Ann. Fr. Med. Urgence (2016) 6:280-283
DOI 10.1007/s13341-016-0654-8

MÉTHODOLOGIE / *METHODOLOGY*

Comprendre et utiliser le logiciel R pour vos analyses statistiques

Understand and Use R Software for your Statistical Analysis

P.-G. Claret · X. Bobbia · J.-E. de La Coussaye · P. Landais

Reçu le 9 février 2016 ; accepté le 12 mai 2016

Introduction : qu'est-ce que R et pour quoi faire ?

L'analyse statistique est souvent considérée par les médecins comme un frein à la recherche clinique. Cependant, une partie des analyses publiées dans les revues médicales ne requièrent pas un niveau avancé en biostatistique. De nombreux logiciels sont disponibles pour réaliser des analyses statistiques. Parmi eux, R est un logiciel libre (licence GNU GPL), gratuit et multiplateforme qui fonctionne sous Windows, Mac ou Linux. Son téléchargement peut se faire à l'adresse suivante : https://cran.r-project.org/. R est grandement utilisé dans les départements de biostatistique et d'informatique médicale hospitaliers. De nombreux tutoriaux sont également disponibles sur Internet, des plus simples [1–3] aux plus avancés. Cette large distribution permet d'obtenir rapidement de l'aide, y compris au sein de la majorité des institutions hospitalières. R fonctionne avec un système de packages. Un package contient des collections de fonctions centrées sur un sujet particulier. Lors de l'installation de R, un grand nombre de packages sont déjà installés par défaut, certains plus spécifiques doivent être téléchargés et installés si besoin. Cette installation peut se faire à partir de l'interface graphique de R, dans le menu « Packages & Données ». Une fois installé, un package doit être chargé avec, en début de script, la fonction « library (nom_du_package) ». Par ailleurs, R est sensible à la casse, c'est-à-dire que « a » et « A » sont deux symboles différents et représentent deux variables différentes.

R est un outil très performant, mais ce n'est pas parce que l'on dispose d'outils puissants que cela exonère de connaître les propriétés des objets que l'on manipule et des règles sous-jacentes. L'objectif de cet article est d'introduire pour le clinicien quelques notions de base d'analyse statistique simple avec R, pas de se substituer à une analyse avancée nécessairement réalisée par un spécialiste des biostatistiques.

Dans cet article, les commandes saisies sous R sont indiquées avec une police à chasse fixe et précédée de l'invite de commande « > » :

```
> Sys.time()
```

Le résultat tel qu'affiché par R, dans cet exemple l'heure et la date du système, est également indiqué avec une police à chasse fixe :

```
[1] 《2015-12-18 09:20:00 CET》
```

Nous illustrerons cet article grâce à l'exemple fictif d'une cohorte de 14 patients hospitalisés pour pneumopathies bactériennes. Les caractéristiques relevées seront le sexe et l'âge des patients. Ces patients seront traités soit avec un antibiotique, soit avec un placebo. Le taux de CRP et le retour à domicile à 48 heures seront les critères de jugement du traitement. Cinq variables seront donc analysées : la variable « age » pour l'âge, « sexe » pour le sexe, « groupe » pour le traitement, « crp » pour le taux de CRP et « rad » pour le retour à domicile. Le fichier est montré dans le Tableau 1.

Comment entrer les données ?

Une première méthode, fastidieuse mais didactique, permet d'entrer, pour chaque variable, les données qui la composent. Pour la variable « age », les 14 âges des patients sont entrés ainsi :

```
> age <- c(48, 56, 76, 89, 45, 23, 87, 72, 91, 71, 58, 20, 69, 93)
```

P.-G. Claret (✉) · X. Bobbia · J.-E. de La Coussaye
Pôle anesthésie-réanimation–douleur–urgences, CHU de Nîmes,
1, place du Professeur-Robert-Debré, F-30029 Nîmes, France
e-mail : pierre.geraud.claret@gmail.com

P.-G. Claret · P. Landais
EA 2415, institut universitaire de recherche clinique,
université de Montpellier, 641, avenue du Doyen-Gaston-Giraud,
F-34093 Montpellier, France

J.-E. de La Coussaye
Université de Montpellier,
Faculté de médecine de Montpellier-Nîmes,
2, rue École-de-Médecine, F-34060 Montpellier, France

P. Landais
Laboratoire de biostatistique, épidémiologie,
santé publique et informatique médicale, CHU de Nîmes,
1, place du Professeur Robert-Debré, F-30029 Nîmes, France

Tableau 1 Caractéristiques des patients.

Sexe	Âge (ans)	Groupe	CRP	Rad
M	48	ATB	26	TRUE
F	56	ATB	31	FALSE
M	76	ATB	54	TRUE
F	89	ATB	27	TRUE
F	45	ATB	82	TRUE
M	23	ATB	83	TRUE
F	87	ATB	47	TRUE
F	72	PCB	209	FALSE
M	91	PCB	127	FALSE

La lettre « c » qui précède la parenthèse ouverte signifie concaténer. Après la ligne de commande, la validation est faite en appuyant sur entrée.

