UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

FACULTAD DE CIENCIAS

CONSTRUCCIÓN DE REGIONES DE CONFIANZA PARA DATOS DIRECCIONALES

T E S I S

QUE PARA OBTENER EL TÍTULO DE:

ACTUARIO

P R E S E N T A :

JUAN PABLO HERNÁNDEZ ROMERO

TUTORA:

MAT. MARGARITA ELVIRA CHÁVEZ CANO

2007

Agradecimiento

Aunque nuestra vida comparada con la del universo es tan solo un parpadeo, en ese tiempo vivimos, crecemos, aprendemos, sentimos, lloramos y amamos, sin estar destinados a una vida solitaria, ya que al menos tenemos siempre a una persona a nuestro lado, aunque a veces lleguemos a dudarlo.

Estoy en deuda con Araceli Romero Oropeza que es mi madre, que aunque no ha tenido una vida llena de gracia y encanto, ha entregado todo lo que ha tenido, así como su vida misma, para darnos una vida placentera a mis hermanos y a mi. Mi madre posee el valor, la fuerza y la esperanza necesaria para realizar sus propósitos, las cuales a mi parecer son los que me han procurado tener una vida feliz.

Asimismo quiero decir que mi vida no sería la misma sin mi hermano Abraham y mi hermana Araceli; estoy tan acostumbrado a su compañía que la verdad no se que sería de mi sin ellos y a pesar del choque de ideas (típicas entre los hermanos) se que estaremos juntos, contando uno con el otro como hermanos.

También quiero agradecer a mi tío Antonio Israel Romero Oropeza, porque no siendo Yo su responsabilidad, ha ayudado a mi madre y se ha hecho cargo de mi como un padre, y se que tanto él como mi tía Carmen Rivera (su esposa) siempre van a estar ahí para mi como padre y madre.

Otros miembros de mi familia que no puedo dejar de mencionar son a mi tía Alma, mi tía Vero y mi tía Yesi, también me han ayudado y he aprendido varias cosas de ellas, y a pesar de algunos apuros, se que puedo contar con ellas.

Yo considero que las ideas y actos de una persona están orientadas por las enseñanzas de la propia familia y por el medio en el que se desarrolla, y no saben la suerte que he tenido, a lo largo de mi vida he conocido a mucha gente buena y de la cual he podido aprender cosas que me han ayudado a crecer.

A mis amigos y amigas doy gracias, porque aprendí tantas cosas de ellos y por eso son parte de mi vida de alguna u otra forma, ya que por suerte nuestras vidas se cruzaron en el mejor momento y pudimos crecer, compartir, soñar, llorar, y vivir.

Quiero dar las gracias a todos los profesores que he tenido a lo largo de mi vida, desde el preescolar hasta la universidad, sin dejar de mencionar a mis maestros de taekwondo; ya que gracias a todos ellos tengo lo mejor que posee el ser humano: el conocimiento. Sin los conocimientos que me han dejado mis profesores yo creo que no pude haber llegado hasta aquí y no sería quien soy ahora.

Gracias a mi tutora de tesis Margarita Elvira Chávez Cano por su comprensión y su tiempo durante la realización de este trabajo, así como también a los sinodales que también tuvieron la atención y tiempo para que este trabajo fuera de la mejor calidad posible.

Por último, le doy gracias a Dios y a la vida misma por permitirme vivir en esta época y haber conocido a todas las personas que he mencionado y a muchas más.

"Para las cosas grandes y fuertes se necesita
combinación sosegada, voluntad decidida, acción vigorosa,
cabeza de hielo, corazón de fuego y mano de hierro"
Jaime Balmes

Índice

Resumen

En la Estadística existe un área de investigación a la que se le denomina Estadística Direccional y más comúnmente llamada Estadística Circular, la cual se caracteriza por tener como espacio muestral una esfera unitaria de dimensión p, y los datos que se encuentran en este espacio muestral se les conoce como datos direccionales o circulares.

Los cálculos y métodos utilizados en la Estadística Circular son similares a los empleados en la Estadística Usual, están adecuados al espacio muestral de los datos direccionales para poder realizar inferencias sobre los mismos.

La media, varianza, desviación estándar y otros parámetros también son de interés para la Estadística Circular, así como todo el análisis que como cualquier parámetro de interés conlleva.

Para la Estadística Usual uno de los principales parámetros de interés es la "media", en la Estadística Circular no deja de ser la excepción y en este espacio el parámetro de interés se le conoce como la "dirección media". Por tal motivo es que se proponen métodos paramétricos y no-paramétricos para la construcción de regiones de confianza para la dirección media, tales métodos basados en una distribución hipotética (de los datos direccionales) proveniente de utilizar el Método Bootstrap para la determinación de la región de confianza.

El Método Bootstrap es un algoritmo que realiza re-muestreos sobre la misma muestra, es decir, toma a la muestra como si fuera la población bajo estudio. El repetir el Método Bootstrap un gran número de veces (miles de veces) hará que la distribución hipotética se aproxime a la distribución real de los datos direccionales.

Los métodos para la construcción de regiones de confianza para datos direccionales utilizados son el Método Gráfico, Método Básico, Método Dúchame, Método Dúchame Pivotal, Método General, Método Pivotal con Vectores Ortogonales y Método Empírico, todos ellos utilizando el Método Bootstrap para determinar la mejor región de confianza para la dirección media de datos direccionales.

Introducción

La *Estadística Circular* se relaciona con observaciones que se representan con vectores unitarios en el plano, o en el espacio tridimensional, a los que llamamos datos direccionales. El espacio muestral es principalmente una esfera unitaria de dimensión p, en caso de que la esfera sea de dimensión $p=2$ el espacio muestral es un círculo unitario, en el caso de que la esfera sea de dimensión $p=3$ el espacio muestral es una esfera en el espacio tridimensional, y así sucesivamente; pero para los cuales constantemente los métodos univariados o multivariados, que se han desarrollado en *Estadística Usual*, no pueden ser utilizados directamente en el análisis de los datos direccionales. Por lo anterior se necesitan métodos para el análisis de datos direccionales que sean adecuados a la estructura del espacio muestral al que pertenecen.

Los datos direccionales se pueden encontrar usualmente en las ciencias naturales tales como astronomía, biología, ecología, geología, medicina, etc.

Muchos métodos que suelen utilizarse, para realizar inferencias sobre datos direccionales, están basados en *Estadística No-Paramétrica* y recurren a la aplicación del *Método Bootstrap* como una herramienta principal.

El *Método Bootstrap* es un procedimiento auxiliar que consiste en la realización de un gran número de remuestreos sobre la muestra original de datos direccionales, es decir, toma a la muestra original como el espacio muestral para tomar de ahí el remuestreo, y se realiza el análisis de interés con base en los remuestreos obtenidos.

Para el desarrollo de procedimientos estadísticos para el análisis de conjuntos de datos direccionales, se plantean procedimientos para la construcción de *Regiones de Confianza* utilizando el método bootstrap en métodos no-paramétricos; así como en métodos paramétricos que utilizan cantidades pivotales que también involucran la utilización del método bootstrap.

Una razón para usar métodos pivotales para la construcción de regiones de confianza por bootstrap en lugar de determinarlos de manera no-paramétrica, es porque en particular un método pivotal da regiones de confianza con un mayor nivel de confianza,

es decir, que para el nivel de confianza $(1-\alpha)$ con $\alpha > 0$ y tomando en cuenta las regiones obtenidas al realizar varios muestreos, el $(1-\alpha)\times 100\%$ de esas regiones contienen el valor real del parámetro desconocido y cuyo error tiende a ser conservativo.

El hecho de que un procedimiento (cualquiera que este sea) utilice el método bootstrap conlleva utilizar recursos computacionales para el análisis que se esté realizando, ya que el procedimiento bootstrap se puede considerar como un tipo de simulación (de hecho un conjunto de simulaciones), y la precisión del resultado obtenido queda determinada por el recurso computacional (tecnología) con que se cuente y se esté utilizando para este fin.

En el capítulo uno, para comprender más de qué se trata la estadística circular y cómo es necesario el desarrollo de métodos para los datos direccionales, se presenta una recopilación de conceptos básicos de estadística circular.

En el capítulo dos se describe con detalle el método bootstrap, para dar una idea más clara del cómo es que trabaja este método, del porqué es considerado como una simulación, y del porqué es una herramienta útil en casos en los que ocurre que las inferencias tradicionales no son aplicables.

En el capítulo tres se presentan métodos estadísticos paramétricos y no-paramétricos para el desarrollo de la construcción de regiones de confianza para datos direccionales.

Por último, en el capítulo cuatro se expone un ejemplo que nos traslada al cuestionamiento de los diferentes métodos de construcción de regiones de confianza, y se expresan aspectos numéricos que se deben de tomar en cuenta en la utilización de los recursos computacionales.

Capítulo 1. Estadística Circular

En varias áreas de la investigación científica se necesita la aplicación de métodos estadísticos para conjuntos de datos donde alguna de sus variables posee características circulares, es decir, donde el rango de valores que puede tomar la variable está dado en ángulos y puede ser representado en un círculo.

Los *datos direccionales* (observaciones circulares o datos circulares) en dos y tres dimensiones son frecuentemente encontrados en las ciencias naturales como astronomía, biología, geología, medicina y meteorología, tales como en la investigación del origen de los cometas, resolviendo problemas de navegación de pájaros, analizando direcciones del viento, etc. Por ejemplo los eventos periódicos pueden ser representados sobre un círculo donde la circunferencia corresponde al periodo como las horas en el día, días de la semana, etc.

Una observación circular puede ser representada como un punto en un círculo de radio ρ, o bien, por un vector (dirección) en el plano. Una dirección y orientación inicial en el círculo se deben de escoger para que cada observación circular pueda ser ubicada por el ángulo que forme respecto a la dirección inicial, así que la representación más simple de los datos direccionales es graficar cada observación como un punto en el círculo unitario en forma desagrupada, es decir, como un ángulo, y del mismo modo se pueden graficar los datos direccionales en forma agrupada.

Un grupo de datos direccionales puede ser representado por un *histograma circular* que es análogo al histograma en la línea real, donde cada barra en el histograma circular está centrada en el punto medio del correspondiente grupo de ángulos, y el área de la barra es proporcional a la frecuencia en ese grupo. Para convertir el histograma circular en un histograma lineal, se corta el histograma circular en un punto sustituible en el círculo y después se desenrolla el histograma circular en el histograma lineal en un intervalo que represente y contenga los 360 grados.

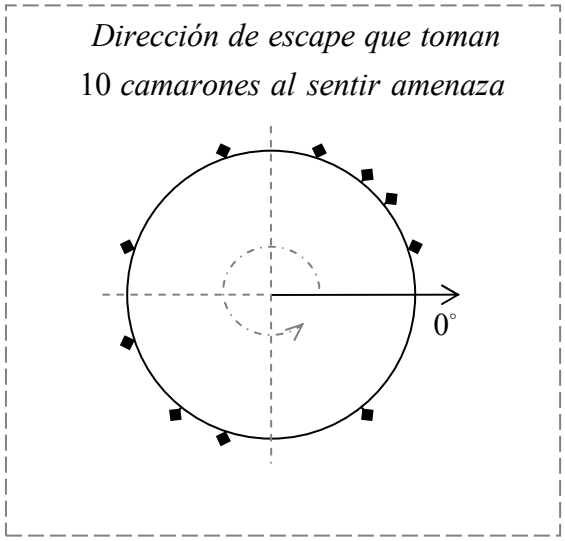

Figura 1.1: Ejemplo de datos direccionales

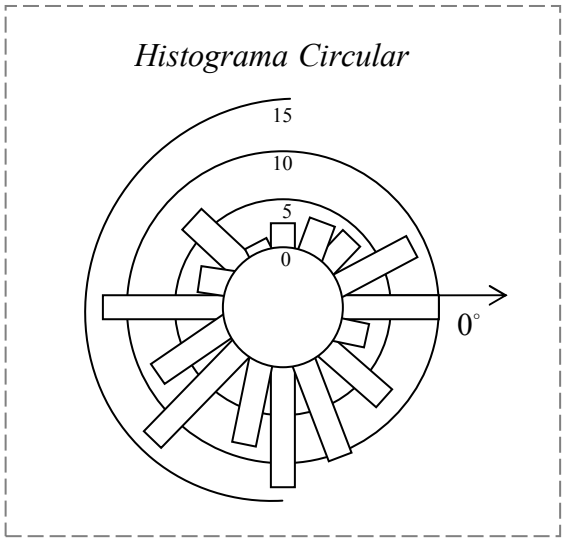

Figura 1.2: Ejemplo de un histograma circular

Entre los datos direccionales hay un tipo de datos que se definen como *datos axiales*. Los datos axiales están dados como observaciones en el círculo donde para cada dirección dada se considera su equivalente en la dirección opuesta, es decir, los ángulos $\alpha \ y \ \alpha+180°$ son equivalentes. Los datos axiales suelen tratarse convirtiéndolos a datos direccionales al duplicar los ángulos, es decir, tomar $\alpha \ como \ 2\alpha$ y así remover la imprecisión en la dirección.

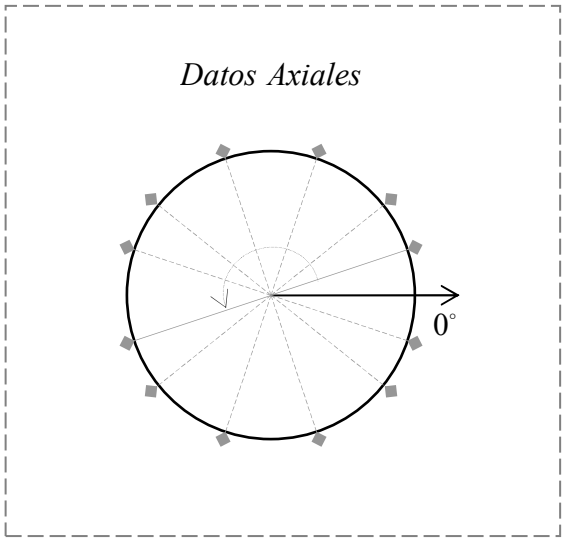

Figura 1.3: Ejemplo de datos direccionales axiales

Escogiendo una dirección y una orientación inicial en el círculo de radio ρ igualmente se está distinguiendo un sistema ortogonal coordenado, así que la posición de un dato direccional x tiene una representación única en un sistema coordenado (x , y) de dos dimensiones.

En coordenadas polares (r,θ), donde r es la distancia del punto al origen y θ es la dirección, se obtiene la representación en ángulo del mismo dato direccional x .

La transformación de coordenadas polares a rectangulares se realiza haciendo $x = r\cos\theta$ y $y = r\text{sen}\theta$, teniendo así:

$$\text{x} = (\text{r}\cos\theta,\text{rsen}\theta)$$

En el análisis de datos direccionales lo que interesa más es la dirección y no la magnitud del vector, por lo cual por convención se utiliza el círculo unitario (círculo de radio uno con centro en el origen) para localizar los datos direccionales.

Por lo tanto, cualquier dato direccional x puede ser representado como (r,θ), el cual en términos de coordenadas rectangulares $\mathrm{x}=(x,y)^T$, al considerar $r=1$, queda expresado como:

$$\mathrm{x}=(\cos\theta,\operatorname{sen}\theta)^T$$

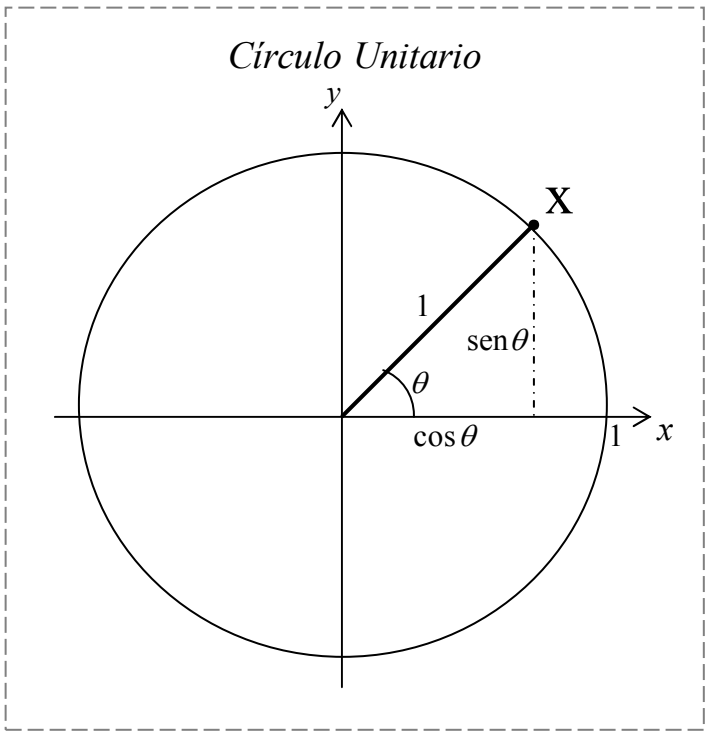

Figura 1.4: Coordenadas Polares

1.1 Dirección Media

Sea $\theta_1, \theta_2, \ldots, \theta_n$ una m.a. de datos direccionales, dados en forma de ángulos, de una población Θ y sea:

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

entonces

$$R = \left\| R^* \right\| = \sqrt{C^2 + S^2}$$

y donde

$$\overline{C} = \frac{1}{n} \sum_{i=1}^{n} \cos\theta_i \quad \text{y} \quad \overline{S} = \frac{1}{n} \sum_{i=1}^{n} \operatorname{sen}\theta_i$$

Teniendo en cuenta que el estimador $\overline{\theta}$ que representa la dirección media de la población esta dada en forma de ángulos, entonces su forma rectangular está compuesta por

$$\cos\overline{\theta} = \frac{C}{R} \quad \text{y} \quad \operatorname{sen}\overline{\theta} = \frac{S}{R} \quad ,$$

entonces

$$C = R\cos\overline{\theta} \quad \text{y} \quad \text{S=}R\operatorname{sen}\overline{\theta}$$

y multiplicando por $\frac{1}{n}$ se tiene que:

$$\overline{C} = \overline{R}\cos\overline{\theta} \quad \text{y} \quad \overline{S} = \overline{R}\operatorname{sen}\overline{\theta}$$

donde $\overline{R}$ es la *longitud media* de las $\theta_1, \theta_2, \ldots, \theta_n$ y que está definida por $\overline{R} = (\overline{C}^2 + \overline{S}^2)^{1/2}$, y dado que $\overline{\theta}$ es la solución para las ecuaciones

$$\overline{C} = \overline{R}\cos\overline{\theta} \quad \text{y} \quad \overline{S} = \overline{R}\operatorname{sen}\overline{\theta}$$

por lo tanto la dirección media es

$$\overline{\theta} = \left(\cos\overline{\theta}, \operatorname{sen}\overline{\theta} \right)$$
$$= \left(\frac{\overline{\text{C}}}{\overline{\text{R}}}, \frac{\overline{S}}{\overline{R}} \right)$$

y para $\overline{\text{R}}$=0 dirección media $\overline{\theta}$ no está definida.

Una definición de $\bar{\theta}$ más explícita está dada por:

$$\bar{\theta} = \begin{cases} \arctan\left(\dfrac{S}{C}\right) & , \ si\ C > 0 \\ \dfrac{\pi}{2} & , \ si\ C = 0\ y\ S > 0 \\ \arctan\left(\dfrac{S}{C}\right) + \pi & , \ si\ C < 0 \\ \arctan\left(\dfrac{S}{C}\right) + 2\pi & , \ si\ C \geq 0\ y\ S < 0 \\ \text{indefinida} & , \ si\ C = 0\ y\ S = 0 \end{cases}$$

y en el contexto de estadística circular $\bar{\theta}$ no es la media tradicional $\frac{1}{n}\sum_{i=1}^{n}\theta_i$, ya que depende del lugar donde se haya definido la dirección inicial.

Otra forma para calcular la dirección media $\bar{\theta}$ para las $\theta_1, \theta_2, \ldots, \theta_n$ es utilizar un método robusto, el cual consiste en que $\bar{\theta}$ es cualquier ángulo ϕ que cumpla:

1. Al menos la mitad de los puntos $\theta_1, \theta_2, \ldots, \theta_n$ estén contenidos en el arco $[\ \phi\ ,\ \phi + \pi\ )$.
2. La mayoría de los puntos en el intervalo $[\ \phi\ ,\ \phi + \pi\ )$ estén más cerca de ϕ que de $\phi + \pi$
3. Cuando el número de puntos contenidos en el intervalo es impar, la media direccional puede ser uno de los datos direccionales, y si es par entonces es un punto medio de dos puntos adyacentes.

Ahora bien, considerando una nueva dirección inicial, haciendo un ángulo γ respecto a la dirección inicial original, entonces $\theta_1, \theta_2, \ldots, \theta_n$ corresponden a los ángulos $\theta_i' = \theta_i - \gamma \quad i = 1, \ldots n$ en un nuevo sistema coordenado (un sistema coordenado original rotado γ grados), y lo que en realidad está ocurriendo es una rotación general de todas las $\theta_1, \theta_2, \ldots, \theta_n$ γ grados. Entonces la longitud media $\bar{\mathbf{R}}' = \bar{\mathbf{R}}$, lo que significa que la longitud media es invariante bajo la rotación, y por lo cual la nueva dirección media es $\bar{\theta}' = \bar{\theta} - \gamma$, es decir, la dirección media es equivariante bajo la rotación.

1.2 Medidas de Concentración

1.2.1 Longitud Media

La *longitud media* $\overline{\mathrm{R}}$ está dada por $\overline{R} = (\overline{C}^2 + \overline{S}^2)^{1/2}$, y al estar trabajando en el círculo unitario entonces $0 \leq \overline{\mathrm{R}} \leq 1$.

