2. Conceptos básicos.

2.1. Conceptos fundamentales de estadística.

Con la finalidad, de facilitar la comprensión por parte del lector del desarrollo posterior, se enuncian las siguientes definiciones:

Variable: Es la asignación de un número a cada punto del espacio muestral o universo. En general una variable x tiene algún significado físico, geométrico o cuantificable.
Variable discreta: si su conjunto de valores está acotado, es numerable, y sí éstos pueden arreglarse en una secuencia que corresponde a los enteros positivos.
Variable continua: cuando los posibles valores se encuentran en el intervalo de los números reales, sin restricción alguna.
Resultado: es el producto de observación de un fenómeno del cual se arroja un dato exclusivamente, en general, implica un valor a la variable X.
Población: Es un conjunto de toda la posible información que caracteriza un fenómeno.
Muestra: Es un subconjunto de resultados obtenidos para la variable x en cuestión, seleccionados de una población.

Para caracterizar estadísticamente una muestra se utilizan parámetros numéricos y representaciones gráficas. En general, es conveniente que dicha caracterización comprenda tanto representaciones gráficas como numéricas.

Los principales parámetros numéricos para caracterizar una muestra son:

2.1.1. Medidas de tendencia central.

Media: Siendo $X_1, X_2, \dots, X_n$, valores medidos de una variable aleatoria. Entonces la media de la muestra, comúnmente conocida como promedio, se define:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_n \qquad (2.1)$$

2.1.2. Medidas de dispersión.

Varianza: segundo momento con respecto a la media. Se representa como S^2:

$$S^2 = \frac{(X_1-\bar{X})^2+(X_2-\bar{X})^2+\cdots+(X_n-\bar{X})^2}{n} \qquad (2.2)$$

Desviación estándar: en términos prácticos es igual a la raíz cuadrada de la varianza.

$$S = \pm\sqrt{S^2} \qquad (2.3)$$

Coeficiente de variación: es un parámetro que nos indica que tan dispersos están los datos con respecto a la media propia de una muestra, y se calcula de la siguiente manera:

$$CV = \frac{S}{\bar{X}} \qquad (2.4)$$

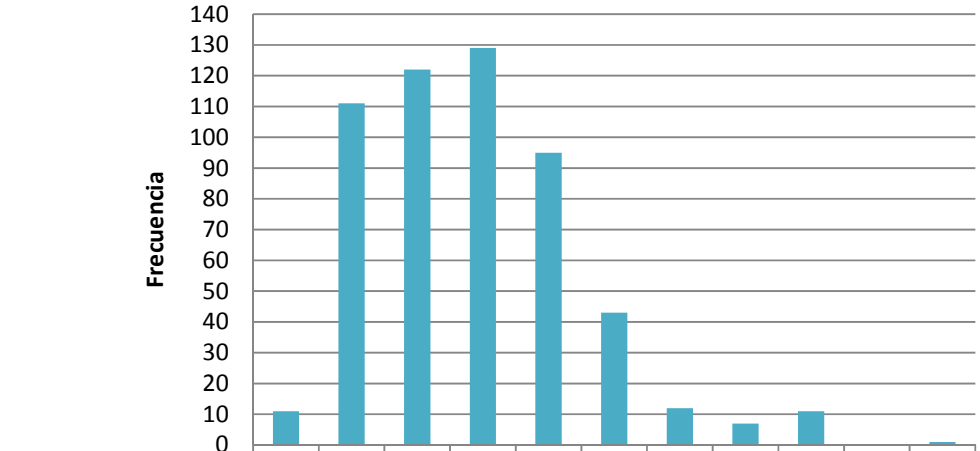
2.1.3. Representación gráfica.

Es frecuente que para observar el comportamiento o la tendencia de una muestra hacia ciertos valores, se utilicen gráficas. Para ello, es necesario organizar los datos de una muestra en intervalos.

