Qu'est-ce que la modélisation statistique ? Une tentative de réponse

Timothée Flutre

LabScience de l'UMR AGAP

Qu'est-ce que la modélisation statistique ? Une tentative de réponse

Timothée Flutre

Institut national de la recherche agronomique

20/03/2017

Résumé (et avertissement)

Loin de prétendre apporter une réponse définitive à cette question, le but de ce LabScience est de discuter autour de la notion de modélisation statistique, en replaçant la quantification de l'incertitude au coeur de toute démarche scientifique. C'est l'occasion d'introduire une vaste littérature d'intérêt pour toutes les personnes impliquées dans des recherches scientifiques, quelle que soit leur discipline de prédilection. Je tenterai d'éviter le langage des équations en donnant la part belle à la communication graphique, primordiale dans le développement des intuitions, tout en laissant du temps à la discussion.

Je ne suis pas statisticien, peut-être seulement un « amateur éclairé », surtout intéressé par comprendre les fondamentaux de mon activité de recherche en génétique quantitative. Cette présentation ne peut donc que bénéficier de votre indulgence.

Plan

Un peu de contexte en partant de loin...

Dans la pratique, avec les modèles paramétriques

Des modèles pour comprendre ? pour prédire ?

Plan

Un peu de contexte en partant de loin...

Dans la pratique, avec les modèles paramétriques

Des modèles pour comprendre ? pour prédire ?

Qu'est-ce que la science ?

Carl Sagan (1990) :

> *La science est une manière de penser bien plus qu'un corpus de connaissance.*

Qu'est-ce que la science ?

Claude Lévi-Strauss (1964) :

> *Le savant n'est pas l'homme qui fournit les vraies réponses ; c'est celui qui pose les vraies questions.*

Qu'est-ce que la science ?

François Jacob (2005) :

> *La science de jour met en jeu des raisonnements qui s'articulent comme des engrenages, des résultats qui ont la force de la certitude. [...] La science de nuit, au contraire, erre à l'aveugle. Elle hésite, trébuche, recule, transpire, se réveille en sursaut. Doutant de tout, elle se cherche, s'interroge, se reprend sans cesse. C'est une sorte d'atelier du possible où s'élabore ce qui deviendra le matériau de la science.*

Qu'est-ce que la science ?

en bref, il semblerait qu'un scientifique ne soit pas un ≪ automate à fournir les vraies réponses ≫...

des réactions ?

Qu'est-ce que la science ?

Karl Popper (1992) :

> *La science peut être décrite comme l'art de systématiquement sur-simplifier, l'art de discerner ce que l'on pourrait avantageusement omettre.*

Qu'est-ce que la modélisation (math/stat) ?

Peter McCullagh (2000) :

L'objectif principal de la modélisation statistique est d'augmenter notre compréhension via de la simplification.

Qu'est-ce que la modélisation (math/stat) ?

George Box (1979) :

Tous les modèles sont faux, mais certains sont utiles.

Qu'est-ce que la modélisation (math/stat) ?

Christian Hennig (2010) :

- revendication des mathématiques : fournir un espace de communication dans lequel l'accord absolu est possible car sans connotation individuelle ni dépendance à l'expérience
- modélisation mathématique : processus d'investigation des manières de penser la (les) réalité(s)
- son utilité : fournir des propositions à propos d'objets concrets en interprétant des résultats mathématiques vrais selon ces objets ; des négotiations étant nécessaires pour décider jusqu'à quel point telle ou telle interprétation peut être acceptée

Qu'est-ce que la modélisation *statistique* ?

Denis Lindley (2006) :

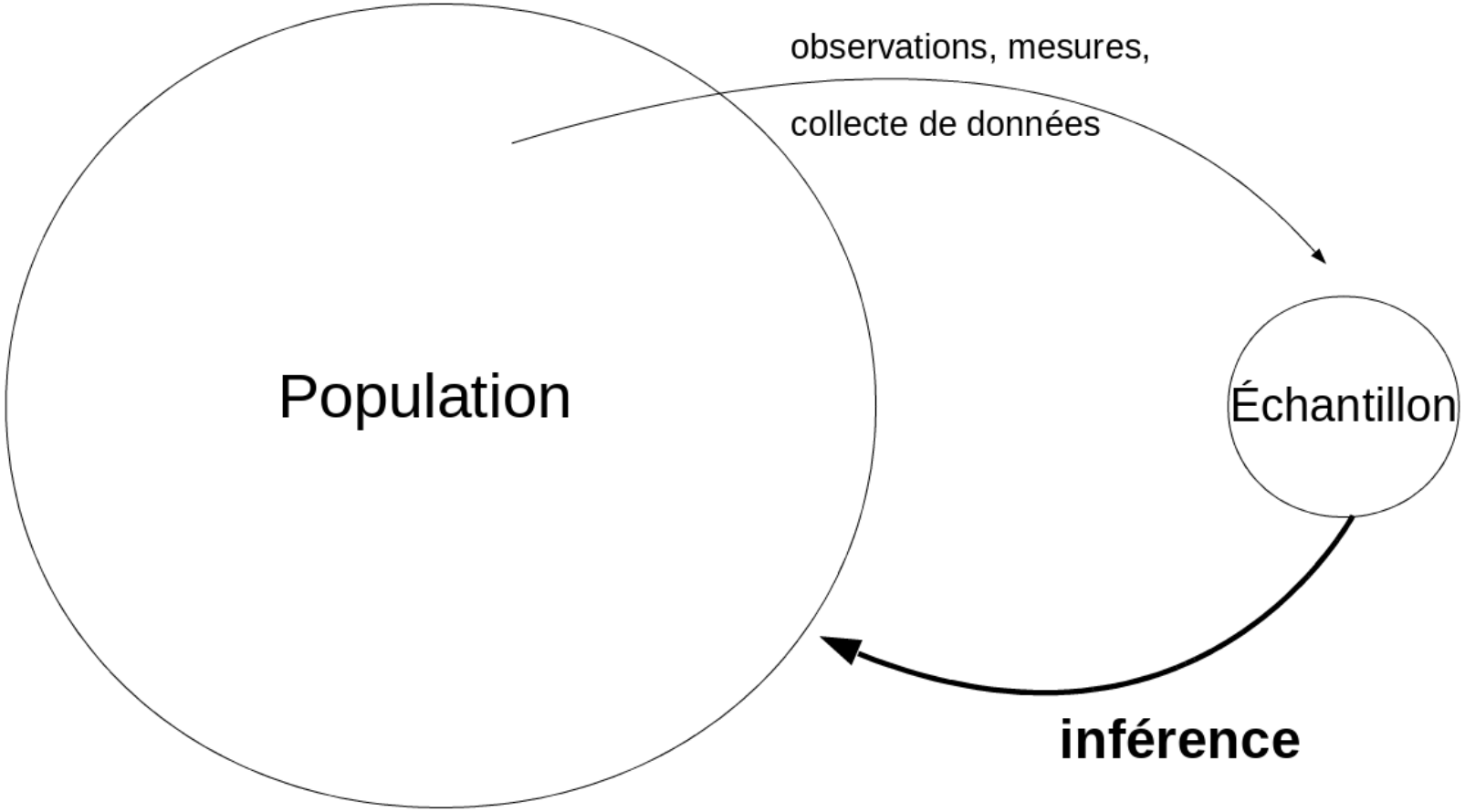
> *Il y a certaines choses que vous, le lecteur de cette préface, savez être vraies, et d'autres que vous savez êtres fausses ; et pourtant, malgré vos importantes connaissances, il reste certaines choses dont la véracité vous est inconnue. Nous disons que vous en êtes incertain. Vous êtes incertain, à des degrés divers, à propos de tout ce qui concerne le futur ; la plupart du passé vous est caché ; et il y a une grande partie du présent dont vous n'avez pas entièrement connaissance. L'incertitude est partout et vous ne pouvez pas y échapper.*

Qu'est-ce que la modélisation *statistique* ?

Denis Lindley (2000) :

- la statistique est l'étude de l'incertitude
- l'une des raisons de collecter des données est de réduire l'incertitude
- la quantification de l'incertitude peut être décrite par le calcul de probabilités

Qu'est-ce que la modélisation *statistique* ?



adapté de Kass (2011)

Qu'est-ce que la modélisation *statistique* ?

Stigler (1986) :

> *La statistique moderne fournit une technologie quantitative pour la science empirique ; c'est une logique et une méthodologie pour la mesure de l'incertitude et pour l'examination des conséquences de cette incertitude sur la planification et l'interprétation de l'expérimentation et de l'observation.*

Cherchez l'intrus

1. Est-ce que la loi Normale, $\frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{(y-\mu)^2}{2\sigma^2}\right)$, décrit bien la distribution de la taille des personnes vivant dans Paris ?

2. Quelle chance a le XV de France de remporter la prochaine Coupe du Monde de rugby ?

3. Aux alentours de quelle date a généralement lieu la floraison de la vigne ?

4. Est-ce que la dérivée de $log(x)$ vaut $1/x$?

5. Combien y a-t-il de lignes de code dans le logiciel Firefox ?

Cherchez l'intrus

1. Est-ce que la loi Normale, $\frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{(y-\mu)^2}{2\sigma^2}\right)$, décrit bien la distribution de la taille des personnes vivant dans Paris ?

2. Quelle chance a le XV de France de remporter la prochaine Coupe du Monde de rugby ?

3. Aux alentours de quelle date a généralement lieu la floraison de la vigne ?

4. Est-ce que $1 + 1$ est égal à 2 ?

5. Combien y a-t-il de lignes de code dans le logiciel Firefox ?

Science de l'incertitude... mais laquelle ?

Jaynes (1989) :

- *lorsque nous effectuons une inférence, nous sommes seulement en train de décrire notre état de connaissances à propos du monde*
- *les distributions de probabilités que nous utilisons pour l'inférence ne décrivent aucune propriété du monde, seulement un certain état de connaissance à propos du monde*

Science de l'incertitude... mais laquelle ?

Lindley (2006) :

- ce qui est incertain pour une personne ne l'est pas forcément pour une autre
- une probabilité est toujours conditionnelle à un certain état de connaissance
- initialement, les scientifiques diffèrent quant à leurs hypothèses, et éventuellement ils se mettent d'accord : l'objectivité apparente est en fait un consensus

Plan

Un peu de contexte en partant de loin...

Dans la pratique, avec les modèles paramétriques

Des modèles pour comprendre ? pour prédire ?

Pré-requis : validité

Gelman et Hill (2006) :

> *le plus important est que les données que vous analysez doivent correspondre à la question de recherche à laquelle vous essayez de répondre*

Pré-requis : planification

Casella (2008) :

> *pour Fisher, un échantillon est un morceau de minerai ; l'analyse statistique la plus fine ne pourra qu'extraire l'or présente dans le minerai ; mais une bonne planification pourra produire un minerai avec plus d'or*

(1) Analyse exploratoire des données (*EDA*)

	A	B	C	D
1	**year**	**geno**	**trait1**	**trait2**
2	2010	geno001	54.9320492152635	17.2175848219739
3	2010	geno002	48.6756226201732	21.5294686265656
4	2010	geno003	54.754035774377	NA
5	2010	geno004	49.4637979051225	18.6418965076544
6	2010	geno005	46.7505461430902	19.9008852499851
7	2010	geno006	54.652190385509	19.7948722701709
8	2010	geno007	54.2576400317033	17.9088460040552
9	2010	geno008	49.4732685681703	20.5136700614508
10	2010	geno009	54.1285913557059	19.7516873537991
11	2010	geno010	57.887870090051	18.7593804357203
12	2010	geno011	50.7148753249191	NA
13	2010	geno012	47.6058412450121	16.623637375049
14	2010	geno013	NA	NA
15	2010	geno014	47.2372813607911	19.5382688837415
16	2010	geno015	44.8651173988901	19.7289094120696
17	2010	geno016	NA	22.5007029439046

(1) Analyse exploratoire des données (*EDA*)

