



**UNIVERSIDAD TECNOLÓGICA DE LA MIXTECA**

**TEORÍA DE VALORES EXTREMOS APLICADO  
A ESTIMACIÓN DE RIESGOS POR OZONO ( $O_3$ )  
EN CDMX**

**T E S I S**

**PARA OBTENER EL TÍTULO DE:  
LICENCIADO EN MATEMÁTICAS APLICADAS**

**PRESENTA:  
ROMÁN EMMANUEL HERNÁNDEZ CARRILLO**

**DIRECTORA DE TESIS:  
M.C. ANA DELIA OLVERA CERVANTES**

*H. CD. HUAJUAPAN DE LEÓN, OAXACA. SEPTIEMBRE 2025*



*Dedicado a  
mis padres y hermanos  
por su amor incondicional.*

*Y a los estudiantes que  
experimentan diversas condiciones  
de salud mental: **no se rindan.***



# Agradecimientos

A **mi madre**, por su amor incondicional y su resiliencia inquebrantable, que ha sido el faro que ha iluminado hasta mis momentos más oscuros. Su valentía para reponerse de las adversidades del pasado y su estricta determinación para brindarnos un futuro mejor sin importar el costo personal, constituyen el mejor legado que puedo recibir. Hoy y siempre serás mi principal motivación para seguir viviendo. Te amo, mamá.

A **mi padre**, por compartirme su sabiduría derivada de sus grandes experiencias. Sus consejos me orientaron a encontrar mi rumbo y su respeto por mi autonomía me enseñaron a tomar mis propias decisiones y aprender de mis errores. Gracias papá, por tu entera confianza, por tu constante apoyo en todo mi proceso escolar, pero sobre todo gracias por cuidar de ti mismo, por buscarme y quererme. Este logro es de los dos.

A **mis queridos hermanos**, pilares fundamentales en mi vida a quienes debo gratitud:

A **Daesy**, por su generoso corazón y su profunda empatía. Su capacidad para mostrarse amable y comprensiva, incluso cuando enfrenta sus propias batallas, es prueba de su gran fortaleza. Saber que puedo contar con mi hermana mayor de manera incondicional es algo que valoro muchísimo.

A **Nancy**, a quien admiro por su inteligencia, su carácter firme y por lo trabajadora que siempre ha sido, desde niño me haz inspirado para esforzarme en mis estudios y para mejorar constantemente. Además, tu entrega abnegada hacia la familia me enseñó el verdadero significado de la gratitud y de la solidaridad para los nuestros.

A **Sehily**, con quien compartí los momentos más divertidos en nuestras “noches de juegos”, contigo aprendí que con concentración, dedicación y persistencia, se pueden superar todos los retos, incluso pasar todos los niveles del Mario Bros en una sola noche. En fin, gracias manita por abrirme las puertas de tu casa y haber sido parte de los recuerdos más felices que tengo; me mostraste que a pesar de tener heridas internas del pasado que nos pueden volver bastante sensibles y vulnerables, siempre podemos reírnos de la vida con ese humor negro tan característico de tu hogar.

---

A **Víctor**, mi hermano mayor. Aunque a veces te ocultas bajo una máscara de seriedad, una capa de apatía y un estandarte de poca paciencia, yo sé que en el fondo late el corazón de un hombre muy sincero, profundamente respetuoso y amoroso. Hoy veo en ti a un gran padre que haría cualquier cosa por su familia, además, eres un protector nato de los animales y eso significa que tienes un corazón muy noble. Contigo he aprendido sobre la complejidad del carácter humano y ahora puedo entender la necesidad de construir varias capas para protegernos de lo que nos hace más vulnerables.

A **Cindy**, siempre optimista, alegre y espontánea. Contigo sigo aprendiendo que la vida hay que disfrutarla como si fuera un gran escenario, dándolo todo sin importar el qué dirán y divirtiéndose como si nadie más existiera. Admiro mucho tu valentía al mostrarte tan auténtica; de chico me tenía que esconder para poder bailar como tú, pero hoy en día con el apoyo de todos, puedo bailar a la par contigo y cantar todo el concierto de RBD si nos da la gana. Gracias por enseñarme a ser más valiente.

A **David**, mi hermano menor, a quien procuro darle un buen ejemplo demostrándole que, sin importar los fantasmas de nuestro pasado y sin importar los problemas que continuamente se presentan, siempre podemos hacer la diferencia y encontrar nuestro propio camino. No debemos estancarnos en el pasado, lo que realmente importa es lo que decidimos hacer para mejorar nuestro futuro.

A **Jenny**, mi hermana mayor, gracias por cuidarme desde el cielo y por darme las fuerzas para seguir apoyando a la familia, estarías muy orgullosa de saber todo lo que he logrado. Te extraño mucho hermanita, siempre te llevaré en mi corazón.

A **mis amigos y compañeros** que hicieron de este proceso una etapa inolvidable.

A **Ingrid** y **Litzy**, por estar a mi lado durante estos cinco años de carrera. Compartir con ustedes tanto los momentos más divertidos como los más difíciles hizo que la estadía aquí fuera más amena y que todo valiera la pena, pues siempre nos acompañamos mutuamente en las buenas y en las malas.

A **Luz**, alias “Miss Pacheco”, por saber escuchar cuando sentía que nadie más lo hacía y quien impidió que me sumergiera en un abismo mental que parecía no tener salida. Gracias por tu comprensión, por tus buenos consejos y por acompañarme cuando más lo necesitaba. Tu amistad es un regalo invaluable en mi vida.

A **Jonas** y **Gil**, mis compas los heteros, siempre voy a atesorar su amistad y los momentos que pasamos juntos. Y a mis compañeros de generación: **Hael**, **Mendiola**, **Ray** y **Yuli**, su compañía en el aula hacía que las clases fueran más llevaderas.

---

A **mi pareja**, a quien admiro profundamente por su dedicación, su espontaneidad, su carisma y su seguridad. Verte cada día es una motivación para crecer y convertirme en una mejor versión de mí mismo. Gracias por cuidar de mi, por tu cariño durante estos años y por toda la paciencia que me brindaste durante la realización de este trabajo.

A **mis tutores y profesores** que me guiaron durante todo mi proceso.

A la **Dra. Ana Delia**, quién me recordó la razón por la que escogí estudiar la carrera de matemáticas y fue quién me inspiró para escoger el área de estadística. Agradezco profundamente su invaluable apoyo, su gran dedicación y la paciencia infinita que tuvo durante el desarrollo de esta investigación. Pero, sobre todo, le agradezco por estar siempre pendiente de mí, por escucharme cuando más preocupado me veía y por depositar su confianza en mí y en este proyecto. Más que mi directora y mi profesora, veo en usted a una mentora, una consejera y a una gran amiga.

A mis tutores **Octavio** y **Tenorio**, quienes siempre me hicieron sentir bienvenido en sus respectivos cubículos. Nuestras conversaciones sobre mi situación académica siempre resultaron reconfortantes, sus palabras de aliento y validación me animaban a seguir adelante y me tranquilizaba escuchar que todo estaría bien.

A todos los **sinodales** por su valioso tiempo que le dedicaron a la revisión de este trabajo y por sus acertadas contribuciones para mejorarlo. Así mismo, extendiendo mi agradecimiento a todos los **profesores** que contribuyeron a mi formación académica a lo largo de la licenciatura, y a todos mis **compañeros** de otras generaciones con quienes compartí muchos momentos agradables dentro y fuera de las áreas de estudio.

A la **Universidad Tecnológica de la Mixteca**, y en especial al **Instituto de Física y Matemáticas**, por brindarme una educación de calidad durante estos cinco años y por proveer las herramientas necesarias para mi desarrollo profesional y personal.

Finalmente, quiero dar gracias a la vida por poner en mi camino a las personas correctas; quienes me impulsaron a no rendirme aún cuando sentía que no podía dar ni un paso más, personas que me enseñaron a disfrutar del proceso en lugar de sufrirlo, y a mis amigos que me acompañaron y me apoyaron cuando perdí por completo mi motivación y las razones para seguir. Gracias a ustedes, se logró.



# Prefacio

La Teoría de Valores Extremos (TVE) tiene sus orígenes en el trabajo pionero de Nicolás Bernoulli en 1709, quien abordó el problema de la esperanza de vida del último superviviente entre un grupo de personas, calculando el máximo de variables aleatorias independientes. De acuerdo a lo mencionado por Emil Gumbel (1958), Bernoulli fue históricamente el primer personaje en introducir las primeras nociones sobre la TVE.

Posteriormente, fue en la década de 1920 cuando la TVE adquirió relevancia formal, gracias a contribuciones clave de Ronald Fisher y Leonard Tippett, con la colaboración de otros autores. Fisher demostró que los valores extremos (máximos y mínimos) siguen distribuciones predecibles, mientras que Tippett, estudiando la resistencia de fibras de algodón, observó la importancia de los eventos extremos.

En 1927, Maurice Fréchet introdujo la distribución que lleva su nombre, y al año siguiente, Fisher y Tippett establecieron el teorema que caracteriza las tres familias de distribuciones límite para máximos, conocido como teorema de Fisher-Tippett [Fisher y Tippett (1928)]. Boris Gnedenko (1943) fundamentó rigurosamente esta teoría, dando lugar al teorema de Fisher-Tippett-Gnedenko.

En los años 50, Emil Gumbel y Waloddi Weibull completaron el marco teórico con sus respectivas distribuciones, destacando el libro *Statistics of Extremes* (Gumbel, 1958) como referencia clave. En los 70, el teorema de Pickands-Balkema-de Haan permitió modelar excesos sobre umbrales mediante la distribución generalizada de Pareto. Durante las décadas de 1980 y 1990, se desarrollaron técnicas de estimación y aplicaciones en procesos multidimensionales. Al día de hoy, han surgido un gran número de publicaciones y la TVE comenzó a difundirse ampliamente en diversos ámbitos científicos.

La teoría de valores extremos es una rama de la estadística que se encuentra en continuo crecimiento, pues ha tenido un desarrollo considerable en las últimas décadas. Se encarga de modelar el comportamiento de eventos asociados a los valores máximos y mínimos de una variable aleatoria. La TVE tiene un amplio campo de aplicación, especialmente en ciencias ambientales, en donde es necesario conocer los cambios extremos de la concentración de contaminantes, precipitaciones pluviales, temperatura, etc.

---

En este trabajo se aplica la teoría de valores extremos considerando como variable aleatoria a la concentración de partículas de ozono ( $O_3$ ) que hay en el aire, registradas por algunas estaciones de monitoreo atmosférico (seleccionadas estratégicamente) ubicadas en distintas zonas de la Ciudad de México.

Se pretende realizar un análisis estadístico para explicar y visualizar el comportamiento de los datos sobre la concentración del contaminante  $O_3$  que se disponen en los registros horarios recabados por la Dirección de Monitoreo Atmosférico de la Ciudad de México en [SEDEMA-SIMAT].

Partiendo de la hipótesis de que la Distribución de Valores Extremos Generalizada (DVEG) se ajusta adecuadamente a las concentraciones máximas de partículas de ozono en la atmósfera de la Ciudad de México, siguiendo el enfoque para el caso univariado; el objetivo de este trabajo es seleccionar el mejor submodelo de la DVEG, realizando estimación de parámetros, calculando intervalos de confianza y aplicando métodos de validación de modelos tanto teóricos como gráficos.

Los métodos gráficos de diagnóstico, los cálculos de funciones, las estimaciones y las simulaciones se realizaron con el lenguaje de programación R. Por otra parte, para la manipulación de los datos se utilizaron herramientas computacionales de **Excel**.

# Índice general

<b>Introducción</b>	<b>XIII</b>
<b>1. Preliminares</b>	<b>1</b>
1.1. Bases de probabilidad . . . . .	1
1.1.1. Variables aleatorias y su distribución . . . . .	2
1.1.2. Vectores aleatorios y generalizaciones . . . . .	5
1.1.3. Independencia y tipos de convergencia . . . . .	6
1.2. Conceptos de estadística . . . . .	10
1.2.1. Estadísticos descriptivos . . . . .	10
1.2.2. Esperanza, varianza, asimetría y curtosis . . . . .	13
1.2.3. Vector de medias, covarianza y correlación . . . . .	18
<b>2. Fundamentos de estimación paramétrica</b>	<b>25</b>
2.1. Principios de estimación . . . . .	25
2.1.1. Estadísticas de orden . . . . .	27
2.1.2. Nivel de retorno . . . . .	30
2.2. Método de máxima verosimilitud . . . . .	31
2.2.1. Estimador de máxima verosimilitud (EMV) . . . . .	31
2.2.2. Propiedades del EMV y condiciones de regularidad . . . . .	33
2.3. Teoría univariada de verosimilitud . . . . .	35
2.3.1. Probabilidad de cobertura . . . . .	37
2.3.2. Normalidad asintótica del EMV . . . . .	38
2.3.3. Razón de verosimilitud . . . . .	40
2.4. Pruebas de hipótesis . . . . .	42
2.4.1. Variables aleatorias estandarizadas . . . . .	43
2.4.2. Metodología para una prueba de hipótesis . . . . .	44

<b>3. Teoría de valores extremos univariante</b>	<b>49</b>
3.1. Funciones asintóticamente equivalentes . . . . .	49
3.1.1. Sucesiones asintóticamente equivalentes . . . . .	52
3.2. Formulación del modelo para máximos . . . . .	53
3.2.1. Aproximación mediante distribuciones límite . . . . .	53
3.3. Distribuciones límite para valores extremos . . . . .	54
3.3.1. Teorema de valores extremos . . . . .	55
3.4. Distribución de valores extremos generalizado . . . . .	58
3.4.1. Teorema de valores extremos generalizado . . . . .	59
3.4.2. Max-estabilidad . . . . .	61
<b>4. Métodos de estimación paramétrica de la DVEG</b>	<b>63</b>
4.1. Método de máximos por bloques . . . . .	65
4.2. Estimación por máxima verosimilitud . . . . .	66
4.3. Inferencia para los niveles de retorno . . . . .	69
4.4. Métodos gráficos de diagnóstico . . . . .	71
4.4.1. Gráfica de probabilidad (P-P) . . . . .	71
4.4.2. Gráfica cuantil (Q-Q) . . . . .	72
4.4.3. Gráfica de nivel de retorno . . . . .	73
<b>5. Aplicación: Concentraciones máximas de ozono</b>	<b>75</b>
5.1. Descripción del problema . . . . .	75
5.2. Descripción y manipulación de los datos . . . . .	77
5.3. Análisis de los datos por estación . . . . .	80
<b>Conclusiones</b>	<b>113</b>
<b>Bibliografía</b>	<b>115</b>
<b>A. Código en el software R</b>	<b>119</b>

# Índice de figuras

1.	Formación de ozono y sus principales efectos en salud. . . . .	XIV
2.	Cambios en los límites de la NOM correspondiente a $O_3$ . . . . .	XVI
1.1.	Comparación de distribuciones simétricas y asimétricas. . . . .	17
1.2.	Comparación de distribuciones por curtosis. . . . .	18
5.1.	Distribución de las estaciones de monitoreo del SIMAT. . . . .	78
5.2.	Caseta de monitoreo de Cuajimalpa (ID: 484090040109) . . . . .	80
5.3.	Gráfica de dispersión y Box-plot: Estación CUA . . . . .	84
5.4.	Gráficas de diagnóstico P-P y Q-Q del modelo Weibull: Estación CUA . . . . .	86
5.5.	Gráfica de niveles de retorno y densidad ajustada (GEVD) . . . . .	87
5.6.	Funciones de distribución acumulada y densidad de probabilidad (GEVD) . . . . .	89
5.7.	Niveles de retorno máximos por día (Weibull): Estación CUA . . . . .	90
5.8.	Caseta de monitoreo de Benito Juárez (ID:484090140309) . . . . .	91
5.9.	Gráfica de dispersión y Box-plot: Estación BJU . . . . .	94
5.10.	Gráficas de diagnóstico P-P y Q-Q del modelo Weibull: Estación BJU . . . . .	95
5.11.	Gráfica de niveles de retorno y densidad ajustada (GEVD) . . . . .	96
5.12.	Funciones de distribución acumulada y densidad de probabilidad (GEVD) . . . . .	97
5.13.	Niveles de retorno máximos por día (Weibull): Estación BJU . . . . .	99
5.14.	Caseta de monitoreo de La Presa (ID:484151040203) . . . . .	99
5.15.	Histograma de frecuencias de los registros horarios: Estación LPR . . . . .	101
5.16.	Gráfica de dispersión y Boxplot: Estación LPR . . . . .	103
5.17.	Gráficas de diagnóstico P-P y Q-Q del modelo Fréchet: Estación LPR . . . . .	106
5.18.	Gráfica de niveles de retorno y densidad ajustada (GEVD) . . . . .	107
5.19.	Funciones de distribución acumulada y densidad de probabilidad (GEVD) . . . . .	108
5.20.	Niveles de retorno máximos por día (Fréchet): Estación LPR . . . . .	109

# Lista de tablas

1.	Escala IMECA establecido por la norma NADF-009-AIRE-2017 . . . .	XVII
2.	Indicadores para aplicar la NADF-009-AIRE-2017. . . . .	XVIII
3.	Escala ppb y $\mu g/m^3$ para $O_3$ (NADF-009-AIRE-2017) . . . . .	XVIII
1.1.	Promedios mensuales de concentraciones de ozono (ppb) del año 2023. .	20
1.2.	Interpretación del coeficiente de correlación de Pearson. . . . .	22
2.1.	Errores Tipo I y Tipo II en pruebas de hipótesis . . . . .	47
5.1.	Información geográfica de las estaciones atmosféricas en estudio. . . . .	80
5.2.	Estadísticas descriptivas generales: Estación CUA . . . . .	82
5.3.	Concentraciones máximas por día (bloques 1-59): Estación CUA . . . .	82
5.4.	Concentraciones máximas por día (bloques 60-120): Estación CUA . . .	83
5.5.	Concentraciones máximas por mes (bloques 121-365): Estación CUA . .	83
5.6.	Descriptivos de los máximos por día: CUA . . . . .	83
5.7.	Periodos y niveles de retorno del modelo Weibull: Estación CUA. . . .	90
5.8.	Estadísticas descriptivas generales: Estación BJU . . . . .	92
5.9.	Intervalos de bloques con registros máximos mensuales: Estación BJU .	93
5.10.	Descriptivos de los máximos por día: BJU . . . . .	93
5.11.	Periodos y niveles de retorno del modelo Weibull: Estación BJU. . . . .	98
5.12.	Estadísticas descriptivas generales: Estación LPR . . . . .	101
5.13.	Intervalos de bloques con registros máximos mensuales: Estación LPR .	102
5.14.	Descriptivos de los máximos por día: LPR . . . . .	102
5.15.	Periodos y niveles de retorno del modelo Fréchet: Estación LPR . . . .	109
5.16.	Programa de contingencias ambientales atmosféricas vigente 2022 . . . .	111
5.17.	Cronología de las contingencias atmosféricas durante 2023 . . . . .	111

# Introducción

En contextos ambientales, la teoría de valores extremos surge como respuesta a la necesidad de cuantificar riesgos asociados a fenómenos meteorológicos y contaminación atmosférica, donde los eventos extremos, aunque poco frecuentes, pueden tener consecuencias significativas en la salud pública y los ecosistemas.

La Secretaría del Medio Ambiente de la Ciudad de México (SEDEMA), afirma que el efecto negativo de la contaminación en la salud humana es el aspecto más preocupante para las autoridades responsables de la salud pública y la protección del medio ambiente. La calidad del aire es un tema de interés general y cada vez se tiene más información respecto a las variables que influyen, como la radiación solar, temperatura, los patrones del viento, las características geográficas, la temporalidad estacional, así como la cantidad de emisiones.

El gobierno federal es el responsable de describir los estándares para la protección de la salud pública y vigilar su cumplimiento. Las Normas Oficiales Mexicanas (NOM) de Salud Ambiental establecen los límites permisibles para la concentración de siete contaminantes: Ozono ( $O_3$ ), Monóxido de Carbono (CO), Dióxido de Nitrógeno ( $NO_2$ ), Dióxido de Azufre ( $SO_2$ ) y Partículas ( $PM_{10}$  y  $PM_{2.5}$ ) menores a 10 y  $2.5 \mu m$  (micrómetros). [SEDEMA (NOM)].

De acuerdo al informe del 2020 de la Calidad del aire en la Ciudad de México, SEDEMA (2023a), el ozono es un componente natural de la atmósfera que se presenta en bajas concentraciones y es indispensable para la vida, el  $O_3$  es una molécula formada por tres átomos de oxígeno. Además, es un poderoso oxidante que reacciona rápidamente con otros compuestos químicos y es inestable en altas concentraciones. Cerca del 90 % del ozono atmosférico se concentra en la capa de ozono que está ubicada en la estratósfera, entre 15 y 30 km de altitud. La capa de ozono absorbe gran parte de los rayos solares ultravioletas UV-A, UV-B y UV-C, los cuales ocasionan efectos adversos en la salud humana: la radiación UV-A causa envejecimiento prematuro de la piel, cataratas y daños en el sistema inmunológico; y la exposición prolongada a los rayos UV-B aumenta el riesgo de contraer cáncer de piel. [Fahey (2002)].

## INTRODUCCIÓN

El ozono troposférico se encuentra en la capa más cercana a la superficie terrestre y se genera a partir de reacciones fotoquímicas de la radiación solar con la combinación de óxidos de nitrógeno ( $\text{NO}_x$ ) y Compuestos Orgánicos Volátiles (COV), como se observa en la Figura 1 (SEDEMA (2023b)). En una publicación del Dr. ?mSánchez, para la Facultad de Farmacia de la UCM<sup>1</sup>, comenta que los COV son producidos por la combustión de gasolina, madera y gas natural con puntos de ebullición que oscilan entre los 50°C y 260°C; también son liberados por disolventes, pinturas, pegamentos y otros productos empleados y almacenados en los hogares y centros de trabajo.

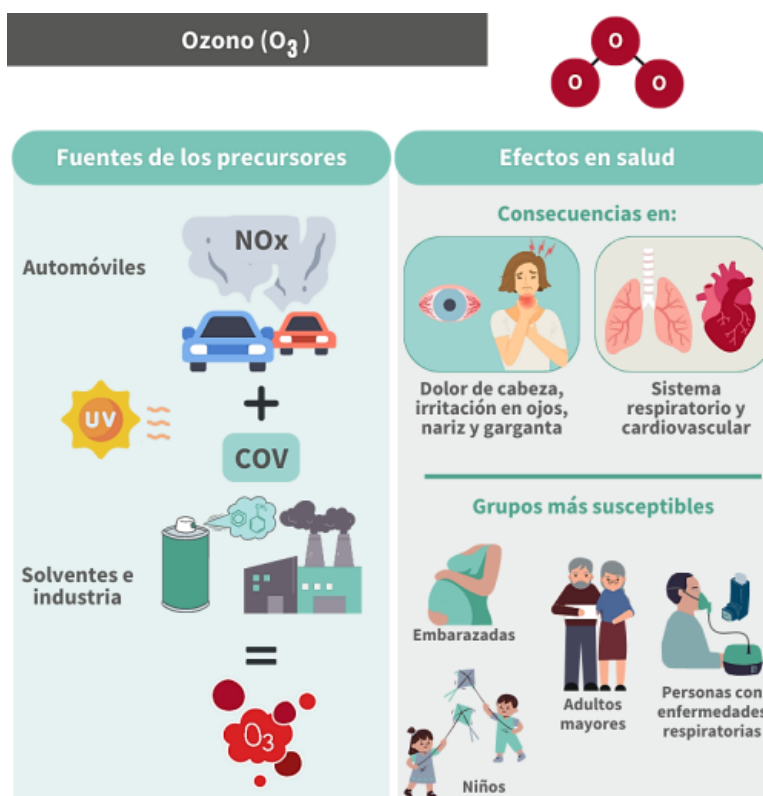


Figura 1: Formación de ozono y sus principales efectos en salud.

La exposición a este contaminante afecta gravemente los pulmones, provocando severas enfermedades respiratorias, dolor de cabeza e irritación en nariz, ojos y garganta. Una exposición prolongada y constante puede aumentar el riesgo de muerte prematura por enfermedades respiratorias en los grupos más susceptibles, los cuales son adultos mayores, niños, mujeres embarazadas y personas asmáticas.

En 1992 y 1993, estudios en niños de edad preescolar mostraron asociación entre la concentración de ozono superior a 0.120 ppm (partes por millón) y el aumento del ausentismo escolar [Romieu et al. (1992)].

<sup>1</sup>Universidad Complutense de Madrid

Una investigación efectuada a escolares de una zona residencial del suroeste de la Ciudad de México, corroboró que la concentración elevada de partículas suspendidas de ozono está asociada con el agravamiento de síntomas de vías respiratorias inferiores y con la disminución del flujo espiratorio máximo [Gold et al. (1999) ].

En 1996, se efectuó un estudio en niños asmáticos de 5 a 7 años de edad, en quienes se identificó una relación entre las concentraciones de  $O_3$  y de partículas  $PM_{10}$  con la presencia de síntomas respiratorios y la incidencia de enfermedades de vías respiratorias bajas [Meneses et al. (1996)].

Ya que las principales fuentes antropogénicas que aumentan la formación de ozono troposférico son los solventes químicos y las emisiones vehiculares e industriales; las grandes ciudades con áreas conurbadas son responsables de gran parte de la contaminación atmosférica debido a su sobre población, por el excesivo consumo de combustibles fósiles y por la intensa actividad industrial. En las grandes ciudades del mundo, como la ZMCM (Zona Metropolitana de la Ciudad de México), constantemente se registran valores que superan los estándares establecidos a nivel mundial.

En el 2023, la Encuesta Nacional de la Dinámica Demográfica (ENADID-INEGI) estimó 129.5 millones de mexicanos de los cuales el Estado de México fue la entidad más poblada con 16.992 millones y la Ciudad de México la segunda con 9.210 millones de habitantes. Esta alta densidad de habitantes en la ZMCM, conlleva a mayores emisiones de contaminantes y, en consecuencia, más presión sobre la calidad del aire. Además de la alta población y las fuentes antropogénicas, hay otros factores meteorológicos que contribuyen a la dinámica de la concentración atmosférica del  $O_3$ .

La Secretaría de Salud (1996), indicó que el Valle de México posee características fisiográficas y climáticas únicas que contribuyen de manera determinante en la severidad de los problemas de contaminación de la Ciudad. La ZMCM es una cuenca hidrológica situada a 2,240 m de altura sobre el nivel del mar en su parte central, tiene una extensión territorial de 1,200  $km^2$  y está rodeada por montañas que tienen una altura promedio de 1,000 m sobre la parte central. Por esta razón la concentración de oxígeno está disminuida en un 23% con relación al nivel del mar, aumentando la concentración de monóxido de carbono e hidrocarburos.

El Dr. Sánchez menciona que el *smog*<sup>2</sup> es originado a partir de contaminantes durante un largo período de altas presiones que provoca el estancamiento del aire y, por tanto, su permanencia en las capas más bajas de la atmósfera. El smog cubre por completo la CDMX y está en contacto directo con toda la población, siendo el ozono troposférico su principal componente.

---

<sup>2</sup>Neblina densa y visible compuesta por contaminantes químicos y partículas suspendidas.

## INTRODUCCIÓN

El crecimiento de la población en la Ciudad de México ha ejercido una acción desfavorable sobre su medio ambiente; la contaminación del aire en la CDMX ha alcanzado dimensiones tan críticas que hoy en día se ubica en el quinto lugar de las urbes más contaminadas del mundo, ubicándose por debajo de Tokio, Delhi, Shanghái y Sau Paulo.

El monitoreo atmosférico es una herramienta indispensable para determinar la calidad del aire en una ciudad y provee de información relevante para los tomadores de decisiones y la población, sobre la concentración de los principales contaminantes en el aire que afectan la salud. De acuerdo al SEDEMA (2023a), el monitoreo está ligado a la normatividad vigente de salud, como ya se mencionó, las NoM establecen los límites permisibles, conocidos como estándares de calidad del aire y se han establecido para no ser superados en un tiempo y área determinada.

La NOM de ozono ha tenido varios cambios desde su primera versión de 1994 en la que sólo se consideraba el límite horario de 110 ppb (Figura 2) [SEDEMA (2024)]; posteriormente, en 2002, se agregó la métrica del promedio móvil de 8h en 80 ppb. La siguiente revisión de la NOM fue en 2014, en la cual se disminuyeron ambos límites: el promedio horario se estableció en 95 ppb (reducción del 14% con respecto al valor anterior) y el promedio móvil de 8h en 70 ppb (reducción del 12%). Finalmente, en 2021 el límite horario se estableció en 90 ppb.

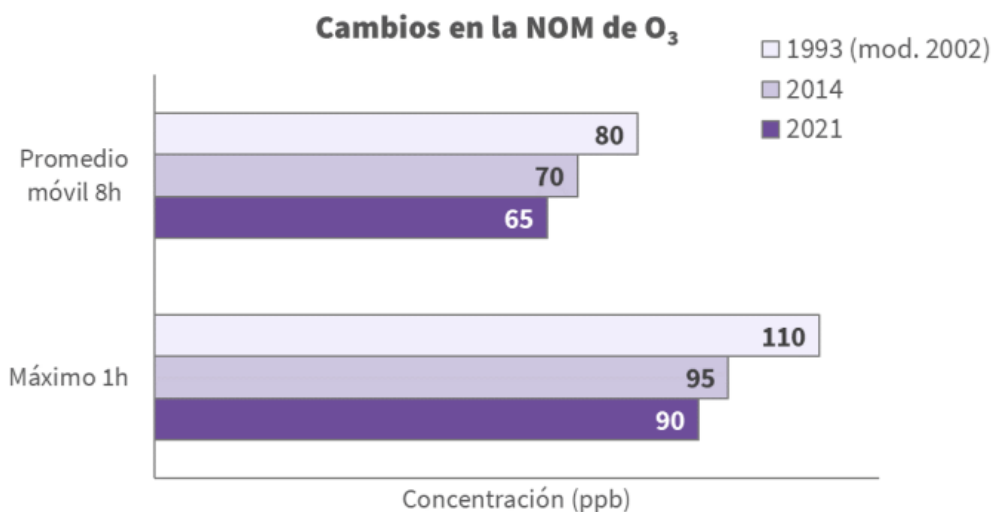


Figura 2: Cambios en los límites de la NOM correspondiente a O<sub>3</sub>.

La concentración de ozono es una problemática constante en la ZMCM y aunque los niveles del contaminante han disminuido considerablemente desde la década de los noventa, todavía se superan los límites de la NOM que son cada vez más estrictos. En diversos estudios se ha buscado un umbral “seguro” de ozono, en el cual los efectos a la salud sean mínimos o nulos, pero existe poca evidencia sobre esto.

La información de la concentración de partículas de  $O_3$  (y otros contaminantes) para obtener el Índice de Calidad del Aire (ICA), proviene de las Estaciones de Monitoreo que cumplen con los criterios de representación física y espacial, y se atienen a los objetivos de monitoreo del SIMAT (SEDEMA-SIMAT (2018)), destinadas a informar y prevenir a la población sobre los niveles de contaminación en la zona representativa de la Ciudad de México y extenderse a los municipios conurbados. El ICA se calcula para cada uno de los contaminantes reportados en las estaciones de monitoreo.

La Dirección de Monitoreo Atmosférico de la Ciudad de México reporta los registros de concentración de ozono cada hora, todos los días del año y se pueden descargar en la página oficial del SIMAT: [http://www.aire.cdmx.gob.mx/estadisticas-consultas/consultas/download\\_imeca.php](http://www.aire.cdmx.gob.mx/estadisticas-consultas/consultas/download_imeca.php) [Datos con derechos reservados].

Es importante aclarar que la base de datos que se descargaron para los objetivos de esta tesis corresponden a las mediciones realizadas en el 2023 y registran el índice de calidad del aire (horario) por sitio de monitoreo de acuerdo a la norma ambiental NADF-009-AIRE-2017 publicada el 14 de noviembre del 2018.

Esta norma atmosférica establece el valor ICA, anteriormente IMECA (Índice Metropolitano de la Calidad del Aire), que es un indicador adimensional utilizado en la ZMCM para medir y difundir de manera más sencilla el nivel de la calidad del aire y los posibles riesgos en la salud como muestra la Tabla 1.

Tabla 1: Escala IMECA establecido por la norma NADF-009-AIRE-2017

IMECA	Calidad del aire	Riesgos y Recomendaciones
0-50	Buena	Sin riesgo para la población.
51-100	Regular	Grupos susceptibles pueden presentar síntomas: se debe limitar actividad al aire libre.
101-150	Mala	Grupos susceptibles presentan efectos en la salud: Evitar exponerse, suspender actividades.
151-200	Muy Mala	La población en general puede experimentar efectos adversos: Contingencia (Fase 1).
>200	Extrem. Mala	Riesgo alto para la población; Emergencia (Fase 2).

El Índice de Calidad del Aire de la CDMX publicado en la NADF-009-AIRE-2017 usa como referencia los límites de las NOM de salud anteriores, para el  $O_3$  se usa la NOM-020-SSA1-2014; y no considera las actualizaciones de dicha normatividad realizadas entre 2019 y 2021. Este índice se conserva con fines comparativos con el IAS (Índice de Aire y Salud). [SEDEMA (NOM)].

## INTRODUCCIÓN

---

La Norma Oficial Mexicana NOM-020-SSA1-2014 establece los valores límite permisibles para la concentración de ozono ( $O_3$ ) en el aire ambiente y los criterios para su evaluación, publicada en el Diario Oficial de la Federación el 19 de agosto de 2014; su aplicación a la norma ambiental NADF-2017 se muestra en la Tabla 2 como lo establece la SEDEMA-SIMAT (2018).

Tabla 2: Indicadores para aplicar la NADF-009-AIRE-2017.

Valores límite permisibles	Forma de integración al Índice de Calidad del Aire
0.095 ppm, promedio horario	Promedio horario al punto de corte 100
0.070 ppm, máximo anual del promedio móvil de 8 horas	Valor considerado como promedio horario referido al punto de corte 50

La concentración de las partículas de ozono se mide en ppm (partículas por millón) y existe una relación con la unidad de concentración  $\mu g/m^3$  (microgramos por metro cúbico) usada por otros contaminantes; aunque es poco común usar esta medida para la concentración de ozono.

Estas unidades de concentración se relacionan mediante la fórmula:

$$\mu g/m^3 = \text{ppm} \times \frac{\text{Masa molar del gas (g/mol)}}{24.45},$$

para el caso particular del  $O_3$ , su masa molar es de 48 g/mol, por lo tanto:

$$\mu g/m^3 = \text{ppm} \times (48/24.45).$$

Finalmente, aunque la concentración de ozono se mide usualmente en ppm, la base de datos que se descargó usa las unidades ppb (partículas por billón); y esa será la unidad de medida empleada en el desarrollo de esta tesis. Para realizar esta conversión, simplemente se usa la equivalencia:  $\text{ppb} = \text{ppm} \times 1000$ .

Tabla 3: Escala ppb y  $\mu g/m^3$  para  $O_3$  (NADF-009-AIRE-2017)

IMECA	$O_3$ (ppb)	$\mu g/m^3$	Calidad del aire
0-50	0-70	0-0.137	Buena.
51-100	71-95	0.138-0.188	Regular.
101-150	96-154	0.189-0.303	Mala.
151-200	155-204	0.304-0.400	Muy mala.
>200	>205	>0.401	Extremadamente mala.

En la Tabla 3 se muestran los resultados de estas equivalencias y su calidad del aire.

La estimación de riesgos ambientales involucra la identificación de umbrales críticos de contaminantes, la evaluación de su frecuencia y la cuantificación de impactos potenciales. En particular, para el caso del ozono troposférico, la estimación precisa de riesgos requiere métodos robustos capaces de caracterizar adecuadamente las colas de distribución.

La TVE ofrece ventajas significativas al proporcionar un marco teórico para modelar valores máximos y superaciones de umbrales. El interés principal está en los eventos asociados a la cola de la distribución de las concentraciones de partículas de ozono registradas. Un enfoque para la modelación de valores extremos es a partir de la Distribución de Valores Extremos Generalizada, que se ajusta a los extremos máximos.

La presencia de ozono ( $O_3$ ) en la atmósfera de ciudades grandes presenta una dinámica temporal volátil, donde sus concentraciones varían significativamente a lo largo del día y del año en función de procesos fotoquímicos y condiciones meteorológicas de la zona. Comprender estos patrones temporales es fundamental para determinar los períodos del día de mayor riesgo ambiental y exposición poblacional.

El análisis de los registros horarios revela cómo la interacción entre emisiones precursoras, como la radiación solar y la presencia de gases compuestos, genera picos característicos en ciertos momentos del día, mientras que el análisis estacional muestra la influencia de factores climáticos en la formación y acumulación de este contaminante. Además, el estudio por puntos de monitoreo revela cómo factores como el tránsito vehicular, la presencia de áreas verdes y las características del terreno influyen en la distribución desigual del ozono troposférico en la ciudad.



# Capítulo 1

## Preliminares

El estudio riguroso de la Teoría de Valores Extremos (TVE) y su aplicación al análisis de riesgos ambientales requiere una sólida fundamentación en conceptos probabilísticos y estadísticos. Este capítulo presenta los elementos esenciales que sustentan el análisis de eventos extremos en series de contaminantes atmosféricos. Particularmente, nos enfocamos en aquellos conceptos de utilidad para la modelación estadística de valores máximos de concentración de ozono en la Ciudad de México. Este trabajo de tesis presupone un conocimiento básico de probabilidad, estadística y de matemáticas en general; por lo que se han omitido desarrollos teóricos que, aunque importantes en probabilidad general, no son esenciales para los fines de esta investigación. Para los temas aquí presentados, se consultó la siguiente bibliografía: Canavos y Urbina (1987), Lladser (2011), Murray y Spiegel (2009), Rincón (2006) y Rincón (2007).

### 1.1. Bases de probabilidad

La teoría de la probabilidad se encarga del estudio de los fenómenos o experimentos aleatorios. Un **experimento aleatorio** (E) es aquel que, bajo las mismas condiciones, no es posible predecir el resultado.

En teoría de probabilidad, el **espacio muestral** ( $\Omega$ ) representa el conjunto de todos los resultados posibles de un experimento aleatorio; cada miembro del espacio muestral se denota por  $\omega$ . Este conjunto puede ser discreto (finito o infinito contable) o continuo (infinito no numerable). Un **evento** es un subconjunto  $A \subset \Omega$ , es decir, es un conjunto de puntos muestrales.

La probabilidad de un evento  $A$ , es un número real en  $[0, 1]$  denotado por  $\mathbb{P}(A)$ , y mide la frecuencia con la que se observa la ocurrencia del evento  $A$  cuando se efectúa el experimento aleatorio. A continuación, se define formalmente.

**Definición 1.1.** Dado un espacio muestral  $\Omega$ , una **medida de probabilidad** es una función  $\mathbb{P} : \Omega \rightarrow [0, 1]$  que asigna a cada evento  $A \subseteq \Omega$  un número real  $\mathbb{P}(A)$ .

La función  $\mathbb{P}$  debe satisfacer los axiomas de probabilidad de Kolmogorov.

Los siguientes tres postulados fueron establecidos por el matemático ruso Andréi Kolmogorov (1933), conocidos como **axiomas de probabilidad de Kolmogorov**.

**Axioma 1.1 (No-negatividad).**  $\mathbb{P}(A) \geq 0$ , para todo evento  $A \in \Omega$ .

**Axioma 1.2 (Normalización).**  $\mathbb{P}(\Omega) = 1$ .

**Axioma 1.3 (Aditividad).**  $\mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(A_i)$ ;  $\{A_i\}_{i=1}^{\infty}$  eventos disjuntos a pares.

**Proposición 1.1.** De los axiomas de Kolmogorov se derivan las siguientes propiedades fundamentales de una función de probabilidad:

- a) Evento imposible:  $\mathbb{P}(\emptyset) = 0$ .
- b) Regla del complemento:  $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$ .
- c) Monotonicidad: Si  $A \subseteq B$ , entonces  $\mathbb{P}(A) \leq \mathbb{P}(B)$ .

### 1.1.1. Variables aleatorias y su distribución

En el estudio de fenómenos aleatorios, el concepto de variable aleatoria permite cuantificar resultados inciertos y asignarles estructura matemática. Una **variable aleatoria (v.a.)** es el valor que resulta de un experimento aleatorio que puede adoptar diferentes valores dentro de un intervalo. Formalmente se define como:

**Definición 1.2.** Dado un experimento aleatorio  $E$  y el espacio muestral  $\Omega$  de  $E$ , una **variable aleatoria** es una función  $X$  del espacio  $\Omega$  al conjunto de números reales  $\mathbb{R}$

$$X : \Omega \rightarrow \mathbb{R}.$$

**Definición 1.3.** Sea  $X$  una v.a. del espacio  $\Omega$ . Si  $\omega \in \Omega$ , con  $X(\omega)$  se denota la imagen de  $\omega$  bajo  $X$ . A la imagen de  $X$  se le denomina recorrido o **soporte de  $X$**  y se denota por  $R(X)$

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

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

Así  $X$  transforma todos los posibles resultados del espacio muestral en cantidades numéricas, lo que permite cuantificar fenómenos aleatorios para facilitar el cálculo de probabilidades mediante herramientas analíticas.

Las variables aleatorias se clasifican en dos categorías según su estructura:

- **Discretas:** Toma valores contables finitos o infinitos numerables.  
Por ejemplo: Número de días que registraron niveles altos de ozono.
- **Continuas:** Toma valores en un intervalo real, infinitos no numerables.  
Por ejemplo: Concentración horaria de ozono en partes por billón.

**Definición 1.4.** Para una variable aleatoria discreta  $X$  se define su **función de masa de probabilidad** (f.m.p.) como:  $p_X(x) = \mathbb{P}(X = x)$ , la cual debe satisfacer las propiedades de No-negatividad y Normalización de los axiomas de Kolmogorov.

El estudio completo de cualquier variable aleatoria requiere la caracterización de su comportamiento probabilístico, lo cual se logra mediante la función de distribución acumulada y, cuando existe, la función de densidad de probabilidad.

**Definición 1.5.** Sea  $X$  una v.a. continua, una función integrable  $f_X : \mathbb{R} \rightarrow \mathbb{R}$ , es la **función de densidad de probabilidad de  $X$**  (f.d.p.) si satisface:

1.  $f_X(x) \geq 0$ , para todo  $x \in \mathbb{R}$ .
2.  $\int_{-\infty}^{\infty} f_X(x) dx = 1$ .
3. Para cualquier intervalo  $[a, b]$ ,  $\mathbb{P}(a \leq X \leq b) = \int_a^b f_X(x) dx$ .

