Les bases de données

Table of contents

| SECTION | COMMANDES ET EXPRESSIONS |
| --- | --- |
| **Affectation d’un répertoire de travail** | cd |
| **Ouverture et sauvegarde d’une base** | use webuse [set] sysuse save saveold import export **usesas** **savasas** set obs insobs |
| **Décrire le contenu d’une base** | describe list codebook labelbook label list |
| **Tri, doublon, position des variables** | sort gsort duplicates order |
| **Decription statistique des variables** | summarize mean tabstat **violinplot** **heatplot** **gjoint** tabulate **fre** **tabm** **catplot** |

* **En gras, commandes externes**
* *En italique, commandes associées à un chapitre ultérieur*

# 1. **Affectation du répertoire de travail**

La commande cd (Current Directory), permet d’indiquer le chemin d’accès du répertoire où se trouve la base à ouvrir ou à enregistrer. Si aucun chemin d’accès n’est spécifié, Stata ira chercher la base dans le répertoire par défaut (normalement C: ou D:).

Syntaxe

cd "path"

**Remarque**  
Avec cette commande, un seul repertoire de travail est actif. On peut élargir les possibilités en affectant des répertoires avec des macros variables locales ou globales

# 2. **Ouverture et sauvegarde d’une base**

## 2.1 Ouverture

**Commande** use  
 Sans chargement d’un répertoire

use "path/nom\_base.dta" [,clear]

L’option clear permet d’effacer une base en cours d’utilisation. Il est conseillé de mettre cette option systématiquement. On peut également utiliser clear comme instruction avant d’ouvrir une base. on ne supprime pas la base du répertoire (commande erase), elle est juste écrasée dans la session.

Syntaxe avec affectation d’un répertoire:

cd "path"  
use "nom\_base.dta", clear

ou

cd "path"  
clear  
use "nom\_base.dta"

Remarque: pour les bases d’exemples préinstallées, on utilise la commande sysuse. Dans les fichiers d’aide, des exemples font également appels a des bases localisées sur des serveurs qui s’ouvrent avec la commande webuse.

sysuse auto, clear

|  |
| --- |
| Note |
| **Ouverture d’une base stockée sur un git [github, gitlab …]**  Dans un premier temps, comme pour cd il faut charger le répertoire où se trouve localisé la base, avec la commande **webuse set**. Par exemple sur mon dépôt git, une base d’exemple (*logement.dta*) pour une commande se trouve à cette adresse: <https://github.com/mthevenin/stata_graphiques/tree/main/programmation/gjoint>  Pour charger ce répertoire à distance:  webuse set "https://raw.githubusercontent.com//mthevenin/stata\_graphiques/master/ressources/gjoint"  On remarque que le chemin n’est pas identique au simple lien (spécificité des dépôt de type git).  Il suffit ensuite de charger la base avec webuse  webuse logement.dta, clear  On revient au dépôt officiel de stata avec webuse set  En résumé avec un seul bloc d’instructions:  webuse set "https://raw.githubusercontent.com//mthevenin/stata\_graphiques/master/ressources/gjoint"  webuse "logement.dta", clear  webuse set  > ster/ressources/gjoint" (prefix now "https://raw.githubusercontent.com//mthevenin/stata\_graphiques/mast > er/ressources/gjoint")   (prefix now "https://www.stata-press.com/data/r17") |

## 2.2 Sauvegarde

**Commandes** save ou saveold

save "path/nom\_base.dta" [, replace]

L’option replace permet d’écraser une version antérieure de la base. Obligatoire à partir de la 2ème sauvegarde, mais on peut l’utiliser dès la première sauvegarde (un message d’avertissement s’affiche).

sysuse auto, clear   
  
save auto, replace

(1978 automobile data)  
  
file auto.dta saved

saveold permettra d’ouvrir une bases avec une version ancienne de Stata non compatibles avec la version courante. Cela commence à devenir moins critique, mais avec save il ne sera plus possible d’ouvrir une base avec une version inférieure à la 13 ou inférieur à la 13 (passage à l’encodage Utf8 avec la version 14).

saveold "path/nom\_base.dta", [replace] [version(#)] // # = numéro de la version de Stata:

Remarque:

* Ecrire l’extension .dta n’est pas obligatoire
* Les doubles quotes ne sont obligatoires que s’il y a un espace dans le chemin d’accès et/ou dans le nom de la base

use "ma base", clear   
use ma\_base, clear

## 2.3 Autres formats

## 2.4 Importation/Exportation

**Excel et fichiers textes (.txt, .csv)**

* Le plus simple est passer par le menu: files + [*Import* ou *Export*] qui dispose d’une fenêtre de prévisualisation.
  + Pour excel les commandes sont import excel et export excel
  + Pour des fichiers textes type csv (R), les commandes sont import delimited et export delimited

