Quantile plots: New planks in an old campaign

Nicholas J. Cox
Department of Geography
Quantile plots

Quantile plots show ordered values (raw data, estimates, residuals, whatever) against rank or cumulative probability or a one-to-one function of the same.

Tied values are assigned distinct ranks or probabilities.
Example with auto dataset

Quantiles of Mileage (mpg)

Fraction of the data
quant i l e default

In this default from the official command quant i l e, ordered values are plotted on the $y$ axis and the fraction of the data (cumulative probability) on the $x$ axis.

Quantiles (order statistics) are plotted against **plotting position** $(i - 0.5)/n$ for rank $i$ and sample size $n$.

Syntax was

```bash
sysuse auto, clear
quant i l e m p g, aspect(1)
```
Quantile plots have a long history


all used quantile plots *avant la lettre*.

In geomorphology, hypsometric curves for showing altitude distributions are a long-established device with the same flavour.
Quantile plots named as such

A relatively long history in Stata


Graph.Kit (February 1986) included commands quantile, qqplot and qnorm.

Thanks to Pat Branton of StataCorp for this history.
Related plots use the same information

Cumulative distribution plots show cumulative probability on the $y$ axis.

Survival function plots show the complementary probability.

Clearly, axes can be exchanged or reflected.

`distplot (Stata Journal)` supports both.
Many people will already know about `sts graph`.
So, why any fuss?

The presentation is built on a long-considered view that quantile plots are the best single plot for univariate distributions.

No other kind of plot shows so many features so well across a range of sample sizes with so few arbitrary decisions.

Example: Histograms require binning choices.
Example: Density plots require kernel choices.
Example: Box plots often leave out too much.
What’s in a name? QQ-plots

Talk of quantile-quantile (Q-Q or QQ-) plots is also common.

As discussed here, all quantile plots are also QQ-plots.

The default quantile plot is just a plot of values against the quantiles of a standard uniform or rectangular distribution.
NJC commands

The main commands I have introduced in this territory are

◊ `quantile2` *(Stata Technical Bulletin)*

◊ `qplot` *(Stata Journal)*

◊ `stripplot` *(SSC)*

Others will be mentioned later.
This command published in *Stata Technical Bulletin* 51: 16–18 (1999) generalized quantile:

◊ One or more variables may be plotted.

◊ Sort order may be reversed.

◊ `by()` option is supported.

◊ Plotting position is generalised to \((i - a) / (n - 2a + 1)\): compare \(a = 0.5\) or \((i - 0.5)/n\) wired into quantile.
qpl ot

The command quant i l 2 was renamed qpl ot and further revised in *Stata Journal* 5: 442–460 and 471 (2005), with later updates:

◊ over() option is also supported.

◊ Ranks may be plotted as well as plotting positions.

◊ The \( x \) axis scale may be transformed on the fly.

◊ recast() to other `twoway` types is supported.
**stripplot**

The command `stripplot` on SSC started under Stata 6 as `onewayplot` in 1999 as an alternative to `graph, oneway` and has morphed into (roughly) a superset of the official command `dotplot`.

It is mentioned here because of its general support for quantile plots as one style and its specific support for quantile-box plots, on which more shortly.
Comparing two groups is basic

superimposed

juxtaposed
Syntax was

```
qplot mpg, over(foreign) aspect(1)
```

```
stripplot mpg, over(foreign) cumulative centre vertical aspect(1)
```
Quantiles and transformations commute

In essence, transformed quantiles and quantiles of transformed data are one and the same, with easy exceptions such as reciprocals reversing order.

So, quantile plots mesh easily with transformations, such as thinking on logarithmic scale.

For the latter, we just add simple syntax such as `ysc(log)`.

Note that this is not true of (e.g.) histograms, box plots or density plots, which need re-drawing.
The shift is multiplicative, not additive?
A more unusual example

Glacier terminus position change may be positive or negative, with possible outliers of either sign.

Cube root transformation pulls in both tails and (fortuitously but fortunately) can separate advancing and retreating glaciers.

Terminus position change (m yr\(^{-1}\))

Pacific Coast of East Antarctica glaciers

Boxplots show 5 25 50 75 95% points
Terminus position change (m yr$^{-1}$)

Pacific Coast of East Antarctica glaciers

Boxplots show 5 25 50 75 95% points

cube root scale

Terminus position change (m yr$^{-1}$)

*multqplot* (Stata Journal)

`multqplot` is a convenience command to plot several quantile plots at once.

It has uses in data screening and reporting.
It might prove more illuminating than the tables of descriptive statistics ritual in various professions.

By default the minimum, lower quartile, median, upper quartile and maximum are labelled on the $y$ axis – so we are half-way to showing a box plot too.

By default also variable labels (or names) appear at the top.

Raw or smoothed?

Quantile plots show the data as they come: we get to see outliers, grouping, gaps and other quirks of the data, as well as location, scale and general shape.

But sometimes the details are just noise or fine structure we do not care about.

Once you register that values of mpg in the auto data are all reported as integers, you want to set that aside.

You can smooth quantiles, notably using the Harrell and Davis method, which turns out to be bootstrapping in disguise. hdquantile (SSC) offers the calculation.
Letter values

Often we do not really need all the quantiles, especially if the sample size is large.

We could just use the letter values, which are the median, quartiles (fourths), octiles (eighths), and so forth out to the extremes, halving the tail probabilities at each step.

`lv` supports letter value displays.
`lvalues` (SSC) is now available to generate variables.

Thanks to David Hoaglin for suggesting letter values at the Chicago meeting and to Kit Baum for posting `lvalues` on SSC.
Parsimony of letter values

For $n$ data values, there are $1 + 2 \cdot \text{ceil}(\log_2 n)$ letter values.

