UNIVERSIDAD TECNOLÓGICA DE LA MIXTECA

Ajuste de la distribución de valores extremos generalizada a las precipitaciones pluviales máximas del Estado de Oaxaca

T E S I S

PARA OBTENER EL TÍTULO DE:
LICENCIADO EN MATEMÁTICAS APLICADAS

PRESENTA:
Luz Antonia Pacheco Santiago

DIRECTOR DE TESIS:
Dr. José del Carmen Jiménez Hernández

CO-DIRECTOR DE TESIS:
Dr. Eyther Matías Martín González

H. CD. HUAJUAPAN DE LEÓN, OAXACA, MARZO DE 2025

Agradecimientos

A la Universidad Tecnológica de la Mixteca (UTM), especialmente al Instituto de Física y Matemáticas, por la formación recibida durante cinco años.

Al Dr. José del Carmen Jiménez Hernández por su apoyo y dedicación durante el desarrollo de esta tesis. Su amplio conocimiento, disposición para resolver mis dudas y sus oportunas sugerencias fueron fundamentales para la realización de este trabajo. Agradezco también su infinita paciencia y el tiempo que dedicó a revisar cada detalle.

Al Dr. Ehyter Matías Martín González por su invaluable experiencia, compromiso y rigor científico, los cuales fueron clave para el desarrollo de este trabajo. Sus observaciones precisas contribuyeron de manera significativa a mi formación. Le agradezco, además, su paciencia inagotable y su permanente disposición para aclarar mis dudas.

A todos los sinodales por el apoyo en las revisiones. A todos los profesores que formaron parte de mi formación como licenciada.

A mis amigos Hael, Raymundo, Mendiola, Yulisa y Román. Su compañía convirtió los desafíos en experiencias más llevaderas y los logros en victorias aún más significativas.

A Gildardo, mi primer amigo en la universidad; por su apoyo durante la elaboración de esta tesis. Gracias por estar siempre presente, brindándome fortaleza y motivación.

A mi padre, por sus palabras de aliento que me motivaron a seguir adelante en momentos de incertidumbre, por su infinito apoyo incondicional y su incansable esfuerzo.

A mi hermana, por su motivación y por enseñarme, con su forma de ser, que siempre se puede ir más allá.

A mi madre, cuyo amor infinito y esfuerzo incansable siguen guiando mi camino, incluso en su ausencia. Sin su dedicación, sus enseñanzas y el inmenso cariño que siempre me brindó, este logro no habría sido posible.

A mis padres, por su amor.
A mi yo del futuro.

◇

Contenido

Lista de tablas

Lista de tablas

Lista de figuras

Introducción

El origen de la teoría de valores extremos (TVE) puede ser bastante antiguo; no fue hasta el año 1950 que la metodología fue propuesta de manera seria y formal para el modelado de fenómenos físicos (Coles *et al*., 2001).

Esta teoría se encuentra en constante crecimiento desde los comienzos del siglo XX y desempeña un papel fundamental en los estudios relacionados con las mediciones físicas, en las que se aplica con la finalidad de describir el comportamiento de eventos atípicos. Históricamente, la TVE se ha aplicado con éxito en el ámbito de las ciencias ambientales en el tratamiento estadístico de datos sobre el nivel del mar, la velocidad del viento, el caudal de los ríos, las temperaturas y las precipitaciones extremas, entre otros. En finanzas, se usa para modelar riesgos de pérdidas extremas y crisis económicas. En ingeniería, se emplea para diseñar infraestructuras resilientes ante cargas extremas, como terremotos o inundaciones. También tiene aplicaciones en biología, seguros y ciencia de datos, donde la predicción de eventos poco frecuentes es crucial para la toma de decisiones y la gestión de riesgos.

Las precipitaciones pluviales máximas son un fenómeno de interés para la TVE y han sido poco estudiadas en la República Mexicana, pero tienen grandes impactos y efectos en el sector agrícola, ganadero, en otros como servicios de energía eléctrica, transporte y en salud (Englehart y Douglas, 2000). El análisis estadístico de datos de lluvia es de suma importancia cuando se desean desarrollar medidas preventivas en casos de eventos que implican algún tipo de riesgo (González y Macías, 2011).

El Estado de Oaxaca, debido a su ubicación geográfica y su diversidad topográfica, enfrenta diversos riesgos asociados a lluvias máximas. Estos fenómenos pueden generar afectaciones significativas en infraestructura, comunidades y el medio ambiente, por ejemplo,

- Las zonas bajas y costeras, como el Istmo y las cuencas de los ríos, son propensas a inundaciones severas debido al desbordamiento de ríos y acumulación de agua en áreas urbanas como Oaxaca de Juárez.
- Las regiones montañosas de la Sierra Sur y Sierra Norte son vulnerables a deslaves debido a la reblandecimiento del suelo por lluvias intensas, afectando caminos, viviendas y comunidades enteras.

- Los ríos como el Atoyac, el Papaloapan y el Tehuantepec pueden incrementar su caudal peligrosamente, afectando comunidades ribereñas y generando corrientes que dañan carreteras y viviendas. Las lluvias pueden provocar derrumbes y bloqueos en carreteras que atraviesan zonas montañosas, afectando el tránsito y el abastecimiento de bienes esenciales. Tramos como la autopista Oaxaca-Cuacnopalan y la carretera a la Costa suelen ser afectados. Oaxaca es afectado por ciclones tropicales y huracanes provenientes del Pacífico, los cuales pueden traer lluvias torrenciales, vientos fuertes y marejadas en la costa, impactando municipios como Salina Cruz, Huatulco y Puerto Escondido.

En este trabajo se realiza un análisis probabilístico para observar el comportamiento de los datos sobre las precipitaciones pluviales que se disponen de los registros meteorológicos diarios recabados por las estaciones climatológicas a cargo del Servicio Meteorológico Nacional (SMN) de la Comisión Nacional del Agua (CONAGUA). Se dispone de un total de 59 estaciones meteorológicas del Estado de Oaxaca, con información actualizada, cuyos datos corresponden al período de enero de 2000 a diciembre de 2017. Sin embargo, se seleccionan 8 estaciones meteorológicas, una por región, para hacer dicho análisis. Se trabaja con la hipótesis de que la distribución de valores extremos generalizada se ajusta de manera adecuada a las precipitaciones pluviales máximas del Estado de Oaxaca.

Un enfoque para la modelación de valores extremos es a partir de la distribución de valores extremos generalizada. Esta distribución se ajusta a los valores máximos y mínimos de datos. Otro enfoque para el análisis de valores extremos es a partir del análisis de excedentes sobre umbrales. En este trabajo se contempla el análisis del primer enfoque para el caso univariado. Por tanto, el objetivo de este trabajo es ajustar la distribución de valores extremos generalizada a datos de precipitación pluvial del Estado de Oaxaca, aplicando el método de máximos por bloques.

Para el desarrollo de este trabajo, en el Capítulo 1 se presentan de forma clara los conceptos básicos de probabilidad y estadística que se utilizarán para el análisis y parte de la Teoría de Valores Extremos.

En el Capítulo 2 se presentan los resultados de la teoría clásica de valores extremos, como la formulación del modelo y la distribución de valores extremos generalizada. Se muestra la descripción del modelo de máximos por bloques, el cual se utilizará en la aplicación del trabajo.

En el Capítulo 3 se explica cómo se realiza la inferencia para la estimación de los parámetros y el análisis que se lleva a cabo para obtener los niveles de retorno. En el Capítulo 4 se muestra la aplicación realizada a los datos de precipitaciones pluviales del Estado de Oaxaca, específicamente se muestra el análisis de tres estaciones meteorológicas del Estado. En él, también se realiza la descripción del problema y se explica cómo se trabajó con los datos.

Finalmente, se exponen las conclusiones en el Capítulo 5 que sintetizan los hallazgos más relevantes del análisis realizado, destacando las principales interpretaciones obtenidas a partir

de los datos y resultados. Además, se incorporaron dos apéndices que contienen información adicional de utilidad, como el código utilizado para el procesamiento y análisis de los datos, así como los registros y resultados correspondientes a las cinco estaciones meteorológicas que no se presentan en el Capítulo 4.

Capítulo 1

Preliminares

Una descripción informativa de cualquier conjunto de datos está dada por la frecuencia de repetición o arreglo distribucional de las observaciones en el conjunto. La probabilidad es un mecanismo por medio del cual pueden estudiarse sucesos aleatorios, cuando éstos se comparan con los fenómenos determinísticos.
La probabilidad tiene un papel crucial en la aplicación de la inferencia estadística porque una decisión, cuyo fundamento se encuentra en la información contenida en una muestra aleatoria, puede estar equivocada. Sin una adecuada comprensión de las leyes básicas de la probabilidad, es difícil utilizar la metodología estadística de manera efectiva.
En este capítulo se introduce la notación básica de probabilidad y estadística que se utilizará a lo largo de esta tesis. Para el desarrollo de este capítulo se utilizó la siguiente bibliografía Contreras y Jiménez-Hernández (2020); Rincón (2006); Canavos y Medal (1987); Lladser (2011); Murray y Spiegel (2009); y Rossi (2018).

1.1. Conceptos de probabilidad

La teoría de la probabilidad se encarga del estudio de los fenómenos o experimentos aleatorios. Por experimento aleatorio se entenderá todo aquel experimento que, cuando se le repite bajo las mismas condiciones iniciales, el resultado que se obtiene no siempre es el mismo. El ejemplo más sencillo y cotidiano de un experimento aleatorio es el de lanzar una moneda o un dado, y aunque estos experimentos pueden parecer muy sencillos, algunas personas los utilizan para tomar decisiones en su vida. En principio, no se sabe cuál será el resultado del experimento aleatorio, así que por lo menos conviene agrupar en un conjunto a todos los resultados posibles.

Definición 1.1.1. *El conjunto de todos los posibles resultados de un experimento aleatorio recibe el nombre de* ***espacio muestral****.*

El conjunto de todos los posibles resultados puede ser finito, infinito numerable o no

numerable. Se dice que un espacio muestral es *discreto* si su resultado puede ponerse en una correspondencia uno a uno con el conjunto de los enteros positivos, y se dice que un espacio muestral es *continuo* si sus resultados consisten en un intervalo de números reales.
El espacio muestral (o espacio muestra) de un experimento se denota generalmente por la letra griega Ω (omega). En algunos textos se usa también la letra S para denotar el espacio muestral. Esta letra proviene del término sampling space de la lengua inglesa, equivalente a espacio muestral.

Definición 1.1.2. *Un **evento** del espacio muestral es un grupo de resultados contenidos en éste que tienen una característica en común.*

En otras palabras, se llamará evento a cualquier subconjunto del espacio muestral y se denotará por las primeras letras del alfabeto en mayúsculas: *A, B, C,* etc. Es importante mencionar que a cada uno de los resultados posibles del experimento aleatorio se le llama elemento o miembro del espacio muestral y se denota por ω (omega minúscula).

Ejemplo 1.1.1. *Considere el experimento aleatorio que consiste en lanzar un dado honesto y observar el número que aparece en la cara superior, entonces el espacio muestral es el conjunto* $\Omega = \{1, 2, 3, 4, 5, 6\}$. *Como ejemplo de un evento para este experimento se define el conjunto* $A = \{2, 4, 6\}$, *que corresponde al suceso de obtener como resultado un número par.*

Definición 1.1.3 (Probabilidad)**.** *Sea* Ω *un espacio muestral y E cualquier evento de este. Se llamará función de probabilidad, medida de probabilidad o simplemente probabilidad sobre el espacio muestral* Ω *a* $\mathbb{P}(E)$ *si satisface los siguientes axiomas,*

1. $\mathbb{P}(E) \geq 0$.
2. $\mathbb{P}(\Omega) = 1$.
3. *Si, para eventos* $E_1, E_2, E_3, \ldots, E_i \cap E_j = \emptyset$ *para toda* $i \neq j$, *entonces* $\mathbb{P}(E_1 \cup E_2 \cup \cdots) = \mathbb{P}(E_1) + \mathbb{P}(E_2) + \cdots$.

Estas propiedades se obtienen a partir de la definición de la probabilidad clásica. La **probabilidad frecuentista** supone que se realizan n repeticiones de un cierto experimento aleatorio y que se registra el número de veces que ocurre un determinado evento E.

Definición 1.1.4 (Probabilidad frecuentista)**.** *Sea* ε *un experimento aleatorio con espacio muestral* Ω *y sea* $E \subseteq \Omega$ *un evento. Suponga que se realizan n repeticiones del experimento aleatorio y sea* n_E *el número de veces que ocurre el evento E en los n ensayos del experimento. La **probabilidad frecuentista** está dada por,*

$$P(E) = \lim_{n \to \infty} \frac{n_E}{n}.$$

Es claro que por la definición anterior, n_E puede tomar valores de 0 y hasta n, ya que no es posible realizar una infinidad de veces dicho experimento aleatorio. En la práctica, no es

posible encontrar de manera exacta la probabilidad de un evento cualquiera. Sin embargo, se puede tener una aproximación empírica, siempre y cuando n sea suficientemente grande, mediante,

$$P(E) \approx \frac{n_E}{n}.$$

La probabilidad frecuentista cumple con las mismas propiedades que la probabilidad clásica. Por último, se menciona la probabilidad axiomática, la cual no establece una forma explícita de obtener la probabilidad de un evento en el espacio muestral; sin embargo, establece un conjunto de reglas que dicha probabilidad debe cumplir.

Definición 1.1.5 (Probabilidad axiomática)**.** *Sea A un evento del espacio muestral Ω.*

1. *La probabilidad de todo evento en el espacio muestral debe ser no negativa, es decir,*

$$P(A) \geq 0, \quad para \quad todo \quad evento \quad A \subseteq \Omega.$$

2. *La probabilidad del evento seguro Ω, es igual a 1, es decir,*

$$P(\Omega) = 1.$$

3. *Para cualquier sucesión infinita de eventos disjuntos dos a dos, $A_1, A_2, ...$ se cumple,*

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i).$$

1.1.1. Variable aleatoria

Una variable aleatoria es una función que asigna un número real a cada elemento del espacio muestral. Esto permite considerar que el resultado del experimento aleatorio es el número real tomado por la variable aleatoria. Así, el interés en el estudio de los experimentos aleatorios se traslada al estudio de variables aleatorias y sus características particulares.

Definición 1.1.6 (Variable aleatoria)**.** *Sea Ω un espacio muestral de algún experimento aleatorio E. Una variable aleatoria X una función del espacio muestral Ω al conjunto de los números reales, es decir, $X : \Omega \to \mathbb{R}$.*
Si $\omega \in \Omega$, con $X(\omega)$ se denota la imagen de ω bajo X. A la imagen de X se le denomina soporte de X y será denotado por $R(X)$, se define como,

$$R(X) = X(\Omega) = \{x \in \mathbb{R} | \omega \in \Omega : X(\omega) = x\},$$

donde $R(X)$ es nuevamente un espacio muestral.

Es decir, X transforma todos los posibles resultados del espacio muestral en cantidades numéricas. Por conveniencia, al escribir v.a. se hace referencia a una variable aleatoria o a

variables aleatorias. Las letras V, W, X, Y, Z denotarán variables aleatorias, en caso de necesitar más v.a. se ocuparán subíndices, es decir $X_1, X_2, X_3, ...X_n$.

Ejemplo 1.1.2. *Considere el experimento de lanzar una moneda 10 veces, de la cual al lanzar una vez la moneda solo se puede obtener cara o cruz, cuyo espacio muestral asociado consiste* 2^{10} *diferentes resultados. Así, si se define X igual al número de caras de los 10 lanzamientos y* $Y = 10 - X$*, que es el número de veces que cae cruz, se tiene que ambas son variables aleatorias.*

Ejemplo 1.1.3. *Considere el experimento donde se registra la asistencia a clase en un curso de treinta alumnos. El espacio muestral asociado a este experimento aleatorio es el conjunto* $\{0, 1, 2, \ldots, 30\}$*. Una v.a. asociada a este experimento es el número de alumnos presentes en clase. Si este número se denomina como X entonces el suceso "ningún alumno se presentó en clase" puede escribirse como* $[X = 0]$*. Por otro lado, el evento "al menos dos tercios de los alumnos asistió a clase" corresponde a* $[20 \leq X \leq 30]$*. También se puede escribir este evento como* $[X \geq 20]$*. Esto se debe a que la preimagen del intervalo cerrado* $[20, 30]$ *bajo la transformación X es la misma que el intervalo* $[20, +\infty)$*.*

Definición 1.1.7. *Una variable aleatoria X es discreta si* $R(X)$ *es un conjunto finito o infinito numerable.*

Los conjuntos finitos o infinitos numerables son llamados conjuntos discretos. Por ejemplo, el conjunto $\{0, 1, 2, ..., n\}$ es un conjunto discreto porque es finito, de la misma manera $\mathbb{N}$ es un conjunto discreto dado que aunque este es un conjunto infinito, es numerable y por lo tanto discreto.

Definición 1.1.8. *Una variable aleatoria X es continua si* $R(X)$ *es un intervalo de la recta de los números reales.*

Esta clasificación de v.a. es la usual, pero no es completa, ya que existen variables que no pertenecen a ninguno de los dos tipos mencionados con anterioridad. En el estudio de este escrito, únicamente se ocuparan v.a. continuas.
Dada una v.a. existen dos funciones que proveen de información acerca de las características de la variable aleatoria: la función de densidad de probabilidad o función de densidad (f.d.p.) y la función de distribución acumulada o acumulativa (f.d.a.); éstas permiten representar al mismo tiempo tanto los valores que puede tomar la v.a. como las probabilidades de los distintos eventos de interés.
Dado que $a \leq X \leq b$ es un evento, la f.d.p. proporciona un medio para calcular la probabilidad de que la variable X tome un valor dentro del intervalo $[a, b]$.

Definición 1.1.9. *Sea X una v.a. continua. La función* $f_X : \mathbb{R} \to [0, \infty)$ *es la función de densidad de probabilidad de X si para cualquier intervalo* $(a, b) \in \mathbb{R}$ *se cumple la igualdad,*

$$P(a \leq X \leq b) = \int_a^b f_X(x)dx.$$

Toda función de densidad $f_X(x)$ de una variable aleatoria continua cumple con las siguientes propiedades,

a) $f_X(x) \geq 0$ *para cualquier valor* $x \in \mathbb{R}$.

b) $\int_{-\infty}^{\infty} f_X(x)dx = 1$.

Puesto que el área total bajo $f_X(x)$ es uno, la probabilidad del intervalo $a \leq X \leq b$ es el área acotada por la función de densidad y las rectas $X = a$ y $X = b$.

Ejemplo 1.1.4. *Suponga que la v.a. X tiene una función de densidad, dada por*

$$f_X((x) = \begin{cases} ce^{-3x}, & si\ x > 0, \\ 0, & x \leq 0. \end{cases}$$

Hallar (a) la constante c que haga a $f_X((x)$ *una f.d.p., (b)* $P(1 < X < 2)$ *y (c)* $P(X \geq 3)$.

Para contestar (a), se debe cumplir que,

$$\int_{-\infty}^{\infty} f_X((x)dx = 1.$$

Así,

$$\int_0^{\infty} ce^{-3x}dx = c\int_0^{\infty} e^{-3x}dx,$$

haciendo

$$u = -3x, du = -3dx, \int \frac{e^u}{3}du = \frac{e^u}{-3}.$$

Luego,

$$\int_0^{\infty} ce^{-3x}dx = c\int_0^{\infty} e^{-3x}dx = c\left[\frac{e^{-3x}}{-3}\right]\Bigg|_0^{\infty} = -\frac{c}{3}(0-1) = \frac{c}{3},$$

por tanto $c = 3$.

Para contestar (b) $P(1 < X < 2) =$

$$\int_1^2 3e^{-3x}dx = 3\int_1^2 e^{-3x}dx = -\left[e^{-3x}\right]\Big|_1^2 = -(e^{-6} - e^{-3}) = e^{-3}(1 - e^{-3}) = 0.0473.$$

Para contestar (c) $P(X \geq 3) =$

$$\int_3^{\infty} 3e^{-3x}dx = 3\int_3^{\infty} e^{-3x}dx = -[e^{-3x}]]_3^{\infty} = e^{-9} = 0.0001.$$

Definición 1.1.10. *Sea X una variable aleatoria continua. La función de distribución acumulativa o acumulada de la variable aleatoria X es la función $F_X : \mathbb{R} \to [0, 1]$, definida como,*

$$F_X(x) = P(X \leq x), \quad para \quad todo \quad x \in \mathbb{R}.$$

En particular, si X es una v.a. continua por definición, la f.d.a. está dada por

$$F_X(x) = P(X \leq x) = \int_{-\infty}^{x} f_X(t)dt,$$

en donde t es una variable artificial de integración. Así, la función acumulada $F_X(x)$ es el área acotada por la función de densidad que se encuentra a la izquierda de la recta $X = x$. Dado que para cualquier variable aleatoria continua X,

$$P(X = x) = \int_{x}^{x} f_X(t)dt = 0,$$

entonces:

$$P(X \leq x) = P(X < x) = F_X(x).$$

Más aún, por el teorema fundamental del cálculo se cumple que $F'_X(x) = f_X(x)$. Así, se puede determinar $f_X(x)$ a partir de $F_X(x)$.

Proposición 1.1.1. *Toda función de distribución acumulada $F_X(x)$ cumple con lo siguiente,*

- $\lim_{x \to -\infty} F_X(x) = F_X(-\infty) = 0$, *el límite de $F_X(x)$ a la izquierda es cero.*
- $\lim_{x \to \infty} F_X(x) = F_X(\infty) = 1$, *el límite de $F_X(x)$ a la derecha es 1.*
- *Si $x_1 \leq x_2$, entonces $F_X(x_1) \leq F_X(x_2)$, $F_X(x)$ es una función monótona creciente.*
- *Si $x_1 \leq x_2$, entonces $P(x_1 < X < x_2) = F_X(x_2) - F_X(x_1)$.*
- *$F_X(x)$ es una función continua por la derecha.*

Definición 1.1.11. *Dada una función de distribución acumulada $F_X(x)$ de una v.a. continua X tal que $F_X(x)$ es estrictamente monótona y continua en el intervalo $(0, 1)$, entonces la función cuantil se denota por $Q(x)$ y se define como la función inversa de $F_X(x)$ en este intervalo, es decir,*

$$Q(x) = F_X^{-1}(x) \quad para \quad todo \quad x \in (0, 1).$$

Definición 1.1.12. *El cuantil de probabilidad acumulada p o cuantil de orden p de la función de distribución $F_X(x)$, con $p \in (0, 1)$, se define como,*

$$Q_p = F_X^{-1}(p).$$

1.1.2. Esperanza, varianza y momentos

A continuación se mencionan algunas características numéricas que ayudarán a identificar las variables aleatorias de manera única.

Definición 1.1.13. *El valor esperado o media de una v.a. continua X, denotado por μ_X o por $E(X)$, es el promedio o valor medio de X y está dado por,*

$$\mu_X = E(X) = \int_{-\infty}^{\infty} x f_X(x) dx,$$

donde $f_X(x)$ es la función de densidad de probabilidad inducida por la variable aleatoria X.

La esperanza es un valor numérico que indica el promedio de los diferentes valores que puede tomar la v.a.

Nota 1.1.1. *La esperanza existe y se dice que X es una v.a. con esperanza finita si la integral impropia de la definición es absolutamente convergente: $\int_{-\infty}^{\infty} |x| f_X(x) dx < \infty$, en caso contrario se dice que la v.a. X no tiene esperanza finita.*

Proposición 1.1.2. *Sea X una v.a. continua y $Y = g(X)$ una v.a. con $g : \mathbb{R} \to \mathbb{R}$, entonces,*

$$\mu_Y = E(Y) = E[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x) dx.$$

Note que la esperanza μ_Y existe si $Y = g(X)$ es una v.a. con esperanza finita.

Definición 1.1.14. *Sea X una v.a. continua, la varianza de X se denota por σ_X^2 o por $Var(X)$ y se define como,*

$$\sigma_X^2 = Var(X) = E[(X - \mu_X)^2] = \int_{-\infty}^{\infty} (x - \mu_X)^2 f_X(x) dx.$$

La varianza es una medida de dispersión de la distribución de probabilidad de una v.a. La desviación estándar es la raíz cuadrada positiva de la varianza, se denota por σ_X. A continuación se define el *r*-ésimo momento de una v.a.

Definición 1.1.15. *Sea X una v.a. continua, el r-ésimo momento de X alrededor del cero se define por,*

$$\mu_x^r = E[X^r] = \int_{-\infty}^{\infty} x^r f_X(x) dx.$$

para $r \in \mathbb{N}$.

Nota 1.1.2. *El primer momento alrededor del cero es el valor esperado o media de la v.a. Se considera además como el valor alrededor del cual los valores de la v.a. tienden a agruparse.*

1.1.2.1. Variables aleatorias estandarizadas y degeneradas

Definición 1.1.16. *La v.a. estandarizada Z de una v.a. X con media μ_X y desviación estándar $\sigma_X > 0$ está definida por*

$$Z = \frac{X - \mu_X}{\sigma_X}.$$

Las propiedades de una v.a. estandarizada son:

- $\mu_Z = 0$.
- $\sigma_Z = 1$.

Definición 1.1.17. *Una variable aleatoria discreta X es degenerada si existe una constante c en $\mathbb{R}$ tal que $P(X = c) = 1$.*

En otras palabras, la definición anterior quiere decir que el soporte de X se limita a un único valor c. Además, se puede decir que X es constante con probabilidad 1.
La distribución asociada a una v.a. degenerada se denomina distribución degenerada (en caso contrario, se conoce como no degenerada).

Nota 1.1.3. *La función de probabilidad de una v.a. X degenerada está dada por $P(X = x) = 1$, $x = c$.*
Mientras que la función de distribución acumulada está dada por,

$$F(x) = \begin{cases} 0, & si\ x < c, \\ 1, & si\ x \geq c. \end{cases}$$

Además , $E[X] = c$, $E[X^r] = c^r$ para $r \in \mathbb{N}$ y $Var(X) = 0$.
La v.a. X es degenerada si y sólo si $Var(X) = 0$.

El concepto de v.a. que toma valores en $\mathbb{R}$ se puede extender a variables aleatorias que toman valores en $\mathbb{R}^n$. Este tipo de funciones se conocen como variables aleatorias multidimensionales o vectores aleatorios.

Definición 1.1.18. *Una variable aleatoria de dimensión n es una función $X : \Omega \to \mathbb{R}^n$ dada por $X = (X_1, X_2, ..., X_n)$, donde cada coordenada $X_1, X_2, ..., X_n$ es una variable aleatoria.*

Sólo se consideran vectores aleatorios cuyas componentes sean todas variables aleatorias discretas o todas variables aleatorias continuas. En tal caso se les conoce como vector aleatorio discreto o continuo.

Definición 1.1.19. *Sea $(X_1, X_2, ..., X_n)$ un vector aleatorio continuo. La función $f_{X_1,X_2,...,X_n}(x_1, x_2, ..., x_n) : \mathbb{R}^n \to [0, \infty)$ es la función de densidad del vector $(X_1, X_2, ..., X_n)$,*

también conocida como la función de densidad de probabilidad conjunta de las variables $X_1, X_2, ..., X_n$ *si cumple,*

$$P(a_1 \leq X_1 \leq b_1, ..., a_n \leq X_n \leq b_n) = \int_{a_1}^{b_1} \cdots \int_{a_n}^{b_n} f_{X_1,...,X_n}(x_1, ..., x_n) dx_n \ldots dx_1,$$

para cualesquiera valores de a_i *y* b_i *en* $\mathbb{R}$ *con* $a_i < b_i$ *para* $i = 1, 2, ..., n$.

La función de densidad de probabilidad conjunta de las variables siempre es mayor o igual que cero para cada valor x_i en $\mathbb{R}$ y la probabilidad descrita en la definición anterior es igual a 1. A continuación se da la definición para la función de distribución acumulada conjunta.

Definición 1.1.20. *Sea* $(X_1, ..., X_n)$ *un vector aleatorio. La función de distribución de probabilidad acumulada del vector* $(X_1, ..., X_n)$ *o función de distribución acumulada conjunta de* $X_1, ..., X_n$ *es la función* $F_{X_1,...,X_n}(x_i, ..., x_n) : \mathbb{R}^n \to [0, 1]$ *dada por,*

$$F_{X_1,...,X_n}(x_i, ..., x_n) = P(X_1 \leq x_1, ..., X_n \leq x_n)$$

para todo $(x_1, ..., x_n) \in \mathbb{R}^n$.

Nota 1.1.4. *Si* $(X_1, ..., X_n)$ *es un vector de variables aleatorias continuas, la función de distribución acumulada conjunta de* $X_1, ..., X_n$ *está dada por,*

$$F_{X_1,...,X_n}(x_1, ..., x_n) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_n} f_{X_1,...,X_n}(u_1, ..., u_n) du_n \ldots du_1,$$

donde $u_1, ..., u_n$ *son variables artificiales de integración. Más aún, conociendo la función de distribución acumulada conjunta se puede obtener la función de densidad conjunta, basta derivar parcialmente n-veces la función de distribución conjunta con respecto a* $x_1, ..., x_n$.

La independencia entre variables aleatorias suele ser conveniente, ya que en el estudio de fenómenos se busca que la probabilidad de ocurrencia de eventos no dependa de otros; de esta forma, es más sencillo poder analizar y modelar el comportamiento.

Definición 1.1.21. *Sea* $(X_1, X_2, ..., X_n)$ *un vector aleatorio continuo. Se dice que las variables aleatorias* $X_1, X_2, ..., X_n$ *son independientes si la función de densidad de probabilidad conjunta* $f_{X_1,X_2,...,X_n}(x_1, x_2, ..., x_n) = \prod_{i=1}^{n} f_{X_i}(x_i)$, *para todo vector* $(x_1, x_2, ..., x_n) \in \mathbb{R}^n$, *donde cada* $f_{X_i}(x_i)$ *es la función de densidad de probabilidad de cada v.a.* X_i.

