统计学（Stata实现）

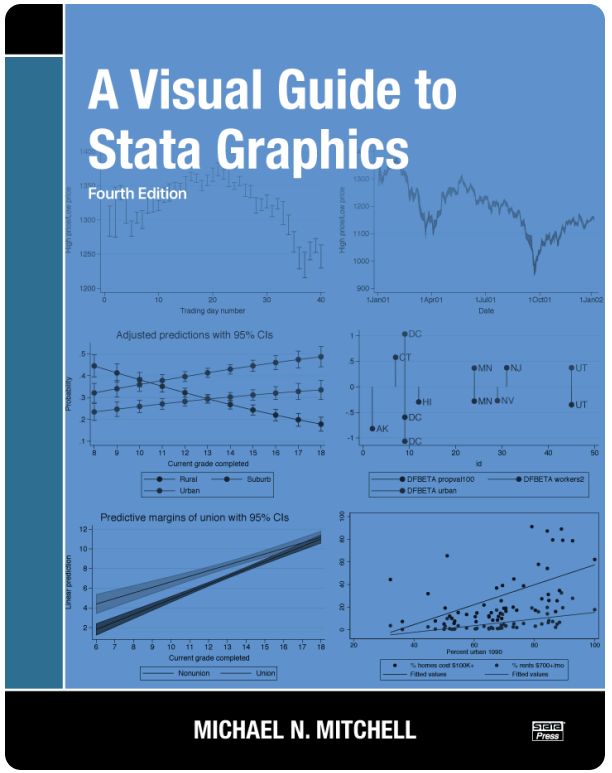
第一次上机课

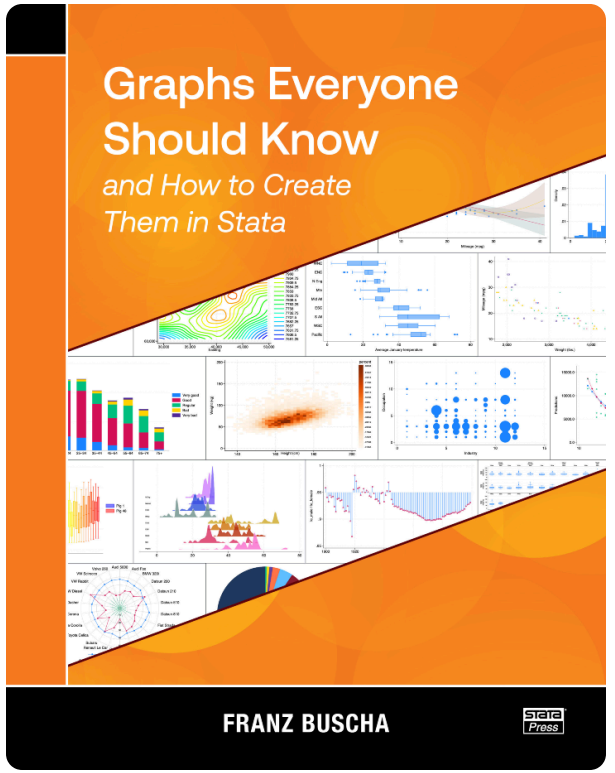
赵震宇（2120253538）
南开大学 国际经济研究所
zzynankai@outlook.com
xishanyu2.github.io
2025 年 10 月 9 日

参考资料

- 张甜，杨维忠编著. Stata统计学与案例应用精解. 清华大学出版社. 2025
- 谢慧明. 经济学实证论文写作讲义：方法与应用. 经济科学出版社. 2024
- 陈强. 计量经济学及Stata应用. 高等教育出版社. 2023
- 陈舒艳编. 统计学：Stata应用与分析. 机械工业出版社. 2020
- 连玉君老师、司继春老师课件、资料等

- Stata Conferences↗:
  - Data visualization with Stata↗
- Stata Bookstore↗:
  - A Visual Guide to Stata Graphics, Fourth Edition↗
  - Graphs Everyone Should Know and How to Create Them in Stata↗
- Documentation | Stata↗






第1章 导论
1.1 统计及其应用领域
1.2 统计数据的类型
1.3 统计中的几个基本概念
第2章 数据的搜集
2.1 数据的来源
2.2 调查方法
2.3 实验方法
2.4 数据的误差
第3章 数据的图表展示
3.1 数据的预处理
3.2 品质数据的整理与展示
3.3 数值型数据的整理与展示
3.4 合理使用图表
第4章 数据的概括性度量
4.1 集中趋势的度量
4.2 离散程度的度量
4.3 偏态与峰态的度量
第5章 概率与概率分布
5.1 随机事件及其概率
5.2 离散型随机变量及其分布
5.3 连续型随机变量的概率分布
第6章 统计量及其抽样分布
6.1 统计量
6.2 由正态分布导出的几个重要分布
6.3 样本均值的分布与中心极限定理
第7章 参数估计
7.1 参数估计的基本原理
7.2 一个总体参数的区间估计
7.3 两个总体参数的区间估计
7.4 样本量的确定
第8章 假设检验
8.1 假设检验的基本问题
8.2 一个总体参数的检验
8.3 两个总体参数的检验
8.4 检验问题的进一步说明


第9章 分类数据分析
9.1 分类数据与统计量
9.2 拟合优度检验
9.3 列联分析：独立性检验
9.4 列联表中的相关测量
9.5 列联分析中应注意的问题
第10章 方差分析
10.1 方差分析引论
10.2 单因素方差分析
10.3 双因素方差分析
第11章 一元线性回归
11.1 变量间关系的度量
11.2 一元线性回归
11.3 利用回归方程进行预测
11.4 残差分析
第12章 多元线性回归
12.1 多元线性回归模型
12.2 回归方程的拟合优度
12.3 显著性检验
12.4 多重共线性
12.5 利用回归方程进行预测
12.6 变量选择与逐步回归
第13章 时间序列分析和预测
13.1 时间序列及其分解
13.2 时间序列的描述性分析
13.3 时间序列预测的程序
13.4 平稳序列的预测
13.5 趋势型序列的预测
13.6 复合型序列的分解预测
第14章 指数
14.1 基本问题
14.2 总指数编制方法
14.3 指数体系
14.4 几种典型的指数
14.5 综合评价指数

CONTENT

目录

1. 引言

1.1 Why use Stata?

1. `Excel`, `SPSS`, `EViews`: 点击式 v.s `Stata`, `R`, `Python`: 命令式——易于复现!
2. 封装好的命令（简洁！）、丰富的内置数据集与外部命令
3. 全新的画图风格（从Stata18开始）
4. 后续学习：计量经济学、因果推断、机器学习（从Stata19开始）、经济学实证论文写作

   ......

1.2 Why visualize data?

> A picture is worth a thousand words. —— Barnard (1927)

1.3 Why use Stata for data visualization?

- Creating graphs in Stata is easy.
- Stata supports a wide variety of plots.
- Stata graphic commands are highly customizable and extensible.

