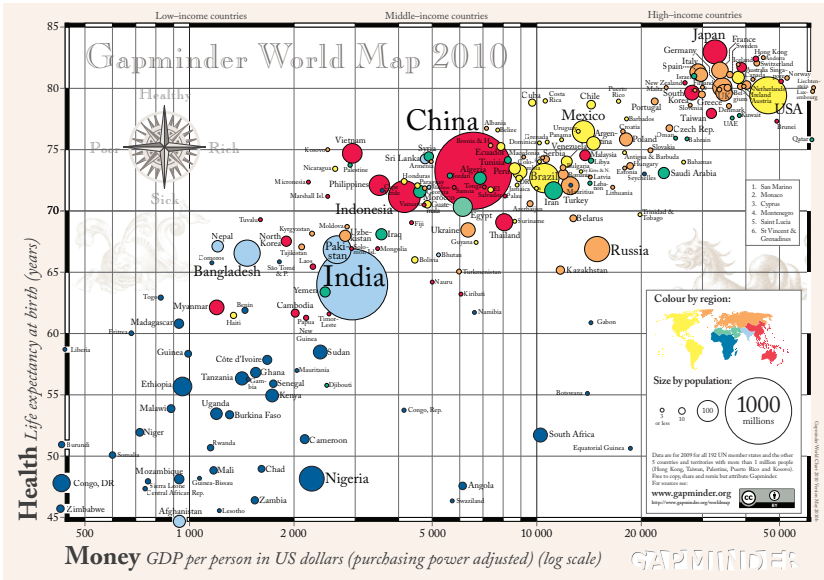
Statistiques descriptives d'un couple de variables

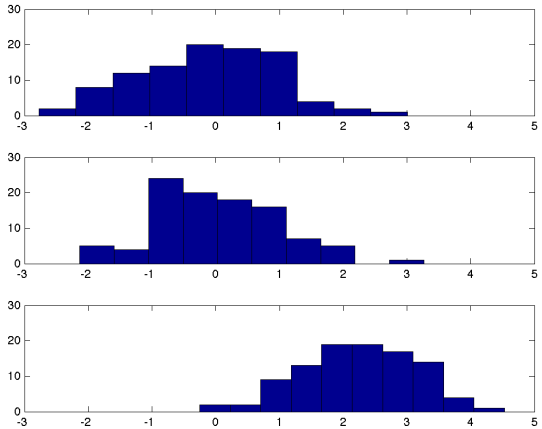
Benoit Gaüzère, Stéphane Canu
benoit.gauzere@insa-rouen.fr

INSA Rouen Normandie - ITI

February 6, 2024

Exemples de couples de variables





Les distances de freinage et les consommations d'alcool

Pourquoi étudier un couple de variables ?

- Il y a t'il des couples aberrants ?
- Ces deux variables sont elles liées ou sont elles indépendantes ?
  - Quelle est la nature de cette dépendance ?

Plan

Deux variables discrètes : tableau de contingence

Considérons un échantillon où chaque observation est une vente, décrite par deux variables qualitatives : le vendeur et le produit.

Vendeur	Produit
Bob	un pc
Percy	un aspirateur
Percy	une télé
Bob	une télé
Percy	une télé
Bob	un mac
John	un mac
Bob	un lave vaisselle

$\Rightarrow$

Deux variables discrètes : tableau de contingence

Considérons un échantillon où chaque observation est une vente, décrite par deux variables qualitatives : le vendeur et le produit.

Vendeur	Produit
Bob	un pc
Percy	un aspirateur
Percy	une télé
Bob	une télé
Percy	une télé
Bob	un mac
John	un mac
Bob	un lave vaisselle

$\Rightarrow$

	Bob	Percy	John
télévision	1	2	0
ordinateurs	2	0	1
aspirateurs	0	1	0
lave vaisselle	0	1	0

Tableau de Contingence

Deux variables discrètes : Tableau de contingence

Définition : Tableau de contingence

Soit X une variable discrète à ℓ modalités et Y une variable discrète à k modalités. Soit

$$S_n = \{(x_1, y_1), (x_2, y_2), ..., (x_i, y_i), ..., (x_n, y_n)\}$$

un échantillon de n observations.
On appelle **Tableau de contingence** de l'échantillon S_n le tableau N de ℓ lignes et k colonnes dont le terme général N_{ij} désigne le nombre d'individus présentant conjointement les modalités a_i pour X et b_j pour Y.

Effectifs marginaux d'un tableau de contingence

Définition : Effectif marginal

L'effectif marginal $n_{i\bullet}$ $(n_{\bullet j})$ d'une modalité d'un tableau de contingence de taille n est le nombre d'individus dans l'échantillon pour lesquels on a observé la modalité a_i (b_j)

- Ligne : $n_{i\bullet} = \sum_{j=1}^{k} n_{ij}$
- Colonne : $n_{\bullet j} = \sum_{i=1}^{\ell} n_{ij}$
- Total : $n = \sum_{i=1}^{\ell} \sum_{j=1}^{k} n_{ij}$

Modalités	b_1	b_2	$\cdots$	b_j	$\cdots$	b_k	marge
a_1	n_{11}	n_{12}	$\cdots$	n_{1j}	$\cdots$	n_{1k}	$n_{1\bullet}$
$\vdots$	$\vdots$					$\vdots$	$\vdots$
a_i	n_{i1}	n_{i2}	$\cdots$	n_{ij}	$\cdots$	n_{ik}	$n_{i\bullet}$
$\vdots$						$\vdots$	$\vdots$
a_ℓ	$n_{\ell 1}$	$n_{\ell 2}$	$\cdots$	$n_{\ell j}$	$\cdots$	$n_{\ell k}$	$n_{\ell\bullet}$
marge	$n_{\bullet 1}$	$n_{\bullet 2}$	$\cdots$	$n_{\bullet j}$	$\cdots$	$n_{\bullet k}$	$n_{\bullet\bullet} = n$

Fréquences

Définition : Fréquence

La fréquence $f_{ij} = \frac{n_{ij}}{n}$, c'est un estimateur du probabilité jointe $\mathbb{P}(a_i \text{ et } b_j)$.

Par construction, nous avons $\sum_{i=1}^{\ell} \sum_{j=1}^{k} f_{ij} = 1$ et $0 \leq f_{ij} \leq 1$.

Profils

Définition : Profil Ligne et Profil Colonne

On appelle **Profil Ligne** les nombres $l_{ij} = \frac{n_{ij}}{n_{i\bullet}}$ et **Profil Colonne** les nombres $c_{ij} = \frac{n_{ij}}{n_{\bullet j}}$.

