Initiation à Stata

Evens SALIES, evens.salies@sciencespo.fr, evens.salies@univ-cotedazur.fr
Observatoire Français des Conjonctures Economiques (Sciences Po),

0. Introduction

Le cours « Initiation à Stata » appartient à l'UE1 « Methodological prerequisites » du Master Expertise Economique de l'UCA-UNSA. L'objet du cours est d'apprendre à utiliser le logiciel Stata, et quelques techniques de programmation en général, pour l'application de méthodes statistiques sur données socio-économiques.

Vous pourrez utiliser ces données, si vous le souhaitez, dans le cadre d'un projet dans un autre cours

Les cours ont lieu à l'ISEM. Les salles ne sont pas équipées de PC. Il y a 18h de cours réparties sur 6 séances : les **03/09** et **06/09** de 13h à 16h, et les **13/09** et **18/09** de 9h à 12h et de 13h à 16h.

Ajouter la salle ici

0.1. Enjeux de l'utilisation de Stata

Comme tout logiciel de statistique, Stata permet d'étudier les données rapidement et avec précision. Les graphiques qui sont difficiles à faire à la main s'obtiennent en quelques secondes (les jolis graphes prennent plus de temps)

[Montrer des ex. de graphiques en .pdf du dossier **Documents**]

```
merge_11_energy
mesri_2
pb_ue_graphics_1
pb_ue_graphics_cr4cr20
responsablequalite
```

Cette performance reflète celle de l'ordinateur, sans lequel d'ailleurs certaines méthodes n'existeraient probablement pas. Il y a cependant un effet pervers à utiliser un logiciel de statistique : logiciel de statistique = boîte noire !

Stata fait gagner du temps dans l'ère des données, et rend créatif …

- Une quantité croissante de données à étudier. Pour reprendre ce qui est dit dans la vidéo du cours Introduction à la Statistique avec R sur https://www.fun-mooc.fr (une nouvelle session commence le **9 septembre**), « il y a plus de données disponibles qu'on ne peut en analyser ». Sur le site du logiciel **Tableau**, on pouvait lire qu'en 2020 la production mondiale de données était 50 fois celle de 2011. Analyser des données est le business de beaucoup d'entreprises du secteur numérique, et existe depuis longtemps dans d'autres secteurs de l'économie.[1] Grâce à l'ordinateur, on peut faire ces analyses avec **Stata** ou autre logiciel (**Excel**, **Gauss**, **Python**, **R**, **RATS**, **SPSS**, **Tableau**, …).

- Des graphiques sophistiqués et des calculs statistiques compliquées. Comme à la main, mais plus vite et en plus joli.

  Les graphiques (brutes ou transformées) peuvent être portés d'une machine à l'autre (.png, .pdf, …) et visualisés en dehors de Stata (mais pas portables entre **Stata**, **R**, **Python**). Les graphiques sont « customisables » en peu de temps dans **Stata**.

  Certaines méthodes statistiques nécessitent des calculs qui prennent du temps à faire à la main, même avec peu d'observations. Par exemple, calculer la médiane ou une moyenne tronquée nécessite d'ordonner les valeurs de la variable concernée. En programmant, on peut simuler les propriétés d'une statistique, faire du Bootstrap, etc.

  Plutôt que de passer du temps à faire ces calculs à la main, il vaut mieux affecter ce temps à bien spécifier un modèle, à chercher les hypothèses pertinentes à tester.

[1] Administration, banque/assurance, service public en réseau, institut de statistique, secteur agricole.

- Susciter rapidement de nouvelles approches de vos données. Certaines manipulations rapides de vos données peuvent révéler des structures/patterns dans les graphiques et les tableaux. Quelles manipulations ?
  - Ordonner (c'est sur cette étape préalable, simple en théorie, mais fastidieuse à la main, que se fonde le tracé d'un histogramme)
  - Permuter
  - Transformer les variables comme prendre la transformation logarithmique pour linéariser les modèles multiplicatifs, prendre des ratios, etc.

[ouvrir « 1999_Wine.xls » et « statainitiation_0.do »]

fichier statainitiation_0_vin.csv

- Une interface *point-and-click*, qui révèle des commandes. Pour les néophytes de Stata, le logiciel a ses commandes accessibles *via* des menus.

