☺A

EXPLORATORY DATA ANALYSIS: AN APPLICATION TO CROSS-SECTION DATA

Maria Elena Bontempi e.bontempi@economia.unife.it

Roberto Golinelli roberto.golinelli@unibo.it

this version: 26/09/2007[§]

1. Introduction

This note presents commands to deal with the exploratory data analysis (EDA) applied to cross-sections.[1]

Commands well operate in releases 7, 8, and 9 of STATA. The only warning concerns the syntax of the graph command that has been completely revised in version 8/9 of Stata. We usually follow the old graph commands of Stata 7 because they are easier. If you are working in Stata 8/9, you may adopt different strategies depending on the situation. Write `graph7` instead of `graph`: given that this last command does not exist anymore in Stata 8/9, if you type `graph` you would obtain an error message. Write `version 7: rvfplot` instead of `rvfplot`: this last command does exist in Stata 8/9, but the way in which options are used is changed; thus, if you type `rvfplot` and use the version 7 options you would obtain an error message.

Just as an idea, the Stata 8/9 graph commands include six different graph types:

```
graph twoway              scatterplots, line plots, etc.
graph matrix              scatterplot matrices
graph bar                 bar charts
graph dot                 dot charts
graph box                 box and whisker plots
graph pie                 pie charts
```

The `graph twoway` type includes the following graphs among others:

```
scatter          scatterplot
line             line plot
connected        connected-line plot
area             line plot with shading
spike            spike plot
dot              dot plot
lowess           LOWESS line plot
lfit             linear prediction plot
qfit             quadratic prediction plot
lfitci           linear prediction plot with CIs
qfitci           quadratic prediction plot with CIs
function         line plot of function
histogram        histogram plot
kdensity         kernel density plot
```

As an example, in Stata 7 you type:

```
. graph y yhat x, symbol([var_name].) connect(.l)
```

The corresponding graph in Stata 8/9 is obtained by:

```
. twoway (lfitci y x) (scatter y x, mlabel(var_name))
```

[§] Very preliminary. Comments welcome.

[1] Tukey J. W. (1970) Exploratory Data Analysis, Reading, MA: Addison-Wesley, v. 1. Hoaglin D. C., Iglewicz B. and Tukey J. W. (1986) Performance of Some Resistant Rules for Outlier Labelling, Journal of the American Statistical Association, v. 81, n. 396, 991-999. Frigge M., Hoaglin D. C. and Iglewicz B. (1989) Some Implementations of the Boxplot, The American Statistician, v. 43, n. 1, 50-54.

or

```
scatter y yhat x, msymbol(i i) connect(. l) mlabel(var_name)
```

The `symbol` option has been changed with the marker symbol, `msymbol`; moreover, note the spaces between each elements in the list of indicators inside brackets of `msymbol` and `connect` (not required in Stata 7).

2. Univariate data analysis: distribution, center, spread, outlier.

Data (source Williams K. R. and Flewelling R. L. (1988) The Social production of Criminal Homicide: a Comparative Case Study of Disaggregated Rates in American Cities, American Sociological Review, 53, 3, 421-431) in urban.dta:

```
cd ….
descr

      Contains data from urban.dta
  obs:            20                          20 cities >100k pop
 vars:            13
 size:         1,120 (99.7% of memory free)
-------------------------------------------------------------------------------
              storage  display     value
variable name   type   format      label      variable label
-------------------------------------------------------------------------------
city            str16  %16s                   City
state           byte   %8.0g       slbl       State code
region          byte   %8.0g       rlbl       Geographical region
divorce         float  %9.3f                  Divorces/1000 ages 15-59
educ            float  %9.0g                  Median years education
inequal         float  %9.0g                  Household inequality index
change          float  %9.0g                  % population change  1970-1980
pop             float  %9.1f                  Population in 1,000s
poor            float  %9.2f                  Percent families below poverty
homic           float  %9.2f                  Homicides/100,000 people
count           byte   %8.0g                  Frequency
poorrank        float  %9.0g                  Poverty rank
homrank         byte   %8.0g                  Homicide rank
-------------------------------------------------------------------------------
```

where the variables of interest are:

City	indicator of individuals: 20 US cities, randomly drawn. i=1,...,N, N=20
Pop	city population, in thousands of 1980
Homic	homicide victims per 100,000 people (average of 1980-1984 years)
Poor	percentage of families with incomes below the poverty line

Aim of the analysis: explain why the homicide rate is higher in some cities than others, supposing that the criminality is a function of social hardships.

