eman ta zabal zazu

Universidad del País Vasco Euskal Herriko Unibertsitatea

CAMPUS OF INTERNATIONAL EXCELLENCE

Example 3.5

Univariate analysis

Pilar González and Susan Orbe

Dpt. Applied Economics III (Econometrics and Statistics)

Contents

1. 3.5.1. Normality tests.
   - Pizza consumption.
   - Visitors to Bilbao.
2. 3.5.2. Frequency distribution.
   - Pizza consumption.
   - Visitors to Bilbao.
   - Simulation.
3. 3.5.3. Estimated density plot.
   - Pizza consumption.
   - Visitors to Bilbao.
4. 3.5.4. Boxplot.
5. 3.5.5. Normal Q-Q plot.
   - Visitors to Bilbao.
   - Simulation.

Contents

1. 3.5.1. Normality tests.
   - Pizza consumption.
   - Visitors to Bilbao.
2. 3.5.2. Frequency distribution.
   - Pizza consumption.
   - Visitors to Bilbao.
   - Simulation.
3. 3.5.3. Estimated density plot.
   - Pizza consumption.
   - Visitors to Bilbao.
4. 3.5.4. Boxplot.
5. 3.5.5. Normal Q-Q plot.
   - Visitors to Bilbao.
   - Simulation.

Example 3.5.1. Normality tests. Pizza consumption.

Questions.

Load the data file `pizza.gdt`.

a. Run normality tests for the variables pizza and income.

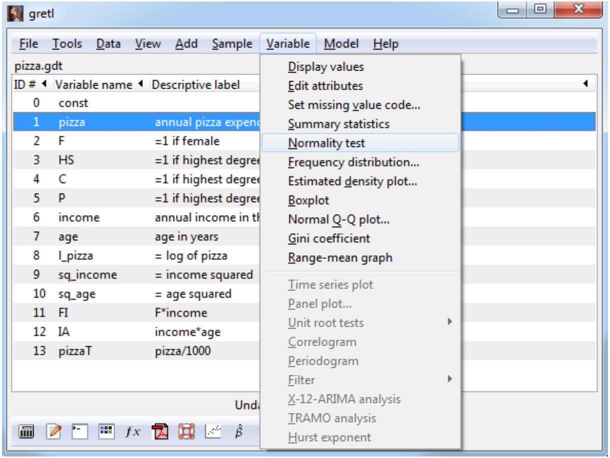
b. Save the results as an icon to your session.

c. Interpret the results and save the session with the name dpizza.

Example 3.5.1. Normality tests. Pizza consumption.

To run normality tests, go up to the Menu Bar and click

```
Variable -> Normality test
```



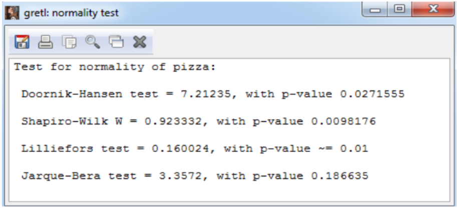
pizza.gdt

ID #	Variable name	Descriptive label
0	const	
1	pizza	annual pizza expend
2	F	=1 if female
3	HS	=1 if highest degree
4	C	=1 if highest degree
5	P	=1 if highest degree
6	income	annual income in th
7	age	age in years
8	l_pizza	= log of pizza
9	sq_income	= income squared
10	sq_age	= age squared
11	FI	F*income
12	IA	income*age
13	pizzaT	pizza/1000

Example 3.5.1. Normality tests. Pizza consumption.

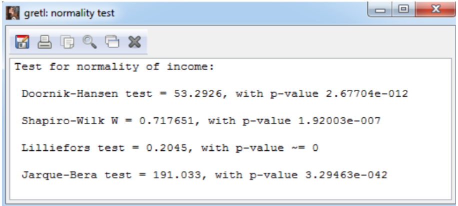
$$\begin{aligned} H_0 &: \text{ variable comes from a normal distribution} \\ H_a &: \text{ variable does not come from a normal distribution} \end{aligned}$$

The results are shown in the figures below.



```
Test for normality of pizza:

 Doornik-Hansen test = 7.21235, with p-value 0.0271555

 Shapiro-Wilk W = 0.923332, with p-value 0.0098176

 Lilliefors test = 0.160024, with p-value ~= 0.01

 Jarque-Bera test = 3.3572, with p-value 0.186635
```

PIZZA



```
Test for normality of income:

 Doornik-Hansen test = 53.2926, with p-value 2.67704e-012

 Shapiro-Wilk W = 0.717651, with p-value 1.92003e-007

 Lilliefors test = 0.2045, with p-value ~= 0

 Jarque-Bera test = 191.033, with p-value 3.29463e-042
```

INCOME

Example 3.5.1. Normality tests. Pizza consumption.

Results.

The output consists of a table with the statistics and the p-values.

The p-value is the estimated probability of rejecting the null hypothesis when it is true.

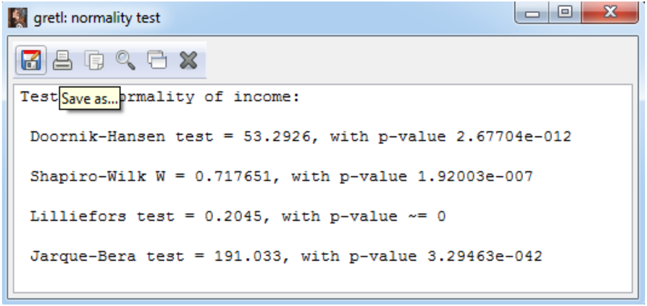
Decision rule: if the p-value is smaller than $\alpha = 0.05$, the null hypothesis is rejected at a $5\,\%$ significance level.

Conclusions:

- Most of the statistics indicate that the variable pizza does not come from a normal distribution.
- All the statistics indicate that the variable income does not come from a normal distribution.

Example 3.5.1. Normality tests. Pizza consumption.

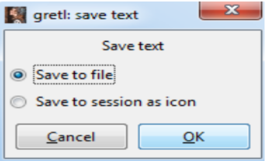
To save the results, go up to the menu bar in the **normality test** window and click on the left icon (*Save as...*).



```
Test for normality of income:

 Doornik-Hansen test = 53.2926, with p-value 2.67704e-012

 Shapiro-Wilk W = 0.717651, with p-value 1.92003e-007

 Lilliefors test = 0.2045, with p-value ~= 0

 Jarque-Bera test = 191.033, with p-value 3.29463e-042
```

Example 3.5.1. Normality tests. Pizza consumption.

To save the results in a file select the first option.



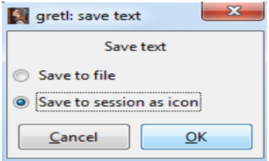
gretl: save text

Save text

● Save to file

○ Save to session as icon

To save the results as an icon to the session select the second option.



gretl: save text

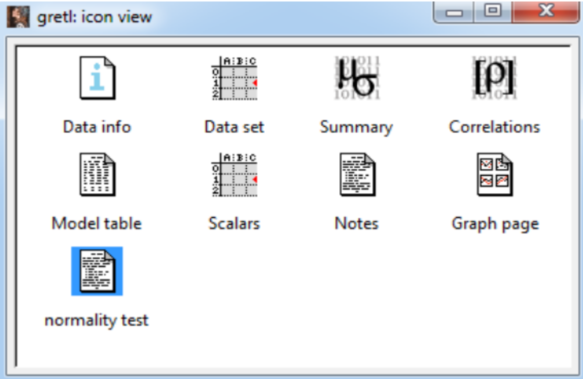
Save text

○ Save to file

● Save to session as icon

Example 3.5.1. Normality tests. Pizza consumption.

By default, Gretl names the icon *normality test*. The results can be retrieved double-clicking on this icon.



Do not forget to save your session!

Example 3.5.1. Normality tests. Visitors to Bilbao.

Questions.

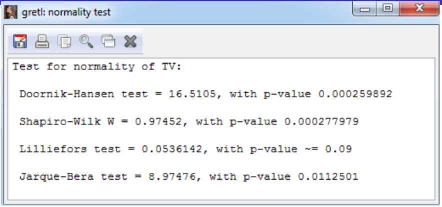
Load the data file `tourism.gdt`.

a. Run normality tests for the three visitors variables.

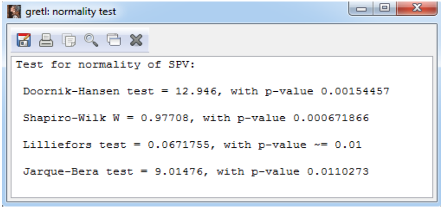
b. Save the results as an icon to your Gretl session.

c. Interpret the results and save the session with the name `dturismo`.

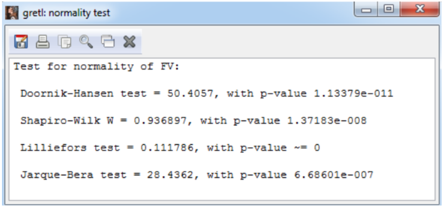
Example 3.5.1. Normality tests. Visitors to Bilbao.



```
Test for normality of TV:

 Doornik-Hansen test = 16.5105, with p-value 0.000259892

 Shapiro-Wilk W = 0.97452, with p-value 0.000277979

 Lilliefors test = 0.0536142, with p-value ~= 0.09

 Jarque-Bera test = 8.97476, with p-value 0.0112501
```



```
Test for normality of SPV:

 Doornik-Hansen test = 12.946, with p-value 0.00154457

 Shapiro-Wilk W = 0.97708, with p-value 0.000671866

 Lilliefors test = 0.0671755, with p-value ~= 0.01

 Jarque-Bera test = 9.01476, with p-value 0.0110273
```



```
Test for normality of FV:

 Doornik-Hansen test = 50.4057, with p-value 1.13379e-011

 Shapiro-Wilk W = 0.936897, with p-value 1.37183e-008

 Lilliefors test = 0.111786, with p-value ~= 0

 Jarque-Bera test = 28.4362, with p-value 6.68601e-007
```

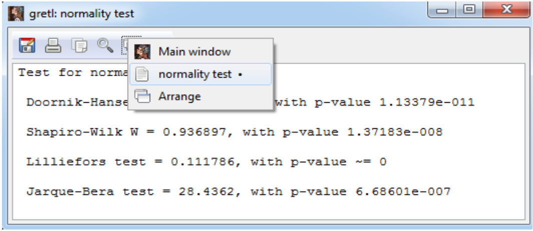
Result.

It may be concluded, at the 5 % significance level, that the series of visitors do not come from a normal distribution.

Do not forget to save these results as icons to your session!

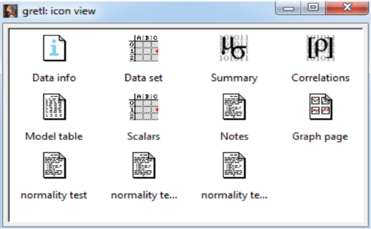
Example 3.5.1. Normality tests. Visitors to Bilbao.

The saved results may be retrieved by clicking on the icon *windows* in the **normality test** window.



```
Test for norma

 Doornik-Hanse                  with p-value 1.13379e-011

 Shapiro-Wilk W = 0.936897, with p-value 1.37183e-008

 Lilliefors test = 0.111786, with p-value ~= 0

 Jarque-Bera test = 28.4362, with p-value 6.68601e-007
```

Or from the session (there is one icon for each test).



Contents

1. 3.5.1. Normality tests.
   - Pizza consumption.
   - Visitors to Bilbao.
2. 3.5.2. Frequency distribution.
   - Pizza consumption.
   - Visitors to Bilbao.
   - Simulation.
3. 3.5.3. Estimated density plot.
   - Pizza consumption.
   - Visitors to Bilbao.
4. 3.5.4. Boxplot.
5. 3.5.5. Normal Q-Q plot.
   - Visitors to Bilbao.
   - Simulation.

Example 3.5.2. Frequency distribution. Pizza consumption.

Questions.

Analyse the frequency distribution of the variables included in the data file `pizza.gdt`.

a. Obtain the frequency distribution of the variable pizza.

b. Obtain the graph of the frequency distribution and run the tests against the normal and gamma distributions for the variable pizza.

c. Save the numerical results as icons to your Gretl session and the graphs in Word format.