Les autres variables sont ensuite entrées :

```
> sexe <- c("M", "F", "M", "F", "F", "M", "F", "F", "M", "F", "M", "F", "F")
> groupe <- c("ATB", "ATB", "ATB", "ATB", "ATB", "ATB", "ATB", "PCB", "PCB", "PCB", "PCB", "PCB", "PCB", "PCB")
> crp <- c(26, 31, 54, 27, 82, 83, 47, 209, 127, 142, 249, 120, 91, 102)
> rad <- c(TRUE, FALSE, TRUE, TRUE, TRUE, TRUE, TRUE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, TRUE)
```

Ces trois dernières lignes permettent d'illustrer la différence d'écriture entre les variables catégorielles (entre guillemets), numériques (telles qu'elles) et logiques (TRUE ou FALSE). Une dernière commande permet de reconstruire le tableau et d'accoler les colonnes les unes à côté des autres :

```
> data <- data.frame(sexe, age, groupe, crp, rad)
```

Automatiser la lecture d'un fichier est indispensable lorsque le set de données est déjà répertorié dans une ou plusieurs bases de données. Dans ce contexte, une deuxième méthode, plus pragmatique, permet de lire un fichier dans un certain format par exemple un fichier CSV (*comma-separated values*) obtenu à partir d'un fichier Excel. Il s'agit d'un format informatique de type texte représentant des données de tableaux Excel sous forme de valeurs séparées par des virgules. Chaque ligne du texte correspond en fait à une ligne du tableau et les virgules correspondent aux séparations entre les colonnes, soit au contenu de chaque cellule du tableau. Il faut préparer le fichier Excel puis le lire dans R. Les séparateurs et les décimales ne sont pas codés de la même façon selon qu'Excel est en version française ou anglo-saxonne. Ainsi, pour les utilisateurs d'une version française d'Excel, il faudra écrire :

```
> data <- read.table("fichier.csv", header = TRUE, sep = ";", dec = ",")
```

Pour les utilisateurs d'une version américaine :

```
> data <- read.table("fichier.csv", header = TRUE, sep = ",", dec = ".")
```

Comment vérifier les données ?

Il est nécessaire de vérifier ses données pour corriger les éventuelles erreurs ou incohérences de saisie. Par exemple, en français, une décimale est définie par une virgule à la place d'un point en format anglo-saxon. Il faut aussi vérifier le format des dates qui doit être analogue dans tout le fichier. La vérification de la qualité des données, avant tout traitement, est primordiale : c'est l'étape du *data management*. Pour visualiser les données d'une variable, il faut taper son nom dans la console et valider :

```
> age
[1] 48 56 76 89 45 23 87 72 91 71 58 20 69 93
```

La variable « age » est bien composée des 14 valeurs de la colonne « age » de notre tableau.

La commande « class » permet d'afficher les caractéristiques d'une variable :

```
> class(sexe)
[1] "character"
> class(age)
[1] "numeric"
> class(rad)
[1] "logical"
```

Ces trois dernières lignes nous montrent que R considère la variable « sexe » comme alphabétique (M ou F), la variable « age » comme numérique et la variable « rad » comme logique (vrai ou faux). Dans l'exemple de la variable « sexe », il est possible de la transformer secondairement en variable catégorielle avec la fonction « as.factor ».

Comme pour afficher le contenu d'une variable, taper directement le nom d'un tableau permet de l'afficher. Les

dimensions de ce tableau (lignes, colonnes) peuvent être obtenues avec la commande « dim » :

```
> dim(data)
[1] 14 5
```

Le tableau est bien composé de 14 lignes qui correspondent aux 14 patients et de cinq colonnes qui correspondent aux cinq variables étudiées.

Comment traiter les données

Il convient de différencier les variables qualitatives et quantitatives. Les variables qualitatives sont nominales ou ordinales. Une variable est qualitative nominale quand ses valeurs sont des éléments d'une catégorie non hiérarchique (exemple : papule, macule, vésicule). Une variable qualitative est dite ordinale quand ses valeurs sont des éléments d'une catégorie hiérarchique (exemple : jamais, rarement, parfois, fréquemment, toujours). Les variables quantitatives sont discrètes ou continues. Une variable discrète correspond à une valeur finie, dans un ensemble énumérable. En revanche, une variable continue peut prendre, en théorie, une infinité de valeurs (exemple : âge, poids, taux de CRP).

