Planification expérimentale

R. Fisher (1890-1962)

Gilles San Martin
gilles.sanmartin@gmail.com

Wallonie recherche
CRA-W

Objectifs

Classiquement planification expérimentale =
1 design expérimental optimisant la récolte de données
+
une analyse pour chaque cas

Pex : Randomized Bloc Designs, Latin Square designs, Split Plot design… + ANOVAs

Ici :
On se contentera de voir les grands principes.
Très peu de mathématiques !

Les GLM (+ extensions) sont ensuite suffisamment flexibles pour s'adapter à la plupart des cas d'analyse

4 parties

1) Qu'est-ce que la science ?
rôle de la planification expérimentale dans la méthode scientifique

2) Principales approches dans la collecte de données
expérimentales - semi/quasi-expérimentales - observatives

3) Les règles d'or de la planification expérimentale
Question et population bien définies
Adéquation des mesures
Réplications indépendantes - Attention à la pseudoréplication !
Randomisation des traitements et des mesures
Témoins

4) Puissance statistique : Combien d'échantillons a-t-on besoin ?
Comment augmenter la puissance statistique ?

Objectifs

1) Évaluer un article/une étude scientifique

être capable de discuter et de sous-peser les forces et les faiblesses des données et des hypothèses d'une étude, d'un protocole de récolte de données existant

2) Concevoir un protocole de collecte de données

être capable de d'imaginer un protocole de récolte de données pour répondre à une question scientifique en en comprenant les limites et en évitant les pièges les plus fréquents

3) La méthode scientifique

Avoir une meilleure vue d'ensemble sur la manière dont se pratique la science et sur la manière dont elle peut être perçue de l'extérieur

Parmi les "faiblesses", la "**pseudoréplication**" est un des problèmes les plus répandus et mérite une attention toute particulière !

(1) Qu'est-ce que la science ?

Quel est le rôle de la planification expérimentale dans la méthode scientifique ?

Qu'est-ce que la science ?

Science
=
Corpus de connaissances
+
"Une" méthode pour acquérir ces connaissances

"*Science is a way of thinking much more than it is a body of knowledge.*"
- Carl Sagan

Autres sources de connaissance/décisions :

Autorité (parents, "leaders", "experts")
Société (imitation - conformisme)
Observations non systématiques / expérience personnelle
Intuition - bon sens
Foi

On fonctionne la plupart du temps de cette manière, pas avec l'approche scientifique pour plusieurs raisons :

1) La méthode scientifique demande des efforts considérables et conscients pour aller contre notre "manière naturelle de penser" et éviter les pièges

2) La méthode scientifique est beaucoup trop lente et complexe dans la plupart des situations du quotidien. Impossible de tout évaluer "scientifiquement"

NB : la science est faite par des êtres humais totalement faillibles et qui sont soumis aux même "tendances naturelles" et contraintes temporelles que les autres...

Voir : Daniel Kahneman (2012) Thinking fast and slow

Autres sources de connaissance :
Autorité (parents, "leaders", "experts")
Société (imitation - conformisme)

En théorie, les arguments d'autorité n'ont aucune place dans une argumentation scientifique. Les faits/données devraient toujours prévaloir ...

Mais l'autorité (sur base de l'expertise) et la communauté/culture (ea les découvertes passées) peuvent aussi jouer un rôle en science.

Dans la vie quotidienne - comme en science - en l'absence de toute information (ou de temps pour l'acquérir) le conformisme est souvent l'approche la plus efficace (mais pas sans risque)...

Autres sources de connaissance :

Intuition - bon sens

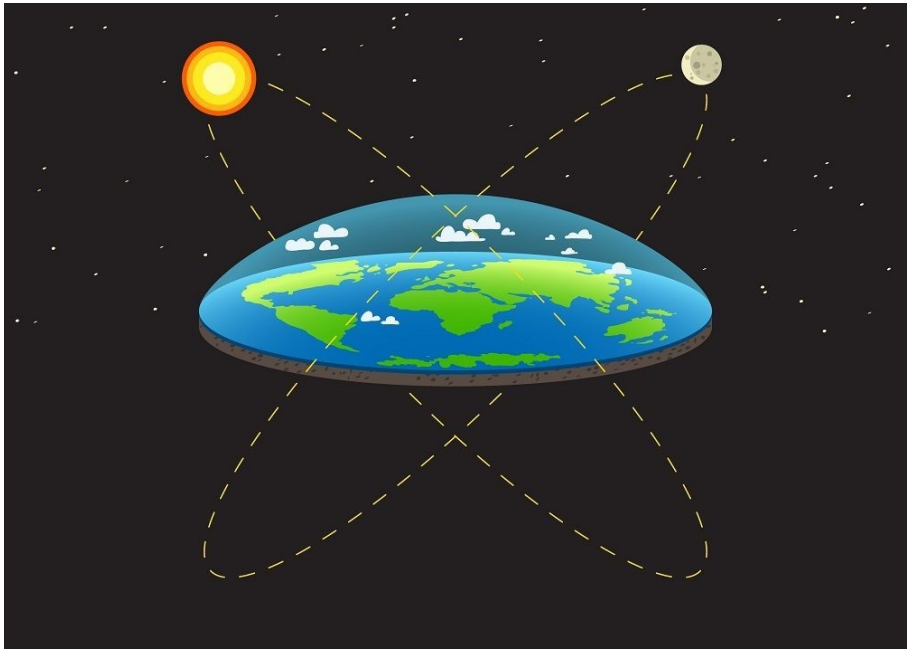
Se méfier de ce qui semble "évident" …

Envisager toutes les explications/hypothèses alternatives

Ex1 : que la terre soit ronde et que la terre tourne autour du soleil ne sont pas spécialement évident au premier abord …

Ni que vous êtes en train de vous déplacer à > 1000km/h (rotation de la terre...)

Ex 2 : les jeux vidéos/images violents comme catharsis



NB : l'intuition est souvent essentielle pour initier le processus scientifique, mais les intuitions doivent ensuite être passées au crible de la méthode scientifique pour être validées/invalidées

Autres sources de connaissance :

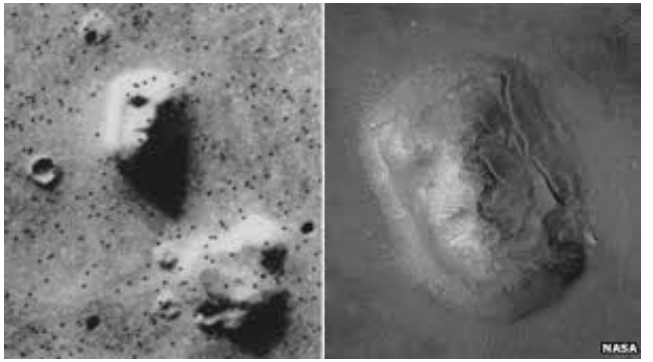
Observations non systématiques / expérience personnelle

Se méfier de ses sens et de sa mémoire

Pex: "Patternicity" : tendance à voir des "patterns" là où il n'y a que du bruit









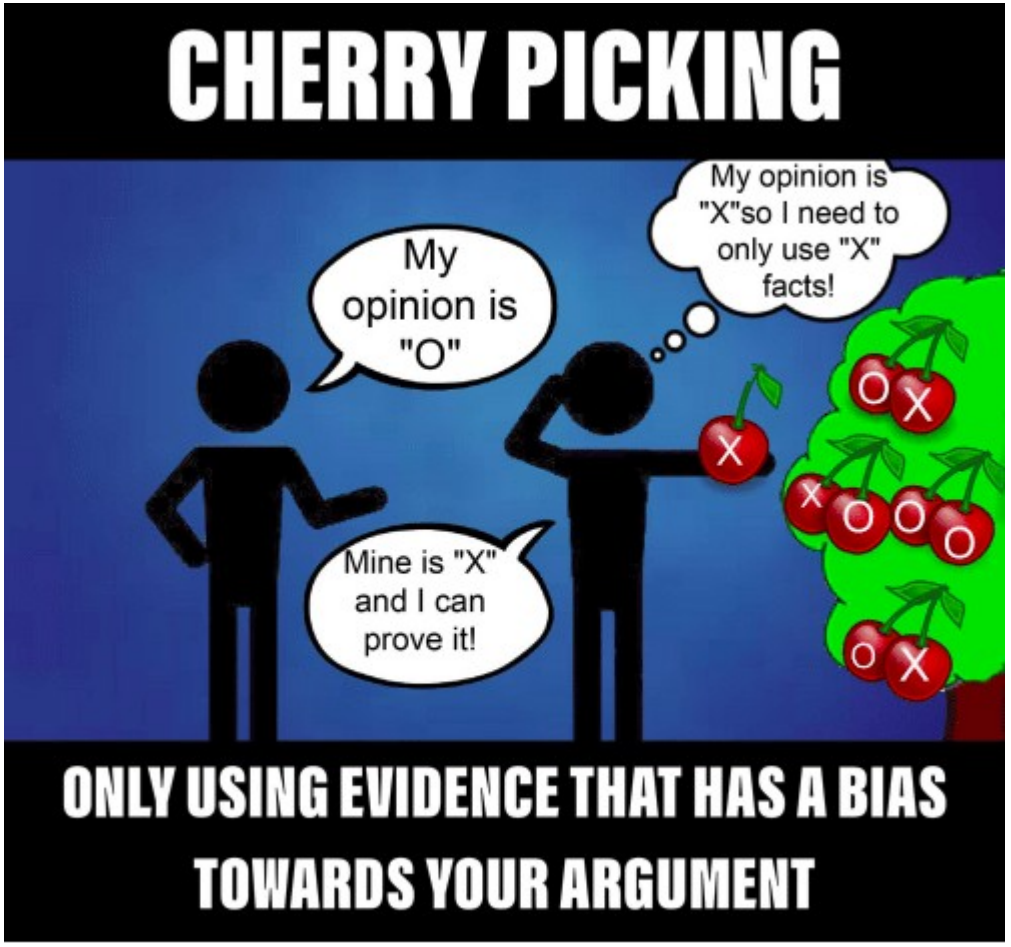
Autres sources de connaissance :

Observations non systématiques / expérience personnelle

Se méfier de ses sens et de sa mémoire

Pex: Tendance à remarquer/retenir ce qui confirme ses idées préconçues

"cherry picking"

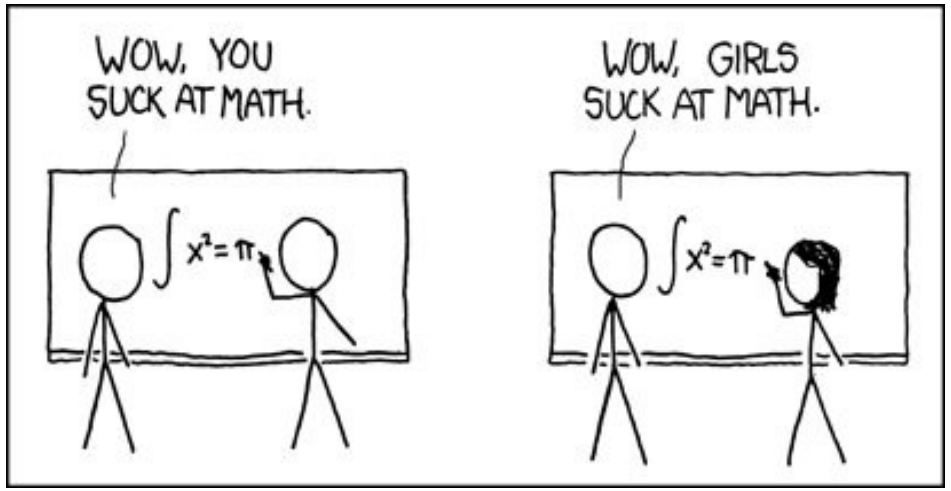


Autres sources de connaissance :

Observations non systématiques / expérience personnelle
Se méfier de ses sens et de sa mémoire

Tendance à généraliser sur base de cas particuliers
Tendance à retenir des faits "marquants" et à sous estimer les cas "banals"
Tendance à remarquer/retenir ce qui confirme ses idées préconçues

"*The plural of anecdote is no data*"
Exemple : l'influence de la lune sur les accouchements



THE SCIENCE NEWS CYCLE

Start Here

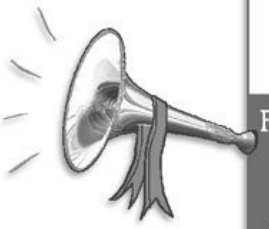
Your Research

Conclusion: **A is correlated with B (ρ=0.56)**, given C, assuming D and under E conditions.

...is translated by...

UNIVERSITY PR OFFICE
(YES, YOU HAVE ONE)

FOR IMMEDIATE RELEASE: SCIENTISTS FIND POTENTIAL LINK BETWEEN A AND B (UNDER CERTAIN CONDITIONS).



...which is then picked up by...

NEWS WIRE ORGANIZATIONS

A CAUSES B, SAY SCIENTISTS.

...who are read by ...

THE INTERNETS

...then noticed by...

Scientists out to kill us again.
POSTED BY RANDOM DUDE
Comments (377)
OMG! i kneeew it!!!
WTH???????
...

We saw it on a Blog!
A causes B all the time
What will this mean for Obama?
BREAKING NEWS BREAKING NEWS BREA

...and caught on ...

CNC Cable NEWS

4 LOCAL EYEWITLESS NEWS

...eventually making it to...

WHAT YOU DON'T KNOW ABOUT "A"... CAN KILL YOU! MORE AT 11...

YOUR GRANDMA

Qu'est-ce que la méthode scientifique ?

Définition exacte très difficile !!
Plus un ensemble d'approches reconnues comme valides par la communauté scientifique dans un domaine (avec de nombreuses exceptions) que "une" méthode.

Science/Scientific method
~=
Self correcting process to build collective knowledge
based on evidences and carefully tested hypotheses

"*Extraordinary Claims Require Extraordinary Evidence.*" - Carl Sagan

"*There are many hypotheses in science which are wrong. That's perfectly all right; they're the aperture to finding out what's right. Science is a self-correcting process. To be accepted, new ideas must survive the most rigorous standards of evidence and scrutiny.*" - Carl Sagan

Qu'est-ce que la méthode scientifique ?

Evidence (English) : pas vraiment d'équivalent en français !!!

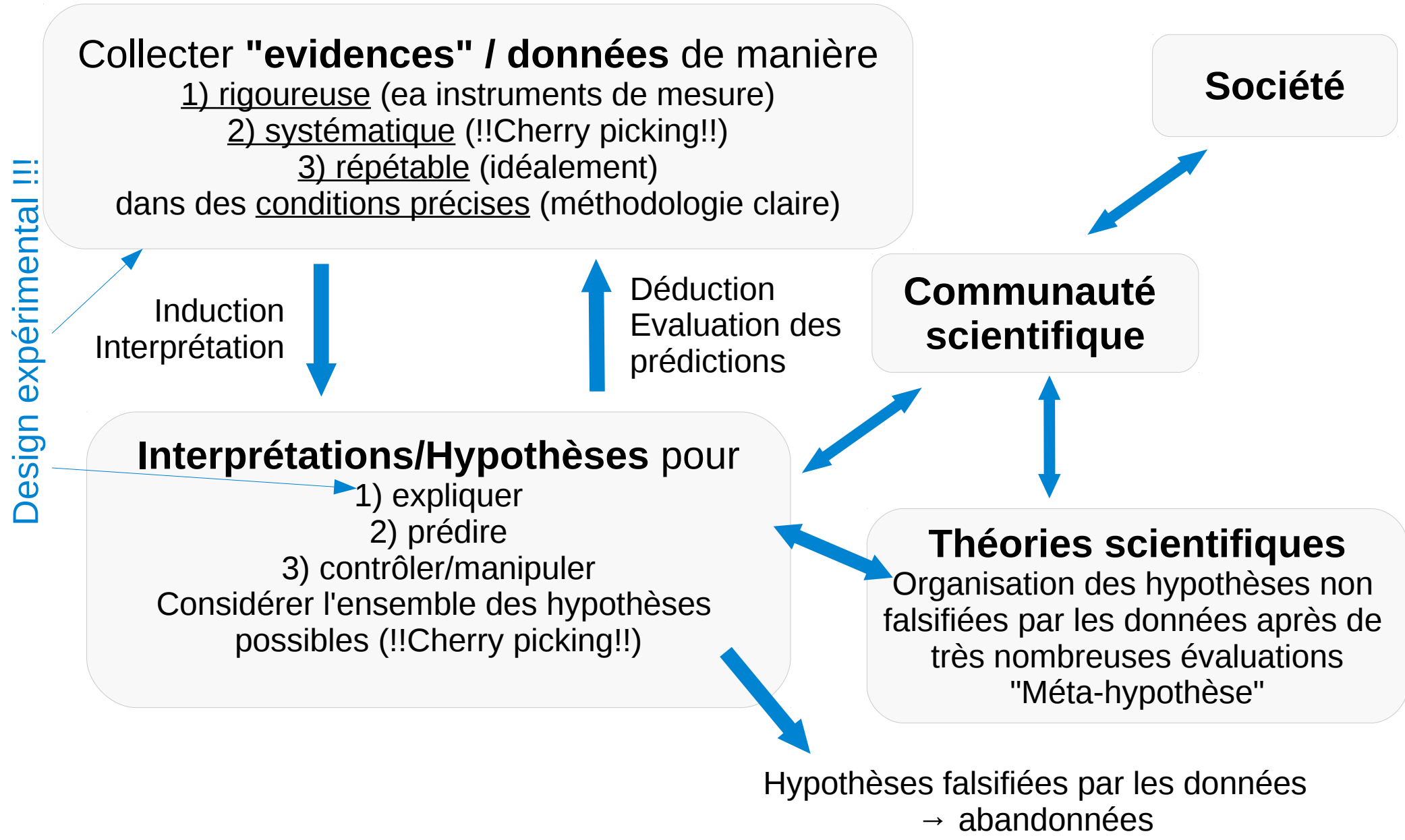
Facts or observations presented in support of an assertion.
(https://en.wiktionary.org/wiki/evidence)
Evidence is anything that you see, experience, read, or are told that causes you to believe that something is true or has really happened
(www.collinsdictionary.com)

Parfois traduit par "**preuve**" :
Ce qui établit la véracité d'une proposition ou d'un fait.
https://fr.wiktionary.org/wiki/preuve

"Preuve" a une connotation beaucoup plus définitive

Preuve : élément qui démontre une hypothèse
Evidence : élément qui fait pencher la balance en faveur d'une hypothèse
→ parfois traduit par "élément de preuve"

Qu'est-ce que la méthode scientifique ?



Qu'est-ce que la méthode scientifique ?

Important de distinguer :

1) les données récoltées (Résultats) +
méthode de récolte (Matériel & Méthodes)

→ nouvelles "evidences" / "faits" a mettre dans le pot commun

2) L'interprétation de ces résultats sur base de la logique et
en comparaison avec ce qu'on sait déjà (Discussion)

→ en général plus spéculatives.

Qu'est-ce que la méthode scientifique ?

Falsification des hypothèses

"~*You only need one white crow to disprove the rule that all crows are black*~ - William James

Réfuter une hypothèse est souvent plus facile que confirmer une hypothèse. Avant qu'une hypothèse puisse être "acceptée" il faut un grand nombre de cas où elle n'a pas été réfutée.

C'est un problème fréquent :
pex il est très difficile de "prouver" qu'un pesticide n'est pas toxique

→ on met souvent l'accent sur la falsification des hypothèses en science (critère de Popper)

Attention à ne pas trop vite éliminer une hypothèse quand les prédictions s'avèrent fausses !

