Traiter ses données proprement : vers un meilleur usage du data cleaning

Aurore Deledalle, Charlotte Rowe

Traiter ses données proprement : vers un meilleur usage du *data cleaning*

Treating your data correctly: towards a better use of data cleaning

Aurore Deledalle

Charlotte Rowe

Aurore Deledalle, Maître de Conférences en Méthodologie, Faculté de Psychologie, Laboratoire de Psychologie des Pays de la Loire LPPL-EA4638, Université de Nantes, France

Thèmes de recherche : Méthodologie quantitative, psychométrie, analyse factorielle, modélisation en équations structurelles

Charlotte Rowe, Etudiante, Faculté de Psychologie, Laboratoire de Psychologie des Pays de la Loire LPPL-EA4638, Université de Nantes, France

Thèmes de recherche : statistiques, méthodologie, modélisation en équations structurelles, trauma, trouble de stress post-traumatique

Auteur correspondant :

Aurore Deledalle, chemin de la Censive du Tertre, 44000 Nantes, France.

Mail : aurore.deledalle@univ-nantes.fr

Téléphone : 02 53 52 26 11

Traiter ses données proprement : vers un meilleur usage du *data cleaning*

Résumé

L'analyse préliminaire des données, appelée *data cleaning* en anglais, est une étape essentielle, mais souvent obscure du traitement statistique. Dans cet article nous étudions l'effet du *data cleaning* avec des préconisations pour la pratique. Les étapes abordées lors du *data cleaning* sont l'examen des données avant toute analyse, le traitement des valeurs atypiques (*outliers*) et celui des valeurs manquantes. Après avoir étudié les différentes procédures possibles et leurs conditions d'application à chaque étape, nous proposons une démarche transparente et systématique de *data cleaning*. Enfin, nous illustrons l'effet du *data cleaning* sur un jeu de données réel.

Mots-clé : data cleaning, valeurs atypiques, données manquantes, normalité, conditions d'application

Treating your data correctly: towards a better use of data cleaning

Abstract

Data cleaning has long been shown to be an essential step in data analysis but its application is not systematic and varies between researchers. The aim of the present article is to study the effect of the different elements of data cleaning and to give recommendations. The steps considered are normality, outliers and missing values. Firstly, we advise a close examination of the data and its normality with a view to transformation. Secondly, we look at outliers and their treatment (trimming and winsorizing). Lastly we turn our attention to missing values and how to deal with them depending on their randomness. We also cover the importance of

visualizing data before analysis. Finally, we illustrate the effect of these data cleaning practices on a real data set. We show that data cleaning can lead to a significant result becoming non-significant and therefore demonstrate the importance of these steps before starting data analysis.

Keywords: data cleaning, outliers, missing data, normality, statistical assumptions

Date de soumission : 5 mai 2018

Date de révision : 24 juin 2019

Traiter ses données proprement : vers un meilleur usage du *data cleaning*

Assez paradoxalement, alors qu'un grand nombre de recherches en psychologie s'appuie sur de larges recueils de données quantitatives, on constate une hétérogénéité de procédures concernant la pratique de l'analyse préliminaire des données ; pourtant celle-ci est un pré-requis au traitement statistique. Par analyse préliminaire, on entend l'ensemble des démarches visant à s'assurer de la validité des données (absence d'erreurs de saisie), à vérifier la nature des variables à partir de la distribution des réponses, à prendre des décisions sur le traitement des valeurs manquantes et celles dites atypiques (*outliers*), ensemble de démarches identifié dans la littérature internationale par l'expression *data cleaning*[1].

Un examen de la littérature démontre néanmoins une production vaste sur ce sujet, on relève en effet : un numéro spécial dans *Frontiers in Psychology* coordonné par Osborne (2013), un ouvrage intégralement consacré au *data cleaning* (Osborne, 2012), un chapitre entier (*Cleaning Up Your Act: Screening Data Prior to Analysis*) dans l'ouvrage de Tabachnick et Fidell (2007), manuel reconnu comme une référence dans l'analyse de données (cité 89048 d'après Google Scholar). Malgré l'existence de cette littérature, un examen sur les pratiques révèle que le domaine reste encore peu maîtrisé par les chercheurs (Hoekstra, Kiers & Johnson, 2012, Keselman et al., 1998). Pour expliquer ce décalage entre ces vastes ressources méthodologiques disponibles d'une part et les pratiques hétérogènes des chercheurs d'autre part plusieurs hypothèses peuvent être avancées. (i) Tout d'abord, le *data cleaning* jouit d'une mauvaise réputation : souvent associé à de la manipulation des données pour obtenir une confirmation des hypothèses, le *data cleaning* semble suspicieux au sein des sciences psychologiques récemment montrées du doigt pour leur manque de reproductibilité et de quelques cas de fraudes médiatisés (Chambers, 2017). (ii) Autre lacune, à une situation

[1] Nous conserverons ce terme dans la suite du manuscrit puisqu'il n'existe pas de traduction satisfaisante en langue française

problème en *data cleaning* peuvent correspondre plusieurs solutions qui nécessitent pour faire un choix éclairé un approfondissement parfois complexe, il s'agit en cela d'un savoir méthodologique disponible mais peu accessible au chercheur. Par exemple, avant de décider du remplacement d'une valeur manquante par une autre valeur ou bien de la suppression de l'observation, encore faut-il préalablement déterminer si la valeur manquante est complètement aléatoire (Missing Completely At Random, MCAR), aléatoire (Missing At Random, MAR) ou encore non aléatoire (Not Missing At Random, NMAR). (iii) Troisièmement, lorsqu'on se penche sur la littérature méthodologique, on constate assez rapidement des prises de position contradictoires qui peuvent démoraliser le chercheur ; par exemple faut-il ou ne faut-il pas transformer une variable ne se distribuant pas normalement ? Ainsi Field (2009) le préconise, puis prend une position beaucoup plus modérée par la suite (Field, 2016). (iv) Enfin, une quatrième piste est à notre sens plus historique : la psychologie actuelle tient d'une grande tradition expérimentale ; or l'expérimentation n'implique pas de travailler sur de grands jeux de données (bien souvent une trentaine de participants par condition expérimentale est suffisant pour mettre en évidence un effet modéré), contrairement à la psychologie reposant sur des méthodes descriptives d'enquête (qui découlent davantage des sciences sociales). En conséquence, une partie des chercheurs ne semble pas concernée par la pratique des procédures de *data cleaning* (*eg*. la probabilité d'avoir des valeurs manquantes est plus élevée pour un jeu de données avec un effectif de 800 participants qu'avec 30), ceci pouvant expliquer que ces procédures ne figurent pas systématiquement dans les programmes de formation.

Les objectifs de cet article sont les suivants : (1) Présenter de manière synthétique l'ensemble des étapes abordées lors du *data cleaning* avec des préconisations pour la pratique, (2) Proposer une démarche transparente et systématique de *data cleaning* en vue d'améliorer la

qualité et la reproductibilité de nos recherches, (3) Illustrer cette démarche par son emploi sur un jeu de données.