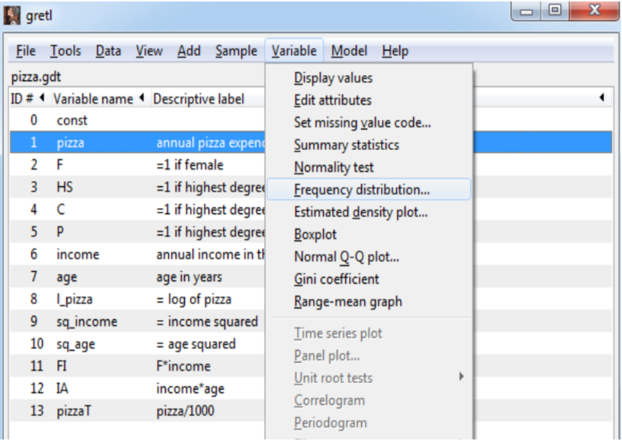
d. Comment on the results.

e. Save the session.

Example 3.5.2. Frequency distribution. Pizza consumption.

To obtain the frequency distribution and its plot, highlight the variable of interest and select the option *Frequency distribution* from the **Variable** pulldown menu.

```
Variable -> Frequency distribution...
```



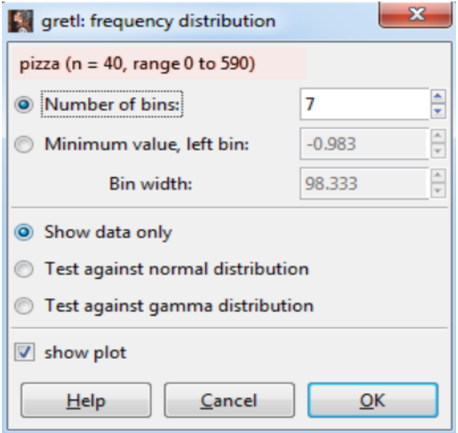
pizza.gdt

ID #	Variable name	Descriptive label
0	const	
1	pizza	annual pizza expen
2	F	=1 if female
3	HS	=1 if highest degree
4	C	=1 if highest degree
5	P	=1 if highest degree
6	income	annual income in th
7	age	age in years
8	l_pizza	= log of pizza
9	sq_income	= income squared
10	sq_age	= age squared
11	FI	F*income
12	IA	income*age
13	pizzaT	pizza/1000

Alternatively, highlight the variable of interest, right click and select the option

Example 3.5.2. Frequency distribution. Pizza consumption.

From the dialog box you will tell Gretl the number of bins you want to use and the type of output you want to get.



gretl: frequency distribution

pizza (n = 40, range 0 to 590)

● Number of bins: 7

○ Minimum value, left bin: -0.983

Bin width: 98.333

● Show data only

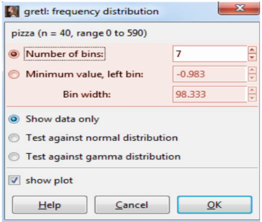
○ Test against normal distribution

○ Test against gamma distribution

☑ show plot

Example 3.5.2. Frequency distribution. Pizza consumption.

By default, Gretl calculates the number of bins depending on the range of values and shows both the numerical results for the frequency distribution and its plot, but you may select any other number of bins. It is also possible to change the number of bins by writing the minimum value of the left bin and its width.



pizza (n = 40, range 0 to 590)

● Number of bins: 7

○ Minimum value, left bin: -0.983

Bin width: 98.333

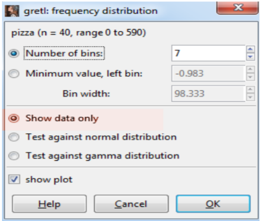
● Show data only

○ Test against normal distribution

○ Test against gamma distribution

☑ show plot

It is possible to run tests against the normal distribution or the gamma distribution.



pizza (n = 40, range 0 to 590)

● Number of bins: 7

○ Minimum value, left bin: -0.983

Bin width: 98.333

● Show data only

○ Test against normal distribution

○ Test against gamma distribution

☑ show plot

Example 3.5.2. Frequency distribution. Pizza consumption.

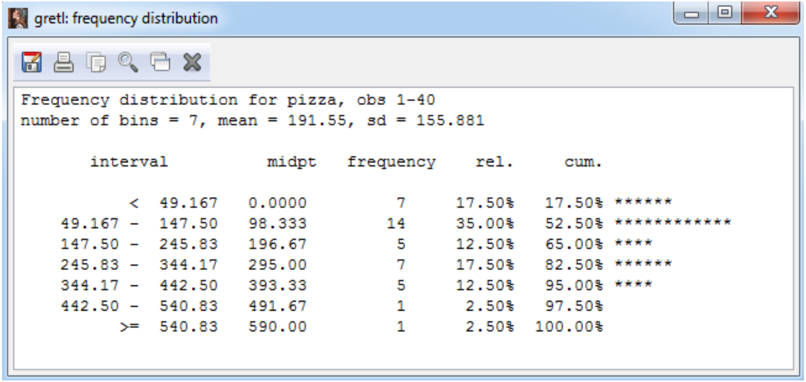
The output consists of a table of numerical results with information on:

- the variable, number of observations and bins, the mean and the standard deviation.
- the bounds of each interval.
- the midpoint of the interval.
- the frequency.
- the relative frequency.
- the cumulative frequency.

and the plot of the frequency distribution.

Example 3.5.2. Frequency distribution. Pizza consumption.

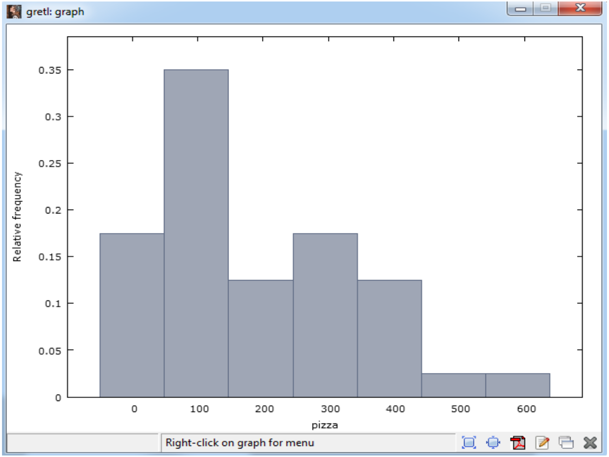
Numerical results (default). Pizza.



```
Frequency distribution for pizza, obs 1-40
number of bins = 7, mean = 191.55, sd = 155.881

       interval          midpt   frequency    rel.     cum.

           <  49.167    0.0000        7      17.50%   17.50% ******
    49.167 -  147.50    98.333       14      35.00%   52.50% ************
    147.50 -  245.83    196.67        5      12.50%   65.00% ****
    245.83 -  344.17    295.00        7      17.50%   82.50% ******
    344.17 -  442.50    393.33        5      12.50%   95.00% ****
    442.50 -  540.83    491.67        1       2.50%   97.50%
          >=  540.83    590.00        1       2.50%  100.00%
```

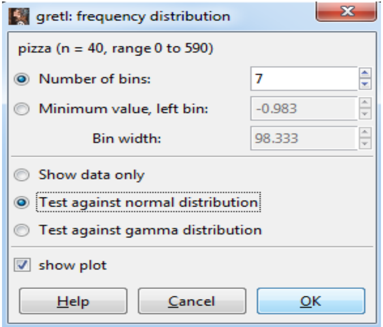
Example 3.5.2. Frequency distribution. Pizza consumption.

Frequency distribution plot (default). Pizza.



Example 3.5.2. Frequency distribution. Pizza consumption.

Frequency distribution with tests.



gretl: frequency distribution

pizza (n = 40, range 0 to 590)

● Number of bins: 7

○ Minimum value, left bin: -0.983

Bin width: 98.333

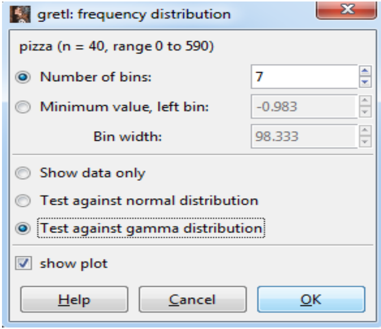
○ Show data only

● Test against normal distribution

○ Test against gamma distribution

☑ show plot

Against the normal (Doornik-Hansen)



gretl: frequency distribution

pizza (n = 40, range 0 to 590)

● Number of bins: 7

○ Minimum value, left bin: -0.983

Bin width: 98.333

○ Show data only

○ Test against normal distribution

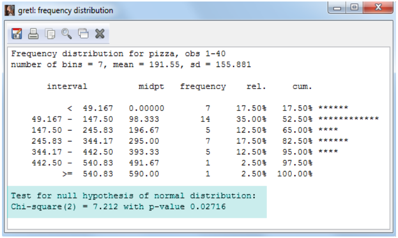
● Test against gamma distribution

☑ show plot

Againts the gamma (Lock)

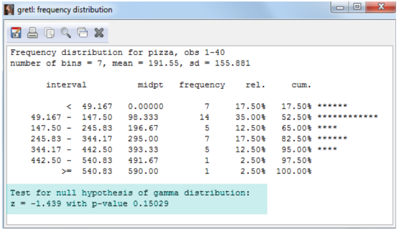
Example 3.5.2. Frequency distribution. Pizza consumption.

The numerical results are shown in the figures below.



```
Frequency distribution for pizza, obs 1-40
number of bins = 7, mean = 191.55, sd = 155.881

       interval          midpt   frequency    rel.     cum.

           <  49.167   0.00000        7     17.50%   17.50% ******
   49.167 -  147.50   98.333        14     35.00%   52.50% ************
   147.50 -  245.83   196.67         5     12.50%   65.00% ****
   245.83 -  344.17   295.00         7     17.50%   82.50% ******
   344.17 -  442.50   393.33         5     12.50%   95.00% ****
   442.50 -  540.83   491.67         1      2.50%   97.50%
          >=  540.83   590.00         1      2.50%  100.00%

Test for null hypothesis of normal distribution:
Chi-square(2) = 7.212 with p-value 0.02716
```

Against the normal (Doornik-Hansen)

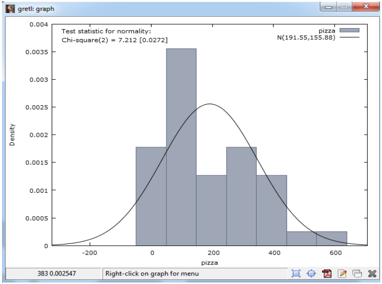


```
Frequency distribution for pizza, obs 1-40
number of bins = 7, mean = 191.55, sd = 155.881

       interval          midpt   frequency    rel.     cum.

           <  49.167   0.00000        7     17.50%   17.50% ******
   49.167 -  147.50   98.333        14     35.00%   52.50% ************
   147.50 -  245.83   196.67         5     12.50%   65.00% ****
   245.83 -  344.17   295.00         7     17.50%   82.50% ******
   344.17 -  442.50   393.33         5     12.50%   95.00% ****
   442.50 -  540.83   491.67         1      2.50%   97.50%
          >=  540.83   590.00         1      2.50%  100.00%

Test for null hypothesis of gamma distribution:
z = -1.439 with p-value 0.15029
```

Against the gamma (Lock)

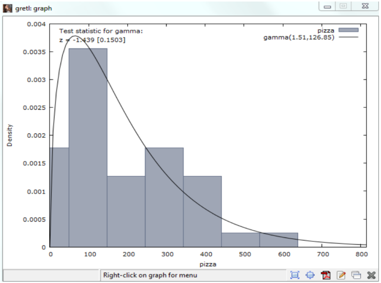
Example 3.5.2. Frequency distribution. Pizza consumption.

The frequency plots appear in the figures below.



Test statistic for normality:
Chi-square(2) = 7.212 [0.0272]

Against the normal (Doornik-Hansen)

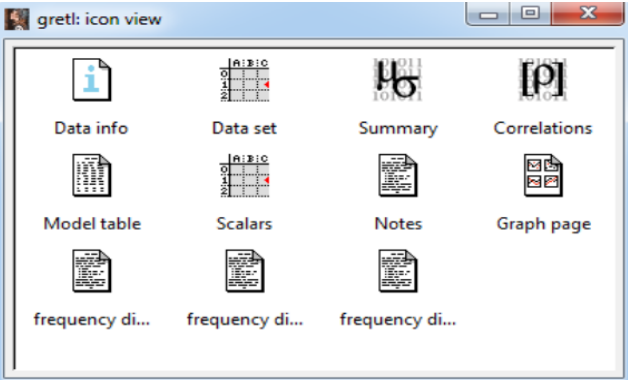


Test statistic for gamma:
z = -1.439 [0.1503]

Against the gamma (Lock)

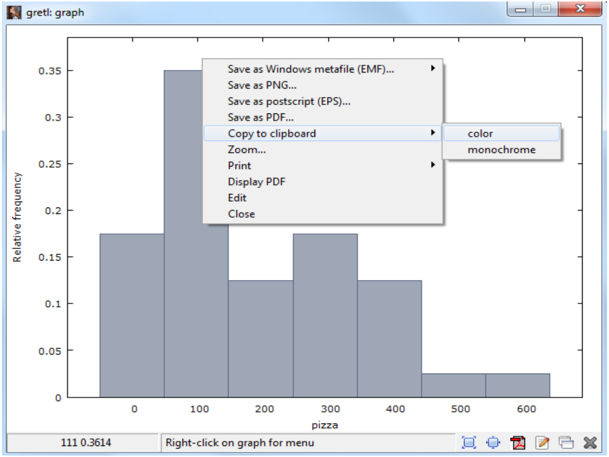
Example 3.5.2. Frequency distribution. Pizza consumption.