Before proceeding in the regression analysis, it is important the exploratory data analysis (EDA). The EDA allows for a number of checks to make sure we can firmly stand behind the regression results. Aims of EDA.

- Description of the sample (randomly or not randomly drawn, how much it is representative of a larger population).
- Investigating the data combining knowledge, judgement, and statistical methods, without imposing any particular assumption or a priori model. Looking for errors in the data and verifying whether the data meet the assumptions of linear regression.
- Analysing the variables distributions (are they normal or not?). Graphs can at once show information about the shape of the variables better than simple numeric statistics can. Again, it is important to detect the pattern in the data rather than to impose pattern on the data.

- Studying measures or statistics summarising important features of variable distributions, like centre (mean, median) and spread (variance, quartiles).
- Verifying if outliers are present (in which case it is important to refer to non parametric characterisation of the variables distributions) and if it is useful to impose cleaning rules (of course, taking into account their impact in terms of selection bias).

2.1. Measurement of the variables of interest

Why the dependent variable is a rate, instead of a level? Below we create *pop1000*, population measured in units, and *nhomic*, number of homicides.

```
g pop1000=pop*1000
g nhomic=homic*pop1000/100000
sort nhomic
list city pop homic pop1000 nhomic
```

```
                 city        pop      homic    pop1000     nhomic
  1. Sterling Heights      109.0       0.55     109000      .5995
  2.         Fullerton      102.0       2.35     102000      2.397
  3.         Sunnyvale      106.6       2.44     106600    2.60104
  4.           Concord      103.3       3.10     103300     3.2023
  5.      Independence      111.8       3.58     111800    4.00244
  6.             Tempe      106.7       4.12     106700    4.39604
  7.         Allentown      103.8       4.24     103800    4.40112
  8.            Peoria      124.2       4.03     124200    5.00526
  9.              Erie      119.1       4.70     119100     5.5977
 10.          Berkeley      103.3       8.52     103300   8.801161
 11.         Salt Lake      163.0       6.01     163000     9.7963
 12.   Virginia Beach      262.2       3.81     262200    9.98982
 13.          Columbus      169.4       9.21     169400   15.60174
 14.       Albuquerque      331.8       6.39     331800   21.20202
 15.         Rochester      241.7      10.84     241700   26.20028
 16.             Tulsa      360.9       8.64     360900   31.18176
 17.          Portland      366.4       8.62     366400   31.58368
 18.          Honolulu      365.0       9.15     365000    33.3975
 19.         Milwaukee      636.2       7.83     636200   49.81446
 20.            Dallas      904.1      29.98     904100   271.0492
```

Albuquerque and Rochester present very similar values for the number of homicides, while population is very different. If the number of homicide victims increase with the population, criminality measured by the number of homicides does not take into account that Albuquerque, despite the large population, has a low criminality if compared to Rochester. The rate normalises respect to the population and allows for the comparison of cities with different population. The homicide rate is measured in thousands in order to allow for a better perceiving:

```
. summ homic pop poor
```

