Recherche Reproductible avec R

Introduction Générale

Ehouman Evans Abidjan RUsers, Yamoussoukro

06 August 2022

# Welcome

## Moi !

CIV, Yamoussoukro, Chercheur, Gestionnaire de Projets, Conférencier, Animateur Culturel, R User.

Twitter [@ehoumanevans](<https://twitter.com/ehoumanevans>)

Ma Biodata sur [@Netlify](<https://ehoumanevans.netlify.app>)

[@LinkedIn](<https://www.linkedin.com/in/ehoumanevans>)

## Recherche

## On s’ammuse un peu …

# Plan

* Reproductibility Research ?
* Idée maîtresse
* Strucuture d’un projet R (.Rproj)
* Le markdown
* Live coding : analyses statistiques
* Version Control
* Outils en ligne : Stockage et analyses de données

# Bibliographie

## Bibliographie 1

* Mike Croucher, Laura Graham, Tamora James, Anna Krystalli & François Michonneau (2017). Guide to Better Science: Reproductible code. *British Ecological Society*. Natural History Museum, and Durham University, UK, 41 p.
* Christophe Lalanne & Bruno Falissard (2018). Le langage R Markdown. <https://lms.fun-mooc.fr/c4x/UPSUD/42001S02/asset/RMarkdown.html>. 07/05/22.
* Barba, L. A. (2018). Terminologies for reproducible research. arXiv preprint arXiv:1802.03311.

## Bibliographie 2

* Christopher Gandrud (2020). Reproducible Research with R and RStudio (Third Edition).Chapman and Hall/CRC, London, UK 298 Pages.
* Jean-Marie DE KETELE & Xavier ROEGIERS (2015). Méthodologie du recueil d’informations : Formation des méthodes d’observation, de questionnaire, d’interview et d’étude de documents, 5ème édition, deBoeck Supérieur, 208 pages.

# Introduction

## Définition 1

Recherche reproductible : Les auteurs fournissent les données et le code nécessaire pour rejouer les analyses et recréer les résultats numériques.

Replication : Une étude arrive aux mêmes résultats scientifiques en collectant de nouvelles données (éventuellement avec des méthodes différentes) et en réalisant une nouvelle analyse pour répondre à la même question de recherche.

## Définition 2

Le principe de *répétabilité* ou *reproductibilité* :

* Production de connaissance validé par la communauté “scientifique” ;
* Reproduite par d’autres chercheurs.

## L’origine de l’analyse reproductible

Transparence

## *Reproductibility Research?* Recherche reproductible ?

# Mise en place

## Démarche

* Avoir un plan pour organiser, stocker et rendre les données et les codes d’analyse disponibles ;
* Transparence dans la production des résultats ;
* Revenir plustard et facilement réaliser ses analyses.

## Resultats : Figures ou Tableaux

## Resultats (Figure/Tableau) + Données

## Résultats + Données + Meta Données

## Resultats + Données + Meta Données

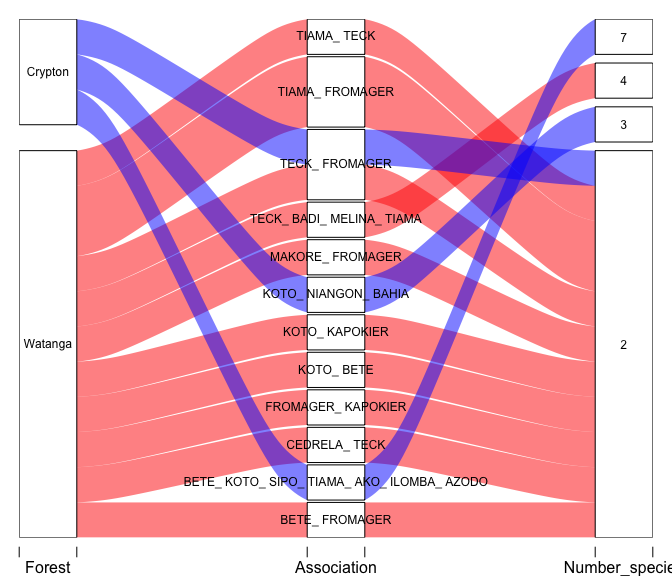
## Warning: package 'dplyr' was built under R version 4.1.2