Si las direcciones $\theta_1, \theta_2, \ldots, \theta_n$ están cercanas entonces $\overline{\mathrm{R}}$ tenderá a 1, mientras que si están alejadas $\overline{\mathrm{R}}$ tenderá a 0, y es por eso que se toma a $\overline{\mathrm{R}}$ como una medida de concentración (dispersión) para los datos direccionales $\theta_1, \theta_2, \ldots, \theta_n$.

El que $\overline{\mathrm{R}} \simeq 0$ no significa que las $\theta_1, \theta_2, \ldots, \theta_n$ estén dispersas por todo el círculo.

1.2.2 Varianza Circular

La longitud media $\overline{\mathrm{R}}$ es la más importante medida de dispersión, pero para propósitos de inferencia y descripción es necesario tener una medida de dispersión que permita obtener más información acerca de la muestra. Por lo anterior se utiliza la *varianza circular* que se define como:

$$\mathrm{V} = 1 - \overline{\mathrm{R}} \quad , 0 \leq \mathrm{V} \leq 1$$

1.2.3 Desviación Estándar Circular

La *desviación estándar circular* se obtiene de la transformación de la varianza circular V y está dada por:

$$\begin{aligned} v &= \{-2\log(1-\mathrm{V})\}^{1/2} \\ &= \{-2\log \overline{R}\}^{1/2} \end{aligned}$$

y v toma valores entre $(0, \infty)$, mientras que V toma valores entre $[0,1]$. Para V muy pequeña, la expresión anterior se reduce a:

$$\begin{aligned} v &\simeq (2\mathrm{V})^{1/2} \\ &= \{2(1-\overline{R})\}^{1/2} \end{aligned}$$

1.2.4 Dispersión Circular Muestral

La *dispersión circular muestral* $\hat{\delta}$ se define como:

$$\hat{\delta} = \frac{1 - \bar{R}_2}{2\bar{R}_2}$$

donde $\bar{R}_2$ es la longitud media de los datos direccionales duplicados, es decir, es la longitud media de $2\theta_1, 2\theta_2, \ldots, 2\theta_n$.

1.2.5 Diferencia Media Circular

La *diferencia media circular* es:

$$\bar{D}_0 = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \left\{ \pi - | (\pi - | \theta_i - \theta_j |) | \right\}$$

es decir, la distancia media entre pares de datos direccionales.

1.2.6 Rango Circular

El *rango circular* es la longitud del arco más chico que contiene a todas las observaciones (datos direccionales).

Para calcular el rango circular se corta el círculo en la dirección inicial y se considera a las $\theta_1, \theta_2, \ldots, \theta_n$ en el rango $0 \le \theta_i \le 2\pi$. Se ordenan las $\theta_1, \theta_2, \ldots, \theta_n$ para obtener las estadísticas de orden $\theta_{(1)}, \theta_{(2)}, \ldots, \theta_{(n)}$.

La longitud de arco entre observaciones adyacentes es

$$T_i = \theta_{(i+1)} - \theta_{(i)} \qquad , \quad i = 1, \ldots, n-1$$

$$T_n = 2\pi - \theta_{(n)} + \theta_{(1)}$$

entonces el rango circular w es:

$$w = 2\pi - \max\{T_1, \ldots, T_n\}$$

1.3 Distribuciones de Probabilidad Circulares

Una distribución circular es aquélla cuya probabilidad total está concentrada sobre la circunferencia de un círculo unitario. El rango (rv) de una variable aleatoria circular θ, medida en radianes, toma valores en $[0, 2\pi)$ ó $[-\pi, \pi)$.

De la misma forma que las distribuciones clásicas de probabilidad, las distribuciones circulares son básicamente de dos tipos:

I. Discretas

II. Continuas

y como cualquier función de densidad de probabilidad (fdp) $f(\theta)$ debe de cumplir las siguientes propiedades:

1) $f(\theta) \geq 0$

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

y una adicional que nos indica la periodicidad de f

3) $f(\theta) = f(\theta + k \cdot 2\pi) \quad \forall k \in \mathbb{Z}$

Una distribución circular algunas veces se puede generar a partir de distribuciones de probabilidad conocidas sobre la recta real o sobre el plano, por medio de diferentes métodos. Algunos de ellos son:

1.- Envolviendo una distribución lineal alrededor del círculo unitario (Wrapping).

2.- Transformando una variable aleatoria lineal bivariada a sus componentes direccionales, las cuales son llamadas distribuciones de desplazamiento (Off Set).

3.- Iniciando una distribución sobre la recta, se aplica una función inyectiva que haga corresponder a cada punto $x \in \mathbb{R}$ un punto θ del círculo unitario, y se ajustan los valores $-\infty$ e ∞ haciéndolos corresponder con 2π.

1.3.1 Distribución Uniforme

Cuando la probabilidad total es extendida uniformemente sobre la circunferencia se obtiene la *distribución circular uniforme* con función de densidad constante:

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

La distribución circular uniforme tiene un papel importante en el análisis de datos circulares porque representa la ausencia de dirección media. Cuando un conjunto de datos circulares no se ajusta a una distribución uniforme se puede considerar la presencia de una o más direcciones medias.

1.3.2 Distribución Von Mises

Una variable aleatoria θ se dice que sigue una distribución *Von Mises* si tiene una función de densidad:

$$f(\theta; \mu, k) = \frac{1}{2\pi I_0(k)} e^{k\cos(\theta-\mu)} \quad , \quad 0 \le \theta \le 2\pi$$

donde $0 \le \mu < 2\pi$ y $k \ge 0$ son los parámetros. El término $I_0(k)$ de la constante de normalización es la función modificada de Bessel de primera clase de orden cero y está dada por:

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

La función de densidad Von Mises tiene las siguientes propiedades:

i. *Simetría*: Debido a la simetría de la función coseno, la densidad es simétrica alrededor de la dirección μ.

ii. *Moda en* μ: Dado que la función coseno tiene máximo en cero, la densidad de Von Mises tiene máximo en $\theta = \mu$,

es decir, μ es la moda direccional cuyo valor máximo es:

$$f(\mu) = \frac{e^{k}}{2\pi I_0(k)}$$

iii. *Antimoda en* $\mu \pm \pi$: Ya que $\cos\pi = -1$ es el valor mínimo, entonces $\theta = \mu \pm \pi$ da el valor mínimo en:

$$f(\mu \pm \pi) = \frac{e^{-k}}{2\pi I_0(k)}$$

así que $\mu \pm \pi$ es la dirección antinormal.

iv. *Parámetro de concentración* k*:*

Al dividir $f(\mu) = \frac{e^{k}}{2\pi I_0(k)}$ entre $f(\mu \pm \pi) = \frac{e^{-k}}{2\pi I_0(k)}$ se obtiene:

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

así que a medida que k aumenta, la razón $\frac{f(\mu)}{f(\mu \pm \pi)}$ es más grande, indicando mayor concentración alrededor de la media direccional poblacional μ. Por lo anterior se define a k como el parámetro de concentración alrededor de la media direccional.

1.4 Datos direccionales de dimensión p

Los datos direccionales no solo se encuentran en dos dimensiones, sino también en tres y p dimensiones, haciéndolos corresponder de acuerdo a su respectiva S^p := hiperesfera unitaria de dimensión p.

Al igual que en el estudio de datos direccionales en el círculo, el objeto de estudio de la estadística circular es la dirección y la magnitud de las observaciones, siendo la dirección el de mayor interés.

Entonces, se pueden ubicar datos direccionales de dimensión $p = 2$ en el círculo unitario, a los datos direccionales de dimensión $p = 3$ en la esfera unitaria, pero en más dimensiones no es posible visualizarlos aunque se entiende que deben de encontrarse en la hiperesfera correspondiente en dimensión p.

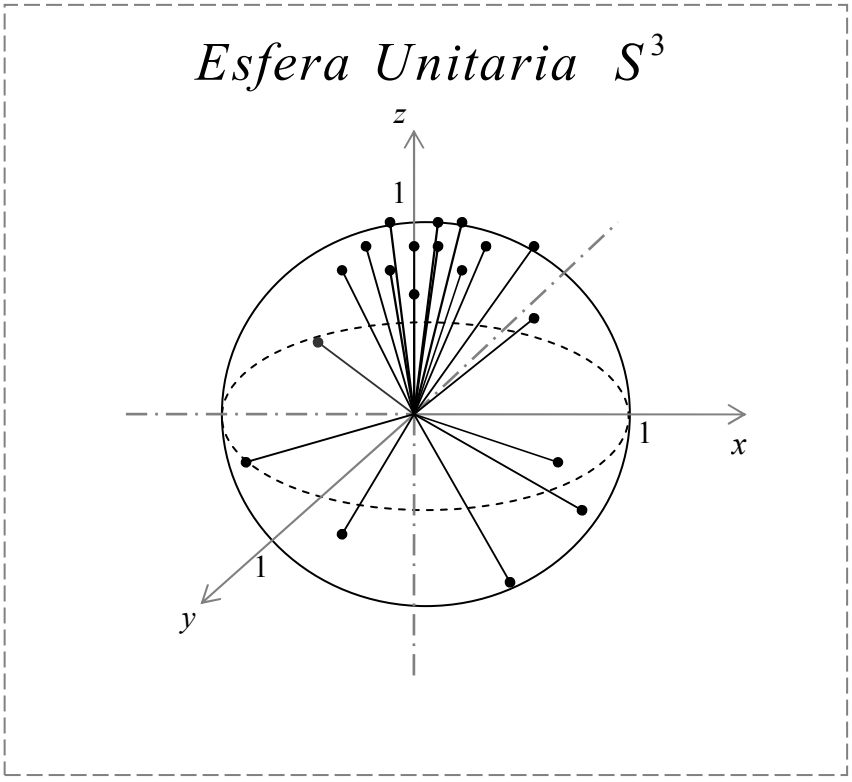

Figura 1.5: Datos direccionales de dimensión $p = 3$

Se denota el vector direccional aleatorio de dimensión p por θ, donde $\theta^T\theta = 1$. El vector unitario θ toma valores en la superficie de S^p (hiperesfera unitaria de dimensión p).

Los datos direccionales se consideran por lo regular en su forma polar esférica $x = r\theta(u)$ con $\theta = (\theta_1, \ldots, \theta_p)^T$ y $r > 0$, donde:

$$u_i = \cos\theta_i \prod_{j=0}^{i-1} \operatorname{sen}\theta_j \quad , \quad i = 1, \ldots, p$$

$$, \quad \operatorname{sen}\theta_0 = 1$$

$$, \quad \cos\theta_p = 1$$

El Jacobiano de la transformación de coordenadas polares esféricas a coordenadas rectangulares de x es:

$$J_p = r^{p-1} \prod_{i=1}^{p-1} \operatorname{sen}^{p-1}\theta_{i-1}$$

$$J_2 = r$$

y siempre $r = 1$, ya que se considera el trabajo en la hiperesfera unitaria.

1.4.1 Dirección Media y Longitud Media

Sea $X_1, X_2, \ldots, X_n$ una m.a. de datos direccionales de una población Θ, en la esfera unitaria S^p. Entonces la localización de esos puntos dejan entrever su dirección media muestral por un vector en $\mathbb{R}^p$, que es:

$$\overline{\mathrm{x}} = \frac{1}{n}\sum_{i=1}^{p} x_i .$$

Como en el caso de la esfera en dos dimensiones, es mejor expresar el vector $\overline{\mathrm{x}}$ en su forma polar como:

$$\overline{\mathrm{x}} = \overline{R}\overline{\mathrm{x}}_0 ,$$

donde $\overline{\mathrm{x}}_0$ es un vector unitario y $\overline{R} \geq 0$ ya que:

$$\overline{R} = || \overline{\mathrm{x}}||$$

por lo tanto

$$\overline{\mathrm{x}}_0 = ||\overline{\mathrm{x}}||^{-1}\overline{\mathrm{x}}$$

El vector unitario $\overline{\mathrm{x}}_0$ es llamado la *dirección media* y $\overline{R}$ es llamado la longitud media.

1.4.2 Distribución Von Mises-Fisher

Un vector aleatorio x sigue una distribución de dimensión p Von Mises-Fisher, si la función de densidad de probabilidad es:

$$f(x;\mu,k) = \left(\frac{k}{2}\right)^{(p/2)-1} \frac{1}{\Gamma(p/2) I_{(p/2)-1}(k)} \exp\{k\mu^T x\}$$

donde $k \geq 0$, $||\mu|| = 1$ y I_v denota la función modificada de Bessel de primer tipo de orden v. Donde μ es la media direccional y k es el parámetro de concentración.

Capítulo 2. Método Bootstrap

El *Método Bootstrap* es una técnica estadística moderna desarrollada (que utiliza implementos computacionales) para la simplificación de pruebas estadísticas o valoración de un punto estimado en situaciones cuando los procedimientos estadísticos tradicionales no son validos y/o no viables; que permite obtener estimación de medidas de precisión así como la realización de pruebas de hipótesis en aquellas situaciones en las que no se dispone de información acerca de la distribución muestral de una estadística o en casos en los que la distribución muestral depende de parámetros desconocidos.

Entonces el *Método Bootstrap* es útil para la descripción de la distribución muestral de aquellos estimadores con propiedades muestrales desconocidas o difícilmente obtenibles por medios analíticos; es usado como una alternativa basada en supuestos paramétricos cuando esos supuestos son dudosos o cuando la inferencia paramétrica es imposible o requiere fórmulas complicadas.

El *Método Bootstrap* depende mucho de la noción de lo que es una *muestra bootstrap*, por ello sea $\mathrm{X} = \{X_1, \ldots, X_n\}$ una m.a. de una población de tamaño n la cual sigue una función de distribución F desconocida (o conocida) y definimos a $\hat{F}$ como la función de distribución empírica de las observaciones de X. Sea $\mathrm{x} = \{x_1, \ldots, x_n\}$ el conjunto de observaciones de la muestra aleatoria X, de las cuales la probabilidad de seleccionar una x_i para alguna i es $\frac{1}{n}$ con $i = 1, \ldots, n$.

Una muestra bootstrap se define como una muestra aleatoria de tamaño n (la misma n de la m.a. X) bajo la distribución $\hat{F}$, y se expresa como:

$$\mathrm{x}^* = \{x_1^*, \ldots, x_n^*\}$$

donde la notación x^* nos indica que este conjunto de observaciones ya no es el mismo conjunto x de la m.a. X, sino que es una versión del conjunto x en la cual las observaciones han sido sorteadas o reemplazadas, es decir, x^* es el resultado de aplicar una vez un re-muestreo a las observaciones de la m.a. X.

Una forma más sencilla de entender lo que es una muestra bootstrap es que los datos $\mathrm{x}^* = \{x_1^*, \ldots, x_n^*\}$ son una m.a. con reemplazo de tamaño n, proveniente de una población de n observaciones $x_1, \ldots, x_n$. El conjunto $\mathrm{x}^* = \{x_1^*, \ldots, x_n^*\}$ de observaciones bootstrap está compuesto por las observaciones de la m.a. X, donde cada observación de la muestra original puede no aparecer en la muestra bootstrap, puede aparecer una vez o dos veces, etc.

Habiendo definido lo que es una muestra bootstrap, hay que comprender su utilidad de la siguiente forma: en el estudio de una población de tamaño *n* nos interesa algún parámetro desconocido θ (o más) de la función de distribución F y la estimación del mismo (mediante una estadística) se realiza a partir de las observaciones de la m.a. X y se denota por $\hat{\theta}$.

Correspondiendo a la muestra bootstrap x^* se puede tener un estimador para θ, que llamaremos *réplica bootstrap* de $\hat{\theta}$ y se denota como:

$$\hat{\theta}^* = s(\mathrm{x}^*)$$

donde la cantidad $s(\mathrm{x}^*)$ es el resultado de aplicar a x^*, como si fuera las observaciones de la m.a. original, la misma función para obtener la estadística $s(\cdot)$ y así obtener el estimador para el parámetro desconocido.

Para tener un buen estimador por bootstrap del valor real del parámetro desconocido θ, no basta con solo obtener una muestra bootstrap y una replica bootstrap, sino es necesario obtener muchas muestras bootstrap independientes con su respectiva réplica bootstrap, es decir, se deben de obtener:

$$\begin{array}{ll} \mathrm{x}^{*1} = \{x_1^{*1}, \ldots, x_n^{*1}\} & \hat{\theta}_1^* = s(\mathrm{x}^{*1}) \\ \vdots \qquad\qquad\qquad\qquad \longrightarrow & \vdots \\ \mathrm{x}^{*B} = \{x_1^{*B}, \ldots, x_n^{*B}\} & \hat{\theta}_B^* = s(\mathrm{x}^{*B}) \end{array}$$

donde B es el número de veces que se realiza el bootstrap, las $\mathrm{x}^{*i}\ \ i=1,...,B$ son B muestras bootstrap independientes y las $\hat{\theta}_i^*\ \ i=1,...,B$ son las correspondientes réplicas bootstrap.

Por lo tanto, ya que tenemos las $\hat{\theta}_i^*\ \ i=1,...,B$ réplicas bootstrap del parámetro θ, estas se pueden utilizar para:

- Utilizar el estimador $\overline{\theta}=\frac{1}{B}\sum_{i=1}^{B}\hat{\theta}_i^*$ como un mejor estimador para el valor real del parámetro θ, y que se permite considerarlo, ya que:
  - Las $\hat{\theta}_i^*\ \ i=1,...,B$ tienen su propia distribución.
  - Si la muestra es un buen indicador de la población, el estimador bootstrap de la población será similar al de la muestra original.
  - Tanto como B tienda a infinito el estimador bootstrap tenderá a parecerse más al valor real del parámetro θ de la población bajo estudio.
- La estimación del error estándar de $\hat{\theta}$.
- La construcción de Regiones de Confianza para el valor real del parámetro θ.
- Realizar pruebas de hipótesis.

2.1 Definición de un algoritmo para el Método Bootstrap

El *Método Bootstrap* puede considerarse como un algoritmo que consiste de cinco pasos:

1. Sea $\mathrm{X}=\{X_1,\ldots,X_n\}$ una m.a. de una población (bajo estudio) de tamaño n, de la cual se define a su respectivo conjunto de observaciones $\mathrm{x}=\{x_1,\ldots,x_n\}$ como el conjunto de observaciones original.

Se debe tener en cuenta el parámetro de interés, por ejemplo θ y utilizando las observaciones de X considerar el estimador $\hat{\theta}$.

Las observaciones de la m.a. original son tratadas como si constituyesen los datos de toda la población, es decir, se utilizan como el universo del que se extraerán muestras con reemplazo.

2. Tomar una muestra aleatoria de tamaño n con reemplazo (re-muestreo) de las observaciones de la m.a. original X, en la cuál puede ocurrir que se tengan observaciones que ocurran más de una vez, y otras no. A este re-muestro se le llamará muestra bootstrap y se le define como x^*.

   Para hacer el re-muestreo se puede considerar alguno de los siguientes métodos:

   I. *Re-muestreo con reemplazo.*

      (Utilizando la función $f : R \to R$ tal que $f(s) = [s]$, donde [s] es el mayor entero menor o igual a 's')

      Sea $u_1, \ldots, u_n$ un conjunto de números aleatorios provenientes de una distribución uniforme U[0,1], y sea:

      $$m_i = [nU_i + 1] \quad , \quad i = 1, \ldots, n$$

      entonces la m.a. con reemplazo (muestra bootstrap) es la m.a. formada por las observaciones:

      $$\mathrm{x}^* = (x_{m_1}, \ldots, x_{m_n})$$

   II. *Re-muestreo balanceado*

      i. Decidir el número *B* de muestras bootstrap a utilizar.

      ii. Crear un vector de longitud $B \times n$, que consiste en la secuencia de las observaciones $x_1, \ldots, x_n$ repetida *B* veces:

      $$(\overbrace{x_1, \ldots, x_n}^{1}, \overbrace{x_1, \ldots, x_n}^{2}, \ \ldots \ , \overbrace{x_1, \ldots, x_n}^{\mathrm{B}})$$

      iii. Aplicar una permutación aleatoria al vector para obtener:

$$\overbrace{x_1^*,\ldots,x_n^*}^{1},\ \overbrace{x_{n+1}^*,\ldots,x_{2n}^*}^{2},\ \ldots,\ \overbrace{x_{Bn-n+1}^*,\ldots,x_{Bn}^*}^{B}$$

y este resultado es B muestras bootstrap que contienen números iguales de las observaciones originales $x_1,\ldots,x_n$

3. Estimar el parámetro de interés con las observaciones de la muestra bootstrap (réplica bootstrap) que se define como $\hat{\theta}^*$.
4. Repetir los pasos 2 y 3 B veces $B>25$, para así conseguir B bootstrap estimadores $\hat{\theta}^*$, es decir, se obtienen $\hat{\theta}_i^*$, $i=1,...,B$.
5. Las $\hat{\theta}_i^*$ $i=1,...,B$ réplicas bootstrap son la base para el análisis de la característica que nos interesa de la población bajo estudio.

2.2 Estimación por Bootstrap del Error Estándar de $\hat{\theta}$

El error estándar de una estadística $\hat{\theta}$, expresado por $se_F(\hat{\theta})$, es un estimador que usa la función de distribución empírica $\hat{F}$ en lugar de la función de distribución F desconocida. Específicamente el estimador bootstrap de $se_F(\hat{\theta})$ está definido por:

$$se_F(\hat{\theta}^*) \quad \text{(estimador ideal bootstrap)},$$

en otras palabras, el estimador de $se_F(\hat{\theta})$ es el error estándar de $\hat{\theta}$ que se tiene al realizar un re-muestreo del conjunto de observaciones de la muestra original.

