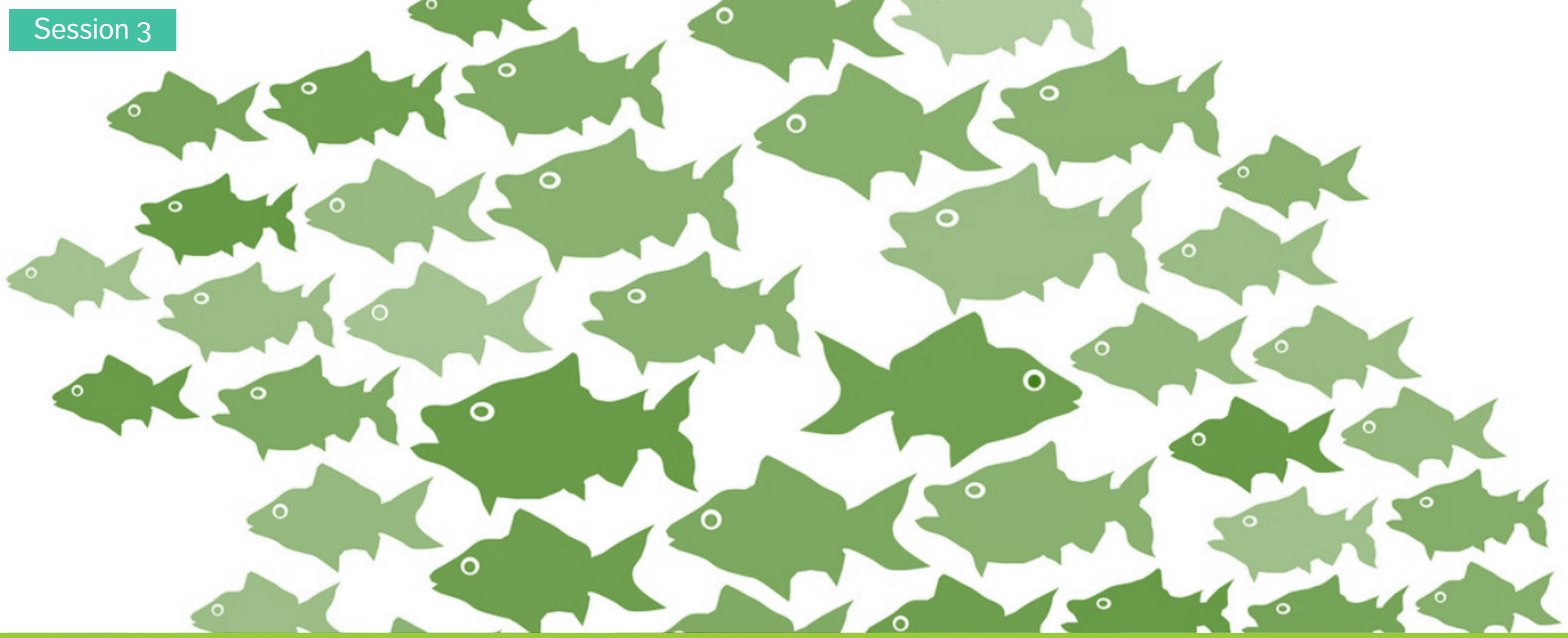


9. Les observations anormales

Les observations anormales

En pratique, une **observation anormale** peut se présenter comme

- un **"mauvais" objet/mesure :** artefacts de données, fautes, valeurs mal imputées, etc. ;
- une **observation mal classée :** selon les modèles de données existants, l'observation aurait dû être étiquetée différemment ;
- une observation dont les mesures se trouvent dans les **queues de distribution** d'un nombre suffisamment grand d'éléments ;
- un **inconnu inconnu** : un type d'observation totalement nouveau dont l'existence était jusqu'alors insoupçonnée.

Les observations anormales

Une observation peut être anormale dans un contexte, mais pas dans un autre

- un homme adulte de 1.80 m se situe dans le 86^{e} percentile pour les hommes canadiens (grand, mais pas inhabituel).
- en Bolivie, le même homme serait dans le 99.9^{e} percentile (très grand, inhabituel)

La détection des anomalies soulève des **questions intéressantes** pour les analystes et les experts en la matière : dans ce cas, pourquoi existe-t-il un écart aussi important entre les deux populations ?