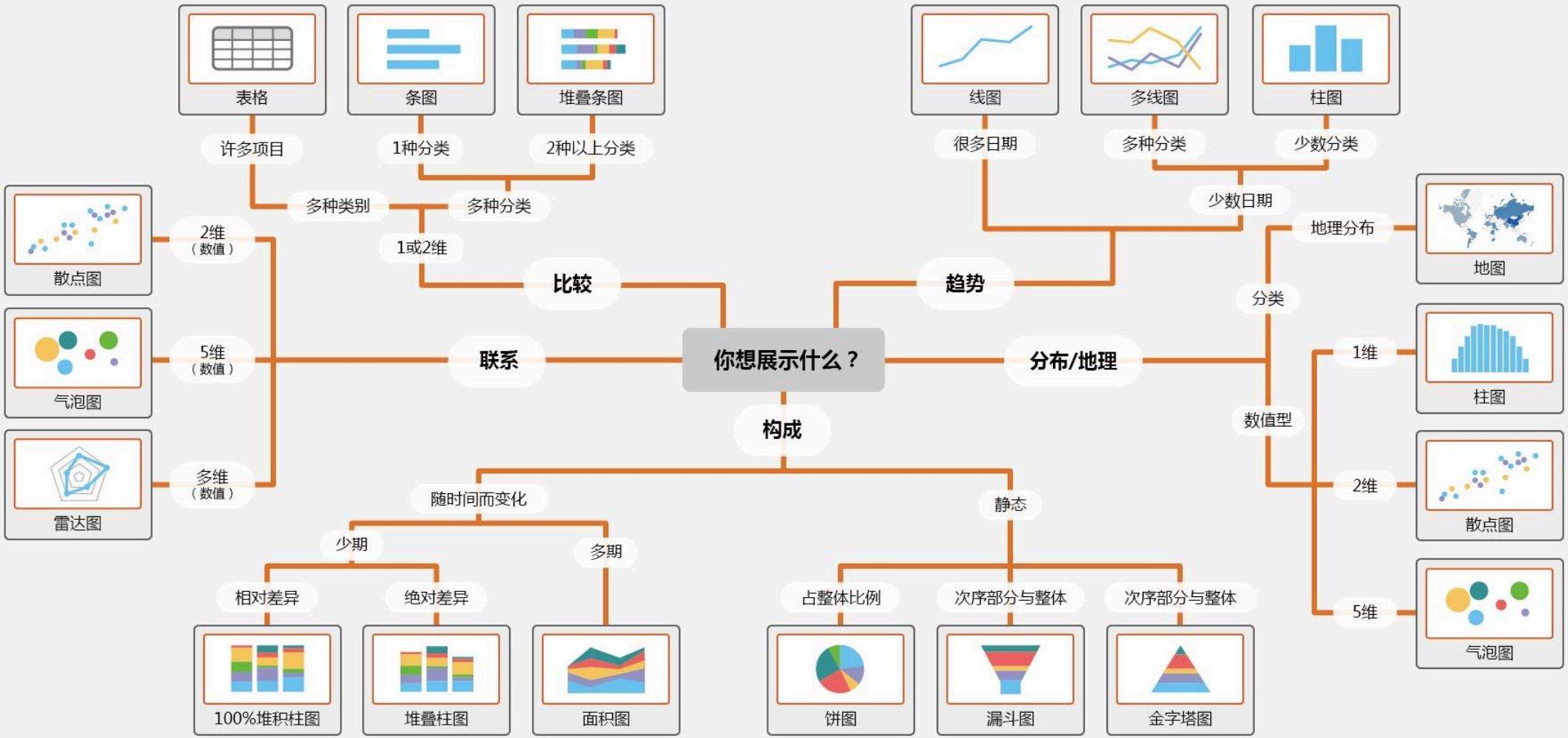
2. 数据可视化

2.0 数据图表简史 video1 video2

> Graphics that help us think faster or see a book's worth of information on a single page are the key to unlocking new discoveries.

> Graphicacy enables us to harness our built-in GPU to process mountains of data and find the veins of gold hiding within.

如何选择图表的类型？


你想展示什么？
比较
联系
构成
趋势
分布/地理
表格
条图
堆叠条图
线图
多线图
柱图
许多项目
1种分类
2种以上分类
很多日期
多种分类
少数分类
多种类别
多种分类
1或2维
少数日期
2维（数值）
5维（数值）
多维（数值）
散点图
气泡图
雷达图
地理分布
分类
地图
1维
柱图
数值型
2维
散点图
5维
气泡图
随时间而变化
静态
少期
多期
相对差异
绝对差异
占整体比例
次序部分与整体
次序部分与整体
100%堆积柱图
堆叠柱图
面积图
饼图
漏斗图
金字塔图

2.1 类别数据可视化

条形图

```
graph bar (mean) numeric_var, over(cat_var)
/*
              y
              |                                  numeric_var must be numeric;
            7 |            +--+                  statistics of it are shown on
              |            |  |                  the y axis.
            5 |  +--+      |  |
              |  |  |      |  |                  cat_var may be numeric or string;
              |  |  |      |  |                  it is shown on the categorical
              |  |  |      |  |                  x axis.
              +------------------------- x
                 First     Second     ...
                 group      group
*/
```

```
graph hbar (mean) numeric_var, over(cat_var)
/*
                        x
                        |
                        |------+
                First group |      |
                        |------+
                        |
                        |----------+
               Second group |          |
                        |----------+
                    .   |
                    .   |
                        +----------------- y
                              5   7
*/
```

same conceptual layout:
numeric_var still appears
on y, cat_var on x

```
graph bar (mean) y, over(division) over(region)
/*
            +-+    +-+    +-+       +-+           +-+
            | |    | |    | |       | |           | |
            | |    | |    | |       | |           | |
           ----------------------------------------------
           div_1  div_2  div_3     div_1  div_2  div_3
                 region_1              region_2
*/
```

```
* Simple bar chart
graph bar (count), over(x)

* Horizontal percentage bar chart
graph hbar (percent), over(x)

* A sorted bar chart
graph hbar (percent), over(x, sort(1) descending)

* Multiple bar charts
graph bar (count), over(x) by(y)

* Multiple bar charts with two or three over() options
graph bar (count), over(x) over(y) over(z)
```

饼图

```
* Basic pie chart
graph pie, over(race)

* Labeling pies
graph pie, over(x) plabel(_all percent)

* Multiple pie charts with by()
graph pie, over(x) by(y)
```

2.2 数据分布可视化

直方图

```
* A basic histogram
histogram tempjan

* Options that change bin sizes
histogram x, bin(10) name(g1, replace)
histogram x, bin(50) name(g2, replace)
histogram x, discrete name(g3, replace)
histogram x, width(10) name(g4, replace)
graph combine g1 g3 g2 g4
```

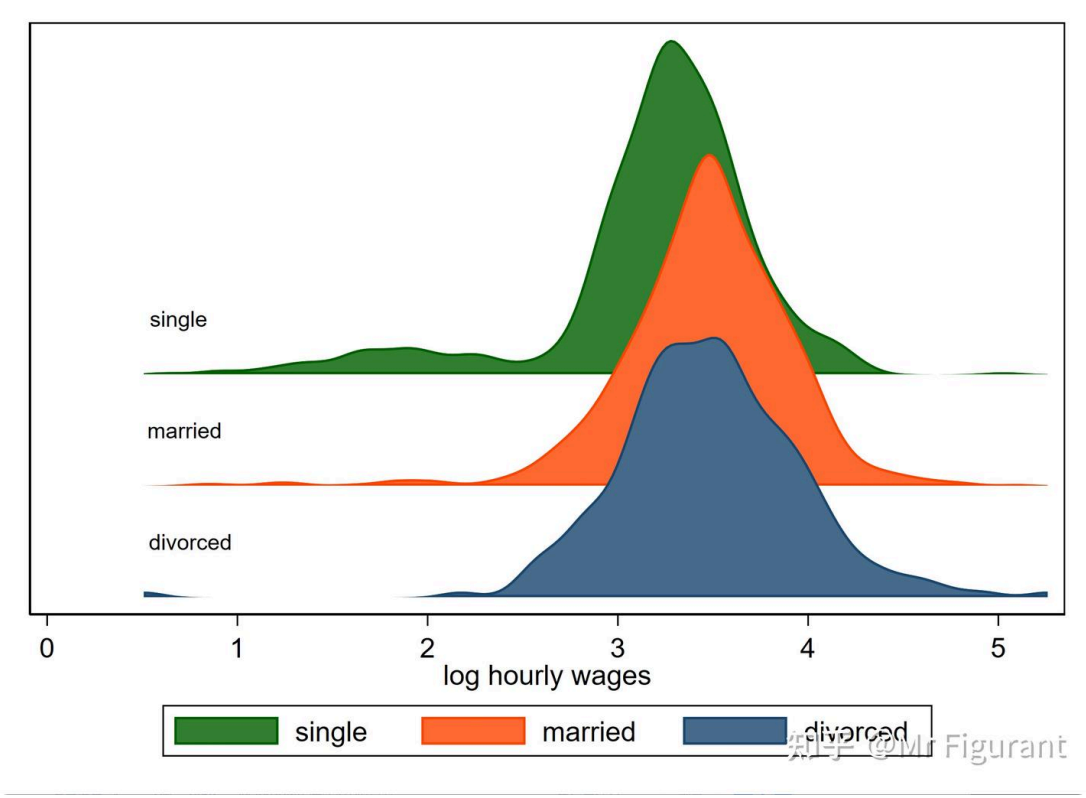
```
* Options that change the y axis
histogram x, density name(g1, replace)
histogram x, fraction name(g2, replace)
histogram x, frequency name(g3, replace)
histogram x, percent name(g4, replace)
graph combine g1 g3 g2 g4

* Adding kernel density plot
histogram x, width(5) kdensity kdenopts(bwidth(5))

* Multiple histograms with by()
histogram x, by(y)

* Multiple two-way histograms
twoway (histogram x if y == 1, color(edkblue%30)) ///
 (histogram x if y == 2, color(erose%30)) ///
 (histogram x if y == 3, color(stone%30)) ///
 (histogram x if y == 4, color(emerald%30))
```