Se faire une idée de la tendance moyenne et de la dispersion, par exemple du premier caractère la première année qu'il a été mesuré :

min	q1	med	mean	q3	max
39.44	49.45	52.64	51.94	54.85	62.16

(1) Analyse exploratoire des données (*EDA*)

trait	year	n	na	min	q1	med	mean	q3	max
trait1	2010	200	36	39.44	49.45	52.64	51.94	54.85	62.16
trait1	2011	200	39	42.31	53.03	55.80	55.60	58.09	67.91
trait1	2012	200	44	45.74	50.87	53.88	53.70	55.98	62.45
trait1	2013	200	40	40.64	50.11	53.04	52.85	55.21	67.37
trait1	2014	200	41	45.75	55.82	58.20	58.20	60.74	69.62
trait2	2010	200	44	14.51	17.37	18.78	18.86	20.17	24.41
trait2	2011	200	32	15.12	20.76	21.91	22.17	23.75	27.25
trait2	2012	200	44	18.42	22.67	23.84	24.18	25.90	29.29
trait3	2010	200	32	25.06	30.32	33.01	33.09	35.31	42.94
trait3	2011	200	19	28.51	35.71	38.77	38.65	41.75	47.83
trait3	2012	200	17	25.04	32.82	35.54	35.67	37.95	47.78
trait3	2013	200	14	28.69	35.68	37.91	37.80	39.93	51.43
trait3	2014	200	18	27.61	34.61	37.49	37.51	39.91	48.77

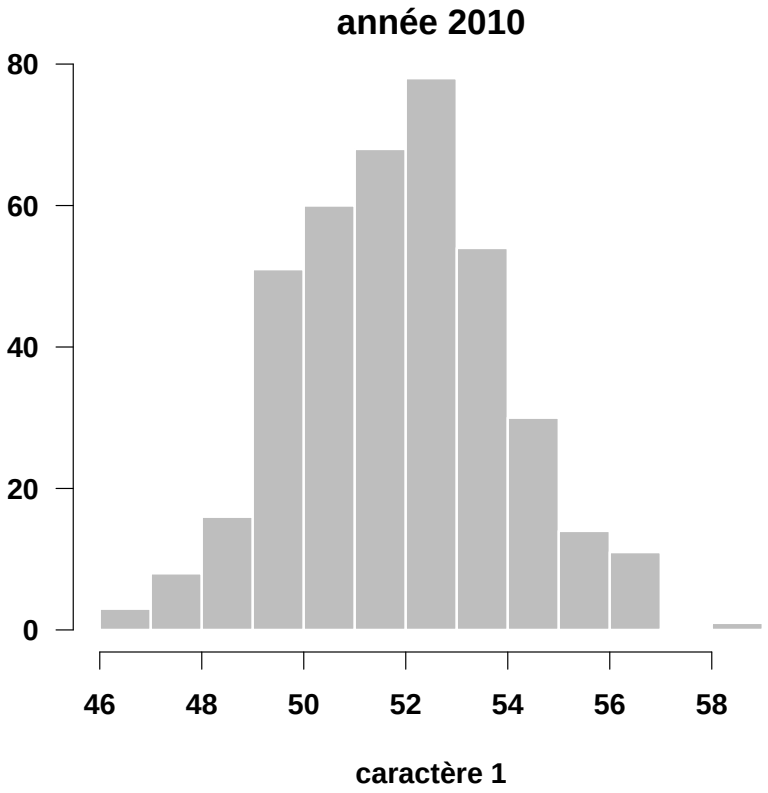
(1) Analyse exploratoire des données (*EDA*)

Gelman et al (2002, 2004) :

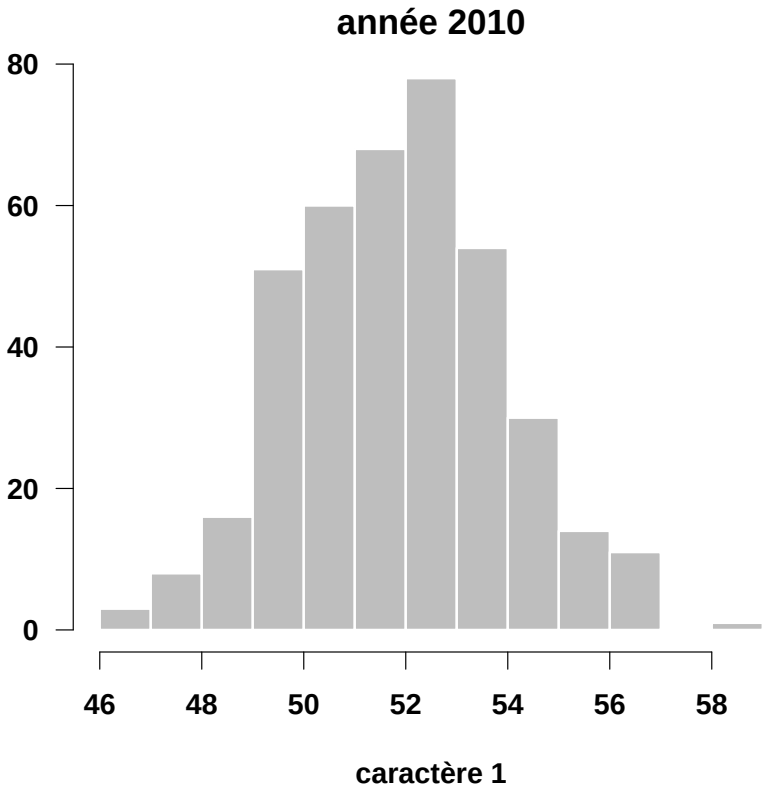
- *l'oeil est attiré par les tendances qui violent les symmétries attendues, ou par les répétitions*
- *à faire et à lire, un graphique nécessite un plus grand effort initial qu'un tableau, mais une fois qu'il est compris, il peut être utilisé pour faire des comparaisons quasi impossibles à faire avec un tableau*

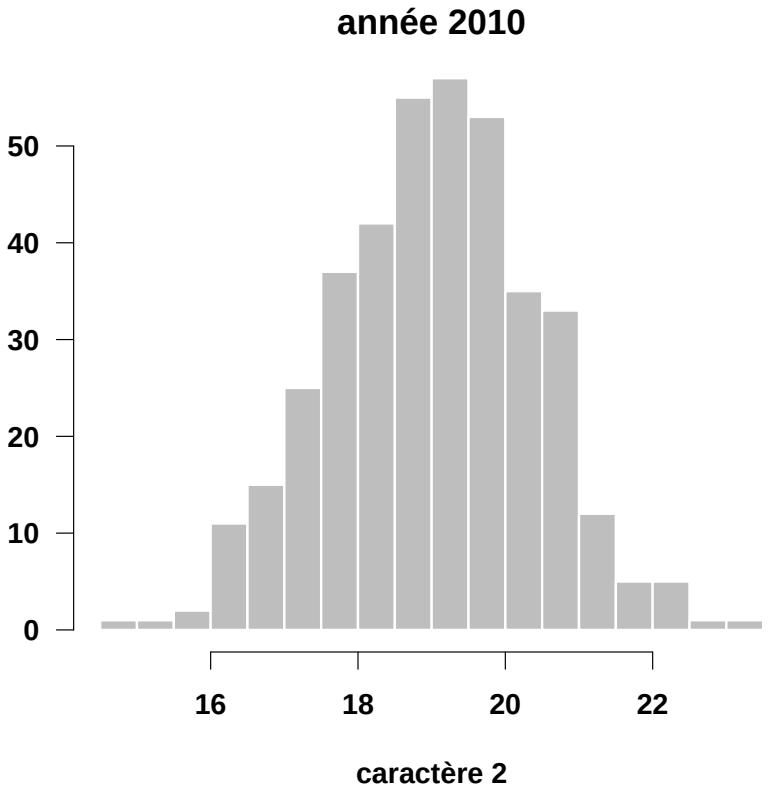
- comprendre que l'idée d'un graphique est de réaliser implicitement une comparaison avec une distribution de référence [modèle implicite]

(1) Analyse exploratoire des données (*EDA*)

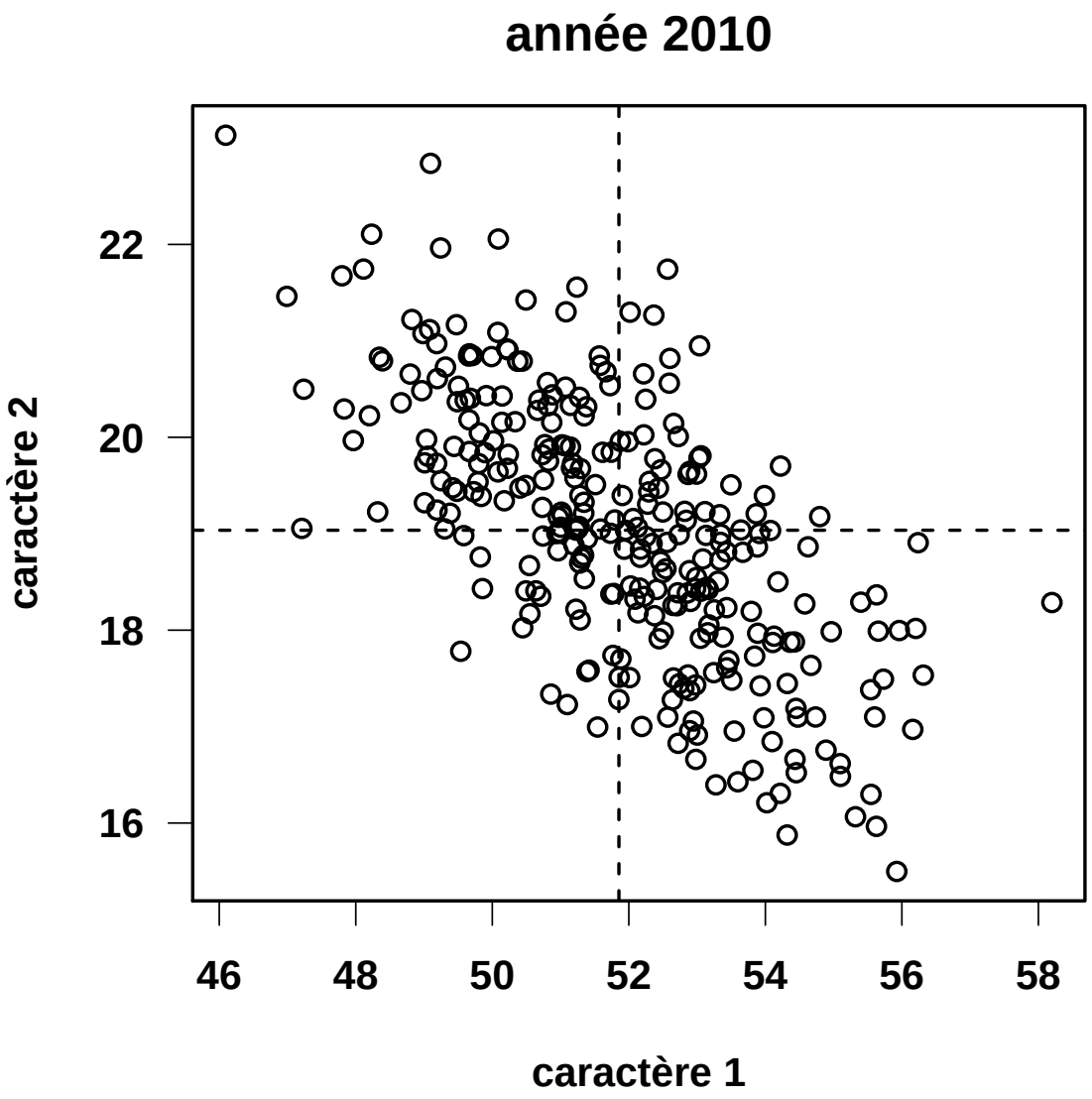


(1) Analyse exploratoire des données (*EDA*)