Sea $\mathrm{X} = \{X_1, \ldots, X_n\}$ una m.a. de una población de tamaño n y su correspondiente conjunto de observaciones $\mathrm{x} = \{x_1, \ldots, x_n\}$:

1. Seleccionar B muestras bootstrap independientes con reemplazo de x

$$\mathrm{x}^{*1}, \mathrm{x}^{*2}, \ldots, \mathrm{x}^{*B}$$

(Para la estimación del error estándar, B se debe escoger de tal manera que $B > 25$)

2. Evaluar las réplicas bootstrap correspondientes a su respectivas muestras bootstrap,

$$\hat{\theta}_i^* = s(\mathrm{x}^{*i}) \qquad i = 1, \ldots, B$$

3. Estimar el error estándar $se_F(\hat{\theta})$ mediante la desviación estándar muestral de las B replicas bootstrap.

$$\widehat{se}_B = \left\{ \frac{1}{B-1} \sum_{i=1}^{B} \left(\hat{\theta}_i^* - \hat{\theta}^*(\cdot) \right)^2 \right\}^{\frac{1}{2}}$$

$$donde\ \hat{\theta}^*(\cdot) = \frac{1}{B} \sum_{i=1}^{B} \hat{\theta}_i^*$$

El límite de $\widehat{se}_B$ cuando B tiende a infinito es el estimador ideal bootstrap de $se_F(\hat{\theta})$,

$$\lim_{B \to \infty} \widehat{se}_B = se_{\hat{F}} = se_{\hat{F}}(\hat{\theta}^*),$$

lo anterior quiere decir que la desviación estándar empírica tiende a la desviación estándar poblacional cuando el número de réplicas tiende a infinito.

El estimador ideal bootstrap $se_{\hat{F}}(\hat{\theta}^*)$ y la aproximación $\widehat{se}_B$ son a veces considerados como *estimadores bootstrap no paramétricos* porque están basados en $\hat{F}$, el estimador no parámetrico de F función de distribución de la población.

2.3 Métodos Bootstrap para Construcción de Regiones de Confianza

La construcción de Regiones de Confianza se realiza como una técnica estadística para la estimación de un parámetro θ, el cual nos indica que para un nivel de confianza $1-\alpha$ con $\alpha > 0$. Si se realiza un número *d* de re-muestreos de la m.a. X=$\{X_1,\ldots,X_n\}$ de tamaño *n* de la población bajo estudio, en el $(1-\alpha)\times 100\%$ de los casos el valor real de parámetro θ se encontrará contenido en el intervalo construido mediante $\hat{\theta}$.

Existen varios métodos para la estimación de las regiones de confianza, los cuales son:

I. Bootstrap Percentil
II. Bootstrap Estándar
III. Bootstrap-t
IV. Bootstrap Sesgo Corregido

2.3.1 Bootstrap Percentil

Sea $\mathrm{X} = \{X_1,\ldots,X_n\}$ una m.a. de una población de tamaño n, con sus respectivas observaciones $\mathrm{x} = \{x_1,\ldots,x_n\}$, con un parámetro θ desconocido.

i. Obtener x^{*i} $i = 1,\ldots,B$ muestras bootstrap independientes
ii. Obtener las $\hat{\theta}_i^*$ $i = 1,\ldots,B$ réplicas bootstrap correspondientes.
iii. Se ordenan las $\hat{\theta}_i^*$ para obtener las estadísticas de orden:

$$\hat{\theta}^{*1} \le \hat{\theta}^{*2} \le \ldots \le \hat{\theta}^{*B}$$

iv. Para el nivel de confianza $1-\alpha$ con $\alpha > 0$, los cuantiles $\alpha/2$ y $1\text{-}\alpha/2$ se definen como:

$$\hat{\theta}^*_{\mathrm{L}} = \alpha/2 \quad \text{cuantil}$$
$$\hat{\theta}^*_{\mathrm{U}} = 1\text{-}\alpha/2 \quad \text{cuantil}$$

donde $\hat{\theta}^*_{\mathrm{L}}$ es la estadística de orden más pequeña y $\hat{\theta}^*_{\mathrm{U}}$ es la estadística de orden más grande, para las cuales se cumple que en el intervalo ($\hat{\theta}^*_{\mathrm{L}}$, $\hat{\theta}^*_{\mathrm{U}}$) se encuentran el $(1-\alpha)\times B$ de las $\hat{\theta}^*_i \quad i=1,...,B$, es decir, si para el $(1-\alpha)\times B$ de las réplicas bootstrap se encuentran en el intervalo con los extremos determinados por las estadísticas de orden $\hat{\theta}^*_{\mathrm{L}}$ y $\hat{\theta}^*_{\mathrm{U}}$.

$\therefore$ La región del $(1-\alpha)\times 100\%$ de confianza para θ es

($\hat{\theta}^*_{\mathrm{L}}$, $\hat{\theta}^*_{\mathrm{U}}$)

2.3.2 Bootstrap Estándar

Sea $\mathrm{X}=\{X_1,\ldots,X_n\}$ una m.a. de una población de tamaño n, con sus respectivas observaciones $\mathrm{x}=\{x_1,\ldots,x_n\}$, con un parámetro θ desconocido.

(Para aplicar este método es necesario suponer que la distribución de las $\hat{\theta}^*_i \quad i=1,...,B$ sea normal)

i. Obtener $\mathrm{x}^{*i} \quad i=1,...,B$ muestras bootstrap independientes
ii. Obtener las $\hat{\theta}^*_i \quad i=1,...,B$ réplicas bootstrap correspondientes.
iii. Calcular el estimador del error estándar $\widehat{SE_B}$ de $\hat{\theta}$ definido por

$$\widehat{SE_B}=\left\{\frac{1}{B-1}\sum_{i=1}^{B}\left(\hat{\theta}^*_i-\hat{\theta}^*(\cdot)\right)^2\right\}^{\frac{1}{2}} \quad , \textit{donde } \hat{\theta}^*(\cdot)=\frac{1}{B}\sum_{i=1}^{B}\hat{\theta}^*_i$$

que esta basado en la desviación estándar estimada de los B estimadores bootstrap.

iv. Dado que las $\hat{\theta}_i^*$ $i=1,...,B$ se distribuyen de manera normal, para el nivel de confianza $1-\alpha$ con $\alpha>0$, se toman los cuantiles $t_{(n-1)}^{\alpha/2}$ y $t_{(n-1)}^{1-\alpha/2}$ de una distribución t con (n-1) grados de libertad, donde $t_{(n-1)}^{\alpha/2}=-t_{(n-1)}^{1-\alpha/2}$.

$\therefore$ La región del $(1-\alpha)\times 100\%$ de confianza para θ es

$$\left(\hat{\theta}-t_{(n-1)}^{1-\alpha/2}\widehat{SE_B} \ , \ \hat{\theta}+t_{(n-1)}^{1-\alpha/2}\widehat{SE_B} \right)$$

donde $\hat{\theta}$ es el parámetro estimado con las observaciones de la muestra original X.

2.3.3 Bootstrap-t

Sea $\mathrm{X}=\{X_1,\ldots,X_n\}$ una m.a. de una población de tamaño n, con sus respectivas observaciones $\mathrm{x}=\{x_1,\ldots,x_n\}$, con un parámetro θ desconocido.

i. Obtener $\mathrm{x}^{*b}=\{x_1^{*b},\ldots,x_n^{*b}\}$ $b=1,\ldots,B$ muestras bootstrap independientes.

ii. Obtener las $\hat{\theta}_i^*$ $i=1,...,B$ réplicas bootstrap correspondientes.

iii. Calcular el error estándar $\widehat{SE_B}$ de $\hat{\theta}$ basado en la desviación estándar estimada de los B estimadores bootstrap.

$$\widehat{SE_B}=\left\{\frac{1}{B-1}\sum_{i=1}^{B}\left(\hat{\theta}_i^*-\hat{\theta}^*(\cdot)\right)^2\right\}^{\frac{1}{2}} \quad , \textit{donde } \hat{\theta}^*(\cdot)=\frac{1}{B}\sum_{i=1}^{B}\hat{\theta}_i^*$$

iv. Obtener la estadística t a partir de las muestras bootstrap de la siguiente forma:

Para $b = 1,\ldots,B$ calcular:

$$\hat{\mu}_b^* = \frac{1}{n}\sum_{i=1}^{n} x_i^{*b}$$

$$SE_b^* = \frac{\sum_{i=1}^{B}(x_i^* - \hat{\mu}_b^*)}{(n-1)\sqrt{n}}$$

$$\textit{donde } \hat{\mu}_b^* = \textit{media de la muestra bootstrap } \mathrm{X}^{*b}$$

$$SE_b^* = \textit{error estándar estimado de la muestra bootstrap } \mathrm{X}^{*b}$$

para así obtener:

$$t_b = \frac{\hat{\mu}_b^* - \hat{\mu}}{SE_b^*} \quad , \mathrm{b}=1,\ldots,\mathrm{B}$$

$$\textit{donde } \hat{\mu} = \textit{media de la muestra original } \mathrm{X}$$

v. Ordenar las t_b para obtener las estadísticas de orden:

$$t_1, t_2, \ldots, t_B$$

vi. Dado el nivel de confianza $1-\alpha$ con $\alpha > 0$, se obtienen los cuantiles

$$t_L = \alpha/2 \quad \text{cuantil}$$

$$t_U = 1-\alpha/2 \quad \text{cuantil}$$

donde t_L es la estadística de orden más pequeña y t_U es la estadística de orden más grande, para las cuales se cumple que en el intervalo (t_L , t_U) se encuentran el $(1-\alpha)\times B$ de las $t_i \quad i = 1,\ldots,B$, es decir, si para el $(1-\alpha)\times B$ de las $t_i \quad i = 1,\ldots,B$ se encuentran en el intervalo con los extremos determinados por las estadísticas de orden t_L y t_U.

$\therefore$ La región del $(1-\alpha)\times 100\%$ de confianza para θ es

$$\left(\hat{\theta} - t_L \widehat{SE_B} \;,\; \hat{\theta} + t_U \widehat{SE_B} \right)$$

donde $\hat{\theta}$ es el parámetro estimado con las observaciones de la muestra original X.

2.3.4 Bootstrap Sesgo Corregido

Sea $\mathrm{X} = \{X_1, \ldots, X_n\}$ una m.a. de una población de tamaño n, con sus respectivas observaciones $\mathrm{x} = \{x_1, \ldots, x_n\}$, con un parámetro θ desconocido.

i. Obtener $\mathrm{x}^{*b} = \{x_1^{*b}, \ldots, x_n^{*b}\}$ $b = 1, \ldots, B$ muestras bootstrap independientes.

ii. Obtener las $\hat{\theta}_i^*$ $i = 1, \ldots, B$ réplicas bootstrap correspondientes.

iii. Se ordenan las $\hat{\theta}_i^*$ para obtener las estadísticas de orden:

$$\hat{\theta}^{*1} \leq \hat{\theta}^{*2} \leq \ldots \leq \hat{\theta}^{*B}$$

iv. Calcular el corrector de sesgo $\hat{Z}_0$

$$\hat{Z}_0 = \Phi^{-1}\left(\frac{\#\{\hat{\theta}^* < \hat{\theta}\}}{B}\right)$$

donde $\Phi^{-1}(\cdot)$ es la función inversa de la función de distribución $N(0,1)$ y $\hat{Z}_0$ es el sesgo medio de $\hat{\theta}^*$.

v. Calcular el acelerador $\hat{a}$

1) Para $i = 1, \ldots, n$ obtener una muestra de observaciones omitiendo la observación i de la muestra de observaciones original.

$$\begin{aligned} A_i &= x_i \\ &= (x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n) \end{aligned}$$

2) Para $i = 1, \ldots, n$ obtener:

$$\hat{\theta}_{-i} = \frac{\sum_{i=1}^{n}\left(A_i - \overline{A}\right)^2}{n}$$

3) Calcular:

$$\hat{\theta}_{\cdot} = \frac{\sum_{i=1}^{n}\hat{\theta}_{-i}}{n}$$

4) El acelerador es:

$$\hat{a} = \frac{\sum_{i=1}^{n}\left(\hat{\theta}_{\cdot} - \hat{\theta}_{(i)}\right)^3}{6\left\{\sum_{i=1}^{n}\left(\hat{\theta}_{\cdot} - \hat{\theta}_{(i)}\right)^2\right\}^{3/2}}$$

vi. Para el nivel de confianza $1-\alpha$ con $\alpha > 0$, tomar de la distribución $N(0,1)$ los cuantiles $Z^{(\alpha)}$ y $Z^{(1-\alpha)}$.

vii. Determinar los valores α_1 y α_2:

$$\alpha_1 = \Phi\left(\hat{Z}_0 + \frac{\hat{Z}_0 + Z^{(\alpha)}}{1-\hat{a}\left(\hat{Z}_0 + Z^{(\alpha)}\right)}\right)$$

$$\alpha_2 = \Phi\left(\hat{Z}_0 + \frac{\hat{Z}_0 + Z^{(1-\alpha)}}{1-\hat{a}\left(\hat{Z}_0 + Z^{(1-\alpha)}\right)}\right)$$

donde $\Phi(\cdot)$ es la función de distribución normal estándar.

$$Si \quad \hat{a} = 0 \;\; y \;\; \hat{Z}_0 = 0 \;\; \Rightarrow \;\; \alpha_1 = \Phi(Z^{(\alpha)}) = \alpha$$
$$\alpha_2 = \Phi(Z^{(1-\alpha)}) = 1-\alpha$$

viii. El intervalo BCa está dado por:

$$BCa = (\; \hat{\theta}^*_L \;,\; \hat{\theta}^*_U \;)$$
$$= (\; \hat{\theta}^{(\alpha_1)} \;,\; \hat{\theta}^{(\alpha_2)}) \;,$$

donde el límite de confianza inferior es la posición del valor para el cual de la lista ordenada de $\hat{\theta}^{*1}, \hat{\theta}^{*2}, \ldots, \hat{\theta}^{*B}$ se excede el $\alpha_1 \times 100\%$ de todas las $\hat{\theta}^*_i \;\; i = 1, \ldots, B$, y el límite de confianza superior es la posición del valor para el cual se excede el $\alpha_2 \times 100\%$ de todas las $\hat{\theta}^*_i \;\; i = 1, \ldots, B$.

$\therefore$ El intervalo del $(1-\alpha)\times 100\%$ de confianza para θ es

$$(\; \hat{\theta} - t_L \widehat{SE_B} \;,\; \hat{\theta} + t_U \widehat{SE_B} \;)$$

donde $\hat{\theta}$ es el parámetro estimado con las observaciones de la muestra original X.

2.3.5 Aspectos importantes de Bootstrap

Al considerar la construcción de una región de confianza, para el parámetro desconocido, por cualquier método de bootstrap, es necesario considerar lo siguiente:

- Las regiones de confianza difieren dependiendo de la aproximación y precisión que se esté usando.
- En general, cuando n es pequeña y el sesgo del estimador crece, se tiene :

  Bootstrap-t > BCa > percentile > standard

- Los extremos del intervalo que utiliza ($\hat{\theta}_{\mathrm{L}}$, $\hat{\theta}_{\mathrm{U}}$) pueden exceder el valor mínimo y el valor máximo de las observaciones de la muestra.

2.4 Un ejemplo de aplicación del Método Bootstrap

Sea $X_1,\ldots,X_n$ una m.a. de una población Ξ con distribución normal con media μ y varianza σ^2 desconocidos, y suponiendo que se quiere determinar una región de confianza para la estimación del parámetro μ mediante una cantidad pivotal $\theta = \theta(\mu)$ (una función de μ).

Sea $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ la media, y $S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$ la varianza muestral de Ξ, una cantidad pivotal adecuada es $\overline{Y} = \frac{\overline{X} - \mu}{S}$, donde $S = \sqrt{S^2}$

Para j, ($j = 1,...,B$) (donde B es el número de veces que se realiza el bootstrap):

1. Sea $X_j^* = \{X_{j1}^*,...,X_{jn}^*\}$ una muestra por bootstrap de Ξ, con media $\bar{X}_j^*$ y varianza S_j^{*2}.
2. Estandarizando $\bar{X}_j^*$ se obtiene $\bar{Y}_j^* = \dfrac{\bar{X}_j^* - \bar{X}}{S_j^*} \sim N(0,1)$ que es el equivalente por bootstrap a $\bar{Y}$.

Ordenando las $\bar{Y}_j^* \quad j = 1,\ldots,n$ para obtener las estadísticas de orden $\bar{Y}_{(1)}^*, \bar{Y}_{(2)}^*, \ldots, \bar{Y}_{(n)}^*$, donde tomando $\bar{Y}_L^*$ y $\bar{Y}_U^*$ que respectivamente serían el valor pequeño y el valor grande que toman las $\bar{Y}_j^* \quad j = 1,\ldots,n$ para hacer el intervalo ($\bar{Y}_L^*$, $\bar{Y}_U^*$).

Sea el nivel de confianza $1-\alpha$ con $\alpha > 0$, entonces si el $(1-\alpha)\times B$ de las $\bar{Y}_j^* \quad j = 1,\ldots,n$ e encuentran el en intervalo ($\bar{Y}_L^*$, $\bar{Y}_U^*$), entonces:

$$\frac{\bar{X} - \mu}{S} \subset (\bar{Y}_L^* , \bar{Y}_U^*)$$

$$\bar{Y}_L^* < \frac{\bar{X} - \mu}{S} < \bar{Y}_U^*$$

$$S\bar{Y}_L^* < \bar{X} - \mu < S\bar{Y}_U^*$$

$$\bar{X} - S\bar{Y}_U^* < \mu < \bar{X} - S\bar{Y}_L^*$$

Por lo tanto la región del $(1-\alpha)\times 100\%$ de confianza para μ es:

$$(\bar{X} - S\bar{Y}_U^*, \bar{X} - S\bar{Y}_L^*)$$

y el intervalo correspondiente con $(1-\alpha)\times 100\%$ de confianza para $\theta = \theta(\mu)$ es:

$$(\theta(\bar{X} - S\bar{Y}_U^*) , \theta(\bar{X} - S\bar{Y}_L^*))$$

ambos basados en la cantidad pivotal $\bar{Y}$.

Capítulo 3. Construcción de Regiones de Confianza para la Dirección Media

Para las áreas de investigación científica que tienen por objeto de estudio a datos direccionales, los problemas que son de interés para las mismas son resueltos en forma Paramétrica y No Paramétrica utilizando el Método Bootstrap para determinar Regiones de Confianza, y dado que se trabaja con datos direccionales el problema más común e importante es el determinar una medida de tendencia central que es la *dirección media*.

Sea $\theta_1,\ldots,\theta_n$ una m.a. de datos direccionales de una población Θ, con dirección media μ y longitud media L, entonces:

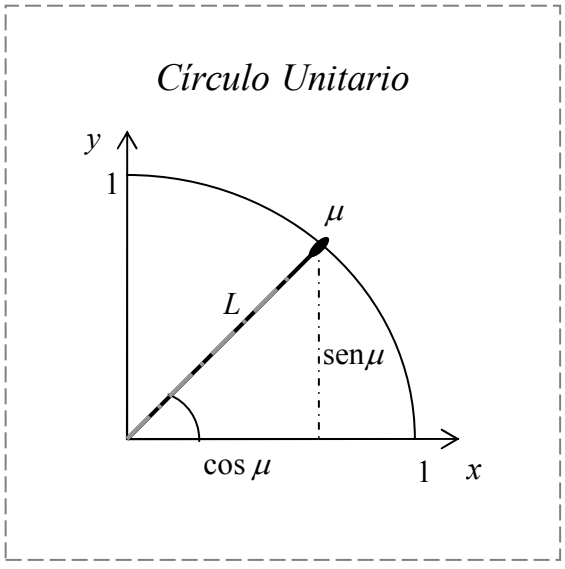

Figura 3.1: Dirección media μ y longitud media L para datos direccionales de dimensión $\rho = 2$ de una población Θ

Dado que $E(\Theta) = E(\cos\Theta, \text{sen}\Theta)$ entonces:

$$E(\cos\Theta) = \cos\theta$$

$$E(sen\Theta) = \text{sen}\theta$$

y $\overline{\theta}$ un estimador para μ, donde:

$$\cos\overline{\theta} = n^{-1}\sum_{i=1}^{n}\cos\theta_i$$

$$\text{sen}\overline{\theta} = n^{-1}\sum_{i=1}^{n}\text{sen}\theta_i$$

sea el nivel de confianza $1-\alpha$ con $\alpha > 0$, y tomando los cuantiles θ_α y $\theta_{1-\alpha/2}$, por lo tanto una región con $(1-\alpha)\times 100\%$ de confianza para μ tendrá la forma:

$$(\ \overline{\theta} - \theta_{1-\alpha/2}\ ,\ \overline{\theta} + \theta_{\alpha/2}\ )$$

3.1 Métodos No Paramétricos

3.1.1 Método Gráfico

Una forma relativamente simple y con un margen de error considerable, pero hasta cierto punto útil para la determinación de una región de confianza para la dirección media, es mediante la graficación de puntos provenientes de la estimación constante de la dirección media al utilizar el algoritmo bootstrap.

La determinación de una región de confianza para la dirección media mediante la graficación está limitada en el aspecto de que solo podemos tener la percepción gráfica hasta una dimensión tridimensional, y que para dimensiones mayores se necesita recurrir a la proyección sobre las dimensiones del espacio que podamos representar. Por lo anterior se va a considerar la utilización de este método para esferas de dimensión $p = 2$ y $p = 3$.