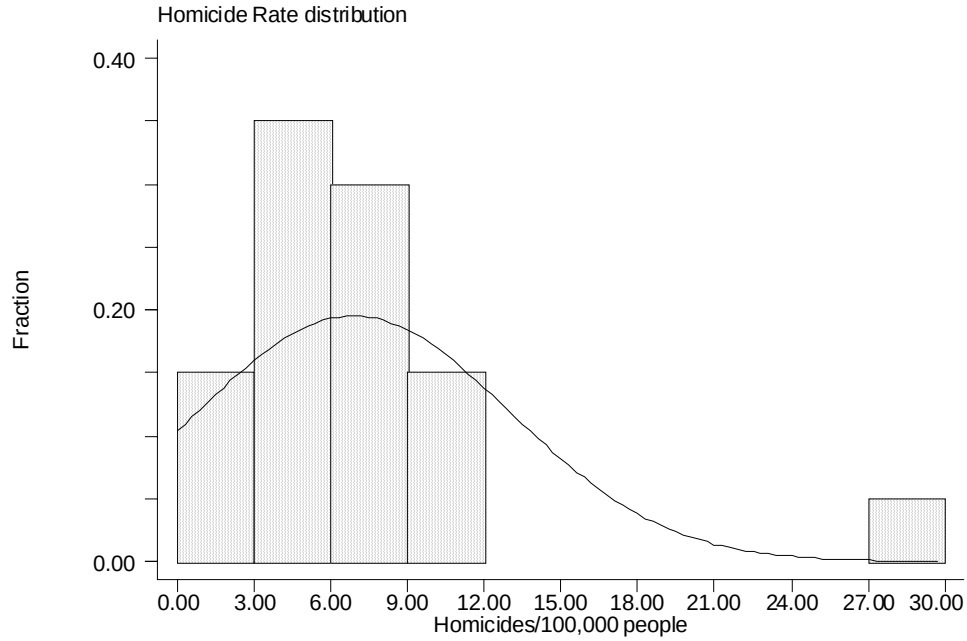
```
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
       homic |        20      6.9055     6.12563        .55      29.98
         pop |        20     244.525    209.7372       102      904.1
        poor |        20        8.18    3.273434       3.1       14.5
```

2.2. Distribution

In what follows we apply EDA techniques to the dependent variable. However, it is important to perform such an inspection for each variable of your model.

The histogram graphically shows a frequency distribution of a variable. The vertical axis indicates frequency, proportion, or percentage.

```
graph7 homic, bin(10) norm xlabel(0, 3 to 30) ylabel t1(Homicide Rate distribution) ytick(0, 0.05 to 0.35)
```



Some useful options are:
`bin` selects the number of equal-width classes (columns) in which data values are grouped (the default is 5). Remember that histograms are sensitive to the number of bins used in the display. The rule of thumb is to use high bin for a variable that assumes many values (like an interest rate) and to use fewer bins for a variable assuming fewer values (like the number of car's models produced by a company). This avoids to loose information and to make o wrong opinion on the data.

`normal` superimposes a normal distribution, calculated on the observed mean and standard deviation. The normal or Gaussian distribution is a bell-shaped pdf, symmetric around the mean=median=mode (skewness=0), mesokurtic (kurtosis=3).

`xlabel ylabel` are useful if you want to select particular values on the axes.

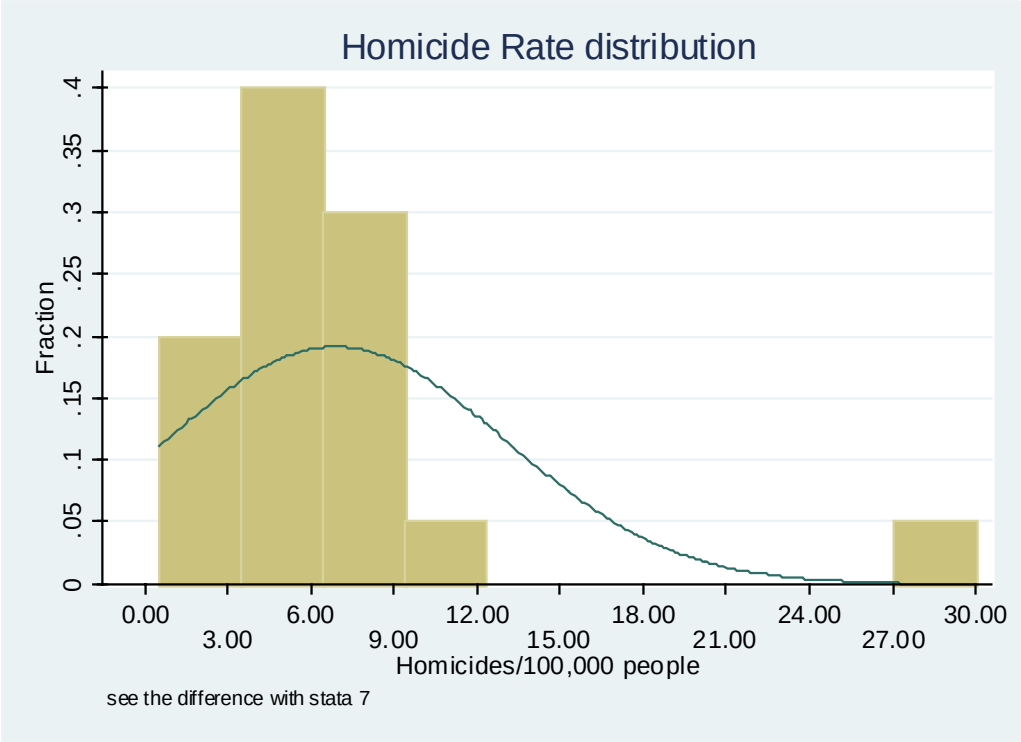
`xtick ytick` show ticks corresponding to the values you selected.

