Universidad de Oviedo

Trabajo Fin de Grado

Técnicas estadísticas para datos circulares

Laura García Díaz

Supervisado por: Arís Fanjul Hevia y Laura Freijeiro González

UNIVERSIDAD DE OVIEDO

Facultad de Ciencias

Grado en Matemáticas

Junio de 2024

Índice general

Capítulo 1

Introducción

La estadística clásica es la disciplina encargada del estudio de variables aleatorias, de recoger, analizar e interpretar los datos procedentes de estas. Sin embargo, existen ciertos tipos de datos cuyas características son algo particulares y es necesario el desarrollo o la modificación de algunas herramientas estadísticas de forma que se ajusten a estas propiedades.

Este es el caso de los datos circulares, que son aquellos datos que pueden ser representados como puntos en una circunferencia de radio unidad o como un vector unitario en el plano, es decir, un ángulo. Del estudio de estos se encarga la estadística circular.

Algunas de las principales características de este tipo de datos ya fueron explotadas en siglos anteriores en distintos ámbitos. Los primeros análisis de estadística circular datan del siglo XVIII, cuando el reverendo John Mitchell estudió la separación angular de las estrellas y vio que había demasiadas parejas cercanas como para considerar correcta la hipótesis de que las direcciones en las que se encontraban estas estrellas estuviesen uniformemente distribuidas. De esta forma, concluyó que estas parejas de estrellas debían estar enlazadas mediante atracción gravitatoria. Años después, la idea de que este tipo de datos debían ser estudiados de forma diferente a los lineales fue introducida por John Playfair (Fisher, 1993).

Posteriormente, durante la Guerra de Crimea (1853–1856), Florence Nightingale, jefa de enfermería de la armada británica, se dio cuenta de que se podrían haber salvado muchas vidas mejorando las condiciones sanitarias en hospitales y barracones. Para demostrar su estudio, utilizó argumentos gráficos, desarrollando lo que denominó como diagrama de área polar (*polar area diagram*). Esta es una representación gráfica especialmente diseñada para datos circulares, muy similar al diagrama de rosa que estudiaremos al final del capítulo 3. Este diagrama de área polar es el que puede verse en la figura 1.1. En él se muestran las muertes producidas por distintas enfermedades en cada mes (sector) de un año concreto: en azul se representan las muertes por enfermedades contagiosas; en rojo, las causadas por heridas y, en negro, las

producidas por otras causas (Fisher, 1993). En esta imagen se representan dos de los diagramas mencionados: el de la derecha se corresponde con las muertes producidas entre abril de 1854 y marzo de 1855, mientras que el de la izquierda con las producidas entre abril de 1855 y marzo de 1856. Florence Nightingale construyó el primero de ellos cuando llegó a la región donde iba a ejercer de enfermera y vio que la mayor parte de las muertes eran provocadas por enfermedades contagiosas (sectores azules), debidas a las malas condiciones sanitarias del momento, en lugar de por heridas de guerra (sectores rojos). Las malas condiciones sanitarias del momento contribuían a la propagación de las infecciones, por lo que introdujo algunas mejoras en el hospital y en los procedimientos utilizados en el cuidado de los pacientes. Tras estas modificaciones, Nightingale volvió a realizar su estudio y representó el segundo de los gráficos mencionados. En este podemos ver cómo a medida que avanzan los meses del año 1855, las muertes producidas por enfermedades contagiosas se reducen considerablemente. Por ello, su estudio fue revolucionario.

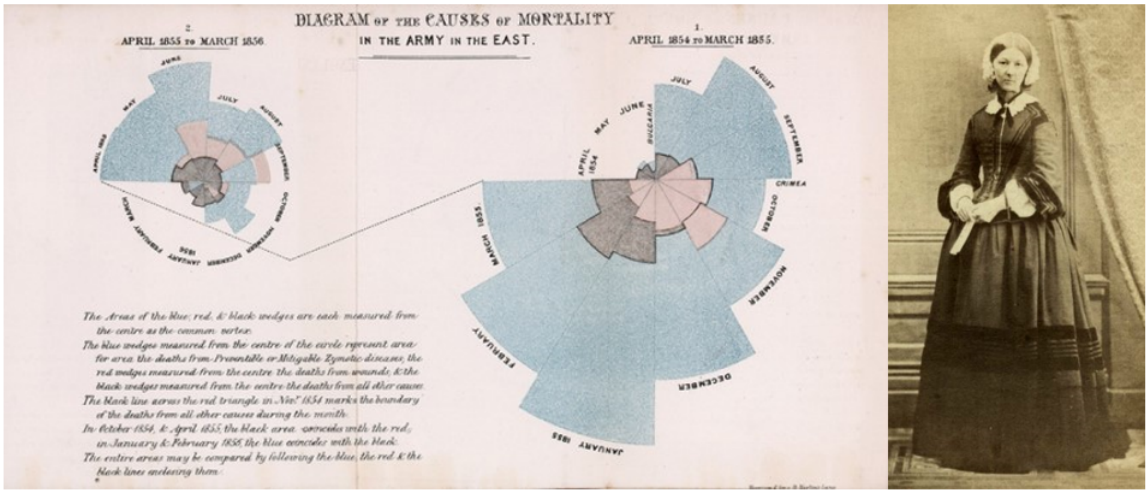


Figura 1.1: En la imagen de la izquierda puede verse el diagrama de área polar ideado por Florence Nightingale durante la Guerra de Crimea. A la derecha se tiene una fotografía de la propia Florence Nightingale.

Además de en Astronomía y Medicina, la estadística circular tiene especial importancia en otras áreas como pueden ser la Biología (como el estudio de la dirección de migración de las aves) o en Ecología (como la dirección del aire que transporta sustancias contaminantes), entre otras. Algunos de estos estudios de estadística circular en biología pueden verse en las obras como Batschelet (1981). Otros ejemplos de aplicaciones en medicina se encuentran en Jammalamadaka y SenGupta (2001), entre otros.

Con lo mencionado hasta ahora, se hace latente la gran utilidad de la estadística circular en diversas materias. Sin embargo, las características específicas de los datos circulares harán que

surjan ciertos problemas que no existen en la estadística convencional.

El principal objetivo de este proyecto es presentar las propiedades más importantes de los datos circulares, así como proporcionar soluciones a las complicaciones que surgen en el estudio de los mismos. Además, compararemos las diferencias y similitudes entre los nuevos resultados obtenidos con los ya conocidos de la estadística clásica. De esta forma, veremos si es posible heredar, adaptar o extender algunos de los procedimientos empleados en esta última y definir unos nuevos, si no lo es. Entre los primeros métodos que veremos si se pueden emplear o es necesario adaptarlos, se encuentran los relacionados con la estadística descriptiva. También introduciremos modelos paramétricos ideados para ajustar la distribución de estos datos y proporcionaremos herramientas para estimar la función de densidad de aquellos que parecen no seguir estos modelos. Todo ello será aplicado finalmente a dos bases de datos reales.

Para este fin, comenzaremos el capítulo 2 con una introducción a los conceptos básicos de la estadística clásica, seguida de la extensión al caso circular. Además, se motivará su necesidad con algunos ejemplos donde los métodos estadísticos clásicos claramente no proporcionan resultados correctos en este contexto. Presentaremos las funciones de distribución, de densidad (en el caso de variables aleatorias continuas) y su función característica asociada apropiadas en estadística circular, comentado también sus propiedades.

Seguiremos con el capítulo 3 destinado a la estadística descriptiva circular, en el que obtendremos algunas medidas de posición y dispersión. Será necesario presentar nuevas definiciones de los parámetros usuales, tales como la media, la mediana o la varianza; y definir algunos conceptos nuevos como los parámetros de concentración. Por otro lado, mostraremos las principales formas de representar gráficamente los datos circulares.

Posteriormente, dedicaremos el capítulo 4 a introducir las distribuciones circulares más comunes. Algunas de ellas no tienen equivalente en la estadística no circular, es decir, se desarrollan específicamente para este tipo datos, mientras que otras se obtienen a partir de adaptaciones de distribuciones conocidas de la estadística clásica.

Continuaremos el estudio hablando de la estimación no paramétrica de la función de densidad (capítulo 5), cuyo objetivo consiste en modelar la tendencia de una serie de datos de los que se desconoce su distribución. Veremos primero una introducción a la estimación no paramétrica en el caso lineal y seguiremos con un análisis para el caso circular. En ambos casos nos vamos a encontrar con las mismas dificultades: seleccionar un parámetro de suavizado adecuado para obtener una buena estimación de las funciones de densidad. Mediante simulaciones, ilustraremos el comportamiento del estimador utilizando distintos parámetros de suavizado.

En el capítulo 6, aplicaremos todos los métodos estudiados a dos bases de datos reales para

conocer mejor y modelar el comportamiento de dichos fenómenos. La primera de ellas consta de un conjunto de medidas de las direcciones de la salida de las partículas resultantes de la colisión de protones en el LHC (*Large Hadron Collider*). La segunda muestra está formada por observaciones de la dirección de las corrientes marinas en el Cabo de Peñas. Para todas las simulaciones y análisis de datos se utiliza el *software R* (R Core Team, 2022).

Finalizaremos con el capítulo 7 en el que expondremos las conclusiones que se pueden extraer de este análisis y mencionaremos algunas posibles ampliaciones del estudio realizado.

Los códigos empleados en las simulaciones mostradas a lo largo del trabajo se pueden ver en el anexo A tras este último capítulo.

Capítulo 2

Conceptos básicos

Para poder interpretar de forma adecuada los nuevos resultados relacionados con los datos circulares, es conveniente hacer una pequeña introducción a la estadística clásica para recordar sus principales características. De esta forma, será más sencillo hacer una comparación entre sus propiedades y las de la estadística circular.

Tras este recordatorio, comenzaremos con el estudio propiamente dicho de los datos circulares definiendo los conceptos necesarios para el posterior análisis.

2.1. Estadística clásica

Una variable aleatoria lineal asociada a un experimento aleatorio se define como una aplicación $X : \Omega \to \mathbb{R}$, es decir, asigna a cada posible resultado del experimento del espacio muestral Ω un valor en la recta real (Gorgas *et al.*, 2011). Se puede estudiar el comportamiento de X a través de su función de distribución $\mathcal{F}$.

Definición 2.1. *Sea X una variable aleatoria lineal. Se define su función de distribución como una función $\mathcal{F} : \mathbb{R} \to \mathbb{R}$ tal que $\mathcal{F}(x) = P(X \leq x)$, es decir, la probabilidad de que X tome valores menores o iguales x. Además, $\mathcal{F}$ debe satisfacer las siguientes propiedades:*

1. *La función de densidad toma valores entre cero y uno, de forma que*

$$\lim_{x\to\infty} \mathcal{F}(x) = 1, \qquad \lim_{x\to-\infty} \mathcal{F}(x) = 0. \tag{2.1}$$

2. *$\mathcal{F}(x)$ es no decreciente, es decir, $\forall x, y \in \mathbb{R}$ con $x \leq y$, se tiene que $\mathcal{F}(x) \leq \mathcal{F}(y)$.*

3. *$\mathcal{F}(x)$ es continua por la derecha, esto es, $\forall x_0 \in \mathbb{R}$:*

$$\mathcal{F}(x_0^+) = \lim_{x\to x_0^+} \mathcal{F}(x) = \mathcal{F}(x_0).$$

En el caso de que X pueda tomar una cantidad numerable de valores $\{x_1, ..., x_n, ...\}$, diremos que es una variable aleatoria discreta y podemos definir su función de probabilidad.

Definición 2.2. *Sea X una variable aleatoria discreta que toma valores $\{x_1, ..., x_n, ...\}$. Denotando por $p_i = P(X = x_i)\ \forall i \in \mathbb{N}$, el conjunto de estas probabilidad $\{p_1, ..., p_n, ...\}$ se denomina función de probabilidad. Para garantizar que se verifica la propiedad (2.1) de la función de distribución se debe cumplir que*

1. $p_i > 0\ \forall i \in \mathbb{N}$,
2. $\sum_{i=1}^{\infty} p_i = 1$.

Así, podemos reescribir dicha distribución como sigue:

$$\mathcal{F}(x) = \sum_{i \in \mathbb{N} |\ x_i \leq x} p_i.$$

Por otro lado, X es una variable aleatoria continua si su función de distribución es absolutamente continua y se puede definir su función de densidad.

Definición 2.3. *Sea X una variable aleatoria continua con función de distribución $\mathcal{F}$. Se define la función de densidad de X como aquella $\mathrm{f} : \mathbb{R} \to \mathbb{R}$ tal que*

$$\mathcal{F}(x) = \int_{-\infty}^{x} \mathrm{f}(t)dt \iff \mathrm{f}(x) = \frac{d}{dx}\mathcal{F}(x).$$

Además, ha de verificar que

1. $\mathrm{f}(t) \geq 0, \forall t \in \mathbb{R}$,
2. $\int_{-\infty}^{\infty} \mathrm{f}(t)dt = 1$.

En ambos casos podemos definir unos parámetros representativos de la variable aleatoria X, como son las medidas de posición y dispersión. Como ejemplo de las primeras tenemos la esperanza o media poblacional, que se define como

$$\mathbb{E}(X) = \begin{cases} \sum_{i=1}^{\infty} x_i P(x = x_i) & \text{si} \quad X \text{ es discreta con valores } x_i \\ \int_{-\infty}^{\infty} x\ \mathrm{f}(x)dx & \text{si} \quad X \text{ es continua con función de densidad } f(x). \end{cases}$$

Como medidas de dispersión caben mencionar la varianza y la desviación típica. Dada una variable aleatoria X su varianza se obtiene mediante la expresión

$$Var(X) = \mathbb{E}([X - \mathbb{E}(X)]^2) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2. \tag{2.2}$$

A partir de la esta se define la desviación típica como

$$\sigma = \sqrt{Var(X)}.$$

Para finalizar esta sección, introducimos a continuación el concepto de función característica.

Definición 2.4. *Dada una variable aleatoria X se define su función característica para cierto $t \in \mathbb{R}$ como*

$$g(t) = \mathbb{E}(e^{itX}), \tag{2.3}$$

donde i es la unidad imaginaria.

Notemos que la derivada k-ésima de esta función característica es

$$g^{(k)}(t) = \mathbb{E}[(iX)^k e^{itX}].$$

Así, evaluando esta ecuación en el punto $t = 0$, se obtienen los momentos:

$$\mathbb{E}(X^k) = i^{-k} g^{(k)}(0). \tag{2.4}$$

Es posible obtener medidas de posición empleando los momentos obtenidos en (2.4) para valores de $k \in \mathbb{N}$. Por ejemplo, cuando $k = 1$, obtenemos la expresión de la media poblacional.

2.2. Estadística circular

Comencemos recordando que hemos definido los datos circulares como aquellos que se pueden ver como puntos en la circunferencia de radio uno o una dirección o vector unitario en el plano (Mardia y Jupp, 2000).

Este tipo de datos tiene algunas características especiales que no poseen los datos lineales[1], entendiendo estos últimos como aquellos que toman valores en la recta real. En primer lugar, al poder representarse los primeros como un vector o un ángulo, es posible que dicha representación no sea única. Para ello, se ha de elegir una dirección preferencial, denominada dirección cero (que se denotará de ahora en adelante como ángulo 0), y un sentido de rotación, que fija el sentido de lectura de los datos.

En la figura 2.1 se muestra un ejemplo de lo explicado: si se toma como dirección cero el norte (N) y el sentido de rotación horario (azul), el dato circular indicado con el vector unitario se correspondería con un ángulo de 30°. Sin embargo, al elegir el este (E) como dirección cero y el sentido de rotación antihorario (verde), este dato se corresponde con un ángulo de 60º.

Es importante mencionar que la elección de la dirección cero y del sentido de rotación para el análisis de los datos no debe influir en los resultados que se obtengan. Puntualicemos que, si

[1]Con el fin de simplificar notación, utilizaremos "datos lineales" para referirnos a los no circulares.

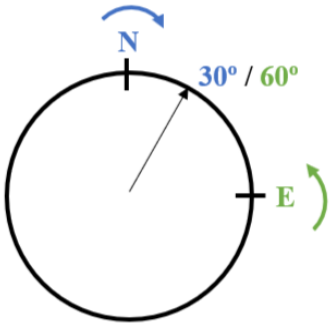


Figura 2.1: Ejemplo de la no unicidad de representación numérica de los datos circulares. Se representa el ángulo señalado con la flecha correspondiente al tomar la dirección cero en el norte y sentido horario en azul y al tomar la dirección cero el este y sentido antihorario en verde.

no se indica lo contrario, se tomará el origen 0 en el este, el sentido de rotación antihorario y los ángulos estarán medidos en radianes.

Además, debido a la representación angular de los datos, es claro que las medidas son periódicas. Obviamente el principio de la circunferencia coincide con su final, es decir, $0° = 360°$ o, lo que es equivalente, 0 rad $= 2\pi$ rad. Por tanto, si a un dato le corresponde un ángulo θ, entonces también se puede expresar como $\theta = \theta + 2\pi k$ con $k \in \mathbb{Z}$.

Por todas estas razones son necesarios nuevos métodos estadísticos para el estudio de los datos circulares. Las definiciones usuales para la media, la mediana, la varianza, etc. en estadística lineal, algunas de ellas vistas en la sección 2.1, no son correctas en el caso circular. Por tanto, es necesario redefinir estos conceptos, así como modificar las distribuciones ya conocidas para datos lineales como puede ser la distribución normal.

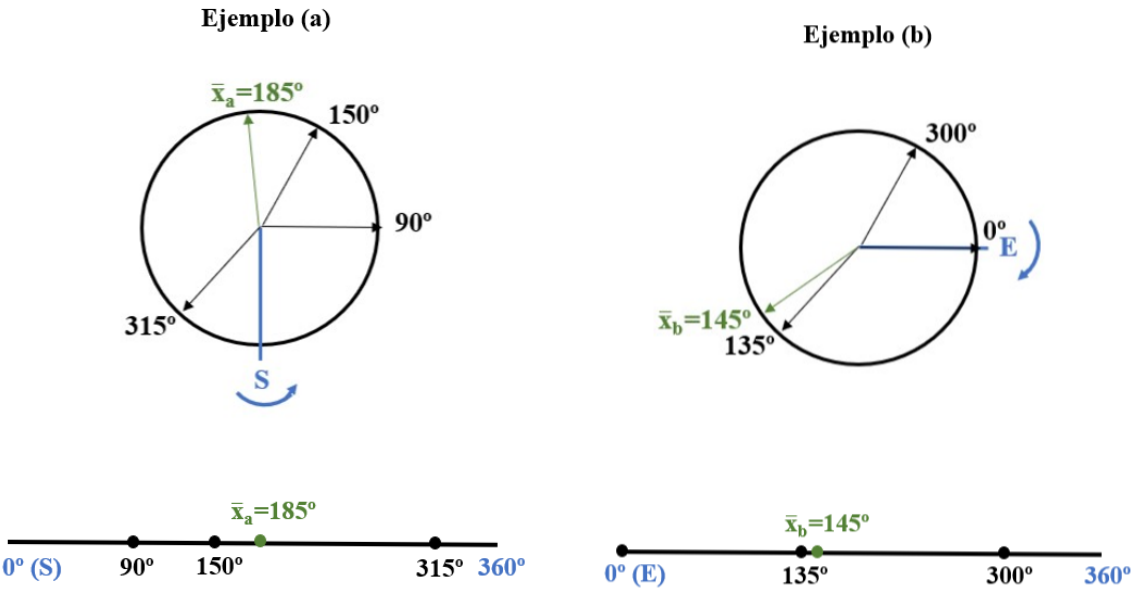
A continuación se muestra un ejemplo de cómo las técnicas habituales en estadística lineal no sirven para datos circulares. Como se puede ver en la figura 2.2, tomamos tres datos medidos desde dos direcciones iniciales diferentes y en sentidos de rotación distintos. Así, aunque se trate de las mismas medidas, al tomar distintos orígenes y sentidos de rotación, estos datos toman valores diferentes sobre la circunferencia. Esta es otra clara diferencia entre los datos circulares y los lineales, que además se puede observar las rectas representadas en las figuras 2.2a y 2.2b.

Dada una muestra aleatoria simple $(X_1,...,X_n)$ con $n \in \mathbb{N}$ de una variable aleatoria lineal X, un estimador de la media poblacional definida en la sección 2.1 se corresponde con la expresión siguiente:

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i. \tag{2.5}$$

Si ahora realizamos este cálculo con cada uno de los datos que se muestran en las figuras 2.2a y 2.2b se tiene que: $\overline{x}_a = 185°$ y $\overline{x}_b = 145°$. Se observa que al representar estos valores en

la circunferencia con sus respectivas referencias, la dirección de las medias no coincide. Como veremos más adelante, la definición de la media muestral para los datos circulares hace que esta dirección sí coincida en ambos casos. Por tanto, la media muestral calculada mediante la expresión anterior depende claramente de la dirección cero elegida, además del sentido de rotación.



(a) Dirección cero: Sur. Sentido antihorario.

(b) Dirección cero: Este. Sentido horario.

Figura 2.2: Ejemplo de tres datos circulares medidos desde direcciones y en sentidos diferentes: en el ejemplo (a) tomamos el sur como dirección cero y el sentido antihorario, mientras que en el ejemplo (b) el este es la dirección cero y el sentido es horario. En las dos circunferencias se representan los mismos datos señalados mediantes flechas, el origen y sentido de cada ejemplo en azul y sus respectivas medias muestrales, calculadas mediante la ecuación (2.5), en verde. En las rectas inferiores se representan estos datos (puntos negros) en sendas rectas reales junto con sus medias (en verde).

Tras este ejemplo es evidente la necesidad de nuevos métodos para el estudio de los datos circulares. En las secciones 3.1 y 3.3 se introducen medidas de posición y dispersión específicas para el caso circular. En particular, se muestra cómo corregir el problema anterior de la media.

2.3. Funciones esenciales de la estadística circular

En el siguiente capítulo definiremos los parámetros poblacionales más relevantes en el estudio de variables circulares. Para ello, será necesario conocer su función de distribución, así como su función de probabilidad (en caso de que sea discreta) o de densidad (si es continua).

Así pues, en esta sección se introducirán las definiciones y las características generales de dichas funciones, junto con la función característica asociada.

Definición 2.5. *(Pewsey y Ruxton, 2013) Sea Θ una variable aleatoria circular medida en radianes. Su función de distribución, F, es una función dada por:*

$$F(\theta) = P(0 < \Theta \leq \theta), \quad \forall \theta \in [0, 2\pi), \tag{2.6}$$

y

$$F(\theta + 2\pi) - F(\theta) = 1, \quad -\infty < \theta < \infty, \tag{2.7}$$

donde $P(0 < \Theta \leq \theta)$ denota la probabilidad de que la variable aleatoria Θ tome valores entre 0 y θ.

Además, la función de distribución cumple que $F(0) = 0$ y $F(2\pi) = 1$, por definición, y es continua por la derecha.

Esta definición presenta diferencias importantes con la que se tiene para variables lineales (definición 2.1). La expresión (2.6) no es totalmente análoga al caso de variables lineales: la función de distribución de estas últimas representa la probabilidad de que tomen valores menores o iguales a uno concreto, mientras que para circulares, se corresponde con la probabilidad de que la variable tome un valor entre el origen fijado y un ángulo θ dado.

Por otro lado, la condición (2.7) se impone para garantizar la periodicidad de las distribuciones circulares. Además, este término puede entenderse como que la probabilidad de que la variable aleatoria Θ tome un valor θ en cualquier intervalo de longitud 2π es 1. Esto último recuerda a la condición (2.1) de la definición de función de distribución lineal 2.1. Sin embargo, para variables aleatorias circulares se tiene que:

$$\lim_{\theta \to -\infty} F(\theta) = -\infty \quad y \quad \lim_{\theta \to \infty} F(\theta) = \infty.$$

Esto implica que los valores que toma $F(\theta)$ no sean probabilidades en general, aunque sí lo son en el intervalo [0,2π), como se muestra en la definición 2.5, que es el que nos interesa. Por tanto, para dos valores α y β tales que $0 \leq \alpha \leq \theta \leq \beta \leq 2\pi$ se tiene que

$$P(\alpha < \Theta \leq \beta) = F(\beta) - F(\alpha) = \int_{\alpha}^{\beta} dF(\theta).$$

Si la función de distribución es absolutamente continua, implica que la variable Θ es continua y es posible definir una función de densidad. En el caso de que Θ sea discreta, se obtendría una función de probabilidad. Sin embargo, en este trabajo nos centraremos en el caso continuo.

Definición 2.6. *(Jammalamadaka y SenGupta, 2001) Sea Θ una variable aleatoria circular continua medida en radianes y definida en el soporte [0,2π). La función de densidad, f, es una función que verifica:*

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

Analicemos cada punto de la definición anterior para observar las diferencias y similitudes entre la función de densidad para variables circulares y no circulares. El primer punto indica que esta debe ser positiva en todo el soporte. Esta característica es común a ambos tipos de variables como se puede ver al comparar esta con la definición 2.3 para variables aleatorias lineales. Así mismo, también comparten el segundo punto de que ambas integran uno en sus respectivos soportes. Sin embargo, la tercera propiedad, que explica que la función de densidad es 2π-periódica, solo se impone para variables circulares.

Además, al igual que en el caso de variables lineales (definición 2.3), si F y f son las correspondientes funciones de distribución y densidad de una variable aleatoria circular continua Θ, respectivamente, entonces dado $\theta \in [0, 2\pi)$ se tiene que:

$$f(\theta) = \frac{d}{d\theta}F(\theta) \iff F(\theta) = \int_0^{\theta} f(\varphi)d\varphi.$$

Hablaremos de estas variables continuas en el capítulo 4, que dedicaremos a estudiar las distribuciones circulares más importantes.

Definimos por último la función característica en el caso circular.

Definición 2.7. *(Jammalamadaka y SenGupta, 2001) Sea Θ una variable aleatoria circular medida en radianes y con soporte [0,2π). Se define la función característica, para $t \in \mathbb{R}$, como*

$$\varphi_\theta(t) \coloneqq \mathbb{E}(e^{it\Theta}) \tag{2.8}$$

Notemos que esta definición para variables circulares no difiere dada por la expresión (2.3) para variables lineales. Sin embargo, para que se verifique la propiedad 3 de la definición 2.6 de la función de densidad circular, la función característica para este tipo de datos debe cumplir además que

$$\varphi_\theta(t) \coloneqq \mathbb{E}(e^{it\Theta}) = \mathbb{E}(e^{it(\Theta+2\pi)}) = e^{it(2\pi k)}\varphi_\theta(t),$$

lo que ocurre cuando $\varphi_\theta(t) = 0$ o $t \in \mathbb{Z}$.

Esto hace que solo utilicemos la función característica cuando $t \in \mathbb{Z}$, a diferencia de lo que sucede en el caso lineal en el que t puede tomar cualquier valor real.

De esta forma, cuando $t \in \mathbb{Z}$, podemos definir los momentos trigonométricos como sigue:

$$\begin{aligned}\varphi_\theta(k) \coloneqq \mathbb{E}(e^{ik\Theta}) &= \int_0^{2\pi} e^{ik\theta} f(\theta) d\theta = \int_0^{2\pi} \cos(k\theta) f(\theta) d\theta + i \int_0^{2\pi} \operatorname{sen}(k\theta) f(\theta) d\theta \\ &= \mathbb{E}(\cos(k\Theta)) + i\mathbb{E}(\operatorname{sen}(k\Theta)),\end{aligned} \tag{2.9}$$

para todo $k \in \mathbb{Z}$.

Estos momentos serán útiles en el capítulo 3 para definir nuevas medidas de posición.

Capítulo 3

Estadística descriptiva circular

La estadística descriptiva trata de representar o resumir las características más importantes de los datos muestrales tomados. Para ello se requiere la utilización de gráficos y ciertas medidas que recogen la información principal que contienen dichos datos (Mendenhall *et al.*, 2010). Estas medidas pretenden concentrar la información de las observaciones en unos pocos valores representativos y nos permiten tener una idea general de como se distribuyen los datos (Gorgas *et al.*, 2011).

Las medidas de posición determinan valores que ocupan ciertas localizaciones (centrales o no) en la muestra. Para saber si estos son realmente representativos en la muestra, se utilizan las medidas de escala que indican la dispersión, es decir, la variabilidad de los datos observados en torno a dichos valores.

Así, queda patente la importancia de estos parámetros. En lo que sigue veremos cómo calcular algunas medidas de posición y escala básicas en estadística circular. En particular, en la sección 3.1, se definirá la media, además de explicar brevemente la mediana, la moda y los cuantiles (epígrafe 3.2), como ejemplos de las primeras. Aquellas que sustituyen a la varianza y a la desviación típica en el caso de datos circulares, como ejemplos de las segundas, se verán en la sección 3.3. Por último, en 3.4, se mostrarán algunas representaciones gráficas válidas para este tipo de datos.

En relación a los resultados teóricos que se introducen a continuación, se sigue denotando la dirección cero por 0 y se asume que las variables angulares se miden en radianes. En cambio, en los ejemplos propuestos, si no se especifica otra cosa, trabajamos con muestras en grados (para facilitar la interpretación visual), fijamos la dirección cero en el este y el sentido de lectura de los datos antihorario.

