Universidad Nacional Autónoma de México

Facultad de Ciencias

DETECCIÓN DE OBSERVACIONES DISCORDANTES PARA DATOS CIRCULARES

T E S I S

QUE PARA OBTENER EL TÍTULO DE:
ACTUARIO

PRESENTA:
CECILIA ISABEL CARMONA SIMON

DIRECTOR DE TESIS:
MAT. MARGARITA ELVIRA CHÁVEZ CANO

2009

Hoja de Datos del Jurado

1. Datos del alumno
Carmona
Simon
Cecilia Isabel
56182427
Universidad Nacional Autónoma de México
Facultad de Ciencias
Actuaría
401062234

2. Datos del tutor
Mat.
Chávez
Cano
Margarita Elvira

3. Datos del sinodal 1
Dra.
Fuentes
García
Ruth Selene

4. Datos del sinodal 2
Act.
Vázquez
Alamilla
Jaime

5. Datos del sinodal 3
M. en P.
Alonso
Reyes
María del Pilar

6. Datos del sinodal 4
Act.
Sánchez
Villareal
Francisco

7. Datos del trabajo escrito.
Detección de observaciones discordantes para datos circulares
83 p.
2009

Como un reconocimiento a la mujer que me dió la vida y
me enseñó a confiar en mis propios pasos:

Carmen Simon,

Por todo el apoyo y amor que me has brindado a lo largo
de este tiempo, tu fuerza y entrega son mi ejemplo a seguir,
como agradecimiento y dedicatoria a ti, mi querida madre.

No hay palabras suficientes que pueden expresar lo agradecida que estoy con la vida, mi familia, mis amigos, profesores, la UNAM, por todos estos maravillosos años de formación y crecimiento: por compartir, enseñar, ayudar, escuchar, ser, estar...

GRACIAS

Con especial afecto y gratitud a ustedes a quien amo y valoro por formar parte de mi vida:

Arthur:

> Por los bellos momentos que hemos atesorado y los que están por venir, agradezco tu presencia y el apoyo que me brindas, tomas un lugar muy importante en mi corazón y sin duda alguna tenemos la oportunidad de hacer más fuerte este hermoso vínculo de padre e hija.

David y Denisse:

> Por todos estos años de convivencia, aprecio cada momento que vivimos juntos y el saber que siempre contaré con ustedes, lejos o cerca este lazo de hermandad siempre estará presente.

Germán:

> Por la apoyo que has brindado a mi familia y cuando siempre lo necesité, gracias por tu amistad fiel.

Bryan:

> Por la ternura, cariño y lealtad de tu corazón que han llenado mi vida de amor y nuevas esperanzas, sin duda, nuestra unión es tan fuerte que venceremos cualquier obstáculo para juntos seguir creciendo y formar lo que siempre hemos soñado: una bella familia.

Dios:

> Por el aliento de vida, la fé, el amor que día a día depositas en mi corazón para seguir adelante y lograr siempre la Victoria.

Por la guía y apoyo para la realización de esta tesis, un sueño realizado, agradezco a:

Mat. Margarita Chávez:

Por brindarme la oportunidad de trabajar con usted, por su confianza y guía, es un claro ejemplo para muchos estudiantes, gracias por su dedicación y el cariño que demuestra en lo que hace.

Ruy Manrique:

Por el tiempo dedicado para escuchar y compartir en todo momento el avance y desarrollo de este trabajo, tu ayuda fue crucial y de gran significancia, pero sobre todo agradezco tu amistad sincera y la nobleza de tu corazón.

Ricardo Ríos:

Por compartir tus conocimientos en Latex y tu disponibilidad por enseñar y ayudar en el formato de mi tesis.

La amistad, una de las cosas más apreciadas de la vida, que otorga al ser humano la capacidad de dar y amar, con un agradecimiento sincero a todos mis amigos que me acompañaron no solo en mi formación académica sino en una de las etapas más bellas de mi juventud:

Alma, Liliana, Maricarmen, Sarah.

Fanny, Chayo, Paulina, Miguel, Alejandro, Poncho, Iván, Yadira, Chores, Cristian y todos mis compañeros de carrera.

Toño, Héctor, Compa, Chucho, Checo, Abraham, Hugo, Miguelón, Paula, Lety, Salvador, Esteban, Adrián, Nico, Alí, todos los Miaus, compañeros de clase y de facultad.

Índice general

Preliminares

La estadística direccional se desarrrolló con el fin de analizar datos direccionales[1] de forma adecuada, pues éstas poseen características propias y especiales que necesitan ser analizados de forma distinta a la estadística tradicional. Hoy en día ha tomado el interés de los investigadores al darse cuenta de su importante aplicación a diversas áreas científicas.

Este trabajo de tesis se centra en el estudio de datos direccionales de dos dimensiones donde los datos son representados como puntos sobre la circunferencia de un círculo unitario centrado en el origen o como vectores unitarios que conectan a estos puntos con el origen. Entonces, de forma análoga a la estadística tradicional, en donde se definen funciones de distribución en la línea recta, también en el contexto de datos direccionales se definen distribuciones en el círculo . Para fines de este estudio, se considera la distribución Von Mises que presenta gran semejanza a la distribución normal pero en el contexto direccional.

El objetivo de este estudio es la detección de valores discordantes en una muestra de datos univariados cuyo modelo de probabilidad es la distribución Von Mises. Así pues, se analiza el concepto de valores atípicos o inconsistentes: outliers, en un conjunto de datos y se examina como su presencia puede ser detectada a través de pruebas de discordancia. Cabe destacar que esta investigación se retoma de la ya realizada por D. Collet y cuyo artículo se titula:"Outliers in Circular Data".[2]

La tesis se divide en cuatro capítulos: En el primer capítulo se analiza el concepto de datos direccionales, mostrando las estadísticas descriptivas que permiten conocer su comportamiento. En el segundo capítulo se muestran los conceptos básicos de las distribuciones circulares y específicamente las características propias de la distribución Von Mises. En el tercer capítulo se expone el concepto de valores atípicos conocidos como "outliers", su naturaleza y la variedad de contextos en los que pueden aparecer y para los cuales se pueden tomar diferentes tipos de acciones como respuesta a la presencia de éstos. Asimismo se explica el concepto y las bases estadísticas para la construcción de herramientas que permiten evaluar si valores inconsistentes en muestras univariadas son o no discordantes, y por esta razón se les conoce como *pruebas de discordancia.*

[1] Medidas angulares.
[2] Collet,1979 [3].

Una vez analizado el concepto de valores inconsistentes en un conjunto de datos lineales y univariados: su detección, tratamiento y diferentes formas de plantear pruebas de discordancia, se introduce el tema para datos circulares, pues es diferente al caso lineal. Es por esto que en el cuarto capítulo se explica la forma de identificar una observación atípica en una muestra circular univariada en términos de la "distancia circular ". Además se definen pruebas de discordancia para detectar si una observación proveniente de una muestra con función de probabilidad Von Mises es discordante o no. Las estadísticas de prueba que Collet define y que se exponen son: Prueba L, Prueba C, Prueba D y Prueba M, las cuales analiza al estudiar sus distribuciones asintóticas y puntos percentiles por medio del Método Monte Carlo a un nivel de sinificancia de $\alpha = .01$ y $\alpha = .05$. Con el fin de entender dicho procedimiento se realizó una simulación en el programa R que genera las distribuciones muestrales empíricas de las pruebas y el valor de sus cuantiles para cualquier parámetro deseado.

Después se explica cómo evaluar el funcionamiento de las pruebas estadísticas. Se menciona que Collet por medio de una simulación genera la función potencia y otras probabilidades de interés, presenta sus conclusiones de cuál prueba es mejor, bajo que parámetros y circunstancias. Asimismo para entender el procedimiento, se realizó una simulación en R para evaluar la potencia de las pruebas en la detección de valores discordantes con el objetivo de dar conclusiones propias y compararlas con las de Collet.

Posteriormente, se muestra un ejemplo propio de la Biología, en donde se estudia una muestra de direcciones que siguen 22 estrellas de mar al ser alejadas de su habitat natural para detectar posibles observaciones atípicas a través de técnicas exploratorias simples: gráfica de dispersión y gráfica P-P, y posteriormente analizar si dicha observación atípica es o no discordante con base en las estadísticas de prueba ya mencionadas, utilizando los cuantiles estimados por Collet y los generados con la simulación propuesta.

Finalmente se presentan las conclusiones y los apéndices con la información básica que se requiere para el estudio de datos direccionales y observaciones discordantes. También se muestran los códigos de las simulaciones hechas en R.

Capítulo 1

Datos direccionales

1.1. Introducción

Al conjunto de observaciones cuyo valor se expresa en ángulos o radianes se le conoce como *datos direccionales*. En el caso de dos dimensiones y dado que las direcciones no tienen magnitud, los datos en su forma más simple son representados como puntos sobre la circunferencia de un círculo unitario centrado en el origen. Debido a su representación en un círculo son llamados "datos circulares". Figura 1.1

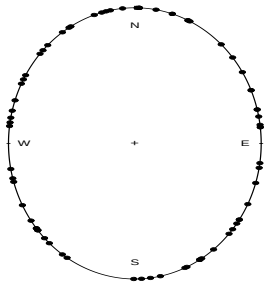


Figura 1.1: Representación de datos direccionales sobre la circunferencia de un círculo

La representación numérica de las direcciones no necesariamente es única ya que el valor del ángulo depende del valor del *sentido de rotación*[1] y de la *dirección cero*[2]. Por ejemplo, un matemático al medir 60^o considera como dirección cero el Este y como rotación positiva el sentido opuesto a las manecillas del reloj. Sin embargo un Geólogo que toma como dirección cero el Norte y como rotación positiva el sentido de las manecillas del reloj, la misma dirección angular es de 30^o; *véase figura 1.2*. Luego entonces, no es posible establecer un orden natural entre observaciones de dos o más muestras ya que éstas dependen de la elección de la dirección cero y del sentido de rotación. Además, cabe destacar que las medidas angulares son cíclicas por el hecho de que $0 = 2\pi$ y se dice que θ es periódica por el hecho de que $\theta = \theta + 2\pi * k$, donde el valor de θ se expresa en radianes y k es cualquier entero.

[1] Mismo sentido a las manecillas del reloj o contrario.
[2] Punto de referencia inicial.

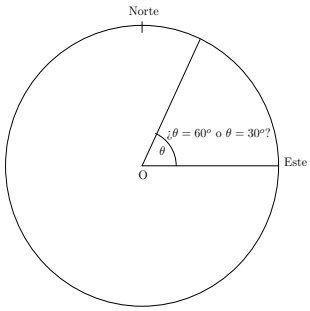


Figura 1.2: Importancia del sentido de rotación y dirección cero

Es necesario señalar que es posible convertir una variable lineal que mide un evento cíclico en una variable circular. Un claro ejemplo sería una variable cuya unidad es el tiempo, se sabe que los datos circulares tienen un rango de 0^o a 360^o, o bien, de 0 a 2π radianes, entonces un período de 24 horas sería equivalente a un ángulo de 360^o, medio día a un ángulo de 180^o y una hora a un ángulo de 15^o. De la misma forma se podría representar un mes, un año y cualquier otro período en la circunferencia de un círculo. Por ejemplo en un círculo se podrían representar los 365 días del año para graficar la frecuencia de accidentes áereos y ver si éstos se distribuyen uniformemente en las diferentes estaciones del año.

1.2. Aplicaciones

En diversos campos científicos se utilizan medidas direccionales, es por eso que la importancia y desarrollo de la estadística circular ha ido en aumento. A continuación se mencionan algunas de las la aplicaciones que existen en las diversas áreas:

- *Biología:* En estudios sobre la orientación y movimientos direccionales que siguen ciertas especies de animales. Como por ejemplo: al migrar las mariposas se estudian la direcciones que éstas siguen en su trayectoría, para determinar si se guían o no por señales tales como la dirección del sol o el campo magnético de la Tierra.
- *Física:* Gracias al estudio de las desviaciones de los pesos atómicos hecha por Von Mises en 1918 se dió a conocer una de las distribuciones básicas de la estadística circular.
- *Psicología:* En el estudio de mapas mentales que la gente utiliza para representar su entorno.
- *Medicina:* En el estudio de muertes ocasionadas por enfermedades en un período de tiempo.
- *Geología:* En el estudio del cambio magnético de los polos, al analizar la direcciones entre las corrientes de los ríos y las direcciones magnéticas de los polos de la tierra en siglos pasados.
- *Ecología:* En el estudio de la contaminación ambiental se considera como factor importante las direcciones que sigue el viento.

1.3. Notación

Para la representación de datos circulares se utiliza el sistema de coordenadas rectangulares o el sistema de coordenadas polares. En el primero, se tienen dos ejes perpendiculares: X y Y, con origen en O, entonces, cualquier punto en el plano P puede ser representado como (x, y) o en términos de coordenadas polares como (r, θ), donde r es la distancia al origen y θ es la dirección.

Es fácil transformar coordenadas polares en coordenadas rectangulares y viceversa, haciendo uso de las funciones trigonométricas *seno* y *coseno*. Figura 1.3. Por ejemplo, si se toma al punto $P = (r, \theta)$ entonces las coordenadas rectangulares están dadas por:

$$x = rcos\theta \qquad y = rsen\theta$$

Debido a que en el análisis direccional es de interés la dirección y no la magnitud de los vectores, entonces $r = 1$. Así pues, a cada dirección le corresponde un punto P sobre la circunferencia del círculo unitario y la conversión de coordenadas polares a rectangulares para este punto P es:

$$(1, \theta) \Longleftrightarrow (x = cos\theta, y = sen\theta)$$

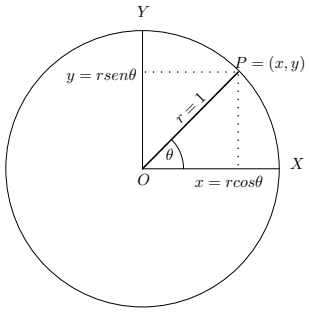


Figura 1.3: Relación entre coordenadas rectangulares y polares

1.4. Estadística descriptiva

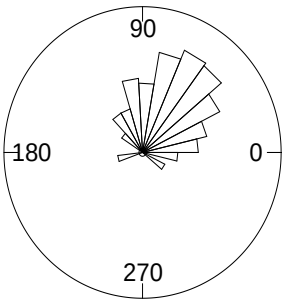
Para analizar un conjunto de datos es importante representarlos gráficamente y resumir de forma numérica sus aspectos más relevantes a través del uso de la estadística descriptiva. De forma intuitiva, se pensaría en cortar el círculo en un punto apropiado y calcular las medidas necesarias para un adecuado resumen numérico en la línea recta. Sin embargo, estas medidas dependen fuertemente de la elección del punto en que se corta el círculo. Para entender este problema, suponga que se tiene una muestra de dos direcciones (con dirección cero en el Este y el sentido de rotación contrario a las manecillas del reloj) $\theta_1 = 1^o$ y $\theta_2 = 359^o$. Si se corta el círculo en 0^o, la media muestral es 180^o y la desviación muestral estándar [3] de 179^o. Por otro lado, si cortará el círculo en 180^o la media muestral sería de 0^o y la desviación estándar de 1^o.

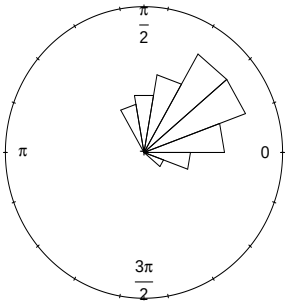
[3] Versión sesgada.

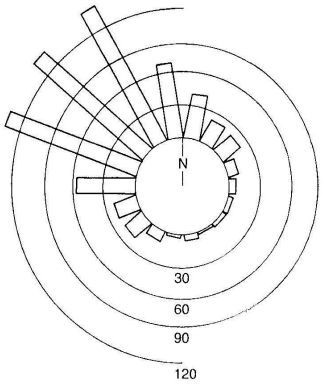
De esta forma, es evidente que se requiere una forma adecuada para el cálculo de medidas descriptivas. Además que la necesidad de utilizar métodos y medidas estadísticas invariantes a la elección de la dirección cero y sentido de rotación, ocasiona que muchas de las técnicas y medidas de la estadística tradicional sean confusas y carezcan de sentido. Gran parte de las medidas descriptivas, tales como la media y la varianza, se vuelven inapropiadas y necesitan ser redefinidas para datos direccionales; de igual forma herramientas analíticas como la generadora de momentos, el coeficiente de correlación, modelos de inferencia, de regresión, etc. necesitan ser redefinidas para datos direccionales.

1.4.1. Representación gráfica

Una forma de representar gráficamente a un conjunto de datos direccionales en una circunferencia, es por medio de los *histogramas circulares*, donde es necesario ordenar y clasificar las observaciones en grupos. Se sigue la misma analogía de un histograma lineal, es decir, la frecuencia para cada grupo de datos se representa por medio de barras rectangulares cuya área es proporcional a la frecuencia de cada grupo. Una variante de estos histogramas son los *Diagramas de rosa*, en los cuales, las barras son reemplazadas por sectores.







1.4.2. Medidas de tendencia central

Dirección media

Para definir la dirección media de forma correcta para datos direccionales, se debe considerar a los datos como vectores unitarios y calcular la dirección del vector resultante.

Sea una muestra de direcciones angulares: $\theta_1, \theta_2, ...\theta_n$ cuyos vectores unitarios son $e_1, e_2, ...e_n$, tal que $|e_i| = 1$, $\forall_{i=1...n}$, ver figura 1.4.

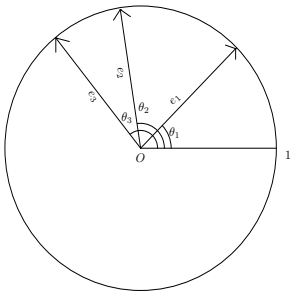


Figura 1.4: Representación de los vectores unitarios

Para obtener la dirección media de los vectores unitarios se tiene que calcular la dirección del vector resultante, la cual se obtiene al sumar los vectores unitarios:

$$R = \sum_{i=1}^{n} e_i$$

Si se considera la transformación de las componentes polares a rectangulares de e_i se tiene que sus coordenadas son:

$$(cos\theta_i, sen\theta_i), i = 1, ...n$$

entonces,

$$R = \left(\sum_{i=1}^{n} cos\theta_i, \sum_{i=1}^{n} sen\theta_i \right) = (C, S)$$

donde:

$$C = \sum_{i=1}^{n} cos\theta_i \qquad y \qquad S = \sum_{i=1}^{n} sen\theta_i$$

siendo el tamaño del vector resultante:

$$\mathcal{R} = \|R\| = \sqrt{C^2 + S^2}$$

Por tanto, el vector medio de la muestra, indica el centro de masa y está dada por:

$$\bar{R} = \frac{R}{n}$$

siendo r la longitud del vector medio:

$$r = \|\bar{R}\| = \sqrt{\bar{C}^2 + \bar{S}^2}$$

Y así el centro de masa de los vectores unitarios o centro de gravedad tiene componentes cartesianas en: $(\bar{C}, \bar{S})$ y se puede encontrar dentro del círculo unitario. No obstante si los ángulos son los mismos: $\theta_i = \theta_j$, $i \neq j$, el centro de masa estará en la circunferencia del círculo.

La dirección del vector medio tiene un ángulo bien definido llamado el *ángulo medio de la muestra:* $\bar{\theta}$, y se obtiene al resolver las ecuaciones:

$$\bar{C} = rcos\bar{\theta}, \quad \bar{S} = rsen\bar{\theta}$$

Entonces $\bar{\theta}$ está dado por[4]:

$$\theta = \begin{cases} \arctan(\bar{S}\backslash\bar{C}), & si \quad \bar{C} > 0, \quad \bar{S} \geq 0 \\ \pi\backslash 2, & si \quad \bar{C} = 0, \quad \bar{S} > 0 \\ \arctan(\bar{S}\backslash\bar{C}) + \pi, & si \quad \bar{C} < 0, \\ \arctan(\bar{S}\backslash\bar{C}) + 2\pi, & si \quad \bar{C} \geq 0, \quad \bar{S} < 0 \\ indefinido, & si \quad C = 0 \quad S = 0 \end{cases}$$

Teorema

Sea $\bar{\theta}$ la dirección del vector medio, entonces se cumple que:

$$\sum_{i=1}^{n} sen(\theta_i - \bar{\theta}) = 0 \tag{1.1}$$

$$\frac{1}{n}\sum_{i=1}^{n} cos(\theta_i - \bar{\theta}) = r \tag{1.2}$$

Prueba:

$$\begin{aligned} \tfrac{1}{n}\textstyle\sum_{i=1}^{n} sen(\theta_i - \bar{\theta}) &= \tfrac{1}{n}\textstyle\sum_{i=1}^{n}(sen\theta_i cos\bar{\theta} - cos\theta_i sen\bar{\theta}) \\ &= \tfrac{1}{n}\textstyle\sum_{i=1}^{n} sen\theta_i cos\bar{\theta} - \tfrac{1}{n}\textstyle\sum_{i=1}^{n} cos\theta_i sen\bar{\theta} \\ &= \bar{S}cos\bar{\theta} - \bar{C}sen\theta_i \\ &= rsen\bar{\theta}cos\bar{\theta} - rcos\bar{\theta}sen\bar{\theta} \\ &= 0 \end{aligned}$$

Como $n \neq 0$, se sigue que:

$$\sum_{i=1}^{n} sen(\theta_i - \bar{\theta}) = 0$$

[4]Recuérdese que la función inversa $arctan(tan^{-1})$ toma valores entre $[-\pi\backslash 2, \pi\backslash 2]$.

De forma similar:

$$\begin{aligned}
\tfrac{1}{n}\sum_{i=1}^{n} cos(\theta_i - \bar{\theta}) &= \tfrac{1}{n}\sum_{i=1}^{n}(cos\theta_i cos\bar{\theta} + sen\theta_i sen\bar{\theta}) \\
&= \tfrac{1}{n}\sum_{i=1}^{n} cos\theta_i cos\bar{\theta} + \tfrac{1}{n}\sum_{i=1}^{n} sen\theta_i sen\bar{\theta} \\
&= \bar{C}cos\bar{\theta} + \bar{S}sen\theta_i \\
&= rcos\bar{\theta}cos\bar{\theta} + rsen\bar{\theta}sen\bar{\theta} \\
&= r(cos^2\bar{\theta} + sen^2\bar{\theta}) \\
&= r
\end{aligned}$$

Por lo tanto:

$$\frac{1}{n}\sum_{i=1}^{n} cos(\theta_i - \bar{\theta}) = r$$

Estos resultados ponen en contexto algunas analogías con la estadística clásica. Dado un conjunto de datos $x_1, x_2, ..., x_n$ con una media muestral: $\bar{x}$, entonces, la ecuación 1.1 es análoga a la suma de las desviaciones alrededor de la media, la cual es cero:

$$\sum_{i=1}^{n}(x_i - \bar{x}) = 0$$

Ahora, analizando el efecto de rotación sobre la dirección del vector medio, se verá que éste es equivariante, es decir, si los ángulos se desplazan por una cierta cantidad, también la dirección del vector medio lo hace a esa misma cantidad.

