METODOS ESTADÍSTICOS

Introducción.

Uno de los objetivos de la asignatura de Hidrología, es mostrar a los alumnos, las herramientas de cálculo utilizadas en Hidrología Aplicada para diseño de Obras Hidráulicas. Una de esas herramientas de cálculo que se utiliza es a través del uso de las técnicas estadísticas para determinar los eventos de diseño máximos, asociados a diferentes periodos de retorno.

Este procedimiento de cálculo se fundamenta en correlacionar los registros históricos de las estaciones hidrométricas, con las diferentes distribuciones de probabilidad que existen.

Sin embargo, el desarrollar en forma completa un ejemplo de aplicación real, a través de esta técnica, conlleva varias horas clase que en muchos casos hace que el alumno pierda de vista el concepto fundamental, por esa razón, en este trabajo, se plantea utilizar el proceso de automatización de los métodos estadísticos, a través de programar en Visual Basic for Application de Excel, las distribuciones de probabilidad más utilizadas, con el objetivo de que el alumno, interactúe con la computadora para darle solución a un problema y preste más atención a los conceptos básicos del tema, así como pueda utilizar el programa para asignaturas consecuentes.

El trabajo consiste en tres partes fundamentales:

La primera, muestra los aspectos teóricos relacionados con los métodos estadísticos para maximización de eventos hidrológicos, la segunda, presenta las características del programa de automatización y finalmente, la tercera parte, presenta un ejemplo de aplicación a un registro hidrométrico real.

Asimismo, este material didáctico incluye un disco que contiene el archivo **"Analisis de frecuencia.xls",** que sirve para aplicar el proceso de automatización a cualquier registro hidrométrico.

I MÉTODOS ESTADÍSTICOS EN HIDROLOGÍA.

I.1 ANÁLISIS DE FRECUENCIA.

Uno de los problemas más importantes en la hidrología consiste en obtener una interpretación de eventos probabilísticos a futuro, asociados a un registro en el pasado.
Ejemplo de este caso, es la estimación de gastos máximos y su procedimiento se conoce con el nombre de análisis de frecuencia.
Muchos procesos en Hidrología deben ser analizados y explicados con base a la ciencia probabilística, por su inherente aleatoriedad. Por lo tanto, no es posible predecir una avenida o una precipitación con base únicamente determinística. Afortunadamente, los métodos estadísticos permiten presentar, organizar y reducir datos para facilitar su interpretación y evaluación. Esta parte del trabajo presenta los gastos máximos anuales cuantificados y presentados con distribuciones de probabilidad continua.
Muchas funciones de densidad de probabilidad continuas son usadas en la Hidrología, sin embargo este trabajo hace énfasis solo en las más comunes. Ellas son:

Distribución Exponencial con dos parámetros
Distribución Gamma de dos parámetros.
Distribución Gamma de tres parámetros (Pearson tipo III).
Distribución General de Valores Extremos (Gumbel)
Distribución Gumbel de dos poblaciones (Gumbel 2p)
Distribución Log-Normal.
Método de Nash.
Distribución Normal.

Para poder correlacionar una muestra de registro hidrométrico a una distribución de probabilidad, se requiere de un método de estimación de parámetros que permita relacionar la información muestral con la poblacional, los métodos de estimación de parámetros que se conocen son:

Momentos. Iguala momentos poblacionales con muestrales.
Máxima Verosimilitud. Supone que el mejor parámetro de una función debe ser aquel que maximiza la probabilidad de ocurrencia de la muestra observada.

Mínimos cuadrados. Minimiza la suma de los cuadrados de todas las desviaciones entre los valores calculados y observados.
Probabilidad Pesada. Deriva expresiones para los parámetros de distribuciones cuyas formas inversas se puede definir inversamente.
Sextiles. El rango de la variable es dividida en 6 intervalos, tal que la probabilidad acumulada en cada intervalo es de un sexto.
Momentos L.

Este material didáctico, considera el método de momentos para la estimación de parámetros en las funciones de distribución.

Se debe recordar que una variable aleatoria, es aquella que no se puede predecir con certeza al realizar un experimento y su comportamiento se describe mediante su ley de probabilidades, la cual se especifica por su función de densidad de probabilidad f(x), o por su función de densidad acumulada F(x) que representa el área bajo la curva de la función de densidad, representando la probabilidad de ocurrencia del evento.