`t1( )` puts a title in the graph.

Adding the option `saving(fig1, replace)` you may save a .gph file containing your graph (after, you can copy your figure in your document).

The same graph in Stata 8/9, with some of the many improvements, is obtainable as:

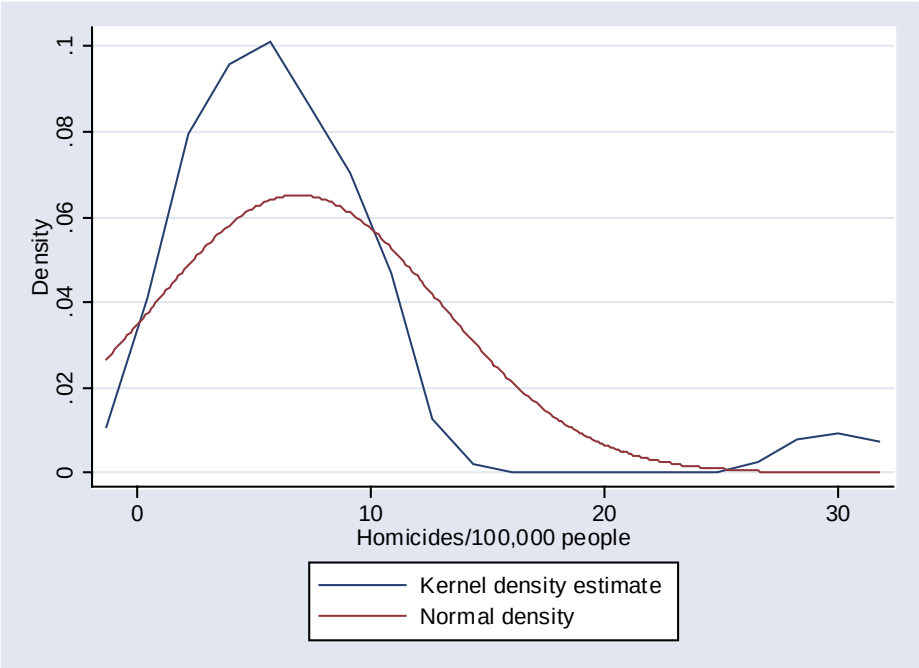
```
. histogram homic, fraction normal bin(10) xlabel(0(3)30, alternate) ytick(0(0.05)0.40, tposition(crossing)) ylabel(#10) title("Homicide Rate distribution") note("see the difference with stata 7")
(bin=10, start=.55000001, width=2.943)
```



The picture shows that *homic* does not follow a normal distribution: positively skewed distributions are frequent in variables with a lower limit of zero, but a no definite upper bound, or in variables affected by outliers.

