

Regard de biais sur l'analyse de données : contributions de la statistique au déroulement de la recherche en santé

Félix Camirand Lemyre

Professeur agrégé

Groupe de recherche interdisciplinaire en informatique de la santé - Uds

Chercheur régulier

Centre de recherche du centre hospitalier universitaire de Sherbrooke

Chercheur honoraire

The University of Melbourne



Statisticien méthodologiste & biostatisticien

Directeur

Centre de consultation statistique de l'Université de Sherbrooke

Directeur scientifique – volet biostatistique

PREVALIS



Objectifs de l'atelier

Objectifs de formation

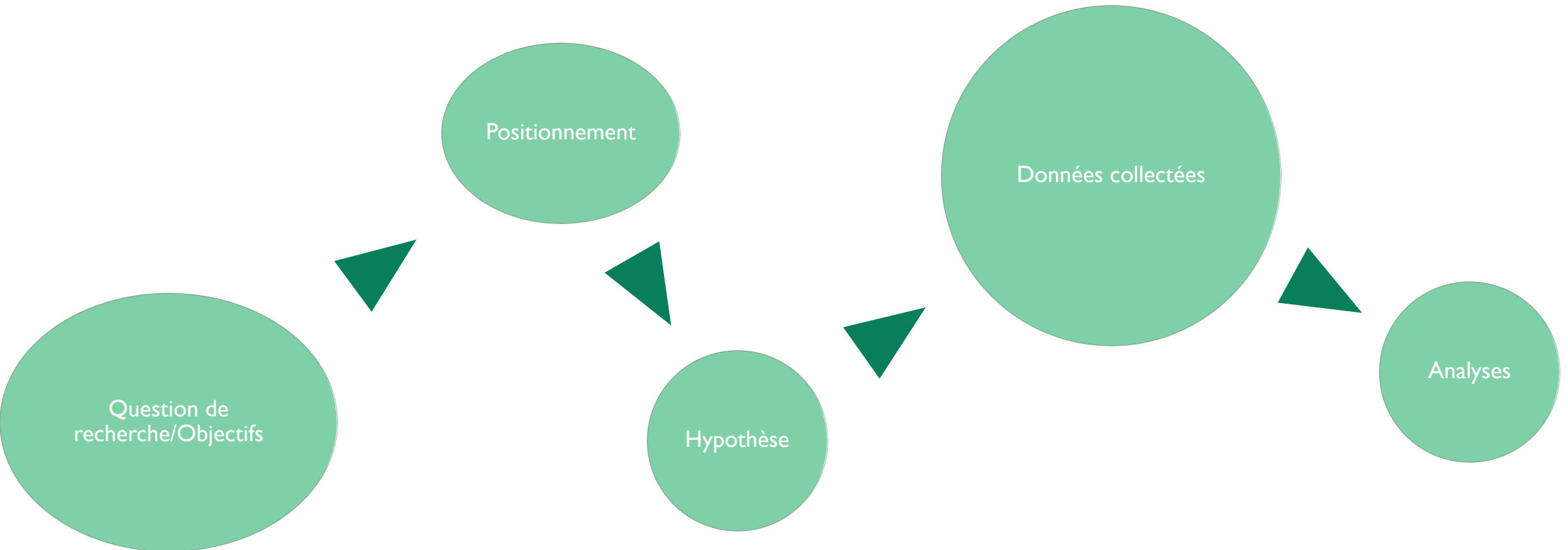
- O1. Mieux cerner le rôle de la statistique inférentielle.
- O2. Approfondir des notions de méthodologie statistique.
- O3. Susciter la réflexion au niveau de biais d'analyse inhérents au contexte de collecte de données.
- O4. Apprendre à cerner la portée de l'utilisation d'éléments d'analyse ou d'IA.





Un mot sur la méthodologie de recherche

Règles générales



Ce qu'on doit retenir de l'atelier

- La façon dont les données sont/ont été collectées doit influencer la manière de les analyser + la portée des conclusions
- Primordial de chercher à nuancer, critiquer, confronter



La statistique inférentielle, qu'est-ce que c'est?

Inférence statistique?

