

Détection et traitement des valeurs hors-normes avec R

Table des matières

1. INTRODUCTION	1
2. DEFINITION DES "HORS-NORMES"	2
3. DETECTION	2
3.1. LE RESUME DES DONNEES	2
3.2. UN GRAPHIQUE EN POINTS	2
3.3. LES GRAPHIQUES XY EN NUAGES DE POINTS.....	3
3.4. LES BOITES DE DISPERSION	3
3.5. LA DISTANCE DE COOK.....	4
3.6. LE PACKAGE { OUTLIERS }	5
4. ANALYSES MULTIFACTORIELLES	6
4.1. ANALYSE EN COMPOSANTES PRINCIPALES (ACP)	6
4.2. ANALYSES FACTORIELLES DES CORRESPONDANCES (AFC & ACM).....	7
5. TRAITEMENT	8
5.1. ENLEVER LES INDIVIDUS CONCERNES.....	8
5.2. IMPUTATION OU REMPLACEMENT	9
REFERENCES	9

1. Introduction

Qu'elles soient très largement au-dessus ou en-dessous de la moyenne, les valeurs hors-normes (*outliers*) affectent à la fois la moyenne, la variance d'une série de valeurs et sa distribution, qui se trouve alors souvent éloignée de la normalité. Elles sont particulièrement nuisibles lors d'analyses de régressions ou de corrélations.

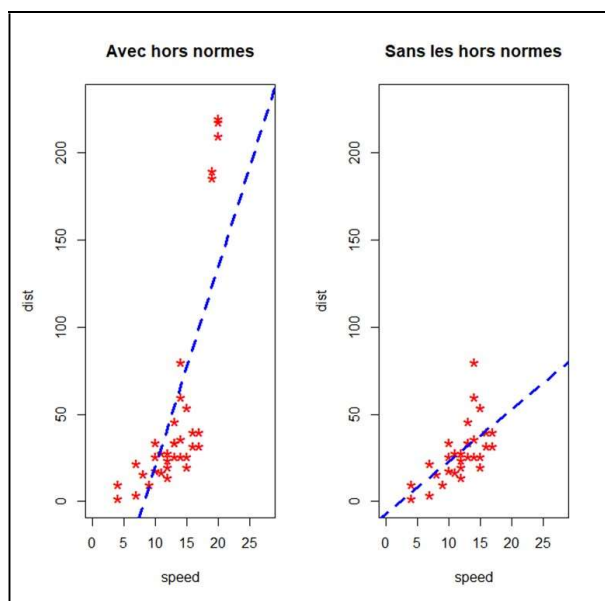
Pour en donner une illustration nous introduisons artificiellement ici une série de valeurs hors-normes dans les données 'cars'.

```
# Chargement des données 'cars' {datasets}
data(cars)

# Ajout de hors-normes
cars1 <- cars[1:30, ] # données d'origine, lignes 1 à 30.
cars_outliers <- data.frame(speed=c(19,19,20,20,20),
                             dist=c(190, 186, 210, 220, 218))
cars2 <- rbind(cars1, cars_outliers) # données avec hors-normes.

# Graphique avec hors-normes
par(mfrow=c(1, 2))
plot(cars2$speed, cars2$dist, xlim=c(0, 28),
     ylim=c(0, 230), main="Avec hors normes",
     xlab="speed", ylab="dist", pch="*", col="red", cex=2)
abline(lm(dist ~ speed, data=cars2), col="blue", lwd=3, lty=2)

# Graphique des données d'origine sans hors-normes.
plot(cars1$speed, cars1$dist, xlim=c(0, 28),
     ylim=c(0, 230), main="Sans les hors normes",
     xlab="speed", ylab="dist", pch="*", col="red", cex=2)
abline(lm(dist ~ speed, data=cars1), col="blue", lwd=3, lty=2)
```



Noter le changement de pente et le meilleur ajustement sans les valeurs hors-normes. Si la modélisation est faite avec les valeurs hors-normes (graphique de gauche) les prédictions de "dist" sont exagérées pour les valeurs élevées de "speed" à cause de la pente excessive.