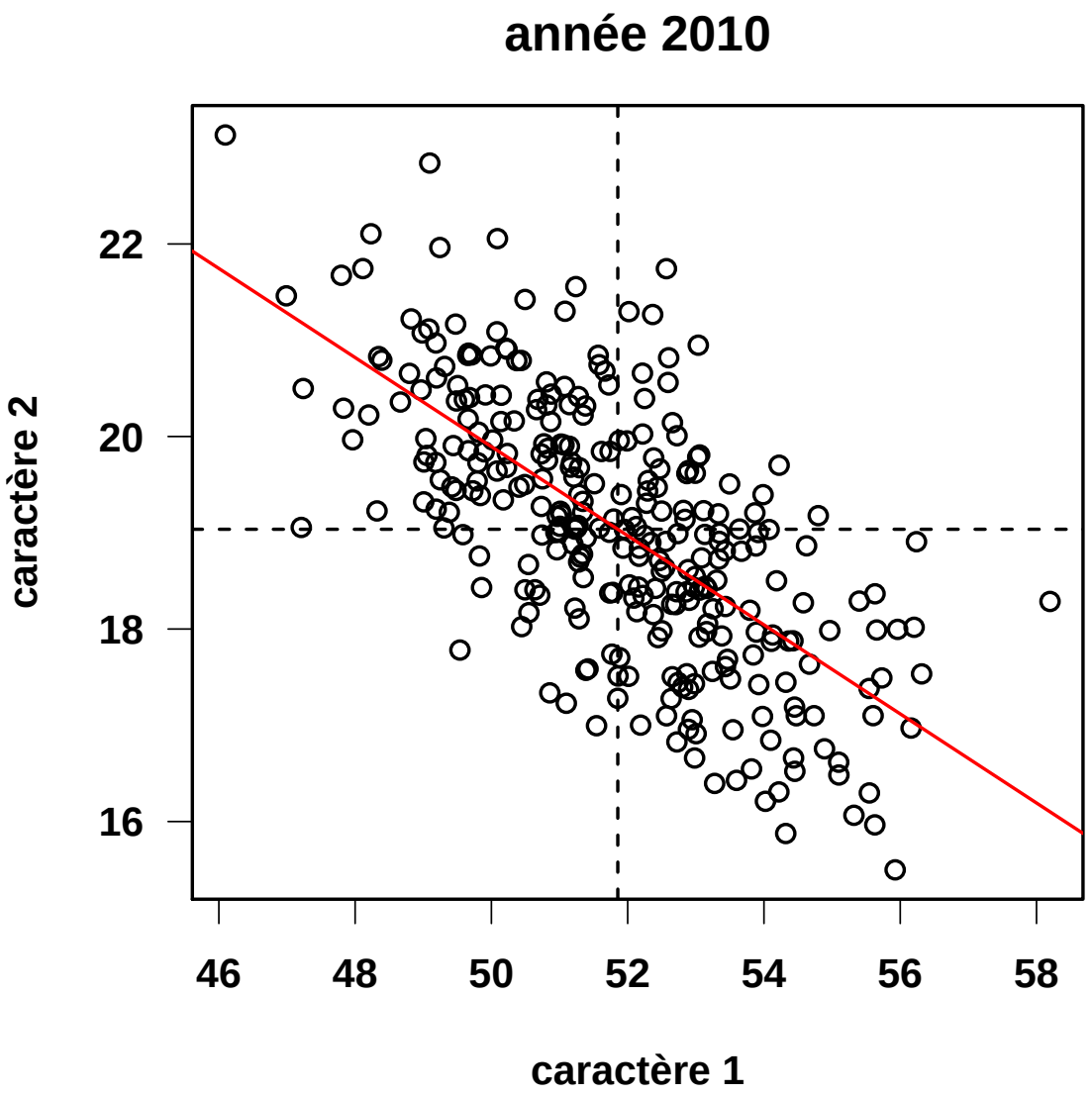




(1) Analyse exploratoire des données (*EDA*)



(1) Analyse exploratoire des données (*EDA*)



(1) Analyse exploratoire des données (*EDA*)

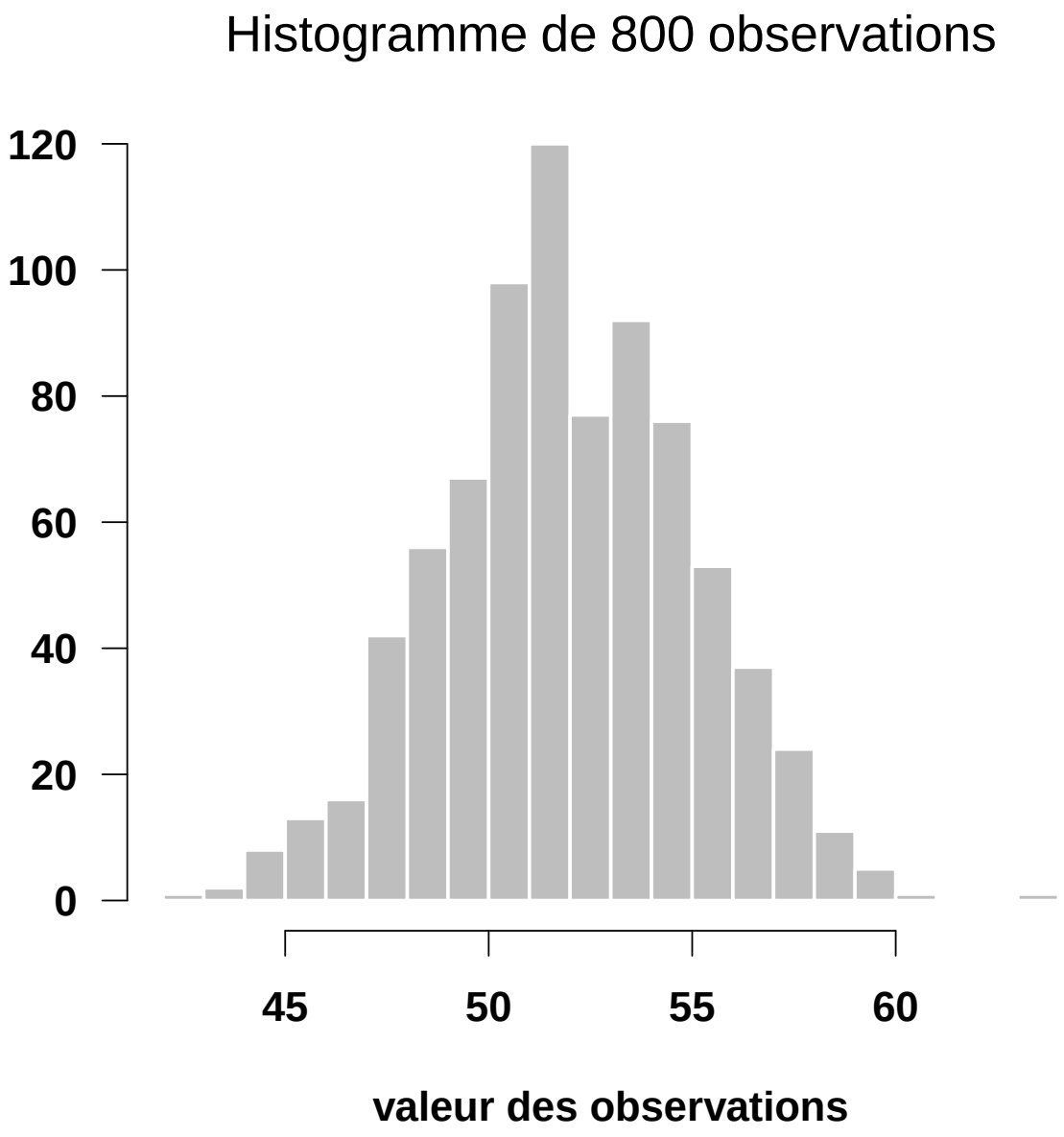
quelles autres questions, et donc graphiques ?

Introduction des variables aléatoires

Kass (2011) :

- *utiliser les probabilités pour décrire la variation dans les données*
- *besoin d'introduire des objets mathématiques appelés « variables aléatoires » [qui suivent une loi de probabilités]*
- *celles-ci sont des abstractions, potentiellement très utiles lorsque leur monde théorique est « bien aligné » avec le monde réel [des données]*

Introduction des variables aléatoires

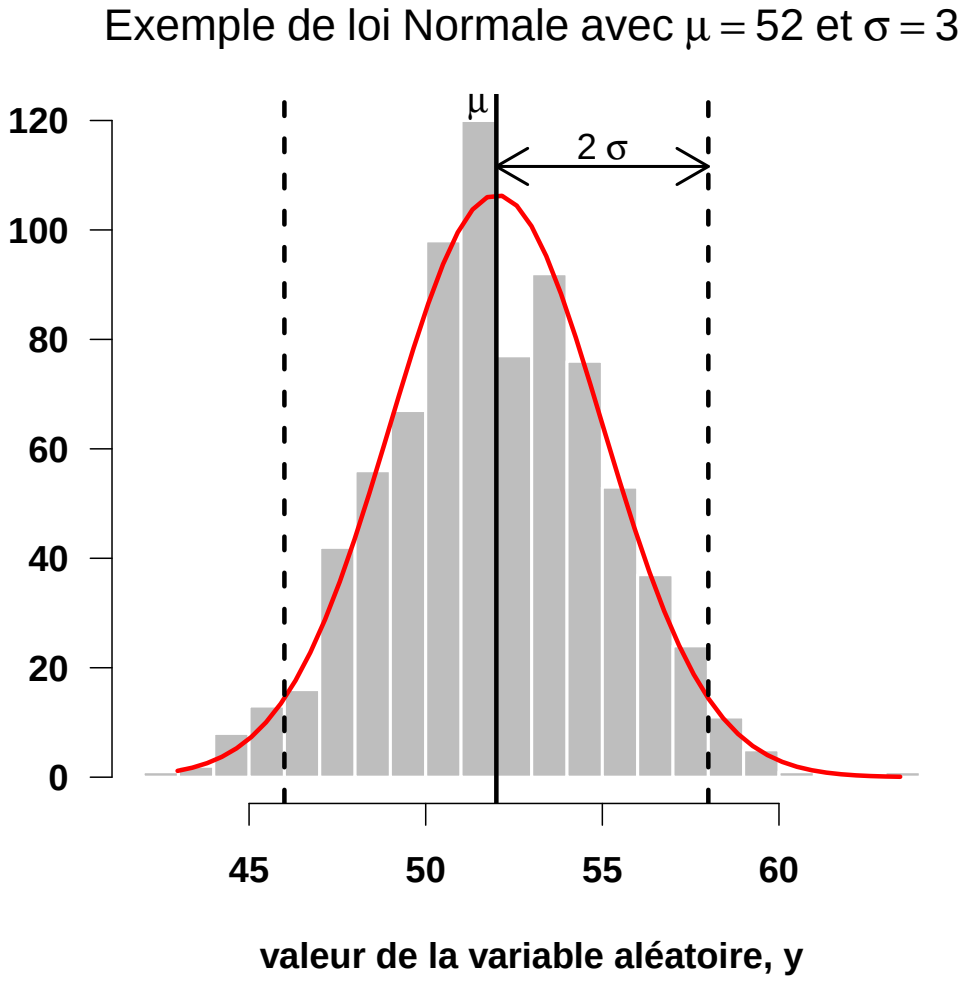


Introduction des variables aléatoires

Variable aléatoire Y :

- qui suit une loi Normale $\mathcal{N}$
- de moyenne μ
- et de variance σ^2

$\Rightarrow$ 95% de chance d'être entre $\mu - 2\sigma$ et $\mu + 2\sigma$



Modèle probabiliste versus statistique

modèle probabiliste	modèle statistique
$Y \sim \mathcal{N}(\mu, \sigma^2)$ avec μ et σ connus	$\forall i \in \{1, \ldots, n\},\ y_i \sim \mathcal{N}(\mu, \sigma^2)$ avec μ et σ inconnus
permet de faire des simulations	permet de faire de l'inférence

Modèle probabiliste versus statistique

modèle probabiliste	modèle statistique
$Y \sim \mathcal{N}(\mu, \sigma^2)$ avec μ et σ connus	$\forall i \in \{1, \ldots, n\},\ y_i \sim \mathcal{N}(\mu, \sigma^2)$ avec μ et σ inconnus
permet de faire des simulations	permet de faire de l'inférence

$\Rightarrow$ un modèle statistique correspond donc à un ensemble de modèles probabilistes

$\Rightarrow$ la modélisation statistique chercher donc à résoudre le « problème inverse » : trouver la distribution de probabilités des variables aléatoires décrivant la variation des données

« Pont » entre mondes théorique et réel(s)

Exemple tiré de Kass (2011) :

1. imaginons que l'on dispose des $n = 800$ observations, que l'on va appeler $\{y_1, \ldots, y_n\}$, dont la moyenne, $\bar{y}$, vaut $51,7$

2. nous introduisons maintenant n variables aléatoires, celles-ci notées $\{Y_1, \ldots, Y_n\}$, chacune correspondant à une observation, en supposant qu'elles sont indépendantes et suivent la même loi Normale, dont la moyenne théorique, elle aussi une variable aléatoire, est notée $\bar{Y}$

$\Rightarrow$ nous « construisons et traversons le pont » lorsque l'on autorise $\bar{y}$ à correspondre à la fois à la valeur observée et à la valeur théorique de la variable aléatoire $\bar{Y}$