1. Les étapes du *data cleaning*

1.1. Examiner attentivement les données

Traiter des données demande du temps. Avant de se précipiter dans la mise en œuvre des tests statistiques liés à la validation des hypothèses, il est nécessaire de se questionner sur la qualité des données collectées, de comprendre l'allure de la distribution des variables de l'étude. Cela passe tout d'abord par la vérification d'éventuelles erreurs de saisie ce qui n'est pas si aisé. Parmi ces erreurs de saisie, les plus faciles à détecter sont celles pour lesquelles la valeur sort de l'étendue vraisemblable (*expected range*) pour une variable : par exemple, lorsqu'en saisissant des données le chercheur maintient son doigt trop longtemps sur une touche. Il suffit de considérer attentivement les valeurs maximales prises par chaque variable pour identifier ces erreurs.

Dans la même veine, il est recommandé d'examiner la distribution de chacune des variables. Cette étape permet tout d'abord de s'assurer de la bonne reconnaissance de la nature de la variable par le logiciel d'analyse statistique. On vérifie notamment que les variables catégorielles codées avec des valeurs numériques sont bien identifiées comme telles. Sous R, la fonction `summary ()` permet de repérer rapidement ce type d'erreur en offrant un résumé des statistiques descriptives ; en effet avec cette fonction, R présente pour les variables numériques des indices de tendance centrale alors que seuls les effectifs sont donnés pour les variables catégorielles. La fonction `str` () est également intéressante car elle permet de vérifier la structure des variables après importation et intègre les notions de variables discrètes/continues, nominale/ordinale.

Il est ensuite primordial de repérer la dispersion des variables. Pourquoi le faire ? Et bien tout d'abord parce que de nombreuses hypothèses de recherche sont de nature corrélationnelle, or si une variable de l'étude ne varie pas ou peu, il est impossible qu'elle varie en même temps qu'une autre (Tabachnick & Fidell, 2007) : pas de covariance sans variance ! Pour les variables continues, il est donc nécessaire de présenter les indices de dispersion (variance, écart-type, quartiles…), en complétant par un examen graphique (histogramme, boîte à moustaches…), une dispersion maximale est assurée par une distribution normale. Pour les variables catégorielles (eg. sexe, niveau d'études), des conditions sont également à vérifier en fonction du test choisi. Par exemple, pour un Chi^2, il faut s'assurer que les modalités de réponses contiennent des effectifs suffisants (i.e., pas de cellule avec moins de 5 observations) ; s'il y a beaucoup de modalités avec à chaque fois de faibles effectifs, il est alors judicieux d'effectuer des regroupements (Husson, Lê & Pagès, 2009).

1.2. Vérifier la normalité

Dans cette étape d'examen préliminaire apparaît alors le questionnement sur la normalité de la distribution pour les variables continues. L'examen de l'allure des distributions des variables est en effet essentiel pour comprendre comment les participants se sont comportés durant l'étude. Dans la pratique actuelle, il existe une focalisation de la vérification de la normalité des variables observées en tant que condition *sine qua none* à la réalisation d'un test statistique paramétrique (Ghasemi & Zahediasl, 2012). Ceci n'est pas tout à fait juste. En effet, la condition de normalité, même si elle existe, ne porte pas sur la normalité des variables observées (*sample distribution*, Field, 2009), mais sur la distribution d'échantillonnage d'un paramètre (*sampling distribution*). Pour bien illustrer ce point, il est nécessaire de faire un bref rappel sur la logique des tests statistiques. Malgré leur diversité apparente, les tests statistiques les plus usuels en psychologie (test-*t*, ANOVA, régression...) reposent sur une

même approche fréquentiste. A partir des données fournies par un échantillon, on dresse la distribution d'échantillonnage d'un paramètre (un coefficient *b* en regression, *r* en corrélation, $\mu 1$-$\mu 2$ dans une comparaison de moyenne,...) en se positionnant sous H_0. Cette distribution d'échantillonnage représente les diverses valeurs prises par le paramètre (*b*, *r*, μ_1-μ_2...) si on le mesurait dans de multiples échantillons, en partant de l'hypothèse que, dans la population, l'effet n'existe pas (H_0, c'est à dire qu'il n'y a pas de différence de moyennes entre des groupes ou bien pas de relation entre des variables). C'est ici que l'on considère que cette distribution d'échantillonnage doit nécessairement suivre une loi normale car ensuite on utilise les probabilités connues de la distribution normale pour apprécier la rareté de la valeur obtenue du paramètre (*b*, *r*, m1-m2...) dans l'échantillon observé. Si la valeur observée pour l'échantillon se situe à l'extrémité de la distribution d'échantillonnage, elle est donc très peu probable (p <.05) sous H_0, ce qui conduit au rejet de cette hypothèse nulle. Donc faire un test de significativité, et dans la même veine calculer des intervalles de confiance, ne prend sens que si la distribution d'échantillonnage du paramètre suit une distribution normale. La difficulté est alors de s'assurer que la distribution d'échantillonnage suit une loi normale car nous n'y avons pas directement accès. Les avis restent partagés sur cette question. Certains considèrent que du moment qu'un échantillon est de taille supérieure à 30, il n'y a pas à s'en préoccuper en raison du Théorème Central Limite (TCL, Çetinkaya-Rundel, 2014, Field, 2016)[2].

[2] Pour illustrer le TCL, le lecteur pourra utiliser l'application suivante proposée par Mine Cetinkaya-Rundel `https://gallery.shinyapps.io/CLT_mean/` En partant d'une population parente normale de moyenne 0 et écart-type 20, on extrait par exemple 200 échantillons de taille 30, on constate alors que la variable se distribue rarement normalement dans les échantillons. Néanmoins la distribution d'échantillonnage est d'allure normale. Le blog d'Andy Field offre également une vision en ce sens de la condition de normalité.

http://discoveringstatistics.blogspot.com/2012/08/assumptions-part-1-normality.html

D'autres, au contraire, soulignent que travailler avec des variables observées d'allure normale favorise la validité des calculs des erreurs standards et tests de significativité (Miles & Shevlin, 2001) et préconisent d'appliquer des transformations en cas d'écart important à la normalité (pour une présentation complète, voir Tabachnick et Fidell, 2007). Le débat n'est pas tranché mais, même si la normalité des variables observées prend une place relative en tant que condition d'application[3], il n'en demeure pas moins qu'une description correcte des variables implique un questionnement sur l'allure des distributions.

Rappelons qu'il existe deux facteurs d'écarts à la normalité (Miles & Shevlin, 2001) : (1) la variable mesurée du fait de sa nature ne se distribue pas normalement, cela est au final assez fréquent, par exemple des variables telles que le niveau de revenu, la dépression, le temps de réaction, présentent habituellement une asymétrie à droite (*positive skew*), les réponses à un questionnaire de satisfaction, une asymétrie à gauche (*negative skew*) en raison de la tendance à l'acquiescement (lorsque le pôle positif de l'échelle de Likert exprime un fort degré de satisfaction) ; (2) le second facteur d'écart est lié à l'existence de valeurs atypiques (*outliers*). Le traitement des *outliers* est suffisamment dense pour que nous le traitions dans une section suivante.