2. Définition des "hors-normes"

Pour une série de valeurs continues, l'espace interquartile (IQR) comprend 50% des valeurs autour de la médiane, c'est à dire les valeurs comprises entre Q3, le 75^{ème} et Q1, le 25^{ème} percentile de la distribution. Dans un graphique en boîte de dispersion (*boxplot*) ce sont les bordures inférieure et supérieure de la boîte. Les "moustaches" du boxplot sont situées à 1.5 fois la différence Q3-Q1 au-delà des 75^{ème} et 25^{ème} percentiles. On considère généralement que les valeurs situées à l'extérieur des "moustaches" sont hors-normes.

3. Détection

Divers stades d'exploration des données peuvent permettre de détecter des valeurs hors-normes.

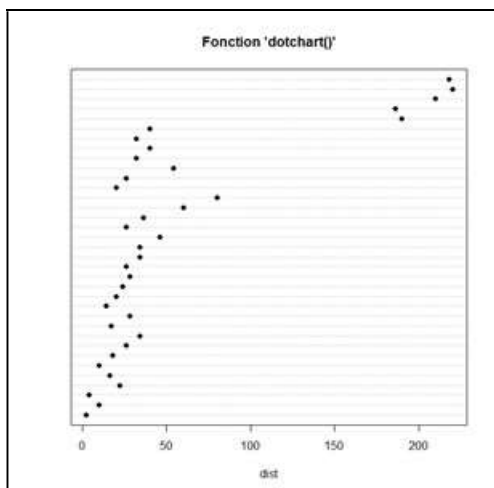
3.1. Le résumé des données

summary(cars2)	
speed	dist
Min. : 4.00	Min. : 2.0
1st Qu.: 10.50	1st Qu.: 20.0
Median : 13.00	Median : 28.0
Mean : 13.03	Mean : 53.8
3rd Qu.: 15.50	3rd Qu.: 43.0
Max. : 20.00	Max. : 220.0

Outre la présence d'une valeur 'Max' très supérieure au 3^{ème} quartile, on remarque une grande différence entre la médiane et la moyenne, cette dernière étant influencée par au moins une valeur hors-norme.

3.2. Un graphique en points

```
dotchart(cars2$dist, main="Fonction 'dotchart()'",
         xlab="dist", pch=16)
```

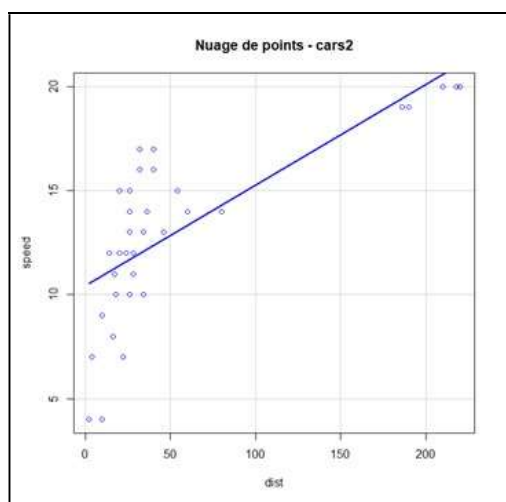


5 valeurs hors normes se repèrent très bien dans les valeurs de la variable 'dist'.

3.3. Les graphiques xy en nuages de points

Souvent utilisés sous forme de matrices de nuages de points, les valeurs hors-normes y sont facilement repérables. En outre ces graphiques permettent de visualiser si les distributions s'écartent de la normalité.

```
scatterplot(speed~dist, regLine=TRUE, smooth=FALSE,
            boxplots=FALSE, data=cars2)
```



D'une part les points hors normes sont facilement repérables. D'autre part on voit nettement leur influence sur la ligne des moindres carrés qui, de leur fait, semble indiquer une forte corrélation positive qui ne serait peut-être pas le cas sans eux.