Exemple : Orbite d'Uranus inconsistante avec la théorie Newtonienne → hypothèse (1846) : présence d'une planète inconnue → découverte en effet plus tard (Neptune)

Théories scientifiques

Différence de sens Science - Langage courant

Dans le langage courant :

Via le latin theoria « spéculation »
Dans le langage ordinaire, toute notion générale, ensemble d'idées, par comparaison avec une théorie scientifique. (Wiktionary)
→ forte connotation spéculative
("c'est juste une théorie")

En science : les théories acceptées par la communauté scientifique ne sont pas considérées comme "vraies" ou "fausses" mais comme la vision du monde la mieux supportée par les données/"evidences" disponibles à un moment donné.
→ ce n'est pas parfait mais c'est ce dont on dispose de mieux
(connotation beaucoup moins spéculative que dans le langage courant)

Dans de rares cas les théories sont confirmées par tellement d'observations qu'elles sont considérées - dans le langage courant - comme des faits objectifs (mais c'est rare).
Pex : Héliocentrisme, Evolution (mais le Darwinisme resterait une théorie)

Théories scientifiques

En science les théories sont un outil utilisé pour organiser les observations et hypothèses non falsifiées par les observations après de nombreuses évaluations.

Steven Goldman fait une excellente analogie avec les **cartes géographiques** :

La carte n'est pas le territoire…

Elles servent à se repérer dans le monde, à faire des prédictions

Elles ne sont pas figées, on les corrige en fonction des meilleures informations disponibles

Selon l'objectif, on peut utiliser des cartes/théories avec différent niveaux de précision/échelles.

Une carte routière de la France n'est pas "fausse" elle représente le monde avec un niveau de simplification utile pour certains usages. Une carte topographique est plus précise mais inutilisable dans bien des situations.

Selon l'objectif, on utilise différents types de cartes qui représentent différents aspects de la réalité (géologique, photo satellite,…) et on peut superposer plusieurs cartes.

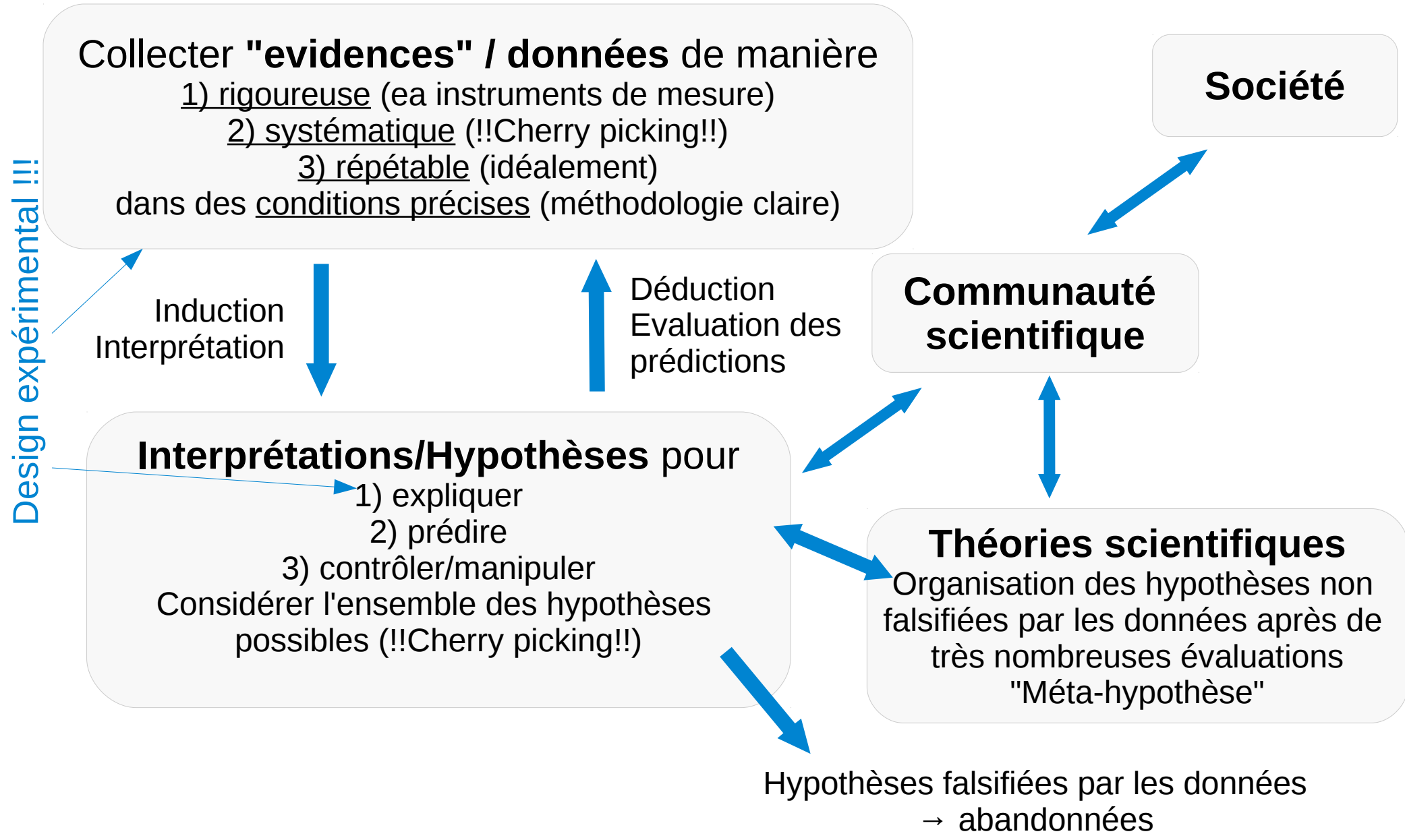
La représentation en 2D d'une sphère cause des distorsions Mais on peut choisir par exemple des projections qui conservent les directions (pex pour la navigation) ou les surfaces

Théories scientifiques

Préférence pour les théories qui :

permettent d'expliquer un maximum de faits
sont parcimonieuses
permettent de faire des prédictions précises de phénomènes diversifiés
ont un large spectre d'applications
sont consistantes avec d'autres théories/domaines
sont génératrices de nouvelles idées

Qu'est-ce que la méthode scientifique ?



Planification expérimentale : Pourquoi ?

Nombreux pièges cognitifs, facteurs de confusion, grande variabilité,…

La planification expérimentale peut servir à :

1) **éviter/limiter certains problèmes** :

- biais dans la récolte de données
(pex : échantillonnage non aléatoire)

- interprétation incorrecte des résultats
(pex : confusion de facteurs, absence de contrôles)

- invalidité des inférences (on généralise alors qu'on ne devrait pas)
(pex : non indépendance des échantillons, pseudoréplication)

2) **Augmenter la puissance d'un test** (diminuer les p-valeurs*)
et/ou la précision des estimations, idéalement à coût égal
autrement-dit : à détecter des effets plus subtils

*NB : définition fausse mais çà donne une idée de l'objectif...

Planification expérimentale : Pourquoi ?

Il est évidemment primordial de penser à tout çà avant de commencer l'étude…

Encore plus pour les études longitudinales où on aura pas la possibilité d'adapter un protocole mal conçu au départ

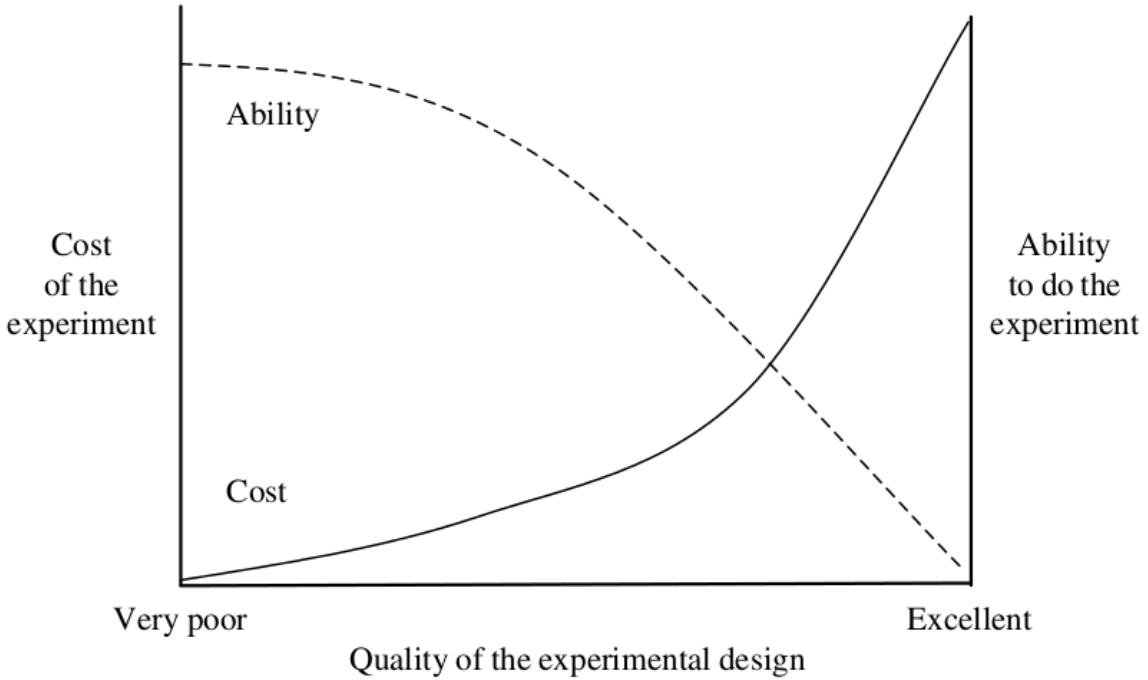
"To call in the statistician after the experiment is done may be no more than asking him to perform a postmortem examination: he may be able to say what the experiment died of." R. Fisher

Ne pensez pas qu'il y aura toujours un "truc" statistique pour contourner tous les problèmes dans la récolte des données !!

Planification expérimentale : Pourquoi ?

Compromis entre :

design expérimental optimal
et
contraintes pratiques - moyens disponibles



Planification expérimentale : Pourquoi ?

→ On doit parfois laisser tomber certaines "règles"
Mais bien avoir en tête ces "règles d'or" permet :

1) d'ajuster quand même son design expérimental au mieux.
Souvent on aurait pu faire mieux avec les mêmes moyens

2) si il y a vraiment certaines règles qu'on doit laisser tomber, il faut le garder en tête pour l'interprétation des résultats
(présence de biais, confusion d'effets, impossibilité de généraliser ses résultats à d'autres cas,...)

3) dans certains cas il faut se rendre à l'évidence avant de commencer l'étude qu'on ne pourra pas répondre à la question

...

(2) Principales approches

Etudes observatives vs expérimentales

Types d'approche :

Expérimentale ("controled/manipulative experiments") :
L'expérimentateur contrôle tous les paramètres et en fait varier certains
--> met en évidence des liens de cause à effet (sous certaines conditions)
Mais : les conditions sont-elles réalistes ?

Semi-expérimentale
On manipule le facteur d'intérêt (traitement) mais on ne contrôle pas les conditions
exemple typique : essais agronomiques en champs, essais cliniques
Souvent plus réalistes mais aussi moins répétables

Observative ("natural experiments") :
On échantillonne dans une variété de conditions existantes sans pouvoir les contrôler
--> met en évidence des corrélations mais plus difficilement des lien de cause à effet
Mais : conditions souvent les plus réalistes
Souvent la seule approche possible !

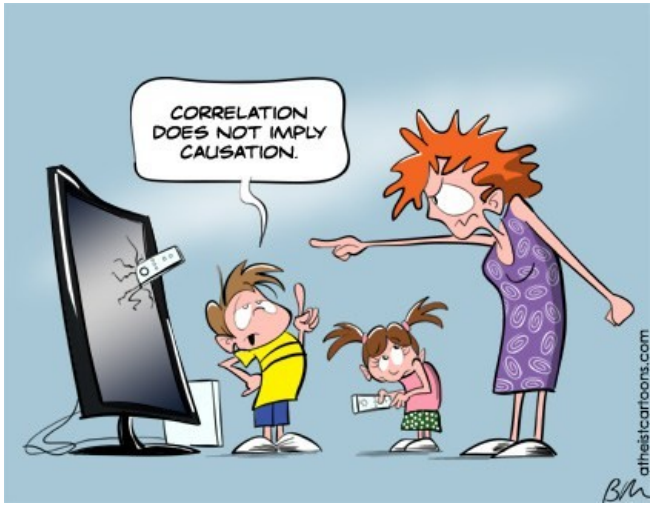
NB : les règles d'or de la planification expérimentale concernent aussi
en partie les études observatives !

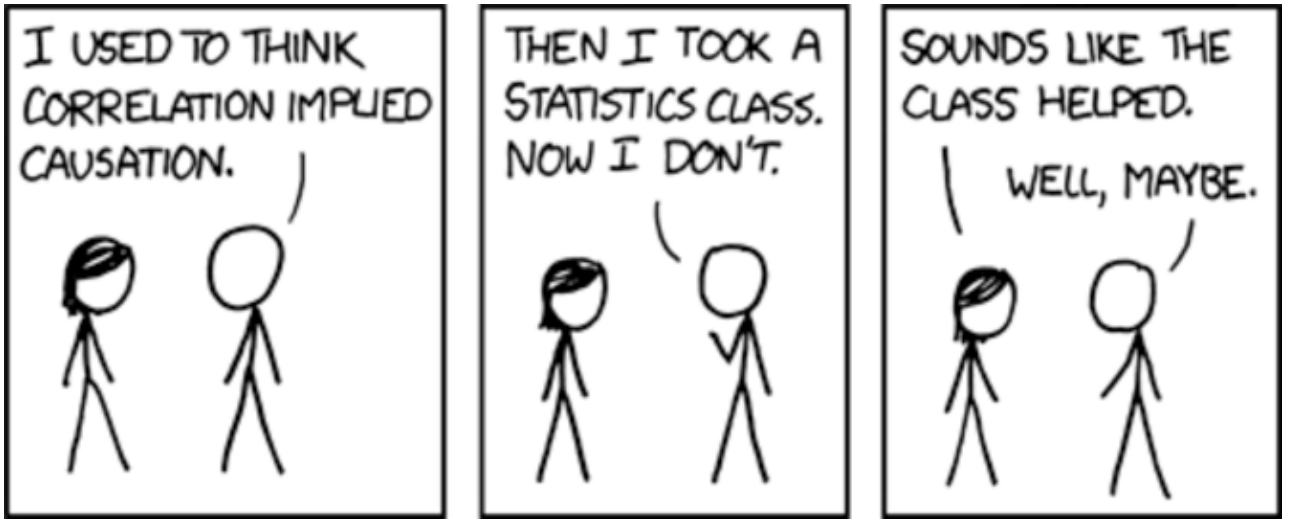
Les études **expérimentales** contrôlées et randomisées sont souvent préférées par les scientifiques. Bien que très puissantes quand elles sont possibles, elles ne sont pas la panacée..

Pex : les études in vitro sur cellules souches montrent les liens de cause à effet mais elles ne sont qu'une étape préliminaire dans un très long processus d'évaluation...

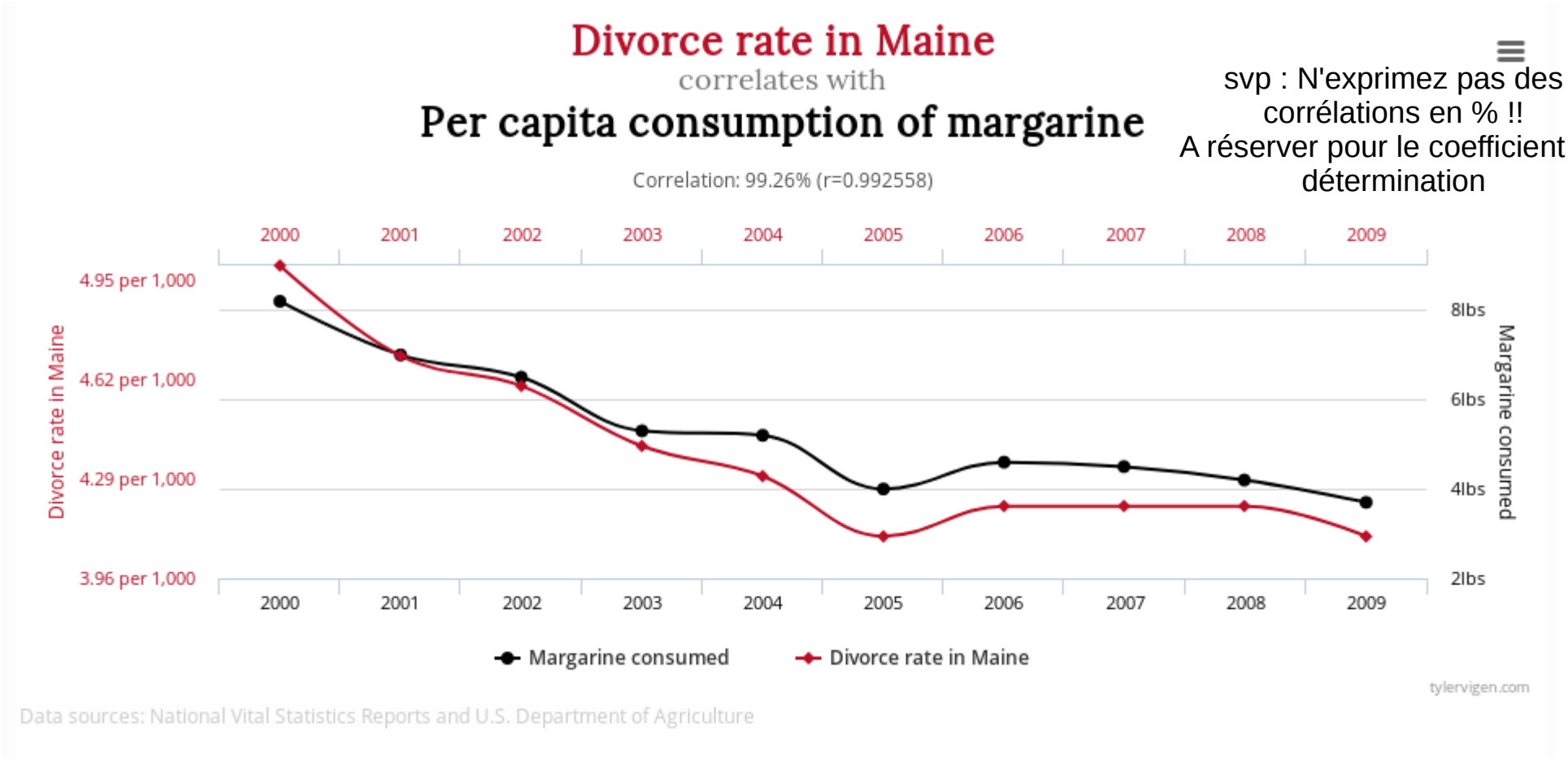
"Corrélation n'implique pas causalité"

Les études **observatives** sont souvent critiquées à cause du lieu commun : "Corrélation n'implique pas causalité"





"Corrélation n'implique pas causalité"



