DATA VISUALISATION, DATA ANALYSIS, AUTOMATION OF WORKFLOW

Kausik Chaudhuri

Thursday 12th April

University of Leeds

DATA VISUALISATION

WHY DATA VISUALISATION IS IMPORTANT?

Helps us inspect raw data, identify patterns in the data, understand distributions

Develop hypotheses about our data.

Helps to communicate **key points** more clearly with a graph than a table

We emphaszie on:

- Histogram
- Scatterplots
- Lineplots
- Bar charts

HISTOGRAMS

PURPOSE OF HISTOGRAMS

Histograms help to understand:

- how variables in your dataset are distributed?
- are distributions skewed? are there outliers?
- is there a lot of variance in the data?
- how do two distributions compare to each other?

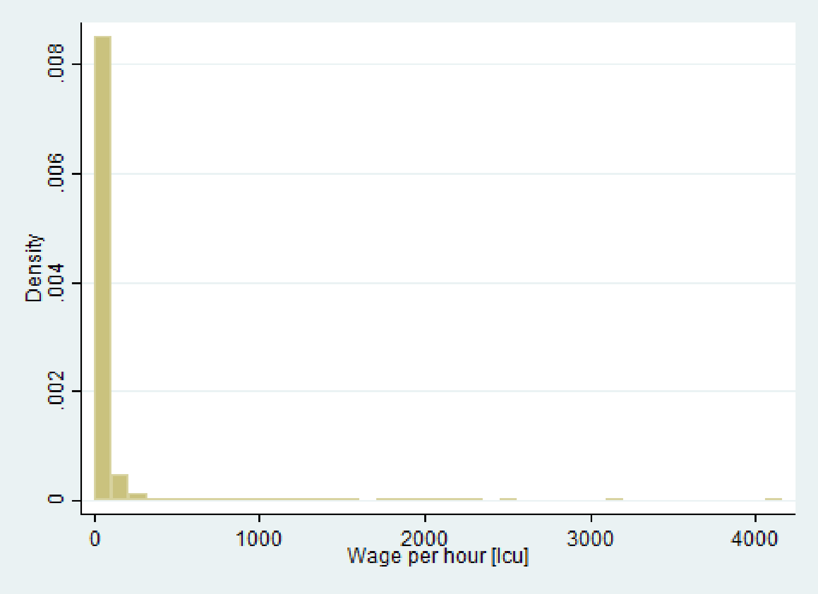
MORE ON HISTOGRAM

CONSTRUCTION OF HISTOGRAM

- estimate of the probability distribution of a continuous variable **(quantitative variable)**
- Step 1: divide the entire range of values into a series of intervals
- Step 2: count how many values fall into each interval
- Step 3: the bins are usually specified as consecutive, non-overlapping intervals of a variable and the bins (intervals) must be adjacent,
- Step 4: bins are often (but are not required to be) of equal size

- We use look at the distribution of hourly wages in Kenya (Wage per hour in 2005 Kenyan Shillings)).
- Plot of histogram: command is: tw histogram variable name

