The Stata Journal (2024)
24, Number 3, pp. 514–534 DOI: 10.1177/1536867X241276114

Speaking Stata: Quantile–quantile plots, generalized

Nicholas J. Cox
Department of Geography
Durham University
Durham, UK
n.j.cox@durham.ac.uk

Abstract. Quantile–quantile plots in the precise sense of scatterplots showing corresponding quantiles of two variables have long been supported by official command `qqplot`. That command is generalized here in several ways in a new command, `qqplotg`. In this article, I explain the major features of `qqplotg` and give several examples of its use. Themes include the use of quantile–quantile plots to explore the possibilities for working on a transformed scale and the value of plotting difference between quantiles versus mean quantile or plotting position. Various historical and methodological remarks are sprinkled throughout.

Keywords: gr0096, qqplotg, quantile–quantile plots, quantile plots, quantiles, plotting positions, additivity, homoskedasticity, symmetry, linearity, transformations, difference versus mean, graphics, distributions

1 Introduction

Quantile–quantile plots, in the sense used in this article, compare precisely two distributions, whether as two groups of one variable or as two variables that are comparable (which at its simplest implies having the same units of measurement). As an immediate concrete example, consider domestic and foreign cars in Stata's `auto.dta`, domestic meaning made in the United States and foreign meaning made elsewhere. `auto.dta` includes several variables that may be compared for such groups, such as price, weight, or miles per gallon. A quantile–quantile plot is a scatterplot of corresponding points in each sample distribution. An easy start is that the minimum and maximum of one group or variable can be plotted against the minimum and maximum of the other group or variable. How this idea is extended to other points in each distribution will become clear shortly.

One way or another, Stata has included official support for quantile–quantile plots since 1985. The *STATA/Graphics User's Guide* of August 1985 included the do-files `qqplot.do` and `quantile.do`. The Graph.Kit of February 1986 included the commands `qqplot`, `quantile`, and `qnorm`.

However, what `qqplot` can do remains fairly limited. The main point of this article is to discuss how the underlying ideas can be extended, as is implemented in the new command `qqplotg`. The extra letter `g` conveniently stands for generalized, group, and `generate`, which is good enough for a new name.

Section 2 expands on the ideas of quantiles and quantile–quantile plots. Readers already familiar with those ideas may feel comfortable skipping or skimming through this section. Section 3 covers the new features introduced in `qqplotg`. Section 4 is a more formal specification of the command. Throughout the article, the intent is also to provide examples of such plotting, a method that still seems widely unknown or at least widely undervalued.

2 Quantiles and quantile–quantile plots

Just about every introductory statistics course or text covers several methods for comparing two-group data, or two-variable data, including various graphs (histogram or box plots, say), various summary measures (means or medians, say), and various significance tests (Student's t or Wilcoxon–Mann–Whitney, say). In contrast, the method of quantile–quantile plots is often not covered in introductions to statistics. Indeed, the method is often not discussed in detail even in more advanced treatments of statistical graphics. Some otherwise excellent books do not get beyond the valuable but very specific method of normal quantile plots (Unwin 2015; Wilke 2019). The resulting plots, however, are easy to understand and may even be more informative, and hence more helpful, than other kinds of graphs.

The term "quantile" is often attributed to Kendall (1940) but can be found in Fisher and Yates (1938). However, it has acquired at least three distinct if related meanings, so we should tease them apart.

1. Quantiles in this context mean the entire set of values of a quantitative variable sorted or ordered from smallest to largest, the "order statistics" if you prefer. Terms such as quantile plot, used in this sense, go back at least to an outstanding article by Wilk and Gnanadesikan (1968). For general exposition, Chambers et al. (1983) and Cleveland (1993, 1994) remain authoritative and lucid. For Stata-related discussions, see, for example, Cox (2005, 2007b).

   Note that it may often be sensible and sufficient to plot selected quantiles, especially if a dataset is very large. That idea is not pursued further here, but for one line of attack, see particularly Cox (2016b).

2. A related but distinct idea is that quantiles are summary measures based on some proportion or percent being smaller and the complementary proportion or percent being larger. Median and quartiles are likely to be very familiar examples. With extra detail about calculation recipes, the median is defined by 50% of values being smaller and 50% being larger; and each quartile, lower or upper, is defined by 25% of values being on one side and 75% of values being on the other side. Many terms have been proposed for measures based on different subdivisions of the cumulative probability range. See Cox (2016b) for an incomplete menagerie. The attraction of quantile as a term is that it covers all such cases without confronting readers with a term that may be unfamiliar. Indeed, the word is already available for any application for which a specific term has not yet been invented.

3. Quantiles are also used to denote bins, classes, or intervals defined by such summary measures in sense 2. The median and quartiles, say, define four quartile bins, each holding about a quarter of any batch of values.

Back to quantile–quantile plots: For simplicity, first suppose that two groups are equal in size, with the same number of observations in each. After sorting or ordering each group separately, we could plot the smallest in one group against the smallest in the other; plot the second smallest in each group as another point; and so on until we plot the largest in each group as our last point. The resulting scatterplot includes all the information in the data about the two distributions, without any arbitrary decisions about binning (compare histograms) or about what to show by way of summary measures or important detail (compare box plots). Two distributions that are very similar would plot close to a reference line of equality $y = x$ (say). Two distributions very similar but for an additive shift (the simplest reference case for comparison of two means or medians) would plot close to a different straight line, as would two distributions very similar but for a multiplicative shift. Real data could easily be more complicated, but that is precisely the point too. Other complications or features, such as grouping, outliers, or distributions being similar in the middle but different in the tails, would be matched by the configuration of the plot.

Often, two groups are not equal in size, but that is not a real barrier to applying these ideas. In `auto.dta`, there are 22 foreign cars and 52 domestic cars. One answer is that we can plot 22 points, the quantiles for foreign cars versus interpolated quantiles for domestic cars for the equivalent cumulative probabilities. As always, there are limits to the usefulness of this device. A plot of 3 values in one group against 3 interpolated between 3,000 values for another group would not be much help, but the same limitation would apply to any other method.