Remarques

- Les f_{ij} définissent une distribution de probabilité empirique suivant la loi jointe $\mathbb{P}(X = a_i, Y = b_j) = f_{ij}$
- Les **profils ligne** définissent des probabilités empiriques de la loi conditionnelle $\mathbb{P}(Y = b_j | X = a_i)$
- Les **profils colonne** les probabilités empiriques de la loi conditionnelle $\mathbb{P}(X = a_i | Y = b_j)$

Exemple de tableau de contingence

Marges, fréquences et profils

	Bob	Percy	John
TV	12	21	12
PC	21	12	14
VC	22	31	14
LV	4	17	5

Exemple de tableau de contingence

Marges, fréquences et profils

	Bob	Percy	John	**Marge**
TV	12	21	12	**45**
PC	21	12	14	**47**
VC	22	31	14	**67**
LV	4	17	5	**26**
Marge	**59**	**81**	**45**	**185**

Exemple de tableau de contingence

Marges, fréquences et profils

Tableau de Contingence avec Fréquences

	Bob	Percy	John	Fréquences
TV	$\frac{12}{185}$	$\frac{21}{185}$	$\frac{12}{185}$	$\frac{45}{185}$
PC	$\frac{21}{185}$	$\frac{12}{185}$	$\frac{14}{185}$	$\frac{47}{185}$
VC	$\frac{22}{185}$	$\frac{31}{185}$	$\frac{14}{185}$	$\frac{67}{185}$
LV	$\frac{4}{185}$	$\frac{17}{185}$	$\frac{5}{185}$	$\frac{26}{185}$
Fréquences	$\frac{59}{185}$	$\frac{81}{185}$	$\frac{45}{185}$	$\frac{185}{185} = 1$

Exemple de tableau de contingence

Marges, fréquences et profils

- Profils colonnes:

	Bob	Percy	John
TV	$\frac{12}{59}$	$\frac{21}{81}$	$\frac{12}{45}$
PC	$\frac{21}{59}$	$\frac{12}{81}$	$\frac{14}{45}$
VC	$\frac{22}{59}$	$\frac{31}{81}$	$\frac{14}{45}$
LV	$\frac{4}{59}$	$\frac{17}{81}$	$\frac{5}{45}$

- Profils lignes:

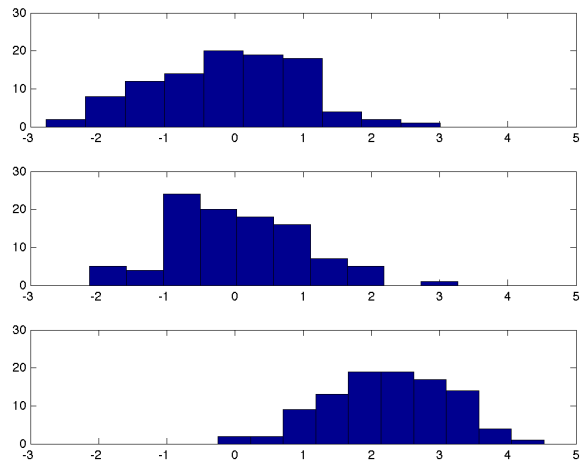
	Bob	Percy	John
TV	$\frac{12}{45}$	$\frac{21}{45}$	$\frac{12}{45}$
PC	$\frac{21}{47}$	$\frac{12}{47}$	$\frac{14}{47}$
VC	$\frac{22}{67}$	$\frac{31}{67}$	$\frac{14}{67}$
LV	$\frac{4}{26}$	$\frac{17}{26}$	$\frac{5}{26}$

Cas mixte : une variable discrète et l'autre continue

Performances de conduite pour trois groupes de personnes

Mesure de la distance de freinage pour 3 groupes de personnes : à jeun, a bu une bière (G1) et a bu trois verres de vin (G2).

Sujet	À jeun	G. 1	G. 2
1	22 m	21 m	27 m
2	20 m	24 m	24 m
⋮	⋮	⋮	⋮
n	19 m	22 m	31 m
moyenne	20,9	22,5	28,3
variance	4,32	5,18	5,87



- Effet de la variable qualitative ?
- Le codage des variables a un effet.

Tendance centrale

Moyenne

Couple des moyennes $(\overline{x}, \overline{y})$:

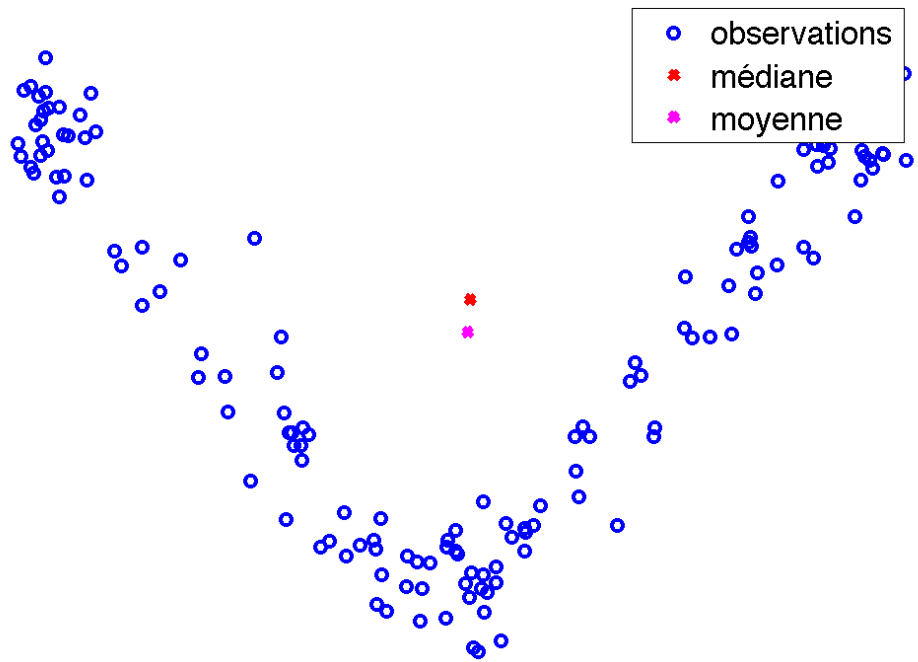
$$\min_{c_1,c_2} \sum_{i=1}^{n} (x_i - c_1)^2 + (y_i - c_2)^2 = \sum_{i=1}^{n} \|\mathbf{z}_i - \mathbf{c}\|^2 \qquad \text{avec : } \begin{array}{c} \mathbf{z}_i = (x_i, y_i) \\ \mathbf{c} = (c_1, c_2) \end{array}$$

Médiane

Couple des médianes (M_x, M_y) :

$$\min_{c_1,c_2} \sum_{i=1}^{n} |x_i - c_1| + |y_i - c_2| = \sum_{i=1}^{n} \|\mathbf{z}_i - \mathbf{c}\|_1 \qquad \text{avec : } \begin{array}{c} \mathbf{z}_i = (x_i, y_i) \\ \mathbf{c} = (c_1, c_2) \end{array}$$