Estas funciones no solo codifican toda la información probabilística de la variable, sino que permiten el cálculo de probabilidades para eventos arbitrarios y establecen los fundamentos para derivar propiedades estadísticas clave.

**Definición 1.6.** La **función de distribución acumulada** (f.d.a.) para una variable aleatoria  $X$ , está dada por  $F_X : \mathbb{R} \rightarrow [0, 1]$  y se define como:

$$F_X(x) = \mathbb{P}(X \leq x) = \mathbb{P}(\{\omega \in \Omega : X(\omega) \leq x\}).$$

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

$$F_X(x) = \mathbb{P}(X \leq x) = \int_{-\infty}^x f_X(u) du.$$

**Teorema 1.1.** Sea  $f(x)$  una función continua en el intervalo  $[a, b]$  y sea  $F(x)$  una función tal que  $F'(x) = f(x)$  para todo  $x \in [a, b]$ . Entonces:

$$\int_a^b f(x) dx = F(b) - F(a).$$

**Observación 1.1.** Esta breve versión del teorema fundamental del cálculo afirma que, para una variable aleatoria  $X$ ,  $F'_X(x) = f_X(x)$ . De manera que, se puede determinar la función de densidad  $f_X(x)$  a partir de la función de distribución  $F_X(x)$ :

$$f_X(x) = \frac{dF_X(x)}{dx}.$$

**Proposición 1.2.** Para cualquier v.a.  $X$ , toda función de distribución acumulada  $F_X(x)$  debe satisfacer las siguientes propiedades:

1.  $\lim_{x \rightarrow -\infty} F_X(x) = 0$  y  $\lim_{x \rightarrow +\infty} F_X(x) = 1$ .
2.  $F_X(x)$  es monótona creciente: Si  $x_1 \leq x_2$  entonces  $F_X(x_1) \leq F_X(x_2)$ .
3. Si  $x_1 \leq x_2$  entonces  $\mathbb{P}(x_1 < X < x_2) = F_X(x_2) - F_X(x_1)$ .
4.  $F_X(x)$  es continua por la derecha:  $\lim_{x \rightarrow a^+} F_X(x) = F_X(a)$ .

**Definición 1.7.** Una v.a.  $X$  es **degenerada** si existe una constante  $c \in \mathbb{R}$  tal que  $\mathbb{P}(X = c) = 1$ , es decir, el soporte de  $X$  se limita a un único valor  $c$ .

La distribución asociada a una v.a. degenerada se denomina distribución degenerada, en caso contrario, se conoce como no degenerada.

**Definición 1.8.** Sea  $X$  una v.a. degenerada, con f.d.p. dada por  $\mathbb{P}(X = x) = 1$ .

La función de distribución acumulada se define como:

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

**Nota:** Aunque los conceptos de Esperanza ( $\mathbb{E}$ ) y Varianza ( $\text{Var}$ ) se verán más adelante, se anticipan los siguientes resultados:

**Proposición 1.3.** Si  $X$  es una variable aleatoria degenerada se cumple:

$$\text{a) } \mathbb{E}[X] = c, \quad \text{b) } \mathbb{E}[X^r] = c^r, \quad r \in \mathbb{N}, \quad \text{c) } \text{Var}(X) = 0.$$

**Lema 1.1.** Una variable aleatoria  $X$  es degenerada si, y sólo si,  $\text{Var}(X) = 0$ .

### 1.1.2. Vectores aleatorios y generalizaciones

Los vectores aleatorios constituyen la generalización natural del concepto de variable aleatoria al caso multivariado, permitiendo modelar fenómenos donde múltiples cantidades aleatorias interactúan simultáneamente. Los vectores aleatorios son esenciales en aplicaciones que requieren analizar dependencias entre varias variables.

**Definición 1.9.** Un **vector aleatorio**  $\mathbf{X}$  de dimensión  $n$  es una función  $\mathbf{X} : \Omega \rightarrow \mathbb{R}^n$  dada por  $\mathbf{X} = (X_1, X_2, \dots, X_n)$ , donde cada coordenada  $X_i : \Omega \rightarrow \mathbb{R}$  es una variable aleatoria para  $i = 1, \dots, n$ .

**Nota 1.1.** 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 1.10.** Sea  $\mathbf{X} = (X_1, X_2, \dots, X_n)$  un vector aleatorio continuo. Se dice que la función integrable y no negativa  $f_{X_1, X_2, \dots, X_n}(x_1, x_2, \dots, x_n) : \mathbb{R}^n \rightarrow [0, \infty)$  es la función de densidad del vector  $\mathbf{X}$ , o bien que es la **función de densidad de probabilidad conjunta** de las variables  $X_1, X_2, \dots, X_n$  si se cumple la igualdad:

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

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

**Proposición 1.4.** Toda f.d.p. conjunta  $f_{X_1, \dots, X_n}$  cumple dos propiedades:

- a)  $f_{X_1, \dots, X_n}(x_1, \dots, x_n) \geq 0, \quad x_1, \dots, x_n \in \mathbb{R}.$
- b)  $\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_{\mathbf{X}}(x_1, \dots, x_n) dx_1 \cdots dx_n = 1.$

**Definición 1.11.** Sea  $\mathbf{X} = (X_1, \dots, X_n)$  un vector aleatorio. La función de acumulación de probabilidad del vector  $\mathbf{X}$   $F_{X_1, \dots, X_n}(x_1, \dots, x_n) : \mathbb{R}^n \rightarrow [0, 1]$ , está dada por:

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

En particular, si  $\mathbf{X}$  es un vector aleatorio continuo, la **función de distribución conjunta** de  $X_1, \dots, X_n$  está dada por:

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

donde  $u_1, \dots, u_n$  son variables de integración.

La función de distribución acumulada conjunta se interpreta como la probabilidad simultánea de eventos para todas las componentes que caracterizan completamente la ley del vector.

**Observación 1.2.** De manera que, la función de densidad conjunta  $X_1, \dots, X_n$  se puede calcular diferenciando  $F_{X_1, \dots, X_n}(x_1, \dots, x_n)$  con respecto a  $x_1, \dots, x_n$ ; es decir,

$$f_{X_1, \dots, X_n}(x_1, \dots, x_n) = \frac{\partial^n F_{X_1, \dots, X_n}(x_1, \dots, x_n)}{\partial x_1 \cdots \partial x_n}.$$

**Proposición 1.5.** Sea el vector aleatorio  $\mathbf{X}$  con f.d.a. conjunta  $F_{\mathbf{X}}(\mathbf{x})$  con  $\mathbf{x} = (x_1, \dots, x_n)$ , se cumplen las siguientes propiedades para cada componente  $x_i$ :

1.  $F_{\mathbf{X}}(x_1, \dots, x_n)$  es una función monótona creciente.
2.  $\lim_{x_i \rightarrow -\infty} F_{\mathbf{X}}(x_i) = 0$       y       $\lim_{x_i \rightarrow +\infty} F_{\mathbf{X}}(x_i) = 1$ .

**Definición 1.12.** Sea  $f_{X_1, \dots, X_n}(x_1, \dots, x_n)$  la f.d.p. conjunta del vector aleatorio continuo  $\mathbf{X}$ . Se define la **función de densidad de probabilidad marginal** de la variable aleatoria  $X_i$  como sigue:

$$F_{X_i}(x_i) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_{X_1, \dots, X_n}(x_1, \dots, x_n) dx_1 \cdots dx_{i-1} dx_{i+1} \cdots dx_n.$$

Las funciones de densidad marginales son funciones de densidad univariadas.

### 1.1.3. Independencia y tipos de convergencia

En muchos análisis de datos se busca que la probabilidad de ocurrencia de los eventos de estudio no dependa de la ocurrencia de otros. Se busca que la probabilidad de ocurrencia de un evento  $A$ , sea la misma independientemente de la ocurrencia de  $B$ .

**Definición 1.13.** Dos eventos  $A$  y  $B$  son **independientes**, si

$$\mathbb{P}(A \cap B) = \mathbb{P}(A) \mathbb{P}(B).$$

**Lema 1.2.** Las siguientes ecuaciones son equivalentes:

1.  $\mathbb{P}(A \cap B) = \mathbb{P}(A) \mathbb{P}(B)$ .
2.  $\mathbb{P}(A|B) = \mathbb{P}(A)$  cuando  $\mathbb{P}(B) > 0$ .
3.  $\mathbb{P}(B|A) = \mathbb{P}(B)$  cuando  $\mathbb{P}(A) > 0$ .

En el análisis de vectores aleatorios, ciertas distribuciones multivariadas destacan por su capacidad para modelar fenómenos complejos con dependencias entre variables. Intuitivamente, dos v.a.  $X$  y  $Y$  son independientes si cada acontecimiento que implica solamente  $X$  es independiente de cada evento que involucra sólo a  $Y$ .

**Definición 1.14.** Sea  $\mathbf{X} = (X_1, \dots, X_n)$  un vector aleatorio continuo. Se dice que las variables aleatorias  $X_1, \dots, X_n$  **son independientes** si la función de densidad de probabilidad conjunta  $f_{X_1, \dots, X_n}(x_1, \dots, x_n)$  es igual al producto de las  $n$  funciones de densidad marginal correspondiente a cada variable  $X_i$  para  $i = 1, \dots, n$ ;

$$f_{X_1, \dots, X_n}(x_1, \dots, x_n) = \prod_{i=1}^n f_{X_i}(x_i),$$

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

Análogamente, se puede definir la independencia en términos de la función de distribución conjunta  $F_{X_1, \dots, X_n}$ . Ambas definiciones son equivalentes y se usan con regularidad en conceptos propios de estadística.

**Proposición 1.6.** Sea  $X_1, \dots, X_n$  una sucesión de variables aleatorias independientes. Para cada  $i = 1, \dots, n$ , sean  $h_i : \mathbb{R} \rightarrow \mathbb{R}$  funciones y  $Y_i$  variables aleatorias dadas por

$$Y_i = h_i(X_i),$$

entonces  $Y_1, \dots, Y_n$  también son variables aleatorias independientes.

**Nota 1.2.** Una sucesión de variables aleatorias  $X_1, \dots, X_n$  es *independiente e idénticamente distribuida* (v.a.i.i.d.) si todas las variables son independientes y tienen la misma distribución de probabilidad.

## Tipos de convergencia

En esta sección se estudia el comportamiento asintótico de sucesiones de variables aleatorias, y se darán las definiciones de algunos tipos de convergencia.

Si se considera una sucesión infinita de variables aleatorias  $\{X_n\}_{n=1}^{\infty}$  denotado por  $\{X_n\}$ , o simplemente  $X_n$ , la variedad de formas en las que puede definirse la convergencia de v.a.'s estará dada por las formas en las que se decida medir la cercanía de la sucesión con el límite a través de la medida de probabilidad.

**Definición 1.15** (Convergencia puntual). La sucesión de variables aleatorias  $\{X_n\}_{n=1}^{\infty}$  **converge puntualmente** a la variable aleatoria  $X$  si para cada  $\omega \in \Omega$ ,

$$\lim_{n \rightarrow \infty} X_n(\omega) = X(\omega).$$

Si una sucesión  $X_n$  converge puntualmente a  $X$ , sólo se denota por  $X_n \rightarrow X$ .

En general, la convergencia puntual es una condición muy **fuerte** ya que pide la convergencia de la sucesión evaluada en todos y cada uno de los elementos de  $\Omega$ .

Se puede ser menos estricto y pedir, por ejemplo, que la convergencia se verifique en todo el espacio  $\Omega$  excepto en un subconjunto de probabilidad cero; este es el caso del concepto de convergencia con probabilidad 1.

**Definición 1.16** (Convergencia casi segura). La sucesión de variables aleatorias infinitas  $\{X_n\}_{n=1}^{\infty}$  **converge con probabilidad 1** (o converge casi seguramente) a la variable aleatoria  $X$ , si

$$\mathbb{P}\left(\omega \in \Omega : \lim_{n \rightarrow \infty} X_n(\omega) = X(\omega)\right) = 1.$$

Para indicar la convergencia casi segura se escribe  $X_n \xrightarrow{\text{c.s.}} X$ , o bien  $\lim_{n \rightarrow \infty} X_n = X$  c.s.

En la convergencia casi segura se permite que para algunos valores de  $\omega$ , la sucesión numérica  $X_1(\omega), X_2(\omega), \dots$  pueda no converger. De este modo, se permite que exista un subconjunto de  $\Omega$  en donde no se verifique la convergencia; así, tal subconjunto debe tener probabilidad cero. El subconjunto de  $\Omega$  que verifique convergencia debe tener probabilidad 1.

Un tipo de convergencia aún menos restrictiva que la convergencia casi segura es la convergencia en probabilidad.

**Definición 1.17** (Convergencia en probabilidad). La sucesión de variables aleatorias  $\{X_i\}_{i=1}^{\infty}$  **converge en probabilidad** a la v.a.  $X$ , o bien  $X_n \xrightarrow{P} X$ , si para cada  $\varepsilon > 0$ ,

$$\lim_{n \rightarrow \infty} \mathbb{P}[|X_n - X| > \varepsilon] = 0.$$

o equivalentemente

$$\lim_{n \rightarrow \infty} \mathbb{P}[|X_n - X| \leq \varepsilon] = 1.$$

**Lema 1.3.** Si una sucesión  $X_n$  converge en probabilidad a  $X$ , entonces cualquier sub-sucesión de  $X_n$  también converge en probabilidad a  $X$ .

**Proposición 1.7.** 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  $\mathbb{P}[X = Y] = 1$ .
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  $X_n^2 \xrightarrow{P} X^2$ .

**Lema 1.4.** Sean  $X_n$  y  $Y_n$  sucesiones de variables aleatorias tales que  $X_n \xrightarrow{P} X$  y  $Y_n \xrightarrow{P} Y$ , entonces  $X_n Y_n \xrightarrow{P} XY$ .

En la práctica, es difícil determinar con precisión la función de distribución de probabilidad. Estas complicaciones pueden originarse porque la distribución es desconocida o por limitaciones analíticas para su cálculo exacto.

En esta situación es posible realizar una aproximación a la distribución real. Esto requiere una definición de convergencia de variables aleatorias. Existen muchas posibilidades, pero la más útil para nuestros propósitos es la convergencia en distribución.

**Definición 1.18** (Convergencia en distribución). La sucesión de variables aleatorias  $X_1, X_2, \dots$  con funciones de distribución  $F_{X_1}, F_{X_2}, \dots$  respectivamente **converge en distribución** a la variable aleatoria  $X$ , con función de distribución  $F_X$ , si cumple que

$$\lim_{n \rightarrow \infty} F_{X_n}(x) = F_X(x),$$

para todo punto  $x$  en donde  $F_X(x)$  es continua.

**Nota 1.3.** Este tipo de convergencia es la menos restrictiva de todas las mencionadas y en contextos más generales también se le llama **convergencia débil**; se puede denotar por:  $X_n \xrightarrow{d} X$ ,  $F_{X_n} \xrightarrow{d} F_X$ , o bien  $X_n \xrightarrow{d} F_X$ .

**Teorema 1.2.** El límite es único en el sentido de igualdad en distribución:

$$\text{Si } X_n \xrightarrow{d} X \text{ y } X_n \xrightarrow{d} Y \text{ entonces } X \stackrel{d}{=} Y, \text{ es decir, } F_X(x) = F_Y(x).$$

En la práctica estadística, identificar  $F_X(x)$  como distribución límite para una sucesión de variables aleatorias  $X_1, X_2, \dots$  justifica su uso como una aproximación a la distribución de  $X_n$  para  $n$  suficientemente grande.

## 1.2. Conceptos de estadística

Una **población** de interés es un conjunto arbitrario de personas, mediciones u objetos en general. Para conocer cierta información de esta población, se procede a tomar un pequeño subconjunto representativo de la población llamado **muestra**. Al número de elementos de una muestra se le llama **tamaño de la muestra**.

Una **variable** es una característica que posee cada elemento de una población y varía de elemento a elemento. Las variables pueden ser **cualitativas**, o **cuantitativas**; dependiendo si describen atributos (cualidades), o si miden magnitudes (cantidades).

**Definición 1.19.** Una **muestra aleatoria** de tamaño  $n$  es una colección de variables aleatorias  $X_1, \dots, X_n$ , de una población, que son independientes e idénticamente distribuidas.

Es decir, que cada una de las v.a.  $X_i$  son independientes y tienen la misma distribución de probabilidad con los mismos parámetros, para  $i = 1, 2, \dots, n$ .

**Definición 1.20.** Una estadística muestral, o simplemente **estadístico**, es una v.a.  $T(X_1, \dots, X_n)$ , en donde  $X_1, \dots, X_n$  es una muestra aleatoria y  $T : \mathbb{R}^n \rightarrow \mathbb{R}$  es una función. Además, dado que  $T$  es una función de v.a.'s,  $T$  también es una v.a.

### 1.2.1. Estadísticos descriptivos

La estadística descriptiva es fundamental en el análisis de datos; permite resumir y describir las características principales de grandes volúmenes de información mediante medidas de dispersión y localización y algunas representaciones gráficas.

En esta sección se describirán estos conceptos de manera general y después se profundizará en las medidas más relevantes para este estudio.

**Medidas de tendencia central:** indican el valor en donde se concentran los datos, pero no pueden describir su variabilidad:

- La **media** proporciona un valor promedio que puede ser representativo, pero es susceptible a valores extremos.
- La **media muestral** para  $n$  variables aleatorias  $X_1, X_2, \dots, X_n$  se calcula:

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i.$$

- La **mediana** representa el valor medio de un conjunto de datos ordenados; es útil para describir distribuciones asimétricas o con valores atípicos, ya que no se ve afectada por estos extremos.
- La **moda** representa el valor más frecuente en un conjunto de datos; es útil cuando se desea identificar el valor más común. Un conjunto puede ser unimodal, bimodal o multimodal.

**Medidas de dispersión:** cuantifican la variabilidad y miden cuánto se alejan los datos del valor central.

- El **rango** mide la amplitud de un conjunto de datos; aunque es sensible a valores atípicos. Se obtiene de la diferencia entre el valor máximo y mínimo del conjunto:

$$R = \text{máx}(X) - \text{mín}(X).$$

- La **varianza** cuantifica la variabilidad de un conjunto de datos, mide la dispersión de los valores en relación con la media:

- \* Un valor alto de la varianza indica que los valores están muy dispersos respecto al valor esperado.
- \* Una varianza baja sugiere que los valores están más agrupados cerca del promedio.

- La **varianza muestral** se puede estimar como:

$$S_n^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2.$$

La varianza se expresa en unidades al cuadrado, lo que puede dificultar su interpretación directa en el contexto de los datos originales.

- La **desviación estándar** se define como la raíz cuadrada de la varianza. Así, la desviación estándar se expresa en las mismas unidades que los datos originales, lo que la convierte en una medida más intuitiva.

**Nota 1.4.** En una distribución normal, la desviación estándar es particularmente útil:

- El 68 % de los datos se ubica dentro de una desviación estándar de la media.
- El 95 % se encuentra dentro de dos desviaciones estándar del promedio.
- El 99.7 % se encuentra dentro de tres desviaciones estándar del valor central.

- El **coeficiente de variación** (CV) expresa la desv. estándar de un conjunto de datos como un porcentaje de su media, se calcula como:  $CV = \frac{\sigma}{\mu} \times 100$ ,  $\mu \neq 0$ .
- El **rango intercuartílico** (IQR) es la diferencia entre el tercer cuartil ( $Q_3$ ) y el primer cuartil ( $Q_1$ ) de un conjunto de datos:  $IQR = Q_3 - Q_1$ .  
Se enfoca en la variabilidad del rango que contiene la mitad central de los datos. A diferencia del rango simple, el IQR es menos sensible a los valores atípicos.

**Medidas de localización:** dividen el conjunto de datos en partes porcentuales y son indispensable para detectar outliers.

- Los **cuartiles** dividen un conjunto de datos en cuatro partes iguales. Existen tres cuartiles:  $Q_1$ ,  $Q_2$  y  $Q_3$  que dividen los datos en 25 %, 50 % (mediana) y 75 %, respectivamente.
- Los **deciles** dividen un conjunto de datos ordenados en diez partes iguales; permiten una comprensión más detallada de la distribución al ofrecer diez puntos de corte que representan el 10 %, 20 %, ... hasta el 90 % de las observaciones.
- Los **percentiles** dividen un conjunto de datos ordenados en cien partes iguales. Cada percentil indica el porcentaje de observaciones que se encuentran por debajo de un valor específico. Por ejemplo, el percentil 20 indica que el 20 % de los datos son menores o iguales a ese valor.

**Gráficos descriptivos:** facilitan la interpretación de los datos y ayudan a comunicar hallazgos de manera visual y más efectiva.

- El **histograma** muestra la distribución de datos continuos agrupándolos en intervalos; es ideal para visualizar la distribución y la frecuencia de los datos.
- El **diagrama de dispersión** representa puntos en un plano cartesiano, donde cada punto corresponde a un par de valores; se utiliza para identificar patrones, tendencias o correlaciones entre dos variables.
- El **box plot** muestra la distribución de un conjunto de datos a través de sus cuartiles. Incluye una caja que representa el rango intercuartílico (IQR) y líneas que indican los valores máximos y mínimos; es útil para detectar outliers<sup>1</sup> y comparar distribuciones entre diferentes grupos.

---

<sup>1</sup>Valores atípicos o anomalías que se desvían significativamente del resto de los datos en una muestra

### 1.2.2. Esperanza, varianza, asimetría y curtosis

Los conceptos de esperanza y varianza constituyen los fundamentos para caracterizar el comportamiento de una variable aleatoria. Además, el coeficiente de asimetría y curtosis proporcionan información detallada sobre la forma y comportamiento de una distribución.

**Definición 1.21.** Sea  $X$  una v.a. continua con una f.d.p.  $f_X(x)$ , el **valor esperado de  $X$**  denotado por  $\mathbb{E}[X]$ , o  $\mu_X$ , se define como el promedio o valor medio de  $X$ :

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

**Observación 1.3.** La esperanza existe y se dice que  $X$  tiene esperanza finita si la integral impropia anterior es **absolutamente convergente**, es decir

$$\int_{-\infty}^{\infty} |x| f_X(x) dx < \infty$$

en caso contrario, la v.a.  $X$  no tiene esperanza finita.

**Proposición 1.8.** Para dos v.a.  $X, Y$  y una constante  $a \in \mathbb{R}$ , la esperanza matemática cumple con las siguientes propiedades:

1. **Linealidad:**  $\mathbb{E}[aX + Y] = a \mathbb{E}[X] + \mathbb{E}[Y]$ .
2. **Monotonicidad:** Si  $X \leq Y$  entonces  $\mathbb{E}[X] \leq \mathbb{E}[Y]$ .

**Teorema 1.3.** Sea  $X$  una v.a. continua y  $Y = g(X)$  una v.a. con  $g : \mathbb{R} \rightarrow \mathbb{R}$ , entonces,

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

La esperanza  $\mu_Y$  existe si  $Y = g(X)$  es una v.a. con esperanza finita.

**Definición 1.22.** La **varianza** mide la dispersión de  $X$  alrededor de su valor medio, se denota por  $\sigma_X^2$  o bien  $\text{Var}(X)$  y se define como:

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

**Definición 1.23.** La **desviación estándar** es la raíz cuadrada positiva de la varianza:

$$\sigma_X = \sqrt{\text{Var}(X)}.$$

**Proposición 1.9.** La varianza cumple con las siguientes propiedades, donde  $a \in \mathbb{R}$ :

1.  $\text{Var}(aX + b) = a^2 \text{Var}(X)$ .
2.  $\text{Var}\left(\sum X_i\right) = \sum \text{Var}(X_i) + 2 \sum_{i < j} \text{Cov}(X_i, X_j)$ ,

donde  $\text{Cov}(X_i, X_j)$  es la Covarianza de  $X_i$  y  $X_j$  que se define más adelante.

**Observación 1.4.** Si  $\mathbb{E}[X^2] = \infty$ , la varianza no está definida. Sin embargo, siempre que se cumpla  $\mathbb{E}[X^2] < \infty$  para una v.a.  $X$ , entonces la varianza se puede expresar como sigue:

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

### Momentos de orden superior y la FGM

**Definición 1.24.** Sea  $X$  una v.a. continua con f.d.p.  $f(x)$  y  $r \in \mathbb{N}$ . El **r-ésimo momento** de  $X$  alrededor del cero (no centrado), si existe, se define por:

$$\mu_r = \mathbb{E}[X^r] = \int_{-\infty}^{\infty} x^r f(x) dx.$$

A los números  $\mu_1, \mu_2, \dots$ , se les llama también **momentos poblacionales**.

**Nota 1.5.** El **primer momento** alrededor del cero es la media o valor esperado de la v.a. y se denota por  $\mu$ ; así,  $\mu = \mu_1 = \mathbb{E}[X]$ . Además, se considera como la cantidad numérica alrededor de la cual los valores de la v.a. tienden a agruparse.

**Definición 1.25.** El **r-ésimo momento central** de  $X$  respecto a  $\mu = \mathbb{E}[X]$  es:

$$\mu_X^r = \mathbb{E}[(X - \mu)^r] = \int_{-\infty}^{\infty} (x - \mu)^r f(x) dx.$$

**Definición 1.26.** Sea  $X$  una v.a. con función de distribución  $F(x)$  y f.d.p.  $f(x)$  conocida. La **función generadora de momentos** (FGM)  $M_X : \mathbb{R} \rightarrow \mathbb{R}^+ \cup \{\infty\}$  se define como:

$$M_X(t) = \mathbb{E}[e^{tX}] = \int_{-\infty}^{\infty} e^{tx} f(x) dx.$$

**Ejemplo 1.1.** Sea  $X$  una v.a. con distribución normal,  $X \sim N(\mu, \sigma)$ , su f.d.p. está dada por

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right), \quad x \in \mathbb{R}.$$

Su FGM se obtiene evaluando la esperanza  $\mathbb{E}[e^{tX}]$ :

$$M_X(t) = \int_{-\infty}^{\infty} e^{tx} \cdot \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx.$$

Completando el cuadrado en el exponente y distribuyendo términos:

$$tx - \frac{(x-\mu)^2}{2\sigma^2} = -\frac{x^2 - 2\mu x + \mu^2 - 2\sigma^2 tx}{2\sigma^2} = -\frac{(x - (\mu + \sigma^2 t))^2}{2\sigma^2} - \frac{(\mu + \sigma^2 t)^2 - \mu^2}{2\sigma^2}$$

Sustituyendo y simplificando:

$$M_X(t) = e^{\mu t + \frac{\sigma^2 t^2}{2}} \cdot \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x - (\mu + \sigma^2 t))^2}{2\sigma^2}} dx.$$

Por ser la f.d.p. normal,  $\int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x - (\mu + \sigma^2 t))^2}{2\sigma^2}} dx = 1$ .

Por lo tanto:

$$M_X(t) = \exp\left(\mu t + \frac{\sigma^2 t^2}{2}\right).$$

**Nota 1.6.** La función generadora de momentos puede no existir para todo  $t \in \mathbb{R}$ . Por ejemplo, en la distribución de Cauchy cuya f.d.p. está dada por  $f(x) = 1/\pi\sigma \left[1 + \left(\frac{x-\mu}{\sigma}\right)^2\right]$ , la correspondiente integral  $M_X(t)$  diverge.

**Proposición 1.10.** Si  $M_X(t)$  existe en un entorno de  $t = 0$ , entonces se cumplen:

1. **Generación de momentos:** El  $r$ -ésimo momento se obtiene como:

$$\mathbb{E}[X^r] = \left. \frac{d^r}{dt^r} M_X(t) \right|_{t=0}.$$

2. **Unicidad:** Si  $M_X(t) = M_Y(t)$ , entonces  $X \stackrel{d}{=} Y$ .

3. **Transformaciones lineales:** Si  $Y = aX + b$ , entonces  $M_Y(t) = e^{bt} M_X(at)$ .

**Definición 1.27.** Sea  $X_1, \dots, X_n$  una m.a. y sea  $r \in \mathbb{N}$ . El  $r$ -ésimo momento muestral no centrado o simplemente el **r-ésimo momento muestral** es la variable aleatoria dada por:

$$M_r = \frac{1}{n} \sum_{i=1}^n X_i^r.$$

El **método de momentos** para la estimación de parámetros consiste en resolver un sistema de ecuaciones planteado al igualar los momentos poblacionales con los momentos muestrales. Este método presupone que el sistema de ecuaciones tiene una solución única y ésta es sencilla de encontrar, pero no hay garantía que esto sea así.

La distribución hipergeométrica, por ejemplo, tiene tres parámetros  $N$ ,  $K$ ,  $n$ , pero solo dos momentos están definidos de manera convencional, esto genera un sistema con múltiples soluciones que dificulta la aplicación directa del método.

**Ejemplo 1.2.** Sea  $X$  una variable aleatoria con distribución normal de parámetros  $\mu$  y  $\sigma^2$ , esto es  $X \sim N(\mu, \sigma^2)$  y  $X_1, \dots, X_n$  una muestra aleatoria.

Para hallar un estimador de los parámetros  $\mu$  y  $\sigma^2$  por el método de momentos, se puede verificar que el primer y segundo momento poblacional están dados por:  $\mu_1 = \mathbb{E}[X] = \mu$  y  $\mu_2 = \mathbb{E}[X^2] = \sigma^2 + \mu^2$ , respectivamente.

Igualando estos momentos poblacionales con sus respectivos momentos muestrales, se tiene el siguiente sistema de ecuaciones;

$$\hat{\mu} = \bar{X}, \quad \hat{\sigma}^2 + \hat{\mu}^2 = \frac{1}{n} \sum_{i=1}^n X_i^2.$$

De la primera ecuación obtenemos directamente el estimador:  $\hat{\mu} = \bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ .

Reemplazando  $\hat{\mu}$  en la segunda ecuación:

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n X_i^2 - \bar{X}^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2,$$

donde la varianza muestral está definida como:

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2.$$

Por lo tanto:  $\hat{\sigma}^2 = \frac{n-1}{n} S^2$ .

Así los estimadores de  $\mu$  y  $\sigma^2$  por el método de momentos son  $\hat{\mu} = \bar{X}$  y  $\hat{\sigma}^2 = \frac{n-1}{n} S^2$ .

## Coeficiente de asimetría y curtosis

El coeficiente de asimetría mide la falta de simetría de una distribución y la curtosis cuantifica el peso de las colas y la agudeza del pico.

**Definición 1.28.** Para una variable aleatoria  $X$  con media  $\mu_X$  y desviación estándar  $\sigma_X$ , el **coeficiente de asimetría** se denota por  $\gamma_1$ , o bien  $S_X$ , y se define como:

$$\gamma_1 = S_X = \frac{\mathbb{E}[(X - \mu_X)^3]}{\sigma_X^3}.$$

- Si  $\gamma_1 = 0$  o  $\gamma_1 \approx 0$  entonces la distribución de  $X$  es simétrica; Media  $\approx$  Mediana.
- Si  $\gamma_1 > 0$ , hay asimetría positiva (cola derecha más larga); Media  $>$  Mediana.
- Si  $\gamma_1 < 0$ , hay asimetría negativa (cola izquierda más larga); Media  $<$  Mediana.

**Ejemplo 1.3.** Comparación de distribuciones simétricas y asimétricas: Considere las gráficas de las siguientes distribuciones: Normal(0,1); Chi-cuadrada (3 gl); y Beta(5,2).



Figura 1.1: Comparación de distribuciones simétricas y asimétricas.

Cada gráfica se generó con una m.a.i.i.d. y se calcularon los respectivos coeficientes de asimetría  $\gamma_1 = S_x$ , de donde se obtuvo lo siguiente:

Distribución Normal(0,1):  $S_x = -0.0039$  approx. 0

Distribución Chi-cuadrada(gl=3):  $S_x = 1.6258 > 0$

Distribución Beta(5,2):  $S_x = -0.5998 < 0$

Los resultados empíricos concuerdan con los valores teóricos esperados:

- La distribución normal presenta simetría perfecta, pues  $\gamma_1 \approx 0$ .
- La distribución  $\chi_3^2$  muestra asimetría positiva, el alto valor ( $\gamma_1 = 1.6258 > 0$ ) refleja la fuerte cola derecha.
- La distribución Beta(5,2) con  $\alpha > \beta$  exhibe asimetría negativa, como el valor absoluto de  $\gamma_1$  no es tan alto, muestra menor grado de asimetría a comparación con la  $\chi_3^2$ .

**Definición 1.29.** Para una v.a.  $X$  con media  $\mu_X$  y desviación estándar  $\sigma_X$ , la **curtosis** se denota por  $\gamma_2$ , o bien  $K_X$ , y se define:

$$\gamma_2 = K_X = \frac{\mathbb{E}[(X - \mu_X)^4]}{\sigma_X^4}.$$

- Si  $\gamma_2 = 3$  o  $\gamma_2 \approx 3$  entonces la distribución de  $X$  tiene forma Mesocúrtica.
- Si  $\gamma_2 > 3$ , tiene forma Leptocúrtica (colas pesadas y pico agudo).
- Si  $\gamma_2 < 3$ , tiene forma Platicúrtica (colas ligeras y pico aplanado).

**Ejemplo 1.4.** Comparación de distribuciones por curtosis: Considere las gráficas de las siguientes m.a.i.i.d.: Normal(0,1); t-Student (10 gl); y Uniforme(-4,4).

Para cada muestra también se obtuvo su respectivo valor de curtosis  $\gamma_2 = K_X$ .

Figura 1.2: Comparación de distribuciones por curtosis.

Distribución Normal(0,1):  $K_x = 2.9901$  approx. 3

Distribución t-Student(gl=10):  $K_x = 3.9418 > 3$

Distribución Uniforme(-4,4):  $K_x = 1.7979 < 3$

Los resultados empíricos concuerdan con los valores teóricos esperados:

- La distribución normal (curtosis estándar) tiene curtosis cercana a 3 (Mesocúrtica).
- t-Student muestra una distribución más puntiaguda con colas pesadas (Leptocúrtica).
- La distribución uniforme tiene una forma más aplanada (Platicúrtica).

### 1.2.3. Vector de medias, covarianza y correlación

Los conceptos de media y varianza se generalizan al caso multivariado por medio del vector de medias y la matriz de covarianza, que extiende la varianza a covarianzas a pares; la matriz de correlación normaliza estas relaciones. Este marco permite analizar simultáneamente dependencias lineales entre múltiples variables.

**Definición 1.30.** Sea  $\mathbf{X} = (X_1, \dots, X_k)$  un vector aleatorio k-dimensional y supóngase que cada coordenada del vector tiene esperanza finita, entonces se define la esperanza de  $\mathbf{X}$ , o el **vector de medias** como:

$$\boldsymbol{\mu}_{\mathbf{X}} = \mathbb{E}[\mathbf{X}] = (\mathbb{E}[X_1], \dots, \mathbb{E}[X_k]).$$

**Definición 1.31.** Si suponemos que las esperanzas  $\mathbb{E}[X_1]$ ,  $\mathbb{E}[X_2]$  y  $\mathbb{E}[X_1X_2]$  son finitas; entonces la **covarianza** entre dos variables aleatorias  $X_1$  y  $X_2$  está dada por:

$$\text{Cov}(X_1, X_2) = \mathbb{E}[(X_1 - \mathbb{E}[X_1])(X_2 - \mathbb{E}[X_2])].$$

Cuando  $X_1$  y  $X_2$  son variables aleatorias continuas, se define:

$$\text{Cov}(X_1, X_2) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x_1 - \mathbb{E}[X_1])(x_2 - \mathbb{E}[X_2]) f_{X_1, X_2}(x_1, x_2) dx_1 dx_2.$$

**Observación 1.5.** Aplicando la linealidad de la esperanza en la definición 1.27, se verifica que:

$$\text{Cov}(X_1, X_2) = \mathbb{E}[X_1X_2] - \mathbb{E}[X_1] \mathbb{E}[X_2].$$

El valor de la covarianza **cuantifica el grado de asociación lineal** entre las variables: un valor positivo indica que ambas variables tienden a moverse en la misma dirección, mientras que un valor negativo sugiere una relación inversa.

**Proposición 1.11.** Para cualquier par de v.a.  $X_i, X_j$ , con  $i \neq j$ , del vector aleatorio  $\mathbf{X} = (X_1, \dots, X_n)$ , la covarianza cumple las siguientes propiedades:

1.  $\text{Cov}(X, X) = \text{Var}(X)$ .
2. Simétrica:  $\text{Cov}(X_i, X_j) = \text{Cov}(X_j, X_i)$ .
3. Linealidad:  $\text{Cov}(aX_i + b, cX_j + d) = ac \text{Cov}(X_i, X_j)$ , con  $a, b, c, d \in \mathbb{R}$ .
4. Si  $X_i$  y  $X_j$  son independientes, entonces  $\text{Cov}(X_i, X_j) = 0$ .
5.  $-\sqrt{\text{Var}(X_i) \text{Var}(X_j)} \leq \text{Cov}(X_i, X_j) \leq +\sqrt{\text{Var}(X_i) \text{Var}(X_j)}$ .

**Nota 1.7.** El recíproco de la propiedad 4 es en general falso; el hecho de que la covarianza sea cero no implica necesariamente que las v.a.'s sean independientes.

**Definición 1.32.** Si cada coordenada del vector aleatorio  $\mathbf{X}$  tiene varianza finita, entonces la varianza de  $\mathbf{X}$  se define como la matriz cuadrada de  $k \times k$  dada por:

$$\Sigma(\mathbf{X}) = \begin{pmatrix} \text{Var}(X_1) & \text{Cov}(X_1, X_2) & \dots & \text{Cov}(X_1, X_k) \\ \text{Cov}(X_2, X_1) & \text{Var}(X_2) & \dots & \text{Cov}(X_2, X_k) \\ \vdots & \vdots & \ddots & \vdots \\ \text{Cov}(X_k, X_1) & \text{Cov}(X_k, X_2) & \dots & \text{Var}(X_k) \end{pmatrix}.$$

La matriz  $\Sigma(\mathbf{X})$  es la **matriz de covarianza** y se denota por  $\Sigma$ .

**Observación 1.6.** Los elementos de la matriz de covarianza  $\Sigma$  cumplen:

1. La diagonal principal contiene las varianzas de cada variable aleatoria.
2. Las covarianzas reflejan las relaciones lineales entre las variables.
3. Su determinante  $|\Sigma|$  cuantifica el "volumen" conjunto de variabilidad.

**Proposición 1.12.** Para cualquier vector aleatorio  $\mathbf{X}$ , la matriz  $\Sigma$  cumple:

1.  $\Sigma$  es simétrica.
2.  $\Sigma$  es definida positiva:  $\mathbf{x} \Sigma \mathbf{x}^t \geq 0$  para todo vector  $\mathbf{x} \in \mathbb{R}^k$ .
3.  $\text{Var}(\mathbf{a}^t \mathbf{X}) = \mathbf{a}^t \Sigma \mathbf{a}$  para  $\mathbf{a} \in \mathbb{R}^k$ .
4. Si  $\mathbf{Y} = \mathbf{A}\mathbf{X} + \mathbf{b}$ , entonces  $\Sigma_{\mathbf{Y}} = \mathbf{A} \Sigma_{\mathbf{X}} \mathbf{A}^t$ .

**Ejemplo 1.5.** Sean  $X_1, X_2$  y  $X_3$  variables aleatorias que representan los promedios mensuales (2023) de las concentraciones de ozono (ppb) registradas por las estaciones de monitoreo atmosférico: CUA (Cuaajimalpa), BJU (Benito Juárez) y CCA (Centro de Ciencias de la Atmósfera); estamos interesados en saber si existe una relación lineal entre las concentraciones de ozono de cada estación.

Para calcular cada covarianza, primero obtenemos los valores de  $\mathbb{E}[X_1X_2]$ ,  $\mathbb{E}[X_1X_3]$  y  $\mathbb{E}[X_2X_3]$  para después aplicar la ecuación  $\text{Cov}(X_i, X_j) = \mathbb{E}[X_iX_j] - \mathbb{E}[X_i] \mathbb{E}[X_j]$ .

Los resultados se muestran en la siguiente tabla:

Tabla 1.1: Promedios mensuales de concentraciones de ozono (ppb) del año 2023.

Mes (2023)	CUA $X_1$	BJU $X_2$	CCA $X_3$	$X_1X_2$	$X_1X_3$	$X_2X_3$
Enero	28.804	19.363	26.166	557.747	753.680	506.657
Febrero	32.352	21.404	28.629	692.472	926.206	612.768
Marzo	33.569	22.929	30.692	769.702	1030.306	703.746
Abril	32.968	24.340	29.784	802.434	981.913	724.953
Mayo	34.237	24.603	30.956	842.354	1059.864	761.633
Junio	33.208	26.726	33.465	887.497	1111.292	894.367
Julio	28.660	20.806	26.411	596.306	756.929	549.518
Agosto	26.950	21.634	23.612	583.021	636.336	510.820
Septiembre	30.524	28.498	26.766	869.863	816.982	762.769
Octubre	25.719	22.673	20.730	583.116	533.153	470.001
Noviembre	24.449	21.041	19.952	514.453	487.816	419.820
Diciembre	18.401	15.287	14.195	281.299	261.204	217.001
$\mathbb{E}[X_i]; \mathbb{E}[X_iX_j]$	29.153	22.442	25.946	665.022	779.640	594.504

De esta forma, calculamos cada covarianza a pares  $\text{Cov}(X_i, X_j)$ :

$$\text{Cov}(X_1, X_2) = \mathbb{E}[X_1 X_2] - \mathbb{E}[X_1] \mathbb{E}[X_2] = 665.022 - (29.153)(22.442) = 10.770 \text{ ppb}^2.$$

$$\text{Cov}(X_1, X_3) = \mathbb{E}[X_1 X_3] - \mathbb{E}[X_1] \mathbb{E}[X_3] = 779.640 - (29.153)(25.946) = 23.236 \text{ ppb}^2.$$

$$\text{Cov}(X_2, X_3) = \mathbb{E}[X_2 X_3] - \mathbb{E}[X_2] \mathbb{E}[X_3] = 594.504 - (22.442)(25.946) = 12.224 \text{ ppb}^2.$$

Además, si obtenemos las varianzas de cada variable:

$$\text{Var}(X_1) = 20.266; \text{Var}(X_2) = 10.886; \text{Var}(X_3) = 27.808.$$

Podemos construir la matriz de covarianza, que está dada por:

$$\Sigma(\mathbf{X}) = \begin{pmatrix} 20.266 & 10.770 & 23.236 \\ 10.770 & 10.886 & 12.224 \\ 23.236 & 12.224 & 27.808 \end{pmatrix}.$$

### Conclusiones del análisis de covarianzas del Ejemplo 1.3:

Referente a la diagonal principal de la matriz de covarianza podemos deducir que la variable  $X_3$ , de la estación CCA, es la que tiene mayor dispersión en los datos registrados de concentración de  $O_3$ ; mientras que  $X_2$ , correspondiente a la estación BJU, es la más estable concentrando sus valores alrededor de su media (22.442 ppb).

