Universidad de San Carlos de Guatemala
Escuela de Ciencias Físicas y Matemáticas
Departamento de Matemática

TEORÍA DE PROBABILIDADES Y MÁXIMA VEROSIMILITUD APLICADA AL PROBLEMA DE ESTIMACIÓN DEL TAMAÑO DE UNA POBLACIÓN

José Daniel Romero

Asesorado por Lic. William Roberto Gutiérrez Herrera

Guatemala, febrero del 2020

UNIVERSIDAD DE SAN CARLOS DE GUATEMALA

ESCUELA DE CIENCIAS FÍSICAS Y MATEMÁTICAS

TEORÍA DE PROBABILIDADES Y MÁXIMA VEROSIMILITUD APLICADA AL PROBLEMA DE ESTIMACIÓN DEL TAMAÑO DE UNA POBLACIÓN

TRABAJO DE GRADUACIÓN
PRESENTADO A LA JEFATURA DEL
DEPARTAMENTO DE MATEMÁTICA
POR

JOSÉ DANIEL ROMERO

ASESORADO POR LIC. WILLIAM ROBERTO GUTIÉRREZ HERRERA

AL CONFERÍRSELE EL TÍTULO DE

LICENCIADO EN MATEMÁTICA APLICADA

GUATEMALA, FEBRERO DEL 2020

UNIVERSIDAD DE SAN CARLOS DE GUATEMALA

ESCUELA DE CIENCIAS FÍSICAS Y MATEMÁTICAS

CONSEJO DIRECTIVO

DIRECTOR	M.Sc. Jorge Marcelo Ixquiac Cabrera
SECRETARIO ACADÉMICO	M.Sc. Edgar Anibal Cifuentes Anléu

TRIBUNAL QUE PRACTICÓ EL EXAMEN GENERAL PRIVADO

EXAMINADOR	Licda. Mariela Lizeth Benavides Lázaro
EXAMINADOR	Lic. William Roberto Gutiérrez Herrera
EXAMINADOR	Lic. Rafael Alejandro Martínez Márquez

USAC TRICENTENARIA Universidad de San Carlos de Guatemala

ECFM ESCUELA DE CIENCIAS FÍSICAS Y MATEMÁTICAS

Escuela de Ciencias Físicas y Matemáticas

Ref. D.DTG. 001-2020
Guatemala 19 de febrero de 2020

El Director de la Escuela de Ciencias Físicas y Matemáticas de la Universidad de San Carlos de Guatemala, luego de conocer la aprobación por parte del Coordinador de la Licenciatura en Matemática Aplicada, al trabajo de graduación Titulado: **"TEORÍA DE PROBABILIDADES Y MÁXIMA VEROSIMILITUD APLICADA AL PROBLEMA DE ESTIMACIÓN DEL TAMAÑO DE UNA POBLACIÓN"** presentado por el estudiante universitario **José Daniel Romero,** autoriza la impresión del mismo.

IMPRÍMASE.

"ID Y ENSEÑAD A TODOS"

M.Sc. Jorge Marcelo Ixquiac Cabrera
Director

ESCUELA DE CIENCIAS FÍSICAS Y MATEMÁTICAS · DIRECTOR · ECFM - USAC

200312867
JMIC/Paola

AGRADECIMIENTOS

A mi madre: Angélica Romero, por su amor incondicional, por su lucha constante en la vida y por la gran visión de proveernos educación a mí y a mis hermanos ante cualquier circunstancia.

A mis hermanos: Vinicio (†), Axel y Wendy, por su apoyo y acompañamiento en todas las etapas de mi vida.

A mi esposa: Evelyn, por su amor, su lealtad, su confianza y apoyo en los momentos alegres y en los momentos difíciles.

A mis asesores: Lic. William Gutiérrez y Dr. Antonio Murillo Salas, por haberme brindado su tiempo y sus conocimientos.

A la Universidad de San Carlos de Guatemala: Por brindarme acceso a la educación superior y por permitirme estudiar la carrera que tanto anhelaba.

A la Universidad de Guanajuato, México: Por su apoyo en la realización de este trabajo.

A mis amigos: Por permitirme compartir con ellos nuestras aficiones en común.

A mi maestro: Dr. Edgardo Cáceres por brindarme su tiempo, su conocimiento, su apoyo y mostrarme un nuevo mundo y pasión a través del juego de Go.

Doña Zoilita (†) y don Ramiro (†): Por su cariño, por contarme historias fantásticas, poemas de la vida y sueños de años pasados.

A mis maestros: A aquellos que me brindaron su pasión por el conocimiento, por explorar, por vivir.

A los defensores: Aquellos que han luchado por la educación pública, en la cual me formé, aquellos que han logrado que la educación no sea privilegio exclusivo de los más afortunados.

DEDICATORIA

A una mujer visionaria, inquebrantable, luchadora, llena de amor, a ti madre.
A mis hijas Ana Paula y Melina.

ÍNDICE GENERAL

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

LISTA DE SÍMBOLOS

Símbolo	Significado
σ-álgebra	sigma álgebra
$\in$	pertenece
E^c	complemento de E
$\mathbb{Z}^+$	números enteros positivos
$\cup$	unión entre conjuntos
$\emptyset$	conjunto vacío
$:=$	es definido por
$\mathcal{P}(X)$	conjunto potencia de X
$\cap$	intersección entre conjuntos
$(X, \mathscr{X})$	espacio topológico
$\mathbb{R}$	números reales
f^{-1}	aplicación inversa de f
$\mapsto$	mapeo
$\sum$	suma
Pr	probabilidad
$\mathscr{S}$	espacio muestral
$\mathfrak{S}$	σ-álgebra de eventos
$(\mathscr{S}, \mathfrak{S}, \Pr)$	espacio de probabilidad
$\circ$	composición
$\subset$	contenido en
$\binom{n}{y}$	combinaciones de n en y
${}_N P_r$	permutaciones de N en r
$\int$	integral
$\lim_{n\to\infty}$	límite cuando n tiende a infinito
μ	media poblacional
σ	desviación estándar poblacional

OBJETIVOS

General

Dar una introducción a los conceptos principales de la teoría de máxima verosimilitud, así como su aplicación a la estimación del tamaño de una población, utilizando el método de captura-recaptura.

Específicos

1. Introducir los conceptos y definiciones modernas utilizadas en la teoría de probabilidades, como lo son σ-álgebra, medidas, espacios medibles y espacios de probabilidad.

2. Introducir los conceptos y propiedades básicas de la teoría de estimación, así como brindar ejemplos prácticos de su uso.

3. Introducir los conceptos y definiciones principales utilizados en teoría de máxima verosimilitud, así como sus propiedades básicas.

4. Plantear, desarrollar y resolver el problema de la estimación del tamaño de una población con el método de captura-recaptura utilizando la función de verosimilitud y la función log-verosimilitud.

INTRODUCCIÓN

La teoría de probabilidades tiene sus inicios en el estudio de los juegos de azar, pues se deseaba determinar que decisiones debían tomarse para ganar en el juego, o bien como debía repartirse una apuesta en un juego si este era interrumpido, a este último se le conoce como el *problema del reparto de apuestas*, los primeros matemáticos importantes en abordar el tema fueron Tartaglia y Cardano en el siglo XVI, Cardano escribe un manual para jugadores donde da recomendaciones acerca de como evitar que los jugadores hagan trampa y también toca el tema de las probabilidades en estos juegos de azar desde un punto de vista de la teoría de probabilidades clásica, además da una solución al *problema del reparto de apuestas* que termina siendo incorrecta; tiempo después surge una importante correspondencia entre Pascal y Fermat en el siglo XVII, donde trataban la resolución de juegos de dados y nuevamente el *problema de repartición de apuestas*, dándole ambos una correcta solución a este último para el caso de dos jugadores; otro matemático importante que se dedicó a tocar el tema de la probabilidad en los juegos de azar fue Huygens, quien generalizaría la solución del *problema del reparto de apuestas* e introduciría el concepto de *esperanza matemática* (véase Salinero [13]).

Otros personajes como Graunt, Petty y Edmund Halley usarían la probabilidad para tocar temas demográficos en el mismo siglo; Jakob Bernoulli (1654-1705) da por primera vez la definición de probabilidad clásica en su libro *El Arte de Predecir* que se publicó póstumamente en el año 1713 y es Daniel Bernoulli en el siglo XVIII quien utiliza por primera vez el método de *máxima verosimilitud* para estimar un parámetro; aunque el concepto de *variable aleatoria* ya lo habían trabajado todos los anteriormente mencionados de una forma intuitiva, fue Poisson en 1832 quien dio la idea de la definición y se dio cuenta de la importancia de hacerlo.

En el siglo XX muchos biólogos y ecólogos empiezan a utilizar la máxima verosimilitud para estimar el tamaño de una población de animales (N) realizando métodos de captura-recaptura, al respecto hay varios artículos como los de J.N. Darroch y D. Ratcliff [4], C.C. Craig [3] y más recientemente Calambokidis [1] y

Soisalo [15], entre otros.

El primer capítulo de este trabajo está dedicado a la definición de los conceptos básicos y formales de la teoría de probabilidad, comenzando con la definición de σ-álgebra, espacios medibles, espacios muestrales, conjuntos medibles, funciones de probabilidad y espacios de probabilidad los cuales son usados en los capítulos posteriores, asimismo se dan ejemplos de las distribuciones binomial y normal las cuales son distribuciones discreta y continua respectivamente, luego se definen los conceptos de convergencia de funciones de probabilidad para finalmente enunciar y demostrar dos importantes teoremas de convergencia como lo son el *Teorema de continuidad de Lévy* y el *Teorema del Límite Central*, este último es uno de los teoremas más importantes sino el más importante de la teoría de probabilidad y la estadística.

El segundo capítulo está dedicado a desarrollar los conceptos básicos en teoría de estimación así como ejemplos de cada uno de ellos, como lo son; estimadores puntuales, estimadores por intervalo, estimadores sesgados e insesgados, también propiedades importantes como eficiencia relativa, consistencia e intervalos de confianza; el capítulo se cierra dando un ejemplo de la utilización de los intervalos de confianza para estimar la media de un conjunto de datos.

El tercer capítulo se centra en desarrollar los conceptos básicos de la teoría de máxima verosimilitud, definiendo primero la *función de verosimilitud* para el caso discreto y luego generalizándola para el caso continuo, se define la función *log-verosimilitud* la *función relativa de verosimilitud* y sus propiedades más importantes, se dan ejemplos prácticos del uso de estos conceptos para estimar parámetros de interés de poblaciones en las cuales supondremos que conocemos su función de distribución de probabilidad y se define el importante concepto de *estimador de máxima verosimilitud* (EMV).

El cuarto y último capítulo está dedicado a la aplicación de la teoría de máxima verosimilitud al método de captura-recaptura, el cual se utiliza para estimar el tamaño de una población, se define el problema a resolver, se deduce la función de verosimilitud correspondiente y con la ayuda de un lenguaje de programación se desarrolla un código que genera los estimadores de máxima verosimilitud para diferentes escenarios, asimismo se realiza una comparación entre la ventaja de utilizar la función de verosimilitud y la función log-verosimilitud; finalmente se realiza una comparación entre los resultados de este trabajo y los de un artículo que estima una población de mariposas.

1. Teoría de probabilidades

> Cualquier teoría matemática debe considerar tres aspectos fundamentales, a) el contenido formal lógico b) la intuición de trasfondo y c) sus aplicaciones; el carácter y el encanto de la estructura completa no pueden ser apreciados sin considerar estos aspectos en la relación correcta. (Feller, William: vol. 1. p. 1)

Al seguir esta línea se discutirán estos tres aspectos fundamentales. La teoría de probabilidades tiene sus inicios en el estudio de los juegos de azar al tratar de determinar cuales eran los posibles resultados en un juego dado y más aún, cuándo y cómo apostar para obtener la mayor ganancia posible, los primeros matemáticos importantes en abordar el tema fueron Tartaglia y Cardano alrededor del siglo XVI, más adelante en el siglo XVII Pascal y Fermat mantuvieron una constante correspondencia resolviendo problemas de repartición de apuestas en juegos de azar, Huygens en el mismo siglo sigue la línea de estos últimos para dar respuestas correctas a problemas abiertos hasta entonces, sobre tiradas de dados y el de repartición de apuestas.

A partir de lo anterior podemos decir de forma intuitiva que la teoría de probabilidades tiene como objetivo determinar los posibles resultados y la medida de ocurrencia que se obtendrá de ellos en un experimento aleatorio, esto es, un experimento en el cual no se tiene certeza acerca del resultado. Claro está que la intuición tiene sus limitaciones, de hecho grandes matemáticos fallaron al resolver problemas concernientes a la probabilidad precisamente por dejarse llevar por su intuición y es ahí donde surge la necesidad de desarrollar la teoría de una forma rigurosa. Algunos de los conceptos más importantes y modernos en teoría de probabilidades como lo son σ-álgebra , medida de probabilidad, espacios de probabilidad y variables aleatorias que fueron desarrollados entre los siglos XIX y XX se definirán a continuación.

1.1. σ-álgebras

Definición 1.1. Una familia $\mathfrak{M}$ de subconjuntos de X es una σ-álgebra sobre X, si cumple que:

1. $X \in \mathfrak{M}$
2. Si $E \in \mathfrak{M}$ entonces $E^c \in \mathfrak{M}$
3. Si $E_i \in \mathfrak{M}$, $i \in \mathbb{Z}^+$ entonces
$$\bigcup_{i=1}^{\infty} E_i \in \mathfrak{M}.$$

A la pareja $(X, \mathfrak{M})$ le llamaremos **espacio medible**, a los miembros de $\mathfrak{M}$ les llamaremos **conjuntos medibles**.

Nota. Si la condición (3) únicamente se cumple para colecciones finitas $E_i, \ldots, E_n \in \mathfrak{M}$ es decir,
$$\bigcup_{i=1}^{n} E_i \in \mathfrak{M}$$
diremos que $\mathfrak{M}$ es un **álgebra**.

De la definición se sigue que:

1. Si $X \in \mathfrak{M}$ entonces $\emptyset = X^c \in \mathfrak{M}$.
2. $\mathfrak{M} := \{\emptyset, X\}$ es una σ-álgebra , asimismo.
3. El conjunto potencia $\mathfrak{M} := \mathcal{P}(X)$ es una σ-álgebra , y son llamadas σ-álgebras triviales.

Podemos además deducir fácilmente el siguiente

Lema 1.1. *La intersección de cualquier colección de σ-álgebras sobre X es una σ-álgebra sobre X.*

A partir de lo anterior se puede demostrar el siguiente

Lema 1.2. *Sea $\mathcal{F}$ una familia de subconjuntos de X, entonces existe una σ-álgebra $\mathfrak{M}^*$ minimal de X tal que $\mathcal{F} \subseteq \mathfrak{M}^*$.*

Demostración. Sea I cualquier subconjunto indexado no vacío (finito, infinito contable o infinito no contable) y sean $\mathfrak{M}_i$ con $i \in I$, el conjunto de todas las σ-álgebras que contienen a $\mathcal{F}$, definamos entonces

$$\mathfrak{M}^* = \bigcap_{i \in I} \mathfrak{M}_i$$

por el lema anterior tenemos que $\mathfrak{M}^*$ es una σ-álgebra la cual obviamente contiene a $\mathcal{F}$. Además de la definición de $\mathfrak{M}^*$ se deduce su minimalidad y unicidad por lo cual queda demostrado. □

Definición 1.2. A la σ-álgebra $\mathfrak{M}^*$ le llamaremos σ-álgebra generada por $\mathcal{F}$ y escribiremos $\mathfrak{M}^* = \mathfrak{M}^*(\mathcal{F})$.

Definición 1.3. Sea $(X, \mathscr{X})$ un espacio topológico entonces a $\mathfrak{B} := \mathfrak{M}^*(\mathscr{X})$ le llamaremos **σ-álgebra de Borel**, a sus elementos les llamaremos **conjuntos de Borel**.

Ejemplo 1.1. Consideremos $(\mathbb{R}, \mathscr{R})$ el conjunto de los números reales con la topología natural (generada por los intervalos abiertos), entonces $\mathfrak{B} := \sigma(\mathscr{R})$ es σ-álgebra de Borel y los intervalos abiertos, cerrados, semicerrados o semiabiertos son conjuntos de Borel, asimismo en $(\mathbb{R}^k, \mathscr{R}^k)$ tenemos $\mathfrak{B}^k := \sigma(\mathscr{R}^k)$.

Definición 1.4. Consideremos $f\colon (X, \mathfrak{A}) \to (Y, \mathfrak{D})$ una aplicación entre espacios medibles. Decimos que f es una **función medible** si para cada $T \in \mathfrak{D}$ se cumple que $f^{-1}(T) \in \mathfrak{A}$ es decir $f^{-1}(\mathfrak{D}) \subseteq \mathfrak{A}$.