3.4. Les boîtes de dispersion

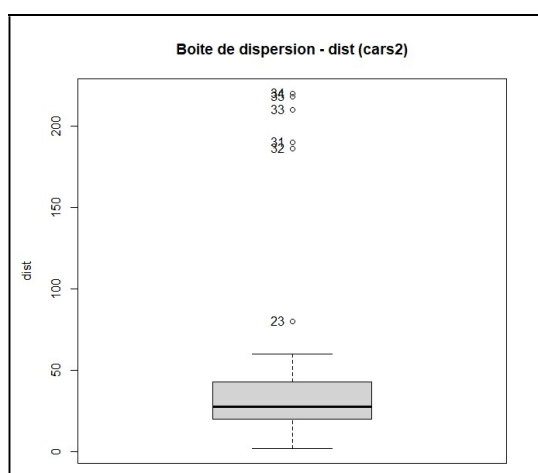
Outre qu'elles renseignent sur une éventuelle dissymétrie de distribution, les boîtes de dispersion ou diagrammes de Tukey (*box & whiskers plot*, *boxplot*) présentent des limites, extrémités des "moustaches", proposées par Tukey :

- on calcule $i = Q1 - 1.5(Q3 - Q1)$; on recherche dans les valeurs celle qui est la plus proche de i ; cette valeur est la limite inférieure.
- on calcule $s = Q3 + 1.5(Q3 - Q1)$; on recherche dans les valeurs celle qui est la plus proche de s ; cette valeur est la limite supérieure.

Des fonctions graphiques permettent d'identifier les valeurs extérieures aux "moustaches" par leur numéro de ligne.

```
Boxplot( ~ dist, data=cars2, id=list(method="y"),
        main="Boite de dispersion - dist (cars2)")
```

```
[1] "23" "31" "32" "33" "34" "35"
```



Les points extérieurs aux "moustaches" sont indiqués sur le graphe et listés dans la console.

3.5. La distance de Cook

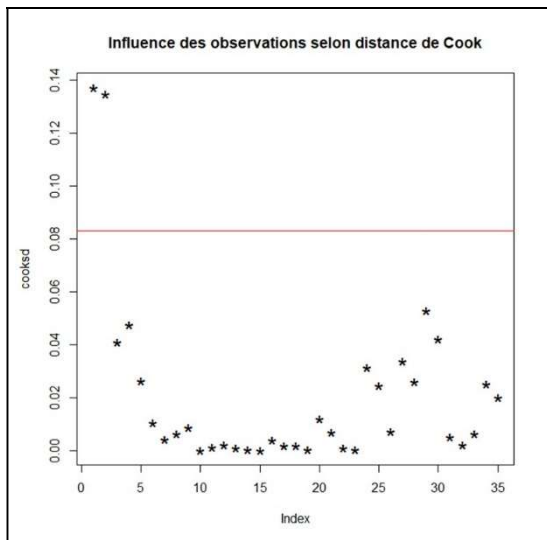
La distance de Cook est une mesure calculée à l'issue de la construction d'un modèle de régression, c'est à dire utilisant à la fois variable à expliquer et variables explicatives. Elle n'est impactée que par les variables explicatives du modèle. Elle calcule l'influence exercée par chaque ligne des données sur la prévision du modèle. Pour chaque observation i , la distance de Cook évalue le changement opéré sur la variable y prédite selon que i est présente ou absente.

Modélisation et calcul des distances de Cook

```
mod <- lm(speed~dist, data=cars2)
cooks_d <- cooks.distance(mod)
```

Graphique.

```
plot(cooks_d, pch="*", cex=2,
      main="Influence des observations selon distance de Cook")
abline(h = 4*mean(cooks_d, na.rm=T), col="red")
```



On considère généralement qu'au-dessus de 4 fois la distance moyenne les observations ont une influence trop forte. Ce seuil est matérialisé par une ligne horizontale ; mais il est arbitraire et ne va pas toujours isoler tous les points très influents sur le graphique. De plus certains peuvent y être superposés.

Il est ensuite très utile de lister les observations. Celles qui ont une trop forte influence seront marquées d'une étoile.

```
influence.measures(mod,
                    infl = influence(mod))
```