Sobre los resultados de las covarianzas, los tres valores positivos nos confirman que sí existe una relación lineal directa de las concentraciones de ozono entre las tres estaciones de monitoreo; sin embargo, se observa una relación de mayor magnitud (23.236 ppb<sup>2</sup>) entre la estación de Cuajimalpa (zona suroeste) y la estación CCA ubicada en Coyoacán (zona sureste). Además, la covarianza entre la estación en Benito Juárez (zona central) y las periféricas es menor, sugiriendo diferencias en los patrones de dispersión de contaminantes.

Estos resultados indican que el comportamiento de las partículas de ozono en la región podría estar influenciado por factores meteorológicos (fuertes vientos en la zona sur; mayor radiación solar) o factores geográficos (zonas menos urbanizadas; afectaciones industriales); provocando una mayor correlación entre zonas periféricas que con el centro urbano de la Ciudad de México.

**Nota 1.8.** La covarianza mide la relación lineal entre dos variables aleatorias, pero tiene limitaciones: depende de las unidades de medida y no está acotada, lo que dificulta interpretar su magnitud. Por ello, se prefiere utilizar el coeficiente de correlación de Pearson ( $\rho$ ), que normaliza la covarianza usando las desviaciones estándar de las variables. A continuación, se define esta medida.

**Definición 1.33.** Sean  $X_1$  y  $X_2$  dos variables aleatorias, se define el **coeficiente de correlación de Pearson** como:

$$\rho_{X_1, X_2} = \frac{\text{Cov}(X_1, X_2)}{\sqrt{\text{Var}(X_1) \text{Var}(X_2)}} \in [-1, 1].$$

**Definición 1.34.** Sea cada  $\rho_{ij}$  el coeficiente de correlación entre  $X_i$  y  $X_j$ , donde  $\rho_{ii} = 1$ ; la **matriz de correlación  $\mathbf{R}$**  se define como:

$$\mathbf{R} = \begin{pmatrix} 1 & \rho_{12} & \cdots & \rho_{1k} \\ \rho_{21} & 1 & \cdots & \rho_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{k1} & \rho_{k2} & \cdots & 1 \end{pmatrix}.$$

El coeficiente de correlación de Pearson  $\rho_{X_1, X_2}$  es una medida estandarizada que permite comparar la fuerza de las asociaciones lineales, cuantificando el grado y dirección de asociación entre dos variables aleatorias. Además, identifica patrones de dependencia lineal, útil para validar supuestos en modelos como regresión lineal múltiple.

**Lema 1.5.** Bajo normalidad:  $X_1$  y  $X_2$  son independientes si, y solo si,  $\rho_{12} = 0$ .

Este resultado no se generaliza a distribuciones arbitrarias.

Tabla 1.2: Interpretación del coeficiente de correlación de Pearson.

$\rho$	Correlación Positiva	Comportamiento de la relación lineal
$\rho = 1$	Perfecta	Aumento directamente proporcional.
(0.7, 1)	Fuerte	Asociación entre altos valores de contaminantes.
(0.3, 0.7)	Moderada	Discernible, afectada por otros factores.
(0, 0.3)	Ausente	Puede haber relaciones no lineales.
$\rho = 0$	Independencia	No existe relación lineal entre las variables.
$\rho$	Correlación Negativa	Comportamiento de la relación lineal
(-0.3, 0)	Ausente	No implica independencia (relación no lineal).
(-0.7, -0.3)	Moderada	Si $X_1$ aumenta, $X_2$ tiende a disminuir.
(-1, -0.7)	Fuerte	Asociación inversa en niveles de contaminación.
$\rho = -1$	Perfecta	Incremento inversamente proporcional. No hay casos reales en sistemas ambientales.

**Ejemplo 1.6.** Siguiendo con el **Ejemplo 1.3**, ahora calcularemos los coeficientes de correlación de las tres variables aleatorias correspondientes a las estaciones de monitoreo atmosférico antes descritas:

$$\rho_{12} = \frac{\text{Cov}(X_1, X_2)}{\sqrt{\text{Var}(X_1) \text{Var}(X_2)}} = \frac{10.770}{\sqrt{(20.266)(10.886)}} = 0.725.$$

$$\rho_{13} = \frac{\text{Cov}(X_1, X_3)}{\sqrt{\text{Var}(X_1) \text{Var}(X_3)}} = \frac{23.236}{\sqrt{(20.266)(27.808)}} = 0.979.$$

$$\rho_{23} = \frac{\text{Cov}(X_2, X_3)}{\sqrt{\text{Var}(X_2) \text{Var}(X_3)}} = \frac{12.224}{\sqrt{(10.886)(27.808)}} = 0.703.$$

Así, la matriz de correlación está dada por:

$$\mathbf{R} = \begin{pmatrix} 1 & 0.725 & 0.979 \\ 0.725 & 1 & 0.703 \\ 0.979 & 0.703 & 1 \end{pmatrix}.$$

#### Conclusiones de la matriz de correlación:

Se observa una correlación lineal prácticamente perfecta entre las variables  $X_1$  y  $X_3$  ( $\rho_{13} = 0.979$ ), lo que indica que las estaciones de Cuajimalpa y la del Centro de Ciencias de la Atmósfera comparten prácticamente la misma información; indicando que están determinadas por factores geográficos o atmosféricos muy similares.

Esta fuerte relación sugiere que, para muchos propósitos prácticos, podría ser redundante incluir ambas estaciones de monitoreo atmosférico en análisis posteriores, ya que proporcionan esencialmente la misma información.

Por otro lado, las correlaciones de la estación Benito Juárez con las demás muestran una relación positiva significativa pero considerablemente más débil, lo que indica que la zona de esta estación de monitoreo contiene información adicional no capturada completamente por las estaciones periféricas.



# Capítulo 2

## Fundamentos de estimación paramétrica

El objetivo principal de un análisis estadístico, tanto descriptivo como inferencial, es extraer información significativa a partir de datos para apoyar la toma de decisiones, responder preguntas de investigación o comprender patrones y relaciones en los datos. Sin embargo, cada tipo de análisis tiene enfoques específicos:

La **estadística descriptiva** se enfoca en resumir, organizar y presentar los datos de manera clara y comprensible. Se enfoca en describir características básicas de los datos e identificar patrones, distribuciones o anomalías en los datos; sin embargo, no busca hacer generalizaciones, sino describir la muestra analizada.

En cambio, la **estadística inferencial** obtiene conclusiones o realiza predicciones sobre una población más grande a partir de una muestra de datos. Se enfoca en estimar parámetros poblacionales con intervalos de confianza, realiza pruebas de hipótesis para comparar grupos o evaluar relaciones causa-efecto y generaliza resultados con un margen de error controlado.

### 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. La inferencia equivale a la estimación de esta distribución para la cual existen dos enfoques distintos: paramétrico y no paramétrico. Para este análisis se desarrollará la teoría para seguir un **enfoque paramétrico**; asumiendo que las variables aleatorias de los datos siguen distribuciones conocidas de la TVE.

**Definición 2.1.** Sea  $X_1, \dots, X_n$  una muestra aleatoria y  $T(X_1, \dots, X_n)$  un estadístico; si se emplea  $T$  para estimar un parámetro desconocido  $\theta$ , entonces  $T$  recibe el nombre de **estimador** de  $\theta$ . Además, el valor específico de  $t$  como un resultado de los datos muestrales recibe el nombre de **estimación** de  $\theta$ .

**Definición 2.2.** Un **estimador puntual** para el parámetro desconocido  $\theta$ , que se denota  $\hat{\theta}$ , es una función de una muestra aleatoria  $X_1, \dots, X_n$  que sirve para estimar el valor del parámetro desconocido  $\theta$ .

Se estudiará el caso de una variable aleatoria continua  $X$  cuya función de densidad de probabilidad existe y está dada por  $f(x; \theta)$ . Suponga que los datos  $x_1, x_2, \dots, x_n$  comprenden una muestra aleatoria de observaciones independientes de  $X$ , cuya f.d.p. pertenece a una *familia* conocida de distribuciones de probabilidad con funciones de densidad que no dependen de parámetros desconocidos:

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

La inferencia se reduce, por tanto, a la estimación del parámetro  $\theta$ . Cuyo valor puede ser un escalar o puede representar un vector de parámetros como  $\theta = (\mu, \sigma)$ , en el caso de la familia normal.

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.

**Definición 2.3.** El **sesgo de un estimador** cuantifica la diferencia sistemática entre el valor esperado del estimador y el verdadero valor del parámetro poblacional. Para un parámetro  $\theta$  y su estimador  $\hat{\theta}$ , el sesgo ( $B$ ) se define como:

$$B(\hat{\theta}) = \mathbb{E}[\hat{\theta}] - \theta.$$

**Lema 2.1.** Si  $\mathbb{E}[\hat{\theta}] = \theta$ , el estimador  $\hat{\theta}$  se denomina **estimador insesgado**; o bien, se dice que tiene **sesgo nulo** si  $B(\hat{\theta}) = 0$ . Además, si  $B(\hat{\theta}) \neq 0$  se tiene:

- **Sesgo positivo:** Si  $B(\hat{\theta}) > 0$ , hay una sobreestimación sistemática de  $\theta$ .
- **Sesgo negativo:** Si  $B(\hat{\theta}) < 0$ , Subestimación sistemática de  $\theta$ .

**Definición 2.4.** El **error cuadrático medio** (ECM) mide la variación del estimador  $\hat{\theta}$  en torno al valor real del parámetro  $\theta$ ; se define como:

$$\text{ECM}(\hat{\theta}) = \mathbb{E}[(\hat{\theta} - \theta)^2].$$

El ECM puede expresarse en términos de la varianza y el sesgo del estimador:

$$\text{ECM}(\hat{\theta}) = \text{Var}(\hat{\theta}) + B(\hat{\theta})^2 = \text{Var}(\hat{\theta}) + [\mathbb{E}[\hat{\theta}] - \theta]^2.$$

El ECM considera tres criterios para su interpretación:

- **Medida de precisión:** Cuantifica la calidad del estimador.
- **Unidades:** Se expresa en unidades al cuadrado del parámetro original.
- **Minimización:** Se buscan estimadores con ECM mínimo.

En cualquier muestra en particular, un ECM bajo implica que probablemente la estimación se acerca al valor real del parámetro.

**Definición 2.5.** Se define el **error estándar** (SE) del estimador  $\hat{\theta}$  como la desviación estándar de su distribución muestral:  $SE(\hat{\theta}) = \sqrt{\text{Var}(\hat{\theta})}$ .

### Interpretación del error estándar

- El error estándar mide la dispersión de  $\hat{\theta}$  alrededor de  $\mathbb{E}[\hat{\theta}]$ .
- Cuantifica precisión: cuanto menor sea SE, mayor será la precisión del estimador.
- Es clave para realizar pruebas de hipótesis y construir intervalos de confianza.

### 2.1.1. Estadísticas de orden

Los estadísticos de orden tienen muchas aplicaciones dentro de la estadística no paramétrica e inferencial. Algunos ejemplos de estadísticos de orden son: el valor mínimo, el máximo, la mediana y otros cuantiles de una muestra.

En estadística, se suele considerar el *estadístico de orden  $i$*  de una muestra aleatoria como el  $i$ -ésimo valor más pequeño. Para una muestra aleatoria de tamaño  $n$ , y sea  $x_1, \dots, x_n$  una realización de esa muestra; el **mínimo** es siempre el valor más pequeño de la muestra, esto es,  $X_{(1)} = \min\{x_1, \dots, x_n\}$ ; mientras que el **máximo** es el valor más grande de la muestra, esto es  $X_{(n)} = \max\{x_1, \dots, x_n\}$ .

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  $X_1, \dots, X_n$ , se puede evaluar cada una de estas variables en cualquier punto muestral  $\omega$  y obtener una colección de números reales  $X_1(\omega), \dots, X_n(\omega)$ . Estos números pueden ser ordenados de menor a mayor, incluyendo repeticiones. Si  $X_{(i)}(\omega)$  denota el  $i$ -ésimo número ordenado, se tiene entonces la colección no decreciente de números reales:  $X_{(1)}(\omega) \leq \dots \leq X_{(n)}(\omega)$ .

Ahora bien, si se hace variar el argumento  $\omega$  para cada  $\omega \in \Omega$ , se obtienen  $n$  v.a. llamadas **estadísticas de orden**. Este proceso de ordenamiento resulta ser bastante útil en muchas aplicaciones, principalmente en la teoría de valores extremos.

**Definición 2.6.** Sean  $X_1, \dots, X_n$ , una muestra aleatoria, a las v.a.'s ordenadas:

$$\begin{aligned} X_{(1)} &= \min\{X_1, \dots, X_n\}, \\ X_{(2)} &= \min\{X_1, \dots, X_n\} \setminus \{X_{(1)}\}, \\ X_{(3)} &= \min\{X_1, \dots, X_n\} \setminus \{X_{(1)}, X_{(2)}\}, \\ &\vdots \\ X_{(n)} &= \max\{X_1, \dots, X_n\}, \end{aligned}$$

se les conoce como **estadísticas de orden**.

**Observación 2.1.** Con base a esta definición se debe de considerar lo siguiente:

1. Las estadísticas de orden no son v.a. independientes, pues deben mantener la relación  $X_{(1)} \leq X_{(2)} \leq \dots \leq X_{(n)}$ .
2. La  $i$ -ésima estadística de orden  $X_{(i)}$  es una función de todas las variables de la muestra aleatoria.

En otras palabras, puntualmente  $X_{(i)}$  es alguna de las variables  $X_1, \dots, X_n$ , pero globalmente no es necesariamente una de ellas. En general,  $X_{(i)}$  se llama  **$i$ -ésima estadística de orden**, para  $i = 1, \dots, n$ .

Así, las estadísticas de orden son nuevas variables aleatorias; suponiendo que  $X_1, \dots, X_n$  es una muestra aleatoria cuya distribución es conocida, y por simplicidad se supondrá absolutamente continua, en donde cada variable tiene función de densidad  $f(x)$  y una función de distribución acumulada  $F(x)$ ; el objetivo es encontrar algunas fórmulas relacionadas con las **distribuciones de probabilidad de las estadísticas de orden**. Por ejemplo, para encontrar las distribuciones del primer y del último estadístico de

orden se calcula:

$$\begin{aligned}
 F_{\min}(x) &= F_{X_{(1)}}(x) = \mathbb{P}(X_{(1)} \leq x) = \mathbb{P}(\min\{X_1, \dots, X_n\} \leq x) \\
 &= 1 - \mathbb{P}(\min\{X_1, \dots, X_n\} > x) \\
 &= 1 - \mathbb{P}(X_1 > x, X_2 > x, \dots, X_n > x) \\
 &= 1 - \prod_{i=1}^n \mathbb{P}(X_i > x) = 1 - [\mathbb{P}(X_i > x)]^n \\
 &= 1 - [1 - \mathbb{P}(X_i \leq x)]^n = 1 - [1 - F(x)]^n.
 \end{aligned}$$

$$\begin{aligned}
 F_{\max}(x) &= F_{X_{(n)}}(x) = \mathbb{P}(X_{(n)} \leq x) = \mathbb{P}(\max\{X_1, \dots, X_n\} \leq x) \\
 &= \mathbb{P}(X_1 \leq x, X_2 \leq x, \dots, X_n \leq x) \\
 &= \prod_{i=1}^n \mathbb{P}(X_i \leq x) = \prod_{i=1}^n F(x) = [F(x)]^n.
 \end{aligned}$$

**Teorema 2.1.** Para  $n \geq 1$ ,

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

La demostración del teorema es inmediata, pues sólo bastaría derivar las funciones de distribución acumulada que se obtuvieron anteriormente.

De manera general, se tiene la siguiente definición de la f.d.p.:

**Definición 2.7.** La **función de densidad** de la  $i$ -ésima estadística de orden es:

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

Las estadísticas de orden representan un conjunto de observaciones independientes de una variable aleatoria  $X$  con función de distribución  $F(x)$ .

En la práctica,  $(X_1, \dots, X_n)$  se trata de una muestra de  $n$  elementos tomada de cierta población; por lo que se puede reordenar de menor a mayor para obtener la secuencia:  $X_{1:n} \leq X_{2:n} \leq \dots \leq X_{n:n}$ , que son los **estadísticos ordenados de la muestra**.

**Definición 2.8.** El  $r$ -ésimo elemento de la secuencia de desigualdades  $X_{r:n}$ , se le conoce como **r-ésimo estadístico ordenado**. Los casos particulares del primer y el último estadístico ordenado, reciben el nombre de **valores extremos** o simplemente **extremos**;  $X_{1:n}$  es conocido como el **mínimo** y  $X_{n:n}$  como el **máximo**.

### 2.1.2. Nivel de retorno

**Definición 2.9.** Sea  $X$  una v.a. continua, tal que su f.d.a.  $F(x)$  es estrictamente monótona y continua en el intervalo  $(0, 1)$ , entonces se define la **función cuantil**  $Q(x)$  como la función inversa de  $F(x)$  en este intervalo:

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

**Definición 2.10.** El cuantil de probabilidad acumulada  $p$  o **cuantil de orden  $p$**  de la función de distribución  $F(x)$ , con  $p \in (0, 1)$ , se define como:

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

Análogamente, a  $F^{-1}(1 - p)$  se le conoce como cuantil de orden  $(1 - p)$ , cuantil de probabilidad acumulada  $(1 - p)$  o **cuantil por exceso**; se denota por  $Q_{1-p} = z_p$ .

Es necesario definir los conceptos de nivel de retorno y período de retorno, ya que es de interés conocer cuál es el valor que, en promedio, se excede una vez cada determinado tiempo. Para tal propósito, se considerarán v.a.i.i.d. con función de distribución común  $F(z)$  conocida, tomadas en periodos de tiempo iguales.

**Definición 2.11.** 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, es decir, es la frecuencia con la que se presenta tal evento.

En general, sea  $\{X_i\}_{i \geq 1}$  una sucesión de v.a.i.i.d. con función de distribución  $F(z)$  y  $w$  un valor dado. Considérese también la sucesión  $\mathbf{1}_{\{X_i > w\}}_{i \geq 1}$  de v.a.i.i.d. con distribución Bernoulli que toman el valor 1 (éxito) si  $X_i > w$  y 0 en otro caso, con probabilidad  $p = 1 - F(w)$  y  $1 - p$ , respectivamente.

Entonces, el **instante del primer éxito** está dado por

$$Y(w) = \min\{i \geq 1 \mid X_i > w\},$$

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

$$P\{Y(w) = k\} = (1 - p)^{k-1}p, \quad k = 1, 2, 3, \dots$$

Por lo tanto,  $T = \frac{1}{p}$ , donde  $T = \mathbb{E}[Y(w)]$ .

**Definición 2.12.** Al valor  $w$  se le llama **nivel de retorno** con periodo de retorno  $T = 1/p$  para los eventos  $\{X_i > w\}_{i \geq 1}$ .

**Ejemplo 2.1.** Si se desea encontrar el nivel de retorno  $w$  correspondiente a un periodo de 20 años, entonces se debe encontrar el valor de  $w$  tal que  $\mathbb{E}[Y(w)] = 20$ , es decir, encontrar  $w$  tal que  $p = 1/20 = 0.05$ .

Como  $p = 1 - F(w)$ , el problema consiste en resolver la ecuación  $F(w) = 1 - 0.05 = 0.95$  para  $w$ . Por lo tanto,  $w$  es el 0.95-cuantil de la distribución  $F(z)$ , esto es,  $Q_{0.95}$

**Nota 2.1.** El nivel de retorno es igual al cuantil por exceso, es decir,  $w = z_p$ .

## 2.2. Método de máxima verosimilitud

El método de máxima verosimilitud es considerado como uno de los mejores métodos en estimación paramétrica debido a sus propiedades óptimas en muestras grandes: consistencia, eficiencia (varianza mínima) y normalidad asintótica.

Intuitivamente, el método selecciona los valores  $\hat{\theta}$  que hacen más plausibles los datos observados. Su versatilidad lo hace aplicable a modelos complejos y su capacidad para derivar intervalos de confianza mediante la teoría asintótica lo consolidan como la herramienta fundamental en inferencia estadística.

### 2.2.1. Estimador de máxima verosimilitud (EMV)

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  $\theta$  se denomina función de verosimilitud.

Los valores de  $\theta$  que tienen alta probabilidad corresponden a modelos que dan alta probabilidad a los datos observados. El objetivo es hallar el estimador que asigna la mayor probabilidad al parámetro de los datos observados.

**Definición 2.13.** Sea  $X_1, \dots, X_n$  una muestra aleatoria de una población con función de densidad  $f_X(x; \theta)$ . La **función de verosimilitud** de la muestra, denotada por  $L(\theta) = L(\theta; x_1, \dots, x_n)$ , se define como la función de densidad de probabilidad conjunta de  $X_1, \dots, X_n$ , es decir:

$$L(\theta) = f_{X_1, \dots, X_n}(x_1, \dots, x_n; \theta) = f_{X_1}(x_1; \theta) \cdots f_{X_n}(x_n; \theta) = \prod_{i=1}^n f_{X_i}(x_i; \theta).$$

**Proposición 2.1.** En la práctica, es más conveniente tomar logaritmos y trabajar con la **función log-verosimilitud**:

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

Cabe mencionar que  $\ell(\theta)$  es máxima en el mismo punto en donde  $L(\theta)$  lo es, esto debido a las propiedades de continuidad y monotonía de la función logaritmo.

El **método de máxima verosimilitud** consiste en obtener el valor de  $\theta$  que maximice la función de verosimilitud  $L(\theta)$ , o bien, que maximice la función  $\ell(\theta)$ .

**Definición 2.14.** El **estimador de máxima verosimilitud (EMV)** de  $\theta$ , denotado por  $\hat{\theta}_{MV}$ , se define como el valor de  $\theta$  en donde  $\ell(\theta)$  alcanza el máximo:

$$\hat{\theta}_{MV} = \arg \max_{\theta \in \Theta} \ell(\theta).$$

**Ejemplo 2.2.** Sea  $X_1, \dots, X_n$  una m.a.i.i.d. de una distribución exponencial con parámetro  $\beta$ , cuya función de densidad de probabilidad es:

$$f(x; \beta) = \frac{1}{\beta} \exp\left(-\frac{x}{\beta}\right), \quad x \geq 0.$$

La función de verosimilitud para la muestra observada es:

$$L(\beta) = \prod_{i=1}^n f(x_i; \beta) = \prod_{i=1}^n \frac{1}{\beta} \exp\left(-\frac{x_i}{\beta}\right) = \left(\frac{1}{\beta}\right)^n \exp\left(-\frac{1}{\beta} \sum_{i=1}^n x_i\right).$$

Tomando la función log-verosimilitud:

$$\ell(\beta) = \ln L(\beta) = -n \ln \beta - \frac{1}{\beta} \sum_{i=1}^n x_i$$

Para encontrar el máximo, derivamos respecto a  $\beta$  e igualamos a cero:

$$\frac{d\ell}{d\beta} = -n\beta + \sum_{i=1}^n x_i = 0$$

Resolviendo para  $\beta$ :

$$n\beta = \sum_{i=1}^n x_i \quad \Rightarrow \quad \beta = \frac{1}{n} \sum_{i=1}^n x_i = \bar{x}.$$

Por lo tanto, el EMV de  $\beta$  corresponde a la media muestral:

$$\hat{\beta}_{\text{MV}} = \frac{1}{n} \sum_{i=1}^n X_i = \bar{X}.$$

**Observación 2.2.** Observaciones generales sobre el método de máxima verosimilitud:

- a) Puede aplicarse para distribuciones discretas y continuas.
- b) No siempre es fácil encontrar el máximo de  $L(\theta)$ .
- c) En algunos casos, el EMV puede no existir.

### 2.2.2. Propiedades del EMV y condiciones de regularidad

La **invariancia funcional** es una de las propiedades fundamentales de un EMV.

**Definición 2.15.** Sea  $\theta = (\theta_1, \dots, \theta_n)$  un vector  $n$ -dimensional de parámetros de una distribución. Si  $\tau : \mathbb{R}^n \rightarrow \mathbb{R}$  es cualquier función, entonces a  $\tau(\theta)$  se le llama **función paramétrica** o parametral.

**Principio de invarianza:** Suponga que  $X_1, \dots, X_n$  es una muestra aleatoria de una población cuya función de densidad está dada por  $f(x; \theta)$  y  $\hat{\theta}$  es el EMV de  $\theta$ . Si se desea estimar una función del parámetro  $\theta$ , es decir, una función paramétrica  $\tau(\theta)$ , se puede hacer mediante  $\tau(\hat{\theta})$ .

**Teorema 2.2.** Sea  $\hat{\theta}$  el estimador de máxima verosimilitud de  $\theta$ , entonces el estimador de máxima verosimilitud de cualquier función paramétrica  $\tau(\theta)$  está dado por  $\tau(\hat{\theta})$ .

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

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

**Teorema 2.3.** Sea  $\hat{\theta} = (\hat{\theta}_1, \dots, \hat{\theta}_k)$  el estimador de máxima verosimilitud de  $\theta = (\theta_1, \dots, \theta_k)$ . Si  $\tau(\theta) = (\tau_1(\theta), \dots, \tau_r(\theta))$  para  $1 \leq r \leq k$  es un vector  $r$ -dimensional de funciones paramétricas, entonces el estimador de máxima verosimilitud de  $\tau(\theta) = (\tau_1(\theta), \dots, \tau_r(\theta))$  es  $\tau(\hat{\theta}) = (\tau_1(\hat{\theta}), \dots, \tau_r(\hat{\theta}))$ , donde  $\tau_j(\hat{\theta})$  es el estimador de máxima verosimilitud de  $\tau_j(\theta)$  para  $j = 1, \dots, r$ .

Los EMV tienen otras propiedades fundamentales como **consistencia**, **insesgajez asintótica** y **normalidad asintótica**; sin embargo, a diferencia de la propiedad de invarianza funcional, estas otras propiedades se cumplen bajo ciertas condiciones generales, llamadas **condiciones de regularidad**.

## Condiciones de regularidad

A continuación se enuncian dichas condiciones para el caso uniparamétrico, es decir, cuando  $\theta$  es un escalar; sin embargo, estas condiciones pueden extenderse al caso donde  $\boldsymbol{\theta} = (\theta_1, \dots, \theta_k)$  es un vector  $k$ -dimensional de parámetros.

**Definición 2.16.** Sea  $X_1, \dots, X_n$  una muestra aleatoria de una población con una función de densidad  $f(x; \theta)$  y  $T(X_1, \dots, X_n)$  es un estimador de la función paramétrica  $\tau(\theta)$ . Entonces las **condiciones de regularidad** de la función de densidad  $f(x; \theta)$  y del estimador  $T(X_1, \dots, X_n)$  son las siguientes:

1. El soporte de la f.d.p.  $f(x; \theta)$  dado por el conjunto  $\{x | f(x; \theta) > 0\}$  no depende del parámetro  $\theta$ .
2. Para todo  $x$  en el soporte de  $f(x; \theta)$ , la función  $\ln[f(x; \theta)]$  es diferenciable respecto al parámetro  $\theta$ .

$$3. \frac{\partial}{\partial \theta} \int \cdots \int \prod_{i=1}^n f(x_i; \theta) dx_1 \cdots dx_n = \int \cdots \int \frac{\partial}{\partial \theta} \prod_{i=1}^n f(x_i; \theta) dx_1 \cdots dx_n.$$

$$4. \frac{\partial}{\partial \theta} \int \cdots \int T(\mathbf{x}) \prod_{i=1}^n f(x_i; \theta) dx_1 \cdots dx_n = \int \cdots \int T(\mathbf{x}) \frac{\partial}{\partial \theta} \prod_{i=1}^n f(x_i; \theta) dx_1 \cdots dx_n,$$

donde  $\mathbf{x} = (x_1, \dots, x_n)$ .

$$5. 0 < \mathbb{E} \left[ \left( \frac{\partial}{\partial \theta} \ln [f_X(x; \theta)] \right)^2 \right] < \infty.$$

Finalmente, se mencionan otras dos propiedades fundamentales del EMV:

■ **Insesgadez asintótica:** Bajo las condiciones de regularidad, si  $\hat{\theta}_n$  es el EMV del parámetro  $\theta$  basado en una m.a. de tamaño  $n$ , entonces  $\hat{\theta}_n$  es *insesgado asintóticamente*, es decir:

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

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

$$\text{a) } \lim_{n \rightarrow \infty} \mathbb{E}(\hat{\theta}_n) = \theta. \quad \text{b) } \lim_{n \rightarrow \infty} \text{Var}(\hat{\theta}_n) = 0.$$

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

### 2.3. Teoría univariada de verosimilitud

La función de verosimilitud permite cuantificar la plausibilidad de algún valor del parámetro desconocido  $\theta$  con base en una muestra observada, es decir, el EMV  $\hat{\theta}$  es el valor de  $\theta$  más probable, más plausible o más verosímil de  $\theta$  en el sentido que este maximiza la probabilidad de lo que se ha observado. Se considera el caso univariado, es decir, cuando la función de verosimilitud depende de un único parámetro desconocido  $\theta$ , también se conoce como caso uniparamétrico.

**Definición 2.17.** Se define la **función score**  $S(\theta)$  como la primera derivada de la función de log-verosimilitud, es decir,

$$S(\theta) = \ell'(\theta) = \frac{d\ell(\theta)}{d\theta}.$$

Por lo tanto, el estimador de máxima verosimilitud  $\hat{\theta}$  es la solución de la ecuación score:

$$S(\theta) = 0.$$

**Observación 2.3.** Dado que  $\hat{\theta}$  es un máximo, la derivada de segundo orden de la función log-verosimilitud evaluada en  $\hat{\theta}$  es negativa, esto es,

$$\left. \frac{d^2\ell(\theta)}{d\theta^2} \right|_{\theta=\hat{\theta}} = \ell''(\hat{\theta}) < 0.$$

**Definición 2.18.** Se define la **curvatura** en  $\hat{\theta}$  como  $\mathcal{J}(\hat{\theta})$ , donde,

$$\mathcal{J}(\theta) = -\ell''(\theta) = -S'(\theta) = -\frac{d^2\ell(\theta)}{d\theta^2}.$$

Así, una condición para que  $\hat{\theta}$  sea un EMV es que  $\mathcal{J}(\hat{\theta}) > 0$ .

**Nota 2.2.** En la teoría de verosimilitud,  $\mathcal{J}(\theta)$  se conoce como **información de Fisher de la muestra**; y el número  $\mathcal{J}(\hat{\theta})$  es una cantidad clave llamada **información de Fisher observada**.

**Definición 2.19.** Se define la **función de verosimilitud relativa** (FVR) de  $\theta$  como la razón de la función de verosimilitud  $L(\theta)$  y su máximo  $L(\hat{\theta})$ :

$$R(\theta) = \frac{L(\theta)}{L(\hat{\theta})},$$

donde  $0 \leq R(\theta) \leq 1$  y  $R(\hat{\theta}) = 1$ .

La función  $R(\theta)$  proporciona una medida relativa de compatibilidad con los datos observados:

- Para valores de  $\theta$  con  $R(\theta)$  cercanos a  $R(\hat{\theta}) = 1$  hacen a la muestra observada casi tan probable como lo hace el EMV  $\hat{\theta}$ .
- Valores de  $\theta$  con  $R(\theta)$  cercanos a cero hacen que la probabilidad de la muestra observada sea pequeña con respecto a su máxima probabilidad alcanzada en  $\hat{\theta}$ .

**Definición 2.20.** La **función de log-verosimilitud relativa** de  $\theta$  es el logaritmo natural de la función de verosimilitud relativa:

$$r(\theta) = \ln[R(\theta)] = \ln L(\theta) - \ln L(\hat{\theta}) = \ell(\theta) - \ell(\hat{\theta}),$$

donde  $\ell(\theta)$  es la función de log-verosimilitud y  $-\infty < r(\theta) \leq 0$  para todo  $\theta$ .

**Definición 2.21.** El conjunto de valores de  $\theta$  para los cuales  $R(\theta) \geq p$ , donde  $p$  es su nivel de verosimilitud y  $p \in (0, 1)$ ; se llama **región de verosimilitud** o **intervalo de verosimilitud** (IV) del  $(100 p) \%$  para  $\theta$ , esto es:

$$IV(p) = \{\theta \mid R(\theta) \geq p\}.$$

A  $IV(p)$  también se le conoce como intervalo de verosimilitud de nivel  $p$  para  $\theta$ .

**Observación 2.4.** Se hacen algunas observaciones sobre el intervalo de verosimilitud.

- a) La región de verosimilitud consiste en un intervalo de valores reales.
- b) Usualmente se consideran los intervalos de verosimilitud del 50 %, 10 % y 1 %.
- c) Se clasifican los valores de  $\theta$  contenidos en los intervalos de verosimilitud de nivel  $p = 0.1$  como valores “plausibles” y a los de nivel  $p = 0.5$  como “muy plausibles”. Por otra parte, cuando  $p = 0.01$  considera los valores fuera del intervalo como prácticamente imposibles.
- d) Los intervalos del 14.7 % y 3.6 % corresponden aproximadamente a los intervalos de confianza del 95 % y 99 %.
- e) Los intervalos de verosimilitud pueden ser determinados con la gráfica de  $r(\theta)$ .
- f)  $r(\theta) \geq -0.69$ ,  $r(\theta) \geq -2.30$  y  $r(\theta) \geq -4.61$  son los intervalos de verosimilitud del 50 %, 10 % y del 1 % respectivamente.
- g) Los puntos finales del IV del  $100p \%$  pueden encontrarse como raíces de la ecuación  $r(\theta) - \ln(p) = 0$ .

### 2.3.1. Probabilidad de cobertura

**Definición 2.22.** Sean  $X_1, \dots, X_n$  una m.a. con función de distribución  $F(x; \theta)$  donde  $\theta$  es un parámetro unidimensional fijo en un valor  $\theta_0$ ,  $A = T_1(X_1, \dots, X_n)$  y  $B = T_2(X_1, \dots, X_n)$  dos estadísticos tales que  $A < B$  y  $x_1, \dots, x_n$  una muestra observada proveniente de la m.a. de tamaño  $n$ . Entonces con base en esta muestra aleatoria se puede construir un intervalo  $[A, B]$  para el valor verdadero  $\theta_0$ ; se denomina como **intervalo de estimación**.

Ahora, si se construye nuevamente un intervalo  $[A, B]$  pero con otra muestra del mismo experimento o fenómeno aleatorio de interés, casi seguramente se obtendrá un intervalo diferente al anterior. Esto ocurre debido a que los intervalos  $[A, B]$  son variables aleatorias. Por tanto, cada vez que se varíe la muestra, los intervalos  $[A, B]$  algunas veces cubrirán el valor verdadero  $\theta_0$  y otras no.

**Definición 2.23.** La **probabilidad de cobertura** (PC) de un intervalo aleatorio  $[A, B]$  de  $\theta$ , es la probabilidad de que dicho intervalo cubra o contenga el verdadero valor  $\theta_0$  del parámetro  $\theta$ , es decir,

$$PC(\theta_0) = P(A \leq \theta_0 \leq B \mid \theta = \theta_0).$$

En otras palabras, la probabilidad de cobertura  $PC(\theta)$  expresa el porcentaje de veces que el intervalo  $[A, B]$  cubre al valor verdadero  $\theta_0$  en un número muy grande de repeticiones de un experimento.

Ahora bien, la distribución de probabilidad de los extremos del intervalo  $A$  y  $B$  se pueden calcular a partir de la distribución de la m.a. pero generalmente depende de  $\theta_0$ .

**Definición 2.24.** Un intervalo aleatorio  $[A, B]$  es llamado **intervalo de confianza** (IC) para  $\theta$  cuando su probabilidad de cobertura no depende de  $\theta_0$ , es decir, cuando el valor de  $PC(\theta_0)$  es el mismo para todo valor del parámetro  $\theta_0$ . En este caso la probabilidad de cobertura es igual a un número  $(1 - \alpha)$  llamado coeficiente de confianza o **nivel de confianza**.

Por tanto, de la definición anterior, se dice que  $[A, B]$  es un IC para  $\theta$  con nivel de confianza  $(1 - \alpha)$  si

$$P(A \leq \theta_0 \leq B \mid \theta = \theta_0) = (1 - \alpha).$$

Además, si  $\alpha \in (0, 1)$  se dirá que  $[A, B]$  es un intervalo de confianza del  $100(1 - \alpha)\%$ .

El nivel de confianza se refiere a la probabilidad de que  $\theta_0$  se encuentre entre las variables aleatorias  $A$  y  $B$  antes de que se observen los datos, y generalmente se utilizan niveles de confianza superiores al 90% ( $\alpha < 0.1$ ).

**Nota 2.3.** Cabe mencionar que una observación del intervalo  $[A, B]$  es un intervalo  $[a, b]$  con  $a = T_1(x_1, \dots, x_n)$  y  $b = T_2(x_1, \dots, x_n)$ .

### 2.3.2. Normalidad asintótica del EMV

El método de máxima verosimilitud ofrece como principal ventaja la normalidad asintótica de sus estimadores, propiedad que posibilita la aproximación de errores estándar y la construcción de intervalos de confianza. A continuación se presenta formalmente este resultado fundamental para los EMV, junto con otros teoremas relacionados y un previo panorama de la distribución normal multivariada.

**Definición 2.25.** Se dice que el vector  $\mathbf{X} = (X_1, \dots, X_n)$  tiene una **distribución normal multivariada** si su función de densidad de probabilidad conjunta es

$$f_{X_1, \dots, X_n}(x_1, \dots, x_n) = \frac{1}{(2\pi)^{n/2} |\Sigma|^{1/2}} \exp \left[ -\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu}) \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu})^t \right],$$

en donde  $\mathbf{x} = (x_1, \dots, x_n) \in \mathbb{R}^n$ ,  $\boldsymbol{\mu} = (\mu_1, \dots, \mu_n)$  es la esperanza del vector aleatorio  $\mathbf{X}$ ,  $\Sigma$  es la matriz de covarianza del vector  $\mathbf{X}$ ,  $|\Sigma|$  es el determinante de la matriz  $\Sigma$  y  $\Sigma^{-1}$  es la matriz inversa de  $\Sigma$ .

Para denotar que un vector aleatorio  $\mathbf{X}$  sigue una distribución normal multivariada con vector de medias  $\boldsymbol{\mu}$  y matriz de covarianza  $\Sigma$ , se escribe:

$$\mathbf{X} \sim NM(\boldsymbol{\mu}, \Sigma).$$

La matriz  $\Sigma$  de la definición anterior es una matriz de dimensión  $n \times n$  definida positiva, es decir,

$$\mathbf{x} \Sigma \mathbf{x}^t \geq 0 \quad \text{para todo vector } \mathbf{x} \in \mathbb{R}^n.$$

**Lema 2.2.** Si el vector  $\mathbf{X} \sim NM(\boldsymbol{\mu}, \Sigma)$ , se verifica que  $X_i$  tiene distribución marginal  $N(\mu_i, \sigma_i^2)$ , donde  $\mu_i$  y  $\sigma_i^2$  es la esperanza y la varianza de la variable aleatoria  $X_i$ , respectivamente, para  $i = 1, \dots, n$ .

**Teorema 2.4** (Normalidad asintótica del EMV). Sea  $X_1, \dots, X_n$  una sucesión de variables aleatorias independientes con función de distribución común  $F(x)$  y sea  $\hat{\boldsymbol{\theta}}$  el estimador de máxima verosimilitud del vector de parámetros  $\boldsymbol{\theta} = (\theta_1, \dots, \theta_q)$ . Entonces, bajo condiciones de regularidad, para  $n$  suficientemente grande

$$\hat{\boldsymbol{\theta}} \sim NM(\boldsymbol{\theta}, I_E^{-1}(\boldsymbol{\theta})),$$

donde

$$I_E(\boldsymbol{\theta}) = \begin{bmatrix} e_{1,1}(\boldsymbol{\theta}) & e_{1,2}(\boldsymbol{\theta}) & \cdots & e_{1,q}(\boldsymbol{\theta}) \\ e_{2,1}(\boldsymbol{\theta}) & e_{2,2}(\boldsymbol{\theta}) & \cdots & e_{2,q}(\boldsymbol{\theta}) \\ \vdots & \vdots & \ddots & \vdots \\ e_{q,1}(\boldsymbol{\theta}) & e_{q,2}(\boldsymbol{\theta}) & \cdots & e_{q,q}(\boldsymbol{\theta}) \end{bmatrix},$$

con

$$e_{i,j}(\boldsymbol{\theta}) = E \left[ -\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \theta_i \partial \theta_j} \right].$$

La matriz simétrica  $I_E(\boldsymbol{\theta})$  es llamada **matriz de información esperada**.

El teorema afirma que el EMV  $\hat{\boldsymbol{\theta}}$  tiene una distribución normal multivariada con vector de medias igual al vector  $q$ -dimensional de parámetros  $\boldsymbol{\theta}$  y matriz de covarianza igual a la inversa de la matriz de información esperada.

**Observación 2.5.** Es posible obtener los intervalos de confianza aproximados para las componentes individuales de  $\boldsymbol{\theta}$ , es decir, los intervalos de confianza para  $\theta_1, \dots, \theta_q$ .

Denotando un término arbitrario  $e_{i,j}(\boldsymbol{\theta})$  en la matriz inversa de  $I_E(\boldsymbol{\theta})$  por  $\psi_{i,j}$  se sigue del Lema 2.2 que, para  $n$  suficientemente grande, se cumple

$$\hat{\theta}_i \sim N(\theta_i, \psi_{i,i}).$$

Por lo tanto, si  $\psi_{i,i}$  es conocido, un intervalo de confianza del  $(1 - \alpha)100\%$  para  $\theta_i$  está dado por

$$\hat{\theta}_i \pm z_{\alpha/2} \sqrt{\psi_{i,i}}$$

donde  $z_{\alpha/2}$  es el  $(1 - \alpha/2)$ -cuantil de la distribución normal estándar.