I.2 DISTRIBUCIÓN EXPONENCIAL CON DOS PARÁMETROS.

La función de distribución exponencial se define como:

$$F(x) = \int_0^x \left(1 - e^{-\beta x}\right) dx \qquad \text{(I.1)}$$

y la función de densidad de probabilidad es:

$$f(x) = \beta e^{-\beta x} \qquad \text{(I.2)}$$

donde, β se conoce como parámetro de escala.

La estimación del parámetro de escala por el método de momentos se hará a través de la siguiente ecuación:

$$\beta = \frac{1}{\hat{x}}$$

(I.3)

donde x, es la media de la muestra, que se calculará a través de la siguiente expresión:

$$\hat{x} = \sum_{i=1}^{n} \frac{x_i}{n} \qquad \text{(I.4)}$$

La ecuación para determinar los gastos calculados a través de la muestra con la distribución Exponencial es:

$$Q_{calc.} = \frac{Ln\left(\frac{1}{T}\right)}{-\beta} \qquad \text{(I,5)}$$

donde, T es el periodo de retorno en años y Qcalc, es el gasto de diseño calculado con la distribución exponencial para un periodo de retorno dado.

I.3 DISTRIBUCIÓN GENERAL DE VALORES EXTREMOS I. (GUMBEL)

Supóngase que se tienen N muestras, cada una de las cuales contiene n eventos. Si se selecciona el máximo x de los n eventos de cada muestra, es posible demostrar que, a medida que n aumenta, la función de distribución de probabilidad de x tiende a:

$$F(x) = \int_{0}^{x} e^{-e^{-\alpha(x-\beta)}} dx \qquad \text{(I.6)}$$

La función de densidad de probabilidad es entonces:

$$f(x) = \alpha e^{\left[-\alpha(x-\beta)-e^{-\alpha(x-\beta)}\right]} \qquad \text{(I.7)}$$

donde α y β son los parámetros de escala y forma de la función, y se estiman por el método de momentos como α = 0.78 s y β = x -0.5772α, donde x

representa la media de la muestra y se valúa con la ecuación I.4 y s es la desviación estándar que se calculará con la siguiente ecuación:

$$s = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \hat{x})}{n-1}} \qquad \text{(I.8)}$$

Despejando "x" de la ec. 1.6, la ecuación para determinar los gastos a través de la distribución Gumbel es:

$$x = Q_{calc} = \beta - \alpha\left[\ln\left(\ln\frac{T}{T-1}\right)\right] \qquad \text{(I.9)}$$

I.4 MÉTODO DE NASH.

Utilizando la función de distribución de probabilidad de Gumbel de una población, Nash propone la siguiente metodología para calcular los parámetros de la función:
Sea:

$$Q_{calc} = a + c\ln\ln\left(\frac{T}{T-1}\right) \qquad \text{(I.10)}$$

Comparando la ec. I.10 con la I.9, a = β y c = -α

Con un cambio de variable, la ec. I.10 queda:

$$Q_{calc} = a + cx \qquad \text{(I.11)}$$

$$x = \ln\ln\left(\frac{T}{T-1}\right) \qquad \text{(I.12)}$$

y "a" y "c" son los parámetros de la función, que se obtendrán a través de un análisis de correlación lineal simple con el criterio de los mínimos cuadrados.

$$a = \frac{\sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} x_i y \sum_{i=1}^{n} x_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} \qquad \text{(I.13)}$$

$$b = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} \qquad \text{(I.14)}$$

El subíndice i representa los datos muestrales.

I.5 DISTRIBUCIÓN GUMBEL DE DOS POBLACIONES (GUMBEL 2P).

En muchos lugares, los gastos máximos anuales pertenecen a dos poblaciones diferentes, la primera es la de los gastos producidos por precipitaciones relacionadas con los fenómenos meteorológicos dominantes en la región en estudio, y la segunda es la de los gastos producidos por precipitaciones ciclónicas, normalmente mayores que los primeros.

