# The Stata Journal

The *Stata Journal*, electronic version (ISSN 1536-8734) is a publication of Stata Press, and Stata is a registered trademark of StataCorp LP.

The Stata Journal (2005)
**5**, Number 3, pp. 442–460

# Speaking Stata: The protean quantile plot

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

**Abstract.** Quantile plots showing by default ordered values versus cumulative probabilities are both well known and also often neglected, considering their major advantages. Their flexibility and power is emphasized by using the `qplot` program to show several variants on the standard form, making full use of options for reverse, ranked, and transformed scales and for superimposing and juxtaposing quantile traces. Examples are drawn from the analysis of species abundance data in ecology. A revised version of `qplot` is formally released with this column. Distribution plots in which the axes are interchanged are also discussed briefly, in conjunction with a revised version of `distplot`, also released now.

**Keywords:** gr0018, qplot, distplot, distributions, quantile plots, statistical graphics, species abundance, ecology, Whittaker plots, broken stick, lognormal, power laws, scaling laws

## 1 Introduction

In my last column (Cox 2005), I started to look at some personal favorite methods that somehow or other lurk beyond the bounds of what is standard. The first method examined was that of density probability plots, which have attracted very few users in the decade since they were proposed. For this column, we stay in the same territory of distribution graphics. The method to be examined now is that of quantile plots for showing one or more distributions. Regarding these as not standard may appear more disputable. The advantages of quantile plots are first discussed, making clear that they provide solutions to the main limitations of the much more popular histograms and box plots. Examples of standard ecological analyses are then used to show how versatile quantile plots can be.

Quantile plots graph a set of ordered values[1] against the so-called plotting positions, in essence the associated cumulative probabilities. They may well be in the routinely used toolkit of several readers. Quantile plots have been described in detail in various leading texts (e.g., Chambers et al. 1983; Cleveland 1993, 1994). An official Stata implementation has been available for some time in `quantile`. An extended user-written implementation has also been available in `qplot`, formerly `quantil2` (Cox 1999b). Moreover, I have previously discussed quantile plots together with other methods for graphing distributions (Cox 2004b). What then are the grounds for regarding quantile plots as neglected?

[1] Otherwise put, the (sample) quantiles for our purposes are identical to the order statistics. This terminology may seem loose if you are accustomed to other definitions of quantiles, but it is standard in the statistical graphics literature. For a review of order statistics, see David and Nagaraja (2003).

## 2 Quantile plots are needed

In broad terms, the case is that quantile plots are even more useful than is widely realized. Specifically, this column coincides with the release of a further-enhanced version of `qplot`. More generally, there are several grounds for promoting their use. The arguments differ to some extent, depending on whether teaching or research is the main issue.

At an introductory level, the most-popular methods taught for showing univariate distributions of variables that are continuous, or nearly so, appear to be histograms and box plots. Dot plots are a popular third alternative in some quarters. The choice of what to teach is usually guided by a need to explain methods that may be encountered by students or practitioners in their later courses or subsequent careers, so selection is often conservative.

The merits of histograms and box plots do not need much emphasis, but their limitations do require brief mention. Without saying everything that could be said, here are some standard comments.

1. The binning that is the basis of histograms is often awkward. Every histogram depends on decisions about bin width and origin. Texts thus have to explain the compromise required between undersmoothing that omits too little detail and oversmoothing that omits too much. Although learners do not find the principle difficult to understand, in practice they often make poor choices with their own histograms. Arguably, good choice of bin width depends on statistical experience—precisely what introductory audiences rarely have. Without that, there is much scope for small or large misinterpretations over key detail in the tails, real and imagined modality, and even gross distribution shape. For example, slight but notable skewness can easily be missed with poor bin choices.

2. Histograms can be excellent for individual distributions, but too often they tend to be dead ends leading nowhere else. Overlay of a curve (especially some theoretical or fitted density function) can be very useful, but the comparison of two or more histograms is frequently difficult. Neither juxtaposition (histograms side by side) nor superimposition (histograms on top of each other) usually works well, except for the crudest propaganda purposes and a small number of histograms. (Look! The distributions are almost the same! Or: Look! The distributions are very different!) Ask yourself: How many histograms could you superimpose and still have your readers come to quantitative conclusions easily and effectively? How many histograms could you juxtapose with the same result?

3. Conversely, box plots can be excellent for comparisons but arguably have been oversold for showing single batches of the data. With, say, dozens of groups in the data, the condensation imparted by box plots can be vital for seeing overall patterns together with gross anomalies. With one or just a few groups of data, box plots can show too little about what is going on both in the central box and in the tails. This is one reason for the steadily growing popularity of dot plots.

The limitations of these most popular choices correspond to the strengths of quantile plots.

1. Quantile plots require no decisions about bin width or origin or similar matters. There is scope for varying axis scales from the default, but this is a feature, not a problem to be solved.
2. Quantile plots show empirical curves that are weakly monotonic. These are relatively easy to compare, certainly considering the difficulties of comparing histograms. Typically, users choose point or line representations. In Stata, the flexibility possible over plottype arises through the scope for using `recast()` with `quantile` or `qplot` to produce a different kind of graph. Moreover, in the latest version of `qplot`, the user can take advantage of either or both an `over()` option and a `by()` option to show superimposed and juxtaposed traces as desired.
3. Quantile plots in principle show all the information in the distribution but without either emphasizing or suppressing any particular data points. Thus fine structure in the tails or indeed anywhere else is evident. Contrast the case of histograms, which may suppress much or all of the fine structure through their use of bins, and the case of box plots, which may flag apparent outliers but suppress virtually all other detail.

