

Statistique

Benjamin Bobbia

ISAE



Point organisation

- Les BE : Sur le logiciel R :
 - Sujet au format notebook disponible sur le LMS.
 - A uploader et traiter sur le Jupyter Lab de l'isae :
<https://jupyter.isae-sup Aero.fr/>
- Évaluation : Projet, à rendre par groupe de 5, pour le **01/10/2025**.
L'objectif est de mettre en oeuvre les méthodes vue dans le cours et les BE sur un jeux de données réelles.
 - Quelle question je me pose sur les données ? Qu'est ce que je veux illustrer ?
 - Pourquoi la méthode que je vais utiliser est pertinente ?
 - Quelles conclusions j'en retire ?
 - Quelles sont les possible limitation de ce que je propose ?

Point organisation : le projet

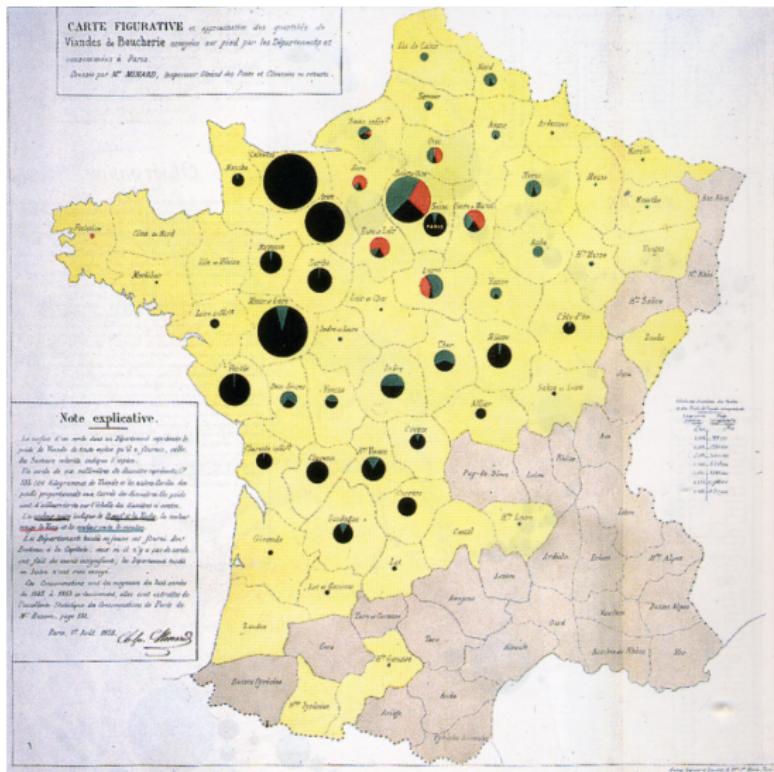
Le projet consiste en l'étude du jeu de donnée **meteoSDD**. Le document rendus, sous la forme d'un **notebook Jupyter**, devra contenir au moins :

- Une présentation des données avec des résumées numériques et graphiques des données. (Plusieurs évidemment)
- Un calcul et étude d'estimateur.
- Des tests (au moins un sur un parametres et une ANOVA).
- Une regression et étude de corrélation.
- Une ACP et une PLS.
- Une méthode de clustering.

Introduction

« Je libérerai les statistiques du mépris dans lequel elles sont longtemps restées. »

Charles-Joseph Minard, La Statistique, 1869



Un peu d' étymologie :

- **Status** : *État* en latin,
- **Statista** : *Homme d'État* en italien (1633),
- **Statistica** : *Description détaillée d'un état relativement à son étendue, à sa population, à ses ressources, ...* en italien (1672),
- **Statistik** : *Connaissances que doit posséder un homme d'État* selon l'économiste allemand Gottfried Achenwall (1785).

Définition de l'académie française

Science qui a pour objet de recueillir et de dénombrer les divers faits de la vie sociale.

La statistique est une discipline qui concerne les affaires de l'État. Il convient de distinguer :

- **les statistiques** : valeurs des informations recueillies,
- **la statistique** : ensemble des techniques utilisées pour l'étude méthodique des faits sociaux.

Dans un cadre statistique, il est commun d'appeler :

- **un individu** un élément individuel considéré dans l'étude,
- **une population** l'ensemble des individus considérés,
- **un échantillon** une partie de la population étudiée,
- **une statistique** une fonction définie sur l'ensemble des échantillons.

Ces définitions se généralisent aux méthodes utilisées pour l'étude d'un ensemble d'observations, appelé **jeu de données**, que les mathématiques permettent de placer dans un cadre formel. Il s'agit de l'**objet de ce cours**.

Statistique mathématique

Domaine des mathématiques dédié à l'étude d'un phénomène par la collecte de données, leur traitement, leur analyse, l'interprétation des résultats et leur présentation.

Ces définitions se généralisent aux méthodes utilisées pour l'étude d'un ensemble d'observations, appelé **jeu de données**, que les mathématiques permettent de placer dans un cadre formel. Il s'agit de l'**objet de ce cours**.

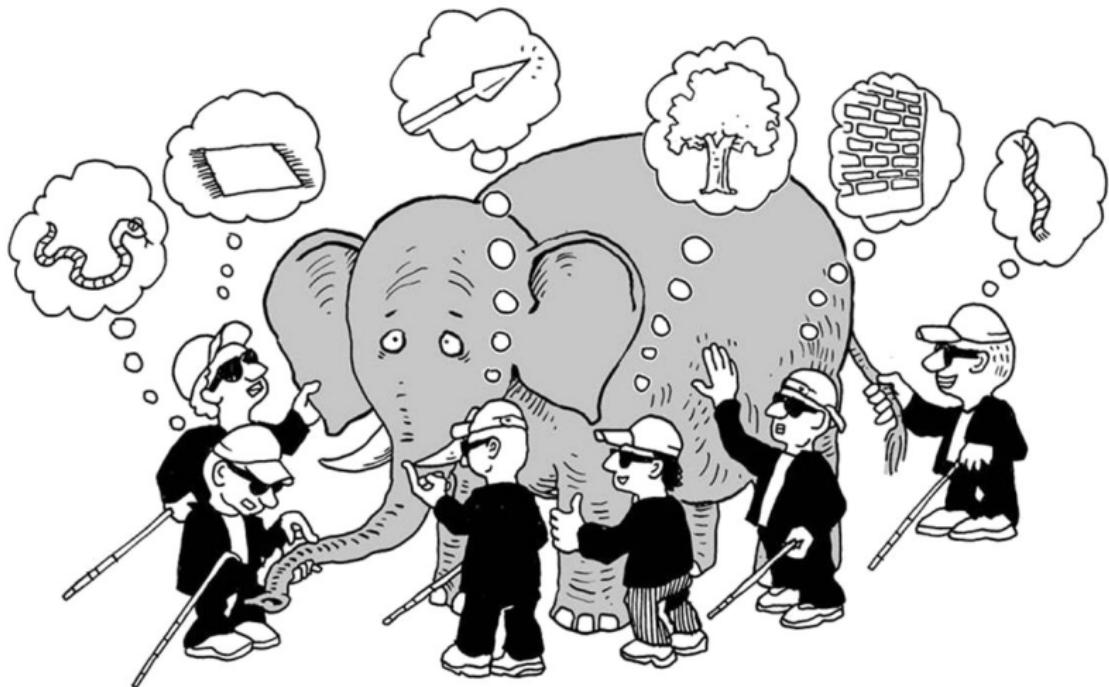
Statistique mathématique

Domaine des mathématiques dédié à l'étude d'un phénomène par la collecte de données, leur traitement, leur analyse, l'interprétation des résultats et leur présentation.

Cette définition ne donne pas de finalité à cette étude :

- si l'objectif est de **décrire** ou de **résumer** l'information contenue dans les données, nous parlerons de **statistique exploratoire**,
- si l'objectif est de **prédirer** ou de **généraliser** à partir des données, elles devront être placées dans un modèle mathématique (généralement probabiliste) et nous parlerons de **statistique inférentielle**.

Statistique exploratoire



1.1 Notions élémentaires

Définitions

Variable et observations

Une **variable** x est une entité à valeurs dans un espace \mathcal{X} qui peut être observée. Dans la suite, un jeu de données issu de n observations de x est appelé un **échantillon** et noté $x_1, \dots, x_n \in \mathcal{X}$.

Une variable x à valeurs dans un espace \mathcal{X} est appelée :

- **quantitative** si \mathcal{X} est un sous-espace de \mathbb{R}^p (température, position spatiale, ...),
- **qualitative** ou **catégorielle** si \mathcal{X} est un ensemble fini (catégorie socio-professionnelle, ...).

Lorsque $\mathcal{X} = \mathbb{R}$, la variable x est dite **réelle**.

Tendance centrale

Considérons une variable réelle x et des observations $x_1, \dots, x_n \in \mathbb{R}$.

Objectif : résumer les données numériques de l'échantillon par un ou plusieurs nombres (mais pas trop).

- **Moyenne** : somme des observations divisée par leur nombre,
- **Médiane** : valeur qui permet de couper les observations en 2 parties de tailles égales,
- **Mode** : valeur la plus fréquente dans l'échantillon,
- **Quartiles** : 3 valeurs qui permettent de couper les observations en 4 parties de tailles égales,
- **Déciles** : 9 valeurs qui permettent de couper les observations en 10 parties de tailles égales,
- **Percentiles** : 99 valeurs qui permettent de couper les observations en 100 parties de tailles égales,
- **Quantiles** : valeurs qui généralisent les indicateurs précédents.

Moyennes

Échantillon réel : $x_1, \dots, x_n \in \mathbb{R}$.

- **Moyenne (arithmétique)** : $\bar{x} = \frac{1}{n} \sum_{k=1}^n x_k$
- **Moyenne géométrique** : $\left(\prod_{k=1}^n x_k \right)^{1/n}$
- **Moyenne harmonique** : $n \left(\sum_{k=1}^n \frac{1}{x_k} \right)^{-1}$
- **Moyenne quadratique** : $\sqrt{\frac{1}{n} \sum_{k=1}^n x_k^2}$

Moyenne et médiane

Soient x et y des variables réelles et des échantillons $x_1, \dots, x_n \in \mathbb{R}$ et $y_1, \dots, y_n \in \mathbb{R}$ associés.