To save the results as icons to the session go up to the menu bar in the **frequency distribution** window and click on the left icon. The name by default is *frequency di...*.



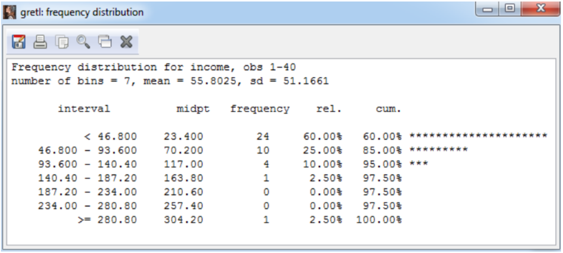
Example 3.5.2. Frequency distribution. Pizza consumption.

To save the plots in Word format, right-click on each plot and select the option *Copy to clipboard* in the pulldown menu. You may choose whether you want to save the plot in colour. Then you can paste the plot to a Word file.

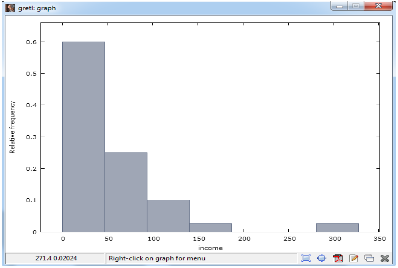


Example 3.5.2. Frequency distribution. Pizza consumption.

The results about the frequency distribution of variable income are shown below.



```
Frequency distribution for income, obs 1-40
number of bins = 7, mean = 55.8025, sd = 51.1661

       interval          midpt    frequency     rel.      cum.

           < 46.800     23.400        24      60.00%    60.00% *********************
    46.800 - 93.600     70.200        10      25.00%    85.00% *********
    93.600 - 140.40     117.00         4      10.00%    95.00% ***
    140.40 - 187.20     163.80         1       2.50%    97.50%
    187.20 - 234.00     210.60         0       0.00%    97.50%
    234.00 - 280.80     257.40         0       0.00%    97.50%
          >= 280.80     304.20         1       2.50%   100.00%
```



Example 3.5.2. Frequency distribution. Pizza consumption.

Results.

- The frequency distribution of the variable pizza consumption shows that 35 % of the individuals in the sample spend around $98.333 in pizza and that 5 % of the individuals spend between 216.45 and 264.55 dollars.
- The variable pizza does not come from a normal distribution but there is evidence in the sample that it comes from a gamma distribution.
- 60 % of the individuals in the sample have an income lower than 23.4 thousands of dollars per year, while there is only one individual in the sample with an income higher than 280.8 thousands of dollars.

Example 3.5.2. Frequency distribution. Visitors to Bilbao.

Questions.

Analyse the frequency distribution of the variables included in the data file `tourism.gdt`.

a. Obtain the frequency distribution for the variable TV. Save the results as an icon to the session and the plot in Word format.

b. Obtain the frequency distribution for the variable foreign visitors.

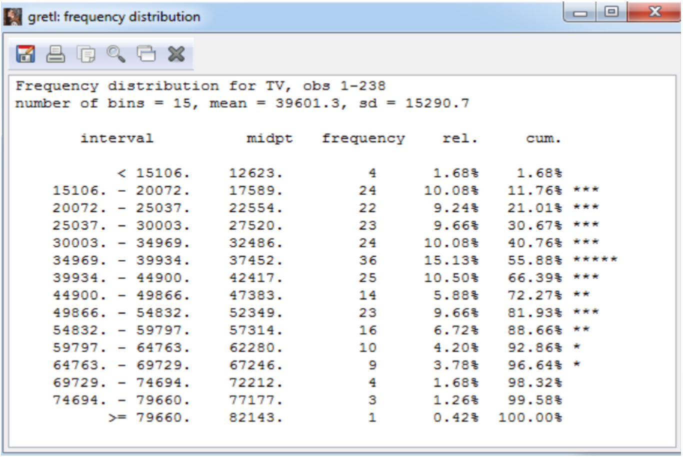
c. Augment the default number of bins by ten units and save the results.

d. Reduce the default number of bins to five units and save the result.

e. Comment on the results.

Example 3.5.2. Frequency distribution. Visitors to Bilbao.

Numerical results (default). Total Visitors.

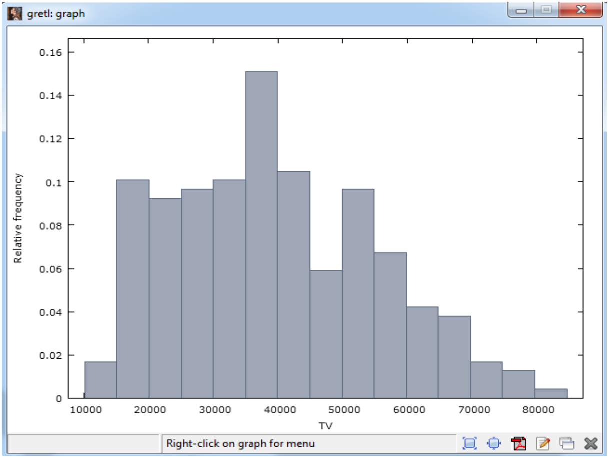


```
Frequency distribution for TV, obs 1-238
number of bins = 15, mean = 39601.3, sd = 15290.7

       interval          midpt   frequency    rel.     cum.

            < 15106.    12623.        4      1.68%    1.68%
    15106. - 20072.    17589.       24     10.08%   11.76% ***
    20072. - 25037.    22554.       22      9.24%   21.01% ***
    25037. - 30003.    27520.       23      9.66%   30.67% ***
    30003. - 34969.    32486.       24     10.08%   40.76% ***
    34969. - 39934.    37452.       36     15.13%   55.88% *****
    39934. - 44900.    42417.       25     10.50%   66.39% ***
    44900. - 49866.    47383.       14      5.88%   72.27% **
    49866. - 54832.    52349.       23      9.66%   81.93% ***
    54832. - 59797.    57314.       16      6.72%   88.66% **
    59797. - 64763.    62280.       10      4.20%   92.86% *
    64763. - 69729.    67246.        9      3.78%   96.64% *
    69729. - 74694.    72212.        4      1.68%   98.32%
    74694. - 79660.    77177.        3      1.26%   99.58%
           >= 79660.    82143.        1      0.42%  100.00%
```

Do not forget to save these results as icons to your session!

Example 3.5.2. Frequency distribution. Visitors to Bilbao.

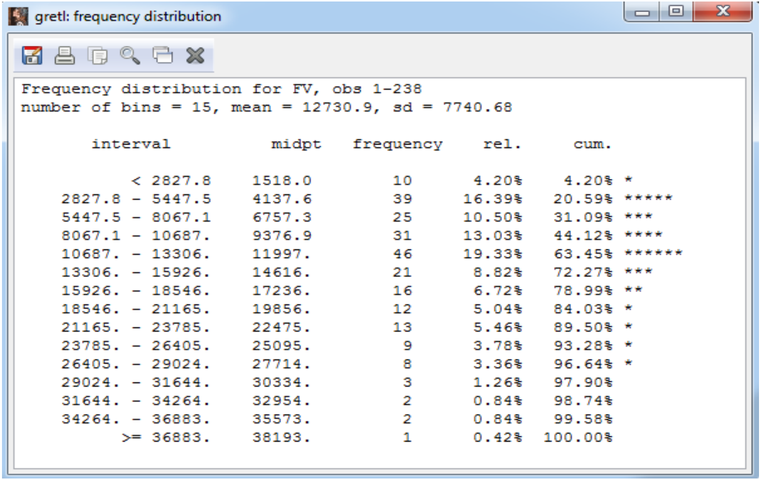
Frequency distribution plot (default). Total visitors.



Do not forget to save this plot in Word format!

Example 3.5.2. Frequency distribution. Visitors to Bilbao.

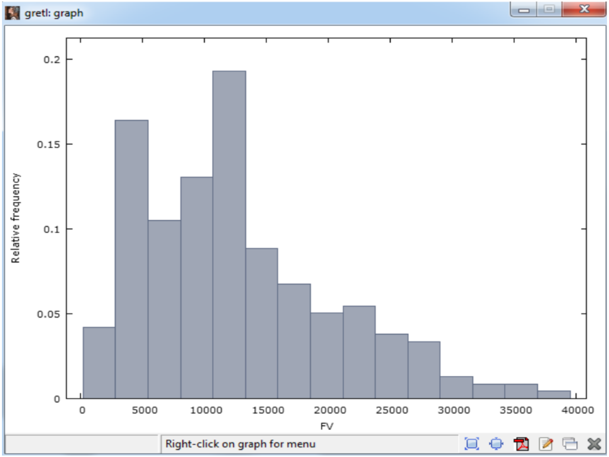
Numerical results (default). Foreign Visitors.



```
Frequency distribution for FV, obs 1-238
number of bins = 15, mean = 12730.9, sd = 7740.68

       interval          midpt   frequency    rel.     cum.

           < 2827.8    1518.0        10      4.20%    4.20% *
    2827.8 - 5447.5    4137.6        39     16.39%   20.59% *****
    5447.5 - 8067.1    6757.3        25     10.50%   31.09% ***
    8067.1 - 10687.    9376.9        31     13.03%   44.12% ****
    10687. - 13306.    11997.        46     19.33%   63.45% ******
    13306. - 15926.    14616.        21      8.82%   72.27% ***
    15926. - 18546.    17236.        16      6.72%   78.99% **
    18546. - 21165.    19856.        12      5.04%   84.03% *
    21165. - 23785.    22475.        13      5.46%   89.50% *
    23785. - 26405.    25095.         9      3.78%   93.28% *
    26405. - 29024.    27714.         8      3.36%   96.64% *
    29024. - 31644.    30334.         3      1.26%   97.90%
    31644. - 34264.    32954.         2      0.84%   98.74%
    34264. - 36883.    35573.         2      0.84%   99.58%
          >= 36883.    38193.         1      0.42%  100.00%
```

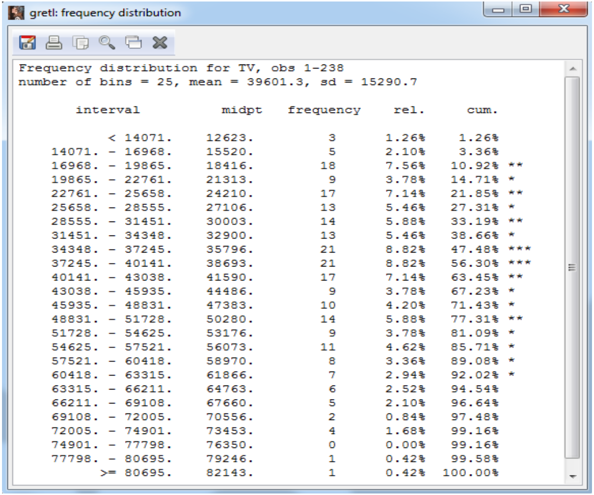
Example 3.5.2. Frequency distribution. Visitors to Bilbao.

Frequency distribution plot (default). Foreign visitors.



Example 3.5.2. Frequency distribution. Visitors to Bilbao.

