COLLECTE ET TRAITEMENT DES DONNÉES

PRÉPARATION DU TERRAIN

« 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

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.

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 justifient 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

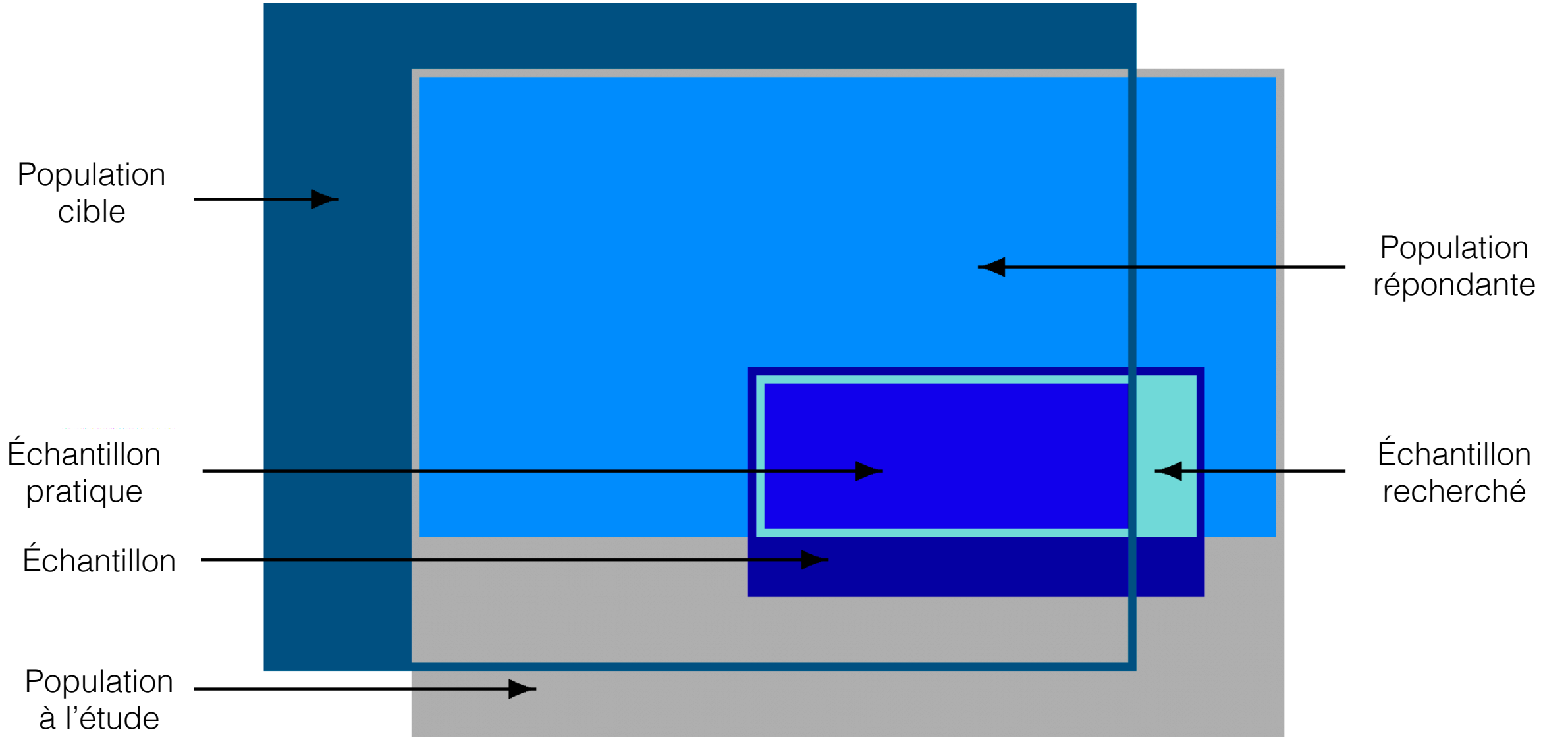
Seules ou combinées, ces deux situations conduisent à des conclusions médiocres (et **potentiellement nuisibles**).

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.


Population cible
Population répondante
Échantillon pratique
Échantillon recherché
Échantillon
Population à l'étude

ERREUR D'ENQUÊTE

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

- erreur d'échantillonnage : enquête, pas recensement
- erreur de mesure : manque d'exactitude dans la mesure des observations
- erreur de non-réponse : non-répondants présentant des différences d'observation systématiques
- erreur de couverture : 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).

PLAN D'ÉCHANTILLONNAGE

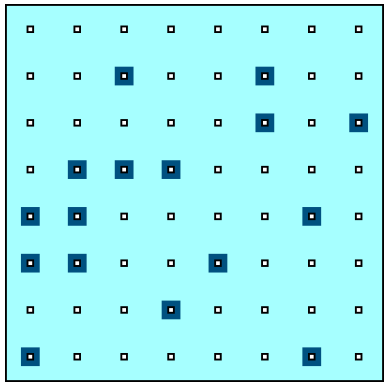
Les différents **plans d'échantillonnage** présentent des avantages et des inconvénients distincts.