Influence measures of `lm(formula = speed ~ dist, data = cars2)`

	dfb.1	dfb.dist	dffit	cov.r	cook.d	hat	inf
1	-0.56496	0.356613	-0.5651	0.775	0.137179	0.0475	*
2	-0.56415	0.321321	-0.5669	0.736	0.134892	0.0421	*
3	-0.28890	0.178122	-0.2891	1.003	0.040886	0.0461	
4	-0.30774	0.141510	-0.3167	0.937	0.047679	0.0357	
...							
29	0.31405	-0.109188	0.3371	0.895	0.052887	0.0319	
30	0.26336	-0.063166	0.2982	0.923	0.042058	0.0299	
31	0.02593	-0.089917	-0.0993	1.261	0.005072	0.1593	*
32	0.01633	-0.058847	-0.0653	1.252	0.002198	0.1518	*
33	0.03417	-0.102274	-0.1104	1.326	0.006280	0.2005	*
34	0.07311	-0.207212	-0.2219	1.355	0.025253	0.2233	*
35	0.06453	-0.184751	-0.1982	1.349	0.020158	0.2186	*

3.6. Le package {outliers}

Ce package propose un certain nombre de fonctions relatives aux données hors-normes, par exemple une fonction pour réaliser le test de Dixon. La fonction `scores()` est particulièrement utile pour identifier les valeurs hors normes d'une variable particulière.

Arguments de la fonction `scores()` :

`x` = un vecteur de données numériques

type : "z" calcule le score normalisé de chaque valeur (différence entre chaque valeur et la moyenne divisée par l'écartype).

"t" calcule le score du t de Student.

"chisq" calcule le score de χ^2 (carré des différences entre chaque valeur et la moyenne divisée par la variance).

"iqr" ne concerne que les valeurs inférieures au 1^{er} quartile ou supérieures au 3^{ème} quartile. La différence entre la valeur et le plus proche quartile, divisée par le quartile est calculée. Toutes les valeurs comprises entre les quartiles sont mises à zéro.

"mad" calcule les différences entre chaque valeur et la médiane, divisée par l'écart absolu à la médiane.

prob : Si renseigné, ce sont les p-values et non les scores qui sont calculées. Si prob=1 toutes les p-values sont indiquées. Si une p-value est indiquée une indication logique (TRUE ou FALSE) est renvoyée pour chaque valeur. Le type "iqr" ne supporte pas l'indication de prob mais l'argument lim peut être indiqué.

lim : ne concerne que les scores de type "iqr". Un vecteur logique est renvoyé indiquant les valeurs qui excèdent cette limite.

Exemples

```
library(outliers)
```

```
scores(cars2$dist, type="z", prob=0.99)
```

```
[1] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[13] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[25] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE TRUE TRUE TRUE
```

```
scores(cars2$dist, type="iqr")
```

```
[1] -0.78260870 -0.43478261 -0.69565217 0.00000000 -0.17391304 -0.43478261
[7] -0.08695652 0.00000000 0.00000000 -0.13043478 0.00000000 -0.26086957
[13] 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000
[19] 0.13043478 0.00000000 0.00000000 0.73913043 1.60869565 0.00000000
[25] 0.00000000 0.47826087 0.00000000 0.00000000 0.00000000 0.00000000
[31] 6.39130435 6.21739130 7.26086957 7.69565217 7.60869565
```

```
scores(cars2$dist, type="iqr", lim=1)
```

```
[1] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[13] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE TRUE FALSE
[25] FALSE FALSE FALSE FALSE FALSE FALSE TRUE TRUE TRUE TRUE TRUE
```

```
scores(cars2$dist, type="mad", prob=1)
```

```
[1] 0.05543993 0.13485901 0.07056227 0.35647239 0.23092375 0.13485901
[7] 0.26988084 0.45119819 0.64352761 0.24999968 0.50000000 0.19532401
[13] 0.31187636 0.40312393 0.50000000 0.45119819 0.64352761 0.64352761
[19] 0.86514099 0.45119819 0.68812364 0.97512778 0.99928494 0.31187636
[25] 0.45119819 0.94456007 0.59687607 0.76907625 0.59687607 0.76907625
[31] 1.00000000 1.00000000 1.00000000 1.00000000 1.00000000
```

On remarque que les valeurs repérées par les différents calculs ne sont pas les mêmes. Le calcul utilisant `type="z"` sera certainement plus pertinent avec des valeurs nombreuses et normalement distribuées.

L'argument `type="iqr"` permet de décider d'une limite à partir de laquelle on considère que la valeur est hors-normes ; avec `lim=1` on demande de repérer les cas où l'écart avec le plus proche quartile vaut au moins une fois le quartile.