Proposición:

Sean los ángulos $\theta_1, \theta_2, ... \theta_n$ cuyo vector medio es $\bar{\theta}$. Sea α el ángulo de desplazamiento, entonces, el nuevo conjunto de vectores $\theta_1 + \alpha,\ \theta_2 + \alpha, ... ,\theta_n + \alpha$ tiene una dirección media de $\bar{\theta} + \alpha$. Supóngase que $\bar{R}'$ es el vector medio del nuevo conjunto de observaciones, es decir, después del cambio. Así que se tiene:

$$\bar{R}' = \left(\frac{1}{n}\sum_{i=1}^{n} cos(\theta_i + \alpha), \frac{1}{n}\sum_{i=1}^{n} sen(\theta_i + \alpha)\right) = (\bar{C}', \bar{S}')$$

de donde:

$$\begin{aligned}
\bar{C}' &= \tfrac{1}{n}\sum_{i=1}^{n} cos(\theta_i + \alpha) \\
&= \tfrac{1}{n}\sum_{i=1}^{n}(cos\theta_i cos\alpha - sen\theta_i sen\alpha) \\
&= \bar{C}cos\alpha - \bar{S}sen\alpha \\
&= rcos\bar{\theta}cos\alpha - rsen\bar{\theta}sen\alpha \\
&= rcos(\bar{\theta} + \alpha)
\end{aligned}$$

Similarmente:

$$\bar{S}' = rsen(\bar{\theta} + \alpha)$$

Entonces:

$$r' = \|\bar{R}'\| = \sqrt{\bar{C}'^2 + \bar{S}'^2}$$

Por tanto, se concluye que:

$$\bar{C}' = r'cos(\bar{\theta} + \alpha) \qquad \bar{S}' = r'sen(\bar{\theta} + \alpha)$$

Y en consecuencia, la dirección del vector medio no depende de la dirección cero.

Mediana

Dada una muestra de datos direccionales sobre el círculo unitario, se puede obtener una versión de la mediana muestral al dividir dicha muestra entre un diámetro tal que la mitad de los datos estén a un lado de éste y la otra mitad del otro lado. Por consiguiente, la mediana se define como el ángulo ϕ tal que la mitad de los datos cae en el arco $[\phi, \phi + \pi)$. Cuando el tamaño de la muestra n es impar, la mediana es una de las observaciones del conjunto de datos. Cuando n es par la mediana muestral es el punto medio de dos datos muestrales apropiados.

1.4.3. Medidas de dispersión y distancia circular

Varianza circular

Al tomar $e_1, e_2, ...e_n$ como vectores unitarios, entonces el tamaño del vector medio r tiene rango entre 0 y 1, es decir: $0 \leq \ r \leq 1$. Si las direcciones $\theta_1, \theta_2, ...\theta_n$ tienden a agruparse, entoncer r tomará un valor cercano a 1, de lo contrario, si están demasiado dispersos r será cercano a 0. Por tanto, r es una medida de concentración del conjunto de datos. A partir de este hecho, se define la varianza circular muestral como:

$$V = 1 - r, \qquad 0 \leq V \leq 1$$

Cabe señalar que la varianza circular también puede escribirse en términos del tamaño del vector resultante $\mathcal{R}$, ya que la dirección del vector resultante es la misma que la del vector medio $\bar{\theta}$. Entonces se define a V' como una medida de dispersión, donde valores de $\mathcal{R}$ cercanos a 0 indican que la dispersión es grande, mientras que valores cercanos a n indican que el conjunto de observaciones tiene una dispersión pequeña o más concentrada hacia el centro.

$$V' = n - \mathcal{R}, \qquad 0 \leq V' \leq n$$

Por otro lado, a veces es útil conocer la desviación estándar análoga a la de la estadística clásica para un conjunto de datos. Así pues la ecuación 1.2

$$\frac{1}{n}\sum_{i=1}^{n} cos(\theta_i - \bar{\theta}) = r$$

se puede escribir de la siguiente forma:

$$\frac{1}{n}\sum_{i=1}^{n} 2[1 - cos(\theta_i - \bar{\theta})] = 2(1 - r)$$

Haciendo uso de resultados trigonométricos se sabe que para una desviación estándar pequeña se tiene que:

$$2[1 - cos(\theta_i - \bar{\theta})] = (\theta_i - \bar{\theta})^2$$

En consecuencia, se puede aproximar la siguiente fórmula :

$$\frac{1}{n}\sum_{i=1}^{n}(\theta_i - \bar{\theta})^2 \approx 2(1 - r)$$

De esta manera se observa que la estadística equivalente en la estadística tradicional es:

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = s^2.$$

Este resultado guía a definir la *varianza circular* equivalente en el modelo clásico como $2(1-r)$. Sin embargo, en ocasiones es preferible permanecer con la definición proveniente de la medida de concentración r en vez de querer encontrar el equivalente al modelo tradicional.

Distancia circular

Una forma razonable para medir la distancia entre dos puntos en la circunferencia, es el tomar la distancia más pequeña de los arcos que se forman entre esos dos puntos.

Sean θ_i, θ_j los ángulos correspondientes a los dos puntos, entonces se define la distancia circular como:

$$\delta_{ij} = \delta(\theta_i, \theta_j) = min(\theta_i - \theta_j, 2\pi - (\theta_i - \theta_j)) = \pi - |\pi - |\theta_i - \theta_j||$$

De forma clara, la distancia entre dos puntos no puede ser más grande que π, en consecuencia, la distancia circular toma valores entre $[0, \pi]$

Por ejemplo, dados dos puntos A y B sobre la circunferencia, la distancia entre ellos podría ser la longitud del arco AB_1 o la del arco AB_2; ver la figura 1.5. Siguiendo la definición de la distancia circular, se observa que el arco AB_1 es más pequeño que el arco AB_2. Por tanto la distancia circular entre A y B es la longitud del arco AB_1.

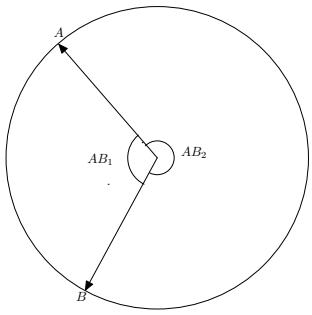


Figura 1.5: La distancia circular es la longitud del arco AB_1

Es importante examinar que δ_{ij} sea una *medida de disimilaridad* para poder utilizarla como instrumento de decisión.

Entonces, δ_{ij} es una medida de disimilaridad si cumple las siguientes propiedades:

- $\delta_{ij} \geq 0 \ \ \forall\, \theta_i, \theta_j$ (Positiva)
- $\delta_{ii} = 0 \ \ \forall\, \theta_i$ (Nulidad)
- $\delta_{ij} = \delta_{ji} \ \ \forall\, \theta_i, \theta_j$ (Simetría)
- $\delta_{ik} \leq max(\delta_{ij}, \delta_{jk}) \ \ \forall\, \theta_i, \theta_j, \theta_k$ (Ultramétrica)

Proposición:

La distancia circular definida como $\delta_{ij} = \pi - |\pi - |\theta_i - \theta_j||$ es una medida de disimilaridad.

Demostración:

Se sabe que el valor máximo de la diferencia $\theta_i - \theta_j \ \epsilon\, (-2\pi, 2\pi) \Rightarrow \pi - |\theta_i - \theta_j| \,\epsilon\, (-\pi, \pi) \Rightarrow \pi - |\pi - |\theta_i - \theta_j||$ tiene como rango $[0, \pi] \Rightarrow \delta_{ij} \geq 0 \ \ \forall\, i, j$. Luego entonces, la distancia δ_{ij} es positiva.

Para probar la nulidad, resulta obvio el hecho de que si $|\theta_i - \theta_i| = 0 \Rightarrow \delta_{ii} = 0 \ \ \forall_i$.

Por otro lado, como se cumple que $|\theta_i - \theta_j| = |\theta_j - \theta_i| \Rightarrow \delta_{ij} = \delta_{ji} \ \ \forall_{i,j}$. De esta forma, δ_{ij} es simétrica.

Por último, como $\delta_{ik} \,\epsilon\, [0, \pi] \ \ \forall_{i,k}$; entonces el $max(\delta_{ij}, \delta_{jk}) \,\epsilon\, [0, 2\pi] \ \ \forall_{i,j,k}$.
Por tanto $\delta_{ij} \leq max(\delta_{ij}, \delta_{jk})$ y se cumple la propiedad ultramétrica.

$\therefore\, \delta_{ij}$ es una medida de disimilaridad

Capítulo 2

Distribuciones circulares

2.1. Conceptos básicos

Las distribuciones univariadas en la estadística clásica se representan en la línea recta. Como ejemplo de éstas, podemos mencionar las distribuciones: Binomial, Poisson, Normal, etc. Estas distribuciones pueden tener un rango finito como la distribución binomial o un rango infinito, como la distribución Poisson y más aún tener un rango de $-\infty$ a $+\infty$ como la distribución normal univariada.

Una distribución circular es una distribución de probabilidad cuya probabilidad total se concentra en la circunferencia del círculo unitario y siempre tiene un rango finito de $[0, 2\pi)$. Dado que cada punto en la circunferencia representa una dirección, este tipo de distribuciones asignan probabilidades a diferentes direcciones o definen una distribución direccional.
Las distribuciones circulares siempre son continuas y por consiguiente existe una densidad de probabilidad, la cual es una función continua para los valores de θ con las siguientes propiedades:

1. $f(\theta) \geq 0$
2. $\int_0^{2\pi} f(\theta)\ \mathrm{d}\theta = 1$
3. $f(\theta) = f(\theta + (k * 2\pi))$, para cualquier entero k

2.1.1. Función característica

Como en la estadística clásica, una distribución de una variable aleatoria θ puede ser descrita a través de la funcion $t \rightarrow E[e^{it\theta}]$. A partir de que θ es una variable aleatoria periódica, entonces $\theta + 2\pi$ tiene la misma distribución y por tanto es necesario restringir el valor de t a valores enteros. Por tanto, para variables aleatorias circulares, la función característica necesita ser definida sólo para valores enteros:

Por definición se tiene que:

$$\varphi_\theta(t) = E[e^{it\theta}] = E[e^{it(\theta+2\pi)}] = e^{it2\pi} * \varphi_\theta(t)$$

Entonces, ya sea que $\varphi_\theta(t) = 0$ o $e^{it2\pi} = 1$, es decir, t debe ser un entero. La función característica en un entero p es llamado el *p-ésimo momento trigonométrico de* θ.

Si $f(\theta)$ es la función de densidad de probabilidad para una muestra aleatoria circular, entonces:

$$\varphi_\theta(p) = E(e^{ip\theta}) = \int_0^{2\pi} e^{ip\theta} f(\theta)\ \mathrm{d}\theta, \quad p = 0, \pm 1, \pm 2, ..$$

Hay que recordar por la relación de Euler que:

$$e^{i\theta} = cos\theta + isen\theta$$

Por tanto se puede escribir:

$$\varphi_\theta(p) = E(e^{ip\theta}) = \int_0^{2\pi} cosp\theta f(\theta)\ \mathrm{d}\theta + \int_0^{2\pi} isenp\theta f(\theta)\ \mathrm{d}\theta$$

claramente,

$$\varphi_0 = 1, \quad |\varphi_p| \leq 1$$

donde:

$$\alpha_p = E[cosp\theta] = \int_0^{2\pi} cosp\theta f(\theta)\ \mathrm{d}\theta \tag{2.1}$$

$$\beta_p = E[senp\theta] = \int_0^{2\pi} senp\theta f(\theta)\ \mathrm{d}\theta \tag{2.2}$$

Las ecuaciones 2.1 y 2.2 se relacionan con las ecuaciones:

$$\rho_p = \sqrt{\alpha_p^2 + \beta_p^2} \quad y \quad \mu_p = arctan\ \frac{\beta_p}{\alpha_p}$$

Entonces:

$$\varphi_p = \alpha_p + i\beta_p = \rho_p * e^{i\mu_p} \qquad p = 0, \pm 1, \pm 2, ..$$

Cabe notar que:

$$\alpha_p = \alpha_{-p}, \ \ \beta_p = \beta_{-p}, \ \ |\alpha_p| \leq 1, \ \ |\beta_p| \leq 1$$

2.1.2. Momentos trigonométricos

Se sabe que las distribuciones lineales pueden ser descritas por sus momentos. El primer momento es la *media* μ o la *esperanza* de una variable aleatoria X. Si la distribución tiene densidad de probabilidad entonces el e-nésimo momento es:

$$E[x^n] = \int_{-\infty}^{\infty} x^n f(x)\ \mathrm{d}x, \qquad n = 1, 2, ..$$

También, para el estudio de las distribuciones circulares es útil el cálculo de los momentos, sin embargo, debido a la periodicidad que tienen las distribuciones circulares, es necesario redefinir los momentos como *momentos trigonométricos*.

Sea $f(\theta)$ la función de densidad de probabilidad de una distribución circular, asi para cada vector unitario con componentes: $x = cos\theta$ $y = sen\theta$, se calcula la media con las siguientes igualdades:

$$\alpha_1 = \int_0^{2\pi} x f(\theta)\ \mathrm{d}\theta = \int_0^{2\pi} cos\theta f(\theta)\ \mathrm{d}\theta$$

$$\beta_1 = \int_0^{2\pi} y f(\theta)\ \mathrm{d}\theta = \int_0^{2\pi} sen\theta f(\theta)\ \mathrm{d}\theta$$

La media poblacional con componentes α_1, β_1 son los componentes del vector medio que apunta al centro de masa. Por tanto, al vector medio también se le conoce como *El primer momento trigonométrico* de la distribución circular y es aquél que contiene más información que cualquier otro momento ordinario:

$$\varphi_1 = (\alpha_1, \beta_1) = (\rho_1, \mu_1)$$

donde ρ es el tamaño del vector medio, μ es la dirección del vector medio.

De manera general para $p = 0, \pm1, \pm2, ..$ se tiene [1]:

$$\alpha_p = \int_0^{2\pi} cosp\theta f(\theta)\ \mathrm{d}\theta, \qquad \beta_p = \int_0^{2\pi} senp\theta f(\theta)\ \mathrm{d}\theta$$

de forma que, se denotará: $\rho_1 = \rho$ y $\mu_1 = \mu$, al considerar el primer momento trigonométrico como:

$$\phi_1 = \alpha_1 + i\beta_1 = \rho * e^{i\mu}$$

El primer momento trigonométrico: ρ y μ, proveen las medidas teóricas o poblacionales de la concentración y dirección media, respectivamente, de θ. Se puede ver que mientras más grande sea el tamaño de ρ, es decir, más cercano a 1, entonces mayor es la concentración hacia la dirección media μ.

[1] La sucesión $[(\alpha_p, \beta_p) : p = o, \pm1, \pm2..]$ de momentos trigonométricos de un vector aleatorio θ es equivalente a la función característica de θ.

2.2. Distribución Von Mises

La distribución Von Mises es de gran importancia en la teoría estadística para datos direccionales, fue introducida por Von Mises en 1918 al estudiar las desviaciones de los pesos atómicos y desde ese momento ha sido estudiada ampliamente y diversas técnicas inferenciales han sido desarrolladas, por lo que su uso es de gran utilidad para distintas áreas. Debido a su semejanza con la distribución normal para el análisis estadístico clásico, también se le conoce como la distribución normal circular.

Como se ha visto, la dirección del vector resultante muestral proporciona una dirección media razonable para la muestra. Von Mises se preguntó si había un modelo circular en el que $\bar{\theta}$ brindara un estimador máximo verosímil, es decir, una forma de caracterizar la distribución circular en el que la dirección media poblacional μ fuera estimada por la dirección del vector resultante con máxima probabilidad. Gauss demostró que la distribución normal se puede derivar de la función de verosimilitud, siguiendo la hipótesis de que la media es el valor más posible. En 1918 Richard Von Mises aplicó el método de Gauss para una variable circular y derivó la distribución Von Mises, el procedimiento a grandes rasgos es el siguiente:

Sean n observaciones $\theta_1, \theta_2 ... \theta_n$ de la densidad $f(z)$, donde $z_i = \theta_i - \mu$. La función de verosimilitud es:

$$L = \prod_{i=1}^{n} f(z_i)$$

Así la ecuación de verosimilitud es:

$$\frac{\mathrm{d}\ logL}{\mathrm{d}\mu} = const * \sum_{i=1}^{n} \frac{f'(\theta_i - \mu)}{f(\theta_i - \mu)} = 0 \tag{2.3}$$

Por otro lado, si μ fuera estimada por $\bar{\theta}$, también se cumple que:

$$\sum_{i=1}^{n} sen(\theta_i - \mu) = 0 \tag{2.4}$$

Por tanto, como las ecuaciones 2.3 y 2.4 se cumplen para todo θ_i y n arbitrario, la igualdad es cierta término a término y entonces:

$$\frac{f'(\theta_i - \mu)}{f(\theta_i - \mu)} = ksen(\theta_i - \mu), \quad k = cte$$

Lo que conduce a la siguiente solución general:

$$f(z_i) = ce^{kcosz_i}, \quad i = 1, 2, ...n$$

donde los parámetros c y k están condicionados por

$$\int_0^{2\pi} f(z)\ \mathrm{d}z = 1$$

y en consecuencia:

$$c = \frac{1}{\int_0^{2\pi} e^{kcosz}\ \mathrm{d}z}$$

$$\Longrightarrow c = \frac{1}{2\pi I_o(k)}$$

Donde $I_o(k)$ es la función modificada de Bessel de primer grado y de orden cero y está dada por:

$$I_o(k) = \frac{1}{2\pi}\int_0^{2\pi} e^{kcos\theta}\ \mathrm{d}\theta$$

2.2.1. Definición

Una variable aleatoria circular se dice que tiene una distribución Von Mises $V(\mu, k)$, si su función de densidad de probabilidad es:

$$f(\theta) = \frac{1}{2\pi I_o(k)} e^{kcos(\theta-\mu)}$$

con dos parámetros: $k \geq 0$, $0 \leq \mu \leq 2\pi$. El parámetro μ es la dirección media, y el parámetro k es conocido como el *parámetro de concentración.*

El centro de masa está dado por el vector medio cuya dirección la determina μ y cuya longitud se denotará como $\rho = A(k)$, donde $A(k)$ es una función definida por:

$$A(k) = \rho = \frac{I_1(k)}{I_o(k)}$$

donde $k \geq 0$, $I_0(k)$, $I_1(k)$ son funciones de Bessel[2]. El parámetro k puede ser convertido a ρ y viceversa.[3]

En la figura 2.1, se muestra la gráfica de una distribución Von Mises, donde la longitud del vector medio es ρ =.5, es decir, k =1.16

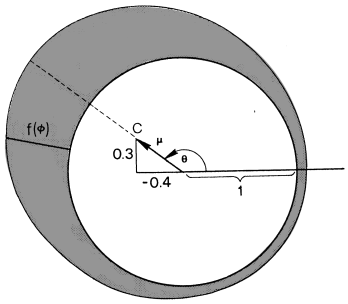


Figura 2.1: Gráfica circular de una distribución Von Mises

[2]Ver apéndice C.1.
[3]Ver apéndice C.2 y C.3.

2.2.2. Estimadores máximo verosímiles

Sea $\theta_1, \theta_2, ..\theta_n$ una muestra aleatoria con f.d.p. $V(\mu, k)$. A continuación se obtendrá el estimador máximo verosímil para el ángulo medio μ cuando k es conocida. La función de densidad de probabilidad es:

$$f(\theta) = c\ exp[kcos(\theta - \mu)], \quad \text{donde} \quad c = \frac{1}{2\pi I_0(k)}$$

La función de verosimilitud es:

$$\begin{aligned} L(\underline{\theta}, \mu) &= \prod_{i=1}^{n} f(\theta_i, \mu) \\ L(\underline{\theta}, \mu) &= (2\pi I_0(k))^{-n}\ exp \sum_{i=1}^{n} k[cos(\theta_i - \mu)] \end{aligned}$$

Aplicando logaritmo natural:

$$lnL(\underline{\theta}, \mu) = -nln(2\pi) - nlnI_0(k) + k \sum_{i=1}^{n} cos(\theta_i - \mu) \tag{2.5}$$

Al derivar con respecto a μ :

$$\frac{\mathrm{d}\ lnL}{\mathrm{d}\mu} = k \sum_{i=1}^{n} sen(\theta_i - \mu)$$

Entonces de la ecuación: $\sum_{i=1}^{n} sen(\theta_i - \bar{\theta}) = 0$, se tiene que el estimador máximo verosímil del parámetro μ de una distribución Von Mises es el ángulo medio de la muestra, es decir:

$$\widehat{\mu} = \bar{\theta}$$

Por otro lado, se obtendrá el estimador máximo verosímil para el parámetro de concentración k cuando μ es conocido. Es necesario señalar que de acuerdo a una propiedad de las funciones de Bessel, se tiene que:

$$I_0'(x) = I_1(x)$$

Ahora bien, partiendo de la ecuación 2.5 y derivando con respecto a k se obtiene:

$$\frac{\mathrm{d}\ lnL}{\mathrm{d}k} = -nA(k) + \sum_{i=1}^{n} cos(\theta_i - \mu)$$

donde:

$$A(k) = \frac{I_1(k)}{I_0(k)}$$

Por consiguiente, lnL' es cero si:

$$A(k) = \frac{1}{n}\sum_{i=1}^{n} cos(\theta_i - \mu)$$

Se observa que el lado derecho de la ecuación anterior, es el tamaño del vector medio muestral r, como lo indica la ecuación 1.2. Por tanto, el estimador máximo verosímil de k es la solución de:

$$A(\widehat{k}) = r \quad \text{o bien} \quad \widehat{k} = A^{-1}(r)$$

La solución de estas dos ecuaciones se puede resolver numéricamente [4], así el estimador máximo verosímil del tamaño del vector medio poblacional es igual al tamaño del vector medio muestral, es decir:

$$\widehat{\rho} = r$$

En el caso de que ambos parámetros fuesen desconocidos, se obtienen las derivadas parciales de la función de verosimilitud de k y μ, y la solución del sistema de ecuaciones son los estimadores máximo verosímiles de μ y k.

