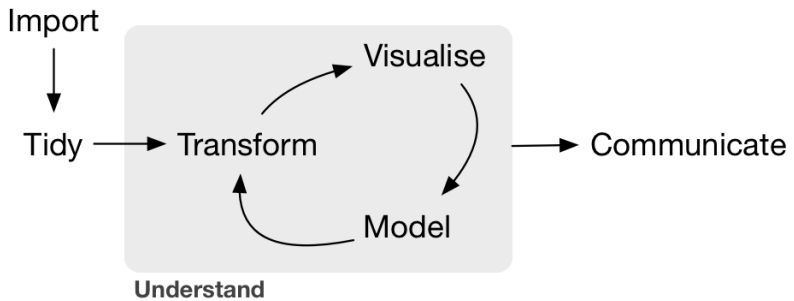
Analyse de données

Économétrie (ECON0212)

Malka Guillot
HEC Liège

Utiliser des données

- Les `données` sont centrales en économétrie.



- Selon un article du New York Times de 2014, "les data scientists [...] passent de ***50 % à 80 % de leur temps*** plongés dans ce travail plus prosaïque de collecte et de préparation de données numériques rebelles, avant de pouvoir les explorer"
- Aujourd'hui : les basiques de l'analyse de données
  - préparation, visualisation et statistiques descriptives

Au menu de cette séance

Stata 101

Nettoyer les données

Transformations avancées (aggrégations, combinaisons...)

Visualisation

Statistiques descriptives

Au menu de cette séance

Stata 101 *Là où les choses sérieuses commencent*

Nettoyer les données

Transformations avancées (aggrégations, combinaisons...)

Visualisation

Statistiques descriptives

Installation de `Stata`

05:00

1. Téléchargement depuis DOX

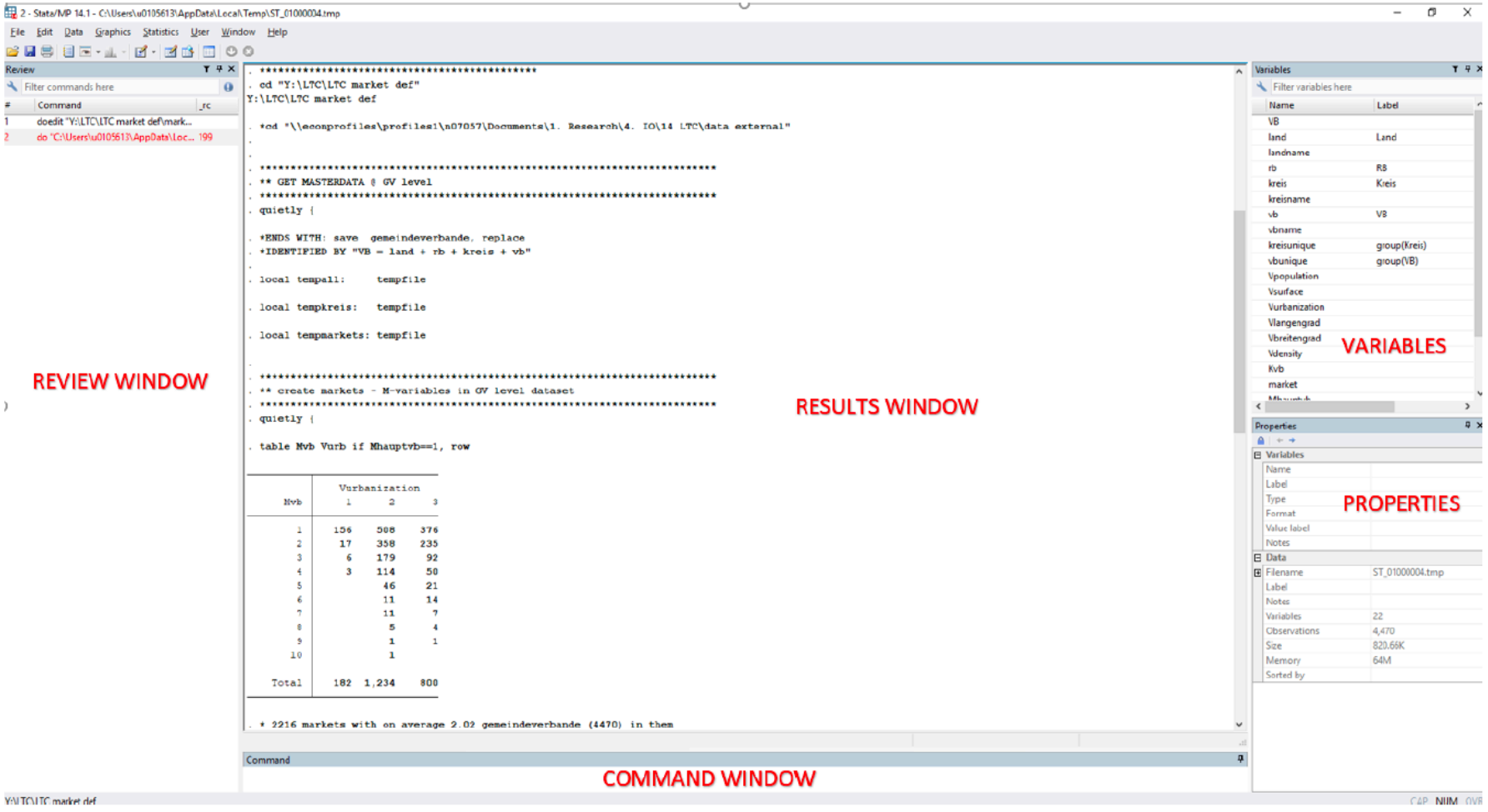
- Mot de passe = stata18hec
- Version Windows : https://dox.uliege.be/index.php/s/WVxyU6JxlW8LpxI
- Version Mac : https://dox.uliege.be/index.php/s/vRFUUNjcGbr0dFr
- Attention, il faut installer la version stata SE

