Title

diagnostic plots — Distributional diagnostic plots

Syntax

Symmetry plot

`symplot` *varname* [*if*] [*in*] [, *options*$_1$]

Ordered values of varname against quantiles of uniform distribution

`quantile` *varname* [*if*] [*in*] [, *options*$_1$]

Quantiles of varname$_1$ *against quantiles of varname*$_2$

`qqplot` *varname*$_1$ *varname*$_2$ [*if*] [*in*] [, *options*$_1$]

Quantiles of varname against quantiles of normal distribution

`qnorm` *varname* [*if*] [*in*] [, *options*$_2$]

Standardized normal probability plot

`pnorm` *varname* [*if*] [*in*] [, *options*$_2$]

Quantiles of varname against quantiles of χ^2 *distribution*

`qchi` *varname* [*if*] [*in*] [, *options*$_3$]

χ^2 *probability plot*

`pchi` *varname* [*if*] [*in*] [, *options*$_3$]

options$_1$	Description
Plot	
marker_options	change look of markers (color, size, etc.)
marker_label_options	add marker labels; change look or position
Reference line	
`rlopts(`*cline_options*`)`	affect rendition of the reference line
Add plots	
`addplot(`*plot*`)`	add other plots to the generated graph
Y axis, X axis, Titles, Legend, Overall	
twoway_options	any options other than `by()` documented in [G-3] ***twoway_options***

options$_2$	Description
Main	
`grid`	add grid lines
Plot	
marker_options	change look of markers (color, size, etc.)
marker_label_options	add marker labels; change look or position
Reference line	
`rlopts(`*cline_options*`)`	affect rendition of the reference line
Add plots	
`addplot(`*plot*`)`	add other plots to the generated graph
Y axis, X axis, Titles, Legend, Overall	
twoway_options	any options other than `by()` documented in [G-3] ***twoway_options***

options$_3$	Description
Main	
`grid`	add grid lines
`df(#)`	degrees of freedom of χ^2 distribution; default is `df(1)`
Plot	
marker_options	change look of markers (color, size, etc.)
marker_label_options	add marker labels; change look or position
Reference line	
`rlopts(`*cline_options*`)`	affect rendition of the reference line
Add plots	
`addplot(`*plot*`)`	add other plots to the generated graph
Y axis, X axis, Titles, Legend, Overall	
twoway_options	any options other than `by()` documented in [G-3] ***twoway_options***

Menu

symplot

Statistics > Summaries, tables, and tests > Distributional plots and tests > Symmetry plot

quantile

Statistics > Summaries, tables, and tests > Distributional plots and tests > Quantiles plot

qqplot

Statistics > Summaries, tables, and tests > Distributional plots and tests > Quantile-quantile plot

qnorm

Statistics > Summaries, tables, and tests > Distributional plots and tests > Normal quantile plot

pnorm

Statistics > Summaries, tables, and tests > Distributional plots and tests > Normal probability plot, standardized

qchi

Statistics > Summaries, tables, and tests > Distributional plots and tests > Chi-squared quantile plot

pchi

Statistics > Summaries, tables, and tests > Distributional plots and tests > Chi-squared probability plot

Description

`symplot` graphs a symmetry plot of *varname*.

`quantile` plots the ordered values of *varname* against the quantiles of a uniform distribution.

`qqplot` plots the quantiles of $varname_1$ against the quantiles of $varname_2$ (Q–Q plot).

`qnorm` plots the quantiles of *varname* against the quantiles of the normal distribution (Q–Q plot).

`pnorm` graphs a standardized normal probability plot (P–P plot).

`qchi` plots the quantiles of *varname* against the quantiles of a χ^2 distribution (Q–Q plot).

`pchi` graphs a χ^2 probability plot (P–P plot).

See [R] **regress postestimation diagnostic plots** for regression diagnostic plots and [R] **logistic postestimation** for logistic regression diagnostic plots.

Options for symplot, quantile, and qqplot

Plot

marker_options affect the rendition of markers drawn at the plotted points, including their shape, size, color, and outline; see [G-3] ***marker_options***.

marker_label_options specify if and how the markers are to be labeled; see [G-3] ***marker_label_options***.