Ils peuvent servir à calculer des estimations

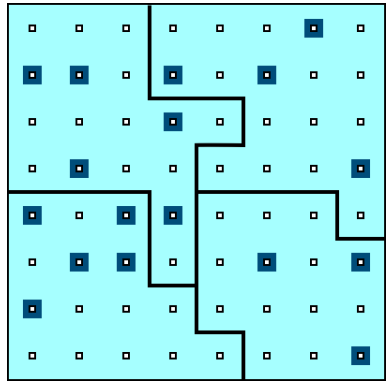
- pour diverses caractéristiques de la population : moyenne, total, proportion, ratio, différence, etc.
- pour l'IC à 95 % correspondant.

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).

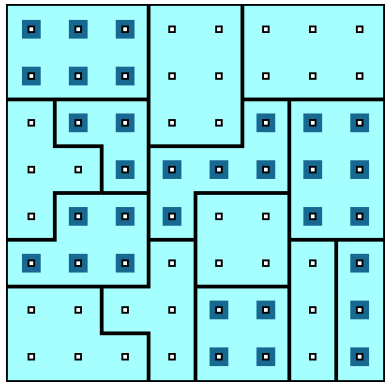
PLANS D'ÉCHANTILLONNAGE



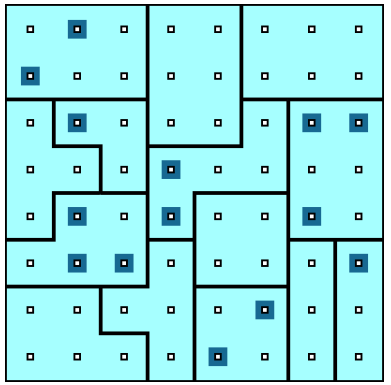
Échantillonnage aléatoire simple



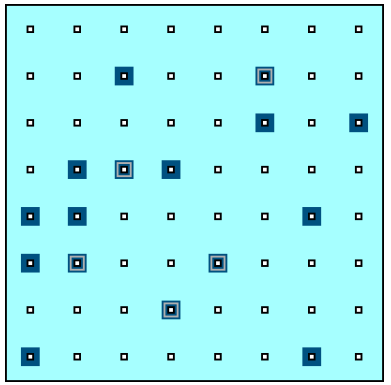
Échantillonnage aléatoire stratifié



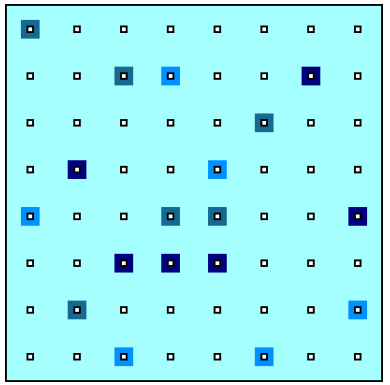
Échantillonnage en grappes



Échantillonnage à plusieurs degrés



Échantillonnage à plusieurs phases



Échantillonnage répété

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

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. »

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.

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

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**
- **Problème dans la saisie des données**
- **Entrée invalide**
- **Vide attendu**

Les méthodes analytiques ne peuvent pas toutes tenir compte aisément des observations manquantes. Il existe deux options :