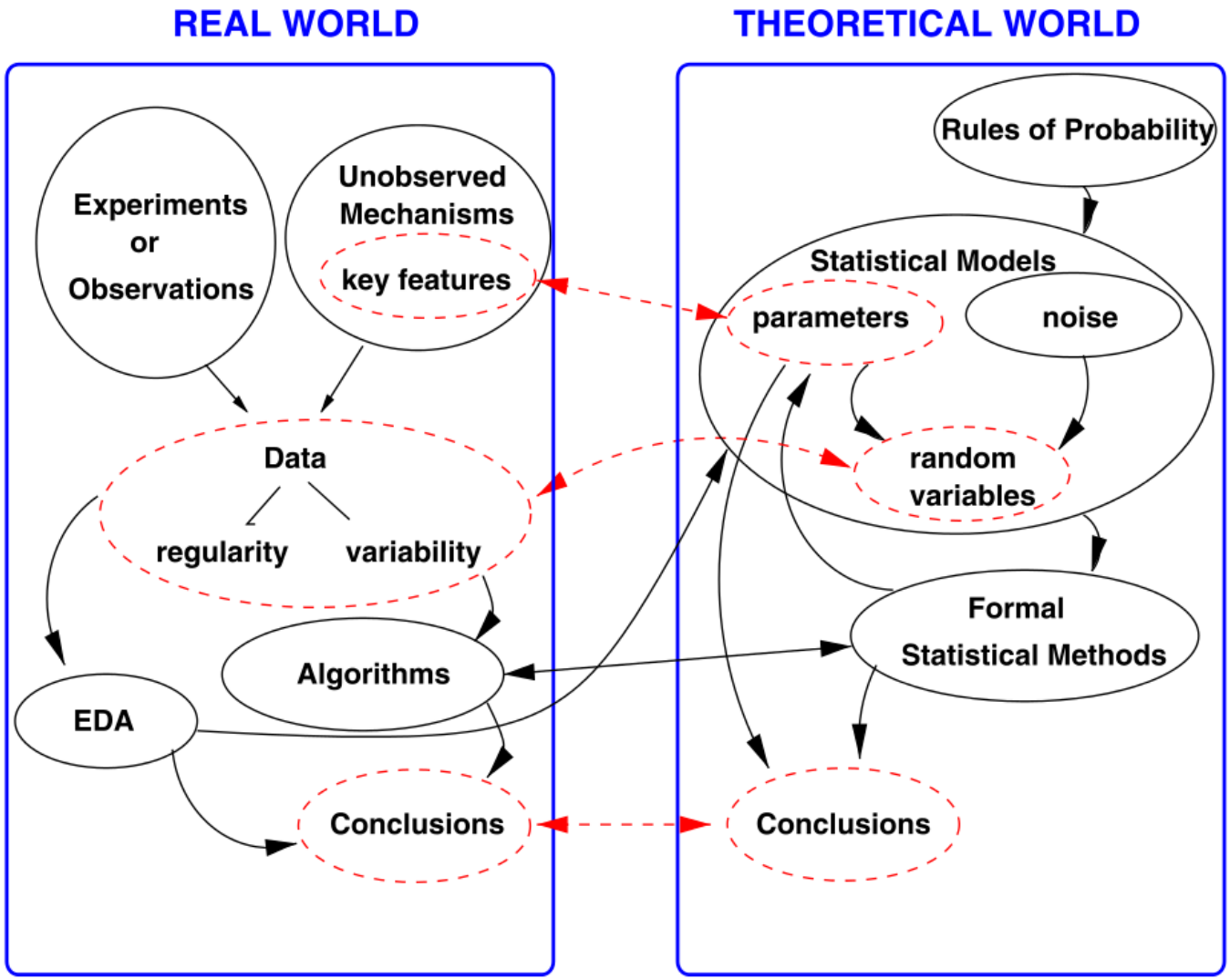
« Pont » entre mondes théorique et réel(s)

Un modèle statistique peut être interprété comme une déclaration au conditionnel.

Kass (2011) :

- *quand un modèle statistique est introduit, nous pouvons dire que l'inférence est basée sur ce qui arriverait si les données sont des variables aléatoires distribuées selon le modèle*
- *si les hypothèses de modélisation sont raisonnables, le modèle décrirait de manière précise la variation dans les données*

« Pont » entre mondes théorique et réel(s)



Notation et vocabulaire

variables observées $\rightarrow$ données (observationnelles/expérimentales)

- ex. : $\mathcal{D} = \{y_1, \ldots, y_n\}$
- si y_i est la taille de l'individu i, on peut raisonnablement supposer que y_i est entre $0,5$ m et $2,5$ m

variables non-observables $\rightarrow$ paramètres, à estimer

- ex. : $\Theta = \{\mu, \sigma\}$ de la loi Normale $\mathcal{N}(\mu, \sigma^2)$
- la moyenne μ prend ses valeurs entre $-\infty$ et $+\infty$, et la variance σ^2 est strictement positive

variables observables mais non-observées $\rightarrow$ prédictions

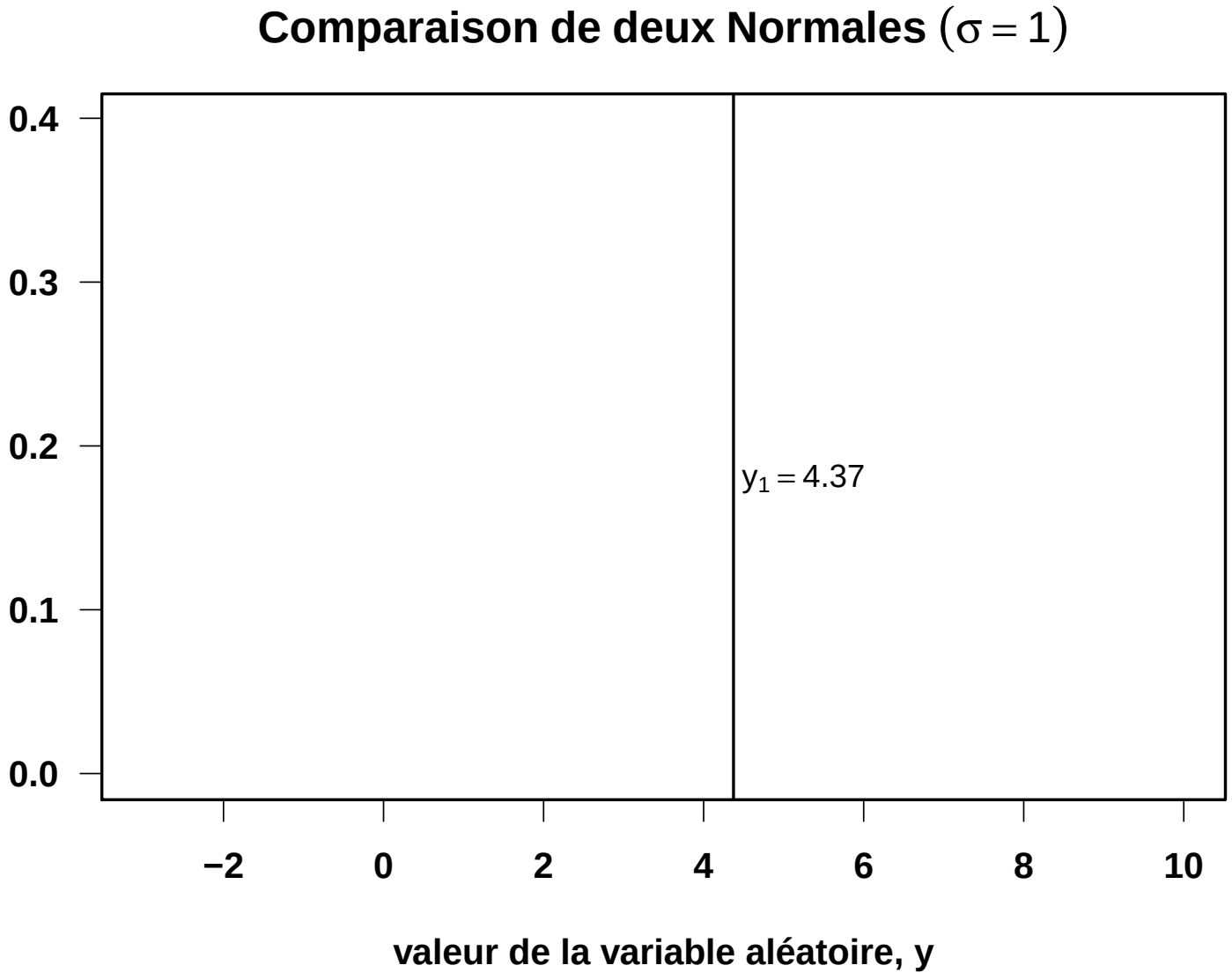
- ex. : $\tilde{y}_{n+1}, \tilde{y}_{n+2}, \ldots$

Notation et vocabulaire

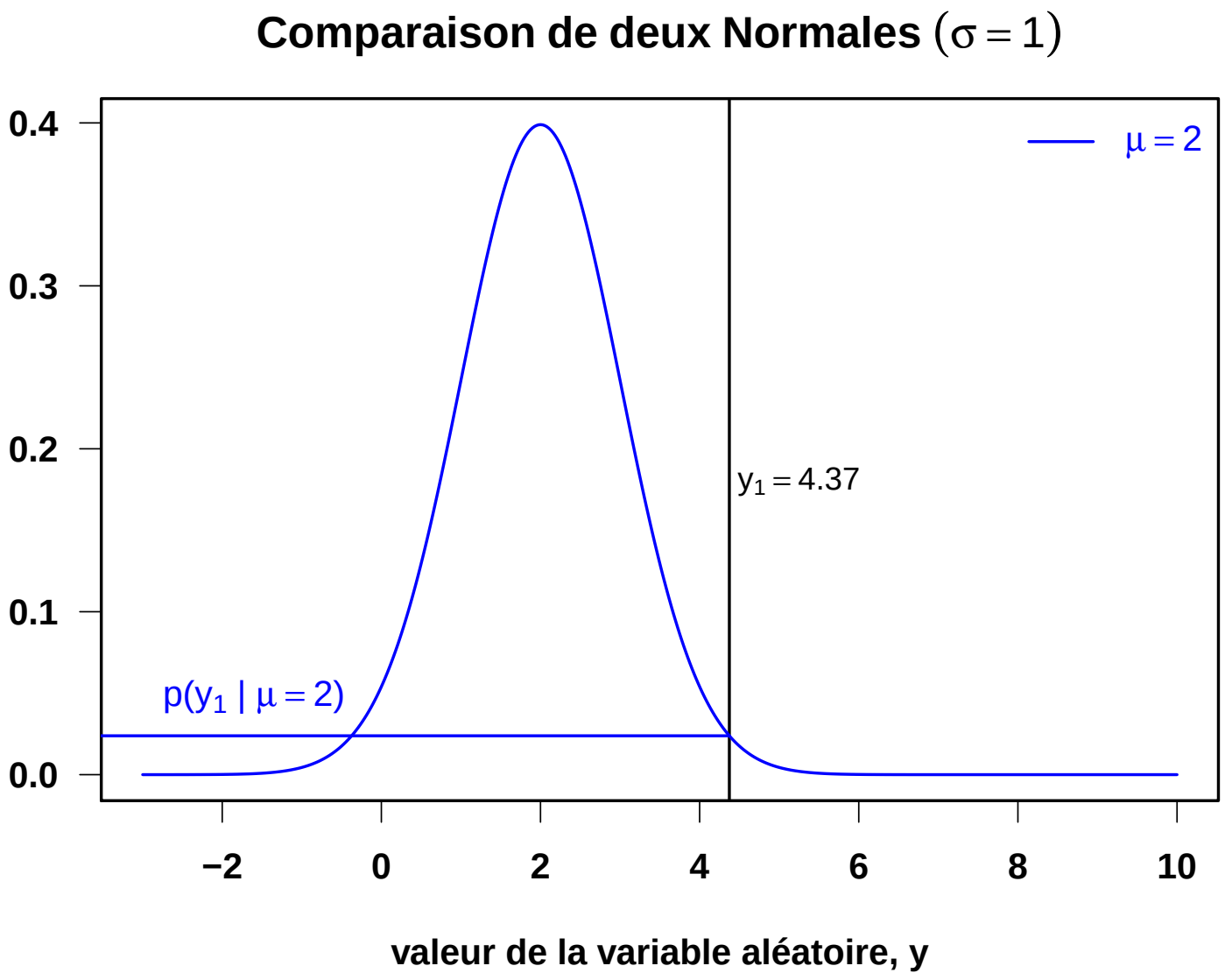
vraisemblance (*likelihood*) :