2. Installation

- Lancer l'installateur et choisir la version Stata/SE18.
- Ensuite encoder les credentials ci-dessous :
  - Authorization : `73$v`
  - Code : `i5kw 5egd 3vL3 m5n0 fp$1 vfr3 4678 92uL awir 5`
  - Serial number : `401809321757`

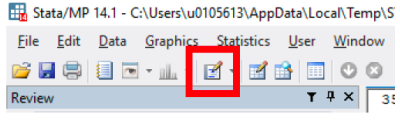
Cf. guide sur lola

Interface utilisateur de `Stata`



Glossaire

- `Stata`: le nom du logiciel
- Une ***commande*** : donnée fournie par l'utilisateur.rice que `Stata` comprend
  - Exemples: ouvrir une base de donnée, créer une variable, calculer une moyenne...
  - ⚠️ Une seule commande par ligne
- Un ***dofile*** : une liste de commandes, contenues dans un fichier texte (extension=`.do`)
  - Chaque commande doit être séparée par une nouvelle ligne
  - L'ordre des commandes dans le script correspond à leur ordre de lecture
  - Ouvrir un dofile :



- Pour lancer une commande dans un dofile :
  - Sélectionner la ligne à lancer et taper `Ctrl`+`D` (Windows) ou `Ctrl`+ `Maj`+`D` (mac)
  - Appuyer sur `Do`

Exercice 1

05:00

1. Créer un nouveau dofile (File → New → Do-file). Sauvegarder quelque part sous le nom `1-analyse-de-donnees`.

2. Ecrire le code suivant dans le dofile et compiler (`Ctrl`+`D` or `Ctrl`+`Maj`+`D`). (Vous pouvez surligner le code à compliler, ou compiler sur tout le dofile)

```
set obs 10 /* Génère 10 observations */
gen x = 1 /* crée une variable nommée x et égale à 1 */
```

Bravo ! Vous avez créé votre première variable !

1. Créer une nouvelle variable `y` égale à x^2.

Trouver de l'aide

- Directement dans `Stata` :

```
help command
```

  - Limite : il faut connaître le nom de la `commande`
  - Par exemple: `help regress`

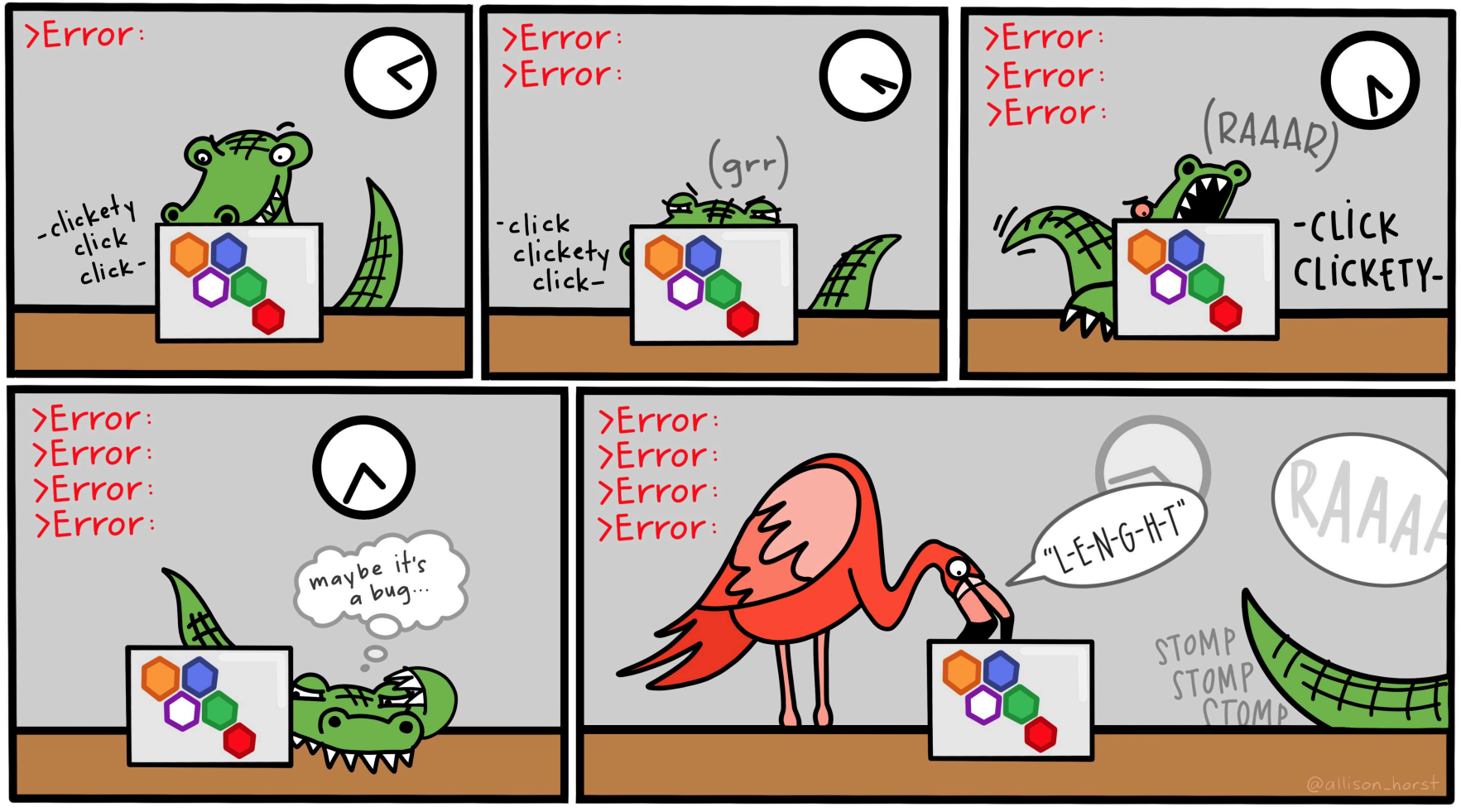
- Sur internet :