L'avantage à avoir des commandes dans des menus est que vous pouvez tomber rapidement sur des représentations graphiques et des méthodes statistiques auxquelles vous n'auriez jamais pensé.

… mais il y a un coût d'entrée …

- Apprendre le langage Stata. Même si on n'a pas à connaître toutes les commandes, il y en a un paquet. **Stata Corp** a un document de deux pages de la quarantaine de commandes que toute utilisatrice devrait connaître, https://www.stata.com/manuals/u28.pdf. Un programme inclut très souvent au moins l'une de ces commandes.

- Faire un fichier plat à partir d'une base de données peut être un peu long.

Même pour un petit fichier de données que vous avez téléchargé d'un site comme celui d'Eurostat ou celui de l'OCDE, les données sont rarement formatées pour une analyse directe sur **Stata**. Par exemple, les variables sont en ligne, **Stata** les veut en colonnes. Il y a du texte inutile partout.

Personnellement, si je veux faire un graphique juste pour un pays et quelques années, je vais souvent plus vite avec une importation des données dans **Excel**, suivie de quelques manipulations (copier, coller, transposer, etc.).

Pour les données Eurostat, **R** a une librairie qui charges directement les variables mises en forme dans la feuille de données. Nous ferons un programme dans **Stata** (**section 1**) qui manipule une base d'Eurostat. Mais faudra d'abord y aller la chercher.

… et ne s'appuyer que sur les commandes a un effet pervers : Stata = boîte noire

Si vous ne passez que par le logiciel pour faire de la statistique, vous allez oublier des techniques. Il faut toujours essayer de comprendre ce que fait une commande **Stata** avant de l'exécuter ; comprendre avant de faire plutôt que faire avant de comprendre, quitte à ne connaître qu'un nombre limité de commandes (vos camarades connaissent les autres). **Stata** ne doit pas vous faire oublier les formules statistiques que vous avez vues en Licence.

Même au niveau des techniques de manipulation des données, il est utile de comprendre les opérations telles que le tri (tri à bulle), la transposition de tableaux (transposer une matrice), la fusion de bases de données venant de différents fichiers (unions, intersection d'ensembles), etc. Dans tous ces cas, vous devez toujours être capables de faire les manipulations à la main, sur un petit exemple.

0.2. Pourquoi Stata ? Pourquoi pas RATS, EViews, Limdep, SPSS, SAS, Gauss, … R voire Python ?

Nous n'éluderons pas la question des avantages et inconvénients de **Stata** relativement à d'autres logiciels de statistique.

Certains sont plutôt spécialisés *time series* (**RATS**, **EViews**), **Limdep** plutôt variables dépendantes limitées. **SPSS** est très bien pour les enquêtes. D'autres logiciels sont avantageux pour le calcul matriciel (**Gauss**). **SAS** est le logiciel historique à l'Insee et d'autres instituts de statistique ; ce n'est pas le plus vieux (1976), **SPSS** est de 1968. **Stata** date de 1985 !

R remonte à 1993 ; il succède à **S**. Je ne vous cache pas que **R** rentre dans toutes les institutions ! Il a plein de qualités, mais Stata aussi … pour qui veut bien passer du temps dessus.

En essayant de me projeter un peu, c'est le degré d'utilité d'un logiciel pour l'apprentissage automatique qui doit déterminer le choix. Dans ce cas, le couple **Python-Stata** ou **Python-R** !

La côte de Stata chez les Internautes

L'intérêt de **Stata** est largement expliqué sur le site https://www.stata.com/why-use-stata/. Stata est aussi puissant que d'autres logiciels pour la plupart des besoins en statistique (enseignement, recherche, études économiques, etc.). Il est livré avec toutes les librairies installées. Cependant, son prix, de 50€ à 500€ selon le type de licence, peut être un problème étant donné qu'il existe des logiciels gratuits ; l'arrivée de **R** et **Python** dans les départements d'économie s'inscrit dans le mouvement du logiciel libre. Quelques requêtes sur votre moteur de recherche préféré :