En raison, de ces deux facteurs, identifier un écart à la normalité consiste alors à décrire l'allure de la distribution et à détecter d'éventuelles valeurs atypiques. Pour ce faire, il est d'usage d'employer une méthode graphique (histogramme, graphe quantile-quantile, boite à moustache). A préciser que l'histogramme offre parfois une vision déformée de la distribution car les valeurs peuvent être agrégées en classes sur l'échelle et donner l'impression d'une sur ou sous représentation de certaines densités. Des méthodes numériques sont également éclairantes : Tabachnick et Fidell (2007) proposent par exemple que des indices d'asymétrie

[3] Il apparaît en cela plus judicieux de traduire le terme *assumption* par condition de généralisation plutôt que condition d'application.

et d'aplatissement inférieurs à |1| assurent d'une distribution en forme de cloche, il n'y a cependant pas de consensus sur les valeurs seuil de ces indices dans la littérature. L'usage de test statistique pour vérifier la normalité (eg. Shapiro-Wilk, Mardia Kurtosis pour la normalité multivariée) est possible mais déconseillé avec un large effectif (Mahibbur Rahman & Govindarajulu, 1997) en effet le test devient alors trop puissant (la puissance statistique étant notamment liée à la taille de l'échantillon) : un moindre écart à la normalité est alors détecté et s'ensuit d'un rejet de l'hypothèse nulle.

1.3. Traiter les valeurs atypiques

Une valeur atypique (*outlier*)[4] est décrite comme une observation « qui dévie tellement des autres observations qu'elle soulève la suspicion d'avoir été généré par un autre mécanisme » (Hawkins, 1980).

La présence de valeurs atypiques dans un jeu de données peut sérieusement nuire à la validité et la généralisabilité des résultats : les valeurs atypiques augmentent la variance d'erreur et réduisent la puissance des tests statistiques ; si elles ne se distribuent pas aléatoirement, elles sont également susceptibles d'augmenter à la fois les risques d'erreur de type I et II ; enfin, les paramètres estimés sont biaisés puisque non généralisables à la population d'intérêt (Osborne, 2010).

Pour détecter une valeur atypique sur une variable, la procédure la plus fréquemment (Ghosh & Vogt, 2012) employée consiste, si la variable se distribue normalement, à repérer les valeurs s'écartant d'au moins 3 écarts-type de la moyenne, soit $z > 3$ si la variable est standardisée. Sachant que la probabilité d'être au moins à cette distance de la moyenne pour

[4] L'expression « valeur atypique » semble une meilleure traduction d'*outlier* que « valeur extrême » car celle-ci n'est correcte que pour le cas univarié : si une valeur atypique pour une variable est assimilable à une valeur extrême, dans le cas multivarié, l'*outlier* renvoie à une configuration atypique de scores (Tabachnick & Fidell, 2007).

un individu est de 0.26 %, il y a là un élément tangible pour suspecter cet individu de ne pas appartenir à la population d'intérêt. Néanmoins, comme la valeur atypique participe au calcul de l'écart-type, Field (2016) recommande plutôt de considérer les valeurs les plus extrêmes de la distribution comme atypiques (par exemple, 5% des valeurs se situant aux deux extrémités). Pour identifier une configuration atypique de scores (*bivariate or multivariate outlier*) et son influence sur les résultats, des indices peuvent être calculés (résidus standardisés ou studentisés, distance de Mahalanobis et de Cook, pour une présentation complète, consulter Miles & Shevlin, 2001). Osborne (2010) identifie six explications à la présence de valeurs atypiques : (i) les erreurs de saisie, que l'on peut corriger par un retour au matériel original, (ii) les réponses volontairement erronées, que ce soit à des fins de sabotage de la recherche ou bien guidées par un besoin de valorisation de soi (biais de désirabilité sociale) ; dans ce cas il est conseillé de retirer les données de l'individu (Judd & McClelland, 1989 ; Osborne, 2010), (iii) une erreur d'échantillonnage, dans cette situation il est également conseillé de retirer les données de l'individu puisqu'il n'appartient pas à la population d'intérêt, (iv) un défaut d'ordre méthodologique avec par exemple un appareil de mesure physiologique qui se dérègle : si l'erreur n'est pas corrigeable, là encore il est plus prudent de supprimer les données concernées, (v) une distribution considérée à tort comme d'allure normale : si la distribution présente une forte asymétrie, il est incorrect d'interpréter une valeur $z > 3$ comme une valeur atypique, dans ce cas une transformation préalable de la variable est souhaitable avant de standardiser et de rechercher les valeurs s'écartant de 3 écarts-type de la moyenne, (vi) les cas rares mais appartenant bien à la population d'intérêt. C'est cette dernière situation qui est la plus problématique puisque la valeur atypique est finalement légitime. On pourrait choisir de la conserver, puisqu'elle n'est ni une erreur de saisie, ni une erreur d'échantillonnage, ni une réponse volontairement erronée, etc. Néanmoins, les tests risquent d'être affectés par ces valeurs. Field (2016) présente différentes stratégies pour traiter les

valeurs atypiques relevant de cette sixième situation telles que la suppression (*trimming*) de la valeur atypique lors du calcul des paramètres ou le remplacement de la valeur atypique par une valeur plus plausible (*winsorizing*) : par exemple en remplaçant les valeurs atypiques par celle des percentiles 5 et 95.

1.4. Considérer les valeurs manquantes

Avoir des valeurs manquantes dans un fichier de données n'est pas une situation rare comme le révèle une étude de Peng et al. (2006) : sur un examen de 1087 recherches quantitatives publiées dans 11 revues de psychologie de l'éducation, 54 % contiennent des valeurs manquantes.

Si la part de ces valeurs manquantes reste raisonnable pour une variable donnée (inférieure à 5% selon le seuil proposé par Schafer, 1999), différentes techniques de prises en charge sont à disposition du chercheur : les techniques de suppression, celles de substitution traditionnelle et celles plus modernes d'imputation (Peugh & Enders, 2004). Lorsque la part de valeurs manquantes est plus élevée, il importe de s'interroger sur la validité du recueil de données (Schlomer, Bauman & Card, 2010).

Les techniques de suppression consistent simplement à ignorer les valeurs manquantes. La suppression par observation (*listwise deletion*) vise à supprimer toutes les données pour un individu dès lors qu'il existe au moins une valeur manquante pour cet individu, alors que la suppression par paire (*pairwise deletion*) n'implique de ne supprimer l'observation que lorsque l'analyse demandée nécessite cette valeur (par exemple pour un calcul de corrélation). Ces techniques restent fréquemment utilisées car souvent paramétrées par défaut dans les logiciels statistiques. Elles questionnent néanmoins sur la représentativité de l'échantillon obtenu suite à ces suppressions et menace substantiellement la puissance statistique des tests

effectués du fait d'une réduction, parfois drastique, de la taille d'échantillon. C'est pourquoi ces techniques sont déconseillées (Peugh & Enders, 2004).

