PROBABILIDADE E ESTATÍSTICA APLICADAS À HIDROLOGIA

Mauro Naghettini

Maria Manuela Portela

DECivil, IST, 2011

PROBABILIDADE E ESTATÍSTICA APLICADAS À HIDROLOGIA

Mauro Naghettini

Professor Associado, Escola de Engenharia da Universidade Federal de Minas Gerais, Belo Horizonte, Brasil.

Maria Manuela Portela

Professora Auxiliar, Instituto Superior Técnico da Universidade Técnica de Lisboa, Portugal.

(Nota: o presente texto foi produzido a partir de capítulo homónimo do livro *Hidrologia Aplicada*, a ser publicado entre 2011 e 2012 pela Associação Brasileira de Recursos Hídricos, ABRH. O intuito é o de proporcionar noções fundamentais de probabilidades e estatística aplicadas à hidrologia, incluindo conceitos relacionados com a análise de incertezas)

Índice do texto

Pág.

Índice de Tabelas

Índice de Figuras

Índice de Exercícios

Pág.

1. Introdução

Os fenómenos naturais, nomeadamente, hidrológicos contêm *incertezas* que lhes são inerentes sendo que existem duas fontes para tais incertezas: (i) a *aleatoriedade natural* associada às possíveis ocorrências (ou *realizações*) de um certo fenómeno; e (ii) e as imperfeições e/ou insuficiências do conhecimento humano sobre os processos que determinam tais ocorrências. As incertezas do primeiro tipo – ou aleatórias – podem ser expressas em termos da maior ou menor variabilidade de uma ou mais das *variáveis* (ou grandezas mensuráveis) associadas ao fenómeno em estudo. As incertezas do segundo tipo resultam da interpretação imperfeita ou imprecisa da realidade subjacente ao referido fenómeno, por parte dos modelos teóricos e/ou físicos utilizados para o caracterizar.

As incertezas aleatórias não podem ser reduzidas ou modificadas porque são intrínsecas à variabilidade dos fenómenos em observação. Em geral, essas incertezas apenas podem ser *parcialmente estimadas* pelo padrão da variabilidade exibido pelas amostras referentes a realizações desses fenómenos ou das variáveis que nele intervêm. Já as incertezas que decorrem das limitações do conhecimento humano acerca dos mencionados fenómenos podem ser reduzidas, seja pela obtenção de dados e de informação adicionais, seja pela especificação de novos modelos teóricos (ou físicos) mais conformes com a realidade. Em ambos os casos, os conceitos e métodos da teoria de probabilidades e da estatística constituem conhecimentos indispensáveis para lidar com as incertezas e para as interpretar (*Ang e Tang*, 2007).

As consequências que as incertezas acarretam no projecto e no planeamento de estruturas e sistemas de engenharia, em geral, e de engenharia de recursos hídricos, com particular ênfase, são muito importantes. De facto, num contexto de incerteza, o projecto e o planeamento de estruturas e sistemas de aproveitamento e de controlo de recursos hídricos envolvem *riscos*, os quais envolvem *probabilidades* de ocorrência de certos acontecimentos críticos e das suas respectivas *consequências*, e, finalmente, a formulação de processos de *tomada de decisões*. De modo ideal, a tomada de uma decisão, por exemplo, quanto às dimensões do descarregador de superfície de uma barragem, deveria levar em consideração: (i) a probabilidade de que, ao longo da vida útil do empreendimento, o caudal máximo para o qual foi projectado seja ultrapassado pelas caudais de cheia que efectivamente se constate ser necessário descarregar; (ii) as possíveis consequências da eventual subestimação do caudal de projecto; e (iii) a formulação de planos de tomada de decisões assentes em soluções de compromisso entre avaliações quantitativas dos riscos, custos e benefícios das diversas soluções alternativas estudadas.

Assim, num quadro completo e racional de tomada de decisões relacionadas com o projecto e o planeamento de infra-estruturas e de sistemas de recursos hídricos, é preciso levar em consideração as incertezas associadas aos fenómenos hidrológicos intervenientes. A teoria de probabilidades e a estatística constituem um campo de saber e fornecem ferramentas adequadas para interpretar as características de alguns desses fenómenos e para equacionar parte da incerteza que lhes possa estar associada.

No presente documento sistematizaram-se alguns dos conceitos daquela teoria mais relevantes e frequentemente intervenientes em estudos do âmbito da engenharia dos recursos hídricos, com ênfase para a hidrologia. Pretendendo-se que se trate de um documento didáctico, foram incluídos exemplos e exercícios de aplicação de modo a tornar mais explícitos aqueles conceitos.

2. Caracterização preliminar das incertezas presentes nos fenómenos hidrológicos

As ocorrências de muitos dos fenómenos relevantes no âmbito da engenharia dos recursos hídricos, incluindo a componente de hidrologia, contêm incertezas aleatórias, que não podem ser previstas com absoluta precisão. Em geral, esses fenómenos são caracterizados por uma ou mais variáveis mensuráveis na natureza (ou em laboratório), de modo normalizado e sistemático. Sob as mesmas condições de observação, os dados ou registos de uma mesma variável podem apresentar valores muito diferenciados entre si, alguns com menor frequência e outros com maior. A variabilidade dos dados apresenta um certo padrão, o qual exemplifica apenas uma realização ou *amostra* da variação intrínseca do fenómeno natural a que se referem tais dados.

Considere a amostra de precipitações diárias máximas anuais, Pdma, apresentadas na Tabela 1, relativa ao posto udométrico de Pavia (20I/01G) (localizado na bacia hidrográfica do rio Tejo) no período de 94 anos hidrológicos, entre 1911/12 e 2004/05. Recorda-se que tal amostra é constituída por um valor por ano hidrológico, a máxima precipitação em 24 h em cada ano. Como é do conhecimento geral, em Portugal o ano hidrológico decorre entre 1 de Outubro e 30 de Setembro.

Tabela 1 – Precipitações diárias máximas anuais, Pdma, no posto udométrico de Pavia (20I/01G), na bacia hidrográfica do rio Tejo, no período de 94 anos hidrológicos, entre 1911/12 e 2004/05.

Ano hidrológico	Pdma (mm)	Ano hidrológico	Pdma (mm)	Ano hidrológico	Pdma (mm)	Ano hidrológico	Pdma (mm)	Ano hidrológico	Pdma (mm)
1911/12	24.2	1930/31	15.3	1949/50	43.8	1968/69	43.7	1987/88	27.0
1912/13	31.3	1931/32	40.2	1950/51	58.2	1969/70	36.2	1988/89	58.0
1913/14	32.5	1932/33	20.4	1951/52	34.6	1970/71	29.8	1989/90	27.8
1914/15	33.5	1933/34	20.2	1952/53	40.2	19710/72	60.2	1990/91	37.5
1915/16	20.2	1934/35	32.8	1953/54	20.8	1972/73	28.0	1991/92	35.2
1916/17	38.2	1935/36	43.2	1954/55	69.0	1973/74	31.4	1992/93	27.5
1917/18	36.7	1936/37	29.8	1955/56	44.0	1974/75	38.4	1993/94	28.5
1918/19	35.2	1937/38	42.8	1956/57	27.2	1975/76	29.4	1994/95	52.0
1919/20	92.3	1938/39	45.0	1957/58	37.2	1976/77	34.0	1995/96	56.8
1920/21	30.0	1939/40	34.2	1958/59	36.7	1977/78	47.0	1996/97	80.0
1921/22	25.2	1940/41	32.8	1959/60	49.0	1978/79	57.0	1997/98	29.0
1922/23	50.4	1941/42	46.3	1960/61	38.9	1979/80	36.5	1998/99	55.2
1923/24	35.7	1942/43	31.9	1961/62	59.6	1980/81	84.2	1999/00	48.4
1924/25	40.5	1943/44	34.2	1962/63	63.3	1981/82	45.0	2000/01	33.2
1925/26	10.3	1944/45	24.3	1963/64	41.2	1982/83	95.5	2001/02	27.4
1926/27	40.2	1945/46	71.4	1964/65	46.6	1983/84	48.5	2002/03	27.4
1927/28	8.1	1946/47	37.4	1965/66	84.2	1984/85	38.0	2003/04	18.2
1928/29	10.2	1947/48	31.4	1966/67	29.5	1985/86	38.6	2004/05	34.2
1929/30	14.2	1948/49	24.3	1967/68	70.2	1986/87	26.0		

O *padrão de variabilidade* temporal das precipitações diárias máximas anuais apresentadas na anterior tabela pode ser visualizado pelo *diagrama de série temporal* ou *diagrama cronológico* da Figura 1 (a) e, de forma mais elaborada, pelo *histograma* da Figura 1 (b).

Para construir o histograma da Figura 1(b) obtiveram-se as ocorrências ou as *frequências absolutas* com que os sucessivos valores da precipitação estão compreendidos entre os limites de diferentes *intervalos de classe* para o que foram consideradas classes com amplitude de 12.5 mm. O resultado, em cada classe, do quociente entre a correspondente frequência absoluta e o número total de valores da amostra ou *dimensão da amostra*, N, a saber no exemplo da Figura 1,

N=94, é a *frequência relativa* nesse intervalo de classe (eixo principal das ordenadas no diagrama do lado direito), que, na figura, foi expressa em percentagem. Para fixar o número de intervalos de classe do histograma adoptou-se a *regra de Sturges*, ou seja, $\mathrm{NIC} = 1 + 3.3\log_{10}(\mathrm{N})$, na qual NIC denota o número recomendado daqueles intervalos e N tem o significado antes explicitado.

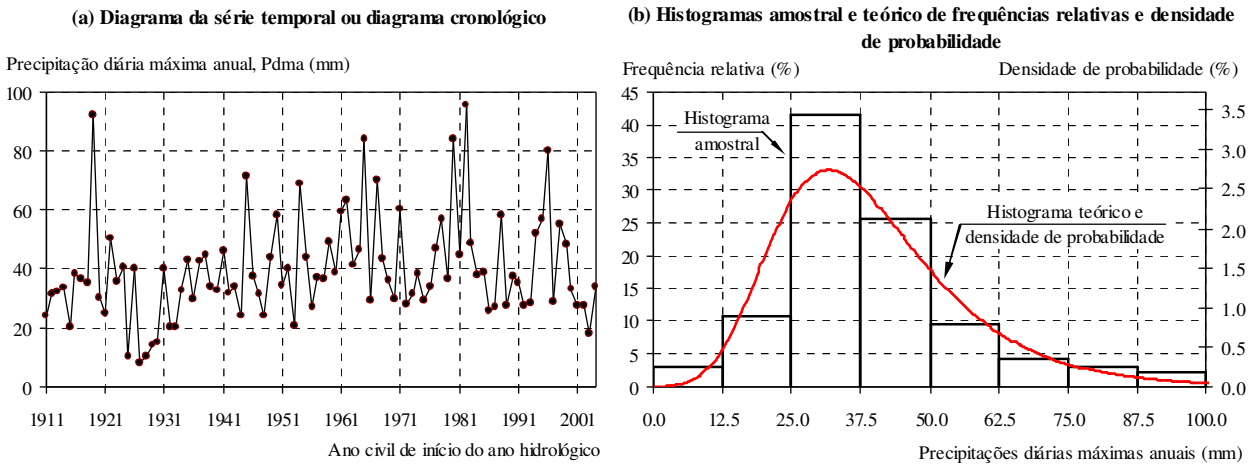

Figura 1 – Variabilidade temporal das precipitações diárias máximas anuais (mm) no posto udométrico de Pavia (20I/01G), na bacia hidrográfica do rio Tejo, no período de 94 anos hidrológicos, entre 1911/12 e 2004/05.

Suponha-se agora que, tendo em vista um problema de análise de cheias, se pretendia estimar o caudal de ponta de cheia para a precipitação diária máxima anual de 103 mm, superior a qualquer valor da amostra da Tabela 1. Com base *unicamente* nessa amostra, poder-se-ia concluir que, *não tendo ocorrido no passado um valor dessa ordem de grandeza, seria improvável que o mesmo se realizasse no futuro*, especialmente estando-se em presença de uma amostra consideravelmente longa. Em contrapartida, poder-se-ia admitir que, não obstante esta última constatação, se a amostra tivesse maior dimensão ou se respeitasse a outro intervalo de tempo, eventualmente conteria valores iguais ou mesmo superiores a 103 mm.

Para averiguar se poderão ou não ocorrer valores para além dos contidos numa dada amostra é necessário obter, de algum modo, o *padrão completo de variabilidade* da variável a que se refere essa amostra (ou seja, o histograma de um número infinito de observações da mesma) através de um função teórica de *distribuição de probabilidade* ou, de modo equivalente, da correspondente função teórica *de densidade de probabilidade*, para o que é necessário estabelecer os modelos matemáticos que exprimem essas funções, com estimação, a partir da amostra, dos respectivos *parâmetros*.

Um exemplo de uma dessas funções, no caso em menção, referente à lei de Gumbel de dois parâmetros (objecto do **item 4**), está indicado na Figura 1(b) pela curva a vermelho que, lida em correspondência com o eixo secundário das ordenadas (eixo de densidade de probabilidade), representa a função densidade de probabilidade de tal lei. A mesma curva lida em correspondência com o eixo principal das ordenadas (eixo de frequência relativa) traduz o histograma teórico, também de acordo com a mencionada lei.

Embora o estudo e o ajuste de modelos paramétricos sejam tratados apenas em itens subsequentes, anota-se, desde já, que a probabilidade de ocorrer uma precipitação diária máxima anual superior a 103 mm segundo a lei de Gumbel com parâmetros estimados a partir da amostra apresentada na Tabela 1, é de 0.5%, ou seja, embora pequena, não é nula. A anterior probabilidade pode ser entendida como significando que, em média, nos próximos 200 anos, poderá ocorrer uma dessas precipitações em um ano qualquer.

Poder-se-ia dar o caso de o critério de projecto requerer uma precipitação mais excepcional, por exemplo, susceptível de ocorrer em qualquer um dos próximos 1000 anos. Uma precipitação de projecto tão elevada asseguraria condições de dimensionamento certamente mais robustas. Contudo, convém sublinhar, que, por regra, a decisão de adoptar um critério de projecto mais excepcional implica, por um lado, maiores custos de construção e, por outro lado, risco de falha ou mesmo de colapso menor. A opção por um dado valor de projecto, para além de reflectir eventuais condicionalismos legais (tais como normas ou regulamentos), deve decorrer de uma análise de *custos/benefícios e riscos*, avaliados tendo em conta o horizonte da vida útil esperada para a estrutura hidráulica em cujo dimensionamento intervém, a par com as consequências da falha/colapso dessa estrutura.

Um processo complementar para caracterizar de modo sintético a variabilidade de uma série temporal de uma variável hidrológica, como a apresentada na Tabela 1, utiliza as designadas *estatísticas amostrais* ou *estatísticas descritivas* que não são mais do que *medidas numéricas*, calculadas a partir da amostra, que "descrevem" as características essenciais do histograma, tais como a abcissa de seu centro geométrico, a dispersão com que os pontos amostrais se distribuem em torno do valor central e a eventual assimetria entre as caudas inferior e superior do diagrama.

A Tabela 2 contém o resumo das principais estatísticas amostrais, as fórmulas de cálculo dessas estatísticas e, especificamente para a amostra de precipitações diárias máximas anuais da Tabela 1, os respectivos valores numéricos. Explicitam-se, ainda, os significados das estatísticas enquanto descritores da forma do histograma.

As principais medidas de *tendência central* são a *média*, a *moda* e a *mediana*. A primeira corresponde à abcissa do centro geométrico do histograma, enquanto a moda é o valor mais frequente da amostra e é dada pela abcissa da maior ordenada do *polígono de frequências*. Este polígono é formado pela junção dos pontos médios dos topos dos rectângulos que constituem o histograma, para o que é necessário considerar duas classes adicionais, uma em cada extremidade, ambas com ordenadas nulas. Por sua vez, a mediana de uma amostra classificada por ordem crescente – $\{x_{(1)}, x_{(2)}, \ldots, x_{(N)}\}$ tal que $x_{(i)}$ é inferior ou igual a $x_{(i+1)}$ – corresponde ao elemento de ordem (N+1)/2, se N é ímpar, ou à média aritmética entre os elementos de ordens (N/2) e [(N/2)+1], se N é par.

Uma das principais *medidas de dispersão* é a *variância*, a qual é dada pela média dos quadrados das diferenças entre os elementos amostrais e a respectiva média, multiplicada pelo factor N/(N-1) para corrigir o chamado *viés*. A raiz quadrada da variância é o *desvio-padrão*, sendo que o quociente entre este desvio e a média recebe a designação de *coeficiente de variação*, grandeza adimensional muito útil para comparar as dispersões relativas de diferentes variáveis.

Outra grandeza adimensional de grande utilidade para a análise estatística de variáveis hidrológicas é o *coeficiente de assimetria*, calculado conforme também indicado na Tabela 2. Relativamente a tal coeficiente, anota-se que, no caso de acontecimentos hidrológicos extremos, a soma das diferenças cúbicas entre os elementos da amostra e a respectiva média é

frequentemente positiva, em consequência de os valores mais elevados estarem muito mais afastados da média do que os valores que lhe são inferiores. Como estão em causa diferenças ao cubo, resulta um coeficiente de assimetria positivo. É este o caso do histograma da Figura 1 (b) e de tantos outros histogramas de amostras de variáveis hidrológicas, o que torna necessário o estudo de distribuições de probabilidade capazes de reproduzir essa assimetria, como, por exemplo a de Gumbel a que se refere a curva de densidade de probabilidade representada naquela figura. Contudo, pode dar-se o caso de uma amostra exibir um coeficiente de assimetria, quer nulo, sendo o correspondente histograma simétrico, quer negativo, traduzido, neste caso, por uma cauda inferior do histograma relativamente mais prolongada/estendida do que a cauda superior.

Tabela 2 – Principais estatísticas amostrais ou descritivas, respectivas fórmulas de cálculo, significados e valores tendo por base a amostra de precipitações diárias máximas anuais da Tabela 1.

Designação	Tipo	Notação	Fórmula cálculo ou conceito	Interpretação	Valor para a amostra da Tabela 1
Média	Tendência central	$\overline{X}$	$\overline{X} = \frac{1}{N}\sum_{i=1}^{N} x_i$	Abcissa do centro geométrico do histograma	39.5 mm
Moda	Tendência central	X_{MO}	Elemento da amostra com maior frequência	Abcissa da maior ordenada do polígono de frequências	40.2 mm
Mediana ou 2º quartil	Tendência central	X_{MD} ou Q_2	50% dos valores ordenados abaixo e 50 % acima	Abcissa que divide ao meio a área do histograma	36.4 mm
1º quartil	Cauda inferior	Q_1	Mediana dos 50% menores valores	Abcissa que divide em 25-75% a área do histograma	34.2 mm
3º quartil	Cauda superior	Q_3	Mediana dos 50% maiores valores	Abcissa que divide em 75-25% a área do histograma	38.4 mm
Amplitude interquartis	Dispersão	AIQ	$AIQ = Q_3 - Q_1$	Amplitude entre as abscissas Q_3 e Q_1	4.2 mm
Momento central de ordem r	-	m'_r	$m'_r = \frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{X})^r$	Potência r da média dos desvios em relação à média	-
Variância	Dispersão	S_X^2	$S_X^2 = \frac{N}{N-1} m'_2$	Média dos desvios quadráticos, em relação à média	295.9 mm^2
Desvio--padrão	Dispersão	S_X	$S_X = \sqrt{S_X^2}$	Raiz quadrada do desvio quadrático médio	17.2 mm
Coeficiente de variação	Dispersão	CV	$CV = \frac{S_X}{\overline{X}}$	Desvio-padrão expresso em fracção da média	0.436
Coeficiente de assimetria	Assimetria	g	$g = \frac{N^2 m'_3}{(N-1)(N-2)(S_X)^3}$	Coeficiente adimensional	1.149
Coeficiente de curtose	Curtose	k	$k = \frac{(N+1)\,N^2\,m'_4}{(N-1)(N-2)(N-3)(S_X)^4} - \frac{3(N+1)^2}{(N-2)(N-3)}$	Coeficiente adimensional (achatamento)	1.699

Em complemento dos elementos precedentes referentes à análise preliminar de dados hidrológicos, recomenda-se a consulta do capítulo 2 do livro de *Naghettini e Pinto* (2007), sendo que tal livro se encontra disponível na sua versão completa, mediante acesso à seguinte URL: http://www.cprm.gov.br/publique/cgi/cgilua.exe/sys/start.htm?infoid=981&sid=36.

A prática profissional associada à engenharia dos recursos hídricos exige a formulação de modelos matemáticos com o objectivo de representar/caracterizar os processos físicos e, assim, possibilitar a tomada de decisões, por exemplo, quanto ao planeamento e ao projecto dos sistemas para aproveitamento e/ou controlo das disponibilidades hídricas de superfície. No essencial, tais modelos podem ser determinísticos e não determinísticos, sendo que, naquele primeiro tipo se incluem os modelos empíricos e os fisicamente baseados, e, no segundo tipo, os modelos probabilísticos e os estocásticos, *Quintela e Portela (2002).*

Uma vez que os modelos são representações imperfeitas e aproximadas da realidade, as estimativas e as previsões a que conduzem estão necessariamente sujeitas a imprecisões e, portanto, contêm incertezas. Como antes mencionado, essas incertezas decorrem da insuficiente monitorização e/ou conhecimento associado ao processo físico em causa e, sempre que possível, devem se consideradas em simultâneo com as incertezas aleatórias, intrínsecas do processo, para assegurar uma completa caracterização das incertezas e das suas implicações nos actos de tomada de decisões de engenharia (*Ang e Tang*, 2007). Algumas dessas incertezas podem ser reduzidas pela aquisição de dados adicionais e/ou pela formulação de modelos alternativos, expectavelmente mais aptos a representar o fenómeno em estudo.

Ao pretender-se caracterizar as precipitações diárias máximas anuais no posto de Pavia (20I/01G) a que se refere a Tabela 1 mediante adopção da lei de probabilidades de Gumbel, conforme antes considerado, introduz-se, necessariamente uma simplificação na interpretação do processo natural que produz tais precipitações que, porventura, poderiam ser melhor descritas por uma outra função de distribuição de probabilidade ou mesmo por uma combinação de várias dessas funções. Mesmo que a distribuição de Gumbel constituísse a verdadeira síntese matemática do processo físico conducente àquelas precipitações, tal distribuição possui *parâmetros*, cujas *estimativas* são obtidas a partir de uma amostra com dimensão sempre muito limitada face à infinitude do universo de onde provém, pelo que aqueles parâmetros necessariamente diferem dos verdadeiros, embora desconhecidos, parâmetros do universo.

Em consequência das anteriores incertezas, ao afirmar-se que à precipitação diária máxima anual de 103 mm (ou seja, ao *quantil* de 103 mm) está associada a *probabilidade de excedência* de 0.5%, está simplesmente a falar-se de um *valor esperado*, ou seja, de um valor médio em torno do qual se pode construir um *intervalo* de valores que conterá o verdadeiro e desconhecido valor do *quantil*, com uma certa *confiança*, por exemplo, de 95%. A inclusão destas e de outras incertezas na prática da engenharia de recursos hídricos requer alguns fundamentos da teoria de probabilidades e estatística que a seguir se descrevem.

3. Definições básicas

3.1. Nota prévia

Apresentam-se a seguir algumas definições básicas e os principais fundamentos que enquadram as aplicações da teoria de probabilidades e estatística à hidrologia.

3.2. Espaço de resultados ou espaço amostral

O *espaço de resultados* ou *espaço amostral* é o conjunto de todos os resultados elementares, mutuamente exclusivos e colectivamente exaustivos de uma experiência aleatória. Em geral, denota-se esse conjunto por Ω distinguindo-se entre espaços *numeráveis* e *não numeráveis* e entre espaços *finitos* e *infinitos.* Um acontecimento é um qualquer subconjunto do espaço amostral.

Exemplos:

(i) Ω_1:{número de dias chuvosos num ano}≡{ 0, 1, 2, ... , 365} → espaço amostral numerável e finito;

(ii) Ω_2:{número de dias consecutivos sem chuva}≡{ 0, 1, 2, ... } → espaço amostral numerável e infinito;

(iii) Ω_3:{precipitação diária máxima anual no posto udométrico de Pavia ≡{P; $P \in \mathbf{R}_+$} → espaço amostral não numerável e infinito.

3.3. Acontecimento aleatório

Um acontecimento aleatório é uma situação específica que se pretende que ocorra cada vez que se realiza uma experiência aleatória. Um acontecimento aleatório pode ser um elemento ou um subconjunto do espaço amostral Ω.

Exemplos:

(i) A:{média da precipitação nos dias com chuva no posto udométrico de Pavia (20I/01G) no ano hidrológico de 1916/17};

(ii) B:{número anual de dias com chuva no posto udométrico de Pavia (20I/01G) durante a década de 1980 a 1990}.

3.4. Complementar de um acontecimento aleatório

O complementar, E^c, de um acontecimento aleatório, E, é o acontecimento que ocorre quando não ocorre E. O complementar é, portanto, o conjunto formado por todos os elementos pertencentes a Ω e que não pertencem a E.

Exemplo:

Se a experiência aleatória consistisse na contagem do número anual de dias com chuva no posto udométrico de Pavia a que se refere a Tabela 1 e se, para o ano hidrológico de 1916/17, resultasse no evento de 82 dias com chuva, ter-se-ia E^c:{0, 1, 2, ... , 80, 81, 83, 84, ... , 365}.