Reference line

`rlopts(`*cline_options*`)` affect the rendition of the reference line; see [G-3] ***cline_options***.

Add plots

addplot(*plot*) provides a way to add other plots to the generated graph; see [G-3] ***addplot_option***.

Y axis, X axis, Titles, Legend, Overall

twoway_options are any of the options documented in [G-3] ***twoway_options***, excluding by(). These include options for titling the graph (see [G-3] ***title_options***) and for saving the graph to disk (see [G-3] ***saving_option***).

Options for qnorm and pnorm

Main

grid adds grid lines at the 0.05, 0.10, 0.25, 0.50, 0.75, 0.90, and 0.95 quantiles when specified with qnorm. With pnorm, grid is equivalent to yline(.25,.5,.75) xline(.25,.5,.75).

Plot

marker_options affect the rendition of markers drawn at the plotted points, including their shape, size, color, and outline; see [G-3] ***marker_options***.

marker_label_options specify if and how the markers are to be labeled; see [G-3] ***marker_label_options***.

Reference line

rlopts(*cline_options*) affect the rendition of the reference line; see [G-3] ***cline_options***.

Add plots

addplot(*plot*) provides a way to add other plots to the generated graph; see [G-3] ***addplot_option***.

Y axis, X axis, Titles, Legend, Overall

twoway_options are any of the options documented in [G-3] ***twoway_options***, excluding by(). These include options for titling the graph (see [G-3] ***title_options***) and for saving the graph to disk (see [G-3] ***saving_option***).

Options for qchi and pchi

Main

grid adds grid lines at the 0.05, 0.10, 0.25, 0.50, 0.75, 0.90, and .95 quantiles when specified with qchi. With pchi, grid is equivalent to yline(.25,.5,.75) xline(.25,.5,.75).

df(*#*) specifies the degrees of freedom of the χ^2 distribution. The default is df(1).

Plot

marker_options affect the rendition of markers drawn at the plotted points, including their shape, size, color, and outline; see [G-3] ***marker_options***.

marker_label_options specify if and how the markers are to be labeled; see [G-3] ***marker_label_options***.

Reference line

rlopts(*cline_options*) affect the rendition of the reference line; see [G-3] ***cline_options***.

Add plots

addplot(*plot*) provides a way to add other plots to the generated graph; see [G-3] ***addplot_option***.

Y axis, X axis, Titles, Legend, Overall

twoway_options are any of the options documented in [G-3] ***twoway_options***, excluding by(). These include options for titling the graph (see [G-3] ***title_options***) and for saving the graph to disk (see [G-3] ***saving_option***).

Remarks and examples

stata.com

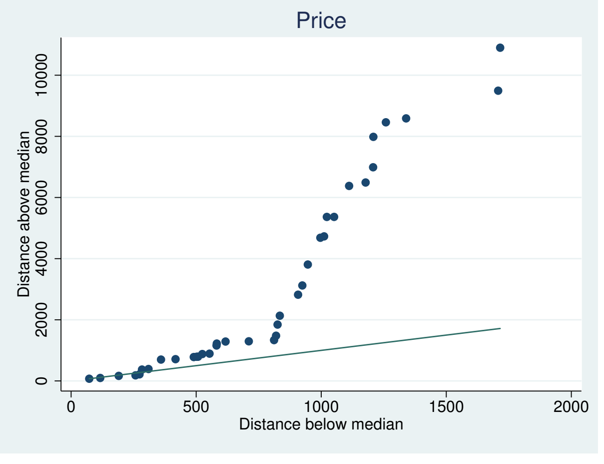
Remarks are presented under the following headings:

symplot

▷ Example 1

We have data on 74 automobiles. To make a symmetry plot of the variable price, we type

```
. use http://www.stata-press.com/data/r13/auto
(1978 Automobile Data)
. symplot price
```



All points would lie along the reference line (defined as $y = x$) if car prices were symmetrically distributed. The points in this plot lie above the reference line, indicating that the distribution of car prices is skewed to the right—the most expensive cars are far more expensive than the least expensive cars are inexpensive.

