Outliers ou données extrêmes, comment les détecter ? que faire de ces observations ?

Claire Chabanet, Fabrice Dessaint

26 mai 2015

1 Objectif de la fiche

L'utilité de la rédaction de cette fiche est née de la constatation d'une hétérogénéité des pratiques concernant la façon de gérer les points dits « aberrants ». De la confrontation de la diversité des pratiques, nait une question : « Que faire face à un/des points extrêmes ? »

Nous écartons ici les cas où l'on s'intéresse spécifiquement aux points extrêmes (crues, grandes marées, fraudes,...). Nous nous intéressons exclusivement aux cas de l'utilisation des méthodes paramétriques les plus courantes : le test de student, l'analyse de variance ou la régression linéaire. Le cadre étant posé, la question est : « comment détecter les points extrêmes et que faire de ces points ? »

Des éléments de réponse à ces questions sont donnés dans la section 3 et la section 5 « Que faire en cas de détection de points extrêmes ? ». Cette dernière constitue le cœur de cette fiche. Cependant, cette question amène naturellement à évoquer des domaines différents du champ des statistiques, comme les conditions d'utilisation des tests statistiques, les transformations de variables, les tests de normalité, ..., qui peuvent *a priori* paraître déconnectés de la question initiale, mais qui sont en fait, intimement liés. Ces questions seront inévitablement évoquées, mais l'objet de cette fiche n'est pas de les traiter.

Merci aux relecteurs : Stéphanie Chambaron,

2 Point de vocabulaire

Outliers, **données aberrantes**, **données extrêmes**,... sont des termes utilisés pour désigner ces données « particulières ». Le terme de données aberrantes n'est pas toujours opportun, une donnée pouvant être extrême (par rapport aux autres observations) pour différentes raisons, y compris des raisons liées à la variabilité naturelle. Et dans ce cas, il n'y a pas de raison de qualifier cette donnée d'aberrante. On préfèrera les termes *outlier* (emprunté à l'anglo-saxon) ou « donnée extrême » en français.

3 Comment repérer les points extrêmes ? . . . graphiquement !

Les représentations graphiques constituent le meilleur moyen de détecter les points extrêmes et on aura toujours intérêt à représenter la distribution des observations [1]. La représentation pourra

1. La représentation doit être faite par groupe ou par traitement et non pas directement en poolant les données de l'ensemble des groupes. En effet, les distributions n'ayant pas nécessairement la même moyenne dans les différents groupes, la distribution de l'ensemble des données des différents groupes est un mélange de distributions qui cache des hétérogénéités. La visualisation de la distribution des données poolées ne permet en aucun cas d'appréhender la normalité des distributions intra-groupe.

prendre la forme d'un histogramme, *boxplot*[2] (Fig. 1a), *stem and leaf*[3], *stripchart* (Fig. 1b), *dotchart*[4] , ou autre graphique.
Le type de représentation dépend de la structure des données et du nombre d'observations : si le nombre d'observations par groupe est faible, une représentation sous forme de *boxplot* ou d'histogramme n'est pas pertinente ; on préférera représenter les observations par un *stripchart* ou un *dotchart*.

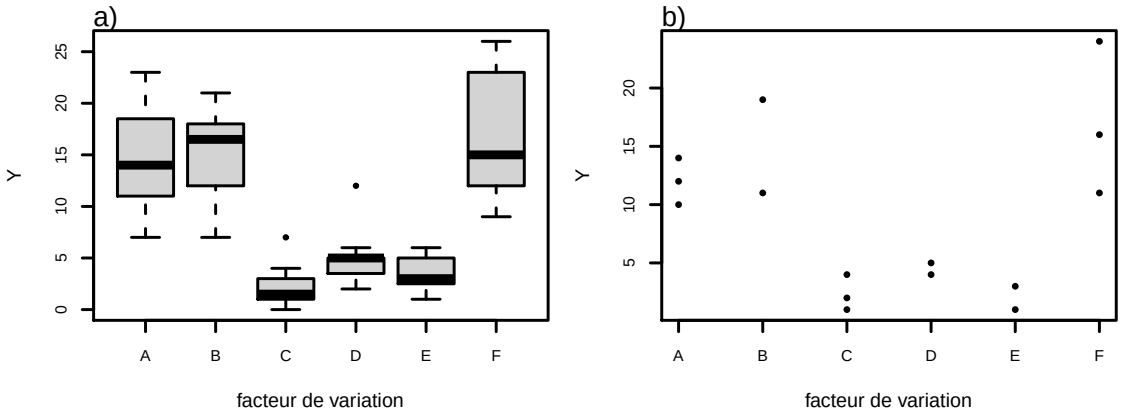

FIGURE 1 – Distribution des observations par groupe :
a) sous forme de *boxplots* : représentation qui permet de résumer les distributions ; à ne pas utiliser si le nombre d'observations est petit (moins d'une dizaine).
b) sous forme de *stripchart* : les observations sont toutes représentées ; à utiliser si le nombre d'observations est petit.

