HIDROLOGIA ESTATÍSTICA vem preencher significativa lacuna na literatura técnica especializada em língua portuguesa no campo dos recursos hídricos.
O conhecimento das ferramentas de estatística é fundamental para a evolução e para a prática da Hidrologia, onde encontra diversificada gama de aplicações nas atividades rotineiras ligadas aos estudos e projetos de engenharia hidrológica, que necessitam das teorias probabilísticas para a sua solução.

Conhecer e investigar as variáveis do meio físico são atributos comuns entre os conceitos aqui registrados e o Serviço Geológico do Brasil – CPRM.
O livro apresenta o material didático capaz de orientar a pesquisa, e, com essa iniciativa, a instituição amplia a visibilidade do seu papel de agente promotor dos levantamentos hidrológicos básicos no país.

HIDROLOGIA ESTATÍSTICA é publicação dirigida para os profissionais do setor, bem como para a formação de alunos de graduação e pós-graduação. Municia o leitor com princípios introdutórios, análise de dados, teoria das probabilidades, variáveis aleatórias discretas e contínuas, análise de freqüência, correlação e regressão. Destaca também técnicas mais sofisticadas de tratamento, manipulação e representação de dados estatísticos, com exemplos práticos reais e selecionados da rede hidrometeorológica operada pela CPRM.

www.cprm.gov.br

CPRM
Serviço Geológico do Brasil

Secretaria de
Geologia, Mineração e
Transformação Mineral

Ministério de
Minas e Energia

BRASIL
UM PAÍS DE TODOS
GOVERNO FEDERAL

Período Comemorativo

planetaterra®
Ciências da Terra para a Sociedade

ANO INTERNACIONAL DO PLANETA TERRA – 2008

AGOSTO
DE 2007

Hidrologia Estatística

ISBN 978-85-7499-023-1

CPRM

CPRM
Serviço Geológico do Brasil

Hidrologia Estatística

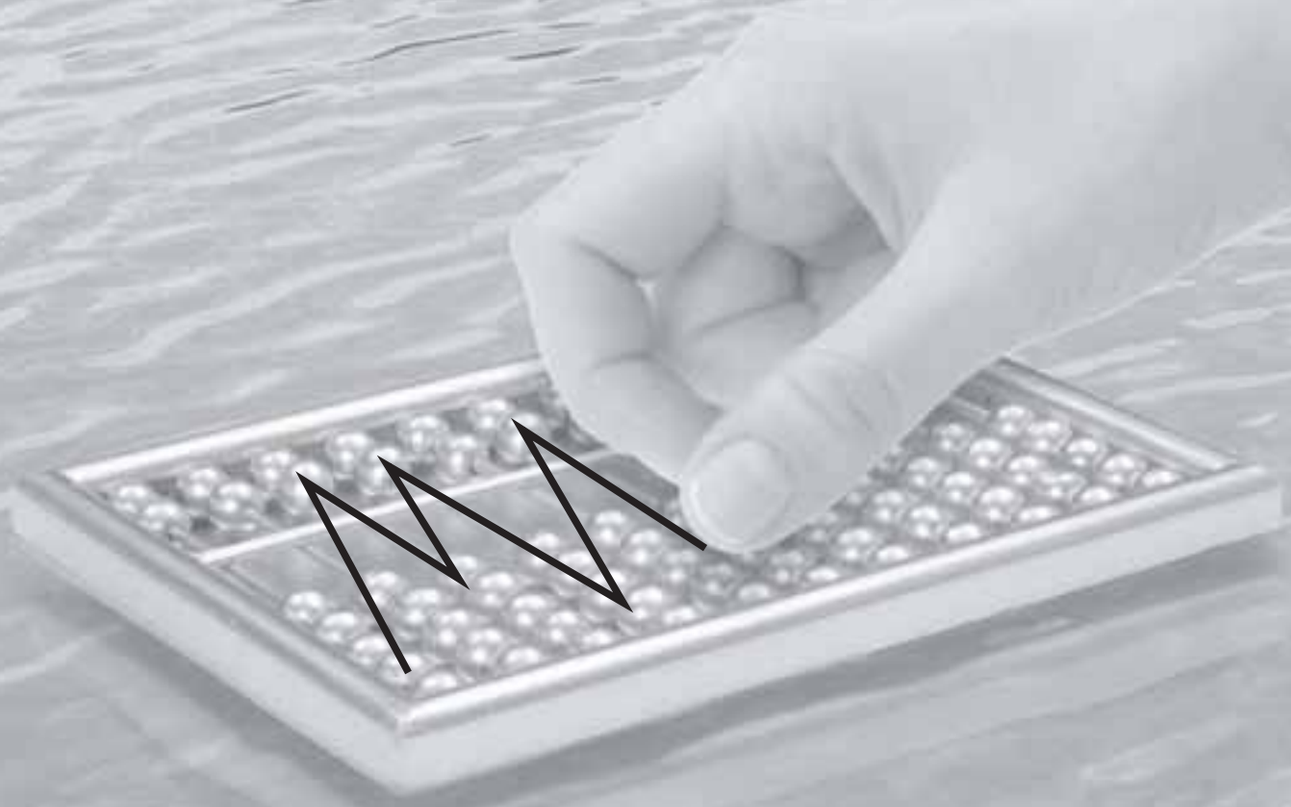
MAURO NAGHETTINI
ÉBER JOSÉ DE ANDRADE PINTO

2007

CPRM
Serviço Geológico do Brasil

Hidrologia Estatística

MAURO NAGHETTINI
ÉBER JOSÉ DE ANDRADE PINTO



2007

MINISTÉRIO DE MINAS E ENERGIA
NELSON HUBNER
Ministro Interino

SECRETARIA DE GEOLOGIA, MINERAÇÃO E TRANSFORMAÇÃO MINERAL
CLÁUDIO SCLIAR
Secretário

SERVIÇO GEOLÓGICO DO BRASIL – CPRM
AGAMENON SÉRGIO LUCAS DANTAS
Diretor-Presidente

MANOEL BARRETTO DA ROCHA NETO
Diretor de Geologia e Recursos Minerais

JOSÉ RIBEIRO MENDES
Diretor de Hidrologia e Gestão Territorial

FERNANDO PEREIRA DE CARVALHO
Diretor de Relações Institucionais e Desenvolvimento

ÁLVARO ROGÉRIO ALENCAR SILVA
Diretor de Administração e Finanças

FREDERICO CLÁUDIO PEIXINHO
Chefe do Departamento de Hidrologia

ERNESTO VON SPERLING
Chefe da Divisão de Marketing e Divulgação

COORDENAÇÃO E AUTORIA
MAURO NAGHETTINI
ÉBER JOSÉ DE ANDRADE PINTO

COLABORAÇÃO
ALICE SILVA DE CASTILHO
ELIZABETH GUELMAN DAVIS
ERNESTO VON SPERLING
FERNANDO ALVES LIMA
FREDERICO CLÁUDIO PEIXINHO
JOSÉ MÁRCIO HENRIQUES SOARES
MARCELO JORGE MEDEIROS
MÁRCIO DE OLIVEIRA CÂNDIDO

CPRM
Serviço Geológico do Brasil

Hidrologia Estatística

Mauro Naghettini
Éber José de Andrade Pinto

Belo Horizonte
Agosto de 2007

Coordenação Editorial a cargo da
Divisão de Marketing e Divulgação
Diretoria de Relações Institucionais e Desenvolvimento
Serviço Geológico do Brasil - CPRM

Publishers
Ernesto von Sperling
José Márcio Henriques Soares

Naghettini, Mauro

N147 Hidrologia estatística. / Mauro Naghettini; Éber José de Andrade Pinto. — Belo Horizonte: CPRM, 2007.
552 p.

Executado pela CPRM – Serviço Geológico do Brasil, Superintendência Regional de Belo Horizonte. Hidrologia. 2. Recursos Hídricos. 3. Engenharia Hidráulica. 4. Estatística.

I. Pinto, Éber José de Andrade.
II. CPRM- Serviço Geológico do Brasil.
III.Título.

ISBN 978-85-7499-023-1

APRESENTAÇÃO

A água, um bem natural de inestimável valor para humanidade, projeta-se, no cenário mundial, como tema central na agenda política das nações, face aos desafios relacionados com a sua escassez e a ocorrência de eventos extremos como secas e inundações, que inibem o desenvolvimento das nações e geram conflitos, degradando a qualidade de vida das populações em várias regiões do planeta.

Torna-se então cada vez mais imperioso o conhecimento sobre a ocorrência da água nos continentes, fundamental para a sua adequada gestão e o conseqüente aproveitamento racional deste valioso recurso.

A Hidrologia, como ciência da Terra que estuda a ocorrência, a distribuição, o movimento e as propriedades da água na atmosfera, na superfície e no subsolo, tem buscado, cada vez mais, uma abordagem sistêmica e interdisciplinar, integrando-se às outras geociências com o objetivo de expandir o conhecimento existente das diversas fases do ciclo da água no planeta. Ao tratar o ciclo hidrológico de forma integrada, visa também descrever o passado e prever o futuro.

Dada à natureza probabilística do fenômeno hidrológico, a Estatística é uma área de conhecimento importante da Hidrologia, utilizada na avaliação do comportamento dos processos hidrológicos.

O Serviço Geológico do Brasil, em consonância com a sua missão de gerar e difundir conhecimento hidrológico teve a iniciativa de produzir esta publicação, a qual representa uma relevante contribuição para a comunidade técnica e científica e a sociedade brasileira.

Agamenon Sérgio Lucas Dantas
Diretor-Presidente
Serviço Geológico do Brasil - CPRM

PREFÁCIO

É para mim um grande prazer escrever um prefácio para este excelente livro. Os dois autores conseguiram produzir um texto de qualidade que deveria ser usado como livro texto para estudantes de hidrologia e engenharia de recursos hídricos não apenas no Brasil, mas de maneira mais abrangente, em todos os países de língua portuguesa. Além de ser extremamente útil para o ensino, o livro também encontrará lugar na biblioteca de profissionais destas áreas, que encontrarão no mesmo um resumo muito útil das características das distribuições de probabilidade largamente encontradas na literatura de recursos hídricos. Começando com noções simples das essenciais análises gráficas dos dados hidrológicos, o livro fornece uma clara visão do papel importante que as considerações sobre probabilidade devem ter durante a modelação, o diagnóstico de ajuste de modelos, a previsão, e a avaliação das incertezas nas previsões fornecidas pelos modelos. Uma excelente apresentação é feita sobre como estabelecer relações entre duas ou mais variáveis e sobre a forma como estas relações são usadas para transferência de informação entre postos através da regionalização.

A grande variedade de exemplos discutidos no livro é especialmente admirável, bem como a inclusão de exercícios que ilustram e estendem o material de cada capítulo. O livro irá certamente se constituir numa base sólida para estudantes e outras pessoas interessadas em explorar não apenas os muitos métodos estatísticos nele descritos mas também outros assuntos associados como "bootstrap methods", análise Bayesiana e Modelos Lineares Generalizados.

A nível pessoal, foi para mim um privilégio ter conhecido muitos dos autores mencionados na longa lista de referências deste livro, e ter trabalhado com vários deles na década de 1970 e início da década de 1980 quando o Institute of Hydrology (IH) do Reino Unido estava na liderança de muitas pesquisas em hidrologia. Entre os autores mencionados no livro incluem-se Cunnane que trabalhou no Flood Studies Report (FSR) do NERC (National Environment Research Council); Reed, que aprimorou a metodologia do FSR, e foi um dos autores da versão atualizada do relatório, intitulada Flood Estimation Handbook (FEH); Hosking, que desenvolveu várias de suas idéias com a colaboração de Wallis durante sua licença sabática passada no IH; Sutcliffe, um dos membros fundadores do IH, e Wiltshire, cujo trabalho é descrito no Capítulo 10. De um grupo de aproximadamente 20 pesquisadores trabalhando no IH em procedimentos estatísticos aplicados a estudos de cheias e estiagem, durante aquele tempo, nada menos que sete tornaram-se, subseqüentemente, professores titulares em universidades britânicas.

Tenho certeza de que os capítulos que seguem este curto Prefácio constituir-se-ão numa sólida fundação para o conhecimento dos estudantes que por sua vez farão grandes contribuições ao gerenciamento dos recursos hídricos no Brasil, e em outros paises, durante as décadas que estão por vir, que são de incertezas sobre as mudanças climáticas, o rápido desenvolvimento urbano, e o fornecimento de energia.

Porto Alegre, Agosto de 2007.

Robin Thomas Clarke

PREFÁCE

It gives me much pleasure to write a Preface to this excellent book. The two authors have succeeded in producing a first-class text which should be prescribed reading for students of hydrology and water resource engineering not only in Brazil but more widely throughout the Lusophone world. Besides being superbly useful for teaching, the book will also find a place on the bookshelves of mature practitioners who will find in it a useful summary of the characteristics of probability distributions widely encountered in the water resource literature. Starting with simple notions of the essential graphical examination of hydrological data, the book gives a very lucid account of the role that probability considerations must play during modeling, diagnosis of model fit, prediction, and evaluating the uncertainty in model predictions. An excellent account is given of how to establish relationships between two or more variables, and of the way in which such relationships are used for the transfer of information between sites by regionalization. The wide range of examples discussed in the book is especially admirable, and the inclusion of exercises which both illustrate and extend the material given in each chapter. The book will provide a very firm basis for students and others who need to explore not only the many statistical methods described within its covers but also the associated fields of bootstrap methods, Bayesian analysis and Generalized Linear Models.

At a personal level, it has been a privilege for me to have known a number of the authors mentioned in the book's extensive list of references, and to have worked with several of them during the decade of the 1970s and early 1980s when the UK Institute of Hydrology (IH) was at the forefront of much hydrological research. Those mentioned in the book include, but are not limited to, Cunnane who worked on the UK Flood Study Report (FSR); Reed, who developed the FSR methodology still further, and was joint author of its successor the Flood Estimation Handbook (FEH); Hosking, who developed many of his ideas during the sabbatical year that Wallis spent collaborating with him at Wallingford; Sutcliffe, one of the founder members of IH; and Wiltshire, whose work is described in Chapter 10. Of a group of about 20 researchers working at IH on statistical procedures and modeling applied to flood and drought studies during that time, no less than seven subsequently held senior chairs at British universities.

I have every confidence that the chapters that follow this short Preface will lay a similar foundation for students who will in their turn make major contributions to the management of water resources in Brazil, and elsewhere, during the coming decades of uncertainty about climate change, rapid urban development, and energy supplies.

Porto Alegre, August 2007.

Robin Thomas Clarke

INTRODUÇÃO

A humanidade, desde seus primórdios, sempre se interessou em observar o comportamento das variáveis hidrológicas, tais como, níveis em curso d'água e as precipitações. O desenvolvimento científico e tecnológico possibilitou o registro desse comportamento ao longo do tempo. O acumulo dessas informações permite a formação de séries, as quais são analisadas utilizando a estatística como uma ferramenta básica e fundamental, de forma que o conhecimento dos conceitos estatísticos é indispensável ao desenvolvimento de estudos em hidrologia e em ciências naturais.

Este livro tem por objetivo fornecer aos profissionais que trabalham com recursos hídricos e as ciências ambientais, bem como aos estudantes de graduação e pós-graduação dessas áreas do conhecimento, um texto em português sobre os conceitos básicos de estatística, enfatizando a sua aplicação em hidrologia e nas ciências naturais.

A publicação foi organizada em dez capítulos, os quais apresentam a teoria, exemplos de emprego em hidrologia e nas ciências naturais de cada tópico analisado e, ao final, exercícios para treinamento e consolidação do aprendizado. O primeiro capítulo, Introdução à Hidrologia Estatística, apresenta brevemente as idéias de processos, variáveis, séries e dados hidrológicos. A análise preliminar de dados hidrológicos é descrita no segundo capítulo. Os fundamentos da teoria de probabilidades são expostos em detalhes no capítulo 3. A descrição dos modelos discretos de distribuição de probabilidades é o escopo do capítulo 4 e os principais modelos contínuos são apresentados no capítulo 5. A estimação pontual e por intervalos dos parâmetros dos modelos probabilísticos é delineada no capítulo 6. As linhas gerais para construção dos testes de hipóteses, a formulação dos testes paramétricos para populações normais, a lógica inerente aos testes não paramétricos, os testes de aderência e de detecção dos pontos amostrais atípicos formam o conteúdo do capítulo 7. No oitavo capítulo são descritos os procedimentos da análise de freqüência local de variáveis hidrológicas. A apresentação dos conceitos básicos que possibilitam a realização de estudos de correlação e regressão linear entre duas ou mais variáveis é efetuada no capítulo 9. Finalmente, no décimo capítulo, são descritos os métodos de análise de freqüência regional, com maior detalhe para o método *index-flood*, utilizando os momentos-L e as estatísticas-L.

A Diretoria de Hidrologia e Gestão Territorial através do Departamento de Hidrologia expressa o compromisso de disseminar o conhecimento geocientífico, ao promover e incentivar a publicação de um livro sobre hidrologia estatística, cujo tema apresenta grande importância no desenvolvimento dos trabalhos em recursos hídricos, uma das áreas fundamentais de atuação do Serviço Geológico do Brasil.

Frederico Cláudio Peixinho
Chefe do Departamento de Hidrologia
Serviço Geológico do Brasil - CPRM

DEDICATÓRIA

Para meus pais, Nilo (in memoriam) e Augusta, meus exemplos permanentes de perseverança e dignidade.

MN

Para meus pais, Dalva Urbano de Resende e José Maria de Andrade Pinto e, ao fraterno amigo do movimento escoteiro, Luiz Tadeu Coelho. Pessoas muito queridas que partiram no último ano hidrológico (2005-2006). E para os meus filhos, Lúcio e Maria Cecília, fontes de alegria e sentido nessa existência.

EJAP

SOBRE OS AUTORES

Mauro Naghettini

Graduou-se em Engenharia Civil pela Universidade Federal de Minas Gerais, em 1977. Mestre em Hidrologia pela École Polytechnique Fédérale de Lausanne, Suíça, em 1979 e PhD em Engenharia de Recursos Hídricos pela University of Colorado at Boulder, Estados Unidos, em 1994. De 1979 a 1989, foi engenheiro da Divisão de Hidrologia da Companhia Energética de Minas Gerais (CEMIG), tendo atuado no planejamento, projeto e operação de aproveitamentos hidrelétricos. Desde 1989 é professor do Departamento de Engenharia Hidráulica e Recursos Hídricos da UFMG, com atividades de pesquisa, ensino e extensão universitária. Atua no Programa de Pós-Graduação em Saneamento, Meio Ambiente e Recursos Hídricos da UFMG, lecionando diversas disciplinas, entre as quais "Hidrologia Estatística", orientando alunos de mestrado e doutorado, com ativa participação nas linhas de pesquisa "Modelos Estocásticos em Hidrologia" e "Modelos de Simulação e Previsão Hidrológica". Autor de vários artigos publicados em periódicos especializados e anais de simpósios e congressos técnicos. É pesquisador do CNPq desde 1996 e membro do conselho editorial da Revista Brasileira de Recursos Hídricos. Consultor de diversas empresas atuantes na área de engenharia de recursos hídricos.

E-mail: naghet@netuno.lcc.ufmg.br.

Eber José de Andrade Pinto

Engenheiro Civil graduado pela Escola de Engenharia da Universidade Federal de Minas Gerais em abril de 1992. Mestre e Doutor em Engenharia de Recursos Hídricos pelo Programa de Pós-Graduação em Saneamento, Meio Ambiente e Recursos Hídricos da Universidade Federal de Minas Gerais em 1996 e 2005, respectivamente. Trabalha, desde fevereiro de 1994, como engenheiro hidrólogo na CPRM – Serviço Geológico do Brasil, onde ingressou por concurso e atuou em projetos de obtenção de dados hidrométricos básicos, de consistência de dados hidrológicos, de avaliação da disponibilidade de recursos hídricos, de definição das relações intensidade-duração-freqüência, de regionalização de variáveis hidrológicas, de avaliação de estruturas de captação de águas de chuva, de implantação de bacias representativas, de operação do sistema de alerta de cheias da bacia do rio Doce, de definição de planícies de inundação, de zoneamento ecológico econômico, entre outras atividades. Autor de artigos publicados em periódicos especializados e anais de simpósios e congressos técnicos. Lecionou disciplinas de Hidrologia em cursos de especialização do CEFET-MG e do Instituto de Educação Continuada da PUC-MG.

E-mail: eber@bh.cprm.gov.br.

AGRADECIMENTOS

"Os autores agradecem o apoio institucional do Serviço Geológico do Brasil – CPRM, sem o qual, este livro não poderia ter sido publicado. Um agradecimento especial ao Marcelo Jorge Medeiros, da CPRM-Brasília, pelas iniciativas de propor a preparação deste livro e de não medir esforços para viabilizá-lo. Os autores agradecem as sugestões e idéias de diversos colegas, entre os quais, destacam Alice Silva de Castilho, Elizabeth Guelman Davis, Márcio de Oliveira Cândido e Fernando Alves Lima. Também agradecem a Frederico Cláudio Peixinho e Ernesto von Sperling pelo suporte e a colaboração durante a elaboração do livro. Finalmente, os autores agradecem aos seus familiares pela compreensão e estímulo."

SUMÁRIO

LISTA DE ANEXOS

LISTA DE FIGURAS

CAPÍTULO 1
INTRODUÇÃO À HIDROLOGIA ESTATÍSTICA

CAPÍTULO 2
ANÁLISE PRELIMINAR DE DADOS HIDROLÓGICOS

CAPÍTULO 3
TEORIA ELEMENTAR DE PROBABILIDADES

CAPÍTULO 4
VARIÁVEIS ALEATÓRIAS DISCRETAS: DISTRIBUIÇÕES E APLICAÇÕES

CAPÍTULO 5
VARIÁVEIS ALEATÓRIAS CONTÍNUAS: DISTRIBUIÇÕES APLICAÇÕES

CAPÍTULO 6
ESTIMAÇÃO DE PARÂMETROS

CAPÍTULO 7
TESTES DE HIPÓTESES

CAPÍTULO 8
ANÁLISE LOCAL DE FREQÜÊNCIA DE VARIÁVEIS HIDROLÓGICAS

CAPÍTULO 9
CORRELAÇÃO E REGRESSÃO

CAPÍTULO 10
ANÁLISE REGIONAL DE FREQÜÊNCIA DE VARIÁVEIS HIDROLÓGICAS

LISTA DE TABELAS

CAPÍTULO 1
INTRODUÇÃO À HIDROLOGIA ESTATÍSTICA

CAPÍTULO 2
ANÁLISE PRELIMINAR DE DADOS HIDROLÓGICOS

CAPÍTULO 3
TEORIA ELEMENTAR DE PROBABILIDADES

CAPÍTULO 4
VARIÁVEIS ALEATÓRIAS DISCRETAS: DISTRIBUIÇÕES E APLICAÇÕES

CAPÍTULO 5
VARIÁVEIS ALEATÓRIAS CONTÍNUAS: DISTRIBUIÇÕES E APLICAÇÕES

CAPÍTULO 6
ESTIMAÇÃO DE PARÂMETROS

CAPÍTULO 7
TESTES DE HIPÓTESES

CAPÍTULO 8
ANÁLISE LOCAL DE FREQÜÊNCIA DE VARIÁVEIS HIDROLÓGICAS

CAPÍTULO 9
CORRELAÇÃO E REGRESSÃO

CAPÍTULO 10
ANÁLISE REGIONAL DE FREQÜÊNCIA DE VARIÁVEIS HIDROLÓGICAS

CAPÍTULO 1

INTRODUÇÃO À HIDROLOGIA ESTATÍSTICA

CAPÍTULO 1 INTRODUÇÃO À HIDROLOGIA ESTATÍSTICA

Esse capítulo apresenta o contexto no qual se insere a hidrologia estatística e introduz brevemente as idéias de processos, variáveis, séries e dados hidrológicos.

1.1 – Caracterização dos Fenômenos e Processos Hidrológicos

A 'Hidrologia' é a geociência que investiga os fenômenos que determinam a distribuição espaço-temporal da água, em nosso planeta, sob os atributos de quantidade, de qualidade e de interação com as sociedades humanas. Os fenômenos hidrológicos são aqueles que definem os mecanismos de armazenamento e transporte entre as diversas fases do ciclo da água em nosso planeta, com atenção especial para as áreas continentais. As intensidades com que esses fenômenos se manifestam apresentam uma marcante variabilidade ao longo do tempo e do espaço, em decorrência das variações, algumas regulares e muitas irregulares, dos climas global e regional, bem como das particularidades regionais e locais, sob os aspectos meteorológicos, geomorfológicos, de propriedades e uso do solo, entre tantos outros. A 'Hidrologia Aplicada' utiliza os princípios da hidrologia para planejar, projetar e operar sistemas de aproveitamento e controle de recursos hídricos; a consecução desses objetivos requer a quantificação confiável das variabilidades espaciais e/ou temporais presentes em fenômenos hidrológicos tais como: precipitação, escoamento e armazenamento superficiais, evapotranspiração, infiltração, escoamento e armazenamento sub-superficiais, propriedades físico-químicas e biológicas da água, conformações geomorfológicas, transporte de sedimentos, etc.

As intensidades com que os fenômenos hidrológicos ocorrem, podem ser postas como funções do tempo, ou do espaço, ou de ambos, em escalas geográficas diversas que vão desde a global até a local, passando pela escala usual da bacia hidrográfica. A tais funções associa-se o conceito de processos hidrológicos. A função do tempo que descreve a evolução contínua das vazões que atravessam uma certa seção fluvial é um exemplo de um processo hidrológico. Os processos associados ao ciclo hidrológico podem ser classificados, grosso modo, em determinísticos ou estocásticos embora, em geral, sejam, de fato, uma combinação de ambos.

Os processos hidrológicos determinísticos são aqueles que resultam da aplicação direta de leis da Física, Química ou Biologia. Em hidrologia, são raríssimas as ocorrências das regularidades inerentes aos processos puramente determinísticos, nos quais as variações espaço-temporais podem ser completamente explicadas

por um número limitado de variáveis, a partir de relações funcionais ou experimentais unívocas. A resposta hidrológica de uma superfície completamente impermeável, de geometria simples e totalmente definida, a um pulso conhecido, uniforme e homogêneo de precipitação, pode ser considerado um raro exemplo de um processo hidrológico puramente determinístico. Uma curva-chave estável, válida para uma seção encaixada em um leito rochoso de um trecho fluvial, com controle hidráulico invariável e inequivocamente definido, para a qual tenha sido precisamente determinada a histerese devida ao escoamento não permanente, é outro raro exemplo de uma relação puramente determinística. Evidentemente, em rios naturais, com leitos móveis ou controle hidráulico variável, a situação anteriormente descrita é de ocorrência muito improvável, estando a relação cota-descarga sujeita à complexa interferência de uma infinidade de fatores aleatórios.