EXAMPLE OF HISTOGRAM

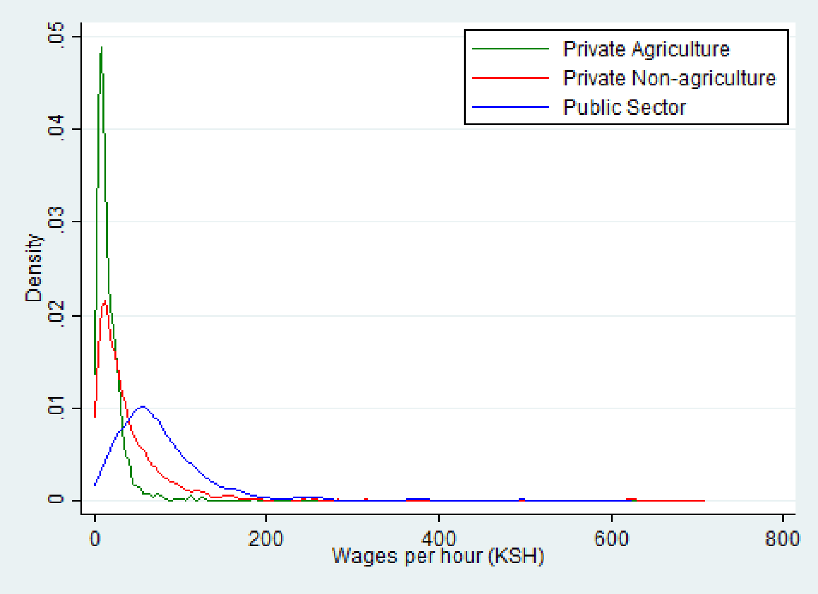


EXAMPLE OF HISTOGRAM

- Use of the bin option by increasing the number of bars drawn enables to see finer changes in the distribution.
- For example, tw histogram variable name, bin(100)
- In stata For example, tw histogram variable name, bin(100)
- Another way to represent distributions is by using the **kernel density (command is kdensity) function**
- Use of kdensity helps to compare different distributions in the same graph (for example wages in different types of employment)
- With several series, STATA automatically adds a legend to differentiate between them
- But we can create a better legend manually
- We can also add titles to the x and y axes

KERNEL DENSITY

KERNEL DENSITY EXAMPLE



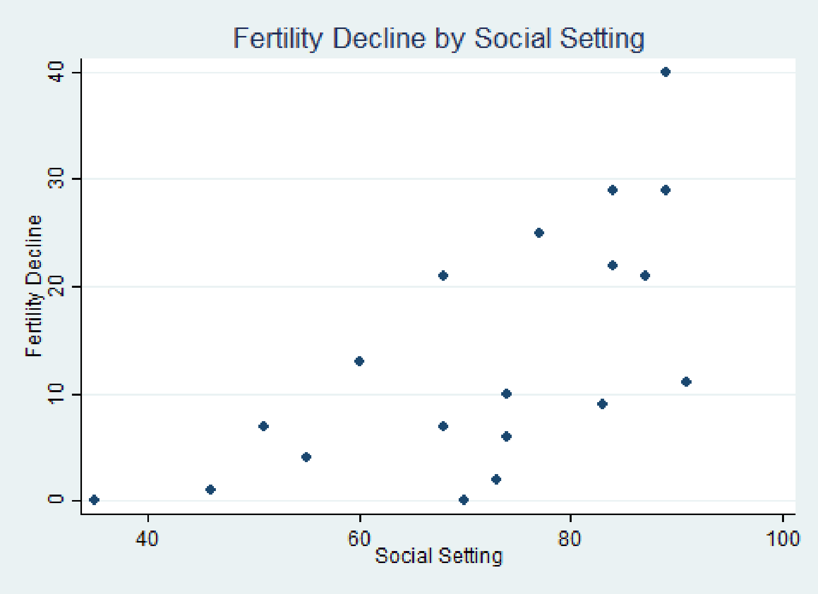
SCATTER DIAGRAMS

SCATTER DIAGRAM

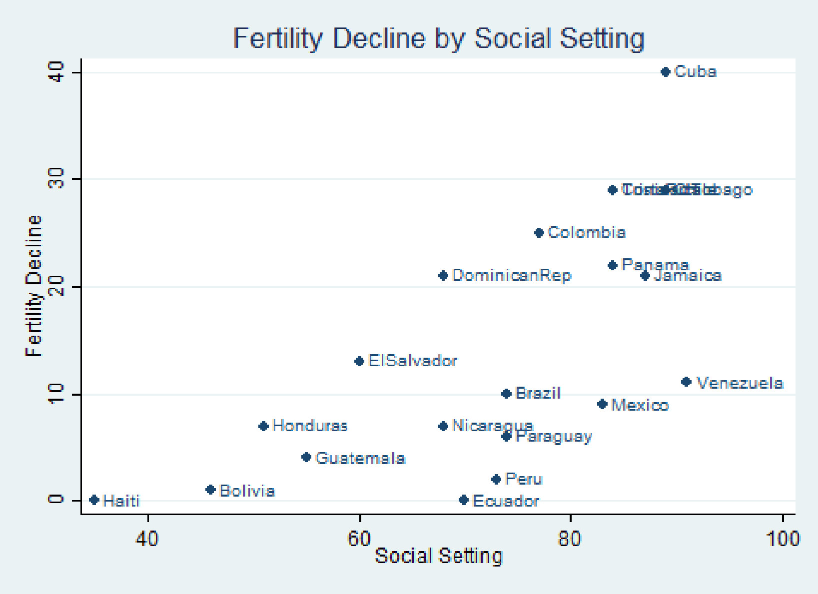
- Scatter diagram plots pairs of numerical data, with one variable on each axis, to look for a relationship between them.
- If the variables are correlated, the points will fall along a line or curve.
- For example, Variable A is the number of employees trained on new software, and variable B is the number of calls to the computer help line. You suspect that more training reduces the number of calls. Plot number of people trained versus number of calls.
- Even if the scatter diagram shows a relationship,we should not think that one variable caused the other. Both may be influenced by a third variable.

