Quelques notions à prendre en compte lorsque plusieurs facteurs sont susceptibles d'impacter la variable étudiée : notion de facteurs de **confusion**, notion d'**interaction** entre facteurs. De l'intéret des méthodes utilisant un modèle linéaire

M.L. Delignette-Muller

14 novembre, 2018

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)

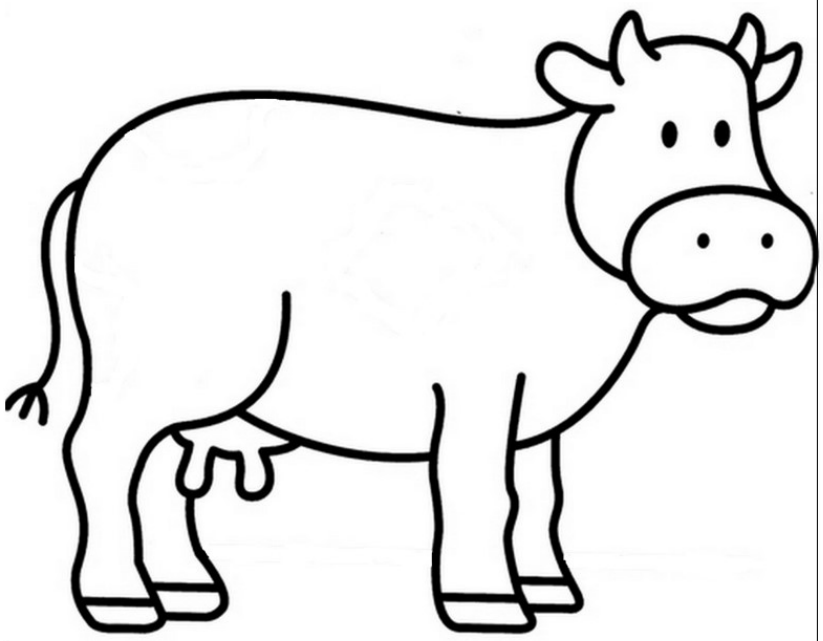
Modélisation de l'effet d'une variable (qualitative ou quantitative) **sur une variable quantitative**

- modélisation de l'effet d'un facteur (variable qualitative) sur une variable quantitative (modèle 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 ?

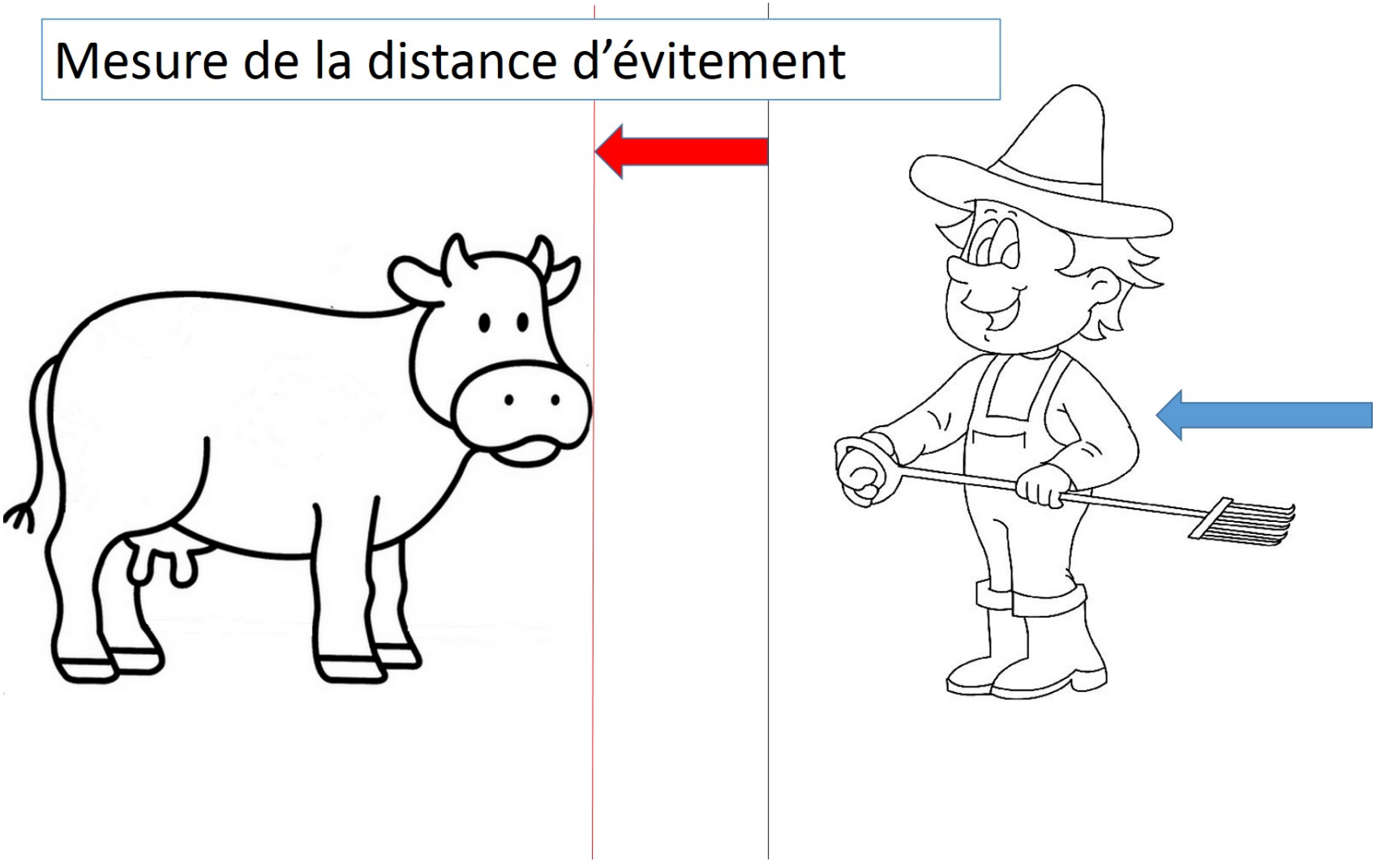
NON! Un premier exemple en bien-être animal pour vous en convaincre

Cette vache est-elle bien traitée par son éleveur ?



Etude de la distance d'évitement face à un observateur

Mesure de la distance d'évitement



Impact de diverses variables sur la distance d'évitement ?

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

Résultats obtenus avec une analyse basique

Analyse basique avec les méthodes vues en S6

Comparaison de moyennes pour tester séparément l'effet de chaque facteur et des tests de corrélation pour tester séparément l'effet de chaque covariable

Effet significatif

- de la **hauteur relative** de l'observateur (plus l'observateur est haut plus la vache recule)
- et de la **parité** (les primipares reculent moins)

Résultats obtenus avec un **modèle linéaire mixte**

Analyse prenant en compte simultanément l'ensemble des variables explicatives ainsi que la variabilité liée aux élevages et aux observateurs (facteurs aléatoires) dans cette étude

Effet significatif

- de la **boiterie** : les vaches boîteuses reculent en moyenne de moins de 9.9 cm que les non boîteuses,
- de la **race** : les vaches de la race 66 reculent en moyenne de plus 6.8 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 52 cm de plus (les hauteurs relatives allaient de 0.87 à 1.59 sur le jeu de données observées)

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

Modèle prenant en compte **simultanément plusieurs variables explicatives** à la fois qualitatives et quantiatives ainsi que l'**effet potentiel de variables aléatoires** (élevage, observateur)

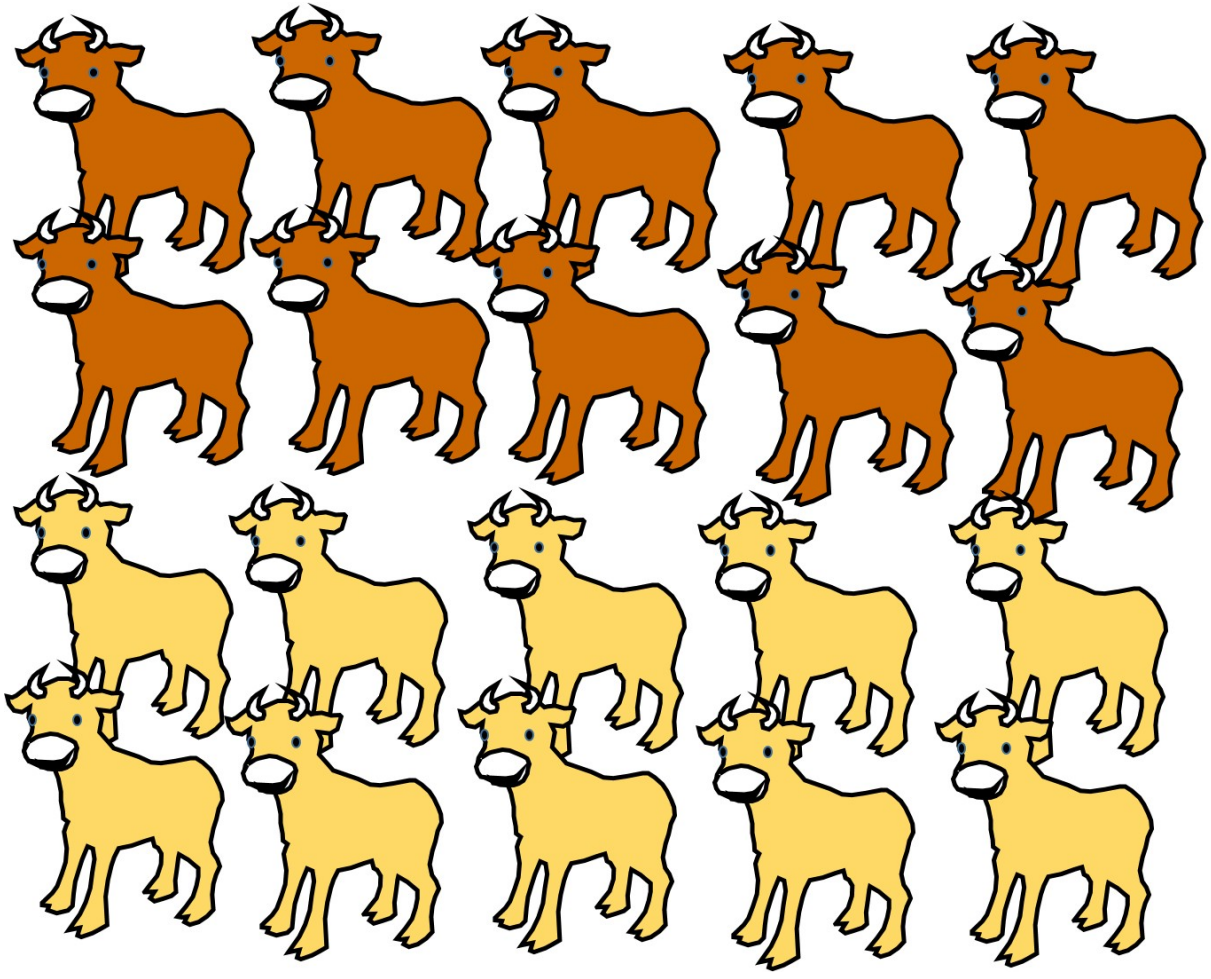
Permet ici de mettre en évidence et de quantifier les effets de variables explicatives qui n'apparaissaient pas significatifs dans une simple analyse où l'effet de chaque variable explicative est testé séparément.