Les valeurs aberrantes (“outliers”)

Les **observations aberrantes** sont des observations qui sont **atypiques** par rapport aux :

- autres caractéristiques à même l'unité (“*within units*”), et
- valeurs des caractéristiques des autres unités (“*between-units*”)

Les valeurs aberrantes sont des observations qui **ne ressemblent pas aux autres cas** ou qui **contredisent des dépendances** ou des règles **connues**.

Une étude minutieuse est nécessaire pour déterminer si ces valeurs aberrantes doivent être conservées ou supprimées de l'ensemble de données.

La détection des anomalies

Les valeurs aberrantes peuvent être anormales par rapport à une variables de l'unité, ou en combinaison.

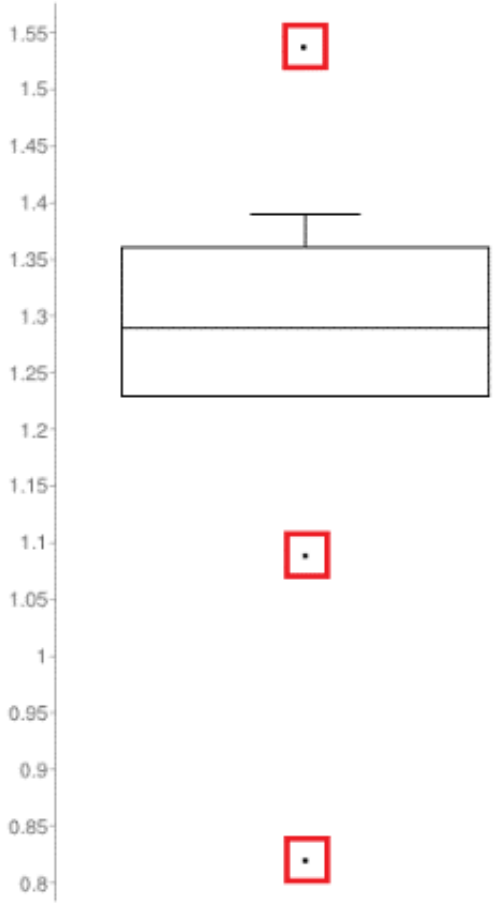
Les anomalies sont par définition **peu fréquentes**, et généralement entourées d'**incertitude en** raison de la petite taille des échantillons.

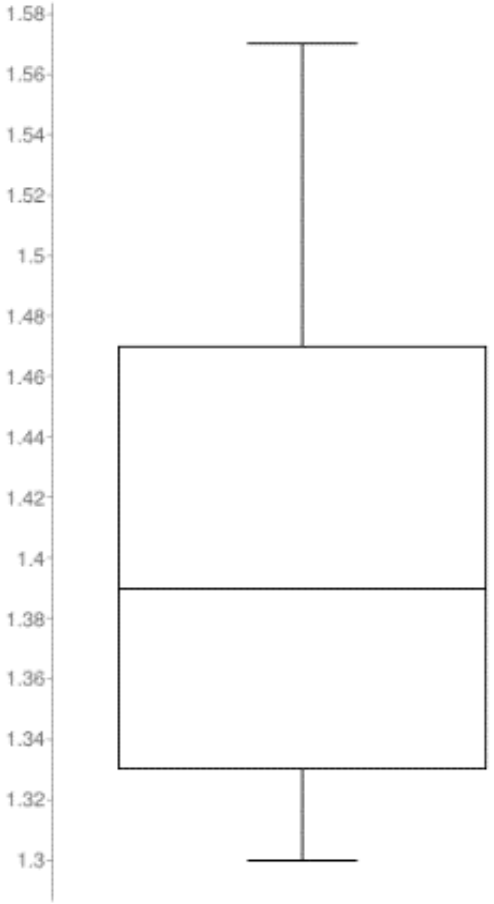
Il est difficile de différencier les anomalies du bruit ou des erreurs de saisie.

