Recherche des substances émergentes dans les eaux et intéressant la santé publique et l'environnement

-

Programme de recherche IMHOTEP

Inventaire des Matières Hormonales et Organiques en Traces dans les Eaux Patrimoniales et Potabilisables

Annexe 4 du rapport final : rapport d'Ephesia Consult – Comparaison des méthodes de traitement des valeurs sous la LQ

Juin 2018

Comparaison des méthodes de traitement des valeurs sous la limite de quantification

Dimitri D'Or, Ephesia Consult

9 juin 2017

1 Introduction

Cette note a pour objectif de comparer trois méthodes de traitement des valeurs sous la limite de quantification (LQ) pour le calcul des statistiques d'un échantillon. Ces trois méthodes sont les suivantes :

- **LQ/2** : les valeurs sous la LQ sont remplacées par la moitié de celle-ci.
- **IGNORE** : les valeurs sous la LQ sont simplement ignorées.
- **NADA** : les statistiques de l'échantillon sont calculées par la méthode NADA (Helsel, 2005).

La comparaison portera sur l'estimation de la moyenne de l'échantillon. La première section décrit la construction de l'expérience et la seconde expose les résultats.

2 Construction de l'expérience

Pour comparer les moyennes estimées par les différentes méthodes, 1000 échantillons de 100 individus sont générés aléatoirement selon une distribution lognormale dont la moyenne et la variance du logarithme des valeurs valent respectivement 0 et 1.

Sur les échantillons générés, les valeurs sous la LQ sont remplacées par LQ/2 pour la méthode *LQ/2* et par des non valeurs (NA) pour la méthode *IGNORE*. Pour la méthode *NADA*, on accompagne simplement l'échantillon d'un vecteur de même taille dont les valeurs sont à 1 pour les valeurs sous la LQ et à 0 pour les autres.

Les moyennes sont ensuite estimées en calculant simplement la moyenne expérimentale pour les méthodes *LQ/2* et *IGNORE*, tandis que pour la méthode *NADA*, la méthode *cenmle* du package R NADA [1] est utilisée. La moyenne est estimée par Maximum de Vraisemblance (MLE en anglais). Avec cette méthode, l'estimateur est donc construit de façon à ne pas être biaisé, mais les individus ne sont pas "corrigés" ; on ne leur attribue pas de valeur représentant le fait qu'ils sont en-dessous de la LQ.

Pour comparer les moyennes, un test classique de comparaison de moyennes de Student est utilisé. [2] Pour chacun des 1000, on enregistre les p-valeurs des comparaisons entre méthodes

1. `https://cran.r-project.org/web/packages/NADA/NADA.pdf`
2. Bien que les données soient appariées dans les faits puisque les valeurs au-dessus de la LQ sont identiques pour toutes les méthodes, il n'est pas possible d'utiliser un test pour données pairées avec la méthode NADA puisque les individus ne sont pas "corrigés".

prises 2 à 2. La p-valeur est la probabilité d'obtenir la même valeur (ou une valeur encore plus extrême) du test si l'hypothèse nulle (ici, c'est l'hypothèse d'égalité des moyennes : $H_0 : \mu_1 = \mu_2$) était vraie. La p-valeur est généralement comparée à une valeur seuil prédéterminée (niveau de confiance, alpha, souvent 0.01 ou 0.05). Si la p-valeur est inférieure à ce seuil, on rejette l'hypothèse nulle en faveur de l'hypothèse alternative (ici, $H_1 : \mu_1 \neq \mu_2$), et le résultat du test est déclaré *statistiquement significatif*. Dans le cas contraire, si la p-valeur est supérieure au seuil, on ne rejette pas l'hypothèse nulle, et on ne peut rien conclure quant aux hypothèses formulées. En d'autres termes, les moyennes ne sont pas significativement différentes, sans pour autant qu'il soit sûr qu'elles sont égales.

L'expérience est répétée pour des valeurs de LQ correspondant respectivement à des proportions allant de 10 à 90% de valeurs sous la LQ, par pas de 10%.

Pour synthétiser les résultats, on calculera la proportion de tests à résultat significatif pour chaque proportion de valeurs sous la LQ. On donnera aussi la distribution des p-valeurs pour chaque niveau, ainsi que la distribution des moyennes.

