FACULTADE DE MATEMÁTICAS

Traballo Fin de Grao

UNHA INTRODUCIÓN Á ANÁLISE DE DATOS CIRCULARES

Sara Álvarez Lorenzo

2024/2025

UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

GRAO DE MATEMÁTICAS

Traballo Fin de Grao

UNHA INTRODUCIÓN Á ANÁLISE DE DATOS CIRCULARES

Sara Álvarez Lorenzo

Xuño, 2025

UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Traballo proposto

Área de Coñecemento: Estatística e Investigación Operativa, Dpto. de Estatística, Análise Matemática e Optimización
Título: Unha introdución á análise de datos circulares
Breve descrición do contido
Os datos circulares son observacións que se poden representar como puntos na circunferencia do círculo unidade. Neste Traballo de Fin de Grao revisaranse ferramentas clásicas para a análise descritiva de mostras de datos circulares, introduciranse algúns modelos clásicos de distribución e presentaranse algúns contrastes clásicos (uniformidade, comparación de distribucións). As distintas ferramentas ilustraranse con datos simulados e datos reais, empregando o software R.

Índice

Resumo

Os datos circulares son observacións que se identifican como puntos ou vectores na circunferencia do círculo unidade. Neste Traballo de Fin de Grao consideraremos ferramentas clásicas para a análise descritiva de mostras de datos circulares, introduciremos algúns modelos destacados de distribución e presentaremos algúns procedementos inferenciais para este tipo de observacións, entre eles contrastes (uniformidade, bondade de axuste) e estimacións. Os distintos instrumentos estatísticos serán ilustrados con datos simulados e reais, empregando o software R.
O traballo organízase en tres capítulos diferenciados. No primeiro deles, exporemos medidas descritivas para mostras circulares (medidas de posición, de dispersión e representacións gráficas); no segundo repararemos nas distribución circulares máis prominentes e nos métodos dos que dimanan; e no derradeiro capítulo, centrarémonos en coñecer e manexar os útiles que a estatística inferencial pon a nosa disposición.

Abstract

Circular data is data that can be identified as points or vectors within the unit circle. In this bachelor thesis we will consider classic statistical tools designed for this kind of data, as well as introduce the most important distribution models and inference procedures, including tests (about uniformity, or goodness-of-fit) and estimates of the parameters. This circular theory will be exemplified using simulated and real data in R software.
This work is structured in three differentiated chapters. In the first one, we will dive in the definitions of descriptive statistics for circular data (measures of location, dispersion and graphical representation); in the second one, we will study how to construct circular distributions and expound the most significant ones; and in the last one, we will show basic inference instruments and model fitting for a single sample.

Introdución

Este traballo inscríbese no que podemos considerar como substancia esencial da humanidade: a traslación da realidade palpable á inmaterialidade das linguas (linguas entendidas no senso máis libre, como aparello inmaterial de comunicación intrahumana). A estatística, como disciplina descritiva, responde a esta arela de aprehender os acaecementos que ocorren ao noso redor e que, só coas mans, non somos quen de abranguer.
Esta arela é a que, a comezos do pasado século, experimentaron as científicas de diversos campos ao decatárense de que os modelos cos que estaban a traballar eran incapaces de bosquexar de xeito atinado a realidade que procuraban describir: a nacenza de modelos direccionais foi a resposta natural para suplir estas insuficiencias.
Estes modelos están deseñados para tratar datos circulares, que son observacións direccionais no plano real $\mathbb{R}^2$ que descansan sobre a circunferencia unidade $\mathbb{S}^1$. Dada a esencialidade da súa natureza direccional, o módulo de cada observación carece de relevancia descritiva, logo pode ser normalizado sen caer en perda de información, e así, ser rexistrado sobre $\mathbb{S}^1$.

Este tipo de datos xorden de xeito orgánico en múltiples escenarios da existencia material na que estamos inmersas: moitos animais expresan condutas fundamentalmente direccionais, coma a disposición de cabaliños do demo (*Sympetrum*) para tomar o Sol ([3], Capítulo 1), a orientación solar do *Fundulus notii*, un peixe común en Norteamérica ([3], Capítulo 4), ou o desplazamento de estrelas mariñas (*Astropecten jonstoni*) durante a migración primaveral ([3], Capítulo 7); eventos xeolóxicos satisfacen modelos circulares, como a distribucións das crebaduras naturais en terreos [15], o estudo das direccións tomadas por un tsunami acontecido en 2004 en Banda Aceh, Indonesia [8]; mesmo hai aplicacións nas ciencias que estudan o corpo, por exemplo, dentro das ramas da bioinformática das proteínas [13] ou na investigación médica, onde se ten estudado o momento óptimo do ciclo menstrual no que operar cancro de mama [5].
Reparemos que, neste último caso, os datos circulares non son só un aparello útil para representar orientacións, senón que tamén resultan eficaces para estudar as evolucións temporais dende a óptica circular que caracteriza os ciclos biolóxicos e horarios, en troques da linealidade cronolóxica habitual en moitas investigacións estatísticas.

A descuberta dos datos direccionais fixo agromar unha nova galla na disciplina estatística, que fraguou ferramentas novas para o seu tratamento.
De xeito preliminar, reparemos na existencia de varios procederes para identificar os datos circulares (que, recordemos, son esencialmente direccións bidimensionais): podemos pensalos ben coma puntos sobre a circunferencia unidade, ben coma ángulos ou ben coma vectores unitarios. Porén, esta identificación non é única, cambia segundo a elección da orixe (de onde comezamos a medir) e do sentido de rotación (como medimos). Por mor disto, as conclusións que destilemos a partir dos datos non deberían depender da escolla que fagamos respecto á orixe e ao sentido de rotación.

Exemplo 0.1. Exemplo extraído de [9], Capítulo 1. Un ángulo de 60° visto por unha matemática que toma a dirección Leste coma a dirección de orixe e sentido de rotación antihorario é visto por unha xeóloga que toma a dirección Norte coma orixe e sentido de rotación horario como un ángulo de 30°.

Para solventar esta contrariedade, empregaremos sistemas de coordenadas que nos permitan determinar de xeito unívoco a posición direccional dos datos: o sistema de coordenadas rectangulares, $P \equiv (x, y)$; e o sistema de coordenadas polares $P \equiv (r, \theta)$. Xa que o interés primordial é a dirección, traballaremos sobre a circunferencia unidade $r = 1$, e, en xeral, mediremos $0 \leq \theta < 2\pi$ en sentido antiohorario dende o eixo horizontal positivo. [1] Vainos ser de grande utilidade coñecer o cambio dunhas coordenadas a outras, que se fai mediante

$$x = \cos\theta, \qquad y = \operatorname{sen}\theta.$$

Outra posible representación é identificando o ángulo ou vector cun número no plano complexo, $z = \cos\theta + \mathbf{i}\operatorname{sen}\theta = x + \mathbf{i}y \in \mathbb{C}$, onde $\mathbf{i}$ é a unidade imaxinaria. Estas tres identidades vannos ser de proveito en capítulos vindeiros.

Atendendo ao anterior, unha mostra de observacións circulares independentes vaise designar coma $\theta_1, \ldots, \theta_n$, sendo $\theta_i \in [0, 2\pi)$ un ángulo medido en radiáns en sentido antihorario.
Unha variable circular aleatoria Θ seguirá unha distribución circular, que reparte a probabilidade total sobre a circunferencia. Neste caso, nomearemos as correspondentes funcións de distribución e densidade coma $F(\theta)$, $f(\theta)$, respectivamente. No caso de ter necesidade de estimalas, o estimador expresarase en termos de $\hat{F}(\theta)$, $\hat{f}(\theta)$. Asemade, os estimadores para os distintos parámetros que precisemos manexar tamén quedarán indicados con $\hat{\mu}, \hat{\rho}, \hat{\sigma}, \ldots$. As funcións de distribución e

[1] Nos exemplos é posible que traballemos con ángulos en grados $[0, 360)$ ou en outros intervalos en radiáns, coma $[-\pi, \pi)$; pero será indicado no momento necesario.

densidade para variables aleatorias lineares X denotámolas por $G(x)$, $g(x)$.

Ao longo deste traballo ilustraremos os conceptos definidos con datos reais, recollidos nos paquetes de R `NPCirc` [18] e `circular` [2], a saber:

- `cycle.changes`: as observacións colectadas rexistran os cambios nos ciclos de temperatura a nivel do chan en Monte Alvear (Arxentina), unha rexión periglaciar. O dataset inclúe 350 observacións que se corresponden ás horas nas que a temperatura troca de positiva a negativa e viceversa, abarcando os días entre Febreiro de 2008 e Marzo de 2009.
  Segundo [17], estas observacións son de extrema utilidade para dar conta dos efectos do cambio climático nas rexións glaciares; en tanto o retroceso dos mesmos é consecuencia dun balance negativo na masa de xeo, isto é, na diferencia entre acumulación e ablación.

- `sandhoppers`: este dataset recolle a orientación, medida baixo condicións naturais e outras variables de interese para analizar a ductilidade da conduta de dúas especies de pulgas de praia, *Talitrus saltator* e *Talorchestia brito*. O experimento foi levado a cabo nunha zona de area sen exposición ás ondas da praia de Zouara, na costa nordés de Túnez.
  O interese de analizar o comportamento destes anfípodos estriba na súa capacidade orientativa, que é indicativa do estado do ecosistema litoral. En xeral, as pulgas de praia permanecen soterradas baixo area húmida durante as mareas altas diúrnas, e retornan á superficie na nocturnidade. Se ao longo do día sofren algún desplazamento ou a area na que se agochan seca, estes artrópodos tentarán tornar cara a ribeira, facéndose guiar polo acimut do Sol ou outras sinais locais. Esta capacidade de adaptación está en estreita relación co estado da praia: nun ecosistema máis protexido e estable, as pulgas de praia presentan menos variación na dirección dos seus saltos ca nunha praia erosionada e exposta. Máis información a este respecto pode atoparse en [11].

- `zebrafish`: os datos son medicións extraídas dun experimento (máis detalles respecto ao mesmo en [14]) sobre larvas de peixes cebra, axexadas por un robot predador camuflado coma un peixe cebra adulto. Recolléronse os ángulos de escape de cada peixe e os ángulos nos que a ameaza foi percibida.
  A análise das estratexias óptimas de escape das presas é vital para o estudo das conductas animais. Neste caso, a toma de datos como direccións é vital para poder conducir con efectividade o estudo e inferir conclusións válidas, tales como as que se destilan en [1].

- `pigeons`: as observacións identifícanse coas direccións no horizonte nas que se deixou de percibir o voo de pombas mensaxeiras que foron trasladadas e logo liberadas en dúas localizacións descoñecidas para elas. Os ángulos foron medidos respecto á dirección da residencia á que tiñan que voltar (identificada coa posición de 360°).

> O estudo da xeolocalización das pombas mensaxeiras ten un longo percorrido na literatura da Bioloxía. A asunción máis estendida respecto deste tema é que as pombas dependen dun mapa magnético interno, que trazan experimentando cambios nos campos magnéticos dos lugares polos que voan, para orientarse con éxito. Porén, Gagliardo e coautores apuntan nun estudo [7] que é a anosmia (incapacidade olfativa) a que supón un verdadeiro impedimento para a retorna ao fogar das pombas.

Estes casos prácticos, baseados en datos reais, agromarán paulatinamente no traballo, que está dividido en tres capítulos: no primeiro deles, introduciremos ferramentas descritivas e gráficas que nos permitirán facer un estudo preliminar dalgún destes datos; no segundo, extenderémonos nas distribucións poboacionais existentes para datos circulares; e no ulterior exporemos métodos e artefactos estatísticos que hannos de posibilitar o facer inferencia nas devanditas distribucións. Tanto o quefacer teórico como empírico están cimentados no coñecemento que temos adquirido en diversas materias do grao de Matemáticas da Universidade de Santiago de Compostela, especialmente en Elementos de Probabilidade e Estatística, Probabilidade Estatística e Inferencia Estatística.

Capítulo 1

Ferramentas descritivas

A estatística descritiva é un instrumento esencial no tratamento preliminar dos datos, xa que permite organizar e sintetizar as características máis sobranceiras dunha mostra e bosquexar o seu comportamento, así como suxerir de que distribucións é verosímil que promane. Este sumario alicérzase sobre medidas de posición e dispersión, e sobre as posibles representacións da mostra. Neste capítulo introduciremos estas ferramentas, cimentándonos na labor de [6], Capítulo 1 e [19], Capítulos 2 e 3.

1.1. Medidas descritivas

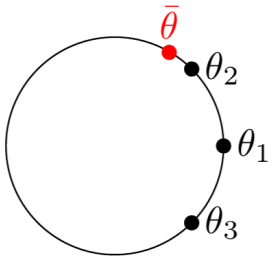
As medidas descritivas de localización e dispersión máis comúns para datos euclídeos (en $\mathbb{R}$ ou $\mathbb{R}^d$) son a media e a varianza (ou covarianza en dimensión superior) mostrais. De xeito intuitivo, poderiamos pensar que as análogas aritméticas do caso real serán de utilidade no tratamento de datos circulares. Non obstante, ningunha das dúas son adecuadas para traballarmos con ángulos.

Exemplo 1.1. Se tiveramos unha mostra de tres ángulos, $\theta_1 = 0$, $\theta_2 = \frac{\pi}{4}$ e $\theta_3 = \frac{3\pi}{4}$, a súa media aritmética sería

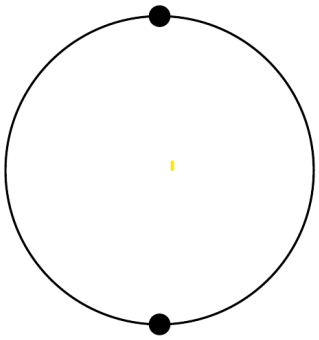
$$\bar{\theta} = \frac{1}{3}\sum_{i=1}^{3}\theta_i = \frac{\pi}{3}.$$

Porén, ao obsevarmos a figura 1.1a, decatarémonos da inadecuación da media aritmética como medida de localización para datos circulares. Ademais, se, en troques de traballar en $[0, 2\pi)$, identificamos os ángulos en $[-\pi, \pi)$, de xeito que a mostra se expresaría coma $\theta_1 = 0$, $\theta_2 = \frac{\pi}{4}$ e $\theta_3 = -\frac{\pi}{4}$, entón a media aritmética mudaría a $\bar{\theta} = 0$, malia que a posición dos datos sobre a circunferencia sería a mesma.

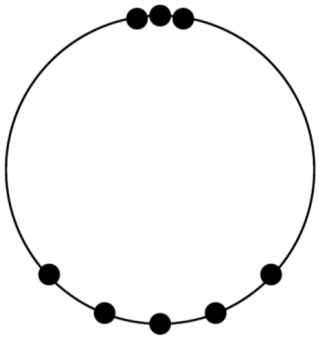
Desta incompatibilidade é nada a necesidade de definir outras medidas resumo adaptadas a



(a) A media arimética da mostra (obtida como se os ángulos foran valores reais) non é adecuada como dirección de referencia.



(b) Non sempre é posible definir a dirección media mostral. Isto acontece en mostras onde temos un ángulo e o seu oposto.



(c) A lonxitude do vector resultante non é a medida de dispersión mais adecuada en presenza de multimodalidade.

Figura 1.1: Ilustración das problemáticas nas medidas de localización e dispersión para datos ciruculares.

este tipo de mostras, que non dependan da elección da orixe e do sentido de rotación.

1.1.1. Medidas de localización

As medidas de localización serven para atopar datos cunha posición destacada na mostra, indicativos en moitas ocasións dunha tendencia no comportamento das observacións.

Dirección Media Mostral $\bar{\theta}$

A dirección media dun conxunto de observacións circulares vai ser a dirección do seu vector resultante. Así, para unha mostra aleatoria independente dunha variable circular, $\theta_1, \dots, \theta_n$, que podemos identificar coa colección de vectores unitarios $\{\cos\theta_i, \operatorname{sen}\theta_i\}_{i=1}^n$, o vector resultante é

$$\mathbf{R} = \left(\sum_{i=1}^{n} \cos\theta_i, \sum_{i=1}^{n} \operatorname{sen}\theta_i \right) := (C, S). \tag{1.1}$$

A súa norma denotarémola por $R = \|\mathbf{R}\| = \sqrt{C^2 + S^2}$.

Definición 1.2. Definimos a dirección media mostral como o argumento do número complexo correspondente ao vector resultante, i.e.

$$\bar{\theta} = \arg\left(\sum_{i=1}^{n} \cos\theta_i + \mathbf{i} \sum_{i=1}^{n} \operatorname{sen}\theta_i \right),$$

ou, equivalentemente, mediante as ecuacións

$$\cos\bar{\theta} = \frac{C}{R}, \quad \operatorname{sen}\bar{\theta} = \frac{S}{R}. \tag{1.2}$$

Podemos calcular o seu valor por medio da seguinte función:

$$\bar{\theta} = \operatorname{atan2}(S, C) = \begin{cases} \arctan(S/C) & \text{se} \quad C > 0, \quad S \geq 0, \\ \pi/2 & \text{se} \quad C = 0 \quad S > 0, \\ \arctan(S/C) + \pi & \text{se} \quad C < 0, \\ \arctan(S/C) + 2\pi & \text{se} \quad C \geq 0 \quad S < 0, \\ \text{indefinido} & \text{se} \quad C = 0, \quad S = 0. \end{cases} \tag{1.3}$$

Así definida, a dirección media mostral elude os problemas que xordían coa media aritmética, xa que independentemente da escolla da orixe e do sentido rotacional que fagamos, $\bar{\theta}$ conservará o seu valor. Ademais, de xeito simétrico ao caso linear, a dirección media é rotacionalmente equivariante.

Proposición 1.3. *A dirección media mostral é rotacionalmente equivariante, é dicir, se xiramos os datos unha certa cantidade, o valor de $\bar{\theta}$ cambia nesa mesma cantidade.*

Demostración. Sexa $\theta_1, \ldots, \theta_n$ unha mostra con dirección media $\bar{\theta}$. Consideremos outra mostra $\theta_1 + c, \ldots, \theta_n + c$ e vexamos que ten dirección media $\bar{\theta} + c$.
O vector resultante para esta segunda mostra será

$$\mathbf{R}' = \left(\sum_{i=1}^{n} \cos(\theta_i + c), \sum_{i=1}^{n} \operatorname{sen}(\theta_i + c) \right) = (C', S').$$

Logo,

$$C' = \sum_{i=1}^{n} \cos(\theta_i + c) = \sum_{i=1}^{n} (\cos\theta_i \cos c - \operatorname{sen}\theta_i \operatorname{sen} c) \overset{(1.1)}{=} C\cos c - S\operatorname{sen} c =$$
$$\overset{(1.2)}{=} R\cos\bar{\theta}\cos c - R\operatorname{sen}\bar{\theta}\operatorname{sen} c = R\cos(\bar{\theta} + c).$$

De xeito análogo, chegamos a que $S' = R\operatorname{sen}(\bar{\theta} + c)$. Agora ben,

$$R' = \|\mathbf{R}'\| = \sqrt{C'^2 + S'^2} = \sqrt{[R\cos(\bar{\theta} + c)]^2 + [R\operatorname{sen}(\bar{\theta} + c)]^2} = R.$$

Así, podemos concluír que

$$\frac{C'}{R'} = \cos(\bar{\theta} + c), \quad \frac{S'}{R'} = \operatorname{sen}(\bar{\theta} + c),$$

tal e como queriamos ver. □

Ao contrario ca no caso linear, esta medida non sempre existe para calquera conxunto de datos. En casos extremos, cando na mostra atopamos un ángulo e o seu oposto antipodal, xorde unha ambigüidade irresoluble, dada esta definición, que causa que a media non estea definida.

Exemplo 1.4. Se tiveramos unha mostra como a representada na figura 1.1b, con dous datos en direccións opostas, non poderiamos achar de modo inapelable a dirección media.

Por este motivo, se o vector resultante ten lonxitude R positiva, a súa dirección $\bar{\theta}$ é tomada como a dirección media circular. Doutra banda, se $R = 0$, non podemos definir a dirección media.

Dirección Mediana Mostral $\tilde{\theta}$

Coma no caso linear, a mediana personifícase coma unha alternativa robusta á media, especialmente se a mostra presenta asimetría.

Definición 1.5. Unha dirección mediana é calquera ángulo $\tilde{\theta}$ para o que a metade dos datos da mostra caian no arco $[\tilde{\theta}, \tilde{\theta} + \pi)$, e a maioría de puntos estean máis próximos a $\tilde{\theta}$ que a $\tilde{\theta} + \pi$.