## Rows: 14  
## Columns: 26  
## $ plot\_ID <chr> "Watanga\_Parcelle 1\_2004", "Watanga\_Parcelle 2\_20…  
## $ entity <chr> "Bois Ivoire", "Bois Ivoire", "Bois Ivoire", "Boi…  
## $ code\_forest <chr> "Watanga", "Watanga", "Watanga", "Watanga", "Wata…  
## $ interview\_date <dttm> 2021-10-20, 2021-10-20, 2021-10-20, 2021-10-20, …  
## $ company <chr> "Bois Ivoire", "Bois Ivoire", "Bois Ivoire", "Boi…  
## $ status\_forest <chr> "Forêt classée", "Forêt classée", "Forêt classée"…  
## $ name\_forest <chr> "Watanga", "Watanga", "Watanga", "Watanga", "Wata…  
## $ town <chr> "Yamoussoukro", "Yamoussoukro", "Yamoussoukro", "…  
## $ village <chr> "Attiégouakro", "Attiégouakro", "Attiégouakro", "…  
## $ name\_plot <chr> "Parcelle 1", "Parcelle 2", "Parcelle 3", "Parcel…  
## $ years\_install <chr> "2004", "2005", "2004", "2004", "2005", "2008", "…  
## $ age <dbl> 17, 16, 17, 17, 16, 13, 15, 15, 12, 12, 12, 22, 2…  
## $ area <dbl> 5.00, 20.00, 5.45, 4.83, 5.44, 5.16, 18.00, 6.84,…  
## $ species\_1 <chr> "KOTO", "TIAMA", "KOTO", "BETE", "TIAMA", "TECK",…  
## $ species\_2 <chr> "BETE", "FROMAGER", "KAPOKIER", "FROMAGER", "FROM…  
## $ species\_3 <chr> NA, NA, NA, NA, NA, NA, NA, NA, "MELINA", NA, NA,…  
## $ species\_4 <chr> NA, NA, NA, NA, NA, NA, NA, NA, "TIAMA", NA, NA, …  
## $ species\_5 <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, "AKO"…  
## $ species\_6 <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, "ILOM…  
## $ species\_7 <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, "AZOD…  
## $ no\_species <chr> "2", "2", "2", "2", "2", "2", "2", "2", "4", "2",…  
## $ association <chr> "KOTO\_ BETE", "TIAMA\_ FROMAGER", "KOTO\_ KAPOKIER"…  
## $ system\_culture <chr> "line/line", "line/line", "line/line", "line/line…  
## $ spacing\_species\_line <chr> "5/10", "5/10", "5/10", "5/10", "5/10", "3/3", "1…  
## $ ratio\_species <chr> "50/50", "50/50", "50/50", "50/50", "50/50", "50/…  
## $ density\_stem\_ha <chr> "200", "200", "200", "200", "200", "1100", "200",…

## Résultats + Données + Code

## Resultats + Données + Meta Data + Code

## Warning: package 'ggplot2' was built under R version 4.1.2



# Strucuture d’un projet R (.Rproj)

## Le problème

## le bon projet R

## Gestion d’un Dossier Projet R à partir de RStudio

## Fait par saven85

## Constituer un dossier projet manuellement

## Constituer un dossier projet automatiquement

# install.packages("prodigenr")  
  
library(prodigenr)  
  
# https://cran.r-project.org/web/packages/prodigenr/vignettes/prodigenr.html

# Le Markdown Rmarkdown

## Petite idée …

*Markdown* est un langage de balisage léger créé en 2004 par *John Gruber* avec l’aide d’ *Aaron Swartz*. Il a été créé dans le but d’offrir une syntaxe facile à lire et à écrire.

*R Markdown* offre une syntaxe simplifiée pour mettre en forme des documents contenant à la fois du texte, des instructions R et le résultat fourni par R lors de l’évaluation de ces instructions. En ce sens, il s’agit d’un outil permettant de produire des rapports d’analyse détaillés et commentés, plutôt que de simples scripts R incluant quelques commentaires.

