Data analysis for unbiased machine learning

Session 2:
Understand data to avoid bias

Alice HELIOU and Vincent THOUVENOT
Laboratoire de Data Science de CortAIx Labs

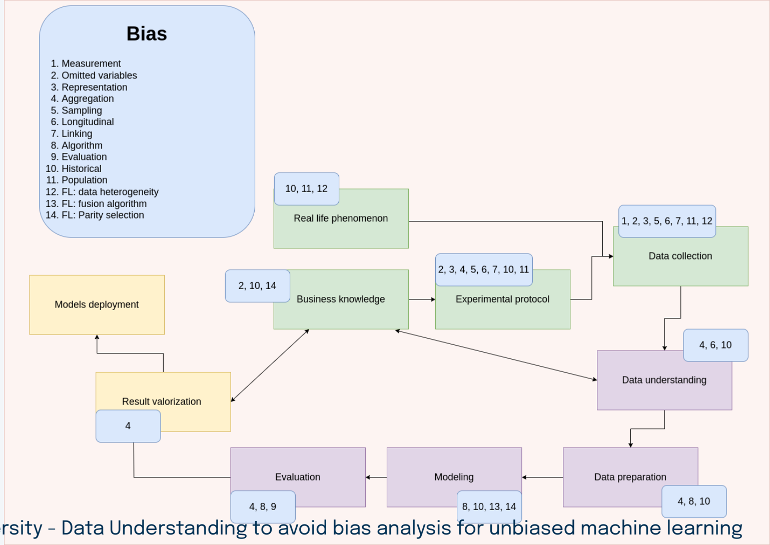
THALES
Building a future we can all trust

Content

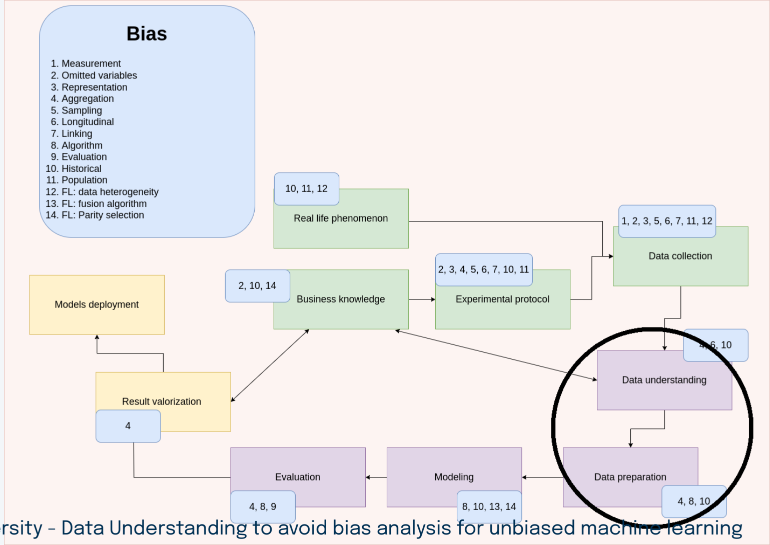
1. Introduction
2. Causalité – adapted from https://fairmlbook.org/pdf/causal.pdf
3. Analyse descriptive
4. Valeur manquante
5. Biais inconscients

Introduction

Lesson position



Lesson position



Les types de variables

Il existe 4 types de variables.

Variable Qualitative

Nominale qui qualifie
Catégorielle/Textuelle
Type de cheveux

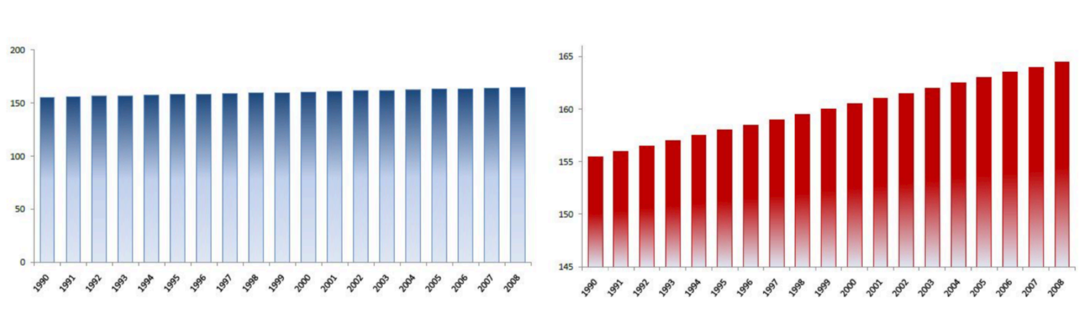
Ordinale qui classe
Echelle de Likert
Fréquence de sortie

Variable Quantitative

Discrète qui compte
Nombre d'enfant à charge

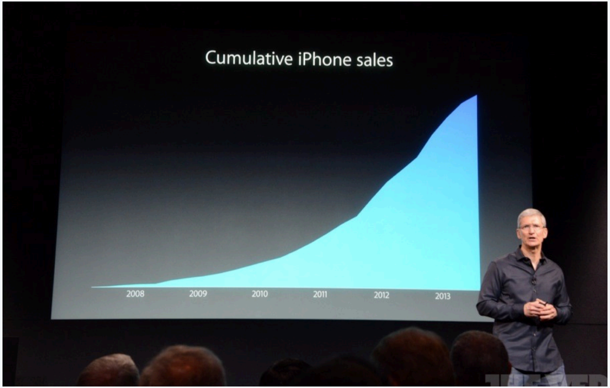
Continue qui mesure
Intervalle/Rapport
Température/Volume de vente

Importance of scale

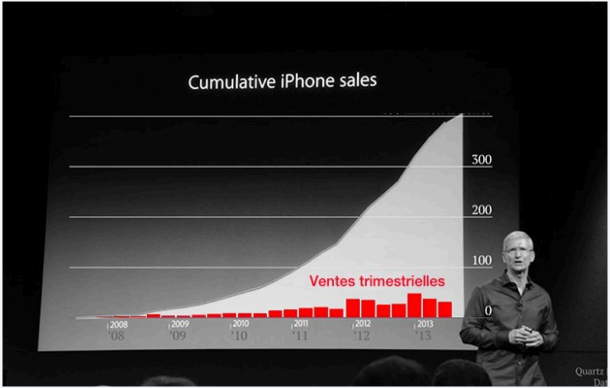


Importance of scale

Importance of scale



Importance of scale



Causalité – adapted from https://fairmlbook.org/pdf/causal.pdf

Les observations

Ce sont des faits que l'on peut tirer directement des données.

