Pratique des Statistiques dans R II

Alboukadel Kassambara

Comparaison de Groupes: Variables Numériques

Résumer, Visualiser, Vérifier les hypothèses, Exécuter des tests, Interpréter, Rapporter

datanovia.com

Edition 1

Pratique des Statistiques dans R II - Comparaison de Groupes: Variables Numériques

Alboukadel KASSAMBARA

Published by Datanovia (`https://www.datanovia.com/en`), Alboukadel Kassambara

Contact: Alboukadel Kassambara <alboukadel.kassambara@gmail.com>

For general information contact Alboukadel Kassambara <alboukadel.kassambara@gmail.com>.

Contents

Préface

0.1 Ce que vous apprendrez

Ce livre de statistiques dans R fournit un solide guide pratique étape par étape des analyses statistiques pour **comparer les moyennes des groupes** en utilisant le logiciel R. De plus, nous avons développé un package R nommé **rstatix** (`https://rpkgs.datanovia.com/rstatix/`), qui fournit un système simple et intuitif, cohérent avec la philosophie de conception `tdyverse`, pour calculer les analyses statistiques les plus courantes, dont le test t, le test de Wilcoxon, ANOVA, Kruskal-Wallis, les analyses de corrélation, l'identification des valeurs extrêmes et autres.

Ce livre est conçu pour vous permettre de faire les tests statistiques dans R le plus rapidement possible. Le livre se concentre sur la mise en œuvre et la compréhension des méthodes, sans avoir à lutter à travers des pages de démonstrations mathématiques.

Vous serez guidé à travers les étapes de synthèse et de visualisation des données, de vérification des hypothèses et de réalisation de tests statistiques dans R, d'interprétation et de reporting des résultats.

0.2 Principales caractéristiques de ce livre

Bien qu'il existe plusieurs bons livres sur les statistiques et les sujets associés, nous estimons que beaucoup d'entre eux sont trop théoriques. Notre but était de rédiger un guide pratique de statistiques dans R avec visualisation, interprétation et rapport des résultats.

Les principales parties du livre sont les suivantes:

- *Tests statistiques et hypothèses* pour la comparaison des moyennes des groupes,
- *comparaison de deux moyennes*,
  - *t-test*,
  - *test de Wilcoxon*,
  - *Test des signes*,
- *comparaison de plusieurs moyennes*,
  - *ANOVA - Analyse des Variances* pour les mesures indépendantes
  - *ANOVA à mesures répétées*,
  - *ANOVA mixte*,
  - *ANCOVA et MANOVA*,
  - *Test de Kruskal-Wallis*
  - *test de Friedman*

Le livre présente les principes de base de ces tâches et donne de nombreux exemples dans R. Ce livre offre de solides conseils en statistiques pour les étudiants et les chercheurs.

Caractéristiques principales:

- Couvre les tests statistiques et les implémentations
- Les hypothèses clés sont présentées et vérifiées
- Chapitres courts et complets avec des exemples pratiques. Cela signifie que vous n'avez pas besoin de lire les différents chapitres dans l'ordre.

Dans chaque chapitre, nous présentons des sections pratiques de R dans lesquelles nous travaillons systématiquement à travers l'application des différentes méthodes discutées dans ce chapitre.

0.3 Comment ce livre est organisé ?

Ce livre contient 3 parties. Après une rapide introduction à R (Chapitre 1), **la partie I** présente quelques questions de recherche et les **tests statistiques** correspondants, ainsi que les **hypothèses** des tests. Bon nombre des méthodes statistiques, dont le test t et l'analyse de variance (ANOVA), supposent certaines caractéristiques des données, notamment **la normalité de la distribution des données** et **l'égalité des variances des groupes**. Ces hypothèses doivent être prises au sérieux pour tirer une interprétation et des conclusions fiables de la recherche. Dans la Partie I, vous apprendrez comment évaluer la normalité à l'aide du test de **Shapiro-Wilk** (chapitre 3) et comment comparer les variances dans R en utilisant le **test de Levene** et plus (chapitre 4).