The logic works as follows: a variable, z, is distributed symmetrically if

$$\text{median} - z_{(i)} = z_{(N+1-i)} - \text{median}$$

where $z_{(i)}$ indicates the ith-order statistic of z. `symplot` graphs $y_i = \text{median} - z_{(i)}$ versus $x_i = z_{(N+1-i)} - \text{median}$.

For instance, consider the largest and smallest values of `price` in the example above. The most expensive car costs $15,906 and the least expensive, $3,291. Let's compare these two cars with the typical car in the data and see how much more it costs to buy the most expensive car, and compare that with how much less it costs to buy the least expensive car. If the automobile price distribution is symmetric, the price differences would be the same.

Before we can make this comparison, we must agree on a definition for the word "typical". Let's agree that "typical" means median. The price of the median car is $5,006.50, so the most expensive car costs $10,899.50 more than the median car, and the least expensive car costs $1,715.50 less than the median car. We now have one piece of evidence that the car price distribution is not symmetric. We can repeat the experiment for the second-most-expensive car and the second-least-expensive car. We find that the second-most-expensive car costs $9,494.50 more than the median car, and the second-least-expensive car costs $1,707.50 less than the median car. We now have more evidence. We can continue doing this with the third most expensive and the third least expensive, and so on.

Once we have all of these numbers, we want to compare each pair and ask how similar, on average, they are. The easiest way to do that is to plot all the pairs.

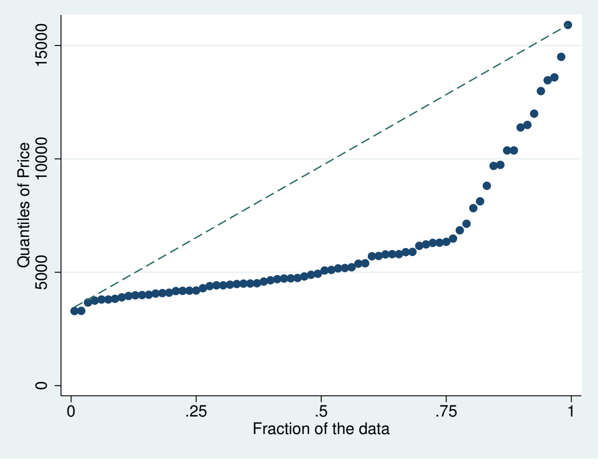
◁

quantile

▷ Example 2

We have data on the prices of 74 automobiles. To make a quantile plot of `price`, we type

```
. use http://www.stata-press.com/data/r13/auto, clear
(1978 Automobile Data)
. quantile price, rlopts(clpattern(dash))
```



We changed the pattern of the reference line by specifying `rlopts(clpattern(dash))`.

In a quantile plot, each value of the variable is plotted against the fraction of the data that have values less than that fraction. The diagonal line is a reference line. If automobile prices were rectangularly distributed, all the data would be plotted along the line. Because all the points are below the reference line, we know that the price distribution is skewed right.

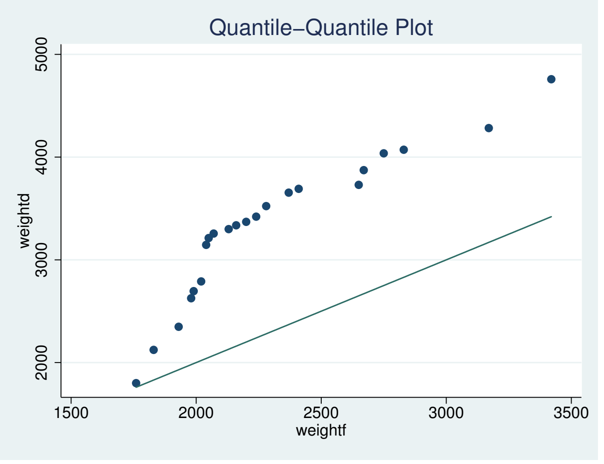
◁

qqplot

▷ Example 3

We have data on the weight and country of manufacture of 74 automobiles. We wish to compare the distributions of weights for domestic and foreign automobiles:

```
. use http://www.stata-press.com/data/r13/auto
(1978 Automobile Data)
. generate weightd=weight if !foreign
(22 missing values generated)
. generate weightf=weight if foreign
(52 missing values generated)
. qqplot weightd weightf
```