Quase todos os processos hidrológicos são considerados estocásticos, ou governados por leis de probabilidades, por conterem componentes aleatórias as quais se superpõem a regularidades eventualmente explicitáveis, tais como as estações do ano ou às variações da radiação solar no topo da atmosfera ao longo da órbita da Terra em torno do Sol. Nesse sentido, em um dado ponto do espaço geográfico, são considerados processos hidrológicos estocásticos a precipitação, a evapo-transpiração, os escoamentos superficial e sub-superficial, os afluxos de sedimento em suspensão, as concentrações de oxigênio dissolvido, as conformações do leito fluvial, as temperaturas da água, as capacidades de infiltração, dentre tantos outros. Rigorosamente, pela forçosa existência de componentes aleatórios, inexistem relações funcionais e unívocas entre as variáveis características de processos hidrológicos. Tomando-se como exemplo as características relevantes das enchentes em uma certa bacia hidrográfica, é notável a presença de forte aleatoriedade por tratar-se de um fenômeno no qual nem todos os fatores causais e/ou influentes, bem como suas interdependências nas escalas espacial e temporal, podem ser precisamente explicitados e determinados. De fato, as distribuições espacial e temporal da precipitação, a velocidade e a direção de deslocamento da tormenta sobre a bacia, as variações temporais e espaciais das perdas por interceptação, evapo-transpiração e infiltração, bem como dos teores de umidade do solo, são exemplos do grande número de fatores interdependentes que podem causar cheias ou influir em sua formação e intensificação.

Nesse ponto, poder-se-ia inferir, então, que, se todos os fatores causais pudessem ser definidos e medidos com precisão e se todas as possíveis dependências entre eles puderem ser explicitadas e determinadas, as características relevantes das enchentes de uma dada bacia hidrográfica poderiam ser tratadas como relações puramente determinísticas do tipo causa-efeito. Entretanto, tal possibilidade esbarra

em restrições práticas associadas ao monitoramento preciso e abrangente dos fatores causais, bem como nos limites do conhecimento humano sobre os processos hidrológicos, muito embora sejam inegáveis os avanços continuados da pesquisa científica e do desenvolvimento tecnológico em tais direções. Ao longo do futuro, esses avanços certamente irão reduzir o grau de aleatoriedade presente nos processos hidrológicos, mas não o farão a ponto de torná-los puramente determinísticos.

Estas constatações conduzem ao emprego simultâneo das abordagens determinística e estocástica para a melhor explicitação e para o correto entendimento das regularidades e também das variabilidades inerentes aos processos hidrológicos, de modo a agregá-las em sólido arcabouço científico e tecnológico capaz de proporcionar elementos para a formulação de propostas racionais para questões relativas ao desenvolvimento dos recursos hídricos. Nesse contexto, posto que aos fenômenos hidrológicos associam-se distribuições da variabilidade espaço-temporal de variáveis aleatórias, relativas à quantidade e à qualidade da água, é forçosa a necessidade do emprego da teoria de probabilidades, aqui resumidamente definida como a área da matemática que investiga os fenômenos aleatórios. A teoria de probabilidades apresenta duas ramificações de grande importância para a hidrologia aplicada: a estatística matemática e o estudo de processos estocásticos. A estatística matemática é o ramo da teoria de probabilidades que permite analisar um conjunto limitado de observações de um fenômeno aleatório e extrair inferências quanto à ocorrência de todas as prováveis realizações do fenômeno em questão. O estudo de processos estocásticos refere-se à identificação e interpretação da aleatoriedade presente em tais processos, em geral por meio de modelos matemáticos que buscam estabelecer as possíveis conexões seqüenciais, no tempo e/ou no espaço, entre suas realizações.

O conjunto {teoria de probabilidades - estatística matemática - processos estocásticos} constitui um amplo corpo teórico que partilha dos mesmos fundamentos e encontra uma diversificada gama de aplicações em hidrologia. Não obstante a fundamentação teórica em comum, é freqüente agruparem-se as aplicações hidrológicas da teoria de probabilidades e da estatística matemática na disciplina 'Hidrologia Estatística', cabendo à 'Hidrologia Estocástica' o estudo dos processos hidrológicos estocásticos. Esta publicação, sob o título 'Hidrologia Estatística', tem por objetivo apresentar os fundamentos da teoria de probabilidades e da estatística matemática, tal como aplicados na identificação e interpretação da aleatoriedade presente nos processos hidrológicos, bem como na formulação e estimação de modelos probabilísticos de suas respectivas variáveis características.

1.2 – Variáveis Hidrológicas

As variações temporais e/ou espaciais dos fenômenos do ciclo da água podem ser descritas pelas variáveis hidrológicas. São exemplos de variáveis hidrológicas o número anual de dias consecutivos sem precipitação, em um dado local, e a intensidade máxima anual da chuva de duração igual a 30 minutos. Outros exemplos são a vazão média anual de uma bacia hidrográfica, o total diário de evaporação de um reservatório ou a categoria dos 'estados do tempo' empregada em alguns boletins meteorológicos.

As flutuações das variáveis hidrológicas, ao longo do tempo ou do espaço, podem ser quantificadas, ou categorizadas, por meio de observações ou medições, as quais, em geral, são executadas de modo sistemático e de acordo com padrões nacionais ou internacionais. Por exemplo, as variações temporais dos níveis d'água médios diários da seção fluvial de uma grande bacia hidrográfica podem ser monitoradas pelas médias aritméticas das leituras das réguas linimétricas, tomadas às 7 e às 17 horas de cada dia. Da mesma forma, as variações dos totais diários de evaporação de um lago podem ser estimadas pelas leituras dos níveis de um tanque evaporimétrico local, tomadas regularmente às 9 horas da manhã. Essas são exemplos de variáveis hidrológicas, as quais, por estarem associadas a processos estocásticos, são descritas por distribuições de probabilidade e consideradas variáveis aleatórias. Ao conjunto das observações de uma certa variável hidrológica, tomadas em tempos e/ou locais diferentes, dá-se o nome de amostra, a qual contem um número limitado de realizações daquela variável. É certo que a amostra não contem todas as possíveis observações daquela variável, as quais estarão contidas na população que reúne a infinidade de todas as possíveis realizações do processo hidrológico em questão. O objeto principal da hidrologia estatística é o de extrair da amostra, os elementos suficientes para concluir, por exemplo, com que probabilidade a variável hidrológica, em questão, irá igualar ou superar um certo valor de referência, o qual ainda não foi observado, encontrando-se, portanto, fora da amplitude estabelecida pelos limites amostrais.

Segundo as características de seus resultados possíveis, as variáveis aleatórias podem ser classificadas em qualitativas ou quantitativas. As primeiras são aquelas cujos resultados possíveis não podem ser expressos por um número e, sim, por um atributo ou qualidade. As variáveis qualitativas ainda podem ser subdivididas em nominais e ordinais, em consonância com as respectivas possibilidades de seus atributos, ou qualidades, não serem ou serem classificados em modo único. O estado do tempo, entre as possibilidades {'bom', 'chuvoso' e 'nublado'}, é exemplo de uma variável hidrológica qualitativa nominal porque seus resultados não são números e, também, por não serem passíveis de ordenação ou classificação.

De outra forma, o nível de armazenamento de um reservatório, tomado entre as possibilidades {A: excessivamente alto; B: alto; C: médio; D: baixo e E: excessivamente baixo}, representa um exemplo de uma variável hidrológica qualitativa ordinal.

As variáveis hidrológicas quantitativas são aquelas cujos resultados possíveis são expressos por números inteiros ou reais, recebendo a denominação de discretas, no primeiro caso, e contínuas no segundo. O número anual de dias consecutivos sem chuva, em um dado local, é um exemplo de uma variável hidrológica discreta cujos valores possíveis estarão compreendidos integralmente no subconjunto dos números inteiros dado por {0, 1, 2, 3, ..., 366}. Por outro lado, a altura diária máxima anual de precipitação, nesse mesmo local, é uma variável hidrológica contínua porque o conjunto de seus resultados possíveis estará totalmente contido no subconjunto dos números reais não negativos. As variáveis hidrológicas quantitativas ainda podem ser classificadas em limitadas e ilimitadas. As primeiras são aquelas em que os resultados possíveis são limitados superior e inferiormente, seja por condicionantes físicas, seja pelo modo como são medidas. A variável concentração de oxigênio dissolvido em um lago, por exemplo, é limitada inferiormente por zero e superiormente pela capacidade de dissolução de oxigênio do corpo d'água, a qual, por sua vez, é dependente de sua temperatura. Do mesmo modo, a direção do vento local, registrada em um anemômetro, será um ângulo compreendido entre 0 e 360º. Por sua vez, as variáveis ilimitadas não possuem limites inferior e superior definidos. Embora a variável vazão média diária de um curso d'água não pode, evidentemente, ter valores negativos, ela não estará limitada, pelo menos do ponto de vista da hidrologia estatística, a um limiar superior conhecido ou definível, sendo, portanto, uma variável hidrológica, quantitativa, contínua e ilimitada.

As variáveis hidrológicas ainda podem ser classificadas em univariadas, quando a elas associam-se os resultados de apenas um único atributo de quantidade ou qualidade da água, ou multivariadas em caso contrário. As alturas horárias de precipitação em um certo local são um exemplo de variável hidrológica univariada, enquanto a variação conjunta das alturas horárias de chuva, observadas simultaneamente em diversos pontos de uma bacia hidrográfica, pode ser descrita por uma variável hidrológica multivariada. Por representarem o objeto de grande parte das aplicações da hidrologia estatística, a presente publicação ocupar-se-á exclusivamente das variáveis hidrológicas aleatórias quantitativas.

1.3 – Séries Hidrológicas

As variáveis hidrológicas e hidrometeorológicas têm sua variabilidade registrada por meio das chamadas séries temporais, as quais reúnem as observações ou medições daquela variável, organizadas no modo seqüencial de sua ocorrência no tempo (ou espaço). Por limitações impostas pelos processos de medição ou observação, as variáveis hidrológicas, embora apresentem variações instantâneas ou contínuas ao longo do tempo, ou do espaço, têm seus registros separados por determinados intervalos de tempo, ou de distância. Em geral, os intervalos de tempo (ou de distância) entre os registros sucessivos de uma série temporal são eqüidistantes, embora possam existir séries temporais com registros tomados em intervalos irregulares. Em uma bacia hidrográfica de alguns milhares de quilômetros quadrados, por exemplo, as vazões médias diárias, tomadas como médias aritméticas das leituras linimétricas instantâneas das 7 e das 17 horas de cada dia, irão constituir a série temporal representativa da variável hidrológica em questão. Em outra bacia hipotética, de apenas algumas dezenas de quilômetros quadrados e com tempos de concentração de poucas horas, as vazões médias diárias serão insuficientes para demonstrar a variabilidade ao longo do dia; nesse caso, a série temporal mais conveniente deveria ser, por exemplo, aquela formada pelos registros consecutivos de vazões médias horárias.

As séries hidrológicas podem incluir todas as observações disponíveis, coletadas em intervalos de tempo regulares ao longo de vários anos de registros, ou apenas alguns de seus valores característicos como, por exemplo, os máximos anuais ou as médias mensais. No primeiro caso, quando nenhum registro é desprezado, trata-se da chamada série hidrológica completa e, no segundo, quando apenas algumas observações do registro são consideradas, ou quando elas são resumidas por meio de valores médios anuais ou mensais, trata-se da série hidrológica reduzida. A série composta por todas as vazões médias diárias observadas em uma estação fluviométrica é um exemplo de uma série completa, enquanto que aquela composta pelas vazões médias anuais, organizadas de acordo com a ordem cronológica das ocorrências, é um exemplo de série reduzida.

No caso específico de eventos hidrológicos extremos, tais como máximos e mínimos, as séries reduzidas podem ser anuais, quando os registros consecutivos são eqüidistantes no tempo, ou de duração parcial, em caso contrário. A Figura 1.1 ilustra a série de máximos anuais do Rio Paraopeba, na estação fluviométrica de Ponte Nova do Paraopeba (código 40800001), localizada na região centro sul do estado de Minas Gerais, a qual é composta por valores extraídos da série hidrológica completa; para um certo ano, extraiu-se somente um único valor, correspondente à máxima vazão média diária entre as 365 ou 366 observações

daquele ano. Embora a série anual, assim construída, contenha menos informação hidrológica que a série completa, ela reúne as observações geralmente consideradas como essenciais em estudos de vazões de enchentes. Observando a série anual da Figura1.1, vê-se que o ano de 1971 foi excepcionalmente seco e que seu valor máximo é muito baixo quando comparado às máximas de outros anos, ou mesmo talvez às suas correspondentes segundas ou terceiras maiores enchentes. Essa constatação remete à construção das chamadas séries de duração parcial, na qual todas as enchentes, que sejam independentes entre si e superiores a um determinado valor limiar, são ali incluídas, de modo não eqüidistante no tempo. Dessa maneira, de volta à Figura 1.1, se o valor limiar fosse fixado em 290m^3/s, as máximas descargas médias diárias dos anos de 1971, 1976 e 1989, respectivamente 246, 276 e 288 m^3/s, não estariam incluídas na série de duração parcial. Por outro lado, poderiam estar incluídas as segundas, as terceiras ou até as quartas maiores enchentes de outros anos, fazendo com que a série de duração parcial, assim formada, pudesse ser constituída, por exemplo, das 73 ou 82 maiores descargas médias diárias distribuídas de forma não eqüidistante ao longo dos 57 anos de registros. Para a seleção dos valores constituintes de uma série de duração parcial, há que se observar que entre dois de seus pontos consecutivos haja um período suficientemente longo de recessão de seus respectivos hidrogramas, de modo que as descargas da série sejam independentes entre si.

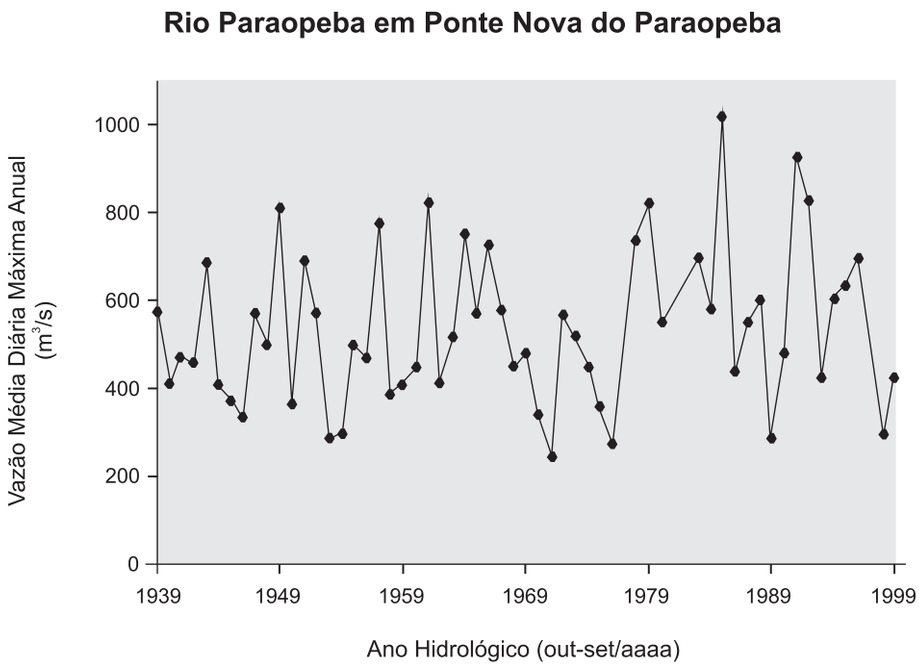


Figura 1.1 – A Série de Máximos Anuais do Rio Paraopeba em Ponte Nova do Paraopeba

As séries hidrológicas podem apresentar uma tendência, ou um 'salto', ou uma periodicidade ao longo do tempo, como resultado de variações naturais do clima ou alterações induzidas pela ação do homem. Nesse caso, as séries hidrológicas seriam ditas não estacionárias ao longo do tempo. Por exemplo, um reservatório de acumulação, de dimensões importantes, construído logo a montante de uma estação fluviométrica, faria com que a série hidrológica correspondente se apresentasse não estacionária e heterogênea no tempo, respectivamente, com descargas não regularizadas e regularizadas, antes e depois da implantação daquele reservatório a montante. Por outro lado, quando certas propriedades estatísticas de uma série hidrológica não se alteram ao longo do tempo, a série é dita estacionária. A série é considerada homogênea se o padrão de variabilidade, em torno de seu valor médio, é único e idêntico, ao longo do tempo. No exemplo do reservatório de acumulação, a série completa é certamente não estacionária e heterogênea, sendo composta por duas sub-séries, possivelmente estacionárias e homogêneas. Na maioria das aplicações da hidrologia estatística, as séries hidrológicas reduzidas devem ter como pré-requisito os atributos de estacionariedade e homogeneidade.

Finalmente, as séries hidrológicas devem ser representativas ou, em outras palavras, que seus valores constituintes sejam representativos da variabilidade presente no fenômeno hidrológico em questão. De volta ao exemplo da série de máximos anuais, ilustrada na Figura 1.1, a sub-série constituída somente pelos anos excepcionalmente secos de 1967 a 1976 não seria representativa porque contem uma seqüência de valores consistentemente mais baixos. Por outro lado, a sub-série constituída apenas pelos máximos anuais, ocorridos entre os anos considerados excepcionalmente molhados de 1978 e 1985, também não seria representativa da variabilidade das enchentes anuais do anuais do Rio Paraopeba em Ponte Nova do Paraopeba. Assim como para os atributos de estacionariedade e homogeneidade, na maioria das aplicações da hidrologia estatística, as séries hidrológicas reduzidas devem ser também representativas. Esses tópicos serão discutidos com maior rigor no capítulo 7, desta publicação.

1.4 – População e Amostra

O conjunto finito ou infinito de todos os possíveis resultados, ou possíveis realizações, de uma variável hidrológica recebe o nome de população. Na maioria das situações, o que se conhece é um sub-conjunto extraído da população, com um número limitado de observações, sub-conjunto ao qual dá-se o nome de amostra. Supondo tratar-se de uma amostra estacionária, homogênea, e representativa da população, nos mesmos sentidos enunciados para as séries

hidrológicas, pode-se dizer que o principal objetivo da hidrologia estatística é o de extrair conclusões válidas sobre o comportamento populacional da variável hidrológica em análise, somente a partir da informação contida na amostra. Como exemplo desse raciocínio, tome-se, por exemplo, a série de máximos anuais do Rio Paraopeba em Ponte Nova do Paraopeba cujos valores mínimo e máximo são 246 e 1017 m^3/s, respectivamente. Com base única e exclusivamente na amostra, poder-se-ia dizer que a probabilidade de se observar valores inferiores ou superiores aos limites amostrais é nula. De modo análogo, poder-se-ia dizer apenas que a enchente que deverá ocorrer no próximo ano, neste local, estaria provavelmente compreendida entre 246 e 1017 m^3/s.

O raciocínio subjacente à hidrologia estatística inicia-se com a proposta de um modelo matemático plausível para a distribuição de freqüências das realizações populacionais; nesse caso, trata-se de um raciocínio dedutivo, no qual faz-se uma tentativa de propor uma idéia geral válida para quaisquer casos particulares. Tal modelo matemático possui parâmetros, cujos verdadeiros valores populacionais devem ser estimados a partir dos valores amostrais. Uma vez estimados os seus parâmetros e, portanto, particularizado para um local ou situação, o modelo matemático pode ser agora usado para inferir sobre probabilidades de cenários não observados, tais como a probabilidade de ocorrer um valor superior a 1350 m^3/s em Ponte Nova do Paraopeba ou mesmo sobre a descarga média diária máxima local, cuja probabilidade de ser igualada ou superada é de apenas 0,01%; nesse caso, trata-se de um raciocínio indutivo, no qual particulariza-se a idéia geral. A Figura 1.2 ilustra as etapas do raciocínio inerente à hidrologia estatística.

Em geral, a amostra é constituída por elementos que são extraídos da população, um a um, de maneira aleatória e independente. Isso significa que em uma amostra, composta pelos elementos $\{x_1, x_2, \ldots, x_N\}$, cada um deles foi extraído da população ao acaso dentre um grande número de escolhas possíveis e equiprováveis. O elemento x_1, por exemplo, teve a mesma chance de ser sorteado da de qualquer outro constituinte da amostra e, inclusive, até mesmo de se repetir como elemento sorteado. Essa última possibilidade, ou seja a amostragem com reposição, implica em se ter independência entre os N elementos constituintes da amostra. Os atributos de eqüiprobabilidade e de independência definem uma amostra aleatória simples (AAS), a qual representa o plano de amostragem mais simples e mais eficaz para se fazer inferências sobre a população. Uma AAS homogênea e representativa é, em geral, o requisito inicial de qualquer aplicação da hidrologia estatística.

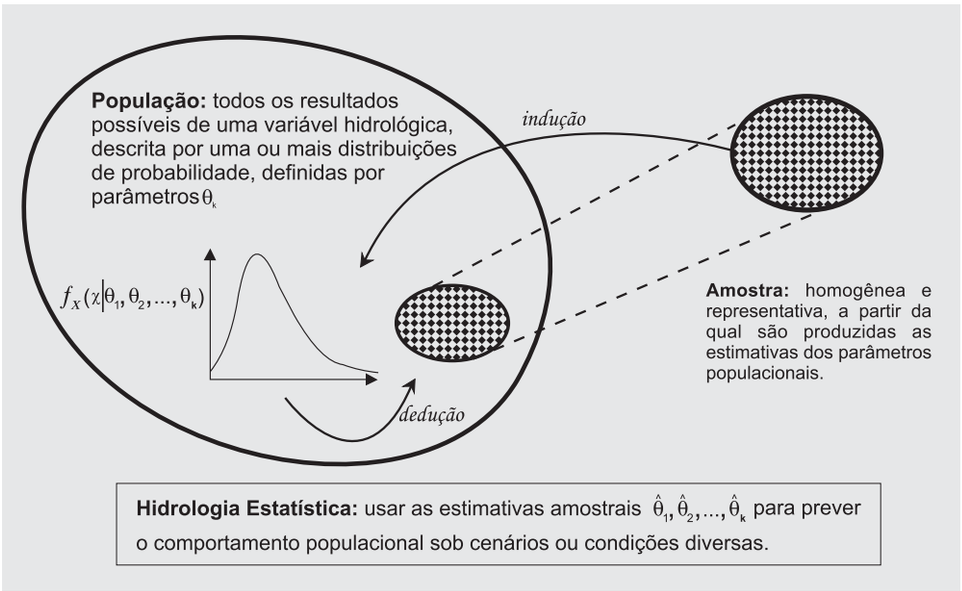


Figura 1.2 – Ilustração do Raciocínio Típico da Hidrologia Estatística

1.5 – Dados Hidrológicos

A quantificação das diversas fases do ciclo hidrológico, das suas respectivas variabilidades e de suas inter-relações, requer a coleta sistemática de dados básicos que se desenvolvem ao longo do tempo ou do espaço. As respostas aos diversos problemas de hidrologia aplicada serão tão mais corretas, quanto mais longos e precisos forem os registros de dados hidrológicos. Esses podem compreender dados climatológicos, pluviométricos, fluviométricos, evaporimétricos, sedimentométricos e de indicadores de qualidade da água, obtidos em instalações próprias, localizadas em pontos específicos de uma região, em intervalos de tempo pré-estabelecidos e com sistemática de coleta definida por padrões conhecidos. O conjunto dessas instalações, denominadas postos ou estações, constituem as redes hidrométricas e/ou hidrometeorológicas, cujas manutenção e densidade são essenciais para a qualidade dos estudos hidrológicos.

No Brasil, as principais entidades produtoras de dados hidrológicos e hidrometeorológicos são a Agência Nacional de Águas (ANA), cuja parte da rede é operada pela CPRM - Serviço Geológico do Brasil, e o Instituto Nacional de Meteorologia (INMET). Outras redes acessórias, de menor extensão, são mantidas por companhias energéticas ou por companhias de serviços de saneamento básico, entre outras. Grande parte dos dados hidrológicos brasileiros encontra-se disponível por meio do Sistema de Informações Hidrológicas da

Agência Nacional de Águas – Hidroweb, mediante acesso à URL http:// hidroweb.ana.gov.br. Alguns fenômenos hidrológicos e algumas de suas variáveis características mais comumente medidas encontram-se listadas na Tabela 1.1, juntamente com suas respectivas unidades.

Tabela 1.1 – Características e Variáveis Hidrológicas - Unidades

Fenômeno	Variável Característica	Unidade
Precipitação	Altura Intensidade Duração	mm, cm mm/h h, min
Evaporação/ Evapotranspiração	Intensidade Total	mm/dia, mm/mês mm, cm
Infiltração	Intensidade Altura	mm/h mm, cm
Escoamento total	Fluxo Volume Altura equivalente (Deflúvio)	l/s, m^3/s m^3, 10^6 m^3 , (m^3/s).mês mm ou cm sobre uma área
Escoamento subterrâneo	Fluxo Volume	l/min, l/h, m^3/dia m^3, 10^6 m^3

Os dados hidrológicos contêm erros aleatórios, sistemáticos e/ou grosseiros. Os primeiros são inerentes aos atos de medir e observar, trazendo consigo as imprecisões das leituras e medições ou, em outras palavras, as flutuações em torno de seus verdadeiros valores. Por exemplo, se em um único dia, forem realizadas 10 medições de descarga líquida em uma seção fluvial, em meio a uma estiagem prolongada com descarga quase constante, empregando o mesmo molinete e o mesmo hidrometrista, teríamos 10 resultados próximos e diferentes, os quais estariam flutuando em torno do verdadeiro valor da descarga líquida naquele local. Os erros sistemáticos, por sua vez, são aqueles que produzem um viés, para cima ou para baixo, nos resultados das observações e podem ter origem em mudanças na técnica de medição empregada, em calibrações incorretas de aparelhos de medição ou nos processos de coleta, transmissão e processamento dos dados. A mudança da posição de um pluviômetro, por exemplo, pode, em decorrência da ação do vento, provocar a ocorrência de erros sistemáticos nas observações das alturas de precipitação em um dado local. Do mesmo modo, a extrapolação errônea de uma curva-chave pode resultar em descargas exageradamente altas ou exageradamente baixas. Os erros grosseiros provêm de falhas humanas e resultam da falta de cuidado na execução de uma medição ou observação de uma variável hidrológica. Leituras linimétricas incorretas ou ilegíveis são exemplos de erros grosseiros.

