

COLLECTE ET TRAITEMENT DES DONNÉES

« Les gens résistent à un recensement, mais présentez-leur une page de profil et ils passeront la journée à vous raconter qui ils sont. »

Max Berry, Lexicon

APERÇU

1. Caractéristiques des données à recueillir : Théorie de l'échantillonnage et plan d'étude
2. Collecte de données moderne : Interfaces de programmation d'applications (API) et moissonnage du Web
3. Utilisation des données : Préparation préalable des données
4. Préparation à l'analyse : Nettoyage des données
5. Simplification de la gestion des données : Transformation des données
6. Préservation de la fiabilité des données : Qualité et validation des données

THÉORIE DE L'ÉCHANTILLONNAGE ET PLAN D'ÉTUDE

COLLECTE ET TRAITEMENT DES DONNÉES

« La dernière enquête indique que trois personnes sur quatre représentent 75 % de la population »

D. Letterman

L'OBJECTIF D'UN PLAN D'ÉTUDE ET D'ÉCHANTILLONNAGE EFFICACE

Nous recherchons des données de nature à :

- donner un aperçu légitime de notre système d'intérêt
- fournir des réponses correctes et précises aux questions pertinentes
- soutenir la formulation de conclusions légitimes et valides, en permettant de nuancer ces conclusions en matière de portée et de précision

Un tel processus commence par le **plan d'étude** – quelles données recueillir et comment les recueillir.

« À l'aide d'un appareil d'imagerie par résonance magnétique (IRM), un diplômé de Dartmouth a étudié l'activité cérébrale d'un saumon lorsqu'on lui montrait des photographies et qu'on lui posait des questions. L'aspect le plus intéressant de cette recherche, ce n'est pas qu'on ait étudié un saumon, mais que ce saumon était mort. Hé oui! On a acheté un saumon mort au marché local, on l'a placé dans un appareil d'IRM, et on a observé certains schémas. Il y avait inévitablement des schémas, mais ils étaient invariablement dénués de sens. »

ÉCHANTILLONNAGE NON PROBABILISTE ET « PÊCHE » AUX TENDANCES

Deux situations distinctes peuvent s'associer pour causer dans **problèmes** d'analyse des données :

- la formulation de conclusions (inférences) à partir d'un échantillon de population qui ne se justifie pas par la méthode de collecte de l'échantillon (symptomatique d'un échantillonnage non probabiliste)
- la recherche d'un quelconque schéma dans les données, puis la formulation d'explications a posteriori concernant ces schémas

Seules ou combinées, ces deux situations conduisent à des conclusions médiocres (et **potentiellement nuisibles**).

ÉTUDES ET ENQUÊTES

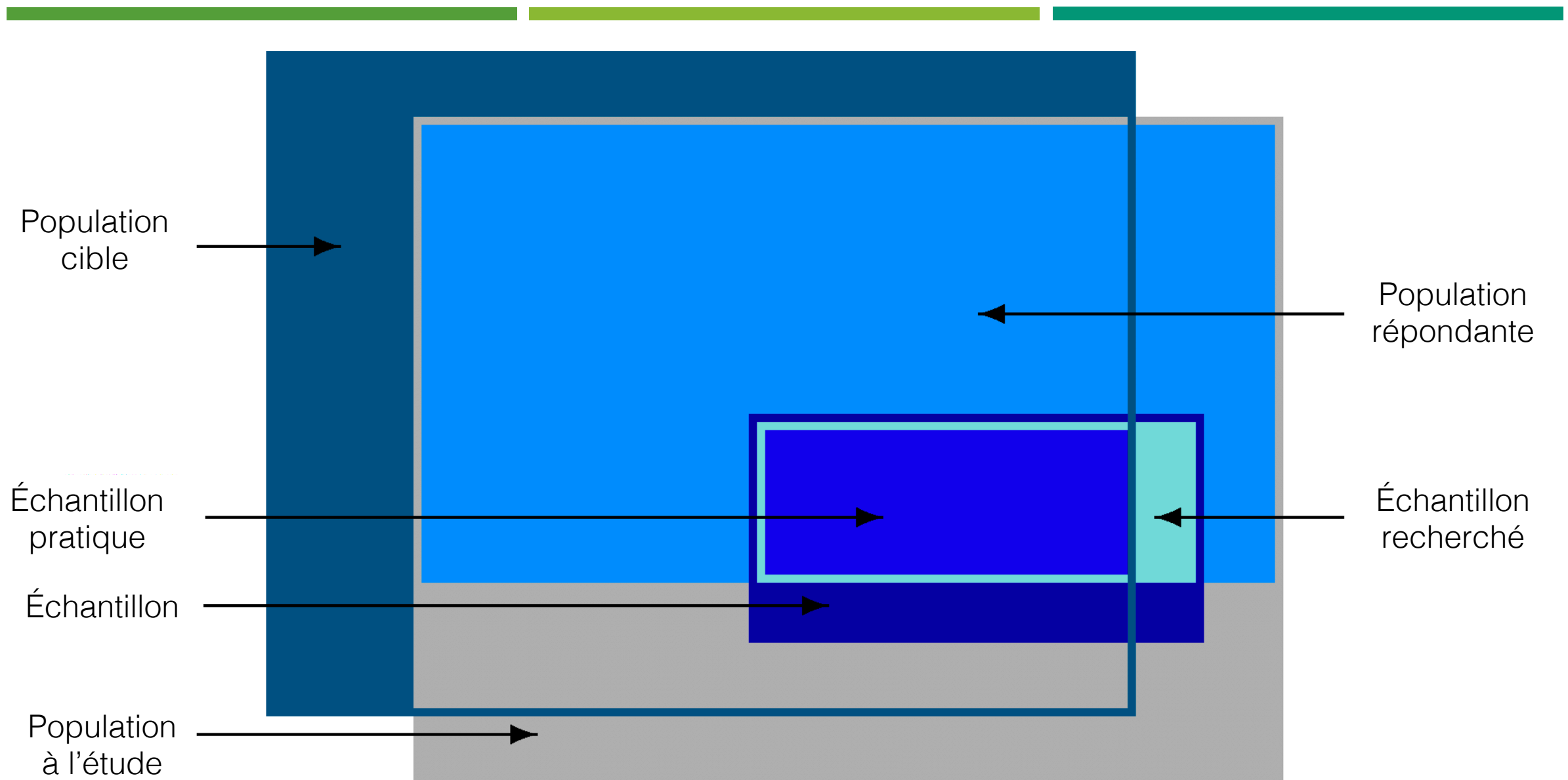
Une **enquête** est une activité qui consiste à recueillir de l'information sur des caractéristiques d'intérêt :

- de manière **organisée** et **méthodique**;
- sur une partie ou la totalité des **unités** d'une population;
- à l'aide de concepts, de méthodes et de procédures **bien définis**;
- grâce à la compilation de renseignements sous forme d'un résumé **significatif**.

MODÈLES D'ÉCHANTILLONNAGE

Un **recensement** est une collecte de données sur toutes les unités d'une population, alors qu'une **enquête sur échantillon** n'utilise qu'une fraction des unités.

Lorsque l'échantillonnage de l'enquête est effectué correctement, il est possible de recourir à diverses **méthodes statistiques** pour faire des **inférences** sur la **population cible** en échantillonnant un (comparativement) petit nombre d'unités dans la **population étudiée**.



BASES D'ENQUÊTE

La base idéale contient les données d'identification, les données de contact, les données de classification, les données de maintenance et les données de couplage, et doit réduire au minimum le risque de **sous-dénombrement** ou de **surdénombrement**, ainsi que le nombre de dédoublements et d'erreurs de classification (même si certains problèmes éventuels peuvent être réglés à l'étape du traitement des données).

Une approche d'échantillonnage statistique est contre-indiquée à moins que la base d'enquête choisie ne soit :

- **pertinente** (autrement dit qu'elle corresponde et permette l'accessibilité à la population cible);
- **exacte** (l'information qu'elle contient est valide);
- **opportune** (elle est à jour);
- **offerte à un prix compétitif.**

ERREUR D'ENQUÊTE

Erreur totale = erreur d'échantillonnage + erreur de mesure + erreur de non-réponse + erreur de couverture

enquête, pas
recensement

manque d'exactitude
dans la mesure des
observations

non-répondants présentant
des différences
d'observation
systématiques

dégradation ou
corruption de la
base

L'échantillonnage statistique permet de fournir des estimations, mais, surtout, il permet aussi de contrôler dans une certaine mesure l'**erreur totale** (ET) dans les estimations.

Idéalement, $ET = 0$. Dans la pratique, deux principaux éléments contribuent à l'ET : les **erreurs d'échantillonnage** (attribuable au choix du plan d'échantillonnage) et les **erreurs non attribuables à l'échantillonnage** (tout le reste).

ERREUR NON ATTRIBUABLE À L'ÉCHANTILLONNAGE

Dans une certaine mesure, il est possible de contrôler une erreur non attribuable à l'échantillonnage :

- l'**erreur de couverture** peut être réduite au minimum en choisissant des bases d'enquête à jour et de grande qualité;
- l'**erreur de non-réponse** peut être atténuée en choisissant soigneusement le mode de collecte des données et le plan du questionnaire, et au moyen de « rappels » et de « suivis »;
- l'**erreur de mesure** peut être grandement diminuée par une conception minutieuse du questionnaire, un essai préliminaire de l'appareil de mesure et une validation croisée des réponses.

Dans les faits, ces suggestions ne s'avèrent pas d'une grande utilité à l'heure actuelle (les bases d'enquête fondées sur la téléphonie filaire perdent de leur pertinence en raison de la démographie, les taux de réponse aux enquêtes non obligatoires en vertu de la loi sont faibles, etc.). Cela explique, en partie, la trop large utilisation du **moissonnage du Web** et de l'**échantillonnage non probabiliste**.

ÉCHANTILLONNAGE NON PROBABILISTE

Les méthodes d'**échantillonnage non probabiliste** sélectionnent les unités d'échantillonnage de la population cible à l'aide d'approches subjectives et non aléatoires.

- L'échantillonnage non probabiliste a le mérite d'être rapide, relativement peu coûteux et pratique (aucune base d'enquête requise).
- Les méthodes d'échantillonnage non probabiliste sont idéales pour l'analyse exploratoire et l'élaboration des enquêtes.

Malheureusement, on a souvent recours aux échantillonnages non probabilistes au lieu des échantillonnages probabilistes (ce qui est problématique).

- Le biais de sélection qui y est associé rend les méthodes d'échantillonnage non probabiliste peu sûres lorsqu'il s'agit d'inférences (elles ne peuvent être utilisées pour fournir des estimations fiables de l'erreur d'échantillonnage, la seule composante de l'ET sur laquelle l'analyste a une emprise directe).
- La collecte automatisée des données tombe souvent dans le champ des échantillonnages non probabilistes – il est toujours possible d'analyser les données recueillies selon cette méthode, mais pas de généraliser les résultats à la population cible.

ÉCHANTILLONNAGE PROBABILISTE

Les plans d'échantillonnage probabiliste sont généralement plus **difficiles** et plus **coûteux** à mettre en place (car ils requièrent une base d'enquête de qualité), et ils prennent plus de temps à réaliser.

Ils fournissent des **estimations fiables** de la caractéristique d'intérêt et de l'**erreur d'échantillonnage**, ouvrant la voie à l'utilisation de petits échantillons pour tirer des inférences sur des populations cibles plus vastes (en théorie, du moins, les composantes de l'erreur non attribuable à l'échantillonnage peuvent tout de même jouer sur les résultats et la généralisation).

INTERVALLES DE CONFIANCE

Si l'estimation $\hat{\beta}$ est non biaisée, $E(\hat{\beta} - \beta) = 0$, **un intervalle de confiance au seuil d'environ 95 %** (IC à 95 %) pour β est alors donné approximativement par

$$\hat{\beta} \pm 2\sqrt{\hat{V}(\hat{\beta})},$$

où $\hat{V}(\hat{\beta})$ est une estimation **propre au plan d'échantillonnage** de $V(\hat{\beta})$.

Mais à quoi correspond exactement un IC à 95 %?

PLAN D'ÉCHANTILLONNAGE

Les différents **plans d'échantillonnage** présentent des avantages et des inconvénients distincts.

Ils peuvent servir à calculer des estimations

- pour diverses caractéristiques de la population : moyenne, total, proportion, ratio, différence, etc.
- pour l'IC à 95 % correspondant.

On pourrait aussi vouloir calculer la taille des échantillons pour une **limite d'erreur** donnée (une limite supérieure à l'intérieur de l'IC à 95 % désiré), et déterminer la **répartition de l'échantillon** (combien d'unités à échantillonner dans divers groupes de sous-population).

PLAN D'ÉCHANTILLONNAGE – L'UNIVERS DU DISCOURS

Population cible :

- N unités et mesures $\mathcal{U} = \{u_1, \dots, u_N\}$

Caractéristiques réelles de la population :

- moyenne μ , variance σ^2 , total τ , proportion p

Échantillon de population :

- n unités et mesures $\mathcal{Y} = \{y_1, \dots, y_n\} \subseteq \mathcal{U}$

Caractéristiques de l'échantillon de population :

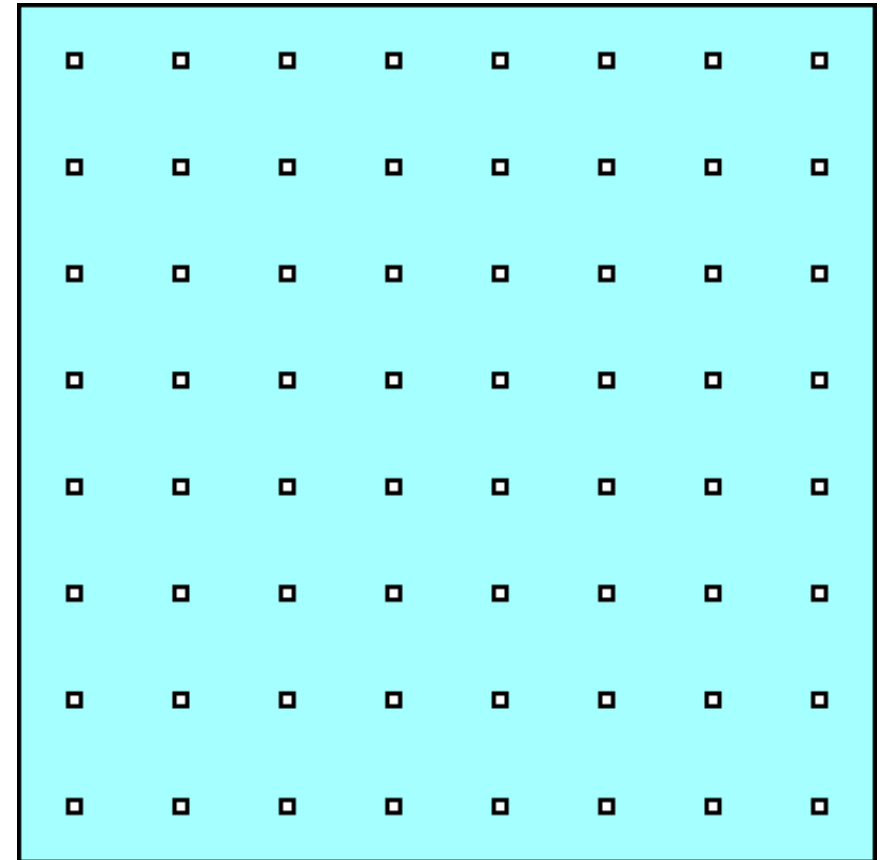
- moyenne de l'échantillon \bar{y} , variance de l'échantillon s^2 , total de l'échantillon $\hat{\tau}$, proportion de l'échantillon \hat{p}

PLAN D'ÉCHANTILLONNAGE – L'UNIVERS DU DISCOURS

Objectif : faire l'estimation des véritables caractéristiques de la population μ, σ^2, τ, p grâce aux caractéristiques de l'échantillon de population $\bar{y}, s^2, \hat{\tau}, \hat{p}, n$, et à la taille N de la population cible.

Pour une caractéristique donnée, on définit δ_i comme prenant la valeur 1 ou 0 selon que l'unité échantillon y_i possède ou non la caractéristique en question.

On utilise la limite d'erreur $B = 2\sqrt{\hat{V}}$.



ÉCHANTILLONNAGE ALÉATOIRE SIMPLE (EAS)

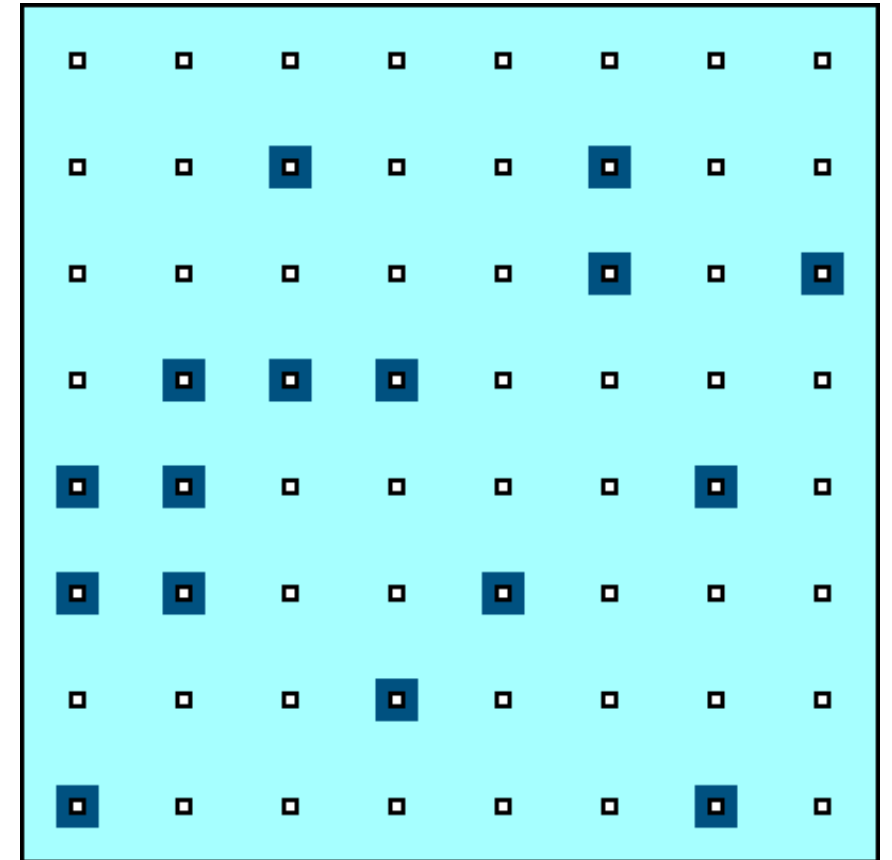
Dans l'EAS, n unités sont sélectionnées au hasard dans la base.

Avantages :

- Plan d'échantillonnage le plus facile à mettre en place
- Erreurs d'échantillonnage bien connues et faciles à estimer
- Pas nécessaire d'avoir de données auxiliaires

Inconvénients :

- Ne fait aucunement appel aux données auxiliaires
- Ne fournit aucune garantie quant à la représentativité de l'échantillon
- Coûteux si l'échantillon est largement réparti géographiquement



ESTIMATEURS DE L'EAS

Estimateurs :

$$\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i, \quad \hat{t} = N\bar{y}, \quad \hat{p} = \frac{1}{n} \sum_{i=1}^n \delta_i$$

Estimations des variances propres au plan d'échantillonnage :

$$\hat{V}(\bar{y}) = \frac{s^2}{n} \left(1 - \frac{n}{N}\right), \quad \hat{V}(\hat{t}) = N^2 \hat{V}(\bar{y}), \quad \hat{V}(\hat{p}) = \frac{\hat{p}(1-\hat{p})}{n} \left(1 - \frac{n}{N}\right)$$

Répartition de l'échantillon :

$$n_{\bar{y}} = \frac{4N\tilde{\sigma}^2}{(N-1)B^2 + 4\tilde{\sigma}^2}, \quad n_{\hat{t}} = \frac{4N^3\tilde{\sigma}^2}{(N-1)B^2 + 4N^2\tilde{\sigma}^2}, \quad n_{\hat{p}} = \frac{4\tilde{p}(1-\tilde{p})}{(N-1)B^2 + 4\tilde{p}(1-\tilde{p})}$$

ÉCHANTILLONNAGE ALÉATOIRE STRATIFIÉ

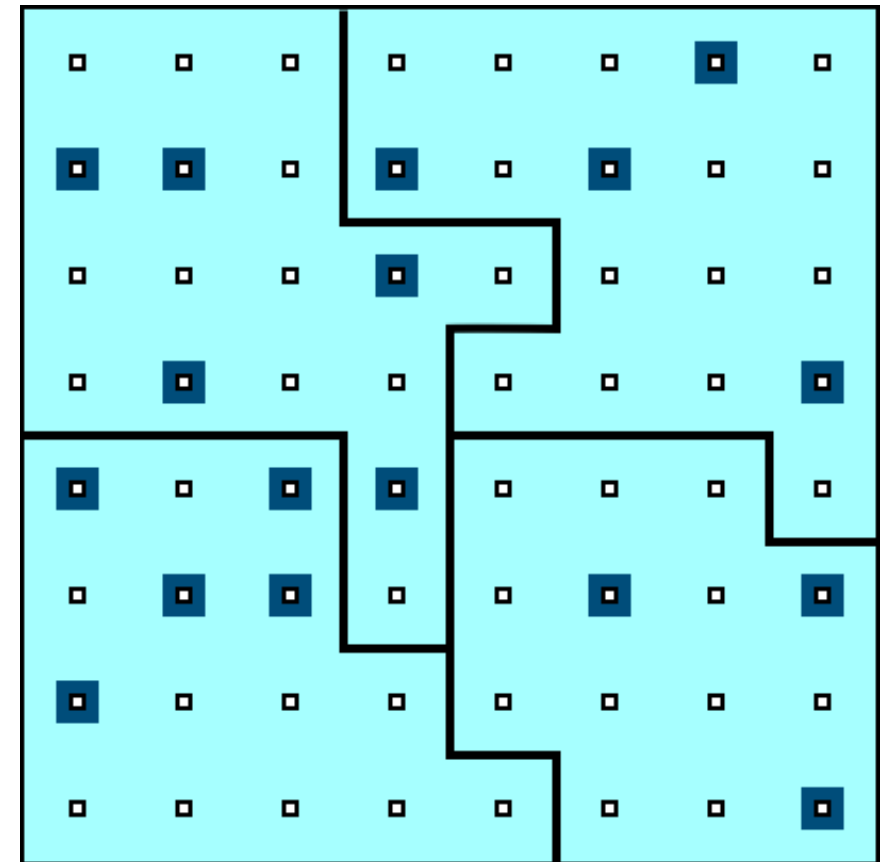
Dans l'échantillonnage aléatoire stratifié, $n = n_1 + \dots + n_k$ unités sont sélectionnées de manière aléatoire à partir de k **strates** de la base.

Avantages :

- Peut produire une limite d'erreur inférieure sur l'estimation, en comparaison de l'EAS;
- Peut être moins coûteux à condition de stratifier adéquatement les éléments;
- Peut fournir des estimations pour des sous-populations.

Inconvénients :

- Aucun inconvénient majeur
- S'il n'existe aucun moyen naturel de stratifier la base d'enquête en groupes homogènes, l'échantillonnage aléatoire stratifié devient à peu près équivalent à l'EAS



ESTIMATEURS DE L'ÉCHANTILLONNAGE ALÉATOIRE STRATIFIÉ

Estimateurs :

$$\bar{y}_{st} = \sum_{j=1}^k \frac{N_j}{N} \bar{y}_j, \quad \hat{t}_{st} = N \bar{y}_{st}, \quad \hat{p}_{st} = \sum_{j=1}^k \frac{N_j}{N} \hat{p}_j$$

Estimations des variances propres au plan d'échantillonnage :

$$\hat{V}(\bar{y}_{st}) = \frac{1}{N^2} \sum_{j=1}^k N_j^2 \hat{V}(\bar{y}_j), \quad \hat{V}(\hat{t}_{st}) = N^2 \hat{V}(\bar{y}_{st}), \quad \hat{V}(\hat{p}_{st}) = \frac{1}{N^2} \sum_{j=1}^k N_j^2 \hat{V}(\hat{p}_j)$$

EXERCICES

On vous demande de faire une estimation du salaire annuel des scientifiques des données au Canada.

Cernez les éléments possibles suivants :

- populations (cible, étude, répondant, bases d'enquête);
- échantillons (prévus, obtenus);
- informations sur l'unité (unité, variable de réponse, caractéristique de population);
- sources de biais (couverture, non-réponse, échantillonnage, mesure) et de variabilité (échantillonnage, mesure).

EXERCICES

Le fichier `cities.txt` contient des informations sur la population urbaine d'un pays. On définit qu'une ville est « petite » si sa population est inférieure à 75 000 habitants, « moyenne » si elle se situe entre 75 000 et 1 million d'habitants, et « grande » au-delà.

1. Trouvez le fichier et téléchargez-le dans l'espace de travail de votre choix. Combien y a-t-il de villes? Combien dans chaque groupe?
2. Affichez les statistiques démographiques sommaires des villes, à la fois globalement et par groupe.
3. Calculez un IC à 95 % pour la moyenne de la population en 1999 en utilisant un EAS de taille $n = 10$.
4. Calculez un IC à 95 % pour la moyenne de la population en 1999 en utilisant un échantillonnage aléatoire stratifié de taille $(n_s, n_m, n_l) = (5, 3, 2)$.
5. Comparez les estimations avec la valeur réelle. Les résultats sont-ils étonnants? Sinon, auraient-ils pu l'être?

Matériel supplémentaire

FACTEURS DÉCISIFS

Dans certains cas, il faut disposer de l'information sur l'**ensemble** de la population pour répondre aux questions, alors que dans d'autres, ce n'est pas nécessaire. Le **type d'enquête** dépend de multiples facteurs, notamment :

- le type de questions auxquelles il faut répondre
- la précision requise
- le coût du sondage d'une unité
- le temps requis pour sonder une unité
- la taille de la population faisant l'objet de l'enquête
- la prévalence des caractéristiques d'intérêt

ÉTAPES DE L'ÉTUDE OU DE L'ENQUÊTE

Les études ou enquêtes suivent les mêmes étapes générales :

1. énoncé de l'objectif
2. sélection de la base d'enquête
3. plan d'échantillonnage
4. plan du questionnaire
5. collecte des données
6. saisie et codage des données
7. traitement des données et imputation
8. estimation
9. analyse des données
10. diffusion
11. documentation

Le processus n'est pas toujours linéaire, mais il existe un cheminement clair depuis l'objectif jusqu'à la diffusion.

BASES D'ENQUÊTE

La **base** d'enquête permet de **sélectionner** et de **contacter** les unités de population visées par l'enquête. Sa création et sa maintenance sont en général coûteux (en fait, il existe des organisations et des entreprises spécialisées dans la constitution ou la vente de telles bases).

Les bases utiles contiennent :

- les données d'identification
- les données de contact
- les données de classification
- les données de maintenance
- les données de couplage

MODES DE COLLECTE DES DONNÉES

Sur support papier ou assisté par ordinateur

- Les **questionnaires auto-administrés** sont utilisés lorsque l'enquête nécessite des renseignements détaillés pour permettre aux unités de consulter les dossiers personnels; associés à un taux de non-réponse élevé.
- Les **questionnaires assistés par un intervieweur** adéquatement formé sont utilisés pour augmenter le taux de réponse et la qualité globale des données; en personne ou au téléphone.
- Les **entrevues assistées par ordinateur** associent la collecte et la saisie des données, ce qui fait gagner du temps.
- Observation directe discrète.
- Journaux à remplir (format papier ou électronique).
- Enquêtes omnibus.
- Courriel, Internet et médias sociaux.

MÉTHODES D'ÉCHANTILLONNAGE NON PROBABILISTE

Au hasard

- Un passant; dépend de la disponibilité des unités et du biais lié à l'intervieweur.

Volontaire

- Biais d'autosélection.

A priori

- Biaisé par des idées préconçues inexactes concernant la population cible.

