Points de vigilance lors du recueil et du rendu des données expérimentales

Marie Laure Delignette-Muller

25 janvier, 2022

Objectifs

- Savoir détecter les erreurs statistiques les plus courantes
- Savoir éviter ces erreurs lors de
  - la **réalisation**,
  - la **présentation**
  - et l'**interprétation des données** issues d'expériences sur l'animal

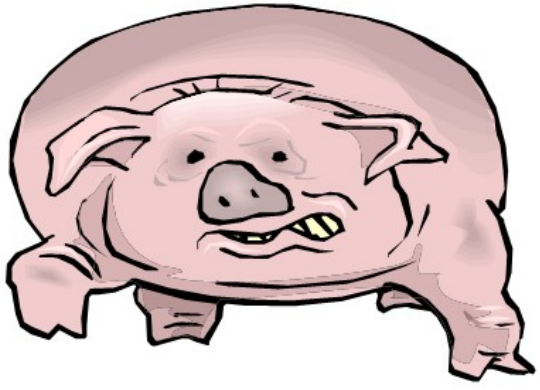
Erreurs lors de la réalisation d'une expérience sur l'animal

Erreurs lors de la réalisation d'une expérience sur l'animal

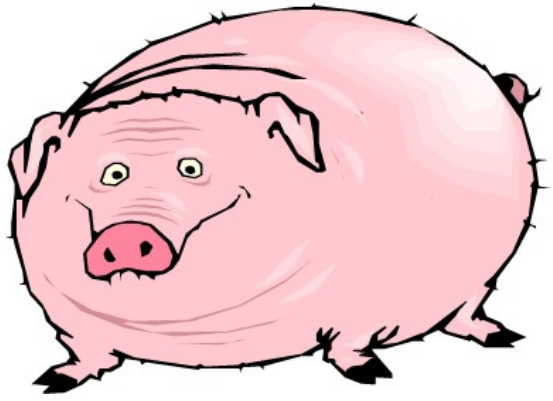
A vous de trouver ce qui ne va pas dans les exemples suivants

Exemple 1 - ???

avant traitement



après traitement

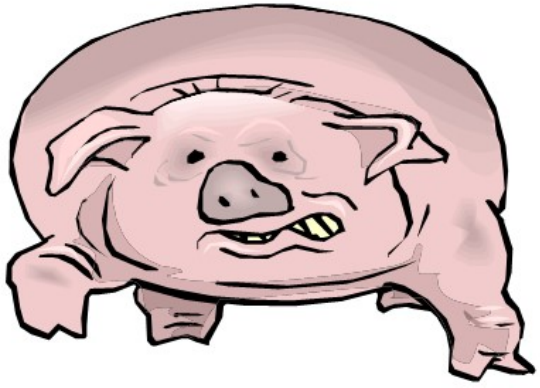


On observe une baisse significative ($p = 0.02$) de la température rectale sur un échantillon de 40 porcs malades après 2 jours de traitement.

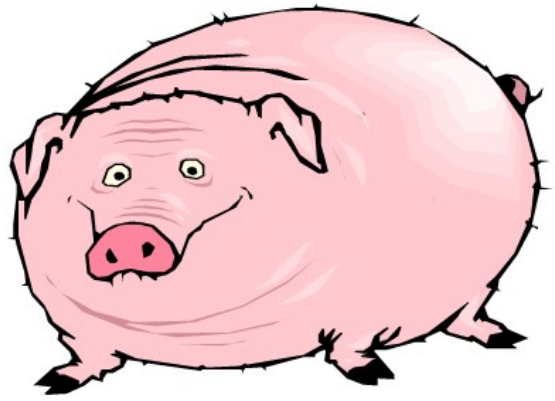
On en déduit un effet significatif du traitement sur la température rectale.

Exemple 1 - problème du témoin historique

avant traitement



après traitement



On ne peut rien conclure d'une expérience sans groupe témoin car l'effet mis en évidence peut être dû à l'évolution naturelle de la maladie ou à l'effet d'un autre facteur.

Nécessité de comparer le groupe traité à un groupe témoin (traité avec un placebo ou traitement de référence suivant les objectifs).

Exemple 2 - ???

On compare deux protocoles opératoires sur des césariennes de vaches :

- en 1998 : 55 vaches opérées selon le protocole P1
- en 1999 : 57 vaches opérées selon le protocole P2

La fréquence de complications est significativement plus faible avec le protocole P2 ($p < 0.001$).

On en conclut à une meilleure efficacité du protocole P2 pour limiter les complications.

Exemple 2 - non comparabilité initiale des groupes

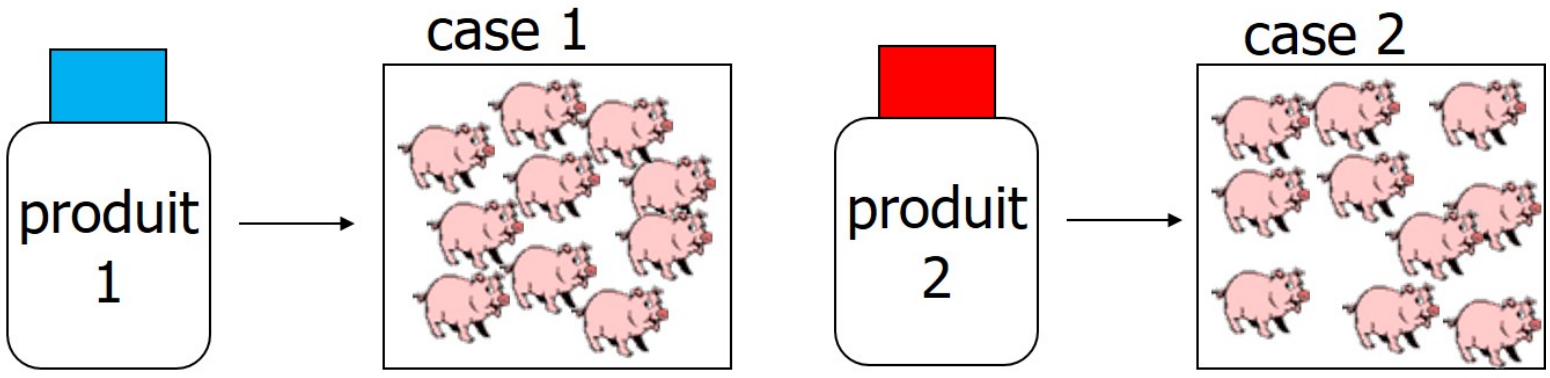
Comparaison de deux protocoles opératoires sur des césariennes de vaches :

- en 1998 : 55 vaches opérées selon le protocole P1
- en 1999 : 57 vaches opérées selon le protocole P2

Il existe dans ce type d'expérience un **biais potentiel de sélection**.

Afin de s'assurer de la comparabilité initiale des groupes, leur constitution doit faire l'objet d'une randomisation.

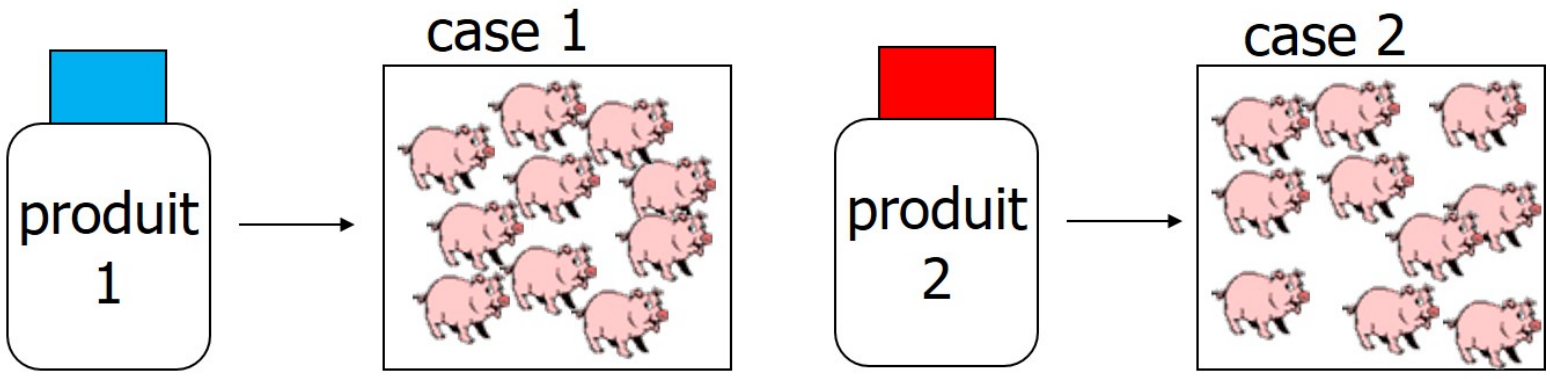
Exemple 3 - ???



On compare des gains de poids moyens sur deux échantillons randomisés de 10 porcs : p = 0.003. La différence est significative (p = 0.003).

On en conclut à des effets significativement différents des deux produits alimentaires sur le gain de poids.