Exemple :

Les conducteurs de 18 ans causent-ils plus d'accidents que ceux de 20 ans ?

Les observations

Ce sont des faits que l'on peut tirer directement des données.

Exemple :

Les conducteurs de 18 ans causent-ils plus d'accidents que ceux de 20 ans ?

La réponse consiste en une probabilité conditionnelle. si l'ensemble de données est bien collecté et suffisamment important alors elle sera proche de la 'vérité'.

La causalité

Souvent la question à laquelle on souhaite répondre n'est pas aussi simple.
La réponse ne pas être obtenue en un seul calcul à partir des données.
Exemple :
Est-ce que le nombre d'accidents diminuerait si on augmentait de 2 ans l'age minimal pour obtenir le permis ?

La causalité

Souvent la question à laquelle on souhaite répondre n'est pas aussi simple.
La réponse ne pas être obtenue en un seul calcul à partir des données.
Exemple :
Est-ce que le nombre d'accidents diminuerait si on augmentait de 2 ans l'age minimal pour obtenir le permis ?

Là la réponse est bien plus complexe, il y a de nombreux autres facteurs qui peuvent intervenir, et de nombreuses hypothèses possibles

- peut être que les conducteurs de 20 ans ont moins d'accidents car ils ont en moyenne plus d'expérience => on peut alors prendre en compte l'expérience

La causalité

Souvent la question à laquelle on souhaite répondre n'est pas aussi simple.
La réponse ne pas être obtenue en un seul calcul à partir des données.
Exemple :
Est-ce que le nombre d'accidents diminuerait si on augmentait de 2 ans l'age minimal pour obtenir le permis ?

Là la réponse est bien plus complexe, il y a de nombreux autres facteurs qui peuvent intervenir, et de nombreuses hypothèses possibles

- peut être que les conducteurs de 20 ans ont moins d'accidents car ils ont en moyenne plus d'expérience => on peut alors prendre en compte l'expérience
- peut être que les novices de 20 ans sont principalement des personnes très prudentes et qui ont donc moins d'accidents que les novices de 18 ans
- ou peut être que ceux qui conduisent à 18 ans sont poussés à conduire car ils vivent par exemple en milieu rural où il y a peu de transport public, moins d'éclairage et des vitesses plus élevées.

Le danger des probabilités conditionnelles

Exemple adapté du papier: Peter J Bickel, Eugene A Hammel, J William O'Connell, et al. Sex bias in graduate admissions: Data from berkeley. Science, 1975.

On considère les taux d'admission par genre a une université ayant 2 départements avec des processus de sélection distincts.

	Hommes		Femmes		Total	
	Candidatures	Admissions (%)	Candidatures	Admissions (%)	Candidatures	Admissions (%)
Total	550	250 (**45**)	650	250 (**38**)	1200	500 (42)

Le danger des probabilités conditionnelles

Exemple adapté du papier: Peter J Bickel, Eugene A Hammel, J William O'Connell, et al. Sex bias in graduate admissions: Data from berkeley. Science, 1975.
On considère les taux d'admission par genre a une université ayant 2 départements avec des processus de sélection distincts.

	Hommes		Femmes		Total	
	Candidatures	Admissions (%)	Candidatures	Admissions (%)	Candidatures	Admissions (%)
Total	550	250 (**45**)	650	250 (**38**)	1200	500 (42)

=> Le taux d'admission global des femmes est significativement plus faible que celui des hommes.

Le danger des probabilités conditionnelles

Exemple adapté du papier: Peter J Bickel, Eugene A Hammel, J William O'Connell, et al. Sex bias in graduate admissions: Data from berkeley. Science, 1975.

On considère les taux d'admission par genre a une université ayant 2 départements avec des processus de sélection distincts.

	Hommes		Femmes		Total	
	Candidatures	Admissions (%)	Candidatures	Admissions (%)	Candidatures	Admissions (%)
Total	550	250 (**45**)	650	250 (**38**)	1200	500 (42)

=> Le taux d'admission global des femmes est significativement plus faible que celui des hommes.

Pouvons-nous aussi simplement conclure qu'il y a un biais dans la sélection des candidatures en défaveur des femmes ? Qu'elles peuvent être les raisons de l'écart observé ?

Il s'agit à nouveau du paradoxe de Simpson

Exemple adapté du papier: Peter J Bickel, Eugene A Hammel, J William O'Connell, et al. Sex bias in graduate admissions: Data from berkeley. Science, 1975.

	Hommes		Femmes		Total	
Dpt.	Candidatures	Admissions (%)	Candidatures	Admissions (%)	Candidatures	Admissions (%)
A	400	200 (**50**)	200	100 (**50**)	600	300 (**50**)
B	150	50 (**33**)	450	150 (**33**)	600	200 (**33**)
Total	550	250 (45)	650	250 (38)	1200	500 (42)

=> Les départements ont des sélections indépendantes et chacun semble indifférent du genre.

Il s'agit à nouveau du paradoxe de Simpson

Exemple adapté du papier: Peter J Bickel, Eugene A Hammel, J William O'Connell, et al. Sex bias in graduate admissions: Data from berkeley. Science, 1975.

	Hommes		Femmes		Total	
Dpt.	Candidatures	Admissions (%)	Candidatures	Admissions (%)	Candidatures	Admissions (%)
A	400	200 (**50**)	200	100 (**50**)	600	300 (**50**)
B	150	50 (**33**)	450	150 (**33**)	600	200 (**33**)
Total	550	250 (45)	650	250 (38)	1200	500 (42)

=> Les départements ont des sélections indépendantes et chacun semble indifférent du genre.
=> la différence observée est due au fait que les femmes (dans cet exemple) candidatent plus au département qui est le plus demandé et le plus sélectif.

Analyse descriptive

Quelques paramètres...- Analyse univariée

Pour les variables quantitatives continues :

- **La moyenne** est égale au quotient de la somme de toutes les valeurs de la série par l'effectif total.

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

- **La variance** est la moyenne des carrés des écarts à la moyenne.

$$s^2{}_x = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

- **L'écart-type** est la racine carrée de la variance.
  NB : Les paramètres précédents sont sensibles aux variations des valeurs aberrantes.
- **Les quartiles** Q_1 , Q_2 (la médiane), Q_3.