\*\*Exemples

\* exportation csv  
export delimited using "D:\D\stata\_temp\export\_csv.csv", replace  
  
\* exportation xls  
export excel using "D:\D\stata\_temp\export\_excel.xls", firstrow(variables) replace  
  
\* importation cxv  
import delimited "D:\D\stata\_temp\export\_csv.csv", clear  
  
  
\* importation xls  
import excel "D:\D\stata\_temp\export\_excel.xls", sheet("Sheet1") firstrow clear

**SAS**

* Depuis la version 16 de Stata il est possible d’importer directement des formats sas7bdat. Pas d’exportation possible.
* Pour les versions antérieure, la solution installée via *sasxport* n’est pas satisfaisante. Il est alors conseillé d’utiliser le package externe savasas
  + Sas => Stata [importation]: commande usesas
  + Stata => Sas [exportation] : commande savasas
  + Si le chemin d’accès à l’exécutable de SAS Windows n’est pas reconnu, il faut récupérer et éditer le fichier sasexe.do (à partir de la ligne 169), dont l’accès est donné dans le fichier d’aide.
  + Pour l’exportation, on peut générer un catalogue de format,en dur, avec l’option format.

**SPSS**

* Depuis la dernière version de Stata (16), il est possible d’importer directement des bases de ce format.
* Pas de possibilité d’exportation directe.

## 2.5 Création d’une base de donnée, ajout d’observations

On peut créer une base de donnée, “vide”, avec la commande **set obs nbre\_observations**

clear   
set obs 100

A une base existante, on peut ajouter des observations (en valeurs manquante) avec la commande insobs nbre\_observation. Par défaut ces observations s’ajouteront après la dernière ligne (option before/after(position) pour renseigner la position de la première observation ajoutée.

insobs 10 // ajout de 10 observations à la base

# 3. **Décrire le contenu d’une base**

## 3.1 Commande describe

Permet, sous forme de tableau, d’avoir des renseignement sur une base de donnée: taille en mémoire, nombre d’observations, descriptif des variables (nom, format, labels). La commande est régulièrement tronquée jusqu’à des

**Base courante (ouverte)**

describe [varlist , short]

sysuse auto.dta, clear  
  
des

(1978 automobile data)  
  
  
Contains data from C:\PROGRA~1\Stata17\ado\base/a/auto.dta  
 Observations: 74 1978 automobile data  
 Variables: 12 13 Apr 2020 17:45  
 (\_dta has notes)  
-------------------------------------------------------------------------------  
Variable Storage Display Value  
 name type format label Variable label  
-------------------------------------------------------------------------------  
make str18 %-18s Make and model  
price int %8.0gc Price  
mpg int %8.0g Mileage (mpg)  
rep78 int %8.0g Repair record 1978  
headroom float %6.1f Headroom (in.)  
trunk int %8.0g Trunk space (cu. ft.)  
weight int %8.0gc Weight (lbs.)  
length int %8.0g Length (in.)  
turn int %8.0g Turn circle (ft.)  
displacement int %8.0g Displacement (cu. in.)  
gear\_ratio float %6.2f Gear ratio  
foreign byte %8.0g origin Car origin  
-------------------------------------------------------------------------------  
Sorted by: foreign

**Base stockée (non ouverte)**

On peut également décrire le contenu d’une base en format .dta en mémoire et non ouverte avec l’argument using "path/nombase"

describe using "https://www.stata-press.com/data/r17/census2.dta"

Contains data 1980 Census data by state  
 Observations: 50 2 Dec 2020 09:21  
 Variables: 15   
-------------------------------------------------------------------------------  
Variable Storage Display Value  
 name type format label Variable label  
-------------------------------------------------------------------------------  
state str13 %-13s State  
state2 str2 %-2s Two-letter state abbreviation  
region byte %-8.0g cenreg Census region  
pop long %12.0gc Population  
poplt5 long %12.0gc Pop, < 5 year  
pop5\_17 long %12.0gc Pop, 5 to 17 years  
pop18p long %12.0gc Pop, 18 and older  
pop65p long %12.0gc Pop, 65 and older  
popurban long %12.0gc Urban population  
medage float %9.2f Median age  
death long %12.0gc Number of deaths  
marriage long %12.0gc Number of marriages  
divorce long %12.0gc Number of divorces  
drate int %9.0g Deathrate  
age byte %9.0g Age  
-------------------------------------------------------------------------------  
Sorted by:

## 3.2 Autres commandes

* **Affichage de la base dans l’output**

Commande list

list [varlist] [expression]

list price mpg turn foreign in 1/10

| price mpg turn foreign |  
 |--------------------------------|  
 1. | 4,099 22 40 Domestic |  
 2. | 4,749 17 40 Domestic |  
 3. | 3,799 22 35 Domestic |  
 4. | 4,816 20 40 Domestic |  
 5. | 7,827 15 43 Domestic |  
 |--------------------------------|  
 6. | 5,788 18 43 Domestic |  
 7. | 4,453 26 34 Domestic |  
 8. | 5,189 20 42 Domestic |  
 9. | 10,372 16 43 Domestic |  
 10. | 4,082 19 42 Domestic |  
 +--------------------------------+

Sauf exceptions, comme la petite base d’exemple utilisée ici, penser à bien filtrer les informations souhaitées en termes de variables et d’observations.

* **Information sur les labels affectés aux variables**

Commande labelbook et label list

labelbook: affiche les informations sur les labels affectés aux modalités des variables.

labelbook [nom\_label]

labelbook origin

Value label origin   
-------------------------------------------------------------------------------  
  
 Values Labels  
 Range: [0,1] String length: [7,8]  
 N: 2 Unique at full length: yes  
 Gaps: no Unique at length 12: yes  
 Missing .\*: 0 Null string: no  
 Leading/trailing blanks: no  
 Numeric -> numeric: no  
 Definition  
 0 Domestic  
 1 Foreign  
  
 Variables: foreign

label list [nom\_label] donne seulement l’affectation des labels aux valeurs.

label list origin

origin:  
 0 Domestic  
 1 Foreign

# 4. **Tri, doublon, position des variables**

## 4.1 Tri d’une base

Commande sort

sort varlist

* La commande sort n’effectue que des tris croissants. Pour faire un tri décroissant, on peut utiliser la commande gsort. Tris croissants et décroissant peuvent se succéder dans une logique de cluster.
  + sort varlist => tri croissant
  + gsort + var1 => croissant *var1*
  + gsort - var1 => décroissant *var1*
  + gsort + var1 - var2 => croissant *var1* et décroissant *var2* dans chaque strate de *var1*
* Le tri d’une peut/doit être réalisé lorsqu’on veut répéter une instruction avec le préfixe bysort: bysort varlist: instruction. Il est imposé lorsqu’on souhaite apparier des bases [voir chapitre 6]

\* tri croissant sur la variable price  
sort price  
  
\* tri décroissant sur la variable prix pour chaque niveau de la variable foreign  
gsort + foreign - price

## 4.2 Repérage et suppression des doublons

**Repérage et suppression des doublons**

Commande duplicates list/tag/drop [varlist]

Permet de lister, repérer (avec gen(varname) ) ou supprimer des observations répliquées. Si la liste de variables n’est pas renseignée, elles toutes sont utilisées.

Syntaxe

duplicates list [varlist]  
  
duplicates tag [varlist], gen(var)  
  
duplicates drop [varlist]

## 4.3 Modifier la place des variables dans la base

Commande order

order varlist, [first/last] [after/before(varname)]

order foreign, first  
order rep78, after(foreign)  
  
des

Contains data from C:\PROGRA~1\Stata17\ado\base/a/auto.dta  
 Observations: 74 1978 automobile data  
 Variables: 12 13 Apr 2020 17:45  
 (\_dta has notes)  
-------------------------------------------------------------------------------  
Variable Storage Display Value  
 name type format label Variable label  
-------------------------------------------------------------------------------  
foreign byte %8.0g origin Car origin  
rep78 int %8.0g Repair record 1978  
make str18 %-18s Make and model  
price int %8.0gc Price  
mpg int %8.0g Mileage (mpg)  
headroom float %6.1f Headroom (in.)  
trunk int %8.0g Trunk space (cu. ft.)  
weight int %8.0gc Weight (lbs.)  
length int %8.0g Length (in.)  
turn int %8.0g Turn circle (ft.)  
displacement int %8.0g Displacement (cu. in.)  
gear\_ratio float %6.2f Gear ratio  
-------------------------------------------------------------------------------  
Sorted by: foreign

Cette opération est particulièrement pour organiser sa base de données après la création de nouvelles variables.