Teorema 1.1.1. *Sea $\mathfrak{D} = \mathfrak{D}(\mathscr{C})$, σ-álgebra generada por $\mathscr{C}$, $\mathscr{C}$ una familia de subconjuntos de Y. Entonces $f\colon (X, \mathfrak{A}) \to (Y, \mathfrak{D})$ es medible si y solo si $f^{-1}(\mathscr{C}) \subseteq \mathfrak{A}$.*

Demostración. Supongamos que f es medible, entonces de la definición se sigue que $f^{-1}(\mathfrak{D}) \subseteq \mathfrak{A}$, en particular ya que $\mathscr{C} \subseteq \mathfrak{D}$, tendremos que $f^{-1}(\mathscr{C}) \subseteq \mathfrak{A}$, por otro lado, supongamos que se cumple $f^{-1}(\mathscr{C}) \subseteq \mathfrak{A}$, ya que $\mathfrak{D}$ es generada por $\mathscr{C}$ tendremos que para cada $T \in \mathfrak{D}$, $f^{-1}(T) \in \mathfrak{A}$, de donde $f^{-1}(\mathfrak{D}) \subseteq \mathfrak{A}$. □

Definición 1.5. Sea $(X, \mathfrak{A})$ un espacio medible y $(Y, \mathscr{Y})$ espacio topológico, y sea $f\colon X \to Y$. Si $f^{-1}(B) \in \mathfrak{A}$ para cada $B \in \mathscr{Y}$ decimos que f es **Borel-medible** o **B-medible**.

1.2. Medidas

Las llamadas «funciones ordinarias» asignan puntos $y \in Y$ a puntos $x \in X$ *i.e.* $x \mapsto y = f(x)$ por otro lado tendremos las «funciones de conjuntos» que asignan puntos a elementos de una clase de conjuntos.

$$A \mapsto y = f(A) \in Y$$

Definición 1.6. Una **medida** m es una función de conjuntos de valor real, no negativa y contablemente aditiva definida sobre una σ-álgebra tal que $m(\emptyset) = 0$.

Recordemos que, si $\mathfrak{A}$ es σ-álgebra , $m\colon \mathfrak{A} \to [0, \infty]$ y si $(E_i) \subset \mathfrak{A}$ sucesión de conjuntos disjuntos entonces

$$m\Big(\bigcup_{i=1}^{\infty} E_i\Big) = \sum_{i=1}^{\infty} m(E_i).$$

Nota. Recordar que dos eventos A y B se dicen que son excluyentes si $A \cap B = \emptyset$ y se tiene que $\Pr(A \cup B) = \Pr(A) + \Pr(B)$.

Para evitar trivialidades asumiremos que existe $A \in \mathfrak{A}$ tal que $m(A) < \infty$.

Definición 1.7. Una medida es **totalmente finita** si $m(X) < \infty$, en donde $(X, \mathfrak{A})$ es un espacio medible.

1.3. Espacio de probabilidad

Ahora definiremos conceptos importantes acerca de los espacios de probabilidad.

Definición 1.8. Un **experimento** es la recreación de un fenómeno para obtener una medida.

Definición 1.9. Un **experimento aleatorio**, es un experimento que tiene un conjunto de resultados bien definidos al que llamaremos **espacio muestral** y denotamos con $\mathscr{S}$.

Definición 1.10. Sea $\mathfrak{S}$ σ-álgebra de eventos asociados a $\mathscr{S}$. Si existe $\Pr\colon \mathfrak{S} \to \mathbb{R}$ con Pr medida totalmente finita sobre $\mathfrak{S}$ y $\Pr(\mathscr{S}) = 1$ entonces diremos que Pr es una **medida de probabilidad**.

Definición 1.11. Si $\mathscr{S}$ conjunto finito, $\mathfrak{S} = \mathscr{P}(\mathscr{S})$, a la tripleta $(\mathscr{S}, \mathfrak{S}, \Pr)$ le llamaremos **espacio de probabilidad**.

Definición 1.12. Sea $X\colon (\mathscr{S}, \mathfrak{S}) \to (\mathbb{R}, \mathfrak{B})$ una aplicación, si X es medible le llamaremos **variable aleatoria** y $X(s) \in \mathbb{R}$ para cada $s \in \mathscr{S}$.

Si $X\colon \mathscr{S} \to \mathbb{R}$ es variable aleatoria usaremos la notación

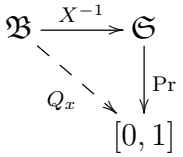
$$(X \in T) = [X \in T] = X^{-1}(T) := \{s \in \mathscr{S} \mid X(s) \in T\}$$

para $T \in \mathfrak{B}, X^{-1} \in \mathfrak{S}$.

Ejemplo 1.2. Sea X la suma de las caras obtenidas al lanzar dos dados y sea $T = \{4\}$ entonces se tiene que

$$(X \in \{4\}) = X^{-1}(4) = \{s \in \mathscr{S} \mid X(s) \in \{4\}\},\ X \in \{4\} = \{(2,2), (3,1), (1,3)\}.$$

Además podemos observar que si X es una variable aleatoria y Pr una medida de probabilidad se tiene que $X\colon (\mathscr{S}, \mathfrak{S}, \Pr) \to (\mathbb{R}, \mathfrak{B})$, pues $X\colon \mathscr{S} \to \mathbb{R}$, $\Pr\colon \mathfrak{S} \to [0,1]$, $X^{-1}\colon \mathfrak{B} \to \mathfrak{S}$, y el siguiente diagrama conmutativo se cumple.



Es decir $Q_x = \Pr \circ X^{-1}$ y así $Q_x\colon \mathfrak{B} \to [0,1]$, $Q_x(B) = (\Pr \circ X^{-1})(B)$, $B \in \mathfrak{B} = \Pr(X^{-1}(B)) = \Pr(x \in B) = \Pr(\{s \in \mathfrak{S} \mid X(s) \in B\})$.

Definición 1.13. Sea $X\colon (\mathscr{S}, \mathfrak{S}) \to (\mathbb{R}^k, \mathfrak{B}^k)$ medible, a X le llamaremos **vector aleatorio k-dimensional**.

Además para $X^{-1}(B) \in \mathfrak{S}$ con $B \subset \mathbb{R}^k$ conjunto de Borel, es válida la construcción.

$$Q(B) = (\Pr \circ X^{-1})(B) = \Pr(X^{-1}(B))$$

en donde $\Pr\colon \mathfrak{S} \to [0,1]$ es una medida de probabilidad, con un subconjunto de tamaño $k\colon X_1, \ldots, X_k$ medidas obtenidas. Sea $X := (X_1, \ldots, X_k)$ un vector aleatorio, entonces la realización de la muestra es denotada por $X = (X_1 = x_1, \ldots, X_k = x_k)$; así que nuestro conjunto de Borel es $B = \{(X_1, \ldots, X_k)\}$.

Teorema 1.3.1. *La función de conjuntos $Q\colon \mathfrak{B}^k \to \mathbb{R}$ es una medida de probabilidad y $(\mathbb{R}^k, \mathfrak{B}^k, Q)$ es un espacio de probabilidad.*

Demostración. Sea $B \in \mathfrak{B}^k$, entonces $Q(B) = \Pr(X^{-1}(B)) \geq 0$, $X^{-1}(B) \in \mathfrak{S}$, para $B = \mathbb{R}^k$ tenemos $Q(\mathbb{R}^k) = \Pr(X^{-1}(\mathbb{R}^k)) = \Pr(\mathscr{S}) = 1$, para una clase disjunta $\{B_j\}$ de conjuntos de Borel en $\mathbb{R}^k$:

$$Q\Big(\bigcup_{i=1}^{\infty} B_i\Big) = \Pr\Big(X^{-1}\Big(\bigcup_{i=1}^{\infty} B_i\Big)\Big) = \Pr\Big(\bigcup_{i=1}^{\infty} X^{-1}(B_i)\Big) = \sum_{i=1}^{\infty} \Pr\Big(X^{-1}(B_i)\Big) =$$

$$\sum_{i=1}^{\infty} Q(B_i).$$

Entonces Q es una medida probabilidad en $(\mathbb{R}^k, \mathfrak{B}^k)$. □

Definición 1.14. Sea X un vector aleatorio k-dimensional, se dice que X es **discreto** si existe un conjunto contable de puntos $z_j \in \mathbb{R}^k$, $j \in I$ contable, tal que $\sum_{j \in I} \Pr(x = z_j) = 1$, en donde $(x = z_j) :\ = \{s \in \mathscr{S} \mid x(s) = z_j\}$.

Definición 1.15. Sea $f(z_j) :\ = \Pr(x = z_j)$, $j \in I$. A f le llamaremos **función de densidad de probabilidad discreta de X** o simplemente la **función de probabilidad**,[1] entonces:

1. $f(z_j) \geq 0$, $j \in I$
2. $\sum_{j \in I} f(z_j) = 1$
3. Si $B \in \mathfrak{B}^k$ entonces $\Pr(X \in B) = \sum_{z_j \in B} f(z_j)$

Definición 1.16. Si X es tal $\Pr(x = z) = 0$, para cada $z \in \mathbb{R}^k$ diremos que X es un **vector aleatorio continuo**.

Definición 1.17. Si existe una aplicación no negativa $f : (\mathbb{R}^k, \mathfrak{B}^k) \to (\mathbb{R}, \mathfrak{B})$ medible, tal que $\Pr(B) = \int_B f \, \mathrm{d}m, B \in \mathfrak{B}^k$ (integral de Lebesgue, m : medida de Lebesgue), decimos que X es **absolutamente continuo**. A f_x le llamaremos **función de densidad continua de probabilidad o densidad de X.**[2]

Teorema 1.3.2. *Sea $X = (x_1, \ldots, x_k) : (\mathscr{S}, \mathfrak{S}) \to (\mathbb{R}^k, \mathfrak{B}^k)$ una aplicación en donde cada x_i, con $i = 1, \ldots, k$ es una función de valor real sobre $\mathscr{S}$. Entonces X es un vector aleatorio si y solo si las x_i son variables aleatorias para $i = 1, \ldots, k$.*

[1] Algunos autores también le llaman función de distribución. Kreyszig, E. (1979). p.88.

[2] Para un abordaje amplio de teoría de la medida consultar Rudin, W. (1987). Real and Complex Analysis.

Demostración. Si X es un vector aleatorio, se tiene que X es medible, es decir que se cumple que $X^{-1}(T) \in \mathfrak{S}$ para cada $T \in \mathfrak{B}^k$ y dado que $X^{-1}(T) = (x_1^{-1}(T), \ldots, x_k^{-1}(T))$ para cada $i = 1, \ldots, k$ tendremos que las x_i son medibles y por lo tanto variables aleatorias para $i = 1, \ldots, k$; la prueba de la recíproca es similar. □

1.4. Algunas funciones de probabilidad

Dado que hemos definido formalmente las funciones de probabilidad definiremos a continuación la distribución binomial y la distribución normal que son ejemplos de funciones de probabilidad discreta y continua respectivamente, además de ser las más conocidas y utilizadas en la mayoría de experimentos aleatorios.

1.4.1. Distribución binomial

Si repetimos un experimento varias veces y en cada una de los ensayos el experimento solamente tiene dos posibles resultados S y F que llamaremos éxito y fracaso respectivamente, y a su vez con probabilidades p y q, a lo largo de todos los ensayos que a su vez son eventos independientes y si p y q se mantienen constantes decimos que es un Ensayo de Bernoulli.

A partir de lo anterior podemos ver que $p, q \geq 0$ y además $p + q = 1$, así pues en un ensayo los posibles resultados son $\mathscr{S} = \{S, F\}$ en dos ensayos $\mathscr{S} = \{(S,S), (S,F), (F,S), (F,F)\}$ y en cinco ensayos un posible resultado sería (S, F, F, S, S) y dado que la probabilidad de cada uno es p ó q y son eventos independientes la probabilidad de este último resultado sería $\Pr\ (x = (S, F, F, S, S)) = p \cdot q \cdot q \cdot p \cdot p = p^3q^2$ y si además no nos importara el orden en el que aparece esta combinación de tres éxitos y dos fracasos tendríamos otras combinaciones, como (S, S, S, F, F) por ejemplo y en total dado que son cinco ensayos y dos resultados posibles tenemos $\binom{5}{2} = 10$ combinaciones diferentes lo que hace que la probabilidad de obtener tres éxitos y dos fracasos sin importar el orden, sea $\Pr = \binom{5}{2}p^3q^2$ siguiendo la misma lógica podemos definir la **distribución binomial**.

Definición 1.18. Dado un experimento con dos posibles resultados con probabilidades p en caso de éxito y $q = 1 - p$ en caso contrario y realizado n veces con $y = y_0$ éxitos y $n - y$ fracasos, su probabilidad de ocurrencia la calculamos con

$$\Pr(y = y_0; p) = \binom{n}{y} p^y (1-p)^{n-y}$$

siendo esta llamada **función de probabilidad binomial**.

1.4.2. Distribución normal

La función definida por

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

es llamada la **función de probabilidad normal** y a su integral de área acumulada

$$F(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{1}{2}y^2} dy$$

le llamamos la **función de distribución normal**, las gráficas de ambas funciones se muestran en las figuras 1.1 y 1.2 respectivamente, ambas gráficas fueron realizadas con el software R.[3]

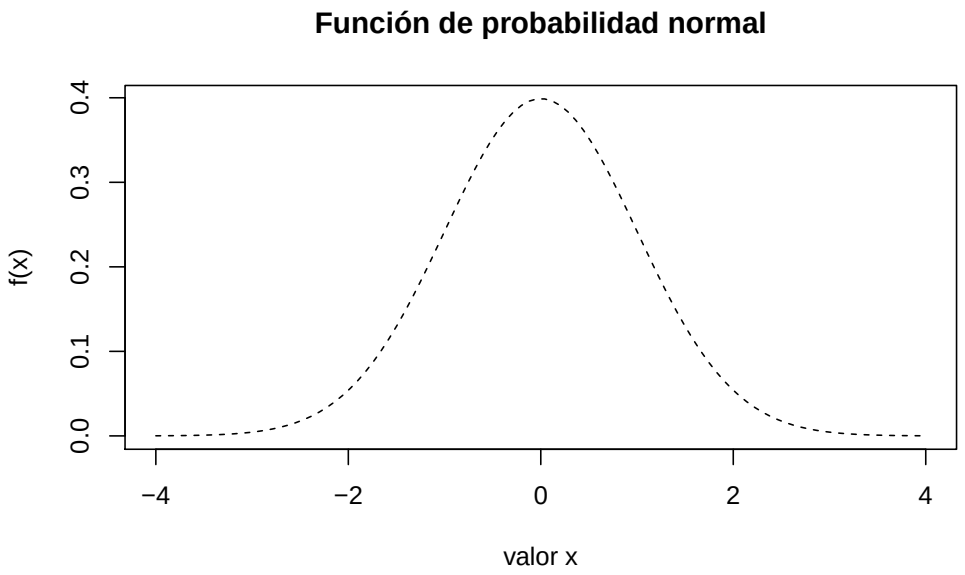


Figura 1.1. Función de probabilidad normal. Fuente: elaboración propia con R.

Una particularidad de $F(x)$ es que no puede calcularse con métodos elementales de integración por lo cual es necesario el uso de métodos numéricos para obtener los valores requeridos, dichos valores pueden calcularse fácilmente en un software especializado como R, una calculadora gráfica o incluso en tablas impresas con tal fin generalmente adjuntas en los libros de estadística.

[3]R es un entorno y lenguaje de programación enfocado en el análisis estadístico.

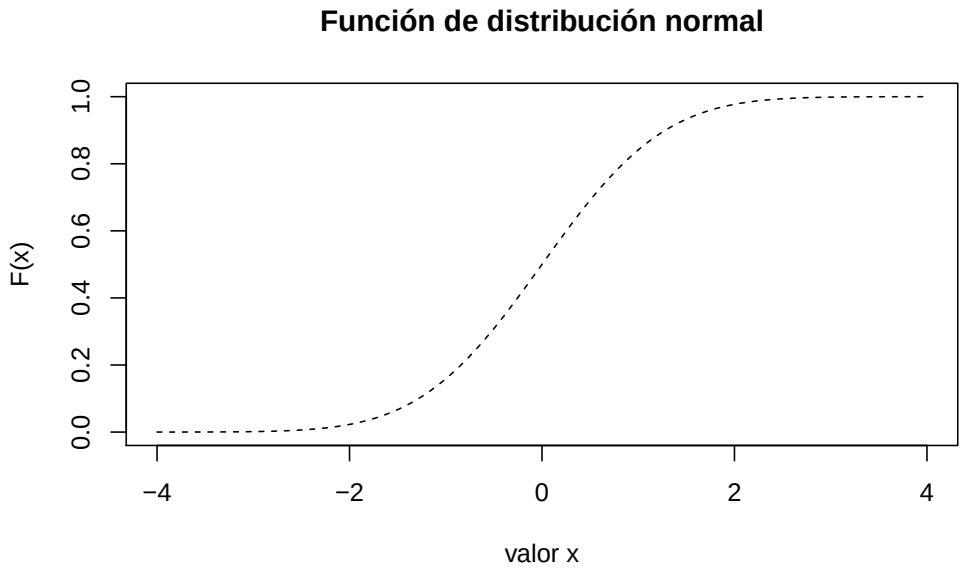


Figura 1.2. Función de distribución normal. Fuente: elaboración propia con R.