Par quotas

- Sondage fait à la sortie de l'isoloir, ignore le biais de non-réponse.

MÉTHODES D'ÉCHANTILLONNAGE NON PROBABILISTE

Modifié

- D'abord probabiliste, puis par quotas en réaction à des taux de non-réponse élevés

En boule de neige

- Plan « pyramidal »

Dans certains contextes, les méthodes d'échantillonnage non probabiliste pourraient répondre aux besoins d'un client ou d'une organisation (et c'est à eux qu'il appartient de prendre la décision en dernier lieu), mais on doit l'informer des inconvénients et lui proposer des solutions probabilistes.

CONCEPTS MATHÉMATIQUES DE BASE

Posons une population finie \mathcal{U} , avec N unités et mesures $\{u_1, \dots, u_N\}$.

La **moyenne** et la **variance** de la population pour la variable d'intérêt sont données par

$$\mu = \frac{1}{N} \sum_{j=1}^N u_j, \quad \sigma^2 = \frac{1}{N} \sum_{j=1}^N (u_j - \mu)^2.$$

Si $\mathcal{Y} \subseteq \mathcal{U}$ représente un **échantillon** de la population avec n unités et mesures $\{y_1, \dots, y_n\}$, alors la **moyenne** et la **variance de l'échantillon** sont données par

$$\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i, \quad s^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2.$$

CONCEPTS MATHÉMATIQUES DE BASE

Soit X_1, \dots, X_n des **variables aléatoires**, $b_1, \dots, b_n \in \mathbb{R}$, et E , V , Cov l'**espérance**, la **variance** et la **covariance**, respectivement, c.-à-d. :

- $E(X_i) = \mu_i$
- $\text{Cov}(X_i, X_j) = E(X_i X_j) - E(X_i)E(X_j)$
- $V(X_i) = \text{Cov}(X_i, X_i) = E(X_i^2) - E^2(X_i) = E(X_i^2) - \mu_i^2 = \sigma_i^2$ and

$$E\left(\sum_{i=1}^n b_i X_i\right) = \sum_{i=1}^n b_i E(X_i) = \sum_{i=1}^n b_i \mu_i$$
$$V\left(\sum_{i=1}^n b_i X_i\right) = \sum_{i=1}^n b_i^2 V(X_i) + \sum_{i \neq j} b_i b_j \text{Cov}(X_i, X_j)$$

CONCEPTS MATHÉMATIQUES DE BASE

Le **biais** d'une composante d'erreur est la moyenne de cette composante d'erreur si l'enquête est répétée plusieurs fois indépendamment et dans les mêmes conditions. Une estimation **sans biais** est une estimation pour laquelle le biais est nul.

La **variabilité** d'une composante d'erreur est la mesure dans laquelle cette composante varierait par rapport à sa valeur moyenne dans le scénario idéal décrit ci-dessus.

L'**erreur quadratique moyenne** d'une composante d'erreur est une mesure de sa taille :

$$\text{MSE}(\hat{\beta}) = V(\hat{\beta}) + \text{Bias}^2(\hat{\beta}),$$

Où $\hat{\beta}$ est un estimateur de β .

PLANS D'ÉCHANTILLONNAGE PROBABILISTE

Échantillonnage aléatoire simple (EAS)

Échantillonnage aléatoire stratifié

Échantillonnage systématique

Échantillonnage en grappes

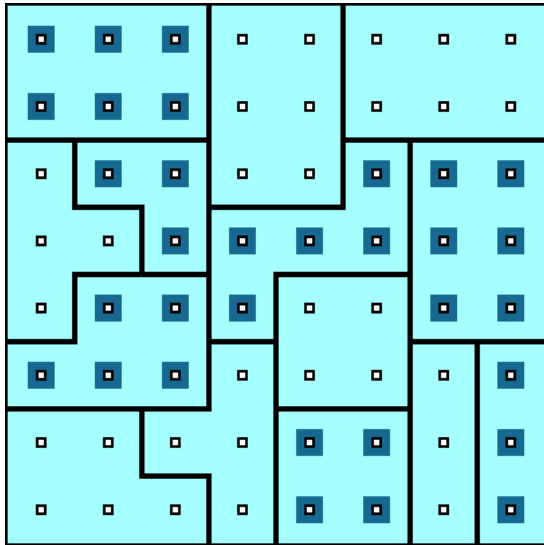
Échantillonnage avec probabilité proportionnelle à la taille (PPT)

Échantillonnage répété

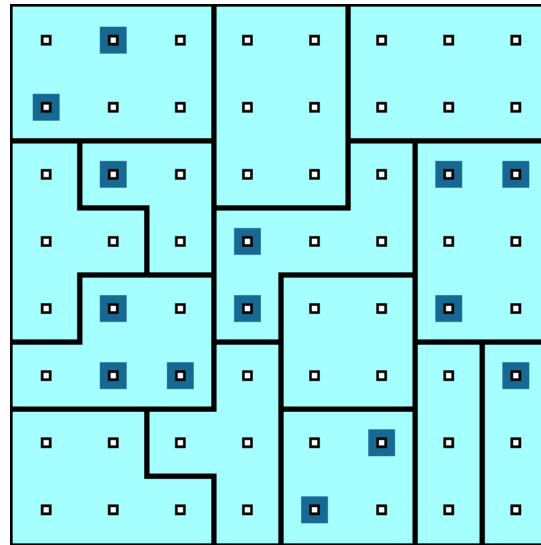
Échantillonnage à plusieurs degrés

Échantillonnage à plusieurs phases

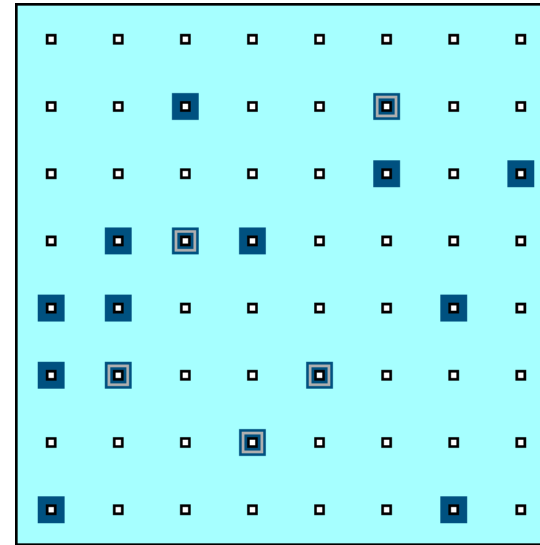
AUTRES EXEMPLES DE PLANS D'ÉCHANTILLONNAGE



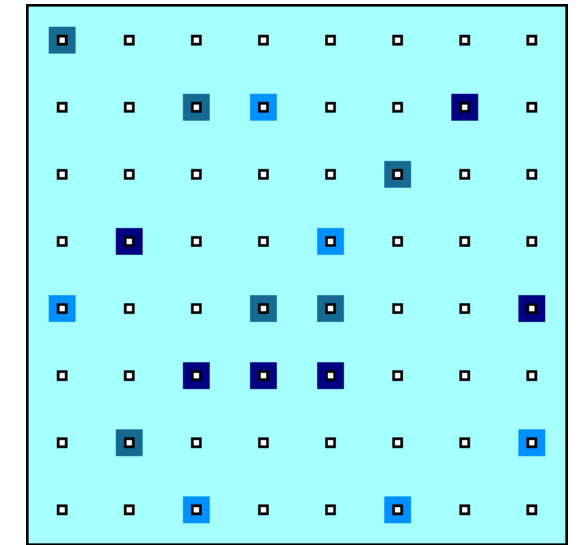
Échantillonnage en
grappes



Échantillonnage à
plusieurs degrés



Échantillonnage à
plusieurs phases



Échantillonnage répété

API ET MOISSONNAGE DU WEB

COLLECTE ET TRAITEMENT DES DONNÉES

« Les rues du Web sont pavées de données qui n'attendent que d'être recueillies. »

Munzart, Rubba, Meissner, Nyhuis, Automated Data Collection with R

WORLD WIDE WEB

Il fut un temps, assez récent, où tant la rareté des données que leur inaccessibilité constituaient un problème pour les chercheurs et les décideurs. Tel n'est **manifestement** plus le cas désormais.

L'abondance des données présente son propre lot de problèmes particuliers :

- des masses de données enchevêtrées
- les méthodes classiques de collecte des données et les techniques usuelles d'analyse des données (en petites quantités) peuvent ne plus suffire aujourd'hui

EXEMPLE DE MOISSONNAGE DU WEB – NOUVEAU TÉLÉPHONE

Supposons que vous aimeriez savoir ce que la population pense d'un nouveau téléphone. Approche standard : étude de marché (p. ex. sondage téléphonique, système de récompenses, etc.).

Pièges :

- échantillon non représentatif : il se pourrait que l'échantillon sélectionné ne représente pas la population visée
- non-réponse systématique : les personnes qui n'aiment pas les sondages téléphoniques pourraient être moins (ou plus) portées à ne pas aimer le nouveau téléphone
- erreur de couverture : à titre d'exemple, il serait impossible de joindre les personnes qui ne disposeraient pas d'un téléphone filaire
- erreur de mesure : les questions du sondage fournissent-elles des renseignements convenant au problème posé?

QUALITÉ DES DONNÉES DU WEB – NOUVEAU TÉLÉPHONE

Ces solutions peuvent être **onéreuses, chronophages, inefficaces**.

Variables de substitution – indicateurs qui sont étroitement reliés à la popularité du produit, sans mesurer directement celle-ci pour autant.

Si la notion de **popularité** renvoie au fait que de grands groupes de gens préfèrent un produit par rapport à un produit concurrent, les statistiques de vente que l'on retrouve sur un site Web commercial pourraient constituer un substitut de la popularité.

Les classements sur Amazon pourraient offrir une idée plus **complète** du marché des téléphones par rapport à ce que permettrait d'obtenir un sondage classique.

PROBLÈMES POTENTIELS – NOUVEAU TÉLÉPHONE

Représentativité des produits répertoriés

- Tous les téléphones sont-ils répertoriés?
- Si tel n'est pas le cas, est-ce parce que le site Web ne les vend pas?
- Y a-t-il une autre raison?

Représentativité des clients

- Certains groupes spécifiques achètent-ils/n'achètent-ils pas de produits en ligne?
- Certains groupes spécifiques achètent-ils sur des sites spécifiques?
- Certains groupes spécifiques laissent-ils ou non des commentaires?

Honnêteté des clients et **crédibilité** des commentaires.

LE MOISSONNAGE DU WEB EST-IL LÉGAL?

Qu'est-ce qu'une araignée?

- Il s'agit d'un programme qui parcourt ou arpente le Web pour en extraire de l'information rapidement
- L'araignée, ou programme collecteur, saute d'une page à l'autre, en en extrayant l'intégralité du contenu

Le **moissonnage** consiste à extraire de l'information spécifique de sites Web spécifiques (c'est le but) : en quoi ces méthodes sont-elles **différentes**?

« Comme, fondamentalement, le moissonnage consiste à **copier** de l'information, l'une des revendications les plus évidentes à l'encontre des dispositifs de récupération de données tient à la violation du droit d'auteur. »

ACTIONS EN JUSTICE – MOISSONNAGE DU WEB

eBay c. Bidder's Edge (BE)

- BE a eu recours à des programmes automatisés pour extraire de l'information de différents sites de vente aux enchères.
- Les utilisateurs pouvaient consulter les listes sur la page Web de BE, plutôt que d'avoir à se rendre sur les différents sites de vente aux enchères.
- En 1999, BE a accédé aux sites d'eBay environ 100 000 fois par jour (1,53 % du nombre de requêtes, 1,1 % de l'ensemble des données transférées par eBay).
- eBay a réclamé des dommages-intérêts allant de 45 000 \$ et 62 000 \$, sur une période de 10 mois.
- BE n'a volé aucune information qui n'était pas déjà publique, mais l'augmentation du trafic a imposé une charge additionnelle aux serveurs d'eBay.
- **Votre verdict?**

COOPÉRATION AMICALE AVEC LES API

Qu'est-ce qu'une API? L'acronyme « API » signifie *application program interface*, ou interface de programmation d'applications, soit un ensemble de routines, de protocoles et d'outils pour la construction d'applications logicielles.

Plusieurs API restreignent l'utilisateur à un certain nombre d'appels d'API par jour (ou à d'autres formes de limites).

Il importe de respecter ces limites.

Matériel supplémentaire

POURQUOI PROCÈDE-T-ON À LA COLLECTE AUTOMATISÉE DES DONNÉES?

En ce qui concerne les données scientifiques sociales :

- caractère limité des ressources financières
- peu de temps ou de désir de recueillir les données manuellement
- désir de travailler avec des sources riches en données à jour et de grande qualité
- documenter le processus du début (collecte des données) à la fin (publication) de sorte qu'il puisse être reproduit

Problèmes que pose la collecte manuelle :

- processus non reproductible
- présente des risques d'erreur en plus d'être lourd
- présente un risque plus élevé de « mourir d'ennui »

POURQUOI PROCÉDER À LA COLLECTE AUTOMATISÉE DES DONNÉES?

Avantages des solutions fondées sur un programme :

- fiabilité
- reproductibilité
- rapidité
- groupe d'ensembles de données de meilleure qualité

LISTE DE VÉRIFICATION APPLICABLE À LA COLLECTE AUTOMATISÉE

Le **moissonnage du Web** ou le **traitement de texte statistique** (collecte automatisée ou semi-automatisée des données) est-il absolument nécessaire?

Critères :

- Prévoyez-vous répéter l'opération de temps à autre, p. ex. pour mettre à jour votre base de données?
- Désirez-vous que d'autres puissent reproduire votre processus de collecte des données?
- Traitez-vous fréquemment avec des sources de données en ligne?
- La tâche est-elle non négligeable en termes de portée et de complexité?

LISTE DE VÉRIFICATION APPLICABLE À LA COLLECTE AUTOMATISÉE

Critères : (suite)

- Si la tâche peut être réalisée manuellement, n'avez-vous pas à votre disposition les ressources nécessaires pour laisser autrui faire le travail?
- Êtes-vous disposé à automatiser le processus au moyen de la programmation?

Si la plupart des réponses sont données par l'affirmative, une méthode automatisée pourrait être la voie à suivre.

WORLD WIDE WEB

La façon dont nous **partageons**, **recueillons** et **publions** les données a changé au cours des dernières années, du fait de l'omniprésence du *World Wide Web* (WWW).

Les **entreprises privées**, les **gouvernements** et les **utilisateurs individuels** publient et partagent toutes sortes de données et d'information.

À tout moment, de nouveaux canaux génèrent de vastes quantités de données sur le comportement humain.

LOGICIEL LIBRE

Une autre tendance :

- la croissance, ainsi que la popularité et la puissance sans cesse plus grandes des **logiciels libres** (le code source peut être inspecté, modifié et amélioré par quiconque)

Aspect communautaire → évolution continue et amélioration constante

Les logiciels **R** et **Python** sont des logiciels libres qui peuvent servir à des fins d'analyse de données dans le domaine des sciences sociales et dans d'autres domaines.

Ils intègrent des **interfaces** avec d'autres langages de programmation et **solutions** logicielles.

NETTOYAGE ET TRAITEMENT DES DONNÉES

La collecte des données, en tant que telle, ne constitue que la pointe de l'iceberg.

Le nettoyage ainsi que le traitement des données sont **essentiels** (en plus de nécessiter du temps).

Tâches :

- Sélection des colonnes (variables) présentant de l'intérêt
- Réétiquetage de ces colonnes
- Modification du type de données des colonnes de sorte que les données puissent être utilisées comme nous le souhaitons

NETTOYAGE ET TRAITEMENT DES DONNÉES

Tâches : (suite)

- Édition et/ou extraction des données d'une colonne
- Décider comment gérer les données manquantes (ce qui peut être délicat)
- De multiples autres tâches, selon les données et leurs utilisations

Certaines tâches peuvent être automatisées, d'autres non.

QUESTIONS AU SUJET DE LA QUALITÉ DES DONNÉES

1. Quel type de données est le plus approprié pour répondre à vos questions?
2. La qualité des données est-elle suffisamment élevée pour répondre à votre question?
3. L'information est-elle systématiquement déficiente?

Pouvez-vous parvenir à éviter la redoutée formule : « Eh bien, ce sont les meilleures données dont nous disposons... »?

QUALITÉ DES DONNÉES

Information de première main : à titre d'exemple, un gazouillis ou un article de nouvelles.

Données de deuxième main : données qui ont été copiées d'une source hors ligne ou extraites d'ailleurs.

Parfois, vous ne pouvez vous souvenir de la source des données ou retrouver celle-ci, lorsqu'il s'agit de données de deuxième main.

Convient-il tout de même de s'en servir? Cela dépend.

La **validation croisée** constitue une procédure standard liée à l'utilisation de toute donnée secondaire.

QUALITÉ DES DONNÉES ET OBJECTIF DE L'UTILISATEUR

La qualité des données est fonction de l'**utilisation**.

Par exemple :

- Un échantillon de gazouillis recueillis au cours d'une journée aléatoire pourrait servir à analyser l'utilisation qui est faite d'un mot-clic ou l'utilisation de termes selon le sexe.
- Pas aussi utiles si elles sont recueillies le jour de l'élection pour prédire les résultats de celle-ci (**biais associé à la collecte**).

SOURCES DE DONNÉES (COMPROMIS)

Automatisée c. classique

Exactitude c. exhaustivité

Couverture c. validité

Vitesse c. coût

etc.

PROCESSUS DE COLLECTE DES DONNÉES (5 ÉTAPES)

1. Savoir exactement de quel type d'information vous avez besoin

- Spécifique : PIB de tous les pays membres de l'OCDE au cours des dix dernières années; ventes des dix principales marques de chaussures en 2017
- Vague : l'opinion des gens sur la marque de chaussures X

2. Déterminer s'il existe des sources de données sur le Web qui pourraient fournir de l'information directe ou indirecte sur votre problème

- Plus facile dans le cas de faits spécifiques : la page Web d'un magasin de chaussures fournira de l'information sur les chaussures qui sont actuellement prisées, c.-à-d. sandales, bottes, etc.
- Les gazouillis peuvent permettre de dégager des tendances en matière d'opinion sur *tout et n'importe quoi*
- Les plateformes commerciales peuvent fournir de l'information sur le niveau de satisfaction à l'égard d'un produit

PROCESSUS DE COLLECTE DES DONNÉES (5 ÉTAPES)

3. Élaborer une théorie quant au processus de production des données lorsque l'on se penche sur des sources éventuelles

- Quand les données ont-elles été générées?
- Quand ont-elles été téléchargées sur le Web?
- Qui a téléchargé les données?
- Y a-t-il d'autres aspects qui pourraient ne pas avoir été couverts? Cohérence? Précision?
- À quelle fréquence les données sont-elles mises à jour?

PROCESSUS DE COLLECTE DES DONNÉES (5 ÉTAPES)

4. Trouver un équilibre entre les avantages et les inconvénients des sources de données potentielles

- Valider la qualité des données utilisées
- Existe-t-il d'autres sources indépendantes qui fournissent de l'information similaire, et par rapport auxquelles il serait possible de procéder à une vérification croisée?
- Pouvez-vous identifier la source originale des données secondaires?

5. Prendre une décision

- Choisir la source de données qui semble la plus appropriée
- Documenter les raisons de cette décision
- Recueillir des données de plusieurs sources afin de valider les sources de données

LE MOISSONNAGE DU WEB EST-IL LÉGAL?

Lignes directrices en matière d'éthique :

- Travailler de manière aussi transparente que possible
- Documenter les sources de données en tout temps
- Accorder le crédit à ceux qui, les premiers, ont recueilli et publié les données
- Si vous n'avez pas recueilli l'information, vous aurez vraisemblablement besoin d'une permission pour la reproduire
- Ne faites rien d'illégal

L'extraction d'information d'une autre entreprise en vue de la traiter et de la revendre constitue un motif de plainte courant.

ACTIONS EN JUSTICE – MOISSONNAGE DU WEB

Associated Press (AP) c. Meltwater

- Meltwater offre un logiciel qui permet de récupérer ou d'extraire des nouvelles au moyen de mots clés spécifiques.
- Les clients commandent des résumés portant sur certains thèmes en particulier dans lesquels figurent des extraits d'articles de nouvelles.
- AP affirmait que son contenu avait été volé et que Meltwater avait besoin d'une licence pour distribuer l'information extraite.
- Le juge a rendu une décision en faveur d'AP, faisant valoir que Meltwater était un concurrent.
- **Votre verdict?**

ACTIONS EN JUSTICE – MOISSONNAGE DU WEB

Facebook c. Pete Warden

- Pete Warden extrayait des renseignements de base des profils d'utilisateurs de Facebook, afin d'offrir des services de gestion des communications et des réseaux.
- Selon lui, son processus allait dans le même sens que robots.txt.
- Après qu'il a eu publié son premier billet de blogue faisant mention des données extraites de Facebook, il a été invité à effacer celles-ci.
- Facebook a fait valoir que robots.txt n'avait aucune valeur juridique et qu'elle pouvait poursuivre quiconque accédait à son site, même si cette personne ou ce groupe se conformait aux instructions en matière de moissonnage, et que la seule façon légale d'accéder à quelque site Web que ce soit au moyen d'un robot Web, c'était en obtenant une autorisation écrite préalable.
- **Votre verdict?**

ACTIONS EN JUSTICE – MOISSONNAGE DU WEB

États-Unis c. Aaron Swartz

- Swartz a participé à la création de RSS, markdown, et Infogami.
- Il a été arrêté en 2011 pour avoir téléchargé illégalement des millions d'articles des archives de JSTOR.
- L'affaire a été classée après qu'il se fut suicidé, en janvier 2013.
- **Votre verdict?**

LEÇONS APPRISSES

On ne peut établir clairement quelles mesures de moissonnage sont illégales et lesquelles sont légales.

On estime que le fait de publier de nouveau du contenu à des fins commerciales est plus grave que ne l'est celui qui consiste à télécharger des pages à des fins de recherche ou d'analyse.

Robots.txt : le *protocole d'exclusion des robots* est un fichier qui indique au logiciel de récupération quelle information peut être recueillie sur le site.

Soyez gentil! Il n'est pas nécessaire de récupérer tout ce qui peut être récupéré. Les programmes de récupération devraient se comporter « gentiment », fournir les données que vous recherchez en plus d'être efficaces, dans cet ordre de priorité.

robots.txt cqads.carleton.ca/robots.txt

```
# This file is to prevent the crawling and indexing of certain parts
# of your site by web crawlers and spiders run by sites like Yahoo!
# and Google. By telling these "robots" where not to go on your site,
# you save bandwidth and server resources.

# This file will be ignored unless it is at the root of your host:
# Used:    http://example.com/robots.txt
# Ignored: http://example.com/site/robots.txt
#
# For more information about the robots.txt standard, see:
# http://www.robotstxt.org/robotstxt.html
```

```
User-agent: *
Crawl-delay: 10
# Directories
Disallow: /includes/
Disallow: /misc/
Disallow: /modules/
Disallow: /profiles/
Disallow: /scripts/
Disallow: /themes/
# Files
Disallow: /CHANGELOG.txt
Disallow: /cron.php
Disallow: /INSTALL.mysql.txt
Disallow: /INSTALL.pgsql.txt
Disallow: /INSTALL.sqlite.txt
Disallow: /install.php
Disallow: /INSTALL.txt
Disallow: /LICENSE.txt
Disallow: /MAINTAINERS.txt
Disallow: /update.php
Disallow: /UPGRADE.txt
Disallow: /xmlrpc.php
```

```
User-agent: Twitterbot
Allow: /
```

```
User-agent: *
Disallow: /esi/
Disallow: /webview
Disallow: /vweb
Disallow: /news/sponsored
Disallow: /search
Disallow: /19849159/
```

theweathernetwork.com/robots.txt

```
User-agent: *
Disallow:
Crawl-delay: 10
```

cfl.ca/robots.txt

COMMUNIQUER AVEC LES FOURNISSEURS DE DONNÉES

Toutes les données auxquelles il est possible d'accéder par le truchement d'un formulaire HTTP sont stockées dans une base de données quelconque.

Demandez tout d'abord aux propriétaires des données s'ils peuvent concéder l'accès à la base de données ou aux fichiers.

Plus la quantité de données qui vous intéressent est importante, **plus il est préférable, pour les deux parties, de communiquer avant le lancement de la collecte de données.**

Pour des petites quantités de données, cela a moins d'importance.

CE QU'IL FAUT ET CE QU'IL NE FAUT PAS FAIRE EN MATIÈRE DE MOISSONNAGE

1. Demeurer identifiable

2. Réduire le trafic

- Accepter les fichiers comprimés
- En cas de moissonnage des mêmes ressources à plusieurs reprises, vérifiez tout d'abord si celles-ci ont changé avant d'y accéder de nouveau
- Ne récupérer que des parties de fichier

CE QU'IL FAUT ET CE QU'IL NE FAUT PAS FAIRE EN MATIÈRE DE MOISSONNAGE

3. Ne pas soumettre de demandes multiples au serveur

- Le fait de soumettre de nombreuses demandes par seconde peut entraîner la mise hors service des serveurs peu puissants
- Les webmaîtres peuvent vous bloquer si votre logiciel de récupération de données se comporte de cette façon
- On considère qu'une ou deux demandes par seconde est un rythme acceptable

4. Concevoir un logiciel de récupération de données modeste (efficient et poli)

- Il est inutile de récupérer des pages quotidiennement ou de répéter la même tâche sans cesse; faites en sorte que votre programme de récupération de données soit aussi efficient que possible
- Ne pas soumettre des pages à un trop grand nombre de demandes récupération
- Sélectionner les ressources que vous souhaitez utiliser et laisser le reste intact

OUTILS DE DÉVELOPPEMENT

Les outils de développement nous permettent (notamment) d'observer la correspondance entre le code HTML d'une page et la version rendue que nous retrouvons dans le navigateur.

Contrairement à la fonction « Afficher la source », les outils de développement présentent la version *dynamique* du code HTML (c. à d. que les instructions HTML apparaissent sans les modifications apportées par JavaScript depuis que la page a été reçue, la première fois).

Il est essentiel, pour récupérer des données des sites Web de façon efficiente, d'inspecter les divers éléments qui composent une page et de déterminer où ils se trouvent dans le fichier HTML.