Unsurprisingly, quantile plots have disadvantages, too. We make two comments now and will say more later.

1. For some groups, they may seem unfamiliar, but this is precisely where we came in. You need to learn how to "read" quantile plots, just as you do any other kind of graph. Managers, advisors, or editors (but not the Editors of the *Stata Journal*, naturally) might insist that you use what is to them a more familiar kind of graph, but exercise of power is not the best way to win an argument.
2. More substantially, if your problem is better illuminated by looking at densities, as when the modality of a distribution is of concern, a density trace (such as that produced by `kdensity`) will clearly be more appropriate.

The definition of quantile plot here has been quite strict: we insist that quantiles are plotted as response on the vertical axis. Many readers would be happy with the idea that distribution plots, in which cumulative probability or frequency is plotted on the vertical axis versus quantiles on the horizontal axis, are just a variant on quantile plots, with the main difference that the axes are interchanged. More will be said on this point later.

More broadly, emphasis on quantiles rather than cumulative distributions is more than a matter of cosmetics or aesthetics. The tendency persists in many quarters of introducing the quantile function rather grudgingly and indirectly as the inverse of the distribution function. Authors are perhaps here recapitulating the historical order, rather than following the most natural scientific route. Be that as it may, whenever the quantiles are essentially the response of interest, they belong on the vertical axis.

## 3 Abundance distributions in ecology

A key problem in ecology is to compare counts from several sites of the numbers of individuals in various different species. Almost always, the attention is not to all organisms present but is restricted to groups of fairly closely related species, say, birds, lizards, fish, or trees. Typically, there is much variation from a few very common species to rather more (even very many more) rarer species at each site. How can we best report such data in terms of distributions fitted or informative summary measures? This is usually preliminary to some kind of assessment of relationships with other ecological or environmental variables.

This can be seen as the prototype for several related problems. Focus at the species level is common, but not universal, as other taxa (units on other taxonomic levels, say, genera or subspecies) could be used. Counting individuals is often the most straightforward method, but with some organisms (e.g., corals, mosses), the definition of an individual can be problematic, and in any case, it may be desired to measure surface cover, biomass, productivity, or some other property. In our examples, we will focus on counting. Little imagination is needed to appreciate that even at restricted sites and with appropriate taxonomic skills, ecologists generally face a sampling problem, as some species may never be seen within a particular sample. Only in a few cases, say, recording all trees present at a temperate forest site, is anything like a complete census possible.

This basic problem is part of the descriptive science underlying the analysis of ecological diversity, or biodiversity, that is so prominent in many environmental discussions. The literature is enormous. Magurran (2004) offers a fairly informal introduction with worked examples. Lande, Engen, and Sæther (2003) relate the analysis of diversity to population dynamics at a deeper level. May (1975) remains an outstanding blend of ecological, mathematical, and statistical thinking. Hutchinson (1978) tastefully embeds the story in the wider history of ideas.

An ecological problem is being used here as a vehicle for some graphical ideas, but no discipline possesses totally idiosyncratic questions. In particular, there are many parallels with problems in what ostensibly is a very different field, analysis in economics of the distribution of incomes, firm sizes, and so forth. Kleiber and Kotz (2003) give an excellent review of the literature. In each case, much of the challenge lies in the handling of very skewed and long-tailed distributions. In each case, problems of sampling, censoring, and truncation can be acute. The jargon in ecology is of diversity, species richness, and species evenness, and in economics of inequality and concentration, but the issues remain sisters under the skin. The pitfalls too show resemblances. Much time and effort have gone in each discipline into searches for summary measures (indexes, coefficients, and what have you) that somehow capture the interesting information, remain interpretable theoretically and are robust to whatever investigators regard as perversities in their data. In each case, skeptics intermittently inject notes of caution about such an enterprise, while others keep on hoping. Despite these similarities, cross-fertilization of literatures has been very slight[2]. What is more common is multiple rediscovery of the same methods, or ignorance of relevant solutions elsewhere. What to ecologists is the Simpson index is to economists the Herfindahl index and to yet others one of the many beasts called a Gini index, to give one small example.

## 4 Birds in the Killarney woods

Let us consider the best structure in Stata for ecological abundance data. If only presence is recorded, each observation might as well be site name, species name, and number of individuals. Thus we represent frequencies in a long structure, to use the terminology of [D] **reshape**. A wide data structure with either each site as a variable, or each species as a variable, has some advantages for another analyses but can be decidedly awkward for graphics, as we are likely to hit against limits on the number of variables that can be shown on a graph. In some problems, the species names can be ignored—or were never supplied in literature reports—so that data take on the form of frequencies of frequencies, that is, how many species are represented by 1, 2, 3, ... individuals.

### 4.1 Whittaker plots

A particular variant on the quantile plot is often now recommended by ecologists as the best single starting point for showing data of this kind. Abundance on a logarithmic scale should be plotted against the rank of each species. Krebs (1989, 344) called such a plot a *Whittaker plot* after the leading American ecologist Robert Harding Whittaker (1920–1980), who used them prominently. Ecologists follow the opposite convention from statisticians and assign rank 1 to the most common species. The fact that the least-common species may be difficult or even impossible to determine is grounds enough for this convention.

Figure 1 shows a simple example of a Whittaker plot. The data on bird species diversity come from Magurran (2004, 237–238), who in turn got them from a study by Batten (1976, 307, 309, 313) in contrasting woodlands in Killarney, Ireland. The rationale for the log scale is clear: even in this rather small and simple dataset, there is a 35-fold variation between the most- and least-common species. On a log scale, with rank on a linear scale, a geometric series would show as a straight line, and a logarithmic series distribution (Fisher, Corbet, and Williams 1943) as very nearly a straight line. On the evidence here, there is some support for going further with either of those models, but also a worry that the dataset is rather small.