2.2.3. Función característica y momentos

A partir de que la distribución Von Mises es simétrica alrededor de la media μ, entonces, se define al p-ésimo momento trigonométrico alrededor de la media como:

$$\bar{\alpha}_p + i\bar{\beta}_p$$

donde:

$$\bar{\alpha_p} = E[cosp(\theta - \mu)], \qquad \bar{\beta}_p = E[senp(\theta - \mu)]$$

Análogo a la ecuación 1.1, se cumple que: $\bar{\beta}_p = E[senp(\theta - \mu)] = 0$ y también:

$$\begin{aligned} \bar{\alpha}_p &= \frac{1}{2\pi I_0(k)} \int_0^{2\pi} cos\ p(\theta - \mu)\ e^{kcos(\theta-\mu)}\ \mathrm{d}\theta \\ &= \frac{I_p(k)}{I_0(k)} \end{aligned}$$

donde I_p es la función de Bessel modificada de primer grado, la cual se define como:

$$I_p(k) = \frac{1}{2\pi} \int_0^{2\pi} cos\ p\theta\ e^{kcos\theta}\ \mathrm{d}\theta$$

Consecuentemente, la función característica de la distribución Von Mises es:

$$\phi_p = e^{ip\mu}\frac{I_p(k)}{I_p(0)}$$

En particular:

$$\alpha = A(k)cos\mu, \quad \beta = A(k)sen\mu, \quad \rho = A(k) = \frac{I_1(k)}{I_0(k)}\ .$$

[4] Valores de $A(\hat{k})$ y $A^{-1}(r)$ se encuentran en las tablas del apéndice C.2 y C.3 respectivamente.

2.2.4. Propiedades

- Simetría: Por la simetría de la función coseno, la distribución Von mises es simétrica alrededor de la media (también para $\mu + \pi$) ·
- Moda en μ: Debido a que la función coseno toma el valor máximo cuando $\theta = 0$ y así $cos\theta = 1$, la función de densidad Von Mises toma su máximo cuando $\theta = \mu$ y por tanto μ es la dirección modal con el máximo valor en:

$$f(\mu) = \frac{e^k}{2\pi I_o(k)} \tag{2.6}$$

- Antimoda en $(\mu+\pi)$: Dado que la función coseno se minimiza cuando $\theta = \pi$, y así $cos\theta = -1$, la función de densidad Von Mises cuando $\theta = \mu + \pi$ toma el valor mínimo y por tanto la antimoda direccional es $\mu + \pi$:

$$f(\mu + \pi) = \frac{e^{-k}}{2\pi I_o(k)} \tag{2.7}$$

- Concentración: k es el parámetro que mide la concentración de la población en dirección a la media. De las ecuaciones 2.6 y 2.7 se observa que:

$$\frac{f(\mu)}{f(\mu + \pi)} = e^{2k}$$

  Por tanto entre más grandes sean los valores de k más grandes será el cociente de $f(\mu)$ y $f(\mu + \pi)$ y por consiguiente la concentración en dirección a μ.

La figura 2.2 muestra la f.p.d. Von Mises con $\mu = 0$ y con diferentes valores para k . Se observa que cuando $k = 4$, con un 99 % de probabilidad los datos caen en el arco (-90º,90º).

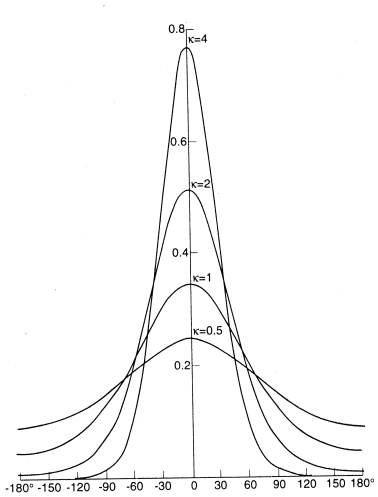


Figura 2.2: Densidad de una distribución Von Mises, $\mu = 0$ y $k = \frac{1}{2}, 1, 2, 4$

2.2.5. Relación con otras distribuciones

- Uniforme: Cuando $k = 0$ la distribución Von Mises se convierte en una Uniforme.

$$f(\theta) = \frac{1}{2\pi I(o)} e^{0cos(\theta-\mu)} = \frac{1}{2\pi * 1} = \frac{1}{2\pi}$$

- Normal bivariada: Considerando que la distribución normal bivariada tiene centro en $(m, 0)$, desviación estándar $s^1 = s^2 = 1$ y coeficiente de correlación igual a cero, se observa que la distribución bajo la restricción $x^2 + y^2 = 1$ es una distribución Von Mises con parámetro de concentración $k = m$

- Cardioide: La aproximación $exp(x) \simeq 1 + x$, muestra que para valores pequeños de k, la distribución Von Mises $V(\mu, k)$ denota una distribución cardioide $C(\mu, k/2)$, cuya f.d.p. es:

$$C(\mu, \rho) = \frac{1}{2\pi}\{1 + 2\rho cos(\theta - \mu)\}; \quad | \rho | < \frac{1}{2}$$

- Normal envuelta: La distribución Von Mises puede ser aproximada a esta distribución al comparar las distribuciones de acuerdo a su primer momento trigonométrico: $V(\mu, k) \simeq WN(\mu, A(k)), \quad k \to \infty$. La f.d.p. de la distribución normal envuelta es:

$$WN(\mu, \rho) = \frac{1}{\sigma\sqrt{2\pi}} \sum_{k=-\infty}^{\infty} exp\left\{\frac{-(\theta - \mu + 2\pi k)^2}{2\sigma^2}\right\}$$

- Cauchy envuelta: También la distribución $V(\mu, k)$ es cercana a esta distribución.

La figura 2.3 muestra una comparación entre las distribuciones Cardioide, Normal envuelta, Cauchy envuelta y Von Mises; las cuales son distribuciones simétricas y unimodales, y cuya gran diferencia es la posición de los puntos de inflexión.

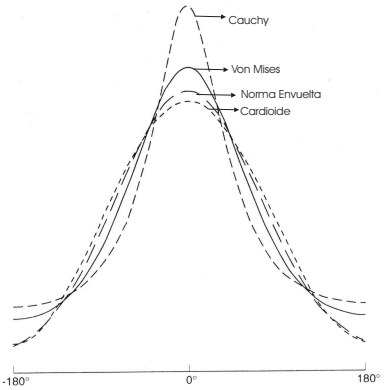


Figura 2.3: Funciones de densidad semejantes a la de Von Mises

Capítulo 3

Observaciones discordantes

3.1. Datos atípicos: outliers

3.1.1. Definición

La existencia de observaciones dudosas o raras en un conjunto de datos han sido estudiadas desde hace mucho tiempo, ciertamente desde la mitad del siglo XVIII. Estos valores se han visto como contaminantes al reducir y distorsionar la información que proporcionan sobre su fuente o mecanismo que los generó. Ha sido natural buscar formas de interpretarlos, categorizarlos y hasta rechazarlos con el fin de restaurar la propiedad de los datos o al menos tomar en cuenta su presencia en cualquier análisis estadístico. Existe una gran diversidad de términos para identificar estas observaciones, tanto en español como en inglés. Por ejemplo en español se les conoce como observaciones discordantes, aberrantes, anormales, inconsistentes etc.; mientras que en inglés como outliers, surprising values, inconsistent data, contaminants, suspect value, etc. Así pues, si se pretende dar una sola definición de estos datos, se verá que diferentes autores expresan de una u otra forma su opinión al respecto y como referencia se tiene:

Grubs (1969):

Un outlier es una observación que aparenta estar notablemente desviada de los otros miembros de la muestra en que se encuentra.

Barnett y Lewis (1978):

Un outlier en un conjunto de datos es una observación (o conjunto de observaciones) que aparenta ser inconsistente con el resto de ese conjunto de datos.

Beckman y Cook(1980)

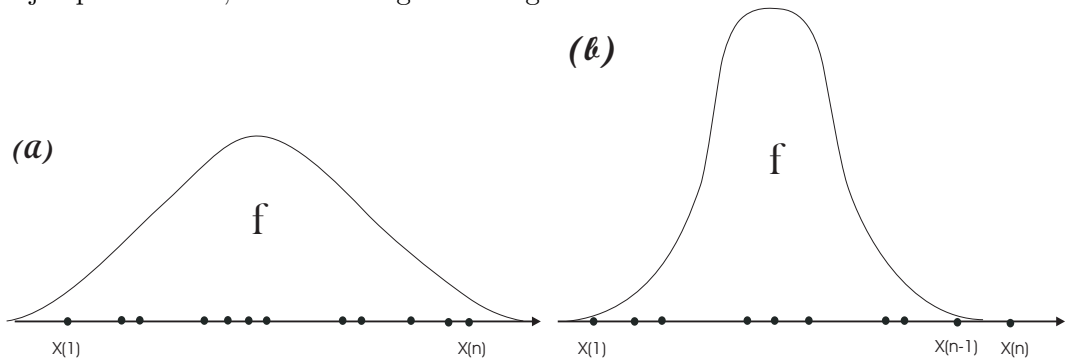
Observación discordante: Es cualquier observación que es inesperada o parece discrepante al observador.

Banett y Lewis (1983)

En sus investigaciones posteriores dan una definición más amplia y un poco más formal de cómo distinguir y definir observaciones extremas, outliers, contaminantes, etc.

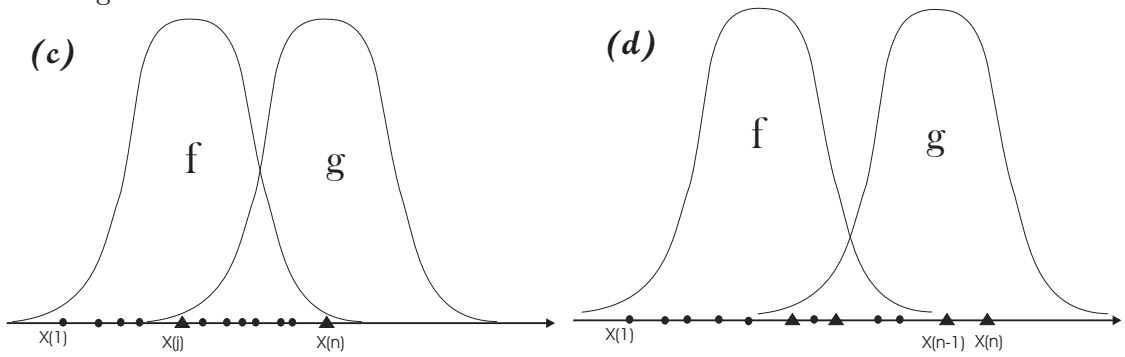
A continuación se explicarán estos términos :

Suponga que se tiene una muestra aleatoria de tamaño n: $X_1, X_2, \ldots, X_n$, cuya función de distribución es denotada como f y donde los estadísticos de orden son: $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$. Las observaciones $x_{(1)}$ y $x_{(n)}$ son valores extremos de la muestra. Si se declara o no uno de ellos como outlier, dependerá de cómo aparecen en relación al modelo postulado f.
Para ejemplificar esto, veánse las siguientes figuras:



En la figura (a), ni $x_{(1)}$ o $x_{(n)}$ parecen ser outliers. Sin embargo, en la figura (b) se observa que $x_{(n)}$ es un outlier superior, mientras que $x_{(1)}$ muestra indicios de ser un outlier inferior. Cabe destacar que se podría declarar al par $x_{(n-1)},\ x_{(n)}$ como outliers superiores. Por tanto, valores extremos pueden ser o no outliers. Sin embargo, cualquier outlier es siempre un valor extremo (o relativamente extremo) en la muestra.

Ahora, suponga que no todas las observaciones provienen de la distribución f, sino que una o dos de ellas provienen de la distribución g, la cual es un desplazamiento de f, es decir: presenta una media más grande o es una distribución diferente a f. Las observaciones de g son llamadas *contaminantes*, las cuales pudieran o no aparecer como valores extremos. En las siguientes figuras se muestra este hecho:



En la la figura (c), se aprecian dos observaciones contaminantes en la distribución f: $x_{(j)}$ y $x_{(n)}$, siendo una de ellas un valor extremo superior, y la otra el valor medio de la muestra. Sin embargo $x_{(n)}$ es un valor extremo y contaminante pero no un outlier. En contraste, en la figura (d) hay observaciones contaminates y no extremas y en particular está la pareja de observaciones x_{n-1} y x_n que son outliers, extremos y contaminantes.

Así pues, *outliers pueden ser o no contaminantes y contaminantes pueden ser o no outliers.* Ciertamente, no hay forma de saber si alguna observación es o no contaminante, todo lo que se puede hacer es considerar al outlier como una posible manifestación de contaminación y entonces examinar a través de métodos estadísticos su comportamiento.

3.1.2. Naturaleza y origen

Las observaciones atípicas: outliers, se pueden presentar por diferentes razones, y por tanto es importante conocer el motivo por el cual se origina la variabilidad o dispersión en un conjunto de datos. Cabe señalar que la naturaleza del outlier se dice aleatoria cuando no se saben las causas que lo ocasionan y es determinista cuando se identifica el por qué estan presentes.

Cuando se originan por razones puramente deterministas: por errores de cálculo o registro, entonces, el remedio es claro: se remueven los valores equívocos de la muestra o se reemplazan los valores (por ejemplo, se corrige la observación o se repite). Por otro lado, pueden presentarse en situaciones menos claras, donde se sospecha pero no se garantiza una explicación tangible de un outlier y por lo tanto no hay alternativa más que el de considerar al outlier de naturaleza aleatoria o inexplicable y por consiguiente debe ser interpretado en términos de las propiedades de variación de cualquier muestra aleatoria generada por el modelo de probabilidad propuesto, siendo algunas veces apropiado cambiar dicho modelo de distribución que represente mejor al conjunto de datos, incluyendo los anómalos.

A continuación se mencionan tres razones por las cuales se originan errores, ocasionando una gran variabilidad en la muestra de datos.

- *Variabilidad Inherente*: Es la expresión de la forma en que las observaciones varían en la población, tal variación es una característica natural de la población y propia del fenómeno que no puede ser controlado, pues refleja las propiedades de un modelo de distribución que describe la generación de los datos.

- *Error de medición*: Se produce cuando no se dispone de la técnica adecuada o cuando no existe un procedimiento en el uso de instrumentos de medición. Sin embargo, es posible controlarlo. En este error se incluye el redondeo de valores y registro de datos.

- *Error del experimentador*: Es el error atribuible al propio experimentador y se basa en la imperfecta recopilación de datos. Algunas precauciones pueden reducir tal variabilidad pero no se está del todo consciente de estos errores. Se puede presentar de dos formas diferentes:

  - *Error de información*: Surge cuando se establece un modelo o estructura matemática no adecuada o precisa a la población, o bien cuando se considera información inicial o hipótesis incorrectas.

  - *Error de planeación*: Se presenta cuando el investigador no llega a delimitar de una forma correcta el espacio y el experimento lo realiza sobre un espacio distinto. Por ejemplo, al incluir individuos que no son del todo representativos de la población o que provienen de una población diferente a la de interés.

3.1.3. Tratamiento

Como se vio anteriormente, el concepto de outlier ha fascinado a estudiosos de diferentes áreas desde el comienzo por querer interpretar datos. Incluso antes del desarrollo formal de métodos estadísticos, hubo diferentes argumentos de cuándo y en qué circunstancia se debían descartar observaciones cuando éstas no eran representativas o incluso erróneas. Así pues, distintas posturas han variado de un extremo a otro: desde rechazarlos para restaurar la propiedad de los datos o incluso adoptar métodos para reducir su impacto en el análisis estadístico. Sin embargo no es necesario adoptar uno de estos extremos, ya que al rechazar se tiene el riesgo de perder información genuina, y al aceptar se tiene el riesgo de contaminar la información. Es por consiguiente fundamental identificar la naturaleza y las razones por las cuales se presentan observaciones atípicas para establecer la estrategia a tomar.

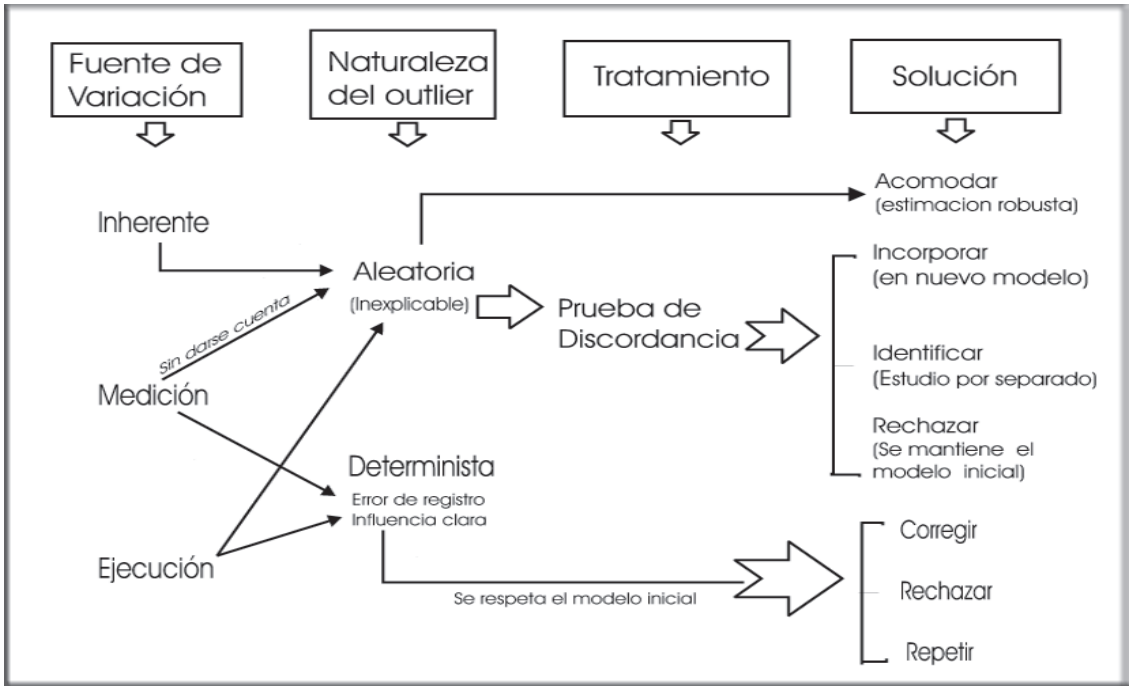
Cabe señalar, que observaciones inconsistentes no son necesariamente malas o erróneas e incluso el investigador pudiera en ciertas situaciones no rechazarlas sino aceptarlas como indicadores de algún inusual e inesperado tratamiento industrial o variedad del área de estudio. No obstante, si se muestra que tales observaciones son estadísticamente irrazonables con base en el modelo inicial de probabilidad planteado, entonces se estaría reflejando la inconveniencia de este modelo.

Cuando la razón por la cual surgen observaciones atípicas es determinista, los valores deben ser removidos o reemplazados por valores corregidos, tal vez por algún equivalente estadístico, por ejemplo, al simular las observaciones de acuerdo a la distribución supuesta. Ahora bien, cuando el problema es aleatorio, podrían examinarse procedimientos estadísticos para detectar discordancia. Sin embargo, no necesariamente se tiene que empezar con una prueba de discordancia, pues existen métodos robustos en donde se busca acomodar estas observaciones, por ejemplo: al ponderarlos con un peso menor al resto de los datos y asi tener una menor influencia en estudios futuros o aplicar el "método de Winzorización"[1] o el de α-truncada.

Con el fin de detectar si hay observaciones discordantes, hay que recordar que éstas no necesariamente se muestran como outliers. Por tanto, tomando en cuenta la definición de Barnett, se sabe que un outlier es aquél que *aparenta ser inconsistente* y lo cual implica el uso de un modelo inicial con respecto al cual es inconsistente. Barnett comenta que la frase es crucial, pues es cuestión de un juicio de carácter subjetivo por parte del observador quien decide eliminar estas observaciones o no, ya que lo que realmente preocupa es saber si estas observaciones son miembros genuinos de la población principal. Luego entonces, con el fin de tener una correcta detección de valores discordantes es necesario llevar a cabo pruebas llamadas de discordancia, y así tomar la decisión de rechazar, o establecer un nuevo modelo de distribución e incorporar los datos analizados, o bien, identificarlos como factores no conocidos de importancia práctica y estudiarlos por separado.

[1] Ver Barnett pag. 25-26 [1].

El esquema que a continuación se muestra resume lo explicado en esta sección.



3.2. Pruebas de discordancia

En las secciones anteriores, se analizó con detalle el término de observación atípica: outlier, su naturaleza y la variedad de contextos en los que pueden surgir. Se han planteado también las diferentes soluciones que se pueden tomar ante la presencia de outliers en relación a la naturaleza de éstos. Se ha visto que la presencia aleatoria de outliers conlleva a pensar si el modelo de probabilidad inicial propuesto es incorrecto Por esto, para decidir si un outlier debe considerarse como miembro de la población principal o no y tomar cualquier medida de rechazo u omisión, es indispensable ejecutar un procedimiento de detección, una prueba estadística, definida como *prueba de discordancia*.

3.2.1. Tipos de pruebas

Pruebas intuitivas

Hay que aclarar, que existen diferentes formas de proponer una prueba de discordancia, de hecho mucho del trabajo existente y de antaño tienen una forma altamente intuitiva y hay poca consideración en la naturaleza de la hipótesis alternativa y nula. Por ejemplo, al analizar pruebas intuitivas para muestras univariadas, se observa que una razón por la que la observación x_j parece aberrante en una conjunto de datos es porque se encuentra "ampliamente separada del resto de la muestra", en relación a la dispersión de ésta. Por tal motivo, se puede pensar en pruebas estadísticas de la forma N/D, donde el numerador es una medida de separación entre x_j del resto de la muestra; y el denominador D es una medida de dispersión total.

Las medidas usadas para N incluyen: desviaciones, extremos, excesos; mientras que las medidas para D incluyen: la desviación estándar, el rango. Así, al observar un outlier superior x_n en una muestra aleatoria, es natural pensar en las estadísticas que a continuación se muestran y que evaluan la discordancia de la observación:

$$\frac{x_n - x_{n-1}}{D}, \qquad \frac{x_n - \bar{x'}}{D}$$

donde x_i es el valor i-ésimo de la muestra ordenada de forma ascendente; $\bar{x'}$ es la media muestral al excluir el outlier x_n; D es alguna medida de dispersión de la muestra [2]

Sin embargo, se pueden distinguir y clasificar diferentes pruebas de base intuitiva, dependiendo si se desea examinar un solo outlier superior o dos outliers inferiores o tal vez uno superior y otro inferior. A continuación se muestran algunos de estos tipos pruebas:

- *Estadísticas de exceso-dispersión*

  Son cocientes que miden la desviación existente entre un outlier y la observación vecina más cercana a él con respecto al rango o alguna otra medida de dispersión de la muestra. Ejemplos para examinar al outlier superior x_n:

  $$\text{Dixon(1951)}: \quad \frac{x_n - x_{n-1}}{x_n - x_1} \qquad \text{Irwin(1925)}: \quad \frac{x_n - x_{n-1}}{\sigma}$$

  donde σ es la desviación estándar en el modelo básico. La estadística de Irwin supone que σ es conocido. Sin embargo, ésta se puede remplazar por algún estimador muestral que excluya la observación outlier.

- *Estadísticas de rango-dispersión*

  En estos, se reemplaza el numerador por el rango de la muestra. Por ejemplo :

  Pearson and Stephens (1964): $\frac{x_n - x_1}{s}$, donde s puede ser remplazada por alguna medida de dispersión total.