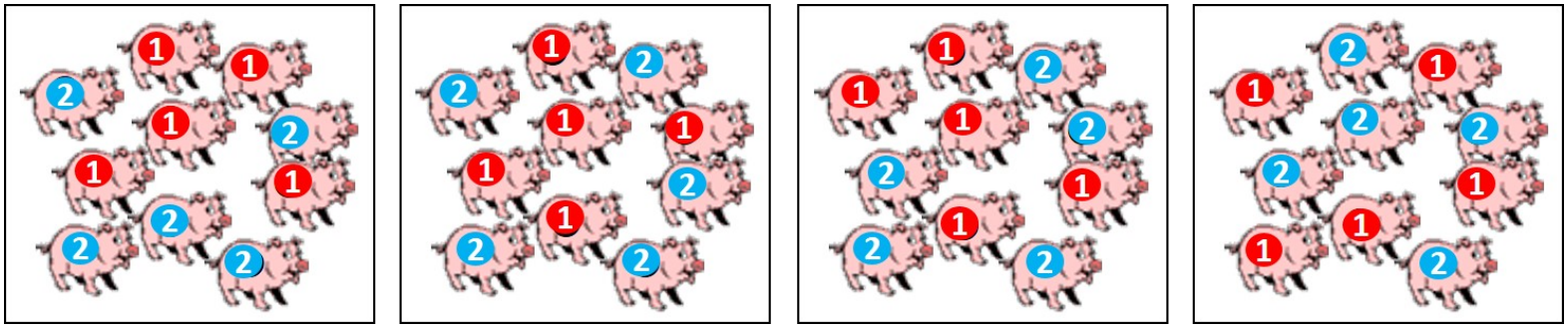
Exemple 3 - non comparabilité des groupes en cours d'étude



On ne prend pas en compte ici l'effet potentiel de facteurs de confusion. Ici l'effet "groupe" (case de porcs) est complètement confondu avec l'effet "traitement" (produit alimentaire).

Il est nécessaire de mettre en place un plan d'expérience contrôlant tous les facteurs concomitants susceptibles d'avoir un effet sur la variable étudiée.

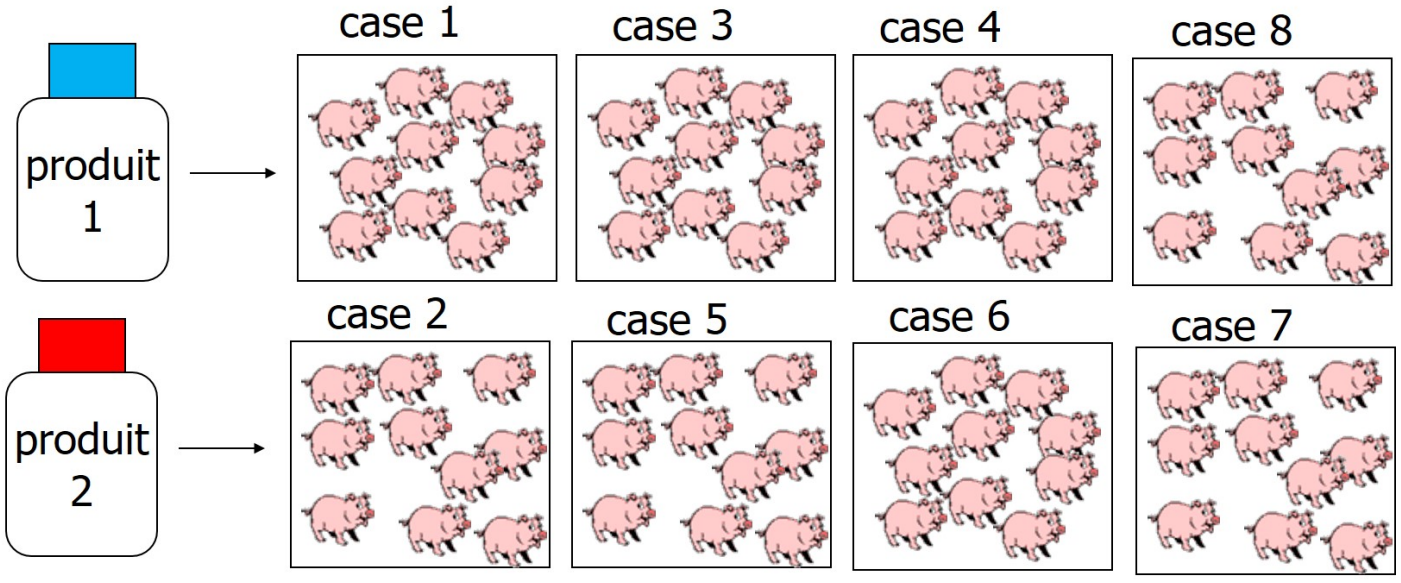
Exemple 3 - solution : randomisation au sein des cases



Une solution consiste à randomiser les porcs au sein de chaque case pour leur administrer l'un ou l'autre des deux produits.

Cette solution n'est pas toujours réalisable d'un point de vue pratique.

Exemple 3 - autre solution : randomisation des cases



Une seconde solution consiste à randomiser les groupes de porcs pour administrer à chaque groupe l'un ou l'autre des deux produits. Ainsi on n'a plus complète confusion entre l'effet "produit" et l'effet "case".

ATTENTION, il faudra alors bien prendre en compte l'effet "case" lors de la représentation et l'analyse des données. Et plus l'effet "case" est important, plus il faudra prendre de cases pour avoir une chance de mettre en évidence un effet.

Exemple 4 - ???

On fait un essai sur 45 chiens reproducteurs afin de comparer deux méthodes d'obtention de sperme :

- Groupe 1 (n = 10) : électroéjaculation
- Groupe 2 (n=35): utilisation d'un vagin artificiel

La comparaison des nombres moyens de spermatozoïdes obtenus ne montre pas de différence non significative ($p > 0.05$).

L'essai ne permet donc pas de mettre en évidence une différence entre les 2 méthodes.

Exemple 4 - perte de puissance due aux effectifs déséquilibrés

On fait un essai sur 45 chiens reproducteurs afin de comparer deux méthodes d'obtention de sperme :

- Groupe 1 (n = 10) : électroéjaculation
- Groupe 2 (n=35): utilisation d'un vagin artificiel

Une meilleure connaissance *a priori* de l'un des traitements ne peut pas être prise en compte en statistique fréquentiste : seules les données obtenues dans une expérience sont prises en compte pour conclure.

Des effectifs équilibrés sont donc généralement préférables pour comparer 2 groupes.

Exemple 5 - ???

Dans le cadre d'une étude multicentrique deux traitements locaux (une crème et une solution) sont comparés dans le cadre d'un essai randomisé comprenant 45 chiens par groupe.

L'aspect de la lésion à l'issue du traitement est évalué par le vétérinaire prescripteur.

Groupe	Aspect (%)		
	propre	souillé, huileux	sec, crevassé
1 (n=40)	65	17.5	17.5
2 (n=43)	16.3	65.1	18.6

Exemple 5 - critères subjectifs non lus en aveugle

Dans le cadre d'une étude multicentrique deux traitements locaux (une crème et une solution) sont comparés dans le cadre d'un essai randomisé comprenant 45 chiens par groupe.

L'aspect de la lésion à l'issue du traitement est évalué par le vétérinaire prescripteur.

Le vétérinaire connaît le traitement qu'il a lui-même prescrit à l'animal. De ce fait son appréciation d'un caractère subjectif peut être biaisée.

La lecture en aveugle s'impose dans le cas de critères subjectifs.

Exemple 5 - non prise en compte des données manquantes

Dans le cadre d'une étude multicentrique deux traitements locaux (une crème et une solution) sont comparés dans le cadre d'un essai randomisé comprenant **45 chiens par groupe**.

Groupe	Aspect (%)		
	propre	souillé, huileux	sec, crevassé
1 (n=40)	65	17.5	17.5
2 (n=43)	16.3	65.1	18.6

Il manque des données : que sont devenus les chiens manquants ? Des données manquantes peuvent être informatives (ex.: guérisons rapides et complètes, décès d'animaux . . .).

Les perdus de vue (ou les exclusions en cours d'étude) doivent être pris en compte dans la collecte et l'analyse des données afin d'éviter le **biais d'attrition** : **la transparence est de rigueur** et une **analyse en intention de traiter (ITT)** peut parfois être envisagée.

A RETENIR !!!