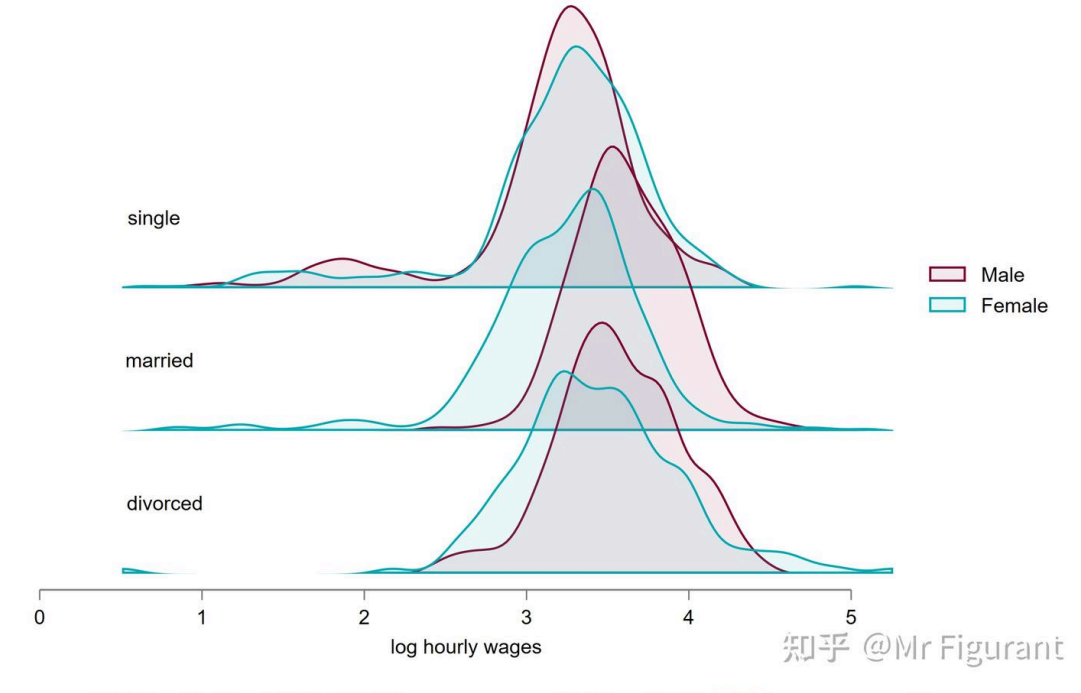
核密度估计图

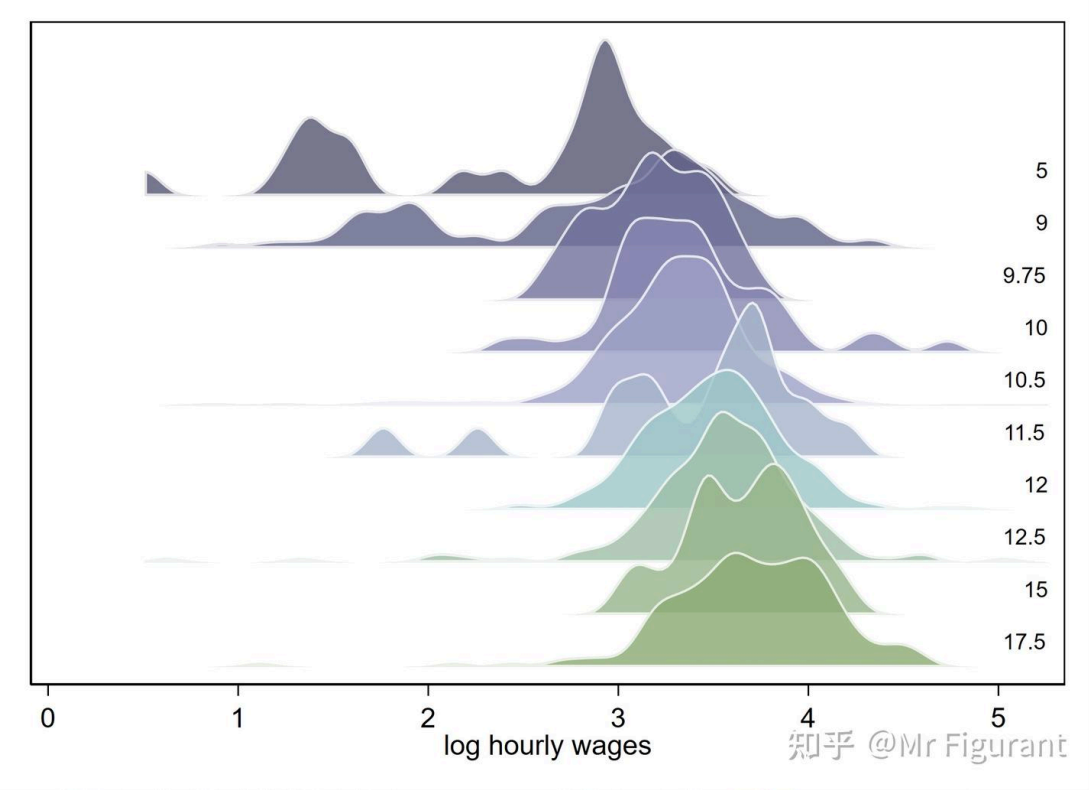
```
* A basic kernel density plot
kdensity x

* Basic visualization options
kdensity x, recast(area) fcolor(eltblue) lcolor(black)

* Density plots with different bandwidths
kdensity x, bwidth(10)

* Multiple kernel density plots
twoway (kdensity x, by(y))

* Multiple kdensity plots with if qualifiers
twoway (kdensity x if y == 1) ///
(kdensity x if y == 2)
```