OUTILS DE DÉVELOPPEMENT

Firefox

- clic du bouton droit de la souris dans la page → Inspecter l'élément

Safari

- Safari → Préférences → Avancées → Afficher le menu Développer dans la barre de menus
- Développer → Afficher inspecteur Web

Chrome

- clic du bouton droit de la souris dans la page → Inspecter



HOME LIVE SHOWS **ERB** MUSIC VIDEOS GALLERY SHOP PRESS ARCHIVE CONTACT

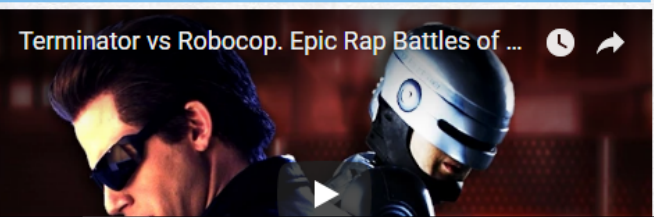
Shaka Zulu vs Julius Caesar



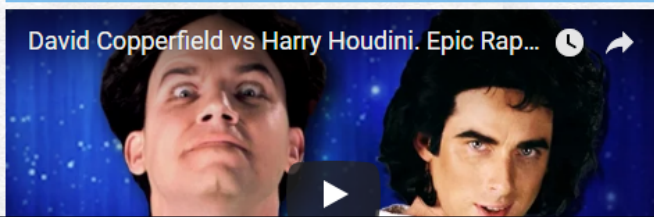
Eastern Philosophers vs Western Philosophers



Terminator vs Robocop



David Copperfield vs Harry Houdini



Nice Peter - ERB

Secure | https://nicepeter.com/erb

FAST Trav f EF

HBC

PC+

Other bookmarks

402

2

Elements Console Sources Network

```

<div id="page-content-wrap">
  <div class="zoogoo-content block" data-arrangement-id="727198" content-width=">100 >120 >140 >160 >180 >200 >220 >240 >260 >280 >300 >320 >340 >360 >380 >400 >420 >440 >460 >480 >500 >520 >540 >560 >580 >600 >620 >640 >660 >680 >700 >720 >740 >760 >780 >800 >820 >840 >860 >880 >900 >920 >940 >960 >980 >1000 >1020 >1040 >1060 >1080 >1100 >1120 >1140 >1160 >1180 >1200 >1220 >1240 >1260 >1280 >1300 >1320 >1340 >1360 >1380 >1400">
    </div>
  <div class="zoogoo-columns zoogoo-columns-2 zoogoo-columns-50-50 default-section-style padding-none title-alignment-left block block-row layout_full zoogoo-columns-first" data-row-id="2869945">
    <div class="zoogoo-columns-inner site-wrap">
      <div class="zoogoo-column zoogoo-column-1-of-2 col_1_of_2 block layout_half" data-column-id="3098090">
        <div class="zoogoo-feature block layout_full block-title-feature" data-block-id="3105122">
          Shaka Zulu vs Julius Caesar. Epic Rap Battle...
        </div>
        <div class="zoogoo-feature block layout_full" data-block-id="3105121">
          Eastern Philosophers vs Western Philosophers
        </div>
        <div class="zoogoo-feature block layout_full" data-block-id="3007634">
          Terminator vs Robocop
        </div>
        <div class="zoogoo-feature block layout_full" data-block-id="3007639">
          David Copperfield vs Harry Houdini
        </div>
      </div>
    </div>
  </div>
</div>

```

Styles

Event Listeners DOM Breakpoints Properties Accessibility

Filter :hov .cls +

element.style { }

#usersite-application-6fb...f72b7b2e.css:1

container .zoogoo-columns-inner {

display: flex;

display: flex;

position: relative;

-webkit-box-pack: justify;

-ms-flex-pack: justify;

-ms-flex-wrap: wrap;

flex-wrap: wrap;

display: flex;

justify-content: space-between;

}

div {

display: block;

}

position 0

margin -

border -

padding -

0 - 1024 x 9516.130 - 0

Filter

Show all

color rgb(0...

display flex

flex-wrap wrap

font-family Mulii, s...

font-size 14px

NICE PETER

HOME LIVE SHOWS ERB MUSIC VIDEOS GALLERY SHOP PRESS ARCHIVE CONTACT

div.zoogoo-feature.block.layout_full | 491.52 x 275.22

Shaka Zulu vs Julius Caesar. Epic Rap Battle...

Eastern Philosophers vs Western Philosophers

Terminator vs Robocop

David Copperfield vs Harry Houdini

XPATH

XPath est un langage d'interrogation (propre à un domaine)

- Il est utilisé pour sélectionner des éléments d'information spécifiques dans des documents balisés, comme HTML, XML ou des variantes telles que SVG et RSS
- L'information stockée dans les documents balisés doit être convertie dans des formats qui se prêtent au traitement et à l'analyse statistique
- Mise en œuvre dans le XML du progiciel R
- Étapes du processus :
 1. Préciser les données présentant de l'intérêt
 2. Les situer dans un document spécifique
 3. Adapter une interrogation au document en vue d'extraire les renseignements souhaités



Robert Gentleman

'What we have is nice, but we need something very different'

Source: Statistical Computing 2003, Reisenburg

Rolf Turner

'R is wonderful, but it cannot work magic'

answering a request for automatic generation of 'data from a known mean and 95% CI'

Source: [R-help](#)

[The book homepage](#)

Bloc-notes : notions fondamentales concernant XPath

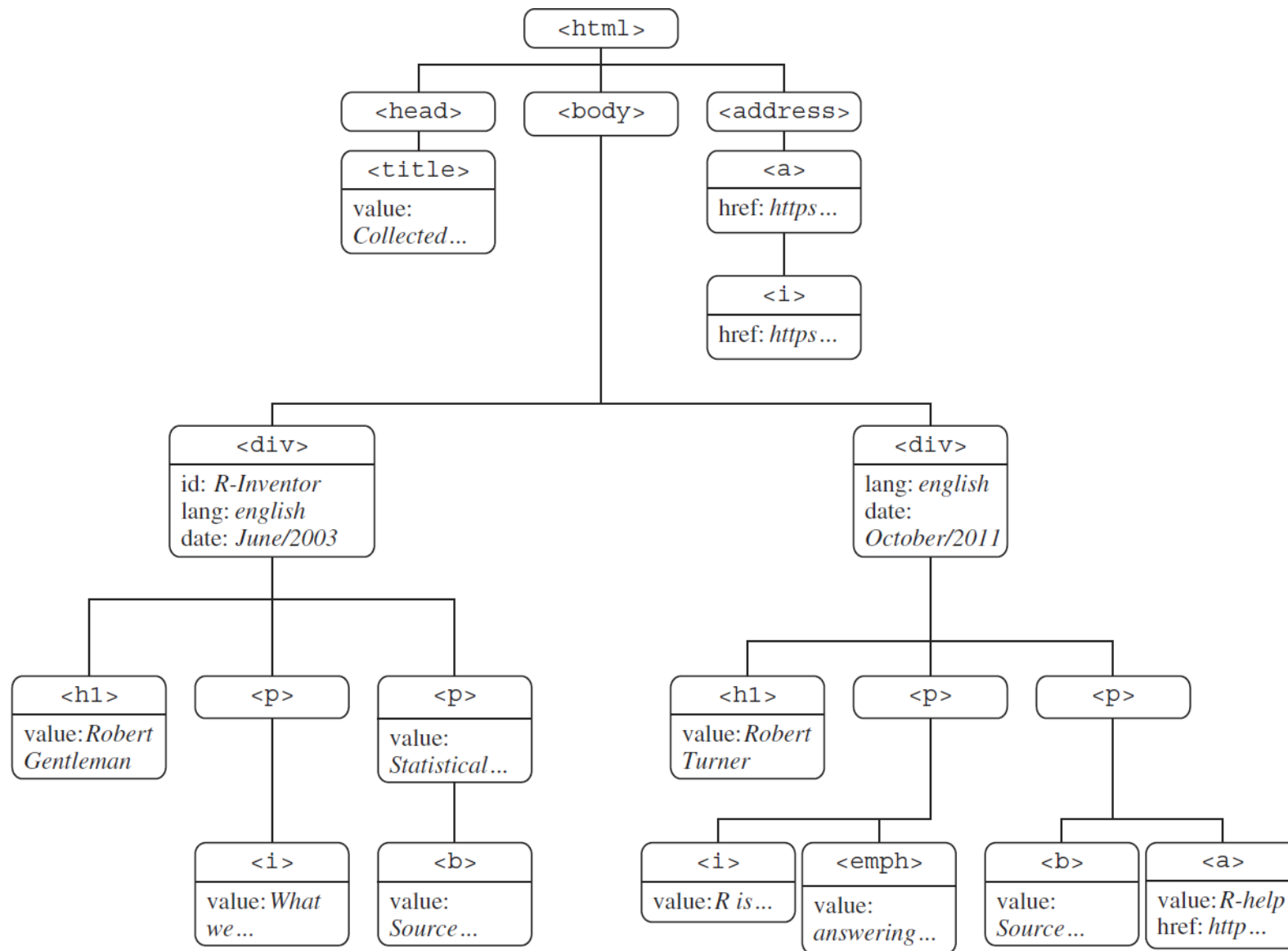


```
<!DOCTYPE HTML PUBLIC "-//IETF//DTD HTML//EN">
<html>
<head><title>Collected R wisdoms</title></head>
<body>
<div id="R Inventor" lang="english" date="June/2003">
  <h1>Robert Gentleman</h1>
  <p><i>'What we have is nice, but we need something very different'</i></p>
  <p><b>Source: </b>Statistical Computing 2003, Reisingburg</p>
</div>

<div lang="english" date="October/2011">
  <h1>Rolf Turner</h1>
  <p><i>'R is wonderful, but it cannot work magic'</i> <br><emph>answering a request
for automatic generation of 'data from a known mean and 95% CI'</emph></p>
  <p><b>Source: </b><a href="https://stat.ethz.ch/mailman/listinfo/r-help">R-help</a>
</p>
</div>

<address>
<a href="http://www.r-datacollectionbook.com"><i>The book homepage</i></a><a></a>
</address>

</body>
</html>
```

XPATH – STRUCTURE DE BASE

Les balises HTML/XML présentent des **attributs** et des **valeurs**.

Les fichiers HTML doivent être analysés avant qu'ils puissent faire l'objet d'une interrogation par XPath.

Les interrogations XPath ont besoin d'un chemin d'accès et d'un document visé par la recherche.

- les chemins d'accès consistent en un mécanisme d'adressage hiérarchique (succession de nœuds, séparés par des barres obliques [« / »])
- les interrogations se présentent selon le format `xpathSApply(doc, path)` :

- `xpathSApply(doc_analysé, "/html/body/div/p/i")`

cette instruction permet d'extraire toutes les balises `<i>` qui se trouvent à l'intérieur d'une balise `<p>` à l'intérieur d'une balise `<div>` dans le `corps` du fichier `html`.

XPATH – RELATIONS DES NŒUDS

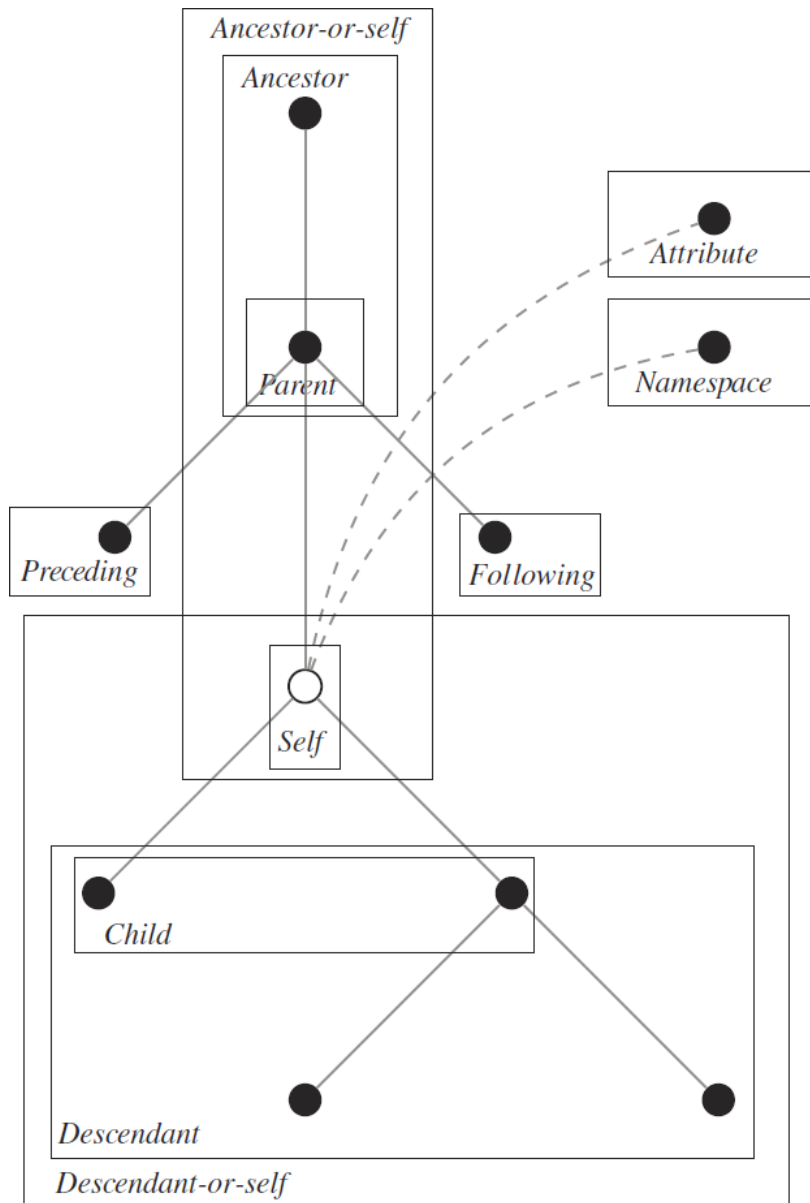
Les chemins d'accès absolus (voire relatifs) ne peuvent pas toujours sélectionner de manière succincte des nœuds dans de gros fichiers ou dans des fichiers complexes.

Analogie de l'arborescence familiale : la place du nœud à l'intérieur de l'arborescence analysée se rapproche fréquemment des relations qu'entretiennent les familles élargies.

Les relations sont désignées par rapport à `node1/relation::node2`.

Exemples :

- « `//a/ancestor::div` » permet d'extraire tous les nœuds `<div>` qui précèdent le nœud `<a>`.
- « `//a/ancestor::div//i` » permet d'extraire tous les nœuds `<i>` qui se trouvent à l'intérieur d'un nœud `<div>` qui précède un nœud `<a>`, etc.



Axis name	Result
ancestor	Selects all ancestors (parent, grandparent, etc.) of the current node
ancestor-or-self	Selects all ancestors (parent, grandparent, etc.) of the current node and the current node itself
attribute	Selects all attributes of the current node
child	Selects all children of the current node
descendant	Selects all descendants (children, grandchildren, etc.) of the current node
descendant-or-self	Selects all descendants (children, grandchildren, etc.) of the current node and the current node itself
following	Selects everything in the document after the closing tag of the current node
following-sibling	Selects all siblings after the current node
namespace	Selects all namespace nodes of the current node
parent	Selects the parent of the current node
preceding	Selects all nodes that appear before the current node in the document except ancestors, attribute nodes, and namespace nodes
preceding-sibling	Selects all siblings before the current node
self	Selects the current node

XPATH – PRÉDICATS

Un prédicat est une fonction qui s'applique au nom, à la valeur ou aux attributs d'un nœud et qui produit une réponse logique *VRAI (TRUE)* ou *FAUX (FALSE)*.

Les prédicats modifient le chemin d'entrée d'une interrogation XPath. Les nœuds pour lesquels la relation s'avère exacte sont sélectionnés par l'interrogation.

Les prédicats sont présentés entre crochets, et suivent un nœud.

Exemples :

- « `//p[position()=1]` » permet d'extraire le premier nœud `<p>` par rapport à son nœud parent `<p>`.
- « `//p[last()]` » permet d'extraire le dernier nœud `<p>` par rapport à son nœud parent `<p>`.
- « `//div[count(./@*)>2]` » permet d'extraire tous les nœuds `<div>` avec deux attributs ou plus, etc.

Function	Description	Example
<code>name(<node>)</code>	Returns the name of <node> or the first node in a node set	<code>//*[name()='title'];</code> Returns: <title>
<code>text(<node>)</code>	Returns the value of <node> or the first node in a node set	<code>//*[text()='The book homepage'];</code> Returns: <i> with value <i>The book homepage</i>
<code>@attribute</code>	Returns the value of a node's <i>attribute</i> or of the first node in a node set	<code>//div[@id='R Inventor'];</code> Returns: <div> with attribute <i>id</i> value <i>R Inventor</i>
<code>string-length(str1)</code>	Returns the length of <code>str1</code> . If there is no string argument, it returns the length of the string value of the current node	<code>//h1[string-length()>11];</code> Returns: <h1> with value <i>Robert Gentleman</i>
<code>translate(str1, str2, str3)</code>	Converts <code>str1</code> by replacing the characters in <code>str2</code> with the characters in <code>str3</code>	<code>//div[translate(./@date, '2003', '2005')='June/2005'];</code> Returns: first <div> node with date attribute value <i>June/2003</i>
<code>contains(str1, str2)</code>	Returns TRUE if <code>str1</code> contains <code>str2</code> , otherwise FALSE	<code>//div[contains(@id, 'Inventor')];</code> Returns: first <div> node with id attribute value <i>R Inventor</i>
<code>starts-with(str1, str2)</code>	Returns TRUE if <code>str1</code> starts with <code>str2</code> , otherwise FALSE	<code>//i[starts-with(text(), 'The')];</code> Returns: <i> with value <i>The book homepage</i>
<code>substring-before(str1, str2)</code>	Returns the start of <code>str1</code> before <code>str2</code> occurs in it	<code>//div[substring-before(@date, '/')='June'];</code> Returns: <div> with date attribute value <i>June/2003</i>
<code>substring-after(str1, str2)</code>	Returns the remainder of <code>str1</code> after <code>str2</code> occurs in it	<code>//div[substring-after(@date, '/')=2003];</code> Returns: <div> with date attribute value <i>June/2003</i>
<code>not(arg)</code>	Returns TRUE if the boolean value is FALSE, and FALSE if the boolean value is TRUE	<code>//div[not(contains(@id, 'Inventor'))];</code> Returns: the <div> node that does not contain the string <i>Inventor</i> in its id attribute value
<code>local-name(<node>)</code>	Returns the name of the current <node> or the first node in a node set—without the namespace prefix	<code>//*[local-name()='address'];</code> Returns: <address>
<code>count(<node>)</code>	Returns the count of a nodeset <node>	<code>//div[count(./a)=0];</code> Result: The second <div> with one <a> child
<code>position(<node>)</code>	Returns the index position of <node> that is currently being processed	<code>//div/p[position()=1];</code> Result: The first <p> node in each <div> node
<code>last()</code>	Returns the number of items in the processed node list <node>	<code>//div/p[last()];</code> Result: The last <p> node in each <div> node

COMMUNIQUÉS DE PRESSE DU GOUVERNEMENT DU ROYAUME-UNI – CONTEXTE

Le gouvernement du Royaume-Uni publie tous ses communiqués de presse en ligne, à l'adresse gov.uk/government/announcements.

Le 29 mars 2018, on dénombrait plus de 65 000 communiqués de presse sur le site.

Questions :

- Pouvons-nous prédire quel organisme a fait une annonce en se fiant uniquement au contenu textuel de cette dernière?
- Y a-t-il des thèmes qui semblent sans cesse revenir à l'avant-plan?



Announcements

You can use the filters to show only results that match your interests

Contains

Announcement type

All announcement types

Policy area

All policy areas

Department

All departments

Person

All people

World locations

All locations

65,716 announcements

Get updates to this list  [email](#) [feed](#)

Welsh innovation is key to Britain's future export success

24 March 2018 WO Speech

Preventing Hunger as a Weapon of War

23 March 2018 FCO Speech

"Our vote today against this resolution is a vote against the politicization of the Commission on the Status of Women."

23 March 2018 FCO Speech

Lord Ahmad welcomes conclusions of the 37th Session of the UN Human Rights Council

23 March 2018 FCO Speech

Rt Hon Mark Field MP speech at Global FinTech Investor Forum

23 March 2018 FCO Speech

Foreign Secretary statement on Iran

The Foreign Secretary has made the following statement on the protests in Iran.

Published 1 January 2018

From: [Foreign & Commonwealth Office](#) and [The Rt Hon Boris Johnson MP](#)



The Foreign Secretary Boris Johnson said:

“ The UK is watching events in Iran closely. We believe that there should be meaningful debate about the legitimate and important issues the protesters are raising and we look to the Iranian authorities to permit this.”

“ We also believe that, particularly as we enter the 70th anniversary year of the Universal Declaration on Human Rights, people should be able to have freedom of expression and to demonstrate peacefully within the law.”

“ We regret the loss of life that has occurred in the protests in Iran, and call on all concerned to refrain from violence and for international obligations on human rights to be observed.”

Chaque communiqué de presse contient ce qui suit :

- titre
- date de publication
- organismes/personnes l'ayant publié
- texte du communiqué de presse

Les communiqués de presse portent principalement sur 2017 et proviennent des bureaux ou organismes suivants :

- Bureau du pays de Galles
- Ministère des Affaires étrangères
- Ministère des Sciences et de la Technologie
- Ministère de l'Environnement, de l'Alimentation et des Affaires rurales

Bloc-notes : communiqués de presse du gouvernement du Royaume-Uni

EXPRESSIONS RÉGULIÈRES

L'objectif principal du moissonnage du Web consiste à recueillir de l'information **utile** pour le problème faisant l'objet de travaux de recherche, à partir d'une quantité considérable de données textuelles.

Nous nous intéressons aux éléments systématiques des données textuelles, tout particulièrement si des méthodes quantitatives seront éventuellement appliquées.

Les structures systématiques peuvent prendre les formes suivantes :

- nombres
- noms (pays, etc.)
- adresses (postales, courriel, URL, etc.)
- chaînes de caractères spécifiques, etc.

EXPRESSIONS RÉGULIÈRES

Les expressions régulières (« regexps ») permettent l'extraction systématique des composantes d'information.

Les **expressions régulières** sont des séquences abstraites de chaînes qui correspondent à des modèles concrets récurrents qui se retrouvent dans le texte.

Elles peuvent servir à extraire de l'information de fichiers en texte brut, voire de type HTML et XML.

Utiles lorsque l'information est dissimulée à l'intérieur de valeurs *atomiques*.

Bloc-notes : expressions régulières en Python et plus encore

BEAUTIFUL SOUP

Pour qu'elles permettent d'extraire une page et le contenu HTML, les demandes Web élémentaires doivent être assorties d'instructions réseau.

Les navigateurs accomplissent énormément de travail pour analyser de manière intelligente une syntaxe HTML absolument inappropriée, comme dans le cas suivant :

```
<a href="crummy.com"> <b>link text<a> </b>
```

Beautiful Soup est une bibliothèque Python qui facilite l'extraction de données de fichiers HTML et XML. Cette bibliothèque analyse les fichiers HTML, même s'ils sont brisés.

BEAUTIFUL SOUP

Beautiful Soup (BS) ne se limite pas à convertir des instructions HTML qui laissent à désirer en instructions XHTML, de sorte que celles-ci puissent être analysées au moyen d'un logiciel d'analyse XML.

BS permet à un utilisateur d'inspecter intégralement la structure HTML (appropriée) qu'elle produit, grâce à un programme.

Une fois que BS a terminé son travail portant sur un fichier HTML, il en résulte une API qui permet de soumettre les éléments du document à un survol, à une recherche et à une lecture.

BEAUTIFUL SOUP

Les éléments HTML qui sont généralement extraits/lus se présentent selon divers formats :

- texte
- tableaux
- valeurs de champs
- images
- vidéos
- etc.

Beautiful Soup offre des façons **idiomatiques** de soumettre l'arborescence d'analyse du fichier HTML à des opérations de navigation, de recherche et de modification (ce qui fait gagner énormément de temps).

```
html_doc = """
<html><head><title>The Dormouse's story</title></head>
<body>
<p class="title"><b>The Dormouse's story</b></p>

<p class="story">Once upon a time there were three little sisters; and their names were
<a href="http://example.com/elsie" class="sister" id="link1">Elsie</a>,
<a href="http://example.com/lacie" class="sister" id="link2">Lacie</a> and
<a href="http://example.com/tillie" class="sister" id="link3">Tillie</a>;
and they lived at the bottom of a well.</p>

<p class="story">...</p>
"""
```

```
from bs4 import BeautifulSoup
soup = BeautifulSoup(html_doc, 'html.parser')

print(soup.prettify())
# <html>
# <head>
# <title>
#   The Dormouse's story
# </title>
# </head>
# <body>
# <p class="title">
#   <b>
#     The Dormouse's story
#   </b>
# </p>
# <p class="story">
#   Once upon a time there were three little sisters; and their names were
#   <a class="sister" href="http://example.com/elsie" id="link1">
#     Elsie
#   </a>
#   ,
#   <a class="sister" href="http://example.com/lacie" id="link2">
#     Lacie
#   </a>
#   and
#   <a class="sister" href="http://example.com/tillie" id="link2">
#     Tillie
#   </a>
#   ; and they lived at the bottom of a well.
# </p>
# <p class="story">
#   ...
# </p>
# </body>
# </html>
```



```
soup.title
# <title>The Dormouse's story</title>

soup.title.name
# u'title'

soup.title.string
# u'The Dormouse's story'

soup.title.parent.name
# u'head'

soup.p
# <p class="title"><b>The Dormouse's story</b></p>

soup.p[ 'class' ]
# u'title'

soup.a
# <a class="sister" href="http://example.com/elsie" id="link1">Elsie</a>

soup.find_all('a')
# [<a class="sister" href="http://example.com/elsie" id="link1">Elsie</a>,
#  <a class="sister" href="http://example.com/lacie" id="link2">Lacie</a>,
#  <a class="sister" href="http://example.com/tillie" id="link3">Tillie</a>]

soup.find(id="link3")
# <a class="sister" href="http://example.com/tillie" id="link3">Tillie</a>
```

```
for link in soup.find_all('a'):
    print(link.get('href'))
# http://example.com/elsie
# http://example.com/lacie
# http://example.com/tillie
```

```
print(soup.get_text())
# The Dormouse's story
#
# The Dormouse's story
#
# Once upon a time there were three little sisters; and their names were
# Elsie,
# Lacie and
# Tillie;
# and they lived at the bottom of a well.
#
# ...
```

SELENIUM

Selenium est un outil qui permet d'automatiser les interactions avec les navigateurs Web (en Python). S'il sert généralement à automatiser les applications Web à des fins de mise à l'épreuve, il offre également d'autres applications.

Il permet principalement à un utilisateur d'ouvrir un navigateur et d'effectuer des tâches analogues à celles qu'exécuterait un humain, comme les suivantes :

- cliquer sur un bouton
- introduire des données dans un formulaire
- rechercher de l'information particulière dans des pages Web
- etc.

SELENIUM

Selenium a besoin d'un pilote pour établir une interface avec le navigateur choisi. À titre d'exemple, Firefox a besoin de *geckodriver*.

Les autres navigateurs pris en charge disposeront de leurs propres pilotes :

Chrome : <https://sites.google.com/a/chromium.org/chromedriver/downloads>

Edge : <https://developer.microsoft.com/en-us/microsoft-edge/tools/webdriver/>

Firefox : <https://github.com/mozilla/geckodriver/releases>

Safari : <https://webkit.org/blog/6900/webdriver-support-in-safari-10/>

SIMULATION D'UN NAVIGATEUR WEB

Selenium contrôle automatiquement l'ensemble du navigateur, y compris en ce qui concerne le rendu des documents Web et l'exécution de JavaScript.

Ceci est utile pour les pages qui intègrent un fort contenu dynamique qui ne se retrouve pas dans le fichier HTML de base.

Selenium peut programmer des actions telles que « cliquer sur ce bouton » ou « taper ce texte », et vous avez en tout temps accès au fichier HTML dynamique qui correspond à l'état actuel de la page, comme ce que l'on retrouve dans les outils de développement.

UTILISATION DES API

Une API constitue la façon, pour un site Web, d'offrir l'accès à ses données à un programme, sans qu'il soit nécessaire d'en récupérer le contenu.

Ainsi donc, une API offre un **accès structuré** à des **données structurées**.

À titre d'exemple, un site à caractère financier pourrait offrir une API assortie de données financières, tout comme le *New York Times* pourrait offrir une API adaptée aux articles de nouvelles.

Dans l'un ou l'autre des cas, les données se présenteraient dans un format prédéfini, structuré (lequel est fréquemment JSON).

UTILISATION DES API

Les API que nous examinerons intègrent des bibliothèques R/Python qui prennent en charge l'ensemble des opérations de réseau et de codage requises.

Cela signifie qu'il suffit de lire la documentation relative à la bibliothèque pour savoir quoi faire.

Exercice : servez-vous de Zomato pour trouver quelle ville canadienne propose les meilleurs restaurants de sushi (<https://github.com/fatihsucu/pyzomato>).

API DE YOUTUBE – KHAN ACADEMY

Des millions de vidéos sont accessibles par **YouTube**.

Il n'est pas évident de déterminer comment l'on s'y prendrait pour extraire, de manière générale, du contenu vidéo du Web (si ce n'est par l'entremise des URL); à certaines vidéos correspond du contenu sous forme de texte (**transcriptions**).

Nous avons recours à l'API de YouTube pour extraire ce contenu.

Bloc-notes : transcriptions YouTube



Home



Trending



History

BEST OF YOUTUBE



Music



Sports



Gaming



Movies



TV Shows



News



Live



360° Video



Browse channels

Sign in now to see your channels and recommendations!