Tendance centrale : illustration 2d



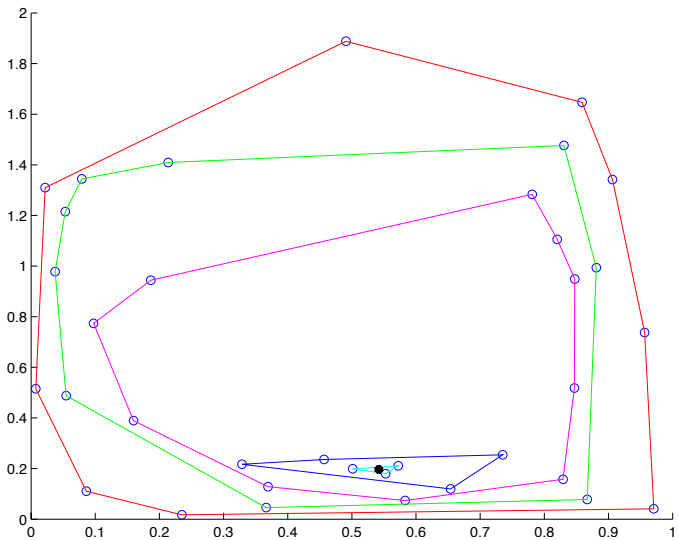
Tendance centrale

Extension au multi-dimensionnel

- Les concepts de moyenne et médiane définis en 1-D peuvent être adaptées en multi-dimensionnel
- Attention à l'interprétation !

Alternative

- Médiane de Tukey
- ...



Médiane de Tukey

Définition : Médiane de Tukey

C'est le centre de gravité des points de profondeur maximale encore appelé le sous espace médian.

La définition originale est de nature géométrique, elle est complexe mais il existe des algorithmes de calcul rapide.

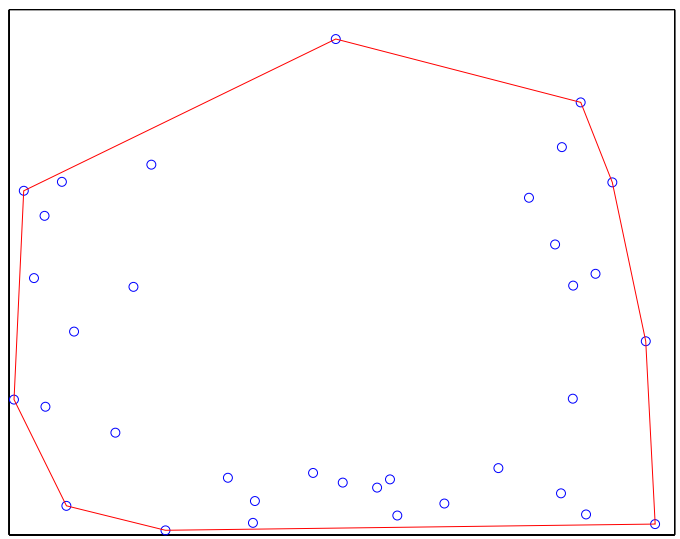
La méthode des pelures d'oignon

D_k l'ensemble des points de profondeurs k

La notion d'enveloppe convexe

Définition : Enveloppe convexe (Convex hull)

L'enveloppe convexe d'un ensemble de points est le plus petit ensemble convexe qui le contienne. Cette enveloppe convexe est un polyèdre convexe dont les sommets sont les points extérieurs.

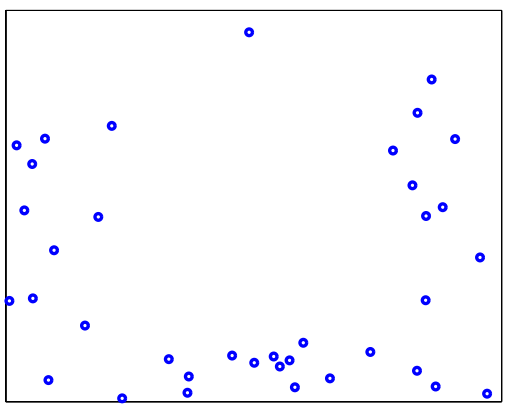


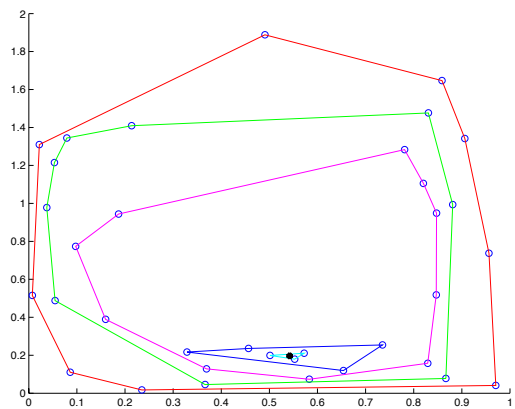
Calcul de la médianne de Tukey

Algorithme du calcul de la Médiane de Tukey

Pour construire la médiane de Tukey on utilise un algorithme de type **épluchage**:

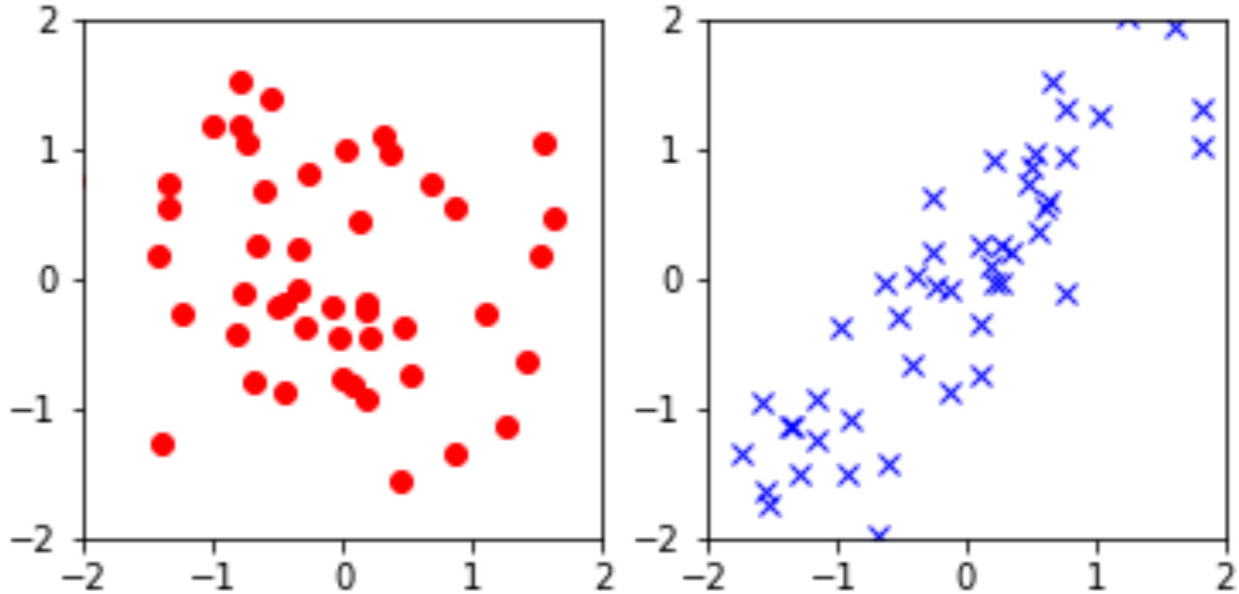
1. Tant qu'il reste des points non étiquetés
   - Considérons l'enveloppe convexe des points non étiquetés
   - on étiquette les points de cette enveloppe convexe
2. La médiane de Tukey est alors le centre de gravité (la moyenne) du dernier ensemble de points





