Introduction aux modèles statistiques et aide à la formalisation d'une question scientique sous la forme d'une question de modélisation

M.L. Delignette-Muller

27 mars, 2025

Un modèle, à quoi ça sert ?

Un modèle, à quoi ça sert ?

Il devient difficile de trouver un article vétérinaire récent basé uniquement sur des analyses statistiques de base (celles enseignées en A2).

La plupart des études modernes utilisent un **modèle statistique**.

Pourquoi ?

A quoi ça sert ?

Ce que vous savez faire à partir des statistiques de base

Test de la corrélation entre deux variables

- qualitative / qualitative : (χ^2, Mc Nemar, Cochran)
- qualitative / quantitative : comparaison de moyennes (Student, Wilcoxon, ANOVA 1)
- quantitative / quantitative : corrélation (Pearson, Spearman)

Les modèles statistiques que vous avez déjà abordés

Modélisation de l'effet d'une variable explicative (qualitative ou quantitative) **sur une variable à expliquer quantitative**

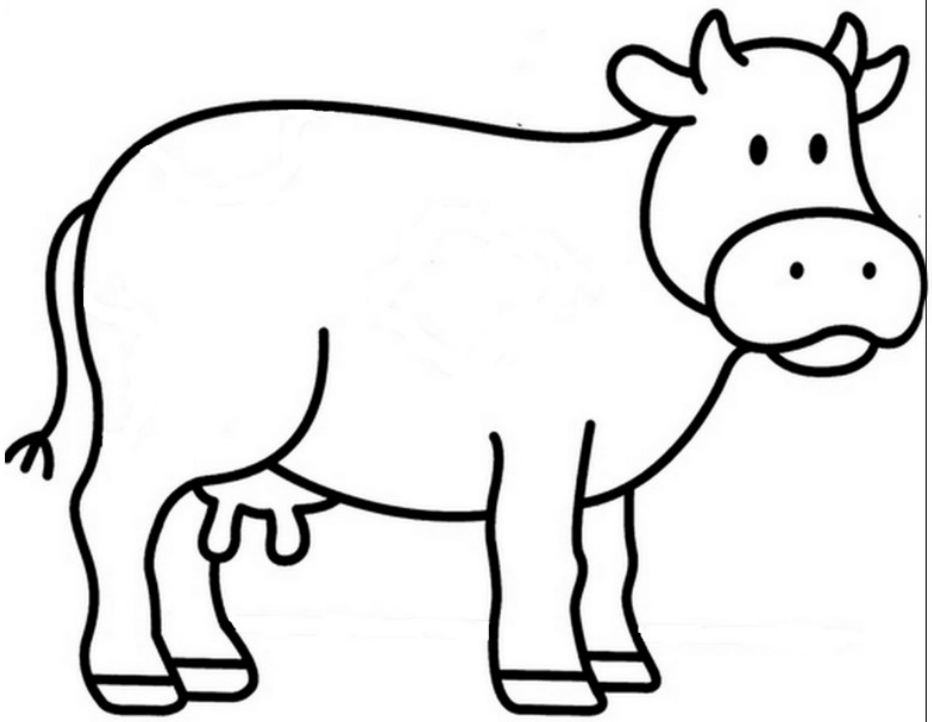
- modélisation de l'effet d'un facteur (variable qualitative) sur une variable quantitative (modèle d'ANOVA 1)
- modélisation de l'effet d'un régresseur (variable quantitative) sur une variable quantitative (modèle de régression linéaire) si la relation est linéaire

Ces méthodes suffisent-elles pour analyser des données si plusieurs facteurs sont susceptibles d'impacter la variable étudiée ?

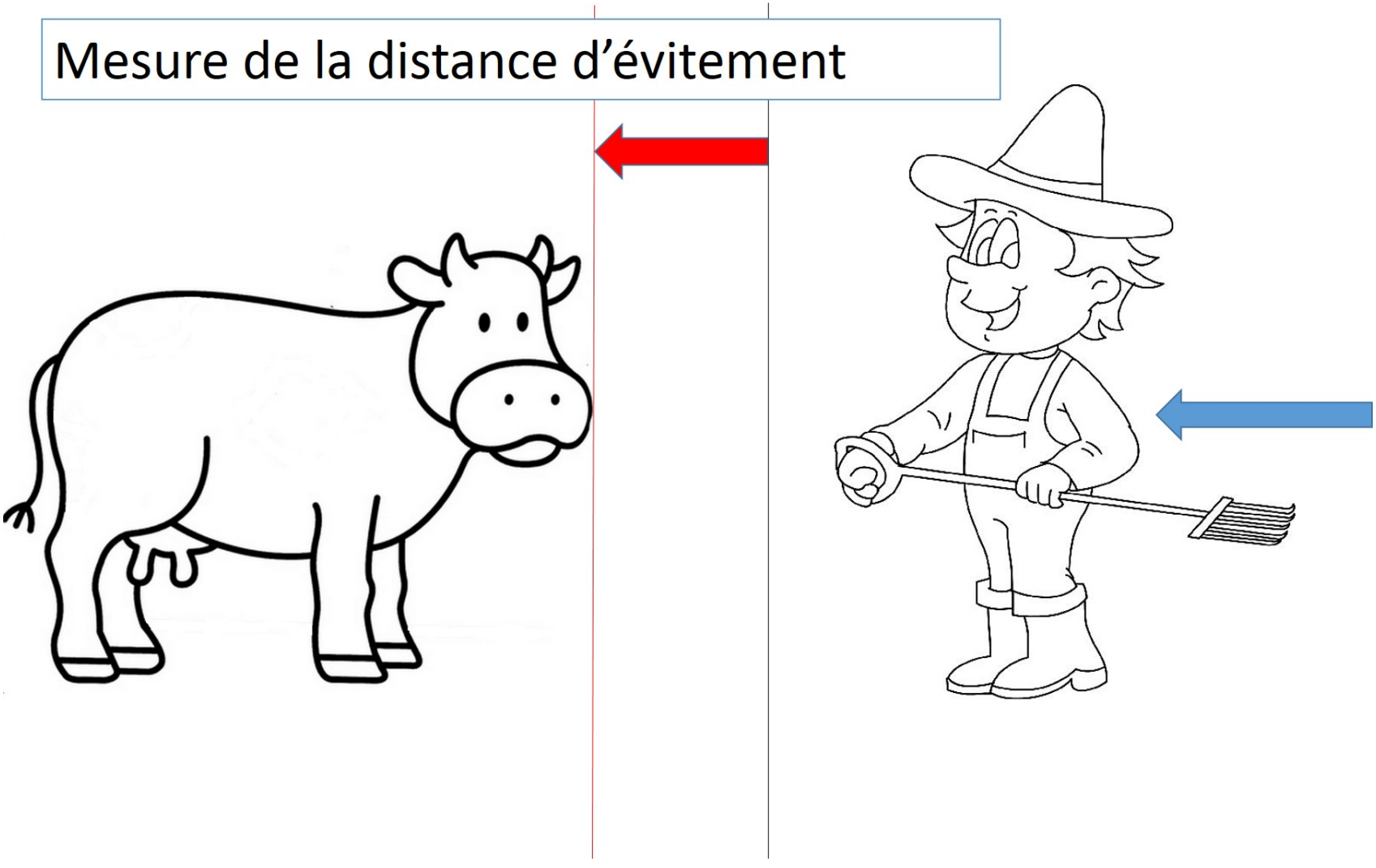
Comment prendre en compte l'effet simultané de plusieurs variables explicatives

NON! Un premier exemple en bien-être animal pour vous en convaincre

Cette vache est-elle bien traitée par son éleveur ?



Etude de la distance d'évitement face à un observateur (un indicateur utilisable)



Etude de ce qui peut impacter la distance d'évitement (variable à expliquer) ?

