Stata users meeting, London, May 2001

Plotting graded data: a Tukey-ish approach

Nicholas J. Cox

Department of Geography, University of Durham,
Durham City, DH1 3LE, UK
n.j.cox@durham.ac.uk

Introduction

This presentation is about `ordplot`, a Stata program which plots cumulative distribution functions. As other such Stata programs exist, including `cdf` (Clayton and Hills, STB-49, 1999) and `distplot` (Cox, STB-51, 1999), this may appear to be only a little deal. However, `ordplot` was developed specifically to incorporate some features which are useful for ordinal data, and it is described here from that point of view. Such features in no way inhibit its use for interval and ratio variables, and users may thus find it helpful beyond its original intended field of application. `ordplot` may be installed now from SSC-IDEAS.

The trigger for this project was the passing of John W. Tukey (1915-2000), who made many deep contributions to statistics. In the latter part of his career he gave an enormous stimulus to exploratory data analysis, including graphics. Broadly speaking, it seems that there are fewer simple and useful exploratory tools for categorical data analysis than in other branches of statistics. I interact with many relatively non-statistical students and researchers who work mainly with questionnaire data which are primarily categorical, and I have been frustrated at the apparent lack of useful graphics for categorical data. Tables of frequencies can be represented by (e.g.) bar charts, but even two-way tables do not always reveal their structure easily and fully through such displays. Re-reading some of Tukey's works uncovered some suggestions, both ideas he was using 40 years ago (Tukey 1960, 1961) and a neat device from his well-known *Exploratory data analysis* (Tukey 1977). The combination of these ideas seems to deserve wider use.

The psychologist S.S. Stevens classified variables into those on nominal, ordinal, interval and ratio scales. Associated with this classification, at least in some applied fields, are a series of debates (and various prejudices, taboos, etc.) about what techniques can or should be applied to variables recorded on different scales. In fact, such debate goes back to Karl Pearson and G.U. Yule at least: how far can categorical variables be taken as approximations to hidden or latent variables on interval or ratio scales?

Despite the near universal adoption of *ordinal* as a general term, it covers a wide range. Once more following Tukey (1977), ordinal data may be split into grades and ranks. Graded data are those possessing an inherent order but falling short of a metric scale, and even falling short of ranking, in which we have a directly given count of how many are higher, and so on. Such grades are common in many fields, especially as the record of some considered judgement. Examples are opinions on a five-point scale such as *strongly disagree*, *disagree*, *neutral*, *agree*, *strongly agree*. Little attention seems to have been given to methods for the easy and effective graphical display of graded data. Indeed, the assumption is often made that such data are fit for tabulation, but not for graphical display.

Cumulative probabilities with `ordplot`

The very simple basis for `ordplot` is the principle that cumulative probabilities are a logical and practical way to represent graded data: this principle is after all the basis for many models for graded responses. In fact, experience indicates that cumulative probabilities may be useful even when you find it difficult to believe that the graded scale should be treated as an approximation to some underlying metric variable. Cumulative probability is shown on the y axis and the variable itself on the x axis.

More specifically, cumulative probability curves for different subsets of the data are often useful in initial description and exploration of the data. In most cases, cumulative distributions will be shown for each category of a classifying variable, specified by the `by()` option. Note that this classifying variable must have no more than 10 categories.

In most cases the graded variable will be regarded as a response variable, but this is not essential for `ordplot` to be useful.

`ordplot` incorporates several special features, most of which remain pertinent to variables on interval and ratio scales:

◇ *Split fractions or ridits*

Given proportions or probabilities p_i at grades indexed $i = 1, \ldots, k$, then cumulative probabilities could be defined most obviously

$$\text{either as} \qquad P_i = \sum_{j<i} p_j \qquad\qquad \text{or as} \qquad P_i = \sum_{j \leq i} p_j.$$

The choice might seem immaterial, but note that with the first convention $P_1 \equiv 0$ and with the second convention $P_k \equiv 1$. Such identities tell us nothing about the data and are awkward when plotting probabilities on some transformed scales (e.g. logits of 0 and 1 are indeterminate). A surprisingly helpful wrinkle, which is the default in `ordplot`, is to work with

$$P_i = \sum_{j<i} p_j + \frac{p_i}{2},$$

which with terminology from Tukey (1977, pp.496-497) could be called a *split fraction below*. It is also a *ridit* as defined by Bross (1958): see also Fleiss (1981, pp.150-7) or Flora (1988). It is psychologically helpful to show k data points for k grades so far as is possible: with say 5 grades, even 1 missing data point is conspicuous by its absence.

If desired, the more conventional alternatives are available through the `lt` and `le` options, respectively.

◇ *Reverse curves*

A plot of the complement of this cumulative probability, $1 - P$, a.k.a. the reliability, survival or survivor function, may be obtained through the `reverse` option. The pertinent alternatives are available through the `ge` or `gt` options.

◇ *Transforms of probability scales*

The `scale()` option of `ordplot` indicates a scale for plotting cumulative distributions. `logit` (synonym `flog` for *folded logarithm*), `froot`, `folded`, `loglog`, `cloglog`, `normal` or `Gaussian`, `percent` and `raw` are allowed. `raw` is the default.

Given cumulative probabilities P, and using log to denote natural logarithm (base e),

`logit` or `flog` means $\log(P/(1-P)) = \log P - \log(1-P)$.

