



Introduction à la science des données

Instructeur: Patrick Boily



uOttawa

Institut de développement professionnel
Professional Development Institute

Patrick Boily

Carrière :

Professeur [uOttawa] (~55 cours/ ~150 journées d'atelier)

Gérant ['12 – '19, CQADS/CAQAD, Carleton]

Fonctionnaire ['08 – '12, ASFC | StatCan | TC | TPSGC]

Clients :

AMC, SGDN, ACSTA, plusieurs autres (~40 projets)

Spécialités :

Visualisation des données, nettoyage des données, application d'un large éventail de méthodes quantitatives.



Meko Deng '17

TECHNIQUES DE BASE D'ANALYSE DES DONNÉES

Patrick Boily

Data Action Lab | uOttawa | Idlewyld Analytics

pboily@uottawa.ca

PIPELINE D'ANALYSE DES DONNÉES

Modélisation des données et analyse conceptuelle

Collecte des données

Transformation des données

Stockage des données

Exploration des données

Analyse des données

Présentation des données

APERÇUS ET CALCULS CONCEPTS DE BASE

TECHNIQUES DE BASE D'ANALYSE DES DONNÉES



SCHÉMAS, GÉNÉRALITÉS, STRUCTURE

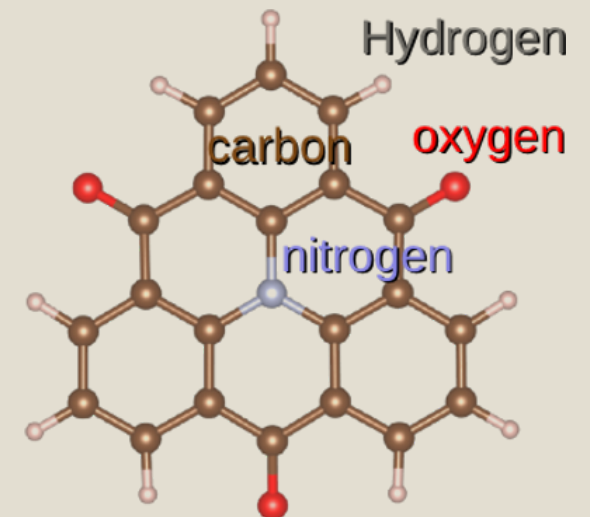
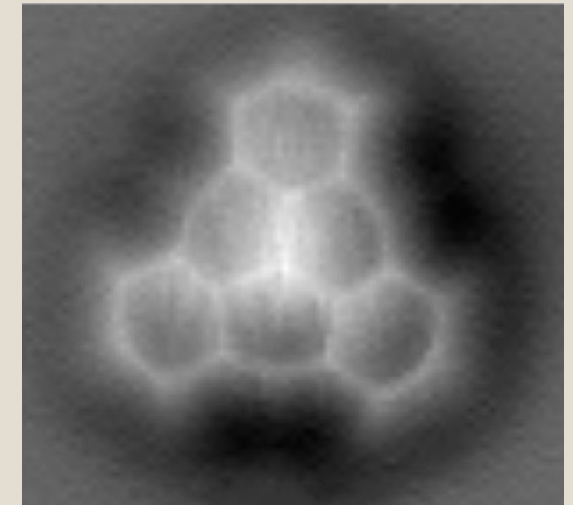
Schéma (« Pattern ») : une régularité prévisible et répétitive

Structure : organisation des éléments d'un système

Généralisation : création de concepts plus généraux ou abstraits à partir de concepts ou d'instances plus spécifiques.

Objectif sous-jacent de l'analyse : trouver des modèles ou des structures dans les données et **tirer des conclusions** à partir de ces modèles ou structures.

Trouver des modèles et des structures n'est pas inutile en soi, mais c'est la façon dont ces découvertes sont **utilisées** pour tirer des conclusions qui est importante.



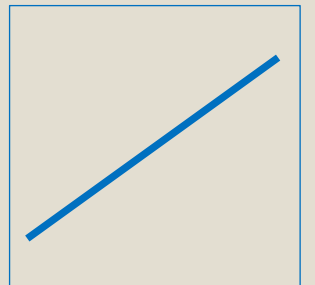
VARIABLES DÉPENDANTES ET INDÉPENDANTES

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 ?

Hauteur des
plantes



Heures d'ensoleillement

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)

CATÉGORIQUE → NUMÉRIQUE

Les données catégorielles peuvent être transformées en données numériques en générant des **comptes de fréquence** des différentes valeurs de la variable catégorielle.

Cela nous permet ensuite d'appliquer des techniques d'analyse numérique.

Couleur de la maison	Fréquence
rouge	40
bleu	13
vert	2

LE RÔLE PARTICULIER DES DONNÉES CATÉGORIELLES

Les données catégorielles jouent un rôle particulier :

- en science des données, les **variables catégorielles** prennent des **valeurs prédéfinies**
- en science expérimentale, un **facteur** est une variable indépendante dont les niveaux sont définis (on peut également la considérer comme catégorie de traitement)
- en analyse d'entreprise, il s'agit de **dimensions** (avec des membres) par rapport à des mesures.

Quelle que soit la façon dont ils sont nommés, elles peuvent être utilisées afin de créer des **sous-ensembles** ou **de résumer** les données.

DONNÉES HIÉRARCHIQUES/ENCHEVÊTRÉES/ MULTI-NIVEAUX

Lorsqu'une variable catégorielle possède plusieurs niveaux d'abstraction, de nouvelles variables catégorielles peuvent être créées à partir de ces niveaux.

La "nouvelle" variable catégorielle a des relations prédéfinies avec le niveau le plus détaillé.

Nous pouvons souvent examiner les variables temporelles et spatiales de manière plus ou moins détaillée.

Granularité des données : quel est le niveau de détail le plus élevé que nous pouvons observer dans les données ?

Année	Trimestre	Compte_Q
2012	1	34
2012	2	12
2012	3	52
2012	4	0
2013	1	21
2013	2	9
2013	3	112
2103	4	8

Année	Compte_A
2012	98
2013	150

APERÇUS ET CALCULS TECHNIQUES DE BASE

TECHNIQUES DE BASE D'ANALYSE DES DONNÉES



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

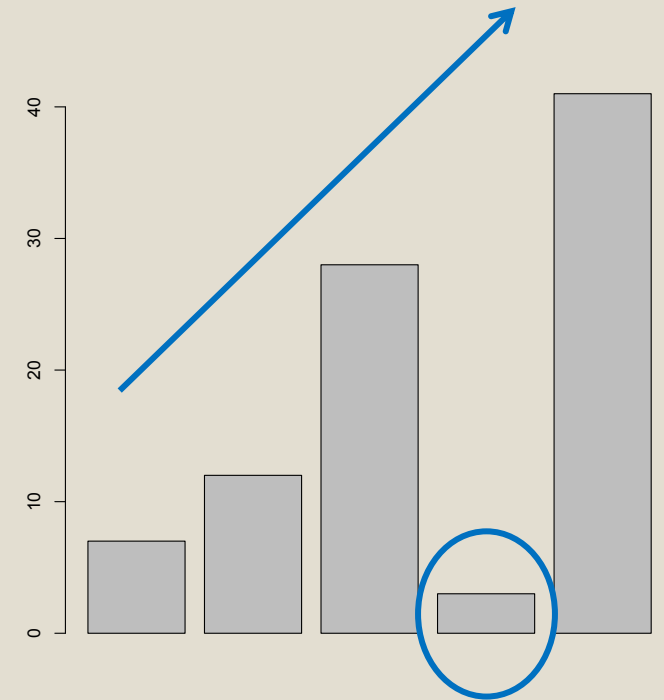
Analyse (au sens large) :

- 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.



DESCRIPTIONS DE DONNÉES (EN PROFONDEUR)

TECHNIQUES DE BASE D'ANALYSE DES DONNÉES



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**

EXEMPLES

- **L'âge** est une variable numérique, mesurée en années, bien qu'il soit souvent rapporté à l'année entière la plus proche, ou dans une tranche d'âge d'années (auquel cas il est ordinal).
- Les variables numériques typiques comprennent la distance en m , le volume en cm^3 , etc.
- **Le diagnostic de la maladie** est une variable catégorielle avec 2 catégories (positif/négatif).
- **La conformité à une norme** est une variable catégorielle : il peut y avoir 2 niveaux (conforme/non conforme) ou plus (conformité, problèmes mineurs de non-conformité, problèmes majeurs de non-conformité).
- Les **variables de comptage** sont des variables numériques.

