

Campus Monterrey

Materia

Inteligencia artificial avanzada para la ciencia de datos II

Actividad integradora I

Estudiante

Jacobo Hirsch Rodríguez - A00829679

Profesor

Blanca R. Ruiz Hernández

Índice

Introducción.....	3
Metodología.....	4
Selección de estado a estudiar y limpieza.....	4
Análisis de frecuencia método gráfico.....	6
Análisis de frecuencias método analítico.....	8
Tabla de resultados de los gráficos y su interpretación.....	10
Tabla de pruebas de bondad de ajuste y su interpretación.....	11
Problemática.....	12
Discusión y conclusiones.....	13
Referencias.....	14

Introducción

En hidrología, prever la magnitud y frecuencia de precipitaciones extremas es clave para diseñar obras hidráulicas resistentes y eficaces, como presas, puentes y sistemas de drenaje. Uno de los conceptos fundamentales para realizar estos diseños es el "periodo de retorno", que representa el intervalo promedio de años entre eventos de lluvia extrema de similar intensidad en una región. Para calcularlo, se parte del análisis de datos históricos de precipitación, obtenidos de instituciones como la Comisión Nacional del Agua (CONAGUA) en México. Esta información proporciona las bases para evaluar tendencias y patrones en la precipitación máxima anual y, así, generar modelos predictivos que ayudan a definir parámetros de seguridad en la infraestructura, para que los ingenieros puedan adecuar las obras de ingeniería a los requisitos que la región exige.

Una de las metodologías más empleadas en estos cálculos es el análisis de frecuencia, donde se ordenan los valores máximos anuales de precipitación, permitiendo calcular la "probabilidad de excedencia", es decir, la probabilidad de que una precipitación anual supere un determinado valor. Este análisis permite definir un periodo de retorno, como podría ser de 50 años, indicando que una precipitación similar o mayor puede ocurrir, en promedio, en el tiempo seleccionado. Como se mencionó anteriormente, Este cálculo permite a los ingenieros anticipar la intensidad de las lluvias y diseñar infraestructuras acordes a las características climáticas de la región.

Para reforzar estos pronósticos, se ajustan los datos a diferentes distribuciones de probabilidad, como la Normal, Log-normal, Exponencial, Gamma, Weibull y Gumbel. Estas distribuciones permiten modelar los datos históricos, proporcionando un marco para predecir la magnitud de futuros eventos extremos. Mediante pruebas de bondad de ajuste, como las pruebas de Kolmogorov-Smirnov y Shapiro-Wilk, se valida el modelo que mejor representa los datos, garantizando así la confiabilidad de los pronósticos.

Se presenta un análisis corto, mediante un análisis estadístico descriptivo y de frecuencia, busca identificar patrones en la precipitación máxima mensual en el estado de Colima, México.