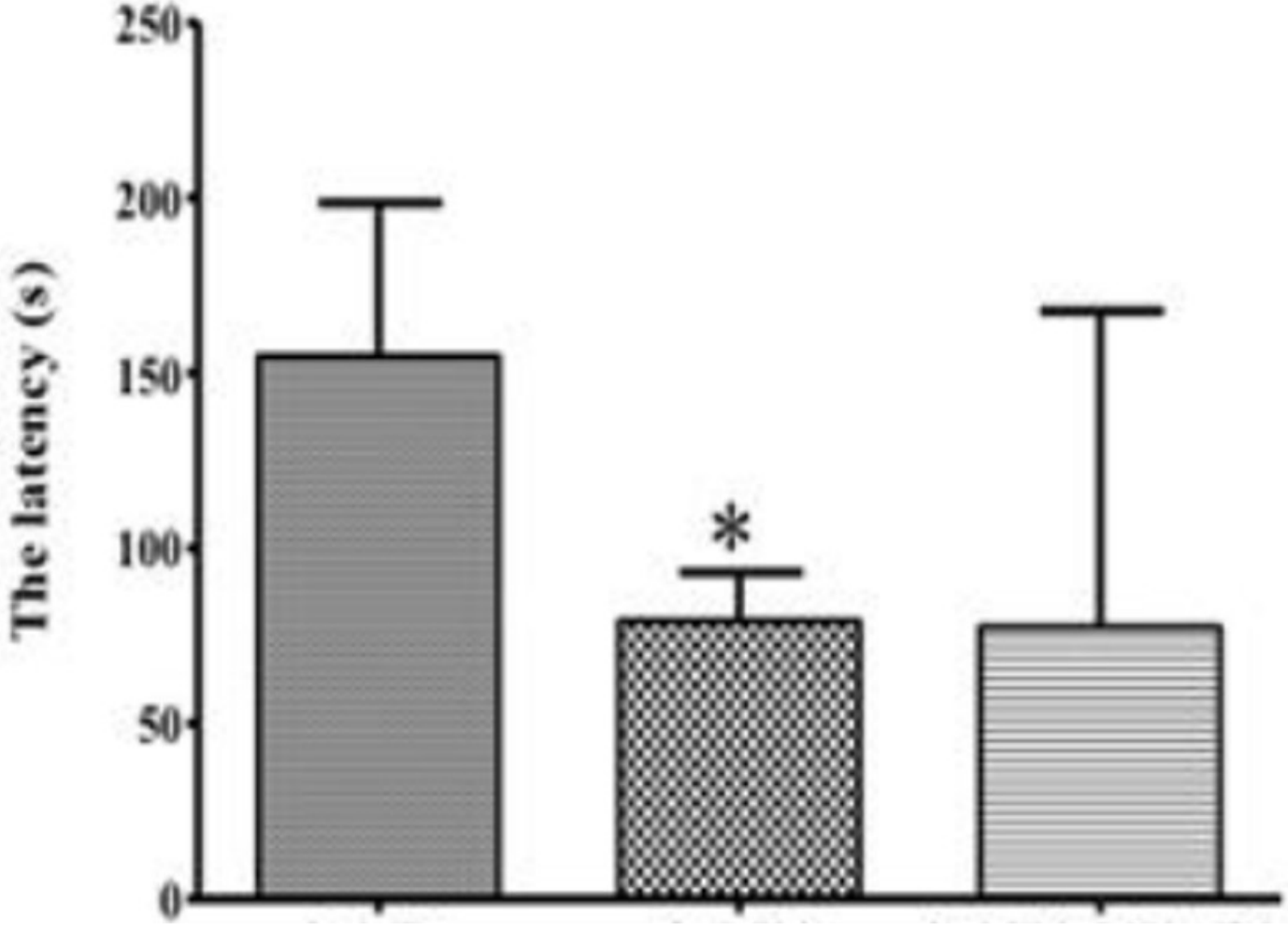
- Un essai doit être **comparatif** avec des **groupes comparables** au départ et tout au long de l'essai.
- Des **effectifs équilibrés** sont généralement préférables.
- Les éventuels **critères subjectifs** doivent absolument être **lus en aveuge**.
- On se doit d'être **transparent au sujet des données manquantes**.

Erreurs lors de la représentation des données

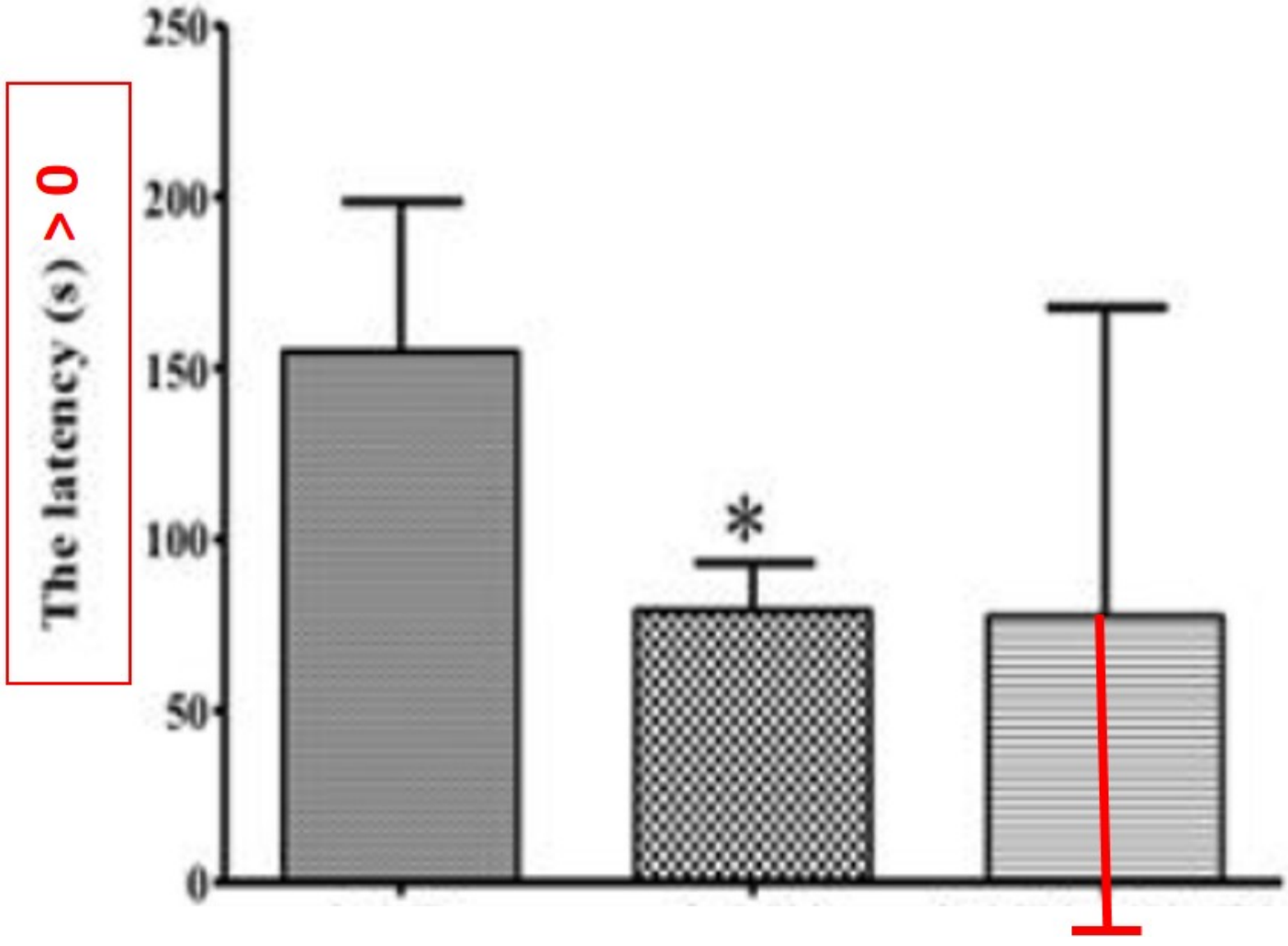
Erreurs lors de la représentation des données

A vous de trouver ce qui ne va pas dans les exemples suivants

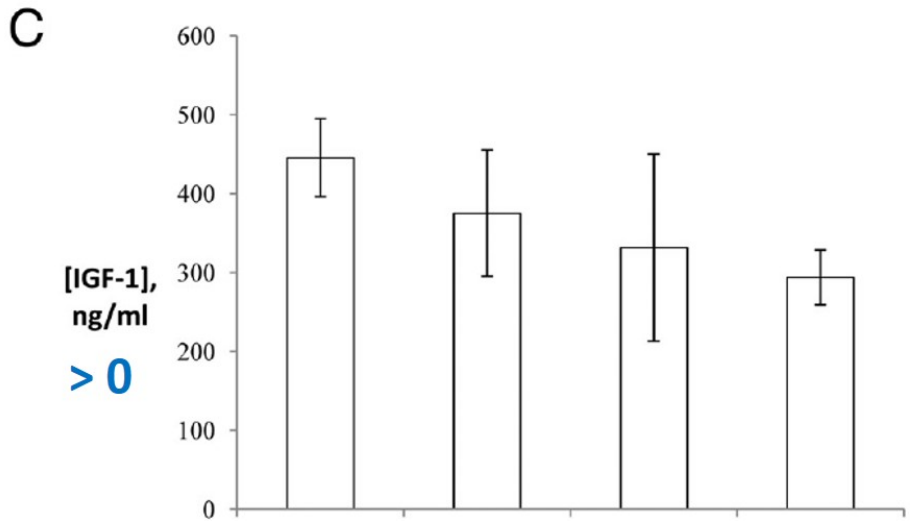
Exemple 6 - ???

