Fiche d'aide à la lecture critique des parties statistiques des articles

Cette fiche liste différentes questions pour vous aider à lire de façon critique les parties statistiques des articles scientifiques et/ou pour bien rédiger les analyses statistiques de base que vous réalisez dans le cadre de votre thèse d'exercices par exemple.

Marie Laure Delignette-Muller – 31 mars 2026

I Les statistiques descriptives proposées (**représentations graphiques** et **paramètres statistiques** utilisés pour résumer les données) vous paraissent-elles bien choisies et de bonne qualité ?

- Représentent-elles bien les **résultats obtenus sur le(s) critère(s) de jugement principal(aux)** ? Sont-elles suffisamment **informatives**, résumant bien les données brutes ?
- Les **légendes** des graphes et/ou tables indiquent-elles de façon claire ce qui a été représenté / calculé ?
- Les figures permettent-elles de visualiser la **forme des distributions** (pour les variables quantitatives) et ainsi de guider le choix des calculs statistiques ultérieurs (choix des paramètres descriptifs – choix entre tests paramétriques ou non paramétriques) ? Voir exemple en fin de document[1].
- Lorsque des **tests non paramétriques sont choisis, des statistiques de rang sont-elles bien aussi utilisées** pour résumer les données (médiane, quartiles), et non des moyennes et SD (ou SEM) ?
- Les figures donnent-elles une idée de la **variabilité sur les données individuelles** et/ou de **l'incertitude sur les paramètres statistiques résumés** ?
- Dans le cas de **données appariées** (ou plus généralement dépendantes), **l'appariement** (ou la dépendance) **apparaît-il bien sur les figures** (voir exemple en fin de document[2]).
- Les figures utilisent-elles une **échelle appropriée** (par ex. choix d'une échelle logarithmique lorsque les valeurs observées varient sur plusieurs ordres de grandeur) ?

II Dans les études de type essai clinique, est-ce qu'un **critère d'efficacité** a été défini,
par ex. une différence minimale, entre le traitement dont on veut prouver l'efficacité et le placebo, sur le critère de jugement, au bout d'un temps donné de traitement, ou autre ex., le retour de ce critère de jugement à des valeurs dans une gamme considérée comme normale au bout d'un temps donné de traitement, …) ? Si oui ce critère d'efficacité vous paraît-il **pertinent d'un point de vue biologique / clinique**, et non défini *a posteriori* à partir des résultats observés (comme c'est parfois le cas).

III Les **méthodes d'analyse** (**tests** ou autres calculs réalisés) sont-elles bien choisies ?

- **Permettent-elles bien de répondre aux questions posées** dans les objectifs de l'article ? Auriez-vous choisi les mêmes méthodes où auriez-vous utilisé des méthodes alternatives vous paraissant plus directes ou plus appropriées ? Q*uand les méthodes proposées sont différentes de celles qu'on aurait naturellement utilisé, cela peut être suspect.*
  Voici un exemple : parfois quand le test qui compare deux traitement entre eux ne donnent pas de différence significative, les auteurs n'en donnent pas le résultat, et préfèrent montrer que l'un diffère significativement du placebo et pas l'autre, et en concluent abusivement que les deux diffèrent entre eux.
- Les **conditions d'utilisation des tests vous semblent-elles vérifiées**, d'après les informations fournies dans l'article ? Notamment le choix entre une approche paramétrique ou non paramétrique est-il argumenté autrement que par un test de normalité (dont on rappelle que la seule conclusion possible est le rejet de la normalité, et non son acceptation), et le choix entre variances égales ou inégales autrement que par un test de comparaison de variances (qui ne permet pas non plus de conclure à l'égalité des variances – cf. rappel qui suit sur l'interprétation des p-values[3] si vous n'avez plus les idées claires sur ce point).

IV Les **résultats** des analyses statistiques **sont-ils correctement présentés et interprétés** ?
Notamment **l'interprétation des p-values** respecte-t-elle bien les préconisations de l' **« ASA statement on p-values** » (résumé pour rappel en fin de document[3]).

- Les **effets estimés** sont-ils bien décrits, qu'ils soient significatifs ou non, et assortis de leurs intervalles de confiance lorsque c'est possible. N'oubliez pas qu'il est très incomplet, voire trompeur, de ne donner que la p-value dans une conclusion, sans indication de la taille d'effet estimée, et éventuellement de son incertitude (même si c'est encore trop courant).
- Les **conclusions sont-elles suffisamment prudentes**, notamment **en cas de non rejet de H_0**, qui doit s'interpréter comme **absence de preuve d'effet/corrélation**, et non preuve de l'absence d'effet/corrélation.
- Les **p-values sont-elles données avec un nombre de chiffres significatifs raisonnable**. Seul son ordre de grandeur importe, donc un voire deux chiffres significatifs sur une p-value suffisent. En donner plus conduit implicitement à donner trop d'importance aux p-values.
- Le **nombre de tests réalisés dans l'article vous semble-t-il raisonnable** (pas trop important) ? Souvenez-vous que plus on cherche, plus on a de chance de trouver. S'il n'y a absolument aucun effet à détecter, un test sur 20 détectera tout de même un effet (5% des cas).

