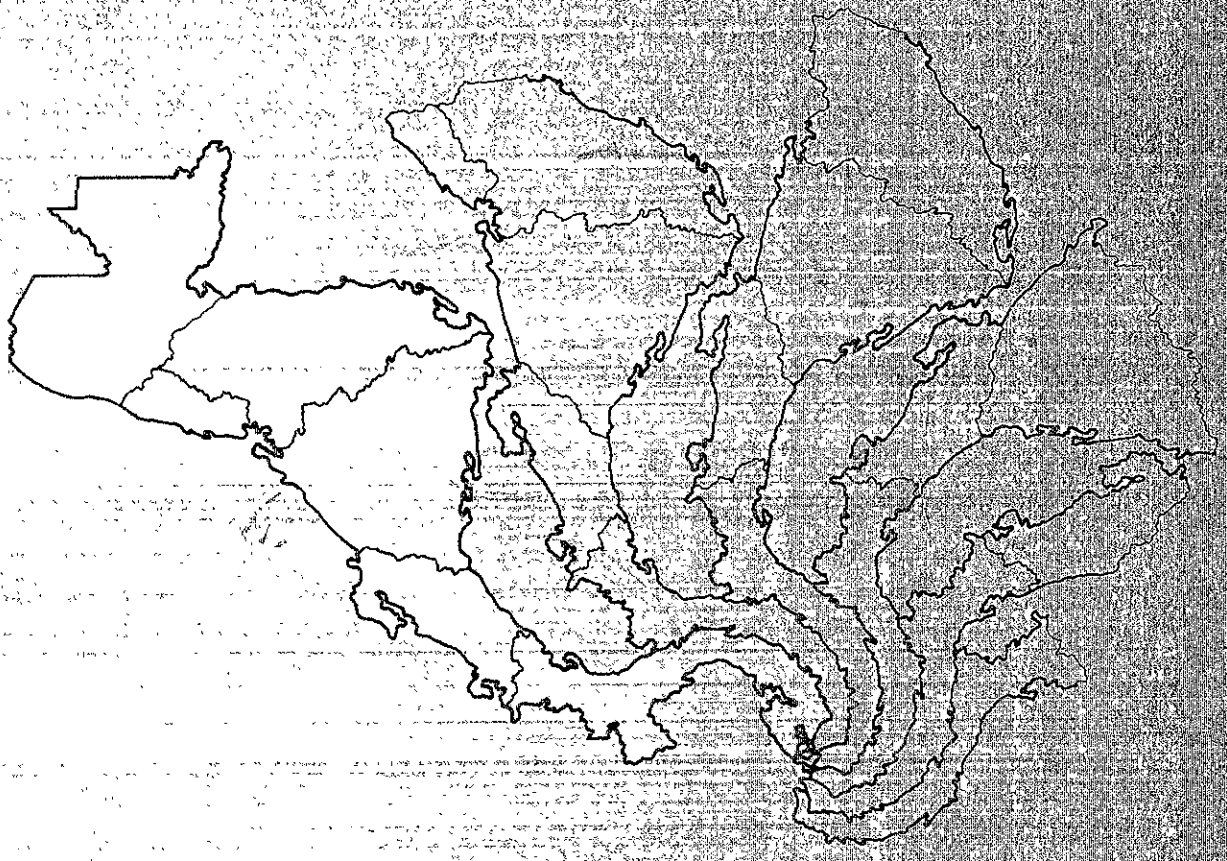
PROGRAMA DE ACTIVIDADES REGIONALES EN EL SUBSECTOR ELECTRICO DEL ISTMO CENTROAMERICANO (PARSEICA)

MODULO DE PLANEAMIENTO OPERATIVO



CONCEPTOS BASICOS DE PLANEAMIENTO OPERATIVO

Seminario I

Febrero de 1992

Documento preparado por PROMON, con el apoyo de ELETROBRAS.

CAPÍTULO 11
REVISIÓN DE PRINCIPIOS DE PROBABILIDAD Y ESTADÍSTICA

11.1 DISTRIBUCIÓN DE PROBABILIDAD

Se presenta un texto introductório, bastante reducido, que no tiene la menor pretensión de discurrir sobre la materia con rigor matemático. Al contrario, el asunto fluye de cuestiones familiares al planeamiento de sistemas hidroeléctricos, presentando simultáneamente conceptos de Probabilidad y de Estadística. En los libros didácticos estas dos materias son usualmente presentadas por separado, una a la vez.

Comencemos por una pregunta para la cual todo planeador de sistemas hidroeléctricos adoraría tener una respuesta:

> ¿"Cuál será el volumen total de agua que un río cualquiera depositará en el embalse de una central hidroeléctrica a lo largo de un año cualquiera del futuro?"

Nadie sabe responder con precisión esta cuestión. Se dice que este volumen desconocido, que dependerá de acontecimientos futuros, es una variable aleatoria, designada usualmente por una letra mayúscula, por ejemplo X. Todo lo que podemos hacer es colectar los volúmenes observados en el pasado y procurar identificar las regularidades estadísticas de esta colección de números.

Por ejemplo, la Tabla 11.1 muestra los valores observados para las afluencias al embalse de Funil, puesto fluviométrico de Barra de Pirai, Río Paraiba do Sul (Brasil), de 1921 a 1970. Los valores están expresos en caudal medio anual, que es el volumen total dividido por el número de segundos de un año. Cada valor de la Tabla 11.1 es una observación de la variable aleatoria X. Usualmente se representan simbólicamente estas observaciones por letras minúsculas. Así $x(1)=336$, $x(2)=329$, $x(3)=361$,... $x(n)=216$.

Serie de caudales anuales del río Paraíba do Sul, en Barra do Piraí, RJ.

Subserie 1		Subserie 2	
Año	Caudal (m^3/s)	Año	Caudal (m^3/s)
21-22	336	46-47	406
22-23	329	47-48	380
23-24	361	48-49	271
24-25	247	49-50	361
25-26	387	50-51	376
26-27	385	51-52	331
27-28	306	52-53	197
28-29	421	53-54	223
29-30	313	54-55	177
30-31	416	55-56	214
31-32	355	56-57	303
32-33	253	57-58	293
33-34	258	58-59	319
34-35	306	59-60	282
35-36	299	60-61	401
36-37	292	61-62	308
37-38	329	62-63	286
38-39	304	63-64	201
39-40	276	64-65	285
40-41	206	65-66	362
41-42	274	66-67	515
42-43	265	67-68	265
43-44	298	68-69	200
44-45	273	69-70	303
45-46	284	70-71	216

Observaciones

1) Media muestral

$$\mu = \frac{1}{n}\sum_{i=1}^{n} q_i$$

2) Desvío padrón muestral

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(q_i - \mu)^2}$$

3) Coeficiente de variación muestral

$$C_v = \frac{\hat{\sigma}}{\hat{\mu}}$$

4) Coeficiente de asimetría muestral

$$\gamma = \frac{1}{\hat{\sigma}^3}\cdot\frac{n}{(n-1)(n-2)}\sum_{i=1}^{n}(q_i - \hat{\mu})^3$$

5) Coeficiente de auto correlación muestral

$$\rho = \frac{\sum_{i=1}^{n-1}(q_i-\hat{\mu})(q_i+1-\hat{\mu})}{\sum_{i=1}^{n}(q_i-\mu)^2}$$

n = número de años

	Subserie 1	Subserie 2	Serie completa
μ	311	299	305
σ	53	79	67
C_v	0,17	0,26	0,22
γ	0,45	0,63	0,49
ρ	0,22	0,29	0,26

Tabla 11.1

Esta colección es designada de *muestra* de la variable aleatoria X, siendo n el tamaño de la muestra. En el caso de la Tabla 11.1, n=50. Es necesario tener plena percepción de la diferencia entre la variable aleatoria X, cuyo valor es desconocido y una *observación* de la variable aleatoria, que es un número conocido, x.

Para cualquier par de números (a, b), se dice que existe una *probabilidad* de que X asuma un valor en el intervalo (a, b). Esta probabilidad es un número entre 0 y 1, designado simbólicamente por $P(a<X\leq b)$. Análogamente, $P(X<a)$ y $P(X>b)$ significan respectivamente la probabilidad de que X sea menor que a y mayor que b.

Para el ejemplo presentado, está claro que el volumen total afluente no puede ser negativo, lo que implica que $P(X<0)=0$. Por otro lado, $P(X>0)=1$. En otras palabras, ciertamente el valor de X será positivo. Imaginándose el dominio de X como la semi-recta positiva, X puede asumir cualquier punto de un número infinito de puntos que componen la semi-recta. Para un valor de x cualquiera, X será ciertamente menor o igual a x, o entonces X será mayor que x. Luego $P(X \leq x) + P(X>x) = 1$.