Inférence statistique:

- Ensemble des techniques pour induire les caractéristiques (paramètres) inconnus d'une population à partir de celles observées auprès d'un échantillon
- Modélisation probabiliste de systèmes
- Estimation des paramètres, intervalles de confiance, tests

Idée phare derrière le processus d'inférence

Patron observé dans l'échantillon → reflet du patron observé dans la Population

Patron: éléments paramétrés d'équations probabilistes

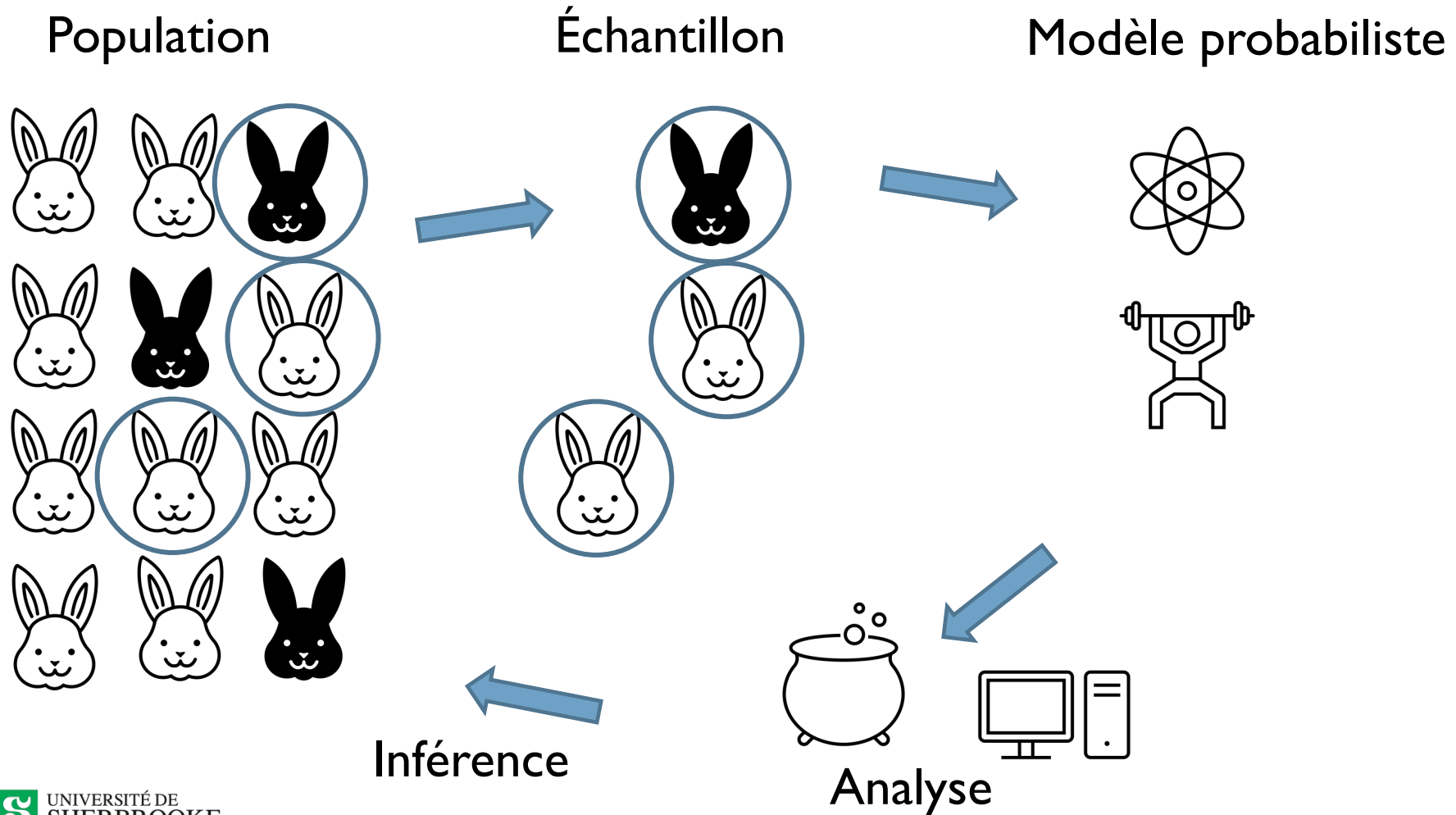
Théorie des probabilités: quantification de la qualité du reflet en terme d'incertitude sur l'estimation des paramètres

Inférence vs prédiction

Inférence: À quel point les caractéristiques de l'échantillon sont-elles représentatives de celles de la population de laquelle il est issu?

Prédiction: À quel point est-il possible de prédire une issue sur la base d'une sélection de variables prédictrices?