La importancia de la distribución normal es crucial en la modelación de muchos fenómenos aleatorios, pues la evidencia ha mostrado que al aumentar el número de observaciones la función de probabilidad que los modela tiende a ser la función normal y eso nos lleva al teorema 1.6.3, que es uno de los teoremas fundamentales en el desarrollo de la teoría de probabilidades y que demostraremos más adelante.

1.5. Convergencia de funciones de probabilidad

Definición 1.19. Sea $\{X_n\}$, $n = 1, 2, \ldots$ una sucesión de variables aleatorias y sea X una variable aleatoria definida en el espacio muestral $\mathscr{S}$ con eventos $\mathfrak{S}$ y función de probabilidad Pr, *i.e.* la secuencia de variables aleatorias $\{X_n\}$ y la variable aleatoria X se encuentran definidas en el espacio de probabilidad $(\mathscr{S}, \mathfrak{S}, \Pr)$, definimos entonces para esta sucesión cuatro diferentes formas de convergencia.

1. Decimos que $\{X_n\}$ **converge casi seguramente** (c.s.) o **con probabilidad uno** a X cuando $n \to \infty$ y escribimos $X_n \xrightarrow[n\to\infty]{\text{c.s.}} X$ o $X_n \underset{n\to\infty}{\longrightarrow} X$ con probabilidad 1, o $\Pr\left[X_n \underset{n\to\infty}{\longrightarrow} X\right] = 1$, si $X_n(s) \underset{n\to\infty}{\longrightarrow} X(s)$ para cada $s \in \mathscr{S}$ excepto para posiblemente un subconjunto S de $\mathscr{S}$ tal que $\Pr(S) = 0$, de tal forma que $X_n \xrightarrow[n\to\infty]{\text{c.s.}} X$ significa que para cada $\epsilon > 0$ y para todo $s \in S^c$ existe un $N(\epsilon, s) > 0$ tal que $|X_n(s) - X(s)| < \epsilon$ para cada $n \geq N(\epsilon, s)$, a esta

convergencia le llamaremos **convergencia fuerte**.

2. Decimos que $\{X_n\}$ **converge en probabilidad** a X cuando $n \to \infty$ y escribimos

   $X_n \xrightarrow[n\to\infty]{\mathrm{P}} X$ si para cada $\epsilon > 0$, $\Pr\left[|X_n - X| > \epsilon\right] \underset{n\to\infty}{\longrightarrow} 0$.

   De tal forma que $X_n \xrightarrow[n\to\infty]{\mathrm{P}} X$ significa que: Para cada $\epsilon, \delta > 0$ existe $N(\epsilon, \delta) > 0$ tal que $\Pr\left[|X_n - X| > \epsilon\right] < \delta$ para cada $n \geq N(\epsilon, \delta)$.

3. Sean $F_n = F_{X_n}$, $F = F_X$ entonces decimos que $\{X_n\}$ **converge en distribución** a X cuando $n \to \infty$ y escribimos $X_n \xrightarrow[n\to\infty]{d} X$ si $F_n(x) \underset{n\to\infty}{\longrightarrow} F(x)$ para cada $x \in \mathbb{R}$ para el cual F es continuo.

   De tal forma que $X_n \xrightarrow[n\to\infty]{d} X$ significa que: Para cada $\epsilon > 0$ y para todo x para el cual F es continua, existe $N(\epsilon, x)$ tal que $|F_n(x) - F(x)| < \epsilon$ para cada $n \geq N(\epsilon, x)$. Siendo este tipo de convergencia llamado **convergencia débil**.

4. Decimos que $\{X_n\}$ converge a X en **media cuadrática** (q.m) cuando $n \to \infty$ y escribimos $X_n \xrightarrow[n\to\infty]{q.m.} X$ si su momento[4] $E|X_n - X|^2 \underset{n\to\infty}{\longrightarrow} 0$, lo cual significa que para cada $\epsilon > 0$ existe $N(\epsilon) > 0$ tal que $E|X_n - X|^2 < \epsilon$ para cada $n \geq N(\epsilon)$.

Nota. La convergencia casi segura es la ya familiar convergencia puntual de la sucesión de números $\{X_n(s)\}$ para cada s fuera de un evento S de probabilidad cero (evento nulo). La convergencia en distribución es también una convergencia puntual de la sucesión de números $\{F_n(x)\}$ para todo x para el cual F es continua. La convergencia en probabilidad posee una naturaleza diferente, colocando $A_n = \{s \in S : |X_n(s) - X(s)| > \epsilon\}$ para un arbitrario pero fijo ϵ, tenemos que $X_n \xrightarrow[n\to\infty]{P}$, si $\Pr(A_n) \underset{n\to\infty}{\longrightarrow} 0$, de tal forma que la sucesión de números $\{\Pr(A_n)\}$ tiende a 0 cuando $n \to \infty$, pero los eventos A_n se mantienen alrededor del espacio muestral $\mathscr{S}$. Y finalmente la convergencia en media cuadrática significa simplemente que los promedios $E|X_n - X|^2$ convergen a 0 cuando $n \to \infty$.

[4]La definición y diferencia entre los momentos $E[X]$, $E(X)$ y $E|X|$ también conocidos como esperanzas matemáticas o promedios puede verse en Roussas p.107.

1.6. Teoremas de convergencia, teorema del límite central

A continuación enunciaremos dos teoremas que engloban las relaciones existentes entre los tipos de convergencia enunciados anteriormente, las pruebas pueden encontrarse en el libro de Roussas. p.183.

Teorema 1.6.1. *Sea X_n una sucesión.*

1. $X_n \xrightarrow[n\to\infty]{c.s.} X$ *implica que* $X_n \xrightarrow[n\to\infty]{P} X$.
2. $X_n \xrightarrow[n\to\infty]{q.m.} X$ *implica que* $X_n \xrightarrow[n\to\infty]{P} X$.
3. $X_n \xrightarrow[n\to\infty]{P} X$ *implica que* $X_n \xrightarrow[n\to\infty]{d} X$. *La conversa también es cierta si X es degenerada, i.e.* $\Pr[X = c] = 1$ *para alguna constante c.*

Previo a enunciar el siguiente teorema se necesita la definición de *función característica* de una variable aleatoria X la cual se enuncia a continuación.

Definición 1.20. Sea X una variable aleatoria con función de densidad de probabilidad f, entonces la **función característica** de X (ch. f. de X) denotada por ϕ_X o solamente ϕ es una función definida en $\mathbb{R}$ que toma valores complejos y que se define como

$$\phi_X(t) = E\Big[e^{itX}\Big] = \begin{cases} \sum\limits_x e^{itX} f(x) = \sum\limits_x \Big[\cos(tx)f(x) + \mathrm{i}\,\mathrm{sen}(tx)f(x)\Big] \\ \int\limits_{-\infty}^{\infty} e^{itX} f(x)dx = \int\limits_{-\infty}^{\infty} \Big[\cos(tx)f(x) + \mathrm{i}\,\mathrm{sen}(tx)f(x)\Big]dx \end{cases}$$

$$= \begin{cases} \sum\limits_x \Big[\cos(tx)f(x)\Big] + \mathrm{i}\sum\limits_x \Big[\mathrm{sen}(tx)f(x)\Big] \\ \int\limits_{-\infty}^{\infty} \cos(tx)f(x)dx + \mathrm{i}\int\limits_{-\infty}^{\infty} \mathrm{sen}(tx)f(x)dx. \end{cases}$$

La función característica ϕ_X es llamada también la **transformada de Fourier** de f.

Teorema 1.6.2. *(Teorema de continuidad de Lévy) Sea $\{F_n\}$ una sucesión de funciones de distribución y sea F una función de distribución, sea ϕ_n la ch. f. correspondiente a F_n y ϕ la ch. f. correspondiente a F, entonces tendremos que:*

1. *Si* $F_n(x) \underset{n\to\infty}{\longrightarrow} F(x)$ *para todos los puntos de continuidad* x *de* F*, implica que* $\phi_n(t) \underset{n\to\infty}{\longrightarrow} \phi(t)$*, para todo* $t \in \mathbb{R}$*.*

2. *Si* $\phi_n(t)$ *converge a una función* $g(t)$*, cuando* $n \to \infty$ *y* $t \in \mathbb{R}$ *y además* $g(t)$ *es continua en* $t = 0$*, entonces* g *es una ch.f. y si* F *es la función de distribución correspondiente, entonces* $F_n(x) \underset{n\to\infty}{\longrightarrow} F(x)$*, para todos los puntos de continuidad* x *de* F*.*

Ahora terminaremos este capítulo con uno de los teoremas más importantes y conocidos en la teoría de probabilidades y estadística, el Teorema del Límite Central que es básicamente en el que nos basamos para realizar inferencias en conjuntos de datos relativamente grandes y por el cual en muchos casos podemos suponer normalidad o realizar pruebas estadísticas para establecerla, en la función de distribución estudiada.

Teorema 1.6.3. *(Teorema del Límite Central) Sean* $X_1, \ldots X_n$ *variables aleatorias independientes e idénticamente distribuidas (iid) con media* μ *(finita) y varianza* σ^2 *(finita y positiva). Sea*

$$\overline{X}_n = \frac{1}{n}\sum_{j=1}^{n} X_j, \quad G_n(x) = \Pr\Big[\frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma} \leq x\Big], \quad y \quad \Phi(x) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{-t^2/2}dt$$

entonces $G_n(x) \underset{n\to\infty}{\longrightarrow} \Phi(x)$ *para cada* $x \in \mathbb{R}$*.*

Nota. i) Frecuentemente se define (informalmente) el Teorema del Límite Central (CLT, por sus siglas en inglés) como

$$\frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma} \approx N(0,1) \quad \text{o} \quad \frac{S_n - E(S_n)}{\sigma(S_n)} \approx N(0,1)$$

para n grande tenemos

$$S_n = \sum_{j=1}^{n} X_j \quad \text{donde} \quad \frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma} = \frac{S_n - E(S_n)}{\sigma(S_n)}.$$

ii) En la parte i) la notación S_n fue usada para denotar la suma de las variables aleatorias $X_1, \ldots, X_n$ esta notación es generalmente aceptada y es la que usaremos.
iii) La notación «o pequeña» será empleada para denotar el residuo en la expansión de la Serie de Taylor.

Demostración. Sea g_n la ch. f. de G_n y ϕ la ch. f. de Φ, esto es $\phi(t) = e^{-t^2/2}$, $t \in \mathbb{R}$, entonces por el teorema 1.6.2 es suficiente probar que $g_n(t) \underset{n\to\infty}{\longrightarrow} \phi(t)$, esto implicará que $G_n(x) \to \Phi(x)$, $x \in \mathbb{R}$.

Consideremos

$$\frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma} = \frac{n\overline{X}_n - n\mu}{\sigma\sqrt{n}} = \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \frac{X_j - \mu}{\sigma} = \frac{1}{\sqrt{n}} \sum_{j=1}^{n} Z_j,$$

donde $Z_j = (X_j - \mu)/\sigma, j = 1, \ldots, n$ son (i.i.d.) con $E(Z_j) = 0, \quad \sigma^2(Z_j) = E(Z_j^2) = 1$ (por simplicidad escribiremos $\sum_j Z_j$ en lugar de $\sum\limits_{j=1}^{n} Z_j$ cuando aparezca como un subíndice), tenemos entonces que

$$g_n(t) = g_{(t/\sqrt{n})\sum_j Z_j}(t) = g_{\sum_j Z_j}\Big(\frac{t}{\sqrt{n}}\Big) = \Big[g_{Z_1}\Big(\frac{1}{\sqrt{n}}\Big)\Big]^n.$$

Si consideramos la expansión en serie de Taylor de g_{Z_1} alrededor del cero hasta el segundo término, tendremos que

$$g_{Z_1}\Big(\frac{t}{\sqrt{n}}\Big) = g_{Z_1}(0) + \frac{t}{\sqrt{n}} g'_{Z_1}(0) + \frac{1}{2!}\Big(\frac{t}{\sqrt{n}}\Big)^2 g''_{Z_1}(0) + o\Big(\frac{t^2}{n}\Big)$$

y dado que

$$g_{Z_1}(0) = 1, \quad g'_{Z_1}(0) = iE(Z_1) = 0, \quad g''_{Z_1}(0) = i^2 E(Z_1^2) = -1$$

obtenemos

$$g_{Z_1}\Big(\frac{t}{\sqrt{n}}\Big) = 1 - \frac{t^2}{2n} + o\Big(\frac{t^2}{n}\Big) = 1 - \frac{t^2}{2n} + \frac{t^2}{n}o(1) = 1 - \frac{t^2}{2n}[1 - o(1)]$$

de modo que

$$g_n(t) = \Big[1 - \frac{t^2}{2n}[1 - o(1)]\Big]^n.$$

Tomando el límite cuando $n \to \infty$ tenemos que, $g_n(t) \underset{n\to\infty}{\longrightarrow} e^{-t^2/2}$, la cual es la ch.f. de Φ quedando demostrado. □

El Teorema 1.6.3 genera el siguiente corolario, de modo que entre ambos justifican varias aproximaciones.

Corolario 1.1. *La convergencia de* $G_n(x) \underset{n\to\infty}{\longrightarrow} \Phi(x)$ *es uniforme en* $x \in \mathbb{R}$*, es decir*

que para cada $x \in \mathbb{R}$ *y cada* $\epsilon > 0$ *existe un* $N(\epsilon) > 0$ *independiente de* x *tal que* $|G_n(x) - \Phi(x)| < \epsilon$ *para cada* $n \geq N(\epsilon)$ *y cada* $x \in \mathbb{R}$ *simultáneamente.*

2. Teoría de estimación

2.1. Poblaciones y muestras

Uno de los principales objetivos de la estadística consiste en describir y predecir el comportamiento de un conjunto de datos, grupo de personas o fenómeno en general, a este grupo de estudio que nos interesa describir o predecir comportamientos le llamamos *población*, por ejemplo nos puede interesar entender que tipo de personas utilizan el sistema de transporte público de una determinada ciudad, así que nuestra población de estudio son todas las personas que utilizan el servicio de transporte público en esa ciudad, tratar de resumir las características principales de esta población cualitativa o cuantitativamente puede resultar extremadamente difícil o de hecho imposible así que una buena técnica es estudiar solo una parte de esta población y suponer que la población total debe de comportarse de manera muy similar a este subconjunto que llamaremos *muestra*.

En nuestro ejemplo una muestra podría ser tomar a todos las personas que utilizan transporte público de una determinada calle o zona, o bien escoger al azar a 1000 personas durante el transcurso de una semana completa en diferentes horarios y en diferentes estaciones de bus o tren, en el primer caso nuestra muestra fue tomada probablemente por conveniencia práctica más que por conveniencia teórica, pues es posible que esta muestra de una calle o zona en particular no represente la realidad de toda la población de la ciudad pues sus características socio-económicas pueden ser muy diferentes del resto de población y podríamos cometer graves errores al tratar de describir el comportamiento de toda la población basándonos únicamente en el comportamiento de este grupo de personas; por otro lado en el segundo caso en el que escogemos 1000 personas al azar de diferentes áreas y en diferentes horarios tiene más probabilidades de representar correctamente las características de toda la población ya que trata de tener una muestra tan heterogénea como sea posible; a una muestra que es escogida por un método aleatorio o al azar como en el segundo caso le llamaremos *muestra aleatoria*, en general siempre preferiremos que nuestra

muestra sea aleatoria pues de tal forma se cumplirán los supuestos probabilísticos que necesitamos para describir de mejor forma nuestra población.