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.

MÉDIANE

La **médiane** d'une variable quantitative comportant n observations est une valeur qui divise les données ordonnées en 2 sous-ensembles égaux : la moitié des observations sont inférieures (ou égales) à la médiane, et l'autre moitié supérieures (ou égales) à celle-ci.

Si n est **impair**, la médiane est la $\frac{n+1}{2}$ —ième observation ordonnée.

Si n est **pair**, la médiane est toute valeur comprise entre les $\frac{n}{2}$ et $\frac{n}{2} + 1$ observations ordonnées (on prend généralement leur moyenne).

La **procédure** est simple : ordonner les données et suivre à la lettre les règles du jeu (paires/impaires).

MÉDIANE

1. Imaginez une variable quantitative avec $n = 5$ observations, prenant les valeurs: 4,6,1,3,7.

Commencez par ordonner les valeurs : 1,3,4,6,7.

$n = 5$ est impair; on cherche la $\frac{n+1}{2} = \frac{5+1}{2} = 3^{\text{e}}$ observation, qui est **4**.

Notez qu'il y a 2 observations en dessous de 4 (1,3) et 2 observations au-dessus de 4 (6,7).

2. Imaginez une variable quantitative avec $n = 6$ observations, prenant les valeurs : 4,6,1,3,7,23.

Commencez par ordonner les valeurs : 1,3,4,6,7,23.

$n = 6$ est pair, on cherche une valeur entre $\frac{n}{2} = \frac{6}{2} = 3^{\text{e}}$ et $\frac{n}{2} + 1 = \frac{6}{2} + 1 = 4^{\text{e}}$ obs., disons **5.2**.

Notez qu'il y a 3 observations en dessous de 5.2 (1,3,4) & 3 observations au-dessus de 5.2 (6,7,23).

MOYENNE

La moyenne d'un échantillon est simplement la **moyenne arithmétique** de ses observations. :

$$\text{moyenne} = \frac{x_1 + \dots + x_n}{n}$$

D'autres moyennes existent, citons la moyenne **harmonique** et la moyenne **géométrique**.

Exemples:

- $\text{moyenne}(4,6,1,3,7) = \frac{4+6+1+3+7}{5} = \frac{21}{5} = 4.2 \approx 4 = \text{moyenne}(4,6,1,3,7)$
- $\text{moyenne}(4,6,1,3,7,23) = \frac{4+6+1+3+7+23}{6} = \frac{44}{6} = 7.3 \approx 7.3 = \text{moyenne}(4,6,1,3,7,23)$

MOYENNE OU MÉDIANE ?

Quelle mesure de centralité doit-on utiliser pour présenter les données ?

La moyenne est soutenue par la **théorie** CLT (qui ne sera pas abordée ici).

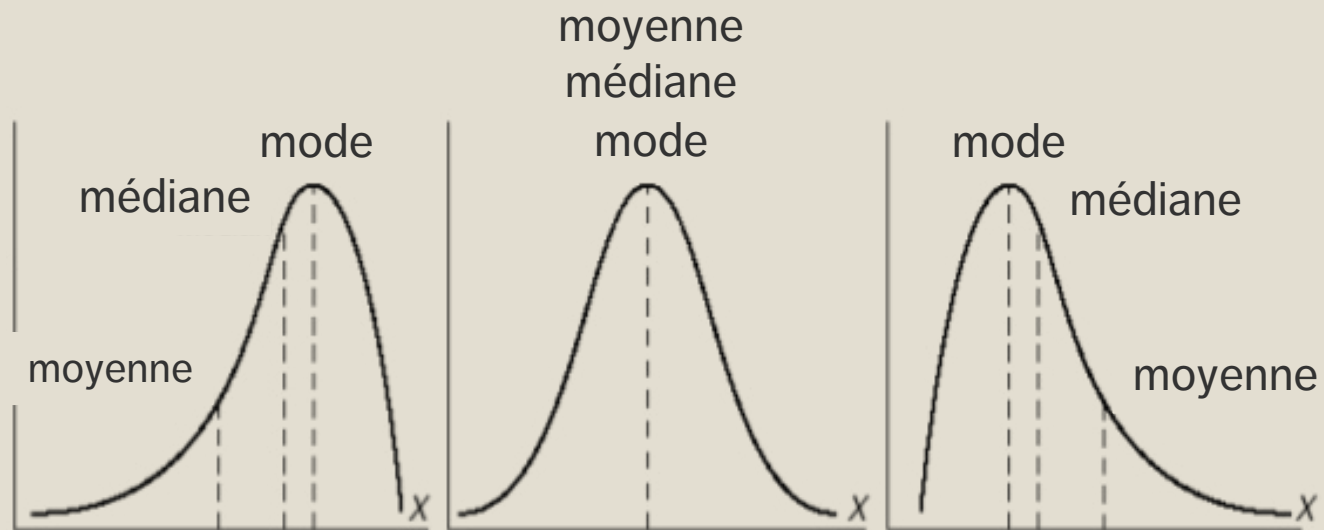
Si la distribution des données est **symétrique**, les deux valeurs seront proches l'une de l'autre.

Si la distribution des données est **asymétrique**, la moyenne est tirée vers la **longue queue** et, par conséquent, donne une image déformée du centre réel.

Par conséquent, les médianes sont utilisées pour les prix des logements, les revenus, etc.

La médiane est **robuste** contre les valeurs aberrantes et les interprétations incorrectes, alors que la moyenne ne l'est pas.

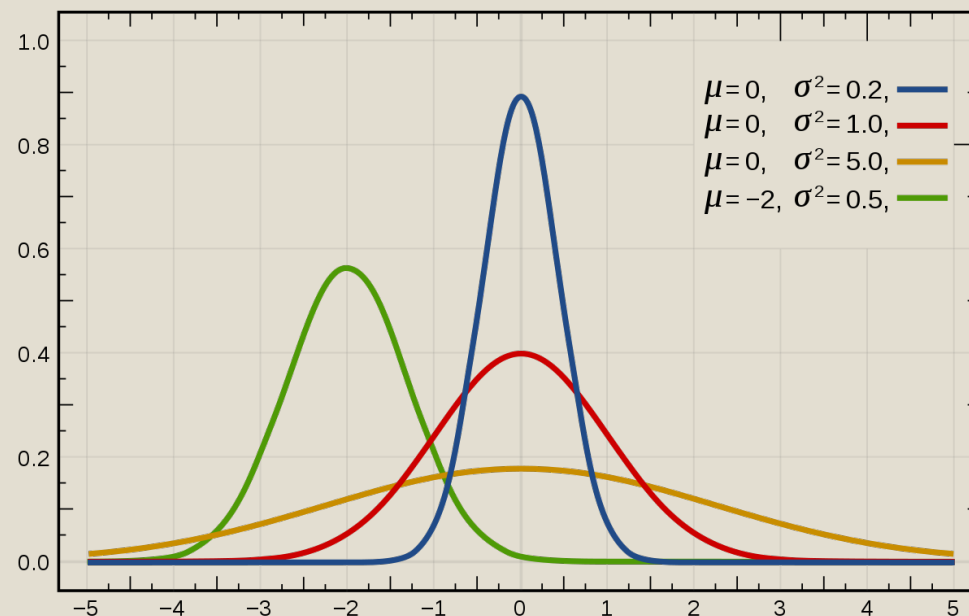
MOYENNE OU MÉDIANE ?



ÉCART-TYPE

Les mesures de centralité donnent une idée de l'endroit où les valeurs de la variable sont "« **regroupées** ».

L'écart-type fournit une notion de sa **dispersion** ; un écart-type plus élevé signifie une dispersion plus élevée.