- *Estadísticas de desviación-dispersión*

  Son cocientes que miden la desviación que hay entre un outlier y alguna medida de tendencia central de la muestra. Por ejemplo, para un outlier inferior:

  Grubs(1950)

  $$\frac{\bar{x} - x_1}{s}, \qquad \frac{max|x_i - \bar{x}|}{s}$$

[2] Es posible considerar a D, omitiendo la observación x_n.

- *Estadísticas de extremo-ubicación*

  Son cocientes entre un valor extremo y alguna otra medida de tendencia central. Por ejemplo, una prueba de discordancia para un outlier superior es:

$$\frac{x_n}{\bar{x}}$$

Pruebas formales

Las pruebas de discordancia cuya base depende de la manera de plantear la hipótesis nula y alternativa, surgen en un nivel más formal de analizar y determinar ciertas características de éstas, como su nivel de significancia (tamaño de la prueba) o su potencia. Cabe destacar que al llevar a cabo una regla de decisión en términos estadísticos es necesario basarse en los principios metodológicos para pruebas de hipótesis [3]. Por tanto, hay que recordar que la hipótesis nula se conserva si no existe evidencia significativa que confirme su rechazo; mientras que la hipótesis alternativa afirma que la hipótesis nula es falsa. En una prueba de discordancia, la hipótesis nula establece como modelo de probabilidad aquel que genera todo el conjunto de datos sin considerar la presencia de outliers; mientras que la hipótesis alternativa expresa la manera en que el modelo debe ser modificado con el fin de incorporar o explicar la presencia de outliers. Por esta razón, es necesario señalar que existen distintas formas de plantear pruebas de discordancia ya que dependen de la manera en que la hipótesis alternativa es especificada. A continuación se explicarán algunas de estas pruebas, considerando que para cada una se establece que la hipótesis nula afirma que todas las observaciones provienen de la distribución F.

$$H_o : x_j \ \sim \ F \qquad \forall j = 1, 2, ..., n$$

a) Alternativa inherente

Supone que los outliers aparecen en un conjunto de datos debido a un grado inesperado de variabiliad inherente en el modelo, por lo que la distribución pudiera en verdad tener colas más pesadas. Se admite la posibilidad de que el modelo de probabilidad es otro, tal vez de la misma familia de distribución o diferente.

La hipótesis alternativa establece que todos los datos provienen de la distribución G,

$$H_a : x_j \ \sim \ G \qquad \forall j = 1, 2, ..., n$$

b) Alternativa mixta

Admite que la muestra refleja un ligero grado de contaminación al aceptar la presencia de algún miembro extraño que se muestra como outlier. La hipótesis alternativa declara que los outliers tienen la posibilidad λ de provenir de algún otro modelo de distribución G y el cual es algo diferente al modelo inicial F y se establece como:

$$H_a : x_j \ \sim \ (1 - \lambda)F \ + \ \lambda G \qquad \forall j = 1, 2, ..., n$$

[3] Ver apéndice B.1.

c) Alternativa de deslizamiento

Plantea que todas las observaciones a excepción de algún número pequeño de ellas k (una o dos) provienen del modelo inicial F, con parámetros de posición y escala; mientras que las k observaciones restantes provienen de una distribución modificada de F, es decir, donde alguno de los parámetros han sido cambiados. Es el tipo de hipótesis alternativa más frecuente para establecer un modelo de contaminación. Por ejemplo, Guttman (1973), declara que en la distribución normal, cuando un parámetro es desconocido, se puede establecer una hipótesis alternativa que considera que el contaminante proviene del mismo modelo pero con media diferente:

$$H_a : x_j \sim N(\mu + a, \sigma^2)$$

De esta forma, se ha analizado la importancia de la hipótesis alternativa en las pruebas de discordancia y de la relevancia del método intuitivo. Sin embargo, además de estos procedimientos básicos para formular pruebas de discordancia, existen otros métodos extensamente aplicables para llevarlas a cabo, como por ejemplo el *Principio del cociente de verosimilitudes* [4], donde la construcción de estas pruebas depende en primera instancia de la hipótesis alternativa empleada para explicar los outliers. En general esta prueba se construye a partir de la función de verosimilitud de la muestra bajo la hipótesis nula y alternativa: $L_{H_0}(\theta)$ y $L_{H_a}(\theta)$. Se basa en encontrar el punto en donde se maximizan dichas funciones: $L_{\widehat{H}_0}(\theta)$, $L_{\widehat{H}_a}(\theta)$; y así, a partir de ellos, obtener la razón de verosimilitudes , la cual es: $lnL_{\widehat{H}_0} - lnL_{\widehat{H}_a}$

3.2.2. Evaluación de la prueba

En general se ha explicado el concepto de pruebas de discordancia y las diversas formas de plantearlas. Ahora bien, es importante conocer el criterio que permite determinar cuando una prueba es mejor que otra. Supóngase que se desea probar si la observación x_n es discordante, utilizando la prueba Z. Así pues, se calcula el valor de Z dada la observación x_n y se denota como Z_n. Entonces, se declara que x_n es discordante si $Z_n > z_\alpha$, donde z_α es el valor crítico de Z dado el nivel de significancia α, el cual se define como:

$$P[Z_n > z_\alpha | H_o] = \alpha$$

Tomando en cuenta la hipótesis alternativa de deslizamiento, se cree que una de las observaciones de la muestra es un contaminante x_c, siendo el valor de la prueba de discordancia Z_c. En el contexto de este particular conjunto de supuestos, David (1981) sugiere las siguientes probabilidadades como "medidas razonables" para evaluar el desarrollo de Z_n bajo la hipótesis alternativa.

[4] Ver apéndice B.2.

La probabilidad de que un outlier sea identificado como discordante.

$$P_1 = P[Z_n > z_\alpha | H_a]$$

La probabilidad de que el contaminante sea identificado como discordante.

$$P_2 = P[Z_c > z_\alpha | H_a]$$

La probabilidad de que el contaminante sea un outlier y que sea identificado como discordante.

$$P_3 = P[Z_c = Z_n, Z_n > z_\alpha | H_a]$$

La probabilidad de que cuando el contaminante es realmente un outlier sea identificado como discordante.

$$P_4 = P[Z_c = Z_n > z_\alpha, Z_{n-1} < z_\alpha | H_a]$$

$$P_5 = P[Z_c > z_\alpha | z_c = Z_n; H_a]$$

Capítulo 4

Pruebas de discordancia en la estadística circular

4.1. Outliers para datos circulares

En la línea, entre más extremo es el valor de un outlier, mayor es la separación del resto de la muestra de datos. Sin embargo, en la circunferencia un valor extremo no implica ser una observación angular extrema. Por ejemplo, considere el siguiente conjunto muestral de datos angulares:

$$10^o, 18^o, 33^o, 48^o, 67^o, 349^o$$

Si estos datos fueran lineales, no se tendría ninguna duda en decir que el valor 349^o es extremo, pero dado de que ésta es una muestra de datos circulares, donde las observaciones varían entre 0^o y 360^o, el valor 349^o es perfectamente consistente con los otros. Si 349^o es de hecho un outlier, entonces éste ha aparentado ser una observación respetable. Claramente, se espera solamente encontrar outliers en muestras de datos angulares cuando la masa principal de los datos se encuentra suficientemente concentrada sobre una dirección en particular. En este ejemplo, la muestra tiene una baja concentración y será difícil encontrar una sola observación que esté suficientemente separada del resto y que muestre evidencia de ser un outlier.

Ahora bien, observe la siguiente muestra:

$$8^o, 10^o, 13^o, 31^o, 32^o, 40^o, 69^o, 135^o, 314^o, 325^o, 344^o, 347^o, 350^o$$

Se pensaría que el valor outlier es el valor extremo de la muestra: 350^o. Sin embargo, al graficar los datos, se observa que el valor más alejado del resto de los datos concentrados es $135\ ^o$, *veáse la figura 4.1.* Esto hace ver que existe una manera de indentificar un valor outlier, por ejemplo, al basarse en las desviaciones que generan los datos angulares con respecto a la dirección del vector medio y el cual se expondrá a continuación.

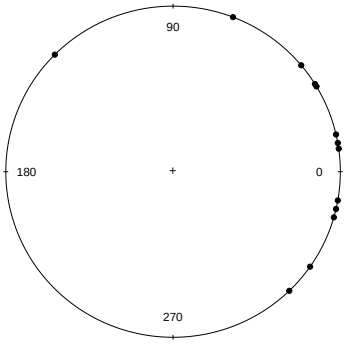


Figura 4.1: Gráfica de una muestra de datos direccionales donde aparece un outlier que presenta la desviación más grande con respecto al resto de la masa de datos y es identificado como 135°

4.2. Identificación de outliers

A continuación se examinará como valores inesperados, se manifiestan en una muestra de datos direccionales. Sea $\theta_1, \theta_2, ..., \theta_n$ una muestra circular de tamaño n, entonces una observación $\theta_k\,(1 \leq k \leq n)$ será inconsistente, cuando las desviaciones angulares de θ_i a la dirección media $\bar{\theta}$, está dado por:

$$\xi_i = min(\theta_i^*, 2\pi - \theta_i^*) = \pi - \mid \pi - \theta_i^* \mid \tag{4.1}$$

donde $\theta_i^* = \theta_i - \bar{\theta}, (mod\, 2\pi)$, es el máximo sobre $i = 1, 2, ..., n$. Lo que significa que la observación θ_k es en efecto un outlier cuando se cumple que: $max_i\{\xi_i\} = \xi_k$.

Si se escribe $C = \sum cos\theta_i$ y $S = \sum sen\theta_i$, entonces la dirección media muestral $\bar{\theta}$ y el tamaño del vector resultante muestral $\|\bar{R}\| = \mathcal{R}$, están dados por: $C = \mathcal{R}cos\bar{\theta}$, $S = \mathcal{R}sen\bar{\theta}$, con $\mathcal{R}^2 = C^2 + S^2$. Si al valor del tamaño del vector resultante muestral se le omite la observación i-ésima se tiene que:

$$\begin{aligned}
\mathcal{R}_i^2 &= (C - cos\theta_i)^2) + (S - sen\theta_i)^2 \\
&= C^2 - 2Ccos\theta_i + cos^2\theta_i + S^2 - 2Ssen\theta_i + sen^2\theta_i \\
&= \mathcal{R}^2 + 1 - 2(Ccos\theta_i + Ssen\theta_i) \\
&= \mathcal{R}^2 + 1 - 2(\mathcal{R}cos\bar{\theta}cos\theta_i + \mathcal{R}sen\bar{\theta}sen\theta_i) \\
&= \mathcal{R}^2 + 1 - 2\mathcal{R}cos(\theta_i - \bar{\theta})
\end{aligned}$$

Consecuentemente, mientras la desviación angular de θ_i a $\bar{\theta}$, ξ_i, incrementa de 0 a π, $cos(\theta_i - \bar{\theta})$ decrece de 1 a -1 y $\mathcal{R}_i^2$ incrementa de $(\mathcal{R} - 1)$ a $(\mathcal{R} + 1)$. En otras palabras, la omisión de θ_i, la observación más alejada de $\bar{\theta}$, conlleva al incremento más grande del valor del tamaño del vector muestral resultante. Entonces el valor de θ más alejado de la dirección media muestral es el candidato para ser examinado como un outlier.

4.3. Pruebas de discordancia para muestras Von Mises

Considerando el caso particular de una muestra aleatoria de tamaño n, de una distribución Von Mises $V(\mu, k)$, se sabe que su función de densidad es:

$$f(\theta) = \frac{1}{2\pi I_o(k)} e^{kcos(\theta-\mu)}, \quad 0 \leq \theta < 2\pi$$

donde $0 \leq \mu < 2\pi$, $k > 0$, y $I_0(k)$ es la función de Bessel modificada de primer grado y de orden cero.

Con el fin de evaluar si un valor inesperado en una muestra Von Mises es o no en realidad un dato discordante, es necesario llevar a cabo una prueba de discordancia con base en la hipótesis alternativa de deslizamiento. Así pues, se establece como hipótesis nula que las n observaciones provienen de la distribución $V(\mu, k)$. Por otro lado, la hipótesis alternativa afirma que $n-1$ observaciones provienen de $V(\mu, k)$, mientras que aquella observación restante procede de $V(\mu^*, k)$, donde $\mu^* \neq \mu$. Cabe señalar, que se está suponiendo que el valor contaminante se deriva de una distribución con diferente dirección media que del resto de la muestra, esto es, un modelo que tiene observaciones contaminantes.

A continuación se describiran cuatro estadísticas: L, C, D, M, para evaluar la posible discordancia de una observación angular considerada como outlier. Dos de ellas: L y M son definidas con una particular referencia del modelo Von Mises; las otras dos: C y D, se basan en consideraciones intuitivas y pueden ser utilizadas para otros modelos.

4.3.1. Estadística L

Suponga que θ_k es aquella observación cuya desviación angular de la dirección media muestral es la máxima de todas las demás observaciones. Entonces el cociente de verosimilitudes basada en la hipótesis alternativa de deslizamiento es:[1]

$$L = (R_k + 1)\widehat{k}_k - \widehat{k}R - nln\left\{\frac{I_0(\widehat{k}_k)}{I_0(\widehat{k})}\right\}$$

donde $\widehat{k}$ es el estimador máximo verosímil de k, y que como ya se vio, está dado por $A(\widehat{k}) = r$, donde $R_k^2 = C_k^2 + S_k^2$, con C_k y S_k que son los valores de C y S basados en $n-1$ observaciones, excluyendo a θ_k y donde el estimador de $\widehat{k}_k$ es tal que $A(\widehat{k}_k) = (R_k + 1)/n$.

[1] Ver ejemplo en el apéndice B.2.

4.3.2. Estadística C

Esta estadística está basada en el incremento relativo del valor del tamaño del vector medio muestral $\bar{\mathcal{R}}\,(=\mathcal{R}/n)$ al omitir el valor dudoso. Si se denota al tamaño del vector medio muestral basado en los $n-1$ valores al omitir θ_i, por $\bar{\mathcal{R}}_i = \mathcal{R}_i/(n-1)$.

Se observa que entre más alejado se encuentre θ_i del resto de los valores muestrales, más grande será el valor de $\bar{\mathcal{R}}_i$ en relación a $\bar{\mathcal{R}}$. De esta forma, un criterio natural es usar $max_i\{\bar{\mathcal{R}}_i/\bar{\mathcal{R}}\}$ o en la forma en la que se trabajará:

$$C = max_i \left\{ \frac{\bar{\mathcal{R}}_i - \bar{\mathcal{R}}}{\bar{\mathcal{R}}} \right\}$$

Si se denota: $max_i\{\bar{\mathcal{R}}_i\} = \bar{\mathcal{R}}_k$, entonces:

$$C = \left\{ \frac{\bar{\mathcal{R}}_k - \bar{\mathcal{R}}}{\bar{\mathcal{R}}} \right\}$$

4.3.3. Estadística D

Un segundo criterio intuitivo está basado en el uso del tamaño relativo de arcos, y es en esencia una estadística de tipo Dixon. Si se forman las estadísticas de orden lineal de la muestra, $\theta_{(1)}, \theta_{(2)}, ..., \theta_{(n)}$ y se denotan los arcos que se forman entre observaciones consecutivas como T_i, entonces:

$$\begin{array}{rcll} T_i & = & \theta_{(i+1)} - \theta_{(i)}, & i = 1, 2, ..., n-1 \\ T_n & = & 2\pi - \theta_{(n)} + \theta_{(1)} & \end{array}$$

De esta forma, una posible estadística para evaluar la discordancia de θ_i puede ser expresada como:

$$D_i = \frac{T_i}{T_{i-1}}, \quad i = 1, 2, ..., n$$

donde queda definido $T_0 \equiv T_n$.

Claramente, D_i es independiente de la localización de los θ's. Si la lejanía de θ_k del conjunto datos es notoria, entonces θ_k debe encontrarse en el arco de mayor tamaño que contiene sólo una observación. Entonces D_k es simplemente la razón de los arcos ya sea de un lado o del otro del valor. Los valores de D_k menores o mayores a uno, sugieren que θ_k es un outlier cuando el resto de los valores de la muestra están suficientemente concentrados. La estadística D_k como fue definida es de dos colas y por tanto, por conveniencia, se trabajará en términos de $D = min(D_k, D_k^{-1})$, $0 < D < 1$.

4.3.4. Estadística M

Fisher [2] sugiere la siguiente prueba estadística para evaluar la discordancia de un outlier:

$$M' = \underset{i}{min} \left\{ \frac{n-1-R_i}{n-R} \right\}$$

Por convenciencia se trabajará en términos de $M = 1 - M'$, donde $R_k = max\{R_i\}$

$$M = \underset{i}{max} \left\{ \frac{R_i - R + 1}{n-R} \right\}$$

$$M = \frac{R_k - R + 1}{n-R}$$

4.3.5. Puntos percentiles de las pruebas estadísticas

La forma algebraica de la distribución de cada prueba estadística es extremadamente difícil de conocer y es por esto que Collet en su artículo utiliza el Método Monte Carlo para obtener los puntos percentiles de las pruebas bajo la distribución nula. Menciona que es posible obtener la distribución asintótica de la estadística M para valores grandes del parámetro k y entonces se puede mostrar que la distribución nula de M tiende a la distribución nula de $n(n-1)b^2*$, donde $b*$ es la estadística que se utiliza para la prueba de discordancia de muestras normales univariadas.

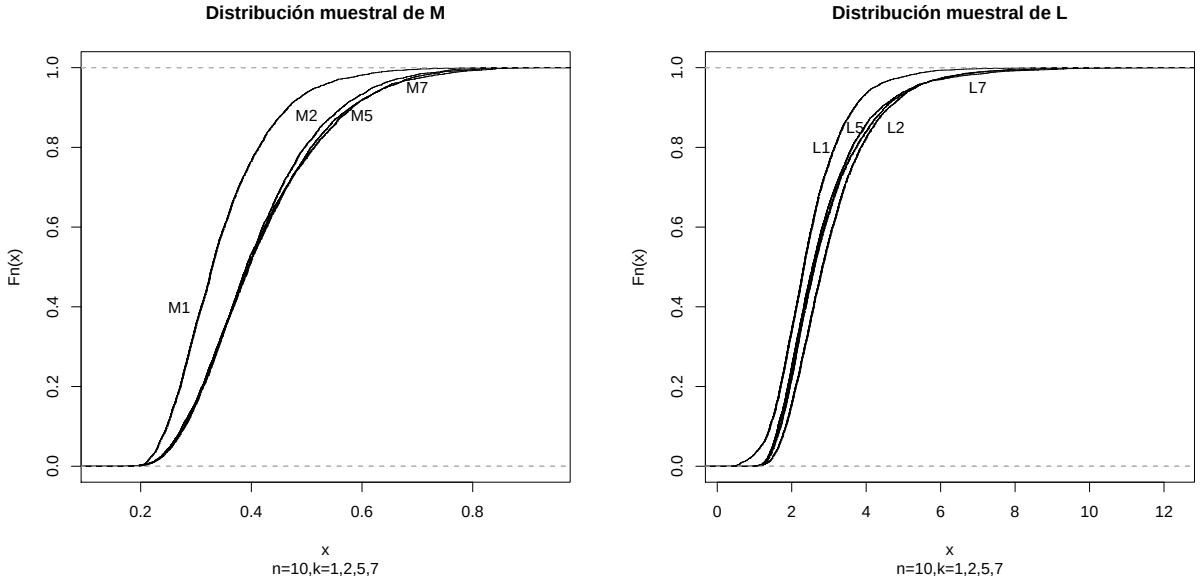
$$b* = max_i \left\{ \frac{|x_i - \bar{x}|}{\sum (x_i - \bar{x})^2} \right\}$$

Collet encontró que los valores percentiles simulados para la estadística M eran exactamente iguales a los valores conocidos de la distribución asintótica de M para valores grandes de k. Por consiguiente los puntos percentiles de la distribución nula de M para valores grandes de k se pueden obtener de aquellos de la distribución nula de $b*$.[3]

Con el fin de simular los puntos percentiles de la distribución nula de las cuatro estadísticas anteriormente expuestas y analizar sus distribuciones muestrales, se realizó una simulación en R^4 para generar 5000 muestras aleatorias de tamaño n de la distribución Von Mises $V(0,k)$ con diferentes valores de n y k. Los valores de las cuatro estadísticas M, C, D, L fueron determinadas para cada una de las 5000 muestras, y cada conjunto de esos 5000 valores fueron ordenados de manera ascendente. Los valores percentiles se obtuvieron por medio de las estadísticas de orden correspondientes y su probabilidad empírica acumulada. Este procedimiento se llevó a cabo para diferentes n y k. Las tablas obtenidas por Collet de los valores percentiles al 5 % y 1 % de la distribución nula de C y D para k=2,3,4,5,7.5,10 se muestran en el apéndice C.4 y C.5 respectivamente.

[2] Fisher, 1995 [6].
[3] Ver tabla en apéndice C.6.
[4] Ver código en apéndice D.2.

Se verificó en la simulación que los valores percentiles estimados de las estadísticas L y M son independientes de los valores de k cuando $k \geq 2$ y para cada valor de n. Con el fin de ilustrar este hecho a continuación se muestran las gáficas de las distribuciones muestrales empíricas de L y M en el caso cuando n=10, $k = 1, 2, 5, 7$. Así pues, se nombró a la distribución L1 cuando k=1, L2 cuando k=2, L5 cuando k=5 y L7 cuando k7; y de igual forma para M. Se observa claramente como las distribuciones convergen a partir de que $k \geq 2$ por lo que la distribución nula de L y M tiende a una distribución asintótica independiente de k para valores grandes de k.



4.3.6. Evaluación de las pruebas de discordancia

Diferentes medidas para evaluar el desempeño de pruebas de discordancia fueron propuestas por David [1970] y posteriormente examinadas por Barnett y Lewis, quienes concluyeron que una "buena" prueba tiene:

1. Alta potencia,
2. Alta probabilidad de identificar un valor contaminante como outlier cuando es en efecto un valor extremo
3. Baja probabilidad de identificar erróneamente una buena observación como discordante.

Usando la notación de David, se escribirá:

- $P1 \rightarrow$ Función potencia
- $P3 \rightarrow$ Probabilidad de que el contaminante sea extremo y sea identificado como discordante
- $P5 \rightarrow$ Probabilidad de que cuando el contaminante es realmente extremo sea identificado como discordante.

Entonces una buena prueba estará caracterizada al tener alto $P1$, alto $P5$ y bajo $P1-P3$, este último mide la probabilidad de identificar erróneamente una buena observación como discordante, cuando el contaminante está presente.

Para evaluar el funcionamiento de las pruebas de discordancia con base en la hipótesis alternativa de deslizamiento, habrá que calcular las probabilidades anteriormente expuestas, suponiendo que $n-1$ observaciones de la muestra de tamaño n provienen de la distribución $VM(\mu, k)$ mientras que la observación restante proviene de la distribución $VM(\mu + \lambda\,\pi, k)$, $0 \leq \lambda \leq 1$. Esta estructura implica que mientras λ incrementa de 0 a 1, el valor contaminante pasa de ser un miembro de la misma población que las restantes $n-1$ observaciones, a un miembro de una distribución cuya dirección media es diametralmente opuesta al resto de los valores. Sin pérdida de generalidad, se supone que $\mu = 0$; así la muestra de tamaño n, consiste en $n-1$ observaciones de la distribución $VM(0, k)$, más la observación proveniente de $VM(\lambda\,\pi, k)$. Cabe señalar que este procedimiento se puede realizar también con referencia a la hipótesis alternativa en donde el outlier proviene de $VM(\mu, k^*)$.