Quelques paramètres...- Analyse univariée

Pour les variables qualitatives ou discrètes:

- **L'effectif** de la modalité m_q d'une variable X est le nombre total n_q d'individus de l'échantillon pour lesquels la variable vaut m_q.
- **La fréquence** f_q est la proportion d'individus pour lesquels la variable vaut m_q.

$$f_q = \frac{n_q}{n}$$

  Le pourcentage est $f_q x 100$

- **Le mode** est la valeur la plus observée.
- **Les quartiles** pour les variables discrètes/qualitatives ordinales.
- **L'effectif ou la fréquence cumulée** permet de connaître le nombre/la proportion d'observations inférieures à une modalité donnée pour les variables discrètes/qualitatives ordinales.

Quelques paramètres...- Analyse univariée

Zoom sur les quartiles :

Les quartiles Q_1 , Q_2 (la médiane), Q_3 divisent une série statistique en 4 parties d'effectifs égaux :

- le quartile zéro (minimum) est celui qui a le rang 1
- le premier quartile est celui qui a le rang (N+3)/4
- la deuxième quartile (médiane) est celui qui a le rang (N+1)/2
- le troisième quartile est celui qui a le rang (3N+1)/4
- le quatrième quartile est celui qui a le rang N

La médiane peut donc être définie comme la valeur “du milieu”. Elle correspond plus précisément à un pourcentage cumulé de 50 % (c'est-à-dire que 50 % des valeurs lui sont supérieures et 50 % lui sont inférieures).

Type de paramètres - Analyse univariée

- **Paramètres de position :** *minimum, maximum, moyenne, quantiles, médiane, mode...*
  Il donne l'ordre de grandeur.

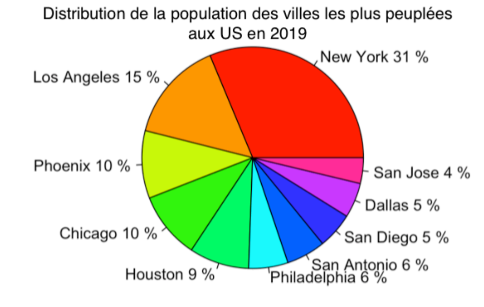
- **Paramètres de dispersion :** *variance, étendue, écart-type, écart-interquartile,...*
  Il donne la répartition autour de l'ordre de grandeur.

- **Paramètres de forme :** *coefficient d'asymétrie ou d'aplatissement*.
  Il donne la tendance, l'allure.

Type de graphiques - Analyse univariée

Pour une variable qualitative (1)

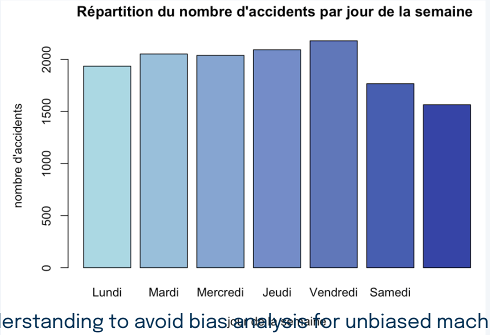
Le diagramme en secteur représente la distribution d'une variable qualitative nominale : les modalités sont représentées par des portions de disque proportionnelles à leur effectif, ou à leur fréquence.



Type de graphiques - Analyse univariée

Pour une variable qualitative (2)

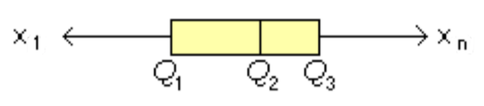
Le diagrammes en barres représente sa distribution : les modalités sont placées en abscisse, formant des bases de rectangles égales et équidistantes, et les effectifs (ou fréquences) en ordonnée, suivant une échelle arithmétique. Les surfaces des rectangles obtenus sont proportionnelles aux effectifs (ou aux fréquences).



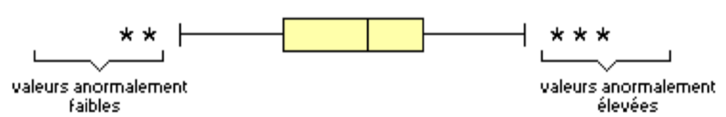
Type de graphiques - Analyse univariée

Pour une variable quantitative (1)

La boite à moustaches positionne les quartiles Q_1 , Q_2 , Q_3, au moyen de rectangles, prolongés par des ”moustaches” de part et d'autre, de longueur au plus égale à une fois et demie l'interquartile.



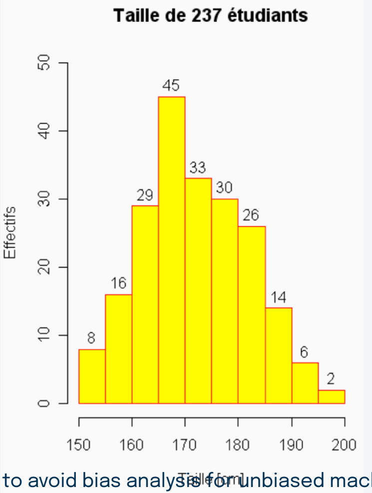
Si la plus petite ou la plus grande valeur observée se trouve à l'intérieur, on raccourcit les moustaches correspondantes ; si elle se trouve à l'extérieur, on positionne à part les valeurs ”aberrantes” qui dépassent des moustaches.



Type de graphiques - Analyse univariée

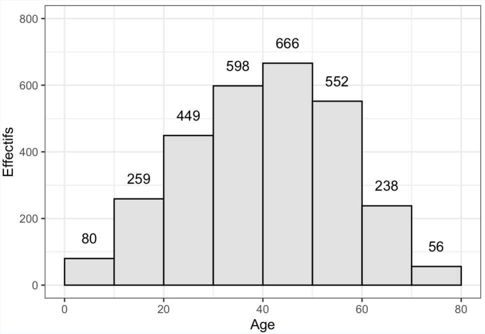
Pour une variable quantitative (2)

L'histogramme représente une distribution continue regroupée en classes : rectangles juxtaposés dont les bases sont les classes, et les surfaces sont proportionnelles aux effectifs (ou fréquences) associés.



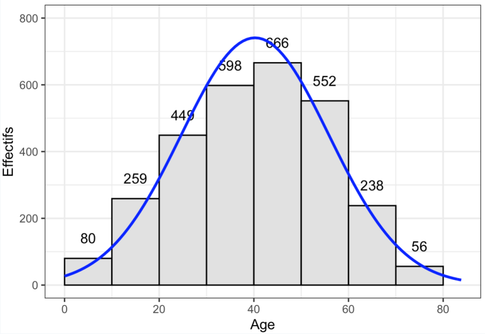
Type de graphiques - Analyse univariée

Lors de la représentation d'une variable quantitative par **l'histogramme** la distribution peut avoir une forme de cloche : on parlera de distribution **normale** ou **gaussienne**.



