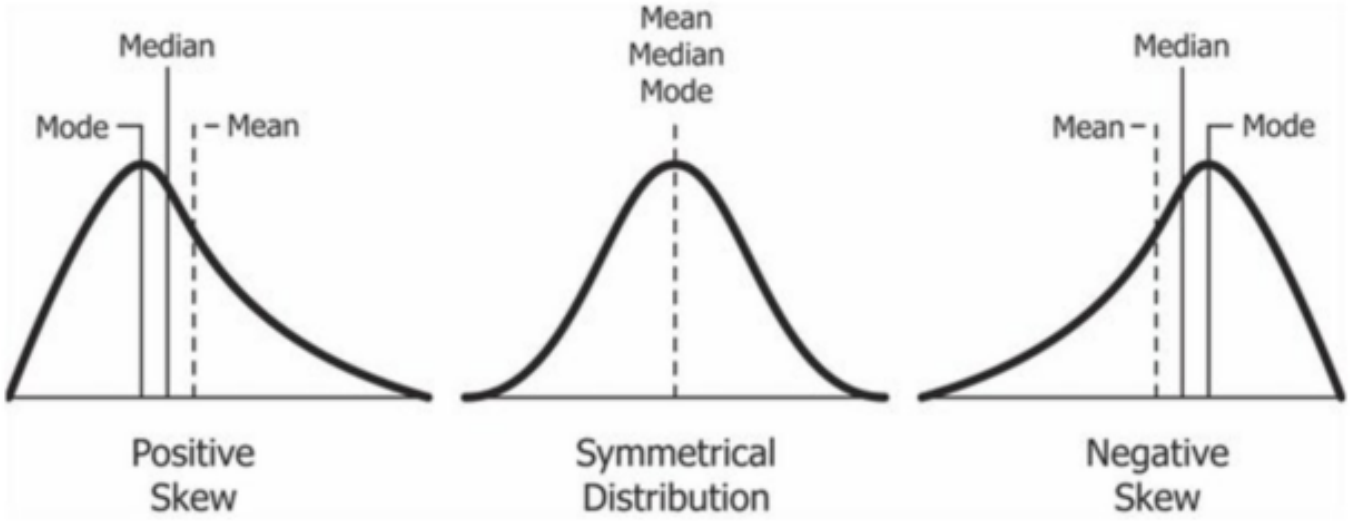
$$\widehat{\gamma}_1 = \frac{\mu^3}{\mu_2^{3/2}} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^3}{\widehat{\sigma}^3} \qquad \gamma_1 = \mathbb{E}\left[\left(\frac{X - \mathbb{E}(X)}{\sigma}\right)^3\right]$$

Le coefficient d'aplatissement (Kurtosis en anglais)

$$\widehat{\beta}_2 = \frac{\mu^4}{\mu_2^2} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^4}{\widehat{\sigma}^4} \qquad \beta_2 = \mathbb{E}\left[\left(\frac{X - \mathbb{E}(X)}{\sigma}\right)^4\right]$$

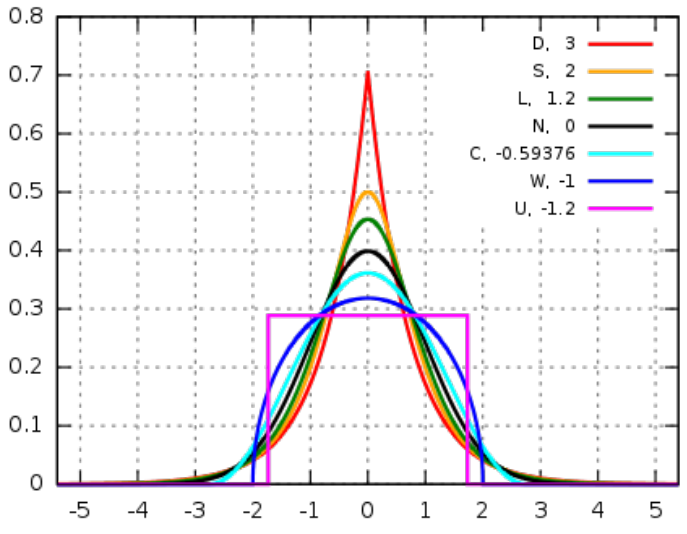
Asymétrie et aplatissement

Asymétrie



Aplatissement (Kurtosis normalisé = excès d'aplatissement = $\beta_2 - 3$)