Mesure de dépendance : motivation

Moyennes et variances égales:



Comment mesurer la dépendance ?

https://moodle.insa-rouen.fr/mod/resource/view.php?id=57129

Mesure de dépendance : cas des variables quantitatives

Variance empirique

$$s_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \qquad s_y^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

Covariance empirique

$$s_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) \qquad s_{xy} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\bar{y}$$

- Mesure comment les variables évoluent de manière conjointe

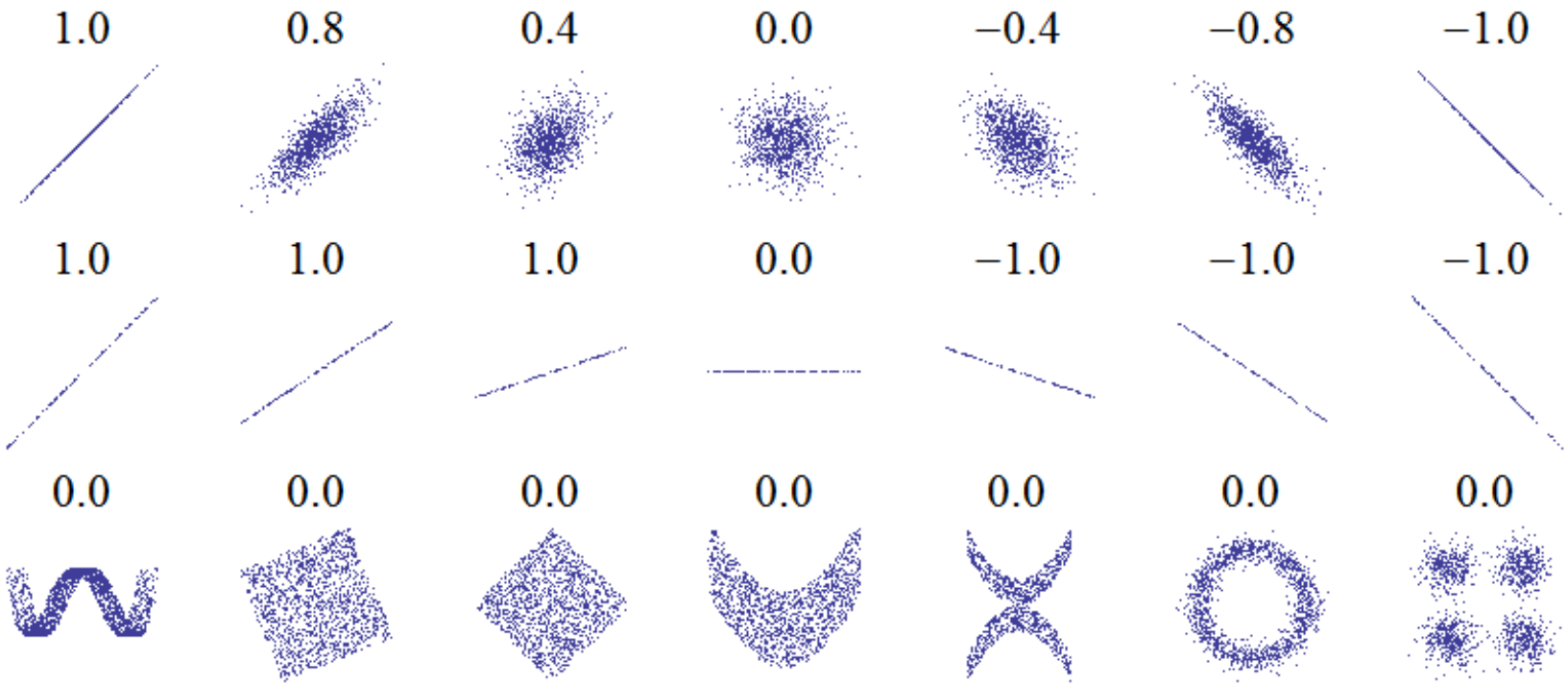
Mesure de dépendance : cas des variables quantitatives

Coefficient de corrélation

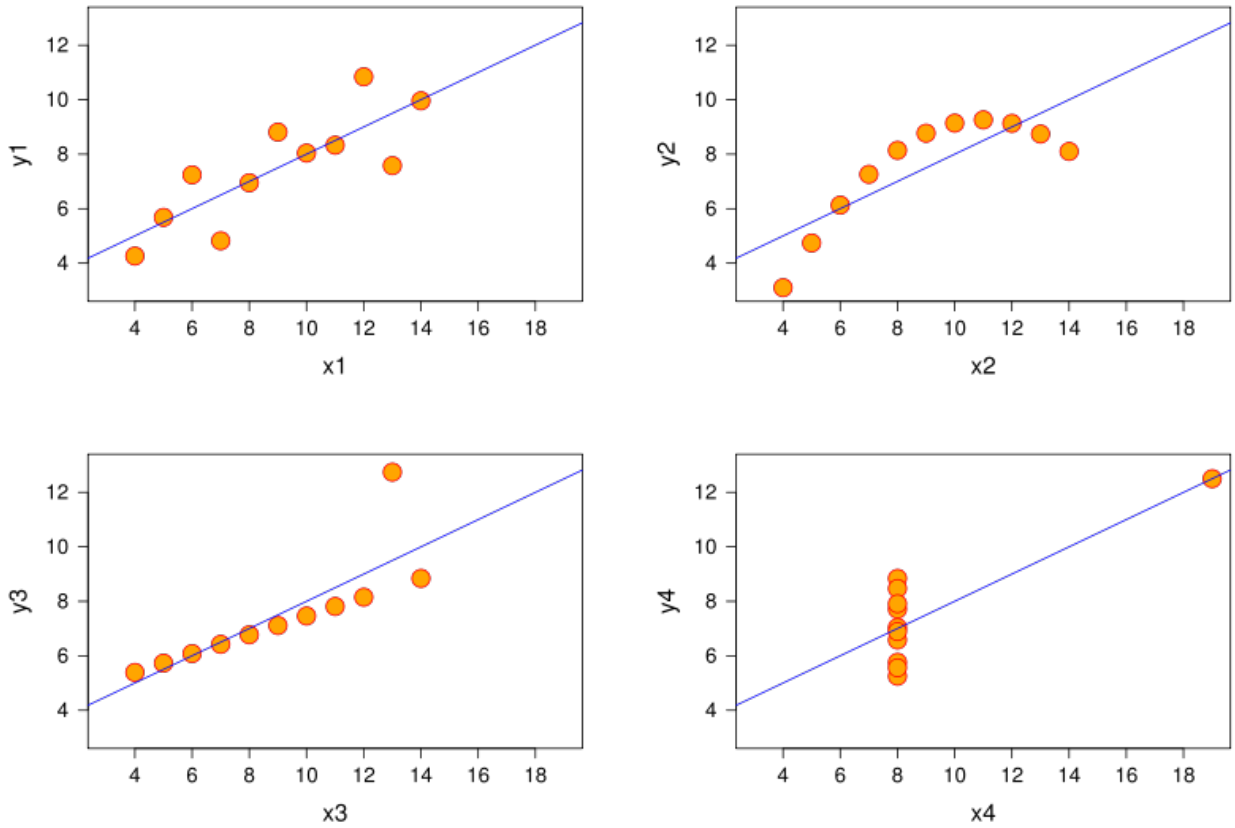
$$r = \frac{s_{xy}}{s_x s_y} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{s_x s_y} = \frac{(\sum_{i=1}^{n} x_i y_i) - n\bar{x}\bar{y}}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