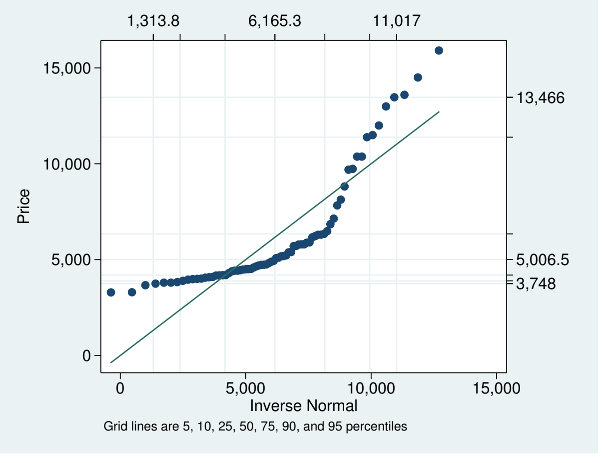
◁

qnorm

▷ Example 4

Continuing with our price data on 74 automobiles, we now wish to compare the distribution of `price` with the normal distribution:

```
. qnorm price, grid ylabel(, angle(horizontal) axis(1))
> ylabel(, angle(horizontal) axis(2))
```



The result shows that the distributions are different.

◁

❑ Technical note

The idea behind qnorm is recommended strongly by Miller (1997): he calls it probit plotting. His recommendations from much practical experience should interest many users. "My recommendation for detecting nonnormality is *probit plotting*" (Miller 1997, 10). "If a deviation from normality cannot be spotted by eye on probit paper, it is not worth worrying about. I never use the Kolmogorov–Smirnov test (or one of its cousins) or the χ^2 test as a preliminary test of normality. They do not tell you how the sample is differing from normality, and I have a feeling they are more likely to detect irregularities in the middle of the distribution than in the tails" (Miller 1997, 13–14).

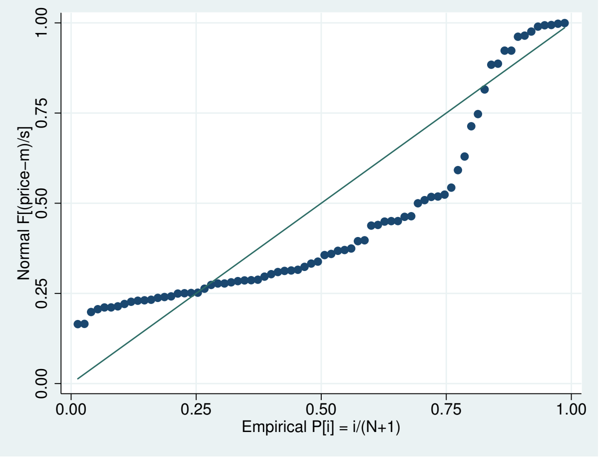
❑

pnorm

▷ Example 5

Quantile–normal plots emphasize the tails of the distribution. Normal probability plots put the focus on the center of the distribution:

```
. pnorm price, grid
```



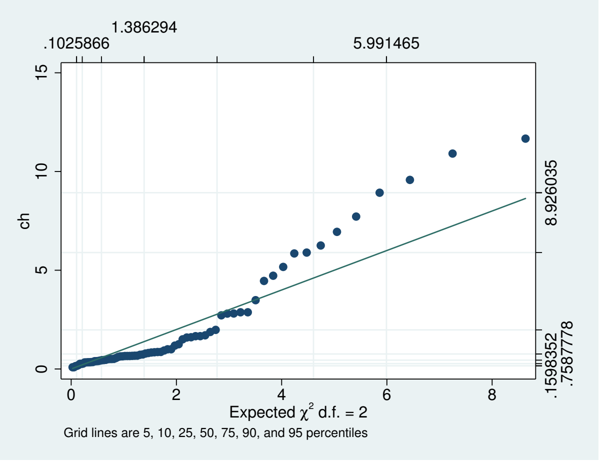
◁

qchi

▷ Example 6

Suppose that we want to examine the distribution of the sum of squares of price and mpg, standardized for their variances.

```
. egen c1 = std(price)
. egen c2 = std(mpg)
. generate ch = c1^2 + c2^2
. qchi ch, df(2) grid ylabel(, alt axis(2)) xlabel(, alt axis(2))
```



The quadratic form is clearly not χ^2 with 2 degrees of freedom.

