# Atelier d'analyse de données

François Briatte
Automne 2014

Cet atelier est composé de huit séances de travaux dirigés qui complètent le cours magistral d'épistémologie et de méthodologie des sciences sociales enseigné par Julien Navarro.

Chaque séance servira à découvrir comment s'effectue le traitement de données quantitatives dans des disciplines de sciences sociales comme la démographie, la santé publique, la science politique ou la macroéconomie. L'accent sera mis sur la visualisation de données d'enquête.

> L'atelier vise à illustrer par la pratique les thèmes abordés en cours magistral : les exercices vous montreront à quoi ressemble concrètement une analyse empirique quantitative, et comment la reproduire dans un environnement professionnel de programmation statistique. Toutes les illustrations fournies en atelier sont basées sur des analyses réelles.

L'atelier requiert d'utiliser les logiciels R et RStudio, dont le fonctionnement sera expliqué et pratiqué en classe. Pour une introduction en langue française à ces logiciels, consultez le guide de Julien Barnier, *Introduction à R pour les sociologues (et assimilés)* (2014). L'installation de R et RStudio fera partie des premiers exercices de l'atelier.

Cet atelier ne requiert pas d'apprendre à programmer en langage R au-delà des manipulations les plus basiques. Si vous souhaitiez toutefois apprendre R en marge des séances, consultez sa documentation, les fiches Quick-R de Robert Kabacoff, les vidéos "Twotorials" d'Anthony Damico et celles de Google, "Intro to R", et le cours "Introduction to Data Analysis".

Enfin, de manière à vous rappeler les fondamentaux de votre cours de statistiques de première année, une sélection de chapitres de *Statistics in Plain English* (2010), de Timothy Urdan, vous sera envoyée via Google Drive. Cette lecture n'est pas obligatoire mais fortement recommandée si vous n'avez suivi qu'un seul enseignement de statistiques introductives par le passé.

Votre présence aux séances de l'atelier est obligatoire. Vous serez tenus de rédiger un certain nombre d'exercices, dont une dissertation d'entraînement à l'épreuve finale du cours magistral, et un examen de contrôle continu qui évaluera votre capacité à

- utiliser votre ordinateur comme un outil de travail,
- reproduire les étapes d'une analyse quantitative en sciences sociales,
- interpréter des traitements graphiques, et
- interpréter des statistiques univariées (distributions) ou bivariées (associations).

# Séances

## 1. Introduction

- Rappels sur l'épreuve de dissertation du cours magistral
- Installation des logiciels R et RStudio
- Travaux pratiques : visualisation des données Piketty et Saez

## 2. Méthodologie

- Dernier rappels sur l'épreuve de dissertation
- Méthodes qualitatives et quantitatives
- Travaux pratiques : visualisation des données Reinhart et Rogoff

## 3. Données

- Définition d'un jeu de données : format, observations, variables et échantillons
- Chapitre du manuel de référence : Urdan, ch. 1
- Travaux pratiques : premiers pas avec l'European Social Survey (enquête ESS 2010)

## 4. Distributions

- Tendances centrales (moyenne, médiane, mode) et dispersion (variance, écart-type)
- Chapitres du manuel de référence : Urdan, ch. 2 et 3
- Travaux pratiques : polarisation idéologique du Congrès américain (données Shor 2012)

## 5. Échantillonnage

- Distribution normale, erreur standard et intervalles de confiance
- Chapitres du manuel de référence : Urdan, ch. 4, 6 et 7
- Travaux pratiques : estimation de l'obésité aux États-Unis (enquête NHIS 2012)

## 6. Récapitulatif

- Rappels sur l'intégralité des séances de travaux dirigés
- Chapitres du manuel de référence : Urdan, ch. 1–7 (relecture)
- Travaux pratiques : entraînement à l'examen final (enquête ESS 2008)

## 7. Modélisation

- Association et corrélation linéaire : moindres carrés ordinaires
- Modèles linéaires et non linéaires : aperçu des modèles de régression
- Travaux pratiques : vote présidentiel aux États-Unis (modèles Hibbs et Bartels)

*Note* — Une séance additionnelle sera consacrée, en cours de semestre, à la dissertation et à l'examen d'une enquête auto-administrée.

# Instructions techniques

Les sections suivantes résument les opérations que vous aurez besoin d'effectuer sur votre ordinateur personnel (à amener en cours) pour suivre chaque séance des travaux dirigés.