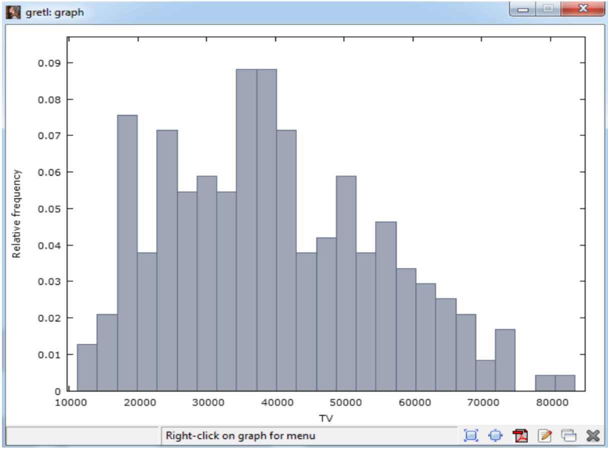
Numerical results (25 bins). Total Visitors.



```
Frequency distribution for TV, obs 1-238
number of bins = 25, mean = 39601.3, sd = 15290.7

       interval          midpt   frequency    rel.     cum.

             < 14071.    12623.         3     1.26%    1.26%
     14071. - 16968.    15520.         5     2.10%    3.36%
     16968. - 19865.    18416.        18     7.56%   10.92% **
     19865. - 22761.    21313.         9     3.78%   14.71% *
     22761. - 25658.    24210.        17     7.14%   21.85% **
     25658. - 28555.    27106.        13     5.46%   27.31% *
     28555. - 31451.    30003.        14     5.88%   33.19% **
     31451. - 34348.    32900.        13     5.46%   38.66% *
     34348. - 37245.    35796.        21     8.82%   47.48% ***
     37245. - 40141.    38693.        21     8.82%   56.30% ***
     40141. - 43038.    41590.        17     7.14%   63.45% **
     43038. - 45935.    44486.         9     3.78%   67.23% *
     45935. - 48831.    47383.        10     4.20%   71.43% *
     48831. - 51728.    50280.        14     5.88%   77.31% **
     51728. - 54625.    53176.         9     3.78%   81.09% *
     54625. - 57521.    56073.        11     4.62%   85.71% *
     57521. - 60418.    58970.         8     3.36%   89.08% *
     60418. - 63315.    61866.         7     2.94%   92.02% *
     63315. - 66211.    64763.         6     2.52%   94.54%
     66211. - 69108.    67660.         5     2.10%   96.64%
     69108. - 72005.    70556.         2     0.84%   97.48%
     72005. - 74901.    73453.         4     1.68%   99.16%
     74901. - 77798.    76350.         0     0.00%   99.16%
     77798. - 80695.    79246.         1     0.42%   99.58%
            >= 80695.    82143.         1     0.42%  100.00%
```

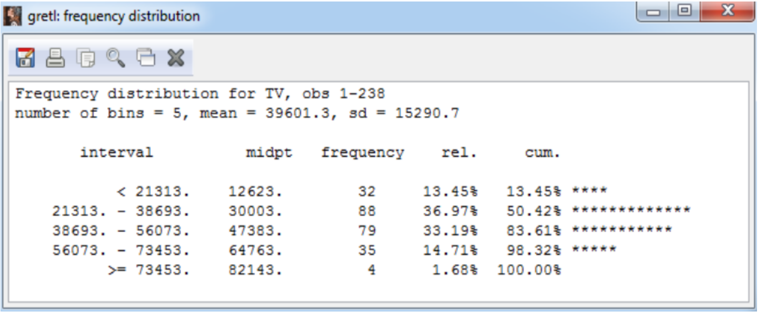
Example 3.5.2. Frequency distribution. Visitors to Bilbao.

Frequency distribution plot (25 bins). Total Visitors.



Example 3.5.2. Frequency distribution. Visitors to Bilbao.

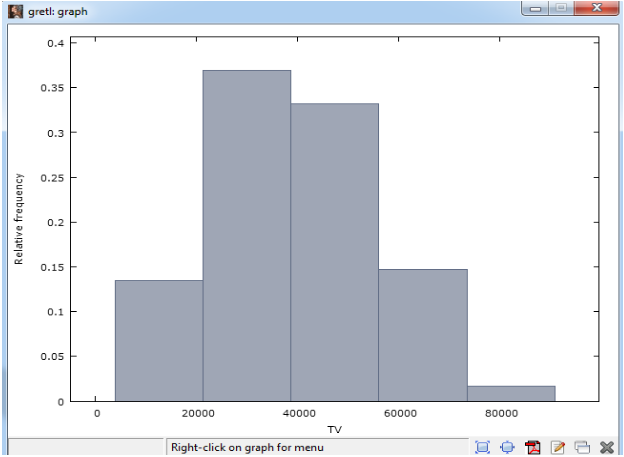
Numerical results (5 bins). Total Visitors.



```
Frequency distribution for TV, obs 1-238
number of bins = 5, mean = 39601.3, sd = 15290.7

       interval          midpt   frequency    rel.     cum.

           < 21313.    12623.        32     13.45%   13.45% ****
    21313. - 38693.    30003.        88     36.97%   50.42% *************
    38693. - 56073.    47383.        79     33.19%   83.61% ***********
    56073. - 73453.    64763.        35     14.71%   98.32% *****
          >= 73453.    82143.         4      1.68%  100.00%
```

Example 3.5.2. Frequency distribution. Visitors to Bilbao.

Frequency distribution plot (5 bins). Total Visitors.



Example 3.5.2. Frequency distribution. Visitors to Bilbao.

Results.

- The frequency distribution of the variable FV suggests that it may come from a bimodal distribution. The two modes contain 18.25 % of the foreign visitors. It seems that there are two relevant points in time for the arrival of foreign visitors.
- The frequency distribution of the variable TV obtained with 25 bins shows that the first part of the frequency distribution is not as homogeneous as it looked like when calculated with only 15 bins.
- On the other hand, the frequency distribution of the variable TV obtained with 5 bins is too smooth. It is not possible to see any detail and it seems much more homogeneous than it really is.
- It is recommended to use the default value of bins if you are not an expert in calculating the optimal number.

Example 3.5.2. Frequency distribution. Simulation.

Questions.

a. Simulate cross-section data: 50 observations from a normal distribution with zero mean and standard deviation 1. Denote this variable by VN. Save the simulated data in the file `Sim50.gdt`.

  a.1. Plot the standard normal distribution and save the graph in pdf format.

  a.2. Obtain the frequency distribution for the variable VN and run the the tests against the normal and gamma distributions. Save the graphs in pdf format.

b. Simulate cross-section data: 1000 observations from a Chi-squared distribution with 10 degrees of freedom. Denote this variable by VX. Save the simulated data in the file `Sim1000.gdt`.

  b.1. Plot the selected Chi-squared distribution and save the graphs in pdf format.

  b.2. Obtain the frequency distribution for the variable VX and run the the tests against the normal and gamma distributions. Save the graphs in pdf format.

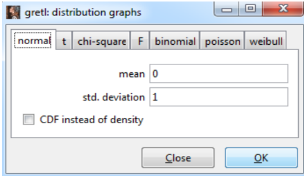
c. Comment on the results.

Example 3.5.2. Frequency distribution. Simulation.

Look up Example 3.2.4 to refresh how to simulate data.

To plot the standard normal distribution, go up to the Menu Bar and click

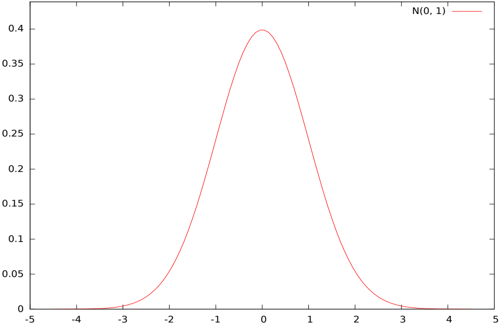
`Tools -> Distribution graphs`


gretl: distribution graphs

mean 0

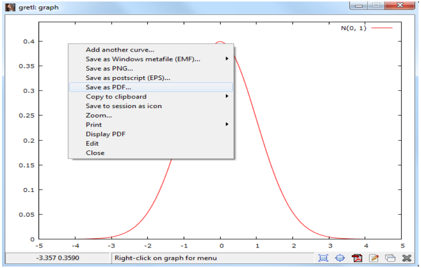
std. deviation 1

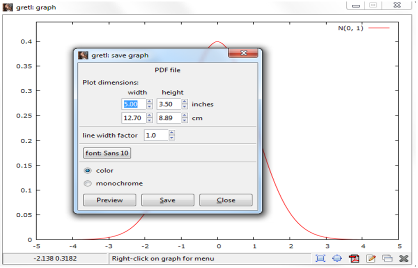
□ CDF instead of density



Example 3.5.2. Frequency distribution. Simulation.

To save this plot, right-click on the graph and select the option *Save as PDF...* from the pulldown menu. This yields a dialog box that enables us to edit the graph.





PDF file

Plot dimensions:

	width	height	
	5.00	3.50	inches
	12.70	8.89	cm

line width factor 1.0

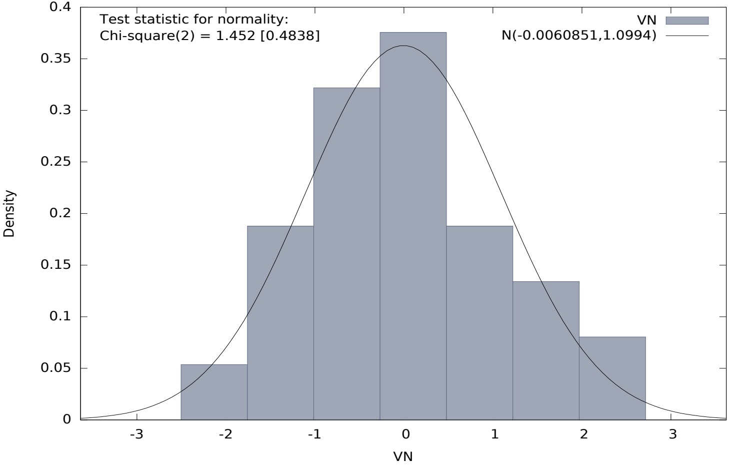
font: Sans 10

● color

○ monochrome

Example 3.5.2. Frequency distribution. Simulation.

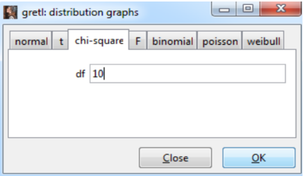
Frequency distribution plots and normality test (Doornik-Hansen). VN.



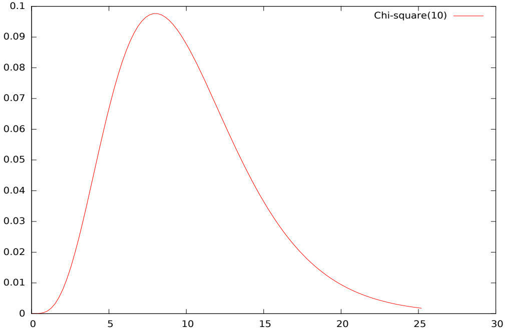
Do not forget to save this plot in pdf format!

Example 3.5.2. Frequency distribution. Simulation.

$\chi^2(10)$ distribution plot.



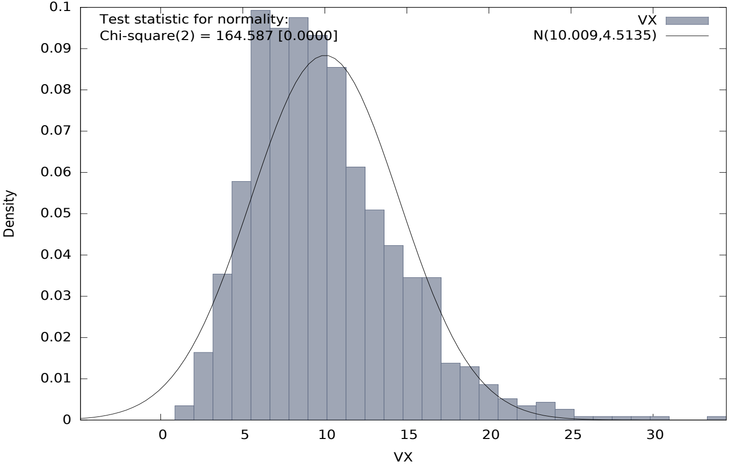
df 10



Do not forget to save this plot in pdf format!

Example 3.5.2. Frequency distribution. Simulation.