SIGN IN



Statistics

68 videos • 3,290,303 views • Last updated on Jul 2, 2014



Khan Academy

SUBSCRIBE

Introduction to statistics. Will eventually cover all of the major topics in a first-year statistics course (not there yet!)

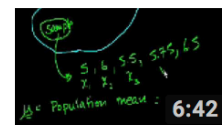
1



Statistics: The average | Descriptive statistics | Probability and Statistics | Khan Academy

Khan Academy

2



Statistics: Sample vs. Population Mean

Khan Academy

3



Statistics: Variance of a population | Probability and Statistics | Khan Academy

Khan Academy

4



Statistics: Sample variance | Descriptive statistics | Probability and Statistics | Khan Academy

Khan Academy

5



Statistics: Standard deviation | Descriptive statistics | Probability and Statistics | Khan Academy

Khan Academy

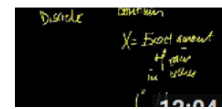
6



Statistics: Alternate variance formulas | Probability and Statistics | Khan Academy

Khan Academy

7



Introduction to Random Variables

Khan Academy

PRÉPARATION PRÉALABLE DES DONNÉES

COLLECTE ET TRAITEMENT DES DONNÉES

OBJECTIFS D'APPRENTISSAGE

Se familiariser avec le format des données ordonnées

Mieux connaître les fonctions de préparation des données

Repérer les paquets R qui facilitent le traitement des données

PRÉPARATION PRÉALABLE DES DONNÉES

Un temps considérable (jusqu'à 80 % peut-être) doit être consacré au traitement des données (nettoyage et manipulation).

Les principaux objectifs de la **préparation préalable des données** sont les suivants :

- rendre les données utilisables par un logiciel particulier
- déceler des informations d'analyse préliminaire dans les données

DONNÉES ORDONNÉES (« TIDY DATA »)

Les données ordonnées ont une structure particulière :

- chaque variable est une colonne
- chaque observation est une rangée
- chaque type d'unité d'observation est un tableau

Country	2011	2012	2013
FR	7000	6900	7000
DE	5800	6000	6200
US	15000	14000	13000

et

Country	Year	n
FR	2011	7000
DE	2011	5800
US	2011	15000
FR	2012	6900
DE	2012	6000
US	2012	14000
FR	2013	7000
DE	2013	6200
US	2013	13000

FONCTIONNALITÉS

Les fonctions de préparation préalable des données devraient permettre à l'analyste de faire ce qui suit :

- extraire un sous-ensemble de variables de la trame de données
- extraire un sous-ensemble d'observations de la trame de données
- trier la trame de données en fonction de n'importe quelle combinaison de variables par ordre croissant ou décroissant
- créer de nouvelles variables à partir de variables existantes
- créer des tableaux croisés dynamiques, par groupes d'observation
- créer des fonctionnalités de base de données (agrégations, etc.)
- etc.

FONCTIONNALITÉS

Dans R, cela peut se faire de plusieurs façons. Les paquets préférés actuels incluent :

- `tidyr`
- `dplyr` (transformation de données)
- `lubridate` (dates et heures)
- `stringr` (manipulation de chaînes)
- `purrr` (fonctions)
- `readr` (importation de données)

Pour les modules Python équivalents, consultez *Data Wrangling with Python* de Kazil et Jarmul.

EXERCICES

À quoi ressemblerait l'ensemble de données suivant dans un format ordonné?

tempête	statistique	valeur
Alex	vent	68
Alex	pression	130
Allison	vent	55
Allison	pression	121
Bobbie	vent	72
Bobbie	pression	118

EXERCICES

Comment iriez-vous du tableau de gauche au tableau de droite?

tempête	statistique	valeur
Alex	vent	68
Alex	pression	130
Allison	vent	55
Allison	pression	121
Bobbie	vent	72
Bobbie	pression	118

statistique	moyenne	écart-type
vent	65	8,9
pression	123	6,2

EXERCICES

Exécutez la section 9 du cahier de l'ÉFPC 04 `R Basics.ipynb` pour découvrir comment les paquets `tidyr` et `dplyr` facilitent la préparation préalable des données dans R.

EXERCICES

Transformez les données trouvées dans `cities.txt` en un ensemble de données ordonnées.

NETTOYAGE DES DONNÉES

COLLECTE ET TRAITEMENT DES DONNÉES

« De toute évidence, la meilleure façon de traiter les données manquantes consiste à n'en avoir aucune. »

T. Orchard et M. Woodbury

« L'expression la plus excitante à entendre, celle qui annonce la plupart des découvertes, n'est pas "Eurêka", mais bien "C'est drôle..." »

OBJECTIFS D'APPRENTISSAGE

Reconnaître les avantages et les inconvénients des deux grandes approches du nettoyage des données

Cerner les méthodes de traitement des observations manquantes

Mieux connaître les différents tests de détection des anomalies et des valeurs aberrantes

QUATRE REMARQUES TRÈS IMPORTANTES

Ne travaillez **JAMAIS** sur l'ensemble de données d'origine. Faites des copies en cours de route.

Consignez **TOUTES** vos étapes et procédures de nettoyage.

Si vous constatez que vous nettoyez trop de vos données, **ARRÊTEZ** votre travail. Il y a peut-être quelque chose qui cloche dans la procédure de collecte des données.

Pensez-y à **DEUX FOIS** avant de supprimer un enregistrement complet.

APPROCHES DU NETTOYAGE DES DONNÉES

Il existe deux approches **philosophiques** du nettoyage et de la validation des données :

- l'approche méthodique
- l'approche descriptive

L'approche **méthodique** consiste à passer en revue une **liste de contrôle** des problèmes possibles et à signaler ceux qui se rapportent aux données.

L'approche **descriptive** consiste à **explorer** l'ensemble de données et à tenter de dégager les schémas improbables et irréguliers.

POINTS À RETENIR

L'approche descriptive s'apparente au fait de remplir une grille de mots croisés avec un stylo et à y inscrire de temps en temps des réponses potentiellement mauvaises, puis voir où cela mène.

L'approche mécanique s'apparente au fait de remplir la grille à l'aide d'un crayon et d'un dictionnaire et à ne jamais inscrire de réponse sans être convaincu qu'elle est exacte.

Vous remplirez plus de grilles (et de façon plus éclatante) avec la première approche, mais avec la seconde approche, vous aurez rarement tort.

Soyez à l'aise avec les deux approches.

TYPES D'OBSERVATIONS MANQUANTES

Il existe quatre catégories de champs vides :

- **Absence de réponse**
On s'attendait à une observation, mais aucune n'a été saisie.
- **Problème dans la saisie des données**
On a consigné une observation, mais celle-ci n'a pas été saisie dans l'ensemble de données.
- **Entrée invalide**
On a consigné une observation, mais elle a été considérée comme étant invalide et elle a été retirée.
- **Vide attendu**
On a laissé un champ vide, comme prévu.

TYPES D'OBSERVATIONS MANQUANTES

Un nombre trop élevé de valeurs manquantes (des trois premiers types) peut révéler l'existence de **lacunes dans le processus de collecte des données** (on y reviendra plus tard).

Un nombre trop élevé de valeurs manquantes (du quatrième type) peut révéler l'existence d'un **questionnaire mal conçu**.

BIEN-FONDÉ DE L'IMPUTATION

Les méthodes analytiques ne peuvent pas toutes tenir compte aisément des observations manquantes.

Il existe deux options :

- **Rejeter** l'observation manquante
 - Cette option n'est pas recommandée, sauf si les données manquantes sont totalement aléatoires dans l'ensemble de données en général.
 - Elle est acceptable dans certaines situations (comme un petit nombre de valeurs manquantes dans un grand ensemble de données).
- Trouver une **valeur de remplacement**
 - Cette option comporte un inconvénient important : nous ne connaissons jamais avec certitude la véritable valeur.
 - Il s'agit souvent de la meilleure option disponible.

MÉCANISMES DE DONNÉES MANQUANTES

Données manquantes complètement aléatoires (Missing Completely at Random ou MCAR)

- L'absence de l'élément ne dépend pas de sa valeur ou des variables auxiliaires.

Données manquantes simplement aléatoires (Missing at Random ou MAR)

- L'absence de l'élément n'est pas entièrement aléatoire; elle peut être attribuée aux variables auxiliaires assorties de renseignements complets.

Données manquantes non aléatoires (Not Missing at Random ou NMAR)

- La raison de la non-réponse se rapporte à la valeur de l'élément (appelée aussi « **non-réponse non ignorable** »).

MÉTHODES D'IMPUTATION

Suppression à partir d'une liste

Imputation par la moyenne ou la plus fréquente

Imputation par régression ou corrélation

Imputation par régression stochastique

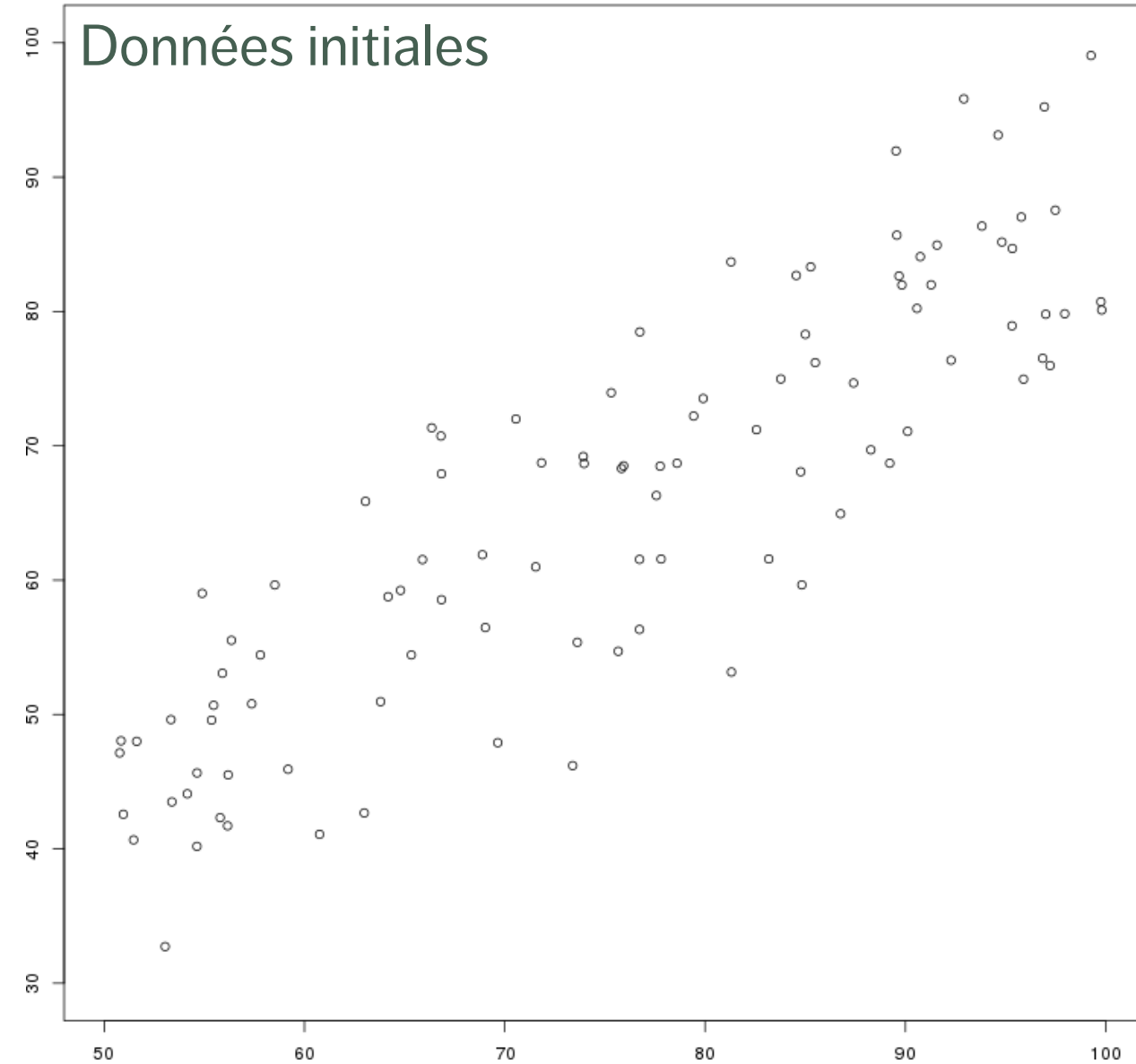
Report en avant de la dernière observation

Imputation par la méthode du plus proche voisin k

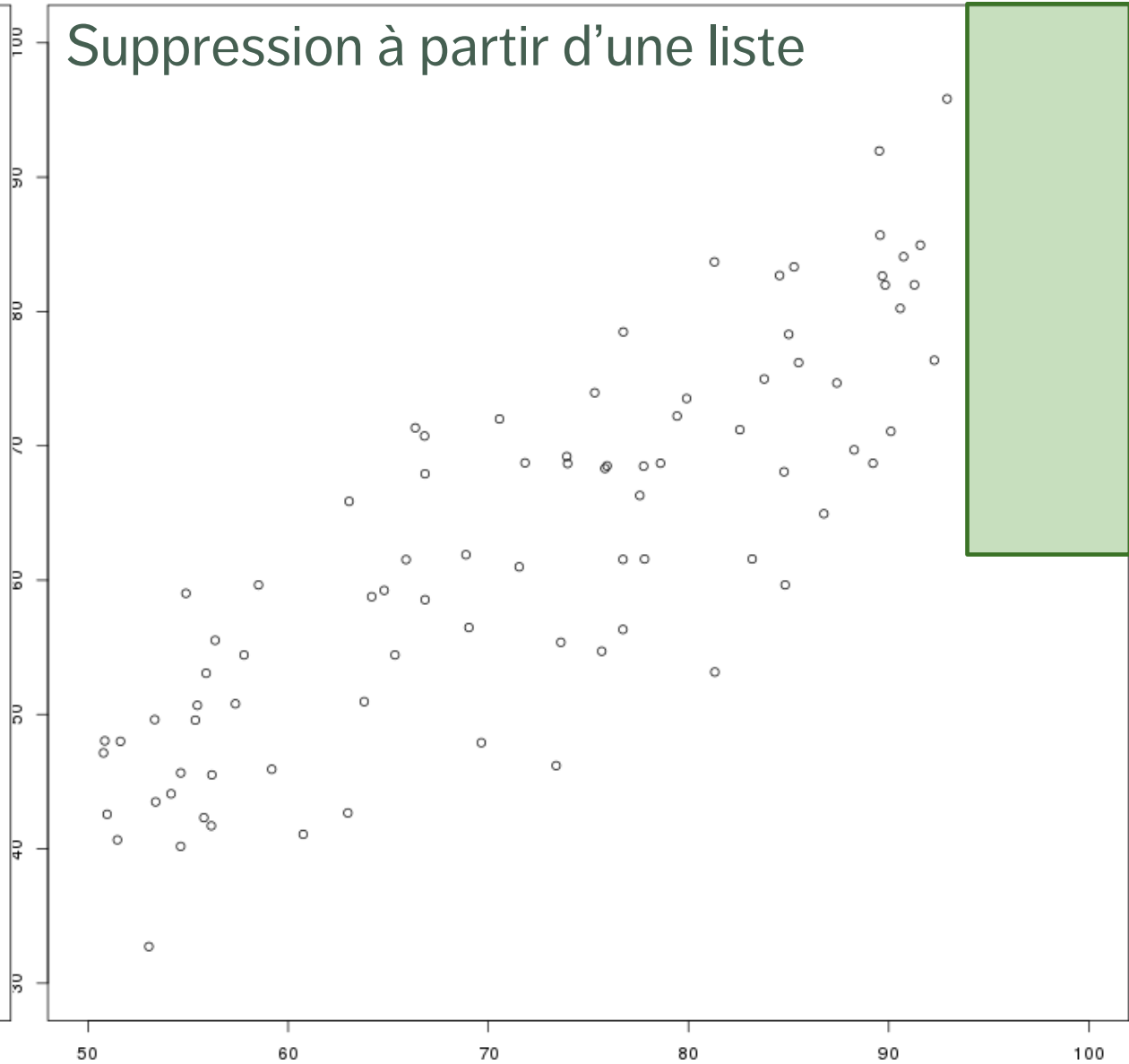
Imputation multiple

Données artificielles : Les valeurs y de tous les points pour lesquels $x > 92$ ont été effacées par erreur.

Données initiales

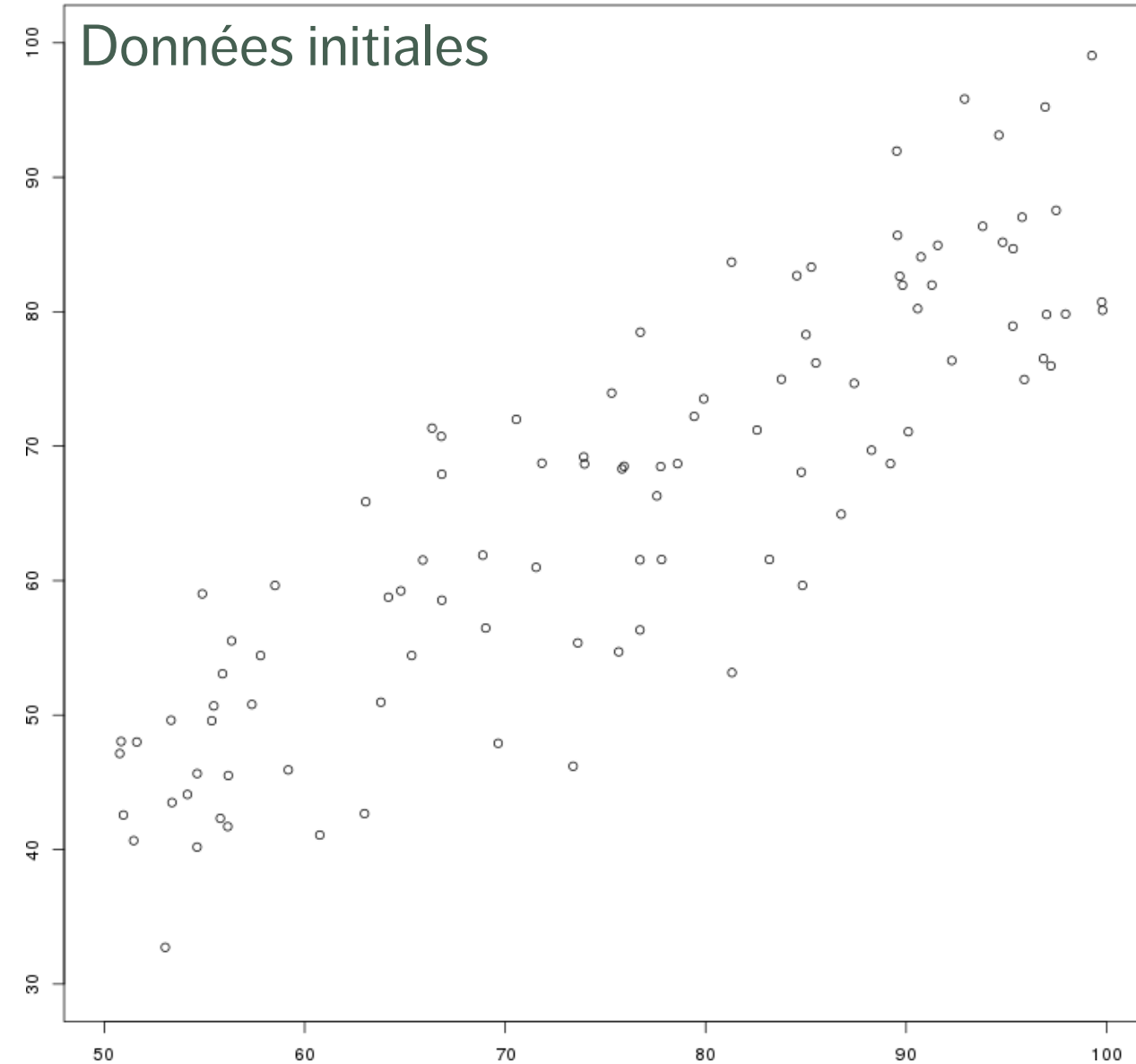


Suppression à partir d'une liste

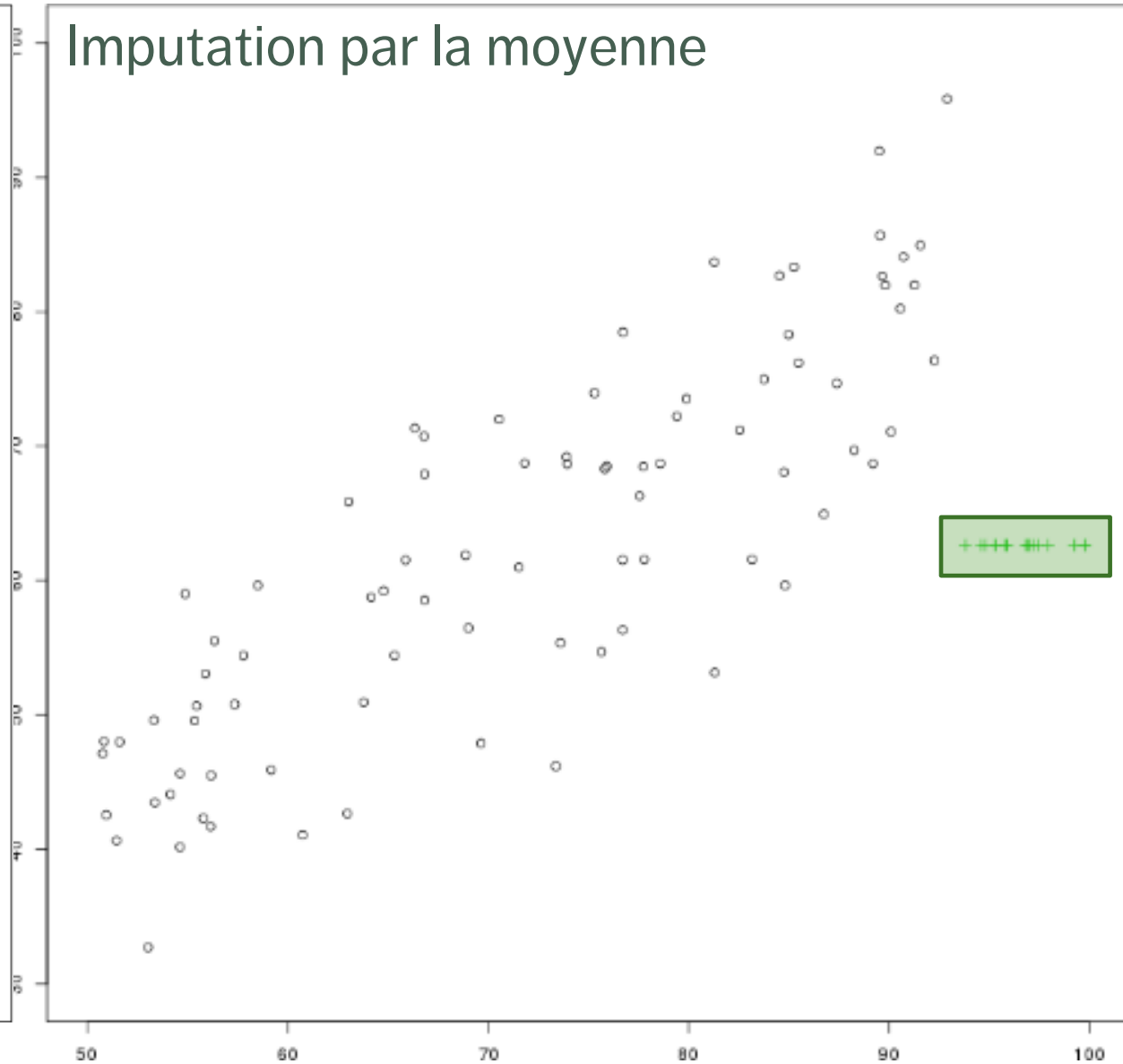


Données artificielles : Les valeurs y de tous les points pour lesquels $x > 92$ ont été effacées par erreur.

Données initiales

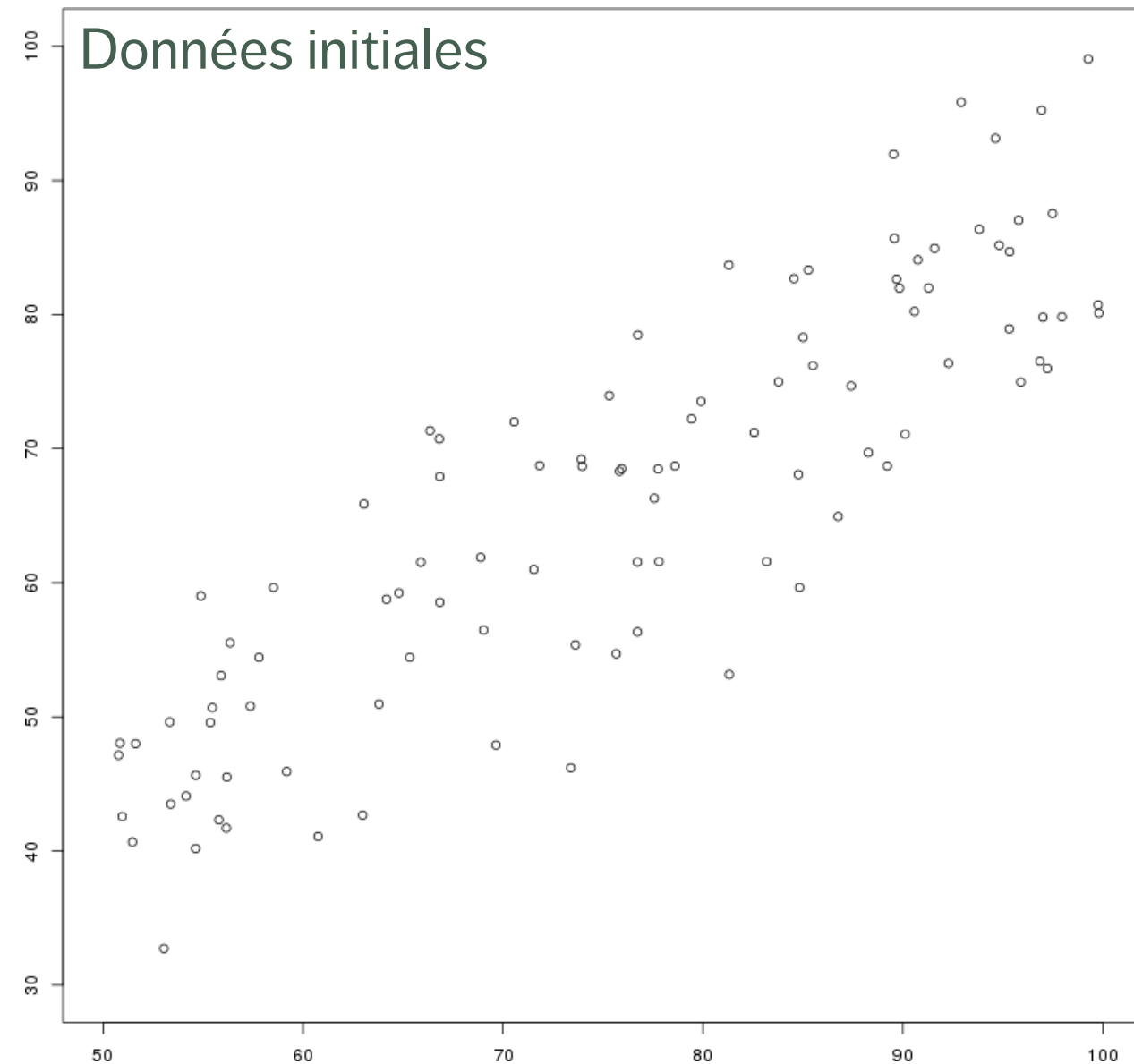


Imputation par la moyenne

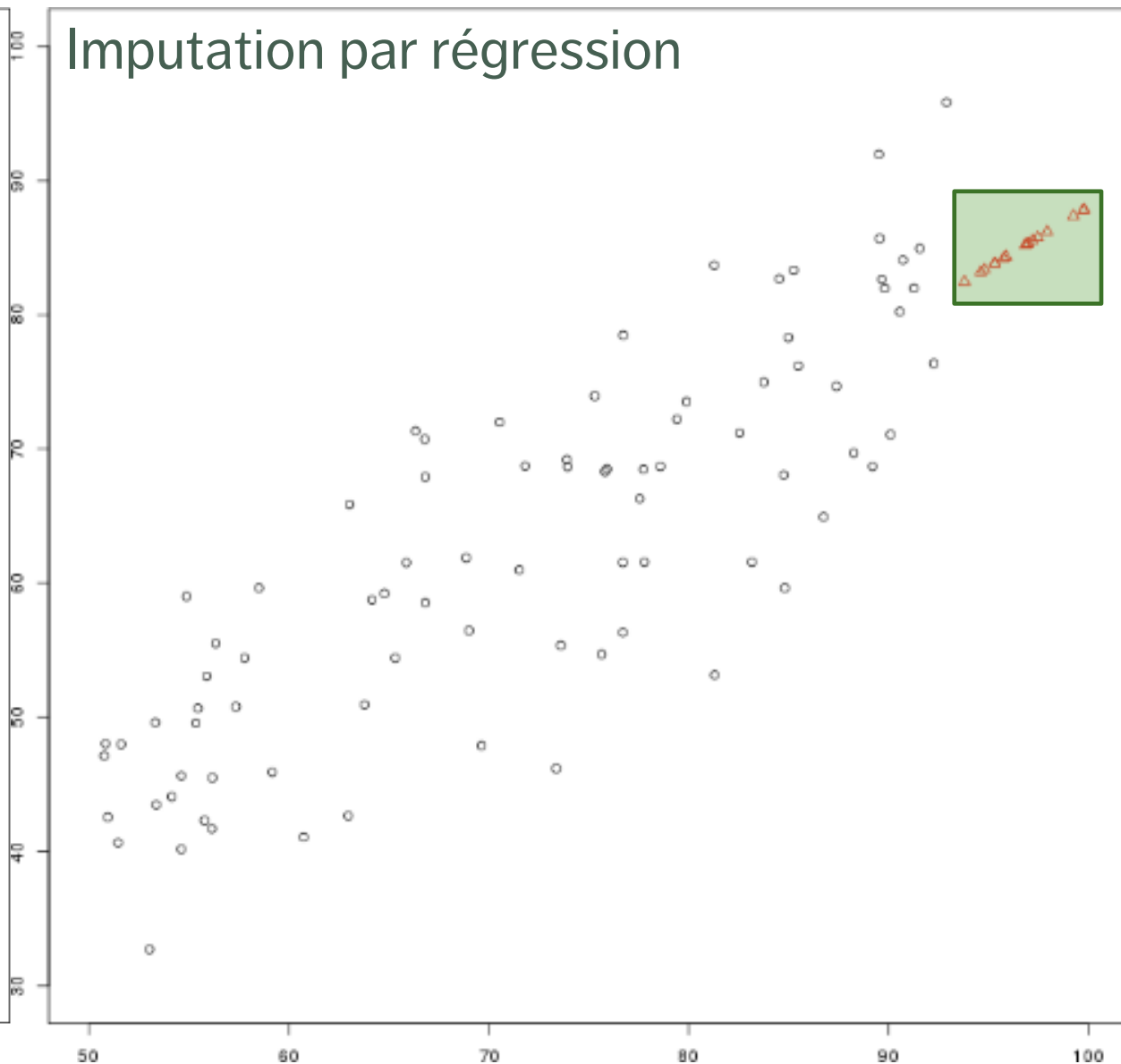


Données artificielles : Les valeurs y de tous les points pour lesquels $x > 92$ ont été effacées par erreur.