Observación 1.6. Denotamos por n o número de datos da mostra. Se n é impar, a mediana coincide con algún dos datos tomados. Por contra, se n é par, usualmente a mediana vaise corresponder co promedio dos puntos inmediatamente a súa esquerda e dereita.

De xeito formal, podemos defini-la mediana como o argumento que minimiza a medida de dispersión seguinte

$$d_2(\psi) = \frac{1}{n} \sum_{i=1}^{n} (\pi - |\pi - |\theta_i - \psi||). \tag{1.4}$$

Dirección Modal Mostral $\breve{\theta}$

Definición 1.7. A dirección modal mostral $\breve{\theta}$ é a dirección correspondente á máxima concentración dos datos da mostra.

Observación 1.8. Notemos que as definicións de mediana e moda mostrais non garanten unicidade: é posible (e ocorrerá en moitos casos prácticos), que unha mostra sexa multimodal ou teña varias medianas.

1.1.2. Medidas de dispersión

As medidas de dispersión son ferramentas estatísticas que serven para dar unha idea da diseminación que teñen os datos respecto da media mostral. Para variables circulares, hai unha variedade ampla de medidas desta categoría que se poden detallar.

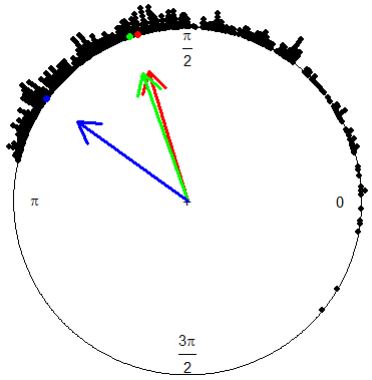


Figura 1.2: Medidas de localización para os estímulos en `zebrafish`: dirección media $\bar{\theta}$ (vermello), dirección mediana $\tilde{\theta}$ (verde) e dirección modal $\check{\theta}$ (azul) mostrais.

Para comezar, o vector resultante (1.1) da mostra non só dá unha medida razoable para a dirección media, senón que a súa lonxitude $R = \|\mathbf{R}\|$ é un aparello que dá conta da concentración respecto de $\bar{\theta}$. Advirtamos que $R \in [0, n]$, de xeito que, baixo unimodalidade, se R é próximo a n os datos apuntarán cara a mesma dirección; mentres que se é próximo a 0, haberá moita máis dispersión, con datos disgregados de xeito equitativo. O análogo ocorre coa lonxitude normalizada $\bar{R} = \dfrac{R}{n} = \sqrt{\bar{C}^2 + \bar{S}^2} \in [0, 1]$, onde $\bar{C} = \frac{C}{n}$, $\bar{S} = \frac{S}{n}$.
Porén, se a mostra é multimodal, estas dúas medidas poden non ser bós indicadores para dispersión, como se pode ver na figura 1.1c, onde, malia que $\bar{R} = 0$, atopámonos con que a dispersión non é uniforme, senón que está estruturada en dous grupos diferenciados.

Definición 1.9. Denotamos por varianza circular mostral a $V = 1 - \bar{R} \in [0, 1]$.

É interesante observar tamén que, a diferenza do caso linear, a varianza circular toma valores entre $[0, 1]$. Canto menor sexa o valor de V, máis concentrada respecto da dirección media estará a nosa mostra; se ben, coma ocorría coa lonxitude do vector resultante, un valor de V próximo a un non é indicativo necesariamente dun caso de dispersión máxima, polo menos en presenza de multimodalidade.

Definición 1.10. Definimos a desviación típica mostral coma $\hat{\sigma} = \sqrt{-2\log(1 - V)} \in [0, \infty)$.

Observación 1.11. Para mostras concentradas, con valores de V miúdos, podemos aproximar $\hat{\sigma} \approx \sqrt{2V}$.

A maiores destas medidas, é produtivo definir outras baseadas en distancias entre ángulos na circunferencia. Se tomamos a distancia

$$d(\psi, \omega) = 1 - \cos(\psi - \omega), \tag{1.5}$$

e identificamos as observacións $\theta_1, \dots, \theta_n$ cos vectores unitarios $\{\mathbf{u}_i\}_{i=1}^n = \{(\cos\theta_i, \operatorname{sen}\theta_i)\}_{i=1}^n$, entón podemos calcular a dispersión mostral respecto dun vector $\mathbf{v}$ mediante a seguinte expresión, sen máis ca computar ψ_i, o ángulo entre o vector $\mathbf{u}_i$ e $\mathbf{v}$,

$$D_{\mathbf{v}}(\mathbf{u}_1, \dots, \mathbf{u}_n) = \sum_{i=1}^n d(\mathbf{v}, \mathbf{u}_i) = n - \sum_{i=1}^n \cos\psi_i.$$

Se poñemos por caso que $\mathbf{v} = \mathbf{R}$ é o vector resultante, entón a dispersión mostral é mínima e $D_{\mathbf{R}}(\mathbf{u}_1, \dots, \mathbf{u}_n) = n - R$, de xeito que o valor $n - R$ (e a súa normalización $1 - \bar{R}$) dan conta da dispersión da mostra relativa á dirección media mostral, proporcionando así motivación para a definición da varianza circular mostral que detallamos con anterioridade.

Doutra banda, se nos fora dada a dirección media poboacional μ, e tomamos o chamado vector polar $\mathbf{P} = (\cos\mu, \operatorname{sen}\mu)$, entón podemos medir a dispersión mostral respecto de μ,

$$D_{\mathbf{P}}(\mathbf{u}_1, \dots, \mathbf{u}_n) = n - \sum_{i=1}^n \cos(\theta_i - \mu) = n - V_0,$$

onde

$$\begin{aligned} V_0 &= \sum_{i=1}^n \cos(\theta_i - \mu) = \sum_{i=1}^n (\cos\theta_i \cos\mu + \operatorname{sen}\theta_i \operatorname{sen}\mu) = C\cos\mu + S\operatorname{sen}\mu = \\ &= R\cos\bar{\theta}\cos\mu + R\operatorname{sen}\bar{\theta}\operatorname{sen}\mu = R\cos(\bar{\theta} - \mu). \end{aligned}$$

Isto é, V_0 representa a lonxitude da proxección do vector resultante $\mathbf{R}$ cara a dirección polar $\mathbf{P}$. Coma queira que $\cos(\bar{\theta} - \mu) \leq 1$, entón $V_0 \leq R$, e a igualdade entre ambos ocorre se e soamente se $\bar{\theta} = \mu$.

Teorema 1.12. *Se $\bar{\theta}$ denota á dirección do vector resultante da mostra $\theta_1, \dots, \theta_n$, entón*

$$\sum_{i=1}^n \operatorname{sen}(\theta_i - \bar{\theta}) = 0 \quad e \quad \sum_{i=1}^n \cos(\theta_i - \bar{\theta}) = R.$$

Demostración. O resultado séguese de que

$$\begin{aligned} \sum_{i=1}^n \operatorname{sen}(\theta_i - \bar{\theta}) &= \sum_{i=1}^n (\operatorname{sen}\theta_i \cos\bar{\theta} - \cos\theta_i \operatorname{sen}\bar{\theta}) = S\cos\bar{\theta} - C\operatorname{sen}\bar{\theta} = \\ &= R\cos\bar{\theta}\operatorname{sen}\bar{\theta} - R\cos\bar{\theta}\operatorname{sen}\bar{\theta} = 0, \\ \sum_{i=1}^n \cos(\theta_i - \bar{\theta}) &= \sum_{i=1}^n (\cos\theta_i \cos\bar{\theta} + \operatorname{sen}\theta_i \operatorname{sen}\bar{\theta}) = C\cos\bar{\theta} + S\operatorname{sen}\bar{\theta} = \\ &= R\cos^2\bar{\theta} + R\operatorname{sen}^2\bar{\theta} = R. \end{aligned}$$

□

Con esta mesma definición de distancia, podemos definir unha medida de dispersión da mostra $\theta_1, \dots, \theta_n$ respecto dun ángulo ψ sen máis ca tomar

$$d_1(\psi) = \frac{1}{n}\sum_{i=1}^{n}[1 - \cos(\theta_i - \psi)]. \tag{1.6}$$

Esta medida de dispersión é mínima cando $\psi = \bar{\theta}$, e nese punto toma valor $d_1(\bar{\theta}) \overset{(1.12)}{=} 1 - \bar{R} = V$. Usando (1.5), a distancia media entre as observacións vén dada por

$$\bar{d_1} = \frac{1}{n^2}\sum_{i,j=1}^{n}(1 - \cos[\theta_i - \theta_j)] \overset{(1.12)}{=} 1 - \bar{R}^2.$$

Outra distancia comunmente empregada na circunferencia é

$$d(\psi, \omega) = \text{mín}\{\psi - \omega,\ 2\pi - (\psi - \omega)\} = \pi - |\pi - |\psi - \omega||. \tag{1.7}$$

A medida de dispersión da mostra respecto dun ángulo ψ asociada a esta distancia xa a adiantabamos anteriormente, en (1.4), a saber:

$$d_2(\psi) = \frac{1}{n}\sum_{i=1}^{n}(\pi - |\pi - |\theta_i - \psi||) = \pi - \frac{1}{n}\sum_{i=1}^{n}|\pi - |\theta_i - \psi||.$$

Esta medida é mínima cando $\psi = \tilde{\theta}$, e o valor que toma neste caso noméase desviación media:

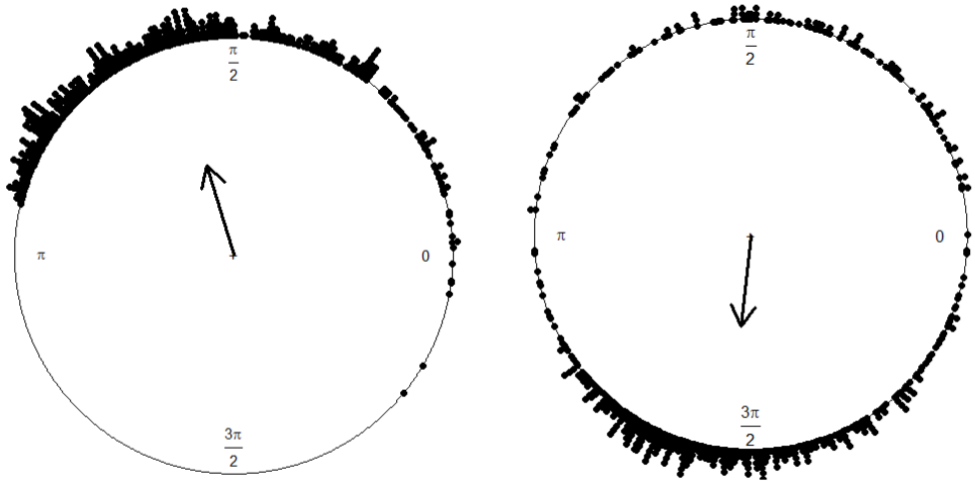
$$d_2(\tilde{\theta}) = \pi - \frac{1}{n}\sum_{i=1}^{n}|\pi - |\theta_i - \tilde{\theta}||.$$

Empregando (1.7), a distancia media entre os datos da mostra exprésase coma segue:

$$\bar{d_2} = \frac{1}{n^2}\sum_{i,j=1}^{n}(\pi - |\pi - |\theta_i - \theta_j||).$$

Por último, outra medida de dispersión que paga a pena nomear é o rango mostral, que é a lonxitude angular do menor arco que abarca todas as observacións.

Para ilustrar o significado práctico deste compendio de medidas, imos fixarnos nos ángulos de estímulo e resposta de peixes cebra, representados na figura 1.3. Segundo o observable nestes grafos, o que podemos esperar é que as medidas indiquen maior dispersión na resposta ca no estímulo. En efecto, a lonxitude normalizada do vector resultante é máis próxima a un nos ángulos de estímulo, indicando unha maior concentración das observacións cara a dirección media. Resulta coherente, en consecuencia, que tanto a varianza circular coma a desviación típica mostrais sexan maiores nos ángulos de resposta. Consonte a isto, a distancia media entre as observacións (empregando como medida (1.6)) será maior no derradeiro grupo, así coma tamén o será o rango mostral.



(a) Ángulos de estímulo. (b) Angulos de resposta.

Figura 1.3: Ángulos de estímulo e resposta de peixes cebra, rexistrados no dataset `zebrafish`, coas súas respectivas medias mostrais (identificadas cunha frecha) de lonxitude $\bar{R}$.

Peixes Cebra	$\bar{R}$	$V = 1 - \bar{R}$	$\hat{\sigma} = \sqrt{-2\log(1-V)}$	$\bar{d}_1 = 1 - \bar{R}^2$	Rango (en radiáns)
Estímulo	0.7775	0.2225	0.7094	0.3954	3.5838
Resposta	0.4232	0.5768	1.3114	0.8209	6.0777

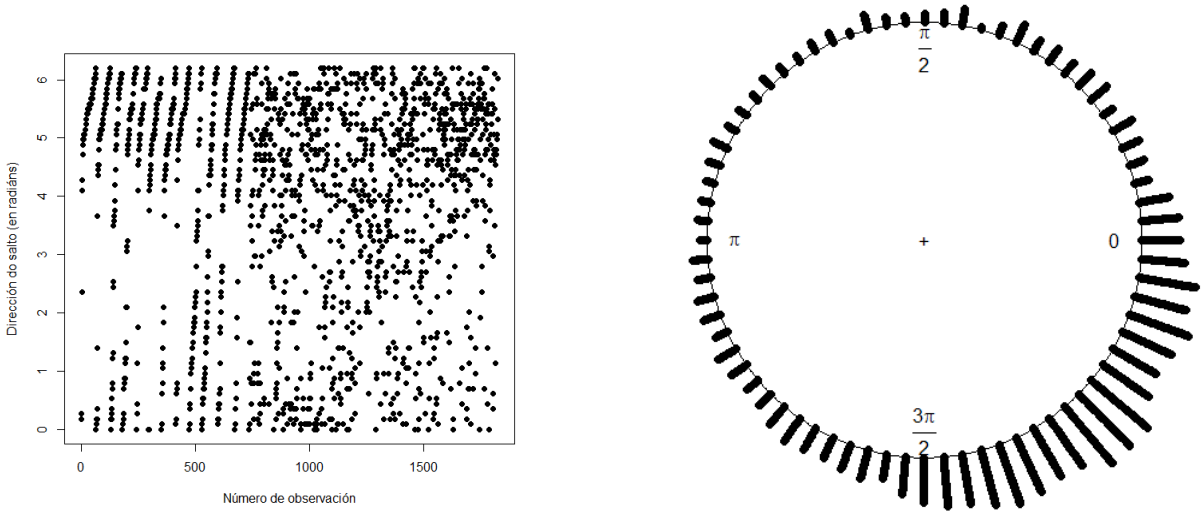
Cadro 1.1: Medidas de dispersión para ángulos de estímulo e resposta de peixes cebra.

1.2. Representacións Gráficas

Reproducir graficamente un conxunto de observacións circulares é valioso na análise inicial dos datos, por mor da facilidade que as representacións visuais teñen para facernos gañar unha idea preliminar das características máis sobresaíntes da mostra, para suxerir posibles modelos de distribución subxacentes, ou para acentuar algunhas medidas descritivas de xeito intuitivo. Deseguido imos afondar nos instrumentos de representación para datos circulares máis habituais no ámbito científico.

1.2.1. Representación de datos en cru

A representación de datos orbiculares máis inmediata que podemos bosquexar baséase en reproducir as observacións mostrais coma puntos sobre a circunferencia. Ademais, para dar conta da densidade de datos que se pode acumular nunha dirección, é necesario amorear oblicuamente as observacións nese sentido.



(a) A representación habitual para datos lineares resulta improdutiva no caso circular.

(b) Os grafos en forma orbicular son máis adecuados para mostras circulares.

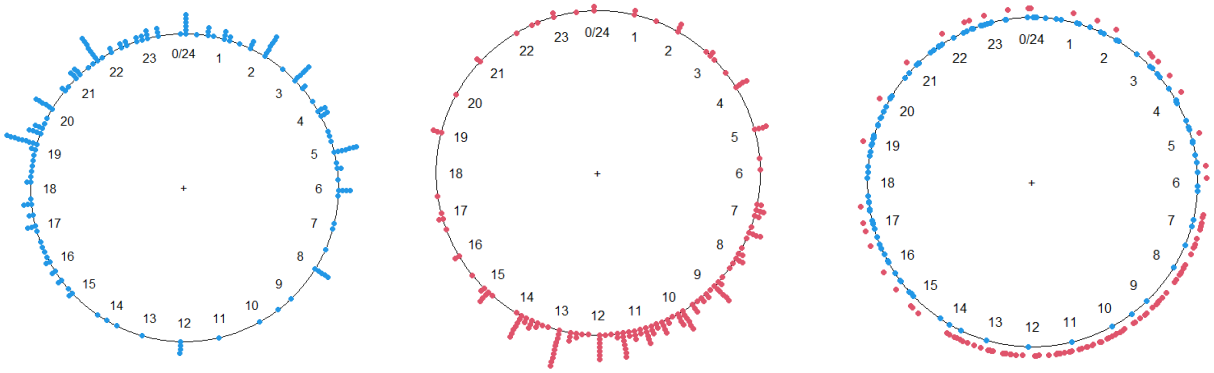
Figura 1.4: Representación linear e circular referida ás direccións de salto de pulgas de area, recollidas no dataset `sandhoppers`.

Como podemos observar na figura 1.4, se tentamos empregar unha representación bidimensional linear para representar datos circulares (1.4a), é difícil extraer conclusións sobre a súa tendencia xeral, a localización dunha posible dirección media, a presenza de pequenos grupos modais ou a existencia dalgunha observación discordante. No entanto, se optamos por unha representación orbicular (1.4b), podemos inferir os valores posibles que poden tomar algunhas das medidas descriptivas que temos tratado anteriormente.

Conforme ao que acabamos de facer notar, cando forzamos a que os datos se acumulen uns enriba doutros, é máis sinxelo detectar detalles sobre a mostra. Non obstante, cando traballamos con distintos grupos dunha mesma poboación pode ser máis interesante comparar as direccións observadas nuns e noutros que a densidade de datos que se pode chegar a acumular nunha dirección dada.

Para ilustrar isto, imos tomar de exemplo os datos recollidos en `cycle.changes`, que, relembremos, rexistraban as horas nas que se producían cambios nos ciclos de temperatura (de xeada a desxeo e viceversa) a nivel do chan na zona periglacial de Monte Alvear, en Arxentina.

Como é de esperar, os cambios no ciclo de temperatura a xeada (1.5a) orixínanse maioritariamente nas horas nas que o Sol dispensa pouca luz e calor ou nas horas nocturnas, mentres que os cambios a desxeo (1.5b) ocorren principalmente en horario diúrno. Esta diferenza temporal entre as variacións gaña énfase ao representarmos as observacións de ambos grupos simultaneamente



(a) Horas nas que se produce cambio de ciclo a xeada.

(b) Horas nas que se produce cambio de ciclo a desxeo.

(c) Comparativa dos cambios de ciclo.

Figura 1.5: Distintas representacións para os cambios nos ciclos de temperatura en Monte Alvear segundo a hora do día.

(1.5c), se ben perdemos información de en que momentos estas variacións danse de xeito máis acusado.

1.2.2. Histogramas

Os histogramas son un artiluxio estatístico moi útil para destilar instintivamente a distribución subxacente dunha mostra, polo que resulta útil procurar o traslado do seu uso ao caso circular. Veremos que deste transvase ao noso material de estudo son nados novos problemas respecto das eleccións que esixe a construción dun histograma, e que se herdan os conflitos usuais no caso linear (elección de lonxitude e número de intervalos).

Histogramas lineares

A idea que subxace na elaboración dun histograma linear con datos orbiculares é a de acernar a circunferencia nun punto e asociar os dous extremos que xorden dese corte aos extremos do histograma. Por causa disto, abrolla un dilema que no caso linear non se daba, a saber: a escolla dun punto de corte adecuado. A toma dunha mala decisión a este respecto pode levar a unha importante distorsión na representación dos datos.