3.1. Medidas de posición: la media

En esta sección daremos la definición de media poblacional de una variable aleatoria circular. Además, también se introduce una estimación de la misma que proporciona resultados adecuados en el contexto circular.

Aplicaremos esta última definición al ejemplo de la figura 2.2 para comprobar que, en efecto, se han corregido los resultados erróneos que habíamos obtenido, consiguiendo ahora unicidad en la estimación sin depender del origen o sentido de giro.

3.1.1. Definición de la media poblacional

Comencemos observando que si Θ es una variable aleatoria circular continua, medida en radianes y que toma valores en el soporte $[0,2\pi)$ sobre el círculo unidad, parece más sencillo utilizar coordenadas polares o representar las observaciones como puntos en el plano complejo, pudiendo utilizar entonces su forma exponencial[1]. Así pues, la esperanza o media poblacional de Θ se obtiene del siguiente modo:

$$\begin{aligned}\mathbb{E}(\Theta) \coloneqq \mathbb{E}(e^{i\Theta}) &= \int_0^{2\pi} e^{i\theta} f(\theta) d\theta = \int_0^{2\pi} \cos f(\theta) d\theta + i \int_0^{2\pi} \sin \theta f(\theta) d\theta \\ &= \mathbb{E}(\cos\Theta) + i\mathbb{E}(\sin\Theta).\end{aligned} \tag{3.1}$$

Si denotamos por:

$$S \coloneqq \mathbb{E}(\sin\Theta) \qquad\qquad C \coloneqq \mathbb{E}(\cos\Theta), \tag{3.2}$$

entonces la dirección media poblacional se corresponde con la siguiente expresión (Pewsey y Ruxton, 2013):

$$\mathbb{E}(\Theta) \coloneqq \arctan^* \frac{S}{C} = \begin{cases} \arctan(S/C) & si \quad C > 0, S \geq 0, \\ \arctan(S/C) + \pi & si \quad C < 0, \\ \arctan(S/C) + 2\pi & si \quad C > 0, S < 0, \\ \pi/2 & si \quad C = 0, S > 0, \\ 3\pi/2 & si \quad C = 0, S < 0, \\ \text{indefinido} & si \quad C = 0, S = 0. \end{cases} \tag{3.3}$$

La definición 3.3 se hace algo complicada debido a que la inversa de la función tangente (denotada en la definición anterior como "arctan") se debe definir en un cuadrante específico.

[1] $\forall z \in \mathbb{C}$, $z = Re^{i\theta} = R(\cos(\theta) + i\sin(\theta))$, donde $R > 0$, la coordenada radial, toma el valor 1 pues se trabaja en el círculo unidad y θ es el ángulo correspondiente a cada observación.

En este caso, habíamos elegido [0,2π). Si se hubiese escogido otro intervalo, la definición anterior cambiaría.

Notemos que aunque estamos asumiendo que la variable aleatoria Θ es continua, también es posible definir la esperanza en el caso discreto a partir de la función característica dada en la definición 2.8. Evaluando la expresión de esta dada en dicha definición en $t = 1$ obtendríamos

$$\varphi_\theta(1) = \mathbb{E}(e^{i\Theta}),$$

es decir, la dirección media de la variable Θ.

3.1.2. Construcción de la media muestral

Sea Θ una variable aleatoria circular medida en radianes y $(\theta_1,...,\theta_n)$ una muestra de observaciones de Θ de tamaño n$\in \mathbb{N}$. Para cada $j = 1, ..., n$, θ_j representa un ángulo en el intervalo $[0, 2\pi)$ sobre el círculo unidad, que se pueden considerar un punto en el plano complejo $e^{i\theta_j} = \cos\theta_j + i \sin\theta_j$, como se ha mencionado en la sección 3.1.1 anterior.

Definimos entonces las siguientes magnitudes:

$$\overline{S} := \frac{1}{n}\sum_{j=1}^{n} \operatorname{sen}\theta_j \qquad\qquad \overline{C} := \frac{1}{n}\sum_{j=1}^{n} \cos\theta_j \tag{3.4}$$

Definición 3.1. *La dirección media muestral se define en Jammalamadaka y SenGupta (2001) como* $\bar{\Theta} = \arg\{\overline{C} + i\overline{S}\}$[2] *o equivalentemente como:*

$$\overline{\Theta} = \arctan^* \frac{\overline{S}}{\overline{C}} = \begin{cases} \arctan(\overline{S}/\overline{C}) & si \quad \overline{C} > 0, \overline{S} \geq 0, \\ \arctan(\overline{S}/\overline{C}) + \pi & si \quad \overline{C} < 0 \\ \arctan(\overline{S}/\overline{C}) + 2\pi & si \quad \overline{C} > 0, \overline{S} < 0, \\ \pi/2 & si \quad \overline{C} = 0, \overline{S} > 0, \\ 3\pi/2 & si \quad \overline{C} = 0, \overline{S} < 0, \\ \textit{indefinido} & si \quad \overline{C} = 0, \overline{S} = 0. \end{cases} \tag{3.5}$$

Una interpretación algo más intuitiva de este valor es que el operador arctan* asigna al par $(\overline{S}, \overline{C})$ el ángulo que forma el vector que une el origen de coordenadas con el punto del círculo $(\overline{S}, \overline{C})$ con el eje de abscisas. Es decir, este ángulo es la dirección media muestral.

Volvamos ahora al ejemplo de la figura 2.2 y calculemos nuevamente la media muestral, esta vez utilizando la ecuación (3.5):

[2]Hemos denotado por "arg" al argumento o ángulo con el que se define un número complejo en su forma polar.

a. $\overline{C} = -0.053$, $\overline{S} = 0.264 \implies \overline{\theta}_a = 101.334°$

b. $\overline{C} = 0.264$, $\overline{S} = -0.053 \implies \overline{\theta}_b = 348.666°$

Aunque estas medias puedan parecer diferentes, en realidad se trata del mismo ángulo, ya que se debe tener en cuenta que estos se miden desde direcciones cero y sentidos de rotación distintos (véase figura 3.1). Recordemos que esto no fue lo que obtuvimos cuando calculamos la media muestral de estos datos utilizando la definición para muestras lineales: en la imagen 2.2 veíamos, al igual que ahora, medias distintas, pero no coincidían sus direcciones sobre la circunferencia.

El motivo de que esta nueva definición proporcione resultados adecuados es que la forma en que está construida la media muestral hace que sea invariante ante translaciones de las observaciones recogidas. Esto también incluye los cambios en el sentido de lectura de los datos (Jammalamadaka y SenGupta, 2001).

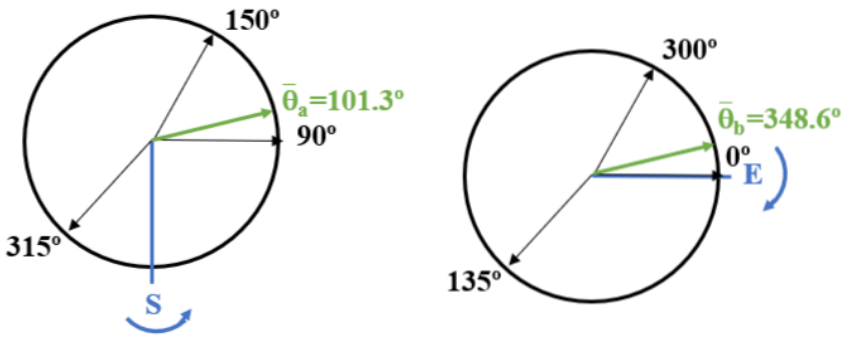


Figura 3.1: Corrección de la media muestral para los datos de la figura 2.2. En las dos imágenes se representan dichos datos señalados mediantes flechas, el origen y sentido de cada ejemplo en azul y sus respectivas medias, calculadas mediante la ecuación (3.1), en verde.

3.2. Otras medidas de posición

Al igual que para variables no circulares, también se pueden definir otras medidas de posición como pueden ser la mediana, los cuantiles y la moda. Por tanto, en esta sección vamos a introducir las definiciones de estas medidas para el caso circular.

Si X es una variable aleatoria lineal definida en un soporte Ω con función de distribución $\mathcal{F}$, la mediana poblacional es un valor $x \in \mathbb{R}$ tal que $\mathcal{F}(x) = P(X \leq x) \geq 0.5$. Es decir, es el punto del soporte en el que se acumula, al menos, el 50 % de la probabilidad. Sin embargo, en el caso de variables circulares, la mediana se define de una forma un tanto diferente.

Definición 3.2. *(Pewsey y Ruxton, 2013) Sea Θ una variable aleatoria circular medida en radianes y que toma valores en el soporte [0,2π). La mediana poblacional para variables circulares se define como aquel ángulo ψ que acumula la mitad de la probabilidad en el intervalo [ψ,ψ + π) mód 2π y, además, minimiza $\mathbb{E}(\pi - |\pi - |\Theta - \psi||)$.*

Notemos que para definir la mediana en el caso circular no basta únicamente con garantizar la acumulación del 50 % de la probabilidad, sino que se debe añadir una condición en la que se busca asegurar que la distancia de los puntos a la mediana sea mínima. Entendemos por distancia entre dos puntos como la longitud del arco que los une, esto es, dados dos ángulos $\alpha, \beta \in [0, 2\pi)$, la expresión

$$\pi - |\pi - |\alpha - \beta||$$

devuelve la distancia entre ellos.

Por otro lado, se puede estudiar la mediana muestral. Esta fue introducida por primera vez en Mardia (1972) y se puede ver también en Fisher (1993). Se trata de aquel ángulo $\tilde{\theta}$ que deja la mitad de los datos en el arco $[\tilde{\theta},\tilde{\theta} + \pi)$ y minimiza la siguiente expresión:

$$d(\tilde{\theta}) = \frac{1}{n}\sum_{j=1}^{n}\{\pi - |\pi - |\theta_j - \tilde{\theta}||\}. \tag{3.6}$$

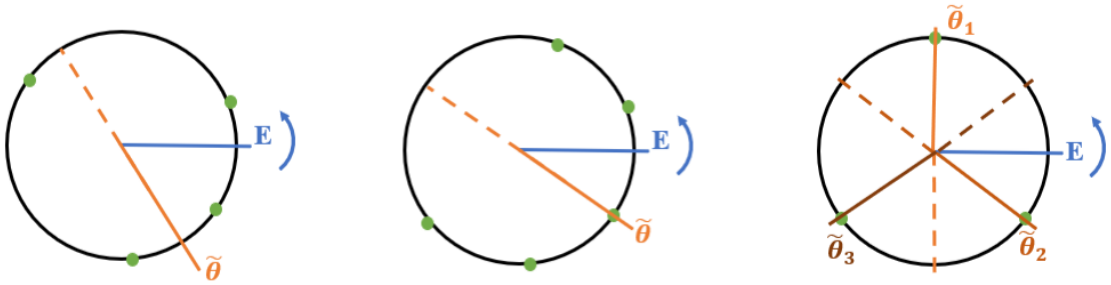
De esta forma, como se puede ver en Pewsey y Ruxton (2013), podemos encontrar una definición de la mediana muestral equivalente a la anterior y algo más práctica. Se trata de sustituir el hecho de tomar el ángulo $\tilde{\theta}$ que minimiza (3.6) por la exigencia de que la mayoría de las observaciones estén más cerca de $\tilde{\theta}$ que de $\tilde{\theta} + \pi$ (la antimediana).

Si la muestra $(\theta_1, ..., \theta_n)$ es de tamaño $n \in \mathbb{N}$ impar, la mediana coincidirá con uno de los puntos de esta, mientras que si n es par, se situará entre dos ángulos de la muestra. Además, la ecuación (3.6) nos asegura que este parámetro es único para datos unimodales (Fisher, 1993), tal y como se observa en los ejemplos 3.2a y 3.2b, para muestras de tamaño $n = 4$ y $n = 5$, respectivamente. En estas figuras se verifica también las observaciones muestrales se encuentran más cerca de sus respectivas medianas que de el ángulo diametralmente opuesto a esta.

Sin embargo, no podemos garantizar la unicidad si la muestra es multimodal o isotrópica[3]. Un ejemplo de este último caso puede verse en la figura 3.2c donde, efectivamente, se tienen tres medianas distintas.

[3]Una muestra es isotrópica cuando sus datos se distribuyen de forma equidistante en torno al círculo.

Esto último es una clara diferencia con la mediana muestral definida para muestras procedentes de variables aleatorias lineales. Ya que, aunque su cálculo también depende del número de datos considerado, n: tras ordenar la muestra, la mediana será el valor central si n es impar y un valor en medio de los dos valores centrales si n es par; siendo siempre este valor único.



(a) La mediana muestral es única (datos pares unimodales).

(b) La mediana muestral es única (datos impares unimodales).

(c) La mediana muestral no es única (datos isotrópicos).

Figura 3.2: Ejemplo de la unicidad y no unicidad de la mediana muestral en datos circulares. Se representan con puntos verdes las observaciones muestrales; en azul, la dirección cero y el sentido de lectura de los datos y en naranja, las medianas muestrales.

También se podría dar una definición general para los cuantiles de una distribución circular.

Definición 3.3. *(Pewsey y Ruxton, 2013) Sea Θ una variable aleatoria circular medida en radianes con función de distribución F. Se define su función cuantil Q, para todo $0 < p < 1$, como*

$$Q(p) \coloneqq \text{ínf}\{\theta : F(\theta) \geq p\},$$

i.e., el valor mínimo de $\theta \in [0, 2\pi)$ tal que $F(\theta) = P(0 < \Theta \leq \theta) \geq p$.
Así mismo, en el caso de que Θ sea continua, se puede obtener el cuantil de orden p $(0 \leq p \leq 1)$ utilizando la función densidad f (Fisher, 1993) como el valor q_p tal que

$$\int_{\tilde{\theta}-\pi}^{q_p} f(\theta)d\theta = p, \tag{3.7}$$

donde $\tilde{\theta}$ es la mediana poblacional.

Es decir, el cuantil de orden p es aquel valor q_p que hace que el arco de extremos $\tilde{\theta} - \pi$ y q_p acumule una probabilidad p. Como la mediana $\tilde{\theta}$ no es necesariamente única, q_p tampoco tiene por qué serlo.

Cabe mencionar que no se dispone de una forma concreta para estimar los cuantiles muestrales, al contrario de lo que ocurre en el caso lineal. Una posible manera de obtenerlos podría

ser estimando la función de distribución (o de densidad en el caso continuo) de los datos y trabajar empleando dicha estimación.

Atendiendo a esta última definición (3.7) y teniendo en cuenta que la mediana $\tilde{\theta}$ es el cuantil 0.5, en el caso continuo, esta se puede expresar como

$$\int_{\tilde{\theta}-\pi}^{\tilde{\theta}} f(\theta)d\theta = 0.5.$$

En cuanto a la moda poblacional $\check{\theta}$ de una variable aleatoria circular Θ, esta será el ángulo de mayor probabilidad, si es discreta, o el máximo de su función de densidad si Θ es continua. Es claro que la unicidad de la moda no está garantizada ya que podrían existir ángulos distintos con la misma probabilidad, en el primer caso, o funciones de densidad con varios máximos, en el segundo. Notemos que esto también ocurre en variables aleatorias lineales.

Por otro lado, no resulta sencillo obtener una expresión de la moda muestral. Una forma de calcularla sería determinar una estimación de la función de densidad y maximizarla (Fisher, 1993). Esta dificultad también está presente principalmente en variables lineales continuas, cuando no se conoce ni su función de densidad ni el número de modas de la distribución.

También es posible definir otras medidas de posición partiendo de los momentos trigonométricos definidos en la sección 2.3. Fijémonos en que al tomar $k = 1$ en la expresión de estos momentos trigonométricos (2.9) recuperamos la ecuación (3.1), es decir, la media poblacional. Además, al fijar $k = 2$, tenemos $\varphi_\theta^{(2)} = \mathbb{E}(e^{i2\Theta}) = \mathbb{E}(\Theta^2)$; con $k = 3$, ($\varphi_\theta^{(3)} = \mathbb{E}(e^{i3\Theta}) = \mathbb{E}(\Theta^3)$, etc. De forma general, $\varphi_\theta^{(k)} = \mathbb{E}(e^{ik\Theta}) = \mathbb{E}(\Theta^k)$, con $k \in \mathbb{Z}$.

3.3. Medidas de dispersión

Al igual que sucede con las medidas de posición, debemos buscar nuevas formas de definir medidas de dispersión válidas para variables circulares. Esto se debe a que las técnicas utilizadas en el contexto lineal a las que estamos acostumbrados vuelven a proporcionar resultados con poco sentido para el caso circular. Veamos esto en el ejemplo que sigue.

Volvamos a retomar el ejemplo de la figura 2.2 (sección 2.2) y veamos qué sucede si calculamos la varianza y la desviación típica muestrales para variables lineales. Obtenemos

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 \qquad \text{y} \qquad \hat{\sigma} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2},$$

respectivamente, donde (X_1,...,X_n) es una muestra aleatoria simple de tamaño n$\in \mathbb{N}$ de una variable aleatoria X.

Recordemos que las dos muestras con las que trabajamos venían dadas por los valores $\vec{\theta}_a = (90°, 150°, 315°)$ y $\vec{\theta}_b = (0°, 135°, 300°)$, con sus respectivas medias $\overline{\theta}_a = 101.334°$ y

$\overline{\theta}_b = 348.666°$. Aplicando las fórmulas vistas para el caso lineal se obtienen los siguientes valores de la varianza y desviación típica muestrales:

a. $\hat{\sigma}_a^2 = 9050$, $\hat{\sigma}_a = 95.131$

b. $\hat{\sigma}_b^2 = 15050$, $\hat{\sigma}_b = 122.678$.

Se observa que los resultados obtenidos al variar la dirección cero y el sentido de lectura de los datos no son los mismos. Comentamos al principio del documento que un estimador adecuado en el caso circular no debe depender de estas elecciones. El ejemplo relacionado con la media muestral visto en el capitulo 2 dejaba constancia de este hecho. En este nuevo ejemplo se observa que, para los mismos datos pero medidos con diferentes dirección cero y sentido de giro, se obtienen distintos valores de dispersión. De esta forma, se tiene que el uso de los estimadores de la dispersión lineales no son adecuados para el caso circular, ya que no se verifica la independencia de los resultados con dicho origen y sentido.

3.3.1. La longitud media resultante y la varianza circular poblacionales

Al igual que se define una dirección media o ángulo medio, se puede definir la longitud media del vector resultante poblacional R como se muestra en Jammalamadaka y SenGupta (2001):

$$\vec{R} \coloneqq (C, S) \implies R \coloneqq \|\vec{R}\| = \sqrt{C^2 + S^2}, \tag{3.8}$$

con C y S dados en (3.2). Es claro que $R \in [0, 1]$.

Esta es una medida de concentración, es decir, mide cómo de condensada está la probabilidad en torno a la dirección media. Cuando $R = 1$, se trata de una distribución degenerada, esto es, toda la probabilidad está concentrada en la dirección media. Por otra parte, si $R = 0$, estaremos en el caso de una distribución uniforme en todas las direcciones.

Por tanto, hablar de concentración lleva a la idea de ver cómo de alejada está la distribución de los datos con respecto a la uniforme. Es interesante este hecho pues proporciona información sobre si dicha distribución posee o no una dirección que acumula mayor probabilidad o si, por el contrario, todas las direcciones son igualmente probables. Así, un valor de R cercano a uno implica una baja dispersión, es decir, que existe una cierta acumulación de dichos datos en torno a la dirección media y no están uniformemente distribuidos a lo largo de la circunferencia. Por ello, según Fisher (1993), se define la varianza circular de la siguiente forma

$$V \coloneqq 1 - R. \tag{3.9}$$

Así, cuando la concentración es alta, la varianza será pequeña. Como $0 \leq R \leq 1$, se sigue que $V \in [0, 1]$. Esta medida de dispersión se introduce con el objetivo de poder hacer una

comparación con el caso lineal , ya que también ocurre que cuando la varianza (dada por la expresión 2.2) es pequeña, hay una cierta acumulación de la probabilidad en torno a la media de la distribución.

Al igual que ocurre con las variables aleatorias lineales, se define la desviación típica circular utilizando la varianza. Sin embargo, en Mardia y Jupp (2000), se muestra que esta no se considera de la forma $\sqrt{V}$, sino como

$$\sigma := \sqrt{-2\log(1-V)} = \sqrt{-2\log R}. \tag{3.10}$$

El motivo de que esta sea la forma de definir la desviación típica es que, así definida, la σ^2 resultante coincide con la varianza de una distribución circular específica, la distribución normal enrollada $\mathcal{WN}(\mu, \rho)$ (Mardia y Jupp, 2000). Sin embargo, aunque esta sea la motivación de esta definición, es válida para todas las muestras de datos, no solo aquellos que siguen la distribución citada, la cual estudiaremos más adelante en el capítulo 4.

Como veremos en la siguiente sección, existen varias definiciones para la versión muestral de la varianza y, consecuentemente, para la desviación típica asociada. Por ello, las medidas de dispersión pueden resultar algo ambiguas. Es por esto que las medidas de concentración son más utilizadas ya que están bien definidas e indican de cierta forma si los datos están o no agrupados, pues si lo están, su distribución no será uniforme.

3.3.2. La longitud resultante media y la varianza circular muestrales

Se puede definir la versión muestral de la longitud del vector resultante de forma análoga al apartado anterior a través de los coeficientes $\bar{S}$ y $\bar{C}$ dados en (3.4). Es decir,

$$\vec{\overline{R}} := (\overline{C}, \overline{S}) \implies \overline{R} := \|\vec{\overline{R}}\| = \sqrt{\overline{S}^2 + \overline{C}^2}. \tag{3.11}$$

Es obvio que $0 \leq \overline{R} \leq 1$. Además, la interpretación sobre qué ocurre cuando $\overline{R} = 1$ o $\overline{R} = 0$ es análoga a la que se ha visto en el apartado anterior para el parámetro poblacional: cuando $R = 1$ teníamos una distribución degenerada, mientras si $R = 0$ la distribución será uniforme. Sin embargo, aunque el valor que proporciona $\overline{R}$ es una estimación del parámetro R, es complicado establecer un límite en el que poder decir que la distribución de los datos está lo suficientemente alejada de la uniforme, es decir, cuánto se debe alejar el valor $\overline{R}$ de cero para que consideremos que los datos no se distribuyen uniformemente en torno a la circunferencia.

Para comprobar si podemos rechazar la uniformidad de los datos existen tests de uniformidad, como puede ser el test de Kuiper (Stephens, 1970), el cual utilizaremos en el capítulo 6 al aplicarlo a dos muestras reales, pero cuyos detalles superan los objetivos de este trabajo.

Además, del hecho de que $\vec{\overline{R}} = (\overline{C}, \overline{S})$ y $\overline{\Theta} = \arg\{\overline{C} + i\overline{S}\}$, se obtiene que $\overline{C} = \overline{R}\cos\overline{\Theta}$ y $\overline{S} = \overline{R}\sin\overline{\Theta}$. Esto relaciona los parámetros introducidos en (3.4) para la construcción de la media muestral con la concentración.

Así pues, se define un primer estimador de la varianza circular muestral como:

$$\hat{V} \coloneqq 1 - \overline{R}. \tag{3.12}$$

La expresión anterior también se puede obtener utilizando la distancia coseno: si ϕ y ψ son dos ángulos, se define su distancia coseno como $1-\cos(\phi - \psi)$. Así, si Θ es una variable aleatoria circular y $(\theta_1, ..., \theta_n)$ una muestra de esta, se puede definir una nueva medida de dispersión de los datos con respecto a un ángulo ξ de la siguiente forma:

$$D(\xi) \coloneqq \frac{1}{n}\sum_{i=1}^{n}(1 - \cos(\theta_i - \xi)). \tag{3.13}$$

Cuando $\xi = \overline{\theta}$, se tiene el siguiente teorema.

Teorema 3.1. *(Jammalamadaka y SenGupta, 2001) Si $\overline{\theta}$ es la dirección media muestral de $(\theta_1, ..., \theta_n)$, entonces*

$$\sum_{i=1}^{n}\sin(\theta_i - \overline{\theta}) = 0 \tag{3.14}$$

y

$$\frac{1}{n}\sum_{i=1}^{n}\cos(\theta_i - \overline{\theta}) = \overline{R}. \tag{3.15}$$

Dem:

$$\begin{aligned}\sum_{i=1}^{n}\sin(\theta_i - \overline{\theta}) &= \sum_{i=1}^{n}(\sin\theta_i\cos\overline{\theta} - cos\theta_i\sin\overline{\theta}) = nS\cos\overline{\theta} + nC\sin\overline{\theta} \\ &= n(\overline{R}\sin\overline{\theta}\cos\overline{\theta} - \overline{R}\cos\overline{\theta}\sin\overline{\theta}) = 0.\end{aligned}$$

Del mismo modo,

$$\begin{aligned}\sum_{i=1}^{n}\cos(\theta_i - \overline{\theta}) &= \sum_{i=1}^{n}(\cos\theta_i\cos\overline{\theta} + \sin\theta_i\sin\overline{\theta}) = nC\cos\overline{\theta} + nS\sin\overline{\theta} \\ &= n(\overline{R}\cos^2\overline{\theta} + \overline{R}\sin^2\overline{\theta}) = n\overline{R}.\end{aligned}$$

En ambos desarrollos hemos utilizado que $\overline{S} = \overline{R}\sin\overline{\theta}$ y $\overline{C} = \overline{R}\cos\overline{\theta}$. ■

Por tanto, evaluando (3.13) en $\overline{\theta}$ y utilizando (3.15), se obtiene que:

$$D(\overline{\theta}) = \frac{1}{n}\sum_{i=1}^{n}(1 - \cos(\theta_i - \overline{\theta})) = 1 - \overline{R} = \hat{V},$$

siendo n el tamaño muestral. Esto es, así como en el caso lineal disponemos de un estimador de la varianza que consiste en calcular el promedio de las distancias a la media muestral, la expresión anterior representa una idea similar para datos circulares utilizando la distancia coseno.

A través del estimador proporcionado en la expresión (3.12) para la varianza, podemos introducir una versión de la desviación típica muestral asociada, de forma análoga a la poblacional. Esto es

$$\overline{\sigma} := \sqrt{-2\log(1-\hat{V})} = \sqrt{-2\log \overline{R}}. \tag{3.16}$$

Puede verse que $\bar{\sigma} \in [0,\infty)$ y solo se anulará cuando todas las observaciones muestrales coincidan, lo que se traduce en una distribución degenerada. Este hecho también ocurre en el caso lineal. En conclusión, este estimador así definido refleja propiedades del estimador lineal.

No obstante, existen otras formas de definir la varianza en el caso circular. Por ejemplo, en Batschelet (1981), se define la varianza circular muestral como

$$\overline{V} := 2(1-\overline{R}), \tag{3.17}$$

que es asintóticamente equivalente a la varianza muestral para variables lineales. Esto es sencillo de ver a partir de la ecuación (3.15) y teniendo en cuenta además que, para ángulos ϕ pequeños, $\cos\phi \approx 1 - \frac{\phi^2}{2}$. Así pues,

$$2(1-\overline{R}) = \frac{2}{n}\sum_{i=1}^{n}(1-\cos(\theta_i-\overline{\theta})) \simeq \frac{2}{n}\sum_{i=1}^{n}(1-1+\frac{(\theta_i-\overline{\theta})^2}{2}) = \frac{1}{n}\sum_{i=1}^{n}(\theta_i-\overline{\theta})^2 = \hat{\sigma}^2,$$

con $2(1-\overline{R}) \simeq \hat{\sigma}^2 \in [0,2]$ (Jammalamadaka y SenGupta, 2001).

Parece más intuitivo utilizar la primera definición (3.12) pues toma valores entre 0 y 1, mientras que esta última los toma entre 0 y 2 y se introduce por las propiedades asintóticas que acabamos de ver. Por otro lado, ninguna de las dos se asemeja a los casos no circulares en los que la varianza toma valores mayores o iguales que cero en toda la recta real.

Esta nueva idea de varianza permite dar una nueva definición de la desviación típica muestral para el caso circular utilizando la expresión (3.17), de forma que también sea asintóticamente equivalente a la lineal. Esto es, simplemente, tomando la raíz cuadrada de la varianza:

$$\overline{\sigma} = \sqrt{\overline{V}} = \sqrt{2(1-\overline{R})}. \tag{3.18}$$

La diferencia más significativa que se aprecia entre esta nueva propuesta, y la vista en la ecuación (3.16), es que esta primera toma valores entre 0 y $\sqrt{2}$, mientras que la segunda toma valores mayores o iguales a cero, siendo no acotada. Por tanto, la versión mostrada en (3.16) comparte soporte con la desviación típica en el caso lineal.