3.5. Combinação de acontecimentos aleatórios. União e intersecção

- ***União***

A união de dois acontecimentos A e B, representada por A∪B, é o conjunto formado pelos elementos pertencentes a A ou a B ou a ambos. Por exemplo, se A se refere aos anos em que, em dada estação hidrométrica, ocorreram caudais instantâneos superiores a 80 m^3/s e B aos anos em que a máxima precipitação diária num posto udométrico situado na bacia hidrográfica daquela estação hidrométrica foi superior a 40 mm, então A∪B representa os elementos de A ou B ou de ambos.

- ***Intersecção***

A intersecção de dois acontecimentos A e B, representada por A∩B, é o conjunto formado pelos elementos que simultaneamente pertencem a A e a B. No exemplo anterior, a intersecção de A com B designa os anos em que simultaneamente ocorreram caudais instantâneos superiores a 80 m^3/s e máximas precipitações diárias superiores a 40 mm. Se a intersecção de A com B é um conjunto vazio, ou seja, se A∩B=∅, então os acontecimentos não ocorrem simultaneamente, recebendo a designação de *acontecimentos mutuamente exclusivos*, *incompatíveis* ou *disjuntos*. Qualquer acontecimento e o seu complementar, A e A^c, constituem exemplos de acontecimentos disjuntos.

3.6. Probabilidade

Uma vez definidos o espaço amostral e os acontecimentos aleatórios, pode associar-se uma *probabilidade* a cada um desses acontecimentos, podendo entender-se por tal uma medida relativa da sua possibilidade de ocorrer, compreendida entre os valores extremos de 0 (*impossibilidade* de ocorrência ou *acontecimento impossível*) e de 1 (*certeza* de ocorrência ou *acontecimento certo).*

Segundo a definição mais usual, a probabilidade de um acontecimento A de um espaço amostral Ω, P(A), é um número não negativo que deve satisfazer os seguintes *axiomas*:

(a) $0 \leq P(A) \leq 1$;

(b) $P(\Omega)=1$; e

(c) para qualquer sequência de acontecimentos mutuamente exclusivos E_1, E_2, ... E_∞, a probabilidade da união desses acontecimentos é igual à soma das respectivas probabilidades individuais, ou seja, $P\left(\bigcup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} P(E_i)$.

Dos anteriores axiomas, decorrem os seguintes corolários:

- $P(A^c)=1-P(A)$
- $P(\emptyset)=0$
- Se A e B são dois acontecimentos do espaço amostral Ω e $A \subset B$, então $P(A) \leq P(B)$.
- *Desigualdade de Boole* (ou limite da união): se A_1, A_2, ... , A_k são acontecimentos definidos num espaço amostral, então, $P\left(\bigcup_{i=1}^{\infty} A_i\right) \leq \sum_{i=1}^{\infty} P(A_i)$.

- *Regra da adição de probabilidades*: se A e B são dois acontecimentos do espaço amostral Ω, então, $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

3.7. Dependência e independência estatísticas

Um acontecimento A *depende estatisticamente* de B se o facto de B ocorrer altera a probabilidade de A ocorrer. Neste caso, a probabilidade de que o acontecimento A ocorra, dado que o acontecimento B ocorreu, é referida como *probabilidade condicional* de A dado B e denotada por $P(A|B)$. Em termos formais, é calculada por $P(A|B) = P(A \cap B)/P(B)$. Ao contrário, se a probabilidade de ocorrência do acontecimento A não é afectada pela ocorrência de B, ou seja, se $P(A|B) = P(A)$, então A é dito *estatisticamente independente* de B sendo a probabilidade da ocorrência simultânea dos acontecimentos A e B dada por $P(A \cap B)=P(A).P(B)$.

Exercício 1 – Considera-se que dois acontecimentos naturais podem produzir a ruptura de uma dada barragem situada numa região pouco monitorizada do ponto de vista hidrológico e sujeita a tremores de terra: a ocorrência de um caudal de ponta de cheia superior ao caudal de projecto do descarregador de superfície (acontecimento A) e o colapso estrutural devido a um tremor de terra (acontecimento B). Admitindo que as probabilidades anuais dos anteriores acontecimentos são, respectivamente, P(A)=0.02 e que P(B)=0.01, estime a probabilidade da barragem romper num ano qualquer.

Solução: A ruptura da barragem pode ser devida a uma cheia, a um tremor de terra ou à acção conjunta dos dois acontecimentos; tratando-se, portanto, de um acontecimento composto pela união dos acontecimentos A e B, a respectiva probabilidade é dada por $P(A \cup B) = P(A) + P(B) - P(A \cap B)$, sendo que não se conhece $P(A \cap B)$. No pressuposto de que, mesmo que exista alguma dependência estatística entre A e B, $P(A \cap B)$ deverá apresentar um valor muito baixo e atendendo à desigualdade de Boole, resulta, de modo conservador, que $P(A \cup B) \cong P(A) + P(B) = 0.02+0.01=0.03$. Admitindo-se que os acontecimentos A e B são independentes, obter-se-ia $P(A \cup B) = P(A) + P(B) - P(A)\,P(B) = 0.0298$.

3.8. Variáveis aleatórias discretas e contínuas

Seja E uma experiência aleatória e Ω o respectivo espaço amostral. Por *variável aleatória* entende-se uma função X que associa a cada elemento $s \in \Omega$ um número x(s).

Para melhor explicitar o significado de X, considere-se a experiência E: {lançamento simultâneo de duas moedas distinguíveis entre si} cujo espaço amostral é Ω:{ff, cc, fc, cf}, onde f simboliza 'face' ou 'cara', e *c* 'coroa'. Se a variável *X* for definida como o número de 'faces'/'caras' decorrentes da mencionada experiência, os seus valores possíveis são os indicados na Tabela 3.

Tabela 3 – Número de 'faces' resultantes do lançamento simultâneo de duas moedas.

Acontecimento	Valores da variável aleatória *X*	Probabilidade de ocorrência
A:{ff}	x=2	0.25
B:{cc}	x=0	0.25
C:{fc}	x=1	0.25
D:{cf}	x=1	0.25

Em condições normais de realização da experiência, os acontecimentos A, B, C e D são considerados equiprováveis, ou seja, P(A)=P(B)=P(C)=P(D)=0.25. As probabilidades de que a variável aleatória X assuma cada um dos seus possíveis valores são: P(X=2)=P(A)=0.25, P(X=0)=P(B)=0.25 e P(X=1)=P(C∪D)=P(C)+P(D)=0.50; observe-se que os acontecimentos C e D são disjuntos e, em consequência, $P(C \cap D) = 0$. Neste exemplo, a variável aleatória X apenas pode assumir valores positivos e inteiros, em conformidade com as possíveis realizações da experiência E, no espaço amostral Ω. Em geral, a notação usada para expressar a probabilidade de uma variável aleatória X assumir um dado valor x é $P(X = x)=p_X(x)$ ou simplesmente $P(X = x)=p(x)$.

- ***Variável aleatória discreta***

Uma *variável aleatória discreta* pode assumir somente *valores inteiros*, correspondendo a espaços amostrais finitos ou infinitos, porém susceptíveis de serem enumerados, ou seja, espaços amostrais *numeráveis*. No caso da experiência E:{lançamento simultâneo de duas moedas distinguíveis entre si} a que se refere a Tabela 3, sendo X o número de 'caras' obtidas num lançamento, X é uma variável aleatória discreta.

- ***Variável aleatória contínua***

Uma *variável aleatória contínua* pode assumir qualquer *valor real* num dado intervalo, correspondendo a espaços amostrais finitos ou infinitos, porém *não numeráveis*. Exemplificando-se, considere a experiência A:{medição da precipitação diária num dado posto udométrico}. A variável aleatória X representativa da precipitação diária máxima anual nesse posto é uma variável aleatória contínua pois, teoricamente, pode assumir qualquer valor real entre 0 e ∞, embora com diferentes probabilidades.

4. Funções de distribuição de probabilidade

As *funções de distribuição de probabilidade* são funções que descrevem o "comportamento" de uma variável aleatória, discreta ou contínua.

Assim, para caracterizar as probabilidades associadas aos possíveis valores de variáveis aleatórias, X, do tipo discreto, $P(X = x)=p_X(x)$, utilizam-se as designadas *funções de probabilidade* ou *funções massa de probabilidade*, fmp. Qualquer fmp tem de satisfazer as seguintes condições:

(i) $p_X(x) \geq 0, \ \forall x$;e

(ii) $\sum p_X(x) = 1, \ \forall x$.

A soma das ordenadas de uma fmp relativas aos sucessivos valores de x, conduz à designada *função acumulada de probabilidades,* FAP ou seja, $F_X(x) = P(X \leq x) = \sum_{x_i \leq x} p_X(x_i) = \sum_{x_i \leq x} p(x_i)$,. A Figura 2 ilustra as duas anteriores funções tendo por base o exemplo da Tabela 3.

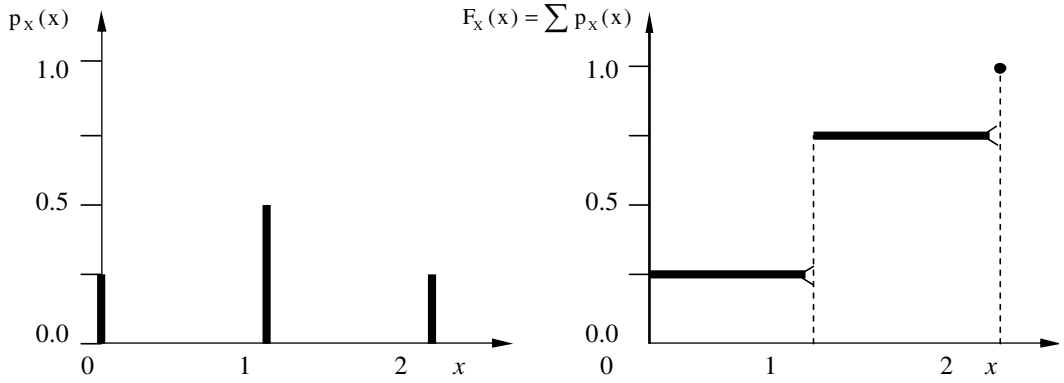

Figura 2 – Funções massa e acumulada de probabilidades da variável aleatória discreta X do exemplo da Tabela 3.

Se a variável aleatória X puder assumir qualquer valor real, ou seja, se for do tipo *contínuo*, a função equivalente à fmp é denominada por *função densidade de probabilidade*, fdp. Esta *função não negativa*, em geral denotada por $f_X(x)$ ou simplesmente por f(x), está exemplificada na Figura 3, representando o caso limite de um polígono de frequências para uma amostra de tamanho infinito e, portanto, com as amplitudes dos intervalos de classe a tender para zero.

É importante notar que, contrariamente à função fmp relativa ao caso discreto, a fdp num dado ponto x_0, $f_X(x_0)$ não fornece a probabilidade de X para o argumento x_0 e, sim, a *intensidade* com que a probabilidade de ocorrerem valores menores ou iguais do que x_0 se altera na vizinhança desse argumento.

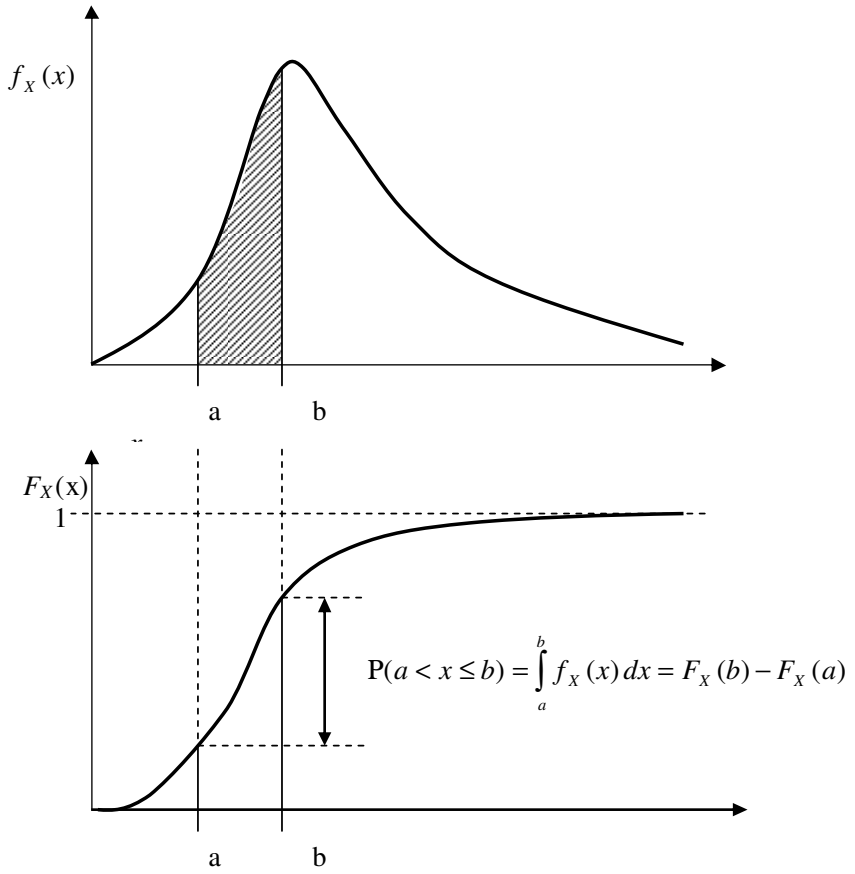

Figura 3 – Funções densidade e acumulada de probabilidades de uma variável contínua.

A área entre dois limites *a* e *b*, definidos no eixo das abcissas representativo dos possíveis valores da variável aleatória contínua, X, fornece a probabilidade de a variável estar compreendida entre esses limites, como ilustrado na Figura 3. Portanto, para uma fdp $f_X(x)$, é válida a equação:

$$P(a < X < b) = P(a \le X \le b) = \int_a^b f_X(x)\,dx = F_X(b) - F_X(a) = F(b) - F(a) \quad \text{.....(1)}$$

Consequentemente, ao fazer-se convergir o limite inferior da anterior integração, *a*, para o correspondente limite superior, *b*, a representação da área do gráfico entre aqueles limites tende, por assim dizer, para uma recta no plano real com área, por princípio, nula. Conclui-se, portanto, que, para uma variável aleatória contínua X, P(X=x)=0.

Em correspondência com o caso discreto, a *função acumulada de probabilidade,* também simplesmente designada por *função distribuição de probabilidade*, FDP, de uma variável aleatória contínua X, representada por $F_X(x)$ ou simplesmente por F(x), fornece a probabilidade associada a valores inferiores ou iguais ao argumento x, ou seja, a *probabilidade de não-excedência* de *x*, $P(X \le x)$. Inversamente, a fdp correspondente pode ser obtida pela diferenciação de $F_X(x)$, em relação a x. Tal como no caso discreto, a FDP de uma variável aleatória contínua é uma função não decrescente, sendo válidas as expressões $F_X(-\infty)=0$ e $F_X(+\infty)=1$.

Exercício 2 – Considere que a Figura 4 representa a função densidade de probabilidade da variável aleatória contínua 'caudal médio diário máximo anual (m^3/s)', numa dada estação hidrométrica. Determine: (a) $P(X<100\ m^3/s)$; (b) $P(X>300\ m^3/s)$.

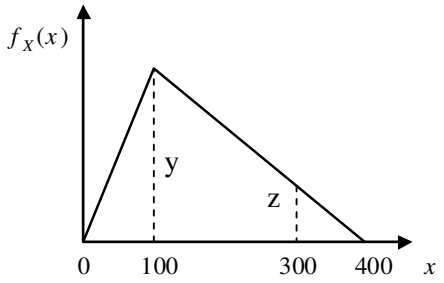

Figura 4 – Função densidade de probabilidade da variável aleatória contínua X.

Solução:

(a) Se $f_X(x)$ é uma função densidade de probabilidades, a área do triângulo deve ser igual a 1. Assim, (400y)/2=1, o que resulta em y=1/200. Logo, $P(X \leq 100\ m^3/s)$, correspondente à área do triângulo até a abcissa 100, é (100y)/2=0.25.

(b) P(X>300), ou [1- $P(X \leq 300)$], corresponde à área do triângulo à direita da abcissa 300. A ordenada z pode ser calculada por semelhança de triângulos, ou seja, (y/z)=300/100, o que resulta em z=1/600. Logo, P(X>300)=0.083.

5. Medidas descritivas populacionais das variáveis aleatórias

5.1. Nota prévia

A *população* de uma variável aleatória X corresponde ao universo ou espaço amostral dos todos os seus possíveis resultados, cujas frequências de ocorrências podem ser sintetizadas por uma fmp $p_X(x)$ ou por uma fdp, $f_X(x)$, consoante X é uma variável aleatória discreta ou contínua, respectivamente. Em ambos os casos e de modo equivalente às estatísticas descritivas de uma amostra extraída daquela população, objecto do **item 2**, as características de forma das funções $p_X(x)$ ou $f_X(x)$ podem ser sintetizadas por meio de *medidas descritivas populacionais*. Tais medidas são obtidas através de *médias*, *ponderadas por* $p_X(x)$ ou $f_X(x)$, de funções da variável aleatória e incluem o *valor esperado*, a *variância* e o *coeficiente de assimetria*, entre outras.

5.2. Valor esperado

O *valor esperado* ou a *esperança matemática* de X é o resultado da soma de todos os valores possíveis da variável aleatória, ponderados por $p_X(x)$ ou por $f_X(x)$. O valor esperado, denotado por E[X], equivale à *média populacional*, μ_X, indicando, portanto, a *abcissa do centro de massa* ou *centróide* das funções $p_X(x)$ ou $f_X(x)$, pelo que tem as mesmas unidades de X. A definição formal de E[X] é dada por:

$$E[X] = \mu_X = \sum_i x_i \, p_X(x_i) \qquad \forall\, x_i \quad \text{.....(2)}$$

para o caso discreto; e por

$$E[X] = \mu_X = \int_{-\infty}^{+\infty} x \, f_X(x) dx \quad \text{.....(3)}$$

para o caso contínuo.

O valor esperado pode ser entendido como um *operador matemático* e ser generalizado para qualquer função g(X) da variável aleatória X, conforme expresso pelas equações (4) e (5) para X discreta ou contínua, respectivamente.

$$E[g(X)] = \sum_i g(x_i) p_X(x_i) \qquad \forall\, x_i \quad \text{.....(4)}$$

$$E[g(X)] = \int_{-\infty}^{+\infty} g(x) f_X(x) dx \quad \text{.....(5)}$$

As principais *propriedades* do operador valor esperado E(.) são:

- E[c]=c, para *c* constante.
- E[cg(X)]=cE[g(X)], para c constante e g(X) com o significado antes apresentado.
- $E[c_1 g_1(X) \pm c_2 g_2(X)] = c_1 E[g_1(X)] \pm c_2 E[g_2(X)]$, para c_1 e c_2 constantes e $g_1(X)$ e $g_2(X)$ funções de X.
- $E[g_1(X)] \geq E[g_2(X)]$, se $g_1(X) \geq g_2(X)$.

Exercício 3 – Calcule o valor esperado para a função massa de probabilidades especificada pela Figura 2.

Solução: A aplicação da equação (2) resulta em $E[X]=\mu_X=0\times0.25+1\times0.50+2\times0.25=1$ que, de facto, é o centróide da função massa de probabilidades.

Exercício 4 – Considere uma variável aleatória contínua *X*, cuja função densidade de probabilidade é dada por $f_X(x)=1/\theta\exp(-x/\theta)$, para $x\geq0$ e $\theta\geq0$, tratando-se, portanto, da distribuição de probabilidade *exponencial*, que, de facto, é uma família de curvas, a depender do valor numérico do parâmetro θ. Nessas condições: (a) calcule o valor esperado de X; (b) supondo que o valor numérico de θ é igual a 2, calcule a probabilidade associada a valores da variável aleatória superiores a 3, ou seja, $P(X>3)$; e (c) supondo que $\theta=2$, calcule a mediana da variável aleatória exponencial X.

Solução: (a) Para a distribuição em questão, $E[X]=\mu_X=\int_0^\infty x\,f_X(x)dx=\int_0^\infty(x/\theta)\exp(-x/\theta)dx$. Esta integração pode ser resolvida por partes, ou seja, $dv=(1/\theta)\exp(-x/\theta)\,dx\Rightarrow v=-\exp(-x/\theta)$ e $u=x\Rightarrow du=dx$. Resulta, assim, $\int_0^\infty udv=uv\Big]_0^\infty-\int_0^\infty vdu=-x\exp(-x/\theta)\Big]_0^\infty-\theta\exp(-x/\theta)\Big]_0^\infty=\theta$. Portanto, para a forma paramétrica exponencial, o valor esperado, ou seja, a média da população μ_X é igual ao parâmetro θ; por outras palavras, a abcissa do centróide da função densidade de probabilidade, fdp, exponencial é θ. (b) A probabilidade pedida é calculada por $P(X>3)=1-P(X\leq3)=1-F_X(3)$ em que $F_X(x)$ é a função distribuição de probabilidade, FDP, dada por $F_X(x)=\int_0^x(1/\theta)\exp(-x/\theta)dx$ e cuja solução é $F_X(x)=1-\exp(-x/\theta)$. Para os dados do exercício, $P(X>3)=1-1+\exp(-3/2)=0.2231$. (c) A mediana é o valor de *x* que corresponde a $P(X\geq x)=P(X\leq x)=F_X(x)=0.50$. Invertendo-se a função $F_X(x)$, obtém-se $x(F)=-\theta\ln(1-F)$. Para os dados do exercício, a mediana é $x(0.50)=-2\ln(1-0.50)=1.39$.

5.3. Variância, desvio-padrão e coeficiente de variação da população

A *variância da população* de uma variável aleatória X, representada por Var[X] ou por σ_X^2, é definida como sendo o *momento central de segunda ordem*, ou μ_2, e corresponde à medida populacional mais frequentemente utilizada para caracterizar a dispersão das funções massa, $p_X(x)$, ou densidade, $f_X(x)$ de probabilidade. Obtém-se, assim:

$$Var[X]=\sigma_X^2=\mu_2=E\left[(X-\mu_X)^2\right]=E\left[(X-E[X])^2\right] \quad \text{.....(6)}$$

Expandindo o quadrado contido na anterior equação e usando as propriedades do operador esperança matemática, resulta:

$$Var[X]=\sigma_X^2=\mu_2=E[X^2]-(E[X])^2 \quad \text{.....(7)}$$

Logo, a variância populacional de uma variável aleatória X é igual ao valor esperado do quadrado dessa variável menos o quadrado do valor esperado de X, ou seja, o quadrado da média de X. A variância de X tem as mesmas unidades de X^2 e as seguintes propriedades:

- $Var[c]=0$, para c constante.
- $Var[cX]=c^2Var[X]$.
- $Var[cX+d]=c^2Var[X]$, para d constante.

De modo equivalente às estatísticas descritivas amostrais, o *desvio-padrão da população* σ_X é a raiz quadrada (positiva) da variância, σ_X^2, possuindo, portanto, as mesmas unidades de X. Define-se, igualmente, uma medida relativa adimensional da dispersão de $p_X(x)$ ou $f_X(x)$ por meio do *coeficiente de variação populacional* CV_X, dado por:

$$CV_X = \frac{\sigma_X}{\mu_X} \quad \text{.....(8)}$$

Exercício 5 – Calcule a variância, o desvio-padrão e o coeficiente de variação para a função massa de probabilidade especificada pela Figura 2.

Solução: A aplicação da equação (7) requer o cálculo de $E[X^2]$ para o qual resulta $E[X^2] = \sum_i x_i^2 \, p_X(x_i) = 0^2 \times 0.25 + 1^2 \times 0.5 + 2^2 \times 0.25 = 1.5$. Atendendo a que, de acordo com o exercício 3, $E[X]=\mu_X=1$, obtém-se para a equação (7), $Var[X] = \sigma^2_X = 1.5 - 1.0^2 = 0.5$. O desvio padrão é, portanto, $\sigma_X = 0.71$ e o coeficiente de variação, $CV_X = 0.71/1.0 = 0.71$.

5.4. Coeficiente de assimetria

O *coeficiente de assimetria* γ_X de uma variável aleatória X é uma grandeza adimensional definida por

$$\gamma_X = \frac{\mu_3}{(\sigma_X)^3} = \frac{E\left[(X-\mu_X)^3\right]}{(\sigma_X)^3} \quad \text{.....(9)}$$

O numerador do segundo membro da equação (9) é o *momento central de ordem 3*, ou seja, é o valor esperado do cubo dos desvios da variável aleatória *X* em relação à respectiva média μ_X, podendo ser positivo, negativo ou nulo. Se tal numerador e, consequentemente, o coeficiente de assimetria, forem nulos, a função densidade (ou massa) de probabilidade será simétrica. Se os valores de *X* superiores à média μ_X estiverem relativamente muito mais afastados do que os inferiores, os cubos dos desvios positivos irão prevalecer sobre os negativos e o coeficiente γ_X será positivo, configurando uma função densidade (ou massa) com assimetria positiva. Caso contrário, ter-se-á uma função densidade (ou massa) de probabilidade com assimetria negativa.

A Figura 5 ilustra três funções densidades de probabilidade: uma simétrica, portanto, com o coeficiente de assimetria nulo, outra com assimetria positiva igual a γ=1.14 a e a terceira com a assimetria negativa de γ=-1.14.

Outras medidas, como os momentos de ordens superiores a 3 e o coeficiente de curtose, embora constituam importantes complementos para a caracterização da forma das funções densidade (ou massa) de probabilidade, encontram aplicações menos frequentes na modelação de variáveis aleatórias hidrológicas. Ao leitor interessado em aprofundar os seus conhecimentos sobre estes tópicos, recomenda-se a consulta dos livros de *Rao e Hamed* (2000) e *Hosking e Wallis* (1997).