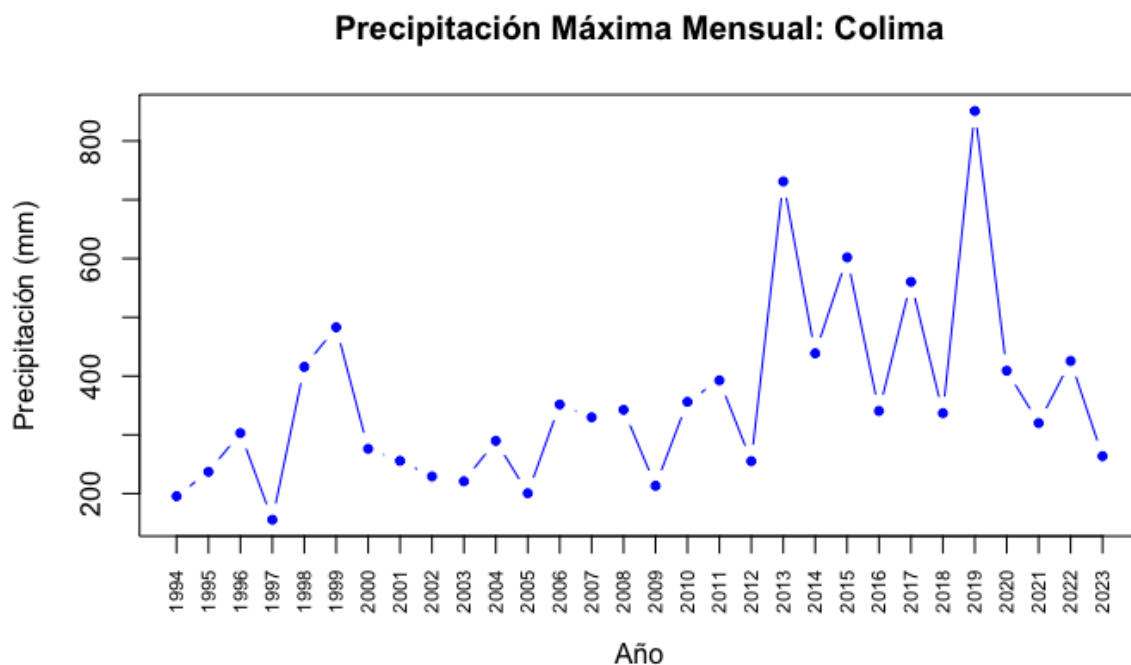
Metodología

Para el siguiente análisis estadístico se tomaron los datos proporcionados por el docente, que a su vez estos fueron obtenidos por la conagua, que es la comisión nacional del agua en México, institución responsable de la administración y preservación de los recursos hídricos en el país, así como de la infraestructura hidráulica y de los sistemas de distribución de agua.

Selección de estado a estudiar y limpieza

El dataset contenía datos sobre la precipitación máxima en mm de todos los estados de la república por mes, desde el año 1994 al año 2023. Para el siguiente análisis se seleccionó el estado de colima por lo que se filtró el dataset para que solo contuviera los datos de dicho estado.

Como se mencionó anteriormente, la necesidad de este análisis es poder predecir las precipitaciones máximas en un periodo de retorno esperado para alguna obra hidráulica, por lo que únicamente nos interesa saber cuales fueron las precipitaciones máximas en dichos años, por lo que se filtro el dataset para que únicamente estuvieran los años desde 1994 al 2023 mapeados con el máximo de precipitación registrado. Se obtuvo el siguiente gráfico de dichas operaciones.



Además se obtuvieron las siguientes medidas de centralización y variación de las precipitaciones máximas anuales en Colima.

Resumen de Medidas para Precipitación Máxima anual en Colima:

Medidas de Centralización:

Media: 359.4467

Mediana: 333.4

Moda: 155.5

Medidas de Dispersión o Variación:

Desviación Estándar: 158.3343

Varianza: 25069.74

Rango (diferencia entre máximo y mínimo): 695.5

Cuartiles:

- Primer Cuartil (Q1, 25%): 255.375

- Mediana (Q2, 50%): 333.4

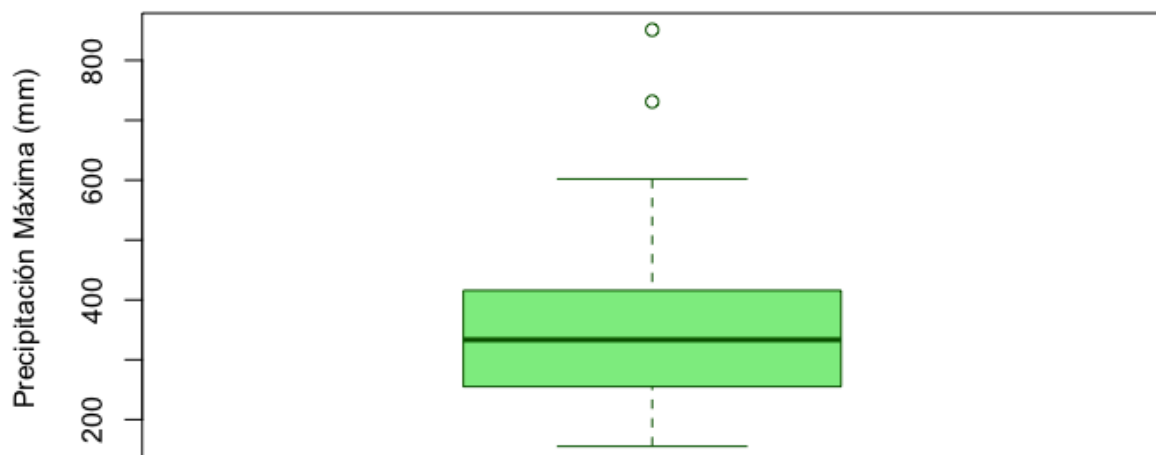
- Tercer Cuartil (Q3, 75%): 414.05

Percentil 90: 564.46

Los resultados muestran que, en promedio, se registran precipitaciones de 359.45 mm, con una mediana de 333.4 mm y una moda de 155.5 mm, lo que sugiere predominio de años con lluvias moderadas. La alta variabilidad se refleja en una desviación estándar de 158.33 mm y un rango de 695.5 mm, indicando fluctuaciones significativas de un año a otro. Los cuartiles (Q1 de 255.38 mm y Q3 de 414.05 mm) ubican la mayoría de los valores en un rango intermedio, aunque algunos años extremos se destacan. Solo el 10% de los años superan los 564.46 mm, evidenciando que las lluvias intensas son raras, pero posibles. Este análisis es clave para la planificación de infraestructura ante eventos climáticos extremos.

El siguiente boxplot ayudará a comprender las medidas estadísticas antes descritas:

Boxplot de Precipitación Máxima en Colima (1994-2023)



Sin embargo, eventos poco comunes con lluvias superiores a 600 mm reflejan una distribución sesgada hacia valores extremos. Esto destaca una variabilidad notable, dando a entender que dentro de este rango de tiempo hubo fenómenos meteorológicos extremos en el estado.

Análisis de frecuencia método gráfico

A continuación una descripción de los pasos que se ejecutaron en el análisis de frecuencia con los resultados de forma gráfica.

Ordenación de Precipitaciones Máximas: Los valores de precipitación máxima en Colima se organizaron en orden descendente, creando la columna precipitación máxima ordenada para destacar los eventos de lluvia más extremos. Este orden es crucial para calcular la frecuencia de estos eventos mediante la probabilidad de excedencia.

Asignación de Rank: A cada valor ordenado se le asignó un número de orden que posiciona cada registro desde el más alto al más bajo. Este ranking facilita el cálculo de la probabilidad de excedencia, esencial para medir la frecuencia de eventos de alta magnitud.

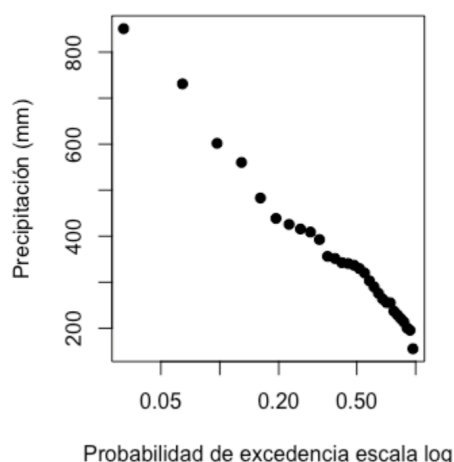
Cálculo de la Probabilidad de Excedencia: Utilizando el método de Weibull, la probabilidad de excedencia se calculó dividiendo el rango entre el total de observaciones más uno, almacenándose de forma separada en forma de columna dentro del dataframe. Este cálculo evalúa con qué frecuencia se podría registrar una precipitación de igual o mayor magnitud, siendo clave en el análisis de riesgos.