Generalmente el verdadero valor de  $\boldsymbol{\theta}$  es desconocido y por tanto es usual aproximar los términos de la matriz de información esperada,  $I_E(\boldsymbol{\theta})$ , con los términos de la información de Fisher observada,  $\mathcal{J}(\boldsymbol{\theta})$  para el caso univariado.

Similarmente si se denotan los términos de la matriz inversa de la matriz de información de Fisher por  $\tilde{\psi}_{i,j}$ , se sigue que un intervalo de confianza aproximado del  $(1 - \alpha)100\%$  para  $\theta_i$  está dado por

$$\hat{\theta}_i \pm z_{\alpha/2} \sqrt{\tilde{\psi}_{i,i}}.$$

Los intervalos de confianza para muestras grandes de este tipo se denominan **intervalos de confianza de Wald**.

### 2.3.3. Razón de verosimilitud

Un método alternativo para cuantificar la incertidumbre en el EMV  $q$ -dimensional  $\hat{\theta}$  de  $\theta$  se basa en la función de devianza, también conocida como **razón de verosimilitud** o índice de probabilidad, definida por

$$D(\theta) = -2 \ln[R(\theta)],$$

donde  $R(\theta)$  es la función de verosimilitud relativa conjunta de  $(\theta_1, \dots, \theta_q)$ .

$$D(\theta) = -2 \ln[R(\theta)] = -2 \ln \left[ L(\theta)/L(\hat{\theta}) \right] = -2[\ln L(\theta) - \ln L(\hat{\theta})] = 2 [\ell(\hat{\theta}) - \ell(\theta)].$$

**Lema 2.3.** La **función de devianza** puede ser definida en términos de la función log-verosimilitud:

$$D(\theta) = 2[\ell(\hat{\theta}) - \ell(\theta)].$$

Recordemos que las regiones de verosimilitud se refieren sólo a la verosimilitud o plausibilidad relativa de los distintos valores del parámetro  $\theta$ , y no a la incertidumbre del intervalo. Sin embargo, en algunos casos es posible aproximar la probabilidad de que estas regiones contengan el verdadero valor del parámetro, y una forma de hacer esto es considerando la función de devianza y haciendo uso del siguiente teorema.

**Teorema 2.5.** Sea  $X_1, \dots, X_n$  una sucesión de variables aleatorias independientes con función de distribución común  $F(x)$  y sea  $\hat{\theta}$  el estimador de máxima verosimilitud del vector de parámetros  $\theta = (\theta_1, \dots, \theta_q)$ . Entonces, bajo condiciones de regularidad, la función de devianza satisface

$$D_n(\theta) \xrightarrow{d} \chi_q^2 \quad \text{cuando } n \rightarrow \infty.$$

Es decir que, bajo condiciones de regularidad, la función de devianza tiene una distribución chi-cuadrada con  $q$  grados de libertad.

A continuación se muestra la metodología para aproximar la probabilidad de cobertura de un intervalo de verosimilitud del  $100p\%$  a través de la distribución de probabilidad de la devianza para un  $\theta$  fijo en  $\theta_0$ .

Para un valor particular  $\theta = \theta_0$  se cumple la siguiente cadena de implicaciones:

$$\theta_0 \in IV(p) \iff R(\theta_0) \geq p \iff -2 \ln R(\theta_0) \leq -2 \ln(p).$$

De aquí que la probabilidad de cobertura del  $IV(p)$  sea

$$PC(\boldsymbol{\theta}_0) = \mathbb{P}(\boldsymbol{\theta}_0 \in IV(p) \mid \boldsymbol{\theta} = \boldsymbol{\theta}_0) = \mathbb{P}(D_n \leq -2 \ln(p) \mid \boldsymbol{\theta} = \boldsymbol{\theta}_0).$$

Así, se tiene que la probabilidad de cobertura del  $IV(p)$  es aproximadamente

$$\mathbb{P}(\chi_q^2 \leq x), \quad \text{donde } x = -2 \ln(p).$$

De esta manera, cuando  $x = c_{\alpha,q}$ , donde  $c_{\alpha,q}$  es el  $(1 - \alpha)$ -cuantil de la distribución  $\chi_q^2$ , se tiene que el intervalo de verosimilitud de nivel  $p$ , con  $p = \exp(-c_{\alpha,q}/2)$ , heredará una probabilidad de cobertura aproximada del  $100(1 - \alpha)\%$ .

Nótese que la probabilidad de cobertura del  $IV(p)$  no depende de  $\theta_0$ , de modo que el intervalo de verosimilitud es también un intervalo de confianza.

**Observación 2.6.** Una **región de confianza** de nivel aproximado  $100(1 - \alpha)\%$  está dada por:

$$C_{\alpha,q} = \{\boldsymbol{\theta} \mid D(\boldsymbol{\theta}) \leq c_{\alpha,q}\}.$$

Y en términos de la verosimilitud relativa, se puede escribir como:

$$C_{\alpha,q} = \{\boldsymbol{\theta} \mid R(\boldsymbol{\theta}) \geq \exp(-c_{\alpha,q}/2)\}.$$

donde  $c_{\alpha,q}$  es el  $(1 - \alpha)$ -cuantil de la distribución  $\chi_q^2$ .

**Ejemplo 2.3.** Consideremos una m.a. de tamaño  $n = 50$ , donde  $X_i \sim \text{Exp}(\theta)$ , para cada  $i = 1, \dots, 50$ , con función de densidad:  $f(x|\theta) = \theta \exp(-\theta x)$ ,  $x > 0$  y una media muestral de  $\bar{X} = 2$ .

■ Primero, debemos calcular el Estimador de Máxima Verosimilitud (EMV):

La función de verosimilitud es

$$L(\theta) = \prod_{i=1}^n \theta \exp(-\theta X_i) = \theta^n \exp(-\theta n\bar{X})$$

La log-verosimilitud

$$\ell(\theta) = n \ln \theta - \theta n\bar{X}$$

Derivando e igualando a cero para encontrar el EMV:

$$\frac{d\ell}{d\theta} = \frac{n}{\theta} - n\bar{X} = 0 \Rightarrow \hat{\theta} = \frac{1}{\bar{X}} \Rightarrow \hat{\theta} = 1/2.$$

■ Ahora, calculamos la función de devianza y construimos un IC al 95%:

La devianza se define como

$$D(\theta) = 2[l(\hat{\theta}) - l(\theta)] = 2 \left[ n \ln \left( \frac{\hat{\theta}}{\theta} \right) - n\bar{X}(\hat{\theta} - \theta) \right] = 2 \left[ 50 \ln \left( \frac{1}{2\theta} \right) - 100(1/2 - \theta) \right]$$

Para  $n$  grande,  $D(\theta) \stackrel{a}{\sim} \chi_1^2$ . Para  $\alpha = 0.05$ , el cuantil es  $\chi_{1,0.95}^2 = 3.841$ .

Resolvemos:

$$D(\theta) \leq 3.841 \iff \left[ \ln \left( \frac{1}{2\theta} \right) - (1 - 2\theta) \right] \leq 0.03841$$

Definimos la función

$$f(\theta) = \ln \left( \frac{1}{2\theta} \right) + 2\theta - 1.03841$$

Nos apoyamos de la gráfica para hallar la solución: IV=(0.374, 0.652).



## 2.4. Pruebas de hipótesis

Antes de revisar la metodología para realizar una prueba de hipótesis estadística, estudiaremos algunos conceptos y teoremas fundamentales que ayudarán a comprender mejor la teoría de valores extremos.

La ley de los grandes números asegura que, bajo ciertas condiciones, el promedio de variables aleatorias converge a la media común  $\mu$  conforme  $n$  tiende a infinito.

**Teorema 2.6** (Ley de los grandes números). Sea  $X_1, X_2, \dots$  una sucesión infinita de variables aleatorias independientes e idénticamente distribuidas (v.a.i.i.d.) con media finita  $\mu$ ; entonces,

$$\frac{1}{n} \sum_{i=1}^n X_i \rightarrow \mu, \quad \text{cuando } n \rightarrow \infty.$$

en donde la convergencia se verifica en el sentido casi seguro (ley fuerte) y también en probabilidad (ley débil).

**Nota 2.4.** Este resultado no establece la distribución de las variables aleatorias; estas pueden tener distribución discreta o continua, pero sí es indispensable que la media  $\mu$  sea finita. Además, nótese que la ley fuerte implica la ley débil, ya que la convergencia casi segura conlleva la convergencia en probabilidad.

El teorema central del límite es de amplio uso en estadística y otras ramas de aplicación de la probabilidad; existen muchas versiones y generalizaciones de este teorema, a continuación se enuncia una de ellas.

**Teorema 2.7** (Teorema del límite central). Sea  $\{X_i\}_{i=1}^{\infty}$ , una sucesión de v.a.i.i.d. tales que  $\mathbb{E}[X_n] = \mu$ ,  $\text{Var}(X_n) = \sigma^2 < \infty$  y considerando  $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ , entonces:

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \xrightarrow{d} N(0, 1), \quad \text{cuando } n \rightarrow \infty.$$

**Corolario 2.1.** Si consideramos las mismas condiciones del teorema anterior, pero ahora con  $\mathbb{E}[X_n] = p$ ,  $\text{Var}(X_n) = pq < \infty$  y considerando  $\hat{p} = \frac{1}{n} \sum_{i=1}^n X_i$ , entonces:

$$\frac{\hat{p} - p}{\sqrt{pq/n}} \xrightarrow{d} N(0, 1), \quad \text{cuando } n \rightarrow \infty.$$

### 2.4.1. Variables aleatorias estandarizadas

**Definición 2.26.** Para una v.a.  $X$  con media  $\mu_X$  y desviación estándar  $\sigma_X > 0$ , la **variable aleatoria estandarizada**  $Z$  está definida por:

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

donde los valores de la media y la desviación estándar de  $Z$  son:  $\mu_Z = 0$  y  $\sigma_Z = 1$ .

**Teorema 2.8.** Sea  $\{X_i\}_{i=1}^n$  una m.a.i.i.d. si  $X_i \sim N(\mu, \sigma^2)$  y  $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ , entonces:

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right).$$

**Teorema 2.9.** Sea  $\{X_i\}_{i=1}^n$  una m.a.i.i.d. donde  $\mathbb{E}[X_i] = \mu$ ,  $\text{Var}(X_i) = \sigma^2 < \infty$ , con la

distribución no necesariamente conocida, y considerando  $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$  entonces

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1), \quad \text{cuando } n \rightarrow \infty.$$

**Lema 2.4.** Sea  $\{X_i\}_{i=1}^n$  una m.a.i.i.d. con  $\mathbb{E}[X_i] = \mu$ ,  $\text{Var}(X_i) = \sigma^2$ ; entonces

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$$

es un estimador insesgado de  $\sigma^2$ , para  $n$  suficientemente grande.

**Corolario 2.2.** Sea  $\{X_i\}_{i=1}^n$  una m.a.i.i.d. con  $n < 30$  y  $X_i \sim N(\mu, \sigma^2)$  donde  $\sigma^2$  es desconocida,  $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$  y  $S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$  entonces

$$\tau = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}.$$

### Utilidad de estandarizar variables aleatorias

- Permite construir intervalos de confianza mediante la distribución normal estándar.
- Facilita el cálculo de valores- $p$  en las pruebas de hipótesis, mediante la tabla normal.
- Identifica outliers ( $|Z| > 3$ ) y compara medidas de distintas distribuciones.

### 2.4.2. Metodología para una prueba de hipótesis

Una prueba de hipótesis permite evaluar la compatibilidad de datos muestrales con afirmaciones respecto a alguna característica desconocida de una población. En esencia, se trata de un proceso para determinar si la evidencia empírica respalda o refuta una hipótesis preestablecida.

**Definición 2.27.** Una **hipótesis estadística** es una afirmación o conjetura acerca de la distribución de una o más variables aleatorias.

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**. En general, se contrastan dos hipótesis de acuerdo al siguiente esquema y notación:

$$H_0 : (\text{hipótesis nula}) \quad \text{vs} \quad H_1 : (\text{hipótesis alternativa}).$$

**Definición 2.28.** Una **prueba de hipótesis** es una regla para decidir si no se rechaza la hipótesis nula ( $H_0$ ) o si se rechaza en favor de la hipótesis alternativa ( $H_1$ ).

La **hipótesis nula**  $H_0$  es una conjetura sobre lo que se espera obtener de una característica poblacional; debe considerarse como verdadera y someterse a una refutación mediante datos muestrales.

La **hipótesis alternativa**  $H_1$  o  $H_A$  representa alguna forma de negación de la hipótesis nula, representa el efecto o la diferencia que el investigador busca detectar. También se le conoce como hipótesis del investigador.

La regla de decisión de una prueba de hipótesis se expresa en términos de una región crítica (RC), que está delimitada por el o los **valores críticos** (VC); la decisión de rechazar  $H_0$  se basa en determinar si el estadístico de prueba cae dentro de la RC.

**Definición 2.29.** El **estadístico de prueba** es una variable aleatoria calculada a partir de datos muestrales cuya distribución bajo  $H_0$  es conocida; permite cuantificar la discrepancia entre  $H_0$  y los datos observados.

El estadístico de prueba no es un parámetro poblacional, sino una herramienta inferencial. Su elección depende del parámetro bajo estudio y de los supuestos sobre la población (normalidad, tamaño muestral, etc.).

**Definición 2.30.** Una región crítica o **región de rechazo** es un subconjunto de valores de una muestra aleatoria para los cuales se rechaza la hipótesis nula.

## Procedimiento para una prueba de hipótesis

1. Plantear la hipótesis nula ( $H_0$ ), con la que se contrastan los datos de la muestra y la hipótesis alternativa ( $H_1$ ).
2. Determinar el estadístico de prueba  $T = T(X_1, \dots, X_n)$  y su distribución.
3. Fijar el o los valores críticos para establecer la región de rechazo.
4. Calcular el estadístico de prueba con base en la muestra.
5. Tomar una decisión: ¿Los datos evidencian que la hipótesis nula es falsas?
  - Si. Se rechaza  $H_0$  si y sólo si  $T \in RC$ .
  - No. No se rechaza  $H_0$ ;  $T \notin RC$ .

Cuando  $T \in RC$  se dice que la prueba es **significativa** y cuando  $T \notin RC$  se dice que la prueba es **no significativa**.

**Nota 2.5.** Rechazar  $H_0$  no necesariamente significa que la hipótesis nula sea falsa; pero la evidencia muestral con base en la cual se toma la decisión proporciona un **grado de confiabilidad** con el que puede procederse como si  $H_0$  fuese falsa.

Por otra parte, no rechazar  $H_0$  no implica demostrar su veracidad, sino la insuficiencia de evidencia en contra.

**Observación 2.7.** La región de rechazo no depende de la muestra  $x_1, \dots, x_n$ , es decir, aún antes de tomar la muestra, la región de rechazo tiene existencia propia. Los datos intervienen para tomar o no la decisión de rechazar  $H_0$ , lo cual se realiza con la región de rechazo, al comparar el valor de  $T$  con el conjunto  $RC$ .

## Errores de tipo I y tipo II

**Definición 2.31.** Los dos tipos de errores que pueden surgir al tomar una decisión en una prueba de hipótesis son los siguientes:

- El **Error tipo I** se comete al rechazar una hipótesis nula cuando en realidad es verdadera; se le conoce como **falso positivo**.
- El **Error tipo II** se comete por no rechazar  $H_0$  cuando en realidad es falsa; también es llamado **falso negativo**.

En cualquiera de los casos, ha habido una decisión errónea, por lo que es necesario diseñar buenas reglas de decisión que **minimicen** estos tipos de errores. Sin embargo, para cualquier tamaño de muestra dado, al tratar de disminuir un tipo de error, suele incrementarse el otro.

**Definición 2.32.** A la probabilidad de cometer el error tipo I se le denota por  $\alpha$ ;

$$\alpha = P(\text{"Error de tipo I"}) = P(\text{"Rechazar } H_0 \text{"} \mid \text{"}H_0 \text{ es verdadera"}).$$

A la probabilidad de cometer el error tipo II se le denota por la letra  $\beta$ , es decir,

$$\beta = P(\text{"Error de tipo II"}) = P(\text{"No rechazar } H_0 \text{"} \mid \text{"}H_0 \text{ es falsa"}).$$

A  $\alpha$  se le llama tamaño de la región crítica o tamaño de la región de rechazo. A esta probabilidad también se le conoce como **nivel de significancia** de la prueba.

Las probabilidades  $\alpha$  y  $\beta$  no son complementarias, es decir, no suman uno y se busca que ambas probabilidades sean pequeñas.

La única manera de reducir los dos tipos de errores es aumentando el tamaño de la muestra, lo que no siempre es posible.

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.

## Prueba de significancia $p$ -valor

En general, las pruebas de hipótesis se realizan con un cierto nivel de significancia  $\alpha$ , con  $0 \leq \alpha \leq 1$ , por tal razón muchas veces se denota a la región de rechazo  $RC$  con el subíndice  $\alpha$  para indicar que se trata de una prueba de nivel  $\alpha$ ; usualmente se toman valores  $\alpha = 0.05$ , o bien  $\alpha = 0.01$ .

Por otro lado, el **nivel de confianza** de una prueba se denota por  $(1 - \alpha)$  y se interpreta como el porcentaje  $100(1 - \alpha)\%$  de que se haya tomado una decisión correcta.

Si al diseñar la regla de decisión se elige el nivel de significancia  $\alpha = 0.05$ , entonces existe el 5% de posibilidad de que se rechace una hipótesis que resulta ser verdadera; es decir, se tiene una **confianza** de aproximadamente 95% de que se ha tomado la decisión correcta. En tal caso se dice que  $H_0$  ha sido rechazada al nivel de significancia 0.05, lo que significa que la hipótesis tiene una probabilidad de 0.05 de ser errónea.

**Definición 2.33.** El **p-valor** (o 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 tomar una decisión en esta prueba, se establece previamente un nivel de significancia  $\alpha$  y luego se calcula el p-valor:

Si el p-valor  $\leq \alpha$  se rechaza  $H_0$ ; en caso contrario, no se rechaza  $H_0$ .

En la Tabla 2.1 se visualizan los casos en los que se ha tomado una decisión correcta y los casos en los que se cometen los errores de tipo I y II.

Tabla 2.1: Errores Tipo I y Tipo II en pruebas de hipótesis

Decisión	$H_0$ Verdadera	$H_0$ Falsa
No rechazar Hipótesis nula	Acción Correcta $p\text{-valor} > \alpha$	Error Tipo II Falso negativo
Rechazar Hipótesis nula	Error Tipo I Falso positivo	Acción Correcta $p\text{-valor} \leq \alpha$

## Pruebas de bondad de ajuste

**Definición 2.34.** Dado un conjunto de observaciones, una **prueba de bondad de ajuste** es una técnica de validación estadística que permite verificar de qué distribución provienen dichas observaciones o cuál es la distribución que mejor ajusta a los datos.

Para realizar una prueba de bondad de ajuste es necesario utilizar la técnica de prueba de hipótesis y contrastar las siguientes dos hipótesis:

- $H_0$ : hipótesis nula: los datos provienen de la distribución  $F(z)$ .
- $H_1$ : hipótesis alternativa: los datos no provienen de la distribución  $F(z)$ .

La prueba de bondad de ajuste no solo permite verificar si una muestra proviene de una distribución teórica específica, sino que también puede extenderse para evaluar **independencia estadística** entre variables categóricas. En este contexto,

- $H_0$  Hipótesis nula: plantea que las variables son independientes.
- $H_1$  Hipótesis alternativa: sugiere asociación entre ellas.

Bajo el marco ya establecido para pruebas de hipótesis, esta técnica se adapta para contrastar patrones de independencia en tablas de contingencia, utilizando distribuciones como la **chi-cuadrada**  $\chi^2$ , o por medio de la **función devianza**.

# Capítulo 3

## Teoría de valores extremos univariante

La teoría de valores extremos (TVE) univariante proporciona el marco teórico para modelar el comportamiento estocástico de eventos raros pero de consecuencias significativas. En el contexto de la calidad del aire en la CDMX, esto se traduce en comprender las concentraciones máximas de ozono que exceden los umbrales de seguridad. Este capítulo desarrolla los fundamentos matemáticos que sustentan el análisis de extremos, comenzando con conceptos de equivalencia asintótica, pasando por la formulación del modelo general y culminando con la distribución límite que caracteriza estos fenómenos.

### 3.1. Funciones asintóticamente equivalentes

El estudio de valores extremos requiere comprender cómo se comportan las funciones en los límites de su dominio. Las funciones asintóticamente equivalentes son útiles para calcular dichos límites mediante la sustitución de funciones de aspecto complejo por versiones aproximadas más simples, preservando el comportamiento límite original.

Antes de formalizar el concepto de funciones asintóticamente equivalentes, se deben establecer algunos fundamentos matemáticos que sustentan esta teoría, que permitirán una mejor comprensión del tema.

Sea  $A \subseteq \mathbb{R}$ , un punto  $z \in \mathbb{R}$  se llama **punto de acumulación de  $A$** , si para todo  $\varepsilon > 0$ , existe algún punto  $x \in A$  con  $x \neq z$  tal que:  $\|x - z\| < \varepsilon$ .

El **conjunto derivado**  $A'$  de  $A$  es la colección de todos sus puntos de acumulación:

$$A' = \{z \in \mathbb{R} \mid z \text{ es un punto de acumulación de } A\}.$$

A continuación, se enuncia la primera definición de funciones asintóticamente equivalentes, la cual analiza la convergencia local de funciones cerca de un punto finito de acumulación  $x_0$  del dominio común  $A$ .

Este enfoque es útil para estudiar aproximaciones locales y límites en puntos específicos.

**Definición 3.1.** Sean  $A \subset \mathbb{R}$ ,  $f, g : A \rightarrow \mathbb{R}$  funciones y  $x_0 \in A'$ ; cuando  $x$  tiende a  $x_0$ , se dice que las funciones  $f$  y  $g$  son **asintóticamente equivalentes** o asintóticamente iguales cuando:

$$\lim_{x \rightarrow x_0} \frac{f(x)}{g(x)} = 1.$$

Se denota por  $f(x) \sim g(x)$  cuando  $x \rightarrow x_0$ .

La segunda definición examina la convergencia global cuando  $x \rightarrow \infty$ , este caso no requiere de dominios idénticos ni condiciones sobre puntos de acumulación.

Se aplica en análisis de crecimiento de funciones y complejidad asintótica.

**Definición 3.2.** Sean  $A, B \subset \mathbb{R}$ ,  $f : A \rightarrow \mathbb{R}$  y  $g : B \rightarrow \mathbb{R}$ ; se dice que las funciones  $f$  y  $g$  son **asintóticamente equivalentes** o asintóticamente iguales, si se cumple:

$$\lim_{x \rightarrow \infty} \frac{f(x)}{g(x)} = 1.$$

En tal caso, solo se denota por  $f(x) \sim g(x)$ .

Todas las funciones que cumplan con estas condiciones, serán funciones asintóticamente equivalentes; por lo que existe una gran variedad de ellas. Los siguientes teoremas mencionan las equivalencias asintóticas más utilizadas en la práctica.

**Teorema 3.1.** Si  $n \in \mathbb{N}$ ,  $a_0, a_1, \dots, a_n \in \mathbb{R}$  con  $a_n \neq 0$  y  $p : \mathbb{R} \rightarrow \mathbb{R}$  dada por:  $p(x) = a_n x^n + a_{n-1} x^{n-1} + \dots + a_1 x + a_0$ ; entonces cuando  $x \rightarrow \infty$ , las siguientes funciones son asintóticamente equivalentes:

$$\mathbf{a)} \quad p(x) \sim a_n x^n. \qquad \mathbf{b)} \quad \ln(p(x)) \sim n \ln x.$$

**Teorema 3.2.** Si  $f(x) \sim g(x)$ , se cumplen las siguientes *equivalencias asintóticas*:

**a)**  $g(x) \sim f(x)$ .

**b)**  $f(x)^r \sim g(x)^r$ , para cualquier exponente  $r$  (posiblemente negativo).

**c)** Si  $g(x) \rightarrow \infty$ ,  $\ln f(x) \sim \ln g(x)$  cuando  $x \rightarrow \infty$ .

## Equivalencias asintóticas notables

A continuación, se enuncian las principales funciones asintóticamente equivalentes:

**Teorema 3.3.** Si  $x \rightarrow 0$ , las siguientes funciones son asintóticamente equivalentes:

- |  |   |
|--|---|
| <p>a) <math>\tan x \sim x</math>.</p> <p>b) <math>\arctan x \sim x</math>.</p> <p>c) <math>\arcsin x \sim x</math>.</p> <p>d) <math>\exp(x) - 1 \sim x</math>.</p> | <p>e) <math>\ln^r(1+x) \sim x^r, r \in \mathbb{R}</math>.</p> <p>f) <math>\sin^r x \sim x^r, r \in \mathbb{R}</math>.</p> <p>g) <math>1 - \cos x \sim x^2/2</math>.</p> <p>h) <math>(1+x)^\alpha - 1 \sim \alpha x, \alpha \in \mathbb{R}</math>.</p> |
|--|---|

**Corolario 3.1.** Si  $x \rightarrow 1$ , se tienen las siguientes equivalencias asintóticas:

- a)  $\ln x \sim (x - 1)$ .      b)  $\exp(x^2 - 1) \sim (x^2 - 1)$ .

**Teorema 3.4.** Sea  $A \subset \mathbb{R}$  y  $f, g : A \rightarrow \mathbb{R}$ ; cuando  $x \rightarrow x_0$ ,  $f(x) \sim g(x)$  si y solo si para toda función  $h(x)$  se cumple:

$$\lim_{x \rightarrow x_0} f(x)h(x) = \lim_{x \rightarrow x_0} g(x)h(x).$$

**Dem:** Si  $f(x) \sim g(x)$  cuando  $x \rightarrow x_0$ , por definición,  $\lim_{x \rightarrow x_0} \frac{f(x)}{g(x)} = 1$ . Así,

$$\begin{aligned} \lim_{x \rightarrow x_0} f(x)h(x) &= \lim_{x \rightarrow x_0} \frac{f(x)}{g(x)} g(x)h(x) \\ &= \lim_{x \rightarrow x_0} \frac{f(x)}{g(x)} \cdot \lim_{x \rightarrow x_0} g(x)h(x) \\ &= \lim_{x \rightarrow x_0} g(x)h(x). \end{aligned}$$

El **teorema de equivalencia asintótica** (3.4) es especialmente útil en cálculo de límites indeterminados, series y aproximaciones asintóticas; apoyándose con las definiciones y teoremas de las funciones asintóticamente equivalentes.

Para aplicar este teorema y calcular límites indeterminados, se realiza lo siguiente:

1. Analizar el límite a resolver  $\lim_{x \rightarrow a} f(x)$  e identificar su forma indeterminada.
2. Buscar una función  $g(x)$  asintóticamente equivalente a la función en el límite que se está calculando:  $f(x) \sim g(x)$ . (Consultar las equivalencias asintóticas).
3. Se sustituye la función  $f(x)$  en el límite por otra asintóticamente equivalente a ella, siempre que sea un factor en el límite; se busca que el nuevo límite sea más sencillo de calcular.

### 3.1.1. Sucesiones asintóticamente equivalentes

Una sucesión de números reales es una función  $\Phi : \mathbb{N} \rightarrow \mathbb{R}$ ; usualmente se denotan por  $\{x_n\}$ ,  $\{y_n\}$ ,  $\{z_n\}$ , o simplemente por  $x_n$ ,  $y_n$ ,  $z_n$ . Para definir cuándo dos sucesiones son asintóticamente equivalentes sólo se van a considerar sucesiones de números reales; así nos referiremos a la *sucesión de números reales* simplemente como *sucesión*.

**Definición 3.3.** Dos sucesiones  $a_n$  y  $b_n$  son **asintóticamente equivalentes** si:

$$\lim_{n \rightarrow \infty} \frac{a_n}{b_n} = 1.$$

Se denota por  $a_n \sim b_n$ .

El **principio de sustitución** afirma que el límite de una sucesión convergente o divergente no se altera al sustituir uno de sus factores o divisores por otro asintóticamente equivalente. Este principio se enuncia en el siguiente teorema.

**Teorema 3.5.** Sean  $x_n, y_n$  dos sucesiones asintóticamente equivalentes ( $x_n \sim y_n$ ), y  $z_n$  una sucesión cualquiera; entonces se cumple que:

- a)  $x_n z_n$  es divergente si y solo si  $y_n z_n$  es divergente.
- b)  $x_n z_n$  es convergente si y solo si  $y_n z_n$  es convergente.

En tal caso, ambas sucesiones tienen el mismo límite:  $\lim_{n \rightarrow \infty} x_n z_n = \lim_{n \rightarrow \infty} y_n z_n$ .

**Definición 3.4.** Una sucesión  $a_n$  se denomina **infinitesimal** o **infinitésimo** si:

$$\lim_{n \rightarrow \infty} a_n = 0.$$

**Proposición 3.1.** Sean  $a_n$  y  $b_n$  sucesiones infinitesimales, se dice que  $a_n$  y  $b_n$  son **infinitésimos equivalentes**, denotado por  $a_n \sim b_n$ ; si se cumple:

$$\lim_{n \rightarrow \infty} \frac{a_n}{b_n} = 1.$$

**Teorema 3.6.** Si  $a_n$  es una sucesión infinitesimal, entonces:

- a)  $\sin(a_n) \sim a_n$ .
- b)  $\tan(a_n) \sim a_n$ .
- c)  $\arctan(a_n) \sim a_n$ .
- d)  $\arcsin(a_n) \sim a_n$ .
- e)  $1 - \cos(a_n) \sim a_n^2/2$ ,  $r \in \mathbb{R}$ .
- f)  $(1 - a_n)^\alpha - 1 \sim \alpha a_n$ .
- g)  $\exp(a_n) - 1 \sim a_n$ .
- h)  $\ln(1 + a_n) \sim a_n$ .

## 3.2. Formulación del modelo para máximos

Para una sucesión de v.a.i.i.d.  $X_1, X_2, \dots, X_n$ , donde  $n \geq 1$ , y con función de distribución  $F(z)$ ; el objetivo de esta sección es desarrollar el modelo que representa el marco teórico esencial de la teoría de valores extremos (TVE), el cual se enfoca en analizar el comportamiento asintótico del **máximo muestral**:

$$M_n = \text{máx}\{X_1, X_2, \dots, X_n\}.$$

Las v.a.  $X_i$  representan valores de un proceso de medición en una escala de tiempo regular. Por ejemplo, las mediciones por hora de la concentración de partículas de  $O_3$  en la atmósfera, o los promedios diarios registrados por las estaciones de monitoreo.

Por lo tanto,  $M_n$  representa el *máximo sobre  $n$  unidades de tiempo* dentro del proceso de observación. Si  $n$  es el número de observaciones tomadas durante un año, un mes o un día; entonces  $M_n$  corresponde al máximo anual, mensual o al máximo diario.

### 3.2.1. Aproximación mediante distribuciones límite

Teóricamente, se puede obtener la distribución de  $M_n$  en forma exacta para todos los valores de  $n$ , teniendo en cuenta las propiedades de independencia:

$$\begin{aligned} F_{M_n}(z) &= \mathbb{P}(M_n \leq z) = \mathbb{P}(\text{máx}\{X_1, X_2, \dots, X_n\} \leq z) \\ &= \mathbb{P}(X_1 \leq z, X_2 \leq z, \dots, X_n \leq z) \\ &= \mathbb{P}(X_1 \leq z) \times \mathbb{P}(X_2 \leq z) \times \dots \times \mathbb{P}(X_n \leq z) \\ &= [F(z)]^n = F^n(z). \end{aligned}$$

Y derivando, se podría obtener su función de densidad:  $f_{M_n}(z) = n[F(z)]^{n-1}f(z)$ .

Pero en la práctica, usualmente se desconoce la función de distribución  $F(z)$ ; aún así, se puede aproximar esta distribución utilizando los datos observados y sustituir esta estimación en la ecuación de  $F_{M_n}(z) = [F(z)]^n$ . Desafortunadamente, discrepancias muy pequeñas en la estimación de  $F$  pueden producir grandes diferencias para  $F^n(z)$ .

### Distribución límite no degenerada de $M_n$

Un enfoque alternativo es aceptar que  $F$  es desconocida y buscar aproximar *familias de modelos* a la distribución de  $F^n(z)$  que puedan estimarse basándose únicamente en los datos extremos; es decir, se busca una distribución límite que aproxime a  $F^n(z)$ .

Para este enfoque, se procede analizando el comportamiento de la distribución del máximo  $F_{M_n} = F^n$  cuando  $n \rightarrow \infty$ . Si definimos a  $\omega$  como el **extremo derecho** de  $F(z)$  tal que  $\omega(F) = \sup\{z : F(z) < 1\} \leq \infty$ ; diremos que  $M_n$  es una sucesión creciente que converge a  $\omega(F)$  con probabilidad 1.

Por lo tanto,  $M_n$  converge en distribución a una v.a. degenerada; sin embargo, esta dificultad se puede evitar realizando una re-normalización lineal de la variable  $M_n$ .

**Nota 3.1.** El método a seguir será similar a lo que ocurre con las sumas de v.a.i.i.d. y el teorema del límite central.

**Definición 3.5.** Para obtener una **distribución límite no degenerada de  $M_n$** , se deben elegir sucesiones constantes  $\{a_n > 0\}$  y  $\{b_n\}$ , de manera que

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

converja a una distribución no degenerada cuando  $n \rightarrow \infty$ .

Elecciones apropiadas de  $\{a_n\}$  y  $\{b_n\}$  estabilizan la *escala* y la *ubicación* de  $M_n^*$  a medida que  $n$  aumenta, evitando las dificultades que surgen con la variable  $M_n$ .

Por lo tanto, para la distribución  $F_{M_n^*}(z) = \mathbb{P}(M_n^* \leq z)$ , cuando  $n \rightarrow \infty$ , se busca:

$$\mathbb{P}\left(\frac{M_n - b_n}{a_n} \leq z\right) = F^n(a_n z + b_n) \xrightarrow{d} G(z),$$

donde  $G(z)$  sea una **distribución límite no degenerada de  $M_n^*$** .

**Nota 3.2.** En la Teoría de Valores Extremos los parámetros de *escala* y de *ubicación* se denotan por  $\mu$  y  $\sigma$ ; además se usa  $\xi$  para el parámetro de *forma*.

### 3.3. Distribuciones límite para valores extremos

En 1928, Fisher y Tippett formularon un teorema fundamental, que más tarde fue probado rigurosamente por Gnedenko en 1943. Este resultado permite aproximar la distribución de una normalización lineal del máximo muestral  $M_n$ , y aparece en el estudio de las posibles distribuciones límite para máximos normalizados de v.a.i.i.d.

Es decir, el teorema establece que los máximos muestrales reescalados  $(M_n - b_n)/a_n$  convergen a una variable que tiene una distribución límite dentro de una de las tres clases de familias denominadas I, II y III.

### 3.3.1. Teorema de valores extremos

**Teorema 3.7** (TVE). Si existen sucesiones de constantes  $\{a_n\} > 0$  y  $\{b_n\} \in \mathbb{R}$  para  $n \geq 1$ , tales que

$$\mathbb{P} \left\{ \frac{M_n - b_n}{a_n} \leq z \right\} = F^n(a_n z + b_n) \xrightarrow{d} G(z),$$

cuando  $n \rightarrow \infty$ , donde  $G(z)$  es una función de distribución no degenerada; para  $\xi > 0$ ,  $G(z)$  pertenece a alguna de las siguientes tres familias de distribuciones:

Tipo I:  $G(z) = \exp(-\exp(-z)), z \in \mathbb{R}.$

Tipo II:  $G(z) = \begin{cases} 0, & z \leq 0, \\ \exp(-z^{-\xi}), & z > 0. \end{cases}$

Tipo III:  $G(z) = \begin{cases} \exp(-(-z)^\xi), & z < 0, \\ 1, & z \geq 0. \end{cases}$

En conjunto, estas tres familias de distribuciones del Tipo I, II y III se denominan **distribuciones de valores extremos (DVE)**; y son ampliamente conocidas como las familias *Gumbel*, *Fréchet* y *Weibull*, respectivamente.

Es importante mencionar que el teorema de valores extremos no garantiza la existencia de un límite no degenerado para  $M_n$  y, cuando el límite existe, no establece cuál es. Además, las sucesiones de constantes  $\{a_n\} > 0$  y  $\{b_n\} \in \mathbb{R}$  no son únicas y existen distintas formas de encontrarlas.

**Nota 3.3.** La característica notable de este resultado es que cuando  $M_n$  puede ser estabilizada con sucesiones de constantes adecuadas  $\{a_n\} > 0$  y  $\{b_n\} \in \mathbb{R}$ , los tres tipos de distribuciones de valores extremos son los *únicos límites posibles* para las distribuciones de  $M_n^*$ , independientemente de la distribución  $F(z)$  para la población. Es en este sentido que el teorema proporciona un resultado análogo al teorema del límite central.

#### Densidades asociadas a la DVE

Las **densidades de las distribuciones de valores extremos** son las siguientes:

Tipo I (Gumbel):  $g(z) = \exp(-z) \exp(-\exp(-z)), z \in \mathbb{R}.$

Tipo II (Fréchet):  $g(z) = \xi z^{-(1+\xi)} \exp(-z^{-\xi}), z > 0.$

Tipo III (Weibull):  $g(z) = |\xi|(-z)^{\xi-1} \exp(-(-z)^\xi), z < 0.$

Las distribuciones de valores extremos son **unimodales**, es decir, hay una única moda y por tanto, el **nivel de asimetría** de estas distribuciones se describe mediante tres categorías conocidas:

- a) Distribuciones simétricas:  $(\gamma_1 \approx 0)$ .
- b) Distribuciones asimétricas positivas, o asimetría a la derecha:  $(\gamma_1 > 0)$ .
- c) Distribuciones asimétricas negativas, o asimetría a la izquierda:  $(\gamma_1 < 0)$ .

En cuanto al **sesgo de las densidades**, las densidades de Fréchet y Gumbel son sesgadas a la derecha; mientras que el sesgo de las densidades Weibull varía de acuerdo al siguiente criterio del *parámetro de forma*  $\xi$ :

Si  $\xi > -3.6$  son sesgadas a la izquierda.

Si  $\xi < -3.6$  presentan sesgo a la derecha.

Para  $\xi \approx -3.6$  son aproximadamente simétricas.

En la TVE cada distribución representa una familia de distribuciones según los valores del parámetro de *localización* o *ubicación*  $\mu$  y del parámetro de *escala*  $\sigma$ .

De manera que, el TVE puede presentarse frecuentemente de la siguiente forma:

**Teorema 3.8 (TVE).** Suponga que existen sucesiones de constantes  $\{a_n\} > 0$  y  $\{b_n\} \in \mathbb{R}$  para  $n \geq 1$  tales que

$$\mathbb{P} \left\{ \frac{M_n - b_n}{a_n} \leq z \right\} = F^n(a_n z + b_n) \rightarrow G(z),$$

cuando  $n \rightarrow \infty$ , donde  $G(z)$  es una función de distribución no degenerada, entonces  $G(z)$  pertenece a algunas de las siguientes tres familias de distribuciones:

$$\text{Tipo I: } G(z) = \exp \left\{ - \exp \left[ - \left( \frac{z - \mu}{\sigma} \right) \right] \right\}, \quad z \in \mathbb{R};$$

$$\text{Tipo II: } G(z) = \begin{cases} 0, & z \leq \mu, \\ \exp \left\{ - \left( \frac{z - \mu}{\sigma} \right)^{-\xi} \right\}, & z > \mu; \end{cases}$$

$$\text{Tipo III: } G(z) = \begin{cases} \exp \left\{ - \left[ - \left( \frac{z - \mu}{\sigma} \right) \right]^\xi \right\}, & z < \mu, \\ 1, & z \geq \mu; \end{cases}$$

para parámetros  $\sigma > 0$ ,  $\mu \in \mathbb{R}$ , y en el caso de las familias II y III,  $\xi > 0$ .

**Corolario 3.2.** Del TVE, si existen sucesiones  $a_n > 0$  y  $b_n \in \mathbb{R}$ , para todo punto  $z$  donde  $G(z)$  es una función continua y no degenerada, si se cumple la condición:

$$\lim_{n \rightarrow \infty} F^n(a_n z + b_n) = G(z),$$

entonces  $G(z)$  es una distribución de valores extremos (DVE).

En principio se deducen las siguientes equivalencias de formulaciones de sucesiones asintóticamente iguales:

$$\begin{aligned} \lim_{n \rightarrow \infty} F^n(a_n z + b_n) = G(z) &\iff \lim_{n \rightarrow \infty} n \ln F(a_n z + b_n) = \ln G(z) \\ &\iff \lim_{n \rightarrow \infty} n(-\ln F(a_n z + b_n)) = -\ln G(z). \end{aligned}$$

**Nota 3.4.** Cabe mencionar que, de estas formulaciones asintóticamente equivalentes, se puede verificar que para cada  $z$  fijo  $F(a_n z + b_n) \rightarrow 1$ .

Dando como resultado el siguiente lema.

**Lema 3.1.** Sean  $a_n > 0$  y  $b_n \in \mathbb{R}$  dos sucesiones de constantes,  $F(z)$  y  $G(z)$  dos funciones de distribución. Para cualquier  $z$  tal que  $0 \leq G(z) \leq 1$  se cumple que:

$$\lim_{n \rightarrow \infty} F^n(a_n z + b_n) = G(z) \iff \lim_{n \rightarrow \infty} n(1 - F(a_n z + b_n)) = -\ln G(z).$$

### Densidades asociadas a la DVE para máximos

Las densidades asociadas a las DVE para máximos, con parámetros  $\mu$ ,  $\sigma$  y  $\xi$ ; se obtienen derivando de manera directa cada distribución.