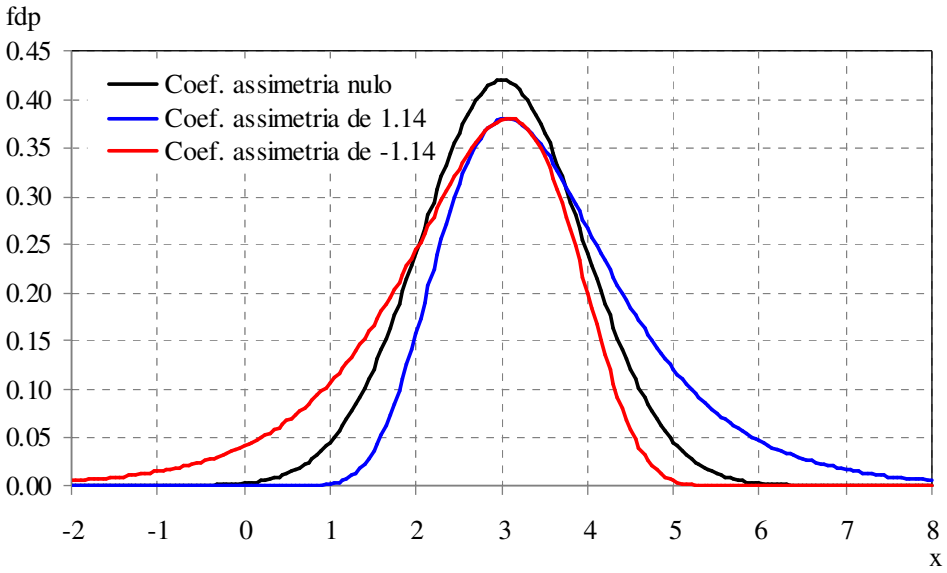

Figura 5 – Exemplos de funções densidade (ou massa) de probabilidade simétricas e assimétrica.

6. Modelos de distribuição de probabilidades de variáveis aleatórias discretas

6.1. Nota prévia

Um *modelo de distribuição de probabilidades* é uma forma matemática abstracta capaz de representar, de modo conciso, as variações contidas numa amostra de uma variável aleatória. Um modelo de distribuição de probabilidades também é uma *forma paramétrica*, ou seja, é um modelo matemático contendo parâmetros, cujos valores numéricos o definem completamente e o particularizam para uma dada amostra de uma variável aleatória. Uma vez estimados os valores numéricos desses parâmetros, o modelo de distribuição de probabilidades passa a caracterizar o comportamento plausível da variável aleatória a que respeita aquela amostra podendo, como tal, ser utilizado para interpolar ou extrapolar probabilidades e/ou quantis não contidos na mesma.

Os principais *modelos de variáveis aleatórias discretas* que encontram aplicações em hidrologia estão relacionados com repetições independentes dos chamados *processos de Bernoulli*. Estes modelos são as distribuições *geométrica e binomial* que a seguir se descrevem de modo sucinto.

6.2. Distribuição geométrica. Período de retorno

Por prova de Bernoulli entende-se a experiência aleatória em que somente dois *resultados dicotómicos* são possíveis: "sucesso" ou "falha", "sim" ou "não", "0" ou "1", "positivo" ou "negativo" são exemplos. Tal conceito serve de base a várias distribuições teóricas.

Suponha-se que a escala temporal associada a uma determinada variável aleatória foi discretizada em intervalos com amplitude definida, por exemplo, em intervalos anuais. Suponha-se também que, em cada intervalo de tempo, possa ocorrer um único 'sucesso', com probabilidade *p*, ou uma única 'falha', com probabilidade (1-*p*), e que essas probabilidades não são afectadas pelas ocorrências anteriores, nem afectem as ocorrências posteriores. O processo composto pela anterior *sequência de repetições independentes* de uma prova de Bernoulli constitui uma *sucessão de provas de Bernoulli*.

Para melhor ilustrar a aplicação dos processos de Bernoulli à hidrologia, considere que o caudal médio diário correspondente ao extravasamento/transbordamento de uma secção transversal de um curso de água é Q_0, conforme se esquematiza na Figura 6. Considere, ainda, que, em tal secção, o regime fluvial se encontra em regime natural (ou seja, não é influenciado pelo Homem), que se dispõe na mesma de registos contínuos durante *N* anos de caudais médios diários - *série completa* de caudais médios diários – e que, para analisar as condições de transbordamento da secção, se constitui a série de caudais médios diários máximos anuais formada em cada ano pelo máximo caudal médio diário nesse ano, Q^{max} – *série reduzida* de Q^{max}, com dimensão N, representada na Figura 6. Em qualquer ano i, com $1 \le i \le N$, o 'sucesso', em termos de transbordamento, é dado pelo acontecimento $S:\left\{Q_i^{max} > Q_0\right\}$, sendo a 'falha' o acontecimento complementar $F:\left\{Q_i^{max} \le Q_0\right\}$. Tratando-se de um problema de génese de cheias num trecho fluvial em regime natural, é válido admitir que a probabilidade de ocorrência de um 'sucesso' (ou de uma 'falha'), em um ano qualquer, não é afectada pelas ocorrências em anos anteriores e em nada afecta as ocorrências em anos posteriores. Supondo que a probabilidade anual do acontecimento $\left\{S: Q_i^{max} > Q_0\right\}$ é igual a p, verifica-se, assim, o preenchimento de todos os requisitos para considerar essa sequência independente como um processo de Bernoulli.

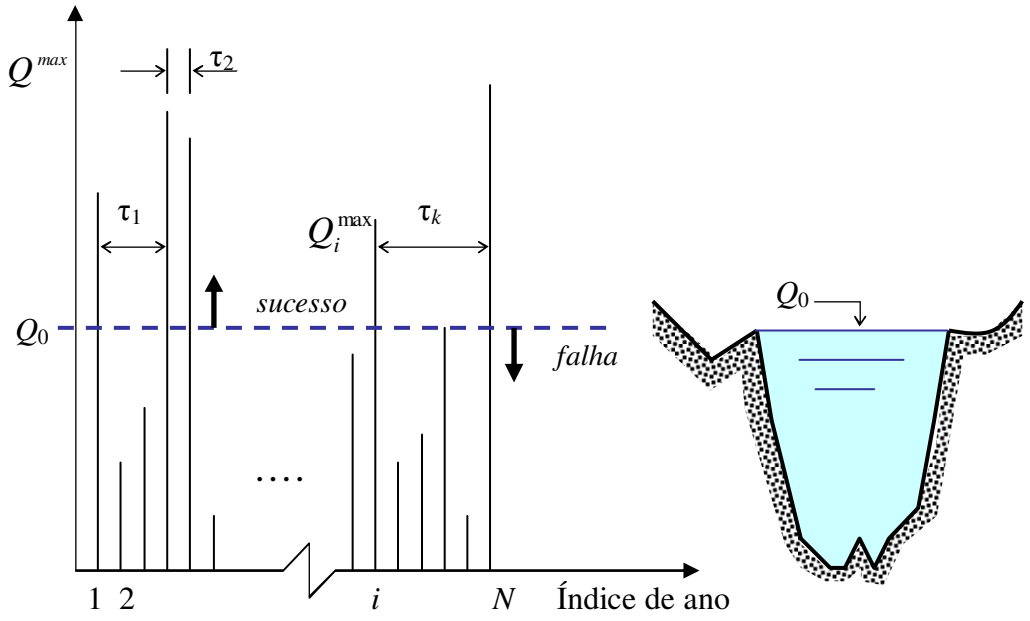

Figura 6 – Cheias máximas anuais como ilustração de um processo de Bernoulli.

A variável aleatória discreta Y correspondente à *distribuição geométrica* refere-se ao *número inteiro* de experiências (ou intervalos discretos de tempo) necessários para que um único 'sucesso' ocorra. Portanto, se o valor da variável é Y=y, isto significa que ocorreram (*y*-1) 'falhas' antes da ocorrência do 'sucesso', exactamente, na *y*-ésima tentativa. As *funções massa e acumulada da distribuição geométrica* são dadas pelas seguintes equações:

$$p_Y(y) = p(y) = p(1-p)^{y-1},\ y = 1,2,3,...\infty \text{ e } 0 < p < 1 \quad \text{.....(10)}$$

$$F_Y(y) = F(y) = \sum_{i=1}^{y} p\,(1-p)^{i-1},\ y = 1,2,3,...,\infty \quad \text{.....(11)}$$

nas quais a probabilidade anual de ocorrência de um 'sucesso', p, representa o único *parâmetro* da distribuição. Demonstra-se que *valor esperado de uma variável geométrica*, resultado da soma infinita de termos, decorrente da aplicação da equação (2), é

$$E[Y] = \frac{1}{p} \quad \text{.....(12)}$$

ou seja, quando o número de repetições (ou intervalos discretos de tempo) tende para infinito, o valor médio de uma variável geométrica é o *inverso da probabilidade de 'sucesso' p*.

Introduza-se, neste ponto, um conceito de grande importância em hidrologia, que é o de *período de retorno*. Para tanto, considere-se que, nas condições da Figura 6, a variável τ designa o número de anos entre 'sucessos' (transbordamentos) consecutivos. Adoptando-se para origem da escala de tempos o ano do primeiro 'sucesso', a Figura 6 indica que seriam necessários τ_1=3 anos para uma nova ocorrência do acontecimento S: $\left\{ Q_{i=4}^{max} > Q_0 \right\}$. A partir do segundo 'sucesso', τ_2=1 ano e assim sucessivamente até τ_k=5 anos. Se, por hipótese, N=50 anos e se nesse período de tempo tivessem ocorrido 5 'sucessos', depreender-se-ia que o número de anos que, *em média*,

separaria as ocorrências de caudais superiores a Q_0 seria de $\bar{\tau}=10$ anos, significando que o caudal Q_0 é superado com a *frequência anual média* de 1 a cada 10 anos.

É fácil verificar que a variável τ se enquadra integralmente na definição de uma variável aleatória discreta geométrica e que, portanto, a ela se podem associar as características populacionais definidas pelas equações (10), (11) e (12). Em particular, pode definir-se o *período de retorno*, denotado por *T* e expresso em anos, como o *valor esperado* da variável geométrica τ. Com essa definição e usando a equação (12), resulta:

$$T = E[\tau] = \frac{1}{p} \quad \text{.....(13)}$$

O período de retorno, T, não se refere, portanto, a um 'tempo cronológico'. De facto, T é uma *medida da tendência central dos 'tempos cronológicos'*. Por outras palavras, o período de retorno, T, associado a um certo acontecimento de referência de um processo de Bernoulli necessariamente definido numa base temporal anual, corresponde ao *número médio de anos necessários para que o acontecimento ocorra num ano qualquer desses anos e é igual ao inverso da probabilidade de esse acontecimento ocorrer num ano qualquer desses anos, ou seja, é igual ao inverso da probabilidade anual de ocorrência desse acontecimento.*

Em hidrologia, o conceito de período de retorno é vulgarmente utilizado, por exemplo, no estudo probabilístico de *acontecimentos máximos anuais*, tais como caudais instantâneos ou diários máximos anuais ou, ainda, precipitações máximas anuais com dada duração. Tais variáveis aleatórias são contínuas e, portanto, têm o seu comportamento definido por funções densidade de probabilidade genericamente designadas por $f_X(x)$. Se, para uma dessas variáveis, denotada por X, se definir um quantil de referência x_T, de modo que o 'sucesso' seja a ocorrência de valores superiores a x_T, então, o período de retorno, T, associado a esse quantil de referência á dado pelo *número médio de anos* necessário para que o acontecimento $\{X>x_T\}$ ocorra uma vez, num *qualquer desses anos*. De acordo com a equação (13), resulta que o período de retorno corresponde ao inverso de $P(X>x_T)$, ou seja, ao inverso de $[1-F_X(x_T)]$.

Exercício 6 – Considere a situação descrita no exercício 2, na qual a variável X se refere ao caudal médio diário máximo anual (m^3/s). Determine: (a) o período de retorno para x=300 m^3/s; e (b) o caudal médio diário máximo anual com o período de retorno T=50 anos.

Solução: (a) Estando-se em presença de uma variável definida numa base anual é válido aplicar a noção de período de retorno. Atendendo a que tal período é dado pelo inverso da probabilidade de excedência e tendo-se estimado no exercício 2 que P(X>300)=0.083 resulta que o período de retorno associado a esse caudal é de T=1/0.083=12.05 anos. (b) Ao período de retorno de T=50 anos corresponderá um caudal x_{50} compreendido entre 300 e 400 m^3/s já que $P(X>x_{50})=[1- P(X \le x_{50})]=0.02$. De entre as possíveis vias de resolução do problema, optou-se por atender à equação da recta que passa pelos pontos (100; 1/200) e (400; 0) dada por $f_X(x)=f(x)=-x/60000+1/150$. De acordo com o pretendido, a área do triângulo com base dada pelo segmento de recta definido pelas abcissas x_{50} e 400 e com altura dada por $f(x_{50})=-x_{50}/60000+1/150$ é igual a 0.02, ou seja $(400-x_{50})(-x_{50}/60000+1/150)/2=0.02$. A anterior equação do segundo grau tem duas raízes, uma maior do que 400 m^3/s e que, portanto, está fora do domínio de definição de X, e a outra de sensivelmente x_{50}=351 m^3/s e que constitui a solução do problema. Nesse ponto, o valor de $f_X(x)$ é de aproximadamente 0.000817, verificando-se que se obtém de facto para a área do triângulo 0.000817 (400-351)/2=0.02.

6.3 Distribuição Binomial. Risco hidrológico

Ainda referente ao processo de Bernoulli anteriormente descrito, considere-se que a variável aleatória discreta Y representa o número de 'sucessos', de entre N possibilidades (ou intervalos discretos de tempo). A variável Y pode ter qualquer valor entre 0, 1, ... , *N*. Em resultado da hipótese de independência entre as experiências de Bernoulli, cada ponto do espaço amostral com *y* 'sucessos' e (N-y) 'falhas' terá probabilidade de ocorrência igual a $p^y(1-p)^{N-y}$. Entretanto, os y 'sucessos' e as (N-y) 'falhas' podem ser combinados de $N!/[y!(N-y)!]$ modos diferentes, cada um deles com probabilidade igual a $p^y(1-p)^{N-y}$. Portanto, a fmp da variável *Y* é dada por

$$p_Y(y) = \frac{N!}{y!(N-y)!}p^y(1-p)^{N-y} = \binom{N}{y}p^y(1-p)^{N-y},\ y = 0,1,...,N \text{ e } 0<p<1 \quad \text{.....(14)}$$

que constitui a distribuição *binomial*, com parâmetros N e p. A FAP da distribuição binomial fornece a probabilidade de X ser menor ou igual ao argumento x e é dada por

$$F_Y(y) = \sum_{i=0}^{y}\binom{N}{i}p^i(1-p)^{N-i},\ y = 0,1,2,...,N \quad \text{.....(15)}$$

O valor esperado e a variância da distribuição binomial são respectivamente iguais a Np e Np(1-p). A fmp binomial é simétrica quando p=0.5 e apresenta assimetria positiva, se p<0.5, e negativa, em caso contrário.

Exercício 7 – Nas condições da Figura 6, suponha-se que a dimensão da séries caudais médios diários máximos caudais, Q^{max}, é de N=10 anos e que o *período de retorno* associado ao caudal Q_0 é de 4 anos. Pergunta-se: (a) qual é a probabilidade de que o caudal Q_0 tenha sido superado exactamente em 2 dos 10 anos? (b) qual é a probabilidade de que o caudal Q_0 tenha sido superado em pelo menos 2 dos 10 anos?

Solução: É fácil verificar que o cenário ilustrado pela Figura 6 se adequa a um processo de Bernoulli e a variável 'número de sucessos em N anos', a uma variável binomial Y. (a) A probabilidade de que o caudal Q_0 tenha sido superado exactamente 2 vezes em 10 anos pode ser calculada directamente pela equação 14, sabendo-se que a probabilidade anual p (de 'sucesso') é o inverso do período de retorno T=4 anos, ou seja, p=0,25. Logo, $p_Y(2) = [10!/(2!\,8!)]0.25^2(1-0.25)^8 = 0.2816$. (b) A probabilidade de que o caudal Q_0 tenha sido excedido pelo menos 2 vezes em 10 anos é igual à probabilidade de que o acontecimento tenha ocorrido 2, 3, 4, ... , 10 vezes, em 10 anos, ou seja, é igual à soma dos resultados da função massa para todos os argumentos compreendidos entre 2 e 10, inclusive. Entretanto, tal cálculo é equivalente ao cálculo do complementar, em relação a 1 ocorrência, da soma das probabilidades de que o acontecimento não tenha ocorrido ou que tenha ocorrido apenas 1 vez. Portanto, nesse entendimento, $P(Y \geq 2) = 1 - P(Y < 2) = 1 - p_Y(0) - p_Y(1) = 0.7560$.

Um conceito associado ao período de retorno refere-se à definição de *risco hidrológico*, tal como aplicado em projectos de estruturas hidráulicas de controlo de cheias ou de desvio provisório de um curso de água durante as obras de construção de uma barragem.

Seja x_T o valor da variável hidrológica, por exemplo, caudal de ponta de cheia, para o período de retorno T. Nestas condições, o risco hidrológico, R, não é mais do que a probabilidade de ocorrer um ou mais valores da variável hidrológica iguais ou superiores a x_T num período de N anos. Em geral, o quantil de referência x_T corresponde à cheia para a qual foi projectada a estrutura hidráulica, enquanto o período de N anos corresponde à sua vida útil da obra ou período durante o qual é necessário assegurar o desvio do curso de água. A dedução da expressão do

risco hidrológico, R, pode recorrer à distribuição binomial. Com efeito, a probabilidade de que pelo menos um 'sucesso' ocorra num período de N anos é equivalente à probabilidade do acontecimento complementar, em relação a 1, de que nenhum 'sucesso' ocorra nesse período. Portanto, usando a notação Y para o número de 'sucessos' em N anos, tem-se que

$$R = P(Y \geq 1) = 1 - P(Y = 0) = 1 - \binom{N}{0} p^0 (1-p)^{N-0} \quad \text{.....(16)}$$

Se o quantil de referência x_T tem período de retorno T, a probabilidade de um 'sucesso', em um ano qualquer, é igual a $1/T$. Substituindo este resultado na equação (16), segue-se que

$$R = 1 - \left(1 - \frac{1}{T}\right)^N \quad \text{.....(17)}$$

Um raciocínio alternativo, embora simplificado, para alcançar a noção de risco hidrológico utiliza fundamentalmente o conceito de período de retorno e a independência temporal dos "sucessos" ou dos "insucessos". Com efeito, representando x_T o valor da variável hidrológica com o período de retorno T, a probabilidade de, em qualquer ano, ocorrer x_T é, como antes afirmado, igual a 1/T. Logo, a probabilidade de x_T não ocorrer em qualquer ano é 1-1/T. Atendendo a que a não ocorrência de x_T num dado ano em nada altera a probabilidade de não ocorrer no ano ou nos anos seguintes (pois os acontecimentos são independentes) concluiu-se que a probabilidade de x_T não ocorrer em nenhum dos N anos do período considerado é de $(1-1/T)^N$. Logo, o risco hidrológico, sendo a probabilidade de x_T ocorrer uma ou mais vezes durante esses N anos, não é mais do que o acontecimento complementar daquele outro acontecimento, correspondendo-lhe, portanto, uma probabilidade complementar, do que precisamente resulta a equação 17.

Se o risco hidrológico foi fixado à priori, por exemplo, em função da tipologia, da importância e das dimensões da estrutura hidráulica, bem como das consequências (incluindo eventual danos materiais e perda de vidas humanas) do seu eventual colapso, pode empregar-se a equação 17 para determinar o período de retorno que deve ser adoptado como critério de projecto, em face do período de vida útil da obra de N anos a que tal critério de projecto se aplica.

Exercício 8 – A Figura 7 mostra o esquema do desvio provisório de um rio durante a construção de uma barragem, compreendendo a execução de duas ensecadeiras A e B e de um túnel de desvio provisório inserido na margem direita e iniciando-se a montante da ensecadeira de montante e finalizando a jusante da ensecadeira de jusante.

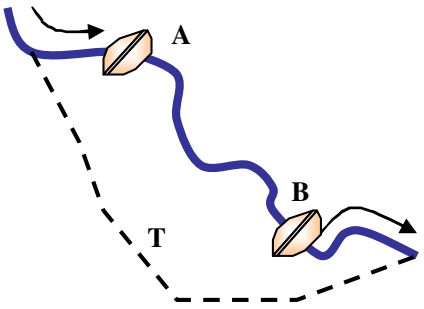

Figura 7 – Esquema de desvio provisório de um rio.

Deste modo e até dadas condições de projecto, não existirão caudais circulantes no trecho fluvial compreendido entre ensecadeiras. Suponha-se que o período de construção da obra é de 5 anos e que o risco de inundação do trecho fluvial entre ensecadeiras foi fixado em 10% (probabilidade de a capacidade de vazão do túnel ser excedida e de as ensecadeiras serem galgadas uma ou mais vezes durante o período de construção de apenas 10%). Com base nesses elementos, determine o período de retorno do caudal de ponta de projecto a considerar no dimensionamento do túnel e na fixação da cota do topo das ensecadeiras.

Solução: A inversão da equação 17 fornece para *T*:

$$T = \frac{1}{1-(1-R)^{1/N}}$$

Para R=0.10 e N=5 a anterior equação conduz a T=47.95 anos. Deste modo, a secção transversal do túnel e a cota do topo das ensecadeiras devem ser dimensionadas para o caudal de ponta de cheia com período de retorno de aproximadamente 50 anos.

7. Modelos de distribuição de probabilidades de variáveis aleatórias contínuas

De modo análogo às variáveis aleatórias discretas, existe um grande conjunto de modelos probabilísticos para as variáveis aleatórias contínuas, com funções densidade de probabilidade, fdp, e distribuição de probabilidade, FDP, definidas por parâmetros. A partir desse conjunto, elaboraram-se as Tabelas 4 e 5 contendo uma lista não exaustiva dos modelos com maior aplicação às variáveis hidrológicas, bem como a especificação dos respectivos parâmetros e características principais.

De acordo com as características intrínsecas mais vulgarmente patentes nas amostras de certas variáveis hidrológicas, especificam-se, seguidamente, alguns dos modelos probabilísticos que previsivelmente melhor se adequam a essas variáveis. Assim, (i) as distribuições Normal e log-Normal ou de Galton são frequentemente aplicáveis a valores anuais da precipitação e do escoamento; (ii) as distribuições log-Normal, de Gumbel para máximos ou Gumbel Max (por regra, referenciada apenas por distribuição de Gumbel), Pearson III, log-Pearson III e Generalizada de Valores Extremos (GEV), a valores extremos máximos, tais como, precipitações máximas anuais com dada duração ou caudais instantâneos máximos anuais; e (iii) os modelos de Gumbel para mínimos ou Gumbel Min e de Weibull, a valores mínimos, por exemplo, de estiagem, tais como caudais médios diários ou, ainda, em períodos de 7 dias, uns e outros, mínimos anuais. A previsível adequação de alguns modelos a dadas variáveis hidrológicas decorre, quer de considerações teóricas, quer de certas características de forma das distribuições de probabilidades, com ênfase, para as referentes à assimetria.

Anota-se que a distribuição log-Normal aplica o formalismo da distribuição Normal à transformada logarítmica da variável aleatória objecto desta última distribuição, passando-se outro tanto entre as distribuições log-Pearson III e Pearson III.

A adequação da distribuição Normal à descrição de algumas variáveis hidrológicas resulta do chamado *teorema do limite central*, segundo o qual a soma (ou a média) de um grande número de variáveis aleatórias independentes tende a ser normalmente distribuída. Raciocínio análogo pode ser elaborado para a distribuição log-Normal, no que respeita ao produto de um grande número de variáveis independentes.

No caso de valores máximos ou mínimos, a *teoria de valores extremos* fornece as bases teóricas para a utilização dos modelos que dela derivam, nomeadamente, as distribuições Gumbel Max e GEV, para máximos, e as de Gumbel Min e Weibull, para mínimos. Apesar de a aplicação dessas considerações teóricas às variáveis hidrológicas não ser isenta de controvérsia – ver, por exemplo, *Benjamin e Cornell* (1970) ou *Naghettini e Pinto* (2007) –, por regra, os modelos das Tabelas 4 e 5 e as indicações de algumas das suas potenciais aplicações são adequadas.

Para ilustrar o cálculo de probabilidades com distribuições de variáveis aleatórias contínuas, considere-se o caso da distribuição Normal a qual descreve o comportamento de uma variável aleatória contínua X que se dispõe simetricamente em torno de um valor central (a média), com funções densidade, fdp, e distribuição, FDP, de probabilidades definidas pelos parâmetros de posição (média), μ_X, e de escala (desvio-padrão), σ_X, de acordo com as equações da Tabela 4.

Tabela 4 – Principais modelos de distribuição de probabilidades de variáveis aleatórias contínuas hidrológicas e hidrometeorológicas.