La moyenne est **linéaire** en l'échantillon,

$$\forall a \in \mathbb{R}, \quad \overline{ax} = \frac{1}{n} \sum_{k=1}^n ax_k = a\overline{x} \quad \text{et} \quad \overline{x+y} = \frac{1}{n} \sum_{k=1}^n x_k + y_k = \overline{x} + \overline{y}.$$

La médiane est **homogène** mais **pas additive** (sauf cas particuliers),

$$\forall a \in \mathbb{R}, \quad \text{Med}(ax) = a\text{Med}(x) \quad \text{et} \quad \text{Med}(x+y) \neq \text{Med}(x) + \text{Med}(y).$$

$$\begin{aligned} x_1 &= 0, x_2 = 1, x_3 = 2 : \text{Med}(x) = 1, \\ y_1 &= 2, y_2 = 0, y_3 = 0 : \text{Med}(y) = 0, \\ \text{Med}(x+y) &= 2 \text{ et } \text{Med}(x) + \text{Med}(y) = 1. \end{aligned}$$

Moyenne et médiane

La moyenne \bar{x} minimise les **écart au carré**,

Moyenne : barycentre L2

$$\forall t \in \mathbb{R}, \sum_{k=1}^n (x_k - \bar{x})^2 \leq \sum_{k=1}^n (x_k - t)^2.$$

La médiane $\text{Med}(x)$ minimise les **écart en valeur absolue**,

Médiane : barycentre L1

$$\forall t \in \mathbb{R}, \sum_{k=1}^n |x_k - \text{Med}(x)| \leq \sum_{k=1}^n |x_k - t|.$$

La médiane est plus **robuste** par rapport aux valeurs extrêmes.

Moyenne et médiane

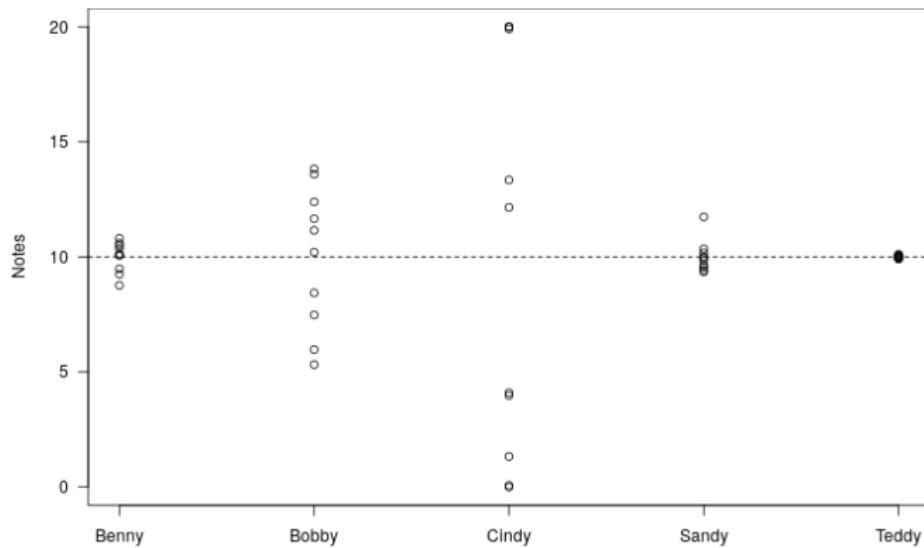
Considérons une entreprise de 12 personnes :

- 8 ouvriers (1201 €)
- 1 chef d'atelier (2000 €)
- 1 directeur technique (5000 €)
- 1 directeur des RH (8000 €)
- 1 directeur général (10000 €)

Salaire moyen : 2884 € – Salaire médian : 1201 €

Autres exemples : gains au Loto, ...

Capacité à résumer



Cette figure représente les notes obtenues par 5 étudiants dans 10 matières différentes. Ces étudiants ont tous une moyenne égale à 10. Donneriez-vous la même appréciation à chacun d'entre eux ? Est-ce que la moyenne est un « bon » résumé de leurs résultats ?

Dispersion

Échantillon réel : $x_1, \dots, x_n \in \mathbb{R}$.

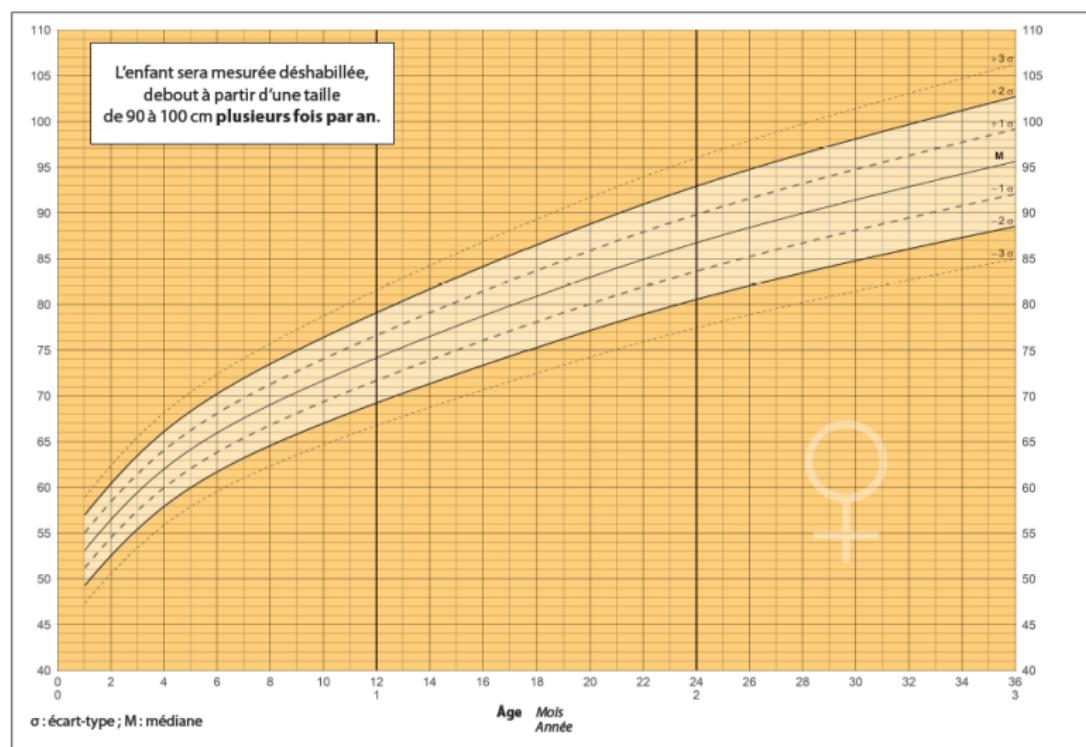
- **Variance** : $\sigma^2(x) = \frac{1}{n} \sum_{k=1}^n (x_k - \bar{x})^2$.
- **Écart-type** : $\sigma(x) = \sqrt{\sigma^2(x)}$, cet indicateur s'exprime dans les mêmes unités que la moyenne \bar{x} . Ainsi, ces quantités peuvent être additionnées pour considérer la dispersion à l'aide d'**intervalles**,

Loi normale : 1sigma 66% de l'information

$\forall a \in \mathbb{R}_+, [\bar{x} - a\sigma(x), \bar{x} + a\sigma(x)]$.

- **Étendue** : $\max_k x_k - \min_k x_k$.
- **Écart interquartile** : distance entre le premier et le troisième quartile.

Exemple d'intervalles



Variance

Échantillon réel : $x_1, \dots, x_n \in \mathbb{R}$.

$$\text{Variance} : \sigma^2(x) = \frac{1}{n} \sum_{k=1}^n (x_k - \bar{x})^2.$$

La variance est **quadratique**,

$$\forall a \in \mathbb{R}, \sigma^2(ax) = \frac{1}{n} \sum_{k=1}^n (ax_k - a\bar{x})^2 = a^2 \sigma^2(x).$$

La variance est **invariante par translation**,

$$\forall b \in \mathbb{R}, \sigma^2(x + b) = \frac{1}{n} \sum_{k=1}^n ((x_k + b) - (\bar{x} + b))^2 = \sigma^2(x).$$

Variance

Échantillon réel : $x_1, \dots, x_n \in \mathbb{R}$.

Variance : $\sigma^2(x) = \frac{1}{n} \sum_{k=1}^n (x_k - \bar{x})^2$.

La variance est **nulle** si et seulement si l'échantillon est **constant**,

$$\sigma^2(x) = 0 \iff (x_k - \bar{x})^2 = 0, \forall k \iff x_k = \bar{x}, \forall k.$$

La variance peut être utilisée comme une **mesure de la capacité à résumer** l'information contenue dans le jeu de données réelles par la moyenne. Autrement dit, la variance quantifie la **quantité d'information non expliquée** par la moyenne.

Centrer et réduire

Échantillon réel : $x_1, \dots, x_n \in \mathbb{R}$.

- **Centrer** : retrancher la moyenne.
- **Réduire** : diviser par l'écart-type.

Version centrée-réduite / z-transformation

Les observations z_1, \dots, z_n centrées et réduites sont données par

$$\forall i \in \{1, \dots, n\}, z_i = \frac{x_i - \bar{x}}{\sigma(x)}.$$

Par définition, $\bar{z} = 0$ et $\sigma^2(z) = 1$.

Ces opérations permettent d'exprimer les données dans une échelle neutre en les débarrassant de leurs unités physiques. Une fois centrées et réduites, les observations s'expriment comme un **nombre d'écart-types par rapport à la moyenne**.

Covariance

Soit un couple (x, y) de variables réelles.

Observations : $(x_1, y_1), \dots, (x_n, y_n) \in \mathbb{R}^2$.

La **covariance** entre les observations de x et y est définie par

$$\sigma(x, y) = \frac{1}{n} \sum_{k=1}^n (x_k - \bar{x})(y_k - \bar{y}).$$

Covariance

Soit un couple (x, y) de variables réelles.

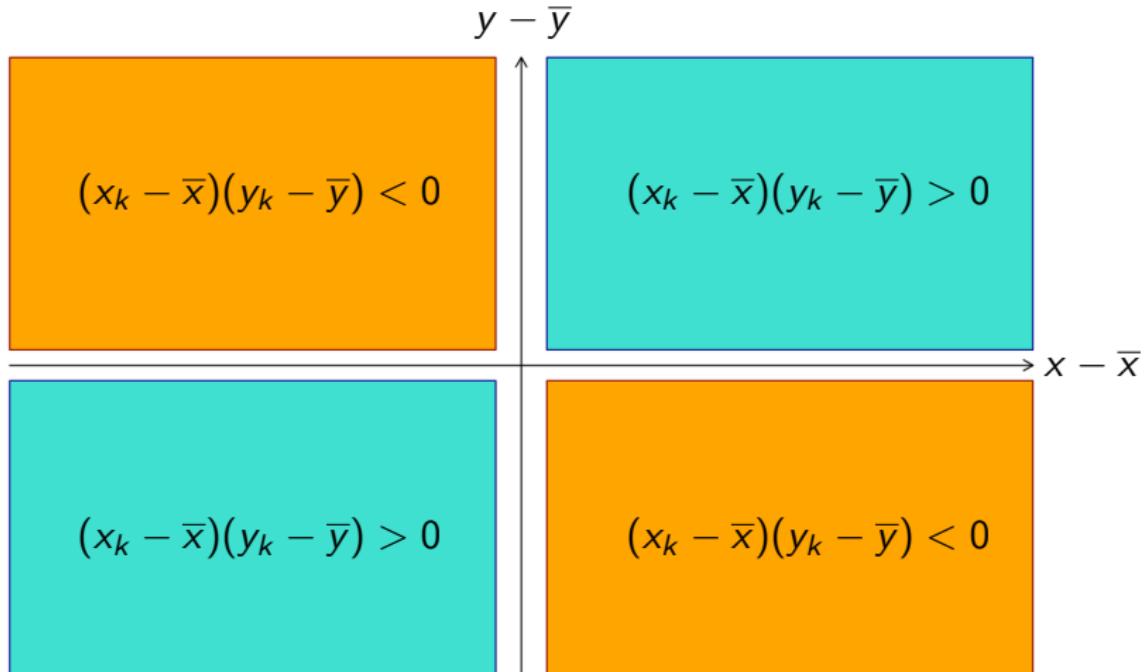
Observations : $(x_1, y_1), \dots, (x_n, y_n) \in \mathbb{R}^2$.

