Affaires mondiales Canada | Global Affairs Canada

Canada

CANADIAN FOREIGN SERVICE INSTITUTE | L'INSTITUT CANADIEN DU SERVICE EXTÉRIEUR

Introduction à l'analyse des données

TECHNIQUES DE BASE D'ANALYSE DES DONNÉES

Patrick Boily
Data Action Lab | uOttawa | Idlewyld Analytics
pboily@uottawa.ca

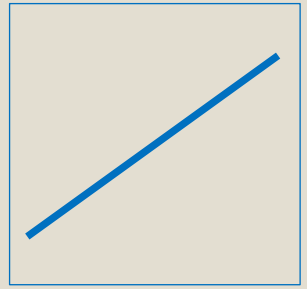
[en collaboration avec Jen Schellinck | Sysabee]

VARIABLES DÉPENDANTES ET INDÉPENDANTES

Dans un *contexte expérimental* :

- **les variables de contrôle/extrinsèque :** nous faisons de notre mieux pour qu'elles restent contrôlées et immuables alors que d'autres variables sont modifiées
- **variables indépendantes :** nous contrôlons leurs valeurs car nous pensons qu'elles influencent les variables dépendantes
- **variables dépendantes :** nous ne contrôlons pas leurs valeurs ; elles sont générées d'une manière ou d'une autre pendant l'expérience et dépendent vraisemblablement de tout.

Comment cela se traduit-il dans d'autres ensembles de données ?



TYPES DE DONNÉES

Données numériques : nombres entiers ou nombres continus

- 1, 7, 34.654, 0.000004

Données textuelles : chaînes de texte - peuvent être limitées à un certain nombre de caractères

- “Bienvenue au parc”, “AAAAA”, “345”, “45.678”

Données catégorielles : un nombre fixe de valeurs, qui peuvent être numériques ou représentées par des chaînes de caractères**. Il n'y a pas d'ordre spécifique ou inhérent**

- (‘rouge’,’bleu’,vert’), (‘1’,’2’,’3’)

Données ordinales : données catégorielles avec un ordre inhérent. Contrairement aux données en nombre entier, **l'espacement entre les valeurs n'est pas défini**

- (très froid, froid, tiède, chaud, très chaud)

COMMENT RÉSUMER LES DONNÉES

Min : la plus petite valeur

Max : la plus grande valeur

Médiane : valeur « centrale »

Mode : valeur la plus fréquente

Valeurs uniques : liste des valeurs uniques

etc.

Signal	Type
4.31	Bleu
5.34	Orange
3.79	Bleu
5.19	Bleu
4.93	Vert
5.76	Orange
3.25	Orange
7.12	Orange
2.85	Bleu

COMMENT RÉSUMER LES DONNÉES

Nous pouvons effectuer des opérations sur un ensemble de données, (généralement sur ses **colonnes**).

Une telle opération revient à **condenser** les nombreuses valeurs des données en une seule valeur représentative.

Exemples : « moyenne », « somme », « compte », « variance », etc.

Nous pouvons appliquer la même fonction d'agrégation à de nombreuses colonnes différentes, ce qui permet d'obtenir un **mapping** (liste) des colonnes vers les valeurs.

Signal	Type
4.31	Bleu
5.34	Orange
3.79	Bleu
5.19	Bleu
4.93	Vert
5.76	Orange
3.25	Orange
7.12	Orange
2.85	Bleu

Compte	Signal moy.	Écart type signal	Mode type
9	4.73	1.33	Bleu/ Orange

TABLEAUX DE CONTINGENCE / TABLEAUX CROISÉS DYNAMIQUES

Tableau de contingence : un tableau qui examine la relation entre deux variables catégorielles par le biais de leur rapport (tableau croisé).

Tableau croisé dynamique : tableau généré en appliquant des opérations (somme, compte, moyenne, etc.) à des variables, éventuellement basées sur une autre variable (catégorielle). Les tableaux de contingence sont des cas particuliers de tableaux croisés dynamiques.