Tableau 0.1. Résultats de requêtes (fréquences) autour de Stata, R et Python, 2018-2024

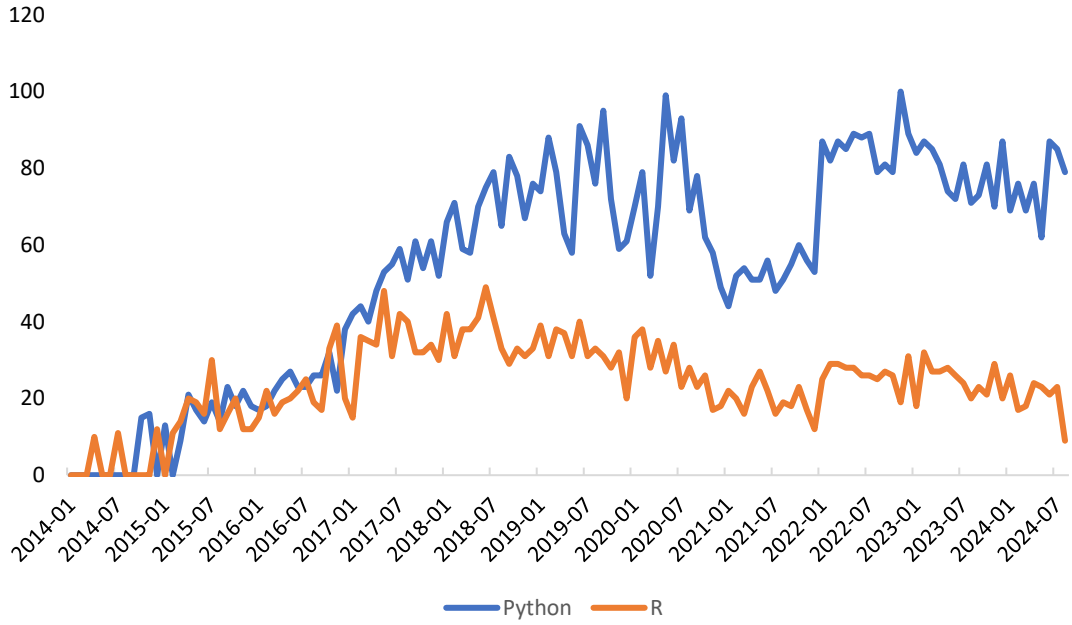
	2018	2019	2020	2021	…	2024
« programmation python »						215000
« logiciel r »	**185000**	↑277000	↓**179000**	**232000**		123000
« logiciel stata »	11200	↑207000	↓24900	29800		26800
« python programming »						**28800000**
« r software »	**959000**	**1370000**	**1540000**	**1960000**		↑13900000
« stata software »	14700	↑341000	321000	391000		386000
« statistique avec python »						3200
« statistique avec r »	**22100**	↑**56600**	**48000**	**30600**		↓**14700**
« statistique avec stata »	566	↑1380	↑4520	↓807		↑1480
« statistics in python »						166000
« statistics with r »	**410000**	**436000**	**584000**	**645000**		↓**246000**
« statistics with stata »	24400	35100	32300	35200		18000
« économétrie avec python »						4
« économétrie avec r »	7	↑758	↑**1900**	↑**3690**		↓408
« économétrie avec stata »	**566**	↑**933**	↓259	223		↑482
« econometrics in python »						12100
« econometrics with r »	**46100**	**48300**	**52500**	**70800**		↓**25600**
« econometrics with stata »	23900	↓4460	4750	6660		2990
« machine learning en python »						**28700**
« machine learning avec r »						7760
« machine learning dans stata »						
« machine learning in python »						**3890000**
« machine learning in r »						557000
« machine learning in stata »						10100

Note : « ↑ » indique une hausse importante et « ↓ » une baisse importante (d'au moins 50% environ, à la hausse ou à la baisse). Un nombre en **gras** signifie que le logiciel a au moins 50% env. de plus d'entrées que l'autre logiciel (avant 2024) et les deux autres logiciels (après 2024) ; par ex., en 2020, « logiciel r » a 618 % plus d'entrées que « logiciel Stata ».