Para las **densidades asociadas a las DVE para máximos**, se denota:

$$\text{Tipo I : } \lambda(\mu, \sigma; z) = \frac{dG(z)}{dz} = \frac{1}{\sigma} \exp \left[ -\exp \left( \frac{\mu - z}{\sigma} \right) + \frac{\mu - z}{\sigma} \right], \quad z \in \mathbb{R}.$$

$$\text{Tipo II : } \phi(\mu, \sigma, \xi; z) = \frac{dG(z)}{dz} = \frac{\xi}{\sigma} \left( \frac{z - \mu}{\sigma} \right)^{-(1+\xi)} \exp \left[ -\left( \frac{z - \mu}{\sigma} \right)^{-\xi} \right], \quad z > \mu.$$

$$\text{Tipo III : } \psi(\mu, \sigma, \xi; z) = \frac{dG(z)}{dz} = \frac{\xi}{\sigma} \left( -\frac{z - \mu}{\sigma} \right)^{\xi-1} \exp \left[ -\left( -\frac{z - \mu}{\sigma} \right)^{\xi} \right], \quad z < \mu.$$

**Nota 3.5.** Si  $Q(x)$  es la **función cuantil**, variando su valor para cada tipo de distribución. Los cuantiles de probabilidad acumulada se obtienen al resolver la ecuación  $G(Q_p) = p$  para  $p \in (0, 1)$ , donde  $G(z)$  es la función de distribución de Gumbel, Fréchet y Weibull respectivamente.

Así, las expresiones para el **cuantil de probabilidad acumulada**  $\mathbf{p}$  son:

$$\text{Tipo I (Gumbel): } Q_p = \mu - \sigma \ln[-\ln(p)].$$

$$\text{Tipo II (Fréchet): } Q_p = \mu + \sigma[-\ln(p)]^{-\frac{1}{\xi}}.$$

$$\text{Tipo III (Weibull): } Q_p = \mu - \sigma[-\ln(p)]^{\frac{1}{\xi}}.$$

En las aplicaciones de la teoría de valores extremos, por lo general se adopta una de las tres familias de DVE para modelar eventos extremos, y luego se estiman los parámetros relevantes o de interés asociados a la distribución elegida. [Molina (2010)]

Desafortunadamente, el TVE no proporciona información sobre cómo discernir entre las tres familias de distribuciones, por lo que es necesario presentar una reformulación que permita la aplicación de técnicas de inferencia estadística para cuantificar la incertidumbre al momento de elegir una de ellas.

### 3.4. Distribución de valores extremos generalizado

Entre los años de 1954 y 1955, Von Mises y Jenkinson propusieron que las familias de distribuciones Gumbel, Fréchet y Weibull del TVE podían ser combinadas en una sola distribución con parametrización común; a este nuevo resultado se le conoce como la parametrización de Jenkinson-Von Mises o comúnmente conocida como la distribución de valores extremos generalizada (DVEG), o GEVD por sus siglas en inglés.

**Definición 3.6.** La forma de la familia de la **DVEG**, definida sobre el conjunto  $\{z : 1 + \xi((z - \mu)/\sigma) > 0\}$  donde  $\sigma > 0$  y  $\mu, \xi \in \mathbb{R}$ , está dada por:

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

donde  $\mu, \sigma, \xi$  son los parámetros de localización, de escala y de forma respectivamente.

El comportamiento de la cola de la DVEG está determinada por la especificación del parámetro de forma  $\xi$ , dependiendo del valor que tome este parámetro se tendrá:

La distribución Fréchet si  $\xi > 0$ .

La distribución Weibull si  $\xi < 0$ .

La distribución Gumbel si  $\xi = 0$ .

Esta última se interpreta como el límite de la DVEG cuando  $\xi \rightarrow 0$ .

**Nota 3.6.** Para representar que una variable aleatoria  $X$  sigue una distribución de valores extremos generalizada de parámetros  $\mu$ ,  $\sigma$  y  $\xi$ , se escribe:  $X \sim \text{VEG}(\mu, \sigma, \xi)$ . Además, por lo mencionado previamente se demuestra que:

$$X \sim \text{Gumbel}(\mu, \sigma) \quad \text{si y sólo si} \quad X \sim \text{VEG}(\mu, \sigma, 0).$$

El **soporte** de la DVEG está dado por los conjuntos de todos los  $z$  tales que:

$$\begin{aligned} z &\in \left( \frac{\mu - \sigma}{\xi}, +\infty \right) && \text{cuando } \xi > 0, \\ z &\in \left( -\infty, \frac{\mu - \sigma}{\xi} \right) && \text{cuando } \xi < 0, \\ z &\in (-\infty, +\infty) && \text{cuando } \xi = 0. \end{aligned}$$

### 3.4.1. Teorema de valores extremos generalizado

**Teorema 3.9** (TVEG). Suponga que existen sucesiones de constantes  $\{a_n\} > 0$  y  $\{b_n\} \in \mathbb{R}$  para  $n \geq 1$  tales que

$$\mathbb{P} \left\{ \frac{M_n - b_n}{a_n} \leq z \right\} = F^n(a_n z + b_n) \rightarrow G(z),$$

cuando  $n \rightarrow \infty$  y  $G(z)$  es una función de distribución no degenerada; entonces  $G(z)$  pertenece a alguna familia de DVEG:

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

definida sobre el conjunto  $\{z \mid 1 + \xi \left( \frac{z - \mu}{\sigma} \right) > 0\}$ , donde  $\sigma > 0$  y  $\mu, \xi \in \mathbb{R}$ .

**Teorema 3.10** (DVEG). La DVEG puede ser escrita de la siguiente forma:

$$G(z) = \begin{cases} \exp \left[ - \left[ 1 + \xi \left( \frac{z - \mu}{\sigma} \right) \right]^{-1/\xi} \right], & \text{si } \xi \neq 0, \\ \exp \left[ - \exp \left( \frac{\mu - z}{\sigma} \right) \right], & \text{si } \xi = 0. \end{cases}$$

### Densidades asociadas a la DVEG

Si  $G(z)$  es la DVEG para el caso  $\xi \neq 0$ , al derivar ambos lados se tiene lo siguiente:

$$\begin{aligned}
 \frac{dG(z)}{dz} &= -\exp\left\{-\left[1 + \xi\left(\frac{z-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}}\right\} \frac{d}{dz} \left[1 + \xi\left(\frac{z-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}} \\
 &= -\exp\left\{-\left[1 + \xi\left(\frac{z-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}}\right\} \left(-\frac{1}{\xi}\right) \left[1 + \xi\left(\frac{z-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}-1} \frac{d}{dz} \left[1 + \xi\left(\frac{z-\mu}{\sigma}\right)\right] \\
 &= \frac{1}{\xi} \exp\left\{-\left[1 + \xi\left(\frac{z-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}}\right\} \left[1 + \xi\left(\frac{z-\mu}{\sigma}\right)\right]^{-\left(\frac{1}{\xi}+1\right)} \left(\frac{\xi}{\sigma}\right) \\
 &= \frac{1}{\sigma} \left[1 + \xi\left(\frac{z-\mu}{\sigma}\right)\right]^{-\left(\frac{1}{\xi}+1\right)} \exp\left\{-\left[1 + \xi\left(\frac{z-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}}\right\}.
 \end{aligned}$$

La cual corresponde a la función de densidad del modelo DVEG para  $\xi \neq 0$ .

Si  $G(z)$  es la DVEG y  $\xi = 0$  entonces  $G(z)$  es la distribución Gumbel, cuya función de densidad  $G'(z)$  ya se obtuvo previamente.

Así, las **densidades del modelo de DVEG** con parámetros  $(\mu, \sigma, \xi)$  están dadas por:

**Si  $\xi \neq 0$** ; para todo  $z$  tal que  $1 + \xi\left(\frac{z-\mu}{\sigma}\right) > 0$ , la densidad es:

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

**Si  $\xi = 0$** ; para todo  $z \in \mathbb{R}$ , la densidad está dada por:

$$g(z) = \frac{1}{\sigma} \exp\left[-\exp\left(\frac{\mu-z}{\sigma}\right) + \frac{\mu-z}{\sigma}\right].$$

La distribución de valores extremos generalizada (DVEG) resulta particularmente útil en la práctica, pues la estimación del parámetro de forma  $\xi$  permite determinar directamente el tipo de modelo apropiado para los datos.

Además, es posible emplear métodos de inferencia estadística que proporcionen una medida cuantitativa de la incertidumbre asociada a la selección del modelo, como se desarrollará posteriormente.

**Nota 3.7.** La distribución de valores extremos generalizada tiene propiedades muy particulares que la caracterizan: la media existe si  $\xi < 1$ , mientras que la varianza existe si  $\xi < \frac{1}{2}$ . Es decir, el  $k$ -ésimo momento existe si  $\xi < \frac{1}{k}$ .

### 3.4.2. Max-estabilidad

La max-estabilidad es una propiedad fundamental en la teoría de valores extremos, estrechamente relacionado con el teorema de valores extremos generalizado.

**Definición 3.7.** Una función de distribución  $F(z)$  se dice que es **max-estable** si, para cada  $n \geq 2$ , existen sucesiones de constantes  $a_n > 0$  y  $b_n \in \mathbb{R}$  tales que:

$$F^n(a_n z + b_n) = F(z).$$

Considérese un conjunto de v.a.i.i.d.  $X_1, \dots, X_n$  con función de distribución común  $F(z)$ . El máximo  $M_n = \max\{X_1, \dots, X_n\}$  sigue entonces una distribución  $F^n(z)$ .

La condición de max-estabilidad establece que existen constantes de normalización  $a_n > 0$  y  $b_n \in \mathbb{R}$  tales que, tras una transformación lineal del máximo:

$$\frac{M_n - b_n}{a_n} \sim F(z),$$

es decir, el **máximo normalizado conserva la misma distribución que las variables originales**. Esta notable propiedad caracteriza a las distribuciones límite en el análisis de valores extremos.

**Teorema 3.11.** Una función de distribución es **max-estable** si y sólo si corresponde a una distribución de valores extremos generalizada DVEG.

La demostración de que toda DVEG es max-estable puede obtenerse mediante operaciones algebraicas elementales. Sin embargo, la prueba del recíproco, que toda distribución max-estable es DVEG, requiere herramientas avanzadas de análisis funcional que exceden los objetivos de este trabajo, por lo que se omite su desarrollo.

Cabe destacar que la DVEG unifica las tres familias de distribuciones Gumbel, Fréchet y Weibull del teorema de valores extremos. Esta unificación permite una formulación alternativa del resultado anterior.

**Teorema 3.12.** Una distribución es max-estable si y sólo si es una distribución de valores extremos DVE.

Este resultado establece que, cuando la distribución  $F(z)$  de la población sigue una DVE, el máximo de muestras independientes preserva el mismo tipo distribucional bajo transformaciones lineales adecuadas.



# Capítulo 4

## Métodos de estimación paramétrica de la DVEG

Según [Coles et al. (2001)] se han propuesto muchas técnicas para la estimación de los parámetros del modelo de valores extremos. Entre ellas, se incluyen técnicas gráficas basadas en probabilidad, el modelo de bloques, el método de momentos, procedimientos donde los parámetros se estiman como funciones especificadas de las estadísticas de orden, así como los métodos basados en máxima verosimilitud.

Cada técnica tiene sus ventajas y sus desventajas; pero la utilidad y la adaptabilidad a la construcción de modelos complejos de técnicas basadas en la verosimilitud hacen que este enfoque sea particularmente atractivo para estimar los parámetros de la DVEG,

Una **serie temporal** se define como un conjunto de observaciones de una variable que consiste en mediciones secuenciales en el tiempo, cuantificada en momentos específicos o períodos determinados. Estas observaciones suelen registrarse en intervalos fijos: horarios, diarios, semanales, mensuales o anuales; aunque se pueden aplicar en cualquier otro periodo.

La dinámica de una serie temporal suele explicarse mediante la interacción de cuatro elementos constitutivos: tendencia, variación cíclica, componente estacional y fluctuaciones irregulares, cuya combinación genera los valores observados.

Cabe destacar que estos componentes de las series temporales no siempre se presentan solos, frecuentemente interactúan de manera conjunta, pudiendo presentarse de manera combinada o incluso, pueden presentarse todas juntas.

A continuación, se definen a grandes rasgos cada uno de ellos:

- **Tendencia:** Es un un cambio progresivo hacia niveles superiores o inferiores que persiste durante un extenso lapso temporal, manifestándose a través de múltiples

períodos de observación. Estas tendencias pueden adoptar formas lineales, no lineales o presentar ausencia de patrón direccional.

- **Variación cíclica:** La serie de tiempo muestra un comportamiento de tendencias periódicas, caracterizadas por oscilaciones recurrentes sobre la línea de tendencia subyacente que superan el umbral anual.
- **Componente estacional:** Cuando existen patrones de fluctuación periódicos con ciclos completos que no exceden el intervalo de un año.
- **Fluctuaciones irregulares:** Componente que captura las variaciones aleatorias observables, no atribuibles a los demás factores identificados.

De acuerdo al TVEG, el máximo  $M_n$  de una sucesión de v.a.i.i.d. se aproxima a una distribución  $G(z)$  de una familia de DVEG, es decir; para  $n$  suficientemente grande,

$$\mathbb{P} \left\{ \frac{M_n - b_n}{a_n} \leq z \right\} \approx G(z).$$

El problema práctico surge en la estimación de las sucesiones constantes de normalización  $a_n$  y  $b_n$ , el cual se puede resolver aplicando la siguiente metodología.

**Lema 4.1.** Si  $Z = \frac{M_n - b_n}{a_n}$  tal que  $Z \sim \text{VEG}(\mu, \sigma, \xi)$  con  $\mathbb{P}(Z \leq z) = G(z)$ ; para  $y = a_n z + b_n$  y por el teorema de convergencia de familias, se sigue que:

$$\begin{aligned} \mathbb{P} \left\{ \frac{M_n - b_n}{a_n} \leq z \right\} \approx \mathbb{P}\{Z \leq z\} &\iff \mathbb{P} \left\{ \frac{M_n - b_n}{a_n} \leq \frac{y - b_n}{a_n} \right\} \approx \mathbb{P} \left\{ Z \leq \frac{y - b_n}{a_n} \right\} \\ &\iff \mathbb{P}\{M_n \leq y\} \approx \mathbb{P} \left\{ Z \leq \frac{y - b_n}{a_n} \right\} \\ &\iff \mathbb{P}\{M_n \leq y\} \approx G \left( \frac{y - b_n}{a_n} \right). \end{aligned}$$

Dado que  $G \left( \frac{y - b_n}{a_n} \right) = G^*(y)$ , se tiene finalmente:

$$\mathbb{P}\{M_n \leq y\} \approx G^*(y),$$

donde  $G^*(y)$  es otro miembro de la familia de DVEG.

Este resultado asegura que, si la distribución del máximo  $M_n^* = (M_n - b_n)/a_n$  se puede aproximar por un miembro de la familia de DVEG para  $n$  suficientemente grande, entonces la distribución de  $M_n$  también puede ser aproximada por un miembro diferente de la familia de DVEG.

Por otra parte, como los parámetros de la distribución tienen que ser estimados, en la práctica no es necesario que los parámetros de la distribución  $G(z)$  sean iguales a los de la distribución  $G^*(z)$ .

**Nota 4.1.** El vector de parámetros  $\theta = (\mu, \sigma, \xi) \in \mathbb{R} \times \mathbb{R}^+ \times \mathbb{R}$  consiste del parámetro de localización  $\mu$ , el parámetro de escala  $\sigma$  y del parámetro de forma  $\xi$ .

Si se tiene una muestra  $X_1, \dots, X_n$  de v.a.i.i.d. de  $G(z; \theta)$  se pueden usar métodos estándar de estimación paramétrica. Sin embargo, la hipótesis de que  $X_i$  tienen exactamente una DVEG puede no ser realista en la mayoría de los casos.

## 4.1. Método de máximos por bloques

El **método de máximos por bloques** consiste en agrupar los datos en  $k$  bloques de igual tamaño o longitud y posteriormente ajustar la DVEG al conjunto de los máximos correspondientes a cada uno de los bloques.

La selección del tamaño de bloque ( $k$ ) es crucial y debe adaptarse a la periodicidad característica del fenómeno bajo estudio. La elección de bloques muy pequeños afectará la calidad de la aproximación del modelo.

Considere una colección de datos  $X_1, \dots, X_n$  que se agrupan en  $k$  conjuntos disjuntos de datos consecutivos llamados bloques  $\mathcal{B}^{(1)}, \dots, \mathcal{B}^{(k)}$  y de igual longitud.

Si se interpreta el parámetro  $s$  como el tiempo, entonces cada bloque  $\mathcal{B}^{(i)}$  contiene la información correspondiente a un período fijo de tiempo  $|\mathcal{B}^{(i)}| = s$ ; a menudo los bloques se eligen para un período de un mes o de un año.

La elección del período  $s$  compensa las variaciones internas, optimiza el balance entre estacionariedad y ofrece un tamaño muestral efectivo  $k = \lfloor n/s \rfloor$ .

Por lo tanto, los datos originales se distribuyen en:

$$\begin{aligned} \mathcal{B}^{(1)} &= (X_1^{(1)}, \dots, X_s^{(1)}) \\ \mathcal{B}^{(2)} &= (X_1^{(2)}, \dots, X_s^{(2)}) \\ &\vdots \\ \mathcal{B}^{(k)} &= (X_1^{(k)}, \dots, X_s^{(k)}) \end{aligned}$$

Donde se supone que  $\mathcal{B}^{(1)}, \dots, \mathcal{B}^{(k)}$  son independientes e idénticamente distribuidos, pero las componentes de cada vector  $\mathcal{B}^{(i)}$  pueden ser dependientes.

La m.a.i.i.d. para  $G(\boldsymbol{\theta}; z)$  sobre la cual se hará inferencia es:

$$Z_i = \max \left( X_1^{(i)}, \dots, X_s^{(i)} \right), \quad i = 1, \dots, k.$$

La DVEG proporciona un modelo para la distribución de cada  $Z_i$ .

**Ejemplo 4.1.** En la Ciudad de México, las estaciones de monitoreo atmosférico realizan mediciones por hora de la concentración de partículas de ozono troposférico y están ubicadas en diferentes zonas de la ciudad.

Supóngase que, por las condiciones geográficas de la zona de Cuajimalpa, estamos interesados en los datos marcados por esta estación en el año 2023 ( $n = 8760$ ).

- a) Si se quiere saber cuáles fueron los registros máximos por día en esa zona, el período será  $s = 24$  (horas en un día) y habrá  $k = 365$  bloques.
- b) En cambio, para conocer los registros máximos por mes, se tendrá un período de  $s = 720$  (horas en un mes) y habrá  $k = 12$  bloques.

## 4.2. Estimación por máxima verosimilitud

Sea  $g(\boldsymbol{\theta}; z)$  la función de densidad de  $G(\boldsymbol{\theta}; z)$  y sean  $Z_1, \dots, Z_k$  v.a.i.i.d. con DVEG.

**Para el caso  $\xi \neq 0$ :**

- La **función de verosimilitud** para esta dada por:

$$\begin{aligned} L(\boldsymbol{\theta}) &= g_{Z_1, Z_2, \dots, Z_k}(\boldsymbol{\theta}; z_1, z_2, \dots, z_n) = g_{Z_1}(\boldsymbol{\theta}; z_1) \cdots g_{Z_n}(\boldsymbol{\theta}; z_n) \\ &= \left( \frac{1}{\sigma} \right)^n \prod_{i=1}^n \left[ 1 + \xi \left( \frac{z_i - \mu}{\sigma} \right) \right]^{-\left(1 + \frac{1}{\xi}\right)} \exp \left\{ - \sum_{i=1}^n \left[ 1 + \xi \left( \frac{z_i - \mu}{\sigma} \right) \right]^{-\frac{1}{\xi}} \right\}. \end{aligned}$$

- Por tanto, la **función de log-verosimilitud** es:

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

con

$$1 + \xi \left( \frac{z_i - \mu}{\sigma} \right) > 0, \quad \text{para } i = 1, \dots, k.$$

**Nota 4.2.** Para las combinaciones de parámetros en las cuales no se cumple la última desigualdad, indica que al menos uno de los datos  $z_i$  se encuentra fuera del rango de valores admisibles, es decir, cae más allá del punto final de la distribución en cuestión; lo que conduce a una *verosimilitud nula*  $L(\boldsymbol{\theta}) = 0$ , y una *log-verosimilitud indefinida*  $\ell(\boldsymbol{\theta}) \rightarrow -\infty$ , invalidando así dicha configuración paramétrica.

■ Las **derivadas parciales** con respecto a  $\mu, \sigma$  y  $\xi$  respectivamente de la función de log-verosimilitud se expresan algebraicamente como:

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

Para el caso  $\xi = 0$  (Gumbel):

■ La **función de verosimilitud** esta dada por:

$$\begin{aligned}L(\boldsymbol{\theta}) &= g_{Z_1, Z_2, \dots, Z_n}(\boldsymbol{\theta}; z_1, z_2, \dots, z_n) = g_{Z_1}(\boldsymbol{\theta}; z_1) \cdots g_{Z_n}(\boldsymbol{\theta}; z_n) \\ &= \left(\frac{1}{\sigma}\right)^n \prod_{i=1}^n \exp \left[ -\exp \left(\frac{\mu - z_i}{\sigma}\right) + \frac{\mu - z_i}{\sigma} \right] \\ &= \left(\frac{1}{\sigma}\right)^n \exp \left\{ \sum_{i=1}^n \left[ -\exp \left(\frac{\mu - z_i}{\sigma}\right) \right] + \sum_{i=1}^n \left[ \frac{\mu - z_i}{\sigma} \right] \right\}.\end{aligned}$$

■ La **función log-verosimilitud** para este caso es:

$$\ell(\boldsymbol{\theta}) = -n \ln(\sigma) - \sum_{i=1}^n \left[ \exp \left(-\frac{z_i - \mu}{\sigma}\right) \right] - \sum_{i=1}^n \left[ \frac{z_i - \mu}{\sigma} \right].$$

■ Las **derivadas parciales** con respecto a  $\mu$  y  $\sigma$  de la función de log-verosimilitud son:

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

Igualando a 0 estas derivadas parciales para el caso Gumbel se obtienen:

$$\begin{aligned}0 &= n - \sum_{i=1}^n \exp \left( -\frac{z_i - \mu}{\sigma} \right). \\ 0 &= n + \sum_{i=1}^n \left( \frac{z_i - \mu}{\sigma} \right) \left[ \exp \left( -\frac{z_i - \mu}{\sigma} \right) - 1 \right].\end{aligned}$$

Nótese que no hay una solución explícita de  $\mu$  y  $\sigma$ , y cuando  $\xi \neq 0$  es aún más complicado, de modo que la maximización de las ecuaciones log-verosimilitud no tienen solución analítica; sin embargo, para un conjunto de datos proporcionado, la maximización se obtiene de manera directa usando algoritmos numéricos de optimización.

Al aplicar este método, se deben considerar un par de dificultades:

1. Al usar algoritmos numéricos, se debe asegurar que las combinaciones de parámetros cumplan con las evaluaciones de  $\ell(\boldsymbol{\theta})$  para ambos casos.
2. Las condiciones de regularidad no son satisfechas por la DVEG, lo que implica que el método de máxima verosimilitud no pueda aplicarse automáticamente.

**Nota 4.3.** Las condiciones de regularidad se requieren para que las propiedades asintóticas asociadas con el EMV sean válidas. Estas condiciones no son satisfechas por la DVEG porque los extremos  $(\mu - \sigma)/\xi$  son una función de los parámetros; es decir, para  $\xi < 0$  es un punto final superior, y cuando  $\xi > 0$  es un punto final inferior de la DVEG.

Se estudió este problema a detalle y se demostraron los siguientes resultados:

- Cuando  $\xi > -0.5$  los estimadores de máxima verosimilitud son regulares, es decir, tienen las propiedades asintóticas usuales.
- Cuando  $-1 < \xi < -0.5$  es posible obtener los estimadores de máxima verosimilitud pero no tienen las propiedades asintóticas usuales.
- Cuando  $\xi < -1$  es posible que no puedan obtenerse los estimadores de máxima verosimilitud.

El caso cuando  $\xi \leq -0.5$ , la distribución presenta una cola superior extremadamente corta y acotada. Sin embargo, esta situación es inusual en la práctica de la teoría de valores extremos, por lo que las limitaciones teóricas del enfoque de máxima verosimilitud tienen poca relevancia en las aplicaciones reales.

Cuando  $\xi$  cumple con las condiciones establecidas, los estimadores  $\hat{\boldsymbol{\theta}} = (\hat{\mu}, \hat{\sigma}, \hat{\xi})$  se aproximan a una distribución normal multivariada centrada en los parámetros reales  $(\mu, \sigma, \xi)$ , con matriz de covarianza dada por la inversa de la matriz de información de Fisher evaluada en los EMV.

Aunque esta matriz se puede calcular analíticamente, es más sencillo calcularla numéricamente. Los intervalos de confianza aproximados y otras formas de inferencia se siguen inmediatamente de la normalidad asintótica del estimador.

### 4.3. Inferencia para los niveles de retorno

Considérese el problema de estimación clásico en el cual se tiene una muestra aleatoria  $X_1, \dots, X_n$  con densidad  $G(\boldsymbol{\theta}; z)$  y supóngase que  $\hat{\boldsymbol{\theta}}$  es el estimador de máxima verosimilitud de  $\boldsymbol{\theta}$ . En esta situación es inmediato obtener un **estimador de los cuantiles**; dado que la DVEG es invertible se tiene que para cualquier  $p \in (0, 1)$  el  $p$ -cuantil está dado por  $Q_p = G^{-1}(\boldsymbol{\theta}; p)$ .

Así, por el principio de invarianza de los estimadores de máxima verosimilitud, un estimador natural para  $Q_p$  basado en la muestra  $X_1, \dots, X_n$  es

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

#### Estimación del p-cuantil de la DVEG

##### ■ Cuando $\xi \neq 0$

La fórmula  $G(Q_p) = p$  está dada por

$$\exp \left\{ - \left[ 1 + \xi \left( \frac{Q_p - \mu}{\sigma} \right) \right]^{-\frac{1}{\xi}} \right\} = p.$$

Por lo tanto, al despejar  $Q_p$ , se obtiene el cuantil por defecto de la DVEG:

$$Q_p = \mu - \frac{\sigma}{\xi} \{ 1 - [-\ln(p)]^{-\xi} \}.$$

■ Cuando  $\xi = 0$

La fórmula  $G(Q_p) = p$  está dada por

$$\exp \left\{ - \exp \left[ - \left( \frac{Q_p - \mu}{\sigma} \right) \right] \right\} = p.$$

De donde, se obtiene el cuantil por defecto para la distribución Gumbel:

$$Q_p = \mu - \sigma \ln[-\ln(p)].$$

Por lo tanto los cuantiles por defecto de la DVEG están dados por:

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

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:

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

## Diagrama de retorno

Si se define  $y_p = -\ln(1-p)$ , los cuantiles por exceso se pueden expresar como:

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

De esto último, se deduce que si  $z_p$  se grafica contra  $y_p$  en una escala logarítmica o equivalentemente, si  $z_p$  se grafica contra  $\ln(y_p)$ , se obtiene un **diagrama de retorno**.

### Interpretación del diagrama de retorno

- Si  $\xi = 0$  la gráfica obtenida es una línea recta.
- Si  $\xi < 0$  la gráfica es cóncava con límite asintótico en  $\mu - \frac{\sigma}{\xi}$  cuando  $p \rightarrow 0$ .
- Si  $\xi > 0$  la gráfica es convexa y no tiene cota finita.

Debido a la simplicidad de la interpretación y a la elección de la escala logarítmica, se comprime la cola de la distribución permitiendo resaltar los efectos de la extrapolación a valores extremos, lo cual es útil para representar y validar el modelo.

## 4.4. Métodos gráficos de diagnóstico

El proceso de modelación estadística tiene como objetivo fundamental encontrar una representación adecuada de fenómenos reales mediante estructuras matemáticas que permitan realizar inferencias confiables. Estas inferencias dependen críticamente del modelo seleccionado, lo que hace esencial lograr el mejor ajuste posible a los datos observados, incluso cuando esto se limite al tamaño de la muestra.

Sean  $X_1, X_2, \dots, X_n$  una muestra aleatoria con función de distribución común  $F(x)$  y si consideramos una *realización ordenada* de observaciones obtendremos la secuencia  $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$ , donde cada  $x_{(i)}$  representa el  $i$ -ésimo valor en esta ordenación.

Una propiedad de las *estadísticas de orden* es que exactamente  $i$  observaciones son menores o iguales a  $x_{(i)}$ . Por lo que una estimación empírica de la probabilidad acumulada en ese punto es  $\tilde{F}(x_{(i)}) = i/n$ .

**Nota 4.4.** En la práctica, se prefiere usar la modificación  $\tilde{F}(x_{(i)}) = i/(n+1)$ , que evita la estimación perfecta  $\tilde{F}(x_{(n)}) = 1$ . Esto lleva a la siguiente definición.

**Definición 4.1.** Dada una muestra ordenada de observaciones independientes

$$x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)},$$

de una población con función de distribución  $F(x)$ , se define la **función de distribución empírica** por

$$\tilde{F}(x) = \frac{i}{n+1} \quad \text{para } x_{(i)} \leq x < x_{(i+1)}.$$

### 4.4.1. Gráfica de probabilidad (P-P)

Suponga que  $\hat{F}(x)$  es una estimación de  $F(x)$  que se obtuvo por el método de máxima verosimilitud, de modo que la distribución empírica  $\tilde{F}(x)$  sirve como referencia para validar el modelo candidato. Cuando  $\hat{F}(x)$  estima adecuadamente  $F(x)$ , se espera congruencia entre  $\tilde{F}(x)$  y  $\hat{F}(x)$ , principio que fundamenta múltiples pruebas de bondad de ajuste mediante comparación directa entre ambas funciones.

**Definición 4.2.** Dada una muestra ordenada de observaciones independientes de una población  $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$ , con función de distribución estimada  $\hat{F}(x)$ , la gráfica de probabilidad, gráfica probabilidad-probabilidad o **gráfica P-P** consiste del siguiente conjunto de puntos

$$\left\{ \left( \hat{F}(x_{(i)}), \frac{i}{n+1} \right) : i = 1, \dots, n \right\}.$$

### Interpretación del Gráfico P-P

- Si  $\hat{F}(x)$  es un modelo razonable para la distribución de la población, se espera que los puntos en la gráfica de probabilidad se alineen cerca de la recta identidad o línea identidad.

- Las desviaciones sustanciales de la linealidad proporcionan incongruencias entre el modelo propuesto y el comportamiento real de los datos, sugiriendo que  $\hat{F}(x)$  no captura correctamente las características de la distribución subyacente.

### 4.4.2. Gráfica cuantil (Q-Q)

**Definición 4.3.** Dada una muestra ordenada de observaciones independientes

$$x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)},$$

de una población con función de distribución estimada  $\hat{F}(x)$ , la gráfica cuantil, gráfica cuantil-cuantil o **gráfica Q-Q** consiste del siguiente conjunto de puntos

$$\left\{ \left( \hat{F}^{-1} \left( \frac{i}{n+1} \right), x_{(i)} \right) : i = 1, \dots, n \right\}.$$

### Interpretación del Gráfico Q-Q

- Si  $\hat{F}(x)$  es una estimación razonable de  $F(x)$ , los puntos en el gráfico cuantil-cuantil (Q-Q) deben distribuirse alrededor de la recta identidad.

- Para la gráfica Q-Q es común agregar intervalos de confianza para los cuantiles para considerar la variabilidad de las observaciones en el ajuste.

**Nota 4.5.** La gráfica de probabilidad y la gráfica de cuantiles contienen la misma información expresada en una escala diferente. Esta diferencia en escalamiento puede afectar la percepción visual del ajuste: un modelo que parece adecuado en una representación podría mostrar inconsistencias evidentes en la otra.

### Validación de un modelo VEG

**Definición 4.4.** Sean  $z_{(1)} \leq z_{(2)} \leq \dots \leq z_{(n)}$  una realización de máximos por bloques ordenada. La **función de distribución empírica** evaluada en  $z_{(i)}$  viene dada por:

$$\tilde{G}(z_{(i)}) = \frac{i}{n+1}.$$

Al sustituir los parámetros estimados en la DVEG se obtiene el **modelo estimado**:

$$\hat{G}(z_{(i)}) = \exp \left\{ - \left[ 1 + \hat{\xi} \left( \frac{z_{(i)} - \hat{\mu}}{\hat{\sigma}} \right) \right]^{-\frac{1}{\hat{\xi}}} \right\}.$$

Si el modelo de valores extremos generalizados funciona bien, entonces para cada  $i$ :

$$\hat{G}(z_{(i)}) \approx \tilde{G}(z_{(i)})$$

- La gráfica de probabilidad en el modelo VEG consiste en los puntos

$$\left\{ \left( \tilde{G}(z_{(i)}), \hat{G}(z_{(i)}) \right) : i = 1, \dots, n \right\},$$

Cualquier desviación sustancial de la línea identidad indica alguna falla en el modelo.

- Por otra parte, la gráfica cuantil, consiste en los puntos

$$\left\{ \left( \hat{G}^{-1} \left( \frac{i}{n+1} \right), z_{(i)} \right) : i = 1, \dots, n \right\},$$

donde

$$\hat{G}^{-1} \left( \frac{i}{n+1} \right) = \hat{\mu} - \frac{\hat{\sigma}}{\hat{\xi}} \left[ 1 - \left\{ -\ln \left( \frac{i}{n+1} \right) \right\}^{-\hat{\xi}} \right].$$

Las desviaciones de la linealidad en el gráfico Q-Q también indican una falla del modelo.

#### 4.4.3. Gráfica de nivel de retorno

Por otro lado, para la DVEG se tiene una gráfica particular que se utiliza como presentación del modelo estimado y para la validación de ajuste del mismo.

Dicha gráfica recibe el nombre de **gráfica de nivel de retorno** y consiste en el lugar geométrico de los puntos

$$\{(\ln(y_p), \hat{z}_p) : 0 < p < 1\},$$

donde

$$\hat{z}_p = \begin{cases} \hat{\mu} - \frac{\hat{\sigma}}{\hat{\xi}} \left[ 1 - y_p^{-\hat{\xi}} \right], & \hat{\xi} \neq 0 \\ \hat{\mu} - \hat{\sigma} \ln(y_p), & \hat{\xi} = 0. \end{cases}$$

Se pueden agregar intervalos de confianza correspondientes a cada punto de la gráfica de nivel de retorno para aumentar su capacidad informativa. La gráfica es lineal para el modelo Gumbel, cóncava para el caso Weibull y convexa para el caso Fréchet.

**Nota 4.6.** Para completar un diagnóstico de métodos gráficos se contrasta la **función de densidad** de probabilidad del modelo ajustado con el **histograma** de los datos. Este último gráfico es generalmente menos informativo que las gráficas anteriores, ya que la forma de un histograma puede variar sustancialmente con la elección de los intervalos de agrupación o intervalos de clase.

El uso de modelos gráficos de diagnóstico son esencialmente útiles para la valuación de modelos. Además, la visualización del comportamiento de la distribución de los datos complementa la validación de los resultados obtenidos por las herramientas teóricas estadísticas explicadas en esta tesis.

# Capítulo 5

## Aplicación: Concentraciones máximas de ozono

En este capítulo se aplica la metodología de la TVE al estudio de las concentraciones máximas diarias de ozono registradas en la Ciudad de México y su zona metropolitana. Los datos analizados provienen de estaciones de monitoreo atmosférico distribuidas en distintas regiones de esta área urbana, las cuales miden de manera continua los niveles por hora de contaminantes críticos, entre ellos el ozono ( $O_3$ ). Se realiza un breve análisis descriptivo de los datos, seguido de la aplicación de la metodología de estimación de parámetros y validación de modelos desarrollada en el capítulo anterior.

Para la implementación de los métodos estadísticos y la generación de gráficos de diagnóstico, se utilizó el software `R-Studio`, con un enfoque particular en el paquete `evd` (*Extreme Value Distributions*), el cual proporciona funciones especializadas para el análisis de extremos. Los resultados obtenidos permitirán evaluar el comportamiento de las concentraciones máximas de ozono por día y su relación con eventos extremos en la zona de estudio.

### 5.1. Descripción del problema

La contaminación atmosférica se define como la presencia en el aire de sustancias o formas de energía que impliquen riesgos para la salud o el medio ambiente. Aunque existen fuentes naturales, la contaminación antropogénica constituye gravemente el problema actual; los procesos de combustión, especialmente el uso de combustibles fósiles, son la principal fuente de emisión de contaminantes primarios como dióxido de carbono ( $CO_2$ ), óxidos de nitrógeno ( $NO_x$ ), partículas ( $PM_{2.5}$ ,  $PM_{10}$ ) y compuestos orgánicos volátiles (COV).

En la Ciudad de México, las fuentes antropogénicas se deben a las grandes industrias y a las aglomeraciones urbanas con múltiples emisores. Además, la dispersión y evolución de los contaminantes dependen de factores atmosféricos y topográficos, lo que determina su distribución espacial y temporal. Históricamente, el monitoreo de los niveles de inmisión<sup>1</sup> ha sido fundamental para estudiar sus efectos en la salud pública y diseñar políticas de control ambiental.

El ozono troposférico ( $O_3$ ) es un contaminante secundario que se forma a partir de reacciones fotoquímicas entre  $NO_x$  y COV en la atmósfera. El  $O_3$  es uno de los contaminantes con mayor importancia y atención a nivel mundial, debido a que tiene efectos adversos en el ambiente y en la salud humana; existen grupos de población que son más susceptibles a la exposición de este contaminante, como son los adultos mayores, niños, embarazadas, personas que trabajan en exteriores, así como aquellas que padecen enfermedades respiratorias.

Existe una cantidad considerable de estudios que aportan evidencia sobre los efectos de la exposición a ozono a corto y largo plazo. De acuerdo a la OMS (2021), los efectos a corto plazo son principalmente problemas respiratorios, de los más comunes son:

- Dificultad para respirar, sibilancias<sup>2</sup> y tos.
- Incremento en la frecuencia de ataques de asma.
- Mayor riesgo de infecciones respiratorias.
- Susceptibilidad a la inflamación pulmonar.

La exposición al  $O_3$  a largo plazo puede aumentar el riesgo de muerte prematura por enfermedades respiratorias [SEDEMA (2023a)]. Además, las investigaciones recientes proporcionan evidencia sobre más efectos de la exposición crónica entre los que se encuentran:

- Aumento de hospitalizaciones por enfermedades respiratorias preexistentes como asma infantil.
- Puede causar daños reproductivos y de desarrollo, como bajo peso al nacer y disminución de la función pulmonar en recién nacidos.
- Incremento del riesgo de trastornos metabólicos como intolerancia a la glucosa, hiperglucemia y diabetes.

---

<sup>1</sup>Concentración real de contaminantes en el aire.

<sup>2</sup>Sonidos al respirar provocado por la obstrucción de las vías respiratorias.

- Afectación del sistema nervioso central.
- Puede causar daños cardiovasculares como ataques cardíacos, accidentes cerebrovasculares, enfermedades e insuficiencia cardíaca.
- Aumenta la respuesta a los alérgenos en personas con rinitis alérgica, también ocasiona mayor sensibilidad a los alérgenos exteriores.
- Respirar  $O_3$  en combinación con  $SO_2$  y  $NO_x$ , puede ocasionar que los pulmones reaccionen con más fuerza que sólo respirar  $O_3$ .
- Los adultos mayores enfrentan un mayor riesgo de muerte prematura incluso a niveles menores al estándar.

Estos hallazgos subrayan la urgencia de implementar estrategias de monitoreo y control de  $O_3$ , en zonas urbanas con alta contaminación primaria de precursores.

En la Ciudad de México, donde la combinación de factores geográficos, meteorológicos y antropogénicos favorece la acumulación de contaminantes, el ozono troposférico representa un riesgo significativo para la salud pública. La exposición crónica a concentraciones superiores a los umbrales establecidos en las normas oficiales puede agravar los efectos adversos documentados, particularmente en una población urbana densa con alta vulnerabilidad.

La vigilancia sistemática y el análisis de valores extremos de  $O_3$  adquieren así gran relevancia, no solo como herramientas de evaluación ambiental, sino como insumos fundamentales para la protección de la salud y la implementación de políticas públicas efectivas. Este escenario justifica plenamente el estudio de las concentraciones máximas de  $O_3$  mediante enfoques estadísticos robustos, como la TVE, que permitan anticipar episodios de alta contaminación y sus potenciales impactos.

## 5.2. Descripción y manipulación de los datos

Los datos de las concentraciones de ozono, se descargaron de la página oficial de la Dirección de Monitoreo Atmosférico de la Ciudad de México, SIMAT: ([http://www.aire.cdmx.gob.mx/estadisticas-consultas/consultas/download\\_imeca.php](http://www.aire.cdmx.gob.mx/estadisticas-consultas/consultas/download_imeca.php)).

Las bases de datos anuales del índice de calidad del aire contienen información de seis contaminantes criterio ( $O_3$ ,  $NO_2$ ,  $SO_2$ ,  $CO$ ,  $PM_{10}$  y  $PM_{2.5}$ ) que se miden a través de la RAMA (Red Automática de Monitoreo Atmosférico), donde se reporta el índice máximo horario por estación de monitoreo, desde 2019. [SEDEMA-SIMAT].

## CAPÍTULO 5. APLICACIÓN: CONCENTRACIONES MÁXIMAS DE OZONO

A continuación, se describe la estructura de la base de datos y se mencionan algunas especificaciones importantes:

- Se eligió la base de datos del año más reciente disponible, la cual fue del año 2023. Se descargó el archivo comprimido imeca-2023.gz.
- La carpeta contiene ocho archivos con extensión “.csv” (Excel); seis de ellos corresponden a la información de cada contaminante, y los otros dos son catálogos de información, uno de mediciones y el otro de las estaciones de monitoreo.
- Nos interesa el archivo “O3.csv” que contiene las concentraciones horarias de ozono del año 2023, medidas en ppb (partes por billón) y clasificadas por estación.
- Cada estación contiene 8760 registros; desde 01/01/2023 de la hora 01:00, hasta 31/12/2023 de la hora 24:00.
- Hay un total de 34 estaciones de monitoreo atmosférico (Fig. 5.1), sin embargo, algunas de ellas no contienen información del contaminante O<sub>3</sub>, y algunas otras tienen una considerable cantidad de datos nulos, etiquetados como **-99**.