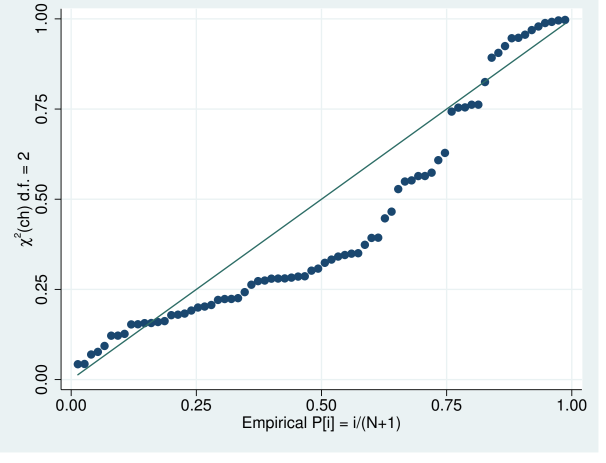
◁

pchi

▷ Example 7

We can focus on the center of the distribution by doing a probability plot:

```
. pchi ch, df(2) grid
```



◁

Methods and formulas

Let $x_{(1)}$, $x_{(2)}$, ..., $x_{(N)}$ be the data sorted in ascending order.

If a continuous variable, x, has a cumulative distribution function $F(x) = P(X \leq x) = p$, the quantiles x_{p_i} are such that $F(x_{p_i}) = p_i$. For example, if $p_i = 0.5$, then $x_{0.5}$ is the median. When we plot data, the probabilities, p_i, are often referred to as plotting positions. There are many different conventions for choice of plotting positions, given $x_{(1)} \leq \cdots \leq x_{(N)}$. Most belong to the family $(i - a)/(N - 2a + 1)$. $a = 0.5$ (suggested by Hazen) and $a = 0$ (suggested by Weibull) are popular choices.

For a wider discussion of the calculation of plotting positions, see Cox (2002).

`symplot` plots median $- x_{(i)}$ versus $x_{(N+1-i)}$ $-$ median.

`quantile` plots $x_{(i)}$ versus $(i - 0.5)/N$ (the Hazen position).

`qnorm` plots $x_{(i)}$ against q_i, where $q_i = \Phi^{-1}(p_i)$, Φ is the cumulative normal distribution, and $p_i = i/(N + 1)$ (the Weibull position).

`pnorm` plots $\Phi\{(x_i - \widehat{\mu})/\widehat{\sigma}\}$ versus $p_i = i/(N + 1)$, where $\widehat{\mu}$ is the mean of the data and $\widehat{\sigma}$ is the standard deviation.

`qchi` and `pchi` are similar to `qnorm` and `pnorm`; the cumulative χ^2 distribution is used in place of the cumulative normal distribution.

`qqplot` is just a two-way scatterplot of one variable against the other after both variables have been sorted into ascending order, and both variables have the same number of nonmissing observations. If the variables have unequal numbers of nonmissing observations, interpolated values of the variable with more data are plotted against the variable with fewer data.

Ramanathan Gnanadesikan (1932–) was born in Madras. He obtained degrees from the Universities of Madras and North Carolina. He worked in industry at Procter & Gamble, Bell Labs, and Bellcore, as well as in universities, retiring from Rutgers in 1998. Among many contributions to statistics he is especially well known for work on probability plotting, robustness, outlier detection, clustering, classification, and pattern recognition.

Martin Bradbury Wilk (1922–2013) was born in Montreal. He obtained degrees in chemical engineering and statistics from McGill and Iowa State Universities. After holding several statistics-related posts in industry and at universities (including periods at Princeton, Bell Labs, and Rutgers), Wilk was appointed Chief Statistician at Statistics Canada (1980–1986). He is especially well known for his work with Gnanadesikan on probability plotting and with Shapiro on tests for normality.