3 Résultats

La Figure 1 montre les boxplots des p-valeurs en fonction de la proportion de valeurs sous la LQ pour les trois comparaisons. Dans les trois cas, les p-valeurs diminuent exponentiellement lorsque la proportion de valeurs sous la LQ augmente. La proportion de p-valeurs passant sous le niveau de confiance α =0.01 augmente en retour (Figure 2). Pour la comparaison entre *LQ/2* et *NADA*, la médiane passe sous la valeur de α =0.01 à partir de 40% de données sous la LQ, indiquant ainsi que les tests donneront des différences significatives dans plus de 50% des cas. Pour les deux autres comparaisons, ce phénomène se produit dès 30% et 20% pour *LQ/2* vs. *IGNORE* et *IGNORE* vs. *NADA*, respectivement.

Les boxplots des moyennes sont donnés à la Figure 3. Les distributions, bien superposées pour des proportions faibles de valeurs sous la LQ, se séparent lorsque cette proportion augmente. Les moyennes de la méthode *IGNORE* augmentent plus vite que celles des méthodes *NADA* et *LQ/2*, respectivement. Lorsque la proportion de valeurs sous la LQ dépasse les 80%, les moyennes de la méthode *NADA* dépassent les deux autres.

4 Discussion

La méthode **LQ/2** produit systématiquement des valeurs plus faibles de moyennes. Cela est dû au fait que toutes les valeurs sous la LQ sont remplacées par LQ/2. Au fur et à mesure que la proportion de valeurs sous la LQ augmente, LQ/2 s'éloigne des autres valeurs de la distribution et "tire" la moyenne vers le bas. On observe donc un **biais de l'estimation de la moyenne vers les valeurs faibles**. Ce biais s'accentue avec la proportion de valeurs sous la LQ.

Avec la méthode **IGNORE**, c'est l'inverse qui se produit : en ignorant de plus en plus de valeurs faibles, et en ne retenant donc que des valeurs élevées, la moyenne est "tirée" vers le haut. On observe donc un **biais de l'estimation de la moyenne vers les valeurs élevées**. Ce biais s'accentue également avec la proportion de valeurs sous la LQ.

La méthode **NADA a été conçue pour corriger ces deux biais**. Elle produit donc des valeurs estimées intermédiaires, sauf lorsque les proportions de valeurs sous la LQ sont égales ou

dépassent les 80%. Dans ces cas-là, le nombre de données au-dessus de la LQ devient insuffisant et l'estimateur perd de sa précision. Elle semble donc devoir être recommandée.

Si, toutefois, des tests d'hypothèses doivent être réalisés sur les données, l'utilisation de l'approche *NADA* rend la mise en oeuvre plus compliquée. En effet, la plupart des logiciels utilisent les échantillons en tant que tels comme variables d'entrée. Or, comme les valeurs sous la LQ ne sont pas remplacées explicitement par une valeur donnée, il est impossible de prendre en compte un échantillon "corrigé" par *NADA*. Il serait alors nécessaire d'estimer les moyennes et variances avec *NADA* et de calculer explicitement la statistique de test avant de la comparer à la valeur seuil dans la table de Student. Tout cela nécessite une compétence statistique d'un niveau supérieur à celui nécessaire pour appliquer les fonctions de test disponibles dans les logiciels.

Dans le présent document, les valeurs nulles éventuellement générées sont considérées comme des valeurs sous la LQ. En opérationnel, ces valeurs nulles peuvent correspondre à des valeurs non détectées et doivent être prises en compte en tant que telles, sinon la moyenne sera surestimée (que ces valeurs soient ignorées ou remises à LQ/2).