- When the data are plotted, the more the diagram resembles a straight line, the stronger the relationship.
- If the scatter diagram shows no relationship between the variables, consider whether the data might be stratified.
- If the diagram shows no relationship, consider whether the independent (x-axis) variable has been varied widely.
- **Drawing a scatter diagram is the first step in looking for a relationship between variables**.

SCATTER PLOT EXAMPLE

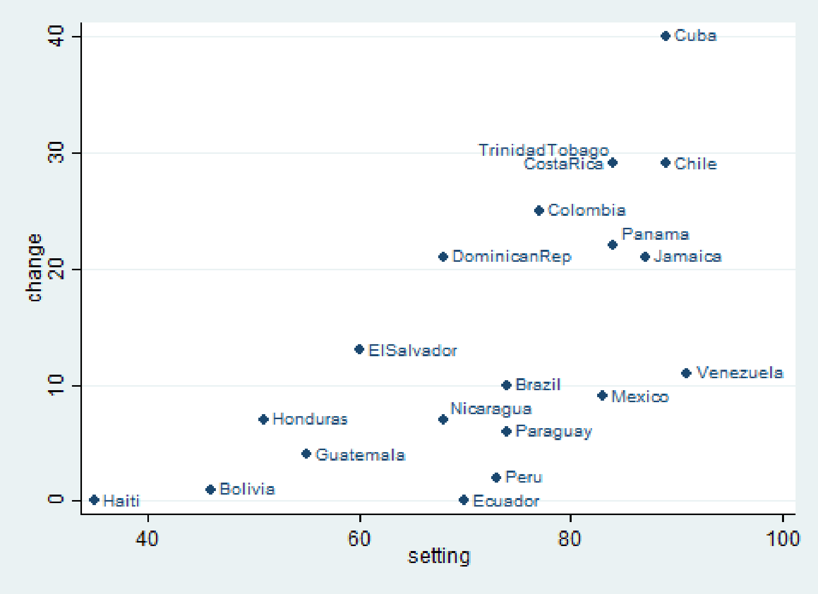


SCATTER DIAGRAM WITH COUNTRY NAME



- Problem with the labels is the overlap of Costa Rica and Trinidad Tobago.
- We can solve this problem by specifying the position of the label relative to the marker using a 12-hour clock.