Cálculo de la Probabilidad de No Excedencia: Como complemento de la probabilidad de excedencia, la probabilidad de no excedencia se registró dentro del data frame para obtener los cálculos posteriores. indicando la frecuencia de eventos con precipitación igual o menor, proporcionando una perspectiva adicional de los eventos menos extremos.

Cálculo del Periodo de Retorno: El periodo de retorno, calculado como el inverso de la probabilidad de excedencia, indica el tiempo promedio esperado entre dos eventos de igual o mayor magnitud.

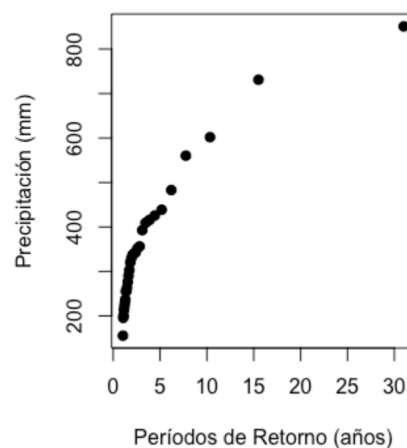
Visualización del Análisis de Frecuencia: Se realizaron dos gráficos para representar la relación entre precipitación máxima, probabilidad de excedencia y periodo de retorno:

Gráfico de Probabilidad de Excedencia: En escala logarítmica, muestra la frecuencia de eventos extremos.



Se observa una tendencia descendente, lo que indica que las precipitaciones máximas de mayor magnitud tienen una baja probabilidad de excedencia, es decir, ocurren con poca frecuencia. Por el contrario, las lluvias menores son más comunes, lo que se refleja en una mayor probabilidad de excedencia, algo esperado ya que en 29 años de registro la mayoría de los datos obtenidos parecían seguir la misma tendencia a excepción de dos valores atípicos.

Gráfico del Periodo de Retorno: Representa la magnitud de las lluvias en función del intervalo esperado entre eventos similares.



A medida que el periodo de retorno aumenta, también lo hace la magnitud de la precipitación máxima. Esto significa que los eventos de lluvia más intensos tienen periodos de retorno más largos, lo que indica que son menos frecuentes pero de mayor magnitud. Se muestra que hay eventos extremos, con precipitaciones superiores a 600 mm, que son poco frecuentes, presentando una probabilidad de excedencia muy baja (inferior a 0.05) y un periodo de retorno largo, de entre 15 a 30 años.

Análisis de frecuencias método analítico

Este método consiste en ajustar los datos a diferentes funciones de densidad de probabilidad con el objetivo de identificar cuál de estas distribuciones representa mejor el comportamiento de las precipitaciones y, así, poder modelar y pronosticar eventos futuros. A continuación, se detallan los pasos seguidos para este análisis:

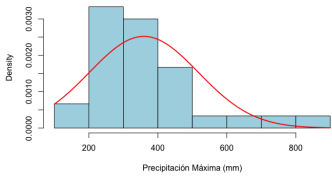
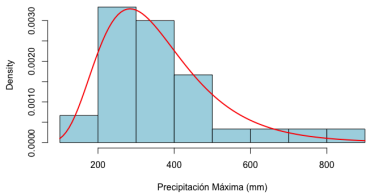
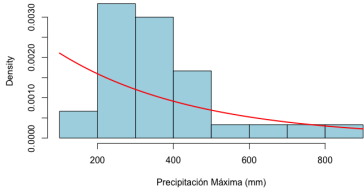
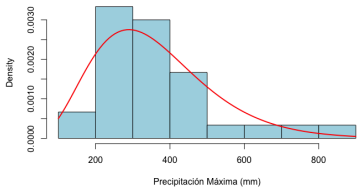
1. Ajuste a Diferentes Distribuciones de Probabilidad

Se evaluó el ajuste de los datos de precipitación máxima mensual a las siguientes distribuciones:

- Distribución Normal.
- Distribución Log-Normal.
- Distribución Exponencial.
- Distribución Gamma.
- Distribución Weibull.