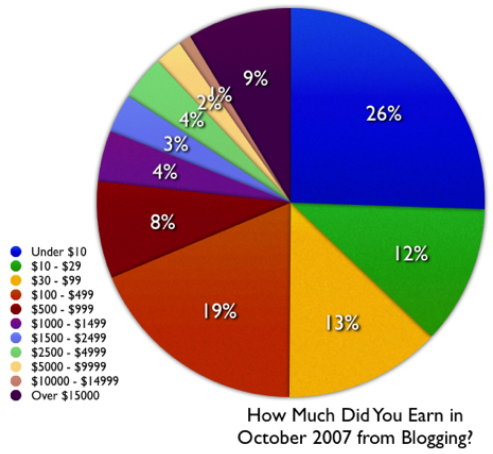
Loi de probabilité	Kurtosis normalisé	Symbole dans la figure	Couleur dans la figure
Loi de Laplace	3	D	Courbe rouge
Loi sécante hyperbolique	2	S	Courbe orange
Loi logistique	1,2	L	Courbe verte
Loi normale	0	N	Courbe noire
Loi du cosinus surélevé	-0,593762...	C	Courbe cyan
Loi triangulaire	-0,6		
Loi du demi-cercle	-1	W	Courbe bleue
Loi uniforme continue	-1,2	U	Courbe magenta

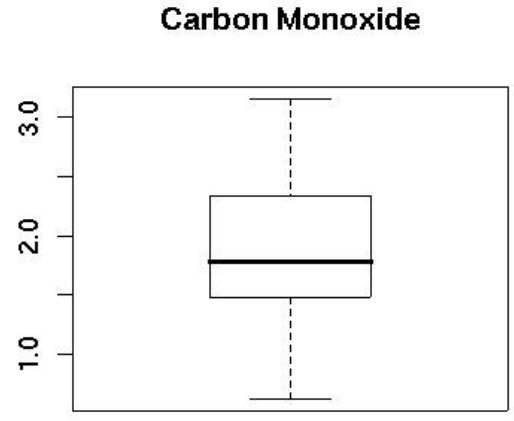


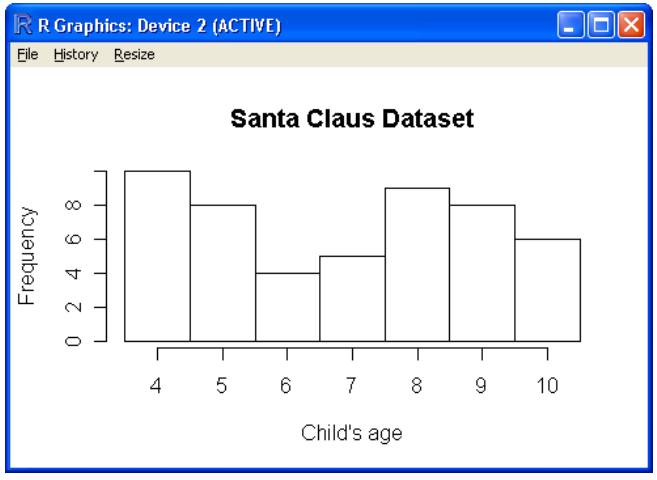
Plan

Résumé graphique des données

- Qualitatives : camemberts
- Quantitatives
  - Continues : boite à moustache
  - Discrètes : histogramme