2.2. Estimadores

Una *estimación* es un ejercicio o regla estadística que consiste en determinar una medida numérica de una población llamada *parámetro*, existen básicamente dos formas de dar una estimación, dando un posible valor numérico o bien dar un intervalo que con una probabilidad tan alta como sea posible de contener el parámetro θ de interés, cuando damos un único valor o un intervalo posible para un parámetro poblacional a partir de una regla estadística en base a las medidas tomadas de una muestra se está calculando un *estimador* del parámetro poblacional θ, si es un único valor, le llamaremos *estimador puntual* y si damos un intervalo le llamaremos *estimador por intervalo*, un estimador puntual de un parámetro poblacional θ será representado como $\hat{\theta}$.

Por ejemplo un estimador puntual del ingreso mensual de los usuarios del transporte público de una ciudad podría ser $\theta =$ Q 5000.00 mientras que un estimador por intervalo diría que la estimación es que el ingreso mensual se encuentre en algún punto entre Q 4000.00 y Q 6000.00 es decir nuestra estimación por intervalo sería el intervalo $(4000, 6000)$.

2.3. Propiedades de estimadores puntuales

En el ejemplo anterior, puede interesarnos determinar el ingreso mensual promedio por persona, si entrevistamos por ejemplo a una persona y esta tiene un ingreso mensual promedio de Q 5000.00 podríamos inferir que el ingreso mensual promedio por persona de toda la población es de Q 5000.00, evidentemente la intuición nos dice que tomar este único dato como válido para realizar una inferencia poblacional es poco útil y confiable de tal forma que lo mejor sería entrevistar más personas, si tomamos 10 personas dentro de nuestra muestra lo más probable es que cada una de ellas reporte un dato diferente o al menos uno de ellos lo haga, de tal forma que obtendremos 10 datos diferentes de ingresos mensuales.

En este punto la pregunta obvia sería ¿cuál de los datos tomo como estimador o que operación debo realizar con todos ellos para obtener esta estimación?, en base a nuestros conocimientos básicos de estadística una de las opciones podría

ser tomar el promedio de ellos o bien la mediana de los datos, pero ¿cuál de estos será preferible?, a continuación veremos dos propiedades deseables que queremos de nuestros estimadores.

Antes de la siguiente definición se debe recordar que, el estimador puntual $\hat{\theta}$ posee una función de probabilidad f.

Definición 2.1. Sea $\hat{\theta}$ un estimador puntual del parámetro θ, decimos que $\hat{\theta}$ es un **estimador insesgado** si se cumple que $E(\hat{\theta}) = \theta$ de lo contrario decimos que es **sesgado**.

De lo anterior una de las propiedades deseables de nuestro estimador es que sea insesgado, pues esto asegura que entre más elementos tenga nuestra muestra su esperanza matemática estará más cerca del parámetro que deseamos estimar, si tenemos un estimar sesgado podemos definir su sesgo de la siguiente forma.

Definición 2.2. El **sesgo** B de un estimador puntual $\hat{\theta}$ está dado por $B = E(\hat{\theta}) - \theta$.

¿Qué sucede cuando tengo dos estimadores puntuales insesgados, cuál de ellos escoger? es acá donde surge la segunda propiedad y es que escogeremos aquel que tenga la menor varianza, así que la segunda propiedad deseable de nuestros estimadores es que sea de *mínima varianza*, esto es porque con una mínima varianza aseguramos que al realizarse un muestreo repetido la mayoría de los valores de $\hat{\theta}$ serán cercanos a θ (Wackerly, D. p. 393 [17]) la figura 2.1 nos muestra por que se desea un estimador con mínima varianza al comparar la forma de una distribución normal con desviación estándar 0.5, 1 y 10 respectivamente, como puede verse entre menor desviación estándar (y varianza en consecuencia) contenga nuestro estimador, mayor área se acumulará en un intervalo pequeño alrededor de la media o esperanza.

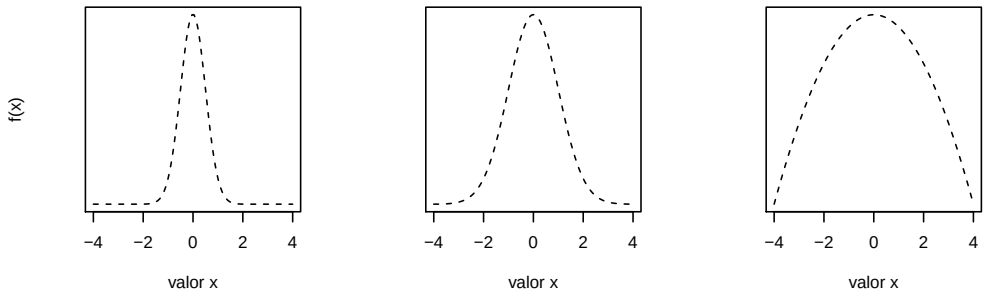


Figura 2.1. Funciones normales con desviación 0.5, 1 y 10. Fuente: Elaboración propia con R.

Algunos de los estimadores insesgados más comunes se presentan en la tabla 2.1.

Tabla 2.1. Estimadores insesgados comunes. Fuente: Wackerly, D.(2010).[tabla]

Parámetro θ	Tamaño de muestra	Estimador puntual $\hat{\theta}$	$E(\hat{\theta})$	$\sigma^2_{\hat{\theta}}$
μ	n	$\bar{Y}$	μ	$\frac{\sigma^2}{n}$
p	n	$\hat{p} = \frac{Y}{n}$	p	$\frac{pq}{n}$
$\mu_1 - \mu_2$	n_1 y n_2	$\bar{Y_1} - \bar{Y_2}$	$\mu_1 - \mu_2$	$\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$
$p_1 - p_2$	n_1 y n_2	$\hat{p_1} - \hat{p_2}$	$p_1 - p_2$	$\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}$

Adicionalmente es importante remarcar las siguientes propiedades de dos importantes estimadores puntuales de la varianza, en el siguiente teorema.

Teorema 2.3.1. *El estimador puntual de la varianza* $S'^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n}$ *es sesgado, mientras que* $S^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}$ *es insesgado.*

Demostración. Sabemos que

$$\sum_{i=1}^{n} (Y_i - \bar{Y})^2 = \sum_{i=1}^{n} Y_i^2 - n\bar{Y}^2$$

De modo que

$$E\Big[\sum_{i=1}^{n} (Y_i - \bar{Y})^2\Big] = E\Big(\sum_{i=1}^{n} Y_i^2\Big) - nE\big(\bar{Y}^2\big) = \sum_{i=1}^{n} E\big(Y_i^2\big) - nE\big(\bar{Y}^2\big). \qquad (2.1)$$

Tomado en cuenta que $E(Y_i^2)$ es igual para toda $i = 1, 2, \ldots, n$ y dado que la

varianza de una variable aleatoria está dada por $V(Y) = E(Y^2) - \mu^2$ obtenemos que $E(Y^2) = V(Y) + \mu^2$ y la expresión 2.1 se convierte en

$$E\Big[\sum_{i=1}^{n}(Y_i - \bar{Y})^2\Big] = \sum_{i=1}^{n}(\sigma^2 + \mu^2) - n\Big(\frac{\sigma^2}{n} + \mu^2\Big)$$

$$= n(\sigma^2 + \mu^2) - n\Big(\frac{\sigma^2}{n} + \mu^2\Big)$$

$$= n\sigma^2 - \sigma^2 = (n-1)\sigma^2$$

Por lo que $E\big(S'^2\big) = \Big(\frac{n-1}{n}\Big)\sigma^2$, de donde S'^2 es sesgado ya que $E\big(S'^2\big) \neq \sigma^2$. Por otro lado tenemos que

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\big(Y_i - \bar{Y}\big)^2$$

$$= \Big(\frac{n}{n-1}\Big)\Big[\frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})^2\Big] = \Big(\frac{n}{n-1}\Big)S'^2$$

$$= \Big(\frac{n-1}{n}\Big)\Big[\Big(\frac{n}{n-1}\Big)\sigma^2\Big] = \sigma^2$$

y concluimos que S^2 es un estimador insesgado para σ^2. □

2.4. Eficiencia relativa

Como ya se mencionó al tener dos estimadores puntuales $\hat{\theta}_1, \hat{\theta}_2$ para un mismo parámetro θ preferimos utilizar el de mínima varianza, así que dados dos estimadores puntuales podemos dar una medida numérica para la eficiencia de uno respecto de otro con el fin de compararlos, de modo que podemos dar la siguiente definición.

Definición 2.3. Dados dos estimadores insesgados $\hat{\theta}_1, \hat{\theta}_2$ del parámetro θ con varianzas $V(\hat{\theta}_1)$ y $V(\hat{\theta}_2)$ respectivamente, entonces **la eficiencia de θ_1 con respecto de** θ_2 denotada por $\text{eff}(\hat{\theta}_1, \hat{\theta}_2)$ se define como la razón

$$\text{eff}(\hat{\theta}_1, \hat{\theta}_2) = \frac{V(\hat{\theta}_2)}{V(\hat{\theta}_1)}.$$

De lo anterior podemos ver que $\text{eff}(\hat{\theta}_1, \hat{\theta}_2) > 1$ si $V(\hat{\theta}_1) < V(\hat{\theta}_2)$ de modo que preferiremos $\hat{\theta}_1$ por otro lado si $\text{eff}(\hat{\theta}_1, \hat{\theta}_2) < 1$ preferiremos $\hat{\theta}_2$ ya que $V(\hat{\theta}_2) < V(\hat{\theta}_1)$.

Puede demostrarse (Wackerly, D.[17]) que la varianza de la mediana muestral

es $V(\hat{\theta}_1) = (1.2533)^2(\sigma^2/n)$ con lo que si $\hat{\theta}_2$ es la media muestral, tendremos que

$$\text{eff}(\hat{\theta}_1, \hat{\theta}_2) = \frac{V(\hat{\theta}_2)}{V(\hat{\theta}_1)} = \frac{\sigma^2/n}{(1.2533)^2\sigma^2/n} = \frac{1}{1.2533^2} = 0.6366,$$

con esto podemos ver porque preferimos utilizar la media muestral como estimador de la media poblacional por encima de la mediana muestral.

2.5. Consistencia

En nuestro ejemplo anterior, nos interesaba conocer el promedio o media de los ingresos mensuales de los usuarios, nos habíamos percatado que con un solo individuo en la muestra con ingresos de Q 5000.00 mensuales nuestra mejor estimación del parámetro poblacional era precisamente $\hat{\theta} =$ Q 5000.00 pero no podemos confiar mucho en esta estimación, sin embargo esperaríamos que al tomar más individuos en la muestra nuestro estimador $\hat{\theta}$ se acerque más al parámetro θ conforme n crezca, o formalmente esperamos que converja a θ, a esta propiedad le llamamos *consistencia*.

Definición 2.4. Se dice que el estimador $\hat{\theta}_n$ es un estimador **consistente** de θ si para cualquier $\epsilon > 0$ tenemos que

$$\lim_{n\to\infty} \Pr(|\hat{\theta}_n - \theta| \leq \epsilon) = 1.$$

2.6. Intervalos de confianza

Antes definimos lo que es un *estimador por intervalo*, a estos estimadores se les conoce más comúnmente como *intervalos de confianza*, estos intervalos de confianza están representados por dos números reales a, b con $a < b$ y representados como (a, b) un intervalo usual, al igual que en los estimadores puntuales deseamos que el estimador por intervalo tenga ciertas propiedades, a saber, que θ esté contenido en (a, b) y que la distancia entre a, b sea tan pequeña como sea posible, sin embargo, dado que los extremos de este intervalo son variables aleatorias en función de la muestra tomada, no podemos asegurar la pertenencia, pero si podemos exigir que la probabilidad que contenga a θ sea tan alta como querramos, a esta probabilidad le llamaremos *coeficiente de confianza* o *nivel de confianza* denotado por $(1 - \alpha)$, de modo que si $\hat{\theta}_L$ y $\hat{\theta}_U$ son los límites inferior y superior respectivamente del intervalo de confianza para un parámetro θ la expresión

$$\Pr(\hat{\theta}_L \leq \theta \leq \hat{\theta}_U) = 1 - \alpha$$

indica que la probabilidad de que θ pertenezca al intervalo de confianza es $1 - \alpha$. El intervalo de confianza $(\hat{\theta}_L, \hat{\theta}_U)$ es conocido como *intervalo de confianza bilateral*, por otro lado llamamos *intervalos de confianza unilaterales* a los intervalos $(\theta_L, +\infty)$ y $(-\infty, \theta_U)$.

Ejemplo 2.1. Supóngase que se desea conocer la estatura promedio de los niños que asisten al tercer grado de primaria de un establecimiento educativo y para ello se toma la estatura de 36 niños escogidos al azar, dando como resultado una media muestral $\bar{y} = 1.28\,\text{m}$ y desviación estándar de $s = 0.09\,\text{m}$, y deseamos calcular un intervalo de confianza bilateral con un nivel de confianza igual a 90 %, debido a que el número de la muestra es mayor a 30,[1] podemos utilizar el hecho de que la distribución muestral tiende a ser una distribución normal y por ende la transformación

$$Z = \frac{\hat{\theta} - \theta}{\sigma_{\hat{\theta}}} \tag{2.2}$$

se aproxima a una distribución normal estándar, nuestro objetivo se traduce entonces a resolver la desigualdad

$$\Pr(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}) = 1 - \alpha = 0.90 \tag{2.3}$$

de modo que a partir de 2.2 y 2.3 despejando θ obtenemos,

$$\Pr(\hat{\theta} - z_{\alpha/2}\sigma_{\hat{\theta}} \leq \theta \leq \hat{\theta} + z_{\alpha/2}\sigma_{\hat{\theta}}) = 1 - \alpha = 0.90 \tag{2.4}$$

de donde los límites superior e inferior son; $\hat{\theta}_L = \hat{\theta} - z_{\alpha/2}\sigma_{\hat{\theta}}$ y $\hat{\theta}_U = \hat{\theta} + z_{\alpha/2}\sigma_{\hat{\theta}}$ respectivamente, en nuestro ejemplo tenemos que:

$\hat{\theta} = \bar{y} = 1.28$, $\alpha = 1 - 0.90 = 0.10$, $z_{\alpha/2} = z_{0.05} \approx 1.645$ y $\sigma_{\hat{\theta}} = \frac{\sigma}{\sqrt{n}} \approx \frac{s}{\sqrt{n}} = \frac{0.09}{\sqrt{36}}$ quedando entonces nuestros límites como:

$$\hat{\theta}_L = \bar{y} - z_{0.05}\Big(\frac{s}{\sqrt{n}}\Big) = 1.28 - 1.645\Big(\frac{0.09}{6}\Big) \approx 1.26$$

y

$$\hat{\theta}_U = \bar{y} + z_{0.05}\Big(\frac{s}{\sqrt{n}}\Big) = 1.28 + 1.645\Big(\frac{0.09}{6}\Big) \approx 1.30$$

[1]En estadística es común considerar muestras mayores a 30 como normales, ver teorema del límite central.

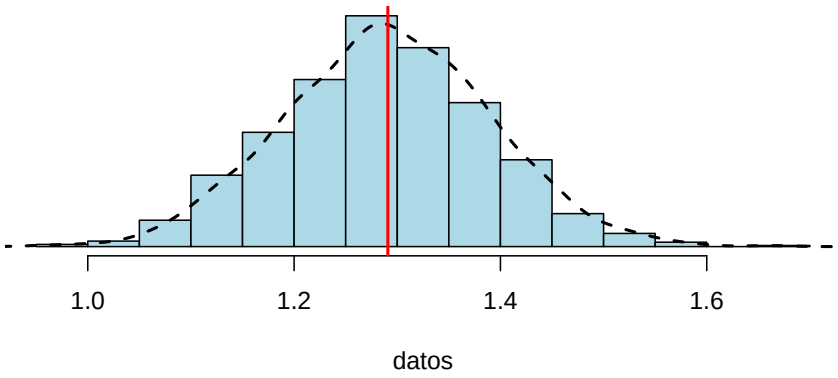


Figura 2.2. Histograma del conjunto de 1000 datos. Fuente: Elaboración propia con R.

así que nuestro intervalo de confianza para la media de la estatura será $(1.26, 1.30)$ con un nivel de confianza del 90 %.