Les techniques traditionnelles de substitution consistent à remplacer la valeur manquante par une valeur plausible. On peut par exemple citer *le remplacement par la moyenne* (la valeur manquante est remplacée par la moyenne arithmétique calculée à partir des valeurs présentes pour la variable) ou le remplacement par régression (une équation de régression est calculée pour prédire les valeurs manquantes de la variable considérée à partir des valeurs des autres variables qui lui sont corrélées, lesquelles sont alors prises comme variables prédictrices). Ces deux techniques restent controversées car elles réduisent la variance de la variables, on sous-estime alors les variances et covariances ; les intervalles de confiance s'en trouvent aussi réduits.

Les limites des techniques évoquées précédemment peuvent être dépassées par des procédures plus actuelles parmi lesquelles on peut citer l'estimation par le maximum de vraisemblance (*Maximum Likelihood*, ML) et l'imputation multiple (*Multiple Imputation*, MI) qui vont être brièvement décrites. L'objectif de l'estimation ML est d'identifier les valeurs des paramètres pour la population qui seraient les plus vraisemblables au vu des données observées dans un échantillon. La méthode s'inscrit dans le contexte plus large des modèles d'équations structurales, que les données de l'échantillon comportent ou non des valeurs manquantes. Elle implique un processus itératif durant lequel différentes valeurs pour l'estimation des paramètres sont successivement testées jusqu'à l'obtention d'une bonne adéquation entre ces paramètres estimés et les valeurs observées dans l'échantillon. Lorsque cette méthode est employée, il n'est pas nécessaire de supprimer les observations comportant des données incomplètes, ni de fixer une valeur de remplacement avant d'effectuer des analyses : les données partielles permettent d'imputer des valeurs probables aux scores manquants à partir

des corrélations entre les variables (Peugh & Enders, 2004). La méthode *d'imputation multiple* est développée plus récemment, elle présente l'avantage de préserver l'incertitude liée à la valeur manquante. Dans l'ensemble des techniques évoquées précédemment, la valeur manquante se trouve remplacée par une valeur donnée et une seule, ce qui ne restitue pas la variabilité de cette valeur. Avec l'imputation multiple, plusieurs jeux de données imputés sont créées (le plus souvent entre 5 et 10), c'est-à-dire que pour chaque jeu, la valeur manquante est remplacée par une valeur différente qui est prédite à partir des valeurs sur les autres variables. Ces jeux de données « complets » vont respecter les caractéristiques de la distribution des données observées (variabilité et corrélations entre les variables). Cette procédure permet de prendre en compte l'incertitude liée aux données manquantes. On effectue ensuite les analyses souhaitées sur ce pool de jeux de données : les résultats obtenus (par exemple, des coefficients de régression) sont une synthèse des résultats de chaque jeu de données (Peugh & Enders, 2004).

Le choix de la technique à employer est orienté par la nature des valeurs manquantes. En effet, il existe différents modèles (*pattern*) de valeurs manquantes qui peuvent être qualifiés selon Little and Rubin (1987) de « missing completely at random » (MCAR), « missing at random (MAR) ou « not missing at random » (NMAR). Les valeurs qualifiées de MCAR sont totalement aléatoires, elles ne sont pas corrélées à d'autres variables de l'étude. Il n'y a aucune relation entre une valeur manquante et toute autre valeur contenue dans les données, autrement dit, la probabilité d'absence est la même pour toutes les données. Si les données sont MCAR, alors choisir de gérer les valeurs manquantes par suppression n'est pas un mauvais choix car le sous-échantillon sera autant représentatif que l'échantillon initial. Il est néanmoins nécessaire de s'assurer que travailler sur un échantillon plus petit ne posera pas de difficulté par rapport aux analyses envisagées et ne réduit pas excessivement la puissance statistique.

Le qualificatif de MAR pour les données manquantes est une expression ambigüe car le terme « manquant au hasard » est ici trompeur : en réalité la probabilité d'avoir des valeurs manquantes est liée à des variables de l'étude. MAR signifie que les données manquantes sont liées à des variables observées dans l'étude et non pas aux valeurs manquantes en elles-mêmes. Dong et Peng (2013) illustrent le cas de données MAR par l'exemple suivant : un chercheur évalue l'effet d'un entraînement au calcul mental, il effectue deux mesures (pré- et post-test) de la performance en calcul mental. Imaginons qu'un étudiant a un niveau très faible en calcul mental, il peut se décourager et quitter l'entraînement, cela va avoir pour conséquence une donnée manquante pour le post-test : la probabilité d'avoir une valeur manquante en post-test est alors dépendante du score en pré-test. Si les données manquantes sont MAR, il est judicieux d'appliquer une technique d'imputation. La suppression de l'observation conduirait à un échantillon biaisé (eg. l'échantillon ne serait constitué que d'élèves les plus performants en calcul mental).

Enfin, les données désignées par NMAR désignent le phénomène que les données dépendent d'elles-mêmes et non pas d'autres variables observées dans l'étude. Voici un exemple (proposé par Dong & Peng, 2013) : imaginons dans un questionnaire un item qui demanderait la valeur du salaire. Ici, les personnes qui ne répondent pas sont par exemple celles qui ont les revenus les plus élevés car elles sont plus réticentes à donner cette information. Les valeurs manquantes ne sont donc pas nécessairement liées à une autre variable observée. Ce type de valeur manquante est plus difficile à détecter, il faut que le chercheur s'interroge sur la raison des valeurs manquantes. Pour les données, NMAR, il n'existe pas de procédure statistique pour les détecter ni de solution pour les prendre en charge : la suppression comme l'imputation ne sont pas des solutions satisfaisantes.

2. Proposition d'une démarche systématique de *data cleaning*

Bien qu'il existe une littérature pertinente et détaillée sur chacune des étapes du *data cleaning*, on constate qu'il existe peu de recommandations sur la globalité de ces étapes. L'objectif de cet article est d'apporter une contribution en proposant une démarche systématique de *data cleaning* basée sur les trois étapes précédemment présentées : traitement préalable des variables, gestion des valeurs atypiques et valeurs manquantes. L'ensemble des préconisations est rappelées dans le tableau 1.

