Circular Statistics II - Asymptotics

We expect to have a central limit theorem on the circle.

Theorem. Let $\theta_1, \ldots, \theta_n$ be independent and identically distributed random variables on the circle. If the distribution is not lattice, the variable $\theta_1 + \ldots + \theta_n$ converges in distribution to a uniform random variable as $n \to \infty$.

Proof. Consider the characteristic function. Since the r.v. is not lattice, $|\varphi_p| < 1$ for all $p \neq 0$. Thus the char. function of $\sum_{i=1}^{n} \theta_i$, φ_p^n, goes to 0 for $p \neq 0$, and has $\varphi_0^n = \varphi_0 = 1$. But this is the c.f. of a uniform distribution.

Poincare's Theorem.

Let x be a random variable on $\mathbb{R}$ with pdf given by f. Define $X_\omega = x \pmod{2\pi}$ to be a corresponding distribution on S^1.

Let $x' = cx$. The distribution of x'_ω converges to uniform as $c \to \infty$.

Proof. We know for x_ω that

$$\alpha_p = E(\cos p\theta) = \int_0^{2\pi} \cos p\theta \, f_\omega(\theta)\, d\theta$$

$$= \int_{-\infty}^{\infty} \cos p\theta \, f_\omega(x)\, dx$$

change vars $u = \frac{1}{c}x$

$$= \int_{-\infty}^{\infty} \cos cp\theta \, f(cu)\, du$$

we can write

$$\varphi_p = \varphi(p)$$

where the Fourier transform

$$\varphi(z) = \int_0^{2\pi} e^{i\frac{z}{\epsilon}\theta} f_w(\theta)\, d\theta$$

$$= \int_{-\infty}^{\infty} e^{izx} f(x)\, dx$$

and $f(x)$ is the pdf for x. The pdf for the scaled variable is given by

$f_\epsilon^c(cx) = \frac{1}{c} f(x)$, so

when we wrap the scaled var,

$$\varphi^c(z) = \int_0^{2\pi} e^{iz\theta} f_w^c(\theta)\, d\theta$$

$$= \int_{-\infty}^{\infty} e^{izx} f^c(x)\, dx$$

Now if $x = cu$, then $dx = c\,du$

and we can change variables

$$\varphi^c(z) = \int_{-\infty}^{\infty} e^{izx} f^c(x)\,dx$$

$$= \int_{-\infty}^{\infty} e^{izcu} f^c(cu)\,c\,du$$

$$= \int_{-\infty}^{\infty} e^{i(zc)u} f(u)\,\tfrac{1}{c}\cdot c\,du$$

$$= \int_{-\infty}^{\infty} e^{i(zc)u} f(u)\,du.$$

$$= \varphi(cz).$$

Now since $f(x)$ ~~is~~ ~~comp~~ is L^2 (with total mass 1), we know its Fourier transform $\varphi(z) \to 0$ as $|z| \to \infty$, so $\varphi^c(p) \to 0$ for $p \neq 0$, $\varphi^c(0) = 1$ for all c. But this shows that $f_\omega^c \to$ uniform. □

Application. Benford's Law.

Suppose we have a collection of data written in decimal notation. Benford's law states

$$\Pr(\text{first digit of } x \text{ is } i) = \log_{10}(i+1) - \log_{10}(i).$$

We observe that the first digit of x is i (for $i \in 1,\dots,9$), if

$$i \times 10^r \le x < (i+1) \times 10^r$$

for some integer r.

This means

$$r + \log_{10} i \le \log_{10} x \le r + \log_{10}(i+1).$$

or that $\log_{10} x \pmod 1$ is between $\log_{10}(i)$ and $\log_{10}(i+1)$.

Now if the distribution of $\log_{10} x$ is

very spread out, the distribution of the wrapped variable $\log_{10} x \pmod 1$ is close to uniform & by Poincare's theorem, and so the 1st digits ~ the given distribution.

(An actual proof was given by Hill in 1995).

We now turn to uniformity testing. Suppose we have a large sample of points on the circle.

Theorem. The joint distribution of $\bar{C} = \frac{1}{n}\sum \cos(\theta_i)$, $\bar{S} = \frac{1}{n}\sum \sin(\theta_i)$ is asymptotically normal. ~~Thus the distribution of $(\bar{R}^2 = \bar{C}^2 + \bar{S}^2$~~ in the rescaled sense that $\sqrt{2n}(\bar{C}, \bar{S}) \sim$ bivariate normal.