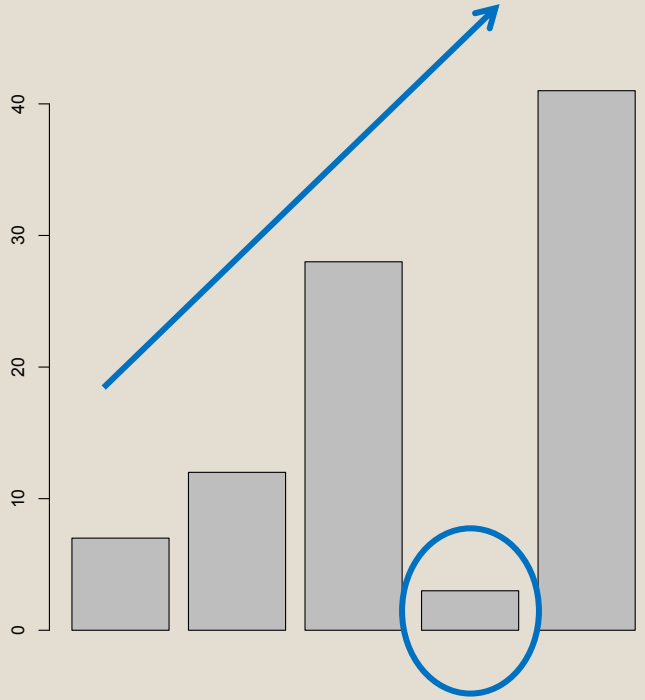
	Grande	Moyen	Petit
Fenêtre	1	32	31
Porte	14	11	0

Type	Compte	Signal moy.	Écart type signal
Bleu	4	4.04	0.98
Vert	1	4.93	N.A.
Orange	4	5.37	1.60

L'ANALYSE PAR LA VISUALISATION

Analyse (au sens large) :

- identifier des modèles ou une structure
- ajouter du sens à ces modèles ou à cette structure en les interprétant dans le contexte du système.



Options:

1. utiliser des méthodes analytiques pour y parvenir.

2. visualiser les données et utiliser le pouvoir analytique du cerveau (perceptuel) pour tirer des conclusions significatives sur ces schémas.

Nous en discuterons plus en détail.

DESCRIPTION DES DONNÉES (REPRISE)

En un sens, la raison sous-jacente de toute **analyse** est de parvenir à la compréhension des données.

Les études et les expériences donnent lieu à des **unités**, qui sont généralement décrites par des **variables** (et des mesures).

Les variables sont soit **qualitatives** (catégoriques), soit **quantitatives** (numériques) :

- les variables catégorielles prennent des valeurs (niveaux) dans un ensemble fini de classes
- les variables numériques prennent des valeurs dans un ensemble (potentiellement infini) de **quantités**

RÉSUMÉS NUMÉRIQUES

Pour commencer, une variable peut être décrite selon deux dimensions : **la centralité** et **la dispersion** (**l'asymétrie** et **l'aplatissement** sont aussi parfois utilisés).

Les mesures de **centralité** comprennent:

- **médiane, moyenne, mode** (moins fréquemment)

Les mesures de **dispersion** comprennent :

- **écart-type, variance, quartiles, écart interquartile, plage de données** (moins fréquemment)

La médiane, la plage de données et les quartiles sont facilement calculés à partir d'une **liste ordonnée** de données.

RÉSUMÉS VISUELS – BOÎTE À MOUSTACHES

Le **boîte à moustache** (« boxplot ») est un moyen rapide de présenter un résumé graphique d'une distribution univariée.

Dessinez une boîte le long de l'axe d'observation, avec des extrémités à Q_1 et Q_3, et avec une « ceinture » à la médiane.