拓展阅读：joy_plot

- Stata学习：如何绘制Fancy的核密度图？ - 知乎

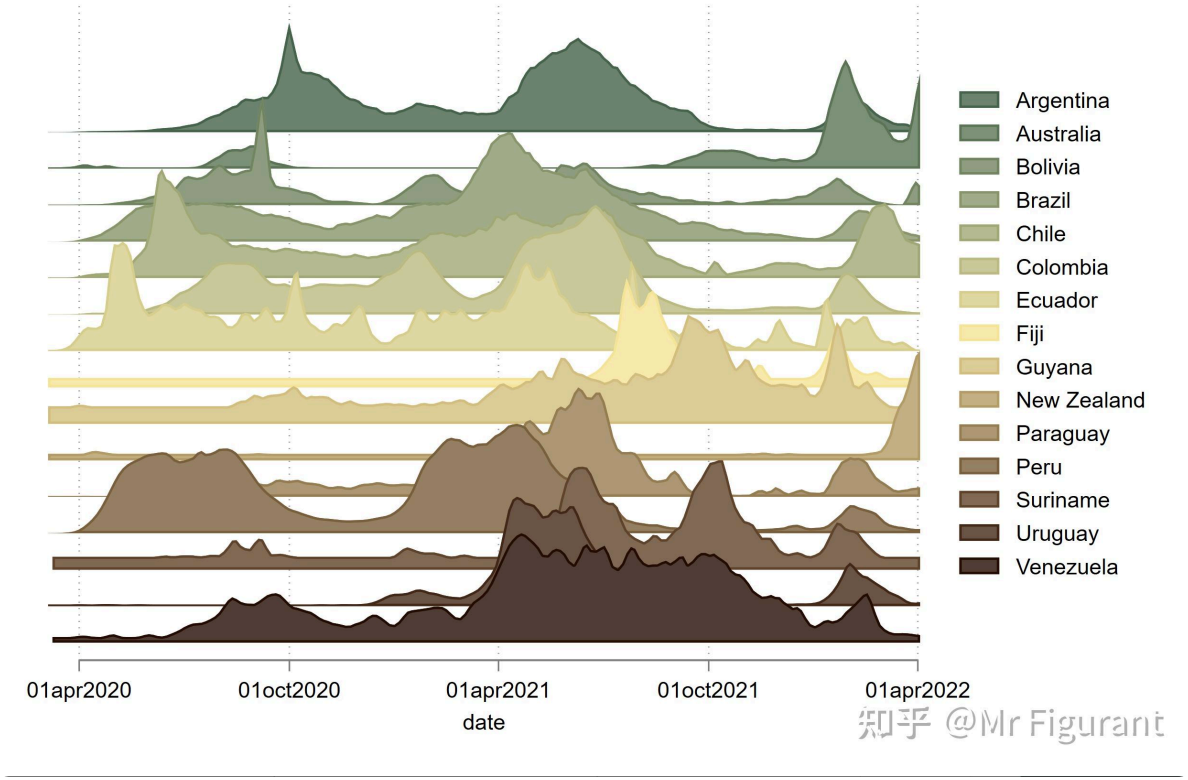


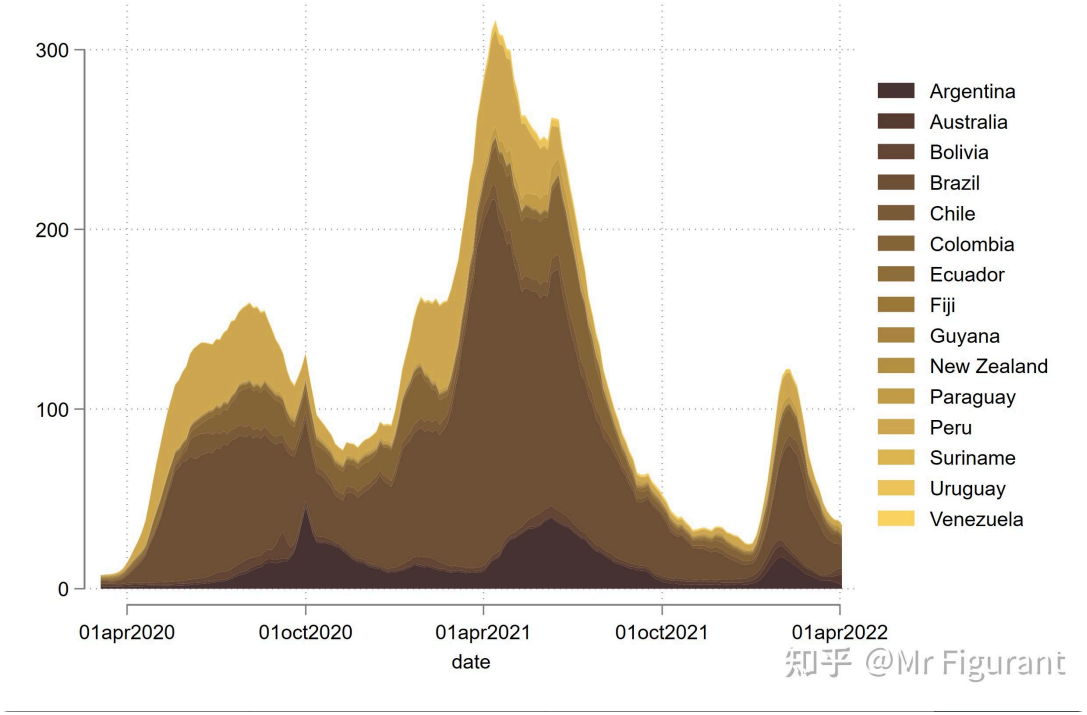

single
married
divorced
Male
Female
0
1
2
3
4
5
log hourly wages
知乎 @Mr Figurant

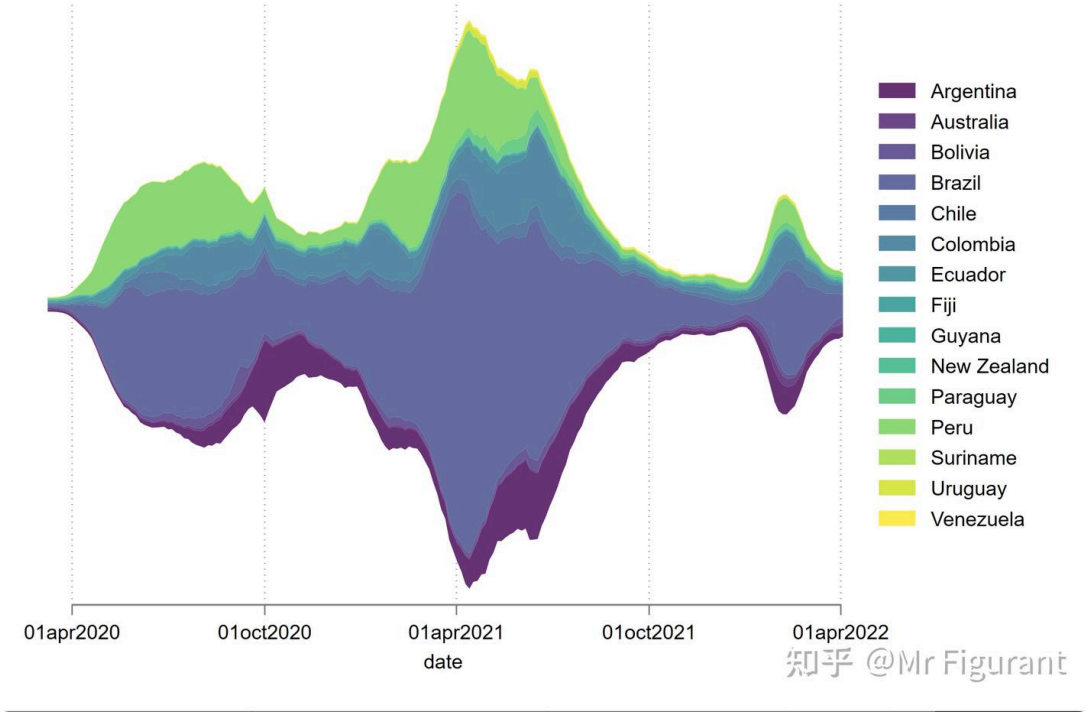

5
9
9.75
10
10.5
11.5
12
12.5
15
17.5
0
1
2
3
4
5
log hourly wages
知乎 @Mr Figurant

拓展阅读：脊线图

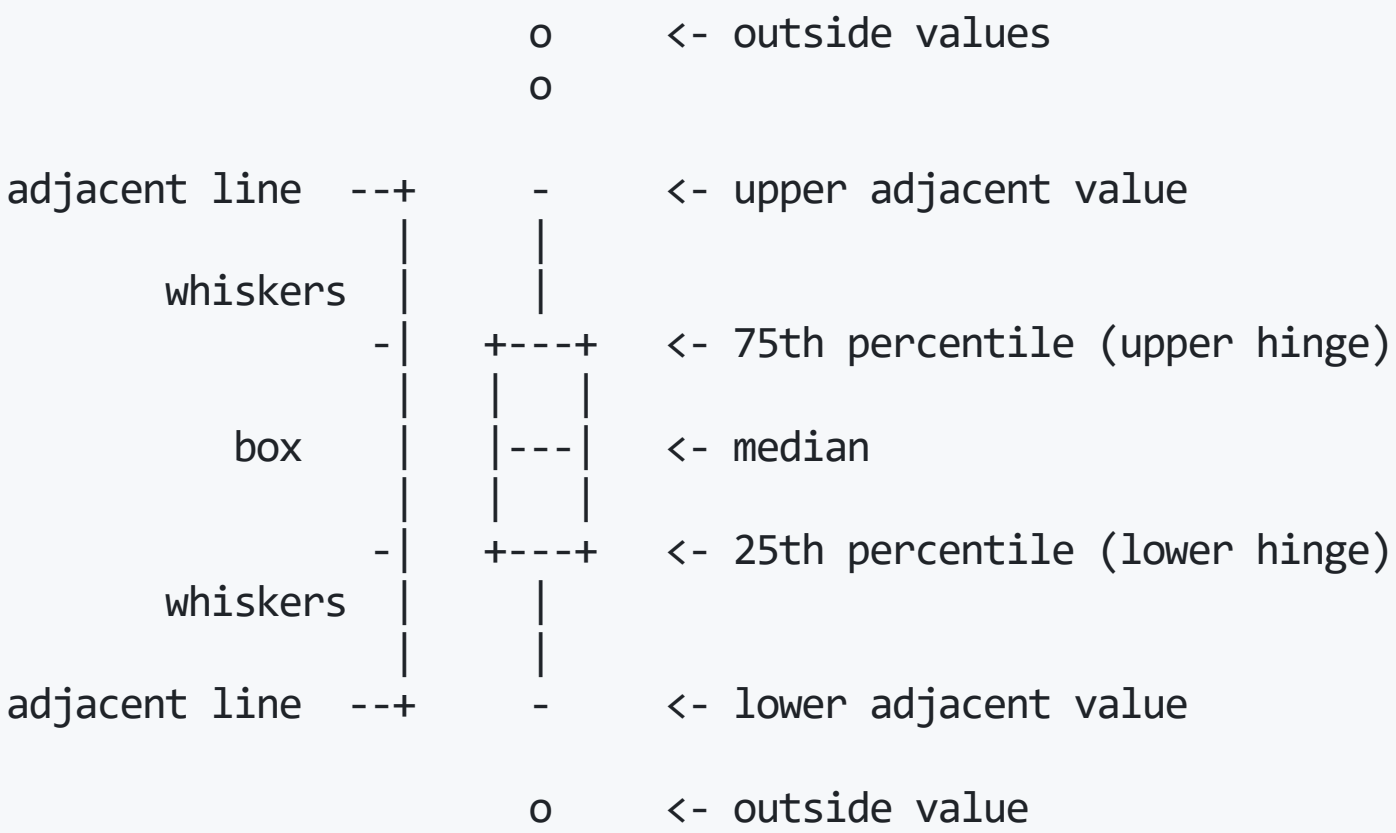
- Stata学习：如何绘制脊线图 ridgeline? - 知乎




300
200
100
0
01apr2020
01oct2020
01apr2021
01oct2021
01apr2022
date
Argentina
Australia
Bolivia
Brazil
Chile
Colombia
Ecuador
Fiji
Guyana
New Zealand
Paraguay
Peru
Suriname
Uruguay
Venezuela
知乎 @Mr Figurant


01apr2020
01oct2020
01apr2021
01oct2021
01apr2022
date
Argentina
Australia
Bolivia
Brazil
Chile
Colombia
Ecuador
Fiji
Guyana
New Zealand
Paraguay
Peru
Suriname
Uruguay
Venezuela
知乎 @Mr Figurant

