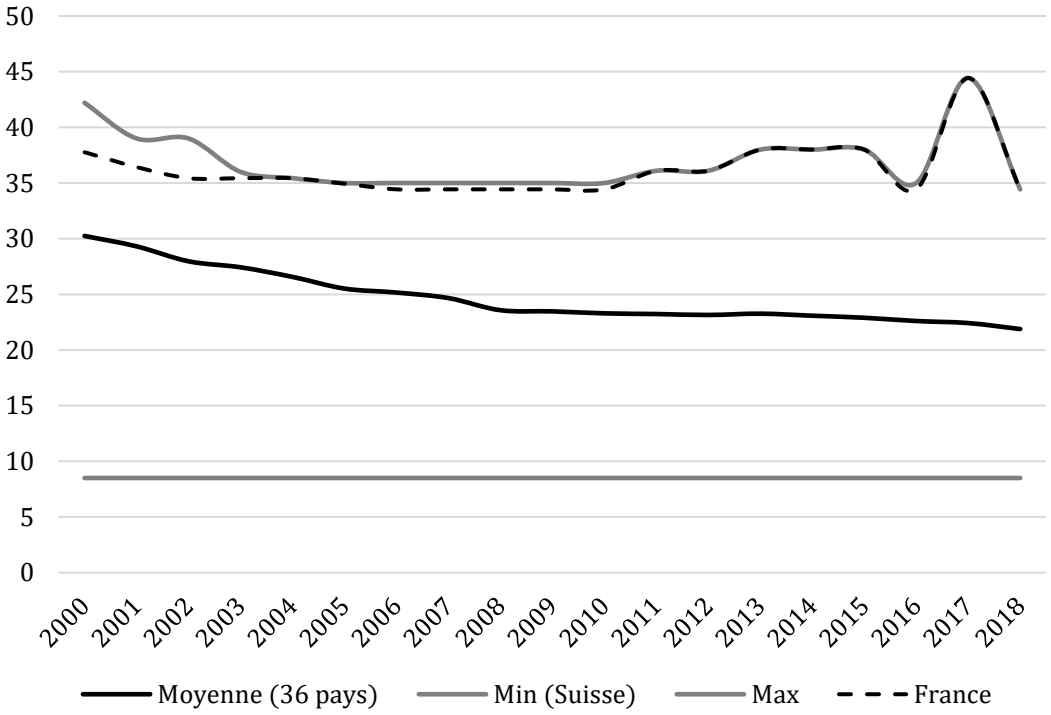
1. Représentation des données et statistiques descriptives

On commence par importer une base de taux d'imposition des sociétés (IS) pour les pays de l'OCDE entre 2000 et 2018 (36 pays en 2018). Ce court exemple montre des capacités de manipulation des données de Stata. Nous verrons ensuite d'autres bases, plus petites, dont nous tirerons des graphiques particuliers (camembert, boîte à moustache, etc.).

Sur le plan économique, cet exemple illustre la tendance des gouvernements à réformer l'IS dans un contexte de concurrence fiscale (**graphique ci-dessous**) ; le taux moyen décroit depuis une vingtaine d'années. Dans cette application, nous allons utiliser quelques commandes importantes : **`cd`**, **`drop`**, **`rename`**, **`local`**, etc., et **`reshape`** qui fait gagner beaucoup de temps.

Graphique. Taux statutaire d'imposition des bénéfices des entreprises (OCDE, 2000-2018)



Source : OECD tax database, url : https://stats.oecd.org/Index.aspx?QueryId=78166, dernier accès le 14/09/2022, et calculs de l'auteur.

Une étape préalable à la réalisation de tableaux, graphiques, estimations, est de créer un fichier plat à partir de données brutes. La base de données qu'on devrait avoir à la fin est **`statainitiation_1_corporatetaxrates.dta`**. Le code est sur Moodle (fichier « Initiation à Stata, section 1, transform OECD raw data to Stata panel (.do) »).

Le graphique ci-dessus a été fait avec Excel, pour une raison pratique : quand (i) on a une base de données brutes de petite taille comme ici, (ii) qu'on travaille sur une variable, deux-trois individus, (iii) que l'on souhaite produire un graphique vite-fait, et (iv) qu'on n'a pas à transposer les données, alors Excel (ou LibreOffice Calc, …) permet d'aller vite. Le **programme suivant** importe des données dans **`Stata`** et crée un fichier plat. Le programme sera redéployable sur les données maj et pour d'autres variables, ce qui pourrait être utile pour vos mémoires (nous en reparlerons dans le cours « **Données du web - micro** »).