ÉCART TYPE

L'écart-type est construit à partir d'une sorte de **moyenne** des observations de la variable :

$$\text{écart-type} = \sqrt{\frac{(x_1 - \text{moyenne})^2 + \dots + (x_n - \text{moyenne})^2}{n}}.$$

Exemples:

- $\text{écart-type}(4,6,1,3,7) = \sqrt{\frac{(4-4.2)^2 + (6-4.2)^2 + (1-4.2)^2 + (3-4.2)^2 + (7-4.2)^2}{5}} \approx \mathbf{2.14}$
- $\text{écart-type}(4,6,1,3,23) = \sqrt{\frac{(4-7.3)^2 + (6-7.3)^2 + (1-7.3)^2 + (3-7.3)^2 + (7-7.3)^2 + (23-7.3)^2}{6}} \approx \mathbf{3.98}$

QUANTILES

Les **centiles**, les **déciles** ou les **quartiles** constituent un autre moyen de fournir des informations sur la dispersion des données.

Le **quartile inférieur** Q_1 d'une colonne comportant n entrées est une valeur numérique qui divise les données ordonnées en 2 sous-ensembles inégaux : 25% des observations sont **en-dessous** (ou au niveau) à Q_1 et 75% des observations sont **en-dessus** (ou au niveau) à Q_1 .

De même, le **quartile supérieur** Q_3 divise les données ordonnées en 75 % des observations **en-dessous** (ou au niveau) de Q_3 et 25 % des observations **au-dessus** (ou au niveau) de Q_3 .

La médiane peut être interprétée comme le **quartile central** Q_2 des données, le minimum comme Q_0 , et le maximum comme Q_4 ; $(Q_0, Q_1, Q_2, Q_3, Q_4)$ représentent le **résumé en 5 points** des données.

AUTRES MESURES

Centralité :

- le **milieu** d'une variable est $\frac{\min + \max}{2} = \frac{Q_0 + Q_4}{2}$.
- la **tri-moyenne** d'une variable est $\frac{Q_1 + 2Q_2 + Q_3}{4}$.

Dispersion:

- la **plage** d'une variable est $\max - \min = Q_4 - Q_0$.
- l'**écart interquartile** d'une variable est $IQR = Q_3 - Q_1$.

En général, nous pouvons mieux comprendre une variable grâce à des mesures **multiples**.

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.

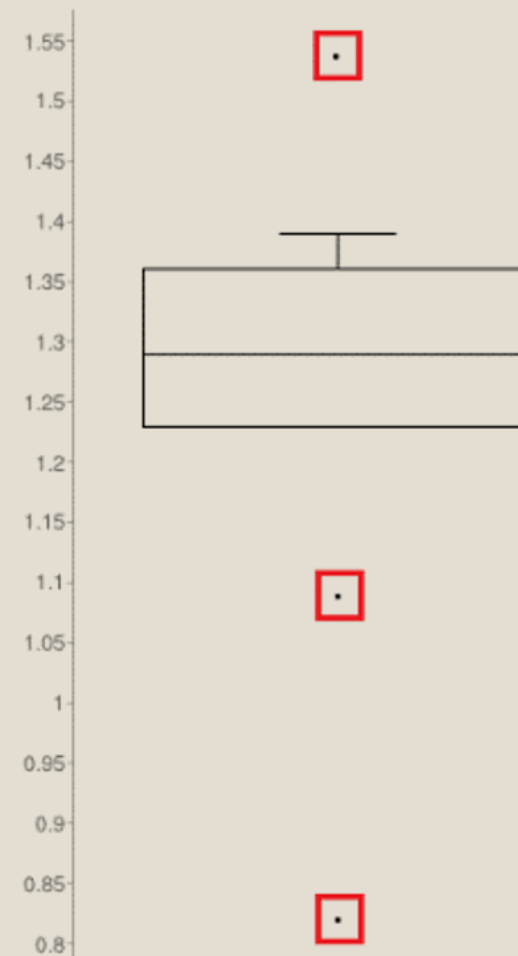
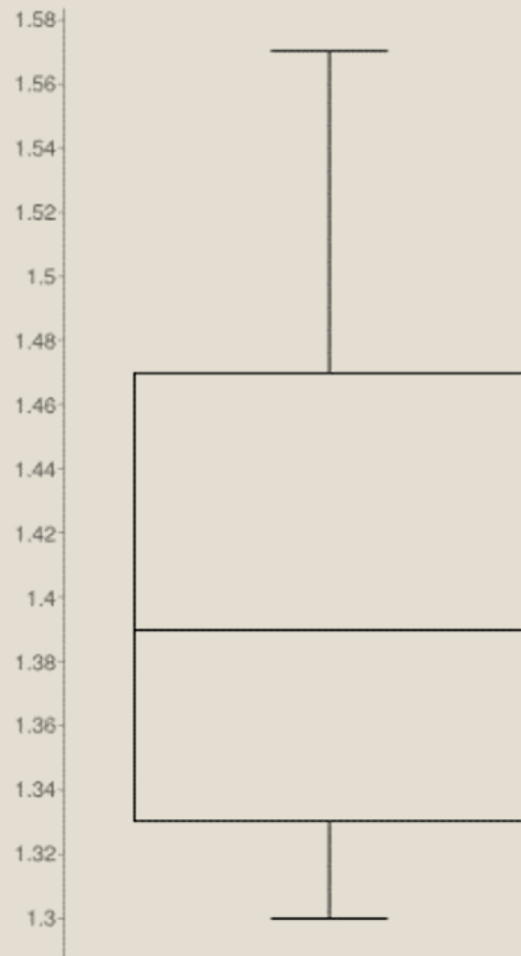
Tracez une ligne s'étendant de Q_1 à la plus petite observation inférieure à $1.5 \times \text{IQR}$ à gauche de Q_1 .

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

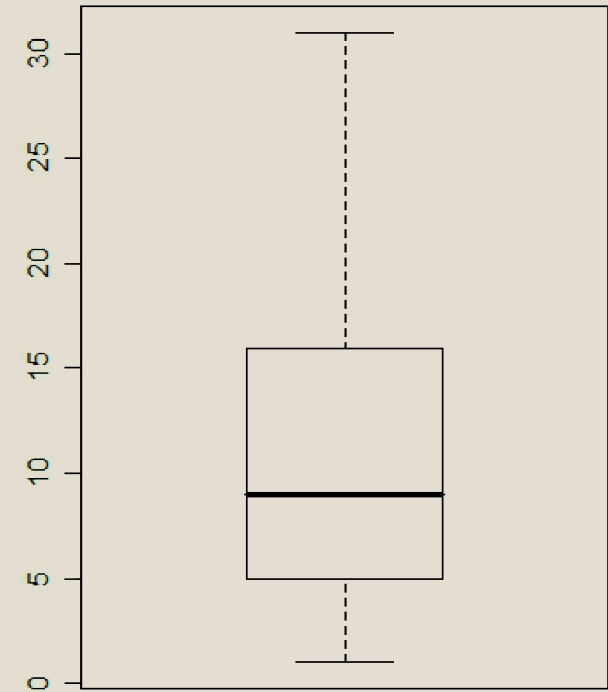
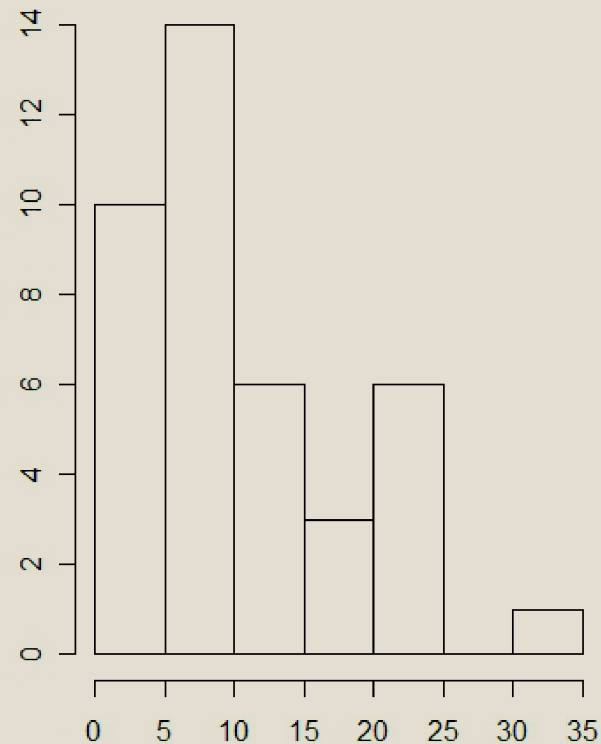
6, 3, 2, 24, 12, 3, 7, 14, 21, 9, 14, 22, 15,
2, 17, 10, 7, 7, 31, 7, 18, 6, 8, 2, 3, 2, 17,
7, 7, 21, 13, 23, 1, 11, 3, 9, 4, 9, 9, 25

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 ?