  - Il vaut mieux faire une recherche en anglais pour avoir plus de réponses
  - Parfois plus efficace
  - Renvoie souvent sur la documentation officielle

- ChatGPT :

  - Potentiellement encore plus efficace !
    - Surtout utile quand on a déjà une idée de ce qu'on fait
  - Example d'utilisation

Collaborer!


>Error:
-clickety click click-
>Error:
>Error:
(grr)
-click clickety click-
>Error:
>Error:
>Error:
(RAAAR)
-CLICK CLICKETY-
>Error:
>Error:
>Error:
>Error:
maybe it's a bug...
>Error:
>Error:
>Error:
>Error:
"L-E-N-G-H-T"
RAAAA
STOMP STOMP STOMP
@allison_horst

Quand ça tourne mal

Du rouge ? C'est une erreur ! Que faire ?

1. La lire $\rightarrow$ souvent, la source de l'erreur est expliquée
2. La rechercher sur internet car $P(\text{Quelqu'un a déjà eu cette erreur}) > 99\%)$

Importer des données

`stata` peut importer des données depuis différents formats :

- Format stata : extension `.dta`

```
use dataset_name.dta, replace
```

- Format Excel : extension `.xls` ou `.xlsx`

```
import excel using dataset_name.xlsx, firstrow replace
```

  - Attention au format du fichier excel: une ligne est une observation, une colonne une variable

- Format csv : extension `.csv`

```
import delimited using dataset_name.txt, delimiters(",") replace
```

  - Chaque élément est séparé de l'ensemble des données par un `delimiter` qui peut être : une tabulation, une virgule, un point virgule...

Inspecter des données

Une fois l'ensemble de données chargé, vous pouvez commencer à l'explorer :

- `browse` : Examinez vos données dans la fenêtre d'exploration.
- `edit` : Produit une liste de toutes les variables, leur type de données et leur étiquette
- `sum nom_de_la_variable` : Donne le nombre d'observations, moyenne, min et max des variables spécifiées après `sum`
- `sum nom_de_la_variable, detail` : Donne un résumé plus détaillé de la variable spécifiée
- `tab nom_de_la_variable` : produit un tableau de fréquences qui donne le nombre d'occurrences pour chaque valeur de la variable

Exercice 2

05:00

1. Importez la base `gapminder` en format dta. Cette base contient les données d'espérance de vie et de PIB par habitant de l'introduction.

2. Quelles variables contiennent les données? Vous pouvez utiliser `codebook` ou `describe`.

3. Inspectez visuellement les données. Plusieurs solutions pour ouvrir la base de donnée grâce à `browse`

    1. Naviguer dans le menu: `data, data Editor`
    2. Bouton `browse`
    3. Commande : `browse [varlist] [if]` dans la fenêtre de commande ou un dofile

    A quoi correspond la variable `pop` ?