[Insérer tableau 1]

3. Illustration de la démarche proposée

3.1. Présentation des données

Les données utilisées pour illustrer ces différentes étapes proviennent d'une recherche plus vaste (auteurs, 2017) visant notamment à comparer l'effet de deux méthodes pédagogiques d'enseignement de l'anglais (intitulées pédagogie A et pédagogie B) sur le niveau d'anxiété scolaire d'élèves de classe de 6ème. La pédagogie A est classique alors que la pédagogie B est davantage innovante. Les chercheurs s'intéressent ici non pas à une amélioration des performances en anglais des élèves mais à l'effet de ces choix pédagogiques sur les émotions ressenties par les élèves (*achievement emotions*, Pekrun, Goetz, Frenzel, Barchfeld & Perry, 2011), en particulier l'anxiété. Ce niveau d'anxiété est mesuré en début d'année scolaire, puis en fin d'année, ce qui conduit à analyser l'effet de la pédagogie par une ANOVA mixte 2(pédagogie, facteur inter-sujets)*2(temps, facteur intra-sujet). Les chercheurs font l'hypothèse que le niveau d'anxiété diminue davantage avec la pédagogie B.

Des données ont été recueillies auprès de 72 participants au premier temps (T1) ; 10 participants n'étaient pas présents lors du second temps (T2), l'échantillon rassemble donc 62 participants.

Le fichier (disponible en matériel supplémentaire[5]) comporte 62 observations et 118 colonnes qui regroupent des variables sociodémographiques (SEX, AGE), les scores aux 15 items évaluant l'anxiété scolaire (AEQ, Pekrun et al., 2011) aux temps 1 (T1) et au temps 2 (T2), différents scores composites illustrant les étapes du *data cleaning* : scores d'anxiété recalculés après que les outliers aient été supprimés (trimmed) ou bien remplacés (winzorised), après une transformation racine carrée (sqrt), après que les valeurs manquantes aient été remplacées soit par la moyenne (MVbyMEAN) soit par imputation multiple (MVimputed).

On propose donc, pour chaque étape du *data cleaning* de recalculer les moyennes des groupes et les résultats de l'ANOVA mixte.

3.2. Etapes du data cleaning

Valeurs atypiques. Puisque les valeurs atypiques peuvent expliquer un écart à la normalité, il s'avère nécessaire de débuter le *data cleaning* par le traitement de ces valeurs. Pour illustrer cette étape, nous avons opté pour deux modalités de traitement, la suppression (*trimming*) et le remplacement par une valeur plus plausible (*winsorizing*). Nous ne rechercherons pas ici l'explication de la présence de ces valeurs atypiques (car pour cela, un retour sur le recueil de données est nécessaire) et les considérerons comme des valeurs atypiques mais légitimes (6ème cas évoqué par Osborne, 2010). Comme préconisé par Field (2016), nous avons pris comme valeur seuil, 2,5 % de chaque coté de la distribution (plutôt que 3 écarts-type), ce qui conduit à supprimer/remplacer 4 valeurs pour la variable anxiété 1 (individus 22, 37, 40 et 60) et 5 valeurs pour la variable anxiété 2 (individus 21, 22, 23, 37 et 50). Les résultats sont présentés dans le tableau 2.

[Insérer tableau 2]

[5] Fichier au format csv. Les séparateurs de champs sont des points-virgules, la marque de décimale est la virgule, les données manquantes indiquées par NA.

On constate qu'en l'absence de *data cleaning* (ligne 1, tableau 2), le résultat de l'ANOVA est significatif ainsi qu'en remplaçant les valeurs atypiques par des valeurs plus plausibles (ligne 3, tableau 2), avec à chaque fois un effet moyen, selon les valeurs seuil proposées par Cohen[6] (1969). La suppression des valeurs atypiques (ligne 2) conduit à un résultat d'ANOVA non significatif (tableau 2) associée à une diminution de la taille d'échantillon de 11%.

[Insérer Figure 1]

Distribution des variables. Les Q-Q plot et histogrammes présentés aux figures 1 et 2 révèlent une asymétrie importante des deux distributions. En cas de forte asymétrie positive, Tabachnik et Fidell (2007) recommande d'appliquer une transformation racine carrée. Cette transformation est donc effectuée sur les deux variables dépendantes (pour lesquelles les valeurs atypiques ont été remplacées) de manière à obtenir une distribution plus proche de la normalité (figures 1 et 2). Par ailleurs, le test de normalité multivariée de Mardia Kurtosis confirme l'intérêt de la transformation des variables observées (avant les transformations z = -0.352, p = .725 ; après les transformations : z = -1.282, p = .200)[7].

Les résultats de l'ANOVA mixte sont à nouveau calculés (ligne 4) et présentés dans le tableau 2.

[Insérer figure 2]

Valeurs manquantes. Dans un premier temps, afin de déterminer le pattern des valeurs manquantes, les variables anxiété 1 et anxiété 2 sont recodées en variables catégorielles dichotomiques à deux modalités : présence ou absence de valeur, conformément aux

---

[6] Cohen (1969, pp.278-280) propose les seuils de .0099, .0588, and .1379 pour les $\eta 2$ partiels en tant qu'indicateurs d'effets petit, moyen et fort, respectivement.

[7] La condition de normalité multivariée n'est pas respectée si z>5 et p<.05.

recommandations de Schlomer, Bauman & Card (2010). Puis deux régressions logistiques sont réalisées avec, pour l'explication des valeurs manquantes sur le score d'anxiété en T1, les prédicteurs suivants : l'âge, le sexe et la pédagogie et, pour T2, ces mêmes variables prédictrices plus la présence de valeurs manquantes en T1. Les analyses effectuées ne révèlent pas de relation entre l'absence/présence de données et les autres variables à l'exception de la variable âge : le fait d'être plus âge rend plus probable la présence de valeurs manquantes en T1 (OR = 0.24, p = .01). Pour illustrer cette étape du *data cleaning*, nous considérerons les données manquantes comme MAR, ce qui signifie que ces valeurs manquantes peuvent être en lien avec d'autres variables du jeu de données. Nous imputerons les valeurs manquantes des items (par la moyenne à l'item dans un premier temps puis par imputation multiple avec le package `mice` sous R) avant de recalculer les scores composites. Nous observons entre 0 et 4.84% de valeurs manquantes pour les items mesurés en T1 et pour T2, entre 0 et 6.45% de valeurs manquantes.

Les résultats de l'ANOVA mixte sont présentés dans le tableau 2 (ligne 5 et 6). Ceux-ci sont statistiquement significatif quelle que soit la méthode d'imputation. On constate que le remplacement des valeurs manquantes par la moyenne offre des écarts-type légèrement plus réduits. L'effet observé est sensiblement plus élevé avec la méthode d'imputation multiple.

3.3. Discussion

L'objectif de cette étude était de présenter puis d'illustrer les principales étapes du *data cleaning* à partir d'un jeu de données. Il ressort des résultats les éléments suivants ; pour le traitement des valeurs atypiques, lorsque celles-ci s'avèrent légitimes, il est préférable de les remplacer par une valeur plus plausible que de les supprimer ou de les conserver. Par ailleurs, s'interroger sur les valeurs qui se trouvent aux extrémités de la distribution est plus pertinent que de prendre en compte les valeurs à trois écart-types de la moyenne. Examiner l'allure de

la distribution est un moment clé du *data cleaning* pour comprendre comment les participants se sont comportés durant une étude. Néanmoins, transformer une variable afin de rendre son allure plus proche d'une distribution normale n'est pas une nécessité. La transformation effectuée sur les données n'a pas modifié le résultat de l'ANOVA. Par contre la transformation a pour inconvénient de dénaturer l'échelle de mesure et donc rend l'interprétation plus délicate (nous avons travaillé sur des racines carrées de score d'anxiété...). Enfin, les traitements effectués sur les valeurs manquantes ont montré que les méthodes telles que l'imputation multiple offre un résultat plus satisfaisant que les méthodes plus anciennes, comme le remplacement par la moyenne.