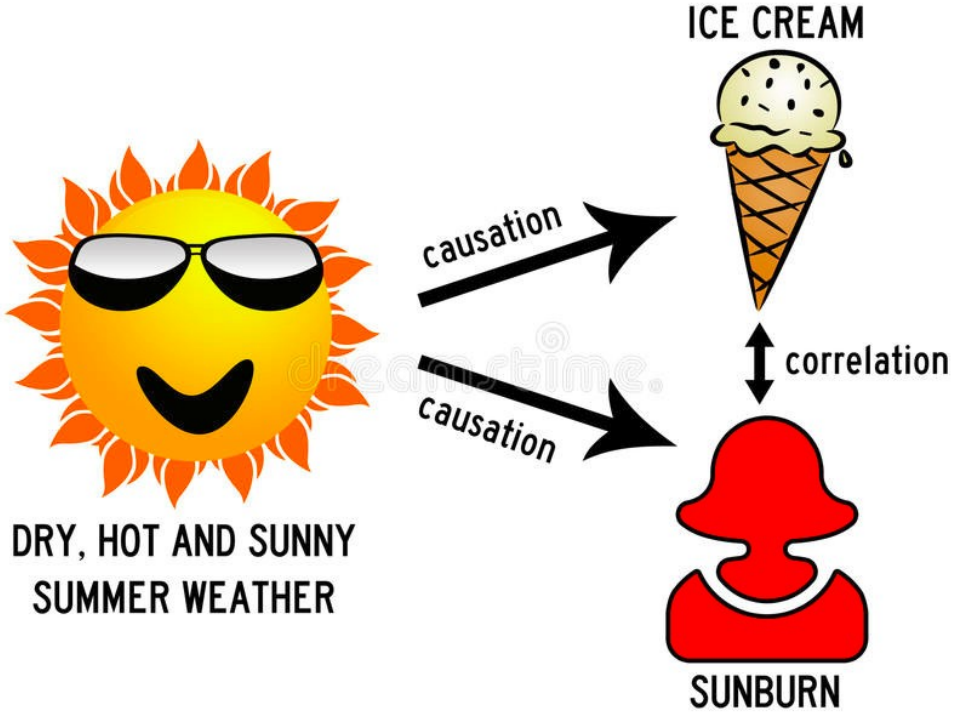
svp : N'exprimez pas des corrélations en % !!
A réserver pour le coefficient de détermination

2 variables sans aucun lien logique
ordonnées dans le temps (simple coïncidence)
seulement 10 points
combien de paires de variables non corrélées ont été comparées avant d'arriver à ce résultat ?

tp://www.tylervigen.com/spurious-correlations

"Corrélation n'implique pas causalité"

"In fact, with few exceptions, correlation does imply causation." More precisely : "a simple correlation implies an unresolved causal structure, since we cannot know which is the cause and which is the effect or if both are common effects of other unmeasured variables."* - Bill Shipley (2016)



A corrélé avec B :
4 possibilités principales

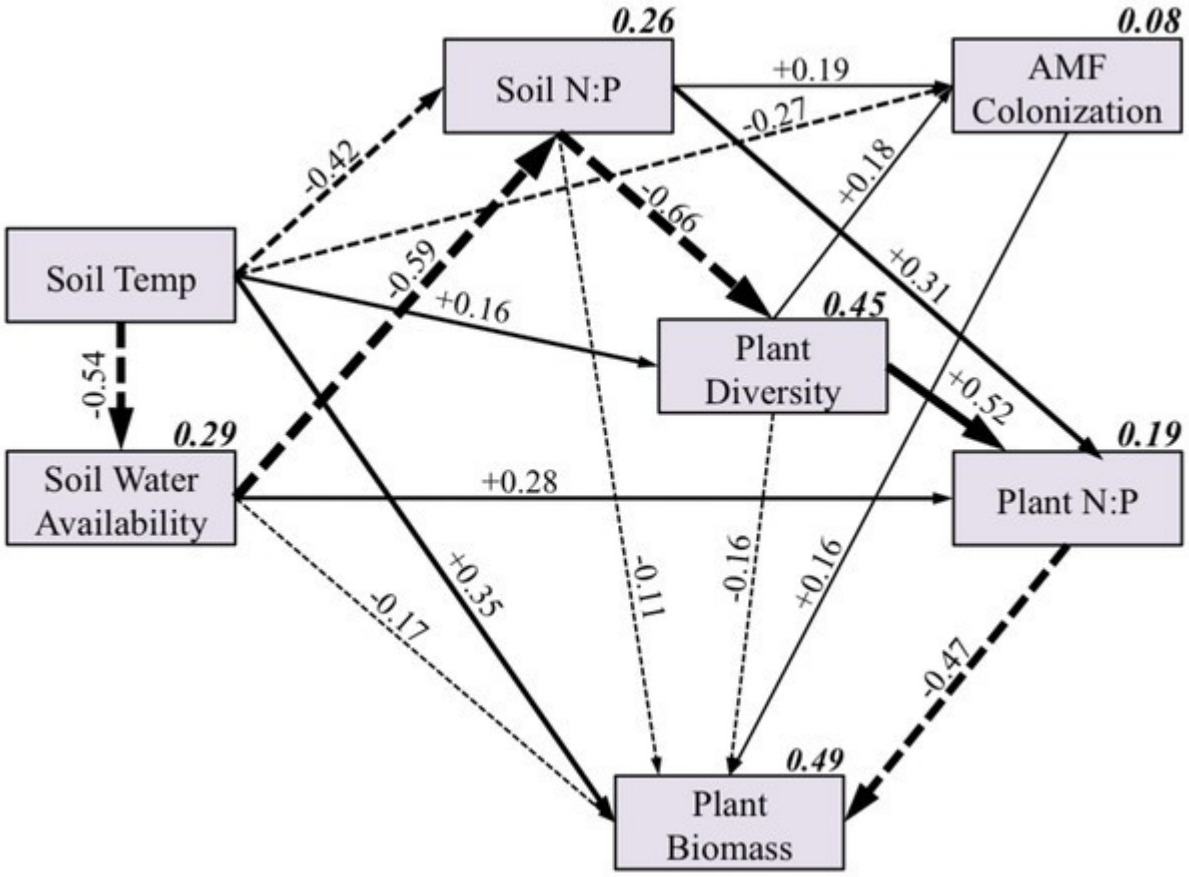
1) A cause B
2) B cause A
3) Une variable extérieure "C" cause A et B
4) Aucun lien entre A et B
la corrélation est juste un hasard de l'échantillonnage

* eg variables that covary only because they are time ordered

Shipley B (2016) Cause and Correlation in Biology: A User's Guide to Path Analysis, Structural Equations and Causal Inference with R. Cambridge University Press

"Corrélation n'implique pas causalité"

Certains outils statistiques peuvent aider à comparer des hypothèses sur les relations de causalité entre variables (structure) et des données observatives.
Eg "Structural Equation Modelling"
Très utilisé en sciences sociales.



ttps://peerj.com/articles/2083/

"Corrélation n'implique pas causalité"

La manière de collecter les données observatives peut aussi avoir un effet sur la facilité à en extraire des "evidences" sur des liens de cause à effet.

Exemple : "Quasi-Experimental designs" (Sciences Sociales)
plus ou moins équivalents aux "BACI designs" en Biologie
Before/After Control/Impact Designs

= Etudes observatives longitudinales (observations répétées dans le temps)
Mesures (O) avant et après un événement/ traitement que l'on veut étudier (X)
Idéalement avec suffisamment de répétitions
Idéalement : X ne se produit pas au même moment
et/ou Idéalement : groupes témoins (sans X)

O O O O O X O O O O O
O O O X O O O O O O O
O O O O O O O O X O O
O O O O O O O O O O O
O O O O O O O O O O O
etc.

Exemple : étude de l'impact d'une espèce invasive sur les populations indigènes
Comparaison des tendances de population avant/après l'arrivée
(décalée dans le temps selon les régions)

Etudes observatives vs expérimentales

Une seule étude, un seul article ne permet presque jamais de "prouver scientifiquement" un fait, en particulier dans les sciences de la vie et en sciences humaines.

Ce n'est que l'accumulation considérable d'études observatives et/ou expérimentales bien menées (voir "règles d'or") qui peuvent permettre à une communauté scientifique d'arriver à un consensus voire une certitude.

La répétabilité des résultats est un élément essentiel !
La "triangulation"* est peut-être encore plus essentielle !

*Munafo MR, Smith GD (2018) Repeating experiments is not enough. Nature 553:399–401

ILLUSTRATION BY DAVID PARKINS

Repeating experiments is not enough

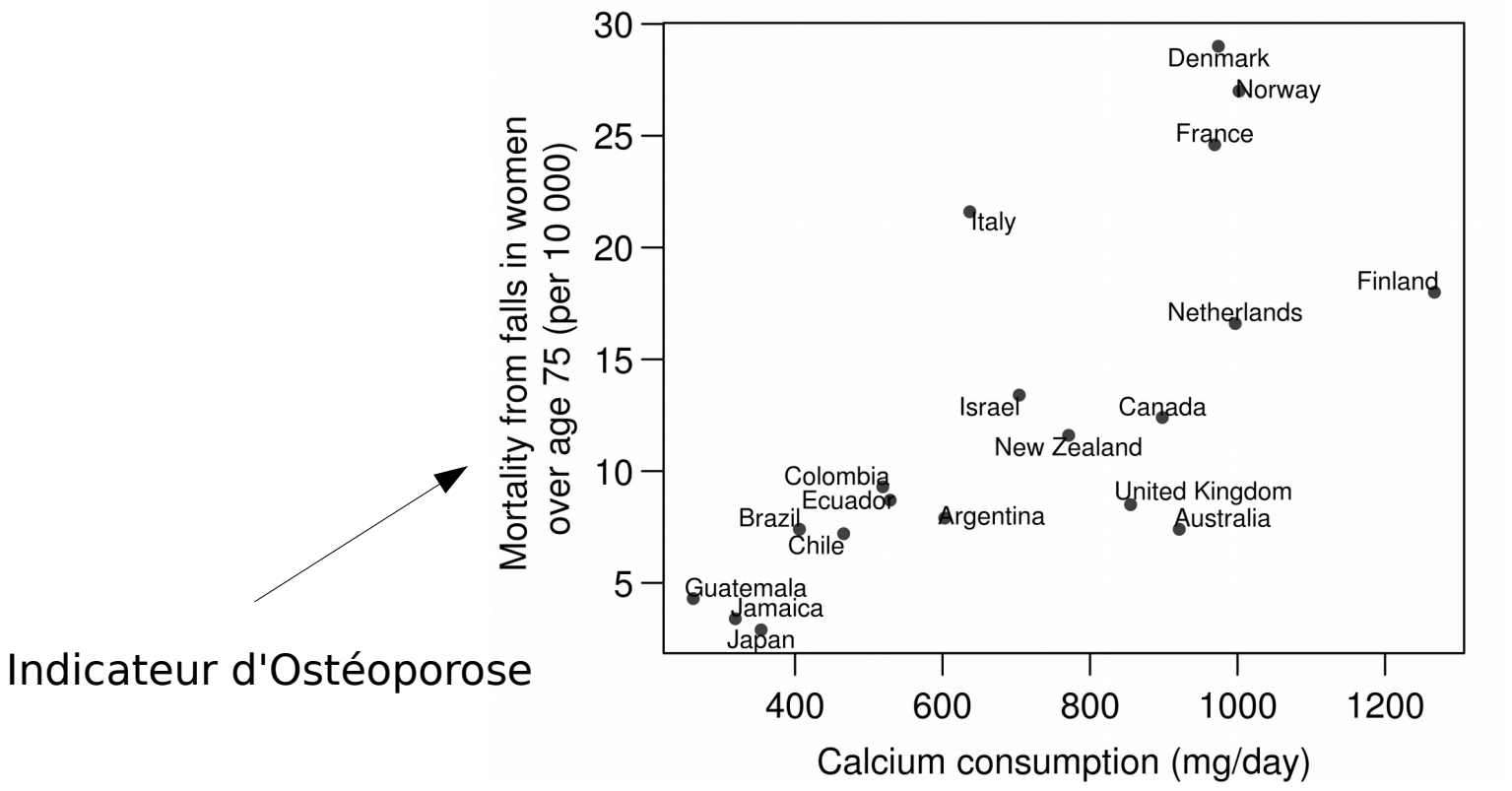
Verifying results requires disparate lines of evidence — a technique called triangulation. **Marcus R. Munafò** and **George Davey Smith** explain.

*Munafo MR, Smith GD (2018) Repeating experiments is not enough. Nature 553:399–401

Exemple : Paradoxe du calcium

Exemple de l'intérêt des études observatives pour se poser les bonnes questions et démêler un réseau de relation causales.

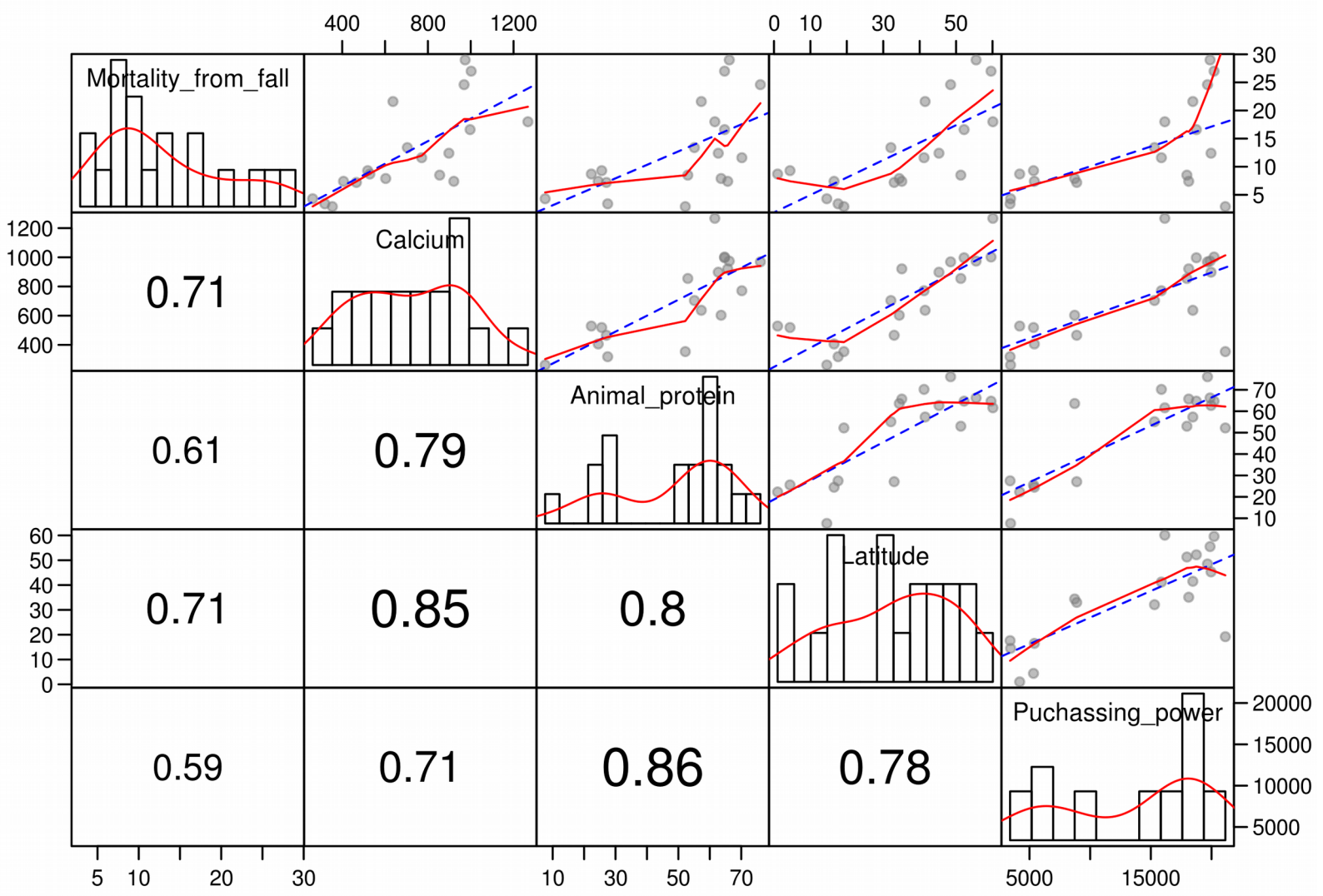
Le "bon sens" voudrait que dans les pays où on consomme le plus de calcium on ait moins de problèmes d'ostéoporose...



Données FAO : http://www.fao.org/docrep/w7336t/w7336t04.htm

Exemple : Paradoxe du calcium

Nombreux facteurs corrélés/confondus...



Données FAO : http://www.fao.org/docrep/w7336t/w7336t04.htm

Exemple : Paradoxe du calcium

Paradoxe du calcium

La calcémie dépend de plusieurs facteurs :

- la quantité de calcium dans la nourriture

- la capacité à absorber ce calcium
qui diminue quand la vitamine D diminue et donc quand la latitude augmente

- la perte de calcium via l'urine
qui augmente par exemple avec la consommation de protéines animales et de sel

- etc...

Ce n'est qu'en combinant des études observatives et expérimentales qu'on a pu comprendre les interactions complexes entre ces facteurs corrélés entre eux...

Exemple : Etudes Abeilles

(Thèse Noa Simon-Delso - CARI)

1) Etude observative :

colonies d'abeilles présentant des symptômes de mortalité hivernale
dosage des pesticides et des virus dans la ruche
caractérisation du paysage agricole autour de la ruche

Résultats :

pas de lien observé avec les virus ni avec les insecticides
une relation inattendue mortalité ~ fongicides
pas de lien observé avec la présence de cultures suspectes (ea colza)
une relation positive avec la surface de cultures et négative avec la surface de prairies

Exemple : Etudes Abeilles

(Thèse Noa Simon-Delso - CARI)

Pas de relation observée ne veut pas dire qu'il n'y a pas de lien !
La charge de la preuve est du côté du chercheur...

On a peut être simplement pas assez de répétitions dans cette étude (pas assez de puissance statistique).
Démontrer une absence d'effet est presque impossible...

Certains critiquent d'ailleurs l’extrême prudence de la communauté scientifique ea le fameux seuil alpha = 0.05

voir pex : Conway & Oreskes (2014)

D'autres au contraire pensent qu'on devrait être plus exigent pour augmenter la la reproductibilité!

Exemple : Etudes Abeilles

(Thèse Noa Simon-Delso - CARI)

Le fait qu'on ait trouvé un lien significatif déclin~fongicides (p = 0.008) n'est pas non plus une preuve définitive

La probabilité de trouver un tel résultat par chance est faible mais pas nulle
Le résultat est peut-être particulier à la population échantillonnée (nord de la Wallonie) et pas extrapolable à d'autres contextes
On ne peut jamais exclure des erreurs, biais, etc... même si l'étude a été faite de manière compétente et honnête

--> ce n'est qu'en trouvant des résultats similaires dans d'autres études, d'autres contextes, que ces observations pourront être confirmées...

Exemple : Etudes Abeilles

(Thèse Noa Simon-Delso - CARI)

Pourquoi ce lien inattendu mortalité hivernale ~ fongicides ?
Étude observative : difficile d'établir un lien de causalité

Hypothèses :

1) Lien de causalité direct ou indirect (par quel mécanisme?)
2) Fongicides comme marqueurs d'insecticides non détectés
3) Fongicides marqueurs de milieux agricoles pauvres en nourriture
4) Fongicides marqueurs de conditions climatiques humides
etc

Exemple : Etudes Abeilles

(Thèse Noa Simon-Delso - CARI)

Lien de causalité direct déclin ~ fongicide ?

2) Test expérimental en labo

Ouvrières exposées au fongicide le plus fréquent

1) protocole standard, suivi 10 jours : pas d'effet
2) si on prolonge le suivi au delà de 10 jours :
mortalité très forte à toutes les doses (pas dans le contrôle)

Questionne les protocoles standard...
On peut probablement conclure à un lien de causalité
MAIS
Est-ce que conditions de labo sont représentatives de la nature ?
Doses ? Condition de vie des abeilles ?
Conditions de "vie" du produit ?