Veamos pues si estos nuevos parámetros muestrales nos proporcionan los resultados esperados de varianza (utilizando la definición introducida en (3.12)) y desviación típica (definición dada en (3.16)) para el ejemplo que estamos tratando (véase la figura 3.1):

a. $\vec{\theta}_a = (90°, 150°, 315°) \implies \overline{R}_a = 0.269,\ \hat{V}_a = 0.730,\ \overline{\sigma}_a = 1.619.$

b. $\vec{\theta}_b = (0°, 135°, 300°) \implies \overline{R}_b = 0.269,\ \hat{V}_b = 0.730,\ \overline{\sigma}_b = 1.619.$

Atendiendo a los resultados anteriores, queda claro que se han solucionado los problemas que teníamos al utilizar las expresiones de la varianza y desviación típica ya conocidas para variables aleatorias lineales: para los mismos datos medidos desde direcciones cero y sentidos distintos, dichos parámetros deben ser iguales. También hemos calculado $\overline{R}$ en ambos casos y, además de confirmar que son idénticos, vemos que las observaciones no están demasiado concentradas en torno a la dirección media, ya que R no tiene un valor cercano a 1. Cabe mencionar que, por la forma en que están definidas la varianza y la desviación (ecuaciones (3.12) y (3.16), respectivamente), en la que la primera toma valores entre cero y uno y la segunda involucra el logaritmo de $1 - \overline{V}$ (que también será un valor entre cero y uno), es normal que ocurra que la desviación resulte ser mayor que la varianza.

Por otro lado, bien podríamos haber calculado la varianza y la desviación típica utilizando la definición (3.17) y la expresión (3.18) respectivamente. De esta forma se obtendría:

a. $\overline{V}_a = 1.461,\ \overline{\sigma}_a = 1.209.$

b. $\overline{V}_b = 1.461,\ \overline{\sigma}_b = 1.209.$

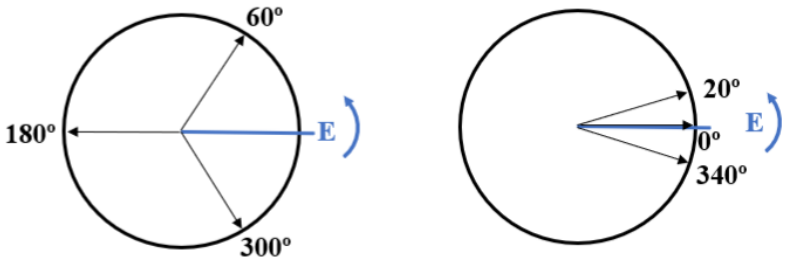
Las definiciones alternativas de la varianza y la desviación típica dadas por las expresiones (3.17) y (3.18), respectivamente, proporcionan resultados que tampoco dependen de la dirección cero o el sentido de lectura de los datos escogidos. Sin embargo, al comparar los valores obtenidos ahora con los anteriores, vemos que la varianza se duplica con respecto a estos, mientras que la desviación es menor. Además, en este caso en este caso el resultado de la varianza es mayor que el de la desviación, que es lo contrario de lo que ocurre en el otro ejemplo. Esto se debe a que $\overline{V} = 2\hat{V}$ y a las muy distintas definiciones de las respectivas desviaciones típicas.

Veamos ahora qué sucede cuando tenemos una muestra en la que las observaciones están lo más alejadas posible unas de otras y otra muestra en la que están juntas. Esto es lo que se observa en las imágenes 3.3a y 3.3b, respectivamente. Calculemos en estos casos los parámetros de concentración y dispersión definidos.

a. $\vec{\theta}_a = (60°, 180°, 300°) \implies \overline{R}_a = 0,\ \overline{V}_a = 1,\ \overline{\sigma}_a = 8.603.$

b. $\vec{\theta}_b = (0°, 20°, 340°) \implies \overline{\theta}_b = 360° = 0°,\ \overline{R}_b = 0.960,\ \overline{V}_b = 0.040,\ \overline{\sigma}_b = 0.286.$

Con esto observamos que cuando las observaciones muestrales están próximas, $\overline{R}$ tiende a uno, mientras que la varianza lo hace a cero. Los datos están concentrados en torno a la media muestral. Ocurre lo contrario cuando los datos son equidistantes y además, atendiendo a la definición 3.5, como $S_a = C_a = 0$, la media muestral no está definida.



(a) (b)

Figura 3.3: Ejemplos de dos muestras donde (a) los datos están lo más alejados posible unos de otros y (b) los datos se encuentran cerca.

3.4. Representaciones gráficas de los datos circulares

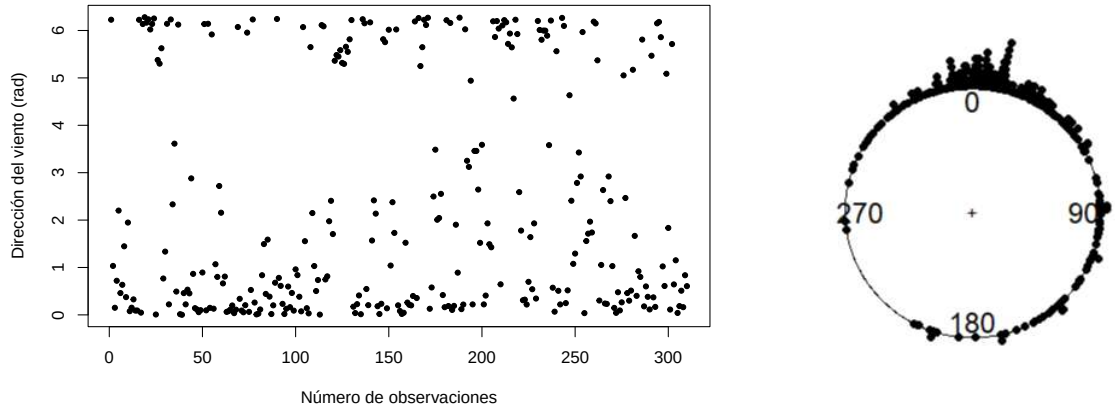
Las representaciones gráficas de los datos muestrales son una potente herramienta descriptiva, muy útil como estudio preliminar al análisis detallado de los datos. Nos permiten resumir información compleja y hacerla más accesible y comprensible. Además, nos permiten conocer en qué rango de valores se mueven las observaciones y de qué forma, permitiendo observar a simple vista características importantes de la muestra, como la existencia o no de modas, o intuir la distribución que podrían seguir los datos.

Al igual que hemos adaptado las definiciones de las medidas de posición y dispersión poblacionales o muestrales, es necesario adecuar los gráficos a las variables circulares, pues hemos visto que estas se representan sobre la circunferencia de radio unidad. Aunque en los ejemplos expuestos anteriormente ya se han utilizado algunas de las representaciones que explicaremos a continuación, vamos a ver cuatro tipos de representaciones. Algunas de ellas tienen una análoga para variables lineales.

3.4.1. Diagramas de puntos

Es una representación bastante sencilla en la que cada observación se representa con un punto sobre la circunferencia unidad. Sin embargo, permiten observar de forma más clara que otros gráficos algunas propiedades de los datos, como detectar pequeños grupos de puntos aislados o algún punto solitario de especial interés (Fisher, 1993).

Esta es la representación utilizada en el ejemplo donde se explicaba la unicidad de la mediana muestral (ver figura 3.2). A continuación veremos un gráfico con más datos y en el que se pueda apreciar su utilidad.



(a) Diagrama de puntos lineal.

(b) Diagrama de puntos circular.

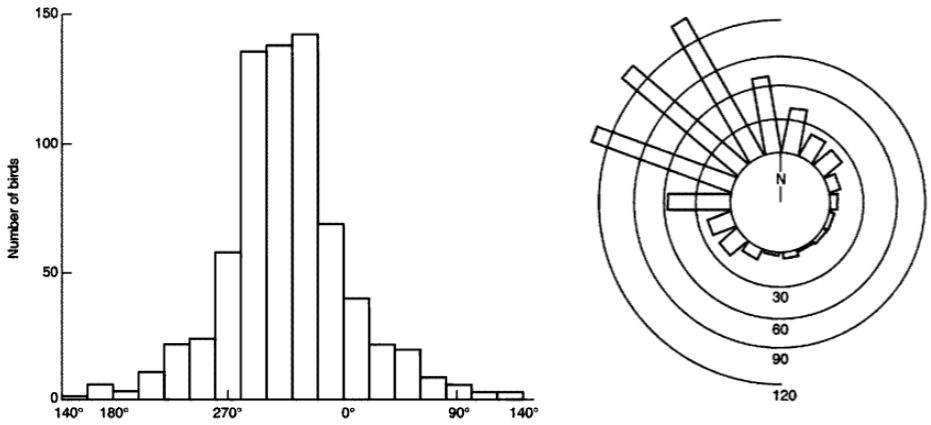
Figura 3.4: Representación de la base de datos "wind" del paquete "circular" de R (Agostinelli y Lund, 2023) (a) sin tener en cuenta que son datos circulares y (b) empleando el diagrama de puntos circular.

En la figura 3.4 vemos dos gráficos en los que se representa la dirección del viento medida en radianes en sentido horario y con el norte como dirección cero. Claramente, al representar la muestra como si fuesen observaciones de una variable aleatoria lineal (diagrama 3.4a), la interpretación de los resultados se presenta complicada. Sin embargo, el diagrama de puntos circular 3.4b facilita en gran parte el estudio de los datos, observándose una clara tendencia del viento hacia el norte.

3.4.2. Histogramas circulares

Los histogramas resultan muy útiles para analizar observaciones de datos lineales continuos. Por ello, parece lógico intentar adaptar este tipo de gráficos al caso circular. Como se muestra en Mardia y Jupp (2000), una forma de obtenerlo es representar un histograma lineal y "envolverlo"

alrededor de la circunferencia unidad (véase la figura 3.5). Las barras se construyen tomando un conjunto de datos como puntos medios de estas y su altura es proporcional al número de observaciones en cada intervalo $[x_0 - h, x_0 + h]$, donde x_0 es el punto medio de cada barra y $h > 0$, un parámetro ventana. La construcción del histograma y su interpretación no paramétrica se explican en profundidad en la sección 5.1. Pueden ser útiles, pero también pueden llevar a resultados erróneos o confusos si no se escogen adecuadamente los cortes considerados en el círculo. Estos se pueden apreciar en la imagen 3.5b. Por ejemplo, cuando la moda es única, una mala disposición de dichos cortes puede hacer que parezca que la distribución es multimodal. Esto mismo ocurre en el caso lineal cuando el histograma no posee suficientes barras. Por el contrario, si se consideran demasiadas barras podemos hacer que aparezcan modas que en realidad no hay. Esto está relacionado con la elección del parámetro de ventana h cuyo efecto estudiaremos también en la sección 5.1.



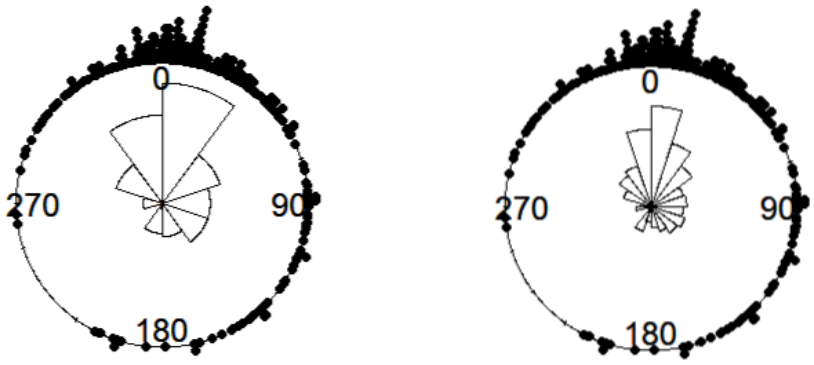
(a) Histograma lineal. (b) Histograma circular.

Figura 3.5: Ejemplos de (a) histograma lineal y (b) circular que representan la dirección de vuelo de un grupo de pájaros (Mardia y Jupp, 2000).

3.4.3. Diagramas de rosa

Otro tipo de representación similar al histograma circular es el diagrama de rosa, en el se sustituyen las barras por sectores circulares. Algunos autores, como Mardia y Jupp (2000), toman el área de cada sector proporcional a la frecuencia del grupo de datos que representa y el radio, a la raíz cuadrada de la frecuencia relativa de dichos datos, aunque no todos siguen este convenio. Al igual que sucede con los histogramas lineales en los que es importante hacer una buena elección de la anchura de las barras, también lo es elegir correctamente el arco de los sectores circulares (normalmente se utiliza 10° o 20°). En caso contrario, la información podría

ser malinterpretada ocultando grupos de datos relevantes. El arco de los sectores depende del parámetro de suavizado ν, que estudiaremos en profundidad en la sección 5.2 y veremos varios métodos para obtener un valor adecuado de este.



(a) Diagrama de rosa con 10 sectores. (b) Diagrama de rosa con 20 sectores.

Figura 3.6: Para este ejemplo de diagrama de rosa se ha utilizado la base de datos "wind" disponible en la librería "circular" de R y hemos empleado (a) 10 sectores circulares y (b) 20 sectores circulares.

Utilicemos los datos de la figura 3.4 y el diagrama de puntos circular para representar los diagramas de rosa de la imagen 3.6. En esta se observan dos gráficos que únicamente se diferencian en el número de sectores circulares que resumen la información de la muestra: en 3.6a se utilizan 10 y el 3.6b, 20. Estos muestran claramente el hecho de que si no se utilizan suficientes sectores, puede que algunos grupos de datos queden ocultos debido a que sus vecinos son más frecuentes como ocurre con los puntos situados en la zona noreste.

Representemos ahora el histograma que utilizaríamos en el caso lineal. Esto es el que tenemos en la figura 3.7. Comparando este gráfico con los de la figura 3.6, vemos que con el primero perdemos parte de la información en el sentido de que no sabemos cuál es la dirección cero ni el sentido de lectura de los datos. Además, ya no se ve a simple vista que estos son periódicos, pues los extremos del histograma no están unidos, como sí ocurre en el diagrama de rosa. Además, en el histograma lineal podría parecer que tenemos dos grupos modales: el primero en 0° y el segundo en 360°. Sin embargo, se trata de un único grupo, tal y como se muestra en el diagrama de rosa debido a la periodicidad.

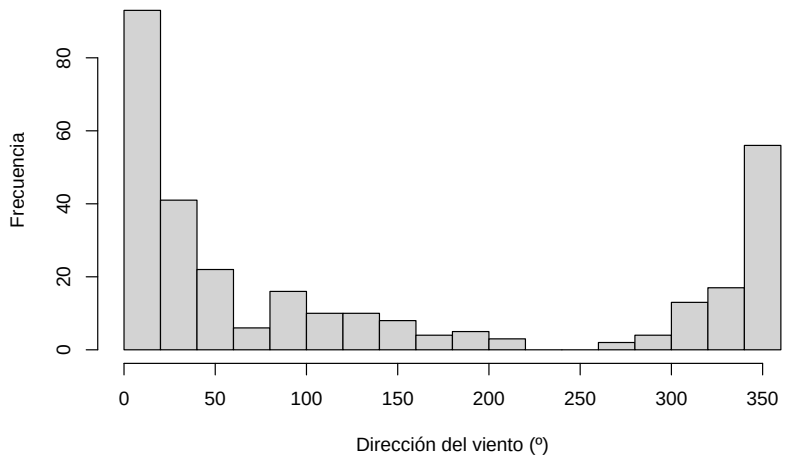


Figura 3.7: Histograma de la dirección que toma el viento entre 0° y 360.

3.4.4. Diagramas de caja

La construcción de una diagrama de cajas para datos circulares resulta algo compleja. No obstante, nosotros veremos la explicación dada en Buttarazzi *et al.* (2018) que se basa en la idea del diagrama de cajas para datos lineales de Tukey (Tukey, 1977). Esta opción utiliza rangos[4] y el concepto de profundidades.

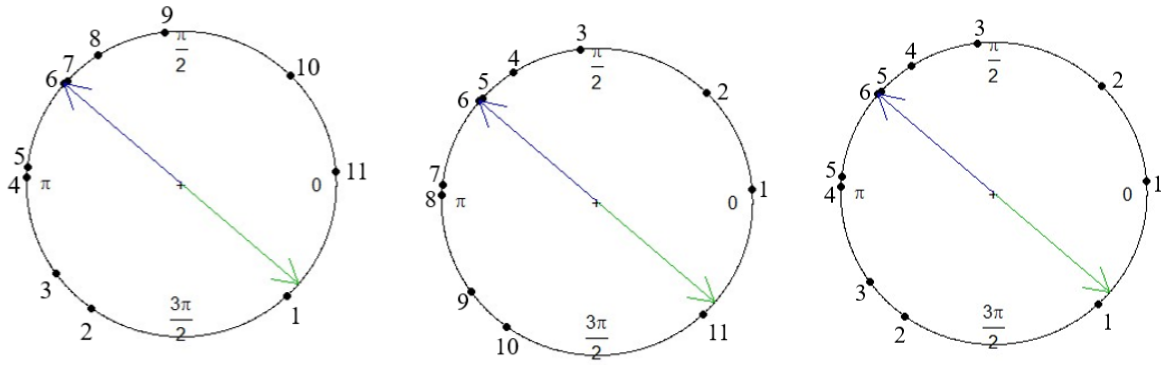
Lo primero que se debe hacer para obtener el diagrama de cajas de una muestra de datos circulares es obtener su mediana. Como ya hemos visto en la sección 3.2, esta no tiene por qué ser única, por ello, se utilizará aquella que minimice (3.6). Posteriormente, se localiza la antimediana, es decir, el punto más alejado de la mediana en el círculo (el que forma 180º con ella). De esta forma, el eje que forman ambas divide los datos en dos.

Tras esto, establecemos los rangos de los datos partiendo de la antimediana en sentido horario y en sentido antihorario (véanse las figuras 3.8a y 3.8b, respectivamente). A continuación, se calcula la profundidad de cada observación tomando el mínimo de ambos rangos, como se puede ver en la imagen 3.8c.

Si la antimediana es un punto de la muestra, los rangos se empiezan a contar a partir de esta. Por ello, la profundidad no está definida para este dato, pero se le da el valor cero.

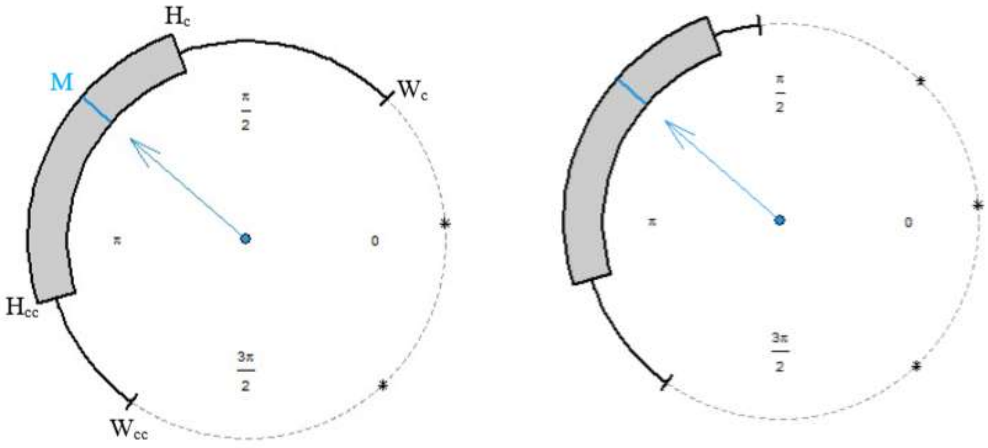
A partir de estos rangos y profundidades se definen los distintos parámetros (H_c, H_{cc}, W_c,

[4]Definimos el rango de una observación muestral como la posición que ocupa en la muestra ordenada



(a) Rangos (horario) (b) Rangos (antihorario) (c) Profundidad

Figura 3.8: En las dos primeras imágenes vemos los rangos muestrales tomados desde la antimediana (flecha verde) en sentido (a) horario y (b) antihorario, respectivamente. En la figura (c) tenemos las profundidades asociadas a los datos. La fecha azul representa la mediana muestral que minimiza 3.6.



(a) MF óptimo (b) MF=0.7

Figura 3.9: Representación de dos diagramas de cajas para los mismos datos. Se toma el valor del parámetro (a) MF óptimo y (b) MF=0.7.

W_{cc}) que están representados en la figura 3.9a. Los coeficientes H_c y H_{cc} se corresponden con el valor[5] $\frac{1+\lfloor(1+n)/2\rfloor}{2}$, con n el tamaño muestral, al obtener los rangos en sentido horario y

[5]Denotamos por $\lfloor x \rfloor$ a la parte entera de x, $\forall x \in \mathbb{R}$.

antihorario, respectivamente. La caja se define en el arco que une ambos puntos. Por otro lado, los "bigotes", W_c y W_{cc}, se obtienen mediante

$$W_c = H_c - MF(H_{cc} - H_c) \quad \textbf{y} \quad W_{cc} = H_{cc} + MF(H_{cc} - H_c),$$

respectivamente; donde MF es un factor de multiplicidad que garantiza que la probabilidad de que uno de los datos esté fuera de los límites que marcan W_c y W_{cc} sea 0.7 %. Estos puntos están representados en las figuras 3.9a y 3.9b con el símbolo *. En el caso lineal se fija MF=1.5 pero, en el circular, MF depende de la concentración de los datos, es decir, está relacionado con la longitud media resultante R (3.8). En el apéndice de Buttarazzi *et al.* (2018) se puede ver un ejemplo de cálculo del valor óptimo de este coeficiente para el caso de la distribución de Von-Mises (distribución que veremos en el capítulo siguiente).

Al representar el diagrama de cajas de datos circulares usando el software R, este toma por defecto el parámetro MF óptimo, que es lo que vemos en la figura 3.9a. Sin embargo, se puede fijar cualquier otro valor. Al tomar MF=0.7 obtenemos el diagrama 3.9b. En ambos observamos la mediana muestral, dibujada en azul. Sin embargo, el "bigote" W_c de la imagen de la derecha es notablemente más corto, lo que hace que haya más observaciones fuera del diagrama.

Capítulo 4

Distribuciones circulares

Tras estudiar los principales parámetros estadísticos y las representaciones gráficas más útiles, veremos ahora las distribuciones de probabilidad definidas para variables aleatorias circulares más importantes.

Por un lado, existen algunas que se desarrollan específicamente para variables circulares. Las más destacadas son la distribución uniforme circular, la cardioide y la de Von-Mises.

Sin embargo, también se pueden adaptar las distribuciones para variables lineales para que sean válidas en el caso circular. Así, estas se pueden "enrollar" alrededor de la circunferencia unidad o, en el caso de que tengamos un vector aleatorio bidimensional, se transforma en su componente angular. Como ejemplo de las primeras veremos la distribución normal enrollada (*Wrapped Normal*) y la Cauchy enrollada (*Wrapped Cauchy*) y, en relación a las segundas, la normal proyectada. Por último, estudiaremos la normal asimétrica (*wrapped skew Normal*) y las mixturas de Von-Mises como ejemplos de modelos asimétricos y/o multimodales.

4.1. Distribuciones circulares propias

4.1.1. Distribución uniforme circular

Esta es la distribución más sencilla. Una variable aleatoria continua circular Θ medida en radianes sigue una distribución uniforme circular si su función de densidad viene dada por

$$f(\theta) = \frac{1}{2\pi}, \qquad 0 \leq \theta < 2\pi, \tag{4.1}$$

cuya representación gráfica puede verse en la figura 4.1 y cuya función de distribución tiene la expresión

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

Se tiene que todas las direcciones acumulan la misma probabilidad y por tanto, no existe un ángulo medio:

$$C = \mathbb{E}(\cos\Theta) = \int_0^{2\pi} f(\theta)\cos\theta = \frac{1}{2\pi}\int_0^{2\pi} \cos\theta = 0$$
$$S = \mathbb{E}(\sin\Theta) = \int_0^{2\pi} f(\theta)\sin\theta = \frac{1}{2\pi}\int_0^{2\pi} \sin\theta = 0.$$

Así, por la definición (3.3), la dirección media no está definida. Por otro lado, la longitud media del vector resultante es R=0 y su desviación típica es $\sigma = \infty$. Vemos ahora el sentido de definir un parámetro que mida la concentración en lugar de la desviación ya que de esta forma habría información de si existe o no una dirección que acumule más datos. En el caso de no haberla, estaríamos ante la distribución uniforme. Fijémonos en la importancia de esta ya que lo que estamos haciendo es comparar cada distribución con la uniforme a través del parámetro R. Veremos una aplicación a un caso real en la sección 6.

Por otra parte, es la única distribución circular invariante bajo rotaciones y reflexiones (Pewsey y Ruxton, 2013).

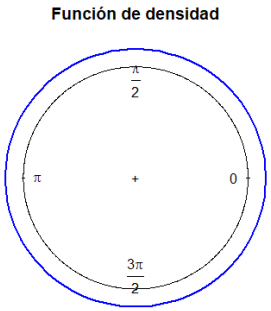


Figura 4.1: Función de densidad de la distribución uniforme circular.

Podemos fijarnos también que las funciones de distribución y densidad dadas por las expresiones (4.1.1) y (4.1), respectivamente, coinciden con las de una variable aleatoria lineal que sigue un distribución uniforme en el intervalo [0,2π).

4.1.2. Distribución cardioide

Una variable aleatoria continua circular Θ, medida en radianes y que toma valores en el soporte [0,2π), sigue una distribución cardioide de parámetros μ y ρ, $\mathcal{C}(\mu,\rho)$, si su función de densidad es

$$f(\theta) = \frac{1}{2\pi}\{1 + 2\rho\cos(\theta-\mu)\}, \tag{4.2}$$

y, por tanto, su función de distribución es

$$F(\theta) = \frac{\rho}{\pi}\sin(\theta-\mu) + \frac{\theta}{2\pi},$$

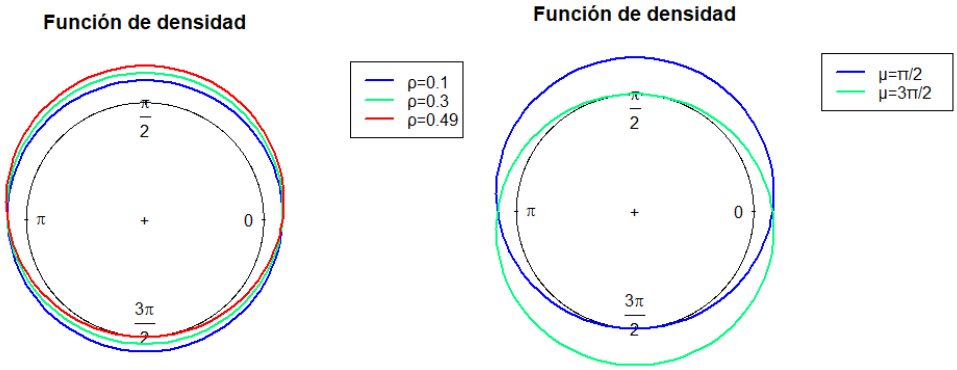
con $0 \leq \mu < 2\pi$, $0 \leq \rho < \frac{1}{2}$ (Fisher, 1993).

La principal idea para introducir esta distribución es conseguir una que sea adecuada para datos con una moda, de forma que sea posible que aparezcan rangos de ángulos que apenas tengan probabilidad pero haciendo que aquellos intervalos más probables acumulen la mayor parte de esta. De esta forma, se obtiene una distribución unimodal y simétrica sin alejarse demasiado de la uniforme.

Por ello, se obtiene añadiendo una perturbación en forma de función coseno a la distribución uniforme circular: cuando $\rho = 0$, la cardioide se reduce a esta última. Además, ρ se corresponde con la longitud media del vector resultante y μ, con la dirección media, la mediana y la única moda. Estas características se pueden apreciar en las figuras 4.2a y 4.2b.

En 4.2a vemos como al aumentar ρ, hay una mayor concentración de la probabilidad en torno a μ, alejándose de la distribución uniforme, lo que concuerda con el hecho de que ρ sea la longitud media resultante y μ sea su media. Además, notemos que la distribución también se desplaza haciendo que haya puntos de la circunferencia que casi no acumulen probabilidad: los ángulos localizados en la parte baja del círculo en la curva representada en rojo en 4.2a tienen menos probabilidad que sus puntos antipodales. Por otro lado, en 4.2b, observamos que al cambiar su media la función de densidad no varía de forma, pero sí cambia su orientación haciendo que el máximo de la curva esté dirigido hacia el ángulo μ.

Otra característica importante, que se puede apreciar en estas dos figuras, es que esta distribución es simétrica con respecto a μ.



(a) $\mathcal{C}(\frac{\pi}{2}, \rho)$ (b) $\mathcal{C}(\mu, 0.49)$

Figura 4.2: Representaciones de la función de densidad de la distribución cardioide variando los parámetros (a) ρ y (b) μ.

Por último, cabe mencionar que las variables aleatorias circulares que siguen esta distribución son reproductivas con respecto a ambos parámetros, es decir, si $\Theta_1 \sim \mathcal{C}(\mu_1, \rho_1)$ y $\Theta_2 \sim \mathcal{C}(\mu_2, \rho_2)$, con Θ_1 y Θ_2 independientes, entonces $\Theta_1 + \Theta_2 \sim \mathcal{C}(\mu_1 + \mu_2, \rho_1\rho_2)$ (Mardia y Jupp, 2000).