Littéralement, **R** continue de grimper, l'écart se creuse avec **Stata** ; surtout en économétrie. Il se maintient en statistique. L'intérêt pour **Stata** en statistique semble stagner dans la communauté francophone. Le succès de **R** (voir les deux premières parties du tableau, surtout la requête « r software ») s'explique par le fait que beaucoup de disciplines scientifiques l'utilisent.[2] Encore plus pour Python, qui est un **langage de programmation généraliste** (*general purpose language*).

En 2024, nous avons ajouté dans le tableau des requêtes pour **Python**. Il est évident qu'en matière d'**apprentissage automatique** (*machine learning*), c'est ce langage qui domine. Et il est possible que **Python** rattrape **R** pour les méthodes statistiques en général. **Stata** se maintient partout ; le logiciel a une communauté de fidèles (votre prof !). Mais la gratuité est un critère important, qui fait que **R** et **Python** sont adoptés plus facilement que **Stata**.

Graphique 0.1. Fréquence relative des requêtes « machine learning in x », x = Python, R.



Source : Google Trend.

Pour ceux qui ~~seraient amenés à traduire des programmes~~ **~~R~~** ~~en~~ **~~Stata~~** ~~et vice versa, il existe un~~ document d'Oscar Torres-Reyna qui contient la traduction de commandes Stata en R. Un document équivalent de Daniel Sullivan existe pour le couple **Stata**-**Python**. **Stata** et **R** permettent d'intégrer du **Python** dans leurs codes.

Lien non-corrompu du document de Torres-Reyna's : https://www.princeton.edu/~otorres/RStata.pdf

[2]. Par soucis de comparaison, nous ajoutons qu'en 2018 **Excel** faisait mieux que **Stata** : « statistique avec excel » (15900) et « statistics with excel » (311000) ; en 2019, ces nombres sont 6480 (↓) et 857000 (↑). En 2021, 8760 et 288000 (↓). En 2021, **Stata** faisait mieux en statistique que **SAS** en français mais moins bien en anglais, et **SAS** se hissait au niveau de **Stata** en économétrie. Ce n'est plus vrai en 2023, année qui n'est pas reportée dans le tableau : « econometrics with sas » donnait 3880, alors que « econometrics with stata », 15200).

Stata est convivial

Fenêtres principales — Résultats, Commandes, Variables (aussi Data Editor), Propriétés (des données, voir **`describe`**), Historique des Commandes tapées, etc.

La fenêtre Propriétés se remplit quand on clique sur une variable de la fenêtre Variable.

Depuis **Stata 18**, on peut avoir plusieurs feuilles de données en mémoire (**`frames.do`**)

Data Editor a deux modes de visualisations : **`edit`** et **`browse`** :

`edit` : copier-coller des données .xlsx, .txt, …, dans la fenêtre Editor, décrire, créer des variables, attacher des labels, etc.

`browse` : *read-only*, visualiser les données en temps réel

Variables Manager — Organise les variables (voir menu Data du Data Editor)

Do-file Editor — Où il faut écrire ses *scripts*, dès lors que les programmes ont beaucoup de commandes. Les différents éléments des commandes sont colorés (*synthax highlighting*), ainsi que les commentaires, etc. Ces couleurs sont paramétrables. Les boucles sont fermables (*code folding*). Il existe une fonction Edit>Find>*Balance Braces* pour ne pas se perdre dans les boucles imbriquées (**`statainitiation_1_tablesgraphs.do`**).

Project Manager — Organise vos fichiers Stata **`.do`**, les place à côté des **`.dta`** et **`.gph`**. Un « projet » est dans un fichier (**`statainitiation.stpr`**)

Graph Editor — S'ouvre quand vous exécutez une commande graphique. Vous pouvez éditer le graphique, le sauver dans plusieurs formats, etc.