[télécharger http://www.evens-salies.com/statainitiation_1_xlsxtodta.do]

1.1. Tableaux à une ou deux entrées

Stata a différentes commandes pour différents types de tableaux : **`table`**, **`tabulate`** et **`tabtsat`**, etc.

1.1.1. **`table`** (abrégé **`tab`**) (tableau flexible de statistiques élémentaires)
1.1.2. **`tabulate`** (tableau des fréquences, fréquences relatives et cumulées)
1.1.3. **`tabtsat`** (tableau compact de statistiques élémentaires)
1.1.4. **`recode`** (créer catégories, intervalles de classes)
1.1.5. **`missing`** (valeurs manquantes)

Ce qui distingue ces commandes c'est notamment :

- ce qu'elles font par défaut
- la prise en compte des valeurs manquantes (*missing values*)
- **`tabtsat`** ne produit que des tableaux à une entrée (*one-way table*) pour une variable

> Définition : une valeur manquante est la valeur non-renseignée d'une variable numérique. Elle a pour symbole « . » dans le **Data Editor**. Une chaîne de caractère vide « » n'a pas le statut de valeur manquante. C'est ce que nous allons voir ci-dessous avec un exemple construit à la main. Supposons une variable Y pour des unités d'observation dans U observées une ou plusieurs années numérotées dans A.

Tableau 1. Variables U, A et Y

unité d'observation	U	A (année)	Y
U_1	a	1	1
U_1	a	2	6
U_2	b	1	6
U_2	b	2	.
U_3		1	1
U_2	b	3	4
U_4	c	1	.

[rentrer le code « statainitiation_1_missing.do »]

Exécuter ce code rempli la fenêtre Data Editor qui s'ouvre avec un **`browse`**. La fenêtre Results comporte différents tableaux. On constate que :

- U_3 n'a pas de valeur manquante (au sens où nous avons défini ce terme), sa fréquence vaut 1. En fait, pour les fréquences des chaînes de caractères, **`table`** et **`tabulate`** comptabilisent les chaînes vides (pas besoin de l'option **`missing`**)
- Concernant les fréquences d'une variable numérique, **`table`** ignore celles manquantes, contrairement à **`tabulate`** (option , **`m`**). **`tabstat`** garde les missing mais leur attribue 0
- Pour le calcul d'autres statistiques (moyenne, etc.), **`table`** avec l'option **`missing`**, et **`tabstat`**, mettent un « . » dans la ligne d'une unité d'observation quand il n'y a pas (ou pas assez) d'observations pour le calcul. Les commandes **`table U, content(mean Y) row`** et **`tabstat Y, statistics(mean) by(U)`** renvoient les mêmes tableaux de résultats

1.2. Statistiques descriptives

[Données OCDE, comment les mettre à plat dans Stata ?]

1.2.1. **`list`**, **`count`** et applications de **`table`**, **`tabulate`** et **`tabstat`**
1.2.2. **`summarize`** (nb. d'observations, centralité, étendue, min, max, …)
1.2.3. **`summarize [...], detail`** (1er, 2e, 3e quartiles, etc.)

[On peut insérer les statistiques descriptives dans un tableau]

1.3. Graphiques à une entrée (*one-way*)

[G]

1.3.1. **`xtline [...],`** (graphique par coupe)
1.3.2. **`xtline [...], overlay`** (graphique qui superpose les coupes)
1.3.3. **`tsline`** (graphique d'une série chronologique)

[On change de données Rubin (1977)]

1.3.4. **`stem`** (graphique tige-feuille *stem and leaf*)
1.3.5. **`histogram`** (histogramme pour variable discrète)
1.3.6. **`box`** (boîte à moustache, *box plot*)

[On change de données Energy Information Administration (2005)]

1.3.7. **`pie`** (camembert, *pie*)

1.4. Transformer les données, combiner les variables, données

[Données OCDE] [On change de données Baromètre du Numérique]

Créer une seule variable de revenu à partir de plusieurs, dans un pseudo-panel

1.4.1. Lags et différences premières
1.4.2. Fusionner deux variables ayant un nombre de catégories différentes (**`label define`**, **`label values`**)

1.5. Local, global, scalar