- $r \in [-1, 1]$

Corrélations



Corrélations



Les quatre exemples ont le même coefficient de corrélation de 0,81
[Tiré de Wikipédia]

Cas des variables qualitatives

Variables indépendantes

$$\mathbb{P}(x, y) = \mathbb{P}(x)\mathbb{P}(y)$$

Mesure du degré d'indépendance entre deux variables

Distance du χ^2 entre la distribution empirique observé et celle théorique calculée en faisant l'hypothèse d'indépendance sur le tableau de contingence.

	Bob	Percy	John
TV	12	21	12
PC	21	12	14
VC	22	31	14
LV	4	17	5

Cas des variables qualitatives

Variables indépendantes

$$\mathbb{P}(x, y) = \mathbb{P}(x)\mathbb{P}(y)$$

Mesure du degré d'indépendance entre deux variables

Distance du χ^2 entre la distribution empirique observé et celle théorique calculée en faisant l'hypothèse d'indépendance sur le tableau de contingence.

	Bob	Percy	John	
TV	12	21	12	45
PC	21	12	14	47
VC	22	31	14	67
LV	4	17	5	26
	59	81	45	185

Cas des variables qualitatives

Variables indépendantes

$$\mathbb{P}(x, y) = \mathbb{P}(x)\mathbb{P}(y)$$

Mesure du degré d'indépendance entre deux variables

Distance du χ^2 entre la distribution empirique observé et celle théorique calculée en faisant l'hypothèse d'indépendance sur le tableau de contingence.

	Bob	Percy	John	
TV	12	21	12	45
PC	21	12	14	47
VC	22	31	14	67
LV	4	17	5	26
	59	81	45	185

	Bob	Percy	John
TV	14.35	19.70	10.94
PC	14.98	20.57	11.43
VC	21.36	29.33	16.29
LV	8.29	11.38	6.32

$$\mathbb{P}(TV, Bob) = \frac{59}{185} * \frac{45}{185} = 14.35$$

Variables qualitatives : distance du χ^2

Définition : Distance du χ^2

On appelle distance du χ^2 du tableau de contingence n:

$$D_{\chi^2}(X,Y) = \sum_{i\in\Omega_x}\sum_{j\in\Omega_y} \frac{(n_{ij} - n_{i\bullet}n_{\bullet j}/n)^2}{n_{i\bullet}n_{\bullet j}/n} = n\sum_{i\in\Omega_x}\sum_{j\in\Omega_y} \frac{\left(\widehat{\mathbb{P}}_{ij} - \widehat{\mathbb{P}}_i\widehat{\mathbb{P}}_j\right)^2}{\widehat{\mathbb{P}}_i\widehat{\mathbb{P}}_j}$$

- $\widehat{\mathbb{P}}_{ij} = n_{ij}/n$: fréquence empirique de la loi jointe
- $\widehat{\mathbb{P}}_i = n_{i\bullet}/n$, $\widehat{\mathbb{P}}_j = n_{\bullet j}/n$: Distributions marginales empiriques
- $D_{\chi^2}(X,Y) = \frac{(12-14.35)^2}{14.35} + \dots$
- À comparer avec des distances de réference (on verra ça plus tard)

Cas mixte : l'analyse de la variance

Découpage de la variance

- X variable qualitative, Y la variable quantitative
- y_{ij} le i-ème de la modalité j.
- Variance totale:

$$s^2 = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{j\in\Omega_x}\sum_{i=1}^{n_j}(y_{ij} - \bar{y})^2 = s^2_{\text{inter}} + s^2_{\text{intra}}$$

- Variance intra-classe

$$s^2_{\text{intra}} = \sum_{j\in\Omega_x}\sum_{i=1}^{n_j}(y_{ij} - \bar{y_j})^2$$

- Variance inter-classe

$$s^2_{\text{inter}} = \sum_{j\in\Omega_x}\sum_{i=1}^{n_j}(\bar{y_j} - \bar{y})^2 = \sum_{j\in\Omega_x} n_j(\bar{y_j} - \bar{y})^2$$

Cas mixte : l'analyse de la variance

Le rapport de corrélation

La dépendance est évaluée par le rapport des variances :

$$\eta^2_{Y/X} = \frac{s^2_{\text{inter}}}{s^2} = \frac{\sum_{j\in\Omega_x} n_j(\bar{y_j} - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

- Carré du rapport de corrélation $\eta^2_{Y/X}$
- $\eta^2_{Y/X} \in [0, 1]$
- $\eta^2_{Y/X} = 1$: la variance entre les classes expliquent toute la variance : les variables sont liées
- $\eta^2_{Y/X} = 0$: Pas de variance inter classe. Toute la variance est "à l'intérieur" des classes.
- Il augmente avec $p = \text{card}(\Omega_x)$