4.1.3. Distribución de Von-Mises

Una variable aleatoria circular Θ, medida en radianes y definida en el soporte $[0,2\pi)$, sigue una distribución de Von-Mises de parámetros μ y κ, $\mathcal{M}(\mu,\kappa)$, si su función de densidad es

$$f(\theta) = \frac{1}{2\pi I_0(\kappa)} e^{\kappa \cos(\theta-\mu)}, \tag{4.3}$$

donde $0 \leq \mu < 2\pi$ es la dirección media, $\kappa \geq 0$ se conoce como parámetro de concentración, y $I_0 = \frac{1}{2\pi}\int_0^{2\pi} e^{\kappa\cos\theta} d\theta = \sum_{r=0}^{\infty} \frac{1}{(r!)^2}(\frac{\kappa}{2})^{2r}$ es la función de Bessel modificada de primera especie y orden cero. La longitud media del vector resultante se define en Mardia y Jupp (2000) como $R = \frac{I_1(\kappa)}{I_0(\kappa)}$, con $I_1(\kappa)$ la función de Bessel modificada de primera especie y orden uno.

Surgen dificultades a la hora de calcular la función de distribución debido a la problemática que surge de integrar la función de densidad (4.3) debido a las funciones de Bessel. Por tanto, se recurre a métodos numéricos para evaluarla y estos valores están tabulados para $0 < \kappa \leq 10$, como ocurre con la distribución normal estándar para variables aleatorias lineales. Sin embargo, para valores de κ suficientemente grandes, se puede obtener una aproximación de la función de distribución de Von-Mises a través de la de la normal. Para ver esto, tomamos $\xi = \kappa^{1/2}(\theta - \mu)$ y, utilizando (4.3), vemos que la función de densidad de ξ es proporcional a

$$\exp(-\kappa(1 - \cos(\kappa^{-1/2}\xi))) \simeq \exp(-1/2\xi^2),$$

donde también hemos utilizado que $\cos x \simeq 1 - \frac{x^2}{2}$ cuando $x \to 0$ y, por tanto, $1 - \cos(\kappa^{-1/2}\xi) \simeq 1/2\xi^2\kappa^{-1}$ cuando $\kappa \to \infty$. Así, $\xi \sim \mathcal{N}(0,1)$.

Resumiendo, si $\Theta \sim \mathcal{M}(\mu,\kappa)$, $\kappa^{-1/2}(\Theta - \mu) \sim \mathcal{N}(0,1)$, cuando κ es suficientemente grande. De esta forma, es posible aproximar la función de distribución de Von-Mises con la de una normal estándar. Se puede encontrar más información sobre este hecho en Mardia (1972).

Además, teniendo en cuenta que $I_0(\kappa)$ es una función par y que $\cos(\alpha + \pi) = -\cos\alpha$, es claro que $\mathcal{M}(\mu+\pi,\kappa) = \mathcal{M}(\mu,-\kappa)$.

Al igual que la distribución cardioide, esta también es unimodal y simétrica con respecto a μ. Esto último se aprecia claramente en la imagen 4.3a, en la que se representa en azul la función de densidad de la distribución $\mathcal{M}(\mu,2)$, para $\mu = \frac{\pi}{4}$ y $\mu = \pi$. Además, μ se corresponde también con la mediana y la moda. Podemos comparar las representaciones de la figura 4.3b, para confirma que la concentración depende de κ (hemos visto que la longitud resultante media dependía de κ): al aumentar este parámetro también lo hace la concentración de la probabilidad a la media (que en este caso es 0).

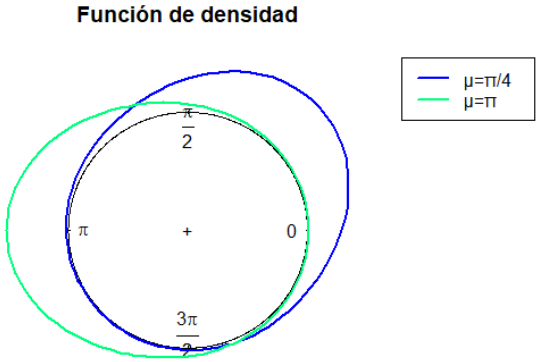
Así mismo, fijémonos en que distribuciones cardiode y de Von-Mises están relacionadas ya

que, teniendo en cuenta que $e^x \simeq 1 + x$ cuando $x \to 0$, tomando $\kappa \to 0$:

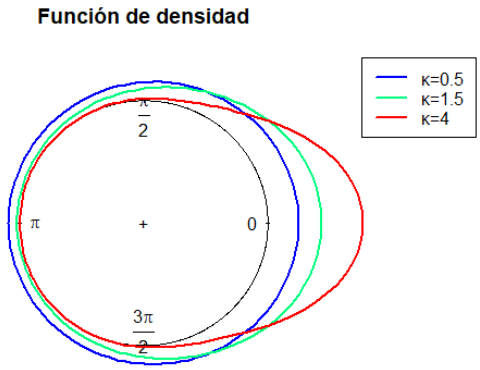
$$f_{\mathcal{M}(\mu,\kappa)}(\theta) = \frac{1}{2\pi I_0(\kappa)} e^{\kappa \cos(\theta-\mu)} \simeq \frac{1}{2\pi}\{1 + \kappa \cos(\theta - \mu) = f_{\mathcal{C}(\mu,\kappa/2)}.$$

En la figura 4.3c, se comparan las funciones de densidad de estas dos distribuciones con $\mu = \pi$ y $\kappa = 0.75$, observando que, efectivamente, se asemejan.

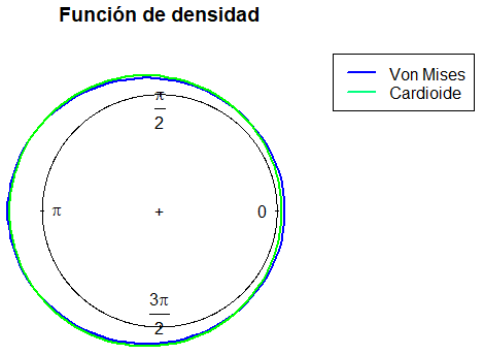
Por otro lado, si κ=0, se observa que la distribución de Von-Mises coincide con la uniforme circular dada por la expresión (4.1).



(a) $\mathcal{M}(\mu, 2)$



(b) $\mathcal{M}(0, \kappa)$



(c) $\mathcal{M}(\pi, 0.75) \sim \mathcal{C}(\pi, 0.375)$

Figura 4.3: Representaciones de la función de densidad de la distribución de Von Mises variando los parámetros (a) μ y (b) κ (superior derecha). En la imagen (c) vemos una comparación de las funciones de densidad de esta y de la cardioide para κ pequeño.

4.2. Distribuciones enrolladas

En esta sección se presenta un método para obtener distribuciones apropiadas para variables aleatorias circulares a partir de las ya conocidas para variables aleatorias lineales, sin más que "enrollarlas" alrededor del soporte de la circunferencia.

Consideramos X una variable aleatoria lineal. Entonces se puede obtener una variable aleatoria circular Θ medida en radianes mediante la transformación: $\Theta = X \mod 2\pi$.

La funciones de distribución $F(\theta)$ y de densidad $f(\theta)$ de Θ se calculan a partir de las de X, que denotamos por $\mathcal{F}(x)$ y f(x) respectivamente, de la siguiente forma:

$$F(\theta) = \sum_{k=-\infty}^{\infty} \{\mathcal{F}(\theta + 2k\pi) - \mathcal{F}(2k\pi)\}, \tag{4.4}$$

$$f(\theta) = \sum_{k=-\infty}^{\infty} \text{f}(\theta + 2k\pi). \tag{4.5}$$

Este tipo de distribuciones presentan una serie de características importantes dignas de mención (Mardia y Jupp, 2000):

I. Si X e Y son variables aleatorias lineales se tiene que

$$(X + Y) \mod 2\pi = X \mod 2\pi + Y \mod 2\pi, \tag{4.6}$$

es decir, no importa si primero sumamos las variables y después, las "enrollamos" o si por el contrario, se "enrollan"primero y luego se suman.

II. Si $\phi(t) = \mathbb{E}(e^{itX})$ es la función característica de una variable aleatoria lineal X y consideramos $\Theta = X \pmod{2\pi}$, su n-ésimo momento trigonométrico $\varphi_\theta^{(n)}$ coincide con $\phi(n)$:

$$\varphi_\theta^{(n)} = \int_0^{2\pi} e^{in\theta} f(\theta) d\theta = \sum_{k=-\infty}^{\infty} \int_{2k\pi}^{2(k+1)\pi} e^{in\theta} f(\theta) d\theta = \int_{-\infty}^{\infty} e^{inx} f(x) dx = \phi(n).$$

III. Si la función $\phi(t)$ del punto anterior es integrable, entonces la función de densidad de Θ puede expresarse como una expansión en serie de potencias de la siguiente forma

$$f(\theta) = \frac{1}{2\pi} \left\{ 1 + 2 \sum_{n=1}^{\infty} (C_n \cos(n\theta) + S_n \operatorname{sen}(n\theta)) \right\}, \tag{4.7}$$

donde $C_n = \mathbb{E}(\cos(nX))$ y $S_n = \mathbb{E}(\sin(nX))$, de forma que $\phi(n) = C_n + iS_n$. Esta forma de representar la función de densidad de Θ puede simplificar la expresión (4.5).

A continuación se presentan un par de ejemplos de distribuciones obtenidas mediante este método: la normal enrollada y la Cauchy enrollada.

4.2.1. Distribución Normal enrollada

Una variable aleatoria $X \sim \mathcal{N}(\mu, \sigma)$ tiene función de densidad

$$\mathrm{f}(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

Por lo tanto, la variable aleatoria circular definida por $\Theta = X \mod 2\pi$ seguirá una distribución normal enrollada $\mathcal{WN}(\mu, \rho)$, y su función de densidad vendrá dada por

$$f(\theta) = \frac{1}{\sigma\sqrt{2\pi}} \sum_{K=-\infty}^{\infty} \exp -\frac{(x - \mu + 2k\pi)^2}{2\sigma^2}. \tag{4.8}$$

Así, $\mu \mod 2\pi$ será la dirección media de Θ y $\rho = e^{-\sigma^2/2}$, la longitud media del vector resultante. En la figura 4.4b vemos cómo al aumentar ρ (disminuir σ) la probabilidad está más concentrada entorno a $\mu \mod 2\pi = 4\pi \mod 2\pi = 0$.

Sin embargo, también se podría obtener otra representación de la función anterior a partir de (4.7). La función característica de X es $\phi(n) = e^{in\mu - n^2\sigma^2/2} = e^{in\mu}\rho^{n^2}\ \forall n \in \mathbb{N}$ y por tanto, $S_n = \rho^{n^2}\sin(n\mu)$ y $C_n = \rho^{n^2}\cos(n\mu)$. Finalmente, sustituyendo en (4.7) y reordenando los términos, obtenemos que

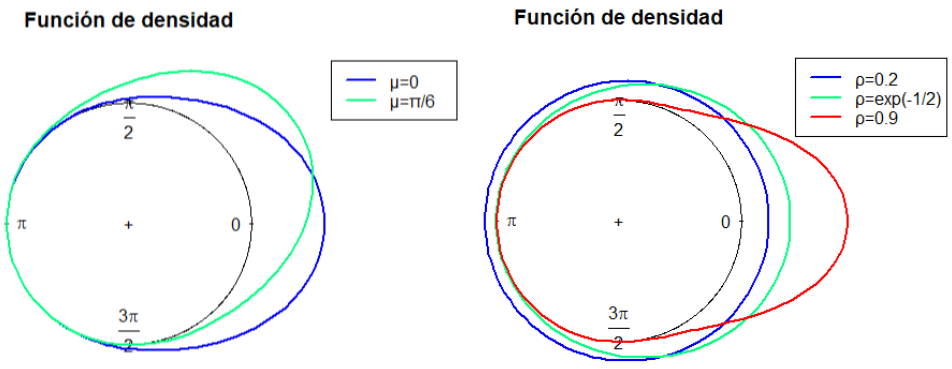
$$f(\theta) = \frac{1}{2\pi}\left\{1 + 2\sum_{n=1}^{\infty} \rho^{n^2} \cos n(\theta - \mu)\right\}.$$

En este caso, esta última expresión no simplifica mucho la que hemos definido primero (4.8).

La distribución normal enrollada también es simétrica y unimodal, con moda $\mu \mod 2\pi$. Podemos observar esto en la figura 4.4a, en la que las funciones de densidad representadas son simétricas con respecto a 0 (curva azul) y $\frac{\pi}{6}$ (curva verde).

Es interesante observar que las funciones de densidad representadas en 4.3b y 4.4b son bastante similares. Sin embargo, el ligero inconveniente que se encuentra en la distribución que estudiamos ahora con respecto a la Von-Mises es que una pequeña variación de ρ provoca grandes cambios en la forma de la función de densidad. Claro que esto puede deberse a que los parámetros de concentración de cada una de ellas pueden tomar valores en soportes distintos. Otra dificultad que puede hacer más conveniente trabajar con la distribución de Von-Mises es la expresión de su densidad, pues es más sencillo trabajar con una expresión reducida en lugar de con una suma infinita de términos.

Así mismo, podemos encontrar alguna relación entre esta distribución y las que hemos visto en el epígrafe anterior. Cuando $\rho \to 0$, $f(\theta) \to \frac{1}{2\pi}$, que es la función de densidad (4.1) de la uniforme circular. Por otro lado, fijémonos en que si $\rho \to 1$, teniendo en cuenta que $\rho = e^{-\sigma^2/2}$, se tiene que $\sigma \to 0$, por lo que la distribución normal $\mathcal{N}(\mu, \sigma)$ de partida será en realidad una



(a) $\mathcal{WN}(\mu, 0.8)$ (b) $\mathcal{WN}(4\pi, \rho)$

Figura 4.4: Representaciones de la función de densidad de la distribución normal enrollada variando los parámetros (a) μ y (b) ρ.

distribución degenerada en el punto μ y la $\mathcal{WN}(\mu, \rho)$ asociada será una distribución degenerada en $\mu \mod 2\pi$.

Además, de (4.6), se verifica que si $\Theta_1 \sim \mathcal{WN}(\mu_1, \rho_1)$ y $\Theta_2 \sim \mathcal{WN}(\mu_2, \rho_2)$ son dos variables aleatorias circulares independientes, entonces $\Theta_1 + \Theta_2 \sim \mathcal{WN}(\mu_1 + \mu_2, \rho_1\rho_2)$.

4.2.2. Distribución de Cauchy enrollada

Una variable aleatoria lineal X sigue una distribución de Cauchy, $\mathcal{C}(\mu, \sigma)$, si su función de densidad de probabilidad es

$$g(x) = \frac{1}{\pi} \frac{\sigma}{\sigma^2 + (x-\mu)^2},$$

con $x, \mu \in (-\infty, \infty)$ y $\sigma > 0$.

Así, la correspondiente variable aleatoria circular Θ, definida como $\Theta = X \mod 2\pi$, tendrá una distribución de Cauchy enrollada $\mathcal{WC}(\mu, \rho)$ con función de densidad

$$f(\theta) = \sum_{k=-\infty}^{\infty} \frac{1}{\pi} \frac{\sigma}{\sigma^2 + (\theta - \mu + 2k\pi)^2}.$$

A pesar de que la media de X no está definida para el caso lineal, la de Θ es $\mu \mod 2\pi$ y la longitud media es $R = \rho = e^{-\sigma}$.

Sin embargo, en este caso sí que es posible simplificar la expresión anterior utilizando (4.7). Teniendo en cuenta que la función característica de X es $\phi(n) = e^{-\sigma n + in\mu} = \rho e^{in\mu}$ $(n \in \mathbb{N})$ y que, por tanto, $S_n = \rho \sin(n\mu)$ y $C_n = \rho \cos(n\mu)$, obtenemos la siguiente función de densidad

$$f(\theta) = \frac{1}{2\pi} \left\{ 1 + 2 \sum_{n=1}^{\infty} \rho^n \cos n(\theta - \mu) \right\}. \tag{4.9}$$

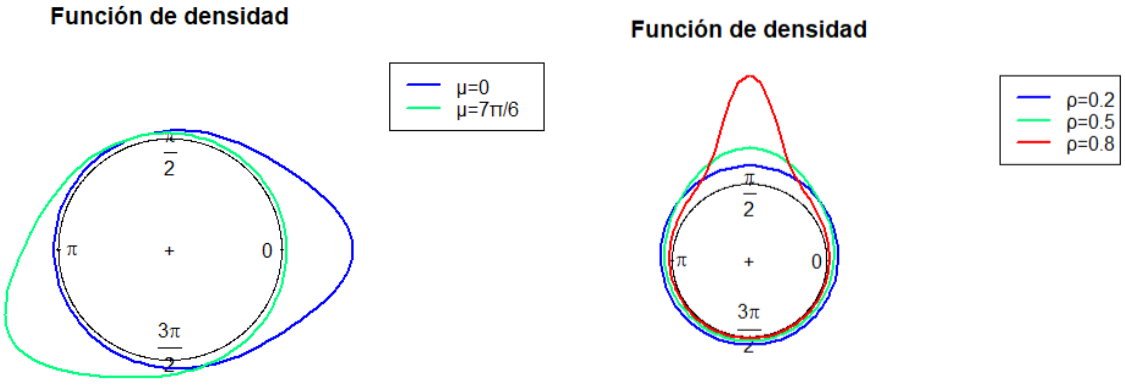
Ahora bien, si nos fijamos, la suma que aparece en esta expresión converge a una función más sencilla ya que

$$\sum_{n=1}^{\infty} \rho^n \cos n(\theta-\mu) = \mathrm{Re}\left(\sum_{n=1}^{\infty} \rho^n e^{in(\theta-\mu)}\right) = \mathrm{Re}\left(\frac{\rho e^{i(\theta-\mu)}}{1-\rho e^{i(\theta-\mu)}}\right) = \frac{\rho(\cos(\theta-\mu)-\rho)}{1+\rho^2-2\rho\cos(\theta-\mu)},$$

donde hemos utilizado que la suma que aparece tras la primera igualdad es la de una serie geométrica de razón menor que 1. Finalmente, sustituyendo este resultado en (4.9) y reordenando los términos, se consigue una representación mucho más compacta de la función de densidad de Θ dada por

$$f(\theta) = \frac{1}{2\pi}\frac{1-\rho^2}{1+\rho^2-2\rho\cos(\theta-\mu)}. \tag{4.10}$$

Notemos que cuando $\rho \to 0$ recuperamos la distribución uniforme circular, mientras que cuando $\rho \to 1$ tenemos una distribución degenerada en el punto μ mód 2π. También en este caso tenemos una distribución unimodal y simétrica respecto a dicho punto, tal y como se puede apreciar en 4.5a: al variar el parámetro μ (sin hacerlo ρ) no cambia la forma de la curva si no únicamente la orientación de esta, de forma que el máximo de la función de densidad coincide con el ángulo medio. En las funciones de densidad representadas en la figura 4.5b podemos ver el hecho de que ρ es el parámetro que se corresponde con la longitud media, ya que, cuánto aumenta este valor, más se aleja la distribución Cauchy enrollada de la uniforme circular. Fijémonos en que la forma de campana de esta distribución es algo diferente a las que hemos visto hasta ahora ya que, aunque sí que observamos que cuando crece ρ más concentración de probabilidad hay en torno a μ (como ocurre con las anteriores), en este caso, el máximo de la curva es más pronunciado, es decir, la probabilidad está aún más concentrada que en los otros ejemplos vistos.



(a) $\mathcal{WC}(\mu, 0.6)$ (b) $\mathcal{WC}(\frac{\pi}{2}, \rho)$

Figura 4.5: Representación de la función de densidad de la distribución Cauchy enrollada variando los parámetros (a) μ y (b) ρ.

4.3. Distribuciones proyectadas

Otra forma de obtener distribuciones apropiadas para variables aleatorias circulares es "proyectándo" las funciones de distribución de vectores aleatorios lineales bidimensionales.

Sea el vector aleatorio (X, Y) con función de densidad $f(x, y)$. Mediante un cambio de variable, podemos obtener su representación en coordenadas polares $f(r, \theta)$ (con r la variable radial y θ, la angular). Así una variable aleatoria circular Θ será aquella que tiene por función de densidad la función de densidad marginada $f(\theta)$, es decir,

$$f(\theta) = \int_0^\infty f(r, \theta) r dr. \tag{4.11}$$

Como ejemplo de este tipo de distribuciones veamos la normal proyectada $\mathcal{NP}(\vec{\mu}, \mathbf{\Sigma})$ que se obtiene a la partir de una distribución normal bivariante $\mathcal{N}_2(\vec{\mu}, \mathbf{\Sigma})$ de un vector aleatorio lineal (X_1, X_2), donde $\vec{\mu} = (\mu_1, \mu_2)^T$ es el vector de medias y $\mathbf{\Sigma}$, la matriz de covarianzas, siendo ρ el coeficiente de correlación de X_1 y X_2, y σ_1^2 y σ_2^2 sus respectivas varianzas. En Jammalamadaka y SenGupta (2001) se puede ver la función de densidad de $\Theta \sim \mathcal{NP}(\vec{\mu}, \mathbf{\Sigma})$ es

$$f(\theta) = \frac{1}{C(\theta)} \left\{ \phi_2(\mu_1, \mu_2) + aD(\theta)\Phi[D(\theta)]\phi\left(\frac{a(\mu_1 \sin\theta - \mu_2 \cos\theta)}{\sqrt{C(\theta)}} \right) \right\}, \tag{4.12}$$

con

$$\begin{aligned} a &= \frac{1}{\sigma_1 \sigma_2 \sqrt{1-\rho^2}}, \\ C(\theta) &= a^2(\sigma_2^2 \cos^2\theta - \rho\sigma_1\sigma_2 \sin 2\theta + \sigma_1^2 \operatorname{sen}^2\theta) \quad \text{y} \\ D(\theta) &= \frac{a^2}{\sqrt{C(\theta)}}[\mu_1\sigma_2(\sigma_2 \cos\theta - \rho\sigma_1 \sin\theta) + \mu_2\sigma_1(\sigma_1 \sin\theta - \rho\sigma_2 \cos\theta)], \end{aligned}$$

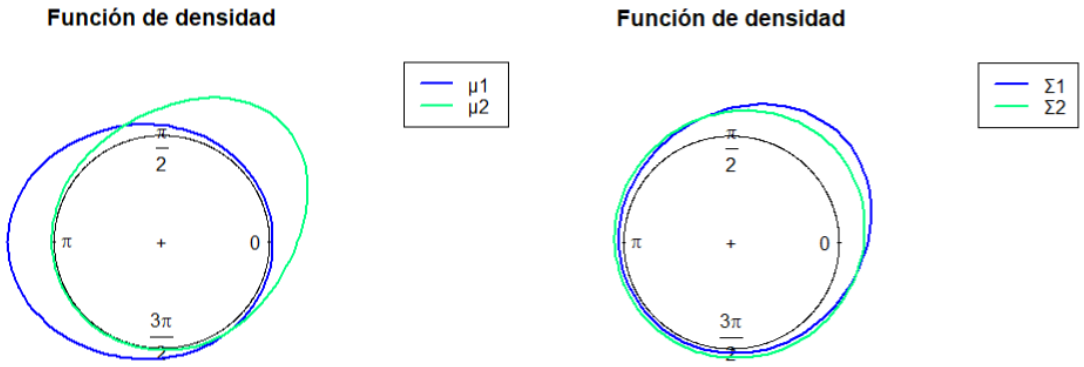
donde $\phi_2(\cdot, \cdot)$ es la función de densidad de $\mathcal{N}_2(\vec{0}, \mathbf{\Sigma})$, y $\phi(\cdot)$ y $\Phi(\cdot)$, las funciones de densidad y distribución de $\mathcal{N}(0, 1)$, respectivamente.

Al considerar el caso en que $\vec{\mu} = \vec{0}$ y $\sigma_1 = \sigma_2 = 1$, la expresión (4.12) se simplifica a

$$f(\theta) = \frac{\sqrt{1-\rho^2}}{2\pi(1 - \rho \sin 2\theta)}.$$

Es obvio que, en este caso, cuando $\rho = 0$, recuperamos la función de densidad de la distribución uniforme circular (4.1). En las imágenes de la figura 4.6 tenemos la representación de la función de densidad introducida en (4.12) para distintos parámetros, para los que se obtienen distribuciones unimodales y simétricas. En la de la izquierda (figura 4.6a), vemos como a pesar de que ambas distribuciones tienen el mismo parámetro $\mathbf{\Sigma}$, parece que la curva verde posee una mayor concentración de probabilidad en torno a la media que la azul. Sin embargo, a la derecha (figura 4.6b), donde hemos tomado dos $\mathbf{\Sigma}$ distintos pero diagonales, se tiene que esta concentración

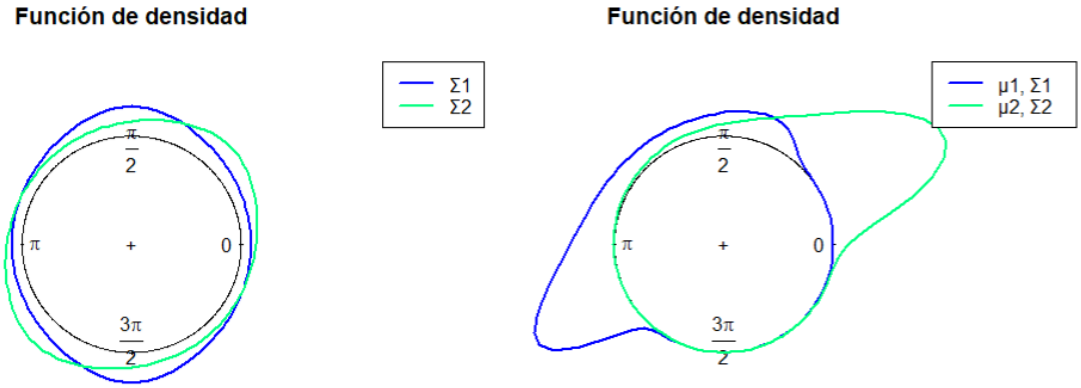
disminuye al aumentar las varianzas en las distribuciones normales bivariantes lineales de las que proceden estas normales proyectadas.



(a) $\mu_1 = (-1, 0)$, $\mu_2 = (1, 1)$, $\Sigma = \left(\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right)$ (b) $\mu = (1, 1)$, $\Sigma_1 = \left(\begin{smallmatrix} 2 & 0 \\ 0 & 2 \end{smallmatrix}\right)$, $\Sigma_2 = \left(\begin{smallmatrix} 6 & 0 \\ 0 & 6 \end{smallmatrix}\right)$

Figura 4.6: Representación de la función de densidad de la distribución normal proyectada variando los parámetros μ (a) y $\mathbf{\Sigma}$ (b), de forma que se obtienen distribuciones unimodales y simétricas.

Por otro lado, a diferencia de las distribuciones anteriores, esta puede ser también bimodal o asimétrica. Esto es lo que observamos en las imágenes 4.7a y 4.7b respectivamente. Observemos que ahora los parámetros $\mathbf{\Sigma}$ utilizados no tienen por qué ser matrices diagonales, sino simétricas.



(a) $\mu = (0, 0)$, $\Sigma_1 = \left(\begin{smallmatrix} 2 & 0 \\ 0 & 6 \end{smallmatrix}\right)$, $\Sigma_2 = \left(\begin{smallmatrix} 1 & 0.5 \\ 0.5 & 1 \end{smallmatrix}\right)$

(b) $\mu_1 = (-1, 0)$, $\Sigma_1 = \left(\begin{smallmatrix} 1 & 0.9 \\ 0.9 & 2 \end{smallmatrix}\right)$, $\mu_2 = (3, 2)$, $\Sigma_2 = \left(\begin{smallmatrix} 4 & 0.3 \\ 0.3 & 2 \end{smallmatrix}\right)$

Figura 4.7: Representación de la función de densidad de la distribución normal proyectada variando los parámetros μ y $\mathbf{\Sigma}$, de forma que se obtienen distribuciones (a) bimodales o (b) asimétricas.

Antes de explicar los últimos ejemplos de distribuciones, cabe hacer una comparación entre

las que hemos visto hasta ahora. Esto es lo que tenemos en la figura 4.8 en la que hemos intentado que las distribuciones tuviesen parámetros similares. Llama la atención que la Cauchy enrollada, junto con la norma proyectada, son las que tienen una mayor concentración de la probabilidad en torno a la media. Por otro lado, podemos ver que la cardioide y la normal enrollada son bastante similares.

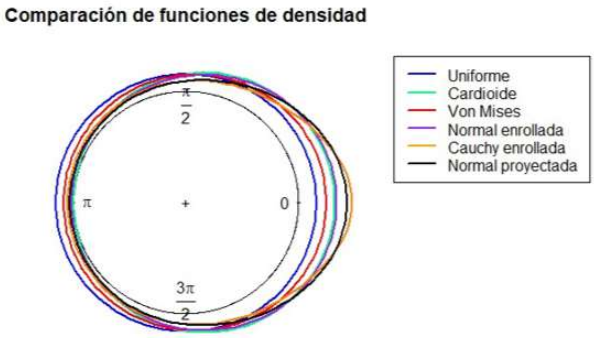


Figura 4.8: Comparación de las funciones de densidad de las distribuciones vistas anteriormente: circular uniforme, cardioide, Von-Mises, normal enrollada, Cauchy enrollada y normal proyectada. Excepto esta última, todas ellas tienen como parámetros $\mu = 0$ y $\rho = 0.5$ ($\kappa = 0.5$, en el caso de la distribución de Von-Mises). Los parámetro de la normal proyectada son $\mu = (1, 0)$ y $\Sigma = \left(\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right)$.