Rigorosamente, os pressupostos da hidrologia estatística não admitem a existência dos erros mencionados. A hidrologia estatística lida com os 'erros' de amostragem ou com as flutuações amostrais de um fenômeno natural que possui uma variabilidade temporal e/ou espacial. Cinco diferentes amostras de uma certa

variável hidrológica, cada uma com o mesmo número de elementos, irão produzir 5 estimativas diferentes de determinadas propriedades estatísticas populacionais. As diferenças entre tais estimativas são os erros de amostragem em torno de seus respectivos e verdadeiros valores populacionais. Esses valores populacionais serão conhecidos somente se toda a população for amostrada. A essência da hidrologia estatística é extrair conclusões válidas a respeito do comportamento populacional, tendo-se em conta a incerteza devida à presença e à magnitude dos erros de amostragem. Nesse sentido, é evidente que quanto maior for a quantidade dos dados hidrológicos disponíveis e quanto mais isentos estiverem de erros de observação e medição, tanto melhores serão as inferências relativas ao comportamento populacional.

Exercícios

1. Enumere as principais razões que tornam estocásticos os fenômenos das precipitações e vazões de uma bacia hidrográfica.

2. Enumere exemplos de processos hidrológicos aproximadamente determinísticos.

3. Dê 3 exemplos de possíveis variáveis discretas e 3 de possíveis variáveis contínuas, associadas ao fenômeno da precipitação.

4. Os anexos 1, 2 e 3 desse boletim técnico referem-se respectivamente às vazões médias mensais, às vazões médias diárias máximas anuais e às alturas diárias de precipitação máximas anuais das estações fluviométrica 40800001 e pluviométrica 01944004, ambas com a mesma denominação de Ponte Nova do Paraopeba, em Minas Gerais. Faça um gráfico de dispersão entre as vazões médias diárias e as alturas de precipitação máximas anuais. Enumere as principais causas da impossibilidade de se estabelecer uma relação funcional do tipo causa e efeito entre tais variáveis. Explique como tais causas prováveis se manifestam no gráfico de dispersão.

5. Com relação às amostras de vazões médias mensais do anexo 1, discuta os atributos de aleatoriedade e independência, necessários ao conceito de amostra aleatória simples.

6. Enumere 3 exemplos de erros grosseiros eventualmente presentes em dados fluviométricos.

7. Enumere 3 exemplos de erros sistemáticos eventualmente presentes em dados pluviométricos.

8. Enumere 3 causas possíveis de presença de heterogeneidade em dados pluviométricos e fluviométricos. O que fazer diante de séries heterogêneas?

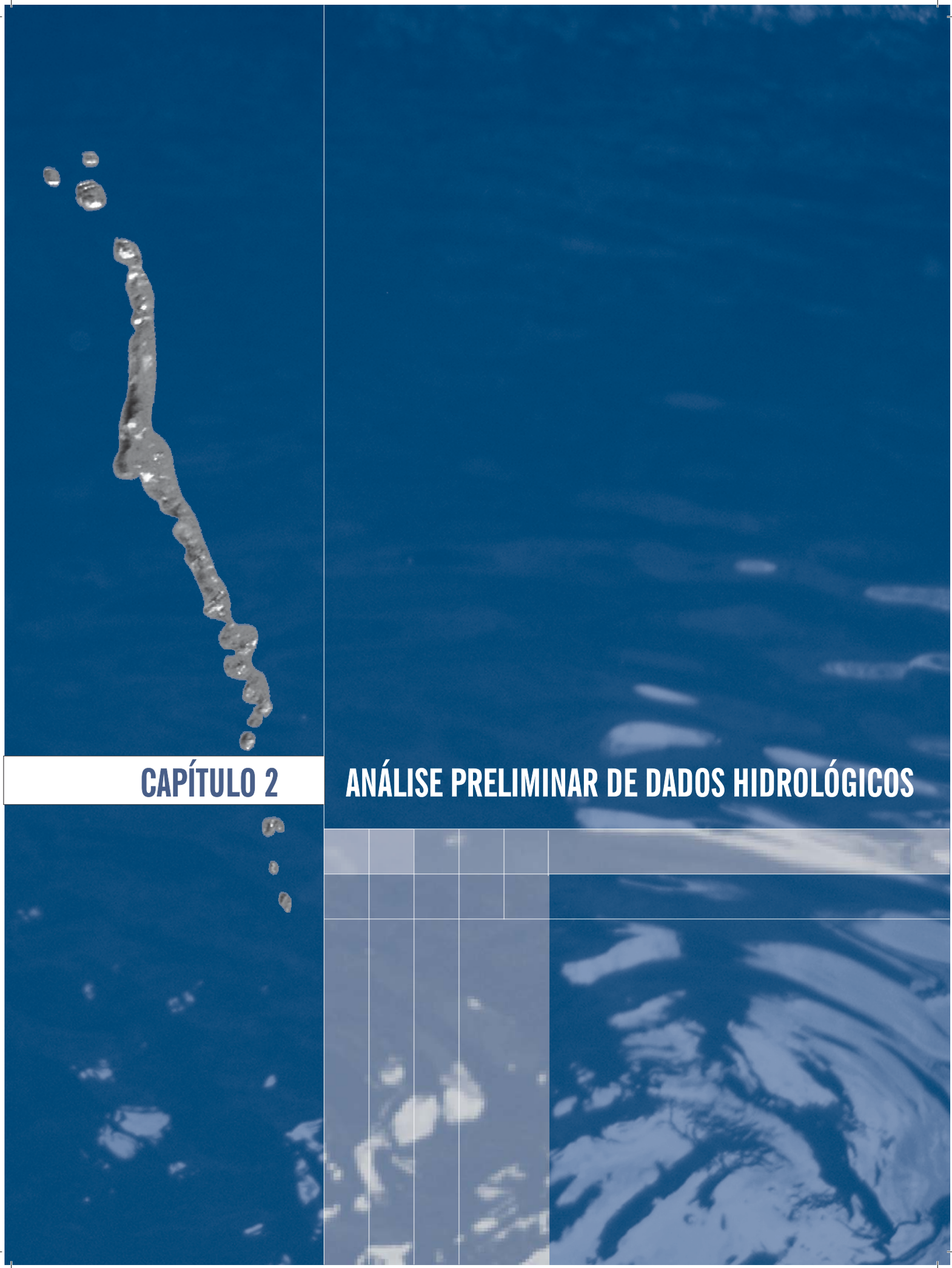
9. Visite o site do Sistema de Informações Hidrológicas – Hidroweb, clique em Dados Hidrológicos' e, em seguida, em 'Séries Históricas'. Faça o download da série hidrológica completa da estação fluviométrica de código 40800001, do Rio Paraopeba em Ponte Nova do Paraopeba, disponível desde 1938. Verifique o número N de anos completos, sem falhas no período chuvoso da região centro-sul de Minas Gerais que vai de Outubro a Março. Construa a série de duração parcial das N maiores vazões médias diárias máximas e compare-a com a série de máximos anuais, também de tamanho N. Em sua opinião, qual delas pode ser considerada mais representativa das cheias do Rio Paraopeba em Ponte Nova do Paraopeba?

10. Tome a série de máximos anuais, obtida no exercício 7, e subdivida-a em 5 sub-series não sobrejacentes de igual tamanho. Calcule e compare as médias aritméticas para cada uma das sub-séries e para a série total. Por que todas são estimativas da cheia média anual? Qual das estimativas é mais confiável? Por que?

11. Com os resultados do exercício 9, discuta a questão da representatividade da amostra.

12. Os rios Tocantins e Araguaia têm sua confluência a montante do reservatório de Tucuruí. Com base nas observações das estações pluvio-fluviométricas de ambas sub-bacias, é possível considerar as afluências a Tucuruí como uma variável hidrológica multivariada? Discuta as dificuldades de se conceber e empregar tal variável na previsão das afluências a Tucuruí.

13. Haan (1977) afirma que um problema hidrológico raramente preenche todos os requisitos necessários à aplicação de um certo método ou técnica estatística. Na seqüência, esse autor aponta duas alternativas. A primeira é a de redefinir a questão de forma que ela preencha os requisitos da teoria estatística e produza uma resposta "exata" para o problema artificial. A segunda é a de alterar a técnica estatística, quando possível, e aplicá-la ao problema real, tendo-se em conta que os resultados serão respostas aproximadas para a questão em foco e que o grau de aproximação irá depender fortemente da severidade com que as premissas da teoria estatística foram violadas. Qual das duas alternativas lhe parece mais adequada? Por que?



CAPÍTULO 2

ANÁLISE PRELIMINAR DE DADOS HIDROLÓGICOS

CAPÍTULO 2 ANÁLISE PRELIMINAR DE DADOS HIDROLÓGICOS

Conforme exposto no capítulo 1, os fenômenos hidrológicos apresentam uma aleatoriedade intrínseca devida à complexa interação e dependência entre inúmeros fatores influentes nas diversas fases do ciclo hidrológico. Para lidar com tais incertezas, o hidrólogo tem como uma de suas primeiras tarefas, obter e analisar uma amostra de dados hidrológicos. A investigação organizada de um conjunto de dados hidrológicos, na busca de evidências e padrões empíricos de variabilidade, é uma aplicação da estatística em um estágio *descritivo* e constitui o objeto do presente capítulo. O estágio seguinte, o qual procura estabelecer o padrão de variabilidade da população de onde foi extraída aquela amostra, é uma aplicação da teoria de probabilidades e dos métodos de inferência estatística, cujos fundamentos serão tratados nos capítulos subseqüentes desta publicação. A análise preliminar de uma amostra de dados hidrológicos compreende um conjunto de métodos e técnicas que visam extrair as características empíricas essenciais do padrão de distribuição de uma variável hidrológica. Esse conjunto pode ser dividido em três grupos: (a) Apresentação Gráfica de Dados Hidrológicos; (b) Sumário Numérico e Estatísticas Descritivas e (c) Métodos Exploratórios. Complementarmente à primeira análise de uma amostra de dados de uma única variável, apresenta-se, ao final desse capítulo, uma breve discussão sobre a associação entre observações simultâneas de duas variáveis.

2.1 – Apresentação Gráfica de Dados Hidrológicos

Em geral, um conjunto de observações de uma variável hidrológica encontra-se disponível em forma tabular (ver, por exemplo, o anexo 1 ou o exercício 9 do capítulo 1), a qual, muitas vezes, não consegue demonstrar, com facilidade e nitidez, a essência do padrão de distribuição da variável em questão. Essa nitidez é mais facilmente conseguida com o emprego de representações gráficas da variável hidrológica. O que se segue é um apanhado não exaustivo de diferentes tipos de gráficos de variáveis hidrológicas discretas e contínuas.

2.1.1 – Diagrama de Linha

O número de ocorrências de uma variável hidrológica *discreta* pode ser convenientemente representado pelo chamado *diagrama de linha*, o qual dispõe os valores possíveis da variável em um eixo horizontal, enquanto os correspondentes números de ocorrências são representados pelas alturas das linhas verticais. A Figura 2.1 exemplifica um diagrama de linha, onde, em abscissas, encontram-se os valores possíveis do número anual de cheias do Rio Magra na estação fluviométrica de Calamazza (Itália) que ultrapassaram a vazão de referência de 300 m^3/s em um período de 34 anos de observação, enquanto as alturas das linhas verticais representam os correspondentes números de ocorrências. A vazão de referência foi estabelecida como aquela, acima da qual os elevados níveis d'água ameaçam vidas e propriedades locais. A observação do diagrama da Figura 2.1 sugere uma distribuição aproximadamente simétrica do número de ocorrências, com valor central em torno de 4 cheias anuais.

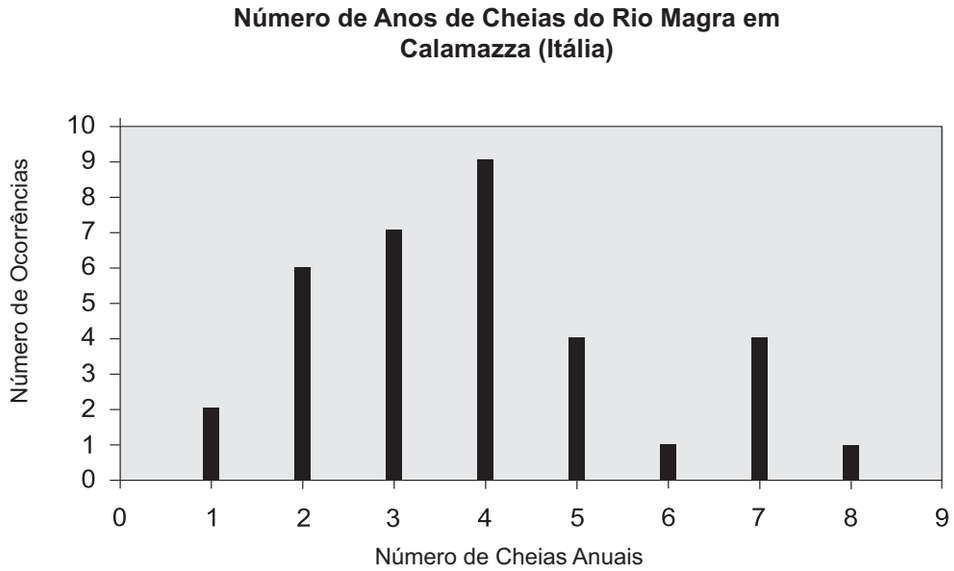


Figura 2.1 – Exemplo de Diagrama de Linha para o número de anos de cheias do Rio Magra em Calamazza, Itália, (adaptado de Kottegoda e Rosso, 1997)

2.1.2 – Diagrama Uniaxial de Pontos

O *diagrama uniaxial de pontos* é uma representação gráfica apropriada para amostras pequenas, de tamanho arbitrado como menor ou igual a 25 ou 30

observações, de variáveis *contínuas*. Os dados são inicialmente classificados em *ordem crescente* e, em seguida, grafados como pontos em um único eixo horizontal. A Tabela 2.1 apresenta as vazões médias anuais do Rio Paraopeba em Ponte Nova do Paraopeba, inicialmente na ordem cronológica de suas ocorrências entre os anos civis de 1938 a 1963, e, em seguida, classificadas em ordem crescente. Esses dados hidrológicos foram empregados para construir o diagrama uniaxial de pontos, ilustrado na Figura 2.2, no qual é possível visualizar a distribuição ligeiramente assimétrica dos elementos da amostra em torno do valor central, próximo a 86 m^3/s, assim como a ocorrência de anos relativamente mais chuvosos como o de 1943.

Tabela 2.1 – Vazões Médias Anuais do Rio Paraopeba em Ponte Nova do Paraopeba (m^3/s)

Ano Civil	Vazões médias anuais	Vazões classificadas	Número de ordem
1938	104,3	43,6	1
1939	97,9	49,4	2
1940	89,2	50,1	3
1941	92,7	57	4
1942	98	59,9	5
1943	141,7	60,6	6
1944	81,1	68,2	7
1945	97,3	68,7	8
1946	72	72	9
1947	93,9	80,2	10
1948	83,8	81,1	11
1949	122,8	83,2	12
1950	87,6	83,8	13
1951	101	87,6	14
1952	97,8	89,2	15
1953	59,9	92,7	16
1954	49,4	93,9	17
1955	57	97,3	18
1956	68,2	97,8	19
1957	83,2	97,9	20
1958	60,6	98	21
1959	50,1	101	22
1960	68,7	104,3	23
1961	117,1	117,1	24
1962	80,2	122,8	25
1963	43,6	141,7	26

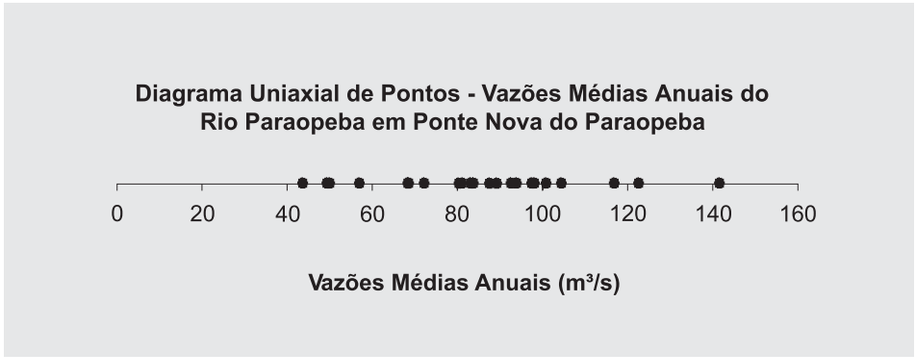


Figura 2.2 – Exemplo de Diagrama Uniaxial de Pontos para as vazões médias anuais do Rio Paraopeba em Ponte Nova do Paraopeba – Período 1938-1963

2.1.3 – Histograma

O *tamanho da amostra* é dado pelo número de elementos (ou itens ou observações) que a compõem e pode ser arbitrariamente considerado como pequeno, médio ou grande, a depender das características da variável em foco e, principalmente, se a série hidrológica disponível é do tipo completa ou do tipo reduzida. A série, apresentada na Tabela 2.2, de 62 anos de vazões médias anuais do Rio Paraopeba em Ponte Nova do Paraopeba, para o período 1938 a 1999, pode ser considerada de tamanho médio. Entretanto, uma amostra de 62 itens seria de tamanho pequeno se ela se referisse a vazões médias diárias. As séries hidrológicas reduzidas podem ser arbitrariamente categorizadas em amostras de tamanho pequeno se o número de elementos (N) for menor ou igual a 25, e de tamanho grande, se $N \geq 70$. Para as amostras médias e grandes, é conveniente classificá-las ou agrupá-las em subconjuntos, de modo a se ter uma melhor compreensão do padrão de variabilidade da variável em questão. Esse expediente dá origem a diversos tipos de gráficos, entre os quais destaca-se o *histograma*.

Para se construir um histograma, é necessário, primeiramente, agrupar as observações em *classes*, definidas por intervalos de *largura fixa ou variável*, e, em seguida, contar o número de ocorrências, ou seja, a *freqüência absoluta* em cada classe. O número de classes a ser considerado, representado por *NC*, depende do tamanho da amostra; de fato, um valor excessivamente pequeno para *NC* não irá permitir a visualização de características importantes da amostra, enquanto um valor excessivamente grande irá produzir flutuações exageradas das freqüências das classes. Kottegoda e Rosso (1977) sugerem que *NC* pode ser

Tabela 2.2 – Vazões Médias Anuais do Rio Paraopeba em Ponte Nova do Paraopeba (m^3/s)

Ano Civil	Vazões médias anuais	Ano Civil	Vazões médias anuais
1938	104,3	1969	62,6
1939	97,9	1970	61,2
1940	89,2	1971	46,8
1941	92,7	1972	79
1942	98	1973	96,3
1943	141,7	1974	77,6
1944	81,1	1975	69,3
1945	97,3	1976	67,2
1946	72	1977	72,4
1947	93,9	1978	78
1948	83,8	1979	141,8
1949	122,8	1980	100,7
1950	87,6	1981	87,4
1951	101	1982	100,2
1952	97,8	1983	166,9
1953	59,9	1984	74,8
1954	49,4	1985	133,4
1955	57	1986	85,1
1956	68,2	1987	78,9
1957	83,2	1988	76,4
1958	60,6	1989	64,2
1959	50,1	1990	53,1
1960	68,7	1991	112,2
1961	117,1	1992	110,8
1962	80,2	1993	82,2
1963	43,6	1994	88,1
1964	66,8	1995	80,9
1965	118,4	1996	89,8
1966	110,4	1997	114,9
1967	99,1	1998	63,6
1968	71,6	1999	57,3

aproximado pelo inteiro mais próximo de $\sqrt{N}$, com um mínimo de 5 e um máximo de 25, argumentando, assim, que não são informativos os histogramas de amostras de tamanho inferior a 25. Uma indicação alternativa é a regra de Sturges (1926) que sugere a seguinte aproximação para o número de classes:

$$NC = 1 + 3{,}3 \log_{10} N \tag{2.1}$$

Para ilustrar a elaboração da *tabela de freqüências*, essencial para a construção do histograma, tomemos a amostra de vazões médias anuais da Tabela 2.2, cujo tamanho é $N = 62$. De acordo com as recomendações mencionadas, o número de classes deve estar compreendido entre 7 e 8; tomemos $NC = 7$, lembrando que o limite inferior da primeira classe deve ser menor ou igual ao mínimo amostral (43,6 m³/s), enquanto o limite superior da sétima classe deve ser maior ou igual ao máximo amostral (166,9 m³/s). Uma vez que a *amplitude A* entre os valores máximo e mínimo da amostra é de 123,3 e que $NC = 7$, pode-se arbitrar a *largura de intervalo de classe* como fixa e igual a $LIC = 20$ m³/s, em decorrência de ser um inteiro próximo a 17,61, resultado do quociente entre a amplitude e o número de classes. A Tabela 2.3 apresenta um resumo do cálculo (a) das freqüências absolutas, obtidas pelo número de ocorrências em cada classe, (b) das freqüências relativas, resultantes da divisão das freqüências absolutas por $N = 62$ e (c) das freqüências relativas acumuladas.

Tabela 2.3 – Tabela de freqüências das vazões médias anuais do Rio Paraopeba em Ponte Nova do Paraopeba – Período 1938 a 1999

Classe j	Intervalo de Classe (m³/s)	Freqüência Absoluta f_j	Freqüência Relativa fr_j	Freqüência Acumulada $F = \sum_j fr_j$
1	(30,50]	3	0,0484	0,0484
2	(50,70]	15	0,2419	0,2903
3	(70,90]	21	0,3387	0,6290
4	(90,110]	12	0,1935	0,8226
5	(110,130]	7	0,1129	0,9355
6	(130,150]	3	0,0484	0,9839
7	(150,170]	1	0,0161	1
Total		62	1	

Com base nos elementos da Tabela 2.3, pode-se construir o histograma, da Figura 2.3, o qual é um simples gráfico de barras tendo, em abscissas, os intervalos de classes e, em ordenadas, as freqüências absolutas e/ou relativas. A observação do histograma da Figura 2.3 mostra algumas características salientes da amostra, tais como: (a) a maior concentração de pontos no terceiro intervalo de classe, o qual provavelmente contem o valor central em torno do qual os pontos restantes se dispersam; (b) uma certa assimetria da distribuição de freqüências, demonstrada pela maior amplitude à direita do bloco de maior freqüência, quando comparada com a amplitude à esquerda e (c) a ocorrência isolada de observações muito superiores ao valor central. É importante ressaltar, entretanto, que a forma do histograma é muito sensível ao número, à largura e aos limites dos intervalos de classe. De volta ao exemplo, note que os dois últimos intervalos de classe contêm respectivamente 3 e 1 pontos amostrais, os quais certamente podem ser

concentrados em uma única classe de largura 40 m^3/s, com limite inferior igual a 130 m^3/s e superior igual a 170 m^3/s.

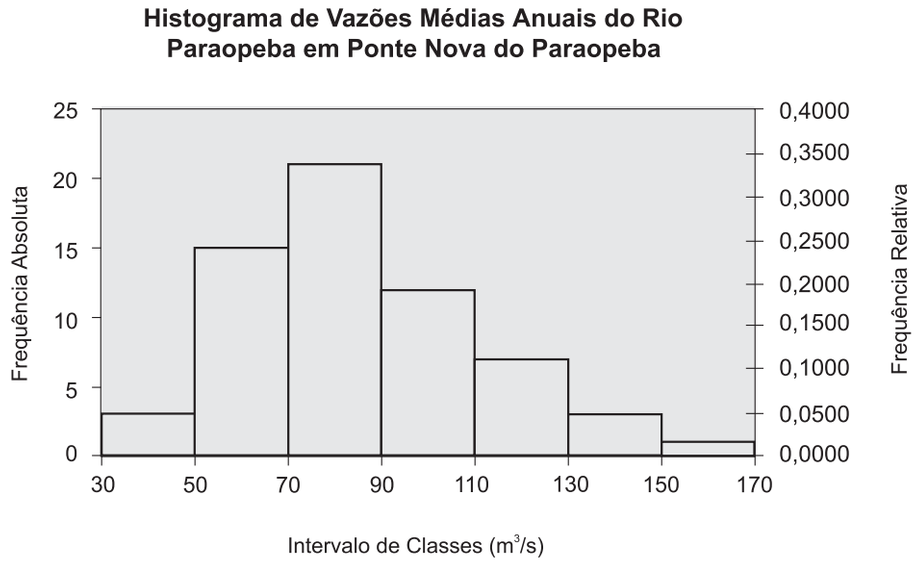


Figura 2.3 – Histograma das vazões médias anuais do Rio Paraopeba em Ponte Nova do Paraopeba – Período 1938 a 1999

2.1.4 – Polígono de Freqüências

O *polígono de freqüências* é outra representação gráfica da tabela de freqüências, sendo muito útil para diagnosticar o padrão de distribuição de uma variável. Esse polígono é aquele formado pela junção dos pontos médios dos topos dos retângulos do histograma, depois de estendê-lo por uma classe adicional de cada um de seus lados. O polígono de freqüências correspondente ao histograma da Figura 2.3 encontra-se ilustrado na Figura 2.4. Observe que, como o polígono de freqüências deve ter ordenadas inicial e final nulas e, por convenção, área igual à do histograma, ele deve começar meio intervalo de classe à esquerda e finalizar meio intervalo à direita. Em conseqüência, o polígono de freqüências da Figura 2.4 inicia com a abscissa 20 m^3/s e termina com 180 m^3/s, ambos com freqüências relativas iguais a zero. O valor que corresponde à maior ordenada do polígono recebe a denominação de *moda*; no caso da Figura 2.4, a moda, ou o valor mais freqüente, é de 80 m^3/s.

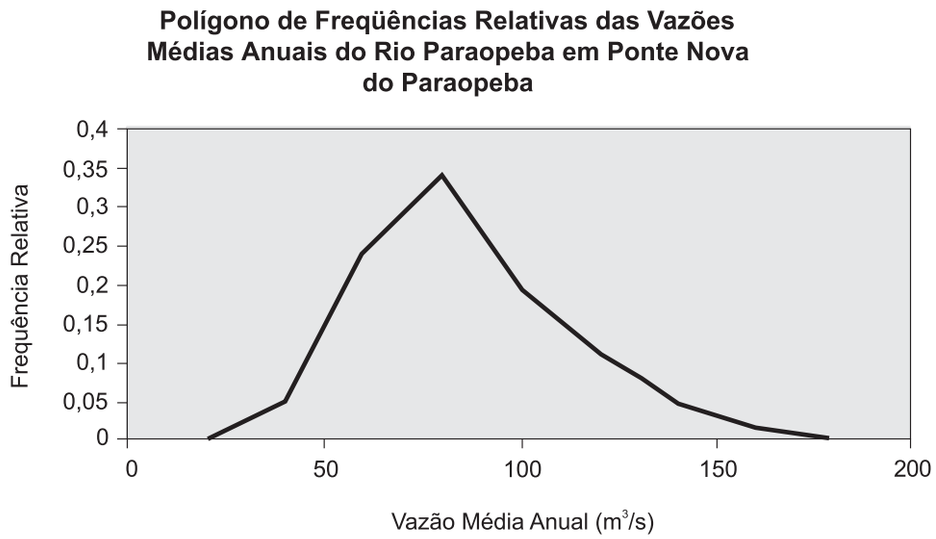


Figura 2.4 – Polígono de Freqüências Relativas das vazões médias anuais do Rio Paraopeba em Ponte Nova do Paraopeba – Período 1938 a 1999

É mais usual construir-se o *polígono de freqüências relativas*, ao invés de se empregar as freqüências absolutas; neste caso, as ordenadas de cada classe são as respectivas freqüências de ocorrência, limitadas entre os valores extremos de 0 e 1. À medida que o número de observações cresce e, em conseqüência, a largura dos intervalos de classe decresce, o polígono de freqüências relativas torna-se uma curva de freqüência. No caso limite de uma amostra de tamanho infinito, esta curva tornar-se-ia a *função densidade de probabilidade* da população, cuja definição formal será um dos objetos do capítulo 3.