Podemos realizar una simulación, que seleccione muestras aleatorias tomadas de un conjunto de datos, con el fin de calcular intervalos de confianza para la media poblacional μ, con cada una de ellas y corroborar la cantidad de veces que μ queda contenida dentro de cada intervalo de confianza, para dicha simulación podemos utilizar el siguiente código en el lenguaje de programación R,[2] que genera 1000 datos de forma aleatoria tomados de una distribución normal, en la figura 2.2 se representa un histograma de estos datos y su media $\mu = 1.29087$.

```
##Fijamos una semilla para que el experimento sea reproducible.
set.seed(11)
##Generamos un conjunto de 1000 datos.
data = round(rnorm(1000,1.29,0.1),2)
data
##Calculamos la media del conjunto de datos.
mean_data = mean(data)
mean_data
#Se establece la cantidad y el tamano de muestras.
total_samp = 100
n = 250
##se establece el valor de alfa
```

[2]Los comentarios al código de R se preceden con los símbolos ##.

```
alpha = 0.10
## se inicializa un vector donde se guardaran los resultados
aciertos = c()
##Este ciclo genera la cantidad de muestras establecidas,
##calcula los intervalos de confianza, verifica si la
##media de los datos queda dentro de los intervalos
##de confianza y guarda los resultados
for (i in 1:total_samp) {
  samp = sample(data,n)
  x = mean(samp)
  s = sd(samp)
  z = qnorm(1 - alpha)

  up = x + z*(s/sqrt(n))
  low = x - z*(s/sqrt(n))

  sucess = (mean_data >= low & mean_data <= up)
  aciertos = c(aciertos,sucess)
}
##Contamos los aciertos totales y calculamos la fraccion de
##veces que el intervalo contiene la media de la poblacion.
total = length(which(aciertos == TRUE))
percent = total/total_samp
```

La fracción resultante es $93/100 = 0.93$ que es lo que esperábamos; es de suponer que si cambiamos la semilla inicial esta fracción podrá variar.

3. Teoría de verosimilitud

En inferencia estadística nos interesa conocer ciertos parámetros desconocidos de una población y para ello nos valemos de la técnica del muestreo, un método útil en esta tarea es el de verosimilitud el cual se define como la probabilidad conjunta entre dos vectores, uno de estos es un vector aleatorio y se construye a partir de los valores de la muestra obtenida de la población en estudio, el otro vector estará conformado por los parámetros de la función de probabilidad f que suponemos modela dicha población, a partir de esto trataremos de inferir el modelo que mejor se ajusta a los datos observados, es decir que, la *verosimilitud* es la probabilidad conjunta de observar la muestra.

De ahora en adelante supondremos dos cosas, la primera es que conocemos la función de probabilidad de la población en estudio y la otra es que las variables aleatorias son independientes e idénticamente distribuidas (iid), esta última propiedad es útil para poder hacer afirmaciones más amplias acerca de la verosimilitud; a continuación se definirá de manera formal la función de verosimilitud.

3.1. Función de verosimilitud

Definición 3.1. Sea y una variable aleatoria discreta con función de probabilidad $f(y;\theta) = \Pr(y;\theta)$, $y = y_0$ un vector aleatorio n-dimensional el cual está conformado por los valores observados en la muestra, $\theta = (\theta_1, \ldots, \theta_k)$ el vector de los k parámetros de f, definimos entonces **la función de verosimilitud de** θ como

$$L(\theta; y_0) = C(y_0) f(y_0; \theta) \propto \Pr(y = y_0; \theta), \tag{3.1}$$

donde $C(y_0)$ es una constante que depende de la muestra observada.

El empleo de la constante C puede resultar de uso práctico al querer encontrar el estimador de máxima verosimilitud el cual se definirá más adelante, esta constante al no depender de θ puede ser escogida a conveniencia.

Ejemplo 3.1. Supongamos que una fábrica de dispositivos electrónicos desea saber que proporción de los productos que produce son defectuosos, en la práctica hacer una prueba de los dispositivos uno a uno sería una tarea que consume demasiado tiempo y recursos, incluso podría ser imposible para cantidades muy grandes, así que, el método utilizado es tomar una muestra de tamaño n e inferir acerca de la proporción poblacional utilizando la proporción de la muestra, este experimento puede ser modelado con una distribución binomial

$$\Pr(y = y_o; \theta) = \binom{n}{y} \theta^y (1-\theta)^{n-y}, \tag{3.2}$$

donde θ es un vector de dimensión 1 y representa la proporción de dispositivos defectuosos, por lo que $0 \leq \theta \leq 1$, por otro lado n es el número de dispositivos tomados aleatoriamente es decir el tamaño de la muestra y y es el número de los dispositivos defectuosos encontrados en ella, también unidimensional, la función de verosimilitud queda entonces como

$$L(\theta; y_0) = C \cdot \binom{n}{y} \theta^y (1-\theta)^{n-y}, \tag{3.3}$$

en este caso es conveniente hacer $C = 1/\binom{n}{y}$ para obtener $L(\theta; y_0) = \theta^y (1-\theta)^{n-y}$, esta elección de C es válida ya que C no depende del parámetro θ.

Al parámetro $\widehat{\theta} = \theta(y)$ que maximiza $L(\theta; y)$ le llamaremos el *estimador de máxima verosimilitud* (EMV) de θ.

En nuestro ejemplo anterior nosotros queremos calcular el mencionado EMV para obtener el valor del parámetro θ más verosímil a la luz de los datos observados, esto se reduce entonces a encontrar el máximo de la función $L(\theta; y)$, hay que hacer énfasis en que este valor máximo podría no ser único y aún más, podría no existir, afortunadamente en el ejemplo anterior éste si existe y además es simple de encontrar, observemos que maximizar $\binom{n}{y}\theta^y(1-\theta)^{n-y}$ es equivalente a maximizar $\theta^y(1-\theta)^{n-y}$, es acá donde se puede apreciar la utilidad de la constante C, por otro lado como hemos supuesto que las muestras son vectores aleatorios independientes, tendremos entonces que la verosimilitud se expresa como función de productos y será más fácil aún maximizar el logaritmo de $L(\theta; y_o)$ esto es;

$$\text{máx}[L(\theta; y_o)] = \text{máx}[\log(L(\theta; y_o))],$$

lo cual nos lleva a la siguiente.

Definición 3.2. A $l(\theta) = \log(L(\theta; y_o))$, le llamaremos **función log-verosimilitud** de θ, además se tendrá que

$$l(\theta) = C_1 + \log(f(\theta; y)), \text{ con } C_1 = \log(C).^1$$

Siguiendo nuestro ejemplo tenemos que

$$l(\theta) = y \log(\theta) + (n - y) \log(1 - \theta) \tag{3.4}$$

el EMV será entonces el valor θ que maximiza $l(\theta)$, calcularemos este valor usando los criterios de derivada[2]

$$l'(\theta) = \frac{dl(\theta)}{d\theta} = \frac{y}{\theta} - \frac{n - y}{1 - \theta} \tag{3.5}$$

igualando a 0 obtenemos que $\theta = \frac{y}{n}$, ahora bien tenemos que mostrar aún que este valor maximiza $l(\theta)$ para ello tenemos que

$$l''(\theta) = \frac{d^2 l(\theta)}{d\theta^2} = -\frac{y}{\theta^2} - \frac{n - y}{(1 - \theta)^2} \tag{3.6}$$

sustituyendo $\theta = \frac{y}{n}$ en (3.6) obtenemos

$$l''(\theta) = -\frac{n^2}{y} - \frac{n - y}{(1 - \frac{y}{n})^2} \tag{3.7}$$

dado que $0 < y \leq n$ tenemos que $l''(\frac{y}{n}) < 0$ por lo que $\widehat{\theta} = \frac{y}{n}$, que es el valor que esperaríamos en base a nuestros conocimientos de inferencia estadística.

A $l'(\theta)$ y $-l''(\theta)$ se les conoce como *función puntuación* y *función información* respectivamente.

3.2. Verosimilitud para distribuciones continuas

Hemos definido la función de verosimilitud para variables aleatorias discretas y esto podría sugerir que la función de verosimilitud no se define para el caso continuo, sin embargo en la práctica, los datos tomados tienen una precisión finita debido a los instrumentos utilizados para tomar dichos datos, de modo que cuando decimos que $Y = y_0$ significa que $y_0 - \frac{1}{2}\epsilon \leq Y \leq y_0 + \frac{1}{2}\epsilon$, donde ϵ está determinado por la

[1] A menos que se indique lo contrario usaremos log para denotar al logaritmo natural.

[2] Usualmente el valor de la derivada en el máximo será 0, pero hay que recalcar que si restringimos el dominio, esto podría no cumplirse.

precisión del instrumento, y es esto lo que motiva dar la definición 3.1, habiendo tomado en cuenta esto, extenderemos la definición de función de verosimilitud para variables aleatorias continuas.

Definición 3.3. Sea Y una variable aleatoria continua con función de densidad $f(y;\theta)$, y sea $Y = (Y_1, \ldots, Y_n)$ una muestra de variables aleatorias iid, de modo que la observación $Y_i = y_i$, entonces la función de verosimilitud de θ es proporcional a la probabilidad conjunta de la muestra observada y se define como

$$L(\theta; y) \propto \prod_{i=1}^{n} \Pr(y_i - \frac{1}{2}\epsilon \leq Y_i \leq y_i + \frac{1}{2}\epsilon; \theta)$$

siendo equivalente a

$$L(\theta; y) \propto \prod_{i=1}^{n} \int_{y_i - \frac{1}{2}\epsilon}^{y_i + \frac{1}{2}\epsilon} f(y;\theta) dy$$

llamada la **verosimilitud exacta** de θ.

Desde este punto, se podrá notar que estamos en aprietos si lo que queremos es encontrar el EMV para una función de este tipo, pues debemos maximizar un producto de integrales, así que necesitaremos hacer algunas aproximaciones para lograr nuestro objetivo.

Si consideramos que

$$\Pr(a < X \leq b) = \int_a^b f(x)dx = F(b) - F(a),$$

donde $F(x)$ es la función de distribución acumulada, podemos ver que

$$L(\theta; y) \propto \prod_{i=1}^{n} [F(y_i + \frac{1}{2}\epsilon) - F(y_i - \frac{1}{2}\epsilon)],$$

dado que ϵ es un valor pequeño, el cómputo de este producto puede causar graves errores de redondeo (Kalbfleisch, J.G. volume 2. p. 26. [7]), por lo que haremos la aproximación

$$F(y_i + \frac{1}{2}\epsilon) - F(y_i - \frac{1}{2}\epsilon) \approx f(y_i)\Delta_i,$$

la motivación de esta aproximación es encontrar el área bajo la función de densidad a través de un rectángulo de altura $f(y_i)$ y ancho Δ_i. A partir de lo anterior obtenemos

$$L(\theta; y) \propto \prod_{i=1}^{n} \Delta_i \prod_{i=1}^{n} f(y_i)$$

y dado que los Δ_i no dependen de θ, la función de verosimilitud es directamente proporcional al producto de las funciones de densidad,

$$L(\theta) = C \prod_{i=1}^{n} f(y_i). \tag{3.8}$$

Nótese que al hacer la aproximación $F(y_i + \frac{1}{2}\epsilon) - F(y_i - \frac{1}{2}\epsilon) \approx f(y_i)\Delta_i$, estamos suponiendo que f es suave, por lo que esta es una buena aproximación y además f no tiene singularidades; esta es una de las críticas que se le hace a la función de verosimilitud ya que si existe alguna singularidad para un valor de θ, $L(\theta)$ no estará definida en ese valor; sin embargo hay que tomar en cuenta que si bien hemos definido $L(\theta)$ para el caso continuo como proporcional a la función de densidad, esto se hace por mera conveniencia ya que es más fácil trabajar con derivadas e integrales que con productos, por lo tanto no hay que olvidar que la función de densidad es una aproximación a la función de probabilidad y no al contrario.

Con esto aseguramos que si surgen problemas con singularidades en la función de densidad se utilizará en su lugar la función de probabilidad correspondiente (Montoya, Díaz-Francés, Sprott. [10]), por último no olvidemos que si bien el método de máxima verosimilitud es una herramienta inferencial muy útil no podemos pretender que sea el mejor método aplicable en todos los casos.

Ejemplo 3.2. Supongamos que tenemos un vector aleatorio $(y_1, y_2, \ldots, y_n)$ asociado a un muestreo que se le realizó a una población que se distribuye de acuerdo a la función normal, supongamos además que la varianza σ^2 es conocida y queremos encontrar μ, entonces la función de verosimilitud de μ está dada por

$$L(\mu) = C \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(y_i - \mu)^2} \tag{3.9}$$

escogiendo adecuadamente la constante C, esto puede quedar simplificado como

$$L(\mu) = e^{-\frac{1}{2\sigma^2}\sum (y_i - \mu)^2} \tag{3.10}$$

y obtenemos

$$l(\mu) = -\frac{1}{2\sigma^2} \sum (y_i - \mu)^2 \tag{3.11}$$

$$l'(\mu) = \frac{1}{\sigma^2} \sum (y_i - \mu) \tag{3.12}$$

$$l''(\mu) = -\frac{n}{\sigma^2} \tag{3.13}$$

de (3.12) y (3.13) obtenemos que $\widehat{\mu} = \frac{1}{n}\sum y_i = \overline{y}$ como esperaríamos. Como se pude apreciar $\widehat{\mu}$ es un estimador insesgado, sin embargo utilizando el ejemplo anterior pero tomando a μ como el parámetro conocido y a σ como el desconocido obtenemos que $\widehat{\sigma}^2 = \frac{1}{n}\sum(y_i - \overline{y})^2$ siendo este un estimador sesgado, cuando $n \to \infty$ el sesgo tiende a 0, sin embargo para $n = 2$ el sesgo puede llegar a ser muy grande (Montoya, p. 45. [9]), aunque ésta pueda ser otra crítica al método de máxima verosimilitud, realmente el problema del sesgo se tiene únicamente para muestras muy pequeñas $n < 10$ y ya dependerá del problema que estamos abordando el máximo error que podamos tolerar.

3.3. Función relativa de verosimilitud

A fin de comparar dos estimadores diferentes θ_1, θ_2 debemos emplear una forma de cuantificar sus respectivas verosimilitudes, para ello calcularemos la razón entre ellas, de modo que si tenemos

$$\frac{L(\theta_1; y_0)}{L(\theta_2; y_0)} < 1$$

el valor de $L(\theta_2; y_0)$ es mayor al de $L(\theta_1; y_0)$ y por lo tanto es más verosímil o creíble. Por lo que para determinar cual de todos los $L(\theta; y_0)$ es más creíble definiremos la función relativa de verosimilitud.

Definición 3.4. La **función relativa de verosimilitud** se define como el cociente

$$R(\theta; y) = \frac{L(\theta; y)}{\sup_\theta L(\theta; y)} = \frac{L(\theta; y)}{L(\widehat{\theta}; y)}, \tag{3.14}$$

dado que $f(y; \theta)$ es una función de probabilidad, es acotada y por lo tanto el denominador de (3.14) existe y es finito, adicionalmente se tiene que $0 \leq R(\theta; y) \leq 1$.

Para la función $R(\theta; y)$ podemos definir también

$$r(\theta) = \log(R(\theta; y)) = \log(L(\theta) - \log(L(\widehat{\theta}))), \tag{3.15}$$

$r(\theta)$ será llamada *función log-relativa de verosimilitud*, la función $r(\theta)$ cuenta con una ventaja y es que podemos hacer una aproximación sencilla de ella.

Sea $l(\theta)$ la función log-verosimilitud de un parámetro continuo θ, supongamos

además que $\widehat{\theta}$ existe y que además $l(\theta)$ tiene expansión en serie de Taylor en $\theta = \widehat{\theta}$ entonces;

$$l(\theta) = l(\widehat{\theta}) + \frac{\theta - \widehat{\theta}}{1!} l'(\widehat{\theta}) + \frac{(\theta - \widehat{\theta})^2}{2!} l''(\widehat{\theta}) + \frac{(\theta - \widehat{\theta})^3}{3!} l'''(\widehat{\theta}) + \cdots . \tag{3.16}$$

Dado que $l'(\widehat{\theta}) = 0$ y $r(\theta) = l(\theta) - l(\widehat{\theta})$, obtenemos que

$$r(\theta) = \frac{(\theta - \widehat{\theta})^2}{2} l''(\widehat{\theta}) + \frac{(\theta - \widehat{\theta})^3}{3!} l'''(\widehat{\theta}) + \cdots \tag{3.17}$$

y definimos la *aproximación normal a* $r(\theta)$ como

$$r_N(\theta) = \frac{(\theta - \widehat{\theta})^2}{2} l''(\widehat{\theta}), \tag{3.18}$$

la motivación de esta definición es que si $\theta - \widehat{\theta}$ es pequeño, a partir del término cúbico en adelante los términos de (3.17) son despreciables, de donde $r(\theta) \approx r_N(\theta)$.

3.4. Propiedades de la verosimilitud

3.4.1. La no-aditividad de la verosimilitud

A diferencia de la probabilidad la adición en la verosimilitud no tiene sentido, una medida de probabilidad como se vio en el capítulo 1 toma valores en una σ-álgebra y los mapea en un intervalo cerrado, los elementos de dicha σ-álgebra son eventos para los cuales está bien definida la probabilidad de la unión de dos eventos, esto es $\Pr(A \cup B) = \Pr(A) + \Pr(B)$ si A y B son disjuntos, sin embargo para la verosimilitud dados θ_1 y θ_2 su unión puede no estar bien definida ya que son hipótesis acerca de los parámetros y esta operación no siempre define una nueva hipótesis (Sprott, D.A. p.13. [16]).