箱线图


o <- outside values
o
adjacent line --+ - <- upper adjacent value
| |
whiskers | |
-| +---+ <- 75th percentile (upper hinge)
| | |
box | |---| <- median
| | |
-| +---+ <- 25th percentile (lower hinge)
whiskers | |
| |
adjacent line --+ - <- lower adjacent value
o <- outside value

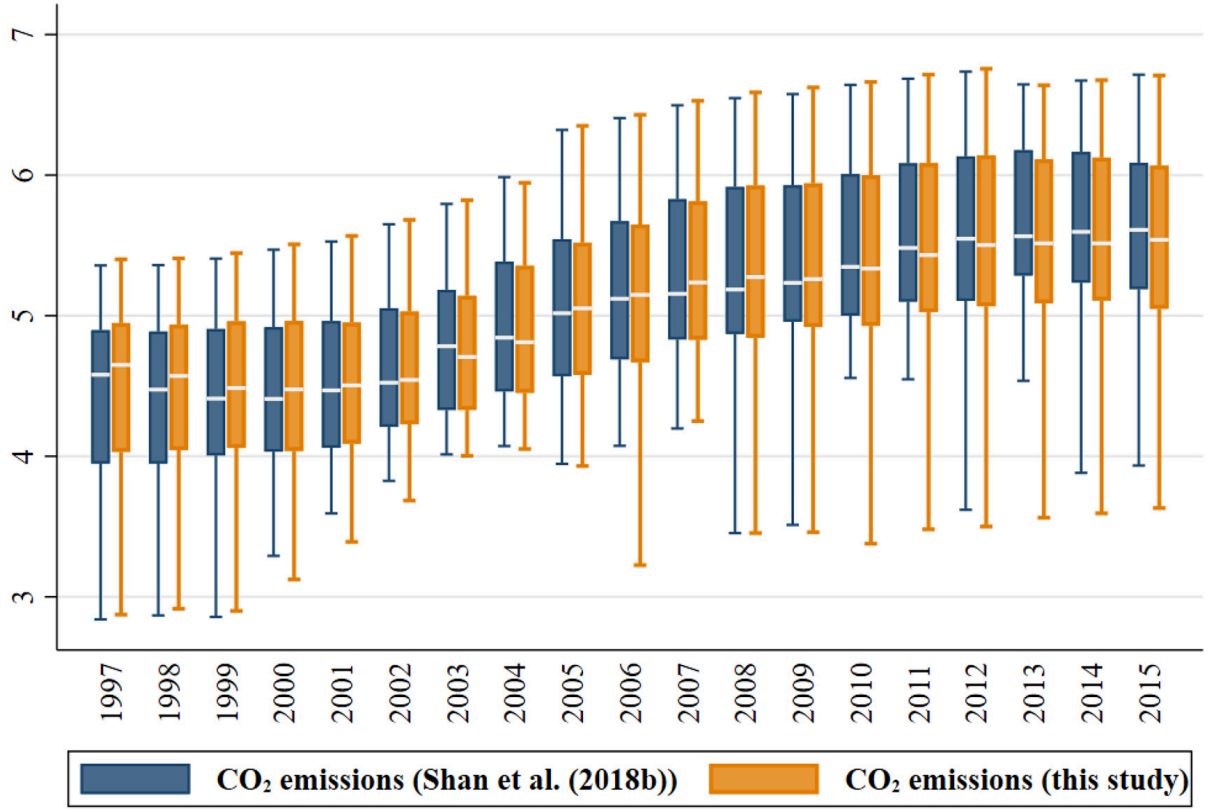
```
graph box y1 y2, over(cat_var)
/*
                    y
                  8 |
                    |              -              o
                    |     o       |              -          y1, y2 must be numeric;
                  6 |           +-+      -       |          statistics are shown on
                    |      -    | |     +-+     +-+         the y axis
                    |    +-+    |-|     |-|     | |
                  4 |    |-|    +-+     +-+     |-|         cat_var may be numeric
                    |    +-+     |       -     +-+          or string; it is shown
                    |     -      -               |          on categorical x axis
                  2 |                            o     -
                    +------------------------------ x
                          First          Second
                          group           group
*/
```

```
graph hbox y1 y2, over(cat_var)
/*
                        x
                        |     +---+
                        |  |--| | |--|  o
                First   |     +---+
                group   |
                        |         +------+
                        |  |-----| |    |-----|       same conceptual layout
                        |         +------+            as for graph box:
                        |
                        |                             y1, y2 appear on y axis
                        |         +---+
                        | o  |--| | |--|              cat_var appears on x axis
                Second  |         +---+
                group   |
                        |         +------+
                        | |-----| |    |-----|o
                        |         +------+
                        |
                        +-------------------------- y
                         2      4      6      8
```

```
* Basic box plot
graph box x

* Basic box plot with outliers
graph box x if y == 1

* boxlook_options
graph box x, medtype(marker) box(1, fcolor(erose)) ///
 cwhiskers lines(lpattern(dash))

* Horizontal box plot
graph hbox x

* Histogram with box plot
graph hbox x, name(g1, replace)
histogram y, name(g2, replace) ytitle("")
graph combine g1 g2, cols(1) xcommon
```

```
* Multiple box plots
graph box x y

* Multiple box plots over a categorical variable
graph box x y, over(z)

* Multiple box plots with ascategory option
graph hbox x y, over(z) scale(0.8) ascategory

* Ordering boxes
graph hbox x, over(z, sort(1))
```

论文阅读



Chen Y, Shao S, Fan M, et al. One man's loss is another's gain: does clean energy development reduce CO2 emissions in China? Evidence based on the spatial Durbin model[J]. Energy Economics, 2022, 107: 105852.

小提琴图

```
ssc install vioplot

* A violin plot
vioplot x

* Violin plot with kernel() option
vioplot x, kernel(gaussian) bwidth(10)

* A horizontal violin plot
vioplot x, horizontal

* Multiple violin plots
vioplot x y, over(z)

* Multiple violin plots with custom widths
vioplot x y, over(z) barwidth(500) dscale(200)
```

自学

点阵图 `dotplot`

茎叶图 `stem`

区分 graph dot 和 dotplot

```
graph dot (mean) numeric_var, over(cat_var)
/*
                        x axis
                          |
                Group 1   |..........o..............
                Group 2   |..............o..........
                Group 3   |..................o.......   <-  lines
                Group 4   |......................o...
                          +-------------------------
                           0    1    2    3    4    5   <-  y axis

                                        marker
                                       /
                                    .............o............
                                    |--distance--|       \
                                                          dots
*/
```

```
graph hbar y, over(group)
/*
                        |---+
                Group 1 |   |
                        |---+
                        |
                        |--------+
                Group 2 |        |
                        |--------+
                        |
                        |--------------+
                Group 3 |              |
                        |--------------+
                        |
                       +----------------
*/
```

```
graph dot y, over(group)
/*
    Group 1 |...o.............
    Group 2 |........o.......
    Group 3 |..............o.
            +-----------------
*/
```

```
graph hbar y1 y2, over(group)
/*
                        |-+
                        | |
                Group 1 |---+
                        |   |
                        |---+
                        |
                        |---+
                        |   |
                Group 2 |-------+
                        |       |
                        |-------+
                        |
                        |-----+
                        |     |
                Group 3 |-------------+
                        |             |
                        |-------------+
                        |
                       +----------------
*/
```

```
graph dot y1 y2, over(group)
/*
    Group 1 |.x.o.............
    Group 2 |...x....o.......
    Group 3 |.....x........o.
            +----------------
*/
```