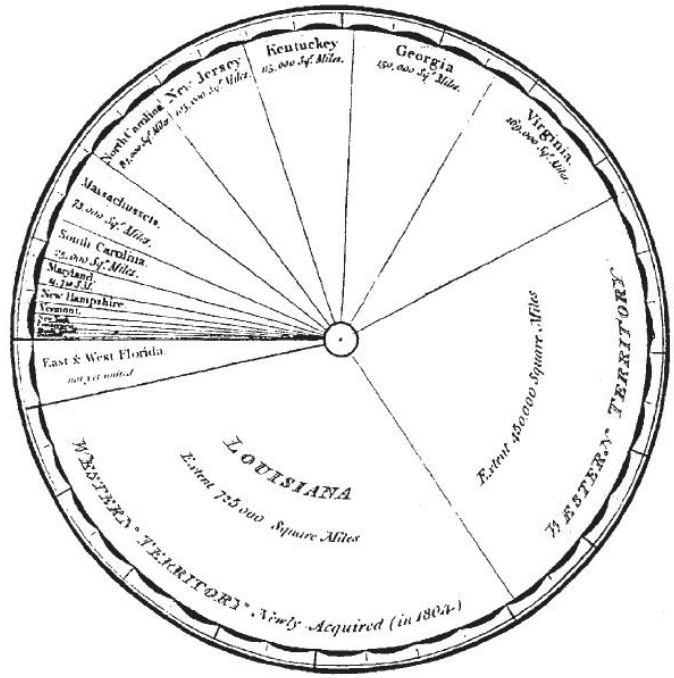






Cas des variables qualitatives

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 2.