La **covariance** entre les observations de x et y est définie par

$$\sigma(x, y) = \frac{1}{n} \sum_{k=1}^n (x_k - \bar{x})(y_k - \bar{y}).$$

- La covariance $\sigma(x, y)$ est **positive** si les x_k et les y_k ont tendance à être simultanément du même côté de leurs moyennes respectives.
⇒ **Relation croissante** (e.g. température et nombre de glaces vendues)
- La covariance $\sigma(x, y)$ est **négative** si les x_k et les y_k ont tendance à être simultanément du côté opposé de leurs moyennes respectives.
⇒ **Relation décroissante** (e.g. température et consommation de chauffage)

Covariance (signe)



Propriétés de la covariance

Échantillon : $(x_1, y_1), \dots, (x_n, y_n) \in \mathbb{R}^2$.

Covariance : $\sigma(x, y) = \frac{1}{n} \sum_{k=1}^n (x_k - \bar{x})(y_k - \bar{y})$.

La covariance est ...

- **symétrique** : $\sigma(x, y) = \sigma(y, x)$,
- **bilinéaire** :

$$\forall a \in \mathbb{R}, \quad \sigma(ax, y) = a\sigma(x, y) \quad \text{et} \quad \sigma(x + x', y) = \sigma(x, y) + \sigma(x', y),$$

- **définie positive** : $\sigma(x, x) = \sigma^2(x) \geq 0$.

La covariance est donc un ???

Propriétés de la covariance

Échantillon : $(x_1, y_1), \dots, (x_n, y_n) \in \mathbb{R}^2$.

Covariance : $\sigma(x, y) = \frac{1}{n} \sum_{k=1}^n (x_k - \bar{x})(y_k - \bar{y})$.

La covariance est ...

- **symétrique** : $\sigma(x, y) = \sigma(y, x)$,
- **bilinéaire** :

$$\forall a \in \mathbb{R}, \sigma(ax, y) = a\sigma(x, y) \quad \text{et} \quad \sigma(x + x', y) = \sigma(x, y) + \sigma(x', y),$$

- **définie positive** : $\sigma(x, x) = \sigma^2(x) \geq 0$.

La covariance est donc un **produit scalaire** (et l'écart-type est la norme associée).

Cauchy-Schwarz

$$|\sigma(x, y)| \leq \sigma(x)\sigma(y)$$

Coefficient de corrélation linéaire de Pearson

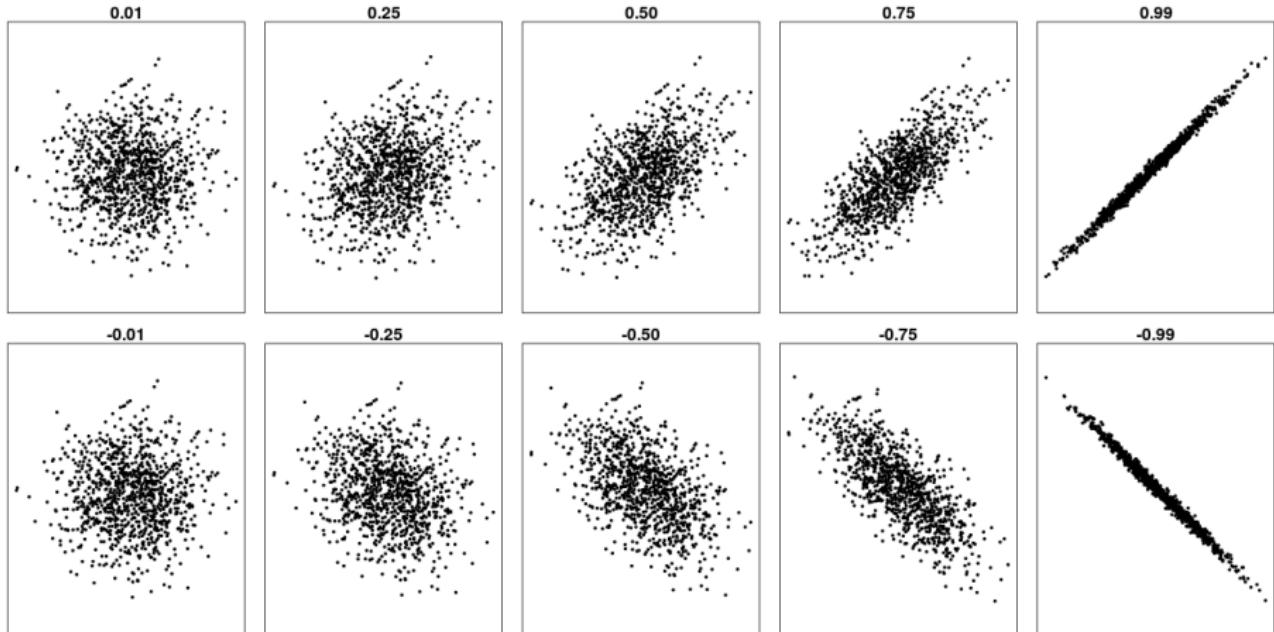
Échantillon : $(x_1, y_1), \dots, (x_n, y_n) \in \mathbb{R}^2$.

Le coefficient de corrélation linéaire de Pearson entre les observations des x et y est défini par

$$\rho(x, y) = \frac{\sigma(x, y)}{\sigma(x)\sigma(y)} \in [-1, 1].$$

- Par linéarité de la covariance, $\rho(x, y)$ est également la covariance des versions centrées-réduites de x et y .
- Le signe de $\rho(x, y)$ s'interprète comme celui de $\sigma(x, y)$.
- Si $|\rho(x, y)| = 1$, alors il y a égalité dans Cauchy-Schwarz et donc colinéarité, i.e. **les observations sont distribuées sur une droite.**

Coefficient de corrélation linéaire (exemple)



Régression linéaire simple

Échantillon : $(x_1, y_1), \dots, (x_n, y_n) \in \mathbb{R}^2$.

Objectif : trouver l'équation de la droite $y = ax + b$ qui passe « au plus près » des points observés.

Critère des moindres carrés

$$\forall a, b \in \mathbb{R}, \gamma(a, b) = \frac{1}{n} \sum_{k=1}^n (y_k - ax_k - b)^2$$

La **droite de régression** est donnée par l'équation $y = \hat{a}x + \hat{b}$ telle que (\hat{a}, \hat{b}) minimise le critère des moindres carrés.

Les écarts au modèle linéaire $\varepsilon_k = y_k - \hat{a}x_k - \hat{b}$ sont appelés les **résidus**.

Régression linéaire simple

Échantillon : $(x_1, y_1), \dots, (x_n, y_n) \in \mathbb{R}^2$.

Objectif : trouver l'équation de la droite $y = ax + b$ qui passe « au plus près » des points observés.

Point d'annulation du gradient :

$$\nabla \gamma(a, b) = 0 \iff \begin{cases} \frac{1}{n} \sum_{k=1}^n x_k(y_k - ax_k - b) = 0 \\ \frac{1}{n} \sum_{k=1}^n (y_k - ax_k - b) = 0 \end{cases}$$

$$\iff \begin{cases} a\sigma^2(x) = \sigma(x, y) \\ b = \bar{y} - a\bar{x} \end{cases}$$

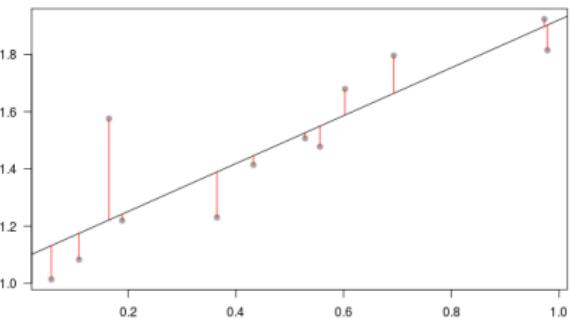
Régression linéaire simple

Échantillon : $(x_1, y_1), \dots, (x_n, y_n) \in \mathbb{R}^2$.

Objectif : trouver l'équation de la droite $y = ax + b$ qui passe « au plus près » des points observés.

Si $\sigma^2(x) > 0$, la **droite de régression** est donnée par l'équation $y = \hat{a}x + \hat{b}$ où

$$\hat{a} = \frac{\sigma(x, y)}{\sigma^2(x)} = \frac{\rho(x, y)\sigma(y)}{\sigma(x)} \quad \text{et} \quad \hat{b} = \bar{y} - \hat{a}\bar{x}.$$



$$\begin{aligned} \bar{x} &= 0.471 & \sigma^2(x) &= 0.090 \\ \bar{y} &= 1.478 & \sigma^2(y) &= 0.080 \\ \rho(x, y) &= 0.882 \end{aligned}$$

$$\Rightarrow y = 0.832x + 1.086$$

Régression linéaire : Qualité de l'ajustement

Comment peut-on mesurer la qualité de cette approximation linéaire ?