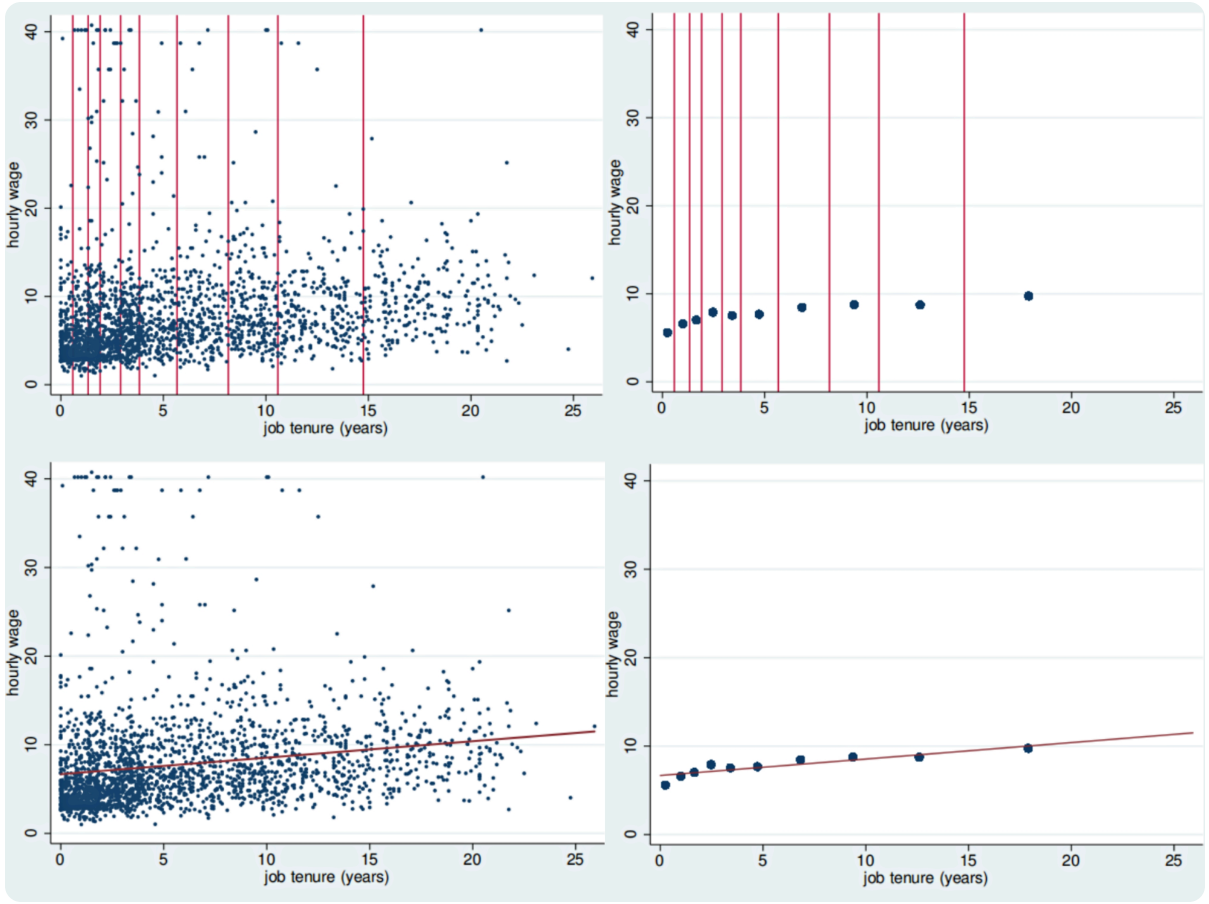
2.3 变量间关系可视化

散点图

```
* Basic scatterplot
twoway (scatter y x)

* Square, colored, and sized markers
twoway (scatter y x in 1/15, msymbol(S) mcolor(edkblue%70) msize(*3) mlabel(make))

* Multiple y variables against one x variable
twoway (scatter x1 x2 x3 y, msymbol(S T D) mcolor(gs1 gs6 gs12))

* Advanced syntax, two axes
twoway (scatter y x, xaxis(1) yaxis(1 2)) ///
(scatter y z, xaxis(2))
```

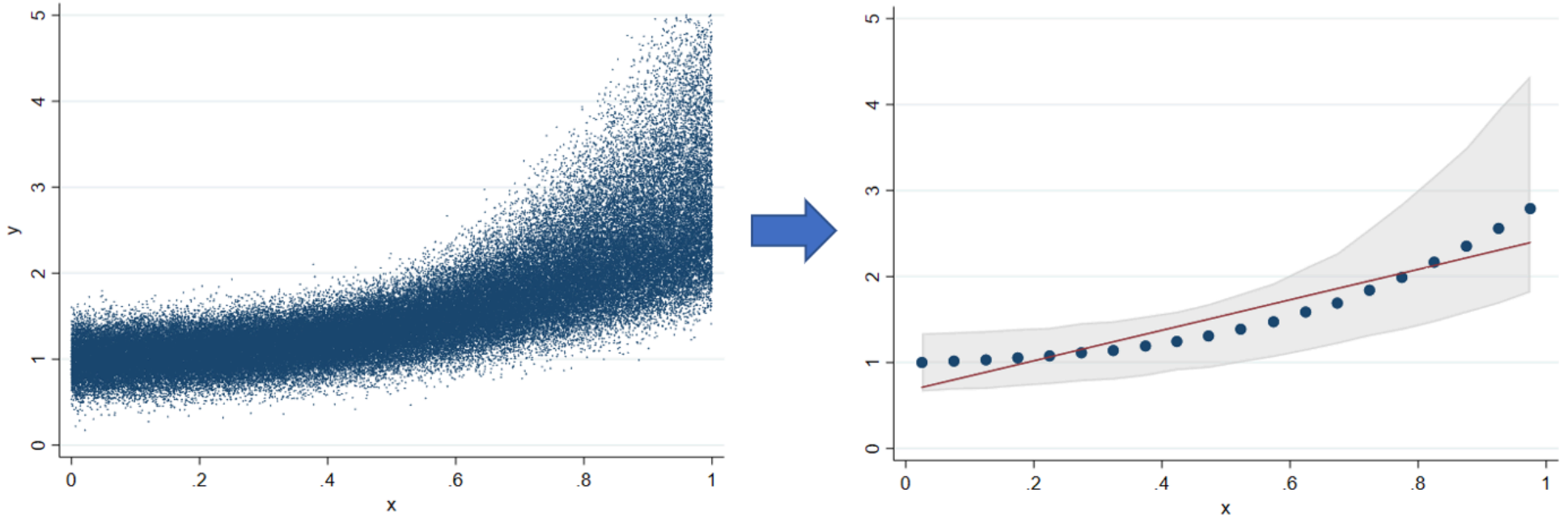
拓展阅读：binscatter

- Stata：分仓散点图绘制-binscatter-binscatter2



拓展阅读：binscatter2

- Stata：分仓散点图绘制-binscatter-binscatter2



相关系数矩阵与散点图矩阵

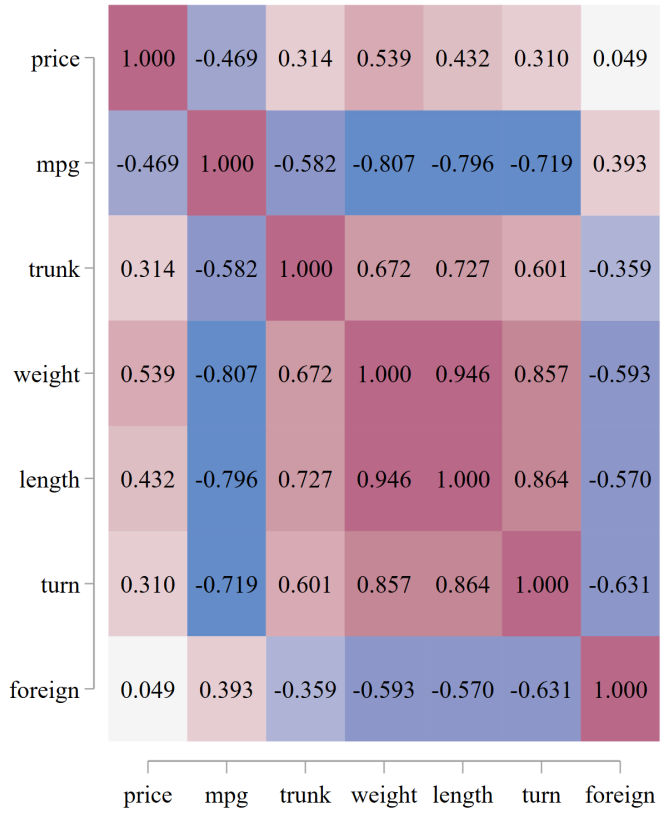
```
*Correlation matrix
correlate v1 v2 v3 v4 v5

* Matrix scatterplot
graph matrix v1 v2 v3 v4 v5

* Half-scatter matrix plot with options
graph v1 v2 v3 v4 v5, mlwidth(vvthin) mfcolor(ltblue%80) msize(medsmall) half scale(0.7)
```