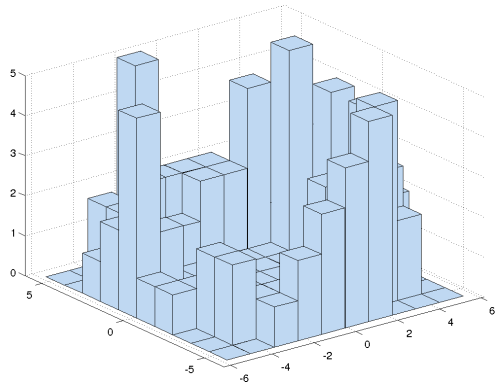
Mesure de ressemblance : Résumé

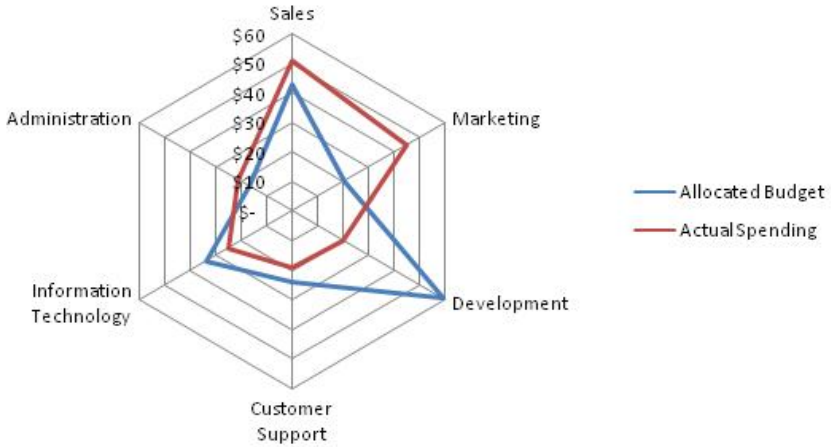
X	Y	Exemple	étude de la dépendance ?
continue	continue	poids/taille	corrélation
continue	discrète	poids/variété	analyse de la variance
discrète	discrète	nombre d'enfants / PCS	distance du χ^2

- Corrélation : $r = \frac{s_{xy}}{s_x s_y} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{s_x s_y}$
- Analyse de la variance : $s^2_{Y/X} = \frac{\sum_{j \in \Omega_x} n_j (\bar{y_j} - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$
- Distance du χ^2 : $D_{\chi^2}(X, Y) = \sum_{i \in \Omega_x} \sum_{j \in \Omega_y} \frac{(n_{ij} - n_{i.} n_{.j}/n)^2}{n_{i.} n_{.j}/n}$

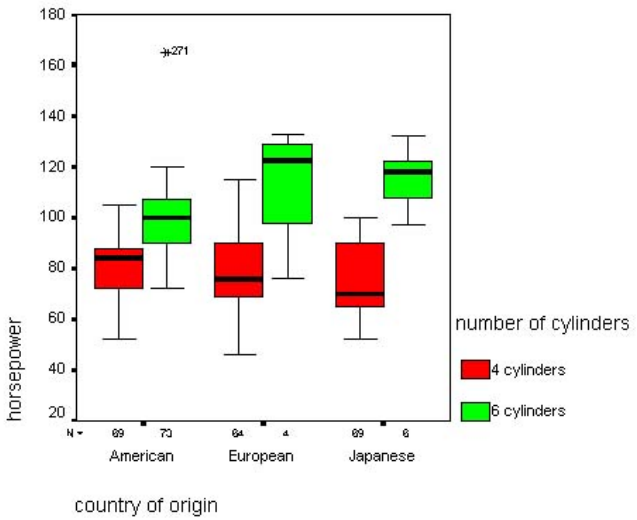
Résumé graphique de deux variables qualitatives

Histogrammes 2d, diagramme de Kiviat (radar, star, ou spider charts)