Distribuição	Aplicação	Variável	Domínio	Função densidade de probabilidade, fdp [$f_x(x)$ ou $f_Y(y)$]	Função distribuição de probabilidade, FDP [$F_x(x)$ ou $F_Y(y)$]	Parâmetro		
						Posição	Escala	Forma
Normal	M/T	X	$(-\infty,+\infty)$	$f_X(x)=\frac{1}{\sigma_X\sqrt{2\pi}}\exp\left[-\frac{1}{2}\left(\frac{x-\mu_X}{\sigma_X}\right)^2\right]$	$\int_{-\infty}^{x} f_X(x)\,dx$ ou $\Phi(z)$ com $Z=\frac{X-\mu_X}{\sigma_X}$	μ_X	σ_X (>0)	-------
log-Normal ou de Galton	M/T Max	$Y=\ln(X)$	$[0,+\infty)$	$f_Y(y)=\frac{1}{\sigma_Y\sqrt{2\pi}}\exp\left[-\frac{1}{2}\left(\frac{y-\mu_Y}{\sigma_Y}\right)^2\right]$	$\int_{-\infty}^{y} f_Y(y)dy$	μ_Y	σ_Y (>0)	-------
Gumbel Max (ou apenas Gumbel)	Max	X	$(-\infty,+\infty)$	$f_X(x)=\frac{1}{\alpha}\exp\left[-\frac{x-\beta}{\alpha}-\exp\left(-\frac{x-\beta}{\alpha}\right)\right]$	$\exp\left[-\exp\left(-\frac{x-\beta}{\alpha}\right)\right]$	β	α (>0)	-------
Pearson III	Max	X	$\alpha\geq 0$: $[\delta,\infty)$ $\alpha<0$: $(-\infty,\delta]$	$f_X(x)=\frac{1}{\lvert\alpha\rvert\Gamma(\beta)}\left(\frac{x-\delta}{\alpha}\right)^{\beta-1}\exp\left(-\frac{x-\delta}{\alpha}\right)$	$\int_{\delta}^{x} f_X(x)dx$	δ	α	β (>0)
log-Pearson III	Max	$Y=\ln(X)$	$\alpha_Y\geq 0$: $[\exp(\delta_Y),\infty)$ $\alpha_Y<0$: $(-\infty,\exp(\delta_Y)]$	$f_Y(y)=\frac{1}{\lvert\alpha_Y\rvert\Gamma(\beta)}\left(\frac{y-\delta_Y}{\alpha_Y}\right)^{\beta_Y-1}\exp\left(-\frac{y-\delta_Y}{\alpha_Y}\right)$	$\int_{\delta_Y}^{y} f_Y(y)dy$	δ_Y	α_Y	β_Y (>0)
GEV	Max	X	$\kappa<0$: $x>(\beta+\alpha)/\kappa$ $\kappa<0$: $x<(\beta+\alpha)/\kappa$ $\kappa=0$: GEV≡Gumbel	$f_X(x)=\frac{1}{\alpha}\left[1-\kappa\left(\frac{x-\beta}{\alpha}\right)\right]^{1/\kappa-1}\exp\left\{-\left[1-\kappa\left(\frac{x-\beta}{\alpha}\right)\right]^{1/\kappa}\right\}$	$\exp\left\{-\left[1-\kappa\left(\frac{x-\beta}{\alpha}\right)\right]^{1/\kappa}\right\}$	β	α (>0)	κ
Gumbel Min	Min	X	$(-\infty,+\infty)$	$f_X(x)=\frac{1}{\alpha}\exp\left[\frac{x-\beta}{\alpha}-\exp\left(\frac{x-\beta}{\alpha}\right)\right]$	$1-\exp\left[-\exp\left(\frac{x-\beta}{\alpha}\right)\right]$	β	α (>0)	-------
Weibull	Min	X	$[0,+\infty)$	$f_X(x)=\frac{\alpha}{\beta}\left(\frac{x}{\beta}\right)^{\alpha-1}\exp\left[-\left(\frac{x}{\beta}\right)^{\alpha}\right]$	$1-\exp\left[-\left(\frac{x}{\beta}\right)^{\alpha}\right]$	-------	β (>0)	α (>0)

Observações: 1) Distribuições adequadas a amostras de valores: M/T, médios ou de totais anuais; Max e Min: extremos, incluindo, respectivamente, máximos anuais e mínimos anuais.

2) $\Gamma(\beta)$=função Gama completa para o argumento β ou $\Gamma(\beta)=\int_0^{\infty} x^{\beta-1}\exp(-x)dx$ (ver resolução do Exercício 10 e o Anexo 4 de *Naghettini e Pinto*, 2007).

3) A distribuição GEV, para $\kappa=0$, torna-se na distribuição de Gumbel Max ou simplesmente de Gumbel.

Tabela 5 – Principais características das distribuições de probabilidades de variáveis aleatórias contínuas hidrológicas e hidrometeorológicas.

Distribuição	Variável	Parâmetro			Média E[X] ou, havendo variável transformada Y, médias E[X] e E[Y]	Variância Var[*X*] ou, havendo variável transformada Y, variâncias Var[*X*] e Var [Y]	Coeficiente de assimetria γ_X ou, havendo variável transformada Y, coeficientes de assimetria γ_X e γ_Y	Função de quantis x(F) ou havendo variável transformada Y, funções de quantis x(F) e y(F)
		Posição	Escala	Forma				
Normal	X	μ_X	σ_X (>0)	-------	μ_X	σ_X^2	0	$\mu_X + z(F)\sigma_X$ com $z(F) = \Phi^{-1}(F)$
log-Normal ou de Galton	X Y = ln(X)	μ_Y	σ_Y (>0)	-------	$\mu_X = \exp\left[\mu_Y + \frac{\sigma_Y^2}{2}\right]$ μ_Y	$\sigma_X^2 = \mu_X^2\left[\exp\left(\sigma_Y^2\right) - 1\right]$ σ_Y^2	$\gamma_X = 3CV_X + (CV_X)^3$ com $CV_X = \frac{\sigma_X}{\mu_X} = \sqrt{\exp\left(\sigma_Y^2\right) - 1}$ $\gamma_Y = 0$	$\exp[\mu_Y + z(F)\sigma_Y]$ $\mu_Y + z(F)\sigma_Y$ com $z(F) = \Phi^{-1}(F)$
Gumbel Max (ou apenas Gumbel)	X	β	α (>0)	-------	$\beta + 0.577216\ \alpha$	$\frac{\pi^2\alpha^2}{6}$	+1,1396	$\beta - \alpha\ln[-\ln(F)]$
Pearson III	X	δ	α	β (>0)	$\alpha\beta + \delta$	$\alpha^2\beta$	$\frac{2}{\sqrt{\beta}}$	Não há forma analítica simples para a função (ver Rao e Hamed, 2000)
log-Pearson III	X Y = ln(X)	δ_Y	α_Y	β_Y (>0)	$\exp(\delta_Y)\left(\frac{1}{1-\alpha_Y}\right)^{\beta_Y}$ $\alpha_Y\ \beta_Y + \delta_Y$	$e^{2\delta_Y}\left[\left(\frac{1}{1-2\alpha_Y}\right)^{\beta_Y} - \left(\frac{1}{1-\alpha_Y}\right)^{2\beta_Y}\right]$ (ver Griffis e Stedinger, 2007) $\alpha_Y{}^2\ \beta_Y$	$\frac{E[X^3] - 3\,E[X]E[X^2] + 2\{E[X]\}^3}{\{Var[X]\}^{3/2}}$ (ver Griffis e Stedinger, 2007) $\frac{2}{\sqrt{\beta_Y}}$	Não há forma analítica simples para a função (ver Rao e Hamed, 2000)
GEV (ver obs. 1)	X	β	α (>0)	κ	$\beta + \frac{\alpha}{\kappa}[1 - \Gamma(1+\kappa)]$	$\left(\frac{\alpha}{\lvert\kappa\rvert}\right)^2\left[\Gamma(1+2\kappa) - \Gamma^2(1+\kappa)\right]$	$[-\Gamma(1+3\kappa) + 3\Gamma(1+\kappa)\Gamma(1+2\kappa) - 2\Gamma^3(1+\kappa)] / [\Gamma(1+2\kappa) - \Gamma^2(1+\kappa)]^{3/2}$ multiplicar o resultado por -1 se κ for negativo.	$\beta + \frac{\alpha}{\kappa}\left\{1 - [-\ln(F)]^k\right\}$
Gumbel Min (ver obs. 1)	X	β	α (>0)	-------	$\beta - 0.577216\alpha$	$\frac{\pi^2\alpha^2}{6}$	-1.1396	$\beta + \alpha\ln[-\ln(F)]$
Weibull (ver obs. 1)	X	-------	β (>0)	α (>0)	$\beta\,\Gamma\left(1 + \frac{1}{\alpha}\right)$	$\beta^2\left[\Gamma\left(1+\frac{2}{\alpha}\right) - \Gamma^2\left(1+\frac{1}{\alpha}\right)\right]$	$\frac{\Gamma\left(1+\frac{3}{\alpha}\right) - 3\Gamma\left(1+\frac{2}{\alpha}\right)\Gamma\left(1+\frac{1}{\alpha}\right) + 2\Gamma^3\left(1+\frac{1}{\alpha}\right)}{\sqrt{\left[\Gamma\left(1+\frac{2}{\alpha}\right) - \Gamma^2\left(1+\frac{1}{\alpha}\right)\right]^3}}$	$\beta\,[-\ln(F)]^{1/\alpha}$

Observação: 1) $\Gamma(\beta)$=função Gama completa para o argumento β ou $\Gamma(\beta) = \int_0^\infty x^{\beta-1}\exp(-x)dx$ (ver resolução do Exercício 10 e o Anexo 4 de *Naghettini e Pinto*, 2007).

A FDP de uma variável normal requer uma *integração sem solução analítica* sendo que a correspondente solução numérica depende, por sua vez, dos valores numéricos dos parâmetros μ_X e σ_X. O cálculo de probabilidades de variáveis aleatórias normais é facilitado pela utilização da variável normal reduzida Z.

Com efeito, se X é uma variável Normal e Z é uma combinação linear de X, da forma $Z=(X-\mu_X)/\sigma_X$, então a variável Z, também é distribuída segundo uma lei Normal com parâmetros μ_Z=0 e σ_Z=1. A distribuição de Z é geralmente referida como *distribuição Normal padrão* **N~**(0,1) e a variável Z, por *normal reduzida*. A integração numérica da função densidade de probabilidade da distribuição **N~**(0,1), para distintos argumentos z, Φ(z), encontra-se tabelada – Tabela 6.

Dada a simetria da fdp da lei Normal e, obviamente, da lei Normal padrão, a um argumento negativo, -z, simétrico de um outro tabelado, z, corresponde uma probabilidade de não-excedência, Φ(-z) complementar da tabelada para aquele outro valor, ou seja, Φ(-z)=1-Φ(z). A função Φ(z) consta também das funções implementadas no *software* Microsoft Excel (DIST.NORMP e NORMSDIST nas versões, respectivamente, em Português e em Inglês). O exercício 9 exemplifica o cálculo de probabilidades para a distribuição Normal.

Exercício 9 – Considere que a variável escoamento anual (m^3/s) num dado curso de água em regime natural é normalmente distribuída com média de 100 m^3/s e desvio-padrão de 50 m^3/s. Calcule (a) a probabilidade de ocorrerem caudais inferiores ou iguais a 50 m^3/s, ou seja, $P(Q\leq 50)=F(50)$; e (b) o escoamento anual com o período de retorno T=50 anos.

Solução: (a) Por meio da transformação $Z=(X-\mu_X)/\sigma_X$, verifica-se que a probabilidade pedida é dada por $P(Q\leq 50)=F(50)=P(z\leq (50-100)/50)=P(z\leq -1)=\Phi(-1)$. A Tabela 6, referente à distribuição Normal padrão, fornece $\Phi(z)$ apenas para valores positivos de z, sendo necessário recorrer à propriedade da simetria da distribuição Normal, ou seja, $\Phi(-1)=1-\Phi(+1)=1-0.8413=0.1587$. (b) De acordo com a definição de período de retorno aplicada a uma variável aleatória definida numa base anual, resulta que T=1/(1-F) em que F designa a probabilidade de não-excedência. Para T=50 anos, obtém-se $F(q)=P(Q\leq q)=0.98$. De acordo com a Tabela 6 para $\Phi(z)=0.98$ obtém-se, por interpolação linear, z=2.054. Logo, o caudal q com T=50 anos corresponde ao quantil $q=100+2.054\times 50\approx 203$ m^3/s.

Conforme antes mencionado, as amostras de algumas variáveis hidrológicas, tais como de precipitações ou de caudais máximos anuais apresentam, em geral, *coeficientes de assimetria positivos* e histogramas assimétricos à direita (ver Figura 5), em consequência de os processos naturais subjacentes aos acontecimentos hidrometeorológicos e hidrológicos raros e extremos serem normalmente caracterizados por desvios, em relação à média, dos valores extremos superiores a essa média, consideravelmente maiores do que os desvios dos valores extremos inferiores à média. Para o caso de valores máximos anuais, as Tabelas 4 e 5 identificam as distribuições mais frequentemente empregadas, a saber, os modelos log-Normal ou de Galton e de Gumbel Max (ou simplesmente de Gumbel), descritos por *dois parâmetros*, e os modelos Pearson III, log-Pearson III e GEV, com *três parâmetros*. Deste grupo, com excepção da distribuição Gumbel Max, cujo coeficiente de assimetria, γ_X, é fixo e igual a +1.1396, as distribuições restantes possuem coeficientes de assimetria variáveis, facto que as torna mais flexíveis no que concerne à forma (ver Tabela 5).

Tabela 6 – Função distribuição de probabilidade, FDP, da distribuição Normal padrão, $\Phi(z) = 1/\sqrt{2\pi}\int_{-\infty}^{z} \exp\left(-z^2/2\right)dz$.

z	**0**	**0.01**	**0.02**	**0.03**	**0.04**	**0.05**	**0.06**	**0.07**	**0.08**	**0.09**
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5606	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8585	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9137	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998

Julga-se pertinente introduzir aqui uma importante ressalva relativa aos modelos Pearson III, Log-Pearson III e GEV. Com efeito, tais modelos podem apresentar *coeficientes de assimetria negativos* (dependendo dos valores numéricos de seus parâmetros), conducentes a funções de distribuição de probabilidade que, de algum modo, definem *limites superiores* para os valores máximos da variável em estudo a que correspondem probabilidades de excedência, para

todos os efeitos, iguais a zero. Nestes casos particulares, atendendo à incerteza inerente à estimação de parâmetros populacionais a partir das amostras, em geral pequenas, de variáveis hidrológicas, é prudente não recomendar o emprego de distribuições limitadas superiormente.

O exercício 10 ilustra o cálculo de probabilidades para a lei Generalizada de Valores Extremos (GEV).

Exercício 10 – Seja X a variável aleatória 'caudal médio diário máximo anual'. Suponha-se que, numa dada secção da rede hidrográfica, E[X]=500 m^3/s, Var[X]=47 025 $(m^3/s)^2$ e γ_X=1.40. Tendo por base a lei Generalizada de Extremos, GEV, calcule o caudal médio diário máximo anual com o período de retorno 100 anos.

Solução: Conforme decorre das equações da Tabela 5 referentes à lei GEV, a relação entre o parâmetro de forma κ e o coeficiente de assimetria γ_X é biunívoca sendo apresentada no gráfico da Figura 8. Para γ_X=1.40 resulta $\kappa \approx$ -0.04. Recorrendo novamente à Tabela 5, nomeadamente, às equações da GEV que relacionam Var[X] com α e E[X] com α e β, obtém-se primeiramente α=159.97 e, seguidamente, fazendo intervir este resultado, β=401.09. Anota-se que o *software* Microsoft Excel dispõe de uma função estatística – LNGAMA, na versão em Português, e GAMMLN, na versão em Inglês – que corresponde ao logaritmo neperiano da função Gama para um dado argumento, pelo que a exponencial dessa função fornece Γ para esse argumento. O caudal médio diário máximo anual com o período de retorno de T=100 anos é dado pela função de quantis da GEV (última coluna da Tabela 5), ou seja, x(100)=1209 m^3/s.

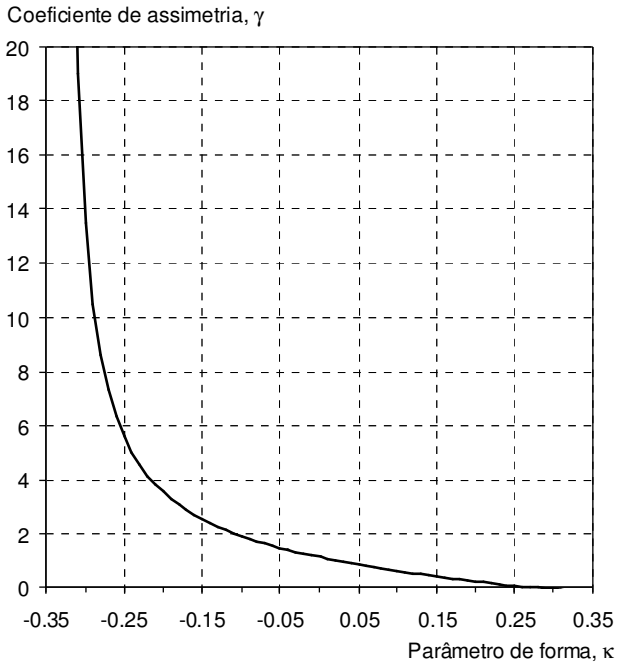

Figura 8 – Modelo GEV: relação entre κ e γ_X.

8. Estimação de parâmetros e de quantis das distribuições de probabilidade

8.1 Procedimento geral. Método dos momentos

Tomada a decisão quanto ao modelo de distribuição de probabilidades a aplicar à amostra de uma variável aleatória e determinados os valores numéricos dos parâmetros que o definem, é possível calcular as probabilidades associadas a quaisquer valores da variável em questão. Importa, contudo, registar que, mesmo que tal modelo represente fidedignamente a variável aleatória, só seria possível conhecer os verdadeiros valores numéricos dos seus parâmetros se toda a população tivesse sido amostrada, o que, na prática e pelo menos no que respeita às variáveis hidrológicas, é impossível.

Assim, na posse de apenas uma amostra finita de observações de uma variável aleatória – como a amostra de precipitações diárias máximas anuais apresentada na Tabela 1 –, pretender-se-á, por regra: (i) identificar o modelo de distribuição de probabilidades da população donde provém a amostra; e (ii) proceder à estimativa dos valores numéricos dos parâmetros que descrevem tal modelo. Os métodos que permitem estabelecer a associação entre a realidade física contida num conjunto de observações (ou seja, numa amostra) e a concepção abstracta de um modelo probabilístico são geralmente denominados de *inferência estatística*.

A população é, de certa forma um conceito abstracto, pois remete para um conjunto infinito de elementos potencialmente observáveis, mas que não existem no sentido físico. Por outro lado, a amostra é constituída por um conjunto de N observações reais $\{x_1, x_2, ..., x_N\}$, as quais se supõem terem sido *aleatoriamente sorteadas*, *uma a uma, de modo independente entre si, de uma única população,* cujo comportamento probabilístico é dado por uma certa função densidade de probabilidades $f_X(x)$ ou $f(x)$, definida por parâmetros $\theta_1, \theta_2, \ldots, \theta_k$. Nas anteriores condições de amostragem, $\{x_1, x_2, ..., x_N\}$ constitui uma *amostra aleatória simples* (*AAS*). As observações $\{x_1, x_2, ..., x_N\}$ representam os factos concretos, a partir dos quais, são obtidas as *estimativas das características populacionais*, tais como a média, a variância e o coeficiente de assimetria, assim como as inferências sobre a respectiva distribuição de probabilidades e sobre os valores dos seus parâmetros.

Em alguns casos, a forma de $f_X(x)$ pode ser deduzida a partir das características físicas do fenómeno em questão ou de algumas estatísticas amostrais. Entretanto, mesmo que $f_X(x)$ tenha sido correctamente postulada, as estimativas $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_i, \ldots, \hat{\theta}_k$, dos seus parâmetros $\theta_1, \theta_2, \ldots, \theta_i, \ldots, \theta_k$, têm de ser necessariamente inferidas a partir de uma amostra. Se outras amostras, todas com a mesma dimensão N da anterior amostra, estivessem disponíveis seria de esperar que cada uma delas produzisse estimativas, $\hat{\theta}_i$, distintas dos parâmetros da distribuição, θ_i. Se as amostras com dimensão N susceptíveis de serem constituídas fossem em grande número, as sucessivas estimativas assim obtidas para cada um daqueles parâmetros constituiriam, elas próprias, uma variável aleatória e, portanto, uma distribuição da *estatística amostral* em causa, a qual teria de conter o verdadeiro valor populacional desse parâmetro, embora de forma mais ou menos dispersa, em conformidade com o grau de incerteza decorrente do processo de estimação dos parâmetros populacionais a partir de amostras finitas de tamanho N.

Há uma variedade de métodos de estimação de parâmetros, entre os quais se destacam: (i) o método dos momentos; (ii) o método da máxima verosimilhança; (iii) o método dos momentos-L; (iv) o método da máxima entropia; (v) o método dos mínimos quadrados; e (vi) o método

generalizado dos momentos. No presente documento apenas o *método dos momentos* será objecto de apresentação, por ser o método mais frequente utilizado e de mais fácil implementação. Ao leitor interessado noutros métodos de estimação de parâmetros de distribuições estatísticas, recomendam-se as seguintes referências: *Rao e Hamed* (2000), *Hosking e Wallis* (1997), *Meylan et al.* (2008) e o capítulo 6 de *Naghettini e Pinto* (2007).

O método dos momentos consiste em *igualar os momentos amostrais aos momentos populacionais*. O resultado dessa operação fornece as estimativas dos parâmetros da distribuição de probabilidades em questão. Formalmente, sejam $\{x_1, x_2, ..., x_N\}$ as observações constituintes de uma *amostra aleatória simples* constituída a partir de uma população de uma variável aleatória com função densidade de probabilidade com k parâmetros, representada por $f_X(x;\theta_1,\theta_2,...,\theta_i,...,\theta_k)$ ou, numa anotação simplificada, por $f(x;\theta_1,\theta_2,..,\theta_i,..,\theta_k)$. Se μ_j e m_j representam, respectivamente, os momentos populacionais e amostrais, o sistema fundamental k equações a k incógnitas do método dos momentos é dado por:

$$\mu_i(\theta_1,\theta_2,..,\theta_i,..,\theta_k) = m_j \text{ com } i=1, 2,..,k \quad \text{.....(18)}$$

As soluções $\hat{\theta}_1,\hat{\theta}_2,..,\hat{\theta}_i,..,\hat{\theta}_k$ do anterior sistema de equações constituem as estimativas dos parâmetros θ_i pelo método dos momentos. Os exercícios a seguir exemplificam a aplicação de tal método.

Exercício 11 – Seja x_1, x_2, x_3, ... , x_N uma amostra aleatória simples retirada da população de uma variável aleatória X, cuja função densidade de probabilidade, com um único parâmetro, θ, é dada por $f_X(x;\theta) = (\theta+1)x^{\theta}$ para $0 \le x \le 1$. Pede-se para: (a) determinar o estimador de θ pelo método dos momentos; e (b) supondo-se que a amostra de X seja constituída pelos seguintes elementos {0.20; 0.90; 0.05; 0.47; 0.56; 0.80; 0.35}, calcular o valor do anterior estimador, ou seja, a estimativa de θ pelo método dos momentos, $\hat{\theta}$ e a probabilidade de X ser maior do que 0.8.

Solução: (a) De acordo com o método dos momentos, havendo apenas um parâmetro a estimar, então, o momento de ordem 1 fornecerá esse parâmetro, ou seja, $\mu_1=m_1$. De acordo com a equação (3), o primeiro momento populacional é dada por $\mu_1=E(X)=\int_0^1 x(\theta+1)x^{\theta}dx=(\theta+1)/(\theta+2)$ sendo que o primeiro momento amostral é a média da amostra, ou seja, $m_1=(1/N)\sum_{i=1}^{N} x_i=\overline{X}$. Logo, $(\hat{\theta}+1)/(\hat{\theta}+2)=\overline{X} \Rightarrow \hat{\theta}=(2\overline{X}-1)/(1-\overline{X})$. A última equação dá o estimador de θ pelo método dos momentos. (b) A amostra fornecida conduz a $\overline{X}=0.4757$. Entrando com este resultado na equação antes determinada para o estimador de θ, obtêm-se $\hat{\theta}=(2\times0.4757-1)/(1-0.4757)=-0.0926$. A função distribuição de probabilidade, FDP, é dada por $F_X(x)=F(x)=\int_0^x(\theta+1)x^{\theta}dx=x^{\theta+1}$. Logo P(X>0.8)=1-P(X≤0.8)=1-F (0.8)=1-0.8167=0.1833.

Exercício 12 – Considere a amostra de precipitações diárias máximas anuais no posto udométrico de Pavia (20I/01G) a que se refere a Tabela 1. Conforme se explicitou na Tabela 2, foram estimadas as seguintes estatísticas amostrais: média, $\overline{X}=39.5$ mm; desvio-padrão, $s_X=17.2$ mm; e coeficiente de assimetria, $g_X=1.149$. Determine: (a) os estimadores dos parâmetros da distribuição de Gumbel para máximos (ou simplesmente distribuição de Gumbel) pelo método dos momentos; e (b) as estimativas dos anteriores parâmetros. Calcule: (c) a probabilidade de a precipitação diária máxima anual ser superior a 65 mm; e (d) a precipitação diária máxima anual com o período de retorno de 100 anos.