Sea $\theta_1,\ldots,\theta_n$ una m.a. de datos direccionales de dimensión $p=2$ de una población Θ con dirección media μ y definimos como θ_0 al valor real de μ, la dirección media estimada esta dada por el vector:

$$\bar{\theta} = \frac{\hat{\theta}}{\|\hat{\theta}\|}$$

donde $$\hat{\theta} = \frac{1}{n}\left(\sum_{i=1}^{n}\cos\theta_i, \sum_{i=1}^{n}\operatorname{sen}\theta_i\right)$$

donde $\bar{\theta}$ es un punto en el círculo unitario.

El método bootstrap consiste en calcular $\bar{\theta}_j^*$ $j=1,\ldots,B$ para un número B grande de re-muestreos y la región de confianza para la dirección media se muestra al graficar las $\bar{\theta}_j^*$ $j=1,\ldots,B$.

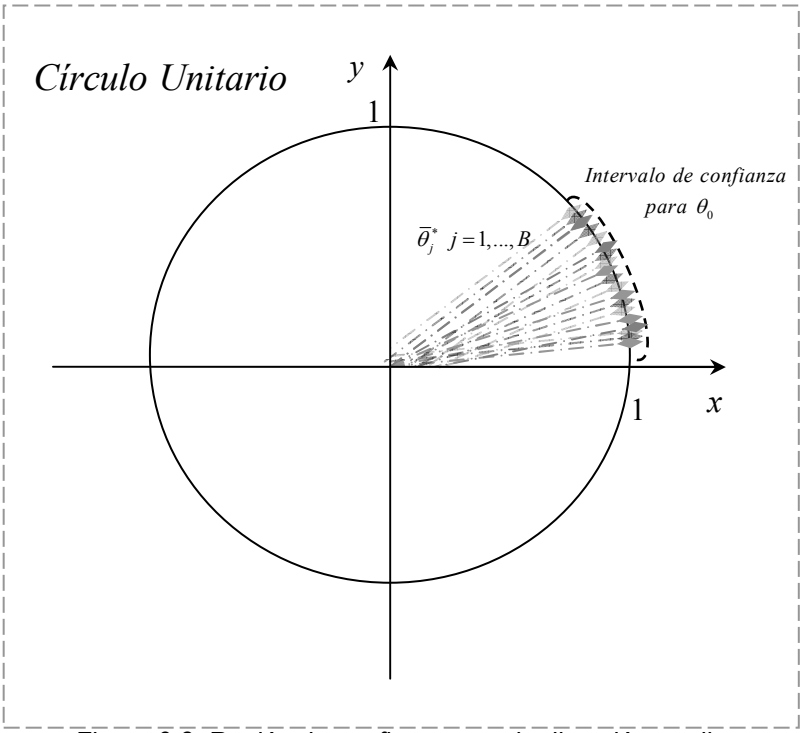

Figura 3.2: Región de confianza para la dirección media de datos direccionales de dimensión $p=2$

Ahora, sea $\theta_1,\ldots,\theta_n$ una m.a. de datos direccionales de dimensión $p=3$ de una población Θ con dirección media μ, la dirección media estimada esta dada por el vector $\bar{\theta}$ y aplicando el método bootstrap para obtener $\bar{\theta}_j^*\quad j=1,\ldots,B$ para un número B grande de re-muestreos y la región de confianza se muestra al graficar las $\bar{\theta}_j^*\quad j=1,\ldots,B$.

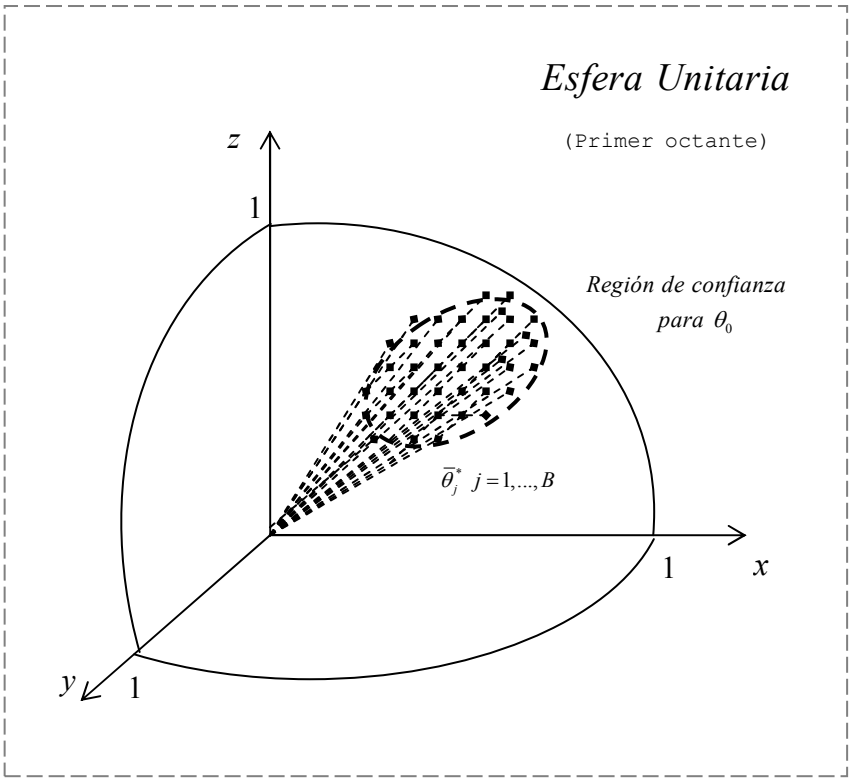

Figura 3.3: Región de confianza para la dirección media de datos direccionales de dimensión $p=3$

La región de confianza se toma a partir de la información gráfica que se obtuvo de aplicar el método bootstrap, donde para un nivel de confianza $1-\alpha$ con $\alpha>0$, la región con $(1-\alpha)\times 100\%$ de confianza queda definida en forma intuitiva.

3.1.2 Método Básico

Sea $\theta_1,\ldots,\theta_n$ una m.a. de datos direccionales de una población Θ con dirección media μ y definimos como θ_0 al valor real de μ, la dirección media estimada $\overline{\theta}$ esta dada por el vector:

$$\overline{\theta} = \frac{\hat{\theta}}{\| \hat{\theta} \|}$$

donde $$\hat{\theta} = \frac{1}{n}\left(\sum_{i=1}^{n}\cos\theta_i, \sum_{i=1}^{n}\operatorname{sen}\theta_i\right)$$

aplicando el algoritmo bootstrap B veces se obtiene:

$$\overline{\theta}_1^*, \overline{\theta}_2^*, \ldots, \overline{\theta}_B^*$$

ahora definiendo:

$$\gamma_j = \overline{\theta}_j^* - \overline{\theta} \quad , \quad j = 1,\ldots,B$$
$$, \quad -\pi \le \gamma_j \le \pi$$

donde a γ_j se debe considerar como diferencia entre ángulos y por lo cual la γ_j es un ángulo. Ordenando en forma creciente las $\gamma_j \quad j = 1,\ldots,B$ para obtener las estadísticas de orden

$$\gamma_{(1)}, \gamma_{(2)}, \ldots, \gamma_{(B)}$$

Dado un nivel de confianza $1-\alpha$ con $\alpha > 0$, para determinar la región de confianza se toman las estadísticas de orden $\gamma_{(l)}$ y $\gamma_{(m)}$ tal que $\gamma_{(l)} < \gamma_{(m)}$, donde $\gamma_{(l)}$ es la estadística de orden más pequeña y $\gamma_{(m)}$ es la estadística de orden más grande para las que se cumple que el $(1-\alpha)\times B$ de las γ_j , $j = 1,\ldots,B$ están contenidas en el intervalo $(\gamma_{(l)}, \gamma_{(m)})$.

Otra forma para seleccionar a $\gamma_{(l)}$ y $\gamma_{(m)}$ es tomando directamente la estadística de orden en la posición l y m, donde $l = \left[\frac{11}{2}B\alpha + \frac{1}{2}\right]$ (donde $[\cdot]$ es el mayor entero menor o igual a $'\cdot'$) y $m = B - l$.

Ahora, si tomando en cuenta que la mayoría de los ángulos (diferencias) γ_j están contenidas en el intervalo $(\gamma_{(l)}, \gamma_{(m)})$, entonces:

$$\bar{\theta} - \theta_0 \subset (\gamma_{(l)}, \gamma_{(m)})$$

Por lo tanto región con $(1-\alpha)\times 100\%$ de confianza para θ_0 es:

$$(\bar{\theta} + \gamma_{(l)}, \bar{\theta} + \gamma_{(m)})$$

3.1.3 Método Básico para Datos Axiales

Región de Confianza para el Eje Polar Media

Si X es un vector unitario aleatorio de dimensión p que proviene de una población axial Θ (es decir, X y $-$X tienen la misma distribución), entonces $E(\mathrm{X}) = 0$ y la dirección media no está bien definida. Por lo anterior, el parámetro de interés para la población axial es el Eje Polar Medio.

Sea $\theta_1, \ldots, \theta_n$ una m.a. de datos direccionales axiales de una población Θ con eje polar medio ε y definimos como θ_0 al valor real de ε, el eje polar medio estimado esta dado por:

$$\bar{\theta} = \frac{\hat{\theta}}{\| \hat{\theta} \|}$$

donde $$\hat{\theta} = \frac{1}{n}\left(\sum_{i=1}^{n} \cos\theta_i, \sum_{i=1}^{n} \operatorname{sen}\theta_i\right)$$

Para aplicar el algoritmo bootstrap, en el re-muestreo se realiza el siguiente cambio:

> Sea $u_1', \ldots, u_n'$ una colección de números aleatorios provenientes de la distribución $U[0,1]$. Para i, $i = 1, \ldots, n$ si $u_i' < 0.5$ entonces $\theta_i^* = 2\bar{\theta} - \theta_i^*$ en la muestra bootstrap para cada B vez.

Al aplicar el algoritmo bootstrap B veces se obtienen:

$$\overline{\theta}_1^*, \overline{\theta}_2^*, \ldots, \overline{\theta}_B^*$$

y definiendo

$$\psi_j = | \overline{\theta}_j^* - \overline{\theta} | \quad , \quad j = 1, \ldots, B$$

donde ψ_j es el valor absoluto de la diferencia entre ángulos, y por lo cual se debe considerar como un ángulo. Ordenando las ψ_j para obtener las estadísticas de orden

$$\psi_{(1)}, \psi_{(2)}, \ldots, \psi_{(B)}$$

Dado $1-\alpha$ con $\alpha > 0$, entonces sea $l = \left[B\alpha + \frac{1}{2} \right]$ y $m = B - l$ para tomar la estadística de orden $\psi_{(m)}$, y dado que se realizó el cambio necesario para determinar una distribución axial (simétrica) sobre θ_0 entonces:

$$\overline{\theta} - \theta_0 \subset (-\psi_{(m)}, \psi_{(m)}) .$$

Por lo tanto la región de $(1-\alpha)\times 100\%$ de confianza para θ_0 es:

$$(\overline{\theta} - \psi_{(m)} , \overline{\theta} + \psi_{(m)})$$

3.2 Métodos Paramétricos

La utilización de métodos paramétricos en la determinación de regiones de confianza para la media direccional se establece por la utilización de métodos pivotales. Una razón para usar métodos pivotales para calcular regiones de confianza por bootstrap en lugar de determinarlos de manera no paramétrica es porque en particular un método pivotal da regiones con un mayor nivel de confianza, es decir, que para el nivel de confianza $1-\alpha$ con $\alpha > 0$, tomando en cuenta intervalos obtenidos a partir de varios muestreos, el $(1-\alpha)\times 100\%$ de esas regiones contienen el valor real del parámetro desconocido y cuyo error tiende a ser conservativo.

Sea $X = \{X_1, \ldots, X_n\}$ una m.a. de datos direccionales de dimensión p de una población Θ con dirección media μ y definimos como θ_0 al valor real de μ.

Suponiendo lo siguiente:

a) La distribución de la población satisface la condición teórica de convergencia.
b) θ_0 está bien definida al tener un gran tamaño de población.

3.2.1 Regiones de Confianza Dúchame
(para datos direccionales utilizando el método bootstrap)

Sea $\theta_1, \ldots, \theta_n$ una m.a. de datos direccionales de una población Θ con dirección media μ y definimos como θ_0 al valor real de μ, la dirección media estimada $\bar{\theta}$.

Para datos direccionales de dimensión $p = 2$, el método de Duchame (DJRT) es usar la estadística $S(\bar{\theta}, \theta_0)$ como pivote y usando una desviación estándar, definida de la siguiente manera:

$$S(\bar{\theta}, \theta_0) = n(1 - \cos(\bar{\theta} - \theta_0))$$

donde $(1 - \cos(\bar{\theta} - \theta_0))$ tiene una distribución asintótica $\sigma^2 \chi_1^2$ cuando $\sigma^2 = \dfrac{\{1 - E(\cos 2(\Theta - \theta_0)\}}{4L^2}$ (de acuerdo con DJRT), y donde L es la longitud media.

El método bootstrap es usado para obtener los valores

$$S(\bar{\theta}_1^*, \bar{\theta}), S(\bar{\theta}_2^*, \bar{\theta}), \ldots, S(\bar{\theta}_B^*, \bar{\theta})$$

después ordenando los valores $S_j^* = S(\bar{\theta}_j^*, \bar{\theta})$, $j = 1, \ldots, B$ para obtener las estadísticas de orden

$$S_{(1)}^*, S_{(2)}^*, \ldots, S_{(B)}^*$$

Dado un nivel de confianza $1 - \alpha$ con $\alpha > 0$, se selecciona la posición $\alpha \times 100$ la estadística de orden $S_{(\alpha)}^*$ que corresponde al valor de $\bar{\theta}_\alpha^*$.

Definiendo

$$\theta_\alpha = \overline{\theta}^*_\alpha - \overline{\theta} = \tan^{-1}\left(1 - \frac{S^*_{(\alpha)}}{n}\right)$$

entonces la región por DJRT (C_B) con $(1-\alpha)\times 100\%$ de confianza para θ_0 es:

$$(\ \overline{\theta} - \theta_\alpha\ ,\ \overline{\theta} + \theta_\alpha\ )$$

Nota.- hay que señalar que el método propuesto por DJRT no es completamente pivotal, ya que la estadística propuesta $S(\overline{\theta}, \theta_0)$ depende del parámetro desconocido σ^2.

3.2.2 Versión Pivotal de la Región Dúchame

Para poder eliminar el problema del parámetro desconocido σ^2 (de la estadística propuesta por DJRT), es necesario estimar σ^2 utilizando las propiedades de los estimadores, para así obtener:

$$\begin{aligned}\hat{\sigma}^2 &= \frac{\widehat{\{1 - E(\cos 2(\Theta - \theta_0)\}}}{4L^2} \\ &= \frac{\{1 - \widehat{E(\cos 2(\Theta - \theta_0)\}}}{4\hat{L}^2} \\ &= \frac{[1 - n^{-1}\sum_{i=1}^{n}\cos 2(\theta_i - \overline{\theta})]^2}{4\hat{L}^2}\end{aligned}$$

donde $\hat{L}$ es la longitud media estimada.

Ahora definiendo

$$\begin{aligned}T(\overline{\theta}, \theta_0) &= \frac{S(\overline{\theta}, \theta_0)}{\hat{\sigma}^2} \\ &= \frac{n(1 - \cos(\overline{\theta} - \theta_0))}{\hat{\sigma}^2}\end{aligned}$$

se resuelve el problema del parámetro desconocido y de esta forma ya se tiene una cantidad pivotal adecuada.

Entonces aplicando el bootstrap B veces se obtienen $T_j^* = T(\overline{\theta}_j^*, \overline{\theta})$ $j = 1,...,B$, es decir:

$$T(\overline{\theta}_1^*, \overline{\theta}), T(\overline{\theta}_2^*, \overline{\theta}), ..., T(\overline{\theta}_B^*, \overline{\theta})$$

Ordenando las T_j^* $j = 1,...,B$ para obtener las estadísticas de orden

$$T_{(1)}^*, T_{(2)}^*, \ldots, T_{(B)}^*$$

Dado un nivel de confianza $1-\alpha$ con $\alpha > 0$, se elige la estadística de orden $T_{(\alpha)}^*$ cuya posición se obtiene del valor $\alpha \times 100$ y que corresponde al valor de $\overline{\theta}_\alpha^*$.

Ahora, definiendo

$$\Delta_{tB} = 1 - \frac{\hat{\sigma}^2 T_{(\alpha)}^*}{n}$$

por lo tanto la región Duchame Pivotal con $(1-\alpha)\times 100\%$ para θ_0 es:

$$(\overline{\theta} - \Delta_{tB} \ , \ \overline{\theta} + \Delta_{tB})$$

Observación.- es necesario mencionar que este método y el anterior tienen una desventaja al presentarse simétricamente, y no se pueda mostrar asimetrías en la distribución de Θ.

3.2.3 Clase General de Regiones Confianza

Sea $\mathrm{X} = \{X_1, \ldots, X_n\}$ una m.a. de datos direccionales en p-dimensiones de una población Θ, con un vector de medias μ y una matriz de varianzas y covarianzas Σ, y definimos como θ_0 al valor real del vector de medias μ.

Utilizando el método bootstrap para obtener X^* la dirección media estimada de X y X^* está dada por $\bar{X}$ y $\bar{X}^*$ respectivamente, y con una matriz de varianzas y covarianzas estimada asociada a X:

$$\hat{D} = n^{-1} \sum_{i=1}^{n} (X_i - \bar{X})(X_i - \bar{X})'$$

Utilizando la expresión anterior de la estimación de la matriz de varianzas y covarianzas de X, se obtiene la matriz de varianzas y covarianzas para X^*:

$$\hat{D}^* = n^{-1} \sum_{i=1}^{n} (X_i^* - \bar{X}^*)(X_i^* - \bar{X}^*)'.$$

Una cantidad pivotal adecuada para X es:

$$\bar{Y} = \mu - D^{1/2} \hat{D}^{-(1/2)} (\bar{X} - \mu)$$

y la versión bootstrap de $\bar{Y}$ es:

$$\bar{Y}^* = \bar{X} - \hat{D}^{1/2} \hat{D}^{*-(1/2)} (\bar{X}^* - \bar{X})$$

El vector $\bar{Y}^*$ no se encuentra normalizado, entonces la dirección media asociada por bootstrap es:

$$\bar{\mathrm{Y}}_0^* = \frac{\bar{Y}^*}{\| \bar{Y}^* \|}$$

Se realiza el bootstrap B veces para obtener

$$\bar{\mathrm{Y}}_{0i}^{*\prime} s \quad , \quad i = 1, \ldots, B$$

ahora dado un nivel de confianza $1-\alpha$ con $\alpha > 0$, se eliminan $\alpha \times B$ de los B puntos $\bar{\mathrm{Y}}_0^*$ para así obtener un conjunto llamado C_{TB} que contenga el $(1-\alpha) \times B$ de los B puntos $\bar{\mathrm{Y}}_0^*$.

Entonces, C_{TB} define una región de confianza en la cual se encuentre el posible valor real del vector de medias μ, ya que de las repeticiones bootstrap $\bar{Y}_0^{*\prime}s$ solo el $\alpha \times B$ de los puntos no pertenecen al conjunto.

$\therefore$ C_{TB} es una región de $(1-\alpha)$x100% de confianza para θ_0.

La desventaja del procedimiento anterior es que es muy flexible y para mostrarlo consideremos lo siguiente:

CASO 1.- Suponiendo que los datos se han rotado del tal manera que $\hat{\mu} = 0$. Sea $\Theta^* \in (-\pi, \pi]$ donde $Y_0^* = (\cos\Theta^*, \operatorname{sen}\Theta^*)$. Entonces consideremos un intervalo de confianza simétrico, es decir, el intervalo $C_{TB} = \{(\cos\theta, \operatorname{sen}\theta) : |\theta| \leq \Delta\}$, donde Δ está determinado por $(1-\alpha)\times B$ de los puntos cumplan $|\Theta^*| \leq \Delta$, y así el intervalo de confianza C_{TB} se ha vuelto simétrico.

*CASO 2.-*También podemos proponer otro caso definiendo las colas, es decir, que el intervalo $C_{TB} = \{(\cos\theta, \operatorname{sen}\theta) : \theta_1 < \theta < \theta_2\}$, donde θ_1 y θ_2 son escogidos de tal forma que se cumpla que $\frac{1}{2}\alpha B$ de los puntos Θ^* que cumplan $\Theta^* \leq \theta_1$ y que solo $\frac{1}{2}\alpha B$ cumplan $\Theta^* \geq \theta_2$.

3.2.4 Método Pivotal utilizando Vectores Ortogonales para la Región de Confianza para la Dirección Media

Sea $X = \{X_1, \ldots, X_n\}$ una m.a. de datos direccionales de una población Θ con dirección media μ y definimos como m_0 al valor real de μ.

La dirección media normalizada de un vector aleatorio está dado por $m = \frac{E(x)}{\| E(x) \|}$, donde $\| E(x) \|$ es la longitud de la media direccional, y m está bien definida suponiendo que $\| E(x) \| \neq 0$.

Un estimador de m, para utilizar el algoritmo bootstrap, es $\hat{m} = \frac{\sum_{i=1}^{n} X_i}{\| \sum_{i=1}^{n} X_i \|}$ y sean $\hat{m}_1, \ldots, \hat{m}_{p-1}$ vectores normalizados de dimensión p que son ortogonales entre si, y que son escogidos por ser ortogonales a $\hat{m}$.