Análisis de Collet

Con el fin de analizar el desempeño de las pruebas, Collet menciona que con la estructura descrita anteriormente, realizó una simulación por medio del método Monte Carlo al generar 2000 muestras de tamaño n, para n=5, 10, 15 ; $k = 1, 2, 5, 10$ y $\lambda = 0$ (.1) 1 y así estimar los valores de $P1$, $P5$ y $P1-P3$, usando un nivel de significancia del 5 %. Utilizó las mismas muestras para cada una de las estadísticas de prueba con el fin de tener una mejor precisión al compararlas.

Al analizar P5, Collet encontró que la prueba basada en la estadística L funciona mejor que la M, mientras que ambas son inferiores a C y D, C siendo mejor que D. Observa que valores de P5 sólo se aproximan a un nivel razonable cuando $k \geq 2$. Sin embargo mientras el tamaño de la muestra aumenta, las diferencias en el desempeño de las cuatro pruebas se vuelven menos notorias.

Al analizar los valores de la función potencia para diferentes valores de n y k, se encontró que para grandes valores de λ, la estadística C es más potente que la D y ambas a la vez, mejores que L y M. Para valores grandes de n hay menos diferencia entre las pruebas.

Al analizar la medida de desempeño P1-P3, se encontró que para las pruebas L y M, los valores de P1-P3 son muy cercanos a cero. (al menos, en la base de las 2000 simulaciones.) Para la prueba C y D, es en general un poco más alta. Sin embargo los valores de P1-P3 nunca exceden el .05, y se observa que mientras k y n aumentan, dichos valores disminuyen.

Así pues, Collet concluye que para tamaños de muestra mayores a 15 y grandes valores de k, la diferencia entre las cuatro pruebas es muy poca. Sin embargo para muestras pequeñas las pruebas más potentes son aquellas cuya distribución muestral no dependen del valor de k. La prueba M se recomienda en ciertos casos pues los puntos percentiles son fáciles de determinar. Sin embargo, para muestras pequeñas y valores moderados de k, las estadísticas C y D son superiores a M en términos de sus desempeños. De hecho, la prueba D es mucho más fácil de calcular que C.

Análisis con una simulación actual

Con el fin de analizar y entender el procedimiento para evaluar las pruebas de discordancia a través de la función potencia y P3, se realizó una simulación en R [5] al generar 3000 muestras con $n1$ elementos provenientes de la distribución $VM(0, k)$, y $n2$ elementos discordantes provenientes de $VM(\lambda\,\pi, k)$, $0 \leq \lambda \leq 1$, $k > 0$, para identificar el valor outlier (extremo) de cada muestra, obtener las estadísticas C, D, M, L, y realizar la prueba de hipótesis para cada valor outlier, considerando que los valores críticos de cada prueba se pueden obtener por medio de la simulación anteriormente explicada, y entonces rechazar o no la hipótesis nula, con el objeto de determinar si se detectó correctamente el outlier discordante como contaminante.

Debido a que sólo se conocen las distribuciones asintóticas de las pruebas C, D, L, M, la forma de estimar la función potencia de las pruebas es por medio del Método Monte Carlo cuyo procedimiento es el siguiente:

1. Se selecciona un valor particular de λ bajo la hipótesis alternativa.
2. Se generan j muestras aleatorias: $\theta_1^j, \theta_2^j, ..., \theta_{n1}^j, ..., \theta_{n1+n2}^j$; cada una de tamaño n=n1+n2, donde $j = 1, ..., m$.
3. Se obtiene el valor outlier out_j para cada muestra.
4. Se observa si el valor outlier proviene de la muestra generada con distribución $VM(\lambda\pi, k)$. —La variable $od_j = 1$ en caso afirmativo o $od_j = 0$ en caso contrario.
5. Se realizan las pruebas de discordancia para la muestra j-ésima.
6. Se cuenta el número de veces en que se rechaza la hipótesis nula para cada prueba.
   —La variable $d_j = 1$ si para la muestra j-ésima se detecta un valor discordante al nivel de significancia α y $d_j = 0$ en caso contrario.
7. Se cuentan los casos para los cuales se rechaza la hipótesis nula y en efecto el valor out_j es un valor discordante.
   — Si las variables $od_j = 1$ y $d_j = 1$, entonces la variable $r_j = 1$ en caso afirmativo, $r_j = 0$ en otro caso.
8. La función potencia es: $\hat{\pi} = \sum_{j=1}^{m} \frac{r_j}{od_j}$

Siguiendo la misma metodología el calculo de P3 es:

$$\hat{P3} = \sum_{j=1}^{m} \frac{r_j}{n2 * m}$$

donde $n2 * m$ es el total de discordantes en la simulación.

[5] Ver código en el apéndice D.3.

Así pues, con el fin de observar el comportamiento de la función potencia, se tomaron en cuenta dos valores de λ bajo la hipótesis alternativa: λ =.5, 1. Asimismo se evaluaron muestras de tamaño $n = 8,\ 15\ ,30$, con $k = 2,\ 3,\ 7$. Los valores simulados de los cuantiles para cada prueba y bajo estos parámetros se encuentran en el apéndice C.7.

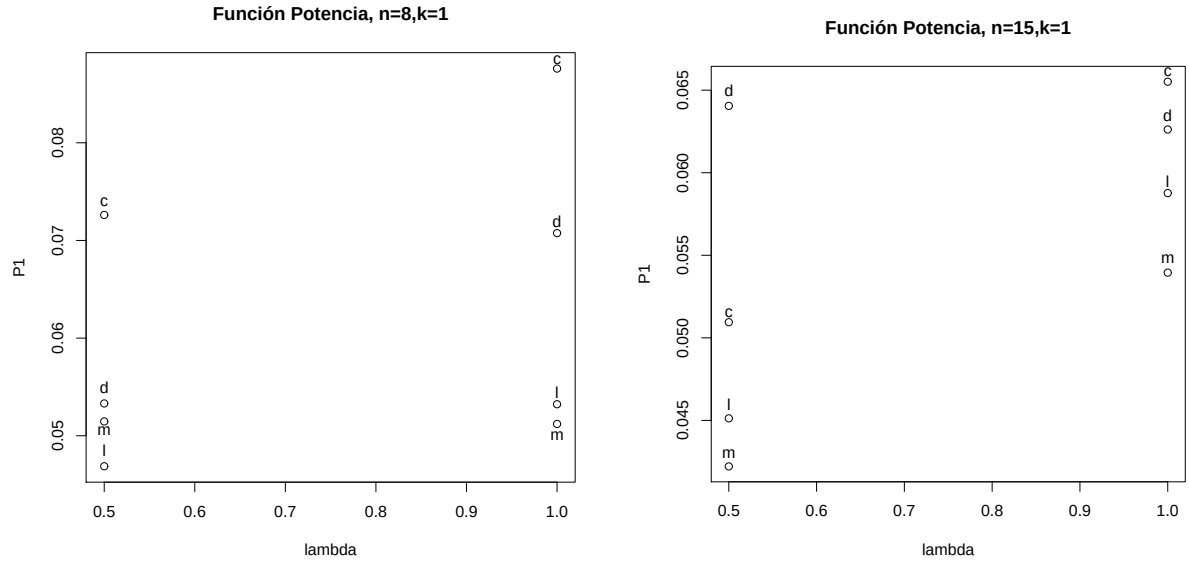
El programa permite hacer un conteo de las siguientes posiblidades:

- Detectar un outlier discordante como discordante=r → Acierto
- Detectar un outlier discordante como no discordante=od-r → Error
- Detectar un outlier No discordante como discordante=d-r → Acierto
- Detectar un outlier No discordante como discordante=m+r-(od+d) → Error

A continuación se muestra una tabla de los totales con los diferentes casos a evaluar considerando la presencia o no de outliers discordantes bajo la hipótesis nula y alternativa.

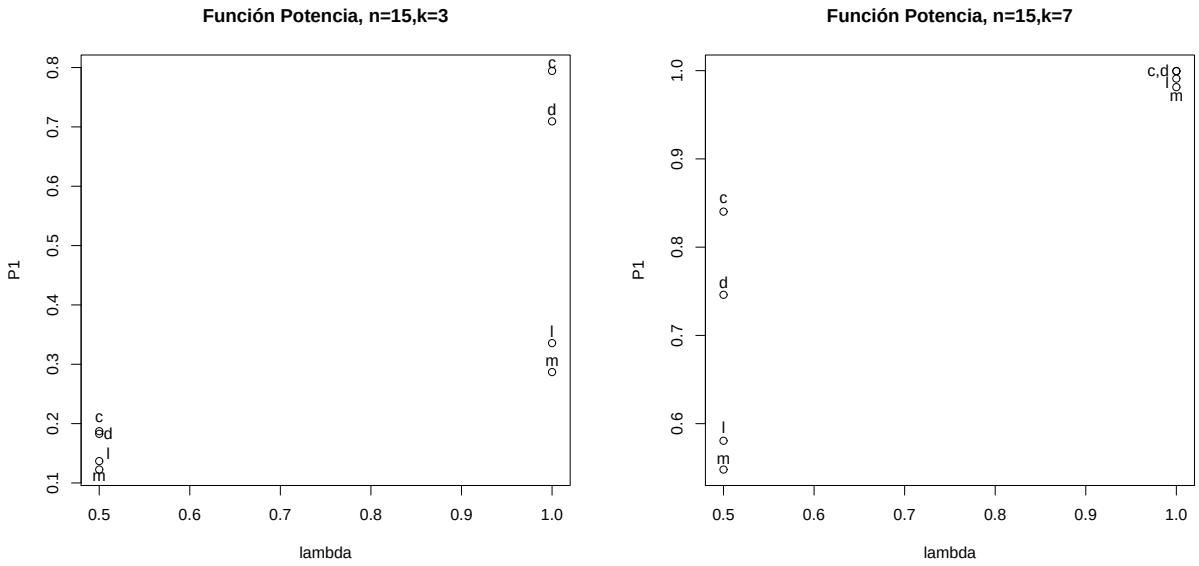
	Hipótesis Alternativa	Hipótesis Nula	
Outliers	Discordantes	No Discordantes	Totales
Detectados	r	d-r	d
No Detectados	od-r	m+r-(od+d)	m-d
Totales	od	m-od	m

Entonces al evaluar la función potencia de las pruebas para las diferentes n y k propuestas, se observa que cuando el parámetro de concentración es bajo, la potencia también lo es, cuando sólo esta presente un contaminante en cualquier muestra de tamaño n. Para ejemplificar esto, a continuación se presentan dos gráficos de la función potencia:



Por otro lado mientras más grande es k, la función potencia aumenta al presentarse un sólo contaminante. Por ejemplo, cuando $k = 3$ y $n = 15$, la potencia es mucho mejor que cuando $k = 1$, sin embargo, siguen siendo mucho mejor las pruebas C y D sobre la M y D.

Pero si se analiza el caso cuando $k = 7$ y $n = 15$, se observa que todas las pruebas son todavía más potentes y tienden a tener la misma potencia cuando $\lambda = 1$.



Entonces entre más grande sea el tamaño y la concentración de la muestra, la potencia de las pruebas serán mucho mejores, y convergerán al mismo valor. Por otro lado, P1-P3 son muy cercanas a cero, y mientras más se acerca λ a uno, la probabilidad de cometer un error e identificar una observación buena como discordante es baja para todas las pruebas.

4.4. Ejemplo: movimiento de estrellas de mar

A continuación se presenta un ejemplo correspondiente a la aplicación práctica de datos direccionales en la rama de la Biología que estudia la orientación que siguen los animales bajo cierta situación específica. Así pues, se muestran las direcciones que 22 estrellas de mar de la especie *Sidinian* tomaron después de 11 días en movimiento al ser desplazadas de su habitat natural. Cabe señalar que la posición de la costa se consideró aproximadamente de cero grados. Los datos se muestran en la siguiente tabla:

Observación (x_i)	***Dirección de estrellas de mar*** $(^\circ)$	***Observación*** (x_i)	***Dirección de estrellas de mar*** $(^\circ)$
1	0	12	45
2	1	13	147
3	3	14	298
4	3	15	329
5	8	16	332
6	13	17	335
7	16	18	340
8	18	19	350
9	30	20	354
10	31	21	356
11	43	22	357

El objetivo del ejercicio, es evaluar la presencia de algún valor discordante en la muestra de acuerdo a las estadísticas de prueba anteriormente explicadas. Por tanto, es necesario establecer la hipótesis nula y la hipótesis alternativa para realizar la prueba de hipótesis que declarará o no al valor alejado como discordante.

Sea:

H_o: Todas las observaciones de la muestra provienen de la distribución $VM(\mu,\ 3.3)$[6]

$$H_o : \mu = 0 \quad \forall\, \theta_1, ..., \theta_n$$

H_a: n-1 observaciones de la muestra provienen de la distribución $VM(\mu$,3.3) a excepción de una de ellas que proviene de la misma distribución pero con parámetro μ diferente, es decir, de la distribución $VM(\mu + \lambda\,\pi$,3.3), donde $0 \leq \lambda \leq 1$

$$H_a : \mu = 0 \quad \forall\, \theta_1, ..., \theta_{n-1} \quad \text{y} \quad \mu = \lambda * \pi, \quad p.a.\ \theta_j,\ \ j = 1, ..., n$$

Ahora bien, con el fin de detectar de forma sencilla la posible presencia de valores inconsistentes en la muestra, es útil realizar una gráfica de dispersión y detectar la observación que más se aleja del conglomerado de datos. Es por esto que en la Figura 4.2 se presenta la gráfica de las direcciones que tomarón las 22 estrellas de mar. La gráfica muestra dos observaciones incosistentes o outliers. Una de las estrellas de mar parece haber tomado una ruta más alejada de la playa y cuya dirección corresponde a los 298°. Por otro lado, una estrella salió del mar, tomando una dirección de 147°.

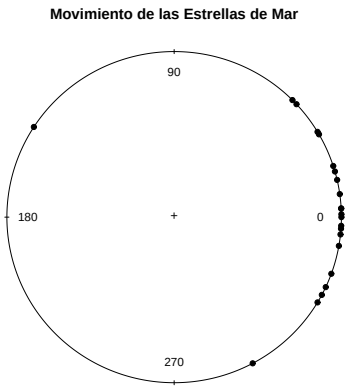


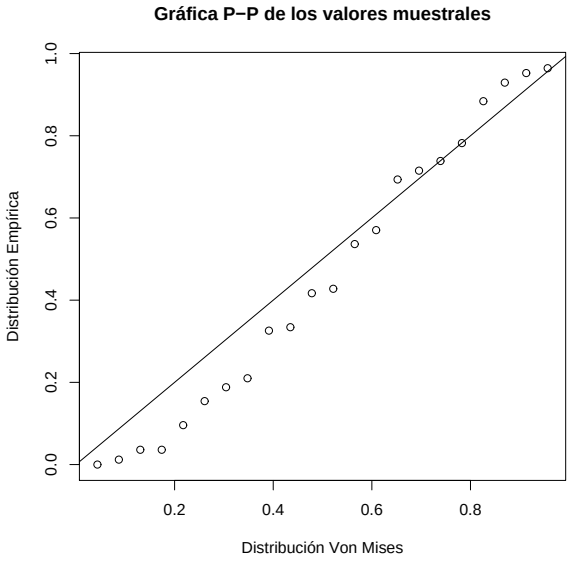
Figura 4.2: Observaciones direccionales del movimiento de 22 estrellas de mar

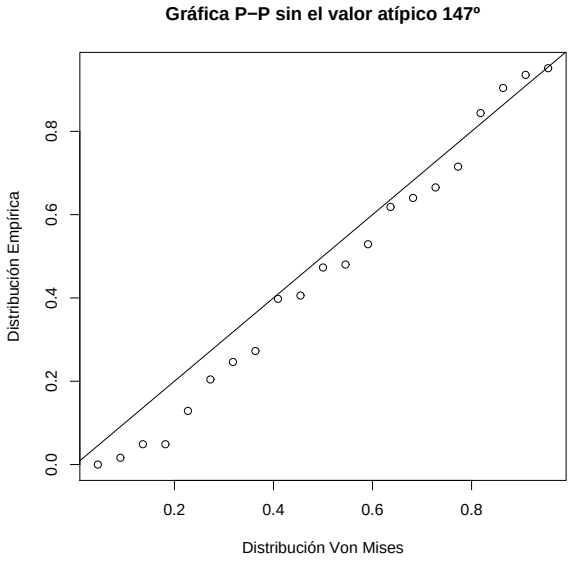
El punto es identificar cuál de esas dos observaciones "inconsistentes" es la que se evaluará como posible dato discordante, y para lo cual, es necesario evaluarlas en términos de la distancia circular, que permitirá identificar el outlier que esta demasiado alejado de la media.

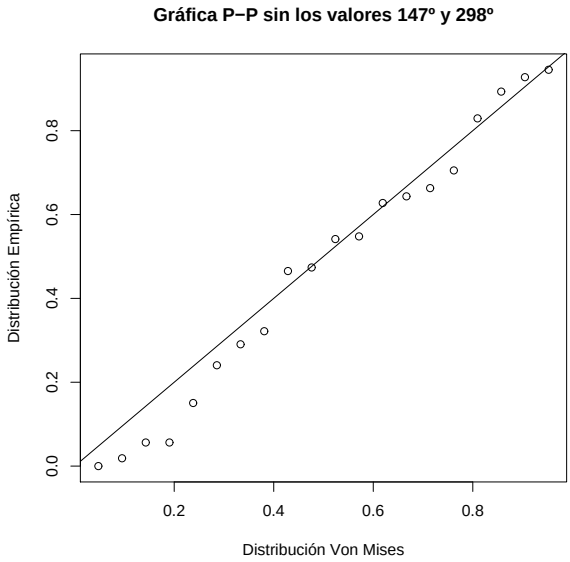
[6]Cabe destacar que para este ejemplo k fue estimada a través de la función $A^{-1}(\rho)$.

Así pues, al evaluar la distancia circular que hay entre las observaciones y la media, el valor atípico es la observación 147°, la cual efectivamente se identificó en la gráfica de dispersión circular. La pregunta que ahora surge es si este dato es un error de medición, o es debida a la variabilidad propia del estudio, o si en efecto la observación es un valor contaminante.

Se empezará por detectar si la muestra se ajusta o no al modelo de distribución Von Mises y en particular detectar la existencia de datos atípicos, por medio de una gráfica de papel probabilístico P-P, donde se muestra la función de distribución empírica acumulada versus la distribución propuesta. Si los datos ajustan bien al modelo propuesto entonces los puntos no se desvían mucho de la recta a 45°, en caso contrario, hay una desviación prominente y para lo cual se concluiría que la muestra no ajusta a la distribución deseada y esto puede ser debida a la observación outlier. Por tanto, a continuación se muestran las gráficas P-P de los datos muestrales, sin la observación atípica y sin los dos datos inconsistentes.







En la gráfica P-P de todos los valores muestrales se observa una desviación prominente en la cola de la distribución lo que advierte una posible presencia de valores atípicos. Al graficar sin el valor ya identificado como outlier, efectivamente los puntos se ajustan mejor a la recta y por tanto a la distribución deseada. Por lo que se puede concluir de primera instancia, que los datos ajustan al modelo de distribución $VM(\mu, k)$ a excepción del valor atípico 147° cuya distribución puede ser $VM(\mu^*, k)$. Por otro lado, al hacer la grafica P-P sin los valores inconsistentes: 147° y 298° se observa todavía un mejor ajuste, sin embargo, es preciso y necesario realizar las pruebas de discordancia correspondientes para verificar si ambos valores son contaminantes.

Se analizará primero si el valor outlier 147° es discordante, por lo que es necesario calcular la dirección y tamaño del vector medio, el parámetro de concentración $\hat{k}$ tanto para la muestra completa, como para la muestra sin dicho valor. Por tanto se tiene que:

Muestra	*Tamaño de* Muestra	*Dirección del vector medio*	*Tamaño del vector medio*	*Paramétro de concentración* $\hat{k}$
Completa	22	3.10°	.829	3.27
Sin Outlier	21	1.33°	.908	5.73

Cabe destacar que el procedimiento para calcular las estadísticas de prueba: C, D, M y L, se llevó a cabo por medio del programa R[7]. Asimismo, se realizó una simulación similar a la que hizo Collet para 5000 muestras, con el fin de generar el cuantil α =.05 de las distribuciones muestrales para las pruebas de discordancia, para este ejemplo en particular, y compararlas con los obtenidos al interpolar en los valores de las tablas dadas por Collet. Por otro lado, es importante señalar que Collet en su artículo no muestra la tabla de cuantiles para la prueba L, por lo que fue de gran utilidad realizar dicha simulación y concluir en esta prueba si el valor atípico era o no discordante.

La siguiente tabla muestra los resultados del valor que toman las estadísticas de prueba al analizar el posible outlier:

Estadística de Prueba	*Valor de la Estadística* C_n	*Cuantil* C_α *Interpolación*	*Cuantil* C_α *Simulación*	*Conclusión de la Prueba*
C	.094	≈.09	.089	Se rechaza H_0
D	.675	≈ .45	.45	Se rechaza H_0
M	.485	≈ .39	.38	Se rechaza H_0
L	7.88		5.8	Se rechaza H_0

Se observa que en los tres casos, como $C_n > C_\alpha$ entonces se rechaza la hipótesis nula, es decir, el dato outlier 147° es en efecto un dato discordante y por tanto debe ser removido de la muestra.

[7] Ver apéndice D.1.

En efecto, tomando en cuenta los cuantiles de cada prueba mostrados en las tablas de Collet e interpolando para $n = 22$ y $k = 3.3$, se observa que estos valores coinciden con los obtenidos en la simulación, lo que verifica la validación de dicha simulación.

Por otro lado al realizar la prueba para evaluar al valor 298° como contaminante, sin la presencia del dato 147°, se tienen los siguientes valores de las pruebas:

Estadística de Prueba	*Valor de la Estadística*	*Conclusión de la Prueba*
C	.026	No se rechaza H_0
D	.122	No se rechaza H_0
M	3.85	No se rechaza H_0
L	.296	No se rechaza H_0

Se concluye que la dirección 298° no es discordante y esto confirma que tal valor es debido a la variabilidad inherente del problema.

Conclusiones

Los valores outliers en la estadística circular se detectan de forma muy diferente que en la estadística tradicional, ya que se utiliza la distancia circular para detectarlos. Las pruebas de discordancia C, D, L y M, detectan valores contaminantes a partir de que éstos son outliers, sin embargo, no necesariamente un valor contaminante es un valor outlier.