Se ha demostrado que, en estos casos, la función de distribución de probabilidad se puede expresar como:

$$F(x) = F_1(x)\left[p + (1 - p)F_2(x)\right] \qquad \text{(I.15)}$$

donde $F_1(x)$ y $F_2(x)$ son, respectivamente, las funciones de distribución de probabilidad de los gastos máximos anuales producidos por tormentas ciclónicas y de los producidos por ellas, y p es la probabilidad de que en un año cualquiera el gasto máximo no sea producido por una tormenta ciclónica. El número de parámetros de la función anterior es:

$$n = n_1 + n_2 + 1 \qquad \text{(I.16)}$$

donde n_1 = número de parámetros de $F_1(x)$, n_2 = número de parámetros de $F_2(x)$ y el parámetro restante es p. El valor de p será entonces:

$$p = \frac{N_n}{N_T} \qquad \text{(I.17)}$$

donde N_n es el numero de años de registro en que el gasto máximo no se produce por una tormenta ciclónica y N_T es el número total de años de registro.

$F_1(x)$ y $F_2(x)$ son del tipo Gumbel, por lo que la función de probabilidad queda así:

$$F(x) = e^{-e^{-\alpha_1(x-\beta_1)}}\left[p + (p-1)e^{-e^{-\alpha_2(x-\beta_2)}} \right] \qquad \text{(I.18)}$$

donde α_1 y β_1 son los parámetros correspondientes a la población no ciclónica y α_2 y β_2 corresponden a la ciclónica.

La estimación de parámetros α_1 , β_1 , α_2 y β_2, por momentos se calculan con el mismo criterio de la distribución Gumbel de 1 población.

En este caso no es posible determinar una ecuación para el cálculo de gastos máximos debido a que la función de distribución de probabilidad de Gumbel de dos poblaciones es implícita, eso implica que la solución de dicha ecuación debe realizarse a través de algún método para determinar raíces en una función.

I.6 DISTRIBUCIÓN NORMAL.

La función de densidad de probabilidad normal se define como:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)} \qquad \text{(I.19)}$$

donde, μ y σ son los parámetros de la distribución. Estos parámetros determinan la forma de la función f(x) y su posición en el eje x.

Los valores de μ y σ son la media y la desviación estándar de la población y pueden estimarse como la media y desviación estándar de los datos. La función de distribución de probabilidad normal es:

$$F(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)} dx \qquad \text{(I.20)}$$

Como se sabe, hoy en día no se conoce analíticamente la integral de la ecuación F(x), por lo que es necesario recurrir a métodos numéricos para valuarla. Sin embargo, para hacer esto se requiere una tabla para cada valor de μ y σ, por lo que se ha definido la variable estandarizada:

$$z = \frac{x - \mu}{\sigma} \qquad \text{(I.21)}$$

que está normalmente distribuida con media cero y desviación estándar unitaria. Así la función de distribución de probabilidad se puede escribir como:

$$F(x) = F(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi\sigma}} e^{\frac{-z^2}{2}} dz \qquad \text{(I.22)}$$

La función F(z) se ha calculado numéricamente y se han publicado tablas de ella. Debido a que la función F(z) es simétrica, en dicha tabla se encuentran únicamente valores de:

$$\int_{0}^{z} \frac{1}{\sqrt{2\pi\sigma}} e^{\frac{-z^2}{2}} dz \qquad \text{(I.23)}$$

con lo que es posible calcular F(z) para cualquier valor de z.

Otra manera más conveniente de estimar f(z) o F(z), es mediante fórmulas aproximadas. La función de densidad f(z) se aproxima, como:

$$z = w - \frac{C_0 + C_1 w + C_2 w^2}{1 + d_1 w + d_2 w^2 + d_3 w^3} \qquad \text{(I.24)}$$

Donde
C_o = 2.515517
C_1 = 0.802853
C_2 = 0.010328
d_1 = 1.432788
d_2 = 0.189269
d_3 = 0.001308

$$w = \sqrt{\ln\left(\frac{1}{(1 - P(t))^2}\right)} \qquad \text{(I.25)}$$

donde:

$$P(t) = 1 - \frac{1}{T} \qquad \text{(I.26)}$$

Para calcular los gastos máximos de diseño con esta distribución, se utiliza la siguiente expresión:

$$Q_{calc} = \overline{Q} + s_z \qquad \text{(I.27)}$$

donde: $\overline{Q}$ y s son respectivamente la media y desviación estándar de la muestra.