- fonction des paramètres (inconnus)
- notée $\mathcal{L}(\Theta) = p(\mathcal{D} \mid \Theta)$
- prend une valeur d'autant plus élevée que la probabilité d'observer les données avec certaines valeurs des paramètres est élevée
- utile pour estimer les paramètres
- ex. : tous les y_i sont indépendants et suivent la loi $\mathcal{N}(\mu, \sigma^2)$

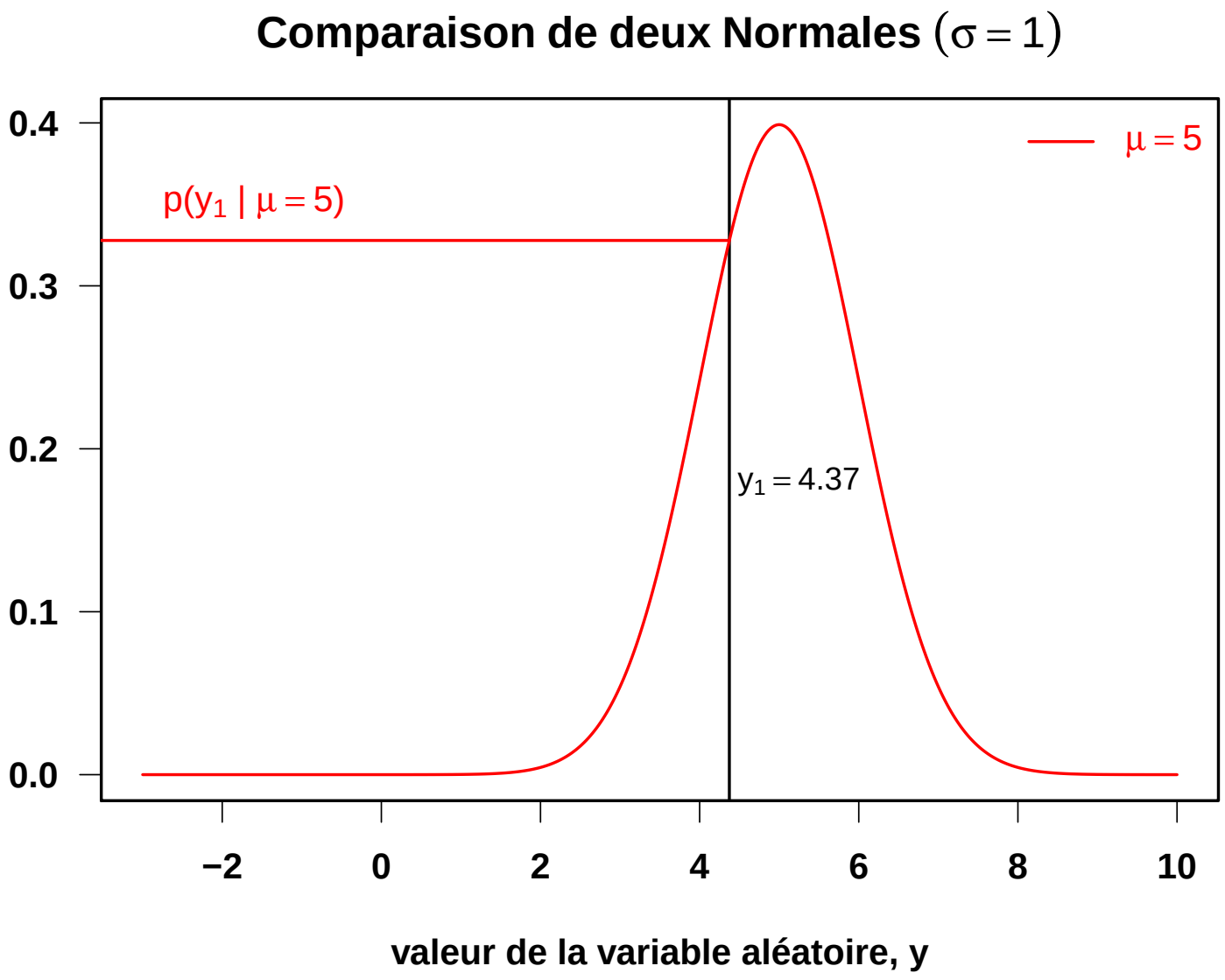
Comparaison de vraisemblances



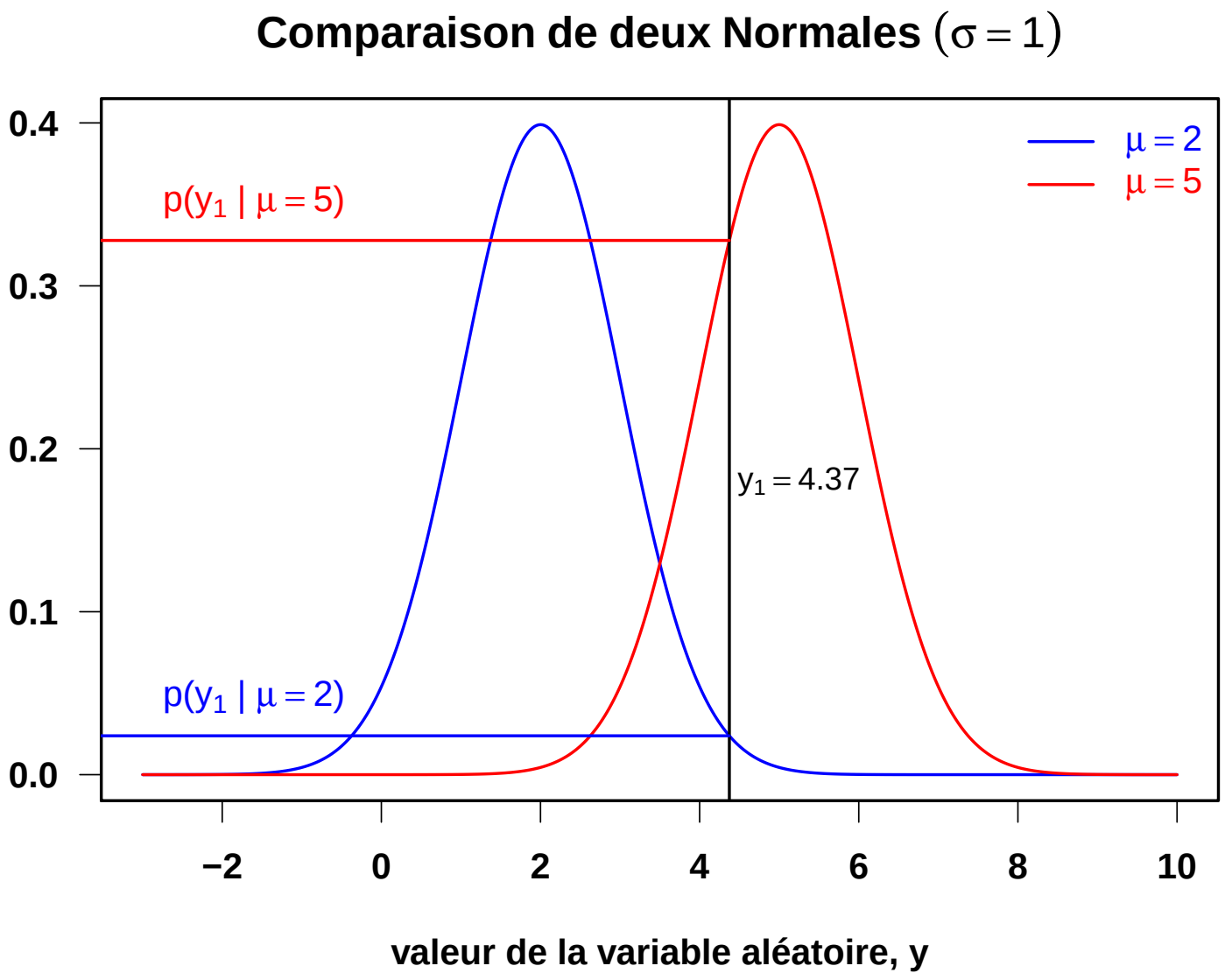
Comparaison de vraisemblances



Comparaison de vraisemblances



Comparaison de vraisemblances



(2) Ecriture du modèle statistique

lister les indices

- ex. : i pour les génotypes, d pour les caractères, j pour les parcelles, k pour les années, t pour le temps (intra-année)...

(2) Ecriture du modèle statistique

lister les indices

- ex. : i pour les génotypes, d pour les caractères, j pour les parcelles, k pour les années, t pour le temps (intra-année)...

décrire à quoi correspond chaque variable aléatoire

- ex. : $y_{d=1,i,j,k}$ est le rendement, $y_{d=2,i,j,k}$ est la qualité, ...

(2) Ecriture du modèle statistique

lister les indices

- ex. : i pour les génotypes, d pour les caractères, j pour les parcelles, k pour les années, t pour le temps (intra-année)...

décrire à quoi correspond chaque variable aléatoire

- ex. : $y_{d=1,i,j,k}$ est le rendement, $y_{d=2,i,j,k}$ est la qualité, ...

écrire la distribution de probabilités reliant les variables aléatoires

- ex. : $\mathcal{L}(\Theta) = p(y_1|\mu,\sigma) \times \ldots \times p(y_n|\mu,\sigma_n)$
  en supposant $p(y_i|\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{(y_i-\mu)^2}{2\sigma^2}\right)$
- à cette étape, beaucoup d'hypothèses sont faites ! mieux vaut en être conscient et savoir les expliciter...

(3) Inférence (aspects techniques)

Quelles tâches reste-il à faire ?

- comparaison et sélection de modèles
- test d'une hypothèse nulle
- estimation de paramètres, et quantification de l'incertitude
- prédiction de données

(3) Inférence (aspects techniques)

Pour tout jeu de données, la difficulté principale réside dans la construction itérative d'un « modèle utile »...

- *la modélisation statistique est un processus*

... car, pour un modèle donné, la théorie des probabilités fournit « automatiquement » la valeur estimée du paramètre, l'intervalle quantifiant son incertitude, la force avec laquelle telle hypothèse nulle est rejetée, etc

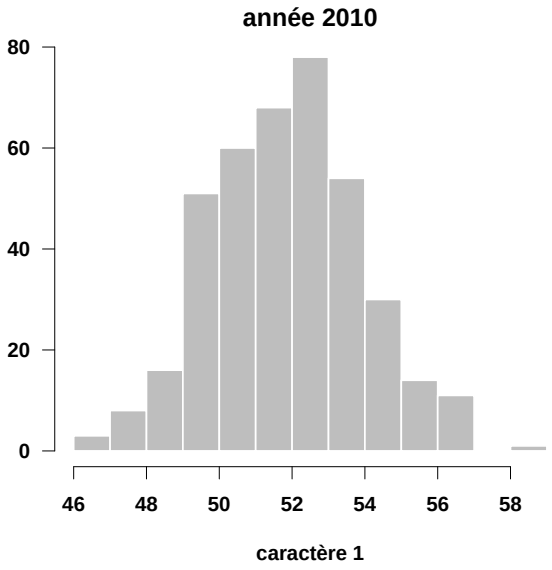
- *en supposant que l'on dispose d'un logiciel mettant en oeuvre la méthode et l'algorithme souhaité en un temps raisonnable !*

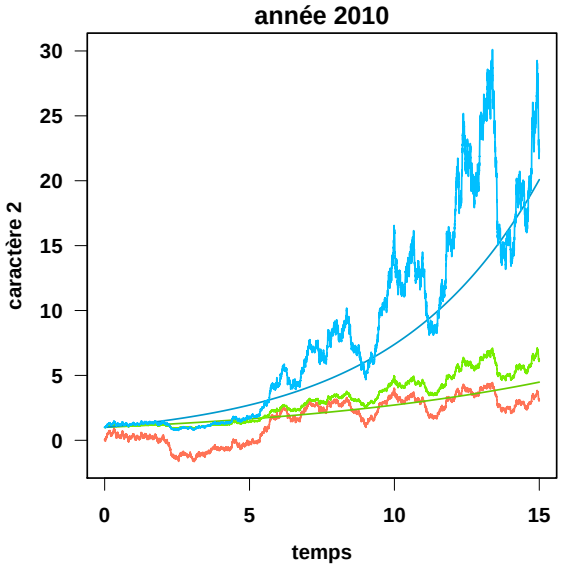
Exemple de front de recherche d'intérêt pour AGAP

$\Rightarrow$ modéliser conjointement l'architecture génétique de plusieurs caractères, tout en tenant compte de l'hétérogénéité spatiale