On constate, à travers l'analyse de ces données, que les choix effectués lors du *data cleaning* ne sont pas sans conséquence sur la confirmation des hypothèses aussi, plutôt que de rejeter ces étapes, il importe au contraire d'adopter la plus grande transparence dans les décisions prises par les chercheurs. Cette transparence dans le *data cleaning* fait écho aux recommandations sur les nouvelles pratiques de recherche portées par l'Open Science Framework[8], celles-ci insistant sur la nécessité d'une science ouverte et collaborative (mise à disposition des données, pré-enregistrement des protocoles de recherche,…). En cela, cette démarche de *data cleaning* devrait dans l'idéal être systématiquement présentée dans les publications scientifiques au sein de la section méthode afin d'assurer une approche standardisée, transparente et reproductible. De meilleures habitudes en ce domaine apporteraient des arguments face aux critiques liées à la difficulté de répliquer et seraient en ce sens bénéfiques aux sciences psychologiques.

REFERENCES

Auteurs (2017). *British Journal of Educational Psychology.*

[8] https://osf.io/

Çetinkaya-Rundel, M. (2014). Data Analysis and Statistical Inference. [MOOC offered by Duke University]. Retrieved from: https://www.coursera.org/learn/inferential-statistics-intro.

Chambers, C. (2017). *The seven deadly sins of psychology: A manifesto for reforming the culture of scientific practice*. Princeton University Press.

Cohen, J. (1969). *Statistical power analysis for the behavioural sciences.* New York: Academic Press.

Dong, Y., & Peng, C. Y. J. (2013). Principled missing data methods for researchers. *SpringerPlus, 2*(1), 222. https://doi.org/10.1186/2193-1801-2-222.

Field, A. (2009). *Discovering Statistics Using SPSS*. Sage publications.

Field, A. (2016). *An Adventure in Statistics: the Reality Enigma*. Sage publications.

Ghasemi, A., & Zahediasl, S. (2012). Normality tests for statistical analysis: a guide for non-statisticians. *International journal of endocrinology and metabolism*, *10*(2), 486. doi:10.5812/ijem.3505

Ghosh, D., & Vogt, A. (2012). Outliers: An evaluation of methodologies. In *Joint statistical meetings* (pp. 3455-3460). San Diego, CA: American Statistical Association.

Hoekstra, R., Kiers, H., & Johnson, A. (2012). Are assumptions of well-known statistical techniques checked, and why (not)?. *Frontiers in Psychology*, *3*, 137. https://doi.org/10.3389/fpsyg.2012.00137

Hawkins, D. M. (1980). *Identification of outliers* (Vol. 11). London: Chapman and Hall.

Husson, F., Lê, S., & Pagès, J. (2009). *Analyse de données avec R*. Rennes : Presse Universitaire de Rennes.

Judd, C. M., & McClelland, G.H. (1989). *Data analysis: A model-comparison approach.* San Diego, CA: Harcourt Brace Jovanovich.

Keselman, H. J., Huberty, C. J., Lix, L. M., Olejnik, S., Cribbie, R., Donahue, B., Kowalchuk, R. K., Lowman, L. L., Petoskey, M. D., Keselman, J. C. & Levin, J. R. (1998). Statistical practices of educational researchers: An analysis of their ANOVA, MANOVA, and ANCOVA analyses. *Review of Educational Research, 68*, 350-386. https://doi.org/10.3102/00346543068003350

Little, R. J., & Rubin, D. (1987). *Statistical analysis with missing data*. New York: Wiley.

Mahibbur Rahman, M., & Govindarajulu, Z. (1997). A modification of the test of Shapiro and Wilk for normality. *Journal of Applied Statistics*, 24(2), 219-236. https://doi.org/10.1080/02664769723828

Miles, J., & Shevlin, M. (2001). *Applying Regression & Correlation: A Guide for Students and Researchers*. London: Sage Publications Ltd.

Osborne, J. W. (2010). Data cleaning basics: Best practices in dealing with extreme scores. *Newborn and Infant Nursing Reviews, 10*(1), 37-43. https://doi.org/10.1053/j.nainr.2009.12.009

Osborne, J. W. (2012). *Best practices in data cleaning: A complete guide to everything you need to do before and after collecting your data*. Sage Publications.

Osborne, J. W. (2013). Is data cleaning and the testing of assumptions relevant in the 21st century?. *Frontiers in Psychology*, *4*, 370. https://doi.org/10.3389/fpsyg.2013.00370

Pekrun, R., Goetz, T., Frenzel, A. C., Barchfeld, P., & Perry, R. P. (2011). Measuring emotions in students' learning and performance: The Achievement Emotions Questionnaire (AEQ). *Contemporary educational psychology*, *36*(1), 36-48. https://doi.org/10.1016/j.cedpsych.2010.10.002

Peng, C. Y. J., Harwell, M., Liou, S. M., & Ehman, L. H. (2006). Advances in missing data methods and implications for educational research. *Real Data Analysis*, 3178.

Peugh, J. L., & Enders, C. K. (2004). Missing data in educational research: A review of reporting practices and suggestions for improvement. *Review of Educational Research, 74*(4), 525-556. https://doi.org/10.3102/00346543074004525

Schafer, J. L. (1999). Multiple imputation: a primer. *Statistical methods in medical research, 8*(1), 3-15. https://doi.org/10.1177/096228029900800102

Schlomer, G. L., Bauman, S., & Card, N. A. (2010). Best practices for missing data management in counseling psychology. *Journal of Counseling psychology, 57*(1), 1-10. http://dx.doi.org/10.1037/a0018082

Suits, D. B. (1957). Use of dummy variables in regression equations. *Journal of the American Statistical Association*, *52*(280), 548-551 DOI: 10.1080/01621459.1957.10501412

Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics*. Allyn & Bacon/Pearson Education.