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

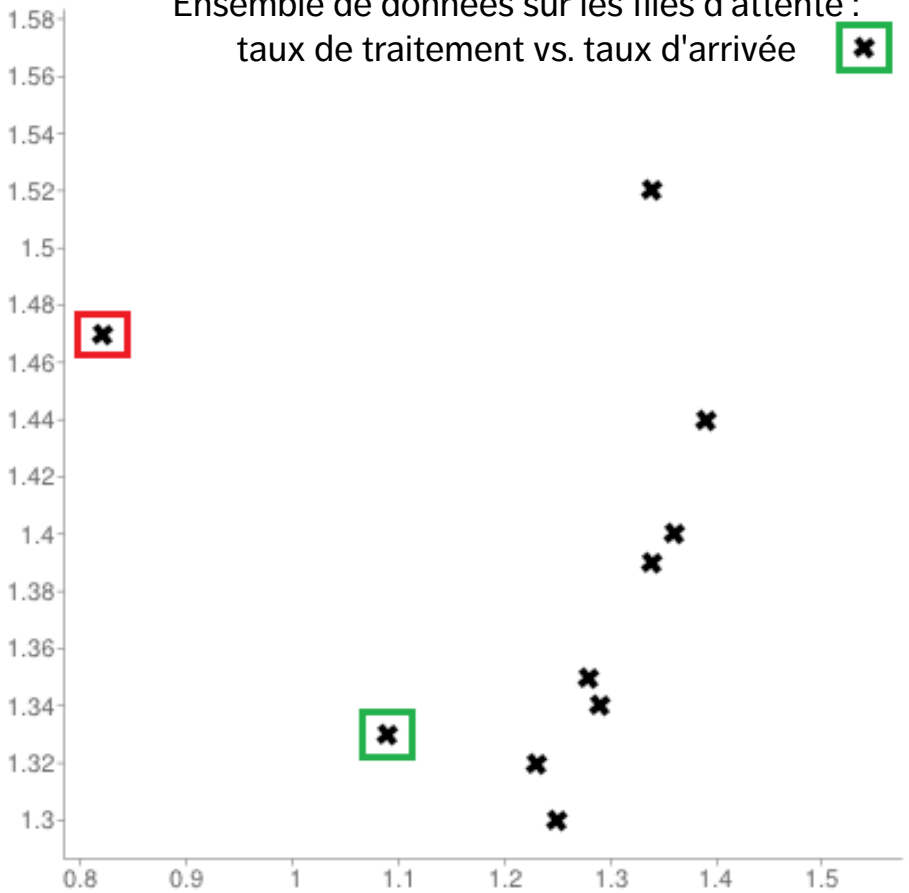
Les anomalies liées à des activités malveillantes sont généralement **déguisées**.

La détection des anomalies





Ensemble de données sur les files d'attente : taux de traitement vs. taux d'arrivée



La détection des anomalies

Il y a de nombreuses méthodes pour identifier les observations anormales ; **aucune d'entre elles n'est infaillible** et il faut faire preuve de discernement

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

- **observations périphériques**

  box-plots, nuages de points, matrices de nuages de points, tour 2D, distance de Cooke, tracés qq normaux

- **données influentes**

  un certain niveau d'analyse doit être effectué (effet de levier)

Attention : si les observations anormales ont été retirées de l'ensemble de données, des unités auparavant "régulières" peuvent devenir anormales !

Algorithmes de détection d'anomalie

Les **méthodes supervisées** utilisent un historique d'observations anormales étiquetées :

- l'expertise du domaine est requise pour étiqueter
- tâche de classification ou de régression
- problème d'occurrence rare

		Prédictions	
		Normales	Anormales
Réalité	Normales	*VN*	*FP*
	Anormales	*FN*	*VP*

Les **méthodes non supervisées** n'utilisent pas d'informations externes :

- méthodes et tests traditionnels
- problème de regroupement ou de règles d'association

Les algorithmes de détection

Le coût des erreurs de classification est souvent supposé être symétrique, ce qui peut conduire à des résultats **techniquement corrects, mais inutiles**.

Par exemple, la grande majorité des passagers aériens (99.99%+) n'emportent pas d'armes ; un modèle qui prédit qu'aucun passager ne fait passer une arme en fraude serait précis à 99.99%+, mais il passerait à côté du problème.