Types de variables

- Numérique (`float`, `int`) vs. texte (`string`)
  - Exemple: pays ou nom $\rightarrow$ `string`; age $\rightarrow$ `int`
- Catégoriel:
  - Exemple: genre,
  - Une variable numérique avec un label pour chaque valeur
    - 1="female; 2="male"

Modifier le types de la variable

- texte $\rightarrow$ numérique

```
destring nom_de_la_variable, replace
```

- numérique $\rightarrow$ texte

```
tostring nom_de_la_variable, replace
```

- texte $\rightarrow$ catégoriel

```
encode nom_de_la_variable, gen(variable_encodee)
```

Modifier des données existantes (variables)

- Supprimer une variable (ou une liste de variables)

```
drop variable_1 variable_2
```

- Sélectionner une variable (ou une liste de variables) [ce qui supprimer les autres]

```
keep variable_1 variable_2
```

- Renommer une variable :

```
rename ancien_nom nouveau_nom
```

- Créer un label :

```
label var nom_de_la variable "Label de la variable"
```

Créer des variables

- On génère une nouvelle variable en lui donnant un nom et en définissant ses valeurs dans une `EXPRESSION`

```
gen nouvelle_variable = EXPRESSION
```

- Une `EXPRESSION` peut être :
  - Mathématique, par exemple :

```
gen variable1 = 200
gen variable2 = variable1*2
```

  - Du texte (`string` en `stata`), par exemple :

```
gen variable1 = "coucou"
```

- Une fonction de variables existantes:

```
gen variable3 = (variable1 + variable2)/2
```

Modifier la valeur d'une variable

- On peut remplacer la valeur d'une variable (très similaire à `gen`)
- Une fonction de variables existantes:

```
replace une_variable_qui_existe = EXPRESSION
```

- Remplacement conditionnel :

```
replace age_category = "old" if age > 70
```

Expressions logiques

- Expressions qui peuvent être utilisées pour préciser la condition :
  - `==` : égal à
  - `!=` : non égal à
  - `<=` : inférieur ou égal à
  - `>=` : supérieur ou égal à
  - `<` : strictement inférieur ou égal à
  - `>` : strictement supérieur ou égal à
- Pour combiner ces conditions, on utilise `&` ou `|` :
  - `&` = *et*: si les 2 conditions doivent être vérifiées toutes les 2

```
gen age_category = "old" if age > 70 & age < 110
```

  - `|` = *ou*: si une des conditions seulement doit être vérifiée

```
gen etudiant_hec = "bac" if filiere=="IG" | filiere=="SEG"
```

Sélection d'observations

- Garder ou supprimer les observations sous certaines conditions

```
keep if CONDITION
drop if CONDITION
```

- Par exemple:

```
/* Sélectionne les pays du continent américain */
keep if continent=="Americas"

/* Supprime les pays américains sauf l'Argentine */
drop if continent=="Americas" & country!="Argentina"
```

⚠ Comme c'est le cas avec de nombreux logiciels, l'expression logique s'écrit `==`

Exercice 3

07:00

1. Combien d'observations y a-t-il dans la base de données ouverte précédement (`gapminder`)?
2. Combien de variables ? De quel type ?
3. Ne gardez que les observations correspondant au continent `Asia`.
4. Chargez à nouveau la base, puis créez la variable `gdppercap` égale au PIB par habitant grâce au code suivant :

```
gen gdppercap = gdp / population
```

Félicitation, vous avez créé votre première variable ! Utilisez la commande `browse` pour la voir dans la base.

Syntaxe de stata

La plupart des commandes ont la syntaxe suivante :

```
command [varlist] [if] [, options]
```

où `[..]` correspond à des options. Exemples :

- `summarize` ou juste `sum`
  - Statistiques descriptives pour toutes les variables
- `summarize gdp`
  - Statistiques descriptives pour le `gdp` uniquement
- `summarize gdp if year == 1960`
  - Statistiques descriptives pour le `gdp` en 1960

Au menu de cette séance

Stata 101

Nettoyer les données

Transformations avancées (aggrégations, combinaisons...)

Visualisation

Statistiques descriptives

Valeurs manquantes

- Quand une valeur est maquante, elle est indiquée
  - `.` pour un variable numérique
  - `""` pour un variable string
- `Stata` propage les manquantes au fur et à mesure:

```
display . + 10
```

- `Stata` assigne la valeur $+\infty$ à une missing :

```
display . > 5
```

- Attention à la propagation des valeurs manquantes:

	country	year	population	gdp	gdppercap
1	Albania	1960	1636054	.	.
2	Algeria	1960	11124892	1.383e+10	1.54e+17
3	Angola	1960	5270844	.	.
4	Antigua and Barbuda	1960	54681	.	.
5	Argentina	1960	20619075	1.083e+11	2.23e+18

Exercice 4: Nettoyage des données

07:00

On utilise la base `gapminder` précédement importée.

- 1. Quels années ont des valeurs manquantes pour le PIB ?

1. Quel pays a l'éspérance de vie la plus élevée pour une population de plus d'un milliard en 2007 ?

2. Quel est la moyenne de l'espérance de vie par continent et par an ? Vous pouvez utiliser la fonction `collapse`

Au menu de cette séance

Stata 101

Nettoyer les données

Transformations avancées (aggrégations, combinaisons...)