Por exemplo, se consideramos as direccións que tomaban as pombas no horizonte, recollidas en `pigeons`, unha intuición inicial pode facer que nos decantemos por cortar a circunferencia no 0. Se optamos por facer isto, o histograma que imos obter (figura 1.7a) vainos dar a idea (errada, polo observable na figura 1.6) de que a nosa mostra é bimodal. Por outro lado, se cambiamos o punto de corte a π radiáns, a dirección na que menos concentración de datos parece haber,

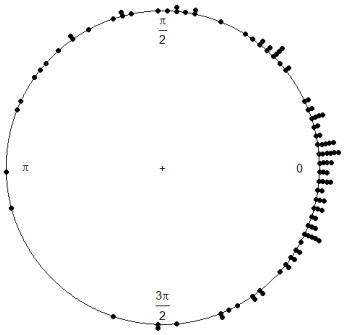
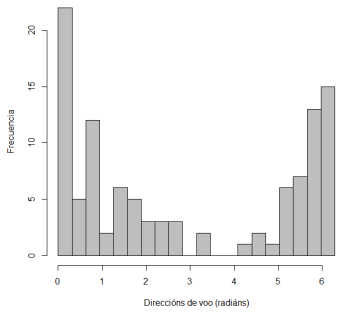
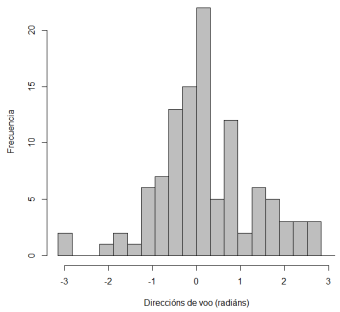


Figura 1.6: Grafo circular da dirección de voo das pombas.

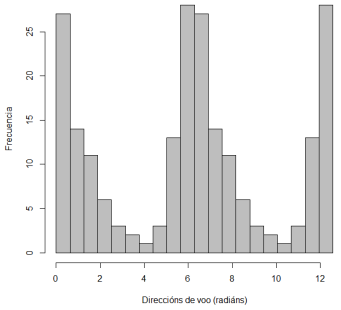
o consecuente histograma (figura 1.7b) semella reflectir con maior fidelidade a distribución das observacións.



(a) Histograma con punto de corte en 0 radiáns.



(b) Histograma con punto de corte en π radiáns.



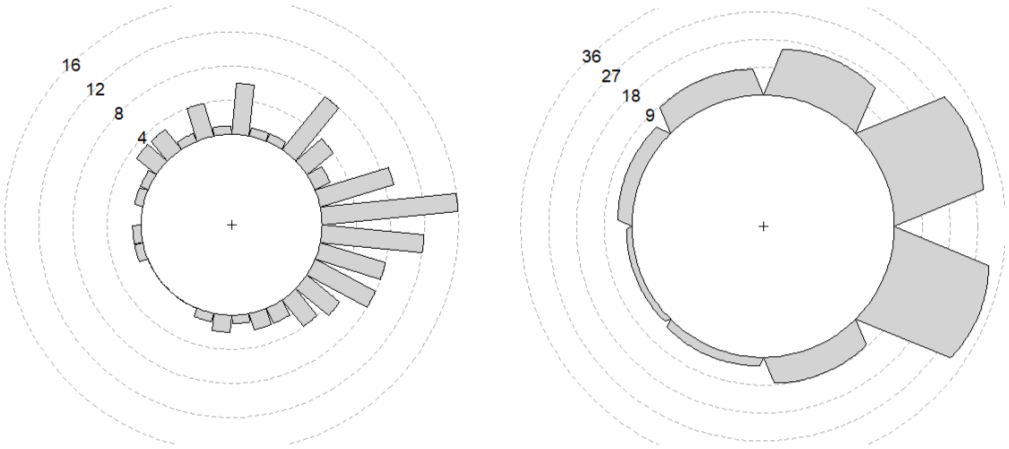
(c) Histograma con corte en 0 radiáns, repetindo ciclo.

Figura 1.7: Histogramas lineares para as direccións de voo das pombas.

Outro xeito de solventar este problema é repetir no histograma un ciclo completo. Facendo isto para o noso exemplo, a figura resultante 1.7c tamén manifféstase coma unha representación adecuada da distribución dos datos observable na figura 1.6.

Histogramas angulares

Un modo diferente de solucionar o problema da elección do punto de corte é enrolar o histograma linear arredor da circunferencia. Deste xeito, as barras estarán centradas no punto medio do intervalo do seu grupo, e, xa que dimanan do histograma linear, a súa altura vai ser proporcional á frecuencia relativa (isto é, a cantidade de datos) do grupo.



(a) Histograma angular da dirección de voo das pombas.

(b) Histograma angular, con menor número de caixas.

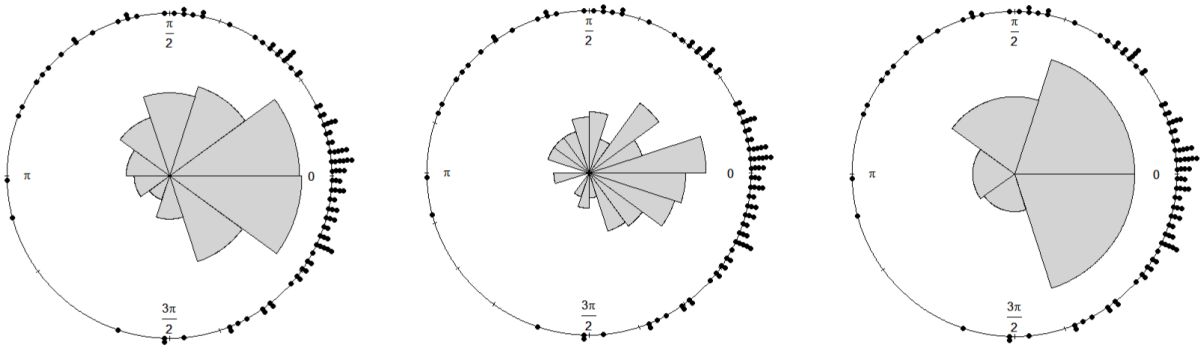
Figura 1.8: Histogramas angulares da dirección de voo das pombas.

En xeral, os histogramas (tanto lineares como angulares) son aparellos de representación proveitosos, en tanto que son simples de construír e poden aportar moita información sobre o comportamento da mostra. Porén, xa que exhiben formas moi dispares segundo a elección que fagamos da localización do punto de corte, ou dos extremos e o ancho dos intervalos, poden levarnos a conclusións erradas.

1.2.3. Diagramas de rosa

Semellantes aos histogramas, os diagramas de rosa, en troques de barras, representan as frecuencias relativas mediante áreas de sectores circulares. A raíz disto, non hai relación linear entre o raio dunha sección e a súa área, ao contrario ca no caso dos histogramas. Con vistas a solucionar esta traba, ímonos servir dunha das seguintes convencións no uso dos diagramas: ou ben tomamos o raio do sector como a raíz cadrada da frecuencia relativa, de xeito que o ratio existente entre as áreas dos sectores é o mesmo que entre as frecuencias; ou ben establecemos unha relación linear entre o raio e a frecuencia, de xeito que podemos comparar directamente entre os raios dos sectores. O habitual é facer uso da primeira das dúas convencións, que é a que temos usado nas representacións das direccións de voo das pombas na figura 1.9.

Malia que os diagramas de rosa semellan ser máis axeitados ca os histogramas para os datos circulares, comparten cos últimos unha contrariedade capital: a elección do número de seccións. En xeral, un criterio aconsellable é o de tomar a ráiz cadrada do tamaño mostral n como o tal



(a) Diagrama de rosa con $\sqrt{n}$ caixas.

(b) Diagrama de rosa con moitas caixas.

(c) Diagrama de rosa con poucas caixas.

Figura 1.9: Diagramas de rosa para as direccións de voo das pombas.

número, en pos de ter a mellor representación posible do comportamento das observacións.

1.2.4. Estimadores non paramétricos da densidade

Outra ferramenta gráfica importante á hora de poder sacar conclusións respecto da mostra é a de obter un estimador non paramétrico da distribución poboacional subxacente; artefacto que decontado exporemos, alicerzándonos sobre o traballo de Oliveira et al. (2012) [16].

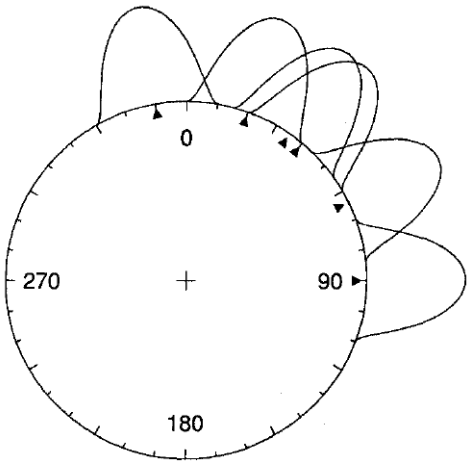
Se optamos por un estimador de densidade de tipo núcleo, o procedemento que se implementa é o dunha media en movemento, é dicir, imos repartir a contribución de cada dato (que é de $\frac{1}{n}$, sendo n o tamaño da mostra) nun arco relativamente exiguo que conteña á tal observación. Deste xeito, a estimación da densidade nunha dirección θ é a suma das contribucións dos puntos esparexidos preto dela, de xeito que a súa influencia queda distribuída nunha veciñanza. Formalmente, esta idea a podemos expresar como atopar un estimador da densidade circular subxacente $f(\theta)$,

$$\hat{f}(\theta) = \frac{1}{n} \sum_{i=1}^{n} \mathcal{K}_\nu \left(\theta - \theta_i\right). \tag{1.8}$$

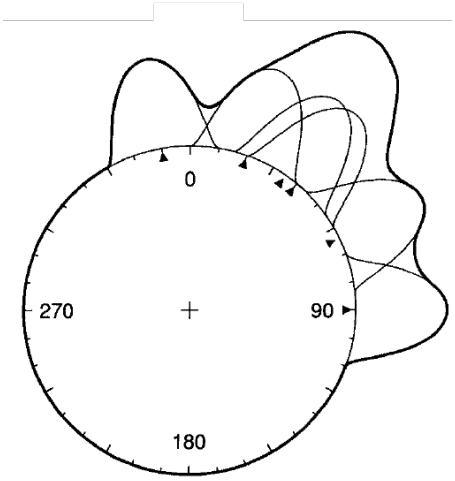
En resumo, o método baséase en facer copias idénticas do núcleo ou kernel elixido centradas sobre cada observación da mostra, deseguido sumar os valores que toman neses puntos, e promedialos por n, tendo tamén en conta un parámetro de concentración $\nu > 0$.

Na expresión (1.8), $\mathcal{K}_\nu$ é a función de núcleo circular, unha función de densidade [1], e ν é o parámetro de concentración ou suavizado.

[1] En xeral, utilizamos a distribución von Mises, sección 2.2.3, como función núcleo.



(a) Repartimos a contribución de cada dato nunha vecindade usando unha distribución.



(b) Sumamos as contribucións anteriores para obter o estimador da densidade.

Figura 1.10: Estimación non paramétrica da densidade usando un estimador de tipo kernel. Imaxes extraídas de *Fisher, (2003)*, Capítulo 2 [6].

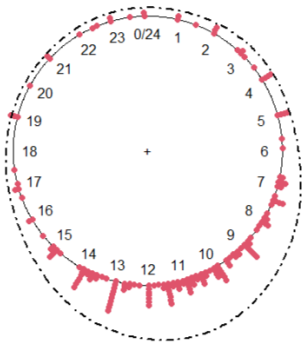
Unha boa elección do parámetro de concentración ha de ter en conta tanto o tamaño da mostra como a súa dispersión. Así, se hai moitos datos, imos poder ser quen de representar con maior miudeza; en troques, se a mostra fora máis pequena, teriamos que traballar inferindo de xeito máis xeral. Doutra banda, temos que reparar en que se os datos están concentrados nun arco estamos traballar en distinta escala ca se están repartidos en toda a circunferencia. Por exemplo, se consideramos a distribución vonMises como núcleo circular, o estimador resultante pódese expresar coma

$$\hat{f}(\theta) = \frac{1}{2\pi I_0(\nu)n} \sum_{i=1}^{n} e^{\nu \cos(\theta - \theta_i)}.$$

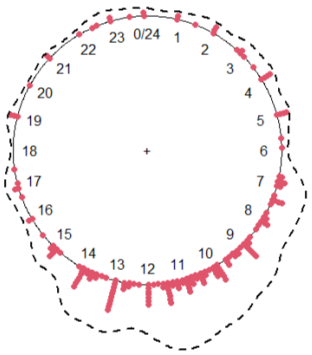
Aquí, como xa adiantabamos, a elección do parámetro de concentración ou suavizado ν tense que facer con sumo receo: valores grandes de ν dan lugar a estimadores con moita variación (infrasuavizados), pero valores miúdos carrexan estimadores demasiado suaves (sobresuavizados).

Para poder elixir o parámetro de suavizado que produza a mellor estimación da densidade subxacente téñense elaborado distintos métodos, se ben neste traballo limitámonos a facer probas con distintos valores do parámetro ata atopar o máis adecuado.

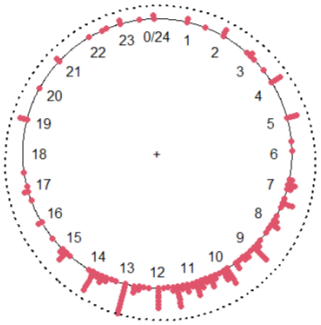
Se voltamos ao exemplo dos cambios nos ciclos de temperatura (en particular, os cambios a desxeo) a nivel do chan na zona periglacial de Monte Alvear, podemos procurar representar unha estimación non paramétrica da súa densidade. Se eliximos un parámetro de suavizado ν



(a) Estimación con parámetro de concentración ν máis adecuado.



(b) Estimación con parámetro ν grande, infrasuavizado.



(c) Estimación con parámetro ν pequeno, sobresuavizado.

Figura 1.11: Estimación non paramétrica da densidade do cambio de ciclo a desxeo en `cycle.changes`.

grande en demasía, atoparémonos (figura 1.11b) con que a estimación obtida é moi variable e se achega moito ás observacións, de xeito que non é frutífera para colixir que comportamento ten a mostra. Pola contra, se seleccionamos un ν miúdo, podemos chegar a suavizar tanto a estimación (figura 1.11c) que a representación que acadamos non achega ningunha información sobre os datos (de feito, achégase a una Uniforme Circular, sección 2.2.1). Así, probar con valores medios de μ permítenos colixir (figura 1.11a) que os cambios de ciclo a desxeo están concentrados principalmente nas horas de maior claridade (preto do mediodía) nunha mostra que semella ter un comportamento esencialmente unimodal.

Capítulo 2

Principais modelos de distribucións circulares

Neste capítulo introduciremos as distribucións máis clásicas para variables circulares, sustentándonos nas publicacións de [6], Capítulo 3, [19], Capítulo 4 e [9], Capítulo 2.

2.1. Conceptos básicos

Unha distribución circular é unha distribución de probabilidade cuxa probabilidade total está condensada na circunferencia unidade; e que serve para asignar probabilidades ás distintas direccións. Para tales distribucións, podemos definir a súa función de distribución acumulativa coma segue:

Definición 2.1. Unha función de distribución acumulativa dunha variable circular θ é unha función $F(\theta) = \mathbb{P}(0 < \Theta \leq \theta)$ verificando:

1. $F(\theta)$ é non decrecente e continua pola dereita.
2. $F(0) = 0$ e $F(2\pi) = 1$.
3. $F(\theta + 2\pi) = F(\theta) + 1$ para $-\infty < \theta < \infty$.

Coma no caso linear, as distribucións circulares son, esencialmente, de dúas clases:

- Discretas: asígnanse masas de probabilidade a un número contable de direccións.
- Absolutamente continuas (con respecto á medida de Lebesgue na circunferencia). Para estas últimas, existe función de densidade, que denotamos por $f(\theta)$.

Definición 2.2. Unha función de densidade dunha variable circular θ verifica:

1. $f(\theta) \geq 0$.
2. $\int_0^{2\pi} f(\theta)d\theta = 1$.
3. $f(\theta) = f(\theta + 2\pi k)$, $k \in \mathbb{Z}$ (f é periódica).

De modo análogo ao caso real linear, unha distribución circular pódese determinar univocamente mediante a súa función característica.

2.1.1. Función Característica

Definición 2.3. Para unha variable circular θ, definimos a súa función característica coma

$$\varphi_\theta(t) = \mathbb{E}[e^{it\theta}], \quad t \in \mathbb{R}.$$

Notemos que, como queira que θ é unha variable aleatoria periódica, esta función característica presenta a seguinte propiedade:

$$\varphi_\theta(t) = \mathbb{E}\left[e^{\mathbf{i}t\theta}\right] = \mathbb{E}\left[e^{\mathbf{i}t(\theta+2\pi)}\right] = e^{\mathbf{i}t2\pi}\varphi_\theta(t) \Rightarrow \begin{cases} \varphi_\theta(t) = 0 \\ \text{ou} \\ e^{\mathbf{i}t2\pi} = 1 \end{cases} \Leftrightarrow t \in \mathbb{Z}.$$

Logo a función característica só pode ser definida en valores enteiros.

Definición 2.4. O valor da función caracteristca en $p \in \mathbb{Z}$ é o momento trigonométrico $p-$ésimo da variable aleatoria circular continua Θ,

$$\begin{aligned} \varphi_\theta(p) := \mathbb{E}[e^{\mathbf{i}p\theta}] = \int_0^{2\pi} e^{\mathbf{i}p\theta} dF(\theta) = \int_0^{2\pi} e^{\mathbf{i}p\theta} f(\theta)d\theta = \int_0^{2\pi} \cos(p\theta)f(\theta)d\theta + \\ + \mathbf{i}\int_0^{2\pi} \operatorname{sen}(p\theta)f(\theta)d\theta = \mathbb{E}[\cos(p\theta)] + \mathbf{i}\mathbb{E}[\operatorname{sen}(p\theta)], \end{aligned} \tag{2.1}$$

onde $F(\theta)$, $f(\theta)$ son as funcións de distribución e densidade, respectivamente. A partires deste punto, e dada esta definición, denotaremos $\varphi_\theta(p) = \varphi_p$.

Estes mometos trigonométricos podemos expresalos en termos de

$$\begin{aligned} \alpha_p = \mathbb{E}[\cos(p\theta)], \qquad \beta_p = \mathbb{E}[\operatorname{sen}(p\theta)]; \\ \rho_p = \sqrt{\alpha_p^2 + \beta_p^2} = |\varphi_p|, \quad \mu_p = \operatorname{atan2}(\beta_p, \alpha_p), \end{aligned} \qquad p \in \mathbb{Z}.$$

Deste xeito, $\varphi_p = \alpha_p + \mathbf{i}\beta_p = \rho_p e^{\mathbf{i}\mu_p}$.

Observación 2.5. Advirtamos que $0 < \rho_p < 1$, xa que $\rho_p = |\sigma_p| = \|\mathbb{E}[e^{\mathbf{i}p\theta}]\| \leq \mathbb{E}\left[\|e^{\mathbf{i}p\theta}\|\right] = 1$.

O primeiro momento trigonómetrico, $\varphi_1 = \alpha_1 + \mathbf{i}\beta_1 = \rho_1 e^{\mathbf{i}\mu_1}$ xoga un papel destacado, xa que a lonxitude ρ_1 e a dirección μ_1 (que denotaremos por ρ e μ, respectivamente) empréganse como medidas teóricas ou poboacionais da concentración e da dirección media de θ, xogando o papel análogo das medidas empíricas $\bar{R} = \frac{R}{n}$ e $\bar{\theta}$. Así, canto máis próximo ρ sexa de 1, maior será a concentración cara a dirección media μ, e se e $\rho = 0$ entón non imos poder determinar μ.

De xeito parello ao caso real linear, tamén podemos definir os momentos trigonométricos centrados. Definimos o momento trigonométrico $p-$ésimo respecto da dirección media μ coma $\bar{\varphi}_p = \bar{\alpha}_p + \mathbf{i}\bar{\beta}_p$, onde

$$\bar{\alpha}_p = \mathbb{E}\left\{\cos[p(\theta - \mu)]\right\}, \quad \bar{\beta}_p = \mathbb{E}\left\{\operatorname{sen}[p(\theta - \mu)]\right\}. \tag{2.2}$$

Observación 2.6. Os momentos trigonométricos φ_p son momentos respecto da dirección nula.