- Distribución Gumbel.
2. Visualización de la Densidad Empírica y Teórica
Para cada una de las distribuciones seleccionadas, se construyó un histograma de la densidad empírica y se sobrepuso la curva de densidad teórica correspondiente con los parámetros estimados de los datos. Esto permitió evaluar, de manera visual, el grado de ajuste de cada distribución a los datos empíricos. Se analizó si la curva teórica se ajustaba adecuadamente a la forma del histograma.
 3. Gráficas Q-Q (Quantile-Quantile Plot)
Se construyeron Q-Q plots para cada distribución, los cuales comparan los cuantiles teóricos con los cuantiles empíricos de los datos. En un ajuste adecuado, los puntos deberían alinearse a lo largo de una línea recta, lo que indica una buena correspondencia entre los datos empíricos y la distribución teórica.
 4. Comparación de Distribuciones de Probabilidad Acumuladas (Ojivas)
Se compararon las distribuciones de probabilidad acumuladas empíricas y teóricas para cada una de las distribuciones evaluadas. Los datos empíricos representan las probabilidades acumuladas observadas, mientras que los datos teóricos provienen de la distribución ajustada. Una buena coincidencia entre ambas ojivas sugiere que la distribución teórica describe correctamente el comportamiento de los datos.
 5. Pruebas de Bondad de Ajuste
Para evaluar el ajuste de cada distribución, se aplicaron dos pruebas estadísticas de bondad de ajuste:
 - Prueba de Shapiro-Wilk: Para evaluar si los datos provienen de una distribución Normal.
 - Prueba Kolmogorov-Smirnov (KS): Para comparar la distribución empírica con la teórica y determinar si existen diferencias significativas.
 6. Para cada prueba, se estableció la hipótesis nula (H_0 : los datos provienen de la distribución propuesta) y se utilizó el p-value para decidir si se acepta o se rechaza dicha hipótesis. Un p-value alto indica que no hay suficiente evidencia para rechazar la hipótesis nula, sugiriendo un buen ajuste.
 7. Estimación de Parámetros de las Distribuciones
Se estimaron los parámetros correspondientes para cada distribución:
 - Distribución Normal: Media y desviación estándar.
 - Distribución Log-Normal: Parámetros derivados de la media y la desviación estándar en escala logarítmica.
 - Distribución Exponencial: Tasa de decaimiento.
 - Distribuciones Gamma, Weibull y Gumbel: Parámetros que caracterizan la forma y la escala.
 8. Los parámetros se calcularon utilizando el método de momentos o mediante el uso de la biblioteca fitdistrplus de R para obtener estimaciones automáticas.
 9. Interpretación y Comparación de Resultados
Luego de realizar los ajustes y evaluaciones, se compararon los resultados visuales y los resultados de las pruebas de bondad de ajuste para todas las distribuciones. Los resultados de todas la metodología anterior se presentan en formato de tabla.