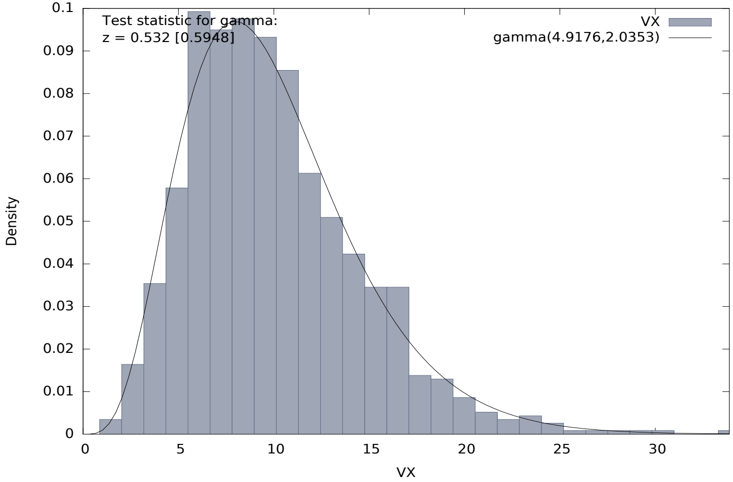
Frequency distribution plot and normality test (Doornik-Hansen). VX.



Do not forget to save this plot in pdf format!

Example 3.5.2. Frequency distribution. Simulation.

Frequency distribution plot and gamma test (Lock). VX.



Do not forget to save this plot in pdf format!

Example 3.5.2. Frequency distribution. Simulation.

Results.

- Comparing the plot of the standard normal distribution and the frequency distribution of the variable VN, it may be concluded that they do not look very similar. Nevertheless, the p-value of the normality test is larger than 0.05 meaning that there is evidence in the sample that the variable VN comes from a normal distribution at the 5 % significance level. It has not been possible to carry out the test against the gamma distribution because Gretl detects that there are some negative values in the sample and therefore the test makes no sense.

- Analysing the frequency distribution of the variable VX, it may observed that it does not look like a normal distribution but it is quite similar to a gamma distribution. The tests reach the same conclusion: the p-value obtained in the normality test is smaller than 0.05 and the p-value obtained in the gamma test is larger than 0.05. Therefore, there is evidence in the sample that the variable VX comes from a gamma distribution at the 5 % significance level. This result was expected since the Chi-squared distribution belongs to the class of gamma distributions.

- Bear in mind that your simulated data are going to be different from ours. Therefore, you will obtain different frequency distributions and your conclusions might not be the same.

Contents

1. 3.5.1. Normality tests.
   - Pizza consumption.
   - Visitors to Bilbao.
2. 3.5.2. Frequency distribution.
   - Pizza consumption.
   - Visitors to Bilbao.
   - Simulation.
3. 3.5.3. Estimated density plot.
   - Pizza consumption.
   - Visitors to Bilbao.
4. 3.5.4. Boxplot.
5. 3.5.5. Normal Q-Q plot.
   - Visitors to Bilbao.
   - Simulation.

Example 3.5.3. Estimated density plot. Pizza consumption.

Questions.

Load the data file `pizza.gdt`.

a. Estimate the density function of the variables pizza and income.

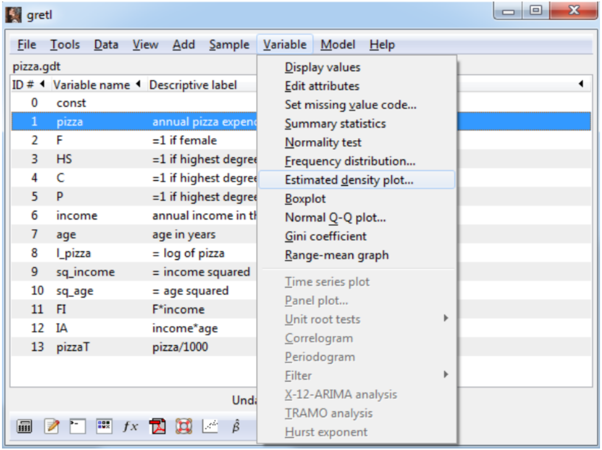
b. Estimate the density function of the variable age using the Gaussian and the Epanechnikov kernels.

c. Comment on the results. Save the session.

Example 3.5.3. Estimated density plot. Pizza consumption.

To obtain the estimated density plot, select the option *Estimated density plot...* from the **Variable** pulldown menu.

```
Variable -> Estimated density plot ...
```

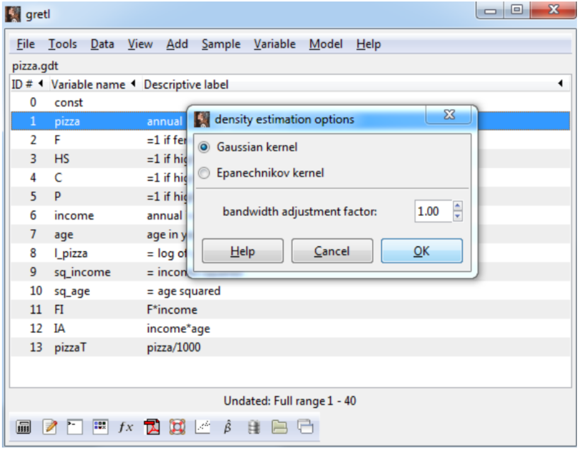


pizza.gdt

ID #	Variable name	Descriptive label
0	const	
1	pizza	annual pizza expend
2	F	=1 if female
3	HS	=1 if highest degree
4	C	=1 if highest degree
5	P	=1 if highest degree
6	income	annual income in th
7	age	age in years
8	l_pizza	= log of pizza
9	sq_income	= income squared
10	sq_age	= age squared
11	FI	F*income
12	IA	income*age
13	pizzaT	pizza/1000

Example 3.5.3. Estimated density plot. Pizza consumption.

The dialog box offers two kernel options.



pizza.gdt

ID #	Variable name	Descriptive label
0	const	
1	pizza	annual
2	F	=1 if fe
3	HS	=1 if hi
4	C	=1 if hi
5	P	=1 if hi
6	income	annual
7	age	age in y
8	l_pizza	= log o
9	sq_income	= incon
10	sq_age	= age squared
11	FI	F*income
12	IA	income*age
13	pizzaT	pizza/1000

density estimation options

● Gaussian kernel

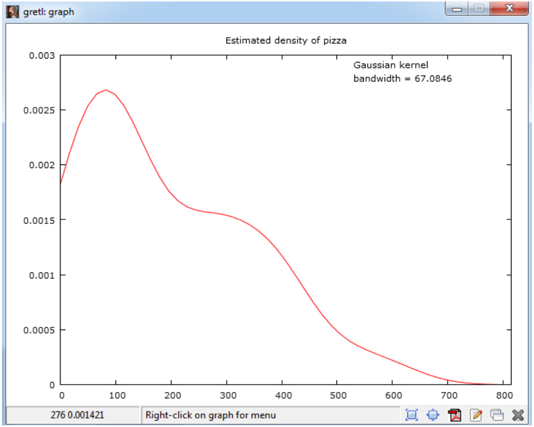
○ Epanechnikov kernel

bandwidth adjustment factor: 1.00

Undated: Full range 1 - 40

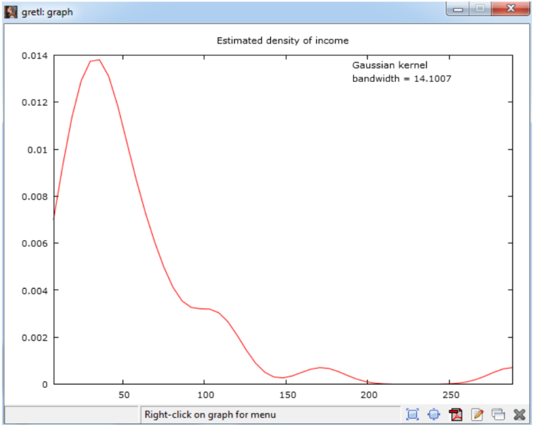
Example 3.5.3. Estimated density plot. Pizza consumption.

Estimated density plot (default bandwidth). *pizza*. Gaussian kernel.



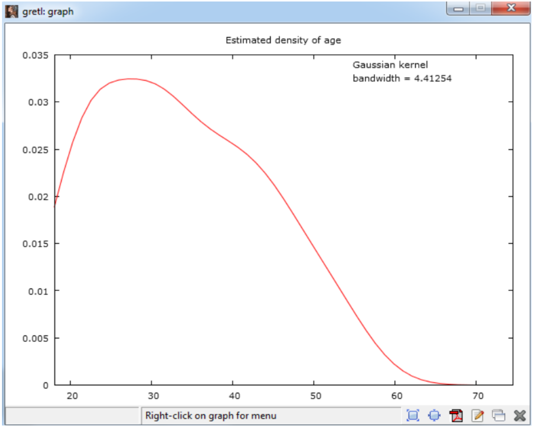
Example 3.5.3. Estimated density plot. Pizza consumption.

Estimated density plot (default bandwidth). *income*. Gaussian kernel.



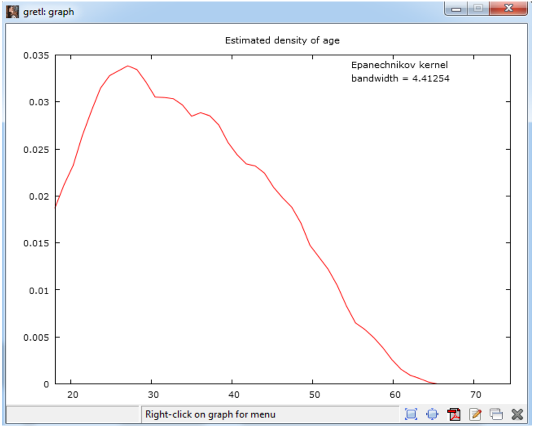
Example 3.5.3. Estimated density plot. Pizza consumption.

Estimated density plot (default bandwidth). *age*. Gaussian kernel.



Example 3.5.3. Estimated density plot. Pizza consumption.

Estimated density plot (default bandwidth). *age*. Epanechnikov kernel.



Example 3.5.3. Estimated density plot. Pizza consumption.

Results.

- Analysing the estimated density functions of the variables pizza and income, it may be concluded that both are asymmetric to the left, that is the values close to the ordinate are more probable.

  In the case of the variable pizza, there is a second mode, a smaller one, around \$350. This density function seems to be bimodal.
  In the case of the variable income, there is a high probability of being around 50 thousand of dollars while the probability of income being higher than 100 thousands of dollars is quite low.

- The Gaussian and Epanechnikov kernels produce the same results asymptotically. However, comparing the estimated density functions of the variable age using these two kernels, it may be observed that the estimated density function obtained using the Gaussian kernel is smoother. This result is due to the sample size and the automatic selection of the bandwidth h.

Example 3.5.3. Estimated density plot. Visitors to Bilbao.

Questions.

Load the data file `tourism.gdt`.

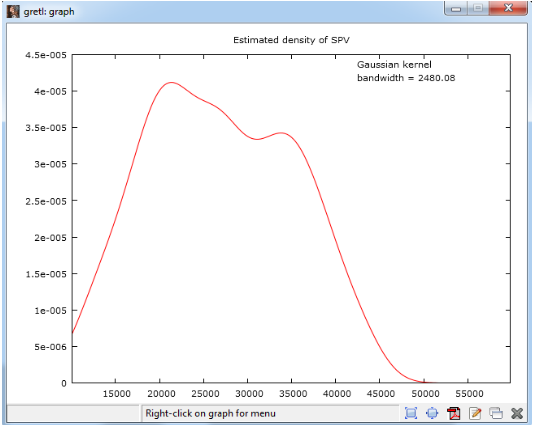
a. Estimate the density function of the variables SPV and FV using the Gaussian kernel.

b. Estimate the density function of the variable $IPIR$ using the Gaussian kernel for several bandwidths: the default, the minimum and the maximum.

c. Discuss and compare the results. Save the session.

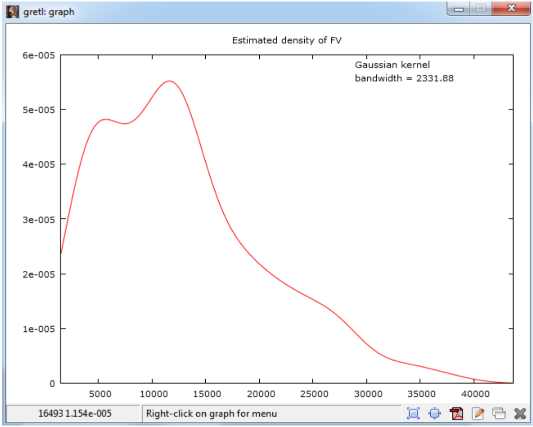
Example 3.5.3. Estimated density plot. Visitors to Bilbao.

Estimated density plot (default bandwith). SPV. Gaussian kernel.