4.4. Distribuciones asimétricas y multimodales

Además de las distribuciones vistas hasta ahora, existen otro modelos para variables aleatorias circulares asimétricas, como puede ser la normal enrollada asimétrica o *wrapped skew normal*. Por otro lado, también es necesaria alguna distribución para variables multimodales en el caso circular, como las mixturas de Von-Mises. A continuación, se presentarán estos dos ejemplos mostrando sus ventajas.

4.4.1. Distribución normal enrollada asimétrica

Esta distribución es útil cuando las variable aleatoria circular Θ que tenemos no es simétrica. Se denota por $\mathcal{WSN}(\mu, \kappa, \lambda)$ y su correspondiente función de densidad viene dada por la siguiente expresión:

$$f(\theta) = \frac{2}{\kappa} \sum_{r=-\infty}^{\infty} \phi\left(\frac{\theta - 2\pi r - \mu}{\kappa}\right) \Phi\left(\lambda\left(\frac{\theta + 2\pi r - \mu}{\kappa}\right)\right),$$

donde $\phi(\cdot)$ y $\Phi(\cdot)$ denotan la funciones de densidad y de distribución de la normal estándar, respectivamente. El coeficiente $\mu \in [0, 2\pi)$ es un parámetro de localización, $\kappa > 0$, un parámetro de escala, y $\lambda \geq 0$, un parámetro de asimetría (Oliveira, 2014). En la figura 4.9 en la que podemos ver cómo varía la simetría de la función de densidad en relación al valor de λ: al aumentar esta cantidad, la simetría disminuye.

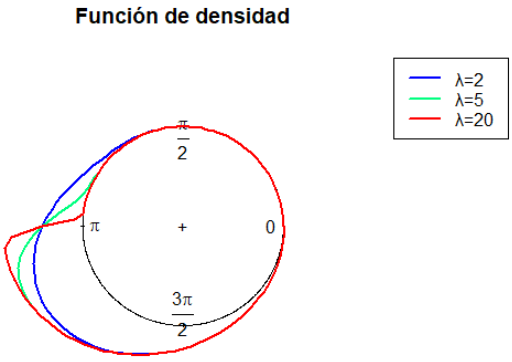


Figura 4.9: Representación de la función de densidad de la distribución normal enrollada asimétrica, con $\mu = \pi$ y $\kappa = 1$, variando el parámetro de asimetría λ.

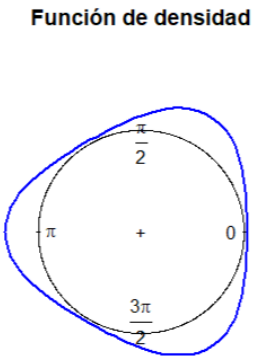
4.4.2. Mixtura de Von-Mises

También se puede obtener una nueva distribución como "mezcla" de otras distribuciones. En este caso, la función de densidad de la variable aleatoria circular Θ se escribe como suma de funciones de densidad de Von-Mises (4.3) de la siguiente forma

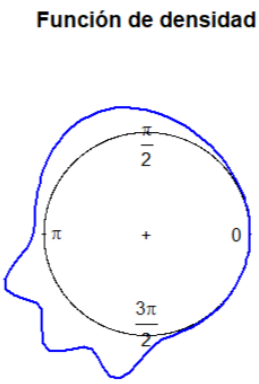
$$f(\theta) = \sum_{i=1}^{k} p_i f_{VM}(\theta; \mu_i, \kappa_i) = \frac{1}{2\pi} \sum_{i=1}^{k} p_i \frac{1}{I_0(\kappa_i)} e^{\kappa_i(\theta - \mu_i)}, \tag{4.13}$$

con $k \geq 2$. Los coeficientes p_i son los pesos de cada componente y deben cumplir: $p_i \geq 0$ y $\sum_{i=1}^{k} p_i$=1. Además, recordemos que $\mu_i \in [0, 2\pi)$ y $\kappa_i \geq 0, \forall i \in \{1, ..., k\}$ (Mardia y Jupp, 2000). De esta forma obtenemos una distribución multimodal y que además, puede ser asimétrica.

Podemos observar las representaciones de esta función de densidad para diferentes combinaciones de sumas de distribuciones de Von-Mises en las figuras de 4.10. En la de la izquierda (4.10a), tenemos la suma de tres de ellas, lo que resulta en una distribución con tres modas, que se corresponden con el parámetro μ de cada Von-Mises de la suma, y es simétrica con respecto a cada una de ellas. Por otro lado, en la imagen de la derecha (4.10b), se han sumado cinco funciones de densidad de Von-Mises y, en este caso, la resultante no es simétrica y tiene varios máximos, ubicados en torno a la circunferencia de forma la distribución es no simétrica.



(a) $\frac{1}{3}\mathcal{VM}(\frac{\pi}{3}, 6) + \frac{1}{3}\mathcal{VM}(\pi, 6) + \frac{1}{3}\mathcal{VM}(\frac{5\pi}{3}, 6)$



(b) $\frac{4}{9}\mathcal{VM}(2, 3)+\frac{5}{36}\mathcal{VM}(4, 3)+\frac{5}{36}\mathcal{VM}(3.5, 50)+\frac{5}{36}\mathcal{VM}(4, 50) + \frac{5}{36}\mathcal{VM}(4.5, 50)$

Figura 4.10: Representaciones de esta función de densidad para diferentes combinaciones de sumas de distribuciones de Von-Mises (Oliveira, 2014).

Capítulo 5

Estimación no paramétrica

En el capítulo 2 hemos analizado las propiedades que deben verificar las funciones de densidad tanto en el caso lineal como circular. Así podemos asegurar que esta caracteriza inequívocamente una variable aleatoria. De ahí la importancia de obtener una estimación de esta a partir de una muestra de dicha variable aleatoria. Esta estimación puede plantearse de dos maneras: de forma paramétrica o no paramétrica. Hemos introducido modelos paramétricos en el capítulo anterior, por ello, en este nos vamos a centrar en la estimación no paramétrica.

La estimación paramétrica se utiliza cuando conocemos la distribución que sigue la variable en cuestión, pero no los parámetros de esta. Por ejemplo, si tenemos $\Theta \sim \mathcal{M}(\mu, \kappa)$, sabemos que Θ sigue una distribución de Von-Mises pero desconocemos su media μ y su parámetro de concentración κ. De esta forma, a través de una muestra de la población, se obtendría una estimación de la función de densidad $\hat{f}$ de Θ a partir de la estimación de estos coeficientes.

El problema de este tipo de estimación radica en la necesidad de asumir que la variable en cuestión sigue una cierta distribución. Es decir, cuando no es posible conocer la distribución de la variable a priori, no se puede emplear el enfoque paramétrico. Los métodos no paramétricos son una alternativa que soluciona este problema, ya que permiten estimar f sin hacer ninguna suposición previa sobre la distribución de la variable. Esta ganancia en flexibilidad se traduce en cierta complejidad en la práctica.

En este capítulo nos centraremos en la estimación no paramétrica de la función de densidad en el caso de variables aleatorias circulares, pero antes haremos una pequeña introducción de los conceptos básicos de este campo en el caso lineal. En particular, utilizaremos la estimación de la función de densidad a través de estimadores tipo *kernel*.

5.1. Introducción a la estimación no paramétrica lineal de la función de densidad

Para comenzar, veremos que es posible estimar la función de densidad de una muestra a través del histograma. Como se muestra en Scott (2015), seleccionado un origen t_k y una constante $h > 0$ (denominada parámetro de ventana), un histograma define un función constante a trozos en los intervalos $\{B_k = [t_k, t_{k+1}) : t_k = t_0 + kh, k \in \mathbb{Z}\}$, denominados *bins*. El valor de esta función en cada B_k se corresponde con el número de observaciones muestrales dentro de cada uno de estos intervalos. De esta forma, el histograma en un punto x se puede expresar como

$$\hat{f}_H(x; t_0, h) = \frac{1}{nh} \sum_{i=1}^{n} 1_{\{X_i \in B_k : x \in B_k\}}, \tag{5.1}$$

donde $X_1, ..., X_n$, con $n \in \mathbb{N}$, es una muestra aleatoria simple de una variable aleatoria lineal X. Si denotamos por v_k el número de observaciones en cada B_k, la expresión anterior es equivalente a

$$\hat{f}_H(x; t_0, h) = \frac{v_k}{nh}, \text{ si } x \in B_k \text{ para algún } k \in \mathbb{Z}. \tag{5.2}$$

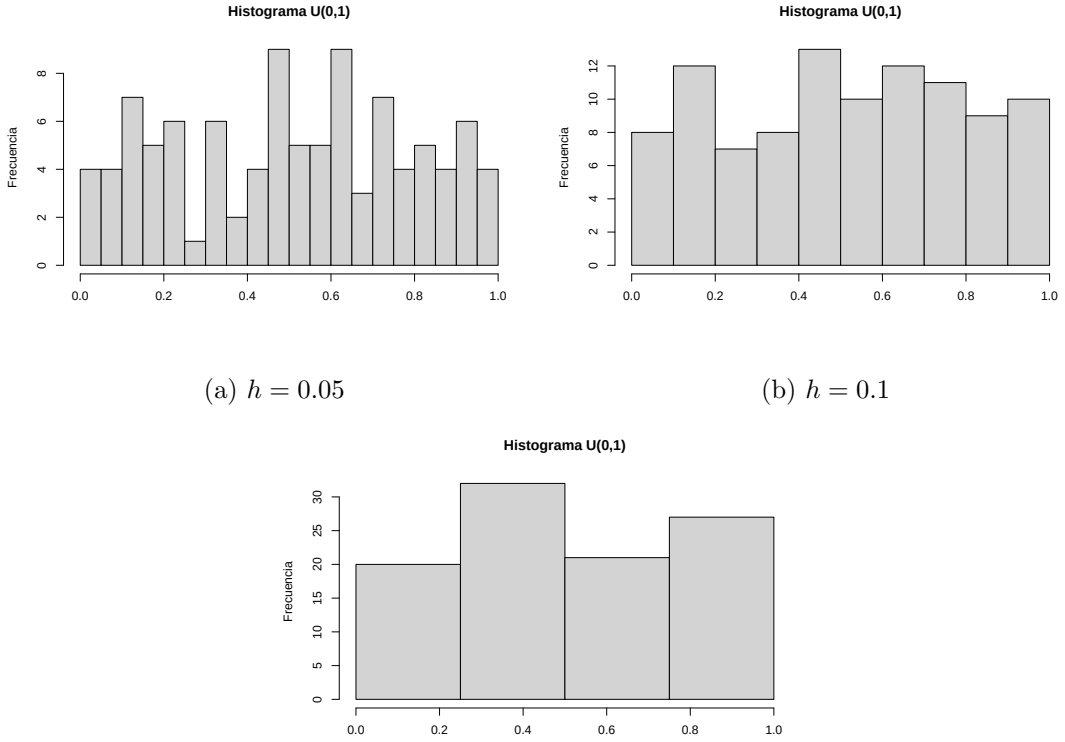
Notemos que la variable aleatoria "número de observaciones v_k en cada *bin* del histograma" sigue una binominal $B(n, p_k)$, con $p_k = P(X \in B_k) = \int_{B_k} f(t)dt$, donde $f(t)$ es la función de densidad de la variable X (continua). Por el Teorema del valor medio, $p_k = hf(\xi_{k,h})$ para algún $\xi_{k,h} \in (t_k, t_{k+1})$. Así, para un $k \in \mathbb{Z}$ y un $x \in B_k$ dados, se tiene que

$$\begin{aligned} \mathbb{E}[\hat{f}_H(x; t_0, h)] &= \frac{np_k}{nh} = f(\xi_{k,h}), \\ Var[\hat{f}_H(x; t_0, h)] &= \frac{np_k(1-p_k)}{n^2h^2} = \frac{f(\xi_{k,h})(1 - hf(\xi_{k,h}))}{nh} \end{aligned}$$

Por tanto, cuando $h \to 0$, $f(\xi_{k,h}) \to f(x)$ y $\mathbb{E}[\hat{f}_H(x; t_0, h)] \to f(x)$, lo que hace de la función dada por 5.2 un estimador asintóticamente insesgado de $f(x)$. No obstante, si $h \to 0$, $Var[\hat{f}_H(x; t_0, h)] \to \infty$, por lo que es necesario que $nh \to \infty$ (Scott, 2015).

Cabe preguntarse cómo varía el histograma y, por tanto, el estimador (5.2) con el parámetro de ventana h y con t_0. Tomamos una muestra de tamaño $n = 100$ de una distribución uniforme $U(0, 1)$ lineal y representamos su histograma tomando $h = 0.05$, $h = 0.1$ y $h = 0.25$ con $t_0 = 0$ (véanse 5.1a, 5.1b y 5.1c, respectivamente). Claramente, lo que estamos haciendo es cambiar en número de *bins* del histograma y vemos que los grupos modales varían de uno a otro. De esta forma, al utilizar un h demasiado pequeño, podrían aparece modas que en realidad no hay y si h es demasiado grande, podrían no verse algunos grupos modales que sí existen. Ya comentábamos en el capítulo 3, cuando estudiamos los histogramas circulares y los diagramas de rosa, que esto

podría suceder si no se selecciona un parámetro de ventana, y consecuentemente un número de *bins*, adecuados. Esto es el efecto de la elección de la ventana.



(a) $h = 0.05$

(b) $h = 0.1$

(c) $h = 0.25$

Figura 5.1: Histogramas correspondientes a una muestra de una distribución uniforme $U(0, 1)$ lineal de tamaño $n = 100$ con parámetro $t_0 = 0$ y parámetros de ventana (a) $h = 0.05$, (b) $h = 0.1$ y (c) $h = 0.25$.

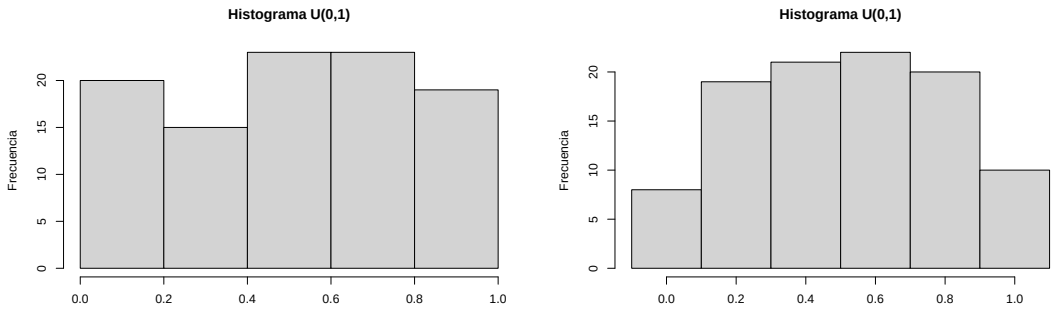
Por otra parte, en la figuras 5.2a y 5.2b hemos representado los histogramas de la muestra anterior para $t_0 = 0$ y $t_0 = -0.1$ (respectivamente), con parámetro de ventana $h = 0.2$. Vemos como no solo varía la forma del histograma, sino que también lo hace el número de *bins* de este.

Para evitar esta dependencia se utiliza el histograma móvil, en el que se utiliza la frecuencia relativa en intervalos de la forma $(x - h, x + h)$ para estimar la densidad de x. De esta forma, se obtiene otro estimador de la función de densidad cuya expresión viene dada por la función constante a trozos

$$\hat{f}_N(x; h) = \frac{1}{2nh} \sum_{i=1}^{n} 1_{\{x-h<X_i<x+h\}}, \tag{5.3}$$

con $X_1, ..., X_n$ una muestra aleatoria simple de una variable aleatoria lineal X (Scott, 2015).

Nuevamente tenemos que, al tratar $\hat{f}_N(x; h)$ como una variable aleatoria, esta sigue una binominal $B(n, p_{x,h})$, con $p_{x,h} = P(x - h < X < x + h) = F(x + h) - F(x - h)$ y F,la función



(a) $t_0 = 0$ (b) $t_0 = -0.1$

Figura 5.2: Histogramas correspondientes a una muestra de tamaño $n = 100$ de una distribución uniforme $U(0, 1)$ lineal con parámetro de ventana $h = 0.2$ y (a) $t_0 = 0$ y (b) $t_0 = -0.1$.

de distribución de X. De esta forma, se tiene que

$$\mathbb{E}[\hat{f}_N(x;h)] = \frac{F(x+h) - F(x-h)}{2h},$$
$$Var[\hat{f}_N(x;h)] = \frac{F(x+h) - F(x-h)}{4nh^2} - \frac{(F(x+h) - F(x-h))^2}{4nh^2}.$$

Así, cuando $h \to 0$, tenemos que $\mathbb{E}[\hat{f}_N(x;h)] \to f(x)$ y $Var[\hat{f}_N(x;h)] \to \infty$, lo que implica que $\hat{f}_N(x;h)$ es también un estimador asintóticamente insesgado de $f(x)$ y es necesario que $nh \to \infty$ si $h \to 0$, de forma que la varianza no diverja.

Fijémonos en que la expresión (5.3) se puede reescribir de la siguiente forma

$$\hat{f}_N(x;h) = \frac{1}{2nh}\sum_{i=1}^{n} 1_{\{x-h<X_i<x+h\}} = \frac{1}{nh}\sum_{i=1}^{n}\frac{1}{2}1_{\{-1<\frac{x-X_i}{h}<1\}} = \frac{1}{nh}\sum_{1=1}^{n} K\left(\frac{x-X_i}{h}\right), \quad (5.4)$$

donde usado la notación $K(z) = \frac{1}{2}1_{\{-1<z<1\}}$, que es la función de densidad de la distribución uniforme $U(-1, 1)$. Por lo que cada $X_1, ..., X_n$ tiene el mismo peso.

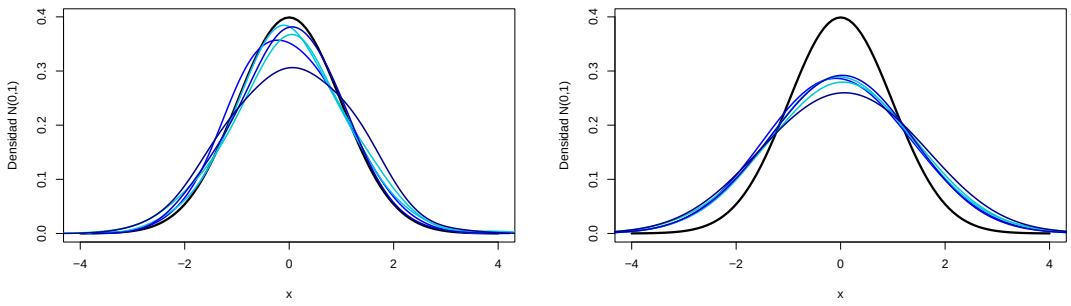
Como podemos ver en Parzen (1962), se posible generalizar la expresión (5.4) sustituyendo K por la función de densidad de una distribución simétrica y unimodal en el cero, como puede ser una normal estándar. De esta forma, K se denomina *kernel* y se define el estimador *kernel* de la función de densidad como

$$\hat{f}(x;h) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right), \quad (5.5)$$

También es frecuente utilizar la función *kernel* reescalada $K_h(u) = \frac{1}{h}K(\frac{u}{h})$. De esta forma la ecuación (5.5) se transforma en

$$\hat{f}(x;h) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - X_i). \quad (5.6)$$

Sin embargo, aunque la elección del *kernel* es importante, lo es aún más la del parámetro de ventana h. Ya hemos comentado esto en el caso de los histogramas, pero veamos un ejemplo donde utilizamos un *kernel* no uniforme: en las figuras 5.3a y 5.3b, tenemos las representaciones de la función de densidad la normal estándar y las estimaciones de estas para cinco muestras de tamaño 200 procedentes de esta distribución. En ambas utilizamos un *kernel* gaussiano, en la primera $h = 0.5$, mientras que en la segunda, $h = 1$. Notamos claramente que para $h = 0.5$ se obtiene, en general, una mejor estimación de las funciones de densidad de cada muestra.



(a) $h = 0.5$ (b) $h = 1$

Figura 5.3: Representaciones de la función de densidad la normal estándar (curvas negras) y las estimaciones de estas para cinco muestras de tamaño 200 procedentes de esta distribución (curvas azules), para los parámetros de ventana (a) $h = 0.5$ y (b) $h = 1$ y *kernel* gaussiano.

Existen varios métodos para obtener un parámetro de ventana adecuado, algunos de los cuales introducimos brevemente a continuación. Para ello, denotemos por f a la función de densidad de una variable aleatoria lineal X y sea $\hat{f}$ un estimador suyo.

Los primeros métodos se corresponden con criterios basados en minimizar algunos errores. Se utilizan el error cuadrático medio integrado, que denotaremos por MISE(h) (del inglés *mean integrated squared error*) que viene dado por

$$\text{MISE}(h) = \mathbb{E}\left[\int (\hat{f}(x;h) - f(x))^2 dx\right], \tag{5.7}$$

y su versión asintótica AMISE (*asymtotic mean integrated squared error*), cuya expresión se puede ver en Marron y Wand (1992).

El problema reside en que para minimizar las expresiones de estos errores es necesario conocer la función de densidad f, la cual queremos estimar y, por tanto, desconocemos.

Una posible solución, que utilizan las denominadas reglas *plug-in*, es asumir que f es la función de densidad de una distribución $\mathcal{N}(\mu, \sigma)$. Al emplear al mismo tiempo un *kernel* gaussiano, se obtiene un caso particular de método *plug-in* llamado regla del pulgar (véase Silverman

(1986)). En cualquier caso, el parámetro de ventana h óptimo depende de la desviación típica σ de la Normal, por lo que se sustituye por su estimador de máxima verosimilitud $\hat{\sigma}$.

Los métodos de validación cruzada se basan en desarrollar la expresión del MISE (5.7) y minimizar aquellos términos que dependen del parámetro de ventana h. De esta forma se obtienen dos métodos: validación cruzada por mínimos cuadrados o LSCV (*least squares cross validation*) y validación cruzada por verosimilitud o LCV (*likelihood cross validation*). La expresiones de estos pueden encontrarse en Scott (2015).

Vemos que la estimación no paramétrica tiene ciertas ventajas frente a la paramétrica ya que no necesitamos conocer a priori la distribución de los datos. Sin embargo, tener una distribución conocida permite obtener más información acerca de los datos que estamos estudiando.

Por otro lado, en ambos modelos surge el problema de qué método emplear. La desventaja que surge en el caso no paramétrico es que los parámetros de ventana que resultan de los diversos métodos nos pueden llevar a las distribuciones de los datos muy distintas. Esto no sucede en el otro caso pues, aunque las distintas técnicas proporcionen parámetros distintos de la distribución que suponemos que siguen los datos, sí conocemos el tipo de distribución.

5.2. Estimación no paramétrica de la función de densidad en el caso circular

Es claro que estos estimadores vistos para el caso lineal en la sección anterior (5.5 y 5.6) no son apropiados para variables circulares pues en primer lugar su función de densidad debe ser periódica. En Hall *et al.* (1987) se muestran dos modificaciones de estos que sí son válidos para este tipo de variables. Uno de ellos está basado en la distancia angular entre dos vectores, pero nosotros nos centraremos en el que se define de forma más o menos análoga a 5.6.

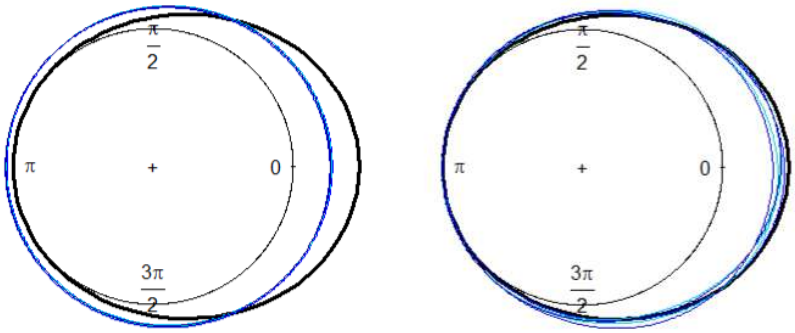
Consideremos $\Theta_1, ..., \Theta_n$ una muestra aleatoria simple de una variable circular Θ, definida en el soporte $[0,2\pi)$. El estimador *kernel* $\hat{f}$ de su función de densidad f viene dado por

$$\hat{f}(x;\kappa) = \frac{1}{n}\sum_{i=1}^{n} K_\nu(\theta - \Theta_i), \tag{5.8}$$

donde $\theta \in [0, 2\pi)$. La función *kernel* circular K_ν se sustituye por una función de densidad circular unimodal y simétrica con parámetro de concentración ν. Este coeficiente ν, que se denomina parámetro de suavizado, sustituye al parámetro de ventana h que teníamos en el caso lineal. De esta forma se mantiene la periodicidad de la función de densidad.

Sin embargo, fijar un valor de ν adecuado es muy importante a la hora de obtener una buena estimación también en este caso. Veamos un ejemplo donde se puede ver esto: en las figuras

5.4a y 5.4b tenemos la representación de la función de densidad de una distribución $\mathcal{WN}(0, 0.7)$ (curva negra), mientras que las curvas azules son las estimaciones de esta función de densidad utilizando en (5.8) la distribución normal enrollada con parámetro de suavizado $\nu = 0.5$ y $\nu = 0.9$, respectivamente, para cinco muestras distintas procedentes de una $\mathcal{WN}(0, 0.7)$. Vemos claramente que la estimación varía en función de ν, de hecho, la primera se queda bastante por debajo de la función real, mientras que la segunda hace un mejor estimación. Aquí está la importancia de una buena elección del parámetro de suavizado. Notemos que en el ejemplo que mostramos para el caso de muestral procedentes de una $\mathcal{N}(0, 1)$ (5.3) también obteníamos una mejor estimación para un parámetro de ventana más pequeño. Sin embargo, un valor se reduce demasiado es posible que tampoco se obtengan resultados satisfactorios.



(a) $\nu = 0.5$ (b) $\nu = 0.9$

Figura 5.4: Representación de la función de densidad de una distribución $\mathcal{WN}(0, 0.7)$ (curva negra) y de las estimaciones de esta función de densidad (curvas azules) utilizando en 5.8 la distribución normal enrollada con parámetro de suavizado (a) $\nu = 0.5$ y (b) $\nu = 0.9$, para cinco muestras distintas de tamaño 200 procedentes de una $\mathcal{WN}(0, 0.7)$.

Sin embargo, al igual que sucede con h cuando se trata de variables lineales, resulta complicado fijar el valor de ν apropiado. Existen varios métodos con los que proceder, aunque sigue siendo un problema abierto. En Hall *et al.* (1987) se presentan dos técnicas, la primera se basa en minimizar el valor esperado de la divergencia de Kullback-Leibler del estimador kernel de la función de densidad que viene dado por

$$\mathcal{L}_{KL}(\nu) = \int_0^{2\pi} f(\theta)\mathbb{E}\left[\log\left(\frac{f(\theta)}{\hat{f}(\theta;\nu)}\right)\right] d\theta,$$

que mide la cantidad de información que hay en la diferencia de $f(\theta)$ y $\hat{f}(\theta;\nu)$. No obstante, nosotros estudiaremos el segundo método en el que se minimiza, al igual que en el caso lineal, el error cuadrático medio integrado del estimador kernel de la función de densidad, que

denotaremos ahora por MISE(ν) y cuya expresión para variables circulares es

$$\begin{aligned}
\text{MISE}(\kappa) =& \mathbb{E}\left[\int_0^{2\pi}(\hat{f}(\theta;\nu)-f(\theta))^2 d\theta\right] = \int_0^{2\pi}\mathbb{E}\left[(\hat{f}(\theta;\nu)-f(\theta))^2\right]d\theta \\
&= \int_0^{2\pi}\mathbb{E}\left[\hat{f}(\theta;\nu)^2+f(\theta)^2-2\hat{f}(\theta;\nu)f(\theta)\right]d\theta \\
&= \int_0^{2\pi}\left(\mathbb{E}\left[\hat{f}(\theta;\nu)^2\right]+f(\theta)^2-2f(\theta)\mathbb{E}\left[\hat{f}(\theta;\nu)\right]+\mathbb{E}\left[\hat{f}(\theta;\nu)\right]^2-\mathbb{E}\left[\hat{f}(\theta;\nu)\right]^2\right)d\theta \\
&= \int_0^{2\pi}\left(\mathbb{E}\left[\hat{f}(\theta;\nu)\right]-f(\theta)\right)^2 d\theta + \int_0^{2\pi}\left(\mathbb{E}\left[\hat{f}(\theta;\nu)^2\right]-\mathbb{E}\left[\hat{f}(\theta;\nu)\right]^2\right)d\theta \\
&= \int_0^{2\pi}[\text{sesgo}(\hat{f}(\theta;\nu))]^2 d\theta + \int_0^{2\pi}\text{Var}(\hat{f}(\theta;\nu))d\theta
\end{aligned}$$

Sin embargo, también en el caso circular, se utiliza normalmente la versión asintótica de este AMISE. Su expresión explícita se puede encontrar en Di Marcio *et al.* (2009) y es similar a la que se puede ver en Hall *et al.* (1987). En la primera de estas referencia se tiene la expresión del AMISE para la distribución de Von-Mises, esta es

$$\text{AMISE}(\nu) = \frac{1}{16}\left(1-\frac{I_2(\nu)}{I_0(\nu)}\right)^2\int_0^{2\pi}(f''(\theta))^2 d\theta + \frac{I_0(\nu)}{2n\pi\kappa(I_0(\nu))^2}, \tag{5.9}$$

donde $I_j(\nu)$ es la función de Bessel modificada de primera especie y orden j, con $j = 0, 2$, y n el tamaño muestral. Tengamos en cuenta que la expresión del AMISE (5.9) es tal si $\nu \to \infty$, $\frac{\sqrt{\nu}}{n} \to 0$ y f'' es continua y de cuadrado integrable.