5 Conclusion

En conclusion, les points suivants doivent être retenus :

1. La méthode *IGNORE* produit une surestimation systématique de la moyenne tandis que la méthode *LQ/2* produit une sous-estimation systématique de celle-ci.
2. Les méthodes *NADA* et *LQ/2* ne montrent pas de différences significatives en-deça de 20% de valeurs sous la LQ. Entre 20% et 40% de valeurs sous la LQ, moins de 50% des tests donnent des différences significatives entre les deux méthodes. Par contre, à partir de 50% de valeurs sous la LQ, les tests donnent des différences significatives dans plus de 90% des cas et les méthodes ne sont donc plus du tout équivalentes.
3. Au-dessus de 80% de valeurs sous la LQ, la méthode *NADA* ne donne plus d'estimations fiables de la moyenne.
4. Quelle que soit la méthode choisie, les vrais zéros (valeurs non détectées) doivent être conservés comme tels sous peine de surestimer la moyenne.
5. Dans un but d'homogénéité et de cohérence, même si les proportions de valeurs sous la LQ varient d'une à l'autre, il est conseillé de n'utiliser qu'une seule et même méthode sur l'ensemble des variables à traiter.

En résumé, en-dessous de 20% de valeurs sous la LQ, la méthode *LQ/2* peut-être utilisée sans perte par rapport à *NADA*. Cette dernière est recommandée si la proportions de valeurs sous la LQ dépasse les 20%. Elle atteint toutefois ses limites lorsque cette proportion dépasse les 80%. Elle peut toutefois demander des compétences statistiques avancées dans le cadre d'une mise en oeuvre dans le contexte de tests d'hypothèses et ne permet pas de représentation graphique correcte de la distribution puisque les individus ne sont pas "corrigés".

6 Références

Helsel, D., 2005. Nondetects And Data Analysis : Statistics for censored environmental data. Wiley.

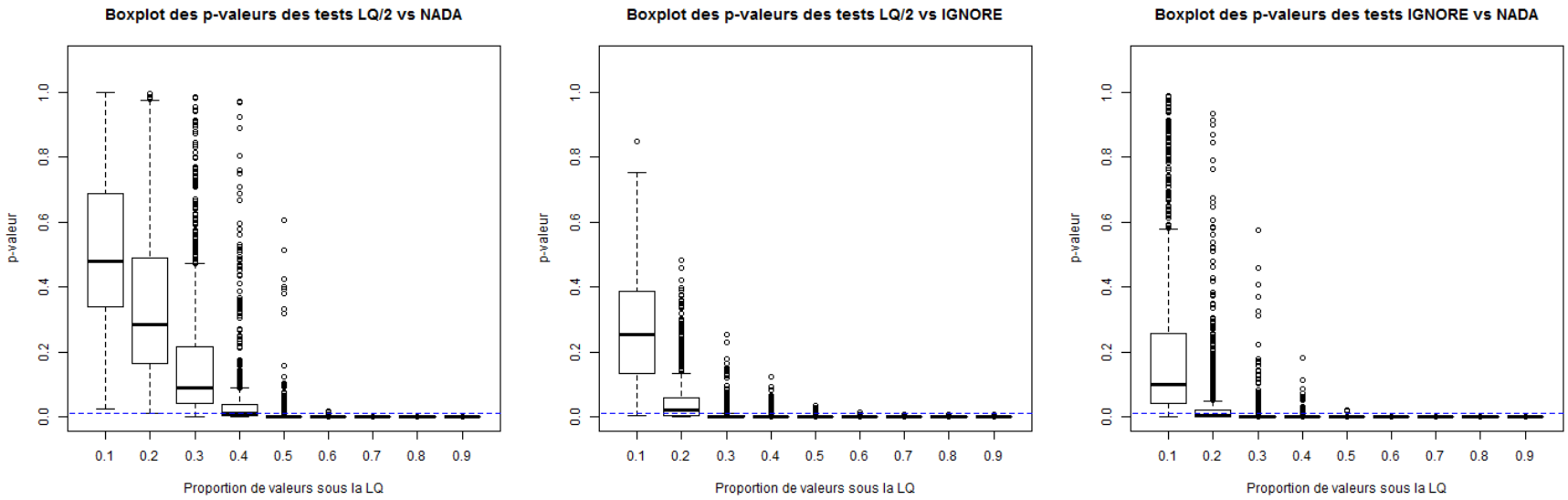

FIGURE 1 – Boxplots des p-valeurs en fonction de la proportion de valeurs sous la LQ pour les trois comparaisons. Le trait pointillé bleu représente le niveau de confiance $\alpha = 0.01$. Le trait central de la boite représente la médiane, les limites inférieures et supérieures de la boite les premier et troisième quartile (quantiles à 25% et 75 %), et les moustaches sont respectivement égales à $min(max(x), Q_3 + 1.5 * IQR)$ et $max(min(x), Q_1 \breve{} 1.5 * IQR)$ ou IQR est l'écart inter-quartile $Q_3 - Q_1$.