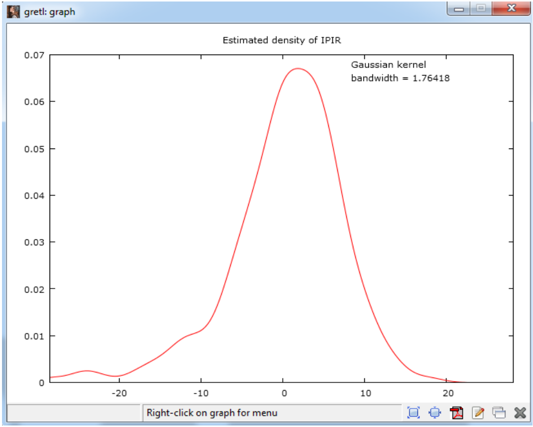
Example 3.5.3. Estimated density plot. Visitors to Bilbao.

Estimated density plot (default bandwith). FV. Gaussian kernel.



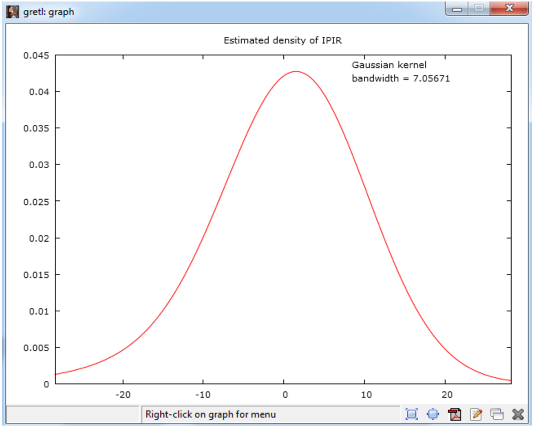
Example 3.5.3. Estimated density plot. Visitors to Bilbao.

Estimated density plot (default bandwith). $IPIR$. Gaussian kernel.



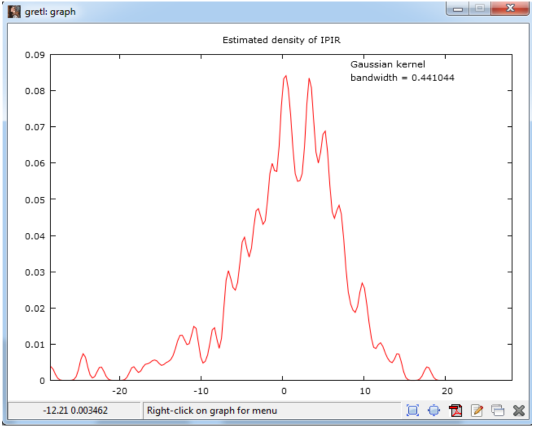
Example 3.5.3. Estimated density plot. Visitors to Bilbao.

Estimated density plot (maximum bandwith). $IPIR$. Gaussian kernel.



Example 3.5.3. Estimated density plot. Visitors to Bilbao.

Estimated density plot (minimum bandwith). $IPIR$. Gaussian kernel.



Example 3.5.3. Estimated density plot. Visitors to Bilbao.

Results.

- The estimated density functions of the variables SPV and FV are both bimodal. In the case of SPV the two modes are quite different, one being much higher than the other. In the case of FV both modes are similar. Comparing both estimated density functions, it may be concluded that the probability of having a high number of visitors is bigger when they come from Spain.

- Comparing the three density functions estimated for $IPIR$, it may be concluded that the maximum bandwidth makes the estimated density function look Gaussian while the minimum bandwidth produces a much more irregular density function. Therefore, if you are not an expert in choosing the bandwidth, it is recommended to use the default bandwidth calculated by Gretl.

Contents

1. 3.5.1. Normality tests.
   - Pizza consumption.
   - Visitors to Bilbao.
2. 3.5.2. Frequency distribution.
   - Pizza consumption.
   - Visitors to Bilbao.
   - Simulation.
3. 3.5.3. Estimated density plot.
   - Pizza consumption.
   - Visitors to Bilbao.
4. 3.5.4. Boxplot.
5. 3.5.5. Normal Q-Q plot.
   - Visitors to Bilbao.
   - Simulation.

Example 3.5.4. Boxplot.

Questions.

Load the data file `pizza.gdt` into Gretl to obtain the boxplots for several variables.

a. Obtain the simple boxplot for the variable pizza, including the confidence interval for the median.

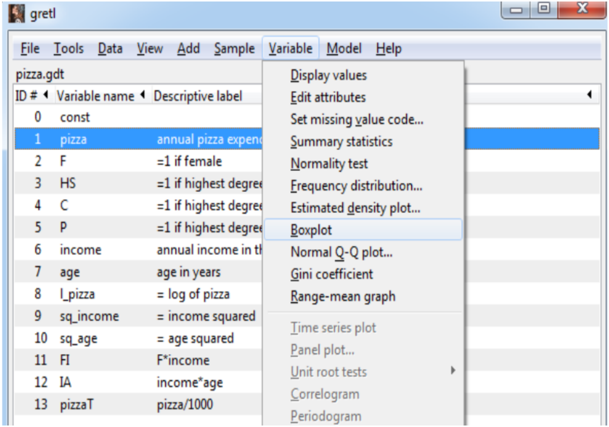
b. Obtain the simple boxplot for the variables pizza, income and age.

c. Obtain a factorized boxplot for the variable pizza using the variable F as factor.

d. Comment on the results.

Example 3.5.4. Boxplot.

To obtain a boxplot, highlight the variable of interest, go up to the Menu Bar and click **`Variable -> Boxplot`**



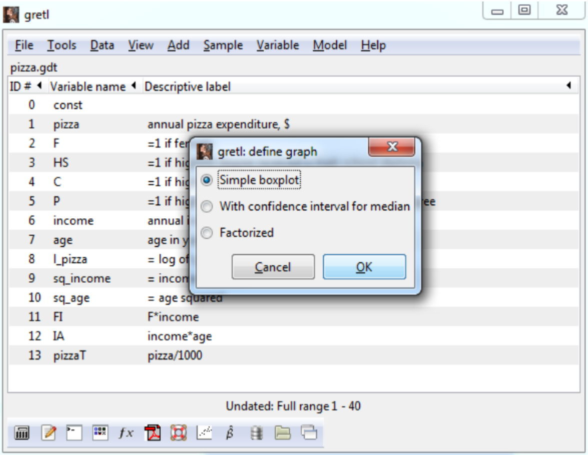
pizza.gdt

ID #	Variable name	Descriptive label
0	const	
1	pizza	annual pizza expend
2	F	=1 if female
3	HS	=1 if highest degree
4	C	=1 if highest degree
5	P	=1 if highest degree
6	income	annual income in th
7	age	age in years
8	l_pizza	= log of pizza
9	sq_income	= income squared
10	sq_age	= age squared
11	FI	F*income
12	IA	income*age
13	pizzaT	pizza/1000

Alternatively, highlight the variable of interest, right-click and select the option *Boxplot* from the pulldown menu.

Example 3.5.4. Boxplot.

The dialog box offers three boxplot options.



pizza.gdt

ID #	Variable name	Descriptive label
0	const	
1	pizza	annual pizza expenditure, $
2	F	=1 if fe
3	HS	=1 if hig
4	C	=1 if hig
5	P	=1 if hig ee
6	income	annual i
7	age	age in y
8	l_pizza	= log of
9	sq_income	= incon
10	sq_age	= age squared
11	FI	F*income
12	IA	income*age
13	pizzaT	pizza/1000

gretl: define graph

● Simple boxplot

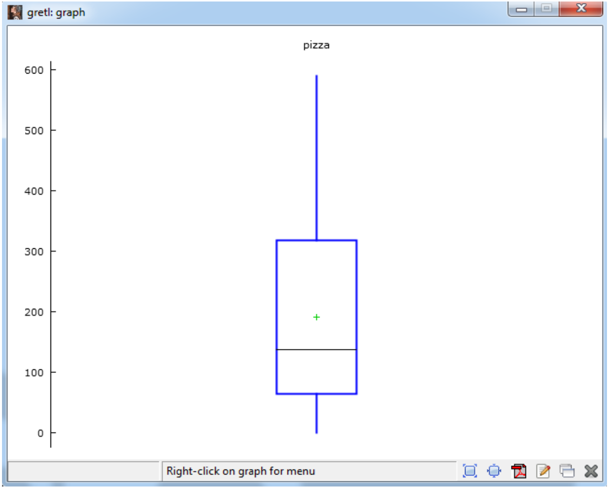
○ With confidence interval for median

○ Factorized

Undated: Full range 1 - 40

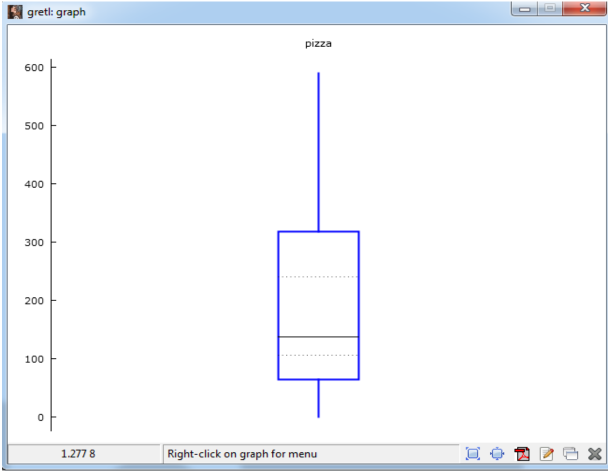
Example 3.5.4. Boxplot.

Simple boxplot. *pizza*.



Example 3.5.4. Boxplot.

Boxplot with confidence interval for the median. *pizza*.

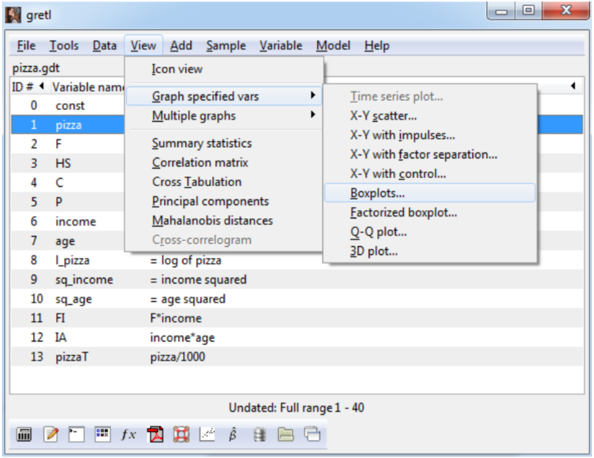


It should be noted that there is no information about the sample mean in this boxplot.

Example 3.5.4. Boxplot.

To obtain a boxplot for several variables on a single graph, go up to the Menu Bar and click

```
View -> Graph specified vars   -> Boxplots...
```



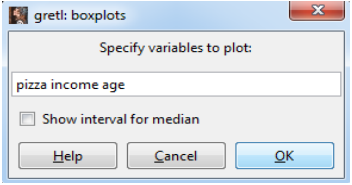
pizza.gdt

ID #	Variable name	
0	const	
1	pizza	
2	F	
3	HS	
4	C	
5	P	
6	income	
7	age	
8	l_pizza	= log of pizza
9	sq_income	= income squared
10	sq_age	= age squared
11	FI	F*income
12	IA	income*age
13	pizzaT	pizza/1000

Undated: Full range 1 - 40

Example 3.5.4. Boxplot.

Select the desired variables: pizza, income and age.



gretl: boxplots

Specify variables to plot:

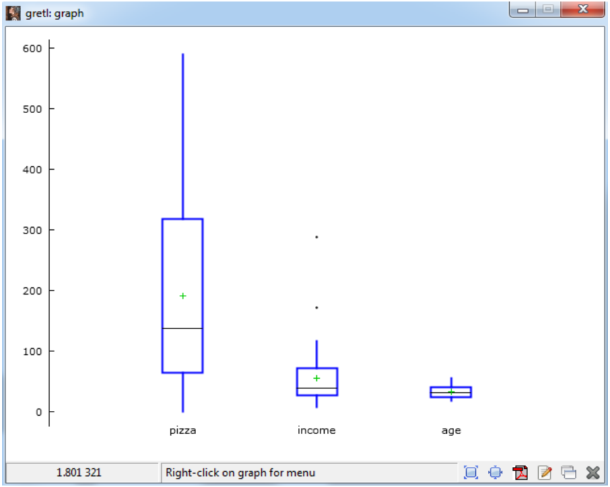
pizza income age

□ Show interval for median

If you want to include the confidence intervals for the medians, mark show interval for median.