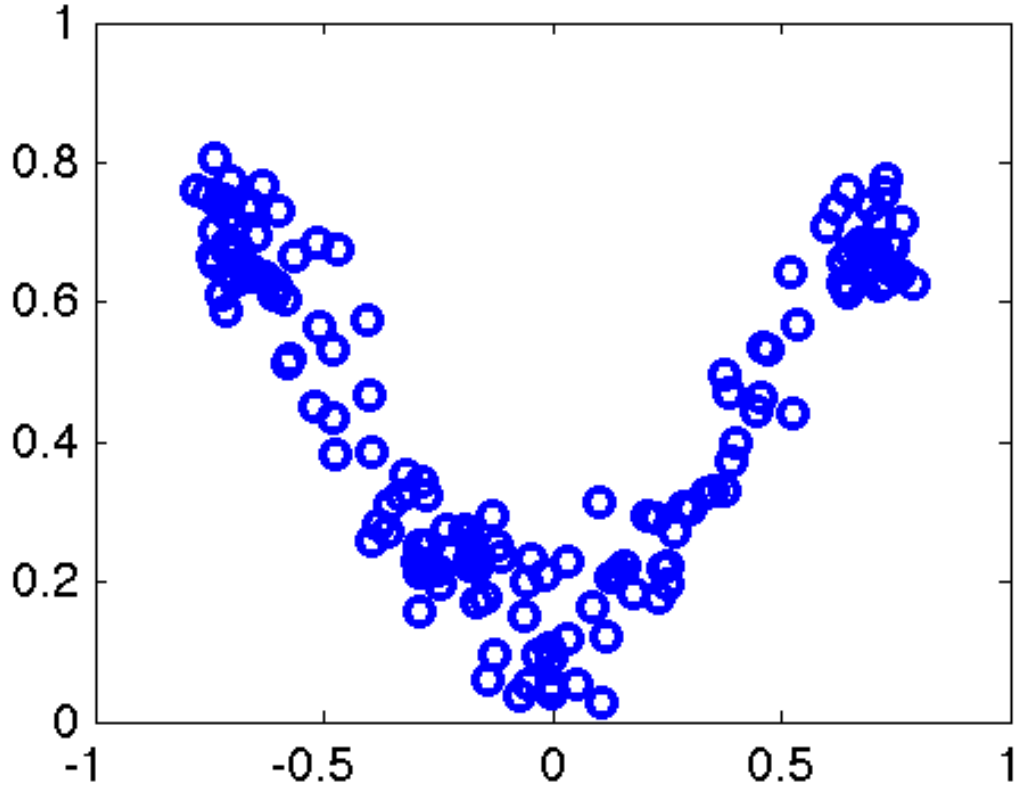




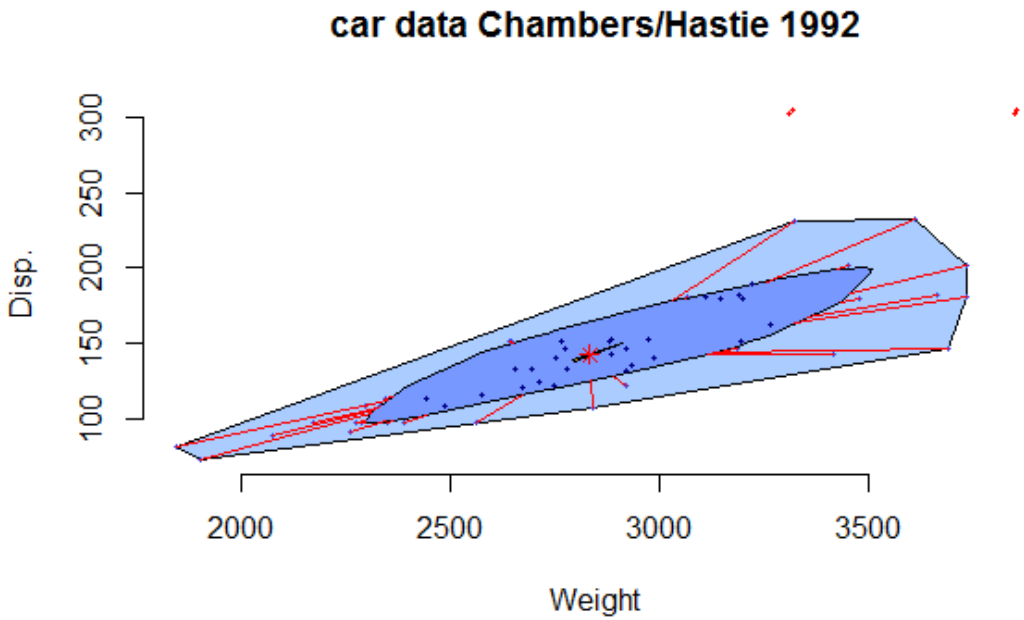
et boites à moustaches



Résumé graphique de deux variables quantitatives



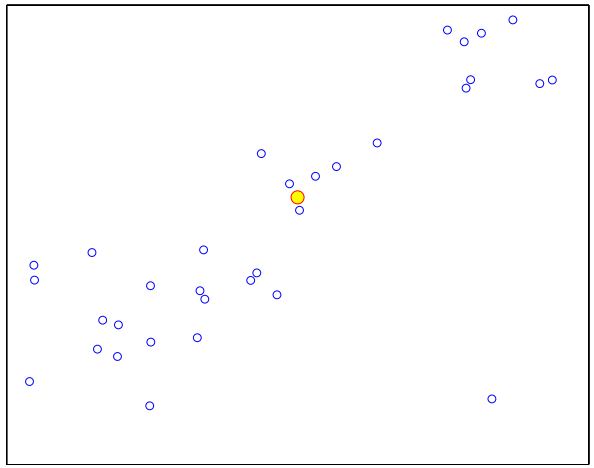
Résumé graphique : Sac à moustache



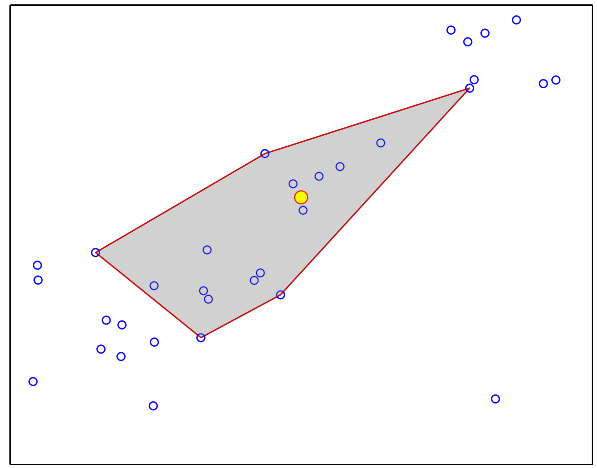
1. Centre (médiane 2d)
2. Boite (sac médian)
3. Épure
4. Moustaches
5. Points aberrant

L'analogue en deux dimensions des boites à moustaches.

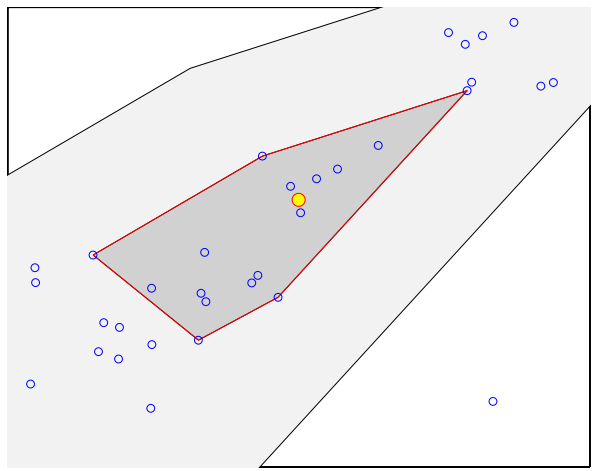
Les étapes du calcul du sac à moustaches



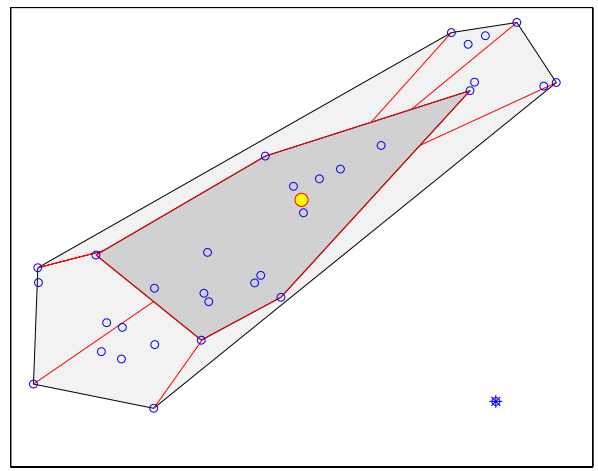
Calcul de la médiane de Tukey



Calcul du sac médian



Calcul de l'épure



Sac à moustaches

Calcul du sac à moustaches

1. Calculer la médiane 2d (le barycentre des points de profondeur maximale)
2. Sélectionner les enveloppes convexes $\#D_k \leq n/2 \leq \#D_{k-1}$ points les plus proches de la médiane, en dessiner l'enveloppe interpolante. Le résultat est ce qu'on appelle le sac médian (*bag plot* en anglais).
3. Calculer l'épure : l'enveloppe du sac médian "gonflé" trois fois
4. Dessiner l'enveloppe convexe des points contenus dans l'épure. Il s'agit des points situés à moins de trois fois la distance entre le sac médian et la médiane. Une moustache est associée à chacun des points de cette enveloppe convexe.
5. Les points en dehors de cette dernière enveloppe son marqués comme étant hors épure.