Liste de variables explicatives potentielles

Variables explicatives qualitatives (facteurs) :

- la race (2 modalités)
- la boiterie (2 modalités)
- le logement (2 modalités)
- le type de traite (2 modalités)
- la parité (2 modalités)

Variables explicatives quantitatives (covariables) :

- la taille du troupeau
- la hauteur relative de l'observateur

Jeu de données sur 2083 vaches

```
## 'data.frame':    2083 obs. of  11 variables:
##  $ distance   : int  110 40 40 20 30 100 0 0 75 0 ...
##  $ Ttroupeau  : int  50 50 50 50 50 50 50 50 50 50 ...
##  $ logement   : Factor w/ 2 levels "Aire_paillee",..: 1
##  $ traite     : Factor w/ 2 levels "Robot","Salle_Traite
##  $ race       : Factor w/ 2 levels "MBT","PH": 1 1 1 1 1
##  $ boiterie   : Factor w/ 2 levels "boiteuse","non_boite
##  $ age        : Factor w/ 2 levels "multipare","primipar
##  $ hauteur    : num  1.3 1.32 1.28 1.3 1.26 1.34 1.39 1.
##  $ elevage    : Factor w/ 101 levels "ACHARD_JOEL",..: 1
##  $ observateur: Factor w/ 5 levels "ALICE","CHRISTOPHE",
##  $ parite     : Factor w/ 2 levels "multipare","primipar
```

Résultats obtenus avec une analyse basique

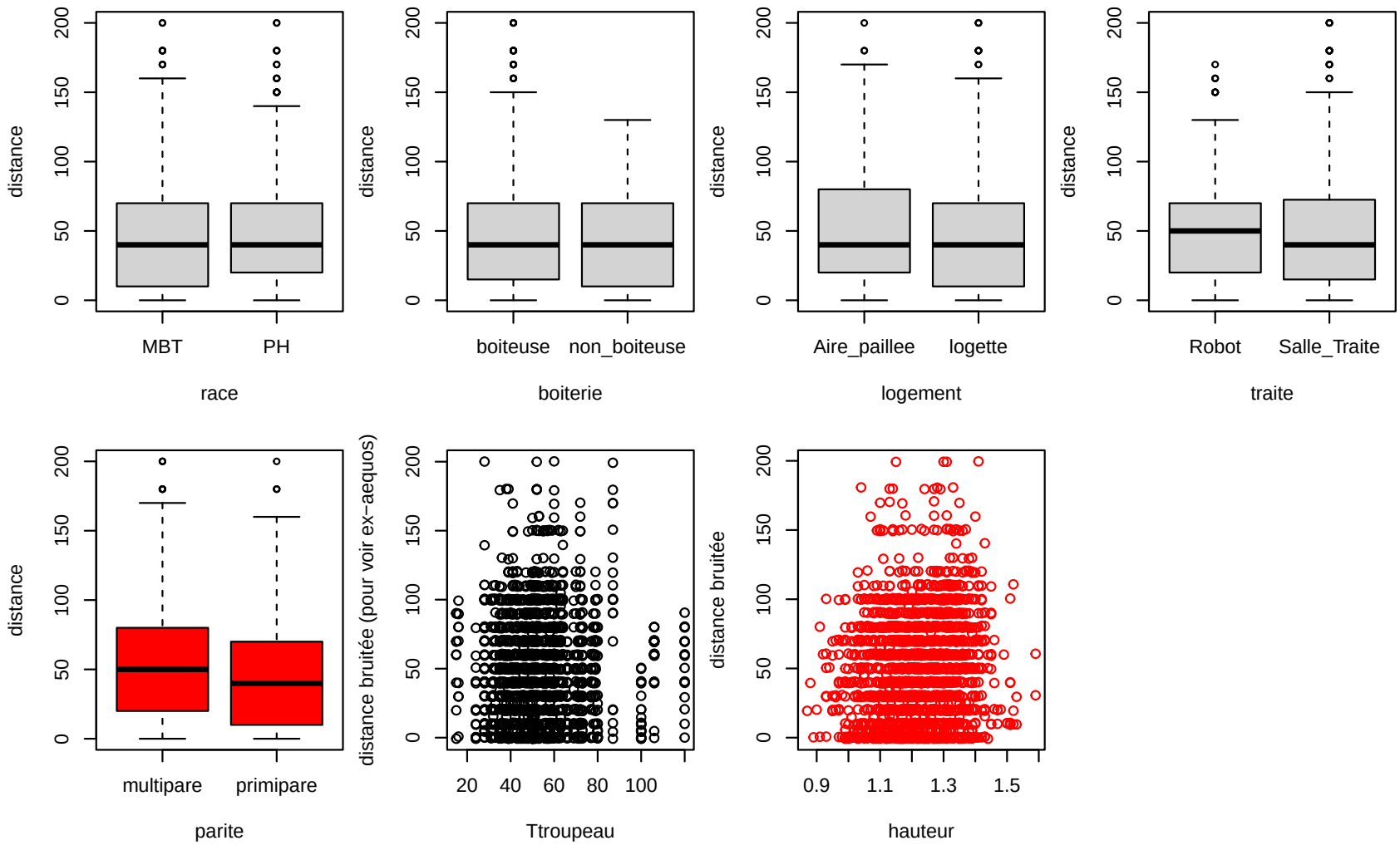
Analyse basique (bivariée) avec les méthodes vues en A2

Comparaison de moyennes pour tester séparément l'effet de chaque facteur et des **tests de corrélation linéaire (+ régression linéaire)** pour tester (modéliser) séparément l'effet de chaque variable quantitative.

Effet significatif

- de la **hauteur relative** de l'observateur : lorsque la hauteur relative prend une unité, les vaches reculent en moyenne de 44 cm de plus (les hauteurs relatives allaient de 0.87 à 1.59 sur le jeu de données observées)
- et de la **parité** (les primipares reculent en moyenne de moins de 6.8 cm que les multipares)

Illustration de l'analyse basique - effets significatifs : age (parité) et hauteur



Résultats obtenus avec un **modèle linéaire mixte**

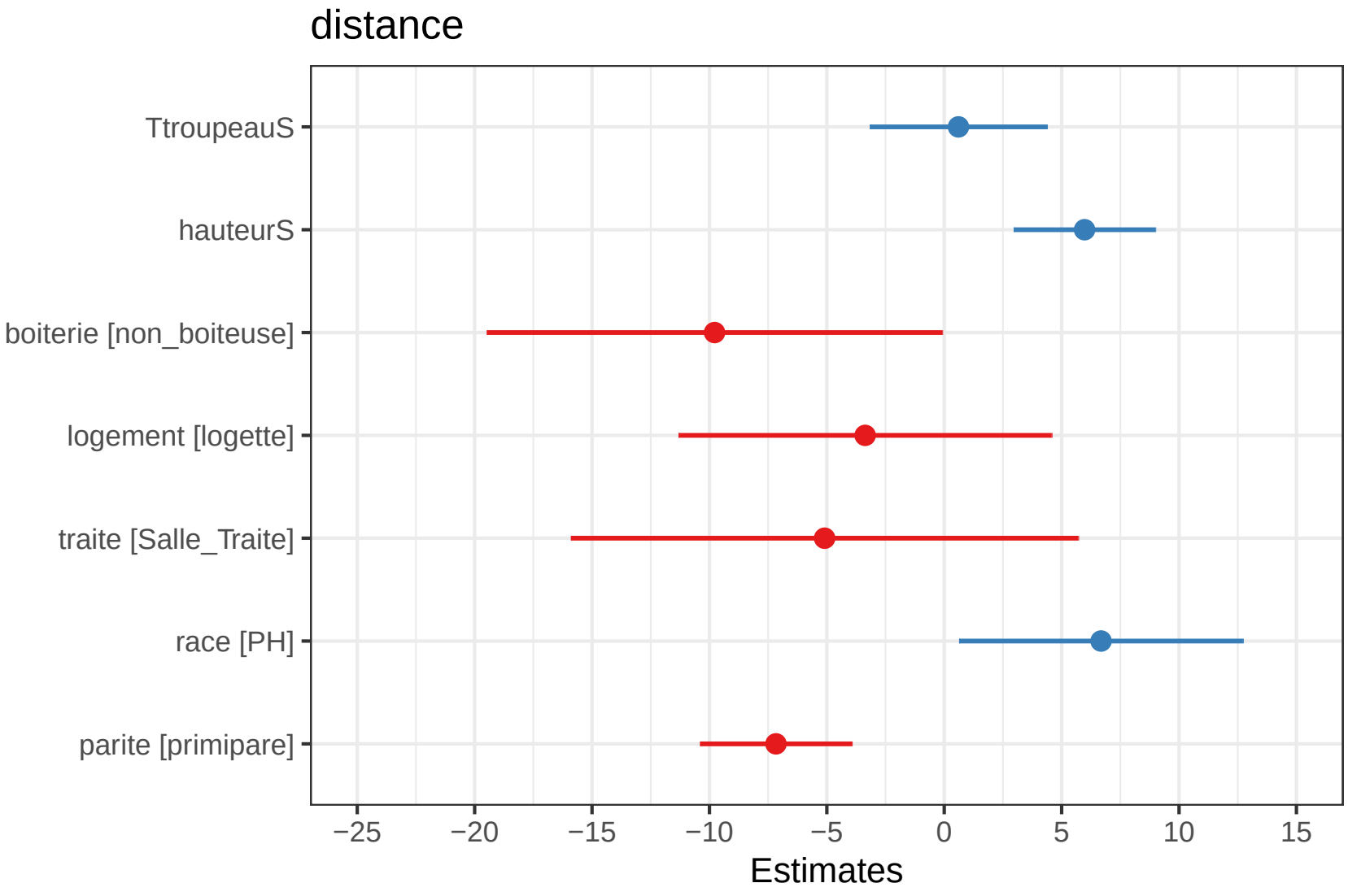
On reviendra plus tard sur la définition d'un modèle mixte

Analyse prenant en compte simultanément l'**ensemble des variables explicatives** (**multivariée**) dans cette étude

Effet significatif

- de la **boiterie** : les vaches boîteuses reculent en moyenne de moins de 9.8 cm que les non boîteuses,
- de la **race** : les vaches de la race 66 reculent en moyenne de plus 6.7 cm que celles de la race 46,
- de la **parité** : les primipares reculent en moyenne de moins de 7.1 cm que les multipares.
- de la **hauteur** : lorsque la hauteur relative prend une unité, les vaches reculent en moyenne de 54 cm de plus (les hauteurs relatives allaient de 0.87 à 1.59 sur le jeu de données observées)

Représentation classique en “forest plot” des effets estimés (significatifs lorsque l’intervalle de confiance de contient pas 0)



Intérêt de l'utilisation d'un modèle dans l'étude présentée

Ici le **modèle a une visée explicative** : mieux comprendre ce qui a un impact sur la distance d'évitement.