**Configuración del Sistema de Monitoreo Atmosférico**

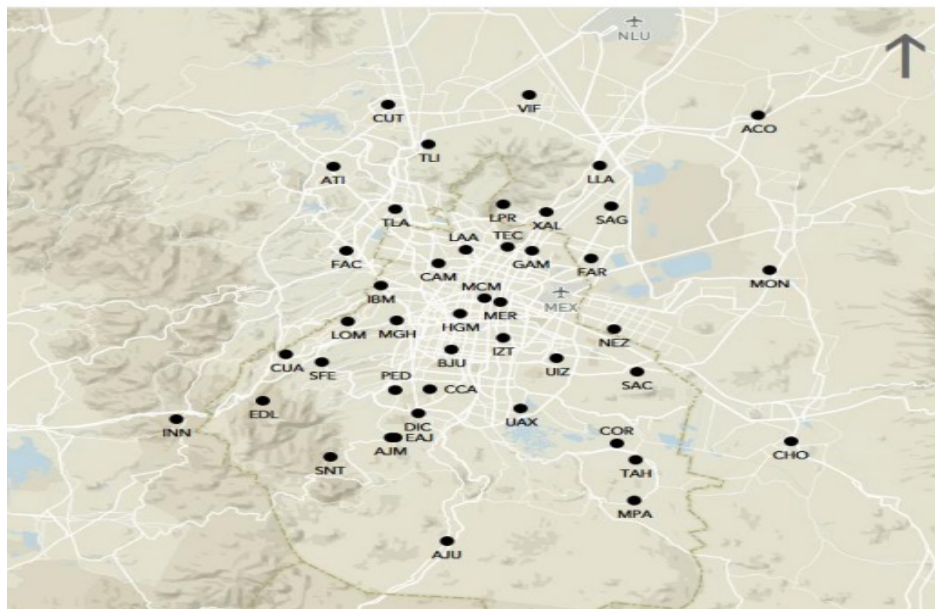


Figura 5.1: Distribución de las estaciones de monitoreo del SIMAT.

Los registros horarios de concentración de ozono nos revela la interacción con emisiones precursoras como la radiación solar y la presencia de gases primarios. Las fuentes antropogénicas también nos brinda información relevante sobre cómo la dinámica diaria de la población alteran las mediciones de O<sub>3</sub>, evidenciando los momentos del día de mayor exposición a este contaminante. Por lo tanto, es de interés analizar las mediciones máximas por día de los datos horarios de concentración atmosférica de ozono.

Al haber demasiados registros perdidos, se realizó una depuración de la base de datos bajo el siguiente criterio; en principio se eliminaron las estaciones que no contenían información existente de la concentración de  $O_3$ ; luego se decidió omitir las estaciones que tenían más del 20% de mediciones perdidas, o bien, más de 1752 datos nulos. Así, de las 34 estaciones disponibles, solo se consideraron 15 centros de medición como candidatos para realizar el análisis descrito.

Se revisaron detalladamente las 15 estaciones y se seleccionaron aquellas que contaban con mayor información, con la finalidad de tener una mejor interpretación del comportamiento de las concentraciones máximas diarias. Cabe mencionar que los pocos datos nulos que se detectaron en estos sitios de monitoreo fueron rellenados con el promedio del resto de las estaciones correspondientes a esa misma hora del mismo día del registro faltante, permitiendo preservar la distribución general de los datos.

Además, para tener un panorama amplio del comportamiento del contaminante  $O_3$  en la ZMCM, también se consideró la ubicación de cada una de las estaciones, ya que las estaciones cercanas podrían contener información similar por las condiciones atmosféricas, geográficas y antropológicas de la región.

Finalmente se seleccionaron únicamente 3 estaciones de monitoreo atmosférico para analizar el comportamiento general de las concentraciones máximas de la ZMCM, las tres estaciones cuentan con suficiente información del contaminante y están ubicadas estratégicamente en distintas zonas de la región.

La hipótesis general es que los datos de concentraciones máximas diarias de ozono 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 observaciones de concentraciones atmosféricas 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 realizar 365 bloques por día de los registros horarios disponibles para cada estación, de esta manera se disminuye la correlación entre las mediciones.

Debemos recordar que nuestros datos están establecidos bajo los estándares de la normatividad NADF-009-AIRE-2017, la cual establece los requisitos para elaborar el Índice de Calidad del Aire en la Ciudad de México. Sin embargo, es importante mencionar que esta norma ambiental no considera las actualizaciones de los valores límite en las NOM (Normas Oficiales Mexicanas) de salud ambiental, realizadas entre 2019 y 2021.

**Nota:** La norma NADF-009-AIRE-2017 establece que el límite máximo horario es de **95 ppb**, valor que no debe ser rebasado en ningún momento del año [SEDEMA (NOM)].

### 5.3. Análisis de los datos por estación

Las tres estaciones de monitoreo atmosférico seleccionadas para el análisis fueron las estaciones: CUA, BJU y LPR; las cuales integran parte de la RAMA (Red Automática de Monitoreo Atmosférico) de la CDMX. La Tabla 5.1 muestra los nombres y la información geográfica<sup>3</sup> de cada sitio de monitoreo. Las alcaldías de Cuajimalpa de Morelos y Benito Juárez están ubicadas dentro de la Ciudad de México, mientras que el municipio de Tlalnepantla de Baz pertenece al Estado de México.

Tabla 5.1: Información geográfica de las estaciones atmosféricas en estudio.

Clave	Nombre	Municipio (Zona)	Latitud	Longitud	Altitud
CUA	Cuajimalpa	C. de Morelos (Suroeste)	19.36531°	-99.2917°	2704 msnm
BJU	Benito Juárez	Benito Juárez (Centro)	19.37167°	-99.1591°	2250 msnm
LPR	La Presa	Tlalnepantla (Norte)	19.53473°	-99.1177°	2302 msnm

Para el análisis individual de cada estación atmosférica se usará la siguiente notación:  $\bar{x}$  ← media muestral,  $S$  ← desv.est. muestral,  $Q_p$  ← p-cuartil ( $Q_2$  ← mediana),  $IQR$  ← rango intercuartílico,  $\gamma_1$  ← coef. de asimetría y  $\gamma_2$  ← curtosis.

#### Estación: Cuajimalpa (CUA)

La estación CUA está ubicada en la dirección: Monte Encino No. 14, Col. Jesús del Monte, C.P. 05260; en la zona suroeste de la ZMCM, cerca de la carretera México-Toluca y zonas boscosas como el Desierto de los Leones.



Figura 5.2: Caseta de monitoreo de Cuajimalpa (ID: 484090040109)

<sup>3</sup>La altitud usa la unidad de medida estándar: *metros sobre el nivel del mar* (msnm).

Esta caseta de monitoreo comenzó a operar desde 1994 y realiza mediciones de la concentración atmosférica de  $O_3$ ,  $CO$ ,  $SO_2$ ,  $NO_2$  y  $PM_{10}$ , considerando variables meteorológicas como temperatura, humedad, radiación solar y velocidad del viento. Para las mediciones de ozono está clasificada como tipo 4-Urbana según estándares de representatividad del SEDEMA-SIMAT, es decir, que la caseta logra registrar la concentración de  $O_3$  dentro de un perímetro de 4 a 50 km.

La Alcaldía de Cuajimalpa cuenta con una población de alrededor de 217,686 habitantes, con una densidad de 2,689 habitantes por  $km^2$ . Esta zona tiene algunas características geográficas y atmosféricas que son importantes mencionar:

Una publicación del Observatorio Territorial del Poniente (OTP) realizado por la UAM-CUA (2022) describe que el terreno es montañoso en la Sierra de las Cruces, lo que genera efectos de canalización de vientos e inversiones térmicas nocturnas frecuentes. Además, se asevera que hay un área cubierta por bosques, predominando árboles de pino-encino y oyamel, lo que contribuye a la formación de COV biogénicos.

De acuerdo a un estudio del panorama geográfico y estadístico de la Alcaldía de Cuajimalpa de Morelos, IPDP (2024), el clima que predomina es templado subhúmedo y con lluvias en verano de mayor humedad. En cuanto a la temperatura, afirman que la media anual oscila entre los  $12^\circ C$  y  $18^\circ C$ , con variaciones según la altitud.

En resumen, se detectaron algunos factores particulares que pueden presentar singularidades en la dinámica del  $O_3$ :

- **Mezcla urbana-forestal.** El área estudiada combina zonas residenciales, comerciales y áreas naturales protegidas.
- **Influencia de bosques.** La vegetación emite compuestos volátiles que, con la luz solar, favorece la formación de ozono troposférico.
- **Tráfico vehicular.** La cercanía a la carretera México-Toluca y avenidas principales contribuye a emisiones de  $NO_x$ .
- **Vientos dominantes.** Los vientos del sur-poniente pueden transportar contaminantes desde otras zonas de la ciudad.

### Análisis descriptivo general de los registros horarios

En la estación CUA, del total de 8,760 mediciones horarias, se identificaron 8,235 registros válidos (94 % de completitud), en contraste se detectaron 525 datos nulos (6 %), los cuales fueron imputados mediante sustitución con el promedio horario.

CAPÍTULO 5. APLICACIÓN: CONCENTRACIONES MÁXIMAS DE OZONO

La Tabla 5.2 muestra los valores teóricos del análisis descriptivo, lo que sugiere algunas inferencias en cuanto a la dinámica de las mediciones horarios de O<sub>3</sub> registradas en Cuajimalpa en el 2023.

Tabla 5.2: Estadísticas descriptivas generales: Estación CUA

$\bar{x}$	$S$	Mín	$Q_1$	$Q_2$	$Q_3$	Máx	IQR	$\gamma_1$	$\gamma_2$
29.11	24.17	0	14	23	36	161	22	1.84	3.87

De los valores que se ubican en los extremos (cuadros verdes) podemos observar que hay una concentración promedio anual de 29.11 ppb con una considerable dispersión en los niveles de ozono oscilando sobre este promedio; más aún, se infiere una distribución en forma leptocúrtica con asimetría positiva pronunciada con la cola derecha muy alargada, debido a registros excepcionalmente altos de ozono. Por otra parte, del valor de los cuartiles, vemos que el 50 % de los datos centrales se encuentra entre 14 ppb y 36 ppb, un intervalo pequeño considerando la gran dispersión, esto sugiere la posible presencia de outliers; por último hacemos énfasis en el valor máximo anual de **161 ppb**.

**Análisis de las máximas concentraciones diarias**

Siguiendo la metodología del modelo de máximos por bloque; se dividieron los 8760 datos en 365 bloques, cada uno con las concentraciones máximas diarias de O<sub>3</sub> registradas en las 24 horas del día correspondiente. Las Tablas 5.3 y 5.4 muestran los primeros 120 bloques de máximos por día, resaltando con rojo los valores máximos por mes. De este modo se identificó que el valor máximo anual de 161 ppb pertenece al bloque 54, del día 23 de febrero del 2023.

Tabla 5.3: Concentraciones máximas por día (bloques 1-59): Estación CUA

ENERO															
<b>Bloq</b>	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
<b>Máx</b>	44	22	30	96	73	65	69	84	94	65	61	71	50	45	104
16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31
101	84	81.23	121	110	65	105	37	36	122	73	104	80	108	105	106
FEBRERO															
<b>Bloq</b>	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46
<b>Máx</b>	130	109	50	115	46	86	115	104	88	118	100	119	104	35	36
<b>Bloq</b>	47	48	49	50	51	52	53	54	55	56	57	58	59		
<b>Máx</b>	109	143	120	33	86	90	36	161	120	42	49	100	106		

## CAPÍTULO 5. APLICACIÓN: CONCENTRACIONES MÁXIMAS DE OZONO

Tabla 5.4: Concentraciones máximas por día (bloques 60-120): Estación CUA

MARZO															
<b>Bloq</b>	60	61	62	63	64	65	66	67	68	69	70	71	72	73	74
<b>Máx</b>	100	47	59	109	120	88	57	59	76	61	82	84	59	50	41
75	76	77	78	79	80	81	82	83	84	85	86	87	88	89	90
32.27	80	31	36	43	129	104	116	104	135	<b>153</b>	104	108	101	117	143
ABRIL															
<b>Bloq</b>	91	92	93	94	95	96	97	98	99	100	101	102	103	104	105
<b>Máx</b>	98	113	<b>139</b>	106	131	102	44	96	61	49	48	39	41	46	49
<b>Bloq</b>	106	107	108	109	110	111	112	113	114	115	116	117	118	119	120
<b>Máx</b>	71	50	67.17	67	105	51	44	43	47	46	51	45	117	63	67

Se mostraron los primeros 120 bloques para identificar el bloque que contiene el máximo anual, sin embargo al haber 365 bloques, es poco práctico mostrar todos los valores máximos diarios encontrados. Por esta razón se muestra la Tabla 5.5 que indica la cantidad de bloques que hay en los meses restantes, así como el máximo mensual registrado y el bloque al que le pertenece dicho valor.

**Nota:** En los próximos análisis se mostrarán los máximos por bloque de esta manera.

Tabla 5.5: Concentraciones máximas por mes (bloques 121-365): Estación CUA

Mes	Bloques	B.Máx	Máx	Mes	Bloques	B.Máx	Máx
<b>Mayo</b>	121-151	123	126	<b>Septiembre</b>	244-273	264	137
<b>Junio</b>	152-181	167	130	<b>Octubre</b>	274-304	294	134
<b>Julio</b>	182-212	196	124	<b>Noviembre</b>	305-334	324	142
<b>Agosto</b>	213-243	223	131	<b>Diciembre</b>	335-365	353	94

Para analizar la dinámica de las concentraciones atmosféricas máximas diarias de  $O_3$  se presenta la Tabla 5.6 de estadísticos descriptivos de los 365 valores máximos por día; además se generó un diagrama de dispersión y un gráfico de "caja y bigotes" para complementar con un análisis teórico y visual.

Tabla 5.6: Descriptivos de los máximos por día: CUA

Mín	$Q_1$	$Q_2$	Media	$Q_3$	Máx	IQR
<b>14.00</b>	<b>46.00</b>	<b>73.00</b>	<b>75.26</b>	<b>104.00</b>	<b>161.00</b>	<b>58.00</b>

La Figura 5.3 muestra ambos gráficos, donde se marca con una línea roja horizontal el valor promedio de los valores máximos diarios de concentración de ozono, que fue de 75.26 ppb, muy cercano a la mediana de 73 ppb.

En el box-plot podemos confirmar que este valor promedio se asemeja al valor de la media, lo cual indica que los máximos diarios se aproximan a una distribución más simétrica, a diferencia de lo indicado en la Tabla 5.2 de los descriptivos generales. En cuanto a la dispersión de los puntos del primer diagrama, no se observa alguna tendencia clara para las concentraciones máximas de ozono, las líneas anaranjadas del mismo gráfico, indican los valores del primer y tercer cuartil, intervalo dentro del cual se encuentra el 50% de los datos centrales.

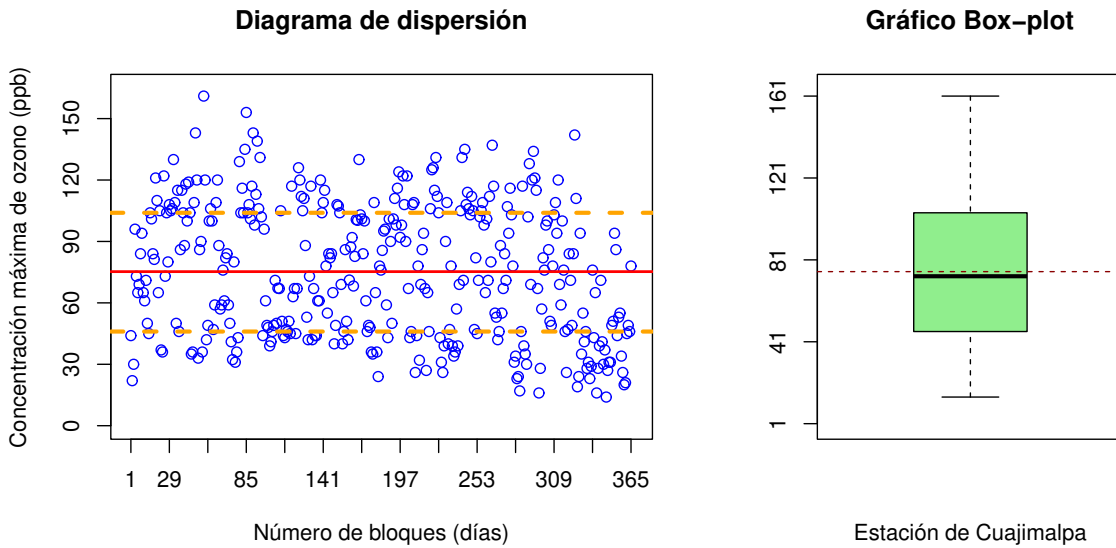


Figura 5.3: Gráfica de dispersión y Box-plot: Estación CUA

El valor mínimo de los máximos diarios fue de 14 ppb, el cual pertenece al bloque 347, del día 13 de diciembre del 2023, lo cual tiene sentido ya que en la Tabla 5.5 se observa que diciembre tiene el menor de los valores máximos por mes; estos niveles de concentración de  $O_3$  están relacionados a las bajas temperaturas decembrinas.

En el box-plot, se visualiza una distribución casi simétrica, por la forma de la caja y la longitud de los bigotes, además la media y la mediana prácticamente se sobreponen una de otra. Más aún, no se visualiza la presencia de valores atípicos y esto se confirma calculando la CSI (Cota Superior Interior), umbral que determina si un dato es considerado atípico y se calcula de la siguiente manera:  $CSI = Q_3 + 1.5 \times IQR = 191$ ; cantidad que sobrepasa el máximo de los datos, por lo que no existe ningún candidato a outlier dentro de la distribución de los máximos diarios.

Por todo lo anterior, se espera una distribución de los máximos aproximadamente simétrica y con forma platicúrtica, de colas no pesadas, con un ligero alargamiento de la cola derecha (asimetría positiva mínima o nula).

### Ajuste de los registros máximos diarios a una DVEG

Para ajustar una base de datos a una DVEG, primero debemos obtener la estimación paramétrica para la función de distribución; usando el método de máxima verosimilitud obtenemos los estimadores de los parámetros de localización, de escala y forma:

$$\hat{\mu} = 63.322, \quad \hat{\sigma} = 31.551, \quad \hat{\xi} = -0.259.$$

Para estimar los errores estándar de cada estimador, podemos apoyarnos de la matriz de covarianza  $\Sigma$  definida en el capítulo 1; suponiendo que  $\hat{\boldsymbol{\theta}} = (\hat{\mu}, \hat{\sigma}, \hat{\xi})$  es el vector de estimadores, la matriz var-cov está dada por:

$$\Sigma(\hat{\boldsymbol{\theta}}) = \begin{pmatrix} 3.536774 & 0.392245 & -0.036485 \\ 0.392245 & 2.008446 & -0.039702 \\ -0.036485 & -0.039702 & 0.001931 \end{pmatrix}.$$

Como ya sabemos, la diagonal principal de la matriz  $\Sigma$  corresponde a las varianzas de cada estimador, aplicando la raíz cuadrada se obtienen los errores estándar de cada uno:

$$SE_{\hat{\mu}} = 1.8806, \quad SE_{\hat{\sigma}} = 1.4172, \quad SE_{\hat{\xi}} = 0.0439.$$

Notemos que el estimador  $\hat{\xi} = -0.259 < 0$ , indicando que el modelo de distribución a la que se ajustan las concentraciones máximas de ozono de la estación CUA, es un modelo Weibull. Sin embargo, al ser un valor cercano a cero, no podemos descartar la posibilidad de que se ajuste a una distribución Gumbel. Por otro lado, podemos descartar el modelo Fréchet, ya que no es adecuado bajo el criterio de estimación puntual.

Lo que sigue es determinar qué parámetros son razonables y cuales son contradichos por los datos; recordemos que un estimador puntual no proporciona información acerca de la incertidumbre en la estimación, por eso se recomienda calcular los intervalos de verosimilitud, o bien, los intervalos de confianza de cada parámetro. El estimador de forma  $\hat{\xi} > -0.5$  establece que los EMV son regulares y por lo tanto cumplen las propiedades asintóticas de normalidad. Así, es posible calcular los intervalos de confianza para cada parámetro a un nivel del 95%; por lo tanto, fijando un valor  $\alpha = 0.05$ , obtenemos:

$$\begin{aligned} \mu &\in (59.5967, 66.9707), \\ \sigma &\in (28.8490, 34.4505), \\ \xi &\in (-0.3468, -0.1749). \end{aligned}$$

Estos intervalos nos brindan un nivel de confiabilidad para asegurar que un verdadero valor de algún parámetro desconocido se encuentre dentro de estos intervalos de estimación. Bajo este argumento y dado que el IC para  $\xi$  no captura el valor 0, podemos descartar la posibilidad, con un nivel de confianza del 95 %, de que la función Gumbel es un buen modelo de ajuste para la distribución de las concentraciones máximas de ozono de la estación CUA. Por lo tanto, los datos proporcionan evidencia suficiente para elegir el modelo Weibull de la familia de DVEG.

### Validación del modelo Weibull

Lo que prosigue es aplicar métodos de validación para determinar si, en efecto, la distribución de las mediciones máximas diarias de  $O_3$  se ajustan a esta función. Para esto, usaremos métodos gráficos de diagnóstico que nos permiten visualizar la proximidad de los datos a la distribución Weibull.

La gráfica de probabilidad (PP-plot) y la gráfica de cuantiles (QQ-plot) son dos métodos de diagnóstico para evaluar la precisión del modelo Weibull ajustado a las concentraciones atmosféricas máximas de ozono medidas por la estación CUA.

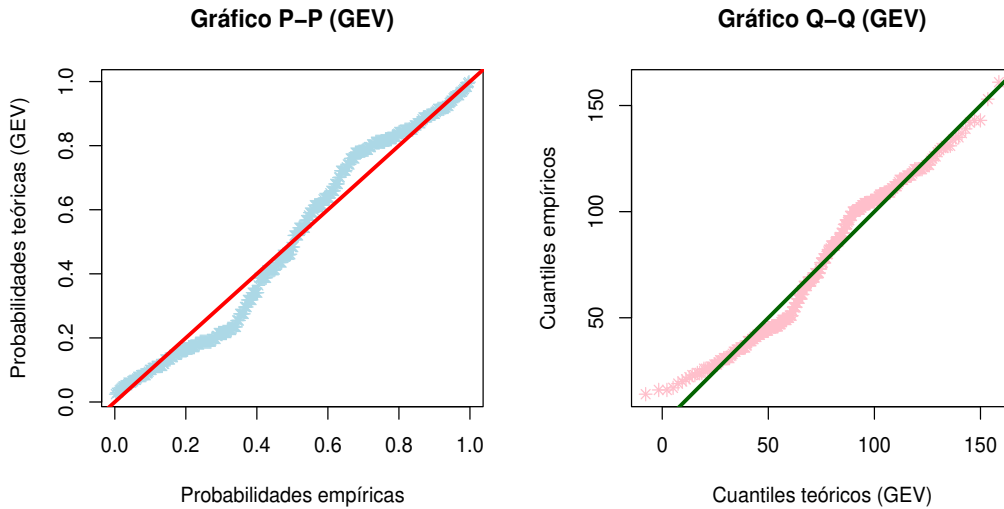


Figura 5.4: Gráficas de diagnóstico P-P y Q-Q del modelo Weibull: Estación CUA

En la Figura 5.4 se presentan ambos gráficos, en donde se observa una considerable proximidad de los datos a las rectas de referencia ( $y = x$ ), afirmando un buen ajuste del modelo Weibull. Sin embargo, a pesar de la cercanía de los registros (puntos azules y rosas) con las líneas de identidad correspondientes (roja y verde), se notan ligeras oscilaciones que sugieren una posible autocorrelación entre los datos, es decir, patrones cíclicos que dependen de mediciones anteriores, lo cual es común en series temporales.

En general, los gráficos de diagnóstico confirman que el modelo Weibull captura adecuadamente la distribución de los datos, indicando que los parámetros estimados  $\hat{\mu}$ ,  $\hat{\sigma}$  y  $\hat{\xi}$  son apropiados, pero la presencia de las oscilaciones no debe ser ignorada; podrían estar señalando variaciones estacionales no capturadas por el modelo.

Siguiendo con el análisis gráfico de la dinámica de la concentración de ozono, en la Figura 5.5 se muestra el diagrama 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.

El gráfico de la función de densidad de probabilidad ajustada consiste en comparar la densidad del modelo ajustado, en nuestro caso del modelo Weibull, con el histograma de los datos de concentraciones atmosféricas máximas de ozono por bloque medidos por la estación de monitoreo de Cuajimalpa de Morelos. 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.

El número de clases del histograma se eligió con el criterio de la raíz ( $\sqrt{365} \approx 19.10$ ), y dado que en esta estación el valor mínimo fue de 14 ppb y el máximo de 161 ppb; se optó por tomar 19 clases en un rango de 12 a 163, así cada clase tendrá la misma dimensión, cubriendo toda la información de los registros máximos.

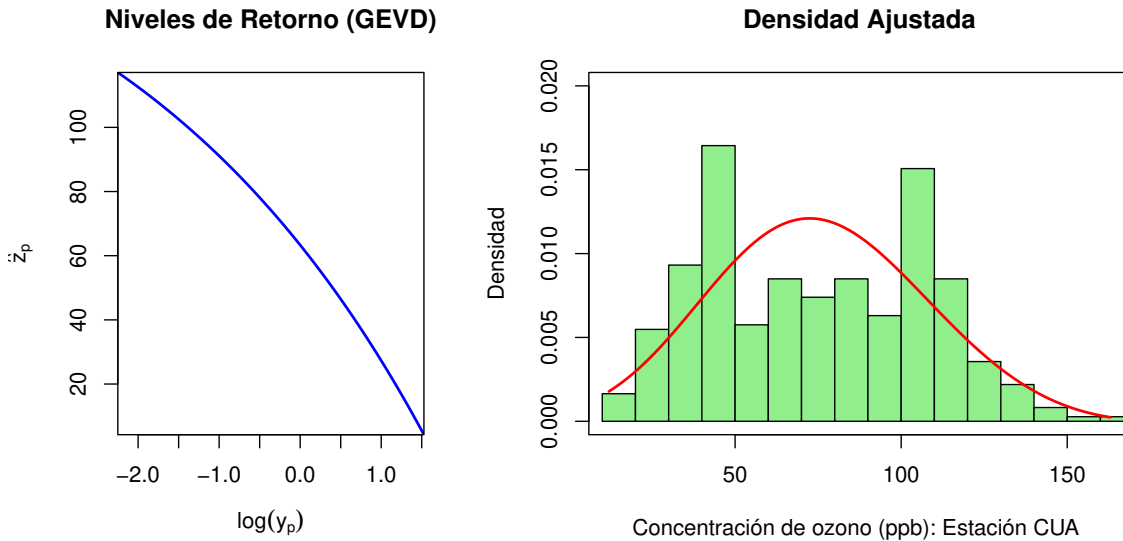


Figura 5.5: Gráfica de niveles de retorno y densidad ajustada (GEVD)

En cambio, la gráfica de niveles de retorno muestra suficiente evidencia de que el modelo Weibull es adecuado para los datos de concentración de  $O_3$  registrados en Cuajimalpa, debido a que la gráfica es cóncava.

Por lo tanto, de los resultados obtenidos en los gráficos de diagnóstico P-P y Q-Q y de la información visual que proporcionan estos dos últimos, se concluye que **el modelo Weibull se ajusta adecuadamente a las concentraciones máximas diarias de concentración de O<sub>3</sub> medidas en la estación CUA.**

### Estimación de riesgos por concentración de O<sub>3</sub> en Cuajimalpa

Es necesario recordar que, en la introducción de esta tesis, se establecieron los umbrales en unidades ppb, la calidad del aire según los límites permisibles, así como los riesgos y las recomendaciones pertinentes. De manera que, una concentración de O<sub>3</sub> por arriba de **95 ppb** cataloga la calidad del aire como **Mala**, ocasionando que los grupos susceptibles presenten efectos dañinos en su salud, y deberán limitar su exposición al aire libre. Más aún, un valor por encima de **154 ppb** indica una **Muy Mala** calidad del aire, lo que provoca que la población en general pueda presentar síntomas, agravando aún más la salud de los grupos susceptibles, entrando a la Fase 1 de contingencia, donde toda la población debe limitar sus actividades al exterior.

Se realizará un ajuste de los registros de las concentraciones máximas diarias de ozono de las mediciones en Cuajimalpa usando la distribución del modelo Weibull; cuya función de distribución está dada por:

$$G(z) = \exp \left\{ - \left[ 1 - 0.259 \left( \frac{z - 63.322}{31.551} \right) \right]^{\frac{1}{(0.259)}} \right\}, \quad z \in (185.1405, \infty).$$

La gráfica de esta función de distribución y la gráfica de la función de densidad de probabilidad ajustada son herramientas que nos ayudan a visualizar el comportamiento de las concentraciones atmosféricas máximas de O<sub>3</sub>. La función de densidad para el modelo Weibull y para  $z \in (185.1405, \infty)$  está dada por:

$$g(z) = \frac{1}{31.551} \left[ 1 - 0.259 \left( \frac{z - 63.322}{31.551} \right) \right]^{2.861} \exp \left\{ - \left[ 1 - 0.259 \left( \frac{z - 63.322}{31.551} \right) \right]^{3.861} \right\}.$$

En la Figura 5.6; la función de distribución indica que la probabilidad de tener una concentración promedio máxima menor a 95 ppb es del 73.06 % ( $\mathbb{P}(z \leq 95) = 0.7306$ ), lo que implica que la probabilidad de tener una Mala calidad del aire un día cualquiera del año es del 26.94 % ( $\mathbb{P}(z > 95) = 1 - \mathbb{P}(z \leq 95) = 0.2694$ ). En la función de densidad podemos ver que en el 2023 hubo una gran cantidad de días que superaron este valor, para ser precisos fueron 127 días (34.79 %) del año 2023 en los que se catálogos una **Mala calidad del aire.**

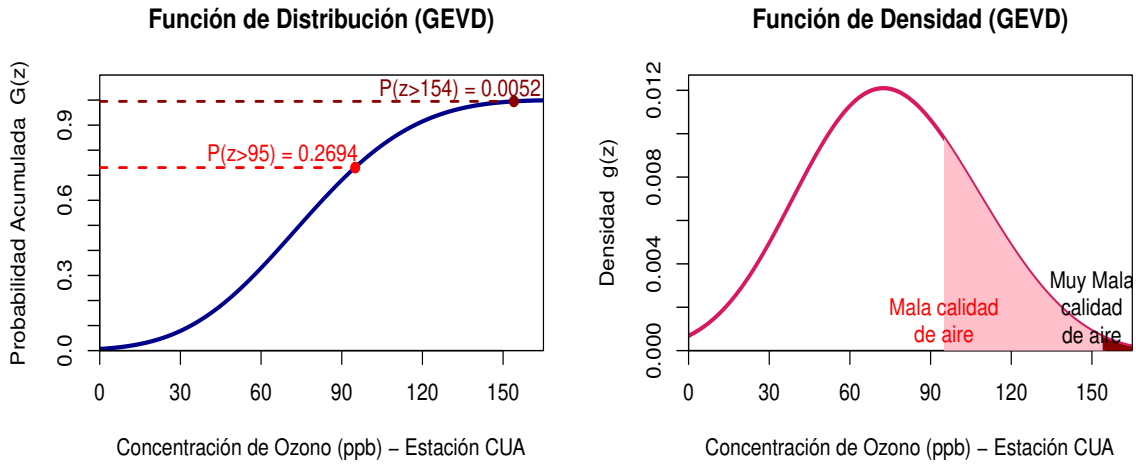


Figura 5.6: Funciones de distribución acumulada y densidad de probabilidad (GEVD)

Por otra parte, la probabilidad de superar el umbral máximo de 154 ppb un día cualquiera del año es del 0.52% ( $\mathbb{P}(z > 154) = 0.0052$ ), aunque esta probabilidad es muy baja, o prácticamente imposible de que suceda, la función de densidad nos indica que en el 2023 hubo por lo menos un caso en el que se superó ese umbral, el cual corresponde al máximo anual marcado por el bloque 54 con una concentración de ozono de 161 ppb el día 23 de febrero del 2023. En este día debió activarse la Fase 1 de Contingencia por una **Muy mala** calidad de aire.

Estos resultados son bastante realistas, y es la razón por la cual las normas oficiales se han ido actualizando constantemente; en la ZMCM la contaminación atmosférica cada vez es mayor y se deben ajustar los valores límite establecidos por el gobierno.

### Niveles de retorno

Los niveles de retorno permiten cuantificar el riesgo de excedencias peligrosas de concentración de ozono, apoyando la gestión ambiental y la protección de la salud pública. En general, en la TVE es de interés conocer niveles de retorno asociados a periodos de 20, 50 y 100 años. Sin embargo, los periodos de retorno pueden variar según convenga al fenómeno natural analizado y a quien lo estudia.

En el caso de las concentraciones atmosféricas máximas diarias de ozono, vamos a considerar periodos de retorno de  $T = 365, 1825, 3650$  y  $7300$  días que equivalen a 1, 5, 10 y 20 años, respectivamente.

La Tabla 5.7 muestra los periodos y niveles de retorno, la varianza asociada a cada  $\hat{z}_p$  calculada con el método delta, y los IC del 95% estimados para cada nivel de retorno.

Tabla 5.7: Periodos y niveles de retorno del modelo Weibull: Estación CUA.

$T$ (días)	$T$ (años)	Nivel de retorno	$Var(z_p)$	IC del 95 %
365	1	158.647 ppb	28.080	(148.26, 169.03)
1825	5	167.649 ppb	44.286	(154.61, 180.69)
3650	10	170.507 ppb	52.018	(156.37, 184.64)
7300	20	172.895 ppb	60.135	(157.69, 188.09)

Para la interpretación de los valores obtenidos, veamos un ejemplo:

El nivel de retorno estimado de 167.649 ppb para un periodo de retorno de 5 años ( $T = 5$ ) indica que, en promedio, se espera que la concentración máxima diaria de ozono exceda este valor al menos una vez en los próximos 5 años. La varianza asociada al estimador que indica un error estándar aproximado de 6.655 ppb junto con el IC del 95 % demuestran una precisión aceptable para esta estimación.

Cabe destacar que, una característica importante de la distribución Weibull es que tiene un límite superior, por lo que las concentraciones máximas de ozono no superarán el valor teórico de  $\mu - \sigma/\xi \approx 185.14$  ppb; esto se refleja en el aplanamiento progresivo de la curva de niveles de retorno conforme aumenta el periodo de retorno (Figura 5.7).

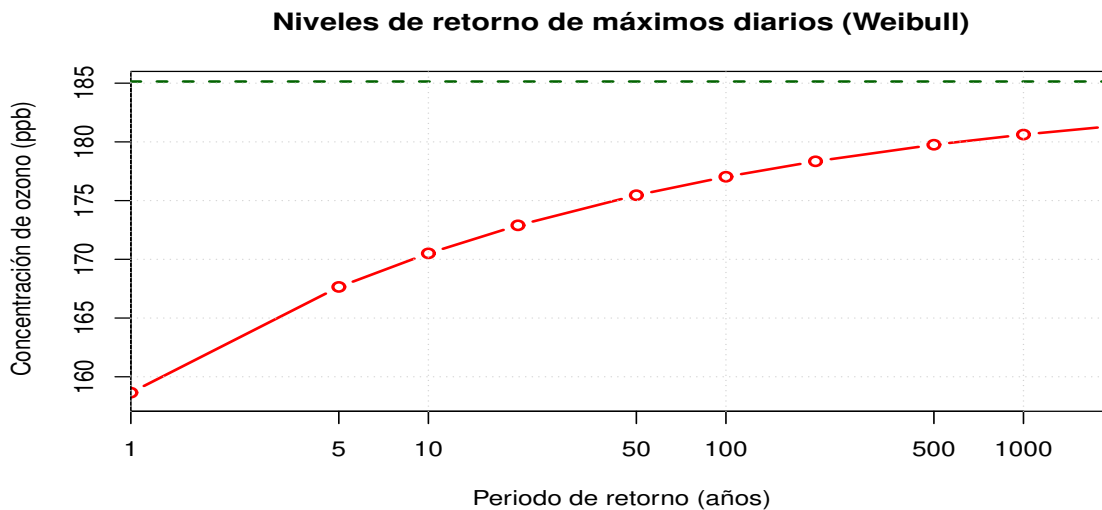


Figura 5.7: Niveles de retorno máximos por día (Weibull): Estación CUA

Estos resultados son relevantes para la evaluación del riesgo ambiental y el diseño de políticas públicas relacionadas con la calidad del aire, ya que permiten cuantificar la probabilidad de ocurrencia de eventos extremos de contaminación por  $O_3$ .

### Estación: Benito Juárez (BJU)

La dirección de la estación BJU es: Municipio libre y Uxmal, Col. Santa Cruz Atoyac, Benito Juárez, C.P. 03310, ubicada en la región central de la ZMCM. Al ser una zona urbana se encuentra cerca de plazas comerciales como Antara y Centro Santa Fe, además dentro de la Alcaldía hay parques locales como el parque Hundido y el Parque de los Venados. Esta caseta comenzó a operar desde el año 2015 y cuenta con registros de la concentración atmosférica de  $O_3$ ,  $CO$ ,  $SO_2$ ,  $NO_2$ ,  $PM_{10}$  y  $PM_{2.5}$ , considerando variables meteorológicas como temperatura, humedad, radiación solar y velocidad del viento. Tiene clasificación tipo 3-Vecinal según estándares de representatividad del SEDEMA-SIMAT, es decir, que la caseta logra registrar la concentración de  $O_3$  dentro de un perímetro de 0.5 a 4 km.



Figura 5.8: Caseta de monitoreo de Benito Juárez (ID:484090140309)

La Alcaldía Benito Juárez tiene una población estimada de 434,153 habitantes, con una densidad poblacional de 18,592 habitantes por  $km^2$ ; de acuerdo al INEGI es una de las densidades más altas de toda la CDMX. Al ser un área tan poblada, a parte de considerar variables meteorológicas, esta estación también considera diversos factores antropológicos que pueden alterar las mediciones de concentración de ozono.

Los principales focos de contaminación que contribuyen en la emisión de  $\text{NO}_x$  y de Compuestos Orgánicos Volátiles en región central son:

- **Avenidas de alto tráfico.** La Av. de los Insurgentes que es una de las más transitadas, y ejes viales como Mixcoac, Félix Cuevas y Río Churubusco generan emisiones de  $\text{NO}_x$ .
- **Zonas industriales y comerciales.** Polanco y Santa Fe concentran edificios corporativos, centros comerciales y pequeñas industrias que usan gran cantidad de solventes.
- **Áreas verdes cercanas.** El Bosque de Chapultepec y los parques locales emiten COV biogénicos.
- **Factores metereológicos.** En la zona central hay fuertes vientos que arrastran contaminantes desde industrias aledañas como Azcapotzalco; presenta un clima templado promedio de  $18^\circ\text{C}$  con variaciones estacionales entre  $6^\circ\text{C}$  y  $28^\circ\text{C}$ .

### Análisis descriptivo general de los registros horarios

La estación BJU tuvo 91.67% de completitud en los datos, es decir que del total de 8,760 mediciones horarias se identificaron 8,030 registros válidos, lo que implica que hubo 730 datos nulos (8.33%) que tuvieron que ser sustituidos por el promedio horario. Los resultados del análisis descriptivo de todos los registros horarios están indicados en la Tabla 5.8 de los cuales podemos hacer algunas observaciones respecto a la dinámica de la base de datos.

Tabla 5.8: Estadísticas descriptivas generales: Estación BJU

$\bar{x}$	<i>SE</i>	Mín	$Q_1$	$Q_2$	$Q_3$	Máx	IQR	$\gamma_1$	$\gamma_2$
22.42	23.08	0	6	15	31	141	25	1.77	6.37

El valor tan alto de la curtosis ( $\gamma_2$ ) indica una distribución en forma leptocúrtica muy pronunciada, esto es debido a que el rango total de los datos es robusto comparado con la distribución de los mismos, dado que el 50% central están entre 6 ppb y 31 ppb; este hecho y la clara asimetría positiva que se observa ( $\gamma_1 > 0$ ) afirman una distribución con un pico muy alto con poca dispersión y la cola derecha muy alargada, infiriendo posibles valores extremos. En este caso hubo una concentración horaria promedio de 22.42 ppb, que supera a la mediana de 15 ppb y el valor máximo anual, que es el de interés para esta tesis, fue de **141 ppb**.

**Análisis de las máximas concentraciones diarias**

Nuevamente se utilizaron 365 bloques con las mediciones máximas horarias, como se mencionó en el análisis anterior, en la Tabla 5.9 se presenta una organización por intervalos de la cantidad de bloques que hay en cada mes, así como el máximo mensual registrado y el bloque al que le pertenece dicho valor. Se logró identificar que la concentración máxima anual pertenece al bloque 241, del día 29 de agosto del 2023.

Tabla 5.9: Intervalos de bloques con registros máximos mensuales: Estación BJU

Mes	Bloques	B.Máx	Máx	Mes	Bloques	B.Máx	Máx
<b>Enero</b>	1-31	19	106	<b>Julio</b>	182-212	185	112
<b>Febrero</b>	32-59	54	120	<b>Agosto</b>	213-243	<b>241</b>	<b>141</b>
<b>Marzo</b>	60-90	87	126	<b>Septiembre</b>	244-273	257	128
<b>Abril</b>	91-120	93	114	<b>Octubre</b>	274-304	295	124
<b>Mayo</b>	121-151	132	109	<b>Noviembre</b>	305-334	325	139
<b>Junio</b>	152-181	152	123	<b>Diciembre</b>	335-365	365	109

En la Tabla 5.10 se muestran los resultados del análisis descriptivo de las máximas concentraciones de ozono diarias de la estación Benito Juárez. Estas cifras junto con el diagrama de dispersión y el gráfico box-plot revelaron mucha información sobre la dinámica de los datos máximos.

Tabla 5.10: Descriptivos de los máximos por día: BJU

Mín	$Q_1$	$Q_2$	Media	$Q_3$	Máx	IQR
<b>9</b>	<b>46</b>	<b>67</b>	<b>69.24</b>	<b>94</b>	<b>141</b>	<b>48</b>

En ambos gráficos de la Figura 5.9 se marca con una línea roja horizontal el valor promedio de los valores máximos diarios de concentración de ozono, que fue de 69.24 ppb; en el boxplot podemos notar que este valor se asemeja más al valor de la media de 67 ppb, lo cual indica que los máximos diarios siguen una distribución casi simétrica, con una posible asimetría positiva muy baja, a diferencia de la distribución de la base de datos original.

En cuanto a la dispersión de los puntos del primer diagrama, no se observa alguna tendencia clara para las concentraciones máximas de  $O_3$ ; las líneas anaranjadas del mismo gráfico indican los valores del primer y tercer cuartil, intervalo dentro del cual se encuentra el 50% de los datos centrales. El hecho que los puntos fuera de este intervalo no se alejen tanto de las líneas de los cuartiles junto con los "bigotes" del box-plot demuestran una distribución de colas pesadas.