3.4.2. Invarianza funcional

Una característica importante que comparten las verosimilitudes con las probabilidades es la invarianza funcional, la cual no posee la función de densidad, esto es de especial interés ya que cualquier afirmación cuantitativa acerca de θ implica una afirmación equivalente para cualquier función inyectiva $\delta = \delta(\theta)$ la cual se deduce de la sustitución algebraica por $\theta = \theta(\delta)$ por ejemplo para $R_\theta(\theta; y)$ verosimilitud

relativa de θ se tiene que $R_\delta(\delta; y)$ es la respectiva verosimilitud relativa de δ.

3.4.3. Intervalos de verosimilitud

Un aspecto interesante de determinar son los rangos dentro de los cuales se encuentran los valores de θ más verosímiles con determinado valor c, en el caso para el cual $k = 1$ estos intervalos se pueden obtener al trazar una línea horizontal en la gráfica de $R(\theta)$ a una distancia c sobre el eje x donde $0 \leq c \leq 1$ si variamos c en dicho intervalo el EMV $\widehat{\theta}$ estará contenido en todos éstos intervalos y además convergen hacia $\widehat{\theta}$ cuando $c \to 1$ los valores de c más utilizados serán los de $c = 0.05$, 0.15 y 0.25 por razones que se expondrán en la siguiente sección; hay que considerar que éstos intervalos de verosimilitud no establecen la incerteza del intervalo, establecen la verosimilitud relativa de los puntos fuera del intervalo escogido.

Definición 3.5. Un **intervalo de verosimilitud** o región de verosimilitud de nivel c para θ, IV(c) se define como

$$\text{IV}(c) = \{\theta \mid R(\theta; y) \geq c\}, \quad \text{donde} \quad 0 \leq c \leq 1. \tag{3.19}$$

Todo valor $\theta \in \text{IV}(c)$ tiene verosimilitud relativa igual o mayor que c y todo valor de $\theta \notin \text{IV}(c)$, tiene verosimilitud relativa menor. «Es decir que el IV(c) separa los valores plausibles o creíbles de θ de los no plausibles a un nivel c» (Sprott, D. A. p. 14. [16]).

Ejemplo 3.3. Consideremos nuevamente la función binomial, a partir de (3.14) tenemos que

$$R(\theta; y, n) = \frac{\binom{n}{y}\theta^y(1-\theta)^{n-y}}{\binom{n}{y}\widehat{\theta}^y(1-\widehat{\theta})^{n-y}} \tag{3.20}$$

del ejemplo (3.1) sabemos que $\widehat{\theta} = \frac{y}{n}$ de donde

$$\frac{\theta^y(1-\theta)^{n-y}}{\frac{y^y}{n^y}(1-\frac{y}{n})^{n-y}} = n^n \left(\frac{\theta}{y}\right)^y \left(\frac{1-\theta}{n-y}\right)^{n-y}. \tag{3.21}$$

Para ejemplificar esto veamos el caso de un estudio clínico en el que se deseaba probar la eficacia de un medicamento llamado Ramipril usado para aumentar la tasa de superviviencia después de un infarto agudo al miocardio (AIRE Study Group 1993), se escogieron 1986 personas al azar de los cuales a 1004 se les suministró

Tabla 3.1. Resultados del ensayo clínico. Fuente: AIRE Study Group.(1993).[tabla]

Tratamiento	Sobrevivientes	Fallecidos	Total
Ramipril	834	170	1004
Placebo	760	222	982
Total	1594	392	1986

Ramipril y a los restantes 982 se les suministró un placebo es decir un grupo de control, los datos se presentan en la tabla de contingencia 3.1.

A partir de los datos anteriores tenemos verosimilitudes relativas de la forma $R_1(\theta; 834, 1004)$ y $R_2(\theta; 760, 982)$, representando a pacientes tratados con Ramipril y placebo respectivamente, sustituyendo datos observamos que

$$R_1(\theta; 834, 1004) = 1004^{1004} \left(\frac{\theta}{834}\right)^{834} \left(\frac{1-\theta}{1004-834}\right)^{1004-834}$$

y

$$R_2(\theta; 760, 982) = 982^{982} \left(\frac{\theta}{982}\right)^{982} \left(\frac{1-\theta}{982-760}\right)^{982-760}$$

los cuales son representados en la figura 3.1.

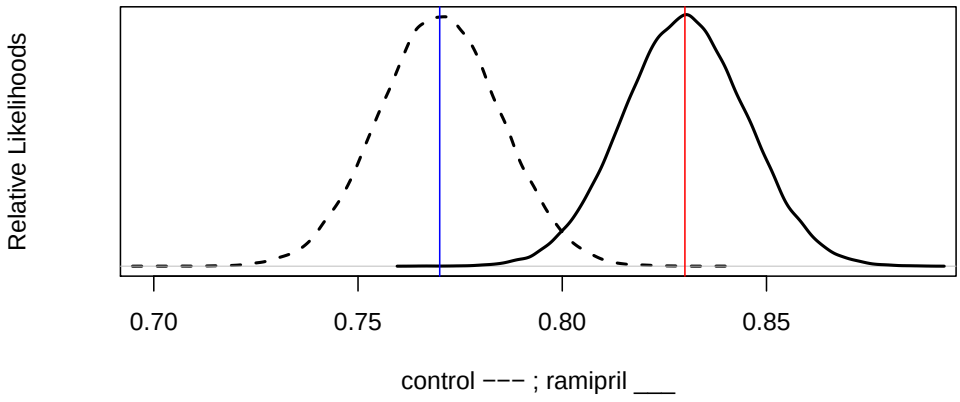


Figura 3.1. Verosimilitudes relativas de pacientes tratados con Ramipril y placebo. Fuente: Elaboración propia con datos de Sprott, D.A. (2000).

Como podemos apreciar las verosimilitudes difieren por mucho y se intersecan en $\theta = 0.803$, lo cual nos muestra que la tasas de supervivencia $\theta > 0.803$ tienen una verosimilitud relativa mayor del 8 % para pacientes tratados con Ramipril y menores al 8 % para pacientes tratados con placebo, el EMV para Ramipril es $\widehat{\theta} = 0.83$ y

para el placebo es $\widehat{\theta} = 0.77$, adicionalmente IV$(0.15) = (0.807, 0.853)$ para Ramipril e IV$(0.15) = (0.747, 0.799)$ para el placebo, todo esto en conjunto sugiere que el tratamiento con Ramipril aumenta la tasa de superviviencia de los pacientes.

3.4.4. Intervalos de verosimilitud confianza

Supongamos que se tiene una muestra $y = (y_1, \ldots, y_n)$ tomada de una distribución de variables aleatorias iid $Y = (Y_1, \ldots, Y_n)$ con función de probabilidad $\Pr(Y = y; \theta)$, donde $\theta = \theta_0$ es un parámetro escalar desconocido y fijo, a partir de esta muestra podemos calcular un intervalo $[A, B]$ para el valor de θ_0; si se hace este mismo procedimiento con muestras diferentes pero tomadas todas de la misma distribución de variables aleatorias iid obtendremos diferentes valores para A y B, y estos extremos a su vez serán variables aleatorias que varían de acuerdo a su muestra correspondiente y por lo tanto en principio su distribución de probabilidad puede ser calculada a partir de la distribución de la variable aleatoria Y, también podemos observar que debido a la variación de los valores de A y B el intervalo $[A, B]$ puede incluir o no al verdadero valor de θ_0, habiendo expuesto lo anterior se procederá a dar las siguientes definiciones.

Definición 3.6. La **probabilidad de cobertura** de un intervalo aleatorio $[A, B]$ es la probabilidad de que el intervalo $[A, B]$ incluya o cubra el verdadero valor del parámetro θ_0 y se denota como

$$\mathrm{PC}(\theta_0) = \Pr(A \leq \theta_0 \leq B; \theta = \theta_0). \tag{3.22}$$

La interpretación de la probabilidad de cobertura es la fracción de veces que el intervalo $[A, B]$ incluirá el verdadero valor de θ_0 para un número muy grande de repeticiones de la muestra, manteniendo el valor de θ fijo en θ_0.

Definición 3.7. Un intervalo $[A, B]$ será llamado **intervalo de confianza** para θ cuando su probabilidad de cobertura no depende de θ_0. Es decir, cuando el valor de $\mathrm{PC}(\theta_0)$ es el mismo para todo valor del parámetro θ_0.

La probabilidad de cobertura de un IV(c) se puede aproximar a través de la distibución de probabilidad de la estadística de la razón de verosimilitud para un $\theta = \theta_0$ fijo, $D_n = -2logR(\theta_0)$, siguiendo esta linea tenemos que

$$\theta_0 \in \mathrm{IV}(c) \Leftrightarrow R(\theta_0) \geq c \Leftrightarrow -2\log R(\theta_0) \leq -2\log(c).$$

En la práctica encontrar la distribución de probabilidad exacta de RV es complicado, la solución a esto se obtiene a través de la teoría asintótica que establece que bajo ciertas condiciones de regularidad,[3] la estadística de la razón de verosimilitud $RV \equiv -2\log R(\theta_0)$, converge en distribución a una ji-cuadrada con un grado de libertad, para cada $\theta_0 \in \Theta \subset \mathbb{R}$, es decir que

$$\lim_{n\to+\infty} \Pr(D_n \leq x; \theta = \theta_0) = \Pr(\chi_1^2 \leq x), \text{ para cada } x > 0.$$

De lo anterior tenemos que

$$\text{PC[IV(c)]} \approx \Pr(\chi_{(1)}^2) \leq x, \text{ donde } x = -2\log(c)$$

de modo que si $x = q_{(\alpha,1)}$ donde $q_{(\alpha,1)}$ es el cuantil $(1-\alpha)$ de la distribución $\chi_{(1)}^2$, el IV(c) con $c = e^{-q_{(\alpha,1)}/2}$, tendrá una probabilidad de cobertura de $(1-\alpha)$ aproximadamente, y dado que IV(c) no depende de θ_0 el intervalo de verosimilitud es a su vez un intervalo de confianza y de ahí el nombre de intervalo de verosimilitud-confianza.

En la tabla 3.2 podemos observar los valores de c que se utilizan para obtener intervalos de verosimilitud-confianza con probabilidad de cobertura del 90, 95 y 99 %. Para el caso del cuantil 90 el valor de $\chi_{(1)}^2$ correspondiente es 2.706 de donde el IV(c) con $c = e^{-2.706/2} = 0.258$ tiene una probabilidad de cobertura del 90 %.

Tabla 3.2. Confianza aproximada de los intervalos de verosimilitud para θ unidimensional. Fuente: Montoya.(2008).[tabla]

$(1-\alpha)$	c	$q_{(\alpha,1)}$
0.90	0.258	2.706
0.95	0.146	3.841
0.99	0.036	6.635

3.5. Parámetros de interés entre parámetros de estorbo

Hasta el momento se han dado ejemplos en los que se busca un parámetro θ de dimensión 1, esto, porque es el caso más sencillo, ahora se abordará el caso en el que la dimensión de θ es mayor o igual a 2. En el ejemplo 3.2 se estudió el caso de la función normal y se supuso que se conocía el parámetro σ, si no suponemos esto

[3]Para detalles de la prueba ver Serfling, R. J. p. 155-156.

tendremos dos parámetros desconocidos, a saber σ y μ lo cual puede escribirse como $\theta = (\sigma, \mu)$ en la práctica muchas veces se necesita estimar únicamente μ de modo que σ vendría a ser un parámetro no necesario o de estorbo, de la misma manera si el parámetro de interés es σ el parámetro μ sería el parámetro de estorbo, en general si se tiene que $\theta = (\theta_1, \ldots, \theta_k) = (\delta, \xi)$ con $\delta = (\theta_1, \ldots, \theta_{k-j})$ y $\xi = (\theta_{k-j+1}, \ldots, \theta_k)$ el parámetro que nos interesa estimar es δ y el parámetro de estorbo es ξ, la motivación de este planteamiento es que si δ y ξ pueden ser separados apropiadamente se podrá estimar δ sin tener que saber nada acerca de ξ, hay que hacer énfasis en que esta idea no siempre funciona, pues estos dos parámetros pueden estar tan estrechamente ligados que sea imposible separarlos sin caer en resultados falaces (Sprott, D. p. 50. [16]).

3.5.1. Verosimilitud condicional

Considere el vector de parámetros $\theta = (\delta, \xi)$, con la estructura de verosimilitud

$$L(\delta, \xi; y) \propto f(y; \delta, \xi) = f(t; \delta, \xi) f(y; \delta \mid t) \propto L_{\text{res}}(\delta, \xi; t) L_c(\delta; y), \tag{3.23}$$

donde δ es un parámetro de interés, y que se tiene un estadístico t suficiente minimal para el parámetro de estorbo ξ; la cantidad $L_c(\delta; y)$ es llamada la *verosimilitud condicional* de δ, dado que está basada en la distribución condicional de la muestra y dado el estadístico t; nótese que el segundo factor no depende de ξ de modo que $L_c(\delta; y)$ nos sirve para hacer inferencias acerca de δ únicamente, cabe mencionar que la calidad de dichas inferencias dependerán de que tan buena es la factorización (3.23) para separar la información del parámetro de interés δ, del parámtero de estorbo ξ, es decir depende de que la cantidad de información que tenga $L_{\text{res}}(\delta, \xi; t)$, al cual llamaremos *función residual de verosimilitud*, del parámetro δ, sea poca cuando desconocemos ξ.

3.5.2. Verosimilitud maximizada o perfil

Para que el método de verosimilitud condicional sea efectivo, se requiere que la función de verosimilitud tenga una estructura especial y esto hace que sea muy restrictiva y poco práctica para la mayoría de casos, por lo que Sprott y Kalbfleisch proponen de manera formal en 1969 la función de verosimilitud maximizada o perfil como un método para estimar parámetros de interés en presencia de parámetros de estorbo.

Definición 3.8. La función de **verosimiltud maximizada** o **perfil** del parámetro de interés δ, $L_p(\delta)$, se define como

$$L_p(\delta; y) = \text{máx} L(\delta, \xi; y) = L[\delta, \widehat{\xi}(\delta, y); y] \tag{3.24}$$

donde $\widehat{\xi}(\delta, y)$ es el estimador de máxima verosimilitud restringido (EMVR) de ξ para un valor específico de δ.

Para obtener la verosimilitud maximizada de δ, $L_p(\delta; y)$ se debe maximizar la función de verosimilitud $L(\delta, \xi; y)$ sobre ξ dejando un valor fijo de δ, de este modo el EMV global de ξ coincidirá con el estimador restringido evaluado en el EMV global $\widehat{\delta}$, $\widehat{\xi} = \widehat{\xi}(\widehat{\delta}, y)$, dicho de otra forma la función de verosimilitud maximizada se obtiene reemplazando el parámetro de estorbo ξ por $\widehat{\xi}(\delta, y)$ que depende de δ en la función de verosimilitud global $L(\delta, \xi; y)$.

En el caso de que tanto δ como ξ sean unidimensionales, la función de verosimilitud global $L(\delta, \xi; y)$ es una superficie en $\mathbb{R}^3$, la verosimilitud es entonces una función definida para la pareja (δ, ξ), si se ve esto de forma gráfica se evidenciará que la sombra o perfil de la verosimilitud global es precisamente la verosimilitud maximizada, y de ahí su nombre. En la figura 3.2 se muestra dicha situación donde el parámetro de interés es θ y el de estorbo p.