An alternative to histogram is the Stata 8/9 kernel density plot, which approximates the probability density of the variable and has the advantage of being smooth and independent of the choice of origin.

```
. kdensity homic, normal
(n() set to 20)
```



2.3. Statistics of centre and spread, and outliers

```
summ homic
```

```
    Variable |     Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
       homic |      20      6.9055    6.12563        .55      29.98
```

Not normal distributions complicate the job of statistical analysis: non-normality may depend on exceptionally high or low values (outliers) that are to be carefully considered. The concept of centre is ambiguous, because the tails tend to affect the mean. Also other parametric statistics used to summarise the distribution are affected by outliers; in particular, spread measures like standard deviation and variance, based on squared deviations from the mean, have even less resistance to extreme values than the mean does. A single outlier can dramatically inflate skewness and kurtosis statistics, which depend on third and fourth powers of deviations from the mean.
In these cases, order statistics may be more useful in summarising a skewed distribution; moreover, they are resistant, i.e. less affected by few extreme values (outliers). Order statistics are defined from the position of values within an ordered list. In Stata, they are obtained by:

```
summ homic, d
```

```
                  Homicides/100,000 people
-------------------------------------------------------------
      Percentiles      Smallest
 1%           .55            .55
 5%          1.45           2.35
10%         2.395            2.44       Obs                  20
25%         3.695             3.1       Sum of Wgt.          20

50%         5.355                       Mean              6.9055
                        Largest         Std. Dev.        6.12563
75%          8.63           9.15
90%        10.025           9.21       Variance         37.52334
95%         20.41          10.84       Skewness         2.759817
99%         29.98          29.98       Kurtosis         11.28847
```

Order statistics are the following.

- Lower and upper bounds (minimum and maximum)
- Median: 50% of cases lie below the median, and the other 50% above.

Given an ordered list of N values, the median equals the value at position (N+1)/2. If N is odd, a single value occupies this position, while, if N is even, the median is usually computed as the mean of the two middle values.

- Quartiles divide the ordered list in quarters.

The first quartile (Q_1 or 25%) is a number theoretically greater than the values of 25% of the cases and lower than the values of the remaining 75%. Analogously, the third quartile (Q_3 or 75%) is a number greater than the values of 75% of the cases and lower than the remaining 25%. Finally, the second quartile (Q_2 or 50%) is the same as the median, and it separates the lower and upper halves of the data.

- Percentiles divide ordered lists into hundreds.

1% of the cases lie below the first percentile, while 99% lie above it. 10% of the cases lie below the tenth percentile, etc. Of course: $Q_1=25^{th}$ percentile; $Q_2=50^{th}$ percentile; $Q_3=75^{th}$ percentile. The value of the p^{th} sample percentile is $(homic_{i-1}+homic_i)/2$ if $cum(P_{i-1})=p$ and $homic_i$ otherwise, where $cum(P_i)$ represent the cumulative percentage of cases up to and including $homic_i$.

The `summ ,d` command also shows the following statistics:
The variance (or the standard deviation, in the same unit of measurement of the mean) that measures the spread of a variable around its average.

The skewness, i.e. the direction and degree of asymmetry in a distribution, is measured by the third central moment. In a normal or symmetrical distribution the skewness = 0, while positive skewness will result if a distribution is skewed to the right, since cubed discrepancies about the mean would be positive. Finally, a left-skewed distribution has skewness < 0.
The kurtosis, i.e. the weight of the tails, is measured by the fourth central moment. In a normal or mesokurtic distribution the kurtosis = 3; a leptokurtic or slim or long-tailed (platykurtic or fat or short-tailed) distribution has kurtosis > (<) 3.

Despite its usefulness in statistical analysis, the mean has a weakness: it can be drastically affected by even one extremely high or low value. The median is an alternative and resistant measure of centre; resistant means that it is less easily affected by extreme values. Indication on the asymmetry of a distribution may also be obtained by comparing mean and median:
if mean = median, the distribution is symmetrical (skewness = 0);
if mean > median, the distribution is affected by positive outliers (skewness > 0);
if mean < median, the distribution is affected by negative outliers (skewness < 0).