Pour les analyses de type ANOVA ou régression, que les effectifs soient petits ou plus grands, la représentation incontournable est la **représentation des résidus en fonction des valeurs ajustées**[5] (Fig. 2). Elle permet non seulement de repérer les points extrêmes, mais aussi de vérifier *a posteriori* certaines des conditions d'utilisation des méthodes d'ANOVA et de régression, notamment l'homogénéité des variances et la symétrie de la distribution[6].

4 Conditions d'utilisation du test de Student, de l'ANOVA et de la régression

Les trois conditions d'utilisation de l'ANOVA ou de la régression simple ou multiple sont, par ordre d'importance décroissante,

1. l'indépendance entre les observations,
2. l'homogénéité des variances,
3. la normalité des distributions.

Inutile d'utiliser un test statistique pour s'assurer de la normalité des distributions, une appréciation graphique suffira, par exemple la représentation des résidus en fonction des ajustés : il s'agira essentiellement de s'assurer de la symétrie de la distribution des résidus et de l'absence de structure systématique. Les résidus doivent paraître distribués aléatoirement autour de 0, sans

2. diagramme en boîtes, boîte à pattes ou boîte à moustaches, fonction boxplot () de R.
3. diagramme branche et feuilles, fonction stem() de R.
4. stripchart() et dotchart() sont des fonctions R.
5. les modèles de régression ou d'ANOVA consistent à décomposer les valeurs observées en la somme d'une valeur prédite et d'une erreur. Les valeurs prédites, encore appelées 'valeurs ajustées' sont les valeurs estimées. Dans le cas de l'ANOVA ce sont simplement les moyennes.
6. en théorie, la normalité est l'une des conditions d'utilisation, mais la vérification graphique de la symétrie est suffisante.

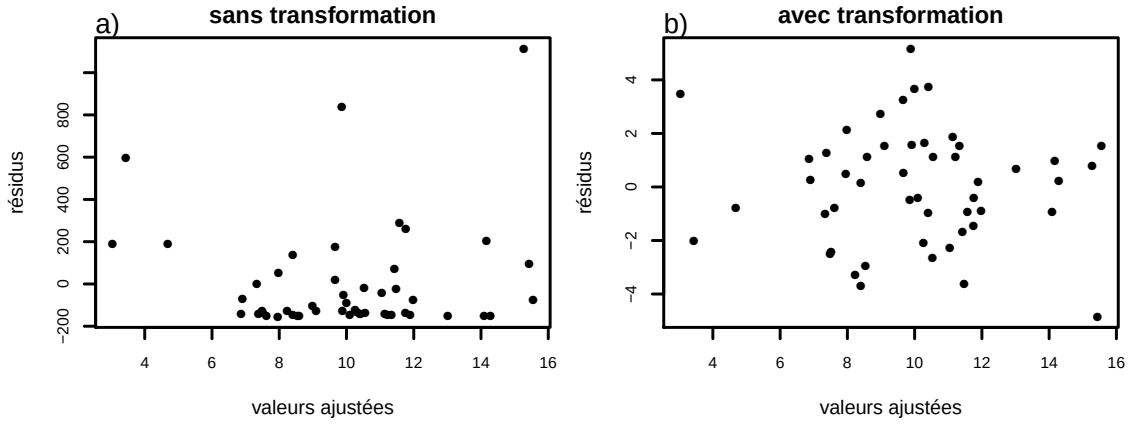

FIGURE 2 – Représentation des résidus en fonction des valeurs ajustées, suite à un modèle d'ANOVA ou de régression (la représentation qu'il faudrait toujours faire).
a) La distribution des résidus est dissymétrique, suggérant une transformation de la variable à expliquer.
b) La distribution des résidus devient symétrique, la variable à expliquer ayant été transformée (transformation logarithme) préalablement au modèle d'ANOVA ou de régression. Les résidus sont distribués aléatoirement autour de 0, avec une variance constante, et ne présentent plus de points extrêmes.

structure systématique, ils doivent s'inscrire dans une « enveloppe constante » : des résidus inscrits dans une enveloppe en forme de cône suggèrent une hétérogénéité des variances, une variance qui augmente ou diminue avec la moyenne.