---------- ***Compléments*** ----------

[1]Petit rappel au sujet de l'utilisation des paramètres statistiques classiques, moyenne, écart type (SD) et erreur standard de la moyenne (SEM).

- Parfois les données sont résumées par des moyennes et des écarts types alors même que la distribution est très dissymétrique. Il n'est pas rare de voir pour une variable par nature positive, des résultats exprimés en mean ± SD avec un SD plus grand que la moyenne, ce qui n'a pas de sens puisque l'intervalle mean ± 2 SD (2 fois plus grand) devrait représenter l'intervalle de fluctuation à 95% si la loi était normale (et mean ± SD l'intervalle de fluctuation à 68%). On trouve aussi fréquemment ce **même problème sur des figures représentant des moyennes sous forme de bâton, avec une barre d'erreur au-dessus du bâton** (cf. illustration en Figure 1a et b).
- Il est parfois utile de changer d'échelle (notamment de faire une transformation logarithmique) avant de pour pouvoir mieux représenter les données et utiliser une approche paramétrique (cf. Figure 1 c et d).
- Pour rappel, l'intervalle mean ± 2 SEM représente l'intervalle de confiance à 95% sur la moyenne (notion d'incertitude sur la moyenne), à condition que le théorème de l'approximation normale s'applique (et l'intervalle mean ± SEM l'intervalle de confiance à 68% sur la moyenne).

Il convient donc, pour bien rédiger les résultats statistiques :

- de **ne pas utiliser les paramètres statistiques classiques (moyennes, SD et SEM) en dehors de leurs conditions d'utilisation** (on utilise les statistiques de rang dans les autres cas, notamment pour des distributions dissymétriques)
- de se souvenir que sous leurs conditions d'utilisation, **SD décrit la variabilité autour de la moyenne**, alors que **SEM décrit l'incertitude sur la moyenne**,
- d'éviter dans tous les cas la notation mean ± SD (préférer mean (SD)) ainsi que le graphe classique associé.
- d'éviter dans tous les cas la notation mean ± SEM et de donner plutôt l'intervalle de confiance à 95% sur la moyenne (mean ± 2 SEM si le théorème de l'approximation normale s'applique).

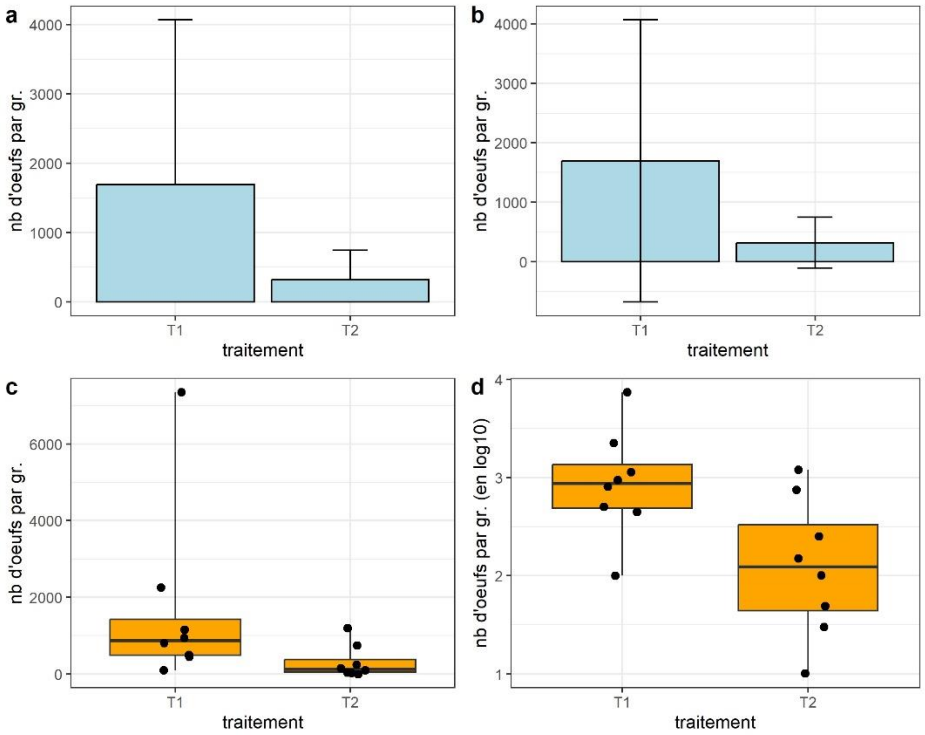

Figure 1 : Différentes représentations, plus ou moins appropriées, des distributions observées d'une variable communément utilisée en parasitologie (résultat d'une coproscopie exprimé en nombre d'œufs du parasite par gramme de feces) à l'issue de deux traitements différents T1 et T2. a) représentation de « mean + SD », b) représentation de « mean $\pm$ SD », c) diagrammes en boîte et points observés et d) diagrammes en boîte et points observés sur le logarithme décimal de la variable observée.