Exemple : Etudes Abeilles

(Thèse Noa Simon-Delso - CARI)

Une solution ?

3) Test semi-expérimentaux au champ

On pulvérise le produit sur une culture et on y place les abeilles pex en tunnel

Très coûteux --> le risque est de ne pas faire assez de répétitions pour mettre en évidence des différences subtiles
(plusieurs dizaines de ruches)
--> importance d'analyser la puissance statistique d'un dispositif expérimental

De plus, le tunnel n'est pas très naturel
Interaction avec les conditions climatiques, la culture, d'autres produits,...

Conclusion

--> Même si toutes les études ne se valent pas,
toutes les études sont critiquables
Le doute fait partie de la méthode scientifique.
Ce n'est pas parce qu'il y a un doute sur les résultats/l'interprétation qu'une étude est sans valeur et doit être oubliée.

Ce n'est qu'en combinant différentes études et approches qu'on peut comprendre un phénomène

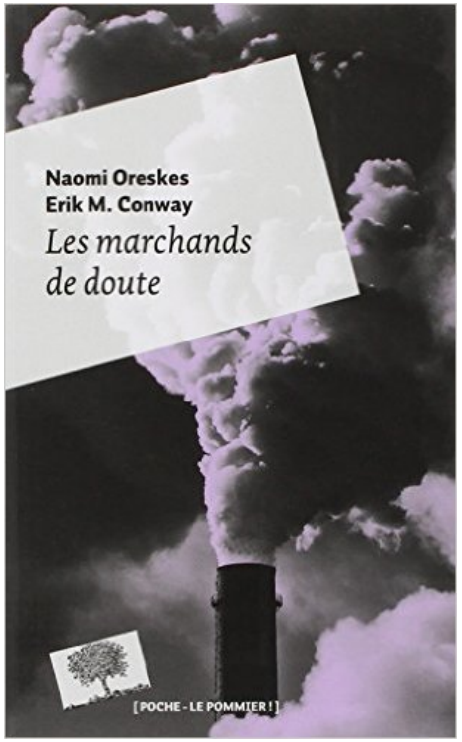
Les outils statistiques peuvent aider a évaluer le degré de certitude d'un résultat.

Pour que les résultats soient exploitables, il faut suivre certaines règles de "design expérimental" garantissant un minimum de qualité
→ sections 3 & 4

Livres

Comment autres acteurs de la société
(citoyens, politiques, journalistes, industries, services publics,...)
voient/utilisent/influencent la science ?

2 livres intéressants sur le sujet :

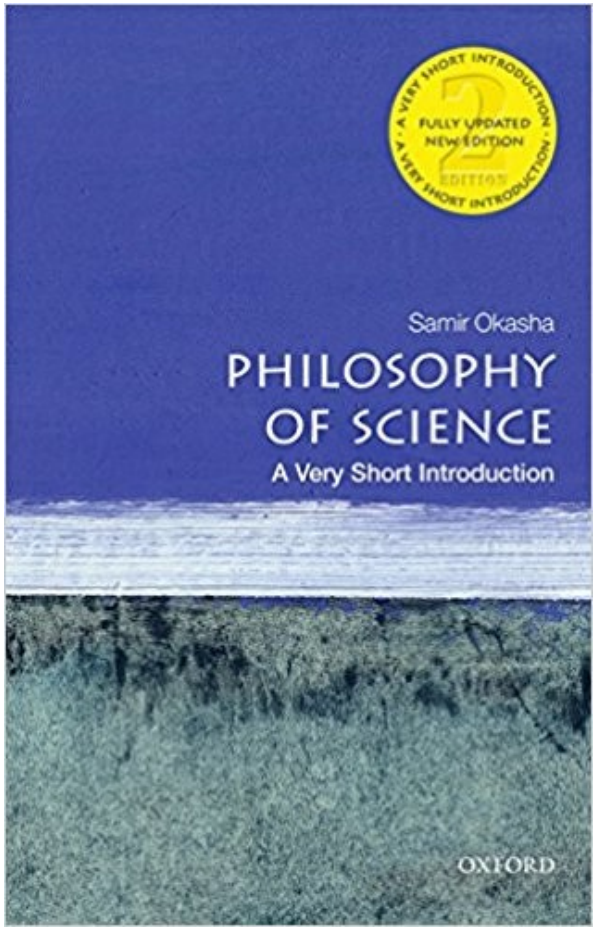


Auteurs : 2 historiens américains

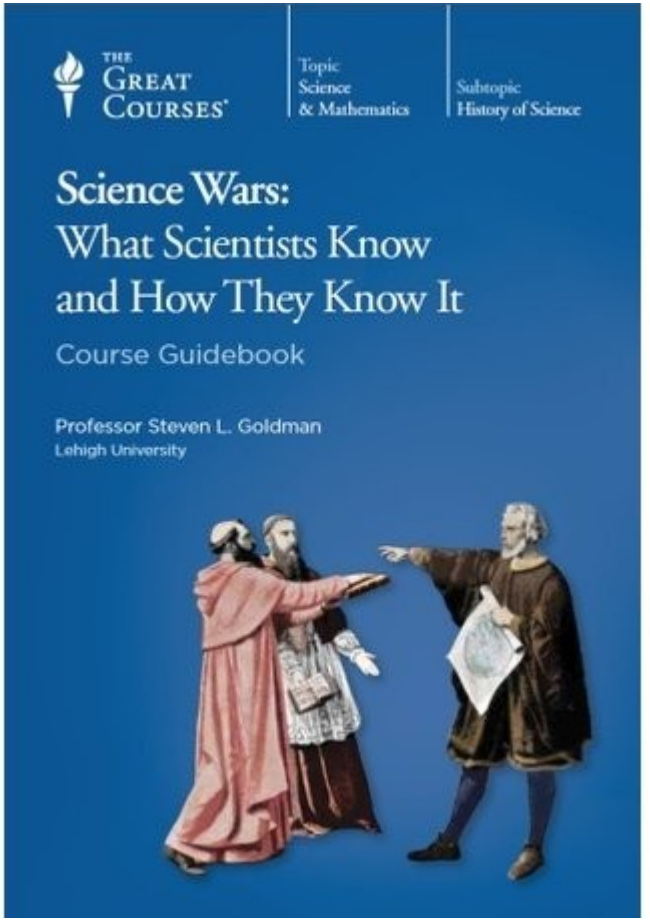
Auteur : 1 journaliste français

Livres

(Histoire de la) Philosophie des sciences
Plus théorique...



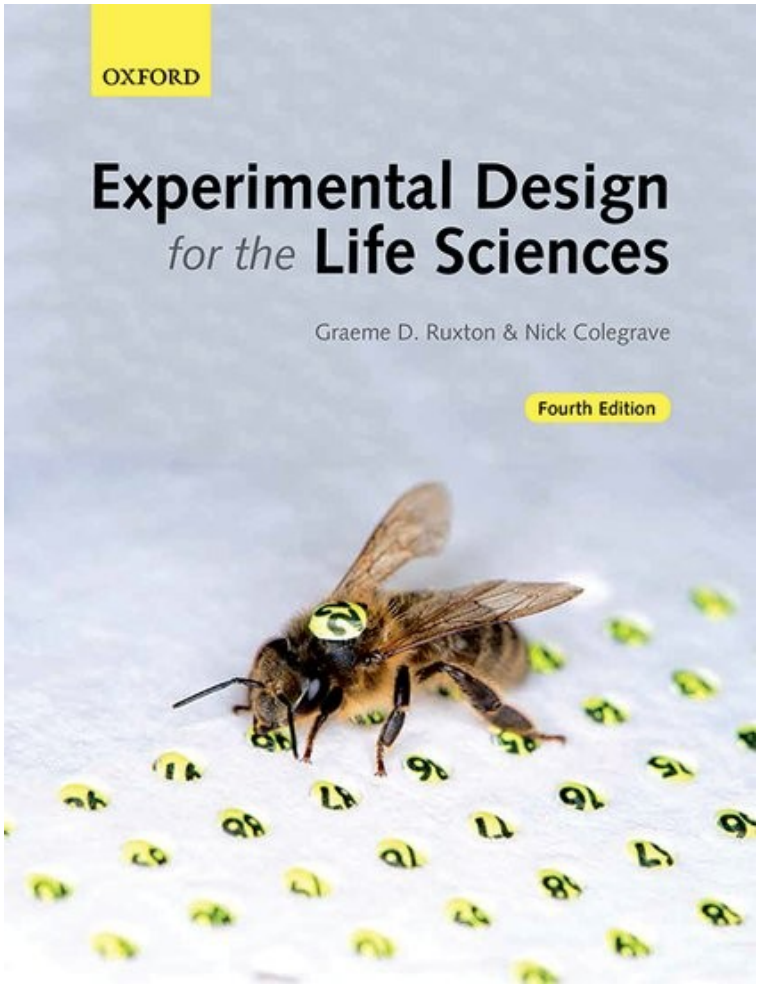
Courte intro agréable et accessible
(Samir Okasha)



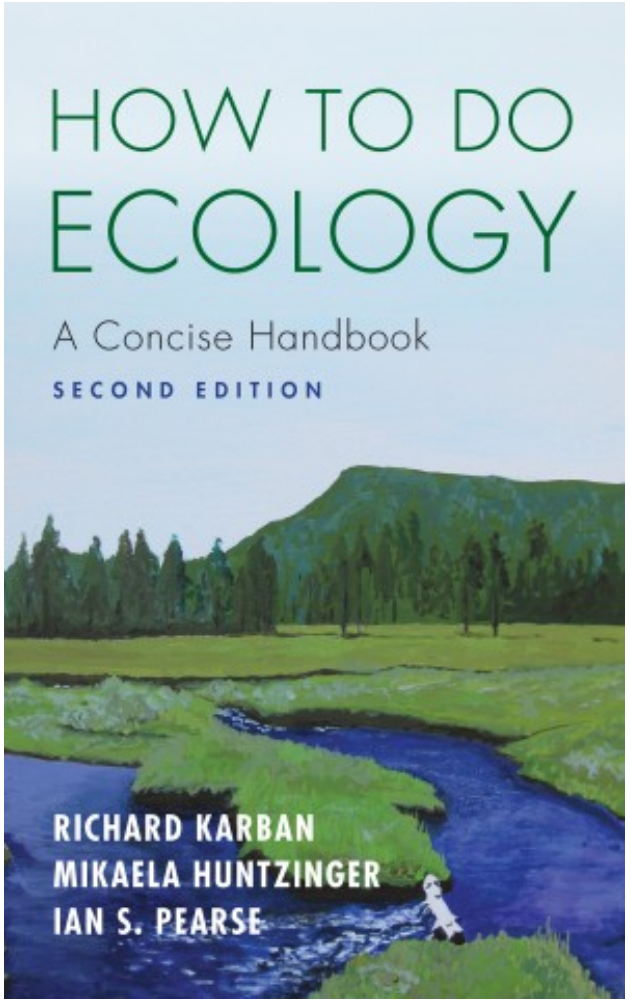
Audiobook - perspective historique
(Steven Goldman)

Livres

Ouvrages pratiques et non mathématiques traitant (au moins en partie) de la pratique des designs expérimentaux.
Pour les scientifiques (biologistes)



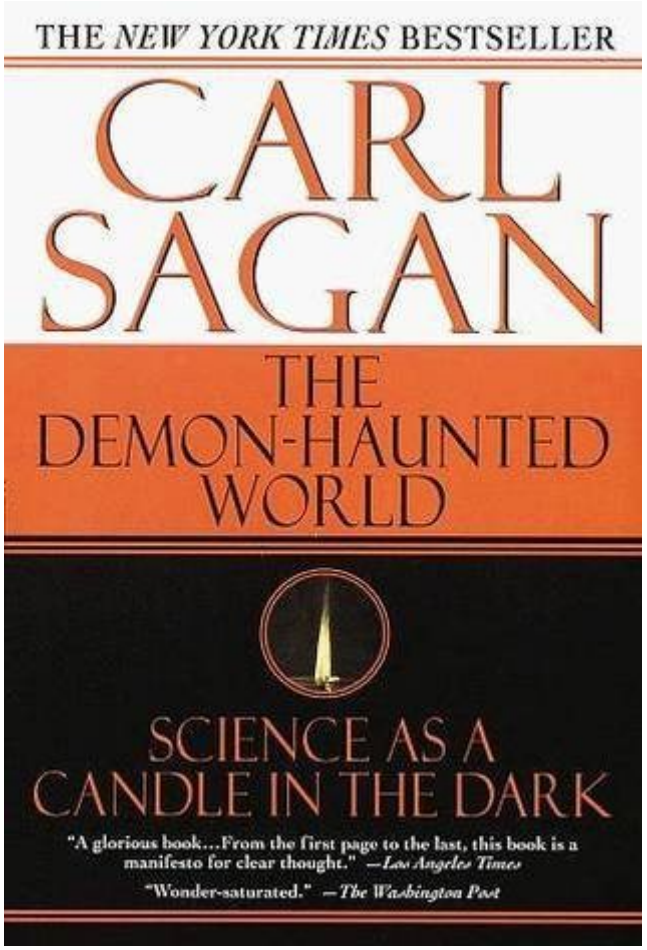
Plus complet



Très court - très accessible !

Livres

Sur le fonctionnement de la science
Pour le grand public



Très peu d'ouvrages convaincants !
Si vous en trouvez : merci de m'en faire part...

Sutherland WJ, Spiegelhalter D, Burgman M (2013) Policy: Twenty tips for interpreting scientific claims. Nature News 503:335.

Carl Sagan : Passion et Humilité
le chapitre le plus connu for "everyday skeptisism":
"The Fine Art of Baloney Detection"
ww.inf.fu-berlin.de/lehre/pmo/eng/Sagan-Baloney.pdf

A ROUGH GUIDE TO SPOTTING

BAD SCIENCE

1. SENSATIONALISED HEADLINES

Headlines of articles are commonly designed to entice viewers into clicking on and reading the article. At best, they over-simplify the findings of research. At worst, they sensationalise and mis-represent them.

2. MISINTERPRETED RESULTS

News articles sometimes distort or misinterpret the findings of research for the sake of a good story, intentionally or otherwise. If possible, try to read the original research, rather than relying on the article based on it for information.

3. CONFLICT OF INTERESTS

Many companies employ scientists to carry out and publish research - whilst this does not necessarily invalidate research, it should be analysed with this in mind. Research can also be misrepresented for personal or financial gain.

4. CORRELATION & CAUSATION

Be wary of confusion of correlation & causation. Correlation between two variables doesn't automatically mean one causes the other. Global warming has increased since the 1800s, and pirate numbers decreased, but lack of pirates doesn't cause global warming.

5. SPECULATIVE LANGUAGE

Speculations from research are just that - speculation. Be on the look out for words such as 'may', 'could', 'might', and others, as it is unlikely the research provides hard evidence for any conclusions they precede.

6. SAMPLE SIZE TOO SMALL

In trials, the smaller a sample size, the lower the confidence in the results from that sample. Conclusions drawn should be considered with this in mind, though in some cases small samples are unavoidable. It may be cause for suspicion if a large sample was possible but avoided.

7. UNREPRESENTATIVE SAMPLES

In human trials, researchers will try to select individuals that are representative of a larger population. If the sample is different from the population as a whole, then the conclusions may well also be different.

8. NO CONTROL GROUP USED

In clinical trials, results from test subjects should be compared to a 'control group' not given the substance being tested. Groups should also be allocated randomly. In general experiments, a control test should be used where all variables are controlled.

9. NO BLIND TESTING USED

To prevent any bias, subjects should not know if they are in the test or the control group. In double-blind testing, even researchers don't know which group subjects are in until after testing. Note, blind testing isn't always feasible, or ethical.

10. 'CHERRY-PICKED' RESULTS

This involves selecting data from experiments which supports the conclusion of the research, whilst ignoring those that do not. If a research paper draws conclusions from a selection of its results, not all, it may be cherry-picking.

11. UNREPLICABLE RESULTS

Results should be replicable by independent research, and tested over a wide range of conditions (where possible) to ensure they are generalisable. Extraordinary claims require extraordinary evidence - that is, much more than one independent study!

12. JOURNALS & CITATIONS

Research published to major journals will have undergone a review process, but can still be flawed, so should still be evaluated with these points in mind. Similarly, large numbers of citations do not always indicate that research is highly regarded.

(3) Les règles d'or de la planification expérimentale

Les "règles d'or" de la planification expérimentale

NB : questions à se poser avant de récolter des données !!!

Quelques règles d'or pour **éviter/limiter les problèmes** :

1) question et population d'intérêt bien définies
2) adéquation des mesures
3) réplication
4) indépendance des échantillons/réplicats
5) échantillonnage aléatoire
6) randomisation des mesures et des traitements
7) contrôles/témoins judicieusement choisis

Quelques règles d'or pour
augmenter la puissance d'un test
et/ou la précision des estimations :

1) taille d'échantillon maximisée
2) variabilité résiduelle minimisée
3) taille de l'effet maximisée

Les "règles d'or" de la planification expérimentale

1) Question et population d'intérêt bien définies

--> détermine tout le reste et le niveau auquel on veut pouvoir généraliser les résultats (la population statistique)

2) Adéquation des mesures

Est-ce que ce qu'on mesure correspond bien à la question ?
Si il y a des erreurs de mesure, quel est leur impact réel ?

3) Réplication

Permet d'estimer la variabilité des résultats et de voir à quel point les résultats sont extrapolables à d'autres cas (inférence stat.)

4) Indépendance des échantillons/réplicats

Pour éviter la pseudoréplication et obtenir des inférences valides

Si les échantillons ne sont pas indépendants on peut dans une certaine limite en tenir compte avec des méthodes statistiques adaptées (pex Modèles mixtes)

Les "règles d'or" de la planification expérimentale

5) Échantillonnage aléatoire
Pour que l'échantillon soit représentatif de la population

6) Randomisation des mesures et des traitements
Pour éviter/limiter les confusion de facteurs explicatifs

7) Contrôles/témoins judicieusement choisis
Pour obtenir un point de comparaison "honnête"
Pour éliminer des hypothèses alternatives qui pourraient expliquer les résultats

Question et population d'intérêt bien définies

Toujours avoir des questions clairement exprimées
et idéalement des hypothèses et prédictions sur les résultats.

Questions générales : définissent le contexte
Ex : quel est l'impact des fongicides sur les abeilles ?

Questions spécifiques : comment traduire cette question générale en questions "actionables" et réellement étudiables ?

Ex : Est-ce que la mortalité après 2 jours augmente lorsque la dose du fongicide A ingéré par voie orale augmente en conditions contrôlées ?
Ex : Est-ce que les colonies d'abeilles présentent plus de mortalité hivernale lorsque le pollen récolté en septembre contient un plus grand nombre de fongicides (ou une plus grande dose?)

Pas toujours simple de passer des questions générales aux questions spécifiques...