I.7 DISTRIBUCIÓN LOG-NORMAL.

En esta función los logaritmos naturales de la variable aleatoria se distribuyen normalmente. La función de densidad de probabilidad es:

$$f(x) = \frac{1}{\sqrt{2\pi}} \frac{1}{x\beta} e^{\frac{-1}{2}\left(\frac{\ln x - \alpha}{\beta}\right)} \qquad \text{(I.28)}$$

donde α y β son los parámetros de la distribución. Los valores de α y β son respectivamente la media y la desviación estándar de los logaritmos de la variable aleatoria.

Esta función no necesariamente es simétrica. Los valores de α y β se estiman a partir de n observaciones x_i, i = 1, 2, 3, ...n, como :

$$\alpha = \sum_{i=1}^{n} \frac{\ln(x_i)}{n} \qquad \text{(I.29)}$$

$$\beta = \sum_{i=1}^{n} \left(\frac{\left(\ln x_i - \alpha\right)^2}{n} \right)^{\frac{1}{2}} \qquad \text{(I.30)}$$

La función de distribución de probabilidad es:

$$F(x) = \int_0^x \frac{1}{\sqrt{2\pi}} \frac{1}{x\beta} e^{-\frac{1}{2}\left(\frac{\ln x - \alpha}{\beta}\right)} dx \qquad \text{(I.31)}$$

Los valores de la función de distribución de probabilidad, se obtienen usando la misma variable estandarizada, que se define para este como:

$$z = \frac{\ln x - \alpha}{\beta} \qquad \text{(I.32)}$$

Por lo que para calcular los gastos máximos de diseño se utiliza la siguiente expresión:

$$Q_{calc} = \overline{Q} + s_z \qquad \text{(I.33)}$$

donde: $\overline{Q}$ y s son respectivamente la media y desviación estándar de los logaritmos de la muestra.

I.8 DISTRIBUCIÓN GAMMA DE DOS PARÁMETROS.

La función de distribución de probabilidad es:

$$F(x) = \int_0^x \left(\frac{x^{\beta-1} e^{\left(\frac{-x}{\alpha}\right)}}{\alpha^{\beta} \Gamma(\beta)} \right) dx \qquad \text{(I.34)}$$

La función de densidad de probabilidad gamma de dos parámetros se define como:

$$f(x) = \frac{x^{(\beta-1)} e^{\left(\frac{-x}{\alpha}\right)}}{\alpha^{\beta} \Gamma\left(\beta\right)} \qquad \text{(I.35)}$$

donde α y β son los parámetros de escala y forma de la función y $\Gamma(\beta)$ es la función Gamma.

Los parámetros α y β se evalúan por el criterio de momentos a partir de las siguientes ecuaciones:

$$\alpha = \frac{s^2}{\hat{x}} \qquad \text{(I.36)}$$

$$\beta = \left(\frac{\hat{x}}{s} \right)^2 \qquad \text{(I.37)}$$

donde x y s son la medía y desviación estándar de los datos.

Para obtener los eventos de diseño para diferentes periodos de retorno la distribución Gamma puede ser obtenida en forma aproximada utilizando la variable estandarizada z de la distribución Normal a través de la siguiente ecuación de aproximación:

$$Q_{calc.} = (\alpha)(\beta)\left[1 - \frac{1}{9\beta} + z\sqrt{\frac{1}{9\beta}}\right]^3 \qquad \text{(I.38)}$$

I.9 DISTRIBUCIÓN PEARSON TIPO III. (GAMMA CON TRES PARÁMETROS)

La función de densidad de probabilidad de la distribución Pearson tipo III se define como:

$$F(x) = \frac{1}{\alpha\Gamma(\beta)}\left(\frac{x - x_0}{\alpha}\right)^{\beta-1} e^{-\left(\frac{x-x_0}{\alpha}\right)} \qquad \text{(I.39)}$$

donde: α, β y x_0 son los parámetros de escala, de forma y de ubicación respectivamente.