Para os momentos trigonométricos (respecto da dirección nula ou da media) existen parellos mostrais,

$$a_p = \frac{1}{n}\sum_{i=1}^{n}\cos(p\theta_i), \qquad b_p = \frac{1}{n}\sum_{i=1}^{n}\operatorname{sen}(p\theta_i), \text{ e}$$
$$\bar{a}_p = \frac{1}{n}\sum_{i=1}^{n}\cos[p(\theta_i - \bar{\theta})], \quad \bar{b}_p = \frac{1}{n}\sum_{i=1}^{n}\operatorname{sen}[p(\theta_i - \bar{\theta})],$$

respectivamente. Manexar estes conceptos será produtivo en vindeiros capítulos.

Doutra banda, estes momentos trigonométricos de θ coinciden cos coeficientes da expansión en serie de Fourier da función de densidade de $f(\theta)$, de modo que se a derradeira é de cadrado integrable en $[0, 2\pi)$, podemos recuperar a función de densidade a partir dos coeficientes de Fourier, mediante a expansión

$$f(\theta) = \frac{1}{2\pi}\sum_{p=-\infty}^{\infty}\varphi_p e^{-\mathbf{i}p\theta} = \frac{1}{2\pi}\left\{1 + 2\sum_{p=1}^{\infty}[\alpha_p\cos(p\theta) + \beta_p\operatorname{sen}(p\theta)]\right\}. \tag{2.3}$$

2.1.2. Medidas poboacionais básicas

Ao longo deste capítulo faremos uso das medidas poboacionais, parellas ás medidas mostrais que temos exposto con anterioridade.

Definición 2.7 (Medidas de localización teóricas)**.** Manexaremos tres medidas de localización:

- Como xa adiantabamos anteriormente, a dirección media poboacional é $\mu = \mu_1 = \operatorname{atan2}(\beta_1, \alpha_1)$.

- A dirección mediana teórica defínese coma $\tilde{\mu} = \arg\min_{\phi} \mathbb{E}[\pi - |\pi - |\Theta - \phi||]$.
- A dirección modal poboacional, $\check{\mu}$, identifícase coma a dirección de máxima probabilidade no caso discreto, e como a dirección tal que a representación polar de $f(\theta)$ é máxima no caso continuo.

Observación 2.8. Fagamos notar que, tanto $\tilde{\mu}$ coma $\check{\mu}$ non teñen por que ser únicas.

Definición 2.9 (Medidas de dispersión teóricas.)**.** Os análogos teóricos para as medidas de dispersión expostas no Capítulo 1 son

- Lonxitude resultante media poboacional $\rho = \rho_1 \in [0, 1]$.
- Varianza circular poboacional $\nu = 1 - \rho \in [0, 1]$.
- Desviación Típica poboacional $\sigma = \sqrt{-2\log(1-\nu)} \in [0, \infty)$.

Outro concepto que vainos a ser de extrema utilidade é o de distribucións simétricas.

Definición 2.10. Diremos que unha distribución é simétrica (por reflexión) se existe un único eixo tal que a reflexión da distribución respecto del é idéntica á orixinal.

Observación 2.11. O segundo momento trigonométrico respecto da media mostral, $\bar{b}_2$ (2.2), dá conta da asimetría dunha mostra unimodal: para valores próximos a cero, a mostra ha de presentar un comportamento simétrico respecto de $\bar{\theta}$.

Proposición 2.12. *Se unha distribución é simétrica respecto de ψ, tamén o será respecto de $\psi + \pi$. A maiores, ocorre que $f(\theta - \psi) = f(\psi - \theta)$, $0 \leq \theta < 2\pi$.*

Proposición 2.13. *Se unha distribución é simétrica e unimodal, entón $\mu = \tilde{\mu} = \check{\mu}$.*

Pode ocorrer que unha distribución multimodal sexa simétrica respecto de varios (poñamos l) eixos, de xeito que a rotación en $\frac{2\pi}{l}$ da distribución sexa igual ca orixinal. Neste caso, diremos que é simétrica en l pregos. No caso en que $l = 2$, diremos que a simetría é antipodal.

2.1.3. Obtención de distribucións circulares

Para poder obter distribucións circulares, os métodos máis xeneralizados son:

- Construír a distribución específicamente para variables circulares (distribucións circulares propias).

- Enrolar unha distribución lineal arredor da circunferencia unidade: se X é unha variable aleatoria linear, pódese transformar nunha circular sen máis ca considerar $\Theta = X(mod\, 2\pi)$. Daquela, se $g(x)$ designa a densidade linear e $f(\theta)$ a circular, a posterior pódese construír acumulando probabilidade sobre os puntos superpostos:

$$f(\theta) = \sum_{m=-\infty}^{\infty} g(\theta + 2\pi m), \quad 0 \leq \theta < 2\pi.$$

  O seguinte resultado vainos permitir, a partir da función característica de X, representar a función de densidade dunha distribución enrolada.

  Proposición 2.14. *Se φ_p é o momento trigonométrico de orden p dunha variable aleatoria circular Θ, e $\phi_X(t)$ a función característica da variable linear X, entón $\varphi_p = \phi_X(p)$.*

  Demostración. Sexan $F(\theta), G$ as funcións de distribución de Θ e X. Entón

$$\varphi_p = \int_0^{2\pi} e^{\mathbf{i}p\theta} dF_\Theta(\theta) = \sum_{k=-\infty}^{\infty} \int_{2\pi k}^{2\pi(k+1)} e^{\mathbf{i}p\theta} dG_X(\theta) = \int_{-\infty}^{\infty} e^{\mathbf{i}px} dG_X(x) = \phi_X(p).$$

  □

- Transformar un vector aleatorio bidimensional nas súas compoñentes direccionais (resultando nas chamadas distribucións *offset*). A transformación do vector (X, Y) en (R, Θ) faise mediante o cambio por coordenadas polares, e deseguida obtemos a densidade circular $f(\theta)$ a partir da densidade conxunta $g(x, y)$, facendo:

$$f(\theta) = \int_0^{\infty} r g(r\cos\theta, r\,\mathrm{sen}\,\theta) dr \ .$$

- Empregar a proxección estereográfica, identificando os puntos $x \in \mathbb{R}$ cos da circunferencia de xeito inxectivo agás por $+\infty, -\infty$, que se fan corresponder con π.

Trataremos a continuación as distribucións máis importantes nadas da aplicación destes métodos.

2.2. Distribucións Circulares Propias

2.2.1. Distribución Uniforme CU

Diremos que unha variable aleatoria circular discreta Θ promana dunha distribución uniforme circular de m puntos, $\Theta \in CU(m, \xi)$ cando a distribución da súa probabilidade toma a seguinte expresión

$$\mathbb{P}\left(\Theta = \xi + \frac{2\pi q}{m}\right) = \frac{1}{m}, \quad q = 0, 1, \ldots m-1. \tag{2.4}$$

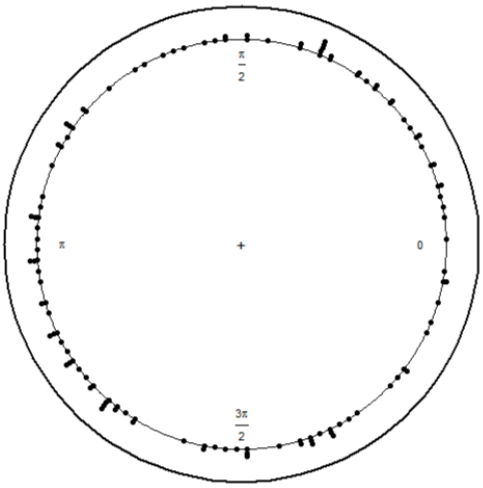
Se $m = 1$, (2.4) define unha distribución dexenerada nun punto, onde toda a masa de probabilidade queda amoreada nunha dirección, ξ; de modo que $\mu = \tilde{\mu} = \check{\mu} \equiv \xi$. No caso de que $m > 1$, (2.4) asigna probabilidades idénticas de $\frac{1}{m}$ a m puntos equivariantes sobre a circunferencia. Consonte a isto, a dirección media neste caso non estará definida, e a mediana $\tilde{\mu}$ e a moda $\check{\mu}$ non serán únicas.

No caso continuo, a probabilidade total queda repartida de xeito uniforme na circunferencia unidade, de xeito que ningunha dirección é máis probable que calquera outra. As funcións de densidade e distribución correspondentes materialízanse coma

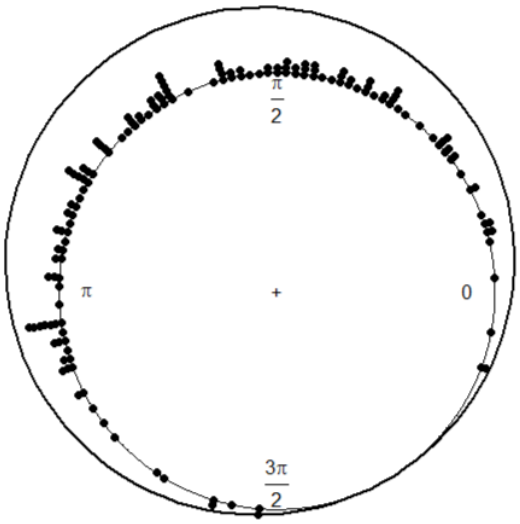
$$f(\theta) = \frac{1}{2\pi}, \quad 0 \leq \theta < 2\pi, \tag{2.5}$$

$$F(\theta) = \frac{\theta}{2\pi}, \quad 0 \leq \theta < 2\pi. \tag{2.6}$$

Observemos que esta distribución non ten unha dirección media ben definida, por causa de $\varphi_1 = \rho e^{\mathbf{i}\mu} = 0$, logo $\rho = 0$ e conforme a isto, μ está indefinida. De xeito similar, non é posible precisarmos unha (ou varias) direccións modais ou medianas baixo a distribución uniforme, daquela, quedan tamén indefinidas.



(a) Distribución uniforme continua circular para unha mostra aleatoria de 100 datos.



(b) Distribución cardioide de parámetros $\rho = \frac{1}{2}$ e dirección media $\mu = \frac{2}{3}\pi$.

Figura 2.1: Representación das funcións de densidade das distribucións circulares uniforme e cardioide, respectivamente.

2.2.2. Distribución Cardioide $C(\mu, \rho)$

A partir da curva cardioide derivamos a devandita distribución, que ten por función de densidade

$$f_{\mu,\rho}(\theta) = \frac{1}{2\pi}\left[1 + 2\rho\cos(\theta - \mu)\right], \quad 0 \leq \theta < 2\pi,$$

onde o parámetro $\rho \in \left[-\frac{1}{2}, \frac{1}{2}\right]$ denota a concentración e μ, $0 \leq \mu < 2\pi$, a dirección media, de xeito que esta distribución é unimodal (hai unha única dirección media ou moda) e simétrica respecto de μ.

2.2.3. Distribución de von Mises $vM(\mu, \kappa)$

Unha variable aleatoria circular Θ segue unha distribución de von Mises de parámetos $0 \leq \mu < 2\pi$ e $\kappa \geq 0$ se a súa función de densidade toma a formulación

$$f_{\mu,\kappa}(\theta) = \frac{e^{\kappa\cos(\theta-\mu)}}{2\pi I_0(\kappa)}, \quad 0 \leq \theta < 2\pi. \tag{2.7}$$

Na expresión anterior, $I_0(k)$ é a función de Bessel modificada de primeira especie e orde cero,

$$I_0(\kappa) = \frac{1}{2\pi}\int_0^2 \pi e^{\kappa\cos\theta} d\theta = \sum_{r=0}^{\infty}\left(\frac{\kappa}{2}\right)^{2r}\left(\frac{1}{r!}\right)^2.$$

Esta distribución é unha das máis relevantes entre as fraguadas para variables circulares, por mor de entrañar un compendio de propiedades de sumo interese, que introduciremos seguidamente.

O primeiro que imos facer notar é a simetría desta distribución respecto de μ, e, como consecuencia da simetría do coseno, de $\mu + \pi$. Ademais de xogar o papel de dirección media, a moda para esta distribución materialízase na dirección μ: posto que o coseno acada un máximo en cero, a densidade dunha normal circular é máxima en $\theta = \mu$, ou o que é o mesmo, μ é unha dirección modal con valor máximo

$$f_{\mu,\kappa}(\mu) = \frac{e^{\kappa}}{2\pi I_0(\kappa)}. \tag{2.8}$$

Seguindo este fío de razoamento respecto das funcións trigonométricas, é claro que, como queira que o coseno abrangue un mínimo en π, se $\theta = \mu \pm \pi$ obtemos o valor mínimo da densidade,

$$f_{\mu,\kappa}(\mu \pm \pi) = \frac{e^{-\kappa}}{2\pi I_0(\kappa)}, \tag{2.9}$$

logo $\mu \pm \pi$ ha de ser a dirección antimodal.

Reparemos arestora no papel do parámetro κ, xa que logo, a partir das propiedades anteriores (2.8), (2.9), inferimos que

$$\frac{f_{\mu,\kappa}(\mu)}{f_{\mu,\kappa}(\mu \pm \pi)} = e^{2\kappa}.$$

Por conseguinte, κ é un parámetro que dá conta da concentración cara a dirección media μ: en efecto, canto maior sexa o valor de κ, maior é a proporción entre $f_{\mu,\kappa}(\mu)$ e $f_{\mu,\kappa}(\mu \pm \pi)$, indicando unha maior concentración cara μ.

Proposición 2.15 (Aproximación por unha distribución normal linear estándar)**.** *Cando facemos tender κ a infinito, podemos asegurar a converxencia en distribución $\sqrt{\kappa}(\theta - \mu) \xrightarrow[d]{\kappa \to \infty} N(0,1)$.*

Demostración. Lembremos que a distribución de von Mises ten densidade (2.7),

$$f_{\mu,\kappa}(\theta) = \frac{e^{k\cos(\theta-\mu)}}{2\pi I_0(\kappa)}, \quad 0 \leq \theta < 2\pi.$$

Denotaremos $\varepsilon = \sqrt{\kappa}(\theta - \mu)$. Serános útil facer uso da expansión en serie de Taylor da función coseno,

$$\cos\theta = \sum_{n=0}^{\infty} \frac{(-1)^n}{(2n)!}\theta^{2n} = 1 - \frac{\theta^2}{2} + \frac{\theta^4}{4!} - \frac{\theta^6}{6!} + \cdots,$$

de xeito que podemos aproximar $\cos\theta \simeq 1 - \frac{\theta^2}{2}$ e así, para κ suficientemente grande

$$\cos(\theta - \mu) = \cos\left(\frac{\varepsilon}{\sqrt{\kappa}}\right) \simeq 1 - \frac{\varepsilon^2}{2\kappa}.$$

Amais disto, contamos coa aproximación $I_0(\kappa) \simeq \frac{e^\kappa}{\sqrt{2\pi\kappa}}$. Logo facendo uso da fórmula de cambio de variable, podemos finalmente inferir que a función de densidade de ε, g, é semellante á dunha normal estándar:

$$g(\varepsilon) = \frac{e^{\kappa\cos\left(\frac{\varepsilon}{\sqrt{\kappa}}\right)}}{2\pi I_0(\kappa)}\frac{1}{\sqrt{\kappa}} \simeq \frac{e^{\kappa\cos\left(\frac{\varepsilon}{\sqrt{\kappa}}\right)}}{2\pi\frac{e^k}{\sqrt{2\pi k}}}\frac{1}{\sqrt{\kappa}} \simeq \frac{e^{\kappa\left(1-\frac{\varepsilon^2}{2\kappa}\right)}}{\sqrt{2\pi}e^\kappa} = \frac{e^{-\frac{\varepsilon^2}{2}}}{\sqrt{2\pi}}$$

□

Proposición 2.16. *Se $\Theta \in vM(\mu,\kappa)$ e $\kappa = 0$ entón $\Theta \in CU$. Doutra banda, cando $\kappa \to \infty$, $vM(\mu,\kappa)$ tende a unha distribución dexenerada nun punto na dirección μ.*

Vimos no capítulo anterior como $\bar{\theta} = \operatorname{atan2}(S,C)$ fornécenos de xeito axuizado dunha dirección media mostral. Vexamos que $\bar{\theta}$ é un estimador de máxima verosimilitude da dirección media poboacional μ, propiedade que a von Mises comparte coa distribución Normal linear, que é a única distribución para datos reais na que a media mostral tamén é o estimador de máxima verosimilitude para a media poboacional. Consonte a isto, é común atopar na literatura sobre estatística circular que a von Mises é designada indistintamente por Normal Circular.

Proposición 2.17. *A media mostral $\bar{\theta}$ é o estimador de máxima verosimilitude da dirección media μ dunha distribución von Mises $vM(\mu, \kappa)$.*

Demostración. Sexan $\theta_1, \theta_2, \ldots, \theta_n$ un conxunto de observacións procedentes dunha densidade $f(\theta - \mu)$. A verosimilitude vén dada por

$$L = \prod_{i=1}^{n} f(\theta_i - \mu),$$

e a log-verosimilitude

$$\log L = \log \left[\prod_{i=1}^{n} f(\theta_i - \mu) \right] = \sum_{i=1}^{n} \log[f(\theta_i - \mu)].$$

Derivando e igualando a cero, obtemos a ecuación de verosimilitude,

$$\frac{\partial}{\partial \mu} \log L = \sum_{i=1}^{n} \frac{f'(\theta_i - \mu)}{f(\theta_i - \mu)} = 0. \tag{2.10}$$

Doutra banda, se $\bar{\theta}$ estima μ, temos que

$$\sum_{i=1}^{n} \operatorname{sen}(\theta_i - \mu) = 0. \tag{2.11}$$

Visto que (2.11), (2.10) hanse de cumprir para θ_i e n arbitrarios, a igualdade debe manterse termo a termo e, por conseguinte, para algunha constante c_1,

$$\frac{f'(\theta - \mu)}{f(\theta - \mu)} = c_1 \operatorname{sen}(\theta - \mu).$$

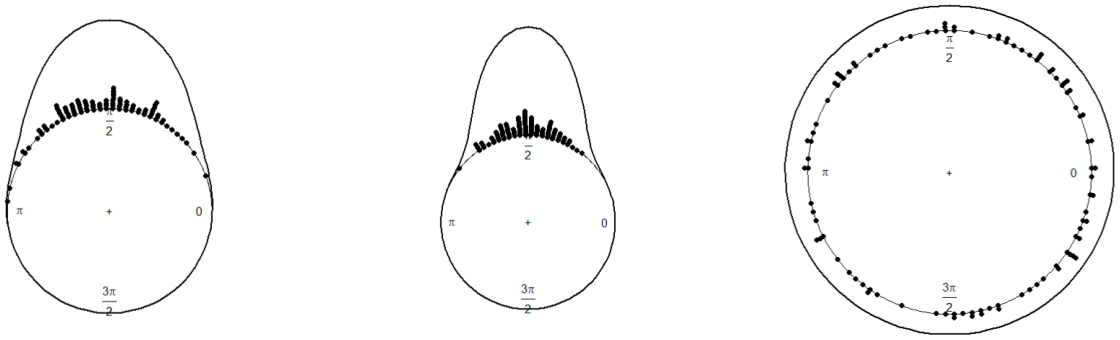
Coliximos que $f(\theta - \mu) = c_2 \cdot e^{c_1 \cos(\theta - \mu)}$, é dicir, $f(\theta - \mu)$ é a densidade de von Mises (2.7). □

2.3. Distribucións Circulares Enroladas

2.3.1. Distribución Normal Enrolada $WN(\mu, \rho)$

Esta distribución é nada de enrolar a distribución normal linear $N(\mu, \sigma^2)$ arredor da circunferencia unidade, aplicando o procedemento exposto en seccións pretéritas deste capítulo, de modo que a expresión da súa función de densidade é

$$f(\theta) = \sum_{m=-\infty}^{\infty} g(\theta + 2\pi m) = \frac{1}{\sigma\sqrt{2\pi}} \sum_{m=-\infty}^{\infty} e^{-\frac{(\theta - \mu - 2\pi m)^2}{2\sigma^2}}. \tag{2.12}$$



(a) $\kappa = 5$ (b) $\kappa = 10$ (c) $\kappa = 0.1$

Figura 2.2: Representación das funcións de densidade dunha von Mises con paráetros $\mu = \frac{\pi}{2}$ e $\kappa \in \{0.1,\ 5,\ 10\}$.