[2] Simon (1991, 367) wrote that "Disciplines, like nations, are a necessary evil that enable human beings of bounded rationality to simplify their goals and reduce their choices to calculable limits".

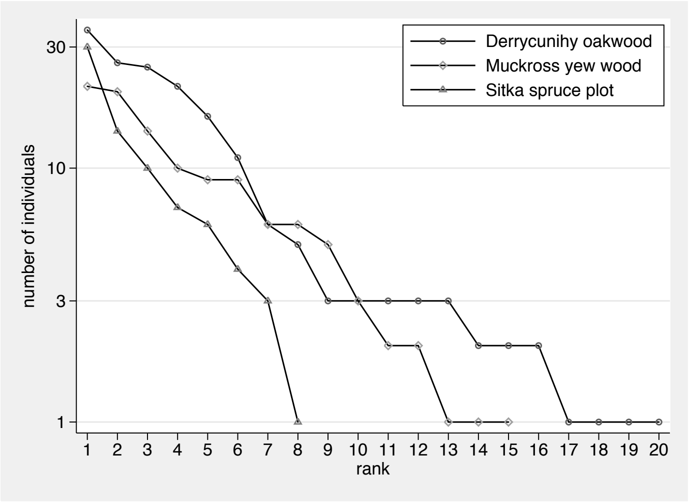

Figure 1: Whittaker plot showing bird diversity in three woodlands in Killarney, Ireland. The vertical log scale gives a good first look at the distributions.

The command for this plot is

```
. qplot count, over(site) rank reverse recast(connected) yscale(log)
> ytitle(number of individuals) ylabel(1 3 10 30, angle(h)) xlabel(1/20)
> legend(position(1) ring(0) column(1))
```

The `over(site)` option gives multiple traces, one for each site, in each panel shown; here just one is shown. An alternative, not so helpful here, would be `by(site)`, giving curves for each site in separate panels. The `rank` and `reverse` options spell out the scale for the horizontal axis. The other options are just standard `twoway` options.

A common alternative is to show abundance as a percent of each site total (figure 2). That requires just a prior `egen` call:

```
. egen pc = pc(count), by(site)
. qplot pc, over(site) rank reverse recast(connected) ysc(log)
> yti(percent of total) yla(1 3 10 30 100, ang(h)) xla(1/20) legend(pos(1)
> ring(0) col(1))
```

(*Continued on next page*)

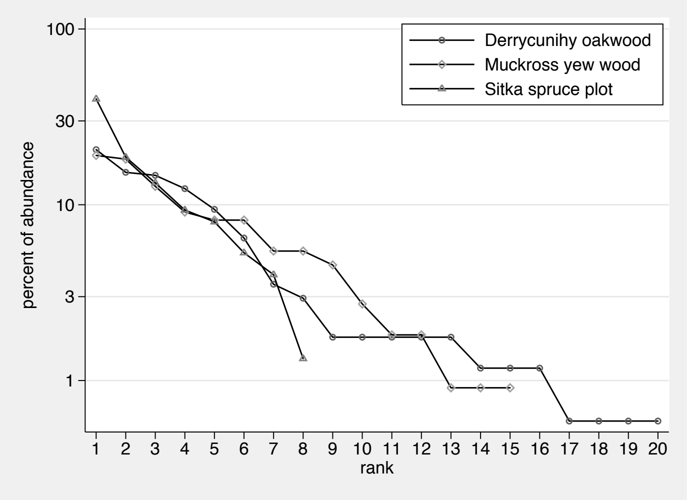

Figure 2: Whittaker plot showing bird diversity for three woodlands in Killarney, Ireland. The vertical scale shows percent of total abundance.

Clearly, proportions between 0 and 1 are possible, as well as percentages.

## 4.2 Broken stick distributions

Several distributions have been entertained in the ecological literature for diversity analysis, in addition to the logarithmic series mentioned in passing above. Extra interest attaches to those with some more or less plausible ecological rationale or interpretation, at least as a kind of null or reference model. We will look in the first instance at two such distributions, the so-called broken stick and the lognormal, to show how quantile plots (and `qplot` in particular) can be used in initial exploration of the applicability of some models. One key point that will be understated is the extent to which the distributions mentioned are special or limiting cases of other more flexible distributions, such as the gamma or the negative binomial (e.g., Lande, Engen, and Sæther 2003).

The name of the broken stick arises from one way of visualizing the splitting of a whole randomly into parts, each part here corresponding to one of $S$ species. Suppose that a stick of unit length is broken at random in $S-1$ places, thus yielding $S$ parts, whose lengths we order as $p_1 \geq p_2 \geq \cdots \geq p_S$. The mean lengths of these ordered proportions $p_i$ can be shown to be

$$p_i \;=\; \frac{1}{S}\sum_{r=i}^{S}\frac{1}{r}$$

and so, given a total of $N$ individuals, the expectation is of

$$Np_i \;=\; \frac{N}{S}\sum_{r=i}^{S}\frac{1}{r}$$

individuals in the $i$th most common species. Note that there are no parameters to be estimated, an interesting detail. This fact naturally does not remove all statistical flavor from the problem: the breaking of the stick is probabilistic, so there will be variance around each mean, a point deserving more attention in a fuller account.