# 5. **Description statistique des variables**

Dans les menus de l’interface principale: **Statistics** => **Summaries, tables & tests**

## 5.1 Variables quantitatives

### 5.1.1 Tableaux d’indicateurs

|  |
| --- |
| Note |
| Les commandes qui sont rapidement décrites afficheront des indicateurs communs, typiquement la moyenne. Elles se distinguent par la forme de leur output facilitant plus ou moins les comparaisons, les possibilités offertes en termes de pondération, et sur la récupération des résultats (macro). |

**Commande summarize**

Comme son l’indique, la commande summarize, avec l’option detail (d) donne un résumé complet de la distribution d’une variable quantitative: moyenne, variance, quantiles, symétrie, applatissement ..nom l’indique l.)

summarize varlist [, detail]

Si on indique pas le nom d’au moins une variable, toutes les variables de la base seront sélectionnées.

La commande peut-être tronquée jusqu’à sum [**Warning**: il existe également une fonction sum pour générer des cumuls lors d’une création de variable - voir chapitre 5].

sum price  
  
sum

Variable | Obs Mean Std. dev. Min Max  
-------------+---------------------------------------------------------  
 price | 74 6165.257 2949.496 3291 15906  
  
 Variable | Obs Mean Std. dev. Min Max  
-------------+---------------------------------------------------------  
 make | 0  
 price | 74 6165.257 2949.496 3291 15906  
 mpg | 74 21.2973 5.785503 12 41  
 rep78 | 69 3.405797 .9899323 1 5  
 headroom | 74 2.993243 .8459948 1.5 5  
-------------+---------------------------------------------------------  
 trunk | 74 13.75676 4.277404 5 23  
 weight | 74 3019.459 777.1936 1760 4840  
 length | 74 187.9324 22.26634 142 233  
 turn | 74 39.64865 4.399354 31 51  
displacement | 74 197.2973 91.83722 79 425  
-------------+---------------------------------------------------------  
 gear\_ratio | 74 3.014865 .4562871 2.19 3.89  
 foreign | 74 .2972973 .4601885 0 1

sum price length, d

Price  
-------------------------------------------------------------  
 Percentiles Smallest  
 1% 3291 3291  
 5% 3748 3299  
10% 3895 3667 Obs 74  
25% 4195 3748 Sum of wgt. 74  
  
50% 5006.5 Mean 6165.257  
 Largest Std. dev. 2949.496  
75% 6342 13466  
90% 11385 13594 Variance 8699526  
95% 13466 14500 Skewness 1.653434  
99% 15906 15906 Kurtosis 4.819188  
  
 Length (in.)  
-------------------------------------------------------------  
 Percentiles Smallest  
 1% 142 142  
 5% 154 147  
10% 157 149 Obs 74  
25% 170 154 Sum of wgt. 74  
  
50% 192.5 Mean 187.9324  
 Largest Std. dev. 22.26634  
75% 204 221  
90% 218 222 Variance 495.7899  
95% 221 230 Skewness -.0409746  
99% 233 233 Kurtosis 2.04156

bysort foreign: sum price, d

-> foreign = Domestic  
  
 Price  
-------------------------------------------------------------  
 Percentiles Smallest  
 1% 3291 3291  
 5% 3667 3299  
10% 3955 3667 Obs 52  
25% 4184 3799 Sum of wgt. 52  
  
50% 4782.5 Mean 6072.423  
 Largest Std. dev. 3097.104  
75% 6234 13466  
90% 11385 13594 Variance 9592055  
95% 13594 14500 Skewness 1.777939  
99% 15906 15906 Kurtosis 5.090316  
  
-------------------------------------------------------------------------------  
-> foreign = Foreign  
  
 Price  
-------------------------------------------------------------  
 Percentiles Smallest  
 1% 3748 3748  
 5% 3798 3798  
10% 3895 3895 Obs 22  
25% 4499 3995 Sum of wgt. 22  
  
50% 5759 Mean 6384.682  
 Largest Std. dev. 2621.915  
75% 7140 9690  
90% 9735 9735 Variance 6874439  
95% 11995 11995 Skewness 1.215236  
99% 12990 12990 Kurtosis 3.555178

* ***Avantage***: récupération des résultats sous forme de macro rapide.
* ***Inconvénients***: pas de sélection des indicateurs avec l’option detail, output pas adapté aux comparaisons.