Outra expresión para (2.12) é posible, a partir da función característica de $X \in N(\mu, \sigma^2)$:

$$\phi_X(p) = e^{\mathbf{i}\mu p - \frac{\sigma^2 p^2}{2}} = e^{\mathbf{i}\mu p}\rho^{p^2} = \varphi_p = \alpha_p + \mathbf{i}\beta_p$$

onde $\rho = e^{-\frac{\sigma^2}{2}}$ e a penúltima igualdade é consecuencia da Proposición 2.14 deste capítulo. Así, $\alpha_p = \rho^{p^2}\cos(\mu p)$ e $\beta_p = \rho^{p^2}\operatorname{sen}(\mu p)$, resultando en que, consonte a expansión en serie de Fourier (2.3),

$$f(\theta) = \frac{1}{2\pi}\left\{1 + 2\sum_{p=1}^{\infty}\rho^{p^2}\cos[p(\theta-\mu)]\right\}.$$

Podemos usar esta expresión para aproximar a densidade mediante os primeiros termos da serie.

Proposición 2.18. *A distribución normal enrolada posúe a propiedade aditiva, i.e., se* $\Theta_1 \in WN(\mu_1, \rho_1)$, $\Theta_2 \in WN(\mu_2, \rho_2)$ *son independentes, entón* $\Theta_1 + \Theta_2 \in WN(\mu_1 + \mu_2, \rho_1\rho_2)$.

Demostración. Visto que $\theta_i = X_i(\mod 2\pi)$, con $X_i \in N(\mu_i, \sigma_i^2)$, $i = 1, 2$,

$$\theta_1 + \theta_2 = X_1(mod\ 2\pi) + X_2(mod\ 2\pi) = (X_1 + X_2)(mod\ 2\pi).$$

Amais, por mor da independencia, $X_1 + X_2 \in N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$, co seu pertinente parámetro de concentración

$$e^{-\frac{\sigma_1^2+\sigma_2^2}{2}} = e^{-\frac{\sigma_1^2}{2}}e^{-\frac{\sigma_2^2}{2}} = \rho_1\rho_2.$$

□

Observación 2.19. A idea que subxace a construcción desta distribución é a que motiva a definición da desviación típica poboacional $\sigma = \sqrt{-2\log(1-\nu)}$, por mor de que o parámetro de concentración para a Normal Enrolada é $\rho = e^{-\frac{\sigma^2}{2}}$, de xeito que $1 - \nu = e^{-\frac{\sigma^2}{2}}$.

2.3.2. Distribución de Cauchy Enrolada $WC(\mu, \rho)$

A distribución de Cauchy no caso linear ten densidade

$$g(x) = \frac{1}{\pi} \cdot \frac{\sigma}{\sigma^2 + (x-\mu)}, \quad -\infty < x < \infty.$$

Obtemos a análoga para o circular enrolando esta última arredor da circunferencia unidade, de xeito que a función de densidade dunha Cauchy Enrolada é

$$f(\theta) = \sum_{m=-\infty}^{\infty} \frac{1}{\pi} \cdot \frac{\sigma}{\sigma^2 + (\theta + 2\pi m - \mu)}, \quad 0 \leq \theta < 2\pi.$$

Como queira que a función característica da distribución de Cauchy é $\phi_X(p) = e^{-\sigma p + \mathbf{i}p\mu} = \rho^p e^{\mathbf{i}p\mu}$, con $\rho = e^{-\sigma}$, os momentos trigonométricos da distribución toman corpo coma

$$\varphi_p = \alpha_p + \mathbf{i}\beta_p = \rho^p e^{\mathbf{i}p\mu} \Rightarrow \begin{cases} \alpha_p = \rho^p \cos(\mu p) \\ \beta_p = \rho^p \operatorname{sen}(\mu p) \end{cases}.$$

En consecuencia, podemos expresar (2.3.2) mediante a súa expansión en serie de Fourier,

$$f(\theta) = \frac{1}{2\pi}\left\{1 + 2\sum_{p=1}^{\infty}[\rho^p \cos(p\mu)\cos(p\theta) + \rho^p \operatorname{sen}(p\mu)\operatorname{sen}(p\theta)]\right\} = \frac{1}{2\pi}\left\{1 + 2\sum_{p=1}^{\infty}\rho^p \cos[p(\theta-\mu)]\right\}.$$

Observemos deseguido que

$$\sum_{p=1}^{\infty}\rho^p \cos[p(\theta-\mu)] = \operatorname{Re}\left\{\sum_{p=1}^{\infty}\left[\rho e^{\mathbf{i}(\theta-\mu)}\right]^p\right\} = \operatorname{Re}\left(\frac{\rho e^{\mathbf{i}(\theta-\mu)}}{1-\rho e^{\mathbf{i}(\theta-\mu)}}\right) = \frac{\rho\cos(\theta-\mu)}{1-\rho\cos(\theta-\mu)}.$$

Atendendo entón a estas derradeiras expresións, a función de densidade para esta distribución pódese escribir de xeito parello coma

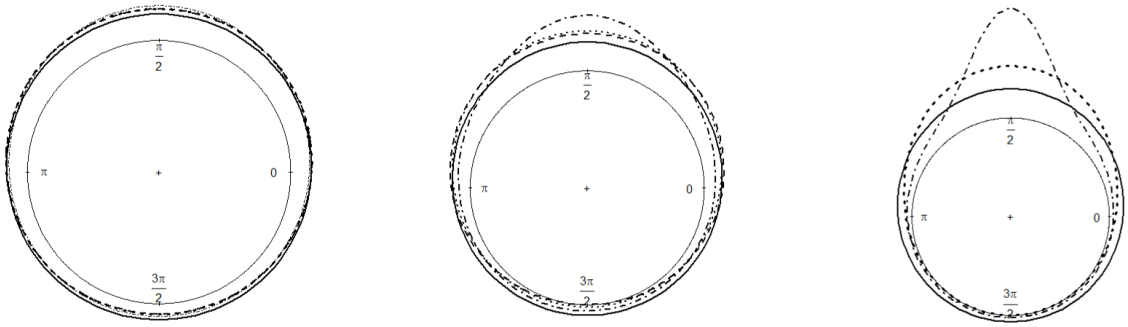
$$f(\theta) = \frac{1}{2\pi}\left(1 + 2\frac{\rho\cos(\theta-\mu)}{1-\rho\cos(\theta-\mu)}\right) = \frac{1}{2\pi}\frac{1-\rho^2}{1+\rho^2-2\rho\cos(\theta-\mu)}, \quad 0 \leq \theta < 2\pi.$$

Observación 2.20. É de proveito notar que a distribución de Cauchy enrolada é unimodal e simétrica, e ademais, posúe a propiedade aditiva 2.3.1.

2.3.3. Distribución Normal Asimétrica Enrolada $WSN(\mu, \rho, \lambda)$

Ata o de agora, temos exposto unha trama de distribucións, que, malia procederen de orixes dispares (algunhas son propias da circunferencia, outras agroman da recta real), comparten unha

[1] Na figura 2.3c non incluímos a distribución Cardioide, xa que $\rho > 0.5$.



(a) $\rho, \kappa = 0.25$ (b) $\rho, \kappa = 0.5$ (c) $\rho, \kappa = 0.75$

Figura 2.3: Comparativa entre as distribucións von Mises (liña sólida), Cardioide (liña discontinua), Normal enrolada (liña a puntos) e Cauchy enrolada (liña discontinua con puntos). Os parámetros toman valores $\mu = \frac{\pi}{2}$ e $\rho, \kappa \in \{0.25,\ 0.5,\ 0.75\}$[1].

particularidade: a simetría. Non obstante, na práctica é frutuoso poder contar con modelos asimétricos. En [20], Pewsey introduce o modelo de distribución normal asimétrica enrolada, partindo dunha variable aleatoria linear provida dunha distribución normal asimétrica con parámetro de asimetría $\lambda \in (-\infty, \infty)$, $X \in SN(\lambda)$, de xeito que a súa función de densidade é

$$g_\lambda(x) = 2\phi\,(x)\,\Phi(\lambda x), \quad -\infty < x < \infty.$$

Onde $\phi(x), \Phi(x)$ designan as funcións de densidade e distribución dunha normal estándar $N(0,1)$. Se transformamos a variable anterior en $Y = \mu + \rho x$, sendo μ un parámetro de localización e ρ un de escala, entón a densidade correspondente será

$$g_{\mu,\rho,\lambda}(y) = \frac{2}{\rho}\phi\left(\frac{y-\mu}{\rho}\right)\Phi\left(\lambda\frac{y-\mu}{\rho}\right), \quad -\infty < y < \infty.$$

Esta parametrización é a que nomearemos como parametrización directa, expresando $Y \in SN_D(\mu, \rho, \lambda)$. Alicerzándonos sobre o anterior, podemos enrolar a distribución $g_{\mu,\rho,\lambda}(y)$ arredor da circunferencia, sen máis que tomar $\Theta \equiv Y\ (mod 2\ \pi)$, obtendo a densidade

$$f_\lambda(\theta) = \frac{2}{\rho}\sum_{m=-\infty}^{\infty}\phi\left(\frac{\theta + 2\pi m - \mu}{\rho}\right)\Phi\left(\lambda\frac{\theta + 2\pi m - \mu}{\rho}\right), \quad 0 \le \theta < 2\pi. \tag{2.13}$$

Consonte a isto, diremos que unha variable aleatoria circular con densidade (2.13) ten distribución normal enrolada antisimétrica, $\Theta \in WSN(\mu, \rho, \lambda)$.

Proposición 2.21. *A distribución normal enrolada antisimétrica verifica:*

1. *Se o parámetro de escala tende a cero, $\rho \to 0$, $WSN(\mu, \rho, \lambda)$ converxe a unha distribución dexenerada nun punto.*

2. *Se o parámetro de escala tende a infinito, $\rho \to \infty$, $WSN(\mu, \rho, \lambda)$ converxe a unha distribución circular uniforme.*

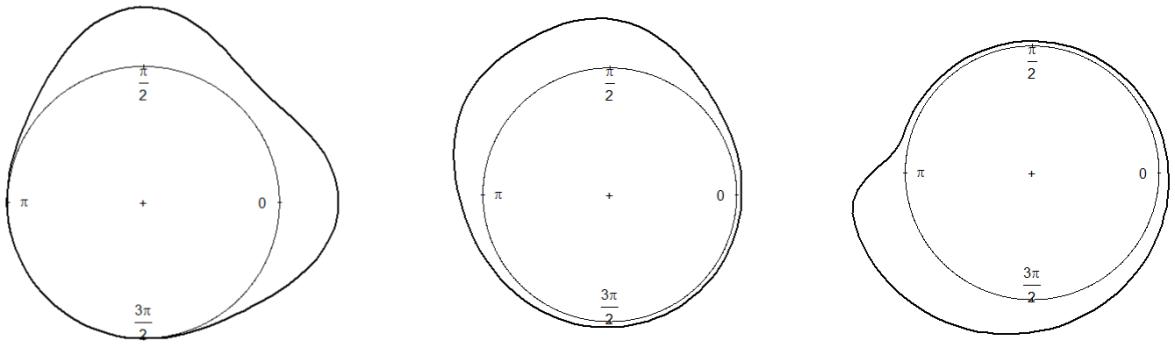
3. *Se o parámetro de asimetría é nulo, $\lambda = 0$, entón a expresión (2.13) é a da función de densidade da normal enrolada.*

2.4. Mesturas

Consideremos $f_{\mu,\rho}(\theta)$ unha función de densidade para unha variable aleatoria circular Θ unimodal con dirección media μ e ρ un parámetro de dispersión respecto de μ. Sexan $p_1, ..., p_K$ unha serie de proporcións, é dicir, $p_i > 0$, $k = 1, ... K$, e tales que $\sum_{k=1}^{K} p_k = 1$. Entón, a mestura de distribucións é a función de densidade

$$f^*(\theta) = \sum_{k=1}^{K} p_k f_{\mu_k,\rho_k}(\theta). \tag{2.14}$$

Observemos que, se $\mu_1 = \mu_2 = ... = \mu_K$, entón $f^*(\theta)$ é unimodal. Pola contra, se $\mu_k \neq \mu_l$ para algún k, l entón, en xeral, $f^*(\theta)$ ha de ser multimodal.



(a) $0.5vM(0, 5) + 0.5vM\left(\frac{\pi}{2}, 5\right)$ (b) $0.3vM\left(\frac{\pi}{2}, 5\right) + 0.7WC\left(\frac{4\pi}{5}, 0.5\right)$ (c) $0.4C(\frac{3\pi}{2}, 0.25) + 0.6WSN(\pi, 1.5)$

Figura 2.4: Mixturas de distribucións circulares.

Capítulo 3

Inferencia en distribucións circulares

Entre outros conceptos, ao longo deste capítulo imonos ocupar dos estimadores de parámetros γ, calculados a partires dunha mostra, aos que denotaremos por $\hat{\gamma}$. Ás mostras circulares sobre as que faremos inferencia esixímoslles que cada observación que as compoñe promane dunha mesma poboación, e que todas elas sexan mutuamente independentes. A unha mostra de tamaño n que verifique estas características nomearémola como mostra aleatoria de variables circulares, $\theta_1, ..., \theta_n$. Se os datos na mesma son independentes e están idénticamente distribuídos denotarémolos de igual modo, agás a inclusión das siglas i.i.d ao seu carón.
Ás veces, sobre unha mostra circular, precisaremos ordear as observacións de algún modo. A fin de establecer unha orde, primeiramente fixaremos a orixe e o sentido de rotación para a mostra, e decontado disporemos os datos seguindo unha ordeación que respecte a elección previa, $\theta_{(1)} \leq ... \leq \theta_{(n)}$.
Outros conceptos clave que é recomendable manexar previo a exposición dos métodos inferencias para variables circulares serán os de nivel de significación e valor crítico, así coma o de estatístico de contraste. Denotamos por $\alpha \in [0, 1]$ o nivel de significación dun contraste, isto é, a probabilidade de rexeitar a hipótese nula condicionado a que a devandita sexa certa. O nivel crítico P representa o mínimo nivel de significación baixo o que podemos rexeitar H_0, ou equivalentemente, o maior α para o que non podemos rexeitala. Intuitivamente, este é un valor que da conta do moito que os datos contradín a hipótese nula. Por último, o estatístico de contraste é unha operación que facemos sobre a mostra de xeito que retrate a compatibilidade das observacións con H_0.

Os contidos deste Capítulo estriban sobre diversas fontes, se ben primordialmente baseámonos nos traballos de Fisher [6], Jammalamadaka e Sen Gupta [9], Pewsey [19] e Mardia e Jupp [12].

3.1. Tests para uniformidade e simetría

Unha das probas preliminares máis importantes sobre a que alicerzar o estudo dunha mostra de datos circulares é a da comprobación de uniformidade. Para este tipo de contrastes, entre as posibles familias de tests, érguense sobranceiramente dúas: as dos test Omnibus e as dos test de Rayleigh. Os primeiros son de extrema utilidade cando o noso interese primordial é detectar calqueira outro modelo alternativo; e os segundos, cando queremos contrastar uniformidade fronte a un modelo unimodal. Baseándonos na labor de Batschelet, [3], Capítulo 4, introduciremos algúns modelos de ambos tipos de test.

Comezamos primeiramente por expor algúns tests Omnibus, en particular, o test de Hodge-Ajne e o test de espazado de Rao. Neles, contrastaremos a hipótese nula de uniformidade contra a alternativa de non uniformidade.
O test de Hodge-Ajne eríxese sobre a asunción de que ou ben os datos non están agrupados ou ben que o número de grupos é grande (respecto do tamaño mostral). A idea sobre a que se fundamenta é a de dividir o círculo en dous semicírculos mediante un diámetro l, que imos ir rotando ata que sexa quen de deixar a un lado o número máximo de observacións posible, e ao outro, o número mínimo, que denotaremos por A_n e que xogará o papel do noso estatístico para este test. O esperado se a mostra promana dunha distribución uniforme circular será que A_n sexa relativamente grande. Así, un criterio de decisión válido é rexeitar a hipótese nula se o P-valor asociado ao tamaño de mostra e ao valor do estatístico A_n é menor que o nivel de significación prefixado.

O test de espazado de Rao considera que, baixo suposición de uniformidade, as observacións circulares deberían corporizarse equidistantemente sobre a circunferencia, de modo que o arco entre dúas observacións veciñas habería de ter lonxitude próxima a $\frac{2\pi}{n}$.
De cara ao entallado do estatístico, temos que arranxar as observacións en orden ascendente, $\theta_{(1)} \leq ... \leq \theta_{(n)}$, e calcular a lonxitude dos arcos entre dúas observacións colindantes coma

$$T_i = \theta_{(i+1)} - \theta_{(i)},\ i = 1, ..., n-1; \quad T_n = 2\pi + \theta_{(1)} - \theta_{(n)}.$$

Así, a desviación de T_i do valor esperado toma corpo en $|T_i - \frac{2\pi}{n}|$. O estatístico para este test construímolo do seguinte modo,

$$U = \frac{1}{2}\sum_{i=1}^{n}\left|T_i - \tfrac{2\pi}{n}\right|.$$

Reparemos en que o feito de que a desviación de T_i do valor esperado sexa grande é indicativo de ausencia de comportamento uniforme dos datos, logo se U fose maior que $U(\alpha)$, o valor crítico recollido nunha táboa para un nivel de significación α, non poderiamos aceptar a hipótese nula.

Con respecto ao Test de Rayleigh, que contrapón uniformidade fronte unha alternativa unimodal, paga a pena reparar en que o contraste da hipótese cambia de xeito dicotómico segundo a dirección media μ sexa coñecida ou non.
No caso de que o seu valor poboacional fósenos ignoto, é razoable tomar como estatístico de contraste $\bar{R} = \frac{R}{n} \in [0,1]$, sendo R a lonxitude do vector resultante da mostra (1.1), que reflicte a concentración dos datos respecto de $\bar{\theta}$. Así, se a mostra presentara un comportamento uniforme, cabería esperar que $\bar{R}$ fose pequeno.
Doutra banda, se o valor da dirección media teórica énos coñecido, $\mu = \mu_0$, acharemos unha medida que da conta da concentración dos datos respecto dese valor no estatístico

$$\overline{V}_0 = \bar{R}\cos(\bar{\theta} - \mu_0). \tag{3.1}$$

Un razoamento análogo ao exposto levaríanos a rexeitar a hipótese nula de uniformidade se o valor de $\overline{V}_0$ fose grande. Este test, con estatístico $\overline{V}_0$ e dirección media poboacional consabida, noméase test V. Observemos que, malia que é útil para contrastar uniformidade, este non é un test adecuado para o contraste dun valor específico para a dirección media poboacional.

Observación 3.1. Cando facemos referencia a valores de estatísticos grandes ou pequenos, falamos dun tamaño relativo ao nivel de confianza que estemos a esixirlle ao test, α, respecto da distribución do estatístico en cuestión.

Se, tras facer un test referido á uniformidade dunha mostra, comprobasemos que esta non o é, unha boa ferramenta supletoria de cara á identificación da distribución poboacional sería o contraste de simetría sobre a mostra. Daquela, o rexeitamento de uniformidade e aceptación de simetría para unha colección de datos circulares comportaría que un test de bondade de axuste, que introduciremos en seccións seguintes, dunha von Mises é razoable.
En [21], Pewsey plantexou un test non paramétrico de simetría (por reflexión) respecto dunha dirección media descoñecida, ciméntandose sobre o segundo momento dos senos respecto da dirección media mostral $\bar{b}_2$ (2.2), que atopa distribución asintótica nunha Normal linear de media e varianza

$$\mathbb{E}\left[\bar{b}_2\right] = \bar{\beta}_2 + \frac{1}{n\rho}\left(-\bar{\beta}_3 - \frac{\bar{\beta}_2}{\rho} + \frac{2}{\rho^3}\bar{\alpha}_2\bar{\beta}_2\right), \tag{3.2}$$

$$\mathrm{Var}(\bar{b}_2) = \frac{1}{n}\left\{\frac{1-\bar{\alpha}_4}{2} - 2\bar{\alpha}_2 - \bar{\beta}_2^2 + \frac{2}{\rho}\bar{\alpha}_2\left[\bar{\alpha}_3 + \frac{\bar{\alpha}_2}{\rho}(1-\bar{\alpha}_2)\right]\right\}. \tag{3.3}$$

Baixo a hipótese nula de simetría nunha distribución con $\rho \in (0,1)$, a dirección media poboacional existe e ademais os momentos dos senos respecto desta, $\bar{\beta}_p$ son nulos, de modo que $\mathbb{E}\left[\bar{b}_2\right] = 0$. Consonte a isto, no seu artigo, Pewsey propón a utilización do seguinte estatístico estudentizado para o contraste de simetría nunha mostra de dirección media poboacional

descoñecida,

$$z = \frac{\bar{b}_2}{\sqrt{\widehat{\mathrm{Var}}(\bar{b}_2)}} \in N(0,1),$$

onde por $\widehat{\mathrm{Var}}(\bar{b}_2)$ denotamos o estimador puntual de $\mathrm{Var}(\bar{b}_2)$ baixo asunción de simetría, isto é, substituíndo en (3.3) os momentos dos cosenos poboacionais, $\bar{\alpha}_p$, polos mostrais, $\bar{a}_p$, e tendo en consideración que, como apuntamos fai un intre, $\bar{\beta}_p = 0$.
Por conseguinte, valores grandes (positivos ou negativos) do estatístico respecto dos cuantís da normal estándar referidos ao nivel de significación escollido comprenden o rexeitamento da hipótese de simetría.