Se define una matriz de dimensión $(p-1) \times p$ de la siguiente forma:

$$\hat{M}_{(p)} = [\hat{m}_1, \ldots, \hat{m}_{p-1}]'.$$

Por ejemplo, cuando $p = 2$, $\hat{m} = (\operatorname{sen}\hat{\theta}, \cos\hat{\theta})^T$ y el vector ortogonal correspondiente es $\hat{m}_1 = (-\cos\hat{\theta}, \operatorname{sen}\hat{\theta})^T$, y cuando $p = 3$ entonces $\hat{m} = (\operatorname{sen}\hat{\theta}\cos\hat{\phi}, \operatorname{sen}\hat{\theta}\operatorname{sen}\hat{\phi}, \cos\hat{\theta})^T$ y los vectores ortogonales correspondientes serían $\hat{m}_1 = (\cos\hat{\theta}\cos\hat{\phi}, \cos\hat{\theta}\operatorname{sen}\hat{\phi}, -\operatorname{sen}\hat{\theta})^T$ y $\hat{m}_2 = (-\operatorname{sen}\hat{\phi}, \cos\hat{\phi}, 0)^T$.

Considerando la estadística:

$$T(m) = nm^T \hat{M}_p^T \hat{G}^{-1} \hat{M}_p m$$

donde $\hat{G} = (\hat{G}_{jk})$ es una matriz de dimensión $(p-1) \times (p-1)$ y

$$\hat{G}_{jk} = \| n^{-1} \sum_{i=1}^{n} X_i \|^{-2} \; n^{-1} \sum_{i=1}^{n} (\hat{m}_j^T X_i)(\hat{m}_k^T X_i)$$

La estadística $T(m)$ se determinó de la siguiente manera:

a. $\hat{M}_{(p)}m$ es la proyección de $m-\hat{m}$ en el plano tangente a $\hat{m}$, es decir,

$$\begin{aligned}\hat{M}_{(p)}(m-\hat{m}) &= \hat{M}_{(p)}\cdot m - \hat{M}_{(p)}\cdot \hat{m} \qquad , \qquad \hat{M}_{(p)} \perp \hat{m} \\ &= \hat{M}_{(p)}\cdot m - 0 \\ &= \hat{M}_{(p)}\cdot m\end{aligned}$$

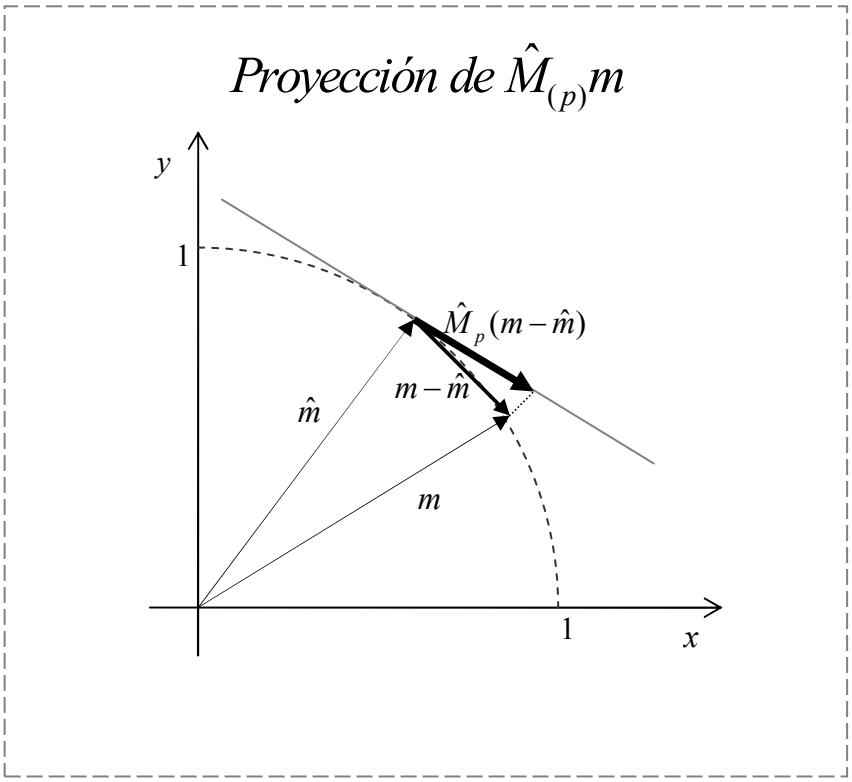

Figura 3.4: Representación geométrica de la proyección de $\hat{M}_{(p)}m$ con $p=2$

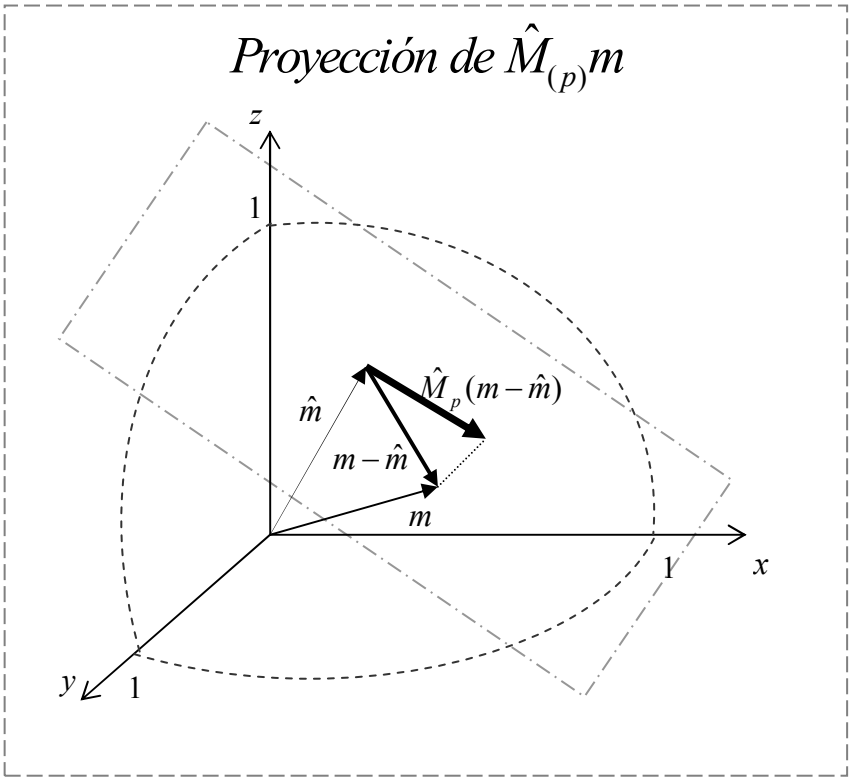

Figura 3.5: Representación geométrica de la proyección de $\hat{M}_{(p)}m$ con $p=3$

b. $\hat{G}$ es un estimador consistente de la matriz de covarianza de $\hat{M}_{(p)}m$.

Por lo tanto T es una cantidad pivotal de acuerdo a como está construido y se utiliza para determinar la región de confianza para μ.

Sea $\mathrm{X}^* = \{X_1^*, \ldots, X_n^*\}$ un re-muestreo obtenido de aplicar una vez el método bootstrap y se calcula $\hat{M}_{(p)}^*$, $\hat{G}^*$ y $T^*(m)$.

Con lo anterior, la versión del bootstrap consiste en calcular $\chi_j^* \quad j = 1, \ldots, B$ para un número grande B de re-muestreo y llegar a obtener $T_1^*(m), \ldots, T_B^*(m)$.

Después se ordenan las $T_j^*(m),\ j = 1, \ldots, B$ para obtener las estadísticas de orden

$$T_{(1)}^*(m), T_{(2)}^*(m), \ldots, T_{(B)}^*(m).$$

Dado un nivel de confianza $1-\alpha$ con $\alpha > 0$, se selecciona la estadística de orden $T_\alpha^*(m)$, que se obtiene a partir del valor $\alpha \times 100$ que determina el valor del cuantil t_α^* que satisface $P(T^*(\hat{m}) \leq t_\alpha^* \mid \chi) = \alpha$.

Por lo tanto la región con $(1-\alpha)\times 100\%$ de confianza para μ es:

$$\Re_\alpha^{(1)} = \{m : T(m) \leq t_\alpha^*\}$$

donde para m_0 que es el valor verdadero de μ, la $P(m_0 \in \Re_\alpha^{(1)}) = 1-\alpha+O(n^{-2})$ y la función $O(n^{-2})$ mide la velocidad de convergencia (ver anexos).

Para la determinación del cuantil t_α^* es necesario realizar una simulación que utilice el Método de Monte Carlo para que se pueda ajustar una distribución a $T_1^*(m), T_2^*(m), \ldots, T_B^*(m)$.

3.2.5 Método Pivotal utilizando Vectores Ortogonales para la Región de Confianza para el Eje Polar Medio

Sea $\mathrm{X} = \{X_1, \ldots, X_n\}$ una m.a. de datos direccionales axiales de una población Θ con eje polar medio ε y definimos como $\mathbb{m}$, donde $\mathbb{m}$ es un vector unitario de dimensión *p*, como el vector que maximiza $\mathbb{m}^T S \mathbb{m}$, donde:

$$\begin{aligned} S &= \widehat{\mathrm{var}(\mathrm{X})} \\ &= E[(\mathrm{X} - \widehat{E(\mathrm{X})})(\mathrm{X} - \widehat{E(\mathrm{X})})^T] \\ &= E[(\mathrm{X} - 0)(\mathrm{X} - 0)^T] \\ &= E[\mathrm{X}\mathrm{X}^T] \end{aligned}$$

Se quiere maximizar $\mathfrak{m}^T S\mathfrak{m}$, entonces:

$$\frac{\partial(\mathfrak{m}^T S\mathfrak{m})}{\partial\mathfrak{m}} = 2S\mathfrak{m} = 0 \Leftrightarrow 2S\mathfrak{m} = 0$$
$$\Leftrightarrow S\mathfrak{m} = 0$$
$$\Leftrightarrow S\mathfrak{m} = 0\mathfrak{m}$$

y sabemos que $S \neq 0$ y $\mathfrak{m} \neq 0$, por lo cual definiendo $\lambda = 0$, entonces λ es el valor propio trivial del vector $\mathfrak{m}$ y así para maximizar $\mathfrak{m}^T S\mathfrak{m}$ es necesario encontrar los vectores propios correspondientes λ no triviales para S, es decir,

$$S\mathfrak{m} = \lambda\mathfrak{m}$$

y donde $\mathfrak{m}$ será el eje polar medio siendo el vector propio asociado con el máximo valor propio λ de S.

Sea $\mathrm{X} = \{X_1, \ldots, X_n\}$ una m.a. de datos direccionales axiales de una población Θ con eje polar medio ε, y definimos a $\mathfrak{m}_0$ como el valor real de ε y obteniendo $\mathfrak{m} = \lambda$ de S.

Haciendo un primer re-muestro que nos de $\chi^* = \{X_1^*, \ldots, X_n^*\}$, entonces aplicando por primera vez el algoritmo bootstrap se obtiene el análogo para $\mathfrak{m}$, es decir, se obtiene el vector propio $\hat{\mathfrak{m}}$ asociado con el valor propio más grande $\hat{S}$.

El percentil t para el intervalo de confianza para el eje polar medio $\mathfrak{m}$ se construye mediante la siguiente estadística:

$$T(\mathfrak{m}) = n\mathfrak{m}^T \hat{M}_{(p)}^T \hat{G}^{-1} \hat{M}_{(p)} \mathfrak{m}$$

donde $\hat{G} = (\hat{G}_{jk})$ que es una matriz de dimensión $(p-1)\times(p-1)$

$$\hat{G}_{jk} = n^{-1}(\hat{\eta}_p - \hat{\eta}_j)^{-1}(\hat{\eta}_p - \hat{\eta}_k)^{-1}\sum_{i=1}^{n}(\hat{\mathfrak{m}}_j^T \mathrm{X}_i)(\hat{\mathfrak{m}}_k^T \mathrm{X}_i)(\hat{\mathfrak{m}}^T \mathrm{X}_i)^2$$

$$\hat{M}_{(p)} = [\hat{\mathfrak{m}}_1, \ldots, \hat{\mathfrak{m}}_{p-1}]^T$$

donde $\hat{\mathfrak{m}}_1,\ldots,\hat{\mathfrak{m}}_{p-1}$ son vectores ortogonales a $\hat{\mathfrak{m}}$ y $\hat{\eta}_1,\ldots,\hat{\eta}_p$ son los valores propios correspondientes a los vectores propios $\hat{\mathfrak{m}}_1,\ldots,\hat{\mathfrak{m}}_{p-1},\hat{\mathfrak{m}}$ de $\hat{S}$.

Ahora, haciendo X_j^* $j=1,...,B$ de remuestreos por bootstrap, para cada X_j^* se calculan $\hat{M}_{(p)_j}^*$ y $\hat{G}_j^*$ para así obtener $T_1^*(\mathfrak{m}),\ldots,T_B^*(\mathfrak{m})$.

Las $T_j^*(\mathfrak{m})$, $j=1,...,B$ se ordenan para obtener las estadísticas de orden

$$T_{(1)}^*(\mathfrak{m}),T_{(2)}^*(\mathfrak{m})\ldots,T_{(B)}^*(\mathfrak{m})$$

y dado un nivel de significancia $\alpha>0$ se toma el valor $\alpha\times100$ para escoger a la estadística de orden $T_{(\alpha)}^*(\mathfrak{m})$ que corresponde al valor del percentil t_α^* que satisface $P(T^*(\hat{\mathfrak{m}})\le t_\alpha^* \mid \chi)=\alpha$.

Por lo tanto la región con $(1-\alpha)\times100\%$ de confianza para el Eje Polar Medio $\mathfrak{m}$ es:

$$\Re_\alpha^{(1)}=\{\mathfrak{m}:T(\mathfrak{m})\le t_\alpha^*\}$$

donde siendo $\mathfrak{m}_0$ el valor verdadero del eje polar medio la $P(\mathfrak{m}_0\in\Re_\alpha^{(1)})=1-\alpha+O(n^{-2})$ y la función $O(n^{-2})$ mide la velocidad de convergencia (ver anexos).

Para la determinación del cuantil t_α^* es necesario realizar una simulación que utilice el Método de Monte Carlo para que se pueda ajustar una distribución a $T_1^*(m),T_2^*(m),\ldots,T_B^*(m)$.

3.2.6 Región de Confianza Empírica General

Sea $\mathrm{X} = \{X_1, \ldots, X_n\}$ una m.a. de datos direccionales de una población Θ con dirección media μ y definimos como m_0 al valor real de μ, y sea $\mathbb{p} = (\mathrm{p}_1, \ldots, \mathrm{p}_n)$ un vector de probabilidad, es decir, $\sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{p}_\mathrm{i} = 1$ y $\mathrm{p}_i \geq 0$, $i = 1, \ldots, n$.

Se define por $\tilde{m}(\mathbb{p})$ al valor que toma parámetro m cuando la distribución de la población es discreta con probabilidad p_i en el punto X_i $1 \leq i \leq n$, entonces la dirección media m es:

$$\tilde{m}(\mathbb{p}) = \frac{\sum_{i=1}^{n} \mathrm{p}_i X_i}{\| \sum_{i=1}^{n} \mathrm{p}_i X_i \|}$$

La verosimilitud empírica L, evaluada en m, está definida como:

$$L(m) = \max_{\mathrm{p}:\tilde{m}(\mathrm{p})=m} \prod_{i=1}^{n} \mathrm{p}_i$$

y sea $\hat{m}$ es el estimador por bootstrap de m, entonces:

$$L(\hat{m}) = n^{-n}$$

por lo tanto la razón de verosimilitudes correspondiente es:

$$\lambda = \frac{L(\hat{m})}{L(m)}$$

De λ podemos notar que es difícil calcular su distribución, por lo cual se utiliza la siguiente forma a la que definiremos como $W(m)$, es decir:

$$W(m) = -2 \log \left\{ \frac{L(\hat{m})}{L(m)} \right\}$$

En particular $W(m)$ satisface en mínimas condiciones regulares para el *Teorema de Wilks (Chernoff 1954; Wilks 1938)*, el cual da el siguiente resultado:

$$W(m_0) \longrightarrow \chi_t^2 \quad \textit{en distribución}$$

donde m_0 es el valor real de m y t es el rango de la matriz de covarianza asintótica de $n^{1/2}(\hat{m}-m_0)$. Hay que observar que el límite de la distribución de $W(m_0)$ no depende de ningún parámetro desconocido.

Una forma simple para construir el intervalo de confianza por verosimilitud empírica es usar la distribución χ_t^2, entonces $W(m) \sim \chi_t^2$, y dado un nivel de confianza $1-\alpha$ con $\alpha > 0$ se tiene que:

La región con $(1-\alpha)\times 100\%$ de confianza para m es:

$$\Re_\alpha^{(2)} = \{m : W(m) \le c_\alpha\}$$

donde c_α satisface $P(\chi_t^2 \le c_\alpha) = \alpha$ y donde para m_0 que es el valor verdadero de m, la $P(m_0 \in \Re_\alpha^{(2)}) = 1-\alpha+O(n^{-1})$ y la función $O(n^{-1})$ mide la velocidad de convergencia.

3.2.7 Región de Confianza por Verosimilitud Empírica para la Dirección Media

Dado un nivel de confianza $1-\alpha$ con $\alpha > 0$, la región con $(1-\alpha)\times 100\%$ de confianza para la dirección media por verosimilitud empírica es

$$\Re_\alpha = \{m : W(m) \le \mathrm{p}_\alpha\}$$

A continuación se explica como se determina $\Re_\alpha$ (para este fin es necesario trabajar con *ecuaciones trascendentes*) en el caso de datos direccionales con dimensión $p=2$ y $p=3$, donde la derivación de $\Re_\alpha$ conlleva a utilizar los *Multiplicadores de Lagrange*.

I. Suponiendo la dimensión $\rho = 2$, entonces $\Re_\alpha$ está determinada por dos ecuaciones:

a. $$(-\cos\theta, \operatorname{sen}\theta)\sum_{i=1}^{n}\{1+k(-\cos\theta, \operatorname{sen}\theta)X_i\}^{-1}X_i = 0$$

b. $$2\sum_{i=1}^{n}\log\{1+k(-\cos\theta, \operatorname{sen}\theta)X_i\} - \mathrm{p}_\alpha = 0$$

y las dos ecuaciones anteriores cuentan con dos soluciones, (θ_1 , k_1) y (θ_2 , k_2), donde $\theta_1 < 0 < \theta_2$.

Por lo tanto la región con $(1-\alpha)\times 100\%$ de confianza resultante está dado por:

$$\Re_\alpha = \{(\operatorname{sen}\theta, \cos\theta)^T : \theta_1 < 0 < \theta_2\}$$

II. Suponiendo la dimensión $\rho = 3$, entonces $\Re_\alpha$ está determinada por tres ecuaciones:

a. $$\sum_{i=1}^{n}\{1+k_1 m_1^T X_i + k_2 m_2^T X_i\}^{-1} m_1^T X_i = 0$$

b. $$\sum_{i=1}^{n}\{1+k_1 m_1^T X_i + k_2 m_2^T X_i\}^{-1} m_2^T X_i = 0$$

c. $$2\sum_{i=1}^{n}\log\{1+k_1 m_1^T X_i + k_2 m_2^T X_i\} - \mathrm{p}_\alpha = 0$$

donde se supone que m_1 y m_2 son ortogonales entre si y se definen como:

$$m_1 = (\cos\theta\cos\phi, \cos\theta\operatorname{sen}\phi, -\operatorname{sen}\theta)$$

$$m_2 = (-\operatorname{sen}\phi, \cos\phi, 0)$$

Una vez fijado ϕ, y con m_1 y m_2, para las tres ecuaciones se tienen tres parámetros desconocidos: $k_1(\phi)$, $k_2(\phi)$ y $\theta(\phi)$.

Por lo tanto la región con $(1-\alpha)\times 100\%$ de confianza resultante está dado por:

$$\Re_\alpha = \{(\operatorname{sen}\theta\cos\phi, \operatorname{sen}\theta\operatorname{sen}\phi, \cos\theta)^T : 0 < \theta < \theta(\phi), \phi \in [0, 2\pi)\}$$

3.2.8 Región de Confianza por Verosimilitud Empírica para el Eje Polar Medio

La región de confianza por verosimilitud empírica para el eje polar medio puede ser determinada utilizando también los *Multiplicadores de Lagrange* y *ecuaciones trascendentes* como en el caso de la dirección media.

I. Suponiendo la dimensión $p = 2$, entonces $\mathcal{R}_\alpha$ está determinada por:

a. $$\sum_{i=1}^{n} \{1 + k(-\cos\theta, \operatorname{sen}\theta) X_i X_i^T m\}^{-1} \times (-\cos\theta, \operatorname{sen}\theta) X_i X_i^T m = 0$$

b. $$2\sum_{i=1}^{n} \log\{1 + k(-\cos\theta, \operatorname{sen}\theta) X_i X_i^T m\} - \mathrm{p}_\alpha = 0$$

cuando $m = (\operatorname{sen}\theta, \cos\theta)^T$ y las dos ecuaciones también tienen dos soluciones, (θ_1 , k_1) y (θ_2 , k_2), donde $\theta_1 < 0 < \theta_2$.