Exemple de front de recherche d'intérêt pour AGAP

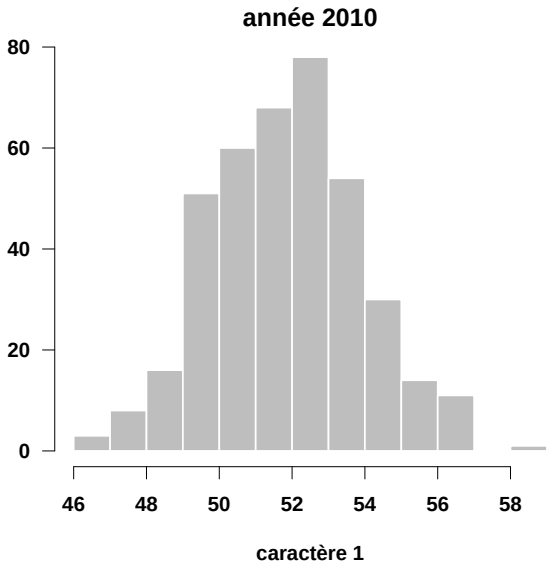
$\Rightarrow$ modéliser conjointement l'architecture génétique de plusieurs caractères, tout en tenant compte de l'hétérogénéité spatiale

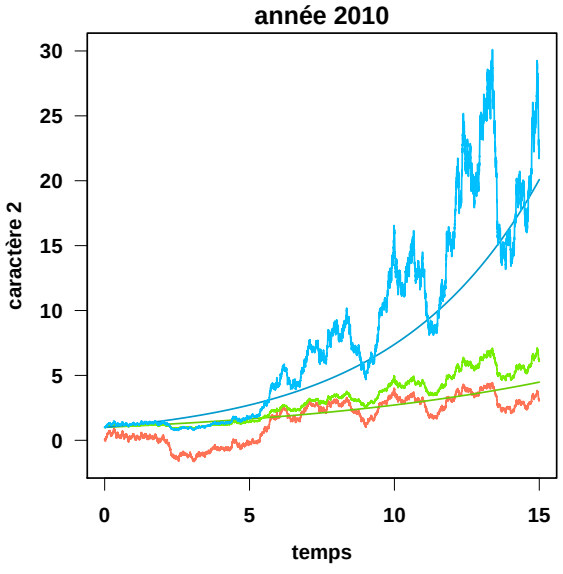


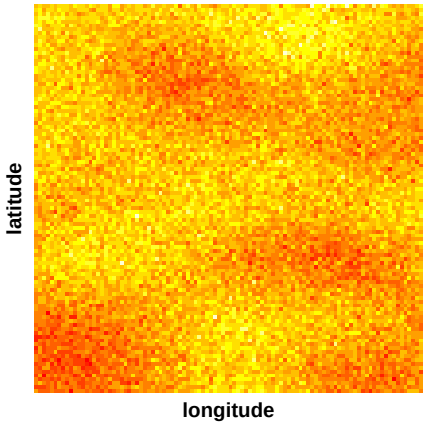


Exemple de front de recherche d'intérêt pour AGAP

$\Rightarrow$ modéliser conjointement l'architecture génétique de plusieurs caractères, tout en tenant compte de l'hétérogénéité spatiale

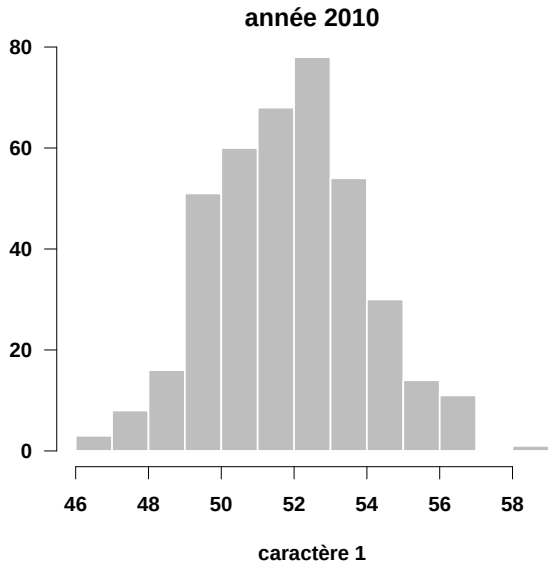


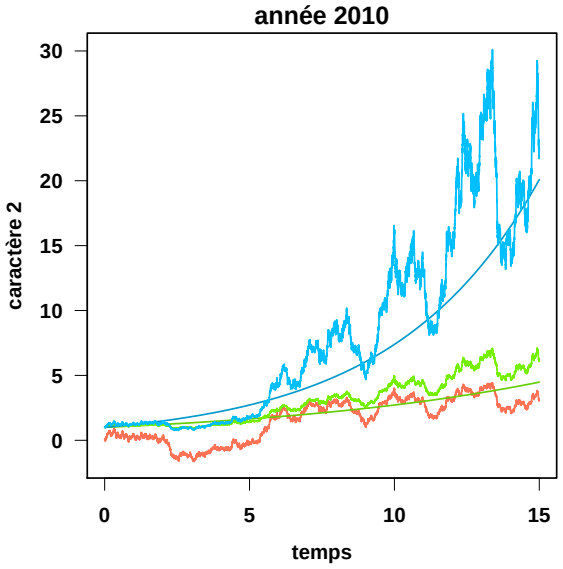


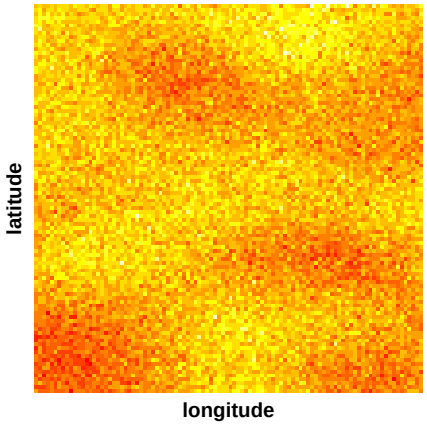


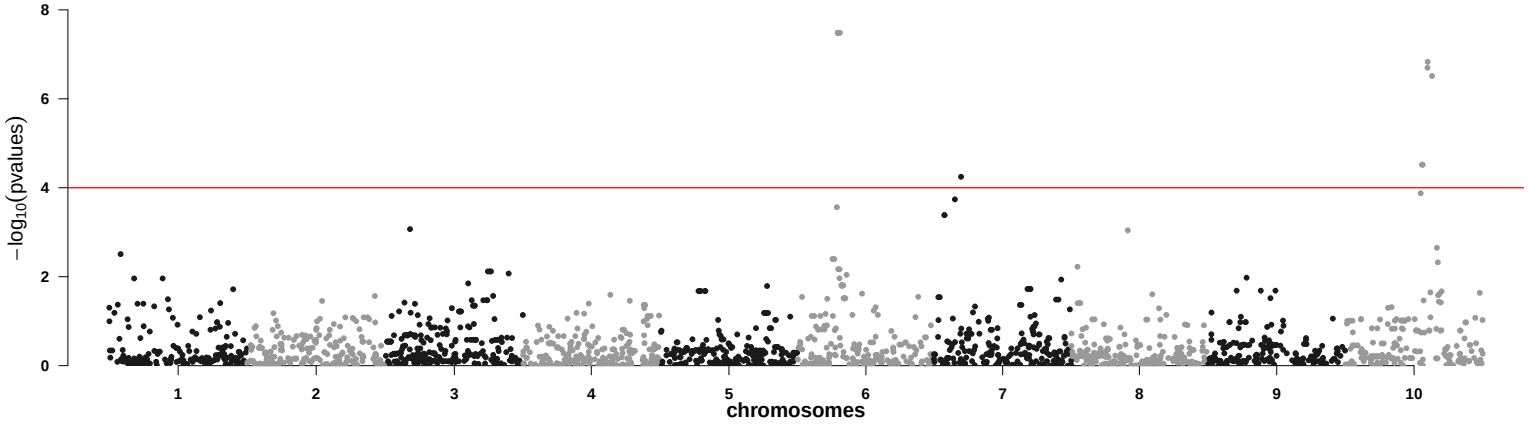
Exemple de front de recherche d'intérêt pour AGAP

$\Rightarrow$ modéliser conjointement l'architecture génétique de plusieurs caractères, tout en tenant compte de l'hétérogénéité spatiale









Plan

Un peu de contexte en partant de loin...

Dans la pratique, avec les modèles paramétriques

Des modèles pour comprendre ? pour prédire ?

Modèle statistique vs mécanistique ?

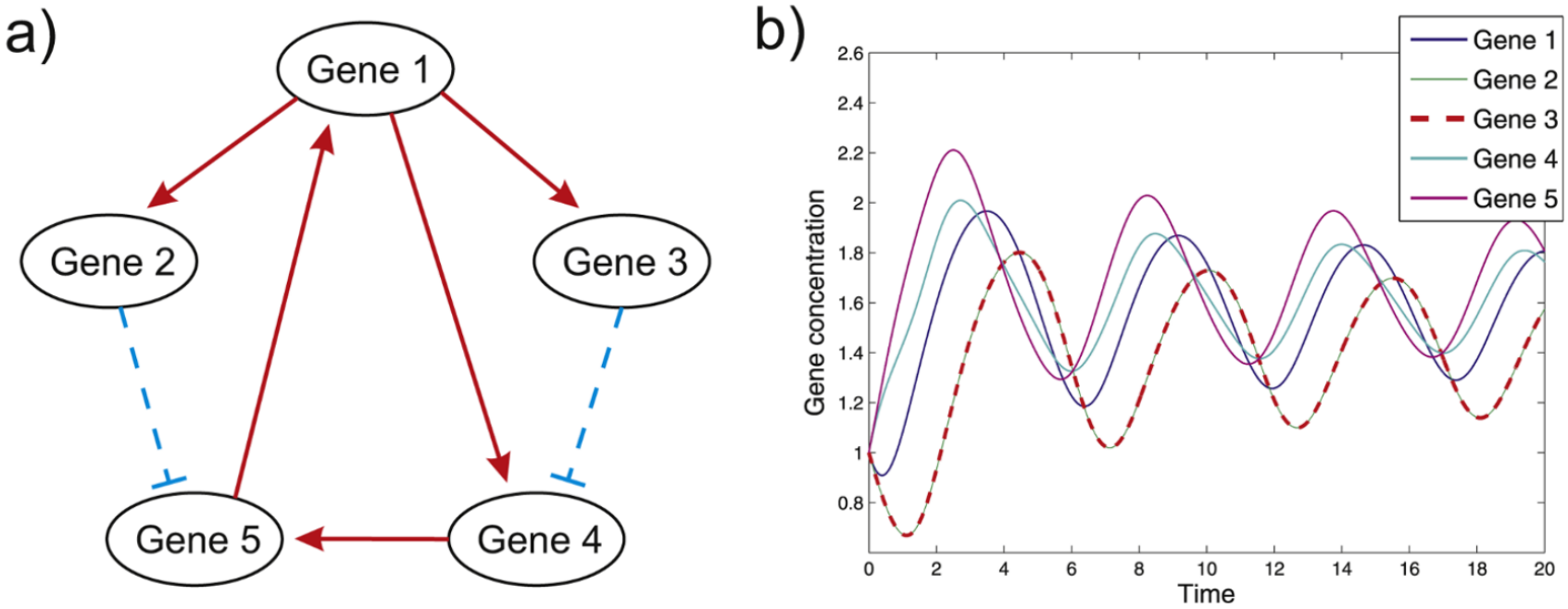
Exemple récent d'un résumé de séminaire (mars 2017) :

> *Cette présentation aura pour but de décrire le rôle des modèles dynamiques basés sur une compréhension biologique, comme moyen de relier les niveaux moléculaire et plante-entière, et comment ceci confère un avantage au-delà des méthodes statistiques pures qui relient les génotypes aux phénotypes pour l'amélioration des plantes, notamment en ce qui concerne les interactions génotype-environnement.*

- les « méthodes statistiques pures » ne prendraient-elles pas en compte la connaissance biologique ? !

Modèle statistique vs mécanistique ?