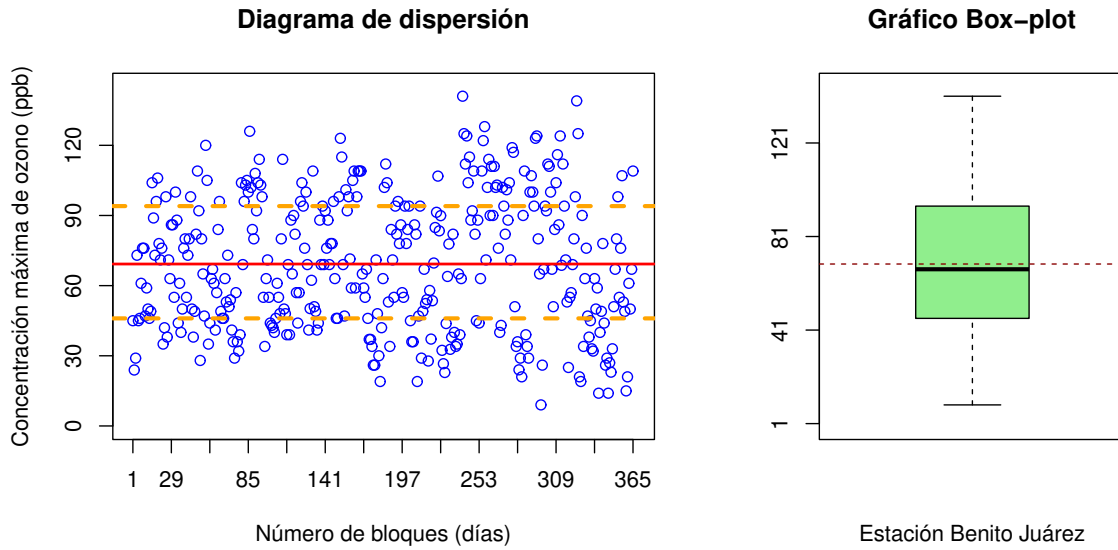


Figura 5.9: Gráfica de dispersión y Box-plot: Estación BJU

Al calcular la CSI,  $Q_3 + 1.5 \times IQR = 166$  ppb, notamos que sobrepasa el valor máximo de 141 ppb, lo que quiere decir que no hay ningún valor candidato a ser atípico; tal y como se visualiza en el box-plot. En este mismo gráfico, la forma peculiar de la "caja" y los no tan alargados "bigotes" sugieren una forma platicúrtica de la distribución de los datos con colas no muy pesadas, al calcular el valor de la curtosis ( $\gamma_2 = 2.04 < 3$ ) se confirma esta aseveración.

### Ajuste de los registros máximos diarios a una DVEG

Usando el método de máxima verosimilitud obtenemos los estimadores de los parámetros de localización, de escala y forma:

$$\hat{\mu} = 58.737, \quad \hat{\sigma} = 27.547, \quad \hat{\xi} = -0.254.$$

Al calcular la raíz cuadrada de los valores de la diagonal principal de la matriz cov-var, obtenemos el error estándar de cada estimador:

$$SE_{\hat{\mu}} = 1.651, \quad SE_{\hat{\sigma}} = 1.235, \quad SE_{\hat{\xi}} = 0.046.$$

El estimador  $\hat{\xi} = -0.254 < 0$ , nuevamente descarta por completo al modelo Fréchet como un buen modelo de ajuste para la distribución de las concentraciones máximas diarias de ozono para la estación Benito Juárez.

Indicando que los posibles modelos de ajuste al comportamiento de los datos son el modelo Weibull o el modelo Gumbel, por lo que debemos calcular los intervalos de confianza para asegurarnos de qué modelo debemos elegir. Como  $\hat{\xi} > -0.5$ , los EMV cumplen las propiedades asintóticas de normalidad, así que es posible calcular los intervalos de confianza del 95 % para cada parámetro:

$$\begin{aligned}\mu &\in (55.4990, 61.9696), \\ \sigma &\in (25.1254, 29.9662), \\ \xi &\in (-0.34294, -0.1642).\end{aligned}$$

Notamos que el IC del parámetro  $\xi$  no contiene el valor 0, por lo que descartamos la probabilidad, con un nivel de confianza del 95 %, de que la distribución Gumbel sea un buen modelo de ajuste para la dinámica de las concentraciones máximas de ozono de la estación BJU. Por lo tanto, los datos proporcionan evidencia suficiente para determinar que el modelo Weibull es el mejor para realizar el ajuste de los valores máximos.

### Validación del modelo Weibull

Los gráficos de diagnóstico de probabilidad (P-P) y de cuantiles (Q-Q) del modelo Weibull para la estación BJU se muestran en la Figura 5.10; en ambos casos se visualiza una clara proximidad de los puntos de los datos a las respectivas rectas identidad de referencia ( $y = x$ ), indicando que la distribución Weibull se ajusta adecuadamente a las concentraciones máximas diarias de  $O_3$ .

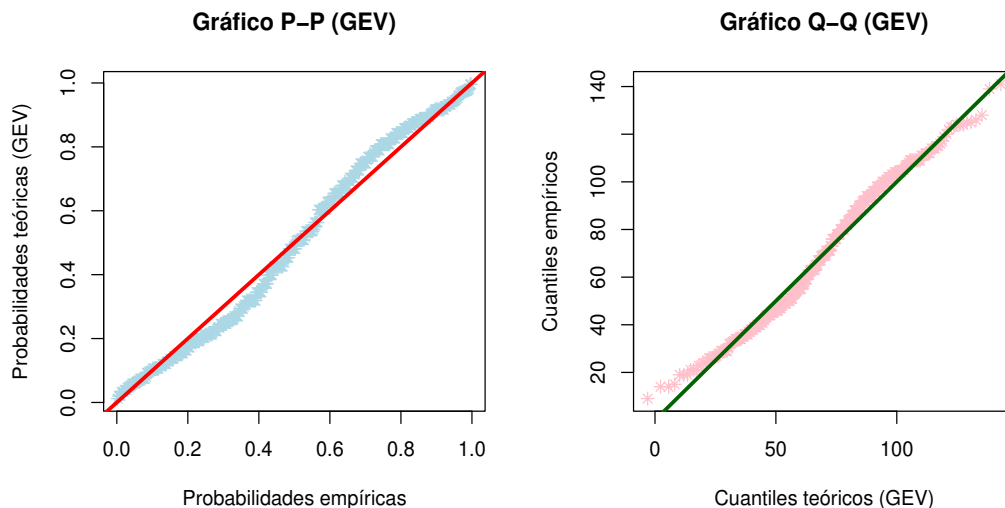


Figura 5.10: Gráficas de diagnóstico P-P y Q-Q del modelo Weibull: Estación BJU

Cabe mencionar que en este caso también se observan ligeras oscilaciones de los puntos, aunque son menos notorias que las que señaló la estación de Cuajimalpa, también indican posibles patrones cíclicos y variaciones estacionales en los datos que no captura el modelo. Sin embargo, esto no es prueba suficiente para refutar que el modelo Weibull se ajuste adecuadamente a los valores máximos de ozono.

Siguiendo con la validación del modelo con métodos gráficos, en la Figura 5.11 se visualiza el diagrama de retorno y el histograma de los datos reales con la función de densidad ajustada del modelo Weibull. Como sabemos, el histograma puede variar dependiendo de la agrupación que se elija; para esta estación también se seleccionaron 19 clases en un rango de 9 ppb hasta 142 ppb, de esta manera se asegura cubrir todos los datos y también se obtienen clases con la misma dimensión.

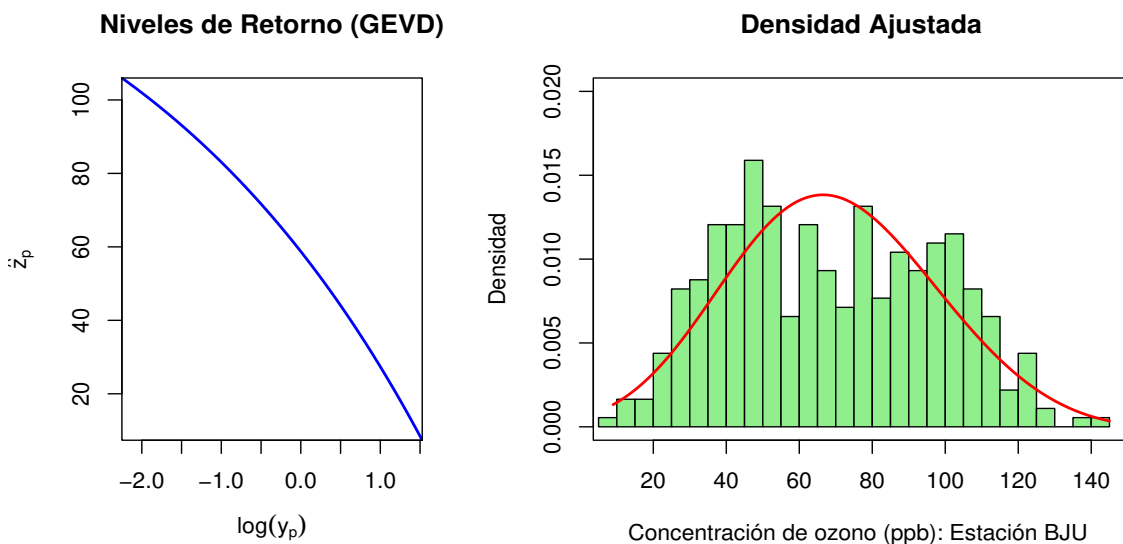


Figura 5.11: Gráfica de niveles de retorno y densidad ajustada (GEVD)

En el segundo gráfico, notamos una clara proximidad de la densidad ajustada del modelo Weibull con el histograma de los datos máximos de ozono. Además, la forma cóncava de la curva que se muestra en la gráfica de niveles de retorno son evidencia suficiente para determinar que la distribución del modelo seleccionado se ajusta adecuadamente a la dinámica de los valores máximos; por está razón, y por los resultados de los gráficos P-P y Q-Q, podemos concluir que **el modelo Weibull es adecuado para un buen ajuste de los registros máximos diarios de concentración de  $O_3$  en la estación de Benito Juárez.**

**Estimación de riesgos por concentración de O<sub>3</sub> en Benito Juárez**

Como ya hemos mencionado antes, los límites permisibles establecidos por la norma NOM-020-SSA1-2014 y que son aplicados por la NADF-009-AIRE-2017 en unidades ppb corresponden a los umbrales **95 ppb** y **154 ppb**. Sin embargo, en esta estación el máximo anual fue de 141, por lo que no hubo ningún día del año 2023 en la que se activara la Fase 1 de Contingencia por sobrepasar el segundo umbral. De manera que, nos enfocaremos en los días en los que hubo una **Mala** calidad del aire y en los efectos dañinos en la salud para los grupos susceptibles y las recomendaciones pertinentes.

A continuación, se realiza el ajuste de los registros de las concentraciones máximas de ozono de las mediciones en la estación BJU usando la distribución Weibull que está dada por:

$$G(z) = \exp \left\{ - \left[ 1 - 0.254 \left( \frac{z - 58.737}{27.547} \right)^{\frac{1}{(0.254)}} \right] \right\}, \quad z \in (167.1898, \infty).$$

Para complementar el análisis visual del comportamiento de las concentraciones máximas por día, se calculó la función de densidad de probabilidad ajustada del modelo Weibull, para  $z \in (167.1898, \infty)$ :

$$g(z) = \frac{1}{27.547} \left[ 1 - 0.254 \left( \frac{z - 58.737}{27.547} \right)^{\frac{1}{(0.254)}} \right]^{2.937} \exp \left\{ - \left[ 1 - 0.254 \left( \frac{z - 58.737}{27.547} \right)^{\frac{1}{(0.254)}} \right]^{3.937} \right\}.$$

Ambas funciones se muestran en la Figura 5.12.

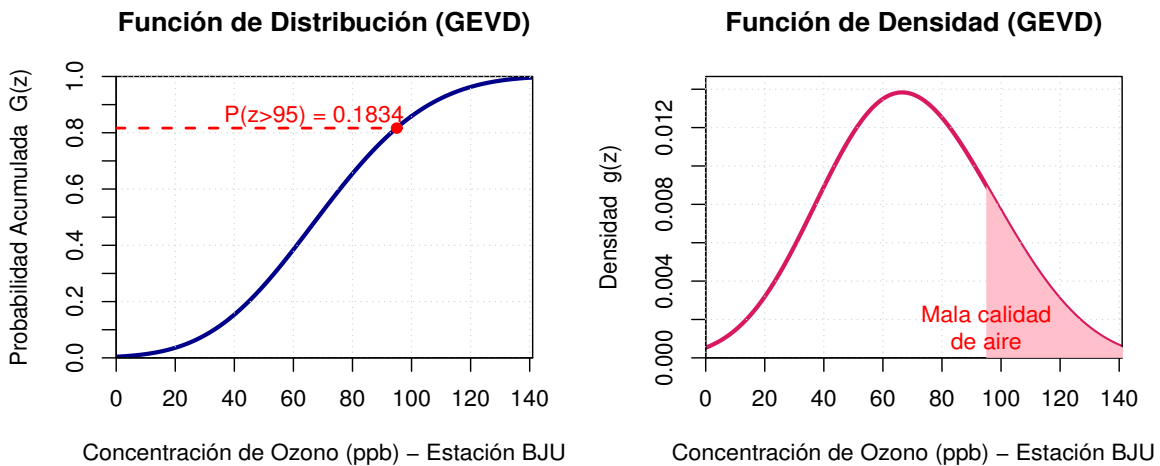


Figura 5.12: Funciones de distribución acumulada y densidad de probabilidad (GEVD)

La función de distribución nos indica que la probabilidad de tener una concentración promedio máxima por debajo de 95 ppb es del 81.66 % ( $\mathbb{P}(X \leq 95) = 0.8166$ ), lo que implica que la probabilidad de tener una **Mala** calidad del aire un día cualquiera del año es del 18.34 %. En la función de densidad podemos ver que en el 2023 hubo una gran cantidad de días que superaron este valor, para ser exactos fueron 84 días de 365 ( $\approx 23\%$ ) en los que se catalogó una **Mala** calidad del aire.

A pesar de todos los factores que se detectaron que pueden generar contaminantes primarios que componen el ozono troposférico; no hubo concentraciones demasiado altas de  $O_3$  registradas por la estación BJU. En general, durante el año 2023 se registró una calidad del aire buena o regular; esto puede deberse a los bajos niveles de temperatura que alcanza la zona o bien, a los fuertes vientos que se llevan parte de la concentración atmosférica hacia otra región de la ZMCM.

### Niveles de retorno

Para cuantificar el riesgo de excedencias peligrosas de concentración de ozono, se calculan los niveles de retorno correspondientes a los periodos de retorno de los días  $T = 365, 1825, 3650$  y  $7300$  que equivalen a 1, 5, 10 y 20 años, respectivamente. La Tabla 5.11 muestra los periodos y niveles de retorno, la varianza asociada a cada  $\hat{z}_p$  calculada con el método delta, y los IC del 95 % estimados para cada nivel de retorno.

Tabla 5.11: Periodos y niveles de retorno del modelo Weibull: Estación BJU.

$T$ (días)	$T$ (años)	Nivel de retorno	$Var(z_p)$	IC del 95 %
365	1	143.035 ppb	21.536	(133.94, 152.13)
1825	5	151.199 ppb	34.10	(139.75, 162.64)
3650	10	153.808 ppb	40.101	(141.40, 166.22)
7300	20	155.997 ppb	46.404	(142.64, 169.35)

Estos resultados estiman que, para un periodo de 10 años, el nivel de retorno es de 153.808 ppb con un IC del 95 % que demuestra su precisión; es decir que, en promedio, se espera que la concentración máxima diaria de ozono exceda este valor al menos una vez en los próximos 10 años. Como  $153.808 \approx 154$  ppb, se espera que dentro de los próximos 10 años se active la Contingencia en la Fase 1 al menos 1 vez.

Por último, se presenta la curva de los niveles de retorno de la distribución Weibull, donde notamos que las concentraciones máximas de ozono no superarán el límite teórico de  $\mu - \sigma/\xi \approx 167.386$  ppb, esto se refleja en el aplanamiento progresivo de la curva en la Figura 5.13.

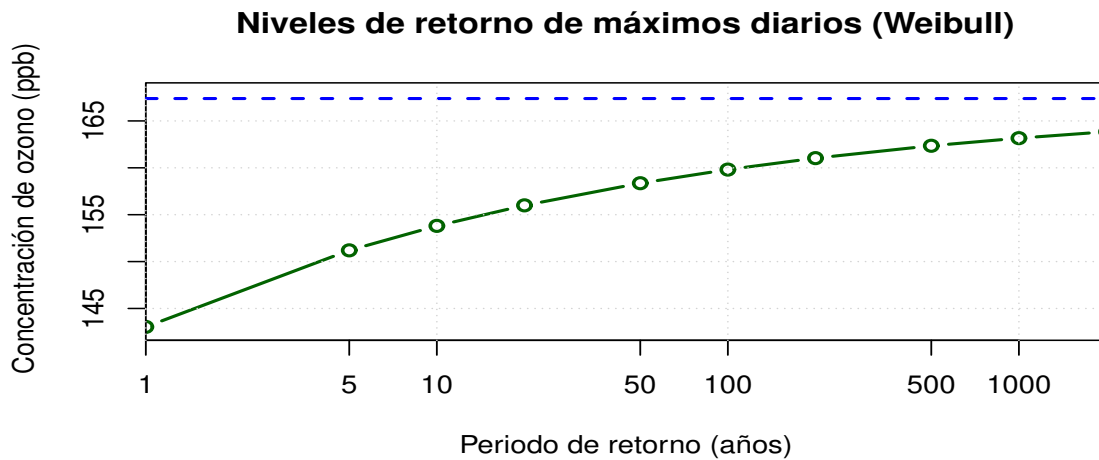


Figura 5.13: Niveles de retorno máximos por día (Weibull): Estación BJU

### Estación: La Presa (LPR)

Este centro de monitoreo está ubicado en el municipio de Tlalnepantla de Baz, en el Estado de México y forma parte de una zona urbana-industrial al norte de la ZMCM con dirección en: Asociación Mexicana de Excursionistas del D.F. s/n, Col. La Presa, C.P. 54189. La caseta de monitoreo comenzó a operar desde 1986 y sólo realiza mediciones de la concentración atmosférica de  $O_3$ ,  $CO$ ,  $NO_2$  y  $SO_2$  considerando variables meteorológicas de temperatura, radiación solar, humedad y velocidad del viento; para los registros de concentración de  $O_3$  está catalogada como tipo 3-Vecinal, teniendo un alcance de medición de 0.5 a 4 km. [SEDEMA-SIMAT].



Figura 5.14: Caseta de monitoreo de La Presa (ID:484151040203)

En el 2023, se estimó una población total de 714,799 habitantes en el municipio de Tlalnepantla, con una de las más altas densidades poblacionales del Edomex con 6,210 habitantes por km<sup>2</sup>. Esta zona urbana-industrial tiene dos infraestructuras cercanas que son críticas en la contaminación del aire; el Parque Industrial Tlalnepantla que se ubica a 3 km al sur y la Autopista México-Querétaro a 2 km al este.

Se ha registrado una temperatura promedio de 18.2°C, alcanzando un máximo histórico de 34.6°C en mayo del 2023 debido a una ola de calor; la media anual de la velocidad del viento se estima en 2.8 m/s con una dirección predominante del noreste que se dirige hacia el suroeste.

Las principales fuentes de COV y NO<sub>x</sub> que contribuyen a la formación del ozono troposférico y que provocan alteraciones en las mediciones de la concentración atmosférica registradas por la caseta LPR son emitidas por:

- **Alta densidad industrial.** Además del Parque Industrial Tlalnepantla, hay más de 2,000 industrias químicas registradas dentro de la zona que son emisoras de gran cantidad de NO<sub>x</sub>.
- **Vialidades de alto tráfico.** Hay congestiones frecuentes en Periférico Norte, tránsito intenso en la Avenida Gustavo Baz y un flujo constante en la Autopista México-Querétaro; todas son emisoras de COV evaporatorios.
- **Áreas verdes.** La principal fuente de COV biogénicos es el Bosque de los Tepetates que se encuentra a 2.3 km al noroeste de la estación LPR.
- **Factores meteorológicos.** Además de los vientos dominantes y de la temperatura promedio, la inversión térmica atrapa contaminantes cerca de La Presa, inhibiendo la dispersión del ozono troposférico.

### Análisis descriptivo general de los registros horarios

La estación LPR tiene un porcentaje de completitud del 95.91 % con 8402 datos válidos del total, siendo el porcentaje más alto (en las mediciones de ozono) de todas las estaciones que forman parte de la RAMA, esto implica que hubo 358 datos (4.09 %) que tuvieron que ser sustituidos con el promedio horario.

En la Tabla 5.12 se observan los resultados del análisis descriptivo de todos los registros horarios de la estación LPR, esta información es relevante ya que nos ofrece un panorama general sobre el comportamiento de la concentración de O<sub>3</sub> registrada en la zona de estudio en el año 2023.

Tabla 5.12: Estadísticas descriptivas generales: Estación LPR

$\bar{x}$	$SE$	Mín	$Q_1$	$Q_2$	$Q_3$	Máx	IQR	$\gamma_1$	$\gamma_2$
18.76	18.78	0	4	14	27	126	23	1.83	7.59

Se obtuvo un promedio anual de 18.76 ppb, valor que supera a la mediana de 14 ppb, indicando nuevamente asimetría positiva; lo cual se confirma con el valor positivo de  $\gamma_1$ . La gran diferencia entre el rango total con el rango intercuartílico hacen suponer poca dispersión de los datos; donde el 50 % de los registros centrales se concentran entre 4 ppb y 27 ppb, por lo cual se deduce un pico alto en las concentraciones más bajas de  $O_3$ . Lo anterior se confirma con el alto valor de la curtosis de 7.59 ppb, aseverando una distribución con forma leptocúrtica con la cola derecha muy alargada, que podría implicar la presencia de valores atípicos. Finalmente, en esta estación el valor máximo anual fue de **126 ppb**.

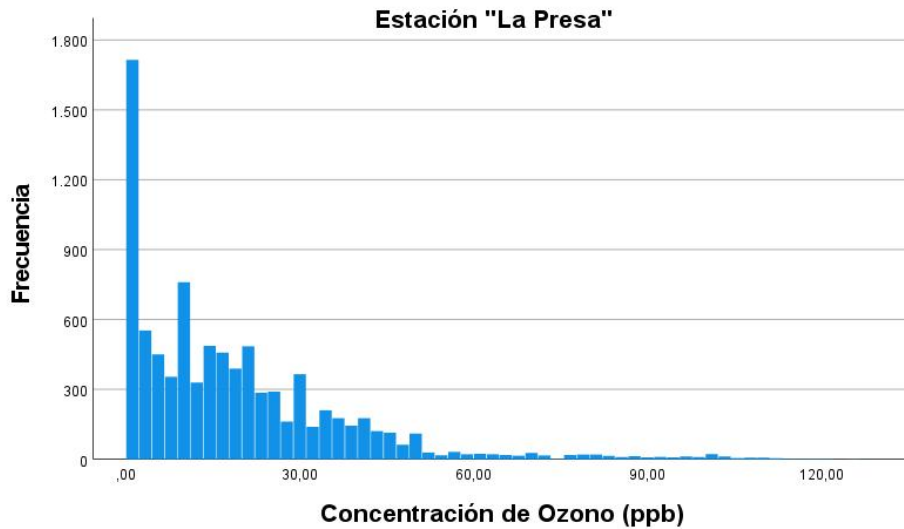


Figura 5.15: Histograma de frecuencias de los registros horarios: Estación LPR

En esta ocasión, se agrega el histograma de frecuencias de los registros horarios de concentración de ozono medidos por la estación LPR en el año 2023 para verificar de manera visual lo que se acaba de inferir con los resultados teóricos de la tabla de estadísticos descriptivos. La gráfica de la Figura 5.15, en efecto, demuestra todo el razonamiento anterior: hay una notoria asimetría positiva en forma leptocúrtica con un pico muy alto en las concentraciones más bajas y una cola derecha muy alargada. En general, los datos de la concentración de  $O_3$  de casi todas las estaciones de monitoreo tienen un comportamiento muy similar; con asimetría positiva y formas leptocúrticas, variando en el peso de la cola derecha y en la dispersión de las mediciones.

### Análisis de las máximas concentraciones diarias

Siguiendo con la misma metodología de los análisis anteriores, se muestra la Tabla 5.13 con la distribución de los 365 bloques organizados en intervalos por mes, donde se muestran los máximos mensuales y el bloque al que pertenece. De este modo, se logró identificar que el máximo anual de 126 ppb le pertenece al bloque 323, proveniente del mes de noviembre; para ser más exactos, el máximo anual de 126 ppb fue registrado el día 19 de noviembre del 2023.

Tabla 5.13: Intervalos de bloques con registros máximos mensuales: Estación LPR

Mes	Bloques	B.Máx	Máx	Mes	Bloques	B.Máx	Máx
<b>Enero</b>	1-31	18	121	<b>Julio</b>	182-212	185	82
<b>Febrero</b>	32-59	43	115	<b>Agosto</b>	213-243	223	118
<b>Marzo</b>	60-90	83	119	<b>Septiembre</b>	244-273	257	112
<b>Abril</b>	91-120	93	112	<b>Octubre</b>	274-304	292	104
<b>Mayo</b>	121-151	143	107	<b>Noviembre</b>	305-334	<b>323</b>	<b>126</b>
<b>Junio</b>	152-181	152	108	<b>Diciembre</b>	335-365	353	84

Siguiendo con el análisis del comportamiento de los máximos diarios de concentración de O<sub>3</sub> registradas por la estación LPR, es necesario presentar la Tabla 5.14 con los resultados teóricos del análisis descriptivo de las concentraciones máximas de ozono. Además, para visualizar la dinámica de estos datos, se presentan en conjunto el diagrama de dispersión y el Box-plot en la Figura 5.16.

Tabla 5.14: Descriptivos de los máximos por día: LPR

Mín	Q <sub>1</sub>	Q <sub>2</sub>	Media	Q <sub>3</sub>	Máx	IQR
<b>8.50</b>	<b>36</b>	<b>46</b>	<b>55.09</b>	<b>73</b>	<b>126</b>	<b>37</b>

En contraste con los análisis de las estaciones CUA y BJU, podemos notar una mayor diferencia entre la media de 55.09 ppb y la mediana de 46 ppb, sugiriendo que hay un ligero sesgo a la derecha en la distribución de los máximos. En ambos gráficos, donde la línea roja representa el valor de la media, es más evidente dicho sesgo: en el diagrama de dispersión se distingue por una mayor cantidad de puntos azules por debajo del promedio y en el box-plot se muestra por la diferencia en el tamaño de los "bigotes", siendo el inferior más corto que el superior, y por el promedio que ya no se sobrepone a la línea de la mediana.

Por todo lo anterior, se confirma que la dispersión de los datos máximos para esta estación presenta una ligera asimetría positiva, con la cola derecha más alargada. El rango intercuartílico de 37 ppb, representado por las líneas naranjas en el diagrama de dispersión y por el tamaño de la "caja" verde en el box-plot, indica que el 50% de los datos se concentra entre 36 ppb y 73 ppb. La "caja" puede parecer estrecha en comparación con la dispersión total sugiriendo una distribución leptocúrtica, sin embargo, al no haber evidencia de valores atípicos y por la poca dispersión de los datos (bigotes cortos) se afirma que la cola derecha de la asimetría de la distribución es ligera, dando indicios de una forma más platicúrtica. Lo anterior se confirma calculando el valor de la curtosis ( $\gamma_2 = 2.5783 < 3$ ), demostrando una forma platicúrtica no muy pronunciada.

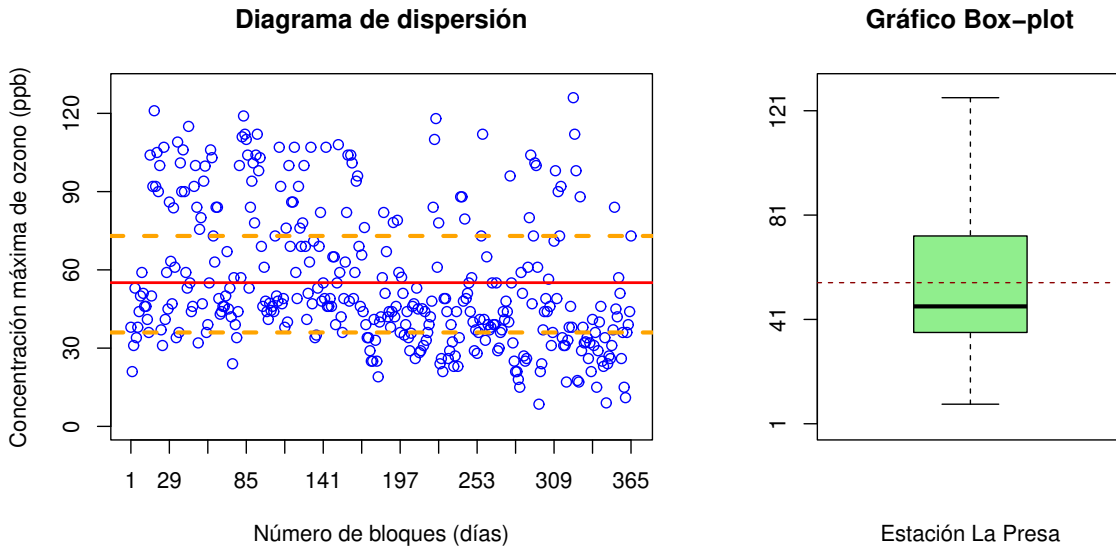


Figura 5.16: Gráfica de dispersión y Boxplot: Estación LPR

Esta forma distintiva que muestran los valores máximos diarios es consecuencia de la gran cantidad de registros bajos en la concentración horaria de ozono que se hicieron en la estación LPR, tal como se mostró en el histograma de frecuencias.

### Ajuste de los registros máximos diarios a una DVEG

Los estimadores de los parámetros de localización, escala y forma que se obtuvieron con el método de máxima verosimilitud fueron:

$$\hat{\mu} = 42.335, \quad \hat{\sigma} = 19.561, \quad \hat{\xi} = 0.068.$$

Los errores estándar se calcularon con la raíz cuadrada de la varianza, mediante la matriz de covarianza  $\Sigma$ , para cada estimador se obtuvo:

$$SE_{\hat{\mu}} = 1.175, \quad SE_{\hat{\sigma}} = 0.886, \quad SE_{\hat{\xi}} = 0.046.$$

Para la estación de La Presa, el estimador  $\hat{\xi} = 0.068 > 0$  sugiere que el modelo de distribución a la que se ajustan las concentraciones máximas diarias de ozono es un modelo Fréchet. Este valor de  $\hat{\xi}$  descarta por completo el modelo Weibull, ya que no es adecuado bajo el criterio de estimación puntual; pero al ser un valor cercano a 0, no podemos descartar la posibilidad de que el modelo Gumbel también se ajuste adecuadamente a la distribución de los datos, por lo que es necesario calcular los IC.

El estimador de forma ( $\hat{\xi} > -0.5$ ) establece que los EMV son regulares y cumplen las propiedades asintóticas de normalidad, por lo que es posible calcular los intervalos de confianza para cada parámetro a un nivel del 95 %:

$$\begin{aligned} \mu &\in (40.0307, 44.6342), \\ \sigma &\in (17.8237, 21.295), \\ \xi &\in (-0.0211, 0.1577). \end{aligned}$$

Se puede notar que el IC del 95 % para  $\xi$  captura el valor 0, lo cual indica la posibilidad de ajustar los máximos diarios con una distribución Gumbel; por otro lado, valores muy negativos de  $\xi$  son poco plausibles, confirmando que el modelo Weibull no es adecuado a la luz de los datos observados. Por lo tanto, los datos no proporcionan evidencia suficiente en contra de los modelos Fréchet y Gumbel de la familia de DVEG.

### Prueba de hipótesis

Es importante recordar que siempre se busca representar adecuadamente un fenómeno, en este caso atmosférico, a través de un modelo estadístico con el cual sea posible realizar inferencias de interés. Estas inferencias pueden variar dependiendo del modelo elegido, por lo cual es indispensable que el modelo seleccionado se ajuste lo “mejor” posible al conjunto de datos disponible.

Ante estos casos de incertidumbre en la elección del mejor modelo de la familia de DVEG que mejor se ajuste a las concentraciones máximas diarias de  $O_3$ , se recomienda realizar una prueba de bondad de ajuste, bajo la hipótesis nula de que la distribución del modelo Gumbel es la más adecuada para ajustar los datos máximos, es decir, bajo la suposición de  $H_0 : \hat{\xi} = 0$  vs  $H_1 : \hat{\xi} \neq 0$ .

Definiendo el vector de estimadores  $\hat{\theta} = (\hat{\mu}, \hat{\sigma}, \hat{\xi})$  y asumiendo  $H_0$  verdadera, deberemos calcular el vector  $\hat{\theta}_0 = (\hat{\mu}_0, \hat{\sigma}_0, 0) = (\hat{\mu}_0, \hat{\sigma}_0)$ , que corresponden a los EMV del modelo Gumbel:

$$\hat{\mu}_0 = 43.050, SE_{\hat{\mu}_0} = 1.105; \quad \hat{\sigma}_0 = 20.089, SE_{\hat{\sigma}_0} = 0.847.$$

Para el estadístico de prueba usaremos la razón de verosimilitud (función de devianza)  $D_0 = -2 \ln[L(\hat{\theta}_0)/L(\hat{\theta})]$ , donde  $L(\hat{\theta}_0)$  y  $L(\hat{\theta})$  son funciones de verosimilitud basadas en los modelos de la DVEG para  $\hat{\xi} = 0$  y  $\hat{\xi} \neq 0$ , respectivamente; en este caso, basadas en los modelos Gumbel y Fréchet.

Bajo condiciones de regularidad, este estadístico de prueba se distribuye como una  $\chi^2$  con 1 grado de libertad, indicado por la diferencia de parámetros entre modelos.

Para esta prueba de hipótesis se elegirá un nivel de significancia  $\alpha = 0.05$ , así la región de rechazo queda determinada por el cuantil 0.95 de una distribución  $\chi_1^2$ , es decir,  $C_\alpha = (3.84, \infty)$ .

Al calcular el valor del estadístico de prueba y el p-valor, obtenemos:

$$\begin{aligned} D_0 &= -2 \ln \left[ \frac{L(\hat{\mu}_0, \hat{\sigma}_0)}{L(\hat{\mu}, \hat{\sigma}, \hat{\xi})} \right] = -2 \ln \left[ \frac{L(43.050, 20.089)}{L(42.335, 19.561, 0.068)} \right] \\ &= -2 [ \ell(43.050, 20.089) - \ell(42.335, 19.561, 0.068) ] \\ &= -2 [ -1688.32 - (-1677.718) ] = 21.205 \in (3.84, \infty). \end{aligned}$$

$$\text{p-valor} = \mathbb{P}[\chi_1^2 > 21.205] = 1 - \mathbb{P}[\chi_1^2 \leq 21.205] = 0.000006 < 0.05 \Rightarrow \text{p-valor} < \alpha.$$

Como  $D_0 = 21.205$  cae dentro de la región de  $C_\alpha$ , **se rechaza la hipótesis nula** eliminando por completo la posibilidad de que el modelo Gumbel sea un buen modelo de ajuste para los datos con un nivel de significancia  $\alpha = 0.05$ .

El resultado del p-valor  $< \alpha$  nos indica que en esta prueba de hipótesis no se cometió el error de ningún tipo. Por lo tanto, los datos proporcionan evidencia suficiente para determinar que el modelo Fréchet es el mejor para ajustar las concentraciones máximas diarias de  $O_3$  provenientes de la estación La Presa.

### Validación del modelo Fréchet

Ahora, se utilizarán las gráficas de probabilidad (P-P) y de cuantiles (Q-Q) como métodos de diagnóstico, su interpretación conjunta permite detectar desviaciones en la forma, colas y tendencia central de los datos respecto al modelo teórico de Fréchet. Posteriormente, se usará el diagrama de retorno para la validación final del modelo.

En la Figura 5.17 se presentan ambos gráficos; en donde la proximidad de los datos a la recta identidad en el gráfico de probabilidad es más evidente que en el gráfico de cuantiles. Es relevante notar que en el gráfico P-P no se detectan desviaciones pronunciadas en los extremos, esto quiere decir que el parámetro de forma  $\xi$  captura adecuadamente el comportamiento de las colas de la distribución ajustada. Sin embargo, se distingue una ligera oscilación que no debería ser ignorada, ya que podría indicar patrones cíclicos en donde las mediciones actuales dependen de mediciones anteriores, lo cual es común en procesos estacionarios.

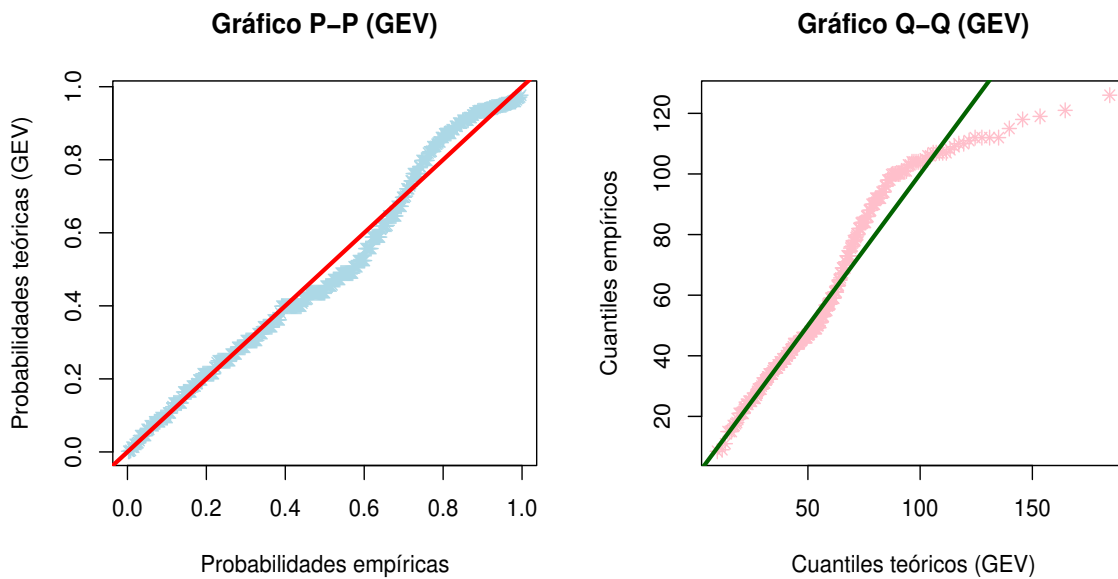


Figura 5.17: Gráficas de diagnóstico P-P y Q-Q del modelo Fréchet: Estación LPR

En contraste con lo anterior, la desviación hacia abajo de la cola derecha del gráfico de cuantiles indica que los valores máximos de los datos tienen colas más ligeras que las de Fréchet, sugiriendo una sobrestimación del parámetro  $\xi$  y que debería ser reducido; esto pudo ser consecuencia del sesgo positivo que presentaron los datos máximos y de la incertidumbre que hubo en la elección del modelo.

Es importante tener en cuenta que el gráfico de cuantiles es más sensible a desviaciones en colas, por lo que la sobrestimación del parámetro  $\xi$  en la gráfica Q-Q no es prueba suficiente para señalarlo como inapropiado. Además, el resultado teórico de la prueba de bondad de ajuste mostró que este parámetro en particular si es el adecuado.

En general, podemos decir que los gráficos de diagnóstico confirman que el modelo Fréchet captura adecuadamente la distribución de los datos, indicando que los parámetros estimados  $\mu$ ,  $\sigma$  y  $\xi$  son apropiados.

Continuando con la validación del modelo con métodos gráficos, en la Figura 5.18 se muestra la gráfica de niveles de retorno y la gráfica de la función de densidad ajustada con el histograma de los datos reales.

La proximidad de la gráfica de niveles de retorno a una línea recta, podría haber sugerido que el modelo Gumbel se ajustaba de manera adecuada a los datos, sin embargo ya se eliminó esa posibilidad con una prueba formal teórica. Por lo tanto, la ligera concavidad que se logra distinguir en la curva es evidencia suficiente de que el modelo Fréchet es adecuado para los máximos diarios de concentración de  $O_3$  registrados en la estación de La Presa.

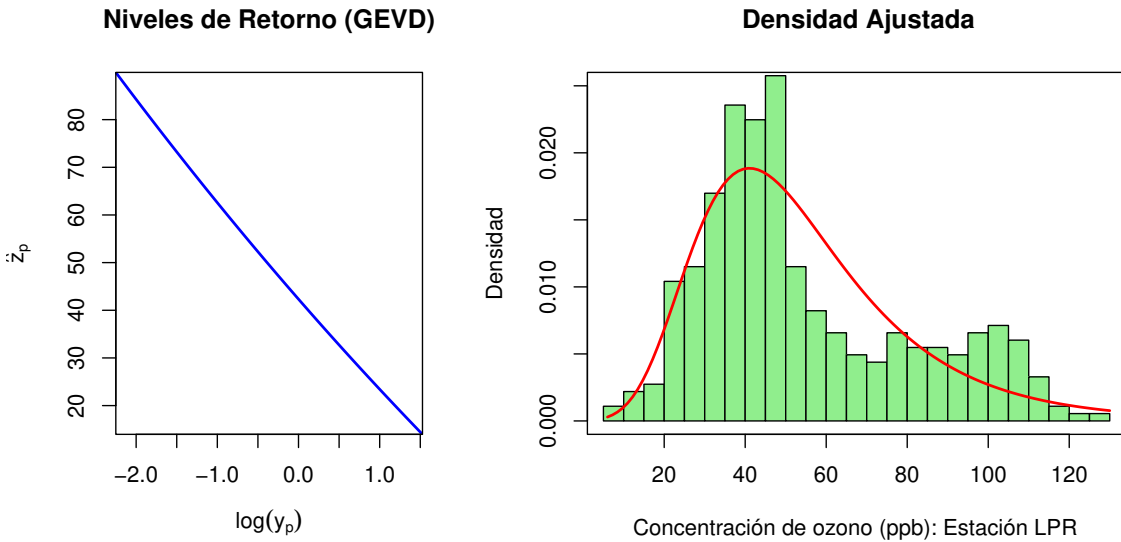


Figura 5.18: Gráfica de niveles de retorno y densidad ajustada (GEVD)

En el gráfico de la función de densidad de probabilidad ajustada se compara la función de densidad del modelo Fréchet con el histograma de los máximos diarios de concentraciones atmosféricas de ozono medidos por la estación de monitoreo LPR.