Données initiales

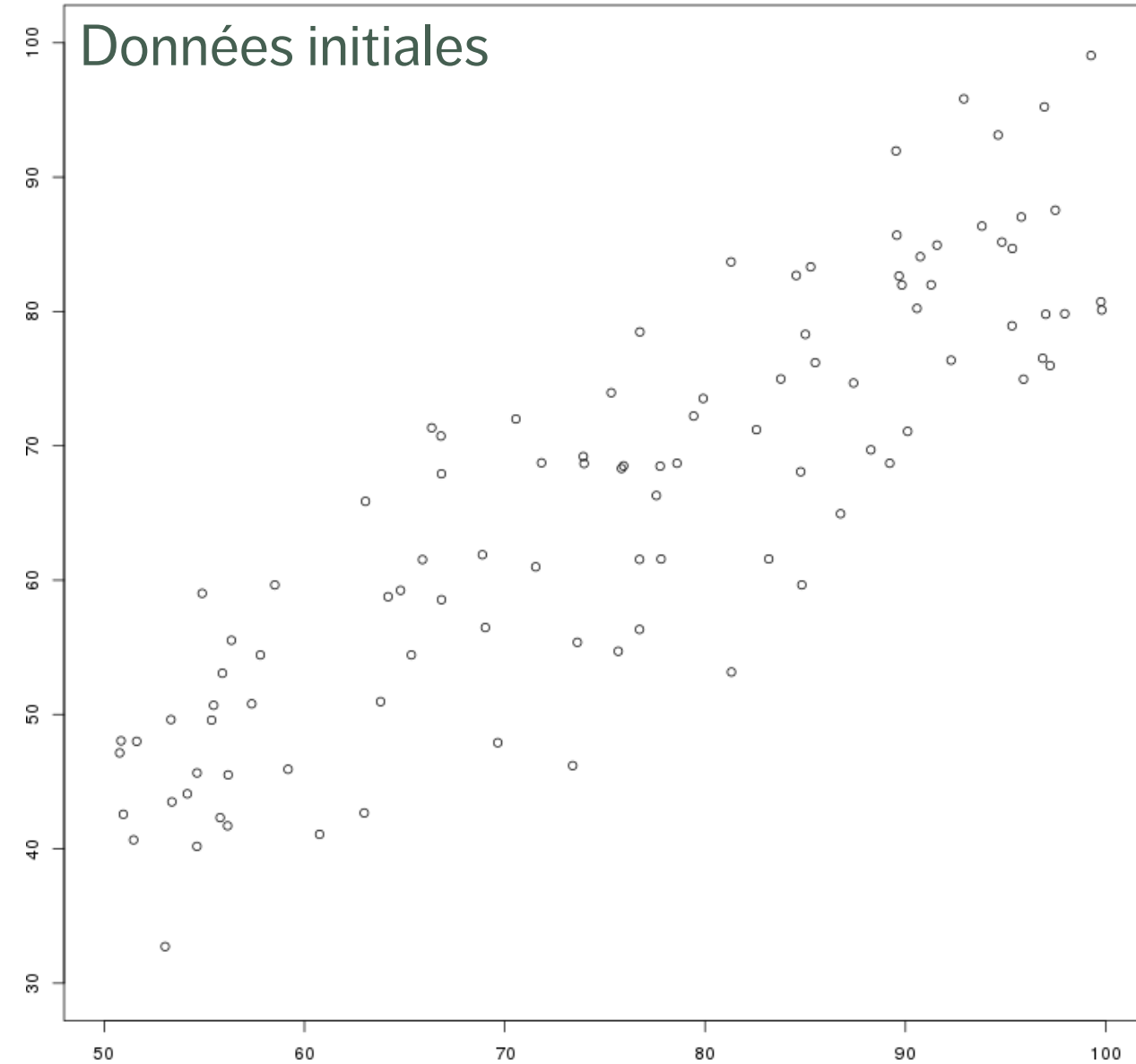


Imputation par régression

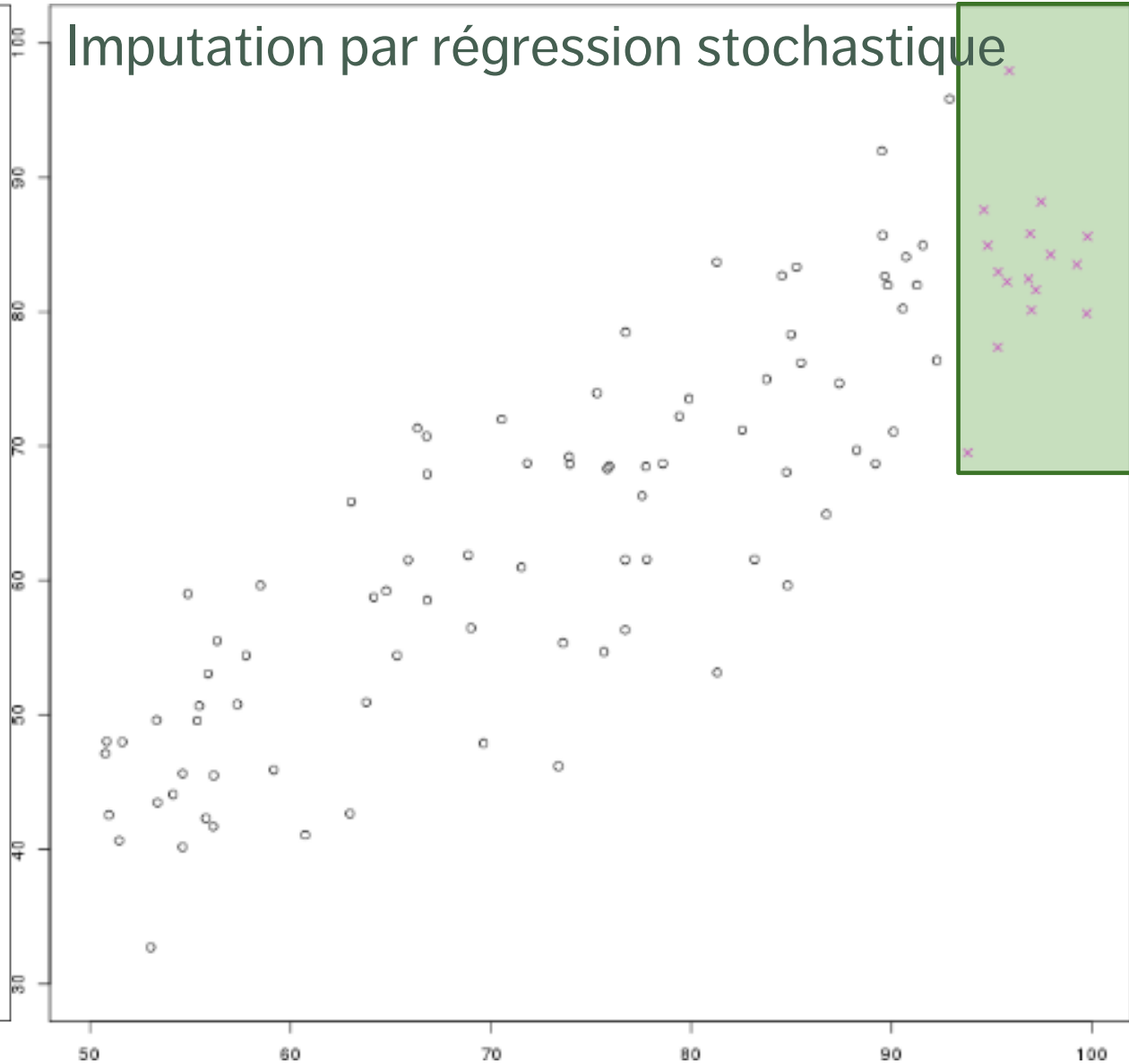


Données artificielles : Les valeurs y de tous les points pour lesquels $x > 92$ ont été effacées par erreur.

Données initiales



Imputation par régression stochastique



IMPUTATION MULTIPLE

Les imputations accentuent le bruit dans les données.

Lors d'une **imputation multiple**, on peut mesurer l'effet du bruit en consolidant le résultat de l'analyse à partir de multiples ensembles de données imputés.

Étapes :

1. Une imputation répétée crée m versions de l'ensemble de données.
2. Chaque ensemble de données est analysé et produit m résultats.
3. Les m résultats sont regroupés en un seul résultat pour lequel la moyenne, la variance et les intervalles de confiance sont connus.

IMPUTATION MULTIPLE

Avantages

- Elle est **souple**. Elle peut servir dans diverses situations (MCAR, MAR et même NMAR dans certains cas).
- Elle tient compte de l'**incertitude** liée aux valeurs imputées.
- Elle est assez facile à mettre en œuvre.

Inconvénients

- m La valeur m doit être assez **grande** quand il y a de nombreuses valeurs manquantes pour une multitude de caractéristiques, ce qui ralentit les analyses.
- Qu'advient-il si le résultat de l'analyse n'est pas une valeur unique, mais un quelconque objet mathématique plus complexe?

POINTS À RETENIR

On ne peut tout simplement pas ignorer les valeurs manquantes.

On ne peut habituellement pas établir le mécanisme de données manquantes avec certitude.

Les méthodes d'imputation sont plus efficaces avec des données manquantes complètement aléatoires ou simplement aléatoires, mais elles tendent à produire des estimations présentant un biais.

Lors d'une imputation unique, les données imputées sont traitées comme des données réelles; l'imputation multiple peut aider à réduire le bruit.

L'imputation stochastique est-elle la meilleure? C'est le cas dans notre exemple, mais méfiez-vous du théorème « *No Free Lunch* » (on n'a rien pour rien)!

OBSERVATIONS SPÉCIALES

Les **observations aberrantes** sont des points de données qui sont **atypiques** par rapport :

- aux caractéristiques restantes de l'unité (*à l'intérieur de l'unité*)
- aux mesures sur le terrain pour d'autres unités (*entre les unités*)

ou en tant qu'éléments d'un sous-ensemble collectif d'observations

Les valeurs aberrantes sont des observations qui sont **différentes des autres cas** ou qui **contredisent les règles ou les liens de dépendance connus**.

Il faut une étude attentive pour déterminer s'il faut conserver ou supprimer les valeurs aberrantes de l'ensemble de données.

OBSERVATIONS SPÉCIALES

Les **points de données influents** sont des observations qui, si elles sont absentes, mènent à des résultats d'analyse **nettement différents**.

La découverte d'observations influentes peut nécessiter la prise de mesures correctives (comme des transformations de données) pour réduire au minimum leurs effets indésirables.

Les valeurs aberrantes peuvent être des points de données influents, mais les points de données influents ne sont pas nécessairement des valeurs aberrantes (données pondérées).

DÉTECTION DES ANOMALIES

Les valeurs aberrantes peuvent s'avérer anormales le long de toute variable de l'unité ou en combinaison.

Les anomalies sont **peu fréquentes** par définition et elles sont habituellement empreintes d'**incertitude** en raison de la petite taille des échantillons.

Il est **difficile** de faire la distinction entre les anomalies et le bruit ou les erreurs de saisie de données.

Les frontières entre les unités normales et les unités déviantes peuvent être **floues**.

Quand les anomalies sont associées à des activités malveillantes, elles sont habituellement **camouflées**.

DÉTECTION DES ANOMALIES

Il existe de nombreux moyens de déceler des observations anormales, mais **aucun n'est infailible**, et il faut savoir exercer son jugement.

Les méthodes graphiques sont faciles à mettre en œuvre et à interpréter.

- **Observations aberrantes**

Diagrammes de quartiles, diagrammes de dispersion, matrices de diagramme de dispersion, visualisation 2D, distance de Cooke, diagrammes Q-Q normaux.

- **Données influentes**

Il faut effectuer un certain niveau d'analyse (levier).

Le retrait des observations anormales de l'ensemble de données peut transformer des unités jusqu'alors « ordinaires » en données aberrantes.

TESTS DE DÉTECTION DES VALEURS ABERRANTES

Les **méthodes supervisées** utilisent un enregistrement historique des observations étiquetées comme étant anormales :

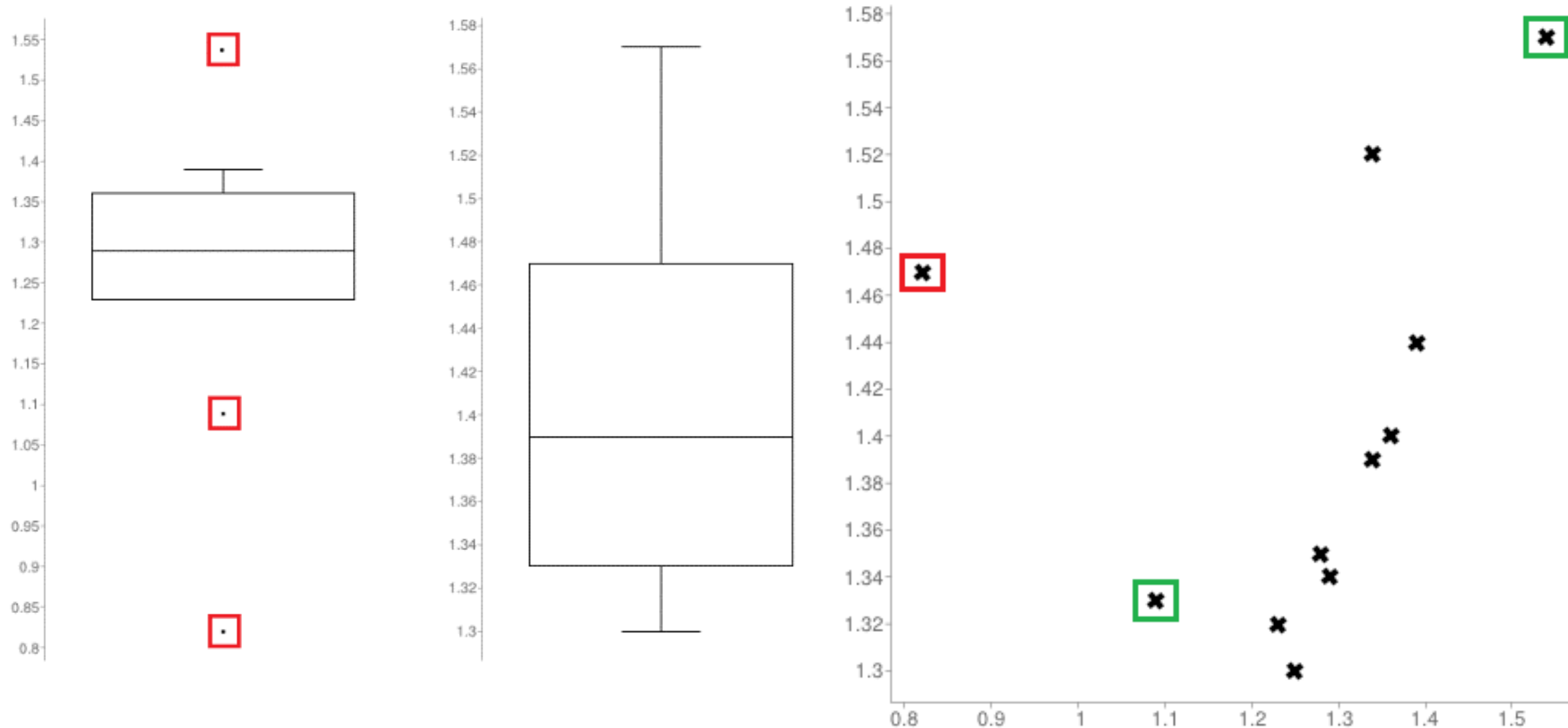
- connaissance du domaine requise pour étiqueter les données
- tâche de classification ou de régression (probabilités et classements des inspections)
- problème d'occurrence rare (plus de détails à venir)

Les **méthodes non supervisées** n'ont pas recours à des renseignements externes :

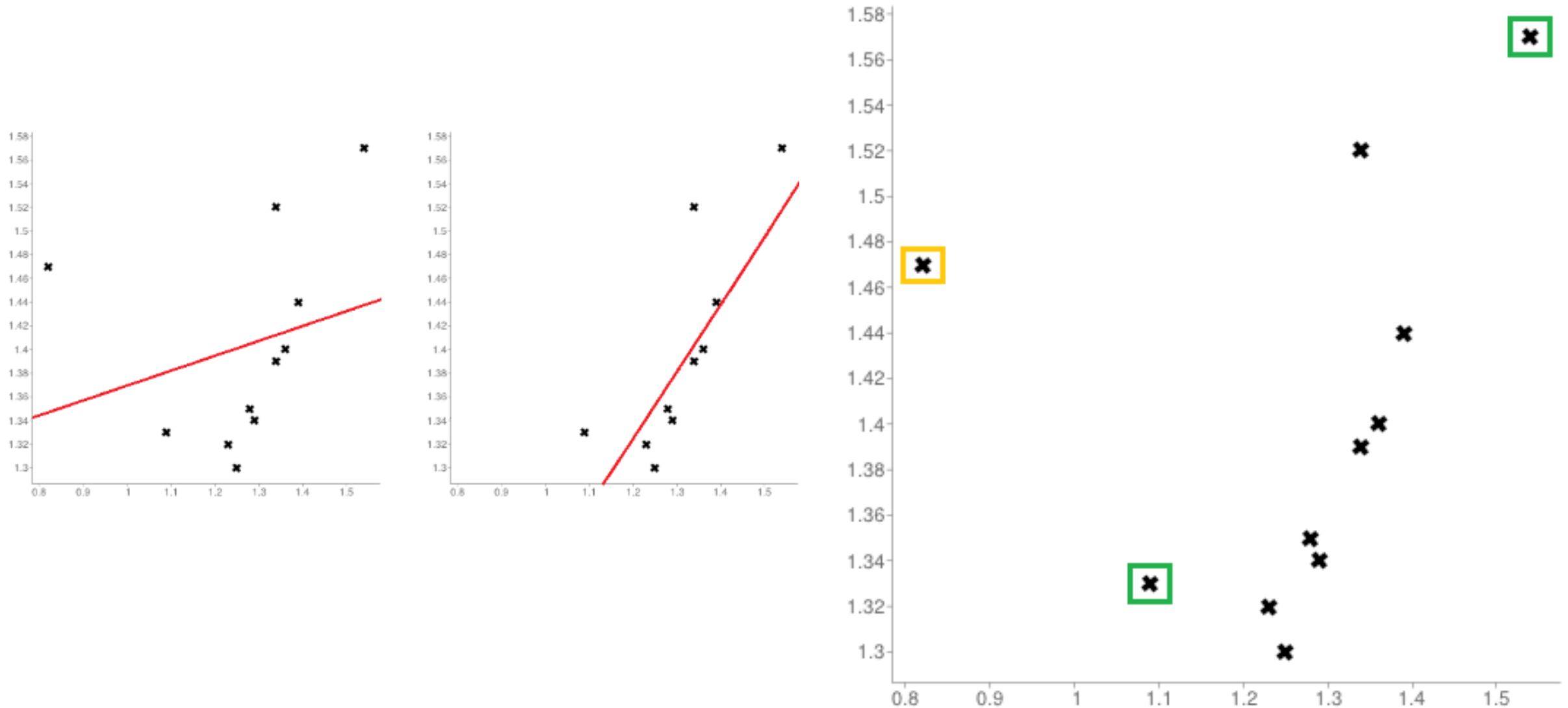
- méthodes et tests classiques
- peuvent aussi être considérées comme un problème lié aux règles de regroupement ou d'association

Il existe aussi des **méthodes semi-supervisées**.

Mise en file d'attente de l'ensemble de données : comparaison entre le taux de traitement et le taux d'arrivée



Mise en file d'attente de l'ensemble de données : comparaison entre le taux de traitement et le taux d'arrivée



POINTS À RETENIR

Le recensement des points influents est un processus itératif, puisqu'il faut effectuer diverses analyses à de nombreuses reprises.

Le recensement et le retrait entièrement automatisés des observations anormales ne sont PAS recommandés.

Il faut recourir à des transformations si les données ne sont PAS distribuées normalement.

Le fait qu'une observation constitue ou non une valeur aberrante dépend de divers facteurs. La détermination des observations constituant des points de données influents dépend de l'analyse à effectuer.

EXERCICES

La capacité de surveiller les diverses proliférations d'algues et de les prévoir tôt est essentielle au contrôle des effets néfastes qu'elles peuvent produire sur l'environnement.

L'ensemble de données `prolifération_algues.csv` sert à améliorer un modèle d'apprentissage qui comprend ce qui suit :

- les **propriétés chimiques** de différents échantillons d'eau provenant de rivières européennes;
- la **quantité de sept algues** dans chacun des échantillons;
- les **caractéristiques du processus de collecte** pour chaque échantillon.

Du point de vue de la science des données, quelle est la raison d'être d'un tel modèle, étant donné que nous **pouvons** effectivement analyser les échantillons d'eau afin de savoir s'il y a présence ou absence d'algues nuisibles?

EXERCICES

La réponse est simple : la surveillance chimique est **peu coûteuse** et **facile à automatiser**, alors que l'analyse biologique d'échantillon est **dispendieuse** et **lente**.

On pourrait aussi répondre que l'analyse des échantillons pour y détecter un contenu nuisible ne permet pas de mieux comprendre les **facteurs** de prolifération des algues; elle nous apprend seulement quels échantillons renferment des algues nuisibles.

Notre modèle permet-il de mieux comprendre la situation des algues?

EXERCICES

Trouvez l'ensemble de données de la prolifération des algues, déterminez sa structure et produisez un résumé de ses caractéristiques.

Calculez le nombre de valeurs manquantes pour chaque enregistrement.

Recensez certaines observations anormales dans le même ensemble de données.

Quelles stratégies pouvez-vous utiliser pour composer avec de telles observations ou de tels enregistrements?

Matériel supplémentaire

AVANTAGES ET INCONVÉNIENTS

Approche **méthodique** (syntaxe)

- Avantages : La liste de contrôle est **indépendante du contexte**. Les pipelines sont **faciles à mettre en œuvre**. Les erreurs communes et les observations invalides sont **faciles à cerner**.
- Inconvénients : Elle peut demander **beaucoup de temps**. Elle ne peut pas recenser de nouveaux types d'erreurs.

Approche **descriptive** (sémantique)

- Avantages : Le processus peut simultanément favoriser la **compréhension des données**. Les faux départs sont (au maximum) aussi coûteux qu'un passage à l'approche mécanique.
- Inconvénients : Elle peut manquer d'importantes sources d'erreurs et d'observations invalides pour les ensembles de données ayant un **nombre élevé de caractéristiques**. Une connaissance du domaine peut exposer le processus à un biais en négligeant les domaines inintéressants de l'ensemble de données.

OUTILS ET MÉTHODES

Approche méthodique

- Liste des problèmes possibles (bingo du nettoyage des données)
- Code pouvant être réutilisé dans différents contextes

Approche descriptive

- Visualisation
- Sommaire des données
- Tableaux de distribution
- Petits multiples
- Analyse des données

Bingo du nettoyage des données

random missing values	outliers	values outside of expected range - numeric	factors incorrectly/inconsistently coded	date/time values in multiple formats
impossible numeric values	leading or trailing white space	badly formatted date/time values	non-random missing values	logical inconsistencies across fields
characters in numeric field	values outside of expected range - date/time	DCB!	inconsistent or no distinction between null, 0, not available, not applicable, missing	possible factors missing
multiple symbols used for missing values	???	fields incorrectly separated in row	blank fields	logical inconsistencies within field
entire blank rows	character encoding issues	duplicate value in unique field	non-factor values in factor	numeric values in character field

TESTS DE DÉTECTION DES VALEURS ABERRANTES

La **normalité** est une hypothèse pour la plupart des tests.