For $n = 1000, 10^6, 10^9$, there are 21, 41, 61 letter values.

We will see examples shortly.
Fitting or testing named distributions

Using quantile plots to compare data with named distributions is common.

The leading example is using the normal (Gaussian) as reference distribution.
Indeed, many statistical people first meet quantile plots as such normal probability plots.

Yudi Pawitan in his 2001 book *In All Likelihood* (Oxford University Press) advocates normal QQ-plots as making sense generally — even when comparison with normal distributions is not the goal.
qnorm is already available as an official command — but it is limited to the plotting of just one set of values.
Named distributions with `qpl ot`

`qpl ot` has a general `trscal e()` option to transform the `x` axis scale that otherwise would show plotting positions or ranks.

For normal distributions, the syntax is just to add `trscal e(invnormal(@))`

`@` is a placeholder for what would otherwise be plotted.

`invnormal()` is Stata’s name for the normal quantile function (as an inverse cumulative distribution function).
A standard plot in support of $t$ tests?

This plot is suggested as a standard for two-group comparisons:

◊ We see all the data, including outliers or other problems.
◊ Use of a normal probability scale shows how far that assumption (read: ideal condition) is satisfied.
◊ The vertical position of each group tells us about location, specifically means.
◊ The slope or tilt of each group tells us about scale, specifically standard deviations.
◊ It is helpful even if we eventually use Wilcoxon-Mann-Whitney or something else.
What if you had paired values?

Plot the differences, naturally.

Nothing stops you plotting the original values too, but at some point the graphics should respect the pairing.
Different axis labelling?

The last plot used a scale of standard normal deviates or $z$ scores.

Some might prefer different labelling, e.g. % points.

`mylabel` (SSC) is a helper command, which puts the mapping in a local macro for your main command:

```
mylabel 1 2 5 10(20) 90 95 98 99, myscale(invnormal(@/100)) local(plabels)
```
quantiles of Mileage (mpg)

exceedance probability (%)

Foreign
Domestic
Syntax for that example

sysuse auto, clear

mylabels 1 2 5 10(20)90 95 98 99, myscale(invnormal(@/100)) local(plabels)

qplot mpg, over(foreign) trscale(invnormal(@) aspect(1) xla(`plabels') xtitle(exceedance probability (%)) xsc(titlegap(*5)) legend(pos(11) ring(0) order(2 1) col(1))
How would letter values do?

For the auto data there are

52 domestic cars  13  letter values
22 foreign cars    11  letter values.

The use of letter values is parsimonious, but respectful of major detail: extremes are always echoed.
quantiles of Mileage (mpg)

Foreign
Domestic

invnormal(P)
Other named distributions?

There are many, many named distributions for which customised QQ-plot commands could be written.

I am guilty of programs for beta, Dagum, Dirichlet, exponential, gamma, generalized beta (second kind), Gumbel, inverse gamma, inverse Gaussian, lognormal, Singh-Maddala and Weibull distributions.

But a better approach when feasible is to allow a distribution to be specified on the fly.
Harold Jeffreys suggested that error distributions are more like $t$ distributions with 7 df than like Gaussians.


Sir Harold Jeffreys 1891–1989

County Durham man established that the Earth’s core is liquid pioneer Bayesian
plotted for probability in [0.001, 0.999]
How to explore?

Simulate with `rt(7, )` and samples of desired size.

`trscale(invt(7, @))` sets up $x$ axis scale on the fly.
quantiles of $t$ with 7 df
Box plot hybrids
Adding a box plot flavour

Earlier we saw how extremes and quartiles could be made explicit on the $y$ axis of a quantile plot. They are the minimal ingredients for a box plot.

Clearly we can also flag cumulative probabilities $0(0.25)1$ on the corresponding $x$ axis scale.
Tracing the box

In multqplot by default the box is shown as part of a double set of grid lines.

This helps underline that half of the points on a box plot are inside the box and half outside, a basic fact often missed in interpreting these plots, even by experienced researchers.
Quantile-box plots


His original examples were not especially impressive, perhaps one reason they have not been more widely emulated.

Emanuel Parzen 1929–2016
Boston housing data

Here for quantile-box plots we use data from


https://archive.ics.uci.edu/ml/datasets/Housing

Number of Figures in original paper: 1
Number of Figures showing raw data: 0
Broad contrast and fine structure

stripplot MEDV, over(CHAS) vertical cumulative centre box cumprob aspect(1)
Some quirks in that dataset

- % residential land zoned lots > 25000 sq. ft
- % non-retail business acres per town
- Accessibility to radial highways
- Full-value property-tax rate per $10,000
Ordinal (graded) data

Ordinal (graded) data can be shown with quantile plots too.

Such data might alternatively be plotted against the midpoints of the corresponding probability intervals.

Statistical discussion was given in *Stata Journal* 4: 190–215 (2004), Section 5.
quantiles of Repair Record 1978

fraction of the data

quantiles of Repair Record 1978

logit(P)

Foreign
Domestic
The `midpoint` option is included in a Software Update in press, *Stata Journal* 16(3) 2016.
Differences of quantiles

Plotting differences of quantiles versus their mean or versus plotting position is often a good idea.

cquantile (SSC) is a helper program.

Much more was said on this at *Stata Journal* 7: 275–279 (2007).
Words from the wise
Graphs force us to note the unexpected; nothing could be more important.

John Wilder Tukey
1915–2000

Using the data to guide the data analysis is almost as dangerous as not doing so.

Frank E. Harrell Jr
Questions?
All graphs use Stata scheme s1col or , which I strongly recommend as a lazy but good default.

This font is Georgia.
This font is Lucida Console.