3.2. Inferencia preliminar para mostras unimodais

Comezaremos por afondar na estimación das medidas descriptivas máis relevantes en datos circulares: as direccións media, mediana e lonxitude normalizada do vector resultante, aputalándonos sobre o descrito por Fisher en [6], Capítulo 4.

Para un conxunto de observacións $\theta_1, \dots \theta_n$, podemos dar cun estimador da mediana poboacional sen máis ca tomar a mediana mostral, $\tilde{\theta}$, que se pode calcular coma o argumento que minimiza a distancia (1.4).
Se quixeramos facer un test directo sobre o posible valor da mediana, isto é, se quixeramos contrastar $H_0 : \tilde{\mu} = \tilde{\mu}_0$ fronte $H_1 : \tilde{\mu} \neq \tilde{\mu}_0$, haberiamos de construír un estatístico de contraste. Con ese obxectivo, imos dar conta do número de datos distintos do valor proposto que caen en $(\tilde{\mu}_0, \tilde{\mu}_0 + \pi)$, e do número de datos iguais ca el, que denotaremos respectivamente coma:

$$m = |\{\theta_i \in (\tilde{\mu}_0, \tilde{\mu}_0 + \pi) \ / \ \theta_i \neq \tilde{\mu}_0\}|, \qquad k = |\{\theta_i / \theta_i = \tilde{\mu}_0\}|.$$

O esperado, se a mediana poboacional coincidise con $\tilde{\mu}_0$, sería que m se achegara a $\frac{n-k}{2}$. Consonte a isto, o estatístico de contraste para este test sería $\frac{[2m-(n-k)]^2}{n-k} \in \chi^2_1$ e, en consecuencia, non aceptaríamos H_0 se o devandito estatístico fose grande respecto do cuantil de χ^2_1 asociado a un nivel de significación prefixado.

De xeito parello, un estimador puntual da dirección media poboacional toma corpo no seu análogo mostral, $\hat{\mu} = \bar{\theta} = \mathrm{atan2}(S, C)$. Para esta medida tamén é posible plantexar un contraste $H_0 : \mu = \mu_0$ fronte $H_1 : \mu \neq \mu_0$, atendendo ao feito de que $\widehat{\mathrm{Var}}(\bar{\theta}) = \dfrac{1 - \bar{a}_2}{2n\bar{R}^2}$, decontado podemos bosquexar o estatístico

$$\frac{\mathrm{sen}(\bar{\theta} - \mu_0)}{\sqrt{\widehat{\mathrm{Var}}(\bar{\theta})}} \in N(0,1),$$

consonte ao cal podemos rexeitar a hipótese nula comparando o seu valor empírico en valor absoluto cos cuantís $\frac{\alpha}{2}$ dunha distribución normal estándar.

Por último, tamén recordarmos (brevemente) que un estimador puntual da lonxitude do vector resultante normalizada poboacional ρ, corporízase no seu análogo mostral, $\bar{R} = \frac{1}{n}\sqrt{C^2 + S^2}$.

3.3. Inferencia sobre a distribución $vM(\mu, \kappa)$

Entre as distribucións presentadas no Capítulo 2 deste traballo, unha das máis relevantes na estatística de datos circulares é a distribución de von Mises, $vM(\mu, \kappa)$. Podérmonos fornecer de ferramentas inferenciais respecto destes modelos é capital no desenvolvemento dos estudos prácticos de observacións circulares.
As seguintes seccións ciméntanse sobre o publicado por Fisher en [6], Capítulo 4 e Mardia e Jupp en [12], Capítulos 5 e 7.

3.3.1. Estimación dos parámetros

Os estimadores dos parámetros dunha distribución von Mises con parámetro de localización μ e de concentración κ poden ser colectados mediante o método de máxima verosimilitude.
No Capítulo 2, na proposición (2.17) xa comprobamos que, para unha mostra de observacións circulares $\theta_1, \dots, \theta_n$ i.i.d, $\hat{\mu}_{mv} = \bar{\theta}$ era o estimador de máxima verosimilitude da dirección media para este modelo. De feito, vimos que esta distribución era a única cuxo estimador de máxima verosimilitude para a dirección media coincidía coa media mostral. De xeito completamente análogo procedemos a deducir o estimador para o parámetro de concentración.
A función de verosimilitude é

$$L = \prod_{i=1}^{n} f_{\mu,\kappa}(\theta_i) = \prod_{i=1}^{n} \frac{e^{\kappa \cos(\theta_i - \mu)}}{2\pi I_0(\kappa)} = \frac{e^{\kappa \sum_{i=1}^{n} \cos(\theta_i - \mu)}}{[2\pi I_0(\kappa)]^n}.$$

Consecuentemente, a log-verosimilitude vén dada por:

$$\log L = \sum_{i=1}^{n} \{-n \log[2\pi I_0(\kappa)] + \kappa \cos(\theta_i - \mu)\}. \tag{3.4}$$

Derivando esta función respecto do parámetro de interese, e tendo en conta que $\frac{\partial}{\partial \kappa} I_0(\kappa) = I_1(\kappa)$, obtemos

$$\frac{\partial}{\partial \kappa} \log L = -n \frac{I_1(\kappa)}{I_0(\kappa)} + \sum_{i=1}^{n} \cos(\theta_i - \mu).$$

Igualando a expresión anterior a cero acadamos a chamada ecuación de verosimilitude, da que podemos despexar o estimador buscado. Con este fin, habemos de substituír nela μ por $\hat{\mu}_{mv} = \bar{\theta}$,

e deseguido observar que $\sum_{i=1}^{n} \cos(\theta_i - \hat{\mu}_{mv}) = R\cos(\theta_i - \hat{\mu}_{mv}) = R$, para así expresar o anterior de xeito equivalente coma

$$\frac{\partial}{\partial \kappa} \log L = 0 \Leftrightarrow \frac{I_1(\kappa)}{I_0(\kappa)} = \frac{R}{n} \Leftrightarrow A(\kappa) = \bar{R}, \tag{3.5}$$

onde por $A(\kappa)$ denotamos $\frac{I_1(\kappa)}{I_0(\kappa)}$. As propiedades das funcións de Bessel modificadas de primeira especie permítennos colixir que a función $A(\kappa)$ é estritamente monótona crecente, así que $\hat{\kappa}_{mv}$ agroma como a única solución da ecuación (3.5).

Observemos que, mentres que o estimador da dirección media $\hat{\mu}_{mv}$ non depende do parámetro de concentración, $\hat{\kappa}_{mv}$ sí depende de μ. Por iso, se o valor poboacional μ nos fose coñecido, computariamos o valor V_0 (3.1), e como consecuencia da monotonía de $A(\kappa)$, poderiamos despexar

$$\hat{\kappa}_{mv} = \begin{cases} A^{-1}\left(\frac{V_0}{n}\right), & \text{se } V_0 \geq 0, \\ 0 & \text{se } V_0 \leq 0. \end{cases}$$

3.3.2. Distribución, nesgo e consistencia dos estimadores

Denotemos por γ un parámetro dunha familia de distribucións circulares e por $\hat{\gamma}$ un estimador do mesmo. Por paralelismo co caso real, poderiamos pensar en definir o nesgo para o tal estimador coma $\mathbb{E}[(\cos\hat{\gamma}, \operatorname{sen}\hat{\gamma})] - (\cos\gamma, \operatorname{sen}\gamma)$. Porén, a convexidade do disco unitario comportaría que a distribución do estimador estivese centrada en γ. Consonte a isto, é máis adecuado tomar unha definición máis débil do nesgo dun estimador circular.

Definición 3.2. O nesgo dun estimador circular defínese coma

$$\text{Nesgo}(\hat{\gamma}) = \frac{\mathbb{E}[(\cos\hat{\gamma}, \operatorname{sen}\hat{\gamma})]}{\|\mathbb{E}[(\cos\hat{\gamma}, \operatorname{sen}\hat{\gamma})]\|} - (\cos\gamma, \operatorname{sen}\gamma).$$

Diremos que un estimador circular $\hat{\gamma}$ dun parámetro γ é inesgado ou non ten nesgo se a dirección media de $\hat{\gamma}$ é o propio parámetro que se está a estimar, isto é, se o seu nesgo é nulo.

Co obxectivo de achar a distribución dos estimadores dunha von Mises, será importante recordar, que para mostras grandes,

$$\sqrt{n}(\hat{\mu}_{mv} - \mu, \hat{\kappa}_{mv} - \kappa) \approx N(0, \mathbf{I}^{-1}),$$

onde $\mathbf{I}$ é a matriz de información de Fisher, a saber:

$$\mathbf{I} = \mathbb{E}\left[-\begin{pmatrix} \frac{\partial^2}{\partial\mu^2}\log L & \frac{\partial^2}{\partial\mu\partial\kappa}\log L \\ \frac{\partial^2}{\partial\kappa\partial\mu}\log L & \frac{\partial^2}{\partial\kappa^2}\log L \end{pmatrix} \right] = \begin{pmatrix} \kappa A(\kappa) & 0 \\ 0 & 1 - A^2(\kappa) - \frac{A(\kappa)}{\kappa} \end{pmatrix}.$$

Así, para un tamaño mostral n considerable, podemos aproximar

$$\mathrm{Var}(\hat{\mu}) \approx \frac{1}{n\kappa A(\kappa)}, \quad \mathrm{Cov}(\hat{\mu},\hat{\kappa}) \approx 0, \quad \mathrm{Var}(\hat{\kappa}) \approx \frac{1}{n\left(1 - A^2(\kappa) - \frac{A(\kappa)}{\kappa}\right)}. \tag{3.6}$$

Daquela, $\hat{\mu}_{mv}, \hat{\kappa}_{mv}$ son aproximadamente independentes e seguen unha distribución normal con medias μ, κ, respectivamente, e as varianzas explicitadas con anterioridade.

Observación 3.3. Como queira que $A(\kappa) \overset{k\to\infty}{\longrightarrow} 1$, $A(\kappa) \overset{k\to 0}{\longrightarrow} 0$, se o parámetro de concentración κ é pequeno, μ estimarase con menos precisión.

Observación 3.4. Dado que $A(\kappa)$ é unha función non linear, $\hat{\kappa}_{mv} = A^{-1}(\bar{R})$ é un estimador con nesgo de κ. Unha aproximación para este nesgo é posible,

$$\mathbb{E}[\hat{\kappa}_{mv} - \kappa] \underset{\kappa\to\infty}{\overset{n\to\infty}{\approx}} \frac{3\kappa}{n}, \quad \mathbb{E}[\hat{\kappa}_{mv} - \kappa] \underset{\kappa\to 0}{\overset{n\to\infty}{\approx}} \frac{3\kappa}{5n}, \tag{3.7}$$

da cal podemos deducir que, para un mesmo tamaño mostral, a estimación de κ será peor canto maior sexa o seu valor poboacional.

Unha proposta para un estimador de κ que non presente tanto nesgo foi feita por Best e Fisher en [4]. Eles foron quen de probar, simulación mediante, que a seguinte alternativa a este estimador aproximábase a ter nesgo nulo, agás no caso de que ambos n, κ fosen pequenos.

$$\begin{cases} \text{máx}\left\{\dfrac{\hat{\kappa}_{mv} - 2}{n\hat{\kappa}_{mv}}, 0\right\} & \text{se } \hat{\kappa}_{mv} < 2, \\ \dfrac{(n-1)^3\hat{\kappa}_{mv}}{n^3 + n} & \text{se } \hat{\kappa}_{mv} \geq 2. \end{cases}$$

Outras alternativas a $\hat{\kappa}_{mv}$ poden acadarse, coñecendo as seguintes expansións da función $A(\kappa)$,

$$\begin{aligned} A(\kappa) &\approx 1 - \frac{1}{2\kappa} - \frac{1}{8\kappa^2} - \frac{1}{8\kappa^3} + o(\kappa^{-3}), \\ A(\kappa) &\approx \frac{\kappa}{2}\left(1 - \frac{1}{8}\kappa^2 + \frac{1}{48}\kappa^4 - \frac{11}{3072}\kappa^6 + \ldots\right). \end{aligned} \tag{3.8}$$

Ao realizarmos a inversión das expresións anteriores, podemos obter outras opcións para o estimador do parámetro de concentración tanto nos casos nos que $\bar{R}$ é grande,

$$\hat{\kappa} \approx \frac{1}{2(1-\bar{R}) - (1-\bar{R})^2 - (1-\bar{R})^3} \approx \frac{1}{2(1-\bar{R})};$$

coma na eventualidade de que $\bar{R}$ sexa pequeno,

$$\hat{\kappa} \approx 2\bar{R} + \bar{R}^3 + \frac{5}{6}\bar{R}^5.$$

A maiores, a primeira das aproximacións en (3.8) permítenos denotar por $\upsilon = \left(\frac{1}{\kappa_0} + \frac{3}{8\kappa_0^2}\right)^{-1}$, de xeito que, para $\kappa > 2$, é razoable aproximar $2\upsilon(n - R) = 2n\upsilon(1 - \bar{R}) \simeq \chi^2_{n-1}$. Segundo o que precede, tamén temos a descomposición

$$\underbrace{2n\upsilon(1-\bar{C})}_{\simeq\chi^2_n} = \underbrace{2n\upsilon(1-\bar{R})}_{\simeq\chi^2_{n-1}} + \underbrace{2n\upsilon(\bar{R}-\bar{C})}_{\simeq\chi^2_1}$$

Estas expresións seránnos de utilidade nas vindeiras páxinas.

Remataremos esta sección estudando a consistencia dos estimadores de máxima verosimilitude; para o cal será presciso coñecer as seguintes deficinións.

Definición 3.5. O erro cadrático medio para un estimador $\hat{\gamma}$ dun parámetro γ materialízase en

$$ECM_{\hat{\gamma}} = \mathbb{E}[(\hat{\gamma} - \gamma)^2] = \text{Nesgo}^2(\hat{\gamma}) + \text{Var}(\hat{\gamma}).$$

Definición 3.6. Nomearemos como estimador consistente a todo aquel estimador que, ao aumentar o tamaño da mostra, aproxime o seu valor ao do parámetro que está a estimar, i.e.

$$\lim_{n\to\infty} \text{ECM}_{\hat{\gamma}} = 0.$$

Atendendo ao que se expuxo de modo pretérito respecto das aproximacións da varianza (3.6) dos estimadores e do nesgo de $\hat{\kappa}_{mv}$ (3.7), podemos colixir que ámbolos dous estimadores de máxima verosimilitude son consistentes.
Todo este compendio de propiedades será observado de xeito empírico, por medio de simulación, en futuras seccións deste capítulo.

3.3.3. Tests sobre os parámetros e intervalos de confianza

Un artiluxio capital para poder elaborar tests sobre os parámetros dunha distribución $vM(\mu, \kappa)$ é o test da razón de verosimilitudes. Na estatística clásica, sendo X unha variable aleatoria real linear con distribución pertencente a unha familia $\{F_\gamma\}_{\gamma\in\Gamma}$, o estatístico da razón de verosimilitudes para o contraste de $H_0 : \gamma \in \Gamma_0$ fronte a $H_1 : \gamma \in \Gamma \setminus \Gamma_0$ definímolo coma

$$\lambda(\mathbf{x}) = \frac{\sup_{\gamma\in\Gamma_0} L(\mathbf{x})}{\sup_{\gamma\in\Gamma} L(\mathbf{x})}.$$

Co obxecto de elaborar os nosos propios tests da razón de verosimilitudes para os parámetros da distribución von Mises, habemos de reescribir a función de log-verosimilitude (3.4),

$$\log L = n\left[-\log 2\pi + \kappa\bar{R}\cos(\bar{\theta} - \mu) - \log I_0(\kappa)\right].$$

A partires desta expresión seranos confortable obter os estatísticos pertinentes para os tests sobre os parámetros.

Supoñamos que queremos comprobar que a dirección media dunha mostra toma un valor específico, isto é, queremos contrastar $H_0 : \mu = \mu_0$, fronte a hipótese alternativa $H_1 : \mu \neq \mu_0$. Se o parámetro de concentración é coñecido, o estatístico de razón de verosimilitudes é $\Upsilon = 2n\kappa(\bar{R} - \bar{V}_0)$, e poderemos aproximar a súa distribución coma $\Upsilon \simeq \chi^2_1$. Pola contra, se ignorasemos o valor poboacional do parámetro de concentración, o devandito estatístico toma corpo en

$$\Upsilon = 2n[\hat{\kappa}_{mv}(\bar{R} - \bar{V}_0) - \log I_0(\hat{\kappa}_{mv}) + \log I_0(\tilde{\kappa})], \quad \tilde{\kappa} = A^{-1}(\bar{V}_0).$$

De querermos simplificalo, bastaría substituír $\tilde{\kappa}$ por $\hat{\kappa}_{mv}$, de xeito que, baixo H_0, $\Upsilon = 2n\hat{\kappa}_{mv}(\bar{R} - \bar{V}_0) \simeq \chi^2_1$. Conforme a isto, rexeitaremos a hipótese nula se o valor do estatístico é maior ca o cuantil correspondente ao nivel de significación α dunha distribución khi cadrado dun grao de liberdade.

Outro xeito de contrastar esta hipótese é elaborando intervalos de confianza para o parámetro de localización e comprobando se o valor μ_0 cae dentro deles ou non. Un modo de facelo é alicerzándonos no estatístico $2n\hat{\kappa}_{mv}(\bar{R} - \bar{V}_0)$, de xeito que o intervalo terá extremos

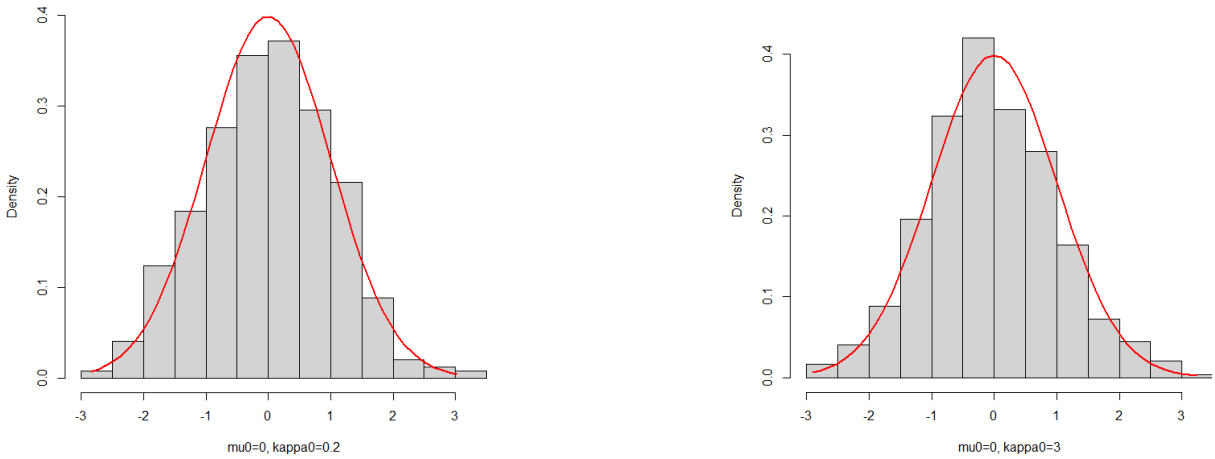
$$\bar{\theta} \pm \arccos\left(\frac{1 - \chi^2_{1,\alpha}}{2R\hat{\kappa}_{mv}}\right), \tag{3.9}$$

onde por $\chi^2_{1,\alpha}$ denotamos o cuantil α superior da antedita distribución.
Asemade, é posible elaborar outro intervalo, utilizando o feito de que $\sqrt{n\bar{R}\hat{\kappa}_{mv}}\,\operatorname{sen}\bar{\theta} \simeq N(0,1)$, e así

$$\bar{\theta} \pm \operatorname{arcsen}\left(\frac{z_{\alpha/2}}{\sqrt{n\bar{R}\hat{\kappa}_{mv}}}\right),$$

con $z_{\alpha/2}$ o cuantil dunha normal estándar que acumula á súa dereita probabilidade $\alpha/2$.