Question et population d'intérêt bien définies

Réfléchir à la manière dont on organiserait les données dans des tableaux et comment on analyserait très approximativement ces données est **très très** utile pour préciser les questions (en particulier avant même d'avoir récolé les données)

Pex :

Quelles sont les variables (colonnes)

Quelles sont les observations/unités d'échantillonnage (lignes)

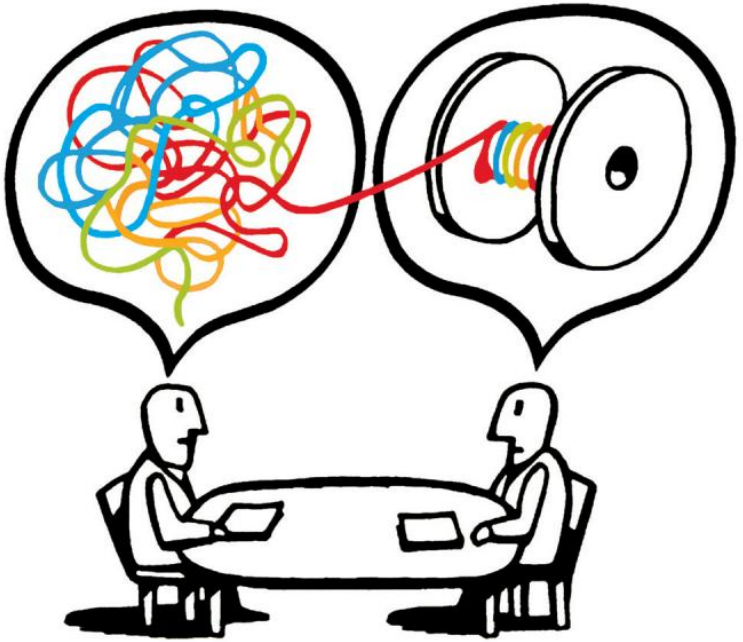
Qu'est-ce qu'on mesure exactement (quels chiffres on met dans les colonnes...) ?

Quelle est la variable dépendante ?

Quelles sont les variables explicatives ?

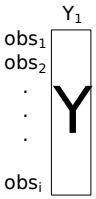
Est-ce qu'on veut juste décrire des groupes d'observations similaires (analyse non supervisée) ?

Souvent on ne sait pas "par quel bout" prendre son jeu de données, ses questions spécifiques,...



Réfléchir en terme de **variables, observations et lien entre les variables** vous aidera à démêler la pelote de laine...

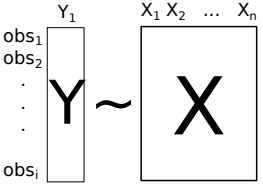
Univariées non supervisées



Une seule variable d'intérêt considérée à la fois

--> statistiques descriptives :
moyenne, médiane, écart type, coefficient de variation,...

Univariées supervisées

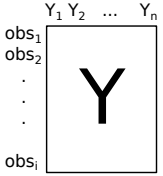


On veut prédire/expliquer une seule variable Y (à la fois) en fonction d'un ou plusieurs prédicteurs X

Régression (& classification) et extentions :
GLMs : incluent régression, ANOVA, ANCOVA, t-test, G test, logistic reg., ...
+ GLMMs Generalized Linear Mixed Models

Machine learning *I* algorithmic approaches
CART (regression trees), Rand. Forest, Boosting, SVM, Neural Net, K-NN, GAMs, MARS,...

Multivariées non supervisées

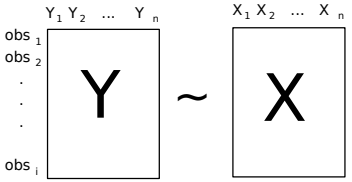


On s'intéresse en général aux similarités/dissimilarités entre colonnes ou lignes
--> **matrice de distance**

Ordinations :
représenter un maximum de variabilité en peu de dimensions
PCA - CA - MDS (=PCoA) - nMDS

Clustering :
diviser les données en groupes similaires
Clustering hiérarchique (dendrograms) , non hiérarchique (K-means), ...

Multivariées supervisées



Liens entre deux (ou plus) matrices

Canonical ordinations (regression like) :
CCA, RDA, dbRDA

Multivariare correlations :
Canonical correlation, Procrustes, Mantel,...

Question et population d'intérêt bien définies

Attention : une erreur très fréquente est de vouloir répondre à trop de questions en même temps avec des données et des moyens limités

Il faut mettre des priorités…
Quelles sont les questions les plus importantes ?

Ex : On veut tester 2 races et 2 fourrages en bio et en conventionnel sur la production de méthane de 8 vaches

Dans ce genre de situation, il faut soit augmenter la taille de l'échantillon soit se demander quelle est la question la plus importante et laisser tomber les autres…

Bien garder en tête l'objectif final peut aider à trier dans les questions.
Pex : le fourrage est sans doute le paramètre que les agriculteurs seront les plus enclins à ajuster → Le fourrage est donc la question qui doit sans doute recevoir le plus d'attention.

Question et population d'intérêt bien définies

En fonction de la question, définir la **population d'intérêt**
c'est à dire celle à laquelle on veut généraliser les résultats

Exemple :
On veut étudier l'effet du pâturage sur populations de papillons de jour des pelouses calcaires.

En Europe ? En Wallonie ?
Sur un site bien particulier dont on doit assurer la gestion ?

En fonction de la réponse l'échantillonnage sera différent.

Une erreur classique consiste à extrapoler les résultats au delà de la population échantillonnée.

Adéquation des mesures

Est-ce que les variables mesurées correspondent bien à la question posée ?

Règle presque triviale !
Vous pouvez avoir le meilleur design du monde, si le critère d'adéquation n'est pas rempli, en général, votre étude ne sert à rien...

Adéquation des mesures

Exemple 1 :

On veut estimer l'effet de l'agriculture biologique sur l'abondance des chauves-souris.
On place des détecteurs d'ultra-sons dans 30 fermes bio et 30 fermes conventionnelles.
On compte le nombre de contacts enregistrés.

--> ce qu'on mesure ici est l'intensité de chasse, pas la taille des populations !

Exemple 2 :

Un chercheur veut savoir si il existe une relation entre la taille et le poids des œufs d'une espèce d'insecte.
Il commence par peser 1500 œufs ensuite il les remet en vrac dans un pot. Il encode les valeurs et les classe par ordre croissant.
Il mesure ensuite 1482 oeufs (il en a perdu en route) et classe les valeurs encodées par ordre croissant.
Il calcule ensuite la corrélation, qui est très élevée et très "significative"...

On compare ici le poids d'un œuf avec la taille d'un autre…
Données irrécupérables !

Adéquation des mesures

Certains cas sont plus subtils...

Exemple 3 :
On veut savoir si la présence d'une espèce de sauterelle est influencée par la hauteur de la végétation.
On parcourt 100 sites où on cherche à vue l'espèce et on note la hauteur de la végétation

Le problème ici est que ce qu'on mesure est à la fois la probabilité de présence mais aussi de détection de l'espèce qui dépend vraisemblablement de la hauteur de la végétation.
Si la détection de l'espèce était indépendante de la variable explicative, çà ne poserait pas de problème (même si la détection varie d'un site à l'autre).
Il s'agit plus ici d'un problème de confusion de facteurs qu'un problème d'adéquation.

Solutions :
Repérer l'espèce au chant
Utiliser des méthodes permettant d'estimer la probabilité de détection (en passant plusieurs fois sur le même site)

Adéquation des mesures

Si il y a des erreurs de mesure, quel est leur impact réel ?

NB : Vocabulaire !!

Langage courant : si on a fait une erreur, on s'est trompé, la mesure/le résultat n'est pas bon, on ne l'utilise pas

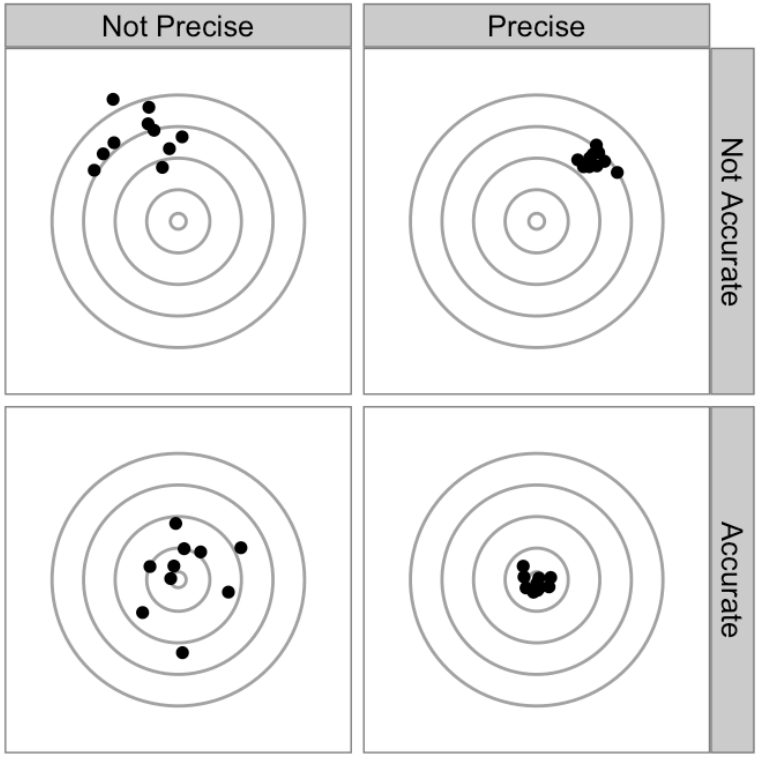
Science (en général) : l'erreur de mesure désigne l'imprécision de la mesure (erreur aléatoire) ou des erreurs systématiques mais n'implique pas automatiquement que la mesure/le résultat n'est pas utilisable !!!

Adéquation des mesures

Si il y a des erreurs de mesure, quel est leur impact réel ?

Biais systématique
L'erreur est toujours dans
le même sens

Erreur aléatoire - "imprécision"
L'erreur est une fois dans
un sens, une fois dans l'autre

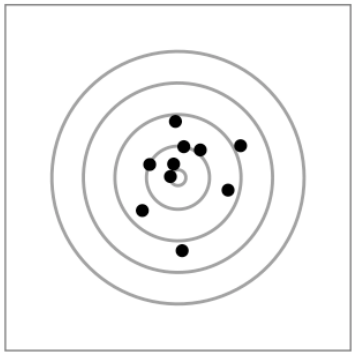


Dans la mesure du possible : éviter les biais et limiter les imprécisions.
Mais quand ils ne peuvent être évités ce n'est pas toujours la fin du monde !

Adéquation des mesures

Si il y a des erreurs de mesure, quel est leur impact réel ?

On a besoin des résultats pour interpréter réellement l'effet de ces erreurs



Erreur aléatoire - "imprécision"
L'erreur est une fois dans
un sens, une fois dans l'autre

L'**erreur aléatoire**, non systématique ajoute du bruit qui peut potentiellement masquer le signal
(ie ce qu'on veut étudier).
Elle diminue la puissance statistique.
L'erreur aléatoire ne peut normalement pas amener à de "faux" résultats positifs (biaisés)

→ **si on obtient un résultat significatif :**
le signal était suffisamment important
par rapport au bruit

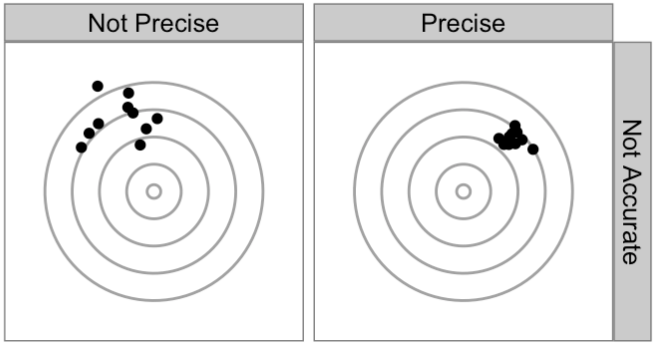
→ **si on obtient pas de résultat significatif :**
il est possible qu'avec une moins grande imprécision dans les mesures on aurait pu détecter un signal…
Difficile de conclure...

idem : si il y a une variable explicative importante qui n'est pas incluse dans un modèle/régression (typiquement parce qu'on ne l'a pas mesurée), cela augmente la variabilité résiduelle. → si on trouve des résultats significatifs pour d'autres variables explicatives, cela ne fait que renforcer ces résultats...

Adéquation des mesures

Si il y a des erreurs de mesure, quel est leur impact réel ?

On a besoin des résultats pour interpréter réellement l'effet de ces erreurs



Biais systématique
L'erreur est toujours dans le même sens

Les biais systématiques peuvent amener à de "faux" résultats positifs mais dans certaines conditions les résultats peuvent néanmoins être informatifs...

Example : présence/absence d'une sauterelle en fonction de la hauteur de la végétation.
On sait que la végétation haute diminue la détectabilité

→ **Conséquence** : on sous-estime la fréquence de l'espèce quand la végétation est haute

Cas 1 : résultat = lien positif entre la présence de la sauterelle et la hauteur de la végétation.
→ malgré la sous-estimation on trouve quand même plus de sauterelles dans la végétation haute.
Le biais ne fait que renforcer le résultat !!

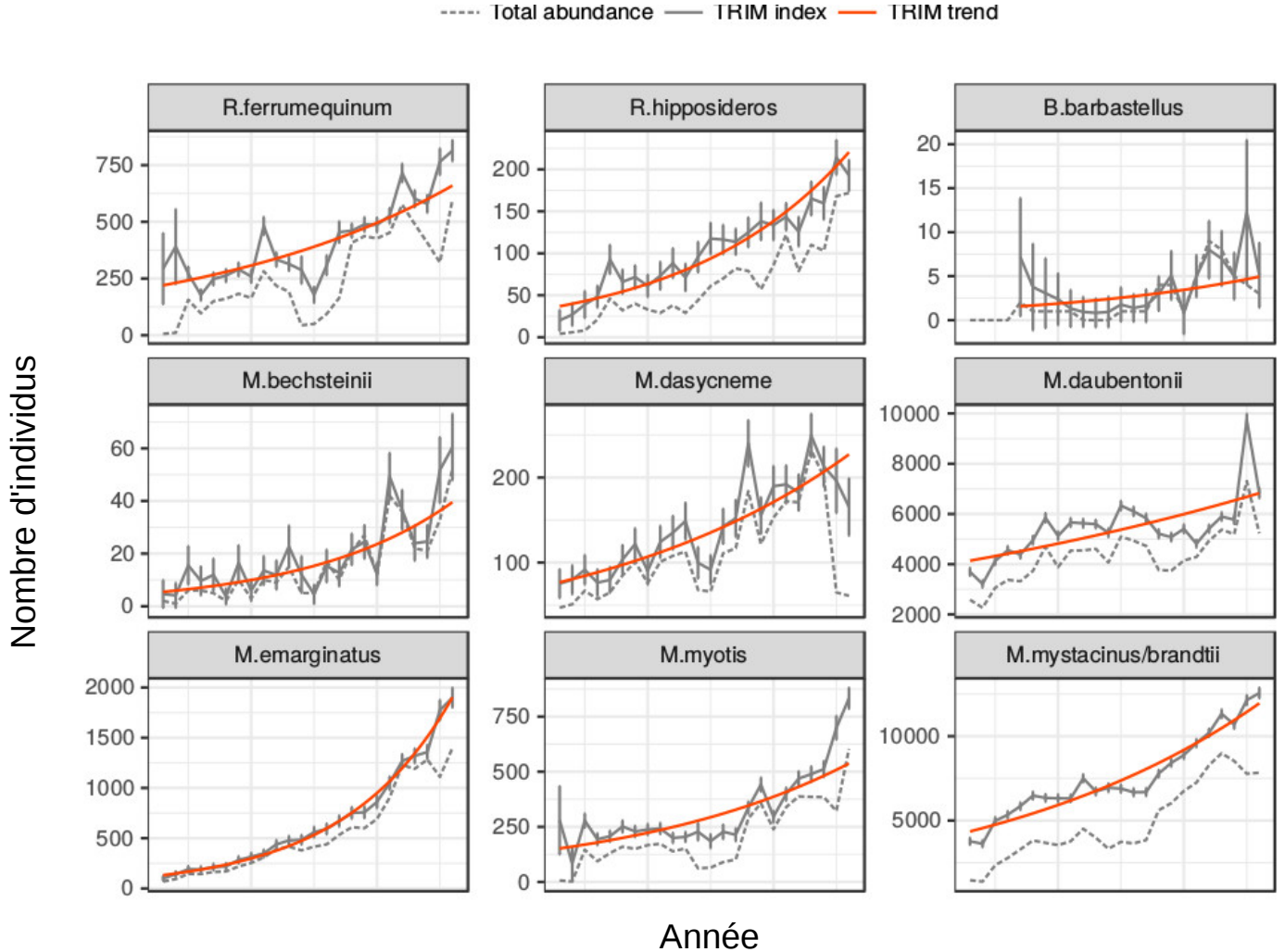
Cas 2 : résultat = pas de lien ou lien négatif avec la hauteur de la végétation
Ces résultats peuvent être dus à la sous-estimation. **Difficile de conclure !**

Adéquation des mesures

Si il y a des erreurs de mesure, quel est leur impact réel ?

On a besoin des résultats pour interpréter réellement l'effet de ces erreurs

Exemple réel : Tendances des populations de chauves-souris (comptages hivernaux en cavités souterraines) au cours du temps. Tendances positives !

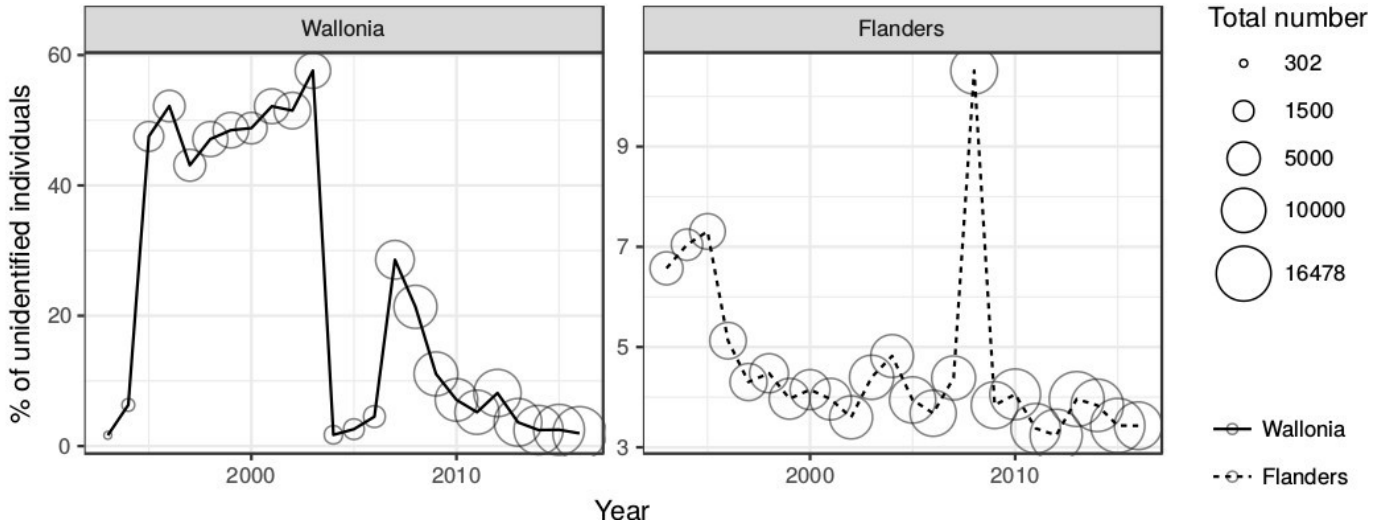


Adéquation des mesures

Si il y a des erreurs de mesure, quel est leur impact réel ?