Por facilidad de notación, se define la "función acumulada de probabilidad" $F(x)$, para todo x del dominio de X, de tal manera que:

$$P(X\leq x) = F(x)$$
$$P(X>x) = 1 - F(x) \qquad (1)$$

La probabilidad de que X pertenezca al intervalo (a, b), es decir $P(a<X\leq b) = P(X\leq b) - P(X<a)$ es dado por:

$$P(a<X\leq b) = F(b) - F(a)$$

La derivada de $F(x)$ en relación a x es también una función de x y es llamada "función densidad de probabilidad de x":

$$f(x) = dF(x)/dx \qquad (2)$$

Como $F(x)$ es no decreciente, $f(x)$ es no negativo. Al contrario de $F(x)$, que es adimensional, $f(x)$ tiene dimensión inversa de x. Así, si x es medido en m^3/s, $f(x)$ será medido en s/m^3. Fuera del dominio de X, $f(x) = 0$.

La probabilidad de que x pertenezca al intervalo (a, b) es también dada por:

$$P(a<X\leq b) = F(b) - F(a) = \int_a^b f(x)\,dx \qquad (3)$$

El menor y el mayor valor de la Tabla 11.1 son respectivamente 177 y 515 (m^3/s). Está claro que nada impide que X asuma un valor fuera de este dominio. O sea, la probabilidad de que en el futuro se observe una afluencia menor o igual que la mínima afluencia registrada en el histórico de caudales no es cero: $F(177)>0$. No obstante, esta ocurrencia no es "probable", como se dice, significando que $F(177)$ es un número mucho más próximo de 0 de que de 1. Sin embargo, la simple observación de la Tabla 11.1 no nos permite precisar el valor de $F(177)$.

En busca de una solución, debemos examinar los datos disponibles. Los valores de la Tabla 11.1 pueden ser ordenados crecientemente, con el propósito de obtener una "nueva" muestra, re-ordenada, $x(1) < x(2) < x(3) < ... < x(n)$. A cada valor de $x(i)$, para i entre 1 y n, se asocia la frecuencia relativa de observaciones menores o iguales a $x(i)$, esto es, i/n. Con los pares de valores $(x(i); i/n)$ se produce el gráfico de la función "escalera" $G(x)$ mostrada en la Figura 11.1. Nótese que $G(x)$ es la frecuencia relativa de observaciones menores o iguales a x, aún cuando x no coincide con ninguna de las observaciones de la muestra. Así, por ejemplo, para $x = 240$, tenemos $G(240) = 8/50 = 0.16$.

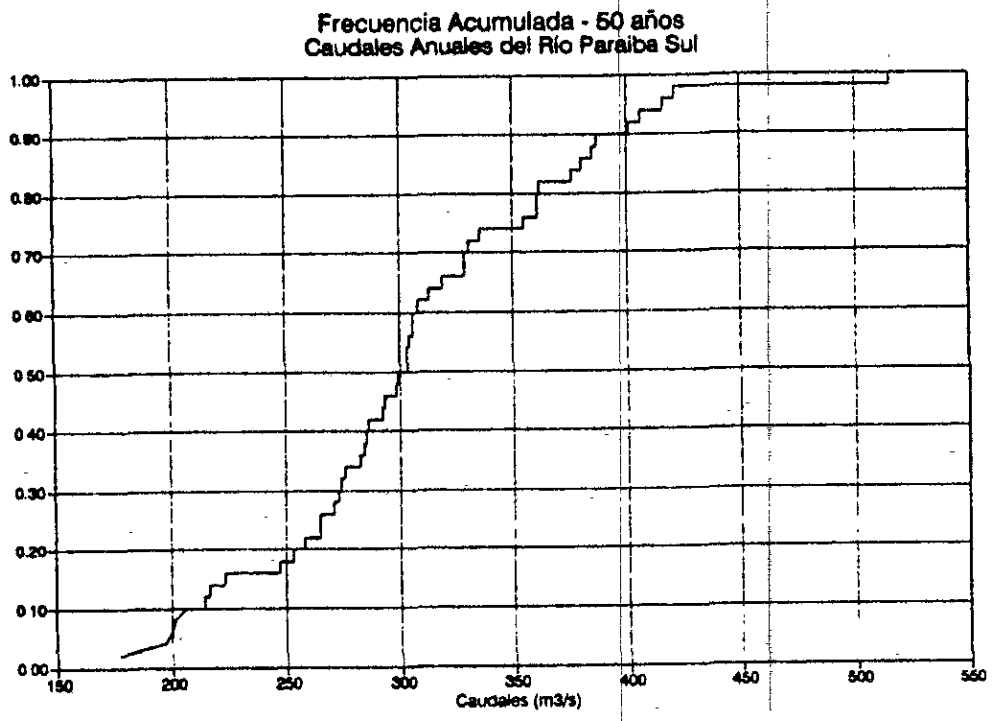


Figura 11.1

En caso que la Naturaleza sólo consiguiese "sortear" cada año uno de los valores de la Tabla 11.1, y ningún valor diferente, $F(240)$ sería igual a $G(240)$. Sin embargo este no es evidentemente el caso, visto que la Naturaleza no se contenta con tan poca variedad. O sea, el futuro no es obligado a repetir el pasado. Sin embargo, $G(240)$ puede ser visto como una "estimación" del valor desconocido de $F(240)$. Es intuitivo que esta estimación será tanto más precisa cuanto mayor sea el tamaño de la muestra n. En la realidad se puede probar que cuando n tiende a infinito, $G(x)$ tiende a $F(x)$.

Conceptualmente esto significa que si tuviésemos el privilegio de haber observado la Naturaleza por un número infinito de años, entonces podríamos conocer con precisión cual es la "ley probabilística" que ella adopta, traducida por la función $F(x)$, también llamada de "población". Como esto nunca será el caso, tendremos que contentarnos con aproximaciones.

En la Figura 11.2 se reproduce la función $G(x)$, junto con las funciones $G_1(x)$ y $G_2(x)$, derivadas respectivamente de la primera mitad (años de 1921 a 1945) y de la segunda mitad (años de 1946 a 1970).

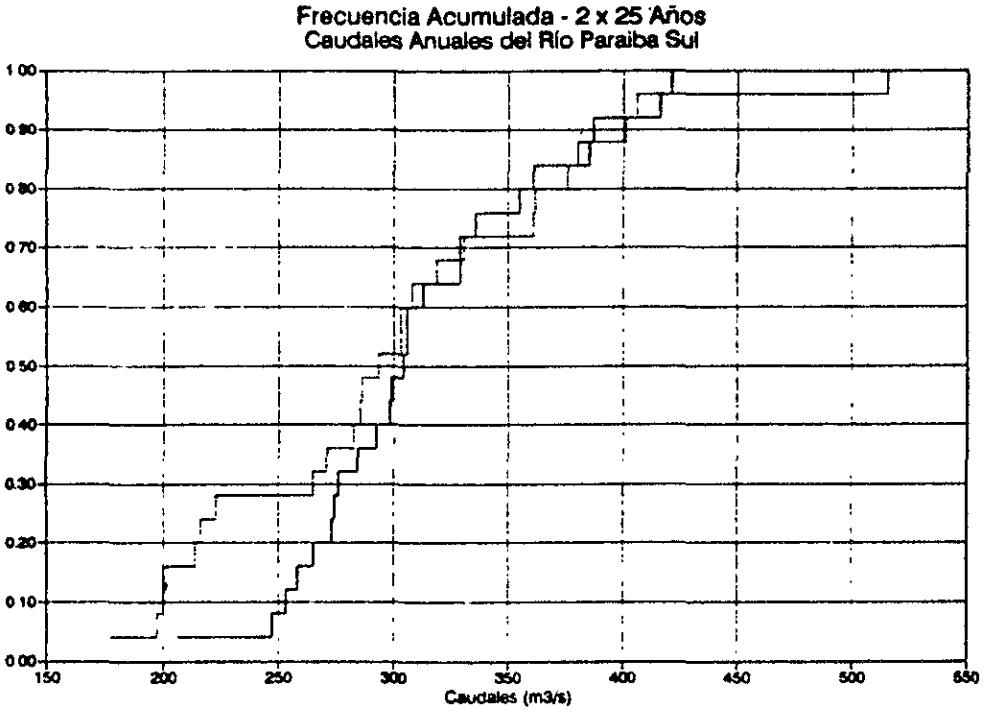


Figura 11.2

Se puede notar que $G_1(240) = 1/25 = 0.04$ es bastante diferente de $G_2(240) = 7/25 = 0.28$. Si tuviésemos otros 25 años obtendríamos un tercero valor, también diferente. A esta diversidad de estimaciones se le llama "variación muestral". Naturalmente $G(x) = (G_1(x) + G_2(x)) / 2$. Por ejemplo, para el caso en que x es igual a 240, tenemos $0.16 = (0.04 + 0.28) / 2$. Si tuviésemos un número infinito de sub-muestras, en vez de apenas dos, $G(x)$ convergiría para $F(x)$.