`froot` for *folded root* means $\sqrt{P} - \sqrt{1-P}$.

`folded` means *folded power* or $P^{\text{power}} - (1-P)^{\text{power}}$. The power to be used must be specified through the `power( )` option and should be non-zero. For reference, note that, apart from scaling constants, good emulations of the angular (arcsine square root) transformation and of the probit transformation are obtained by powers of 0.41 and 0.14 respectively. As power approaches 0, the folded power tends to the logit.

`loglog` means $-\log(-\log P)$.

`cloglog` means $\log(-\log(1-P))$.

`normal` or `Gaussian` means `invnorm(`P`)`. See help in Stata on `functions`.

`percent` means $100P$. (In my experience, many users much prefer a percent scale to a probability scale.)

Under `reverse`, P is replaced by $1-P$, and *vice versa*, in these operations.

Further information on working with counted fractions and folded transformations for probability scales is available in Tukey (1960, 1961, 1977), Atkinson (1985), Cox and Snell (1989) and Emerson (1991). Some of the transformations used here appear as link functions in the literature on generalized linear models (e.g. McCullagh and Nelder 1989; Aitkin *et al.* 1989).

◇ *Labels on original scales*

Cumulative probabilities shown on a transformed scale can be difficult to think about. Even experienced statisticians might well prefer to see a label of 0.95 or 95% rather than of logit(0.95), which is 2.944.

The options `plabel(`*numlist*`)`, `pline(`*numlist*`)` and `ptick(`*numlist*`)` are for use if the `scale` is `logit`, `flog`, `froot`, `folded` with `power`, `loglog`, `cloglog`, `normal` or `Gaussian`. They specify labels, lines or ticks on the y axis on a probability or percent scale. If the largest number in one or more of these *numlists* is > 1, numbers are treated as percents. Otherwise, numbers are treated as probabilities. Numbers which are not plottable on the chosen scale, such as logit of 0 or 1, are ignored.

For `scale` of `raw` or `percent`, use `ylabel()`, `yline()` or `ytick()` instead.

◇ *Changing scores on the fly*

Grades will usually be represented numerically by integer scores, such as 1(1)5. Occasionally, it can be interesting to experiment with different scores. For example, in some surveys *strongly agree* and *strongly disagree* appear to be quite extreme compared with *tend to agree*, *neutral* and *tend to disagree*. (Politically, the extreme right-wing and extreme left-wing may make all the different flavours of liberals and moderates appear much of a muchness.)

The option `asscores()` specifies an ascending *numlist* to use as alternative scores in plotting values on the x axis. The elements of the *numlist* must match one-to-one with the distinct values of the graded variable occurring in the observations used and put into ascending order.

For example, given values 1 2 3 4 5, `asscores(1 4 5 6 9)` will map 1 to 1, 2 to 4, etc. Any value labels will be mapped alongside, unless the target score is not an integer.

There is some danger in being able to do this: users should preferably have, or be able to concoct, some substantive story which makes an alternative scoring seem plausible, if not natural.

As a matter of implementation, `ordplot` is a wrapper for `keyplot`, also available from SSC-IDEAS. The main feature of `keyplot` is that keys appear for up to 10 curves plotted, rather than up to 4 as would be the case with `graph, twoway` at present.

References

Aitkin, M., Anderson, D., Francis, B. and Hinde, J. 1989. *Statistical modelling in GLIM.* Oxford: Oxford University Press.

Atkinson, A.C. 1985. *Plots, transformations, and regression.* Oxford: Oxford University Press.

Bross, I.D.J. 1958. How to use ridit analysis. *Biometrics* 14, 38-58.

Cox, D.R. and Snell, E.J. 1989. *Analysis of binary data.* London: Chapman and Hall.

Emerson, J.D. 1991. Introduction to transformation. In Hoaglin, D.C., Mosteller, F. and Tukey, J.W. (eds) *Fundamentals of exploratory analysis of variance.* New York: John Wiley, 365-400.

Fleiss, J.L. 1981. *Statistical methods for rates and proportions.* New York: John Wiley.

Flora, J.D. 1988. Ridit analysis. In Kotz, S. and Johnson, N.L. (eds) *Encyclopedia of statistical sciences.* Wiley, New York, 8, 136-139.

McCullagh, P. and Nelder, J.A. 1989. *Generalized linear models.* London: Chapman and Hall.

Tukey, J.W. 1960. The practical relationship between the common transformations of percentages or fractions and of amounts. Reprinted in Mallows, C.L. (ed.) 1990. *The collected works of John W. Tukey. Volume VI: More mathematical.* Pacific Grove, CA: Wadsworth & Brooks-Cole, 211-219.

Tukey, J.W. 1961. Data analysis and behavioral science or learning to bear the quantitative man's burden by shunning badmandments. Reprinted in Jones, L.V. (ed.) 1986. *The collected works of John W. Tukey. Volume III: Philosophy and principles of data analysis: 1949-1964.* Monterey, CA: Wadsworth & Brooks-Cole, 187-389.

Tukey, J.W. 1977. *Exploratory data analysis.* Reading, MA: Addison-Wesley.

Acknowledgments

Elizabeth Allred and Ronan Conroy made very helpful comments. The implementation of `plabel`, etc., is based on an idea of Patrick Royston.