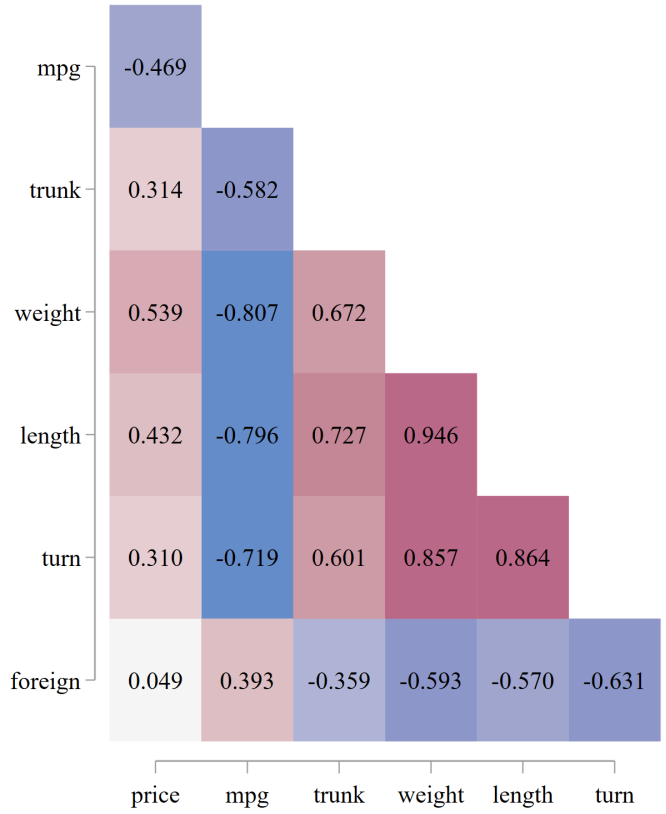
```
graph matrix v1 v2 v3 v4 v5, ylabel(0(5)40, axis(1))
/*
                          x                x
                       axis(2)          axis(4)
                  +---------------------------------------+
                  |       | v1/v2 | v1/v3 | v1/v4 | v1/v5 |  y axis(1)
                  |-------+-------+-------+-------+-------|
          y axis(2) | v2/v1 |       | v2/v3 | v2/v4 | v2/v5 |
                  |-------+-------+-------+-------+-------|
                  | v3/v1 | v3/v2 |       | v3/v4 | v3/v5 |  y axis(3)
                  |-------+-------+-------+-------+-------|
          y axis(4) | v4/v1 | v4/v2 | v4/v3 |       | v4/v5 |
                  |-------+-------+-------+-------+-------|
                  | v5/v1 | v5/v2 | v5/v3 | v5/v4 |       |  y axis(5)
                  +---------------------------------------+
                     x               x               x
                   axis(1)         axis(3)         axis(5)
*/
```

```
graph matrix v1 v2 v3 v4 v5, half
/*
                    +-------+
          y axis(1) |       |
                    |-------+-------+
          y axis(2) | v2/v1 |       |
                    |-------+-------+-------+
          y axis(3) | v3/v1 | v3/v2 |       |
                    |-------+-------+-------+-------+
          y axis(4) | v4/v1 | v4/v2 | v4/v3 |       |
                    |-------+-------+-------+-------+-------+
          y axis(5) | v5/v1 | v5/v2 | v5/v3 | v5/v4 |       |
                    +---------------------------------------+
                       x       x       x       x       x
                     axis(1) axis(2) axis(3) axis(4) axis(5)
*/
```

拓展：用热力图可视化相关系数矩阵

- Stata 可视化：heatplot-热力图





气泡图

```
* A basic bubble plot
twoway (scatter y x [fweight = pop])
```

雷达图

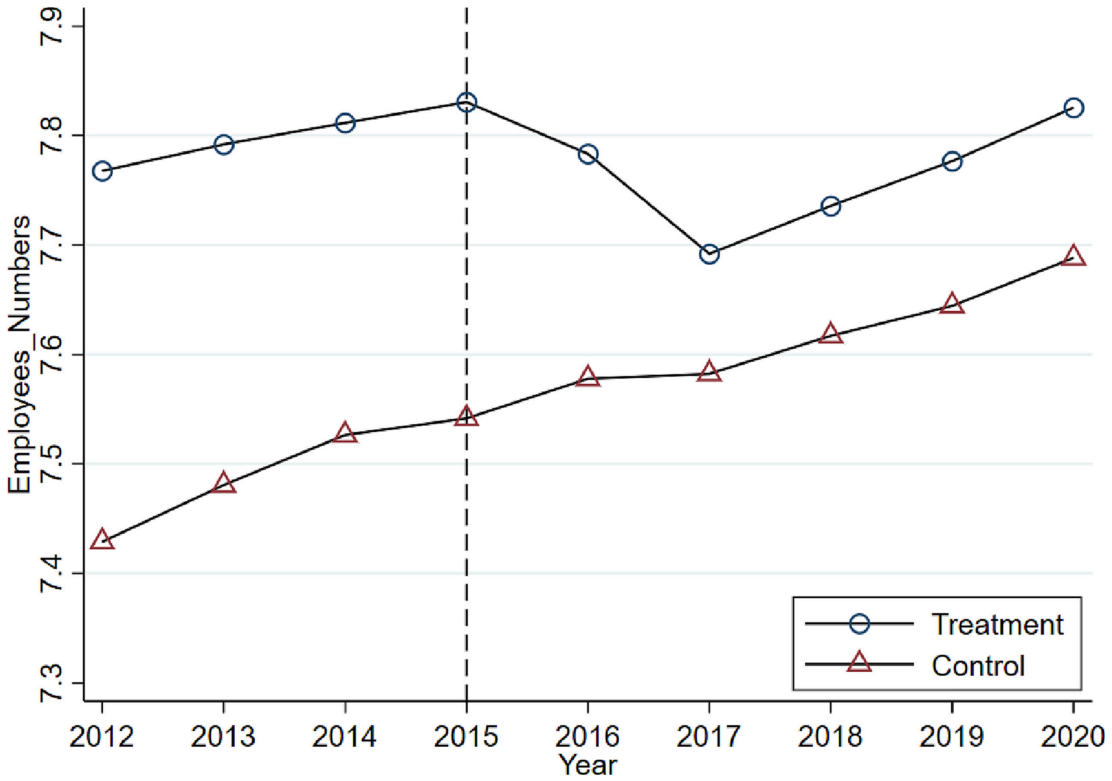
```
ssc install radar

* Radar chart with aspect ratio
radar z x in 1/12, aspect(1) graphregion(color(white)) xlabel(, nogrid)

* Radar charts with multiple variables
radar z x1 x2 x3, rlabel(0 10 20 30 40 50) aspect(1) legend(position(6) row(1)) xlabel(, nogrid)
```

2.4 时间序列可视化

折线图

```
* A basic line plot
line x year

* Multiple line plots via line
line x1 x2 year, lcolor(gs12 gs4)

* Multiple line plots via twoway
twoway (line x1 year, lcolor(gs12)) (line x2 year, lcolor(gs4))

* Line plots with markers
twoway (line x year) (scatter x year, msymbol(D) msize(*0.5) mcolor(%50))
```

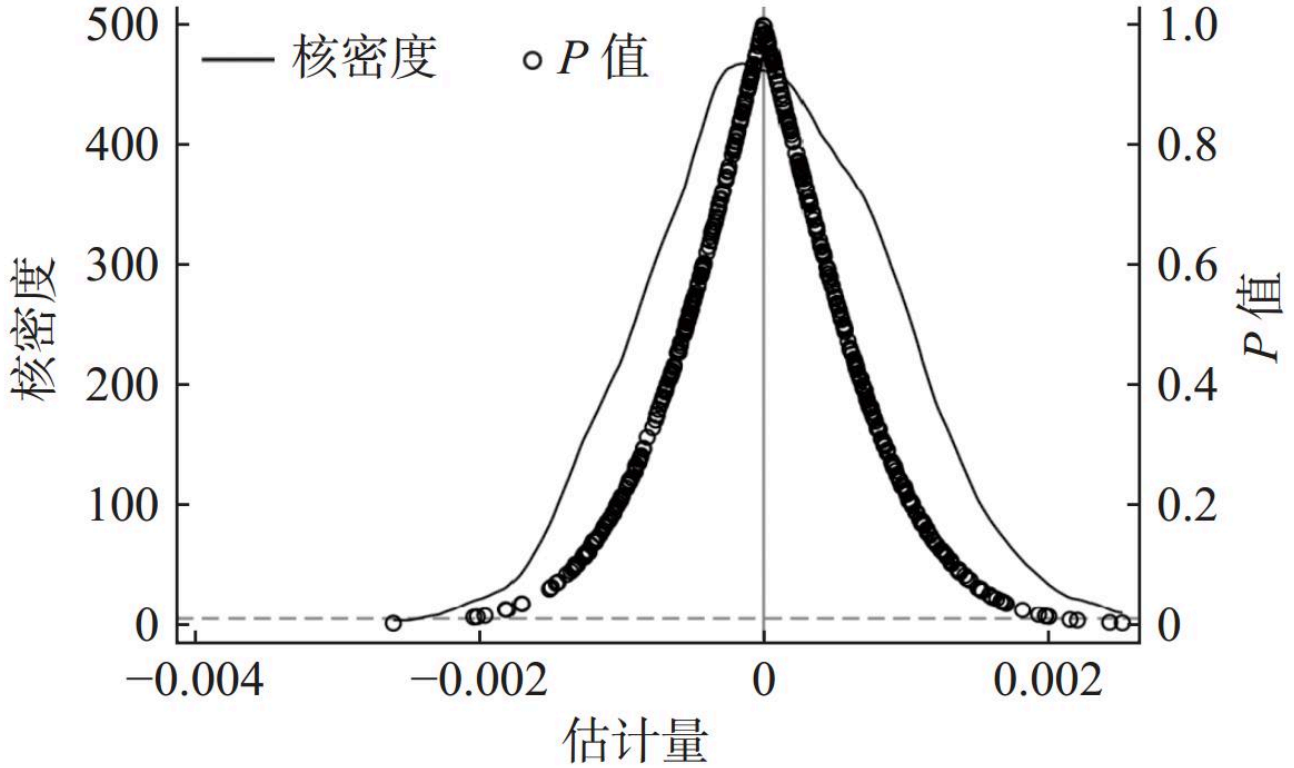
论文阅读



Liao T, Liu G, Liu Y, et al. Environmental regulation and corporate employment revisited: New quasi-natural experimental evidence from China's new environmental protection law[J]. Energy Economics, 2023, 124: 106802.