Régression linéaire : Qualité de l'ajustement

Comment peut-on mesurer la qualité de cette approximation linéaire ?
On peut décomposer la variance des $(y_i)_{i=1}^n$:

$$\underbrace{\sum_{i=1}^n (y_i - \bar{y})^2}_{SCT} = \underbrace{\sum_{i=1}^n (y_i - \hat{y}_i)^2}_{SCR} + \underbrace{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}_{SCE}.$$

Régression linéaire : Qualité de l'ajustement

Comment peut-on mesurer la qualité de cette approximation linéaire ?
On peut décomposer la variance des $(y_i)_{i=1}^n$:

On regarde la proportion de variance expliquée par le régression :

$$R^2 = \frac{SCE}{SCT}.$$

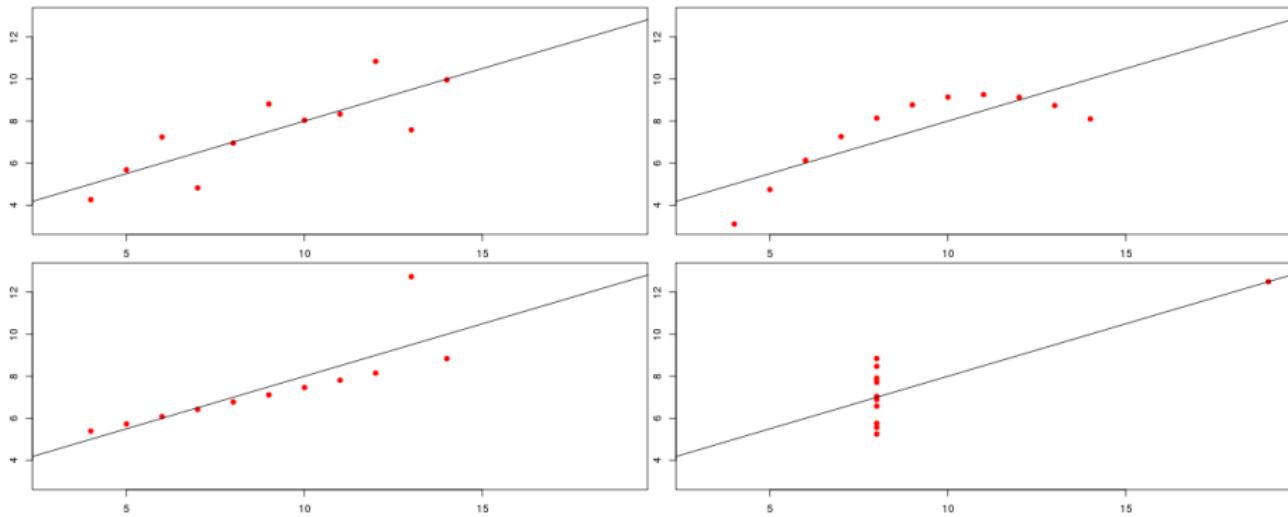
R^2 est appelé coefficient de détermination :

- R^2 proche de 1 \Rightarrow Régression pertinente.
 - R^2 proche de 0 \Rightarrow Régression non pertinente.

Coefficient de corrélation linéaire (contre-exemple)

... make both calculations and graphs. Both sorts of output should be studied; each will contribute to understanding.

F. J. Anscombe, 1973



$$\bar{x} = 9.0 \quad \sigma^2(x) = 10.0 \quad \bar{y} = 7.5 \quad \sigma^2(y) = 3.75 \quad \rho(x, y) = 0.8165$$

Coefficient de corrélation de Spearman

Échantillon : $(x_1, y_1), \dots, (x_n, y_n) \in \mathbb{R}^2$.

Objectif : étudier l'existence d'un lien entre les observations de x et de y sans faire l'hypothèse de linéarité.

Rangs

Soit $k \in \{1, \dots, n\}$, rx_k (resp. ry_k) désigne le rang de x_k (resp. y_k) dans la séquence des observations triées par ordre croissant. En cas d'égalité, le rang est donné par le rang moyen des observations égales.

Exemple : si $x_1 = 17$, $x_2 = 8$, $x_3 = 19$ et $x_4 = 81$, alors

$$rx_1 = 2, \quad rx_2 = 1, \quad rx_3 = 3, \quad rx_4 = 4.$$

Exemple : si $x_1 = 17$, $x_2 = 8$, $x_3 = 17$ et $x_4 = 81$, alors

$$rx_1 = 2.5, \quad rx_2 = 1, \quad rx_3 = 2.5, \quad rx_4 = 4.$$

Coefficient de corrélation de Spearman

Échantillon : $(x_1, y_1), \dots, (x_n, y_n) \in \mathbb{R}^2$.

Coefficient de corrélation de Spearman

$$\rho_S(x, y) = \rho(rx, ry) = \frac{\sigma(rx, ry)}{\sigma(rx)\sigma(ry)}$$

Si **tous les rangs sont distincts**, alors

$$\rho_S(x, y) = 1 - \frac{6}{n(n^2 - 1)} \sum_{k=1}^n (rx_k - ry_k)^2.$$

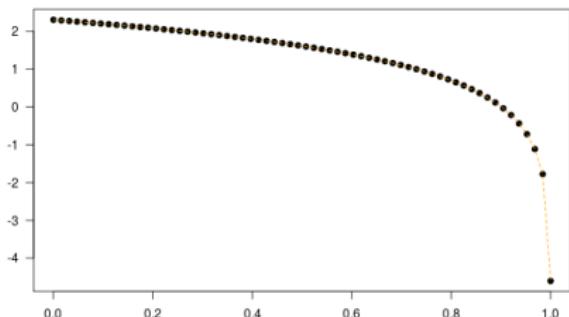
En effet, il est facile de voir dans ce cas que $\bar{rx} = \bar{ry} = (n + 1)/2$ et que $\sigma^2(rx) = \sigma^2(ry) = (n^2 - 1)/12$. Le résultat découle du calcul suivant,

$$\frac{1}{n} \sum_{k=1}^n (rx_k - ry_k)^2 = \frac{n^2 - 1}{6} - 2\sigma(rx, ry).$$

Coefficient de corrélation de Spearman

Échantillon : $(x_1, y_1), \dots, (x_n, y_n) \in \mathbb{R}^2$.

- Par construction, $\rho_S(x, y) \in [-1, 1]$.
- Si $|\rho_S(x, y)|$ est proche de 1, le lien entre les observations de x et y est donné par une **fonction monotone**.
- Le **signe** de $\rho_S(x, y)$ donne le sens de monotonie : croissante si $\rho_S(x, y) > 0$ et décroissante si $\rho_S(x, y) < 0$.



$$\rho(x, y) = -0.816$$
$$\rho_S(x, y) = -1$$

Coefficient de corrélation de Kendall

Échantillon : $(x_1, y_1), \dots, (x_n, y_n) \in \mathbb{R}^2$.

Objectif : étudier l'existence d'un lien entre les observations de x et de y sans faire l'hypothèse de linéarité.

Concordance

Soit $k_1, k_2 \in \{1, \dots, n\}$, les couples d'observations (x_{k_1}, y_{k_1}) et (x_{k_2}, y_{k_2}) sont dits **concordants** si

$x_{k_1} < x_{k_2}$ et $y_{k_1} < y_{k_2}$ ou $x_{k_1} > x_{k_2}$ et $y_{k_1} > y_{k_2}$,

ou **discordants** si

$x_{k_1} < x_{k_2}$ et $y_{k_1} > y_{k_2}$ ou $x_{k_1} > x_{k_2}$ et $y_{k_1} < y_{k_2}$.

Si $x_{k_1} = x_{k_2}$ ou $y_{k_1} = y_{k_2}$, le couple n'est ni concordant, ni discordant.

Coefficient de corrélation de Kendall

Échantillon : $(x_1, y_1), \dots, (x_n, y_n) \in \mathbb{R}^2$.

Coefficient de corrélation de Kendall

$$\rho_K(x, y) = \frac{2(R_+ - R_-)}{n(n-1)}$$

où R_+ (resp. R_-) est le nombre de couples concordants (resp. discordants).

- Un simple argument combinatoire donne $\rho_K(x, y) \in [-1, 1]$.
- Une valeur $|\rho_K(x, y)|$ proche de 1 suggère un lien **monotone** entre les observations de x et de y .
- L'interprétation est similaire à celle du coefficient de corrélation de Spearman.

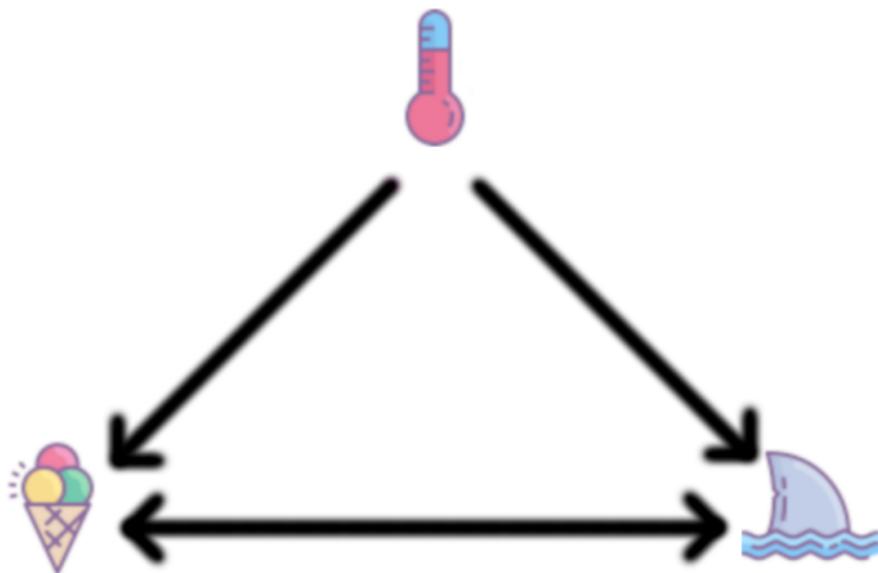
Cum hoc sed non propter hoc (Avec ceci mais pas à cause de ceci)

« Corrélation n'est pas causalité »

Quelques exemples :

- Le fait de dormir avec ses chaussures est fortement corrélé avec celui de se réveiller avec la « gueule de bois ».
→ Dormir avec des chaussures donne-t-il la « gueule de bois » ? Un autre facteur est-il impliqué ?
- La fréquence des attaques de requins est fortement corrélée avec la vente de glaces sur la plage.
→ Est-on plus appétissant pour les requins en ayant mangé de la glace ?
- Les personnes qui meurent ont très souvent vu un médecin dans les jours qui ont précédé.
→ Est-ce dangereux de rencontrer un médecin ?

Corrélation n'est pas causalité



Corrélation partielle

Échantillon : $(x_1, y_1, z_1), \dots, (x_n, y_n, z_n) \in \mathbb{R}^3$.

La **corrélation partielle** entre x et y **conditionnellement** à z est la corrélation linéaire entre les résidus ε^x et ε^y des régressions linéaires sur (x, z) et (y, z) respectivement.

La corrélation partielle permet d'évaluer la corrélation entre les observations de deux variables après avoir contrôlé l'effet perturbateur d'une ou de plusieurs autres variables.

Corrélation partielle

Échantillon : $(x_1, y_1, z_1), \dots, (x_n, y_n, z_n) \in \mathbb{R}^3$.

La **corrélation partielle** entre x et y **conditionnellement** à z est la corrélation linéaire entre les résidus ε^x et ε^y des régressions linéaires sur (x, z) et (y, z) respectivement.