As mentioned in the *Introduction*, `qqplot` as an official command has been paired since birth with `quantile` as an official command, which also remains fairly limited in what it can do. `quantile` offers a plot of the sorted or ordered values of a variable against (loosely) the corresponding cumulative probabilities. The result of `quantile` can be (and in Stata documentation is) also presented as a quantile–quantile plot: The quantiles on the vertical axis are the values of the variable supplied, while those on the horizontal axis are those of a uniform (rectangular, flat) distribution on $[0, 1]$.

Interpreting those quantities as in essence cumulative probabilities is to my taste as valid and indeed more helpful. So let's explore that idea: exactly what `quantile` does and shows is worth explaining.

To make this concrete, imagine a toy sample of 5 distinct values that we can rank 1 to 5. Stata follows a convention that rank 1 corresponds to the lowest value. A rule that cumulative probability is taken to be rank / sample size would give such probabilities as 0.2, 0.4, 0.6, 0.8, and 1, which would not treat the distribution symmetrically. A rule of (rank − 1) / sample size would give us 0, 0.2, 0.4, 0.6, and 0.8, which is no better. Splitting the difference, with a rule (rank − 1/2) / sample size, an idea that goes back at least to Galton (1883, 1907), gives us 0.1, 0.3, 0.5, 0.7, and 0.9, which is pleasingly symmetric about its middle, and indeed a good solution. Assigning cumulative probability of 0.5 to the middlemost value with rank 3 of 5, corresponding to the median in this case, is an evident bonus. There are other solutions to this small problem, but the rule (rank − 1/2) / sample size is used by `quantile`, and so we stop there for now. Such quantities are often called "plotting positions".

What the official command `quantile` does has been extended through the community-contributed command `qplot`. This was released as `quantil2` (Cox 1999), but soon renamed.[1] The latest update to `qplot` is announced in a Software Update in this issue of the *Stata Journal*. Its full functionality need not be summarized here, but an immediate comment is that it offers another way to compare two groups, because the two sets of quantiles can be compared, directly and without interpolation, either superimposed or juxtaposed.

The help for `qplot` includes much more discussion and many more references.

3 New features

3.1 Two groups

Let's look first at using `qqplotg` with the concrete example flagged in the *Introduction*, comparison of any variable as between domestic and foreign cars in the auto data. Here, and often, there is no pairing of domestic and foreign cars, so there is no other information about the distributions.

```
. sysuse auto
(1978 automobile data)
```

We will look at miles per gallon. It is not essential here, but will help mightily soon, to have automated methods of defining and working with axis labels. The commands `nicelabels` and `mylabels` are discussed at length in Cox (2022). The idea they share is putting a set of axis labels into a local macro.

```
. nicelabels mpg, local(la) tight
step:       10
labels:     20 30 40
```

The `over()` option, or if you prefer the equivalent `group()` option, is needed to specify the variable that defines the two groups. It must have two and only two distinct values. If you have a variable with more than two distinct values, but are nevertheless interested in comparing distributions for two of those, ensure that you select those two using an `if` qualifier. Otherwise, if you are interested in comparing three or more sets of quantiles, then you need `qplot`.

[1] The original name was not a typo but was trimmed to fit maximum filename lengths that could be no longer than the pattern `filename.ext` in the DOS operating system.

Figure 1 is a first quantile–quantile plot.

```
. qqplotg mpg, over(foreign) flip xla(`la') yla(`la') subtitle(raw scale)
> name(QQG1)
```

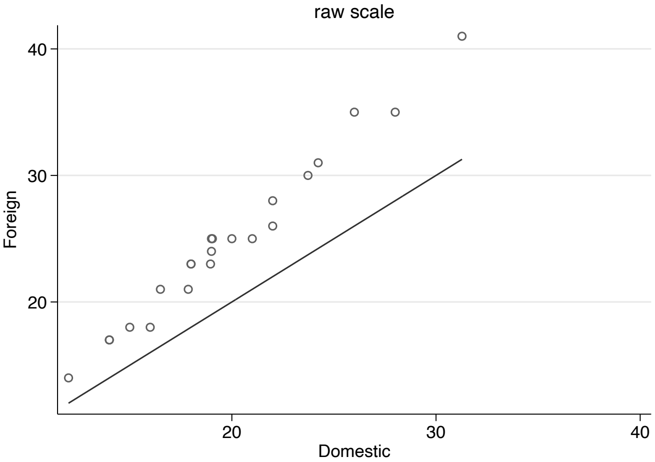

Figure 1. Quantile–quantile plot for miles per gallon comparing domestic and foreign cars. The solid line defines equality of distributions.

As often in such plots, there is a reference line of equality. If the two distributions were essentially the same, all data points would lie very close to that line. Clearly, there is extra structure in these data: quantile for quantile, foreign cars have higher miles per gallon than domestic cars, but the pattern of differences also shows a tilt compared with equality. The pattern is more complicated than just an additive shift.

The `name()` option is used here because we will shortly compare this plot with other plots.

Here and in many other examples you might want to enforce `aspect(1)` as an extra option.

3.2 Flipping axes

Now comes a point that applies widely. The command for figure 1 included a `flip` option that swaps axes as compared with the default. I first ran this command without the `flip` option but then decided that I preferred it with axes swapped, which is what `flip` does for you. So this option removes any need to recode the group variable, in this example swapping values 0 and 1. Further, it reduces any need to think about which way up the points will lie. If you do not much care about which way round the axes are, or you never make silly mistakes, you will not need this option. I tend to prefer plots in which most of or all the points lie above the line of equality whenever I also want to say that one group is typically higher in value than the other, but none of that is binding on anyone else.

3.3 Transformations

Many researchers want to explore whether comparisons are easier or more effective on a transformed scale. That interest does raise other questions too. Often, an analysis needs to be based on much more than the marginal distributions of two groups or two variables. Conversely, comparison of two groups on one outcome sometimes is the entire problem.

In the immediate context of comparing two distributions, key issues include how far each distribution is symmetric or skewed; whether distributions have approximately equal spread or very different spread; and whether distributions are related by an additive shift, or something more complicated. More formally, the "ideal conditions", ideal in the sense that they would make analysis and interpretation easier if they were true, are symmetry, homoskedasticity, and additivity.