2.1.5 – Diagrama de Freqüências Relativas Acumuladas

O *diagrama de frequências relativas acumuladas* resulta da união, por linhas contínuas, dos pares formados pelos limites superiores dos intervalos de classe e pelas ordenadas consecutivamente acumuladas do histograma, desde a menor até a maior. No eixo das ordenadas, o diagrama fornece a freqüência de não superação do valor correspondente da variável, lido no eixo das abscissas. De modo alternativo, o diagrama de freqüências relativas acumuladas pode também ser elaborado sem a prévia construção do histograma ou da tabela de freqüências. Para isso, basta (a) classificar os dados em ordem crescente; (b) associar aos dados classificados os seus respectivos números de ordem da classificação m, com $1 \leq m \leq N;$ e (c) associar aos dados classificados as correspondentes

freqüências ou probabilidades empíricas de não superação, calculadas pelo quociente m/N. Esse modo alternativo foi aqui usado para construir o diagrama de freqüências relativas acumuladas das vazões médias anuais do Rio Paraopeba em Ponte Nova do Paraopeba, ilustrado na Figura 2.5.

O diagrama de freqüências acumuladas permite a identificação imediata da *mediana* Q_2, qual seja do valor correspondente à freqüência de não superação de 0,5, assim como do *primeiro quartil* Q_1 e do *terceiro quartil* Q_3, que correspondem respectivamente às freqüências de 0,25 e 0,75; no diagrama da Figura 2.5, Q_2= 82,7, Q_1= 68,2 e Q_3= 99,1 m^3/s. A *amplitude inter-quartis*, representada por AIQ, é dada pela diferença entre Q_3 e Q_1 e tem sido usada como parte de um critério para a identificação de *pontos atípicos* (ou '*outliers*') eventualmente presentes na amostra. Segundo tal critério, é considerado um ponto atípico superior todo elemento da amostra superior a (Q_3+1,5AIQ) e, analogamente, um ponto atípico inferior é todo e qualquer elemento menor do que (Q_1-1,5AIQ). Como o próprio nome indica, um ponto atípico afasta-se de modo singular e dramático da tendência geral de variação dos outros elementos da amostra, podendo ser resultado de observações com erros grosseiros ou simplesmente a manifestação de eventos muito raros. Comprovado o primeiro caso, a sua remoção da amostra estaria plenamente justificada; no segundo caso, entretanto, sua remoção seria uma decisão incorreta ou, pelo menos, controvertida. De volta ao exemplo da Figura 2.5, e segundo o critério exposto, a vazão média anual de 166,9 m^3/s, correspondente ao ano civil de 1983, é considerada um ponto amostral atípico.

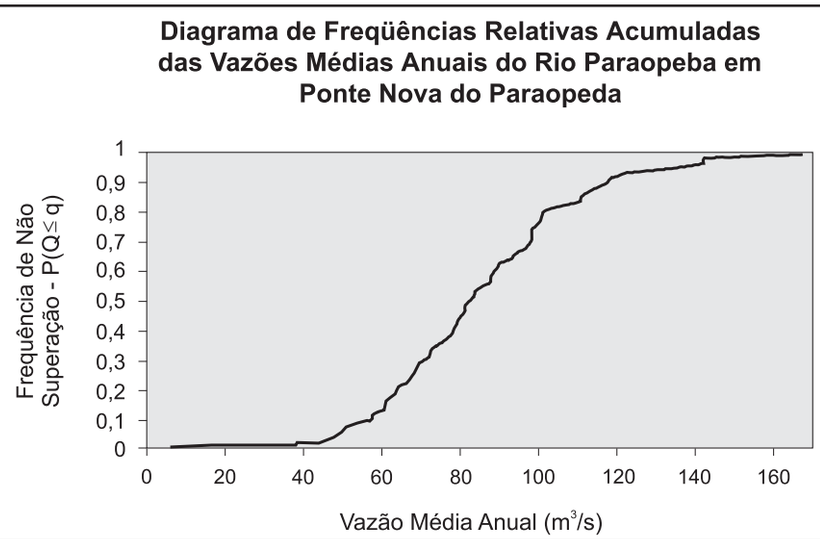


Figura 2.5 – Diagrama de Freqüências Relativas Acumuladas das vazões médias anuais do Rio Paraopeba em Ponte Nova do Paraopeba – Período 1938 a 1999

Do modo análogo aos quartis, pode-se fazer referência aos *decis*, para freqüências acumuladas múltiplas de 0,1, aos *percentis* para freqüências múltiplas de 0,01 e, mais genericamente, aos *quantis*. Convém ressaltar que se houver a inversão dos eixos horizontal e vertical de um diagrama de freqüências acumuladas, resulta o assim denominado *gráfico de quantis*. Novamente, à medida que o número de observações cresce, o diagrama de freqüências relativas acumuladas vai se tornando uma curva de distribuição de freqüências. No caso limite de uma amostra de tamanho infinito, esta curva tornar-se-ia a *função de distribuição de probabilidades acumuladas* da população.

2.1.6 – Curva de Permanência

A chamada *curva de permanência* é uma variação do diagrama de freqüências relativas acumuladas, na qual a freqüência de não superação é substituída pela porcentagem de um intervalo de tempo específico em que o valor da variável, indicado em abscissas, foi igualado ou superado. Em hidrologia, a curva de permanência é muito usada para ilustrar o padrão de variação de vazões, assim como o é para indicadores de qualidade da água, tais como turbidez de um trecho fluvial, dureza da água e concentrações de sedimento em suspensão, entre outros. Em particular, é freqüente o emprego da curva de permanência de vazões para o planejamento e projeto de sistemas de recursos hídricos e, também, como instrumento de outorga de direito de uso da água em alguns estados brasileiros. Por exemplo, a Superintendência de Recursos Hídricos do Estado da Bahia pode outorgar, para um novo usuário dos recursos hídricos de domínio daquele estado, até 80% da vazão denotada por Q_{90}, ou seja, a vazão local que é igualada ou superada em 90% do tempo.

Genericamente, a curva de permanência de vazões de uma dada seção fluvial, para a qual se dispõe de N dias de registros fluviométricos, pode ser construída do seguinte modo: (a) ordene as vazões Q em *ordem decrescente*; (b) atribua a cada vazão ordenada Q_m a sua respectiva ordem de classificação m; (c) associe a cada vazão ordenada Q_m a sua respectiva freqüência ou probabilidade empírica de ser igualada ou superada $P(Q \geq Q_m)$, a qual pode ser estimada pela razão (m/N) e (d) lance em um gráfico as vazões ordenadas e suas respectivas porcentagens $100(m/N)$ de serem igualadas ou superadas no intervalo de tempo considerado. Para exemplificar a construção da curva de permanência, tomemos as vazões médias diárias observadas no Rio Paraopeba em Ponte Nova do

Paraopeba, durante o ano hidrológico de Outubro de 1962 a Setembro de 1963; o fluviograma anual correspondente está ilustrado na Figura 2.6. Efetuando as etapas necessárias e com $N = 365$ dias, a curva de permanência correspondente é aquela ilustrada na Figura 2.7.

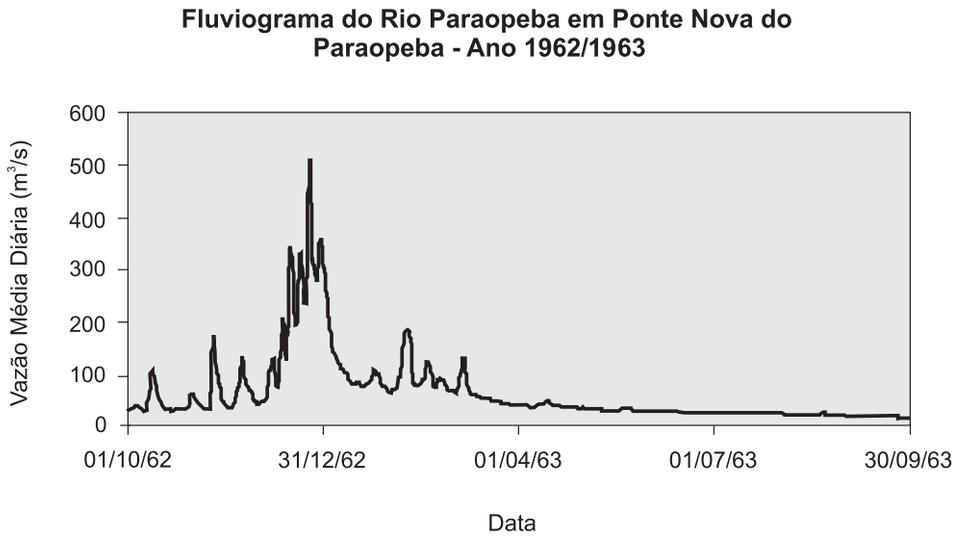


Figura 2.6 – Fluviograma do Rio Paraopeba em Ponte Nova do Paraopeba 1962/1963

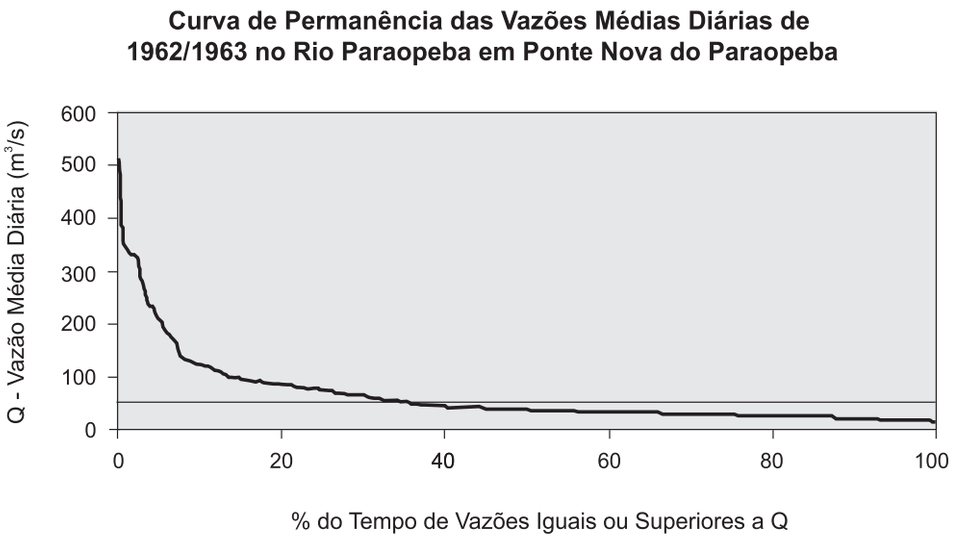


Figura 2.7 – Curva de Permanência das Vazões do Rio Paraopeba em Ponte Nova do Paraopeba

A curva de permanência da Figura 2.7 revela, por exemplo, que a vazão Q_{90}, ou seja a vazão que é excedida em 328,5 dias do ano, é de 23,4 m³/s. Além de seu

uso para cálculo da vazão referencial de outorga, a curva de permanência possui outras utilizações de interesse prático. Uma delas é a estimativa preliminar do volume sazonal de um possível reservatório destinado a manter, por exemplo, um calado mínimo para navegação, ou uma certa vazão mínima Q_r superior à mínima anual, a jusante da seção fluvial em questão. No exemplo da Figura 2.7, supondo que $Q_r = 50\ m^3/s$, tal como indicado pela linha horizontal, o volume a ser acumulado durante o período chuvoso poderia ser estimado pela diferença entre a área compreendida entre a linha horizontal e o eixo das abscissas, e a área abaixo da curva de permanência, ambas calculadas a partir do ponto da interseção das linhas correspondentes. Evidentemente, o volume afluente durante o período chuvoso, o qual pode ser obtido pela área da curva de permanência acima da linha horizontal, deve ser suficiente para suprir o déficit dos meses de estiagem.

2.2 – Sumário Numérico e Estatísticas Descritivas

As características essenciais de forma do histograma ou do polígono de freqüências relativas podem ser sumariadas por meio de *estatísticas descritivas* de uma amostra de dados hidrológicos, as quais são medidas-resumo que sintetizam, de modo simples e econômico, o padrão de distribuição da variável em questão. Além disso, as estatísticas descritivas apresentam uma importante vantagem, em relação à apresentação gráfica de dados, que é a representada pelo seu uso na estatística inferencial, ou seja, o de extrair da amostra as informações necessárias para inferir o comportamento populacional. As estatísticas descritivas podem ser agrupadas em 3 tipos distintos: (a) *medidas de tendência central*; (b) *medidas de dispersão* e (c) *medidas de assimetria e de curtose.*

2.2.1 – Medidas de Tendência Central

Os dados hidrológicos, em geral, se aglomeram em torno de um *valor central*, tal como no diagrama uniaxial da Figura 2.2. O valor central representativo de uma amostra pode ser calculado por uma das *medidas de tendência central ou de posição*, entre as quais, as mais conhecidas são a *média*, a *moda* e a *mediana*. A escolha entre tais medidas depende do uso pretendido do valor central.

Média

Se uma amostra de tamanho N é constituída pelos elementos $\{x_1, x_2, ..., x_N\}$, a *média aritmética*, ou simplesmente *média*, de X é dada por

$$\bar{x} = \frac{x_1 + x_2 + ... + x_N}{N} = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{2.2}$$

Se, das N observações da variável X, N_1 forem iguais a x_1, N_2 forem iguais a x_2 e assim por diante até o k-ésimo valor amostral, então a média de X pode ser obtida por

$$\bar{x} = \frac{N_1 x_1 + N_2 x_2 + ... + N_k x_k}{N} = \frac{1}{N}\sum_{i=1}^{k} N_i x_i \tag{2.3}$$

Analogamente, se f_i denotar a freqüência relativa da observação x_i, a equação 2.3 pode ser re-escrita como

$$\bar{x} = \sum_{i=1}^{k} f_i x_i \tag{2.4}$$

A média é a medida de posição mais freqüentemente usada e tem um significado teórico importante como estimativa da média populacional μ. Conforme mencionado no item 2.1.4, no caso limite de uma amostra de tamanho infinito de uma variável contínua X e, conseqüentemente, do polígono de freqüências tornar-se a função densidade de probabilidade, a média μ irá corresponder à coordenada, no eixo das abscissas, do *centróide* da área abaixo da curva de freqüências.

Alternativamente à media aritmética, porém dentro da mesma idéia por ela sugerida, existem duas outras medidas de tendência central que são úteis em alguns casos especiais. São elas: *a media harmônica*, representada por $\bar{x}_h$, e a *média geométrica* $\bar{x}_g$. A média harmônica é o recíproco da média aritmética dos recíprocos dos elementos da amostra. Formalmente, é definida por

$$\bar{x}_h = \frac{1}{(1/N)\left[(1/x_1) + (1/x_2) + ... + (1/x_N)\right]} \tag{2.5}$$

Tipicamente, a média harmônica apresenta uma noção mais apropriada de 'média' em situações que envolvem *proporções de variação*. Por exemplo, se a primeira

metade de um trecho fluvial é percorrida por um flutuador, a uma velocidade de 0,4 m/s, e a outra metade a 0,60 m/s, a média aritmética seria $\bar{x} = 0{,}50$ m/s e a média harmônica seria $\bar{x}_h = 0{,}48$ m/s, a qual é de fato a velocidade média do flutuador ao longo de todo o trecho fluvial. Por outro lado, a média geométrica é mais apropriada para estimar o valor central de variáveis que possuem um *desenvolvimento geométrico*, ou seja, aquelas cujos valores sucessivos guardam entre si um fator de crescimento ou decrescimento, tais como aumento populacional ou de carga orgânica das afluências a uma estação de tratamento de esgotos. A média geométrica, a qual é consistentemente menor ou igual à média aritmética , é dada pela raiz N-ésima do produto dos N valores amostrais, ou seja,

$$\bar{x}_g = \sqrt[N]{x_1 . x_2 . \ldots . x_N} = \prod_{i=1}^{N} (x_i)^{1/N} = \exp\left(\frac{1}{N} \sum_{i=1}^{N} \ln x_i \right) \tag{2.6}$$

sendo equivalente ao antilogaritmo da média aritmética dos logaritmos dos elementos x_i.

Mediana

A média aritmética de uma amostra, por levar em conta todos os seus elementos, apresenta a desvantagem de ter seu valor afetado pela eventual presença de pontos atípicos. Uma outra medida de posição mais *resistente* do que a média aritmética, por ser imune à eventual presença de valores extremos discordantes na amostra, é a *mediana* x_{md}. Essa é definida como o valor da variável X que separa a freqüência total em duas metades iguais, sendo, portanto, equivalente ao segundo quartil Q_2. Se as observações amostrais são ordenadas de modo que $\left\{ x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(N)} \right\}$, a mediana pode ser calculada por

$$x_{md} = x_{\left(\frac{N+1}{2}\right)} \text{ se } N \text{ for ímpar ou } x_{md} = \frac{x_{\left(\frac{N}{2}\right)} + x_{\left(\frac{N}{2}+1\right)}}{2} \text{ se } N \text{ for par} \tag{2.7}$$

Moda

A moda x_{mo} é o valor amostral que ocorre com maior freqüência, sendo geralmente obtido a partir do polígono de freqüências relativas, tal como o da Figura 2.3. No caso limite de uma amostra de tamanho infinito de uma variável contínua X e, conseqüentemente, do polígono de freqüências tornar-se a função densidade de probabilidade, a moda irá corresponder à coordenada, no eixo das abscissas, do

ponto de derivada nula da curva de freqüências, ressalvando que pode haver mais de um desses pontos em funções não unimodais. Em polígonos de freqüências ou histogramas assimétricos, quais sejam aqueles que apresentam amplitudes diferentes à direita e à esquerda da moda, as medidas de tendência central apresentam características peculiares. Quando a amplitude à direita da moda é muito maior do que à esquerda, trata-se de um histograma com assimetria positiva, caso em que $x_{mo} < x_{md} < \bar{x}$. Quando a amplitude à esquerda da moda é muito maior, a assimetria é dita negativa e $\bar{x} < x_{md} < x_{mo}$. Quando ambas amplitudes aproximadamente se equivalem, o histograma é simétrico e as três medidas de tendência central têm valores próximos entre si.

2.2.2 – Medidas de Dispersão

O grau de variabilidade dos pontos, em torno do valor central de uma amostra, é dado pelas medidas de dispersão. Entre essas, a mais simples e mais intuitiva é a *amplitude*, dada por $A = x_{(N)} - x_{(1)}$, onde $x_{(N)}$ e $x_{(1)}$ são, respectivamente, o *N*-ésimo e o primeiro dos elementos classificados em ordem crescente. A diferença entre o máximo e o mínimo da amostra, tal como expressa pela amplitude, depende exclusivamente de tais pontos. Esses, por sua vez, podem ser muito discordantes dos outros elementos da amostra e tornar a amplitude uma medida não representativa da dispersão ali contida. Uma outra medida mais imune à eventual presença de tais pontos e, portanto, mais resistente, é a amplitude inter-quartis *AIQ*, dada pela diferença entre o terceiro e o primeiro quartis, respectivamente Q_3 e Q_1.

As medidas de dispersão já mencionadas, embora fáceis de calcular, são pouco representativas porque ignoram os elementos restantes da amostra. Essa inconveniência pode ser superada pelo emprego de outras medidas de dispersão que têm como base o desvio médio de todos os pontos amostrais em relação a um valor central representativo. As principais são: o desvio médio absoluto e o desvio padrão.

Desvio Médio Absoluto

O *desvio médio absoluto*, aqui denotado por *d*, representa a média aritmética dos valores absolutos dos desvios amostrais, em relação à média. Para uma amostra $\{x_1, x_2, ..., x_N\}$, *d* é definido por

$$d = \frac{|x_1 - \bar{x}| + |x_2 - \bar{x}| + \ldots |x_N - \bar{x}|}{N} = \frac{1}{N}\sum_{i=1}^{N}|x_i - \bar{x}| \qquad (2.8)$$

Embora seja uma medida intuitiva, o desvio médio absoluto pondera de modo linearmente proporcional tanto os pequenos como os grandes desvios em relação à média. Além disso, o emprego do operador 'valor absoluto', na equação 2.8, torna o cálculo de *d* ligeiramente trabalhoso, do ponto de vista computacional.

Desvio Padrão

Uma prática alternativa ao uso do valor absoluto nas medidas de dispersão, é elevar ao quadrado os desvios em relação à média. Para uma amostra , define-se a *variância amostral* como o desvio quadrático médio, dado pela seguinte equação:

$$s^2 = \frac{(x_i - \bar{x})^2 + (x_2 - \bar{x})^2 + \ldots + (x_N - \bar{x})^2}{N} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2 \qquad (2.9)$$

Analogamente à média μ, a variância populacional, denotada por σ^2, pode ser estimada sem *viés* por meio da seguinte correção da equação 2.9:

$$s^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2 \qquad (2.10)$$

O termo 'viés' é aqui usado livremente para indicar que, *em média*, não existe diferença entre σ^2 e sua estimativa pela equação 2.10, diferentemente do resultado da equação 2.9. Diz-se, nesse caso, que houve a redução de 1 *grau de liberdade* [de *N* para (*N*-1)] pelo fato da média populacional μ haver sido estimada pela média amostral $\bar{x}$, previamente à estimativa de σ^2 por meio da equação 2.10. Os termos 'viés' e 'graus de liberdade' serão formalmente definidos no capítulo 6.

A variância é expressa em termos do quadrado das dimensões da variável original. Para conservar as unidades da variável, define-se o *desvio padrão s* como a raiz quadrada do desvio quadrático médio, ou seja, a raiz quadrada da variância s^2, tal como calculada pela equação 2.10. Formalmente, o desvio padrão é definido pela seguinte expressão:

$$s = \sqrt{\frac{(x_i - \bar{x})^2 + (x_2 - \bar{x})^2 + ... + (x_N - \bar{x})^2}{N-1}} = \sqrt{\frac{1}{(N-1)} \sum_{i=1}^{N} (x_i - \bar{x})^2} \qquad (2.11)$$

Diferentemente do desvio médio absoluto, o desvio padrão é fortemente influenciado pelos menores e maiores desvios, constituindo-se na medida de dispersão mais freqüentemente usada. A expansão do segundo membro da equação 2.11 facilita o cálculo do desvio padrão por meio da seguinte expressão equivalente:

$$s = \sqrt{\frac{1}{(N-1)} \left(\sum_{i=1}^{N} x_i^2 - 2\bar{x} \sum_{i=1}^{N} x_i + N\bar{x}^2 \right)} = \sqrt{\frac{1}{(N-1)} \sum_{i=1}^{N} x_i^2 - \frac{N}{(N-1)} \bar{x}^2} \qquad (2.12)$$

Quando se pretende comparar a variabilidade ou a dispersão de amostras de duas ou mais variáveis diferentes, é comum o emprego do chamado *coeficiente de variação CV*, resultado do quociente entre o desvio padrão *s* e a média $\bar{x}$. O coeficiente de variação é um número adimensional positivo, devendo ser aplicado somente nos casos em que as médias são diferentes de zero e as observações são sempre positivas; caso sejam sempre negativas, o respectivo *CV* deve ser calculado com base no valor absoluto da média.

2.2.3 – Medidas de Assimetria e Curtose

Outras caracterizações importantes da forma de um histograma ou do polígono de freqüências são dadas pelas medidas de assimetria e curtose, ambas baseadas em valores acumulados de potências superiores a 2 dos desvios dos pontos amostrais em relação à média. A principal medida de assimetria é denominada *coeficiente de assimetria*, enquanto a de curtose é dada pelo *coeficiente de curtose*.

Coeficiente de Assimetria

Para uma amostra $\{x_1, x_2, ... x_N\}$, define-se o *coeficiente de assimetria* pelo número adimensional dado por

$$g = \frac{N}{(N-1)(N-2)} \frac{\sum_{i=1}^{N} \left(x_i - \bar{x} \right)^3}{s^3} \qquad (2.13)$$

Na equação 2.13, à exceção do primeiro quociente do segundo membro, o qual contém as correções para fazer do coeficiente de assimetria amostral uma estimativa mais acurada da correspondente medida populacional γ, o coeficiente *g* reflete e acentua a contribuição acumulada dos desvios positivos e negativos, em relação à média amostral. De fato, desvios positivos muito grandes, ou negativos muito grandes, quando elevados à terceira potência, serão grandemente acentuados; a predominância, ou a equivalência, desses desvios, quando somados, irá determinar se o coeficiente de assimetria será positivo, negativo ou nulo. Se o coeficiente *g* é positivo, diz-se que o histograma (ou o polígono de freqüências) possui assimetria positiva, tal como ilustrado pelas Figuras 2.3 e 2.4. Nesse caso, observa-se que a moda amostral é inferior à mediana, a qual, por sua vez, é inferior à média; o contrário seria observado caso o coeficiente *g* determinasse um histograma com assimetria negativa. Caso os desvios positivos e negativos se equivalessem, o coeficiente *g* teria valor nulo (ou próximo de zero) e as 3 medidas de tendência central tenderiam a se concentrar em um único valor de *X*. O coeficiente de assimetria é um número limitado; de fato, a despeito de quão positivos ou negativos sejam os desvios em relação à média, é válida a inequação $|g| \leq \sqrt{N-2}$.

As séries hidrológicas referentes a eventos máximos, em geral, possuem coeficientes de assimetria positivos. Essa constatação é particularmente verdadeira para as séries de vazões máximas anuais. De fato, para tais séries, há uma grande concentração de valores não muito inferiores, ou não muito superiores, à cheia média anual, que, em geral, correspondem aos níveis d'água contidos pelo leito menor da seção fluvial. Entretanto, a rara combinação de condições hidrometeorológicas excepcionais e de elevado teor de umidade do solo pode determinar a ocorrência de uma grande enchente, com vazão máxima muitas vezes superior ao valor modal. Bastam apenas algumas ocorrências de tais grandes enchentes para determinar a forma assimétrica do polígono de freqüências das vazões máximas anuais e, conseqüentemente, valores positivos para o coeficiente *g*. Do exposto, é certo concluir que a prescrição de modelos matemáticos positivamente assimétricos para as funções densidade de probabilidade da população explica-se pelo mecanismo de formação das enchentes de um rio. Vale ressalvar, entretanto, que o coeficiente *g*, por não ser uma medida resistente e, conseqüentemente, ser muito sensível à presença de extremos em amostras de tamanho reduzido, não deve constituir um balizador único ou inequívoco para a prescrição de modelos distributivos positivamente assimétricos.