Nota 1.1.5. *Una sucesión de variables aleatorias* $X_1, ..., X_n$ *es independiente e idénticamente distribuida (i.i.d.) si todas las variables son independientes y tienen la misma distribución de probabilidad.*

1.2. Conceptos de estadística

La estadística se encarga de recolectar, organizar, resumir y analizar datos para después obtener conclusiones a partir de ellos. Principalmente, la estadística se divide en dos áreas: la estadística descriptiva (es una colección de métodos para la organización, resumen y presentación de datos) y la estadística inferencial (técnicas que ayudan a conocer, con determinado grado de confianza, cierta información de la población con base en la información de la muestra obtenida).
Suponiendo que se tiene una **población** de interés, es decir, un conjunto arbitrario de personas, mediciones u objetos cualesquiera (en este caso, la población de interés son las mediciones de precipitaciones pluviales del estado de Oaxaca) y se desea conocer cierta información de esta población. Debido a la imposibilidad o no conveniencia de tener información de cada elemento de la población, se procede a tomar un pequeño subconjunto de la población, el cual se conoce como **muestra**.

Definición 1.2.1. *Una muestra aleatoria es una colección de variables aleatorias $X_1, ..., X_n$, de una población, que son i.i.d.*

Es decir, que cada una de las v.a. X_i para $i = 1, ..., n$, son independientes y tiene la misma distribución de probabilidad con los mismos parámetros ($E(X_i) = \mu$ y $Var(X_i) = \sigma^2$, para $i = 1, 2..., n$). El tamaño de la muestra está dado por el valor que tome n.

Definición 1.2.2. *Una estadística muestral (o simplemente estadístico) es una variable aleatoria $T(X_1, ..., X_n)$ en donde $X_1, ..., X_n$ es una muestra aleatoria y $T : \mathbb{R}^n \to \mathbb{R}$ es una función.*

Dado que T es una función de variables aleatorias, es en sí misma una variable aleatoria. Si se emplea una estadística T para estimar un parámetro desconocido θ, entonces T recibe el nombre de estimador de θ, y el valor específico de t como un resultado de los datos muestrales recibe el nombre de estimación de θ.
En resumen, un estimador es una estadística que identifica el mecanismo funcional por medio del cual se obtiene una estimación.

1.2.1. Principios de estimación

Un objetivo en el modelado estadístico es usar información muestral para hacer inferencias sobre la estructura de probabilidad de la población de la cual surgieron los datos. En el caso más simple, se supone que los datos $x_1, ..., x_n$ son independientes de la distribución de la población. La inferencia equivale a la estimación de esta distribución para la cual existen dos enfoques distintos: paramétrico o no paramétrico. A lo largo del desarrollo de la tesis se ocupará el enfoque paramétrico.

Se estudiará el caso de una variable aleatoria continua cuya función de densidad de probabilidad existe. Suponiendo que los datos $x_1, ..., x_n$ comprenden realizaciones independientes de una variable aleatoria X cuya función de densidad de probabilidad pertenece a una familia conocida de distribuciones de probabilidad con funciones de densidad,

$$\mathcal{F} = \{f(x;\theta) : \theta \in \Theta\}.$$

La inferencia se reduce, por tanto, a la estimación del parámetro θ. El valor del parámetro θ puede ser un escalar, como $\theta = p$ en la familia binomial o puede representar un vector de parámetros como $\theta = (\mu, \sigma)$ en la familia normal.
Supóngase por el momento que el parámetro $\theta \in \mathcal{F}$ es escalar en lugar de un vector. Una función de variables aleatorias que se utiliza para estimar el verdadero valor del parámetro θ se denomina estimador; el valor particular del estimador para un conjunto de datos se conoce como estimación. Dado que los datos son resultados de variables aleatorias, las repeticiones del experimento generarían datos diferentes y, por lo tanto, una estimación diferente. Así, la aleatoriedad en el proceso de muestreo induce aleatoriedad en el estimador. La distribución de probabilidad inducida en un estimador se dice que es su distribución muestral. Dado que es deseable que las estimaciones estén cerca del valor del parámetro que se están estimando, se define el sesgo de un estimador $\hat{\theta}_0$ de θ_0 por,

$$Bias(\hat{\theta}_0) = E(\hat{\theta}_0) - \theta_0,$$

y el error cuadrático medio por,

$$MSE(\hat{\theta}_0) = E\{(\hat{\theta}_0 - \theta_0)^2\}.$$

Se dice que un estimador cuyo sesgo es cero es insesgado; esto corresponde a un estimador cuyo valor, en promedio, es el verdadero valor del parámetro. Un criterio más común para la evaluación del estimador es que su error cuadrático medio (MSE) debe ser pequeño. Dado que el MSE mide la variación del estimador en torno al valor real del parámetro, un MSE bajo implica que, en cualquier muestra en particular, es probable que la estimación se acerque al valor real del parámetro.
La desviación estándar de la distribución muestral de $\hat{\theta}_0$ se llama **error estándar** y se denota por $SE(\hat{\theta}_0)$.

1.2.2. Estimador de máxima verosimilitud

Un método general para la estimación del parámetro desconocido θ_0 dentro de una familia $\mathcal{F}$ es el método por máxima verosimilitud. Cada valor $\theta \in \Theta$ define un modelo en $\mathcal{F}$ que asocia diferentes probabilidades a los datos observados. La probabilidad de los datos observados como función de θ se denomina función de verosimilitud. Los valores de θ que tienen alta probabilidad corresponden a modelos que dan alta probabilidad a los datos observados. Este método, en general, es el que asigna la mayor probabilidad al parámetro de los datos observados. Con mayor detalle, a continuación se define la función de verosimilitud.

Definición 1.2.3. *Sea $X_1, ..., X_n$ una muestra aleatoria de una población con función de densidad $f_X(x, \theta)$. La función de verosimilitud de la muestra se denota por $L(\theta) = L(\theta, x_1, ..., x_n)$ y se define como la función de densidad de probabilidad conjunta de $X_1, ..., X_n$, es decir,*

$$L(\theta) = \prod_{i=1}^{n} f_{X_i}(x_i; \theta).$$

A menudo es más conveniente tomar logaritmos y trabajar con la **función log-verosimilitud**,

$$\ell(\theta) = \log L(\theta) = \sum_{i=1}^{n} \log f_{X_i}(x_i; \theta).$$

El estimador de máxima verosimilitud (EMV) de $\hat{\theta_0}$ de θ_0 se define como el valor de θ que maximiza la función de verosimilitud. Dado que la función logaritmo es monótona, la log-verosimilitud toma su máximo en el mismo punto que la función de verosimilitud, de modo que el estimador de máxima verosimilitud también maximiza la correspondiente función log-verosimilitud.

1.2.3. Inferencia utilizando la función de verosimilitud perfil

En la subsección anterior se describe un método para hacer inferencias sobre una componente particular θ_i de un vector de parámetros θ. Un método alternativo, y normalmente más preciso, se basa en la verosimilitud perfil. La log-verosimilitud perfil de θ se puede escribir formalmente como $\ell(\theta_i, \theta_{-i})$, donde θ_{-i} denota todas las componentes de θ excluyendo θ_i. Además, la log-verosimilitud perfil para θ_i se define como,

$$\ell_p(\theta_i) = \max_{\theta - i} \ell(\theta_i, \theta_{-i}).$$

Es decir, para cada valor de θ_i, la log-verosimilitud perfil es la probabilidad logarítmica máxima con respecto a todos los demás componentes de θ. En otras palabras, $\ell_p(\theta_i)$ es el perfil de la superficie de probabilidad logarítmica vista desde el eje θ_i. Esta definición se generaliza a la situación en la que θ se puede dividir en dos componentes, $(\theta^{(1)}, \theta^{(2)})$, de los cuales $\theta^{(1)}$ es el vector k-dimensional de interés y $\theta^{(2)}$ corresponde a las componentes restantes $(d - k)$. La log-verosimilitud perfil de $\theta^{(1)}$ ahora se define como,

$$\ell_p(\theta^{(1)}) = \max_{\theta^{(2)}} \ell(\theta^{(1)}, \theta^{(2)}).$$

Si $k = 1$ esto se reduce a la definición anterior.
El siguiente resultado conduce a un procedimiento para inferencias aproximadas sobre el estimador de máxima verosimilitud (EMV) de $\theta^{(1)}$.

Teorema 1.2.1. *Sean $x_1, ..., x_n$ realizaciones independientes de una distribución dentro de una*

familia paramétrica $\mathcal{F}$, $\hat{\theta}_0$ el estimador de máxima verosimilitud del parámetro del modelo d-dimensional $\theta_0 = (\theta^{(1)}, \theta^{(2)})$, donde $\theta^{(1)}$ es un subconjunto k-dimensional de θ_0. Luego, en condiciones de regularidad adecuadas, para n suficientemente grande,

$$D_p(\theta^{(1)}) = 2\{\ell(\hat{\theta}_0) - \ell_p(\theta^{(1)})\} \sim \chi^2_k.$$

El Teorema 1.2.1 se utiliza con frecuencia en dos situaciones diferentes. En primer lugar, para una componente $\theta_i, C_\alpha = \{\theta_i : D_p(\theta_i) \leq c_\alpha\}$ es un intervalo de confianza $(1-\alpha)$, donde c_α es el cuantil $(1-\alpha)$ de la distribución χ^2_k. La segunda aplicación es la selección de modelos, es decir, si $\mathcal{M}_1$ es un modelo con un vector de parámetros θ, y el modelo $\mathcal{M}_0$ es el subconjunto del modelo $\mathcal{M}_1$ que se obtiene al restringir k de las componentes de θ para que sean cero, entonces θ puede dividirse como $\theta = (\theta^{(1)}, \theta^{(2)})$, donde la primera componente, de dimensión k, es cero en el modelo $\mathcal{M}_0$. Ahora, si sucede que $\ell_1(\mathcal{M}_1)$ es el EMV para el modelo $\mathcal{M}_1$, y $\ell_0(\mathcal{M}_0)$ el EMV para el modelo $\mathcal{M}_0$, además definiendo,

$$D = 2\{\ell_1(\mathcal{M}_1) - \ell(\mathcal{M}_0)\},$$

como el estadístico de desviación; por el Teorema 1.2.1, $C_\alpha = \{\theta^{(1)} : \mathcal{D}_p(\theta^{(1)}) \leq c_\alpha\}$ comprende una región de confianza $(1-\alpha)$ para el verdadero valor de $\theta^{(1)}$, donde $\mathcal{D}_p$ es la desviación del perfil y c_α es el cuantil $(1-\alpha)$ de la distribución χ^2_k. Por lo tanto, para comprobar si $\mathcal{M}_0$ es una reducción plausible del modelo $\mathcal{M}_1$, basta con comprobar si 0 se encuentra en C_α, lo que equivale a comprobar si $D < c_\alpha$. Esto se denomina una prueba de **razón de verosimilitud**, que se resume de la siguiente manera.

Teorema 1.2.2. *Supóngase que $\mathcal{M}_0$ con parámetro $\theta^{(2)}$ es el submodelo de $\mathcal{M}_1$ con parámetro $\theta_0 = (\theta^{(1)}, \theta^{(2)})$ bajo la restricción de que el subvector k-dimensional $\theta^{(1)} = 0$. Sean $\ell_0(\mathcal{M}_0)$ y $\ell_1(\mathcal{M}_1)$ los valores maximizados de la log-verosimilitud para los modelos $\mathcal{M}_0$ y $\mathcal{M}_1$ respectivamente. Una prueba de la validez del modelo $\mathcal{M}_0$ en relación con $\mathcal{M}_1$ al nivel de significación α consiste en rechazar $\mathcal{M}_0$ a favor de $\mathcal{M}_1$ si $D = 2\{\ell_1(\mathcal{M}_1) - \ell(\mathcal{M}_0)\} > c_\alpha$ donde c_α es el cuantil $(1-\alpha)$ de la distribución χ^2_k.*

Por último, se observa que, bajo regularidad adicional, cada una de las aproximaciones descritas es válida cuando $x_1, \ldots, x_n$ son observaciones independientes pero no idénticamente distribuidas de una familia indexada por un parámetro θ. Por ejemplo, en un modelo de regresión clásico, $X_i \sim \mathcal{D}(\alpha + \beta\omega_i)$ para $i = 1, \ldots, n$, donde $\mathcal{D}(\theta)$ denota una distribución con parámetro θ y $\omega_1, \ldots, \omega_n$ son constantes conocidas. Aunque cada uno de los X_i tiene una distribución diferente, el estimador de máxima verosimilitud de (α, β) sigue satisfaciendo las propiedades indicadas en el Teorema 1.2.2.

1.2.4. Propiedades de los EMV y condiciones de regularidad

Una de las propiedades fundamentales de un EMV es la invarianza funcional.

Definición 1.2.4. *Sea $\theta = (\theta_1, ..., \theta_n)$ un vector n-dimensional de parámetros de una distribución. Si $\tau : \mathbb{R}^n \to \mathbb{R}$ es cualquier función, entonces a $\tau(\theta)$ se le llama función parametral o paramétrica.*

Es decir, a cualquier función de un parámetro o de un vector de parámetros se le conoce como función paramétrica.

Ejemplo 1.2.1. *Algunos ejemplos de funciones paramétricas son los siguientes:*

1. *$\tau(n, p) = np$, para el caso de la distribución binomial de parámetros n y p.*
2. *$\tau(\theta) = E(X)$, donde X es la variable aleatoria en estudio.*
3. *Los cuantiles de orden p de la distribución de estudio.*

Suponga que $X_1, ..., X_n$ es una muestra aleatoria de una población cuya función de densidad está dada por $f_X(x; \theta)$ y $\hat{\theta}$ es el EMV de θ. Si se desea estimar una función del parámetro θ, es decir, una función parametral $\tau(\theta)$, se puede hacer mediante $\tau(\hat{\theta})$. A este resultado se le conoce como principio de invarianza, y viene dado por el siguiente teorema.

Teorema 1.2.3 (Principio de invarianza)**.** *Sea $\hat{\theta}$ el estimador de máxima verosimilitud de θ, entonces el estimador de máxima verosimilitud de cualquier función paramétrica $\tau(\theta)$ está dado por $\tau(\hat{\theta})$.*

Este principio de invarianza establece que una vez calculado el EMV $\hat{\theta}$, el EMV de cualquier función de θ se obtiene por simple sustitución.

Ejemplo 1.2.2. *Dada $X_1, ..., X_n$ una muestra aleatoria de una población X con $X \sim N(\mu, \sigma^2)$, suponga que el EMV de μ es $\hat{\mu} = \bar{X}$, por el principio de invarianza, los EMV de las funciones paramétricas $\tau_1(\mu) = 3\mu$, $\tau_2(\mu) = 3\mu^2$ y $\tau_3(\mu) = 1/\mu$ están dados por $\tau_1(\hat{\mu}) = 3\bar{X}$, $\tau_2(\hat{\mu}) = 3\bar{X}^2$ y $\tau_3(\hat{\mu}) = 1/\bar{X}$.*

El principio de invarianza se puede generalizar de la siguiente manera.

Teorema 1.2.4. *Sea $\hat{\theta} = (\hat{\theta_1}, ..., \hat{\theta_k})$ el estimador de máxima verosimilitud de $\theta = (\theta_1, ..., \theta_k)$. Si $\tau(\theta) = (\tau_1(\theta), ..., \tau_r(\theta))$ para $1 \leq r \leq k$ es un vector r -dimensional de funciones paramétricas, entonces el estimador de máxima verosimilitud de $\tau(\theta) = (\theta_1, ..., \theta_k)$ es $\tau(\hat{\theta}) = (\tau_1(\hat{\theta}), ..., \tau_r(\hat{\theta}))$, donde $\tau_j(\hat{\theta})$ es el estimador de máxima verosimilitud de $\tau_j(\theta)$ para $j = 1, ..., r$.*

Ejemplo 1.2.3. *Si $X_1, ..., X_n$ es una muestra aleatoria de una población X con $X \sim N(\mu, \sigma^2)$ donde μ y σ^2 son desconocidas, $\theta = (\mu, \sigma^2)$ y el EMV de θ es*

$$\hat{\theta} = \left(\bar{X}, (1/n)\sum_{i=1}^{n}(X_i - \bar{X})^2\right),$$

entonces, por el teorema anterior se tiene que el EMV de $\tau(\theta) = \ln(\mu) + 5\sigma$ *está dado por*

$$\tau(\hat{\theta}) = \ln(\bar{X}) + 5\sqrt{(1/n)\sum_{i=1}^{n}(X_i - \bar{X})^2}.$$

A continuación se dará una definición para poder hablar de otras propiedades fundamentales.

Definición 1.2.5 (Condiciones de regularidad)**.** *Sea* $X_1, ..., X_n$ *una muestra aleatoria de una población con una función de densidad* $f_X(x;\theta)$ *y* $T(X_1, ..., X_n)$ *un estimador de la función paramétrica* $\tau(\theta)$*. Las condiciones de regularidad de la función de densidad* $f_X(x;\theta)$ *y del estimador* $T(X_1, ..., X_n)$ *son las siguientes:*

1. *El soporte de la función de densidad* $f_X(x;\theta)$ *dado por el conjunto* $\{x|f_X(x;\theta) > 0\}$ *no depende del parámetro* θ*.*
2. *Para todo* x *en el soporte de* $f_X(x;\theta)$*, la función* $\ln f_X(x;\theta)$ *es diferenciable respecto de* θ*.*
3. $\dfrac{\partial}{\partial\theta}\int\cdots\int\prod_{i=1}^{n} f_X(x_i;\theta)dx_1\cdots dx_n = \int\cdots\int\dfrac{\partial}{\partial\theta}\prod_{i=1}^{n} f_X(x_i;\theta)dx_1\cdots dx_n.$
4. $\dfrac{\partial}{\partial\theta}\int\cdots\int T(x)\prod_{i=1}^{n} f_X(x_i;\theta)dx_1\cdots dx_n = \int\cdots\int T(x)\dfrac{\partial}{\partial\theta}\prod_{i=1}^{n} f_X(x_i;\theta)dx_1\cdots dx_n.$
5. $0 < E\left[\left(\dfrac{\partial}{\partial\theta}\ln\left[f_X(x;\theta)\right]\right)^2\right] < \infty.$

Dada la definición anterior, se pueden mencionar otras propiedades fundamentales del EMV.
Insesgadez asintótica: Bajo las condiciones de regularidad, si $\hat{\theta}_n$ es el EMV del parámetro θ basado en una m.a. de tamaño n, entonces $\hat{\theta}_n$ es insesgado asintóticamente, es decir,

$$\lim_{n\to\infty} E(\hat{\theta}_n) = \theta.$$

Consistencia: Bajo las condiciones de regularidad, si $\hat{\theta}_n$ es el EMV del parámetro θ basado en una m.a. de tamaño n, entonces $\hat{\theta}_n$ es consistente para θ, es decir, se cumplen las siguientes dos condiciones.

- $\lim_{n\to\infty} E(\hat{\theta}_n) = \theta.$
- $\lim_{n\to\infty} Var(\hat{\theta}_n) = 0.$

Por lo tanto, todo EMV es insesgado asintóticamente y su varianza converge a cero.

1.2.5. Prueba de hipótesis estadística

Una hipótesis estadística es una afirmación con respecto a alguna característica desconocida de una población de interés. La esencia de probar una hipótesis estadística es decidir si la afirmación se encuentra apoyada por la evidencia experimental que se obtiene a través de una muestra aleatoria. La afirmación involucra ya sea algún parámetro o alguna forma funcional no conocida de la distribución de interés a partir de la cual se obtiene una muestra aleatoria.

Definición 1.2.6. *Si una hipótesis estadística asigna valores particulares a todos los parámetros desconocidos e identifica la forma funcional de la distribución de interés recibe el nombre de hipótesis sencilla o simple. De otra forma, se conoce como hipótesis compuesta.*

Una hipótesis nula H_0 es una conjetura sobre lo que se espera obtener. Además, debe considerarse como verdadera a menos que exista suficiente evidencia en contra.
Si se decide rechazar H_0 entonces puede que se rechace algo que es cierto (decisión incorrecta) o que se rechace algo que en realidad es falso (decisión correcta). Si no se puede rechazar H_0, entonces no puede rechazarse algo que es cierto (decisión correcta), o no puede rechazarse algo que en realidad es falso (decisión incorrecta). De manera que existen dos posibilidades de tomar una decisión equivocada con respecto al verdadero estado de la naturaleza. La decisión de rechazar H_0 no necesariamente significa que H_0 sea falsa; pero la evidencia muestral con base en la cual se toma la decisión proporciona un grado de confiabilidad (paralelo al de la estimación de intervalo) con el que puede procederse como si H_0 fuese falsa.

Definición 1.2.7. *Una prueba de una hipótesis estadística con respecto a alguna característica desconocida de la población de interés es cualquier regla para decidir si se rechaza la hipótesis nula con base en una muestra aleatoria de la población.*

La decisión se basa en alguna estadística apropiada, la cual recibe el nombre de estadística de prueba. Para ciertos valores de la estadística de prueba, la decisión será el rechazar la hipótesis nula. Estos valores constituyen lo que se conoce como la región crítica de la prueba.
Para construir una regla de decisión apropiada en la prueba de una hipótesis estadística, también es necesario establecer una *hipótesis alternativa* que refleje el valor posible o intervalo de valores del parámetro de interés si la hipótesis nula es falsa. La hipótesis alternativa representa alguna forma de negación de la hipótesis nula y se denota por H_1 o por H_A. Por ejemplo, si una hipótesis nula es $p = 0.5$, la hipótesis alternativa puede ser $p = 0.7$, $p \neq 0.5$ o $p > 0.5$.

Errores de tipo I y tipo II
Si se rechaza una hipótesis que debería aceptarse, se dice que se comete un error tipo I. Si, por otro lado, se acepta una hipótesis que debería rechazarse, se comete un error tipo II. En cualquiera de los casos, ha habido una decisión errónea o se ha hecho un juicio erróneo.
Para que las reglas de decisión (o pruebas de hipótesis) sean buenas, deben diseñarse de manera que se minimicen los errores de decisión. Esto no es sencillo, ya que para cualquier

tamaño de muestra dado, al tratar de disminuir un tipo de error suele incrementarse el otro tipo de error. En la práctica, un tipo de error puede ser más importante que otro y habrá que sacrificar uno con objeto de limitar al más notable. La única manera de reducir los dos tipos de error es aumentando el tamaño de la muestra, lo que no siempre es posible.
Cuando se prueba determinada hipótesis, a la probabilidad máxima con la que se está dispuesto a cometer un error tipo I se le llama nivel de significancia de la prueba. Esta probabilidad acostumbra denotarse por α y por lo general se especifica antes de tomar cualquier muestra para evitar que los resultados obtenidos influyan sobre la elección del valor de esta probabilidad.
En la práctica, se acostumbran los niveles de significancia 0.05 o 0.01, aunque también se usan otros valores. Si, por ejemplo, al diseñar la regla de decisión se elige el nivel de significancia 0.05 (o bien 5 %), entonces existen 5 posibilidades en 100 de que se rechace una hipótesis que debía ser aceptada; es decir, se tiene una confianza de aproximadamente 95 % de que se ha tomado la decisión correcta. En tal caso se dice que la hipótesis ha sido rechazada al nivel de significancia 0.05, lo que significa que la hipótesis tiene una probabilidad de 0.05 de ser errónea.

Valor *p* en pruebas de hipótesis
El valor p es la probabilidad de obtener un estadístico muestral tan extremo o más extremo que el obtenido, suponiendo que la hipótesis nula sea verdadera. Para probar una hipótesis empleando este método se establece un valor α; se calcula el valor p y si el valor $p \leq \alpha$, se rechaza H_0. En caso contrario, no se rechaza H_0. En pruebas para medias empleando muestras grandes ($n > 30$), el valor p se calcula como sigue:

1. Para $H_0 : \mu = \mu_0$ y $H_1 : \mu < \mu_0$, valor $p = P(Z <$ el estadístico de prueba calculado).
2. Para $H_0 : \mu = \mu_0$ y $H_1 : \mu > \mu_0$, valor $p = P(Z >$ el estadístico de prueba calculado).
3. Para $H_0 : \mu = \mu_0$ y $H_1 : \mu \neq \mu_0$,
   $p = P(Z <$ |estadístico de prueba calculado|) + $P(Z >$ |estadístico de prueba calculado|).

El estadístico de prueba calculado es $\dfrac{\bar{x} - \mu_0}{(s/\sqrt{n})}$, donde $\bar{x}$ es la media de la muestra, s es la desviación estándar de la muestra y μ_0 es el valor que se ha especificado para μ en la hipótesis nula. Observe que σ no se conoce, se estima a partir de la muestra y se usa s. Este método para pruebas de hipótesis es equivalente al método de hallar el o los valores críticos y si el estadístico de prueba cae en la región de rechazo, rechazar la hipótesis nula. Usando cualquiera de estos métodos se llega a la misma decisión.

1.2.6. Pruebas de bondad y ajuste

La prueba de bondad del ajuste compara los resultados de una muestra aleatoria con aquellos que se esperaba observar si la hipótesis nula es correcta. La comparación se hace mediante

la clasificación de los datos que se observan en cierto número de categorías y entonces se comparan las frecuencias observadas con las esperadas en cada categoría. Para un tamaño específico del error tipo I, la hipótesis nula será rechazada si existe una diferencia suficiente entre las frecuencias observadas y las esperadas.

1.2.6.1. La prueba de bondad y ajuste chi-cuadrada

Una prueba de bondad y ajuste se emplea para decidir cuándo un conjunto de datos se apega a una distribución de probabilidad dada. Considere una muestra aleatoria de tamaño n de la distribución de una v.a. X dividida en k clases exhaustivas y mutuamente excluyentes, y sea N_i, $i = 1, 2, ..., k$, el número de observaciones en la $i-$ésima clase. Considere la verificación de la hipótesis nula

$$H_0 : F(x) = F_0(x),$$

con $F_0(x)$ especificado completamente. Para $k \geq 2$ categorías distintas, la estadística:

$$\sum_{i=1}^{k} \frac{(N_i - np_i)^2}{np_i},$$

tienen una distribución en forma aproximada, chi-cuadrada con $k - 1$ grados de libertad, si n tiene un valor suficientemente grande. Nótese que N_i es la frecuencia observada en la i-ésima clase, y np_i es la frecuencia correspondiente que se espera bajo la hipótesis nula. Así, la estadística es la suma sobre todas las k clases de los cocientes de los cuadrados de las diferencias entre las frecuencias observadas y las esperadas. Si existe una concordancia perfecta entre las frecuencias que se observan y las que se esperan, la estadística tendrá un valor igual a cero, en caso contrario, si las frecuencias presentan una gran discrepancia, la estadística tomará un valor muy grande.

1.2.6.2. La estadística de Kolmogorov-Smirnov

La prueba de Kolmogorov-Smirnov no necesita que los datos se encuentren agrupados y es aplicable a muestras de tamaño pequeño. Ésta se basa en una comparación entre las funciones de distribución acumulada que se observa en la muestra ordenada y la distribución propuesta bajo la hipótesis nula. Si esta comparación revela una diferencia suficientemente grande entre las funciones de distribución acumulada muestral y la propuesta, entonces la hipótesis nula de que la distribución es $F_0(x)$, se rechaza. Si se denotan por $X_{(1)}, X_{(2)}, ..., X_{(n)}$ a las observaciones ordenadas de una muestra aleatoria de tamaño n y si se define la función de distribución acumulada como:

$$S_n(x) = \begin{cases} 0, & x < x_{(1)}, \\ k/n, & x_{(k)} \leq x < x_{(k+1)}, \\ 1, & x \geq x_n. \end{cases}$$

Y dado que $F_0(x)$ se encuentra completamente especificada, es posible evaluar $F_0(x)$ para algún valor deseado de x, y entonces comparar con el valor correspondiente de $S_n(x)$. Si la hipótesis nula es verdadera se espera que la diferencia sea relativamente pequeña. La estadística de Kolmogorov-Smirnov se define como:

$$D_n = \text{máx}\,|S_n(x) - F_0(x)|.$$

La estadística D_n tiene una distribución que es independiente del modelo propuesto bajo la hipótesis nula. Esto da como resultado que la función de distribución D_n pueda evaluarse sólo en función del tamaño de la muestra y después usarse para cualquier $F_0(x)$.

1.3. Estadísticos de orden

Los estadísticos de orden son una herramienta fundamental de la estadística no paramétrica y de inferencia. Algunos ejemplos de estadísticos de orden son: el mínimo y el máximo valor de una muestra, la mediana y otros cuantiles de la muestra, el estadístico de orden k es el k-ésimo elemento más pequeño de una muestra. El ordenamiento de datos es el proceso de reorganizar un conjunto de datos en una secuencia específica, como ascendente o descendente. Esto puede hacerse de manera numérica, alfabética o alfanumérica. Dada una muestra aleatoria $X_1, ..., X_n$, se puede evaluar cada una de estas variables en un punto muestral ω cualquiera y obtener una colección de números reales $X_1(\omega), ..., X_n(\omega)$. Estos números pueden ser ordenados de menor a mayor, incluyendo repeticiones. Si $X_{(i)}(\omega)$ denota el i-ésimo número ordenado, tenemos entonces la colección no decreciente de números reales $X_{(1)}(\omega) \leq \cdots \leq X_{(n)}(\omega)$.
Ahora, si el argumento ω varía, lo que se obtiene son las llamadas estadísticas de orden. Este proceso de ordenamiento resulta ser de importancia en algunas aplicaciones.

Definición 1.3.1. *Sean $X_1, ..., X_n$, una muestra aleatoria. A las variables aleatorias ordenadas*

$$\begin{aligned}
X_{(1)} &= \text{mín}\{X_1, ..., X_n\},\\
X_{(2)} &= \text{mín}\{X_1, ..., X_n\}\backslash\{X_{(1)}\},\\
X_{(3)} &= \text{mín}\{X_1, ..., X_n\}\backslash\{X_{(1)}, X_{(2)}\},\\
&\vdots\\
X_{(n)} &= \text{máx}\{X_1, ..., X_n\},
\end{aligned}$$

se les conoce con el nombre de estadísticas de orden. A $X_{(1)}$ se llama primera estadística de orden, a $X_{(2)}$ se llama segunda estadística de orden, etc. A $X_{(i)}$ se llama i-ésima estadística de orden, $i = 1, ..., n$.

Aunque los elementos de la muestra aleatoria son variables aleatorias independientes, las estadísticas de orden no lo son, pues deben mantener la relación $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$.