Tableau 1

Etapes de la démarche systématique de *data cleaning*

Etapes	Vérification *Pourquoi ?*	Préconisation *Quoi ?*	R *Comment ?*
Traitement préliminaire des variables	1. S'assurer que la nature des variables (catégorielle ou continue) est correctement reconnue par le logiciel et qu'il n'y a pas d'erreur de saisie	Examen des statistiques descriptives (moyenne, écart-type, minimum, maximum)	`str(Dataset)` `summary(Dataset)`
	2. Pour les variables catégorielles : examiner la répartition des effectifs par modalité	Envisager des regroupements si des modalités comportent de faibles effectifs	`library(questionr)` `irec(Dataset, x)`
	3. Pour les variables continues, effectuer les renversements de scores si nécessaire et calculer les variables composites		`y<-with(Dataset,` x_{max}`-x)` `y<-with(Dataset,` x_1`+`x_2`)`
Vérification de la normalité	4. Examiner l'allure de la distribution des variables continues	Dresser une représentation graphique	`hist(x)` `QQnorm(x)`
	- S'assurer que la variable est uni modale (4a)	Découper la variable en plusieurs catégories si elle s'avère multimodale pour la traiter ensuite en tant que variable catégorielle	`binvar()`
	- Examiner les indices de dispersion (4b)	Suspendre les analyses si la variable ne varie pas ou trop peu	`sd (x)`

	- Examiner les indices d'asymétrie (4c)	Envisager des transformations
		- si forte asymétrie positive
		- si asymétrie positive modérée
		- si asymétrie négative
			`skewness (x)`
			`log(x)`
			`sqrt(x)`
			`renverser les scores puis log(x) ou sqrt(x)`

Gestion des valeurs atypiques	5. Identifier la présence de valeurs atypiques univariées et multivariées	- Outlier univarié : standardiser la variable et identifier les valeurs > |3|
		Ou bien considérer comme atypiques les valeurs aux extrémités de la distribution
		- Outlier multivarié : résidus standardisés ou studentisés, distance de Mahalanobis, distance de Cook, leverage, etc.
			`scale(x)`
			
			`quantile(x,probs = c(0.025,0.975), na.rm=TRUE)`

	6. Définir l'origine de la valeur atypique et appliquer le traitement conséquent.		
	- Erreur de saisie (6a) : Vérifier si des valeurs maximales excèdent l'étendue vraisemblable.	Corriger l'erreur de saisie si un retour au matériel original est possible, sinon supprimer l'observation.	
	- Réponse volontairement erronées (6b) : Examiner l'ensemble des réponses produites par un individu	Si l'ensemble des réponses apparait suspicieux, supprimer l'observation.	
	- Erreur d'échantillonnage (6c) : Vérifier les caractéristiques sociodémographiques du répondant	Si l'individu n'appartient pas à la population cible, supprimer l'observation.	
	- Défaut méthodologique (6d) : s'assurer des bonnes conditions du recueil de données	En cas de défaut confirmé, supprimer l'observation	
	- Distribution asymétrique (6e)	Transformer la distribution	`log(x), sqrt(x)`
	- Valeur atypique mais légitime (6f)	Remplacer la valeur atypique par une valeur plus plausible	`voir appendice`
Gestion des valeurs manquantes	7. Identifier pour chaque variable la part de valeurs manquantes :	Reporter systématiquement la part de ces valeurs dans les résultats Lorsque des valeurs manquantes existent pour un item impliqué dans un calcul de score total, imputer d'abord la valeur manquante avant de calculer le score total.	`summary(Dataset)`

- si elle est supérieure à 5%	Envisager de supprimer la variable quand elle n'est pas cruciale pour la recherche, sinon intégrer ce statut comme modalité dans les analyses	
- si elle est inférieure à 5%	Choisir de supprimer ou de remplacer la VM en fonction du modèle suivi.	`library(mice)` `voir appendice`
8. S'interroger sur la possibilité de valeurs manquantes non aléatoires (NMAR)		
9. Pour distinguer les valeurs MAR des MCAR : Créer une nouvelle variable dichotomique présence/absence de VM pour chaque variable (qui présente des VM) puis tester la relation entre cette variable et les autres variables du jeu de données	Effectuer une série de régressions logistiques	`glm(`x_{DV} `~` x_{IV1} `+` x_{IV2}`, family=binomial(logit), data=Dataset)`
- Si des relations sont trouvées, les données sont MAR	Remplacer les valeurs manquantes	`library(mice)` `voir appendice`
- Si on ne trouve pas de relation les données sont MCAR	Supprimer les observations avec VM ou remplacer les valeurs manquantes	`library(mice)` `voir appendice`

Note. Dataset désigne l'objet contenant le jeu de données, x désigne une variable, x_{max}, la valeur maximale prise par une variable.

Tableau 2

Moyennes, écart-types et résultats de l'ANOVA mixte en fonction des choix effectués lors du *data cleaning*

Traitement	Pédagogie A		Pédagogie B		*F*	*p*	Part. $\eta 2$
	T1	T2	T1	T2			
1. Absence de *data cleaning*	31.48	32.52	40.94	34.25	5.54	.024	0.119
	(10.47)	(10.34)	(14.71)	(11.55)			
2. Suppr. des *outliers* 5%	33.17	32.87	36.50	32.21	1.71	.200	0.047
	(10.05)	(8.54)	(8.92)	(10.35)			
3. Rempl. des *outliers* 5%	32.13	32.60	40.34	34.17	4.54	.039	0.102
	(10.12)	(10.13)	(13.39)	(11.38)			
4. Transf. racine carrée + (3)	5.60	5.64	6.28	5.77	4.40	.042	0.099
	(0.87)	(0.89)	(1.01)	(0.92)			
5. VM rempl. par la moy + (4)	5.62	5.63	6.43	5.98	4.16	.046	0.065
	(0.90)	(0.89)	(1.03)	(0.89)			
6. VM rempl. par IM + (4)	5.60	5.63	6.51	5.99	5.74	.02	0.087
	(0.91)	(0.89)	(1.07)	(0.93)			

Appendice

Identifier les valeurs atypiques

```
#Calculer les quantiles 2.5% et 97.5%

> numSummary(Dataset[,c("ANXIETY_T1", "ANXIETY_T2")],
statistics=c("quantiles"), quantiles=c(.025,.975))

             2.5% 97.5%  n NA

ANXIETY_T1 19.275 67.25 52 10

ANXIETY_T2 18.000 55.65 50 12


#Créer deux nouvelles variables sur lesquelles les valeurs atypiques
seront identifiées

> Dataset$ANXIETY_T1_outliers <- with(Dataset, ANXIETY_T1)

> Dataset$ANXIETY_T2_outliers <- with(Dataset, ANXIETY_T2)


#Remplacer les valeurs inférieures au quantile 2.5% et supérieure au
quantile 97.5% par le texte "outlier" pour les variables ANXIETY_T1
et ANXIETY_T2

> Dataset$ANXIETY_T1_outliers[Dataset$ANXIETY_T1_outliers<=19.275]<-
"Outlier"

> Dataset$ANXIETY_T1_outliers[Dataset$ANXIETY_T1_outliers>=67.25]<-
"Outlier"

> Dataset$ANXIETY_T2_outliers[Dataset$ANXIETY_T2_outliers<=18]<-
"Outlier"

> Dataset$ANXIETY_T2_outliers[Dataset$ANXIETY_T2_outliers>=55.65]<-
"Outlier"
```