Tabla de resultados de los gráficos y su interpretación

Distribución	Gráfico de densidad empírica vs teórica	Resultados del Q-Q plot	Resultados de la Comparación de Distribuciones de Probabilidad Acumuladas
Normal	<p>Histograma de Precipitación Máxima con Ajuste Normal</p> 	<p>La muestra se aproxima a una distribución normal en el centro, ya que la mayoría de los puntos se alinean con la línea de referencia. Sin embargo, las desviaciones en las colas sugieren valores extremos o una mayor dispersión, indicando que la muestra no sigue una distribución normal en los extremos.</p>	<p>Se observaron puntos que siguen de cerca la línea de la distribución teórica, indicando un buen ajuste en la mayor parte de los datos. Sin embargo, se identificaron desviaciones en los extremos de la precipitación, especialmente en los valores bajos y altos, sugiriendo diferencias entre la distribución empírica y la teórica en esas áreas.</p>
Log-normal	<p>Histograma de Precipitación Máxima con Ajuste Log-Normal</p> 	N.A	<p>Se observaron puntos que siguen de cerca la línea de la distribución teórica log-normal, lo cual indica un buen ajuste en la mayoría de los datos. Sin embargo, se identificaron pequeñas desviaciones en los valores más bajos y altos de precipitación, lo que sugiere ligeras diferencias entre la distribución empírica y la teórica en esos extremos.</p>
Exponencial	<p>Histograma de Precipitación Máxima con Ajuste Exponencial</p> 	N.A	<p>Se observaron desviaciones notables entre los datos empíricos y la distribución teórica exponencial, especialmente en valores intermedios y altos de precipitación, donde los datos superan la curva teórica. Esto sugiere que la distribución exponencial no captura adecuadamente la dispersión en estos valores.</p>
Gamma	<p>Histograma de Precipitación Máxima con Ajuste Gamma</p> 	N.A	<p>Se observó un buen ajuste entre los datos empíricos y la distribución teórica Gamma, con la mayoría de los puntos alineados a lo largo de la curva teórica. Sin embargo, hay ligeras desviaciones en los valores más bajos y altos de precipitación, aunque en general la distribución Gamma modela adecuadamente los datos.</p>

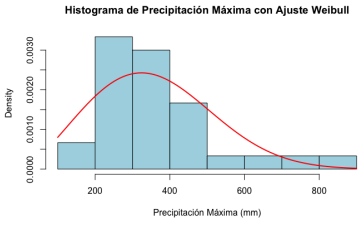
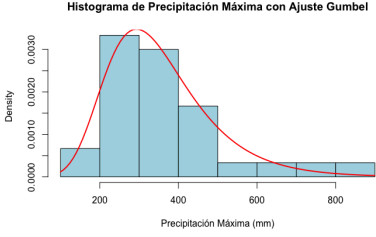
Weibull		N.A	Se observó un buen ajuste entre los datos empíricos y la distribución teórica Weibull, con la mayoría de los puntos alineados a lo largo de la curva. Sin embargo, existen pequeñas desviaciones en los valores bajos y medios de precipitación, indicando ligeras diferencias en estas áreas, aunque la distribución Weibull modela adecuadamente la mayor parte de los datos.
Gumbel		N.A	Se observó un buen ajuste entre los datos empíricos y la distribución teórica Gumbel, con la mayoría de los puntos alineados a la curva, sin embargo se notaban pequeñas desviaciones.

Tabla de pruebas de bondad de ajuste y su interpretación

Distribución	Resultados de la prueba KS	Interpretación de los resultados	Se rechazó la hipótesis nula
Normal	valor D=0.17485 p-value=0.2837	No hay evidencia suficiente para concluir que los datos se desvían significativamente de la distribución teórica dado el bajo valor d y alto p-value	No
Log-normal	valor D = 0.095859 p-value = 0.9211	El valor d indica una pequeña distancia entre la distribución empírica de los datos y la distribución teórica esperada. muy superior a 0.05, no se rechaza la hipótesis nula, lo que sugiere que no hay evidencia estadística para	No
Exponencia	valor D=0.38634 p-value = 0.0001534	El valor del indica una mayor distancia entre la distribución empírica de los datos y la distribución teórica esperada. El valor pr nos indica que se rechaza la hipótesis nula.	Si

Gamma	valor D = 0.11614 p-value= 0.7707	por el valor d ,indican una baja distancia entre la distribución empírica de los datos y la distribución teórica esperada. y el p-value mayor al umbral 0.05, indica que no se rechaza la hipótesis nula.	No
Weibull	valor D = 0.14627 p-value = 0.4966	por el valor de indicando una distancia moderada entre la distribución empírica de los datos y la distribución teórica esperada. Con un p-value mayor a 0.05, no se rechaza la hipótesis nula.	No
Gumbel	valor D=0.090992 p-value = 0.9458	por el valor indica una pequeña distancia entre la distribución empírica de los datos y la distribución teórica esperada. Con un valor p de 0.9458 muy superior a 0.05, no se rechaza la hipótesis nula.	No