Coeficiente de Curtose

Uma medida de quão pontiagudo ou achatado é o histograma (ou o polígono de freqüências) em torno da média amostral, pode ser calculada pelo *coeficiente de curtose*. Esse número adimensional é formalmente definido por

$$k = \frac{N^2}{(N-1)(N-2)(N-3)} \frac{\sum_{i=1}^{N} (x_i - \bar{x})^4}{s^4} \tag{2.14}$$

Por tratar-se de um coeficiente cuja base de cálculo é a soma das quartas potências dos desvios em relação à média, a amostra deve ser de tamanho suficientemente grande, digamos $N \cong 200$, para produzir estimativas confiáveis do grau de achatamento da correspondente função de distribuição de freqüências. O coeficiente de curtose possui maior relevância para distribuições aproximadamente simétricas e também é um indicador do chamado *peso relativo das caudas* de tais distribuições. Com efeito, como o valor do coeficiente k indica quão aglomerados estão os pontos amostrais em torno da média, tem-se também a noção da distribuição dos valores muito distantes daquele valor central e, por conseguinte, das freqüências que se concentram nas caudas inferior e superior. Às vezes, subtrai-se o valor 3 da equação 2.14 para estabelecer o *coeficiente de excesso de curtose* k_e, em relação a uma distribuição padrão perfeitamente simétrica cujo valor de k é igual a 3. Nesse caso, se $k_e = 0$, a distribuição é dita mesocúrtica; se $k_e < 0$, é leptocúrtica; e se $k_e > 0$, é platicúrtica. A Figura 2.8 ilustra esquematicamente as situações mencionadas.

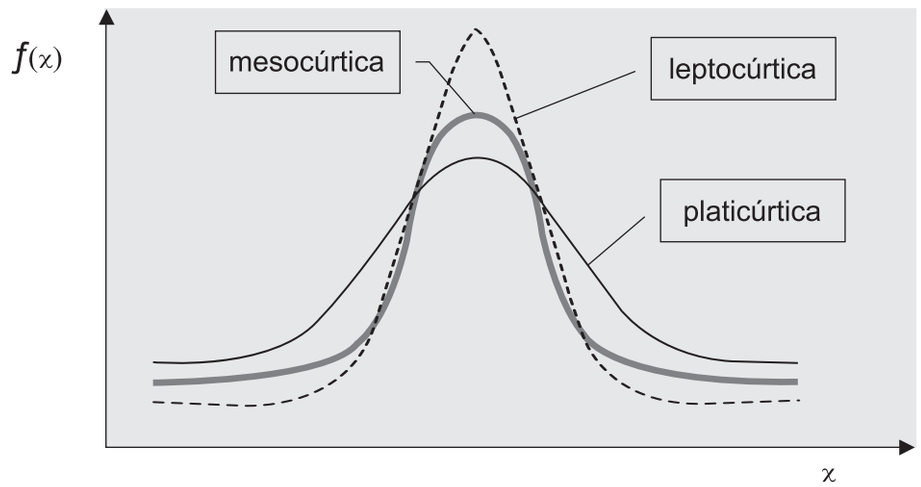


Figura 2.8 – Categorização das distribuições de freqüências com respeito à curtose

Em se tratando de séries hidrológicas, com amostras típicas de tamanho muito limitado, as estatísticas descritivas mais freqüentemente usadas, e consideradas representativas da forma do polígono de freqüências, são a média, o desvio padrão e o coeficiente de assimetria. De fato, essas estatísticas oferecem um sumário numérico conciso da informação contida em uma amostra. A título de exemplo, apresenta-se na Tabela 2.4 o cálculo das principais estatísticas descritivas das vazões médias anuais do Rio Paraopeba em Ponte Nova do Paraopeba, listadas na Tabela 2.2. Os resultados da Tabela 2.4 mostram que a moda é inferior à mediana, a qual, por sua vez, é menor do que a média, indicando, assim, uma assimetria positiva. Tal fato é comprovado pelo exame da Figura 2.3 e pelo coeficiente de assimetria amostral positivo de 0,808. Embora a amostra contenha apenas 62 observações, o coeficiente de excesso de curtose sugere uma distribuição platicúrtica, ou seja, relativamente menos pontiaguda em torno do valor central.

Tabela 2.4 – Estatísticas descritivas das vazões médias anuais do Rio Paraopeba em Ponte Nova do Paraopeba – Período 1938-1999

Estatística Amostral	Notação	Valor	Unidades	Cálculo
Média	$\bar{x}$	86,105	m^3/s	Equação 2.2
Moda	x_{mo}	80	m^3/s	Polígono Freqüências
Mediana	x_{md}	82,7	m^3/s	Equação 2.7
Média Harmônica	$\bar{x}_h$	79,482	m^3/s	Equação 2.5
Média Geométrica		82,726	m^3/s	Equação 2.6
Amplitude	A	123,3	m^3/s	(Máximo-Mínimo)
Primeiro Quartil	Q_1	68,2	m^3/s	Eq. 2.7 (1ª metade da série)
Terceiro Quartil	Q_3	99,1	m^3/s	Eq. 2.7 (2ª metade da série)
Ampl. Inter-Quartis	AIQ	30,9	m^3/s	(Q_3-Q_1)
Desvio Abs. Médio	d	19,380	m^3/s	Equação 2.8
Variância	s^2	623,008	$(m^3/s)^2$	Equação 2.10
Desvio Padrão	s	24,960	m^3/s	Equação 2.11
Coef. de Variação	CV	0,290	Adimensional	$s/\bar{x}$
Coef. de Assimetria	g	0,808	Adimensional	Equação 2.13
Coef. de Curtose	k	3,857	Adimensional	Equação 2.14
Excesso de Curtose	k_e	0,857	Adimensional	$(k-3)$

2.3 – Métodos Exploratórios

Tukey (1977) cunhou a denominação '*análise exploratória de dados*', tradução livre da terminologia de língua inglesa '*EDA - exploratory data analysis*', para identificar uma coleção de técnicas quantitativas e gráficas de exame e interpretação

de um conjunto de observações de uma variável aleatória, sem a preocupação prévia de formular premissas ou modelos matemáticos. A abordagem EDA baseia-se na idéia de que os dados revelam, por si mesmos, sua estrutura subjacente. Entre as técnicas gráficas propostas pela abordagem EDA, destaca-se o diagrama *box plot*, conhecido também pela denominação *desenho esquemático*, e o gráfico *ramo-e-folha*, tradução livre de 'stem-and-leaf'.

2.3.1 – O diagrama *Box Plot*

O diagrama *box plot* consiste em um retângulo definido pelo primeiro e pelo terceiro quartis, contendo a mediana em seu interior, tal como ilustrado na Figura 2.9, relativa às vazões médias anuais do Rio Paraopeba em Ponte Nova do Paraopeba. A partir do lado superior do retângulo, traça-se uma linha até o ponto que não exceda (Q_3+1,5AIQ), considerado limite superior para a identificação de *ouliers.* De modo análogo, traça-se outra linha a partir do lado inferior do retângulo até o limite dado por (Q_1-1,5AIQ). As observações que estiverem acima ou abaixo desses limites são identificadas no diagrama e consideradas *outliers* ou valores atípicos. Para a construção dos diagramas do tipo *box plot*, existem outras alternativas, tais como estender as linhas verticais até os pontos de máximo e mínimo, os quais são assinalados no gráfico por barras horizontais; nesse caso, o diagrama recebe a denominação de *box & whisker*.

Os diagramas do tipo *box plot* são muito úteis por permitirem uma visão geral do valor central, da dispersão, da assimetria, das caudas e de eventuais pontos amostrais discordantes. O valor central é dado pela mediana e a dispersão pela amplitude inter-quartis. A simetria ou assimetria da distribuição pode ser visualizada pelas posições relativas de Q_1, Q_2 e Q_3. Pode-se ter uma idéia das caudas superior e inferior por meio dos comprimentos das linhas verticais que saem do retângulo de quartis. Os diagramas do tipo *box plot* são particularmente úteis para comparar as características de duas ou mais amostras diferentes.

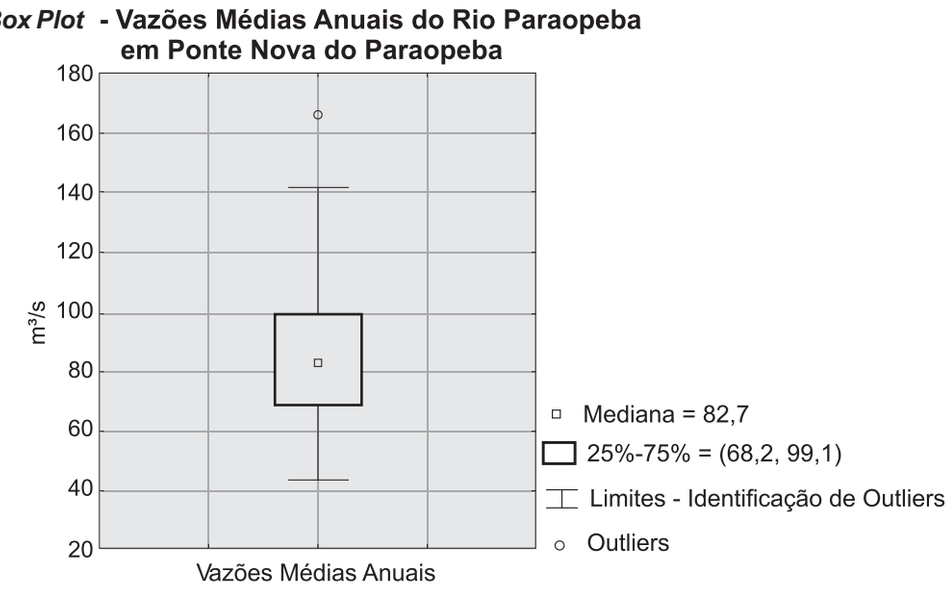


Figura 2.9 – Diagrama *Box Plot* para as vazões médias anuais do Rio Paraopeba em Ponte Nova do Paraopeba – Período 1938-1999

2.3.2 – O diagrama Ramo-e-Folha (*Stem-and-Leaf*)

Para amostras de tamanho médio a grande, o histograma é um procedimento gráfico eficaz para ilustrar a forma da distribuição de freqüências de uma variável. Para amostras menores, uma interessante alternativa ao histograma é dada pelo diagrama *ramo-e-folha*. De fato, esse diagrama agrupa os dados de tal modo, que há nenhuma ou pouca perda da informação contida em cada elemento amostral, realçando a presença de pontos extremos. Para exemplificar a construção de um diagrama ramo-e-folha, tomemos novamente a amostra de vazões médias anuais do Rio Paraopeba em Ponte Nova do Paraopeba, listadas na Tabela 2.2. Inicialmente, as 62 observações são classificadas em ordem crescente, entre o valor mínimo de 43,6 m^3/s e o máximo de 166,9 m^3/s, com grande concentração em torno de 80 m^3/s. Embora não exista uma regra fixa para a construção de um diagrama ramo-e-folha, a idéia central é dividir cada observação classificada em duas partes: a primeira, chamada de ramo, é posta à esquerda de uma linha vertical, enquanto a segunda, denominada folha, é colocada à direita, tal como mostra a Figura 2.10.

Freqüência Acumulada		RAMO	FOLHA					
0		2						
0		3						
0		3						
1		4	36					
3		4	68	94				
5		5	01	31				
8		5	70	73	99			
13	Q_1	6	06	12	26	36	42	
18		6	68	72	82	87	93	
22		7	16	20	24	48		
27		7	64	76	80	89	90	
(6)	Mediana	8	02	09	11	22	32	38
29		8	51	74	76	81	92	98
23		9	27	39				
21	Q_3	9	63	73	78	79	80	91
15		10	02	07	10	43		
11		10						
11		11	04	08	22	49		
7		11	71	84				
5		12	28					
4		12						
4		13	34					
3		13						
3		14	17	18				
1		14						
1		15						
1		15						
1		16						
1		16	69					
0		17						

Figura 2.10 – Diagrama Ramo-e-Folha para as vazões médias anuais do Rio Paraopeba em Ponte Nova do Paraopeba – Período 1938-1999

O ramo indica o dígito inicial, ou os dígitos iniciais, de cada observação, enquanto a folha mostra os dígitos complementares; no exemplo da Figura 2.10, o valor mínimo de 43,6 m^3/s é apresentado na quarta linha, com o ramo 4 e a folha 36, enquanto o máximo, na penúltima linha, tem ramo 16 e folha 69. Observe que, nesse exemplo, os ramos correspondem às dezenas e centenas, enquanto as folhas às unidades, multiplicadas por 10m^3/s. Um ramo com muitas folhas significa um número maior de ocorrências daquele ramo, tal como os dois ramos identificados pelo dígito inicial 8, na Figura 2.10. As freqüências das folhas são acumuladas da primeira linha até aquela que contém a mediana, de cima para baixo, e da última até a linha da mediana, de baixo para cima, e anotadas à esquerda da linha vertical, tal como ilustrado na Figura 2.10. Observe que a freqüência da linha da mediana

não é acumulada; note, também, a anotação complementar das linhas que contêm o primeiro e o terceiro quartis.

O diagrama ramo-e-folha, depois de sofrer uma rotação de 90º à esquerda em torno de seu centro, tem a aparência de um histograma, porém sem perda da informação individualizada por cada observação. Por meio do diagrama ramo-e-folha, é possível visualizar a posição da mediana, as amplitudes total e inter-quartis, a dispersão e a simetria (ou a assimetria) com que os pontos se dispõem em torno do valor central, os intervalos sem observações e a eventual presença de *outliers*. Na Figura 2.10, por conveniência, os ramos tiveram seus dígitos duplicados para melhor definição da concentração das folhas. Algumas vezes, o primeiro dos dígitos duplicados é marcado por um asterisco (*), para identificar que contém as folhas que iniciam de 0 a 4, enquanto o segundo o é por um ponto (•), para as folhas de 5 a 9. Em outras situações, poderia não haver tal duplicação. Em outros casos, as folhas também poderiam sofrer arredondamento para o inteiro mais próximo.

2.4 – Associação entre Variáveis

Nos itens precedentes, foram vistos os principais métodos de como organizar e resumir informações de uma amostra de dados de uma única variável. É freqüente, entretanto, o interesse em analisar o comportamento simultâneo de duas ou mais variáveis, buscando estabelecer eventuais associações entre elas. No presente item, examinaremos o caso mais simples de amostras de somente duas variáveis X e Y, geralmente observadas simultaneamente, ou organizadas em pares, os quais são denotados por $\{(x_1,y_1), (x_2,y_2), \ldots , (x_N, y_N)\}$. O que segue é apenas uma introdução ao tópico sobre regressão e correlação entre variáveis aleatórias, a ser detalhado no capítulo 9 desta publicação. Nesta introdução, destacamos os *diagramas de dispersão e de quantis-quantis* (Q-Q) de duas variáveis X e Y.

2.4.1 – Diagrama de Dispersão

Um diagrama de dispersão consiste em um gráfico onde são lançados em coordenadas cartesianas os pares $\{(x_1,y_1), (x_2,y_2), \ldots , (x_N, y_N)\}$ de observações simultâneas das variáveis X e Y. Para ilustrar a construção e as possibilidades de

um diagrama de dispersão, considere as variáveis X = altura anual de precipitação, em mm, e Y = vazão média anual, em m^3/s, cujas observações simultâneas na localidade de Ponte Nova do Paraopeba, tendo como base de cálculo o ano hidrológico regional de outubro a setembro, encontram-se listadas na Tabela 2.5. As Figuras 2.11 e 2.12 ilustram duas possibilidades interessantes de gráficos de dispersão: a primeira, acompanhada dos histogramas, e a segunda, com os diagramas do tipo *box-plot* grafados nos eixos correspondentes a cada uma das variáveis.

Tabela 2.5 – Vazões medias anuais e alturas anuais de precipitação (ano hidrológico Outubro-Setembro) – Estação Ponte Nova do Paraopeba (Flu:40800001, Plu:01944004)

Ano Hidrológico	Precipitação (mm)	Vazão média (m^3/s)	Ano Hidrológico	Precipitação (mm)	Vazão média (m^3/s)
1941/42	1249	91,9	1970/71	1013	34,5
1942/43	1319	145	1971/72	1531	80,0
1943/44	1191	90,6	1972/73	1487	97,3
1944/45	1440	89,9	1973/74	1395	86,8
1945/46	1251	79,0	1974/75	1090	67,6
1946/47	1507	90,0	1975/76	1311	54,6
1947/48	1363	72,6	1976/77	1291	88,1
1948/49	1814	135	1977/78	1273	73,6
1949/50	1322	82,7	1978/79	2027	134
1950/51	1338	112	1979/80	1697	104
1951/52	1327	95,3	1980/81	1341	80,7
1952/53	1301	59,5	1981/82	1764	109
1953/54	1138	53,0	1982/83	1786	148
1954/55	1121	52,6	1983/84	1728	92,9
1955/56	1454	62,3	1984/85	1880	134
1956/57	1648	85,6	1985/86	1429	88,2
1957/58	1294	67,8	1986/87	1412	79,4
1958/59	883	52,5	1987/88	1606	79,5
1959/60	1601	64,6	1988/89	1290	58,3
1960/61	1487	122	1989/90	1451	64,7
1961/62	1347	64,8	1990/91	1447	105
1962/63	1250	63,5	1991/92	1581	99,5
1963/64	1298	54,2	1992/93	1642	95,7
1964/65	1673	113	1993/94	1341	86,1
1965/66	1452	110	1994/95	1359	71,8
1966/67	1169	102	1995/96	1503	86,2
1967/68	1189	74,2	1996/97	1927	127
1968/69	1220	56,4	1997/98	1236	66,3
1969/70	1306	72,6	1998/99	1163	59,0

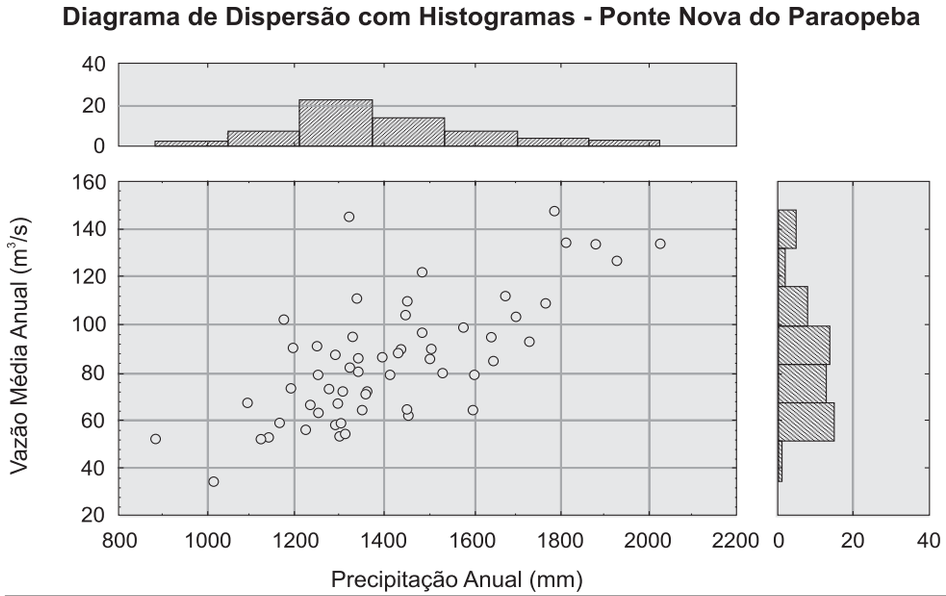


Figura 2.11 – Diagrama de Dispersão com Histogramas – Ponte Nova do Paraopeba

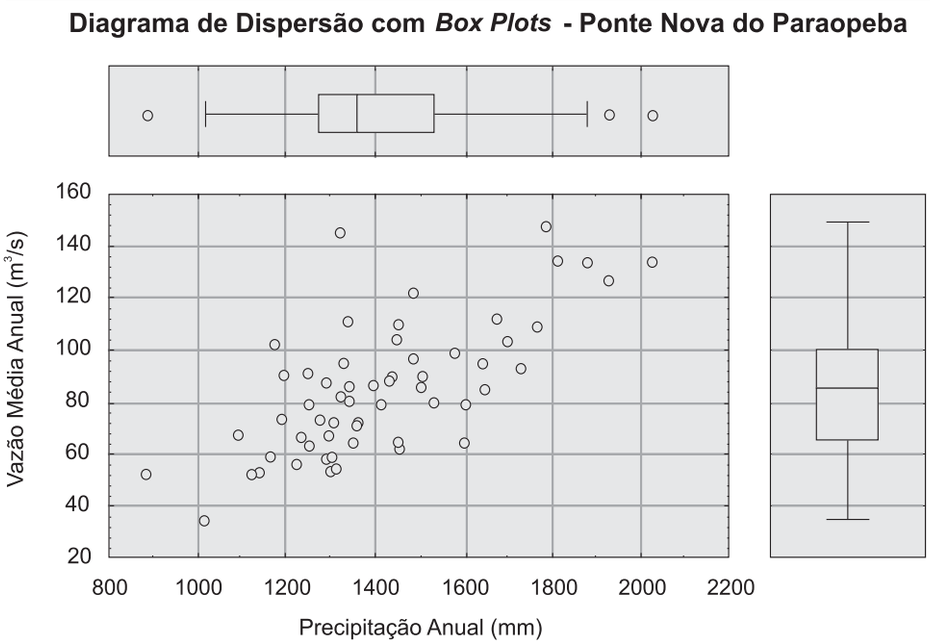


Figura 2.12 – Diagrama de Dispersão com *Box Plots* – Ponte Nova do Paraopeba

O exame dos diagramas de dispersão, das Figuras 2.11 e 2.12, mostra que, em geral, a maiores alturas de precipitação anual, correspondem maiores vazões médias anuais, indicando uma associação positiva entre as duas variáveis. Entretanto, observa-se também uma considerável dispersão entre os pares, demonstrando, com clareza, que a aleatoriedade presente em *Y* não pode ser explicada unicamente pela variação de *X*. De fato outras variáveis, como, por exemplo, a evapotranspiração, poderiam reduzir o grau de dispersão. Além disso, a bacia do Rio Paraopeba em Ponte Nova do Paraopeba drena uma área de 5.680 km^2, com considerável variação espacial das características climáticas e geomorfológicas, das propriedades do solo e das alturas pluviométricas. Os histogramas e os diagramas *box plots*, por sua vez, demonstram a presença de 3 *outliers* entre as alturas pluviométricas anuais, assim como a maior dispersão e a maior assimetria dessa variável, relativamente às vazões.

O grau de *associação linear* entre um conjunto de *N* pares de observações simultâneas de duas variáveis *X* e *Y* pode ser quantificado pelo *coeficiente amostral de correlação*, dado pela seguinte equação:

$$r_{X,Y} = \frac{s_{X,Y}}{s_X\, s_Y} = \frac{1}{N} \frac{\sum_{i=1}^{N} \left(x_i - \bar{x} \right) \left(y_i - \bar{y} \right)}{s_X\, s_Y} \tag{2.15}$$

Esse coeficiente *adimensional* é o resultado da padronização da *covariância amostral*, representada na equação 2.15 por $s_{X,Y}$, pelo produto $s_X\, s_Y$ entre os desvios-padrão das variáveis. Trata-se de um coeficiente que satisfaz a desigualdade $-1 \le r_{X,Y} \le 1$ e traduz o grau de associação linear entre as variáveis *X* e *Y*, a saber, nos casos extremos, 1 ou –1 para associações perfeitas positivas e negativas, respectivamente, e 0, para nenhuma associação.

A Figura 2.13-a mostra o caso de associação parcial positiva, quando *Y* cresce com o aumento de *X*, enquanto as Figuras 2.13-b e 2.13-c ilustram, respectivamente, a associação parcial negativa e nenhuma associação. A Figura 2.13-c mostra que um coeficiente de correlação nulo não implica, necessariamente, em nenhuma relação de dependência entre as variáveis; de fato, nesse caso, a relação de dependência existe, mas é não linear. Finalmente, é preciso ressaltar que uma eventual associação entre duas variáveis, medida por um alto valor do coeficiente de correlação, não implica em uma relação causa-efeito. Essa é clara em alguns casos, tais como a relação entre as precipitações e vazões médias anuais do Rio Paraopeba. Em outros, entretanto, tal relação de dependência física

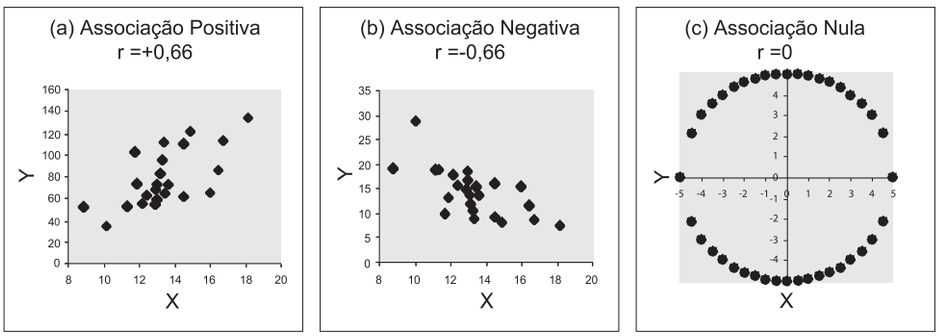


Figura 2.13 – Tipos de associação entre duas variáveis

não é evidente, mesmo que o coeficiente de correlação entre as variáveis tenha um valor elevado.

2.4.2 – Diagrama Quantis-Quantis (Q-Q)

O diagrama *quantis-quantis*, ou diagrama Q-Q, é outra representação gráfica que permite visualizar a associação entre duas variáveis X e Y. Diferentemente do diagrama de dispersão entre observações *simultâneas* das variáveis, o gráfico Q-Q é uma representação dos *dados ordenados* (*ou quantis*) do conjunto $\{x_1, x_2, ..., x_N\}$ contra os *dados ordenados* (*ou quantis*) da amostra de mesmo tamanho $\{y_1, y_2, ..., y_N\}$. Para elaborar um diagrama Q-Q, é necessário: (a) classificar os dados de X (e Y) em ordem crescente; (b) associar aos dados classificados os seus respectivos números de ordem da classificação m, com $1 \leq m \leq N$; e (c) associar aos dados classificados as correspondentes freqüências ou probabilidades empíricas de não superação. Em seguida, os dados de X e Y, com igual freqüência ou probabilidade empírica de não superação, são lançados em coordenadas cartesianas, formando, assim, o diagrama Q-Q. A Figura 2.14 é um exemplo de um diagrama Q-Q elaborado para os dados da Tabela 2.5.