Dans la **Partie II**, nous examinons comment comparer deux moyennes en utilisant le **test t** (méthode paramétrique, Chapitre 8) et le **test de Wilcoxon** (méthode non paramétrique, chapitre 9). Le contenu principal, inclut:

1. **Comparaison de la moyenne d'un échantillon à une valeur théorique standard**:
   - Test T pour échantillon Unique (paramétrique)
   - Test des Rangs Signés de Wilcoxon (non paramétrique)
2. **Comparaison des moyennes de deux groupes indépendants**:
   - Test T pour Echantillons Indépendants (paramétrique)
   - Test de la Somme des Rangs de Wilcoxon (non paramétrique)
3. **Comparaison des moyennes des échantillons appariés**:
   - Test T pour Echantillons Appariés (paramétrique)
   - Test des Rangs Signés de Wilcoxon sur des échantillons appariés (non paramétrique)

Dans cette partie, nous avons également décrit comment vérifier les hypothèses du test t et comment calculer la taille de l'effet du test t (le **d de Cohen**). Vous apprendrez également comment calculer la taille de l'effet du test de Wilcoxon. De plus, nous présentons le **test des signes** (Chapitre 10), une alternative au *test t sur échantillons appariés* et au *test des rangs signés de Wilcoxon*, dans le cas où la distribution des différences entre les valeurs des données appariées n'est ni normale (dans le test t) ni symétrique (dans le test de Wilcoxon).

La partie III décrit comment comparer plusieurs moyennes dans R en utilisant la méthode **ANOVA** (Analyse de variance) et les variantes (Chapitres 12 - 18).

Le chapitre 12 décrit comment calculer et interpréter les différents types d'ANOVA pour comparer les mesures indépendantes, notamment:

- **ANOVA à un facteur**, une extension du test t sur échantillons indépendants pour comparer les moyennes dans une situation où il y a plus de deux groupes.
- **ANOVA à deux facteurs** pour évaluer un effet d'interaction entre deux variables catégorielles indépendantes sur une variable-réponse continue.
- **ANOVA à trois facteurs** pour évaluer un effet d'interaction entre trois variables catégorielles indépendantes sur une variable de résultat continue.

Nous fournissons également un code R pour vérifier les hypothèses de l'ANOVA et effectuer des analyses post-hoc. De plus, nous présenterons le test de **Kruskal-Wallis** (Chapitre 17), qui est une alternative non paramétrique au test ANOVA à un facteur.

Le chapitre 13 présente l'**ANOVA à mesures répétées**, qui est utilisé pour analyser les données lorsque les mêmes sujets sont mesurés plus d'une fois. Vous apprendrez différents types d'ANOVA à mesures répétées, notamment:

- **ANOVA à un facteur sur mesures répétées** pour comparer les moyennes de trois niveaux ou plus d'une variable *intra-sujets*.
- **ANOVA à deux facteurs sur mesures répétées** utilisée pour évaluer simultanément l'effet de deux facteurs intra-sujet sur une variable-réponse continue.
- **ANOVA à trois facteurs sur mesures répétées** utilisées pour évaluer simultanément l'effet de trois facteurs intra-sujet sur une variable-réponse continue.

Vous apprendrez également comment calculer et interpréter le **test de Friedman** (Chapitre 18), qui est une alternative non paramétrique au test ANOVA à un facteur à mesures répétées.

Le chapitre 14 montre comment exécuter l'**ANOVA mixte**, qui est utilisé pour comparer les moyennes des groupes classés par au moins deux facteurs, l'un étant un facteur "*intra-sujets*" (mesures répétées) et l'autre un facteur "*inter-sujets*".

Les chapitres 15 et 16 décrivent, respectivement, quelques extensions avancées d'ANOVA, notamment:

- **ANCOVA** (analyse de la covariance), une extension de l'ANOVA à un facteur qui incorpore une covariable.
- **MANOVA** (multivariate analysis of variance ou analyse multivariée de la variance), une analyse de variance avec deux ou plusieurs variables-réponses continues.