## Installation de R et RStudio

R est un langage de programmation statistique, et RStudio une interface qui permet d'utiliser ce langage. Les instructions ci-dessous expliquent comment télécharger puis installer R et RStudio sur votre ordinateur, sous Mac OS X ou sous Windows.[1]

*Instructions pour Mac OS X*

- Téléchargez R à cette adresse : http://cran.rstudio.com/bin/macosx/R-3.1.1-snowleopard.pkg (~ 71 MB)
- Téléchargez RStudio à cette adresse : http://download1.rstudio.org/RStudio-0.98.1062.dmg (~ 38 MB)

Vous aurez besoin de Mac OS X 10.6+ et d'un processeur 64-bit pour installer ces logiciels. Si vous ne connaissez pas votre numéro de version de Mac OS X ou votre largeur de processeur, suivez cette procédure. Si votre système a été réglé pour n'installer que des applications du Mac App Store, désactivez cette "fonctionnalité" en suivant cette procédure.

*Instructions pour Windows*

- Téléchargez R à cette adresse : http://cran.rstudio.com/bin/windows/base/R-3.1.1-win.exe (~ 56 MB)
- Téléchargez RStudio à cette adresse : http://download1.rstudio.org/RStudio-0.98.1062.exe (~ 45 MB)

Vous aurez besoin de Windows XP, Vista, 7 ou 8 pour installer ces logiciels. Si vous ne connaissez pas votre version de Windows, suivez cette procédure. Vous aurez également besoin de pouvoir exécuter un programme en "mode administrateur". Si vous ne savez pas comment fonctionne cette procédure sur Windows 7 ou 8, lisez ces instructions.

*Procédure d'installation*

Commencez par installer R en double-cliquant sur le fichier d'installation "R-3.1.1" que vous avez téléchargé. Répondez "Oui / Yes / OK" à toutes les étapes de l'installation. Si l'installation crée des raccourcis sur votre Bureau, sous la forme d'icônes "R" mauves et plutôt moches, vous pouvez les supprimer, : vous n'en aurez pas besoin pendant le cours.

Suivez ensuite la même procédure pour RStudio – répondez "Oui / Yes / OK" à toutes les étapes de l'installation. Une fois cette procédure accomplie, vous avez terminé l'installation des deux logiciels. Les instructions pour lancer RStudio et pour l'épingler dans votre Dock (Mac OS X) ou dans votre barre des tâches (Windows) seront expliquées en cours.

[1] R et RStudio sont aussi disponibles pour la plupart des distributions récentes de Linux. Si vous savez utiliser Linux, vous n'avez probablement pas besoin d'instructions à suivre ; contactez-moi si ce n'est pas le cas.

## Installation de *packages*

R est un langage auquel il est possible de rajouter des extensions, qui s'installent sous la forme de "packages". Dans RStudio, la manière la plus simple d'installer un *package* est d'utiliser le menu "Tools", de sélectionner "Install Packages…", d'entrer le nom d'un *package*, puis de sélectionner "OK". Cette procédure requiert d'être connecté à Internet.

Les premiers *packages* qui seront installés au cours des séances des travaux dirigés incluent le package `ggplot2`, qui sert à produire des graphiques de haute qualité, et le package `survey`, qui sert à analyser des données d'enquête. L'installation sera effectuée en cours.

## Utilisation de RStudio

L'interface et l'utilisation de RStudio seront expliquées à chaque séance des travaux dirigés, et la procédure pour reproduire les analyses du cours sera également réexpliquée lors de chaque séance. Cette procédure peut se résumer de la manière suivante, en prenant l'exemple de l'exercice "Piketty et Saez" vu en première séance :

- téléchargez l'archive `piketty-saez.zip` à partir du dossier des exercices d'entraînement du cours
- repérez l'endroit où a été téléchargée l'archive `piketty-saez.zip` sur votre disque dur (en général, le dossier "Téléchargements")
- décompressez l'archive `piketty-saez.zip` si elle n'a pas été décompressée automatiquement lors du téléchargement
- ouvrez le dossier créé par la décompression, et vérifiez qu'il contient plusieurs fichiers, dont un fichier se terminant par l'extension `.r`
- déplacez le dossier `piketty-saez` sur votre Bureau, de manière à pouvoir le localiser facilement par la suite
- sélectionnez ce dossier comme dossier de travail dans RStudio, à partir du menu "Session > Set Working Directory > Choose Working Directory…"
- dans l'onglet "Files" de l'interface RStudio (en bas à droite), ouvrez le fichier R – que l'on appelle un "script" – intitulé `incomes.r` en cliquant dessus
- exécutez le script en sélectionnant toutes les lignes, puis en appuyant sur Ctrl-Entrée (Windows) ou Cmd-Entrée (Mac)
- le(s) graphique(s) de l'analyse s'affiche(nt) dans l'onglet "Plots" (en bas à droite) : utilisez le bouton "Zoom" pour le(s) voir en plein écran

Lors des travaux dirigés, cette procédure sera enseignée à travers les exercices d'entraînement listés dans le plan du cours : "Piketty et Saez", "Reinhart et Rogoff", "European Social Survey"… En fin de semestre, vous aurez besoin de pouvoir la mettre en oeuvre de manière autonome pour l'examen final des travaux dirigés.