By using empirical measures of skewness and kurtosis, it is possible to test normality of a distribution. A simple test of normality is to find out whether the computed values of skewness and kurtosis depart from the norms of 0 and 3. Thus, the null hypothesis is H_0: S = EK = 0, where S and EK respectively indicate the theoretical measures of skewness and excess of kurtosis (kurtosis minus 3).
This is the logic underlying the Jarque-Bera (1980)[2] test of normality, based on the following test-statistic:

$$JB = N\left[\frac{\hat{S}^2}{6} + \frac{E\hat{K}^2}{24}\right] \sim \chi^2_2, \text{ where } \hat{S} = \frac{\frac{1}{N}\sum_{i=1}^{N}(y_i - \bar{Y})^3}{\left[\frac{1}{N}\sum_{i=1}^{N}(y_i - \bar{Y})^2\right]^{3/2}} \text{ and } E\hat{K} = \frac{\frac{1}{N}\sum_{i=1}^{N}(y_i - \bar{Y})^4}{\left[\frac{1}{N}\sum_{i=1}^{N}(y_i - \bar{Y})^2\right]^{2}} - 3.$$

Under the null hypothesis of normality, JB is distributed as a chi-square statistic with 2 df.

```
sktest homic
```

```
                    Skewness/Kurtosis tests for Normality
                                                    ------- joint ------
    Variable |  Pr(Skewness)   Pr(Kurtosis)  adj chi2(2)    Prob>chi2
-------------+-------------------------------------------------------
       homic |      0.000         0.000           23.39       0.0000
```

The null is rejected, both when separately tested on skewness and kurtosis taken alone and when jointly tested.

Technical note
How can you understand that the null is rejected? How can you use the statistical tables in Stata?
For example, in the case, as above, of the chi-distribution:

```
. di invchi2tail(2, .05)
5.9914645
```

gives you the critical value of the chi-squared with 2 degrees of freedom and for a significance level equal to 5%. Given that your test-statistic, distributed as a chi-squared, is equal to 23.39 and that 23.39>5.99, you reject H_0.
Alternatively (more immediate) you may ask which is the P-value associated to your test-statistic:

[2] Jarque C. M. and Bera A. K. (1980) Efficient Tests for Normality, Homoskedasticity and Serial Independence of Regression Residuals, Economics Letters, 6, 255-259.

```
. di chi2tail(2, 23.39)
8.335e-06
```

This P-value is exactly the same reported by the `sktest` command; given that it is less than the size, 5%, again you reject H_0.

Establishing whether the dependent and explanatory variables follow normal distributions is important. While the regression coefficients do not require normally distributed residuals to be unbiased, the residuals need to be normally distributed if we are interested in having valid inference: remember that our hypothesis-testing procedure, as in the t and F tests, is based on the assumption (at least in small or finite samples) that the underlying distribution of the variable (or sample statistic) is normal.

Another useful command for preliminary analysis, written by Lawrence C. Hamilton, Dept. of Sociology, Univ. of New Hampshire, is available by adding to your directory the iqr.ado and iqr.hlp files:

```
iqr homic

mean=  6.905           std.dev.=  6.126          (n= 20)
median=  5.355    pseudo std.dev.=  3.658        (IQR=  4.935)
10 trim=  5.899
                                              low         high
                                              -------------------
                                   inner fences   -3.708       16.03
                               # mild outliers    0           0
                               % mild outliers    0.00%       0.00%

                                   outer fences   -11.11       23.44
                               # severe outliers  0           1
                               % severe outliers  0.00%       5.00%
```

Other order statistics are the following.

- IQR is the interquartile range, equal to the distance between the first and third quartiles, or between the 75th and the 25th percentiles: IQR=Q_3-Q_1, where this distance spans the middle 50% of the data.

Median and IQR respectively measure centre and spread of a distribution. They are analogous to mean and standard deviation, but more resistant, thus providing better summaries when the data are skewed or contain outliers.

- Pseudo-standard deviation, PSD = IQR/1.349, where 1.349=2×0.674 is the interval containing 50% of cases in a normal distribution. The PSD is another resistant measure of spread.

If the distribution appears roughly symmetrical, the comparison between standard deviation and pseudo-standard deviation provide information:
if standard deviation = PSD, the distribution has normal tails;
if standard deviation > PSD, the distribution has heavier than normal tails (kurtosis > 3);
if standard deviation < PSD, the distribution has lighter than normal tails (kurtosis < 3).