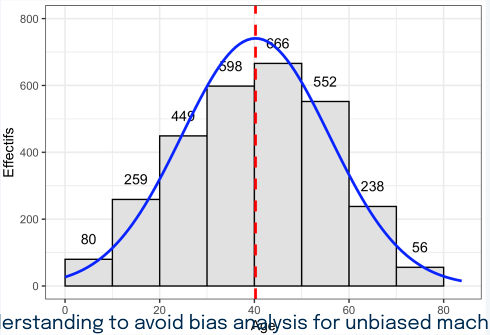
Type de graphiques - Analyse univariée

On peut approximer la distribution à partir d'une courbe appelée **la loi normale.** Cette loi permet d'approximer la distribution des valeurs de certaines variables quantitatives continues.



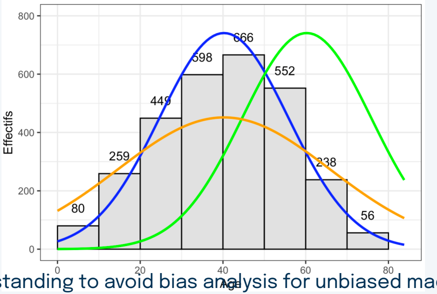
Type de graphiques - Analyse univariée

La courbe de la loi normale est **symétrique.** Le centre de la loi normale équivaut à la fois à la moyenne, au mode et à la médiane.
Son comportement est **asymptotique** de part et d'autres, autrement dit la courbe s'approche de l'axe des abscisses jusqu'à l'infini.



Type de graphiques - Analyse univariée

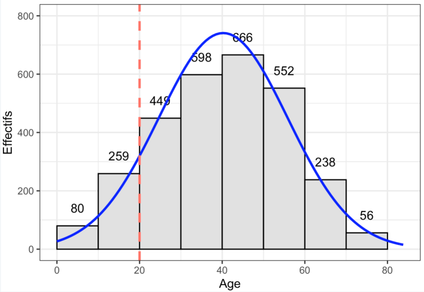
Ici, l'âge suit une loi normale de paramètre μ=40,2 et σ^2=15,6 représentée en bleu.
Si on fait varier la **moyenne**, on observe une translation sur l'axe horizontal (droite/gauche). En vert, μ a été augmenté de 20.
Si on fait varier **l'écart type**, on observe une dilatation verticale (écrasé/tiré vers le haut). En orange, σ^2 a été augmenté de 10.



Type de graphiques - Analyse univariée

- On peut déterminer **le pourcentage** ou **la probabilité**, qu'un patient ait moins de 20 ans :

Selon l'histogramme : (80+259)/2900=0,12 soit 12%.



Quelques paramètres - Analyse bivariée

Pour deux variables qualitatives (1)

On présente généralement ces données sous la forme d'une **table de contingence** reportant les effectifs de chaque couple de modalités (m_{q_1}, m_{q_2}), appelés **effectifs conjoints** et notés $n_{q_1 q_2}$.
Les sommes en ligne et en colonne des effectifs conjoints s'appellent **marges colonnes** et **marges lignes** (correspondant respectivement au vecteur $(n_{1.}, n_{2.}, n_{k_1.})$ et $(n_{.1}, n_{.2}, n_{.k_2})$.

Illustration : Type de dystrophie en fonction de la myotonie

	Absente	Percussion Seule	Préhension légère	Préhension sévère	Somme
DM1	244	282	1048	748	2322
DM2	61	23	17	7	108
Somme	305	305	1065	755	2430

Quelques paramètres - Analyse bivariée

Pour deux variables qualitatives (2)

Afin d'apprécier le lien entre deux variables qualitatives, on compare les distributions conditionnelles d'une variable en fonction des niveaux de l'autre. Pour comparer les distributions conditionnelles, on construit à partir de la table de contingence la **table des profils-lignes (ou des profils-colonnes)** en divisant les effectifs conjoints par les marges colonnes (ou marges lignes).

Illustration : Table des profils-lignes

	Absente	Percussion Seule	Préhension légère	Préhension sévère	Somme
DM1	0.11	0.12	0.45	0.32	1
DM2	0.56	0.21	0.16	0.06	1

Quelques paramètres - Analyse bivariée

Pour deux variables quantitatives (1)

Le lien entre deux variables quantitatives se mesure classiquement à l'aide du **coefficient de corrélation linéaire** appelé **coefficient de Pearson** (si la relation est linéaire...).

$r = \frac{\sigma_{xy}}{s_x * s_y}$ avec $\sigma_{xy} = \frac{1}{n} \sum_{i=1}^{n} ((x_i - \bar{x}) * (y_i - \bar{y}))$

r est compris entre -1 et 1.

Attention, l'indépendance entre les variables implique que $r = 0$ mais $r = 0$ n'implique généralement pas l'indépendance entre les variables, mais simplement une absence de lien linéaire !

Quelques paramètres - Analyse bivariée

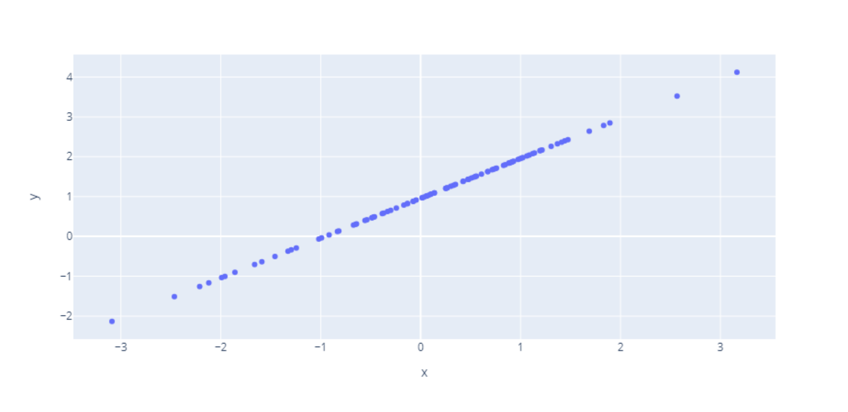
Pour deux variables quantitatives (2)

Si la relation n’est pas linéaire, le lien entre deux variables quantitatives peut se mesurer par le **coefficient de Spearman**. Il évalue la corrélation linéaire sur les rangs des observations.
On le calcule en déterminant les rangs de chaque valeur dans les deux séries (en déterminant un rang moyen en cas d’égalité) puis en calculant le coefficient de corrélation linéaire sur ces rangs.