Una de las cuestiones fundamentales de la Estadística es estimar la función $F(x)$, conocida una muestra de X. En general, se define arbitrariamente para $F(x)$ una expresión matemática simple, con pocos parámetros a ser estimados a partir de los datos existentes. $F(x)$ debe ser una función no decreciente, ya que $P(X \leq a) \leq P(X \leq b)$, para $a < b$. Además, $F(x)$ debe tender para 0 cuando x tiende para el límite inferior del dominio de X, que puede ser menos infinito. Análogamente, F(x) debe tender para 1 cuando x tiende para el límite superior del dominio de X, que puede ser más infinito.

Por ejemplo, vamos suponer que la variable aleatoria X tenga la distribución de probabilidades "exponencial", dada por:

$$F(x) = 1 - e^{-\lambda x} \qquad (4)$$

definida para valores de x en el intervalo $(0, \infty)$, siendo λ el único parámetro.

Observe que $F(x)$ es una función creciente , además de ser positiva y que:

$$F(0) = 0 \text{ y } F(\infty) = 1$$

La función densidad de probabilidad de X es:

$$f(x) = \frac{\partial F(x)}{\partial x} = \frac{\partial (1 - e^{-\lambda x})}{\partial x} = \lambda e^{-\lambda x} \qquad (5)$$

Existen diversos métodos para estimar el (los) parámetro(s) de $F(x)$ a partir de la muestra. El más simple y más utilizado es el método de los momentos. Veamos un ejemplo.

La media aritmética de los valores de una muestra es llamada "media muestral" o "primer momento muestral".

$$\bar{x} = 1/n \sum_{i=1}^{n} x(i) \qquad (6)$$

En el caso de la Tabla 11.1, $\bar{x}$ = 305 m^3/s. Como el propio nombre lo dice, la media muestral depende de la muestra. Si considerásemos sólo la primera mitad de los datos de la Tabla 11.1, la media muestral sería 311 m^3/s.

La "media poblacional", el valor esperado de la variable aleatoria, o aún, el primer momento de población, es definida por:

$$E(X) = \int_R x f(x)\, dx$$

donde R es el dominio de la variable aleatoria X. Para la expresión matemática arbitraria seleccionada se tiene:

$$E(X) = \int_0^{\infty} x \lambda e^{-\lambda x} dx = 1/\lambda \qquad (7)$$

El método de los momentos, para este caso, consiste en igualar el valor numérico del primer momento muestral (la media aritmética de las observaciones) con el primer momento de población (expresión paramétrica del valor esperado de la variable aleatoria). Para el caso específico se tiene:

$$1/\hat{\lambda} = 305 \text{ o } \hat{\lambda} = 0.00328 \qquad (8)$$

Se utiliza el acento circunflejo "^" sobre λ para resaltar que 0.00328 es apenas una estimación del valor verdadero y desconocido de λ.

Adoptándose el valor numérico de la estimación de la Ecuación 4, es posible trazar el gráfico de la estimativa de la función acumulada de probabilidad $F(x)$, que surge en la Figura 11.3, juntamente con el gráfico de $G(x)$.

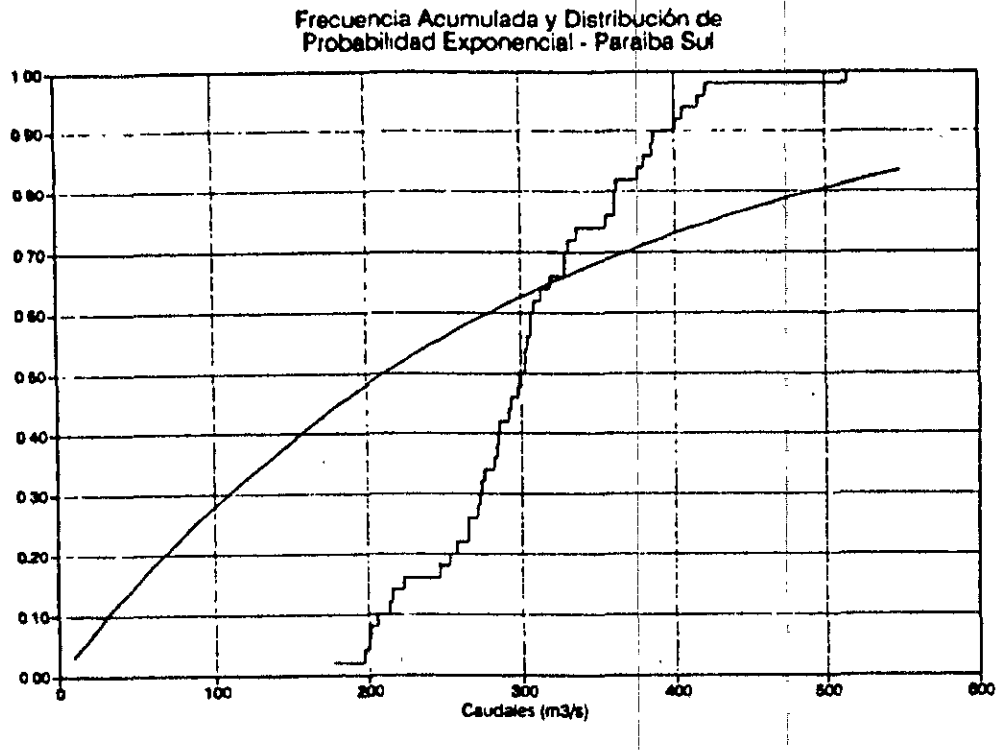


Figura 11.3

Ahora ya estamos en mejores condiciones para estimar la probabilidad de que en un año cualquiera del futuro ocurra una afluencia anual inferior a la mínima del histórico.

$$\hat{F}(177) = 1 - e^{-0{,}00328\,(177)} = 0.44$$

El lector tendrá razón al sospechar que el resultado arriba está incorrecto. En fin, si a lo largo de 50 años no se observó valor alguno inferior a 177 m^3/s, no parece aceptable que la probabilidad de que en un año cualquier del futuro ocurra una super-sequía sea de 0.44 (o como algunos prefieren, de 44%). De hecho, la elección de la Expresión 4 fue totalmente arbitraria, basada en la simplicidad matemática. No tenemos razón alguna para imaginar que la Naturaleza haya hecho la misma elección, lo que para el caso específico definitivamente no ocurrió. Esta impropiedad en la elección de la Ecuación 4 puede ser confirmada observándose la Figura 11.3, donde se nota una concordancia insatisfactoria entre $\hat{F}(x)$ y $G(x)$.

Este aparente fracaso en la utilización de la Ecuación 4 no nos debe inhibir de utilizarla en otras circunstancias. Por ejemplo, si estuviéramos interesados en caracterizar la variable aleatoria Y, definida por el total precipitado a lo largo de un día en el mes de enero, en la ciudad de Rio de Janeiro, Puesto Pluviométrico de Realengo, dado que ocurre alguna precipitación (es decir, descartados los casos en que $Y=0$).

Para el período de 1965 a 1972, se observaron 138 días lluviosos, de un total de 248 días. La media muestral de las precipitaciones en estos 138 días fue de 13.14 mm, lo que resulta en $\hat{\lambda} = 0.0761$ mm^{-1}. En la Figura 11.4 se nota una concordancia satisfactoria entre $\hat{F}(x)$ y $G(x)$, lo que muestra el acierto de la elección de la Ecuación 4 para modelar la precipitación diaria en Rio de Janeiro. Más adelante, cuando fuera discutida la teoría de pruebas de hipótesis, será explicado como se decide la clasificación de una concordancia como "satisfactoria" o "insatisfactoria". En la Figura 11.5 se presenta la estimación de la densidad de probabilidad de Y, juntamente con el gráfico de frecuencia relativa "escalada" de las 138 observaciones de días lluviosos.