- **Aggréger des données:** `collapse`
- **Combiner des données:** `merge` & `append`
- **Reformater des données:** `reshape`

Visualisation

Statistiques descriptives

Aggrégation des données `collapse`

La commande collapse transforme la base de données en mémoire en statistiques essentielles sur celle-ci (`sum`=somme ; `mean`=moyenne ; `sd`=ecart-type; `median`=médiane).

country	year	lifeexp
A	1	70
A	2	71
B	1	80
B	2	82
B	3	85
C	1	79
C	2	82
C	3	81

year	lifeexp
1	76,33
2	78,33
3	83,00

collapse (mean) lifeexp, by(year)

- Attention, `collapse` remplace les données en mémoire

Combiner des bases de données : append

economy2004.dta

country	GDP per Capita	year
ARG	12,468	2004
FRA	27,913	2004
GER	28,889	2004
ITA	28,172	2004
USA	39,498	2004

economy2005.dta

country	GDP per Capita	year
ARG	13,153	2005
FRA	29,203	2005
GER	30,15	2005
ITA	29,414	2005
USA	41,557	2005

Appended

country	GDP per Capita	year
ARG	12,468	2004
FRA	27,913	2004
GER	28,889	2004
ITA	28,172	2004
USA	39,498	2004
ARG	13,153	2005
FRA	29,203	2005
GER	30,15	2005
ITA	29,414	2005
USA	41,557	2005

```
use economy2004.dta, replace
append using economy2005.dta
```

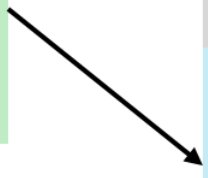
Combiner des bases de données : `merge`

economy2004.dta

country	GDP per Cap 2004
ARG	12,468
FRA	27,913
GER	28,889
ITA	28,172
USA	39,498

economy2005.dta

country	GDP per Cap 2005
ARG	13,153
FRA	29,203
GER	30,15
ITA	29,414
USA	41,557



Merged 1:1

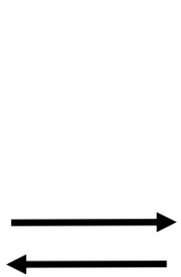
country	GDP per Cap 2004	GDP per Cap 2005
ARG	12,468	13,153
FRA	27,913	29,203
GER	28,889	30,15
ITA	28,172	29,414
USA	39,498	41,557

```
use economy2004.dta, replace
merge 1:1 country using economy2005.dta
```

Formats wide et long et transposition des données : `reshape`

Long

country	year	lifeexp
A	1	70
A	2	71
B	1	80
B	2	82
B	3	85
C	1	79
C	2	82
C	3	81



Wide

year	lifeexpA	lifeexpB	lifeexpC
1	70	80	79
2	71	82	82
3	.	85	81

- Long -> Wide

```
reshape (wide) lifeexp, i(year) j(country)
```

- Wide -> Long

```
reshape (long) lifeexpA lifeexpB lifeexpC, i(year) j(country)
```

Au menu de cette séance

Stata 101

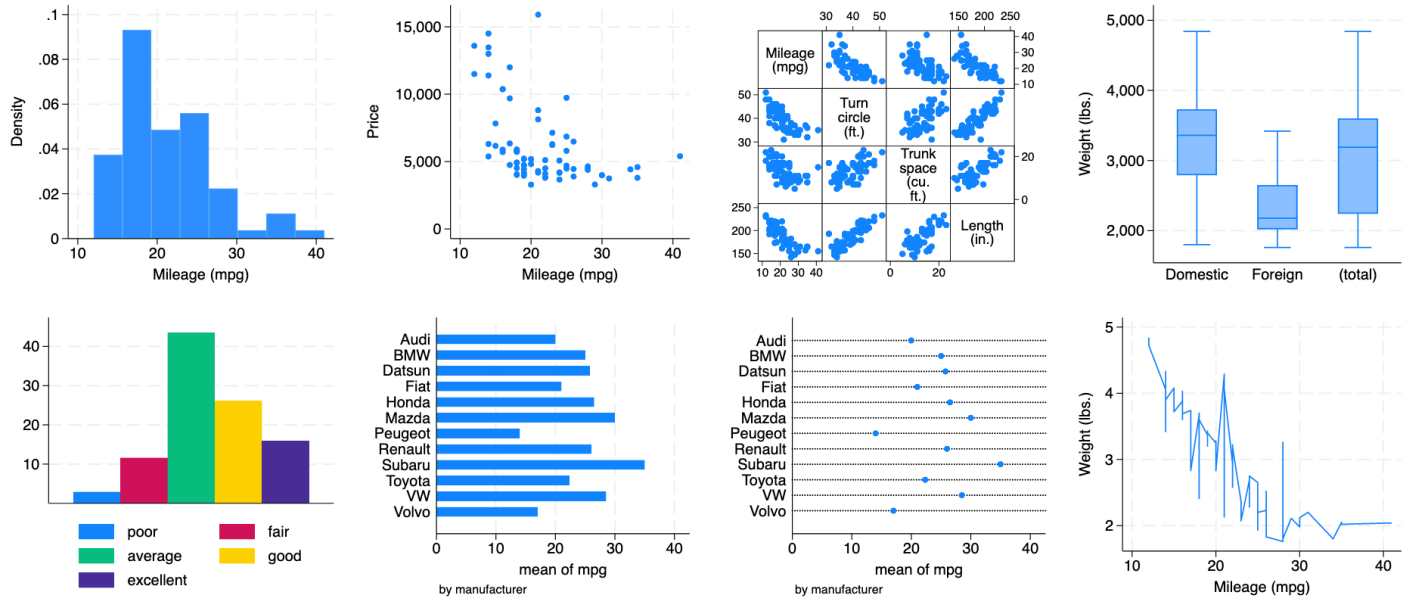
Nettoyer les données

Transformations avancées (aggrégations, combinaisons...)