Por lo tanto la región con $(1-\alpha)\times 100\%$ de confianza resultante está dado por:

$$\mathcal{R}_\alpha = \{(\operatorname{sen}\theta, \cos\theta)^T : \theta_1 < 0 < \theta_2\}$$

II. Suponiendo la dimensión $\rho = 3$, entonces $\mathcal{R}_\alpha$ está determinada por:

a. $$\sum_{i=1}^{n} \{1 + (k_1 m_1^T X_i + k_2 m_2^T X_i)(m^T X_i)\}^{-1} \times m_1^T X_i X_i^T m = 0$$

b. $$\sum_{i=1}^{n} \{1 + (k_1 m_1^T X_i + k_2 m_2^T X_i)(m^T X_i)\}^{-1} \times m_2^T X_i X_i^T m = 0$$

c. $$2\sum_{i=1}^{n} \log\{1 + k_1 m_1^T X_i + k_2 m_2^T X_i\} - \mathrm{p}_\alpha = 0$$

donde $m = (\text{sen}\theta\cos\phi, \text{sen}\theta\text{sen}\phi, \cos\theta)^T$ y tanto m_1 y m_2 son ortogonales entre si y siguen siendo:

$$m_1 = (\cos\theta\cos\phi, \cos\theta\text{sen}\phi, -\text{sen}\theta)$$

$$m_2 = (-\text{sen}\phi, \cos\phi, 0)$$

continuando con tres ecuaciones y teniendo tres parámetros desconocidos: $k_1(\phi)$, $k_2(\phi)$ y $\theta(\phi)$.

Por lo tanto la región con $(1-\alpha)\times 100\%$ de confianza resultante está dado por:

$$\Re_\alpha = \{(\text{sen}\theta\cos\phi, \text{sen}\theta\text{sen}\phi, \cos\theta)^T : 0 < \theta < \theta(\phi), \phi \in [0, 2\pi)\}$$

Nota: Para resolver las ecuaciones trascendentes obtenidas al aplicar el método de verosimilitud empírica se puede recurrir a la utilización de métodos numéricos y paquetería especializada.

Capítulo 4. Aspectos Numéricos y Un Ejemplo

4.1 Aspectos Numéricos

En la implementación de un algoritmo computacional para poder determinar una región de confianza para datos direccionales (mediante alguno de los métodos tratados en el capítulo tres o algún otro), se debe tomar en cuenta ciertos aspectos numéricos que ayudarán a entender los posibles resultados obtenidos, ya que estos pueden llegar a ser incongruentes con lo que se está estudiando. Por lo anterior a continuación se enlistan algunos aspectos importantes en la implementación de cualquier algoritmo:

1.- Como cualquier método numérico, siempre la precisión de los resultados depende de la capacidad computacional con el que se cuente y que se esté utilizando.
2.- La exactitud de cada método se incrementa cuando n (el tamaño de la muestra) aumenta.
3.- La dispersión de los datos de la muestra es muy importante, ya que si la dispersión es muy grande, ésta tiene un efecto sobre la longitud de la región, es decir, que una región de confianza obtenida por un método más preciso tal vez puede llegar a ser de mayor tamaño que una región obtenida por un método más general.
4.- El número B de repeticiones bootstrap es muy importante, ya que se obtiene mayor precisión del tamaño de la región al ser B grande, aunque a veces en la práctica al ser B muy grande puede ocurrir que la región se extienda.
5.- La confiabilidad de una región es mejor cuando el parámetro de concentración k es grande, ya que cuando el parámetro de concentración k

tiende a ser pequeño la región de confianza para la dirección media tiende a agrandarse y ser variable, y por lo tanto la región pierde exactitud.

6.- Teniendo en cuenta parámetros de concentración $k > 0.1$, una distribución que esté muy cerca de la distribución uniforme en la esfera, las regiones bootstrap mantienen un error similar.

7.- Cuando el método bootstrap es utilizado en problemas multivariados utilizando un número B pequeño (por ejemplo $B = 200$) se puede obtener en el re-muestreo resultados erráticos durante la ejecución del mismo. Lo anterior es consecuencia del cambio de dimensión, $B = 200$ puntos en una región de una muestra dada en un espacio de dimensión p pueden estar muy dispersos, aún en la dimensión $p = 3$. Por ejemplo, imaginando que los $B = 200$ puntos se encuentran contenidos en el primer octante de una esfera de dimensión $p = 3$, para un análisis posterior es posible que los $B = 200$ puntos no solo se encuentren en el primer octante. Lo que ocurre es que no se conservan las características de las regiones de confianza y por lo tanto se obtiene mucha imprecisión

8.- El método de Bootstrap Percentil t y el de Método de Verosimilitud Empírica con la aplicación de una *calibración bootstrap* (propuesta por Owen en 1998 como un refinamiento a la aproximación Ji-Cuadrada) tienen teóricamente una velocidad de convergencia de $O(n^{-2})$, recordando que lo anterior sucede para un tamaño de muestra n grande y que no tiene tanta precisión para un tamaño de muestra pequeño.

9.- El método de *Bootstrap Percentil* t que muestra simetría rotacional ó elíptica tiende a ser conservativo, es decir, la región de confianza es tan grande como se quiera.

10.- El Método de Verosimilitud Empírica con una calibración bootstrap es muy preciso aún para un tamaño de muestra pequeño y con parámetros de concentración muy chicos. La diferencia entre la aplicación de verosimilitud empírica normal y una con calibración bootstrap es mostrado cuando se utiliza un tamaño pequeño de muestra.

4.2 Un Ejemplo

Para dar un ejemplo de la construcción de regiones de confianza para datos direccionales se va a utilizar la información proveniente del siguiente experimento:

Experimento.- A campo abierto se dibuja un círculo de radio $r = 0.5\ m$ y se traza un eje coordenado de tal manera que coincida con los puntos cardinales, y nuestra dirección cero coincide con la dirección cardinal este. Se coloca en el centro del círculo una tortuga de agua dulce y se registra la dirección de escape que toma al estar en campo abierto. Lo anterior se realiza 25 veces para determinar la dirección media de escape de la tortuga.

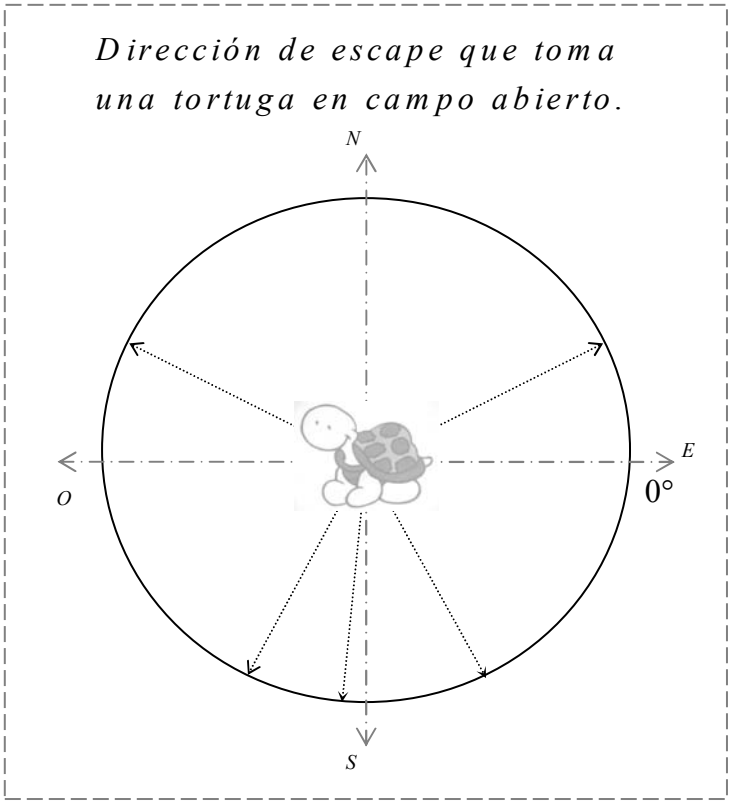

Figura 4.1: Ilustración del experimento

Sobre la muestra obtenida al realizar el experimento 25 veces se construirán *regiones de confianza* para la dirección media de escape de la tortuga.

Sea μ_0 el valor real de la dirección media de escape que toma la tortuga al estar en campo abierto, la muestra de tamaño $n = 25$ que se obtuvo es la siguiente:

muestra = {316°, 289°, 230°, 328°, 257°, 146°, 241°, 230°, 232°, 207°, 28°, 163°, 142°, 213°, 217°,317°, 138°, 169°, 161°, 123°, 182°, 184°, 31°, 150°, 351°}

su dirección media estimada es $\hat{\mu} = 311.076°$.

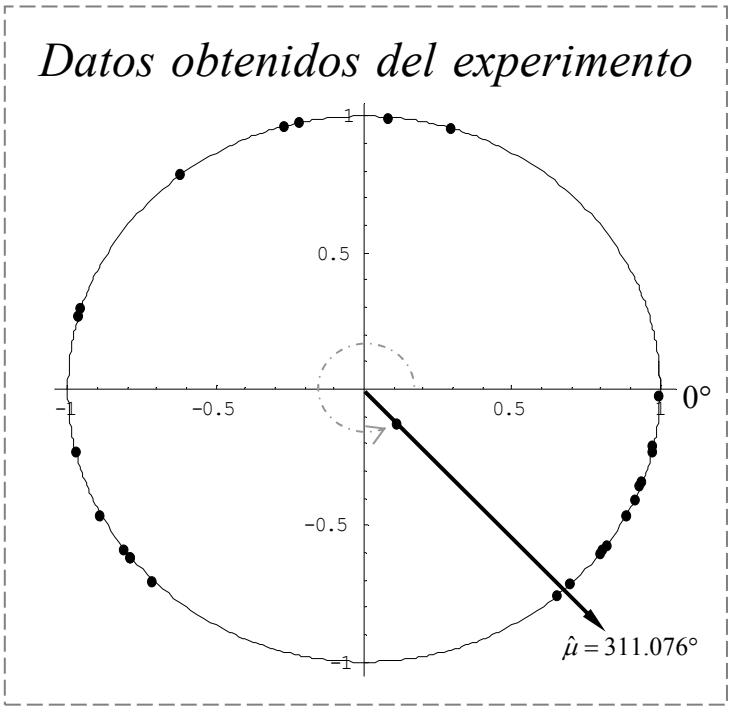

Figura 4.2: Gráfica de los datos direccionales de la muestra obtenida y su respectiva dirección media $\hat{\mu}$.

Para la determinación de las regiones de confianza para la dirección media de escape se utilizan los siguientes métodos:

- *Método Gráfico*
- *Método Básico*
- *Método Clase General de Regiones de Confianza*
- *Método Duchame*
- *Método Duchame Pivotal*

a los cuales se codificó un algoritmo en el programa en *Matemática 5* que utiliza el método bootstrap (ANEXO D). Los programas se ejecutan con $B = 50, 100, 200$ *y* 500 de repetición bootstrap y nivel de confianza $(1-\alpha)\times 100\% = 90\%,\ 95\%$ *y* 99%

Método Gráfico

Como se mencionó en capítulo tres, el método gráfico para la determinación de regiones de confianza por bootstrap no tiene una confiabilidad concretamente establecida, por ello este método lo utilizamos como un auxiliar para poder tener una vista gráfica e intuición del arco (lugar) donde puede encontrarse el valor real de la dirección media μ_0.

I. Repetición de bootstrap $B = 50$

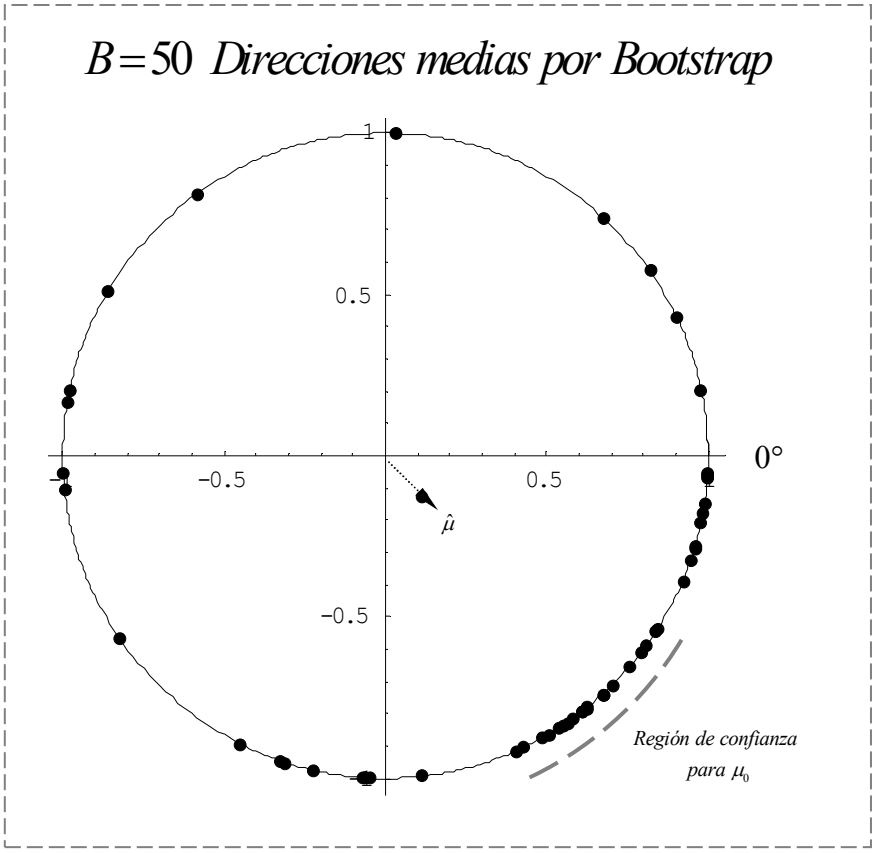

Figura 4.3: Gráfica de las $B = 50$ direcciones medias obtenidas por bootstrap y su respectiva región de confianza para μ_0.

Como se puede observar con un número B pequeño de repeticiones bootstrap se intuye una buena región de confianza para la dirección media de escape del experimento.

II. Repetición de bootstrap $B = 100$

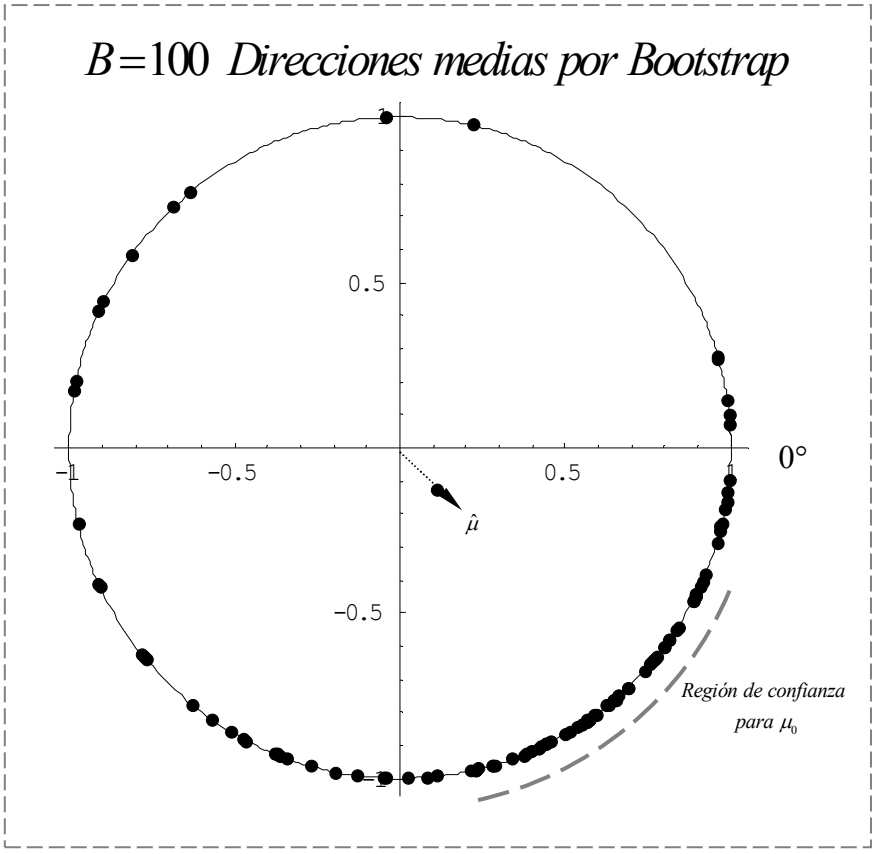

Figura 4.4: Gráfica de las $B = 100$ direcciones medias obtenidas por bootstrap y su respectiva región de confianza para μ_0.

En este caso podemos observar que al duplicar B su tamaño la región de confianza ha aumentado también de tamaño, pero aún se puede considerar a la región obtenida como una buena región de confianza.

III. Repetición de bootstrap $B = 200$

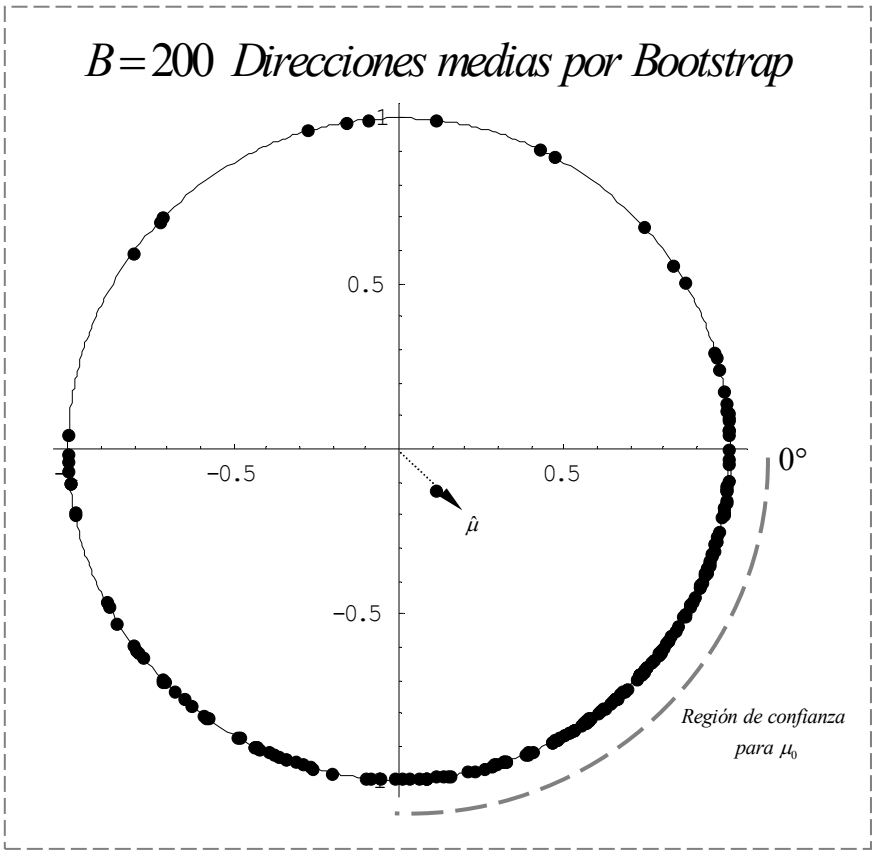

Figura 4.4: Gráfica de las $B = 200$ direcciones medias obtenidas por bootstrap y su respectiva región de confianza para μ_0.

Para este caso se observa que al ser B más grande la región de confianza ha aumentado más su tamaño y de hecho abarca por completo el cuarto cuadrante y la información que proporciona es muy escasa. Por lo anterior el método gráfico mediante el bootstrap se empieza a ser impreciso tanto como B comienza a aumentar de tamaño.

IV. Repetición de bootstrap $B = 500$

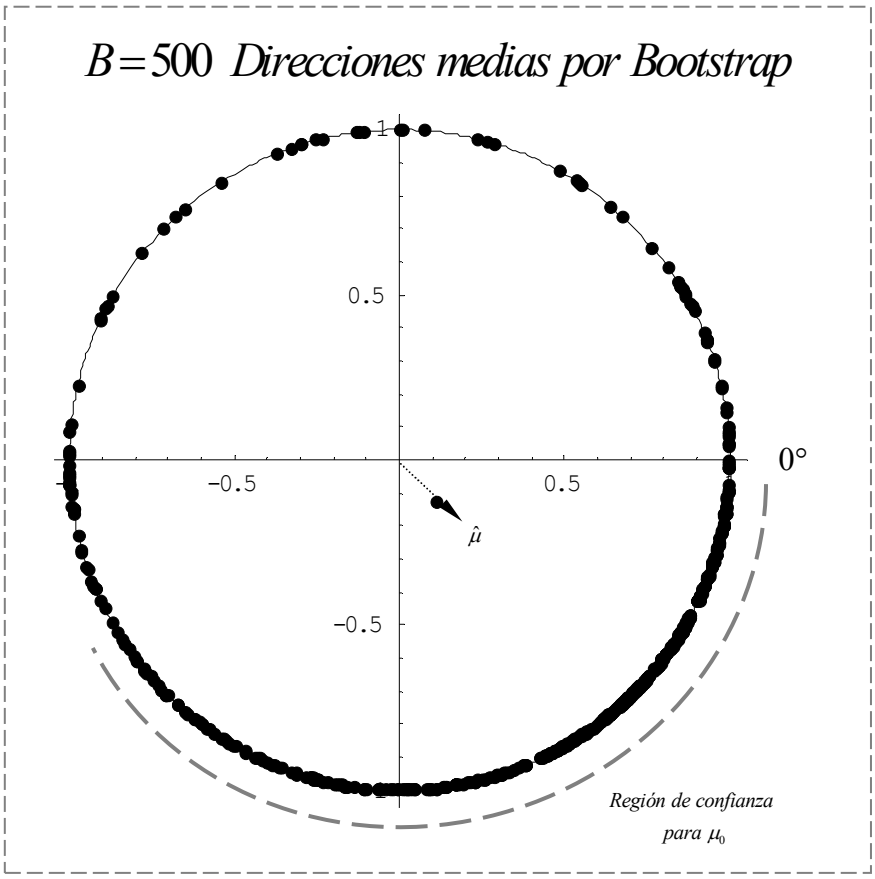

Figura 4.4: Gráfica de las $B = 500$ direcciones medias obtenidas por bootstrap y su respectiva región de confianza para μ_0.