Procédure inférentielle illustrée



Questions typiques de nature inférentielles

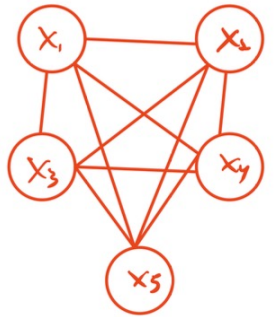
- Dans la population A, le facteur d'exposition X est-il associé à l'issue Y ?
- La distribution dans la population A des caractères X est-elle différente de celle de ces caractères dans la population B?
- Dans la population A, y a-t-il une variation dans la variable Z à travers le temps?

Ingrédients essentiels

Modèle

Régression/Classification

Distribution (jointe)

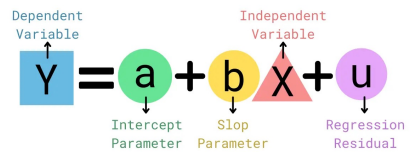


$$Y = f_{\theta}(X) + \epsilon$$

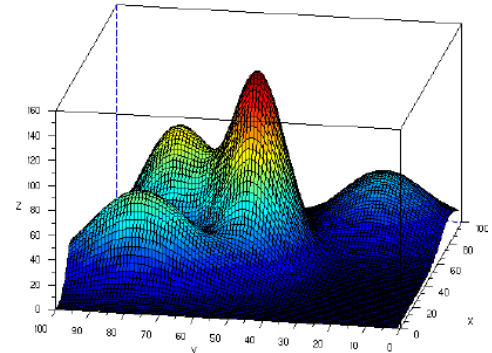
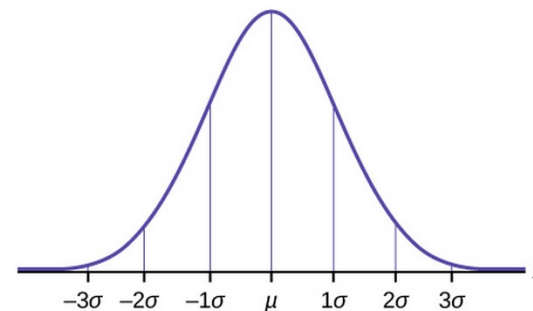
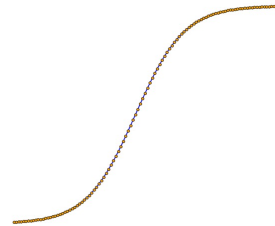
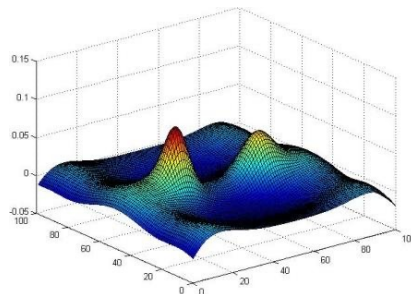
$$\mathbb{P}(Y = 1 \mid X) = f_{\theta}(X)$$

$$X \sim f_{\theta}$$

Regression Analysis



www.erp-information.com



Estimation

Étant donné un échantillon

$$(Y_1, X_1), \dots, (Y_n, X_n)$$

Modèle

Étant donné un échantillon

$$X_1, \dots, X_n$$

Régression/Classification

Distribution (jointe)

$$Y = f_{\theta}(X) + \epsilon$$

$$\mathbb{P}(Y = 1 \mid X) = f_{\theta}(X)$$

$$X \sim f_{\theta}$$

Maximisation de la vraisemblance

Moindres carrés

Fonctions de perte (locales)

Maximisation de la vraisemblance a posteriori

Maximisation de la vraisemblance conditionnelle

Maximisation de la vraisemblance locale

Résultat

$$\hat{f}_\theta$$

Statistique inférentielle



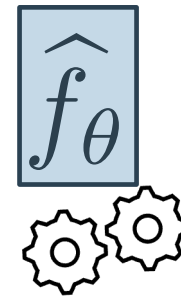
Qualité de \hat{f}_θ vis-à-vis du modèle de la population de laquelle l'échantillon est issu

Pour statuer sur la qualité d'estimation

- Hypothèses sur le modèle
 - Forme fonctionnelle (p.ex: linéaire, nonlinéaire, dimensionalité creuse etc)
 - Régularité
- Hypothèses sur l'échantillonnage
 - Lien entre les observations, p.ex: i.i.d., corrélation sérielle
 - Lien entre les données manquantes (au hasard, complètement au hasard, systématique...)