- **Rejeter** l'observation manquante
- Trouver une **valeur de remplacement**

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

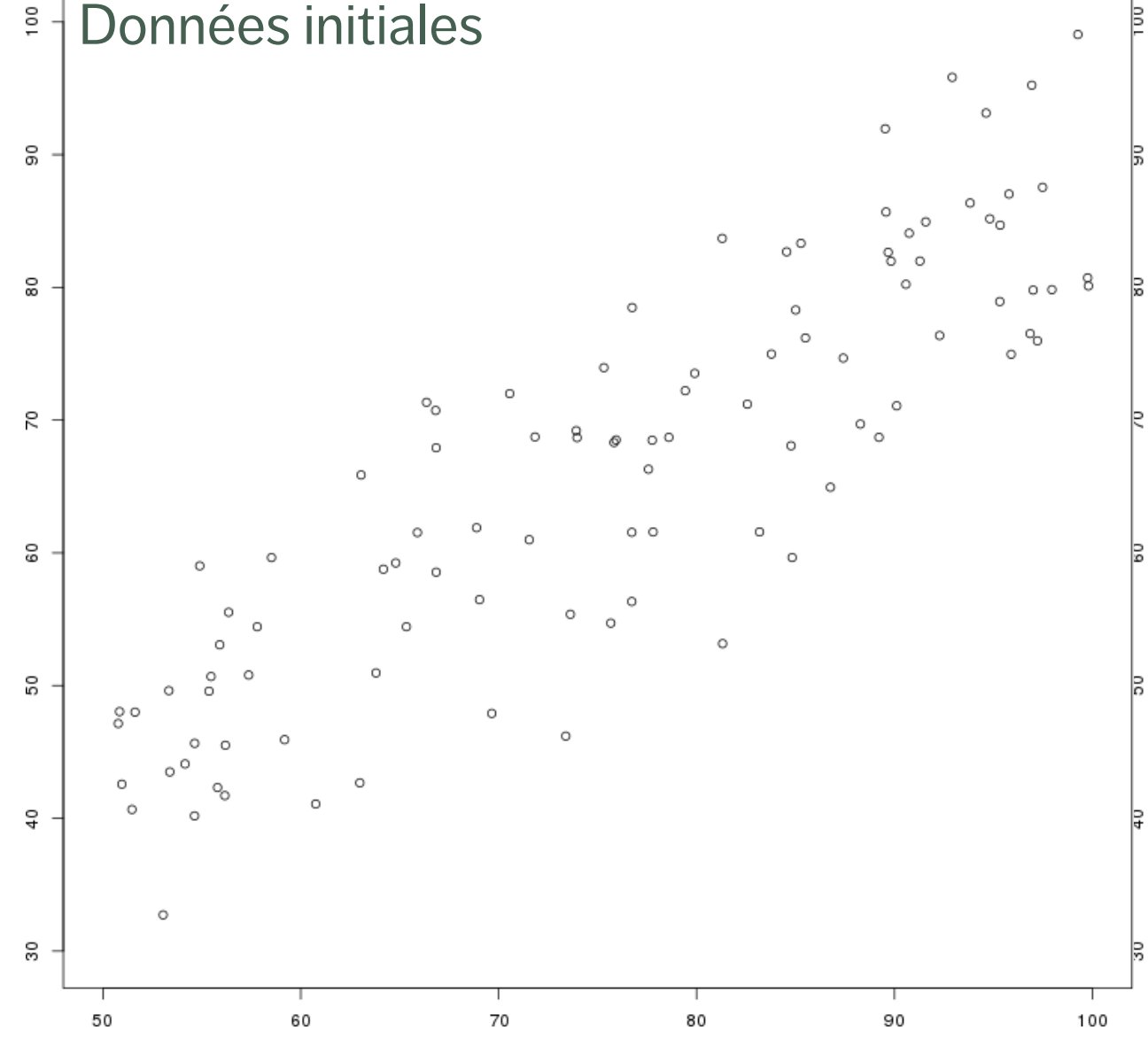
Report en avant de la dernière observation