- 10 trim is the 10% trimmed mean, i.e. the average of cases between 10th and 90th percentiles. It is a less radical (if compared to the median, a 50% trimmed mean), but still resistant summary measure. In particular, it retains much of the normal-distribution efficiency of a mean, but it performs better than means with heavy-tailed distributions. In a symmetrical distribution, trimmed mean = median = mean.

Inner fences = $Q_1 - 1.5 \times IQR$ and $Q_3 + 1.5 \times IQR$

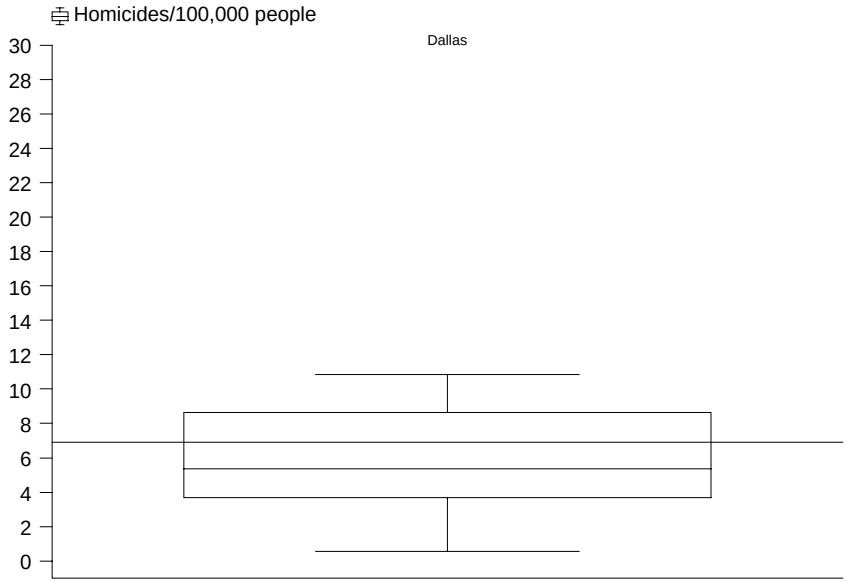
Outer fences = $Q_1 - 3 \times IQR$ and $Q_3 + 3 \times IQR$ [3]

These cut-offs are used to define mild and severe outliers.
Mild outliers may be no cause for alarm, because they make up about 0.7% of a normal distribution.[4] They are defined as:
$Q_1 - 3 \times IQR \leq$ *mild outliers* $< Q_1 - 1.5 \times IQR$
$Q_3 + 1.5 \times IQR <$ *mild outliers* $\leq Q_3 + 3 \times IQR$.

Severe outliers make up about 0.0002% (two per million) of a Gaussian population[5]; Thus, they have substantial effects on means, standard deviations and other statistics. They are defined as:
severe outliers $< Q_1 - 3 \times IQR$
severe outliers $> Q_3 + 3 \times IQR$

The following graph is a boxplot, that graphically displays the median, the IQR, the shape, and the outliers of a distribution.

```
graph7 homic, box yline(6.9055) ylabel(0, 2 to 30) s([city])
```



The corresponding command in Stata 8/9 is `graph box homic, options`.
The upper and lower horizontal lines of the box are Q_3 and Q_1, respectively so the height of the box is the IQR. The horizontal line inside the box is the median: its location (in this case low in the box) suggests positive skew. This is confirmed by the mean (displayed by the option `yline(6.9055)`), higher than the median.
Vertical lines extend from each quartile to adjacent values, i.e. values of the last cases nearest to but not beyond the inner fences, $Q_3 + 1.5 \times IQR$ and $Q_1 - 1.5 \times IQR$. By sorting and listing *homic*, you will note that these adjacent values are 0.55 e 10.84, respectively (so, do not be surprised if the vertical lines are not symmetric!!). These values serve as dividing lines for identifying positive and negative outliers. In our case, a positive outlier (Dallas) is identified by name using a string variable as plotting symbol (option `s([city]`).