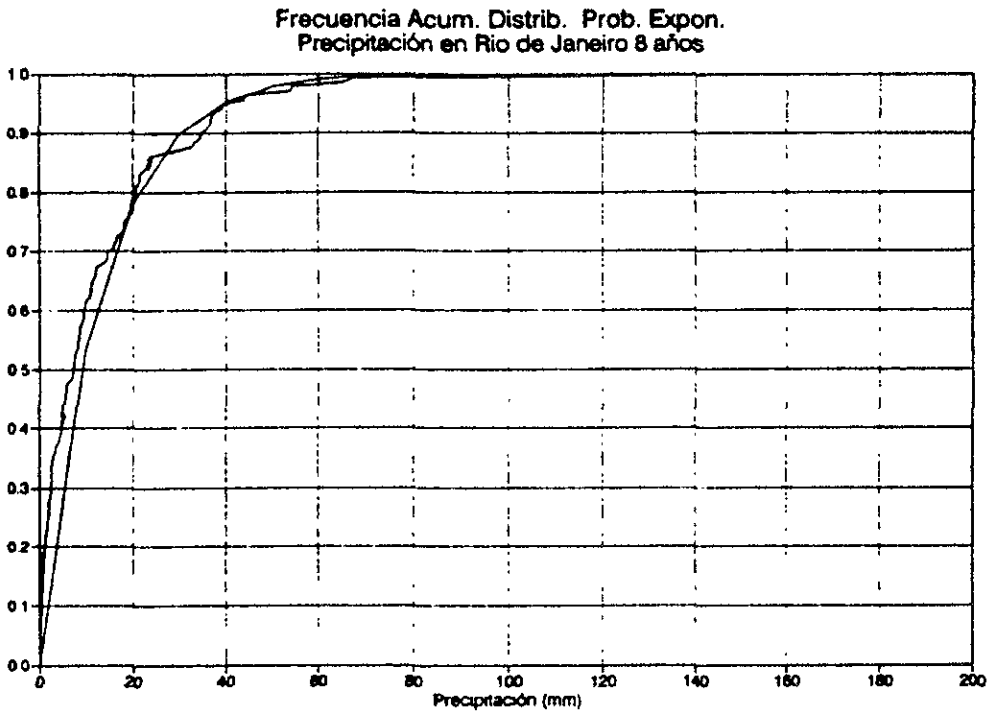


Figura 11.4

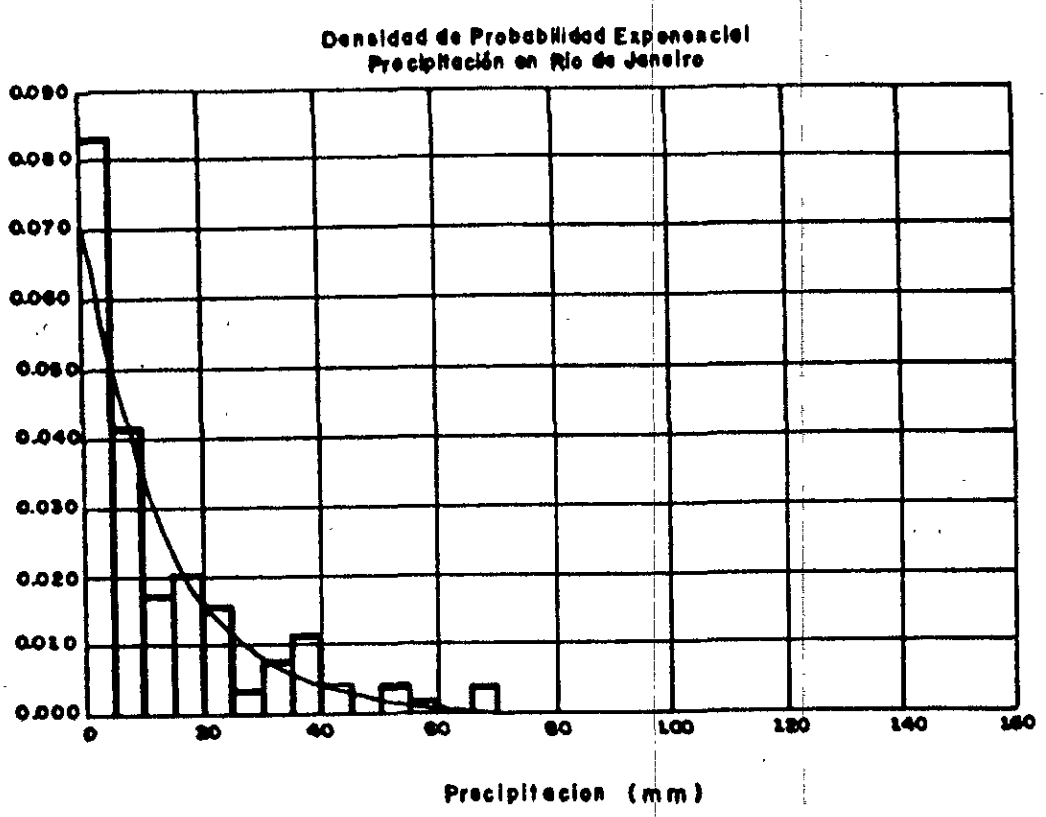


Figura 11.5

Volviendo a la cuestión aún sin respuesta: ¿Cuál función analítica $F(x)$ debe ser escogida para modelar volúmenes anuales afluentes al aprovechamiento hidroeléctrico, visto que la adopción de la Ecuación 4 no fue satisfactoria?

Un importante resultado de la Teoría Estadística ayuda a responder esta pregunta. Se trata del Teorema del Límite Central. Este teorema dice que el resultado de la suma de un gran número de variables aleatorias,

$$X = Y1 + Y2 + Y3 + ... + Yn$$

para n grande, es otra variable aleatoria cuya función densidad de probabilidad está dada, para todos los efectos prácticos, por:

$$f(x) = (1/\ \sigma\sqrt{2\pi})\ \exp\ (-0.5\ (\frac{x-\mu}{\sigma})^2) \qquad , -\infty < x < \infty \qquad (9)$$

Obs.: El dominio de la precipitación fue dividido en intervalos de Δy = 5mm. Si n_i es el número de observaciones en el i-ésimo intervalo, entonces $f_i = \frac{n_i}{n}$ es la frecuencia relativa de observaciones en este intervalo.

La variable aleatoria X que tenga la función de densidad de probabilidad (9) es llamada "normal" o gaussiana, y los parámetros de la función $f(x)$, deben ser estimados en cada caso real.

La Figura 11.6 muestra un gráfico de la ecuación (9) juntamente con el gráfico de frecuencia relativa "escalada" de las 50 observaciones de afluencias al embalse de funil, para $\Delta y = 60 m^3/s$ (Tabla 11.1). Es instructivo comparar la densidad de probabilidad de una variable aleatoria con distribución exponencial, un ejemplo del cual se encuentra en la Figura 11.5, con la densidad de probabilidad de una variable aleatoria con distribución normal, Figura 11.6. En un caso se observa la función tremendamente asimétrica, mientras en el otro ocurre simetría perfecta. El Teorema del Límite Central asegura que aún cuando las parcelas tienen función densidad de probabilidad bastante asimétrica, como en la Figura 11.5, la densidad de probabilidad de la suma va quedando cada vez más simétrica, en dirección a la forma de la Figura 11.6, en la medida en que aumenta el número de parcelas. Casi todas las variables aleatorias relevantes para el planeamiento hidroeléctrico tienen densidad de probabilidad situada entre los dos límites: distribución exponencial y distribución normal.