Ahora bien, teniendo en cuenta que lo que necesitamos el minimizar (5.9) para obtener el valor óptimo de ν y conseguir la mejor estimación de la función de densidad f, la dificultad que se presenta es obvia: la ecuación a minimizar depende de la segunda derivada de f, es decir, de la función que queremos estimar, la cual no conocemos.

Cabría pensar que es posible adaptar la regla del pulgar desarrollada para variables lineales al caso de circulares. Esta noción es la que se explica en Taylor (2008): suponemos que los datos siguen una distribución de Von-Mises. En esta referencia se muestra que el AMISE para este caso es

$$\text{AMISE}(\nu) = \frac{3\kappa^2 I_2(2\kappa)}{32\pi\nu^2 I_0(\kappa)^2} + \frac{\sqrt{\nu}}{2n\sqrt{\pi}},$$

donde κ es el parámetro de concentración de la distribución de Von-Mises que hemos supuesto que siguen los datos y n, el tamaño muestral. Así el parámetro de suavizado óptimo que se obtiene para el criterio de minimización del AMISE es

$$\hat{\nu}_{AMISE} = \left(\frac{3n\hat{\kappa}^2 I_2(2\hat{\kappa})}{4\sqrt{\pi}I_0(\hat{\kappa})^2}\right)^{\frac{2}{5}}, \tag{5.10}$$

donde hemos denotado por $\hat{\kappa}$ al estimador de máxima verosimilitud de κ.

Aunque los resultados que obtenemos por este método son buenos cuando los datos son unimodales, no proporciona un valor adecuado cuando estamos ante datos bimodales o con distribuciones asimétricas. Esto se debe a que en este tipo de muestras la estimación de κ es cercana a cero y, por tanto, el valor de ν también, debido a la relación mostrada en (5.10). Como consecuencia, $\hat{f} = \frac{1}{2\pi}$, es decir, tendríamos una distribución uniforme que no se corresponde con la de la muestra. A pesar de ello, existen algunas modificaciones de este procedimiento que llevan a mejores estimaciones de f. Dos de ellas pueden encontrarse en Taylor (2008) y Oliveira *et al.* (2013).

Otra técnica, basada también en minimizar el AMISE (5.9), se obtiene al suponer que los datos tienen como distribución una mixtura de Von-Mises. Este método *plug-in* se explica en Oliveira (2014) y se resume como sigue: se elige el número de distribuciones de Von-Mises k que conforma la mixtura y se estiman sus parámetros, es decir, μ_i, κ_i y p_i $\forall i = 1, ..., k$. Se introducen estas estimaciones en (4.13) y se calcula su derivada segunda, la cual hará el papel de $\hat{f}''$ en (5.9). De esta forma se obtiene un estimador del AMISE que se minimiza para obtener el parámetro de suavizado ν óptimo.

El método de validación cruzada es otro de los procedimientos que hemos visto para el caso lineal que se puede adaptar para estimar el parámetro de suavizado ν. Esta adaptación de este al caso de variables circulares está propuesto en Hall *et al.* (1987). Sin embargo, nosotros utilizaremos las ideas dadas en Oliveira (2014).

Comenzamos introduciendo el error cuadrático integrado del estimador kernel (5.8), que denotamos por ISE(ν) (*integrated squared error*),:

$$\text{ISE}(\nu) = \int_0^{2\pi} (\hat{f}(\theta;\nu) - f(\theta))^2 d\theta = \int_0^{2\pi} \hat{f}(\theta;\nu)^2 d\theta - 2\int_0^{2\pi} \hat{f}(\theta;\nu) f(\theta) d\theta + \int_0^{2\pi} f(\theta)^2 d\theta. \tag{5.11}$$

Para obtener el parámetro ν que proporcione una mejor estimación de f, también se debe minimizar esta expresión (5.11). Notemos que, como la minimización se hace con respecto al parámetro ν y el último término no depende de este, basta con hacerlo para

$$\int_0^{2\pi} \hat{f}(\theta;\nu)^2 d\theta - 2\int_0^{2\pi} \hat{f}(\theta;\nu) f(\theta) d\theta. \tag{5.12}$$

También en este caso la ecuación a minimizar depende de la función de densidad f que queremos estimar y, por tanto, no conocemos. Sin embargo, fijémonos en que

$$\int_0^{2\pi} \hat{f}(\theta;\nu) f(\theta) d\theta = \mathbb{E}(\hat{f}(\theta;\nu)).$$

Cabría pensar que se podría aproximar $\mathbb{E}(\hat{f}(\theta;\nu))$, a través de un estimador suyo de la forma

$$\widehat{\mathbb{E}(\hat{f}(\theta;\nu))} = \frac{1}{n}\sum_{i=1}^{n} \hat{f}(\theta_i;\nu)),$$

donde θ_i denota la observación i-ésima de la muestra y n, el tamaño muestral. Sin embargo, estaríamos utilizando esta muestra en dos ocasiones: esta última y la estimación de $\hat{f}$. Por esta razón, se emplea otro estimador de $\mathbb{E}(\hat{f}(\theta;\nu))$, que viene dado por

$$\widehat{\mathbb{E}(\hat{f}(\theta;\nu))} = \frac{1}{n}\sum_{i=1}^{n}\hat{f}^{-i}(\theta_i;\nu)),$$

donde $\hat{f}^{-i}$ es el estimador *kernel* circular mostrado en (5.8) y calculado utilizando todas las observaciones muestrales excepto la i-ésima.

En este punto, podemos obtener el parámetro de suavizado ν por medio de dos métodos de validación cruzada. El primero se sigue directamente de lo visto hasta ahora para este procedimiento y se trata de la validación cruzada por mínimos cuadrados (LSCV), en el que se minimiza

$$\hat{\nu}_{LSCV} = \int_0^{2\pi}\hat{f}(\theta;\nu)^2 d\theta - \frac{2}{n}\sum_{i=1}^{n}\hat{f}^{-i}(\theta_i;\nu)). \tag{5.13}$$

Por otra parte, también podemos utilizar la validación cruzada por verosimilitud (LCV) en la que se maximiza la expresión

$$\hat{\nu}_{LCV} := \arg\min_{\nu}\prod_{i=1}^{n}\hat{f}^{-i}(\theta_i;\nu)), \tag{5.14}$$

con n el tamaño muestral.

En Di Marzio *et al.* (2011) se muestra que la validación cruzada por verosimilitud es más estable asintóticamente. Además, también se puede encontrar en esta referencia otro procedimiento para obtener ν basado en métodos *bootstrap*.

5.3. Estudio de la estimación no paramétrica circular mediante simulaciones

Ahora que ya conocemos algunos métodos para obtener un parámetro de suavizado adecuado, apliquémoslos a distintas muestras de distintos tamaños para algunas de las distribuciones vista en el capítulo 4.

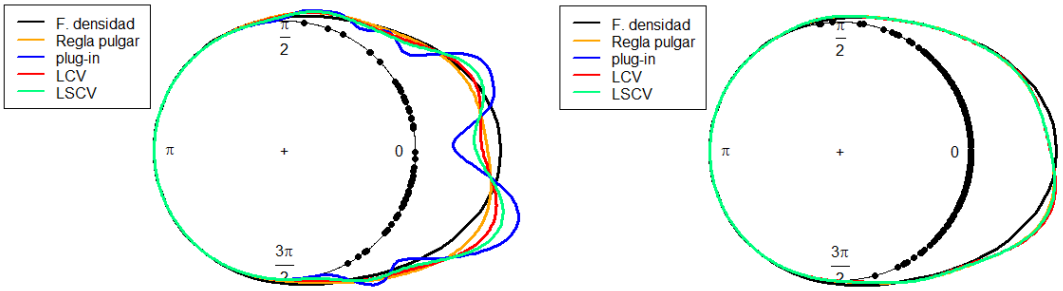
Comencemos analizando una distribución unimodal y simétrica. En este caso, vamos a utilizar dos muestras, una de tamaño n=50 y otra de n=200, de la distribución de Von-Mises de parámetros $\mu = 0$ y $\kappa = 3$. También vamos a obtener el parámetro de suavizado mediante tres métodos: la regla del pulgar, la regla *plug-in*, validación cruzada LCV y LSCV. Esta situación es la que tenemos en la figuras 5.5a y 5.5b, respectivamente. Las diferencias entre las estimaciones de la función de densidad que se obtienen en ambos casos son evidentes. Mientras que

en la primera tenemos estimaciones distintas para cada método utilizado para calcular ν, en la segunda apenas se diferencian. A continuación mostramos los resultados de ν obtenidos:

- Muestra de tamaño 50: ν_{RP}=13.322, ν_{PI}=99.965, ν_{LCV}=23.266, ν_{LSCV}=35.481.
- Muestra de tamaño 200: ν_{RP}=22.680, ν_{PI}=25.917, ν_{LCV}=28.288, ν_{LSCV}=24.633.

Notemos que mientas que los parámetros de suavizado obtenidos para n=50 son también muy distintos entre sí, los de n=200 son similares.

Concluimos, por tanto, que el tamaño muestral n influye en la estimación de la función de densidad de la variable aleatoria de la que proceden los datos, de forma que esta mejora al aumentar n. Esto es lo que cabría esperar ya que tendremos más información de la muestra y también ocurre en el caso lineal.



(a) $n = 50$ (b) $n = 200$

Figura 5.5: Representación de la función de densidad de una distribución $\mathcal{VM}(0, 3)$ (curva negra) y de las estimaciones de esta obtenidas con distintos parámetros de suavizado. Se utiliza una muestra de tamaño (a) $n = 50$ y (b) $n = 200$.

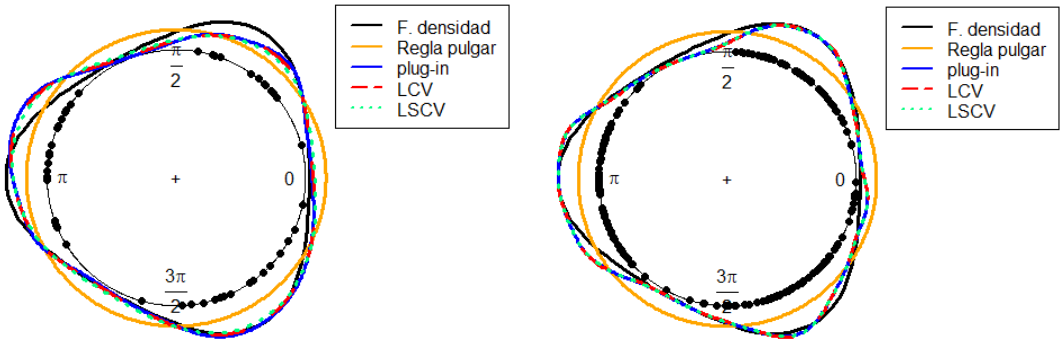
Ahora bien, como hemos comentado anteriormente, todos los métodos propuestos de cálculo de ν funcionan más o menos bien para el caso de distribuciones unimodales y simétricas, como en el ejemplo anterior. Sería interesante analizar qué sucede al aplicar estos procedimientos para distribuciones multimodales o asimétricas.

Así, tomemos dos muestras una de tamaño 50 y otra de 200 de la mixtura de Von-Mises que veíamos en la figura 4.10a (función simétrica y multimodal) y tratemos de estimar su función de densidad. Esto es lo que muestra en las figuras 5.6a y 5.6b, en las que nuevamente notamos la diferencia entre las estimaciones al utilizar tamaños muestrales distintos: se obtiene una más apropiada al aumentar n. Estas diferencias también se aprecian en los valores de los parámetro de suavizado:

- Muestra de tamaño 50: ν_{RP}=0.120, ν_{PI}=20.779, ν_{LCV}=14.419, ν_{LSCV}=10.306.

- Muestra de tamaño 200: ν_{RP}=0.192, ν_{PI}=38.117, ν_{LCV}=43.915, ν_{LSCV}=38.160.

Vemos que, tal y como mencionamos al principio, al tratarse de una distribución multimodal, el parámetro de suavizado calculado mediante la regla del pulgar es muy pequeño para ambos tamaños muestrales y por ello, la estimación obtenida es similar a la función de densidad de una uniforme circular. Con respecto a los otros métodos, estos devuelven valores de ν que originan densidades más parecidas a la real, recogiendo mejor el comportamiento de los datos.



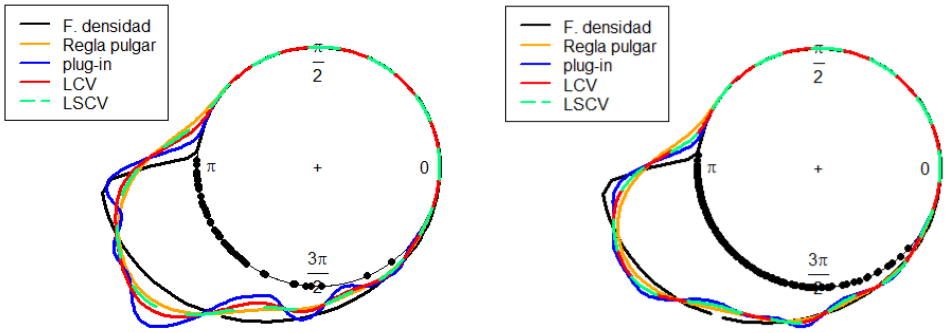
(a) $n = 50$ (b) $n = 200$

Figura 5.6: Representación de la función de densidad de una distribución mixtura de Von-Mises (curva negra) y de las estimaciones de esta obtenidas con distintos parámetros de suavizado. Se utiliza una muestra de tamaño (a) $n = 50$ y (b) $n = 200$.

Por otro lado, al calcular las estimaciones de dos muestras de tamaño 50 y 200 de una distribución normal enrollada asimétrica de parámetros $\mathcal{WSN}(\pi, 1, 20)$ vista en la figura 4.9, obtenemos las representaciones 5.7a y 5.7b, respectivamente. Volvemos a ver la clara diferencia entre las estimaciones obtenidas para cada tamaño muestral y vemos de nuevo que al aumentar este, la estimación mejora. En este caso, es el método *plug-in* el que proporciona una peor estimación, mientras que las otras son muy parecidas entre sí. Los parámetros de suavizado en este caso son:

- Muestra de tamaño 50: ν_{RP}=17.346, ν_{PI}=99.934, ν_{LCV}=34.118, ν_{LSCV}=23.345.
- Muestra de tamaño 200: ν_{RP}=24.356, ν_{PI}=99.935, ν_{LCV}=49.996, ν_{LSCV}=49.995.

Como curiosidad, notemos que los métodos de validación cruzada LSCV (5.13) y LCV (5.14) no tienen por qué proporcionar valores de ν similares.



(a) $n = 50$ (b) $n = 200$

Figura 5.7: Representación de la función de densidad de una distribución normal enrollada asimétrica (curva negra) y de las estimaciones de esta obtenidas con distintos parámetros de suavizado. Se utiliza una muestra de tamaño (a) $n = 50$ y (b) $n = 200$.

Capítulo 6

Aplicación a datos reales

En este capítulo vamos a aplicar los resultados vistos a algunos datos reales con el fin de poner el práctica los resultados que hemos introducido a lo largo del trabajo.

Para ello, hemos tomado dos muestras con alrededor de 300 observaciones de dos bases de datos distintas. En la primera estudiaremos la dirección de salida de las partículas que se producen en el LHC (*Large Hadron Collider*), mientras que en la segunda analizaremos la dirección de las corrientes marinas en Cabo de Peñas.

Comenzaremos con un análisis descriptivo de los datos para conocer sus principales características y observar la forma que podría tener su distribución. Posteriormente, estimaremos su función de densidad utilizando los métodos aprendidos en capítulo anterior y la compararemos con las distribuciones conocidas.

6.1. Dirección de las partículas que se producen tras la colisión en el LHC

El LHC (*Large Hadron Collider*) es el mayor acelerador de partículas hasta el momento. Está en funcionamiento desde septiembre de 2008 y se encuentra en el CERN, en la frontera entre Francia y Suiza. Está formado por una circunferencia de 27 kilómetros de longitud y se ubica a 100 metros bajo tierra. El objetivo de esta instalación es propulsar a las partículas a una gran velocidad para, posteriormente, hacerlas colisionar: adquieren dicha velocidad a través de distintos aceleradores para después pasar al LHC y hacer que colisionen en puntos específicos donde se encuentran los detectores, que recogen los rastros de las partículas resultantes. En

la figura[1] 6.1 se observa un dibujo esquemático de los distintos aceleradores y detectores que forman el CERN, en particular se pueden ver los del LHC. La toma de los datos utilizados para este estudie se hace en CMS (*Compact Muon Solenoid*).

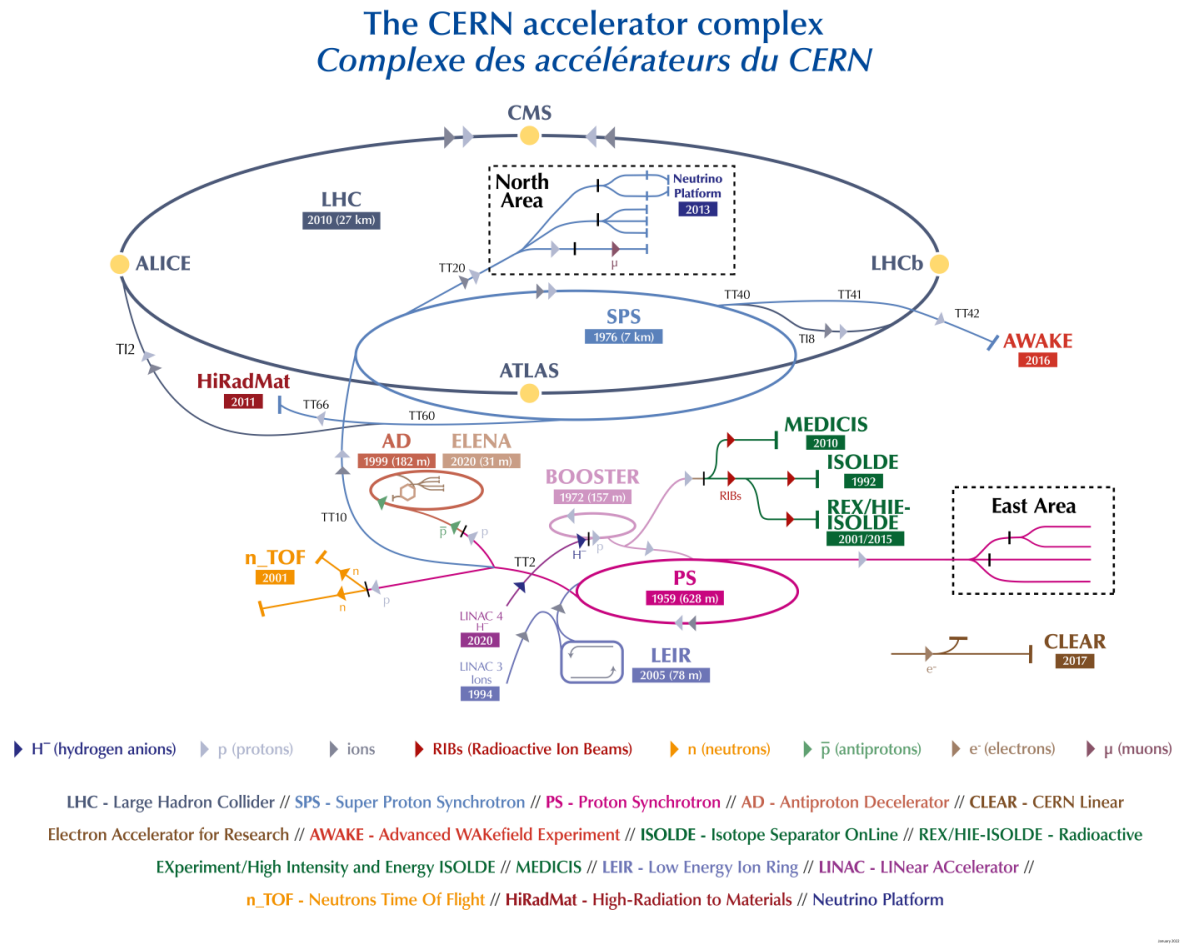


Figura 6.1: Esquema que muestra los distintos aceleradores y detectores que forman el complejo del CERN.

El producto de dicha colisión es un gran número de partículas entre las que se podrían encontrar algunas no conocidas hasta ahora. Es claro que esto no es lo común, pero este experimento también sirve para conocer mejor los procesos que ocurren a escalas de la materia muy pequeñas, así como tener mayor información acerca de las partículas fundamentales (como los electrones, los muones, etc.).

[1]Imagen extraída de la página web del CERN: https://home.cern/science/accelerators/accelerator-complex

En el LHC se hacen colisionar haces de protones principalmente, produciéndose, como hemos dicho, distintas partículas que salen impulsadas en diferentes direcciones entorno al eje en el que se produce la colisión. A este ángulo de salida, medido en radianes, se le denota como ϕ y toma valores en el intervalo $[0,2\pi)$. Esta es la variable del proceso que vamos a estudiar. Para comprender mejor cómo se toma este ángulo, podemos ver en la figura 6.2 el sistema de referencia utilizado en estas mediciones: representamos en verde el eje de colisión de los protones; en azul, los vectores de las distintas componentes del momento y en rojo, los distintos ángulos utilizados (en la imagen 6.2b la variable ϕ que vamos a analizar). Vamos a tomar como dirección cero el este geográfico y el sentido de giro antihorario. La muestra de tamaño 296 utilizada se ha tomado de la base de datos disponible en el Departamento de Física Experimental de Altas Energías de la Universidad de Oviedo[2].

(a) (b)

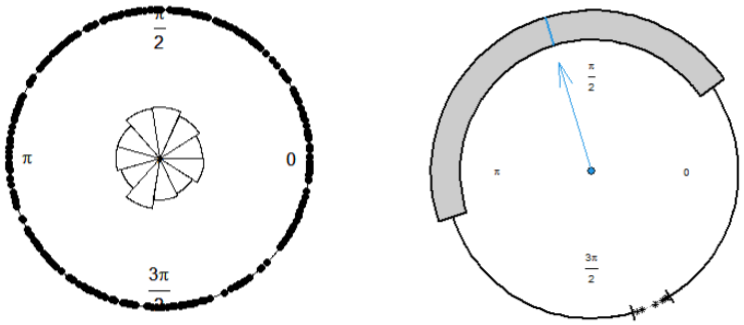
Figura 6.2: Sistema de referencia utilizado para hacer las mediciones en el experimento CMS del LHC. Vemos el eje de colisión de los protones representado en verde. Además, en (a) se observa en azul el vector del momento de una partícula producida tras la colisión y en rojo, en ángulo θ que forma este con el eje Z, y en (b) los vectores de las componentes del momento en el eje X e Y, junto con el momento transverso ($p_T = \sqrt{p_x^2 + p_y^2}$), en azul y el ángulo ϕ que forma este último con el eje X, en rojo.

Este estudio nos puede servir para confirmar que el experimento o el instrumental no tiene fallos pues cabría esperar que no exista una dirección preferencial, es decir, que todas las direcciones sean igualmente probables. Por tanto, si se encontrase una dirección en la que se concentrasen los ángulos de salida, podría existir un error de calibración en el sistema.

Empezamos representado el diagrama de rosa y de cajas de la muestra para observar la

[2]Página web del grupo de Física Experimental de Altas Energía de la Universidad de Oviedo: https://www.hep.uniovi.es/

forma de los datos. Atendiendo a dicho diagrama de rosa (véase la figura 6.3a) ya podemos intuir que los datos siguen una distribución uniforme circular, ya que, aunque hay direcciones con mayor número de observaciones, ninguna sobresale demasiado. Por otro lado, en la figura 6.3b vemos que la mediana (representada en azul) está cerca del centro de la caja, lo que nos lleva a pensar que la distribución de los datos puede ser simétrica, lo que nos proporciona más indicios a favor de la distribución uniforme. Sin embargo, recordemos que, en el caso de estar ante una distribución uniforme, este tipo de diagramas de caja no serían apropiados para representar estos datos.



(a) (b)

Figura 6.3: Diagramas de rosa (izquierda) y de cajas (derecha) asociados a la muestra de la dirección de salida de las partículas resultantes de la colisión en el LHC.

Continuemos obteniendo algunos de los parámetros de localización muestrales. Aplicando las fórmulas de las medias de los senos y los cosenos muestrales dadas en (3.4), tenemos que $\overline{S} = 0.059$ y $\overline{C} = -0.015$. Notemos que ambos coeficientes son cercanos a cero, por lo que cabría pensar que estamos cerca del último caso de la definición de la media muestral dada por la expresión (3.5). Esto último podría ser otra evidencia de que los datos se distribuyen uniformemente en torno a la circunferencia. No obstante, utilizando la última ecuación, tenemos que la media muestral es $\overline{\phi} = 1.828$ rad, que está cerca de la mediana muestral $\tilde{\phi} = 1.863$ rad, que hemos calculado mediante el *software* R.

Por otro lado, al utilizar la ecuación de la longitud media resultante muestral (3.11), obtenemos que nuestro parámetro de concentración es $\overline{R} = 0.061$. Este resultado nos indica que las observaciones muestrales están muy cerca de encontrarse uniformemente distribuidas en torno a la circunferencia, ya que $\overline{R}$ tiene un valor cercano a cero.

Por último, vamos a utilizar los métodos no paramétricos para obtener una estimación de la función de densidad de la muestra $\hat{f}$ y compararla con la de la distribución uniforme. Esto es lo que podemos ver en la figura 6.4, donde apenas observamos diferencias entre las

estimaciones y la función de densidad de la uniforme. Notemos que, como en este caso parece que tenemos una distribución simétrica, el parámetro de suavizado calculado mediante la regla del pulgar proporciona un resultado de $\hat{f}$ adecuado. De hecho, excepto la regla *plug-in*, los métodos empleados para obtener ν devuelven resultados similares de este: $\nu_{RP} = 0.180$, $\nu_{PI} = 2.916$, $\nu_{LCV} = 0.756$, $\nu_{LSCV} = 0.947$.

Fijémonos también en que los parámetros de suavizado obtenidos para esta muestra son bastante más pequeños. Esto concuerda con el hecho de que ν es un parámetro de concentración que hemos mencionado en el capítulo 5, ya que parece que tenemos una distribución uniforme, los datos no están concentrados en torno a ningún punto (ν menor).

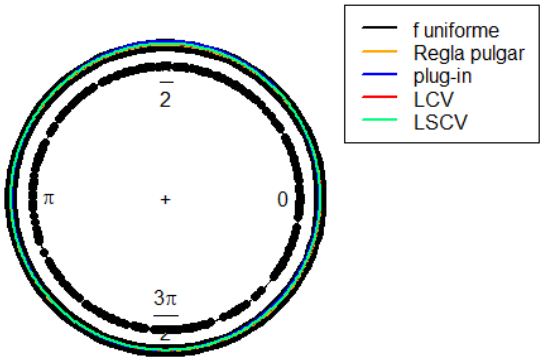


Figura 6.4: Comparación entre la función de densidad de una distribución uniforme y las estimaciones de la función de densidad de la que procede la muestra de la dirección de salida de las partículas resultantes de la colisión en el LHC.

Para poder confirmar que realmente las observaciones se distribuyen uniformemente, no basta con lo visto hasta ahora, sino que es necesario realizar un test de bondad de ajuste. Esto es, un test que indique si los datos siguen una distribución uniforme. Así, nuestras hipótesis nula y alternativa serán respectivamente:

H_0: los datos siguen una distribución uniforme,
H_1: los datos no siguen una distribución uniforme.

En este caso vamos a utilizar el test de Kuiper, cuya información se puede consultar en Stephens (1970), que se trata de un test de uniformidad en el círculo para una muestra. Fijamos el nivel de significación[3] $\alpha = 0.05$ y realizamos este test a la muestra mediante el *software* R, en

[3]El nivel de significación es una cota superior del error de tipo I, es decir, de rechazar la hipótesis nula (los datos siguen una distribución uniforme) cuando esta es cierta.

el que se obtiene que no se rechaza la hipótesis nula. Por tanto, efectivamente, los datos siguen una distribución uniforme.

Con todo lo visto en esta sección podemos concluir que las partículas resultantes de las colisiones en el LCH no se dispensan formando un ángulo concreto, sino que lo hacen aleatoria y uniformemente en todas las direcciones. Esto concuerda con lo esperado ya que en la física de este proceso no existe una dirección preferencial.