This general rule can be cross-checked against other arguments for the simplest interesting case, splitting into $S = 2$ species. The two parts have average proportions

$$p_1 = (1 + 1/2)/2 = 0.75 \qquad \text{and} \qquad p_2 = (1/2)/2 = 0.25$$

which matches an arm-waving argument that on average the break will be midway from the end of the half in which it occurs. Mosteller (1965, 62–65) goes carefully through the cases of $S = 2$ and $S = 3$ and gives the general result. Cohen (1966) is an elegant monograph (a reworking of the author's Bachelor's dissertation) linking the broken stick to ecological and economic theory and examples.

The broken-stick distribution has been rediscovered in various ways over the last century or so, and you may know it under other names. The trail goes back at least as far as William Allen Whitworth (1840–1905) and his textbook *Choice and Chance* (Whitworth 1905). Whitworth studied mathematics at Cambridge and then taught the subject in schools and University College, Liverpool. His later career developed as a minister in the Church of England. Whitworth's other books range from *Trilinear Coordinates* to the posthumous collection of sermons *The Sanctions of God*. For other information, see Irwin (1967) and Domb (1990).

The calculation of the broken-stick frequencies can be achieved in a few lines of Stata, but some small details deserve commentary.

```
. gsort site -count
. by site: gen double bstick = 1/_n
. by site: egen N = total(count)
. by site (bstick), sort: replace bstick = N * sum(bstick)/_N
```

The reciprocals of ranks $1/r$ are to Stata `1/_n`, so long as we have the right sort order. Recall that ecologists use the reverse-ranking convention, so that `gsort` is used to get the sort order needed, the minus sign specifying sorting from largest downwards. `double` is a precaution, as the reciprocals can be rather small and we want to carry as much precision as possible.

There is a clash between Stata's notation and that customary in the ecological literature. As each observation records a species frequency, the number of species $S$ is given directly by `_N`. The number of individuals $N$ is obtained by an `egen` call. From Stata 9, what used to be `egen, sum()` has been renamed `egen, total()` in an attempt to distinguish it more clearly from the `sum()` function.

The remainder of the calculation is a cumulative sum using `sum()`, together with division by $S$ and multiplication by $N$. A delicate point here is that cumulation from the lowest ranking species upwards requires reversing the sort order from that produced by `gsort`. It is tempting to think (as I did in one stab at this problem) that `sort site count` reverses `gsort site -count` within each site, but that would overlook the effect of ties, which can be very common with this kind of data. Although tied frequencies will by definition be identical, the associated reciprocals `1/_n` will not be. Reversing by `sort site bstick` is safe.

If this were a routine problem, then naturally, thought would be given to encapsulating code in a program yielding broken-stick predictions directly.

A partial check on calculations is given by seeing whether observed and predicted frequencies have the same total:

```
. tabstat bstick count, s(sum) by(site)
```

This example of broken-stick calculation points up a standard Stata moral. You can do a lot of work with careful use of `by:`, `_n`, and `_N`, but you have to get exactly the right sort order for the problem. An earlier column (Cox 2002) discussed these features in more detail.

That done, we can now put the broken-stick quantiles on the plot. We have two variables, observed and fitted, and three sites, so we need to use a `by()` option to specify separate panels. We revert to the original scale for the quantiles (figure 3).

```
. qplot count bstick, by(site, row(1)) rank reverse recast(connected)
> yti(number of individuals) yla(, ang(h)) xla(1 5(5)20)
> legend(order(1 "observed" 2 "broken stick") pos(6) row(1) ring(0))
```

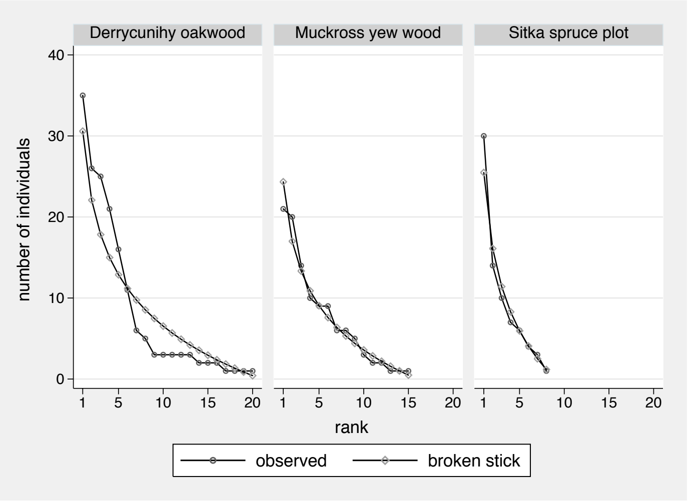

Figure 3: Broken sticks fitted for three woodlands in Killarney, Ireland. We revert to arithmetic scale for responses.

A good tip for broken sticks is that they are close to linear if plotted with a logarithmic scale for ranks (figure 4). The agreement I would describe as fair for the oakwood and better for the other two sites.

```
. qplot count bstick, by(site, row(1) note("")) rank reverse recast(connected)
> yti(number of individuals) yla(, ang(h)) xsc(log) xla(1/5 7 10 15 20)
> legend(order(1 "observed" 2 "broken stick") pos(6) row(1) ring(0))
```

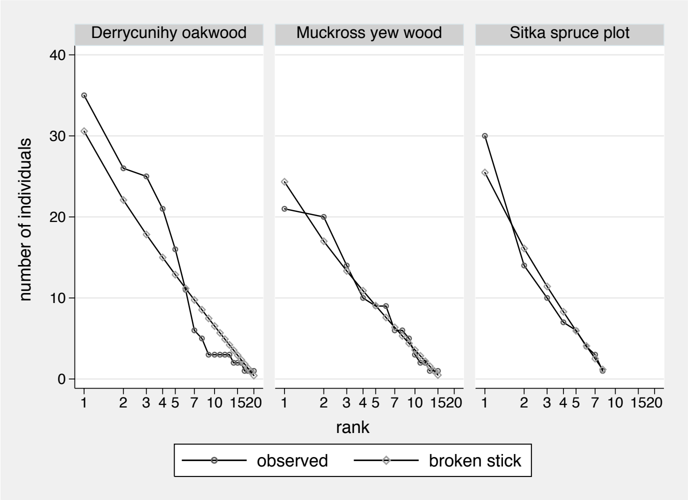

Figure 4: Broken sticks fitted for three woodlands in Killarney, Ireland. The horizontal log scale makes the broken sticks more nearly linear.