Posteriormente, se ubica el valor máximo x_M y mínimo x_m y el tamaño de la diferencia Δx_T entre éstos. Se obtiene el tamaño de los intervalos Δx o número de intervalos i, de tal forma que $i = \frac{\Delta x_T}{\Delta x}$ oscile entre los valores 5 y 20, según el tamaño de la muestra y la precisión requerida. Si el número de datos es relativamente pequeño, el número de clases será cercano a 5, pero en general nunca menor a este valor. Si la cantidad de datos es grande, se espera que sean hasta 20 clases, de lo contrario se puede ocultar la distribución real de un conjunto determinado de datos, pero al mismo tiempo se busca que todos los intervalos tengan una representación. Para definir los intervalos o clases, se utilizará la siguiente secuencia:

$$\boldsymbol{Clases\ o\ intervalos}\begin{cases}\boldsymbol{1er.Intervalo}; & \boldsymbol{x_m \le x < x_m + \Delta x}\\ \boldsymbol{2do.Intervalo}; & \boldsymbol{x_m + \Delta x \le x < x_m + 2\cdot\Delta x}\\ \vdots & \vdots\\ \boldsymbol{i-}\text{é}\boldsymbol{simo\ Int.}; & \boldsymbol{x_m + (i-1)\Delta x \le x < x_m + i\cdot\Delta x}\\ \vdots & \vdots\\ \text{Ú}\boldsymbol{ltimo\ Int.}\ ; & \boldsymbol{x_m + (l-1)\Delta x \le x < x_m}\end{cases} \qquad (2.5)$$

Se obtiene la frecuencia absoluta f_i, que corresponde al número de datos asignados a cada intervalo i. La gráfica que relaciona los intervalos y los valores de frecuencia relativa se conoce como “histograma de frecuencias”.



Intervalo	14 - 40	40 - 66	66 - 92	92 - 118	118 - 144	144 - 170	170 - 196	196 - 222	222 - 248	248 - 274	274 - 300
Frecuencia Absoluta	11	111	122	129	95	43	12	7	11	0	1

Figura 2.1.3.1.Histograma de Frecuencia de registros máximos de Nayarit.

Posteriormente, se obtiene la frecuencia relativa fr_i, que se obtiene de dividir cada valor de frecuencia absoluta f_i de cada intervalo i entre el número total de datos. Esto produce que la suma de las frecuencias relativas sea igual a 1.

Finalmente, sumando los valores de cada una de las frecuencias relativas en cada intervalo i, se obtiene la curva de frecuencias relativas acumuladas. La gráfica resultante se denomina “curva de frecuencias acumuladas”. En términos prácticos, se obtiene uniendo los puntos medios del histograma de frecuencias relativas acumuladas.

Las frecuencias relativas de una variable, resultan una buena aproximación a la probabilidad de que la variable mencionada se encuentre en los intervalos formados.