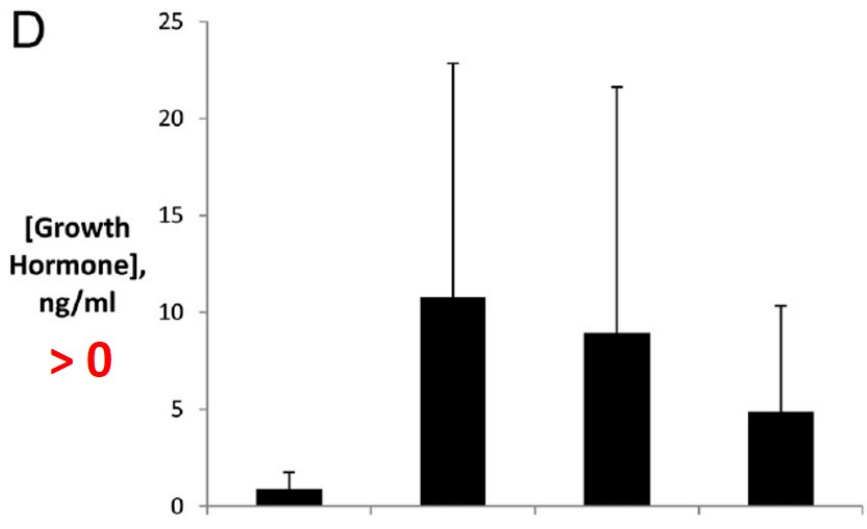


Exemple 6 - Barres d'erreur qui n'ont pas de sens

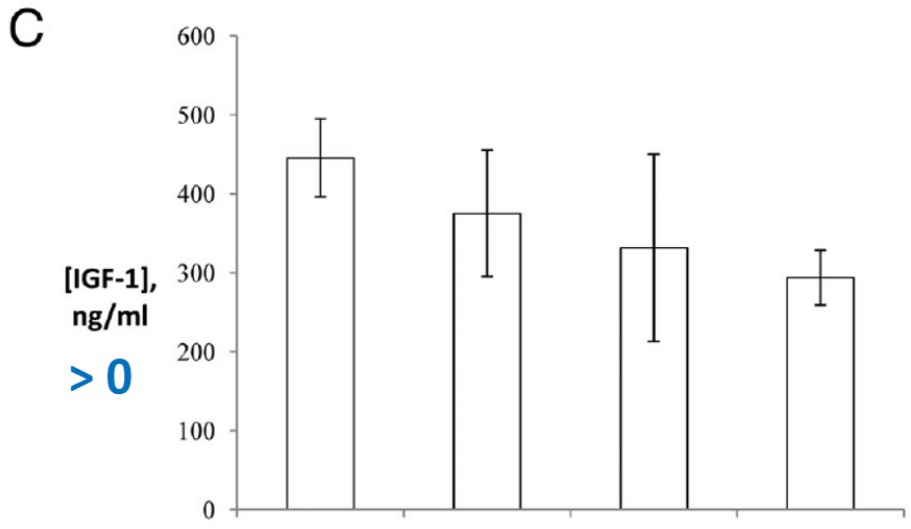


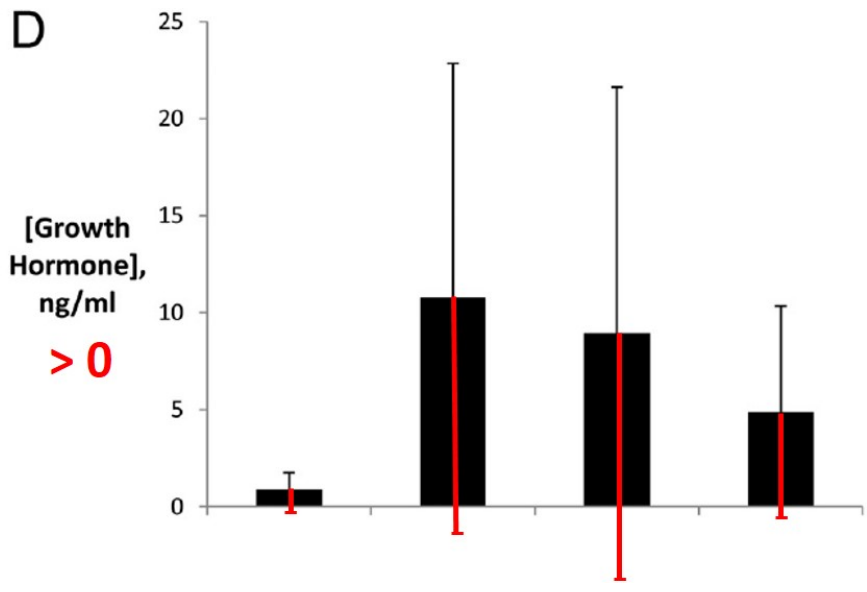
Exemple 7 - ???


C
[IGF-1], ng/ml
> 0
600
500
400
300
200
100
0


D
[Growth Hormone], ng/ml
> 0
25
20
15
10
5
0

Exemple 7 - Même souci, et ici on ne peut croire à la simple ignorance des auteurs





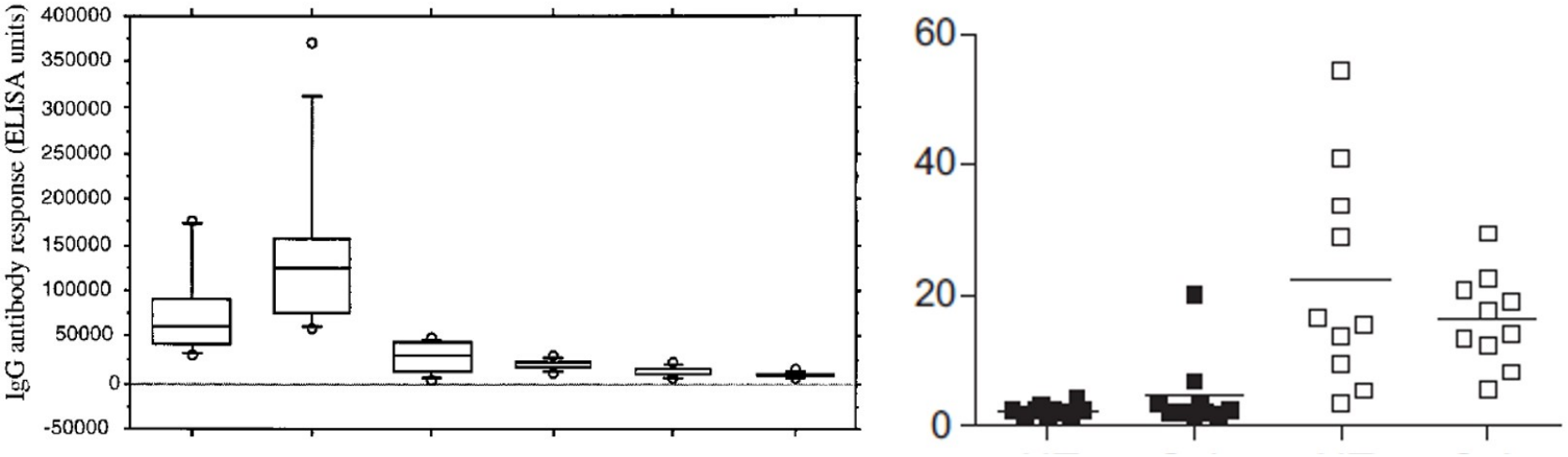
Petit rappel sur les **intervalles de fluctuation** et les **intervalles de confiance** sur la moyenne

Pour une distribution normale (Gaussienne),

- $m \pm 2SD$ représente l'**intervalle de fluctuation à 95%** (intervalle contenant 95% des observations)
- $m \pm SD$ représente l'**intervalle de fluctuation à 68%** (intervalle contenant 68% des observations)
- $m \pm 2SEM$ représente l'**intervalle de confiance à 95%** de la moyenne (indication d'incertitude sur la moyenne estimée : quand on calcule des intervalles de confiance à 95% on sait qu'une fois sur 20 on se trompe)
- $m \pm 2SEM$ représente l'**intervalle de confiance à 68%** de la moyenne

avec SD l'écart type (standard deviation) et $SEM = \frac{SD}{\sqrt{(n)}}$

Exemple 8 et 9 - Graphes alternatifs plus informatifs



- Les diagrammes en boîte (boîtes à moustache) représentant les valeurs minimales, maximales et les quartiles (médiane et quantiles à 25 et 75%) sont tout aussi synthétiques et nettement plus informatifs.
- Lorsque le nombre de points observés est très limités, il convient de montrer tous les points.

Quel paramètre résumé a été ajouté au dernier graphe ? Par quel autre plus robuste aurait-on dû le remplacer ?

Attention aux conditions d'utilisation des méthodes statistiques

Quand les distributions sont **non normales**

et **non normalisables par transformation de variable**,

et que le théorème central limite n'est pas applicable

(du fait d'effectifs petits ou de lois très éloignées de lois normales)