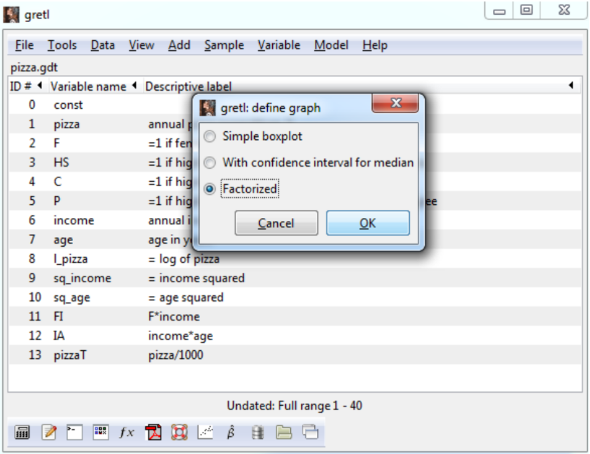
Example 3.5.4. Boxplot.

Boxplot for specified variables.



Example 3.5.4. Boxplot.

Select the third option in the dialog box to obtain a factorized boxplot.


gretl
File Tools Data View Add Sample Variable Model Help
Undated: Full range 1 - 40

pizza.gdt

ID #	Variable name	Descriptive label
0	const	
1	pizza	annual p
2	F	=1 if fen
3	HS	=1 if hig
4	C	=1 if hig
5	P	=1 if hig ee
6	income	annual i
7	age	age in y
8	l_pizza	= log of pizza
9	sq_income	= income squared
10	sq_age	= age squared
11	FI	F*income
12	IA	income*age
13	pizzaT	pizza/1000

gretl: define graph

○ Simple boxplot

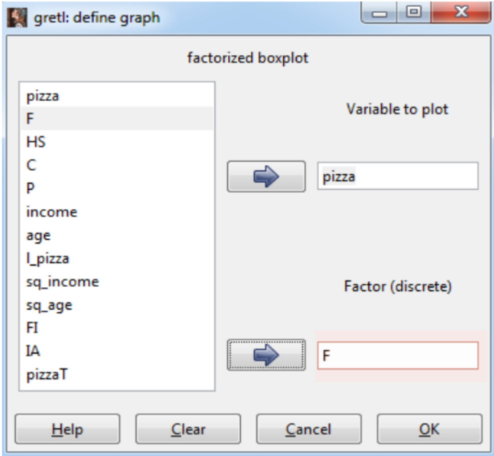
○ With confidence interval for median

● Factorized

Cancel OK

Example 3.5.4. Boxplot.

Select the variable to plot and the variable that will be used as factor. **It has to be a discrete variable.** In this case, select F (female).



gretl: define graph

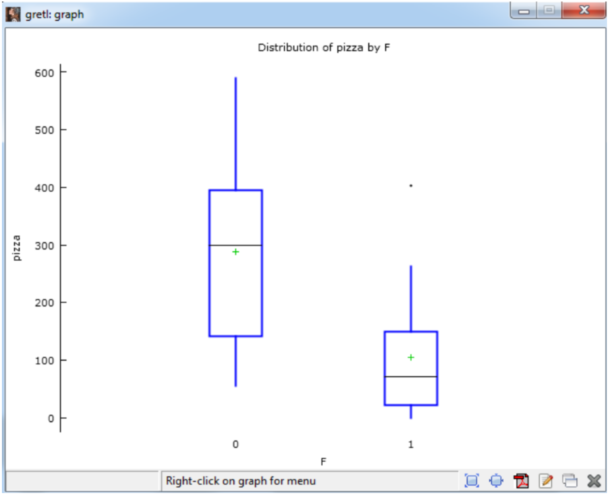
factorized boxplot

Variable to plot: pizza

Factor (discrete): F

Example 3.5.4. Boxplot.

Factorized boxplot (factor F). *pizza*.



Example 3.5.4. Boxplot.

Results.

- Looking at the first boxplot of pizza, it may be concluded than the median is smaller than the sample mean. But it has to be taken into account that only the point estimates of the median and the mean are plotted. The second boxplot of variable pizza gives more information because it includes the confidence interval for the median. In this graph, it can be observed that the sample mean is inside the confidence interval for the median. Therefore, there are not significant differences between the sample mean and the median.
- When the boxplots of pizza, income and age are shown on the same graph, it is difficult to compare the results because the units of measurement are quite different: dollars, thousands of dollars and years. Therefore, it may be concluded that it is only sensible to represent boxplots of several variables on a single graph if the units of measurement are similar.
- Analysing the factorized boxplot of pizza using variable F (1 if female) as factor, it may be concluded that women consume less pizza than men: both the sample mean and the median are smaller for women, the first quartile for women is similar to the third quartile for men and the IQ range is much narrower for women than for men.

Example 3.5.4. Factorized boxplots.

Questions.

Load the data file `pizza.gdt` into Gretl.

a. Create three discrete variables depending on age as follows:
   - $group1$: clients who are less than thirty one years old.
   - $group2$: clients who are over thirty years old and less than forty one years old.
   - $group3$: clients who are over forty years old.

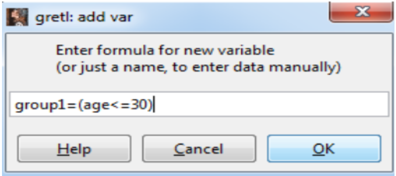
b. Obtain a factorized boxplot for the variable pizza, using as factor the discrete variable $group1$.

c. Obtain a factorized boxplot for the variable income, using as factor the discrete variable $group3$.

d. Comment on the results.

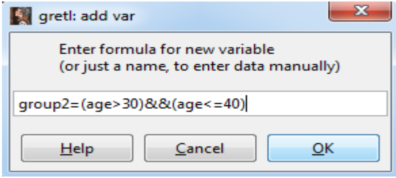
Example 3.5.4. Factorized boxplots.

To create the discrete variables $group1$, $group2$ and $group3$, go up to the Menu Bar and click **`Add -> Define new variable`**.



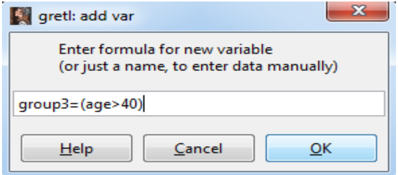
Enter formula for new variable
(or just a name, to enter data manually)

group1=(age<=30)



Enter formula for new variable
(or just a name, to enter data manually)

group2=(age>30)&&(age<=40)



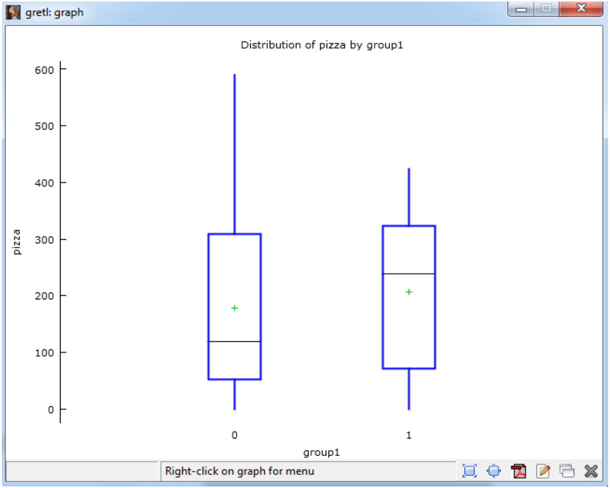
Enter formula for new variable
(or just a name, to enter data manually)

group3=(age>40)

Edit the attributes of these variables and mark Treat this variable as discrete.

Example 3.5.4. Factorized boxplots.

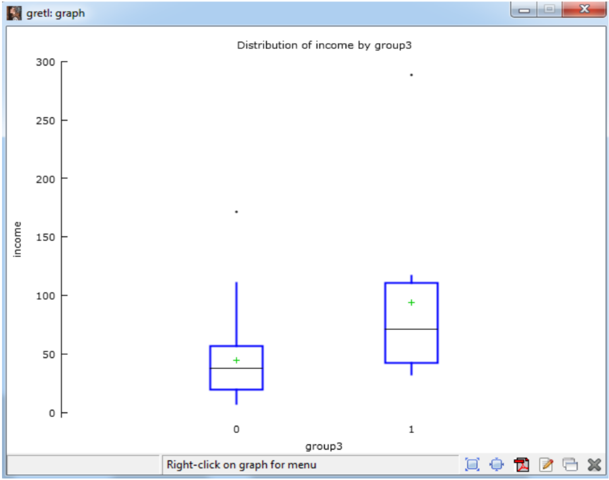
Factorized boxplot (factor $group1$). *pizza*.



gretl: graph
Distribution of pizza by group1
600
500
400
300
200
100
0
pizza
0
1
group1
Right-click on graph for menu

Example 3.5.4. Factorized boxplots.

Factorized boxplot (factor $group3$). *income*.



Example 3.5.4. Factorized boxplots.

Results.

- The quantitative variable age is used to factorize the boxplot. Given that the factor must be a discrete variable, it is necessary to transform the quantitative variable into a discrete variable. This is done by splitting the variable age into three intervalas defining three discrete variables: $group1$ (less than 31 years old), $group2$ (over 30 years old and less than 41) and $group3$ (over forty years old).

- Analysing the factorized boxplot of pizza obtained using $group1$ as factor, it may be observed that the clients less than 31 years old consume more pizza than the rest of the individuals in the whole sample (over 30 years old clients): the mean, the median, the first and the third quartiles are larger for young clients.

- Analysing the factorized boxplot of income obtained using $group3$ as factor, it may be observed that the clients over 40 years old have higher income than the rest of the individuals in the whole sample: the mean, the median, the first and the third quartiles are larger for these clients.

Example 3.5.4. Boxplot. Discrete variables.

Questions.

Open the data file `pizza.gdt`:

a. Create the discrete variable $groups$ that takes the value 1 if the client belongs to $group1$, the value 2 if the client belongs to $group2$ and the value 3 if the client belongs to $group3$.

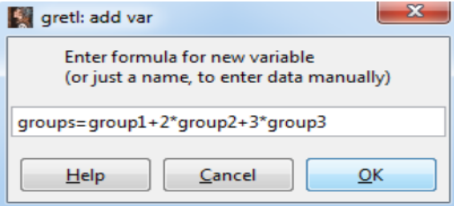
- $group1$: clients who are less than 31 years old.
- $group2$: clients who are over thirty years old and less than forty one years old.
- $group3$: clients who are over forty years old.

b. Obtain a factorized boxplot of pizza and income, using the discrete variable $groups$ as factor.

c. Comment on the results. Save the changes in the data and the Gretl session.

Example 3.5.4. Boxplot. Discrete variables.

To create the discrete variable $groups$ that takes the value 1 if the client belongs to $group1$, the value 2 if the client belongs $group2$ and the value 3 if the client belongs to $group3$, select the option *Define a new variable* from the **Add** pulldown menu.

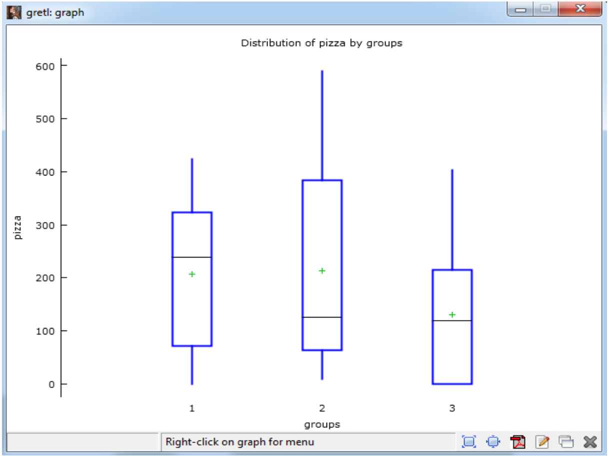

gretl: add var
Enter formula for new variable
(or just a name, to enter data manually)
Help
Cancel
OK

groups=group1+2*group2+3*group3

Edit the attributes of this variable and mark Treat this variable as discrete.

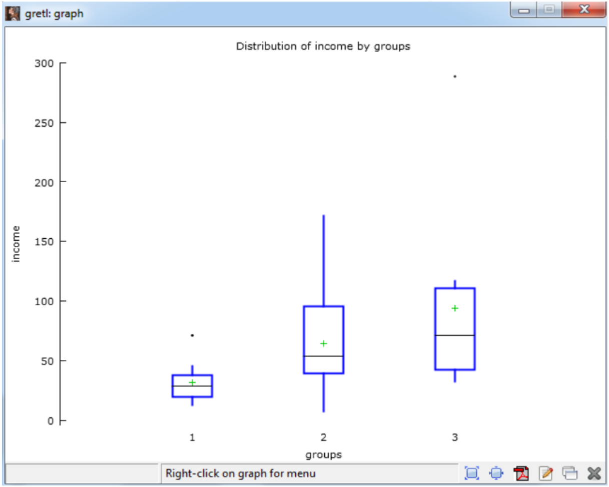
Example 3.5.4. Boxplot. Discrete variables.