De modo diverso de um diagrama de dispersão, o qual estabelece uma associação global entre as variáveis, o gráfico Q-Q demonstra se os valores mais baixos, médios e mais altos de X estão relacionados aos seus correspondentes de Y. Em um caso limite, se as distribuições dos dois conjuntos de dados fossem idênticas, a menos de suas medidas de posição e escala (ou dispersão), os pontos estariam

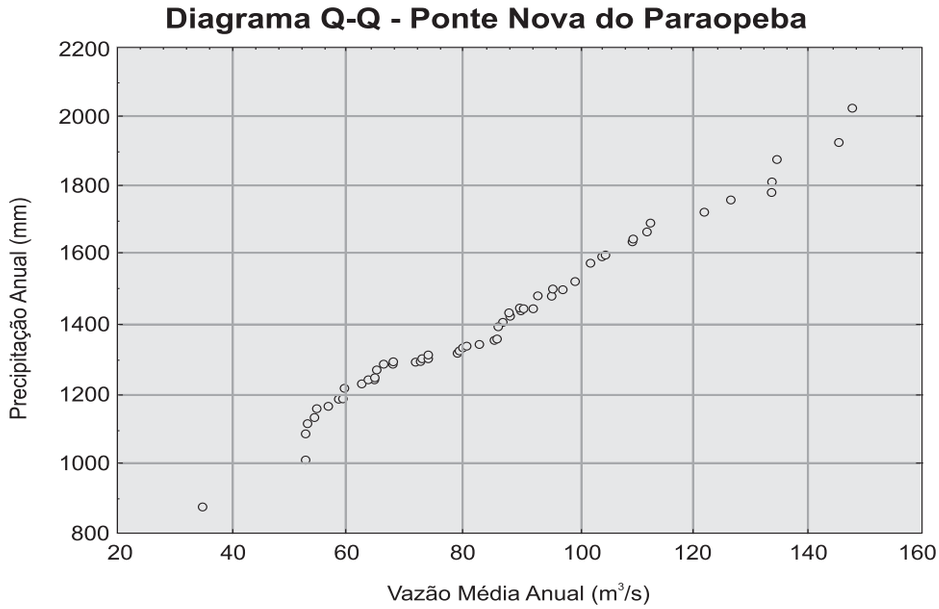


Figura 2.14 – Diagrama Quantis-Quantis entre Vazões Médias Anuais e Alturas Anuais de Precipitação de Ponte Nova do Paraopeba

sobre a reta $y = x$. O modo como os pontos se afastam dessa linearidade revelam as diferenças entre as distribuições de X e Y.

Exercícios

1) Com referência à série parcial das N maiores vazões média diárias, em N anos de registros, do Rio Paraopeba em Ponte Nova do Paraopeba, objeto do exercício 9 do Capítulo 1, faça uma diagrama de linha para a variável discreta 'número de cheias anuais', tal como o da Figura 2.1.

2) Na Tabela 2.5, tome a série de vazões médias anuais do Rio Paraopeba em Ponte Nova do Paraopeba, calculadas com base no ano hidrológico de Outubro a Setembro, e faça os seguintes gráficos:

- diagrama uniaxial de pontos;
- histograma;
- polígono de freqüências relativas;
- diagrama de freqüências relativas acumuladas; e
- diagrama de quantis.

3) Compare os gráficos elaborados no exercício 2 com os mostrados no item 2.1 do presente capítulo. Interprete as diferenças entre eles. Em se tratando da variável vazão média anual, é mais representativo trabalhar com séries reduzidas em ano civil ou ano hidrológico?

4) Com referência à curva de permanência da Figura 2.7, qual seria o *máximo valor teórico* da vazão Q_r a ser constantemente mantida a jusante de um hipotético reservatório de regularização sazonal? Por que esse valor seria o 'máximo teórico'? Calcule o volume do reservatório para a situação descrita.

5) Volte aos dados do exercício 2 e faça um sumário numérico completo da amostra em questão, calculando todas as medidas de posição, dispersão, assimetria e curtose. Interprete e compare os resultados com aqueles apresentados no item 2.2 desse capítulo.

6) Se o primeiro terço de um trecho fluvial é percorrido por um flutuador, a uma velocidade de 0,3 m/s, o segundo a 0,5 m/s e o terceiro a 0,60 m/s, prove que a média harmônica é mais representativa da velocidade média do flutuador, calculada ao longo de todo o trecho fluvial, do que a média aritmética.

7) A população de uma cidade aumenta geometricamente com o tempo. Suponha que no censo de 1980, a população dessa cidade era de 150.000 habitantes, enquanto em 2000 cresceu para 205.000 habitantes. Com a finalidade de verificar as condições de projeto do sistema local de abastecimento de água, um engenheiro sanitarista necessita estimar o consumo de água *per capita* no período intermediário e, portanto, a população em 1990. Calcule o valor central a ser usado. Justifique sua resposta.

8) Uma variável aleatória pode sofrer transformações lineares e não lineares. Um exemplo de transformação linear de X é alterá-la para a *variável central reduzida* Z, por meio de $z_i = (x_i - \bar{x})/s_x$. De fato, nesse caso, X é centrada pela subtração da medida de posição e tem sua escala reduzida pela divisão pelo desvio padrão. Agora, volte aos dados do exercício 2, calcule $\bar{z}$, s_z, g_z e k_z e compare com as mesmas medidas de X, já calculadas no exercício 5. Quais conclusões se pode tirar de uma variável que sofreu uma transformação linear?

9) Um exemplo de transformação não linear é dado pela logaritmização de X, ou seja, $z_i = \log_{10} x_i$ ou $z_i = \ln x_i$. Repita o exercício 8 para essa nova transformação

e tire suas conclusões. Sob a ótica de sua resposta ao exercício 13 do capítulo 1, em que situações você espera verificar uma utilidade prática em uma transformação não linear de uma variável aleatória?

10) Uma família de possibilidades de transformação de uma variável aleatória é dada pela fórmula de transformações potenciais de Box-Cox, ou seja, $z_i = (x_i^{\lambda} - 1)/\lambda$, se $\lambda \neq 0$, ou $z_i = \ln x_i$, se $\lambda = 0$. A escolha correta da potência de transformação λ pode tornar dados originais assimétricos em aproximadamente simétricos. Usando a expressão de Box-Cox com λ = -1, -0,5, 0, +0,5, +1 e +2, transforme os dados da Tabela 2.2, calcule os coeficientes de assimetria e curtose, e verifique qual é o valor de λ que os torna os dados aproximadamente simétricos. Refaça o polígono de freqüências relativas para os dados transformados e compare-o com o da Figura 2.4.

11) Para construir um diagrama de freqüências relativas acumuladas, é necessário, como se viu no item 2.1.5, estimar a probabilidade empírica de não superação $P(X \leq x)$ por meio dos números de ordem de classificação m. No exemplo do item 2.1.5, foi usada a expressão m/N para se estimar $P(X \leq x)$. Contudo, tal estimativa é precária porque implica que é nula a probabilidade da variável produzir um valor maior do que o máximo amostral. Para evitar tal inconveniente, foram propostas diversas fórmulas alternativas para a estimativa de $P(X \leq x)$; na literatura hidrológica, tais fórmulas são conhecidas por fórmulas de "posição de plotagem", decorrente de adaptação do termo em inglês '*plotting position*'. Uma das mais conhecidas é a de Weibull, dada pela expressão $m/(N+1)$. Refaça o diagrama de Figura 2.5, usando a fórmula de Weibull.

12) No anexo 1 desse livro, você encontrará as vazões médias mensais do Rio Paraopeba em Ponte Nova do Paraopeba, de 1938 a 1999. Coloque em um mesmo gráfico os diagramas *box plot* das vazões médias mensais de Janeiro e de Setembro. Interprete os diagramas.

13) Faça e interprete o diagrama ramo-e-folha para as alturas anuais de precipitação observadas na estação de Ponte Nova do Paraopeba, listadas na Tabela 2.5.

14) Interprete o diagrama Q-Q da Figura 2.14.

15) A tabela abaixo se refere aos dados de concentração de sólidos totais dissolvidos e vazão, observados no Rio Cuyahoga na estação de Independence

(código USGS 4208000), no estado americano de Ohio, tais como publicados por Helsel e Hirsch (1992). Os símbolos M e T representam, respectivamente, o mês e o tempo decimal (ano-1000), da realização das medições. A vazão *Q* está expressa em pés cúbicos por segundo e a concentração de sólidos totais *SDT* está em mg/l. Pede-se:

Tabela 2.6 – Exercício 15

Mês	T	SDT	Q	Mês	T	SDT	Q	Mês	T	SDT	Q	Mês	T	SDT	Q
1	74,04	490	458	2	78,12	680	533	10	79,79	410	542	7	81,54	560	444
2	74,12	540	469	3	78,21	250	4930	11	79,87	470	499	8	81,62	370	595
4	74,29	220	4630	4	78,29	250	3810	12	79,96	370	741	9	81,71	460	295
7	74,54	390	321	5	78,37	450	469	1	80,04	410	569	10	81,79	390	542
10	74,79	450	541	6	78,46	500	473	2	80,12	540	360	12	81,96	330	1500
1	75,04	230	1640	7	78,54	510	593	3	80,21	550	513	3	82,21	350	1080
4	75,29	360	1060	8	78,62	490	500	4	80,29	220	3910	5	82,37	480	334
7	75,54	460	264	9	78,71	700	266	5	80,37	460	364	6	82,46	390	423
10	75,79	430	665	10	78,79	420	495	6	80,46	390	472	8	82,62	500	216
1	76,04	430	680	11	78,87	710	245	7	80,54	550	245	11	82,87	410	366
4	76,29	620	650	12	78,96	430	736	8	80,62	320	1500	2	83,12	470	750
8	76,62	460	490	1	79,04	410	508	9	80,71	570	224	5	83,37	280	1260
10	76,79	450	380	2	79,12	700	578	10	80,79	480	342	8	83,62	510	223
1	77,04	580	325	3	79,21	260	4590	12	80,96	520	732	11	83,87	470	462
4	77,29	350	1020	4	79,29	260	4670	1	81,04	620	240	2	84,12	310	7640
7	77,54	440	460	5	79,37	500	503	2	81,12	520	472	5	84,37	230	2340
10	77,79	530	583	6	79,46	450	469	3	81,21	430	679	7	84,54	470	239
11	77,87	380	777	7	79,54	500	314	4	81,29	400	1080	11	84,87	330	1400
12	77,96	440	1230	8	79,62	620	432	5	81,37	430	920	3	85,21	320	3070
1	78,04	430	565	9	79,71	670	279	6	81,46	490	488	5	85,37	500	244

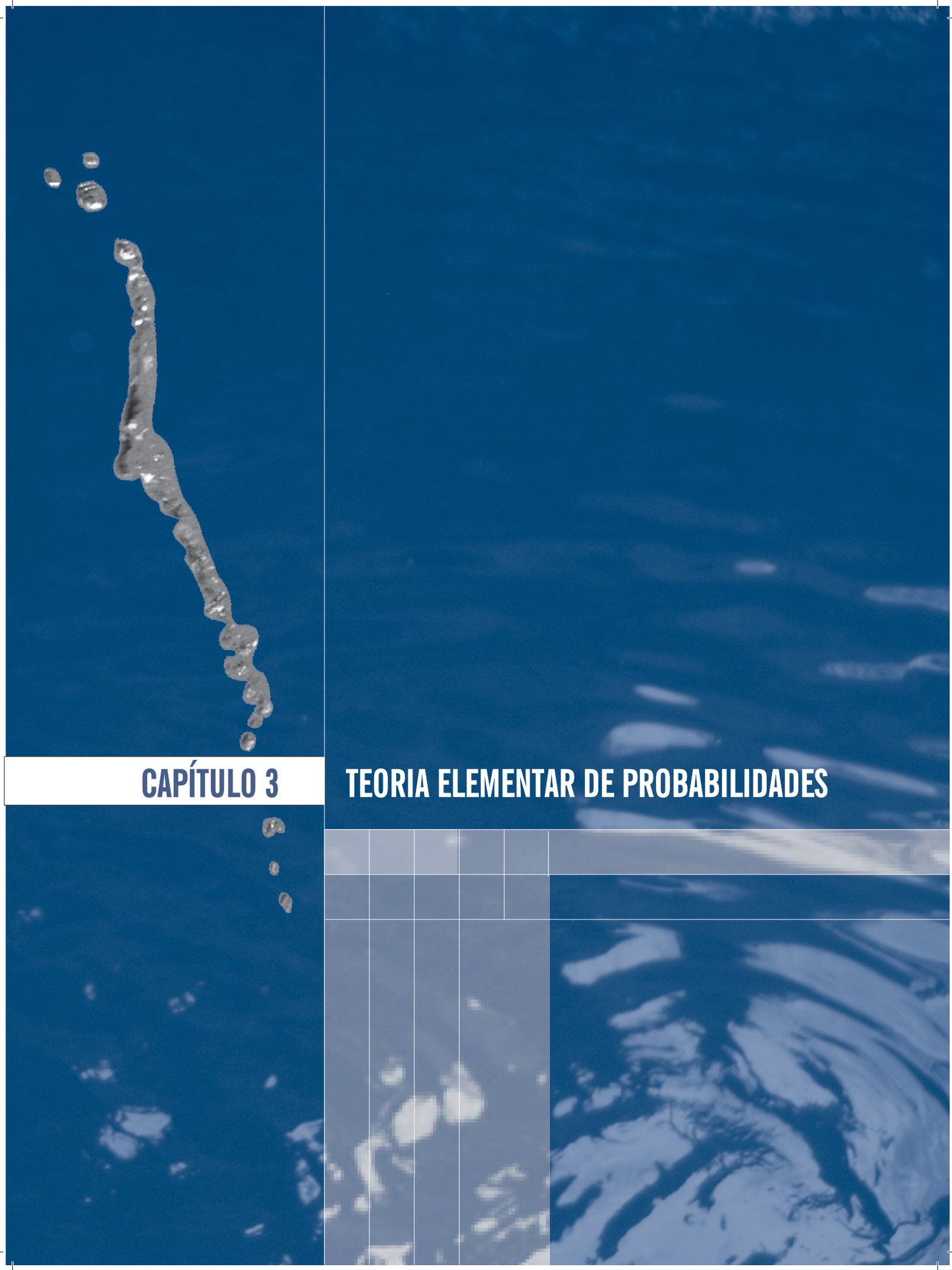
a) registrar em um único gráfico a variação temporal das variáveis *Q* e *SDT*;

b) elaborar e interpretar os diagramas de dispersão, com histogramas e com gráficos do tipo *box plot*, para as variáveis *Q* e *SDT*;

c) calcular o coeficiente de correlação linear entre as variáveis *Q* e *SDT*;

d) no caso em foco, dar a justificativa física do sinal do coeficiente de correlação; e

e) elaborar e interpretar o diagrama quantis-quantis para as variáveis *Q* e *SDT*.



CAPÍTULO 3

TEORIA ELEMENTAR DE PROBABILIDADES

CAPÍTULO 3 TEORIA ELEMENTAR DE PROBABILIDADES

No capítulo 2, viu-se que a análise preliminar de uma amostra de dados hidrológicos, por meio de um conjunto de técnicas numéricas e gráficas, permite que se tenha uma idéia inicial da distribuição de freqüências da variável em questão. Entretanto, as medidas de posição, dispersão, assimetria e curtose são meras estimativas de quantidades populacionais desconhecidas, enquanto as freqüências calculadas o são das *probabilidades* de ocorrência de certos eventos. Para extrair conclusões de uma amostra de dados hidrológicos, que sejam úteis à tomada de decisões no planejamento e projeto de sistemas de recursos hídricos, é necessário estabelecer um *modelo matemático* que contenha os principais elementos do processo hidrológico que determinou a ocorrência daquelas observações. Como visto no capítulo 1, tal modelo deve ser *probabilístico* pela impossibilidade de se sintetizar em um conjunto de equações a lei que descreve rigorosamente a variação de um certo fenômeno hidrológico. Um modelo probabilístico, embora seja incapaz de prever com exatidão a data e a magnitude de uma enchente, por exemplo, revela-se muito útil no estudo do regime local de cheias, especificando com que probabilidade uma certa vazão irá ser igualada ou superada, em um ano qualquer. O presente capítulo tem por objetivo estabelecer os princípios da teoria de probabilidades, necessários à construção de modelos probabilísticos de fenômenos hidrológicos.

3.1 – Eventos Aleatórios

A teoria de probabilidades lida com a realização de *experimentos*, naturais ou planejados pelo homem, cujos *resultados* não podem ser previstos com exatidão. Embora os resultados de um experimento, realizado sob condições uniformes e não tendenciosas, não possam ser antecipados com exatidão, é possível estabelecer o conjunto que contem todos os resultados possíveis ou esperados de tal experimento. A esse conjunto, denotado por *S*, dá-se o nome de *espaço amostral*, o qual contem os chamados *pontos* ou *elementos amostrais*. Suponha, por exemplo, que o experimento se referisse à identificação e contagem do número anual de dias *Y* com alturas diárias de chuva iguais ou superiores a 0,1 mm, observados em uma certa estação pluviométrica; nesse caso, o espaço amostral seria dado pelo conjunto finito $S \equiv S_D = \{ y = 0,1,2,...,366 \}$, cuja composição é de elementos extraídos do conjunto **N** dos números naturais. Por outro lado, se o experimento se referisse ao monitoramento das vazões *X*, em uma certa estação fluviométrica, o espaço amostral seria $S \equiv S_C = \{ x \in \mathbf{R}_+ \}$, ou seja o conjunto infinito dos números reais não negativos.

Qualquer subconjunto do espaço amostral S é chamado de *evento*. No espaço amostral S_C, das vazões X, poderíamos distinguir os valores inferiores a um certo limiar x_0 e agrupá-los no evento $A = \{x \in \mathbf{R}_+ | 0 \le x < x_0\}$, tal que A esteja contido em S_C. O *complemento* de um evento A, denotado por A^c, consiste de todos os elementos de S_C que não estão incluídos em A; em outras palavras, $A^C = \{x \in \mathbf{R}_+ | x \ge x_0\}$ implica na não ocorrência do evento A. Da mesma forma, de volta ao espaço amostral S_D, do número anual de dias chuvosos Y, poderíamos, a título de exemplo, categorizar como anos secos aqueles em que $y < 30$ dias e definir o evento $B = \{y \in \mathbf{N} | y < 30\}$; nesse caso, o complemento de B seria dado pelo conjunto finito $B^c = \{y \in \mathbf{N} | 30 \le y \le 366\}$. Nos exemplos dados, os eventos A e A^c, assim como os eventos B e B^c, quando considerados dois a dois, são denominados *disjuntos* ou *mutuamente excludentes* porque a ocorrência de um implica na não ocorrência do outro; em outras palavras, nenhum dos elementos amostrais contidos em um evento está contido no outro.

Os eventos contidos em um espaço amostral podem estar relacionados entre si pelas operações de *interseção* e de *união*. Se dois eventos *não mutuamente excludentes* A_1 e A_2 possuem *elementos amostrais em comum*, o subconjunto que contem tais elementos constitui a interseção, a qual é representada por $A_1 \cap A_2$ .Contrariamente, se os eventos A_1 e A_2 são disjuntos, sua interseção $A_1 \cap A_2 = \emptyset$, onde Ø representa o *conjunto vazio*; Ø é rigorosamente definido como o complemento S^c do espaço amostral. O subconjunto que contem *todos os elementos amostrais* de A_1 e A_2, incluindo os comuns a ambos, constitui a união, a qual é representada por $A_1 \cup A_2$. A operação de interseção está associada ao operador lógico "e", indicando ocorrência conjunta ou simultânea, enquanto a união associa-se a "e/ou", ou seja, A_1 ou A_2 ou ambos conjuntamente. De volta ao espaço amostral S_C, do exemplo das vazões, considere a existência de alguns eventos hipotéticos, definidos por $A_1 = \{x \in \mathbf{R}_+ | 0 \le x \le 60\ m^3/s\}$,
$A_2 = \{x \in \mathbf{R}_+ | 30\,m^3/s \le x \le 80\ m^3/s\}$ e $A_3 = \{x \in \mathbf{R}_+ | x \ge 50\ m^3/s\}$.

Nesse caso, pode-se extrair as seguintes conclusões:

i. $A_1 \cap A_2 = \{x \in \mathbf{R}_+ | 30\,m^3/s \le x \le 60\ m^3/s\}$

ii. $A_2 \cap A_3 = \{x \in \mathbf{R}_+ | 50\,m^3/s \le x \le 80\ m^3/s\}$

iii. $A_1 \cap A_3 = \{x \in \mathbf{R}_+ | 50\,m^3/s \le x \le 60\ m^3/s\}$

iv. $A_1 \cup A_2 = \{x \in \mathbf{R}_+ | 0\,m^3/s \le x \le 80\ m^3/s\}$

v. $A_2 \cup A_3 = \{x \in \mathbf{R}_+ | 30\,m^3/s \le x \le \infty\}$

vi. $A_1 \cup A_3 = \{x \in \mathbf{R}_+ | x \ge 0\} \equiv S_C$

As operações de interseção e união podem ser estendidas a mais de dois eventos e estão sujeitas às propriedades *associativa* e *distributiva*, de modo análogo às regras que se aplicam à adição e à multiplicação de números. Os seguintes eventos compostos são exemplos de aplicação da propriedade associativa:

$(A_1 \cup A_2) \cup A_3 = A_1 \cup (A_2 \cup A_3)$ e $(A_1 \cap A_2) \cap A_3 = A_1 \cap (A_2 \cap A_3)$.

As operações $(A_1 \cup A_2) \cap A_3 = (A_1 \cap A_3) \cup (A_2 \cap A_3)$

e $(A_1 \cap A_2) \cup A_3 = (A_1 \cup A_3) \cap (A_2 \cup A_3)$ resultam da aplicação da propriedade

distributiva. Referindo-se ao espaço amostral S_C, pode-se escrever,

i. $A_1 \cap A_2 \cap A_3 = \{x \in \mathbf{R}_+ | 50\,m^3/s \le x \le 60\ m^3/s\}$

ii. $A_1 \cup A_2 \cup A_3 = \{x \in \mathbf{R}_+ | x \ge 0\} \equiv S_C$

iii. $(A_1 \cup A_2) \cup A_3 = A_1 \cup (A_2 \cup A_3) = S_C$

iv. $(A_1 \cap A_2) \cap A_3 = A_1 \cap (A_2 \cap A_3) = \{x \in \mathbf{R}_+ | 50\,m^3/s \le x \le 60\,m^3/s\}$

v. $(A_1 \cup A_2) \cap A_3 = (A_1 \cap A_3) \cup (A_2 \cap A_3) = \{x \in \mathbf{R}_+ | 50\,m^3/s \le x \le 60\,m^3/s\}$

vi. $(A_1 \cap A_2) \cup A_3 = (A_1 \cup A_3) \cap (A_2 \cup A_3) = \{x \in \mathbf{R}_+ | x \ge 30\,m^3/s\}$

As operações entre eventos simples e compostos, dispostos em um espaço amostral, podem ser mais facilmente visualizadas, por meio dos chamados diagramas de Venn, como o ilustrado pela Figura 3.1. Esses diagramas, entretanto, não são completamente apropriados para mensurar ou interpretar relações de probabilidades entre eventos.

Como decorrência das operações entre eventos, é possível expressar o espaço amostral como resultado da *união de um conjunto exaustivo de eventos mútua e coletivamente excludentes*. De fato, com referência à Figura 3.1, os eventos $(A \cap B^c), (A \cap B), (A^c \cap B)$ e $(A^c \cap B^c)$ são mútua e coletivamente excludentes, sendo intuitivo verificar que a união de todos eles resulta no espaço amostral *S*.

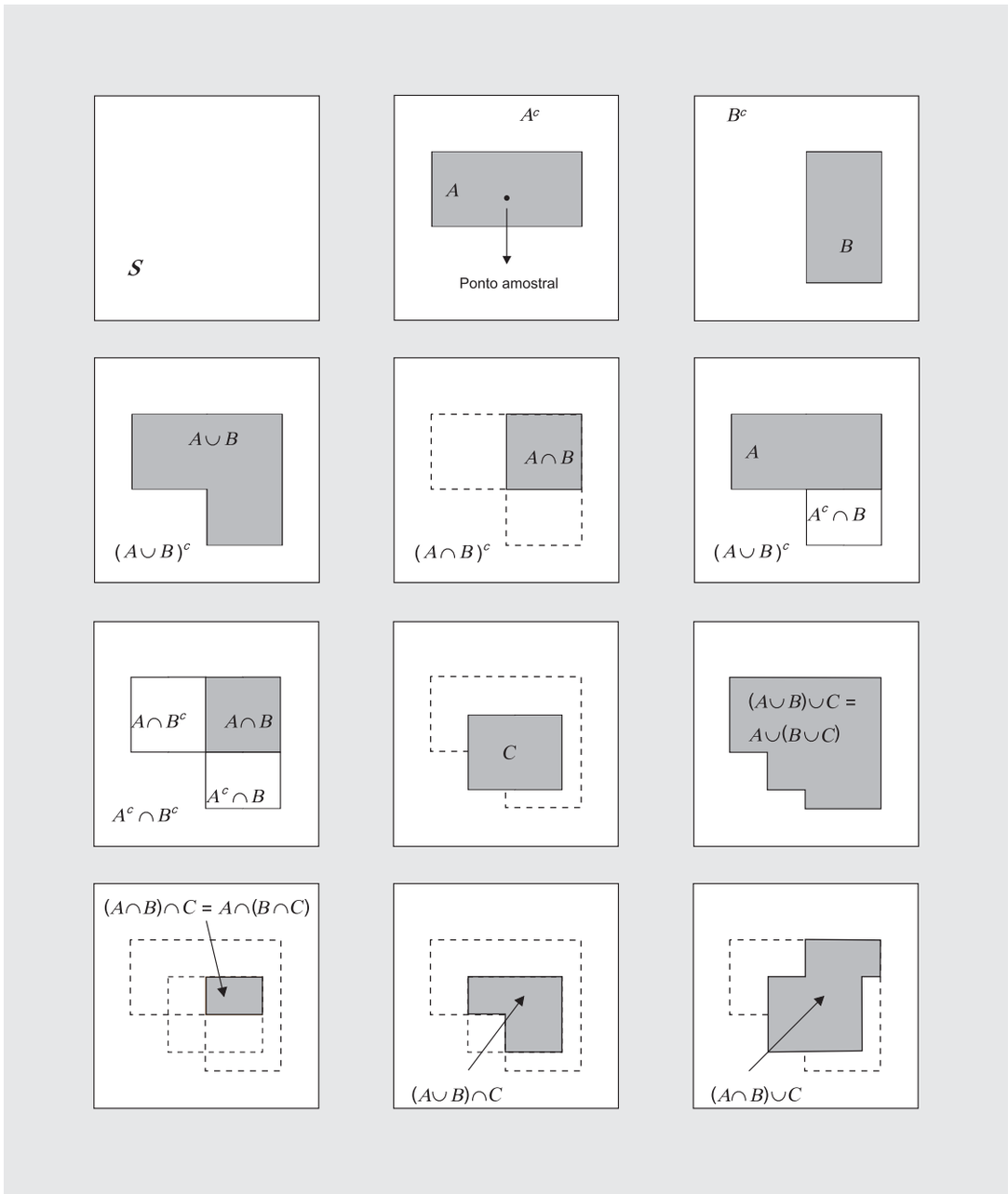


Figura 3.1 – Diagramas de Venn e operações com eventos em um espaço amostral [adap. de Kottegoda e Rosso (1997)]

Quando o experimento envolve observações simultâneas de diversas variáveis, a noção anterior deve ser estendida para a de um *espaço amostral multidimensional*. Em hidrologia, são inúmeros os exemplos de associação entre observações simultâneas de duas ou mais variáveis: número de dias chuvosos e alturas de precipitação em um certo intervalo de tempo; número anual de cheias, vazões de ponta e volumes dos hidrogramas correspondentes, entre outros. O exemplo 3.1 ilustra o espaço bi-dimensional formado pelas vazões de dois rios a montante de sua confluência.