Pour l'**agence de sécurité**, le coût de penser à tort qu'un passager :

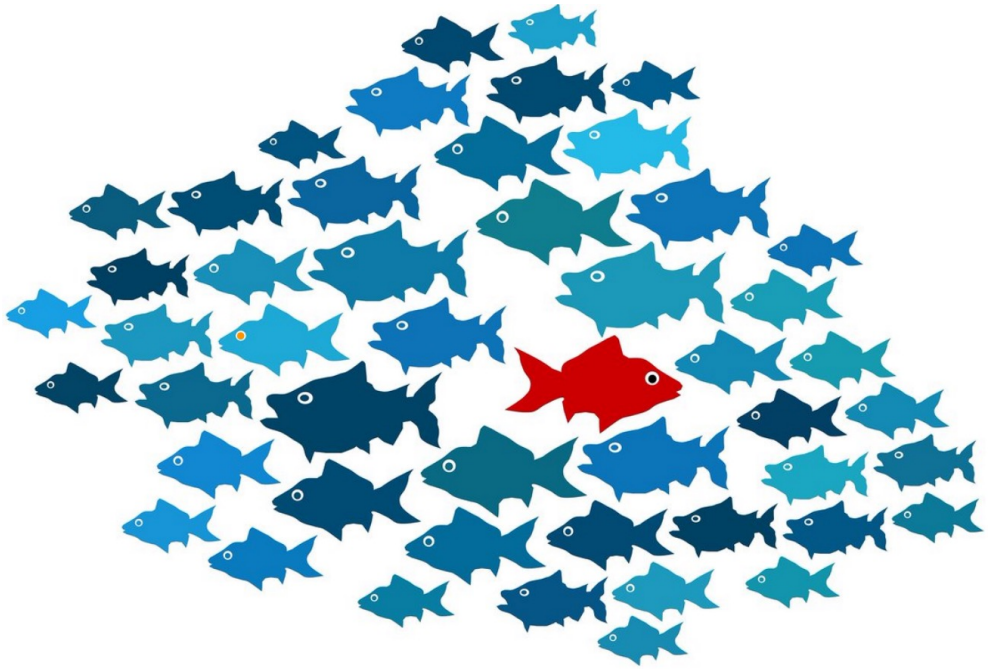
- introduit clandestinement une arme ⟹ coût d'une seule fouille
- ne fait pas passer une arme en fraude ⟹ catastrophe (potentiellement)

Les personnes injustement visées auront un point de vue différent à ce sujet !

Les algorithmes de détection

Si tous les participants à un atelier à l'exception d'un seul, peuvent visionner les conférences par vidéo, cette personne, cette connexion Internet et cet ordinateur sont **anormaux**, car ils ne se comportment pas comme les autres.

Mais cela **NE SIGNIFIE PAS** nécessairement que le comportement différent est celui qui nous intéresse...



Tests simples de valeurs aberrantes

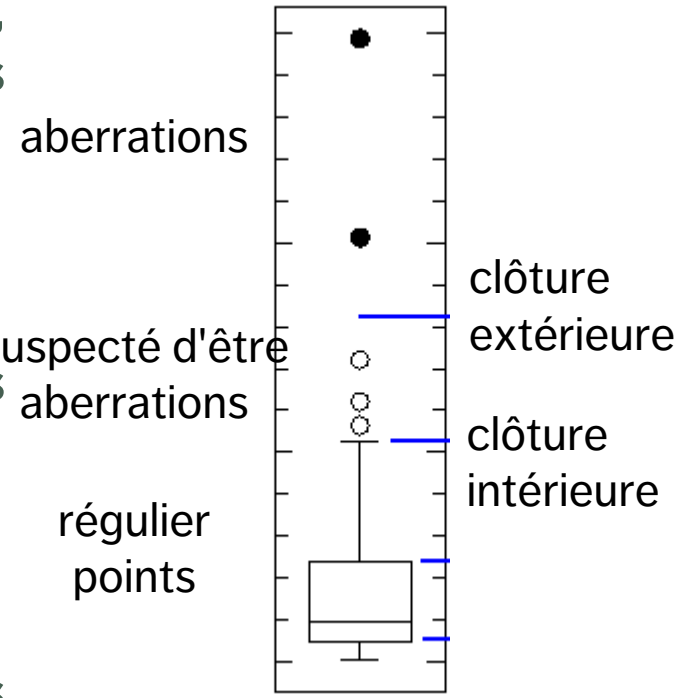
Test Boxplot de Tukey : pour les données normalement distribuées, les observations régulières se situent généralement entre les **clôtures intérieures :**

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

Les **valeurs aberrantes suspectes** se situent entre les **clôtures intérieures** et les **clôtures 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 **clôtures extérieures**.