Often, but I think unhelpfully, these ideal conditions are called assumptions, with sometimes undertones that failure to match assumptions invalidates an analysis. But statistics with data is applied mathematics, not pure mathematics or logic. Even simulated data always fail to match the ideals exactly, and, with real data, assumptions are at best matched approximately. A personal preference to talk about ideal conditions has splendid precedents in Anscombe (1961) and Anscombe and Tukey (1963).[2]

Of these ideals, additivity is perhaps the most valuable goal if you have to choose. More generally, whatever brings data closer to a simpler systematic structure is most important in choosing a transformed scale. Here that means additivity is the main goal. In other applications, linearity of relationships is also a major goal. The relative importance of different ideal conditions (assumptions, if you must) is discussed in (for example) most better texts on regression (for example, Gelman, Hill, and Vehtari [2021]).

Sometimes, however, a specific transformation will help us move closer to all ideals, and we do not have to choose at all. Sometimes, but not always: McCullagh and Nelder (1989, 22) give a salutary example of how different transformations may be needed for different goals. Given a Poisson error distribution, square roots give approximate symmetry, the power 2/3 gives approximate homoskedasticity, while logarithms impart approximate additivity of systematic effects.

I tend to be positive about transformations, while also considering that the number of transformations that are really helpful is rather small. See McCullagh (2022) for similar remarks in a recent text. For that reason, and others, I do not put much trust in more formal approaches such as Box–Cox transformation, despite its wonderful name

[2] Francis John Anscombe (1918–2001) and John Wilder Tukey (1915–2000) married sisters; Tukey's term was brother-in-squared law. In other family news, Paul A. Tukey (1945–), also cited in this article, and John W. Tukey were fifth cousins.

and the undoubted eminence of its authors (Box and Cox 1964). Nor, for different reasons, do I tend to use any of the official commands `ladder`, `gladder`, or `qladder`, which are based on trying out many transformations in the same exercise, a shotgun style differing from that in this article.

A transformation should spring out of a graph as imparting simpler structure—and be independently defensible.

With that high ideal in mind, let's try out some transformations on these data. It is clear from basic summaries or graphs that `mpg` is moderately positively (right) skewed. Although getting closer to symmetry is as said not the most important goal, transformations that do that often take you closer to additive structure.

Cube roots, logarithms, and reciprocals are plausible candidates for transformations in this example. They were listed just now in order of increasing strength, namely, how much they change the shape of a distribution. Naturally, cube roots and reciprocals are powers, 1/3 and -1, respectively. It is not quite so well known that logarithms can be thought of as members of the same family, which is one of the main points of Box and Cox (1964). Additionally, or alternatively, see Tukey (1957, 1977).

A neat way to encapsulate family resemblance comes from a remark by Mosteller and Tukey (1977, 80). Transformations indexed by power p come from $\int y^{p-1} dy$, setting aside constants that do not affect distribution shape. Thus, cube root comes from $\int y^{-2/3} dy$, $\ln p$ comes from $\int y^{-1} dy$, and reciprocal comes from $\int y^{-2} dy$, and those transformations are ordered thus.

The motivation for cube roots is most often pulling in positively skewed distributions. Like square roots, cube roots map zeros to zeros and otherwise pull in large positive values relative to small positive values. Specifically, cube roots to a good approximation map gamma-like distributions to normal distributions. Cube roots can have other advantages too (Cox 2011).

Anscombe (1981, 215) sets as an exercise showing that for an exponential distribution, transformation with the power 0.307 imparts equality of median and mode; 0.302 equality of mean and mode; 0.290 equality of mean and median. We can take this example in different ways. One is to underline that any ideal (here symmetry) is a little fuzzy until made more precise. Another is more optimistic and pragmatic: that nearby transformations will have similar effects. These powers are also close in practice to cube roots.[3]

Logarithms work harder at pulling in positively skewed distributions. Of greater interest is that they imply multiplicative rather than additive relationships, so that logarithms map multiplicative structure to additive structure.

[3] Anscombe's (1981) book tried to persuade statistical people to make more use of the APL programming language. That project was doomed to failure, but the book is delightfully quirky and remains original and provocative. On page 126 can be found the most subtle and succinct epitome of the aim of statistics that I know.

Reciprocals work even harder at doing that. They do reverse order of positive values, so that the smallest value of miles per gallon (in this example) becomes the largest value of gallons per mile, and vice versa. Some researchers correct that change of order by using negative reciprocals, but it seems to me simpler to work with reciprocals and accept that higher and lower groups are reversed. Mentioning units of measurement just now raises an even more crucial point: reexpression in reciprocals changes the units of measurement but does so in a way that may be helpful, or at any rate not especially confusing. In this specific case, although using miles per gallon for gas or petrol consumption is standard in the United States and some other countries, in many countries a reciprocal scale is more nearly standard, with units such as liters per 100 km.

More generally, keeping track of units and dimensions is always a good habit, which comes naturally to people with some background in applied mathematics, physics, or engineering, who are accustomed to principles of dimensional analysis (for example, Gibbings [2011], Mahajan [2010; 2014], Santiago [2019]). Finney (1977) gave a splendidly concise and incisive overview of key dimensional principles for statistics.

So there are grounds for trying out each of those transformations and for comparing them with the original or raw scale. I recommend working a bit at keeping axis labels in the original units, which is a job for `mylabels` (Cox 2022)—unless you are not bothered by units on transformed scales.

```
. mylabels `la', myscale(@^(1/3)) local(la2)
2.714417616594906 "20" 3.107232505953859 "30" 3.419951893353394 "40"
. qqplotg mpg, over(foreign) flip transform(@^(1/3)) xla(`la2')
> yla(`la2') subtitle(cube root scale) name(QQG2, replace)
. mylabels `la', myscale(ln(@)) local(la3)
2.995732273553991 "20" 3.401197381662155 "30" 3.688879454113936 "40"
. qqplotg mpg, over(foreign) flip transform(ln) xla(`la3') yla(`la3')
> subtitle(log scale) name(QQG3, replace)
. mylabels `la', myscale(1/@) local(la4)
.05 "20" .0333333333333333 "30" .025 "40"
. qqplotg mpg, over(foreign) flip transform(1/@) xla(`la4') yla(`la4')
> ysc(reverse) xsc(reverse) subtitle(reciprocal scale) name(QQG4, replace)
. graph combine QQG1 QQG2 QQG3 QQG4, name(QQG5, replace)
```