6.2. Dirección de las corrientes marinas en el Cabo de Peñas

En este segundo ejemplo vamos a estudiar la dirección de las corrientes marinas en el Cabo de Peñas (Asturias) entre los días 13 y 25 de marzo de 2024. La muestra está formada por 311 observaciones que se han tomado de la página oficial de Puertos del Estado[4] y se miden desde una boya 2242 ubicada al norte del Cabo de Peñas a 615 m de la costa, como se puede ver en la figura[5] 6.5. La frecuencia de muestreo es de una observación por hora, es decir, cada dato se corresponde con una hora de los días mencionados. Están medidos en grados y se toma como dirección cero el norte geográfico y el sentido de giro horario.

Comenzamos con un análisis exploratorio de la muestra. En las figura 6.6a y 6.6b, vemos el diagrama de rosa y el de cajas asociado a esta, respectivamente. Atendiendo al primero parece que los datos no proceden de una distribución simétrica, pero este hecho se puede apreciar mejor en el segundo ya que la mediana (representada en azul) no coincide con el centro de la caja. También tenemos una mayor concentración de datos en el tramo 270° - 360°, aunque en torno a los 90° parece haber otro grupo modal, menor que el anterior. Esto sugiere que la moda puede no ser única en este caso.

Calculamos ahora los distintos parámetros muestrales para obtener algo más de información de las corrientes en Cabo de Peñas.

Utilizando las fórmulas de las medias de los senos y los cosenos muestrales dadas en (3.4), obtenemos que $\overline{S} = -0.146$ y $\overline{C} = 0.269$. Así, la media muestral es $\overline{\theta} = 331.51°$ (representado con un punto rojo en la figura 6.6a) y el parámetro de concentración, $\overline{R} = 0.306$, que resultan de aplicar las expresiones de la media muestral y la longitud resultante muestral, (3.5) y (3.11),

[4]Esta información ha sido sacada de la página web de Puertos del Estado perteneciente al Gobierno de España https://www.puertos.es/es-es/oceanografia/Paginas/portus.aspx

[5]Imagen obtenida de la página web de Puertos del estado.

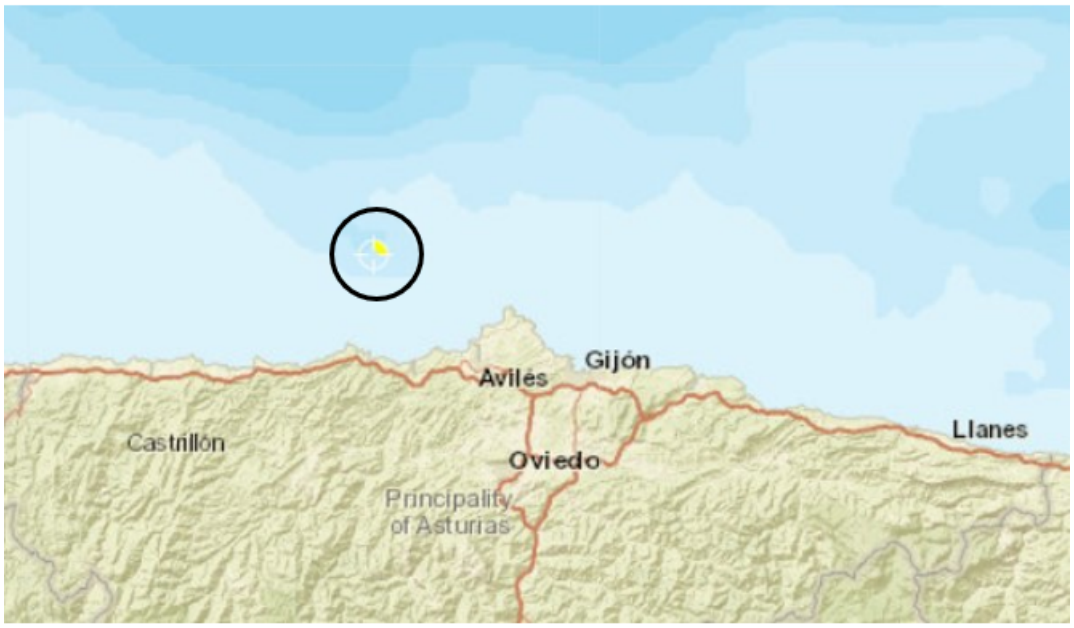
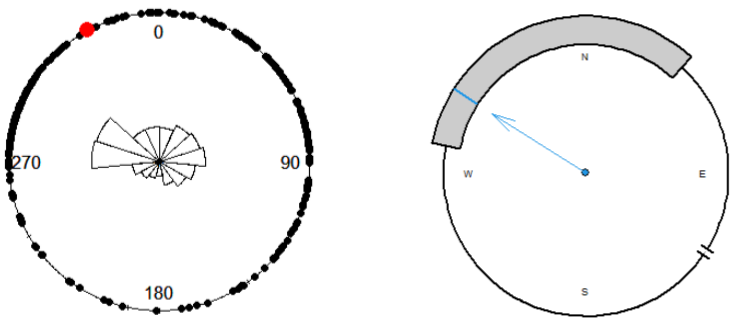


Figura 6.5: Ubicación de la boya (rodeada con un círculo negro) desde la que se han tomado los datos de este estudio.



(a) (b)

Figura 6.6: Diagramas de rosa (izquierda) y de cajas (derecha) asociados a la muestra de dirección de las corrientes marinas en Cabo de Peñas. En el primero señalamos con un punto rojo la dirección media y en el segundo, la mediana con una flecha azul.

respectivamente. Observamos que $\overline{\theta}$ se encuentra cerca del rango de ángulos donde veíamos que había mayor número de observaciones, pero se ve claramente influenciada por la posible existencia de otra moda, cayendo la media entre ambas modas. Por otra parte, la mediana sí parece coincidir con la mayor de las dos modas. Además, el valor de $\overline{R}$ nos indica que parece que los datos no están uniformemente distribuidos en torno a la circunferencia.

Sin embargo, como el valor de $\overline{R}$ tampoco es cercano a uno, que es el caso en el que la distribución de los datos se alejaría mucho de la uniforme, conviene realizar un test de unifor-

midad, al igual que en el ejemplo anterior. Fijando también el nivel de significación $\alpha = 0.05$ y pasando el test de Kuiper a muestra, se rechaza la hipótesis de que la distribución de los datos sea uniforme.

Podemos plantearnos ahora buscar una estimación de una función de densidad apropiada para estos datos y también para ver si estos métodos detectan una o dos modas. En la figura 6.7 tenemos varias de estas estimaciones $\hat{f}$ calculadas utilizando los distintos parámetros de suavizado vistos en el capítulo 5, que resultan: $\nu_{RP} = 2.558$, $\nu_{PI} = 99.961$, $\nu_{LCV} = 35.414$, $\nu_{LSCV} = 49.995$.

Claramente, la $\hat{f}$ obtenida mediante la regla del pulgar es bastante diferente a las demás. De hecho, es el único método que no refleja la existencia de un segunda moda cerca de los 90°. Ya hemos mencionado anteriormente que cuando tenemos una muestra que no procede de una distribución unimodal y simétrica, la regla del pulgar puede no proporcionar una estimación acertada. Además, este efecto se acentúa aún más cuando tenemos una muestra bimodal con modas antipodales. Aunque esto no ocurre en la muestra que estamos estudiando, sí que hemos visto que cerca del punto de la circunferencia opuesto a la moda hay un pequeño grupo modal, lo que puede afectar del mismo modo a la estimación.

Por otro lado, las funciones de densidad estimadas utilizando la regla *plug-in*, LSCV y LCV son bastante similares. Todas ellas muestran que la distribución posee dos grupos modales, uno con mayor probabilidad situado cerca de 270° y otro menor en torno a 90°. Sin embargo, la posición de estos en la circunferencia hace que la distribución no sea simétrica. También se puede ver que en la parte baja de la circunferencia, donde veíamos en 6.6a que había menor cantidad de observaciones, apenas hay densidad de probabilidad. Esto es lo que cabría esperar ya que, observando el mapa de la figura 6.5, tenemos que al sur de la boya está la costa cantábrica lo que dificulta q haya corrientes procedentes de esta dirección.

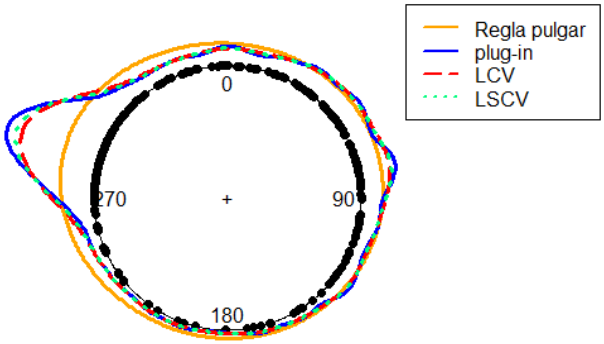


Figura 6.7: Representación de las estimaciones de la función de densidad de la que procede la muestra de las direcciones de las corrientes marinas en Cabo de Peñas.

Por tanto, podemos concluir que en Cabo de Peñas se producen corrientes en todas direcciones, aunque llegan un mayor número de ellas en el noroeste y en el este, aunque en menor medida. Por otro lado, también podemos decir que en la zona costera (al sur) es donde se originan menos corrientes. Como se puede ver en el estudio Ruiz-Villarreal *et al.* (2004) sobre las corrientes marinas en el mar Cantábrico, aproximadamente de marzo a julio las corrientes son ecuatoriales y en sentido noroeste. La fecha en las que han sido tomados los datos que estamos estudiando coinciden con esta época del año, así como las conclusiones obtenidas. Otra posible explicación de estos resultados podría ser que la boya desde la que se están tomando los datos se localiza en una zona marítima algo más profunda, denominada Cañón de Avilés, tal y como se aprecia en la figura 6.5.

También podemos ver en este estudio que de octubre a enero las corrientes son polares y más orientadas hacia el este, lo que podría explicar el pequeño grupo modal observado en esta dirección.

Capítulo 7

Conclusiones

A lo largo de todo este trabajo hemos estudiado en profundidad las principales características de los datos circulares, así como algunos de los métodos estadísticos utilizados en su análisis. Hemos visto cómo ha sido posible adaptar varias de las definiciones utilizadas en la estadística clásica al caso circular. Este es el caso de las funciones de distribución y densidad (vistas en el capítulo 2), entre otras, cuyas definiciones en ambas ramas poseen propiedades comunes, pero a las que se les debe imponer algunas nuevas en el caso circular (como la periodicidad).

En cambio, los estimadores derivados de la estadística no circular no son apropiados en este nuevo contexto, ya que no tienen en cuenta las particularidades de los datos que estamos estudiando. Por ello, en el capítulo 3, se muestran nuevos métodos considerando dichas características y se muestra su buen funcionamiento a través de varios ejemplos. Concretamente, la necesidad de una correcta definición de la media queda patente en el ejemplo 2.2, donde se observa que para la misma muestra, simplemente variando la dirección cero y el sentido de rotación, la dirección media cambia, lo que no es lógico en un contexto circular. Una expresión adecuada de esta para datos circulares fue introducida en la sección 3.1.1. Con respecto a las medidas de posición, también hemos estudiado otros parámetros como la mediana, que es única cuando se le impone el cumplimiento algunas propiedades, o los cuantiles, complicados de definir e identificar y para los cuales todavía no existe una forma concreta de estimarlos.

Por otro lado, la medida de dispersión más utilizada en este tipo de datos es el parámetro de concentración R (introducido en la sección 3.3), en lugar de la varianza o la desviación típica empleadas más en estadística clásica. Este parámetro de concentración, que mide cuánto se aleja la distribución de los datos de la uniforme, toma más en consideración las características propias de las muestras circulares. Aún así, también se han definido medidas similares a la varianza y desviación típica al caso circular, teniendo en cuenta las particularidades de este. La definición de esta medida de concentración hace que la distribución uniforme adquiera una importancia que no

posee en estadística no circular, ya que siempre se compara la distribución de los datos con ella a través de este parámetro. Cuando estamos ante datos uniformemente distribuidos, no existe ninguna dirección que acumule más probabilidad, su media no está definida y no existe ninguna moda y no podemos extraer más información sobre la muestra. Sin embargo, al rechazar la uniformidad de los datos podemos hacer un estudio más completo de ellos obteniendo el número de modas que posee la muestra, su media, su parámetro de concentración, su distribución, etc.

En el capítulo 4 también hemos considerado otros modelos apropiados en el contexto circular. Hemos visto que las distribuciones utilizadas en la estadística clásica tampoco son adecuadas en el caso que estamos estudiado ya que, por ejemplo, estas no respetan la periodicidad de los datos. Algunos de los nuevos modelos introducidos se definen específicamente para variables circulares, como la Von-Mises o la cardioide, mientras que otras se obtienen a partir de algunas funciones de densidad de desarrolladas para variables lineales, como la normal o Cauchy enrolladas. Observamos que todas estas distribuciones tienen en común la característica de ser unimodales y simétricas y, por ello, surge la necesidad de incluir otras que puedan ser útiles cuando las muestras son multimodales o asimétricas. Así, se introducen distribuciones proyectadas, la normal enrollada asimétrica y las mixturas de Von-Mises.

La función de densidad caracteriza inequívocamente el comportamiento de una variable aleatoria, por lo que cuando estamos ante una muestra de la cual no conocemos distribución, es inevitable tener que estimarla a través de métodos no paramétricos. Así, en el capítulo 5 estudiamos estos métodos. Sin embargo, hemos visto que obtener una buena estimación no es sencillo, ya que el estimador utilizado depende de un parámetro de suavizado que se debe calcular. El problema reside en escoger un valor adecuado de este. En el proceso de búsqueda de un valor óptimo de dicho parámetro suele estar involucrada la propia función de densidad de los datos, que al ser desconocida complica el cálculo. Una opción en este punto es introducir la densidad de Von-Mises en lugar de la de los datos. Los diferentes métodos proporcionan diferentes parámetros de suavizado y algunos de ellos no con apropiados en algunos casos, como la regla del pulgar cuando estamos ante una muestra multimodal o asimétrica.

La utilidad de todos los resultados mostrados a lo largo del trabajo se muestra en el capítulo 6, donde ponemos en práctica estos resultados con dos ejemplos de muestras reales muy diferentes entre sí. La primera de las muestras consta datos procedentes del experimento CMS del CERN. Más concretamente, son ángulos de salida de 300 partículas producidas en las colisiones de protones en el LHC. Observamos que los datos seguían una distribución uniforme, de forma que no existía una dirección de salida preferencial de dichas partículas, tal y como cabría esperar. Comprobamos así que, efectivamente, el experimento está bien calibrado y funciona

adecuadamente, sin defectos que lleven a resultados físicos erróneos.

Con la segunda muestra estudiamos las direcciones de corrientes marinas en el Cabo de Peñas producidas en doce días del pasado mes de marzo. Estos datos se recogen en una boya situada al norte de la costa asturiana con una frecuencia de una observación por hora y, a través de ellos, queremos analizar si existe algún punto geográfico donde se concentre una mayor densidad de corrientes. Aplicando un análisis no paramétrico fuimos capaces de estimar la función de densidad de esta muestra y observar una moda notoria y la posibilidad de una segunda más pequeña. La existencia de estas modas implica que se producen más corrientes en el noroeste lo que coincide con lo observado en Ruiz-Villarreal *et al.* (2004) en la época del año en la que se han tomado los datos (marzo). Sin embargo, no hemos tenido en cuenta otras variables que pueden afectar a la corrientes marinas como podrían ser los vientos, el mapa mundial de corrientes procedentes de otro mares u océanos o, como ya hemos mencionado, el período del año en el que se tomen los datos. La incorporación de más de una variable de estudio y su análisis conjunto nos llevaría a hablar de regresión con datos circulares, que se puede ver en Yu *et al.* (2014).

Otro aspecto que no hemos considerado es que las corrientes no solo se producen en la superficie del mar, si no también en zonas más profundas. Por tanto, sería conveniente realizar un estudio de estas en las tres dimensiones del espacio. En este caso, estaríamos ante datos esféricos o, más generalmente, datos direccionales, los cuales se estudian en Mardia y Jupp (2000).

El estudio de la regresión con datos circulares o de dichos datos direccionales serían unas posibles vías de ampliación de este trabajo. También son interesantes los modelos de distribuciones circulares multivariantes como las que se muestran en Kim *et al.* (2016) junto con algunas de sus aplicaciones. Estos serían útiles, por ejemplo, si tuviésemos en cuenta variables atmosféricas además de la dirección, lo que nos lleva nuevamente a modelos de regresión.

Apéndice A

Código implementado

En este apéndice se recogen los códigos (en el lenguaje de programación R) utilizados en las simulaciones en las que nos hemos apoyado para realizar los análisis mostrados en el trabajo. Se han empleado varias librerías del *software* R, estas son: "Circular" (Agostinelli y Lund, 2023), "bpDir" (Buttarazzi, 2021) y "NPCirc" (Oliveira *et al.*, 2014).

El apéndice consta de cuatro secciones que se corresponden con las simulaciones de los análisis realizados en los capítulo 3, 4, 5 y 6, respectivamente.

A.1. Estadística descriptiva circular

- Cálculo de las medidas de posición y dispersión de las muestras de datos vistas en la figura 3.1.

```
#Muetras de datos
a<-c(90,150,315)
b<-c(0,135,300)

#Parámetros C y S de cada muestra
Ca=1/3*sum(cos(a*pi/180))
Sa=1/3*sum(sin(a*pi/180))
Cb=1/3*sum(cos(b*pi/180))
Sb=1/3*sum(sin(b*pi/180))

#Dirección media de cada muestra
Ma<-(atan(Sa/Ca)+pi)*180/pi
Mb<-(atan(Sb/Cb)+2*pi)*180/pi

#Parámetro de concentración
Ra<-sqrt(Ca^2+Sa^2)
```

```
Rb<-sqrt(Cb^2+Sb^2)

#Varianza y desviación típica de cada muestra
Va<-1-Ra
Vb<-1-Rb
dsa<-sqrt(-2*log(Ra))
dsb<-sqrt(-2*log(Rb))

#Varianza y desviación típica con la definición alterna de cada muestra
VaB<-2*Va
VbB<-2*Vb
dsaB<-sqrt(VaB)
dsbB<-sqrt(VbB)
```

- Cálculo de las medidas de posición y dispersión de las muestras de datos vistas en la figura 3.3.

```
#Muetras de datos
a<-c(60,180,300)
b<-c(0,20,340)

#Parámetros C y S de las muestras
Ca=1/3*sum(cos(a*pi/180))
Sa=1/3*sum(sin(a*pi/180))
Cb=1/3*sum(cos(b*pi/180))
Sb=1/3*sum(sin(b*pi/180))

#Dirección media de las muestras
Ma<-(atan(Sa/Ca)+pi)*180/pi
Mb<-(atan(Sb/Cb)+2*pi)*180/pi

#Parámetro de concentración de las muestras
Ra<-sqrt(Ca^2+Sa^2)
Rb<-sqrt(Cb^2+Sb^2)

#Varianza y desviación típica de cada muestra
Va<-1-Ra
Vb<-1-Rb
dsa<-sqrt(-2*log(Ra))
dsb<-sqrt(-2*log(Rb))
```

- Diagrama de puntos lineal y circular representados en la figura 3.4.

```
library(circular)

aire<-circular(wind*180/pi, type='angles',units='degrees',zero=pi/2,rotation = 'clock')
plot(wind, xlab='Número de observaciones', ylab='Dirección del viento (rad)',pch=20)
plot(aire, cex=1.5, bin=720, stack=TRUE, sep=0.035, shrink=1.3,pch=20)
```

- Diagramas de rosa representados en la figura 3.6

```
library(circular)

aire<-circular(wind*180/pi, type='angles',units='degrees',zero=pi/2,rotation = 'clock')
rose.diag(aire, bins=10, cex=1.5, prop=1.3,add=T)  #10 sectores
rose.diag(aire, bins=20, cex=1.5, prop=1.3,add=T)  #20 sectores
```

- Histograma lineal representado en la figura 3.7.

```
hist(wind*180/pi,breaks=20,ylab='Frecuencia',xlab='Dirección del viento (º)',main='')
```

- Rangos representados en la figura 3.8.

```
library(circular)

x<-rcircularuniform(11)
plot(x)
arrows.circular(median.circular(x),add=T,col='Blue')
arrows.circular(median.circular(x)-pi,add=T,col='Green3')
```

- Diagramas de cajas representados en la figura 3.9.

```
library(circular)
library(bpDir)

x<-rcircularuniform(11)
CircularBoxplot(x,template='radians',units='radians')
CircularBoxplot(x,template='radians',units='radians',constant=0.7)
```

A.2. Distribuciones circulares

- Función de densidad de la distribución uniforme representada en la figura 4.1.

```
library(circular)

curve.circular(dcircularuniform, join=TRUE, ylim=c(-1.05, 1.05), lwd=2,col='blue')
title(main='Función de densidad')
```

- Funciones de densidad de la distribución cardioide variando el parámetro ρ representadas en la figura 4.2a.

```
library(circular)

mu=circular(pi/2)
rho1=0.1
rho2=0.3
rho3=0.49
x<-seq(0,2*pi,length.out=1000)
x<-circular(x)
curve.circular(dcardioid(x,mu,rho1), join=TRUE, ylim=c(-1.2, 1.2), lwd=2,col='blue')
curve.circular(dcardioid(x,mu,rho2), join=TRUE, ylim=c(-1.2, 1.2), lwd=2, add=T,
    col='springgreen')
curve.circular(dcardioid(x,mu,rho3), join=TRUE, ylim=c(-1.2, 1.2), lwd=2, add=T,
    col='red')
legend(x = "topright",legend = c('rho=0.1','rho=0.3','rho=0.49'),lty = c(1,1,1),
    col=c('blue','springgreen','red'), lwd = 2)
title(main='Función de densidad')
```

- Funciones de densidad de la distribución cardioide variando el parámetro μ representadas en la figura 4.2b.

```
library(circular)

mu1<-circular(pi/2)
mu2<-circular(3*pi/2)
rho<-0.49
curve.circular(dcardioid(x,mu1,rho), join=TRUE, ylim=c(-1.2, 1.2), lwd=2, col='blue')
curve.circular(dcardioid(x,mu2,rho), join=TRUE, ylim=c(-1.2, 1.2), lwd=2, add=T,
    col='springgreen')
legend(x = "topright",legend = c('mu=pi/2','mu=3pi/2'),lty = c(1, 1, 1),
    col=c('blue','springgreen'), lwd = 2)
title(main='Función de densidad')
```

- Funciones de densidad de la distribución de Von-Mises variando el parámetro μ representadas en la figura 4.3a.

```
library(circular)

mu1<-circular(pi/4)
mu2<-circular(pi)
kappa<-2
curve.circular(dvonmises(x,mu1,kappa), join=TRUE, ylim=c(-1.3, 1.3), lwd=2, col='blue')
```

```
curve.circular(dvonmises(x,mu2,kappa), join=TRUE, ylim=c(-1.3, 1.3), lwd=2, add=T,
    col='springgreen')
legend(x = "topright",legend = c('mu=pi/4','mu=pi'),lty = c(1, 1, 1)
    ,col=c('blue','springgreen'),lwd = 2)
title(main='Función de densidad')
```

- Funciones de densidad de la distribución de Von-Mises variando el parámetro κ representadas en la figura 4.3b.

```
library(circular)

mu<-circular(0)
kappa1<-0.5
kappa2<-1.5
kappa3<-4
x<-seq(0,2*pi,length.out=1000)
x<-circular(x)
curve.circular(dvonmises(x,mu,kappa1), join=TRUE, ylim=c(-1.2, 1.2), lwd=2,col='blue')
curve.circular(dvonmises(x,mu,kappa2), join=TRUE, ylim=c(-1.2, 1.2),
    lwd=2,add=T,col='springgreen')
curve.circular(dvonmises(x,mu,kappa3), join=TRUE, ylim=c(-1.2, 1.2),
    lwd=2,add=T,col='red')
legend(x = "topright",legend = c('kappa=0.5','kappa=1.5','kappa=4'),lty = c(1, 1, 1),
    col=c('blue','springgreen','red'),lwd = 2)
title(main='Función de densidad')
```

- Ejemplo de la elación entre la distribución de Von Mises y la cardioide (figura 4.3c).

```
library(circular)

mu<-circular(pi)
kappa<-0.75
x<-seq(0,2*pi,length.out=1000)
x<-circular(x)
curve.circular(dvonmises(x,mu,kappa), join=TRUE, ylim=c(-1.2, 1.2), lwd=2,col='blue')
curve.circular(dcardioid(x,mu,kappa/2), join=TRUE, ylim=c(-1.2, 1.2), lwd=2, add=T,
    col='green')
legend(x = "topright",legend = c('Von Mises','Cardioide'),lty = c(1, 1, 1),
    col=c('blue','springgreen'),lwd = 2)
title(main='Función de densidad')
```

- Funciones de densidad de la distribución normal enrollada variando el parámetro μ representadas en la figura 4.4a.

```
library(circular)

mu1<-circular(0)
mu2<-circular(pi/6)
rho<-0.8
curve.circular(dwrappednormal(x,mu1,rho), join=TRUE, ylim=c(-1.2, 1.2), lwd=2, col='blue')
curve.circular(dwrappednormal(x,mu2,rho), join=TRUE, ylim=c(-1.2, 1.2), lwd=2, add=T,
    col='springgreen')
legend(x = "topright",legend = c('mu=0','mu=pi/6'),lty = c(1, 1, 1),
    col=c('blue','springgreen'), lwd = 2)
title(main='Función de densidad')
```

- Funciones de densidad de la distribución normal enrollada variando el parámetro ρ representadas en la figura 4.4b.

```
library(circular)

mu<-circular(4*pi)
rho1<-0.2
rho2<-exp(-1/2)
rho3<-0.9
x<-seq(0,2*pi,length.out=1000)
x<-circular(x)
curve.circular(dwrappednormal(x,mu,rho1), join=TRUE, ylim=c(-1.2, 1.2), lwd=2, col='blue')
curve.circular(dwrappednormal(x,mu,rho2), join=TRUE, ylim=c(-1.2, 1.2), lwd=2, add=T,
    col='springgreen')
curve.circular(dwrappednormal(x,mu,rho3), join=TRUE, ylim=c(-1.2, 1.2), lwd=2, add=T,
    col='red')
legend(x = "topright",legend = c('rho=0.2','rho=exp(-1/2)','rho=0.9'),lty = c(1, 1, 1),
    col=c('blue','springgreen','red'), lwd = 2)
title(main='Función de densidad')
```

- Funciones de densidad de la distribución Cauchy enrollada variando el parámetro μ representadas en la figura 4.5a.

```
library(circular)

mu1<-circular(0)
mu2<-circular(7*pi/6)
rho<-0.6
curve.circular(dwrappedcauchy(x,mu1,rho), join=TRUE,ylim=c(-1.5, 1.5), lwd=2,col='blue')
```

```
curve.circular(dwrappedcauchy(x,mu2,rho), join=TRUE,ylim=c(-1.5, 1.5), lwd=2, add=T,
    col='springgreen')
legend(x = "topright",legend = c('mu=0','mu=7pi/6'),lty = c(1, 1, 1),
    col=c('blue','springgreen'), lwd = 2)
title(main='Función de densidad')

```

- Funciones de densidad de la distribución Cauchy enrollada variando el parámetro ρ representadas en la figura 4.5b.

```
library(circular)

mu<-circular(pi/2)
rho1<-0.2
rho2<-0.5
rho3<-0.8
x<-seq(0,2*pi,length.out=1000)
x<-circular(x)
curve.circular(dwrappedcauchy(x,mu,rho1), join=TRUE,ylim=c(-2.15, 2.15), lwd=2,
    col='blue')
curve.circular(dwrappedcauchy(x,mu,rho2), join=TRUE,ylim=c(-1.5, 1.5), lwd=2, add=T,
    col='springgreen')
curve.circular(dwrappedcauchy(x,mu,rho3), join=TRUE,ylim=c(-1.5, 1.5), lwd=2, add=T,
    col='red')
legend(x = "topright",legend = c('rho=0.2','rho=0.5','rho=0.8'),lty = c(1, 1, 1),
    col=c('blue','springgreen','red'), lwd = 2)
title(main='Función de densidad')
```

- Funciones de densidad de la distribución normal proyectada variando el parámetro μ representadas en la figura 4.6a.

```
library(circular)

mu1<-c(-1,0)
mu2<-c(1,1)
sigma<-matrix(c(1,0,0,1),2)
curve.circular(dpnorm(x,mu1,sigma),join=TRUE,ylim=c(-1.5, 1.5), lwd=2, col='blue')
curve.circular(dpnorm(x,mu2,sigma),join=TRUE,ylim=c(-1.5, 1.5), lwd=2, add=T,
    col='springgreen')
legend(x = "topright",legend = c('mu1','mu2'),lty = c(1, 1), col=c('blue','springgreen'),
    lwd = 2)
title(main='Función de densidad')
```

- Funciones de densidad de la distribución normal proyectada variando el parámetro $\boldsymbol{\Sigma}$ representadas en la figura 4.6b.

```
library(circular)

mu<-c(1,1)
sigma1<-matrix(c(2,0,0,2),2)
sigma2<-matrix(c(6,0,0,6),2)
curve.circular(dpnorm(x,mu,sigma1),join=TRUE,ylim=c(-1.5, 1.5), lwd=2,col='blue')
curve.circular(dpnorm(x,mu,sigma2),join=TRUE,ylim=c(-1.5, 1.5), lwd=2, add=T,
    col='springgreen')
legend(x = "topright",legend = c('sigma1','sigma2'),lty = c(1, 1),
    col=c('blue','springgreen'), lwd = 2)
title(main='Función de densidad')
```