\*Extrait de l'aide de summarize  
  
/\* help summarize  
  
 summarize stores the following in r():  
  
 Scalars   
 r(N) number of observations  
 r(mean) mean  
 r(skewness) skewness (detail only)  
 r(min) minimum  
 r(max) maximum  
 r(sum\_w) sum of the weights  
 r(p1) 1st percentile (detail only)  
 r(p5) 5th percentile (detail only)  
 r(p10) 10th percentile (detail only)  
 r(p25) 25th percentile (detail only)  
 r(p50) 50th percentile (detail only)  
 r(p75) 75th percentile (detail only)  
 r(p90) 90th percentile (detail only)  
 r(p95) 95th percentile (detail only)  
 r(p99) 99th percentile (detail only)  
 r(Var) variance  
 r(kurtosis) kurtosis (detail only)  
 r(sum) sum of variable  
 r(sd) standard deviation  
\*/

qui sum price, d  
  
return list

scalars:  
 r(N) = 74  
 r(sum\_w) = 74  
 r(mean) = 6165.256756756757  
 r(Var) = 8699525.97426879  
 r(sd) = 2949.495884768919  
 r(skewness) = 1.653433511704859  
 r(kurtosis) = 4.819187528464004  
 r(sum) = 456229  
 r(min) = 3291  
 r(max) = 15906  
 r(p1) = 3291  
 r(p5) = 3748  
 r(p10) = 3895  
 r(p25) = 4195  
 r(p50) = 5006.5  
 r(p75) = 6342  
 r(p90) = 11385  
 r(p95) = 13466  
 r(p99) = 15906

**mean**

N’affiche que la moyenne et ses statistiques associées. L’option over permet de comparer les valeurs moyennes des modalités d’une variable catégorielle (over(varname)) ou un croisement des modalités de plusieurs variables (over(varlist))

mean price   
mean price, over(foreign)  
mean price, over(foreign rep78)

Mean estimation Number of obs = 74  
  
--------------------------------------------------------------  
 | Mean Std. err. [95% conf. interval]  
-------------+------------------------------------------------  
 price | 6165.257 342.872 5481.914 6848.600  
--------------------------------------------------------------  
  
  
Mean estimation Number of obs = 74  
  
-----------------------------------------------------------------  
 | Mean Std. err. [95% conf. interval]  
----------------+------------------------------------------------  
c.price@foreign |  
 Domestic | 6072.423 429.491 5216.449 6928.398  
 Foreign | 6384.682 558.994 5270.608 7498.756  
-----------------------------------------------------------------  
  
  
Mean estimation Number of obs = 69  
  
-----------------------------------------------------------------------  
 | Mean Std. err. [95% conf. interval]  
----------------------+------------------------------------------------  
c.price@foreign#rep78 |  
 Domestic#1 | 4564.500 369.500 3827.174 5301.826  
 Domestic#2 | 5967.625 1265.494 3442.372 8492.878  
 Domestic#3 | 6607.074 704.611 5201.044 8013.104  
 Domestic#4 | 5881.556 530.673 4822.614 6940.497  
 Domestic#5 | 4204.500 220.500 3764.499 4644.501  
 Foreign#1 | 0.000 (no observations)  
 Foreign#2 | 0.000 (no observations)  
 Foreign#3 | 4828.667 742.249 3347.532 6309.801  
 Foreign#4 | 6261.444 632.031 5000.247 7522.642  
 Foreign#5 | 6292.667 921.876 4453.091 8132.242  
-----------------------------------------------------------------------

* **Avantage**: output synthétique si la moyenne de plusieurs groupes comparées
* **Inconvénients**: récupération des résultats via une matrice (on oublie)

**tabstat**

Permet de sélectionner les indicateurs avec l’option stat() (par défaut la moyenne). L’option by() permet de comparer le ou les indicateurs pour chaque niveau d’une variable catégorielle. Dans ce cas tabstat affiche également les résultats sur l’ensemble des observations (sinon ajouter l’option nototal).

tabstat price   
tabstat price mpg, by(foreign)

Variable | Mean  
-------------+----------  
 price | 6165.257  
------------------------  
  
  
Summary statistics: Mean  
Group variable: foreign (Car origin)  
  
 foreign | price mpg  
---------+--------------------  
Domestic | 6072.423 19.82692  
 Foreign | 6384.682 24.77273  
---------+--------------------  
 Total | 6165.257 21.2973  
------------------------------

\*Extrait de l'aide de summarize  
  
/\* help tabstat  
  
 mean mean  
 count count of nonmissing observations  
 n same as count  
 sum sum  
 max maximum  
 min minimum  
 range range = max - min  
 sd standard deviation  
 variance variance  
 cv coefficient of variation (sd/mean)  
 semean standard error of mean (sd/sqrt(n))  
 skewness skewness  
 kurtosis kurtosis  
 p1 1st percentile  
 p5 5th percentile  
 p10 10th percentile  
 p25 25th percentile  
 median median (same as p50)  
 p50 50th percentile (same as median)  
 p75 75th percentile  
 p90 90th percentile  
 p95 95th percentile  
 p99 99th percentile  
 iqr interquartile range = p75 - p25  
 q equivalent to specifying p25 p50 p75  
\*/