Le modèle multivarié proposé prend en compte **simultanément plusieurs variables explicatives** à la fois qualitatives et quantitatives.

Il permet ici de mettre en évidence et de quantifier les effets de variables explicatives qui n'apparaissaient pas significatifs dans une simple analyse où l'effet de chaque variable explicative est testé séparément.

Si vous voulez aller voir de plus près cet exemple dans sa version finale

Article publié en utilisant un modèle mixte sur ces données

Mounier, L., Veissier, I., Rimbaud, J., Boivin, X., Rebout, N., & De Boyer des Roches, A. (2025). Cow factors to address when performing avoidance distance tests at the feeding rack. Animal, the international journal of animal biosciences, 19(4), 101461.

Revenons sur la notion de modèle mixte

Le modèle présenté en amont est dit mixte car il prend en fait aussi en compte la **variabilité liée aux élevages et aux observateurs**.

Variables explicatives qualitatives (facteurs fixes) :

- la race (2 modalités)
- la boiterie (2 modalités)
- le logement (2 modalités)
- le type de traite (2 modalités)
- la parité (2 modalités)

Variables explicatives quantitatives (covariables) :

- la taille du troupeau
- la hauteur relative de l'observateur

Facteurs aléatoires :

- **l'observateur**
- **l'élevage**

Mais qu'appelle-t-on au juste un facteur aléatoire ?

Un **facteur** est considéré comme **aléatoire** si seul un **échantillon aléatoire des modalités possibles** du facteur apparaît dans les données (le cas ici pour l'observateur et l'élevage). Il est suceptible d'augmenter la variabilité de la réponse, mais son effet est **imprévisible** d'une expérimentation à l'autre. On ne peut prédire que l'ampleur de cet effet, quantifié sous la forme d'un écart type (estimé dans un modèle mixte).

Quelques autres exemples de facteurs aléatoires :

- **l'animal** si plusieurs observations sont réalisées sur chaque animal, par exemple au cours du temps, ou sur différents endroits de son corps,
- **le jour d'expérience**, si toutes les données n'ont pas pu être collectées le même jour et que des conditions non contrôlées sont susceptibles de varier d'un jour à l'autre,
- la **clinique vétérinaire** si les données proviennent d'une étude multicentrique.

Revenons sur l'interprétation des coefficients d'un modèle explicatif

Dans un modèle explicatif on teste l'effet de chaque variable en prenant en compte l'effet des autres variables (principe d'ajustement).

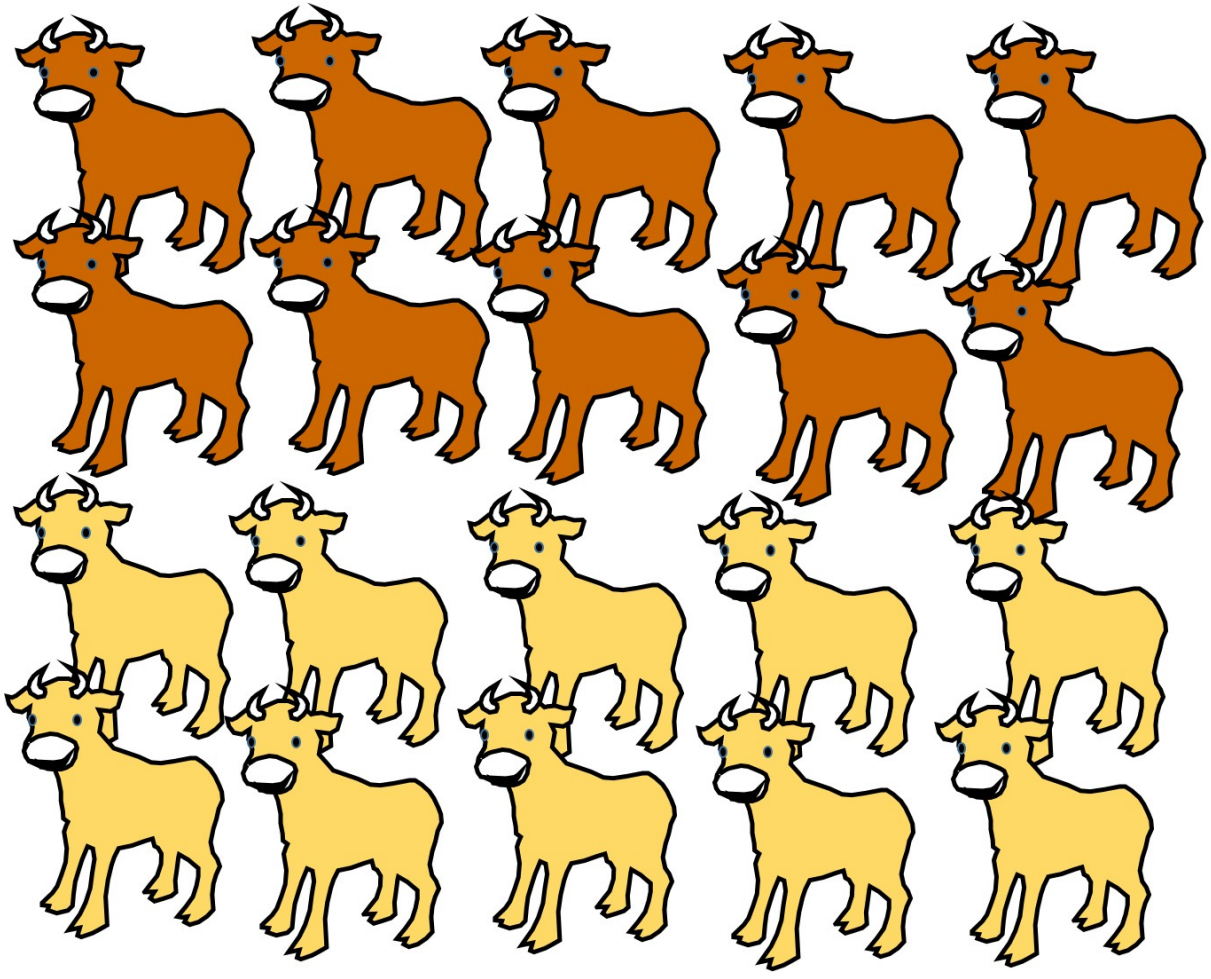
Ex.: ayant déjà pris en compte l'effet de la hauteur, y a-t-il aussi un effet additionnel de la boiterie ?

Comment éviter le biais de confusion, en particulier (mais pas seulement), dans les études observationnelles, dans lesquelles on ne contrôle aucune des variables étudiées.

Comment éviter le biais de confusion, en particulier (mais pas seulement), dans les études observationnelles, dans lesquelles on ne contrôle aucune des variables étudiées.

L'utilisation d'un modèle permet aussi d'éviter le biais de confusion que l'on va illustrer dans un second exemple.