## 4.3 Lognormal distributions

Yet another commonly used distribution as a null or reference case is the lognormal. When fitting this distribution, ecologists most frequently draw histograms, often agonizing at some length about the choice of base of logarithms and how to bin. Whatever is chosen, there will be some loss of detail. Unsurprisingly, we remain with quantile plots and choose a log scale for the responses and an inverse-normal scale for the horizontal axis. This is done by specifying the `trscale()` option, which allows on-the-fly transformation of the horizontal-axis scale. `@` is a place-holder for what would be plotted otherwise, probability or rank. Stata 8 users should here type `invnorm()`, not `invnormal()`. The online help for `qplot` explains how to get a labeling of the horizontal axis in terms of percent points, such as 1 2 5 10(10)90 95 98 99. Hydrologists would call such a display for stream discharges a flow-duration curve (e.g., Gordon et al. 2004).

```
. qplot count, over(site) reverse trscale(invnormal(@))
> xti(standard normal scale) recast(connected) ysc(log)
> yti(number of individuals) yla(1 3 10 30, ang(h))
> legend(pos(1) ring(0) col(1))
```

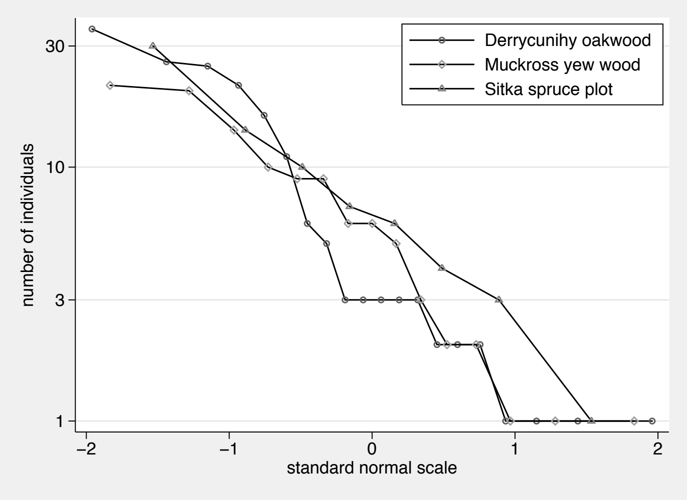

Figure 5: Lognormal distributions fitted for three woodlands in Killarney, Ireland. Note the vertical log scale and the horizontal inverse normal scale.

Again the fits look fairly good, leaving us with an embarrassment of choices and a feeling that we would need more data to decide confidently between these distributions. A further worry is whether we are being too cavalier in ignoring the discreteness of the counts. In particular, some ecologists have fitted Poisson lognormals (Bulmer 1974).

## 5 Macrolepidoptera at Rothamsted 1935 and the temptations of power laws

The main purpose here is to show off quantile plots, so leaving the previous example in midair does not feel too treacherous. Let us look at a more extensive dataset on macrolepidoptera (butterflies and the larger moths) caught in a light trap at Rothamsted (Hertfordshire, England) in 1935 and reported by Williams (1964, 25) (and also by Magurran 2004, 216). 197 species and 6815 individuals were caught, so the data are reported compactly as frequencies of frequencies; that is, 37 species are represented by a single individual, 22 species by two individuals, and so forth. We read them into Stata in that form and then used `expand count` to get an observation for each species.

The name Rothamsted will suggest to many readers the names of statisticians who have worked there at one time or another, including Sir Ronald Fisher, Frank Yates, and John Nelder. The overlap at Rothamsted of Fisher and the statistically minded entomologist Carrington Bonsor Williams (1889–1981) undoubtedly aided their collaboration[3]. This culminated in the key logarithmic series distribution paper of Fisher, Corbet and Williams (1943), one of the main roots of statistical ecology[4].

We will show just one histogram to make a point; it does show the main feature of this distribution clearly (figure 6). Moment-lovers may note calculated skewness of 10.51 and kurtosis of 127.99. The display is spectacular but useless. Not surprisingly, many ecologists would prefer a histogram based on log counts. We revert to quantile plots.

```
. histogram count, w(1) discrete freq yla(, ang(h)) yti(number of species)
> xti(number of individuals)
```

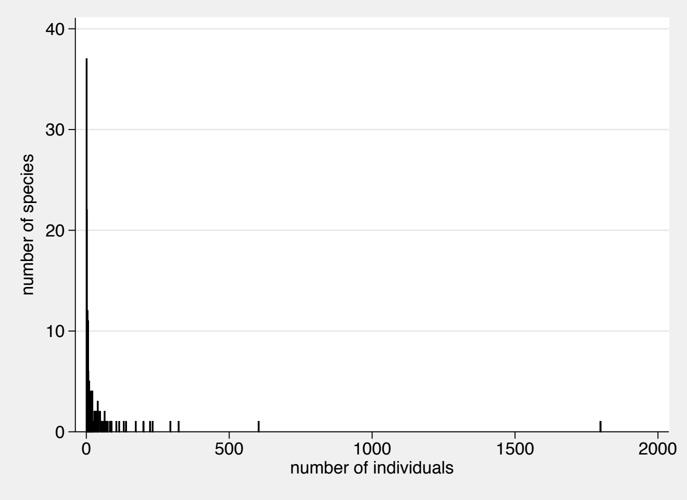

Figure 6: Macrolepidoptera at Rothamsted 1935. A highly skewed distribution for species with varying numbers of individuals.