Collet en su artículo no muestra el procedimiento para realizar sus simulaciones y encontrar los puntos percentiles de las pruebas, se limita a dar una explicación muy breve y general del proceso, por lo que fue de gran ayuda investigar la forma de desarrollar un programa que replica el experimento Monte Carlo para cualquier valor de k, n y λ deseado.

Es importante señalar el reto computacional que implica el generar las simulaciones, ya que no solo se necesitan conocimientos básicos de programación sino también de los paquetes *Circstats* y *Circular* para el manejo de datos circulares en R. Además el hecho de contar con una computadora con un buen procesador y memoria RAM facilitan el proceso de la información, ya que entre mayor sea el número y tamaño de muestra requerido mayor será el tiempo que se tardará en realizar la simulación. En particular se utilizó una computadora con memoria RAM de 2 gigas y procesador pentium de 1.8 GHz y el tiempo estimado en realizar la simulación que calcula las estadísticas de prueba y genera la función potencia para 3000 muestras con $n = 8$ fue de medio minuto, $n = 15$ de un minuto, $n = 30$ de dos minutos y medio.

Al analizar el desempeño de las pruebas a través de la función potencia generada con una simulación para 3000 muestras, se concluye que las pruebas tienen una mejor potencia mientras el parámetro de concentración y el tamaño de muestra aumentan. Se observa que a partir de que $k > 3$ y $n > 15$, la potencia de las pruebas tienden al mismo valor mientras λ crece de 0 a 1, por lo que el desempeño de las pruebas son similares y no hay diferencias entre escoger una prueba u otra. Sin embargo, se recomienda el uso de las pruebas C y D, particularmente cuando la muestra y la concentración son bajas, ya que sus distribuciones son independendientes del valor de k y tienen más potencia que L y M.

Collet en su investigación recomienda el uso de la estadística M en algunos casos pues sus puntos percentiles son fáciles de determinar, y dice que la prueba D es más fácil de calcular que la C. Sin embargo, gracias al avance computacional de hoy en día, todas las estadísticas estudiadas son fáciles de determinar, lo cual quedó comprobado en la simulación realizada.

Se determinó que las pruebas no son potentes cuando la muestra tiene más de un valor contaminante, por lo que el estudio de nuevas estadísticas para detectar el problema de contaminación múltiple debe ser desarrollado.

Apéndice A

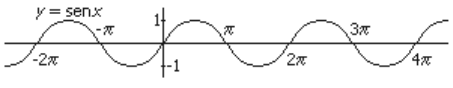
A.1. Funciones trigonométricas

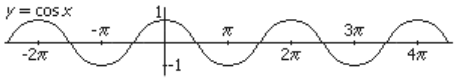
En geometría elemental, un ángulo es sencillamente la unión de dos semirrectas con un punto común inicial. Más útiles para la trigonometría son los "ángulos dirigidos", los cuales pueden ser considerados como pares (l_1, l_2) de semirrectas con el mismo punto inicial.

Ahora bien, Sea P un punto en el plano cartesiano con coordenadas (x, y), al que se le puede asignar un ángulo θ siempre y cuando P se encuentre sobre la circunferencia de un círculo unitario y se elija a l_1 en la mitad positiva del eje horizontal X, entonces, el ángulo dirigido queda totalmente descrito mediante la segunda semirecta l_2, que resulta al rotar la semilínea l_1 (en sentido opuesto a las manecillas del reloj) hasta que ésta pasa por primera vez por el punto P. Por consiguiente, el seno del ángulo se define como y, y el coseno como x, es decir, $y = sen\theta$ y $x = cos\theta$.

No obstante, lo que se desea es definir $senx$ y $cosx$ para cada número x. Para determinar completamente esto, hay que considerar al ángulo en radianes, donde los 360^o de una circunferencia son 2 radianes. Entonces, dado cualquier número x, se elige un punto P sobre el círculo unitario, tal que x es la longitud del arco del círculo que empieza en (1,0) y que se dirige hacia P en sentido contrario al de las agujas del reloj. El ángulo dirigido determinado por P recibe el nombre de "ángulo de x radianes".

Las funciones trigonométricas se obtienen a través de razones trigonométricas, que se relacionan con las razones de los ángulos comprendidos en el intervalo $[0, 2\pi)$ del siguiente modo: si $\theta - \theta' = k * 2\pi$, con k número entero, entonces $sen\theta = sen\theta$', $cos\theta = cos\theta$', $tg\theta = tg\theta$'. Es decir, si dos números difieren por un número entero de 2π, entonces tienen las mismas razones trigonométricas. Se considera que el dominio de θ puede ser cualquier número real, por tanto x y y son funciones periódicas de θ. De este modo se obtienen las funciones trigonométricas *seno* y *coseno*, cuyo período es de 2π y su representación gráficas es:





Las razones trigonométricas $sen\theta$ y $cos\theta$ de un mismo ángulo guardan la siguiente relación fundamental:

$$sen^2\theta + cos^2\theta = 1$$

También se tienen las siguientes relaciones:

$$sen(\theta + \phi) = sen\theta cos\phi + cos\theta sen\phi \qquad sen(\theta - \phi) = sen\theta cos\phi - cos\theta sen\phi$$

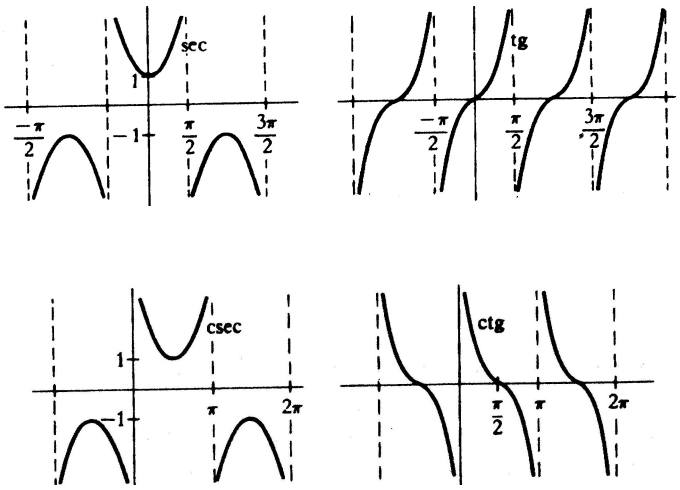
$$cos(\theta + \phi) = cos\theta cos\phi - sen\theta sen\phi \qquad cos(\theta - \phi) = cos\theta cos\phi + sen\theta sen\phi$$

Las demás funciones trigonométricas no presentan absolutamente ninguna dificultad, y se definen como:

$$sec\theta = \frac{1}{cos\theta} \qquad tan\theta = \frac{sen\theta}{cos\theta}$$

$$csc\theta = \frac{1}{sen\theta} \qquad cot\theta = \frac{cos\theta}{sen\theta}$$

Las gráficas respectivas se presentan en la figura siguiente:



Todas las funciones trigonométricas son periódicas: *sen*, *cos*, *sec* y *cosec* tienen período 2π; mientras que *tan* y *cot* tienen período de π. Se observa que para las funciones $sec\theta$ y $tan\theta$ el ángulo θ es diferente de $k\pi + \pi/2$; mientras que para las funciones $csc\theta$ y $cot\theta$, el ángulo es diferente a $k\pi$

En estadística circular se utiliza con mayor frecuencia las funciones seno, coseno y tangente, por lo que a continuación se restringirá el amplio estudio de funciones trigonométricas a éstas funciones.

Las inversas de las funciones trigonométricas se derivan fácilmente y para esto es necesario restringir las funciones trigonométricas primero a intervalos convenientes, ya que no son funciones uno a uno. Los intervalos que generalmente se eligen son:

$$\begin{array}{rcl} -\pi/2 \leq \theta \leq \pi/2 & \text{para} & sen\theta \\ 0 \leq \theta \leq \pi & \text{para} & cos\theta \\ -\pi/2 < \theta < \pi/2 & \text{para} & tan\theta \end{array}$$

Entonces, la función trigonométrica inversa para $y = sen\theta$, equivale a decir: *el ángulo cuyo seno es igua a y*, y se expresa como $\theta = arcseny$ o también como $\theta = sen^{-1}y$. La función $arcsen$ es la función inversa o recíproca de la función. De la misma manera se definen las funciones inversas para $y = cos\theta$ y $y = tan\theta$. De esta forma se definen:

$$\begin{array}{lcl} \theta = arcseny & \text{ó} & \theta = sen^{-1}y \\ \theta = arccosy & \text{ó} & \theta = cos^{-1}y \\ \theta = arctany & \text{ó} & \theta = tan^{-1}y \end{array}$$

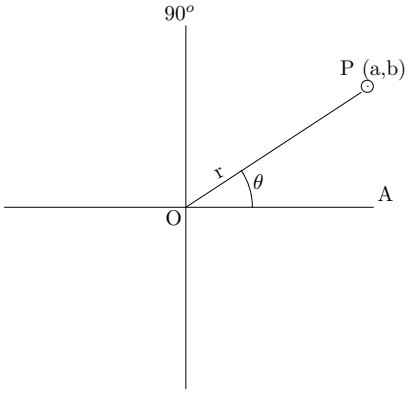
Por otro lado, es importante señalar algunas aproximaciones de θ medido en radianes, cuyos valores son cercanos a cero:

$$\begin{array}{lcl} sen\theta & \approx & \theta \\ 2(1 - cos\theta) & \approx & \theta^2 \\ cos\theta & \approx & 1 - \frac{1}{2}\theta^2 \end{array}$$

A.2. Coordenadas polares

Por medio de un sistema de coordenadas en un plano, es posible localizar cualquier punto en el plano. En el sistema rectangular esto se efectúa al referir el punto a dos rectas fijas perpendicuales llamadas *ejes de coordenadas.*

En el sistema polar, un punto se localiza especificando su posición relativa con respecto a una recta fija y a un punto fijo de esa recta. La recta fija se llama *eje polar* y el punto fijo se llama *polo*. En la siguiente figura se muestra un sistema de coordenadas donde la recta horizontal OA es el eje polar y O el polo. Sea P cualquier punto en el plano coordenado. Si se traza el segmento de recta OP, designando su longitud por r , entonces θ es el ángulo AOP. Así pues, la posición del punto P con relación al eje polar y al polo se determina cuando se conocen r y θ, cantidades comúnmente llamadas *coordenadas polares* del punto P, donde r se llama *radio vector* y θ *ángulo polar*. Por tanto las coordenadas de P se escriben (r, θ).



El ángulo polar se mide como en trigonometría partiendo del eje polar hacia el radio vector. La medida de un ángulo puede expresarse en radianes o en unidades angulares. Así θ tiene un rango entre 0^o y 360^o o bien entre 0 y 2π; donde π el área de un disco de radio 1 y cuyo valor aproximado es de 3.14159. La medida de un ángulo en radianes está definido por $\theta = \frac{s}{r}$, donde θ es el ángulo central que intercepta un arco de longitud s sobre un círculo de radio r. De esta definición se tiene de inmediato la siguiente relación:

$$180^o = \pi \text{ radianes}$$

de donde,

$$1 \text{ radian} = \frac{180^o}{\pi} = 57{,}2958^o \ (aprox)$$

$$1^o = \frac{\pi}{180^o}\text{rad} = {,}017453 \text{ rad} \ (aprox)$$

Entonces se tienen la siguientes relaciones entre ángulos y radianes:

$$\begin{aligned} 360^o &= 2\pi \quad \text{rad} \\ 90^o &= \pi/2 \quad \text{rad} \\ 60^o &= \pi/3 \quad \text{rad} \\ 45^o &= \pi/4 \quad \text{rad} \\ 30^o &= \pi/6 \quad \text{rad} \end{aligned}$$

Es evidente que un par de coordenadas polares (r, θ) determina un solo punto en el plano cartesiano. Sin embargo, al revés no es cierto, porque un punto P con coordenadas (r, θ), está también determinado por los pares de coordenadas $(r, \theta + 2\pi k)$, k entero.

Ahora bien, las coordenadas rectangulares (a, b) de cualquier punto de un plano se pueden transformar a coordenadas polares, por ejemplo, al considerar a r como la distancia del punto $(0, 0)$ al punto (a, b), se tiene:

$$r = \sqrt{a^2 + b^2}$$

Ahora, se definen:

$$sen\theta = \frac{b}{r} = \frac{b}{\sqrt{a^2 + b^2}} \quad \text{y} \quad cos\theta = \frac{a}{r} = \frac{a}{\sqrt{a^2 + b^2}}$$

Por tanto:

$$a = rcos\theta, \qquad b = rsen\theta, \qquad \theta = arctan\frac{b}{a}$$

Como ya se vio, $arctan\frac{b}{a}$ toma valores entre -90º y 90º. Cuando $a > 0$ se producen valores polares en el primer y cuarto cuadrante; y en el segundo y tercer cuadrante cuando $a < 0$. Por consiguiente, es necesario hacer una conversión de coordenadas cartesianas a polares:

$$\theta = \begin{cases} arctan\frac{b}{a} & \text{si} \quad a > 0 \quad \text{y} \quad b > 0 \\ 180^o - arctan\frac{b}{a} & \text{si} \quad a < 0 \quad \text{y} \quad b > 0 \\ 180^o + arctan\frac{b}{a} & \text{si} \quad a < 0 \quad \text{y} \quad b < 0 \\ 360^o - arctan\frac{b}{a} & \text{si} \quad a > 0 \quad \text{y} \quad b < 0 \end{cases}$$

Y en casos excepcionales se tiene:

$$\theta = \begin{cases} 90^o & \text{si} \quad a = 0,\ b > 0 \\ 270^o & \text{si} \quad a = 0,\ b < 0 \\ \text{indeterminado} & \text{si} \quad a = 0,\ b = 0 \end{cases}$$

En particular, se puede seleccionar al punto (a, b) a cualquier distancia del origen. Por muchos propósitos es conveniente seleccionar (a,b) en un círculo de radio 1, y en este caso $sen\theta = b$ y $cos\theta = a$. En consecuencia, por definición, las coordenadas polares de un punto (a, b) en un círculo de radio 1 es: $(cos\theta, sen\theta)$.

Apéndice B

B.1. Prueba de hipótesis

Una prueba de hipótesis es un proceso científico que examina si un supuesto es factible o no. Por esto, se deben inevitablemente examinar dos hipótesis: hipótesis nula e hipótesis alternativa. Se asume que la primera es verdadera siempre y cuando no exista evidencia suficiente para rechazarla, de lo contrario, se considera a la hipótesis alternativa como verdadera, bajo un cierto grado de incertidumbre. Con el fin de llevar a cabo una regla de decisión que permita aceptar o rechazar un supuesto, se deben considerar en cualquier prueba de hipótesis, los siguientes elementos esenciales: hipótesis nula, hipótesis alternativa, estadística de prueba y región de rechazo.

Las partes esenciales de una prueba estadística son la estadística de prueba y la región de rechazo asociada. *La estadística de prueba* es una función de la muestra que sirve como fundamento para la toma de decisiones. *La región de rechazo*, especifica los valores de la estadística de prueba que establecen la veracidad de la hipótesis alternativa. Para determinar un criterio que evalue la decisión de aceptar o no la hipótesis nula, se debe determinar la probabilidad de incurrir en un error. De esta forma, se define al *error tipo I* como la probabilidad de rechazar la hipótesis nula cuando es verdadera, se denota como α, y se le conoce como el tamaño de la prueba o nivel de significancia. El *error tipo II* es la probabilidad de aceptar la hipótesis nula cuando es falsa, se denota como β. Por otro lado $1 - \beta$ mide la probabilidad de cometer un acierto, al rechazar Ho cuando es en efecto falsa y se le conoce como la potencia de la prueba.

	Ho es verdadera	Ho es falsa
Rechazar Ho	Error Tipo I α=Tamaño de la prueba	*Decisión Correcta* 1-β: Potencia de la Prueba
Aceptar Ho	*Decisión Correcta* 1-α: Nivel de confianza	Error Tipo II β

Hay que considerar que el hecho de aceptar o rechazar la hipótesis nula depende del valor que tome la prueba estadística. Es por esto, que surge la pregunta de cuál prueba es preferible a otra ante cualquier situación en particular y cómo se podría evaluar el desarrollo funcional de dichas pruebas.

Hay cuatro aspectos a considerar:

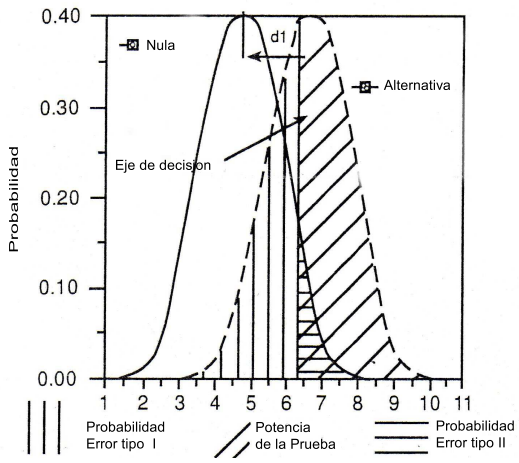
1. El modelo de probabilidad inicial
2. Nivel de significancia
3. Hipótesis alternativa
4. Evaluación del desarrollo de la prueba y la naturaleza del concepto de potencia

Es importante conocer lo que constituye un criterio apropiado para evaluar el funcionamiento de las pruebas de discordancia. Generalmente no hay una única prueba, por eso para escoger la mejor, se necesitan medidas que permitan examinar su funcionamiento. El nivel de significancia, es una medida fundamental, la comparación de las pruebas con el mismo nivel de significancia debe por supuesto depender de la hipótesis nula, ya que de esta manera se puede conocer la distribución de la prueba estadística, o al menos la probabilidad de alguna parte de la cola de la distribución para un valor particular de la prueba estadística. Sin embargo, no siempre en todos los caso es fácil y por tanto la técnica de simulación es requerida.

B.1.1. Potencia de la prueba

Una forma de medir el funcionamiento de una prueba es a través de su función potencia. El término *potencia* se refiere a la habilidad de detectar el hecho de que la hipótesis nula sea falsa. Entonces la *potencia* es la capacidad de una prueba de no cometer el error tipo II. En términos más formales la potencia de la prueba es la probabilidad de que la prueba rechace la hipótesis nula dado que es falsa, y se denota como $1 - \beta$.

Como se podría esperar, la potencia de una prueba depende en parte de qué tan falsa es la hipótesis nula, ya que si ésta se encuentra un poco alejada de la realidad, la prueba no tendrá mucha potencia, de lo contrario, la prueba será potente. Otro factor importante en la potencia de una prueba es el tamaño de la muestra. Entre más grande sea la muestra, más potente es la prueba. Esto es porque entre más grande sea la muestra, más cercana es la estadística al parámetro. El hecho de que el tamaño muestral afecta de cierta forma a la potencia de la prueba, hace posible determinar el tamaño de la muestra necesaria para tener cierto grado de potencia, considerando el riesgo de cometer el error tipo I.



B.2. Pruebas de discordancia: Principio del cociente de verosimilitudes

Como ya se vio en el capítulo concerniente a pruebas de discordancia, existen procedimientos con base intuitiva que permiten establecer pruebas para la detección de observaciones contaminantes . Sin embargo existen otros métodos más formales cuya base parte de un fundamento estadístico y entre ellos esta el *Principio del cociente de verosimilitudes.* Naturalmente la construcción de las pruebas depende en primera instancia de la hipótesis alternativa empleada para explicar los outliers. Para fines de este trabajo se utilizó la hipótesis alternativa de deslizamiento para detectar a través de la estadística de prueba L observaciones discordantes provenientes de una distribución Von Mises. Por esto, es importante recordar la importancia y el significado del principio de verosimilitud con el fin de entender el propósito de este método. Así pues, al considerar la hipótesis A entonces la probabilidad de que la variable aleatoria X tome el valor de x es $p_A(x)$; mientras que al considerar la hipótesis B entonces la probilidad es $p_B(x)$. De esta manera, la observación $X = x$ tiene evidencia y apoyo en la hipótesis A si y solo si $p_A > p_B$ y el cociente de verosimilitudes p_A/p_B ó $p_A - p_B$ mide la fuerza de tal evidencia.

Ahora pues, para entender el procedimiento de este método se explica a continuación un ejemplo de una muestra univariada en donde se desea probar si la observacion x_n un outlier superior.

Suponga que se tiene una muestra aleatoria $X_1, X_2, .., X_n$ proveniente de una función de probabilidad exponencial $\theta e^{-\theta x}$ $(x > 0)$, con θ desconocido. Se desea probar si toda la muestra proviene de la función de distribución exponencial denotada como F o si $n-1$ observaciones provienen de F y la restante, digamos x_n, de la distribución G, con función de densidad $\lambda\theta e^{-\lambda\theta x}$ $(x > 0,\ \lambda < 1)$. Luego entonces, la prueba de hipótesis es la siguiente:

$$H_o : \lambda = 1 \qquad vs \qquad H_a : \lambda < 1$$

La función de verosimilitud bajo la hipótesis nula es:

$$\begin{aligned} f(x_1, x_2, .., x_n; \theta) &= \prod_{i=1}^{n} \theta e^{-\theta x} \\ &= \theta^n e^{-\sum_{i=1}^{n} \theta_i x} \end{aligned}$$

Al obtener el logaritmo natural de la función de verosimilitud se tiene que:

$$lnL_{H_o}(\theta) = nln\theta - n\theta\bar{x}$$

Así pues, se busca el punto que maximiza la función $L_{H_o}(\theta)$ al igualar a cero la siguiente ecuación:

$$\frac{\mathrm{d}\, lnL_{H_o}}{\mathrm{d}\,\theta} = n\frac{1}{\theta} - n\bar{x}$$

Por tanto $L_{H_o}(\theta)$ se maximiza cuando: $\theta = \frac{1}{\bar{x}}$; y el valor máximo que alcanza la función de verosimilitud es:

$$ln\hat{L}_{H_o} = -nln\bar{x} - n$$

Por otro lado, la función de verosimilitud bajo la hipótesis alternativa es:

$$\begin{aligned} f(x_1, x_2, .., x_n; \theta, \lambda) &= \prod_{i=1}^{n-1} \theta e^{-\theta_i x} \ * \ \lambda\theta e^{-\lambda\theta x_n} \\ &= \theta^{n-1} \, e^{-(n-1)\theta\bar{x}} \ * \ \lambda\theta e^{-\lambda\theta x_n} \end{aligned}$$

Al obtener logaritmo natural:

$$lnL_{H_a}(\theta, \lambda) = nln\theta + ln\lambda - (n-1)\theta\bar{x'} - \lambda\theta x_n$$

donde $\bar{x'}$ es la media de $x_1, x_2, .., x_{n-1}$.