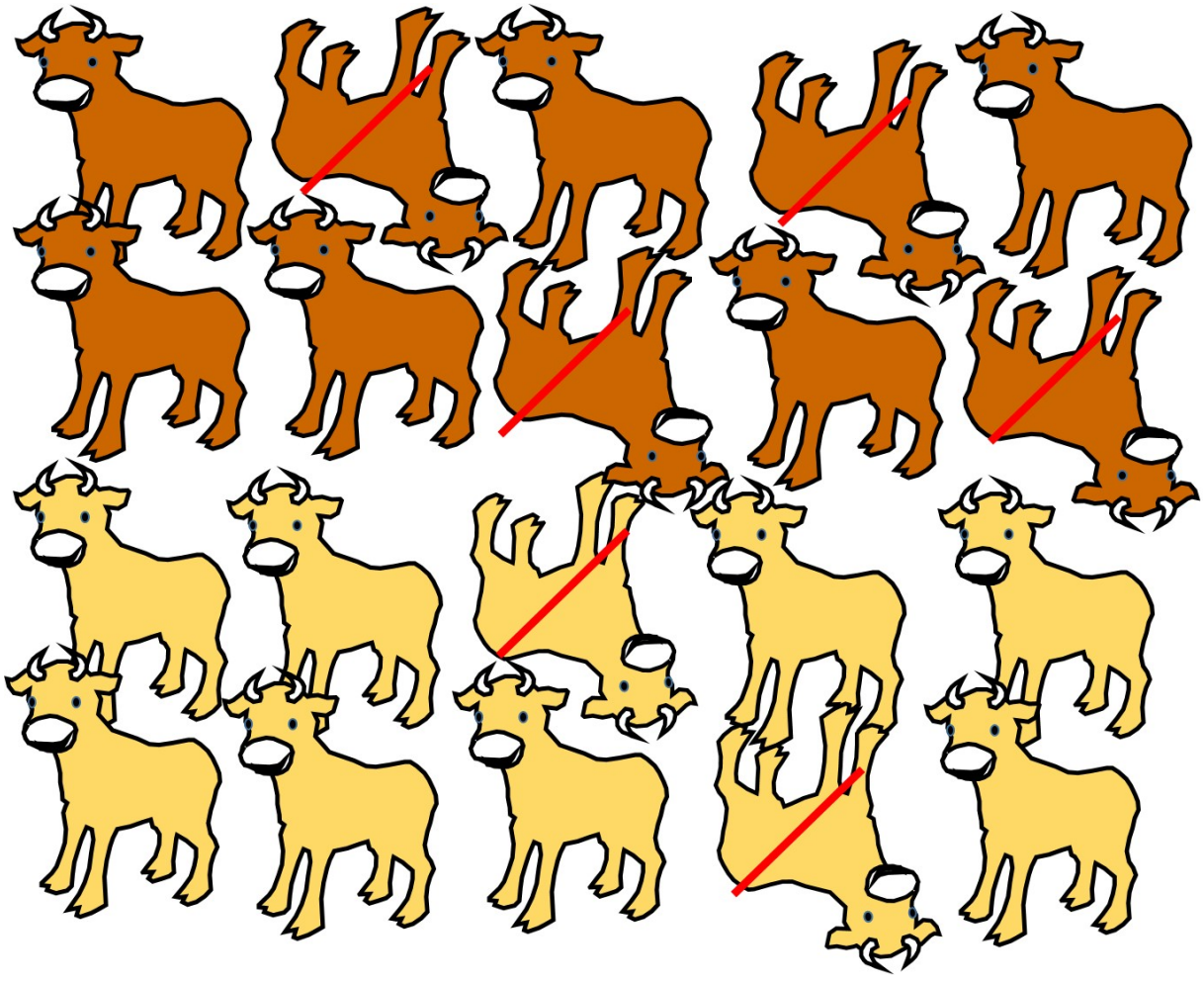
Etude de la survie de veaux atteints de diarrhée



Veaux sans acidose

Veaux avec acidose

Impact positif de l'acidose sur la survie ?

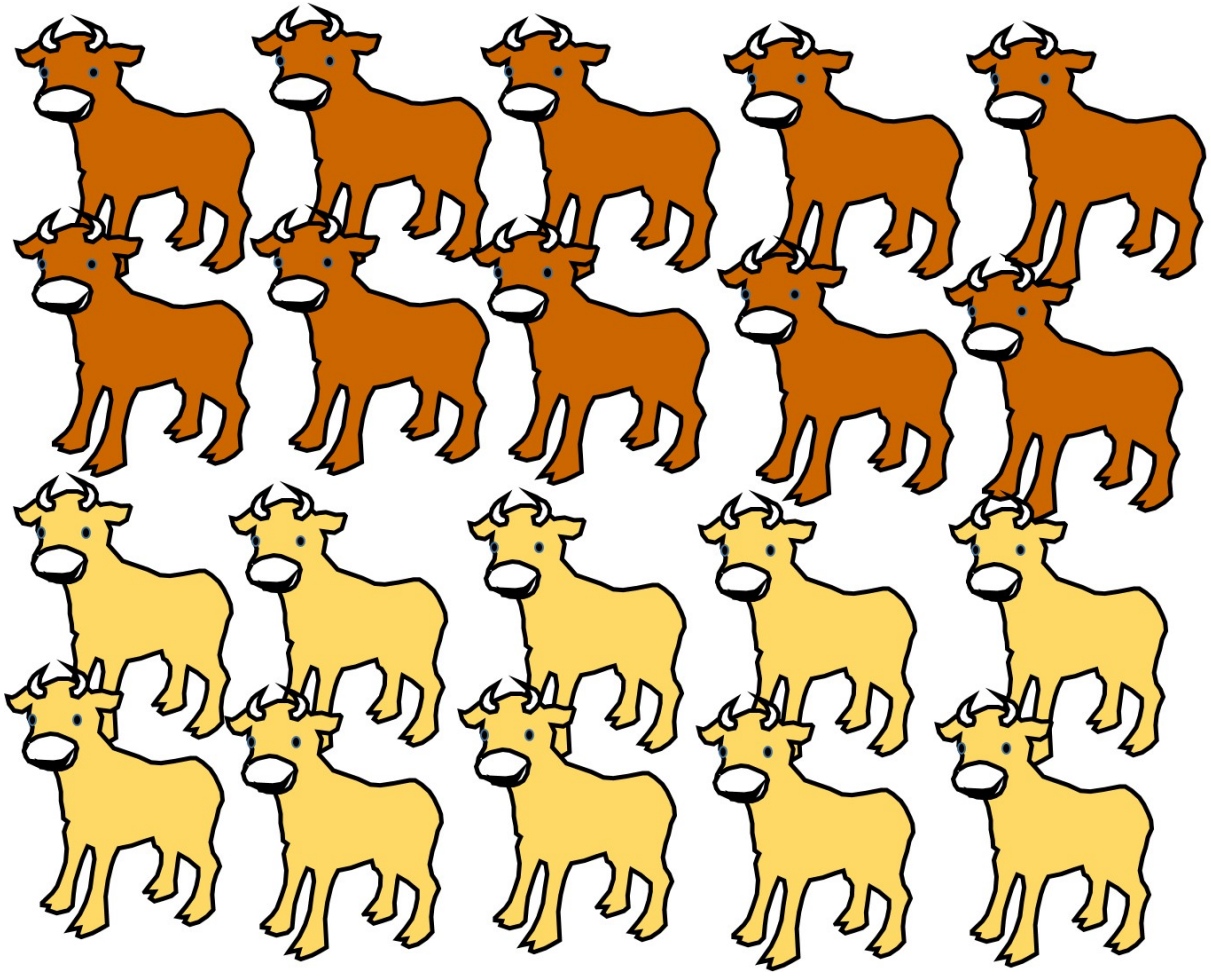


Veaux sans acidose
survie de 60%

Veaux avec acidose
survie de 80%

Meilleure survie en
cas d'acidose ?

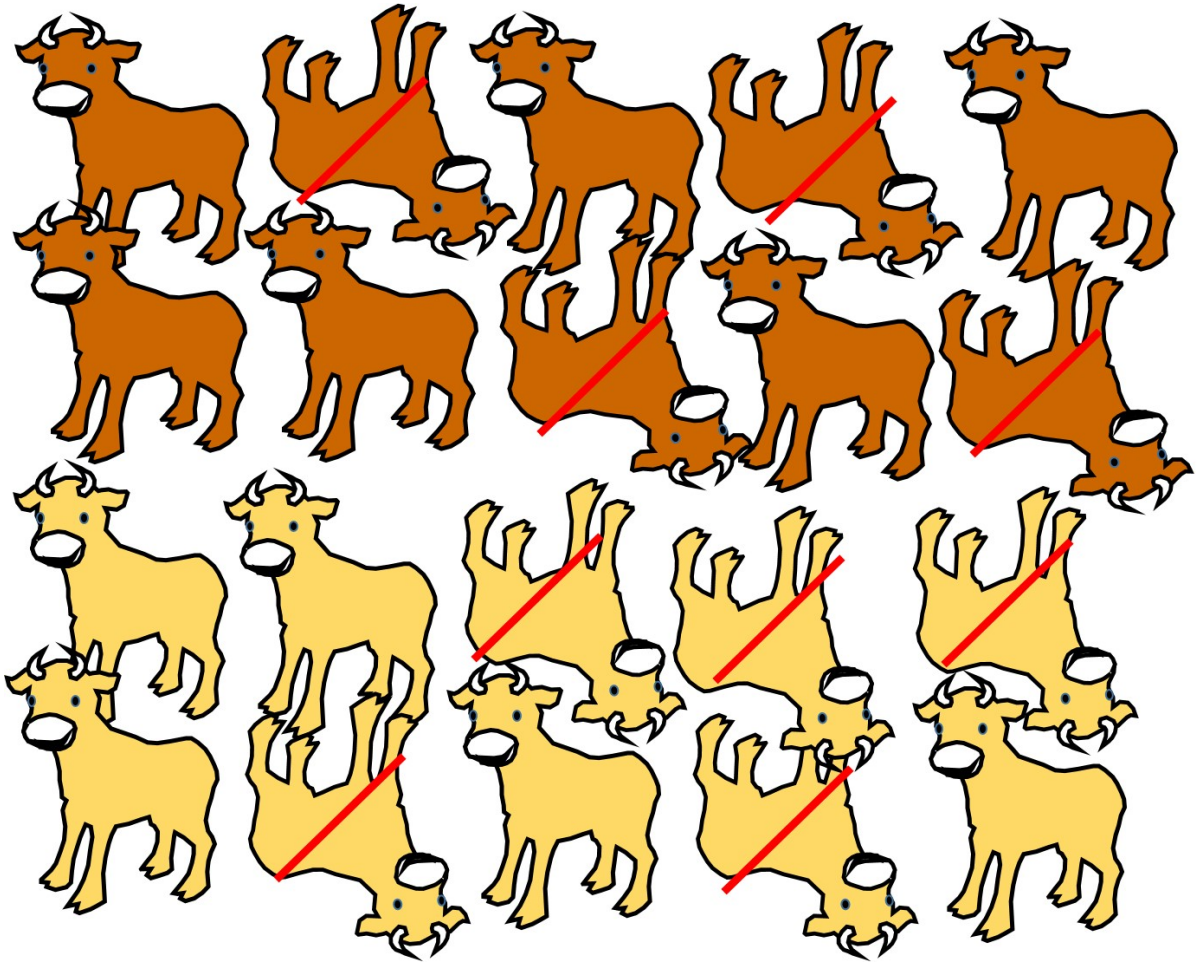
Focalisons-nous sur les veaux de moins de 5 jours ?