Visualisation

Statistiques descriptives

Visualiser des données



Comment choisir la représentation graphique adaptée ?

Ca dépend des variables étudiées, et notamment :

- 1 2 (ou 3) variables
- Continue ou discrète

Visualiser des données - 1 variable

Variable continue

- ***Histogramme*** : visualiser la distribution d'une variable

```
histogram variable_1
```

- ***Densité*** : visualiser la distribution d'une variable

```
kdensity variable_1
```

Visualiser des données - 1 variable

Discrète

- ***Diagramme* en bâton** (*bar chart*): variable
  - Si la variable discrete est numérique:

```
hist variable_qualitative, discrete
```

  - Sinon, la transformer en numérique 😉.

Visualiser des données - 2 variables

- ***Scatter plot*** : montrer les valeurs d'une variable en fonction de l'autre:

```
twoway scatter variable_1 variable_2
```

- ***Diagramme* en bâton** (*bar chart*): variable quantitative selon les valeurs d'une variable qualitative (ou catégorielle)

```
graph bar (mean) variable_quantitative, over(variable_qualitative)
```

Une base de donnée "propre"

1. Chaque ***colonne*** correspond à une variable
2. Chaque ***ligne*** correspond à une observation

Avant de procéder à la visualisation, il faut se demander :

1. Quelle information veux-je représenter ?
    - Faire en particulier attention à la présence de potentielles variables manquantes
2. Est ce que les données contiennent cette information telle que ***une colonne/ligne*** correspond à ce que je veux représenter ?
    - Eventuellement utiliser `collapse` et ou `reshape` pour transformer les données

Types de visualisation

Quelques exemples des types de graphiques les plus utilisés:

- Commande `twoway` + type de graphique

Type	Function
Point	`scatter`
Line	`line`, `connected`
Histogram	`histogram`
Density	`kdensity`

- Commande `graph` + type de graphique

Type	Function
Bar	`bar`
Boxplot	`box`

Exercice 6 : Premiers graphiques

07:00

On utilise les données `gapminder`.

1. L'histogramme des espérances de vie en 2007. Ensuite, précisez la couleur en ajoutant l'option `color(green)` (*il faut mettre les options après une virgule*).

2. Un **scatter plot** du taux de fertilité (y-axis) vs. pib par habitant (x-axis) en 2007. Ensuite, spécifiez les titres des axes.

3. [CHALLENGE] Représentez l'évolution de l'espérance de vie au cours du temps par continent.

    - Il va falloir aggréger les données au niveau continent X year => vous pouvez faire cela avec une commande `collapse`
    - Il est ensuite plus simple de changer la forme des données pour les transformer de "long" en "wide" en utilisant un `reshape`

Au menu de cette séance

Stata 101

Nettoyer les données

Transformations avancées (aggrégations, combinaisons...)

Visualisation

Statistiques descriptives

Synthétiser les données

- En général, on apprend à connaître ses données par des visualisations + calculs de statistiques descriptives
- C'est l'heure des **statistiques descriptives** !
  - Utilisées pour décrire les données et simplifier l'information à partir de différentes mesures, tableaux et graphiques
  - (VS. ***statistiques inférentielles*** qui permettent ensuite de généraliser les résultats à la population d'intérêt.)
- En particulier, nous nous intéressons aux **tendances centrales** et à la **dispersion**.

Tendances centrales

`mean(x)`: la moyenne de toutes less valeurs de `x`.

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

```
set obs 100
gen y = rnormal()
egen y_mean=mean(y)
```

`Médiane`: la valeur de x_j qui partage les observations en 2 parties égales (50% au dessus, 50% en dessous). m est la médiane si

$$\Pr(X \leq m) \geq 0.5 \text{ and } \Pr(X \geq m) \geq 0.5$$

La médiane est robuste à la présence de valeurs *extrêmes*.

```
egen y_p50=median(y)
```