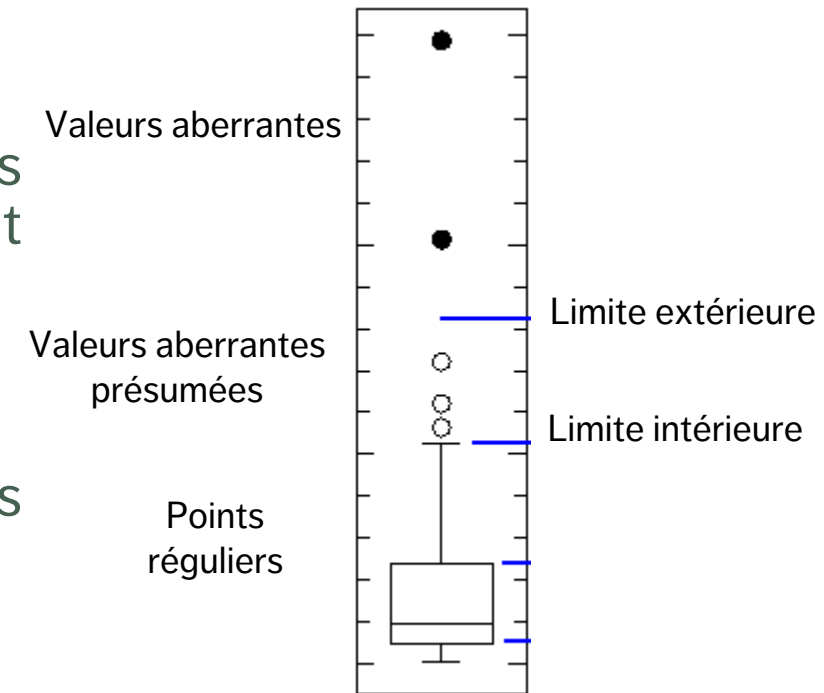
Test du diagramme de quartiles de Tukey : Pour les données distribuées normalement, les observations régulières se trouvent habituellement entre les **limites intérieures**.

$$Q_1 - 1.5 \times (Q_3 - Q_1) \text{ et } Q_3 + 1.5 \times (Q_3 - Q_1).$$

Les **valeurs aberrantes présumées** se trouvent entre les limites intérieures et les **limites extérieures**.

$$Q_1 - 3 \times (Q_3 - Q_1) \text{ et } Q_3 + 3 \times (Q_3 - Q_1).$$

Les **valeurs aberrantes** se trouvent au-delà des limites extérieures.



TESTS DE DÉTECTION DES VALEURS ABERRANTES

Le **test de Grubbs** est un test à une variable. Il faut tenir compte des éléments suivants :

- x_i : la valeur de la caractéristique X pour la i^{e} unité $1 \leq i \leq N$
- \bar{x} : la valeur moyenne de la caractéristique X
- s_x : l'écart-type de la caractéristique X
- α : le niveau de signification
- $T(\alpha, N)$: la valeur de la distribution t avec une signification de $\alpha/2N$

La i^{e} unité est une **valeur aberrante avec la caractéristique X** si

$$|x_i - \bar{x}| \geq \frac{s_x(N-1)}{\sqrt{N}} \times \sqrt{\frac{T^2(\alpha, N)}{N-2+T^2(\alpha, N)}}$$

TESTS DE DÉTECTION DES VALEURS ABERRANTES

Le **test Q de Dixon** est utilisé par les sciences expérimentales pour trouver les valeurs aberrantes de (très) petits ensembles de données (validité douteuse).

La **distance de Mahalanobis** (liée au levier) peut servir à trouver des valeurs aberrantes multidimensionnelles (quand les relations sont linéaires).

Autres tests :

- **Test de Tietjen-Moore** (pour un nombre précis de valeurs aberrantes)
- **Test généralisé de la déviation extrême de Student** (pour un nombre inconnu de valeurs aberrantes)
- **Test du chi carré** (valeurs aberrantes ayant un effet sur la qualité de l'ajustement)
- **DBSCAN, OR_h et LOF** (détection non supervisée des valeurs aberrantes)

MÉTHODES D'IMPUTATION

Suppression à partir d'une liste : Retrait des unités ayant au moins une valeur manquante.

- Hypothèse : MCAR
- Inconvénient : Introduction possible d'un biais (si non MCAR), réduction de la taille de l'échantillon, augmentation de l'écart-type.

Imputation par la moyenne ou la plus fréquente : Remplacement des valeurs manquantes par la valeur moyenne ou la plus fréquente.

- Hypothèse : MCAR
- Inconvénients : Distorsions de la distribution (pointe à la moyenne) et relations entre les variables.

MÉTHODES D'IMPUTATION

Imputation par régression ou corrélation : Remplacement des valeurs manquantes à l'aide d'une régression fondée sur d'autres variables (avec renseignements complets).

- Hypothèse : MAR
- Inconvénients : Réduction artificielle de la variabilité, surestimation de la corrélation.

Imputation par régression stochastique : Imputation par régression avec ajout de termes d'erreur aléatoires.

- Hypothèse : MAR
- Inconvénients : Risque accru d'une erreur de type I (faux positifs) en raison de l'écart-type.

MÉTHODES D'IMPUTATION

Report en avant de la dernière observation : Remplacement des valeurs manquantes par des valeurs antérieures (étude longitudinale)

- Hypothèse : MCAR. Les valeurs ne varient pas beaucoup au fil du temps.
- Inconvénients : Peut s'avérer trop « généreux », selon la nature de l'étude.

Imputation par la méthode du plus proche voisin k : Remplacement de l'entrée manquante par la moyenne du groupe des k répondants complets les plus **similaires**.

- Hypothèse : MAR
- Inconvénients : Choix difficile de la valeur pertinente pour k . Distorsion possible de la structure des données ($k > 1$).

RÉDUCTION ET TRANSFORMATION DES DONNÉES

COLLECTE ET TRAITEMENT DES DONNÉES

OBJECTIFS D'APPRENTISSAGE

Connaître les concepts suivants :

- Dimensionnalité des données
- Fléau de la dimension
- Sélection des caractéristiques
- Analyse en composantes principales (ACP)
- Transformation des données
- Mise à l'échelle des données
- Discrétisation

DIMENSIONNALITÉ DES DONNÉES

Dans l'analyse des données, la dimension des données est le nombre de variables (ou attributs) collectées dans un ensemble de données, représenté par le nombre de colonnes.

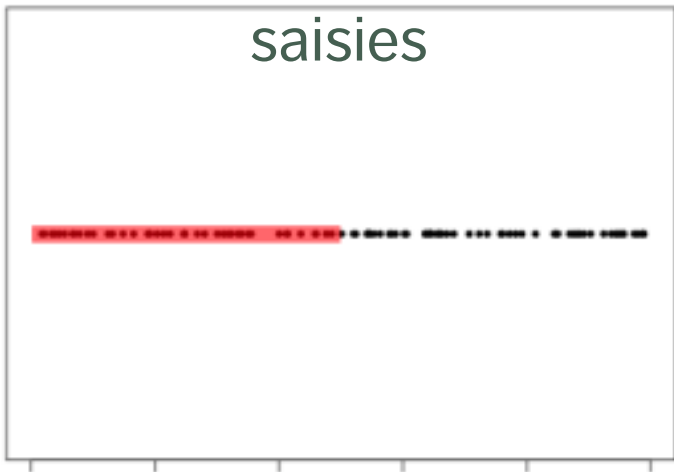
Ici, le terme « dimension » est une extension de l'utilisation du terme pour référer à la taille d'un vecteur.

Nous pouvons penser au nombre de variables utilisées pour décrire chaque objet (ligne) en tant que vecteur décrivant cet objet.

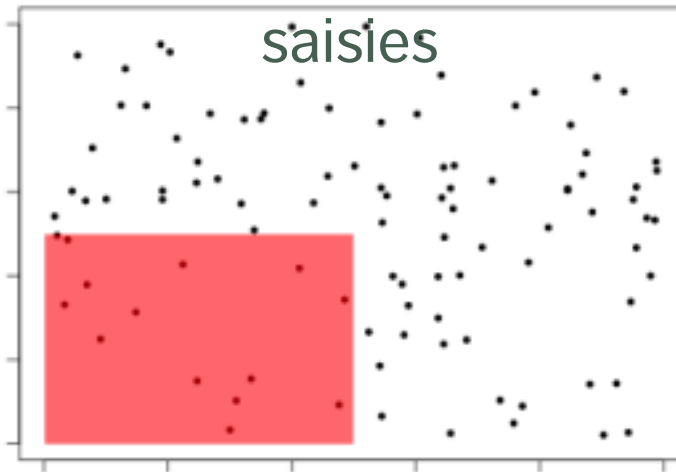
(Remarque : le terme dimension est utilisé différemment dans les contextes d'informatique décisionnelle.)

FLÉAU DE LA DIMENSION

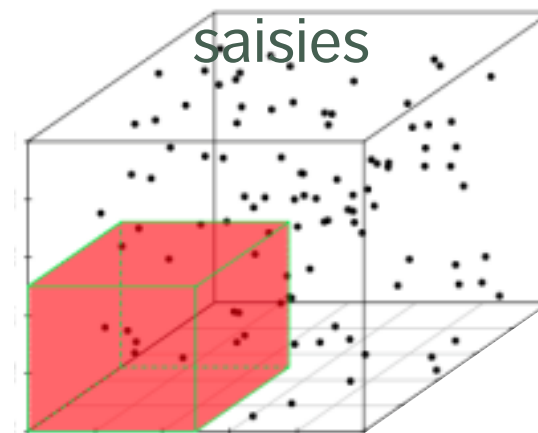
42 % des données sont
saisies



14 % des données sont
saisies



7 % des données sont
saisies



$N = 100$ observations, uniformément réparties sur $[0,1]^d$, $d = 1, 2, 3$.

% des observations saisies par $[0,1/2]^d$, $d = 1, 2, 3$.

OBSERVATIONS DE L'ÉCHANTILLONNAGE

Question : est-ce que chaque ligne de l'ensemble de données doit être utilisée?

Si les lignes sont sélectionnées de manière aléatoire (avec ou sans remplacement), l'échantillon résultant peut être **représentatif** de l'ensemble de données.

Désavantages :

- si le signal d'intérêt est rare, l'échantillonnage peut le noyer complètement
- si une agrégation se produit ultérieurement, l'échantillonnage aura nécessairement une incidence sur les nombres (passagers par rapport aux vols)
- Même des opérations simples sur un fichier volumineux (trouver le nombre de lignes, par exemple) peuvent être pénalisantes en termes de mémoire et de temps de calcul – **des informations préalables sur la structure de l'ensemble de données peuvent aider**

SÉLECTION DES CARACTÉRISTIQUES

La suppression de variables **non pertinentes** ou **redondantes** est une tâche courante de traitement de données.

Motivations :

- Les outils de modélisation ne traitent pas bien ces données (inflation de la variance due à la multicollinéarité, etc.)
- Réduction de la dimension (nombre de variables \gg nombre d'observations)

Approches :

- Filtrage et méthode enveloppante
- Non supervisé ou supervisé

ANALYSE EN COMPOSANTES PRINCIPALES

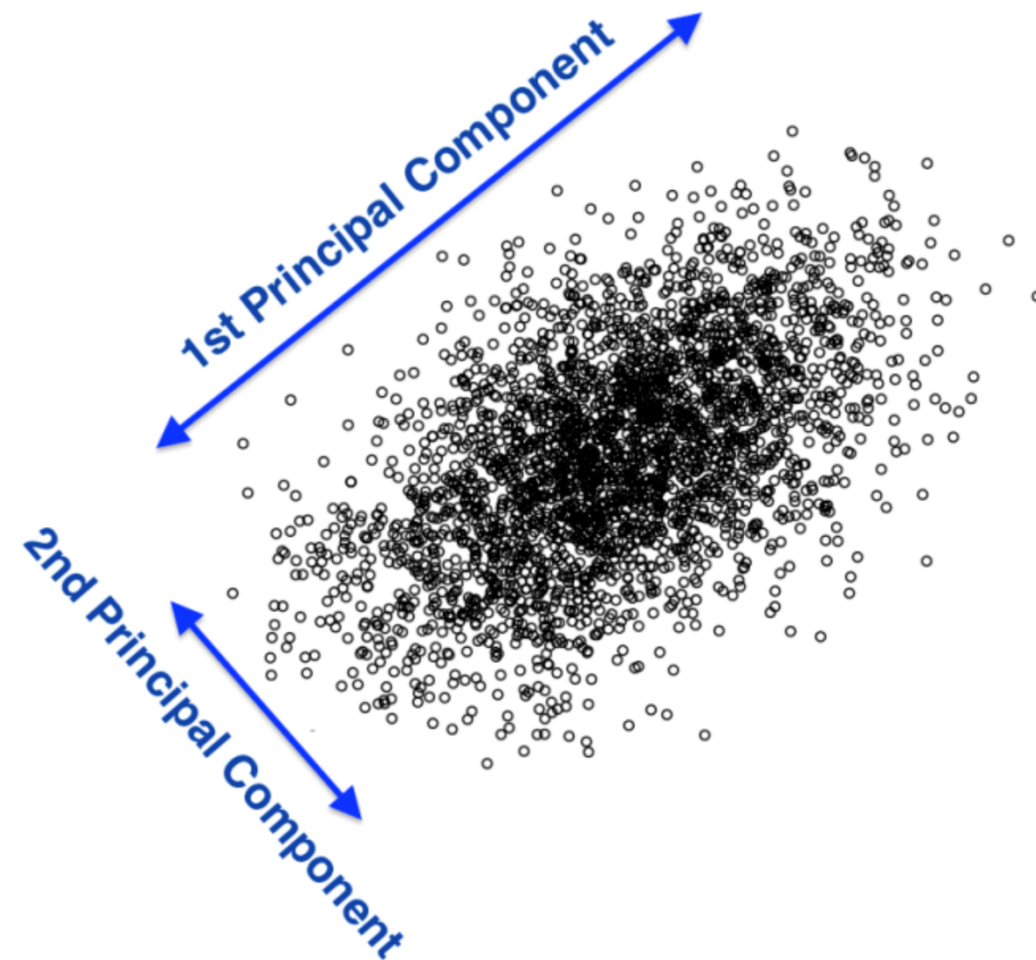
Exemple de motivation :

Contenu nutritionnel des aliments

Quelle est la meilleure façon de différencier les aliments? La teneur en vitamines, en lipides ou en protéines?

Un peu de tout?

L'**analyse en composantes principales** (ACP) peut être utilisée pour trouver les combinaisons de variables avec lesquelles les points de données sont **le plus dispersés**.



Vitamine C

- Parsley
- Kale

- Broccoli

- Cauliflower
- Soybeans
- Yam
- Guinea Hen

Vitamine C - Lipides

- Parsley
- Kale

- Broccoli

- Cauliflower
- Cabbage
- Spinach
- Yam

- Sweet Corn
- Guinea Hen
- Bluefish
- Mackerel
- Chicken

- Beef

- Pork
- Lamb

DIFFÉRENCIATION

La vitamine C est présente à différents niveaux dans les fruits et légumes, mais pas dans les viandes. On peut **séparer** les légumes des viandes, et des légumes particuliers les uns des autres (dans une certaine mesure) en fonction de leur teneur en vitamine C, mais les viandes sont **groupées** (à gauche).

C'est la situation inverse pour les niveaux de *lipides*, donc une combinaison de la teneur en vitamine C et en lipide permet de **séparer** les légumes des viandes et de les distribuer également (à droite).

TRANSFORMATIONS COURANTES

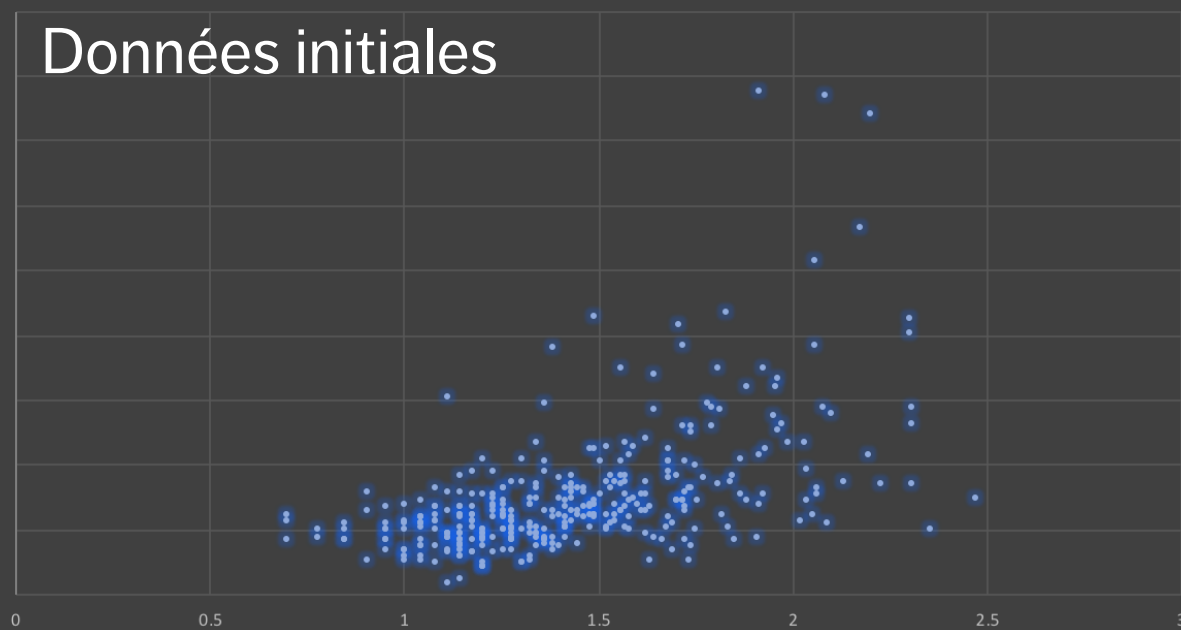
Les modèles exigent parfois que certaines hypothèses de données soient satisfaites (normalité des résidus, linéarité, etc.).

Si les données brutes ne répondent pas aux exigences, nous pouvons soit

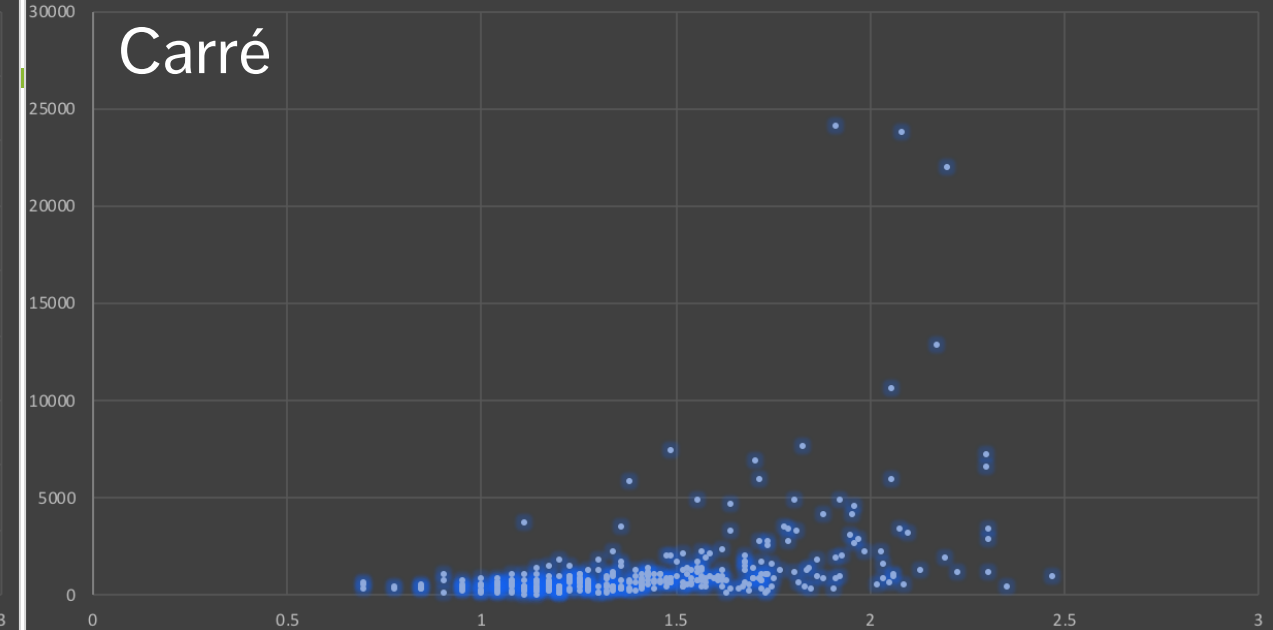
- abandonner le modèle
- tenter de **transformer** les données

La seconde approche nécessite une transformation inverse pour pouvoir tirer des conclusions sur les données d'origine.

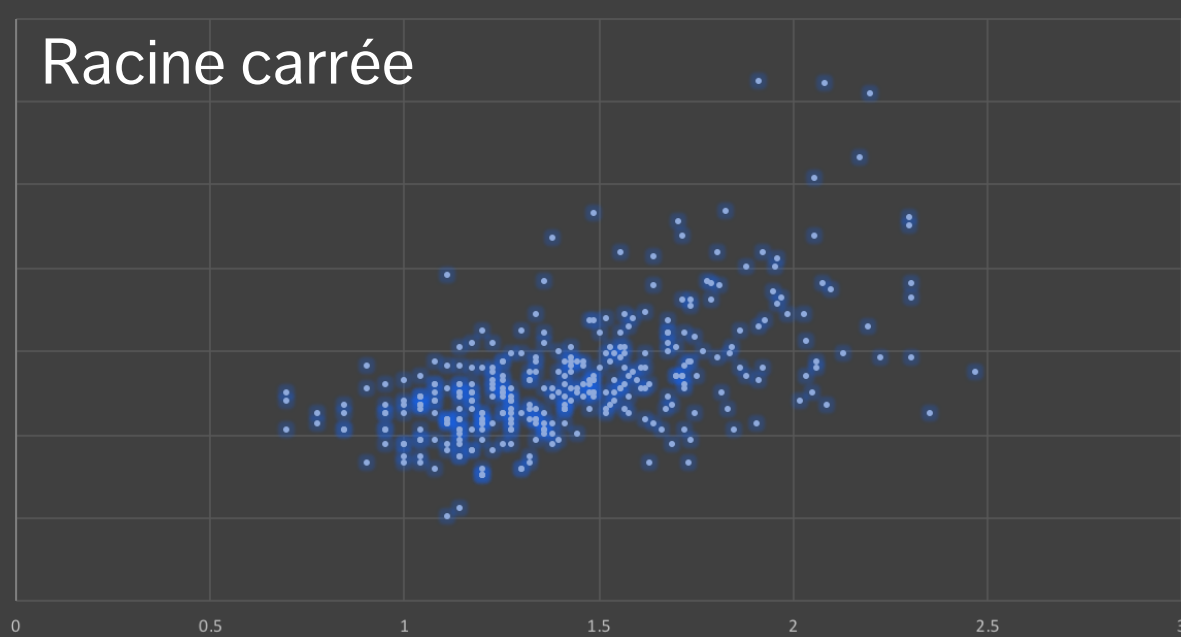
Données initiales



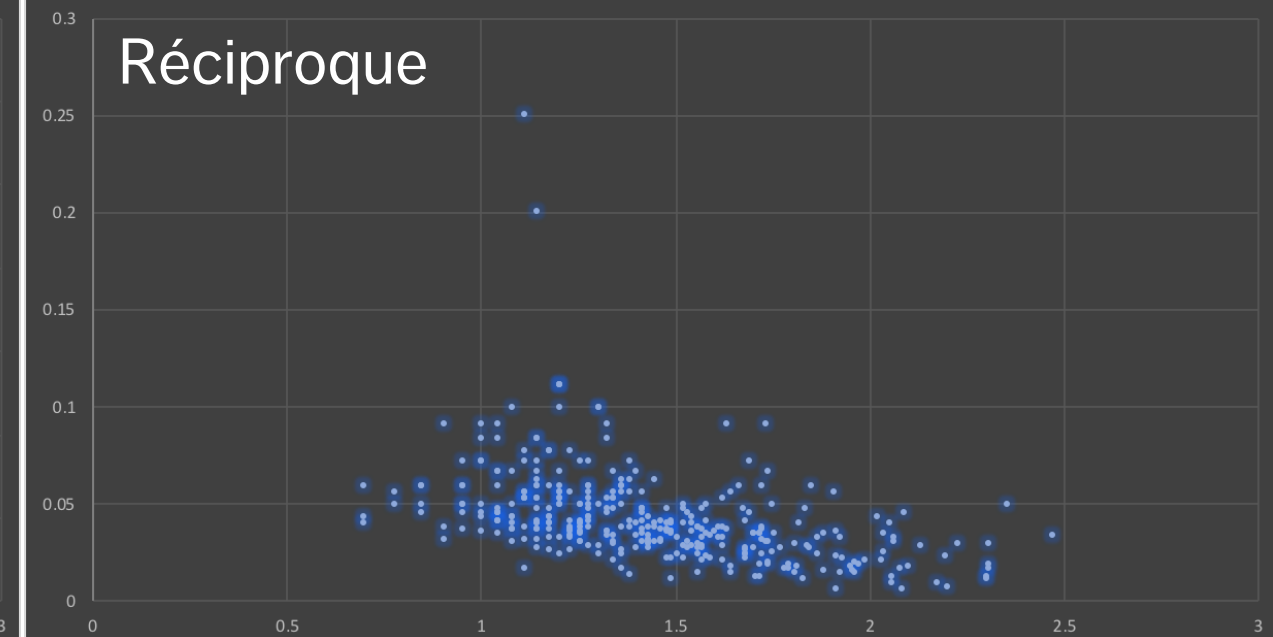
Carré



Racine carrée



Réciproque



MISE À L'ÉCHELLE DES DONNÉES

Les variables numériques peuvent avoir différentes **échelles** (poids et hauteurs, par exemple).

La variance d'une variable à large intervalle est généralement supérieure à celle d'une variable à petit intervalle, ce qui introduit une distorsion (par exemple).

La **standardisation** crée une variable dont la moyenne est 0 et l'écart-type 1 :

$$Y_i = \frac{X_i - \bar{X}}{s_X}$$

La **normalisation** crée une nouvelle variable dans l'intervalle [0,1]: $Y_i = \frac{X_i - \min X}{\max X - \min X}$

DISCRÉTISATION

Pour réduire la complexité des calculs, il peut être nécessaire de remplacer une variable numérique par une variable **ordinaire** (de la valeur de la *hauteur* à « *petit* », « *moyen* », « *grand* », par exemple).

On peut se fonder sur l'expertise du domaine pour déterminer les limites des catégories (bien que cela puisse introduire une distorsion inconsciente dans les analyses).

En l'absence d'une telle expertise, des limites peuvent être fixées pour que, soit

- les catégories contiennent chacune le même nombre d'observations
- les catégories aient chacune la même largeur
- la performance de certains outils de modélisation soit maximisée

CRÉATION DE VARIABLES

Il peut falloir introduire de nouvelles variables :

- en tant que **relations fonctionnelles** d'un sous-ensemble de caractéristiques disponibles
- parce que l'outil de modélisation peut nécessiter que **les observations soient indépendantes**
- parce que l'outil de modélisation peut nécessiter que **les caractéristiques soient indépendantes**
- pour simplifier l'analyse en consultant des **résumés agrégés** (souvent utilisés dans l'analyse de texte)

Dépendances temporelles → analyse des séries temporelles

Dépendances spatiales → analyse spatiale

Matériel supplémentaire

MÉTHODES LOCALES POUR LES NOMBREUSES DIMENSIONS

Un modèle est dit **local** s'il dépend uniquement des données qui sont *proches* du vecteur d'entrée (la méthode k NN est locale, la régression linéaire ne l'est pas).

Avec un **grand ensemble d'apprentissage**, nous pourrions augmenter k (dans un modèle k NN, par exemple) et obtenir suffisamment de points de données pour fournir une approximation fiable de la frontière théorique.

Le **fléau de la dimension** est la faillite de cette approche dans les espaces à nombreuses dimensions : lorsque le nombre de caractéristiques augmente, le nombre d'observations nécessaires pour maintenir la puissance prédictive augmente également... **mais à un taux nettement plus élevé.**

MANIFESTATIONS DU FLÉAU DE LA DIMENSION

Soit $x_i \sim U^1(0,1), i = 1, \dots, N$, des variables indépendantes et identiquement distribuées

Pour tout $z \in [0,1]$ et $\varepsilon > 0$ tels que

$$I_1(z; \varepsilon) = \left[z - \frac{\varepsilon}{2}, z + \frac{\varepsilon}{2} \right] \subseteq [0,1],$$

On s'attend à ce que $|I_1(z; \varepsilon) \cap \{x_i\}_{i=1}^N| \approx \varepsilon \cdot N$

Autrement dit, un sous-ensemble dont l'arête correspond à ε pour cent de l'ensemble d'origine dans \mathbb{R} contient ε pour cent des observations.