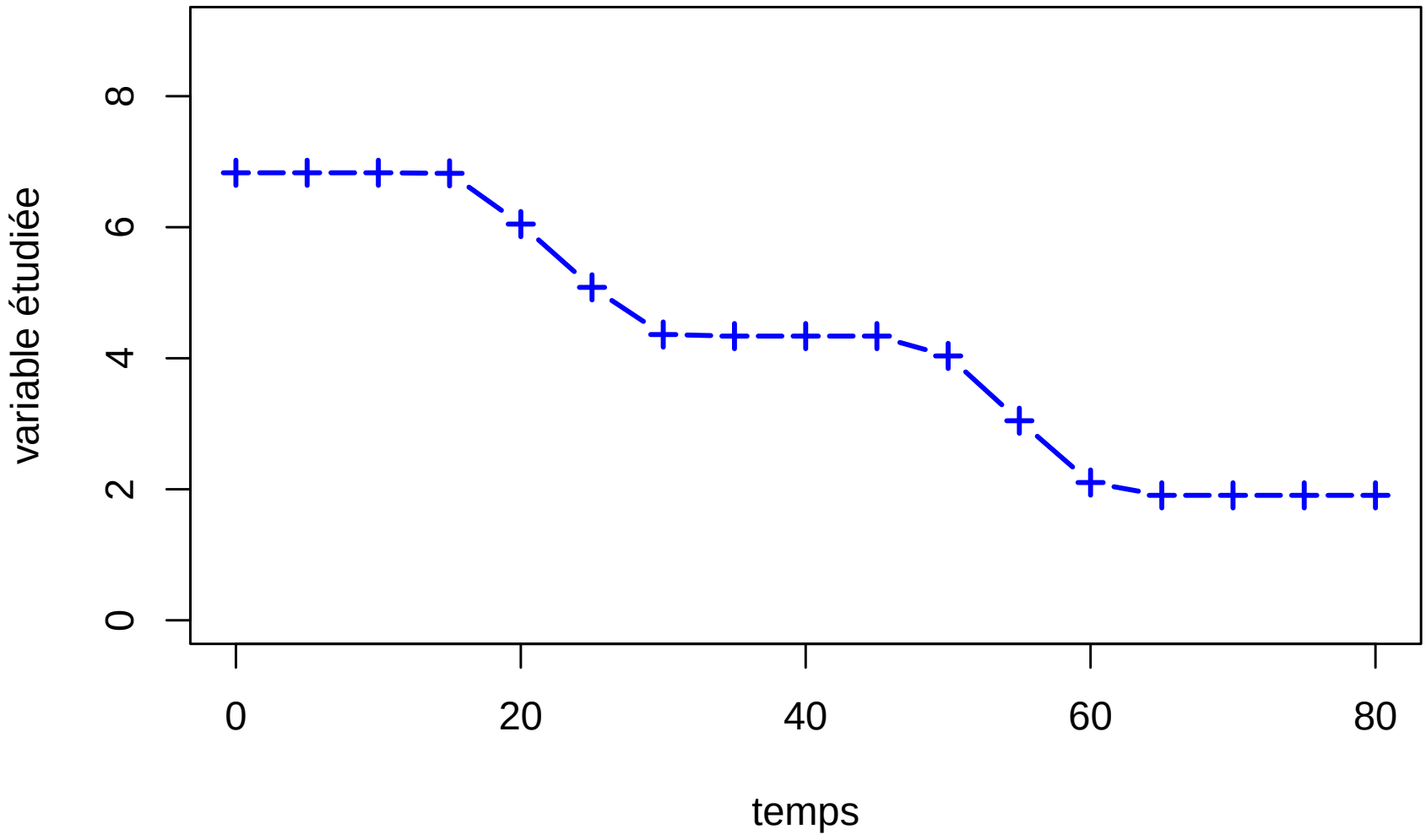
mieux vaut utiliser des **statistiques de rang**,

y compris pour résumer / représenter les données

(tests non paramétriques, tendance centrale résumée par la médiane, . . .)

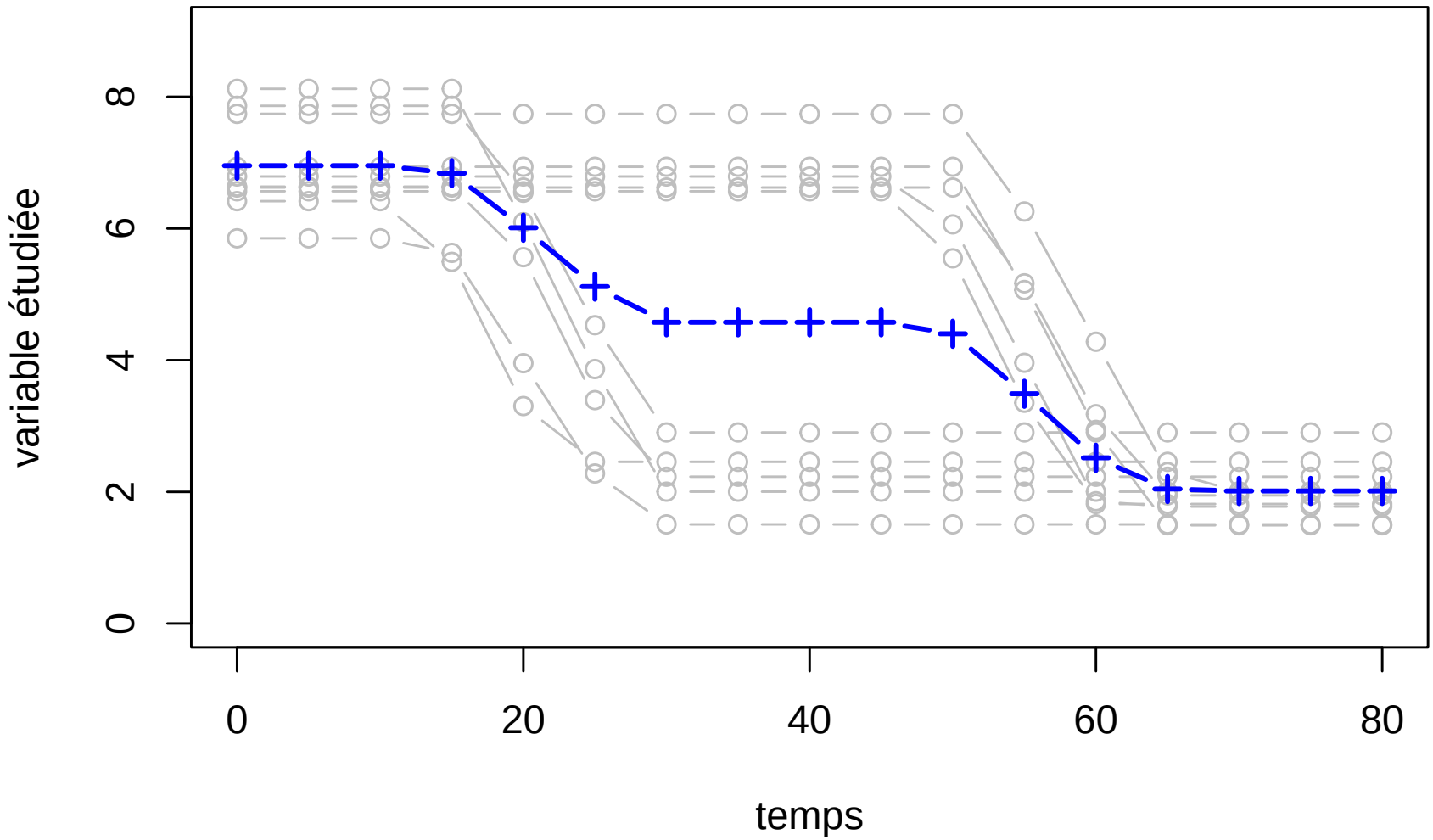
Exemple 10 - ???

Représentation d'une courbe moyenne : ici moyenne de la réponse sur tous les animaux à chaque temps.

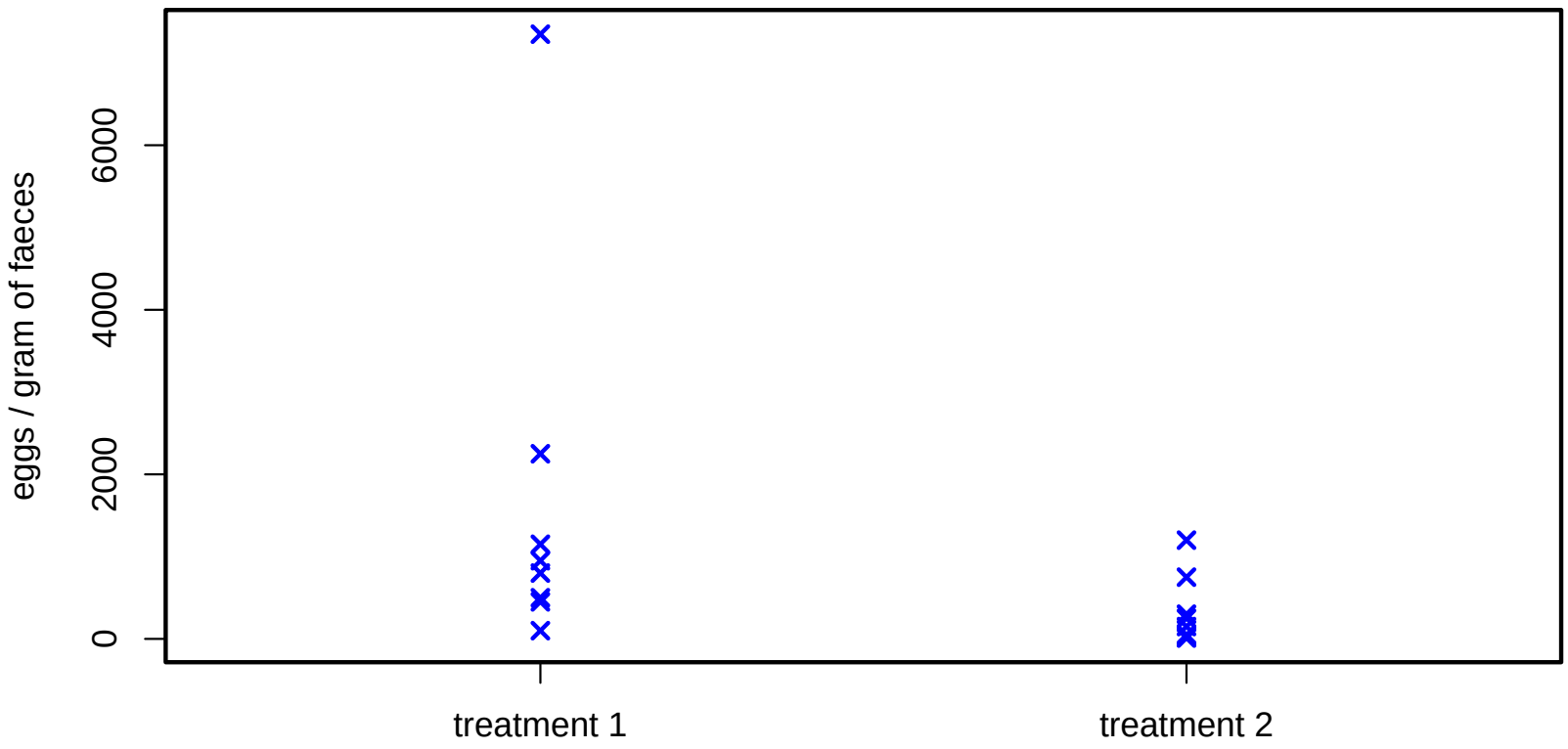


Exemple 10 - les courbes individuelles sont bien mal résumées par la courbe moyenne

Il est capital de montrer les courbes individuelles, ne serait-ce que pour s'assurer que la courbe moyenne les résument bien !