L'argument `type="mad"` associé à `prob=1` fournit des p-values ; plus la p-value est faible plus la valeur est proche de la médiane. On retrouve ici les 5 points très écartés de la médiane et 3 points (soulignés ci-dessus) dont la p-value est très proche de 1.

4. Analyses multifactorielles

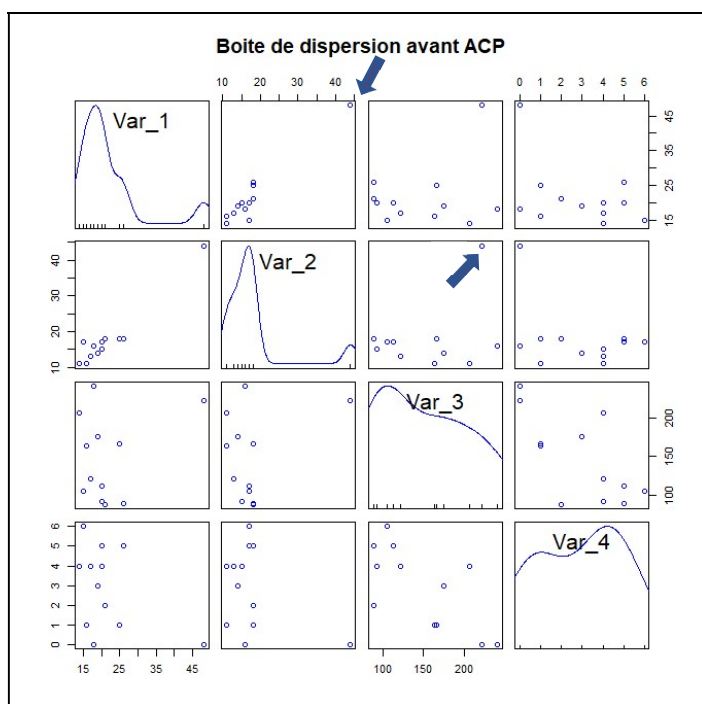
Comme leur nom l'indique, ces analyses portent sur plusieurs variables *à la fois*, c'est à dire que leurs résultats ne sont pas imputables à une seule variable. Dans ces conditions on parlera plutôt *d'individus* (ou de *modalités*) hors-normes ou "extraordinaires" dans la mesure où les "normes" vues plus haut (boxplots par exemple) ne concernent que chaque variable prise séparément.

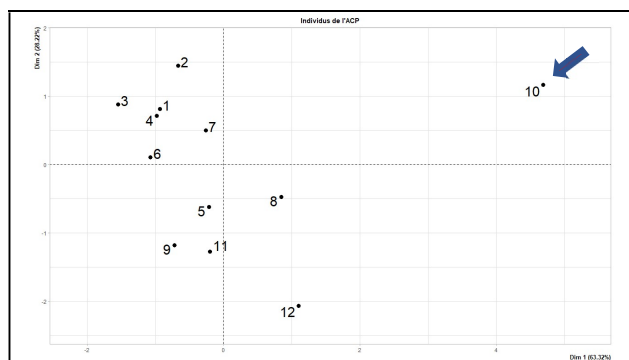
4.1. Analyse en composantes principales (ACP)

Une matrice de nuages de points constitue souvent un premier regard sur les données préalablement à une ACP. Des individus hors-normes y seront souvent repérables.

sujets	Var_1	Var_2	Var_3	Var_4
S1	20	17	112	5
S2	26	18	89	5
S3	15	17	105	6
S4	20	15	92	4
S5	19	14	175	3
S6	17	13	121	4
S7	21	18	88	2
S8	25	18	166	1
S9	14	11	207	4
S10	48	44	222	0
S11	16	11	164	1
S12	18	16	241	0

Ils seront en outre souvent responsables d'écarts à la normalité, comme ci-contre pour les variables 1 et 2.