ASYMÉTRIE

Si la distribution des données est **symétrique**, alors médiane = moyenne, et Q_1 et Q_3 sont équidistants de la médiane : $Q_3 - Q_2 \approx Q_2 - Q_1$.

Autrement :

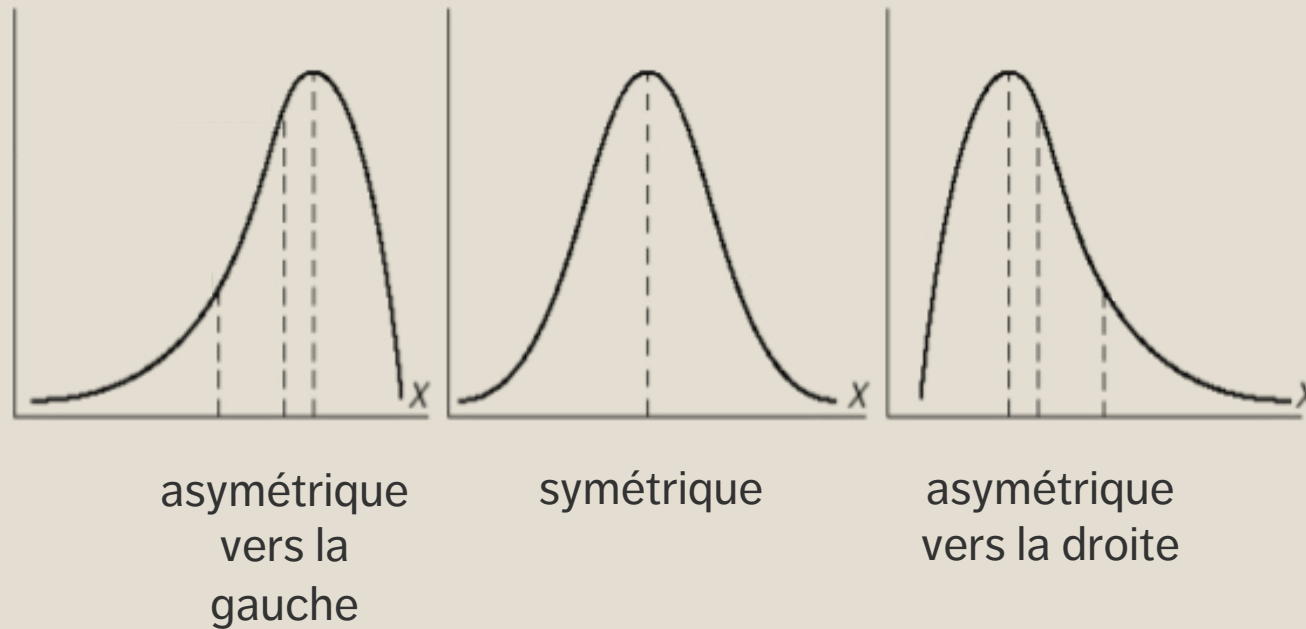
- si $Q_3 - Q_2 > Q_2 - Q_1$, la distribution des données est **asymétrique vers la droite**.
- si $Q_3 - Q_2 < Q_2 - Q_1$, la distribution des données est **asymétrique vers la gauche**

Dans l'exemple précédent ,

$$Q_3 - Q_2 = 15.5 - 9 = 6.5 > 3.5 = 9 - 5.5 = Q_2 - Q_1,$$

donc la distribution est asymétrique vers la droite.

ASYMÉTRIE



La forme d'un ensemble de données peut être utilisée pour suggérer un modèle analytique pour la situation d'intérêt.

CORRÉLATION

TECHNIQUES DE BASE D'ANALYSE DES DONNÉES



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

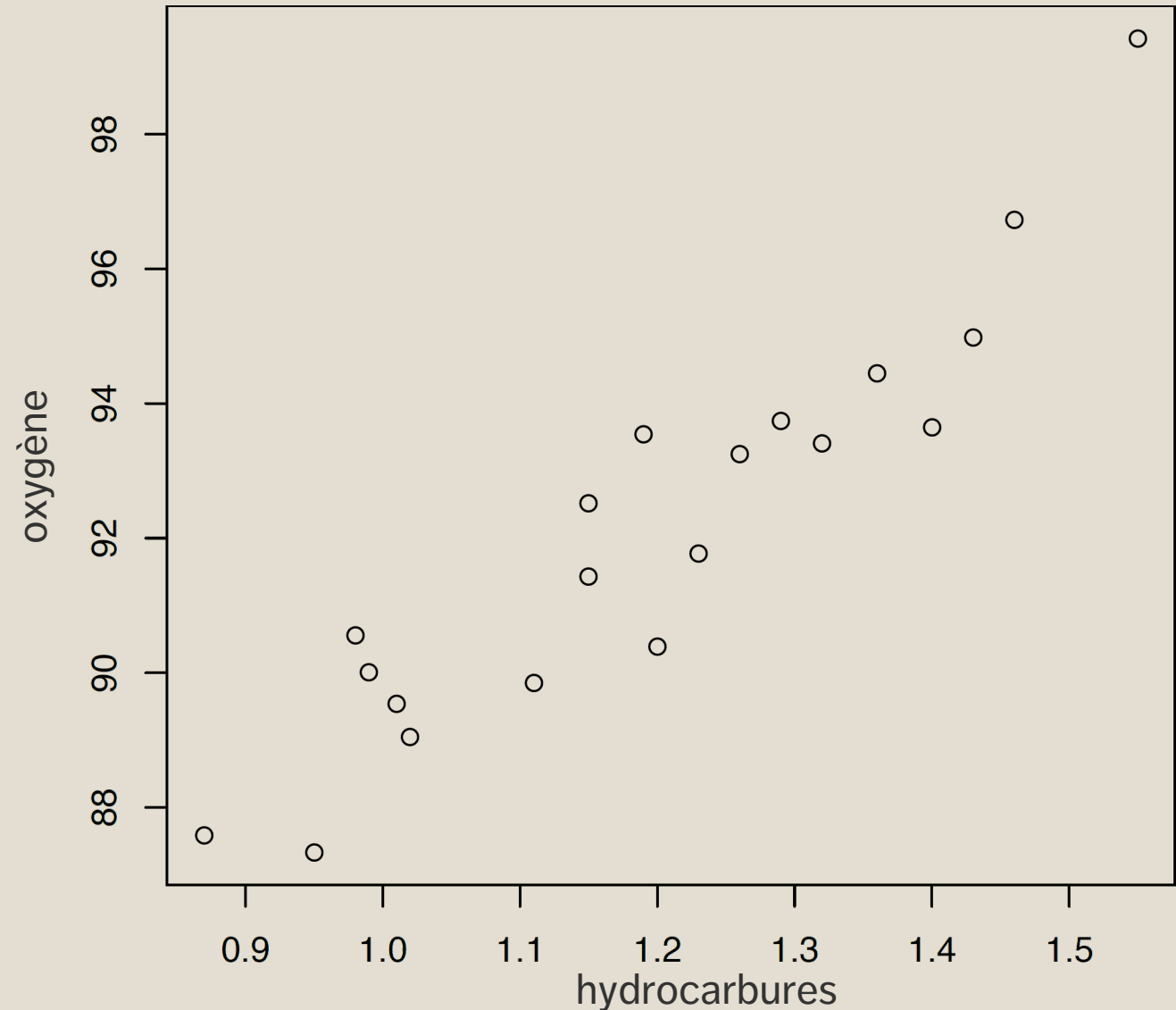
Objectifs :

- 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 !



COEFFICIENT DE CORRÉLATION

Pour les données appariées (x_i, y_i) , $i = 1, \dots, n$, le **coefficient de corrélation** de x et y est de

$$\rho_{XY} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}} = \frac{S_{xy}}{\sqrt{S_{xx} \cdot S_{yy}}}$$

Cette corrélation est définie uniquement si $S_{xx}, S_{yy} \neq 0$, et si ni les x_i ni les y_i ne sont constants.