On a besoin des résultats pour interpréter réellement l'effet de ces erreurs

Problème : le % d'individu non identifié à l'espèce a fortement diminué au cours du temps en particulier en Wallonie
(meilleur matériel, plus de naturalistes, effort de formation)…
Conséquence : on peut s'attendre à voir augmenter le nombre d'individus de certaines espèces simplement parce qu'il y a plus d'individus identifiés jusqu'à l'espèce

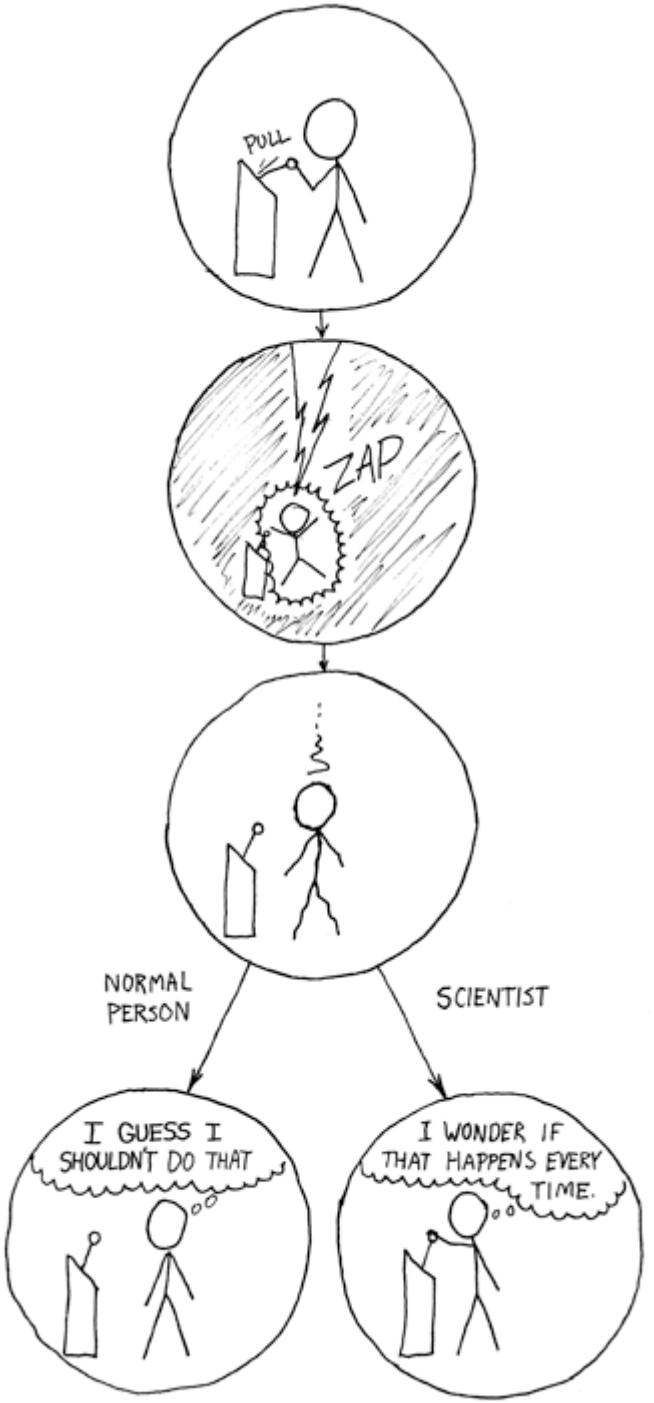


On a que ces données là. On dois faire avec les biais existants...
→ élimination des sites où le % d'identification était particulièrement faible (choix pas trivial !!)
Rhinolophes : toujours identifiés à l'espèce : augmentent aussi
Flandre : % identification + stable : les chauves-souris augmentent aussi
Le nombre total de Chauves-souris (toutes espèces confondues) augmente aussi
→ **il y a vraisemblablement une réelle augmentation des populations de certaines chauves-souris**

Réplication

Nécessaire pour estimer la variabilité des résultats et voir à quel point les résultats sont extrapolables à d'autres cas ou si ils sont dû uniquement au hasard de l'échantillonnage (inférence statistique).

Particulièrement important en sciences de la vie où la variabilité est typiquement très importante...



https://xkcd.com/242/

Réplication

Exemple d'étude sans réplication

On a fait trois aménagements piscicoles différents sur 3 rivières (une échelle à poisson sur la première rivière, un aménagement naturel des berges sur la seconde, et un aménagement de frayères sur la 3ème).
On fait une pêche électrique avant et après l'aménagement.
On veut savoir quel type d'aménagement est le plus favorable aux poissons.

On peut décrire ce qui s'est passé dans ces 3 cas particuliers mais on ne pourra en tirer aucune conclusion généralisable et on ne pourra établir aucune inférence valide. Les différences observées pourraient être dues à d'autres facteurs que les aménagements et même simplement au hasard de l'échantillonnage.

Réplication

Mais les traitements ne peuvent pas toujours être répliqués :-(

Conditions pour que la non réplication soit sans doute "acceptable" :

1) Études à très large échelle apportant un "plus" qualitatif car elles sont plus représentative de ce qui se passe en conditions réelles que en condition expérimentale à plus petite échelle
Ex : étude de la production de méthane dans des grands silos vs dans des petits bidons

2) ou : Réplication éthiquement pas possible
Pex : Quel est l'effet d'une pollution aux métaux lourds sur la faune aquatique d'un lac ?

3) Elles sont présentées comme des études descriptives ("études de cas") pas spécialement généralisables à d'autres cas.
On essaye pas à tout prix de faire des test statistiques inappropriés…

4) Les données sont disponibles publiquement de façon à ce les résultats puissent être utilisés plus tard dans une méta-analyse en conjonction avec des études similaires.

Indépendance des échantillons - Pseudoréplication

Toutes les inférences sont basées sur la quantité d'information indépendante disponible
= "degrés de liberté"

Si les échantillons ne sont pas réellement indépendants
on "ment" sur la quantité d'information, et les p-valeurs, erreurs standards, intervalles de confiance sont faux.

On parle souvent de "**Pseudoréplication**"

Indépendance des échantillons - Pseudoréplication

Très souvent c'est le design d'échantillonnage lui même qui implique la pseudoréplication

Exemple extrême :

Question *:*
On veut savoir si les hommes ont en moyenne des cheveux plus longs ou plus courts que les femmes.

Design expérimental :
On sélectionne un homme et une femme et on mesure 1000 cheveux sur chaque tête. On compare les moyennes avec un test de student.

Où est le problème ?

Indépendance des échantillons - Pseudoréplication

Très souvent c'est le design d'échantillonnage lui même qui implique la pseudoréplication

Question :

On veut savoir si les hommes ont en moyenne des cheveux plus longs ou plus courts que les femmes.

Design expérimental :

On sélectionne un homme et une femme et on mesure 1000 cheveux sur chaque tête. On compare les moyennes avec un test de student.

Où est le problème ?

L'échantillonnage est répliqué mais pas le "traitement" (homme/femme).
L'échantillonnage est répliqué mais pas au bon niveau...
La population échantillonnée ici sont les cheveux de ces deux personnes en particulier et pas des hommes/femmes en général.

Un test statistique "naïf" répondra donc à la question "Est-ce qu'il y a une différence de longueur de cheveux plus grande que ce qu'on pourrait attendre par hasard entre cet homme et cette femme en particulier et pas entre les hommes et les femmes en général "

Indépendance des échantillons - Pseudoréplication

Très souvent c'est le design d'échantillonnage lui même qui implique la pseudoréplication

Il n'est pas rare de voir des études comme celle-ci :

On veut caractériser l'effet de l'arboriculture biologique sur l'abondance totale des carabes.

On sélectionne un verger "bio" et un verger "conventionnel" et on place 30 pièges "pitfall" dans chacun. On compare ensuite l'abondance des carabes par piège.

Ce qu'on mesure ce sont les différences entre ces deux sites particuliers qui diffèrent par le mode d'agriculture mais aussi sans doute par de nombreuses autres caractéristiques

Indépendance des échantillons - Pseudoréplication

Que vaut-il mieux (pour un total de 60 pièges) ? :

2 vergers avec 30 pièges par verger ?

12 vergers avec 5 pièges par verger ?

60 vergers avec 1 piège par verger ?

Indépendance des échantillons - Pseudoréplication

Que vaut-il mieux (pour un total de 60 pièges) ? :

2 vergers avec 30 pièges par verger ?

12 vergers avec 5 pièges par verger ?

60 vergers avec 1 piège par verger ?

Dans la grande majorité des cas, il faut maximiser les réplicats indépendants (ici les vergers) quitte à n'avoir aucun pseudoréplicat (ici les pièges).

C'est particulièrement vrai quand la variation entre pièges d'un même verger est faible. Dans ce cas on ne fait que mesurer plusieurs fois la même chose. Si la variation au sein d'un verger est grande il peut être intéressant d'avoir plusieurs pièges par verger pour mesurer cette variation.

Indépendance des échantillons - Pseudoréplication

Problème :

Même si au total on a le même nombre de pièges,
échantillonner 60 vergers avec 1 piège sera sans doute plus coûteux
que 12 vergers avec 5 pièges.

Ajouter des pseudoréplicas est souvent peu coûteux par rapport aux vrais réplicats et apporte malgré tout une information potentiellement utile.

--> ces designs "hiérarchisés" (nested designs) sont très fréquents

Indépendance des échantillons - Pseudoréplication

Ces designs hiérarchisés/répétés sont fréquents et ne sont pas obligatoirement mauvais en eux-mêmes
mais on rencontre 2 problèmes classiques :

1) On a trop peu de vrais réplicats
(pex 4 vergers avec 15 pièges par verger)

2) On compare les 30 pièges "bio" aux 30 pièges conventionnels comme si ils étaient indépendants
(on ignore la pseudoréplication --> les inférences ne sont pas valides)

2 solutions principales pour ce dernier point:

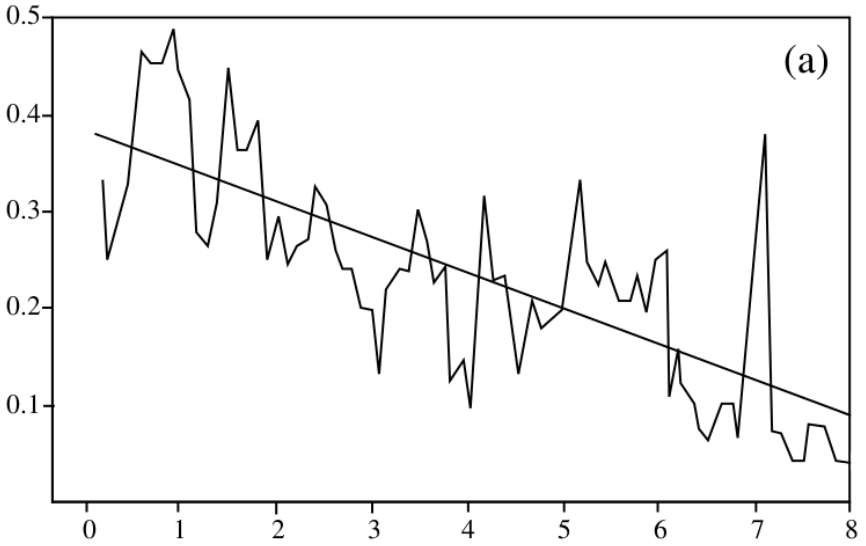
a) prendre la moyenne (ou la somme) par verger (mais on perd de l'info)
b) inclure la variable "verger" comme facteur aléatoire dans l'analyse
--> modèles mixtes
la corrélation entre mesures d'un même verger est incluse dans le modèle

Indépendance des échantillons - Pseudoréplication

Autres cas fréquents de non indépendance

Corrélation temporelle

exemple : on veut comparer l'évolution des populations de chevreuils au cours du temps entre plusieurs modes de gestion cynégétique.
On nous finance grassement et on peut donc mesurer précisément les populations de chevreuils toutes les secondes pendant 10 ans...
On obtient donc plus de 315 milliards de points par population mais ces points sont-ils réellement indépendants ?



Indépendance des échantillons - Pseudoréplication

Autres cas fréquents de non indépendance

Corrélation spatiale

Pour différentes raisons qui seront pas détaillées ici, deux points d'échantillonnage très proches dans l'espace sont *a priori* susceptibles d'avoir des valeurs similaires.

Idéalement il faut éviter les points d'échantillonnage trop proches

Les modèles mixtes et d'autres méthodes statistiques peuvent parfois permettre de prendre compte ces corrélations spatiales et temporelles également.

Échantillonnage aléatoire

Un échantillonnage aléatoire dans l'ensemble de la population est nécessaire pour que **l'échantillon** soit **représentatif**.

Aléatoire signifie que chaque élément de la population a la **même probabilité de se faire échantillonner** que les autres.

On ne prend souvent pas assez de précautions pour rendre l'échantillonnage réellement aléatoire.
L'observateur peut très souvent introduire des biais.

Échantillonnage aléatoire

Exemple 1 :

Des agronomes ont mis au point un modèle prédisant la production de sucre dans les champs de betterave en allant les prélever eux-mêmes dans des champs d'essai. Le modèle fonctionne très bien.

Pour garantir un échantillonnage aléatoire l'agriculteur doit se placer où il veut en bordure du champ, faire entre 20 et 50 pas vers l'intérieur du champ, jeter par dessus l'épaule un bâton et prélever la betterave qui se trouve la plus proche de la pointe. Il doit ensuite recommencer 15 fois et envoyer les betteraves pour analyse.

Lors de la mise en pratique chez les agriculteurs on se rend compte que le modèle sur-estime systématiquement la production

Que se passe-t-il ?

Les agriculteurs introduisent un biais systématique en éliminant consciemment ou non les betteraves les plus chétives et en évitant des parties du champ moins belles.

Échantillonnage aléatoire

Bien souvent on prétend avoir récolté des données aléatoirement ou "au hasard" alors qu'il n'en est rien en réalité

Exemple 2 :

On dispose "aléatoirement" des cadrats sur un site pour faire des relevés botaniques. En fait on se ballade sur le site et on décide à un moment "tiens, je vais mettre mon quadrat ici" ou au mieux, on le jette par dessus l'épaule. Mais en pratique on va souvent (comme dans le cas des betteraves) éviter inconsciemment certaines zones qui nous semblent "pas représentatives"

Exemple 3 :

On sélectionne "aléatoirement" 15 sites avec des pins noirs de plus 5 m de haut pour y réaliser une étude sur les communautés de coccinelles
En pratique, ces sites ne sont pas très faciles à trouver, on les cherche et on prend les 15 premiers qu'on trouve... Souvent on choisit aussi des sites accessibles ou pas trop éloignés de son domicile/lieu de travail, etc...C'est un problème difficile à éviter pour des raisons pratiques

A chaque fois que c'est possible (en particulier en labo, plus difficile sur le terrain...), on devrait utiliser des générateurs de nombre aléatoires pour sélectionner réellement aléatoirement ses sites, la position des quadrats, etc. ou l'ordre dans lequel on traite les échantillons au labo

Échantillonnage aléatoire

Dans certains cas l'expérimentateur choisi délibérément de ne pas faire un échantillonnage aléatoire en particulier quand il veut échantillonner des événements rares.
C'est parfois une bonne stratégie mais il faut être conscient des conséquences...

Exemple 4 :

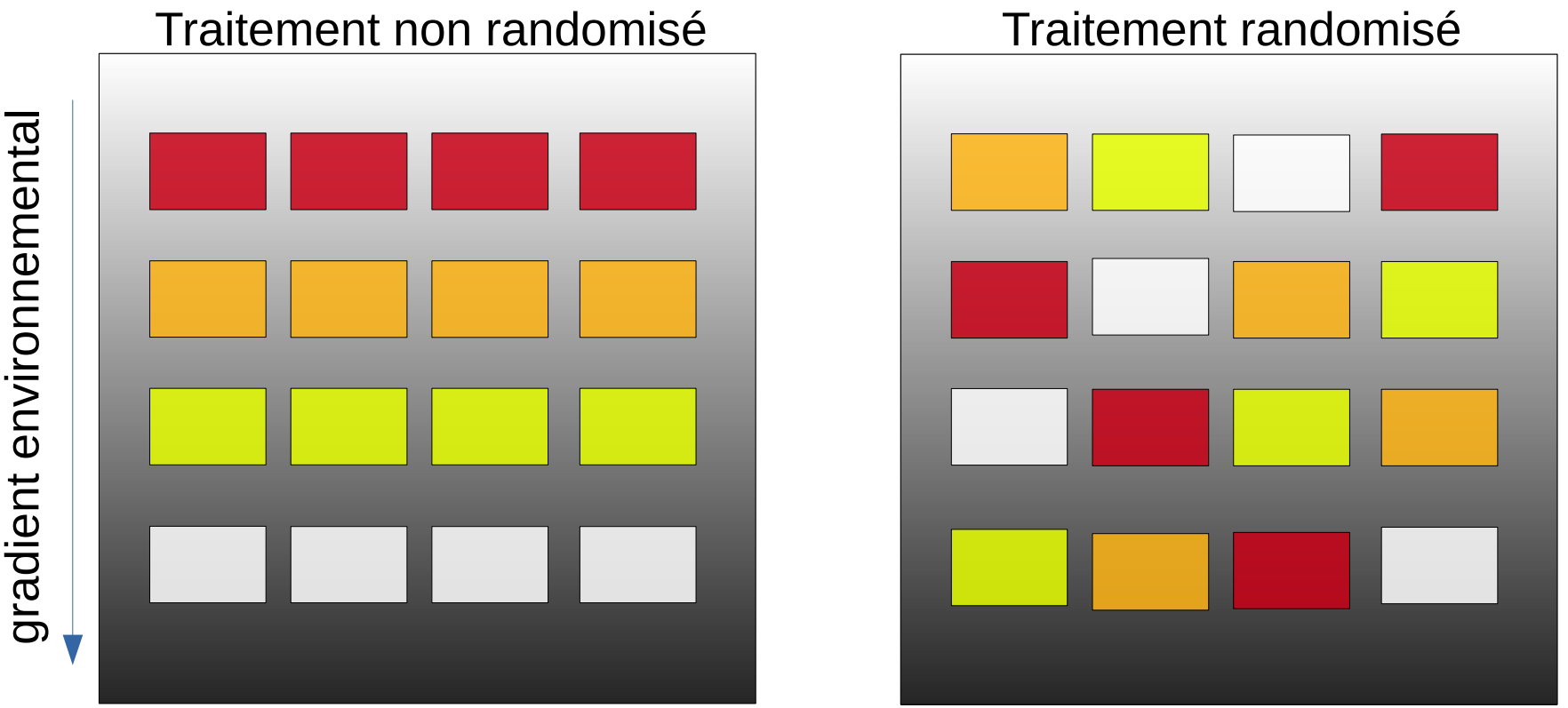
Une chercheuse veut caractériser les sites de ponte d'un papillon. Elle veut ensuite produire une carte prédictive de la qualité des sites de pontes afin de savoir quels sont les sites les plus propices et ceux qui peuvent être améliorés par exemple par des mesures de gestion.
Elle sélectionne 50 plantes hôtes avec des œufs et 50 plantes hôtes sans œufs. Elle caractérise ensuite ces plantes (taille, nombre de feuilles, exposition,...) et met en relation ces valeurs avec la présence/absence des oeufs.

Avec cette approche on pourra mettre en évidence les variables les plus importantes pour caractériser les sites de pontes mais on ne pourra pas faire de carte prédictive (ou alors uniquement de qualité relative) parce que les plantes avec des œufs sont vraisemblablement surreprésentées dans notre échantillon par rapport à un échantillon aléatoire.
On peut estimer un effet relatif mais pas absolu...