Para este último caso en que se utiliza el método gráfico con $B = 500$ de repeticiones bootstrap la intuición sobre la disminución en la precisión e ineficiencia a medida de que B aumenta es un hecho, la región de confianza casi abarca dos cuadrantes del círculo y efectivamente no nos está dando ningún tipo de información. No se toma el caso para $B = 1000$ ya que ocurre el mismo fenómeno que en $B = 500$, es decir, ya no proporciona ninguna información y por eso se ha omitido dicho caso.

Las Regiones de Confianza

En todo momento hay que tomar en cuenta los aspectos numéricos, que se mencionaron al principio de este capítulo, al realizar el análisis de los resultados que se llegan a obtener mediante los diferentes métodos para la construcción de las regiones de confianza, para que así al tomar la región de confianza que más se apegue a nuestras necesidades ésta sea la mejor región para el parámetro de interés.

Al revisar los resultados obtenidos por los diferentes métodos y al realizar en varias ocasiones la simulación, se tienen las siguientes observaciones al respecto:

1. El tamaño de la muestra es importante, siempre se trabajará mejor con un tamaño de muestra lo suficientemente grande.
2. La dispersión de los datos de la muestra influye mucho en el tamaño de la región de confianza.
3. No siempre la dirección media estimada se encuentra dentro de la región de confianza.
4. En la estadística usual se espera que una región de confianza disminuya su tamaño al aumentar la confiabilidad del mismo, pero esto no necesariamente ocurre al utilizar el método bootstrap para auxiliarnos.
5. Al utilizar un número *B* grande de repeticiones de bootstrap es natural esperar más precisión en los resultados de la simulación, pero tampoco necesariamente esto es así, es considerable lo distante que pueden estar los resultados al ir aumentando el número *B* de repeticiones bootstrap, al igual que si se varía algún otro parámetro que sea importante durante la simulación.

A continuación se presentan los resultados obtenidos para la determinación de regiones de confianza para la dirección media por algunos de los métodos mencionados en el capítulo tres, recordando que la dirección media estimada es $\hat{\mu} = 311.076°$.

Método Básico

Nivel de confianza	Repeticiones bootstrap	Región de confianza
90%	B=50	(295.265° , 319.009°)
	B=100	(316.122° , 324.344°)
	B=200	(309.921° , 316.84°)
	B=500	(311.939° , 323.54°)
95%	B=50	(291.853° , 336.287°)
	B=100	(280.329° , 330.267°)
	B=200	(291.532° , 338.544°)
	B=500	(289.207° , 337.668°)
99%	B=50	(212.668° , 4.864°)
	B=100	(269.398° , 19.906°)
	B=200	(221.229° , 30.663°)
	B=500	(220.826° , 43.129°)

Tabla 1: Resultados obtenidos al aplicar el método básico para el cálculo de regiones de confianza utilizando bootstrap con repeticiones $B = 50, 100, 200$ *y* 500 a un nivel de confianza del $90,\ 95$ y 99%.

Método Clase General de Regiones de Confianza

Nivel de confianza	Repeticiones bootstrap	Región de confianza
90%	B=50	(282.043° , 339.923°)
	B=100	(281.235° , 339.612°)
	B=200	(281.814° , 340.702°)
	B=500	(281.957° , 340.948°)
95%	B=50	(281.536° , 341.754°)
	B=100	(281.184° , 341.260°)
	B=200	(280.536° , 341.369°)
	B=500	(281.320° , 341.557°)
99%	B=50	(279.825° , 341.374°)
	B=100	(280.137° , 341.901°)
	B=200	(280.751° , 343.045°)
	B=500	(278.554° , 342.701°)

Tabla 2: Resultados obtenidos al aplicar el método clase general de regiones de confianza utilizando bootstrap con repeticiones $B = 50, 100, 200\ y\ 500$ a un nivel de confianza del $90,\ 95$ y 99%.

Método Duchame

Nivel de confianza	Repeticiones bootstrap	Región de confianza
90%	B=50	(286.722° , 335.430°)
	B=100	(297.617° , 324.534°)
	B=200	(299.255° , 322.897°)
	B=500	(307.136° , 315.015°)
95%	B=50	(298.913° , 323.239°)
	B=100	(290.178° , 331.974°)
	B=200	(302.150° , 320.002°)
	B=500	(307.613° , 314.539°)
99%	B=50	(300.741° , 321.411°)
	B=100	(303.691° , 318.461°)
	B=200	(309.807° , 312.345°)
	B=500	(310.747° , 311.405°)

Tabla 3: Resultados obtenidos al aplicar el método Duchame para regiones de confianza utilizando bootstrap con repeticiones $B = 50, 100, 200\ y\ 500$ a un nivel de confianza del $90,\ 95$ y 99%.

Método Duchame Pivotal

Nivel de confianza	Repeticiones bootstrap	Región de confianza
90%	B=50	(310.079° , 312.073°)
	B=100	(310.077° , 312.075°)
	B=200	(310.076° , 312.076°)
	B=500	(310.076° , 312.076°)
95%	B=50	(310.076° , 312.076°)
	B=100	(310.076° , 312.076°)
	B=200	(310.076° , 312.076°)
	B=500	(310.076° , 312.076°)
99%	B=50	(310.076° , 312.076°)
	B=100	(310.077° , 312.077°)
	B=200	(310.077° , 312.077°)
	B=500	(310.077° , 312.077°)

Tabla 4: Resultados obtenidos al aplicar el método Duchame Pivotal para regiones de confianza utilizando bootstrap con repeticiones $B = 50, 100, 200\ y\ 500$ a un nivel de confianza del $90,\ 95$ y 99%.

Durante el cálculo de las regiones de confianza para la dirección media del experimento, se obtuvieron resultados considerables dado el pequeño tamaño de la muestra, y aún así, en las tablas se puede observar que para distintos valores que toma el número B de repeticiones bootstrap y un mismo nivel de confianza, las regiones resultantes se volvieron conservativas y en otros casos la región de confianza es muy grande. Por lo anterior es que se omitieron los resultados obtenidos al considerar el número $B = 1000$ de repeticiones bootstrap, ya que realmente ya no esta proporcionando información alguna.

Las mejores regiones de confianza para la dirección media de escape de la tortuga en campo abierto son:

1. $(308° , 315°)$ **con el Método Gráfico**
2. $(310.76° , 312.076°)$ **con el Método Duchame Pivotal con** 90% **confianza y** $B = 500$**.**
3. $(310.76° , 312.076°)$ **con el Método Duchame Pivotal con** 95% **de confianza y** $B = 500$**.**
4. $(310.747° , 311.405°)$ **con el Método Duchame con** 99% **de confianza y** $B = 500$**.**

Conclusiones

El propósito de este trabajo ha sido el mostrar un campo más de investigación de la Estadística, en el cual se trabaja con conjuntos de datos direccionales y por lo cual a este campo se le denomina *Estadística Circular*.

Los conceptos y métodos utilizados en la Estadística Circular son en esencia los mismos que se tienen en la Estadística Usual, solo que estos han sido ajustados al espacio muestral en que se encuentran los datos direccionales, siendo el espacio muestral de los datos direccionales una esfera unitaria de dimensión p.

Asimismo, se utilizó una herramienta estadística llamada *Método Bootstrap* que es de gran ayuda cuando el análisis del problema en el que se está trabajando se ha vuelto complicado o cuando no existen fórmulas explícitas para el cálculo de varianzas, ya que al aplicar el método para la solución del problema, este se traslada a un espacio hipotético en el cual la inferencia en el problema es menos complicada y es posible obtener algún resultado.

Por la forma en que trabaja el método bootstrap durante la aplicación del mismo en la solución del problema, tanto como el número B de repeticiones bootstrap tienda a infinito las condiciones hipotéticas se aproximan más a las condiciones reales del problema, y el resultado que se obtenga acerca de la inferencia del parámetro desconocido será más parecido al valor real del parámetro.

Se propusieron métodos paramétricos y no-paramétricos que se apoyan en el método bootstrap, así como modificaciones de los mismos, para la Construcción de Regiones de Confianza para la dirección media y el eje polar medio. De dichas regiones se obtienen regiones de confianza que llegan a tener una velocidad de convergencia de $O(n^{-1})$ y $O(n^{-2})$, dadas las ventajas que se obtienen al utilizar el método bootstrap.

Del experimento que se expuso la dirección media de escape de la tortuga es $\hat{\mu} = 311.076°$ y teniendo en cuenta que al pedir mayor confiabilidad el tamaño de la región de confianza puede aumentar, la mejor región de confianza para la dirección media que se obtuvo fue: 1) $(310.76°, 312.076°)$ con el Método Duchame Pivotal con 90% confianza; 2) $(310.76°, 312.076°)$ con el Método Duchame Pivotal con 95% de confianza y 3) $(310.747°, 311.405°)$ con el Método Duchame con 99% de confianza.
La integración del método bootstrap en la implementación de programas computacionales para el cálculo de regiones de confianza permite distinguir la eficiencia y precisión entre los diferentes métodos para el cálculo de las mismas, lo cual es notorio al trabajar en los programas que se desarrollan.

Durante la ejecución y análisis de cada programa para determinar la región de confianza por cada método propuesto para la dirección media de los datos direccionales, se obtuvo además de las regiones, varios resultados que permiten concluir lo siguiente: 1) el método gráfico nos permite deducir la posible ubicación de la dirección media, pero no es del todo el mejor método para determinar una región de confianza; 2) el tamaño de la muestra de la población bajo estudio debe de ser de $n > 50$; 3) un inconvenientes de utilizar el Método Bootstrap es que como el número B de repeticiones tiende a infinito, en las repeticiones de se llega presentar varias veces eventos que son considerados como muy poco probables, lo que provoca que la frecuencia de estos eventos muy poco probables aumente y después de todo este evento ha dejado de ser tan poco probable como se creía, influyendo en el resultado directamente; 4)la dispersión de los datos influye directamente en el tamaño de la región de confianza; 5) una gran dispersión de los datos junto con un número B grande de repeticiones bootstrap aumenta la frecuencia de los eventos que son muy poco probables y provoca que la región de confianza aumente de tamaño; y 6) el método bootstrap es una herramienta estadística eficaz y eficiente, siendo de gran ayuda para la inferencia en problemas relacionados con la Estadística Circular.

Por último, esta tesis fue realizada con el fin de proponer y ampliar métodos para la construcción de regiones de confianza para la dirección media y eje polar medio de datos direccionales, esperando aportar un poco al conocimiento y a *La Ciencia*.

Anexos

Anexo A.- Argumento asintótico a favor del pivoteo

El problema va a ser tratado sobre un parámetro univariado θ, estimado por $\hat{\theta}$ y con varianza $n^{-1}\sigma^2$. Sea $\hat{\sigma}$ que es el estimador para σ. La distribución de $T = n^{1/2}(\hat{\theta} - \theta)$ y $S = n^{1/2}(\hat{\theta} - \theta)/\hat{\sigma}$ (sin estandarizar y estandarizada respectivamente) admite la expansión

$$P(T \leq x) = \Phi(x/\sigma) + n^{1/2} p(x)\phi(x/\sigma) + O(n^{-1}) \qquad \text{(A.1)}$$

y

$$P(S \leq x) = \Phi(x) + n^{1/2} q(x)\phi(x) + O(n^{-1}) \qquad \text{(A.2)}$$

donde Φ es la distribución normal estándar y ϕ es la función de densidad, p y q son polinomios de grado 2.

Los estimadores bootstrap de esas distribuciones están basadas en $T^* = n^{1/2}(\hat{\theta}^* - \hat{\theta})$ y $S^* = n^{1/2}(\hat{\theta}^* - \hat{\theta})/\hat{\sigma}^*$, y su expansión es

$$P(T^* \leq x \mid \times) = \Phi(x/\hat{\sigma}) + n^{1/2} \hat{p}(x)\phi(x/\hat{\sigma}) + O_p(n^{-1}) \qquad \text{(A.3)}$$

y

$$P(S^* \leq x \mid \times) = \Phi(x) + n^{1/2} \hat{q}(x)\phi(x) + O_p(n^{-1}) \qquad \text{(A.4)}$$

donde $\hat{p}$ y $\hat{q}$ son obtenidos de p y q después de utilizar los estimadores bootstrap, y $P(\cdot \mid \times)$ es la probabilidad condicional en la muestra.

Dado que $\hat{p} - p$, $\hat{q} - q$ y $\hat{\sigma} - \sigma$ tienen un error de $O_p(n^{-1/2})$, entonces $(A.1) - (A.4)$ da

$$P(T \leq x) - P(T^* \leq x \mid \times) = \Phi(x/\sigma) - \Phi(x/\hat{\sigma}) + O_p(n^{-1/2}) \qquad \text{(A.5)}$$

y

$$P(S \leq x) - P(S^* \leq x \mid \times) = O_p(n^{-1}) \qquad \text{(A.6)}$$

Para (A.5) y (A.6) podemos ver que la aproximación bootstrap a la distribución de S tiene un error de n^{-1} a diferencia de la aproximación normal que tiene un error de $n^{-1/2}$, pero la aproximación bootstrap en la distribución de T tiene un error de $n^{-1/2}$, ya que $\hat{\sigma}$ dista de σ en $n^{-1/2}$. Lo anterior hace explicito lo ventaja de el pivoteo.

Anexo B.- Distribución Asintótica de $T(\mathrm{m}_0)$

Considere la función $g(\mathrm{x}) = \mathrm{x} / \| \mathrm{x} \|$, donde x es un vector de dimensión p. Entonces $\hat{\mathrm{m}} = g(n^{-1}\sum \mathrm{X}_i)$, $\mathrm{m}_0 = g(E(\mathrm{X}))$ y $\frac{\partial g(\mathrm{x})}{\partial \mathrm{x}^{\mathrm{T}}} = \| \mathrm{x}\|^{-1}\{I_p - \mathrm{x}\mathrm{x}^{\mathrm{T}} / \| \mathrm{x}\|^2\}$, donde I_p es la matriz identidad de dimensión $p \times p$.

Sea $M_{(p)}$ una matriz de dimensión $(p-1)\times p$ que satisface $M_{(p)}M_{(p)}^T = I_{(p-1)}$ y $M_{(p)}\mathrm{m}_0 = 0$. Entonces por el teorema del límite central y el teorema de Taylor se tiene que

$$n^{1/2}M_{(p)}(\hat{\mathrm{m}} - \mathrm{m}_0) \longrightarrow N_{(p-1)}(0, G) \quad \text{(B.1)}$$

donde $G = M_{(p)}\mathrm{E}(\mathrm{X}\mathrm{X}^{\mathrm{T}})M_{(p)}^T / \|\mathrm{E}(\mathrm{X})\|^2$.

Si se reemplaza $M_{(p)}$ y G por sus bootstrap análogos $\hat{M}_{(p)}$ y $\hat{G}$, entonces siguiendo (B.1) se tiene que

$$T(\mathrm{m}_0) = n\mathrm{m}_0^{\mathrm{T}}\hat{M}_{(p)}^{\mathrm{T}}\hat{G}^{-1}\hat{M}_{(p)}\mathrm{m}_0 \xrightarrow{\mathrm{P}} \chi^2_{(p-1)}.$$

Anexo C.- Aproximación de Convergencia Teórica

A continuación se da la justificación teórica sobre la temprana afirmación concerniente a la de convergencia de las regiones de confianza que utilizan el pivoteo.

C.1 Supuestos

Sea F_0 la distribución de la población definida en la esfera unitaria $S_p = \{\mathrm{x} = (x^1, \ldots, x^p) : \mathrm{x}\mathrm{x}^{\mathrm{T}} = 1\}$. Los siguientes supuestos son acerca de F_0:

1. $F_0 = \lambda F_{ac} + (1-\lambda)F_S$ para alguna $\lambda \in (0,1]$, donde F_{ac} y F_S son distribuciones de probabilidad en S_p que son absolutamente continuas y singulares con respecto a la distribución uniforme de S_p.
2. Existe un subconjunto $A \subset S_p$ y una versión $f_{ac}(\mathrm{x})$ que es la densidad de F_{ac} tal que $f_{ac} > 0$ para todo $\mathrm{x} \in A$.
3. La relevante medida de centralidad de la población, ya sea la dirección media o el eje polar medio, está bien definida en F_0.

C.2 Valides de la expansión polinomial

Para $i = 1, \ldots, k$, sea $P_i(\mathrm{x})$ un polinomio en las componentes $x^1, \ldots, x^p$ de $\mathrm{x} \in S_p$. Se dice que $P_1, P_2, \ldots, P_k$ son polinomios independientes en S_p si $\sum_{i=1}^{k} \lambda_i P_i(\mathrm{x})$=constante para toda $\mathrm{x} \in S_p$ e implica que $\lambda_1 = \cdots = \lambda_k = 0$.

Lema. Suponer que F_0 satisface las condiciones 1 y 2 en *C.1*, y suponer que $\mathrm{X} \sim F_0$.

Entonces para cualquier colección finita de polinomios $P_1, P_2, \ldots, P_k$ en S_p son mutuamente independientes, la distribución común de $(P_1(\mathrm{x}), \ldots, P_k(\mathrm{x}))$, visto como un vector aleatorio de dimensión k, satisface la condición de Cramer, esto es:

$$\lim_{\|t\| \to \infty} \sup |\varkappa(t)| = \lim_{\|t\| \to \infty} \sup \left| \int_{S_p} \exp\{i \sum t_j P_j(\mathrm{x})\} dF_0 \right| < 1 \quad \text{(C.1)}$$

donde $t = (t_1, \ldots, t_k)$.

Demo:

Esto sigue fácilmente del lema 2.2 de Bhattacharya y Ghosh (1978).

El significado de (C.1) es el siguiente. Sea μ y V que denotan la media y la matriz de covarianza de $(P_1(\mathrm{x}),\ldots,P_k(\mathrm{x}))$ bajo F_0, y sea Q_n la distribución de probabilidad de

$$Z_n = n^{1/2}\mathrm{V}^{-1/2}\left\{n^{-1}\sum_{i=1}^{n}(P_1(\mathrm{x}_i),\ldots,P_k(\mathrm{x}_i))^{\mathrm{T}} - \mu\right\} \qquad \text{(C.2)}$$

donde $\mathrm{x}_1,\ldots,\mathrm{x}_n$ es una muestra aleatoria de F_0. Hay que ver que la condición de Cramer (C.1) implica que V es no-singular. Porque todos los momentos de Z_1 (definido en (C.2) con $n=1$) son finitos. Lo que sigue proviene de (C.1) y el corolario 20.4 de Bhattacharya y Rao (1976), que para cualquier entero $s>0$,

$$\sup_{A\in\mathcal{A}}\left|Q_n(A) - \int_A \left\{1+\sum_{j=1}^{s} r_j(y)\right\}\phi_k(y)dy\right| = O(n^{-(s+1)/2}) \quad \text{(C.3)}$$

En el contexto de r_j son polinomios en las componentes $y^1,\ldots,y^p$ de y cuyos coeficientes dependen de la acumulación de Z_1, $\phi_k(y)$ es la densidad estándar Gaussiana en k dimensiones, y $\mathcal{A}$ es una gran clase de conjuntos que no se pueden especificar (ver Bhattacharya y Rao 1976 para mayor detalle). El punto importante es que todas las probabilidades que se necesitan calcular están dadas de la forma $Q_n(A)$, $A\in\mathcal{A}$.

C.3 Aproximación por polinomios

La estadística $T(\mathrm{m}_0)$ y $W(\mathrm{m}_0)$ tienen una distribución asintótica $\chi^2_{(p-1)}$ en C.1. En el caso de $T(\mathrm{m}_0)$ ver Anexo B; y en el caso de $W(\mathrm{m}_0)$ se utiliza también el teorema de Wilk (ver Hall y Lascala 1990 o Owen 1990). Sea Y alguna de las estadísticas anteriores. Se considera la expansión estocástica para la raíz cuadrada de Y:

$$Y = R^T R$$

$$\text{donde } R = R_0 + n^{-1/2}R_1 + n^{-1}R_2 + O_p(n^{-3/2}) \quad \text{(C.4)}$$

donde las componentes $R_i^1,\ldots,R_i^{(p-1)}$ de R, con $i=0,1,2$, son todas funciones polinomiales de cantidades de la forma $n^{-1}\sum_{j=1}^{n}P(\mathrm{x}_j)$, donde $P(\mathrm{x})$ es un generador polinomial. (Ver DiCiccio 1991 para más detalles en el caso de $W(\mathrm{m}_0)$). Cuando $Y=T(\mathrm{m}_0)$ un expansión similar existe, con diferencias entre R_1 y R_2. Como sea, en ambos casos, la distribución asintótica de R_0 es normal estándar multivariada en $(p-1)$ dimensiones.

Hay que también mencionar que (C.4) se mantiene en un subconjunto del espacio muestral. Por ejemplo, en el caso de $T(\mathrm{m}_0)$, (C.4) se restringe a la forma

$$\left\{\left\|n^{-1}\sum_{i=1}^{n}\mathrm{X}_i\right\|>\varepsilon_1,\ \min_{u:u^Tu=1}u^TGu>\varepsilon_2\right\} \quad \text{(C.5)}$$

donde $\varepsilon_1>0$ y $\varepsilon_2>0$ son suficientemente pequeños. Como sea, teniendo los supuestos 1,2 y 3 de C.1 el complemento de (C.5) probabilidad muy pequeña cuando $\varepsilon_1>0$ y $\varepsilon_2>0$ son suficientemente pequeños.

Riguroso trabajo se realizó en el cálculo y aplicación del teorema de Taylor, desigualdad de Chebychev y desigualdad de Bernstein para obtener los resultados que necesitan para esta sección.