Además, la i-ésima estadística de orden $X_{(i)}$ no necesariamente es igual a alguna variable de la muestra aleatoria en particular, sino que, en general, es una función de todas las variables de la muestra aleatoria.
El objetivo es encontrar algunas fórmulas relacionadas con las distribuciones de probabilidad de las estadísticas de orden cuando se conoce la distribución de las variables de la muestra aleatoria (por simplicidad se supondrá absolutamente continua), suponiendo que $X_1, ..., X_n$ es una muestra aleatoria en donde cada variable tiene función de densidad $f(x)$ y una función de distribución $F(x)$. Encontrando la distribución de la primera y de la última estadística de orden de manera individual, se tiene la siguiente proposición.

Proposición 1.3.1. *Para $n \geq 1$,*

1. $f_{X_{(1)}}(x) = nf(x)[1 - F(x)]^{n-1}$.
2. $f_{X_{(n)}}(x) = nf(x)[F(x)]^{n-1}$.

Demostración
Note que:

$$\begin{aligned} F_{X_{(1)}}(x) &= P(X_{(1)} \leq x) \\ &= P(\text{mín}\{X_1, ..., X_n\} \leq x) \\ &= 1 - P(\text{mín}\{X_1, ..., X_n\} > x) \\ &= 1 - P(X_1 > x, ..., X_n > x) \\ &= 1 - [P(X_1 > x)]^n \\ &= 1 - [1 - F(x)]^n \end{aligned}$$

Derivando ambas partes, la primera y última igualdad, se obtiene 1: $f_{X_{(1)}}(x) = nf(x)[1 - F(x)]^{n-1}$.
De manera análoga, para probar 2:

$$\begin{aligned} F_{X_{(n)}}(x) &= P(X_{(n)} \leq x) \\ &= P(\text{máx}\{X_1, ..., X_n\} \leq x) \\ &= 1 - P(X_1 \leq x, ..., X_n \leq x) \\ &= 1 - [P(X_1 \leq x)]^n \\ &= [F(x)]^n \end{aligned}$$

Derivando, se obtiene el resultado $f_{X_{(n)}}(x) = nf(x)[F(x)]^{n-1}$.

De manera general, se tiene la siguiente proposición.

Proposición 1.3.2. *La función de densidad de la i-ésima estadística de orden es*

$$f_{X_{(i)}}(x) = \binom{n}{i} i f(x)[F(x)]^{i-1}[1-F(x)]^{n-i}.$$

1.4. Tipos de convergencia

Definición 1.4.1 (Convergencia en distribución)**.** *Si $\{X_n\}_{n\in\mathbb{N}}$ es una sucesión de variables aleatorias (no necesariamente definidas en el mismo espacio de probabilidad) con distribución F_n, $n \in \mathbb{N}$, se dice que la sucesión $\{X_n\}$ converge en distribución a una variable aleatoria X con distribución F si se cumple que $\lim_{n\to\infty} F_n(x) = F(x)$ para todo x que es punto de continuidad de F. Este tipo de convergencia se denota por $X_n \xrightarrow{d} X$.*

Ejemplo 1.4.1. *Si X_n es una sucesión de variables aleatorias que siguen una distribución uniforme, es decir $X_n \sim \mathcal{U}[0, 2-1/n]$, entonces $X_n \xrightarrow{d} X$, donde $X \sim \mathcal{U}[0,2]$.*
Note que la distribución de X es $F(x) = \frac{x}{2}\mathbb{1}_{[0,2)}(x) + \mathbb{1}_{[2,\infty)}(x)$. Esta función de distribución es continua en todo $\mathbb{R}$, vale 0 si $x < 0$, $\frac{x}{2}$ si $0 \le x < 2$ y 1 si $x \ge 2$. Con ello se tienen los siguientes tres casos:
Caso 1: *Si $x < 0$, entonces*

$$\lim_{n\to\infty} F_n(x) = 0,$$

pues $F_n(x) = 0$ cuando $x < 0$ para toda $n \in \mathbb{N}$.
Caso 2: *Si $0 \le x < 2$, entonces*

$$\lim_{n\to\infty} F_n(x) = \lim_{n\to\infty} \frac{x}{2-1/n} = \frac{x}{2},$$

pues $F_n(x) = \frac{x}{2-1/n}$ cuando $0 \le x < 2 - 1/n$ para toda $n \in \mathbb{N}$.
Caso 3: *Si $x \ge 2$, entonces*

$$\lim_{n\to\infty} F_n(x) = 1,$$

pues $F_n(x) = 1$ cuando $x > 2 - 1/n$ para toda $n \in \mathbb{N}$.
Por los casos anteriores, efectivamente $X_n \xrightarrow{d} X$.

Nota 1.4.1. *En adelante, si F es una función de distribución acumulada, se escribirá:*

$$\omega_F := \sup\{x \in \mathbb{R} : \ F(x) < 1\}.$$

*A ω_F se le conoce como el **extremo derecho de F** y toma valores en $\overline{\mathbb{R}} \setminus \{-\infty\}$ (no puede valer $-\infty$).*

Definición 1.4.2 (Convergencia en probabilidad)**.** *Se dice que una sucesión $\{X_n\}_{n\in\mathbb{N}}$ de*

variables aleatorias converge en probabilidad a la variable aleatoria X si,

$$\lim_{n\to\infty} P[|X_n - X| > \varepsilon] = 0,$$

para cualquier $\varepsilon > 0$. En este caso se escribirá $X_n \xrightarrow{P} X$.

Nota 1.4.2. *Si una sucesión X_n converge en probabilidad a X, entonces cualquier subsucesión de X_n también converge en probabilidad a X.*

Proposición 1.4.1. *Sean X_n y Y_n sucesiones de variables aleatorias y c una constante. Se cumplen las siguientes proposiciones,*

1. *Si $X_n \xrightarrow{P} X$ y $X_n \xrightarrow{P} Y$, entonces $P[X = Y] = 1$.*
2. *Si $X_n \xrightarrow{P} X$ entonces $cX_n \xrightarrow{P} cX$.*
3. *Si $X_n \xrightarrow{P} X$ y $Y_n \xrightarrow{P} Y$, entonces $X_n + Y_n \xrightarrow{P} X + Y$.*
4. *Si $X_n \xrightarrow{P} X$ entonces $X_n^2 \xrightarrow{P} X^2$.*

Corolario 1.4.1. *Sean X_n y Y_n sucesiones de variables aleatorias tales que $X_n \xrightarrow{P} X$ y $Y_n \xrightarrow{P} Y$, entonces $X_nY_n \xrightarrow{P} XY$.*

Ejemplo 1.4.2. *Sea $\Omega = (0, 1]$ y P la medida de Lebesgue sobre Ω, es decir, la medida de probabilidad sobre el intervalo $(0, 1]$ que asigna a cada intervalo su longitud. Para cada $n \in \mathbb{N}$ se define $X_n = I_{(0,\frac{1}{n})}$, es decir:*

$$X_n(\omega) = \begin{cases} 1, & si\ \omega < \frac{1}{n}, \\ 0, & en\ otro\ caso. \end{cases}$$

Dada $\varepsilon > 0$, se tiene:

$$P[|X_n| > \varepsilon] = P[X_n > \varepsilon] = \begin{cases} \frac{1}{n}, & si\ \varepsilon < 1, \\ 0, & en\ otro\ caso. \end{cases}$$

Así que, en cualquier caso, $\lim_{n\to\infty} P[|X_n| > \varepsilon] = 0$, Por lo tanto, $X_n \xrightarrow{P} 0$.

1.5. Intervalos de verosimilitud

La estimación puntual produce una única estimación de un parámetro θ, determinada según un criterio particular; se utiliza el MSE para medir la precisión de la estimación. Un método alternativo para estimar θ es la estimación por intervalos, que produce un intervalo de estimaciones plausibles de θ.

Definición 1.5.1. *Sea $X_1, ..., X_n$ una muestra aleatoria y sean $L(\overrightarrow{X})$ y $U(\overrightarrow{X})$ estadísticos con $L(\overrightarrow{X}) < U(\overrightarrow{X})$. Un intervalo $\left[L(\overrightarrow{X}), U(\overrightarrow{X})\right]$ utilizado para estimar θ se denomina intervalo de estimación.*

Los intervalos de estimación suelen basarse en la distribución de muestreo de un estimador puntual. Por ejemplo, un intervalo de estimación que se utiliza a menudo es $\hat{\theta} \pm 2\sqrt{MSE(\hat{\theta})}$, donde $\hat{\theta}$ es un estimador puntual de θ. Los estimadores de intervalo considerados en esta sección tienen puntos finales aleatorios $L(\overrightarrow{X})$ y $U(\overrightarrow{X})$ y utilizan la probabilidad de cobertura como medida de fiabilidad.

Definición 1.5.2. *Sea $X_1, ..., X_n$ una muestra aleatoria y sea $\left[L(\overrightarrow{X}), U(\overrightarrow{X})\right]$ un intervalo de estimación de θ. La probabilidad de cobertura asociada con el intervalo de estimación $\left[L(\overrightarrow{X}), U(\overrightarrow{X})\right]$ es $P\left(L(\overrightarrow{X}) \leq \theta \leq U(\overrightarrow{X})\right)$.*

Es útil señalar que la probabilidad de cobertura se refiere a la fiabilidad de un intervalo de estimación como procedimiento, no a la fiabilidad de una estimación de intervalo que produce, es decir, antes de que se observen los datos, un intervalo de estimación tiene puntos finales aleatorios y una probabilidad de cobertura, sin embargo, una vez que se han recogido los datos y se ha calculado un intervalo de estimación, los puntos finales del intervalo son números, no variables aleatorias. Por lo tanto, para una muestra observada $\overrightarrow{X}$, la probabilidad de que el intervalo $\left[L(\overrightarrow{X}), U(\overrightarrow{X})\right]$ contenga a θ es 0 ó 1.
Un intervalo de estimación que tiene una probabilidad de cobertura conocida y preespecificada es un intervalo de confianza.

Definición 1.5.3. *Sea $X_1, \ldots, X_n$ una muestra aleatoria y sean $L(\overrightarrow{X})$ y $U(\overrightarrow{X})$ estadísticos con $L(\overrightarrow{X}) \leq U(\overrightarrow{X})$. El intervalo $\left[L(\overrightarrow{X}), U(\overrightarrow{X})\right]$ se denomina intervalo de confianza $(1-\alpha) \times 100\,\%$ para el parámetro θ cuando*

$$P(L(\overrightarrow{X}) \leq \theta \leq U(\overrightarrow{X})) = 1 - \alpha.$$

Un intervalo de confianza de límite inferior $(1-\alpha) \times 100\,\%$ para θ es un intervalo de confianza de la forma $[L(\overrightarrow{X}), \infty)$, y un intervalo de confianza de límite superior $(1-\alpha) \times 100\,\%$ para θ es un intervalo de confianza de la forma $(-\infty, U(\overrightarrow{X})]$.
La probabilidad de cobertura preespecificada para un intervalo de confianza se denomina nivel de confianza del intervalo. El nivel de confianza se refiere a la probabilidad de que θ se encuentre entre las variables aleatorias $L(\overrightarrow{X})$ y $U(\overrightarrow{X})$ antes de que se observen los datos, y generalmente se utilizan niveles de confianza superiores al $90\,\%$.
Los límites inferior y superior de un intervalo de confianza de $(1-\alpha) \times 100\,\%$ se obtienen resolviendo $P(L(\overrightarrow{X}) < \theta) = \alpha_1$ y $P(L(\overrightarrow{X}) < \theta) = \alpha_2$ sujeto a $\alpha_1 + \alpha_2 = \alpha$. La elección de α_1 y α_2 no es única, y cuando $\alpha_1 = \alpha_2 = \frac{\alpha}{2}$ un intervalo de confianza se denomina intervalo de confianza de colas iguales. En la práctica, se suelen utilizar intervalos de confianza de colas iguales.

Ejemplo 1.5.1. *Sea $X_1, \ldots, X_n$ una muestra aleatoria y sea $\hat{\theta}$ un estimador de θ. Entonces, $\hat{\theta} \pm 2\sqrt{MSE(\hat{\theta})}$ es un intervalo de estimación de θ. La probabilidad de cobertura asociada a este intervalo de estimación varía de un modelo de probabilidad a otro. En el mejor de los casos, la desigualdad de Chebyshev puede utilizarse para establecer un límite inferior en la probabilidad de cobertura, ya que:*

$$P\left(|\hat{\theta} - \theta| \leq 2MSE(\hat{\theta})\right) \geq 1 - \frac{1}{2^2} = 0.75,$$

siempre que $MSE(\hat{\theta}) < \infty$. Así, el estimador de intervalo $\hat{\theta} \pm 2\sqrt{MSE(\hat{\theta})}$ tiene una probabilidad de cobertura mínima de 0.75.

En algunos casos, no se dispone de una cantidad fundamental ni de una distribución de muestreo conocida que pueda utilizarse para determinar un intervalo de confianza exacto para θ. Cuando n es suficientemente grande y la f.d.p. subyacente de una muestra aleatoria satisface las condiciones de regularidad dadas en la definición 1.2.5, se puede calcular un intervalo de confianza aproximado. En particular, un intervalo de confianza aproximado $(1-\alpha)\times 100\,\%$ para θ puede basarse en el MLE de θ.
Recordando que cuando se cumplen las condiciones de regularidad y $\hat{\theta}$ es el MLE de θ (Rossi, 2018), la distribución asintótica de $\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N(0, I(\theta)^{-1})$. Por lo tanto, para n suficientemente grande, $\hat{\theta}$ se distribuye aproximadamente como $N(0, AsyVar(\theta)^{-1})$, y

$$P\left(\hat{\theta} - z_{1-\frac{\alpha}{2}}\sqrt{AsyVar(\hat{\theta})} \leq \theta \leq \hat{\theta} + z_{1-\frac{\alpha}{2}}\sqrt{AsyVar(\hat{\theta})}\right) \to 1-\alpha,$$

a medida que $n \to \infty$. Por lo tanto, para muestras suficientemente grandes,

$$\left[\hat{\theta} - z_{1-\frac{\alpha}{2}}\sqrt{AsyVar(\hat{\theta})}\ ,\ \hat{\theta} + z_{1-\frac{\alpha}{2}}\sqrt{AsyVar(\hat{\theta})}\right],$$

es un intervalo de confianza aproximado $(1-\alpha)\times 100\,\%$ para θ donde $AsyVar(\hat{\theta})$ es la cota inferior de Cramér-Rao para estimadores insesgados de θ evaluados en $\theta = \hat{\theta}$. Los intervalos de confianza para muestras grandes de este tipo se denominan intervalos de confianza de Wald y se deben a Wald y Wolfowitz.

Ejemplo 1.5.2. *Sea $X_1, ..., X_n$ una muestra aleatoria de una población con distribución de Poisson de parámetro θ con $\Theta = \mathbb{R}^+$. Dado que la distribución de Poisson con $\Theta = \mathbb{R}^+$ pertenece a la familia de distribuciones exponenciales regulares, se cumplen las condiciones de regularidad. La MLE de θ es $\hat{\theta} = \bar{X}$, y puesto que la cota inferior de Cramér-Rao para estimadores insesgados de θ es $CRLB = \frac{\theta}{n}$, un intervalo de confianza aproximado $(1-\alpha)\times 100\,\%$ de gran muestra para θ basado en la MLE es*

$$\left[\hat{\theta} - z_{1-\alpha}\sqrt{\frac{\hat{\theta}}{n}}\ ,\ \hat{\theta} + z_{1-\alpha}\sqrt{\frac{\hat{\theta}}{n}}\ \right] = \left[\bar{X} - z_{1-\alpha}\sqrt{\frac{\bar{X}}{n}}\ ,\ \bar{X} + z_{1-\alpha}\sqrt{\frac{\bar{X}}{n}}\ \right].$$

Por ejemplo, cuando $n = 50$ *y* $\bar{x} = 1.34$*, un intervalo de confianza aproximado del* 95 % *para* θ *es* $[1.34 - 1.96 \times \sqrt{0.0268}, 1.34 + 1.96 \times \sqrt{0.0268}]$ *o* $[1.019, 1.661]$.

Ejemplo 1.5.3. *Sea* $X_1, \ldots, X_n$ *una muestra aleatoria de una población con distribución exponencial de parámetro* θ *con* $\Theta = \mathcal{R}^+$*. Dado que la distribución exponencial con* $\Theta = \mathcal{R}^+$ *pertenece a la familia de distribuciones exponenciales regulares, se cumplen las condiciones de regularidad. El MLE de* θ *es* $\hat{\theta} = \bar{X}$*, el límite inferior de Cramér-Rao es* $CRLB = \frac{\theta^2}{n}$ *, y la varianza asintótica estimada es* $C\hat{R}LB = \sqrt{\frac{\hat{\theta}}{n}} = \frac{\bar{X}}{\sqrt{n}}$*. Por lo tanto, un intervalo de confianza aproximado del* 96 % *de la gran muestra para* θ *es:*

$$\left[\bar{X} - 2.05\frac{\bar{X}}{\sqrt{n}}\ ,\ \bar{X} + 2.05\frac{\bar{X}}{\sqrt{n}}\ \right].$$

Para un modelo de probabilidad multiparamétrico con $\Theta \subset \mathcal{R}^d$ para $d > 1$, los intervalos de confianza aproximados de muestra grande para θ_i todavía pueden basarse en el MLE de θ_i siempre que se cumplan las condiciones de regularidad. En este caso, un intervalo de confianza aproximado $(1 - \alpha) \times 100\,\%$ para θ_i es:

$$\left[\hat{\theta}_i - z_{1-\frac{\alpha}{2}}\sqrt{AsyVar(\hat{\theta}_i)}\ ,\ \hat{\theta}_i + z_{1-\frac{\alpha}{2}}\sqrt{AsyVar(\hat{\theta}_i)}\right],$$

donde $\hat{\theta}_i$ es el MLE de θ_i y $AsyVar(\hat{\theta}_i)$ es el límite inferior de Cramér-Rao para estimadores insesgados de θ_i evaluados en $\overrightarrow{\theta} = (\hat{\theta}_1, \ldots, \hat{\theta}_d)$ que es $[I_n(\hat{\theta})^{-1}]_{ii}$.

Definición 1.5.4. *Una región de probabilidad* p % *es el conjunto* $\left\{\theta \in \Theta : \frac{L(\theta)}{L(\hat{\theta})} \geq \frac{p}{100}\right\}$ *donde* $\hat{\theta}$ *es el MLE de* θ*.*

Cuando $X_1, ..., X_n$ es una muestra aleatoria de una población X con f.d.p. que satisface las condiciones de regularidad de la definición 1.2.5 y n es suficientemente grande,

$$\begin{aligned} P\left(\theta \in \Theta : \frac{L(\theta)}{L(\hat{\theta})} \geq \frac{p}{100}\right) &= P\left(\theta \in \Theta : \ln\frac{L(\theta)}{L(\hat{\theta})} \geq \ln\frac{p}{100}\right) \\ &= P\left(\theta \in \Theta : \ell(\theta)\ell(\hat{\theta}) \geq \ln\frac{p}{100}\right) \\ &= P\left(\theta \in \Theta : -2[\ell(\theta) - \ell(\hat{\theta})] \leq -2\ln\frac{p}{100}\right). \end{aligned}$$

Para n suficientemente grande, $2[\ell(\theta) - \ell(\hat{\theta})] \sim \chi^2_1$, y por tanto:

$$P\left(\theta \in \Theta : -2(\ell(\theta) - \ell(\hat{\theta})) \leq -2\ln\frac{p}{100}\right) \approx P\left(\theta \in \Theta : \chi^2_1 \leq -2\ln\frac{p}{100}\right).$$

Por ejemplo, una región de verosimilitud del 14.7 % produce un intervalo de confianza aproximado del 95 % para θ siempre que n sea suficientemente grande y se cumplan las

condiciones de regularidad, ya que:

$$P\left(\theta \in \Theta : \frac{L(\theta)}{L(\hat{\theta})} \geq \frac{14.7}{100}\right) = P(\theta \in \Theta : -2[\ell(\theta) - \ell(\hat{\theta})] \geq 3.835)$$
$$= \underbrace{P(\chi^2_1 \leq 3.835)}_{\textit{para n grande}} = 0.9498.$$

Capítulo 2

Teoría clásica de valores extremos

Para desarrollar este capítulo se ocuparon las referencias de Coles *et al*. (2001); Haan y Ferreira (2006); y Contreras y Jiménez-Hernández (2020). En este capítulo se desarrolla el modelo que representa la piedra angular de la teoría de valores extremos.

2.1. Formulación del modelo

El modelo se centra en el comportamiento estadístico de:

$$M_n = \text{máx}\{X_1, \ldots, X_n\},$$

donde $X_1, ..., X_n$, es una sucesión de variables aleatorias independientes que tienen una función de distribución común F. En las aplicaciones, los X_i suelen representar valores enviados de un proceso medido en una escala de tiempo regular (tal vez mediciones del nivel del mar por hora, o temperaturas medias diarias), de modo que M_n representa el máximo del proceso en n unidades de tiempo de observación. Si n es el número de observaciones en un año, entonces M_n corresponde al número anual máximo.
En teoría, la distribución de M_n se puede obtener exactamente para todos los valores de n, teniendo en cuenta las propiedades de independencia,

$$\begin{aligned} F_{M_n}(z) = P\{M_n \leq z\} &= P\{\text{máx}\{X_1, X_2, \ldots, X_n\} \leq z\} \\ &= P\{X_1 \leq z, X_2 \leq z, \ldots, X_n \leq z\} \\ &= P\{X_1 \leq z\} \times P\{X_2 \leq z\} \times \cdots \times P\{X_n \leq z\} \\ &= \{F(z)\}^n \\ &= F^n(z). \end{aligned} \tag{2.1}$$

Sin embargo, esto no resulta inmediatamente útil en la práctica, ya que se desconoce la función de distribución F. Una posibilidad es utilizar técnicas de estadísticas para estimar F a

partir de datos observados y luego sustituir esta estimación en la ecuación (2.1). Desafortunadamente, discrepancias muy pequeñas en la estimación de F pueden dar lugar a discrepancias sustanciales para F^n.
Un enfoque alternativo es aceptar que F es desconocido y buscar familias aproximadas de modelos para F^n, que puedan estimarse basándose únicamente en los datos extremos. Esto es similar a la práctica habitual de aproximar la distribución de las medias muestrales mediante la distribución normal, como lo justifica el teorema del límite central. Los argumentos de este capítulo son esencialmente un análogo de valores extremos de la teoría del teorema del límite central.
Se procede observando el comportamiento de F^n cuando $n \longrightarrow \infty$. Pero esto por sí solo no es suficiente: para cualquier $z < z_+$, donde z_+ es el punto final superior de F, $F^n(z) \longrightarrow 0$ cuando $n \longrightarrow \infty$, de modo que la distribución de M_n degenera a una masa puntual en z_+. Esta dificultad se evita permitiendo una renormalización lineal de la variable M_n,

$$M_n^* = \frac{M_n - b_n}{a_n},$$

para sucesiones de constantes $\{a_n > 0\}$ y $\{b_n\}$. Elecciones apropiadas de $\{a_n\}$ y $\{b_n\}$ estabilizan la ubicación y escala de M_n^* a medida que n aumenta, evitando las dificultades que surgen con la variable M_n. Por lo tanto, se buscan distribuciones límite para M_n^*, con opciones apropiadas de $\{a_n\}$ y $\{b_n\}$, en lugar de M_n.

2.2. Teorema de los tipos de extremos

El rango completo de posibles distribuciones límite para M_n^* viene dado por el teorema de tipos extremos 2.2.1. Este teorema fue propuesto en 1928 por Fisher y Tippett y probado rigurosamente por Gnedenko en 1943 (Gnedenko, 1943).

Teorema 2.2.1. *Si existen sucesiones de constantes $\{a_n\}_n$ y $\{b_n\}_n$, con $a_n > 0$ tales que*

$$P\{(M_n - b_n)/a_n \leq z\} \longrightarrow G(z),$$

cuando $n \longrightarrow \infty$, donde G es una función de distribución no degenerada y z un punto de continuidad de G, entonces G pertenece a una de las siguientes familias:

$$I : G(z) = \exp\{-\exp[-(\tfrac{z-b}{a})]\}, \qquad -\infty < z < \infty,$$

$$II : G(z) = \begin{cases} \exp\{-(\tfrac{z-b}{a})^{-\alpha}\}, & z > b, \\ 0, & z \leq b, \end{cases}$$

$$III : G(z) = \begin{cases} \exp\{-(-\tfrac{z-b}{a})^{\alpha}\}, & z < b, \\ 1, & z \geq b, \end{cases}$$

para parámetros $a > 0, b \in R$ y en el caso de las familias II y III, $\alpha > 0$.

El teorema anterior establece que los máximos muestrales reescalados $(M_n - b_n)/a_n$ convergen en distribución a una variable que tiene una distribución dentro de una de las familias denominadas I, II y III. En conjunto, estas tres clases de distribución se denominan distribuciones de valores extremos, con tipos I, II y III ampliamente conocidas como las familias Gumbel, Fréchet y Weibull, respectivamente. Cada familia tiene un parámetro de ubicación y escala, b y a respectivamente. Además, las familias Fréchet y Weibull tienen un parámetro de forma α.
El Teorema 2.2.1 implica que, cuando M_n puede estabilizarse con las sucesiones adecuadas $\{a_n\}$ y $\{b_n\}$, la variable normalizada correspondiente M_n^* tiene una distribución límite que corresponde a una de los tres tipos de distribuciones de valores extremos. La característica notable de este resultado es que los tres tipos de distribuciones de valores extremos son los únicos límites posibles para las distribuciones de M_n^*, independientemente de la distribución F para la población. Es en este sentido que el teorema proporciona un resultado análogo al teorema del límite central.

2.3. La distribución de valores extremos generalizada

Los tres tipos de límites que surgen en el Teorema 2.2.1 tienen distintas formas de comportamiento, correspondientes a las diferentes formas de comportamiento de la cola para la función de distribución F de X_i. Esto se puede precisar considerando el comportamiento de la distribución límite G en z_+, su punto final superior. Para la distribución de Weibull, z_+ es finito, mientras que para las distribuciones de Fréchet y Gumbel, $z_+ = \infty$. Sin embargo, la densidad de G decae exponencialmente para la distribución de Gumbel y polinomialmente para la distribución de Fréchet, lo que corresponde a tasas de decaimiento relativamente diferentes en la cola de F. De ello se deduce que en las aplicaciones las tres familias diferentes dan representaciones bastante diferentes de comportamiento de valor extremo. En las primeras aplicaciones de la teoría de los valores extremos, era habitual adoptar una de las tres familias y luego estimar los parámetros relevantes de esa distribución. Pero hay dos debilidades: primero, se requiere una técnica para elegir cuál de las tres familias es la más apropiada para los datos disponibles; en segundo lugar, una vez que se toma tal decisión, las inferencias posteriores suponen que esta elección es correcta y no tienen en cuenta la incertidumbre que dicha selección implica, aunque esta incertidumbre pueda ser sustancial.
Un mejor análisis lo ofrece una re-formulación de los modelos en el Teorema 2.2.1. Es sencillo comprobar que las familias Gumbel, Fréchet y Weibull se pueden combinar en una única familia de modelos que tienen funciones de distribución de la forma:

$$G(z) = \exp\left\{-\left[1 + \xi\left(\frac{z-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}}\right\}, \tag{2.2}$$

definido en el conjunto $\{z : 1 + \xi(z-\mu)/\sigma > 0\}$, donde los parámetros satisfacen $-\infty < \mu < \infty$, $\sigma > 0$ y $-\infty < \xi < \infty$. Esta es la familia de distribuciones de valores extremos generalizados (DGVE). El modelo tiene tres parámetros: un parámetro de ubicación, μ; un

parámetro de escala, σ; y un parámetro de forma, ξ. Las clases de distribución de valores extremos tipo II y tipo III corresponden respectivamente a los casos $\xi > 0$ y $\xi < 0$ en esta parametrización. El subconjunto de la familia DGVE con $\xi = 0$ se interpreta como el límite de (2.2) cuando $\xi \longrightarrow 0$, lo que lleva a la familia Gumbel con función de distribución:

$$G(z) = \exp\left[-\exp\left\{-\left(\frac{z-\mu}{\sigma}\right)\right\}\right], \qquad -\infty < z < \infty.$$

La unificación de las tres familias originales de distribución de valores extremos en una sola familia simplifica enormemente la implementación estadística. A través de la primera inferencia, los datos mismos determinan el tipo más apropiado de comportamiento de cola, y no hay necesidad de hacer juicios subjetivos a priori sobre qué familia de valores extremos adoptar. Además, la incertidumbre en el valor inferido de ξ mide la falta de certeza sobre cuál de los tres tipos originales es el más apropiado para un conjunto de datos determinado. Por conveniencia, se reformula el Teorema 2.2.1 en forma modificada.

Teorema 2.3.1. *Si existen sucesiones de constantes* $\{a_n\}_n$ *y* $\{b_n\}_n$*, con* $a_n > 0$ *tales que,*

$$P\{(M_n - b_n)/a_n \leq z\} \longrightarrow G(z), \tag{2.3}$$

cuando $n \longrightarrow \infty$*, donde G es una función de distribución no degenerada, entonces G pertenece a la familia DGVE:*

$$G(z) = \exp\left\{-\left[1 + \xi\left(\frac{z-\mu}{\sigma}\right)\right]^{-1/\xi}\right\},$$

definida en el conjunto $\{z : 1 + \xi(z-\mu)/\sigma > 0\}$*, donde los parámetros satisfacen* $-\infty < \mu < \infty$*,* $\sigma > 0$ *y* $-\infty < \xi < \infty$*.*

Interpretar el límite del Teorema 2.3.1 como una aproximación para valores grandes de n sugiere el uso de la familia DGVE para modelar la distribución de máximos de sucesiones largas. La dificultad aparente de que en la práctica se desconozcan las constantes de normalización se resuelve fácilmente, suponiendo (2.3),

$$P\{(M_n - b_n)/a_n \leq z\} \approx G(z),$$

para n suficientemente grande, de manera equivalente,

$$P\{M_n \leq z\} \approx G\{(z - b_n)/a_n\} = G^*(z),$$

donde G^* es otro miembro de la familia DGVE. En otras palabras, el Teorema 2.3.1 permite la aproximación de la distribución de M_n^* por un miembro de la familia DGVE para n grande, la distribución del propio M_n también puede aproximarse mediante un miembro diferente de la misma familia. Dado que los parámetros de la distribución deben estimarse de todos modos, en la práctica es irrelevante que los parámetros de la distribución G sean diferentes de los de G^*.