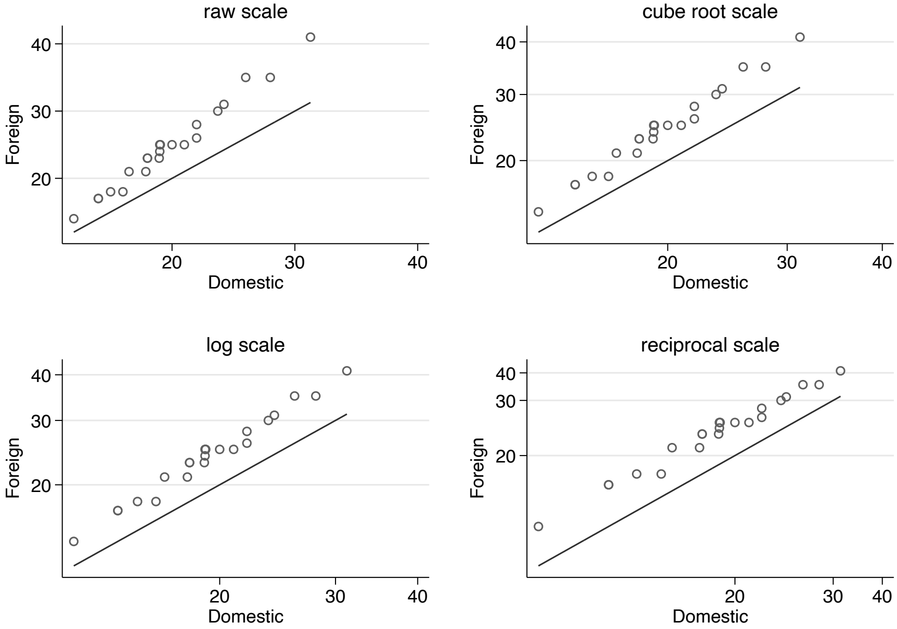

Figure 2. Quantile–quantile plots comparing miles per gallon for domestic and foreign cars on raw, cube root, logarithmic, and reciprocal scales

In specifying a transformation to `qqplotg`, we are limited only to what can be expressed using Stata functions or operators. Even less common transformations are often supported already through Stata functions (for example, `asinh()`, `log1p()`). Otherwise, they can still be expressed concisely. For example, folded roots for proportions in $[0, 1]$ as suggested by Tukey (1977) would be obtained using `sqrt(@) - sqrt(1 - @)`. In contrast, `ladder` and its siblings are restricted to a particular set of transformations.

Figure 2 shows the composite result, which seems fairly clear-cut. Cube root is not strong enough as a transformation to impart additive structure, but both logarithm and reciprocal work better, and the choice is between them. Formalizing choice as, say, a significance test procedure is best avoided. If you can see that one transformation works better, go for it. If you cannot see that, the choice of scale may not matter or can be made on other grounds. For example, there are many arguments that $y = \exp(Xb)$ is a more natural and more flexible pattern for systematic structure than the more traditional $y = Xb$, independently of the details of univariate or multivariate distributions. But that is a big theme that deserves longer and deeper discussion. Gould (2011) dives straight in on the key question.

If you wanted to proceed further with some simple model, you would not be committed to transforming the variable and then (for example) carrying out a Student's *t* test. You could work with a generalized linear model with a suitable link function, which would often be preferable.

In general, the effectiveness of a transformation of an entirely positive variable is tied up with the ratio of largest and smallest values, sometimes called the dynamic range. (A more subtle analysis is needed if values can be zero or negative as well as positive.) For miles per gallon, the dynamic range is 41/12 or about 3.4, which is large enough for transformation to be worth considering but not large enough for it to make an enormous difference in what you see. Some variables range over several orders of magnitude to the extent that use of a transformed scale is well nigh essential, at least for effective visualization. Examples are income or wealth, areas, populations, and plant height (Cox 2018).

Another simple guide to how much difference a transformation can make is given by plots of candidate functions over the observed range. The results are not shown here, but the commands just below show some technique. If a graph looks almost straight, the corresponding transformation does very little to change how the data are represented.

```
. twoway function raw = x, range(12 41)
. twoway function cube_root = x^(1/3), range(12 41)
. twoway function logarithm = ln(x), range(12 41)
. twoway function reciprocal = -1/x, range(12 41)
```

3.4 Two variables

Let us move now to comparing two variables. Our first example concerns comparison of one variable with expected quantiles from some reference distribution. Official Stata commands in this territory are `qnorm` and `qchi` for comparison with normal and chi-squared distributions, respectively. Many more commands have been community contributed, just as there are many more named distributions that might be candidates for comparison with data. More crucial than either fact is that typically generating such quantiles is a few lines of Stata (Cox 2007b, 2016a). The procedure boils down to calculating plotting positions, possibly obtaining parameter estimates, and pushing plotting positions through code for a quantile function.

Plots of quantiles against equivalent quantiles of a normal or Gaussian distribution are perhaps best now known as normal quantile plots. Other names in use (or disuse) are normal probability plot, normal scores plot, normal plot, probit plot, and fractile diagram.