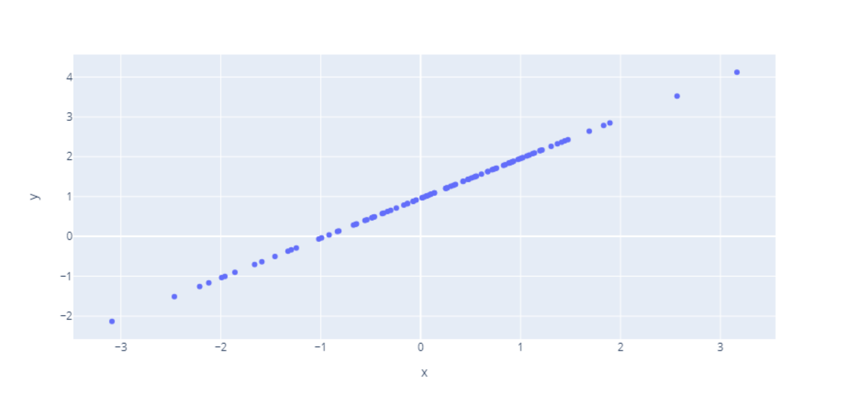
Il est aussi compris entre -1 et 1.

Ce coefficient est robuste aux valeurs aberrantes, mais ne permet de détecter que des relations **monotones**.

Quelques paramètres - Analyse bivariée

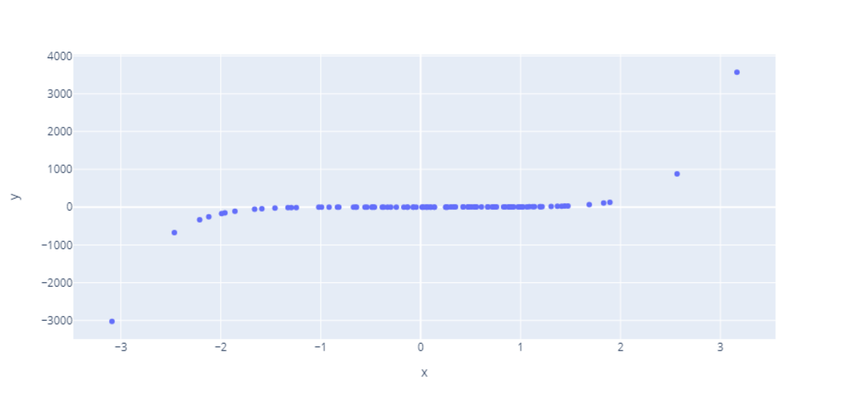


Quelques paramètres - Analyse bivariée

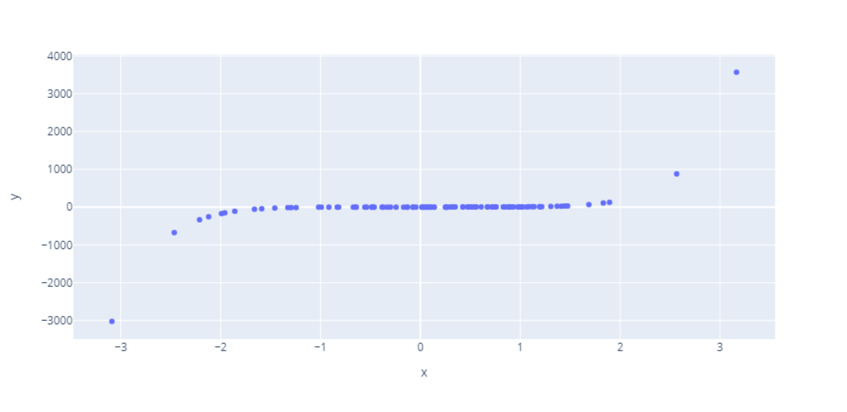


- Pearson: 1.0
- Spearman: 0.99

Quelques paramètres - Analyse bivariée

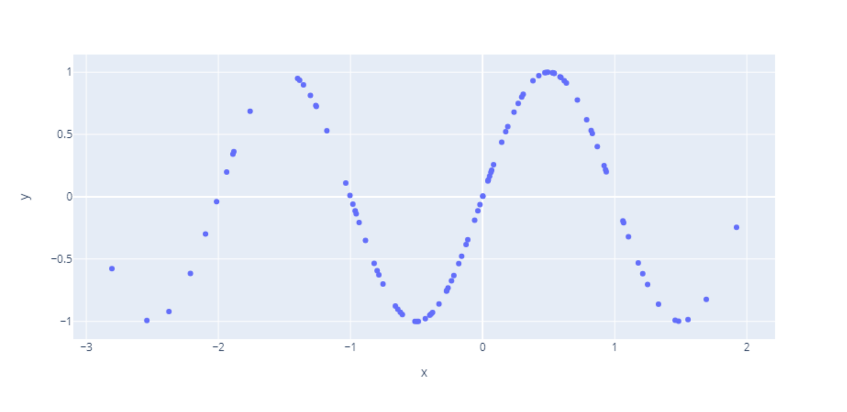


Quelques paramètres - Analyse bivariée

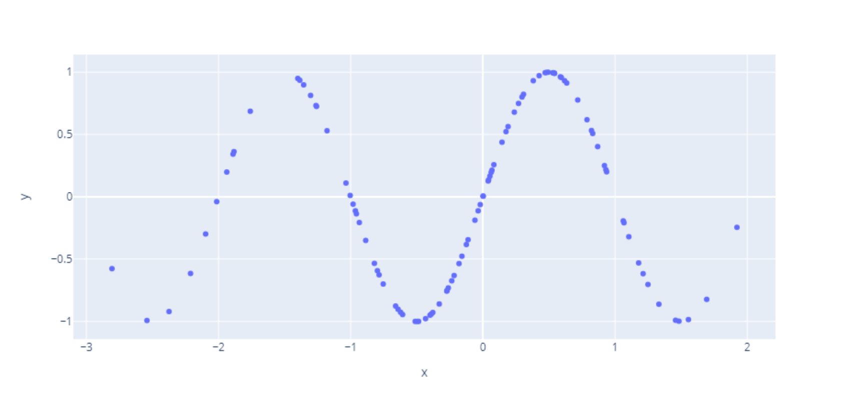


- Pearson: 0.53
- Spearman: 0.99

Quelques paramètres - Analyse bivariée



Quelques paramètres - Analyse bivariée



- Pearson: 0.11
- Spearman: 0.17

Quelques paramètres - Analyse bivariée

- Corrélation nulle n'implique pas l'indépendance
- Corrélation forte entre deux variables n'impliquent pas causalité
  - Les ventes de glace et l'occurrence de coup de soleil sont fortement corrélé
  - Cependant, il n'y a pas de relation de causalité entre les deux phénomènes

Quelques paramètres - Analyse bivariée

Pour une variable quantitative
et une variable qualitative (1)

Les indicateurs vus pour les variables quantitatives dans le cadre des analyses univariées peuvent être présentés par modalité.
Par exemple, la moyenne peut être calculée dans chacune des modalités de la variable qualitative.