Solução: (a) Conforme se explicita na Tabela 5, a distribuição de Gumbel é definida pelos parâmetros α e β, os quais se relacionam com os dois primeiros momentos da população pelas equações $\mu_1=E[X]=\beta+0.5772\,\alpha$ e

$\mu_2 = \mathrm{Var}[X] = \pi^2\alpha^2/6$. Igualando a variância da amostra, s_X^2, à da população, μ_2, e resolvendo a segunda das anteriores equações em ordem a α obtém-se o estimador $\hat{\alpha} = \sqrt{6}s_X/\pi$. Igualando a média da amostra, $\overline{X}$, à da população, μ_1, introduzindo o estimador $\hat{\alpha}$ na primeira das anteriores equações e resolvendo-a em ordem a β obtém-se o estimador $\hat{\beta} = \overline{X} - 0.5772\,\hat{\alpha}$. (b) $\hat{\alpha}$ e $\hat{\beta}$ constituem os estimadores pelo método dos momentos dos parâmetros da lei de Gumbel. As correspondentes estimativas obtêm-se muito simplesmente atendendo aos valores numéricos dos momentos amostrais, $\overline{X}$ e s_X, em conformidade com a amostra em estudo, do que resulta $\hat{\alpha} = \sqrt{6}s_X/\pi = \sqrt{6}\,17.2/\pi = 13.4$ e $\hat{\beta} = \overline{X} - 0.5772\,\hat{\alpha} = 39.5 - 0.5772\times 13.4 = 31.8$. (c) Atendendo a que P(X>65)=1-F(65), bem como à equação de quantis da lei de Gumbel (apresentada na Tabela 5) e às anteriores estimativas dos parâmetros α e β dessa distribuição, obtém-se que P(X>65)=1-F(65)= =1-0.9196=0.0804. (d) Para o período de retorno T=100 anos, correspondente à probabilidade de não-excedência de $F(x_T)=0.99$, a função de quantis da Tabela 5 fornece x_{100}=93.5 mm. Nota: o anterior procedimento de cálculo pode ser estendido às distribuições log-Normal ou de Galton e GEV, a partir das equações de momentos e de quantis da Tabela 5. No caso particular da distribuição GEV, pode usar-se a Figura 8, para uma primeira estimativa do parâmetro de forma κ a partir da estimativa do coeficiente de assimetria g=1.149. Em seguida ou alternativamente, pode obter-se uma maior precisão na estimativa de κ, com base em aproximações sucessivas, mediante uso da função LNGAMA/GAMMLN do *software* Microsoft Excel, referida a propósito daquela figura.

8.2 Factores de probabilidade

Uma abordagem, introduzida por *Chow* (1954), e que facilita muito o cálculo dos quantis, refere-se à utilização dos *factores de probabilidade*. Segundo essa abordagem, o quantil x_T, da variável aleatória X, para a probabilidade de não-excedência, F, ou, de modo equivalente para o correspondente período de retorno, T tal que T=1/(1-F), pode ser estimado através de:

$$x_T = \mu_X + K_{DIST}^F\,\sigma_X \quad \text{.....(19)}$$

em que K_{DIST}^F denota o *factor de probabilidade*, dependente de F e da distribuição estatística para a qual se pretende estimar quantis. Se a média e o desvio-padrão populacionais, a saber, μ_X e σ_X, forem substituídos pelas suas respectivas estimativas amostrais, $\overline{X}$ e s_X, a abordagem passa a ser uma *extensão do método dos momentos* e a equação (19) toma a forma:

$$x_T = \overline{X} + K_{DIST}^F\,s_X \quad \text{.....(20)}$$

a qual exprime o facto de os produtos dos factores de probabilidade pelo desvio-padrão, s_X, representam desvios crescentes, em relação à média amostral, $\overline{X}$, à medida que as probabilidades de não-excedência e, consequentemente, os períodos de retorno, aumentam.

A Tabela 7 apresenta as equações mais vulgares para cálculo dos factores de probabilidade para as distribuições Normal, log-Normal, de Gumbel, GEV, Pearson III e log-Pearson III. Observa-se que, no caso das leis Normal e log-Normal o factor de probabilidade para dado valor da probabilidade de não-excedência, F, ou, de modo equivalente, para o valor correspondente do período de retorno, T, é igual ao valor da normal reduzida para esse valor de F, z, conforme se sistematizou na Tabela 7. Anota-se que as equações da Tabela 7 fornecem exactamente os mesmos resultados para os quantis estimados pelas funções de quantis da Tabela 5.

Tabela 7 – Expressões de cálculo dos factores de probabilidade K_{DIST}^{F} para diversas distribuições.

Distribuição (DIST)	Factor de probabilidade (K_{DIST}^{F})	Equação de quantis (x_F)	Observação
Normal	$K_{Normal}^{F} = z(F)$	$x_F = \overline{X} + K_{Normal}^{F}\, s_X$	Z(F): Tabela 6
log-Normal ou de Galton		$x_F = \exp\left(\overline{Y} + K_{Normal}^{F}\, s_Y\right)$ com $Y = \ln(X)$	Z(F): Tabela 6
Gumbel	$K_{Gumbel}^{F} \cong -\frac{\sqrt{6}}{\pi}\{0.577216 + \ln[\ln(1/F)]\}$ [rigorosamente, K_{Gumbel}^{F} depende da dimensão da amostra, N, Kite (1988)]	$x_F = \overline{X} + K_{Gumbel}^{F}\, s_X$	-------
GEV	$K_{GEV}^{F} = \left(\frac{\kappa}{\lvert\kappa\rvert}\right)\frac{\{\Gamma(1+\kappa) - [-\ln(F)]^{\kappa}\}}{\sqrt{\Gamma(1+2\kappa) - \Gamma^2(1+\kappa)}}$	$x_F = \overline{X} + K_{GEV}^{F}\, s_X$	-------
Pearson-III	Transformação de Wilson-Hilferty $K_{Pearson}^{F} \cong \frac{2}{g_X}\left\{\left[\left(K_{Normal}^{F} - \frac{g_X}{6}\right)\frac{g_X}{6} + 1\right]^3 - 1\right\}$ Alternativa $K_{Pearson}^{F} \cong K_{Normal}^{T} + (K_{Normal}^{T\,2} - 1)\,k + \frac{1}{3}(K_{Normal}^{T\,3} - 6\,K_{Normal}^{T})\,k^2 - (K_{Normal}^{T\,2} - 1)\,k^3 + K_{Normal}^{T}\,k^4 + \frac{1}{3}k^5$	$x_F = \overline{X} + K_{Pearson}^{F}\, s_X$	Z(F): Tabela 6 Na transformação de Wilson-Hilferty $\lvert g_X\rvert < 2$. Para outras assimetrias consultar Rao e Hamed (2000) Na equação alternativa $k = \frac{g_x}{6}$
log-Pearson III	Transformação de Wilson-Hilferty $K_{Pearson}^{F} \cong \frac{2}{g_Y}\left\{\left[\left(K_{Normal}^{F} - \frac{g_Y}{6}\right)\frac{g_Y}{6} + 1\right]^3 - 1\right\}$ Alternativa $K_{Pearson}^{F} \cong K_{Normal}^{T} + (K_{Normal}^{T\,2} - 1)\,k + \frac{1}{3}(K_{Normal}^{T\,3} - 6\,K_{Normal}^{T})\,k^2 - (K_{Normal}^{T\,2} - 1)\,k^3 + K_{Normal}^{T}\,k^4 + \frac{1}{3}k^5$	$x_F = \exp\left(\overline{Y} + K_{Pearson}^{F}\, s_Y\right)$ com $Y = \ln(X)$	Z(F): Tabela 6 Na transformação de Wilson-Hilferty $\lvert g_Y\rvert < 2$ Para outras assimetrias consultar Rao e Hamed (2000) Na equação alternativa $k = \frac{g_y}{6}$

Exercício 13 – Estime a precipitação média diária máxima anual com o período de retorno de 100 anos a que se refere a alínea (d) do exercício 12 no pressuposto de aplicação da lei de Pearson III.

Solução: Conforme se especificou no exercício 12, as estatísticas amostrais são $\overline{X} = 39.5$ mm, $s_X = 17.2$ mm e $g_X = 1.149$. Assim, recorrendo ao factor de probabilidade e às expressões pertinentes da Tabela 7, obtém-se sucessivamente: T=100 anos; F=1-1/T=0.99; $K_{Normal}^{0.99}$=z(0.99)=2.326 (Tabela 6); $K_{Pearson}^{0.99} = 3.1266$. Portanto, a precipitação diária máxima anual com o período de retorno de 100 anos de acordo com a lei de Pearson III é dada por: $x_T = \overline{X} + K_{Pearson}^{100}\, s_X = 39.5 + 3.1266 \times 17.2 = 93.3$ mm.

9. Análise de frequência de variáveis hidrológicas

9.1 Nota prévia

A análise de frequência de amostras de variáveis hidrológicas tem por objectivo estimar valores dessas variáveis para dadas probabilidades de não-excedência, F, ou, de modo equivalente, para dados períodos de retorno, T, adoptados como critério de projecto para o que utiliza distribuições de probabilidade supostamente capazes de descrever as variáveis. Os resultados de tal análise intervêm na solução de inúmeros problemas da engenharia hidráulica e não só, tais como a caracterização das ocorrências extremas associadas a cheias e a secas; o projecto de descarregadores de cheias de barragens; o dimensionamento de albufeiras de regularização, de diques de protecção marginal ao longo dos cursos de água ou de obras de drenagem de vias de comunicação; o projecto de pontes, por exemplo, no que respeita à fixação do vão livre ou da cota do tabuleiro ou, ainda, o estudo das erosões em torno dos pilares; etc.

As amostras utilizadas na análise de frequência devem ser *representativas* da variável a que se referem, não apresentando erros de observação ocasionais e/ou sistemáticos[1], devendo ter um número suficiente de elementos que permita realizar *extrapolações merecedoras de confiança*. Além disso, é necessário assegurar que se tratam de amostras aleatórias simples, ou seja, que os dados são *homogéneos*[2] e *independentes*, além de 'sorteados' ao acaso.

A condição de homogeneidade pretende assegurar que todas as observações tenham sido extraídas de uma mesma população, descrita por uma única distribuição de probabilidades. Por exemplo, para o caso de análise de escoamento, em condições de cheia ou não, pretende-se assegurar que o uso e a ocupação da bacia hidrográfica não tenham sido significativamente modificados ou, ainda, que não tenham sido implantadas estruturas hidráulicas que tenham alterado o regime do escoamento natural. Por outro lado, a condição de independência procura assegurar que não existe dependência serial entre os elementos que constituem a amostra, tornando-a apta a ser analisada mediante aplicação de procedimentos da análise estatística. Os *testes estatísticos de significância* para verificar a adequação das amostras aos anteriores requisitos encontram-se descritos no capítulo 7 de *Naghettini e Pinto* (2007).

9.2. Análise de frequência com base na apreciação visual do ajustamento (em gráficos de probabilidade). Probabilidade empírica de não-excedência

Para proceder à análise de frequência de uma amostra, concretamente, para identificar as distribuições estatísticas susceptíveis de serem aplicadas a essa amostra é frequente recorrer-se ao ajustamento visual, tendo por base a representação gráfica dos pontos da amostra e das leis

[1] Uma amostra de uma variável aleatória é consistente se, ao longo do respectivo período de observação, não existe alteração do erro sistemático de medição da grandeza a que se refere a amostra. Constituem exemplos de quebra de consistência a mudança de local do aparelho de medição da precipitação (udómetro) ou a criação de obstáculos junto ao mesmo ou o incorrecto nivelamento na mudança do sistema de registos de alturas ou níveis hidrométricos (Quintela, 1996).

[2] Uma amostra de uma variável hidrológica diz-se homogénea quando, ao longo do respectivo período de observação, não existirem alteração nos factores que condicionam o fenómeno traduzido pela grandeza a que se refere a amostra. No pressuposto de que, à escala do tempo abrangido pela amostra, não ocorreram mudanças climáticas, as quebras de homogeneidade, a registarem-se, devem-se a alterações em factores físicos, tais como os associados à desflorestação ou ainda os decorrentes da construção de barragens. Em certas circunstâncias, é possível eliminar uma quebra de homogeneidade, procedendo à reconstituição da amostra natural (Quintela, 1996).

teóricas postuladas para representar essa amostra. Para o efeito, é necessário atribuir a cada ponto da amostra uma probabilidade empírica de não-excedência, *F* (na designação inglesa, *plotting position)*. Em geral, o ajustamento gráfico utiliza os designados *papéis de probabilidade*, nos quais os eixos das ordenadas estão graduados nas unidades dos elementos das amostra e os eixos das abcissas, em *escalas transformadas de probabilidades*, tais que, para a lei a que se refere cada um desses papéis, a relação entre os valores da variável aleatória e as respectivas probabilidades teóricas de não-excedência é linear. Os principais papéis de probabilidade referem-se às distribuições Exponencial, Normal, log-Normal e de Gumbel, e todos assentam no mesmo princípio: escala das abcissas de modo a linearizar a mencionada relação para a *distribuição de probabilidades* a que se refere o papel.

A Figura 9 exemplifica o papel de probabilidades Normal sendo que o segmento de recta aí representada fornece as probabilidades de não-excedência para os valores da amostra a que se refere o eixo das ordenadas. No caso da lei Normal, a linearização da relação resulta muito simplesmente de atribuir a cada estimativa da variável aleatória o valor da normal reduzida para a probabilidade de não-excedência correspondente a essa estimativa. Para melhor elucidar o conceito de papel de probabilidade incluíram-se na Figura 9, por assim dizer, três eixos das abcissas: dois na parte inferior do gráfico – um linear em valores da normal reduzida, z – e outro, com os valores correspondentes da probabilidade de não-excedência, F, a qual, no eixo superior foi transcrita em termos dos períodos de retorno, T, que lhe correspondem.

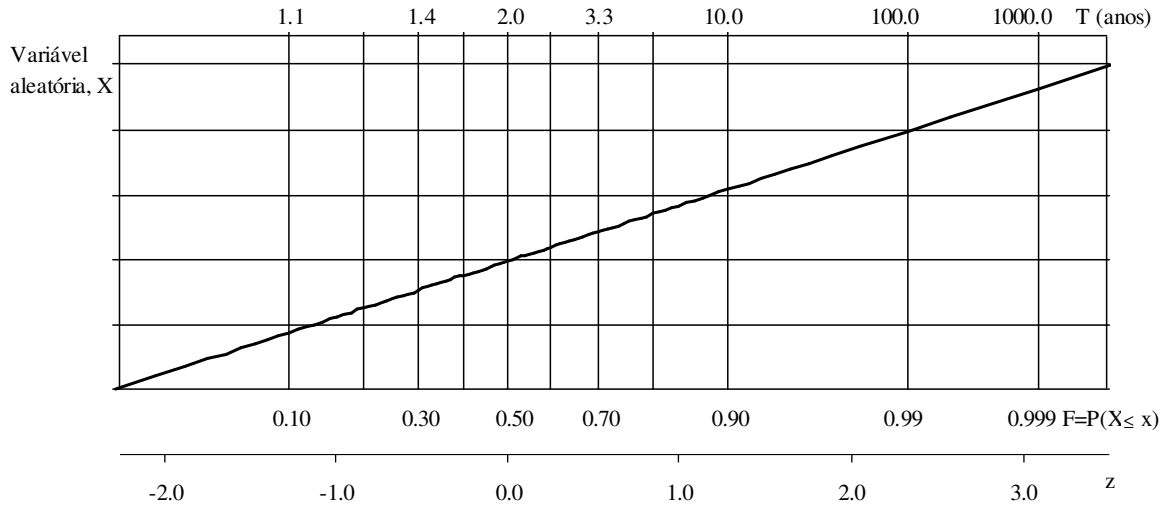

Figura 9 – Papel de probabilidade da lei Normal.

Como mencionado, na representação em papel de probabilidade, a cada valor de uma amostra é associada uma *probabilidade empírica de não-excedência*, F (*plotting position*, como antes especificado). Se a amostra representasse toda a população, a probabilidade de não-excedência associada a cada elemento seu seria dada pelo quociente entre o número de elementos da amostra inferiores ou iguais ao considerado e a dimensão da amostra, N (ou seja, seria a fracção dos elementos da amostra com valor inferior ou igual a cada elemento seu).

Numa amostra sem valores repetidos, se i representasse o número de ordem de um dado elemento após ordenação dos elementos dessa amostra por valores crescentes, tal probabilidade

seria simplesmente dada por i/N. De acordo com essa noção, a probabilidade de ocorrerem elementos com valor, tanto inferior ao elemento com menor valor da amostra, como superior ao elemento com maior valor da amostra seria nula (acontecimentos impossíveis). Em face de amostras finitas representativas de populações infinitas, o pressuposto de que nunca poderão ocorrer elementos com valores para além da gama de valores patente na amostra não tem sentido. Surgiram, assim, fórmulas de estimação de probabilidades empíricas que corrigem esse pressuposto. Tais fórmulas fazem intervir o número de ordem i de cada elemento da amostra, após ordenação dos elementos da mesma por valores crescentes (i igual a 1 para o menor valor da amostra e igual a N, para o maior valor) e são frequentemente casos particulares da seguinte fórmula geral, em que i e N têm os significados antes especificados e ω é uma constante compreendida entre 0 e 1 e que determina a qualidade do ajustamento entre probabilidades empíricas e teóricas de acordo com as leis postuladas:

$$F = P(X \leq x) = \frac{i - \omega}{n + 1 - 2\,\omega} \quad \text{.....(21)}$$

A fórmula a aplicar deve atender à distribuição teórica que se supõe ser válida para a população de onde provém a amostra em estudo. A Tabela 8 apresenta algumas das fórmulas de cálculo de probabilidades empíricas de não-excedência, os correspondentes valores de ω e recomendações quanto à sua aplicabilidade.

Tabela 8 – Fórmulas para estimação de probabilidades empíricas de não-excedência.

Fórmula	Autor	Valor de ω. Atributos de aplicação
$F = \frac{i}{N+1}$	Weibull	ω=0.000. Probabilidades de excedência não enviesadas para todas as distribuições
$F = \frac{i-0.44}{N+0.12}$	Gringorten	ω=0.440. Usada para quantis das distribuições de Gumbel, GEV e Weibull
$F = \frac{i-0.375}{N+0.25}$	Blom	ω=0.375. Quantis não enviesados para as distribuições Normal e Log-Normal
$F = \frac{i-0.5}{N}$	Hazen	ω=0.500. Usada para quantis da distribuição Pearson III
$F = \frac{i-0.40}{N+0.20}$	Cunnane	ω=0.400. Quantis aproximadamente não enviesados para todas as distribuições

Na Figura 10 comparam-se as probabilidades empíricas de não-excedência obtidas pelas fórmulas da Tabela 8 para duas amostras, uma com 50 elementos (gráfico do lado esquerdo) e outra com 20 elementos (gráfico do lado direito). À semelhança do papel de probabilidade da lei Normal, os eixos das abcissas de ambos os gráficos foram graduados numa escala linear de valores da normal reduzida. Como se pode observar, os resultados fornecidos pelas diferentes fórmulas apenas surgem diferenciados (pontos representativos das diferentes probabilidades nitidamente não coincidentes) para probabilidades extremas – muito baixas ou muito elevadas –, distinguindo-se tanto mais, quanto menor a dimensão da amostra a que respeitam.

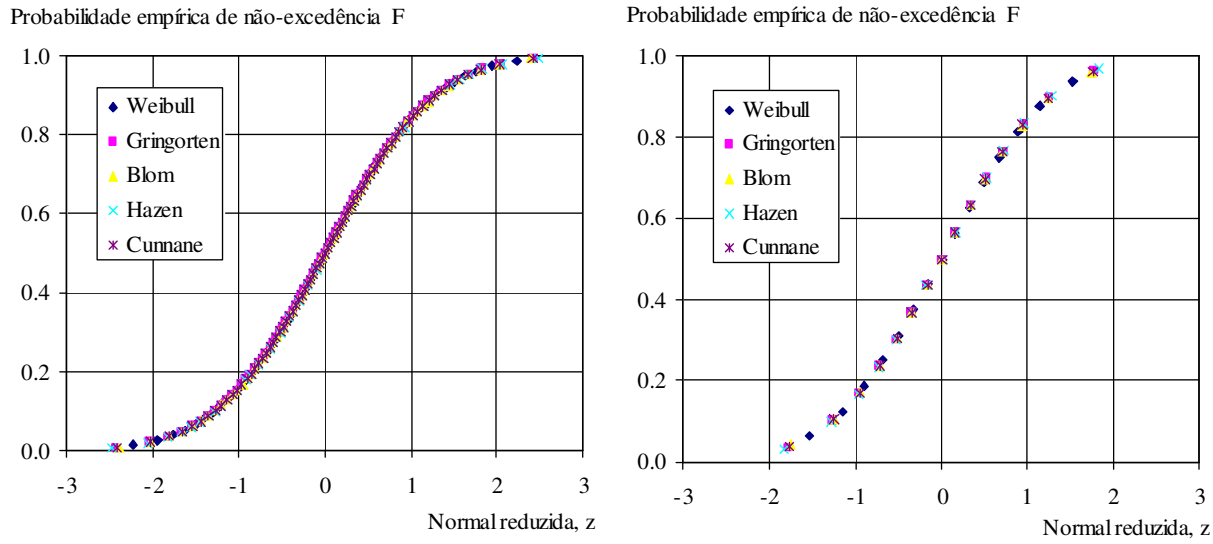

Figura 10 – Probabilidades empíricas de não-excedência fornecidas pelas fórmulas da Tabela 8 para duas amostras, uma, com 50 elementos (à esquerda) e, outra, com 20 elementos (à direita).

Sistematizam-se, seguidamente, as etapas requeridas pela representação, para uma dada amostra, da distribuição das probabilidades empíricas de não-excedência:

(i) ordenação dos valores da amostra por valores crescentes;

(ii) atribuição, a cada valor já ordenado, x_i, da respectiva probabilidade empírica de não-excedência, F_i por aplicação de uma das fórmulas da Tabela 8;

(iii) selecção de um tipo de papel de probabilidades consoante a expectativa da lei com melhor ajuste (exponencial, Normal, log-Normal ou Gumbel), embora, desconhecendo-se tal lei, se possa adoptar o papel de probabilidades da lei Normal;

(iv) representação gráfica dos pares de valores (F_i, x_i).

A Tabela 9 e a Figura 11 exemplificam a estimação da distribuição empírica das precipitações diárias máximas amuais no posto udométrico de Pavia (20I/01G) a que se refere a Tabela 1, mediante o recurso à fórmula de Gringorten e aos papéis de probabilidade das leis Normal (gráfico superior) e de Gumbel (gráfico inferior).

Nos gráficos da Figura 11 os eixos das abcissas são lineares tendo sido completados por um segundo eixo secundário, no topo de cada gráfico, graduado em probabilidades de não-excedência, F. Incluíram-se, nos gráficos, as curvas resultantes do ajustamento das distribuições Normal, de Gumbel e log-Normal aos pontos da amostra. Tais curvas foram calculadas recorrendo à técnica dos factores de probabilidade, conforme antes descrito. Como resulta da observação dos gráficos, no papel de probabilidade da lei Normal (gráfico superior) tal lei é representada por um segmento de recta, acontecendo outro tanto com a lei de Gumbel, quando é utilizado o papel dessa lei (gráfico inferior).

Tabela 9 – Precipitações diárias máximas anuais no posto udométrico de Pavia, de acordo com a Tabela 1. Probabilidades empíricas de não-excedência, P(X≤ x)=F(x), de acordo com a fórmula de Gringorten apresentada na Tabela 8.

i	Pdma (mm)	P(X≤ x)= =F(x)	*i*	Pdma (mm)	P(X≤ x)= =F(x)	*i*	Pdma (mm)	P(X≤ x)= =F(x)	*i*	Pdma (mm)	P(X≤ x)= =F(x)	*i*	Pdma (mm)	P(X≤ x)= =F(x)
1	8.1	0.0059	20	27.5	0.2078	39	34.0	0.4097	58	38.9	0.6116	77	50.4	0.8134
2	10.2	0.0166	21	27.8	0.2184	42	34.2	0.4416	61	40.2	0.6434	78	52.0	0.8241
3	10.3	0.0272	22	28.0	0.2291	42	34.2	0.4416	61	40.2	0.6434	79	55.2	0.8347
4	14.2	0.0378	23	28.5	0.2397	42	34.2	0.4416	61	40.2	0.6434	80	56.8	0.8453
5	15.3	0.0484	24	29.0	0.2503	43	34.6	0.4522	62	40.5	0.6541	81	57.0	0.8559
6	18.2	0.0591	25	29.4	0.2609	45	35.2	0.4734	63	41.2	0.6647	82	58.0	0.8666
8	20.2	0.0803	26	29.5	0.2716	45	35.2	0.4734	64	42.8	0.6753	83	58.2	0.8772
8	20.2	0.0803	28	29.8	0.2928	46	35.7	0.4841	65	43.2	0.6859	84	59.6	0.8878
9	20.4	0.0909	28	29.8	0.2928	47	36.2	0.4947	66	43.7	0.6966	85	60.2	0.8984
10	20.8	0.1016	29	30.0	0.3034	48	36.5	0.5053	67	43.8	0.7072	86	63.3	0.9091
11	24.2	0.1122	30	31.3	0.3141	50	36.7	0.5266	68	44.0	0.7178	87	69.0	0.9197
13	24.3	0.1334	32	31.4	0.3353	50	36.7	0.5266	70	45.0	0.7391	88	70.2	0.9303
13	24.3	0.1334	32	31.4	0.3353	51	37.2	0.5372	70	45.0	0.7391	89	71.4	0.9409
14	25.2	0.1441	33	31.9	0.3459	52	37.4	0.5478	71	46.3	0.7497	90	80.0	0.9516
15	26.0	0.1547	34	32.5	0.3566	53	37.5	0.5584	72	46.6	0.7603	92	84.2	0.9728
16	27.0	0.1653	36	32.8	0.3778	54	38.0	0.5691	73	47.0	0.7709	92	84.2	0.9728
17	27.2	0.1759	36	32.8	0.3778	55	38.2	0.5797	74	48.4	0.7816	93	92.3	0.9834
19	27.4	0.1972	37	33.2	0.3884	56	38.4	0.5903	75	48.5	0.7922	94	95.5	0.9941
19	27.4	0.1972	38	33.5	0.3991	57	38.6	0.6009	76	49.0	0.8028			

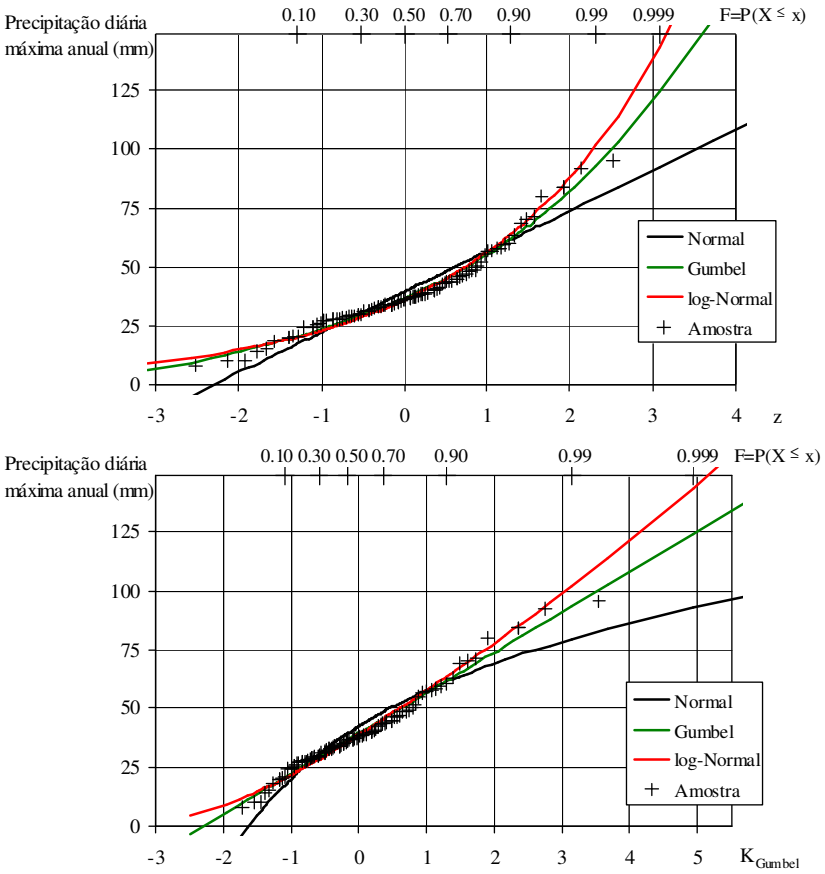

Figura 11 – Precipitações diárias máximas anuais no posto udométrico de Pavia, de acordo com a Tabela . Probabilidades de não-excedência empíricas (fórmula de Gringorten) e de acordo com as leis Normal, de Gumbel e log-Normal para papeis de probabilidade das leis Normal – gráfico superior – e de Gumbel – gráfico inferior.

