

# Formation R initiation

22 et 23 janvier 2018



Martin CHEVALIER (Insee)

## Objectifs et pédagogie

### Objectifs de la formation

1. Acquérir des points de repères et des réflexes dans l'utilisation du logiciel R ;
2. Savoir travailler de façon autonome sur des données statistiques dans une perspective Insee ;
3. Maîtriser suffisamment les principes fondamentaux du logiciel pour être en mesure de se perfectionner par la suite.

### Principes pédagogiques

1. Pratique permanente et orientée métier ;
2. Autonomie dans l'apprentissage et progression à son rythme ;
3. Accompagnement personnalisé par le formateur.

### Supports (1) : pages web de la formation

Le support principal de la formation est un ensemble de pages web accessibles à l'adresse [r.slmc.fr](http://r.slmc.fr).

Ces pages web contiennent l'ensemble du contenu de la formation :

- ▶ présentation, explication et illustration des notions avec des exemples de code ;
- ▶ cas pratiques corrigés.

### Remarques

- ▶ Ces pages web peuvent être sauvegardées sous forme de fichiers .html et consultées hors connexion ;
- ▶ Les supports figurent dans le répertoire de la formation sur AUS : `Y:\Documentation\R\R_initiation\`.

### Supports (2) : livret et présentations

Le contenu de la formation est également disponible dans un **livret imprimé** :

- ▶ faciliter l'appropriation des notions ;
- ▶ réduire la fatigue visuelle ;
- ▶ constituer une référence à l'issue de la formation.

De **courtes présentations** ponctuent la formation :

- ▶ rythmer la progression ;
- ▶ présenter les objectifs des modules ;
- ▶ insister sur les notions les plus importantes.

**Remarque** Tous les supports de la formation ont été produits depuis R avec R Markdown (cf. R perfectionnement).

## Progression

La formation est articulée autour de **trois modules** :

1. **Prise en main du logiciel** : se repérer dans l'interface et savoir explorer des données ;
2. **Manipuler les éléments fondamentaux du langage** : acquérir une connaissance solide des briques élémentaires de R ;
3. **Travailler avec des données statistiques** : articuler les briques du module 2 pour la statistique appliquée.

## Articulation générale

- ▶ Les modules 1 et 3 sont **orientés métier** : travail sur des données Insee dans une perspective « chargé d'études ».
- ▶ Le module 2 est un **détour nécessaire** pour maîtriser les manipulations effectuées dans le module 3.

## Organisation pratique

### Travail en autonomie

- ▶ partie « cours » à partir du livret ;
- ▶ cas pratiques corrigés sur AUS ;
- ▶ **vous m'appellez en cas de difficulté.**

### Horaires

- ▶ proposition : 9h30-12h20 puis 13h40-16h30 ;
- ▶ pauses : 11h, 15h et quand vous sentez que c'est nécessaire !

### Conditions de travail La formation est **intensive** :

- ▶ prenez le temps de bien paramétrer votre espace de travail (éclairage, siège, etc.) et faites des pauses régulièrement ;
- ▶ n'hésitez pas à m'indiquer tout ce qui peut améliorer votre confort.

# Module 1 : Prise en main du logiciel

# Objectifs et organisation

## Objectifs

1. Acquérir des points de repère dans l'interface de R ;
2. Mener quelques traitements simples pour observer le fonctionnement du logiciel ;
3. Introduire des problématiques métier : travail sur des données, importation de fichiers de données SAS.

## Organisation

1. Un peu d'histoire et quelques grands principes
2. Découverte de l'interface
3. Charger et explorer des données
4. Importer des données à l'aide de *packages*



## Prise en main du logiciel

### Un peu d'histoire et quelques grands principes

Insister sur les **spécificités de R**, notamment par rapport à SAS :

- ▶ R est sensible à la casse ;
- ▶ dans R, les chemins doivent être indiqués avec des / et non des \.

Plus généralement, R s'apparente davantage à un **langage de programmation « classique »** (Python par exemple) :

*To understand computations in R, two slogans are helpful :*

- ▶ *Everything that exists is an object.*
- ▶ *Everything that happens is a function call.*

*John Chambers*

## Prise en main du logiciel

### Découverte de l'interface

Faire de **premières manipulations** dans les deux interfaces de R disponibles sur AUS :

- ▶ R « classique » : programme très dépouillé, essentiellement utilisé en mode « console » ;
- ▶ Rstudio : environnement de développement intégré qui facilite considérablement l'**écriture de scripts** (colorisation du code, auto-complétion, etc.).

Acquérir un **vocabulaire de base** :

- ▶ création d'objets simples et opérations arithmétiques ;
- ▶ affichage et manipulation des objets stockés en mémoire ;
- ▶ utilisation de l'aide intégrée dans le logiciel ;
- ▶ écriture d'une première fonction personnalisée.

# Charger et explorer des données

Savoir **utiliser des données** dans R :

- ▶ chargement d'un fichier de données `.RData` ;
- ▶ principales caractéristiques des données : nombre d'observations, affichage des premières lignes, etc.