Este argumento conduce al siguiente enfoque para modelar los extremos de una serie de observaciones independientes $X_1, X_2, \ldots$. Los datos se agrupan en bloques en sucesiones de

observaciones de longitud n, para algún valor grande de n, generando una serie de máximos de bloque, $M_{n,1}, ..., M_{n,m}$, por decir, al que se le puede adaptar la DGVE. A menudo, los bloques se eligen para que correspondan a un periodo de tiempo de un año de duración, en cuyo caso n es el número de observaciones en un año y los máximos del bloque son máximos anuales. Luego se obtienen estimaciones de los cuantiles extremos de la distribución máxima anual invirtiendo la ecuación:

$$z_p = \begin{cases} \mu - \frac{\sigma}{\xi}[1 - \{-\log(1-p)\}^{-\xi}], & \text{para} \quad \xi \neq 0, \\ \mu - \sigma \log\{-\log(1-p)\}, & \text{para} \quad \xi = 0, \end{cases} \tag{2.4}$$

donde $G(z_p) = 1 - p$. En la terminología común, z_p es el nivel de retorno asociado con el período de retorno $1/p$, ya que con un grado razonable de precisión, se espera que el nivel z_p se supere una vez cada $1/p$ años. Más concretamente, z_p es superado por el máximo anual en un año concreto con una probabilidad p.
Dado que los cuantiles permiten expresar los modelos de probabilidad a escala de los datos, la relación del modelo DGVE con sus parámetros se interpreta más fácilmente en términos de cuantiles. En particular, definiendo $y_p = -\log(1-p)$, de modo que:

$$z_p = \begin{cases} \mu - \frac{\sigma}{\xi}\left[1 - y_p^{-\xi}\right], & \text{para} \quad \xi \neq 0, \\ \mu - \sigma \log y_p, & \text{para} \quad \xi = 0, \end{cases}$$

se deduce que, si z_p se representa gráficamente frente a y_p en una escala logarítmica (o equivalentemente, si z_p se representa gráficamente frente a $\log y_p$), el gráfico es lineal en el caso $\xi = 0$. Si $\xi < 0$ el gráfico es convexo con límite asintótico cuando $p \to 0$ en $\mu - \sigma/\xi$; si $\xi > 0$ el gráfico es cóncavo y no tiene límite finito. A este tipo de gráfico se le conoce como gráfico de nivel de retorno. Debido a la sencillez de su interpretación y a que la elección de la escala comprime la cola de la distribución de forma que se pone de relieve el efecto de la extrapolación, los gráficos de nivel de retorno son especialmente convenientes tanto para la presentación como para la validación de modelos.

2.4. Modelo de máximos por bloques

Es uno de los principales métodos en la teoría de valores extremos para obtener una distribución adecuada para estimar la probabilidad de valores grandes, el tamaño de bloque suele ser seleccionado para reflejar la posible periodicidad intrínseca del fenómeno estudiado. Por ejemplo, semanal, mensual, anual, dependiendo del estudio.

El procedimiento consiste en agrupar los datos en bloques de igual tamaño o longitud y posteriormente ajustar la DGVE al conjunto de los máximos correspondientes a cada uno de los bloques. El principal problema que presenta este método reside en la elección del tamaño de los bloques, ya que la elección de bloques muy pequeños conducirá a una pobre aproximación del modelo; sin embargo, para cuestiones prácticas, en sucesiones de datos

temporales mensuales, se suelen tomar bloques de longitud anual, de esta manera los máximos se distribuyen de manera similar en cada uno de los bloques.

2.4.1. Descripción del modelo

Supóngase que se tiene una muestra aleatoria de tamaño n de una población X con cierta distribución F que pertenece a alguno de los tres dominios de atracción. Dicha muestra estará denotada por $X_1, X_2, ..., X_n$. Se debe tomar un valor k tal que la muestra se pueda dividir de la siguiente manera:

$$\begin{aligned} B[1] &= \{X_1, ..., X_k\} \\ B[2] &= \{X_{k+1}, ..., X_{2k}\} \\ &\vdots \\ B[m] &= \{X_{(m-1)k+1}, ..., X_{m_0}\} \end{aligned}$$

para algún $m \in \mathbf{N}$ y $m_0 \leq mk$. Entonces cada $B[j]$ es un **bloque** de la muestra original.

Una vez hecha esta división por bloques, la muestra aleatoria de interés sobre la cual se hará la inferencia es $M_1, M_2, M_3, ..., M_m$, donde $M_j = \text{máx}\, B[j]$ para cada $j = 1, ..., m$, y además se supone que cada M_j es una v.a. i.i.d., sin embargo, las componentes de cada vector $B[j]$ pueden ser dependientes.

2.5. Ejemplos del teorema de la distribución de valores extremos generalizada

Los ejemplos a continuación ilustran cómo la elección cuidadosa de sucesiones normalizadoras conduce a una distribución límite dentro de la familia DGVE, como implica el Teorema 2.3.1.

Ejemplo 2.5.1. *Si $X_1, X_2, ...$ es una sucesión de v.a. i.i.d. con función de distribución exponencial y parámetro $\lambda = 1$, entonces $F(x) = 1 - e^{-x}$ para $x > 0$. En este caso, dejando $a_n = 1$ y $b_n = \ln n$,*

$$\begin{aligned} P\{(M_n - b_n)/a_n \leq z\} &= F^n(z + \ln n) \\ &= \left[1 - e^{(z+\ln n)}\right]^n \\ &= \left[1 - n^{-1}e^{-z}\right]^n \\ &\to \exp(-e^{-z}), \end{aligned}$$

cuando $n \to \infty$, para cada $z \in \mathbb{R}$ fijo. Por lo tanto, con a_n y b_n elegidos, la distribución límite de M_n cuando $n \to \infty$ es la distribución Gumbel, correspondiente a $\xi = 0$ en la familia DGVE.

Ejemplo 2.5.2. *Si $X_1, X_2, ...$ es una sucesión de v.a i.i.d. de Fréchet estándar. $F(x) = \exp(-1/x)$ para $x > 0$, y si $a_n = n$ y $b_n = 0$,*

$$\begin{aligned} P\{(M_n - b_n)/a_n \leq z\} &= F^n(nz) \\ &= [\exp\{-1/(nz)\}]^n \\ &= \exp(-1/z), \end{aligned}$$

cuando $n \to \infty$, para cada $z > 0$. Por lo tanto, en este caso (que es un resultado exacto para todos los n, debido a la máxima estabilidad de F) es también la distribución estándar de Fréchet, con $\xi = 1$ en la familia DGVE.

Ejemplo 2.5.3. *Si $X_1, X_2, ...$ son una sucesión de v.a i.i.d. uniformes $U(0, 1)$, $F(x) = x$ para $0 \leq x \leq 1$. Para z fijo con $z < 0$, supóngase que $n > -z$ y sean $a_n = 1/n$ y $b_n = 1$, entonces,*

$$\begin{aligned} P\{(M_n - b_n)/a_n \leq z\} &= F^n(n^{-1}z + 1) \\ &= \left(1 + \frac{z}{n}\right)^n \\ \to &= e^z, \end{aligned}$$

cuando $n \to \infty$. Por lo tanto, el límite de la distribución es de tipo Weibull, con $\xi = -1$ en la familia DGVE.

En estos ejemplos, la elección de $\{a_n\}$ y b_n tiene cierta libertad. Sin embargo, las diferentes opciones que conducen a un límite no degenerado siempre dan como resultado una distribución límite en la familia DGVE con el mismo valor de ξ, aunque posiblemente con otros valores de los parámetros de escala y localización.

Capítulo 3

Inferencia para la distribución generalizada de valores extremos

En este Capítulo se emplea la inferencia estadística para estimar los parámetros de la DGVE haciendo uso del método de máxima verosimilitud. La bibliografía para este Capítulo es Coles *et al*. (2001); Contreras y Jiménez-Hernández (2020); y Haan y Ferreira (2006).

3.1. Consideraciones generales

La distribución GVE proporciona un modelo para la distribución de los máximos de bloque. Su aplicación consiste en dividir los datos en bloques de igual longitud y ajustar la distribución generalizada de valores extremos al conjunto de máximos de bloque. Sin embargo, a la hora de aplicar este modelo a un conjunto de datos concreto, la elección del tamaño de los bloques puede ser fundamental. La elección equivale a un compromiso entre sesgo y varianza: si los bloques son demasiado pequeños, es probable que la aproximación por el modelo límite del Teorema 2.3.1 sea deficiente, lo que provocará sesgos en la estimación y la extrapolación; si los bloques son grandes, se generan pocos bloques máximos, lo que provoca una gran varianza en la estimación.
Las consideraciones pragmáticas llevan a menudo a adoptar bloques de un año de duración. Por ejemplo, es posible que sólo se hayan registrado los datos máximos anuales, por lo que el uso de bloques más cortos no es una opción. Incluso cuando éste no sea el caso, es probable que un análisis de los datos máximos anuales sea más sólido que un análisis basado en bloques más cortos si se incumplen las condiciones del Teorema 2.3.1. Por ejemplo, es probable que las temperaturas diarias varíen según la estación, violando la suposición de que las X_i; tienen una distribución común. Si los datos se dividen en bloques de unos 3 meses, es probable que el máximo del bloque de verano sea mucho mayor que el del bloque de invierno, y una inferencia que no tuviera en cuenta esta falta de homogeneidad daría resultados

inexactos. En cambio, si se toman bloques de un año de duración, la hipótesis de que los máximos de cada bloque tienen una distribución común es plausible, aunque la justificación formal de la aproximación GEV sigue siendo inválida.
Ahora simplificando la notación denotando los máximos de bloque $Z_1, ..., Z_m$. Se supone que son variables independientes de una distribución GEV cuyos parámetros deben estimarse. Si las X_i son independientes, entonces las Z_i también son independientes. Sin embargo, es probable que la independencia de las Z_i sea una aproximación razonable incluso si las X_i constituyen una serie dependiente. En este caso, aunque no esté cubierto por el Teorema 2.3.1, la conclusión de que las Z_i tiene una distribución GEV puede seguir siendo razonable (Coles *et al.*, 2001).
Se han propuesto muchas técnicas para la estimación de parámetros en modelos de valores extremos. Entre ellas se incluyen técnicas gráficas basadas en versiones de diagramas de probabilidad; técnicas basadas en momentos en las que las funciones de los momentos del modelo se equiparan con sus equivalentes empíricos; procedimientos en los que los parámetros se estiman como funciones especificadas de estadísticas de orden; y métodos basados en la verosimilitud. Cada técnica tiene sus pros y sus contras, pero la utilidad general y la adaptabilidad a la construcción de modelos complejos de las técnicas basadas en la verosimilitud hacen que este enfoque resulte especialmente atractivo.

Una dificultad más que se presenta es el uso de métodos de verosimilitud para el ajuste de la DGVE la cual se refiere a las condiciones de regularidad que se requieren para que las propiedades asintóticas usuales asociadas con el estimador de máxima verosimilitud sean válidas. El modelo de la DGVE no satisface dichas condiciones porque los puntos finales de la distribución GVE son funciones de los valores de los parámetros: $\mu - \sigma/\xi$ es un punto final superior de la distribución cuando $\xi < 0$, y un punto final inferior cuando $\xi > 0$. Esta violación de las condiciones de regularidad usuales significa que los resultados de verosimilitud asintótica estándar no son automáticamente aplicables. Smith (1985) estudió este problema en detalle y obtuvo los siguientes resultados:

- Cuando $\xi > -0.5$, los estimadores de máxima verosimilitud son regulares, en el sentido de tener las propiedades asintóticas habituales;
- Cuando $-1 < \xi < -0.5$, generalmente se pueden obtener estimadores de máxima verosimilitud, pero no tienen las propiedades asintóticas estándar;
- Cuando $\xi < -1$, es poco probable que se puedan obtener estimadores de máxima verosimilitud.

El caso $\xi \leq -0.5$ corresponde a distribuciones con una cola superior acotada muy corta. Esta situación rara vez se encuentra en aplicaciones de modelado de valores extremos, por lo que las limitaciones teóricas del enfoque de máxima verosimilitud no suelen ser un obstáculo en la práctica.

3.2. Estimación de los parámetros

Suponiendo que $Z_1, ..., Z_m$ son variables aleatorias i.i.d. que tienen la función de densidad $g(\theta; z)$ de la DGVE $G(\theta; z)$, la estimación de los parámetros por máxima verosimilitud cuando $\xi \neq 0$ está dada por,

$$\begin{aligned} L(\theta) =& g_{Z_1,...,Z_m}(\theta; z_1, ..., z_m) = g_{Z_1}(\theta; z_1), ..., g_{Z_m}(\theta; z_m) \\ =& \Big(\frac{1}{\sigma}\Big)^m \prod_{i=1}^{m} \Big[1 + \xi\Big(\frac{z_i - \mu}{\sigma}\Big)\Big]^{-(1+\frac{1}{\xi})} \prod_{i=1}^{m} \exp\Big\{ - \Big[1 + \xi\Big(\frac{z_i - \mu}{\sigma}\Big)\Big]^{-\frac{1}{\xi}}\Big\} \\ =& \Big(\frac{1}{\sigma}\Big)^m \prod_{i=1}^{m} \Big[1 + \xi\Big(\frac{z_i - \mu}{\sigma}\Big)\Big]^{-(1+\frac{1}{\xi})} \exp\Big\{ - \sum_{i=1}^{m} \Big[1 + \xi\Big(\frac{z_i - \mu}{\sigma}\Big)\Big]^{-\frac{1}{\xi}}\Big\}. \end{aligned}$$

Así, la función de log-verosimilitud es,

$$\ell(\theta) = n[\log(1) - \log(\sigma)] + \sum_{i=1}^{m} \Big\{ \log\Big(\Big[1 + \xi\Big(\frac{z_i - \mu}{\sigma}\Big)\Big]^{-(1+\frac{1}{\xi})}\Big)\Big\} - \sum_{i=1}^{m} \Big[1 + \xi\Big(\frac{z_i - \mu}{\sigma}\Big)\Big]^{\frac{1}{\xi}},$$

o bien,

$$\ell(\mu, \sigma, \xi) = -m \log \sigma - \Big(1 + \frac{1}{\xi}\Big) \sum_{i=1}^{m} \log\Big[1 + \xi\Big(\frac{z_i - \mu}{\sigma}\Big)\Big] - \sum_{i=1}^{m} \Big[1 + \xi\Big(\frac{z_i - \mu}{\sigma}\Big)\Big]^{-1/\xi}, \qquad (3.1)$$

siempre que $1 + \xi\left(\frac{z_i - \mu}{\sigma}\right) > 0$, para $i = 1, ..., m$.

Nota 3.2.1. *Cuando uno de los datos observados cae más allá de un punto final de la distribución, la probabilidad es cero y la log-verosimilitud es igual a* $-\infty$.

En el caso $\xi = 0$, la función de máxima verosimilitud está dada por,

$$\begin{aligned} L(\theta) =& g_{Z_1,...,Z_m}(\theta; z_1, ..., z_m) = g_{Z_1}(\theta; z_1), ..., g_{Z_m}(\theta; z_m) \\ =& \Big(\frac{1}{\sigma}\Big)^m \prod_{i=1}^{m} \exp\Big[- \exp\Big\{\frac{\mu - z_i}{\sigma}\Big\} + \frac{\mu - z_i}{\sigma}\Big] \\ =& \Big(\frac{1}{\sigma}\Big)^m \exp\Big\{ \sum_{i=1}^{m} \Big[- \exp\Big\{\frac{\mu - z_i}{\sigma}\Big\}\Big] + \sum_{i=1}^{m} \Big[\frac{\mu - z_i}{\sigma}\Big]\Big\}, \end{aligned}$$

así, la función log-verosimilitud es,

$$\ell(\theta) = -m \log(\sigma) - \sum_{i=1}^{m} \Big[\exp\Big\{\frac{\mu - z_i}{\sigma}\Big\}\Big] + \sum_{i=1}^{m} \Big[\frac{\mu - z_i}{\sigma}\Big],$$

o bien,

$$\ell(\mu,\sigma) = -m\log\sigma - \sum_{i=1}^{m}\exp\left\{-\frac{z_i-\mu}{\sigma}\right\} - \sum_{i=1}^{m}\left(\frac{z_i-\mu}{\sigma}\right). \tag{3.2}$$

Las derivadas parciales con respecto a μ, σ y ξ de la función log-verosimilitud dada en la ecuación (3.1) son las siguientes.

$$\begin{aligned}\frac{\partial\ell(\theta)}{\partial\mu} &= -\left(1+\frac{1}{\xi}\right)\sum_{i=1}^{m}\left[\frac{-\frac{\xi}{\sigma}}{[1+\xi(\frac{z_i-\mu}{\sigma})]}\right] - \sum_{i=1}^{m}\left(-\frac{1}{\xi}\right)\left[1+\xi\left(\frac{z_i-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}-1}\left(-\frac{\xi}{\sigma}\right),\\ &= \frac{\xi(1-\frac{1}{\xi})}{\sigma}\sum_{i=1}^{m}\left[1+\xi\left(\frac{z_i-\mu}{\sigma}\right)\right]^{-1} - \frac{1}{\sigma}\sum_{i=1}^{m}\left[1+\xi\left(\frac{z_i-\mu}{\sigma}\right)\right]^{-(1+\frac{1}{\xi})}.\end{aligned}$$

$$\begin{aligned}\frac{\partial\ell(\theta)}{\partial\sigma} &= -\frac{m}{\sigma} - \left(1+\frac{1}{\xi}\right)\sum_{i=1}^{m}\left[\frac{-\frac{\xi(z_i-\mu)}{\sigma^2}}{[1+\xi(\frac{z_i-\mu}{\sigma})]}\right] - \sum_{i=1}^{m}\left(-\frac{1}{\xi}\right)\left[1+\xi\left(\frac{z_i-\mu}{\sigma}\right)\right]^{-\frac{1}{c}-1}\left(-\frac{\xi(z_i-\mu)}{\sigma^2}\right),\\ &= -\frac{m}{\sigma} + \frac{\xi(1+\frac{1}{\xi})}{\sigma^2}\sum_{i=1}^{m}\frac{z_1-\mu}{[1+\xi(\frac{z_i-\mu}{\sigma})]} - \frac{1}{\sigma^2}\sum_{i=1}^{m}(z_i-\mu)\left[1+\xi\left(\frac{z_i-\mu}{\sigma}\right)\right]^{-(1+\frac{1}{\xi})}.\end{aligned}$$

$$\begin{aligned}\frac{\partial\ell(\theta)}{\partial\xi} &= -\left\{\left(-\frac{1}{\xi^2}\right)\sum_{i=1}^{m}\log\left[1+\xi\left(\frac{z_i-\mu}{\sigma}\right)\right] + \sum_{i=1}^{m}\frac{\frac{z_i-\mu}{\sigma}}{\left[1+\xi(\frac{z_i-\mu}{\sigma})\right]}\left(1+\frac{1}{\xi}\right)\right\} - \frac{\partial}{\partial\xi}\sum_{i=1}^{m}\left[1+\xi\left(\frac{z_i-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}},\\ &= \frac{1}{\xi^2}\sum_{i=1}^{m}\log\left[1+\xi\left(\frac{z_i-\mu}{\sigma}\right)\right] - \frac{(1+\frac{1}{\xi})}{\sigma}\sum_{i=1}^{m}\frac{z_i-\mu}{\left[1+\xi(\frac{z_i-\mu}{\sigma})\right]} + \frac{1}{\xi}\sum_{i=1}^{m}\left(\frac{z_i-\mu}{\sigma}\right)\left[1+\xi\left(\frac{z_i-\mu}{\sigma}\right)\right]^{-(1+\frac{1}{\xi})}\\ &\quad - \frac{1}{\xi^2}\sum_{i=1}^{m}\left[1+\xi\left(\frac{z_i-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}}\log\left[1+\xi\left(\frac{z_i-\mu}{\sigma}\right)\right].\end{aligned}$$

Por otro lado, las derivadas parciales de μ, σ y ξ de la función log-verosimilitud dada por la ecuación (3.2), están dadas por las siguientes expresiones algebraicas.

$$\begin{aligned}\frac{\partial\ell(\theta)}{\partial\mu} &= -\frac{1}{\sigma}\sum_{i=1}^{m}\left[\exp\left(-\frac{z_i-\mu}{\sigma}\right)\right] + \frac{m}{\sigma}.\\ \frac{\partial\ell(\theta)}{\partial\sigma} &= -\frac{m}{\sigma} - \left[\sum_{i=1}^{m}\left(\frac{z_i-\mu}{\sigma^2}\right)\left(\exp\left\{-\frac{z_i-\mu}{\sigma}\right\}-1\right)\right].\end{aligned}$$

Igualando a 0 las derivadas parciales $\partial\ell(\theta)/\partial\mu$ y $\partial\ell(\theta)/\partial\sigma$ para el caso Gumbel, se obtienen las

siguientes ecuaciones respectivamente.

$$m - \sum_{i=1}^{m} \exp\Big(- \frac{z_i - \mu}{\sigma} \Big) = 0,$$

$$m + \sum_{i=1}^{m} \Big(\frac{z_i - \mu}{\sigma} \Big) \Big(\exp\Big\{ - \frac{z_i - \mu}{\sigma} \Big\} - 1 \Big) = 0.$$

Para maximizar las ecuaciones (3.1) y (3.2), con respecto al vector de parámetros (μ, σ, ξ), se utilizará la estimación por máxima verosimilitud sobre toda la familia de DGVE. No hay una solución analítica, pero para cualquier conjunto de datos dado, la maximización del vector es sencilla utilizando algoritmos de optimización numérica estándar. Es necesario tener cuidado para garantizar que dichos algoritmos no pasen a combinaciones de parámetros que contradigan la ecuación (3.1). Para evitar que no se cumpla la condición de la ecuación (3.1), es decir, para valores de ξ que son muy cercanos a cero, se utilizará la ecuación (3.2) en lugar de la (3.1).
Además, la distribución aproximada de $(\hat{\mu}, \hat{\sigma}, \hat{\xi})$ es normal multivariada con media (μ, σ, ξ) y matriz de covarianza igual a la inversa de la matriz de información observada evaluada en la estimación por máxima verosimilitud. Aunque la matriz se puede calcular analíticamente, es más fácil utilizar técnicas de diferenciación numérica para evaluar las segundas derivadas. Los intervalos de confianza se deducen inmediatamente de la normalidad aproximada del estimador.

A continuación se presentan las funciones de verosimilitud y de log-verosimilitud para los parámetros de las distribuciones de valores extremos.
Sea $\theta = (\mu, \sigma, \xi)$ el vector de parámetros a estimar. En el caso de la distribución Gumbel, el vector de parámetros a estimar es $\theta = (\mu, \sigma)$. Las funciones de verosimilitud para las distribuciones de valores extremos son:

$$\textit{Gumbel}: \qquad L(\theta) = \Big(\frac{1}{\sigma}\Big)^n \exp\Big\{ \sum_{i=1}^{n} \Big[- \exp\Big(\frac{\mu - z_i}{\sigma}\Big)\Big] + \sum_{i=1}^{n} \Big[\frac{\mu - z_i}{\sigma}\Big] \Big\}.$$

$$\textit{Fréchet}: \qquad L(\theta) = \Big(\frac{\xi}{\sigma}\Big)^n \prod_{i=1}^{n} \Big(\frac{z_i - \mu}{\sigma}\Big)^{-(1+\xi)} \exp\Big\{ \sum_{i=1}^{n} \Big[- \Big(\frac{z_i - \mu}{\sigma}\Big)^{-\xi}\Big] \Big\}.$$

$$\textit{Weibull}: \qquad L(\theta) = \Big(\frac{\xi}{\sigma}\Big)^n \prod_{i=1}^{n} \Big(- \Big(\frac{z_i - \mu}{\sigma}\Big)\Big)^{\xi - 1} \exp\Big\{ \sum_{i=1}^{n} \Big[- \Big(- \Big(\frac{z_i - \mu}{\sigma}\Big)\Big)^{\xi}\Big] \Big\}.$$

Y las funciones de log-verosimilitud están dadas por:

$$Gumbel: \quad \ell(\theta) = -n\log(\sigma) - \sum_{i=1}^{n}\left[\exp\left(\frac{\mu - z_i}{\sigma}\right)\right] + \sum_{i=1}^{n}\left[\frac{\mu - z_i}{\sigma}\right].$$

$$Fréchet: \quad \ell(\theta) = n[\log(\xi) - \log(\sigma)] - (1+\xi)\sum_{i=1}^{n}\left[\log(z_i - \mu) - \log(\sigma)\right] - \sum_{i=1}^{n}\left(\frac{z_i - \mu}{\sigma}\right)^{-\xi}.$$

$$Weibull: \quad \ell(\theta) = n[\log(\xi) - \log(\sigma)] - (1+\xi)\sum_{i=1}^{n}\log\left(\frac{\mu - z_i}{\sigma}\right) - \sum_{i=1}^{n}\left(-\left(\frac{z_i - \mu}{\sigma}\right)\right)^{\xi}.$$

Se requiere de métodos numéricos para hallar los estimadores de máxima verosimilitud.

3.3. Niveles de retorno

Sea $\{X_i\}_{i\geq 1}$ una sucesión de v.a.i.i.d., tomadas en periodos de tiempo iguales; con función de distribución $F(z)$ conocida y ω un valor dado. Considérese también la sucesión $\{1_{\{X_i>\omega\}}\}_{i\geq 1}$ de v.a.i.i.d. con distribución Bernoulli que toma el valor de 1 (éxito) si $X_i > \omega$ y 0 en otro caso, con probabilidad $p = 1 - F(\omega)$ y $1 - p$. respectivamente. Entonces, el instante del primer éxito está dado por

$$Y(\omega) = \text{mín}\{i \geq 1 | X_i > \omega\},$$

es decir, el instante de la primera excedencia del valor ω, es una variable aleatoria con distribución geométrica y función de probabilidad,

$$P[Y(\omega) = k] = (1-p)^{k-1}p, \quad k = 1, 2, 3, ...$$

Por lo tanto,

$$T = \frac{1}{p},$$

donde $T = E[Y(\omega)]$.
Al valor ω se le llama nivel de retorno con periodo de retorno $1/p$ para los eventos $\{X_i > \omega\}_{i\geq 1}$. Así, si se desea encontrar el nivel de retorno ω correspondiente a un periodo de 20 años, entonces se debe encontrar el valor de ω tal que $E[Y(\omega)] = 20$, es decir, encontrar ω tal que $p = 1/20 = 0.05$. Ahora, como $p = 1 - F(\omega)$, el problema consiste en resolver la ecuación $F(\omega) = 1 - 0.05 = 0.95$ para ω. Por lo tanto, ω es el 0.95-cuantil de la distribución $F(Z)$, esto es, $Q_{0.95}$. Por este motivo el nivel de retorno es igual al cuantil por exceso, es decir, $\omega = z_p$.

Ahora, considérese el problema de estimación clásico en el cuál se tiene una muestra aleatoria $X_1, ..., X_n$ con f.d.a. $G(\theta; z)$ y supóngase que $\hat{\theta}$ es el EMV de θ. Dado que la DGVE es invertible, se tiene que para cualquier $p \in (0, 1)$, el cuantil por defecto está dado por $Q_p = G^{-1}(\theta; p)$ así,

por el principio de invarianza de los EMV un estimador natural para Q_p basado en $X_1, ..., X_n$ es

$$\hat{Q}_p = G^{-1}(\hat{\theta}; p).$$

. Los cuantiles por defecto de la DGVE están dados por,

$$Q_p = \begin{cases} \mu - \frac{\sigma}{\xi}\{[1 - [-\log(p)]^{-\xi}\}, & \xi \neq 0, \\ \\ \mu - \sigma \log[-\log(p)], & \xi = 0. \end{cases}$$

Sea $Q_{1-p} = z_p$ para todo $0 < p < 1$ donde $G(z_p) = 1 - p$, entonces los cuantiles por exceso están dados por,

$$z_p = \begin{cases} \mu - \frac{\sigma}{\xi}\{[1 - [-\log(1-p)]^{-\xi}\}, & \xi \neq 0, \\ \\ \mu - \sigma \log[-\log(1-p)], & \xi = 0. \end{cases}$$

Definición 3.3.1. *El período de retorno T de cualquier evento extremo, se define como el lapso de tiempo que en promedio se cree que será igualado o excedido dicho evento.*

En otras palabras, el periodo de retorno T es la frecuencia con la que se presenta tal evento extremo (lluvias torrenciales, temperaturas extremas, huracanes, terremotos etc.). En la TVE y en algunas áreas de aplicación como ingeniería, al cuantil por exceso z_p se le conoce como el nivel de retorno asociado con el periodo de retorno $T = 1/p$, es decir, se espera que el nivel z_p sea excedido en promedio una vez cada $T = 1/p$ unidades de tiempo.

Por sustitución de las estimaciones de máxima verosimilitud de los parámetros de la DGVE en (2.4), los estimadores de máxima verosimilitud de z_p para $0 < p < 1$, con el nivel de retorno $\frac{1}{p}$, se obtienen como,

$$\hat{z}_p = \begin{cases} \mu - \frac{\hat{\sigma}}{\hat{\xi}}[1 - y_p^{-\hat{\xi}}], & \hat{\xi} \neq 0, \\ \\ \hat{\mu} - \hat{\sigma} \log y_p, & \hat{\xi} = 0, \end{cases} \tag{3.3}$$

donde $y_p = -\log(1-p)$. Además, por el método delta,

$$Var(\hat{z}_p) \approx \nabla z_p^T V \nabla z_p, \tag{3.4}$$

donde V es la matriz de varianza-covarianza de $(\hat{\mu}, \hat{\sigma}, \hat{\xi})$ y

$$\begin{aligned} \nabla z_p^T &= \left[\frac{\partial z_p}{\partial \mu}, \frac{\partial z_p}{\partial \sigma}, \frac{\partial z_p}{\partial \xi}\right] \\ &= [1, -\xi^{-1}(1 - y_p^{-\xi}),\ \sigma\xi^{-1}(1 - y_p^{-\xi}) - \sigma\xi^{-1} y_p^{-\xi} \log y_p] \end{aligned}$$

evaluado en $(\hat{\mu}, \hat{\sigma}, \hat{\xi})$.

