Les distributions

Table of contents

# 1. Une variable

Principaux outils de visualisations:

* **Histogrammes**
* Report des **densités** sous forme de courbes ou d’aires
* **Boxplot**: résumé synthétique d’indicateurs de distributions (quantiles)
* **violinplot**: visualisation unifiant boxplot et densités

## 1.1 Histogramme

* Graphique traditionnel
* Quantité utilisé sur Y facilement compréhensible: fréquence, pourcentage
* Modification des intervalles sur l’axe X
* Points de vigilance:
  + Empilement de plusieurs histogrammes peu convaincant. Utiliser de préférences des graphiques combinés type *small multiple*
    - La définition des bornes peut différer d’un logiciel à l’autre, c’est le cas entre Stata R (ggplot)
* **Commandes Stata**:
  + **histogram**
  + **tw histogram** (de préférence). Avec l’option by permet de faire des graphiques de type *small multiple* lorsqu’on souhaite comparer une distribution entre différentes catégories.

**Exemples**

\* style appliqué avec grstyle  
  
grstyle init  
grstyle set mesh, compact horizontal

tw histogram wage, fc("225 50 67%80") lc(black) lw(\*.2) percent  
  
\* Empilement histogramme: not good  
  
local opts lc(black) lw(\*.2) percent   
tw histogram wage if !south, fc("247 208 181%80") `opts' ///  
|| histogram wage if south, fc("225 50 67%80") `opts' ///  
|| , legend(order(1 "Not South" 2 "South") region(color(%0)))  
  
  
local opts fc(%0) lw(\*1) percent   
tw histogram wage if !south, lc("247 208 181%80") `opts' ///  
|| histogram wage if south, lc("225 50 67%80") `opts' ///  
|| , legend(order(1 "Not South" 2 "South") region(color(%0)))  
  
  
\* Facettes histogrammes: good  
tw histogram wage , fc("225 50 67%80") lc(black) lw(\*.2) percent by(south, note(" "))

|  |  |  |
| --- | --- | --- |
| |  | | --- | | Histogramme simple | |  |

|  |  |  |
| --- | --- | --- |
| |  | | --- | | Empilement Histogrammes: not good | |  |

|  |  |  |
| --- | --- | --- |
| |  | | --- | | Empilement Histogrammes: not good | |  |

|  |  |
| --- | --- |
| |  | | --- | | Facettes Histogrammes: good | |

|  |
| --- |
| Options by: à savoir |
| * Avec l’option by, plusieurs options générales du graphiques: titres, notes… sont à mette en option de by. Ici pour supprimer la note: **by(south, note(""))** * Il n’est pas possible d’affecter une couleur différente à chaque sous graphique. Si on souhaite modifier les couleurs, il convient alors de faire un graphique combiné, en générant tous les histogrammes individuels. Au final, une opération assez fastidieuse qui n’apporte pas d’information supplémentaire. |

## 1.2 Densités

* Graphique plébiscité actuellemeny en visualisation
* Basé sur des techniques de lissage, plus d’hypthèses que pour l’histogramme
* Comparaison par empilement des densités sur un même graphique ne pose pas de problème si leur nombre n’est pas trop élevé.
* Variantes:
  + densité en mirroir
  + Ajout d’une box plot (*violinplot* - voir plus loin)
* Point de vigilance:
  + les valeurs sur l’axe des abscisses peuvent sortir des bornes acceptables, par exemple un âge négatif.
* **Comandes Stata**
  + pour le report d’une seule densité, la commande traditionnelle est **kdensity**. Ajouter l’option recast(area) si l’on souhaite reporter la densité sous forme d’aire.
  + **violinplot** (B.Jann). Très récente, et très complète.
  + **joyplot** (A.Naqvi) ou **gridge** (MT). Si le nombre de comparaisons devient trop élevé, disons au moins 5-6, on peut utiliser ce type de visualisation, qui standardise les valeurs des densités, et qui permet par de jouer sur le niveau de superposition.