Factorized boxplot (factor $groups$). *pizza*.



Example 3.5.4. Boxplot. Discrete variables.

Factorized boxplot (factor $groups$). *income*.


gretl: graph
Distribution of income by groups
income
300
250
200
150
100
50
0
1
2
3
groups
Right-click on graph for menu

Example 3.5.4. Boxplot. Discrete variables.

Results (I).

Consider the new variable $groups$ as a factor.

- The three discrete variables created before are not useful.
- We have to generate a new discrete variable that encompasses the three groups considered. Thus, $groups$ takes the value 1 for the clients less than 31 years old, the value 2 for the clients over 30 years old and less than 41, and the value 3 for the clients over 40 years old.
- We could have chosen any other numerical values for the variable $groups$, for instance, 1, 5 and 7. It is only required to assign a different value to each category.

Example 3.5.4. Boxplot. Discrete variables.

Results (II).

Factorized boxplot of pizza using $groups$ as factor.

- The clients over 40 years old consume the least pizza. It may be observed than the first and the third quartiles are below those of the rest of the sample. Nevertheless, since the median of this group is close to the median of $group2$, it would be advisable to analyse the confidence intervals.
- The sample means of the individuals in $group1$ and $group2$ are very similar.
- The median for clients less than 31 years old is clearly larger than the median for the rest of the sample. On the other hand, it is close to the median of $group2$. Even though the first quartile is slightly larger than the one corresponding to $group2$, we do not get the same result for the third quartile.
- Bearing in mind the last two comments, we should consider whether we have chosen the most appropriate intervals to split the variable age. A better option might have been to define the first group as the clients less than 33/35 years old. To do that we should redefine the variable $groups$ and repeat the analysis.

Example 3.5.4. Boxplot. Discrete variables.

Results (III).

Factorized boxplot of income using $groups$ as factor.

- The clients less than 31 years old have the smallest incomes: the sample mean, the median and the third quartile are much smaller than for the rest of the groups. As a matter of fact, the third quartile is quite similar to the first quartile of $group1$ and $group2$.
- The IQ range of $group1$ is the smallest, meaning that the income of these individuals is quite similar. There is only one individual (it may be considered an outlier) with an income close to the sample mean and the median of $group2$.
- Looking at the minimum and maximum values of income, it may be concluded that the individuals in $group2$ present the highest variability. Both the mean and the median of this group is slightly smaller than the mean and the median of $group3$. However, in order to analyse the significance of this result it would be necessary to use other tools, for instance, the confidence intervals for the median.

Example 3.5.4. Boxplot. Discrete variables.

Results (IV).

Factorized boxplot of income using $groups$ as a factor.

- It may be observed that, in general, the income of the clients in $group3$ is higher than the income of clients in $group2$: both the median and the sample mean are higher in $group3$ than in $group2$.
- The individual with the largest income in the sample is less than forty one years old. His/her income is much higher than the sample mean.
- Same as before, it might be advisable to analyse if the age intervals considered here are adequate. It might be better to redefine the last two age groups.

It may be concluded that it would be better to split the variable age into different intervals for the variables pizza and income.

Contents

1. 3.5.1. Normality tests.
   - Pizza consumption.
   - Visitors to Bilbao.
2. 3.5.2. Frequency distribution.
   - Pizza consumption.
   - Visitors to Bilbao.
   - Simulation.
3. 3.5.3. Estimated density plot.
   - Pizza consumption.
   - Visitors to Bilbao.
4. 3.5.4. Boxplot.
5. 3.5.5. Normal Q-Q plot.
   - Visitors to Bilbao.
   - Simulation.

Example 3.5.5. Normal Q-Q plot. Visitors to Bilbao.

Questions.

Open the data file `tourism.gdt`.

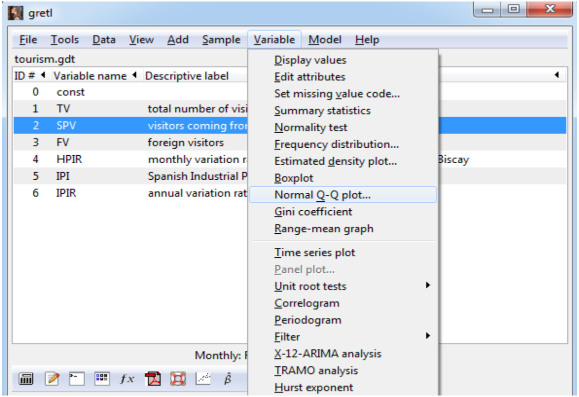
a. Obtain the Q-Q plot for the the variables SPV and FV.

b. Comment on the results.

Example 3.5.5. Normal Q-Q plot. Visitors to Bilbao.

To obtain the normal Q-Q plot, highlight the variable of interest, go up to the Menu Bar and click

```
Variable -> Normal Q-Q plot...
```



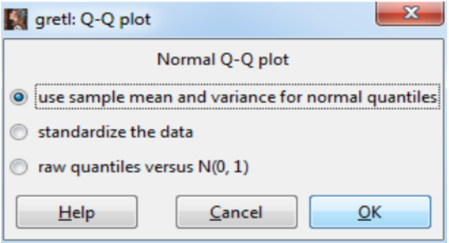
tourism.gdt

ID #	Variable name	Descriptive label
0	const	
1	TV	total number of visi
2	SPV	visitors coming fro
3	FV	foreign visitors
4	HPIR	monthly variation r ... Biscay
5	IPI	Spanish Industrial P
6	IPIR	annual variation rat

Monthly: F

Example 3.5.5. Normal Q-Q plot. Visitors to Bilbao.

In the dialog box, you have to select an option for the plot axis.



gretl: Q-Q plot

Normal Q-Q plot

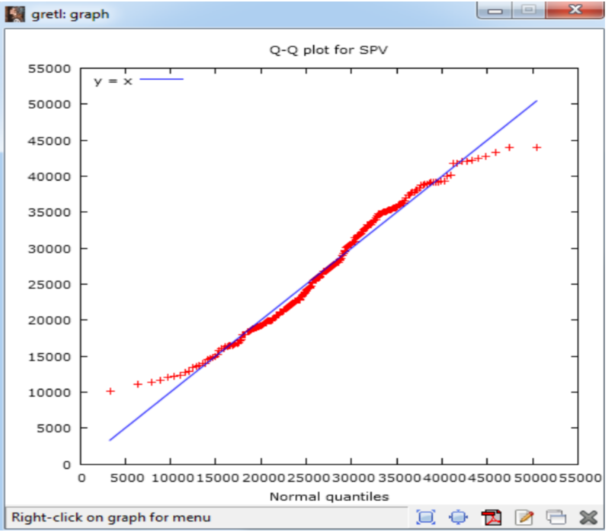
● use sample mean and variance for normal quantiles

○ standardize the data

○ raw quantiles versus N(0, 1)

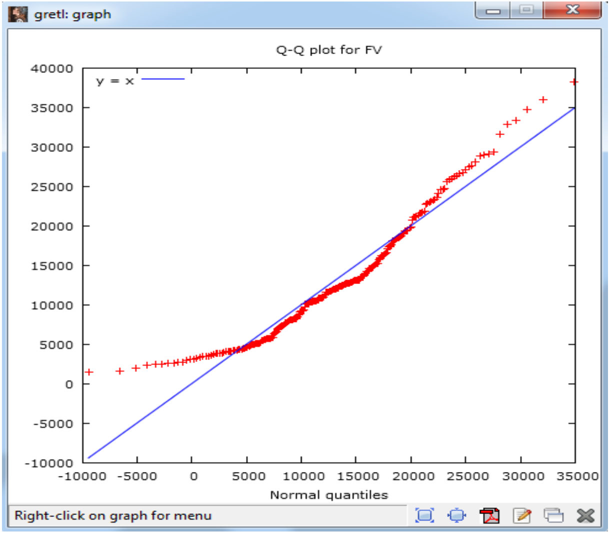
Example 3.5.5. Normal Q-Q plot. Visitors to Bilbao.

Q-Q plot. Variable SPV.



Example 3.5.5. Normal Q-Q plot. Visitors to Bilbao.

Q-Q plot. Variable FV.



Example 3.5.5. Normal Q-Q plot. Visitors to Bilbao.

Results.

- The observations are far away from the main diagonal in both plots. This result is clearer for variable FV. It may concluded that there is no evidence in these plots that these variables come from a normal distribution.
- The second and third options change the scale of the axis (either ordinate or abscissa). Nevertheless, the interpretation of the plot is similar: the more observations are close to the main diagonal of the box, the more evidence in the sample in favour of the normal distribution.

Example 3.5.5. Normal Q-Q plot. Simulation.

Questions.

Open the data files you created in the Example 3.5.2 (`SimT50.gdt` and `SimT1000.gdt`).

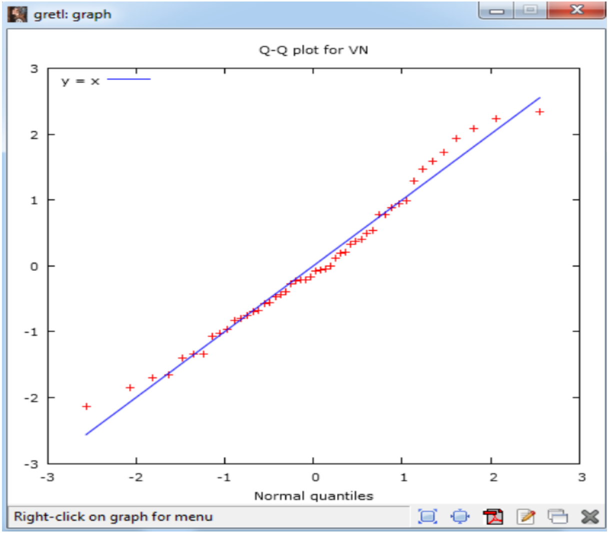
a. Obtain the Q-Q plot for the simulated variable in data file `Sim50.gdt`.

b. Obtain the Q-Q plot for the simulated variable in data file `Sim1000.gdt`.

c. Discuss and compare the results.

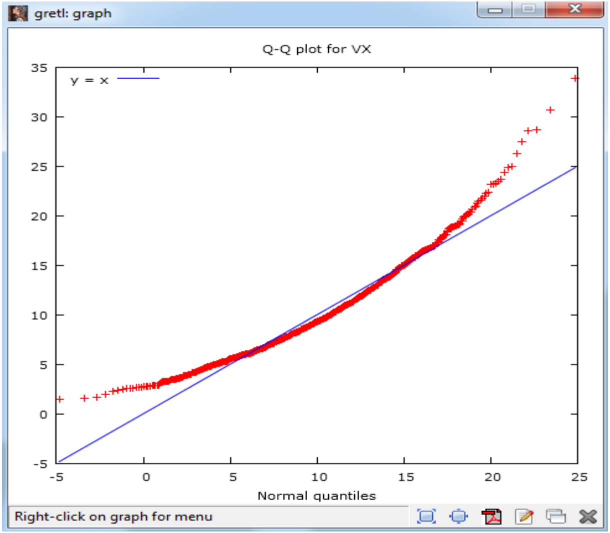
Example 3.5.5. Normal Q-Q plot. Simulation.

Q-Q plot (first option). Variable VN.



Example 3.5.5. Normal Q-Q plot. Simulation.

Q-Q plot (first option). Variable VX.



Example 3.5.5. Normal Q-Q plot. Simulation.

Results.

- A visual inspection of the Q-Q plot for variable VN (T=50) shows that it is not too close to the diagonal in the tails. Even though the variable VN comes from a normal distribution, the Q-Q plot is unclear and it is risky to reach any conclusion. This result can be due to the small sample size.

- Analysing the Q-Q plot for variable VX (T=1000), it can be observed that it moves away from the diagonal in the tails. This result was expected since the variable VX comes from a χ^2 distribution.

- In short, it may be concluded that the results derived from the normal Q-Q plots depend on the sample size and they are more reliable for large samples.

- Bear in mind that your simulated data are going to be different from ours. Therefore, you will obtain different Q-Q plots and your conclusions might be different.