Lors de l'ACP ils auront une position excentrée sur le graphe des individus et seront facilement identifiables, ainsi que sur le listing des contributions :

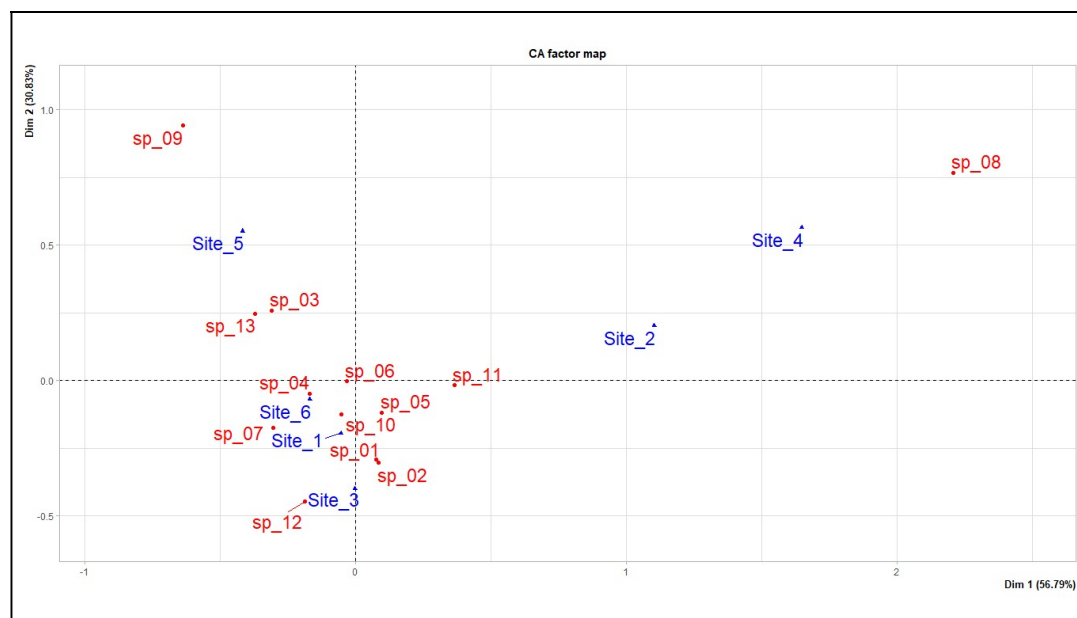
```
$contrib
      Dim.1      Dim.2
1  2.8899474  4.88528891
2  1.4836231 15.41347232
...
9  1.7318390 10.32315509
10 72.1890181 10.01251433
11  0.1338773 11.98490789
12  3.9629161 31.50489716
```

Après avoir vérifié qu'il ne s'agit pas d'erreurs de saisie ou de notation, les valeurs pourront être corrigées par imputation (voir plus bas § 5) ou les individus concernés pourront soit être enlevés de l'analyse, soit placés en supplémentaires de manière à faciliter la lecture des autres points.

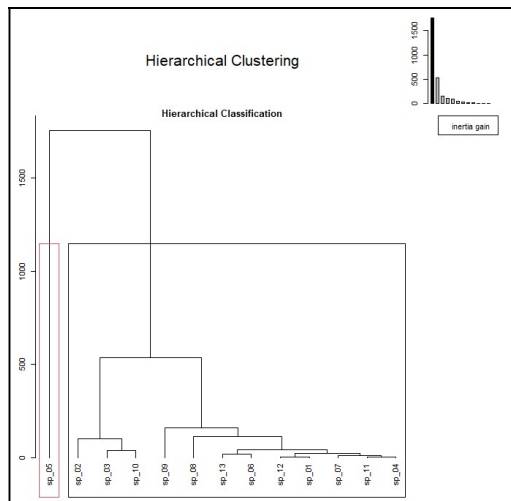
4.2. Analyses factorielles des correspondances (AFC & ACM)

Les positions relatives des individus en lignes et des variables ou des modalités en colonnes dans une AFC ou une ACM sont ici fonction de leurs différences de profils, ce qui est parfois difficile à voir dans le tableau des données, surtout s'il comporte beaucoup de cellules. En outre, des profils très originaux peuvent aussi bien être le fait de lignes que de colonnes du tableau. C'est la notion même de "norme" qui doit être repensée ici.

On donne ci-dessous un exemple imaginaire d'une AFC de comptages d'individus de 13 espèces dans 6 sites différents. Une fois de plus, outre les graphes, les listings de contributions des lignes et des colonnes permettront de repérer l'origine des particularités.