Dans un modèle 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 ?

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

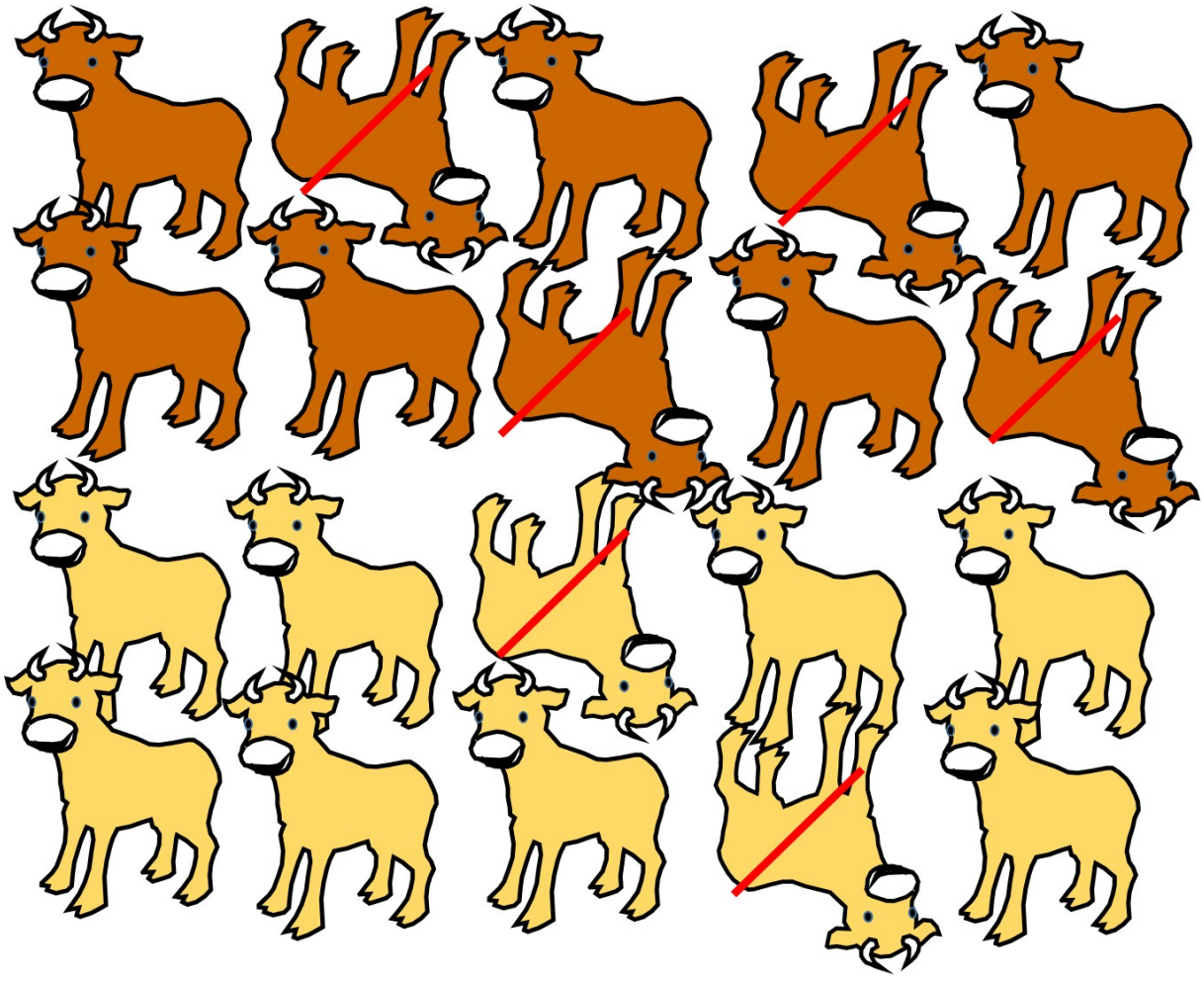
Etude de la survie de veaux atteints de diarrhée



Veaux sans acidose

Veaux avec acidose

Impact positif de l'acidose sur la survie ?

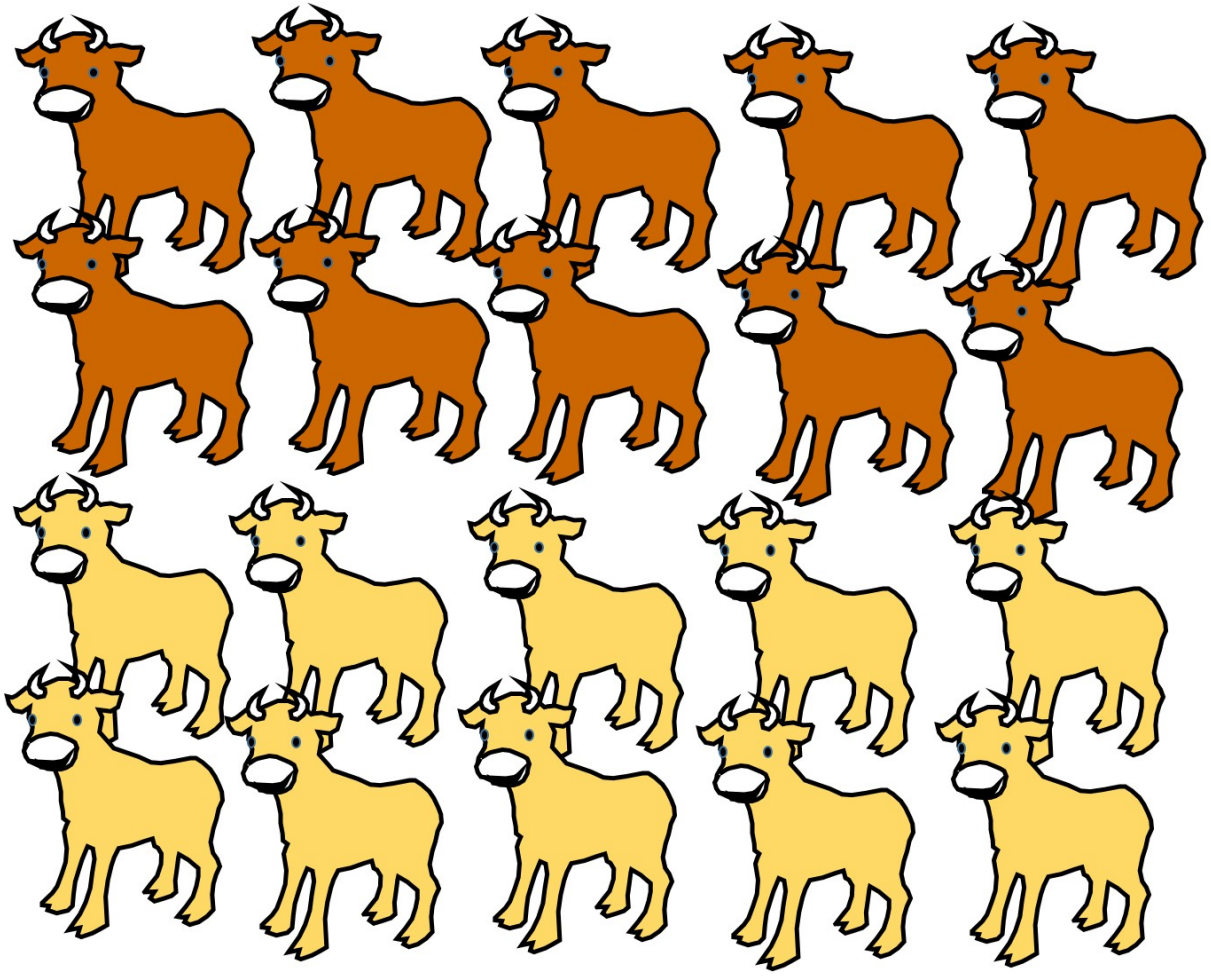


Veaux sans acidose
survie de 60%

Veaux avec acidose
survie de 80%

Meilleure survie en
cas d'acidose ?

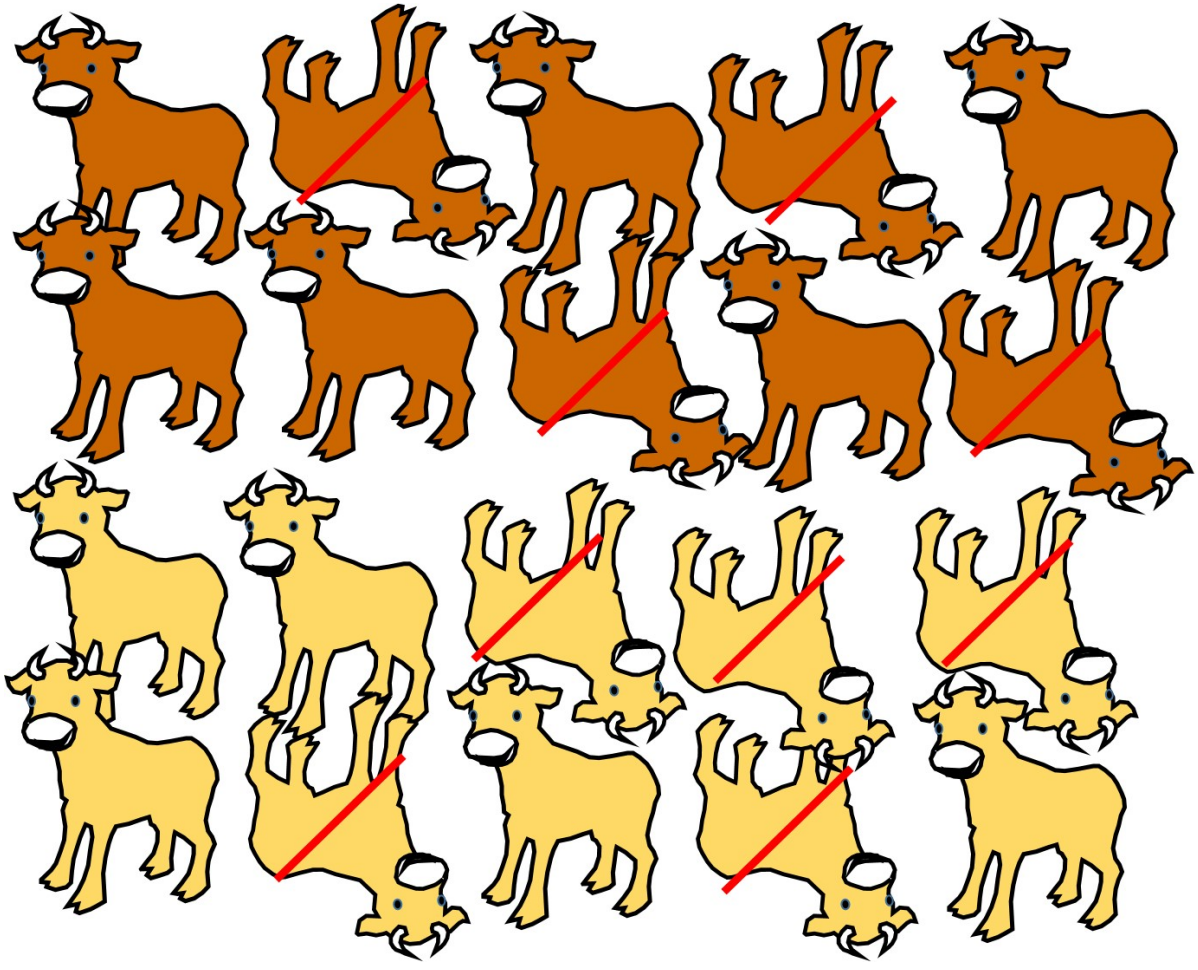
Focalisons-nous sur les veaux de moins de 5 jours ?