9.3. Apreciação da qualidade do ajustamento e escolha do modelo distributivo. Teste de Kolmogorov-Smirnov e do Qui-Quadrado

Em face de uma dada amostra, a representação gráfica em papel de probabilidade das distribuições, por um lado, empírica e, por outro lado, teóricas referentes às leis se afiguram capazes de representar aquela amostra permite *avaliar visualmente a adequação de cada uma daquelas leis à amostra* e, assim, apreciar a *qualidade do ajustamento* de um dado modelo distributivo teórico relativamente a outro(s), principalmente no ramo das curvas de frequência que maiores consequências poderão ter nas decisões de engenharia: cauda superior, para máximos e para valores extremos, e cauda inferior, para mínimos.

A opção por um modelo distributivo apela a alguma *prudência* e a um certo *conservadorismo*, do que deve resultar a escolha, em circunstâncias praticamente equivalentes, do modelo mais exigente em termos de valores de projecto, facto tanto mais justificável quanto as decisões de engenharia requerendo a análise de probabilidades contêm incertezas intrínsecas. Outra importante preocupação na comparação de modelos probabilísticos refere-se ao número dos respectivos parâmetros. Em geral, os modelos de três parâmetros apresentam maior flexibilidade e, com isso, maior adequação ou "aderência" aos pontos das amostras. Entretanto, a maior aderência é obtida à custa de um terceiro parâmetro, cuja estimação a partir da amostra, introduz incertezas adicionais. Se não há grande diferença entre os quantis dos modelos de dois ou de três parâmetros, deve ser dada preferência ao modelo com o menor número de parâmetros, a despeito da sua relativamente menor aderência aos dados – princípio da *parcimónia de parâmetros*.

Além da apreciação visual baseada na representação gráfica em papel de probabilidades, existem diversos testes estatísticos de significância aplicáveis à avaliação da qualidade do ajustamento de um modelo distributivo teórico a uma certa amostra os quais, em linhas gerais, verificam se os dados dessa amostra são compatíveis com aquele modelo.

Os testes mais conhecidos são os *testes de aderência* ou *de ajustamento* do Qui-Quadrado, de Kolmogorov-Smirnov, de Anderson-Darling e de Filliben. Embora propiciem uma avaliação quantitativa do grau de aderência, estes testes apresentam as seguintes deficiências: (i) não são objectivamente decisivos no que respeita à qualidade do ajustamentos das caudas superiores das distribuições de valores máximos, onde, em geral existem poucos pontos amostrais; e (ii) não foram concebidos para comparar, em termos relativos e por meio das suas estatísticas, as diferentes distribuições teóricas aplicadas a uma dada amostra.

O presente item aborda apenas a aplicação dos testes de Kolmogorov-Smirnov, KS, e do Qui-Quadrado, χ^2, ao ajustamento de leis teóricas a amostras. Ao leitor interessado noutros testes e meios para avaliar a qualidade do ajustamento (diagramas de momentos convencionais e de momentos-L), recomenda-se a consulta das referências *Rao e Hamed* (2000), *Meylan et al.* (2008) e do capítulo 7 de *Naghettini e Pinto* (2007).

Os testes de ajustamento "confrontam" (por meio "operadores" designados por *estatísticas dos testes*) a informação contida numa amostra com a que decorre do pressuposto de uma função de distribuição de probabilidades, mediante a análise da chamada *hipótese nula* (H_0) de que o modelo distributivo teórico se ajusta bem aos pontos daquela amostra e que as diferenças encontradas são fortuitas, ou seja, decorrentes de meras flutuações amostrais, não sendo, portanto, *estatisticamente significativas*.

Uma de duas decisões resulta do anterior confronto: a de *'não rejeitar'* ou a de *'rejeitar'* a veracidade da hipótese H_0 de a lei teórica postulada se ajustar aos pontos da amostra. Importa realçar que o teste nunca permite "aceitar" tal lei teórica uma vez que a decisão de "não rejeitar" implica apenas que *não existem elementos significativos que invalidem a hipótese nula* H_0.

Na aplicação de um teste de ajustamento é necessário fixar *a priori* um certo *nível de significância*, α, ou seja, a probabilidade, por regra pequena – entre 1 e 5% –, de se tomar uma decisão incorrecta (rejeitar H_0 ajustando-se bem o modelo distributivo). Ao complementar do *nível de significância*, α, ou seja, a (1-α) atribuiu-se a designação de *nível de confiança*.

A *estatística* do teste de ajustamento de Kolmogorov-Smirnov, KS, é dada pela *máxima diferença* entre as funções de probabilidades acumuladas empírica e teórica de variáveis aleatórias contínuas. O teste não é aplicável a variáveis aleatórias discretas.

Considere-se que X representa uma variável aleatória contínua, de cuja população se extraiu a amostra $\{x_1, x_2, ..., x_N\}$. A *hipótese nula* a ser testada é dada por $H_0 : P(X \le x) = F_X(x) = F(x)$, ou seja, pretende-se averiguar se F(x) é uma distribuição de probabilidade adequada à descrição do comportamento probabilístico da variável X. Para implementar o teste KS, classificam-se os elementos da amostra $\{x_1, x_2, ..., x_N\}$ por *ordem crescente*, de modo a constituir a sequência $\{x_{(1)}, x_{(2)}, ..., x_{(m)}, \ldots, x_{(N)}\}$ na qual $1 \le m \le N$ denota a ordem de classificação. Para cada elemento $x_{(m)}$ a distribuição empírica é fornecida pela proporção de valores amostrais inferiores ou iguais a $x_{(m)}$, ou seja, é igual a m/N. Para tal elemento calcula-se também a respectiva probabilidade de não-excedência teórica, $F(x_{(m)})$, aplicando os métodos anteriormente descritos, por ventura, baseados na inversão, em ordem à variável aleatória, das equações que utilizam o factor de probabilidade. Os anteriores cálculos são efectuados para os sucessivos valores $x_{(m)}$, A *estatística do teste* KS, D, é dada por

$$D_N = \max_{-\infty < x < \infty} \left| m/N - F\left(x_{(m)}\right) \right| \quad \text{.....(22)}$$

correspondendo, portanto, ao valor absoluto da *maior diferença* entre as probabilidades empírica e teórica.

Se H_0 é verdadeira, quando $N \to \infty$, a estatística D_N tenderá para zero. Por outro lado, se N é um valor finito, a estatística D_N deverá ser da ordem de grandeza de $1/\sqrt{N}$ e, portanto, a quantidade $\sqrt{N}\, D_N$ não irá tender a zero, mesmo para valores muito elevados de N. Para amostras com dimensão superior a 40, os *valores críticos* da estatística de teste D_N são $1.3581/\sqrt{N}$, para o nível de significância de α=0.05, e $1.6276/\sqrt{N}$, para α=0.01. Para amostras com dimensão inferior a 40, os valores críticos de D_N devem ser obtidos na Tabela 10. Se a estatística calculada pela equação (21) for *maior do que o valor crítico* tabelado, as diferenças são, de facto, significativas para o nível de significância α e, portanto, a decisão é a de *rejeitar a hipótese* H_0. Em caso contrário, a hipótese nula não deve ser rejeitada.

Tabela 10 - Valores críticos da estatística do teste de Kolmogorov-Smirnov em função da dimensão da amostra, N, e do nível do significância, α, $D_{N,\alpha}$.

N	$D_{N,\ 0.10}$	$D_{N,\ 0.05}$	$D_{N,\ 0.02}$	$D_{N,\ 0.01}$	N	$D_{N,\ 0.10}$	$D_{N,\ 0.05}$	$D_{N,\ 0.02}$	$D_{N,\ 0.01}$
10	0.369	0.409	0.457	0.489	26	0.233	0.259	0.290	0.311
11	0.352	0.391	0.437	0.468	27	0.229	0.254	0.284	0.305
12	0.338	0.375	0.419	0.449	28	0.225	0.250	0.279	0.300
13	0.325	0.361	0.404	0.432	29	0.221	0.246	0.275	0.295
14	0.314	0.349	0.390	0.418	30	0.218	0.242	0.270	0.290
15	0.304	0.338	0.377	0.404	31	0.214	0.238	0.266	0.285
16	0.295	0.327	0.366	0.392	32	0.211	0.234	0.262	0.281
17	0.286	0.318	0.355	0.381	33	0.208	0.231	0.258	0.277
18	0.279	0.309	0.346	0.371	34	0.205	0.227	0.254	0.273
19	0.271	0.301	0.337	0.361	35	0.202	0.224	0.251	0.269
20	0.265	0.294	0.329	0.352	36	0.199	0.221	0.247	0.265
21	0.259	0.287	0.321	0.344	37	0.196	0.218	0.244	0.262
22	0.253	0.281	0.314	0.337	38	0.194	0.215	0.241	0.258
23	0.247	0.275	0.307	0.330	39	0.191	0.213	0.238	0.255
24	0.242	0.269	0.301	0.323	40	0.189	0.210	0.235	0.252
25	0.238	0.264	0.295	0.317	>40	$1.22/\sqrt{N}$	$1.36/\sqrt{N}$	$1.52/\sqrt{N}$	$1.63/\sqrt{N}$

Na aplicação do teste do Qui-Quadrado, χ^2, o domínio da função de distribuição é dividido em M intervalos de partição sendo que o teste compara os números de elementos da amostra efectivamente contidos nos sucessivos intervalos com as esperanças matemáticas, ou seja, com os valores esperados, dos números desses elementos, avaliados em conformidade com o modelo postulado. A *estatística do teste* χ^2 é definida por:

$$\chi^2 = \sum_{j=1}^{M} \frac{(O_j - E_j)^2}{E_j} \quad \text{.....(23)}$$

em que O_j é o número de elementos da amostra efectivamente contidos no intervalo j e E_j, o valor esperado do número de elementos no mesmo intervalo j, dado por $E_j = N\,P_j$ em que P_j é a amplitude do intervalo j expressa em probabilidade e N, a dimensão da amostra.

O teste estatístico pode formular-se do seguinte modo: rejeitar H_0 com um *nível de confiança* (1-α) se $\chi^2 > \chi^2_{(1-\alpha)}$, em que $\chi^2_{(1-\alpha)}$ é o quantil (1-α) da distribuição χ^2 – Tabela 11.

Os valores da estatística χ^2 dependem do número de limites, M, e dos limites dos intervalos de partição do domínio da função de distribuição de probabilidade, F. Não existem, contudo, regras para seleccionar o número de intervalos e a amplitude de cada intervalo. Mann e Wald (1942), citados em Henriques (1990), recomendam a partição dos M intervalos de modo a que as probabilidades associadas a cada intervalo sejam idênticas. Sendo M o número de intervalos, os limites de cada intervalo devem ser definidos por forma a se ter $E_j = N/M$ (j=1, 2, …, M). Atendendo a este critério, a *estatística do teste* χ^2 simplifica-se para:

$$\chi^2 = \frac{M}{N} \sum_{j=1}^{M} O_j^2 - N \quad \text{.....(24)}$$

Na Tabela 12 apresentam-se as partições da função de distribuição de probabilidade, $F_X(x)$ ou $F(x)$, em função da dimensão da amostra, N, sugeridas por Henriques (1990).

Tabela 11 – Quantis da distribuição do Qui-Quadrado em função do número de graus de liberdade, ν, e do nível de confiança, $(1-\alpha)$, $\chi^2_{\nu,(1-\alpha)}$..

Graus de liberdade, ν	Nível de significância, α									
	0.995	0.975	0.900	0.500	0.100	0.050	0.025	0.010	0.005	0.001
	Nível de confiança, 1- α									
	0.005	0.025	0.100	0.500	0.900	0.950	0.975	0.990	0.995	0.999
1	0.000	0.001	0.016	0.455	2.706	3.841	5.024	6.635	7.879	10.827
2	0.010	0.051	0.211	1.386	4.605	5.991	7.378	9.210	10.597	13.815
3	0.072	0.216	0.584	2.366	6.251	7.815	9.348	11.345	12.838	16.266
4	0.207	0.484	1.064	3.357	7.779	9.488	11.143	13.277	14.860	18.466
5	0.412	0.831	1.610	4.351	9.236	11.070	12.832	15.086	16.750	20.515
6	0.676	1.237	2.204	5.348	10.645	12.592	14.449	16.812	18.548	22.457
7	0.989	1.690	2.833	6.346	12.017	14.067	16.013	18.475	20.278	24.321
8	1.344	2.180	3.490	7.344	13.362	15.507	17.535	20.090	21.955	26.124
9	1.735	2.700	4.168	8.343	14.684	16.919	19.023	21.666	23.589	27.877
10	2.156	3.247	4.865	9.342	15.987	18.307	20.483	23.209	25.188	29.588
11	2.603	3.816	5.578	10.341	17.275	19.675	21.920	24.725	26.757	31.264
12	3.074	4.404	6.304	11.340	18.549	21.026	23.337	26.217	28.300	32.909
13	3.565	5.009	7.041	12.340	19.812	22.362	24.736	27.688	29.819	34.527
14	4.075	5.629	7.790	13.339	21.064	23.685	26.119	29.141	31.319	36.124
15	4.601	6.262	8.547	14.339	22.307	24.996	27.488	30.578	32.801	37.698
16	5.142	6.908	9.312	15.338	23.542	26.296	28.845	32.000	34.267	39.252
17	5.697	7.564	10.085	16.338	24.769	27.587	30.191	33.409	35.718	40.791
18	6.265	8.231	10.865	17.338	25.989	28.869	31.526	34.805	37.156	42.312
19	6.844	8.907	11.651	18.338	27.204	30.144	32.852	36.191	38.582	43.819
20	7.434	9.591	12.443	19.337	28.412	31.410	34.170	37.566	39.997	45.314
21	8.034	10.283	13.240	20.337	29.615	32.671	35.479	38.932	41.401	46.796
22	8.643	10.982	14.041	21.337	30.813	33.924	36.781	40.289	42.796	48.268
23	9.260	11.689	14.848	22.337	32.007	35.172	38.076	41.638	44.181	49.728
24	9.886	12.401	15.659	23.337	33.196	36.415	39.364	42.980	45.558	51.179
25	10.520	13.120	16.473	24.337	34.382	37.652	40.646	44.314	46.928	52.619
26	11.160	13.844	17.292	25.336	35.563	38.885	41.923	45.642	48.290	54.051
27	11.808	14.573	18.114	26.336	36.741	40.113	43.195	46.963	49.645	55.475
28	12.461	15.308	18.939	27.336	37.916	41.337	44.461	48.278	50.994	56.892
29	13.121	16.047	19.768	28.336	39.087	42.557	45.722	49.588	52.335	58.301
30	13.787	16.791	20.599	29.336	40.256	43.773	46.979	50.892	53.672	59.702

Tabela 12 – Partições (número e limites) do domínio da função distribuição de probabilidade, F(x), na aplicação do teste do Qui-Quadrado em função da dimensão da amostra, N (adaptada de Henriques, 1990)

N	M	Probabilidades F(x) correspondentes aos limites dos M intervalos de partição											
15-20	5	0.000	0.200	0.400	0.600	0.800	1.000						
20-25	6	0.000	0.167	0.333	0.500	0.667	0.833	1.000					
25-30	7	0.000	0.143	0.286	0.429	0.571	0.714	0.857	1.000				
30-40	8	0.000	0.125	0.250	0.375	0.500	0.625	0.725	0.875	1.000			
40-50	9	0.000	0.111	0.222	0.333	0.444	0.556	0.667	0.778	0.889	1.000		
>50	10	0.000	0.100	0.200	0.300	0.400	0.500	0.600	0.700	0.800	0.900	1.000	

Quando E_j depende de m parâmetros estimados a partir da amostra por um método diferente do método da máxima verosimilhança, a estatística do teste χ^2 tem, aproximadamente a distribuição χ^2 com um número de graus de liberdade compreendido entre M-1 e M-m-1, se H_0 for verdadeira.

Observa-se que o teste de Kolmogorov-Smirnov, KS, faz uso mais completo dos dados disponíveis do que o teste do Qui-Quadrado, χ^2. Com efeito, sendo a distribuição postulada contínua, o teste KS examina o ajustamento em cada um dos pontos da amostra, enquanto que o teste do Qui-Quadrado apenas o faz para cada uma das partições do domínio da função de distribuição.

Exercício 14 – Considere a amostra de precipitações diárias máximas anuais no posto udométrico de Pavia (20I/01G) a que se refere a Tabela 1. Conforme se explicitou na Tabela 2, foram estimadas as seguintes estatísticas amostrais: média, $\overline{X} = 39.5$ mm, desvio-padrão, $s_X = 17.2$ mm e coeficiente de assimetria, $g_X = 1.149$. Por aplicação dos testes de Kolmogorov-Smirnov (KS) e do Qui-Quadrado, χ^2, aprecie a qualidade do ajustamento da lei Gumbel à mencionada amostra. Adopte o nível de significância de 5%.

Solução: A primeira parte da Tabela 13, incluída na página seguinte, contém os sucessivos resultados da aplicação do teste de Kolmogorov-Smirnov, KS, à amostra em estudo. Tais resultados estão parcialmente representados na Figura 12, que permite visualizar o valor da estatística do teste.

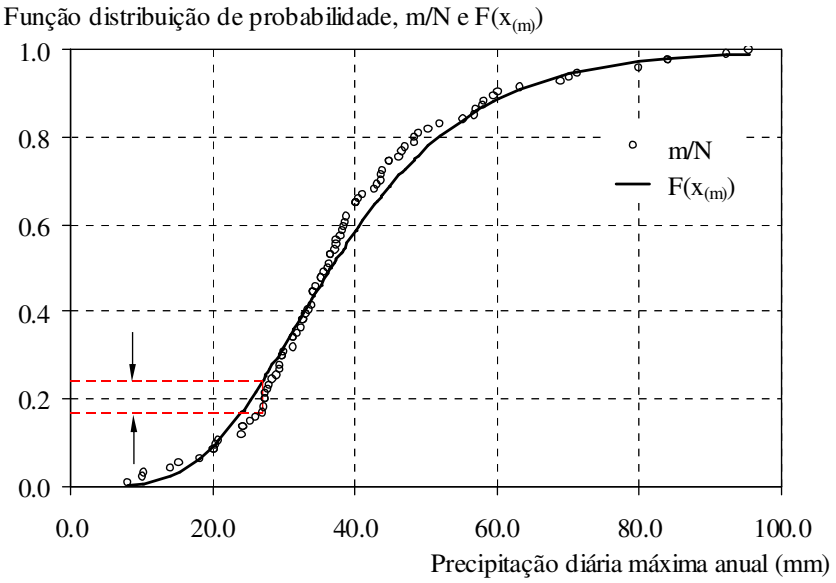

Figura 12 – Aplicação do teste de Kolmogorov-Smirnov, KS, à amostra de precipitações diárias máximas anuais no posto udométrico de Pavia (20I/01G) da Tabela 1. Representação gráfica do valor da estatística do teste.

Conforme se indica na Tabela 13, para o nível de significância adoptado, a estatística do teste (0.0704) é inferior ao correspondente valor crítico (0.1403) pelo que a decisão é a de não rejeitar o ajustamento da distribuição de Gumbel à amostra de precipitações diárias máximas anuais no posto udométrico de Pavia (20I/01G)
A segunda parte da Tabela 13 refere-se à aplicação do teste do Qui-Quadrado. Atendendo à dimensão da amostra (94), foram adoptadas 10 partições com amplitude de 0.10, conducentes a um número esperado de elementos da amostra por intervalo de 9.4. Para a nível de significância de 5%, a estatística do teste (6.8511) é inferior ao valor da distribuição χ^2, tanto para M-1=9 como para M-m-1=7 graus de liberdade, uma vez que foram estimados dois parâmetros a partir da amostra. Q decisão é também a de não rejeitar o ajustamento da distribuição de Gumbel à amostra de precipitações diárias máximas anuais no posto udométrico de Pavia (20I/01G).