Por lo general, los períodos de retorno largos, correspondientes a valores pequeños de p, son los de mayor interés. Si $\hat{\xi} < 0$, es posible hacer inferencia sobre el extremo superior de la distribución, que es efectivamente el ”período de retorno de observación infinito”, correspondiente a z_p con $p = 0$. La estimación de máxima verosimilitud es,

$$\hat{z}_0 = \hat{\mu} - \frac{\hat{\sigma}}{\hat{\xi}},$$

y (3.4) sigue siendo válido con

$$\nabla z_0^T = [1, -\xi^{-1}, \sigma\xi^{-2}],$$

evaluado de nuevo en $(\hat{\mu}, \hat{\sigma}, \hat{\xi})$. Cuando $\hat{\xi} \geq 0$ el estimador de máxima verosimilitud del punto final superior es infinito.

Es necesario ser precavido a la hora de interpretar las inferencias del nivel de retorno, especialmente en el caso de los niveles de retorno correspondientes a períodos de retorno largos. En primer lugar, la aproximación normal a la distribución del estimador de máxima verosimilitud puede ser deficiente. Por lo general, se obtienen mejores aproximaciones a partir de la función de verosimilitud de perfil adecuada. Aún más, las estimaciones y sus medidas de precisión se basan en la suposición de que el modelo es correcto. Aunque el modelo de la DGVE se apoya en argumentos matemáticos, su uso en la extrapolación se basa en suposiciones no verificables, y las medidas de incertidumbre sobre los niveles de rendimiento deberían considerarse adecuadamente como límites inferiores que podrían ser mucho mayores si se tuviera en cuenta la incertidumbre debida a la corrección del modelo.

Capítulo 4

Aplicación: Precipitaciones pluviales máximas

En este Capítulo se analizan los datos de las precipitaciones pluviales máximas en el Estado de Oaxaca. Se aplicó la metodología de inferencia estadística presentada en el capítulo anterior a diversos conjuntos de datos, todos ellos pertenecientes a distintas estaciones meteorológicas ubicadas en las ocho regiones del Estado, se muestra el análisis de 3 estaciones meteorológicas, cada una de ellas en una región distinta.
En este Capítulo se utilizó el software estadístico R (R Core Team, 2024; Posit team, 2024), que tiene implementado el paquete `evd` (Stephenson, 2002), el cual se empleó para llevar a cabo el análisis, adicionalmente se implementaron algunos programas para complementar el análisis, los cuales se muestran en el Anexo A.
Para cada conjunto de datos bajo estudio se presentan algunas estadísticas descriptivas, se estiman los parámetros de la distribución de valores extremos generalizada, se realiza la evaluación del modelo ajustado empleando un método gráfico de diagnóstico y se estiman los niveles de retorno que son de gran interés para la gestión de riesgos.

4.1. Descripción del problema

El cambio climático se define como "todo cambio que ocurre en el clima a través del tiempo, resultado de la variabilidad natural o de las actividades humanas" (Change *et al.*, 2014), diversos estudios han señalado que existen claras evidencias de que el calentamiento global registrado en los últimos 50 años es atribuible a las actividades humanas (Houghton, 2009; Dessler, 2021). Este fenómeno impacta en precipitaciones, temperaturas y eventos extremos, afectando ecosistemas y sociedades. Entre las variables del cambio climático se encuentran:

- Concentración de dióxido de carbono CO_2 (la quema de combustibles fósiles y los

cambios en la cobertura del suelo aumentan la concentración de este gas en la atmósfera).

- Metano CH_4 (el desarrollo de petróleo y gas genera metano, un gas de efecto invernadero más potente que el dióxido de carbono).
- Óxido nitroso N_2O y ozono O_3 (de los gases más abundantes en la atmósfera).
- Temperatura (el grado de frío o calor de una sustancia).
- Humedad relativa (la cantidad de vapor de agua que puede contener el aire).
- Precipitación (la cantidad de lluvia que recibe una región).

Por otro lado, la **precipitación pluvial** se refiere a cualquier forma de agua, sólida o líquida, que cae de la atmósfera y alcanza a la superficie terrestre. Esto incluye básicamente: lluvia, nieve y granizo (también rocío y escarcha, que en algunas regiones constituyen una parte pequeña pero apreciable de la precipitación total). Mientras que la intensidad de precipitación es la relación que existe entre la precipitación pluvial y el tiempo, generalmente suele expresarse como $mm/hora$. En relación con su origen, se distinguen los siguientes tipos:

- Las *ciclónicas* son las provocadas por los frentes asociados a una borrasca o ciclón.
- Las de *convección* se producen por el ascenso de bolsas de aire caliente; son las tormentas de verano.
- Las precipitaciones *orográficas* se presentan cuando masas de aire húmedo son obligadas a ascender al encontrar una barrera montañosa.

El estudio de las precipitaciones pluviales es básico dentro de cualquier estudio hidrológico regional, para cuantificar los recursos hídricos, puesto que constituyen la entrada de agua a una cuenca. También es fundamental en la previsión de avenidas, diseño de obras públicas, estudios de erosión, etc. Las precipitaciones caídas se cuantifican en un punto mediante cualquier recipiente de paredes rectas, midiendo después la lámina de agua recogida. La unidad de litros sobre metro cuadrado utilizada en los medios de comunicación es equivalente al milímetro (mm), es decir, un litro repartido por una superficie de $1m^2$ origina una lámina de agua de $1mm$, por esta razón se considera que la unidad de medida para las precipitaciones es el mm. Para poder leer con más precisión el agua recogida se hace uso del pluviómetro; es un mecanismo que básicamente funciona de la siguiente forma: el pluviómetro recoge el agua en una “bureta”de sección menor que la de la boca del pluviómetro. La lectura del agua recogida se efectúa una vez al día y es registrada de manera digital; anteriormente se hacía uso del pluviograma. Dependiendo de los objetivos del trabajo, se registran precipitaciones diarias, mensuales y anuales.

Debido a que no hay registros sobre un análisis de precipitaciones pluviales para el Estado de Oaxaca, en este trabajo se realiza un análisis probabilístico sobre los máximos de éstas para

observar el comportamiento de las precipitaciones pluviales máximas que se disponen de los registros meteorológicos diarios recabados por las estaciones climatológicas a cargo del Servicio Meteorológico Nacional (SMN) de la Comisión Nacional del Agua (CONAGUA).
Para ello se ajustó la distribución de valores extremos generalizada aplicando el método de máximos por bloques. Ya que se cuenta con una base de datos con 6024 mediciones diarias de las precipitaciones pluviales del Estado de Oaxaca en 89 lugares distintos, el método de bloques máximos es adecuado para obtener el análisis deseado considerando períodos anuales.

4.2. Descripción y manipulación de los datos

Los datos diarios de las precipitaciones pluviales se descargaron de la página web (https://smn.conagua.gob.mx), los cuales son datos recopilados diariamente expresados en milímetros (*mm*) por la Coordinación General del Servicio Meteorológico Nacional.

Originalmente, los datos considerados en este análisis consistían en registros diarios de precipitaciones pluviales, correspondientes a 89 estaciones con rangos de años desde 2000 al 2017. Esta base de datos tuvo que depurarse debido a que, por el método que se desean estudiar los datos, los días en los que la medición de la precipitación fue igual a cero se eliminaron.
Posterior a esta primera depuración de la base de datos se eliminaron las estaciones de las cuales no se encontró el registro. Así, se contó con información correspondiente a 59 estaciones, con rangos de años desde 2000 al 2017. La Figura 4.1 muestra la ubicación geográfica de estas 59 estaciones meteorológicas. La organización de los datos después de las depuraciones se guardó en un archivo de Excel por estación meteorológica, ya que cada una de las diferentes estaciones del Estado contiene diferentes períodos de información, y se optó de esta manera para no perder registro de la fecha por la eliminación de filas que guardan un registro de precipitación pluvial igual a 0.

Al revisar todas las estaciones de monitoreo se seleccionaron únicamente ocho estaciones meteorológicas (una por región), esto con la finalidad de analizar el comportamiento general de las precipitaciones pluviales máximas en estas regiones. Cabe mencionar que, aunque diversas de estas estaciones contienen registros del 01 de enero del 1961 al 30 de junio del 2024, para los datos estudiados sólo se consideran a partir del 01 de enero de 2000 al 31 de diciembre de 2017, teniendo así un número total de dieciocho años disponibles con registros de precipitaciones pluviales diarias en cada uno de los años.

La hipótesis general es que los datos de precipitaciones pluviales máximas se ajustan o pueden aproximarse con la distribución de valores extremos generalizada. En la teoría de valores extremos se contemplan observaciones aleatorias e independientes; sin embargo, las

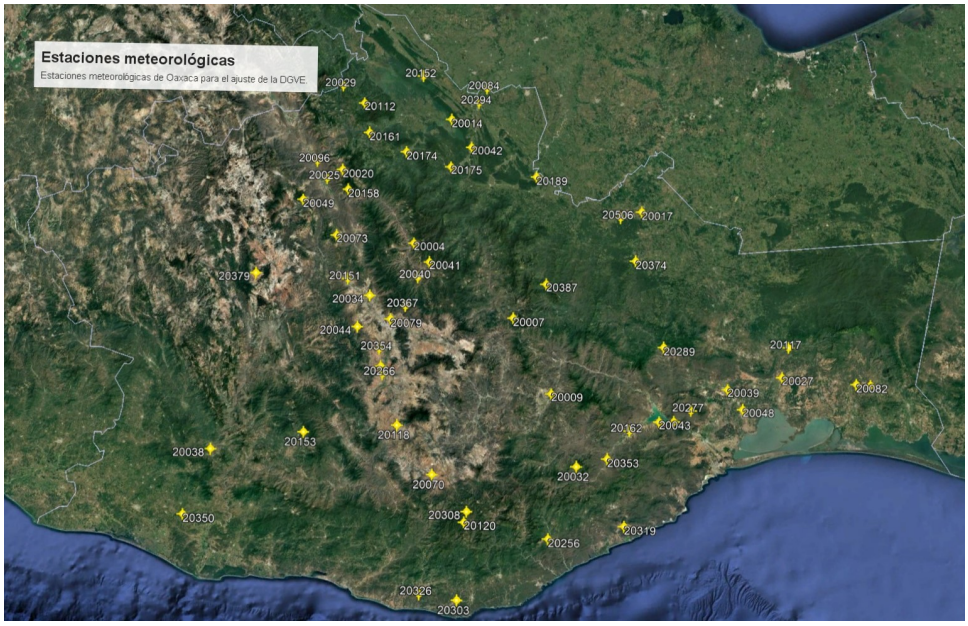

Figura 4.1: Estaciones meteorológicas en el Estado de Oaxaca.

observaciones de precipitaciones pluviales que se analizan no cumplen con esta propiedad debido a que los datos son consecutivos y están bajo las mismas condiciones ambientales, por tanto, se procede a tomar doce bloques anuales de los datos disponibles para cada estación y de esta manera disminuir la correlación entre los datos de precipitaciones pluviales.

4.3. Análisis de los datos

En la Tabla 4.1 se muestra el nombre de la estación meteorológica, región, clave, las coordenadas geográficas y la altitud de cada estación bajo estudio. Esta información permite ubicar en el mapa del Estado de Oaxaca a las estaciones meteorológicas seleccionadas; en la Figura 4.2 se muestra la ubicación geográfica de éstas.

Nota 4.3.1. *Los metros sobre el nivel del mar (m.s.n.m.) son una unidad de medida estándar del sistema métrico decimal para describir la elevación de un lugar del planeta Tierra respecto del nivel medio del mar de ese lugar.*

Nota 4.3.2. *Las Figuras 4.1 y 4.2 se obtuvieron mediante Google Earth Pro, que es un programa gratuito e informático similar a un sistema de información geográfico que permite ver imágenes en 3 dimensiones del planeta Tierra combinando imágenes satelitales, mapas y que además cuenta con funciones avanzadas, como importar y exportar datos.*

A continuación se hace un análisis de tres estaciones meteorológicas seleccionadas en las ocho regiones del Estado de Oaxaca. Las estaciones que a continuación serán analizadas son la de

Estación meteorológica	Región	Clave	Longitud	Latitud	Altitud m.s.n.m
Ayutla	Sierra Norte	20007	-96.099722°	17.016667°	2014
Santa María Ecatepec	Sierra Sur	20032	-95.883056°	16.283056°	1869
Santa María Jacatepec	Papaloapan	20042	-96.200000°	17.866667°	47
Quiotepec	Cañada	20096	-96.9905556°	17.890000°	543
San Miguel Chimalapa	Istmo	20117	-94.748333°	16.711667°	120
San Francisco Telixtlahuaca	Valles Centrales	20151	-96.900000°	17.300000°	2260
Cozoaltepec	Costa	20326	-96.723333°	15.789444°	145
Yodocono de Porfirio Díaz	Mixteca	20379	-97.357500°	17.380833°	2310

Tabla 4.1: Información geográfica de las estaciones meteorológicas en estudio.

Figura 4.2: Estaciones meteorológicas seleccionadas por región para el ajuste de la DGVE.

Quiotepec; de la región de la Cañada, la de San Miguel Chimalpa; de la región Istmo y la de San Francisco Telixtlahuaca de la región de Valles Centrales.

Región Cañada

La región Cañada es la región más pequeña del estado de Oaxaca, abarca una superficie de 4,273 km^2, se subdivide en 45 municipios agrupados en dos distritos: Teotitlán y Cuicatlán. Esta región representa la séptima concentración poblacional en el Estado de Oaxaca y constituye el 5 % de su población total (199,755 habitantes). En general tiene un clima, dentro de los secos muy cálidos y semicálido y templado, con lluvias mínimas de 372.8 *mm* y

máximas de 643.7 *mm* total anual. La economía de la región Cañada se basa principalmente en la gran cantidad de cultivos, como de maíz, café y de frutas como el chicozapote, mango, papaya, sandía, limón, ciruelas y melón. Cabe destacar que el maíz es el producto que registra la contribución más importante con 46 % del valor de la producción total, seguido del café que participa con el 26 %, y el limón con el 11 %.
La estación seleccionada para esta región es la estación meteorológica de Quiotepec. Se encuentra ubicada en el municipio de San Juan Bautista Cuicatlán; el municipio se encuentra ubicado en el centro-norte del Estado de Oaxaca. Forma parte del distrito de Cuicatlán y tiene una extensión territorial de 498.008 km^2 que representa el 0.53 % de la extensión total del estado.

Las precipitaciones pluviales máximas anuales sobre las cuales se llevó a cabo el ajuste para la estación de Quiotepec se muestran en la Tabla 4.3. Por otro lado, la Tabla 4.2 muestra algunas de las estadísticas descriptivas de dicha estación de monitoreo. Para el desarrollo de esta tesis, el primer cuartil se denotará por Q_1, el segundo cuartil (Q_2) es la mediana y el Q_3 hace referencia al tercer cuartil.

Mínimo	Q_1	**Mediana**	**Media**	Q_3	**Máximo**
23.20	37.41	50.65	58.40	74.30	120.60

Tabla 4.2: Estadísticas descriptivas de la estación Quiotepec.

Para analizar el comportamiento de las precipitaciones pluviales máximas se hace un diagrama de dispersión y un boxplot como se muestra en la Figura 4.3. La línea recta horizontal de color azul en el diagrama de dispersión es la media de las precipitaciones pluviales de la estación de Quiotepec, es decir, 58.40. En dicho gráfico se puede observar que las precipitaciones pluviales no siguen un comportamiento en general, vemos que el nivel más alto es de 120.6 *mm* en el bloque 11. Por otra parte, se presentan registros parecidos en el bloque 6 y 8, con precipitaciones pluviales de alrededor de 83 *mm*, el dato mínimo es del bloque 7 igual a 23.2 *mm*. En el diagrama de caja y bigote; Figura 4.3, se observa que las precipitaciones pluviales están centradas entre 37.41 *mm* y 74.30 *mm*, mientras que la mediana es de 50.65 *mm*. Además, se observa que no hay valores atípicos para esta estación. Ahora bien, si se calcula el rango intercuartilíco (IQR), se tiene que: $IQR = Q_3 - Q_1 = 74.30 - 37.41 = 36.89$, luego, la precipitación pluvial máxima de 120.6 *mm* no es un outlier ligero (ya que el valor de 120.60 se ubica a menos de 1.5 veces el IQR por sobre el tercer cuartil; $120.60 < Q_3 + 1.5IQR = 120.635$). Por lo tanto, dado que el valor del máximo no discrepa de los anteriores, se espera una distribución de máximos de colas no pesadas y aproximadamente simétrica.

Ahora se procede a ajustar la distribución de valores extremos generalizada a las observaciones. Por tanto, lo que se hace es estimar los tres parámetros de los que depende la función de distribución, dichas estimaciones son:

$$\hat{\mu} = 44.7034, \quad \hat{\sigma} = 18.2648 \quad \text{y} \quad \hat{\xi} = 0.1608.$$

Bloque	Precipitación máxima (*mm*)
1	46.4
2	35.6
3	53.5
4	36.0
5	53.0
6	83.0
7	23.2
8	83.6
9	62.3
10	47.0
11	120.6
12	113.1
13	37.0
14	64.7
15	77.5
16	27.7
17	48.3
18	38.62

Tabla 4.3: Datos de precipitaciones pluviales máximas por bloque en la estación de Quiotepec.

La matriz de varianza-covarianza aproximada de las estimaciones de los parámetros está dada por,

$$V = \begin{pmatrix} 25.6861 & 11.6794 & -0.4872 \\ 11.6794 & 16.1017 & -0.2524 \\ -0.4872 & -0.2524 & 0.0558 \end{pmatrix}$$

La diagonal de la matriz de varianza-covarianza corresponde a las varianzas de los parámetros individuales, $\hat{\mu}, \hat{\sigma}$ y $\hat{\xi}$, respectivamente. Al tomar la raíz cuadrada de los valores de la diagonal se obtiene que los errores estándares de $\hat{\mu}, \hat{\sigma}$ y $\hat{\xi}$ son 5.0691, 4.0149 y 0.2364 respectivamente.

De las estimaciones, el modelo o la función de distribución de las precipitaciones máximas es una Fréchet ya que el valor de $\hat{\xi} = 0.1608$ es mayor que cero. Sin embargo, $\hat{\xi}$ también es cercano a cero; lo cual sugiere que existe la posibilidad de que un modelo Gumbel también puede ser razonable a la luz de los datos. En contraste, el modelo Weibull no es adecuado bajo el criterio de estimación puntual.
Un estimador puntual no proporciona información acerca de la incertidumbre en la estimación. Para mostrar que valores de un parámetro son razonables o contradichos por los datos, se recomienda hallar los intervalos de verosimilitud perfil o los intervalos de confianza. Dado que la muestra es pequeña, conviene utilizar los intervalos de confianza, usando la normalidad de los EMV. Nótese que el valor de $\hat{\xi}$ es mayor que -0.5 y por lo tanto los EMV

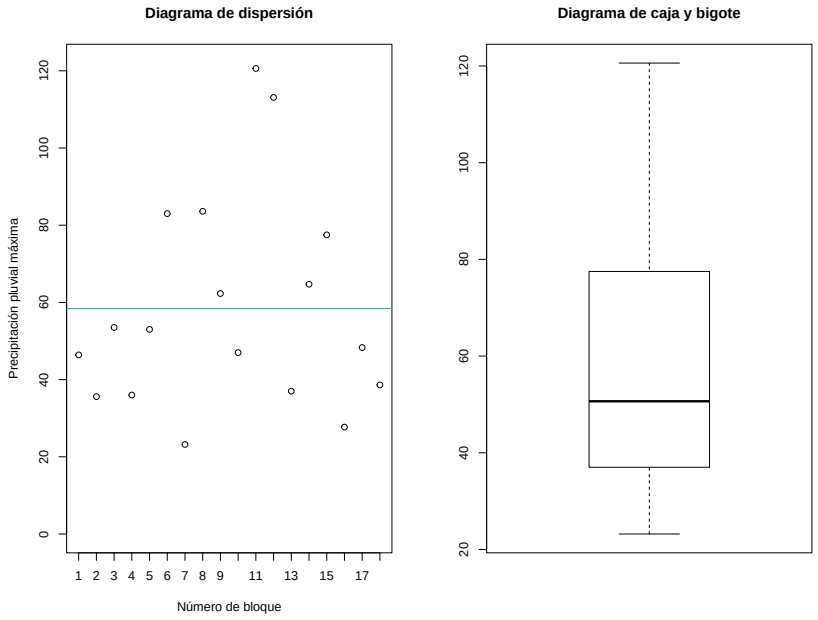

Figura 4.3: Gráfico de dispersión y gráfico de caja y bigote de la estación de Quiotepec.

cumplen las condiciones asintóticas usuales según Smith (1985), por lo que es posible calcular los intervalos de confianza de cada uno de los parámetros estimados para un $\alpha = 0.05$. Al combinar estimaciones y errores estándar se tiene que los intervalos de confianza aproximados del 95 % para cada parámetro están dados por,

$$\mu \in (34.7658,\ 54.6325),$$
$$\sigma \in (10.3947,\ 26.1242),$$
$$\xi \in (-0.3023,\ 0.6243).$$

Recordar que un intervalo de confianza es un intervalo que contiene el verdadero valor de algún parámetro desconocido con cierto grado de confiabilidad. De esta manera se puede observar que el intervalo de confianza del 95 % para ξ, captura el valor $\xi = 0$, lo cual indica la posibilidad de ajustar los máximos anuales a través de una función de distribución Gumbel. Es decir que, los datos no proporcionaron evidencia en contra del modelo Fréchet y Gumbel.

En seguida se aplican los métodos gráficos de diagnóstico para ver el ajuste de la distribución Fréchet.

Al apreciar las gráficas de la Figura 4.4 se observa que la validez del modelo ajustado es adecuada, ya que cada conjunto de puntos trazados quedan cerca de la línea identidad. Nótese que en la gráfica de cuantiles sólo los últimos dos registros se alejan de la línea identidad, sin embargo, no dista mucho de la recta $y = x$.

En seguida se ajustan los datos de precipitaciones pluviales máximas de la estación de

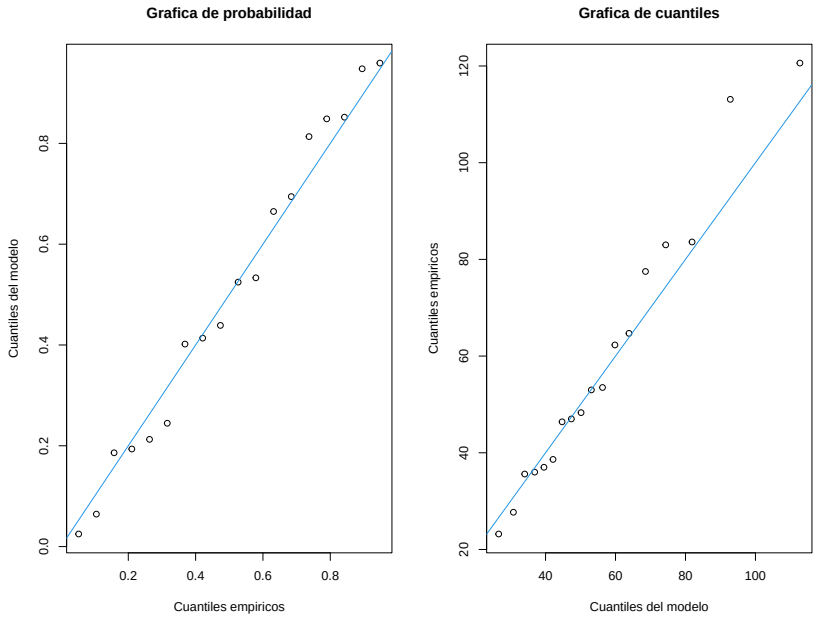

Figura 4.4: Gráfico de probabilidad y gráfico de cuantiles de la estación de Quiotepec.

Quiotepec usando una distribución Fréchet. La función de distribución Fréchet está dada por,

$$G(z) = \exp\left\{-\left[1 + 0.1608\left(\frac{z - 44.7034}{18.2648}\right)\right]^{-\frac{1}{0.1608}}\right\},$$

para $z \in (-68.8837, \infty)$.

Para visualizar el comportamiento de las precipitaciones pluviales máximas, se muestra la gráfica de la función de distribución acumulada y la gráfica de la función de densidad de probabilidad ajustada, en la Figura 4.5.

Observe que de la función de distribución acumulada de probabilidad ajustada se puede decir que, hay poca probabilidad de que la precipitación pluvial máxima en un año cualquiera haya sido menor de 20 *mm* o que es casi imposible. Por otro lado, la probabilidad de que la precipitación pluvial máxima anual en un año cualquiera haya sido menor de 120 *mm* es de 0.9743, luego la probabilidad de que la precipitación pluvial máxima anual haya sido mayor a 120 *mm* es de $1 - 0.9585 = 0.0415$, es decir, es muy poco probable que la precipitación pluvial máxima anual en un año cualquiera sea mayor a 120 *mm* lo cual es bastante realista. Lo anterior también se verifica analizando la función de densidad de probabilidad.

En la Figura 4.6 se muestra la gráfica de niveles de retorno y la gráfica de la función de densidad ajustada con el histograma de los datos reales, estas dos gráficas adicionales se utilizan para la presentación y validación del modelo ajustado. La gráfica de niveles de retorno muestra suficiente evidencia de que el modelo Fréchet es adecuado para los datos de

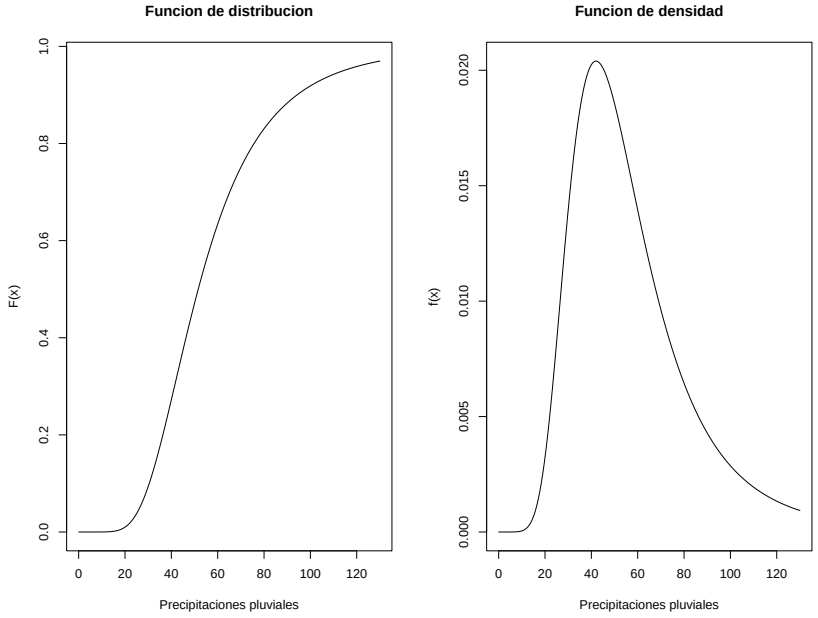

Figura 4.5: Función de distribución acumulada y función de densidad de probabilidad del modelo Fréchet para la estación de Quiotepec.

precipitación pluvial de Quiotepec, debido a que la gráfica es convexa. Para completar los gráficos de diagnóstico se muestra finalmente el gráfico de la función de densidad de probabilidad ajustada; dicho gráfico consiste en la comparación de la función de densidad de probabilidad del modelo ajustado (en nuestro caso el modelo Fréchet) con un histograma de los datos de precipitaciones pluviales máximas por bloque. Este último gráfico es menos informativo que los gráficos anteriores, ya que la forma de un histograma puede variar sustancialmente con la elección de los intervalos de agrupación. En el caso de la estación de Quiotepec se tienen 18 datos de precipitaciones pluviales máximas, y dado que el dato mínimo es de 23.20 y el máximo de 120.60 se optó por tomar 6 clases para el histograma, dichas clases en un rango de 20 a 140. Por lo tanto, de estos últimos dos gráficos se concluye que el modelo Fréchet ajusta adecuadamente a los datos.

A continuación se calculan los niveles de retorno para la estación de Quiotepec, cabe destacar que en general, en la teoría de valores extremos es de interés conocer niveles de retorno asociados a periodos de 20 y 100 años, respectivamente. Sin embargo, los periodos de retorno pueden variar dependiendo de la conveniencia de quien lo utiliza y del fenómeno natural que se estudia.

Para datos de precipitaciones pluviales máximas usualmente se calculan los niveles de retorno correspondientes a periodos de retorno de $T = 5, 10, 20, 50$ y 100 años. También se calculan los intervalos de confianza del 95 % correspondientes a cada nivel de retorno, cabe mencionar que generalmente se alcanza una mayor precisión de los intervalos de confianza del 95 % con el método de máxima verosimilitud, por lo que se aplicó ese método para estimar dichos

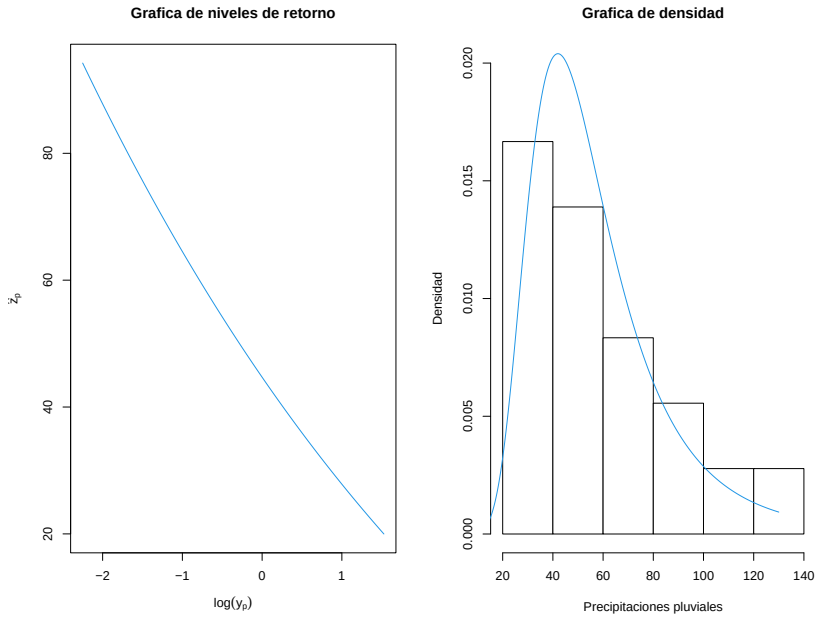

Figura 4.6: Gráfico de niveles de retorno e histograma con la función de densidad de probabilidad ajustada del modelo Fréchet para la estación de Quiotepec.

intervalos de confianza.