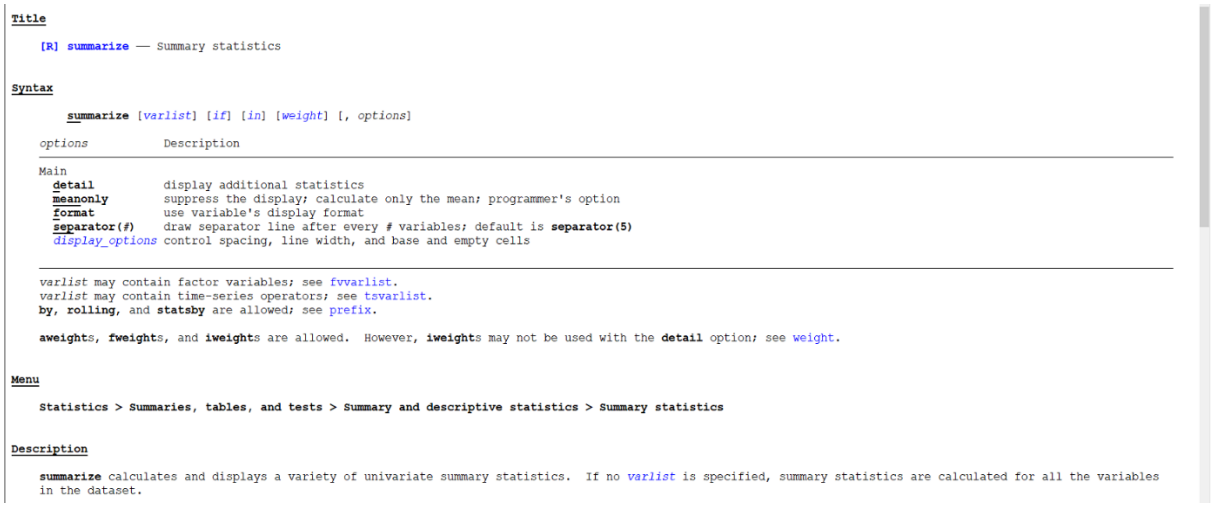
Stata 18 a un mode transparence dans les graphiques

Viewer — C'est la fenêtre d'aide, permettant de chercher des commandes, programmes et données des utilisateurs.

0.3. Ressources supplémentaires pour mieux comprendre le cours

Il y a plusieurs façons complémentaires d'apprendre ce cours : le suivre et consulter les ressources pédagogiques suivantes :

- **`help`** nom de la commande. Par exemple, **`help summarize`** affiche la fenêtre **`Viewer`** :



Title

[R] summarize — Summary statistics

Syntax

summarize [*varlist*] [*if*] [*in*] [*weight*] [, *options*]

options	Description
Main	
detail	display additional statistics
meanonly	suppress the display; calculate only the mean; programmer's option
format	use variable's display format
separator(#)	draw separator line after every # variables; default is **separator(5)**
display_options	control spacing, line width, and base and empty cells

varlist may contain factor variables; see fvvarlist.
varlist may contain time-series operators; see tsvarlist.
by, **rolling**, and **statsby** are allowed; see prefix.

aweights, **fweights**, and **iweights** are allowed. However, **iweights** may not be used with the **detail** option; see weight.

Menu

Statistics > Summaries, tables, and tests > Summary and descriptive statistics > Summary statistics

Description

summarize calculates and displays a variety of univariate summary statistics. If no *varlist* is specified, summary statistics are calculated for all the variables in the dataset.

Cliquer en haut à gauche sur « **`[R] summarize`** »,[3] permet d'accéder à la doc **Stata** officielle de la commande **`summarize`**. C'est la documentation qui accompagne votre logiciel **Stata** (la documentation change en partie d'une version à l'autre).

[3] La couleur des liens vers la documentation .pdf de Stata Corp est celle des liens hypertexte 0 0 255.

- La documentation Stata officielle : https://www.stata.com/features/documentation/. Les manuels généraux :

[GSW] **Getting Started with Stata** : installation, commandes d'importation de données pour s'essayer rapidement à **Stata**, etc.

[U] **User's Guide** : une description de **Stata** (langage, commandes les plus utilisées, pour estimer un paramètre, ou après avoir estimé un paramètre), des nouveautés dans **Stata 18**, des ressources disponibles en ligne (par exemple, la page **StataCorp** de vidéos sur YouTube https://www.youtube.com/user/statacorp), sur la gestion de la mémoire, les messages d'erreur, la programmation, la fusion de données, etc.