MANIFESTATIONS DU FLÉAU DE LA DIMENSION

Soit $x_i \sim U^2(0,1)$, $i = 1, \dots, N$, des variables indépendantes et identiquement distribuées

Pour tout $z \in [0,1]^2$ et $\varepsilon > 0$ tels que

$$I_2(z; \varepsilon) = \left[z_1 - \frac{\varepsilon}{2}, z_1 + \frac{\varepsilon}{2} \right] \times \left[z_2 - \frac{\varepsilon}{2}, z_2 + \frac{\varepsilon}{2} \right] \subseteq [0,1]^2,$$

On s'attend à ce que $|I_2(z; \varepsilon) \cap \{x_i\}_{i=1}^N| \approx \varepsilon^2 \cdot N$

Autrement dit, un sous-ensemble dont l'arête correspond à ε pour cent de l'ensemble d'origine dans \mathbb{R}^2 contient ε^2 pour cent des observations.

MANIFESTATIONS DU FLÉAU DE LA DIMENSION

Soit $x_i \sim U^p(0,1)$, $i = 1, \dots, N$, des variables indépendantes et identiquement distribuées

Pour tout $z \in [0,1]^p$ et $\varepsilon > 0$ tels que

$$I_p(z; \varepsilon) = \prod_{j=1}^p \left[z_j - \frac{\varepsilon}{2}, z_j + \frac{\varepsilon}{2} \right] \subseteq [0,1]^p,$$

On s'attend à ce que $|I_p(z; \varepsilon) \cap \{x_i\}_{i=1}^N| \approx \varepsilon^p \cdot N$

Autrement dit, un sous-ensemble dont l'arête correspond à ε pour cent de l'ensemble d'origine dans \mathbb{R}^p contient ε^p pour cent des observations.

MANIFESTATIONS DU FLÉAU DE LA DIMENSION

Pour saisir r pour cent des observations uniformément réparties dans un p -cube unité, nous avons besoin d'un hyper-sous-ensemble dont l'arête est définie par

$$\varepsilon_p(r) = r^{1/p}.$$

Par exemple, pour $r = 33\%$, nous avons besoin d'un sous-ensemble dont l'arête est

- $\varepsilon_1(1/3) \approx 0.33$ dans \mathbb{R}
- $\varepsilon_2(1/3) \approx 0.58$ dans \mathbb{R}^2
- $\varepsilon_{10}(1/3) \approx 0.90$ dans \mathbb{R}^{10}

On perd la localité!

MÉTHODES DE FILTRAGE SUPERVISÉES

Corrélation entre une caractéristique X et une variable cible Y :

$$\rho_{X,Y} = \frac{\sum_{i=1}^N (y_i - \bar{y}) (x_i - \bar{x})}{\sqrt{\sum_{i=1}^N (y_i - \bar{y})^2} \cdot \sqrt{\sum_{i=1}^N (x_i - \bar{x})^2}}$$

Les caractéristiques fortement corrélées avec la variable cible sont conservées, mais cette approche est limitée si la relation avec la variable cible est **non linéaire**.

MÉTHODES DE FILTRAGE SUPERVISÉES

Information mutuelle de la cible nominale Y de la caractéristique nominale X :

$$I(Y; X) = H(Y) - H(Y|X)$$

où l'**entropie** et l'**entropie de classe conditionnée** sont définies par

$$H(Y) = - \sum_c P(Y = c) \cdot \log P(Y = c) \quad (v, c \text{ représentent les niveaux de } X, Y).$$

$$H(Y|X) = - \sum_{v,c} P(X = v, Y = c) \cdot \log \frac{P(X=v, Y=c)}{P(X=v)}$$

$I(Y; X)$ Mesure la quantité d'informations qui peut être obtenue à propos de Y en sachant X .

LASSO ET SES VARIANTES

La **sélection par étapes** est une forme de *rasoir d'Occam* : à chaque étape, on considère une nouvelle fonctionnalité pour l'**inclure** ou la **supprimer** de l'ensemble des fonctionnalités actuelles en fonction de certains critères (test F , test t , etc.).

Limites :

- Les tests sont biaisés, car ils sont basés sur les mêmes données.
- L'ajustement de R^2 ne prend en compte que le nombre de caractéristiques de l'ajustement final, et non les caractéristiques de données utilisées dans l'ensemble du modèle.
- Si la validation croisée est utilisée, la sélection par étapes doit être répétée pour chaque sous-modèle (ce n'est généralement pas le cas).
- Exemple classique du triturage des données (p -hacking) (résultats sans hypothèses).

LASSO ET SES VARIANTES

Dans l'exemple qui suit, nous supposons que N est **centré** et **normalisé** $x_i = (x_{1,i}, \dots, x_{p,i})^T$ et une observation cible y_i .

Soit $\hat{\beta}_{LS,j} = [(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}]_j$ le j^{e} coefficient MCO et un seuil défini $\lambda > 0$, dont la valeur dépend de l'ensemble de données d'apprentissage.

En général, il n'y a **aucune restriction** sur les valeurs prises par les coefficients $\hat{\beta}_{LS,j}$ – une plus grande magnitude implique que les caractéristiques correspondantes **jouent un rôle important** dans la prédiction de la cible.

LASSO ET SES VARIANTES

La **régression d'arête** est une méthode pour **régulariser** les coefficients de régression (l'effet est de réduire les valeurs des coefficients)

Le problème consiste à résoudre

$$\arg \min_{\boldsymbol{\beta}} \{ \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + N\lambda \|\boldsymbol{\beta}\|_2^2 \},$$

qui permet d'obtenir les coefficients d'arête (dans le cas où $\mathbf{X}^T \mathbf{X} = \mathbf{I}$):

$$\hat{\beta}_{RR,j} = \frac{\hat{\beta}_{LS,j}}{1 + N\lambda}$$

LASSO ET SES VARIANTES

La **régression avec la meilleure sélection de sous-ensembles** est une méthode où certains coefficients de régression sont nuls (potentiellement).

Le problème consiste à résoudre

$$\arg \min_{\boldsymbol{\beta}} \{ \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + N\lambda \|\boldsymbol{\beta}\|_0 \}, \text{ où } \|\boldsymbol{\beta}\|_0 = \sum_j \text{sign}(|\beta_j|),$$

permet d'obtenir les coefficients (dans le cas où $\mathbf{X}^T \mathbf{X} = \mathbf{I}$):

$$\hat{\beta}_{BS,j} = \begin{cases} 0 & \text{if } |\hat{\beta}_{LS,j}| < \sqrt{N\lambda} \\ \hat{\beta}_{LS,j} & \text{if } |\hat{\beta}_{LS,j}| \geq \sqrt{N\lambda} \end{cases}$$

LASSO ET SES VARIANTES

LASSO (Least Absolute Shrinkage and Selection Operator) est une méthode de régression pour **sélectionner** et **régulariser les caractéristiques**.

Le problème consiste à résoudre

$$\arg \min_{\boldsymbol{\beta}} \{ \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + N\lambda \|\boldsymbol{\beta}\|_1 \}$$

qui permet d'obtenir les coefficients lasso (dans le cas où $\mathbf{X}^T \mathbf{X} = \mathbf{I}$):

$$\hat{\beta}_{L,j} = \hat{\beta}_{LS,j} \cdot \max \left(0, 1 - \frac{N\lambda}{|\hat{\beta}_{LS,j}|} \right)$$

LASSO ET SES VARIANTES

LASSO combine les propriétés de la **régression d'arête** (réduction) et de **sélection des meilleurs sous-ensembles** (sélection des caractéristiques).

La régression d'arête peut être vue comme une régression linéaire avec des **distributions antérieures normales** assignées aux coefficients; ce sont des **distributions de Laplace** en régression LASSO.

LASSO sélectionne **au plus** $\max\{p, N\}$ caractéristiques, et ne sélectionne généralement pas plus d'une caractéristique dans un groupe de variables hautement corrélées.

Extensions : filets élastiques; lassos de groupe, fusionnés et adaptatifs; régression « bridge »

COMPOSANTES PRINCIPALES

La présence d'éléments nutritifs semble être corrélée entre les aliments.

Dans le (petit) échantillon composé d'agneau, de porc, de chou frisé et de persil, les *teneurs en lipides* et en *protéines* semblent concorder, de même que les *teneurs en fibres* et en *vitamine C*.

Dans un plus grand ensemble de données, les corrélations sont $r = 0.56$ et $r = 0.57$.

Que peuvent expliquer deux variables?



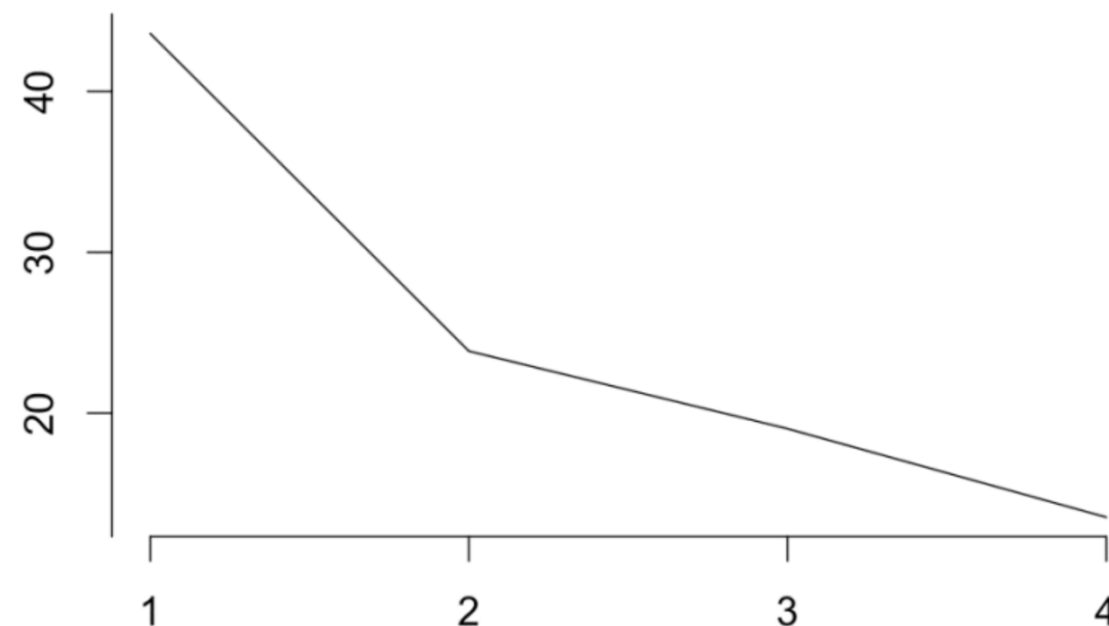
CONSERVER LES COMPOSANTES PRINCIPALES

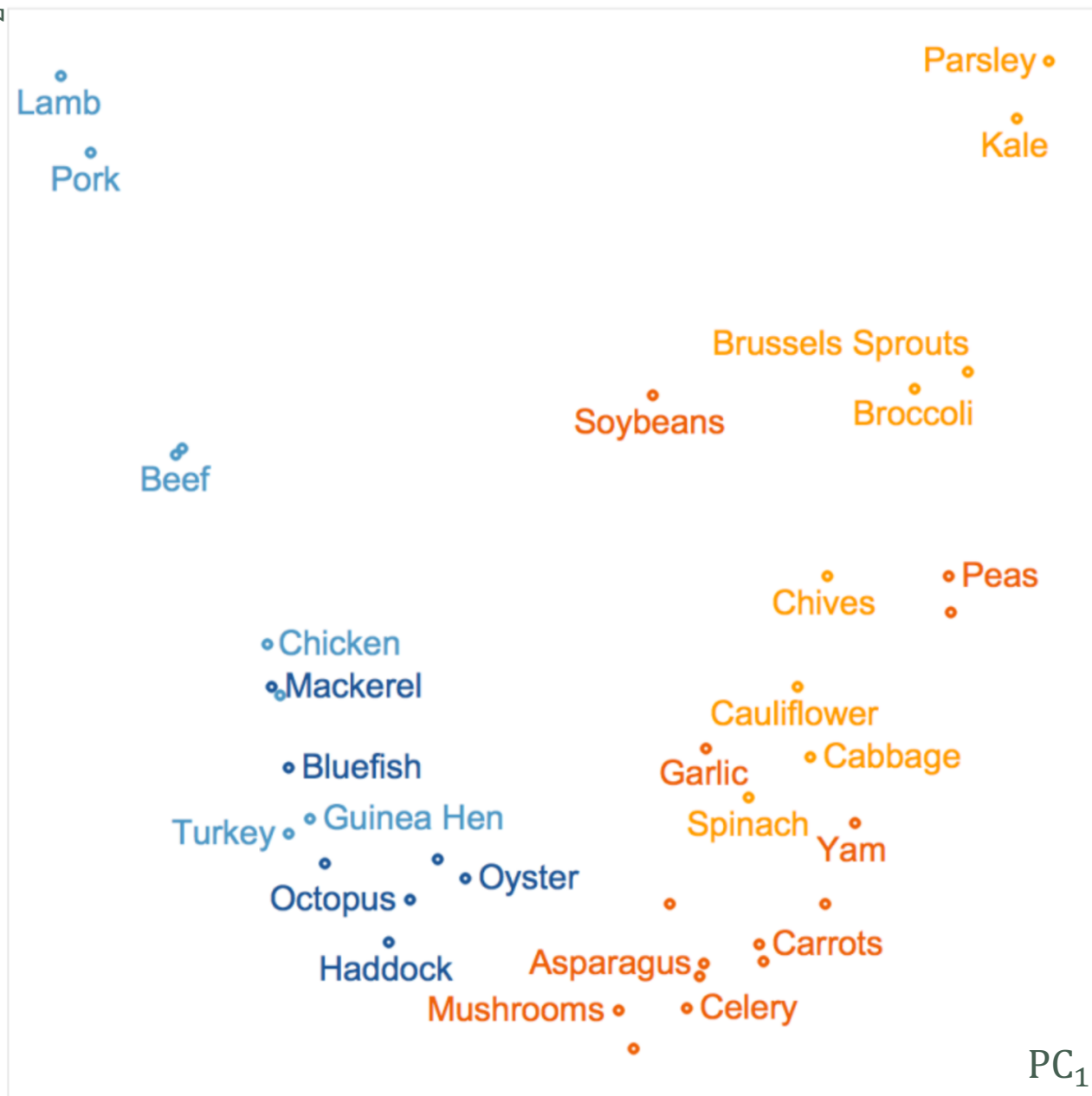
La **proportion de la dispersion** des données qui peut être expliquée par chaque composante principale est indiquée dans le scree plot.

Combien de composantes principales sont retenues pour l'analyse?

- conserver les composantes principales dont la proportion cumulée est inférieure à un certain seuil
- garder les composantes principales menant à un coude

Ici, 2 composantes principales $\approx 68\%$ de la dispersion.





DIFFÉRENTIATION (REPRISE)

PC₁ différencie les légumes des viandes; PC₂ différencie deux **sous-catégories** à l'intérieur de ces ensembles :

- Les **viandes** sont concentrées à gauche (valeurs de PC₁ faibles).
- Les **légumes** sont concentrés à droite (valeurs de PC₁ élevées).
- Les **fruits de mer** ont une teneur en *lipides* plus faible (valeurs de PC₂ faibles) et sont concentrés vers le bas.
- Les **légumes sans feuilles** ont une teneur plus faible en *vitamine C* (valeurs de PC₂ faibles) et sont également concentrés vers le bas.

ACP EN THÉORIE

L'ACP tente d'adapter un **ellipsoïde** p à des données centrées et normalisées*. Les axes ellipsoïdaux représentent les composantes principales des données. Les petits axes sont des composantes le long desquelles la variance est « petite »; la suppression de ces composantes entraîne une « petite » perte d'informations.

Procédure :

1. Centrer et normaliser les données : matrice X
2. Calculer la matrice de covariance des données $K = X^T X$
3. Calculer les valeurs Λ propres de K et la matrice de vecteurs propres orthonormée W
4. Chaque vecteur propre w représente un axe, dont la variance est donnée par la valeur propre associée λ

ACP EN THÉORIE

Les vecteurs propres \mathbf{w} sont également appelés **poids**.

En règle générale, les valeurs propres sont classées par ordre **décroissant**, de sorte que le premier poids correspond au plus grand axe.

\mathbf{K} positif semi-défini \Rightarrow les valeurs propres $\lambda = s^2$ sont positives; s est une valeur particulière de \mathbf{X} (entrée diagonale de $\mathbf{\Sigma}$ dans la **décomposition en valeurs singulières** $\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{W}^T$).

La décomposition ACP de \mathbf{X} est $\mathbf{T} = \mathbf{X}\mathbf{W} = \mathbf{U}\mathbf{\Sigma}$.

ACP EN THÉORIE

Le lien entre les composantes principales et les vecteurs propres peut être rendu explicite :

- la **première** composante principale est le poids qui maximise la variance de la première colonne de T

$$\mathbf{w}^1 = \arg \max_{\|\mathbf{w}\|=1} \{\text{Var}(\mathbf{t}^1)\}$$

- mais T est centré de sorte que la variance est simplement

$$\mathbf{w}^1 = \arg \max_{\|\mathbf{w}\|=1} \{t_{1,1}^2 + \dots + t_{1,N}^2\}$$

- en utilisant la décomposition en composantes principales de X , cela devient

$$\mathbf{w}^1 = \arg \max_{\|\mathbf{w}\|=1} \{\langle \mathbf{x}_1 | \mathbf{w} \rangle^2 + \dots + \langle \mathbf{x}_N | \mathbf{w} \rangle^2\} = \arg \max_{\|\mathbf{w}\|=1} \{\|\mathbf{X}\mathbf{w}\|^2\}$$

ACP EN THÉORIE

- qui, par définition de la norme, équivaut à $\mathbf{w}^1 = \arg \max_{\|\mathbf{w}\|=1} \{\mathbf{w}^T \mathbf{X}^T \mathbf{X} \mathbf{w}\}$
- puisque $\|\mathbf{w}\| = 1$, le poids satisfait également

$$\mathbf{w}^1 = \arg \max_{\|\mathbf{w}\|=1} \left\{ \frac{\mathbf{w}^T \mathbf{X}^T \mathbf{X} \mathbf{w}}{\mathbf{w}^T \mathbf{w}} \right\}$$

- en utilisant des multiplicateurs de Lagrange, on peut montrer que les points critiques de $\frac{\mathbf{w}^T \mathbf{X}^T \mathbf{X} \mathbf{w}}{\mathbf{w}^T \mathbf{w}}$ sont exactement les vecteurs propres de $\mathbf{K} = \mathbf{X}^T \mathbf{X}$ (il y en a p)
- si \mathbf{w} (unité) et $\lambda^* \geq 0$ sont tels que $\mathbf{K} \mathbf{w} = \lambda \mathbf{w}$, alors

$$\frac{\mathbf{w}^T \mathbf{X}^T \mathbf{X} \mathbf{w}}{\mathbf{w}^T \mathbf{w}} = \frac{\mathbf{w}^T \mathbf{K} \mathbf{w}}{\mathbf{w}^T \mathbf{w}} = \frac{\mathbf{w}^T \lambda \mathbf{w}}{\mathbf{w}^T \mathbf{w}} = \lambda \frac{\mathbf{w}^T \mathbf{w}}{\mathbf{w}^T \mathbf{w}} = \lambda$$

REMARQUES SUR L'ACP

Le poids qui explique la plus grande variance le long d'un seul axe est le vecteur propre de la matrice de covariance empirique correspondant à la plus grande valeur propre et cette variance est proportionnelle à la valeur propre.

Le processus est répété pour obtenir les composantes principales **orthonormées** PC_1, \dots, PC_r , où $r = \text{rank}(X)$.

GÉNÉRALISATIONS

Les méthodes non linéaires de type ACP tentent de trouver des **variétés principales**.

- cartes auto-organisées
- auto-encodeurs
- analyse en composantes curvilignes
- sculpture de variété

Plutôt que de réduire la dimensionnalité, nous pouvons **l'étendre** avec une ACP à noyau (cela revient à remplacer le produit interne habituel par des objets plus exotiques).

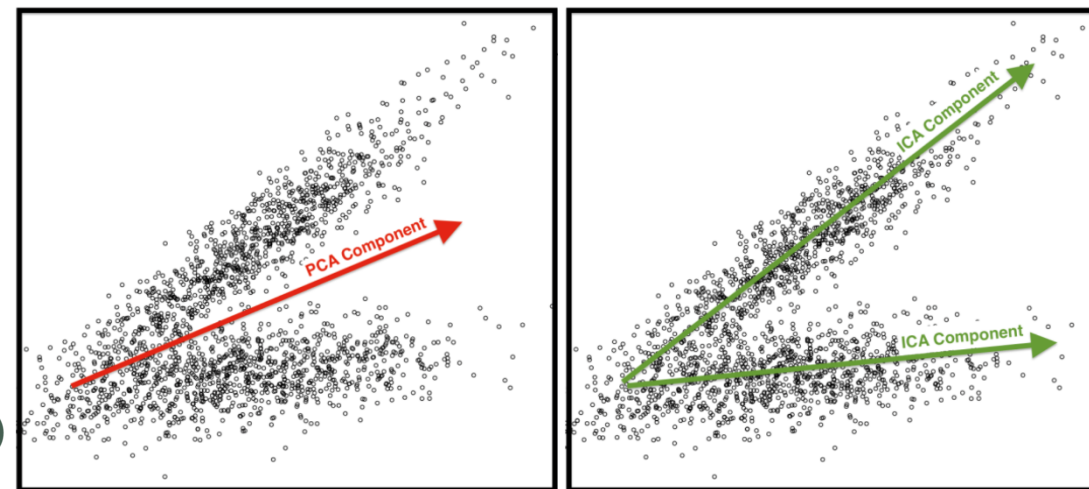
LIMITES

L'ACP dépend de la mise à l'échelle des données (pas unique).

En l'absence de connaissances préalables dans le domaine, l'interprétation des composantes principales peut s'avérer difficile.

Les hypothèses ne sont **pas toujours satisfaites**

- Les structures importantes et la dispersion sont en corrélation (comptage des crêpes)
- Les composantes principales sont orthogonales (qu'en est-il de l'ACI?)
- Changement de cadre de base (suivi du déplacement d'une personne sur la grande roue)



Sensible aux valeurs aberrantes

DIMENSIONNALITÉ ÉLEVÉE ET MÉGADONNÉES

Les ensembles de données peuvent être « volumineux » de différentes manières :

- trop volumineux pour que le **matériel** puisse les traiter (impossibles à enregistrer ou d'y accéder correctement en raison du trop grand nombre d'observations, de caractéristiques ou de la taille totale)
- les dimensions peuvent aller à l'encontre d'**hypothèses de modélisation** spécifiques (nombre de caractéristiques >> nombre d'observations)

Exemples :

- Plusieurs capteurs enregistrant plus de 100 observations par seconde dans une grande zone géographique sur une longue période = **un très grand ensemble de données**.
- Dans la *matrice documents-termes* d'un corpus (colonnes = termes, lignes = documents), le nombre de termes est généralement beaucoup plus élevé que le nombre de documents, ce qui entraîne une **trop grande densité de données**.

MÉTHODES DE SÉLECTION DES CARACTÉRISTIQUES

Les **méthodes de filtrage** inspectent chaque variable individuellement et les notent selon une **mesure d'importance**.

Les caractéristiques les moins pertinentes (c.-à-d. dont la note d'importance est inférieure à un seuil défini) sont ensuite éliminées.

Les **méthodes enveloppantes** recherchent des sous-ensembles de caractéristiques pour lesquels le critère d'évaluation utilisé par la méthode analytique éventuelle est « optimisé ».

Le processus est **itératif** et nécessite généralement beaucoup de calcul : des sous-ensembles candidats sont utilisés dans l'analyse jusqu'à ce que l'on produise une mesure d'évaluation acceptable pour l'analyse.

MÉTHODES DE SÉLECTION DES CARACTÉRISTIQUES

Les **méthodes non supervisées** déterminent l'importance d'une caractéristique uniquement en fonction de ses valeurs.

Les **méthodes supervisées** évaluent l'importance de chaque caractéristique en étudiant la relation avec une **caractéristique cible** (corrélation, etc.).

Les méthodes enveloppantes sont généralement supervisées.

Méthodes de filtrage non supervisées : suppression des variables constantes, variables de type ID (différentes pour toutes les observations), caractéristiques à faible variabilité, etc.

MÉTHODES DE FILTRAGE SUPERVISÉES

Corrélation entre une caractéristique X et une variable cible Y :

$$\rho_{X,Y} = \frac{\sum_{i=1}^N (y_i - \bar{y}) (x_i - \bar{x})}{\sqrt{\sum_{i=1}^N (y_i - \bar{y})^2} \cdot \sqrt{\sum_{i=1}^N (x_i - \bar{x})^2}}$$

Les caractéristiques fortement corrélées avec la variable cible sont conservées, mais cette approche est limitée si la relation avec la variable cible est **non linéaire**.

AUTRES MESURES SUPERVISÉES

Tâches de classification

- Ratio de gain
- Gain d'information
- Gini
- MDL, etc.

Tâches de régression

- Erreur quadratique de la moyenne
- Erreur absolue de la moyenne
- Relief (évalue les caractéristiques simultanément), etc.

TRANSFORMATIONS COURANTES

Dans le contexte de la régression, les transformations sont **monotones** :

- logarithmique
- racine carrée, inverse, puissance : W^k
- exponentielle
- Box-Cox, etc.

Les transformations sur X peuvent atteindre la linéarité, mais cela a généralement un coût (les corrélations ne sont pas préservées, par exemple). Les transformations sur Y peuvent être utiles en cas de non-normalité et de variance inégale des termes d'erreur.