Note that using `egen` is a little over the top here, but that device does extend easily to groups.

```
. egen rank = rank(mpg), unique
. egen n = count(mpg)
. summarize mpg
    Variable |        Obs        Mean    Std. dev.       Min        Max
-------------+---------------------------------------------------------
         mpg |         74     21.2973    5.785503         12         41
. generate normal = r(mean) + r(sd) * invnormal((rank - 0.5)/n)
. label variable normal "Expected normal quantiles"
. qqplotg mpg normal, name(QQG6, replace)
```

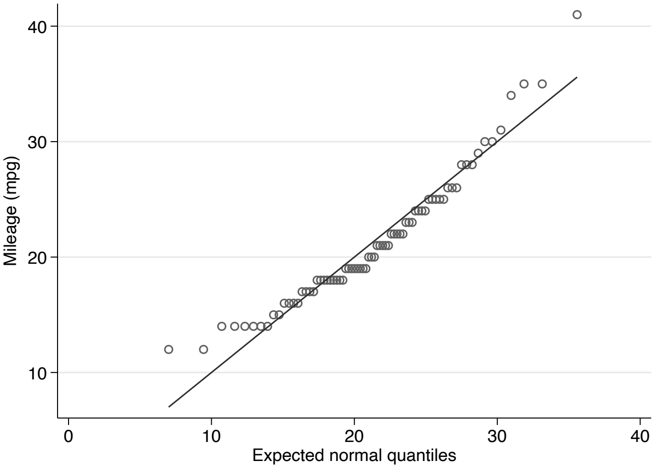

Figure 3. Normal quantile plot of miles per gallon. Note the systematic downward curvature imparted by positive skewness.

Figure 3 is a normal quantile plot for mpg, lumping domestic and foreign cars together. As before, a good fit to a normal distribution would mean that data points are close to the line of equality with only patternless variability around it. That is not what we see. Systematic downward curvature is symptomatic of a right-skewed distribution.

So we have produced a close copy of what qnorm can do. One reason for this demonstration is to underline that the calculations are simple in principle (Cox 2007b). A more immediate payoff is now being able to do something quite different with qqplotg.

3.5 Plotting differences between quantiles

When a reference situation is equality, say, $y = x$, we can recast any comparison as a comparison with a reference situation that differences are zero, $y - x = 0$. The reference line will now be horizontal if we plot the differences $y - x$ against something else, commonly the mean $(y + x)/2$. It could also be the sum $y + x$, a choice that yields the same plot but with different horizontal axis labels. The point is psychological and pragmatic. It can be easier to think about deviations from a horizontal line than from a sloping line. The rotated configuration may also make better use of the graph space.

Plotting difference versus mean is the leading example of a small graphical strategy, as urged by Tukey (1977, 153): "flattening by subtraction makes it much easier to see what is going on at more subtle levels."

Plots of difference versus mean go back at least to the early 1950s. I found a reference to Neyman, Scott, and Shane (1953) in Brillinger (2008) and would be delighted to hear of earlier examples. Indeed, multiple independent inventions might be expected. Other borrowings, rediscoveries, or reinventions in present practice include Bland–Altman plots (especially in medical statistics) and MA-plots in genomics (M and A stand for minus and average).[4]

The idea of plotting difference between quantiles versus mean quantile appears together with the idea of plotting difference between quantiles versus plotting position or cumulative probability in Wilk and Gnanadesikan (1968). Such plots are known as delta plots in psychology (De Jong, Liang, and Lauber 1994; Speckman et al. 2008). The options for such plots in `qqplotg` are `dvm` (as in "difference versus mean") and `dvp` (as in "difference versus cumulative probability" or "difference versus plotting position"). You can look at either or both: you just need to run the command twice for both, but as in section 3.3, you can name each plot and combine them.

Using plotting position for the horizontal coordinate raises the question of how it is calculated. Almost all practices are included within a single-index family: for sample size n, a indexes choices within $(\text{rank} - a)/(n + 1 - 2a)$. In particular, $a = 1/2$ leads to $(\text{rank} - 0.5)/n$, which as flagged in section 2 is the default for `quantile`; it is also the default in `qqplotg`. The family also includes other choices such as $a = 1/3$ and $a = 0$, which are both often advocated and often applied. The choice should almost never make any discernible difference for exploratory work, but some researchers have strong views on the best choice or at least may wish to follow a choice conventional in their field. The option `a()` is provided to accommodate any taste within this family. Whatever the choice, note that rank means distinct or unique ranks, so that all integers from 1 to n are assigned as ranks and tied values are shaken apart.

[4] The search for earliest uses or mentions of such simple ideas might be thought foolish. Informal uses before publication seem likely, the potentially relevant literature is enormous, and anyone smart enough to think of the idea would be likely to think it so obvious as not to be worth even a trumpet toot. Still, as supposedly said by Gershom Scholem, "Nonsense is always nonsense, but the history of nonsense is scholarship" (Pelikan 1988, 224).

Let us return to the mpg data. Here are two sample plots.

```
. qqplotg mpg normal, dvm lpolyopts(kernel(biweight) bw(2))
. qqplotg mpg normal, dvp lpolyopts(kernel(biweight) bw(0.1))
```

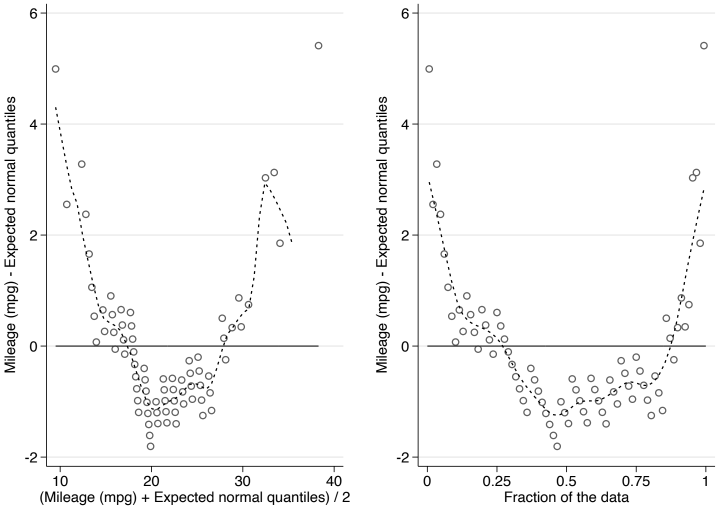

Figure 4. Normal quantile plot for miles per gallon, recast as difference versus mean and difference versus cumulative probability or plotting position. The downward curvature indicates positive skewness. See text for comments on the added smooth curves.

Figure 4 combines these plots. The syntax to produce the figure used `name()` options for the previous two commands followed by `graph combine`, as in section 3.3.