- Funciones de densidad de la distribución normal proyectada bimodales y simétricas representadas en la figura 4.7a.

```
library(circular)

mu<-c(0,0)
sigma1<-matrix(c(2,0,0,6),2)
sigma2<-matrix(c(1,0.5,0.5,1),2)
curve.circular(dpnorm(x,mu,sigma1),join=TRUE,ylim=c(-1.5, 1.5), lwd=2, col='blue')
curve.circular(dpnorm(x,mu,sigma2),join=TRUE,ylim=c(-1.5, 1.5), lwd=2, add=T,
    col='springgreen')
legend(x = "topright",legend = c('sigma1','sigma2'),lty = c(1, 1),
    col=c('blue','springgreen'), lwd = 2)
title(main='Función de densidad')
```

- Funciones de densidad de la distribución normal proyectada asimétricas representadas en la figura 4.7.

```
library(circular)

mu1<-c(-1,0)
sigma1<-matrix(c(1,0.9,0.9,1),2)
mu2<-c(3,2)
sigma2<-matrix(c(4,0.3,0.3,1),2)
curve.circular(dpnorm(x,mu1,sigma1),join=TRUE,ylim=c(-1.5, 1.5), lwd=2, col='blue')
curve.circular(dpnorm(x,mu2,sigma2),join=TRUE,ylim=c(-1.5, 1.5), lwd=2, add=T,
    col='springgreen')
legend(x = "topright",legend = c('mu1, sigma1','mu2, sigma2'), lty = c(1, 1),
    col=c('blue','springgreen'), lwd = 2)
```

```
title(main='Función de densidad')
```

- Comparación de funciones de densidad representadas en la figura 4.8.

```
library(circular)

mu<-0
rho<-0.5
kappa<-0.5
mu2<-c(1,0)
sigma<-matrix(c(1,0,0,1),2)

curve.circular(dcircularuniform, join=TRUE, ylim=c(-1.2, 1.2), lwd=2, col='blue')
curve.circular(dcardioid(x,mu,rho), join=TRUE, ylim=c(-1.2, 1.2), lwd=2, add=T,
    col='springgreen')
curve.circular(dvonmises(x,mu,kappa), join=TRUE, ylim=c(-1.2, 1.2), lwd=2, add=T,
    col='red')
curve.circular(dwrappednormal(x,mu,rho), join=TRUE, ylim=c(-1.2, 1.2), lwd=2, add=T,
    col='purple')
curve.circular(dwrappedcauchy(x,mu,rho), join=TRUE,ylim=c(-1.5, 1.5), lwd=2, add=T,
    col='darkorange1')
curve.circular(dpnorm(x,mu2,sigma),join=TRUE,ylim=c(-1.5, 1.5), lwd=2,add=T, col='177')
legend(x = "topleft",legend = c('Uniforme','Cardioide','Von Mises','Normal enrollada',
    'Cauchy enrollada', 'Normal proyectada'), lty = c(1, 1, 1, 1, 1, 1),
    col=c('blue','springgreen','red','purple','orange', '177'), lwd = 2)
title(main='Comparación de funciones de densidad')
```

- Funciones de densidad de la distribución normal enrollada asimétrica variando el parámetro λ representadas en la figura 4.9.

```
library(circular)
library(NPCirc)

mu<-pi
kappa<-1
lambda1<-0
lambda2<-5
lambda3<-20
curve.circular(dwsn(x,mu,kappa,lambda1),join=TRUE,ylim=c(-1.5, 1.5), lwd=2, col='blue')
curve.circular(dwsn(x,mu,kappa,lambda2),join=TRUE,ylim=c(-1.5, 1.5), lwd=2, add=T,
    col='springgreen')
curve.circular(dwsn(x,mu,kappa,lambda3),join=TRUE,ylim=c(-1.5, 1.5), lwd=2, add=T,
    col='red')
```

```
legend(x = "topright",legend = c('lambda=2','lambda=5','lambda=20'),lty = c(1, 1, 1),
    col=c('blue','springgreen','red'),lwd = 2)
title(main='Función de densidad')
```

- Mixturas de Von-Mises representadas en la figura 4.10.

```
library(circular)

f1<-function(theta) {1/3*dvonmises(theta,pi/3,6) + 1/3*dvonmises(theta,pi,6) +
    1/3*dvonmises(theta,5*pi/3,6)}
curve.circular(f1(x),join=TRUE,ylim=c(-1.5, 1.5), lwd=2, col='blue')
title(main='Función de densidad')

f2<-function(theta) {4/9*dvonmises(theta,2,3) + 5/36*dvonmises(theta,4,3) +
    5/36*dvonmises(theta,3.5,50)+
    5/36*dvonmises(theta,4,50) + 5/36*dvonmises(theta,4.5,50)}
curve.circular(f2(x),join=TRUE,ylim=c(-1.5, 1.5), lwd=2, col='blue')
title(main='Función de densidad')
```

A.3. Estimación no paramétrica

- Histogramas lineales que muestran el efecto de la elección del parámetro de ventana h representados en la figura 5.1.

```
#Muestra uniforme
u<-runif(100)

#Bins del histograma
Bk1<-seq(0, 1, by=0.05)       #h=0.05
Bk2<-seq(0, 1, by=0.1)        #h=0.1
Bk3<-seq(0, 1, by=0.25)       #h=0.25

#Representación de los histogramas
hist(u, breaks=Bk1,main='Histograma U(0,1)',xlab='',ylab='Frecuencia')
hist(u, breaks=Bk2,main='Histograma U(0,1)',xlab='',ylab='Frecuencia')
hist(u, breaks=Bk3,main='Histograma U(0,1)',xlab='',ylab='Frecuencia')
```

- Histogramas lineales que muestran el efecto de la elección del parámetro t_0 representados en la figura 5.2.

```
#Muestra uniforme
u<-runif(100)

#Bins del histograma
Bk1<-seq(0, 1, by=0.2)        #t0=0
Bk2<-seq(-0.1,1.1, by=0.2)    #t0=-0.1

#Representación de los histogramas
hist(u, breaks=Bk1,main='Histograma U(0,1)',xlab='',ylab='Frecuencia')
hist(u, breaks=Bk2,main='Histograma U(0,1)',xlab='',ylab='Frecuencia')
```

- Representaciones de las estimaciones de la función de densidad de una normal estándar para distintas elecciones del parámetro de ventana h representadas en la figura 5.3.

```
#Muestras procedentes de una normal estándar
m1<-rnorm(200)
m2<-rnorm(200)
m3<-rnorm(200)
m4<-rnorm(200)
m5<-rnorm(200)

#Parámetros de ventana
bw1<-1
bw2<-0.5

#Estimaciones de la función de densidad para bw=1
plot(dnorm,from=-4,to=4,col='black',lwd=3,ylab='Densidad N(0,1)')
lines(density(m1,bw=bw1),kernel='gaussian',lwd=2, col='blue')
lines(density(m2,bw=bw1),kernel='gaussian',lwd=2, col='deepskyblue')
lines(density(m3,bw=bw1),kernel='gaussian',lwd=2, col='cyan3')
lines(density(m4,bw=bw1),kernel='gaussian',lwd=2, col='blue3')
lines(density(m5,bw=bw1),kernel='gaussian',lwd=2, col='blue4')

#Estimaciones de la función de densidad para bw=0.5
plot(dnorm,from=-4,to=4,col='black',lwd=3,ylab='Densidad N(0,1)')
lines(density(m1,bw=bw2),kernel='gaussian',lwd=2, col='blue')
lines(density(m2,bw=bw2),kernel='gaussian',lwd=2, col='deepskyblue')
lines(density(m3,bw=bw2),kernel='gaussian',lwd=2, col='cyan3')
lines(density(m4,bw=bw2),kernel='gaussian',lwd=2, col='blue3')
lines(density(m5,bw=bw2),kernel='gaussian',lwd=2, col='blue4')
```

- Representaciones de las estimaciones de la función de densidad de una normal enrollada $\mathcal{WN}(0, 0.7)$ para distintas elecciones del parámetro de suavizado ν representadas en la figura 5.4.

```
library(circular)
library(NPCirc)
set.seed(1)

#Muestras procedentes de una distribución normal enrollada de media 0 y parámetro
#de concentración 0.7
m1<-rwrappednormal(200,0,0.7)
m2<-rwrappednormal(200,0,0.7)
m3<-rwrappednormal(200,0,0.7)
m4<-rwrappednormal(200,0,0.7)
m5<-rwrappednormal(200,0,0.7)

#Parámetros de suavizado
bw1=0.5
bw2=0.9

#Estimaciones de la función de densidad para bw=0.5
curve.circular(dwrappednormal(x,mu=circular(0),rho=0.7),join=TRUE, ylim=c(-1.2, 1.2),
    lwd=3,col='black')
lines(density.circular(x=m1,bw=bw1,kernel='wrappednormal'), lwd=1.5, col='blue')
lines(density.circular(x=m2,bw=bw1,kernel='wrappednormal'), lwd=1.5, col='deepskyblue')
lines(density.circular(x=m3,bw=bw1,kernel='wrappednormal'), lwd=1.5, col='cyan3')
lines(density.circular(x=m4,bw=bw1,kernel='wrappednormal'), lwd=1.5, col='blue3')
lines(density.circular(x=m5,bw=bw1,kernel='wrappednormal'), lwd=1.5, col='blue4')


#Estimaciones de la función de densidad para bw=0.9
curve.circular(dwrappednormal(x,mu=circular(0),rho=0.7),join=TRUE, ylim=c(-1.2, 1.2),
    lwd=3,col='black')
lines(density.circular(x=m1,bw=bw2,kernel='wrappednormal'), lwd=1.5, col='blue')
lines(density.circular(x=m2,bw=bw2,kernel='wrappednormal'), lwd=1.5, col='deepskyblue')
lines(density.circular(x=m3,bw=bw2,kernel='wrappednormal'), lwd=1.5, col='cyan3')
lines(density.circular(x=m4,bw=bw2,kernel='wrappednormal'), lwd=1.5, col='blue3')
lines(density.circular(x=m5,bw=bw2,kernel='wrappednormal'), lwd=1.5, col='blue4')
```

- Estimaciones de la función de densidad de la distribución de Von-Mises $\mathcal{VM}(0,3)$ obtenidas a partir de los parámetros de suavizado que resultan de los distintos métodos estudiados utilizando distintos tamaños muestrales representadas en la figura 5.5.

```
library(circular)
library(NPCirc)

set.seed(1)
mu<-circular(0)
kappa<-3

#Muestra Von-Mises de tamaño 50
mc1<-rvonmises(50,mu,kappa)

#Parámetros de suavizado
ps1<-bw.rt(mc1)                    #Regla del pulgar
ps2<-bw.pi(mc1)                    #Regla plug-in
ps3<-bw.CV(mc1,method='LCV')       #Validación cruzada por verosimilitud
ps4<-bw.CV(mc1,method='LSCV')      #Validación cruzada por mínimos cuadrados
plot(mc1)

#Estimaciones de la función de densidad
curve.circular(dvonmises(x,mu,kappa),add=TRUE,lwd=3)
lines(density.circular(x=mc1,bw=ps1,kernel='vonmises'),lwd=3,col='orange', add=TRUE)
lines(density.circular(x=mc1,bw=ps2,kernel='vonmises'),lwd=3,col='blue', add=TRUE)
lines(density.circular(x=mc1,bw=ps3,kernel='vonmises'),lwd=3,col='red', add=TRUE)
lines(density.circular(x=mc1,bw=ps4,kernel='vonmises'),lwd=3,col='springgreen', add=TRUE)
legend(x = "topleft",legend = c('F. densidad','Regla pulgar','plug-in','LCV', 'LSCV'),
    lty = c(1, 1, 1, 1, 1), col=c('black','orange','blue','red','springgreen'), lwd = 2)


#Muestra Von-Mises de tamaño 200
mc2<-rvonmises(200,mu,kappa)

#Parámetros de suavizado
ps5<-bw.rt(mc2)                    #Regla del pulgar
ps6<-bw.pi(mc2)                    #Regla plug-in
ps7<-bw.CV(mc2,method='LCV')       #Validación cruzada por verosimilitud
ps8<-bw.CV(mc2,method='LSCV')      #Validación cruzada por mínimos cuadrados

#Estimaciones de la función de densidad
plot(mc2)
curve.circular(dvonmises(x,mu,kappa),add=TRUE,lwd=3)
```

```
lines(density.circular(x=mc2,bw=ps5,kernel='vonmises'),lwd=3,col='orange', add=TRUE)
lines(density.circular(x=mc2,bw=ps6,kernel='vonmises'),lwd=3,col='blue', add=TRUE)
lines(density.circular(x=mc2,bw=ps7,kernel='vonmises'),lwd=3,col='red', add=TRUE)
lines(density.circular(x=mc2,bw=ps8,kernel='vonmises'),lwd=3,col='springgreen', add=TRUE)
legend(x = "topleft",legend = c('F. densidad','Regla pulgar','plug-in','LCV', 'LSCV'),
    lty = c(1, 1, 1, 1, 1), col=c('black','orange','blue','red','springgreen'), lwd = 2)
```

- Estimaciones de la función de densidad de la distribución de una mixtura de Von-Mises obtenidas a partir de los parámetros de suavizado que resultan de los distintos métodos estudiados utilizando distintos tamaños muestrales representadas en la figura 5.6.

```
library(circular)
library(NPCirc)
set.seed(1)

#Muestra mixtura de Von-Mises de tamaño 50
mm<-rcircmix(50,model=11)

#Parámetros de suavizado
ps1<-bw.rt(mm)                    #Regla del pulgar
ps2<-bw.pi(mm)                    #Regla plug-in
ps3<-bw.CV(mm,method='LCV')       #Validación cruzada por verosimilitud
ps4<-bw.CV(mm,method='LSCV')      #Validación cruzada por míninimos cuadrados
plot(mm,ylim=c(-1.2, 1.2))

#Estimaciones de la función de densidad
curve.circular(dcircmix(x,model=11),add=TRUE,ylim=c(-1.5, 1.5), lwd=3,col='black')
lines(density.circular(x=mm,bw=ps1,kernel='vonmises'),lwd=3,col='orange', add=TRUE)
lines(density.circular(x=mm,bw=ps2,kernel='vonmises'),lwd=3,col='blue', add=TRUE)
lines(density.circular(x=mm,bw=ps3,kernel='vonmises'),lwd=3,col='red', add=TRUE, lty='aa')
lines(density.circular(x=mm,bw=ps4,kernel='vonmises'),lwd=3,col='springgreen', add=TRUE,
    lty=3)
legend(x = "topright",legend = c('F. densidad','Regla pulgar','plug-in','LCV', 'LSCV'),
    lty = c(1, 1, 1, 5, 3), col=c('black','orange','blue','red','springgreen'), lwd = 2)


#Muestra mixtura de Von-Mises de tamaño 200
mm<-rcircmix(200,model=11)

#Parámetros de suavizado
ps1<-bw.rt(mm)                     #Regla del pulgar
ps2<-bw.pi(mm)                     #Regla plug-in
ps3<-bw.CV(mm,method='LCV')        #Validación cruzada por verosimilitud
```

```
ps4<-bw.CV(mm,method='LSCV')      #Validación cruzada por mínimos cuadrados

#Estimaciones de la función de densidad
plot(mm,ylim=c(-1.2, 1.2))
curve.circular(dcircmix(x,model=11),add=TRUE,ylim=c(-1.5, 1.5), lwd=3,col='black')
lines(density.circular(x=mm,bw=ps1,kernel='vonmises'),lwd=3,col='orange',add=TRUE)
lines(density.circular(x=mm,bw=ps2,kernel='vonmises'),lwd=3,col='blue',add=TRUE)
lines(density.circular(x=mm,bw=ps3,kernel='vonmises'),lwd=3,col='red',add=TRUE,lty='aa')
lines(density.circular(x=mm,bw=ps4,kernel='vonmises'),lwd=3,col='springgreen',add=TRUE,lty=3)
legend(x = "topright",legend = c('F. densidad','Regla pulgar','plug-in','LCV',
    'LSCV'),lty = c(1, 1, 1, 5,
    3),col=c('black','orange','blue','red','springgreen'),lwd = 2)
```

- Estimaciones de la función de densidad de la distribución normal enrollada asimétrica obtenidas a partir de los parámetros de suavizado que resultan de los distintos métodos estudiados utilizando distintos tamaños muestrales representadas en la figura 5.6.

```
library(circular)
library(NPCirc)

mu<-circular(pi)
kappa<-1
lambda<-20

#Muestra normal enrollada asimétrica de tamaño 50
set.seed(1)
ms<-rwsn(50,mu,kappa,lambda)

#Parámetros de suavizado
ps1<-bw.rt(ms)                  #Regla del pulgar
ps2<-bw.pi(ms)                  #Regla plug-in
ps3<-bw.CV(ms,method='LCV')     #Validación cruzada por verosimilitud
ps4<-bw.CV(ms,method='LSCV')    #Validación cruzada por mínimos cuadrados
plot(ms,ylim=c(-1.2, 1.2))

#Estimaciones de la función de densidad
curve.circular(dwsn(x,mu,kappa,lambda),add=TRUE,ylim=c(-1.3, 1.3), lwd=3,col='black')
lines(density.circular(x=ms,bw=ps1,kernel='vonmises'),lwd=3,col='orange',add=TRUE)
lines(density.circular(x=ms,bw=ps2,kernel='vonmises'),lwd=3,col='blue',add=TRUE)
lines(density.circular(x=ms,bw=ps3,kernel='vonmises'),lwd=3,col='red', add=TRUE)
lines(density.circular(x=ms,bw=ps4,kernel='vonmises'),lwd=3,col='springgreen', add=TRUE,
    lty='aa')
```

```
legend(x = "topleft",legend = c('F. densidad','Regla pulgar','plug-in','LCV', 'LSCV'),
    lty = c(1, 1, 1, 1, 5), col=c('black','orange','blue','red','springgreen'), lwd = 2)


#Muestra normal enrollada asimétrica de tamaño 200
set.seed(1)
ms<-rwsn(200,mu,kappa,lambda)

#Parámetros de suavizado
ps1<-bw.rt(ms)                  #Regla del pulgar
ps2<-bw.pi(ms)                  #Regla plug-in
ps3<-bw.CV(ms,method='LCV')     #Validación cruzada por verosimilitud
ps4<-bw.CV(ms,method='LSCV')    #Validación cruzada por mínimos cuadrados

#Estimaciones de la función de densidad
plot(ms,ylim=c(-1.2, 1.2))
curve.circular(dwsn(x,mu,kappa,lambda),add=TRUE,ylim=c(-1.3, 1.3), lwd=3,col='black')
lines(density.circular(x=ms,bw=ps1,kernel='vonmises'),lwd=3,col='orange',add=TRUE)
lines(density.circular(x=ms,bw=ps2,kernel='vonmises'),lwd=3,col='blue',add=TRUE)
lines(density.circular(x=ms,bw=ps3,kernel='vonmises'),lwd=3,col='red',add=TRUE)
lines(density.circular(x=ms,bw=ps4,kernel='vonmises'),lwd=3,col='springgreen', add=TRUE,
    lty='aa')
legend(x = "topleft",legend = c('F. densidad','Regla pulgar','plug-in','LCV', 'LSCV'),
    lty = c(1, 1, 1, 1, 5), col=c('black','orange','blue','red','springgreen'), lwd = 2)
```

A.4. Aplicación a datos reales

- Diagramas de rosa y de cajas de la muestra de la dirección de salida de las partículas resultantes de la colisión en el LHC representados en la figura 6.3. Cálculo de los parámetros de localización y concentración de estos datos.

```
library(bpDir)
library(circular)

#Cargamos los datos
datos<-read.csv('Particulas.csv',sep=';')
datos2<-as.numeric(datos$l0.f)
datosc<-circular(datos2+pi, units = 'radians', zero=0, rotation='counter')

#Diagrama de rosa
plot(datosc,alpha=0.5)
rose.diag(datosc,bins=11,add=T)
```

```
#Diagrama de cajas
CircularBoxplot(datosc,template = 'radians')

#Parámetros muestrales
n<-length(datosc)                  #Tamaño muestral
S<-1/n*sum(sin(datosc))
C<-1/n*sum(cos(datosc))
Media<-(atan(S/C)+pi)              #Dirección media muestral
R<-sqrt(S^2+C^2)                   #Parámetro de concentración muestral
Mediana<-median.circular(datosc)   #Mediana muestral
```

- Estimaciones de la función de densidad de a muestra de la dirección de salida de las partículas resultantes de la colisión en el LHC obtenidas a partir de los parámetros de suavizado que resultan de los distintos métodos estudiados representadas en la figura 6.4. Aplicación del test de Kuiper a estos datos.

```
library(circular)
library(NPCirc)

#Cargamos los datos
datos<-read.csv('Particulas.csv',sep=';')
datos2<-as.numeric(datos$l0.f)
datosc<-circular(datos2+pi, units = 'radians', zero=0, rotation='counter')

#Parámetros de suavizado
ps1<-bw.rt(datosc)                  #Regla del pulgar
ps2<-bw.pi(datosc)                  #Regla plug-in
ps3<-bw.CV(datosc,method='LCV')     #Validación cruzada por verosimilitud
ps4<-bw.CV(datosc,method='LSCV')    #Validación cruzada por mínimos cuadrados

#Estimación de la función de densidad
plot(datosc,ylim=c(-1.3, 1.3))
curve.circular(dcircularuniform, join=TRUE, ylim=c(-1.05, 1.05), lwd=10,col='black',
    add=TRUE)
lines(density.circular(x=datosc,bw=ps1,kernel='vonmises'),lwd=2,col='orange',add=TRUE)
lines(density.circular(x=datosc,bw=ps2,kernel='vonmises'),lwd=2,col='blue',add=TRUE)
lines(density.circular(x=datosc,bw=ps3,kernel='vonmises'),lwd=2,col='red',add=TRUE)
lines(density.circular(x=datosc,bw=ps4,kernel='vonmises'),lwd=2,col='springgreen',
    add=TRUE)
legend(x = "topright",legend = c('f uniforme', 'Regla pulgar','plug-in','LCV','LSCV'),
    lty = c(1, 1, 1, 1, 1), col=c('black','orange','blue','red','springgreen'), lwd = 2)

```

```
#Test de bondad de ajuste a una uniforme (test de Kuiper)
kuiper.test(datosc,alpha=0.05)
```

- Diagramas de rosa y de cajas de la muestra de la dirección de las corrientes marinas en el cabo de Peñas representados en la figura 6.6. Cálculo de los parámetros de localización y concentración de estos datos.

```
library(bpDir)
library(circular)

#Cargamos los datos
datos<-read.delim('Puertos.csv')
datos2<-as.numeric(datos$X.1[8283:8593])
datosc<-circular(datos2, units = 'degrees', zero=pi/2, rotation='clock')

#Diagrama de rosa
plot(datosc)
points.circular(mean.circular(datosc),col='red',cex=2,add=T)
rose.diag(datosc,bins=15,add=T)
#Diagrama de cajas
CircularBoxplot(datosc,template = 'geographics')

#Parámetros muestrales
n<-length(datosc)                  #Tamaño muestral
S<-1/n*sum(sin(datosc*pi/180))
C<-1/n*sum(cos(datosc*pi/180))
Media<-(atan(S/C)+2*pi)*180/pi     #Dirección media muestral
R<-sqrt(S^2+C^2)                   #Parámetro de concentración muestral
Mediana<-median.circular(datosc)   #Mediana muestral
```

- Estimaciones de la función de densidad de a muestra de la dirección las corrientes marinas en el Cabo de Peñas obtenidas a partir de los parámetros de suavizado que resultan de los distintos métodos estudiados representadas en la figura 6.7. Aplicación del test de Kuiper a estos datos.

```
library(circular)
library(NPCirc)

#Cargamos los datos
datos<-read.delim('Puertos.csv')
datos2<-as.numeric(datos$X.1[8283:8593])
datosc<-circular(datos2, units = 'degrees', zero=pi/2, rotation='clock')
```

```
8
9  #Parámetros de suavizado
10 ps1<-bw.rt(datosc)                    #Regla del pulgar
11 ps2<-bw.pi(datosc)                    #Regla plug-in
12 ps3<-bw.CV(datosc,method='LCV')       #Validación cruzada por verosimilitud
13 ps4<-bw.CV(datosc,method='LSCV')      #Validación cruzada por mínimos cuadrados
14
15 #Estimación de la función de densidad
16 plot(datosc,ylim=c(-1.3, 1.3))
17 lines(density.circular(x=datosc,bw=ps1,kernel='vonmises'),lwd=3,col='orange',add=TRUE)
18 lines(density.circular(x=datosc,bw=ps2,kernel='vonmises'),lwd=3,col='blue',add=TRUE)
19 lines(density.circular(x=datosc,bw=ps3,kernel='vonmises'),lwd=3,col='red',add=TRUE, lty=5)
20 lines(density.circular(x=datosc,bw=ps4,kernel='vonmises'),lwd=3,col='springgreen',
       add=TRUE, lty=3)
21 legend(x = "topright",legend = c('Regla pulgar','plug-in','LCV', 'LSCV'), lty = c(1, 1,
       5, 3), col=c('orange','blue','red','springgreen'), lwd = 2)
22
23 #Test de bondad de ajuste a una uniforme (test de Kuiper)
24 kuiper.test(datosc,alpha=0.05)
```

Bibliografía

Agostinelli, C. y Lund, U. (2023). *R package* `circular`*: Circular Statistics (version 0.5-0)*.

Batschelet, E. (1981). *Circular Statistics in Biology*. Mathematics in biology. Academic Press.

Buttarazzi, D. (2021). *bpDir: Boxplots for Directional Data*. R package version 0.1.2.

Buttarazzi, D., Pandolfo, G., y Giovanni, C. P. (2018). A Boxplot for Circular Data. *Journal of Statistical Software*, 74(4):1–10.

Di Marcio, M., Panzera, A., y Taylor, C. C. (2009). Local polynomial regression for circular predictors. *Statistics Probability Letters*, (74):2066–2075.

Di Marzio, M., Panzera, A., y Taylor, C. C. (2011). Kernel density estimation on the torus. *Journal of Statistical Planning and Inference*, (141):2156–2173.

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

Gorgas, J., Cardiel, N., y Zamorano, J. (2011). *Estadística básica para estudiantes de ciencias*. Universidad Complutense de Madrid.

Hall, P., Watson, G. S., y Cabrera, J. (1987). Kernel Density Estimation with Spherical Data. 74(4):751–762.

Jammalamadaka, S. R. y SenGupta, A. (2001). *Topics in Circular Statistics*. World Scientific.

Kim, S., SenGupta, A., y Arnold, B. C. (2016). A multivariate circular distribution with applications to the protein structure prediction problem. *Journal of Multivariate Analysis*, 143:374–382.

Mardia, K. y Jupp, P. (2000). *Directional Statistics*. Academic Press Inc. (London).

Mardia, K. V. (1972). *Statistics of Directional Data*. London: Academic Press.

Marron, J. S. y Wand, M. P. (1992). Exact Mean Integrated Squared Error. *The Annals of Statistics*, 20(2):712 – 736.

Mendenhall, W., Beaver, R., y Beaver, B. (2010). *Introducción a la probabilidad y estadística*. Cengage Learning.

Oliveira, M. (2014). *Nonparametric Circular Methods for Density and Regression*. Tesis doctoral, Universidade de Santiago de Compostela.

Oliveira, M., Crujeiras, R. M., y Rodríguez-Casal, A. (2014). NPCirc: An R Package for Nonparametric Circular Methods. *Journal of Statistical Software*, 61(9):1–26.

Oliveira, M., Crujeiras, R. M., y Rodríguez-Casal, A. (2013). Nonparametric circular methods for exploring environmental data. *Environmental and Ecological Statistics*, (10):1–17.

Parzen, E. (1962). On Estimation of a Probability Density Function and Mode. *The Annals of Mathematical Statistics*, 33(3):1065 – 1076.

Pewsey, A., N. M. y Ruxton, G. D. (2013). *Circular statistics in R*. Oxford University Press.

R Core Team (2022). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Ruiz-Villarreal, M., Coelho, H., Río, G., y Nogueira, J. (2004). Slope current in the Cantabrian: Observations and modeling of seasonal variability and interaction with Aviles Canyon.

Scott, D. (2015). *Multivariate Density Estimation: Theory, Practice, and Visualization*. Hoboken: John Wiley Sons.

Silverman, B. W. (1986). *Density estimation for statistics and data analysis*. Chapman and Hall.

Stephens, M. (1970). Use of the Kolmogorov-Smirnov, Cramer-Von Mises and related statistics without extensive tables. *Journal of the Royal Statistical Society*, (32):115–122.

Taylor, C. C. (2008). Automatic Bandwidth Selection for Circular Density Estimation. *Computational Statistics and Data Analysis*, (52):3493–3500.

Tukey, J. W. (1977). *Exploratory data analysis*. Pearson.

Yu, Z., Dong, Y., y Huang, M. (2014). General directional regression. *Journal of Multivariate Analysis*, 124:94–104.