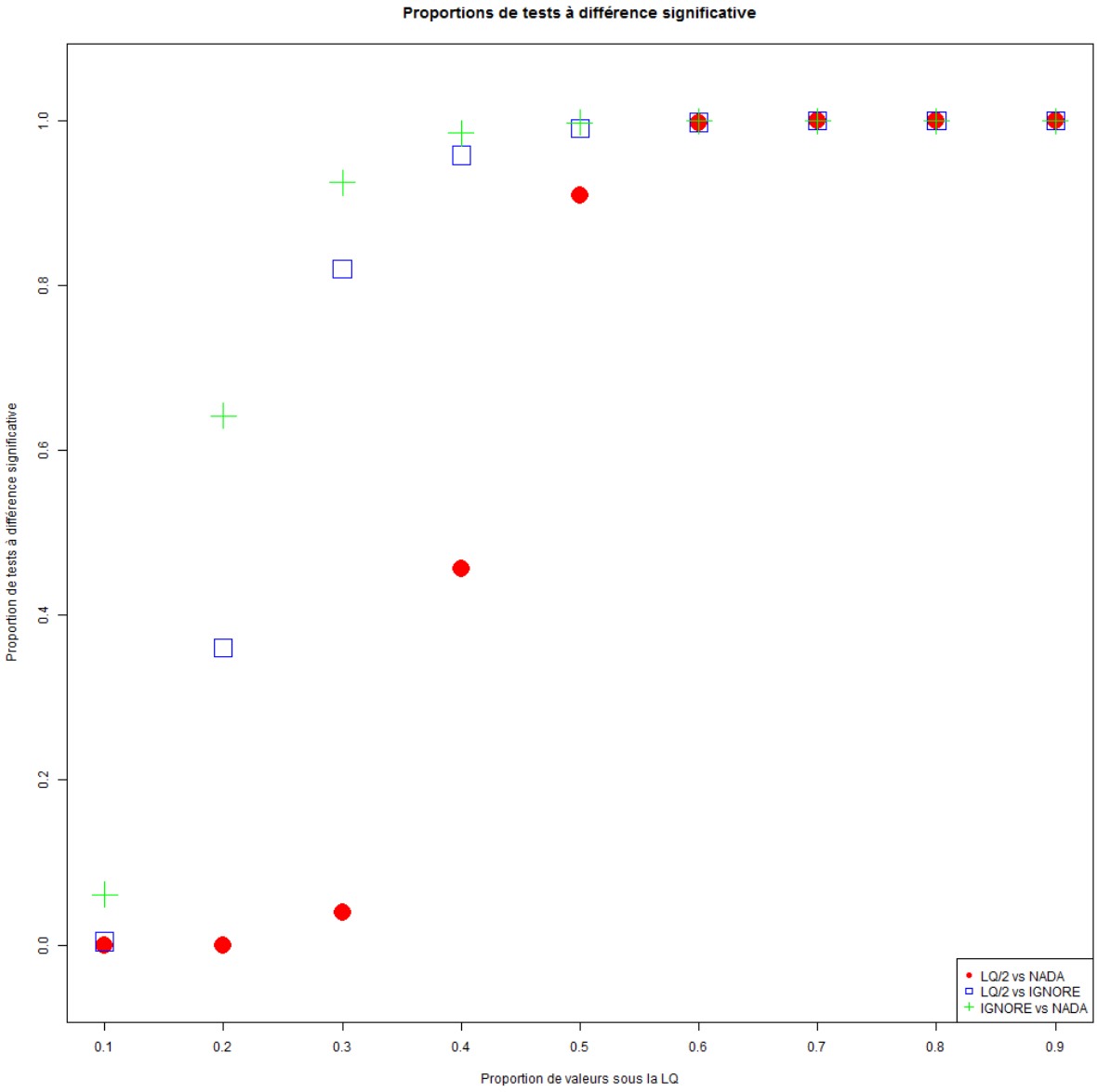

FIGURE 2 – Proportion de p-valeurs sous le niveau de confiance $\alpha = 0.01$ en fonction de la proportion de valeurs sous la LQ pour les trois comparaisons.