\* Penser à installer les dépendances, en particulier colorpalette  
\* Version github  
  
net from https://raw.githubusercontent.com/benjann/violinplot/main/  
net install violinplot, replace

**Sans comparaison**

\* une densité avec kdensity  
kdensity wage, note("") recast(area) fc("225 50 67") lc(black) lw(\*.5) title("")  
  
\* une densité avec violinplot  
violinplot wage, colors("225 50 67") fill p1(lc(black) lw(\*.2)) pdf(ll(0)) ///  
left nomed nobox nowhisker  
  
\* une densité en mirroir avec liolinplot  
violinplot wage, colors("225 50 67") fill p1(lc(black) lw(\*.2)) ///  
nomed nobox nowhisker

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| |  | | --- | | Avec **kdensity** | |  | |  | | --- | | Avec **violinplot** | |

|  |  |  |
| --- | --- | --- |
|  | |  | | --- | | **violinplot**: densité en mirroir | |

**Avec comparaison**

\*\*\*\* Comparaison  
\* Overlay demi Remplissage  
violinplot wage, ///  
over(south) overlay horizontal left nomed ///  
colors("247 208 181" "225 50 67", opacity(80)) lcolors(black) fill p1(lw(\*.2)) p2(lw(\*.2)) ///  
legend(order(1 "Not South" 2 "South") region(color(%0)))  
  
\* Overlay demi sans remplissage  
violinplot wage, ///  
over(south) overlay horizontal nomed left ///  
colors("247 208 181" "225 50 67", opacity(80)) ///  
legend(order(1 "Not South" 2 "South") region(color(%0)))   
  
\* Split  
violinplot wage, split(south) nomed nobox nowhisker ///  
 colors("247 208 181" "225 50 67") ///  
 legend(order(1 "Not South" 2 "South") region(color(%0)))   
   
\* Overlay mirroir sans remplissage  
violinplot wage, ///  
over(south) overlay nomed ///  
colors("247 208 181" "225 50 67") ///  
legend(order(1 "Not South" 2 "South") region(color(%0)))

|  |  |  |
| --- | --- | --- |
| |  | | --- | | **violinplot**: comparaison aire | |  |

|  |  |  |
| --- | --- | --- |
| |  | | --- | | **violinplot**: comparaison courbe | |  |

|  |  |  |
| --- | --- | --- |
| |  | | --- | | **violinplot**: comparaison split | |  |

|  |  |
| --- | --- |
| |  | | --- | | **violinplot**: comparaison mirroir overlay | |

Quelques précision sur les options de la commande violinplot

* Comme l’objectif n’est que de reporter les densités, on a retiré les éléments figurants dans une *boxplot* avec **nomed**, **nob** et **now** (syntaxe tronquée).
* **over()** permet de représenter les densités avec les différentes modalités sur un axe discret. On peut empiler sur une même origine l’axe discret avec l’option **overlay**.
* Les estimations de la densités peuvent sortir des bornes acceptables, ici par exemple un salaire horaire négatif. On peut renseigner les limites des bornes avec **ll(min)**, **ul(max)** ou **range(min max)** comme arguments de l’option **pdf()**. Dans l’exemple: ll(0) comme valeur minimal de x.
* Par défaut les densités sont reportées en mirroir (principe de ce type de graphique). on peut retirer une des deux densités avec **left** ou **right**.
* Remarque: pour les couleurs de remplissage, visiblement B.Jann a réduit l’intensité. Je ne suis pas arrivé à rétablir les couleurs d’origine de la palette choisie.

Si le nombre de modalité s’élève, il est préférable de ne pas activer l’option overlay, ou d’utiliser un graphique type **ligne de ridge**. Pour cette dernière option j’ai pris ma solution [[**gridge**](https://mthevenin.github.io/stata_graphiques/programmation/gridge/gridge.html)] qui permet de contrôler l’ordre des variables reportées sur l’axe discret. On peut également et sans soucis se diriger vers la commanfe **joyplot** d’A.Naqvi

\* Les dépendances sont installées si nécessaire  
  
net install gridge, from("https://raw.githubusercontent.com/mthevenin/stata\_graphiques/master/ressources/gridge/") replace

recode occupation (9 10 11 12 = 13 )  
  
\* violinplot  
violinplot wage, over(occupation) nomed nobox nowhisker ///  
colors(tableau) pdf(ll(0)) fill n(99)  
  
\* gridge  
gridge wage, over(occupation) sortrev(mean) range(0 50) palette(flare) bw(.5)

|  |  |  |
| --- | --- | --- |
| |  | | --- | | **axe discret avec violinplot** | |  |

|  |  |
| --- | --- |
| |  | | --- | | **Lignes de ridge avec gridge** | |

## 1.3 Graphique synthétique: boxplot et violinplot

[14-03-2023]