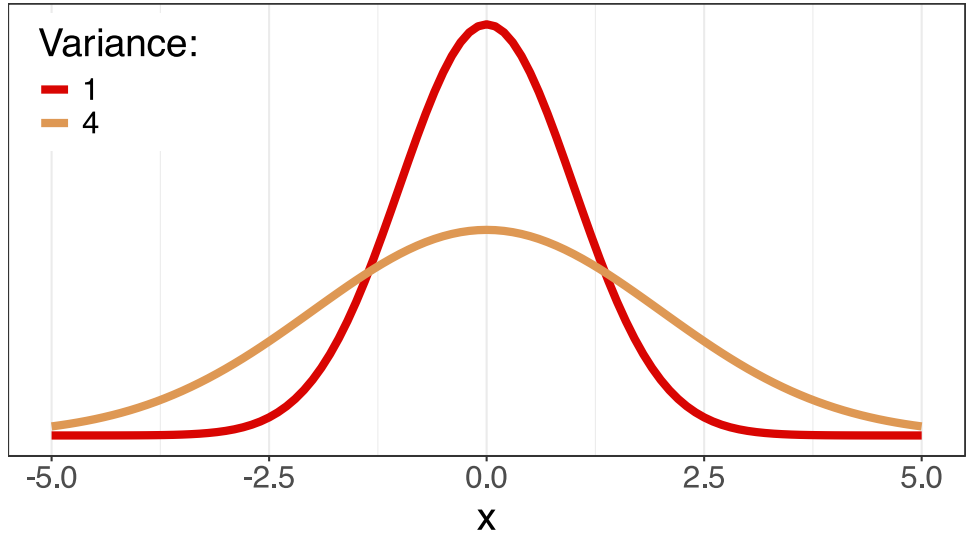
Dispersion

Une autre caractéristique intéressante est la mesure de l'écart d'une variable par rapport à son centre (la moyenne dans ce cas).

La *variance* est une telle mesure.

$$Var(X) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2$$

Soit deux `distributions normales` ayant une moyenne égale à `0`:



Commande :

```
var(x)
```

Statistiques descriptives: les commandes stata

Variables numériques

La commande `summarize` (`sum`) permet d'afficher des statistiques de base sur une ou plusieurs variables numériques de la base de données :

- la moyenne, l'écart-type et les valeurs extrêmes (min et max).
- L'option `detail` permet d'afficher des statistiques plus précises sur la distribution de la ou des variables: médiane et les autres centiles.

```
summarize gdp population, detail
```

Statistiques descriptives: les commandes stata

Variable discrète

`tabulate` (`tab`) permet d'afficher les tableaux de fréquences (i.e., la distribution) des variables numériques catégorielles ou des variables textuelles :

```
tabulate country
```

Statistiques croisées

Deux variables catégorielles

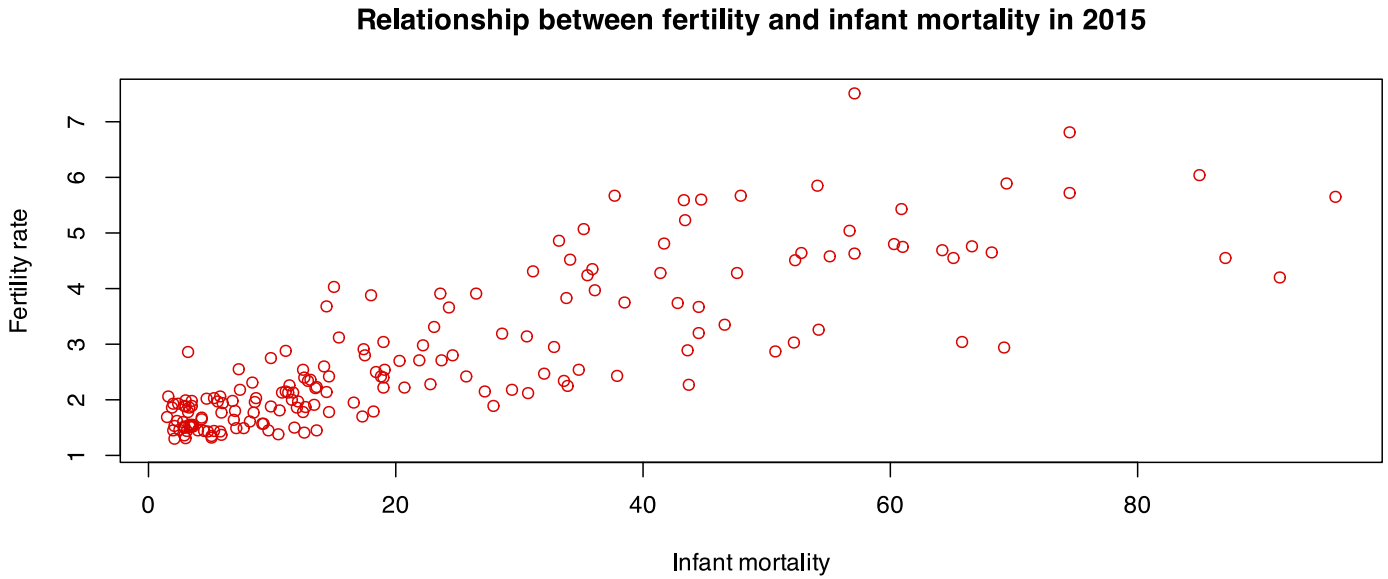
Dans le cas où il y a 2 variables spécifiées, `table` produit une table de contingence :

```
table year country
```

Deux variables : catégorielle vs. quantitative

```
tabstat fertility, by(year)
```

Comment x et y sont associées ? Covariance et corrélation



On s'intéresse principalement à 2 statistiques pour caractériser la relation entre x et y:

1. Covariance
2. Corrélation

Covariance

- La covariance est une mesure de la **variabilité jointe** de deux variables.

$$Cov(x, y) = \frac{1}{N} \sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y})$$

- Difficile à interpréter car sensible à la dispersions des variables par rapport à la moyenne.