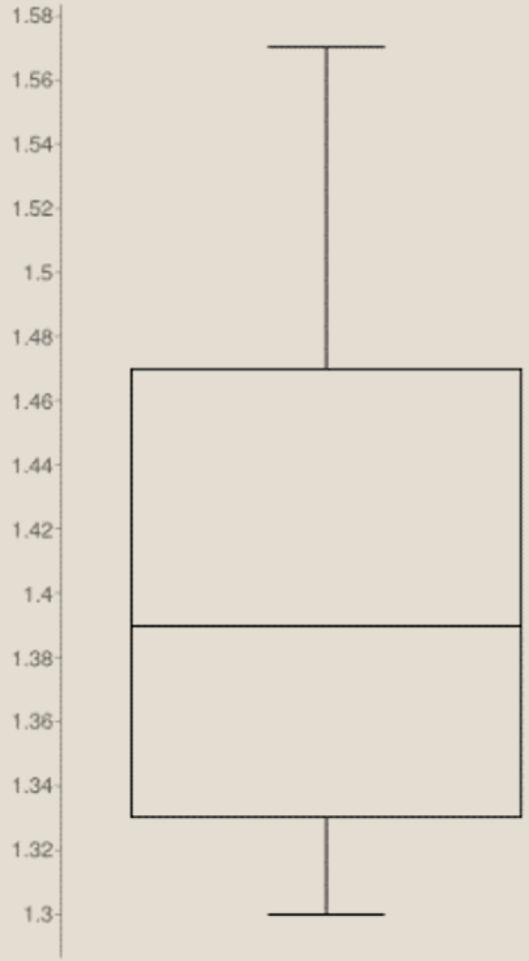
Tracez une ligne s'étendant de Q_1 à la plus petite observation inférieure à $1.5 \times \mathrm{IQR}$ à gauche de Q_1.

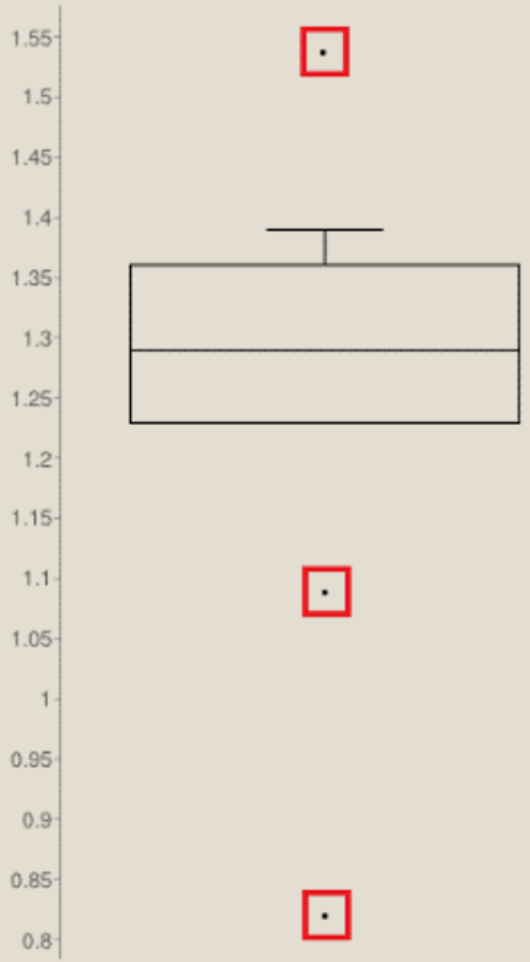
Tracez une ligne s'étendant de Q_3 à la plus petite observation située à plus de $1.5 \times \mathrm{IQR}$ à droite de Q_3.

Toute valeur aberrante présumée est tracée séparément.

EXEMPLES

Ensemble de données sur les files d'attente : taux d'arrivée (à gauche), taux de traitement (à droite)





RÉSUMÉS VISUELS – HISTOGRAMME

Les histogrammes peuvent également fournir une indication de la distribution d'une variable.

Ils doivent inclure/contenir les informations suivantes :

- la plage de l'histogramme est $r = Q_4 - Q_0$;
- le nombre de cases (« bins ») doit approcher $k = \sqrt{n}$, où n est le nombre d'observations ;
- la largeur du case doit approcher r/k, et
- la fréquence des observations dans chaque case doit être ajoutée au graphique.