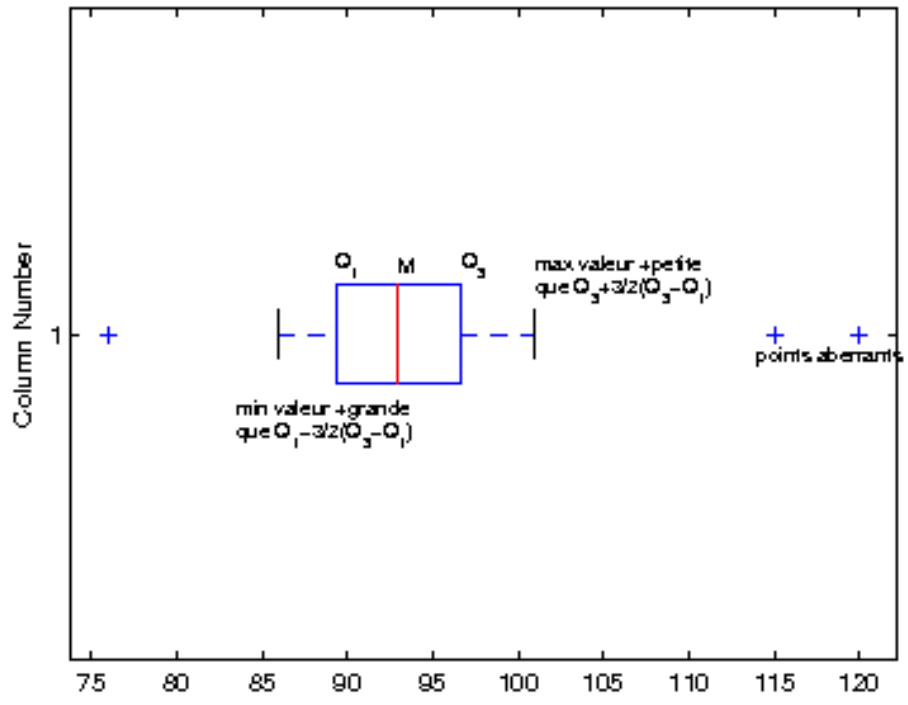
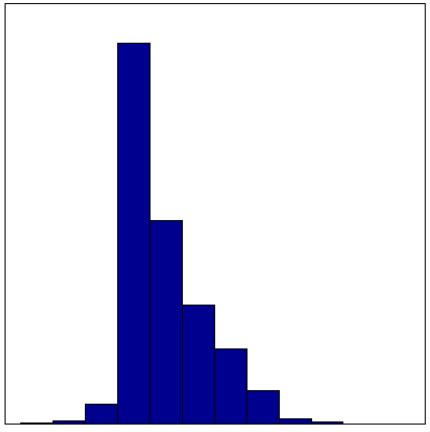
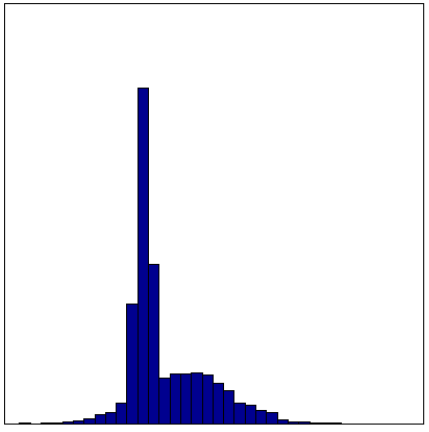
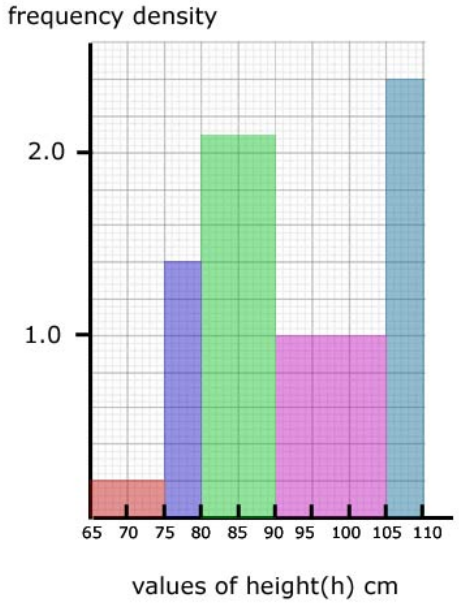