Corrélation linéaire

	Temp	Glace	Requin
Temp	1.00	0.95	0.93
Glace	.	1.00	0.88
Requin	.	.	1.00

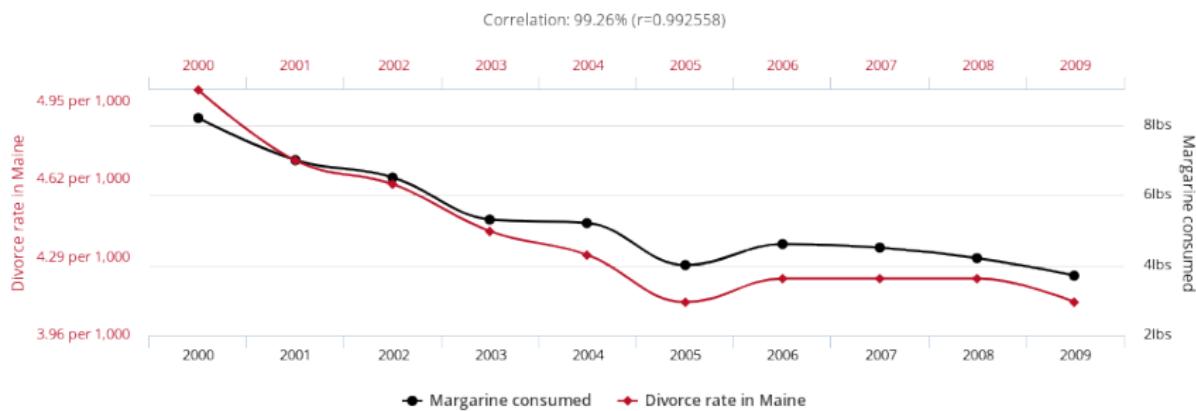
Corrélation partielle

	Temp	Glace	Requin
Temp	1.00	0.74	0.66
Glace	.	1.00	0.04
Requin	.	.	1.00

Régression linéaires entre x et y et y et z puis on fait la corrélation sur les résidus

Spurious correlation

Divorce rate in Maine
 correlates with
Per capita consumption of margarine



Data sources: National Vital Statistics Reports and U.S. Department of Agriculture

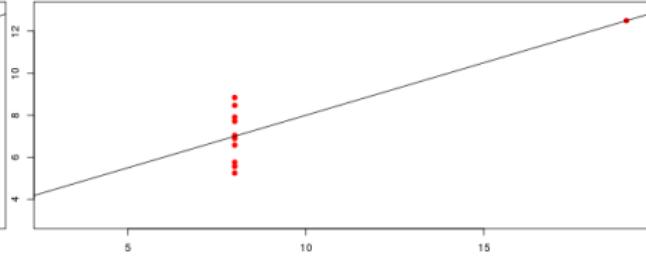
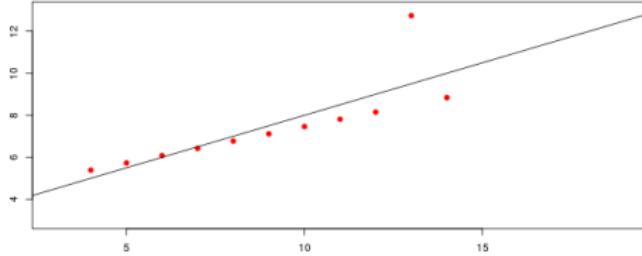
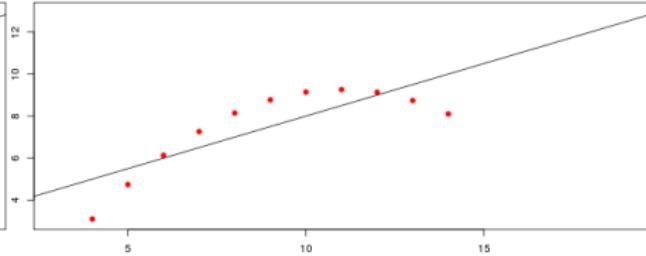
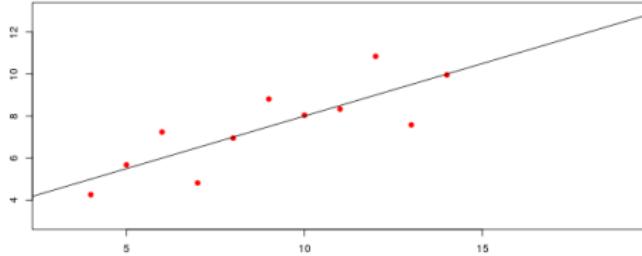
<http://www.tylervigen.com/spurious-correlations>

1.2 Représentaions graphiques

Pourquoi visualiser ?

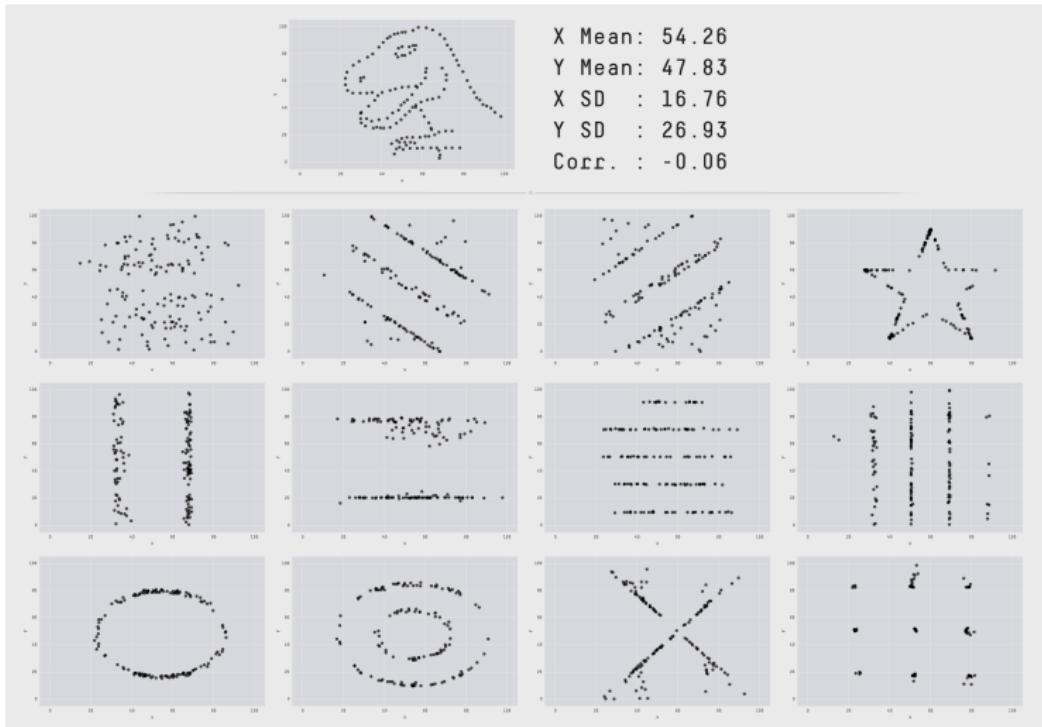
... make both calculations and graphs. Both sorts of output should be studied; each will contribute to understanding.

F. J. Anscombe, 1973



$$\bar{x} = 9.0 \quad \sigma^2(x) = 10.0 \quad \bar{y} = 7.5 \quad \sigma^2(y) = 3.75 \quad \rho(x, y) = 0.8165$$

Datasaurus (Alberto Cairo)



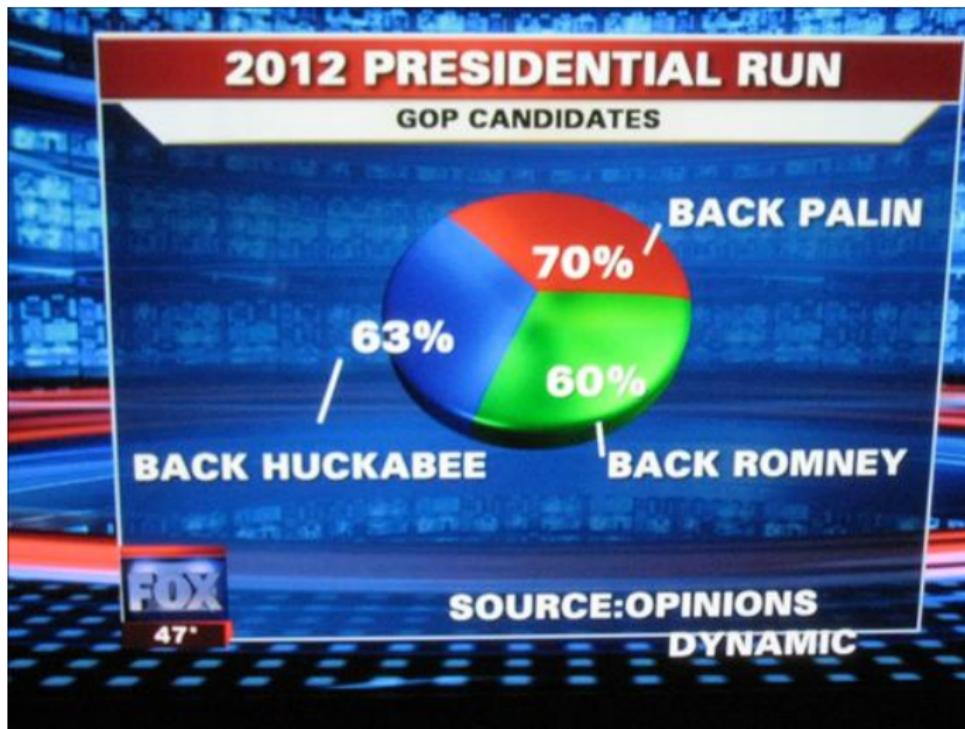
<https://www.autodeskresearch.com/publications/samestats>

The Visual Display of Quantitative Information (E. Tufte, 1983)

Tufte a popularisé plusieurs bonnes pratiques graphiques :

- Montrer les données.
- Inciter celui ou celle qui regarde à penser.
- Éviter de distordre ce que les données ont à dire.
- Présenter beaucoup de données sur une petite surface.
- Révéler les données à des niveaux différents : d'un aperçu global à des structures plus fines.
- Servir un objectif clair et raisonnable.
- Être étroitement intégré à une description statistique du jeu de données.