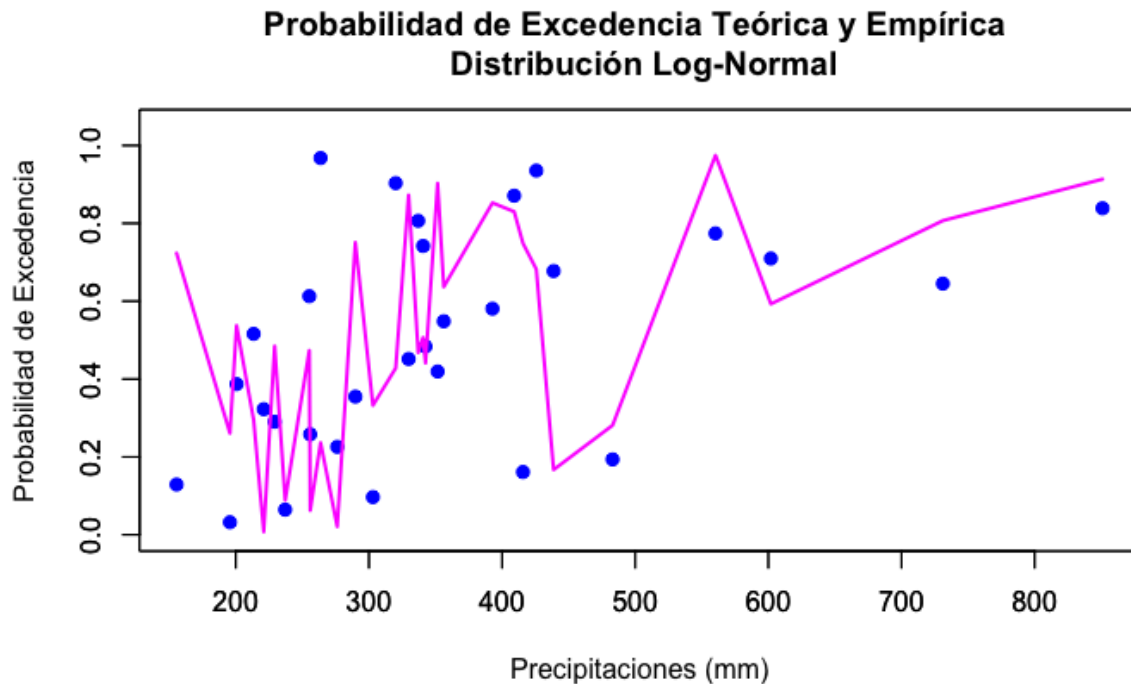
Problemática

Para diseñar una presa derivadora para una zona de riego mediana el periodo de retorno es de 100 a 500 años para una superficie de entre 1,000 y 10,000 hectáreas.

como la distribución log-normal fue la más representativa para los datos empíricos, se utilizó esta distribución para la predicción. Para ajustar una distribución Log-Normal a los datos históricos de precipitación máxima mensual en Colima. El ajuste se realizó mediante máxima verosimilitud para estimar los parámetros de la distribución (meanlog y sdlog), representando el promedio y la desviación estándar de los logaritmos naturales de los datos.

Con los parámetros obtenidos, se calculó la probabilidad de excedencia teórica para cada valor de precipitación usando la distribución Log-Normal. Esta probabilidad representa la probabilidad de que una precipitación alcance o exceda un valor dado. Se comparó esta probabilidad teórica con la probabilidad de excedencia empírica en un gráfico, observando la concordancia entre ambos para evaluar la idoneidad del ajuste.