Para encontrar el punto donde se maximiza $L_{H_a}(\theta, \lambda)$ entonces:

$$\begin{aligned} \frac{\mathrm{d}lnL_{H_a}}{\mathrm{d}\theta} &= \frac{n}{\theta} - (n-1)\bar{x'} - \lambda x_n \\ \frac{\mathrm{d}lnL_{H_a}}{d\lambda} &= \frac{1}{\lambda} - \theta x_n \end{aligned}$$

Tomando en cuenta que $\lambda < 1$, entonces se iguala a cero y se resolve el sistema de ecuaciones. Por tanto, $L_{H_a}(\theta, \lambda)$ se maximiza cuando:

$$\theta = \frac{1}{\bar{x'}} \qquad y \qquad \lambda = \frac{\bar{x'}}{x_n}$$

Lo cual se cumple cuando $x_n \geq \bar{x'}$. Al sustituir estos valores para λ y θ en la función $lnL_{H_a}(\theta, \lambda)$, este alcanza su valor máximo en:

$$ln\hat{L}_{H_a} = -(n-1)ln\bar{x'} - lnx_n - n$$

Así, la prueba estadística que se basa en el principio del cociente de verosimilitudes es

$$\{ln\hat{L}_{H_a} - ln\hat{L}_{H_a}\}$$

Y para términos de este ejemplo el cociente de verosimilitudes es :

$$-(n-1)ln\bar{x'} - lnx_n + nln\bar{x}$$

$$-(n-1)ln\frac{n-T}{n-1} \, lnT, \qquad T = \frac{x_n}{\bar{x}}$$

Se sigue que la prueba del cociente de verosimilitudes es equivalente a rechazar H_o cuando T es grande.

Apéndice C

C.1. Funciones de Bessel

El astrónomo Friedrich W. Bessel (1784-1846) definió una familia de funciones $J_n(z)$ como:

$$J_n(z) = \frac{1}{2\pi}\int_0^{2\pi} cos(n\theta - zsen\theta)\ \mathrm{d}\theta$$

para el orden $n = 0, \pm 1, \pm 2....$ La variable independiente z puede ser real o compleja.

Las funciones $J_n(z)$ tienen diferentes propiedades:

$$J_n(-z) = (-1)^n J_n(z)$$

Particularmente:

$$\begin{aligned} J_0(-z) &= J_0(z) \\ J_1(-z) &= -J_1(z) \\ J_{n+1}(z) &= \frac{2n}{z}\, J_n(z) - J_{n-1}(z), \quad (z \neq 0) \end{aligned}$$

Entonces, para n=1, se tiene la siguiente fórmula de recursión:

$$J_2(z) = \frac{2}{z} J_1(z) - J_0(z), \quad (z \neq 0)$$

Distintas integrales pueden ser expresadas como Funciones de Bessel, ejemplos:

$$\begin{aligned} \int_0^{2\pi} cos(zsin\theta)\, cosn\theta\ \mathrm{d}\theta &= \begin{cases} 2\pi J_n(z) & \text{si} \quad n = 0, 2, 4, 6.. \\ 0 & \text{si} \quad n = 1, 3, 5, 7.. \end{cases} \\ \int_0^{2\pi} cos(zcos\theta) cosn\theta\ \mathrm{d}\theta &= 2\pi cos\frac{n\pi}{2} J_n(z) \end{aligned}$$

Existe una segunda familia de funciones Bessel igualmente importantes en la estadística circular. Se le conoce como la *función modificada de Bessel del n-ésimo momento*, $I_n(z)$, $n = 0, \pm 1, \pm 2...$, y se definen como:

$$I_n(z) = \frac{1}{2\pi} \int_0^{2\pi} exp(z\, cos\theta) cos\, n\theta \ \mathrm{d}\theta$$

Y se relacionan con la familia $J_n(z)$ por la ecuación:

$$I_n(z) = i^{-n} J_n(iz)$$

donde $i = \sqrt{-1}$ es la unidad imaginaria. Por esta razón, $I_n(z)$, tambien es conocida como la *función de Bessel de argumento imaginario puro.*

$I_n(z)$ tiene dos propiedades muy similares a $J_n(z)$:

$$I_n(-z) = (-1)^n I_n(z)$$
$$I_{n+1}(z) = I_{n-1}(z) - \frac{2n}{z} I_n(z), \quad (z \neq 0)$$

Adicionalmente, se utiliza la propiedad:

$$\frac{\mathrm{d}I_0(z)}{\mathrm{d}z} = I_1(z)$$

Y se destaca que la funcin $I_o(k)$ puede ser expresada como una expansión de series de potencias:

$$I_o(k) = \sum_{r=0}^{\infty} \frac{1}{(r!)^2} \left(\frac{k}{2}\right)^{2r}$$

Para la distribución Von Mises, una función de importancia es la función $A(x)$ definida como :

$$y = A(x) = \frac{I_1(x)}{I_0(x)}, \quad (x \geqslant 0)$$

Esta función es monótona creciente a partir de que el valor $A(0) = 0$ en adelante. Alcanza asintóticamente el valor de 1 mietras x tiende a infinito. Hay que recordar que cuando x es interpretado como el parámetro de concentración k, entonces $A(k)$ es igual a la distancia del vector medio ρ.

C.2. Valores de $y = A_1^{-1}(x)$, $0 \leq x \leq 1$

La siguiente tabla muestra los valores de $y = A_1^{-1}(x)$
donde:
$x = \rho$, es decir x es el tamaño del vector medio de la distribución Von Mises,
$y = k$, es decir, y es el valor del parámetro de concentración de ésta distribución.
o bien,
$x = \hat{\rho}$, es decir, x es el tamaño del vector medio de una muestra aleatoria con función de distribución Von Mises.
$y = \hat{k}$, es decir, y es el estimador máximo verosímil de k

y	$A_1(y)$	y	$A_1(y)$	y	$A_1(y)$	y	$A_1(y)$	y	$A_1(y)$
0.001	0.0005	0.30	0.1483	1.15	0.4970	2.0	0.6978	6.0	0.9124
0.005	0.0025	0.35	0.1724	1.20	0.5128			7.0	0.9255
0.01	0.0050	0.40	0.1961	1.25	0.5280	2.1	0.7135	8.0	0.9352
		0.45	0.2195	1.30	0.5427	2.2	0.7280	9.0	0.9427
0.02	0.0100	0.50	0.2425	1.35	0.5568	2.3	0.7414	10.0	0.9486
0.03	0.0150	0.55	0.2651	1.40	0.5704	2.4	0.7536		
0.04	0.0200	0.60	0.2873	1.45	0.5835	2.5	0.7649	15.0	0.9661
0.05	0.0250	0.65	0.3090	1.50	0.5961	2.6	0.7754	20.0	0.9747
0.06	0.0300	0.70	0.3302	1.55	0.6083	2.7	0.7850		
0.07	0.0350	0.75	0.3509	1.60	0.6199	2.8	0.7939	30.0	0.9832
0.08	0.0400	0.80	0.3711	1.65	0.6311	2.9	0.8021	40.0	0.9874
0.09	0.0450	0.85	0.3907	1.70	0.6418	3.0	0.8096	50.0	0.9899
0.10	0.0499	0.90	0.4098	1.75	0.6521				
		0.95	0.4284	1.80	0.6620	3.5	0.8397	100.0	0.9950
0.15	0.0748	1.00	0.4464	1.85	0.6715	4.0	0.8635		
0.20	0.0995	1.05	0.4638	1.90	0.6806	4.5	0.8803	500.0	0.9990
0.25	0.1240	1.10	0.4807	1.95	0.6894	5.0	0.8934		

Figura C.1: La tabla muestra las soluciones de la ecuación $x = A_1(y)$, donde la función $A_1(y) = I_1(y\ )/\ I_0(y)$ es el cociente de dos funciones modificadas de Bessel.

C.3. Valores de $x = A_1(y)$, $y \geq 0$

La siguiente tabla muestra los valores de $x = A_1(y)$.
donde:
$y = k$, es decir, y el parámetro de concentración de la distribución Von Mises.
$x = \rho$, es decir, x es el valor del tamaño del vector medio de ésta distribución.
o bien,
$y = \hat{k}$, es decir, y es el estimador máximo verosímil del parámetro de concentración de una muestra aleatoria con función de distribución Von Mises.
$x = \hat{\rho}$, es decir, x es el valor del tamaño del vector medio muestral.

x	$A_1^{-1}(x)$	x	$A_1^{-1}(x)$	x	$A_1^{-1}(x)$	x	$A_1^{-1}(x)$	x	$A_1^{-1}(x)$
0.0005	0.001	0.22	0.451	0.47	1.07	0.72	2.14	0.930	7.43
		0.23	0.473	0.48	1.10	0.73	2.21	0.935	7.97
0.001	0.002	0.24	0.495	0.49	1.13	0.74	2.29	0.940	8.61
0.005	0.010	0.25	0.516	0.50	1.16	0.75	2.37	0.945	9.37
0.01	0.020	0.26	0.539	0.51	1.19	0.76	2.46	0.950	10.3
0.02	0.040	0.27	0.561	0.52	1.22	0.77	2.55	0.955	11.4
0.03	0.060	0.28	0.584	0.53	1.26	0.78	2.65	0.960	12.8
0.04	0.080	0.29	0.606	0.54	1.29	0.79	2.76	0.965	14.6
0.05	0.100	0.30	0.629	0.55	1.33	0.80	2.87	0.970	16.9
0.06	0.120	0.31	0.652	0.56	1.36	0.81	3.00	0.975	20.3
0.07	0.140	0.32	0.676	0.57	1.40	0.82	3.15	0.980	25.3
0.08	0.161	0.33	0.700	0.58	1.44	0.83	3.32	0.985	33.6
0.09	0.181	0.34	0.724	0.59	1.48	0.84	3.51	0.990	50.3
0.10	0.201	0.35	0.748	0.60	1.52	0.85	3.74		
0.11	0.221	0.36	0.772	0.61	1.56	0.86	3.91	0.991	55.8
0.12	0.242	0.37	0.797	0.62	1.60	0.87	4.18	0.992	62.7
0.13	0.262	0.38	0.823	0.63	1.65	0.88	4.49	0.993	71.7
0.14	0.283	0.39	0.848	0.64	1.69	0.89	4.86	0.994	83.6
0.15	0.303	0.40	0.874	0.65	1.74	0.90	5.30	0.995	100
0.16	0.324	0.41	0.900	0.66	1.79			0.996	125
0.17	0.345	0.42	0.927	0.67	1.84	0.905	5.56	0.997	167
0.18	0.366	0.43	0.954	0.68	1.90	0.910	5.85	0.998	250
0.19	0.387	0.44	0.982	0.69	1.95	0.915	6.18	0.999	500
0.20	0.408	0.45	1.01	0.70	2.01	0.920	6.54		
0.21	0.430	0.46	1.04	0.71	2.08	0.925	6.95	0.9995	1000

Figura C.2: La tabla muestra los valores del cociente $A_1(y) = I_1(y) \ / \ I_0(y)$, donde $I_1(y)$ y $I_0(y)$ son dos funciones modificadas de Bessel.

C.4. Tabla de cuantiles de la distribución C

La siguiente tabla muestra los valores críticos simulados para $\alpha = .05$ y $\alpha = .01$ de la estadística C, para probar si un outlier es discordante en una muestra aleatoria de tamaño n de la distribución Von Mises.

Tamaño de la Muestra	Valor de k					
	2·0	3·0	4·0	5·0	7·5	10·0
5	0·71	0·38	0·22	0·16	0·094	0·066
	1·05	0·69	0·39	0·27	0·15	0·10
6	0·55	0·33	0·19	0·14	0·081	0·057
	0·78	0·53	0·33	0·22	0·12	0·090
7	0·46	0·28	0·17	0·12	0·072	0·051
	0·62	0·42	0·29	0·18	0·11	0·078
8	0·39	0·25	0·15	0·11	0·065	0·046
	0·51	0·35	0·26	0·16	0·096	0·070
9	0·34	0·23	0·14	0·099	0·059	0·043
	0·43	0·31	0·23	0·15	0·086	0·063
10	0·30	0·20	0·13	0·091	0·054	0·039
	0·37	0·27	0·20	0·14	0·078	0·057
12	0·25	0·17	0·11	0·079	0·046	0·034
	0·29	0·22	0·17	0·12	0·066	0·049
14	0·21	0·15	0·095	0·069	0·040	0·029
	0·24	0·19	0·14	0·10	0·057	0·042
16	0·18	0·13	0·085	0·062	0·036	0·026
	0·21	0·16	0·13	0·091	0·051	0·037
18	0·16	0·12	0·076	0·057	0·032	0·024
	0·18	0·14	0·11	0·081	0·045	0·034
20	0·14	0·11	0·069	0·052	0·030	0·021
	0·16	0·12	0·10	0·073	0·041	0·031
25	0·11	0·086	0·056	0·043	0·024	0·018
	0·12	0·097	0·080	0·059	0·034	0·025
30	0·091	0·072	0·047	0·038	0·021	0·015
	0·10	0·081	0·068	0·050	0·029	0·021
35	0·079	0·062	0·041	0·033	0·018	0·014
	0·086	0·071	0·059	0·046	0·025	0·018
40	0·070	0·055	0·036	0·030	0·016	0·012
	0·075	0·064	0·053	0·043	0·022	0·016

Figura C.3: Los datos que se muestran en la tabla en la parte superior indican el valor de los cuantiles para $\alpha = .05$, para diferentes valores de k. Mientras que los datos inferiores son los cuantiles para $\alpha = .01$.

C.5. Tabla de cuantiles de la distribución D

La siguiente tabla muestra los valores críticos simulados para α =.05 y α = .01 de la estadística D, para probar si un outlier es discordante en una muestra aleatoria de tamaño n de la distribución Von Mises.

Tamaño de la Muestra	Valor de k					
	2·0	3·0	4·0	5·0	7·5	10·0
5	0·69	0·40	0·28	0·22	0·16	0·13
	0·92	0·78	0·43	0·31	0·23	0·18
6	0·70	0·41	0·27	0·22	0·16	0·13
	0·93	0·75	0·45	0·31	0·22	0·18
7	0·73	0·42	0·27	0·22	0·15	0·12
	0·93	0·74	0·46	0·31	0·21	0·17
8	0·75	0·43	0·28	0·21	0·15	0·12
	0·94	0·74	0·47	0·31	0·21	0·17
9	0·77	0·43	0·28	0·21	0·15	0·12
	0·94	0·76	0·48	0·31	0·21	0·17
10	0·78	0·44	0·28	0·21	0·14	0·12
	0·95	0·77	0·48	0·32	0·21	0·16
12	0·79	0·46	0·28	0·21	0·14	0·11
	0·95	0·80	0·49	0·32	0·20	0·16
14	0·80	0·48	0·28	0·21	0·14	0·11
	0·96	0·82	0·50	0·33	0·20	0·16
16	0·81	0·50	0·29	0·21	0·14	0·11
	0·96	0·84	0·51	0·33	0·20	0·16
18	0·82	0·52	0·29	0·21	0·14	0·11
	0·96	0·85	0·52	0·33	0·20	0·16
20	0·83	0·53	0·29	0·21	0·14	0·11
	0·96	0·86	0·54	0·33	0·20	0·16
25	0·85	0·58	0·31	0·21	0·14	0·11
	0·97	0·87	0·57	0·34	0·20	0·16
30	0·88	0·61	0·32	0·22	0·14	0·11
	0·97	0·89	0·60	0·34	0·20	0·16
35	0·91	0·64	0·33	0·22	0·14	0·11
	0·97	0·90	0·62	0·35	0·20	0·16
40	0·93	0·67	0·34	0·22	0·13	0·11
	0·97	0·91	0·65	0·35	0·20	0·16

Figura C.4: Los datos que se muestran en la tabla en la parte superior indican el valor de los cuantiles para α = .05, para diferentes valores de k. Mientras que los datos inferiores son los cuantiles para α = .01.

C.6. Tabla de cuantiles de la distribución M

La siguiente tabla muestra el valor de los cuantiles de la distribución de la estadística M, para diferentes tamaños de la muestra, con $\alpha = .05$ y $\alpha = .01$

	Tamaño de la muestra										
α	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*12*	*15*	*20*
.05	.661	.73	.731	.714	.680	.648	.611	.583	.528	.464	.387
.01	.661	0.75	.774	.776	.763	.743	.708	.683	.629	.564	.474

C.7. Valores simulados de los cuantiles para L, C, D, M

Prueba C	k=1	k=3	k=7
n=8	.8329	.2417	.0692
n=15	.3659	.1394	.0426
n=30	.1517	.0745	.0238

Prueba D	k=1	k=3	k=7
n=8	.8963	.4061	.1554
n=15	.8962	.4687	.1518
n=30	.9008	.5942	.1406

Prueba L	k=1	k=3	k=7
n=8	4.559	5.89	5.31
n=15	3.80	5.58	5.32
n=30	3.29	5.96	5.60

Prueba M	k=1	k=3	k=7
n=8	.6385	.7519	.7301
n=15	.3487	.4978	.4992
n=30	.1660	.3021	.3036

Apéndice D

D.1. Sintaxis del ejemplo: estrellas de mar

```
# Muestra direccional de 22 estrellas de mar al ser alejarlas de
#su habitat natural:

star=c(0,1,3,3,8,13,16,18,30,31,43,45,147,298,329,332,335,
340,350,354,356,357)

# Gráfica de los datos en radianes

plot.circular(rad(star),main="Movimiento de las Estrellas de Mar")

# O bien, para que representarlos en grados:

x=circular(c(star))
plot(rad(x),cex=.9,units="degrees",main="Movimiento de las Estrellas de Mar")

# Muestra sin el valor atípico 147°

star13=c(0,1,3,3,8,13,16,18,30,31,43,45,298,329,332,335,340,
350,354,356,357)

# Muestra sin los valores inconsistentes 147° y 298°

star2=c(0,1,3,3,8,13,16,18,30,31,43,45,329,332,335,340,350,354,356,357)

# Se hace la gráfica P–P donde se grafica la función de distribución
#muestral versus la distribución Von Mises

plot.edf(rad(star))
lines(rad(star),pvonmises(rad(star),0,3),lty=3)

pp.plot(rad(star),main="Gráfica P–P de los valores muestrales",
xlab= "Distribución Von Mises",ylab="Distribución Empírica")
```

```
# Gráfica P–P sin el valor outlier

pp.plot(rad(star13),main="Gráfica P–P sin el valor atípico 147°",
xlab= "Distribución Von Mises",ylab="Distribución Empírica")

# Gráfica P–P sin los dos valores atípicos

pp.plot(rad(star2),main="Gráfica P–P sin los valores 147° y 298°",
xlab= "Distribución Von Mises",ylab="Distribución Empírica")

# Se calculan las componentes del vector resultante:

C=sum(cos(rad(star)))
S=sum(sin(rad(star)))
atan2(S,C)

# Se calcula la dirección del vector medio

medio=circ.mean(rad(star))

# Se calcula la dirección del vector medio y el tamaño de éste

circ.summary(rad(star))

# El tamaño del vector medio es:

rho=circ.summary(rad(star))[3]

# Se obtiene el estimador del parámetro de concentración,


k=A1inv(rho)

# Para rectificar el parámetro de concentración se calcula la
#función A(k), es decir, el tamaño del vector medio:

I.1(3.3)
I.0(3.3)
I.1(3.3)/I.0(3.3)

#Para detectar la observación outlier se calculan las desviaciones
#de los ángulos a la media (por ambos lados)

d1=abs(deg(medio)-star)
d11=360-d1
```

```
# El mínimo de las distancias calculadas anteriormente para
# cada observación i, es la verdadera distancia del ángulo a su media

desv1=c()
for (i in 1:22){
desv1[i]=min(d1[i],d11[i])
}
desv1

# El outlier es aquel que genera la máxima desviación angular con
#respecto a la media

md=max(desv1)

# Rectificando con la fórmula general, la máxima desviación
#generada por el outlier es:

md1=max(pi-abs(pi-rad(d1)))

# Ahora, hay que checar de qué elemento i-ésimo proviene la
#máxima desviación, el cual, se guardará en el vector llamado outi.

obsout=c()
for(i in 1:22){
obsout[i]=if(max(desv1)>desv1[i])(0)else(assign("outi",i))
}
obsout
outi

#Se prosigue a detectar, cuál es el valor de la observación outlier,
#el cual se guardará en el vector llamado out

out1=c()
for(i in 1:22){
out1[i]=if (obsout[i]!=0)(assign("out",star[i]))else(0)
}
out1
out


****La observación outlier es el elemento 13 de la muestra de datos,
que corresponde al valor de 147°. Hay que checar si el outlier: 147°,
es un valor discordante, de acuerdo a las estadísticas de prueba****
```

```
# Estadísticas de prueba para la detección de valores discordantes

#**********Estadística C**********

# Recordando, C = (rk-r)/r, donde r es el tamaño del vector medio
#y rk es el tamaño del vector medio sin la observaciòn k-ésima,
#es decir, sin la observación outlier

# Se debe calcular el tamaño del vector medio sin el valor 147°.
# La muestra sin el valor outlier es:

n=c(star[1:(outi-1)],star[(outi+1):length(star)])

# El vector medio y resultante de la muestra "star13" es

circ.summary(rad(star13))

# Calculando el tamaño del vector medio de la muestra con todas
#las observaciones y el de la muestra sin la observación outlier

r13=(circ.summary(rad(star13))[3])
r=(circ.summary(rad(star))[3])

# La estadística C es:

C=(r13-r)/r

# En las tablas de los cuantiles para la distribución C, con alpha=.05,
#se encuentra el valor de C para n=22 y k=3, pero no para k=3.3,
#entonces se  hace una interpolación lineal

# Interpolación lineal.

#Dados dos puntos  A(x1,x2) y B(y1,y2), se sabe que la ecuación de
#una línea recta es: y-y1= m(x-x1), donde m es la pendiente,
#entonces: y= m(x)-m(x1)+y1

# Da el valor de A
x1=20
y1=.11

# Da el valor de B
x2=25
y2=.086
```

```
# Se calcula la pendiente

m= (y1-y2)/(x1-x2)

# Se calcula el valor de Y al asignar el valor de X
x=22
y= m*x-(m*x1)+y1

# Se tiene que el valor del cuantil C_alpha= .1004, con alpha=.05,
#n=22, k=3, pero como k=3.3, entonces seguramente la C_alpha que
#se busca es menor a .09. Así pues, como  C_{n} > C_alpha, entonces
#se rechaza Ho y la observación  147° es en efecto un
#outlier discordante.

********** Estadística D**********

# La Estadística D se basa en las proporciones de los arcos :
#Di=Ti/Ti-1,
#donde:
#Ti=theta_(i+1)-theta_(i) y  Tn=2pi-theta_(n)+theta_(1)

# Se sabe que la observación k-ésima donde Tk es la máxima, es aquél
#que corresponde al outlier. La estadística D es el min(Dk,D^-1)

Tk=star[14]-star[13]
Tk1=star[13]-star[12]

Dk=Tk/Tk1
Dk1=1/Dk

# La estadística D es es min(dk,dk^-1)

D=min (Dk,Dk1)

# Por tanto D vale .6745

# Para obtener el cuantil de D, se hace interpolación con n=22, k=3.

# El valor de A
x1=20
y1=.53

# El valor de B
x2=25
y2=.58
```

```
m= (y1-y2)/(x1-x2)
x=22
y= m*x-(m*x1)+y1

# Se tiene que el valor del cuantil D_alpha=.55, con alpha=.05,
#n=22,k=3, pero como k=3.3, entonces seguramente la D que se busca
#es menor o igual a .45. Por tanto como  D_{n} > D_alpha se rechaza
# Ho y la observación 147° es en efecto un  outlier discordante.

********** Estadística M*********

#Recordando, la estadística M= (R_{k}-R+1)/n-R, donde R es el tamaño
#del vector resultante, R_{k} el tamaño del vector resultante sin
#el elemento k-ésimo.

# Tamaño del vector resultante para la muestra completa

R=r*length(star)

# Tamaño del vector resultante sin la observación outlier

R13=r13*(length(star)-1)

# El valor de la estadística M es:

M=(R13-R+1)/(length(star)-R)

# El valor de M_{n}= .4851 y M_alpha<.38 con alpha=.05, n=22; entonces,
se rechaza Ho y la observación 147° es en efecto un outlier.

#************ Estadística L************

# Se calcula el parámetro de concentración con outlier

k1=est.kappa(rad(star))

# Se calcula el parámetro de concentración sin outlier

k2=est.kappa(rad(star13))

L=((R13+1)*k2)-(k1*R)-( 22 * ( log( I.0(k2)/I.0(k1) ) ) )

# Éste ejercicio fue ejecutado con el sintaxis de la simulación y
#los valores para las estadísticas coincidieron.
```