On peut simplement évaluer, à l'œil, le caractère exceptionnel d'une observation ou d'un résidu, en appréciant la manière dont il s'écarte du nuage, ceci par rapport aux autres observations : il existe toujours un point pour lequel l'écart à la moyenne est le plus grand. Ce qui importe, c'est de savoir si cet écart est beaucoup plus grand ou pas, de l'écart qui le suit, si on ordonne les écarts par ordre décroissant.

On peut, de plus, associer une probabilité critique (*pvalue*) à chacun des résidus, bien que ce ne soit pas indispensable. La probabilité critique associée à chaque résidu est la probabilité d'obtenir un résidu aussi extrême, si les erreurs sont distribuées selon une loi normale. Ne pas oublier qu'il existe toujours un résidu plus grand que les autres, qu'il est bien normal d'obtenir en moyenne 5% de résidus ayant une probabilité critique inférieure à 5%, avant correction liée à la multiplicité des tests (Bonferroni ou autre).
Un résidu associé à une probabilité critique égale à 1% signifie simplement que l'on avait une chance sur 100 d'obtenir une telle valeur. On peut se poser des questions sur cette observation, mais rien ne prouve pas qu'il y ait un problème quelconque avec cette observation. Sur 500 observations, on attend, en moyenne, 5 observations associées à une probabilité critique égale à 1%. Cela peut être un peu plus, comme cela peut être un peu moins, si tout se passe dans des conditions normales.

Pourquoi un test de normalité n'est pas toujours utile ?

L'utilisation « à l'aveugle » d'un test de normalité est en général inutile et peut même se révéler néfaste à plusieurs égards : une telle pratique peut masquer une distribution dissymétrique qui pourrait être rendue symétrique par une transformation de variable ; elle peut conduire à la suppression à tort des données les plus extrêmes (une distribution dissymétrique présente forcément des observations extrêmes) ; elle peut conduire à la décision d'utiliser des tests non paramétriques plutôt que des tests paramétriques, ce qui n'est pas un problème en soi, mais qui ne résoudra pas réellement la question de l' hétérogénéité des variances [a]. Elle peut aussi masquer un point extrême qui peut provenir d'une erreur de transcription des données et qu'il est important de détecter.
Il faut relativiser l'importance de cette condition d'utilisation qu'est la condition de normalité. Un écart à la condition de normalité n'est pas un très gros problème pour plusieurs raisons. Ce n'est pas la condition d'utilisation la plus importante des trois. L'homogénéité des variances et l'indépendance des observations sont plus importantes que la normalité. De plus, même dans le cas où la distribution initiale est non normale, la distribution de la moyenne est rapidement normale, quand la taille de l'échantillon augmente. En résumé, la condition de normalité n'est vraiment pas la condition sur laquelle il faut se focaliser !
Oubliez les tests, faites des graphiques pour apprécier la forme des distributions (unimodalité, symétrie, points extrêmes).

a. Si les tests non paramétriques ne nécessitent pas la normalité des distributions, ils font parfois l'hypothèse de distributions dont la forme est identique, c'est à dire de distributions qui ne diffèrent que par un paramètre de décalage.

5 Que faire en cas de détection de points extrêmes ?

Il faut dans un premier temps s'assurer qu'il ne s'agit pas d'une **erreur de saisie ou de retranscription**, en retournant à l'enregistrement d'origine et corriger les erreurs le cas échéant.

Après correction des éventuelles erreurs de saisie ou de retranscription, il peut s'agir d'un **problème identifié** lié à l'unité expérimentale (individu malade ou n'ayant pas bien compris la tâche, cellule malade, ...) ou à la mesure (appareil de mesure déficient, mal réglé, conditions météo particulières, ...). Un problème identifié peut justifier la mise à l'écart de la donnée. Cela doit être noté et reporté (rapport d'analyse, publication). Dans le même ordre d'idée, on peut, *a posteriori*, se rendre compte que toutes les mesures réalisées le même jour par exemple, sont particulièrement hautes ou basses, ce qui peut laisser soupçonner un problème relatif à la calibration de l'appareil ce jour là. Cela peut aussi justifier la suppression de ces données, mais pas nécessairement. Cette décision est à l'appréciation du chercheur, qui doit, en tout état de cause, en garder la trace, **reporter et justifier** ces décisions.