Tabela 13 – Aplicação dos testes de Kolmogorov-Smirnov, KS, e do Qui-Quadrado, χ^2, à amostra de precipitações diárias máximas anuais no posto udométrico de Pavia (20I/01G) da Tabela 1.

a) Teste de Kolmogorov-Smirnov, KS

m	$x_{(m)}$	m/N	$F(x_{(m)})$	$\lvert m/N - F(x_{(m)})\rvert$	m	$x_{(m)}$	m/N	$F(x_{(m)})$	$\lvert m/N - F(x_{(m)})\rvert$	m	$x_{(m)}$	m/N	$F(x_{(m)})$	$\lvert m/N - F(x_{(m)})\rvert$
1	8.1	0.0106	0.0029	0.0077	33	31.9	0.3511	0.3721	0.0211	64	42.8	0.6809	0.6450	0.0359
2	10.2	0.0213	0.0068	0.0144	34	32.5	0.3617	0.3886	0.0269	65	43.2	0.6915	0.6533	0.0382
3	10.3	0.0319	0.0071	0.0248	36	32.8	0.3830	0.3968	0.0138	66	43.7	0.7021	0.6636	0.0385
4	14.2	0.0426	0.0247	0.0178	36	32.8	0.3830	0.3968	0.0138	67	43.8	0.7128	0.6656	0.0471
5	15.3	0.0532	0.0331	0.0201	37	33.2	0.3936	0.4077	0.0141	68	44.0	0.7234	0.6696	0.0538
6	18.2	0.0638	0.0642	0.0004	38	33.5	0.4043	0.4159	0.0116	70	45.0	0.7447	0.6892	0.0555
8	20.2	0.0851	0.0940	0.0089	39	34.0	0.4149	0.4295	0.0146	70	45.0	0.7447	0.6892	0.0555
8	20.2	0.0851	0.0940	0.0089	42	34.2	0.4468	0.4349	0.0119	71	46.3	0.7553	0.7133	0.0420
9	20.4	0.0957	0.0973	0.0016	42	34.2	0.4468	0.4349	0.0119	72	46.6	0.7660	0.7187	0.0473
10	20.8	0.1064	0.1042	0.0022	42	34.2	0.4468	0.4349	0.0119	73	47.0	0.7766	0.7257	0.0509
11	24.2	0.1170	0.1729	0.0559	43	34.6	0.4574	0.4456	0.0118	74	48.4	0.7872	0.7491	0.0381
13	24.3	0.1383	0.1752	0.0369	45	35.2	0.4787	0.4617	0.0170	75	48.5	0.7979	0.7507	0.0471
13	24.3	0.1383	0.1752	0.0369	45	35.2	0.4787	0.4617	0.0170	76	49.0	0.8085	0.7586	0.0499
14	25.2	0.1489	0.1961	0.0472	46	35.7	0.4894	0.4749	0.0144	77	50.4	0.8191	0.7797	0.0394
15	26.0	0.1596	0.2155	0.0560	47	36.2	0.5000	0.4880	0.0120	78	52.0	0.8298	0.8018	0.0280
16	**27.0**	**0.1702**	**0.2407**	**0.0704**	48	36.5	0.5106	0.4959	0.0148	79	55.2	0.8404	0.8403	0.0001
17	27.2	0.1809	0.2458	0.0649	50	36.7	0.5319	0.5010	0.0309	80	56.8	0.8511	0.8569	0.0059
19	27.4	0.2021	0.2509	0.0488	50	36.7	0.5319	0.5010	0.0309	81	57.0	0.8617	0.8589	0.0028
19	27.4	0.2021	0.2509	0.0488	51	37.2	0.5426	0.5139	0.0287	82	58.0	0.8723	0.8683	0.0040
20	27.5	0.2128	0.2535	0.0408	52	37.4	0.5532	0.5189	0.0342	83	58.2	0.8830	0.8701	0.0128
21	27.8	0.2234	0.2613	0.0379	53	37.5	0.5638	0.5215	0.0423	84	59.6	0.8936	0.8822	0.0114
22	28.0	0.2340	0.2666	0.0325	54	38.0	0.5745	0.5341	0.0404	85	60.2	0.9043	0.8871	0.0172
23	28.5	0.2447	0.2798	0.0351	55	38.2	0.5851	0.5390	0.0461	86	63.3	0.9149	0.9093	0.0056
24	29.0	0.2553	0.2932	0.0378	56	38.4	0.5957	0.5440	0.0518	87	69.0	0.9255	0.9397	0.0142
25	29.4	0.2660	0.3039	0.0380	57	38.6	0.6064	0.5489	0.0575	88	70.2	0.9362	0.9447	0.0086
26	29.5	0.2766	0.3066	0.0300	58	38.9	0.6170	0.5562	0.0608	89	71.4	0.9468	0.9493	0.0025
28	29.8	0.2979	0.3147	0.0169	61	40.2	0.6489	0.5872	0.0617	90	80.0	0.9574	0.9730	0.0155
28	29.8	0.2979	0.3147	0.0169	61	40.2	0.6489	0.5872	0.0617	92	84.2	0.9787	0.9802	0.0015
29	30.0	0.3085	0.3202	0.0117	61	40.2	0.6489	0.5872	0.0617	92	84.2	0.9787	0.9802	0.0015
30	31.3	0.3191	0.3557	0.0365	62	40.5	0.6596	0.5942	0.0654	93	92.3	0.9894	0.9891	0.0002
32	31.4	0.3404	0.3584	0.0180	63	41.2	0.6702	0.6101	0.0601	94	95.5	1.0000	0.9914	0.0086
32	31.4	0.3404	0.3584	0.0180										

Estatística do teste, máx $\lvert m/n - F(x_{(m)})\rvert$:	**0.0704**
Valor crítico da estatística do teste para o nível do significância, α, de 5%:	**0.1403**

b) Teste do Qui-Quadrado, χ^2

Partição do domínio da função distribuição de probabilidade		Factor de probabilidade da lei de Gumbel		Valor da variável aleatória		Número de elementos por classe	
F_i	F_{i+1}	K_i	K_{i+1}	x_i	x_{i+1}	Efectivo, O_j	Esperado, E_j=M/N
0.0	0.1	--	-1.1003	- ∞	20.56	9	9.4
0.1	0.2	-1.1003	-0.8211	20.56	25.36	5	
0.2	0.3	-0.8211	-0.5948	25.36	29.25	10	
0.3	0.4	-0.5948	-0.3819	29.25	32.92	12	
0.4	0.5	-0.3819	-0.1643	32.92	36.66	12	
0.5	0.6	-0.1643	0.0737	36.66	40.75	14	
0.6	0.7	0.0737	0.3538	40.75	45.57	8	
0.7	0.8	0.3538	0.7194	45.57	51.86	7	
0.8	0.9	0.7194	1.3046	51.86	61.93	8	
0.9	1.0	1.3046	--	61.93	+ ∞	9	

Estatística do teste, $\chi^2 = \sum_{j=1}^{M} \frac{(O_j - E_j)^2}{E_j} = \frac{M}{N}\sum_{j=1}^{M} O_j^2 - N$:		**6.851**
Valor crítico da estatística do teste para o nível de significância, α, de 5%	**Para ν=M-1=9 graus de liberdade**	**16.919**
	Para ν=M-m-1=7 graus de liberdade	**14.067**

9.4. Avaliação das incertezas associadas às estimativas de quantis

A estimativa do quantil $\hat{x}_F$, relativo à probabilidade de não excedência, F, obtido por um método de estimação contém, independentemente desse método, erros que são inerentes às incertezas presentes na estimação das características e dos parâmetros populacionais a partir de amostras de dimensão N, necessariamente reduzida face à infinitude daquela população, como já repetidamente afirmado. Uma medida frequentemente usada para quantificar a variabilidade intrínseca de $\hat{x}_F$, e, portanto, indicar a *confiança* das estimativas de quantis de variáveis hidrológicas, é dada pelo *erro padrão da estimativa*, S_F, definido por:

$$S_F = \sqrt{E\left[\left\{\hat{x}_F - E\left[\hat{x}_F\right]\right\}^2\right]} \quad \text{.....(25)}$$

O erro padrão da estimativa leva em conta apenas os erros oriundos do processo de estimação a partir de amostras finitas e, portanto, não considera o eventual erro devido à selecção de uma distribuição de probabilidades inadequada. Admitindo que a distribuição $F_X(x)$ tenha sido correctamente especificada, o erro padrão da estimativa compreende os erros inerentes às *estimativas dos parâmetros* de $F_X(x)$. Os diferentes métodos de estimação produzirão diferentes erros-padrão das estimativas. O método de estimação com maior *eficiência,* do ponto de vista estatístico, é o que resultar no menor valor de S_F.

A teoria estatística de amostragem demonstra que a distribuição de $\hat{x}_F$ é assimptoticamente Normal, com média igual à *estimativa do quantil*, $\hat{x}_F$, e desvio-padrão S_F, quando a dimensão da amostra tende para infinito, ou seja, $N \to \infty$. No que respeita a amostras finitas com dimensão N, o anterior resultado teórico pode ser usado para construir *intervalos de confiança aproximados,* para o nível $100(1-\alpha)\%$, cujos limites são expressos por:

$$\hat{x}_F \pm \left|z_{\alpha/2}\right| S_F \quad \text{.....(26)}$$

onde $z_{\alpha/2}$ representa a variável Normal padrão para a probabilidade de não-excedência de $\alpha/2$.

A dificuldade de aplicar o procedimento descrito para estimar intervalos de confiança associados a estimativas de quantis decorre do cálculo de S_F que é muito complexo para todos os métodos de estimação e para quase todas as distribuições, com particular ênfase para as de três parâmetros – ver *Kite* (1988), *Rao e Hamed* (2000) e o capítulo 6 de *Naghettini e Pinto* (2007).

Uma alternativa para associar intervalos de confiança a quantis, muito menos complexa do que a aproximação expressa pela equação (26), utiliza a *geração, por recurso à técnica de Monte Carlo, de um grande número de amostras* com o mesmo tamanho N da amostra original – amostras *sintéticas da dimensão* N – com estimação a partir de cada uma dessas amostras, do quantil pretendido, ao qual é posteriormente associado uma distribuição empírica de probabilidades.

Suponha-se que, à amostra $\{x_1, x_2, ..., x_N\}$, se ajustou uma distribuição de probabilidades genérica $F_X(x)$, cujos parâmetros $\theta_1, \theta_2, ..., \theta_k$ foram estimados a partir de um método qualquer de estimação designado por EM. A aplicação da técnica de Monte Carlo tendo em vista construir intervalos de confiança em torno das estimativas de k quantis $\hat{x}_{F_k}$ processa-se de acordo com as seguintes etapas sequenciais:

i Para cada valor de (j), variável entre 1 (primeira amostra sintética) e W (última amostra sintética, com W muito grande, da ordem dos milhares, por exemplo, 5000), geração da amostra sintética de ordem (j) com dimensão N, mediante a geração de N *números aleatórios uniformes* entre 0 e 1, $u_i^{(j)}$, com i=1, …, N, sendo N a dimensão das amostra, quer original, quer sintética de ordem (j).

ii No entendimento de que, para a amostra sintética de ordem (j), cada um dos anteriores N valores de $u_i^{(j)}$ representa uma probabilidade de não-excedência, ou seja, $u_i^{(j)} = F^{(j)}(x_i) = F_i^{(j)}$, cálculo dos N quantis $\hat{x}_i^{(j)}$, com i=1, …, N, seja por inversão directa da função $F_i^{(j)}$, seja recorrendo ao método dos factores de probabilidade, conforme Tabela 7, num e noutro caso, tendo por base as estimativas dos parâmetros obtidas a partir da amostra original, $\theta_1, \theta_2, ..., \theta_k$.

iii Da etapa precedente resulta uma amostra sintética de dimensão N, $\hat{x}_i^{(j)}$, de um conjunto W dessas amostras, com W muito elevado, da ordem dos milhares, conforme antes explicitado.

iv Estritamente com base na amostra sintética de ordem (j) e mediante utilização do método de estimação EM, cálculo das estimativas dos parâmetros $\theta_1^{(j)}, \theta_2^{(j)}, ..., \theta_k^{(j)}$, e, conhecidos estes parâmetros, dos quantis pretendidos, $\hat{x}_{F_k}^{(j)}$, seja por inversão da função $F_i^{(j)}$, seja recorrendo ao método dos factores de probabilidade, conforme Tabela 7, num e noutro caso, tendo por base as estimativas dos parâmetros obtidas a partir da amostra sintética de ordem (j), $\theta_1^{(j)}, \theta_2^{(j)}, ..., \theta_k^{(j)}$;

v Repetição das etapas (i) a (iv) para W amostras sintéticas (W=1, …, 5000, …).

vi No final do anterior processo, dispõe-se, para cada quantil $\hat{x}_{F_k}$, de W estimativas, $\hat{x}_{F_k}^{(j)}$, com j=1, … W, as quais são ordenadas por ordem, por exemplo, crescente.

vii Sendo W muito grande, para definir os limites do intervalo de confiança a 100 (1-α/2)% para cada um desses quantis basta reter os quantis com ordens de classificação W(α/2) e W(1-α/2).

A Tabela 14 e a Figura 13 ilustram a obtenção, segundo a lei de Gumbel, com parâmetros estimados pelo método dos momentos, dos intervalos de confiança a 95% dos quantis das precipitações diárias máximas anuais no posto udométrico de Pavia (20I/01G) a que se refere a Tabela 1.

A anterior tabela exemplifica o procedimento de cálculo de acordo com as etapas anteriormente descritas, no pressuposto de geração de W=5000 séries sintéticas de precipitações diárias máximas anuais, cada uma com dimensão igual à da série histórica, ou seja, com N=94 valores. Por razões óbvias, apenas se incluíram uns escassos resultados referentes às primeiras cinco e às últimas cinco séries sintéticas, nomeadamente, alguns dos números aleatórios uniformes gerados entre 0 e 1 (primeiro quadro da tabela) e as correspondentes estimativas de precipitações diárias máximas anuais avaliadas por recurso ao método dos factores de probabilidade para a lei de Gumbel, atendendo à média e ao desvio-padrão da série histórica.

Tabela 14 – Intervalo de confiança a 95%, para a estimativa fornecida pela lei de Gumbel para a precipitação diária máxima anual no posto udométrico de Pavia (20I/01G) com a probabilidade de não-excedência de 99% (período de retorno de 100 anos).

Ano	Séries de 94 números aleatórios uniformes entre 0 e 1										
	Número de ordem da série										
	1	2	3	4	5	...	4996	4997	4998	4999	5000
1	0.548	0.142	0.710	0.254	0.832	...	0.869	0.336	0.502	0.805	0.908
2	0.585	0.618	0.658	0.266	0.088	...	0.283	0.609	0.296	0.361	0.527
3	0.542	0.975	0.426	0.697	0.077	...	0.040	0.553	0.549	0.858	0.545
4	0.882	0.053	0.356	0.572	0.705	...	0.841	0.930	0.586	0.920	0.911
5	0.511	0.069	0.901	0.151	0.540	...	0.247	0.426	0.165	0.164	0.945
6	0.749	0.896	0.113	0.873	0.105	...	0.476	0.028	0.679	0.681	0.552
7	0.755	0.567	0.847	0.426	0.788	...	0.999	0.832	0.647	0.908	0.572
8	0.759	0.525	0.278	0.890	0.728	...	0.538	0.814	0.809	0.379	0.363
9	0.624	0.335	0.127	0.078	0.661	...	0.204	0.286	0.394	0.251	0.299
10	0.040	0.009	0.111	0.849	0.742	...	0.205	0.662	0.525	0.185	0.298
...	...	...	...	...	...	...	...	...	...	...	...
84	0.964	0.425	0.936	0.589	0.622	...	0.134	0.746	0.095	0.038	0.125
85	0.400	0.512	0.390	0.015	0.899	...	0.463	0.556	0.204	0.382	0.288
86	0.659	0.332	0.499	0.547	0.683	...	0.940	0.445	0.209	0.951	0.096
87	0.632	0.692	0.121	0.290	0.226	...	0.361	0.312	0.620	0.603	0.503
88	0.638	0.301	0.016	0.873	0.071	...	0.196	0.678	0.674	0.981	0.581
89	0.148	0.973	0.716	0.706	0.154	...	0.151	0.694	0.765	0.012	0.551
90	0.430	0.094	0.428	0.482	0.332	...	0.404	0.584	0.602	0.176	0.196
91	0.389	0.601	0.750	0.731	0.671	...	0.665	0.058	0.017	0.843	0.440
92	0.308	0.074	0.944	0.410	0.880	...	0.995	0.958	0.519	0.791	0.150
93	0.732	0.631	0.692	0.072	0.165	...	0.076	0.818	0.256	0.548	0.869
94	0.597	0.599	0.876	0.083	0.700	...	0.695	0.922	0.105	0.206	0.472

Ano	Séries sintéticas de precipitações diárias máximas anuais com dimensão de 94 anos (mm)										
	Número de ordem da série										
	1	2	3	4	5	...	4996	4997	4998	4999	5000
1	38.58	22.79	46.13	27.52	54.44	...	58.03	30.57	36.75	52.27	63.17
2	40.12	41.53	43.44	27.99	19.83	...	28.61	41.16	29.10	31.50	37.71
3	38.32	80.82	33.86	45.40	19.10	...	16.06	38.76	38.59	56.96	38.46
4	59.56	17.25	31.33	39.57	45.84	...	55.26	67.02	40.16	64.99	63.59
5	37.10	18.58	62.06	23.18	38.23	...	27.25	33.88	23.86	23.79	70.26
6	48.36	61.35	21.27	58.56	20.83	...	35.75	14.69	44.46	44.58	38.71
7	48.79	39.35	55.78	33.86	50.96	...	123.16	54.43	42.90	63.11	39.57
8	49.03	37.64	28.43	60.56	47.12	...	38.16	52.95	52.51	32.16	31.56
9	41.81	30.53	22.01	19.17	43.58	...	25.55	28.73	32.71	27.41	29.23
10	16.02	11.09	21.16	56.00	47.98	...	25.58	43.61	37.63	24.73	29.18
...	...	...	...	...	...	...	...	...	...	...	...
84	76.17	33.83	68.12	40.28	41.75	...	22.38	48.22	20.25	15.82	21.94
85	32.91	37.13	32.54	12.49	61.85	...	35.27	38.91	25.51	32.27	28.81
86	43.49	30.43	36.64	38.52	44.68	...	68.96	34.58	25.72	71.86	20.35
87	42.18	45.17	21.72	28.88	26.42	...	31.49	29.70	41.65	40.90	36.79
88	42.47	29.30	12.70	58.52	18.72	...	25.18	44.44	44.22	84.52	39.92
89	23.07	79.83	46.47	45.92	23.34	...	23.21	45.24	49.39	11.75	38.70
90	34.03	20.19	33.94	35.96	30.44	...	33.06	40.05	40.83	24.32	25.18
91	32.52	40.81	48.45	47.30	44.04	...	43.75	17.72	12.88	55.48	34.40
92	29.55	18.91	70.14	33.27	59.31	...	103.30	73.92	37.41	51.21	23.17
93	47.37	42.16	45.14	18.74	23.84	...	19.07	53.26	27.60	38.55	58.02
94	40.61	40.72	58.91	19.51	45.56	...	45.33	65.42	20.85	25.61	35.59
Média (mm)	41.52	38.23	38.02	37.25	40.32	...	40.18	40.23	39.55	37.86	41.08
Desvio-padrão (mm)	22.22	16.45	17.36	15.26	17.33	...	20.99	16.16	16.98	16.15	16.94
Estimativa para F=0.99 (mm)	111.22	89.84	92.46	85.11	94.68	...	106.01	90.94	92.82	88.52	94.20

Ordem da série sintética	Estimativa para F=0.99, ou seja, para T=100 anos	
	Para cada série sintética (mm)	Ordenada por valores crecentes (mm)
1	111.22	71.20
2	89.84	72.00
3	92.46	72.50
...	...	...
118	86.53	76.90
119	104.94	77.00
120	87.80	77.08
121	100.62	79.60
122	98.83	80.52
123	93.92	80.74
124	102.46	80.77
125	88.63	**80.79**
125	88.46	82.45
126	88.39	82.74
127	91.60	82.87
128	92.87	82.98
129	97.35	83.32
130	103.22	83.75
...	...	...
4869	94.10	106.90

Ordem da série sintética	Estimativa para F=0.99, ou seja, para T=100 anos	
	Para cada série sintética (mm)	Ordenada por valores crecentes (mm)
4870	78.31	106.53
4871	96.75	107.00
4872	92.96	107.30
4873	95.96	107.40
4874	84.25	107.50
4875	87.20	**107.56**
4876	93.10	108.00
4877	90.49	108.30
4878	89.64	108.50
4879	93.83	108.80
4880	91.25	109.20
4881	92.78	109.50
4882	99.82	110.00
4883	91.27	110.84
4884	90.65	111.00
4885	105.09	111.10
...	...	...
4998	92.82	113.70
4999	88.52	113.90
5000	94.20	114.00

Nota: Os limites do intervalo de confiança estão destacados a negrito e sombreado.

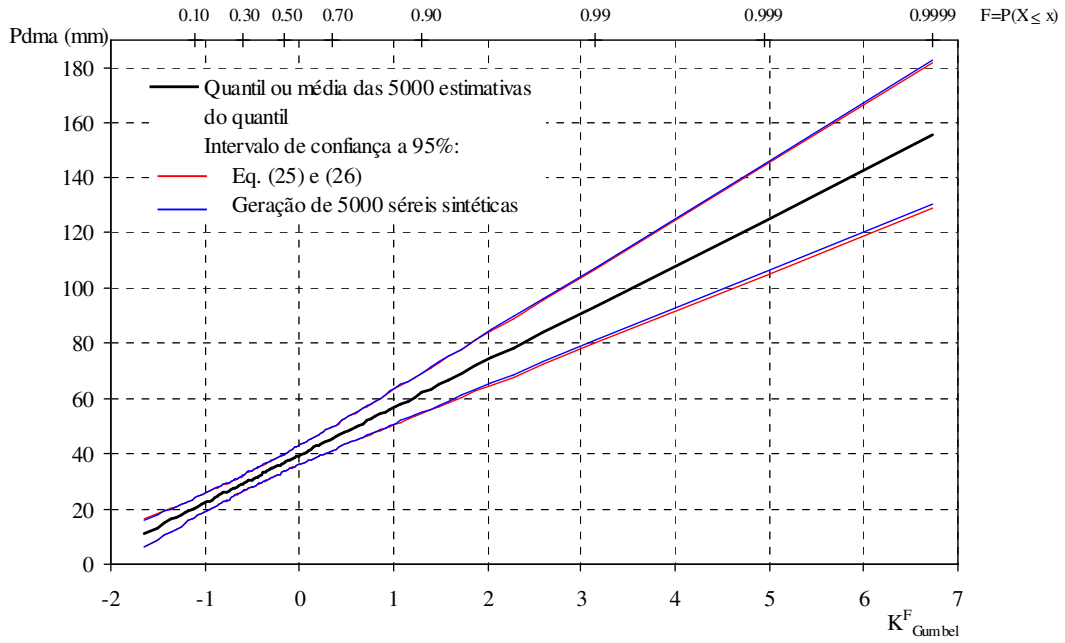

Figura 13 – Intervalos de confiança a 95%, para os quantis fornecidos pela lei de Gumbel para as precipitações diárias máximas anuais no posto udométrico de Pavia (20I/01G).

Resultaram, assim, 5000 séries sintéticas de precipitações diárias máximas anuais (segundo quadro da tabela), sendo que se incluíram na tabela as médias e os desvios-padrões das séries aí parcialmente exemplificadas, bem como as respectivas estimativas das precipitações máximas diárias anuais para a probabilidade de não-excedência de 0.99 (período de retorno de 100 anos), obtidas também por aplicação do método dos factores de probabilidade, mas, agora, fazendo intervir as estatísticas amostrais (média e desvio-padrão) de cada uma das sucessivas séries sintéticas.

O terceiro quadro incluído na Tabela 14 exemplifica o procedimento de cálculo do intervalo de confiança da precipitação para o quantil de 99%. Conforme aí indicado, obtidas as 5000 estimativas das precipitações máximas diárias anuais para a probabilidade de não-excedência de 0.99 e ordenadas tais estimativas por valores crescentes, os limites do intervalo de confiança, por exemplo, a 5% são dados pelas estimativas que ocupam as posições ordenadas $0.025\times5000=125$ e $0.975\times5000=4875$, com os valores de, respectivamente, 80.79 e 107.56 mm, destacados na tabela. Recorda-se que, em conformidade com o Exercício 12, a estimativa da precipitação diária máxima anual fornecida pelo método dos momentos baseado no factor de probabilidade para aquela probabilidade de não-excedência foi de 93.5 mm.

A Figura 13 contém as curvas que definem os limites do intervalo de confiança a 95% para a generalidade dos quantis fornecidos, por um lado, pela geração de 5000 séries sintéticas, de acordo com o procedimento exemplificado na Tabela 15 e, por outro lado, por aplicação das equações (25) e (26), para o que foi necessário especificar o erro padrão, S_F, para o que se utilizou a seguinte equação, válida no caso de aplicação do método dos momentos a uma distribuição estatística de dois parâmetros:

$$S_T^2 = \frac{S_X{}^2}{N}\left\{1 + K^F\ \gamma_1 + \frac{K^{F2}}{4}(\gamma_2 - 1)\right\} \quad \text{.....(27)}$$

em que N é a dimensão da amostra; K^F, o factor de probabilidade; S_X^2, a variância da amostra; e γ_1 e γ_2 os coeficientes de assimetria e de curtose da população que, para a lei de Gumbel, são iguais a, respectivamente, 1.1396 e 5.4.

A Figura 13 suscita algumas observações pertinentes, a primeira das quais relativa ao segmento de recta assinalada a preto. Conforme se explicitou na legenda da figura, tal segmento representa:

- os quantis estimados por aplicação do método dos momentos baseado no factor de probabilidade a partir das estimativas da média e do desvio-padrão da amostra (39.5 mm e 17.2 mm, respectivamente, conforme Tabela 2), sendo que coincide exactamente com o segmento de recta referente à lei de Gumbel incluído no gráfico inferior da Figura 11;
- a menos de desvios praticamente imperceptíveis, a média das 5000 estimativas de cada um dos sucessivos quantis.

Importa recordar que está em causa um segmento de recta e não uma curva pois trata-se de uma representação da função de distribuição de probabilidade da lei de Gumbel em papel de probabilidade dessa mesma lei.

Concluiu-se, assim, que sendo o número de séries sintéticas suficientemente elevado, as médias das estimativas dos sucessivos quantis que resultam das séries sintéticas coincidem com as estimativas desses quantis fornecidas pela amostra histórica.

A título exemplificativo, obteve-se a Figura 14 que contém o histograma das 5000 estimativas da precipitação que decorrem das séries sintéticas para a probabilidade de não-excedência de 0.99. A tais estimativas ajustou-se a lei Normal, conforme representado na figura. A média dessas estimativas – com o valor indicado na figura de cerca de 93.2 mm – é praticamente coincidente com a estimativa do quantil obtida a partir da amostra, dada por:

$$x_T = \overline{X} + K^F_{Gumbel}\ s_X = 39.5 + 3.137 \times 17.2 = 93.5 \text{ mm} \quad \text{.....(28)}$$

resultado, aliás, antes obtido no Exercício 12, não obstante o método de estimação então aplicado ter sido diferente.

Retomando a análise da Figura 13, verifica-se que os limites fornecidos pelo recurso à geração de 5000 de séries sintéticas ou por aplicação das equações (25) e (26) são praticamente coincidentes, sendo que aquela técnica, embora computacionalmente exigente, assenta num formalismo matemático simples e facilmente aplicável a diferentes distribuições obviando a grande complexidade de cálculo do erro padrão da estimativa, S_F, conforme antes referido. Importa anotar que o esforço computacional exigido pela técnica de Monte Carlo pode ser minimizado pela aplicação *Pythia-Statistical Analysis* do *software* gratuito *Hydrognomon*, desenvolvido pela Universidade Técnica de Atenas e disponível para *download* a partir de acesso à URL http://hydrognomon.org/download.html.

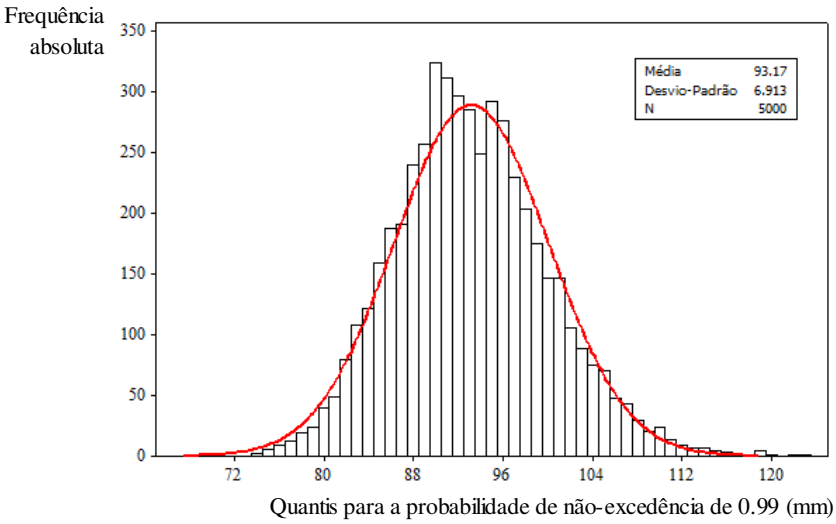

Figura 14 – Histograma das estimativas fornecidas pelas séries sintéticas (em número de W=5000) da precipitação diária máxima anual no posto udométrico de Pavia (20I/01G) para a probabilidade de não-excedência de 99%.