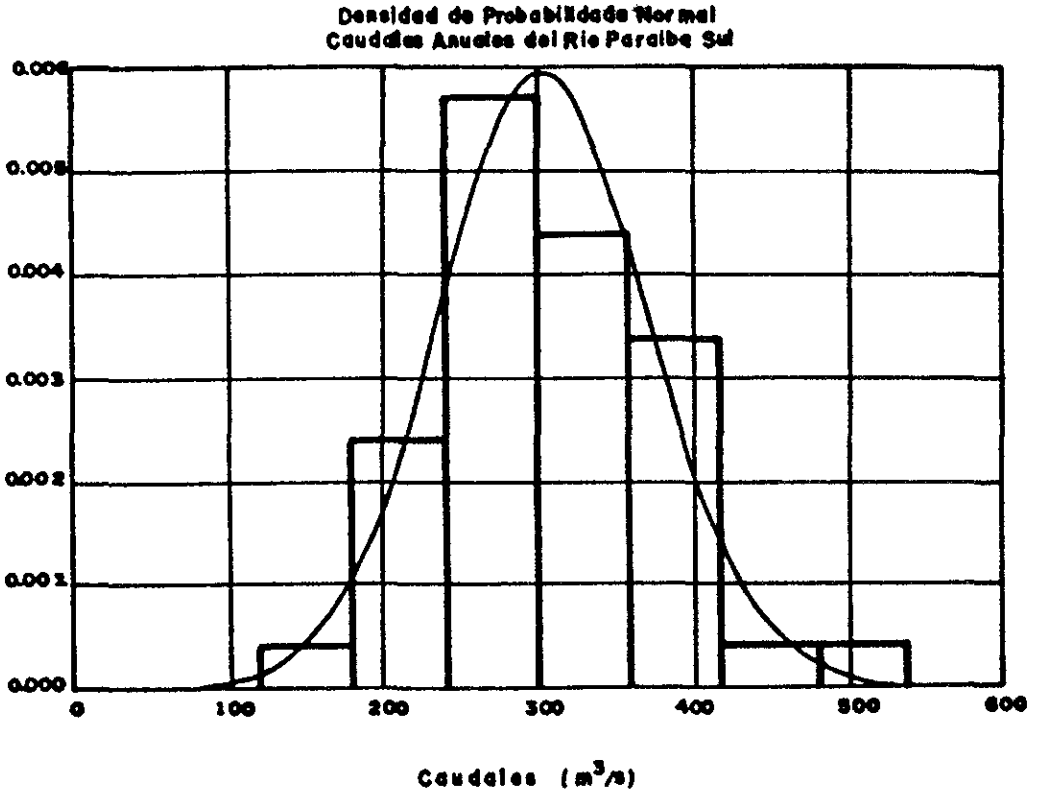


Figura 11.6

Por ejemplo, ya vimos que la distribución exponencial es un modelo satisfactorio para la altura pluviométrica diaria y puntual. Como la precipitación total sobre un área puede ser vista como el resultado de la suma de un gran número de precipitaciones puntuales, cuanto mayor sea esta área, tanto más la distribución de probabilidades deberá alejarse de la "forma exponencial" en dirección a la "forma normal". Si el área de interés fuera todo el Continente Americano, es de suponer que la precipitación diaria total será una variable aleatoria que, para todos los efectos prácticos, tiene distribución normal. Ya la precipitación diaria sobre el área de drenaje contribuyente a una planta hidroeléctrica deberá quedar a medio camino entre la exponencial y la normal. Si consideramos la precipitación sobre el área de drenaje, a lo largo del año, (suma de 365 parcelas), estaremos aproximándonos más aun de la distribución normal.

Es bien posible que la distribución normal sea apropiada en el caso del volumen total afluyente a un embalse, a lo largo de un año, dado que esta variable resulta de la suma a lo largo del tiempo (365 días) y a lo largo del espacio (cuenca de contribución) de las precipitaciones puntuales efectivas, es decir, descontada la evapotranspiración.

Número total de observaciones. Naturalmente, $\sum_i ni = n$ y $\sum_i fi = 1$. La frecuencia relativa "escalada" es $Gi = \frac{fi}{\Delta y}$. Naturalmente, $\Delta y \sum_i gi = 1$.

Como la distribución de probabilidades normal tiene dos parámetros (Ecuación 9), la estimación de los parámetros no puede ser hecha igualando sólo el primer momento muestral (la media aritmética) con el primer momento poblacional (el valor esperado de la variable aleatoria). Es necesario echar mano también de los segundos momentos. Se puede mostrar que, para el caso de la distribución normal, las expresiones paramétricas para los momentos de población son:

$$E[X] = \int_{-\infty}^{\infty} x f(x)\, dx = \mu$$

y

$$E[X^2] = \int_{-\infty}^{\infty} x^2 f(x)\, dx \times f(x)\, dx = \sigma^2 + \mu^2 \qquad (10)$$

Se puede mostrar que $\sigma^2 = E[(X - \mu)^2]$; σ^2 es llamada de varianza de la población y σ es llamada desviación patrón ó desviación "standard".

Para los valores de la Tabla 11.1 ya vimos que la media muestral es 305 m^3/s. La media aritmética de las observaciones elevada al cuadrado (segundo momento muestral) es 97514 $(m^3/s)^2$. Luego tenemos las siguientes estimativas para los parámetros:

$$\hat{\mu} = 305$$
$$\hat{\sigma}^2 = 97514 - 305^2 = 4489 = 67^2 \qquad (11)$$

Se puede ahora re-estimar la probabilidad de que en un año cualquiera del futuro ocurra una afluencia anual inferior a la mínima del histórico:

$$\hat{P}(X<177) = \int_{-\infty}^{177} (1/\ 67\sqrt{2\pi})\ \exp(-0.5\ (\frac{x-305}{67})^2)\ dx \qquad (12)$$

Ocurre que no existe una expresión analítica para la integral anterior. Por otro lado, es común encontrar en diversos libros de texto y "handbooks" una tabla que proporciona el resultado de la integración arriba indicada, que equivale a $F(x)$, para el caso particular en que $\mu = 0$ y $\sigma = 1$, en cuyo caso la variable aleatoria es llamada "normal patrón". Para hacer uso de este recurso, reproducido en la Tabla 11.2, es necesario notar que:

$$P(X \le 177) = F(177) = P(\frac{X - \mu}{\sigma} \le \frac{177 - \mu}{\sigma}) = N(\frac{177 - \mu}{\sigma}) \qquad (13)$$

donde $N(.)$ es la función tabulada. Sustituyendo los valores de los parámetros para el ejemplo en pauta se tiene que:

$$\hat{P}(X \le 177) = N(-1.91) = 0{,}0281\ (\text{ó } 2.81\%) \qquad (14)$$

Aplicándose el mismo procedimiento adoptado en x=177 para otros valores de x, es posible hacer el gráfico de $F(x)$, que aparece en la Figura 11.7, junto con el gráfico de $G(x)$. Se nota ahora una concordancia "satisfactoria".

El lector podrá indagar ¿Cuál alternativa tomaría si el ajuste no fuese "satisfactorio"? Ya tenemos la pista que, en este caso la mejor distribución debería tener una forma a medio camino entre la exponencial y la normal, pero ¿cuál?

En los libros de Estadística son presentadas las diversas distribuciones de probabilidades que "se quedan a medio camino". Para efecto de este texto necesariamente sintético, basta presentar la función densidad de probabilidad de la distribución gama:

$$f(x) = \frac{\lambda k_x k-1}{\Gamma(k)} \exp(-\lambda x) \qquad x \geq 0; \lambda > 0 \qquad (15)$$

donde λ y k son parámetros y $\Gamma(.)$ es la función gama (tabulada). Cuando k es entero, $\Gamma(k) = (k\text{-}1)!$

Es fácil percibir que cuando k=1 la Ecuación 15 colapsa en la Ecuación 5, es decir, la distribución exponencial es un caso particular de la distribución gama. En realidad, se puede mostrar que la función densidad de probabilidad de la variable aleatoria "suma de k parcelas, teniendo cada parcela distribución exponencial" es exactamente la Ecuación 15, aunque k no necesite ser, necesariamente, un entero. Debido al "Teorema del Límite Central", en la medida en que k crezca, la distribución gama se aproximará cada vez más de la distribución normal. Para efecto de estimativa de parámetros, se puede mostrar que:

$$E[X] = k/\lambda \qquad (16)$$

y

$$E[X^2] = k\frac{1+k}{\lambda^2} = \left(\frac{K}{\lambda}\right)^2 + \sigma^2 \quad - \quad \sigma^2 = \frac{K}{\lambda^2}$$

Luego, las estimativas de los parámetros de la distribución gama para los datos de la Tabla 11.1 son:

$$\hat{\lambda} = 0.0679$$

$$\hat{k} = 20.7239$$

Se puede mostrar que:

$$\sigma^2 = E(X^2) - E^2(X) = k\frac{1+k}{\lambda^2} - \frac{K^2}{\lambda^2} = \frac{K}{\lambda^2}$$

11.2 PRUEBAS DE ADHERENCIA

11.2.1 Prueba Chi-Quadrado

Pruebas de adherencia, también llamadas pruebas de adecuación de ajuste, pretenden determinar si una cierta distribución postulada es razonable en presencia de los datos. Por ejemplo, cuando el hidrólogo piensa en adoptar la distribución normal como modelo para describir las afluencias anuales de un río, es lógico probar la adecuación de este procedimiento.

Volvemos nuestra atención más una vez para la Figura 11.6, que servirá como ejemplo de aplicación de la prueba de chi-quadrado. En esta figura se observa que la función densidad de probabilidad parece ajustarse bien al gráfico de las frecuencias relativas escaladas. Pero necesitamos una variable que cuantifique el grado de este ajuste. Para esto debemos dividir el dominio de la variable de interés (caudales, en este ejemplo) en un cierto número de clases. Cada clase debe tener por lo menos cinco observaciones. Suponga que escojamos clases con amplitud de 60 m^3/s, resultando las siguientes frecuencias absolutas y relativas (la frecuencia relativa es la frecuencia absoluta dividida por el número de observaciones, en este caso 50).