With such an extreme distribution shape, we might even consider the idea of a power or scaling law. Over the last century or so, many scholars have fallen for the temptations of power laws. Economists may think of Vilfredo Pareto; linguists, sociologists, psychologists, and geographers of George Kingsley Zipf; physicists of Per Bak; and everyone numerate of Benoît Mandelbrot. The literature extends to grandiose works showing how power laws offer keys to understanding the universe, or at least some large segment of it, such as *Human behavior and the principle of least effort* (Zipf 1949) and *How nature*

[3] Letters from Fisher to Williams copied on the Internet at *http://www.library.adelaide.edu.au/digitised/fisher/corres/williamscb/* show his nuanced progression over the years through greater degrees of informality: from “Dear Dr Williams”, to “Dear Williams”, “My dear Williams”, and “Dear C.B.” to “My dear C.B.”. Notice the mutation through a variety of forms, none of which included the first names so thoughtfully bestowed by Williams’ parents.

[4] The third member of the trio, the British scientist Alexander Steven Corbet (1896–1948), started as a chemist and biochemist and ended as an entomologist, publishing the definitive work on the butterflies of the Malay peninsula. In between, he established the production of $N_2O$ by denitrifying soil bacteria and wrote a book on biological processes in tropical soils (Riley 1948; Hutchinson 1978, 231).

*works* (Bak 1997)[5]. For a level-headed discussion of power laws and how far they have been oversold, see Perline (2005).

In the simplest case, power laws have been suggested of the form

$$\text{count} \ \times \ \text{rank} \ = \ \text{constant}$$

or

$$\text{count} \ \propto \ \text{rank}^{-1}$$

although the greater generality of

$$\text{count} \ \propto \ \text{rank}^{-b}$$

has seemed an easy step further[6]. Here "count" is the right word, but other applications extend to measured sizes, such as earthquake magnitudes or avalanche volumes. Note again that this formulation depends on ranking the largest as 1.

The connection to quantile plots is immediate. With a power law of this form, quantiles expressed as a function of rank will plot as a straight line with downward slope $b$ if both axis scales are logarithmic. For a variety of reasons, some mathematical and some empirical, a variety of further adjustments have been made to the form of the power law, but approximate linearity on a log-log plot remains the general expectation.

Figure 7 shows the result for our dataset. The enthusiast will note a substantial straightening; the skeptic will note remaining systematic curvature. (A regression of log frequency on log rank gives a downward slope of 1.59.)

```
. qplot count, reverse rank ysc(log) yti(number of individuals)
> yla(1 3 10 30 100 300 1000, ang(h)) xsc(log) xla(1 3 10 30 100)
```

[5] Names must be part of the explanation. Suppose that I write a treatise on the negative binomial, using a wide variety of examples to show how it is astonishingly applicable in all sorts of problems across many sciences. Then I leave out most of the equations and add seasoning and garnish in the form of entertaining and poignant anecdotes and quirky connections. The project would never go public, as no publishing person would buy the notion that anything with such a bizarre name as "negative binomial" was the key to much worth knowing about. As in Hollywood, a new name would be needed, or else nothing. On the other hand, the names power or scaling laws, by accident or design, suggest the stuff of legend.

[6] My haphazard sampling of a large literature scattered across time and discipline suggests a conjecture: as the exponent in the power law approaches 1, the proponent approaches single-mindedness.

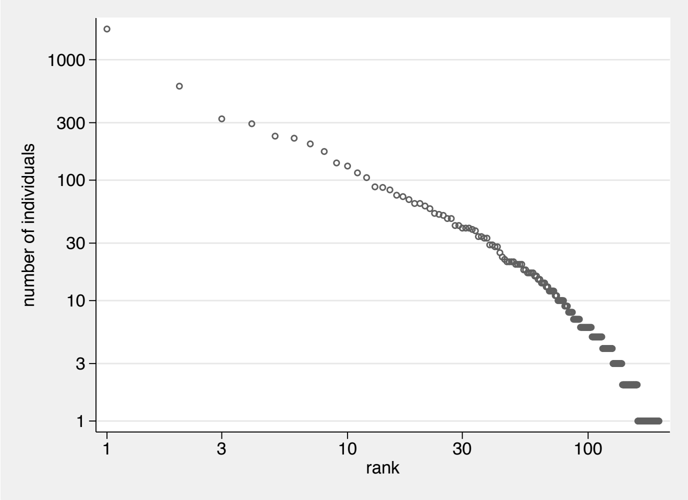

Figure 7: Power law fitted to macrolepidoptera data, Rothamsted 1935. The fit is good by some standards, but systematic curvature is also evident.