Tests simples de valeurs aberrantes

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

La **distance de Mahalanobis** (liée à l'effet de levier) peut être utilisée pour trouver des valeurs aberrantes multidimensionnelles (lorsque les relations sont linéaires).

Autres tests simples :

- **Grubbs** (univarié)
- **Tietjen-Moore** (pour un nombre spécifique de valeurs aberrantes)
- **écart généralisé extrême studentisé** (pour un nombre inconnu de valeurs aberrantes)
- **chi-deux (**les valeurs aberrantes affectant la qualité de l'ajustement)

Test sophistiqués des valeurs aberrantes

- **DBSCAN**, **OR$_h$** , et **LOF** (détection non supervisée des valeurs aberrantes)
- méthode **rang-puissance** (détection supervisée des valeurs aberrantes)
- méthodes **basées sur la distance** ou la **densité** (avec des mesures de distance exotiques)
- **autoencodeurs et erreur de reconstruction** (méthode d'apprentissage profond)
- méthodes d'**occurrences rares** (suréchantillonnage, sous-échantillonnage, CREDOS, PN, SHRINK, SMOTE, DRAMOTE, SMOTEBoost, RareBoost, MetaCost, AdaCost, CSB, SSTBoost, etc.)
- **AVF**, algorithmes **Greedy** (données catégoriques)
- **PCA**, **DOBIN** et autres méthodes de **projection** (pour les données à haute dimension)
- méthodes **subspatiales** et méthodes d'**ensemble**

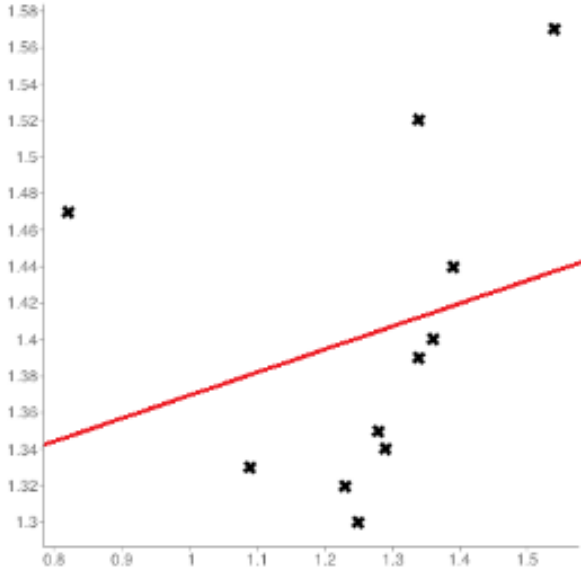
Observations influentes

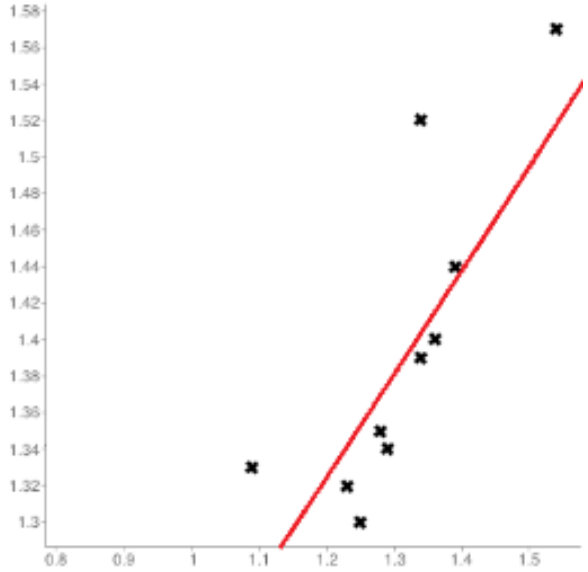
Les **observations influentes** sont des observations dont l'absence entraîne des résultats d'analyse **nettement différents**.