The `mpg` variable is being used as an example for several reasons beyond mere Stata tradition. It is simple enough in distribution to be a modest challenge but also shows some complications that are common in practice. One is the occurrence of tied values arising from the convention that values of miles per gallon are reported only as integers. Being honest and direct about ties is one worthwhile feature of quantile plots. In figure 3, for example, ties are evident as little horizontal runs of data points. In the graphs in figure 4, ties are evident as sloping lines, either exact or approximate straight lines.

The natural guess is that tied values would be shaken apart if we were given some decimal places beyond the integer values. Some values would be a little higher and some a little lower. Researchers can do this shaking apart mentally. Jittering quantile plots is always an option too, but not an option I find especially helpful.

In general, differencing can amplify noise. That together with the specific challenge of ties encourages the display of a smoothed curve. `qqplotg` uses local polynomial smoothing as implemented in `twoway lpoly`. Such smoothing should be taken as seriously as it deserves, but not more seriously. In essence, the goal is just exploratory, to yield a curve to guide the eye and brain. Neither `twoway lpoly` nor the `qqplotg` code sitting on top tries to optimize your curve, and indeed neither can have any sense of your goals or preferences. You should feel free to change the kernel type, degree, and bandwidth. For example, biweight kernels just happen to be a personal favorite that I find easy to explain to colleagues and students (Cox 2007a). Like Goldilocks, the researcher needs to avoid extremes, smoothing too much or smoothing too little, and like her, the researcher is the judge of what is about right.

By construction, data points on the `dvp` plot are equally spaced horizontally, which can help interpretation; but then again, the `dvm` plot might be thought closer to the original data. In the example here, we are not trying to say anything new about the data, which we already know to be nonnormal and positively skewed. The point is rather to show technique that may help with your own data, which will be much more interesting.

A keen reader may now wish to revisit the examples of sections 3.1 and 3.3, plotting differences versus the coordinate of your choice.

3.6 A different example

Let us look at a different example. A dataset supplied in the media for this issue of the *Stata Journal* holds measurements of maximum daily ozone concentrations for Yonkers, New York, and Stamford, Connecticut, in May to September 1974. Yonkers and Stamford are 21 miles (34 km) apart, although no winds promise to blow straight between them. The dataset was used by Chambers et al. (1983) and Cleveland (1993, 1994) to show quantile plot technique. The data are time series and are paired, but as the authors just mentioned, we will not touch on those key aspects of a full analysis.

```
. use ozone, clear

. qqplotg stamford yonkers, xla(0(50)150) yla(0(50)250) subtitle(raw scale)

. mylabels 10 20 50 100 200, myscale(ln(@)) local(la)
2.302585092994046 "10" 2.995732273553991 "20" 3.912023005428146 "50"
4.605170185988092 "100" 5.298317366548036 "200"
```

```
. qqplotg stamford yonkers, xla(`la') yla(`la') transform(ln) subtitle(log scale)
```

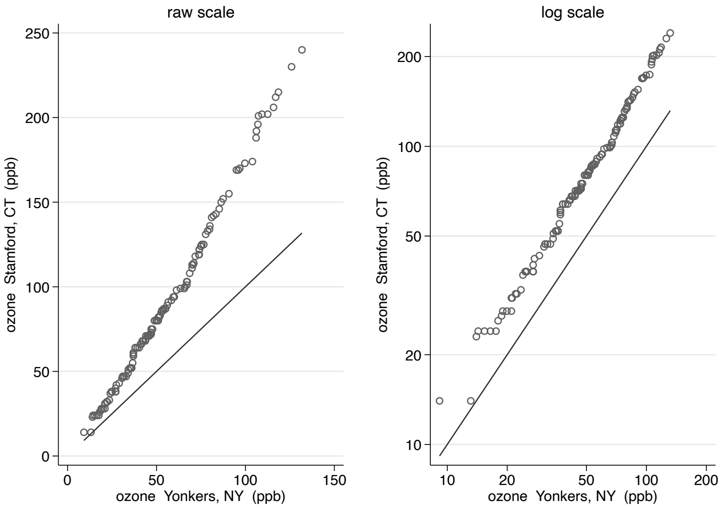

Figure 5. Quantile–quantile plots for ozone concentrations at Stamford and Yonkers in summer 1974, on both original and logarithmic scales

Figure 5 makes a simple case. Logarithmic scale seems appropriate. In essence, the pattern is of multiplicative structure on the raw scale and additive structure on the logarithmic scale. Dynamic ranges are much greater here than in the miles per gallon example: 240/14 or about 17 for Stamford, 132/9 or about 15 for Yonkers, and 240/9 or about 27 when the variables are pooled. A strong effect of transformation is expected with those values.

3.7 Support for by()

`qqplotg` also supports a `by()` option. This option does not contradict the principle that `qqplotg` is about comparing precisely two distributions: you can just do that repeatedly for different subsets defined by some other variable.

The ozone data contain two variables that could be used for experiment. Month of year could be a handle for exploring seasonality; it is a little arbitrary meteorologically. Day of week is for exploring a different kind of dependence and less arbitrary if, say, domestic or industrial emissions are different at weekends, or for other reasons in any weekly cycle.

The syntax might be something like

```
. qqplotg stamford yonkers, by(month, subtitle(log scale)) transform(ln)
> xla(`la') yla(`la')
```

The result is not shown here, but the option is there for experiment, preferably using your own more interesting data.

3.8 Support for generate

A `generate()` option is provided to allow saving calculated variables more permanently, say, for quite different graphs or calculations.

4 Details of qqplotg

4.1 Syntax

`qqplotg` *varname1 varname2* [*if*] [*in*] [`,` `a(`*str*`)` `flip` `rlopts(`*options*`)`
`transform(`*specification*`)` `by(`*byvar*`,` *byopts*`)` `missing` [`dvm`|`dvp`]
`lpolyopts(`*options*`)` *graph_options* `generate(`*stub*`)`]

`qqplotg` *varname* [*if*] [*in*]`,` `over(`*groupvar*`)` [`a(`*str*`)` `flip` `rlopts(`*options*`)`
`transform(`*specification*`)` `by(`*byvar*`,` *byopts*`)` `missing` [`dvm`|`dvp`]
`lpolyopts(`*options*`)` *graph_options* `generate(`*stub*`)`]

4.2 Description

`qqplotg` plots the quantiles of one distribution against the quantiles of another distribution. Here quantiles means ordered values. It is a generalization of official command `qqplot`. Names for this plot include quantile–quantile plot and q–q or Q–Q plot.