Dans certains cas, on peut rencontrer des valeurs que l'on juge comme très très extrêmes, improbables si tout s'était déroulé normalement, et n'avoir aucun moyen d'identifier la source du problème éventuel. On peut émettre des **hypothèses** (« on a pu piquer à côté », ...). Il est possible que l'on n'ait aucun moyen de confirmer l'hypothèse. Attention à ne pas chercher des explications à tout prix, pour chaque valeur jugée un peu éloignée. Cela peut conduire à « trouver » de fausses explications. Ne pas oublier que la variabilité naturelle existe, il est tout à fait normal d'avoir 5% de valeurs qui s'éloignent de la moyenne de plus de 2 écart-types !

On peut aussi n'avoir aucune hypothèse à formuler. En présence d'une hypothèse comme en absence d'hypothèse, de telles valeurs ne doivent pas être passées sous silence. On peut réaliser l'analyse après suppression de ces valeurs, mais il doit être reporté, voire discuté, le fait que de telles valeurs ont été obtenues. De telles valeurs pourraient parfois cacher d'autres raisons que celles qui sont a priori envisagées.

Que l'on ait une hypothèse ou pas, on peut toujours réaliser l'**analyse avec et sans les données extrêmes** et comparer les conclusions. Si les conclusions sont identiques, pas de problèmes. Si elles sont différentes, cela signifie que les conclusions sont bien fragiles ! ! !

Il est essentiel de garder la **trace des suppressions éventuelles** de données, des raisons qui ont conduit à la (les) supprimer, des analyses réalisées avec et sans ces données. Ceci peut se faire de différentes façons : cahier de laboratoire, script de commandes, fichier descriptif en format libre (README.txt ou Lisez-moi.txt)...

Il faut avoir conscience que le fait de supprimer les observations extrêmes modifie les moyennes et **biaise la variance résiduelle**, qui est diminuée de façon systématique. Le fait de reporter une variabilité résiduelle sous-estimée entraîne une sous-estimation du nombre de répétitions ou du nombre de sujets nécessaires dans les calculs de puissance [7].

6 Les transformations de variables

Si dans la majorité des cas, les tests et analyses réalisés portent directement sur la variable initiale (non transformée), une transformation de la variable à expliquer peut être nécessaire dans certains cas [8]. En effet, lorsque la distribution de la variable à expliquer (variable Y, variable endogène) est dissymétrique, l'hypothèse de normalité des distributions n'est pas vérifiée. Bien souvent, dans ces cas là, l'hypothèse d'homogénéité des variances ne l'est pas non plus, et c'est ce point là qui pose problème : la probabilité critique (*pvalue*) des tests réalisés n'est pas celle que l'on croit.

Une transformation préalable à la modélisation (ANOVA ou régression par exemple) permet de se conformer aux conditions d'utilisation de la méthode, l'objectif étant que les probabilités critiques soient évaluées correctement, et reflètent correctement le risque de conclure à un effet significatif à tort.
Remarquons qu'il ne s'agit pas de « bidouiller » les données, mais bien d'une méthode classiquement utilisée en statistique, justifiée par le fait qu'il n'existe pas de méthode de mesure universelle, de même qu'il n'existe pas d'échelle de mesure universelle, ni même d'unité de mesure universelle. Toute transformation monotone (c'est à dire croissante ou décroissante, autrement dit conservant l'ordre des valeurs) convient, pourvu qu'après transformation, les conditions de symétrie des distributions et d'homogénéité des variances soient remplies. Ceci se vérifie graphiquement, soit *a priori* par la représentation des données transformées, soit *a posteriori*, par représentation des résidus en fonction des valeurs ajustées.

On peut choisir la transformation dans l'échelle des transformations puissance :

$$\ldots, -Y^{-2}, \quad -Y^{-1}, \quad -Y^{-1/2}, \quad Y^{0}, \quad Y^{1/2}, \quad Y^{1}, \quad Y^{2}, \quad Y^{3}, \ldots$$

On remplacera Y^0 par $\log(Y)$. La transformation $Y^{1/2}$ est la transformation $\sqrt{}(Y)$ et la transformation $-Y^{-1}$ est l'opposé de l'inverse $1/Y$.
Dans le cas où la variable Y est le temps nécessaire pour effectuer une tâche, la transformation $-1/Y$ est l'opposé de la vitesse de réalisation. On prend l'opposé si l'on souhaite que les plus grandes valeurs restent les plus grandes et ne deviennent pas les plus petites, mais ce n'est pas une obligation.

Notons qu'il existe des méthodes permettant de choisir une transformation, la méthode de Box-Cox ou celle de Yeo-Johnson. Mais, au risque d'insister, l'évaluation graphique *a posteriori* est la meilleure des méthodes, puisqu'elle seule permet de s'approprier les données, de détecter les régularités (*patterns*), les points extrêmes, les variances hétérogènes, les distributions dissymétriques, les groupes d'observations.