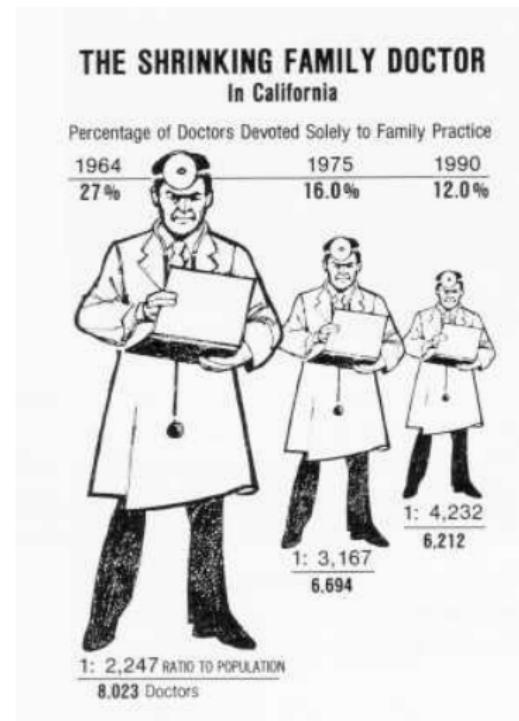
Mauvaise visualisation I



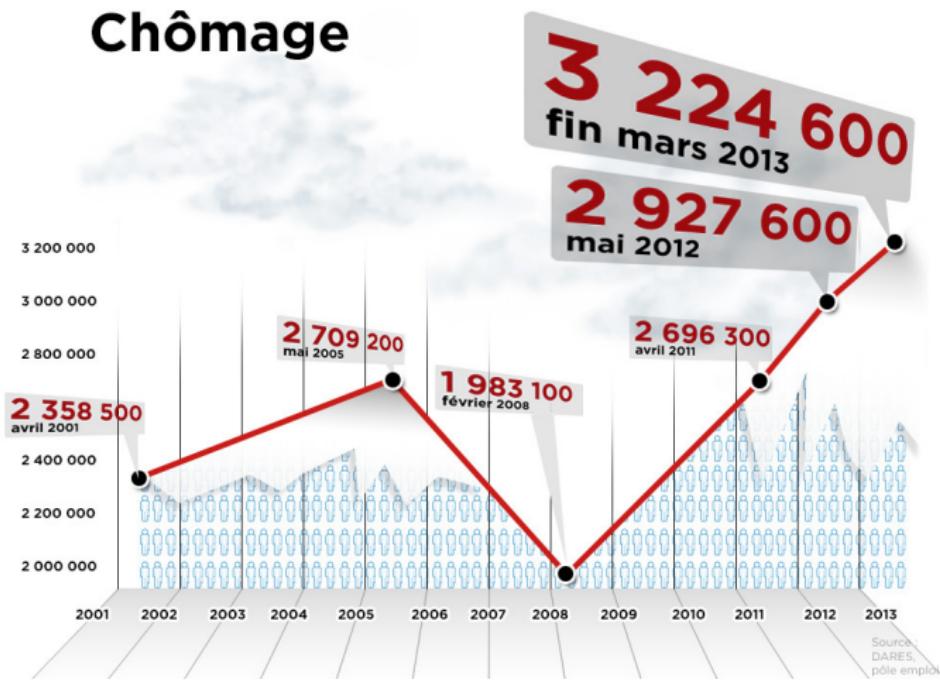
Mauvaise visualisation II



Mauvaise visualisation III



Mauvaise visualisation IV



Données de type « effectif »

A	30
B	15
C	30
D	20
E	25



Diagramme en bâtons (pile)

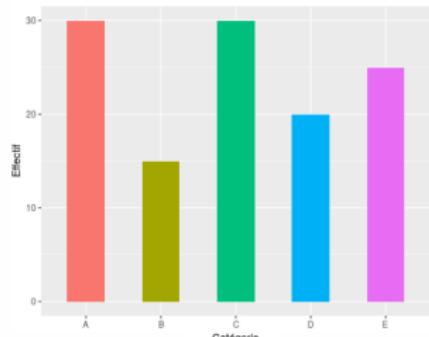
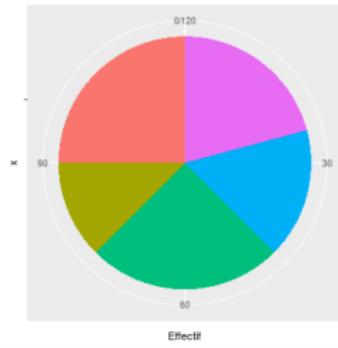
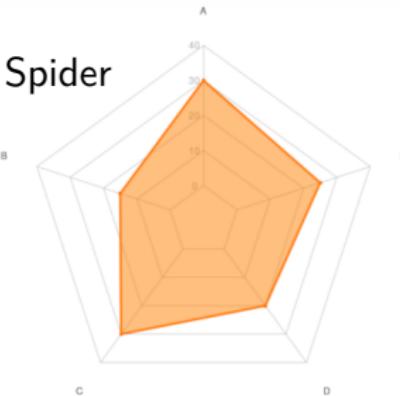


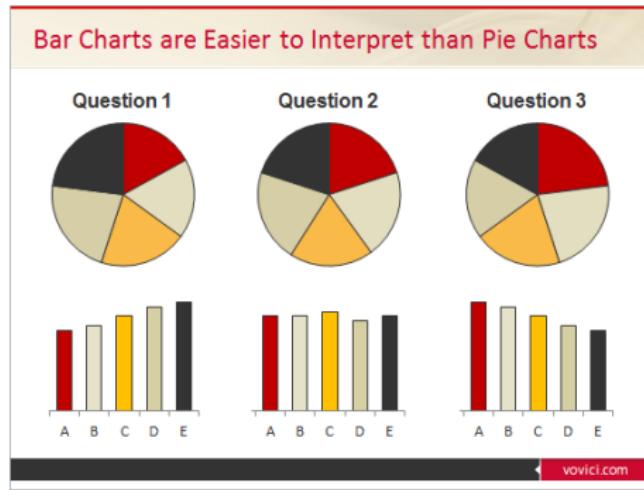
Diagramme en bâtons

Radar / Spider



Camembert
(Pie)

Les problèmes du camembert



Lisibilité difficile : les diagrammes en bâtons sont souvent préférables

Attention aux représentations en 3D.
Est-ce évident que A et C ont la même valeur ?



Données de type « effectif » (2D)

	X	Y
A	30	25
B	15	15
C	30	20
D	20	30
E	25	35

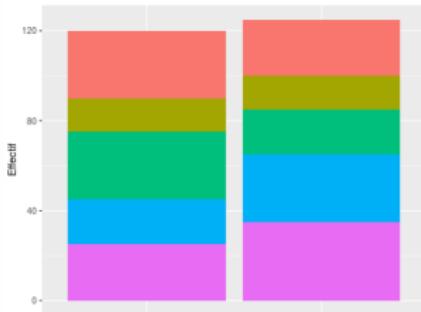


Diagramme en bâtons (pile)

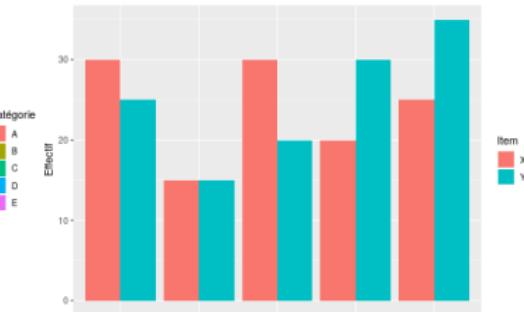
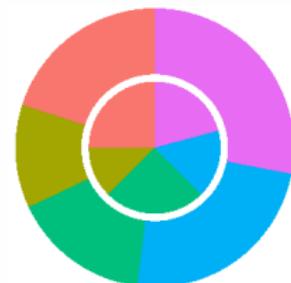
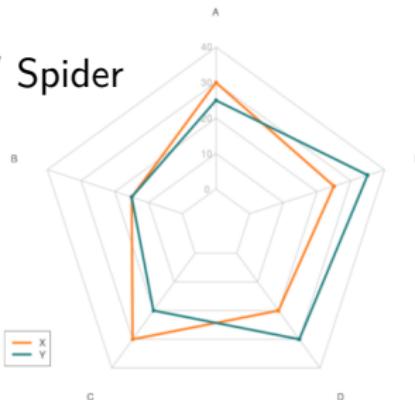


Diagramme en bâtons

Radar / Spider



Catégorie

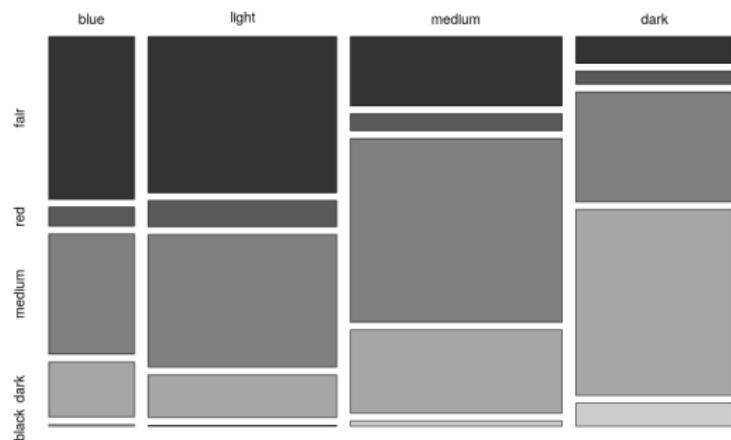
Tore
(Donut)

Effectifs croisés de deux variables qualitatives

Données MASS::caith de R (**table de contingence**) :

	fair	red	medium	dark	black
blue	326	38	241	110	3
light	688	116	584	188	4
medium	343	84	909	412	26
dark	98	48	403	681	85

Diagramme mosaïque

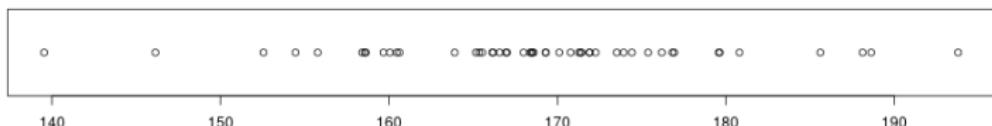


Données réelles

Taille de 50 individus en centimètres :

166.9	159.7	174.4	188.1	166.2	158.6	139.5	193.8	167.0	168.5
171.3	179.6	165.2	171.4	169.3	176.8	168.0	158.6	160.1	168.6
155.8	170.1	158.4	173.5	172.3	146.1	170.8	176.2	185.6	165.4
171.4	160.6	168.5	171.9	169.3	171.9	166.6	163.9	154.5	179.6
176.9	180.8	175.4	166.1	165.5	188.6	168.4	173.9	152.6	160.5

Nuage de points
(strip chart)

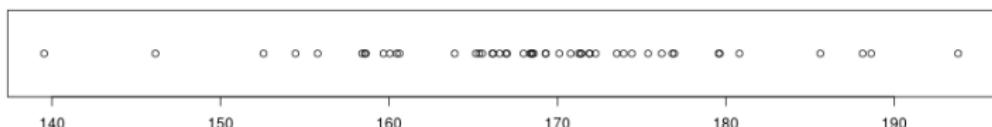


Données réelles

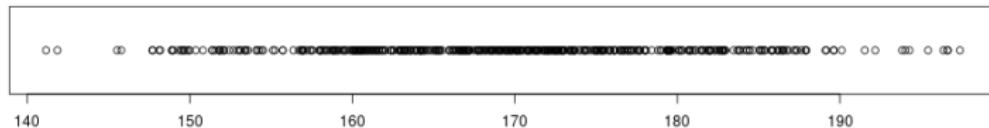
Taille de 50 individus en centimètres :

166.9	159.7	174.4	188.1	166.2	158.6	139.5	193.8	167.0	168.5
171.3	179.6	165.2	171.4	169.3	176.8	168.0	158.6	160.1	168.6
155.8	170.1	158.4	173.5	172.3	146.1	170.8	176.2	185.6	165.4
171.4	160.6	168.5	171.9	169.3	171.9	166.6	163.9	154.5	179.6
176.9	180.8	175.4	166.1	165.5	188.6	168.4	173.9	152.6	160.5

Nuage de points
(strip chart)



Avec 500 observations, la lisibilité devient délicate ...