Exemplo 3.1- O rio R_3 é formado pela confluência dos ribeirões R_1 e R_2. Durante a estação seca, as vazões X de R_1, imediatamente a montante da confluência, variam entre 150 l/s e 750 l/s, enquanto as vazões Y do ribeirão R_2, também a montante da confluência, variam no intervalo de 100 a 600 l/s. O espaço amostral bi-dimensional é dado por $S = \{(x,y) \in \mathbf{R}_+ | 150 \le x \le 750, 100 \le y \le 600\}$ e está ilustrado na Figura 3.2. Os eventos A, B e C, ilustrados na Figura 3.2, são definidos da seguinte forma: A = {as vazões de R_3 superam 850 l/s}, B = {as vazões de R_1 superam as de R_2} e C = {as vazões de R_3 são inferiores a 750 l/s}. A interseção entre A e B corresponde ao evento $A \cap B = \{(x,y) \in S | x + y > 850 \; e \; x > y\}$ e está indicada na Figura 3.2 pelo polígono formado pelos pontos 3,6,9 e 10. A união $A \cup B = \{(x,y) \in S | x + y > 850 \text{ e/ou } x > y\}$ corresponde ao polígono formado pelos pontos 1,4, 9, 10 e 3, enquanto o evento $A \cap C = \emptyset$. Aproveite o exemplo para definir e identificar graficamente os seguintes eventos: $(A \cup C)^c$, $(A \cup C)^c \cap B$ e $A^c \cap C^c$.

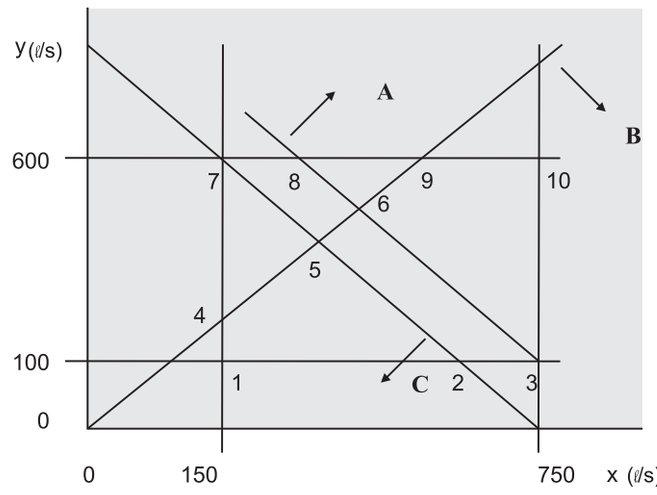


Figura 3.2 – Espaço amostral bi-dimensional para os eventos do exemplo 3.1

3.2 – Noção e Medida de Probabilidade

Uma vez definidos o espaço amostral e os eventos aleatórios, a etapa seguinte é a de associar uma "probabilidade" a cada um desses eventos, ou seja, uma medida relativa de sua chance de ocorrer, entre os extremos de 0 (*impossibilidade*) e 1 (*certeza*). Apesar de tal medida ser algo intuitiva, sua definição matemática teve uma evolução histórica lenta, incorporando modificações graduais, necessárias à acomodação das diferentes noções e interpretações do conceito de probabilidade.

A primeira definição, denominada *clássica* ou *a priori*, teve suas origens nos trabalhos de matemáticos do século XVII, como Blaise Pascal (1623-1662) e Pierre de Fermat (1601-1665), no contexto dos jogos de azar. Segundo essa definição, se um espaço amostral finito S contem n_S formas equiprováveis e mutuamente excludentes dos resultados de um experimento, das quais n_A estão associadas a um determinado atributo A, a probabilidade de ocorrência do evento de atributo A é:

$$P(A) = \frac{n_A}{n_S} \tag{3.1}$$

Essa é a chamada definição *a priori* porque pressupõe, antes dos fatos, que os eventos são equiprováveis e mutuamente excludentes. Por exemplo, no lançamento de uma moeda, a qual sabe-se ser *não tendenciosa*, a probabilidade de resultar 'cara' ou 'coroa' é 0,5.

Existem muitas situações em que a definição clássica é completamente apropriada, enquanto, em outras, duas limitações são óbvias. A primeira refere-se à impossibilidade de acomodar o cenário em que os resultados do experimento não sejam equiprováveis, enquanto a segunda diz respeito à não contemplação de espaços amostrais infinitos. Essas limitações determinaram a formulação da definição de probabilidade, denominada *empírica* ou *a posteriori*, mais abrangente e, geralmente, atribuída ao matemático austríaco Richard von Mises (1883-1953). Segundo tal definição, se um experimento é realizado um *grande número* de vezes n, sob condições rigorosamente idênticas, e o evento de atributo A, contido no espaço amostral S, ocorre n_A vezes, então, a probabilidade empírica ou *a posteriori* de A é dada por

$$P(A)=\lim_{n\to\infty} \frac{n_A}{n} \tag{3.2}$$

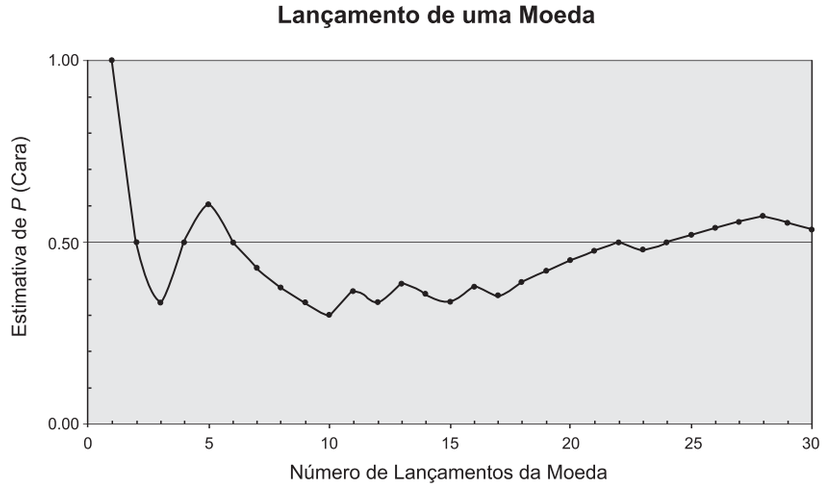


Figura 3.3 – Ilustração da definição empírica ou *a posteriori* de probabilidade

Essa definição é ilustrada pelo gráfico da Figura 3.3, referente à probabilidade do resultado 'cara', em função do número de lançamentos de uma moeda, em relação à qual, nenhuma suposição inicial é feita.

A definição empírica, embora mais abrangente, também possui limitações. A primeira refere-se à determinação de quão grande deve ser o valor de n para proporcionar uma estimativa adequada de P(A); no caso ilustrado pela Figura 3.3, essa limitação fica evidenciada pela impossibilidade de concluir categoricamente a probabilidade do resultado 'cara', ao final dos 30 lançamentos da moeda. Outra limitação refere-se à impossibilidade física de se repetir um experimento um número infinito de vezes, sob condições rigorosamente idênticas. Além dessas limitações, nem a definição *a priori* ou a definição *a posteriori* podem acomodar a noção de *probabilidade subjetiva*, qual seja, aquela que decorre da atribuição de uma ponderação relativa a um evento, com base na experiência ou julgamento pessoal de um especialista. Por exemplo, um engenheiro geotécnico pode usar de sua experiência técnica para atribuir uma probabilidade subjetiva de ocorrência de fraturas na rocha sobre a qual se apóia uma barragem de gravidade. Tais inconsistências proveram a motivação necessária para a formulação de probabilidade como uma função que se comporta de acordo com um determinado conjunto de postulados ou axiomas.

Em 1933, o matemático russo Andrei Kolmogorov (1903-1987) formulou a chamada definição *axiomática* de probabilidade, a qual estabelece a essência lógica do comportamento da função de probabilidade P(.), com base em somente três postulados. A probabilidade de um evento A, contido em um espaço amostral

S, é um número não negativo, denotado por P(*A*), que satisfaz as seguintes condições:

i. $0 \leq P(A) \leq 1$

ii. $P(S) = 1$

iii. Para qualquer seqüência de eventos mutuamente excludentes $E_1, E_2, \ldots$, a probabilidade da união desses eventos é igual à soma das respectivas probabilidades individuais, ou seja,

$$P\left(\bigcup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} P(E_i)$$

As 3 condições enumeradas são, de fato, axiomas sobre os quais todas as propriedades matemáticas da função de probabilidade P(.) podem ser deduzidas. A definição axiomática de probabilidade forma a essência lógica da moderna teoria de probabilidades e acomoda não somente as definições anteriores, como também a noção de probabilidade subjetiva.

São decorrências dos 3 axiomas de Kolmogorov, as seguintes proposições:

i. $P(A^c) = 1\text{-}P(A)$

ii. $P(\emptyset) = 0$

iii. Se *A* e *B* são dois eventos no espaço amostral e $A \subset B$, então $P(A) \leq P(B)$

iv. Para qualquer evento *A*, $P(A) \leq 1$

v. Se $A_1, A_2, \ldots, A_k$ são eventos definidos em uma espaço amostral, então, $P\left(\bigcup_{i=1}^{k} A_i\right) \leq \sum_{i=1}^{k} P(A_i)$. Essa é a chamada desigualdade de Boole.

vi. Se *A* e *B* são dois eventos no espaço amostral, então, $P(A \cup B) = P(A) + P(B) - P(A \cap B)$. Essa é a chamada regra da adição de probabilidades.

Exemplo 3.2 - Em uma área sujeita a terremotos, dois eventos naturais podem produzir a ruptura de uma barragem, a saber: a ocorrência de uma enchente maior do que a cheia de projeto do vertedouro (evento *A*) ou o colapso estrutural devido a um terremoto destrutivo (evento *B*). Suponha que, com base em dados anuais observados em um dado local, foram

estimadas as seguintes probabilidades P(A) = 0,02 e P(B) = 0,01. Com base apenas nesses valores, estime a probabilidade da barragem se romper em um ano qualquer.

Solução: O rompimento da barragem pode ser provocado pela ação das cheias, pela ação dos terremotos ou pela ação de ambos; em outras palavras, o rompimento é um evento composto pela união dos eventos A e B. A probabilidade de rompimento é dada por $P(A \cup B) = P(A) + P(B) - P(A \cap B)$, mas não se conhece $P(A \cap B)$. Entretanto, sabe-se que $P(A \cap B)$ deve ser um valor extremamente baixo. Com base nessas considerações e na desigualdade de Boole, pode-se fazer uma estimativa conservadora de que $P(A \cup B) \cong P(A) + P(B) = 0{,}02+0{,}01 = 0{,}03$

3.3 – Probabilidade Condicional e Independência Estatística

A probabilidade de um evento A pode ser alterada pela ocorrência de um outro evento B. Por exemplo, a probabilidade de que a vazão média de uma bacia irá superar 50 m^3/s, nas próximas 6 horas, é certamente alterada pelo fato de que ela já superou 20 m^3/s. Esse e vários outros são exemplos de *probabilidade condicional*, ou seja, a probabilidade P($A|B$) de ocorrência de um evento A, dado que outro evento B já ocorreu ou que é certo de ocorrer. Desde que a probabilidade de ocorrência de B exista e não seja nula, P($A|B$) é definida por

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \tag{3.3}$$

O diagrama de Venn, mostrado na Figura 3.4, ilustra a noção imposta pela equação 3.3. De fato, se o evento B já ocorreu, ou é certo de ocorrer, o espaço amostral deve ser reduzido para essa nova realidade e a probabilidade de ocorrência de A deve ser recalculada. As seguintes propriedades se aplicam à noção de probabilidade condicional:

i. Se $P(B) \neq 0$, então, para qualquer evento A, $0 \leq P(A|B) \leq 1$

ii. Se dois eventos A_1 e A_2 são disjuntos em B e se $P(B) \neq 0$, então $P(A_1 \cup A_2|B) = P(A_1|B) + P(A_2|B)$

iii. Como particularidade de (ii), segue-se que $P(A|B) + P(A|B^c) = 1$

iv. Se $P(B) \neq 0$, $P(A_1 \cup A_2|B) = P(A_1|B) + P(A_2|B) - P(A_1 \cap A_2|B)$

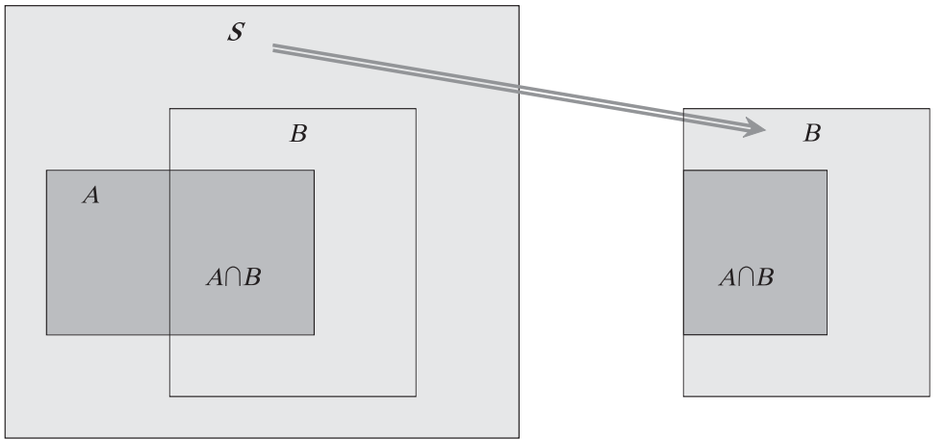


Figura 3.4 – Diagrama de Venn com ilustração do conceito de probabilidade condicional

A equação 3.3 pode ser re-escrita da forma $P(A \cap B) = P(B)\,P(A|B)$ e, como $P(A \cap B) = P(B \cap A)$, segue-se que $P(B \cap A) = P(A)\,P(B|A)$. Essa é a chamada *regra da multiplicação* que pode ser generalizada para o caso de mais de dois eventos; por exemplo, para três eventos, a regra da multiplicação é dada por

$$P(A \cap B \cap C) = P(A)\,P(B|A)\,P(C|A \cap B) \tag{3.4}$$

Se a probabilidade de ocorrência de *A* não é afetada pela ocorrência de *B* e vice-versa, ou seja, se P(*A*|*B*) = P(*A*) e P(*B*|*A*) = P(*B*), então esses eventos são considerados *estatisticamente independentes* e a regra da multiplicação torna-se

$$P(A \cap B) = P(B \cap A) = P(B)\,P(A) = P(A)\,P(B) \tag{3.5}$$

Generalizando, pode-se dizer que se existem *k* eventos mútua e coletivamente independentes em um espaço amostral, denotados por A_1, A_2, ... , A_k, a probabilidade de sua ocorrência simultânea é dada por

$$P(A_1 \cap A_2 \cap ... \cap A_k) = P(A_1)\,P(A_2)...\,P(A_k)$$

Exemplo 3.3 – Suponha que uma cidade, localizada a jusante da confluência de dois rios R_1 e R_2, sofre inundações devidas à ocorrência de enchentes em R_1 (evento A), ou em R_2 (evento B) ou em ambos. Se P(*A*) é o triplo de P(*B*), se P(A|B) = 0,6 e se a probabilidade da cidade sofrer inundações é

de 0,01, calcule (a) a probabilidade de ocorrência de enchentes no rio R_2 e (b) a probabilidade de ocorrência de enchentes *apenas* no rio R_1, dado que a cidade sofreu inundações.

Solução:

(a) A probabilidade da cidade sofrer inundações é dada por $P(A \cup B) = P(A) + P(B) - P(A \cap B)$. Daí,

$$P(A \cup B) = 3P(B) + P(B) - P(B)P(A|B) \Rightarrow$$

$$\Rightarrow 0{,}01 = 3P(B) + P(B) - 0{,}6P(B) \Rightarrow P(B) = 0{,}003 \text{ e } P(A) = 0{,}009$$

(b) A probabilidade de ocorrência de enchentes *apenas* no rio R_1, dado que a cidade sofreu inundações, pode ser escrita da seguinte forma:

$$P[(A \cap B^c) | (A \cup B)] = \frac{P[(A \cap B^c) \cap (A \cup B)]}{P(A \cup B)} = \frac{P[(A \cap B^c)]}{0{,}01} = \frac{P(A)[1 - P(B|A)]}{0{,}01}$$

Nessa equação, apenas a quantidade $P(B|A)$ é desconhecida, mas pode ser deduzida das probabilidades dadas por meio das relações

$$P(A)P(B|A) = P(B)P(A|B) \Rightarrow 3P(B)P(B|A) = P(B)P(A|B) \Rightarrow$$

$$\Rightarrow P(B|A) = P(A|B)/3 = 0{,}2.$$

Com esse valor na equação anterior, tem-se que $P[(A \cap B^c) | (A \cup B)] = 0{,}72$.

3.4 – Teoremas da Probabilidade Total e de Bayes

Suponha que o espaço amostral *S* de um certo experimento seja o resultado da união de *k* eventos mútua e coletivamente excludentes B_1, B_2, ..., B_k, cujas probabilidades de ocorrência são diferentes de zero. Considere também um evento *A*, tal como ilustrado na Figura 3.5, cuja probabilidade de ocorrência é $P(A) = P(B_1 \cap A) + P(B_2 \cap A) + ... + P(B_k \cap A)$. Usando a definição de probabilidade condicional em cada termo do segundo membro dessa equação, segue-se que

$$P(A) = P(B_1)P(A|B_1) + P(B_2)P(A|B_2) + ... + P(B_k)P(A|B_k) = \sum_{i=1}^{k} P(B_i)P(A|B_i) \quad (3.6)$$

A equação 3.6 é a expressão formal do chamado *teorema da probabilidade total*.

Exemplo 3.4 – O sistema de abastecimento de água de uma cidade é composto por dois reservatórios distintos e complementares: o de número 1 com volume de 150.000 l, cuja probabilidade de funcionamento é 0,7, e o de número 2, com 187.500 l, cuja probabilidade de ser usado é 0,3. A demanda diária de água para abastecimento da cidade é uma variável

aleatória cujas probabilidades de igualar ou superar 150.000 l e 187.500 l são respectivamente 0,3 e 0,1. Sabendo-se que quando um reservatório é ativado, o outro encontra-se desativado, pergunta-se: (a) qual é a probabilidade de não atendimento da demanda em um dia qualquer? e (b) supondo que as condições sejam tais que permitam a consideração de independência estatística dos eventos entre dois dias consecutivos, qual é a probabilidade de não atendimento da demanda em uma semana qualquer?
Solução: (a) Considere que o não atendimento da demanda em um dia qualquer seja representado pelo evento A, enquanto os eventos B e B^c denotam o funcionamento dos reservatórios 1 e 2. A aplicação da equação 3.6, com $k = 2$, resulta em

$$\mathrm{P}(A) = \mathrm{P}(A|B)\,\mathrm{P}(B) + \mathrm{P}(A|B^c)\,\mathrm{P}(B^c) = 0{,}3\times 0{,}7 + 0{,}1\times 0{,}3 = 0{,}24 \ .$$

(b) A probabilidade de não atendimento da demanda em uma semana qualquer equivale à probabilidade de haver pelo menos uma falha em 7 dias, a qual por sua vez é igual ao complemento da probabilidade de não haver nenhuma falha em uma semana, em relação a 1. Logo, a resposta é dada por $[1-(0{,}76)^7] = 0{,}8535$.

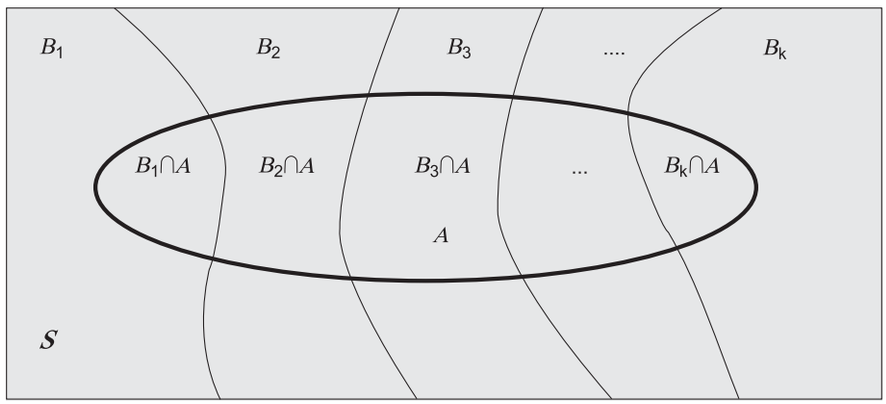


Figura 3.5 – Diagrama de Venn para o Teorema da Probabilidade Total.

O *teorema de Bayes*, devido ao matemático inglês Thomas Bayes (1702-1761), resulta de uma interessante combinação da regra da multiplicação e do teorema da probabilidade total. Considerando novamente a situação ilustrada pela Figura 3.5, podemos expressar a probabilidade de qualquer um dos eventos mutuamente excludentes, por exemplo, B_j, condicionada à ocorrência de A, por meio da equação

$$\mathrm{P}(B_j|A) = \frac{\mathrm{P}(B_j \cap A)}{\mathrm{P}(A)} \qquad (3.7)$$

Pela regra da multiplicação, o numerador do segundo membro da equação 3.7 pode ser expresso por $\mathrm{P}\left(A\middle|B_j\right)\mathrm{P}\left(B_j\right)$, enquanto o denominador pode ser posto na forma do teorema da probabilidade total. A equação resultante é a expressão do teorema de Bayes, a saber,

$$\mathrm{P}\left(B_i|A\right)=\frac{\mathrm{P}\left(A\middle|B_j\right)\mathrm{P}\left(B_j\right)}{\sum_{i=1}^{k}\mathrm{P}\left(A|B_i\right)\mathrm{P}\left(B_i\right)} \tag{3.8}$$

O teorema de *Bayes* constitui um quadro lógico importante para a revisão ou a atualização de probabilidades previamente estabelecidas, à luz de novas informações. Para exemplificar tal possibilidade, considere a necessidade hipotética de cálculo da probabilidade da temperatura mínima de um dia qualquer de Janeiro, em um dado local, estar acima de 15º C, como parte das informações contidas em um boletim de previsão meteorológica. Nesse caso, denotamos por B_1 o evento das temperaturas superiores a 15º C e por B_1^c o evento complementar, de tal modo que esses sejam mútua e coletivamente excludentes e que, portanto $B_1 \cup B_1^c = S$. Se nenhuma outra informação encontra-se disponível, é natural que se estime a probabilidade P(B_1) pela freqüência relativa dos dias de Janeiro com temperaturas superiores a 15º C, digamos (25/31) ou 80,64%. Dentro do contexto do teorema de Bayes, essa estimativa é denominada probabilidade *a priori* ou *subjetiva*, indicando o grau de confiança inicial que tem o meteorologista, referente à ocorrência de B_1. Entretanto, a temperatura mínima diária pode ser afetada pela ocorrência de precipitações naquele dia e, supondo que se preveja um dia chuvoso, tal cenário certamente irá modificar a probabilidade *a priori* P(B_1). Para incorporar tal modificação, é preciso conhecer as estimativas de P($A|B_1$) e P(A), respectivamente as probabilidades de ocorrer chuva nos dias com temperaturas superiores a 15º C e em todos os dias de Janeiro. Suponha que a análise de freqüência dos registros históricos produza as seguintes estimativas P($A|B_1$) = (15/25) e P(A) = (18/31). Com tais estimativas na equação 3.8 e lembrando que o denominador dessa equação é de fato P(A), tem-se P($B_1|A$) = [(15/25).(25/31)]/(18/31) = (15/18) ou 83,33%. Essa é a probabilidade *a posteriori*, revisada pela incorporação da ocorrência do evento *A*.

Exemplo 3.5 – Um satélite meteorológico envia um conjunto de códigos binários (‘0’ ou ‘1’) para descrever o desenvolvimento de uma tempestade. Entretanto, interferências diversas no sinal emitido pelo satélite podem provocar erros de transmissão. Suponha que uma certa mensagem binária, contendo 80% de dígitos ‘0’, tenha sido transmitida e que exista uma probabilidade de 85% de que um dado ‘0’ ou ‘1’ tenha sido recebido

corretamente. Se houve a recepção de um '1', qual é a probabilidade de que um '0' tenha sido transmitido?
Solução: Vamos representar os eventos de que o dígito '0' ou '1' tenha sido transmitido, respectivamente por T_0 ou T_1. Analogamente, R_0 ou R_1 denotam a recepção de um '0' ou de um '1', respectivamente. De acordo com os dados do problema, $P(T_0) = 0,8$, $P(T_1) = 0,2$, $P(R_0|T_0) = 0,85$, $P(R_1|T_1) = 0,85$, $P(R_0|T_1) = 0,15$ e $P(R_1|T_0) = 0,15$. A probabilidade pedida é $P(T_0|R_1)$, a qual pode ser calculada por meio do teorema de Bayes. No caso presente, $P(T_0|R_1) = [P(R_1|T_0)\,P(T_0)] / [P(R_1|T_0)\,P(T_0) + P(R_1|T_1)\,P(T_1)]$. Com os dados do problema, $P(T_0|R_1) = (0,15 \times 0,8)/(0,15 \times 0,8 + 0,85 \times 0,2) = 0,4138$.