EXEMPLE

Considérons le nombre quotidien d'accidents de voiture à Sydney sur une période de 40 jours :

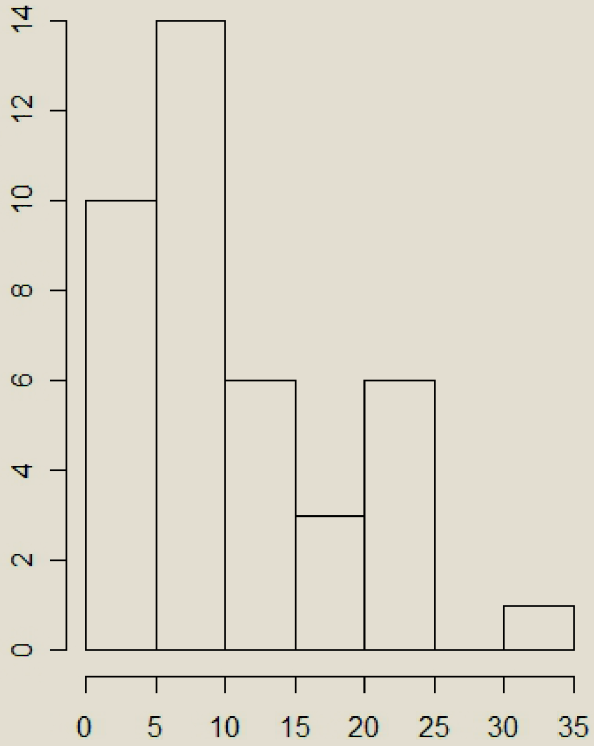
6, 3, 2, 24, 12, 3, 7, 14, 21, 9, 14, 22, 15,
2, 17, 10, 7, 7, 31, 7, 18, 6, 8, 2, 3, 2, 17,
7, 7, 21, 13, 23, 1, 11, 3, 9, 4, 9, 9, 25

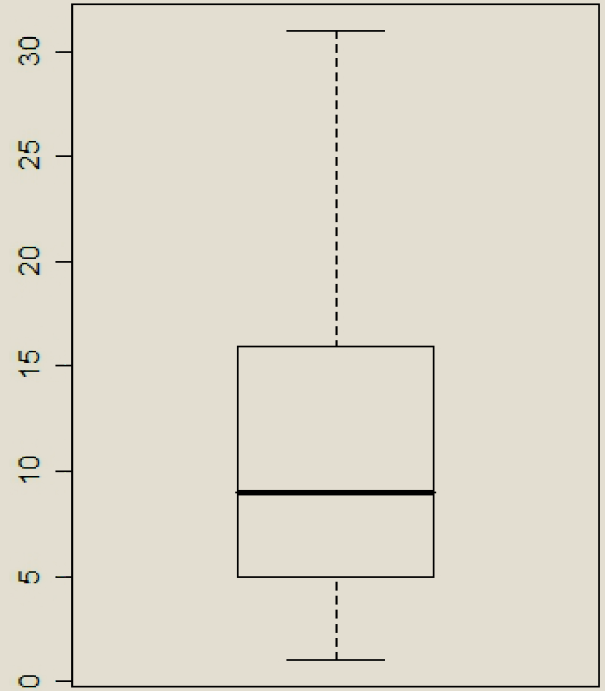
Les valeurs ordonnées sont :

1 2 2 2 2 3 3 3 3 4 6 6 7 7 7 7 7
7 8 9 9 9 9 10 11 12 13 14 14 15 17
17 18 21 21 22 23 24 25 31

min	Q_1	med	Q_3	max
1	5.5	9	15.5	31

Est-il plus probable que l'on observe entre 5 et 15 accidents un jour donné, ou entre 25 et 35 ?





EXEMPLE

Considérons les données suivantes, constituées de $n = 20$ mesures appariées (x_i, y_i) des niveaux d'hydrocarbures (x) et des niveaux d'oxygène pur (y) dans les carburants :

```
x:  0.99  1.02  1.15  1.29  1.46  1.36  0.87  1.23  1.55  1.40
y: 90.01 89.05 91.43 93.74 96.73 94.45 87.59 91.77 99.42 93.65

x:  1.19  1.15  0.98  1.01  1.11  1.20  1.26  1.32  1.43  0.95
y: 93.54 92.52 90.56 89.54 89.85 90.39 93.25 93.41 94.98 87.33
```