Quelques paramètres - Analyse bivariée

Pour une variable quantitative et une variable qualitative (2)

Le rapport de corrélation permet d'évaluer le lien entre une variable quanti et une variable quali (à K modalités). C'est la part de variation de Y expliquée par X dans la variation totale de Y.

$$\eta^2_{(x,y)} = \frac{s_E^2}{s_T^2} \qquad \text{(Il est compris entre 0 et 1)}$$

La variance résiduelle est la moyenne pondérée des variances des sous-populations : $s_R^2 = \frac{1}{n}\sum_{k=1}^{K} n_k s_k^2$, avec s_k^2 variance de y dans le groupe k (variance inter)

La variance expliquée par X est la moyenne pondérée des carrés des variations des sous-populations : $s_E^2 = \frac{1}{n}\sum_{k=1}^{K} n_k(\bar{y_k} - \bar{y})^2$ (variance intra)

La variance totale est $s_T^2 = s_R^2 + s_E^2$

Quelques paramètres - Analyse bivariée

Pour une variable quantitative
et une variable qualitative ordinale (3)

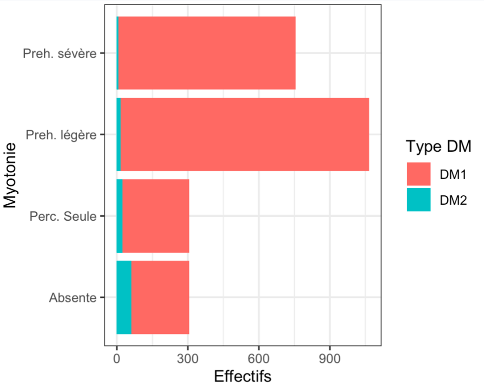
Pour mesurer une progression, on note la valeur de départ V_D et d'arrivée V_A.

- **La variation absolue** : $\Delta V = V_A - V_D$.
- **Le taux d'évolution** : $\frac{V_A - V_D}{V_D}$
  Ce taux est souvent exprimé en pourcentage, il faut donc le multiplier par 100.

Type de graphiques - Analyse bivariée

Pour deux variables qualitatives

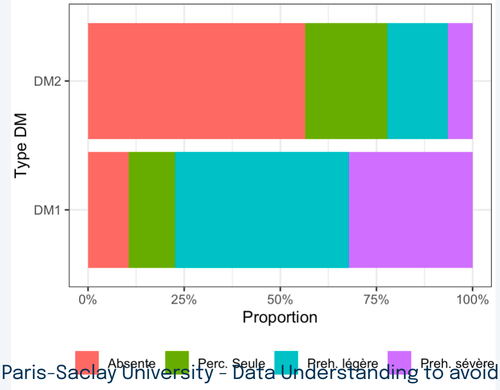
Le diagramme en barre selon les effectifs : si les groupes ne sont pas de tailles égales, il ne permet pas de visualiser les différences.



Type de graphiques - Analyse bivariée

Pour deux variables qualitatives

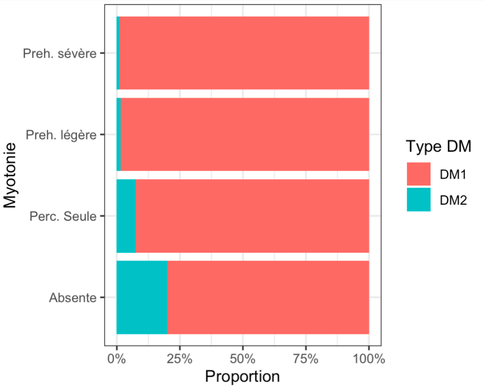
Le diagramme en barre selon les profils lignes :
Lorsque l'on s'intéresse principalement aux variations d'une variable selon une autre, *par exemple ici au type de DM selon le type de myotonie*.



Type de graphiques - Analyse bivariée

Pour deux variables qualitatives

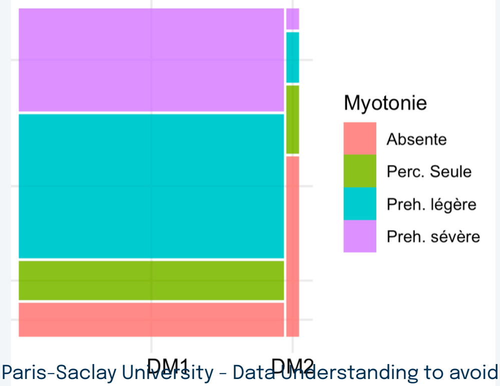
Le diagramme en barre selon les profils colonnes.
Ici par exemple on s'intéresserait au type de myotonie selon le type de DM.



Type de graphiques - Analyse bivariée

Pour deux variables qualitatives

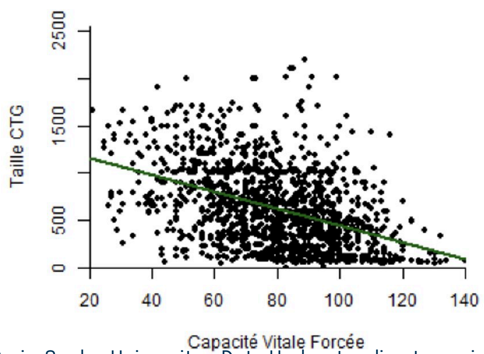
Le graphique en mosaïque permet de visualiser les différences entre les profils : Cette représentation consiste à représenter chaque effectif conjoint par un rectangle dont l'aire est proportionnelle à l'effectif associé.