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.



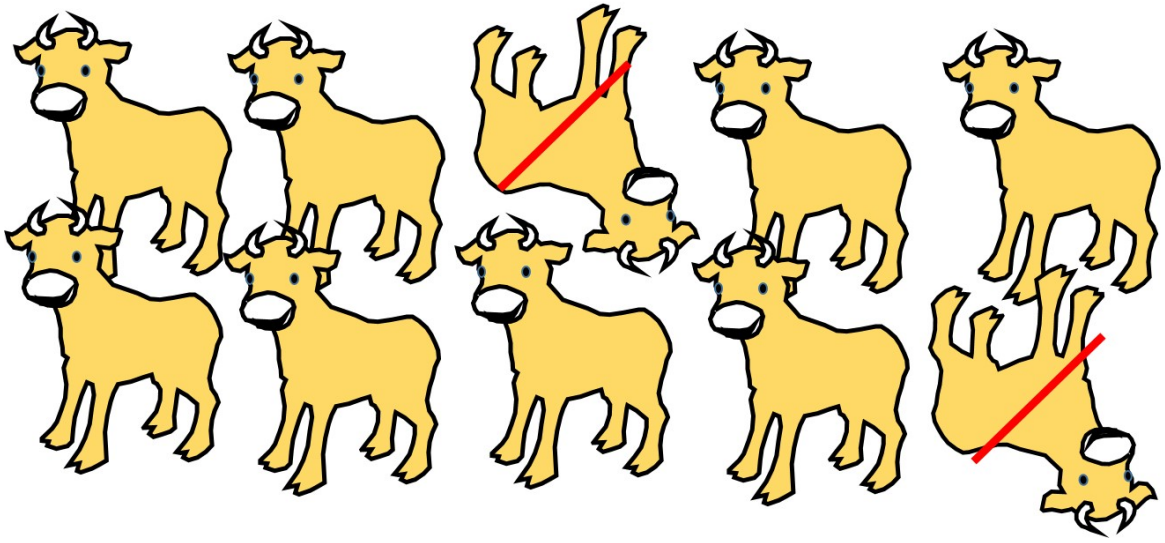
Jeunes veaux sans acidose (< 5 jours) survie de 60%

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.

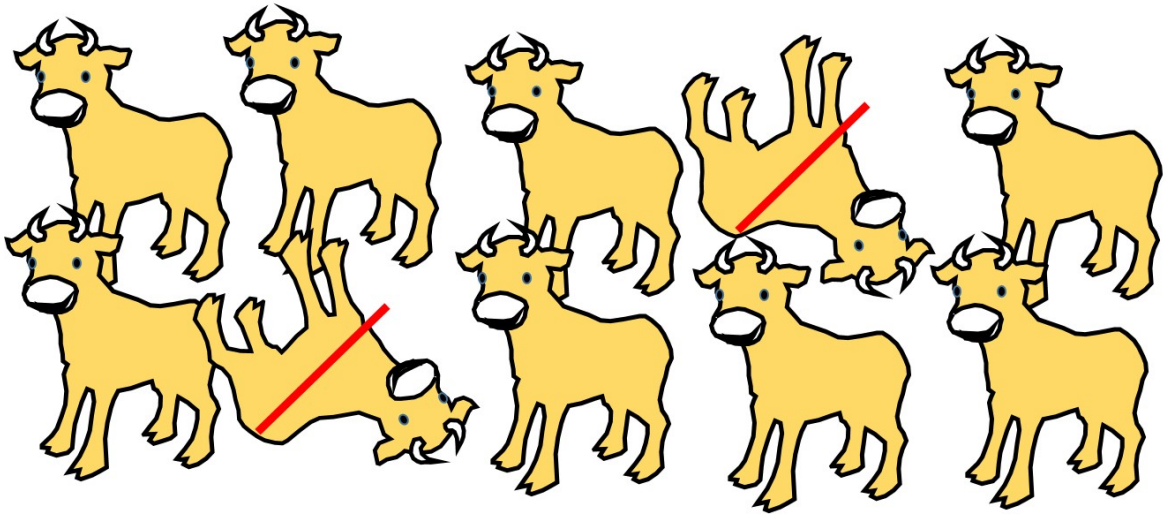
Comment expliquer ces conclusions contradictoires sur l'effet de l'acidose ?

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%

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

Examen du plan d'expérience

```
d <- read.table("DATA/diarreeveau.txt", header = TRUE)
xtabs(~ acidose + age, data = d)
```

```
##         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 prognostic prenant en compte d'autres variables explicatives

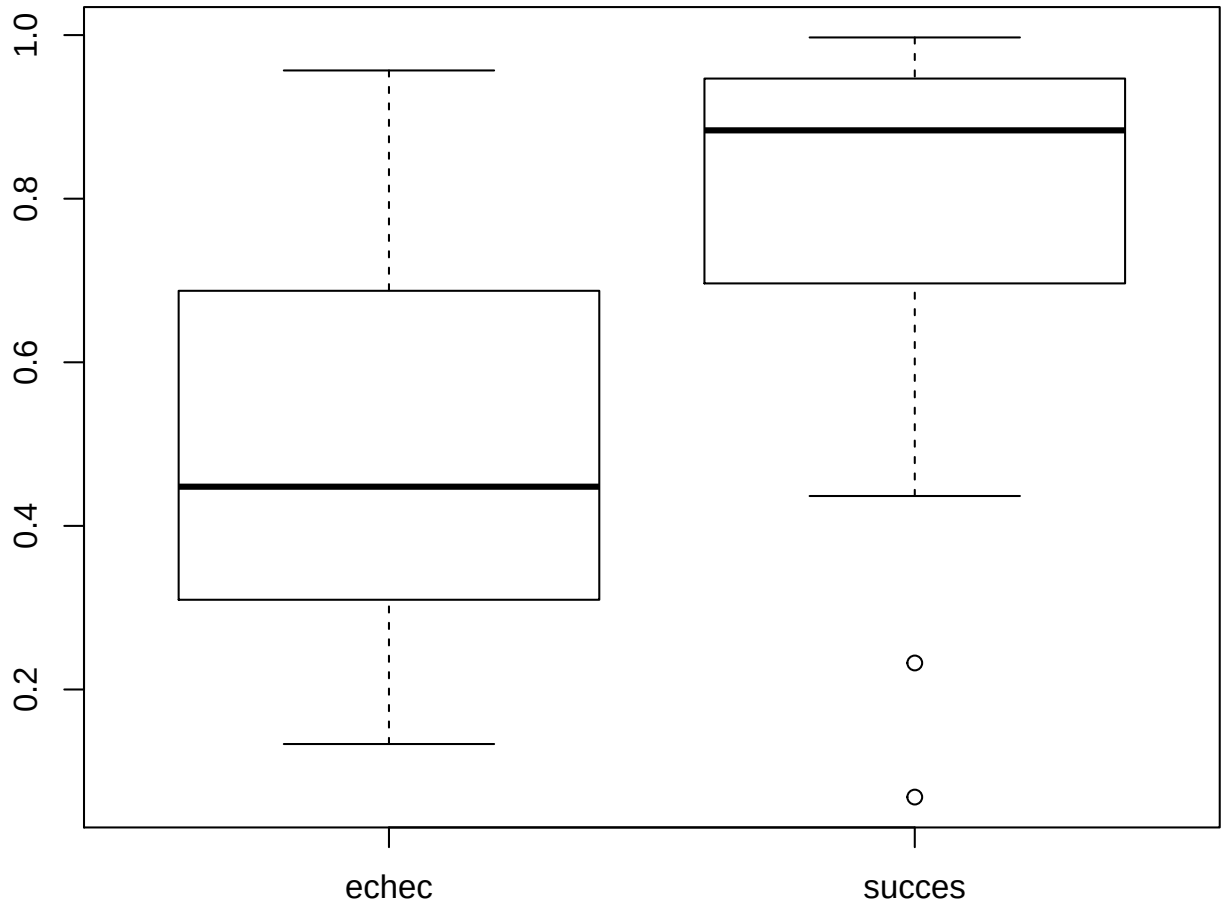
Variable à expliquer: la survie ou non du veau (= succès du traitement)

Variables explicatives (ici toutes quantitatives):

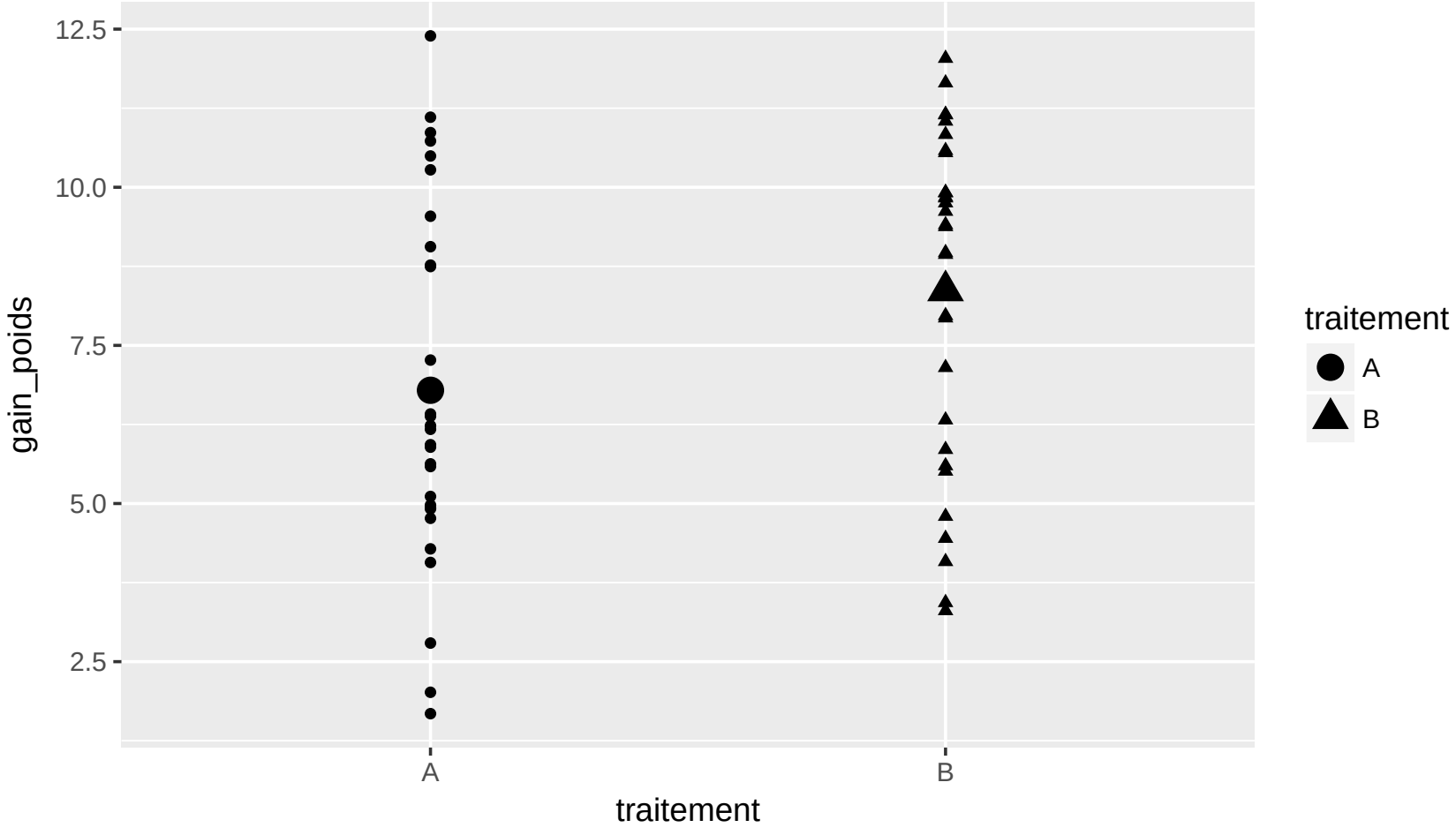
- l'excès de base
- l'âge
- la température
- la déshydratation
- la glycémie
- le score clinique

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

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

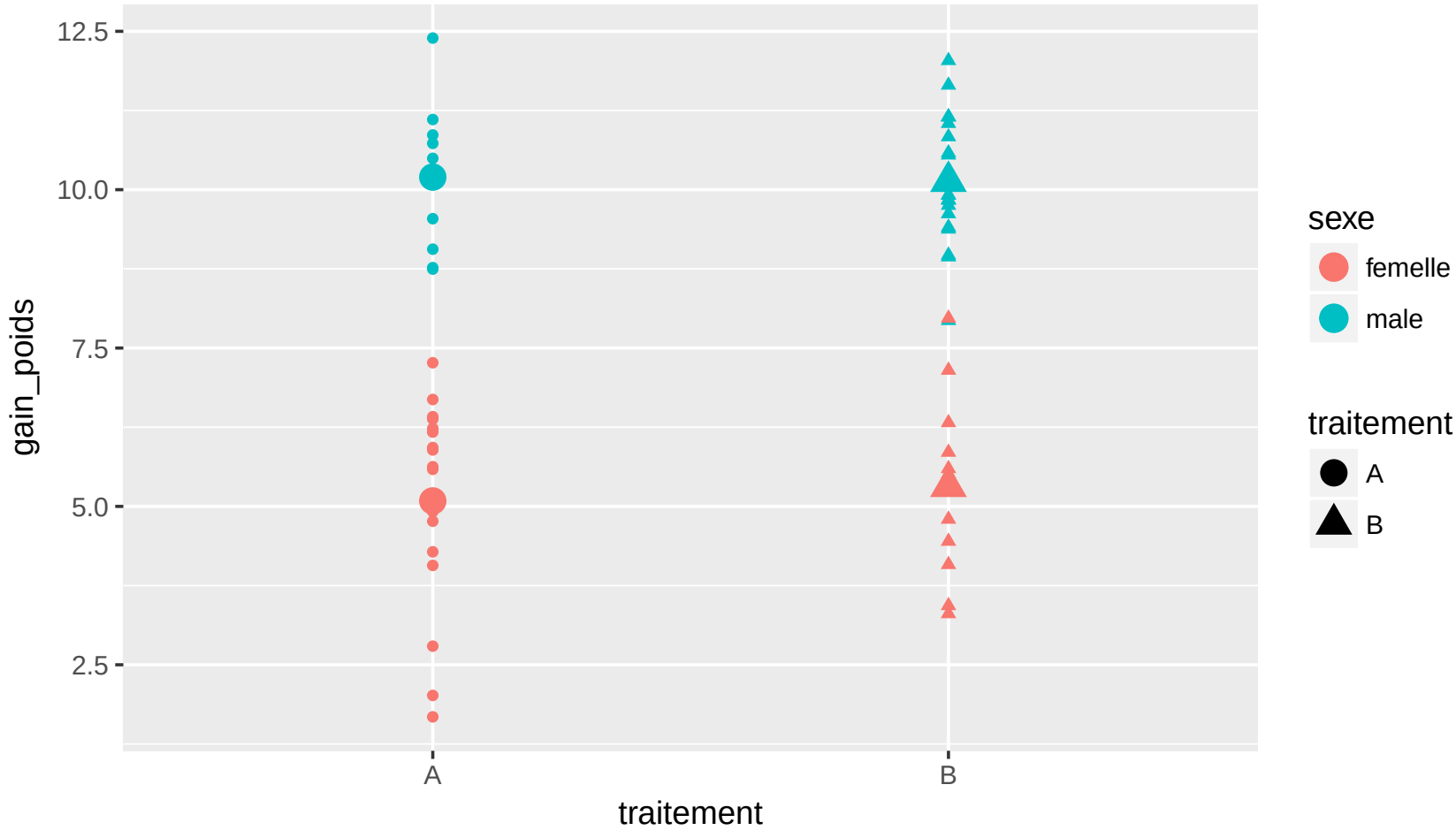


Un autre exemple si nécessaire pour vous convaincre des limites des approches basiques



On serait tenté de voir un effet du traitement sur le gain de poids.

Regardons maintenant l'effet du traitement et du sexe



Pourquoi voyait-on alors un effet du traitement précédemment ?

Examen du plan d'expérience