Jeunes veaux sans acidose (< 5 jours)

Jeunes veaux avec acidose (< 5 jours)

Chez les très jeunes veaux l'acidose serait plutôt un facteur légèrement aggravant.



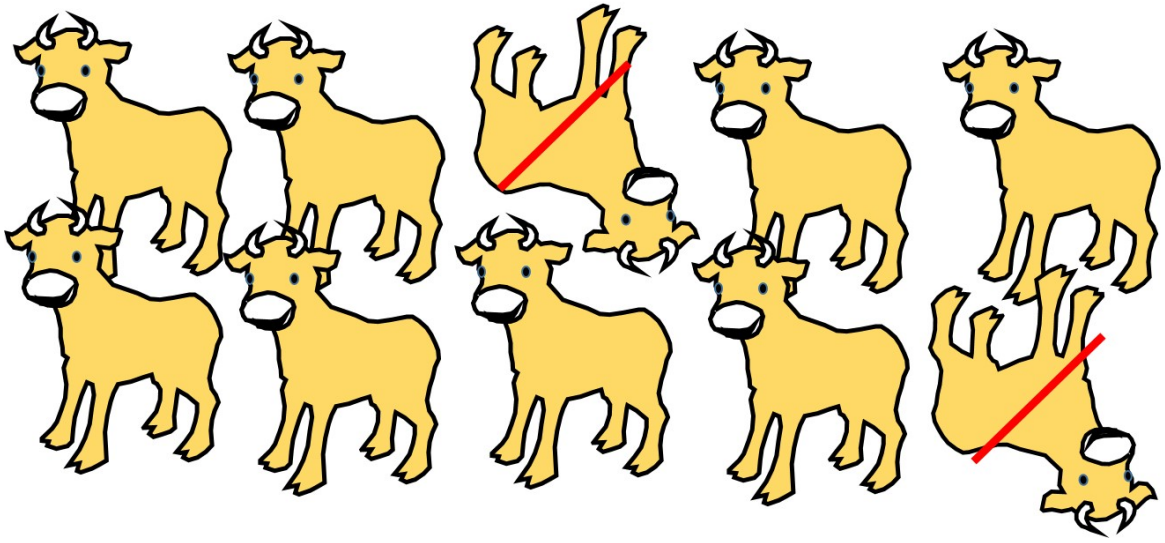
Jeunes veaux sans acidose (< 5 jours) survie de 60%

Jeunes veaux avec acidose (< 5 jours) survie de 50%

Moins bonne survie en cas d'acidose chez les jeunes veaux

Regardons maintenant les veaux de 5 jours ou plus

Veaux de 5 jours ou plus quasi tous avec acidose



Survie des veaux de 5 jours ou plus avec acidose : plus de 80%

Cela ne nous renseigne pas sur un potentiel effet de l'acidose.

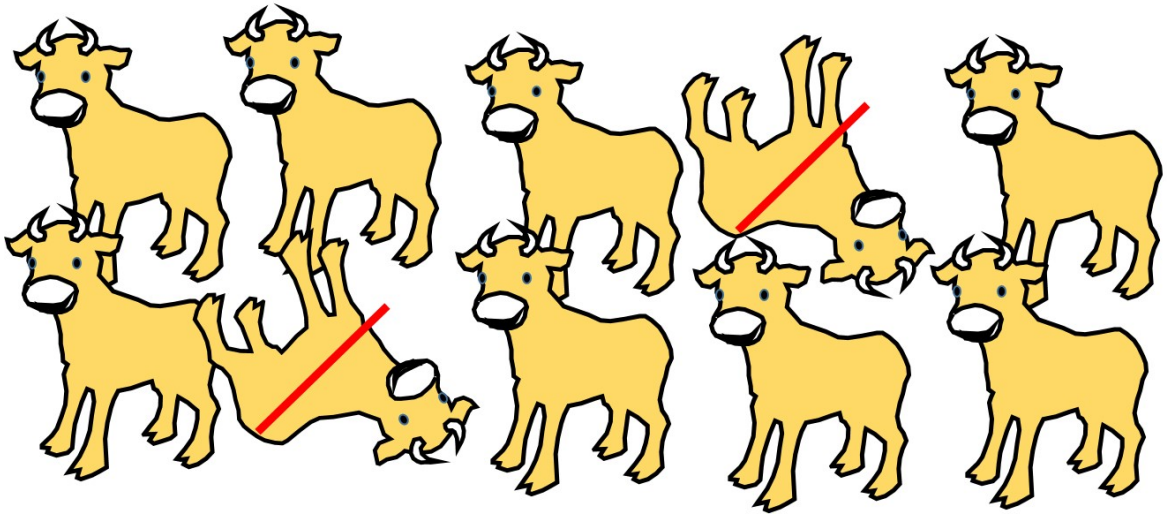
Comment expliquer ces conclusions contradictoires sur l'effet de l'acidose ?

Notion de **facteur de confusion**

Quel est le facteur de confusion qui a un effet sur la survie et qui nous a fait croire à un effet de l'acidose ?

Facteur de confusion = l'âge des veaux

Veaux de moins de 5 jours
survie < 60%



Veaux de 5 jours ou plus
survie > 80%

Jeu de données concernés (77 veaux) - cf. thèse d'Alexis Duthu en 2017

Données issues d'une base de données concernant des veaux en diarrhée pris en charge par des praticiens vétérinaires.

```
## 'data.frame':    77 obs. of  9 variables:
##  $ Resultat : Factor w/ 2 levels "echec","succes": 2 2 1
##  $ scorecalc: int  8 8 11 5 12 7 9 11 12 6 ...
##  $ Age      : int  8 7 15 8 3 7 3 4 4 15 ...
##  $ Temp     : num  39.4 38.4 38 38.4 37.5 39.4 39 37 37
##  $ Deshyd   : int  6 4 0 3 7 0 8 6 8 4 ...
##  $ Be       : int  -14 -13 -27 -13 -9 -21 -3 -16 -11 -19
##  $ Gly      : num  0.93 0.1 0 1.2 0 0.64 0.2 0.69 2.36 0
##  $ acidose  : logi  TRUE TRUE TRUE TRUE TRUE TRUE ...
##  $ age      : Factor w/ 2 levels "moins5jrs","plus5jrs":
```

Pourquoi voyait-on alors un effet de l'acidose dans la première analyse ?

Examen de la répartition des veaux dans les différents groupes d'âge et d'acidose

```
##         age
## acidose moins5jrs plus5jrs
##   FALSE        18         1
##   TRUE         17        41
```

L'effet qu'on aurait pu maladroitement imputer à l'acidose était en fait dû à l'âge (meilleure survie chez les veaux de 5 jours et plus) et à une proportion nettement plus élevée de veaux de plus de 5 jours parmi ceux en acidose.

Analyse des données de cet exemple par régression logistique (modèle linéaire généralisé)

Modèle linéaire généralisé = extension du modèle linéaire classique (Gaussien) à la modélisation d'une variable non gaussienne, ici qualitative à deux modalités (**données binaires**).

variable à expliquer: la survie ou non du veau (= succès du traitement)

variables explicatives:

- l'acidose (variable qualitative à deux modalités)
- l'âge (variable qualitative à deux modalités : moins de 5 ans ou 5 ans et plus)

Seul l'effet de l'âge apparaît significatif.