Imputation par la méthode du plus proche voisin *k*

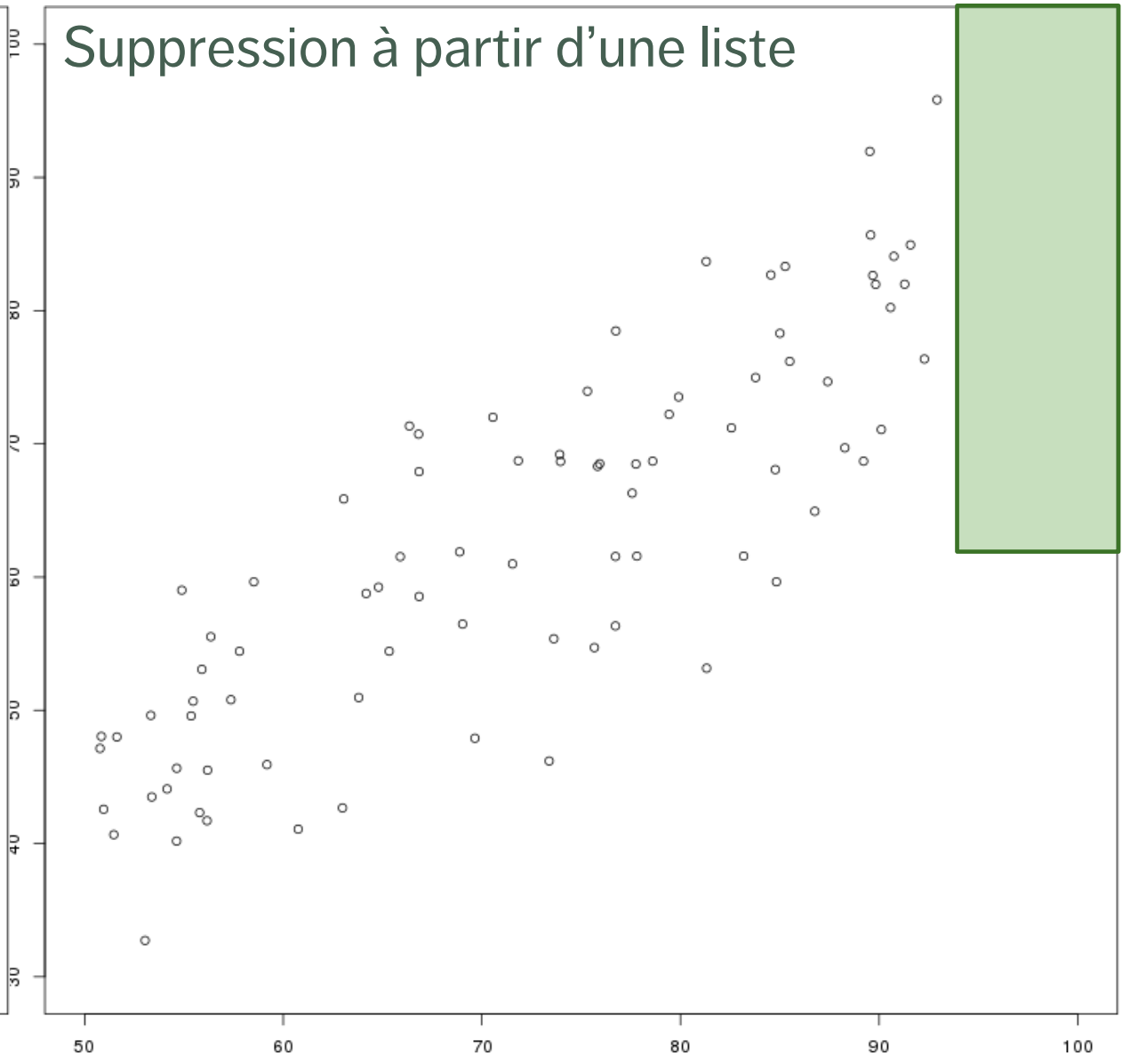
Imputation multiple

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

Données initiales