The two distributions may be of unequal size: if so, corresponding quantiles are calculated by interpolation.

There are two main syntaxes. In the first, emulating `qqplot`, the two distributions are given by the values of two variables, *varname*1 and *varname*2.

In the second, the distributions are given by the values of *varname* for two distinct groups of *groupvar* named in the compulsory option `over()`. The help for `qqplot` explains how to set up such a plot, but a one-line command may be convenient.

By default, a reference line of equality is shown to aid in identifying any systematic or random differences between the two distributions.

Optionally, the distributions may be plotted as differences between corresponding quantiles versus their means; or as differences between corresponding quantiles versus a fraction of the data (also known as cumulative probability or plotting position). In each case, a smooth will be added using `twoway lpoly` of the difference over its support.

Transformations on the fly are supported. It is suggested as essential practice to supply an informative note or title (unless an informative text caption is given otherwise); and as good practice to use axis labels on the original scale. See `nicelabels` and `mylabels` (Cox 2022) for support.

4.3 Options

`over(`*groupvar*`)` is a required option whenever you need to specify a group variable that takes on precisely two distinct numeric or string values.

`group(`*groupvar*`)` is allowed as a synonym.

`a(`*str*`)` specifies a within the plotting position recipe $(i-a)/(n+1-2a)$ for distinct or unique ranks i running over the integers from 1 to sample size n. The default is `a(0.5)`, yielding $(i-0.5)/n$. Alternatives should specify a number such as `a(0)` or a numeric expression such as `a(1/3)`. For more detail, see Cox (2016a).

`flip` swaps axes as compared with the default. This can be especially helpful when a first pass shows that two groups would be better plotted the other way but you have no desire to recode *groupvar*.

`rlopts(`*options*`)` may be used to tune the rendering of reference lines.

`transform(`*specification*`)` specifies a transformation to apply to what is plotted on both axes. There are two syntaxes. 1) A bare function name such as `ln` or `sqrt` will be applied directly. Do not supply parentheses `()`. 2) An expression mentioning `@` will be applied with `@` replaced on the fly with the appropriate variable name. Hence, `@^(1/3)` specifies cube roots of zero or positive values, and `1/@` specifies reciprocals. If `dvm` or `dvp` is also specified, then transforms are calculated first.

A warning will be displayed if the transform creates missing values. For example, taking logarithms of zero or negative values would do that.

`by(`*byvar*`,` *byopts*`)` is supported to produce separate plots for the distinct values of a variable *byvar*. By default, *byopts* includes `legend(off) note("")`. Missing values of *byvar* will be ignored unless the further option `missing` is specified.

`dvm` plots differences between corresponding quantiles versus their means as an alternative to plotting quantile versus quantile. The reference becomes the horizontal line defining difference zero.

`diffvsmean` is allowed as a synonym.

`dvp` plots differences between corresponding quantiles versus their plotting positions as an alternative to plotting quantile versus quantile. The reference becomes the horizontal line defining difference zero.

`dvm` and `dvp` may not be specified together.

`lpolyopts(`*options*`)` are options of `twoway lpoly` that tune the smooth that appears with option `dvm` or `dvp`. Note that `lpolyopts(nodraw)` suppresses display of such graphs.

graph_options are other options allowed with `scatter`. Specifically, `aspect(1)` may be a good idea with quantile–quantile plots.

`generate(`*stub*`)` generates the quantiles as two new variables, variously *stub*`1` and *stub*`2`; or *stub*`d` and *stub*`m` if `dvm` is also specified; or *stub*`d` and *stub*`p` if `dvp` is also specified.

5 Conclusions

A recurrent issue with Stata, or any comparable software, is the small tension between writing a few command lines yourself to travel a modest distance—say, from A to E via B, C, and D—as compared with having access to a command that does B, C, D for you. I recall a conversation with a medical statistician at a very early Stata users' meeting. I was enthusing about the scope for programming some task that was not wired into an existing Stata command. Their reply to the effect that needing to write code was a distraction from their main focus on doing statistics was entirely correct too.

Quantile plots (or quantile–quantile plots) are a case in point. Cox (2007b) was all about plots being accessible through simple steps. This article is mostly about a new command that, as it were, hides much small nitty-gritty from the user.

Quantile plots generally are, as readers will realize by now, a strong personal favorite. `qqplotg` is the latest evangelizing effort to persuade Stata users to try out such plots in a campaign that has so far extended over 25 years. Quantile plots carry all the information in the data and often answer the key question of comparison directly.

6 Acknowledgment

Pat Branton of StataCorp kindly provided historical information on commands in 1985 and 1986.

7 Programs and supplemental materials

To install a snapshot of the corresponding software files as they existed at the time of publication of this article, type

```
. net sj 24-3
. net install gr0096      (to install program files, if available)
. net get gr0096          (to install ancillary files, if available)
```

8 References

Anscombe, F. J. 1961. Examination of residuals. In *Proceedings of the 4th Berkeley Symposium on Mathematical Statistics and Probability.* Vol. 1, *Contributions to the Theory of Statistics*, ed. J. Neyman, 1–36. Berkeley, CA: University of California Press.

———. 1981. *Computing in Statistical Science through APL.* New York: Springer. https://doi.org/10.1007/978-1-4613-9450-1.

Anscombe, F. J., and J. W. Tukey. 1963. The examination and analysis of residuals. *Technometrics* 5: 141–160. https://doi.org/10.2307/1266059.

Box, G. E. P., and D. R. Cox. 1964. An analysis of transformations. *Journal of the Royal Statistical Society,* B ser., 26: 211–243. https://doi.org/10.1111/j.2517-6161.1964.tb00553.x.

Brillinger, D. R. 2008. The 2005 Neyman Lecture: Dynamic indeterminism in science. *Statistical Science* 23: 48–64. https://doi.org/10.1214/07-STS246.

Chambers, J. M., W. S. Cleveland, B. Kleiner, and P. A. Tukey. 1983. *Graphical Methods for Data Analysis.* Belmont, CA: Wadsworth. https://doi.org/10.1201/9781351072304.