Construction d'un modèle à visée prognostique prenant en compte d'autres variables explicatives

On construit ici un **modèle à visée prédictive** (aide au praticien vétérinaire dans sa prise en charge d'un veau en diarrhée)

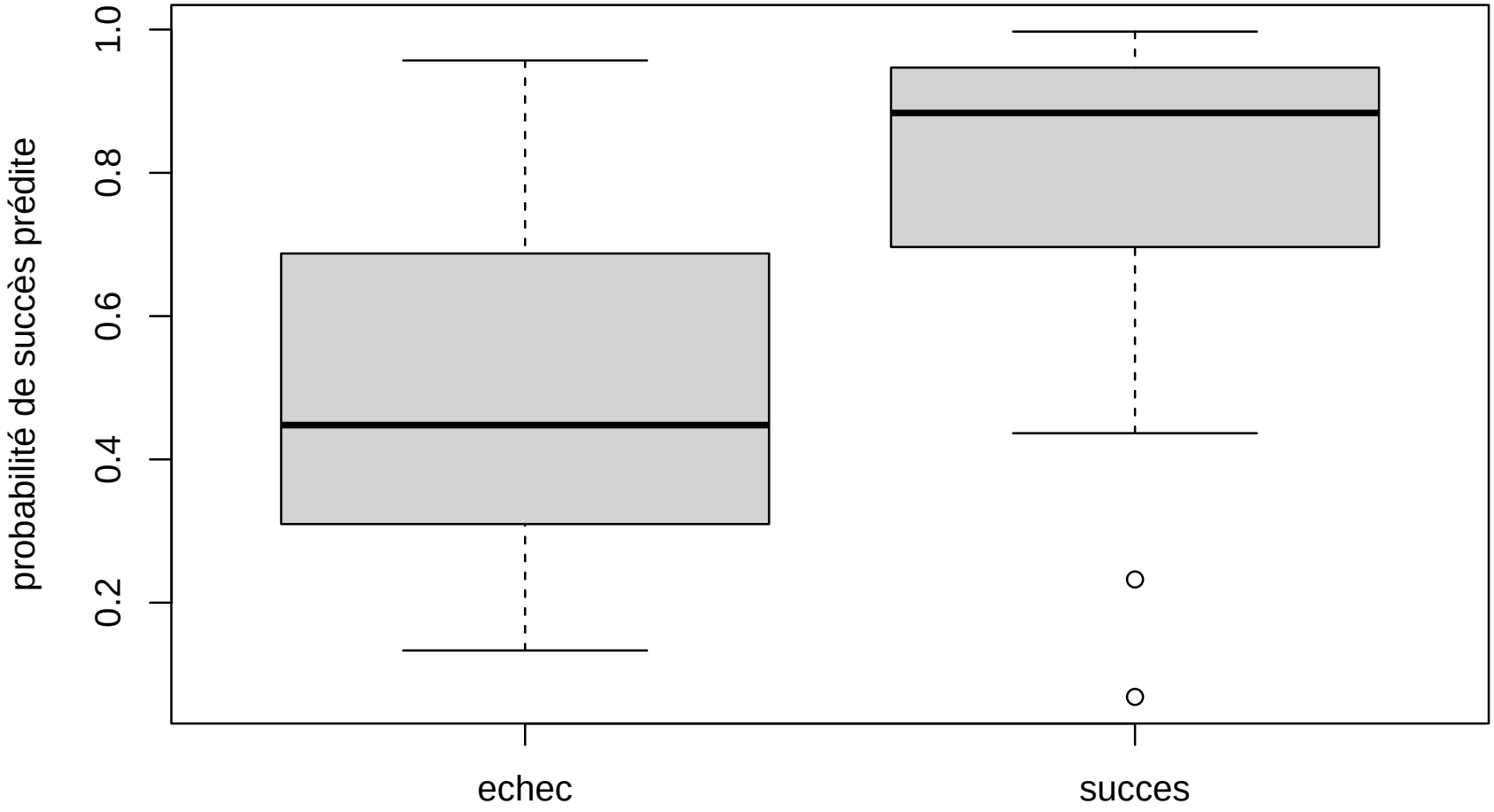
Variable à expliquer: la survie (= succès) ou non (= échec) du veau

Variables explicatives (ici toutes quantitatives):

- l'excès de base
- l'âge
- la température
- la déshydratation
- la glycémie
- le score clinique

Trois variables explicatives retenues dans le modèle prédictif : **l'âge** (corrélé positivement au succès), le **score clinique** (corrélé négativement au succès) et la **glycémie** (corrélée positivement au succès)

Distribution de la variable prédite par le modèle (probabilité de succès) dans les groupes succès (survie) et échec (mort)



Modèle pas parfait mais fournissant tout de même une bonne information pronostique (Alexis avait ensuite amélioré son modèle en prenant en compte des effets non linéaires)!

A retenir de ces exemples

Il n'est pas judicieux d'analyser des données multivariées à l'aide de méthodes de base ne permettant de modéliser que l'effet d'une variable explicative à la fois sur la variable étudiée.

Il est d'autant plus dangereux de le faire si le plan d'expérience n'est pas équilibré, ou si certains facteurs ne sont pas contrôlés (cas des données d'observation rarement équilibrées).

Nécessité de vous former à des méthodes plus complexes si vous devez prendre en compte l'effet simultané de plusieurs variables / facteurs sur la variable étudiée.

Offre de formation en modélisation statistique à VetAgro Sup

Offre de formation en modélisation statistique à VetAgro Sup

Changement à partir de l'année universitaire 2025-26

- Plus de modules de biostatistique d'approfondissement en janvier de la A6
- Un nouvel**EP de A5** de **modélisation statistique** (ouverts aussi à d'autres publics, doctorants, enseignants-chercheurs, ...) qui se déroulera sur des **jeudis après-midi du premier semestre**.

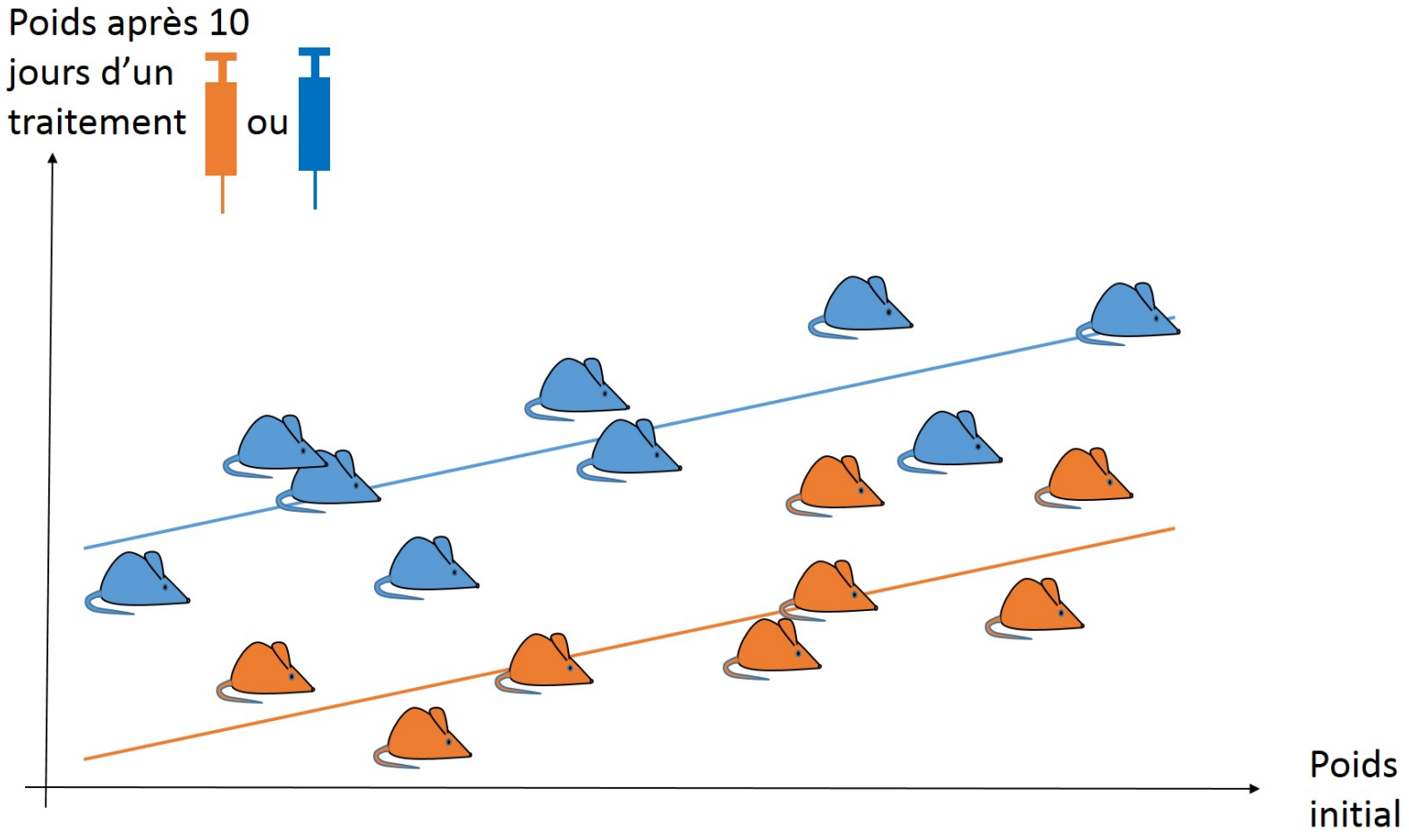
Programme de l'EP de A5

La fiche de cet EP est en cours de construction pour validation par la commission en mai, et sera diffusée dès sa validation.