```
xtabs(~ traitement + sexe, data = d)
```

```
##             sexe
## traitement femelle male
##          A      20   10
##          B      11   19
```

L'effet qu'on aurait pu maladroitement imputer au traitement était en fait dû à l'effet du sexe et à une plus forte proportion de mâles traités avec le traitement B.

A retenir de ces exemples

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

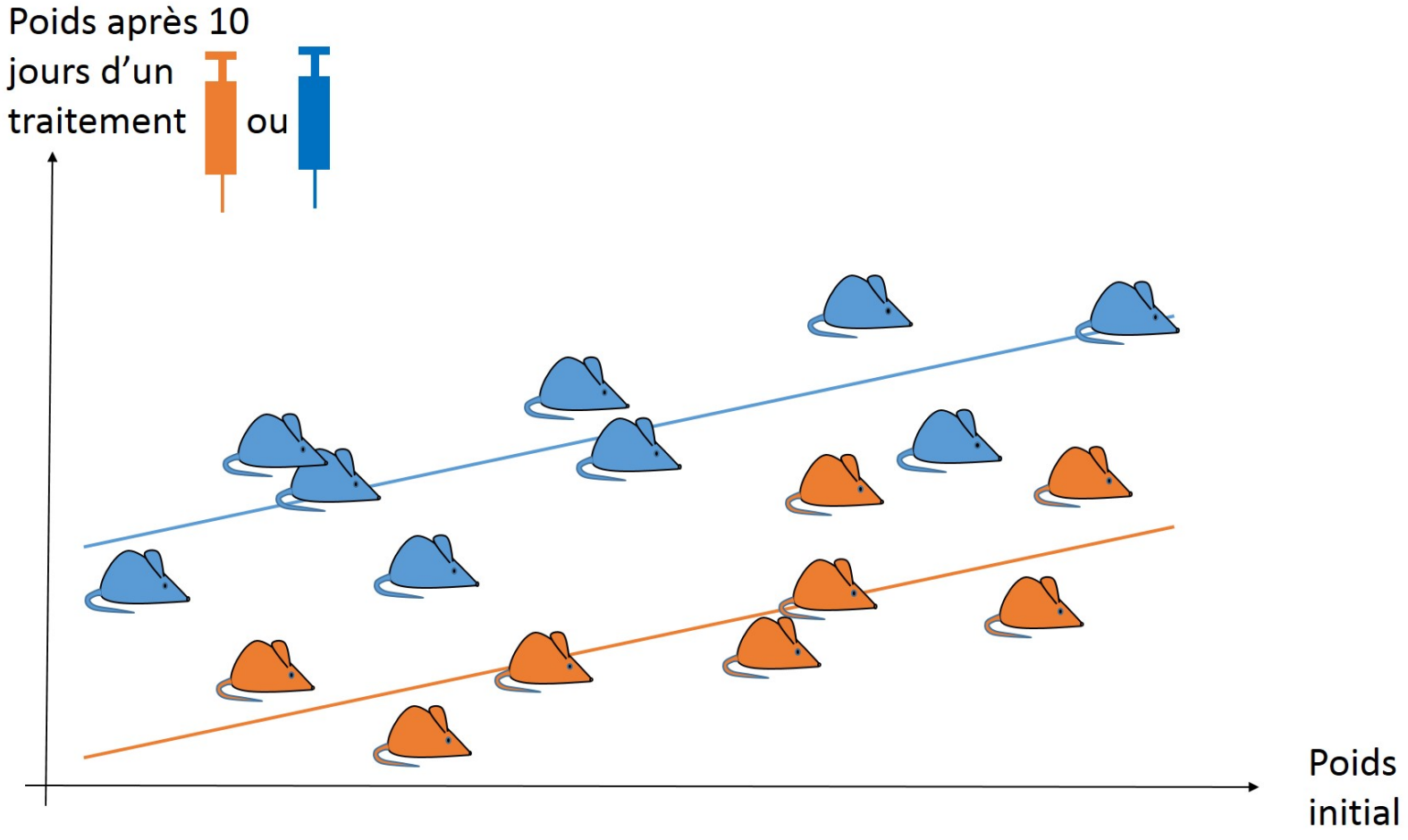
Il est d'autant plus dangereux de le faire si le plan d'expérience n'est pas équilibré, ou si certains facteurs ne sont pas contrôlés (cas des données d'observation rarement équilibrées).

Nécessité de vous former à des méthodes plus complexes si vous devez prendre en compte l'effet simultané de plusieurs facteurs.