Aunque este último gráfico suele ser menos informativo que los anteriores, para este análisis en particular nos ayuda a visualizar la forma asimétrica que se infirió en la distribución de los datos máximos, verificando la existencia de un pequeños sesgo positivo y demostrando que la cola derecha es ligera provocada por la forma platicúrtica que casi no se distingue.

Finalmente, por los resultados del análisis de verificación con métodos gráficos y por el resultado teórico de la prueba de hipótesis concluimos que **el modelo Fréchet es un buen modelo de ajuste.**

### Estimación de riesgos por concentración de O<sub>3</sub> en La Presa

Debemos recordar que los límites permisibles establecidos por la normativa bajo la cual se regula la base de datos, cataloga una **Mala** calidad del aire a partir de concentraciones de O<sub>3</sub> por arriba de **95 ppb**, ocasionando que los grupos susceptibles presenten efectos dañinos en su salud, y deberán limitar su exposición al aire libre.

Para visualizar la dinámica de las concentraciones atmosféricas máximas de O<sub>3</sub> registradas en la estación LPR usaremos la función de distribución acumulada y la función de densidad de probabilidad ajustada del modelo Fréchet.

La función de distribución está dada por:

$$G(z) = \exp \left\{ - \left[ 1 + 0.068 \left( \frac{z - 42.335}{19.561} \right) \right]^{-\frac{1}{0.068}} \right\}, \quad z \in (-245.3268, \infty).$$

La función de densidad de probabilidad ajustada para  $z \in (-245.3268, \infty)$  es:

$$g(z) = \frac{1}{19.561} \left[ 1 + 0.068 \left( \frac{z - 42.335}{19.563} \right) \right]^{-15.706} \exp \left\{ - \left[ 1 + 0.068 \left( \frac{z - 42.335}{19.563} \right) \right]^{-14.706} \right\}.$$

La gráfica de cada función se muestra en la Figura 5.19.

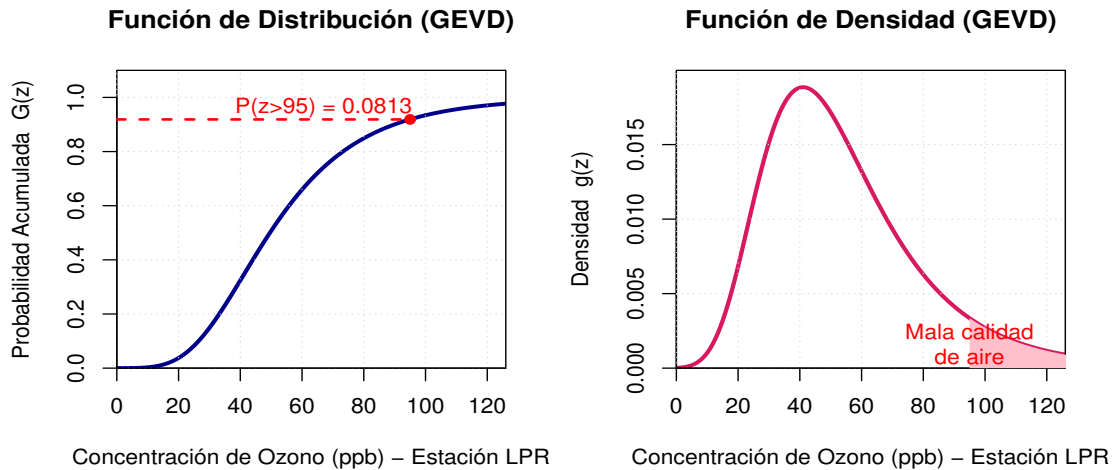


Figura 5.19: Funciones de distribución acumulada y densidad de probabilidad (GEVD)

La función de distribución indica que la probabilidad de tener una concentración promedio máxima por debajo de 95 ppb es del 91.87% ( $\mathbb{P}(X \leq 95) = 0.9187$ ), esto implica que la probabilidad de tener una **Mala** calidad del aire un día cualquiera del año es del 8.13%. La función de densidad determina que en el año 2023 hubo 46 días en los que superó este valor umbral.

En general, se visualiza que durante el año 2023 la calidad del aire en la Col. La Presa, se catalogó entre buena y moderada debido a los bajos registros que se obtuvieron de esta estación. Además, no hubo ningún día del año 2023 en el que se haya activado una Contingencia Fase 1 por superar el umbral de 154 ppb, por tanto, los datos de la estación LPR no muestran probabilidad de tener una Mala calidad del aire.

### Niveles de retorno

Para cuantificar el riesgo de excedencias peligrosas de concentración de ozono se calculan los niveles de retorno para  $T = 1, 5, 10$  y  $20$  años, respectivamente. Los resultados se muestran en la Tabla 5.15 junto con la varianz y los IC del 95% asociados a cada  $\hat{z}_p$ .

Tabla 5.15: Periodos y niveles de retorno del modelo Fréchet: Estación LPR

$T$ (días)	$T$ (años)	Nivel de retorno	$Var(z_p)$	IC del 95%
365	1	184.389 ppb	3.518	(180.712, 188.066)
1825	5	234.189 ppb	5.105	(229.761, 238.617)
3650	10	257.369 ppb	5.819	(252.642, 262.098)
7300	20	281.671 ppb	6.545	(276.656, 286.685)

La gráfica muestra un crecimiento sublineal debido al bajo valor de  $\xi = 0.068$ , esta distribución no alcanza un límite superior teórico, por lo que los niveles de retorno seguirán creciendo indefinidamente; este comportamiento y los resultados de la tabla reflejan mayor incertidumbre en estimaciones más extremas, es recomendable evaluar la sensibilidad del modelo debido a la naturaleza asintótica de la distribución Fréchet.

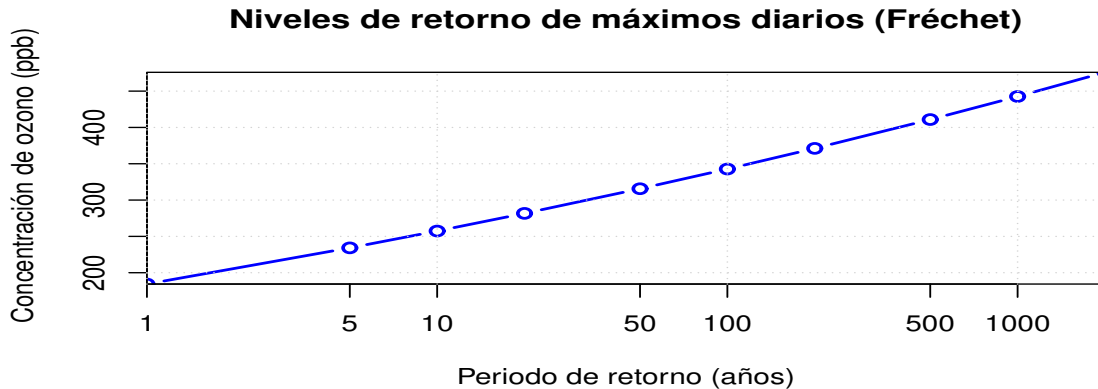


Figura 5.20: Niveles de retorno máximos por día (Fréchet): Estación LPR

## Activación de Contingencias por Ozono en el 2023

En la actualidad la contaminación atmosférica es considerado uno de los riesgos ambientales que tiene mayor impacto en la salud. Los hallazgos científicos en el mundo, incluyendo estudios realizados en México, demuestran que el aumento de la contaminación del aire incrementa la morbilidad y mortalidad en las personas por causas respiratorias o cardiovasculares, siendo el ozono troposférico el principal causante.

La magnitud de los efectos depende de diversos factores como el tiempo, la frecuencia de exposición, la concentración del contaminante en el aire y las características de la población expuesta. La dinámica de la ZMCM es un factor importante que influye directamente en la calidad del aire, en la CDMX se rebasan los límites máximos permisibles de ozono en varios días al año. Por esta razón, se requieren mecanismos para informar a la población de manera adecuada y oportuna sobre los niveles de contaminación y su variación en el tiempo, con el fin de salvaguardar la salud pública.

El SIMAT es uno de los principales instrumentos de gestión de las políticas ambientales del aire de la Secretaría del Medio Ambiente del Gobierno de la Ciudad de México, con él se evalúa la eficacia de las acciones implementadas en la región metropolitana. La metodología para el cálculo del Índice de Calidad del Aire utiliza las Normas Oficiales Mexicanas en materia de Salud Ambiental vigentes, este índice debe ser revisado periódicamente y modificado cuando existan cambios en las NOM o cuando se presente una nueva evidencia de riesgo.

En el contexto de la calidad del aire, la TVE se aplica para estimar riesgos asociados a concentraciones extremas de  $O_3$ , un contaminante crítico que afecta la salud respiratoria y cardiovascular. Mediante el análisis de concentraciones máximas de  $O_3$ , la TVE permite diseñar alertas tempranas y políticas de reducción de emisiones, lo cual es clave para evaluar el cumplimiento de normativas ambientales y proyectar impactos del cambio climático en la formación de  $O_3$ .

Las contingencias ambientales de ozono son eventos atípicos y transitorios en los que se alcanzan valores muy altos de concentración de  $O_3$  y son el resultado de una combinación específica de factores meteorológicos como la estabilidad atmosférica, altas temperaturas, viento débil, intensa radiación solar y cielos despejados que son causados por los sistemas de alta presión.

La Comisión Ambiental de la Megalópolis (CAME), es la entidad responsable de activar o desactivar las contingencias ambientales atmosféricas y considera los datos provenientes de la RAMA del SIMAT, así como la revisión de los pronósticos meteorológicos para el día actual y el siguiente; con la finalidad de estimar si las condiciones serán favorables para la dispersión de contaminantes.

## CAPÍTULO 5. APLICACIÓN: CONCENTRACIONES MÁXIMAS DE OZONO

En la Tabla 5.16 se presentan los umbrales establecidos por los límites permisibles de acuerdo a la normativa NOM-020-SSA1-2014 para activar el programa de contingencias, el cual establece bajo qué condiciones se determina la activación y suspensión de las Fases de contingencia por O<sub>3</sub>.

Tabla 5.16: Programa de contingencias ambientales atmosféricas vigente 2022

FASES PREVENTIVAS		
Condiciones de Activación		Suspensión
<b>Pronóstico O<sub>3</sub> &gt; 142 ppb</b> <b>Prob. de ocurrencia del 70 %</b>		<b>Al día siguiente (19:00 h) o con la activación de Fase de Contingencia.</b>
FASES DE CONTINGENCIA		
Contingencia	Activación	Suspensión
<b>FASE I (FI)</b>	<b>&gt; 154 ppb</b>	<b>Concentración menor a la de la Fase con pronóstico meteorológico favorable sig. día.</b>
<b>FASE II (FII)</b>	<b>&gt; 204 ppb</b>	

En el 2023 se presentaron cuatro Contingencias Fase I por ozono en la ZMCM. En la Tabla 5.17 se presenta la cronología de estos eventos, en los que las estaciones meteorológicas de la RAMA registraron una concentración de O<sub>3</sub> mayor a 154 ppb, se presenta el registro máximo de las casetas de monitoreo ubicadas por zona. Se puede realizar la consulta en la página oficial del SIMAT [SEDEMA (CONSULTAS)].

Tabla 5.17: Cronología de las contingencias atmosféricas durante 2023

Índice de Calidad del Aire mayores a 154 ppb para O <sub>3</sub> por Zona					
Fecha	Noroeste	Noreste	Central	Suroeste	Sureste
<b>23/02/2023</b>	<b>123.32</b>	<b>122.14</b>	<b>155</b>	<b>171</b>	<b>144.56</b>
<b>25/03/2023</b>	<b>161</b>	<b>131.58</b>	<b>138.66</b>	<b>132.76</b>	<b>124.5</b>
<b>26/03/2023</b>	<b>137.48</b>	<b>130.4</b>	<b>145.74</b>	<b>157</b>	<b>120.96</b>
<b>20/11/2023</b>	<b>122.14</b>	<b>158</b>	<b>99.72</b>	<b>139.84</b>	<b>68.6</b>

Se observó que la primer Contingencia FI se presentó el día 23 de febrero con una concentración máxima de 171 ppb registrada al Suroeste de la ZMCM, la región Central también activó la Fase I con una concentración de 155 ppb y en la zona Sureste solo se activó la Fase Preventiva ya que sobrepasó el umbral de 142 ppb.

El segundo evento fue a finales del mes de marzo y tuvo 2 días de duración con un registro máximo de 161 ppb en las estaciones de la zona Noroeste y otra activación FI al día siguiente en el Suroeste con 157 ppb, en el segundo día la zona Central activó una Contingencia Preventiva por registrar una concentración de O<sub>3</sub> mayor a 142 ppb.

El 20 de noviembre se activó por última vez la Contingencia FI por una concentración de 158 ppb reportada en el Noreste de la ZMCM. En contraste, ese mismo día se catalogó una Buena calidad del aire en la zona Sureste, pues no sobrepasó el límite de concentración de 70 ppb para  $O_3$ .

En el análisis de las estaciones CUA, BJU y LPR solo hubo un registro mayor a 154 ppb para la concentración de  $O_3$ , detectado el día 23 de febrero en la Alcaldía de Cuajimalpa de Morelos (CUA), en la zona Suoeste con una medición de 161 ppb.

Para la elaboración de esta tesis se estimaron los niveles de retorno para períodos de 1, 5, 10 y 20 años (365, 1825, 3650 y 7300 días). Sin embargo, para este fenómeno natural, es de mayor interés realizar el pronóstico con periodos de retorno de 24 a 48 horas, para estimar la probabilidad de ocurrencia de superar los umbrales al siguiente día y tomar las medidas preventivas necesarias.

Finalmente, es importante comprender que estos pronósticos son herramientas complementarias que funcionan como una guía para la toma de decisiones, y que su fiabilidad dependerá del desempeño que tengan a lo largo del tiempo. Sin embargo, es necesario buscar tener la menor incertidumbre posible con el objetivo de ofrecer una mejor información y otorgar mayor protección a la salud de la población.

# Conclusiones

En este trabajo de tesis se implementó la TVE para modelar tendencias de contaminación atmosférica provocada por  $O_3$  aplicada al caso particular de la contaminación ambiental de la CDMX en el año 2023. Siguiendo un enfoque paramétrico, se supuso que los registros máximos de cada día de medición de ozono se pueden modelar de acuerdo a una DVEG, cumpliendo con las condiciones de regularidad y aplicando normalidad asintótica para los estimadores.

Para realizar este análisis, se seleccionaron estratégicamente tres estaciones meteorológicas que forman parte de la RAMA; la estación CUA de Cuajimalpa de Morelos, la estación BJU de la Alcaldía Benito Juárez en CDMX y la estación LPR de la colonia La Presa, ubicada en el municipio de Tlalnepantla de Baz, en el Edomex.

En general, las concentraciones máximas diarias de ozono condujeron a modelos adecuados de ajuste de la familia de DVEG, aunque se detectaron ciertas peculiaridades que sugieren independencia de los datos o patrones cíclicos de temporalidad no identificados por el modelo. Además, se aplicó inferencia estadística estableciendo periodos de retorno y calculando los niveles de retorno y la probabilidad de ocurrencia, los cuales están asociados a la gestión o estimación de riesgos.

# Acrónimos

Acrónimo	Definición
CDMX	Ciudad de México
COV	Compuestos Orgánicos Volátiles
EDOMEX	Estado de México
IAS	Índice Aire y Salud
ICA	Índice de Calidad del Aire
IMECA	Índice Metropolitano de la Calidad del Aire
NADF	Norma Atmosférica del Distrito Federal
NOM	Norma Oficial Mexicana
OMS	Organización Mundial de la Salud
RAMA	Red Automática de Monitoreo Atmosférico
SEDEMA	Secretaría del Medio Ambiente de la Ciudad de México
SIMAT	Sistema de Monitoreo Atmosférico de la Ciudad de México
SMN	Servicio Meteorológico Nacional
SSA	Secretaría de Salud del Gobierno Federal
UV-A y UV-B	Radiación Ultravioleta A y Ultravioleta B
ZMCM	Zona Metropolitana de la Ciudad de México
ZMVM	Zona Metropolitana del Valle de México

# Bibliografía

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

Castillejos M., Gold D., Dockery D., et al. (1992). *Effects of ambient ozone on respiratory function and symptoms in México City*. American Review of Respiratory Disease, 145, 276-282.

Coles, S., Bawa, J., Trenner, L. Dorazio, P. (2001). *An introduction to statistical modeling of extreme values*. London: Springer, 208.

Colson, R., Fenelon, D., Smith, S. (1999). *What is ozone?* [PDF]. Farmingdale State College. Recuperado de <https://www.gnydm.com/images/posters/2021/8-Smith,%20S.pdf>

Contreras Ruíz, J. L. (2020). *Análisis de temperaturas máximas en el estado de Oaxaca* [Tesis de licenciatura]. Universidad Tecnológica de la Mixteca, Oaxaca, México.

Cuevas, R. G. I. (2011). *Teoría de valores de extremos empleada en la gestión de riesgo financiero* [Tesis de ingeniería]. Universidad de Chile, Santiago de Chile.

Fahey, D. (2002). *Veinte preguntas y respuestas sobre la capa de ozono* [PDF]. Evaluación Científica del Agotamiento de Ozono. Recuperado de <https://www.gob.mx/cms/uploads/attachment/file/31180/20PreguntasdeOzono.pdf>

Fisher, R. A., Tippett, L. H. C. (1928). *Limiting forms of the frequency distribution of the largest or smallest member of a sample*. In *Mathematical Proceedings of the Cambridge Philosophical Society*. 24(2), 180-190.

Gold, D., Damokosh A. I., Dockery D., et al. (1999). *Particulate and ozone pollutant effects on the respiratory function of children in Southwest Mexico City*. Epidemiology, 10(1), 8-16.

Gumbel, E. J. (2012). *Statistics of Extremes*. Dover Publications. (Original publicado en 1958).

## BIBLIOGRAFÍA

---

- Instituto de Planeación Democrática y Prospectiva, IPDP. (2024). *Panorama geográfico y estadístico de la Alcaldía Cuajimalpa de Morelos* [PDF]. Gobierno de la Ciudad de México, (1), 9.
- Lladser, M. (2011). *Variables aleatorias y simulación estocástica*. JC Sáez Editor, (2). (Original publicado en Santiago, Chile)
- Luna F. V., González S. H. (2007). *Teoría de valores extremos: Una introducción* [PDF]. Revista de Ciencias Básicas UJAT, 6(1), 10-16.
- Meneses, F., Romieu, I., Ruiz, S., et al (1996). *Effects of air pollution on the respiratory health of asthmatic children living in Mexico City*. American Journal of Respiratory and Critical Care Medicine, 154(2), 300-307.
- Molina A. G. (2010). *Teoría de valores de extremos aplicada a la gestión de riesgos en inundaciones*. Universidad de Sonora, México.
- Murray, R. y Spiegel, L. (2009). *Estadística*. McGraw Hill, (4).
- Organización Mundial de la Salud, OMS. (2021). *Directrices mundiales de la OMS sobre la calidad del aire: partículas (PM<sub>2,5</sub> y PM<sub>10</sub>), ozono, dióxido de nitrógeno, dióxido de azufre y monóxido de carbono*. Organización Mundial de la Salud.
- Ortega, S. J. (2009). *Elementos de Probabilidad y Estadística* [Notas de curso]. Centro de Investigación en Matemáticas (CIMAT). Recuperado de <https://www.cimat.mx/~jortega/MaterialDidactico/EPyE09/Libro1.pdf>
- Pacheco S. L. (2025). *Ajuste de la distribución de valores extremos generalizada a las precipitaciones pluviales máximas del Estado de Oaxaca* [Tesis de licenciatura]. Universidad Tecnológica de la Mixteca, Oaxaca, México.
- Perez, R. (2010). *Nociones Básicas de Estadística*. Departamento de Economía Aplicada. Universidad de Oviedo.
- Rincón, L. (2006). *Una introducción a la probabilidad y estadística*. UNAM, Facultad de Ciencias.
- Rincón, L. (2007). *Curso intermedio de probabilidad*. UNAM, Facultad de Ciencias.
- Romieu, I., Lugo, M. C., Velasco, S. R., et al. (1992). *Air pollution and school absenteeism among children in Mexico City*. American Journal of Epidemiology, 136(12), 1524-1531.

Sánchez M. J. M. (2009). *Compuestos orgánicos volátiles en el medio ambiente* [PDF]. Monografías de la Real Academia Nacional de Farmacia. Recuperado de <https://www.ritsq.org/wp-content/uploads/reach-uah/Sanchez-UAH-2008.pdf>

Secretaría de Salud México. (1996). *Programa para Mejorar la Calidad del Aire en el Valle de México*. Departamento del D.F. Gobierno del Estado de México Secretaría de Medio Ambiente Recursos Naturales y Pesca, (2), 46-53.

Secretaría del Medio Ambiente de la Ciudad de México, SEDEMA. (2023). *Calidad del aire en la Ciudad de México, Informe 2020*. Dirección General de Calidad del Aire, Dirección de Monitoreo de Calidad del Aire.

Secretaría del Medio Ambiente de la Ciudad de México, SEDEMA. (2023). *Calidad del aire en la Ciudad de México, Informe 2021*. Dirección General de Calidad del Aire, Dirección de Monitoreo de Calidad del Aire.

Secretaría del Medio Ambiente de la Ciudad de México, SEDEMA. (2024). *Calidad del aire en la Ciudad de México, Informe 2022*. Dirección General de Calidad del Aire, Dirección de Monitoreo de Calidad del Aire,

Secretaría del Medio Ambiente, SEDEMA. *Normas Oficiales Mexicanas establecidas por el gobierno de la CDMX*. Gobierno de la Ciudad de México. Recuperado de <http://www.aire.cdmx.gob.mx/default.php?opc=%27ZaBhnmI=%27&dc=Yw==>. Datos con derechos reservados.

Secretaría del Medio Ambiente, SEDEMA. *Consultas del Índice de Calidad del Aire de la Ciudad de México*. Gobierno de la Ciudad de México. Recuperado de <http://www.aire.cdmx.gob.mx/default.php?opc=%27aqBjnmU=%27>. Datos con derechos reservados.

Secretaría del Medio Ambiente de la Ciudad de México, SEDEMA. (2018). *GACETA OFICIAL DE LA CIUDAD DE MÉXICO* [PDF]. Gobierno de la Ciudad de México. 25-42. Recuperado de <http://www.aire.cdmx.gob.mx/descargas/monitoreo/normatividad/NADF-009-AIRE-2017.pdf>.

Secretaría del Medio Ambiente de la Ciudad de México, SEDEMA. *Datos del Índice de Calidad del Aire del SIMAT* [Conjunto de datos]. Gobierno de la Ciudad de México. Recuperado de [http://www.aire.cdmx.gob.mx/estadisticas-consultas/consultas/download\\_imeca.php](http://www.aire.cdmx.gob.mx/estadisticas-consultas/consultas/download_imeca.php). Datos con derechos reservados.

## BIBLIOGRAFÍA

---

Secretaría del Medio Ambiente de la Ciudad de México, SEDEMA. *Especificaciones de las bases de datos del Índice de Calidad del Aire por Zona* [PDF]. Gobierno de la Ciudad de México. Recuperado de <http://aire.cdmx.gob.mx/estadisticas-consultas/descargas/INDICExls-csv.pdf>.

Secretaría del Medio Ambiente de la Ciudad de México, SEDEMA. *Entornos de la estación de monitoreo CUA (Cuajimalpa)* [Conjunto de datos]. Gobierno de la Ciudad de México. Recuperado de [http://www.aire.cdmx.gob.mx/entornos/entorno\\_detalle.php?est=dIdx](http://www.aire.cdmx.gob.mx/entornos/entorno_detalle.php?est=dIdx). Datos con derechos reservados.

Secretaría del Medio Ambiente de la Ciudad de México, SEDEMA. *Entornos de la estación de monitoreo BJU (Benito Juárez)* [Conjunto de datos]. Gobierno de la Ciudad de México. Recuperado de [http://www.aire.cdmx.gob.mx/entornos/entorno\\_detalle.php?est=c3yF](http://www.aire.cdmx.gob.mx/entornos/entorno_detalle.php?est=c3yF). Datos con derechos reservados.

Secretaría del Medio Ambiente de la Ciudad de México, SEDEMA. *Entornos de la estación de monitoreo LPR (La Presa)* [Conjunto de datos]. Gobierno de la Ciudad de México. Recuperado de [http://www.aire.cdmx.gob.mx/entornos/entorno\\_detalle.php?est=fYKC](http://www.aire.cdmx.gob.mx/entornos/entorno_detalle.php?est=fYKC). Datos con derechos reservados.

Sousa, S. I. V., Alvim-Ferraz, M. C. M., Martins, F. G. (2013). *Health effects of ozone focusing on childhood asthma: what is now known—a review from an epidemiological point of view*. Chemosphere, 90(7), 2051-2058.

Universidad Autónoma Metropolitana Unidad Cuajimalpa, UAM-CUA. (2022). *Observatorio de Recursos Territoriales UAM Cuajimalpa*. Observatorio Territorial del Poniente, Ciudad de México.

Vallejo, M., Jáuregui-Renaud, K., Hermosillo, A. G., et al. (2003). *Efectos de la contaminación atmosférica en la salud y su importancia en la Ciudad de México* [PDF]. Gaceta Médica de México, 139(1), 57-63. Recuperado de [https://www.anmm.org.mx/bgmm/1864\\_2007/2003-139-1-57-63.pdf](https://www.anmm.org.mx/bgmm/1864_2007/2003-139-1-57-63.pdf)

Vilchis, J. E. (2006). *Modelación estadística de máximos por bloques* [Tesis de maestría]. Centro de Investigación en Matemáticas, CIMAT, Mexico.

# Apéndice A

## Código en el software R

### Código para la Estación CUA (Cuajimalpa)

El código para las estaciones BJU y LPR son análogos, salvo etiquetas en las gráficas.

```
1 # -----
2 # ANÁLISIS DE LA ESTACIÓN DE MONITOREO CUA (CUAJIMALPA)
3 # PARA LAS CONCENTRACIONES MÁXIMAS DIARIAS DE OZONO
4 # -----
5 library(readxl)
6 # Se cargan los datos de la estación CUA
7 datos<- read_excel("C:/Users/REALICE/Downloads/LMA TESIS/Ozono_CUA.xlsx")
8 x <- datos$CUA
9 tamx<-length(x) # Tamaño de 8760 datos
10 # Obtiene el vector de máximos
11 n <- x
12 B = 365 # Número de bloques
13 h = trunc(tamx/B) # Número de datos por bloque (h=24)
14 M <- numeric(length = B)
15 #Se llena el vector de máximos por día
16 for(j in 1:B){
17   a<-(j-1)*h+1
18   b<-j*h
19   aux<-n[a:b]
20   M[j]<-max(aux) # Vector de máximos de tamaño B
21 }
22 # Obtiene los máximos diarios clasificados por mes
23 bloque <- 1:365
```

## APÉNDICE A. CÓDIGO EN EL SOFTWARE R

---

```
24 Mr <- round(M, digits = 2) # Máximos redondeados a dos decimales
25 # Nombres de todos los meses
26 meses <- c("Enero", "Febrero", "Marzo", "Abril", "Mayo", "Junio",
27 "Julio", "Agosto", "Septiembre", "Octubre", "Noviembre", "Diciembre")
28 # Días por mes (año 2023)
29 dias_mes <- c(31, 28, 31, 30, 31, 30, 31, 31, 30, 31, 30, 31)
30 # Inicialización
31 k <- 1
32 dias <- list()
33 maxis <- list()
34 for (i in 1:length(meses)) {
35   # Calcular rango de días para el mes actual
36   fin <- k + dias_mes[i] - 1
37   # Asegurarnos de no pasarnos del día 365
38   fin <- min(fin, 365)
39   # Generar secuencia por mes
40   mes <- bloque[k:fin]
41   maxmes <- Mr[k:fin]
42   # Almacenar resultado
43   dias[[meses[i]]] <- mes
44   maxis[[meses[i]]] <- maxmes
45   # Mostrar resultados divididos en mitades (por quincena)
46   mitad1 <- bloque[k:(k+14)]
47   maxmitad1 <- Mr[k:(k+14)]
48   mitad2 <- bloque[(k+15):min(fin, k+30)]
49   maxmitad2 <- Mr[(k+15):min(fin, k+30)]
50   cat(meses[i], "\n")
51   cat("Bloques:", mitad1 , " \n", sep = " ")
52   cat("Máximos:", maxmitad1 , " \n", sep = " ")
53   cat("Bloques:", mitad2 , " \n", sep = " ")
54   cat("Máximos:", maxmitad2 , " \n\n", sep = " ")
55   # Actualizar contador
56   k <- fin + 1
57   # Si ya llegamos al día 365, salir del bucle
58   if (k > 365) break
59 }
60 # Para acceder a cualquier mes específico:
61 # dias$Enero maxis$Enero
```

```

62 # -----
63 # ESTADISTICAS DESCRIPTIVAS DE LOS MAXIMOS DIARIOS
64 library(moments)
65 summary(M)
66 asimetria <- skewness(M, na.rm = TRUE) # Asimetría
67 curtosis <- kurtosis(M, na.rm = TRUE) # Curtosis
68 cat("Coeficiente de asimetría:", round(asimetria, 4), "\n")
69 cat("Coeficiente de curtosis:", round(curtosis, 4), "\n")
70 # Hallar a que bloque le pertenece el dato mínimo
71 diamin <- which.min(M)
72 diamin
73 # -----
74 # DIAGRAMA DE DISPERSIÓN Y BOXPLOT
75 # Diagrama de dispersión
76 par(mfrow = c(1, 2)) # Para mostrar dos gráficos en una ventana
77 layout(matrix(c(1, 2), nrow = 1, ncol = 2), widths = c(6, 4))
78 plot(bloque, M, xlab = 'Número de bloques (días)',
79      ylab='Concentración máxima de ozono (ppb)',
80      main='Diagrama de dispersión', xlim = c(1, 366),
81      ylim = c(0, 165), col="blue", axes = F)
82 axis(1, at = seq(1, 365, 28), cex.axis = 1) # Para editar ejes
83 axis(2, at = seq(0, 165, 30), cex.axis = 1, las = 3)
84 box()
85 abline(h=75.26 , col="red", lwd=2) # Para la línea de media
86 abline(h=46 , col="orange", lwd=3, lty=2)
87 abline(h=104 , col="orange", lwd=3, lty=2)
88 # Boxplot
89 boxplot(M, xlab = "Estación de Cuajimalpa", col="lightgreen",
90         main = "Gráfico Box-plot", ylim = c(0, 165), axes = F)
91 axis(2, at = seq(1, 161, 40), cex.axis = 1, las = 3)
92 box()
93 abline(h=75.26, lty=2, col="darkred")
94 par(mfrow = c(1, 1)) # Volvemos al estado original
95 layout(1) # Restablecer layout
96 # -----
97 # AJUSTE DE LOS VALORES MÁXIMOS A UNA DVEG
98 # install.packages("evd") # Instalar si es necesario
99 # install.packages("ismev") # Instalar si es necesario

```

## APÉNDICE A. CÓDIGO EN EL SOFTWARE R

---

```
100 library(evd)
101 library(ismev)
102 # Ajustar DVEG por Máxima Verosimilitud
103 dismv <- gev.fit(M)
104 param <- dismv$ml
105 se <- dismv$se
106 # Extraer parámetros estimados:
107 mu1 <- param[1]      # Estimador mu
108 sigma1 <- param[2]  # Estimador sigma
109 xi1 <- param[3]     # Estimador xi
110 se_mu1 <- se[1]     # Error estándar de mu
111 se_sigma1 <- se[2]  # Error estándar de sigma
112 se_xi1 <- se[3]     # Error estándar de xi
113 # -----
114 # Método Nelder-Mead
115 z <- fgev(x=M, std.err = TRUE, corr=TRUE, method= "Nelder-Mead")
116 # Mostrar resultados
117 # summary(z)
118 estim <- z$estimate
119 estimse <- z$std.err
120 mcov <- z$var.cov
121 mcorr <- z$corr
122 estim # Estimadores
123 estimse # Errores estándar
124 mcov # Matriz de covarianza
125 mcorr # Matriz de correlación
126 # Se calculan los intervalos de confianza a un 95%
127 IC <- fgev(M)
128 confint(IC ,level=0.95)
129 # -----
130 # MÉTODOS GRÁFICOS DE DIAGNÓSTICO PARA DVEG
131 par(mfrow = c(1,2)) # Divide la ventana gráfica
132 orderM <- sort(M)  # Ordena los datos de menor a mayor
133 # Usar B           # Número de bloques
134 mu <- estim[1]     # Estimador de ubicación (mu)
135 sigma <- estim[2]  # Estimador de escala (sigma)
136 xi <- estim[3]     # Estimador de forma (xi)
137 # Gráfica de Probabilidad o Gráfica P-P
```

```

138 # Probabilidades empíricas (fórmula de Weibull)
139 empiricaP <- (1:B) / (B + 1)
140 # Probabilidades teóricas (FDA de GEV)
141 modeloP <- exp(-(1 + xi * ((orderM - mu)/sigma))-1/xi)
142 # Crear gráfico P-P
143 plot(x = empiricaP, y = modeloP, type = "p", pch = 8, col = "lightblue",
144      main = "Gráfico P-P (GEV)", xlab = "Probabilidades empíricas",
145      ylab = "Probabilidades teóricas (GEV)" )
146 abline(a = 0, b = 1, col = "red", lwd = 3) # Línea de referencia
147 # Gráfica de Cuantiles o Gráfica Q-Q
148 # Función del cuantil teórico GEV (inversa de la FDA)
149 quantile_gev <- function(p, mu, sigma, xi) {
150   if (abs(xi) < 1e-6) {
151     # Caso Gumbel (xi = 0)
152     mu - sigma * log(-log(p))
153   } else {
154     # Caso general GEV
155     mu + (sigma / xi) * ((-log(p))-xi - 1)
156   }
157 }
158 # Cuantiles teóricos
159 modeloQ <- sapply(empiricaP, quantile_gev, mu = mu, sigma = sigma, xi = xi)
160 # Crear gráfico Q-Q
161 plot(x = modeloQ, y = orderM, type = "p", pch = 8, col = "pink",
162      main = "Gráfico Q-Q (GEV)", xlab = "Cuantiles teóricos (GEV)",
163      ylab = "Cuantiles empíricos")
164 abline(a = 0, b = 1, col = "darkgreen", lwd = 3) # Línea de referencia
165 par(mfrow = c(1, 1)) # Resetea formato gráfico
166 # -----
167 # FUNCIÓN DE DISTRIBUCIÓN Y FUNCIÓN DE DENSIDAD (Weibull o Frechet)
168 par(mfrow = c(1, 2))
169 ozono <- seq(from = 0, to = 165, by = 0.1)
170 # 1. Función de Distribución Acumulada (FDA)
171 dis <- pgev(ozono, loc = mu, scale = sigma, shape = xi)
172 plot(x = ozono, y = dis,
173      main = "Función de Distribución (GEVD)",
174      xlab = "Concentración de Ozono (ppb) - Estación CUA",
175      ylab = "Probabilidad Acumulada F(x)",

```

## APÉNDICE A. CÓDIGO EN EL SOFTWARE R

---

```
176     type = "l", lwd = 3, col = "darkblue",
177     xlim = c(0, 165), ylim = c(0, 1.1),
178     xaxs = "i", yaxs = "i",      # desactiva el espacio adicional
179     xaxt = "n", yaxt = "n",      # Suprime el ejes automáticos
180   )
181 axis(1, at = seq(0, 165, by = 30))
182 axis(2, at = seq(0, 1, by = .1))
183 dis_95 <- round(1-dis[950], digits = 4) # Probab de superar 95 ppb
184 dis_154 <- round(1-dis[1540], digits = 4) # Probab de superar 154 ppb
185 lines(x=c(0,95), y=c(dis[950], dis[950]), col="orange", lty=2, lwd=2)
186 points(95, dis[950], pch = 19, col = "orange")
187 text(68, 0.78, paste(expression("P(X>95) ="),dis_95), col = "orange",cex = 1)
188 lines(x=c(0, 154), y=c(dis[1540], dis[1540]), col="red", lty=2, lwd=2)
189 points(154, dis[1540], pch = 19, col = "red")
190 text(134, 1.04, paste(expression("P(X>154) ="),dis_154), col = "red",cex = 1)
191 # 2. Función de Densidad (FDP)
192 den <- dgev(ozono, loc = mu, scale = sigma, shape = xi)
193 plot(x = ozono, y = den,
194     main = "Función de Densidad (GEVD)",
195     xlab = "Concentración de Ozono (ppb) - Estación CUA",
196     ylab = "Densidad f(x)",
197     type = "l", lwd = 3, col = "#D81B60",
198     xlim = c(0, 165), ylim = c(0, max(den) * 1.05),
199     xaxs = "i", yaxs = "i",
200     xaxt = "n",
201   )
202 axis(1, at = seq(0, 165, by = 30))
203 # Área bajo la curva
204 polygon(c(ozono[ozono >= 95], 154, 95),
205     c(den[ozono >= 95], 0, 0),
206     col = "pink", border = NA)
207 text(95, 0.0014, "Mala calidad\nde aire", col = "#D81B60", cex = 1)
208 polygon(c(ozono[ozono >= 154], 165, 154),
209     c(den[ozono >= 154], 0, 0),
210     col = "darkred", border = NA)
211 text(150, 0.002, "Muy Mala\ncalidad\nde aire", col = "black", cex = 1)
212 par(mfrow = c(1, 1)) # Restaurar parámetros gráficos
213 # -----
```

```

214 # DIAS QUE SUPERARON LOS UMBRALES CORRESPONDIENTES
215 u95 <- 0
216 for (i in 1:365){
217   if(i == 1){cat("Dias que superaron el umbral 95 ppb en 2023:\n")}
218   if(M[i] > 95){
219     cat("Bloque:", bloque[i],"ppb:", M[i],"\n")
220     u95 <- u95+1
221   }
222 }
223 cat("\nCantidad de días que excedieron 95 ppb:", u95,"\n")
224 cat("\nProbabilidad de exceder 95 ppb:", round(dis_95*100, 2),"%\n")
225 u154 <- 0
226 for (i in 1:365){
227   if(i == 1){cat("Dias que superaron el umbral 154 ppb en 2023:\n")}
228   if(M[i] > 154){
229     cat("\n Bloque:", bloque[i],"ppb:", M[i],"\n")
230     u154 <- u154+1
231   }
232 }
233 cat("\nCantidad de días que excedieron 154 ppb:", u154,"\n")
234 cat("\nProbabilidad de exceder 154 ppb:", round(dis_154*100, 2),"%\n")
235 # -----
236 # GRÁFICAS DE NIVELES DE RETORNO E HISTOGRAMA DE DENSIDAD AJUSTADA GEVD
237 # Gráfica de Niveles de Retorno
238 ps <- seq(from = 0.1, to = 0.99, by = 0.001) # Probab acumuladas
239 periodo <- log(-log(1 - ps)) # Transformación para eje x
240 niveles <- qgev(1 - ps, loc = mu, scale = sigma, shape = xi) # N. de retorno
241 par(mfrow = c(1, 2))
242 layout(matrix(c(1, 2), nrow = 1, ncol = 2), widths = c(4, 6)) # Proporción
243 plot(periodo, niveles, main = "Niveles de Retorno (GEVD)",
244       type = "l", col = "blue", lwd = 2, xaxs = "i", yaxs = "i",
245       xlab = expression(log(y[p])), ylab = expression(hat(z)[p]))
246 # Histograma y Densidad Ajustada
247 hist(M, freq = FALSE, breaks = 19, main = "Densidad Ajustada",
248      ylab="Densidad", xlab="Concentración de ozono (ppb): Estación CUA",
249      xlim= c(12,163), ylim=c(0,0.02), col = "lightgreen")
250 curve(dgev(x, loc = mu, scale = sigma, shape = xi), add=T, col="red", lwd=2)
251 box()

```

## APÉNDICE A. CÓDIGO EN EL SOFTWARE R

---

```
252 par(mfrow = c(1, 1)) # Restaurar parámetros gráficos
253 layout(1) # Restablecer layout
254 # -----
255 # NIVELES DE RETORNO y METODO DELTA GEVD (FRECHET O WEIBULL)
256 # Niveles de retorno
257 T_años <- c(1, 5, 10, 20)
258 T_dias <- T_años * 365
259 p <- 1-1/T_dias
260 niveles <- qgev(p, loc = mu, scale = sigma, shape = xi)
261 niveles
262 # Calculo de la varianza a traves del metodo Delta
263 Mcov <- z$var.cov
264 for (i in 1:4){
265     vector <- matrix(c(1, -xi^{-1}*(1-(-log(1-p[i])))^{-xi}),
266         sigma*xi^{-2}*(1-(-log(1-p[i])))^{-xi})-sigma*xi^{-1}*
267         (-log(1-p[i]))^{-xi}*log(-log(1-p[i]))), nrow = 1, ncol = 3)
268     vector_t <- t(vector)
269     varianza <- vector%%Mcov%%vector_t
270     # Intervalos de confianza del 95%
271     iclow <- niveles[i]-1.96*sqrt(varianza)
272     icup <- niveles[i]+1.96*sqrt(varianza)
273     cat("Intervalos de confianza del año:", T_años[i],"\n")
274     cat("Inf:", iclow,"Sup:", icup,"\n")
275     cat("Nivel de retorno:", niveles[i],"ppb \n")
276     cat("Varianza:", varianza,"\n\n")
277 }
278 # Gráfica de niveles de retorno
279 plot(T_años, niveles, log = "x", type = "b", col = "red", lwd=2,
280     xaxs = "i", yaxs = "i", # desactiva el espacio adicional
281     xlab = "Periodo de retorno (años)", ylab = " Concentración de ozono
282     ↪ (ppb)",
283     main = "Niveles de retorno (máximos diarios)")
284 grid()
285 # -----
286 # ROMÁN EMMANUEL HERNÁNDEZ CARRILLO
287 # LICENCIATURA EN MATEMÁTICAS APLICADAS
288 # UNIVERSIDAD TECNOLÓGICA DE LA MIXTECA
289 # -----
```