Lorsque des observations influentes sont identifiées, des **mesures correctives** (telles que des transformations de données) peuvent être nécessaires pour minimiser leurs effets indus.

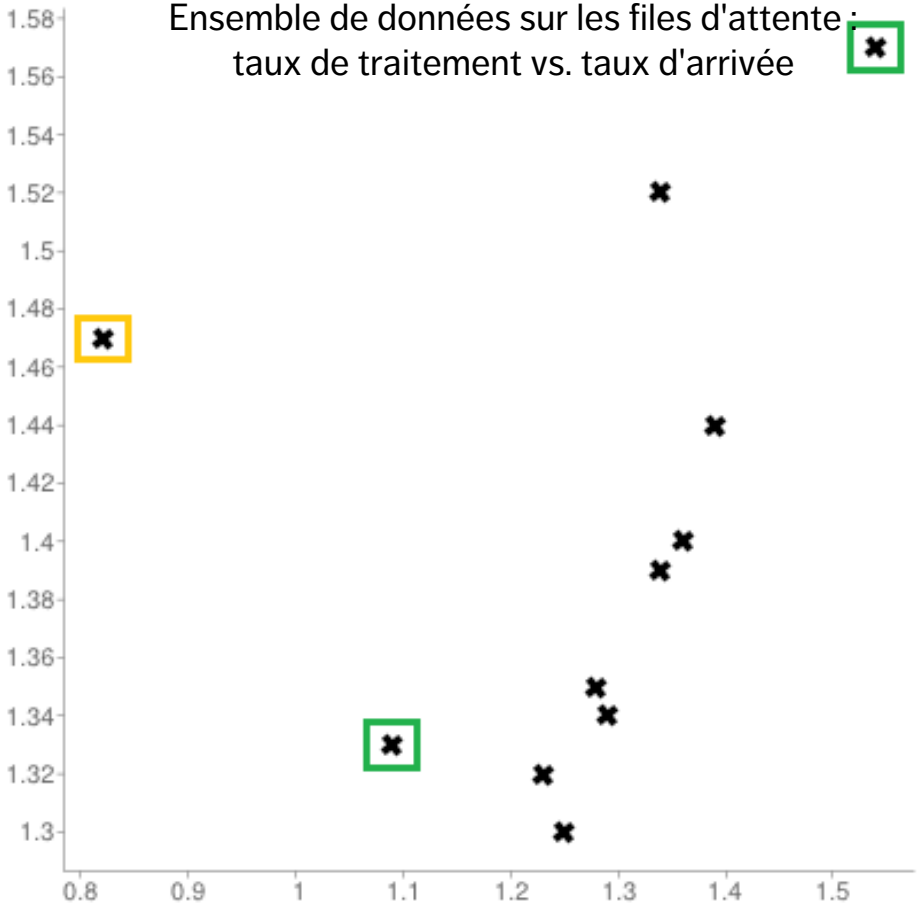
Les valeurs aberrantes peuvent être des observations influentes ; les observations influentes ne sont pas nécessairement des valeurs aberrantes (et *vice-versa)*.

Observations influentes





Ensemble de données sur les files d'attente : taux de traitement vs. taux d'arrivée



Remarques

L'identification des observations influentes est un **processus itératif** car les différentes analyses doivent être exécutées à plusieurs reprises.

L'identification et la suppression entièrement automatisées des observations anormales **ne sont PAS recommandées**.

Utilisez des transformations de données si les données **ne** sont **PAS normalement distribuées**.

Le fait qu'une observation soit une valeur aberrante ou non dépend de **divers facteurs** ; les observations qui finissent par être influentes dépendent de l'**analyse spécifique à effectuer**.

Lectures suggérées

Les observations anormales

Data Understanding, Data Analysis, Data Science
Data Preparation

Anomalous Observations

- Anomaly Detection
- Outlier Tests
- Visual Outlier Detection

***Anomaly Detection and Outlier Analysis** (avancé)

Exercices

Les observations anormales

1. Recréez le processus de détection des anomalies utilisé dans Example: Algae Bloom.
2. Trouvez les observations anormales dans les ensembles de données cities.txt et grades (le cas échéant).
3. Trouvez les observations anormales dans un ensemble de données de votre choix.