Traduction en terme d'hypothèses
structurelles probabilistes

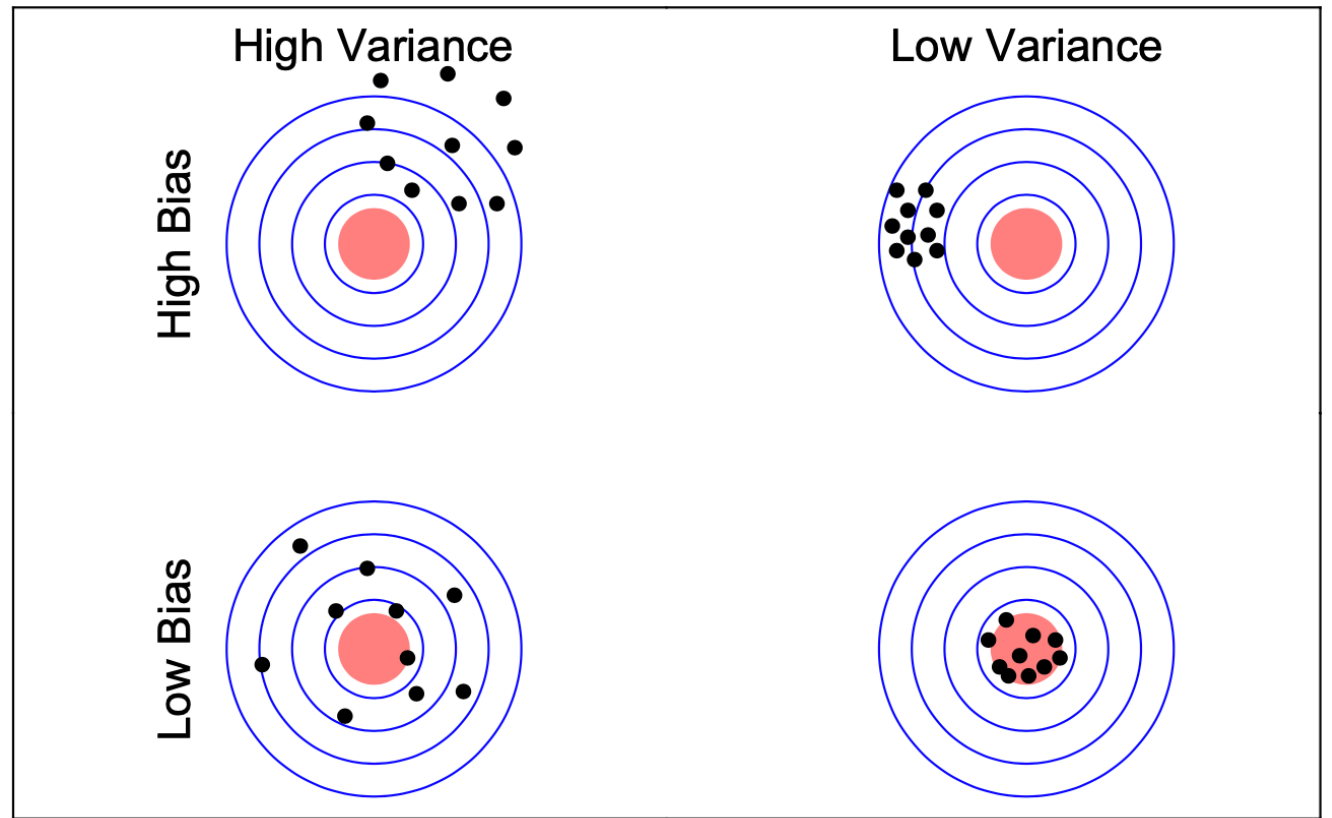


Biais?
Variance?
Convergence?
Intervalle de confiance?