Type de graphiques - Analyse bivariée

Pour deux variables quantitatives

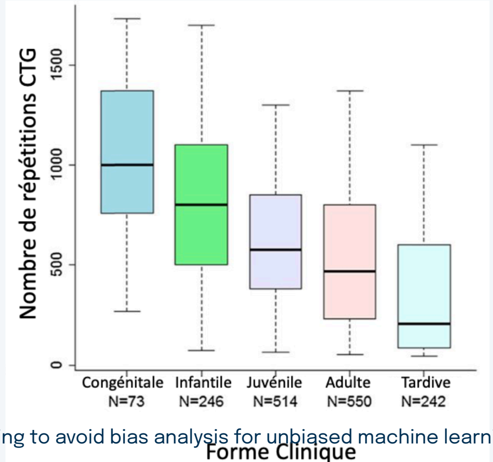
Le nuage de points représente la liaison entre deux variables quantitatives. Chaque individu est représenté par un point de coordonnées (x_i, y_i) qui sont les valeurs respectives des variables X et Y observées chez l'individu *i*.



Type de graphiques - Analyse bivariée

Pour une variable quantitative
et une variable qualitative

Les boîtes à moustaches représentent les distributions de la variable quantitative par modalité de la variable qualitative : la lecture se fait de la même manière que dans le cas univarié.
Deux histogrammes peuvent aussi être présentés juxtaposés, la comparaison peut s'avérer plus difficile.



Valeur manquante

Origine des valeurs manquantes

- L'utilisateur a oublié de remplir un champ
- Un capteur était en panne
- Les données ont été perdues lors du transfert manuel d'une base de données existante
- Il y a eu une erreur de programmation
- Les utilisateurs ont volontairement choisi de ne pas renseigner un champ lié à leurs convictions sur la manière dont les résultats pourraient être utilisés ou interprétés (vie privée, etc.).

A Vous !

- Tirez au sort 1000 valeurs d'une variable aléatoire gaussienne centrée réduite
- Calculez sa moyenne et sa variance. Tracez l'histrogramme des valeurs
- Supprimez aléatoirement 250 valeurs du vecteur. Calculez la moyenne, la variance et l'histogramme. Que constatez-vous ?
- Retirez ensuite des valeurs surtout dans les valeurs basses de l'échantillon : 225 valeurs parmi les valeurs basses et 25 parmi les valeurs hautes. Calculez la moyenne, la variance et l'histogramme. Que constatez-vous ?

Caractérisation de l'origine des valeurs manquantes

- **MCAR (missing completely at random)**: Une donnée est MCAR (manquante de façon complètement aléatoire) si la probabilité d'absence est la même pour toutes les observations.
    - Par exemple : si chaque participant à un sondage décide de répondre à la question du revenu en lançant un dé et refusant de répondre si la face 6 apparaît
    - Si la quantité de données MCAR n'est pas trop importante, ignorer les cas avec des données manquantes ne biaisera pas l'analyse mais risque de baisser la précision des modèles
- **MAR (Missing at random)**: la probabilité d'absence est liée à une ou plusieurs autres variables observées, on parle de missingness at random. Utilisation de méthodes statistiques pour éviter de biaiser l'analyse
- **MNAR (Missing not at random)**: La donnée est manquante de façon non aléatoire (MNAR) si la probabilité d'absence dépend de la variable en question.
    - Par exemple: le cas où des personnes avec un revenu important refusent de le dévoiler.
    - Les données MNAR induisent une perte de précision et un biais

MCAR (missing completely at random)

Soit M la matrice d'indication des valeurs manquantes, Y_{obs} les données observées, Y_{mis} les données manquantes, $Y = \{Y_{obs}, Y_{mis}\}$

- La probabilité qu'une valeur de X_1 soit manquante **ne dépend pas** des autres variables $X_{j \neq 1}$, qu'elles soient manquantes ou pas
- Pas possible de définir un profil des individus ayant des valeurs manquantes
- La probabilité de ces données est uniforme
- $P(M|Y) = P(M)$

MAR (Missing at random)

Soit M la matrice d'indication des valeurs manquantes, Y_{obs} les données observées, Y_{mis} les données manquantes, $Y = \{Y_{obs}, Y_{mis}\}$

- La probabilité qu'une valeur de X_1 soit manquante **dépend de valeurs observées** d'autres variables $X_{j \neq 1}$, mais pas des valeurs manquantes
- Exemple: Il existe une différence de non réponse entre les salariés de deux entreprises concernant la question du revenu, mais parmi les salariés d'une des deux entreprises données, la probabilité d'avoir une non réponse est la même quelque soit le niveau de revenu
- $P(M|Y) = P(M|Y_{obs})$

MNAR (Missing not at random)

Soit M la matrice d'indication des valeurs manquantes, Y_{obs} les données observées, Y_{mis} les données manquantes, $Y = \{Y_{obs}, Y_{mis}\}$

- La donnée est manquante pour une raison intrinsèque à sa valeur
- Exemple: les salariés les mieux payés d'une équipe refusent de répondre à une enquête sur le revenu
- Les données manquantes vont à la fois dépendre de Y_{obs} et Y_{mis}

Exclusion des valeurs manquantes

- Analyse des cas concrets (List Wise Deletion): consiste à ne considérer que les individus pour lesquels toutes les données sont disponibles, i.e. en supprimant les lignes comportant des valeurs manquantes. C'est ce qui est fait automatiquement avec R (na.action=na.omit).
- Analyse des cas disponibles (Pair Wise Deletion): consiste à faire les analyses avec toutes les cases dont les variables sont disponibles. Son désavantage est d'utiliser différentes tailles d'échantillons pour les différentes variables

Valide uniquement en cas de MCAR

Imputation simple

- Consiste à substituer une valeur à la valeur manquante
- De très nombreuses méthodes existantes
- Méthodes très “séduisantes”, mais…

Typologie des méthodes d'imputation

- Méthodes déterministes: Méthodes qui fournissent une valeur fixe étant donné l'échantillon
- Méthodes stochastiques: Méthodes d'imputation ayant une composante aléatoire (et donc qui ne donnent pas nécessairement la même valeur étant donné l'échantillon si la méthode est répétée)

Imputation par une valeur fixe

- Méthode déterministe
- Imputation de la valeur manquante par la moyenne/la médiane pour les variables quantitatives ou le mode pour les variables qualitatives
- Méthode à éviter...

A vous !

Nous nous intéressons ici à la moyenne, la médiane, la variance et la distribution d'un échantillon.

- Tirez au sort 1000 valeurs d'une variable aléatoire gaussienne centrée réduite
- Calculez la moyenne, la médiane, la variance. Tracez l'histrogramme des valeurs
- Supprimez aléatoirement 250 valeurs du vecteur. Calculez la moyenne, la médiane, la variance et l'histogramme. Que constatez-vous ?
- Remplacez les 250 valeurs manquantes par la moyenne des valeurs des 750 observations observées. Calculez la moyenne, la médiane, la variance et l'histogramme. Que constatez-vous ?
- Faites de même lorsque les valeurs manquantes sont majoritairement tirées dans les observations les plus faibles de l'échantillon.