TRANSFORMATION BOX-COX

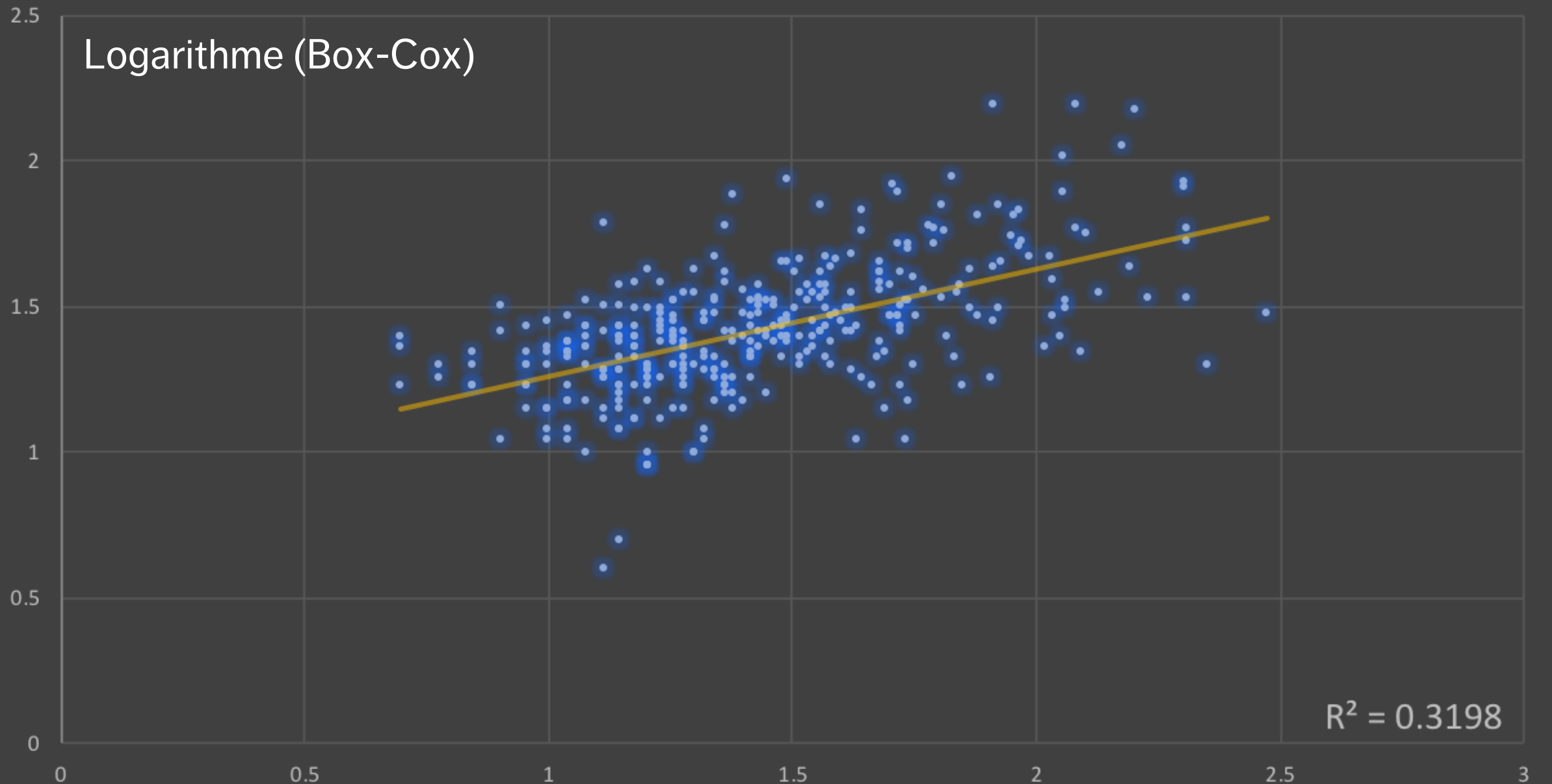
Supposons le modèle habituel $Y_j = \sum_i \beta_i X_{j,i} + \varepsilon_j$ avec soit

- des résidus asymétriques
- une variance non constante
- une tendance non linéaire

La **transformation Box-Cox** $Y_j \mapsto Y_j'(\lambda)$ suggère un choix : choisissez λ de sorte qu'il maximise la vraisemblance logarithmique correspondante

$$Y_j'(\lambda) = \begin{cases} \text{gm}(\mathbf{Y}) \times \ln(Y_j), & \lambda = 0 \\ \lambda^{-1} \text{gm}(\mathbf{Y})^{1-\lambda} \times (Y_j^\lambda - 1), & \lambda \neq 0 \end{cases}$$

Logarithme (Box-Cox)



TRANSFORMATION BOX-COX

La procédure fournit un **guide** pour sélectionner une transformation.

Des raisons théoriques peuvent exister pour choisir un λ particulier.

Une analyse résiduelle est toujours nécessaire pour s'assurer que le choix était approprié.

Les paramètres résultants ont la propriété des moindres carrés uniquement par rapport aux points de données transformés.

QUALITÉ ET VALIDATION DES DONNÉES

COLLECTE ET TRAITEMENT DES DONNÉES

Martin : Nos données sont en désordre.

Allison : Même après avoir été nettoyées?

Martin : Surtout après avoir été nettoyées.

P. Boily, The Great Balancing Act

OBJECTIFS D'APPRENTISSAGE

Comprendre les sources d'erreur de données fréquentes et les types de problèmes potentiels

Expliquer la différence entre « exactitude » et « précision »

Comprendre, de manière générale, certaines techniques de détection des problèmes de données

Se familiariser avec quelques exemples de problèmes de validité des données

SOLIDITÉ DES DONNÉES

L'ensemble de données idéal présentera le moins de problèmes possible en ce qui a trait aux caractéristiques suivantes :

- **Validité** : type de données, plage, réponse obligatoire, unicité, valeur, expressions régulières
- **Exhaustivité** : observations manquantes
- **Exactitude et précision** : liées à des erreurs de mesure ou de saisie des données; [diagrammes de cible](#) (exactitude en matière de subjectivité, précision en matière d'erreur standard)
- **Cohérence** : observations conflictuelles
- **Uniformité** : les unités sont-elles utilisées uniformément?

Vérifier que les données ne posent pas de problème de qualité à un stade précoce peut éviter des maux de tête plus tard dans l'analyse.



exactes et
précises



précises, mais
non exactes



exactes, mais
non précises



ni exactes ni très
précises

SOURCES D'ERREUR FRÉQUENTES

Lorsque vous traitez des ensembles de données **anciens**, **hérités** ou **combinés** (c'est-à-dire des ensembles de données sur lesquels vous avez peu de contrôle) :

- Données manquantes dans un code donné
- « S.O. » ou champ vierge dans un code donné
- Erreur de saisie des données
- Erreur de codage
- Erreur de mesure
- Entrées en double
- Accumulation

DÉTECTER LES ENTRÉES NON VALIDES

Les entrées potentiellement non valides peuvent être détectées à l'aide de :

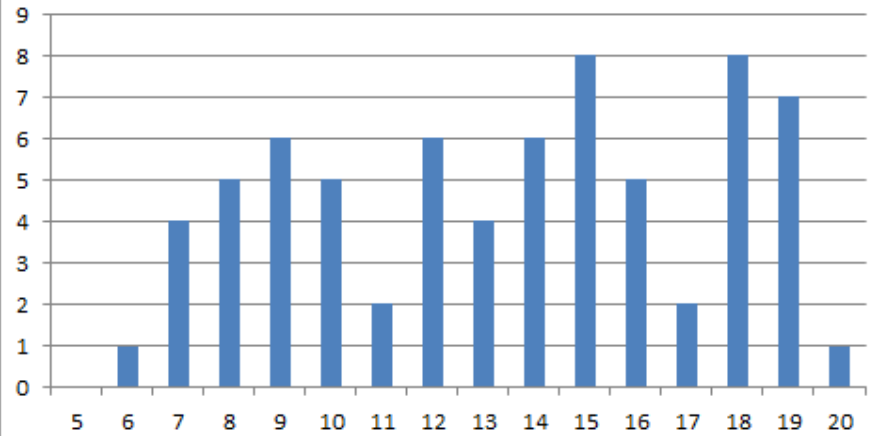
- **Statistiques descriptives univariées**
dénombrement, plage, score z , moyenne, médiane, écart-type, vérification de la logique
- **Statistiques descriptives multivariées**
tableau à n dimensions, vérification de la logique
- **Visualisation des données**
diagramme de dispersion, matrice de diagramme de dispersion, histogramme, histogramme conjoint, etc.

Cette étape pourrait permettre de repérer d'éventuelles valeurs aberrantes.

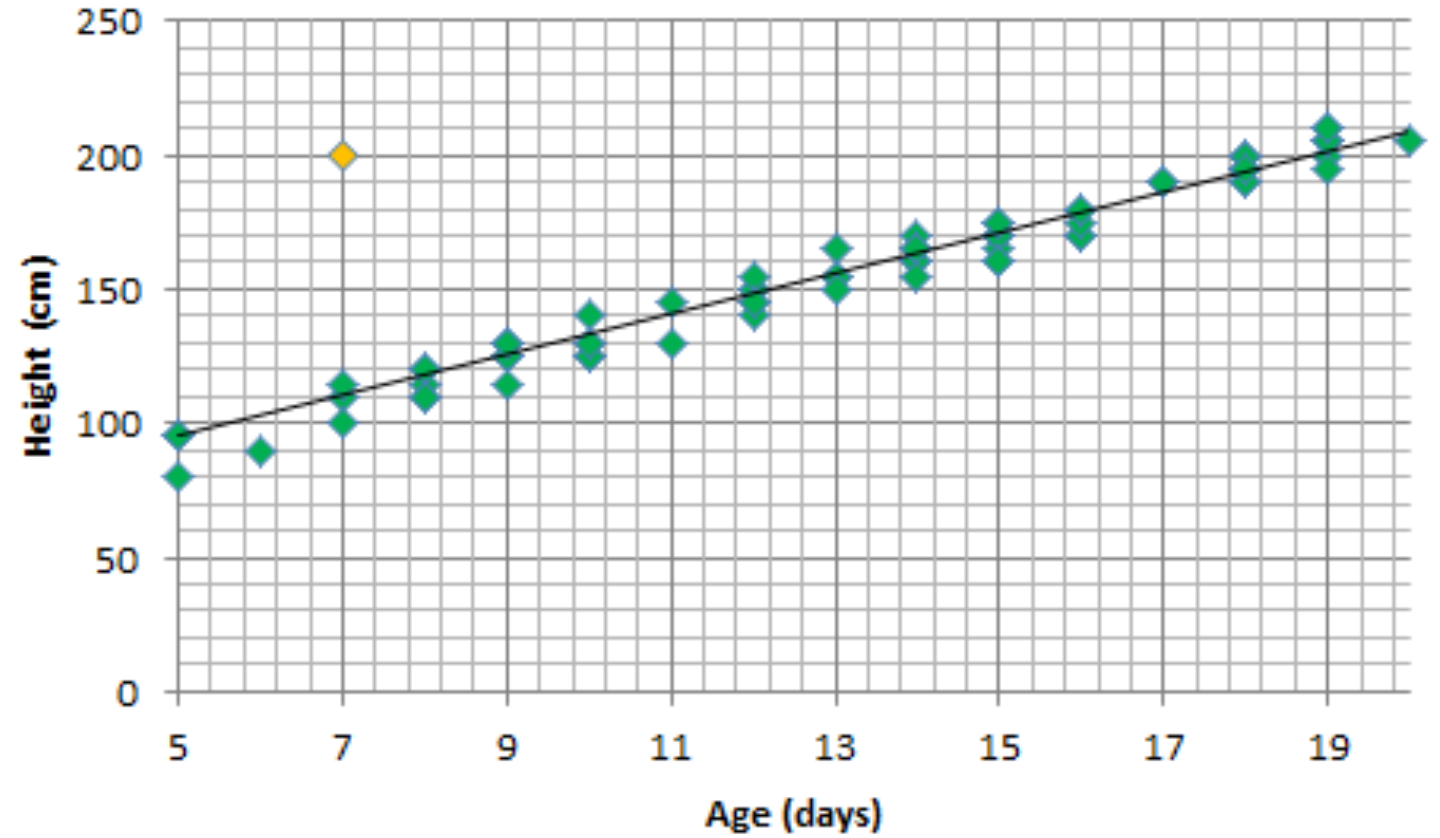
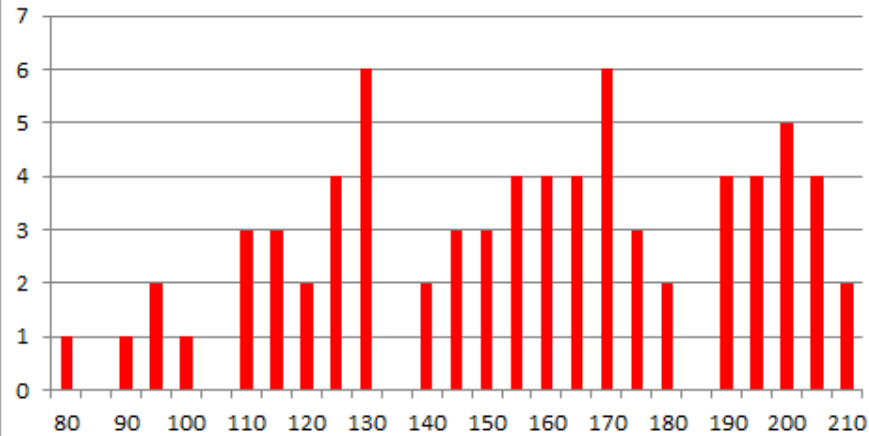
Impossibilité de détecter des entrées non valides \neq toutes les entrées sont valides.

Petit nombre d'entrées non valides enregistrées comme « manquantes ».

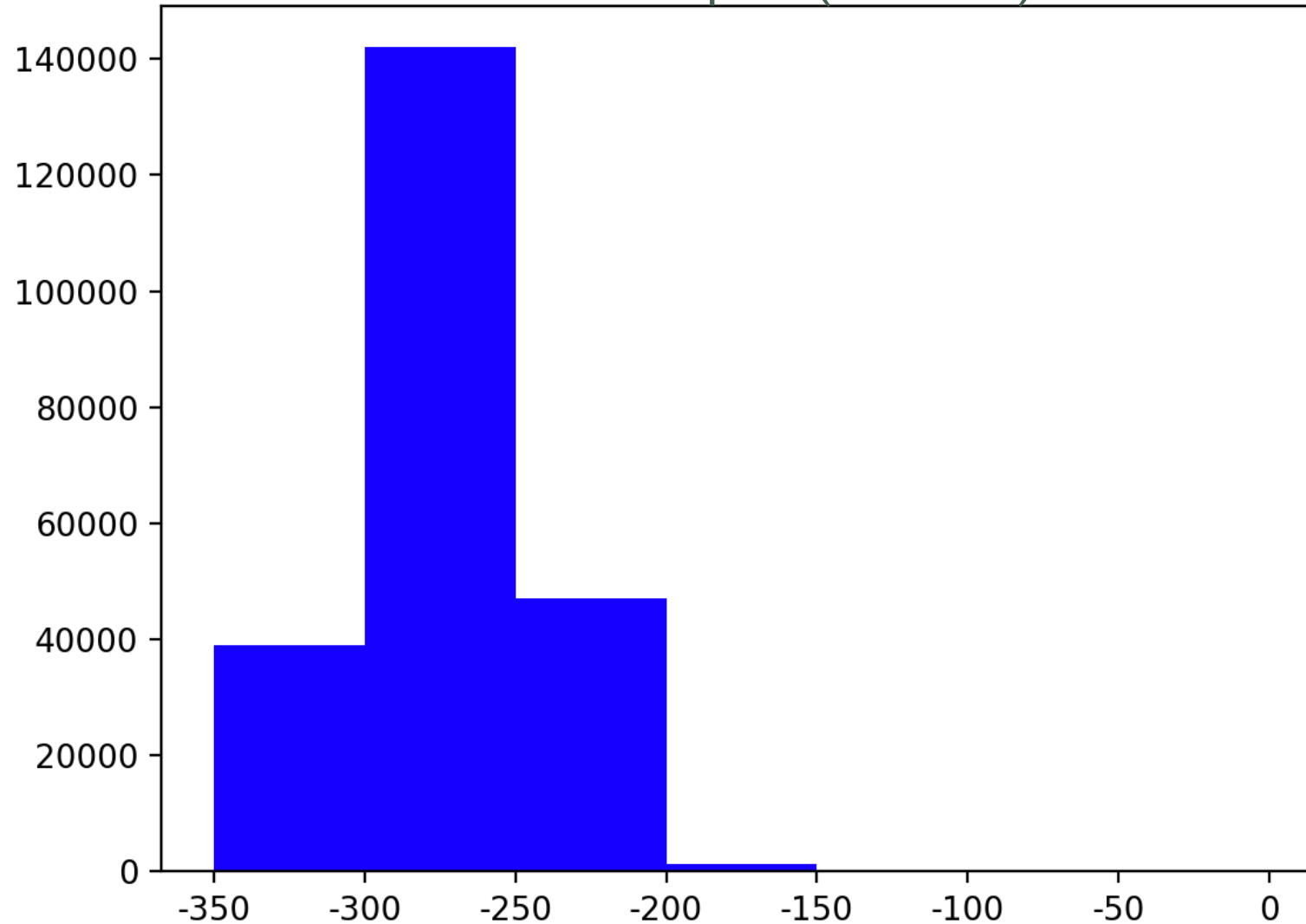
Plant Age (weeks)



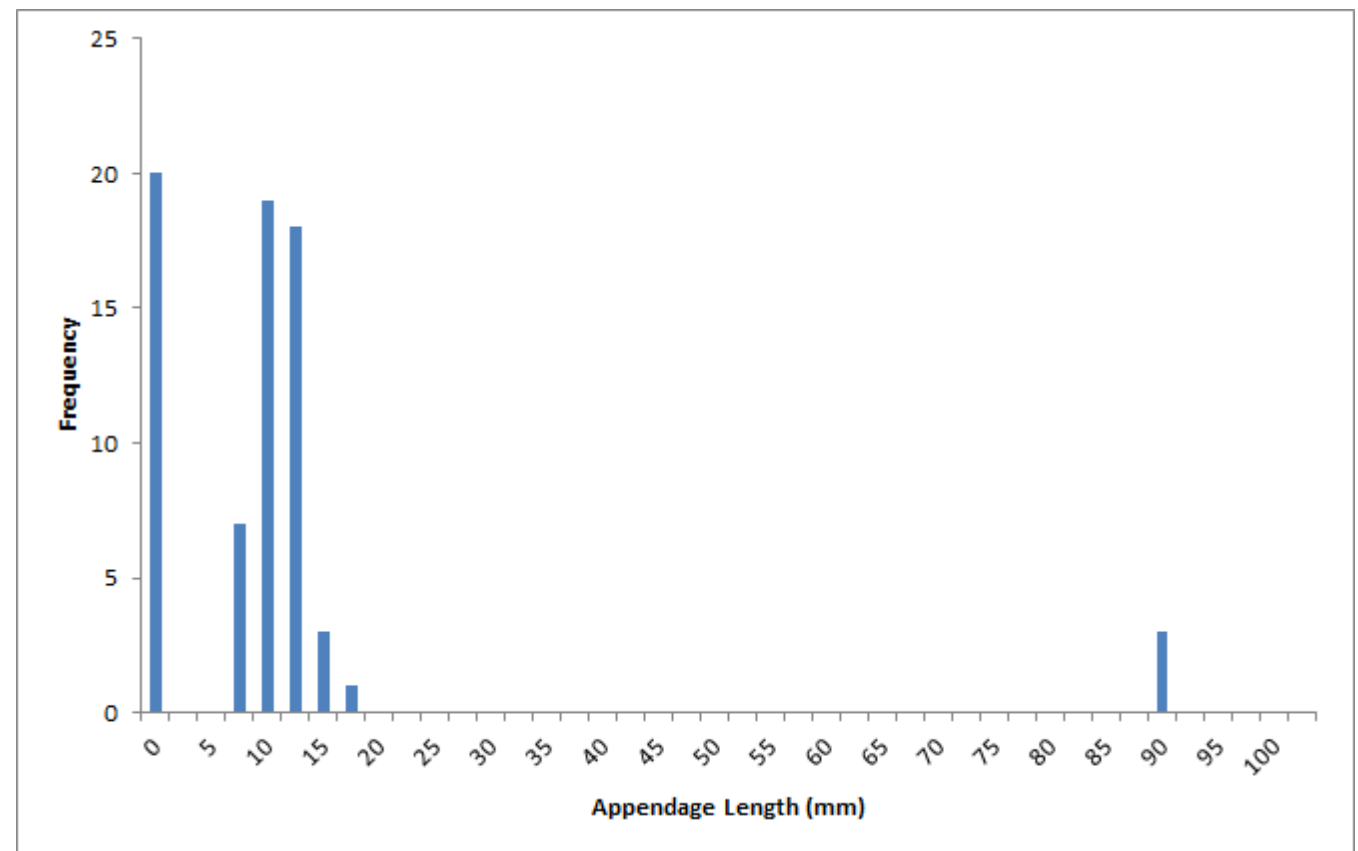
Plant Height (cm)



Heure d'arrivée au poste de tri, avant le départ (minutes)



<i>Appendage length (mm)</i>	
Mean	10.35
Standard Deviation	16.98
Kurtosis	16.78
Skewness	4.07
Minimum	0
First Quartile	0
Median	8.77
Third Quartile	10.58
Maximum	88
Range	88
Interquartile Range	10.58
Mode	0
Count	71



POINTS À RETENIR

N'attendez pas d'avoir fini l'analyse pour découvrir que la qualité des données pose problème.

Les tests univariés ne sont pas toujours suffisants.

Les visualisations peuvent aider.

Le contexte est crucial — vous aurez peut-être besoin de plus de contexte lié aux données afin de comprendre ce que vous voyez... mais, quelle que soit la situation, vous devez comprendre la qualité de l'ensemble de données.

RÉFÉRENCES

COLLECTE ET TRAITEMENT DES DONNÉES

RÉFÉRENCES

Chapman, A. *Principles and Methods of Data Cleaning – Primary Species and Species-Occurrence Data*, Copenhagen, 2005. Rapport pour le Système mondial d'informations sur la biodiversité.

van Buuren, S. *Flexible Imputation of Missing Data*, Boca Raton, CRC Press, 2012.

Orchard, T. et Woodbury, M. *A Missing Information Principle: Theory and Applications*, Proceedings of the Sixth Symposium on Mathematical Statistics and Probability, Berkeley, 1972.

Hagiwara, S. *Nonresponse Error in Survey Sampling – Comparison of Different Imputation Methods*, Ottawa, Université Carleton, 2012. Thèse de spécialisation.

Raghunathan, T., Lepkowski, J., Van Hoewyk, J. et Solenberger, P. « Une technique multidimensionnelle d'imputation multiple des valeurs manquantes à l'aide d'une séquence de modèles de régression », *Techniques d'enquête*, vol. 27, n° 1, Statistique Canada, 2001, p .85 à 95. N° 12-001 au catalogue.

Méthodes et pratiques d'enquête, Statistique Canada. N° 12-587-X au catalogue.

RÉFÉRENCES

Rubin, D.B. *Multiple Imputation for Nonresponse in Surveys*, New York, Wiley, 1987.

Kutner, M., C. Nachtsheim, J. Neter et W. Li. *Applied Linear Statistical Models*, 5^e éd., New York, McGraw-Hill/Irwin, 2004.

Green, S. et N. Salkind. *Using SPSS for Windows and Macintosh – Analyzing and Understanding Data*, 6^e éd., Upper Saddle River, Prentice Hall, 2011.

Wikipédia. « [Nettoyage de données](#) ».

Wikipédia. « [Imputation](#) ».

Wikipédia. « [Données aberrantes](#) ».

Torgo, L. *Data Mining with R*, 2^e éd., CRC Press, 2017.

McCallum, Q.E. *Bad Data Handbook*, O'Reilly, 2013.

RÉFÉRENCES

Kazil, J. et K. Jarmul. *Data Wrangling with Python*, O'Reilly, 2016.

de Jonge, E. et M. van der Loo. *An Introduction to Data Cleaning with R*, Bureau central de la statistique des Pays-Bas, 2013.

Pyle, D. *Data Preparation for Data Mining*, Morgan Kaufmann Publishers, 1999.

Weiss, S.M. et I. Indurkha. *Predictive Data Mining: A Practical Guide*, Morgan Kaufmann Publishers, 1999.

Buttrey, S.E. *A Data Scientist's Guide to Acquiring, Cleaning, and Managing Data in R*, Wiley, 2017.

Aggarwal, C.C. *Outlier Analysis*, Springer, 2013.

Chandola, V., A. Banerjee et V. Kumar. [2007], *Outlier detection: a survey*, Département d'informatique et d'ingénierie, Université du Minnesota, 2007. Rapport technique TR 07-017.

Hodge, V. et J. Austin. « A survey of outlier detection methodologies », *Artificial Intelligence Review*, vol. 22, n° 2, 2004, p. 85 à 126.

RÉFÉRENCES

Feng, L., G. Nowak, A.H. Welsh et T. O'Neill. *ImputeR: A General Imputation Framework in R*, 2014.

Steiger, J.H. [Transformations to Linearity](#). Notes de cours.

Wood, F. [Remedial Measures Wrap-Up and Transformations](#). Notes de cours.

Dougherty, J., R. Kohavi et M. Sahami. « Supervised and unsupervised discretization of continuous features », *Machine Learning: Proceedings of the Twelfth International Conference*, A. Prieditis et S. Russell (éditeurs), Morgan Kaufmann Publishers, 1995.

Orchard, T. et M. Woodbury. [A Missing Information Principle: Theory and Applications](#), *Proceedings of the Sixth Symposium on Mathematical Statistics and Probability*, Berkeley, University of California Press, 1972.

Height Percentile Calculator, by Age and Country, <https://tall.life/height-percentile-calculator-age-country/>.

Dua, D. et E. Karra Taniskidou. « Liver Disorders Data Set », *UCI Machine Learning Repository*, 2017.

RÉFÉRENCES

<http://www.roymfrancis.com/scraping-instagram-choosing-hashtags/>

Munzert, S., C. Rubba, P. Meissner et D. Nyhuis. *Automated Data Collection with R: A Practical Guide to Web Scraping and Text Mining*, Wiley, 2015.

Mitchell, R. *Web Scraping with Python: Collecting Data From the Modern Web*, O'Reilly, 2015.

https://www.w3schools.com/xml/xpath_intro.asp

<https://www.w3schools.com/>

https://fr.wikipedia.org/wiki/Extensible_Hypertext_Markup_Language

<https://medium.com/the-andela-way/introduction-to-web-scraping-using-selenium-7ec377a8cf72>

<https://pypi.python.org/pypi/selenium>

RÉFÉRENCES

Guyon, I. et A. Elisseeff. « [An Introduction to Variable and Feature Selection](#) », *Journal of Machine Learning Research*, vol. 3, mars 2003, p. 1157 à 1182.

Cawley, G.C. et N.L.C. « [Gene selection in cancer classification using sparse logistic regression with Bayesian regularization](#) », *Bioinformatics*, vol. 22, n° 19, 2006, p. 2348 à 2355.

Ambroise, C. et G.J. McLachlan. « [Selection bias in gene extraction on the basis of microarray gene-expression data](#) », *PNAS*, vol. 99, n° 10, 2002, p. 6562 à 6566.

Liu, H. et Motoda, H., éd. *Computational Methods of Feature Selection*, Chapman Hall/CRC Press.

Kononenko, I. et M. Kukar. *Machine Learning and Data Mining: Introduction to Principles and Algorithms*, Horwood Publishing, 2007. Chapitre 6.

Wikipédia. « [Lasso \(statistiques\)](#) ».

Aggarwal, C.C. *Data Mining: the Textbook*, Springer, 2016. Section 2.4.3.

RÉFÉRENCES

Robnik-Sikonja, M. et P. Savicky. [Package CORElearn](#), Comprehensive R Archive Network (CRAN). Documentation. Version 1.51.2.

Ng, A. et K. Soo. [Principal Component Analysis Tutorial](#), 15 juin 2016.

Wikipédia. « [Analyse en composantes principales](#) ».

Hastie, T., R. Tibshirani et J. Friedman. [The Elements of Statistical Learning](#), 2^e éd., Springer, 2009. Chapitre 2.

Smith, L.I. [A Tutorial on Principal Component Analysis](#), 2002.

Shlens, J. [A Tutorial on Principal Component Analysis](#), 2014. ArXiv.org.

Wikipédia. « [Nonlinear dimensionality reduction](#) ».

Leskovec, J., A. Rajaraman et J. Ullman. [Mining of Massive Datasets](#), Cambridge University Press, 2015.

RÉFÉRENCES

Skillicorn, D. *Understanding Complex Datasets: Data Mining with Matrix Decomposition*, Chapman and Hall/CRC Press, 2007.

Document « [CORElearn](#) »

Wikipédia. « [Sélection de caractéristiques](#) ».

<https://simplystatistics.org/2014/10/24/an-interactive-visualization-to-teach-about-the-curse-of-dimensionality/>

Grolemund, G. *Data Wrangling with R: how to work with the structures of your data*, 2015. Webinaire : bit.ly/wrangling-webinar.

<https://www.rstudio.com/resources/cheatsheets/>

Farrell, P. *STAT 4502 Survey Sampling Course Package*, Université Carleton, automne 2008.

RÉFÉRENCES

Lessler, J. et Kalsbeek, W. *Nonsampling Errors in Surveys*, New York, Wiley, 1992.

Oppenheim, N. *Questionnaire Design, Interviewing, and Attitude Measurement*, St. Martin's, 1992.

Hidiroglou, M., J. Drew et G. Gray. « A Framework for Measuring and Reducing non-response in Surveys », *Survey Methodology*, vol. 19, n° 1, 1993, p. 81 à 94.

Gower, A. « Questionnaire Design for Business Surveys », *Survey Methodology*, vol. 20, n° 2, 1994.

Méthodes et pratiques d'enquête, Statistique Canada. No 12-587-X au catalogue.

Boily, P., J. Schellinck, S. Hagiwara et coll. *Introduction to Quantitative Consulting*. En cours d'élaboration.