SCATTER DIAGRAM WITH COUNTRY NAME

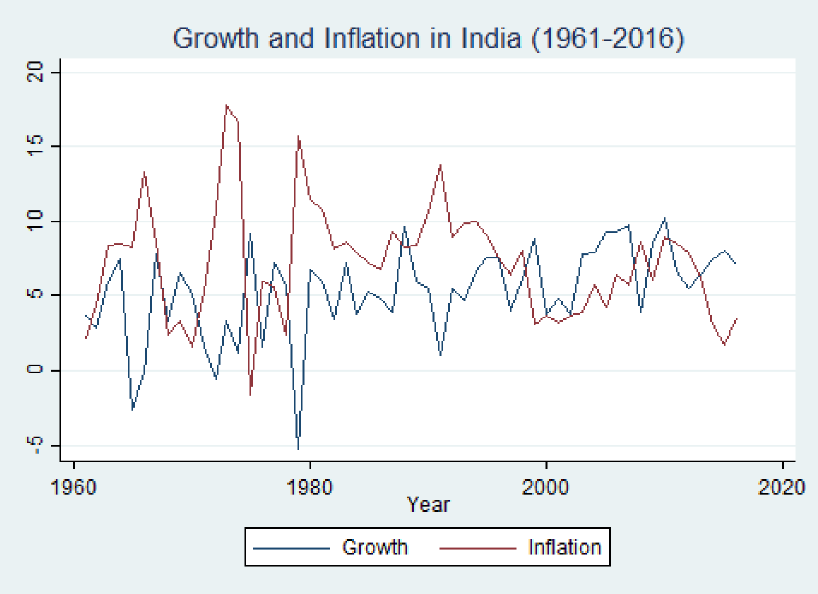


LINE PLOT

LINE PLOT

- A line plot is a graph that shows frequency of data along a number line.
- It is best to use a line plot when comparing fewer than 25 numbers.
- It is a quick, simple way to organize data.

LINE PLOT EXAMPLE

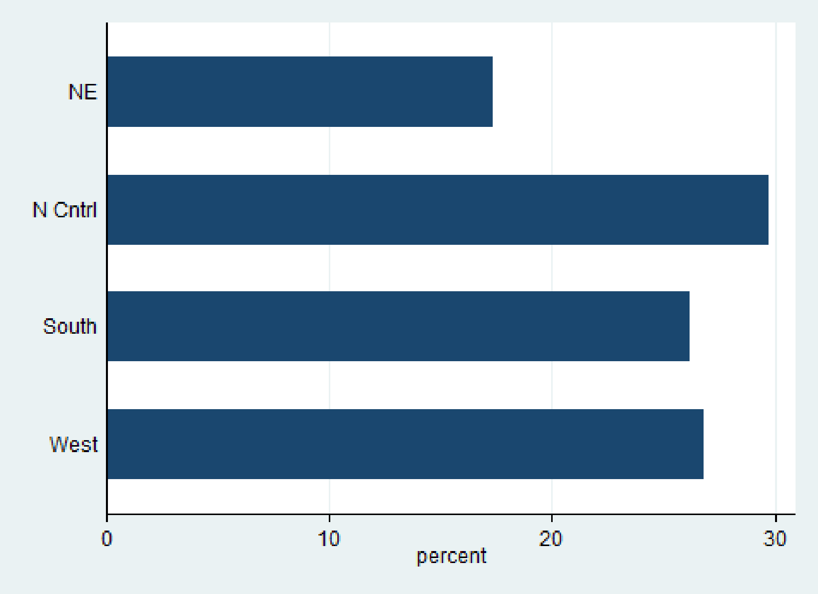


BAR PLOT

BAR PLOT

- Bar graphs may be used to plot the frequency distribution of a categorical variable, or to plot descriptive statistics of a continuous variable within groups defined by a categorical variables.
- The first step is to draw the basic bar graph – achieved using tw bar where tw refers to the twoway class of graphs, bar refers to bar chart.
- The first argument is the variable that contains the values that determine the height of the bars, and the second argument, the variable that contains the value of your categories (when we use twoway graphics, this variable needs to be numeric).
- In case of numeric, we want the bars to be horizontal (a good option when your categories, are long string variables).
- We use the horizontal option to achieve this.