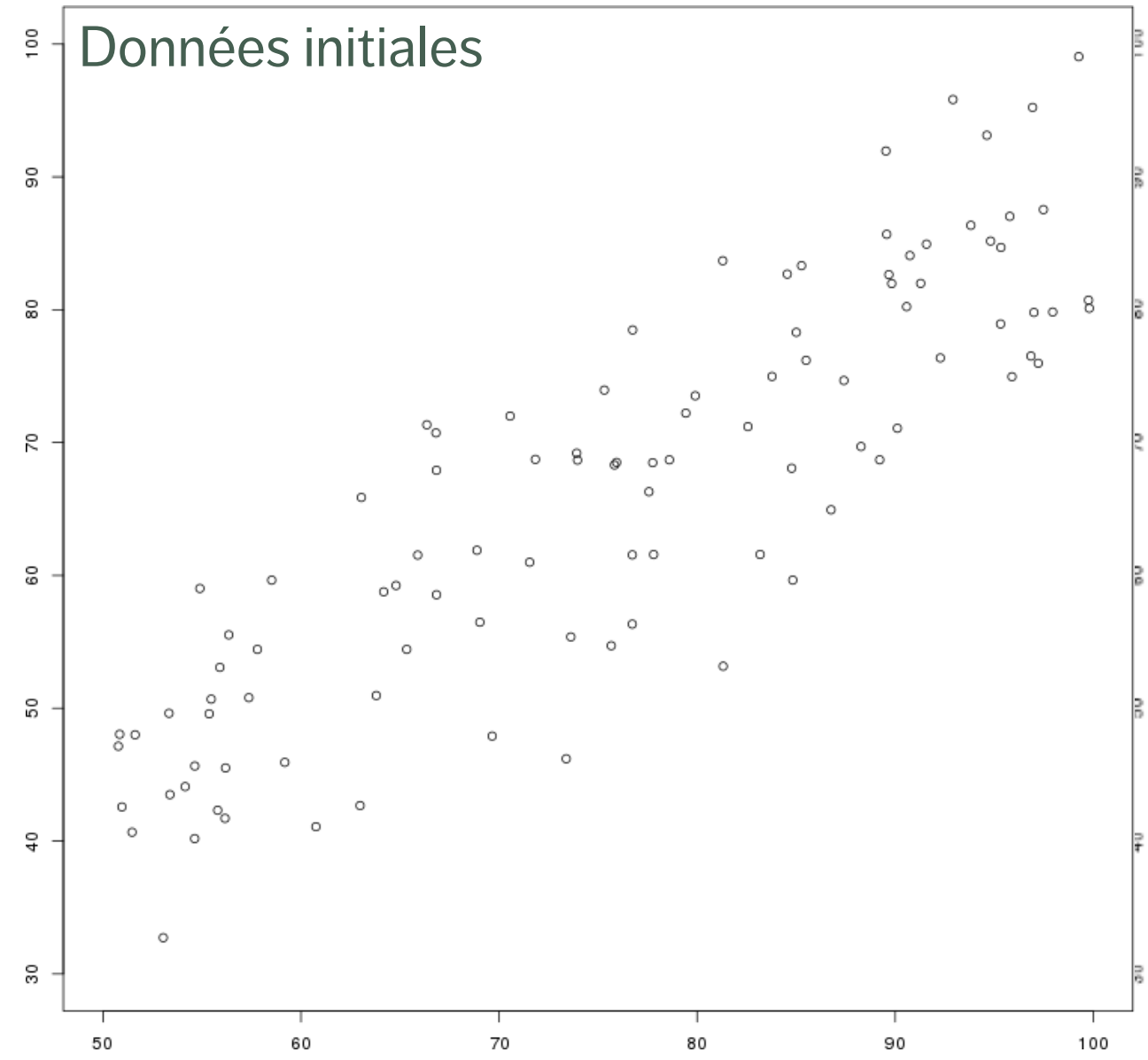


Suppression à partir d'une liste

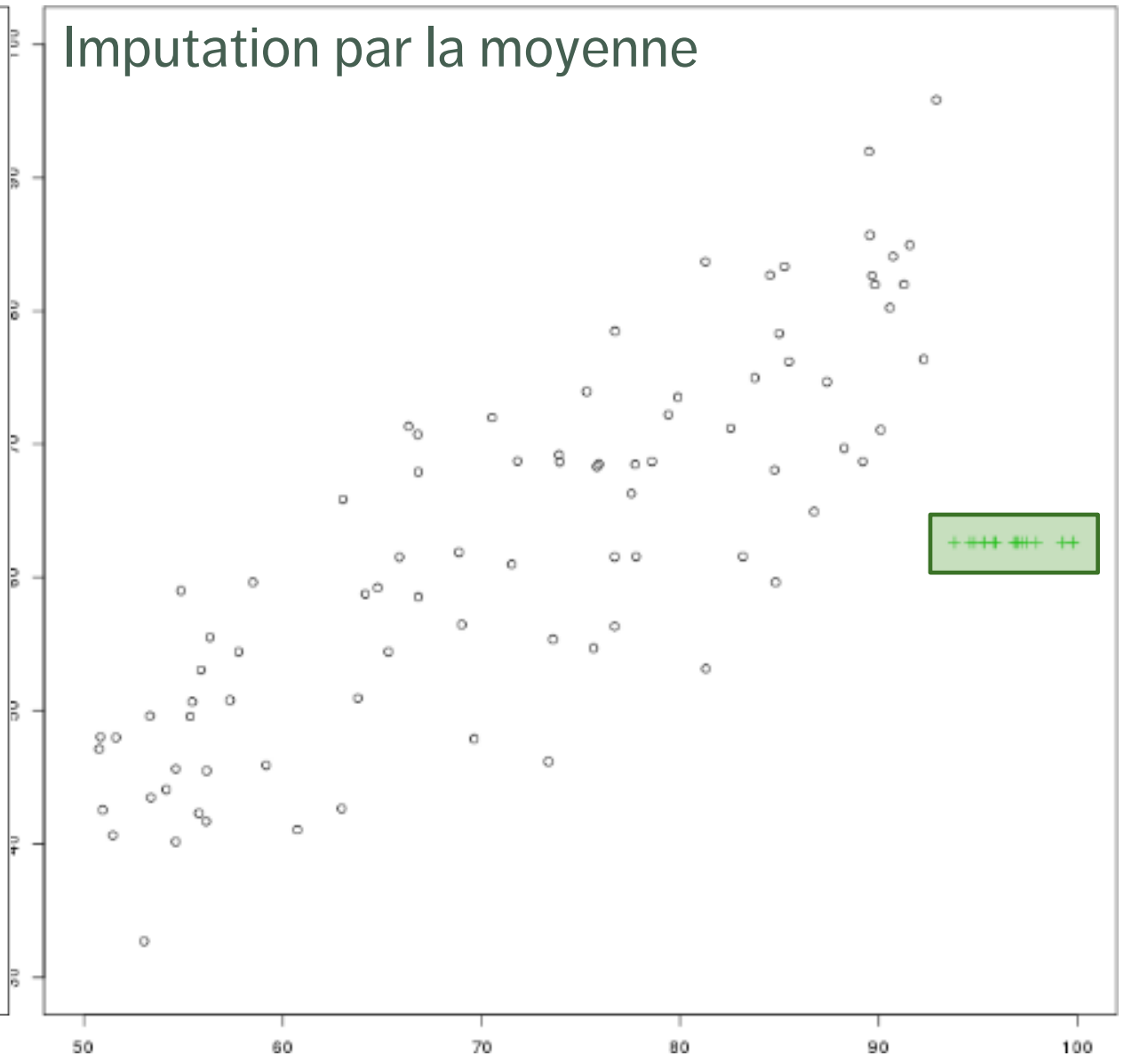


Données artificielles : Les valeurs y de tous les points pour lesquels $x > 92$ ont été effacées par erreur.