0.4 Site du livre

Datanovia : `https://www.datanovia.com/en`

0.5 Exécution des codes R à partir du PDF

Pour un code R en une seule ligne, vous pouvez simplement copier le code du PDF vers la console R.

Pour un code R sur plusieurs lignes, une erreur est parfois générée lorsque vous copiez et collez directement le code R depuis le PDF vers la console R. Si cela se produit, une solution consiste à:

- Collez d'abord le code dans votre éditeur de code R ou dans votre éditeur de texte

- Copiez le code de votre éditeur texte/code dans la console R

De plus, si votre lecteur pdf dispose d'un outil de sélection qui vous permet de sélectionner du texte dans un rectangle, cela fonctionne mieux avec certains lecteurs.

0.6 Remerciements

Je remercie sincèrement tous les développeurs pour leurs efforts derrière les packages dont dépend ce livre, à savoir, bookdown et plus encore.

0.7 Colophon

Ce livre a été construit avec R 3.3.2 et les packages suivants :

```
##        name    version                       source
## 1  bookdown       0.16                         CRAN
## 2     broom      0.5.2                         CRAN
## 3  datarium  0.1.0.999                        local
## 4   emmeans      1.3.3                         CRAN
## 5    ggpubr      0.2.4                         CRAN
## 6   rstatix  0.3.0.999                        local
## 7 tidyverse 1.2.1.9000 Github:tidyverse/tidyverse
```

A propos de l'auteur

Alboukadel Kassambara est docteur en bioinformatique et biologie du cancer. Il travaille depuis de nombreuses années sur l'analyse et la visualisation de données génomiques (pour en savoir plus : `http://www.alboukadel.com/`).

Il a une expertise forte dans l'identification des signatures de biomarqueurs pronostiques et prédictifs par l'analyse intégrative des données génomiques et cliniques à grande échelle.

Il est l'auteur de:

1) l'outil bioinformatique **GenomicScape** (www.genomicscape.com), un outil Web facile à utiliser pour l'analyse et la visualisation des données d'expression de gènes.
2) les sites Web **Datanovia** (`https://www.datanovia.com/en/`) et **STHDA** (`http://www.sthda.com/english/`), qui contiennent de nombreux cours et **tutoriels** sur l'exploration de données et les statistiques d'aides à la décision.
3) plusieurs packages **R** populaires pour l'analyse de données multivariées, l'analyse de survie, la visualisation de matrices de corrélation et la visualisation basique des données (`https://rpkgs.datanovia.com/`).
4) de nombreux **livres** sur l'analyse des données, la visualisation et l'apprentissage automatique (`https://www.datanovia.com/en/shop/`)

Chapter 1

Introduction à R

R est un logiciel statistique gratuit et puissant pour **analyser** et **visualiser** des données. Si vous voulez apprendre facilement l'essentiel de la programmation R, visitez notre série de tutoriels disponibles sur STHDA : `http://www.sthda.com/english/wiki/r-basics-quick-and-easy`.

Dans ce chapitre, nous fournissons une très brève introduction à **R**, pour installer R/RStudio ainsi que pour importer vos données dans R et installer les packages requis.

1.1 Installer R et RStudio

1.1.1 Installation standard

R et RStudio peuvent être installés sur les plates-formes Windows, MAC OSX et Linux. RStudio est un environnement de développement intégré pour R qui facilite l'utilisation de R. Il comprend une console, un éditeur de code et des outils de traçage.

1. R peut être téléchargé et installé à partir de la page Web du Comprehensive R Archive Network (CRAN) (`http://cran.r-project.org/`)
2. Après avoir installé le logiciel R, installez également le logiciel RStudio disponible sur : `http://www.rstudio.com/products/RStudio/`.
3. Lancez RStudio et commencez à utiliser R à l'intérieur de Rstudio.