Observación 3.7. Esta aproximación é apropiada sempre que $\hat{\kappa}_{mv} \geq 2$ ou $\hat{\kappa}_{mv} \leq 0.4,\ n \geq 10$.
En ausencia de comprobación teórica na nosa bibiliografía para o afirmado, procedemos a facer unha ilustración empírica mediante simulación en R: imos xerar 500 mostras de tamaño 100 dunha $vM(0, \kappa_0)$ con $\kappa_0 \in \{0.2, 3\}$, conforme as hipóteses do enunciado. Para cada mostra, determinaremos $\bar{R}, \hat{\kappa}_{mv}, \bar{\theta}$ e almacenaremos o estatístico $\sqrt{n\bar{R}\hat{\kappa}_{mv}}\,\operatorname{sen}\bar{\theta}$ nun vector de xeito iterativo. A este lle podemos aplicar un test de Shapiro-Wilk para contraste de normalidade, do que obtemos valores críticos de 0.7 para $\kappa_0 = 0.2$, e de 0.8 para $\kappa_0 = 3$, isto é: podemos aceptar a hipótese de normalidade a todos os niveis habituais de significación. Na figura 3.1 observamos os histogramas da simulación resultante en cada caso, acompañados da función de densidade dunha Normal estándar, indicada con liña vermella. Coliximos que o resultado presentado é factible.



(a) $vM(0, 0.2)$ (b) $vM(0, 3)$

Figura 3.1: Simulación de $\sqrt{n\bar{R}\hat{\kappa}_{mv}}\,\mathrm{sen}\,\bar{\theta}$ e aproximación á distribución normal.

Atendendo á aproximación asintótica anterior, e sempre que esta sexa realizable, podemos elaborar un test paramétrico para a dirección media sen máis ca tomar $\hat{\sigma}_\mu = \frac{1}{\sqrt{n\bar{R}\kappa}}$ e o estatístico $E_n = \frac{\mathrm{sen}(\bar{\theta}-\mu_0)}{\hat{\sigma}_\mu} \simeq N(0,1)$, que variará segundo o parámetro de concentración nos sexa coñecido ou non. No primeiro caso, substituiríamos en E_n κ por κ_0, o valor poboacional; e no ulterior, por $\hat{\kappa}_{mv}$. Así, poderemos rexeitar $H_0 : \mu = \mu_0$ fronte a

$$\begin{cases} H_1 : \mu \neq \mu_0, & \text{se } |E_n| > z_{\alpha/2}, \\ H_1 : \mu > \mu_0, & \text{se } \mu_0 - \pi < \hat{\mu}_{mv} < \mu_0,\ E_n < -z_\alpha, \\ H_1 : \mu < \mu_0, & \text{se } \hat{\mu}_{mv} < \mu_0 + \pi,\ E_n > -z_\alpha. \end{cases}$$

Se tornamos agora a nosa atención aos tests referidos ao parámetro de concentración, en particular aos baseados no estatístico da razón de verosimilitudes, decatáremonos de que esta última, cando a dirección media non nos é consabida e a aproximamos por $\hat{\mu}_{mv}$, toma a expresión

$$\Upsilon = 2n\left[(\hat{\kappa}_{mv} - \kappa_0)\bar{R} - \log\frac{I_0(\hat{\kappa}_{mv})}{I_0(\kappa_0)}\right],$$

e logo o test depende monotonamente de $R = \sqrt{C^2 + S^2}$, de xeito que rexeitaremos $H_0 : \kappa > \kappa_0$ se R toma valores grandes; e $H_0 : \kappa < \kappa_0$ se R toma valores pequenos.

Observación 3.8. Baixo a hipótese nula, a distribución do parámetro R non nos é ignota.

A elaboración doutros tests aproximados é posible: tomando a expansión da función $A(\kappa)$, (3.8), é razoable aproximar $2v(n - R) \simeq X^2_{n-1}$, para $\kappa > 2$. Así, poderemos non aceptar H_0 :

$\kappa = \kappa_0$ fronte a

$$\begin{cases} H_1 : \kappa \neq \kappa_0, & \text{se } 2\upsilon(n-R) < \chi^2_{n-1,\alpha/2} \text{ ou } 2\upsilon(n-R) > \chi^2_{n-1,1-\alpha/2} \\ H_1 : \kappa > \kappa_0, & \text{se } 2\upsilon(n-R) > \chi^2_{n-1,1-\alpha} \\ H_1 : \kappa < \kappa_0, & \text{se } 2\upsilon(n-R) < \chi^2_{n-1,\alpha}, \end{cases}$$

no caso de que a dirección media nos sexa ignota. Na eventualidade na que $\mu = \mu_0$ fose consabida, se $\kappa > 2$, repetiriamos o test anterior, pero substituíndo R por $R_0 = \sum_{i=1}^{n} \cos(\theta_i - \mu_0)$. Así, non aceptariamos $H_0 : \kappa = \kappa_0$ fronte a

$$\begin{cases} H_1 : \kappa \neq \kappa_0, & \text{se } 2\upsilon(n-R_0) < \chi^2_{n,\alpha/2} \text{ ou } 2\upsilon(n-R_0) > \chi^2_{n,1-\alpha/2} \\ H_1 : \kappa > \kappa_0, & \text{se } 2\upsilon(n-R_0) > \chi^2_{n,1-\alpha} \\ H_1 : \kappa < \kappa_0, & \text{se } 2\upsilon(n-R_0) < \chi^2_{n,\alpha}. \end{cases}$$

Podemos tamén aproveitar este método para fabricar os intervalos de confianza de nivel $1-\alpha$ para o parámetro de concentración, sen máis ca tomar

$$a = \frac{n-R}{\chi^2_{n-1,1-\alpha/2}}, \quad b = \frac{n-R}{\chi^2_{n-1,\alpha/2}},$$

e así obter

$$\left(\frac{1+\sqrt{1+3a}}{4a}, \frac{1+\sqrt{1+3b}}{4b} \right). \tag{3.10}$$

3.4. Bondade de axuste

Chegados a este punto, é importante retornar a unha das ideas primordiais que subxace neste traballo: a de atopar artiluxios inscritos na linguaxe matemática que nos permitan aprehender a realidade. Sabemos, porén, que calquera acaecemento produto da estocástica incisa no mundo natural non pode ser encaixado con absoluta precisión nos nosos modelos teóricos. Porén, podemos esixir certos requerimentos razoables que os nosos modelos (neste caso, as nosas distribucións circulares) teñen que respectar para podelos utilizar coma ferramentas de comprensión da realidade: que presenten afinidade coas características definitorias dos datos que estamos tratar (con datos circulares, precisamos que as distribucións sexan periódcas), e que non contraveñan o comportamento dos datos observados. Este último requisito é o que motiva a creación de métodos que analicen a bondade do axuste dunha distribución sobre unha mostra.

Nesta sección centrarémonos en tres tests de bondade de axuste para mostras circulares, nos que a hipótese nula formúlase coma $H_0 : F(\theta) = F_0(\theta)$: o test χ^2, o test de Kuiper e o test de Watson. A información a este respecto foi extraída de Batschlet [3], Capítulo 4, e os artigos respectivos que citaremos no debido momento.

O test χ^2 neste caso ciméntase sobre ideas parecidas ás do análogo que atopamos na estatística real linear. O procedemento de cara a súa construción comeza por subdividir a circunferencia en k arcos (que poden ser de lonxitude variable, pero aos que pedimos, en xeral, que conteñan sempre 4 ou máis observacións mostrais[1]); e en cada un deles computar a frecuencia observada, que denotaremos coma n_j, e a frecuencia esperada baixo a hipótese nula, e_j, $j = 1, ..., k$. O estatístico para o contraste erixímolo coma

$$\sum_{j=1}^{k} \frac{(n_j - e_i)^2}{e_i} \in \chi^2_{k-1}. \tag{3.11}$$

Así pois, non aceptaremos a hipótese nula de que a nosa mostra dimana dunha distribución dada $F_0(\theta)$ para un nivel de significación α se o nivel crítico é menor ca éste.

Observación 3.9. Este test comporta dúas eleccións (o número de arcos k, e candaseus puntos extremos), que, como comprobaremos máis adiante, ocasinan que o antedito non sexa consistente para comprobar bondade de axuste.

De cara á confección do test de Kuiper, supoñamos que temos unha mostra $\theta_1, ..., \theta_n$ i.i.d con función de distribución baixo a hipótese nula $F_0(\theta)$. Para esta mostra, imos definir a súa función de distribución empírica.

Definición 3.10. Dada uha mostra $\theta_1, ..., \theta_n$, ordeada de xeito que $\theta_{(1)} \leq ... \leq \theta_{(n)}$, definimos a función de distribución empírica coma

$$F_n(\theta) = \begin{cases} 0, & \text{se } \theta < \theta_{(1)} \\ \frac{i}{n}, & \text{se } \theta_{(i)} \leq \theta < \theta_{(i+1)},\ i = 1, ... n-1 \\ 1, & \text{se } \theta \geq \theta_{(n)}. \end{cases}$$

Observación 3.11. A función de distribución empírica depende da elección da orixe e do sentido de rotación.

A partires deste concepto, imos construír o estatístico de Kolmogorov-Smirnov, xa coñecido por nós na asignatura de Inferencia Estatística, para mostras de datos circulares. Definimos as seguintes cantidades:

$$D_n^+ = \sqrt{n} \sup_\theta [F_n(\theta) - F(\theta)], \quad D_n^- = \sqrt{n} \sup_\theta [F(\theta) - F_n(\theta)], \quad D_n = \text{máx}\{Dn^+, Dn^-\}.$$

Decontado bosquexamos o estatístico para o test de Kuiper, $V_n = D_n^+ + D_n^-$; daquela, un valor pequeno do estatístico é indicativo dun bó axuste á mostra. Seguindo este fío de razoamento, non aceptaremos a hipótese nula se o valor do estatístico na mostra supera o valor crítico teórico para o nivel de significación prefixado.

[1] O motivo tras este requerimento é que o erro de test χ^2 é considerable cando hai frecuencias miúdas.

Observación 3.12. Notemos que, así definido, este estatístico é invariante por cambio na orixe.

Outra alternativa ao test de Kolmogorov-Smirnov no caso real linear é a do test de Cramer-von Mises, baseado no estatístico

$$C_n^2 = n \int_{-\infty}^{\infty} (G_n - G)^2 dF,$$

con $G(x)$, $G_n(x)$ as funcións de distribución e distribución empírica dunha variable real linear X. Watson [24] plantexou un test parello a este, adaptado a observacións circulares, cimentado no estatístico

$$W_n^2 = \int_0^{2\pi} \left[(F_n - F) - \int_0^{2\pi} (F_n - F) dF \right]^2 dF.$$

Observación 3.13. O estatístico de test de Watson é invariante por rotación.

Rexeitaremos a hipótese nula se o valor do estatístico é maior ca o cuantil correspondente ao nivel de significación desexado para o test.

Observación 3.14. Reparemos en que facer un test de bondade de axuste para unha certa distribución $F(\theta)$ sobre as observacións $\theta_1, \dots, \theta_n$ i.i.d. é equivalente a estudar a uniformidade da mostra transformada $\{2\pi F(\theta_i)\}_{i=1}^n$.

3.5. Ilustración dos resultados por simulación

É de grande interés poder ilustrar os resultados inferenciais respecto dos parámetros μ, κ dunha von Mises expostos preteritamente. O software R, e en particular o conglomerado de funcións anidadas no paquete `circular`, permítenos poder facer un gran número de simulacións da von Mises e tamén inferencia sobre os seus parámetros para poder comprobar empíricamente as propiedades expostas en seccións anteriores.

Un posible esquema para poder levar a cabo este obxectivo consta dos seguintes pasos:

1. Xeramos unha mostra dunha distribución $vM(\mu_0, \kappa_0)$, fixando de antemán os valores de μ_0, κ_0 e tamén o tamaño mostral $n \in \{50, 100, 500\}$, facendo uso da función `rvonmises`.

2. Para a devandita colección de observacións simuladas, determinamos os estimadores de máxima verosimilitude dos parámetros de interés, mediante `mle.vonmises`, que, en cada iteración, iremos acumulando nos vectores $\hat{\boldsymbol{\mu}}$, $\hat{\boldsymbol{\kappa}}$.

3. Repetimos os dous pasos anteriores $B = 500$ veces, de xeito que podamos completar $\hat{\boldsymbol{\mu}}' = (\hat{\mu}_{mv}^1, \dots, \hat{\mu}_{mv}^B)$, $\hat{\boldsymbol{\kappa}}' = (\hat{\kappa}_{mv}^1, \dots, \hat{\kappa}_{mv}^B)$.

4. Calculamos o nesgo e o erro cadrático medio dos estimadores de μ, κ, atendendo a súa natureza circular ou numérica (linear),

$$\text{Nesgo}(\hat{\mu}_{mv}) = \text{atan2}\left(S_{\hat{\boldsymbol{\mu}}}, C_{\hat{\boldsymbol{\mu}}}\right) - \mu_0, \quad \text{ECM}_{\hat{\mu}_{mv}} = \text{Nesgo}^2(\hat{\mu}_{mv}) + (1 - \bar{R}_{\hat{\boldsymbol{\mu}}});$$

$$\text{Nesgo}(\hat{\kappa}_{mv}) = \frac{1}{B}\sum_{j=1}^{B} \hat{\kappa}_{mv}^j - \kappa_0, \qquad \text{ECM}_{\hat{\mu}_{mv}} = \text{Nesgo}^2(\hat{\kappa}_{mv}) + \text{Var}(\hat{\kappa}_{mv}),$$

onde estamos a denotar $C_{\hat{\boldsymbol{\mu}}} = \sum_{j=1}^{B} \cos \hat{\mu}_{mv}^j$, $S_{\hat{\boldsymbol{\mu}}} = \sum_{j=1}^{B} \text{sen}\, \hat{\mu}_{mv}^j$, as compoñentes do coseno e seno do vector resultante asociado ás observacións abeiradas en $\hat{\boldsymbol{\mu}}$, e $\bar{R}_{\hat{\boldsymbol{\mu}}} = \frac{1}{n}\sqrt{C_{\hat{\boldsymbol{\mu}}}^2 + S_{\hat{\boldsymbol{\mu}}}^2}$, a súa lonxitude normalizada.

Con todo isto, poderiamos elaborar un cadro análogo a 3.1, onde temos recollido os resultados de distintas simulacións da von Mises con diversos tamaños mostrais e valores dos parámetros iniciais. Este compendio de resultados simulados permítenos observar en casos empíricos moitas

Modelo	Nesgo		Erro Cadrático Medio	
$\mu = 0 \quad \kappa = 1$	$\hat{\mu}_{mv}$	$\hat{\kappa}_{mv}$	$\hat{\mu}_{mv}$	$\hat{\kappa}_{mv}$
$n = 50$	0.01511	0.03923	0.02306	0.05752
$n = 100$	0.01159	0.02769	0.01098	0.02894
$n = 500$	-0.00641	0.00368	0.00241	0.00531
$\mu = \pi/2 \quad \kappa = 5$	$\hat{\mu}_{mv}$	$\hat{\kappa}_{mv}$	$\hat{\mu}_{mv}$	$\hat{\kappa}_{mv}$
$n = 50$	0.00201	0.25382	0.00247	1.11550
$n = 100$	-0.00130	0.09635	0.00108	0.45922
$n = 500$	0.00062	0.01417	0.00024	0.08801
$\mu = \pi \quad \kappa = 10$	$\hat{\mu}_{mv}$	$\hat{\kappa}_{mv}$	$\hat{\mu}_{mv}$	$\hat{\kappa}_{mv}$
$n = 50$	-0.00065	0.44506	0.00115	4.37187
$n = 100$	-0.00091	0.19999	0.00054	2.17563
$n = 500$	-0.00011	0.11246	0.00010	0.42878

Cadro 3.1: Simulación dos resultados inferenciais dunha von Mises con R.

das propiedades que comentabamos na sección 3.3 deste capítulo.

De modo primixenio, convén reparar en que o patrón dos tres modelos estudados en 3.1 é moi similar: a maior tamaño da mostra, menor é o nesgo (en valor absoluto) e o erro cadrático medio dos estimadores de máxima verosimilitude de ambos os dous parámetros, dando conta da consistencia dos mesmos que xa se expuxo con anterioridade. A maiores, é frutífero comparar os modelos entre sí segundo o tamaño do parámetro de concentración poboacional: para un mesmo

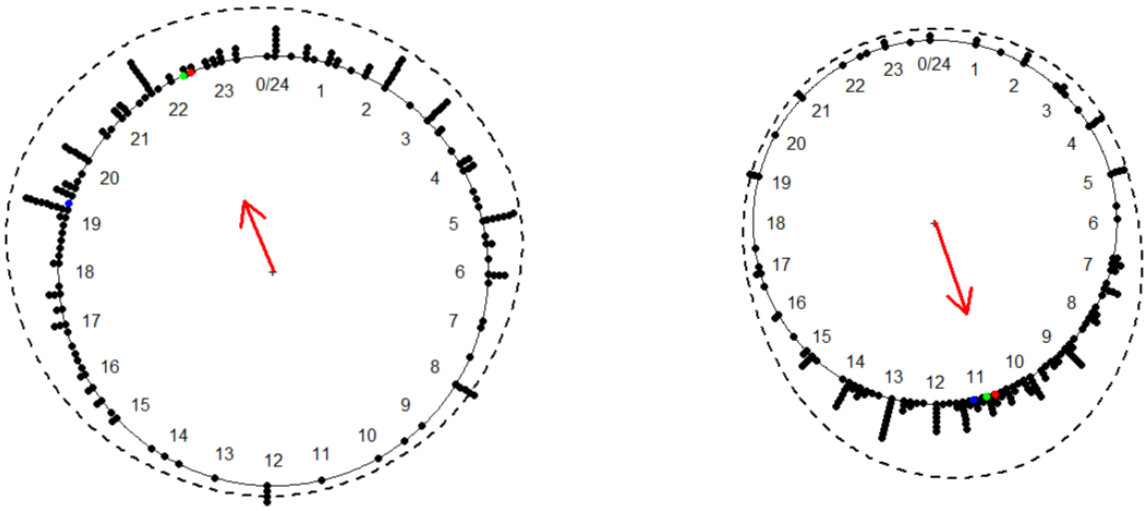
tamaño mostral, un valor grande de κ_0 comporta, como xa apuntabamos, un nesgo menor para o estimador do parámetro de localización; e un valor magno supón unha peor estimación de sí, xa que o nesgo de $\hat{\kappa}_{mv}$ e $\text{ECM}_{\hat{\kappa}_{mv}}$ van en aumento.

3.6. Ilustración das técnicas inferenciais con datos reais

Os instrumentos inferenciais que temos presentado permitiránnos exudar conclusións dos datos que xa introducidos e tratados preliminarmente en anteriores capítulos.

3.6.1. Cambios nos ciclos de temperatura en Monte Alvear

Como xa explicamos ao principio deste traballo, os cambios no ciclo de temperatura desta rexión periglaciar arxentina dan unha idea da afectación que sofre a rexión glaciar a causa do quentamento do noso planeta.
Posto que estamos medindo os momentos temporais ao longo dun día nos que cambian os ciclos, as observacións recollidas están medidas en radiáns no intervalo $[0, 2\pi)$, respecto dunha orixe situada en $\frac{\pi}{2}$ (a hora cero do reloxo) en sentido horario.
Ao longo desta sección, nas distintas figuras, estamos a indicar $\bar{\theta}$ como un punto vermello e



(a) Cambios de ciclo a xeada.

(b) Cambios de ciclo a desxeo.

Figura 3.2: Inferencia sobre as mostras de cambios de ciclo de temperatura.[2]

tamén como una flecha da mesma cor e lonxitude $\bar{R}$; $\tilde{\theta}$ como un punto verde, e $\check{\theta}$ en cor azul.