Si on souhaite ajouter la médiane

tabstat price mpg, by(foreign) stat(mean median)

Summary statistics: Mean, p50  
Group variable: foreign (Car origin)  
  
 foreign | price mpg  
---------+--------------------  
Domestic | 6072.423 19.82692  
 | 4782.5 19  
---------+--------------------  
 Foreign | 6384.682 24.77273  
 | 5759 24.5  
---------+--------------------  
 Total | 6165.257 21.2973  
 | 5006.5 20  
------------------------------

### 5.1.2 Graphiques

**[MAJ EN COURS: bcp de nouveautés]**

Juste une rapide présentation de quelques visualisations permettant d’explorer des distributions.

Une seule distribution: **graph box/hbox** , **histogram**, **violinplot** (externe)  
Plusieurs distributions: **violinplot** (externe) Deux distribution croisée: **hexplot** (externe), **gjoint** (externe)

**Boxplot**

graph hbox mpg,   
graph hbox mpg, over(foreign)

|  |  |  |
| --- | --- | --- |
| |  | | --- | | Sans comparaison | |  |

|  |  |
| --- | --- |
| |  | | --- | | Avec comparaison | |

**Histogramme**

histogram mpg, percent   
histogram mpg, percent by(foreign)

|  |  |  |
| --- | --- | --- |
| |  | | --- | | Sans comparaison | |  |

|  |  |
| --- | --- |
| |  | | --- | | Avec comparaison | |

* Dans le langage de Stata, ce type de graphique est appelé *oneway*.
* L’altération de son aspect avec les options n’est pas très flexible, surtout au niveau des couleurs.
* Pour les histogrammes, il y a une version *twoway* qui permet d’empilé plusieurs histogrammes dans un même graphique, mais généralement la visualisation n’est pas optimale. Préférer en ce cas là une approche par les **densités** (voir violinplot - comparaison III).
* Conseil de sémiologie graphique: garder l’axe quantitatif/continu sur les abcisses .

**Violinplot (Ben Jann)**

Toujours beaucoup d’options dans les commandes de *magik* B.Jann. Se reporter à son tutoriel sur github pour l’installation (nécessite l’installation de dépendances) [Lien](https://github.com/benjann/violinplot/)

violinplot mpg, fill  
  
violinplot mpg, nobox over(foreign) left overlay nomedian dscale(.)   
  
violinplot mpg, fill over(foreign)  
  
violinplot mpg, fill split(foreign) horizontal

|  |  |  |
| --- | --- | --- |
| |  | | --- | | Sans comparaison | |  |

|  |  |  |
| --- | --- | --- |
| |  | | --- | | Comparaison I | |  |

|  |  |  |
| --- | --- | --- |
| |  | | --- | | Comparaison II | |  |

|  |  |
| --- | --- |
| |  | | --- | | Comparaison III | |

**Deux variables quantitatives**

* Une nuage de point pêche rapidement lorsque le nombre d’observations augmente, par exemple audelà de 200.
* Solutions:
  + Courbes de niveaux
  + Heatplot/hexplot: l’idée est de visualiser un histogramme “vu du dessuss”, la hauteurs des barres étant données par un différentiel de couleur issues d’une palette séquentielle (du clair au foncé par exemple).

ssc install heatplot, replace

* Il peut-être intéressant d’ajouté les distributions marginales des deux variables. J’ai programmé une petite commande (encore en version très alpha): **gjoint**. Tout le mérite revient à B.Jann pour la commande hexplot (j’ai juste combiné hexplot avec des histogrammes).

net install gjoint, from("https://raw.githubusercontent.com/mthevenin/stata\_graphiques/master/ressources/gjoint/") replace

hexplot price mpg, colors(flare)  
gjoint price mpg, palette(flare)

|  |  |  |
| --- | --- | --- |
| |  | | --- | | hexplot | |  |

|  |  |
| --- | --- |
| |  | | --- | | hexplot | |

## 5.2 Variables catégorielles

La principale commande est tabulate (tab). On peut l’utiliser avec des variables de type string.

Syntaxe (tableau croisé)

tab var1 var2 [, mis nofreq row col sort ......]