Observa-se, por fim, que, tal como representado na Figura 13, os limites do intervalo de confiança a 95% se distanciam progressiva e significativamente da curva de quantis, à medida que a probabilidade de não-excedência e, logo, o período de retorno aumentam. Com efeito e tal como anteriormente explicitado, para F=0.99 (T=100 anos), o intervalo a 95% associado à correspondente estimativa da precipitação diária máxima anual, de 93.5 mm, é [80.79 mm, 107.56 mm], apresentando, portanto, desvios, relativamente àquela estimativa, sensivelmente entre -13.6 e +15.0 %, de modo a conter as incertezas devidas à estimação de parâmetros e de quantis. O entendimento associado ao anterior intervalo é de que o mesmo contém o verdadeiro, embora desconhecido, quantil da precipitação para o período de retorno de 100 anos, com a probabilidade de 95%.

O afastamento progressivo das curvas que definem os intervalos de confiança para probabilidades de não-excedência crescentes reflecte as incertezas progressivamente maiores subjacentes à análise de frequência com amostras finitas de tamanho N. Esta constatação aponta no sentido de ser necessário um cuidado especial na *extrapolação da curva de frequências* para probabilidades de não-excedência correspondentes a períodos de retorno muito superiores à dimensão, N, da amostra disponível. Embora dependendo da qualidade do ajustamento a uma amostra de tamanho N, de modo geral, *não se recomenda a extrapolação da curva de frequências para períodos de retorno superiores a 4N*. Se essa extrapolação for mesmo necessária, poder-se-á recorrer a métodos complementares, incluindo a análise regional de frequências, que, de algum modo, introduz alguma compensação nas amostras de pequeno tamanho, pela agregação de informações referentes a outras estações de monitorização, localizadas numa mesma região considerada homogénea no que respeita ao fenómeno traduzido pela variável hidrológica para a qual se pretendem estimar quantis. Para detalhes sobre a análise regional de frequências, o leitor pode consultar *Hosking e Waliis* (1997) e o capítulo 10 de *Naghettini e Pinto* (2007).

10. Correlação e regressão simples de variáveis hidrológicas

Na prática da engenharia de recursos hídricos, com alguma frequência, é necessário estabelecer a *forma e o grau da associação* entre duas ou mais variáveis, como, por exemplo, no estudo das relações entre: (i) as intensidades médias, as durações e as frequências associadas a precipitações intensas; (ii) os módulos dos caudais médios diários em diferentes bacias e as áreas de drenagem dessas bacias; (iii) as alturas anuais médias da precipitação e as altitudes dos postos udométricos; ou (iv) os níveis hidrométricos e os caudais afluentes numa estação hidrométrica, entre outros exemplos.

Para tanto é necessário analisar o comportamento simultâneo das duas variáveis aleatórias em presença, Y e X, verificando se a variação (no sentido do aumento ou da diminuição) de uma delas está associada à variação (no mesmo sentido ou em sentidos contrários) da outra, ou mesmo, se não há qualquer dependência estatística entre as variáveis.

Uma medida quantitativa do grau de *associação linear* entre Y e X é dada pelo *coeficiente de correlação de Pearson* (frequentemente, designado apenas por *coeficiente de correlação*), cuja estimativa, a partir de uma amostra de pares de valores $\{x_i, y_i; i=1, 2, \ldots, N\}$, é dada por:

$$r_{xy} = \frac{S_{XY}}{S_X S_Y} = \frac{[1/(N-1)]\sum_{i=1}^{N}(x_i - \overline{X})(y_i - \overline{Y})}{S_X S_Y} \quad \text{.....(29)}$$

onde S_{XY} denota a covariância de X e Y (expressa, portanto, nas unidades de X e de Y) e S_X e S_Y, os respectivos desvios-padrão.

Contrariamente à covariância, o coeficiente de correlação linear de Pearson é *adimensional* e varia entre -1 e +1. Deste modo, as unidades de X e Y não afectam o valor do coeficiente de correlação. Caso os pares $\{x_i, y_i\}$ se alinhem perfeitamente ao longo de uma recta com declive positivo, ter-se-á uma correlação linear positiva perfeita, sendo o coeficiente de correlação igual a 1. A correlação linear negativa perfeita ocorre quando os pares $\{x_i, y_i\}$ se alinham perfeitamente ao longo de uma recta com declive negativo, sendo o coeficiente de correlação neste caso é igual a -1. O significado de valores intermediários do coeficiente é fácil e intuitivamente perceptível.

A Figura 15 apresenta alguns hipotéticos diagramas de dispersão de duas variáveis, com as respectivas estimativas do coeficiente de correlação. Nota-se que um valor nulo para o coeficiente de correlação não implica que não haja nenhuma associação entre X e Y. De facto, tal como ilustrado na Figura 15 apesar de r=0, pode haver *associação não linear* entre as variáveis.

Ainda a respeito de coeficiente de correlação, cabe sublinhar que um elevado valor de r, embora estatisticamente significativo, *não implica necessariamente numa relação de causa e efeito* entre as variáveis. De facto, um elevado coeficiente de correlação indica simplesmente que há uma associação na variação conjunta daquelas variáveis, a qual pode ser explicada, por exemplo, por ocorrências de um factor causal comum a ambas.

A simples visualização de um diagrama de dispersão pode sugerir, muitas vezes, a existência de uma relação funcional entre as variáveis Y e X, o que introduz o problema de se determinar a função que formaliza essa dependência. Uma técnica estatística para o efeito disponível é a *análise de regressão*.

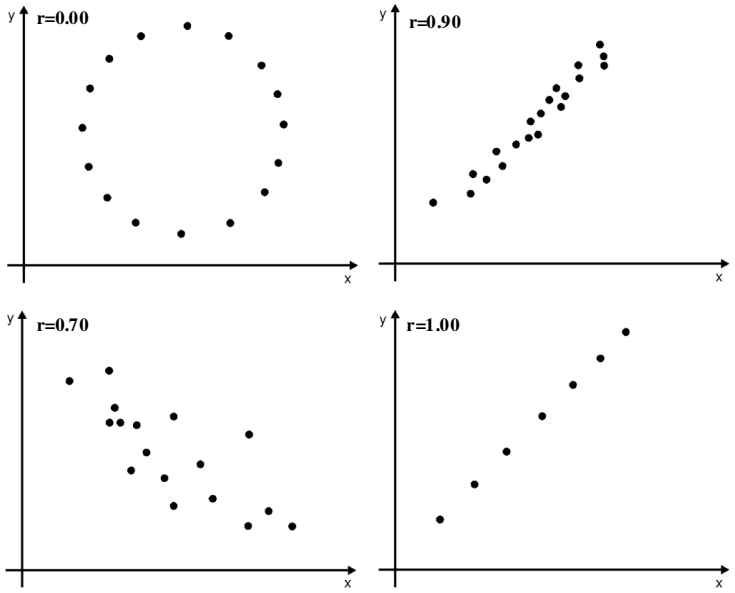

Figura 15 – Alguns exemplos de associações denotando correlação entre as variáveis Y e X.

Nesse contexto, suponha-se que a variação de Y, denominada *variável dependente* (de *resposta* ou *explicada*), possa ser compreendida e modelada a partir da variação de X, chamada *variável independente* (ou *explicativa*). A *forma funcional*, ou *modelo de regressão*, que relaciona Y e X, deve ser capaz de explicar uma parcela significativa da variação conjunta de ambas variáveis. Contudo, pela natureza das dependências estatísticas, parte da variação pode permanecer inexplicada, devendo ser atribuída ao acaso. Noutros termos, admite-se a existência de uma função que explica, em *termos médios*, a variação de Y a partir de X. Os pares de observação $\{x_i, y_i\}$ apresentarão uma variação aleatória em torno da linha estabelecida pela função de regressão, que é denominada *variação residual*. Portanto, a equação que define o modelo de regressão fornece o *valor médio* de Y em função de X. Se a forma funcional do modelo de regressão for conhecida (ou prescrita), haverá que estimar os coeficientes (ou parâmetros) da equação (ou modelo) de regressão.

Admita-se que a equação de regressão entre Y e X seja descrita por uma recta:

$$Y = \alpha + \beta X + \varepsilon \quad \text{.....(30)}$$

onde α e β são os coeficientes de regressão e ε denota os erros ou *resíduos* da regressão. Os coeficientes α e β têm de ser estimados a partir dos pares de observações $\{x_i, y_i; i=1, 2, \ldots, N\}$, resultando na seguinte estimativa:

$$\hat{y}_i = \hat{\alpha} + \hat{\beta} x_i = a + b x_i \quad \text{.....(31)}$$

onde $\hat{y}_i$ é o valor estimado da variável dependente a partir de valor observado x_i da variável independente e $\hat{\alpha} = a$ e $\hat{\beta} = b$ as estimativas dos coeficientes de regressão.

O método mais usual para realizar a estimação de α e β é o método dos *mínimos quadrados*, cujo objectivo é encontrar a função de regressão que *minimiza a soma dos quadrados dos desvios* (ou *resíduos quadráticos*) entre os pontos observados e os calculados pela função ajustada, como se esquematiza na Figura 16.

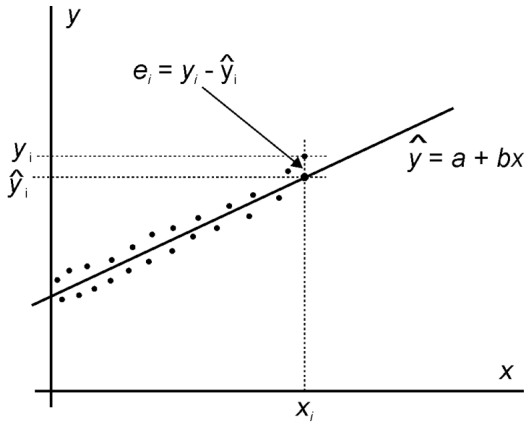

Figura 16 – Coeficientes de regressão pelo método dos mínimos quadrados

De acordo com o anterior método, para o ponto com ordem de i, a distância quadrática é dada por:

$$e_i^2 = (y_i - a - bx_i)^2 = y_i^2 - 2y_i a - 2y_i bx_i + a^2 + 2abx_i + b^2 x_i^2 \quad \text{.....(32)}$$

Logo, para todos os N elementos da amostra, resulta:

$$Z = \sum_{i=1}^{N} \varepsilon_i^2 = \sum_{i=1}^{N} y_i^2 - 2a\sum_{i=1}^{N} y_i - 2b\sum_{i=1}^{N} (x_i y_i) + \\ + Na^2 + 2ab\sum_{i=1}^{N} x_i + b^2 \sum_{i=1}^{N} x_i^2 \quad \text{.....(33)}$$

Como $Z = f(a,b)$, os valores dos coeficientes a e b que minimizam a soma dos quadrados dos desvios são obtidos igualando a zero as derivadas parciais de *Z* em relação àqueles coeficientes. Esta operação resulta no seguinte sistema de duas equações e duas incógnitas:

$$\begin{cases} \sum_{i=1}^{N} y_i - Na - b\sum_{i=1}^{N} x_i = 0 \\ \sum_{i=1}^{N} (x_i y_i) - a\sum_{i=1}^{N} x_i - b\sum_{i=1}^{N} x_i^2 = 0 \end{cases} \quad \text{.....(34)}$$

cujas soluções são as estimativas de α e β, dadas pelas seguintes equações:

$$a = \frac{\sum_{i=1}^{N} y_i}{N} - b\frac{\sum_{i=1}^{N} x_i}{N} = \overline{Y} - b\overline{X} \quad \text{.....(35)}$$

$$b = \frac{N\sum_{i=1}^{N} (x_i y_i) - \sum_{i=1}^{N} y_i \sum_{i=1}^{N} x_i}{N\sum_{i=1}^{N} x_i^2 - \left(\sum_{i=1}^{N} x_i\right)^2} \quad \text{.....(36)}$$

Algumas funções não lineares podem ser linearizadas mediante o uso de transformações adequadas, permitindo, assim, a aplicação das equações da regressão linear simples. Um exemplo é a função potencial do tipo $Y = aX^b$, a qual, mediante aplicação de logaritmos pode ser transformada no modelo linear $Z = k + bV$, no qual $Z = \ln Y$, $k = \ln a$ e $V = \ln X$. As equações (35) e (36) podem, então, ser aplicadas às variáveis transformadas Z e V.

Para modelos lineares e não lineares, a qualidade do ajustamento é avaliada pelo *coeficiente de determinação*, R^2, dado pela equação:

$$R^2 = 1 - \frac{\sum_{i=1}^{N} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{N} (y_i - \overline{Y})^2} \quad \text{.....(37)}$$

O coeficiente de determinação representa a *fracção da variância total de Y que foi explicada pelo modelo de regressão*. Um valor próximo de 1 significa que o modelo de regressão é quase perfeito. Ao contrário, se próximo de zero, significa que o modelo de regressão tem pouquíssimo valor ao tentar explicar a variância de Y. No caso de um modelo linear, o quadrado do coeficiente de correlação, R, avaliado pela equação (29), corresponde também ao coeficiente de determinação. Ao leitor interessado em detalhes sobre outras funções não lineares, testes estatísticos referentes aos coeficientes de correlação e de regressão, análise dos resíduos da regressão e extensão dos conceitos aqui apresentados para mais de uma variável explicativa, sugere-se a consulta de *Montgomery e Peck* (1992) e do capítulo 9 de *Naghettini e Pinto* (2007).

Exercício 15 – Deduzir as *equações normais* para o seguinte modelo parabólico $Q = a + b\,h + c\,h^2$.

Solução: A variável dependente é Q e a independente é h, com três coeficientes de regressão (a, b, c). Procedendo exactamente de acordo com as equações (32) e (33), obtém-se a uma expressão relativa à soma dos desvios quadráticos Z, a qual, neste caso, é uma função de *a*, *b* e *c*. As *equações normais* resultam de se igualarem a zero as derivadas parciais de Z, em relação a *a*, *b* e *c*, com obtenção do seguinte sistema de equações:

$$\begin{cases} \sum Q = N\,a + b\sum h + c\sum h^2 \\ \sum (Q\,h) = a\sum h + b\sum h^2 + c\sum h^3 \\ \sum (Q\,h^2) = a\sum h^2 + b\sum h^3 + c\sum h^4 \end{cases}$$

Exercício 16 - A Tabela 15 contém os caudais instantâneos, Q, considerados no estabelecimento da curva de vazão numa dada estação hidrométrica, bem como as alturas hidrométricas, h, para esses caudais. Estabeleça a equação da curva de vazão: a) usando o modelo de regressão parabólica dado por $Q = a + b\,h + c\,h^2$; b) considerando que a sua forma é do tipo $Q = a\,(h - h_0)^b$.

Tabela 15 – Pares de valores de caudais instantâneos, Q, e das correspondentes alturas hidrométricas, h, relativos a uma estação hidrométrica.

h(m)	Q (m^3/s)	h(m)	Q (m^3/s)	h(m)	Q (m^3/s)	h(m)	Q (m^3/s)
0.5	12	1.91	170	4.73	990	8.21	2540
0.8	40	2.36	240	4.87	990	8.84	2840
1.19	90	2.7	300	5.84	1260	9.64	3320
1.56	120	4.07	680	7.19	1920	----	----

Solução: a) A solução das equações normais de regressão (ver exercício 15) necessita dos seguintes valores N=15, $\sum Q = 15512.00$ m^3/s, $\sum h = 64.41$ m, $\sum (Q\,h) = 113432.00$ m^4/s, $\sum (Q\,h^2) = 905380.75$ m^5/s, $\sum h^2 = 408.18$ m^2, $\sum h^3 = 3045.57$ m^3 e $\sum h^4 = 24564.94$ m^4. A substituição destes valores nas equações normais do modelo parabólico conduz às estimativas dos coeficientes de regressão a=-33.1195, b=53.6034 e c=30.7612. A Figura 17 mostra o gráfico do modelo de regressão ajustado à amostra de pares de valores (h,Q). O coeficiente de

determinação, calculado pela equação (37), resulta em R^2=0.9989 e significa a parcela da variância dos caudais instantâneos que foi explicada pelas alturas hidrométricas.

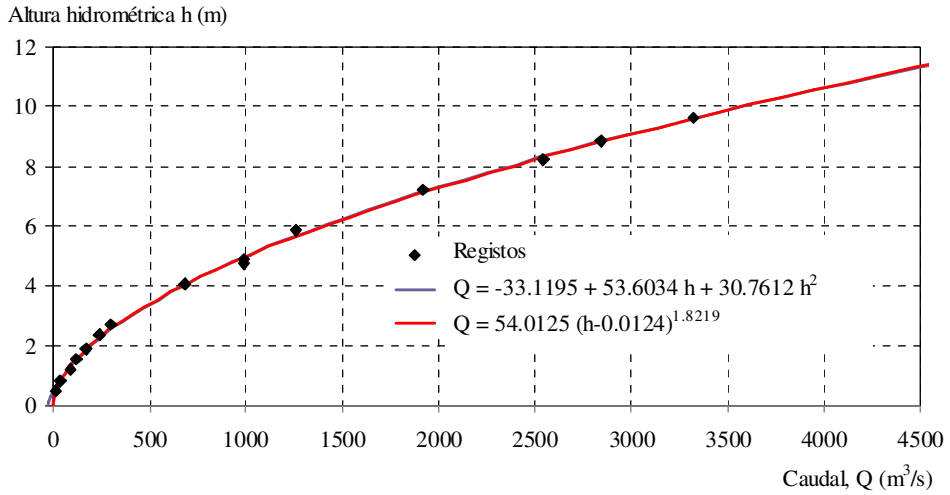

Figura 17 – Curvas de vazão para os dois possíveis modelos definidos no exercício 16.

b) Aplicando logaritmos à equação $Q = a\,(h - h_0)^b$, resulta $\ln Q = \ln a + b \ln (h - h_0)$ que traduz a equação de uma recta em que as abcissas são os valores de $\ln (h - h_0)$ e as ordenadas, os de $\ln Q$. Deste modo, é válido aplicar a análise de regressão linear simples aos pares de valores $(\ln (h - h_0), \ln Q)$. Existem, contudo, três parâmetros da curva de vazão a estimar – a, b, h_0 – ou seja, mais um do que os susceptíveis de serem directamente obtidas por aquela análise. Para resolver o problema, basta arbitrar o valor de h_0 que mais aproxima de uma recta a relação entre $\ln (h - h_0)$ e ln Q, e aplicar a análise de regressão linear para estimar os restantes dois parâmetros. Para cada valor de h_0 arbitrado resulta uma equação para a curva de vazão que, aplicada às alturas hidrométricas utilizadas no seu estabelecimento, h, conduz a estimativas de caudais, $\hat{Q}$, que naturalmente diferem dos caudais que também foram utilizados naquele estabelecimento, Q. A solução do problema traduzir-se-á no conjunto de três parâmetros - a, b, h_0 - que obedeçam a um certo critério de optimização, por exemplo, maximizar a correlação entre os caudais observados, Q, e os estimados a partir da curva de vazão, $\hat{Q}$. A Tabela 16 exemplifica o cálculo descrito.

Tabela 16 – Cálculo dos parâmetros da curva de vazão definida por $Q = a\,(h - h_0)^b$.

h (m)	Q (m^3/s)	h_0(m) (m)	ln Q	ln (h-h_0)	$\hat{Q}$ (m^3/s)	h_0(m) (m)	ln Q	ln (h-h_0)	$\hat{Q}$ (m^3/s)	…	h_0(m) (m)	ln Q	ln (h-h_0)	$\hat{Q}$ (m^3/s)
0.5	12		2.48	0.41	25.17		2.48	-0.69	14.78			2.48	-0.72	14.59
0.8	40		3.69	0.59	4.90		3.69	-0.22	35.00			3.69	-0.24	35.97
1.19	90		4.50	0.78	13.78		4.50	0.17	72.49			4.50	0.16	74.15
1.56	120		4.79	0.94	27.88		4.79	0.44	119.10			4.79	0.44	121.43
1.91	170		5.14	1.07	47.21		5.14	0.65	172.63			5.14	0.64	175.59
2.36	240		5.48	1.21	81.87		5.48	0.86	254.45			5.48	0.85	258.16
2.7	300		5.70	1.31	116.20		5.70	0.99	325.68			5.70	0.99	329.89
4.07	680	-1.000000	6.52	1.62	338.06	0.000000	6.52	1.40	691.24	…	0.012404	6.52	1.40	696.76
4.73	990		6.90	1.75	499.83		6.90	1.55	910.56			6.90	1.55	916.20
4.87	990		6.90	1.77	539.25		6.90	1.58	960.59			6.90	1.58	966.20
5.84	1260		7.14	1.92	865.08		7.14	1.76	1340.28			7.14	1.76	1345.19
7.19	1920		7.56	2.10	1486.19		7.56	1.97	1962.52			7.56	1.97	1964.83
8.21	2540		7.84	2.22	2098.92		7.84	2.11	2503.03			7.84	2.10	2502.02
8.84	2840		7.95	2.29	2544.18		7.95	2.18	2866.46			7.95	2.18	2862.78
9.64	3320		8.11	2.36	3187.53		8.11	2.27	3360.01			8.11	2.26	3352.24
ln a			2.1707				3.9645					3.9892		
a			8.7644				52.6951					54.0125		
b			2.6022				1.8338			…		1.8219		
Coeficiente de correlação entre Q e $\hat{Q}$			0.989470				0.999459					0.999462		

Conforme indicado na tabela, partiu-se de um valor inicial de h_0=-1.000, depois do que se alterou para h_0=0.000, com obtenção, após várias iterações, do valor final de sensivelmente h_0=0.0124, correspondente aos valores, também finais, dos parâmetros da curva de vazão de a=54.0125 e b=1.8219, obtidos a partir dos valores intermédios de $\sum y = \sum \ln Q = 90.6970$, $\sum x = \sum \ln (h - h_0) = 16.9381$, $\sum (x\ y) = \sum \left[\ln (h - h_0) \ln Q\right] = 124.1603$ e $\sum x^2 = \sum \left[\ln (h - h_0)\right]^2 = 31.0621$. Na solução, o coeficiente de correlação entre caudais observados, Q, e estimados a partir da curva de vazão, $\hat{Q}$, é igual a R=0.99946 e o correspondente coeficiente de determinação de R^2=0.9989, ou seja, para a precisão numérica adoptada, igual ao do modelo parabólico. A curva de vazão para o modelo definido por $Q = a\ (h - h_0)^{\ b}$ está também representada na Figura 17. Anota-se que, estando-se em presença de um problema de análise de regressão linear, embora no campo de transformada logarítmicas, os coeficientes de regressão que figuram nas equações (35) e (36) podem ser obtidos a partir da amostra de pares de valores utilizados naquela análise por funções implementadas no *software* Microsoft Excel, designadamente pela função INTERCEPÇÃO (versão em Português) ou INTERCEPT (versão em Inglês) para a ordenada na origem, *a*, e função INCLINAÇÃO (versão em Português) ou SLOPE (versão em Inglês), para o declive da recta de regressão, *b*.

Referências bibliográficas

Ang, A.H.S.; W. T. Tang (2007). *Probability concepts in engineering. Emphasis on Applications to Civil and Environmental Engineering*, 2ª Edição, John Wiley & Sons Inc., Nova Iorque, EUA.

Benjamin, J.; C. A. Cornell (1970). *Probability, statistics and decisions for Civil Engineers*, McGraw-Hill, Nova Iorque, EUA.

Chow, V. T. (1954). "The log-probability law and its engineering applications", *Proceedings of the American Society of Civil Engineers* 80, Paper No. 536, p. 1-25.

Griffis, V. W.; J. R. Stedinger (2007). "Log-Pearson type 3 distribution and its application in flood frequency analysis. II: parameter etimation methods", *Journal of Hydrologic Engineering*, Vol. 12, Nº 5, p. 492-500.

Henriques, A. G. (1990). *Modelos de distribuição de frequências de caudais de cheia*. Dissertação de Doutoramento em Engenharia Civil, Instituto Superior Técnico, Lisboa.

Hosking, J. R. M.; J. R. Wallis (1997). *Regional frequency analysis: an approach based on L-moments*. Cambridge University Press, Cambridge, Reino Unido.

Kite, G.W. (1988). *Frequency and risk analysis in Hydrology*, Water Resources Publications, Littleton (CO), EUA.

Meylan P., A. C. Favre; A. Musy (2008). *Hydrologie fréquentielle – une science prédictive*, Presses Polytechniques et Universitaires Romandes, Lausanne, Suiça.

Montgomery D. C.; E. A. Peck (1992). *Introduction to linear regression analysis*, 2ª Edição, John Wiley & Sons, Nova Iorque, EUA.

Naghettini M.; E. J. A. Pinto (2007). *Hidrologia estatística*, CPRM, Belo Horizonte (MG).

Rao A. R.; K. Hamed (2000). *Flood frequency analysis*, CRC Press, Boca Raton (FL), EUA.

Quintela, A.C.; Portela, M.M. (2002). "A modelação hidrológica em Portugal nos últimos 25 anos do século XX nas perspectivas determinística, probabilística e estocástica", *Revista Brasileira de Recursos Hídricos*, RBRH, Vol. 7 (4) Edição Comemorativa, pp. 51-64, ISSN 1414 381X, Brasil.