Thus

$$2n\bar{R}^2 = 2n(\bar{C}^2 + \bar{S}^2)$$

$$= \left((\sqrt{2n}\,\bar{C})^2 + (\sqrt{2n}\,\bar{S})^2\right)$$

must be ~~χ^2_2~~ asymptotically χ^2.

We can get an even better estimate from (Jupp, 1999)

$$S^* = \left(1 - \frac{1}{2n}\right) 2n\,\bar{R}^2 + \frac{n\bar{R}^4}{2}$$

or (Wilkie, 1983).

$$Pr(n\bar{R}^2 > K) \approx e^{\sqrt{1+4n+4(n^2-nK)} - (1+2n)}$$

A second test is given by Watson's U^2 test, which is based on the mean square deviation between an empirical cdf for the data an the cdf of the uniform distribution.

Given n observations $\theta_1, \ldots, \theta_n$, we can define

$S_n(\theta) = \#$ of observations between 0 and θ/n.

the test statistic is then

$$U^2 = n \int_0^{2\pi} \left(S_n(\theta) - \frac{\theta}{2\pi} - \nu \right)^2 \frac{1}{2\pi} d\theta$$

where

$$\nu = \int_0^{2\pi} \left(S_n(\theta) - \frac{\theta}{2\pi} \right) \frac{1}{2\pi} d\theta.$$

Carry out the integrations, and we get (if $U_i = \frac{\theta_i}{2\pi}$ and $\bar{U} = \frac{1}{n}\sum U_i$)

$$U^2 = \sum U_i^2 - n\bar{U}^2 - \frac{2}{n}\sum i U_i + (n+1)\bar{U} + \frac{n}{12}.$$

The large-sample distribution U^2 for uniform data is

$$Pr(U^2 > u) = 2 \sum_{m=1}^{\infty} (-1)^{m-1} e^{-2m^2\pi^2 u}$$

It's also useful to consider

$$(U^*)^2 = \left(U^2 - \frac{0.1}{n} - \frac{0.1}{n^2}\right)\left(1 + \frac{0.8}{n}\right)$$

which is (almost) invariant in n for $n \geq 8$, and has quantiles

0.10	0.05	0.025	0.01
0.152	0.187	0.221	0.267.

Circular variables can also be tested for correlation. A standard model of dependence might be

$$\Theta = g(\Phi)$$

where Θ, Φ are circular random variables. In this model, the ordering of points in the samples $\Theta_1, \ldots, \Theta_n$ and $\Phi_1, \ldots, \Phi_n$ should be the same, so we can detect this kind of association by computing a new statistic.

Let $\beta_i = \frac{2\pi}{n} r_i$, where r_i is the rank ordering of θ_i on the circle. Similarly, $\gamma_i = \frac{2\pi}{n} s_i$ where s_i is the rank ordering of ϕ_i on the circle.

Define new circular variables by

$$\beta_1 - \gamma_1, \beta_2 - \gamma_2, \ldots, \beta_n - \gamma_n$$

and

$$\beta_1 + \gamma_1, \beta_2 + \gamma_2, \ldots, \beta_n + \gamma_n.$$

One of these should be highly concentrated if $\theta = g(\varphi)$ is orientation preserving, the other is $\theta = g(\varphi)$ is orientation reversing, so we compute

$$n^2 \bar{R}_+^2 = \left(\sum \cos(\beta_i - \gamma_i)\right) + \left(\sum \sin(\beta_i - \gamma_i)\right)$$

and

$$n^2 \bar{R}_-^2 = \left(\sum \cos(\beta_i + \gamma_i)\right) + \left(\sum \sin(\beta_i + \gamma_i)\right).$$

Since these depend only on order, they are invariant under changes in the distributions of θ, φ when θ and φ are independent.

Proposition (Mardia, 1975b, MJ p. 251),
The circular-circular rank correlation coefficient $r_0 = \max(\bar{R}_+, \bar{R}_-) \to 0$ as $n \to \infty$ if θ, φ are indepent. For $n > 10$,

$$P(2(n-1)r_0 > u) \simeq 1 - (1 - e^{-u/2})^2.$$

Another method is to consider some measure of the distance between the empirical distribution function and the product of its marginals, such as defining r_i, s_i to be the circular ranks of the θ_i, φ_i and

$$T_{i,j} = n \min(r_i, s_j) - r_i s_j$$

then defining

$$C_n = \frac{1}{n^4} \sum_{i,j} (T_{i,i} - T_{i,j} - T_{j,i} + T_{j,j})^2$$

This is called "Rothman's C_n^* test," and the approximation

$$P(16\pi^4 C_n > x) \simeq (1.466x - 0.322)e^{-x/2}$$

is used in practice.

Of course, mutual information on T^2 would be another option!