Finalmente, se calculó la precipitación de diseño para un periodo de retorno de 200 años, se obtuvo como resultado que la precipitación máxima anual para dicho periodo de retorno es de 902.04 mm. Además, se obtuvo el siguiente gráfico donde se observa como los puntos empíricos tratan de seguir a la distribución teórica, siendo útil en el futuro para la predicción de lluvias.



Discusión y conclusiones

Este tipo de estudio resulta esencial en ingeniería hidráulica, especialmente en el diseño de infraestructuras críticas como presas derivadoras. Para este análisis, se emplearon datos de precipitaciones máximas en Colima, México, con el propósito de identificar la distribución que mejor modele la frecuencia y magnitud de eventos de lluvia extrema, así como calcular el periodo de retorno necesario para el diseño de obras hidráulicas.

En el caso de la distribución Log-Normal, si bien la curva teórica parecía ajustarse al histograma, la comparación entre las probabilidades de excedencia empírica y teórica mostró discrepancias significativas. Por otro lado, el ajuste con la distribución Gumbel, mostró una representación de los datos igual de buena frente a la Log-Normal. Sin embargo, los resultados de las pruebas de bondad de ajuste no fueron concluyentes en algunos casos, lo que indica que, aunque ninguna distribución obtuvo un ajuste perfecto, algunas se aproximaron más que otras.

El análisis también del periodo de retorno, que representa el intervalo promedio entre eventos de magnitud igual o superior a un umbral de precipitación, que es esencial en la planificación de infraestructuras hidráulicas, ya que permite estimar la frecuencia de eventos extremos. Para el diseño de una presa derivadora en una zona de riego mediana en Colima, se empleó la distribución de mejor ajuste para calcular la precipitación de diseño con un periodo de retorno de 200 años, estimada en 902.45 mm. Esto significa que, en promedio, un evento de esta magnitud o mayor podría esperarse cada 200 años.

Es importante resaltar que tanto el periodo de retorno como la probabilidad de excedencia son herramientas clave para evaluar el riesgo y la seguridad en obras hidráulicas. Aumentar el periodo de retorno reduce la probabilidad de falla ante un evento extremo, aunque también incrementa el tamaño y costo de la obra, haciendo esencial encontrar un balance entre seguridad y viabilidad económica. En este análisis, la información derivada de las diferentes distribuciones proporciona una base cuantitativa para decisiones de diseño en la presa derivadora de Colima. En conclusión, el ajuste de distribuciones de probabilidad para modelar eventos extremos de precipitación es crucial para diseñar infraestructuras hidráulicas seguras y eficientes.

Referencias

Universidad Estatal de San Diego. (n.d.). *Periodos de retorno recomendados para obras hidráulicas*. Recuperado el 29 de octubre de 2024, de https://pon.sdsu.edu/periodos_de_retorno_cna.html

Descripción: El recurso de la Universidad Estatal de San Diego sobre “Periodos de retorno recomendados para obras hidráulicas” ofrece una guía sobre cómo calcular cuánto podrían ocurrir lluvias extremas y cómo esto ayuda a diseñar estructuras como presas y drenajes.

Comisión Nacional del Agua (CONAGUA), Servicio Meteorológico Nacional (SMN). (n.d.). *Servicio Meteorológico Nacional*. Recuperado el 29 de octubre de 2024, de <https://smn.conagua.gob.mx/es/>

Descripción: El SMN proporciona datos meteorológicos actualizados y resúmenes históricos sobre clima, incluyendo información sobre lluvias, temperaturas, y fenómenos naturales en México.

Hong Kong Observatory. (n.d.). *Return period: "Once in N years"*. Hong Kong Observatory. Recuperado el 29 de octubre de 2024,, de <https://www.hko.gov.hk/en/education/climate/climate-change/00672-Return-Period-Once-in-N-Years.html>

Descripción: Explica el concepto de periodo de retorno, un término utilizado en meteorología y estudios climáticos