Données initiales

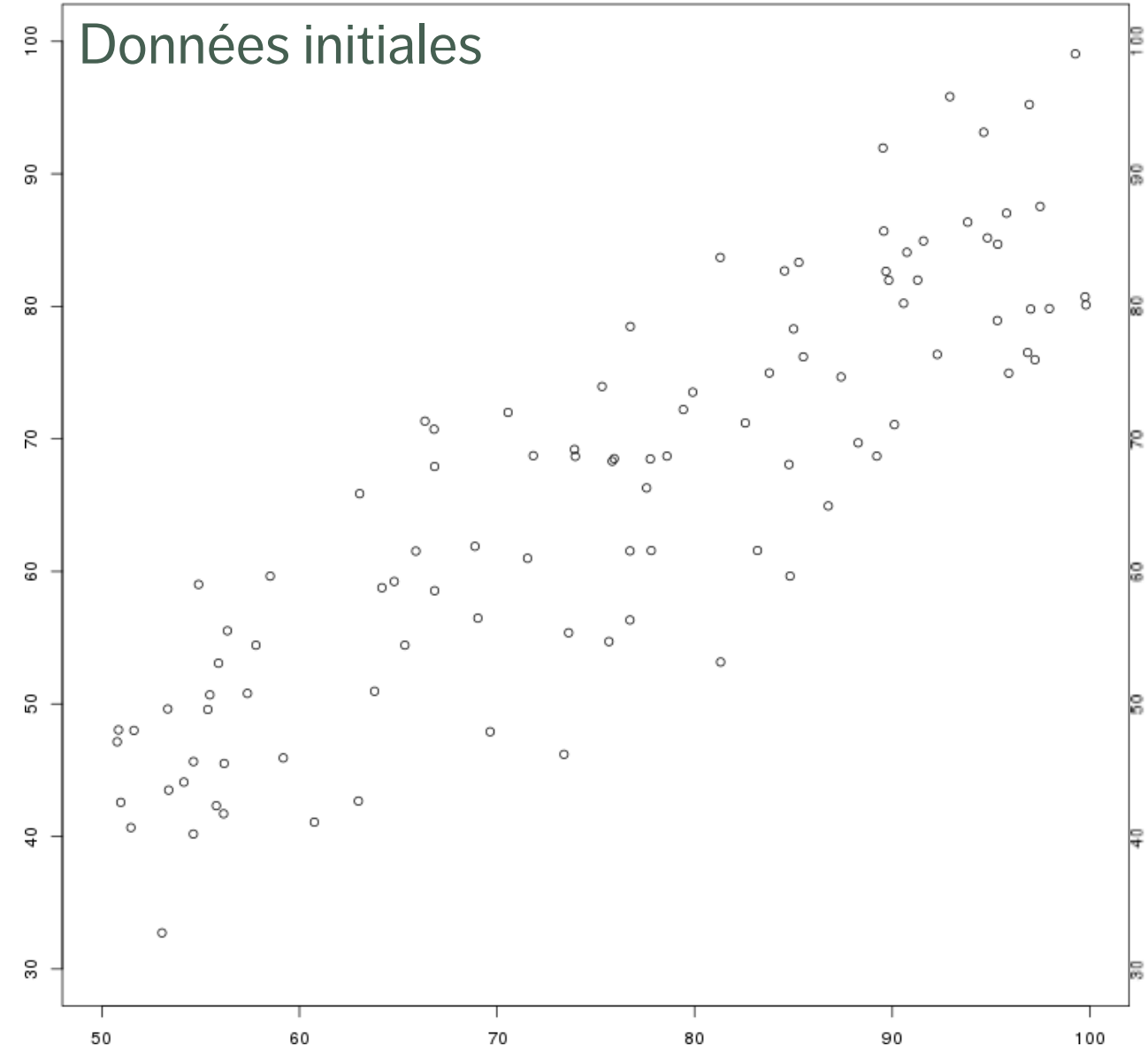


Imputation par la moyenne



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

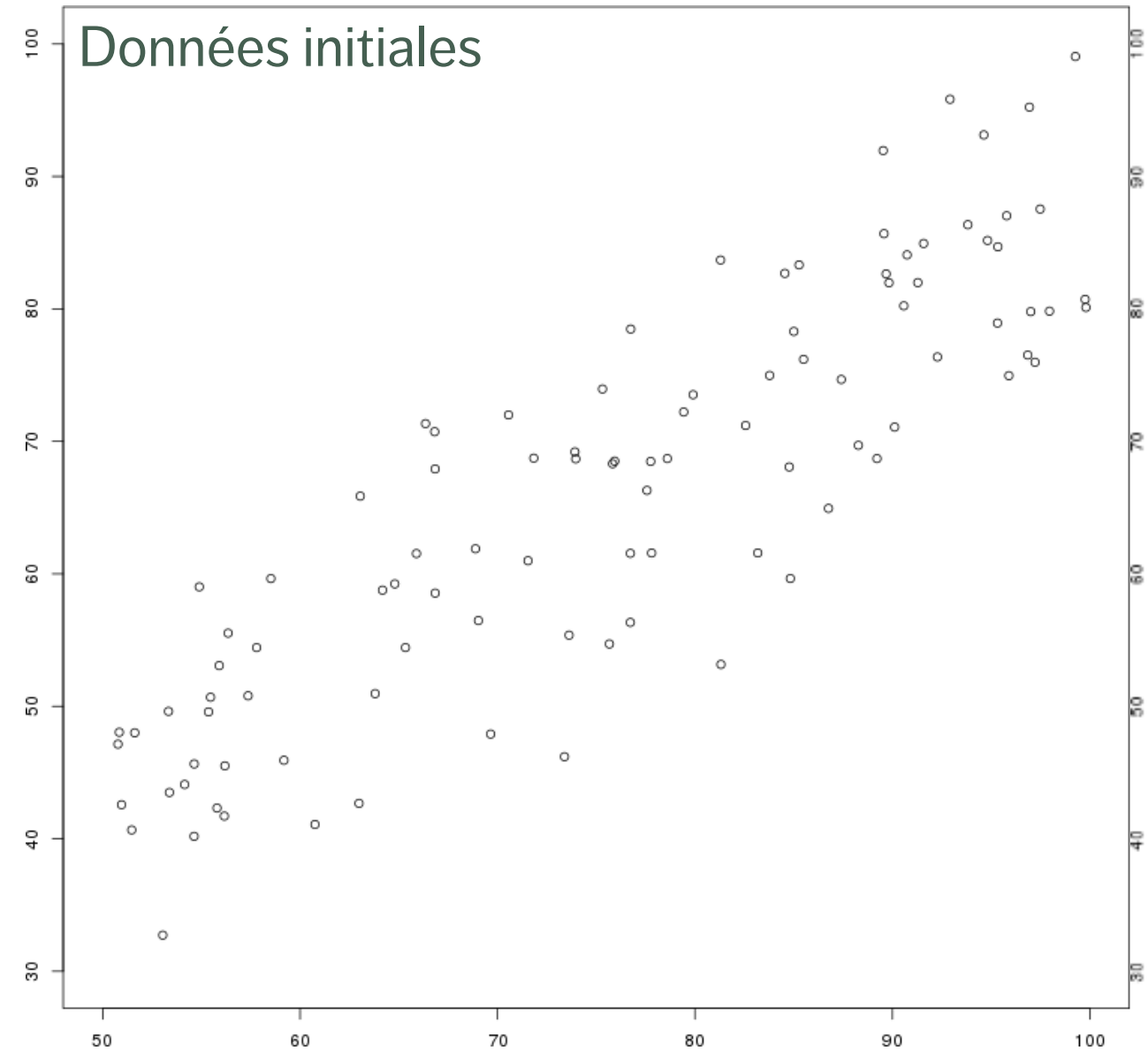
Données initiales



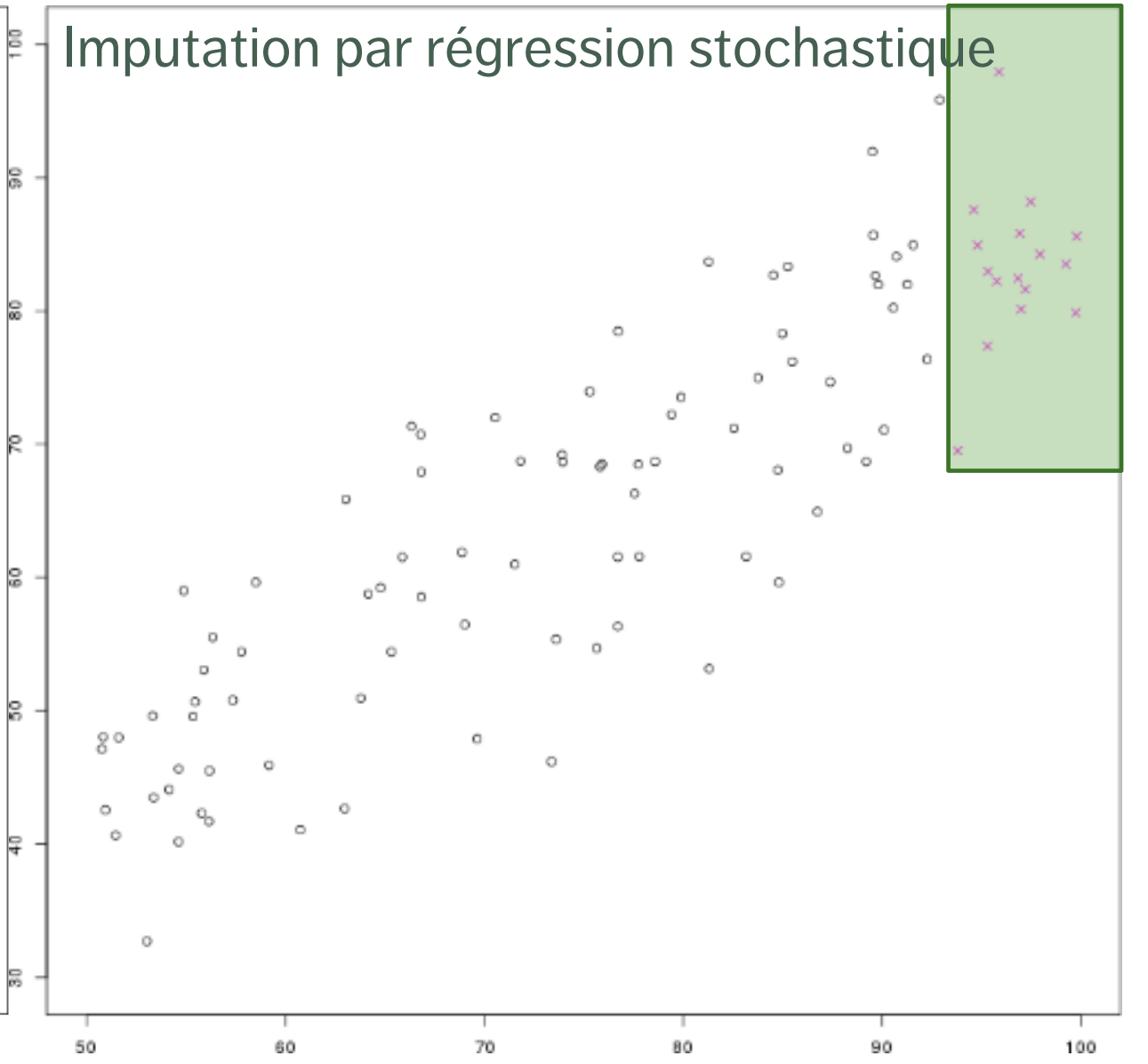
Imputation par régression

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

Données initiales



Imputation par régression stochastique



OBSERVATIONS SPÉCIALES

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.

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.

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 infaillible**, 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)