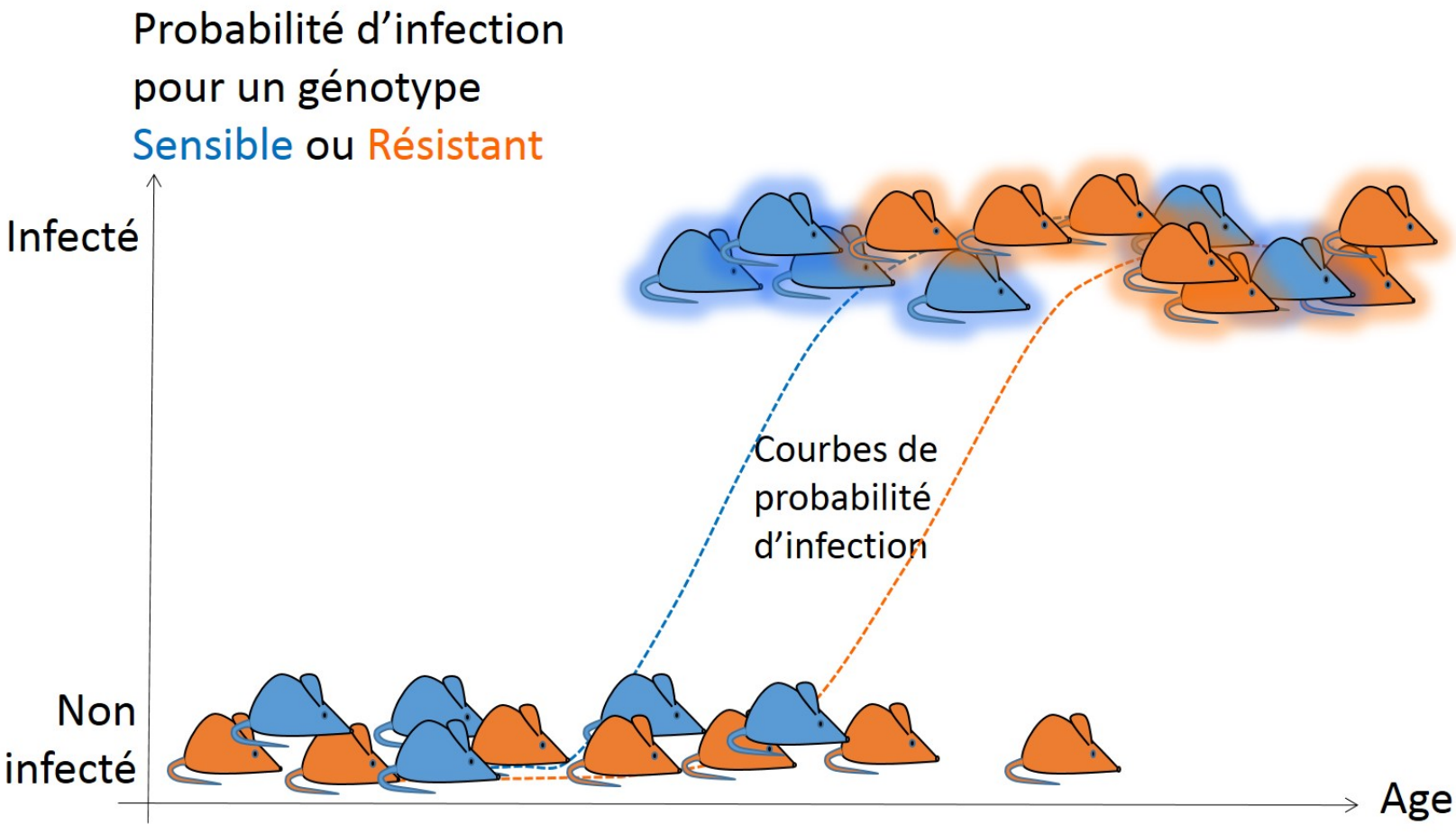
Ce à quoi vous pourrez vous former en 5A

- 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 analyses des mesures répétées dans le temps (**données longitudinales**)
- analyse de courbes de survie (**données de survie**)
- représentation de données spatialisées (**données spatialisées**)
- exploration d'un jeu de données issu d'une étude d'observation (ex. enquête) avec un grand nombre de variables (**analyse multivariée**)

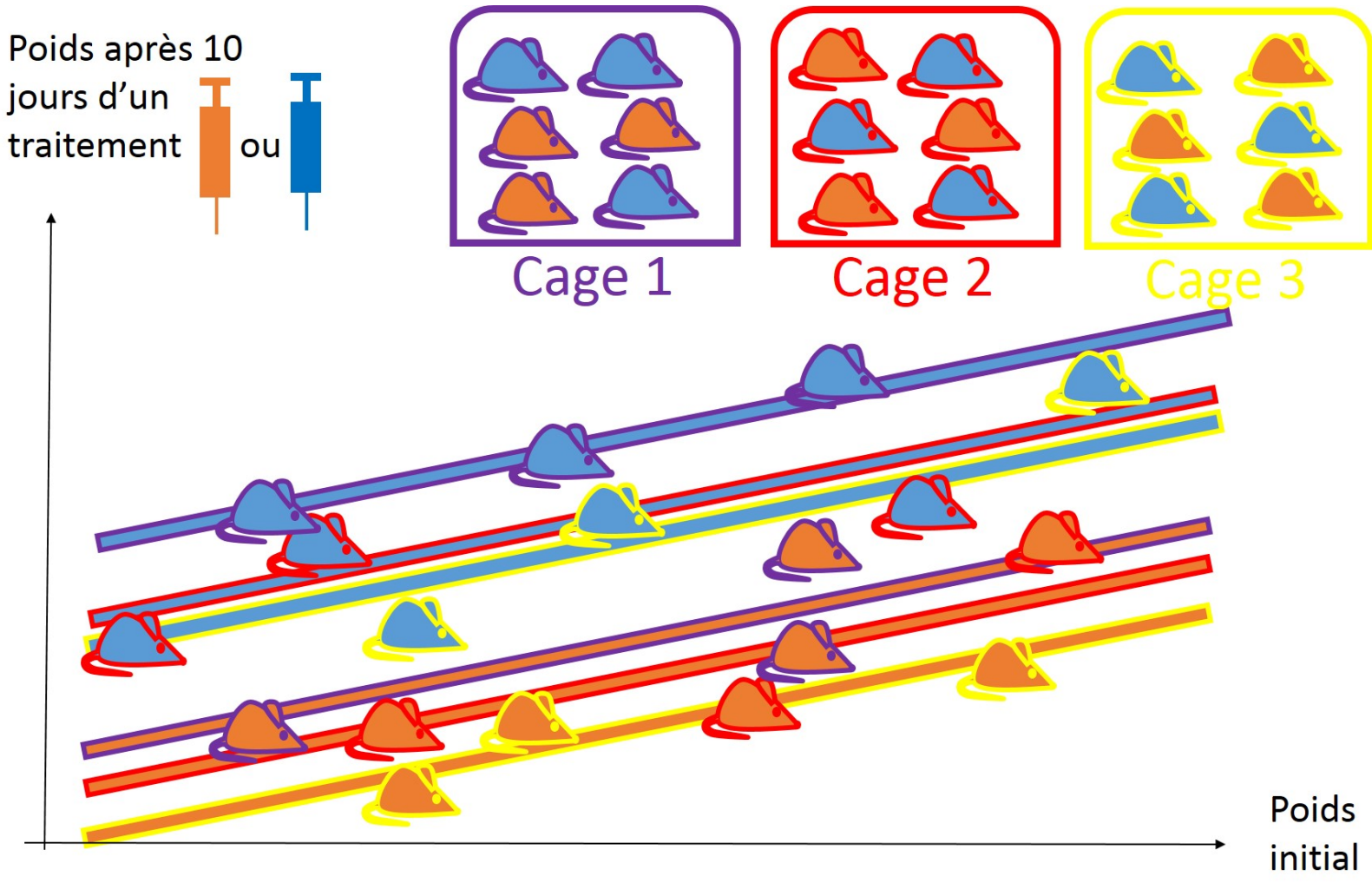
Module modèle linéaire



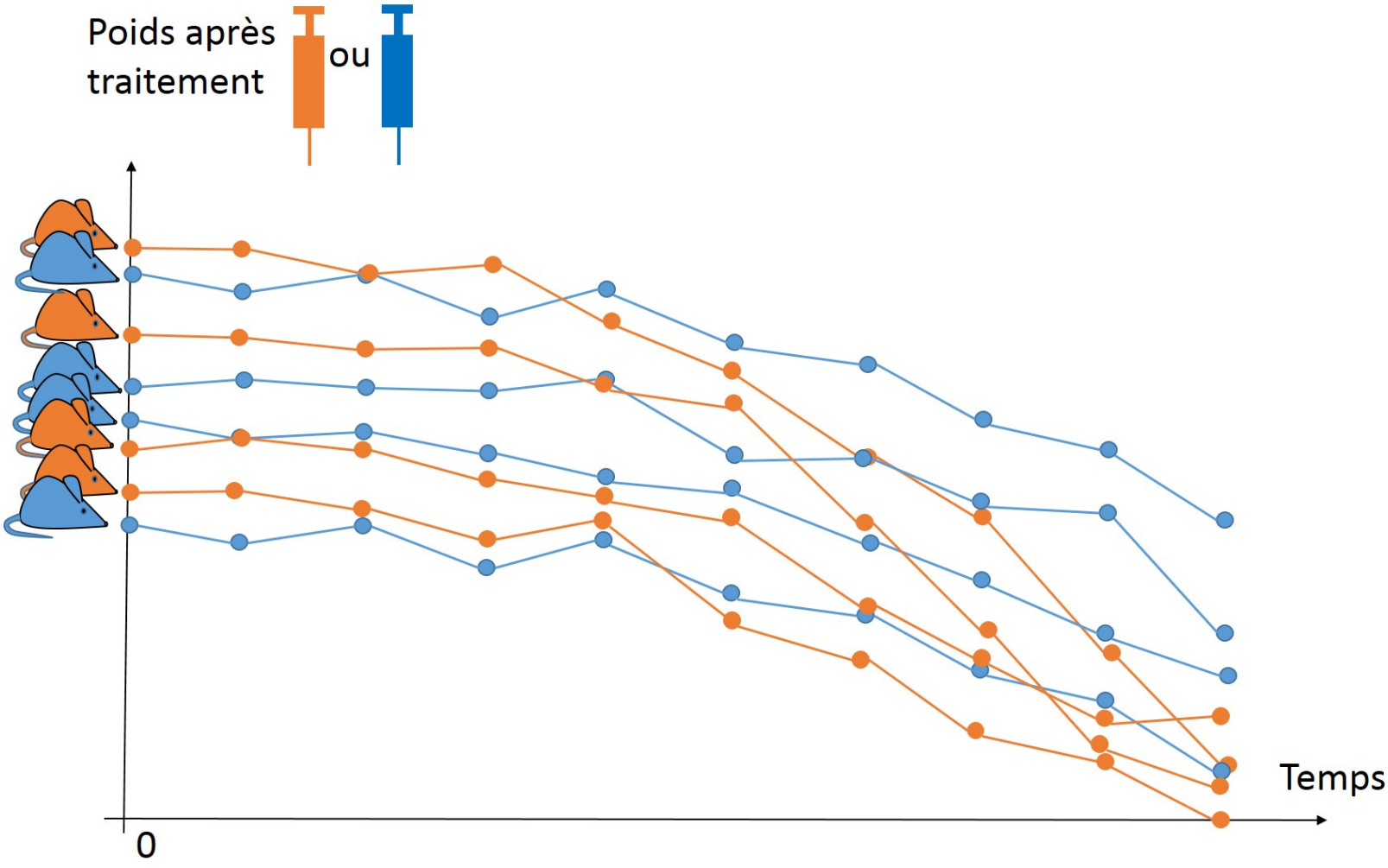
Module régression logistique



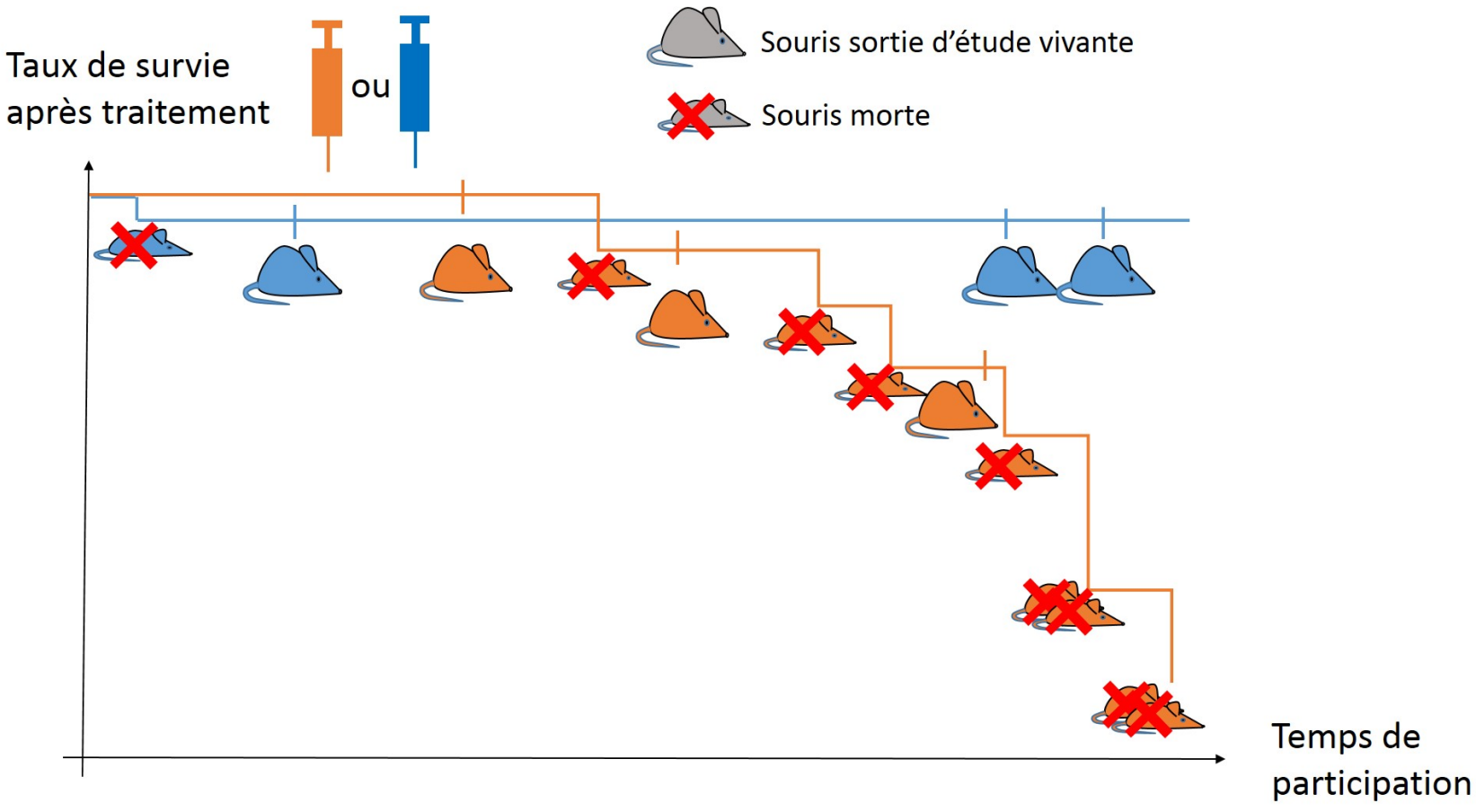