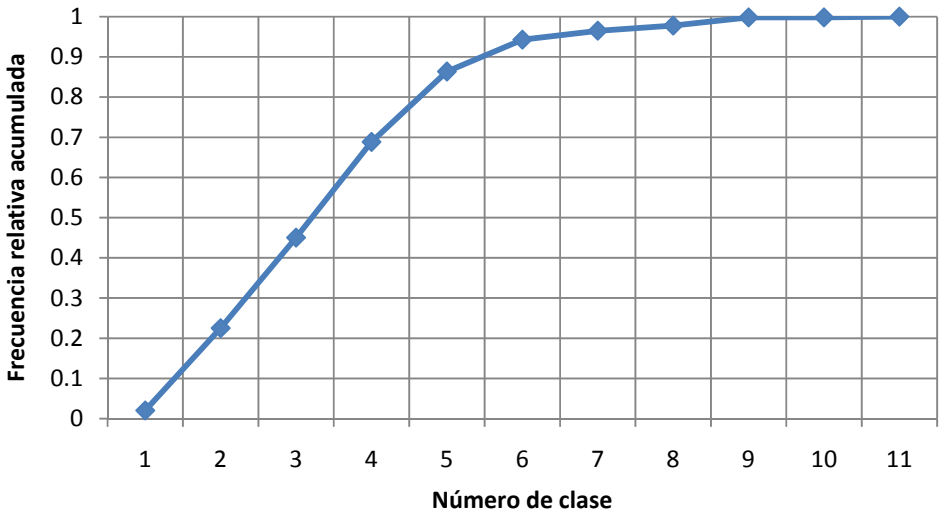


Figura 2.1.3.2.Curva de frecuencia relativa acumulada del Estado de Nayarit

En términos estadísticos, con la metodología de regionalización que se explica en el capítulo III, se pretende encontrar una población representativa de varias muestras, a diferencia de la forma tradicional en la que se analiza cada muestra por separado.

2.2. Conceptos fundamentales de probabilidad

2.2.1. Principales axiomas de probabilidad

Cuando el espacio muestral es continuo, los sucesos corresponden a subconjuntos. Es decir, a cada suceso A del espacio, se asocia un numero real P(A). En otras palabras, P, es una función del valor real definida en el espacio. Así P, se denomina función de probabilidad, y P(A) es la probabilidad del suceso A, si se cumplen los siguientes axiomas.

1. La probabilidad toma valores entre 0 y 1.

$$0 < P < 1 \qquad (2.6)$$

2. La suma de probabilidades de todos los eventos que forman el espacio muestral es igual a 1.

$$P(A) + P(B) + \ldots + P(n) = 1 \qquad (2.7)$$

3. La probabilidad de la unión de dos eventos que son mutuamente excluyentes es igual a la suma de las probabilidades.

$$P(A \cup B) = P(A) + P(B) \qquad (2.8)$$

4. La probabilidad de la unión de dos eventos que no son mutuamente excluyentes es igual a la suma de las probabilidades de los eventos menos la probabilidad de que ocurran ambos.

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \qquad (2.9)$$

O bien para más de dos eventos:

$$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C) \qquad (2.10)$$

5. La probabilidad de que no ocurra un evento A es igual a 1 menos la probabilidad de que ocurra el evento A.

$$P(\bar{A}) = 1 - P(A) \qquad (2.11)$$

2.3. Distribuciones de probabilidad de variables aleatorias continuas.

La distribución de probabilidad de una variable aleatoria continua X está caracterizada por una función $f(x)$ que recibe el nombre de función de densidad de probabilidad. La probabilidad de que X tome el valor específico x es cero, esto es, la función de densidad de probabilidad no representa la probabilidad de que $X = x$, es decir, tenga un valor específico. Más bien, ésta proporciona un medio para determinar la probabilidad de un intervalo $a \leq X \leq b$. Gráficamente, esta función se representa por el polígono de frecuencia relacionado con la idea del histograma previamente analizado en la sección 2.1.3. El polígono de frecuencia, se puede representar uniendo los puntos medios del histograma de frecuencias relativas.

Una función de probabilidad debe cumplir con las siguientes propiedades:

1. $$f(x) \geq 0 \qquad (2.12)$$

2. $$\int_{-\infty}^{\infty} f(x)dx = 1 \qquad (2.13)$$

La segunda propiedad, es una proposición del hecho de que una variable aleatoria continua de valor real debe encontrarse entre ∞ e $-\infty$. Por lo anterior, se da origen a la tercera propiedad, en la cual la probabilidad de que la variable X se encuentre en el intervalo (a, b).

3. $$P(a < X < b) = \int_a^b f(x)dx \qquad (2.14)$$

Es posible demostrar que la expresión anterior cumple los axiomas de probabilidad enunciados con anterioridad.

2.3.1. Obtención de Función de Probabilidad.

Con las ideas desarrolladas en el inciso 2.1.3 para variables discretas, se establece la función que representa el "polígono de frecuencias" definida por:

$$f_s(x) = \frac{n_i}{n_T} \qquad (2.15)$$

Esta ecuación, arroja una probabilidad, asociada a la ocurrencia de un evento determinado. Al realizar estimaciones probabilísticas en relación con alguna variable hidrológica, es necesario primeramente, caracterizar la población. Por esto último, el procedimiento subsecuente, se centra en la caracterización de la misma.