Exemple 11 - Devrait-on enlever la valeur extrême ?



Exemple 11 - Ne jamais enlever une observation juste parce qu'elle apparaît extrême !

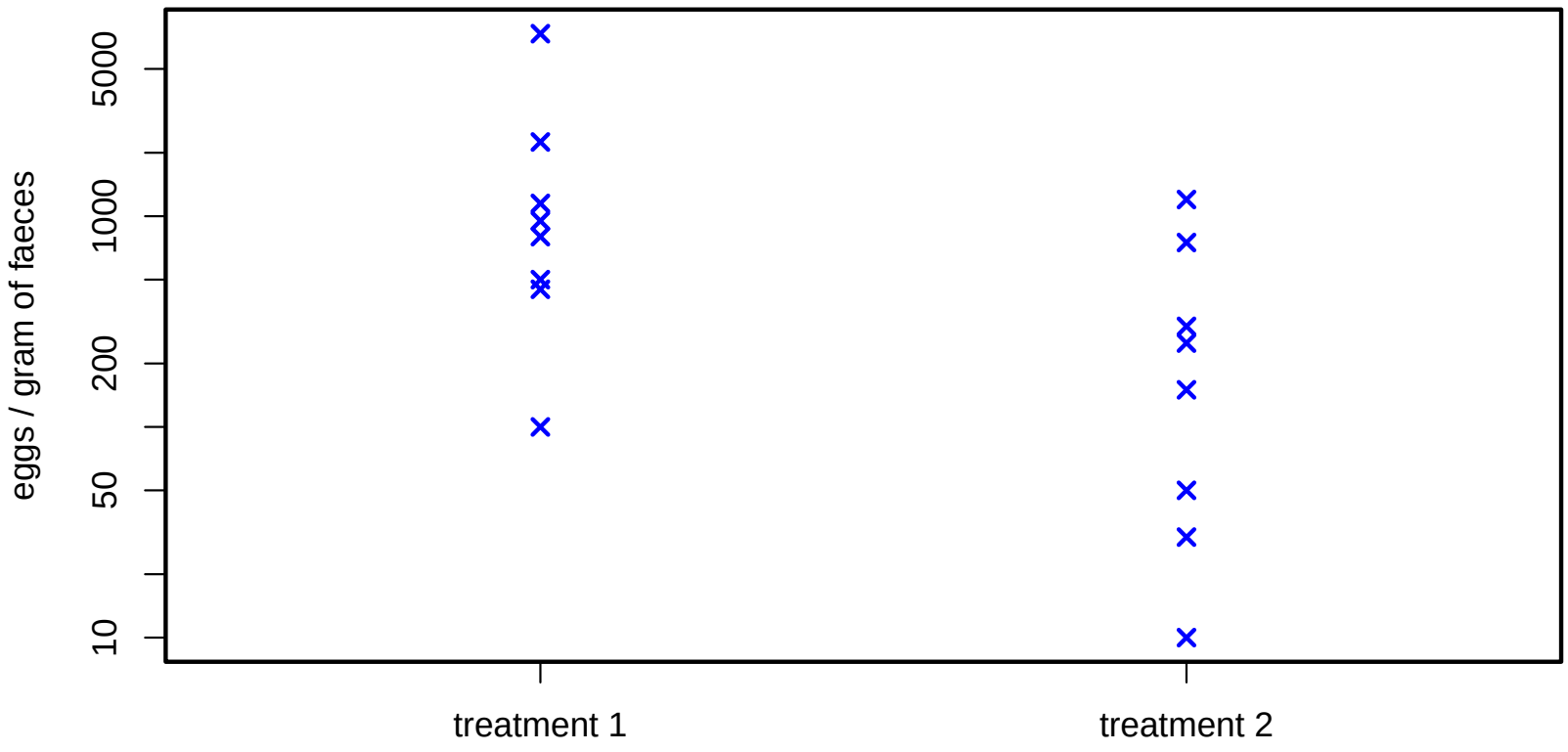
Devrait-on enlever la valeur extrême ?

Bien sûr que non !

- Les valeurs extrêmes peuvent être informatives
- Ce ne sont pas forcément des erreurs de mesures
- Elles peuvent correspondre au comportement différent d'un minorité
- Des données ne doivent pas être exclues pour la seule raison
- Parfois c'est juste que l'échelle sur laquelle on représente les données n'est pas adaptée ...

Exemple 11 - sur une échelle logarithmique

A la fois pour regarder et analyser les données, le choix de l'échelle est très importante.



A RETENIR !!!

- La moyenne et l'écart-type SD (ou l'erreur standard de la moyenne SEM) ne sont pas systématiquement de bons résumés de vos données. **Ne jamais les calculer d'emblée sans avoir bien regarder au préalable la distribution des données.**
- Il est parfois important de **changer d'échelle** (ex. transformation logarithmique) pour représenter et analyser les données, ou à si aucune transformation de variable n'est probante, d'utiliser des **statistiques de rang**.
- Une bonne figure doit être simple mais **informative**.
- On attend d'une figure est qu'elle donne un **bon résumé de la distribution des données brutes**.

Erreurs lors de l'interprétation des analyses statistiques

Erreurs lors de l'interprétation des analyses statistiques

La valeur-*p* : un problème significatif

Les méthodes statistiques pour jauger la pertinence d'un résultat expérimental sont attaquées de toutes parts. Résisteront-elles ?

En 1925, le généticien et statisticien britannique Ronald Fisher publiait un livre intitulé *Statistical Methods for Research Workers* (*Les Méthodes statistiques adaptées à la recherche scientifique*). Il n'avait *a priori* rien d'un best-

Ainsi naissait l'idée qu'une valeur-*p* inférieure à 0,05 apposait le sceau « statistiquement significatif » sur des recherches.

Aujourd'hui, presque un siècle plus tard, une valeur-*p* inférieure à 0,05 est considérée comme la référence pour jauger la qualité statistique d'une expérience. Elle ouvre aux cher-

Exemple 12 - ???

Essai randomisé sur 30 jeunes bovins visant à comparer 2 traitements anti-parasitaires sur la croissance de jeu bovins

Gains de poids à J30 :

- lot1 ($n = 15$): $m_1 = 25$ kg $\sigma_1 = 19$ kg
- lot2 ($n = 15$): $m_2 = 18$ kg $\sigma_2 = 17$ kg

Différence entre les 2 moyennes non significative ($p > 0.05$).

Les auteurs en concluent que les 2 traitements sont équivalents et qu'on peut donc utiliser le moins cher.

Exemple 12 - Abus d'interprétation de la p-value

Essai randomisé sur 30 jeunes bovins visant à comparer 2 traitements anti-parasitaires sur la croissance de jeu bovins

Gains de poids à J30 :

- lot1 ($n = 15$): $m_1 = 25$ kg $\sigma_1 = 19$ kg
- lot2 ($n = 15$): $m_2 = 18$ kg $\sigma_2 = 17$ kg

Différence entre les 2 moyennes non significative ($p > 0.05$).

On ne peut pas conclure à une équivalence à partir d'un tel essai, d'autant que les effectifs sont faibles et la variabilité importante.

Un test de signification (une p-value) ne permet jamais de montrer l'équivalence entre 2 traitements.

Rappel quant au principe d'un test de signification

μ_1 : moyenne théorique du traitement 1

μ_2 : moyenne théorique du traitement 2

Différence observée sur un échantillon: $d_{obs} = m_2 - m_1$

Hypothèse nulle H_0 : $\mu_1 - \mu_2 = 0$

Calcul de $p =$ probabilité, sous H_0, d'observer une différence au moins aussi grande que $d_{obs} = Pr(|d| \geq d_{obs}|H_0)$

Décision : rejet de H_0 si $p < 0.05$

MAIS ATTENTION !

On ne peut pas accepter H_0 si $p > 0.05$

On n'a pas calculé la probabilité de H_0 connaissant les données (bien plus difficile à calculer !)