[R] **Reference** : commandes *built-in*, dans l'ordre alphabétique (par exemple, `drop`, `generate`, `oneway`, `regress`, `summarize`, `ttest`, etc.) mais aussi des mots clés **Stata** qui désignent une famille de commandes (`diagnosic plots`, `maximize`, etc.) ou des messages après l'exécution d'une commande (`error messages`, etc.)

[D] **Data Management** : descriptions, création de variables élaborées, fusion de fichiers (`merge`, `append`, etc.), transposition, etc.

[G] **Graphics** : représentation des données.

[F] **Functions** : fonctions pour transformer les variables (non-linéaires, matricielles, nombres aléatoires, manipulation de chaînes de caractères, etc.). Important !

Jusqu'à la version 12, **Stata** était livré avec les manuels généraux listés ci-dessus au format papier et gratuits. En revanche, la documentation spécialisée (« Time Series », « survey Analysis » …) était payante. Depuis la version 13, toute la documentation est gratuite. Les articles de la revue Stata Journal, où sont publiées de nouvelles commandes (fichiers `.ado`) créées par la communauté et approuvés par Stata corp, sont payants, mais pas les commandes disponibles en tapant `findit` dans la fenêtre de commande (par ex., `findit pscore`)

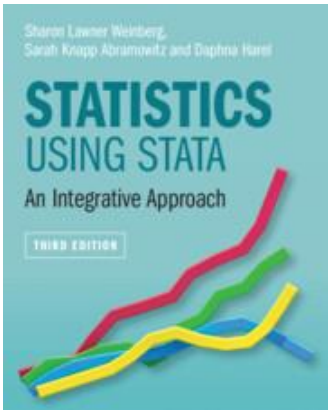
- « stata how to […] » sur Internet vous permettra de trouver des sites, des vidéos, des forums, la StataList (forum de Stata) : https://www.statalist.org/forums/ ; vous devrez créer un compte. Dans la partie Overview de [**U**], p. 13, il y a des liens …

- Me demander ou contacter mes collègues, échangez entre camarades.

- Petit document de commandes du cours

[~~http://evens-salies.com/Dico_Stata_old.txt~~]

- Ouvrage : vous pouvez vous procurer un ouvrage, récent si possible (pas en deçà de **Stata 13** ; avec une date de publication de 2016 ou au-delà). Celui-ci par exemple :



Je n'ai jamais réussi à faire un document définitif. Les autres sources de documentation sont plus pertinentes.

Il fait plus de 700 pages. La dernière version, de 2023, s'appuie sur **Stata 17** !

0.4. Données utilisées dans ce cours et pour le projet

Ce cours utilise les données issues de livres, d'articles, de ressources sur Internet, obtenues :

- directement auprès du producteur de données (par ex., enquête Baromètre du numérique du Credoc, etc.),
- indirectement (par ex., l'enquête R&D menée par les 27 pays de l'UE et disponible sur Eurostat),
- sur le terrain (l'ex. vinicole de cette section, l'enquête TICELEC menée à Biot).

Vous êtes libres d'utiliser ces données pour vos projets. Elles seront appelées par les différents programmes que nous écrirons dans le cours. Par exemple, pour les fichiers de données **Stata** ou **Excel** : http://www.evens-salies.com/[nom].dta ou http://www.evens-salies.com/[nom].xls.

0.5. Evaluation des étudiant·e·s en contrôle continu

Vous avez le choix entre :

- Un QCM d'1h à rentre le dernier jours de l'Eté afin de démarrer l'automne léger·ère !
- Une étude sur un base faite maison de rémunérations de cadres, en relation avec le diplôme, l'année d'obtention du diplôme, le genre, le pays (environ 140 observations).
- Faire un programme de quelques lignes qui importe dans Stata des données Eurostat avec la commande sdmxuse. Je l'ai installée mais je n'arrive pas à la faire marcher.

Vous êtes individuellement libres de choisir l'option qui vous intéresse le plus.

0.6. Plan du cours

Les sections et le plan du cours sont les suivants.