Los parámetros α, β y x_0 se estiman a través del criterio de momentos con las siguientes ecuaciones:

$$\alpha = \frac{gs}{2} \qquad \text{(I.40)}$$

$$\beta = \frac{4}{g^2} \qquad \text{(I.41)}$$

$$x_0 = \hat{x} - \alpha\beta \qquad \text{(I.42)}$$

donde, x, s y g son la media, la desviación estándar y el coeficiente de asimetría respectivamente de la muestra.
Para determinar los eventos de diseño para cualquier periodo de retorno la distribución Pearson tipo III puede ser evaluada a través de una aproximación con la variable estandarizada z de la distribución normal, utilizando la siguiente ecuación:

$$Q_{calc.} = (\alpha)(\beta)\left[1 - \frac{1}{9\beta} + z\sqrt{\frac{1}{9\beta}}\right]^3 + x_0 \qquad \text{(I.43)}$$

I.10 MÉTODO DEL ERROR CUADRÁTICO MÍNIMO.

Consiste en calcular, para cada función de distribución, el error cuadrático como:

$$E = \left[\sum_{i=1}^{n} (Qci - Qmi)^2 \right]^{\frac{1}{2}} \qquad \text{(I.44)}$$

Donde, Qci es el i-ésimo dato calculado con la distribución de probabilidad, Qmi es i-ésimo dato del registro hidrométrico en cuestión y E es el error cuadrático mínimo.

La función de probabilidad de mayor ajuste al registro hidrométrico será entonces aquella que cumpla un valor de E cercano a cero.

I.11 PROCEDIMIENTO DE CÁLCULO PARA CORRELACIONAR UNA MUESTRA A UNA DISTRIBUCIÓN DE PROBABILIDAD.

1. Obtener los valores de la muestra.

2. Ordenar los gastos de la muestra de mayor a menor (Ya que el objetivo es determinar gastos máximos).

3. Calcular el periodo de retorno para cada año de registro, a través de la ecuación:

$$T = \frac{n+1}{m} \qquad \text{(I.45)}$$

donde

n, número de años del registro
m, número de orden que se asigna a la muestra
T, periodo de retorno.

4. Determinar la probabilidad de excedencia

$$P(x) = \frac{1}{T} \qquad \text{(I.46)}$$

(En algunos casos será necesario determinar la probabilidad de no excedencia)

5. Calcular el gasto máximo de acuerdo a la distribución de probabilidad elegida.
6. Determinar los errores al cuadrado, con base en las diferencias entre gasto calculado y medido.
7. Finalmente, evaluar la sumatoria de errores al cuadrado.

Ejemplo:
Los siguientes gastos anuales han sido obtenidos de los registros hidrométricos de un río. Estime la magnitud del gasto para un periodo de retorno de 20 años.

Año	Q(m^3/s)	Q	m	T	T/T-1	Xi	X^2	$Q_i^2*10^6$	XiQi
1967	4000	5100	1	13	1.08	-2.50	6.55	26.01	-12877
1968	5100	4400	2	6.5	1.18	-1.79	3.20	19.36	-7871.6
1969	3270	4000	3	4.33	1.30	-1.34	1.80	16.00	-5352.0
1970	2860	3690	4	3.25	1.44	-1.00	1.00	13.63	-3690
1971	2660	3460	5	2.60	1.63	-0.72	0.52	11.97	-2501.6
1972	4400	3270	6	2.17	1.85	-0.49	0.24	10.69	-1566.3
1973	3690	3120	7	1.86	2.16	-0.26	0.07	9.73	-801.8
1974	3120	2990	8	1.63	2.59	-0.05	0.0025	8.94	-137.5
1975	3460	2860	9	1.44	3.27	0.17	0.0289	3.18	471.9
1976	2570	2760	10	1.30	4.33	0.38	0.14	7.62	1057.1
1977	2760	2660	11	1.18	6.55	0.63	0.40	7.08	1665.2
1978	2990	2570	12	1.08	13.50	0.96	0.92	6.61	2423.5

$\sum 40880 \qquad \sum -6.04 \qquad \sum 14.642 \qquad \sum 145.8 \qquad \sum -29180.7$

sustituyendo en la ecuación:

$$Qm = \frac{40880}{12} = 3406.7\left[m^3 / s\right]$$