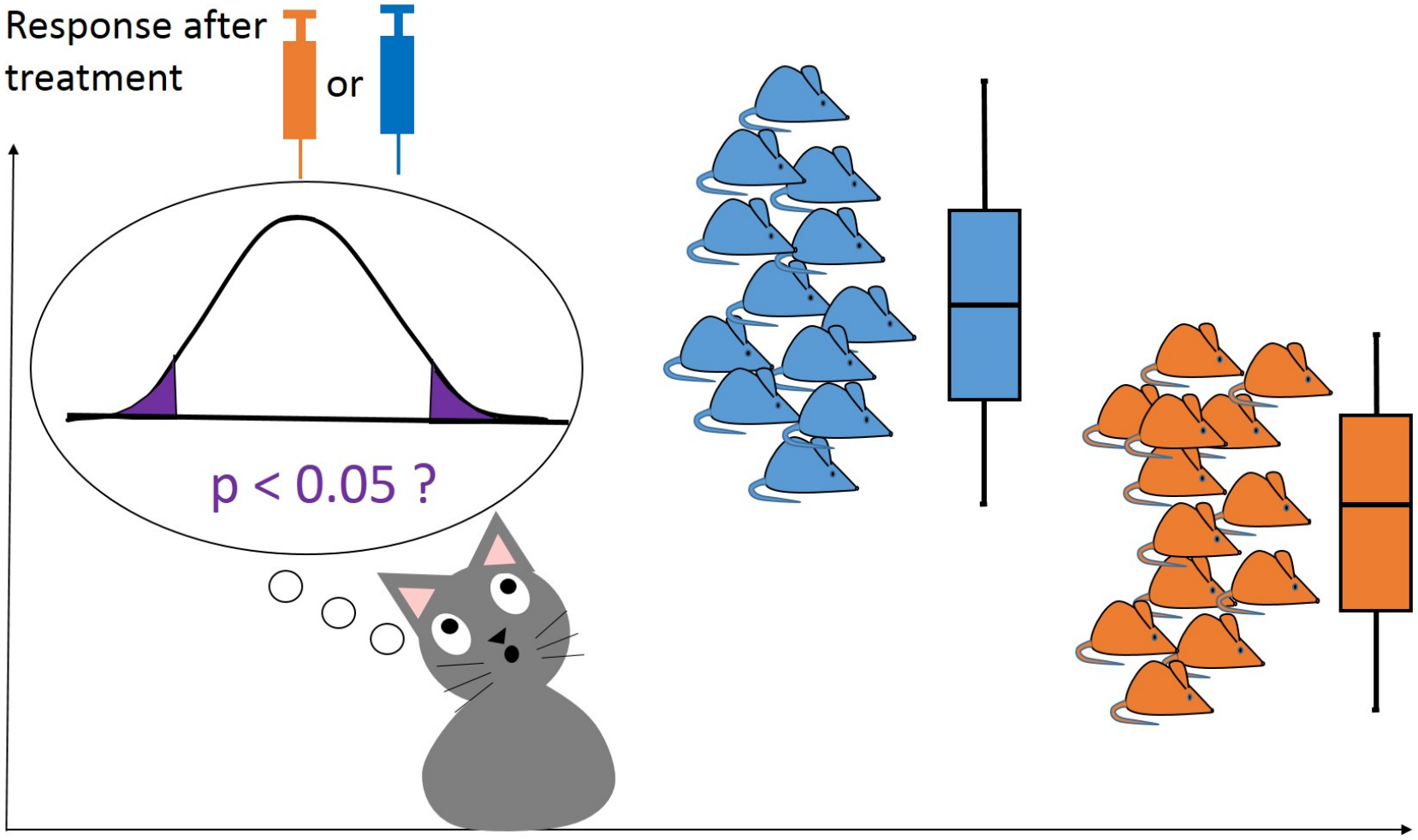
Exemple 12 bis - ???

Essai randomisé sur 300 jeunes bovins visant à comparer 2 traitements anti-parasitaires sur la croissance de jeu bovins.

Différence entre les 2 moyennes des gains de poids après 30 jours de traitement très significative ($p < 0.0001$).

Les auteurs en concluent qu'il existe une forte différence entre les gains de poids après 30 jours de chacun des deux traitements.

La p-value, résultat ultime d'une analyse statistique ?



Exemple 12 bis - Autre abus d'interprétation de la p-value

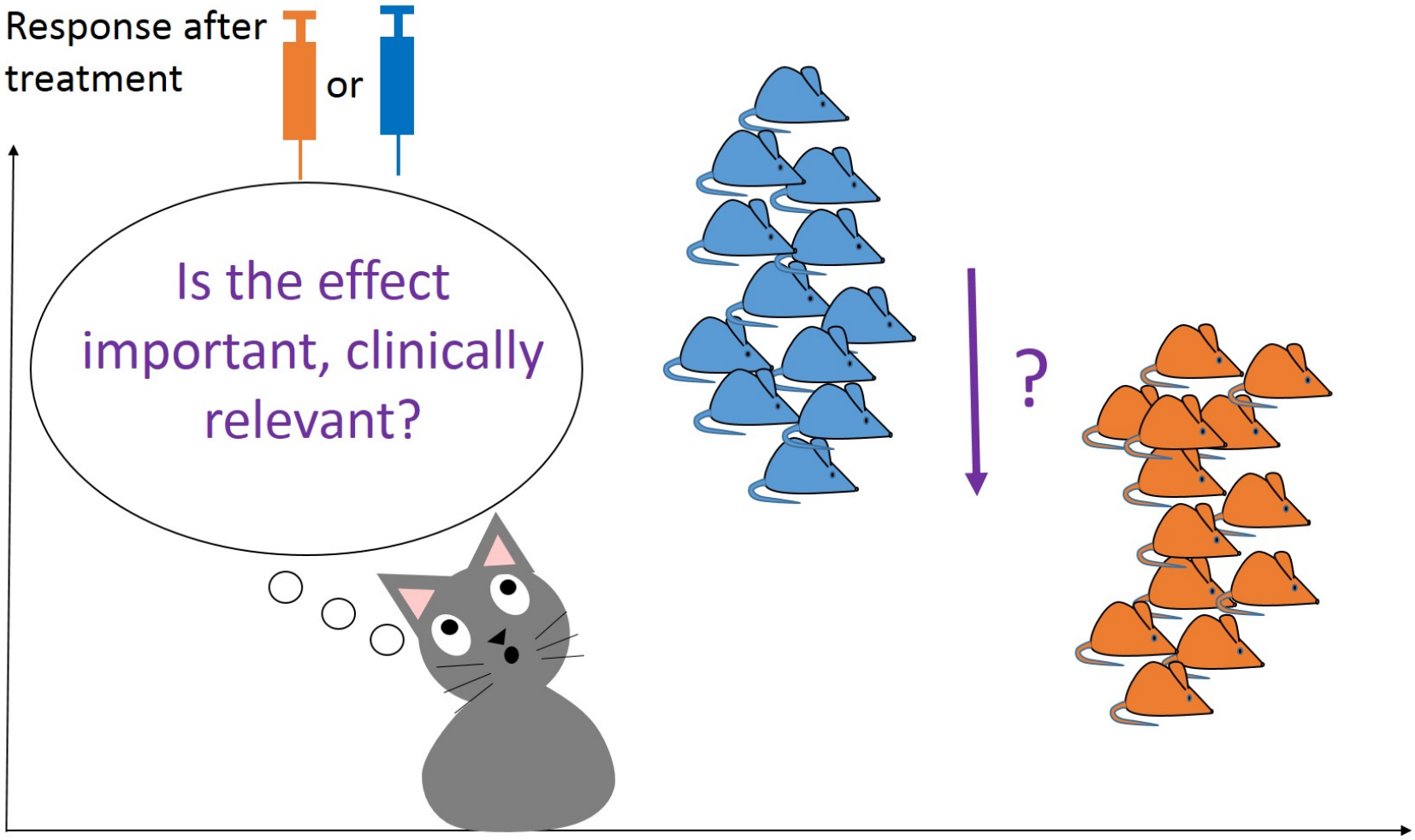
Essai randomisé sur 300 jeunes bovins visant à comparer 2 traitements anti-parasitaires sur la croissance de jeunes bovins.

Différence entre les 2 moyennes des gains de poids après 30 jours de traitement très significative ($p < 0.0001$).

On peut en conclure qu'on est quasi sûr qu'il existe une différence sur le gain de poids après 30 jours entre les deux traitements, mais cette différence n'est pas forcément importante ! Il convient de la regarder !

On ne doit jamais s'arrêter à la p-value. Ce n'est pas le résultat ultime d'une analyse statistique.

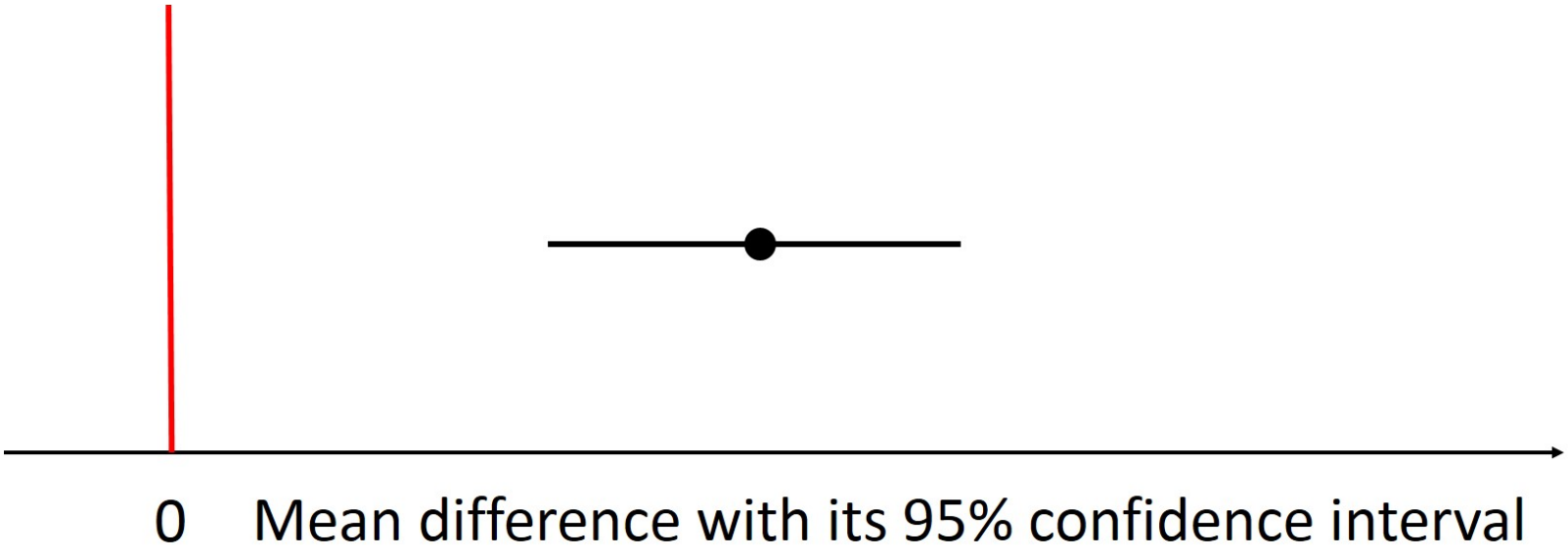
Ne souhaite-t-on pas plutôt caractériser un effet ?