Randomisation des mesures et traitements

La randomisation permet de limiter la confusion de facteurs càd les cas ou les effets observés peuvent être dus à plusieurs facteurs en même temps sans qu'on puisse les distinguer

Exemple 1:
Confusion entre le traitement et un gradient environnemental inconnu



Cas particulier : "carré latin"

Randomisation des mesures et traitements

La randomisation des mesures

Exemple 2:

Riri, Fifi et Loulou font une étude sur l'effet de pesticides sur des mouches cécidogènes du froment. Ils ont pulvérisé des parcelles avec 8 produits différents (et un témoin). Ils fauchent les parcelles et prélèvent aléatoirement 90 plants sur lesquels il faut chercher des galles pas toujours faciles à trouver...

Cas1 :

Par souci d'efficacité Riri s'occupe des 3 premiers traitements, Fifi des 3 suivants et Loulou des 3 derniers. Mais en fait Riri est beaucoup plus expérimenté et trouve beaucoup plus de galles, et Loulou en a marre de tout le temps chercher ces galles et en trouve beaucoup moins... Il y a confusion entre l'effet du traitement et l'effet observateur.

Cas 2 :

Pour éviter la confusion traitement-observateur, ils prennent d'abord le premier traitement et chacun s'occupe de 30 plans, ensuite ils passent au traitement suivant. Le problème est que au cours du temps leur capacité à trouver des galles s'améliore. il y a donc une confusion entre l'effet du traitement et l'ordre dans lesquels ils sont mesurés.

Conclusion : utiliser un générateur de nombres aléatoires pour distribuer les lots de 30 plantes...

Contrôles judicieusement choisis

Les **"contrôles" ou "témoins"** sont des traitements qui n'ont généralement pas d'intérêt en eux-mêmes mais qui **permettent d'éliminer certains facteurs de confusion**, certaines explications autres que le traitement d'intérêt qui permettraient d'expliquer le résultat.

Un ou plusieurs contrôles judicieusement choisis peuvent faire toute la différence au niveau de l'interprétation.

Certaines études observatives peuvent mettre en évidence des lien de cause à effet si les contrôles et l'échantillonnage sont bien choisis ("quasi-experimental designs")

Les témoins doivent en général être les plus proches possibles des traitements à l'exception de l'hypothèse que l'on veut éliminer.
Les témoins sont rarement des échantillons où on a "rien fait".

Contrôles judicieusement choisis

Exemple ecotox :

On teste la toxicité d'un produit sur des insectes et on observe 100 % de mortalité. On peut être tenté de dire que le produit est toxique. Mais qu'est-ce qui nous prouve que notre pulvérisateur n'était pas contaminé par un autre produit ou que notre souche d'insectes était malade ?

-> on ajoute un contrôle négatif : on pulvérise quelques réplicats avec de l'eau

On teste un autre produit et on observe 0 % de mortalité.
Est-ce que le produit n'est pas toxique ? Peut-être.
Ou alors notre pulvérisateur était défectueux, ou il y avait trop de vent et le produit n'est pas arrivé sur les insectes ou bien on a une souche d'insectes particulièrement résistants.

--> on ajoute un contrôle positif : on pulvérise un produit dont la toxicité est bien connue.

--> plusieurs contrôles différents peuvent être utiles pour éliminer des facteurs de confusion différents

L'expérimentateur a souvent tendance à être trop bienveillant avec ses propres études !
--> se mettre dans la peau de qqn de mal veillant pour choisir ses contrôles !

Contrôles judicieusement choisis

Exemple "BACI" = "Before-After Control-Impact" designs
Quasi-experimental design

On veut évaluer l'effet de la fauche dans des prairies humides sur les population d'une espèce de papillons.

On choisi 30 sites que l'on fauche chaque année pendant 10 ans et on compte chaque année les papillons.
On constate une augmentation des effectifs. Peut-on en conclure que ce mode de gestion est favorable ? Peut-être, à moins que les populations n'étaient déjà en augmentation avant le début du traitement.
--> idéalement on devrait récolter aussi 10 années de données avant de commencer à faucher. C'est rarement possible...

On recommence avec une autre espèce de papillon et des données pendant 10 ans avant et après le début du fauchage. On ne constate aucune évolution du nombre d'individus (la population stagne). Est-ce qu'on peut conclure que ce mode de gestion n'a aucun effet ? Peut-être, sauf si dans les sites où il n'y a aucun fauchage les populations sont en augmentation. Dans ce cas l'effet est probablement négatif.
Il faut donc idéalement ajouter des parcelles contrôles (sur les mêmes sites ? dans des sites différentes les plus proches et similaires possibles?). Surtout quand on a pas de données avant...

(4) **Puissance statistique**

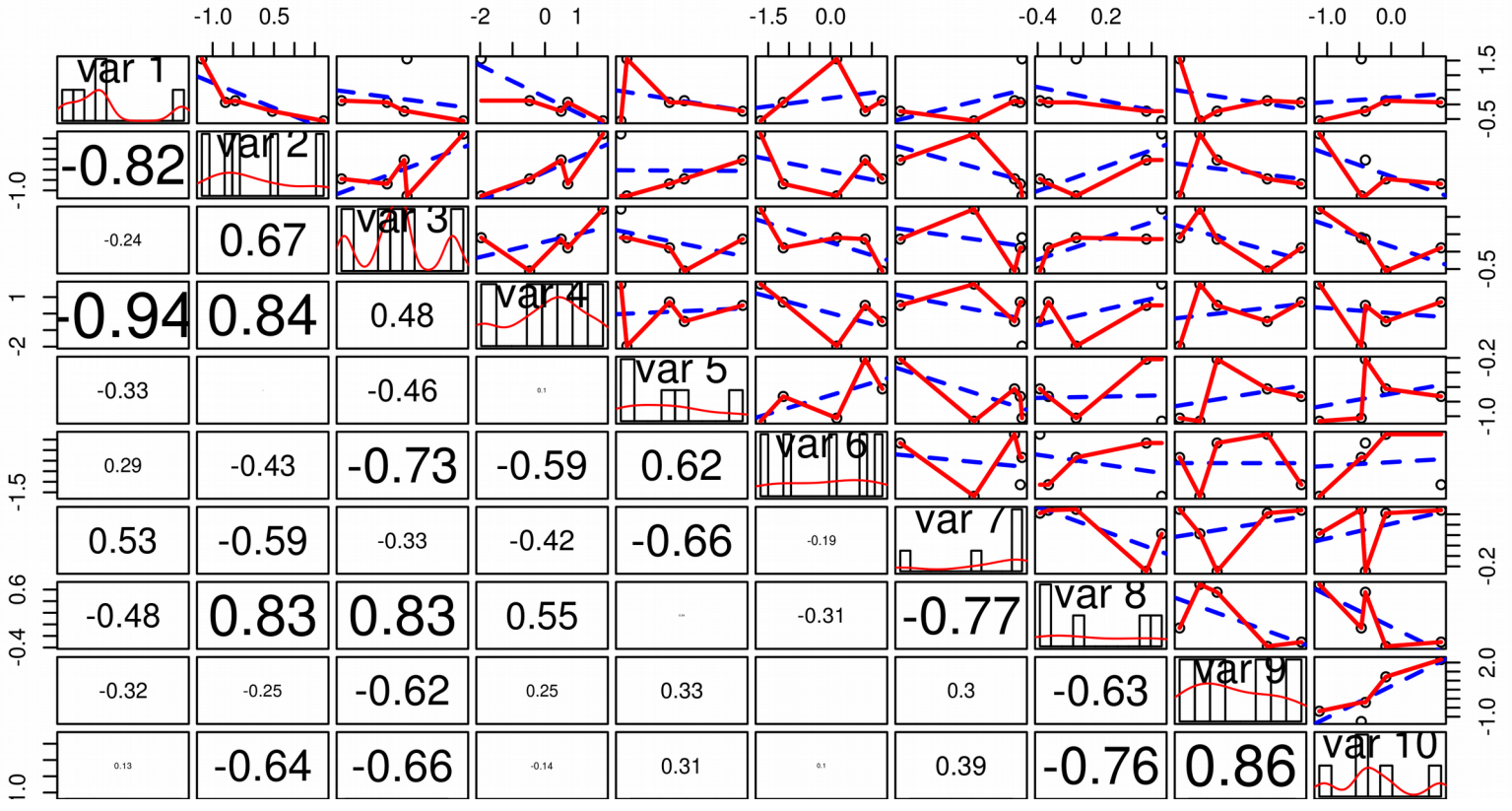
Rappels sur l'inférence statistique : tests d'hypothèse (p-valeurs), intervalles de confiance

Inférence : tests d'hypothèse nulle

Tests d'Hypothèse nulle : le problème

On génère 10 variables complètement aléatoires (donc indépendantes)
Chaque variable contient 5 éléments
On calcule la corrélation entre chaque paire de variable

Certaines corrélations peuvent être extrêmement élevées
mais c'est uniquement dû au hasard



Inférence : tests d'hypothèse nulle

Tests d'Hypothèse nulle : le raisonnement

Si je mesure deux variables réelles et que j'observe une corrélation de 0.8, comment savoir si il y a réellement une relation entre ces variables ou si ce résultat est dû uniquement au hasard ?

C'est pour répondre à cette question qu'on estime une **"p valeur**" qui est la probabilité d'obtenir un tel résultat (ou une corrélation encore plus forte) dans l'hypothèse (la fameuse **hypothèse nulle**) où il n'y aurait en fait aucune relation entre les deux variables

Inférence : précision des paramètres

On utilise les **paramètres calculés sur les échantillons**
pour estimer les **paramètres de la population**

MAIS :

Les valeurs estimées dans les échantillons sont variables
(à cause du hasard de l'échantillonnage)
On a besoin d'estimer leur précision pour savoir
à quel point on peut avoir confiance en ces estimations
et à quel point on peut les extrapoler à l'ensemble de la population.

-> c'est le rôle de :
l'erreur standard
l'intervalle de confiance

Pex : si on calcule une moyenne de 10 avec un intervalle de confiance à 95 % de [7,13], ce la signifie que si on recommençais l'échantillonnage 1000 fois on estime que la moyenne de cahque échantillon tomberait entre 7 et 13 dans 950 échantillons.

Tests d'hypothèse nulle : attention à l'interprétation !

Exemple 1 : On teste l'effet d'un nouveau traitement assez cher sur la production de lait chez les vaches

On compare avec un test de Student la production moyenne entre les vaches d'un groupe témoin et d'un groupe traité

```
> t.test(control, treatment, var.equal = TRUE)

    Welch Two Sample t-test

data:  control and treatment
t = -2.9558, df = 19998, p-value = 0.003123
```

Est-ce que vous investiriez dans ce traitement ?

Tests d'hypothèse nulle : attention à l'interprétation !

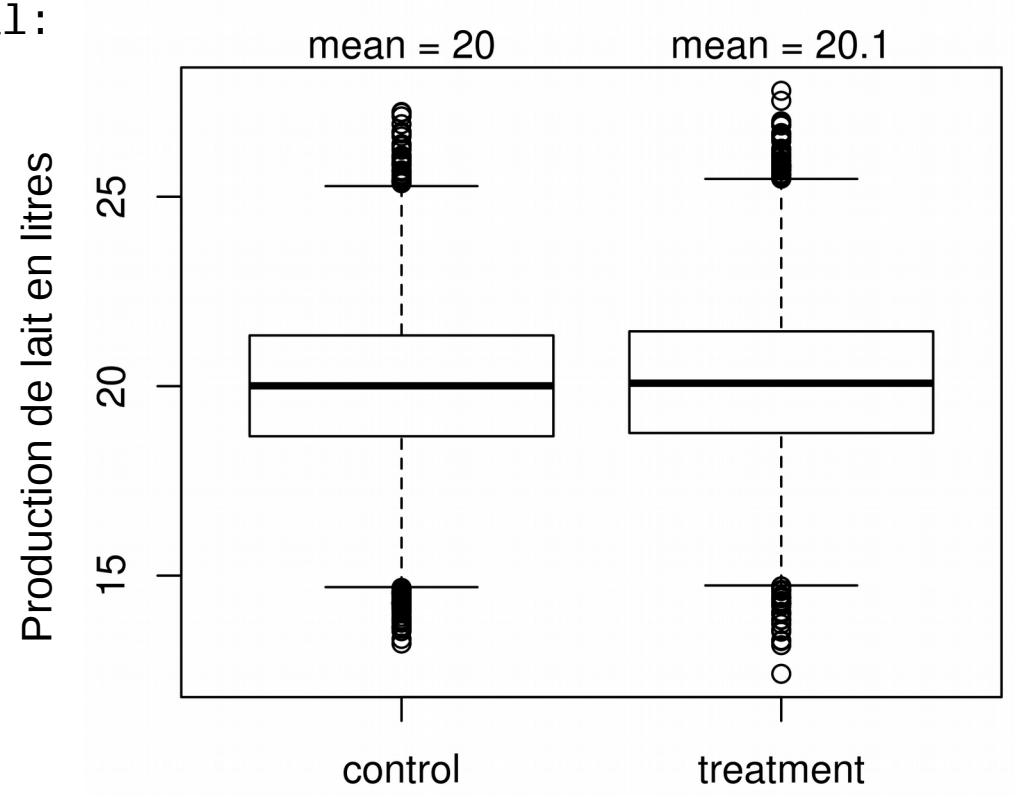
Significativité statistique ≠ significativité biologique

Erreur fréquente : p-valeur petite ne veut pas dire que la différence ou la corrélation est grande !!

```
> t.test(control, treatment, var.equal = TRUE)

data:  control and treatment
t = -2.9558, df = 19998, p-value = 0.003123
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.13808170 -0.02796792
sample estimates:
mean of x mean of y
 20.01223  20.09526
```

La différence entre groupe traité et témoin est à peine de ~ 0.08 litres de lait



Tests d'hypothèse nulle : attention à l'interprétation !

Exemple 2 : autre expérience avec un autre traitement

```
> t.test(control, treatment, var.equal = TRUE)

    Welch Two Sample t-test

data:  control and treatment
t = -1.3331, df = 28, p-value = 0.1933
```

Est-ce que vous investiriez dans ce traitement ?

Tests d'hypothèse nulle : attention à l'interprétation !

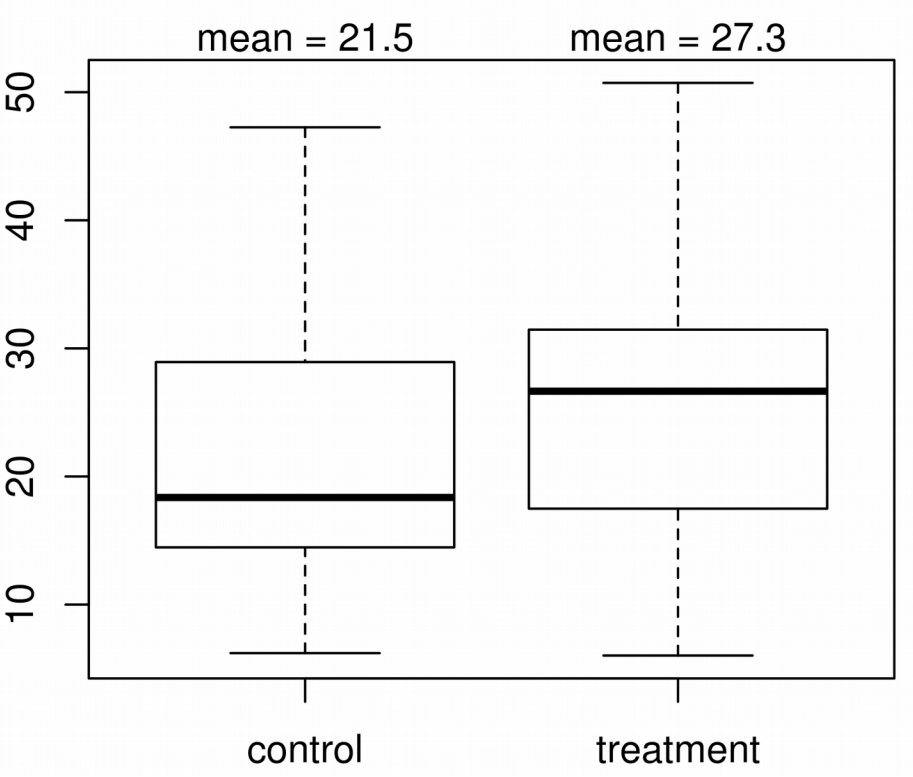
Non significatif ne veut pas dire qu'il n'y a pas d'effet !

```
> n <- 15
> set.seed(12345)
> control <- abs(rnorm(n, 20, 15))
> set.seed(123)
> treatment <- abs(rnorm(n, 25,15))

> t.test(control, treatment, var.equal = TRUE)

data:  control and treatment
t = -1.3331, df = 28, p-value = 0.1933
alternative hypothesis: true difference i
95 percent confidence interval:
 -14.723968   3.114799
sample estimates:
mean of x mean of y
 21.48118  27.28577
```

Il y a peut-être un effet mais les données sont insuffisantes pour conclure...
D'où viens cette variabilité ?
Qu'est-ce que çà donne avec un échantillonnage plus grand ?



Tests d'hypothèse nulle : attention à l'interprétation !

la p-valeur dépend de :

- la taille de l'échantillon
- la variabilité de la population
- la taille de l'effet

"*Si on augmente suffisamment la taille de l'échantillon, les p-valeurs finiront toujours par devenir significatives, à moins que l'on teste des hypothèses stupides, ce qui est rarement le cas*"
(d'après Burnham & Anderson 2000)

p valeur : outil d'aide à la décision à utiliser à bon escient !
Certains préfèrent les intervalles de confiances qui sont moins sujets à une mauvaise interprétation.

Intervalles de confiance vs p valeurs

Est-ce qu'un paramètre est significativement différent de 0 au seuil alpha = 5 % ?

Oui si son intervalle de confiance à 95 % ne comprend pas le 0

Le raisonnement fonctionne pour n'importe quelle valeur fixe autre que 0.

Est-ce que 2 paramètres (par exemple 2 moyennes) sont significativement différents au seuil alpha = 5 % ?

Si leurs intervalles de confiance à 95 % ne se recouvrent pas : oui.
Attention : si les intervalles se recouvrent (jusqu'à ~25% pour une moyenne), çà ne veut pas dire automatiquement que les paramètres ne sont pas significativement différents !!!