Les variables x et y sont **non corrélées** si $\rho_{XY} = 0$ (ou est très petit, en pratique), et elles sont **corrélées** si $\rho_{XY} \neq 0$ (ou $|\rho_{XY}|$ est « grand », en pratique).

Pour les données sur les hydrocarbures, $S_{xy} \approx 10.18$, $S_{xx} \approx 0.68$, $S_{yy} \approx 173.38$ et

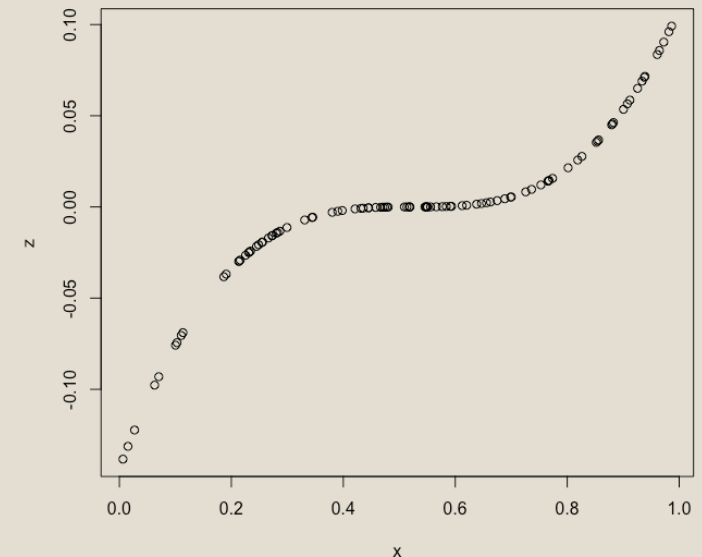
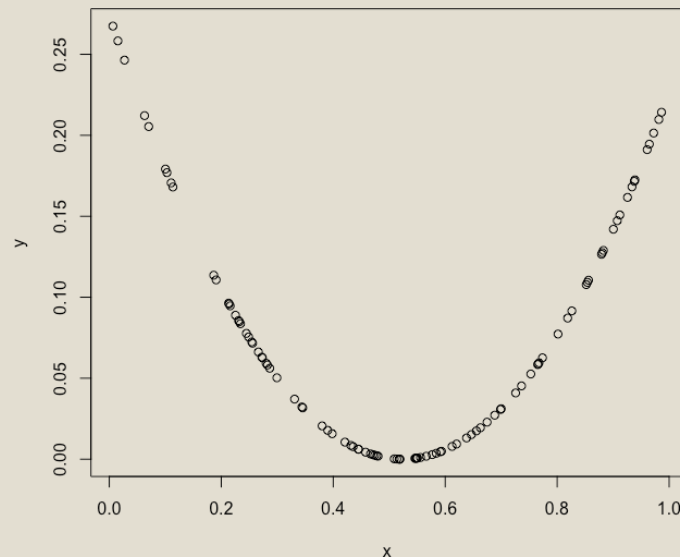
$$\rho_{XY} = \frac{10.18}{\sqrt{0.68 \cdot 173.38}} \approx \mathbf{0.94} \text{ (corrélation élevée).}$$

PROPRIÉTÉS ET INTERPRÉTATION

Le signe de ρ_{XY} reflète la tendance des points.

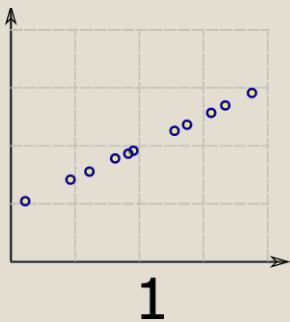
IMPORTANT : une valeur élevée du coefficient de corrélation $|\rho_{XY}|$ n'implique pas nécessairement une **relation de cause à effet** entre les deux variables ;

Notons que x et y peuvent avoir une relation **non linéaire** très forte sans que ρ_{XY} ne la reflète (-0.12 à gauche, 0.93 à droite).

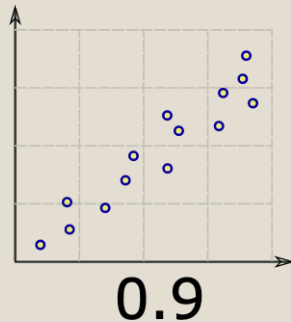


PROPRIÉTÉS ET INTERPRÉTATION

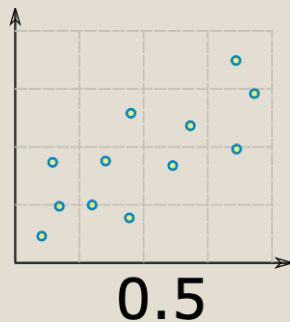
corrélation
positive
parfaite



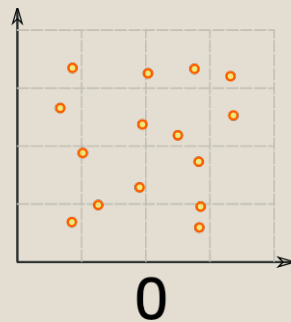
corrélation
positive
élevée



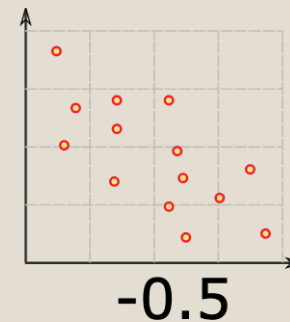
faible
corrélation
positive



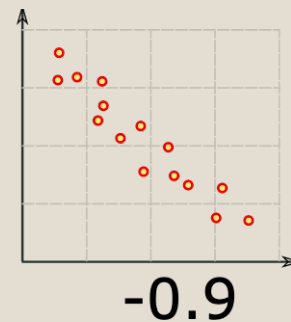
aucune
corrélation



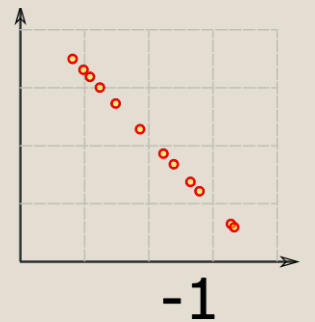
faible
corrélation
négative

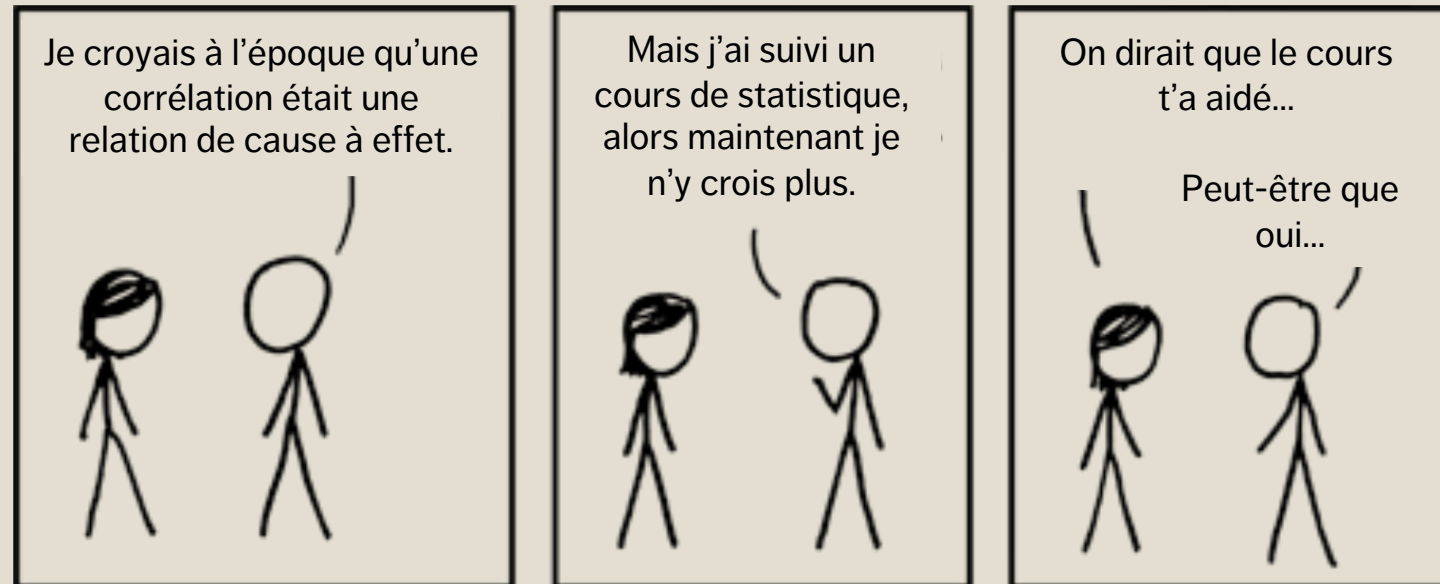


corrélation
négative
élevée



corrélation
négative
parfaite





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 ».