Distribution des variables

```
#Appliquer une transformation racine carrée

Dataset$ANXIETY_T1_sqrt <- with(Dataset,
sqrt(ANXIETY_T1_winzorised))
```

```
Dataset$ANXIETY_T2_sqrt <- with(Dataset,
sqrt(ANXIETY_T2_winzorised))

#Pour calculer le coefficient de Mardia

>library(semTools)

> MARDIA<-mardiaKurtosis(Dataset);round(MARDIA,3)#au préalable,

construire un dataset avec uniquement les variables à tester
```

Traitement des valeurs manquantes

```
#Compter le nombre de valeurs manquantes dans chacune des colonnes

sapply(Dataset,function(x) sum(is.na(x))) # NA counts

#Compter le nombre de NA dans une colonne en particulier, ici la

variable "age"

sum(is.na(Dataset$age))

#Afficher les lignes dont la valeur dans la colonne "age" est NA

Dataset[is.na(Dataset$age),]

#recoder les variables ANXIETY 1 et ANXIETY 2 en variables
dichotomiques pour identifier le pattern des VM

Dataset$ANXIETY_T1_dicho <- with(Dataset, ANXIETY_T1_sqrt)

Dataset <- within(Dataset, {

  ANXIETY_T1_dicho <- Recode(ANXIETY_T1_dicho, 'NA="missing";
4.390:8.201="not_missing"', as.factor.result=TRUE)

})

Dataset$ANXIETY_T2_dicho <- with(Dataset, ANXIETY_T2_sqrt)

Dataset <- within(Dataset, {

  ANXIETY_T2_dicho <- Recode(ANXIETY_T2_dicho, 'NA="missing";
4.242:7.46="not_missing"', as.factor.result=TRUE)

})

#Effectuer une régression logistique expliquant la présence/absence
de valeurs manquantes à partir des autres variables du jeu de
données (AGE, PEDAGOGY, SEXE)
```

```
GLM.1 <- glm(ANXIETY_T1_dicho ~ age + pedagogy + sexe,
family=binomial(logit), data=Dataset)

summary(GLM.1)

exp(coef(GLM.1))  # Exponentiated coefficients ("odds ratios")

GLM.2 <- glm(ANXIETY_T2_dicho ~ age + pedagogy + sexe +
ANXIETY_T1_dicho, family=binomial(logit), data=Dataset)

summary(GLM.2)

exp(coef(GLM.2))  # Exponentiated coefficients ("odds ratios")


#Imputer les valeurs manquantes

library(mice)

init = mice(Dataset, maxit=0)

meth = init$method

predM = init$predictorMatrix

set.seed(103)

imputed = mice(Dataset, method=meth, predictorMatrix=predM, m=5)

imputed <- complete(imputed)

#Pour vérifier

sapply(Dataset, function(x) sum(is.na(x)))
```

Figure 1

Distribution de la variable Anxiété au temps 1 avant (à gauche) et après (à droite) une transformation racine carrée et un traitement des *outliers* sur QQ-plot (en haut) et histogramme (en bas)

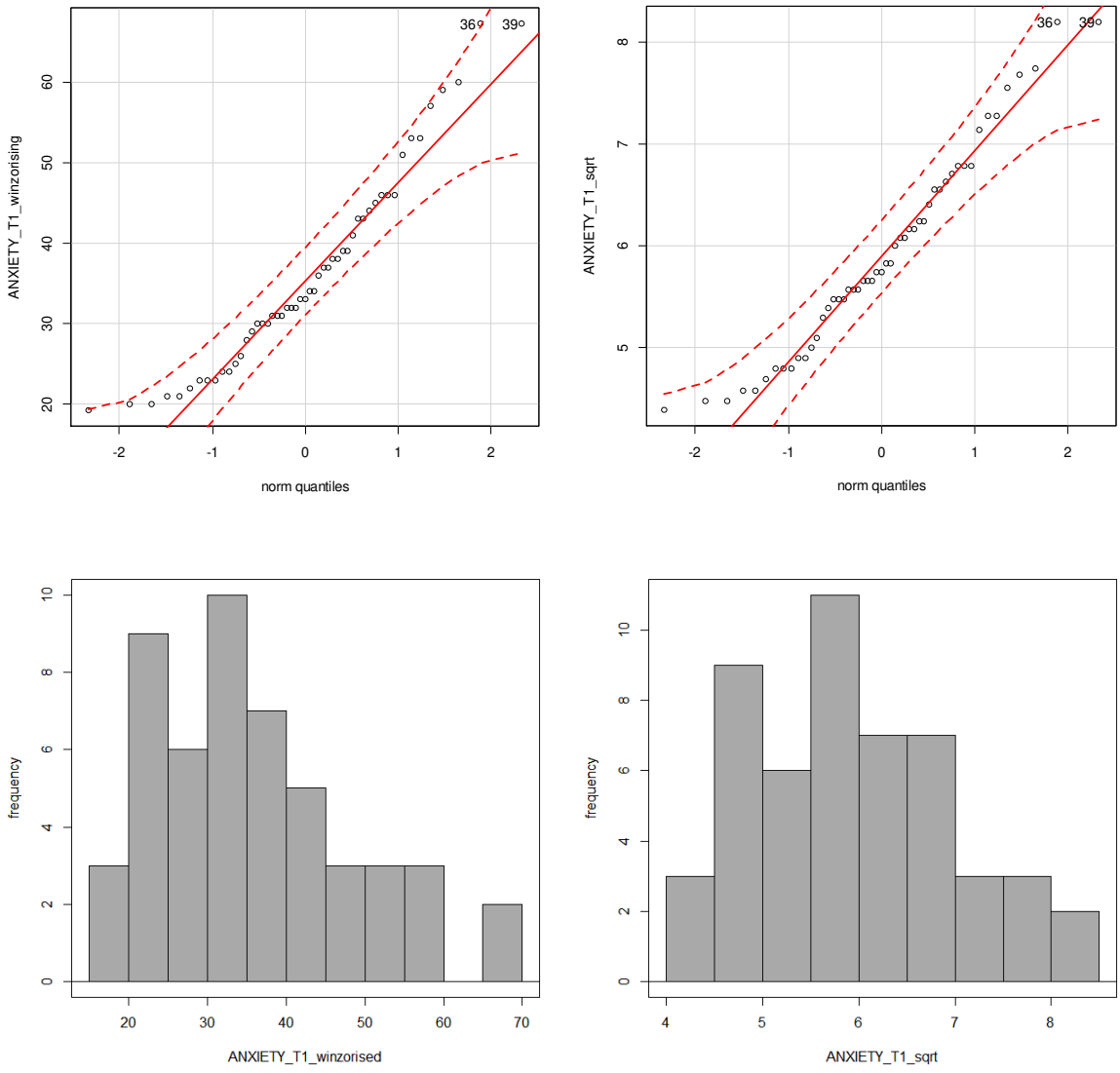

Figure 2

Distribution de la variable Anxiété au temps 2 avant (à gauche) et après (à droite) une transformation racine carrée et un traitement des *outliers* sur QQ-plot (en haut) et histogramme (en bas)