安慰剂检验——散点图+核密度图

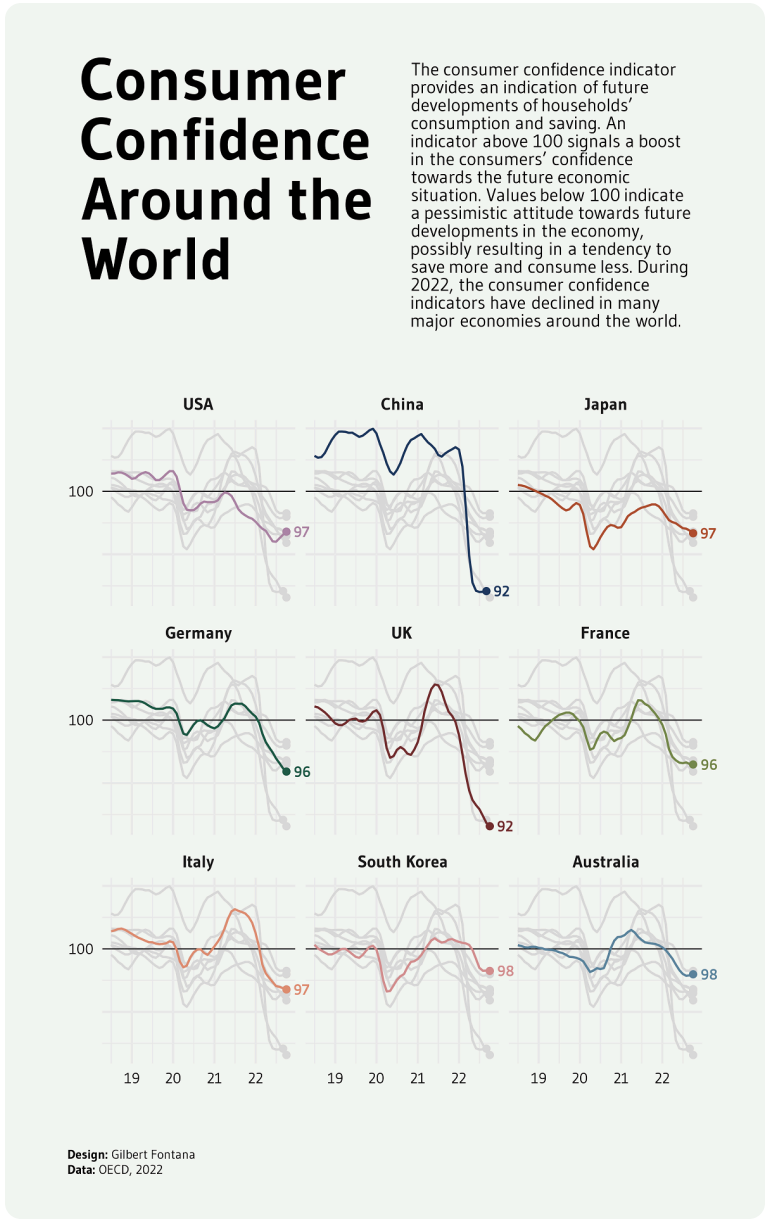
- 平行趋势检验：DID的前提
- 安慰剂检验：DID是否稳健



2.5 图形叠加：twoway，addplot

- grunfeld.dta，

```
use "grunfeld.dta", clear

gen lnInv = ln(invest)
gen lnMV = ln(mvalue)

local y "lnInv"
local y "lnMV"
twoway line `y' time if company==3, ///
  lw(*2) lc(black) lp(solid)
foreach i of numlist 1 2 4/10{
  addplot: line `y' time if company==`i', ///
  lc(black*0.6%50) lp(solid) legend(off)
}
```



3. 描述性统计

3.1 描述水平的统计量

```
sysuse auto
sum price
/*
    Variable |        Obs        Mean    Std. dev.       Min        Max
-------------+---------------------------------------------------------
       price |         74    6165.257    2949.496       3291      15906
*/
```

```
sum price, detail
/*
                            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
*/
```

简单平均数

```
tabstat x, stats(mean) //tabstat x 默认是mean
```

加权平均数

```
gen xbar = m1 if x < x1 //m1为组中值
replace xbar = m2 if x < x2 & x >= x1
replace xbar = m3 if x < x3 & x >= x2
replace xbar = m4 if x < x4 & x >= x3
tabstat xbar, stats(mean)
```

中位数

```
tabstat x, stats(median)
```

四分位数

```
tabstat x, stats(q)
```

百分位数

```
tabstat x, stats(p10)
```

频数和众数

```
tabulate rep78
/*
     Repair |
record 1978 |      Freq.     Percent        Cum.
------------+-----------------------------------
          1 |          2        2.90        2.90
          2 |          8       11.59       14.49
          3 |         30       43.48       57.97
          4 |         18       26.09       84.06
          5 |         11       15.94      100.00
------------+-----------------------------------
      Total |         69      100.00
*/
```

- N = 74?

3.2 描述差异的统计量

极差

```
tabstat x, stats(range)
```

四分位差

```
tabstat x, stats(iqr)
```

方差

```
tabstat x, stats(var)
```

标准差

```
tabstat x, stats(sd)
```

变异系数/离散系数

```
tabstat x, stats(cv)
```

异常值

```
sum x //返回r类命令
gen z = (x - r(mean)) / r(sd)
list x z if z > 3
drop z if z > 3
```

拓展：r类命令、e类命令

e类命令：估计（estimation）命令，比如regress，通过e()调用

r类命令：其他命令，比如sum x，通过r()调用

3.3 描述分布形状的统计量

偏度系数

```
tabstat x, stats(skewness)
```

峰度系数

```
tabstat x, stats(kurtosis)
```

3.4 数据标准化

中心化

```
qui sum x
gen c_x = x - r(mean)
```

标准化

```
qui sum x
gen z_x = (x - r(mean)) / r(sd)
```

标准化2

```
egen z_x = std(x)
```

归一化

```
qui sum x
gen n_x = (x - r(min)) / (r(max) - r(min))
```

拓展：center（外部命令）

```
center x, prefix(c_) # 可加standardize, 简写至s
```

3.5 正态性检验

直方图图示

```
hist x, norm
```

正态分位图

```
qnorm x, grid //add grid lines
```

正态性统计检验

```
sktest x
```

幂阶梯转换

```
ladder x
qladder x //散点图矩阵
gladder x //直方图矩阵
```

4. 课后习题

习题1（数据可视化）

本题使用exercise3_3.csv（请在pydata文件夹中寻找）

数据集来自R语言的mtcars，摘自1974年《美国汽车趋势》杂志，包括32款汽车（1973-1974年款）的油耗、汽车设计和性能等共11个变量。根据该数据集绘制以下图形。

(1)绘制气缸数量(cyl)的直方图，并为直方图叠加核密度曲线。

(2)绘制每加仑油行驶的英里数(mpg)和汽车自重(wt)两个变量的箱线图和小提琴图。

(3)绘制该数据集的散点图矩阵和相关系数矩阵图。

(4)绘制每加仑油行驶的英里数(mpg)、总马力(hp)和汽车自重(wt)3个变量的3D散点图和气泡图。

习题2（描述性统计）

本题使用exercise4_1.csv（请在rdata或pydata文件夹中寻找）

随机抽取50个网络购物的消费者，调查他们某月的网购金额（单位:元）。

(1)计算平均数、标准差、极差和四分位差。

(2)计算10%、50%、75%的分位数。

(3)计算偏度系数和峰度系数，分析网购金额的分布特征。

(4)计算标准分数和极值标准化值，检测数据的离群点。

习题3（选做）

你还知道哪些数据可视化图表，它们在Stata里如何实现？
收集并整理代码，解释图形含义、优缺点、适用场景与案例等。

下面的网站可以帮助你在种类繁多的图表中选择合适的一种：

- 图之典（提供matplotlib和ggplot2代码）
- Chart.Guide - helps you choose and design the Perfect Chart
- Data Viz Project | Collection of data visualizations to get inspired and find the right type
- The Data Visualisation Catalogue

作业要求

- 提交代码+对应图形+必要解释，注意标明题号
- 整理成一个pdf文档，命名为学号-姓名如“2410000-张三”
- **DDL**：10月16日24点前发送至助教飞书（2120253538）
- 视代码完整性、作图美观性、答案准确性评分
- 逾期提交、抄袭雷同记0分
- 可以使用AI

欢迎交流~

zzynankai@outlook.com

Bilibili: 西山yu

xishanyu2.github.io