Usando los resultados en la sección anterior y en esta, y omitiendo fuertes detalles técnicos, se concluye que

$$\sup_{t\geq 0}\left|P(R\in tB_1)-\int_{tB_1}\{1+n^{-(1/2)}q_1(y)+n^{-1}q_2(y)+n^{-(3/2)}q_3(y)\}\phi_{(p-1)}(y)dy\right|=\sup_{t\geq 0}\left|P(R\in tB_1)-\int_{tB_1}\{1+n^{-1}q_2(y)\}\phi_{(p-1)}(y)dy\right|=O(n^{-2}) \quad \text{C.6)}$$

donde P es calculado bajo F_0, $B_1=\{y:y^Ty\leq 1\}$, $tB_1=\{ty:y\in B_1\}$, y q_1,q_2 y q_3 son polinomios en las componentes $y^1,\ldots,y^p$ de y con q_j para j. Los coeficientes de q_j dependen del mayor orden de cumulantes de R, incluyendo términos de orden $O(n^{-(1/2)})$ en $E(R)$ y términos de orden $O(n^{-1})$ en cov(R). La primera igualdad en (C.6) sigue de la propiedad de paridad de q_j mencionado antes, combinado con el factor de la región de integración tB_1, es simétrico alrededor del origen. La segunda igualdad de (C.6) implica que el poco frecuente $n^{-(1/2)}$ no aparezca en la expansión para $P(R\in tB_1)$.

C.4 Expansión para las cantidades bootstrap

Sea $X^* = \{X_1^*, \ldots, X_n^*\}$ una muestra generada por bootstrap, obtenida por el re-muestreo con reemplazo de la muestra original $X = \{X_1, \ldots, X_n\}$. Sea $R^* = R_0^* + n^{-(1/2)} R_1^* + n^{-1} R_2^*$ para la cantidad R en (C.4), pero basado en X^* en lugar de X. Si los supuestos 1,2 y 3 de C.1 se cumplen, entonces el bootstrap análogo de (C.6) se mantiene en el mismo sentido: para $\lambda > 0$, existe una constante $C < \infty$ tal que

$$\left\{ \sup_{t \geq 0} \left| P(R^* \in tB_1 \mid X) - \int_{tB_1} \{1 + n^{-(1/2)} q_1^*(y) + n^{-1} q_2^*(y) + n^{-(3/2)} q_3^*(y)\} \phi_{(p-1)}(y) dy \right| > Cn^{-2} \right\}$$

$$= P\left\{ \sup_{t \geq 0} \left| P(R^* \in tB_1 \mid X) - \int_{tB_1} \{1 + n^{-1} q_2^*(y)\} \phi_{(p-1)}(y) dy \right| > Cn^{-2} \right\} = O(n^{-\lambda}) \quad \text{(C.7)}$$

donde P es calculada bajo F_0 y $q_j^*(y)$ es $q_j(y)$ pero con la muestra de la población reemplazada.

C.5 Convergencia en probabilidad

Dada la expansión de la forma (C.6) y (C.7), a continuación se prueba la afirmación concerniente a la convergencia. Los puntos clave son los siguientes: si t_α se elige de tal forma que $P(R \in t_\alpha B_1) = \alpha$, entonces t_α tiene una expansión de la forma

$$t_\alpha = c_\alpha^{1/2} + n^{-1} h + O(n^{-2}) \qquad \text{(C.8)}$$

donde h depende en los cumulantes de R, Y $c_\alpha^{1/2}$ es tal que $P(\chi^2_{(p-1)} \leq c_\alpha) = \alpha$.

Para obtener (C.8) se realizó lo siguiente: la segunda igualdad en (C.6) implica que si t_α es tal que $P(R \in t_\alpha B_1) = \alpha$, entonces

$$\begin{aligned} \alpha &= \int_{t_\alpha B_1} \{1 + n^{-1} q_2(y)\} \phi_{(p-1)}(y) dy + O(n^{-2}) \\ &= P(\chi^2_{(p-1)} \leq t_\alpha^2) + n^{-1} g(t_\alpha) + O(n^{-2}) \end{aligned} \qquad \text{(C.9)}$$

donde $g(t_\alpha) = \int_{t_\alpha B_1} q_2(y) \phi_{(p-1)}(y) dy$. Finalmente se invirtió la expresión (C.9) para obtener la expansión deseada de t_α.

Un argumento similar, usando (C.7) mejor dicho (C.6), si t_α es tal que $P(R^2 \in t_\alpha^* B_1 / X) = \alpha$, entonces t_α^* tiene una expansión de la forma $t_\alpha^* = c_\alpha^{1/2} + n^{-1}h^* + O_p(n^{-2})$, donde h^* es el mismo que h pero con la población cumulantes reemplazada por la correspondiente muestra de cumulantes. Después $h^* = h = O_p(n^{-2})$.

Si la cantidad de bootstrap t_α^* es obtenida exactamente, entonces la convergencia en probabilidad está dada por

$$P(R \in t_\alpha^* B_1) = P((t_\alpha^*)^{-1} t_\alpha R \in t_\alpha B_1) = P(S \in t_\alpha B_1)$$

donde $S = \{1 - (nc_\alpha^{1/2})^{-1}(h^* - h) + O_p(n^{-2})\}R = \{1 + O_p(n^{-(3/2)})\}R$. Con un poco más de trabajo, lo siguiente puede ser mostrado: S tiene una expansión de la forma dada en la segunda línea de (C.6), pero con el polinomio q_2 en reemplazado por un polinomio diferente u_2, donde el coeficiente polinomial de este u_2 difiere de q_2 en términos de orden $O_p(n^{-(3/2)})$, porque $R - S = O_p(n^{-(3/2)})$. Colocando todos esos factores juntos, se concluye que $P(R \in t_\alpha^* B_1) = \alpha + O(n^{-2})$ (Ver Hall 1992 para más detalles de la afirmación anterior).

En práctica, la cantidad t_α^* es usualmente estimada mediante simulación y solo la aproximación satisface $P(R \in t_\alpha^* B_1) = \alpha$. Dado B^*, el número de el número de re-muestreos usados para estimar t_α^*, no incrementa despacio en proporción de n, el error de convergencia $O(n^{-2})$ se mantiene archivado. Un resultado dado por Hall (1986) indica que si α es un entero múltiplo de $(B^* + 1)^{-1}$, entonces la dependencia del error de convergencia en B^* será muy frágil.

Finalmente, el error de convergencia de las regiones por verosimilitud empírica obtenida sin una calibración bootstrap es de tamaño $O(n^{-1})$.

Anexo D.- Códigos

```
(*Código Método Gráfico*)
muestra = {316, 289, 230, 328, 257, 146, 241, 230, 232,
   207, 28, 163, 142, 213, 217, 317, 138, 169, 161, 123,
   182, 184, 31, 150, 351};
n = Length[muestra];
μ = N[1/n * {Sum[Cos[muestra[[i]]], {i, 1, n}], Sum[Sin[muestra[[i]]], {i, 1, n}]}];
normalμ = μ / Norm[μ];
medengradnegt = 180 * ArcTan[normalμ[[1]], normalμ[[2]]] / π;
medengrad = medengradnegt + 360;
Print["La dirección media estimada es ", medengrad, "°"]
Print["y todos los demás puntos son las B direcciones
medias obtenidas al aplicar el Método Bootstrap"]
B = 50;
remuestreo = muestra;
posicion[x_] := Random[ Integer, {1, n}]
s[x_] := Catch[{{vecttemp = Array[posicion, n], For[i = 0, i < n,
     {loc = vecttemp[[i]], remuestreo[[i]] = muestra[[loc]]}, i++]},
   {medtemp = N[1/n * {Sum[Cos[remuestreo[[i]]], {i, 1, n}], Sum[Sin[remuestreo[[i]]], {i, 1, n}]}],
    medtempnor = medtemp / Norm[medtemp]}, Throw[medtempnor]}]
medbootstrap = Array[s, B];
listamed = Insert[medbootstrap, μ, 1];
circ = Graphics[Circle[{0, 0}, 1], AspectRatio -> Automatic,
   Axes → Automatic];
Show[circ, ListPlot[listamed, PlotStyle → PointSize[0.02],
  AspectRatio -> Automatic, Axes → Automatic]]
MemoryInUse[]
Clear[muestra, n, μ, normalμ, medengradnegt, medengrad, B,
 remuestreo, vecttemp, medtemp, medtempnor, medbootstrap,
 listamed, circ]
```

```
(*Código de Región de Confianza Básica*)
muestra = {316, 289, 230, 328, 257, 146, 241, 230, 232,
   207, 28, 163, 142, 213, 217, 317, 138, 169, 161, 123,
   182, 184, 31, 150, 351};
n = Length[muestra];
μ = N[1/n * {Sum[Cos[muestra[[i]]], {i, 1, n}], Sum[Sin[muestra[[i]]], {i, 1, n}]}];
normalμ = μ / Norm[μ];
medengradnegt = 180 * ArcTan[normalμ[[1]], normalμ[[2]]] / π;
medengrad = medengradnegt + 360;
α = .1;
B = 50;
remuestreo = muestra;
posicion[x_] := Random[ Integer, {1, n}]
s[x_] := Catch[{{vecttemp = Array[posicion, n], For[i = 0,
     i < n, {loc = vecttemp[[i]], remuestreo[[i]] = muestra[[loc]]}, i++]},
   {medtemp = N[1/n * {Sum[Cos[remuestreo[[i]]], {i, 1, n}], Sum[Sin[remuestreo[[i]]], {i, 1, n}]}],
    medtempnor = medtemp / Norm[medtemp]}, Throw[medtempnor]}]
medbootstrap = Array[s, B];
difs[x_] := Catch[{{radtemp = ArcTan[medbootstrap[[x]][[1]],
      medbootstrap[[x]][[2]]],
    medbootengrad = 180 * radtemp / π,
    difmed = medbootengrad - medengrad}, {Throw[difmed]}}]
meddifs = Array[difs, B];
For[i = 0, i < B, meddifs[[i]] = meddifs[[i]] + 360, i++]
medord = Sort[meddifs];
l = IntegerPart[11/2 * B * α + 1/2];
m = B - l;
If[l == 0, y_l = medord[[1]], y_l = medord[[l + 1]]];
y_m = medord[[m]];
regionbasica = {medengrad + y_l, medengrad + y_m};
If[regionbasica[[1]] > 360 || regionbasica[[2]] > 360,
  If[regionbasica[[1]] > 360, If[regionbasica[[2]] > 360,
    regionbasica = regionbasica - 360,
    regionbasica[[1]] = regionbasica[[1]] - 360],
   If[regionbasica[[2]] > 360,
    regionbasica[[2]] = regionbasica[[2]] - 360, False]], False];
If[regionbasica[[2]] < regionbasica[[1]], {temp = regionbasica[[1]],
   regionbasica[[1]] = regionbasica[[2]],
   regionbasica[[2]] = temp}, False];
Print["La región básica con ", (1 - α) * 100, "% de confianza
para la dirección media es ", regionbasica]
MemoryInUse[];
Clear[muestra, n, μ, normalμ, medengradnegt, medengrad, B,
 remuestreo, vecttemp, medtemp, medtempnor, medbootstrap,
 medbootengrad, difmed, meddifs, medord, l, m, regionbasica, temp]
```

```
(*Código de Región de Confianza Clase General*)
muestra = {316, 289, 230, 328, 257, 146, 241, 230, 232, 207, 28, 163, 142, 213, 217,
   317, 138, 169, 161, 123, 182, 184, 31, 150, 351};
n = Length[muestra]; μ = N[1/n * {Sum[Cos[muestra[[i]]], {i, 1, n}], Sum[Sin[muestra[[i]]], {i, 1, n}]}];
normalμ = μ / Norm[μ];
medengradnegt = 180 * ArcTan[normalμ[[1]], normalμ[[2]]] / π;
medengrad = medengradnegt + 360
origD = N[1/n * Sum[({N[Cos[muestra[[i]]]], N[Sin[muestra[[i]]]]}
        - normalμ).({N[Cos[muestra[[i]]]], N[Sin[muestra[[i]]]]} - normalμ), {i, 1, n}]];
α = .1; B = 50;
remuestreo = muestra;
posicion[x_] := RandomInteger[{1, n}]
s[x_] := Reap[Catch[{{vecttemp = Array[posicion, n], For[i = 0, i < n,
      {loc = vecttemp[[i]], remuestreo[[i]] = muestra[[loc]]}, i++]},
    {medtemp = N[1/n * {Sum[Cos[remuestreo[[i]]], {i, 1, n}], Sum[Sin[remuestreo[[i]]], {i, 1, n}]}],
     medtempnor = medtemp / Norm[medtemp]},
    {Dcov = 1/n * Sum[({N[Cos[remuestreo[[i]]]], N[Sin[remuestreo[[i]]]]} - medtempnor).
          ({N[Cos[remuestreo[[i]]]], N[Sin[remuestreo[[i]]]]} - medtempnor),
     Sow[medtempnor], Throw[Dcov]}}]]
listDyMedboot = Array[s, B];
listDtemp[x_] := Extract[listDyMedboot[[x]], 1] listD = Array[listDtemp, B];
listMedboottemp[x_] := Extract[listDyMedboot[[x]], 2]
listMedboot = Array[listMedboottemp, B];
listMed[x_] := Extract[listMedboot[[x]], 1]
listM = Array[listMed, B]; listmedfinal[x_] := Extract[listM[[x]], 1]
listmedbootstrap = Array[listmedfinal, B];
Y[x_] := Catch[{Ytemp = normalμ - (Sqrt[origD / listD[[x]]] * (listmedbootstrap[[x]] - normalμ)),
   Ynor = Ytemp / Norm[Ytemp], Throw[Ynor]}]
listYnor = Array[Y, B];
Yengrad[x_] := ArcTan[listYnor[[x]][[1]], listYnor[[x]][[2]]] * (180 / π)
listYengrad = Array[Yengrad, B];
listYengradord = Sort[listYengrad];
l = IntegerPart[1/2 * B * α + 1/2]; m = B - l;
If[l == 0, Ycero_l = listYengradord[[1]], Ycero_l = listYengradord[[l + 1]]];
Ycero_m = listYengradord[[m]]; regionclasegeneral = {Ycero_l + 360, Ycero_m + 360};
Print["La región clase general con ", (1 - α) * 100, "% de confianza para la
dirección media es ", regionclasegeneral]
MemoryInUse[];
Clear[muestra, n, μ, normalμ, medengradnegt, medengrad, origD, α, B, remuestreo,
 vecttemp, medtemp, medtempnor, Dcov, listDyMedboot, listD, listMedboot, listM,
 listmedbootstrap, listYnor, listYengrad, listYengradord, l, m, regionclasegeneral]
```

```
(*Código de Región de Confianza Duchame*)
muestra = {316, 289, 230, 328, 257, 146, 241, 230, 232, 207, 28, 163,
   142, 213, 217, 317, 138, 169, 161, 123, 182, 184, 31, 150, 351};
n = Length[muestra];
μ = N[1/n * {Sum[Cos[muestra[[i]]], {i, 1, n}], Sum[Sin[muestra[[i]]], {i, 1, n}]}];
normalμ = μ / Norm[μ];
medengradnegt = 180 * ArcTan[normalμ[[1]], normalμ[[2]]] / π;
medengrad = medengradnegt + 360; α = .1; B = 50;
remuestreo = muestra; posicion[x_] := Random[ Integer, {1, n}]
s[x_] := Catch[{{vecttemp = Array[posicion, n], For[i = 0, i < n,
     {loc = vecttemp[[i]], remuestreo[[i]] = muestra[[loc]]}, i++]},
   {medtemp = N[1/n * {Sum[Cos[remuestreo[[i]]], {i, 1, n}], Sum[Sin[remuestreo[[i]]], {i, 1, n}]}],
    medtempnor = medtemp / Norm[medtemp]}, Throw[medtempnor]}]
listmedbootstrap = Array[s, B];
bootgrad[x_] := ArcTan[listmedbootstrap[[x]][[1]],
   listmedbootstrap[[x]][[2]]] * {180 / π}
listmedbootstrapengrad = Array[bootgrad, B];
For[i = 0, i < B, If[listmedbootstrapengrad[[i]][[1]] < 0,
  listmedbootstrapengrad[[i]] = listmedbootstrapengrad[[i]] + 360,
  If[listmedbootstrapengrad[[i]] > 360,
   listmedbootstrapengrad[[i]] =
    listmedbootstrapengrad[[i]] - 360, False]], i++]
listmedbootstrapengradord = Sort[listmedbootstrapengrad];
S[θ_] := Catch[{
   difmed = listmedbootstrapengradord[[θ]] - medengrad,
   tempS = n * {1 - Cos[difmed]}, Throw[tempS]}]
listestS = Array[S, B]; listestSord = Sort[listestS];
alfaS = listestSord[[α * 100]]; alfaθ = ArcCos[1 - alfaS/n] * 180 / π;
regionduchame = Catch[If[alfaθ[[1]][[1]] < 0, Throw[{medengrad + alfaθ[[1]][[1]],
      medengrad - alfaθ[[1]][[1]]}], Throw[{medengrad - alfaθ[[1]][[1]],
      medengrad + alfaθ[[1]][[1]]}]]];
Print["La región duchame con ", (1 - α) * 100, "% de confianza para
 la dirección media es ", regionduchame]
MemoryInUse[];
Clear[muestra, n, μ, normalμ, medengradnegt, medengrad, B, remuestreo,
 vecttemp, medtemp, medtempnor, listmedbootstrap, listmedbootstrapengrad,
 listmedbootstrapengradord, difmed, tempS, listestS, listestSord,
 alfaS, alfaθ, regionduchame]
```

```
(*Código de Región Duchame Pivotal*)
muestra = {316, 289, 230, 328, 257, 146, 241, 230, 232, 207, 28, 163,
   142, 213, 217, 317, 138, 169, 161, 123, 182, 184, 31, 150, 351};
n = Length[muestra]; μ = N[1/n * {Sum[Cos[muestra[[i]]], {i, 1, n}], Sum[Sin[muestra[[i]]], {i, 1, n}]}];
normalμ = μ / Norm[μ];
medengradnegt = 180 * ArcTan[normalμ[[1]], normalμ[[2]]] / π;
medengrad = medengradnegt + 360;
L̂ = N[Sqrt[(Sum[Cos[muestra[[i]]], {i, 1, n}])^2 + (Sum[Sin[muestra[[i]]], {i, 1, n}])^2]];

σ̂ = (1 - n^-1 Sum[Cos[2 * (muestra[[i]] - medengrad)], {i, 1, n}])^2 / (4 * L̂^2);
α = .1;
B = 50;
remuestreo = muestra; posicion[x_] := Random[ Integer, {1, n}]
s[x_] := Catch[{{vecttemp = Array[posicion, n], For[i = 0, i < n,
     {loc = vecttemp[[i]], remuestreo[[i]] = muestra[[loc]]}, i++]},
   {medtemp = N[1/n * {Sum[Cos[remuestreo[[i]]], {i, 1, n}], Sum[Sin[remuestreo[[i]]], {i, 1, n}]}],
    medtempnor = medtemp / Norm[medtemp]}, Throw[medtempnor]}]
listmedbootstrap = Array[s, B];
bootgrad[x_] := ArcTan[listmedbootstrap[[x]][[1]], listmedbootstrap[[x]][[2]]]
* (180 / π)
listmedbootstrapengrad = Array[bootgrad, B];
For[i = 0, i < B, If[listmedbootstrapengrad[[i]] < 0, listmedbootstrapengrad[[i]]
   = listmedbootstrapengrad[[i]] + 360, If[listmedbootstrapengrad[[i]] > 360,
   listmedbootstrapengrad[[i]] = listmedbootstrapengrad[[i]] - 360, False]], i++]
listmedbootstrapengradord = Sort[listmedbootstrapengrad];
S[θ_] := Catch[{
   difmed = listmedbootstrapengradord[[θ]] - medengrad, tempS =
    n * (1 - Cos[difmed]), Throw[tempS]}]
listestS = Array[S, B]; T[θ_] := listestS[[θ]] / σ̂
listestT = Array[T, B]; listestTord = Sort[listestT];
alfaT = listestTord[[α * 100]]; delta = (1 - σ̂^2 * alfaT / n);
regionduchamepivotal = {medengrad - delta, medengrad + delta};
Print["La región Duchame Pivotal con ", (1 - α) * 100, "% de confianza para
 la dirección media es ", regionduchamepivotal]
MemoryInUse[];
Clear[muestra, μ, normalμ, medengradnegt, medengrad, B, remuestreo, vecttemp,
 medtemp, medtempnor, listmedbootstrap, listmedbootstrapengrad,
 listmedbootstrapengradord, difmed, tempS, listestS, listestT, listestTord,
 alfaT, delta, regionduchamepivotal]
```

Bibliografía

- Batschelet, E. (1981). Circular Statistics in Biology. Academic Press, London.
- Bhattacharya, G. K. and Jhonson, R. A. (1977). Statistical Concepts and Methods. Wiley, New York.
- Efron, B. and Tibsshirani, R. (1993). An Introducition to the Bootstrap. Chapman and Hall, New York.
- Fisher, N. I. (1993). Statistical Analysis of Circular Data. University Press, Cambrige.
- Fisher, N. I. and Hall, P. (1989). "Bootstrap Confidence Regions for Directional Data". *Journal of the American Statistical Association.* **84**, 408.
- Fisher, N. I., Hall, P., Jing, B. and Wood, A. (1996) "Improved Pivotal Methods for Constucting Confidence Regions with Directional Data". *Journal of the American Statistical Association.* **91**, 435.
- Mardia, K. V. (1975). Statistics of Directional Data. Academic Press, London.
- Mardia, K. V. and Jupp, P. E. (2000). Directional Statistics. John Wiley & Sons, LTD.