- modélisation de l'effet d'une ou plusieurs variables qualitatives et/ou quantitatives sur une variable quantitative (**modèle linéaire**)
- modélisation de l'effet d'une ou plusieurs variables qualitatives et/ou quantitatives sur une variable qualitative (**régression logistique**)
- prise en compte de facteurs aléatoires (animal, groupe d'animaux) dans les modèles précédents (**modèle mixte**)
- Utilisation des modèles pour l'analyses des mesures répétées dans le temps (**données longitudinales**)
- analyse de courbes de survie (**données de survie**)

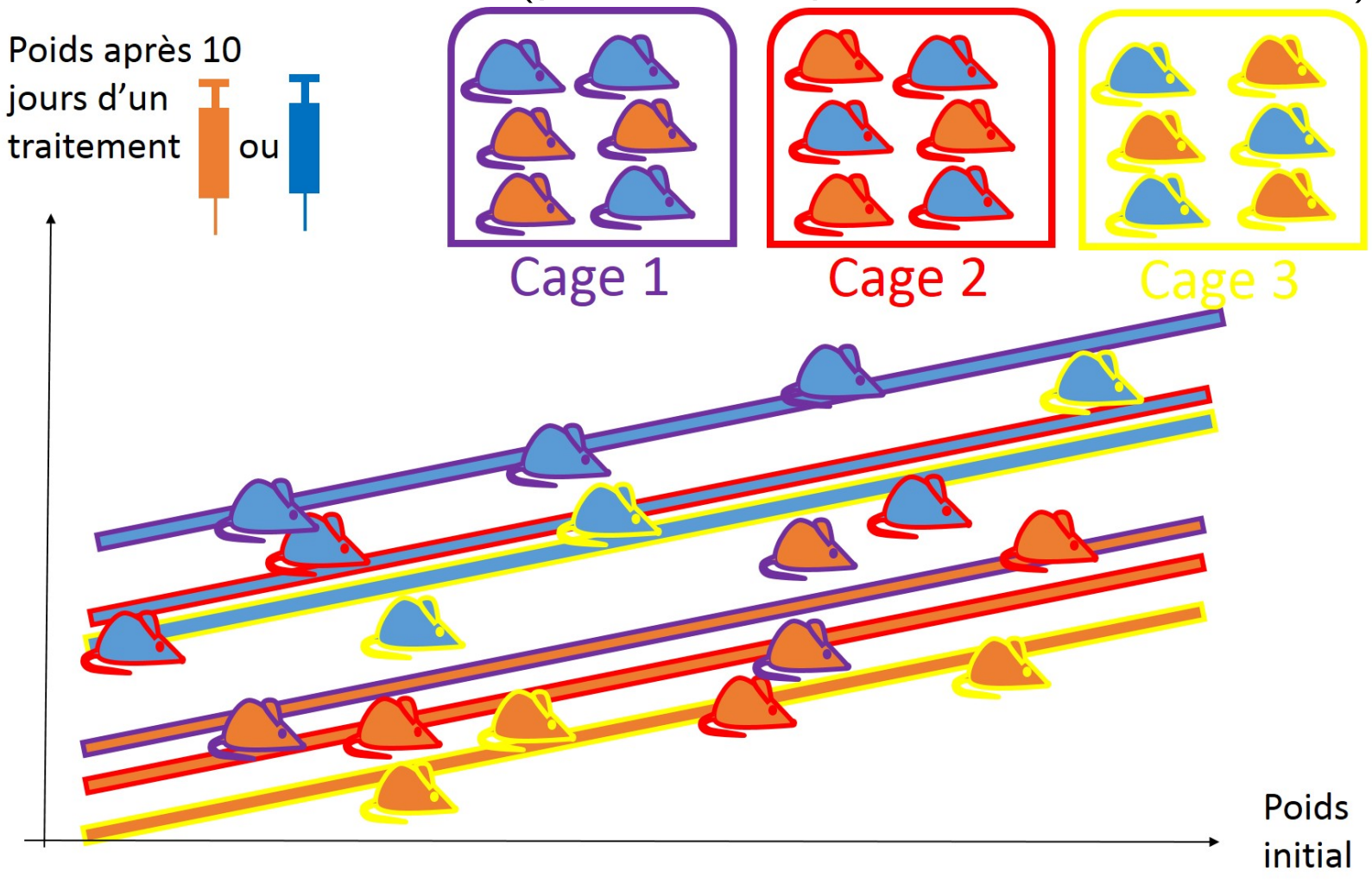
Module modèle linéaire

Modèle linéaire

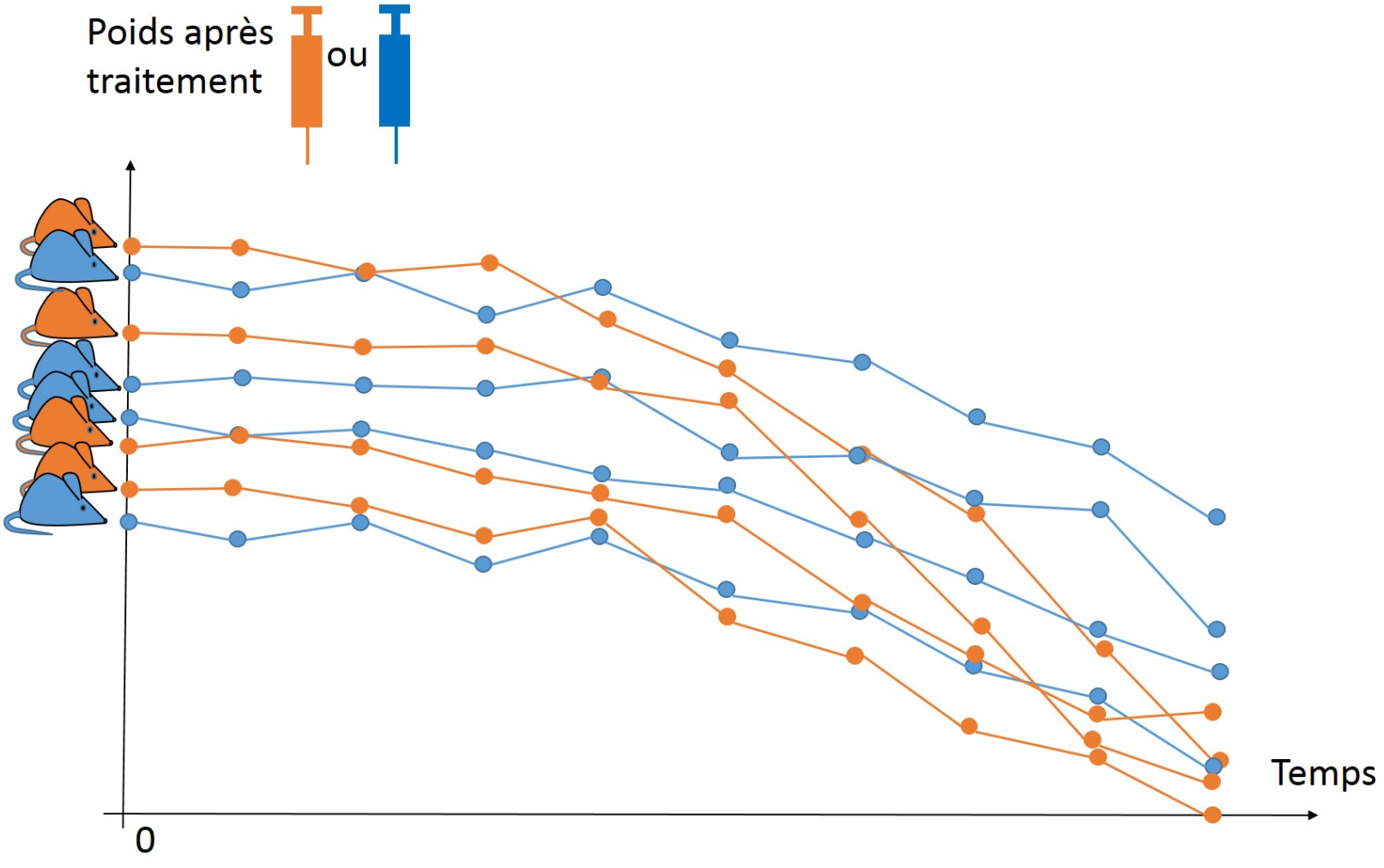


Modèle linéaire linéaire mixte

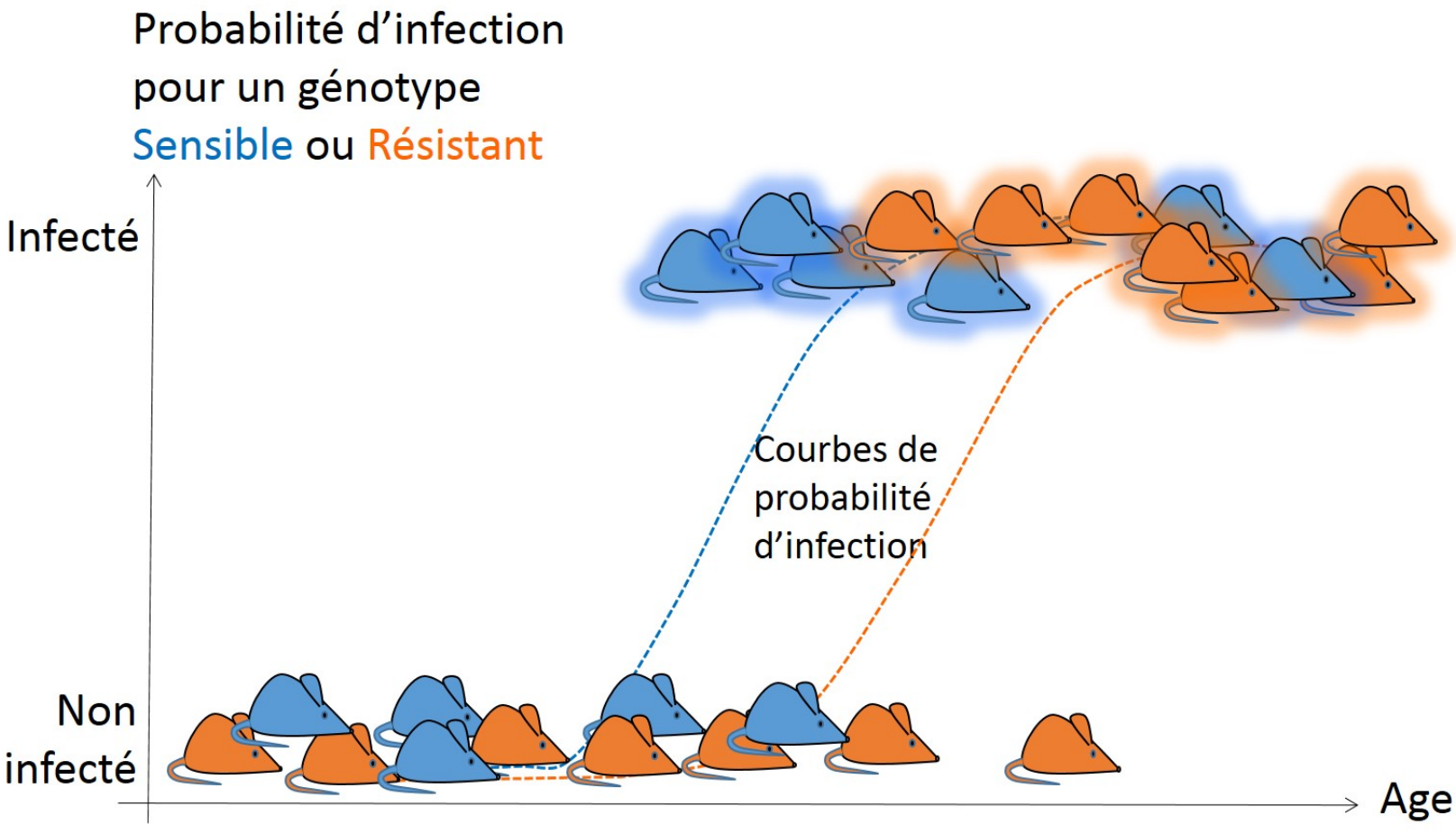
Modèle linéaire mixte (prise en compte de facteurs aléatoires)



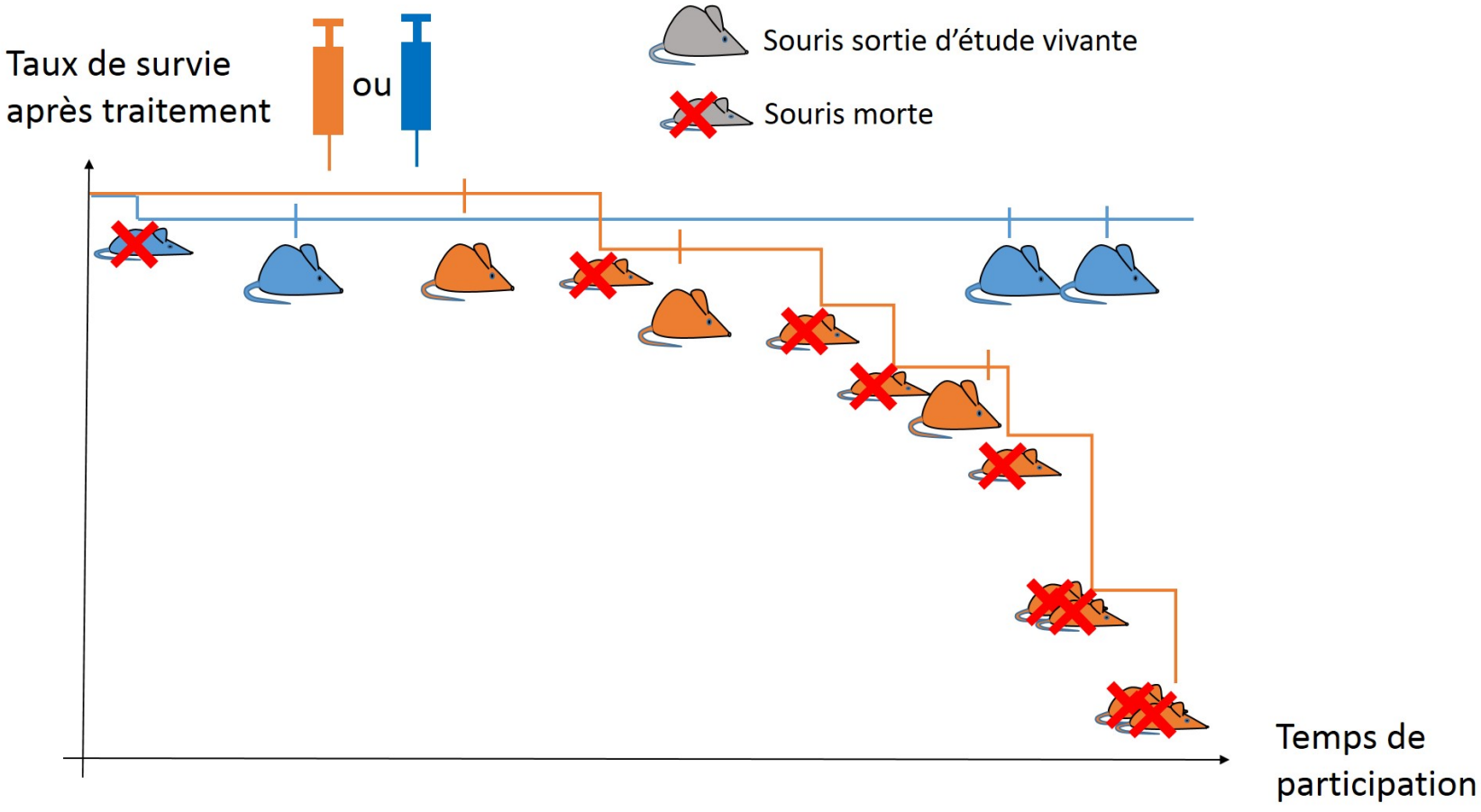
Modèle linéaire et modèle linéaire mixte appliqué au cas très classique des **données longitudinales**



Données binaires et régression logistique



Données de survie et modèles associés (modèle de Cox et modèles paramétriques)



Identification de l'outil statistique adapté à votre problématique de thèse

Identification de l'outil statistique adapté à votre problématique de thèse

Comment savoir si j'aurai besoin d'un modèle, et si oui duquel ?

Nous allons maintenant nous exercer à **formaliser une question scientifique (PICO)**,

puis à la traduire en **question de modélisation**,