* Par défaut, l’ordre d’affichage suis la valeur de la modalité si la variable est de type numérique et l’ordre alphabétique pour une variable de type caractère. On peut utiliser l’option sort pour afficher par ordre croissant des effectifs observés [ou utiliser la commande externe **tabsort** qui offre plus de possibilités]
* Autres commandes (externe): **fre** [B.Jann], **tabm** [NJ.Cox], **tabsort** [NJ.Cox]

ssc install fre  
ssc install tabsort  
ssc install tabm

tab rep78, mis  
tab rep78, mis sort   
tab rep78 foreign, nolab mis  
  
fre rep78   
  
tabsort rep78

Repair |  
record 1978 | Freq. Percent Cum.  
------------+-----------------------------------  
 1 | 2 2.70 2.70  
 2 | 8 10.81 13.51  
 3 | 30 40.54 54.05  
 4 | 18 24.32 78.38  
 5 | 11 14.86 93.24  
 . | 5 6.76 100.00  
------------+-----------------------------------  
 Total | 74 100.00  
  
 Repair |  
record 1978 | Freq. Percent Cum.  
------------+-----------------------------------  
 3 | 30 40.54 40.54  
 4 | 18 24.32 64.86  
 5 | 11 14.86 79.73  
 2 | 8 10.81 90.54  
 . | 5 6.76 97.30  
 1 | 2 2.70 100.00  
------------+-----------------------------------  
 Total | 74 100.00  
  
 Repair |  
 record | Car origin  
 1978 | 0 1 | Total  
-----------+----------------------+----------  
 1 | 2 0 | 2   
 2 | 8 0 | 8   
 3 | 27 3 | 30   
 4 | 9 9 | 18   
 5 | 2 9 | 11   
 . | 4 1 | 5   
-----------+----------------------+----------  
 Total | 52 22 | 74   
  
  
rep78 -- Repair record 1978  
-----------------------------------------------------------  
 | Freq. Percent Valid Cum.  
--------------+--------------------------------------------  
Valid 1 | 2 2.70 2.90 2.90  
 2 | 8 10.81 11.59 14.49  
 3 | 30 40.54 43.48 57.97  
 4 | 18 24.32 26.09 84.06  
 5 | 11 14.86 15.94 100.00  
 Total | 69 93.24 100.00   
Missing . | 5 6.76   
Total | 74 100.00   
-----------------------------------------------------------  
  
 Repair |  
record 1978 | Freq. Percent Cum.  
------------+-----------------------------------  
 3 | 30 43.48 43.48  
 4 | 18 26.09 69.57  
 5 | 11 15.94 85.51  
 2 | 8 11.59 97.10  
 1 | 2 2.90 100.00  
------------+-----------------------------------  
 Total | 69 100.00

**Graphiques**

Niveau graphique, les possibilités restent toujours assez limitées pour les variables discrètes (et on abandonne l’idée des horribles camemberts).

On privilégiera ici la commande de NJ.Cox, **catplot**

ssc install catplot

catplot rep78, percent   
catplot rep78, percent over(foreign)

|  |  |  |
| --- | --- | --- |
| |  | | --- | | Sans comparaison | |  |

|  |  |
| --- | --- |
| |  | | --- | | Avec comparaison | |

# 6. **Introduction aux frames**

* Depuis la version 16 (2019)
* Les frames permettent de travailler en parallèle sur plusieurs bases, sans switcher avec des opérations successives d’enregistrement/ouverture (save/use).
  + Sur l’interface principale, le contenu d’une frame (base de données) est affiché de manière traditionnelle. On peut faire des opérations sur les autres frames déclarées simultanément.
  + La première base ouverte lors de l’ouverture d’une session est déclarée comme frame par défaut.
* Les frames peuvent être liées entre elles avec une clé d’identification commune.
  + Importations partielles de variables ou d’observations d’une frame à l’autre.
  + Permet de générer une variable dans une frame en utilisant des variables d’une ou plusieurs autres frames. Il n’est donc pas nécessaire d’apparier les bases entre elles en amont.

On ne verra ici que quelques manipulations de bases, la liaison de frames sera traitée dans le chapitre 6.