1.1.2 R en ligne

R est également accessible en ligne sans aucune installation. Vous trouverez un exemple à l'adresse `https://rdrr.io/snippets/`. Ce site comprend des milliers de modules complémentaires.

1.2 Installer et charger les package R requis

Un package R est un ensemble de fonctionnalités qui étend les capacités de base de R. Par exemple, pour utiliser le code R fourni dans ce livre, vous devez installer les packages R suivants:

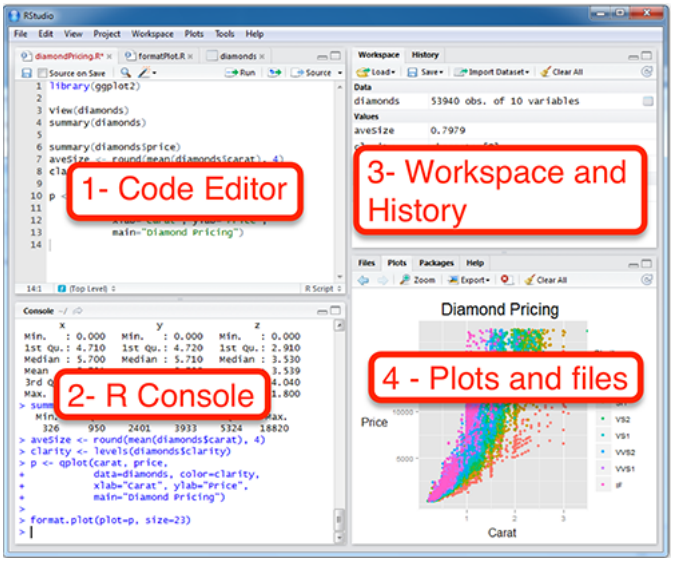

Figure 1.1: Interface Rstudio

- `tidyverse` : collection de packages R partageant la même philosophie de programmation. Ces packages comprennent:
  - `readr`: pour importer des données dans R
  - `dplyr`: pour la manipulation des données
  - `ggplot2`: pour la visualisation des données.
- `ggpubr` package, qui facilite, pour les débutants, la création de graphiques prêt-à-publication
- `rstatix` offre des fonctions R conviviales pour des analyses statistiques faciles à réaliser
- `datarium`: contient les jeux de données requis pour ce chapitre
- `emmeans`: effectuer des analyses post-hoc à la suite de tests ANOVA

1. **Installer le package tidyverse**. L'installation de tidyverse installera automatiquement readr, dplyr, ggplot2 et plus encore. Tapez le code suivant dans la console R:

```
install.packages("tidyverse")
```

2. **Installer les packages ggpubr, rstatix, datarium et emmeans**.

```
install.packages("ggpubr")
install.packages("rstatix")
install.packages("datarium")
install.packages("emmeans")
```

3. **Charger les packages requis**. Après l'installation, vous devez d'abord charger le package pour utiliser les fonctions qu'il contient. La fonction `library()` est utilisée pour cette tâche. Une autre fonction est `require()`. Par exemple, pour charger les packages tidyverse et ggpubr, tapez ceci:

```
library("tidyverse")
library("ggpubr")
```

Maintenant, nous pouvons utiliser des fonctions R, telles que *ggscatter*()[dans le package ggpubr] pour créer un nuage de points.

Si vous voulez de l'aide sur une fonction donnée, par exemple ggscatter(), tapez ceci dans la console R: `?ggscatter`.

1.3 Format des données

Vos données doivent être de format rectangulaire, où les colonnes sont des variables et les lignes des observations (individus ou échantillons).

- Les noms des colonnes doivent être compatibles avec les conventions d'appellation R. Évitez les colonnes avec des espaces vides et des caractères spéciaux. Bons noms de colonnes: `long_jump` or `long.jump`. Mauvais nom de colonne: `long jump`.