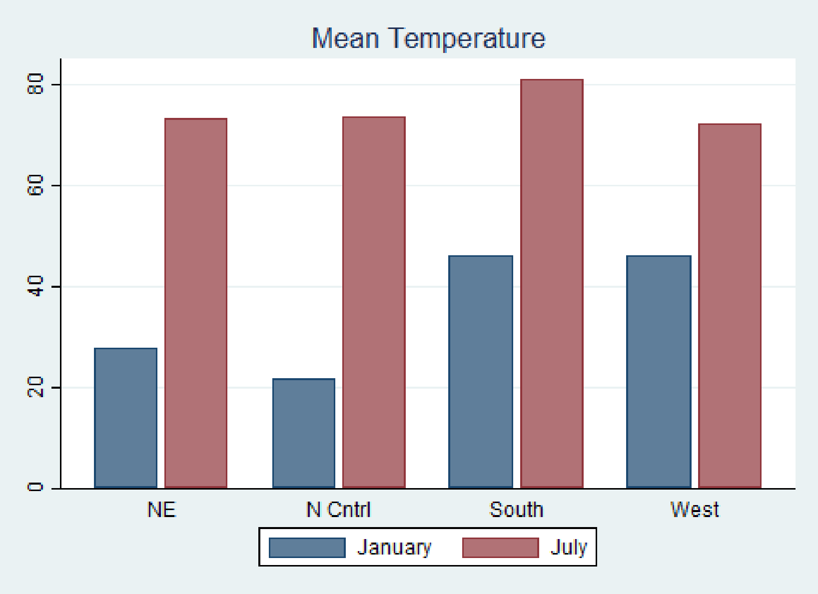
BAR PLOT EXAMPLE



IMPROVING BAR PLOT

- Let us show instead the average temperatures in January and July.
- To do this we could specify (mean) tempjan (mean) tempjuly, (the default statistic is the mean)
- We use over() so the regions are overlaid in the same graph; using by() instead, would result in a graph with a separate panel for each region.
- The bargap() option controls the gap between bars for different statistics in the same over group; here I put a small space.
- The gap() option, not used here, controls the space between bars for different over groups.
- The intensity of the color fill to 70

BAR PLOT EXAMPLE: MEAN TEMPERATURE



DATA ANAYLSIS

- Analysis Of Variance, popularly known as the ANOVA, can be used in cases where there are more than two groups.
- When we have only two samples we can use the t-test to compare the means of the samples but it might become unreliable in case of more than two samples.
- If we only compare two means, then the t-test (independent samples) will give the same results as the ANOVA.
- It is used to compare the means of more than two samples.

Download ”add-on” anova command anovaplot.

In STATA command window type: findit anovaplot

describe

```
. describe

Contains data from C:\statatraining\course_slides\course_slides\hers_640anova.dta
  obs:           612
 vars:             3                          11 Apr 2018 15:00
 size:         2,448
--------------------------------------------------------------------------------
              storage   display    value
variable name   type    format     label      variable label
--------------------------------------------------------------------------------
raceth          byte    %16.0g     raceth     race/ethnicity
physact         byte    %20.0g     physact    comparative physical activity
sbp             int     %9.0g                 systolic blood pressure
--------------------------------------------------------------------------------
Sorted by:
```

DESCRIPTIVE OF THE DATA

```
Summary for variables: sbp
     by categories of: raceth (race/ethnicity)
```

raceth	N	mean	sd
White	300	136.0133	18.55138
African American	218	138.2339	19.99252
Other	94	135.1809	21.25977
Total	612	136.6765	19.50878

ANALYSIS OF VARIANCE MODEL ESTIMATION

ANOVA IN STATA

- Stata offers at least 2 commands for a one way anova: anova or oneway.
- The command ANOVA uses deviation from means parameterization
- * anova YVARIABLE FACTOR
- The command ONEWAY uses deviation from means and provides Bartlett test of equal variances.
- * oneway YVARIABLE FACTOR

ANOVA RESULTS

Number of obs =	612	R-squared =	0.0037
Root MSE =	19.5042	Adj R-squared =	0.0005

Source	Partial SS	df	MS	F	Prob>F
Model	871.00017	2	435.50009	1.14	0.3190
raceth	871.00017	2	435.50009	1.14	0.3190
Residual	231670.94	609	380.41205		
Total	232541.94	611	380.59238		

Analysis of Variance

Source	SS	df	MS	F	Prob > F
Between groups	871.000171	2	435.500085	1.14	0.3190
Within groups	231670.941	609	380.412054		
Total	232541.941	611	380.592375		

Bartlett's test for equal variances: chi2(2) = 3.1766 Prob>chi2 = 0.204

TESTS OF EQUALITY OF VARIANCES

EQUALITY OF VARIANCES

- BARTLETT's Test is provided in output from command oneway
- This test is sensitive to the assumption of normality
- LEVENE and BROWN-FORSYTHE tests are obtained using the command robvar
- These are good choices when assumption of normality is in question.
- $*W_0 =$ Levene test
- $*W_50 =$ Forsythe-Browne modification of Levene test (mean is replaced by median)
- $*W_10 =$ Fosythe-Browne modification of Levene test (mean is replaced by 10
- $*$ robar(YVAR), by(FACTOR)