Imputation par hot-deck aléatoire

- Méthode stochastique
- Tirer au sort aléatoirement (avec remise) parmi les valeurs disponibles de la variable la valeur imputée
- Exemple: Si je veux imputer l'âge et que les valeurs observées sont (20, 20, 30, 40), je tire au sort un âge, sachant 20 aura une proba 1/2 d'être choisi, 30 une proba 1/4 et 40 une proba 1/4.

A vous !

Nous nous intéressons ici à la moyenne, la médiane, la variance et la distribution d'un échantillon.

- Tirez au sort 1000 valeurs d'une variable aléatoire gaussienne centrée réduite
- Calculez la moyenne, la médiane, la variance. Tracez l'histrogramme des valeurs
- Supprimez aléatoirement 250 valeurs du vecteurs. Calculez la moyenne, la médiane, la variance et l'histogramme. Que constatez-vous ?
- Imputez les 250 valeurs manquantes par hot-deck aléatoire. Calculez la moyenne, la médiane, la variance et l'histogramme. Que constatez-vous ?
- Faites de même lorsque les valeurs manquantes sont majoritairement tirées dans les observations les plus faibles de l'échantillon.

Imputation par régression

- Méthode déterministe
- Chaque valeur manquante est remplacée par la prédiction d'un modèle de régression utilisant des variables auxiliaires

Imputation par les K plus proches voisins

- Méthode stochastique/déterministe
- Calculer les observations les plus proches de l'observation dont on veut imputer la valeur d'une variable manquante et imputer par la moyenne des K plus proches voisins

Autres méthodes d'imputation simple

- Régression Locale (LOESS)
- Nonlinear Iterative Partial Least Squares (NIPALS)
- Décomposition en valeurs singulières (SVD)
- MissForest
- Inférence Bayésienne

Défauts de l'imputation simple

- Une unique valeur imputée ne peut pas représenter toute l'incertitude à propos de la valeur à imputer
- Les analyses qui considèrent les valeurs imputées de manière équivalente aux valeurs observées sous-estiment l'incertitude
- Ce handicap peut conduire entre autres à des variances nettement sous-estimées

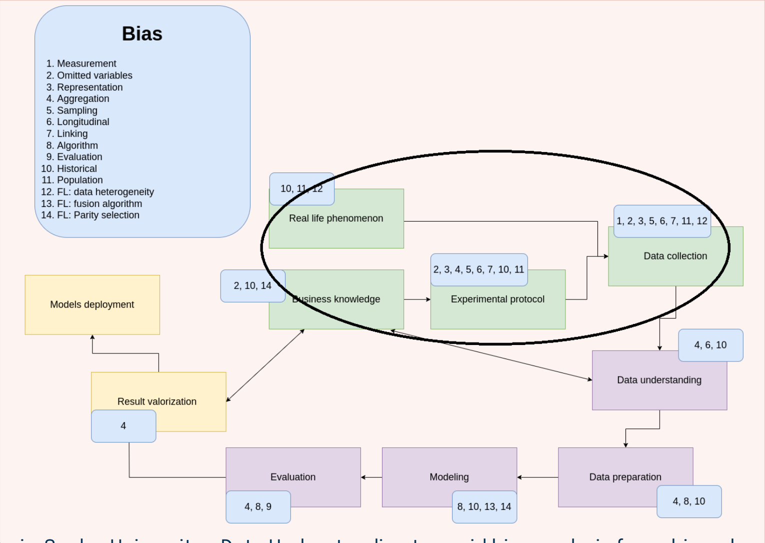
Imputation multiple

- Procéder à $m > 1$ imputations afin d'obtenir m valeurs pour chaque donnée manquante
- Combiner ensuite les statistiques calculées indépendamment sur les m jeux de données

Conclusion

- Les jeux de données contiennent souvent des valeurs manquantes
- Les données sont précieuses, supprimer les observations ou les variables manquantes peut être couteux
- Pour éviter de supprimer des données, on peut tenter d'imputer les valeurs manquantes
- Les valeurs manquantes peuvent être de trois types différentes. Selon leur type, certaines méthodes seront adaptées ou au contraire inadaptées
- Attention aux méthodes imputant par une valeur unique les valeurs manquantes d'une variable (par exemple, par la moyenne ou la médiane). Elles vont distorde la distribution des données
- L'imputation multiple permet une meilleure prise en compte de l'incertitude introduite. Cependant, elle peut être couteuse à mettre en place

Position


Bias
1. Measurement
2. Omitted variables
3. Representation
4. Aggregation
5. Sampling
6. Longitudinal
7. Linking
8. Algorithm
9. Evaluation
10. Historical
11. Population
12. FL: data heterogeneity
13. FL: fusion algorithm
14. FL: Parity selection
10, 11, 12
Real life phenomenon
1, 2, 3, 5, 6, 7, 11, 12
Data collection
2, 3, 4, 5, 6, 7, 10, 11
Experimental protocol
2, 10, 14
Business knowledge
Models deployment
4, 6, 10
Data understanding
Result valorization
4
Evaluation
4, 8, 9
Modeling
8, 10, 13, 14
Data preparation
4, 8, 10

Quelques célèbres biais cognitifs

- Biais des survivants: surévaluer les chances de succès en se concentrant sur les sujets ayant réussi qui sont des exceptions statistiques
  - Exemple: les bombardiers anglais renforcés au mauvais endroit
- Biais de confirmation: tendance à rechercher des informations confirmant ses hypothèses préexistantes en minimisant les autres informations
  - Personne croyant à un régime alimentaire recherchant uniquement des études soutenant son opinion
- Effet de halo: tendance à laisser une impression générale positive ou négative d'une personne influencer son jugement sur ses caractéristiques spécifiques
  - Un employeur influencé par les habits d'un candidat
- Biais de disponibilité: tendance à évaluer la probabilité des événements en fonction de la facilité avec laquelle des exemples nous viennent à l'esprit.
  - Accident d'avion versus accident de voiture
- Biais d'autocomplaisance: tendance à attribuer ses succès à ses propres capacités et ses échecs à des facteurs externes
  - Coureur de cross qualifié grâce à son niveau ou pas qualifié à cause de ses chaussures

Introspection

Test sur vos biais inconscients

https://implicit.harvard.edu/implicit/canadafr/takeatest.html