[3] 1.5 and 3 are the values in a standardised normal that respectively leave 7% and 0.2% of probability in the tail. Type `display invnorm(0.07)` and `display invnorm(0.002)`.
[4] See Monte Carlo simulation in Hoaglin-Iglewicz-Tukey (1986).
[5] See Monte Carlo simulation in Hoaglin-Iglewicz-Tukey (1986).

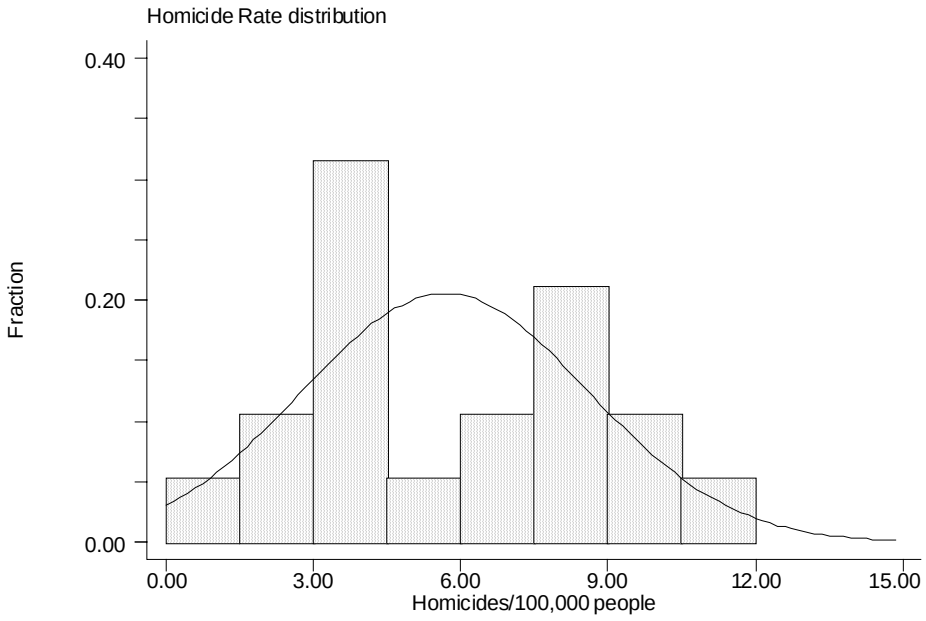
Dallas appears as an outlier due to its exceptionally high homicide rate; however, Dallas' poverty rate is not unusual.
If we drop the observation corresponding to Dallas, we obtain what follows.

```
summ homic if city!="Dallas", d
```

```
                  Homicides/100,000 people
-------------------------------------------------------------
      Percentiles      Smallest
 1%           .55           .55
 5%           .55          2.35
10%          2.35          2.44       Obs                  19
25%          3.58           3.1       Sum of Wgt.          19

50%           4.7                     Mean           5.691053
                        Largest       Std. Dev.      2.910595
75%          8.62          8.64
90%          9.21          9.15       Variance       8.471566
95%         10.84          9.21       Skewness       .1193361
99%         10.84         10.84       Kurtosis       1.850951
```

```
graph7 homic if city!="Dallas", bin(10) norm xlabel(0, 3 to 15) ylabel
t1(Homicide Rate distribution) ytick(0, 0.05 to 0.35)
```




```
sktest homic if city!="Dallas"
```

```
                     Skewness/Kurtosis tests for Normality
                                                    ------- joint ------
    Variable |  Pr(Skewness)   Pr(Kurtosis)  adj chi2(2)    Prob>chi2
-------------+-------------------------------------------------------
       homic |      0.793         0.142            2.52       0.2838
```

This preliminary analysis let us supposing that the errors of our regression model will not be normally distributed. This implies that we can no more apply t and F distributions for making inference, especially in small samples. Moreover, since OLS is affected by outliers, we can have great sample-to-sample variation.
In conclusion, it is important to detect if non normal errors derive from: outliers (lecture_OLS_bivariate and lecture_OLS_multivariate), heteroskedasticy (lecture_GLS), or curvilinearity (lecture_nonlinearity_Chow).