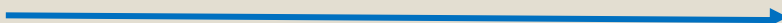
ANALYSE DE RÉGRESSION

TECHNIQUES DE BASE D'ANALYSE DES DONNÉES



MODÉLISATION PAR RÉGRESSION

Structure de données d'une tâche de modélisation générale est représenté par



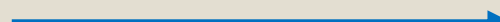
X_1	X_2	$\dots$	X_p	Y
x_{11}	x_{12}	$\dots$	x_{1p}	y_1
x_{21}	x_{22}	$\dots$	x_{2p}	y_2
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
x_{n1}	x_{n2}	$\dots$	x_{np}	y_n

Nous tenons compte de p **variables indépendantes** X_i (les prédicteurs) afin d'essayer de prédire la **variable dépendante** Y (la réponse).

On simplifie la discussion en utilisant la notation matricielle

$$\mathbf{X}_{[n \times p]}, \mathbf{Y}_{[n \times 1]}, \boldsymbol{\beta}_{[p \times 1]},$$

où n est le nombre d'observations et p est le nombre de variables indépendantes.

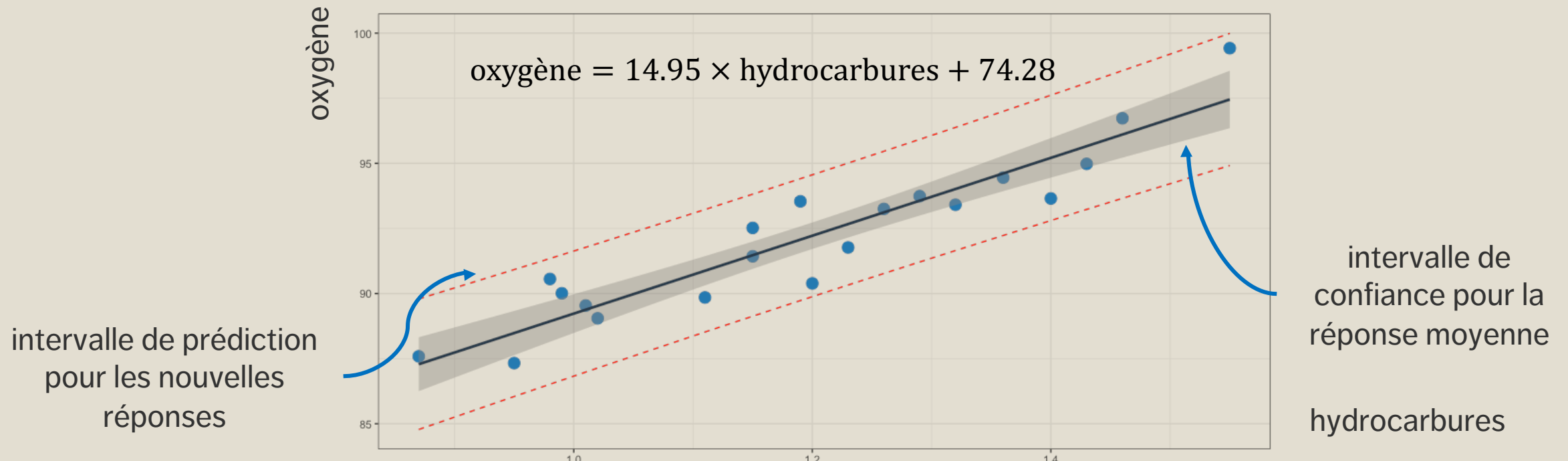


X_1	X_2	$\dots$	X_p	Y
$\mathbf{X}_{[n \times p]}$				$\mathbf{Y}_{[n \times 1]}$

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

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



SÉRIES CHRONOLOGIQUES ET CARTES DE CONTRÔLE

TECHNIQUES DE BASE D'ANALYSE DES DONNÉES



« Les ingénieurs de la NASA n'ont pas identifié le lien entre les températures basses inattendues de la rampe de lancement et les défaillances des joints toriques des fusées d'appoint de la navette spatiale.

Ils ont interprété ce signal critique comme une simple variation fortuite dans la défaillance des joints.

L'absence de ce constat a été déterminante dans la décision de lancer la navette Challenger pour son dernier et désastreux vol. »

Vaughan, D. [1997], *La décision du lancement de Challenger : Technologie risquée, culture et déviance à la NASA*, p.383

SUIVI DE PROCESSUS STATISTIQUE

Les processus sont souvent sujets à la **variabilité** :

- la variabilité due à **l'effet cumulatif** de nombreuses petites causes essentiellement inévitables (un processus qui ne fonctionne qu'avec de telles causes communes est dit en **maîtrise** [statistique]) ;
- la variabilité due à des **causes particulières**, telles que des machines mal réglées, des opérateurs mal formés, des matériaux défectueux, etc. (la variabilité est généralement beaucoup plus importante pour les causes particulières, et on dit que ces processus sont **hors contrôle** [statistique]).

L'objectif du **suivi des processus statistiques** est d'identifier l'occurrence des causes spéciales.

SÉRIES CHRONOLOGIQUES

Considérons quelques observations $\{x_1, \dots, x_n\}$, issues d'un certain processus.

En pratique, l'indice i est souvent un **indice temporel** ou un **indice de localisation**, c'est-à-dire que les x_i sont observés en **séquence** ou dans des **régions**.

Dans le premier cas, les observations forment une **série chronologique**.

Les processus qui génèrent les observations peuvent changer dans le temps et l'espace en raison de :

- **de facteurs externes** (guerre, pandémie, élection, etc.), ou
- **de facteurs internes** (changement de politique, modification du processus de fabrication, etc.)

SÉRIES CHRONOLOGIQUES

La moyenne et l'écart-type pourraient ne pas fournir un résumé utile de la situation.

Pour avoir une idée de ce qui se passe, il pourrait être préférable de **représenter les données** dans **l'ordre où elles ont été recueillies** (ou selon les régions géographiques).

La coordonnée horizontale représente :

- le moment de la collecte t (ordre, jour, semaine, trimestre, année, etc.), ou bien
- le lieu i (pays, province, ville, branche, etc.).