La integral de los valores de las frecuencias relativas hasta determinado punto es igual a la función de frecuencia acumulada $F(x)$. La gráfica de esta, guarda una relación con la "curva de frecuencias acumuladas" previamente mencionada en 2.1.3. Se puede definir la función de distribución acumulada $F(x)$ para una variable aleatoria continua como la función que determina la probabilidad de que la variable aleatoria X tome valores menores o iguales que un valor específico x:

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f_s(x)\, dx \qquad (2.16)$$

De acuerdo con su definición, una función de distribución de probabilidad acumulada es siempre mayor o igual que cero y no decreciente.

Por conceptos del cálculo integral, debe cumplirse que

$$f_s(x) = \frac{dF(x)}{dx} \qquad (2.17)$$

Además, según la tercera propiedad resulta:

$$P(a < X < b) = P(X \leq b) - P(X \leq a) = \int_{-\infty}^{b} f_s(x)dx - \int_{a}^{\infty} f_s(x)dx = F(b) - F(a) \qquad (2.18)$$

Finalmente, según el primer axioma de probabilidad.

$$0 < F(x) < 1 \qquad (2.19)$$

2.4. Métodos para calcular los parámetros de las funciones de probabilidad.

Tras haber seleccionado el tipo de función de probabilidad, deben encontrarse los parámetros de ésta. Esto, con el fin de aproximar la función de distribución a la curva de frecuencias acumuladas, es decir, que la función de distribución resulte representativa de los datos. En la práctica, se acostumbra utilizar los métodos gráficos, de momentos, máxima verosimilitud y mínimos cuadrados que se enuncian a continuación.

2.4.1. Gráficos

Un primer método para verificar la representatividad de la función de probabilidad con los datos medidos es mediante una gráfica donde se hayan dibujado cada una de las diferentes funciones junto con los puntos correspondientes a los registros hidrológicos. La función de distribución de probabilidad que se seleccione será aquella que se apegue mejor visualmente a los datos medidos.

2.4.2. Momentos

El método consiste en igualar los valores de los parámetros estadísticos de la muestra, con los de la población. Es decir, que se planteará un sistema de ecuaciones, cuyo tamaño depende del número de parámetros por estimar. El desarrollo se puede observar a detalle en la referencia[1].

2.4.3. Máxima Verosimilitud

Tras formular una función de densidad de probabilidad $f(x, a_1, a_2, \dots, a_m)$ para x con los parámetros desconocidos $a_i; i = 1, \dots, m$. Debido a que existe una muestra aleatoria $x_1, x_2, \dots, x_n$ se obtiene una función de densidad conjunta $f(x_1, x_2, \dots, x_n; a_1, a_2, \dots, a_m)$. La función de densidad conjunta se puede escribir como:

$$f(x_1, x_2, \dots, x_n; a_1, a_2, \dots, a_m) = \prod_{i=1}^{n} f(x_i, a_1, a_2, \dots, a_m) \qquad (2.20)$$

Por otro lado, la probabilidad de obtener un valor aleatorio de la población X, es proporcional al producto de sus densidades de probabilidad individual. Esta función conjunta se conoce como función de verosimilitud L.

$$L = \prod_{i=1}^{n} f(x_i, a_1, a_2, \dots, a_m) \qquad (2.21)$$

El método estima los parámetros al maximizar la función L, esto es, maximizando la probabilidad de que la muestra en estudio sea el resultado de obtener n números aleatorios a partir de la función de densidad de probabilidad. Los valores de los parámetros así obtenidos son conocidos como los estimadores por máxima verosimilitud.

[1] Técnicas Estadísticas en Hidrología. p. 26

2.4.4. Mínimos cuadrados

El método obtiene los distintos parámetros al minimizar la suma de los cuadrados de todas las desviaciones entre los datos y los valores calculados. Matemáticamente, esta suma se puede expresar como:

$$S = \sum_{i=1}^{n} [P(x_o)_i - F(x_c)_i]^2 \qquad (2.22)$$

Donde

x_o, son los datos de muestra.

x_c, son los valores que dependen de los parámetros $a_i; i = 1, \ldots, m$.