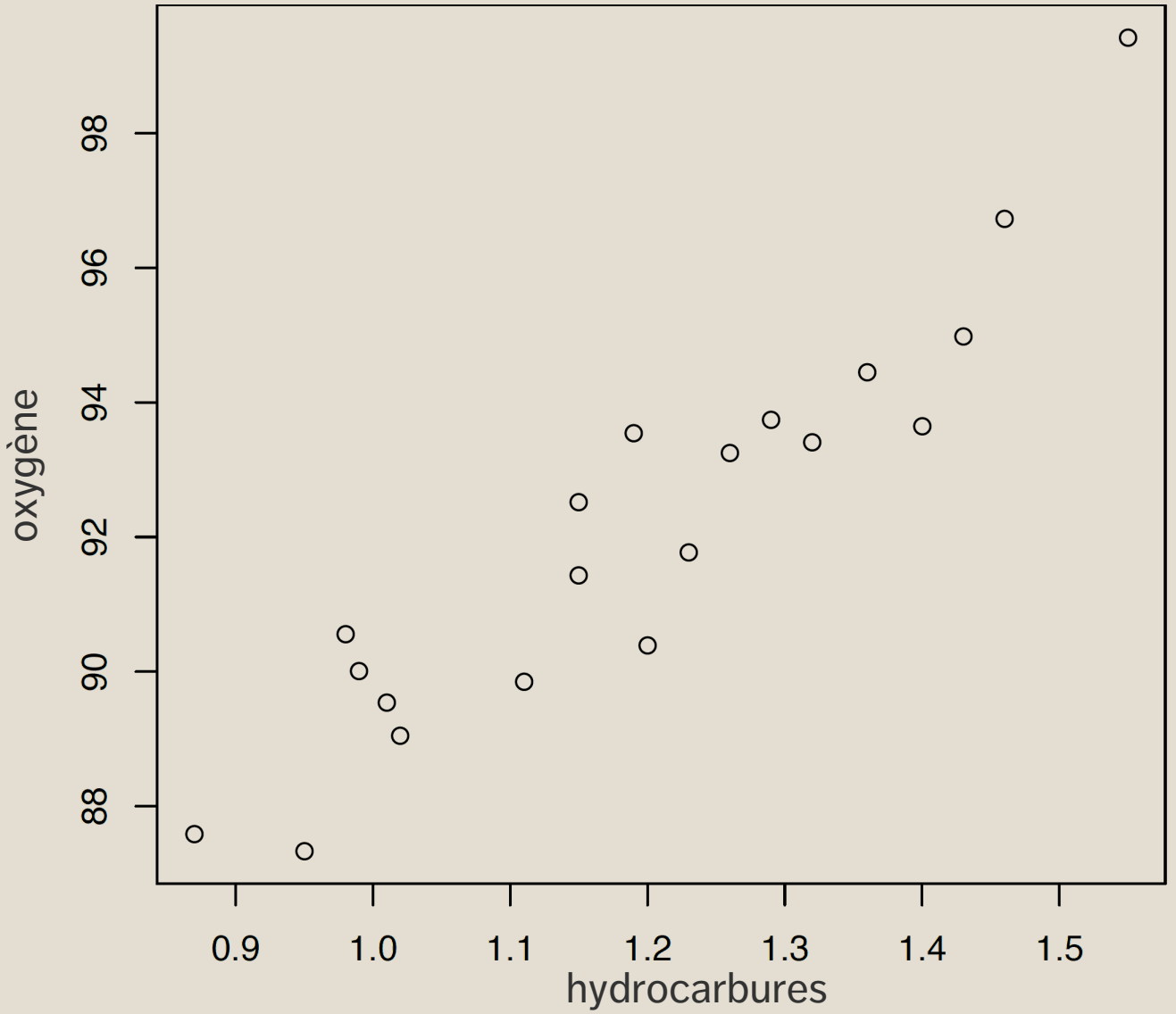
Objectifs :

- mesurer la **force de l'association** entre x et y
- **décrire** la relation entre x et y