ANOVA RESULTS 2

race/ethnicity	Summary of systolic blood pressure Mean	Std. Dev.	Freq.
White	136.01333	18.551379	300
African A	138.23394	19.992518	218
Other	135.18085	21.259767	94
Total	136.67647	19.508777	612

```
W0  =  1.4143305   df(2, 609)     Pr > F = 0.24388559

W50 =  1.4701779   df(2, 609)     Pr > F = 0.23069929

W10 =  1.4741613   df(2, 609)     Pr > F = 0.22978655
```

BASIC PROGRAMMING STRUCTURES

LOCAL VERSUS GLOBAL MACRO

- A macro simply associates a name with some text (or numbers).
- Macros are objects stored in memory (they are not variables in the dataset!).
- The macro can be referenced anywhere in a program.
- Macros can either be local or global in scope.

LOCAL MACROS

LOCAL MACRO

- Local macros are defined as follows: local name [=] text
- Local macros are evaluated as follows: ‘name’
- To capture results, we use the second type of macro definition: local name = text
- The use of the equal sign tells stata to treat the text on the right hand side as an expression, evaluate it and store a representation of the result under the given name.

LOCAL MACRO: EXAMPLE

- We need to run a regression to explore the relationship between income and education with a bunch of control variables.
- We can store those control variables in a local called controls.
- local controls age agesquared male urban
- So instead of running the regression as follows:
- regress income education age agesquared male urban
- We can use the local to reference the controls:
- regress income education ‘controls’

- local controls age agesquared male urban
- regress lnwage EDYEARS ‘controls’.

```
. reg LN_WAGE_LCU_HR EDYEARS `controls'

      Source |       SS           df       MS      Number of obs   =     6,444
-------------+----------------------------------   F(4, 6439)      =   1130.34
       Model |  3943.56084         4  985.89021   Prob > F        =    0.0000
    Residual |  5616.13126     6,439 .872205508   R-squared       =    0.4125
-------------+----------------------------------   Adj R-squared   =    0.4122
       Total |   9559.6921     6,443  1.48373306   Root MSE        =    .93392

------------------------------------------------------------------------------
LN_WAGE_LC~R |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     EDYEARS |   .1373874   .0029856    46.02   0.000     .1315346    .1432402
        AGEY |   .0712808   .0038746    18.40   0.000     .0636852    .0788764
      AGEY_2 |  -.0005281    .000046   -11.47   0.000    -.0006184   -.0004379
       URBAN |     .37721   .0246409    15.31   0.000     .3289058    .4255143
       _cons |   .1291912   .0757213     1.71   0.088    -.0192477    .2776301
------------------------------------------------------------------------------
```

LOCAL MACRO: STORING RESULTS EXAMPLE

- sum
- return list
- Stata displays 8 “scalar” quanitites.
- local meanwage = r(mean)
- display ‘meanwage’

```
. su lnwage

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
      lnwage |      6,805    3.281032    1.228599  -4.941642   8.334871

. return list

scalars:
                  r(N) =  6805
              r(sum_w) =  6805
               r(mean) =  3.281032280263701
                r(Var) =  1.509454811377814
                 r(sd) =  1.228598718613126
                r(min) =  -4.941642284393311
                r(max) =  8.334871292114258
                r(sum) =  22327.42466719449

.  local meanwage = r(mean)

. display `meanwage'
3.2810323
```

LOOPS

LOOPS

- Loops are used to do repetitive tasks.
- Stata has commands that allow looping over sequences of numbers and various types of lists, including lists of variables.
- Looping over sequences of numbers
- Looping over elements in a list
- Looping over variables

DEMONSTRATION OF LOOPS OVER SEQUENCES OF NUMBERS

```
. forvalues number = 1000(50)2000 {
  2.
.     di `number'
  3.
. }
1000
1050
1100
1150
1200
1250
1300
1350
1400
1450
1500
1550
1600
1650
1700
1750
1800
1850
1900
1950
```

DEMONSTRATION OF LOOPS OVER ELEMENTS IN A LIST

```
. foreach software in R, SPSS, SAS, STATA  {
  2.
.     display "`software'"
  3.
. }
R,
SPSS,
SAS,
STATA
```

QUESTIONS??

QUESTIONS