Biais et variance

Chaque point:
Échantillonnage +
estimation



Cadre classique de la statistique inférentielle

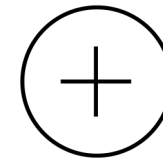
Étant donné:

1. Devis échantillonnal
2. Modèle
3. Technique d'estimation

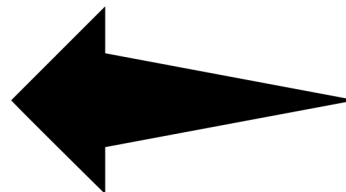


Ensemble minimal
d'hypothèses pour
garantir la validité de

$$\hat{f}_\theta$$



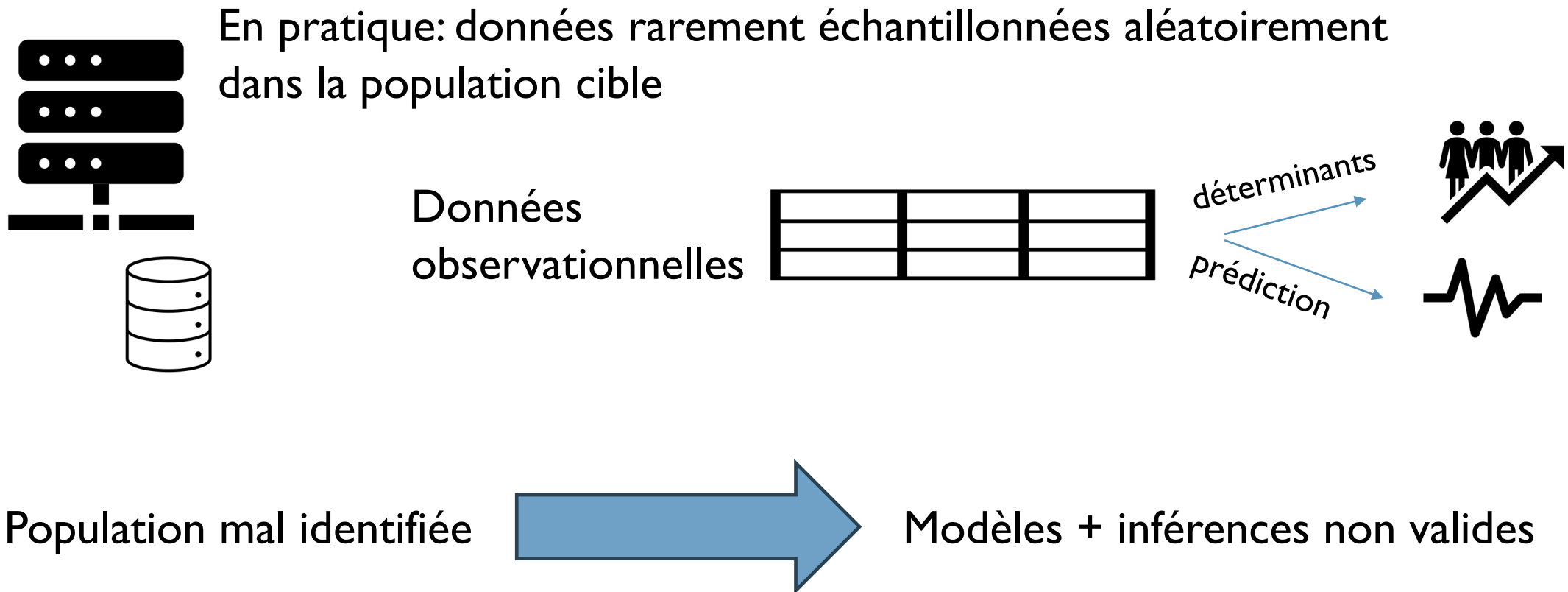
Estimation de biais +
variance



**Valide seulement vis-à-vis de
la population de laquelle est
issu l'échantillon**

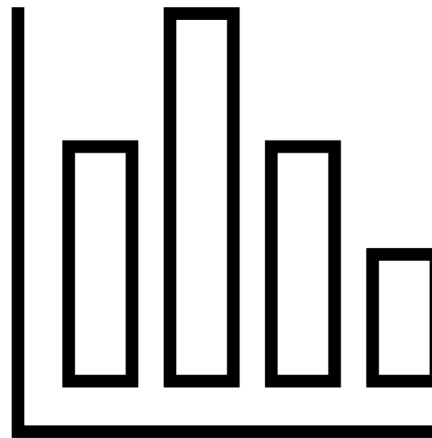
À propos du devis échantillonnal...

Cerner la population: un enjeu en recherche



Population?

Ensemble des individus ayant une probabilité non nulle de se retrouver dans l'échantillon ou la base de données



Exemples

« Après analyse, le taux de mortalité se situe à 40%, avec un intervalle de confiance de [35,45] à un niveau $\alpha = 0,05$. »

- Peut-être normal si l'échantillon provient des soins intensifs...