afin d'identifier **quel type de modèle** serait utile pour y répondre.

Liste de points à aborder pour définir la question de modélisation

1. Bien identifier dans quel objectif concrêt / pratique vous construisez un modèle (modèle **explicatif** ou **prédictif** ?) et sur quelle population (cible) vous envisagez de l'utiliser (**P**ICO)
2. Bien identifier les critères de jugement, ou plus généralement la **variable à expliquer** (PIC**O**), et les **variables explicatives** dont l'effet sur la variable à expliquer vous intéresse (**PI**CO)
3. Identifier aussi toutes les **variables concomittantes**, dont l'effet vous intéresse moins, mais qu'il est important de prendre en compte pour eviter les confusions d'effet. Parmi ces variables concomittantes identifier les éventuels **facteurs aléatoires** à prendre en compte
4. Préciser la nature de la variable à expliquer et des variables explicatives (quantitative / qualitative, combien de modalités si qualitative)

Exerçons-nous sur des problématiques abordées dans des thèses vétérinaires en cours ou passées

- Outil : la **fiche d'aide à la formalisation d'une question de modélisation** et son **lexique** associé
- L'exercice qui vous est proposé à partir de cette fiche : formaliser la question scientifique des exemples que je vais vous proposer (par groupes de 3 étudiants, max 4)

Vous serez évalué par groupe sur la formalisation du dernier exemple proposé.