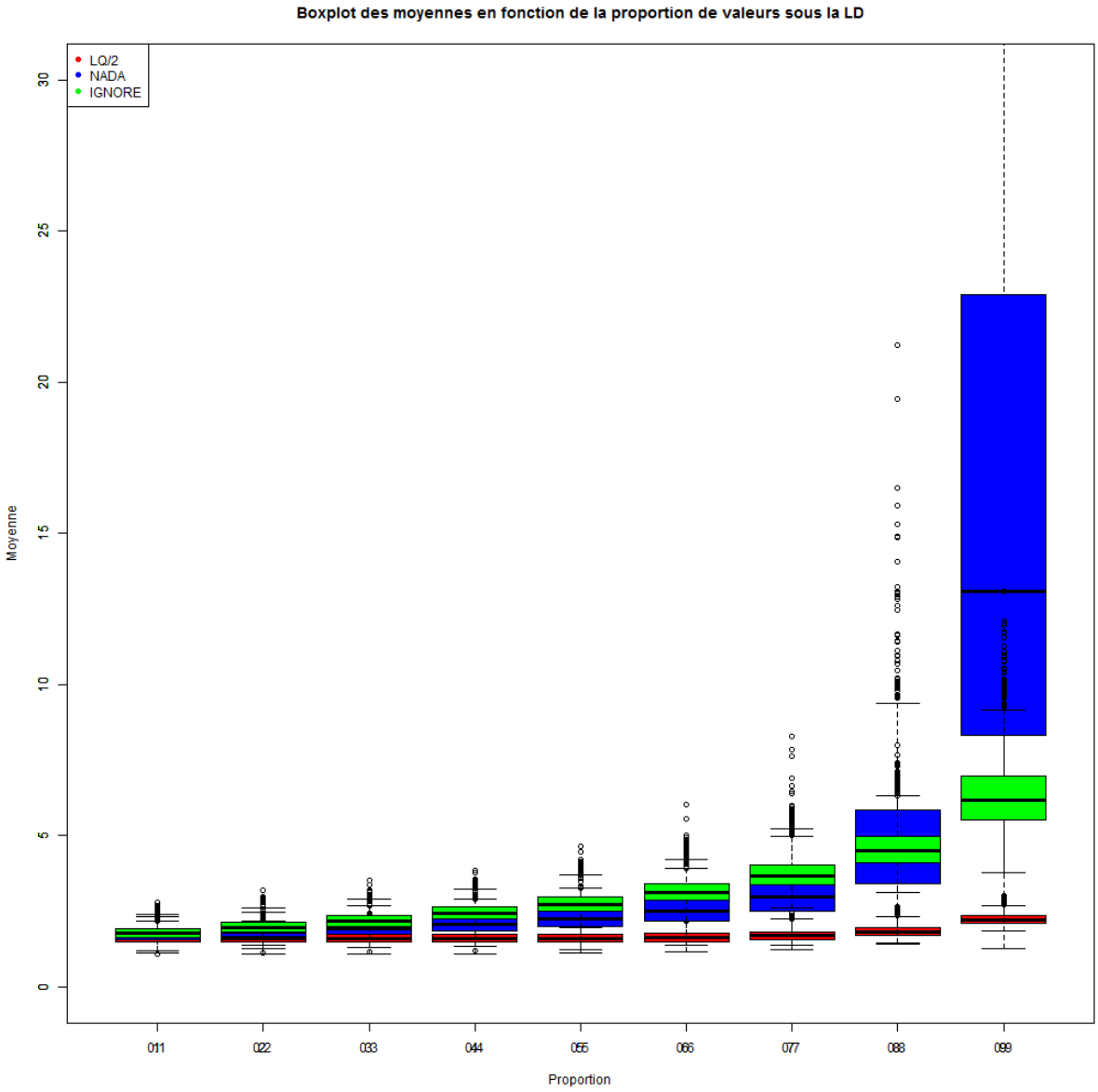

FIGURE 3 – Boxplot des moyennes en fonction de la proportion de valeurs sous la LQ pour les trois méthodes.