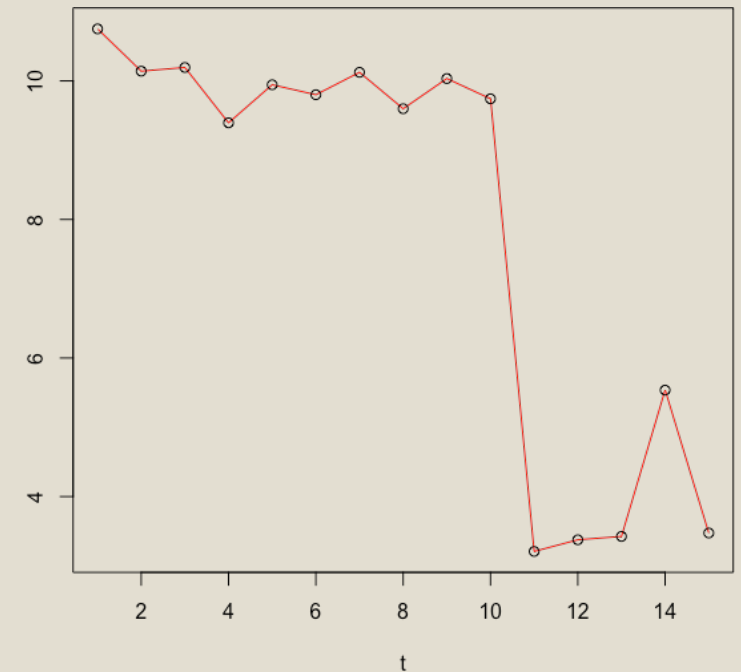
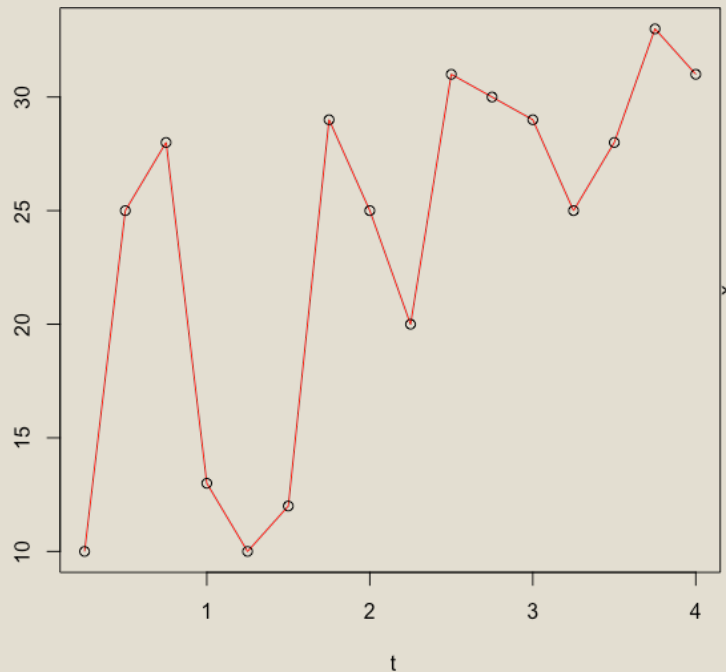
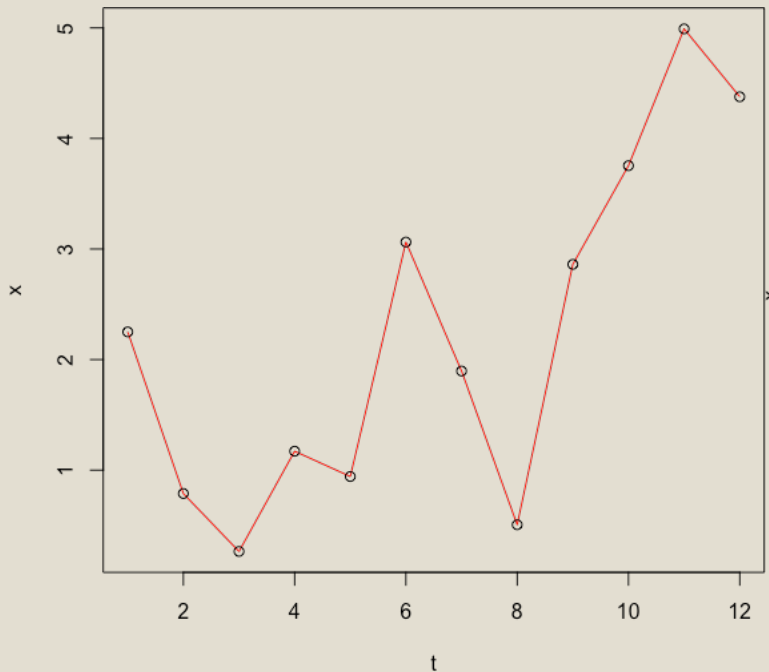
La coordonnée verticale représente les observations d'intérêt x_t ou x_i .

On recherche alors **des tendances, des cycles, des décalages**, etc.

EXEMPLES

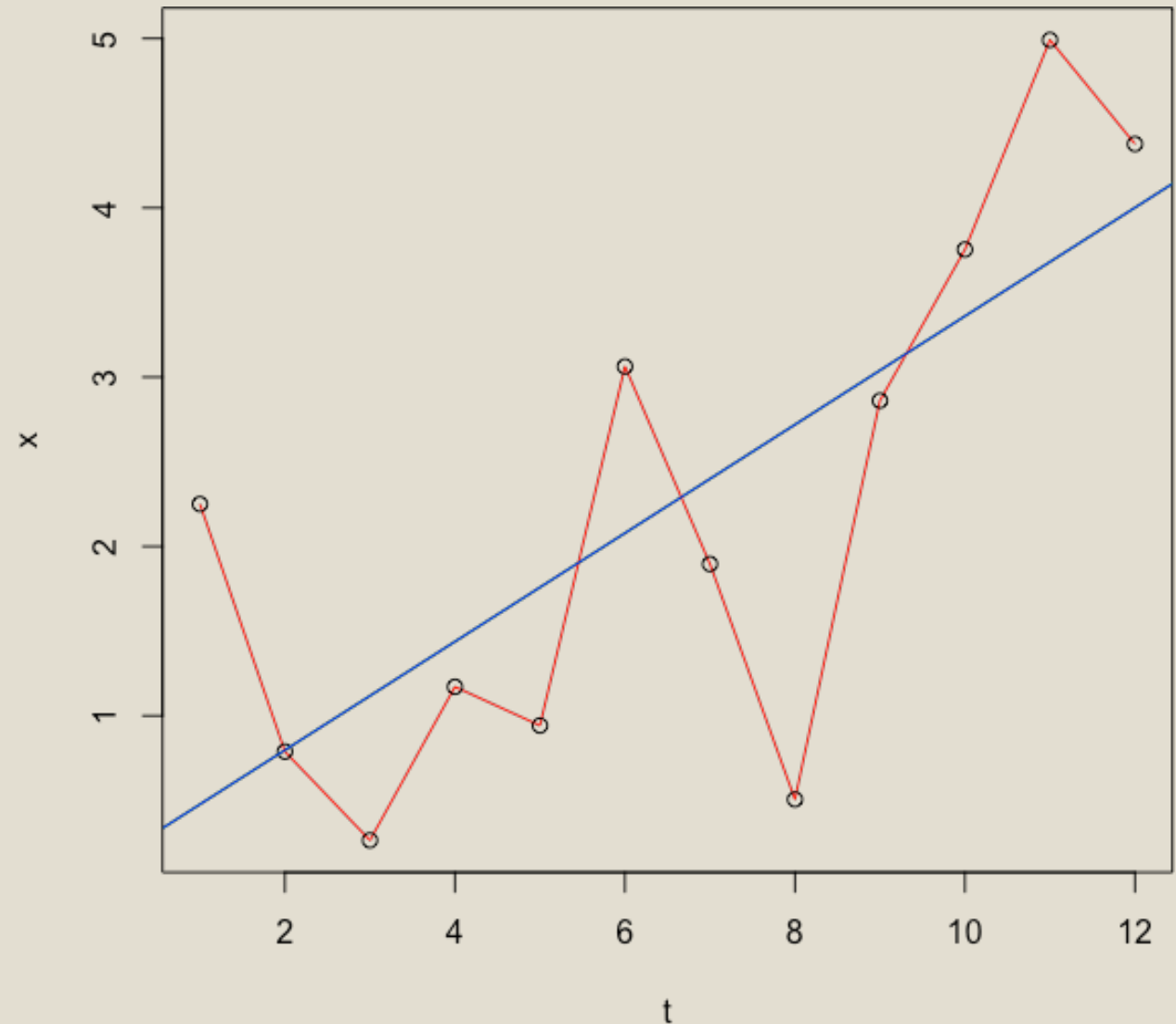


La série temporelle suivante enregistre les ventes x (en 10 000\$) pour 3 produits différents, en fonction du passage du temps t en années (à gauche), trimestres (au milieu), semaines (à droite). Doit-on intervenir ?