En la Tabla 4.4 se presentan los diferentes niveles de retorno ($\hat{z_p}$) asociados a los periodos de retorno (T) mencionados previamente. Además, se calculan las varianzas de cada nivel de retorno, $Var(z_p)$, usando el método delta y se presentan los IC del 95 % correspondientes a z_p. Notar que los niveles de retorno gradualmente se incrementan para periodos de retorno cada vez más grandes. También los IC del 95 % se hacen más anchos conforme el periodo de retorno se incrementa.

T en años	Nivel de retorno (mm)	$Var(z_p)$	IC del 95 %
5	75.68	100.1047	(56.07, 95.29)
10	94.22	264.15	(62.37, 126.08)
20	114.25	719.94	(61.66, 166.84)

Tabla 4.4: Periodos de retorno y niveles de retorno del modelo Fréchet para la estación de Quiotepec.

Para ilustrar como se obtiene la información de la Tabla 4.4, considere un periodo de retorno de 20 años, es decir, $T = 20$, entonces $p = \frac{1}{T} = \frac{1}{20} = 0.05$. El EMV para el nivel de retorno $\hat{z_p}$

cuando $\xi \neq 0$ está dado por,

$$\hat{z_p} = \hat{\mu} - \frac{\hat{\sigma}}{\hat{\xi}}\Big\{1 - \Big[-\log(1-p)\Big]^{-\hat{\xi}}\Big\}, \tag{4.1}$$

así, al sustituir el valor de $p = 0.05$, $\hat{\mu} = 44.7034$, $\hat{\sigma} = 18.2648$ y $\hat{\xi} = 0.1608$ en (4.1) se tiene que el nivel de retorno asociado al periodo de retorno $T = 20$ está dado por,

$$\hat{z}_{0.05} = 44.7034 - \frac{18.2648}{0.1608}\Big\{1 - \Big[-\log(1-0.05)\Big]^{-0.1608}\Big\} = 114.2512.$$

Para calcular la varianza de $\hat{z}_{0.05}$, es decir, $Var(\hat{z}_{0.05})$ se aplica el método delta para el caso $\xi \neq 0$. De aquí que,

$$Var(\hat{z}_p) \approx \nabla z_p^T V \nabla z_p,$$

donde V es la matriz de varianza-covarianza de $(\hat{\mu}, \hat{\sigma}, \hat{\xi})$ y

$$\nabla z_p^T = \Big(1, -\frac{1}{\hat{\xi}}\Big(1-[-\log(1-p)]^{-\hat{\xi}}\Big), \frac{\hat{\sigma}}{\hat{\xi}^2}\Big(1-[-\log(1-p)]^{\hat{\xi}}\Big) - \frac{\hat{\sigma}}{\hat{\xi}}[-\log(1-p)]^{-\hat{\xi}}\log(-\log(1-p))\Big). \tag{4.2}$$

Al sustituir las estimaciones de μ, σ y ξ y el valor de $p = 0.05$ en (4.2) se tiene que

$$\nabla z_{0.05}^T = (1,\ 3.809,\ 111.49),$$

luego,
$Var(\hat{z}_{0.05}) \approx \nabla z_{0.05}^T V \nabla z_{0.05} =$

$$\begin{pmatrix}1 & 3.809 & 111.49\end{pmatrix}\begin{pmatrix}25.6861 & 11.6794 & -0.4872\\ 11.6794 & 16.1017 & -0.2524\\ -0.4872 & -0.2524 & 0.0558\end{pmatrix}\begin{pmatrix}1\\ 3.809\\ 111.49)\end{pmatrix} = 719.9469.$$

Por lo tanto, el intervalo de confianza del 95 % para $z_{0.05}$ está dado por

$$\hat{z}_{0.05} \pm 1.96\sqrt{Var(\hat{z}_{0.05})} = 114.25 \pm 1.96\sqrt{719.9469} = (61.66, 166.84).$$

De manera similar para los demás periodos de retorno de la Tabla 4.4. La interpretación de la Tabla 4.4 es la siguiente: Por ejemplo, la estimación de máxima verosimilitud del nivel de retorno asociado a un periodo de retorno de 20 años fue $\hat{z}_{0.05} = 114.25$. Es decir, se estima que en 20 años se espera obtener una precipitación máxima de 114.25 *mm*. Por otro lado, el intervalo de confianza del 95 % para $\hat{z}_{0.05}$ es (61.66, 166.84), esto indica que con una confianza del 95 % la precipitación pluvial máxima que será sobrepasada una vez en 20 años se encuentra entre 61.66*mm* y 166.84*mm*. La interpretación es análoga para las diferentes estimaciones de los niveles de retorno correspondientes a los otros periodos de retorno.

Región Istmo

La región Istmo abarca una superficie de 20,755.26 km^2; lo que corresponde al 18 % del

territorio estatal, se subdivide en 41 municipios agrupados en dos distritos: Tehuantepec y Juchitán. La región representa la segunda concentración poblacional del Estado y constituye 15.9 % de su población total. Esta región tiene gran potencial con la industria eólica debido a los grandes vientos que predominan, los cuales provienen del golfo. La región también genera energía a través de otras fuentes alternativas, como la hidrológica y la solar.
Además, los principales productos de la región del Istmo son la producción del pasto y el maíz según el *Sistema de Información Agroalimentaria y Pesquera* (SIAP, 2015). Más aún, el Istmo de Tehuantepec se encuentra en el tercer lugar de participación de las ocho regiones del estado de Oaxaca en el valor de la producción agrícola después de la región Costa y la Cuenca del Papaloapan.

La región completa se encuentra en una zona de clima cálido tropical, sin embargo, debido a las elevaciones de la Sierra Atravesada y las montañas de Los Chimalapas, se presenta un marcado contraste climático. La precipitación anual en la vertiente atlántica del istmo de Tehuantepec alcanza los 3000 *mm*. La vertiente del Pacífico suele tener clima notablemente más seco. La precipitación anual en esta vertiente es de unos 900 *mm*.

La estación seleccionada para esta región es la estación meteorológica de San Miguel Chimalapa. Se encuentra ubicada en el municipio del mismo nombre y pertenece al distrito de Juchitepec. El rango de precipitación pluvial media anual es de 1800 a 2000 *mm* y los meses de lluvias son de octubre a mayo.

Las precipitaciones pluviales máximas sobre las cuales se llevó a cabo el ajuste para la estación de San Miguel Chimalapa se muestran en la Tabla 4.5. Por otro lado, la Tabla 4.6 muestra algunas de las estadísticas descriptivas de dicha estación de monitoreo.

Para analizar el comportamiento de las precipitaciones pluviales máximas anuales se hace un diagrama de dispersión y un boxplot como se muestra en la Figura 4.7. La línea recta horizontal de color azul en el diagrama de dispersión es la media de las precipitaciones pluviales máximas de la estación de San Miguel Chimalapa, es decir, 123.47. En dicho gráfico se puede observar que las precipitaciones pluviales no siguen un comportamiento o tendencia muy marcada, se observa que el nivel más alto es de 464.09 *mm* en el bloque 4. Por otra parte, se presentan registros iguales a 100 *mm* en el bloque 1 y 2, y se observan registros similares en el bloque 12 y 13, con precipitaciones pluviales de 51.6 *mm* y 50.5 *mm* respectivamente; el dato mínimo se ubica en el bloque 5 igual a 41.6 *mm*.

En el diagrama de caja y bigote, Figura 4.7, se observa que las precipitaciones pluviales están centradas entre 65.50 *mm* y 107.50 *mm*, mientras que la mediana es de 86.75 *mm*. Además, se observa que hay valores atípicos para esta estación. Ahora bien, si se calcula el rango intercuartilíco (IQR), se tiene que: $IQR = Q_3 - Q_1 = 107.50 - 65.50 = 42$, luego, la precipitación pluvial máxima de 464.09 *mm* es un outlier extremo (ya que el valor de 464.096 se ubica a más de 3 veces el IQR por sobre el tercer cuartil; $464.09 > Q_3 + 3IQR = 233.5$). Comparando las mediciones de los bloques 7 y 8; 317.6 *mm* y 207.7 *mm* respectivamente, se obtiene que ambos son outliers extremos. Esto indica que se espera una distribución de máximos de colas pesadas y nada simétrica.

Bloque	Precipitación máxima (*mm*)
1	100.0
2	100.0
3	160.0
4	464.1
5	41.6
6	46.2
7	317.6
8	207.7
9	80.0
10	83.5
11	90.0
12	51.6
13	50.5
14	108.0
15	64.5
16	82.4
17	68.5
18	106.0

Tabla 4.5: Datos de precipitaciones pluviales máximas por bloque en la estación de San Miguel Chimalapa.

Mínimo	Q_1	Mediana	Media	Q_3	Máximo
41.61	65.50	86.75	123.47	107.50	464.09

Tabla 4.6: Estadísticas descriptivas de la estación de San Miguel Chimalapa.

Debido a la ubicación de la estación, las ondas tropicales que se originan en el Océano Atlántico y que realizan un largo recorrido hasta el Océano Pacífico provocan fuertes precipitaciones en el Estado de Oaxaca, siendo el Istmo una de las regiones más afectadas. Los outliers que se presentan en los datos de estudio corresponden a la tormenta tropical "Arthur" y a las lluvias provocadas por la onda tropical No. 29 del huracán "Stan" (Gobierno de México, 2005, 2008; Diario Oficial de Federación, 2008).

Ahora se procede a ajustar la distribución de valores extremos generalizada a las precipitaciones pluviales máximas de la Tabla 4.5. Por tanto, lo que se hace es estimar los tres parámetros de los que depende la función de distribución; dichas estimaciones están dadas en la Tabla 4.7, junto con el error estándar y los intervalos de confianza.

La matriz de varianza-covarianza aproximada de las estimaciones de los parámetros está dada

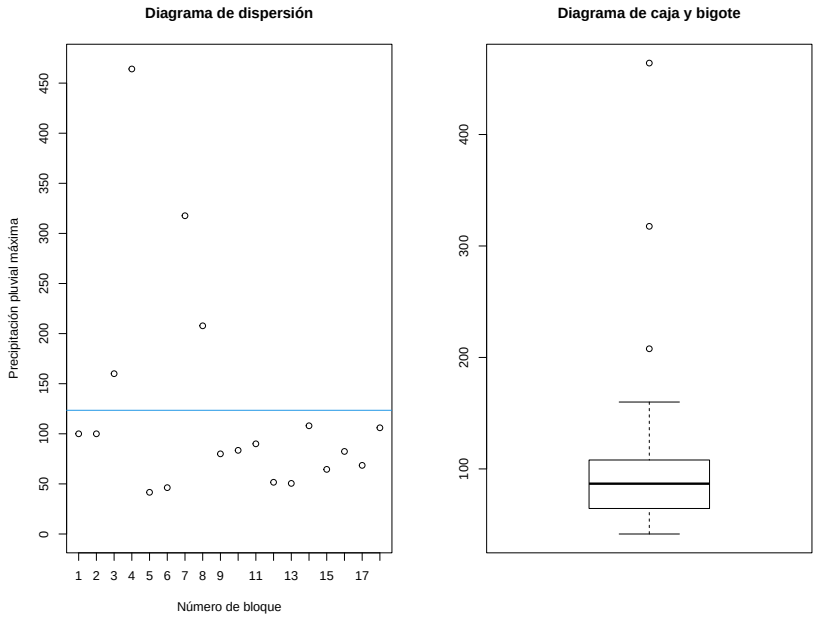

Figura 4.7: Gráfico de dispersión y gráfico de caja y bigote de la estación de San Miguel Chamalapa.

Parámetro	Estimación	Error estándar	IC del 95 %
$\hat{\mu}$	70.59	9.26	(52.44, 88.74)
$\hat{\sigma}$	32.73	9.55	(14.02, 51.43)
$\hat{\xi}$	0.63	0.302	(0.042, 1.23)

Tabla 4.7: Estimaciones de máxima verosimilitud e intervalos de confianza de los parámetros del modelo de la DGVE a la estación de San Miguel Chimalapa.

por,

$$V = \begin{pmatrix} 85.76 & 70.39 & -0.93 \\ 70.39 & 91.12 & 0.089 \\ -0.93 & 0.089 & 0.091 \end{pmatrix}$$

De las estimaciones, la función de distribución de las precipitaciones máximas es una Fréchet ya que el valor de $\hat{\xi} = 0.63$ es mayor que cero. Además, dado que el valor de $\hat{\xi}$ es mayor que -0.5, los EMV cumplen las condiciones asintóticas usuales, lo que implica que se puedan calcular los intervalos de confianza de cada uno de los parámetros estimados para un $\alpha = 0.05$. Sin embargo, el intervalo de confianza del 95 % para $\hat{\xi}$ rechaza la posibilidad de ajustar los máximos anuales a través de una función de distribución Gumbel. En contraste, el modelo Weibull no es adecuado bajo el criterio de estimación puntual.

En la Figura 4.8 se observa que el ajuste del modelo Fréchet a los datos es adecuado; pues el conjunto de puntos trazados quedan cerca de la línea identidad. Nótese que en la gráfica de cuantiles, los últimos tres registros se alejan de la línea identidad, lo que ya esperábamos que ocurriera porque se trata de datos atípicos.

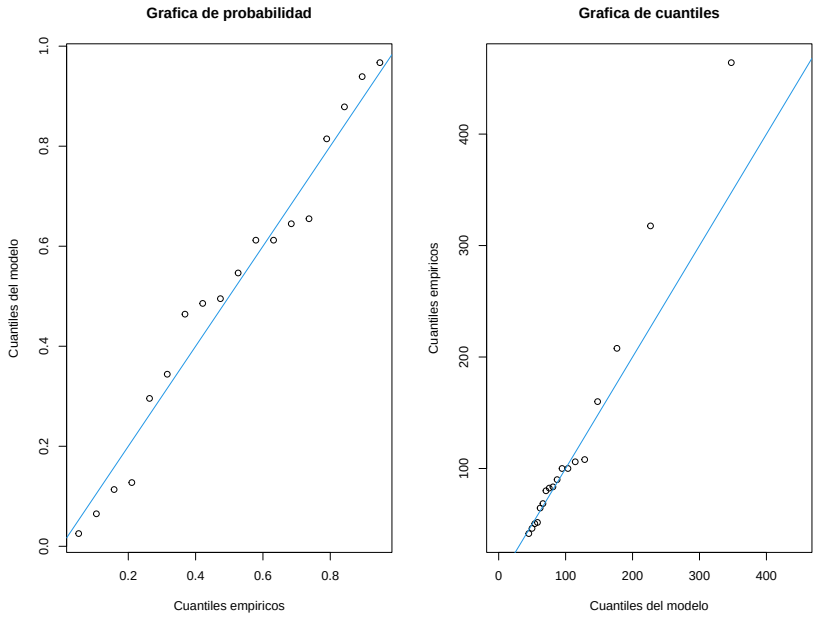

Figura 4.8: Gráfico de probabilidad y gráfico de cuantiles de la estación de San Miguel Chimalapa.

En seguida se ajustan los datos de precipitaciones pluviales máximas de la estación de San Miguel Chimalpa usando una distribución Fréchet. La función de distribución Fréchet está dada por,

$$G(z) = \exp\left\{-\left[1 + 0.63\left(\frac{z - 70.59}{32.73}\right)\right]^{-\frac{1}{0.63}}\right\},$$

para $z \in (18.64, \infty)$.

Para tener una idea del comportamiento de las precipitaciones pluviales máximas, se muestra la gráfica de la función de distribución acumulada y la gráfica de la función de densidad de probabilidad en la Figura 4.9.

Observe que de la función de distribución acumulada de probabilidad ajustada se puede decir que hay poca probabilidad de que la precipitación pluvial máxima en un año cualquiera haya sido menor de 41.61 *mm* o que es casi imposible. Por otro lado, es muy poco probable que la precipitación pluvial máxima anual en un año cualquiera sea mayor a 400 *mm* lo cual es bastante realista.

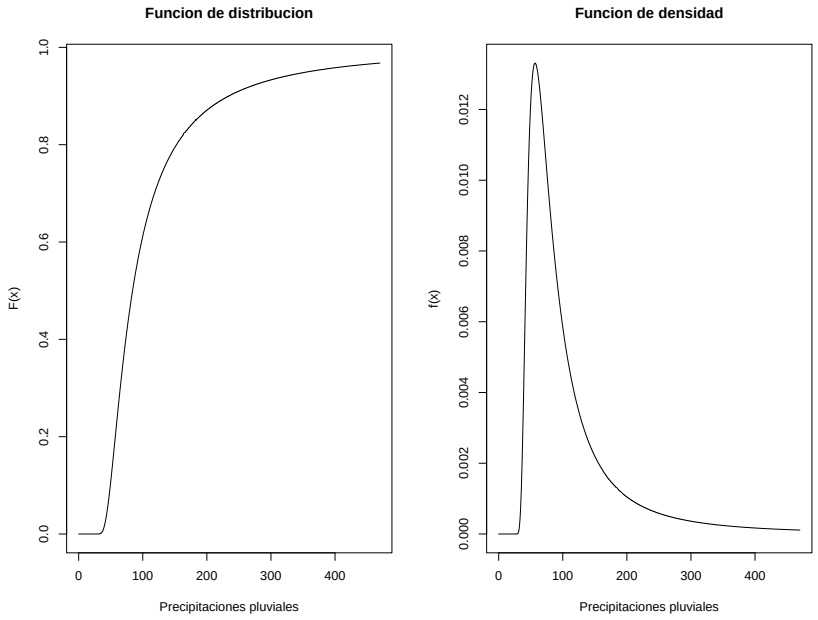

Figura 4.9: Función de distribución acumulada y función de densidad de probabilidad ajustada del modelo Fréchet para la estación de San Miguel Chimalapa.

En la Figura 4.10 se muestra la gráfica de niveles de retorno y el gráfico de la función de densidad de probabilidad ajustado en contraste a los datos empíricos; estas dos gráficas adicionales se utilizan para la presentación y validación del modelo ajustado. La gráfica de niveles de retorno muestra suficiente evidencia de que el modelo Fréchet es adecuado para los datos de precipitación pluvial de San Miguel Chimalapa, debido a que la gráfica es convexa. Finalmente, se muestra el gráfico de la función de densidad de probabilidad ajustada, que consiste en la comparación de la función de densidad de probabilidad del modelo ajustado (en nuestro caso el modelo Fréchet) con un histograma de los datos de precipitaciones pluviales máximas por bloque. De estos últimos dos gráficos se concluye que el modelo Fréchet ajusta adecuadamente a los datos.

A continuación se calculan los niveles de retorno para la estación de San Miguel Chimalapa, correspondientes a periodos de retorno de T igual a 5, 10 y 20 años. También se calculan los intervalos de confianza del 95 % correspondientes a cada nivel de retorno, ocupando el método de máxima verosimilitud.

En la Tabla 4.8 se presentan los diferentes niveles de retorno ($\hat{z_p}$) asociados a los periodos de retorno (T) mencionados previamente. Además, se calculan las varianzas de cada nivel de retorno, $Var(z_p)$, usando el método delta y se presentan los IC del 95 % correspondientes a z_p. Notar que los niveles de retorno gradualmente se incrementan para periodos de retorno cada vez más grandes. También los IC del 95 % se hacen más anchos conforme el periodo de retorno se incrementa.

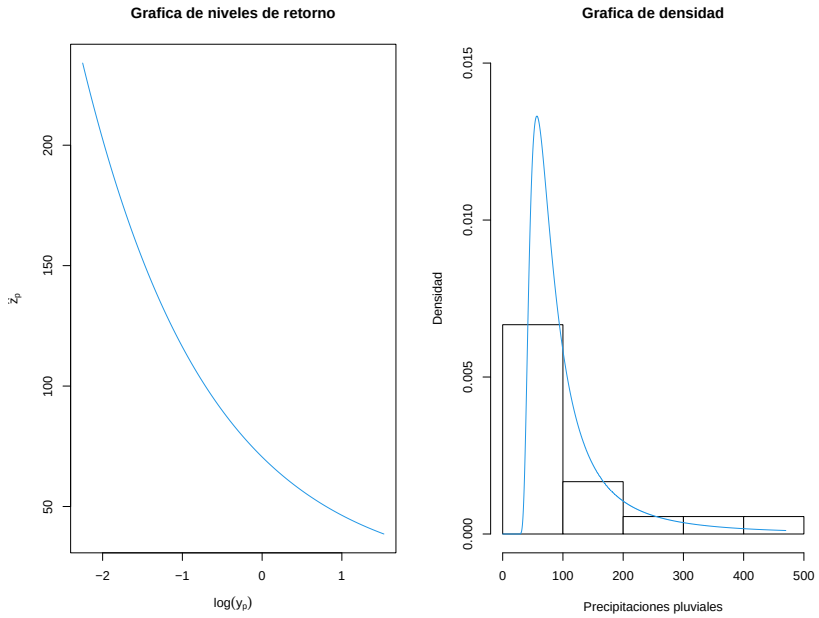

Figura 4.10: Gráfico de niveles de retorno e histograma con la función de densidad de probabilidad ajustada del modelo Fréchet para la estación de san Miguel Chimalapa.

T en años	Nivel de retorno (mm)	$Var(z_p)$	IC del 95 %
5	152.59	1372.75	(79.97, 225.21)
10	234.09	7524.24	(64.079, 404.11)
20	358.67	36342.04	(-14.97, 732.32)

Tabla 4.8: Periodos de retorno y niveles de retorno del modelo Fréchet para la estación de San Miguel Chimalapa.

Con la información obtenida de la Tabla 4.8 se puede decir que, la estimación de máxima verosimilitud del nivel de retorno asociado a un periodo de retorno de 10 años es $\hat{z}_{0.05} = 234.09$. Es decir, se estima que en 10 años se espera obtener una precipitación máxima de 234.09 *mm*. Por otro lado, el intervalo de confianza del 95 % para $\hat{z}_{0.1}$ es (64.079, 404.11), esto indica que con una confianza del 95 % la precipitación pluvial máxima que será sobrepasada una vez en 10 años se encuentra entre 64.079 *mm* y 404.11 *mm*. La interpretación es análoga para las diferentes estimaciones de los niveles de retorno correspondientes a los otros periodos de retorno.

Región Valles Centrales

La región de Valles Centrales de Oaxaca es una región geográfica y cultural del centro del

estado de Oaxaca con una extensión de 330,495 hectáreas y 103 municipios. Se trata de un conjunto de tres valles fluviales localizados entre el Nudo Mixteco, la Sierra Juárez y la Sierra Madre del Sur. Estos tres valles conforman una especie de ”Y”, cada uno de cuyos brazos posee un nombre específico: al noroeste se encuentra el valle de Etla; al oriente, el valle de Tlacolula; y al sur, el valle de Zimatlán-Ocotlán o valle Grande. El valle de Oaxaca se ha distinguido siempre por su producción textil y alfarera.
El área urbana se concentra en los valles y es la principal usuaria de los servicios ambientales que proporcionan los bosques de las zonas altas rurales. La mancha urbana se conforma principalmente por propiedad privada y concentra la mayoría de la población. Aquí se ubica la capital del estado que cuenta con 265,033 habitantes (7.6 % de la población estatal).

El acuífero de Valles Centrales está integrado por un sistema de cuatro microcuencas ubicadas en Coyotepec, Tlacolula, Oaxaca y Ocotlán, parte de la cuenca del río Atoyac. En el acuífero se han identificado 143 núcleos agrarios. La región consume anualmente 121.8 millones de m^3 de agua. El acuífero se está agotando debido a la excesiva extracción ocasionada por la construcción de pozos profundos con infraestructura inadecuada. Las montañas que rodean el valle incluyen bosques templados de pino-encino, selvas medianas, extensos territorios de bosque tropical caducifolio, bosque espinoso y una compleja asociación de cactáceas, matorrales y chaparral.

La estación seleccionada para esta región es la estación meteorológica de San Francisco Telixtlahuaca. Se encuentra ubicada en el municipio del mismo nombre y es cabecera del distrito de Etla. La precipitación total anual reportada para esta estación es de 794.1 *mm*; los meses de menor humedad son enero, febrero y diciembre.

Las precipitaciones pluviales máximas sobre las cuales se llevó a cabo el ajuste para la estación de San Francisco Telixtlahuaca se muestran en la Tabla 4.9. Por otro lado, la Tabla 4.10 muestra algunas de las estadísticas descriptivas de dicha estación.

Para analizar el comportamiento de las precipitaciones pluviales máximas anuales se hace un diagrama de dispersión y un boxplot como se muestra en la Figura 4.11. La línea recta horizontal de color azul en el diagrama de dispersión es la media de las precipitaciones pluviales máximas de la estación de San Francisco Telixtlahuaca, es decir, 53.63. En dicho gráfico se puede observar que las precipitaciones pluviales no siguen un comportamiento o tendencia muy marcada, se observa que el nivel más alto es de 97 *mm* en el bloque 12. Por otra parte, se presentan registros iguales a 48 *mm* en el bloque 9 y 16, y se observan registros similares en el bloque 12 y 13, con precipitaciones pluviales de 51.6 *mm* y 50.5 *mm* respectivamente; el dato mínimo se ubica en el bloque 5 igual a 41.6 *mm*.

En el diagrama de caja y bigote, Figura 4.11, se observa que las precipitaciones pluviales están centradas entre 43.50 *mm* y 58.50 *mm*, mientras que la mediana es de 48.0 *mm*. Además, se observa que hay un valor atípico para esta estación. Ahora bien, si se calcula el rango intercuartilíco (IQR), se tiene que: $IQR = Q_3 - Q_1 = 58.50 - 43.50 = 15$, luego, la precipitación pluvial máxima de 97.0 *mm* es un outlier extremo ($Q_3 + 1.5IQR = 81 < 97.0 < Q_3 + 3IQR = 103.5$).

Bloque	Precipitación máxima (*mm*)
1	51.0
2	54.0
3	59.0
4	45.0
5	57.0
6	42.0
7	47.0
8	77.0
9	48.0
10	43.0
11	40.0
12	97.0
13	70.0
14	40.0
15	65.0
16	48.0
17	37.3
18	45.0

Tabla 4.9: Datos de precipitaciones pluviales máximas por bloque en la estación de San Francisco Telixtlahuaca.

Mínimo	Q_1	Mediana	Media	Q_3	Máximo
37.31	43.50	48.00	53.63	58.50	97.00

Tabla 4.10: Estadísticas descriptivas de la estación de San Francisco Telixtlahuaca.

Ahora se procede a ajustar la distribución de valores extremos generalizada a las precipitaciones pluviales máximas de la Tabla 4.9. Por tanto, lo que se hace es estimar los tres parámetros de los que depende la función de distribución, dichas estimaciones están dadas en la Tabla 4.11, junto con el error estándar y los intervalos de confianza.

La matriz de varianza-covarianza aproximada de las estimaciones de los parámetros está dada por,

$$V = \begin{pmatrix} 4.79 & 2.92 & -0.2019 \\ 2.92 & 3.69 & -0.047 \\ -0.2019 & -0.047 & 0.065 \end{pmatrix}$$

De las estimaciones, la función de distribución de las precipitaciones máximas es una Fréchet

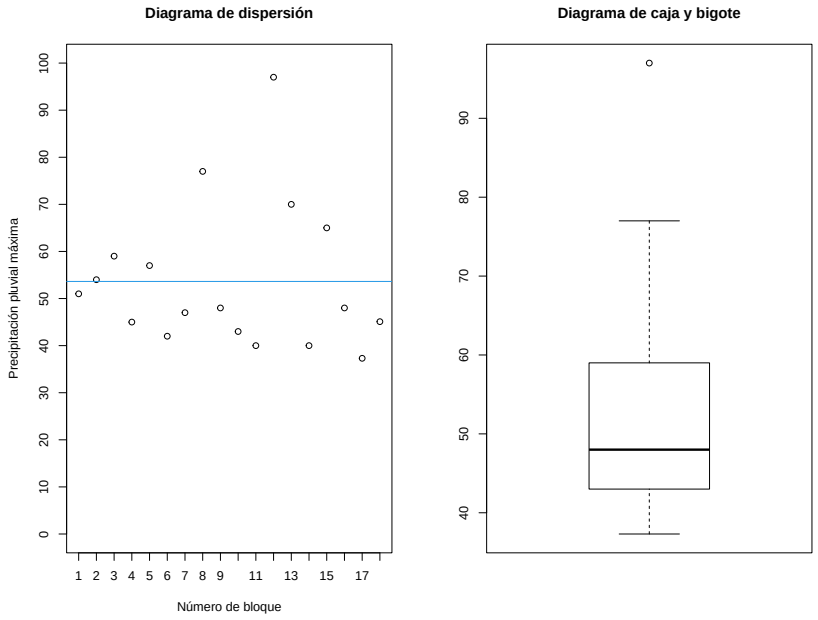

Figura 4.11: Gráfico de dispersión y gráfico de caja y bigote de la estación de San Francisco Telixtlahuaca.

Parámetro	Estimación	Error estándar	IC del 95 %
$\hat{\mu}$	45.66	2.18	(41.37, 49.95)
$\hat{\sigma}$	7.85	1.92	(4.09, 11.62)
$\hat{\xi}$	0.36	0.25	(-0.13, 0.86)

Tabla 4.11: Estimaciones de máxima verosimilitud e intervalos de confianza de los parámetros del modelo de la DGVE a la estación de San Francisco Telixtlahuaca.

ya que el valor de $\hat{\xi} = 0.36$ es mayor que cero. Además, dado que el valor de $\hat{\xi}$ es mayor que -0.5, los EMV cumplen las condiciones asintóticas usuales, lo que implica que se puedan calcular los intervalos de confianza de cada uno de los parámetros estimados para un $\alpha = 0.05$. Sin embargo, el intervalo de confianza del 95 % para $\hat{\xi}$ no rechaza la posibilidad de ajustar los máximos anuales a través de una función de distribución Gumbel.

En la Figura 4.12 se nota que el ajuste del modelo Fréchet a los datos es bueno; ya que el conjunto de puntos trazados quedan cerca de la línea identidad. Note que en la gráfica de cuantiles, sólo el último punto se aleja de la línea identidad, esto ocurre debido a que se trata de un outlier.