Module modèle mixte



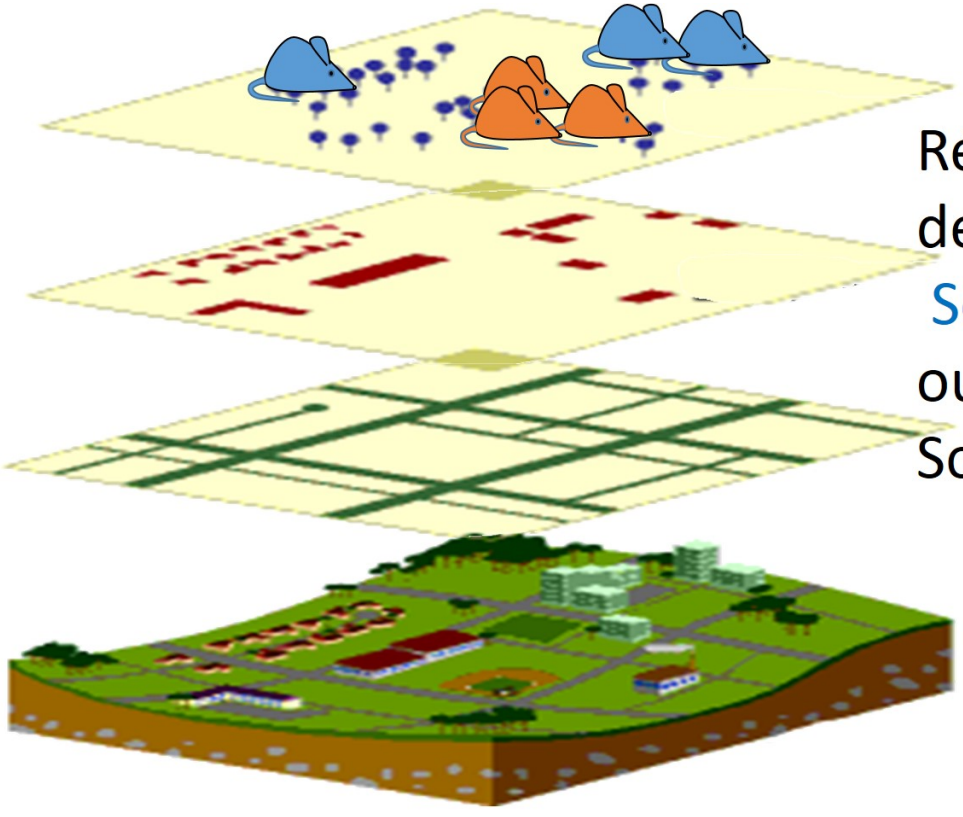
Module données longitudinales



Module données de survie

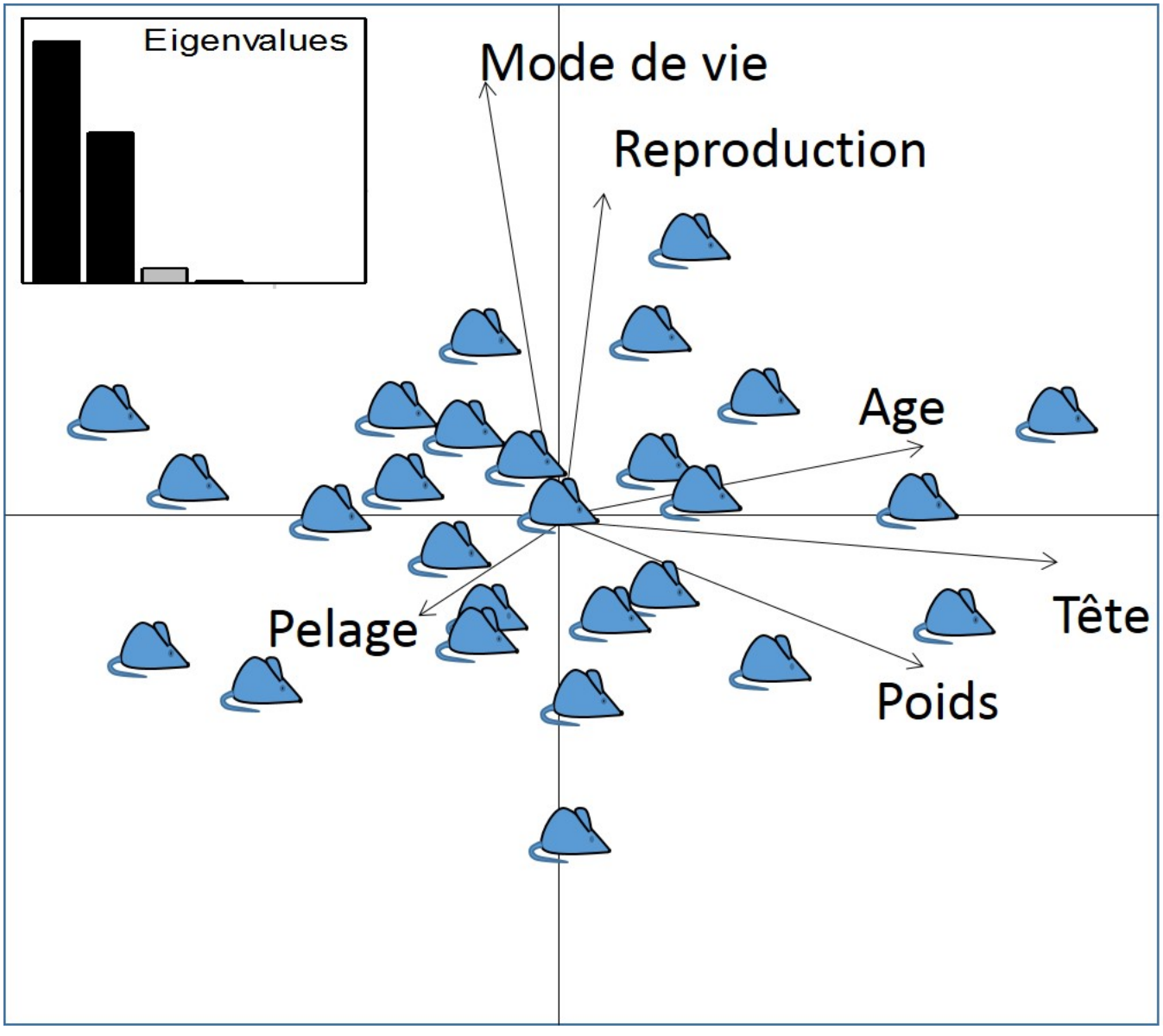


Module données spatialisées

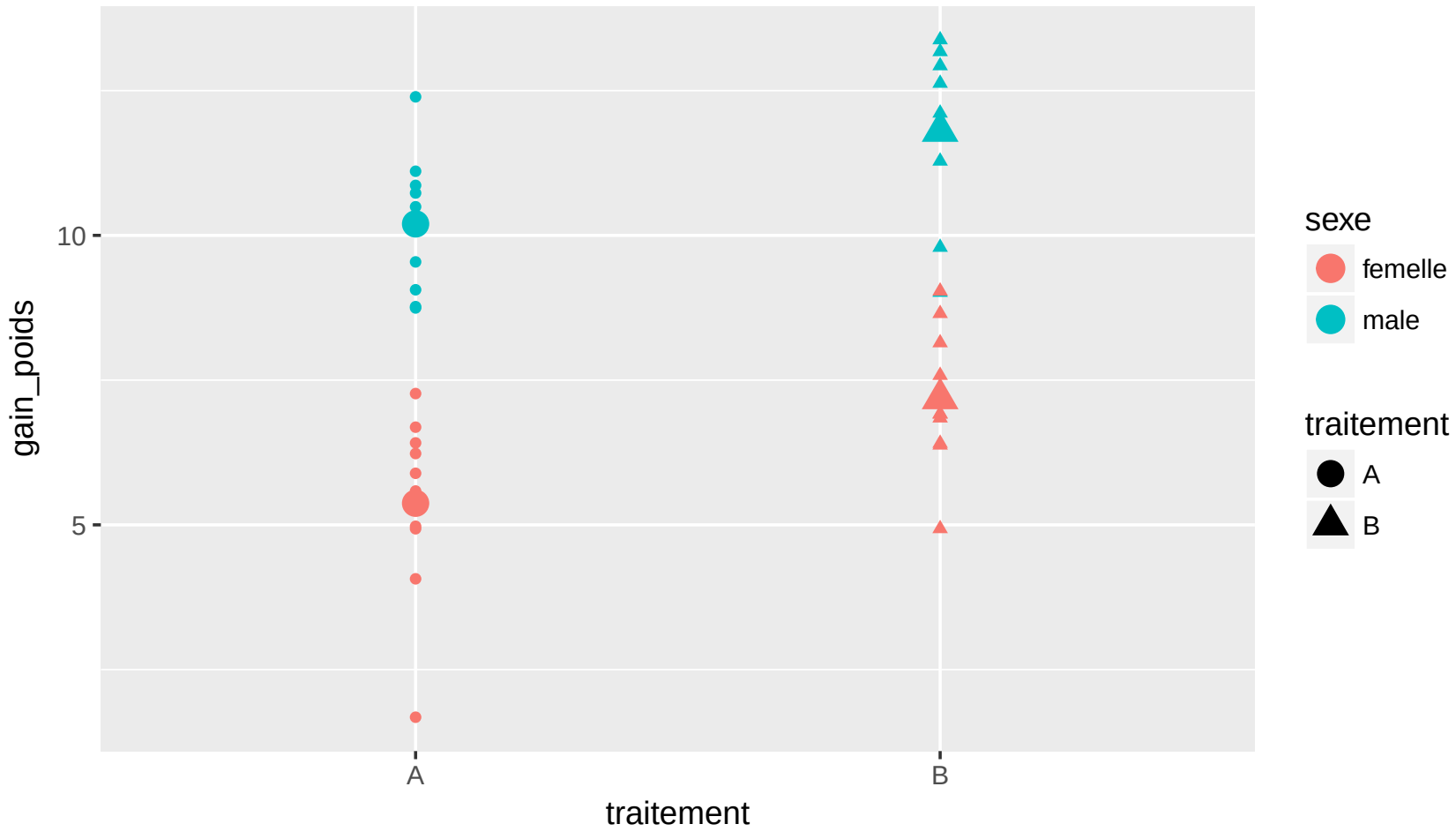


Répartition géographique
des génotypes
Sensible
ou Résistant
Sont-ils répartis au hasard ?

Module analyse multivariée

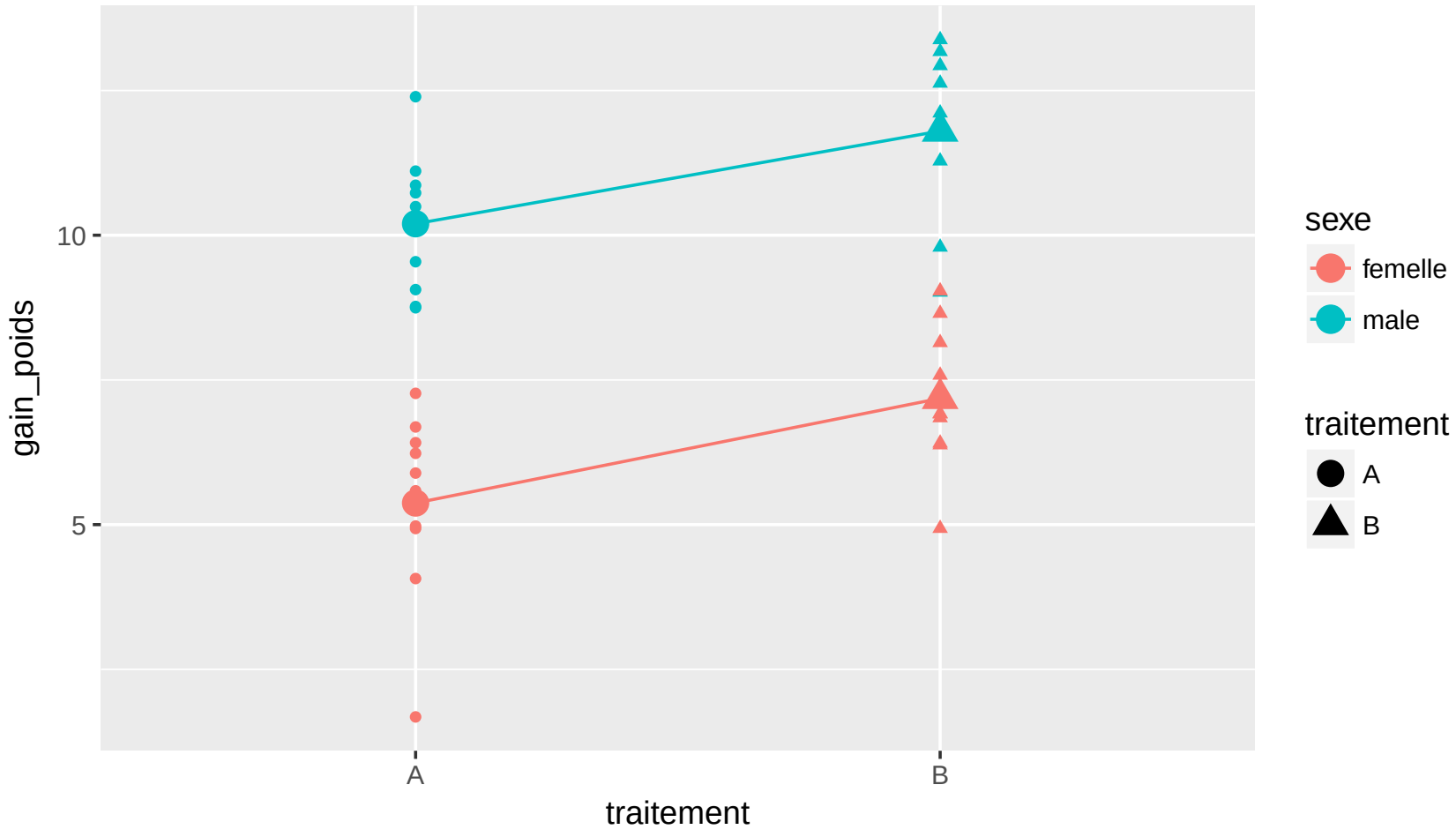


Changeons un peu l'exemple précédent



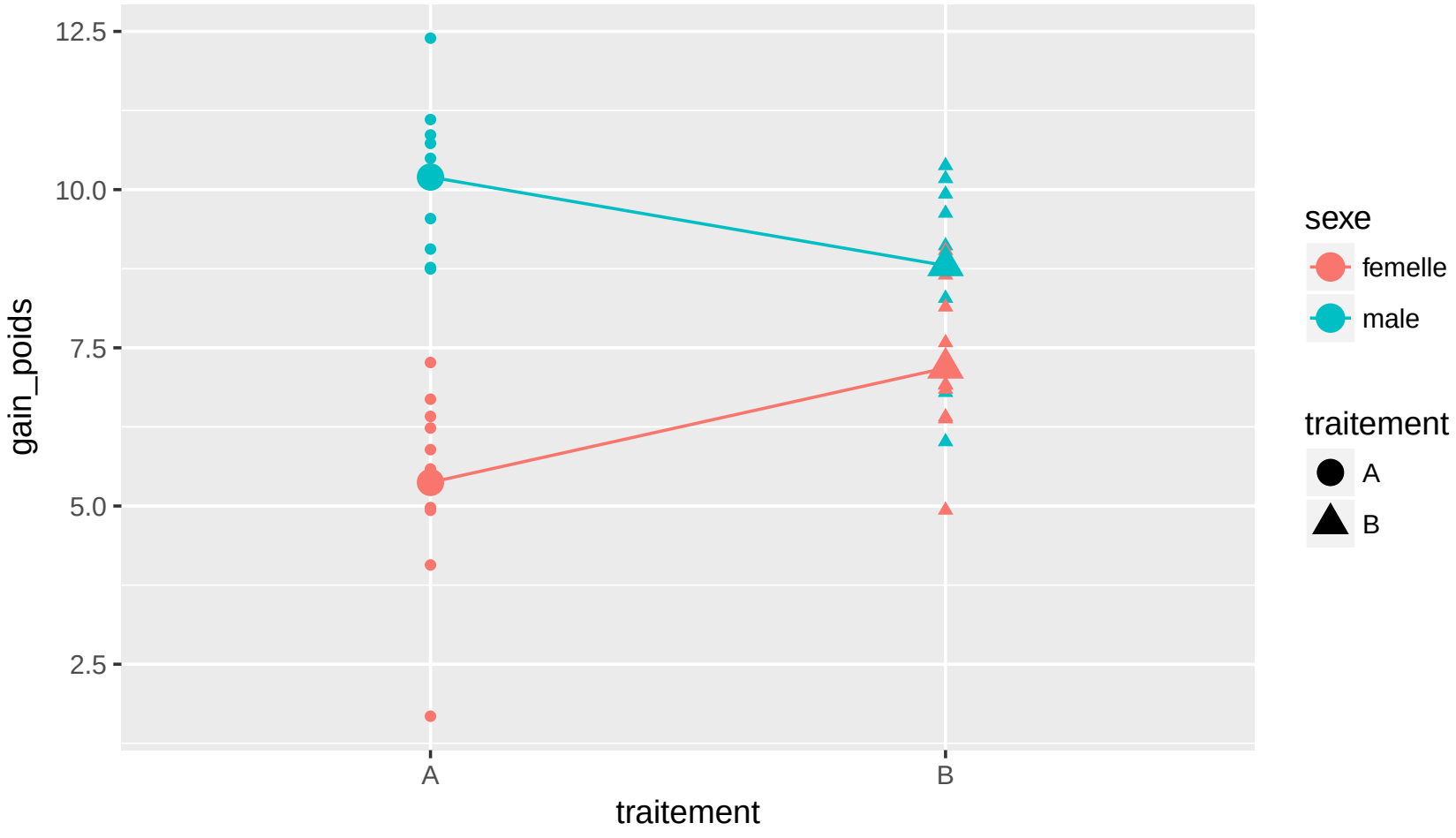
Il semblerait qu'il y ait à la fois un effet sexe et un effet traitement, sans interaction entre les deux (effets additifs).

Graphe d'interaction



Modèle additif (absence d'interaction) si segments parallèles.