EXEMPLE

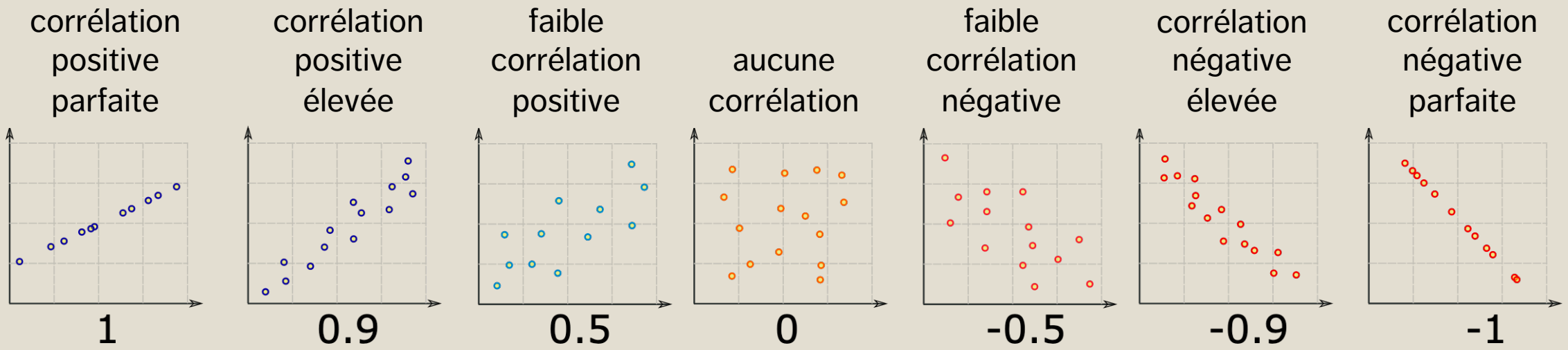
Une représentation graphique fournit une première description de la relation.

Il semble que les points se situent autour d'une ligne cachée !

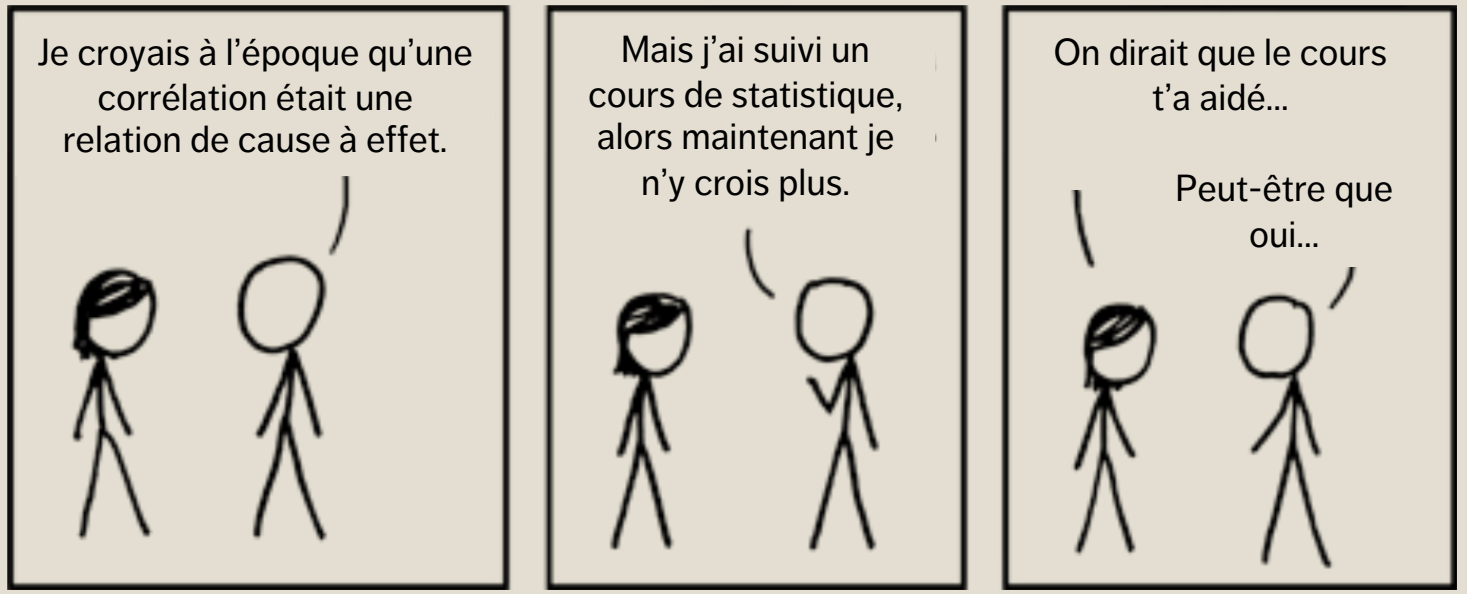


[Auteur inconnu]