- Évitez de commencer les noms de colonnes par un nombre. Utilisez plutôt une lettre. Bons noms de colonnes: `sport_100m` or `x100m`. Mauvais nom de colonne: `100m`.
- Remplacer les valeurs manquantes par `NA` (pour non disponible)

Par exemple, vos données devraient ressembler à ceci:

```
  manufacturer model displ year cyl      trans drv
1         audi    a4   1.8 1999   4   auto(l5)   f
2         audi    a4   1.8 1999   4 manual(m5)   f
3         audi    a4   2.0 2008   4 manual(m6)   f
4         audi    a4   2.0 2008   4   auto(av)   f
```

Plus d'informations ici : [Meilleures pratiques pour la préparation des fichiers de données en vue de leur importation dans R] (`http://www.sthda.com/english/wiki/best-practices-in-preparing-data-files-for-importing-into-r`)

1.4 Importez vos données dans R

Tout d'abord, enregistrez vos données au format txt ou csv et importez-les comme suit (il vous sera demandé de choisir le fichier):

```
library("readr")

# Lecture des fichiers délimités par des tabulations (.txt tab)
my_data <- read_tsv(file.choose())

# Lit les fichiers délimités par des virgules (,) (.csv)
my_data <- read_csv(file.choose())

# Lit les fichiers delimités par des points-virgules (;) (.csv)
my_data <- read_csv2(file.choose())
```

Pour en savoir plus sur l'importation de données dans R, consultez le lien suivant : `http://www.sthda.com/english/wiki/importing-data-into-r`

1.5 Données de démonstration

R est livré avec plusieurs jeu de données de démonstration pour jouer avec les fonctions R. Les jeu de données de démo R les plus utilisés sont les suivants : **USArrests**, **iris** and **mtcars**. Pour charger un jeu de données de démonstration, utilisez la fonction **data**() comme suit. La fonction `head()` est utilisée pour inspecter les données.

```
data("iris")   # Chargement
head(iris, n = 3)  # Affichage des premières n = 3 lignes
```

```
##   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
## 1          5.1         3.5          1.4         0.2  setosa
## 2          4.9         3.0          1.4         0.2  setosa
## 3          4.7         3.2          1.3         0.2  setosa
```

Pour en savoir plus sur le jeu de données iris, tapez ceci:

```
?iris
```

Après avoir tapé le code R ci-dessus, vous verrez la description du jeu de données “iris” : ce jeu de données iris donne respectivement les mesures en centimètres des variables longueur et largeur des sépales, longueur et largeur des pétales, pour 50 fleurs de chacune des 3 espèces d’iris. Les espèces d’Iris sont setosa, versicolor et virginica.

1.6 Manipulation des données

Après avoir importé vos données dans R, vous pouvez facilement les manipuler en utilisant le package `dplyr` (Wickham et al., 2019), qui peut être installé avec le code R: `install.packages("dplyr")`.

Après avoir chargé dplyr, vous pouvez utiliser les fonctions R suivantes:

- `filter()`: Sélectionner des lignes (observations/échantillons) en fonction de leurs valeurs.
- `distinct()`: Supprimer les lignes en double.
- `arrange()`: Trier les lignes.
- `select()`: Sélectionner les colonnes (variables) par leur nom.
- `rename()`: Renommer les colonnes.
- `mutate()`: Ajouter/créer de nouvelles variables.
- `summarise()`: Calculer des statistiques descriptives (p. ex., calculer la moyenne ou la somme)
- `group_by()`: Analyser par sous-groupe de données.

Notez que le package dplyr permet d’utiliser l’opérateur d’enchaînement (%>%) pour combiner plusieurs opérations. Par exemple, x %>% f est équivalent à f(x). A l’aide de l’opérateur (%>%), la sortie de chaque opération est transmise à l’opération suivante. Ceci facilite la programmation de R.

Pour en savoir plus sur la manipulation des données, cliquez sur ce lien : `https://www.datanovia.com/en/courses/data-manipulation-in-r/`

1.7 Fermez votre session R/RStudio

Chaque fois que vous fermez R/RStudio, il vous sera demandé si vous souhaitez sauvegarder les données de votre session R. Si vous décidez de sauvegarder, les données seront disponibles dans les prochaines sessions R.