Lors d'une AFC ou d'une ACM les points ayant des profils particuliers sont souvent des révélateurs intéressants de particularités qui n'étaient pas toujours décelables dès la collecte des données. Une réflexion approfondie sera donc nécessaire avant de décider de modifier ou non les valeurs hors-normes ou de placer certains individus ou certaines variables en supplémentaires.



Si les analyses factorielles sont suivies d'une classification sur facteurs, comme ci-contre pour les lignes du tableau de l'AFC précédente, des classes de très faible effectif apparaîtront.

5. Traitement

Une fois les valeurs hors-normes repérées, la première chose à faire est bien sûr de vérifier qu'il ne s'agit pas d'erreurs de saisie ou de mesure. Dans ces cas bien sûr il faudra les corriger, à condition que l'information soit disponible ou que la mesure soit encore possible. Mais cela n'est pas obligatoirement le cas. Il se peut aussi qu'elles reflètent une variation naturelle et peuvent parfois être l'occasion d'intéressantes découvertes. Le fait d'enlever ou de conserver des valeurs hors-normes nécessite une certaine réflexion. **En tout état de cause il n'est pas conseillé d'enlever des valeurs simplement parce qu'elles sont hors-normes.**

5.1. Enlever les individus concernés

Pour ne conserver que les individus inclus entre les limites définies par les "moustaches" du boxplot :

Calculer et stocker l'écart interquartile

```
iqr <- IQR(cars2$dist)
```

Calculer et stocker les limites supérieure et inférieure

```
up <- Q[2]+1.5*iqr # Limite supérieure
```

```
low <- Q[1]-1.5*iqr # Limite inférieure
```

Créer un nouveau tableau de données

```
cars2.1 <- subset(cars2, cars2$dist > low & cars2$dist < up)
```

Vérifications

```
dim(cars2)
```

```
[1] 35  2
```

```
dim(cars2.1)
```

```
[1] 29  2 # 6 lignes ont été enlevées
```

5.2. Imputation ou remplacement

Contrairement aux cas de données manquantes, le remplacement d'une valeur hors-normes par une valeur choisie (ou calculée par un algorithme) ne doit pas se faire sans réflexion. Si une valeur extrême est porteuse d'une information importante elle peut très bien être conservée. Des procédures statistiques non paramétriques utilisant les rangs des valeurs pourront alors parfois être mises en œuvre. La plupart du temps, soit la valeur hors-normes est intéressante et on la conservera, soit l'individu sera simplement supprimé.

Le remplacement par la moyenne, la médiane ou le mode de la variable peut être effectué avec une méthode comparable à celle utilisée pour l'imputation des valeurs manquantes (voir document "Gestion des données manquantes avec R.pdf"). Il est aussi possible d'utiliser la fonction `replace()`, par exemple pour remplacer quelques valeurs anormalement élevées par la valeur de la "moustache" supérieure d'un boxplot.

Exemple

Calcul de la limite supérieure

```
iqr <- IQR(cars2$dist)
```

```
Q[2]+1.5*iqr
```

```
75%
```

```
68.5 # limite supérieure
```

Nouvelle variable corrigée : remplacement par la valeur de la moustache supérieure.

```
y <- cars2$dist
```

```
vec <- as.vector(replace(y, y > 68.5, 68.5))
```

Attachement du vecteur créé

```
cars2$distNew <- vec
```

Vérification

```
summary(cars2)
```

speed		dist		distNew	
Min.	: 4.00	Min.	: 2.0	Min.	: 2.0
1st Qu.	:10.50	1st Qu.	: 20.0	1st Qu.	:20.0
Median	:13.00	Median	: 28.0	Median	:28.0
Mean	:13.03	Mean	: 53.8	Mean	:34.0
3rd Qu.	:15.50	3rd Qu.	: 43.0	3rd Qu.	:43.0
Max.	:20.00	Max.	:220.0	Max.	:68.5

Références

<https://www.r-bloggers.com/2016/12/outlier-detection-and-treatment-with-r/>

<https://www.r-bloggers.com/2020/01/how-to-remove-outliers-in-r-2/>

Logiciel utilisé : R 4.0.4.

Remarques et suggestions : <https://www.anastats.fr/equipe-et-contact/>