Figure: Exemple de boite à moustache

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 ?

- trouver p par équi-répartition des valeurs telle que le min $s_i \geq 5$. $p \geq 1 + \log n$ par exemple, $p = 1 + \frac{10}{3} \log_{10} n$
- la règle de Scott: $p = \frac{3,5\widehat{\sigma}}{n^{1/3}}$
- la règle de Freedman Diaconis $p = 2\frac{DIQ}{n^{1/3}}$

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

- équirépartition des individus par classe :

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

- équirépartition des intervalles :

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

Comment construire une histogramme (variable continue)

Construction d'un histogramme équiréparti :

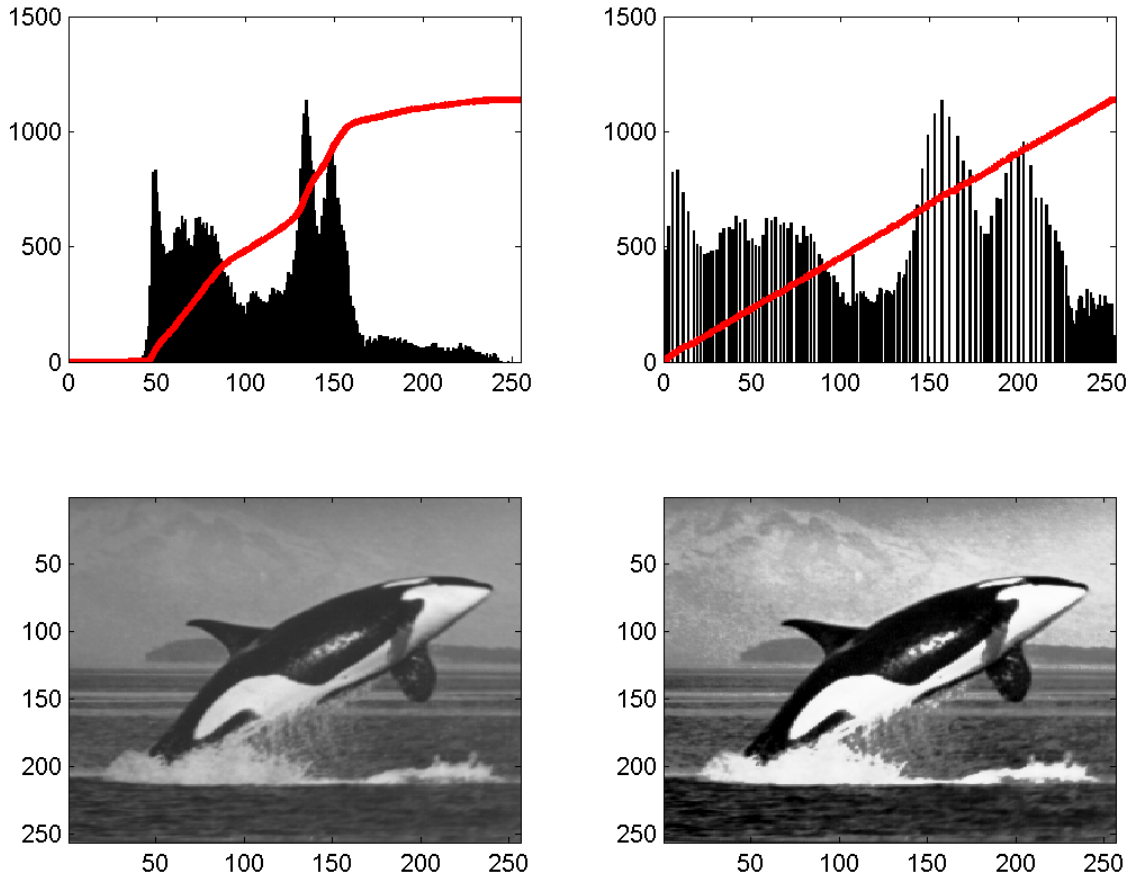
1. choisir p le nombre de classes,
2. calculer les $\{b_i\}_{i=0,\dots,p}$, les bornes des intervalles,
3. en déduire les s_i le nombre d'individus par classe. Si ce nombre est inférieur à cinq, on fusionne des classes.
4. calculer enfin les $h_i = \frac{s_i}{b_i - b_{i-1}}$ les hauteurs.

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[5] : $b_n(i) = \widehat{F}^{-1}\left(\frac{i}{p}\right)$