Part I

Tests Statistiques et Hypothèses

Chapter 2

Introduction

Dans ce chapitre, nous présenterons quelques questions de recherche et les **tests statistiques** correspondants, ainsi que les **hypothèses** des tests.

2.1 Questions de recherche et statistiques

Les questions de recherche les plus populaires sont les suivantes:

1. est-ce que **deux variables** (n = 2) sont **corrélées** (c.-à-d. associées)?
2. est-ce que **plusieurs variables** (n > 2) sont **corrélées**?
3. est-ce que **deux groupes** (n = 2) d'échantillons **différent** les uns des autres?
4. est-ce que **plusieurs groupes** (n >= 2) d'échantillons **différenciés** les uns des autres?
5. est-ce que la **variabilité** de deux échantillons ou plus diffère?

On peut répondre à chacune de ces questions à l'aide des tests statistiques suivants:

1. **Test de corrélation** entre deux variables
2. **Matrice de corrélation** entre plusieurs variables
3. **Comparaison des moyennes de deux groupes**:
   - **Test t de Student** (paramétrique)
   - **Test de Wilcoxon** (non paramétrique)
4. **Comparaison des moyennes de plus de deux groupes**
   - **Test ANOVA** (analyse de variance, paramétrique) : extension du test t pour comparer plus de deux groupes.
   - **Test de Kruskal-Wallis** (non paramétrique) : extension du test de Wilcoxon pour comparer plus de deux groupes
5. **Comparaison des variances**:
   - Comparaison des variances de deux groupes : **Test F** (paramétrique)
   - Comparaison des variances de plus de deux groupes : **Test de Bartlett** (paramétrique), **Test de Levene** (paramétrique) et **Test de Fligner-Killeen** (non-paramétrique)

2.2 Hypothèses des tests statistiques

Bon nombre des méthodes statistiques, notamment la corrélation, la régression, le test t et l'analyse de la variance, supposent certaines caractéristiques des données. En général, ils supposent que:

- les données suivent une **distribution normale**
- et les **variances** des groupes à comparer sont **homogènes** (égales).

Ces hypothèses doivent être prises au sérieux pour tirer une interprétation et des conclusions fiables de la recherche.

Ces tests - corrélation, t-test et ANOVA - sont appelés **tests paramétriques**, car leur validité dépend de la distribution des données.

Avant d'utiliser les tests paramétriques, certains tests préliminaires doivent être effectués pour s'assurer que les hypothèses de test sont respectées. Dans les cas où les hypothèses ne sont pas respectées, il est recommandé d'utiliser des tests **non paramétriques**.

2.3 Évaluer la normalité

1. Avec **des échantillons suffisamment grands** ($n > 30$), la violation de l'hypothèse de normalité ne devrait pas poser de problèmes majeurs (théorème central limite). Cela implique que nous pouvons ignorer la distribution des données et utiliser des tests paramétriques.
2. Cependant, par souci de logique, nous pouvons utiliser le **test statistique de Shapiro-Wilk** comparant la distribution de l'échantillon à une distribution normale afin de déterminer si les données montrent ou non un écart important par rapport à la disribution normale (Ghasemi and Zahediasl, 2012).

2.4 Évaluer l'égalité des variances

Le **test t standard de Student** (comparaison de deux échantillons indépendants) et le test ANOVA (comparaison de plusieurs échantillons) supposent également que les échantillons à comparer ont des variances égales.

Si les échantillons, à comparer, suivent une distribution normale, alors il est possible d'utiliser:

- **Test F** pour comparer les variances de deux échantillons
- **Test de Bartlett** ou **Test de Levene** pour comparer les variances de plusieurs échantillons.

2.5 Résumé

Ce chapitre présente les tests statistiques les plus couramment utilisés et leurs hypothèses.