$F(x_c)$, es la función de distribución de probabilidad buscada para $(x_c)_i$

$P(x_o)$, es la probabilidad de que un dato cualquiera sea menor que x_0, y normalmente se estima con la fórmula de Weibull:

$$P(x_o) = \frac{n - m + 1}{n + 1} \qquad (2.23)$$

Donde

n, es el número de datos de registro.

m, número de orden que ocupa x_o en la serie de datos ordenados de mayor a menor.

2.5. Funciones de distribución de probabilidad recurrentes en la hidrología.

En la actualidad, se maneja una amplia cantidad de funciones de distribución de probabilidad. Sin embargo, dado la asimetría (frecuentemente positiva) de los datos hidrológicos, sólo algunas funciones representan de manera cercana este tipo de información.

Las funciones más frecuentes en hidrología son:

- Distribución exponencial.
- Distribución Normal.
- Distribución Lognormal.
- Distribución Gamma.
- Distribución Pearson tipo III(Gamma de tres parámetros)
- Distribución Log-Pearson tipo III.
- Distribución General de Valores Extremos tipo I (Gumbel)
- Distribución Gumbel de dos poblaciones (Doble Gumbel)

Para los fines de este trabajo se tomaron en cuenta valores extremos, es decir, resultados máximos o mínimos en un conjunto de registros. Es por ello, que basado en la experiencia y la literatura[2] se aplicaron las distribuciones de probabilidad General de Valores Extremos Tipo I (o Gumbel) y Doble Gumbel(o Gumbel de dos poblaciones). Sin embargo, para fines de comparación en los ajustes, se mostrará el desarrollo de la función log normal, que se presenta de manera recurrente en el análisis de las diversas regiones del país.

2.5.1. Distribución Lognormal

La función de Distribución Lognormal es comúnmente aplicada para variables aleatorias que cubren todo el rango de resultados positivos del fenómeno en estudio. La distribución Lognormal tiene la ventaja de trabajar únicamente para valores positivos $(X > 0)$ y la función logaritmo natural, reduce considerablemente la asimetría positiva, propia de la información hidrológica.

La Función de densidad de probabilidad Lognormal se expresa de la siguiente manera:

$$f_s(x) = \frac{1}{x\beta\sqrt{2\pi}} e^{\left(-\frac{(ln-\alpha)^2}{2\beta^2}\right)} \qquad (2.24)$$

Donde

α, es un parámetro de ubicación propio de la distribución.

β, es un parámetro de escala

Es posible, deducir que α y β , respectivamente, son la media y desviación estándar de los logaritmos de la variable aleatoria.

Con la ecuación (2.16) se obtiene la función de distribución de probabilidad:

$$F(x) = \int_0^x \frac{1}{x\beta\sqrt{2\pi}} e^{\left(-\frac{(ln-\alpha)^2}{2\beta^2}\right)} dx \qquad (2.25)$$

2.5.2. Distribución Gumbel

Dada la naturaleza aleatoria de los registros hidrológicos, estos tienen un comportamiento asimétrico positivo, por ello se apegan muy bien a la teoría desarrollada por Gumbel para valores extremos donde se ha demostrado su buen comportamiento con series de registros máximos anuales[3].

Se tienen N muestras, cada una de las cuales contiene n eventos. Si se selecciona el máximo x_i ; i = 1,2,3,….N de los n eventos de cada muestra, es posible demostrar que, a medida que N y n aumentan, la función de distribución de probabilidad de x tiende a:

$$F(x) = e^{-e^{-\alpha(x-\beta)}} \qquad (2.26)$$

[2] Hidrología Aplicada, p. 396. Distribuciones de valores extremos.
[3] Flood Frequency Analysis. Ramachandra P. 229

Donde

α, resulta el parámetro de forma de la curva.

β, es el parámetro de escala.

x, es la variable aleatoria.