Mener des **traitements simples** avec R :

- ▶ indicateurs statistiques usuels : moyenne d'une variable quantitative, distribution d'une variable qualitative ;
- ▶ ventilation des traitements selon les modalités d'une variable qualitative ;
- ▶ production de graphiques simples.

Prise en main du logiciel

## Importer des données à l'aide de *packages*

**Importer des données** qui ne sont pas en format R natif :

- ▶ fichiers plats : .txt, .csv, .d1m;
- ▶ fichiers SAS : .sas7bdat.

**Exporter des données** au format R natif.

Percevoir l'**importance des *packages*** dans R :

- ▶ installation ;
- ▶ chargement pour accéder à de nouvelles fonctions.

## Module 2 : Manipuler les éléments fondamentaux du langage

# Manipuler les éléments fondamentaux du langage

## Objectifs et organisation

### Objectifs

1. Introduire progressivement les briques élémentaires du langage de R ;
2. Connaître leurs propriétés et savoir les manipuler ;
3. Enrichir son vocabulaire de fonctions.

### Organisation

1. Manipuler les vecteurs
2. Manipuler les matrices
3. Manipuler les listes

# Manipuler les éléments fondamentaux du langage

## Manipuler les vecteurs

Les vecteurs sont les éléments fondamentaux du langage de R.  
Ils servent notamment :

- ▶ à coder l'information statistique : les variables d'une table sont des vecteurs ;
- ▶ à modifier le contenu d'une table : créer de nouvelles variables, sélectionner des variables et des observations ;
- ▶ à calculer des indicateurs statistiques : les *inputs* de la plupart des fonctions statistiques sont des vecteurs.

## Progression de la partie

1. Création de vecteurs et sélection d'éléments ;
2. Spécificités des différents types de vecteurs et manipulations ;
3. Compléments : valeurs spéciales, conversion de type

# Manipuler les éléments fondamentaux du langage

## Manipuler les matrices

Les matrices sont une généralisation directe des vecteurs en deux dimensions (ou plus, on parle alors de array).

Cette structure à deux dimensions les rapproche par certains égards des tableaux de données statistiques.

**Objectif de la partie** Création de matrices et sélection d'éléments.



# Manipuler les éléments fondamentaux du langage

## Manipuler les listes

Les listes sont des objets plus complexes que les vecteurs ou les matrices.

En particulier, elles peuvent contenir des **éléments de types différents** (numérique, caractère, logique, etc.), voire d'autres listes.

Cela en fait un type d'objet particulièrement souple pour **stocker et exploiter une information riche et structurée**.

**Exemples** Résultats d'un modèle de régression ou d'une méthode de classification.

## Progression de la partie

1. Création de listes et sélection d'éléments ;
2. Calculs sur les listes.

## Module 3 : Travailler avec des données statistiques

## Objectifs et organisation

### Objectifs

1. Revenir à des problématiques métiers courantes : sélection d'observations et de variables, tri d'une table, etc. ;
2. Mobiliser les fonctions abordées dans le module 2 ;
3. Présenter le calcul de statistiques descriptives avec R.

### Organisation

1. Manipuler les `data.frame`
2. Calculer des statistiques descriptives
3. Quelques liens pour aller plus loin

## Travailler avec des données statistiques

### Manipuler les `data.frame`

Le type `data.frame` est le type le plus souvent utilisé pour exploiter des données statistiques.

Il s'agit d'un **cas particulier de listes** qui **partage beaucoup de propriétés avec les matrices**.

### Progression

1. Création d'un `data.frame` et sélection d'éléments : sélection d'observations et de variables ;
2. Création ou modification de variables dans un `data.frame` ;
3. Modification de la structure d'un `data.frame` : tri, concaténation, fusions, etc. ;
4. Calculs sur un `data.frame` : application d'une fonction à toutes les variables, application d'une fonction par groupe.

# Travailler avec des données statistiques

## Calculer des statistiques descriptives

Plusieurs fonctions permettant d'effectuer des statistiques descriptives sont introduites dans les modules 1 et 2.

Cette partie présente ces fonctions de façon **plus systématique**, notamment autour de la question des **statistiques descriptives pondérées**.

### Progression

1. Statistiques descriptives sur variables quantitatives ;
2. Statistiques descriptives sur variables qualitatives ;
3. Création et paramétrisation de graphiques ;
4. Application à l'enquête Pisa 2012.

## Travailler avec des données statistiques

### Quelques liens pour aller plus loin

Les méthodes statistiques plus avancées sortent du cadre de cette formation au logiciel R.

Des liens sont néanmoins fournis pour approfondir deux aspects souvent utiles en pratique :

- ▶ **analyse de données multidimensionnelle** avec le [package FactoMineR](#) ;
- ▶ [modèles de régression](#) avec les fonctions `lm()` et `glm()`.

Un lien vers la [formation R perfectionnement](#) est également proposé :

- ▶ outils et méthodes pour se perfectionner en R ;
- ▶ traitements avancés sur des données dans R ;
- ▶ graphiques et *reporting* avec R.