« Le taux de mortalité lié à l'administration du vaccin contre la COVID est de 5%. »

- Peut-être normal si l'échantillon est constitué de dossiers rapportés à un organe de pharmacovigilance.

« L'âge moyen estimé des personnes utilisant la piste cyclable du Lac des Nations est de 65 ans. »

- Peut-être conséquent avec un échantillonnage ayant eu lieu les lundis de septembre entre 9h30 et 11h.



À propos du choix du modèle

Conséquence d'une erreur de spécification...

- Rejeter une hypothèse nulle à tort
- Ne pas rejeter une hypothèse nulle à tort
- ...

Compromis dans la flexibilité permise au niveau de la gamme d'effets explorés

Exemple

« Rejeter l'hypothèse d'une variation dans le taux d'hormone de régulation de l'appétit à travers les trimestres de grossesse en se basant sur un modèle linéaire mixte incorporant le temps en facteur linéaire, alors qu'on se serait attendu a priori à observer une augmentation du 1^{er} au 2^e trimestre, puis une diminution du 2^e au 3^e trimestre. »

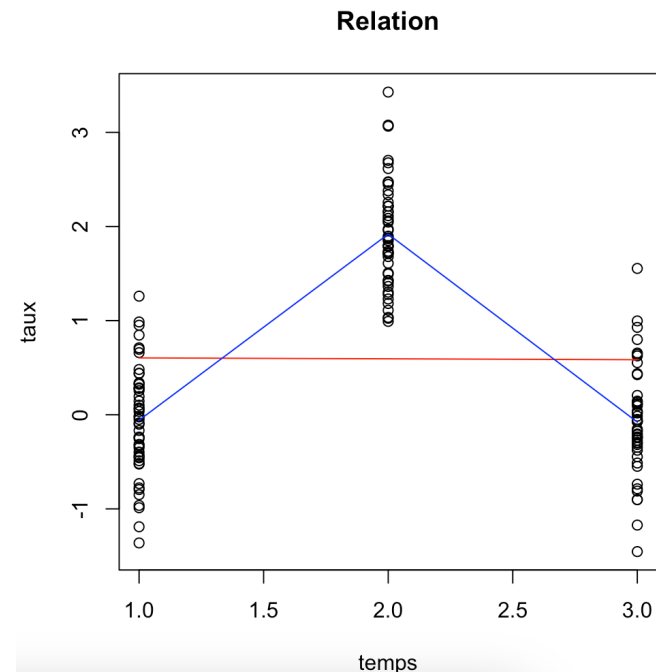
$Y \sim \text{temps} + \text{effet aléatoire}$

vs

$Y \sim \text{temps} + (\text{temps} == 1) + \text{effet aléatoire}$

p-valeur > 0.5

p-valeur < 0.005



Ce que plusieurs modèles supposent

X_1, \dots, X_n I.I.D. • Indépendantes
• De même distribution, égale à celle n'importe quel individu dans la population

ou

$Y_1 | X_1, \dots, Y_n | X_n$ Indépendantes + mêmes constantes distributionnelles

Degré d'adhérence à ces hypothèses → Nuance les résultats observés



Principales conclusions à tirer jusqu'ici

Cadre méthodologique de l'inférence statistique

Au-delà de l'utilisation d'un logiciel statistique et du calcul de p-valeurs

- Évaluation du devis échantillonnal et de la population concernée
- Analyse de la capacité du modèle à discerner les tendances pressenties
- Évaluation de l'adhérence aux critères de validité des estimés
- Documentation de la robustesse de l'approche

Qu'en est-il de la prédiction?

Apprentissage automatique/Apprentissage statistique

1. Devis échantillonnal
(données d'entraînement)
2. Modèle
3. Technique d'estimation



Ensemble minimal
d'hypothèses pour
garantir la validité de

$$\hat{f}_{\theta}$$

Exemple: interpolation/extrapolation

« Pourrait-on se fier à un algorithme entraîné sur la base de données longitudinales d'enfants entre 0 et 12 ans pour prédire la mortalité en CHSLD? »