La función de densidad de probabilidad es entonces:

$$f(x) = \alpha e^{-[-\alpha(x-\beta)-e^{-\alpha(x-\beta)}]} \qquad (2.27)$$

Ambos parámetros se estiman por el método de momentos de la siguiente manera:

$$\alpha = \frac{1.2825}{S} \qquad (2.28)$$

$$\beta = \bar{x} - 0.45S \qquad (2.29)$$

Bajo el método de máxima verosimilitud, se resuelven las expresiones:

$$\sum_{i=1}^{n} x_i e^{-\alpha x_i} - \left(\bar{x} - \frac{1}{\alpha}\right)\sum_{i=1}^{n} e^{-\alpha x_i} = 0 \qquad (2.30)$$

$$\beta = \frac{1}{\alpha}\ln\frac{n}{\sum_{i=1}^{n} e^{-\alpha x_i}} \qquad (2.31)$$

La Figura 2.5.2.1 muestra la gráfica típica de la función de distribución Gumbel.

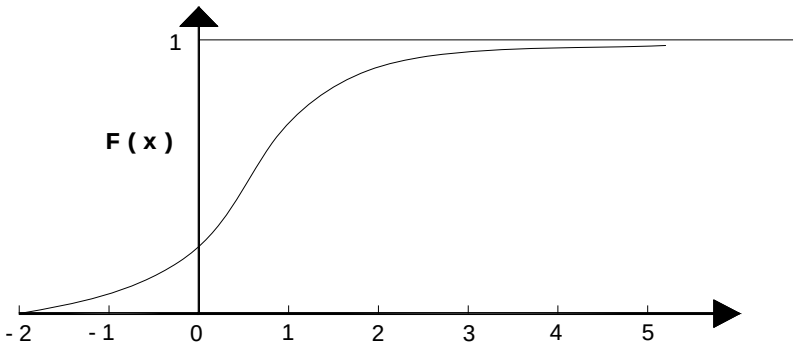


Figura 2.5.2.1. Gráfica típica de la distribución Gumbel.

Por otro lado, debido a la escala de doble logaritmo que maneja la función $F(x)$, para manejar una misma escala aritmética en las gráficas, permitiendo comparaciones con los datos y otras gráficas de distribución de probabilidad, es necesario llevar a cabo el siguiente procedimiento:

Se aplica la función logaritmo natural una vez:

$$\ln F(x) = \ln e^{-e^{-\alpha(x-\beta)}} \quad (2.32)$$

Obteniendo

$$\ln F(x) = -e^{-\alpha(x-\beta)} \quad (2.33)$$

Por propiedades logarítmicas, se puede expresar de la siguiente forma

$$\ln \frac{1}{F(x)} = e^{-\alpha(x-\beta)} \quad (2.34)$$

Se aplica nuevamente la función logaritmo en ambos extremos de la ecuación:

$$\ln \ln \frac{1}{F(x)} = \ln e^{-\alpha(x-\beta)} \quad (2.35)$$

Y, finalmente, se obtiene una expresión que relaciona linealmente x con $\ln \ln \frac{1}{F(x)}$:

$$-\ln \ln \frac{1}{F(x)} = \alpha(x-\beta) \quad (2.36)$$

En la Figura 2.5.2.2 se muestra una gráfica obtenida en una escala aritmética, con el procedimiento previamente descrito.

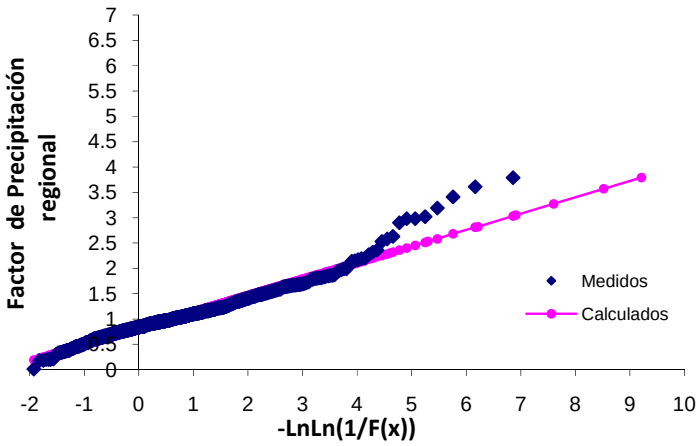


Figura 2.5.2.2. Gráfica Gumbel de la región Centro de Guerrero.