En seguida se ajustan los datos de precipitaciones pluviales máximas de la estación de San Francisco Telixtlahuaca usando una distribución Fréchet. La función de distribución Fréchet

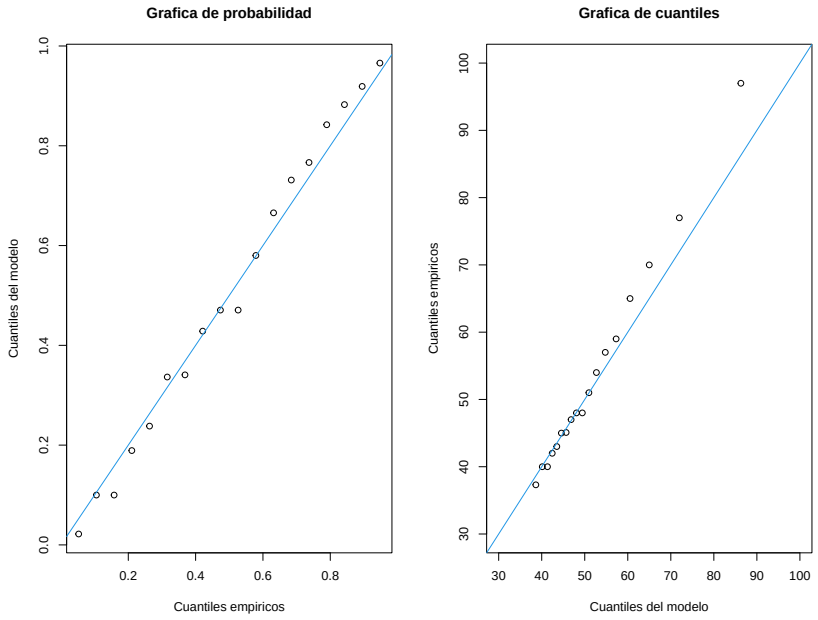

Figura 4.12: Gráfico de probabilidad y gráfico de cuantiles de la estación de San Francisco Telixtlahuaca.

está dada por,

$$G(z) = \exp\left\{-\left[1 + 0.36\left(\frac{z - 45.66}{7.85}\right)\right]^{-\frac{1}{0.36}}\right\},$$

para $z \in (23.85, \infty)$.

Para tener una idea del comportamiento de las precipitaciones pluviales máximas, se muestra la gráfica de la función de distribución acumulada y la gráfica de la función de densidad de probabilidad ajustada en la Figura 4.13.

En la Figura 4.14, la gráfica de niveles de retorno muestra suficiente evidencia de que el modelo Fréchet es adecuado para los datos de precipitación pluvial de San Francisco Telixtlahuaca, debido a que la gráfica es convexa. Para completar los gráficos de diagnóstico se muestra finalmente el gráfico de la función de densidad de probabilidad del modelo ajustado (en nuestro caso, el modelo Fréchet) con un histograma de los datos de precipitaciones pluviales máximas por bloque. De estos últimos dos gráficos se concluye que el modelo Fréchet ajusta adecuadamente a los datos.

A continuación se calculan los niveles de retorno para la estación de San Francisco Telixtlahuaca, correspondientes a periodos de retorno de T igual a 5, 10 y 20 años. También se calculan los intervalos de confianza del 95 % correspondientes a cada nivel de retorno, ocupando el método de máxima verosimilitud.

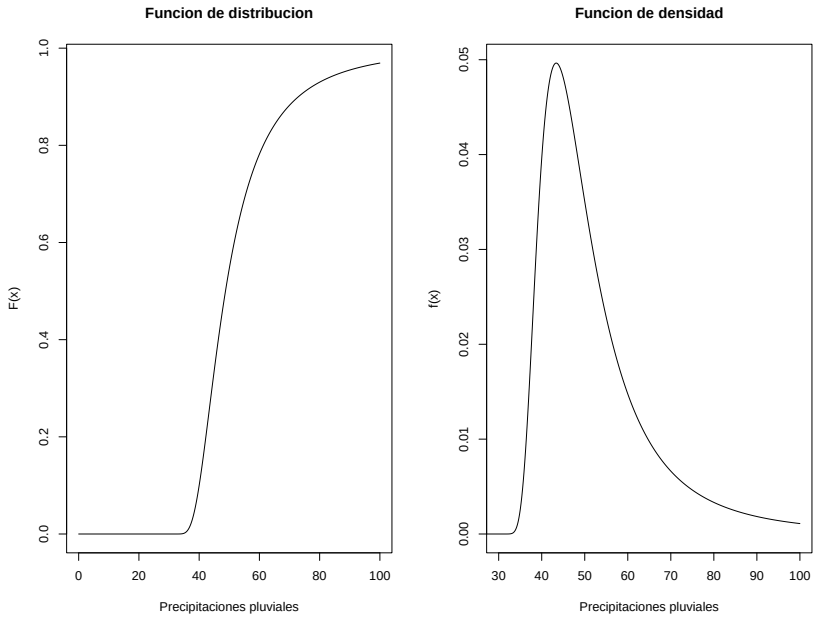

Figura 4.13: Función de distribución acumulada y función de densidad de probabilidad ajustada del modelo Fréchet para la estación de San Francisco Telixtlahuaca.

T en años	Nivel de retorno (mm)	$Var(z_p)$	IC del 95 %
5	61.29	34.08	(49.85, 72.73)
10	72.93	122.64	(51.22, 94.63)
20	87.47	419.93	(47.31, 127.64)

Tabla 4.12: Periodos de retorno y niveles de retorno del modelo Fréchet para la estación de San Francisco Telixtlahuaca.

En la Tabla 4.12 se presentan los diferentes niveles de retorno ($\hat{z_p}$) asociados a los periodos de retorno (T) mencionados previamente. Además, se calculan las varianzas de cada nivel de retorno, $Var(z_p)$, usando el método delta y se presentan los IC del 95 % correspondientes a z_p.

Con la información obtenida de la Tabla 4.12 se puede decir que, la estimación de máxima verosimilitud del nivel de retorno asociado a un periodo de retorno de 5 años es $\hat{z}_{0.2} = 61.29$. Es decir, se estima que en 5 años se espera obtener una precipitación máxima de 61.29 *mm*. Por otro lado, el intervalo de confianza del 95 % para $\hat{z}_{0.2}$ es (49.85, 72.73), esto indica que con una confianza del 95 % la precipitación pluvial máxima que será sobrepasada una vez en 10 años se encuentra entre 49.85 *mm* y 72.73 *mm*. La interpretación es análoga para las diferentes estimaciones de los niveles de retorno correspondientes a los otros periodos de retorno.

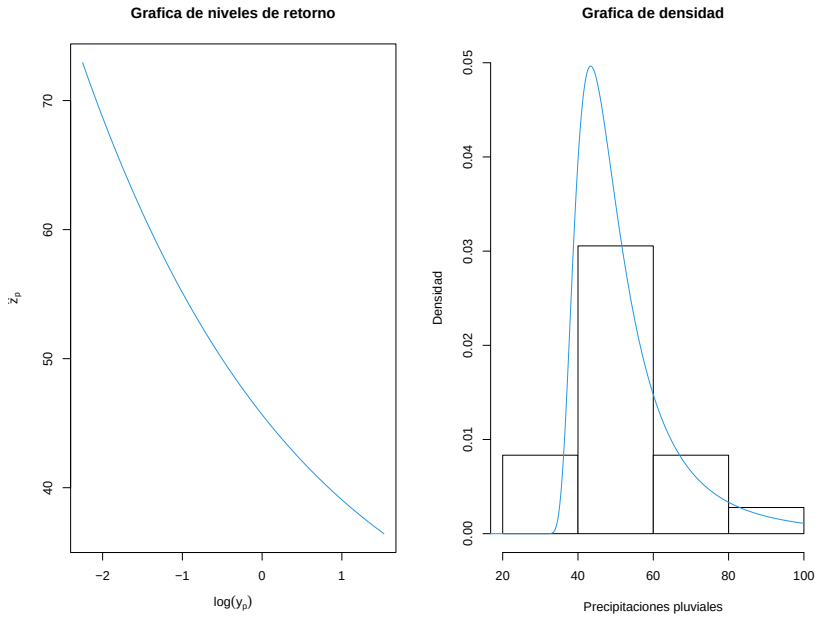

Figura 4.14: Gráfico de niveles de retorno e histograma con la función de densidad de probabilidad ajustada del modelo Fréchet para la estación de San Francisco Telixtlahuaca.

Capítulo 5

Conclusiones

La teoría de valores extremos es una rama de la estadística que estudia el comportamiento de los valores más altos o más bajos observados en una muestra. Posee un sustento probabilístico, el cual ha sido el resultado de varios años de estudio y aportaciones de una gran variedad de autores que poco a poco fueron contribuyendo a la sustentación, validación y formalización de dicha teoría. Su objetivo principal es modelar y predecir eventos raros y extremos, proporcionando herramientas para evaluar riesgos en situaciones donde ocurren fenómenos extremos. Se basa en resultados fundamentales como el Teorema de Fisher-Tippett-Gnedenko y el enfoque de la distribución generalizada de valores extremos (GEV), permitiendo una mejor comprensión de eventos extremos en diferentes contextos.

Esta teoría abarca diversas aplicaciones: en climatología, permite analizar eventos como huracanes, olas de calor y precipitaciones extremas. En finanzas, se usa para modelar riesgos de pérdidas extremas y crisis económicas. En ingeniería, se emplea para diseñar infraestructuras resilientes ante cargas extremas, como terremotos o inundaciones. También tiene aplicaciones en biología, seguros y ciencia de datos, donde la predicción de eventos poco frecuentes es crucial para la toma de decisiones y la gestión de riesgos.

Debido a la ubicación geográfica y su diversa topografía, el Estado de Oaxaca enfrenta diversos riesgos asociados a las precipitaciones pluviales máximas. Este fenómeno puede generar afectaciones significativas en infraestructura, comunidades y el medio ambiente; por ejemplo, inundaciones en las zonas bajas y costeras, como el Istmo, debido al desbordamiento de ríos y acumulación de agua en áreas urbanas como Oaxaca de Juárez. Las regiones montañosas de la Sierra Sur y Sierra Norte son vulnerables a deslaves debido al reblandecimiento del suelo por lluvias intensas, afectando caminos, viviendas y comunidades enteras. Los ríos como el Atoyac, el Papaloapan y el Tehuantepec pueden incrementar su caudal peligrosamente, afectando comunidades, carreteras y viviendas. Las lluvias pueden provocar derrumbes y bloqueos en carreteras que atraviesan zonas montañosas, afectando el tránsito y el abastecimiento de bienes esenciales. Tramos como la autopista Oaxaca-Cuacnopalan y la carretera a la Costa suelen ser afectados.

En este trabajo se realizó una aplicación de esta teoría, la cual consiste en el ajuste de la distribución generalizada de valores extremos a las precipitaciones pluviales máximas del Estado de Oaxaca, debido a la falta de estudios de precipitaciones pluviales en el Estado de Oaxaca.
Durante el proceso de la aplicación se encontró que,

- En general los ajustes que se realizaron a las estaciones de Quiotepec; de la región de la Cañada, San Miguel Chimalapa; de la región del Istmo y San Francisco Telixtlahuaca de la región de Valles Centrales, condujeron a modelos adecuados ocupando 18 bloques máximos, ya que se contó con información del 1 de enero de 2000 al 31 de diciembre del 2017 para las estaciones meteorológicas seleccionadas.

- A pesar de que la estación meteorológica de San Miguel Chimalapa presentara outliers extremos, el ajuste de la DGVE fue razonable para el conjunto de datos.

- Es posible que el ajuste de alguna de las distribuciones de valores extremos no sea adecuado. Esto puede deberse a un tamaño muestral insuficiente o a deficiencias en la recopilación de datos. Asimismo, esta situación puede surgir si no se cumplen ciertas hipótesis en la modelación de máximos por bloques, como la independencia de los datos o la utilización de bloques demasiado pequeños, lo que impide reducir la correlación entre las observaciones.

Referencias

Canavos, G. C. y Medal, E. G. U. (1987). *Probabilidad y estadística*. McGraw Hill México.

Change, I. C. *et al*. (2014). Mitigation of climate change. 1454, 147.

Coles, S., Bawa, J., Trenner, L. y Dorazio, P. (2001). tomo 208. Springer.

Contreras, R. J. L. y Jiménez-Hernández, J. C. (2020). *Análisis de temperaturas máximas en el Estado de Oaxaca*. Universidad Tecnológica de la Mixteca.

Dessler, A. E. (2021). *Introduction to Modern Climate Change*.

Diario Oficial de Federación (2008). Declaratoria de emergencia por la ocurrencia de lluvia atípica el dia 9 de octubre de 2008, en 16 municipios del Estado de Oaxaca. https://dof.gob.mx.

Englehart, P. J. y Douglas, A. V. (2000). *Dissecting the macro-scale variations in Mexican maize yields (1961-1997)*. 4, 1, 65–81.

Gnedenko, B. (1943). Sur la distribution limite du terme maximum dúne serie aleatoire. 44, 3, 423–453.

Gobierno de México (2005). Declaratoria de desastre natural, por las lluvias exrtemas ocasionadas por el huracan "Stan" los días 3, 4 y 5 de octubre de 2005, en diversos municipios del Estado de Oaxaca. https://www.ordenjuridico.gob.mx.

Gobierno de México (2008). Declaratoria de desastre natural por la ocurrencia de lluvias extremas los días 6 y 7 de julio de 2008, en 8 municipios del Estado de Oaxaca. https://www.dof.gob.mx.

González, E. y Macías, E. (2011). *Análisis de máximos para datos espaciales de lluvias*. Master's thesis, Cimat.

Haan, L. y Ferreira, A. (2006). tomo 3. Springer.

Houghton, J. (2009). *Global warming: the complete briefing*. Cambridge university press.

Lladser, M. (2011). *Variables aleatorias y simulación estocástica*. JC Sáez Editor.

Murray y Spiegel (2009). *Estadística*. MC GRAW HILL.

Posit team (2024). *RStudio: Integrated Development Environment for R*. Boston, MA.

R Core Team (2024). *R: A Language and Environment for Statistical Computing*. Vienna, Austria.

Referencias

Rincón, L. (2006). *Una introducción a la PROBABILIDAD Y ESTADÍSTICA*. *Distrito de México: Facultad de Ciencias UNAM*.

Rossi, R. J. (2018). *Mathematical statistics: an introduction to likelihood based inference*. John Wiley & Sons.

Smith, R. L. (1985). *Maximum likelihood estimation in a class of nonregular cases*. 72, 1, 67–90.

Stephenson, A. G. (2002). evd: Extreme Value Distributions. 2, 2, 31–32.

Anexos

Anexo A: Códigos en R

A continuación, se presenta el código que se ha utilizado en el software R para el análisis de datos presentado en el Capítulo 4.

Código para la estación de Quiotepec

```
# ---------- Análisis para la región de la Cañada-----------------------------

# Se cargan los datos para esta estación
datos<- read_excel("D:/Lenovo-Pc/Desktop/DatosLluvias/Datos sin
 ↪  cero/20096.xlsx")
x<-datos$`20096` #asigna datos a la variable
tamx<-length(x)#obtiene los datos del vector x

# Obtiene el vector de máximos ..........................................
n<-x
B=18#numero de bloques
h=trunc(tamx/B)#numero de datos por bloque
M<-numeric(length = B)#vector de maximos de tamaño B

for(j in 1:B){#Se utiliza para llenar el vector de maximos
  a<-(j-1)*h+1
  b<-j*h
  aux<-n[a:b]
  M[j]<-max(aux)
}
M
bloque<-(1:18)

```

```r
# Estadistica descriptiva ..................................................
summary(M)

# Diagrama de dispersión y boxplot de los datos con los que trabajamos.....

par(mfrow = c(1, 2)) #Para mostrar dos gráficos en una ventana
plot(bloque, M, xlab = 'Número de bloque', ylab='Precipitación pluvial máxima',
     main='Diagrama de dispersión', xlim = c(1, 18), ylim = c(0, 122),
     axes = F) #Gráfico izquierda
axis(1, at = seq(1, 18, 1), cex.axis = 1) #Para editar ejes
axis(2, at = seq(0, 120, 20), cex.axis = 1, las = 3.5)
box()
abline(h=58.40 , col=4) # Para asignar la línea de media de datos
boxplot(M, main = "Diagrama de caja y bigote", col=0) # Gráfico de derecha
par(mfrow = c(1, 1)) # Volvemos al estado original

# Obtiene los parámetros estimados de la DGVE ..............................

dism<-gev.fit(M)
pt<-dism$mle
mu<-pt[1]
sigma<-pt[2]
xi<-pt[3]

z<-fgev(x=M, std.err = TRUE,corr=TRUE, method= "Nelder-Mead")
z

# Los valores de la matriz de varianza-covarianza ..............................
M1<-fgev(M)
v<-M1$var.cov
v

# Obtiene los IC de los parametros estimados para un 95% usando propiedades de
# normalidad del estimador.....................................................

confint(M1,level=0.95)

# Graficos QQ y PP para el modelo Weibull y Fréchet ............................

par(mfrow = c(1,2))
orderM<- sort(M)
m<-length(M)
```

```r

# Grafica de probabilidad o grafica P-P
empirica<-rep(0,m)

fuction1<-function(i){
  i/(m+1)}
for(k in 1:m){
  empirica[k]<-fuction1(k)
}
empirica
modelo <-seq(from=1, to=m, by=1)
function2 <- function(z){
  exp((-1)*(1+xi*((z-mu)/sigma))^{-1/(xi)})
}
for(k in 1:m){
  modelo[k]<-function2(orderM[k])
}
modelo
plot(x=empirica, y=modelo, type="p", main="Grafica de probabilidad",lwd=1,
     xlab="Cuantiles empiricos", ylab="Cuantiles del modelo")
abline(0,1,lwd=1,col=4)

# Grafica de cuantiles o grafica Q-Q
fuction3<-function(i){
  mu+(-1)*(sigma/xi)*(1-(-log(i/(m+1)))^{-xi})
}
Empirica1<-rep(0,m)
Empirica1
for(k in 1:m){
  Empirica1[k]<-fuction3(k)
}
Empirica1
plot(x=Empirica1,y=orderM, main="Grafica de cuantiles",
     type="p",lwd=1,xlab="Cuantiles del modelo",ylab="Cuantiles empiricos")
abline(0,1,lwd=1,col=4)
par(mfrow = c(1,2))

# Grafica de la funcion de distribucion Weibull o
 ↪  Frechet.....................
pre_plu<-seq(from=0, to=130, by=0.01)
dis<-pgev(pre_plu,loc=mu, scale=sigma, shape=xi, lower.tail = TRUE)
dis
plot(x=pre_plu,y=dis,main="Funcion de distribucion", xlab ="Precipitaciones
 ↪  pluviales",
     ylab="F(x)", col="black",type="l",lwd=1)
# Grafica de la funcion de densidad Weibull o Frechet
 ↪  ........................
```

```
dis1<-dgev(x=pre_plu,loc=mu, scale=sigma, shape=xi)
dis1
plot(x=pre_plu,y=dis1,main="Funcion de densidad", xlab ="Precipitaciones pluviales",
     ylab="f(x)", col="black", type="l", lwd=1)

# Grafica de niveles de retorno para el caso Weibull y Frechet.................
par(mfrow = c(1,2))

ps <- seq(from=0.1, to=0.99, by=0.001)
ps
l <- length(ps)
fuction4 <- function(p){
  log((-1)*log(1-p))
}
periodo <- rep(0,l)
for(k in 1:l){
  periodo[k] <- fuction4(ps[k])
}
periodo
fuction5 <- function(p){
  mu-(sigma/xi)*(1-(-log(1-p))^{-xi})
}
niveles <- rep(0,l)
for(k in 1:l){
  niveles[k] <- fuction5(ps[k])
}
niveles
plot(x=periodo, y=niveles, xlab=expression(log(y[p])), ylab=expression(
  hat(z)[p]), main="Grafica de niveles de retorno",type="l",lwd=1, col=4)

# Histograma con la grafica de densidad ajustada del modelo Weibull y Fréchet...

hist(M,freq=F,main="Grafica de densidad",nclass=4,ylab="
Densidad", xlab="Precipitaciones pluviales", ylim=c(0, 0.020), col=0)
points(x=Tem,y=dis1,type="l",lwd=1, col=4)

# Calculo de periodos y niveles de retorno para las distribuciones
# Weibull y Frechet
# ..........................................................

p <- 1/t
p
fuction5 <- function(p){
  mu-(sigma/xi)*(1-(-log(1-p))^{-xi})
```

```
}
valor <- fuction5(p)
valor
# Calculo de la varianza a traves del metodo Delta
 ↪  ..............................
Matrizvarcov<-z$var.cov
Matrizvarcov
vector <- matrix(c(1,-xi^{-1}*(1-(-log(1-p))^{-xi}),sigma*xi
                   ^{-2}*(1-(-log(1-p))^{-xi})-sigma*xi^{-1}*(-log(1-p))^{-xi}
                   *log(-log(1-p))), nrow = 1, ncol = 3)
vector
vector_transpuesto<-t(vector)
vector_transpuesto
varianza <- vector%*%Matrizvarcov%*%vector_transpuesto
varianza
# Intervalos de confianza del 95%
 ↪  ........................................
valor-1.96*sqrt(varianza)
valor+1.96*sqrt(varianza)
```

Anexo B: Análisis adicionales para las estaciones meteorológicas presentadas en el Capítulo 4

A continuación se incluyen los resultados del análisis de las cinco estaciones meteorológicas restantes seleccionadas en las ocho regiones del Estado de Oaxaca. Las estaciones que a continuación serán analizadas son la de Ayutla; de la región Sierra Norte, la de Santa María Ecatepec; de la región Sierra Sur, la de Santa María Jacatepec; de la región del Papaloapan, la de Cozoaltepec; de la región de la costa y la de Yodocono de Porfirio Díaz de la región de la Mixteca.

Región Sierra Norte

La región de la Sierra Norte abarca una superficie de 8,944.77 km^2, se subdivide en 68 municipios agrupados en tres distritos: Ixtlán, Villa Alta y Mixe.

La estación seleccionada para esta región es la estación meteorológica de Ayutla. Se encuentra ubicada en el municipio de San Pedro y San Pablo Ayutla. El clima de San Pedro y San Pablo Ayutla varía de templado a frío húmedo. La época de lluvias va de mayo a octubre, pero llueve todo el año; la precipitación total anual es de 1480.6 *mm*.

Las precipitaciones pluviales máximas sobre las cuales se llevó a cabo el ajuste para la estación de Ayutla se muestran en la Tabla B1. Por otro lado, la Tabla B2 muestra algunas de las estadísticas descriptivas de dicha estación de monitoreo.

Para analizar el comportamiento de las precipitaciones pluviales máximas se hace un diagrama de dispersión como se muestra en la Figura B1. La línea recta horizontal de color azul en el diagrama de dispersión es la media de las precipitaciones pluviales de la estación de Ayutla, es decir, 70.72.

Se procede a ajustar la distribución de valores extremos generalizada a las precipitaciones pluviales máximas de la Tabla B1. Por tanto, se estiman los tres parámetros de los que depende la función de distribución; dichas estimaciones están dadas en la Tabla B3, junto con el error estándar y los intervalos de confianza.

La matriz de varianza-covarianza aproximada de las estimaciones de los parámetros está dada por,

$$V = \begin{pmatrix} 13.46 & 8.43 & -0.20 \\ 8.43 & 11.02 & 0.07 \\ -0.20 & 0.07 & 0.04 \end{pmatrix}$$

De las estimaciones, la función de distribución de las precipitaciones máximas es una Fréchet ya que el valor de $\hat{\xi} = 0.41$ es mayor que cero. Además, dado que el valor de $\hat{\xi}$ es mayor que -0.5, los EMV cumplen las condiciones asintóticas usuales, lo que implica que se puedan calcular los intervalos de confianza de cada uno de los parámetros estimados para un $\alpha = 0.05$. Sin embargo, el intervalo de confianza del 95 % para $\hat{\xi}$ no rechaza la posibilidad de ajustar los máximos de bloque a través de una

Bloque	**Precipitación máxima** (*mm*)
1	65.0
2	55.0
3	71.0
4	58.0
5	58.0
6	158.2
7	60.5
8	46.6
9	55.2
10	66.0
11	123.5
12	45.4
13	60.0
14	81.5
15	40.0
16	56.0
17	47.0
18	126.0

Tabla B1: Datos de precipitaciones pluviales máximas por bloque en la estación de Ayutla.

Mínimo	Q_1	**Mediana**	**Media**	Q_3	**Máximo**
40.00	55.05	59.00	70.72	69.75	158.20

Tabla B2: Estadísticas descriptivas de la estación Ayutla.

Parámetro	Estimación	Error estándar	IC del 95 %
$\hat{\mu}$	54.80	3.67	(47.61, 62.0)
$\hat{\sigma}$	13.68	3.32	(7.17, 20.19)
$\hat{\xi}$	0.41	0.22	(-0.02, 0.84)

Tabla B3: Estimaciones de máxima verosimilitud e intervalos de confianza de los parámetros del modelo de la DGVE a la estación de Ayutla.

función de distribución Gumbel.

En la Figura B2 se observa que el ajuste del modelo Fréchet a los datos es adecuado, pues el conjunto de puntos trazados queda cerca de la línea identidad.

Para tener una idea del comportamiento de las precipitaciones pluviales máximas, se muestra la gráfica

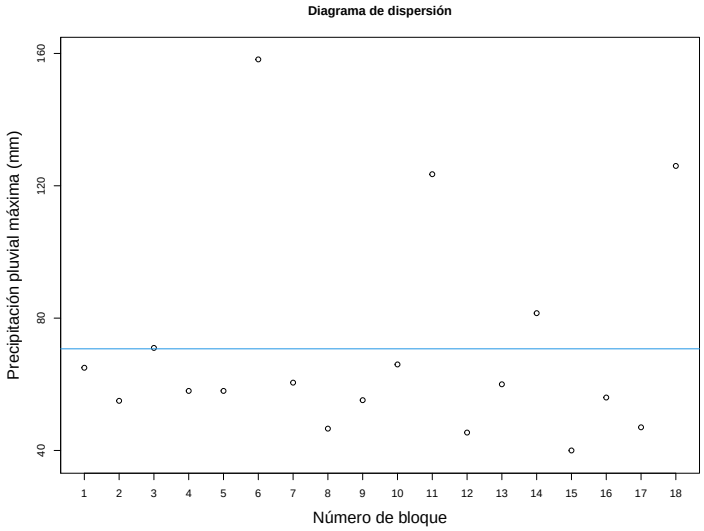

Figura B1: Gráfico de dispersión de la estación de Ayutla.

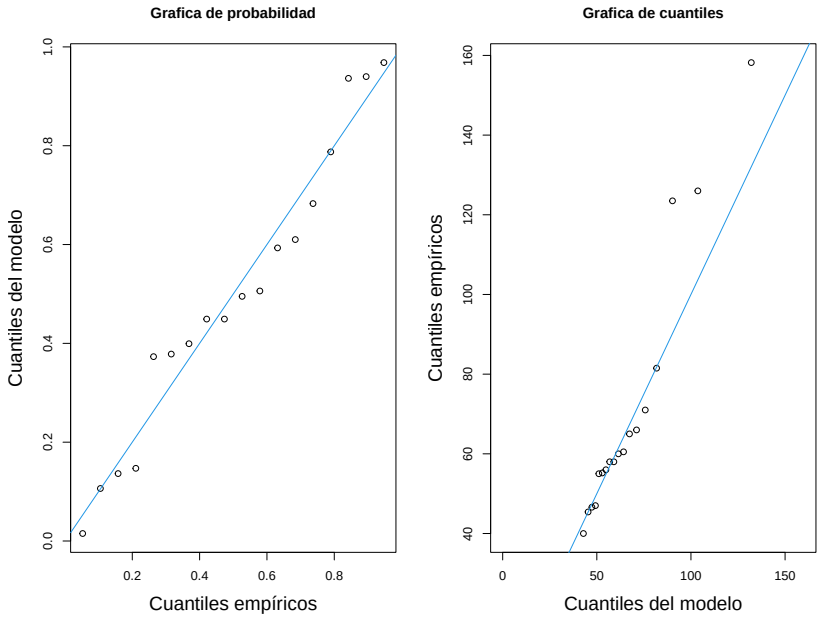

Figura B2: Gráfico de probabilidad y gráfico de cuantiles de la estación de Ayutla.

de la función de distribución acumulada en la Figura B3.

En la Figura B4 se muestra la gráfica de niveles de retorno, la cual muestra suficiente evidencia de que el modelo Fréchet es adecuado para los datos de precipitación pluvial de Ayutla; debido a que la gráfica es convexa.

En la Tabla B4 se presentan los diferentes niveles de retorno ($\hat{z_p}$) asociados a los periodos de retorno

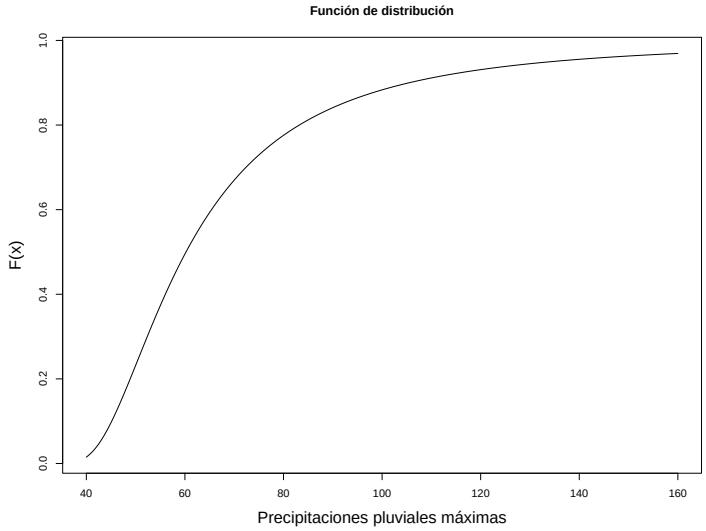

Figura B3: Función de distribución acumulada del modelo Fréchet para la estación de Ayutla.