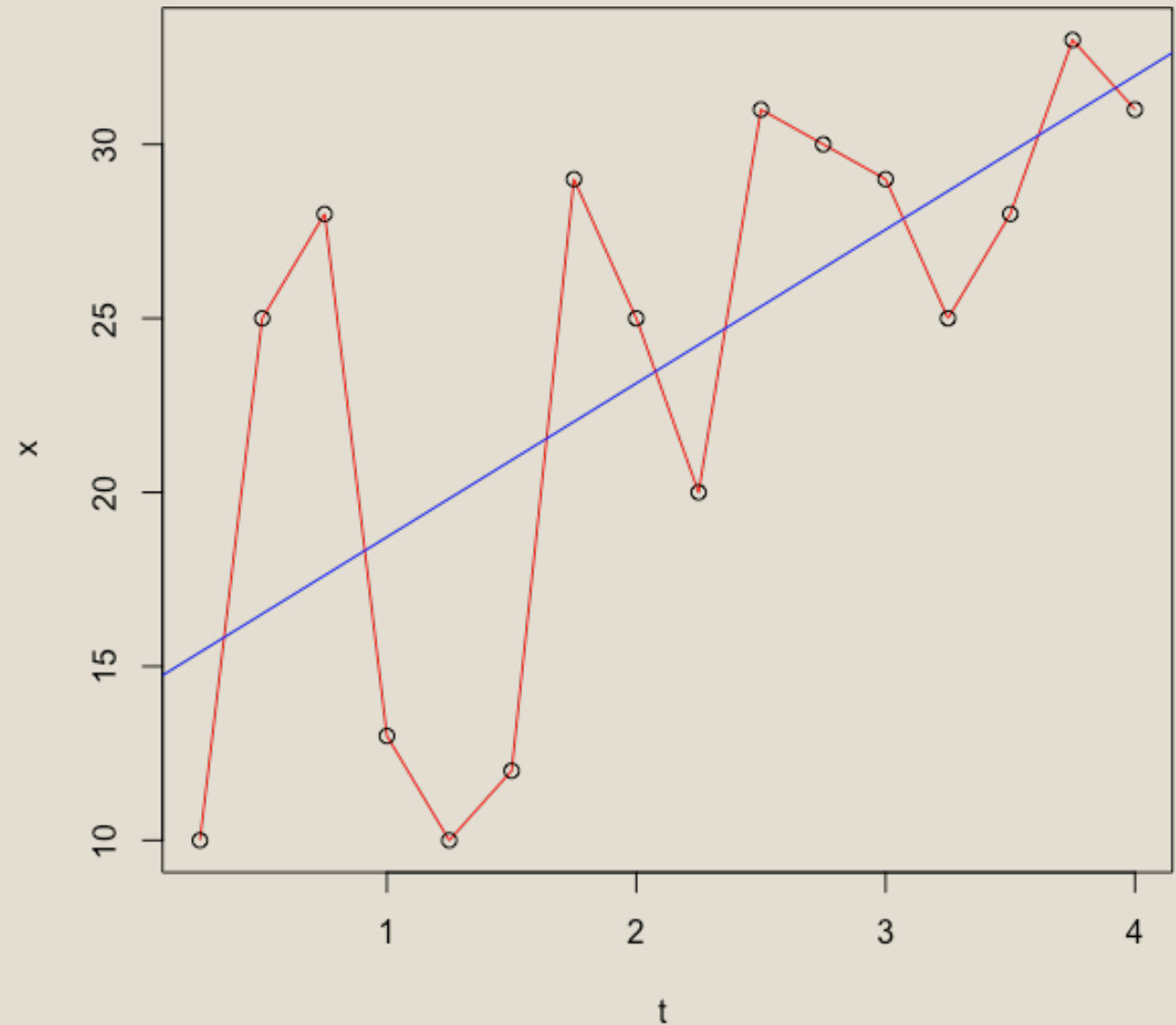
Il y a des baisses occasionnelles des ventes d'une année sur l'autre, mais une tendance claire à la hausse.

Si seuls les deux derniers points sont présentés aux actionnaires, ils pourraient penser qu'il y a des problèmes et que des changements doivent être apportés.



Il y a un effet cyclique avec des augmentations de Q1 à Q2, et de Q2 à Q3, mais des diminutions de Q3 à Q4, et de Q4 à Q1.

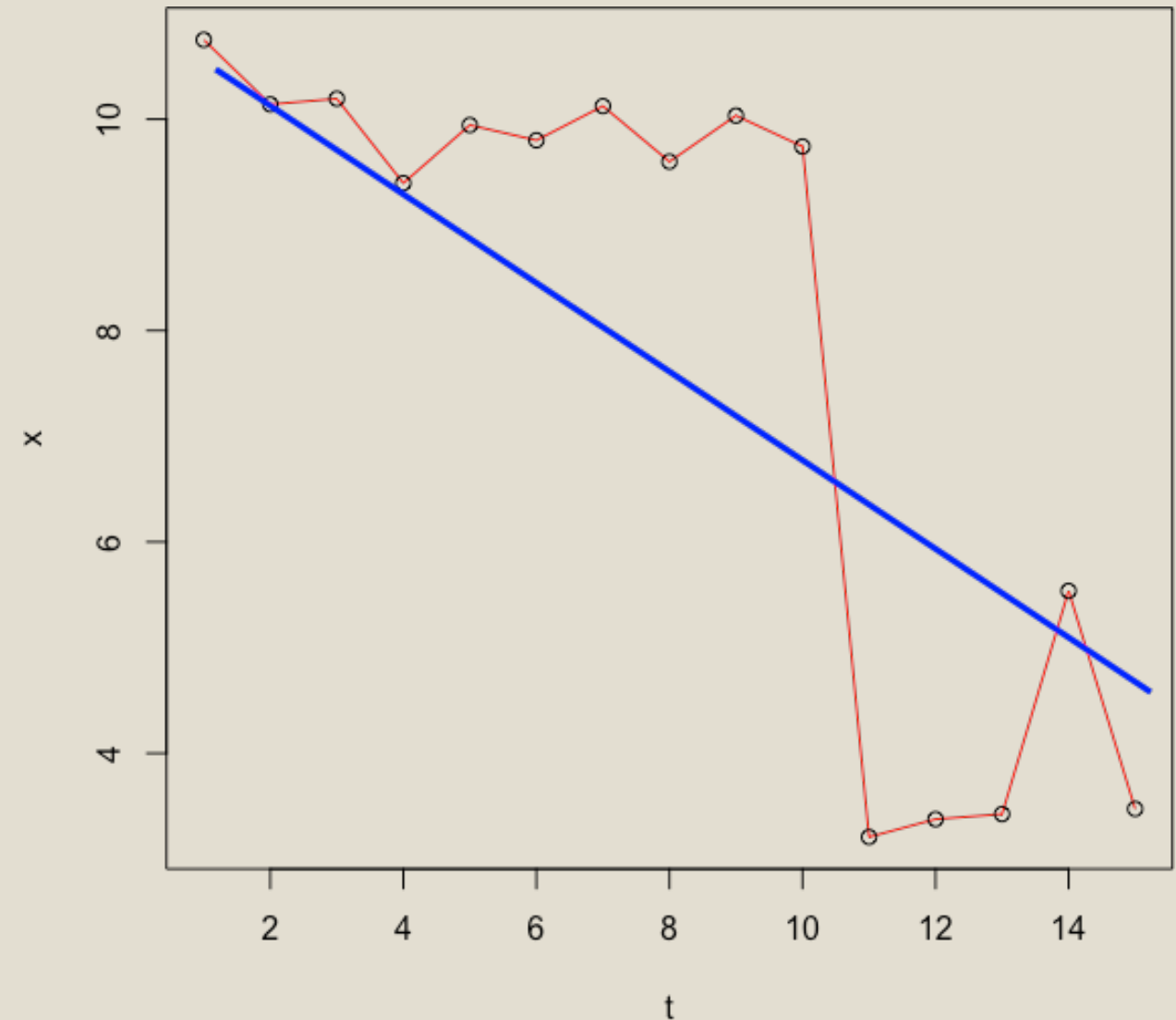
Globalement, il semble y avoir une tendance à la hausse, comme l'indique la ligne de meilleur ajustement.



Il est clair que quelque chose s'est produit après la dixième semaine.

Que les causes particulières soient internes ou externes dépend du contexte (que nous n'avons pas à notre disposition).

Une action semble être nécessaire.



Il est clair que quelque chose s'est produit après la dixième semaine.

Que les causes particulières soient internes ou externes dépend du contexte (dont nous ne disposons pas).

Une action semble être nécessaire.

Les **cartes de contrôle** peuvent aider à identifier les points de rupture ou les situations particulières.