2.5.3. Distribución Doble Gumbel

En muchos lugares del mundo, los registros climatológicos, en especial los relacionados a lluvia han demostrado una clara división en dos poblaciones. La primera población corresponde a las lluvias normales de verano relacionadas con fenómenos meteorológicos propios del lugar; la segunda población corresponde fenómenos ciclónicos, por lo regular, mayores que las primeras. En nuestro país, la Costa del Océano Pacífico y la Costa del Golfo de México, en los eventos más recientes Veracruz y Quintana Roo, son afectadas constantemente por ciclones.

La función Doble Gumbel se utiliza para caracterizar dos poblaciones de distinto origen. Es decir, cuando los registros tienen un origen hidrológico distinto de un año a otro. Se ha demostrado que la función de probabilidad se puede expresar:

$$F(x) = pF_1(x) + (1-p)F_2(x) \qquad (2.37)$$

Donde

$F_1(x)$, función de distribución de probabilidad de los registros producidos por lluvias normales de verano.

$F_2(x)$, función de distribución de probabilidad de los registros producidos por ciclones.

p, es la probabilidad de que en un año cualquiera el registro no sea producido por un ciclón.

Para la determinación de los parámetros de la función se puede utilizar el método de mínimos cuadrados para la minimización del error estándar. Para ello, es necesario determinar p.

$$p = \frac{N_n}{N_T} \qquad (2.38)$$

Donde

N_n, número de años de registro ciclónicos.

N_T, número de años de registro en total.

En general, es posible observar con cierta facilidad un salto brusco en valores de registros ciclónicos. Cuando se tienen dudas, es conveniente probar con varios valores de N_n, de tal manera que se logre un buen ajuste. Dada la experiencia con otros estudios hidrológicos en México, se recomienda un valor tentativo de $p = 0.84$. Al aplicar esta proporción, se está suponiendo que aproximadamente cada 6 años se presentan tormentas ciclónicas.

Una vez estimado p, se evalúan el resto de los parámetros con los métodos expuestos en 0. Se acepta que $F_1(x)$ y $F_2(x)$ tienen una distribución de probabilidad Gumbel respectivamente.

$$F(x) = p\left(e^{-e^{-\alpha_1(x-\beta_1)}}\right) + (1-p)\, e^{-e^{-\alpha_2(x-\beta_2)}} \qquad (2.39)$$

Donde

α_1, parámetro que se determina con la ecuación (2.28)

β_1, parámetro que se determina con la ecuación (2.29)

Y su función de densidad es:

$$f(x) = p\,\alpha_1\, e^{-e^{-\alpha_1(x-\beta_1)}} + (1-p)\,\alpha_2\, e^{-e^{-\alpha_2(x-\beta_2)}} \qquad (2.40)$$

En la Figura 2.5.3.1 se muestra una típica distribución Doble Gumbel, donde se muestra en el eje inferior de las abscisas la probabilidad de ocurrencia, y en la parte superior el periodo de retorno. En el eje de las ordenadas se muestra, en este caso, el gasto pero como se ha mencionado con anterioridad podría tratarse de una precipitación.