**frame dir** ou **frame mist**

frame dir  
frame list

default 74 x 12; 1978 automobile data  
  
 default 74 x 12; 1978 automobile data

sysuse auto, clear  
  
frame dir

(1978 automobile data)  
  
 default 74 x 12; 1978 automobile data

**frame rename**

Comme son nom l’indique on renomme une frame avec ancien nom nouveau nom

frame rename default auto  
  
frame dir

auto 74 x 12; 1978 automobile data

**frame nom\_frame:** ou **frame nom\_frame {}**

* On peut exécuter une commande en la préfixant par **frame nom\_frame:**
* Pour une série de commandes, il suffit d’enchasser cette série dans des crochets

frame auto: mean price

Mean estimation Number of obs = 74  
  
--------------------------------------------------------------  
 | Mean Std. err. [95% conf. interval]  
-------------+------------------------------------------------  
 price | 6165.257 342.872 5481.914 6848.600  
--------------------------------------------------------------

frame auto {  
mean mpg  
table rep78 foreign   
}

Mean estimation Number of obs = 74  
  
--------------------------------------------------------------  
 | Mean Std. err. [95% conf. interval]  
-------------+------------------------------------------------  
 mpg | 21.297 0.673 19.957 22.638  
--------------------------------------------------------------  
  
------------------------------------------------  
 | Car origin   
 | Domestic Foreign Total  
-------------------+----------------------------  
Repair record 1978 |   
 1 | 2 2  
 2 | 8 8  
 3 | 27 3 30  
 4 | 9 9 18  
 5 | 2 9 11  
 Total | 48 21 69  
------------------------------------------------

**frame copy**

Permet de copier à l’idendique une frame **frame copy nom\_frame nouveau\_nom**.

Dans l’exemple qui suit on va une frame à partir de la base auto, frame *prix*, en conservant avec seulement les variables *foreign* et *price*.

frame copy auto prix  
  
frame dir

auto 74 x 12; 1978 automobile data  
 prix 74 x 12; 1978 automobile data

Comme indiqué précédemment, on est pas obligé de charger une base déclarée en frame pour effectuer des opérations dessus. On peut donc conserver les deux seules variables *foreign* et *price* tout en gardant la base *auto* chargée.

frame prix: keep foreign price  
frame dir

auto 74 x 12; 1978 automobile data  
\* prix 74 x 2; 1978 automobile data  
  
Note: Frames marked with \* contain unsaved data.

La frame *prix* ne comporte donc plus que deux variables.

On va supprimer la variable *make* de la base/frame *auto*

drop make

**frame change**

Permet de switcher d’une frame à une autre. Ici ce sont les informations de la frame *prix* qui seront chargée dans l’interface de Stata

frame change prix   
  
mean price, over(foreign)

Mean estimation Number of obs = 74  
  
-----------------------------------------------------------------  
 | Mean Std. err. [95% conf. interval]  
----------------+------------------------------------------------  
c.price@foreign |  
 Domestic | 6072.423 429.491 5216.449 6928.398  
 Foreign | 6384.682 558.994 5270.608 7498.756  
-----------------------------------------------------------------

Du côté de la frame *auto*, base chargée initialement, on note que la suppression de la variable *make* a bien été enregistrée malgré le change de frame.

frame auto: des

Contains data from C:\PROGRA~1\Stata17\ado\base/a/auto.dta  
 Observations: 74 1978 automobile data  
 Variables: 11 13 Apr 2020 17:45  
 (\_dta has notes)  
-------------------------------------------------------------------------------  
Variable Storage Display Value  
 name type format label Variable label  
-------------------------------------------------------------------------------  
price int %8.0gc Price  
mpg int %8.0g Mileage (mpg)  
rep78 int %8.0g Repair record 1978  
headroom float %6.1f Headroom (in.)  
trunk int %8.0g Trunk space (cu. ft.)  
weight int %8.0gc Weight (lbs.)  
length int %8.0g Length (in.)  
turn int %8.0g Turn circle (ft.)  
displacement int %8.0g Displacement (cu. in.)  
gear\_ratio float %6.2f Gear ratio  
foreign byte %8.0g origin Car origin  
-------------------------------------------------------------------------------  
Sorted by: foreign  
 Note: Dataset has changed since last saved.

**frame drop** et **frame reset**

Permettent de supprimer une ou la totalité des frames

**frame drop nom\_frame**, permet de supprimer une frame, à l’exception de celle chargée dans l’interface.

frame change auto  
  
frame drop prix  
  
frame dir

\* auto 74 x 11; 1978 automobile data  
  
Note: Frames marked with \* contain unsaved data.

On a charger dans l’interface la frame *auto*, puis on a supprimé la frame prix.

Par contre il n’est pas possible de supprimer la frame active dans l’interface.

frame drop auto

may not drop current frame  
r(119);  
  
end of do-file  
r(119);

**frame reset**

On peut supprimer toutes les frames, dont celle chargée dans l’interface avec **frame reset**. Dans ce cas il n’y a plus de base chargée dans la session.

frame reset   
  
des

Contains data  
 Observations: 0   
 Variables: 0   
Sorted by: