Stata Users' Meeting London June 2004

Circular statistics in Stata, revisited

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 those dependent on time of day (in biomedical statistics: hospital visits or times of birth) or time of year (in applied economics: unemployment or sales variations). 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`), while a user-written package called CircStats is available for S-Plus and R. This talk describes the development and use of some routines that have been written in Stata, primarily to allow graphical and exploratory analyses. In 2004, such routines are being rewritten, especially to allow use of the new graphics in Stata 8. As they become presentable, they are being placed in the `circular` package on SSC.

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 arithmetic mean is likely to be a poor summary: the mean of 1° and 359° cannot sensibly be 180°. The solution is to use the vector mean direction as circular mean. If θ is direction and there are n observations, form the sums $S = \sum_{i=1}^{n} \sin \theta_i, C = \sum_{i=1}^{n} \cos \theta_i$. Then 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. 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 and fix cosmetic details such as graph axis labels. In due course I may implement, possibly through characteristics modified by some `circset` command, user setting of different scales. Stata expects angles to be in radians (π radians $= 180°$), but I have never seen radians used for reporting data. In Stata, the factors `_pi/180` and `180/_pi` are thus useful for conversion between angles and radians.

1. **Utilities**

`circcentre` rotates a set of directions to a new centre: the result is on $[-180°, 180°]$.

`circdiff` measures difference between circular variables or constants 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. An `egen` function `atan2()` fills the gap.

`fourier` generates pairs of sine and cosine variables $\sin j\theta, \cos j\theta$ for $j = 1, \cdots, k$.

2. **Summary statistics and significance tests**

`circsummarize` 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 abbreviation `circsu` is also allowed.)

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

A circular uniform distribution is often useful as a reference distribution, at least as a zeroth approximation. The Rayleigh test is a test of a null hypothesis of uniformity against an alternative hypothesis of unimodality. It is based on the vector strength. The Kuiper test has a Kolmogorov-Smirnov flavour and is a test of a null hypothesis of uniformity against any alternative.

`circrao` carries out a uniformity test. Sort the n values and calculate spacings as differences between successive ordered values: the last spacing is calculated from the last value to the first. Then calculate $U = \frac{1}{2} \sum |\text{spacing} - 360/n|$, which has the following interpretation: Place n arcs of fixed length $360/n$ degrees on the circumference, starting with each of the sample points. The circumference would be completely covered by these arcs only if the sample points were uniformly (equally) spaced. U is the total uncovered portion of the circumference, or equivalently the extent to which the arcs overlap each other. Large values of U indicate clustering of the sample points or evidence for rejecting the null hypothesis of uniformity. One merit of this test is that it works well for data which are not unimodal.

`circmedian` 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.

`circovm` and `circovstr` show the effect of omitting individual values on the vector mean and vector strength.

`circtwosample` and `circwwmardia` offer nonparametric tests for comparing two or more subsets of directions. `circtwosample` 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^*. `circwwmardia` 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 and bivariate graphics**

`circrplot` loosely resembles `spikeplot`; `circdplot` loosely resembles `dotplot`. `circvplot` shows the ordered directions added end to end with the vector mean as resultant. Many users like such intrinsically circular representations, but note that it may be necessary to use `graph display`, typically with equal `xsize()` and `ysize()`, to fix the aspect ratio.

Another approach is to wrap around the scale, showing up to two full cycles on a linear graph. `circhistogram` is a wrapper for `histogram`, adding a pad of values (default 180°) to both extremes. `circscatter` is a wrapper for `scatter` that adds a pad to both extremes on either or both of x and y axes. (The abbreviations `circhist` and `circsc` are also allowed.)

Note that a quantile plot of directions can be useful: `quantile` – or alternatively, `qplot` (*Stata Journal* 4(1): 97, 2004) – is already available for this purpose. `circqqplot` is a circular version of `qqplot`.

4. **Smoothing, relationships and modelling**

`circkdensity` drives a nonparametric density estimation routine with biweight kernel. Despite the name, it does not call `kdensity`.

For exploratory smoothing, `circylowess` is for circular response and non-circular covariate and `circxlowess` is for non-circular response and circular covariate. Both are wrappers for `lowess`. With `circylowess`, 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 `circxlowess`, the recipe is to smooth around the circle by temporarily adding sufficiently large pads at each end.

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

Note that regression of a non-circular response on various terms of a Fourier series requires nothing extra in Stata (although `fourier` can help). It is often extremely useful, and can be extended to include non-circular covariates.

5. **von Mises distributions**

`circvm` fits a von Mises distribution, the most important unimodal reference distribution on the circle, using an approximate maximum likelihood method. (Doing it properly with `ml` is on the agenda.) First introduce $I_0(z)$, known as the modified Bessel function of the first kind and order 0, which is defined by

$$\sum_{i=0}^{\infty} (z/2)^{2i} \ / \ i! \ i! \qquad \text{or} \qquad \frac{1}{2\pi} \int_0^{2\pi} \exp(z \cos \phi) \, d\phi,$$

but in practice best calculated using polynomial approximations. The von Mises distribution for θ with mean μ is then defined to have probability density function

$$f(\theta) \ = \ \frac{1}{2\pi I_0(\kappa)} \ \exp[\kappa \cos(\theta - \mu)],$$

and so probability distribution function

$$F(\theta) \ = \ \frac{1}{2\pi I_0(\kappa)} \ \int_0^{\theta} \exp[\kappa \cos(\phi - \mu)] \, d\phi.$$

Here $\kappa \geq 0$ is a shape parameter measuring precision or concentration. As κ tends to zero, the von Mises distribution approaches the uniform, while as κ becomes arbitrarily large, it increasingly resembles a normal distribution: hence the name 'circular normal' sometimes used. For positive κ, the distribution is symmetric and unimodal with a single peak at μ. Thus $I_0(\kappa)$, despite its splendid origin, is merely a scaling constant.

The density takes on a maximum proportional to e^{κ} when $\theta = \mu$ and thus $\cos(\theta - \mu)$ is 1, and correspondingly a minimum proportional to $e^{-\kappa}$ opposite μ when $\theta - \mu = \pi$ and $\cos(\theta - \mu) = -1$. The ratio of the densities, a measure of the peakedness of the distribution, is precisely $e^{2\kappa}$, which increases with κ.

Generalised linear model enthusiasts will note membership of the family of exponential distributions.

`circqvm` shows a quantile-quantile plot for data versus a fitted von Mises distribution. Data are rotated so that the mean is at the centre of the plot. `circpvm` gives corresponding P-P plots and `circdpvm` gives density probability plots.

`egen` functions `invvm()`, `vm()`, `vmden()` and `rndvm()` and a calculator function `i0kappa` are supporting utilities, sometimes used directly.

Biographical notes

Friedrich Wilhelm Bessel (1784–1846) German mathematician and astronomer
Jean Baptiste Joseph Fourier (1768–1830) French pure and applied mathematician
Nicolaas Hendrik Kuiper (1920–1994) Dutch mathematician
Kanti Vardichand Mardia (1935–) Indian statistician, at Leeds since 1973
J.S. Rao (= Sreenivasa Rao Jammalamadaka) (1944–) Indian statistician, at Santa Barbara since 1976
John William Strutt, third Baron Rayleigh (1842–1919) British physicist and mathematician; Nobel Prize in Physics 1904 for discovery of argon
Richard von Mises (1883–1953) Austrian applied mathematician and philosopher, latterly at Harvard
Geoffrey Stuart Watson (1921–1998) Australian statistician, at Princeton from 1970