> Rappelez-vous que l'objectif des travaux dirigés n'est pas de vous apprendre à programmer en langage R, mais de vous familiariser avec les manipulations fondamentales servant à reproduire une recherche empirique en sciences sociales, à partir de données quantitatives, dans un environnement de travail professionnel. Vous aurez besoin de votre ordinateur et de toute votre concentration à chaque séance.

## Examen final

L'examen final de travaux dirigés portera sur l'analyse empirique d'un jeu de données, issu de l'enquête comparative *European Social Survey* (ESS), également utilisée lors du semestre pour l'exercice `ess5-france`.

Pour cet examen, vous aurez besoin de :

- connaître la méthodologie, la philosophie et la terminologie des sciences sociales (objectifs pédagogiques du cours magistral)
- savoir reproduire une analyse empirique quantitative en utilisant R et RStudio (objectifs pédagogiques des travaux dirigés)
- savoir rédiger, en langue anglaise ou française, un rapport synthétique en citant vos sources (objectifs pédagogiques transversaux à la Licence)

L'examen se présentera sous la forme d'une archive ZIP à télécharger et à décompresser, dans laquelle vous trouverez un script R à exécuter pour reproduire une analyse empirique simple et visualiser ses graphiques. Il vous sera d'abord demandé de :

- décrire le jeu de données : référence bibliographique complète de la source, période et méthode de collecte des données, type d'échantillon et nombre d'observations[2]
- décrire la distribution d'une variable, dite "variable dépendante", qui sera au coeur de l'analyse : définition de la variable, de ce qu'elle mesure, et de sa distribution[3]
- décrire plusieurs autres variables, dites "indépendantes", afin de résumer la structure socio-démographique de l'échantillon

Il vous sera ensuite demandé de formuler des hypothèses explicatives sur les relations possibles entre la variable dépendante et les variables indépendantes de l'échantillon, en tenant compte du contexte (pays et période des données, biais de mesure).

De manière à mettre vos hypothèses à l'épreuve des données, vous aurez enfin à interpréter plusieurs croisements de variables, sur le modèle des exercices présentés en cours (lecture de tableaux croisés et de graphiques dans RStudio).

Cette dernière étape vous demandera de savoir observer une distribution normale, lire une erreur standard, et comment se construit un intervalle de confiance. Ces notions, comme les précédentes, seront abordées en cours, et sont détaillées dans le manuel de Timothy Urdan.[4]

Votre examen terminal prendra la forme d'un document-gabarit à remplir avec vos réponses, puis à me renvoyer avant une date-limite fixée en cours. L'orthographe et l'expression écrite feront partie des critères de notation, et vos citations devront être formatées au format Harvard.

---

[2] Vous aurez besoin, pour compléter cette étape, d'effectuer des recherches précises dans la documentation de l'enquête ESS, et de comprendre la terminologie utilisée dans le chapitre 1 du manuel de Timothy Urdan.

[3] Vous aurez besoin, pour compléter cette étape, de savoir interpréter les statistiques descriptives vues en cours : moyenne, médiane, mode, minimum et maximum (étendue), quantiles/quartiles, variance, écart-type. Ces statistiques sont expliquées en détail dans les chapitres 2 et 3 du manuel de Timothy Urdan.

[4] Voir les chapitres 4 (distribution normale), 6 (erreur standard) et 7 (intervalles de confiance).

## Liens supplémentaires

Sur l'épistémologie :

Daniel Little, Understanding Society (blog) :

- Catégorie "Epistemology"
- Catégorie "Disciplines"
- Catégorie "Methodology"

Sur la philosophie des sciences :

Stanford Encyclopedia of Philosophy :

- Entrée sur Karl Popper
- Entrée sur Thomas Kuhn
- Entrée sur Max Weber

Sur les statistiques (avec R) :

- Foundation for Open Access Statistics
- Daniel Navarro, *Learning Statistics with R* (livre)
- OpenIntro Statistics et OpenIntro Labs (cours)

Sur la visualisation :

- Alberto Cairo, *The Functional Art* (2012)
- Jeffrey Heer *et al.*, "A Tour through the Visualization Zoo" (2010)
- Edward Tufte, *The Visual Display of Quantitative Information* (2001)