## Exemple de Rapport

* Une étude sur la mise en place d’une clé dichotomique à partie d’un jeu de données d’inventaires floristique réalisée dans la forêt classée de la Téné (Oumé, Côte d’Ivoire).
* Recherche Reproductible avec R : Introduction Générale

## Jeu de données de HepatitisCdata.csv

## Rows: 615  
## Columns: 14  
## $ X <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18…  
## $ Category <chr> "0=Blood Donor", "0=Blood Donor", "0=Blood Donor", "0=Blood D…  
## $ Age <int> 32, 32, 32, 32, 32, 32, 32, 32, 32, 32, 32, 33, 33, 33, 33, 3…  
## $ Sex <chr> "m", "m", "m", "m", "m", "m", "m", "m", "m", "m", "m", "m", "…  
## $ ALB <dbl> 38.5, 38.5, 46.9, 43.2, 39.2, 41.6, 46.3, 42.2, 50.9, 42.4, 4…  
## $ ALP <dbl> 52.5, 70.3, 74.7, 52.0, 74.1, 43.3, 41.3, 41.9, 65.5, 86.3, 5…  
## $ ALT <dbl> 7.7, 18.0, 36.2, 30.6, 32.6, 18.5, 17.5, 35.8, 23.2, 20.3, 21…  
## $ AST <dbl> 22.1, 24.7, 52.6, 22.6, 24.8, 19.7, 17.8, 31.1, 21.2, 20.0, 2…  
## $ BIL <dbl> 7.5, 3.9, 6.1, 18.9, 9.6, 12.3, 8.5, 16.1, 6.9, 35.2, 17.2, 5…  
## $ CHE <dbl> 6.93, 11.17, 8.84, 7.33, 9.15, 9.92, 7.01, 5.82, 8.69, 5.46, …  
## $ CHOL <dbl> 3.23, 4.80, 5.20, 4.74, 4.32, 6.05, 4.79, 4.60, 4.10, 4.45, 3…  
## $ CREA <dbl> 106, 74, 86, 80, 76, 111, 70, 109, 83, 81, 78, 79, 78, 65, 63…  
## $ GGT <dbl> 12.1, 15.6, 33.2, 33.8, 29.9, 91.0, 16.9, 21.5, 13.7, 15.9, 2…  
## $ PROT <dbl> 69.0, 76.5, 79.3, 75.7, 68.7, 74.0, 74.5, 67.1, 71.3, 69.9, 7…

## [1] "0=Blood Donor" "0s=suspect Blood Donor" "1=Hepatitis"   
## [4] "2=Fibrosis" "3=Cirrhosis"

La moyenne d’âge varie t-elle d’une catégorie à une autre ?

## Jeu de données de HepatitisCdata.csv

### Test de normalité variable ALB

##   
## Shapiro-Wilk normality test  
##   
## data: HepatitisCdata$ALB  
## W = 0.92826, p-value < 2.2e-16

### Test de comparaison

Nous pouvons à présent procéder à la comparaison des groupes (FvsH).

Nous sommes partie de l’hypothèse H0 que la valeur en albumine est la même dans les deux catégorie (M, F). Puisque 1.417082310^{-7} <0.05, nous rejetons l’hypothèse H0 pour accepter l’hypothèse alternative. Conclusion, la valeur médiane d’albumine de l’homme est significativement différente à celle de la femme.

## Jeu de données de HepatitisCdata.csv : Médianes (ALB)

## # A tibble: 2 × 4  
## Sex count median mean  
## <chr> <int> <dbl> <dbl>  
## 1 f 238 40.1 40.6  
## 2 m 377 42.9 42.3

La valeur moyenne chez la femme est de 40.5592437 et chez l’homme ?

Que dire de la variable Categorie et le taux d’ALB

## Comparaison de médianes (ALB)

##   
## Kruskal-Wallis rank sum test  
##   
## data: ALB by Category  
## Kruskal-Wallis chi-squared = 70.072, df = 4, p-value = 2.192e-14

Nous pouvons donc procéder à une post-hoc de 1001 manières après avoir déterminer la médiane dans de chaque catégorie.