Représentation du profil d'individus pour lesquels une prédiction est requise

Autres enjeux liés à l'échantillonnage

- Facteurs de confusion

$$X \xrightarrow{Z} Y$$

- Événements rares
- Données manquantes
- Validité temporelle/transportabilité

Identification de facteurs causaux

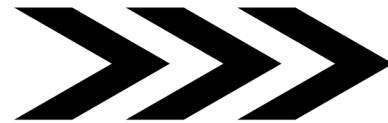
Associations



Causalité

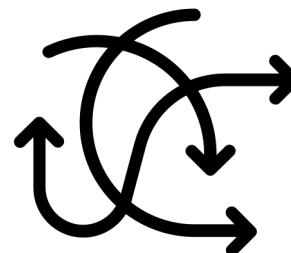


Associations
observationnelles



Autre échantillon

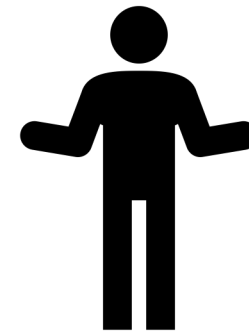
Prédicteurs



Issue de santé

Traitement des symptômes sévères de la COVID

- Cohorte ISARIC (International Severe Acute Respiratory and emerging Infection Consortium)
- >800 000 individus
- Administration de stéroïdes → **Mortalité**
- « Ajustement » pour la sévérité: compliqué



En résumé

Inférences et prédictions non valides si:

- Modèle erroné
- Échantillon non représentatif

Exemples d'études ou de déploiements d'algorithmes qui
ont présentés des enjeux de représentativité
échantillonnale?

Exercice 1

- Quels sont les critères qui vous permettent d'évaluer si les conclusions d'une étude s'appliquent à vous ou non?
- Quels sont les critères qui vous permettent d'évaluer si les prédictions d'un algorithme pourraient être applicables à vous ou non?

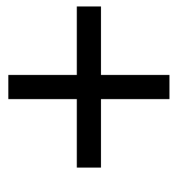
Exercice 2

- Dans quel(s) contexte(s) est-il possible d'estimer des effets causaux à partir de données observationnelles?



Inférence vs prédiction

Collecte de données



Modèle



Analyses/prédictions

Différents objectifs poursuivis

Inférence: À quel point les caractéristiques de l'échantillon sont-elles représentatives de celles de la population de laquelle il est issu?

Prédiction: À quel point est-il possible de prédire une issue sur la base d'une sélection de variables prédictrices?

Des cadres méthodologiques différents

Inférence:

- Prépécification des tests/analyses
- Proscrit d'itérer jusqu'à l'obtention d'effets significatifs...

Prédiction:

- Optimisation d'un critère en lien avec l'application projetée
Pouvoir prédictif global, compromis sensibilité/spécificité,...
- Choix du meilleur modèle possible à partir des données d'entraînement

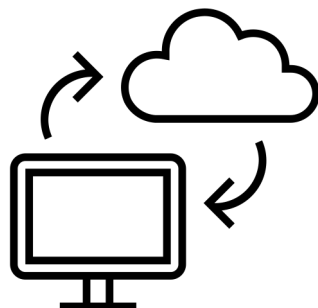
40

Scandal des p-valeurs...

- Tests :
 - Calibrés pour conclure à tort à un effet significatif suivant la réalisation d'une expérience dans au plus α % des cas de même type
 - Probabilité de trouver trouver un effet faussement significatif suivant la réalisation de K expériences indépendantes: $1-(1-\alpha)^K$
 - Si $K = 5$ et $\alpha = 0.05$: 23%
 - Si $K = 10$ et $\alpha = 0.05$: 41%
 - Si $K = 100$ et $\alpha = 0.05$: 99.4%

Un problème du même type en prédiction

Calibration itérative d'un modèle suite à l'évaluation suivant le même ensemble de données tests



Bonnes pratiques

- Planification des expériences et des stratégies anticipées
- Documenter et rapporter ce qui a été réalisé
- Nuancer en conséquence

La place de la recherche exploratoire

Analyses à visées exploratoires vs confirmatoires

Un question de définition d'objectifs vis-à-vis de l'état de la littérature...

Exercice 3

Vous dirigez un centre de soins d'urgence. On vous propose l'implantation d'un outil prédicteur de la mortalité basé sur les informations patients disponibles à l'admission pour guider la mobilisation de ressources.

Sur quels critères baseriez-vous la décision de procéder à cette implantation?

Comment évalueriez-vous la qualité de cet outil?

Comment jugeriez-vous la pertinence de l'outil pour votre milieu de soin?

Conclusion

Exercice 3

- Chercher à documenter les limites des conclusions
- Plusieurs biais difficiles à mitiger