Clase	Dominio	Frecuencia Absoluta	Frecuencia Relativa	Probabilidad
			fi	pi
1	X < 240	8	0.16	0.17
2	240 < X < 300	17	0.34	0.31
3	300 < X < 360	13	0.26	0.32
4	360 <X	12	0.24	0.20
Suma		50	1.00	1.00

La última columna es la probabilidad de que la variable aleatoria sea sorteada en el dominio correspondiente a cada clase. Así, con auxilio de la Tabla 11.2 podemos calcular.

$$\Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-u^2/2}\, du = P(Z \leq z)$$

z	0	1	2	3	4	5	6	7	8	9
−3,0	0,001 3	0,001 0	0,000 7	0,000 5	0,000 3	0,000 2	0,000 3	0,000 1	0,000 1	0,000 0
−2,9	0,001 9	0,001 8	0,001 7	0,001 7	0,001 6	0,001 6	0,001 5	0,001 5	0,001 4	0,001 4
−2,8	0,002 6	0,002 5	0,002 4	0,002 3	0,002 3	0,002 2	0,002 1	0,002 1	0,002 0	0,001 9
−2,7	0,003 5	0,003 4	0,003 3	0,003 2	0,003 1	0,003 0	0,002 9	0,002 8	0,002 7	0,002 6
−2,6	0,004 7	0,004 5	0,004 4	0,004 3	0,004 1	0,004 0	0,003 9	0,003 8	0,003 7	0,003 6
−2,5	0,006 2	0,006 0	0,005 9	0,005 7	0,005 5	0,005 4	0,005 2	0,005 1	0,004 9	0,004 8
−2,4	0,008 2	0,008 0	0,007 8	0,007 5	0,007 3	0,007 1	0,006 9	0,006 8	0,006 6	0,006 4
−2,3	0,010 7	0,010 4	0,010 2	0,009 9	0,009 6	0,009 4	0,009 1	0,008 9	0,008 7	0,008 4
−2,2	0,013 9	0,013 6	0,013 2	0,012 9	0,012 6	0,012 2	0,011 9	0,011 6	0,011 3	0,011 0
−2,1	0,017 9	0,017 4	0,017 0	0,016 6	0,016 2	0,015 8	0,015 4	0,015 0	0,014 6	0,014 3
−2,0	0,022 8	0,022 2	0,021 7	0,021 2	0,020 7	0,020 2	0,019 7	0,019 2	0,018 8	0,018 3
−1,9	0,028 7	0,028 1	0,027 4	0,026 8	0,026 2	0,025 6	0,025 0	0,024 4	0,023 8	0,023 3
−1,8	0,035 9	0,035 2	0,034 4	0,033 6	0,032 9	0,032 2	0,031 4	0,030 7	0,030 0	0,029 4
−1,7	0,044 6	0,043 6	0,042 7	0,041 8	0,040 9	0,040 1	0,039 2	0,038 4	0,037 5	0,036 7
−1,6	0,054 8	0,053 7	0,052 6	0,051 6	0,050 5	0,049 5	0,048 5	0,047 5	0,046 5	0,045 5
−1,5	0,066 8	0,065 5	0,064 3	0,063 0	0,061 8	0,060 6	0,059 4	0,058 2	0,057 0	0,055 9
−1,4	0,080 8	0,079 3	0,077 8	0,076 4	0,074 9	0,073 5	0,072 2	0,070 8	0,069 4	0,068 1
−1,3	0,096 8	0,095 1	0,093 4	0,091 8	0,090 1	0,088 5	0,086 9	0,085 3	0,083 8	0,082 3
−1,2	0,115 1	0,113 1	0,111 2	0,109 3	0,107 5	0,105 6	0,103 8	0,102 0	0,100 3	0,098 5
−1,1	0,135 7	0,133 5	0,131 4	0,129 2	0,127 1	0,125 1	0,123 0	0,121 0	0,119 0	0,117 0
−1,0	0,158 7	0,156 2	0,153 9	0,151 5	0,149 2	0,146 9	0,144 6	0,142 3	0,140 1	0,137 9
−0,9	0,184 1	0,181 4	0,178 8	0,176 2	0,173 6	0,171 1	0,168 5	0,166 0	0,163 5	0,161 1
−0,8	0,211 9	0,209 0	0,206 1	0,203 3	0,200 5	0,197 7	0,194 9	0,192 2	0,189 4	0,186 7
−0,7	0,242 0	0,238 9	0,235 8	0,232 7	0,229 7	0,226 6	0,223 6	0,220 6	0,217 7	0,214 8
−0,6	0,274 3	0,270 9	0,267 6	0,264 3	0,261 1	0,257 8	0,254 6	0,251 4	0,248 3	0,245 1
−0,5	0,308 5	0,305 0	0,301 5	0,298 1	0,294 6	0,291 2	0,287 7	0,284 3	0,281 0	0,277 6
−0,4	0,344 6	0,340 9	0,337 2	0,333 6	0,330 0	0,326 4	0,322 8	0,319 2	0,315 6	0,312 1
−0,3	0,382 1	0,378 3	0,374 5	0,370 7	0,366 9	0,363 2	0,359 4	0,355 7	0,352 0	0,348 3
−0,2	0,420 7	0,416 8	0,412 9	0,409 0	0,405 2	0,401 3	0,397 4	0,393 6	0,389 7	0,385 9
−0,1	0,460 2	0,456 2	0,452 2	0,448 3	0,444 3	0,440 4	0,436 4	0,432 5	0,428 6	0,424 7
−0,0	0,500 0	0,496 0	0,492 0	0,488 0	0,484 0	0,480 1	0,476 1	0,472 1	0,468 1	0,464 1

* B. W. Lindgren, *Statistical Theory*, The Macmillan Company, 1960.

$$\Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-u^2/2}\, du = P(Z \leq z)$$

z	0	1	2	3	4	5	6	7	8	9
0,0	0,500 0	0,504 0	0,508 0	0,512 0	0,516 0	0,519 9	0,523 9	0,527 9	0,531 9	0,535 9
0,1	0,539 8	0,543 8	0,547 8	0,551 7	0,555 7	0,559 6	0,563 6	0,567 5	0,571 4	0,575 3
0,2	0,579 3	0,583 2	0,587 1	0,591 0	0,594 8	0,598 7	0,602 6	0,606 4	0,610 3	0,614 1
0,3	0,617 9	0,621 7	0,625 5	0,629 3	0,633 1	0,636 8	0,640 6	0,644 3	0,648 0	0,651 7
0,4	0,655 4	0,659 1	0,662 8	0,666 4	0,670 0	0,673 6	0,677 2	0,680 8	0,684 4	0,687 9
0,5	0,691 5	0,695 0	0,698 5	0,701 9	0,705 4	0,708 8	0,712 3	0,715 7	0,719 0	0,722 4
0,6	0,725 7	0,729 1	0,732 4	0,735 7	0,738 9	0,742 2	0,745 4	0,748 6	0,751 7	0,754 9
0,7	0,758 0	0,761 1	0,764 2	0,767 3	0,770 2	0,773 4	0,776 4	0,779 4	0,782 3	0,785 2
0,8	0,788 1	0,791 0	0,793 9	0,796 7	0,799 5	0,802 3	0,805 1	0,807 8	0,810 6	0,813 3
0,9	0,815 9	0,818 6	0,821 2	0,823 8	0,826 4	0,828 9	0,831 5	0,834 0	0,836 5	0,838 9
1,0	0,841 3	0,843 8	0,846 1	0,848 5	0,850 8	0,853 1	0,855 4	0,857 7	0,859 9	0,862 1
1,1	0,864 3	0,866 5	0,868 6	0,870 8	0,872 9	0,874 9	0,877 0	0,879 0	0,881 0	0,883 0
1,2	0,884 9	0,886 9	0,888 8	0,890 7	0,892 5	0,894 4	0,896 2	0,898 0	0,899 7	0,901 5
1,3	0,903 2	0,904 9	0,906 6	0,908 2	0,909 9	0,911 5	0,913 1	0,914 7	0,916 2	0,917 7
1,4	0,919 2	0,920 7	0,922 2	0,923 6	0,925 1	0,926 5	0,927 8	0,929 2	0,930 6	0,931 9
1,5	0,933 2	0,934 5	0,935 7	0,937 0	0,938 2	0,939 4	0,940 6	0,941 8	0,943 0	0,944 1
1,6	0,945 2	0,946 3	0,947 4	0,948 4	0,949 5	0,950 5	0,951 5	0,952 5	0,953 5	0,954 5
1,7	0,955 4	0,956 4	0,957 3	0,958 2	0,959 1	0,959 9	0,960 8	0,961 6	0,962 5	0,963 3
1,8	0,964 1	0,964 8	0,965 6	0,966 4	0,967 1	0,967 8	0,968 6	0,969 3	0,970 0	0,970 6
1,9	0,971 3	0,971 9	0,972 6	0,973 2	0,973 8	0,974 4	0,975 0	0,975 6	0,976 2	0,976 7
2,0	0,977 2	0,977 8	0,978 3	0,978 8	0,979 3	0,979 8	0,980 3	0,980 8	0,981 2	0,981 7
2,1	0,982 1	0,982 6	0,983 0	0,983 4	0,983 8	0,984 2	0,984 6	0,985 0	0,985 4	0,985 7
2,2	0,986 1	0,986 4	0,986 8	0,987 1	0,987 4	0,987 8	0,988 1	0,988 4	0,988 7	0,989 0
2,3	0,989 3	0,989 6	0,989 8	0,990 1	0,990 4	0,990 6	0,990 9	0,991 1	0,991 3	0,991 6
2,4	0,991 8	0,992 0	0,992 2	0,992 5	0,992 7	0,992 9	0,993 1	0,993 2	0,993 4	0,993 6
2,5	0,993 8	0,994 0	0,994 1	0,994 3	0,994 5	0,994 6	0,994 8	0,994 9	0,995 1	0,995 2
2,6	0,995 3	0,995 5	0,995 6	0,995 7	0,995 9	0,996 0	0,996 1	0,996 2	0,996 3	0,996 4
2,7	0,996 5	0,996 6	0,996 7	0,996 8	0,996 9	0,997 0	0,997 1	0,997 2	0,997 3	0,997 4
2,8	0,997 4	0,997 5	0,997 6	0,997 7	0,997 7	0,997 8	0,997 9	0,997 9	0,998 0	0,998 1
2,9	0,998 1	0,998 2	0,998 2	0,998 3	0,998 4	0,998 4	0,998 5	0,998 5	0,998 6	0,998 6
3,0	0,998 7	0,999 0	0,999 3	0,999 5	0,999 7	0,999 8	0,999 8	0,999 9	0,999 9	1,000 0

Tabla 11.2

p1 = P (X < 240) = N ((240 - 305)/67) = 0.17
p2 = P (240 < X < 300) = N ((300 - 305)/67) - N ((240 - 305)/67) = 0.31

y así sucesivamente.