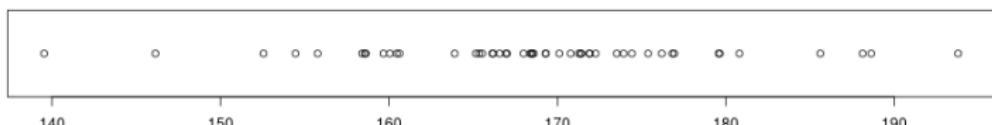


Données réelles

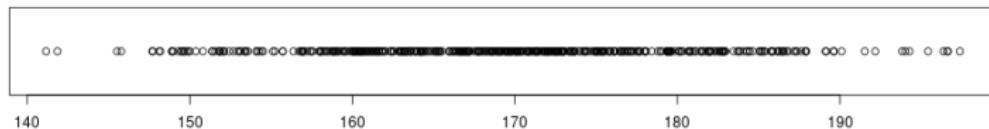
Taille de 50 individus en centimètres :

166.9	159.7	174.4	188.1	166.2	158.6	139.5	193.8	167.0	168.5
171.3	179.6	165.2	171.4	169.3	176.8	168.0	158.6	160.1	168.6
155.8	170.1	158.4	173.5	172.3	146.1	170.8	176.2	185.6	165.4
171.4	160.6	168.5	171.9	169.3	171.9	166.6	163.9	154.5	179.6
176.9	180.8	175.4	166.1	165.5	188.6	168.4	173.9	152.6	160.5

Nuage de points
(strip chart)



Avec 500 observations, la lisibilité devient délicate ...



... et il est nécessaire de **résumer** l'information affichée.



Boxplot

Préciser les valeurs utilisées car pas de convention

Points extrêmes

++

$$q_1 - 1.5(q_3 - q_1) \quad (\text{ou minimum})$$



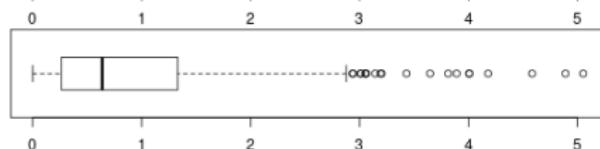
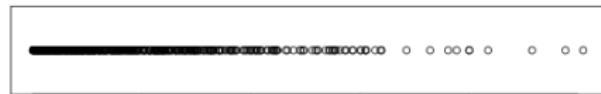
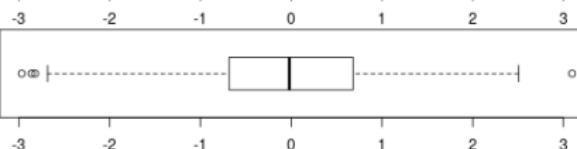
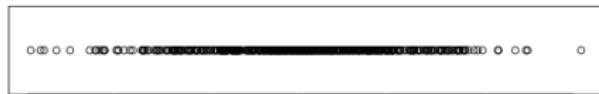
$$q_3$$

$$q_3 + 1.5(q_3 - q_1) \quad (\text{ou maximum})$$

Écart interquartile $q_3 - q_1$

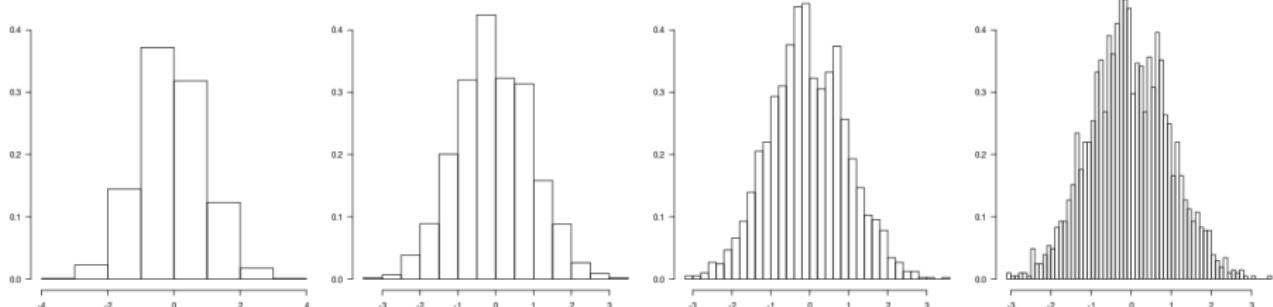


Notations : premier quartile q_1 , médiane q_2 et troisième quartile q_3 .



Histogramme et noyau

Échantillon réel : $x_1, \dots, x_n \in \mathbb{R}$.



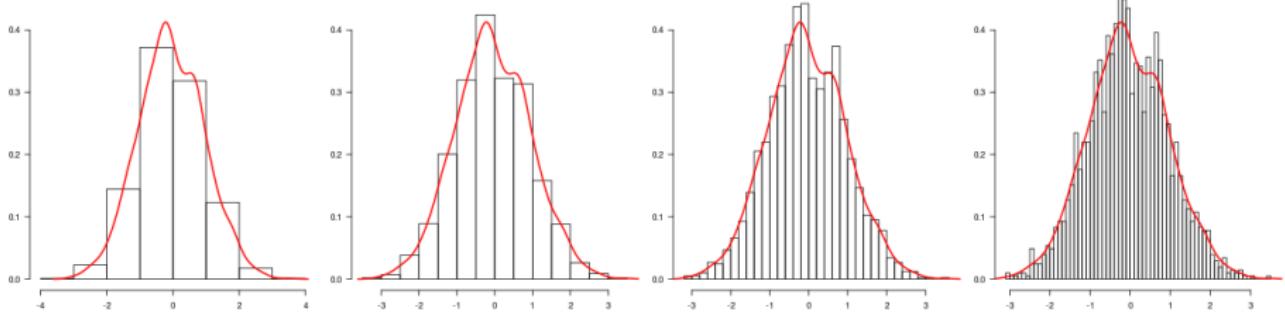
Pour un **histogramme** construit sur des blocs de taille $h > 0$, la valeur prise sur le segment $[b - h/2, b + h/2[$ vaut

$$\frac{\text{Nombre de points dans } [b - h/2, b + h/2[}{\text{Nombre de points total}} = \frac{1}{n} \sum_{k=1}^n \mathbf{1}_{|b-x_k| \leqslant h/2}$$

⇒ La surface de l'histogramme vaut 1 (**fonction de densité**)

Histogramme et noyau

Échantillon réel : $x_1, \dots, x_n \in \mathbb{R}$.



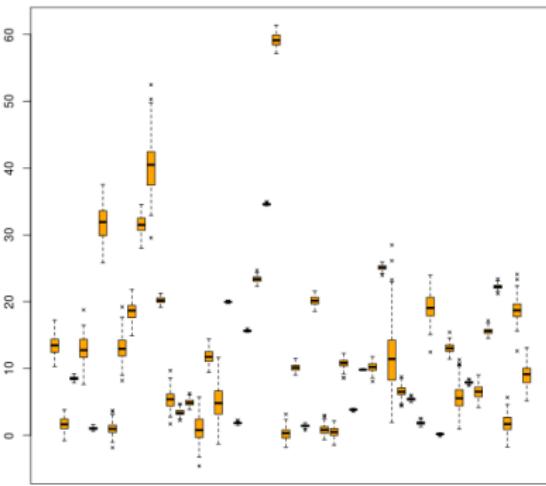
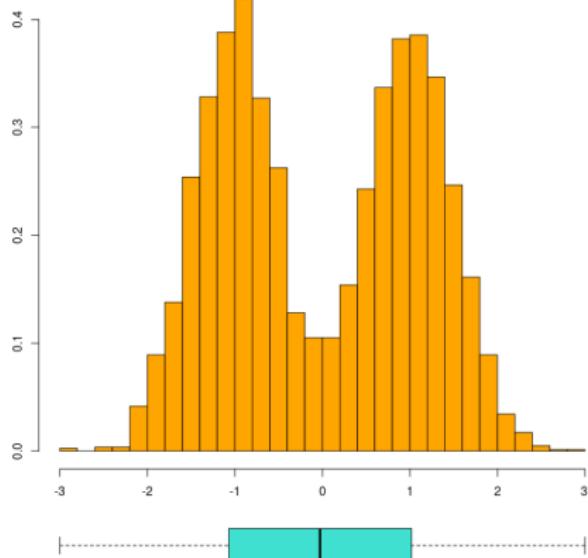
Un **noyau** $K : \mathbb{R} \rightarrow [0, \infty[$ est une fonction paire telle que $\int_{\mathbb{R}} K(t)dt = 1$.
 L'estimateur par noyau de la distribution des observations est donné par

$$\hat{f}_K(t) = \frac{1}{n} \sum_{k=1}^n K(t - x_k).$$

⇒ La surface sous la courbe vaut 1 (**fonction de densité**)

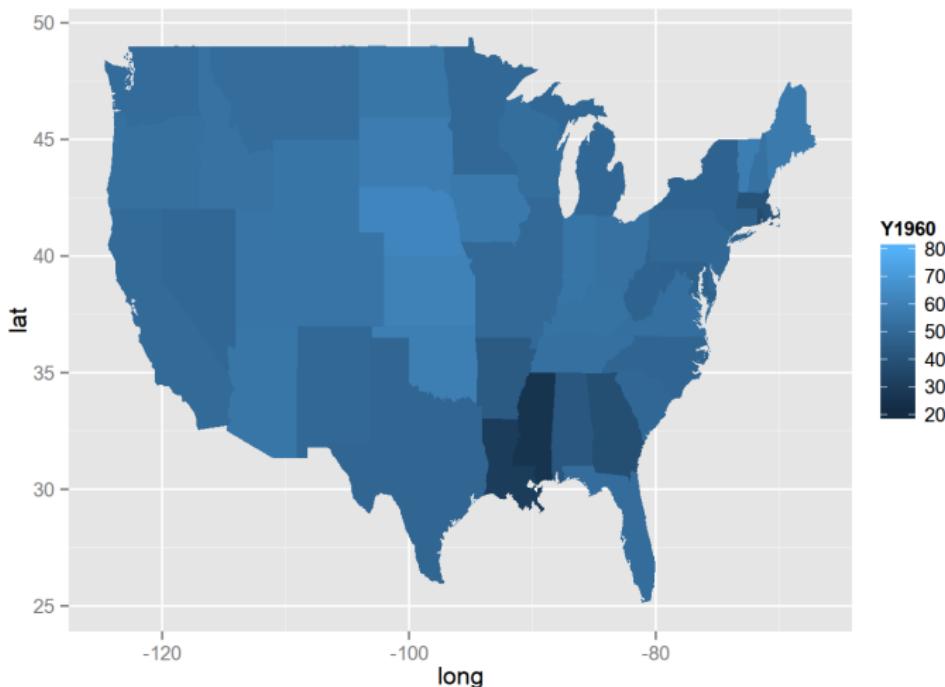
Boxplot versus Histogramme

Un histogramme permet de capter des distributions multimodales ...