## Visulaisation des gorupes (ALB)

## # A tibble: 5 × 4  
## Category count median mean  
## <chr> <int> <dbl> <dbl>  
## 1 0=Blood Donor 533 42.2 42.2  
## 2 0s=suspect Blood Donor 7 21.6 24.4  
## 3 1=Hepatitis 24 43.5 43.8  
## 4 2=Fibrosis 21 41 41.8  
## 5 3=Cirrhosis 29 33 32.5

## Test post-hoc (ALB)

## [1] "1=Hepatitis" "0=Blood Donor" "2=Fibrosis"   
## [4] "3=Cirrhosis" "0s=suspect Blood Donor"

## Réalisation du test de Dunn

## Compact letter display

Affecter des lettres aux groupes

## Comparison Z P.unadj P.adj  
## 1 0=Blood Donor - 0s=suspect Blood Donor 3.5219561 4.283750e-04 8.567501e-04  
## 2 0=Blood Donor - 1=Hepatitis -1.8540331 6.373442e-02 9.104917e-02  
## 3 0s=suspect Blood Donor - 1=Hepatitis -4.0198279 5.824067e-05 1.456017e-04  
## 4 0=Blood Donor - 2=Fibrosis 0.5346925 5.928625e-01 6.587361e-01  
## 5 0s=suspect Blood Donor - 2=Fibrosis -2.7975049 5.149898e-03 8.583164e-03  
## 6 1=Hepatitis - 2=Fibrosis 1.6928492 9.048417e-02 1.131052e-01  
## 7 0=Blood Donor - 3=Cirrhosis 7.3294542 2.310919e-13 2.310919e-12  
## 8 0s=suspect Blood Donor - 3=Cirrhosis 0.1370035 8.910280e-01 8.910280e-01  
## 9 1=Hepatitis - 3=Cirrhosis 6.4665675 1.002541e-10 5.012703e-10  
## 10 2=Fibrosis - 3=Cirrhosis 4.4623858 8.105213e-06 2.701738e-05

## Group Letter MonoLetter  
## 1 BloodDonor a a   
## 2 ssuspectBloodDonor b b  
## 3 1Hepatitis a a   
## 4 2Fibrosis a a   
## 5 3Cirrhosis b b

## Jeu de données de HepatitisCdata.csv : test post-hoc (ALB) : Autres méthodes

## Warning in wilcox.test.default(xi, xj, paired = paired, ...): cannot compute  
## exact p-value with ties  
  
## Warning in wilcox.test.default(xi, xj, paired = paired, ...): cannot compute  
## exact p-value with ties  
  
## Warning in wilcox.test.default(xi, xj, paired = paired, ...): cannot compute  
## exact p-value with ties  
  
## Warning in wilcox.test.default(xi, xj, paired = paired, ...): cannot compute  
## exact p-value with ties  
  
## Warning in wilcox.test.default(xi, xj, paired = paired, ...): cannot compute  
## exact p-value with ties  
  
## Warning in wilcox.test.default(xi, xj, paired = paired, ...): cannot compute  
## exact p-value with ties

## Warning: package 'pgirmess' was built under R version 4.1.2

## Warning: multiple methods tables found for 'direction'

## Warning: multiple methods tables found for 'gridDistance'

## Tester une autre variable (BIL)

Nous sommes partie de l’hypothèse H0 que la valeur en albumine est la même dans les deux catégorie (M, F). Puisque 1.417082310^{-7} < 0.05, nous rejetons l’hypothèse H0 pour accepter l’hypothèse alternative. Conclusion, la valeur médiane d’albumine de l’homme est significativement différente à celle de la femme.

## Jeu de données Chernobyl NPP Accident

Description d’un jeu de données d’une analyse d’échantillon de sol prélévé à Chernobyle