PROPRIÉTÉS ET INTERPRÉTATION



[https://xkcd.com/552/]

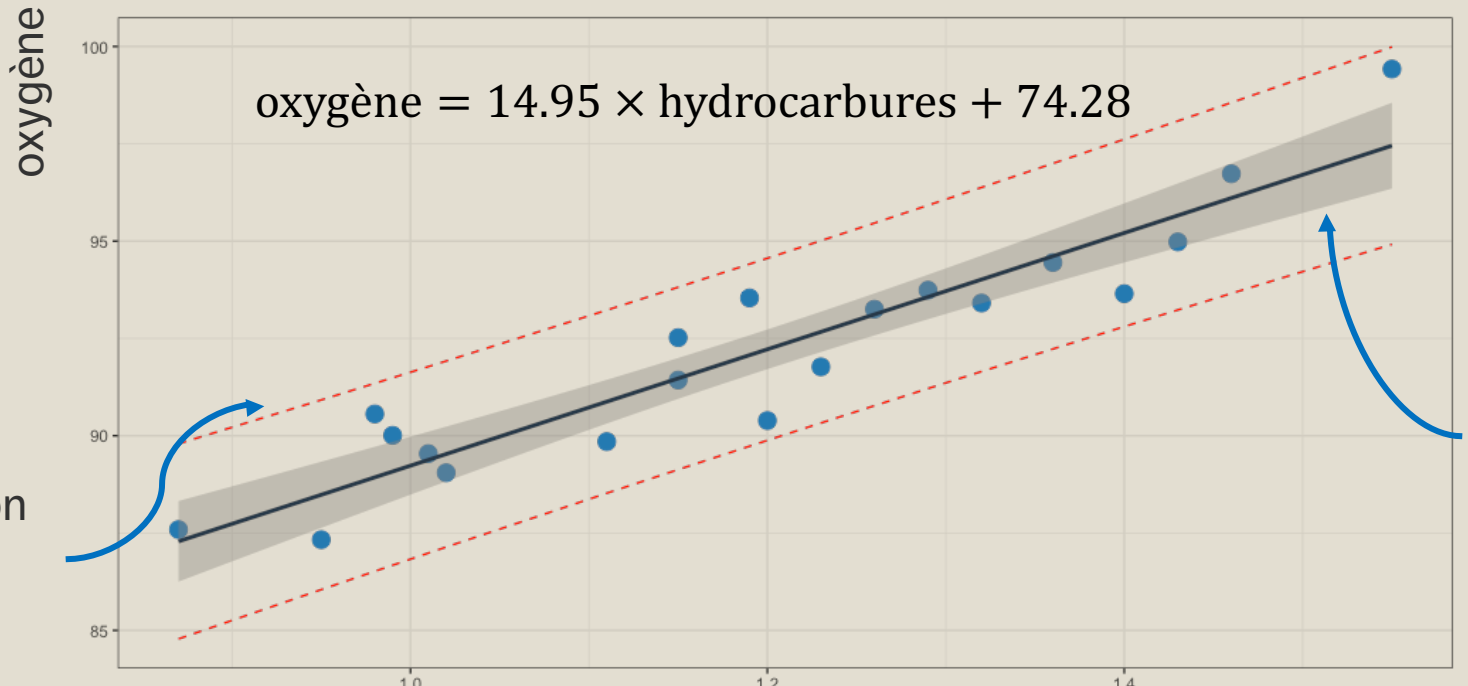


La corrélation n'implique pas la causalité, mais elle agite les sourcils de manière suggestive et fait des gestes furtifs en disant « regardez par là ».

RÉGRESSION LINÉAIRE

Si $\hat{\beta}_i$ est l'estimation du coefficient β_i réel, le modèle de **régression linéaire** associé aux données est le suivant

$$\widehat{\boldsymbol{Y}}(\boldsymbol{x}) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_p x_p = \boldsymbol{\beta x}$$



intervalle de prédiction pour les nouvelles réponses

intervalle de confiance pour la réponse moyenne

hydrocarbures