La medida del ajuste es dado por la suma:

$$c = \sum_{i=1}^{n} \frac{(fi - pi)^2}{pi} \qquad (17)$$

0.20
Cuanto menor sea *c*, mejor será el ajuste. En este ejemplo *c* es igual a ~~0.02~~. Está claro que *c* es una variable aleatoria porque el valor que asume depende de una tabla de observaciones aleatorias. Se puede demonstrar que para cualquier distribución de probabilidad de *X*, la distribución de probabilidades de *c* es aproximadamente chi-quadrado con el único parámetro, llamado "grados de libertad" igual a *m* - *k* - 1, donde:

m = número de clases en que el dominio fue dividido; en este ejemplo, *m* = 4;

k = número de parámetros que fueron estimados; para la distribución normal, *k* = 2.

La Tabla 11.3 reproduce la distribución de probabilidad chi-quadrado para diferentes valores del parámetro "grados de liberdad". Para *m* - *k* - 1 = 4 - 2 - 1 = 1 grado de liberdad nótase que:

$$P\,(C < 3.84) = 0.95$$

Esto significa que si de hecho la distribución de probabilidad escogida fuese correcta, normal para caudales anuales en este ejemplo, entonces la probabilidad de que la medida de ajuste C sea inferior al valor crítico 3.84 es igual a 0.95. Se dice que el nivel de significancia de la prueba de adherencia es de 95%. Como el valor calculado *c* = 0.00 es inferior al valor crítico, decimos que no tenemos razones para rechazar la distribución propuesta (normal, en este caso). Este resultado no nos sorprende porque ya sospechábamos que la distribución normal es un buen modelo matemático para los caudales anuales del Río Paraíba del Sul. Pero, ¿Qué pasa cuando escogimos un mal modelo para el caso bajo examen, como la distribución exponencial?

Vamos a repetir el procedimiento para verificar la adherencia de la distribución exponencial a los datos de caudales anuales. Para esto, calculamos las probabilidades correspondientes a cada clase con auxilio de la ecuación (4), para el valor del parámetro $\hat{\lambda}$ ya calculado e igual a 0.00328. Por consiguiente, tendríamos en este caso que el número de parámetros estimados es $m = 1$ y por lo tanto debemos entrar en la Tabla 11.3 en la segunda línea, correspondiente a 4 - 1 - 1 = 2 grados de libertad, resultando el valor crítico de 5.99, para el mismo nivel de significancia de 95%. Además,

p1 = 1 - e = 0.54
p2 = (1 - e) - (1 - e) = 0.09
p3 = (1 - e) - (1 - e) = 0.06
p4 = 1 - (1 - e) = 0.31

Aplicándose la ecuación 17 para los nuevos valores de p podemos calcular c = 1.64, que aún sea superior al valor de c = 0.20 anteriormente calculado, es también inferior al valor crítico de 5.99.

Este resultado es decepcionante, puesto que sabemos que la distribución exponencial es un mal modelo para los caudales anuales del Río Paraíba del Sul y por lo tanto esperábamos un valor de C superior al valor crítico. O sea, la prueba del chi-quadrado en esto caso fue inútil para indicar que el modelo propuesto es malo. Por esto decimos apenas que "no tenemos elementos para rechazar el modelo propuesto, con base en el resultado de la prueba". No sabemos cual es la probabilidad de estar cometiendo un error al tomar esta decisión. Este error - dejar de rechazar un falso modelo - es llamado "error tipo 2".

Por otro lado, si el valor de C hubiese sido superior al valor crítico, entonces deberíamos rechazar la distribución propuesta. En este caso diríamos que "el modelo debe ser rechazado y la probabilidad de que esta decisión sea equivocada es inferior a 0.05". Este tipo de error, rechazar un modelo verdadero, es llamado "error tipo 1". En pruebas de adherencia en general solamente se cuantifica la probabilidad de cometer error del tipo 1, que es lo complementar del nivel de significancia de la prueba, en esto caso 1 - 0.95 = 0.05.

11.2.2 Prueba de Kolmogorov-Smirnov

Otra prueba de adherencia muy utilizada es la de Kolmogorov-Smirnov, apesar de ser menos general que la prueba chi-quadrado, pues rigorosamente es aplicable solamente para testar la adecuación del ajuste de distribuciones en que los parámetros sean conocidos, en vez de estimados a partir de datos observados. Cuando los parámetros son estimados a partir de los datos, la prueba es apenas aproximada.

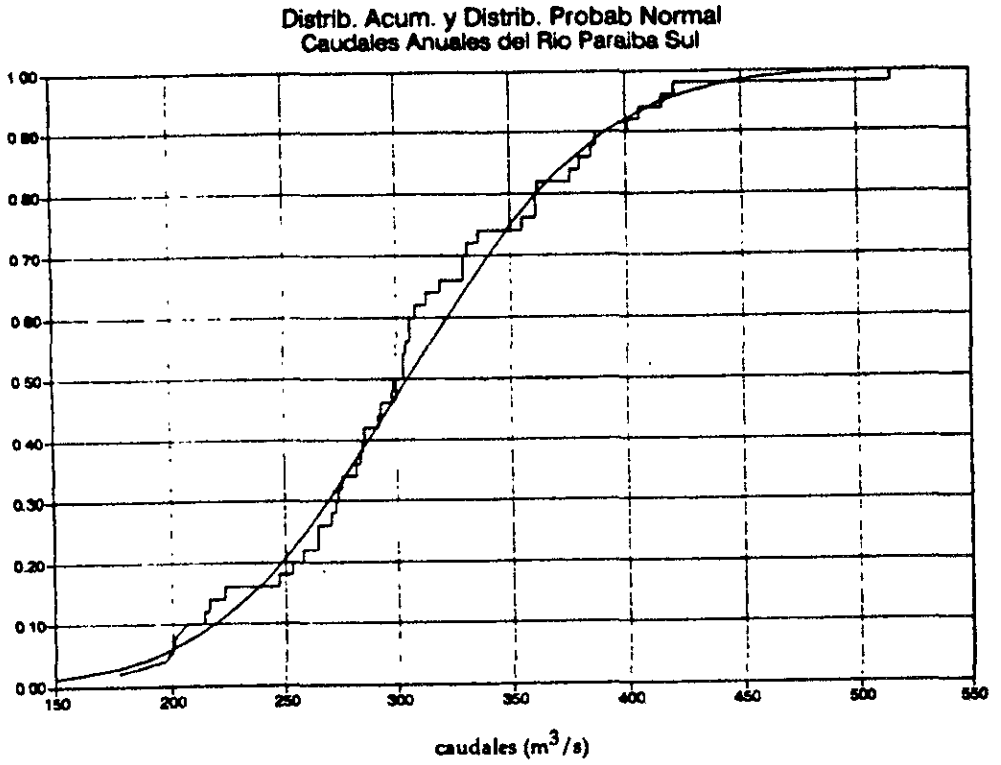


Figura 11.7

La prueba concentra atención en la comparación de las funciones $F(x)$ y $G(x)$. Examinemos la Figura 11.7 en que las dos funciones están graficadas juntas, para el caso en que $F(x)$ fue calculada por la distribución normal ajustada a los caudales anuales del Río Paraíba del Sul. La adherencia entre las dos curvas parece satisfactoria. Una medida de la adherencia entre las dos curvas es el máximo desvío vertical entre ellas. Para este caso, el máximo desvío ocurre para el valor $x = 313$. Consultando la Tabla 11.1, se tiene:

$$G(313) = \frac{\text{Número de observaciones iguales o inferiores a 313}}{50} = \frac{32}{50}$$