Once more, we need to consider alternatives but will go no further than the lognormal. It fits very well, except in the far tail of very rare species, but here the discreteness of the variable really causes problems (figure 8).

```
. qplot count, reverse trscale(invnorm(@)) xti(standard normal scale) ysc(log)
> yti(number of individuals) yla(1 3 10 30 100 300 1000, ang(h))
```

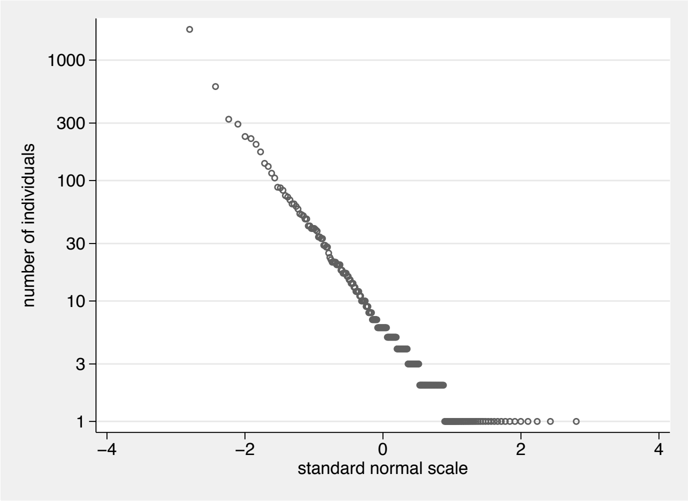

Figure 8: Lognormal distributions fitted to macrolepidoptera data, Rothamsted 1935. The fit is good except for species with counts near 1.

## 6 A Tukeyish aside

The discussion has moved from showing observed distributions, represented as quantile traces, to comparing those with fitted distributions, represented in the same way. Naturally, this is far from the only way to make the comparison. One alternative is a quantile–quantile plot, namely a scatter plot of observed and fitted quantiles.

One disadvantage of quantile plots used in this way is that both observed and fitted traces are necessarily monotone. In a sense, this can make agreement appear better than it really is, and one challenge is therefore to reduce or at least to remove the sloping behavior and make traces more nearly horizontal. This principle has been discussed in a previous column (Cox 2004c).

The particular case of power laws, or rather of very skewed so-called J-shaped distributions, is of interest because other graphical methods have been suggested for this situation, although they appear to have received little attention. Tukey (1977, chapter 18) noted a vague symmetry between two counts, the number of appearances and the number of individuals appearing at least that many times (the rank), and suggested the need for an analysis that treated these counts symmetrically. Given his attitude that square roots are the "first-aid" transformation for counts, Tukey thus proposed a plot of

$$\sqrt{\text{count } \times \text{ rank}} \text{ versus } \log(\text{count}/\text{rank})$$

which he called a product-ratio plot. It is immediate that data for which

$$\text{count} \times \text{ rank} = \text{constant}$$

follow a horizontal line on such a plot. An alternative in the same spirit as Tukey's proposal is a plot of

$$\log(\text{count } \times \text{ rank}) \text{ versus } \log(\text{count}/\text{rank})$$

for which the same property holds. Thus power laws became more nearly horizontal. The idea can be implemented in a few lines:

```
. egen rank = rank(-count), unique
. gen y = log(count) + log(rank)
. gen x = log(count) - log(rank)
. label var y "log(count X rank)"
. label var x "log(count/rank)"
. scatter y x, xsc(reverse)
```

Although no results are shown here, this can be a helpful way of looking in finer detail at the structure of such distributions. How to handle tied counts is a small issue. Tukey (1977) suggested one procedure: the code above is based on another, which at least shows the frequency of ties fairly explicitly.

See also Parunak (1979) for other ideas on the same problem.

## 7 Distribution plots

The plot of cumulative probability versus ordered values we call here a distribution plot. Workers differ on plotting either the cumulative distribution or its complement, especially in the guise of a survival function. Official Stata offers many graphical and other tools for showing and analyzing the survival function for survival time data. It stops short at [R] **cumul**, which just calculates the usual cumulative distribution function, the subsequent graphics being a trivial application of whatever `twoway` plottype is desired. Nevertheless for convenience the `distplot` program has been offered (Cox 1999a), especially to make it easier to show several distributions simultaneously.

Broadly speaking, distribution plots and quantile plots have different bases. The first is based on rank/sample size as the estimate of cumulative probability. For some constant $a$, the second is based on so-called plotting positions defined by (rank $-$ $a$)/(sample size $-$ $2a + 1$): note that this definition does not include rank/sample size as a special case. The difference should not be dismissed as a mere nuance. With the latter convention, estimated cumulative probabilities will not usually be zero or one, and thus they will yield determinate results with some transformations, notably the logarithm and the logit. (The user is at liberty to insist that $a = 1$, however.) `distplot` has a `midpoint` option designed to reduce the number of zeros and ones, so that transformations as far as possible still apply, but the difference remains.

The difference between distribution plots and quantile plots is otherwise essentially one of choice of axes, which may seem to boil down to a question of habit or taste. However, Mandelbrot (1997, 207) makes a simple but valuable point: plots are not neutral, so different presentations of a dataset emphasize different aspects. Hence, for example, when the values that matter most are the largest, they are seen best in rank-size plots.

## 8 Software updates

This is a more formal statement flagging the release of new releases of `qplot` (`gr42_3`) and `distplot` (`gr41_3`) with this column. Previous versions of these programs appeared in Cox (1999b, 2001, 2004a) (`qplot`) and Cox (1999a, 2003a,b) (`distplot`). The main changes in this release are

1. Simplification of the main syntax. Previously, command calls would specify plottype as a subcommand, as in `qplot scatter` or `distplot line`. Now the plottype is by default as just exemplified and is otherwise changeable by use of `recast()`. See [G] ***advanced_options***.

2. Previously the `by()` option specified that two or more groups be summarized by quantile curves within an individual graph panel. This was out of step with current Stata practice, so the `by()` option has been renamed `over()`. A true `by()` option, specifying use of different panels, has now also been added.