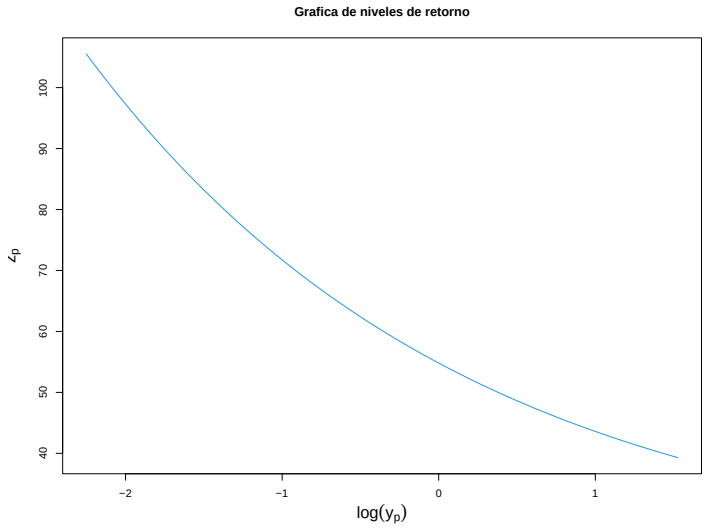

Figura B4: Gráfico de niveles de retorno para la estación de Ayutla.

T en años	Nivel de retorno (mm)	$Var(z_p)$	IC del 95 %
5	83.20	120.09	(61.72, 104.68)
10	105.51	446.06	(64.12, 146.91)
20	134.47	1526.81	(57.88, 211.05)

Tabla B4: Periodos de retorno y niveles de retorno del modelo Fréchet para la estación de Ayutla.

(T) igual a 5, 10 y 20 años. Además, se calculan las varianzas de cada nivel de retorno, $Var(z_p)$, usando el método delta y se presentan los IC del 95 % correspondientes a z_p. Notar que los niveles de retorno gradualmente se incrementan para periodos de retorno cada vez más grandes. También, los IC del 95 % se hacen más anchos conforme el periodo de retorno se incrementa.

Región Sierra Sur

La región Sierra Sur abarca una superficie de 14,753.26 km^2, se subdivide en 70 municipios agrupados en cuatro distritos: Putla, Sola de Vega, Miahuatlán y Yautepec. La región representa la sexta concentración poblacional en el Estado y constituye 8.5 % de su población total, la cual es 336, 421 habitantes. Las precipitaciones pluviales tienen un promedio anual que va de los 800 a 2000 *mm*.

La estación seleccionada para la región Sierra Sur es la estación Santa María Ecatepec, la cual se encuentra ubicada en el distrito de Yautepec. Las precipitaciones pluviales máximas sobre las cuales se llevó a cabo el ajuste para la estación de Santa María Ecatepec se muestran en la Tabla B5. La Tabla B6 muestra algunas de las estadísticas descriptivas de dicha estación de monitoreo.

Bloque	**Precipitación máxima** (*mm*)
1	52.0
2	25.6
3	40.0
4	39.1
5	18.3
6	165.0
7	21.3
8	55.5
9	190.0
10	86.0
11	62.5
12	79.0
13	128.5
14	105.0
15	160.0
16	38.0
17	45.0
18	120.0

Tabla B5: Datos de precipitaciones pluviales máximas por bloque en la estación de Santa María Ecatepec.

Para analizar el comportamiento de las precipitaciones pluviales máximas se hace un diagrama de dispersión como se muestra en la Figura B5. La línea recta horizontal de color azul en el diagrama de dispersión es la media de las precipitaciones pluviales de la estación de Santa María Ecatepec, es decir, 79.50.

Mínimo	Q_1	Mediana	Media	Q_3	Máximo
18.34	39.33	59.00	79.50	116.25	190.00

Tabla B6: Estadísticas descriptivas de la estación Santa María Ecatepec.

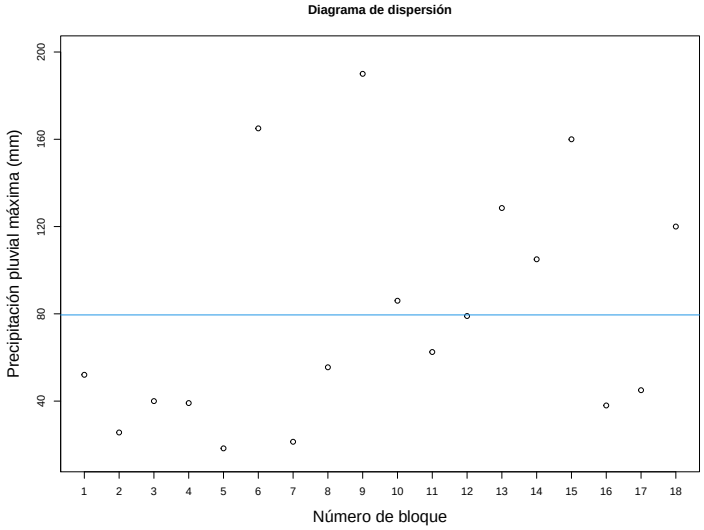

Figura B5: Gráfico de dispersión de la estación de Santa María Ecatepec.

Se procede a ajustar la distribución de valores extremos generalizada a las precipitaciones pluviales máximas de la Tabla B5. Por tanto, se estiman los tres parámetros de los que depende la función de distribución; dichas estimaciones están dadas en la Tabla B7, junto con el error estándar y los intervalos de confianza.

Parámetro	Estimación	Error estándar	IC del 95 %
$\hat{\mu}$	49.46	9.55	(30.74, 68.19)
$\hat{\sigma}$	31.94	8.34	(15.59, 48.28)
$\hat{\xi}$	0.33	0.32	(-0.29, 0.96)

Tabla B7: Estimaciones de máxima verosimilitud e intervalos de confianza de los parámetros del modelo de la DGVE a la estación de Santa María Ecatepec.

De las estimaciones, la función de distribución de las precipitaciones máximas es una Fréchet ya que el valor de $\hat{\xi} = 0.33$ es mayor que cero. Además, dado que el valor de $\hat{\xi}$ es mayor que -0.5, los EMV cumplen las condiciones asintóticas usuales, lo que implica que se puedan calcular los intervalos de confianza de cada uno de los parámetros estimados para un $\alpha = 0.05$. Sin embargo, el intervalo de confianza del 95 % para $\hat{\xi}$ no rechaza la posibilidad de ajustar los máximos de bloque a través de una función de distribución Gumbel.

En la Figura B6 se observa que el ajuste del modelo Fréchet a los datos es adecuado, pues el conjunto de puntos trazados queda cerca de la línea de identidad.

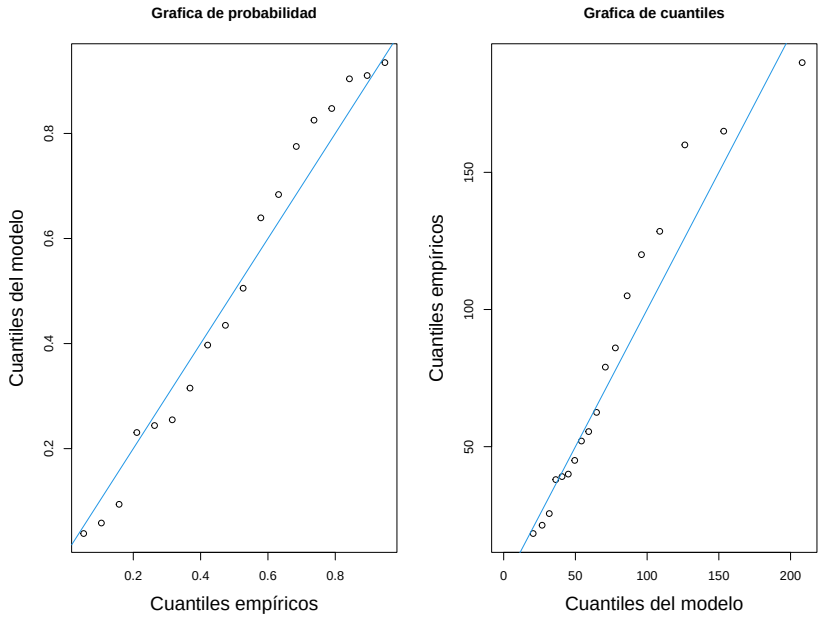

Figura B6: Gráfico de probabilidad y gráfico de cuantiles de la estación de Santa María Ecatepec.

Para tener una idea del comportamiento de las precipitaciones pluviales máximas, se muestra la gráfica de la función de distribución acumulada en la Figura B7.

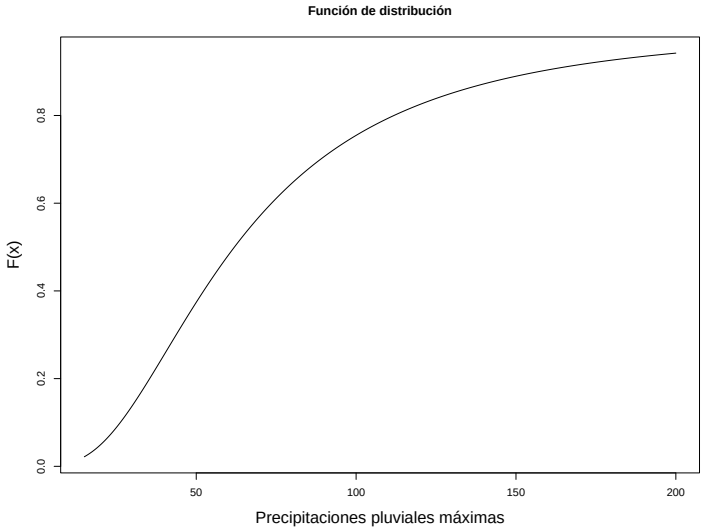

Figura B7: Función de distribución acumulada del modelo Fréchet para la estación de Santa María Ecatepec.

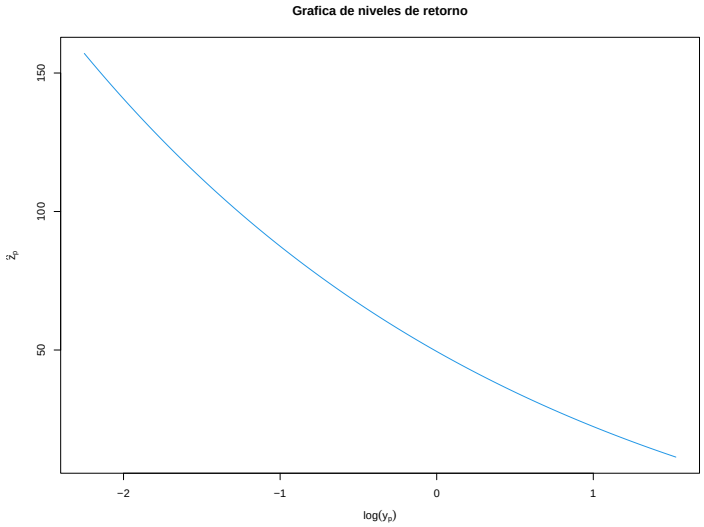

Figura B8: Gráfico de niveles de retorno para la estación de Santa María Ecatepec.

En la Figura B8 se muestra la gráfica de niveles de retorno, la cual muestra suficiente evidencia de que el modelo Fréchet es adecuado para los datos de precipitación pluvial de Santa María Ecatepec, debido a que la gráfica es convexa.

En la Tabla B8 se presentan los diferentes niveles de retorno ($\hat{z_p}$) asociados a los períodos de retorno (T) igual a 5, 10 y 20 años. Además, se calculan las varianzas de cada nivel de retorno, $Var(z_p)$, usando el método delta y se presentan los IC del 95 % correspondientes a z_p. Notar que los niveles de retorno gradualmente se incrementan para períodos de retorno cada vez más grandes. También, los IC del 95 % se hacen más anchos conforme el período de retorno se incrementa.

T en años	Nivel de retorno (mm)	$Var(z_p)$	IC del 95 %
5	111.83	528.74	(66.76, 156.90)
10	157.06	2008.74	(69.22, 244.91)
20	212.64	7467.03	(43.27, 382.0)

Tabla B8: Periodos de retorno y niveles de retorno del modelo Fréchet para la estación de Santa María Ecatepec.

Región del Papaloapan

La región de la Cuenca del Papaloapan abarca una superficie de $8,496.79 km^2$, se subdivide en 20 municipios agrupados en dos distritos: Tuxtepec y Choapam. La región representa la cuarta concentración de población en el estado y constituye 12.2 % de su población total, teniendo 482,149 habitantes. Su clima es estable, predominando el cálido húmedo con lluvias todo el año, semicálido

húmedo y el templado húmedo. La precipitación media anual es de 500 a 2,500 *mm* y las lluvias pueden ser escasas o intensas, dependiendo de las condiciones meteorológicas.

La estación seleccionada para la región del Papaloapan es la estación de Santa María Jacatepec; se encuentra en el municipio del mismo nombre y pertenece al distrito de Tuxtepec. La Tabla B9 muestra algunas de las estadísticas descriptivas de dicha estación de monitoreo.

Mínimo	Q_1	**Mediana**	**Media**	Q_3	**Máximo**
128.0	179.8	213.9	211.5	233.8	318.9

Tabla B9: Estadísticas descriptivas de la estación Santa María Jacatepec.

Para analizar el comportamiento de las precipitaciones pluviales máximas se hace un diagrama de dispersión como se muestra en la Figura B9. La línea recta horizontal de color azul en el diagrama de dispersión es la media de las precipitaciones pluviales de la estación de Santa María Jacatepec, es decir, 211.5.

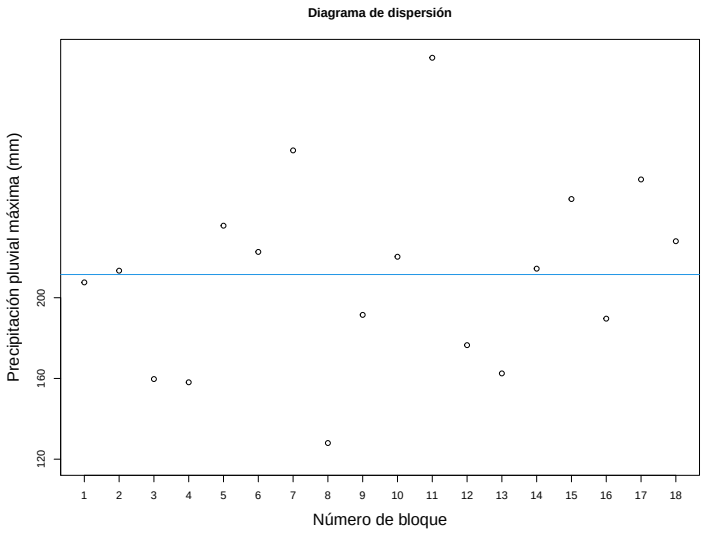

Figura B9: Gráfico de dispersión de la estación de Santa María Jacatepec.

Las estimaciones de los parámetros de los que depende la función de distribución se encuentran en la Tabla B10, junto con el error estándar y los intervalos de confianza de cada uno de los parámetros estimados.

De las estimaciones, la función de distribución de las precipitaciones máximas es una Weibull ya que el valor de $\hat{\xi} = -0.17$ es menor que cero. Además, dado que el valor de $\hat{\xi}$ es mayor que -0.5, los EMV cumplen las condiciones asintóticas usuales, lo que implica que se puedan calcular los intervalos de confianza de cada uno de los parámetros estimados para un $\alpha = 0.05$. Sin embargo, el intervalo de confianza del 95 % para $\hat{\xi}$ no rechaza la posibilidad de ajustar los máximos de bloque a través de una

Parámetro	Estimación	Error estándar	IC del 95 %
$\hat{\mu}$	193.49	11.09	(171.71, 215.20)
$\hat{\sigma}$	42.23	7.78	(26.97, 57.50)
$\hat{\xi}$	-0.17	0.15	(-0.49, 0.13)

Tabla B10: Estimaciones de máxima verosimilitud e intervalos de confianza de los parámetros del modelo de la DGVE a la estación de Santa María Jacatepec.

función de distribución Gumbel.

En la Figura B10 se observa que el ajuste del modelo Weibull a los datos es adecuado, pues el conjunto de puntos trazados queda cerca de la línea de identidad.

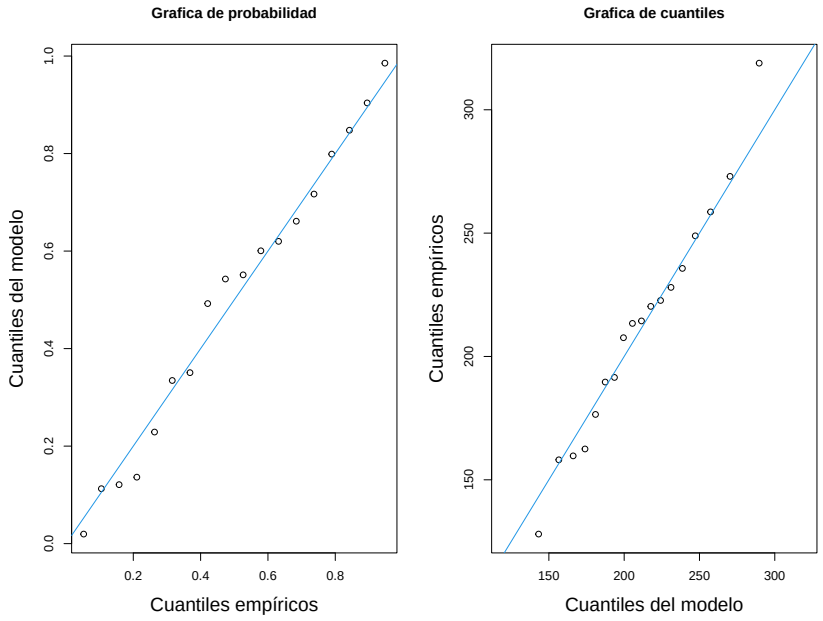

Figura B10: Gráfico de probabilidad y gráfico de cuantiles de la estación de Santa María Jacatepec.

Para tener una idea del comportamiento de las precipitaciones pluviales máximas, se muestra la gráfica de la función de distribución acumulada en la Figura B11.

En la Figura B12 se muestra la gráfica de niveles de retorno, la cual muestra suficiente evidencia de que el modelo Weibull es adecuado para los datos de precipitación pluvial de Santa María Jacatepec, debido a que la gráfica es cóncava.

En la Tabla B11 se presentan los diferentes niveles de retorno ($\hat{z_p}$) asociados a los períodos de retorno (T) igual a 5, 10 y 20 años. Además, se calculan las varianzas de cada nivel de retorno, $Var(z_p)$, usando el método delta y se presentan los IC del 95 % correspondientes a z_p.

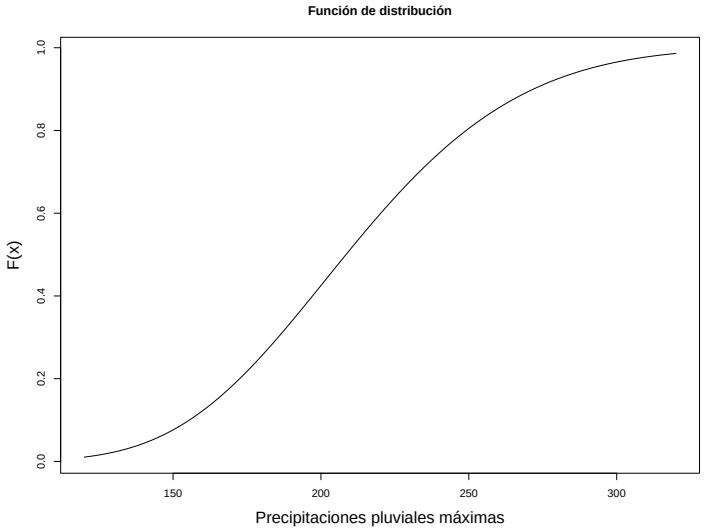

Figura B11: Función de distribución acumulada del modelo Weibull para la estación de Santa María Jacatepec.

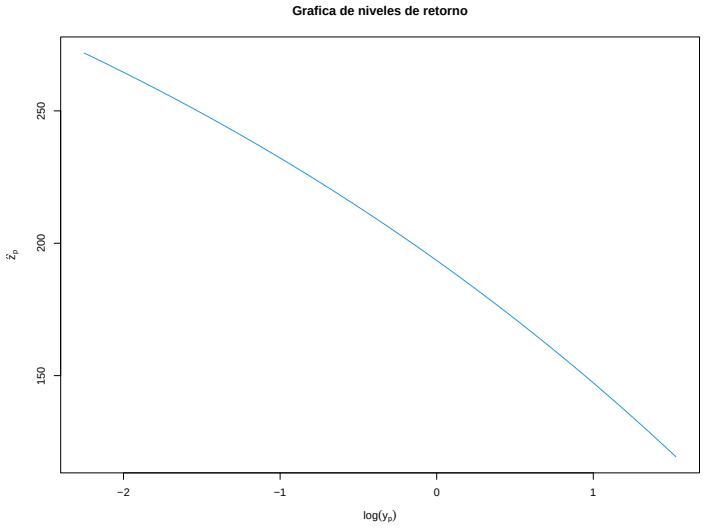

Figura B12: Gráfico de niveles de retorno para la estación de Santa María Jacatepec.

Región de la Costa

La región de la Costa abarca una superficie de 11,605.06 km^2, se subdivide en 50 municipios agrupados en tres distritos: Jamiltepec, Juquila y Pochutla. La región de la Costa representa la tercera concentración de población en el Estado y constituye 14 % de la población total. El clima de esta región es cálido subhúmedo y semicálido húmedo con lluvias en verano. La lluvia anual máxima es de 2,054 y 731.9 *mm* relativamente.

T en años	Nivel de retorno (mm)	$Var(z_p)$	IC del 95 %
5	249.10	189.9	(222.08, 276.11)
10	271.82	263.93	(239.98, 303.67)
20	290.95	414.36	(251.06, 330.85)

Tabla B11: Periodos de retorno y niveles de retorno del modelo Weibull para la estación de Santa María Jacatepec

La estación seleccionada para la región de la Costa es la estación de Cozoaltepec; se encuentra en el municipio de Santa María Tonameca y pertenece al distrito de Pochutla. La Tabla B12 muestra algunas de las estadísticas descriptivas de dicha estación de monitoreo.

Mínimo	Q_1	**Mediana**	**Media**	Q_3	**Máximo**
47.00	79.25	100.75	122.34	142.50	349.20

Tabla B12: Estadísticas descriptivas de la estación de Cozoaltepec.

Las estimaciones de los parámetros de los que depende la función de distribución se encuentran en la Tabla B13, junto con el error estándar y los intervalos de confianza de cada uno de los parámetros estimados.

Parámetro	Estimación	Error estándar	IC del 95 %
$\hat{\mu}$	89.55	9.43	(71.08, 108.12)
$\hat{\sigma}$	35.33	7.78	(20.01, 50.73)
$\hat{\xi}$	0.27	0.20	(-0.11, 0.66)

Tabla B13: Estimaciones de máxima verosimilitud e intervalos de confianza de los parámetros del modelo de la DGVE a la estación de Cozoaltepec.

De las estimaciones, la función de distribución de las precipitaciones máximas es una Fréchet ya que el valor de $\hat{\xi} = 0.27$ es mayor que cero. Además, dado que el valor de $\hat{\xi}$ es mayor que -0.5, los EMV cumplen las condiciones asintóticas usuales, lo que implica que se puedan calcular los intervalos de confianza de cada uno de los parámetros estimados para un $\alpha = 0.05$.

En la Figura B13 se observa que el ajuste del modelo Fréchet a los datos es adecuado, pues el conjunto de puntos trazados queda cerca de la línea de identidad.

En la Tabla B14 se presentan los diferentes niveles de retorno ($\hat{z_p}$) asociados a los períodos de retorno (T) igual a 5, 10 y 20 años. Además, se calculan las varianzas de cada nivel de retorno, $Var(z_p)$, usando el método delta y se presentan los IC del 95 % correspondientes a z_p.

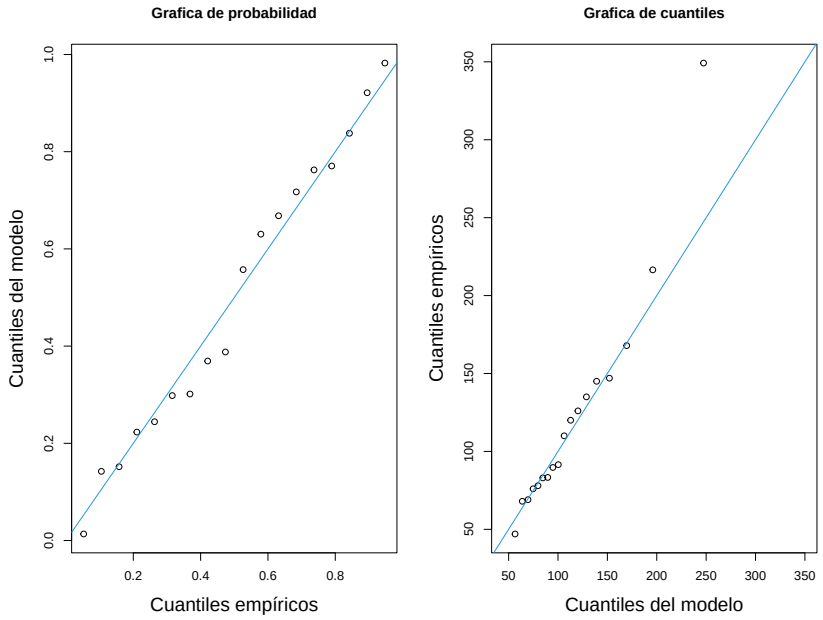

Figura B13: Gráfico de probabilidad y gráfico de cuantiles de la estación de Cozoaltepec.

T en años	Nivel de retorno (mm)	$Var(z_p)$	IC del 95 %
5	155.10	522.42	(110.30, 199.90)
10	199.48	1563.79	(121.97, 276.99)
20	251.54	4486.54	(120.26, 382.83)

Tabla B14: Periodos de retorno y niveles de retorno del modelo Fréchet para la estación de Cozoaltepec

Región de la Mixteca

La región de la Mixteca al noroeste del estado tiene una extensión territorial de 16,333 km^2, comprende siete distritos: Coixtlahuaca, Huajuapan, Juxtlahuaca, Nochixtlán, Silacayoapan, Teposcolula y Tlaxiaco.

Los climas son variados, entre cálido subhúmedo, semicálido subhúmedo y el templado subhúmedo. Las lluvias son desiguales, van de 550 mm a 2,177 mm total anual, predominando la carencia de lluvias. La estación seleccionada para la región de la Mixteca es la estación de Yodocono de Porfirio Díaz; se encuentra en el municipio del mismo nombre y pertenece al distrito de Nochixtlán. La Tabla B15 muestra algunas de las estadísticas descriptivas de dicha estación de monitoreo.

Las estimaciones de los parámetros de los que depende la función de distribución se encuentran en la Tabla B16, junto con el error estándar y los intervalos de confianza de cada uno de los parámetros estimados.

Mínimo	Q_1	**Mediana**	**Media**	Q_3	**Máximo**
30.00	40.10	50.00	53.21	57.88	136.30

Tabla B15: Estadísticas descriptivas de la estación de Yodocono de Porfirio Díaz.

Parámetro	Estimación	Error estándar	IC del 95 %
$\hat{\mu}$	42.60	3.12	(36.48, 48.70)
$\hat{\sigma}$	11.45	2.58	(6.39, 16.49)
$\hat{\xi}$	0.27	0.21	(-0.15,0.69)

Tabla B16: Estimaciones de máxima verosimilitud e intervalos de confianza de los parámetros del modelo de la DGVE a la estación de Yodocono de Porfirio Díaz.

De las estimaciones, la función de distribución de las precipitaciones máximas es una Fréchet ya que el valor de $\hat{\xi} = 0.27$ es mayor que cero. Además, dado que el valor de $\hat{\xi}$ es mayor que -0.5, los EMV cumplen las condiciones asintóticas usuales, lo que implica que se pueden calcular los intervalos de confianza de cada uno de los parámetros estimados para un $\alpha = 0.05$.

En la Figura B14 se observa que el ajuste del modelo Fréchet a los datos es adecuado, pues el conjunto de puntos trazados queda cerca de la línea de identidad.

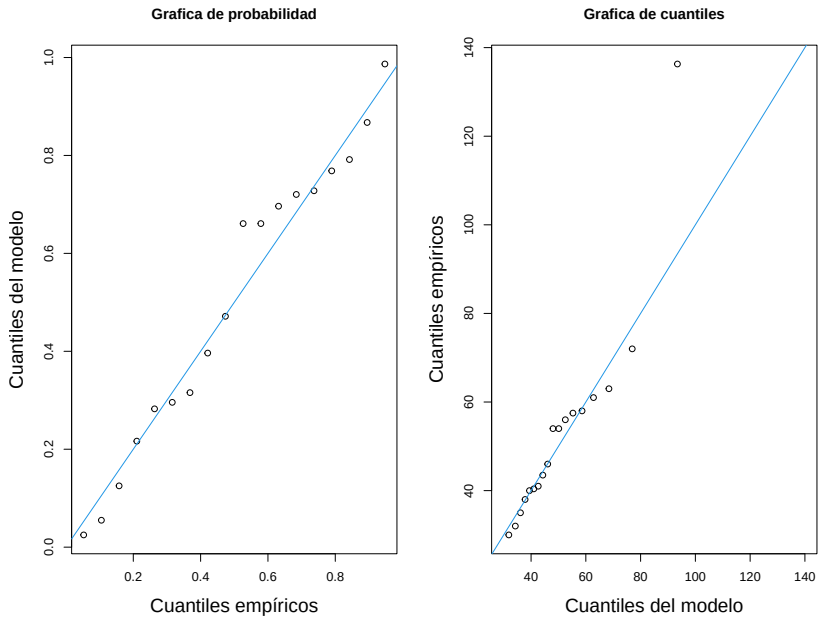

Figura B14: Gráfico de probabilidad y gráfico de cuantiles de la estación de Yodocono de Porfirio Díaz.

En la Tabla B17 se presentan los diferentes niveles de retorno ($\hat{z_p}$) asociados a los períodos de retorno

(T) igual a 5, 10 y 20 años. Además, se calculan las varianzas de cada nivel de retorno, $Var(z_p)$, usando el método delta y se presentan los IC del 95 % correspondientes a z_p.

T en años	Nivel de retorno (mm)	$Var(z_p)$	IC del 95 %
5	63.78	54.19	(49.35, 78.21)
10	78.07	163.59	(121.97, 276.99)
20	251.54	4486.54	(53.00, 103.14)

Tabla B17: Periodos de retorno y niveles de retorno del modelo Fréchet para la estación de Yodocono de Porfirio Díaz