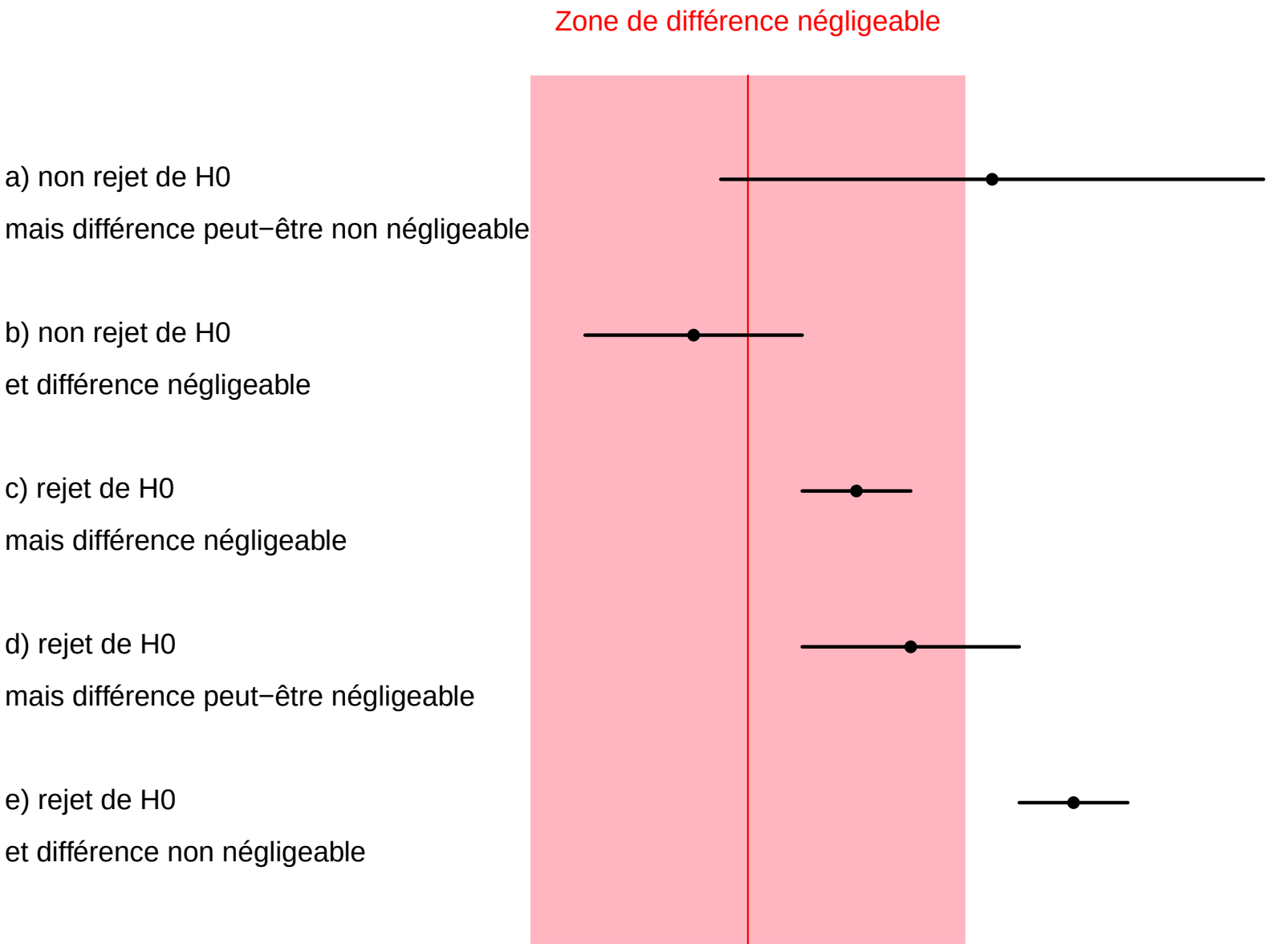
Estimation d'un effet et test de signification

Dans la plupart des tests classiques, la p-value est inférieure à 5% dès que l'intervalle de confiance autour de la différence estimée ne contient pas la valeur 0.

Mais l'estimation ponctuelle de la différence assortie de son intervalle de confiance est plus informative que la p-value !



Intervalle de confiance plus informatif que la p-value



The ASA statement on p-values

Article publié en juin 2016 et cité plus de 4500 fois en janvier 2022 (chiffre issu de Scholar Google) exprimant un **consensus des statisticiens au sujet de la p-value**.

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA statement on p-values: context, process, and purpose. The American Statistician, 70(2), 129-133.

1. P-values can indicate how compatible the data are with a specified statistical model.

- Plus la p-value est petite et plus l'incompatibilité statistique entre les données et l'hypothèse nulle est grande.
- On peut voir la p-value comme un indicateur de discordance entre les données et l'hypothèse nulle.

2. P-values do not measure the probability that the studied hypothesis is true.

La p-value ne doit surtout pas être interprétée comme la probabilité de l'hypothèse nulle connaissant les données, même si cela est très tentant.

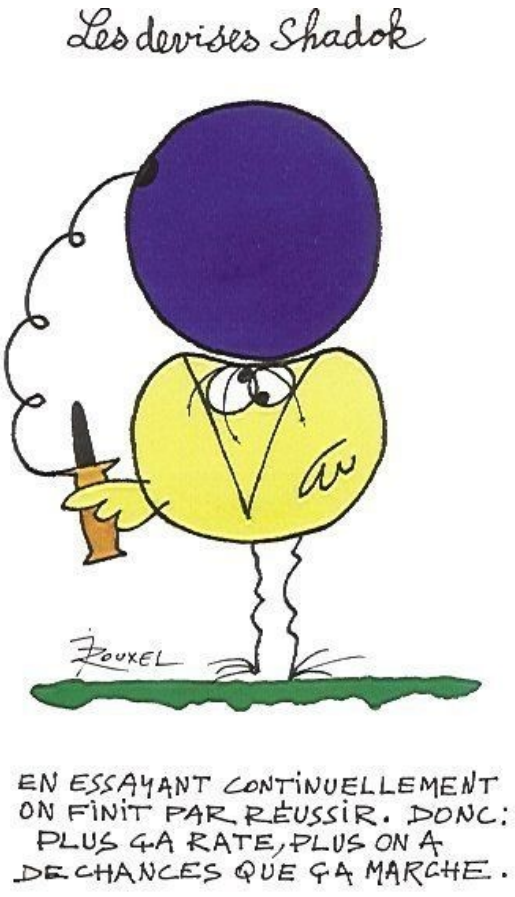
On ne peut pas inverser les probabilités aussi facilement !

3. Scientific conclusions and decisions should not be based only on whether a p-value passes a specific threshold.

- Actuellement les scientifiques donnent souvent trop de poids à la p-value et au résultat du test en terme de différence significative ou non, parfois sans même regarder la différence estimée.
- Il convient plutôt de considérer le **test comme un garde fou, nous empêchant d'interpréter hâtivement une différence qui ne serait pas significative**.

4. Proper inference requires full reporting and transparency.

- Les résultats de **tous les tests réalisés doivent être reportés**, et non seuls les résultats significatifs.
- En moyenne dans tous les cas où H_0 est vraie, une fois sur 20 on a $p < 0.05$.
  A force de chercher on finit par trouver !



5. A p-value does not measure the size of an effect or the importance of a result.

- Une p-value petite n'implique pas forcément la mise en évidence d'une différence d'intérêt biologique.
- Une différence importante peut ne pas apparaître significative du fait du manque de puissance de l'analyse (par ex. en cas d'effectifs faibles).

Il est capital, lorsque cela est possible, **d'interpréter in fine l'estimation de la différence (estimation ponctuelle et intervalle de confiance)**.

6. By itself, a p-value dose not provide a good measure of evidence regarding a hypothesis.

Ne jamais utiliser un test d'hypothèse pour montrer pour montrer une hypothèse et en particulier pour montrer une équivalence mais privilégier les tests d'équivalence basés sur les intervalles de confiance dans ce cas.

Conclusion par rapport aux 3R

Réduction des effectifs, Raffinement des manipulations, Remplacement de l'expérimentation

- Il vaut mieux une bonne expérimentation (exploitable) qu'une mauvaise !
- On se doit dans ce domaine d'être le plus rigoureux et vigilant possible, à toutes les étapes de l'expérimentation !