[2]Importance d'une représentation graphique des données la plus informative possible

Prenons un exemple dans lequel on a des séries appariées correspondant à deux traitements testés successivement sur chaque animal. La figure 2a illustre le manque d'information que l'on aurait sans faire apparaître cet appariement par rapport à la figure 2b plus complète/informative.

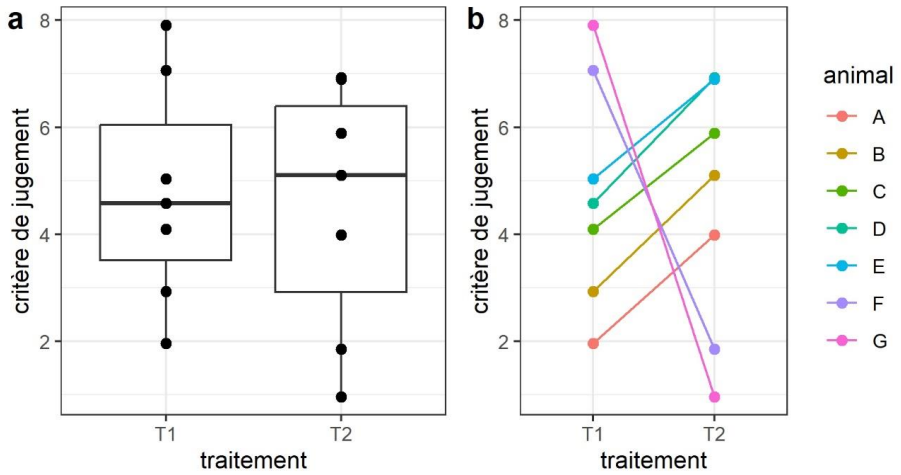

Figure 2 : Différentes représentations d'un critère de jugement observé sur deux séries appariées (correspondant à la réponse à deux traitements), a) en diagrammes en boîte avec les observations individuelles sans faire apparaître l'appariement, b) en points reliés pour faire apparaître l'appariement (facteur animal).

[3]Résumé des consignes du « ASA statement on p-values » :

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA statement on p-values: context, process, and purpose. *The American Statistician*, *70*(2), 129-133. https://www.amstat.org/asa/files/pdfs/p-valuestatement.pdf

1. « P-values can indicate how compatible the data are with a specified statistical model. »
Plus la valeur de p est petite et plus l'incompatibilité statistique entre les données et l'hypothèse nulle est grande. On peut voir la valeur de p 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 valeur de p ne doit surtout pas être interprétée comme la probabilité de l'hypothèse nulle connaissant les données, même si cela est très tentant et trop souvent fait plus ou moins explicitement. On ne peut pas inverser les probabilités aussi facilement (nécessité d'utiliser le théorème de Bayes pour cela) !

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 valeur de p et au résultat du test en termes 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 ($p > 0.05$ à interpréter comme « lack of evidence », c'est-à-dire manque de preuve).

4. « Proper inference requires full reporting and transparency. »

Les résultats de tous les tests réalisés doivent être reportés, et non les seuls résultats significatifs. En moyenne dans tous les cas où H_0 est vraie (différence nulle / corrélation nulle), une fois sur 20 on obtient $p < 0.05$. Donc sur un très grand nombre de tests réalisés, on en a toujours qui conduisent au rejet de H_0 !

5. « A p-value does not measure the size of an effect or the importance of a result. »
Une valeur de p petite n'implique pas forcément la mise en évidence d'une différence d'intérêt biologique ET une différence importante peut ne pas apparaître significative du fait du manque de puissance de l'analyse (par ex. en cas d'effectifs faibles). Dans tous les cas, il est capital, d'interpréter in fine l'estimation de l'effet étudié (différence / corrélation), en donnant des estimations ponctuelles, et des intervalles de confiance lorsque c'est possible (plus compliqués de donner les intervalles de confiance lorsque des statistiques non paramétriques sont utilisées).

6. « By itself, a p-value dose not provide a good measure of evidence regarding a hypothesis. »
Il ne convient pas d'utiliser un test d'hypothèse pour montrer une hypothèse, et en particulier pour montrer une équivalence : les tests d'équivalence basés sur les intervalles de confiance sont à utiliser dans ce cas.
Rappel du principe d'un **test d'équivalence** :
- on définit une zone d'équivalence sur des critères biologiques / cliniques (« quelle différence maximum est considérée comme négligeable ? »), et

- on conclut à l'équivalence si l'intervalle de confiance sur la différence observée est entièrement contenu dans cette zone.