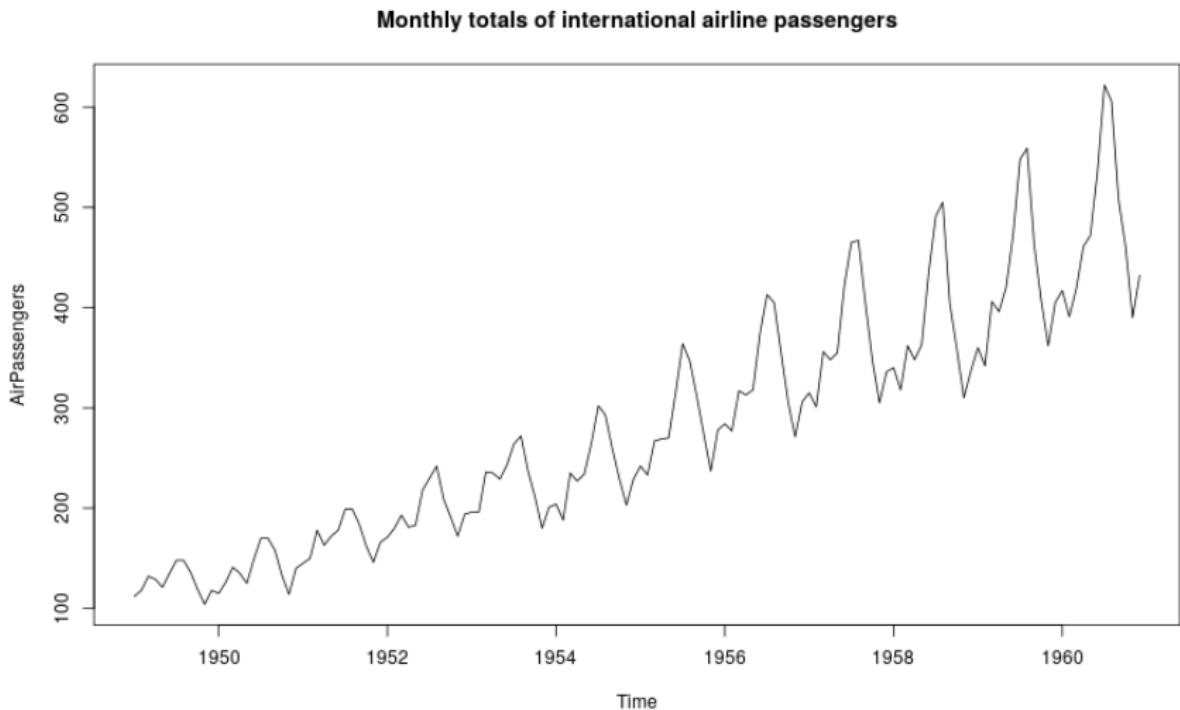
... mais des boxplots permettent de rendre lisibles de nombreuses variables sur un même graphique.

Données réelles dans l'espace



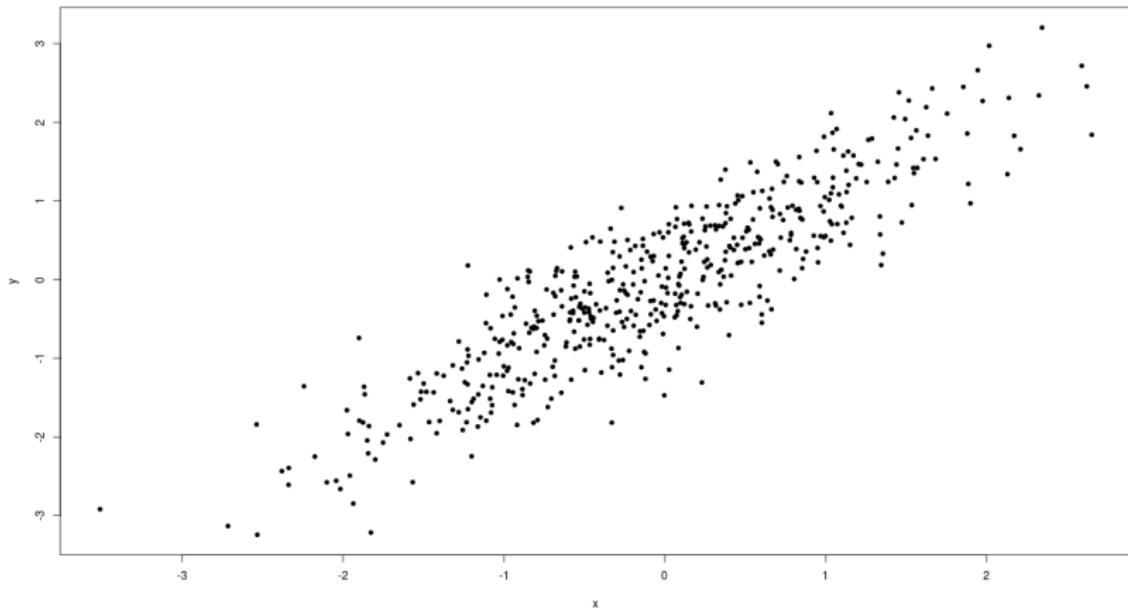
Carte choroplète

Données réelles dans le temps (Série chronologique)



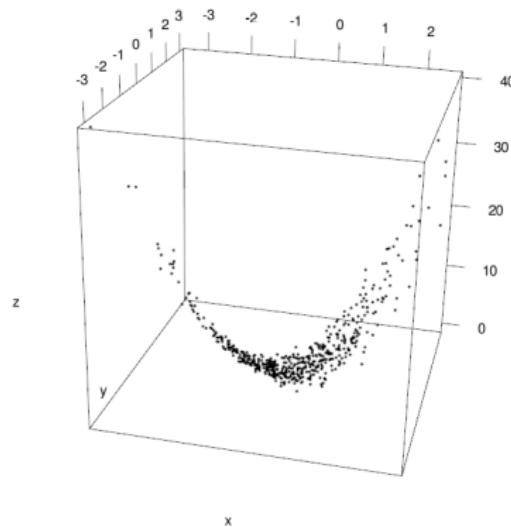
Données quantitatives (2D)

Dans le cas d'observations bidimensionnelles $(x_1, y_1), \dots, (x_n, y_n) \in \mathbb{R}^2$, il est possible de tracer un **nuage de points** (ou **scatter plot**).



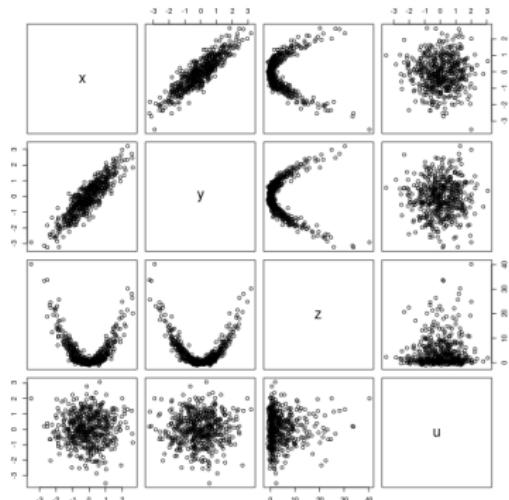
Données quantitatives (3D)

Pour des observations tridimensionnelles $(x_1, y_1, z_1), \dots, (x_n, y_n, z_n) \in \mathbb{R}^3$, il est encore possible de considérer un **nuage de points** en utilisant des outils de visualisation en 3D (e.g. le package rgl avec R).



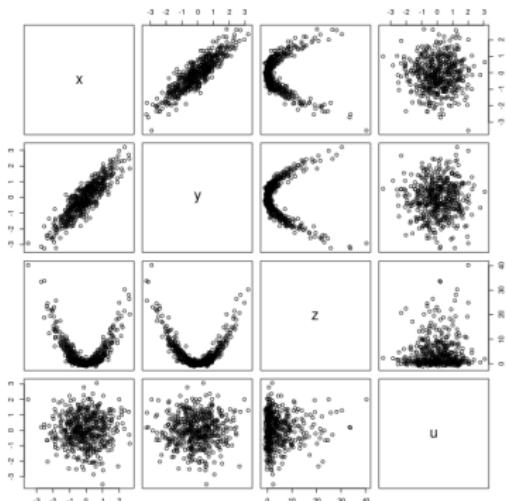
Données quantitatives multidimensionnelles

À partir de la dimension 3, il devient difficile de proposer une représentation graphique simple des données. Une possibilité consiste à tracer les nuages de points associés à chaque paire de variables observées.



Données quantitatives multidimensionnelles

À partir de la dimension 3, il devient difficile de proposer une représentation graphique simple des données. Une possibilité consiste à tracer les nuages de points associés à chaque paire de variables observées.

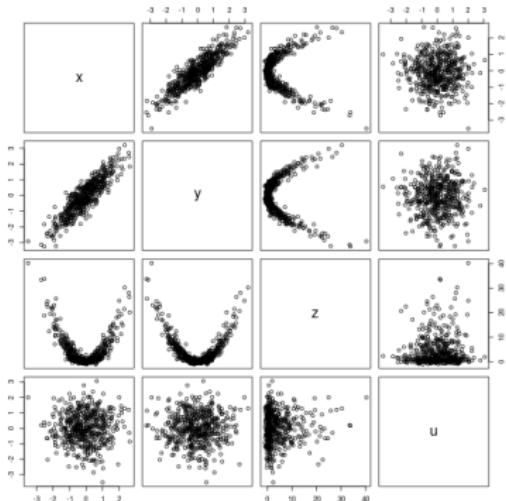


Acceptable pour une première exploration grossière mais plusieurs inconvénients :

- lisibilité difficile pour un nombre important de variables,
- étude limitée à des paires de variables,
- projections uniquement sur des plans parallèles aux axes,
- ...

Données quantitatives multidimensionnelles

À partir de la dimension 3, il devient difficile de proposer une représentation graphique simple des données. Une possibilité consiste à tracer les nuages de points associés à chaque paire de variables observées.



Nous avons besoin d'une méthode systématique et plus souple.