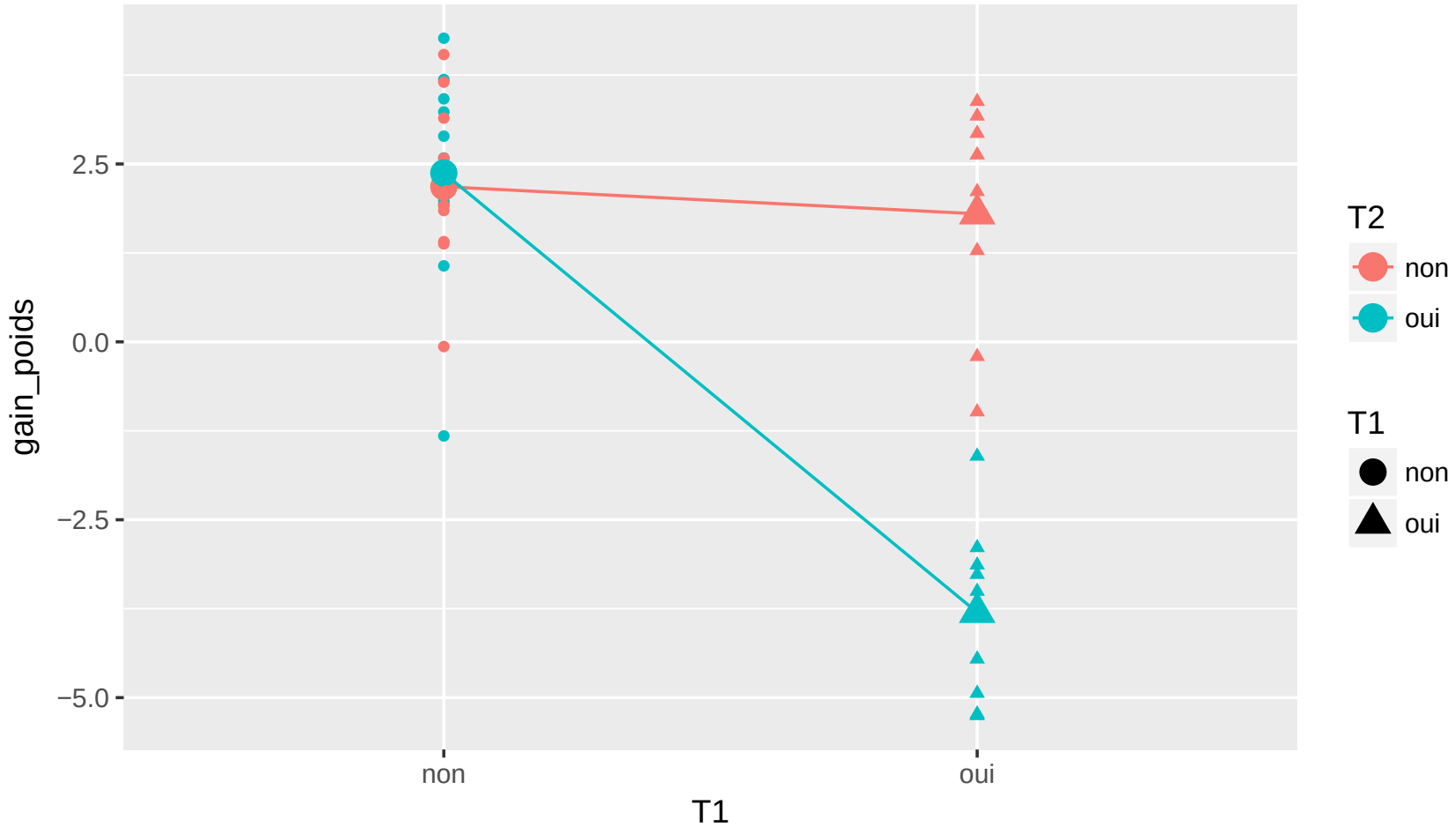
Changeons encore un peu l'exemple



Que dire de l'effet du traitement en présence d'interaction?

Difficile de conclure sans séparer mâles et femelles.

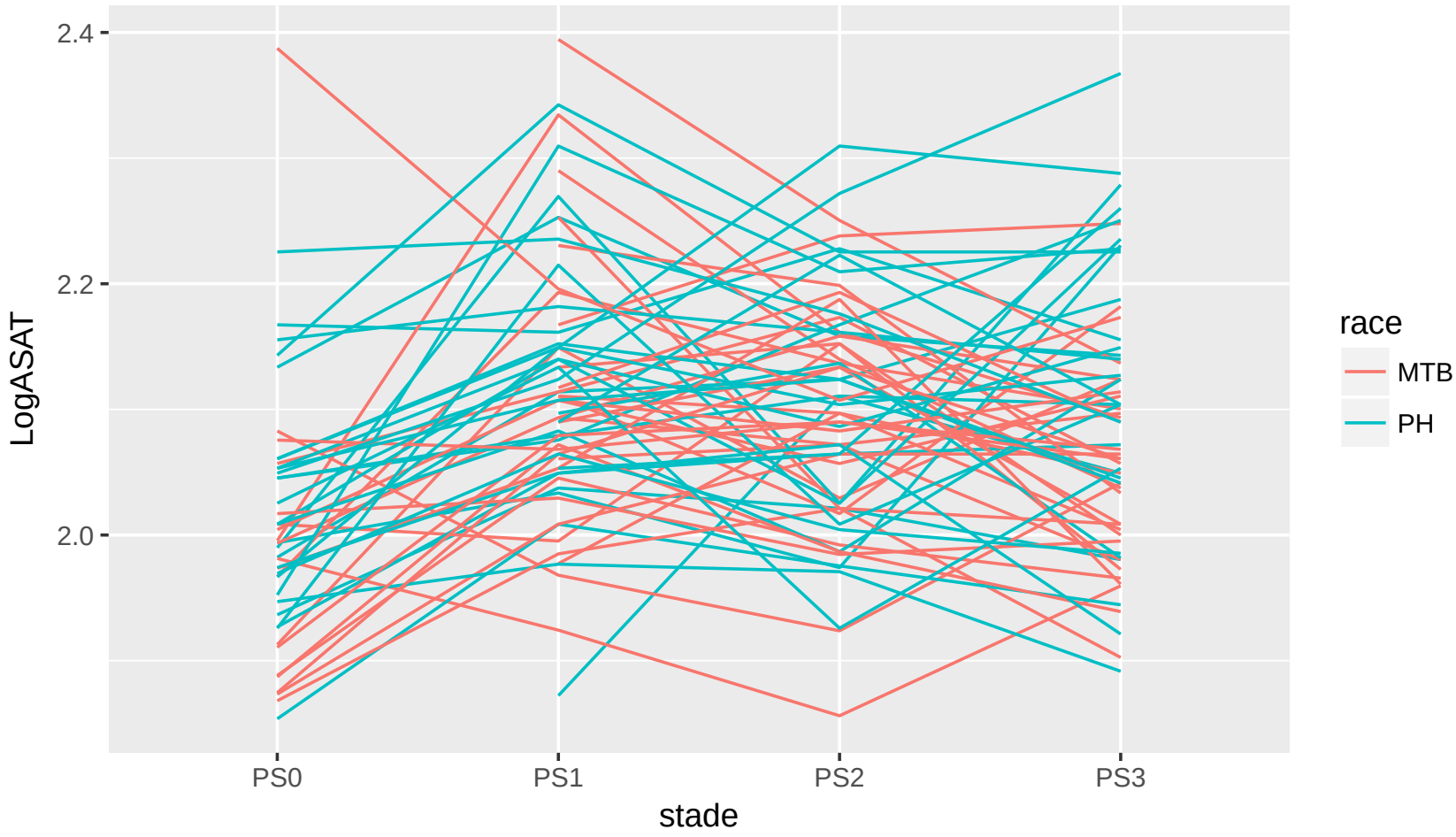
Parfois l'enjeu est de mettre en évidence une interaction



Mise en évidence d'une interaction de type médicamenteuse.

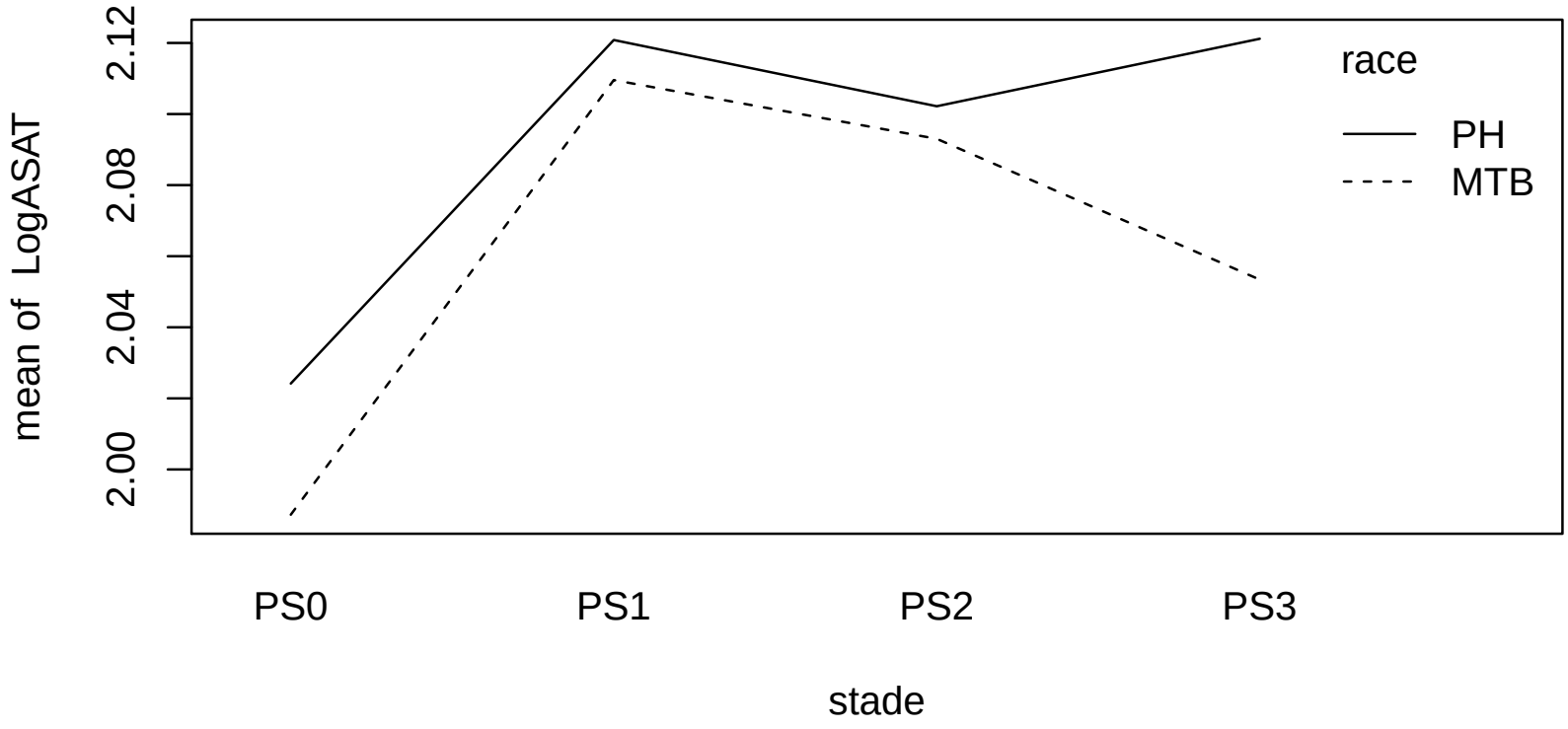
Mais souvent on est plutôt gêné par les interactions

Un exemple réel : effet du stade de lactation et de la race sur la teneur sanguine en ASAT (aspartate aminotransférase)



Mais souvent on est plutôt gêné par les interactions - graphe d'interaction

Un exemple réel : effet du stade de lactation et de la race sur la teneur sanguine en ASAT (aspartate aminotransférase)



En conclusion

- Dès que **plusieurs facteurs** peuvent avoir un effet sur la variable étudiée on ne peut plus se contenter des méthodes statistiques de base, il faut utiliser des **méthodes plus sophistiquées** permettant de **prendre en compte d'éventuels facteurs de confusion**.
- Dès qu'on modélise l'effet de plusieurs facteurs sur une variable étudiée, il faut envisager la possibilité d'une **interaction entre facteurs**, qui complique généralement l'interprétation de l'analyse des données.
- Dans le cadre expérimental, mieux vaut généralement limiter le nombre de facteurs étudiés et choisir des dispositifs équilibrés !