Pour les variables qualitatives :

La commande « table » permet d'obtenir un tableau de contingence à un ou plusieurs niveaux :

```
> table(sexe)
sexe
F M
9 5
> table(groupe, rad)
rad
groupe FALSE TRUE
ATB 1 6
PCB 6 1
```

Ce tableau peut être lui-même entré comme un objet :

```
> tableau <- table(groupe, rad)
```

La commande « prop.table » permet d'obtenir les proportions de chaque cellule de ce nouvel objet :

```
> prop.table(tableau)
rad
groupe FALSE TRUE
ATB 0.07142857 0.42857143
PCB 0.42857143 0.07142857
```

Pour analyser ce dernier tableau de contingence, la commande « fisher.test » permet de réaliser un test exact de Fisher :

```
> fisher.test(groupe, rad)
```

Cette commande est équivalente à la commande :

```
> fisher.test(tableau)
```

Les dernières lignes du résultat de cette commande sont :

```
p-value = 0.02914
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
0.000531383 0.799817637
sample estimates:
odds ratio
0.04256025
```

La *p-value* de ce test est donc $p = 0{,}029$ avec un odds ratio de 0,04 dont l'intervalle de confiance est de 0,0005 à 0,7998. L'intervalle de confiance de l'odds ratio ne contient pas la valeur 1, il est donc significativement différent de 1. Le sens du codage de la variable indique que plus de patients rentrent à domicile à 48 heures lorsqu'ils sont traités par une antibiothérapie.

Pour les variables quantitatives :

La commande « summary » permet d'obtenir en une seule commande la valeur minimum, le premier quartile, la médiane, la moyenne, le troisième quartile et la valeur maximale de la variable :

```
> summary(age)
```

L'écart-type est obtenu avec la commande « sd » :

```
> sd(age)
```

Pour obtenir la moyenne des taux de CRP, seulement pour les patients ayant bénéficié d'une antibiothérapie, il est nécessaire d'utiliser une commande d'indexation :

```
> mean(crp[groupe == "ATB"])
```

D'autres commandes permettent de décrire une variable comme les commandes « range », « min », « max », etc.

Un test de Mann-Whitney-Wilcoxon est réalisé grâce à la commande « wilcox.test » :

```
> wilcox.test(crp[groupe == "ATB"], crp[groupe == "PCB"])
```

Les dernières lignes du résultat de cette commande sont :

```
data: crp[groupe == "ATB"] and crp[groupe == "PCB"]
W = 0, p-value = 0.0005828
alternative hypothesis: true location shift is not equal to 0
```

La *p-value* de ce test est donc $p = 0{,}0006$. On peut conclure que la moyenne du taux de CRP à 48 heures est plus basse pour les patients traités par une antibiothérapie.

Conclusion : quelles sont les limites ?

La recherche biomédicale est critiquée quant à sa qualité parfois discutable [4]. Lorsque le clinicien fait lui-même ses statistiques, il se place au centre de la recherche, augmente ses connaissances et contribue à améliorer par son expérience sa recherche future. Par contre, il est nécessaire de savoir rester dans le champ de ses compétences. Si les statistiques de base sont accessibles à l'ensemble des cliniciens, comprendre et réaliser des analyses statistiques

avancées nécessite une formation tout aussi avancée. Par analogie, lorsque les cliniciens ont compris l'utilité de l'échographie clinique, ils se sont progressivement formés pour qu'aujourd'hui cette pratique fasse partie des recommandations. Ces cliniciens ne remplacent pas pour autant les radiologues qui font maintenant moins d'échographies, mais plus d'examens avancés comme la TDM ou l'IRM. La collaboration entre cliniciens et radiologues s'en est trouvée renforcée ainsi que la prise en charge du patient. Si les urgentistes peuvent en effet faire une échographie de qualité, cela ne se résume pas à poser une sonde sur un ventre. Il faut connaître les caractéristiques de l'appareil, savoir ce que l'on peut ainsi mettre en évidence le cas échéant, connaître l'anatomie et interpréter les images. De même, la recherche clinique fait appel à un ensemble de processus qui ne peuvent se limiter à appliquer des recettes de cuisine statistique. Il faut intégrer la démarche, le design de l'étude, les hypothèses testées, la nature des variables retenues et les méthodes retenues. Le calcul n'est qu'une petite partie du processus. Loin de déposséder les biostatisticiens de leurs compétences, l'appropriation par le clinicien d'outils statistiques simples doit permettre un rapprochement entre les deux disciplines, une meilleure compréhension de la recherche clinique et une amélioration de celle-ci [5].

Liens d'intérêts : les auteurs déclarent ne pas avoir de lien d'intérêt.

Références

1. Dillmann C, Devilliers H (2009) Introduction à R. http://moulon.inra.fr/~mag/modelstat/data/tutoriel_R.pdf (dernier accès le 9 février 2016)
2. Paradis E (2005) R pour les débutants. https://cran.r-project.org/doc/contrib/Paradis-rdebuts_fr.pdf (dernier accès le 9 février 2016)
3. Barnier J (2013) Introduction à R. https://cran.r-project.org/doc/contrib/Barnier-intro_R.pdf (dernier accès le 9 février 2016)
4. Macleod MR, Michie S, Roberts I, et al (2014) Biomedical research: increasing value, reducing waste. Lancet 383:101–4
5. Ioannidis JP (2014) How to make more published research true. PLoS Med 11:e1001747