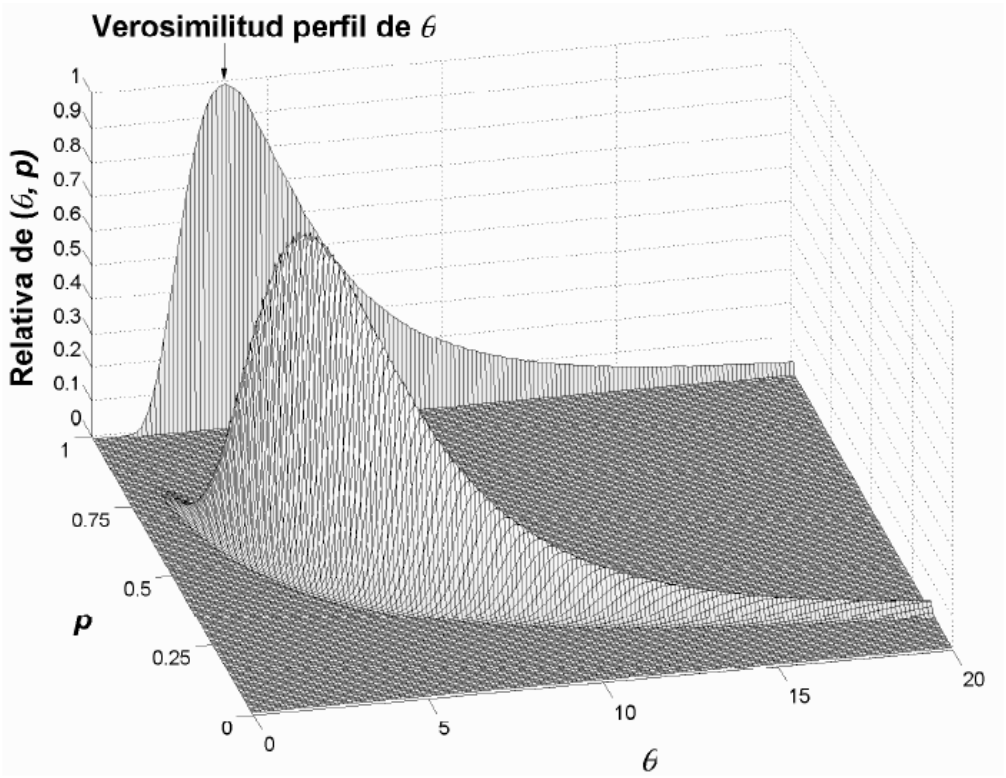


Figura 3.2. Representación de la verosimilitud global y perfil. Montoya, L.(2008). Función de verosimilitud perfil de θ [figura].

4. Aplicación de la función de verosimilitud

Un problema con el cual los ecólogos y biólogos se encuentran frecuentemente es el de estimar el tamaño N de una población de animales (véase Calambokidis [1] y Soisalo [15]) y así determinar si una especie está en situación de vulnerabilidad o peligro de extinción, para ello existen varios modelos, algunos más sencillos que otros, generalmente los modelos sencillos ofrecen una buena aproximación cuando N es grande, sin embargo, cuando N es pequeño puede que exista mucho sesgo en la estimación, a continuación analizaremos un modelo de captura-recaptura desde el punto de vista de la función de verosimilitud para estimar el tamaño N de una población.

4.1. El modelo captura-recaptura

Los modelos de captura-recaptura se basan en tomar una muestra inicial de individuos y marcarlos, luego se devuelven a su hábitat, después de esto se vuelve a tomar una segunda muestra y se les coloca una segunda marca, en esta segunda muestra tendremos dos casos, que el individuo tenga la marca de la primera toma o que no la tenga, si no la tiene quedará con una sola marca, si tiene ya una marca el individuo tendrá una segunda marca y se considera como un individuo con repetición, en el modelo que se propone a continuación se hará este proceso de captura-recaptura y marcaje varias veces capturando un individuo a la vez, intuitivamente nos inclinamos a suponer que la probabilidad de capturar a un mismo individuo varias veces dependerá del tamaño N de la población, siendo más difícil que dicha repetición se dé para N grande, en el siguiente modelo supondremos lo siguiente:

1. La población tiene tamaño $N > 0$, $N \in \mathbb{Z}^+$, la cual no varía durante el tiempo en el que se toma la muestra.

2. Se toma un individuo a la vez, se marca y este es devuelto a la población antes de tomar al siguiente individuo.

3. Los individuos son tomados aleatoriamente y cada uno tiene la misma probabilidad $\frac{1}{N}$ de ser escogido en cada muestra que se toma, de modo que tendremos tomas independientes e idénticamente distribuidas.

4. El número de capturas realizadas se representará con n y la cantidad de individuos diferentes capturados con r.

El procedimiento es el siguiente; se captura un individuo de la población de forma aleatoria y se le coloca una marca, luego este es devuelto a la población y cuando se esté seguro de que se haya mezclado completamente dentro del grupo procedemos a realizar la segunda captura y así sucesivamente, con esto nos aseguramos que cada individuo tiene la misma probabilidad de ser escogido en cada captura incluso aquel que ya haya sido capturado y marcado anteriormente, de modo que la probabilidad de tomar a cada individuo en cada una de las capturas será $\frac{1}{N}$, este proceso se repite n veces, el número n a utilizar se determina por el propio investigador de acuerdo a la población particular que se está estudiando, para nuestro caso definiremos dos conceptos, el de *marcar* y el de *registrar*; *marcar* se refiere a realizar una marca física al individuo capturado, es decir una forma fácil de identificar que el individuo ya fue capturado anteriormente si es que se vuelve a capturar; *registrar* se refiere al registro interno que el investigador lleva y en este caso será un vector ordenado que describe los números de los individuos capturados en cada una de las capturas, el proceso se detalla a continuación y se ilustra con el diagrama 4.1.

1. El primer individuo capturado se marcará con un 1 y se registrará el vector (1).

2. En la segunda captura tendremos dos casos, volver a capturar al mismo individuo, en cuyo caso lo registramos nuevamente con un 1 y el vector registrado será $(1, 1)$ y marcarlo nuevamente no es estrictamente necesario; o bien es un nuevo individuo, en cuyo caso se le marca como 2 y el vector registrado será $(1, 2)$.

3. En tres capturas tenemos cinco casos posibles, a saber; $(1, 1, 1)$, $(1, 1, 2)$, $(1, 2, 1)$, $(1, 2, 2)$, $(1, 2, 3)$.

4. En las siguientes capturas se sigue con este procedimiento, por ejemplo, para cuatro capturas los casos ascienden a quince.

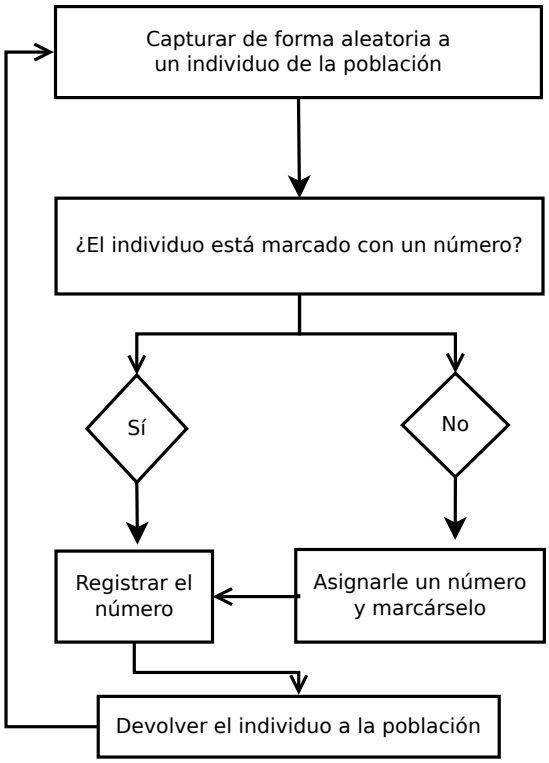


Figura 4.1. Procedimiento del método de captura-recaptura. Fuente: Elaboración propia con programa Dia.

4.2. Deducción de la función de verosimilitud

Hay que tomar en cuenta que la probabilidad de capturar al individuo 1 es $\frac{1}{N}$, al individuo 2 es $1-\frac{1}{N}$ la primera vez y $\frac{1}{N}$ a partir de la segunda vez, la probabilidad de capturar al individuo 3 la primera vez es $1-\frac{2}{N}$ y $\frac{1}{N}$ a partir de la segunda, en general la probabilidad de capturar al individuo i la primera vez, será $1-\frac{i-1}{N}$ y a partir de la segunda será $\frac{1}{N}$; en la tabla 4.1 resumimos todos los casos posibles para las primeras 4 capturas y la probabilidad Pr de ocurrencia de cada caso en términos del tamaño de la población N.

Para $n = 1$ la única posibilidad que se tiene es tener un individuo marcado una sola vez, con probabilidad 1 de que esto suceda, para $n = 2$ podremos tener un individuo con dos marcas o bien dos individuos diferentes cada uno con una marca, con probabilidades de ocurrencia $\frac{1}{N}$ y $1-\frac{1}{N}$ respectivamente, en todo caso la cantidad de individuos diferentes capturados será r, con $r \in \{1, ..., n\}$, además como puede verse en la tabla 4.1 la probabilidad Pr de ocurrencia de cada caso depende únicamente de r, de hecho reordenando términos y haciendo $1-\frac{i}{N} = \frac{N-i}{N}$ puede verse claramente que dadas n capturas y r individuos diferentes capturados en un

$n=1$	Pr	$n=3$	Pr	$n=4$	Pr
(1)	1	$(1,1,1)$	$\frac{1}{N^2}$	$(1,1,1,1)$	$\frac{1}{N^3}$
		$(1,1,2)$	$\frac{1}{N}(1-\frac{1}{N})$	$(1,1,1,2)$	$\frac{1}{N^2}(1-\frac{1}{N})$
$n=2$	Pr	$(1,2,1)$	$(1-\frac{1}{N})(\frac{1}{N})$	$(1,1,2,1)$	$\frac{1}{N}(1-\frac{1}{N})\frac{1}{N}$
		$(1,2,2)$	$(1-\frac{1}{N})(\frac{1}{N})$	$(1,1,2,2)$	$\frac{1}{N}(1-\frac{1}{N})\frac{1}{N}$
$(1,1)$	$\frac{1}{N}$	$(1,2,3)$	$(1-\frac{1}{N})(1-\frac{2}{N})$	$(1,2,1,1)$	$(1-\frac{1}{N})\frac{1}{N^2}$
$(1,2)$	$1-\frac{1}{N}$			$(1,2,1,2)$	$(1-\frac{1}{N})\frac{1}{N^2}$
				$(1,2,2,1)$	$(1-\frac{1}{N})\frac{1}{N^2}$
				$(1,2,2,2)$	$(1-\frac{1}{N})\frac{1}{N^2}$
				$(1,1,2,3)$	$\frac{1}{N}(1-\frac{1}{N})(1-\frac{2}{N})$
				$(1,2,1,3)$	$(1-\frac{1}{N})(\frac{1}{N})(1-\frac{2}{N})$
				$(1,2,2,3)$	$(1-\frac{1}{N})(\frac{1}{N})(1-\frac{2}{N})$
				$(1,2,3,1)$	$(1-\frac{1}{N})(1-\frac{2}{N})\frac{1}{N}$
				$(1,2,3,2)$	$(1-\frac{1}{N})(1-\frac{2}{N})\frac{1}{N}$
				$(1,2,3,3)$	$(1-\frac{1}{N})(1-\frac{2}{N})\frac{1}{N}$
				$(1,2,3,4)$	$(1-\frac{1}{N})(1-\frac{2}{N})(1-\frac{3}{N})$

Tabla 4.1. Probabilidades en función del tamaño de la población N para diferentes casos y valores de n. Fuente: Elaboración propia.

orden en particular, su probabilidad de ocurrencia es

$$\Pr = \frac{(N-1)(N-2)(N-(r-1))}{N^{n-1}} = \frac{N(N-1)(N-2)(N-(r-1))}{N^n} = \frac{{}_NP_r}{N^n} \tag{4.1}$$

así pues los casos anteriores pueden resumirse en la tabla 4.2.

$n=1$	Pr	$n=2$	Pr	$n=3$	Pr	$n=4$	Pr
$r=1$	1	$r=1$	$\frac{N}{N^2}$	$r=1$	$\frac{N}{N^3}$	$r=1$	$\frac{N}{N^4}$
		$r=2$	$\frac{N(N-1)}{N^2}$	$r=2$	$3\cdot\frac{N(N-1)}{N^3}$	$r=2$	$7\cdot\frac{N(N-1)}{N^4}$
				$r=3$	$\frac{N(N-1)(N-2)}{N^3}$	$r=3$	$6\cdot\frac{N(N-1)(N-2)}{N^4}$
						$r=4$	$\frac{N(N-1)(N-2)(N-3)}{N^4}$

Tabla 4.2. Resumen de probabilidades para diferentes casos y valores de n. Fuente: Elaboración propia.

A partir de lo anterior podemos notar que las funciones de probabilidad y de verosimilitud pueden expresarse como:

$$\Pr = \lambda_{(n,r)}\frac{{}_NP_r}{N^n} \tag{4.2}$$

$$L(N) = C_{(n,r)}\cdot\frac{{}_NP_r}{N^n}. \tag{4.3}$$

4.3. Cálculo de los estimadores de máxima verosimilitud

El siguiente paso es encontrar los EMV de N para n, r dados; las gráficas 4.2, 4.3, 4.4 y 4.5 muestran las funciones de verosimilitud presentadas en la tabla 4.2, para las cuales sus máximos representan los EMV.

En estas situaciones, encontrar los máximos de las funciones de verosimilitud a través de un software facilita la tarea; por ejemplo, para $n = 4$ y $r = 3$ usando el siguiente código en R, encontramos que el EMV es 5.

```
##La libreria gtools permite usar la funcion permutations,
##se instala y se carga al entorno de trabajo.
install.packages("gtools")
library(gtools)
##Se miden los tiempos de ejecucion de instrucciones por el CPU
##(user), el tiempo del sistema operativo (system) y el tiempo
##real (elapsed).
t=proc.time()
##Se define un maximo N supuesto, un numero de
##capturas n y un numero de individuos r.
topN=50
n=4
r=3
##Se incicializa un vector que guarda los resultados
l=c()
##Se define un ciclo que evalua los diferentes valores de N.
for (N in r:topN) {l=c(l,(nrow(permutations(N,r,repeats.allowed
   = FALSE))/N^(n)))}
##Se calcula el EMV de todos los N evaluados.
EMV = (r-1)+which(l==max(l))
EMV
##Se cierra la medicion del tiempo
proc.time()-t
```

Encontrar el máximo de la ecuación (4.3) comienza a tornarse complicado para valores de n, r cercanos a 10, de hecho, en el código anterior donde se estableció un valor máximo a evaluar para $N = 50$ y con valores de $n = 4$ y $r = 3$, el tiempo de usuario (user), medido en R fue de 6.72 s,[1] por lo que se hace necesario trabajar con

[1]Medido en una computadora Core i3 con Windows 10.

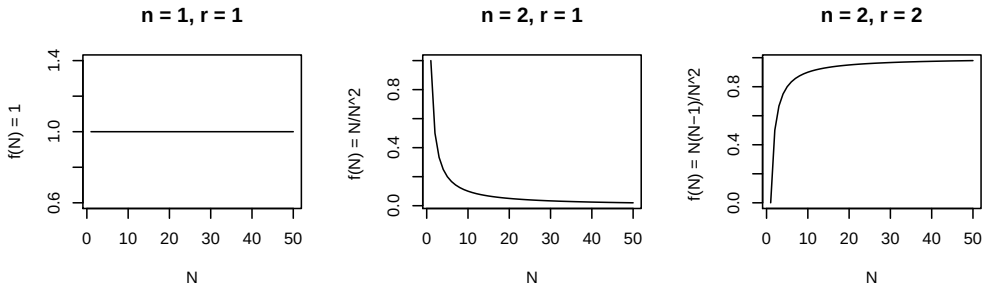


Figura 4.2. Función de verosimilitud correspondiente a $n = 1, 2$. Fuente: Elaboración propia realizada en R.

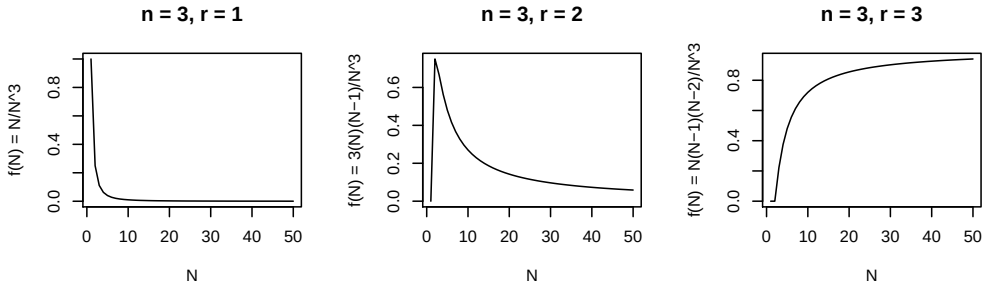


Figura 4.3. Función de verosimilitud correspondiente a $n = 3$. Fuente: Elaboración propia realizada en R.