## 9 Conclusions

Ecological examples have been used to underline the flexibility of quantile plots, going beyond the possibilities of `quantile` and the defaults of `qplot`. Heavy use was made of the options of `qplot` that produce reversed, ranked, or transformed scales. It can be useful to take logarithms of quantiles, logarithms of ranks, or both. It is possible to carry out transformations of the probability or rank axis on the fly. Juxtaposed and superimposed comparisons are both straightforward. Quantile plots are indeed protean.

## 10 Acknowledgment

I thank Sarah A. Corbet for information on her father, A. Steven Corbet.

## 11 References

Bak, P. 1997. *How Nature Works: The Science of Self-Organized Criticality*. Oxford: Oxford University Press.

Batten, L. A. 1976. Bird communities of some Killarney woodlands. *Proceedings, Royal Irish Academy, Series B* 76: 285–313.

Bulmer, M. G. 1974. On fitting the Poisson lognormal distribution of species abundance data. *Biometrics* 30: 101–110.

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

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

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

Cohen, J. E. 1966. *A Model of Simple Competition*. Cambridge, MA: Harvard University Press.

Cox, N. J. 1999a. gr41: Distribution function plots. *Stata Technical Bulletin* 51: 12–16. In *Stata Technical Bulletin Reprints*, vol. 9, 108–112. College Station, TX: Stata Press.

—. 1999b. gr42: Quantile plots, generalized. *Stata Technical Bulletin* 51: 16–18. In *Stata Technical Bulletin Reprints*, vol. 9, 113–116. College Station, TX: Stata Press.

—. 2001. gr42.1: Quantile plots, generalized: update to Stata 7.0. *Stata Technical Bulletin* 61: 10–11. In *Stata Technical Bulletin Reprints*, vol. 10, 55–56. College Station, TX: Stata Press.

—. 2002. Speaking Stata: How to move step by: step. *Stata Journal* 2(1): 86–102.

—. 2003a. Software update: gr41_1: Distribution function plots. *Stata Journal* 3(2): 211.

—. 2003b. Software update: gr41_2: Distribution function plots. *Stata Journal* 3(4): 449.

—. 2004a. Software update: gr42_2: Quantile plots, generalized. *Stata Journal* 4(1): 97.

—. 2004b. Speaking Stata: Graphing distributions. *Stata Journal* 4(1): 66–88.

—. 2004c. Speaking Stata: Graphing agreement and disagreement. *Stata Journal* 4(3): 329–349.

—. 2005. Speaking Stata: Density probability plots. *Stata Journal* 5(2): 259–273.

David, H. A. and H. N. Nagaraja. 2003. *Order Statistics.* Hoboken, NJ: Wiley.

Domb, C. 1990. On Hammersley's method for one-dimensional covering problems. In *Disorder in Physical Systems*, ed. G. R. Grimmett and D. J. A. Welsh, 33–53. Oxford: Oxford University Press.

Fisher, R. A., A. S. Corbet, and C. B. Williams. 1943. The relation between the number of species and the number of individuals in a random sample of an animal population. *Journal of Animal Ecology* 12: 42–58.

Gordon, N. D., T. A. McMahon, B. L. Finlayson, C. J. Gippel, and R. J. Nathan. 2004. *Stream Hydrology: An Introduction for Ecologists.* Chichester, UK: Wiley.

Hutchinson, G. E. 1978. *An Introduction to Population Ecology.* New Haven, CT: Yale University Press.

Irwin, J. O. 1967. William Allen Whitworth and a hundred years of probability. *Journal of the Royal Statistical Society, Series A* 130: 147–176.

Kleiber, C. and S. Kotz. 2003. *Statistical Size Distributions in Economics and Actuarial Sciences.* Hoboken, NJ: Wiley.

Krebs, C. J. 1989. *Ecological Methodology.* New York: Harper and Row.

Lande, R., S. Engen, and B.-E. Sæther. 2003. *Stochastic Population Dynamics in Ecology and Conservation.* Oxford: Oxford University Press.

Magurran, A. E. 2004. *Measuring Biological Diversity.* Malden, MA: Blackwell.

Mandelbrot, B. B. 1997. *Fractals and Scaling in Finance: Discontinuity, Concentration, Risk.* New York: Springer.

May, R. M. 1975. Patterns of species abundance and diversity. In *Ecology and Evolution of Communities*, ed. M. L. Cody and J. M. Diamond, 81–120. Cambridge, MA: Harvard University Press.

Mosteller, F. 1965. *Fifty Challenging Problems in Probability with Solutions.* Reading, MA: Addison–Wesley.

Parunak, A. 1979. Graphical analysis of ranked counts (of words). *Journal of the American Statistical Association* 74: 25–30.

Perline, R. 2005. Strong, weak, and false inverse power laws. *Statistical Science* 20: 68–88.

Riley, N. D. 1948. Dr. A. S. Corbet. *Nature* 161: 1003.

Simon, H. A. 1991. *Models of My Life*. New York: Basic Books.

Tukey, J. W. 1977. *Exploratory Data Analysis*. Reading, MA: Addison–Wesley.

Whitworth, W. A. 1905. *Choice and Chance with One Thousand Exercises*. Facsimile edition, New York: Hafner, 1951. Previous editions 1867, 1870, 1878, 1886.

Williams, C. B. 1964. *Patterns in the Balance of Nature and Related Problems in Quantitative Ecology*. London: Academic Press.

Zipf, G. K. 1949. *Human Behavior and the Principle of Least Effort*. Cambridge, MA: Addison–Wesley.

### 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 co-authored fifteen commands in official Stata. He was an author of several inserts in the *Stata Technical Bulletin* and is an Editor of the *Stata Journal*.