Comezaremos por estudar con detemento os datos recollidos nos cambios do ciclo de temperatura a xeada. Se esbozamos un grafo preliminar das observacións acompañadas dun estimador non paramétrico da densidade (figura 3.2a), pode agromar unha intuición de que a mostra semella ter un comportamento unimodal, non uniforme, concentrado nas horas de nocturnidade total (dez, once e doce da noite), se ben é perceptible un pico arredor das sete da tarde.
Do cálculo dos estimadores puntuais para dirección media, mediana, moda e lonxitude normalizada do vector resultante, empregando `mean.circular`, `median.circular`, `rho.circular` paquete `circular`, e `mode_est_circ` do paquete `bnpreg`, agroman os valores $\bar{\theta} = 5.88$ (en horas, 22:27 aproximadamente), $\tilde{\theta} = 5.85$, (22:20), $\check{\theta} = 5.03$ (19:12) e $\bar{R} = 0.35$. Reparemos en que estes valores semellan ratificar as intuicións iniciais que esbozamos no parágrafo anterior, por mor da semellanza entre media e mediana, a localización da moda no luscofusco, e un parámetro de concentración significativamente alonxado do cero (caso uniforme).
Os valores críticos correspondentes aos tests de uniformidade, calculados en R mediante os comandos `rayleight.test`, `rao.spacing.test`, `kuiper.test`, son ínfimos (menores de 0.001, de feito), co cal rexeitamos uniformidade para esta mostra. Conforme estes resultados, semella oportuno contrastar a posible simetría da mostra. Levando a cabo o test deseñado por Pewsey para este propósito, que programamos de xeito manual, facendo uso das funcións `trigonometric.moment` para mostras circulares, acadamos un p-valor segundo o que deducimos que non hai evidencias para rexeitar simetría aos niveis de significación do $0.1,\ 0.05,\ 0.01$.
É razoable, a vista destes resultados, comprobar se a mostra podería promanar dunha distribución von Mises. De cara a implementar este contraste, podemos elixir entre dúas opcións: introducir os nosos datos na función `watson.test` ou ben executar as referidas aos tests de uniformidade cos datos transformados $\{2\pi F(\theta_i)\}_{i=1}^n$, con $F(\theta)$ a distribución von Mises con parámetros estimados por máxima verosimilitude a partir da mostra. En calqueira caso, os valores críticos correspondentes a estes tests permítennos colixir que esta distribución axústase ben ao comportamento da mostra.
Visto isto, aproveitaremos os artiluxios inferenciais que temos á nosa disposición deseñados para a antedita distribución. O cálculo dos estimadores por máxima verosimilitude, mediante `mle.vonmises` devolve $\hat{\mu}_{mv} = 5.88$, $\hat{\kappa}_{mv} = 0.75$.

Para os cambios de ciclo a desxeo, podemos seguir un esquema semellante: comezamos por observar, sobre un grafo parello ao do caso anterior (figura 3.2b), que os datos parecen agromar dun modelo unimodal, simétrico e non uniforme; e están concentrados nas horas preto do mediodía. En efecto, computando os análogos mostrais de media, mediana, moda e ρ, colectamos os valores $\bar{\theta} = 2.80$ (en horas, 10:40 aproximadamente), $\tilde{\theta} = 2.85$, (10:53), $\check{\theta} = 2.92$ (11:09) e $\bar{R} = 0.53$. Paga a pena apercibirse de que a concentración respecto da dirección media é máis acusada no caso do desxeo ca no caso da xeada.

Respecto da uniformidade, os tests que temos dispoñibles corroboran as nosas sospeitas: os valores críticos que devolven son irrisorios. Non é así no caso da simetría, onde o p-valor é de 0.07, puidendo, en consencuencia, non rexeitala para niveis de significación menores ao 5 %. Por conseguinte, é axuizado realizar un test de bondade de axuste para a distribución von Mises. O test de Watson achega un valor crítico entre 0.5 e 0.01, e o resto, valores superiores ao 0.10: a un nivel de significación do 1 %, non rexeitamos a hipótese de simetría e tamén que a devandita distribución axústase ben á nosa mostra. Asemade, podemos achegar os valores de $\hat{\mu}_{mv} = 2.80$ e $\hat{\kappa}_{mv} = 1.25$, así como elaborar intervalos de confianza do 95 % para ambos os dous parámetros, compuntando directamente en R as expresións (3.9), (3.10). Deste modo, os intervalos

$$(1.223,\ 4.389), \quad (1.131,\ 1.599)$$

conteñen aos parámetros poboacionais μ, κ (respectivamente) cun 95 % de probabilidade.

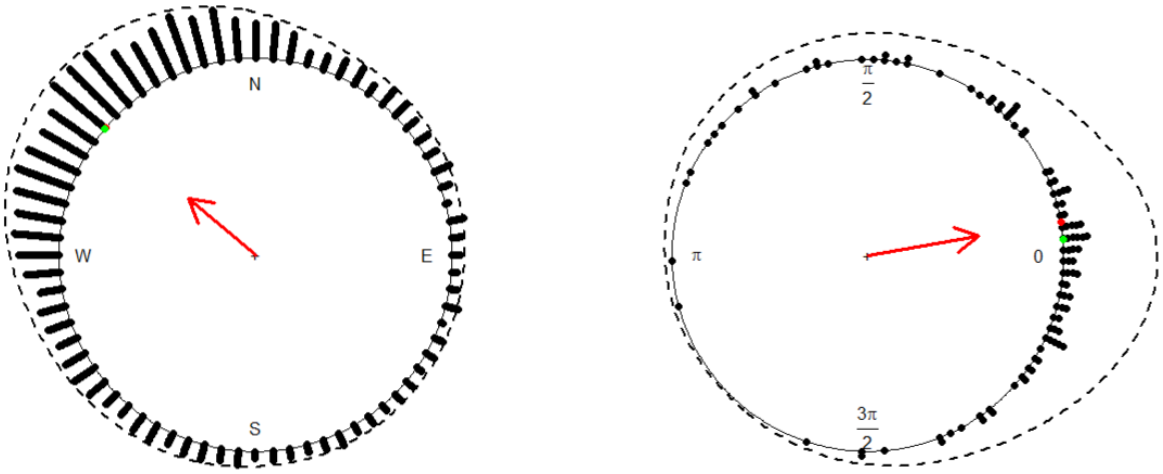
As conclusións que podemos extraer a partir destes resultados respecto ao comportamento nos cambios de ciclo a xeada e a desxeo é que ambas teñen un comportamento unimodal, cos cambios, como era de esperar, para xeada concentrados na nocturnidade, e para desxeo nas horas de máis luz; se ben os primeiros están máis espallados no circo ca os últimos.

3.6.2. Conduta das pulgas de praia

De acordo co que se comentou con anterioridade, a orientación das pulgas da praia ao voltar cara a ribeira reflicte o estado de conservación do ecosistema litoral. Neste caso, estamos a estudar a conduta destes anfípodos na praia de Zouara, na costa nordés de Túnez.
Cara o nordés é precisamente a orientación que toman maioritariamente os saltos, segundo o observable na figura 3.3a. Así o reflicte tamén a estimación non paramétrica da densidade, indicada con liña descontinua, ademais de achegar a idea de que o comportamento é unimodal, simétrico e non uniforme. Comprobemos se estas afirmacións correspóndense realmente coa poboación subxacente dos datos recollidos.

En efecto, o cálculo das medidas descritivas mostrais fainos ver que media, mediana e moda son coincidentes, $\bar{\theta} = 5.41$ (dirección nordeste), e que o parámetro de concentración $\bar{R} = 0.44$ está bastante alonxado do caso uniforme.
Podémonos cerciorar da veracidade desta última afirmación sen máis ca executar os mesmos tests ca en casos anteriores. Como queira que todos os valores críticos son extremadamente miúdos (menores ca 0.001), poderiamos rexeitar uniformidade. Procedemos inmediatamente co contraste de simetría: con un P-valor de 0.09, poderemos aceptar a hipótese nula aos niveis de significación habituais de 5 %, 1 %.

Logo que temos constatado, a eses niveis, que a mostra é simétrica e non uniforme, convén comprobar a bondade de axuste dunha von Mises: os tests devolven valores críticos insignificantes, polo que rexeitamos que a distribución subxacente se poida identificar coa antedita. Porén, podemos probar con outra das distribucións simétricas e unimodais que temos visto, por exemplo, a Normal Enrolada $WN(\mu, \rho)$: aplicaremos contraste de uniformidade sobre a mostra transformada $\{2\pi F(\theta_i)\}_{i=1}^n$, con $F(\theta)$ a función de distribución de $WN(\mu, \rho)$ (`pwrappednormal`). Decontado comprobamos que a Normal Enrolada tampouco se adapta ben ás nosas observacións, en tanto todos os valores críticos dos tests (Kuiper, Rayleigh, Rao e Watson) son ínfimos ou directamente nulos.



(a) Pulgas de praia.

(b) Voo das pombas mensaxeiras.

Figura 3.3: Inferencia preliminar sobre as mostras `sandhoppers`, `pigeons` do paquete `NPCirc` .

3.6.3. Xeolocalización das pombas mensaxeiras

Tal como adiantamos ao encomezo do traballo, a literatura biolóxica centrada na investigación dos mecanismos de xeolocalización en animais, en particular, nas pombas domésticas é ampla e conta con varias hipóteses para dar explicación ao seu funcionamento. Nas observacións recollidas en `pigeons`, e que trataremos deseguido, a dirección de residencia á que deberían voltar as pombas correspóndese co ángulo 2π na circunferencia.

Un cálculo preliminar dos valores mostrais das medidas descritivas máis importantes revela

que $\bar{\theta} = 0.17$, e que hai dous posibles direccións medianas, $\tilde{\theta}_1 = 0.1$, $\tilde{\theta}_2 = 0.07$. Isto estanos a advertir de que a mostra non ten por que ser unimodal; se ben a estimación non paramétrica da densidade semella contravir esa idea (figura 3.3b). Doutra banda, a lonxitude do vector resultante normalizada $\bar{R} = 0.57$ deixa entrever unha dispersión considerable.
Parellamente aos outros casos, podemos comezar por analizar un posible comportamento uniforme: os valores críticos dos test correspondentes a este contraste permítennos colixir que a mostra non é uniforme. Tampouco é simétrica: o P-valor do test correspondente, 0.002, é ínfimo e, en consencuencia, rexeitariamos simetría para esta mostra.

Xa que non podemos aceptar simetría, non contamos con evidencias estatísticamente significativas de que a mostra promane dunha distribución von Mises, polo cal este é un exemplo suxestivo no que pararse a estudar a fiabilidade do test khi cadrado, sinalado en seccións anteriores. O principal problema que agroma da utilización deste test de bondade de axuste é nado da elección implícita que supón cortar a circunferencia en arcos: tanto da escolla dos puntos inicial e final de cada arco como a súa lonxitude poden levar a resultados contraditorios.
Para comprobar a veracidade do test χ^2, imos realizar unha colección deles en R: bastará xerar algúns números aleatorios $a_1, \dots, a_{k+1}$ no intervalo $[0, 2\pi)$ e contar cantas observacións da nosa mostra caen en cada arco $[a_l, a_{l+1}]$, número que denotamos por n_j, $j = 1, \dots, k$.

Observación 3.15. Consideramos que unha colección de puntos $a_1, \dots, a_{k+1}$ é satisfactoria para levar a cabo o test cando $n_j \geq 4$, $j = 1, \dots, k$.

Unha vez fixados os arcos, procedemos a calcular os valores esperados en cada un coma $e_j = n[F(a_{l+1}) - F(a_l)]$, $j = 1, \dots, k$, con $F(\theta)$ a función de distribución da von Mises cos parámetros estimados por máxima verosimilitude a partires da mostra.
Decontado computamos o estatístico de contraste (3.11) e o valor crítico asociado calcúlase como $P(\chi^2_{k-1-2} > D)$.

Observación 3.16. Notemos que, xa que para este test estimamos dous parámetros, os graos de liberdade da distribución serán $k - 1 - 2$.

Implementando este compendio de ideas no código, é sinxelo iterar o test para distintas coleccións de puntos, obtendo en cada caso valores críticos moi dispares para a nosa mostra: nun caso acadamos un valor crítico de 0.29, que permitiría aceptar a hipótese nula de que a von Mises axústase ben ao comportamento das pombas aos niveis de significación habituais; mentres que noutro caso obtivemos 0.004, un valor crítico ínfimo atendendo ao cal rexeitariamos o axuste.

Capítulo 4

Conclusións

Nas páxinas que constitúen o noso traballo esperamos ser quen de imbuír na lectora con ideas introductorias de cara ao tratamento de datos de natureza circular, e da súa utilidade para tratar moitos dos fenómenos que acaecen derredor noso.

No capítulo co que abrimos o texto xenou a importancia de definir unhas medidas descritivas que non dependesen da elección da orixe e do sentido de xiro dos datos e asemade tamén agromou unha das diferenzas capitais entre a estatística clásica e a circular: a imposibilidade de poder definir sempre unha dirección media. É esta a problemática que fai gañar relevancia como medida de dispersión á lonxitude do vector resultante R (ou á normalizada, $\bar{R}$) fronte á varianza ou desviación típicas, xa que R toma o papel de indicador da existencia da media: ao tomares un valor nulo, a media é imposible de determinar.
Asemade, as representacións gráficas clásicas (histogramas, represetación no plano) das observacións resultan ineficaces para retratar adecuadamente a conduta dos datos circulares, polo que xorde a necesidade de buscar outras ou adaptar as existentes co obxecto de lograr unha efixie máis fiel. Neste texto, resaltamos algunhas destas ferramentas: os grafos de datos en crú, os histogramas, os diagramas de rosa e os estimadores non paramétricos da densidade.

No segundo capítulo, reparamos nas funcións de distribución e densidade para este tipo de datos, que, de xeito parello ao caso linear real, quedan unívocamente determinadas mediante a función característica. Porén, a índole periódica dos datos oblíganos a procurar novos modelos, que quer definimos atendendo a esa natureza (distribución uniforme, cardioide, von Mises), quer enrolando distribucións lineares coñecidas na circunferencia (distribución normal enrolada, cauchy enrolada, normal asimétrica enrolada), quer mesturando as devanditas.
Entre todas estas, dúas distribucións adoptan unha importancia capital no estudo dos datos cir-

culares, a saber: a distribución circular uniforme e a distribución von Mises. Para as observacións que subseguen un comportamento uniforme non é realizable a estipulación da dirección media, en tanto ningunha dirección acumula máis probabilidade ca outra. Doutra banda, a von Mises xoga un papel parello á distribución normal na estatística clásica; tanto é así que a primeira é a única para a cal o estimador de máxima verosimilitude do parámetro de localización coincide coa media mostral (analogamente ao que ocorre coa distribución normal).

A inferencia é precisamente a que ocupa o derradeiro dos capítulos que compón o corpus principal do traballo, e nos que afondamos en estimación puntual para mostras unimodais e de máxima verosimilitude para a von Mises (nas que non só profundamos de modo teórico, senón que a maiores, ilustramos mediante simulación), e tamén nos contrastes de uniformidade, simetría e bondade de axuste, que hannos de posibilitar a identificación do comportamento dos datos circulares. Exemplificamos estes dispositivos inferenciais mediante a súa aplicación a datos reais, recollidos nos paquetes `NPcirc` e `circular` de R, puidendo así destilar conclusións respecto deles.

Secasí, os resultados que puidemos extraer destes datos teñen unha postura case preliminar, pois o exposto nestas páxinas toma un carácter meramente introdutorio, xa que na actualidade, malia ser un campo en constante expansión e desenvolvemento, contamos con conceptos e ferramentas sólidos para datos circulares que non temos recollidos aquí, pero que posibilitan análises moito máis complexas e ricas a través dun compendio de ferramentas estatísticas, entre elas, a comparativa entre dúas ou máis mostras que promanen dunha mesma poboación, correlación e regresión, inferencia preditiva, ou detección de datos atípicos.

Por último, quixeramos aproveitar estas derradeiras liñas para salientar non só a utilidade dos datos circulares como espello (matemático, estatístico) idóneo para moitas das condutas inscritas na nosa realidade material; senón tamén o aspecto intrínsecamente humano de intentar bosquexar con verbas o que non abrangue coas mans, que rebule, e sustenta co seu rebulir, as achegas teóricas e prácticas que integran este traballo; e que é a corporización exacta da arela descrita na súa nacenza: a traslación da substantividade ao tecido lingüístico.

Bibliografía

[1] Alonso-Pena, M. e Crujeiras, R. M. (2023). Analizing animal escape data with circular nonparametric multimodal regression. *Annals of Applied Statistics*, **17**(1), 130-152 .

[2] Agostinelli, C. e Lund, U. (2024). R package 'circular': Circular Statistics (version 0.5-1). URL `https://CRAN.R-project.org/package=circular`

[3] Batschelet, E. (1981). *Circular Statystics in Biology*. Academic Press.

[4] Best, D. J., Fisher, N. I. (1981). The BIAS of the maximum likelihood estimators of the von Mises-Fisher concentration parameters. *Communication in Statistics- Simulation and Computation*, **10**, 493-502.

[5] Corder, A. P., Cross, M., Julious, S. A., Mullee, M. A. e Taylor, I. (1994). The timing of breast cancer surgery within the menstrual cycle. *Postgraduate Medical Journal*, **70**, 281-284.

[6] Fisher, N. I. (1993). *Statistical Analysis of Circular Data*. Cambridge University Press.

[7] Gagliardo, A., Ioale, P., Savini, M. e Wild, M. (2008). Navigational abilities of homing pigeons deprived of olfactory or trigeminally mediated magnetic information when young. *Journal of Experimental Biology*, **211**, 2046–2051.

[8] Grancher, D., Bar-Hen, Avner, Paris, R. , Lavigne, F. e Brunstein, D. (2012). Spatial interpolation of circular data: application to tsunami of December 2004. *Advances and Applications in Statistics*. **30**(1), 19-29.

[9] Jammalamadaka, S.R. e SenGupta, A. (2001). *Topics in Circular Statistics*. World Scientific.

[10] Landler L., Ruxton G.D., Malkemper E.P. (2018). Circular data in biology: advice for effectively implementing statistical procedures. Behavioral Ecology and Sociobiology. **72**(8), 128.

[11] Marchetti, G. M. e Scapini, F. (2003). Use of multiple regression models in the study of sandhopper orientation under natural conditions. *Estuarine, Coastal and Shelf Science*, **58**, 207-215.

[12] Mardia, K. V. e Jupp, P. E. (2000). *Directional Statistics*. John Wiley, Chichester.

[13] Mardia, K. V., Taylor, C. C. e Subramaniam, G. K. (2006). Protein Bioinformatics and Mixtures of Bivariate von Mises Distributions for Angular Data. *Biometrics*, **63**, 505-512.

[14] Nair, A., Changsing, K., Stewart, W.J. and McHenry, M.J. (2017). Fish prey change strategy with the direction of a threat. *Proceedings of the Royal Society B: Biological Sciences*, **284**.

[15] Nam H. Tran (2007), Fracture orientation characterization: Minimizing statistical modelling errors. *Computational Statistics and Data Analysis*, **51** (6), 3187-3196.

[16] Oliveira, M., Crujeiras, R. M. e Rodriguez-Casal,A. (2013). A plug-in rule for bandwidth selection in circular density estimation. Computational Statistics and Data Analysis, **56**(12), 3898-3908

[17] Oliveira, M., Crujeiras, R. M. e Rodriguez-Casal,A. (2013). Nonparametric circular methods for exploring environmental data.*Environmental and Ecological Statistics*, **20**, 1–17.

[18] Oliveira, M., Crujeiras, R. M. e Rodriguez-Casal,A. (2014). NPCirc: An R Package for Nonparametric Circular Methods. Journal of Statistical Software, **61** (9), 1-26. URL `http://www.jstatsoft.org/v61/i09/`.

[19] Pewsey, A., Neuhäuser, M., e Ruxton, G. D. (2013). *Circular statistics in R*. OUP Oxford.

[20] Pewsey, A. (2000) The wrapped skew-normal distribution on the circle. *Communications in Statistics- Theory and Methods*, **29**(11), 2459-2472.

[21] Pewsey, A. (2002). Testing circular symmetry. *The Canadian Journal of Statistics*, **30** (4), 591-600.

[22] Pewsey, A. (2004). The large-sample joint distribution of key circular statistics. *Metrika* **60**, 25–32.

[23] R Core Team (2020). *R: A language and environment for statistical computing. R Foundation for Statistical Computing*, Vienna, Austria. URL `https://www.R-project.org/`.

[24] Watson, G. S. (1961), Goodness-Of-Fit Tests on a Circle. *Biometrika*, **48** (1), 109-114.

[25] Xu D, Wang Y (2022). cplots: Plots for Circular Data. R package (version 0.5-0). URL `https://CRAN.R-project.org/package=cplots`.