Acknowledgments

We thank Peter A. Lachenbruch of the Department of Public Health at Oregon State University for writing the original versions of `qchi` and `pchi`. Patrick Royston of the MRC Clinical Trials Unit, London, and coauthor of the Stata Press book *Flexible Parametric Survival Analysis Using Stata: Beyond the Cox Model* also published a similar command in the *Stata Technical Bulletin* (Royston 1996).

References

Chambers, J. M., W. S. Cleveland, B. Kleiner, and P. A. Tukey. 1983. *Graphical Methods for Data Analysis*. Belmont, CA: Wadsworth.

Cox, N. J. 1999. gr42: Quantile plots, generalized. *Stata Technical Bulletin* 51: 16–18. Reprinted in *Stata Technical Bulletin Reprints*, vol. 9, pp. 113–116. College Station, TX: Stata Press.

——. 2001. gr42.1: Quantile plots, generalized: Update to Stata 7. *Stata Technical Bulletin* 61: 10. Reprinted in *Stata Technical Bulletin Reprints*, vol. 10, pp. 55–56. College Station, TX: Stata Press.

——. 2002. Speaking Stata: On getting functions to do the work. *Stata Journal* 2: 411–427.

——. 2004a. Speaking Stata: Graphing distributions. *Stata Journal* 4: 66–88.

——. 2004b. gr42_2: Software update: Quantile plots, generalized. *Stata Journal* 4: 97.

——. 2005a. Speaking Stata: Density probability plots. *Stata Journal* 5: 259–273.

——. 2005b. Speaking Stata: The protean quantile plot. *Stata Journal* 5: 442–460.

——. 2005c. Speaking Stata: Smoothing in various directions. *Stata Journal* 5: 574–593.

——. 2007. Stata tip 47: Quantile–quantile plots without programming. *Stata Journal* 7: 275–279.

——. 2012. Speaking Stata: Axis practice, or what goes where on a graph. *Stata Journal* 12: 549–561.

Daniel, C., and F. S. Wood. 1980. *Fitting Equations to Data: Computer Analysis of Multifactor Data*. 2nd ed. New York: Wiley.

Gan, F. F., K. J. Koehler, and J. C. Thompson. 1991. Probability plots and distribution curves for assessing the fit of probability models. *American Statistician* 45: 14–21.

Genest, C., and G. J. Brackstone. 2013. Obituary: Martin B. Wilk, 1922–2013. *IMS Bulletin* 42(4): 7–8.

Hamilton, L. C. 1992. *Regression with Graphics: A Second Course in Applied Statistics*. Belmont, CA: Duxbury.

——. 2013. *Statistics with Stata: Updated for Version 12*. 8th ed. Boston: Brooks/Cole.

Hoaglin, D. C. 1985. Using quantiles to study shape. In *Exploring Data Tables, Trends, and Shapes*, ed. D. C. Hoaglin, C. F. Mosteller, and J. W. Tukey, 417–460. New York: Wiley.

Kettenring, J. R. 2001. A conversation with Ramanathan Gnanadesikan. *Statistical Science* 16: 295–309.

Miller, R. G., Jr. 1997. *Beyond ANOVA: Basics of Applied Statistics*. London: Chapman & Hall.

Nolan, D., and T. Speed. 2000. *Stat Labs: Mathematical Statistics Through Applications*. New York: Springer.

Royston, P. 1996. sg47: A plot and a test for the χ^2 distribution. *Stata Technical Bulletin* 29: 26–27. Reprinted in *Stata Technical Bulletin Reprints*, vol. 5, pp. 142–144. College Station, TX: Stata Press.

Scotto, M. G. 2000. sg140: The Gumbel quantile plot and a test for choice of extreme models. *Stata Technical Bulletin* 55: 23–25. Reprinted in *Stata Technical Bulletin Reprints*, vol. 10, pp. 156–159. College Station, TX: Stata Press.

Wilk, M. B., and R. Gnanadesikan. 1968. Probability plotting methods for the analysis of data. *Biometrika* 55: 1–17.

Also see

[R] **cumul** — Cumulative distribution

[R] **kdensity** — Univariate kernel density estimation

[R] **logistic postestimation** — Postestimation tools for logistic

[R] **lv** — Letter-value displays

[R] **regress postestimation diagnostic plots** — Postestimation plots for regress