NASUG, Boston, MA, March 2001

Analysing circular data in Stata

Nicholas J. Cox

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

Introduction

Circular data are a large class of directional data, which are of interest to scientists in many fields, including biologists (movements of migrating animals), meteorologists (winds), geologists (directions of joints and faults) and geomorphologists (landforms, oriented stones). Such examples are all recordable as compass bearings relative to North. Other examples include phenomena that are periodic in time, including daily and seasonal rhythms. The analysis of circular data is an odd corner of statistical science which many never visit, even though it has a long and curious history. Moreover, it seems that no major statistical language provides direct support for circular statistics. There is a commercially available special-purpose program called Oriana (see `http://www.kovcomp.co.uk`). This paper describes the development and use of some routines which have been written in Stata, primarily to allow graphical and exploratory analyses. Collectively they offer about as many facilities as does Oriana.

The elementary but also fundamental property of circular data is that the beginning and end of the scale coincide: for example, $0^\circ = 360^\circ$. An immediate implication is that the classic arithmetic mean is likely to be a poor summary: the mean of 1° and 359° cannot be sensibly be 180°. The solution is to use the vector mean direction as circular mean. If θ is direction and there are n observations, each with unit weight, then form the sums

$$S = \sum \sin \theta, \qquad C = \sum \cos \theta.$$

The vector mean direction is

$$\bar{\theta} = \arctan(S/C)$$

and the strength of the resultant vector (a.k.a. mean resultant length) is

$$\bar{R} = \sqrt{S^2 + C^2}/n.$$

$\bar{R}$ varies between 0 and 1 and is an inverse analogue of the variance: however, $\bar{R}$ near 0 can arise in very different ways, as with a circular uniform distribution or with clusters of values 180° apart.

Sometimes data come as axes, undirected lines: one end of a joint in rock cannot be distinguished from the other. The convention with such axial data is to double them, reduce them modulo 360°, analyse these data and finally back-transform them.

Existing programs

The programs written rest, so far, on the assumption that data are recorded in degrees from North. Users working with other scales (e.g. time of day on a 24 hour clock, day or month of year) could write their own trivial preprocessor. In due course it is intended to implement, possibly through characteristics modified by some `circset` command, user setting of different scales. Stata internally expects angles to be in radians (π radians = 180°), but I have not seen radians used for reporting data. In Stata, the factors `_pi/180` and `180/_pi` are thus useful for conversion between angles and radians.

Programs fall into four fairly distinct classes:

1. **Utilities**

`circcent` rotates a set of directions to a new centre: the result is on either $[-180°, 180°]$ or $[0°, 360°]$.

`circdiff` measures difference between values as the shorter arc around the circle.

Also needed is arctangent code. Stata's `atan()` function takes a single argument and has range $-\pi/2$ to $\pi/2$ radians, whereas circular statistics needs an arctangent function which takes two arguments and returns an angle on the whole circle between 0 and 2π radians.

2. **Summary statistics and significance tests**

`circsumm` is a basic workhorse that calculates vector mean and strength and the circular range and offers, as options, approximate confidence intervals for the vector mean and Rayleigh and Kuiper tests of uniform distribution on the circle.

The circular range is the length of the shortest arc which includes all observations.

The Rayleigh test is a test of a null hypothesis of uniformity against an alternative hypothesis of unimodality. The Kuiper test is a test of a null hypothesis of uniformity against any alternative.

`circmed` calculates the circular median and mean deviation from the median. Define the circular distance $d(\theta, \phi)$ as the length of the shorter arc joining θ and ϕ, whether clockwise or anticlockwise: here length is always taken as positive or zero. Then the median is that $\tilde{\theta}$ which minimises the mean deviation

$$\frac{1}{n} \sum d(\theta, \tilde{\theta}).$$

(More precisely, it is the vector mean of any such minimising values.) In practice, the circular median is not as useful as the vector mean, partly because on the circle outliers have less space in which to hide: an outlier can be at most 180° from the next value.

`circ2sam` and `circwwm` offer nonparametric tests for comparing two or more subsets of directions. `circ2sam` offers two test statistics based on empirical distribution functions to test whether two distributions are identical, namely Watson's U^2 and Kuiper's k^*. `circwwm` carries out a homogeneity test due to Wheeler and Watson and to Mardia given subdivision into $r \geq 2$ groups. The test statistic is based on the circular ranks of the data, $2\pi \, \text{rank}/n$, and can be compared with χ^2 with $2r - 2$ degrees of freedom, so long as there are 10 or more

values in each group. Randomisation is recommended otherwise to get an estimate of the P-value.

3. **Univariate graphics**

`circrplt`, `circdplt` and `circvplt` are all written making heavy use of the `gph` features introduced in Stata 5.0. `circrplt` loosely resembles `spikeplt` or `spikeplot` (Stata 6.0 on); `circdplt` loosely resembles `dotplot` (Stata 5.0 on). `circvplt` shows the ordered directions added end to end with the vector mean as resultant.

Many users like such intrinsically circular representations, but another approach is to wrap around the scale, showing up to 2 full cycles on a linear graph:

◇ `circhist` is a wrapper for `graph, histogram`, adding a pad of values with up to 180° range to both extremes. Optionally, `histplot` (SSC-IDEAS) may be called instead, as is essential if the number of bins required exceeds 50, or if unshaded bins are desired.

◇ `circnpde` drives a nonparametric density estimation routine with biweight kernel. (This is self-contained and does not call `kdensity`.)

Note that a quantile plot of directions can be useful: `quantile` – or alternatively, `quantil2` (STB-51, STB-61) – is already available for this purpose. (A circular uniform distribution is often useful as a reference distribution.)

4. **Bivariate relationships**

`circplot` is a wrapper for `graph, twoway` that adds a pad of up to 180° range to both extremes on either or both of x and y axes.

For exploratory smoothing, `circysm` is for circular response and non-circular covariate and `circxsm` is for non-circular response and circular covariate. Both are wrappers for `ksm` and allow in particular lowess (loess) smoothing. With `circysm`, the recipe is to smooth sine and cosine components and to recombine using arctangent:

$$\text{smooth of } \theta = \arctan(\text{smooth of } \sin\theta, \text{smooth of } \cos\theta).$$

With `circxsm`, the recipe is to smooth around the circle by temporarily adding sufficiently large pads at each end.

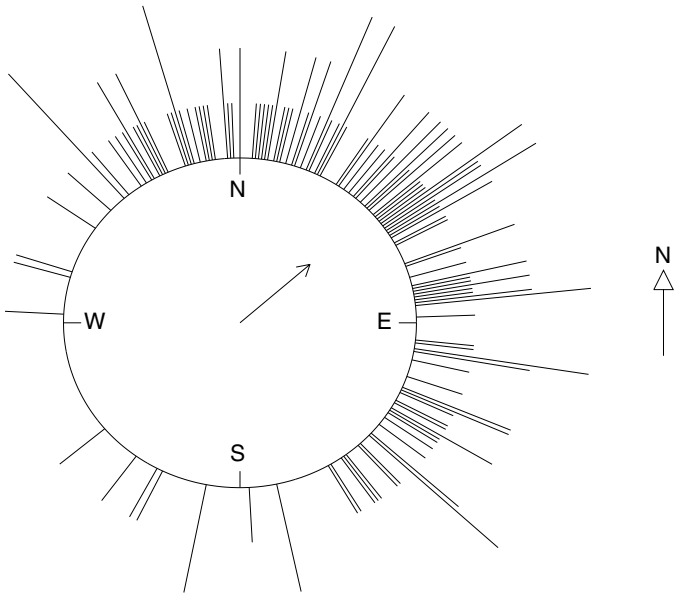
`circcorr` and `circlcco` implement correlation methods for cases where one or both variables are circular.

Regression of a non-circular response on various terms of a Fourier series requires nothing extra in Stata. A utility `fourier` has been written to allow easy generation of as many terms as are required.

With even a modest number of routines supporting a shared need a collective help file, in this case called `circstat`, is useful. Hypertext links to other help files make life easier, even for the program developer, let alone new users.

What next?

Other areas of present interest include circular data in time and space. Some work in circular statistics makes use of pertinent probability distributions such as the von Mises.



Lake District cirques axis aspects
mean direction 48.3 : vector strength 0.532