D.2. Simulación de los puntos percentiles de C, D, L, M.

```
# Simulación que genera m muestras de tamaño n de la distribución Von Mises
#y que analiza cada muestra para detectar un valor outlier y examinarlo
#a través de pruebas de discordancia: C, D, M y L; por lo que se obtienen
# m valores de dichas estadísticas con las cuales se obtiene la función de
#distribución muestral para cada una de ellas y su cuantil correspondiente
#de alpha=.05

# Consideraciones: El tamaño de muestra debe ser suficientemente grande
#para generar el cuantil que acumule exactamente el 95% de probabilidad.

# Da el tamaño de cada  muestra
n=22

# Da el parámetro de concentración
k=3.3

# Da el número de muestras
m=5000

# A continuación se generan m  muestras de la distribucion Von mises de
tamaño n donde los datos generados estan en radianes

# Se define la función que genera las muestras de la distribución Von Mises

f=function(){
rvm(n,0,k)
}

# Se generan m muestras, donde la muestra j-ésima se guarda en
#la columna j de una matriz

muestra1=matrix(nrow=n, ncol=m)
for (i in 1:n){
for (j in 1:m){
muestra1[,j]=f()
}
}

muestra=matrix(nrow=n,ncol=m)
for (j in 1:m){
muestra[,j]=sort(muestra1[,j])
}

# Se calcula la dirección del vector medio para cada muestra
```

```
v1=c()
for (i in 1:m)
{
v1[i]=circ.mean(muestra[,i])
}

# Como el vector medio está definido por el arctan, es necesario saber
#en qué cuadrante cae el ángulo, pues el arctan tiene un rango de
#-90 a 90 grados.

# Para verificar el valor del ángulo tangente en cualquier cuadrante
#entonces se calcula el vector medio a través de las componentes del
#vector resultante.

c=c()
for (j in 1:m){
c[j]=sum(cos(muestra[,j]))
}

s=c()
for (j in 1:m){
s[j]=sum(sin(muestra[,j]))
}

# El vector medio es:

vm=atan2(s,c)

# Ubicando el cuadrante al cual pertenece el ángulo entonces:

v=c()
for (j in 1:m)
{
v[j]=if(c[j]&s[j]>0)(vm[j]) else
if(c[j]>0 & s[j]<0)(2*pi+vm[j]) else
if(c[j]<0 & s[j]>0)(vm[j]) else
if(c[j]<0 & s[j]<0)(2*pi+vm[j])
}

# Se calcula el tamaño  del  vector medio

tv=c()
for (j in 1:m){
tv[j]=(circ.summary(muestra[,j])[3])
}
```

```
# A continuación se detectará el valor outlier de cada muestra
#a través de la distancia circular para determinar el valor que
#se aleja más del conglomerado  de datos.

# Se calcula la desviación que hay entre la media y cada observación
#de la muestra j-ésima

desvt=matrix(nrow=n, ncol=m)
for(i in 1:n){
for(j in 1:m){
desvt[i,j]=abs(muestra[i,j]-v[j])
}}
desvt

# Se calcula la desviación del otro lado

desvt2=matrix(nrow=n, ncol=m)
for(i in 1:n){
for(j in 1:m){
desvt2[i,j]=abs(2*pi-desvt[i,j])
}}
desvt2

# Rectificando lo anterior, la suma de las desviaciones debe sumar 2*pi

desvt+desvt2

# Ahora hay que identificar la desviación mínima entre ambas
#desviaciones calculadas anteriormente

# En la matriz desvt3 se localiza la desviación mínima entre desvt y
desvt2


desvt3=matrix(nrow=n, ncol=m)

for (i in 1:n)
{
        for (j in 1:m)
        {
                desvt3[i,j]=min(desvt[i,j],desvt2[i,j])
        }
}
```

```
# Se detecta cual es la  máxima desviación de cada muestra

maxdesv=c()
for (i in 1:m){
maxdesv[i]=max(desvt3[,i])
}

# Se detecta al elemento i-ésimo en la muestra que genera
#la máxima desviación

h=c()
obsout=matrix(nrow=n, ncol=m)
for(i in 1:n){
        for (j in 1:m)
{
obsout[i,j]=if(maxdesv[j]>desvt3[i,j])(0)
                  else(i)
if(obsout[i,j]>0)
(h[j]=obsout[i,j])
}
}

# Se prosigue a detectar, cuál es el valor de la observación outlier
#para cada muestra

out2=c()
out1=matrix(nrow=n, ncol=m)
for(i in 1:n){
        for (j in 1:m)
{
out1[i,j]=if (obsout[i,j]!=0)(muestra[i,j])else(0)
if(out1[i,j]>0)
(out2[j]=out1[i,j])
}
}

# Se quita el valor outlier de la muestra para analizar el vector medio
#y tamaño del vector resultante sin el efecto del valor outlier

# Para la primera muestra:

n3=c(muestra [1:(h[1]-1),1], muestra [(h[1]+1):length(muestra[,1]),1])

# Ejemplo para hacer  el if y tener las muestras sin outliers.

r=c()
```

```
for(j in 1:m){
if(h[j]==length(muestra[,j]))
(r[j]=h[j])
else
(r[j]=0)
}

# Por muestra, si hay varios if, entonces la forma de escribir es
#if()(hazlo)else if()(hazlo)else(hazlo)

n5=matrix(nrow=n-1, ncol=m)
for(j in 1:m){
        if(h[j]==length(muestra[,j]))
(n5[,j]=c(muestra[1    :    (h[j]-1),j]))
        else
        if(h[j]==1)
(n5[,j]=c(muestra[(h[j]+1):      length(muestra[,j]),j]))
else
(n5[,j]=c(muestra[1    :    (h[j]-1),j]     ,     muestra[(h[j]+1):
length(muestra[,j]),j]))
}
n5

# Se calcula el tamaño del vector medio sin el outlier

tv2=c()
for (j in 1:m)
{
tv2[j]=est.rho(n5[,j])
}

# Se calcula el tamaño del vector resultante sin el outlier

tv3=c()
for (j in 1:m)
{
tv3[j]=tv2[j]*(n-1)
}

# Se calcula  el tamaño del vector resultante con el outlier

tv4=c()
for (j in 1:m)
{
tv4[j]=est.rho(muestra[,j])*n
}
```

```
# Se calcula  el parámetro de concentración sin el outlier

pc=c()
for (j in 1:m){
pc[j]=est.kappa(n5[,j])
}

# Se calcula el parámetro de concentración con el outlier

pcc=c()
for (j in 1:m)
{
pcc[j]=est.kappa(muestra[,j])
}

#******Estadísticas de  prueba******

#******Estadística C *********

C=c()
for (j in 1:m)
{
C[j]=(tv2[[j]]-tv[[j]])/tv[[j]]
}

# Se ordena la muestra de C, para tener las estadísticas de orden

sc=sort(C)

# Pasos para hacer la  gráfica de la distribución empírica
# Se calcula la probabilidad  muestral acumulada de la muestra

proba=c()

proba[1]=0
for (i in 2:m){
proba[i]=(i-1)/m
}
proba

# Se grafica la función de distribución empírica.

plot(sc,proba,type="s")
plot(ecdf(sc), mai="C.d.f.s")
```

```
# Se identifica el elemento i-ésimo correspondiente al cuantil
# de alpha=.05

cuantil=c()
for (j in 1:m)
{
cuantil[j]=if (proba[j]==.950)
(assign("x",j))else(0)
}
x

# Se identifica el valor del cuantil de alpha=.05 de la estadística C

sc[x]

#******Estadística M *********

M=(tv3-tv4+1)/(n-tv4)

# Se ordena la muestra de M, para tener las estadísticas de orden

sm=sort(M)

# Se grafica la función  de distribución empírica.

plot(sm,proba,type="s")
plot(ecdf(sm), mai="C.d.f.s")

# Se identifica el valor del cuantil de alpha=.05 de la estadística M

sm[x]

#******Estadística L *********

L=((tv3+1)*pc)-(pcc*tv4)-(n* ( log(I.0(pc)/I.0(pcc)) ) )

# Se ordena la muestra de L, para tener las estadísticas de orden

sl=sort(L)

# Se grafica la función  de distribución empírica.

plot(sl,proba,type="s")
plot(ecdf(sl), mai="C.d.f.s")

# Se identifica el valor del cuantil de alpha=.05 de la estadística L
```

```
sl[x]

#******Estadistica D *********

Dk=c()
for(j in 1:m){
        if(h[j]!=n)
(Dk[j]=muestra[(h[j]+1),j]-muestra[h[j],j])
else
(Dk[j]=2*pi-muestra[h[j],j]+muestra[1,j])
}
Dk


Dk1=c()
for(j in 1:m){
if(h[j]!=1)
(Dk1[j]=muestra[h[j],j]-muestra[(h[j]-1),j])
else
(Dk1[j]=2*pi-muestra[n,j]+muestra[1,j])
}
Dk1


D1=Dk/Dk1
D2=1/D1

D=c()
for(j in 1:m){
D[j]=min(D1[j],D2[j])
}
D

# Se ordena la muestra de D, para tener las estadísticas de orden

sd=sort(D)

# Se grafica la función de distribución empírica.

plot(sd,proba,type="s",xlim=range(-0,1))
plot(ecdf(sd), mai="C.d.f.s")

# Se identifica el valor del cuantil de alpha=.05 de la estadística D

sd[x]
```

D.3. Simulación para la evaluación de las pruebas

```
# El siguiente programa genera  muestras que contienen elementos
#discordantes con el fin de calcular la probabilidad de que  una
#observación contaminante outlier sea detectado como discordante.

# Se generan m muestras de tamaño n1 bajo la hipótesis nula,
#y tambien m muestras de tamaño n2 bajo la hipótesis alternativa.
#El objetivo es juntar ambas muestras j-ésimas en una sola de tamaño
#n1+n2 y determinar para cada muestra si existen observaciones
#discordantes a través de las estadísticas C, D, M, L,
#y así analizar la función potencia.

# Da el tamaño de la muestra bajo la hipótesis nula
n1=29

# Da el tamaño de la muestra bajo la hipótesis alternativa, es decir,
#el total de valores discordantes.
n2=1

# Da el parámetro de concentración
k=7

# Da el número de muestras
m=3000

# Da el valor de lambda bajo la hipótesis nula
l1=0

# Da el valor de lambda bajo la hipótesis alternativa
l2=1*pi

#****A continuación se generan muestras de la distribución Von Mises****

# Se generan m muestras de tamaño n1, en radianes

# Se define la función que genera las muestras Von Mises bajo la Ho.

f1=function(){
rvm(n1,l1,k)
}

# Se generan las muestras de tamaño n1 bajo la hipótesis nula

muestra1=matrix(nrow=n1, ncol=m)
for (i in 1:n1){
```

```
for (j in 1:m){
muestra1[,j]=f1()
}}

# Se define la función que genera las muestras Von Mises bajo la
#hipótesis nula

f2=function(){
rvm(n2,l2,k)
}

# Se generan las muestras bajo la hipótesis alternativa de tamaño n2

muestra2=matrix(nrow=n2, ncol=m)
for (i in 1:n2){
for (j in 1:m){
muestra2[,j]=f2()
}}

# Ahora se prosigue a juntar muestra por muestra los valores bajo la
#hipótesis nula y los valores bajo la hipótesis alternativa

# Entonces se tiene una muestra de tamano n1+n2
tm=n1+n2

# Se juntan las muestras

muestra=matrix(nrow=tm, ncol=m)
for (j in 1:m){
muestra[,j]=c(muestra1[,j],muestra2[,j])
}}

# Por muestra se analiza cuál es el valor outlier.

# Se calcula la dirección del vector medio para cada muestra

v1=c()
for (i in 1:m)
{
v1[i]=circ.mean(muestra[,i])
}

#Se verifica el valor del ángulo tangente en cualquier cuadrante:

c=c()
for (j in 1:m){
```

```
c[j]=sum(cos(muestra[,j]))
}

s=c()
for (j in 1:m){
s[j]=sum(sin(muestra[,j]))
}

# El vector medio es:

vm=atan2(s,c)

# Se ubica el cuadrante al cual pertenece el ángulo:

v=c()
for (j in 1:m)
{
v[j]=if(c[j]&s[j]>0)(vm[j]) else
if(c[j]>0 & s[j]<0)(2*pi+vm[j]) else
if(c[j]<0 & s[j]>0)(vm[j]) else
if(c[j]<0 & s[j]<0)(2*pi+vm[j])
}

# Se calcula el tamaño del vector medio

tv=c()
for (j in 1:m)
{
tv[j]=(circ.summary(muestra[,j])[3])
}

# A continuación se detectará el valor outlier en términos
#de la distancia circular entre éste y la media.

# Desviación de la media a la observación

desvt=matrix(nrow=tm, ncol=m)
for(i in 1:tm){
for(j in 1:m){
desvt[i,j]=abs(muestra[i,j]-v[j])
}}

# Desviación del otro lado

desvt2=matrix(nrow=tm, ncol=m)
for(i in 1:tm){
```

```
for(j in 1:m){
desvt2[i,j]=abs(2*pi-desvt[i,j])
}}

# Rectificando lo anterior, la suma de las desviaciones debe sumar 2*pi

desvt+desvt2

# Ahora hay que identificar la desviación mínima entre ambas desviaciones
#calculadas anteriormente

# En desvt3 se busca el mínimo de las matrices desvt y desvt2

desvt3=matrix(nrow=tm, ncol=m)
for (i in 1:tm){
for (j in 1:m){
desvt3[i,j]=min(desvt[i,j],desvt2[i,j])
}}
desvt3

# Se detecta cual es la  máxima desviación de cada muestra

maxdesv=c()
for (i in 1:m){
maxdesv[i]=max(desvt3[,i])
}

# Se detecta el elemento i-ésimo en la muestra que genera la
#máxima desviación

h=c()
obsout=matrix(nrow=tm, ncol=m)
for(i in 1:tm){
        for (j in 1:m)
{
obsout[i,j]=if(maxdesv[j]>desvt3[i,j])(0)
                  else(i)
if(obsout[i,j]>0)
(h[j]=obsout[i,j])
}
}
obsout
h
```

```
#Para identificar si es un outlier discordante pongamos uno, sino cero.

od=c()
for (j in 1:m)
{
od[j]=if(h[j]>=n1)(1)else(0)
}

# Se prosigue a detectar, cuál es el valor de la observación outlier

out2=c()
out1=matrix(nrow=tm, ncol=m)
for(i in 1:tm){
        for (j in 1:m)
{
out1[i,j]=if (obsout[i,j]!=0)(muestra[i,j])else(0)
if(out1[i,j]>0)
(out2[j]=out1[i,j])
}}

# Ahora, hay que quitar el valor outlier de la muestra

n5=matrix(nrow=tm-1, ncol=m)
for(j in 1:m){
        if(h[j]==length(muestra[,j]))
(n5[,j]=c(muestra[1    :    (h[j]-1),j]))
        else
        if(h[j]==1)
(n5[,j]=c(muestra[(h[j]+1):      length(muestra[,j]),j]))
else
(n5[,j]=c(muestra[1:(h[j]-1),j],muestra[(h[j]+1):length(muestra[,j]),j]))
}
n5


# Se calcula el tamaño del vector medio sin el outlier

tv2=c()
for (j in 1:m){
tv2[j]=est.rho(n5[,j])
}

# Se calcula el tamaño del vector resultante sin el outlier

tv3=c()
for (j in 1:m){
```

```
tv3[j]=tv2[j]*(tm-1)
}

# Se calcula  el tamaño del vector resultante con el outlier

tv4=c()
for (j in 1:m){
tv4[j]=est.rho(muestra[,j])*tm
}

# Se calcula  el parámetro de concentración sin el outlier

pc=c()
for (j in 1:m){
pc[j]=est.kappa(n5[,j])
}

# Se calcula el parámetro de concentración con el outlier

pcc=c()
for (j in 1:m){
pcc[j]=est.kappa(muestra[,j])
}

#******Calculo de las estadísticas de prueba******

#******Estadística C *********

C=c()
for (j in 1:m)
{
C[j]=(tv2[[j]]-tv[[j]])/tv[[j]]
}

#******Estadística M *********

M=(tv3-tv4+1)/(tm-tv4)

#******Estadística L *********

L=((tv3+1)*pc)-(pcc*tv4)-(tm* ( log(I.0(pc)/I.0(pcc)) ) )

#******Estadística D *********

# Para calcular D hay que ordenar las muestras, pues se trabaja con las
#estadísticas de orden
```

```
mo=matrix(nrow=tm, ncol=m )
for (j in 1:m){
mo[,j]=sort(muestra[,j])
}

# Para detectar en la muestra ordenada el valor outlier

o=c()
oo=matrix(nrow=tm,ncol=m)
for (i in 1:tm){
for (j in 1:m){
oo[i,j]=if(out2[j]!=mo[i,j])(0)
else(i)

if(oo[i,j]>0)
o[j]=oo[i,j]
}}
oo
o

Dk=c()
for(j in 1:m){
          if(o[j]!=tm)
(Dk[j]=mo[(o[j]+1),j]-mo[o[j],j])
else
(Dk[j]=2*pi-mo[o[j],j]+mo[1,j])
}

Dk1=c()
for(j in 1:m){
if(o[j]!=1)
(Dk1[j]=mo[o[j],j]-mo[(o[j]-1),j])
else
(Dk1[j]=2*pi-mo[tm,j]+mo[1,j])
}

D1=Dk/Dk1
D2=1/D1

D=c()
for(j in 1:m){
D[j]=min(D1[j],D2[j])
}
```

```
#Se debe calcular el cuantil de las pruebas con el correspondiente
#parámetro de concentracion de la muestra, tomando en cuenta lambda
#bajo la hipótesis nula. Esto se hace con el programa de "simulación".

# Ahora el punto es ver si se rechaza o no la hipótesis nula para cada
#prueba, entonces se verá para el valor outlier de cada muestra si
#se detecta como discordante o no

# Estadística C

#lo que tengo en cada vector "d" con valor 1 son los valores outliers
#que detectan las estadísticas como discordantes.

dc=c()
for (i in 1:m){
dc[i]=if(C[i]>.0238)(1)else(0)
}
dc

# Es de interés saber cuales de esos valores que se detectan discordantes
#son realmente discordantes

rc=c()
for(j in 1:m){
if (od[j]==1 & dc[j]==1)
(rc[j]=1)
else (rc[j]=0)
}
rc

#Se desea calcular  la función potencia, al dividir el número de
#observaciones discordantes outliers detectadas como discordantes
#entre el número de discordantes outliers.

p1c=sum(rc)/sum(od)
p3c=sum (rc)/(n2*m)

# Se calculan las demás estadísticas de prueba siguiendio la analogía
#anterior

# Estadística M

dm=c()
for (i in 1:m)
dm[i]=if(M[i]>.3036)(1)else(0)
dm
```

```
rm=c()
for(j in 1:m){
if (od[j]==1 & dm[j]==1)
(rm[j]=1)
else (rm[j]=0)
}
rm

p1m=sum(rm)/sum(od)
p3m=sum (rm)/(n2*m)

# Estadística L

dl=c()
for (i in 1:m)
dl[i]=if(L[i]>5.60)(1)else(0)
dl

rl=c()
for(j in 1:m){
if (od[j]==1 & dl[j]==1)
(rl[j]=1)
else (rl[j]=0)
}
rl

p1l=sum(rl)/sum(od)
p3l=sum (rl)/(n2*m)

# Estadística D

dd=c()
for (i in 1:m)
dd[i]=if(D[i]>.1406)(1)else(0)

rd=c()
for(j in 1:m){
if (od[j]==1 & dd[j]==1)
(rd[j]=1)
else (rd[j]=0)
}
rd

p1d=sum(rd)/sum(od)
p3d=sum (rd)/(n2*m)
```

Bibliografía

[1] Barnett, V. & Lewis, T. *Outliers in Statistical Data.* Second edition. London: Wiley.

[2] Batschelet, E. *Circular Statistics in Biology.* 1981. London: Academic Press.

[3] Collet, D. *Outliers in Circular Data.*1980. Appl. Statist. 29, 50-57 (88)

[4] Conover, W.J. *Practical Nonparametric Statistics.* Third Edition. 1999. Jhonn Wiley & Sons, inc.

[5] Chow, L. *Statistical Significance: Rationale, Validity and Utility.* 1996. Sage Publications.

[6] Fisher, N.I. *Statistical Analysis of Circular Data.* 1995. Cambridge, University Press.

[7] Hogg, R. *Introduction to Mathematical Statistics.* Third edition. 1970.

[8] Jammalamadaka, S.R. *Topics in Circular Statistics.* 2001. World Scientific.

[9] Lehmann, C. *Geometría Analítica.* 1990. Editorial Limusa.

[10] Lehmann, E.L. & Romano P.J. *Testing Statistical Hypotheses.* Third edition. 2005. Springer.

[11] Mardia, K.V. *Statistics of Directional Data.* 1975. London: Academic Press.

[12] Mardia, K.V. & Jupp P. E. *Directional Statistics.* 2000. John Wiley and Sons, LTD.

[13] Rizzo, M. *Statistical Computing with R.* 2008. Chapman& Hall/CRC.

[14] Royall, R. *Statistical Evidence: A likelihood paradigm.* 1997. Chapman&Hall.

[15] Spivak, M. *Calculo Infinitesimal.* 2000. Editorial Reverté, S.A.

[16] Verzani, J. *Using R for Introductory Statistics.* 2005. Chapman& Hall/CRC.

[17] Wackerly, D. & Mandenhall, W. & Scheaffer, R. *Estadística Matemática con Aplicaciones.* Sexta edición. 2002. Thomson.