P. J. Rousseeuw, I. Ruts & J. W. Tukey (1999): The Bagplot: A Bivariate Boxplot, The American Statistician, 53:4, 382-387

https://www.tandfonline.com/doi/abs/10.1080/00031305.1999.10474494

Exemple d'utilisation du sac à moustaches

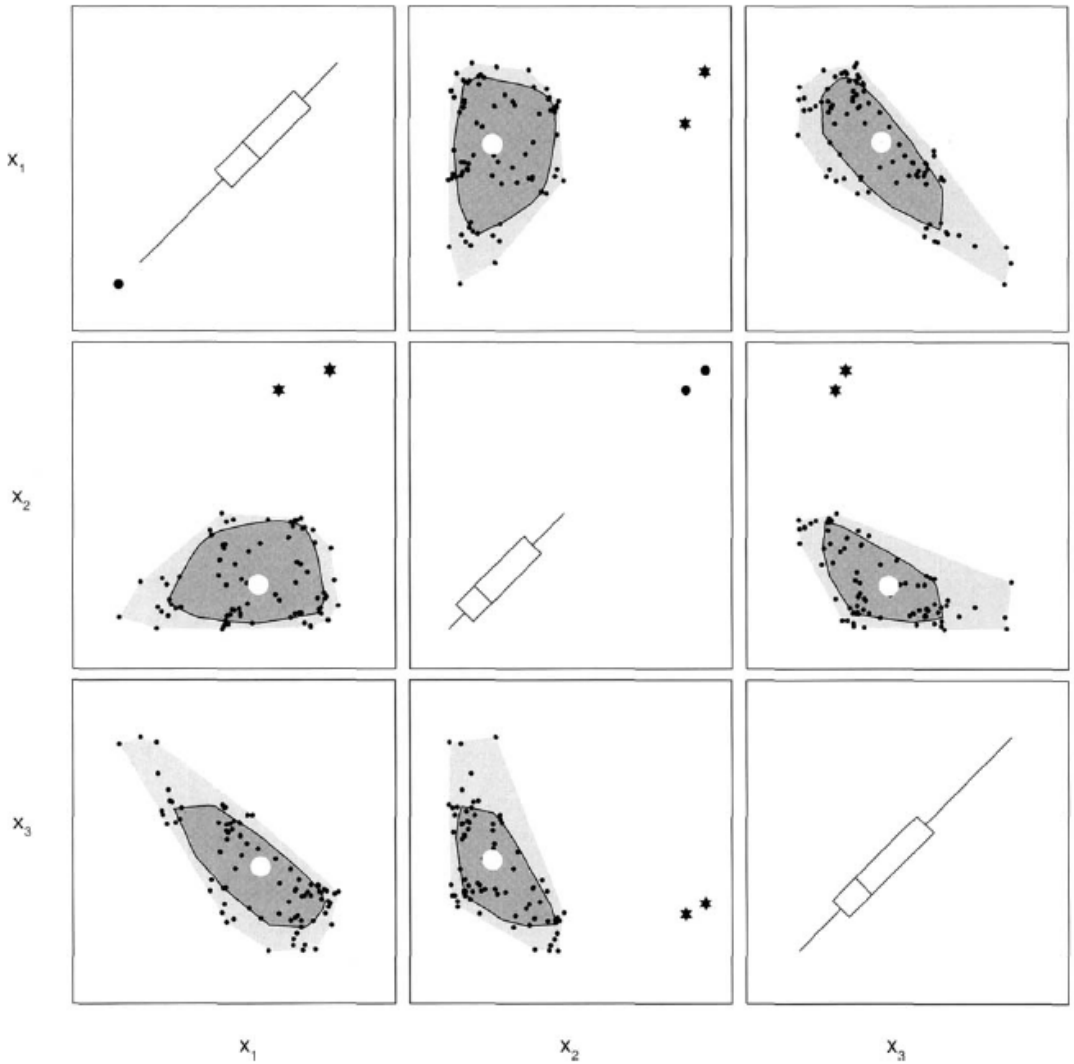


Figure 7. Bagplot matrix of the three-dimensional aquifer data with 85 data points.

P. J. Rousseeuw, I. Ruts & J. W. Tukey (1999): The Bagplot: A Bivariate Boxplot, The American Statistician, 53:4, 382-387

Conclusion

Tableau

- Qualitative et quantitative discrète : Tableau de contingence
- Quantitative (continue) : pseudo variables

Résumé quantitatif

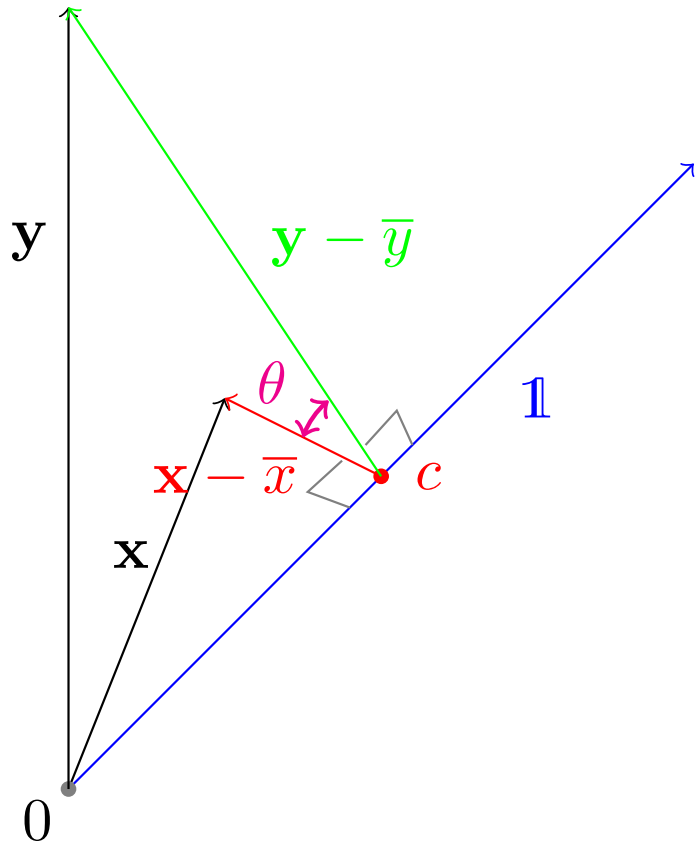
- ordre 1 : centre (calcul)
- ordre 2 : corrélation, Anavar (anova) et χ^2
- 2D est différent de 1D

Résumé qualitatif (graphique)

- Qualitative et quantitative discrète : histogramme
- Quantitative qualitative : diagramme de Kiviat
- Quantitative (continue) : nuage de points et bag plot.

Bonus

Corrélation : interprétation géométrique



Sans perte de généralité supposons que $\overline{x} = \overline{y} = c$

$r \in [-1, 1]$ analogue à $cos(\theta)$, ou θ c'est l'angle formé par $\mathbf{x} - (\bar{x} \times \vec{1})$ et $\mathbf{y} - (\bar{y} \times \vec{1})$

$$r = \frac{s_{xy}}{s_x s_y} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{s_x s_y} = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$