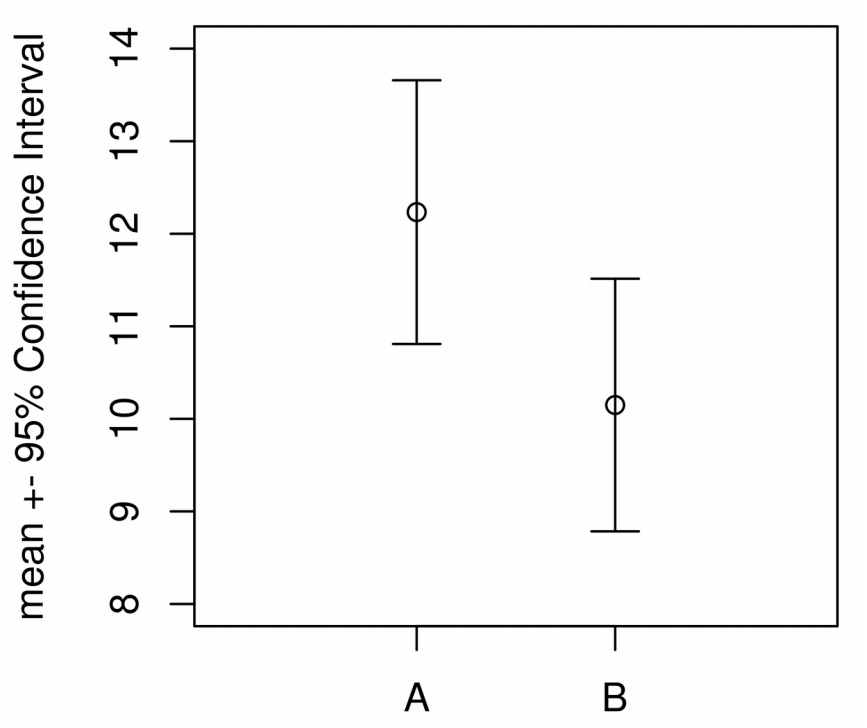
Intervalles de confiance vs p valeurs

Deux intervalles de confiances peuvent se chevaucher alors que la différence de leur paramètres est significative

```
> set.seed(1234)
> A <- rnorm(10, 13, sd=2)
> set.seed(123)
> B <- rnorm(10, 10, sd=2)
> t.test(A,B)

t = 2.3902, df = 17.967, p-value = 0.02801

> t.test(A)$conf.int
[1] 10.80900 13.65837
attr(,"conf.level")
[1] 0.95
> t.test(B)$conf.int
[1]  8.784659 11.513843
attr(,"conf.level")
[1] 0.95
```



(4) **Puissance statistique**

Puissance statistique

La puissance d'un test est la probabilité de rejeter l'hypothèse nulle quand elle est effectivement fausse.

C'est la **proportion de vrais positifs que l'on arrive à détecter**.

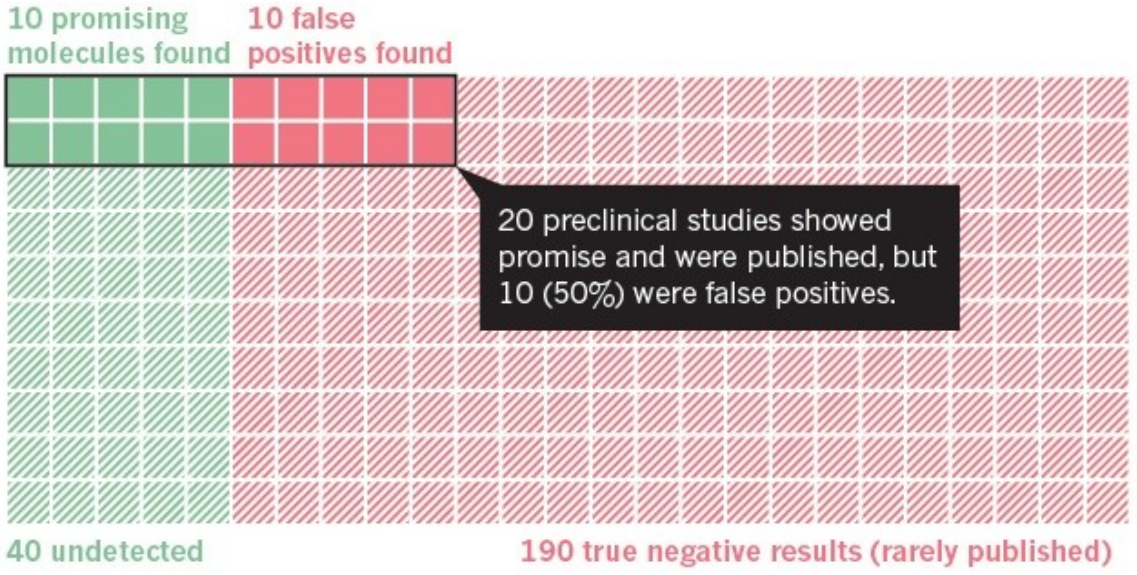
Alors que le seuil de significativité alpha (typiquement 0.05) représente la proportion de faux positifs que l'on accepte...

Autrement dit plus un test est puissant plus il arrivera à "détecter" les relations/différences réelles entre échantillons.
On veut en général maximiser la puissance des tests à coût égal et en tout cas avoir une puissance suffisante pour détecter des effets biologiquement significatifs

Puissance statistique de 0.20 : On rate 80 % des vrais positifs

alpha = 0.05 : on accepte 5 % de faux positifs

STATUS QUO: Most studies have a statistical power of only 20% and a *P* value of 0.05, meaning many more false findings (PPV of 50%). This reflects a sample size of about 10 mice per study.



Puissance statistique de 0.80 : On rate 20 % des vrais positifs

alpha = 0.01 : on accepte 1 % de faux positifs

PROPOSED STANDARDS: To achieve a PPV of 95%, study results would need a *P* value of 0.01 and a large enough sample size to reach 80% statistical power (typically >75 mice per study).

Mogil JS, Macleod MR (2017) No publication without confirmation. Nature News 542:409.

Puissance d'un test

La puissance dépend :

du type de test
du seuil de significativité alpha
de la taille de l'effet
de la variabilité
de la taille de l'échantillon

NB : La précision des estimateurs dépend
principalement de ces 2 derniers points
qui sont globalement ceux sur lesquels on peut le plus jouer

Puissance d'un test

La puissance dépend : du type de test

Pour un échantillon parfaitement identique certains tests seront parfois plus puissants.
Par exemple si on a bien une relation linéaire entre deux variables, le test de corrélation de Pearson est plus puissant que le test de corrélation de rang de Spearman.

Mais en général on choisi le test le plus adapté à son cas.
--> on joue rarement sur ce point volontairement.

Puissance d'un test

La puissance dépend : du seuil de significativité alpha

Un moyen trivial d'augmenter la puissance est simplement de considérer par exemple que tout p<0.1 est significatif (au lieu du traditionnel p<0.05).

On augmente cependant dans ce cas la probabilité de dire qu'il y a un effet alors qu'il n'y en a pas
(plus de faux positifs , erreur de type I).

Si on diminue alpha par contre, on augmente les erreurs de type II : c'est à dire qu'on a plus de chance de "rater" un effet réel (plus de faux négatifs).

Puissance d'un test

La puissance dépend : du seuil de significativité alpha

C'est rare mais il arrive que l'on joue sur ce paramètre.

Par exemple on peut considérer qu'il est moins grave de dire qu'une espèce est en déclin alors qu'elle ne l'est pas en réalité que de de manquer une espèce réellement en déclin en prétendant qu'elle ne l'est pas.

Ou qu'il est plus grave de dire qu'une substance n'est pas nocive pour la santé humaine alors qu'elle l'est en réalité que d'affirmer qu'une substance est nocive alors qu'en réalité elle ne l'est pas.

--> dans ces deux cas on choisirait plutôt un seuil à 0.1

Puissance d'un test

La puissance dépend : de la taille de l'effet

Toutes choses étant égales par ailleurs, on "détectera" plus vite une corrélation réelle de 0.8 que de 0.1.

Dans certains cas on peut jouer sur ce point en choisissant des traitements plus extrêmes

Par exemple si on veut voir si la température a un effet sur la ponte d'un insecte on choisira des 2 températures les plus extrêmes possibles (soit en contrôlant la température, soit en choisissant des sites extrêmes).

Si on veut vérifier si la relation est linéaire on rajoutera quelques échantillons à une température intermédiaire.

C'est aussi aux extrémités qu'il faut mettre le plus de réplicats car c'est là qu'on a le moins de précision.

Puissance d'un test

La puissance dépend : de la variabilité

Il y a trois moyens principaux de jouer sur ce point :

1) On peut essayer de contrôler au mieux les conditions expérimentales, choisir des sites les plus similaires possibles, ...

2) on peut mesurer une covariable qui ne nous intéresse pas en tant que telle mais que l'on sait avoir un effet sur la variable d'intérêt
On pourra alors enlever au moyen d'outils statistiques la variabilité due à cette covariable avant d'examiner l'effet de notre traitement.

Exemple :
On veut mesurer en champ l'effet de divers traitements azotés sur la production céréalière en Wallonie. Les tests sont répartis dans plusieurs champs en Wallonie. On sait que le rendement est aussi influencé par le nombre d'heures d'ensoleillement que l'on mesure donc sur chaque champ. On peut alors enlever la variabilité due à l'ensoleillement avant d'examiner l'effet des traitements azotés.

Puissance d'un test

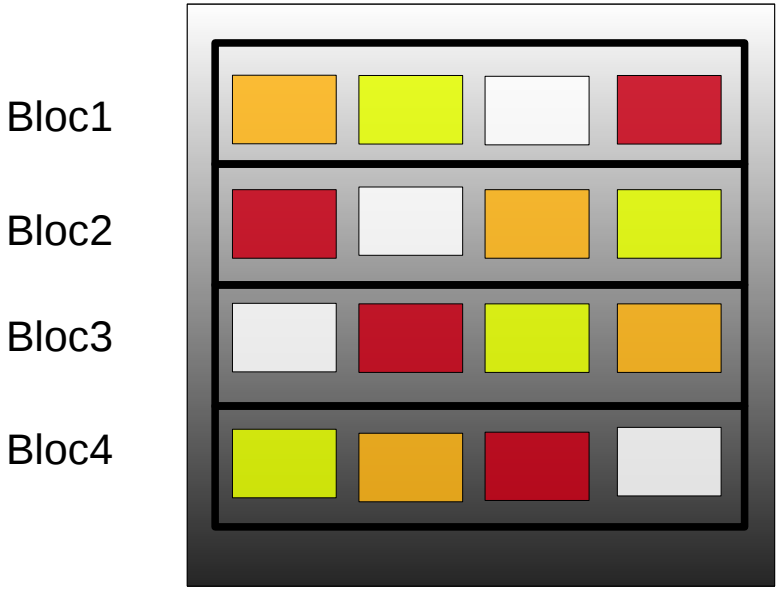
La puissance dépend : de la variabilité

3) Dans certains cas, on suspecte une hétérogénéité environnementale mais on ne sait pas exactement quels facteurs entrent en jeu ou on ne sait pas les mesurer.
Dans ce cas une technique très utile consiste à créer ce qu'on appelle des "blocs".

Puissance d'un test

La puissance dépend : de la variabilité

Un bloc est une surface ou période de temps ou toute autre unité que l'on considère comme relativement homogène. On assigne ensuite au sein de chaque bloc de manière aléatoire les différents traitements.
Les blocs vont permettre de tenir compte statistiquement la variation environnementale entre blocs sans devoir mesurer directement les covariables.



Cas particulier : "carré latin" : on pourrait ajouter des blocs verticaux qui seraient moins utiles dans ce cas

Puissance d'un test

La puissance dépend : de la taille d'échantillon

Plus on a d'échantillons indépendants plus on aura de puissance mais en général le coût augmente en proportion.
C'est souvent le facteur clé pour augmenter la puissance…

Dans certains cas cependant on peut faire des choix à coût égal.

Par exemple pour les comptages hivernaux de chauves-souris est-ce qu'il vaut mieux suivre 50 grottes chaque année ou 100 grottes une année sur deux ?

Puissance d'un test

La puissance dépend : de la taille d'échantillon

Une question très fréquente est de savoir quelle taille d'échantillon on a besoin. Dans ce cas il faut bien sûr définir la taille d'effet que l'on veut pouvoir détecter (et il faut avoir une idée de la variabilité)

Par exemple : combien de grottes faut-il suivre pour pouvoir détecter un déclin de 30 % en 10 ans ?
Ou bien : si on suit 100 grottes, combien d'années faudra-t-il avant de détecter un déclin significatif de 30 % ?

--> c'est le rôle de l'**analyse de puissance**

Puissance d'un test

L'analyse de puissance

En fixant la taille de l'effet (sur base de seuils biologiquement importants), la variabilité (issue d'expériences précédentes), le niveau alpha et le test, on peut déterminer la puissance pour différentes tailles d'échantillon.

En général on essaye d'atteindre une puissance de 0.8 (pure convention)

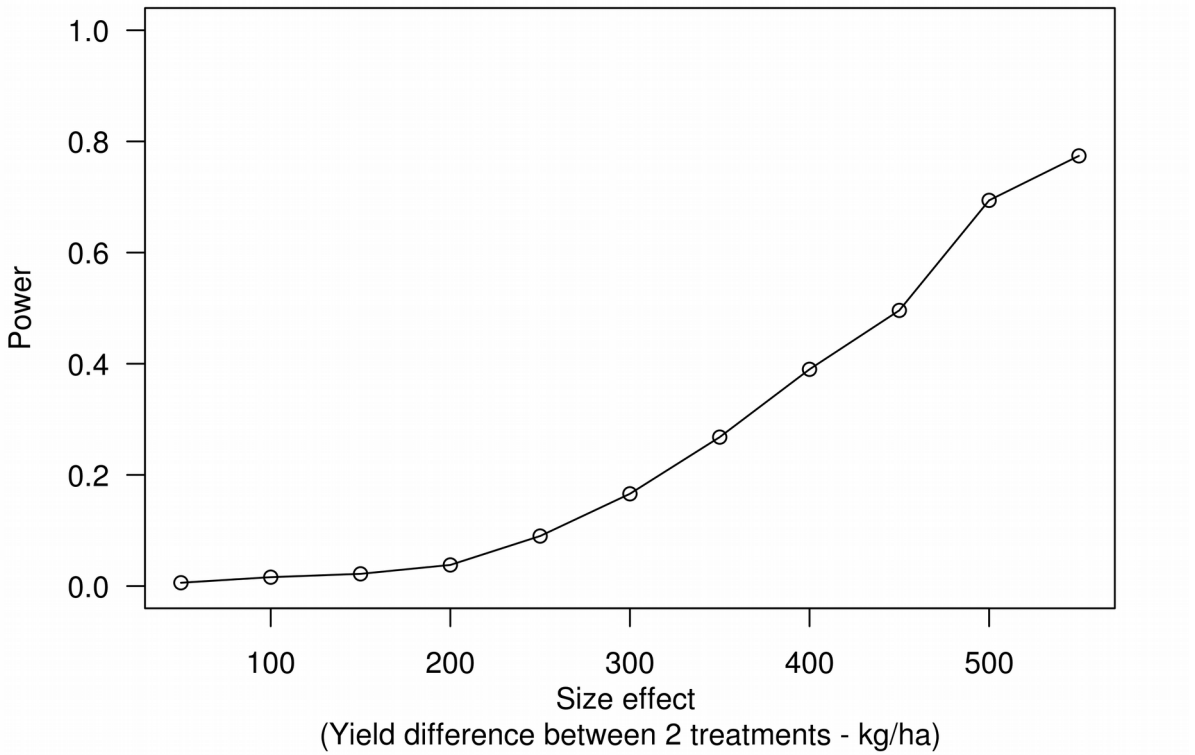
Autrefois il fallait dériver des formules mathématiques complexes pour chaque type de test. Aujourd'hui on peut utiliser des méthodes de simulation qui peuvent s'appliquer à des cas beaucoup plus complexes.

Puissance d'un test

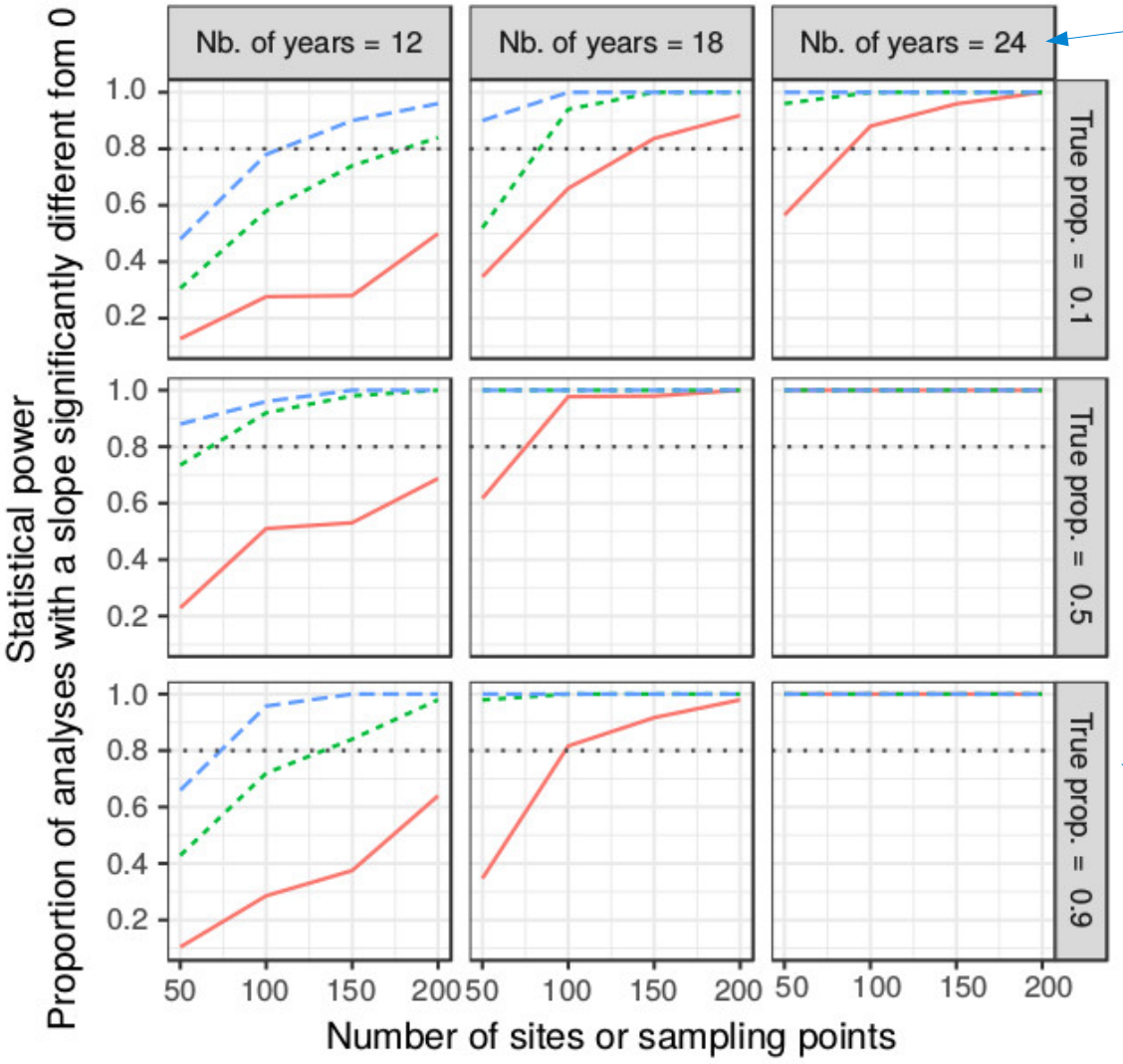
L'analyse de puissance

On peut aussi par exemple fixer la taille d'échantillon (au maximum possible) et estimer la puissance pour différentes tailles d'effet pour déterminer si une expérience vaut le coup d'être menée.

NB : certains statisticiens recommandent l'analyse de puissance pour établir un design avant l'expérience pas pour l'interpréter à postériori



Exemple d'analyse de puissance par simulation - (binomial random slope mixed models)
On fait des relevés de végétation dans des sites Natura 2000 pour évaluer leur état de conservation (bon/pas bon). On veut pouvoir détecter une pente de -0.1 (~ perte de 20 % de sites en bon état de conservation en 10 ans) .
Combien de sites a-t-on besoin de parcourir ?
Quel est l'impact si on fait 1, 3 ou 5 quadrats répétés sur chaque site
La variance est inconnue… On a fixé des valeurs assez élevées



Résultats après 12,18, 24 ans

Proportion de sites en bon état de conservation à l'année 0 (paramètre inconnu également → on teste plusieurs valeurs)