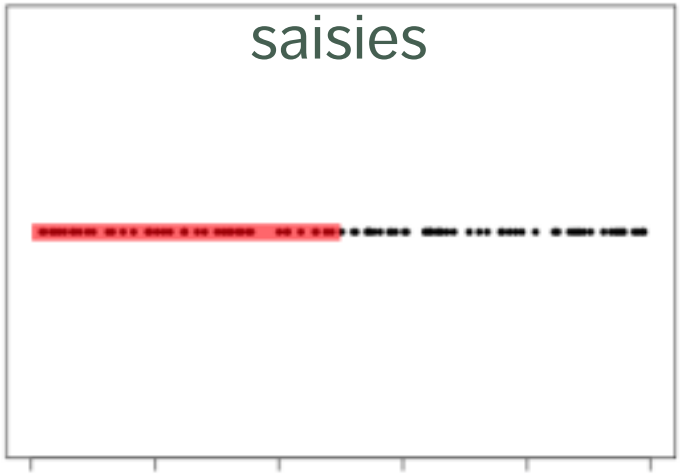
Les **méthodes non supervisées** n'ont pas recours à des renseignements externes :

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

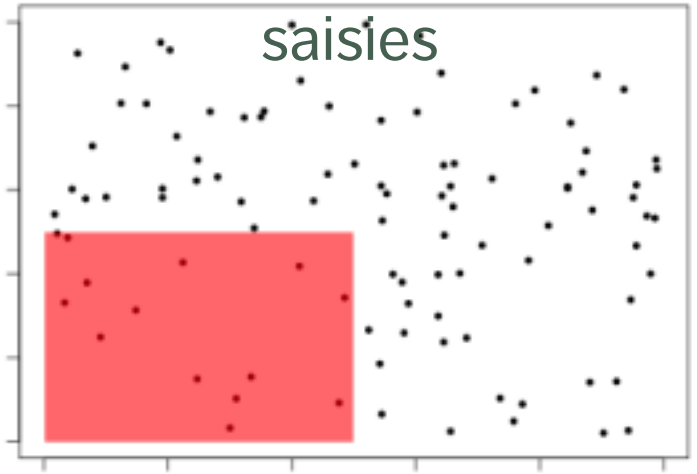
Il existe aussi des **méthodes semi-supervisées**.

FLÉAU DE LA DIMENSION

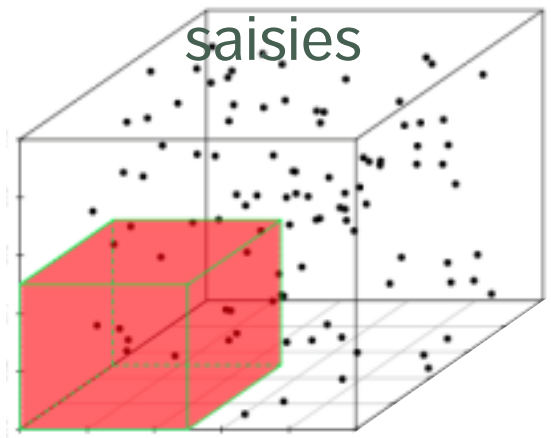
42 % des données sont saisies



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 $[0,1/2]^d, d = 1,2,3.$

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 multicolinéarité, etc.)
- Réduction de la dimension (nombre de variables ≫ nombre d'observations)

Approches :

- Filtrage et méthode enveloppante
- Non supervisé ou supervisé

Réduction de la dimension : PCA, UMAP, apprentissage des variétés, etc.

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.

MISE À L'ÉCHELLE DES DONNÉES

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

La **standardisation** crée une variable de moyenne est 0 et d'écart-type 1 : $Y_i = \frac{X_i - \bar{X}}{s_X}$

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

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

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.