Exemple d'inférence de réseau de régulation génique (paramètres de synthèse, dégradation et force de régulation) :



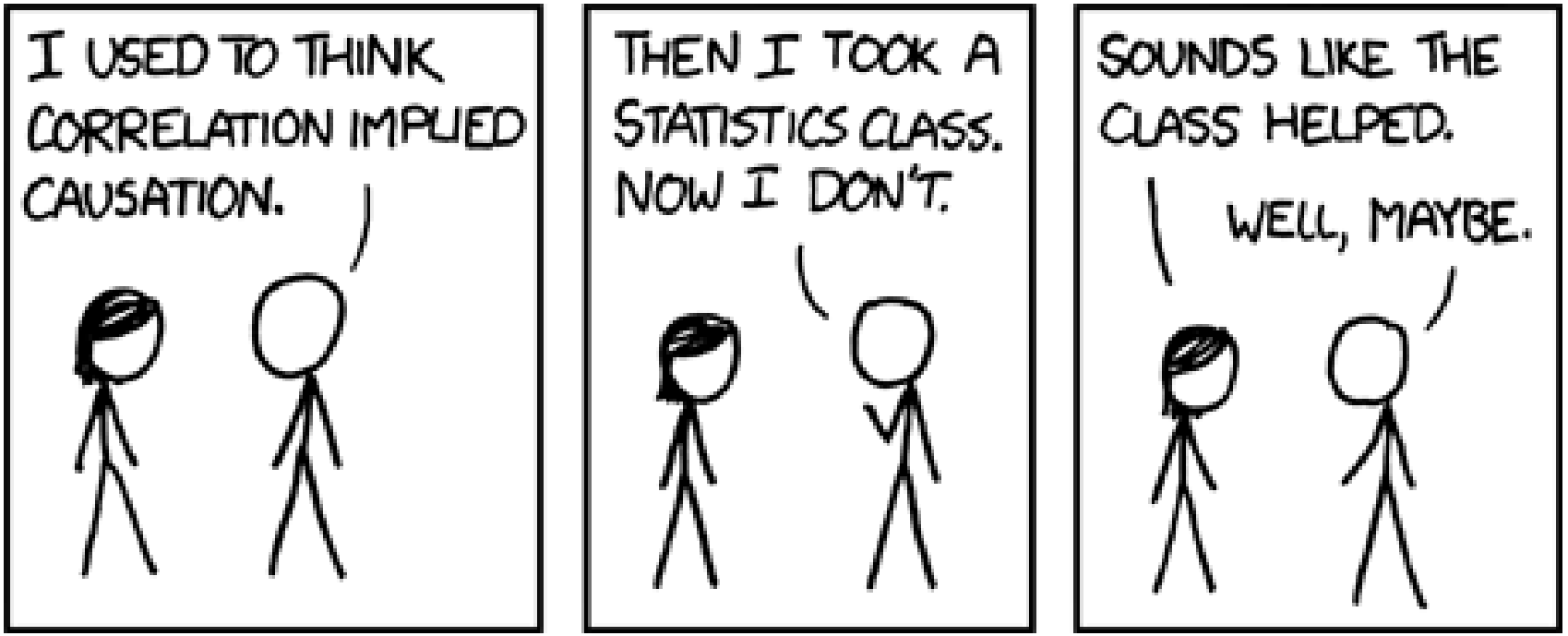
Mazur et coll. (2009)

Modèle statistique vs mécanistique ?

McCullagh (2002) :

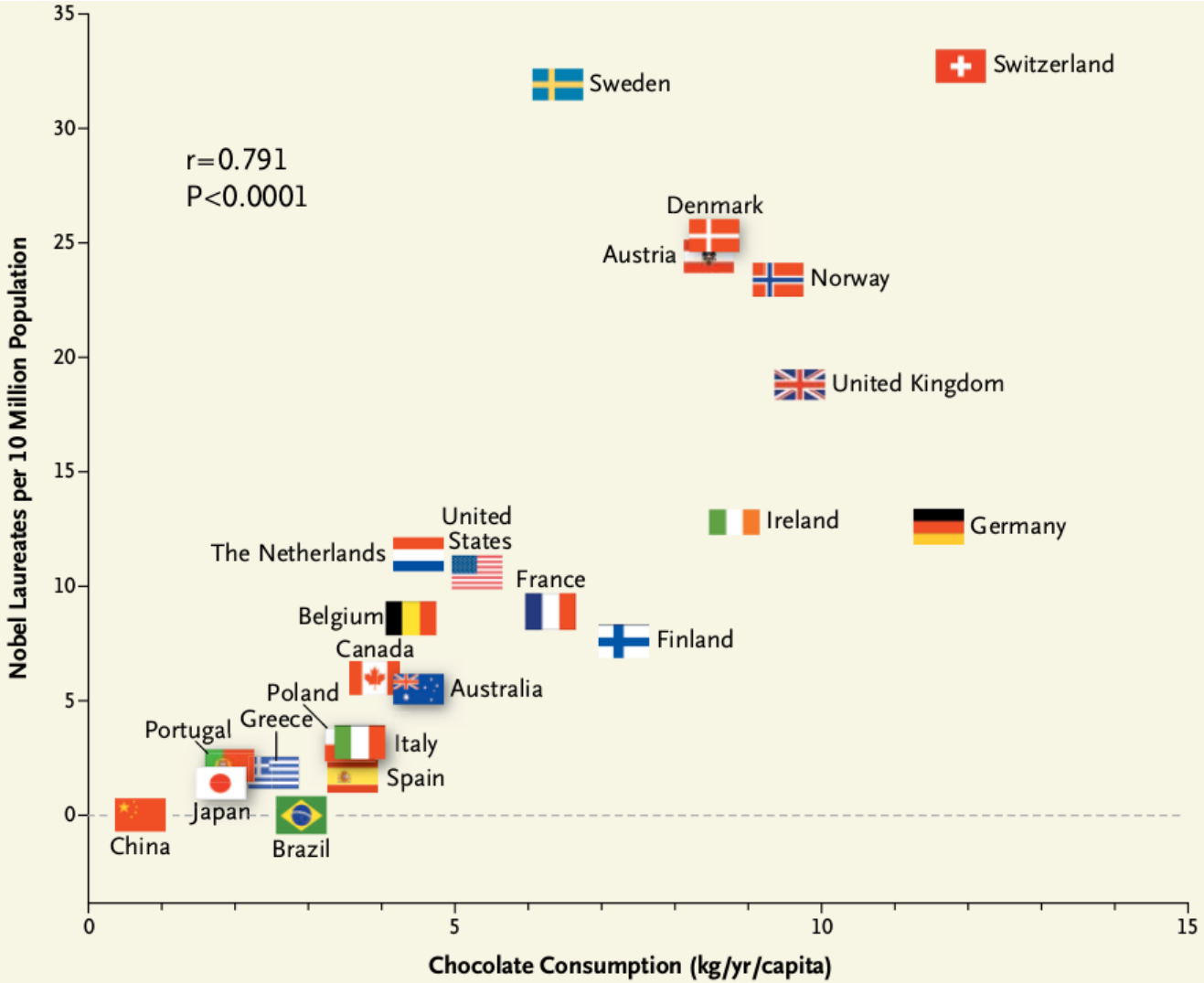
- *un certain modèle [statistique] utilisé dans des domaines d'application très différents pourrait avoir de nombreuses interprétations causales différentes*
- *les scientifiques demandent une explication physique ou mécanistique, dans notre cas [à nous, les statisticiens] une explication de l'origine de l'aléatoire dans la nature*

Corrélation n'est pas causalité



https ://www.xkcd.com/552/

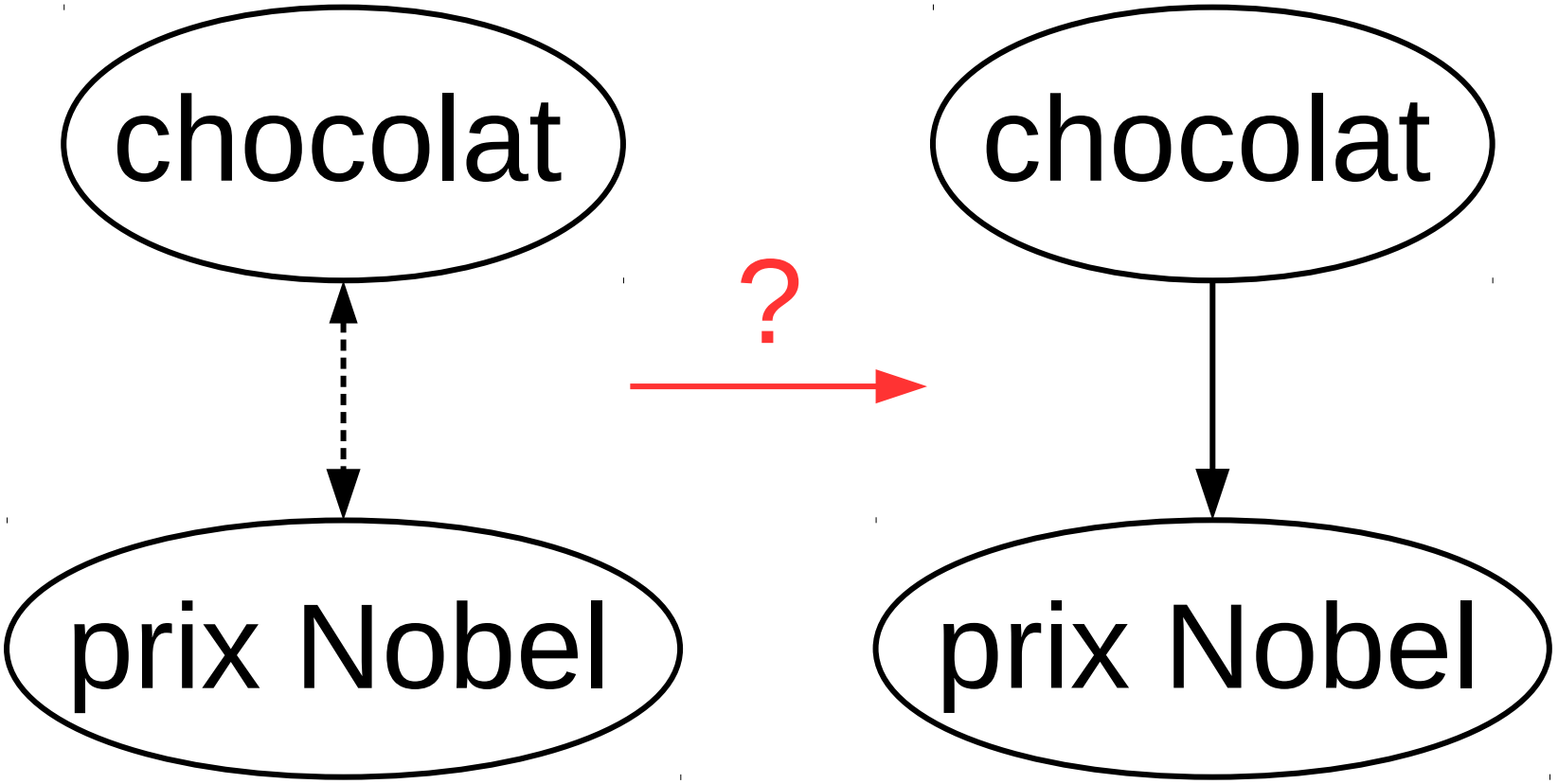
Corrélation n'est pas causalité



Messerli (2012)