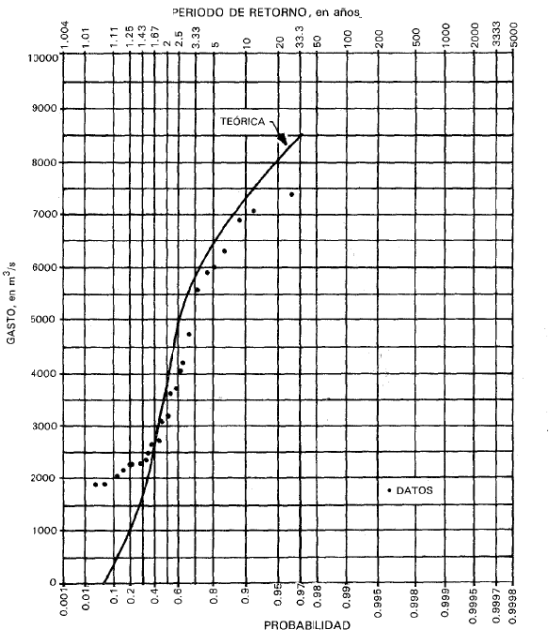


Figura 2.5.3.1. Gráfica Doble Gumbel o de dos poblaciones.

En la Figura 2.5.3.2, se muestra la gráfica correspondiente a la estación Comala, con clave 6007 del Estado de Colima con la transformación propuesta en el subcapítulo anterior 2.5.2.

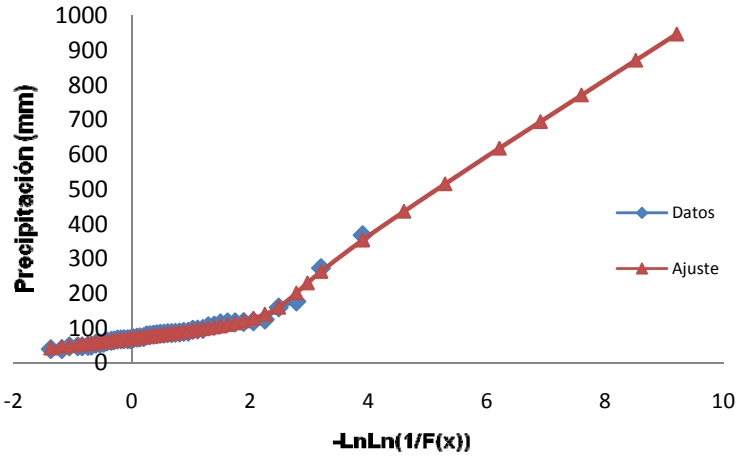


Figura 2.5.3.2. Gráfica Doble Gumbel de la estación Comala con P=0.84.

2.6. Periodo de retorno.

En el análisis hidrológico, resulta de primer interés conocer la frecuencia de una serie de registros de gastos o lluvia de una determinada estación. El objetivo es determinar el intervalo de recurrencia o periodo de retorno Tr en años, de un evento hidrológico de una magnitud x.

El periodo de retorno se define como el número de años que transcurren en promedio para que un evento de magnitud x sea igualado o excedido por lo menos una vez en ese periodo de tiempo.

$$Tr = \frac{1}{P(X \geq x)} = \frac{1}{[1 - P(X \leq x)]} \qquad (2.41)$$

2.6.1. Relación función de distribución-periodo de retorno.

En un análisis de frecuencia de datos, es importante establecer la relación entre su magnitud y periodo de retorno. Cuando se trabaja con series de registros máximos anuales y funciones de distribución de una variable continua, la relación de la ecuación (2.16), se puede reescribir,

$$Tr = \frac{1}{1 - F(x)} \qquad (2.42)$$

De tal forma que

$$\frac{1}{F(x)} = \frac{Tr}{Tr - 1} \qquad (2.43)$$

En hidrología, se utiliza este concepto por encontrarse en la misma unidad de tiempo (años regularmente) que la vida útil de las obras. Debido a que en este trabajo se utilizaron ajustes Gumbel y Doble Gumbel, se muestra que al sustituir la ecuación (2.43) en la ecuación (2.36) se obtiene la ecuación de una recta, pero ahora en términos del periodo de retorno. Esto es más útil ya que al momento de realizar un diseño, se requiere conocer el Tr de un evento y no tanto su probabilidad de excedencia u ocurrencia como se muestra a continuación.

$$-\ln \ln \frac{Tr}{Tr - 1} = \alpha(x - \beta) \qquad (2.44)$$