$G(313) = 0.64$
Consultando la Tabla 11.2,

$$F(313) = N\left(\frac{313 - 305}{67}\right) = N(0.12) = 0.55$$

Por lo tanto la máxima distancia vertical para las dos curvas es:

$$s = 0.64 - 0.55 = 0.09$$

Vamos a repetir la operación para el caso de la Figura 11.3 en que intentábamos ajustar la distribución exponencial a los caudales anuales del Río Paraíba del Sul, con resultados notoriamente insatisfactorios. En este caso la máxima distancia vertical ocurre para $x = 177$ y se tiene:

$$G(177) = 1/50 = 0.02$$

$$F(177) = 0.44$$

Luego la máxima distancia vertical es en este caso:

$$s = 0.44 - 0.02 = 0.42$$

Naturalmente esta medida de máxima distancia vertical entre las dos curvas $F(x)$ y $G(x)$ es también una variable aleatoria S, en la medida en que $G(x)$ es una función muestral. Si el modelo matemático $F(x)$ es verdadero, $G(x)$ deberá aproximarse de $F(x)$ en la medida en que n, el tamaño de la muestra, crezca. Por lo tanto, el valor crítico para S debe variar con n de forma decreciente. La Tabla 11.3 muestra los valores críticos para el nivel de significancia del 95%. Notase que para $n = 50$, que es el número de observaciones de caudales anuales del Río Paraíba del Sul, el valor crítico es igual a 0.19. Como $0.09 < 0.19$, decimos que "no tenemos elementos para rechazar la distribución normal". Por otro lado, como $0.42 > 0.19$, decimos que "la distribución exponencial debe ser rechazada y la probabilidad de que esta decisión sea equivocada es inferior a 0.05".

Tabla 11.3

Prueba de Kolmogurov - Smirnov para Nivel de Significancia de 95%

n	Valor Crítico
5	0.56
10	0.41
15	0.34
20	0.29
25	0.27
30	0.24
35	0.23
40	0.21
45	0.20
50	0.19
.>50	$\frac{1.36}{\sqrt{n}}$

11.3 CORRELACIÓN Y REGRESIÓN

Un caso particular de gran interés en Hidrología es el de evaluar la relación entre dos variables. Los análisis de tales relaciones son generalmente denominadas de *estudios de correlación y regresión.*

Dados n pares de valores observados $\{(xi, yi); i = 1,2, ...,n\}$, el paso inicial en tales estudios consiste en ubicar eses puntos utilizando ejes cartesianos. Las ordenadas yi son interpretadas como valores asumidos por la variable aleatoria Yi, dado que la variable Xi (no necesariamente aleatoria) es igual a la abcisa xi. Por ejemplo, yi puede ser el total anual de precipitación pluvial en un local de altitud $Xi = xi$ (portanto, Xi acá no es variable aleatoria). Como otro ejemplo, yi puede ser el valor asumido por la afluencia anual de un río en cierta sección, y xi el valor asumido por el vaciamiento anual del mismo río en otra sección (portanto, Xi acá es una variable aleatoria).

A partir de n pares de valores observados, es fácil evaluar el coeficiente de correlación muestral, definido como:

$$n = \frac{\sum_{i=1}^{n} (xi - \bar{x})\,(yi - \bar{y})}{\sqrt{\sum_{i=1}^{n} (xi - \bar{x})^2 \sum_{i=1}^{n} (yi - \bar{y})^2}} = \frac{\sum_{i=1}^{n} xi\,yi - n\bar{x}\bar{y}}{\sqrt{\sum_{i=1}^{n} (xi^2 - n\bar{x})^2 \sum_{i=1}^{n} (yi^2 - n\bar{y})^2}} \qquad (18)$$

donde $\bar{x}$ e $\bar{y}$ son las medias muestrales de X e Y, respectivamente. De la misma forma que $\bar{x}$ e $\bar{y}$ estiman las medias poblacionales μ_x y μ_y, respectivamente, r estima el coeficiente de correlación poblacional ρ. Se puede demostrar que r, bien como ρ asume valores entre -1 y +1. Un gran valor absoluto de r indica fuerte asociación lineal entre X e Y. Un valor negativo de r indica que valores grandes de X están generalmente asociados a valores pequeños de Y.

La *curva de regresión* de la variable aleatoria Y, en términos de la variable X es definida como el valor esperado de la variable Y, conocido el valor de X, $E[Y \mid X = x]$.

Por ejemplo, si X fuera la altitud de cierto local, e Y el total anual de precipitación pluvial, $E[Y \mid X = x]$ es el total anual *medio* (poblacional) de precipitación pluvial correspondiente al local de altitud x.

Caso esa relación sea lineal,

$$E[Y \mid X = x] = \beta_o + \beta_1 x; \qquad (19)$$

la regresión es llamada *regresión lineal simple*. Los parámetros β_o y β_1 son desconocidos, y, por tanto, deben ser estimados con base en los pares de valores observados $\{(xi, yi);\ i = 1,2...,n\}$. Los estimadores basados en el *método de los mínimos cuadrados* son los valores β_o y β_1 que minimizan la siguiente función:

$$\Psi = \sum_{i=1}^{n} (Yi - \beta_o - \beta_1 xi)^2 \qquad (20)$$

Así, β_o y β_1 son la solución del siguiente sistema de ecuaciones:

$$\frac{\partial \Psi}{\partial \beta_o} = (-2) \sum_{i=1}^{n} (Yi - \beta_o - \beta_1 xi) = 0 \qquad (21)$$

$$\frac{\partial \Psi}{\partial \beta_1} = (-2) \sum_{i=1}^{n} ni\,(Yi - \beta_o - \beta_1 xi) = 0 \qquad (22)$$

entonces:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (xi - \bar{x})(yi - \bar{y})}{\sum_{i=1}^{n} (xi - \bar{x})^2} \qquad (23)$$

y

$$\hat{\beta}_o = \bar{y} - \hat{\beta}_1 \bar{x} \qquad (24)$$

de modo que la recta de regresión puede ser estimada por:

$$\hat{E}\,(Y \mid X = x) = \hat{\beta}_o + \hat{\beta}_1\, x \qquad (25)$$

Además de que esta recta corresponde al mínimo valor de la función, como fue impuesto, ella posee dos propiedades interesantes: como atesta la ecuación (24), ella pasa por el punto $(\bar{x},\ \bar{y})$ y, de acuerdo con la ecuación (25), la suma de las distancias verticales entre la recta y los puntos (distancia esa frecuentemente denominada de residuo) es nula.

La recta de regresión es "óptima" solamente en el sentido de que minimiza la suma de los cuadrados de los residuos. Para discutir en detalles la adecuación de tal ajuste, es común separar la llamada *suma total de cuadrados* en dos componentes, *la suma de cuadrados explicada por la regresión y la suma de cuadrados residual*, como sigue:

$$\sum_{i=1}^{n} (yi - \bar{y})^2 = (\sum_{i=1}^{n} \hat{E}(Y \mid X = x) - \bar{y})^2 + \sum_{i=1}^{n} (yi - \hat{E}(Y \mid X = x))^2 \qquad (26)$$

Es importante notar que la suma total de cuadrados ofrece una idea de la variabilidad de variables Y, en la ausencia de los valores *xi* (en efecto, la suma total de cuadrados es proporcional a la varianza de Y), y que la suma de cuadrados residual es simplemente el mínimo valor de la función, definida anteriormente [ecuación 20]. Queda a cargo del lector verificar se la ecuación anterior es realmente una identidad.

La ecuación (25) es muy útil cuando se quiere llenar una información faltante. Suponga por ejemplo que se disponga del siguiente registro de caudales anuales para el Río Celeste, situado cerca del Río Paraíba do Sul:

Ano	Caudal (m^3/s)
50-51	113
51-52	118
53	154
54	60
55	39
56	66
57	88
58	95
59	109
60	87
61	142
62	98
63	84
64	61
65	106
66	124
67	174
67-68	106
68-69	49
69-70	104
70-71	60
71-72	130
72-73	70

Observese que el Río Celeste y el Río Paraíba do Sul tienen en común el período 50-51 hasta 70-71, 21 años. Además el Río Celeste tiene 2 años de datos que no están disponibles en el Río Paraíba do Sul.

Vamos utilizar los datos del Río Celeste para llenar la laguna de información en el Río Paraíba del Sul. Por lo tanto los datos del Río Celeste serán llamados X y los del Río Paraíba do Sul Y. Aplicandose las ecuaciones (18), (23) y (24) obténse:

$$r = 0.96$$
$$\hat{\beta}_o = 70.86$$
$$\hat{\beta}_1 = 2.36$$

Por lo tanto la estimativa para el caudal en el Río Paraíba del Sul para el año de 71-72 es 70.86 + 2.36 (130) = 377 m^3/s y para el año 72-73 es 70.86 + 2.36 (70) = 236 m^3/s.