[5] http://en.wikipedia.org/wiki/Histogram_equalization

Histogramme et traitement d'images

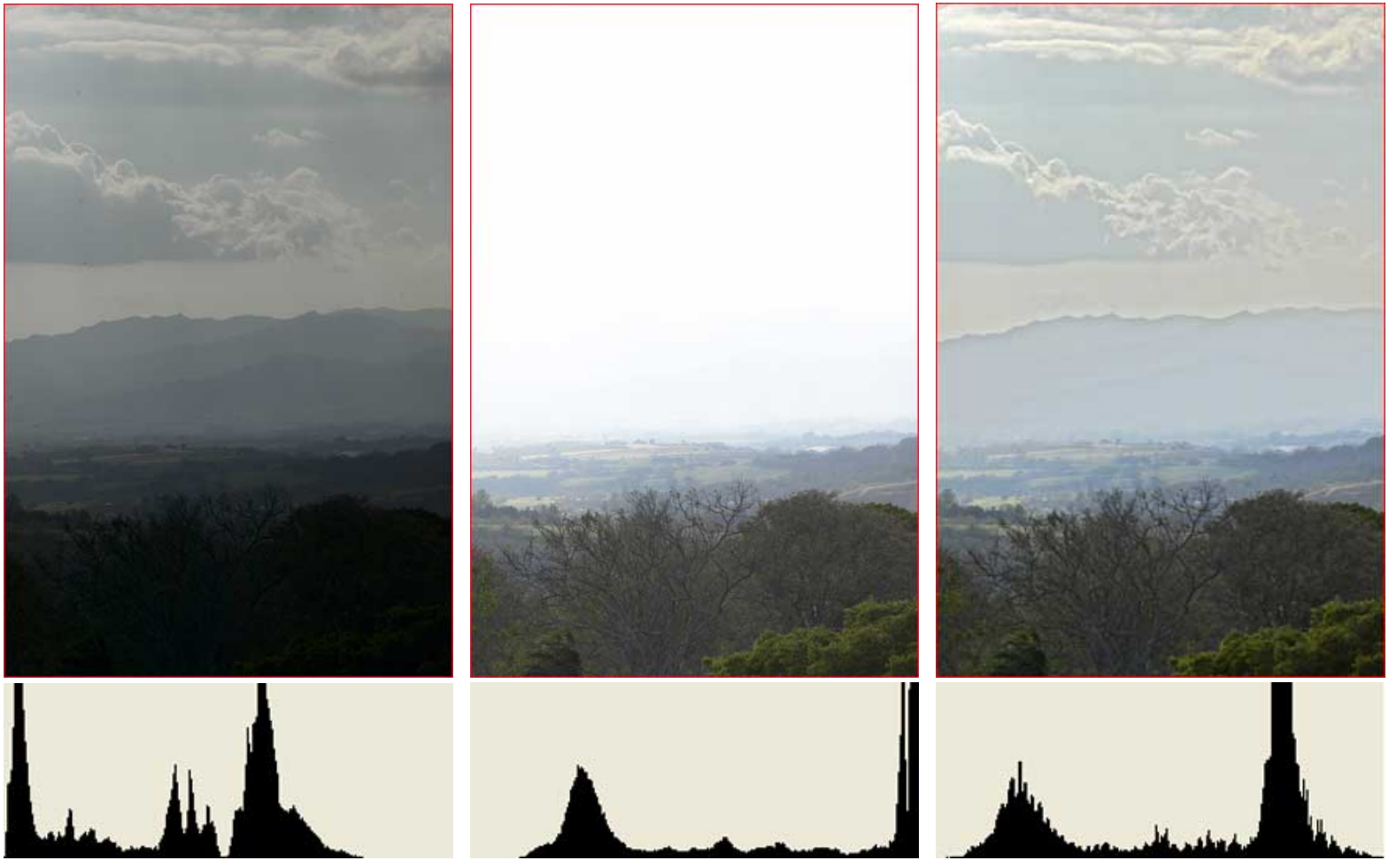


Figure: Exemple d'histogrammes `http://www.llvj.com/tutorials/understanding-series/understanding-histograms.shtml`.

Plan

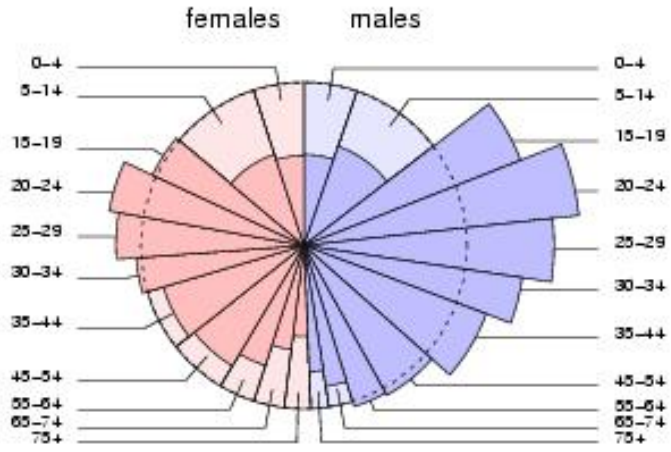
Empirique vs. OPM

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