## Jeu de données Chernobyl NPP Accident : Explorer le jeu de données

## Rows: 33  
## Columns: 32  
## $ N <chr> "G017603", "G017601", "G0176…  
## $ Site <chr> "Inner sampling area", "Inne…  
## $ GP\_Point <chr> "801", "802", "803", "804", …  
## $ Latitude <chr> "51.380494", "51.379942", "5…  
## $ Longitude <chr> "30.024562", "30.025445", "3…  
## $ Date\_Soil\_sampled <chr> "May\_June\_2014", "May\_June\_2…  
## $ Dose\_rate\_microSv\_per\_hour\_measurement\_1 <chr> "4.66", "6.48", "6.53", "5.3…  
## $ Dose\_rate\_microSv\_per\_hour\_measurement\_2 <chr> "4.85", "6.52", "6.48", "4.7…  
## $ Dose\_rate\_microSv\_per\_hour\_measurement\_3 <chr> "4.68", "6.65", "6.38", "4.9…  
## $ Dose\_rate\_microSv\_per\_hour\_measurement\_4 <chr> "n/a", "n/a", "n/a", "n/a", …  
## $ Dose\_rate\_microSv\_per\_hour\_measurement\_5 <chr> "n/a", "n/a", "n/a", "n/a", …  
## $ Dose\_rate\_microSv\_per\_hour\_measurement\_6 <chr> "n/a", "n/a", "n/a", "n/a", …  
## $ Dose\_rate\_microSv\_per\_hour\_measurement\_7 <chr> "n/a", "n/a", "n/a", "n/a", …  
## $ Dose\_rate\_microSv\_per\_hour\_measurement\_8 <chr> "n/a", "n/a", "n/a", "n/a", …  
## $ Dose\_rate\_microSv\_per\_hour\_measurement\_9 <chr> "n/a", "n/a", "n/a", "n/a", …  
## $ Dose\_rate\_microSv\_per\_hour\_measurement\_10 <chr> "n/a", "n/a", "n/a", "n/a", …  
## $ Dry\_mass\_of\_soil\_in\_grams <chr> "10", "10.05", "10", "10.01"…  
## $ Cs.137\_Soil\_Bq\_per\_sample\_DM <chr> "7.75E+01", "5.76E+02", "2.3…  
## $ error\_Cs.137\_Bq\_per\_Sample\_DM\_P.0.95 <chr> "1.02E+01", "7.62E+01", "3.1…  
## $ Am.241\_Soil\_Bq\_per\_sample\_DM <chr> "n/a", "n/a", "8.01E+00", "2…  
## $ error\_Am.241\_Bq\_per\_Sample\_DM\_P.0.95 <chr> "n/a", "n/a", "3.10E+00", "7…  
## $ Sr.90\_Soil\_Bq\_per\_sample\_DM <chr> "4.91E+01", "3.32E+02", "1.3…  
## $ error\_Sr.90\_Bq\_per\_Sample\_DM\_P.0.95 <chr> "1.04E+01", "7.05E+01", "2.9…  
## $ Pu.239\_240\_Soil\_Bq\_per\_sample\_DM <chr> "n/a", "n/a", "n/a", "n/a", …  
## $ error\_Pu.239\_240\_Bq\_per\_Sample\_DM\_P.0.95 <chr> "n/a", "n/a", "n/a", "n/a", …  
## $ Pu.238\_Soil\_Bq\_per\_sample\_DM <chr> "n/a", "n/a", "n/a", "n/a", …  
## $ error\_Pu.238\_Bq\_per\_Sample\_DM\_P.0.95 <chr> "n/a", "n/a", "n/a", "n/a", …  
## $ Cs.137\_Soil\_Bq\_kg\_DM <dbl> 7750, 57300, 23800, 51900, 2…  
## $ Sr.90\_Soil\_Bq\_kg\_DM <dbl> 4910, 33100, 13700, 13400, 6…  
## $ Am.241\_Soil\_Bq\_kg\_DM <chr> "n/a", "n/a", "8.01E+02", "2…  
## $ Pu.239\_240\_Soil\_Bq\_kg\_DM <chr> "n/a", "n/a", "n/a", "n/a", …  
## $ Pu.238\_Soil\_Bq\_kg\_DM <chr> "n/a", "n/a", "n/a", "n/a", …

## Comparaison des paramètres selon le site

### Préparation des données

### Tests statistiques

##   
## Shapiro-Wilk normality test  
##   
## data: Chernobyl\_npp\_soil$Dose\_rate\_microSv\_per\_hour\_measurement\_1  
## W = 0.89552, p-value = 0.009

##   
## Kruskal-Wallis rank sum test  
##   
## data: Dose\_rate\_microSv\_per\_hour\_measurement\_1 by Site  
## Kruskal-Wallis chi-squared = 5.8276, df = 1, p-value = 0.01578

Puisque la p-value est plus petite que la valeur seuil de 0.05, on peut en conclure qu’il y a une différence significative entre les groupes, les sites.

# Version Control

## Git avec RStudio

## Git avec Github Desktop

# Analyse en ligne : RStudio Cloud

[RStudio Cloud](https://rstudio.cloud)

# Stockage les données

[Kaggle](http://Kaggle.com)

[zenedo](http://zenedo.com)

[Havarddataverse](http://verse.harvard.edu)

# Merci !