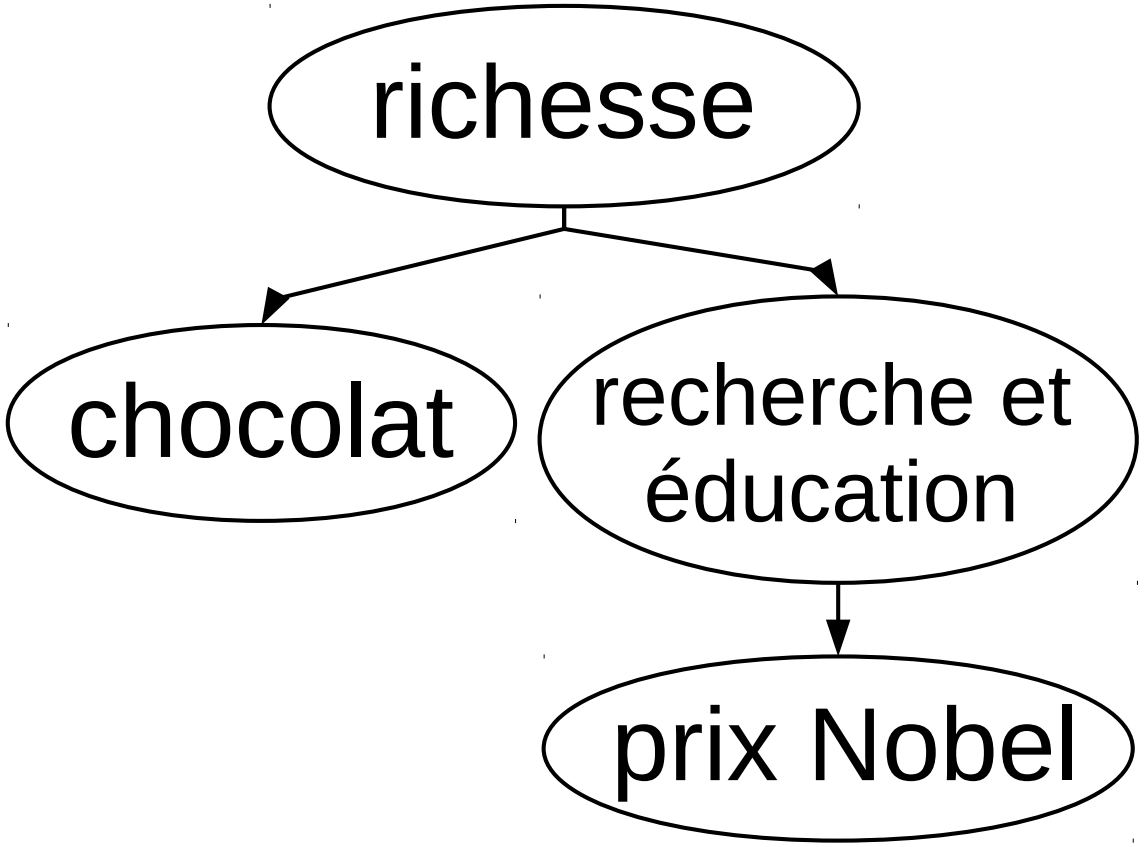
Corrélation n'est pas causalité

cette corrélation peut mener certains à cette hypothèse causale...



Corrélation n'est pas causalité

... mais il existe des alternatives plus probables, par exemple la variation de richesse explique la corrélation :



Corrélation n'est pas causalité

Pearl (2000) :

- *un modèle causal du processus ayant généré les données est une description formelle de comment la valeur de chaque variable observée est déterminée*
  - *modèle statistique dont les variables sont ordonnées via un graphe orienté acyclique*
- *les relations causales sont plus « stables » : elles sont ontologiques, elles décrivent des contraintes physiques objectives de notre monde, alors que des relations probabilistes sont épistémiques, elles reflètent ce que nous connaissons ou croyons à propos du monde*

PNAS

Model-free forecasting outperforms the correct mechanistic model for simulated and experimental data

Charles T. Perretti[a,1], Stephan B. Munch[b], and George Sugihara[a]

[a]Scripps Institution of Oceanography, University of California at San Diego, La Jolla, CA 92093; and [b]Fisheries Ecology Division, Southwest Fisheries Science Center, Santa Cruz, CA 95060

Edited by Peter J. Bickel, University of California, Berkeley, CA, and approved January 10, 2013 (received for review September 14, 2012)

- but : prédire l'abondance d'espèces (dynamique non-linéaire)
- question : est-ce qu'un modèle méchanistique correct prédit mieux qu'une méthode « générique » ?
- réponse (dans ce cas) : non...

Veut-on bien comprendre, mais devrait-on bien prédire ?

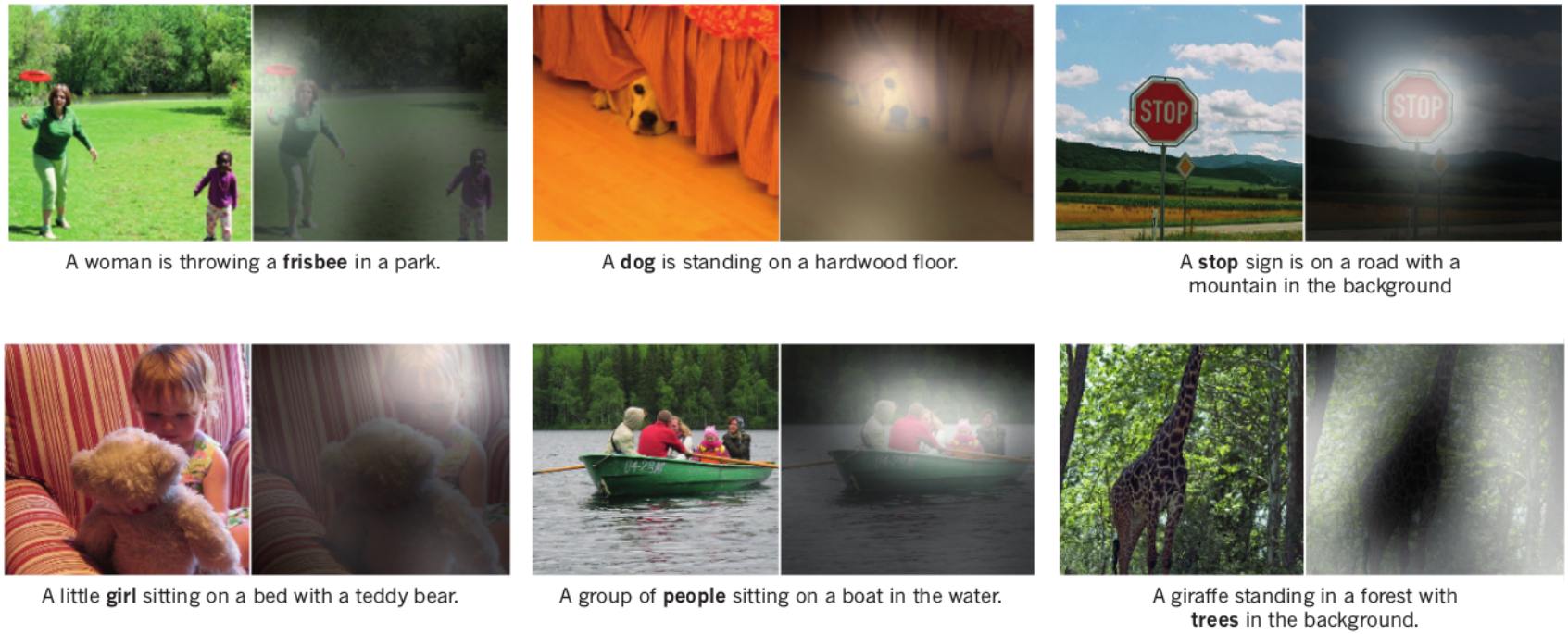
Breiman (2001) :

- *quand un modèle est ajusté à des données pour en tirer des conclusions quantitatives, ces conclusions concernent le mécanisme du modèle, pas celui de la nature*
- *3 leçons : la multiplicité des bons modèles ; le conflit entre simplicité et précision (de prédiction) ; la malédiction de la dimensionnalité*

exemples de « modèles algorithmiques » (semi-, voire non-, paramétriques) : forêts aléatoires, réseaux neuronaux

Cas de l'« apprentissage profond » (*deep learning*)

ex. d'apprentissage supervisé : le réseau de neurones à plusieurs couches est entraîné sur une collection d'images déjà annotées



A woman is throwing a **frisbee** in a park.

A **dog** is standing on a hardwood floor.

A **stop** sign is on a road with a mountain in the background

A little **girl** sitting on a bed with a teddy bear.

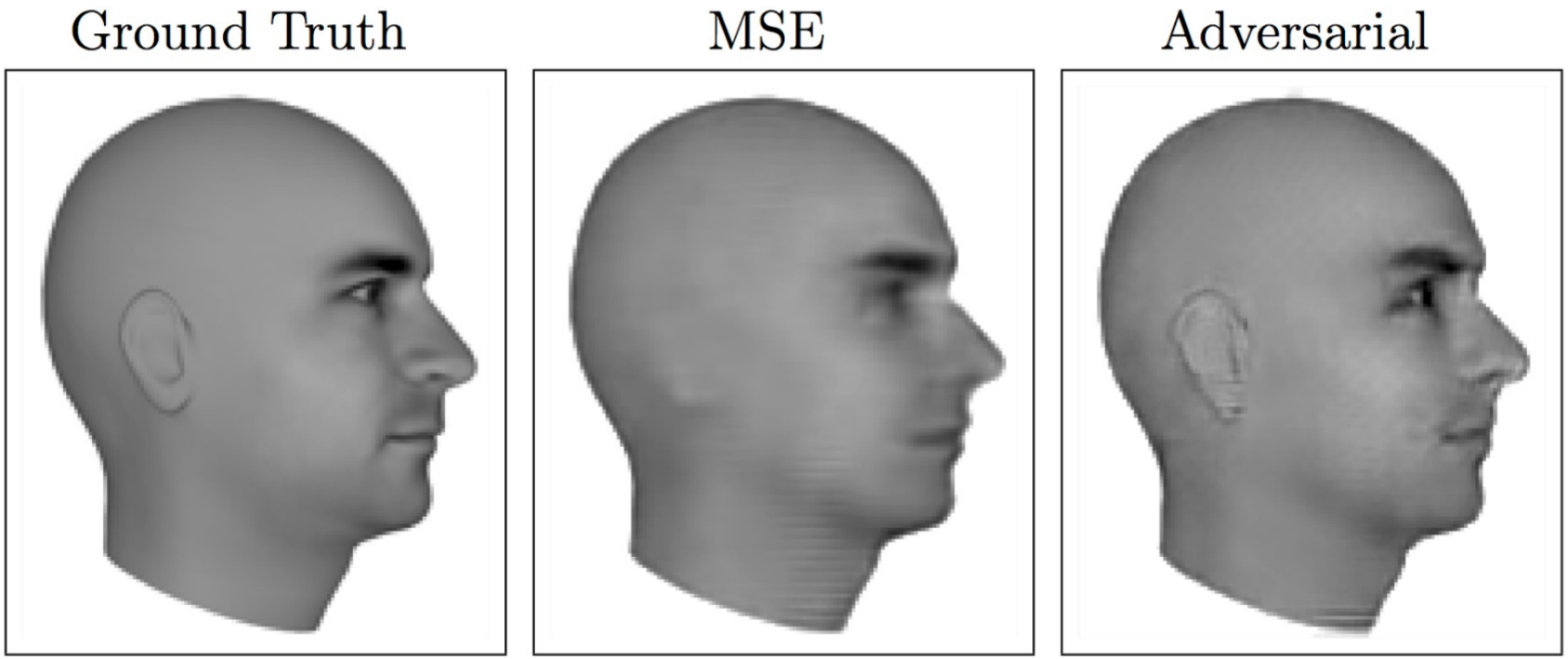
A group of **people** sitting on a boat in the water.

A giraffe standing in a forest with **trees** in the background.

LeCun et coll. (2015)

Cas de l'« apprentissage profond » (*deep learning*)

ex. d'apprentissage non-supervisé via deux réseaux générateurs adversaires



Goodfellow (2016)

Dans quels cas est-ce pertinent ?

- données observationnelles et non expérimentales (peu/pas de planification possible)
- dimension gigantesque des données (très) hétérogènes, donc dimension gigantesque de l'espace des paramètres

Remarque : parfois, on peut écrire une vraisemblance, mais pas la calculer en un temps raisonnable, d'où les méthodes ABC (*approximate Bayesian computation*) beaucoup utilisées en génétique des populations

Une position médiane ?

Gelman et Shalizi (2011) :

- *on construit un modèle à partir de ce qui est diponible, et on le pousse aussi loin qu'il peut nous mener, et même au-delà ; quand il échoue, on le déconstruit, on essaie de comprendre ce qui ne va pas et on bricole avec, ou alors on en essaie un complètement différent*
- *le but de falsifier un modèle n'est pas d'apprendre qu'il est faux, mais plutôt d'apprendre en quoi il est faux*

Remerciements

- Matthew Stephens, pour l'importance de l'intuition statistique

- code du graphique d'hétérogénéité spatiale : Rob Trangucci
- code du graphique du mouvement brownien : `delta9hedge`

Remarque

Ce document a été présenté par T. Flutre le 20 mars 2017 aux agents de l'UMR Agap à l'invitation des organisateurs du LabScience. Il est gratuitement accessible sur le site Prod'INRA répertoriant les productions des agents de l'Inra. En cas de réutilisation de certaines informations issues de ce document, vous êtes tenus de le citer, ainsi que les sources primaires indiquées. N'hésitez pas non plus à contacter l'auteur pour toute précision afin d'éviter les erreurs d'interprétation, ou pour accéder à certaines références.