Section 0	Introduction
Section 1	Type de données, importation et transformation des données, création d'un fichier plat, présentation (tableaux, graphiques, statistiques descriptives les plus courantes), variables locales/globales/scalaires, missing values
Section 2	Lois de probabilités, méthode Monte Carlo, théorème central limite, fonctions de variables aléatoires
Section 3	Protocole, enquête, recensement, modèle d'échantillonnage, distribution d'échantillonnage, échantillon bootstrap, expérimentation contrôlée
Section 4	Type de variable, Spécification (univarié, multivarié, logit/probit, Poisson, MES, autorégressive, erreur de mesure, VO)
Section 5	Estimation (MCO, MV, MM, Bootstrap, Bayésienne), Apprentissage supervisé
Section 6	Test de spécification (ANOVA, permutation, Wald, exogénéité à la Hausman, non-linéarité à la Box-Cox), test de diagnostic dans le modèle linéaire (hétéroscédasticité, normalité)

La première séance commence par deux section : une **section 0** d'introduction générale sur les avantages et inconvénients de Stata (**section 0**) et la **section 1** sur une série d'applications de commandes Stata pour faire des graphiques, des tableaux et calculer des statistiques descriptives courantes. Quelques notions de 'commandes du programmeur' seront également introduites en complément de commandes Stata. La **section 2** est assez standard pour un cours de statistique puisqu'elle introduit des distributions usuelles que l'on rencontre en économie. Nous introduisons dans cette section la méthode Monte Carlo sur laquelle repose la production des nombres aléatoires, des fractiles et probabilités associées aux

distributions usuelles. Nous montrons également l'intérêt de la méthode Monte Carlo pour simuler la distribution de probabilité de fonctions de variables dont on ne connaît pas les moments théoriques.

L'organisation des sections 3 à 6 s'articule autour des thèmes suivants, dans l'ordre : **Protocole**, **Spécification/Estimation**, **Test**. La **section 3** introduit l'échantillonnage, vu selon l'approche modèle ou traditionnelle. Nous verrons ce qu'est un échantillon Bootstrap, et reviendrons sur un protocole expérimental introduit dans la section 1. Les **sections 4-5** portent sur des méthodes d'estimation, après avoir vu différents types de variables et le concept de spécification d'un modèle. Une justification est qu'une estimation dépend largement de la nature des variables (discrète, dichotomique, continue) et de la forme des relations entre elles (log-linéaire, autorégressive, système d'équations). Nous verrons rapidement un problème d'apprentissage supervisé en utilisant Python dans Stata. Enfin, la **section 6** porte sur les tests, étape ultime qui nécessite de connaître les propriétés, en échantillon fini ou 'infini', de l'estimateur (biais, consistance, etc.). Quelle que soit la sophistication du modèle de départ, le test porte généralement sur la valeur d'un ou plusieurs paramètres estimés du modèle.

0.7. Un mot sur le mémoire

Une fois achevé, un mémoire peut être présenté à la Société française d'évaluation (SFE). Le mémoire doit avoir pour objectif d'évaluer une politique publique. L'ouverture de la plateforme de candidature est fin juin 2025.

Si vous faites un mémoire avec moi, ce sera sur une question économique dans mes domaines de recherche. La partie théorie économique est plus béton quand c'est dans mes domaines de recherche :

- Question difficile dans un domaine que je connais bien :
    - Impact des aides directes et indirectes à la R&D en Europe : une méta-analyse. Le sujet est l'efficacité et l'efficience des politiques de soutien à la recherche et l'innovation dans l'UE-28, y compris les politiques nationales. J'ai beaucoup étudié le crédit d'impôt recherche (politique française). Ce serait bien de travailler sur d'autres pays, ainsi que les aides de l'UE. Il ne s'agit pas d'étudier ces politiques de fond en large, mais de faire une étude statistique des effets et de l'efficience des aides trouvés dans la littérature, sur la R&D des entreprises de recherche du secteur privé. Je n'ai jamais fait de méta-analyse, ce serait l'occasion.
- Questions faciles dans des domaines que je connais un peu, et que je n'ai pas communiquées à M. Jobert :
    - Relation innovation-emploi dans l'échantillon tronqué du Scoreboard
    - La coopération scientifique est-elle un facteur de paix ?