Corrélation

- La corrélation est une mesure de la force de l' **association linéaire** entre deux variables.

$$Cor(x,y) = \frac{Cov(x,y)}{\sqrt{Var(x)}\sqrt{Var(y)}}$$

- La fonction `cor` calcule la corrélation:

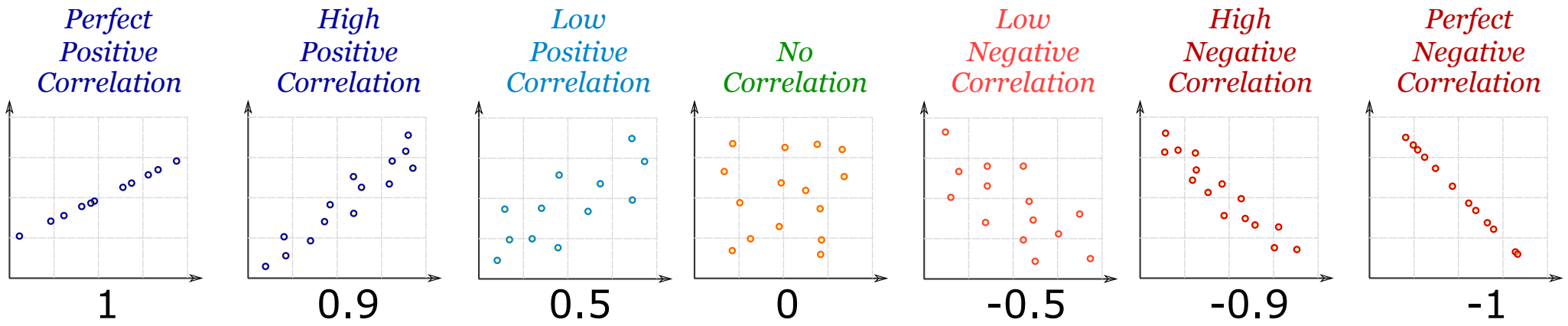
```
cor infant_mortality fertility
```

- On ajoute l'option `cov` si on veut la covariance

```
cor infant_mortality fertility, cov
```

Corrélation

- La **Corrélation est toujours entre -1 et 1!**



Source: mathisfun

Exercice 7: statistiques descriptives

07:00

1. Quelle est la moyenne du PIB en 2011 ?
2. La médiane ?
3. Calculez la moyenne de l'espérance de vie par année.
4. Calculez la corrélation entre la fertilité et la mortalité infantile en 2015.

[À retenir] Utiliser des do files

Dofile = un script contenant une suite de commandes stata

- **Commande vs. script**
  - Ligne de commande: pratique pour tester
  - Script de commandes (= `dofile`) :
    - permet à l'analyse d'être reproductible
  - Commenter ses dofiles: `//`

```
di 1+1 // commentaire sur une ligne

/*
Commentaire sur plusieurs lignes
*/
```

⇒ **toujours** utiliser un script

- Organiser son travail en **plusieurs dofiles**
  - ayant une suite logique :
    - créer la base de donnée → statistiques descriptives → analyse graphique
  - ayant des noms significatifs :
    - `1-construction-data.do` → `2-stat-des.do` → `3-graph.do`

[À retenir] Commandes principales en stata

Tâches	Commandes
Obtenir de l'aide	`help`, `findit`, `lookfor`
Utiliser les données Stata	`use`, `save`, `append`, `merge`
Data management	`reshape`, `collapse`
Gérer variables	`replace`, `rename`, `encode`, `sort`, `keep`, `drop`
Gérer observations	`keep if`, `drop if`
Créer/modifier des variables	`generate`, `replace`, `egen`
Visualiser les données	`describe`, `list`, `tabulate`, `summarize`
Calculatrice	`display`

Question de comprehension [Groupe de 2]

10:00

Choisissez un graphique issu d'un média grand public (article, journal, site internet, rapport) qui concerne une statistique ou une donnée économique.
Précisez les éléments suivant:

- La source et le contexte du graphique.
- Une description de ce que représente ce graphique.
- Ce que vous trouvez pertinent ou problématique dans ce graphique.
- Comment ce graphique peut être interprété ou discuté du point de vue de l'analyse économique et statistique.
- Suggérez une analyse statistique complémentaire

lien

Liste de sources

Où en sommes nous de notre quête de la causalité

✅ **Comment gérer les données?** Regardez-les, ordonnez-les, visualisez-les...

❌ Comment résumer une relation entre plusieurs variables?

❌ Qu'est ce que la causalité ?

❌ Comment faire si nous n'observons qu'une partie de la population ?

❌ Nos résultats sont ils uniquement dus au hasard?

❌ Comment trouver de l'exogénéité en pratique ?

À LA SEMAINE PROCHAINE !

mguillot@uliege.be

MERCI À

Florian Oswald et à toute l'équipe de ScPoEconometrics pour le livre et leurs ressources