7. Calcul du nombre de répétitions nécessaires pour avoir une probabilité données, 80% par exemple, de mettre en évidence une différence donnée, pour un risque α de première espèce donné.
8. `http://en.wikipedia.org/wiki/Data_transformation_(statistics)`

7 Compléments, sujets connexes

Quelques sujets connexes, qui ne sont pas développés dans cette fiche, mais cités pour mémoire :

moyenne tronquée (*trimed mean*, ou *troncated mean*) : au même titre que la moyenne, ou la médiane, c'est une mesure de « tendance centrale ». Le calcul de la moyenne se fait après exclusion d'une certaine proportion des données (les plus extrêmes). Si l'estimateur existe et peut être employé (je n'ai jamais vu d'application réelle), cela n'autorise pas pour autant à supprimer une part des données avant de faire une ANOVA.

winsorising : méthode qui consiste à ramener les données les plus extrêmes à un certain percentile. Je n'ai jamais rencontré d'exemple qui justifie cette modification des données.

donnée censurée : ce sont des données dont on n'a pas la mesure exacte. On sait simplement qu'elles sont inférieures ou supérieures à un certain seuil. Ce type de données peut résulter de valeurs inférieures au seuil de sensibilité d'un appareil par exemple.

critère d'exclusion : certains individus peuvent ne pas appartenir à l'échantillon considéré, ceci parce qu'ils ne remplissent pas les critères de recrutement fixés à l'avance. Mais il ne s'agit pas d'exclusion de données a posteriori mais bien de la définition de la population considérée et donc des critères qui doivent être remplis pour qu'une unité expérimentale fasse partie de l'échantillon, ou qu'une personne soit recrutée.

méthode résistante : une méthode est résistante vis à vis de données extrêmes si les résultats sont peu influencés par la présence de données extrêmes.

méthode robuste : une méthode est dite robuste lorsque les propriétés statistiques sont assez bien conservées en cas d'écart aux conditions d'utilisation.

8 Commandes R : quelques exemples

Les commandes ci-dessous sont présentées pour servir d'exemple. Elles utilisent les jeux de données diffusés dans R, et peuvent donc être tapées ou collées dans R, et s'exécuter normalement, sans produire de message d'erreur (R 3.1.2).

Représentation des données brutes :

```
boxplot(count~spray, data=InsectSprays)      # résumé en boxplot (fig.1a)
stripchart(count~spray, data=InsectSprays)   # données individualisées (fig.1b)
```

Voici l'exemple d'une ANOVA. On réalise la représentation incontournable, celle des résidus en fonction des valeurs ajustées. Une autre représentation possible est celle des résidus studentisés en fonction des valeurs ajustées. Les valeurs -2 et 2 constituent alors des repères [9].

```
lm1=lm(count~spray, data=InsectSprays)  # ANOVA (modèle linéaire : lm)
plot(lm1, which=1)                      # résidus en fonction des valeurs ajustées
plot(rstudent(lm1) ~ predict(lm1))
abline(h=c(-2,0,2), lty=c(33,1,33))
```

Voici l'exemple d'une régression linéaire multiple. On peut éventuellement [10] compléter la représentation des résidus en fonction des valeurs ajustées, par la détection des résidus associés à une probabilité critique inférieure à un seuil donné (ici α =0,05), après correction de Bonferonni [11].

9. -1.96 et 1.96 sont les quantiles d'une loi normale centrée réduite, associés aux probabilités 2.5% et 97.5%, donc associés à un test à 5%.
10. Ceci n'est absolument pas une obligation.
11. Le but de la correction de Bonferroni est de tenir compte de la multiplicité des tests réalisés.

```
library(car)    # pour accéder au dataframe Prestige et à la fonction outlierTest
lm1=lm(prestige~education+income+type, data=na.omit(Prestige)) # régression multiple
plot(lm1, which=1)
outlierTest(lm1, cutoff=0.05)
```

9 En résumé

- **explorer les données** : représentations graphiques : distribution, graphique des résidus en fonction des valeurs ajustées....
- la suppression de données doit être l'exception.
- **traçabilité** : garder la trace des suppressions éventuelles et des raisons [12].
- **transparence** : reporter toute correction, modification, suppression, dans les rapports et publications.
- se préoccuper de l'homogénéité des variances et de la symétrie des distributions et oublier les tests de normalité.

12. Si des modifications sont faites directement dans un fichier excel, conserver la version d'origine, et créer une nouvelle version (en suffixant par le numéro de version fichier-v2.csv par ex).