Cleveland, W. S. 1993. *Visualizing Data.* Summit, NJ: Hobart.

———. 1994. *The Elements of Graphing Data.* Rev. ed. Summit, NJ: Hobart.

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.

———. 2005. Speaking Stata: The protean quantile plot. *Stata Journal* 5: 442–460. https://doi.org/10.1177/1536867X0500500312.

———. 2007a. Kernel estimation as a basic tool for geomorphological data analysis. *Earth Surface Processes and Landforms* 32: 1902–1912. https://doi.org/10.1002/esp.1518.

———. 2007b. Stata tip 47: Quantile–quantile plots without programming. *Stata Journal* 7: 275–279. https://doi.org/10.1177/1536867X0700700213.

———. 2011. Stata tip 96: Cube roots. *Stata Journal* 11: 149–154. https://doi.org/10.1177/1536867X1101100112.

———. 2016a. FAQ: How can I calculate percentile ranks? https://www.stata.com/support/faqs/statistics/percentile-ranks-and-plotting-positions/.

———. 2016b. Speaking Stata: Letter values as selected quantiles. *Stata Journal* 16: 1058–1071. https://doi.org/10.1177/1536867X1601600413.

———. 2018. Speaking Stata: Logarithmic binning and labeling. *Stata Journal* 18: 262–286. https://doi.org/10.1177/1536867X1801800116.

———. 2022. Speaking Stata: Automating axis labels: Nice numbers and transformed scales. *Stata Journal* 22: 975–995. https://doi.org/10.1177/1536867X221141058.

De Jong, R., C.-C. Liang, and E. Lauber. 1994. Conditional and unconditional automaticity: A dual-process model of effects of spatial stimulus-response concordance. *Journal of Experimental Psychology: Human Perception and Performance* 20: 731–750. https://doi.org/10.1037/0096-1523.20.4.731.

Finney, D. J. 1977. Dimensions of statistics. *Journal of the Royal Statistical Society, C ser.*, 26: 285–289. https://doi.org/10.2307/2346969.

Fisher, R. A., and F. Yates. 1938. *Statistical Tables for Biological, Agricultural and Medical Research.* Edinburgh: Oliver and Boyd.

Galton, F. 1883. *Inquiries into Human Faculty and its Development.* London: Macmillan. https://doi.org/10.1037/14178-000.

———. 1907. Grades and deviates: Including a table of normal deviates corresponding to each millesimal grade in the length of an array, and a figure. *Biometrika* 5: 400–404. https://doi.org/10.1093/biomet/5.4.400.

Gelman, A., J. Hill, and A. Vehtari. 2021. *Regression and Other Stories.* Cambridge University Press: Cambridge. https://doi.org/10.1017/9781139161879.

Gibbings, J. C. 2011. *Dimensional Analysis.* London: Springer. https://doi.org/10.1007/978-1-84996-317-6.

Gould, W. 2011. Use poisson rather than regress; tell a friend. *The Stata Blog: Not Elsewhere Classified.* https://blog.stata.com/2011/08/22/use-poisson-rather-than-regress-tell-a-friend/.

Kendall, M. G. 1940. Note on the distribution of quantiles for large samples. *Supplement to the Journal of the Royal Statistical Society* 7: 83–85. https://doi.org/10.2307/2983633.

Mahajan, S. 2010. *Street-Fighting Mathematics: The Art of Educated Guessing and Opportunistic Problem Solving.* Cambridge, MA: MIT Press.

———. 2014. *The Art of Insight in Science and Engineering: Mastering Complexity*. Cambridge, MA: MIT Press.

McCullagh, P. 2022. *Ten Projects in Applied Statistics*. Cham, Switzerland: Springer. https://doi.org/10.1007/978-3-031-14275-8.

McCullagh, P., and J. A. Nelder. 1989. *Generalized Linear Models*. 2nd ed. London: Chapman and Hall/CRC. https://doi.org/10.1201/9780203753736.

Mosteller, F., and J. W. Tukey. 1977. *Data Analysis and Regression: A Second Course in Statistics*. Reading, MA: Addison–Wesley.

Neyman, J., E. L. Scott, and C. D. Shane. 1953. On the spatial distribution of galaxies: A specific model. *Astrophysical Journal* 117: 92–133. https://doi.org/10.1086/145671.

Pelikan, J. 1988. *The Melody of Theology: A Philosophical Dictionary*. Cambridge, MA: MIT Press.

Santiago, J. G. 2019. *A First Course in Dimensional Analysis: Simplifying Complex Phenomena Using Physical Insight*. Cambridge, MA: MIT Press.

Speckman, P. L., J. N. Rouder, R. D. Morey, and M. S. Pratte. 2008. Delta plots and coherent distribution ordering. *American Statistician* 62: 262–266. https://doi.org/10.1198/000313008X333493.

Tukey, J. W. 1957. On the comparative anatomy of transformations. *Annals of Mathematical Statistics* 28: 602–632. https://doi.org/10.1214/aoms/1177706875.

———. 1977. *Exploratory Data Analysis*. Reading, MA: Addison–Wesley.

Unwin, A. 2015. *Graphical Data Analysis with R*. Boca Raton, FL: CRC Press.

Wilk, M. B., and R. Gnanadesikan. 1968. Probability plotting methods for the analysis of data. *Biometrika* 55: 1–17. https://doi.org/10.2307/2334448.

Wilke, C. O. 2019. *Fundamentals of Data Visualization: A Primer on Making Informative and Compelling Figures*. Sebastopol, CA: O'Reilly.

About the author

Nicholas Cox is a statistically minded geographer at Durham University. He contributes talks, postings, FAQs, and programs to the Stata user community. He has also coauthored 16 commands in official Stata. He was an author of several inserts in the *Stata Technical Bulletin* and is Editor-at-Large of the *Stata Journal*. His "Speaking Stata" articles on graphics from 2004 to 2013 have been collected as *Speaking Stata Graphics* (2014, College Station, TX: Stata Press). He is the Editor of *Stata Tips, Volumes I and II* (2024, also Stata Press).