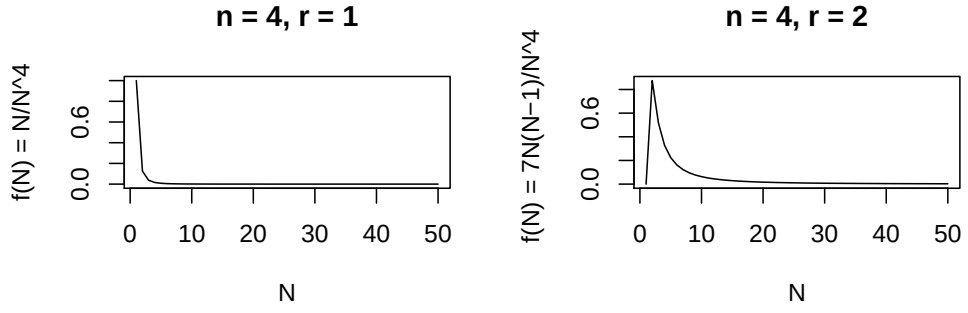


Figura 4.4. Función de verosimilitud correspondiente a $n = 4$, $r = 1, 2$. Fuente: Elaboración propia realizada en R.

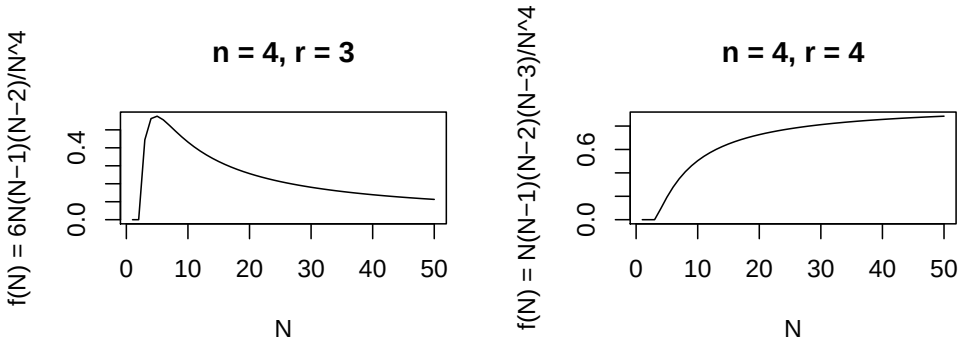


Figura 4.5. Función de verosimilitud correspondiente a $n = 4$, $r = 3, 4$. Fuente: Elaboración propia realizada en R.

la *función log-verosimilitud de* N para disminuir este tiempo, a saber,

$$l(N) = \log C + \log N + \log(N-1) + \cdots + \log(N-(r-1)) - n \log N. \qquad (4.4)$$

De modo que podemos sustituir en el código anterior la función $L(N)$ por $l(N)$ para calcular el EMV, como se muestra a continuación.

```
t=proc.time()
topN=50
n=4
r=3
l=c()
for (N in r:topN ) {
l=c(l,sum(log10((N:(N-(r-1))))))-n*log10(N))
}
EMV = (r-1)+which(l==max(l))
EMV
proc.time()-t
```

El tiempo de usuario en este caso es imperceptible, ya que se registraron 0.01 s, lo cual contrasta evidentemente con el tiempo que tardó encontrar el máximo de la ecuación (4.3).

Si se desea automatizar el proceso de encontrar los EMV para diferentes valores de n, r podemos utilizar el siguiente código que genera una matriz con dichos EMV, en este caso se utilizará $n = 5$ y un valor supuesto máximo a evaluar de $N = 50$.

```
##Se mide el tiempo de ejecucion
t=proc.time()
##Se definen los valores maximos de N y n a evaluar
##Se inicializa un vector y una matriz que guarda los
##resultados
topN=50
topn=5
l=c()
m=c()
result=matrix(data = NA, ncol = topn, nrow = topn,
byrow = FALSE)
result[1,1]<-topN
##Se define un ciclo que evalua los diferentes valores de
##N, n y r
for (n in 2:topn) {
  for (r in 1:n) {
    for (N in r:topN ) {
      l=c(l,sum(log10((N:(N-(r-1)))))-n*log10(N))
    }
    maximo = (r-1)+which(l==max(l))
    l=c()
    result[r,n]<-maximo
  }
}
##Se visualizan los resultados
result
##Se cierra la medicion del tiempo
proc.time()-t
```

El tiempo de usuario medido en este caso fue de 0.06 s, la tabla 4.3 muestra los resultados obtenidos por el código y el análisis de las gráficas anteriores.[2]

Recordemos que N debe ser entero, por lo que el EMV se debe calcular con esta restricción, de lo anterior se pueden observar algunos patrones los cuales se demuestran en el siguiente.

[2]Cuando $n = r$, el EMV calculado por el código es igual a `topN=50`, acá se ha sustituido por $\nexists$ cuando $n = 1$ y por ∞ cuando $n > 1$.

$n=1$	EMV	$n=2$	EMV	$n=3$	EMV	$n=4$	EMV	$n=5$	EMV
$r=1$	$\hat{N}=\nexists$	$r=1$	$\hat{N}=1$	$r=1$	$\hat{N}=1$	$r=1$	$\hat{N}=1$	$r=1$	$\hat{N}=1$
		$r=2$	$\hat{N}=\infty$	$r=2$	$\hat{N}=2$	$r=2$	$\hat{N}=2$	$r=2$	$\hat{N}=2$
				$r=3$	$\hat{N}=\infty$	$r=3$	$\hat{N}=5$	$r=3$	$\hat{N}=3$
						$r=4$	$\hat{N}=\infty$	$r=4$	$\hat{N}=8$
								$r=5$	$\hat{N}=\infty$

Tabla 4.3. Estimadores de máxima verosimilitud de N, para diferentes valores de n y r.

Teorema 4.3.1. *Supongamos que $N > 0$ y además $N \in \mathbb{Z}^+$, entonces para la ecuación* (4.3) *se cumple que:*

1. *Si $r = n = 1$, entonces $\hat{N}$ no existe.*
2. *Si $r = 1$ y $n \geq 2$, entonces $\hat{N} = 1$.*
3. $\lim\limits_{N\to\infty} \frac{{}_NP_n}{N^n} = 1$.
4. *Sean $n, U \in \mathbb{Z}^+$ y $n < U$ entonces es cierta la desigualdad $(1 - n/U) \leq (1 - 1/U)^n$*
5. *Si $N \geq n$ la función $L(N) = C_{(n,r)} \cdot \frac{{}_NP_n}{N^n}$ es monótona creciente.*
6. *Si $N < n$ la función $L(N) = C_{(n,r)} \cdot \frac{{}_NP_n}{N^n} = 0$ para cada $N \in \{1, \ldots, n-1\}$.*
7. *Si $r = n$ y $r > 1$ entonces $\hat{N} = \infty$.*

Demostración. Sea $L(N)$ la *función de verosimilitud* dada en (4.3), con $C = 1$.[3]

1. $L(N) = \frac{{}_NP_r}{N^n} = \frac{{}_NP_1}{N} = \frac{N}{N} = 1$, una función constante la cual no posee un máximo.
2. $L(N) = \frac{{}_NP_1}{N^n} = \frac{N}{N^n} = \frac{1}{N^{n-1}}$, la cual es una función monótona decreciente dado que $N < N+1$ implica que $\frac{1}{N^{n-1}} > \frac{1}{(N+1)^{n-1}}$, de donde $\hat{N} = 1$ necesariamente.
3. $$\lim_{N\to\infty} \frac{{}_NP_n}{N^n} = \lim_{N\to\infty} \frac{\frac{N!}{(N-n)!}}{N^n} = \lim_{N\to\infty} \frac{N(N-1)\cdots(N-(n-1))}{N^n}$$
$$= \lim_{N\to\infty} \frac{N}{N} \cdot \lim_{N\to\infty} \frac{(N-1)}{N} \cdots \lim_{N\to\infty} \frac{N-(n-1)}{N}$$
$$= 1 \cdot 1 \cdots 1 = 1.$$

[3]Recordemos que una adecuada escogencia de la constante hace esto posible.

4. Procederemos a demostrar por inducción:

   Para $n = 1$ tenemos que $(1 - 1/U) \leq (1 - 1/U)$.

   Por hipótesis de inducción para $n = k$ será cierto que $(1 - k/U) \leq (1 - 1/U)^k$.

   Para $n = k + 1$ tenemos que mostrar que $\left(1 - \frac{k+1}{U}\right) \leq \left(1 - \frac{1}{U}\right)^{k+1}$, esto es cierto si y solo si

$$\left(1 - \frac{k}{U} - \frac{1}{U}\right) \leq \left(1 - \frac{1}{U}\right)^k \left(1 - \frac{1}{U}\right) \Leftrightarrow$$

$$\left(1 - \frac{k}{U}\right) - \frac{1}{U} \leq \left(1 - \frac{1}{U}\right)^k \left(1 - \frac{1}{U}\right) \Leftrightarrow$$

$$\left(1 - \frac{k}{U}\right) - \frac{1}{U} \leq \left(1 - \frac{1}{U}\right)^k - \left(1 - \frac{1}{U}\right)^k \frac{1}{U}$$

   .

   Aplicando la hipótesis de inducción esto es equivalente a

$$\left(1 - \frac{1}{U}\right)^k \frac{1}{U} \leq \frac{1}{U} \Leftrightarrow \left(1 - \frac{1}{U}\right)^k \leq 1$$

   quedando demostrado.

5. A partir de que $N < N + 1$ debemos mostrar que $\frac{{}_N P_n}{N^n} \leq \frac{{}_{N+1} P_n}{(N+1)^n}$, esto es cierto si y solo si

$$\frac{\frac{N!}{(N-n)!}}{N^n} \leq \frac{\frac{(N+1)!}{(N+1-n)!}}{(N+1)^n} \Leftrightarrow$$

$$\frac{N(N-1)(N-2)\cdots(N-n+1)}{N^n} \leq \frac{(N+1)(N)(N-1)\cdots(N+1-n+1)}{(N+1)^n}$$

$$\Leftrightarrow$$

$$\frac{N-n+1}{N^n} \leq \frac{N+1}{(N+1)^n} = \frac{1}{(N+1)^{n-1}}$$

$$\Leftrightarrow$$

$$\log\left(\frac{N-n+1}{N^n}\right) \leq \log\left(\frac{1}{(N+1)^{n-1}}\right)$$

$$\Leftrightarrow$$

$$\log(N+1-n) - n\log(N) \leq -(n-1)\log(N+1) =$$

$$= -n\log(N+1) + \log(N+1)$$

$$\Leftrightarrow$$

$$\log(N+1-n) - \log(N+1) \leq n\log(N) - n\log(N+1)$$

$$\Leftrightarrow$$

$$\log\left(\frac{N+1-n}{N+1}\right) \leq n\log\left(\frac{N}{N+1}\right),$$

haciendo la sustitución $U = N+1$ obtenemos

$$\log\left(\frac{U-n}{U}\right) \leq n\log\left(\frac{U-1}{U}\right) \Leftrightarrow$$

$$\log(1-n/U) \leq n\log(1-1/U) \Leftrightarrow$$

$$\log(1-n/U) \leq \log(1-1/U)^n \Leftrightarrow$$

$$(1-n/U) \leq (1-1/U)^n$$

que ya se demostró cierta en el inciso anterior.

6. Dado que $\frac{{}_NP_n}{N^n} = \frac{\frac{N!}{(N-n)!}}{N^n} = \frac{N(N-1)(N-2)\cdots(N-n+1)}{N^n}$ cualquier valor que tome N entre 1 y $n-1$ hará el numerador igual a 0.

7. Se deduce de 3, 5 y 6.

□

4.4. Estimación del tamaño de una población de mariposas

Finalmente como último ejemplo tenemos el problema de estimar el tamaño N de una población de mariposas, el cual es analizado en Craig, C.C. [3], quien para diferentes valores de n, r observados, estima los valores de N utilizando el método de máxima verosimilitud y suponiendo una distribución de Poisson; dichos datos podemos verlos en la tabla 4.4.

Modificando el código anterior con `topN=1060` y `topn=250`, generamos una matriz con los EMV para valores de $n = 1, \ldots, 250$, comparando estos resultados con los presentados por Craig se verifica que son los mismos.

```
t=proc.time()
topN=1060
topn=250
l=c()
m=c()
result=matrix(data = NA, ncol = topn, nrow = topn,
byrow = FALSE)
result[1,1]<-topN
for (n in 2:topn) {
  for (r in 1:n) {
    for (N in r:topN ) {
      l=c(l,sum(log10((N:(N-(r-1)))))-n*log10(N))
    }
  maximo = (r-1)+which(l==max(l))
  l=c()
  result[r,n]<-maximo
  }
}
result
proc.time()-t
```

Valores observados		Estimaciones de N
r	n	Resultados presentados por Craig C.C.
69	72	828
93	108	348
159	187	557
144	161	703
63	76	193
56	66	192
48	74	78
341	435	853
276	330	892
222	249	1059
154	225	275
148	180	442
63	91	114
60	98	90
71	118	104
46	89	58

Tabla 4.4. Estimación del tamaño de una población de mariposas. Fuente: Elaboración propia con datos de Craig, C. C. (1953).

CONCLUSIONES

1. Los conceptos de σ-álgebra, conjuntos medibles, espacios medibles, medidas, medidas de probabilidad, espacios de probabilidad y funciones de probabilidad son la base fundamental en la construcción de la teoría de probabilidades.

2. La teoría de verosimilitud en estadística inferencial, nos brinda herramientas simples y eficientes para estimar parámetros desconocidos en diferentes ámbitos de aplicación.

3. En general, es preferible el uso de la *función log-verosimilitud* a la *función de verosimilitud* para encontrar los EMV, debido a que la *función log-verosimilitud* requiere menos tiempo de ejecución en máquina.

4. El método de captura-recaptura presentado en este trabajo es útil para estimar el tamaño de una población de individuos, y presenta resultados similares a los presentados en el artículo de Craig C.C.[3]

RECOMENDACIONES

1. Profundizar más en la construcción axiomática de la teoría de probabilidades.

2. Consultar la bibliografía recomendada para ampliar los conocimientos en teoría de verosimilitud.

3. Valorar las ventajas y desventajas que la *función de verosimilitud* ofrece, al estimar parámetros poblacionales, comparado con otros métodos de estimación.

4. Formar equipos multi-disciplinarios para implementar el método de captura-recaptura en la estimación del tamaño de una población animal, que se encuentre vulnerable, en peligro de extinción o en peligro crítico de extinción en el país.

BIBLIOGRAFÍA

[1] Calambokidis, John. (Jan. 2004). Abundance of blue and humpback whales in the eastern north pacific estimated by capture-recapture and line-transect methods. *Marine Mammal Science*, 20 No. 1, 63-85.

[2] Coles, S. (2001). *An Introduction to Statistical Modeling of Extreme Values*. Londres: Springer.

[3] Craig, C.C. (Jun. 1953). On the Utilization of Marked Specimens in Estimating Populations of Flying Insects. *Biometrika Trust*, 40 No. 1/2, 170-176.

[4] Darroch, J.N. and Ratcliff, D. (Mar. 1980). A Note on Capture-Recapture Estimation. *Biometrics*, 36 No.1, 149-153.

[5] Feller, William. (1967). *An Introduction to Probability Theory and Its Applications*, Volume 1 . Estados Unidos: Wiley.

[6] Gutierrez, William. (2010). Introducción a TEX y a LATEX 2_ε. Guatemala: Facultad de Ingeniería, USAC.

[7] Kalbfleisch, J.G. (1985). *Probability and Statistical Inference*, Volume 2. Estados Unidos: Springer.

[8] Kreyszig, Erwing. (1979). *Introducción a la Estadística Matemática Principios y Métodos*. México: Limusa.

[9] Montoya, J. (2008). *La verosimilitud perfil en la Inferencia Estadística*. 2008: CIMAT.

[10] Montoya, J. A., Díaz-Francés, E. y Sprott, D. A. (2009). On a criticism of the profile likelihood function. *Statistical Papers*, V.50, 195-202.

[11] Roussas, G. (1997). *A Course in Mathematical Statistics*. Estados Unidos: Academic Press.

[12] Rudin, W. (1987). *Real and Complex Analysis*. Estados Unidos: McGraw-Hill

[13] Salinero, Ruiz. (s.f.) *Historia de la Teoría de la Probabilidad*. Recuperado de https://slidex.tips/download/historia-de-la-teoria-de-la-probabilidad.

[14] Serfling, R.J. (1980). *Approximation Theorems of Mathematical Statistics*. Estados Unidos: John Wiley & Sons.

[15] Soisalo, Marianne K. and Cavalcanti, Sandra M.C. (Jan. 2006). Estimating the density of a jaguar population in the Brazilian Pantanal using camera-traps and capture-recapture sampling in combination with GPS radio-telemetry. *Elsevier*, I29, 487-496.

[16] Sprott, D.A. (2000). *Statistical Inference in Science*. Canadá/México: Springer.

[17] Wackerly, D., Mendenhall, W. y Scheaffer, R. (2010). *Estadística Matemática con Aplicaciones*. México: Cengage Learning.