3.5 – Variáveis Aleatórias

Uma *variável aleatória* é uma função X que associa um valor numérico a cada resultado de um experimento. Embora diferentes resultados do experimento possam compartilhar o mesmo valor associado a X, há um único valor numérico da variável aleatória, associado a cada resultado. Para facilitar o entendimento do conceito de variável aleatória, considere o lançamento simultâneo de duas moedas, distinguíveis uma da outra; o espaço amostral, correspondente a esse experimento, é S={ff, cc, fc, cf}, onde f simboliza 'face' (ou 'cara') e c 'coroa'. Por suposição, os eventos mutuamente excludentes A={ff}, B={cc}, C={fc} e D={cf} são considerados equiprováveis, cada qual, portanto, com probabilidade de ocorrência igual a 0,25. Suponha, ainda, que a variável aleatória X seja definida como o número de 'faces' (ou 'caras') decorrentes da realização do experimento. O mapeamento do espaço amostral S permite associar à variável X os seguintes possíveis valores numéricos: $x = 2$, $x = 1$ ou $x = 0$. Os valores extremos de X, quais sejam 0 e 2, estão, respectivamente, associados à ocorrência de A e B, enquanto $x = 1$ corresponde à união dos eventos C e D.

Além de associar as ocorrências possíveis aos valores da variável aleatória X, é preciso atribuir probabilidades a eles. Logo, $P(X=2) = P(A) = 0,25$, $P(X=0) = P(B) = 0,25$ e, finalmente, $P(X=1) = P(C \cup D) = P(C)+P(D) = 0,50$. Essas probabilidades são representadas genericamente por $p_X(x)$, equivalentes a $P(X = x)$, e ilustradas nos gráficos da Figura 3.6

No exemplo da Figura 3.6, a variável aleatória X é classificada como *discreta* porque ela pode assumir apenas valores numéricos inteiros e, também, por estar associada a um espaço amostral finito e numerável. Nesse caso, $p_X(x)$ representa a *função massa de probabilidades* (FMP) e indica com que probabilidade a variável X assume o valor do argumento x. Por outro lado, $P_X(x)$ denota a *função*

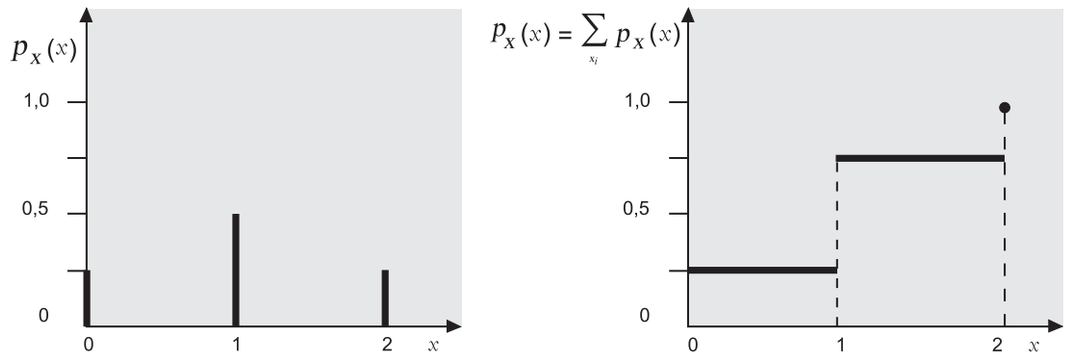


Figura 3.6 – Distribuições de probabilidade da variável aleatória *X*.

acumulada de probabilidades (FAP), ou função de distribuição de probabilidades, e indica com que probabilidade a variável *X* é menor ou igual ao argumento *x*, ou seja, $P_X(x) = \mathrm{P}(X \le x) = \sum_{\text{todos } x_i \le x} p_X(x_i)$. Uma função massa de probabilidades possui as seguintes propriedades:

i. $p_X(x) \ge 0$ *para todo e qualquer valor de x*

ii. $\sum_{\text{todos } x} p_X(x) = 1$

Inversamente, se uma função $p_X(x)$ possui essas propriedades, então ela pode ser considerada uma função massa de probabilidades. Por outro lado, se a variável aleatória *X* pode assumir qualquer valor real, ela é do tipo *contínuo* e, nesse caso, a função equivalente à FMP é denominada *função densidade de probabilidade* (FDP). Essa *função não negativa*, aqui denotada por $f_X(x)$ e ilustrada na Figura 3.7, representa o caso limite de um polígono de freqüências para uma amostra de tamanho infinito e, portanto, com as larguras dos intervalos de classe tendendo a zero. É importante notar que $f_X(x_0)$ não fornece a probabilidade de *X* para o argumento x_0 e, sim, a *intensidade* com que a probabilidade de não superação de x_0 é alterada na vizinhança do argumento indicado. A área entre dois limites *a* e *b*, no eixo dos argumentos da variável aleatória, dá a probabilidade de *X* estar compreendida no intervalo, tal como ilustrado na Figura 3.7. Portanto, para a FDP $f_X(x)$, é válida a equação

$$\mathrm{P}(a < X \le b) = \int_a^b f_X(x)\,dx \tag{3.9}$$

Se fizermos o limite inferior dessa integração se aproximar de *b*, a ponto de ambos se confundirem, o resultado seria equivalente à 'área de uma reta' no plano real que, por definição, é nula. Generalizando, pode-se concluir que para uma variável aleatória contínua X, $\mathrm{P}(X = x) = 0$.

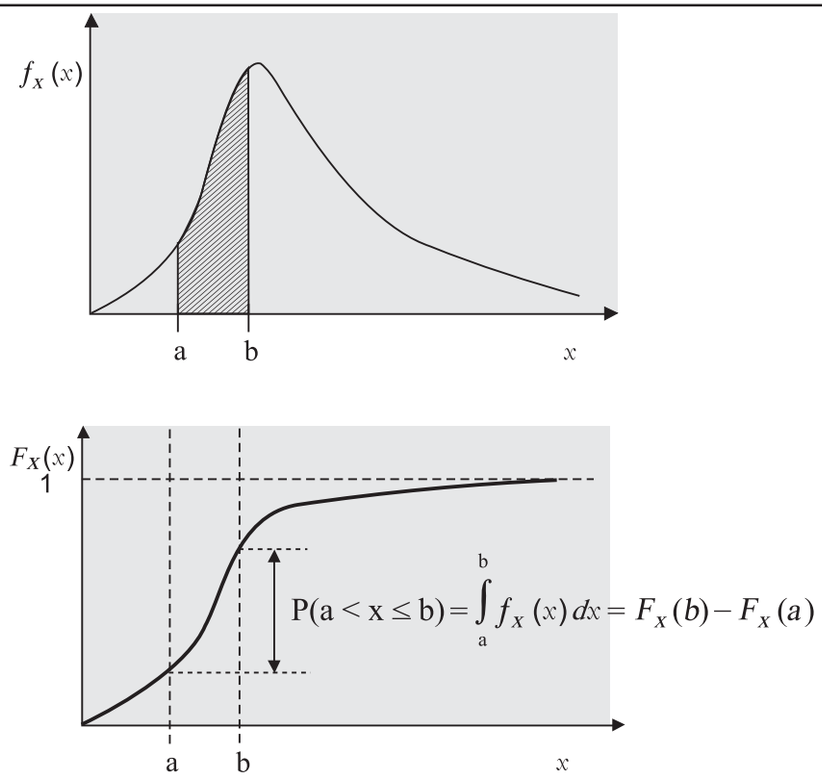


Figura 3.7 – Funções densidade e acumulada de probabilidades de uma variável contínua

Analogamente ao caso discreto, a *função acumulada de probabilidades* (FAP) de uma variável aleatória contínua X, aqui representada por $F_X(x)$, fornece a probabilidade de não superação do argumento x, ou seja, $\mathrm{P}(X \leq x)$ ou $\mathrm{P}(X < x)$. Formalmente,

$$F_X(x) = \int_{-\infty}^{x} f_X(x)\,dx \tag{3.10}$$

Inversamente, a FDP correspondente pode ser obtida pela diferenciação de $F_X(x)$, ou seja,

$$f_X(x) = \frac{d\,F_X(x)}{dx} \tag{3.11}$$

A FAP de uma variável aleatória contínua é uma função não decrescente, sendo válidas as expressões $F_X(-\infty) = 0$ e $F_X(+\infty) = 1$.

As funções massa e densidade de probabilidades, assim como suas correspondentes FAP's, descrevem completamente o comportamento estatístico das variáveis aleatórias discretas e contínuas, respectivamente. Em particular, a função densidade de probabilidade de uma variável contínua X pode ter uma grande variedade de formas, algumas delas ilustradas na Figura 3.8. Como requisito geral, para que se trate de uma densidade de probabilidade de uma variável contínua X pode ter uma grande variedade de formas, algumas delas ilustradas na Figura 3.8. Como requisito geral, para que se trate de uma densidade de probabilidades, a função deve ser não negativa e o resultado de sua integração, ao longo de todo o domínio de variação de X, deve ser igual a 1.

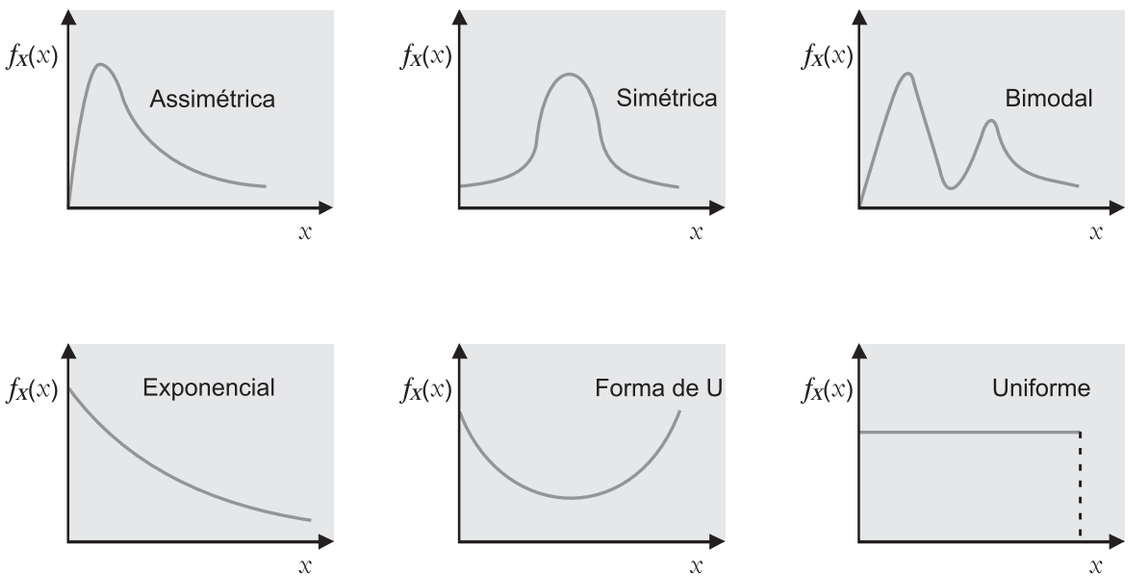


Figura 3.8 – Formas variadas de uma função densidade de probabilidades

Exemplo 3.6 – Considere que a variável aleatória 'vazão media diária máxima anual', em m^3/s, em uma certa estação fluviométrica, seja representada por X e que sua função densidade de probabilidade seja dada pela Figura 3.9. Pede-se (a) $P(X < 100\ m^3/s)$ e (b) $P(X > 300\ m^3/s)$.
Solução: (a) Se $f_X(x)$ é uma função densidade de probabilidades, a área de todo o triângulo deve ser igual a 1. Assim, (400y)/2 = 1, o que resulta em y = 1/200. Logo, $P(X < 100\ m^3/s)$, correspondente à área da do triângulo até a abscissa 100, é (100y)/2 = 0,25.

(b) P($X > 300$), ou [1- P($X < 300$)], corresponde à área do triângulo à direita da abscissa 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.

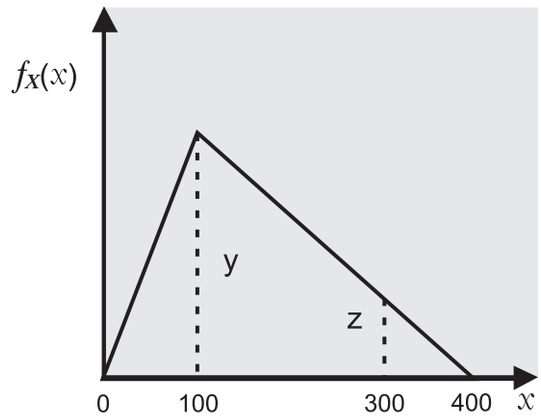


Figura 3.9 – Função Densidade de *X*

Exemplo 3.7 – A função definida por $f_x(x) = \frac{1}{\theta}\exp\left(-\frac{x}{\theta}\right)$, para $x \geq 0$ e $\theta \geq 0$, é a forma paramétrica que define a família exponencial de funções densidade de probabilidades, ou seja, uma FDP para cada valor numérico do parâmetro θ. Pede-se: (a) provar que, independentemente do valor de θ, trata-se de uma função densidade de probabilidade; (b) expressar a função acumulada F_X(x); (c) calcular P($X > 3$), para o caso de θ = 2 e (c) elaborar um gráfico de $f_x(x)$ e $F_x(x)$, versus x, para θ = 2.

Solução: (a) Uma vez que $x \geq 0$ e $\theta \geq 0$, trata-se de uma função não negativa. Em conseqüência, a condição necessária e suficiente para que $f_x(x)$ seja uma função densidade de probabilidades é

$\int_0^\infty \frac{1}{\theta} exp\left(-\frac{x}{\theta}\right) dx = 1$. A integral pode ser resolvida do seguinte modo:

$$\int_0^\infty \frac{1}{\theta} exp\left(-\frac{x}{\theta}\right) dx = -exp\left(-\frac{x}{\theta}\right)\Big]_0^\infty = 1$$

demonstrando, portanto, que se trata de uma FDP.

(b) $$F_X(x) = \int_0^x \frac{1}{\theta} exp\left(-\frac{x}{\theta}\right) dx = -exp\left(-\frac{x}{\theta}\right)\Big]_0^x = 1 - exp\left(-\frac{x}{\theta}\right)$$

(c) $P(X > 3) = 1\text{-}P(X < 3) = 1\text{-}\ F_X(3) = 1 - \left[1 - \exp\left(-\frac{3}{2}\right)\right] = 0{,}2231$

(d) Gráficos: Figura 3.10

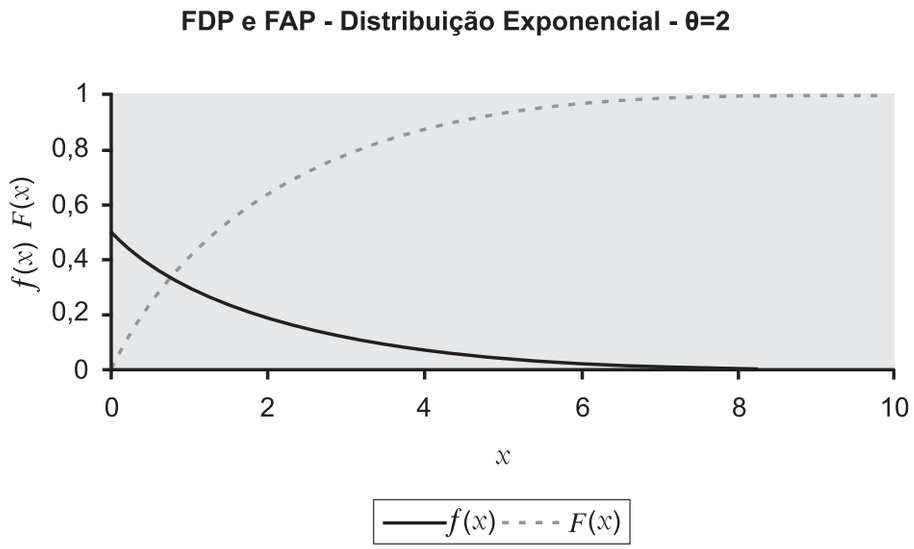


Figura 3.10 – FDP e FAP para a distribuição exponencial com parâmetro θ=2

3.6 – Medidas Descritivas Populacionais de Variáveis Aleatórias

A população de uma variável aleatória *X* é integralmente conhecida, sob o ponto de vista estatístico, pela completa especificação da função massa de probabilidades $p_X(x)$, no caso discreto, ou da função densidade de probabilidades $f_X(x)$, no caso contínuo. Analogamente às estatísticas descritivas de uma amostra extraída da população, objeto do capítulo 2, as características de forma das funções $p_X(x)$ ou $f_X(x)$ podem ser sumariadas por *medidas descritivas populacionais*. Essas são obtidas por meio de médias, ponderadas por $p_X(x)$ ou $f_X(x)$, de funções da variável aleatória e incluem o valor esperado, a variância, os coeficientes de assimetria e de curtose, entre outros.

3.6.1 – Valor Esperado

O *valor esperado* de *X* é o resultado da ponderação por $p_X(x)$, ou $f_X(x)$, dos valores possíveis da variável aleatória. O valor esperado, denotado por $E[X]$,

equivale à média populacional μ_X, indicando, portanto, a abscissa do centróide das funções $p_X(x)$ ou $f_X(x)$. A definição formal de $E[X]$ é dada por

$$E[X] = \mu_X = \sum_{\text{todos } x_i} x_i \; p_X(x_i) \tag{3.12}$$

para o caso discreto; e

$$E[X] = \mu_X = \int_{-\infty}^{+\infty} x \, f_X(x) \, dx \tag{3.13}$$

para o caso contínuo.

Exemplo 3.8 – Calcule o valor esperado para a função massa de probabilidades especificada pela Figura 3.6.
Solução: A aplicação da equação 3.12 resulta em $E[X] = \mu_X = 0\times0{,}25+1\times0{,}5+2\times0{,}25 = 1$ que, de fato, é a abscissa do centróide da função massa de probabilidades.

Exemplo 3.9 – Considere uma variável aleatória exponencial X, cuja função densidade de probabilidade é dada por $f_X(x) = \frac{1}{\theta}\exp\left(-\frac{x}{\theta}\right)$, para $x \geq 0$ e $\theta \geq 0$, tal como no Exemplo 3.7. Pede-se (a) calcular o valor esperado de X e (b) empregando somente as medidas populacionais de tendência central, a saber, a média, a moda e a mediana, comprovar que se trata de uma distribuição com assimetria positiva.
Solução: (a) Para a distribuição em questão,

$$E[X] = \mu_X = \int_0^{+\infty} x \, f_X(x) \, dx = \int_0^{+\infty} \frac{x}{\theta} exp\left(-\frac{x}{\theta}\right) dx \, .$$

Essa integração deve ser resolvida por partes, ou seja, faz-se

$dv = \frac{1}{\theta} exp\left(-\frac{x}{\theta}\right) dx \Rightarrow v = -exp\left(-\frac{x}{\theta}\right)$ e $u = x \Rightarrow du = dx$. Na seqüência,

$$\int_0^{\infty} u \, dv = uv\Big]_0^{\infty} - \int_0^{\infty} v \, du = -x \, exp\left(-\frac{x}{\theta}\right)\Big]_0^{\infty} - \theta \, exp\left(-\frac{x}{\theta}\right)\Big]_0^{\infty} = \theta \cdot$$

Portanto, para a forma paramétrica exponencial, a média populacional é dada pelo parâmetro θ; para outras formas paramétricas, μ_X é, em geral, uma função simples dos parâmetros que especificam a distribuição. No caso de $\theta = 2$ (ver gráficos do Exemplo 3.7), a abscissa do centróide da FDP é $x = 2$.

(b) A média μ_X de uma variável exponencial é θ, portanto, um número positivo. A moda m_X é o valor da variável correspondente à maior ordenada da FDP e, portanto, no caso de uma variável exponencial $m_X = 0$. A mediana u_X corresponde ao valor x para o qual $F_X(x) = 0,5$. Como, nesse caso, $F_X(x) = 1 - exp\left(-\frac{x}{\theta}\right)$ (ver Exemplo 3.7), a função inversa de $F_X(x)$, também denominada *curva de quantis*, é dada por $x = -\theta\, ln(1-F)$. Para $F_X(x) = 0,5$, $u_X = -\theta\, ln(1-0,5) = 0,6932\,\theta$. Logo, pode-se concluir que $m_X < u_X < \mu_X$, o que caracteriza uma distribuição assimétrica positivamente. De fato, como será visto na seqüência do presente item, o coeficiente de assimetria da distribuição exponencial é igual a +2.

Pode-se generalizar a idéia de valor esperado para uma função $g(X)$ da variável aleatória X, ou seja, usar a ponderação de $p_X(x)$ ou $f_X(x)$ para calcular a chamada *esperança matemática* de $g(X)$ ou, simbolicamente, $E[g(X)]$. Em termos formais,

$$E[g(X)] = \sum_{\text{todos } x_i} g(x_i)\, p_X(x_i) \tag{3.14}$$

para uma variável aleatória discreta. No caso contínuo, $E[g(X)]$ é definido por

$$E[g(X)] = \int_{-\infty}^{+\infty} g(x)\, f_X(x)\, dx \tag{3.15}$$

Na equação 3.15, observa-se o requisito de que $E[g(X)]$ existe desde que a integral seja convergente. O operador esperança matemática apresenta as seguintes propriedades:

i. $E[c] = c$, para c constante.

ii. $E[cg(X)] = cE[g(X)]$, para c constante.

iii. $E[c_1 g_1(X) \pm c_2 g_2(X)] = c_1 E[g_1(X)] \pm c_2 E[g_2(X)]$, para c_1 e c_2 constantes, e funções $g_1(X)$ e $g_2(X)$.

iv. $E[g_1(X)] \geq E[g_2(X)]$, se $g_1(X) \geq g_2(X)$.

Exemplo 3.10 – A esperança matemática $E[X - \mu_X]$ é denominado momento central de ordem 1 e corresponde à média das distâncias de x, em relação à média μ_X, ponderada pela FDP ou pela FMP de X. Use as propriedades do operador esperança matemática para mostrar que é nulo o momento central de ordem 1.
Solução: $E[X - \mu_X] = E[X] - E[\mu_X]$. Como μ_X é uma constante, conclui-se que $E[X - \mu_X] = \mu_X - \mu_X = 0$.

A aplicação do operador esperança matemática a potências de ordem *k* das distâncias da variável aleatória *X*, em relação a uma posição de referência *a*, ou seja $E\left[(X-a)^k\right]$, dá origem ao conceito de *momento de ordem k*. Dois casos se destacam: (i) se a posição de referência *a* é igual a zero, os momentos são ditos em relação à origem e denotados por μ_X, se $k=1$ e μ_k', se $k \geq 2$; e (ii) se $a = \mu_X$, os momentos são denominados centrais e representados por μ_k. Os *momentos em relação à origem* são formalmente definidos por

$$\mu_X = E[X] \text{ e } \mu_k' = \sum_{\text{todos } x_i} x^k \, p_X(x_i) \tag{3.16}$$

se a variável aleatória é discreta. No caso de variável contínua,

$$\mu_X = E[X] \text{ e } \mu_k' = \int_{-\infty}^{+\infty} x^k f_X(x)\,dx \tag{3.17}$$

Paralelamente, os momentos centrais são dados por

$$\mu_1 = 0 \text{ e } \mu_k = \sum_{\text{todos } x_i} (x-\mu_X)^k \, p_X(x_i), \text{ se } k \geq 2 \tag{3.18}$$

se *X* é discreta; caso seja contínua,

$$\mu_1 = 0 \text{ e } \mu_k = \int_{-\infty}^{+\infty} (x-\mu_X)^k f_X(x)\,dx, k \geq 2 \tag{3.19}$$

Essas grandezas são denominadas momentos, em analogia aos momentos da mecânica. Em particular, μ_X corresponde à abscissa do centróide da FMP ou FDP, de modo análogo à abscissa do centro de massa de um corpo sólido, enquanto μ_2 equivale ao momento de inércia em relação a um eixo vertical que passa pelo centróide.

3.6.2 – Variância Populacional

A *variância populacional* de uma variável aleatória *X*, representada por Var[*X*] ou σ_X^2, é definida como sendo o momento central de segunda ordem, ou μ_2, e corresponde à medida populacional mais freqüentemente empregada para caracterizar a dispersão das funções $p_X(x)$ ou $f_X(x)$. Portanto, Var[*X*], também denotada por σ_X^2, é dada por

$$\text{Var}[X] = \sigma_X^2 = \mu_2 = E\left[(X-\mu_X)^2\right] = E\left[(X-E[X])^2\right] \tag{3.20}$$

Expandindo o quadrado contido nessa equação e usando as propriedades do operador esperança matemática, pode-se reescrevê-la como

$$\mathrm{Var}[X] = \sigma_X^2 = \mu_2 = E[X^2] - (E[X])^2 \tag{3.21}$$

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

i. $\mathrm{Var}[c] = 0$, para c constante.

ii. $\mathrm{Var}[cX] = c^2\,\mathrm{Var}[X]$.

iii. $\mathrm{Var}[cX+d] = c^2\,\mathrm{Var}[X]$, para d constante.

De modo análogo às estatísticas descritivas amostrais, define-se o *desvio padrão populacional* σ_X como a raiz quadrada positiva da variância, 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 pela expressão

$$\mathrm{CV}_X = \frac{\sigma_X}{\mu_X} \tag{3.22}$$

Exemplo 3.11 – Calcule a variância, o desvio padrão e o coeficiente de variação para a função massa de probabilidades especificada pela Figura 3.6.
Solução: A aplicação da equação 3.21 requer o cálculo de $E[X^2]$. Portanto, calculando tal grandeza, $E[X^2] = 0^2\times0{,}25+1^2\times0{,}5+2^2\times0{,}25 = 1{,}5$. De volta à equação 3.21, $\mathrm{Var}[X] = \sigma^2_X = 1{,}5-1^2 = 0{,}5$. O desvio padrão, portanto, é $\sigma_X = 0{,}71$ e o coeficiente de variação é $\mathrm{CV}_X = 0{,}71/1 = 0{,}71$.

Exemplo 3.12 - Considere a variável aleatória exponencial X, tal como no Exemplo 3.9. Calcule a variância, o desvio padrão e o coeficiente de variação de X.
Solução: O valor esperado de uma variável exponencial é θ (ver exemplo 3.9). Novamente, a aplicação da equação 3.21 requer o conhecimento de $E[X^2]$. Por definição, $E[X^2] = \int_0^{+\infty} x^2 f_X(x)\,dx = \int_0^{+\infty} \frac{x^2}{\theta} exp\left(-\frac{x}{\theta}\right) dx$, a qual, mais uma vez, pode ser resolvida por partes, ou seja, faz-se, $dv = \frac{x}{\theta} exp\left(-\frac{x}{\theta}\right) dx \Rightarrow v = -x\,exp\left(-\frac{x}{\theta}\right) - \theta\, exp\left(-\frac{x}{\theta}\right)$, tal como no exemplo 3.9, e $u = x \Rightarrow du = dx$. Na seqüência,

$$\int_0^\infty u\,dv = uv\Big]_0^\infty - \int_0^\infty v\,du = 0 - \int_0^\infty \left[-x\,exp\left(-\frac{x}{\theta}\right) - \theta\,exp\left(-\frac{x}{\theta}\right)\right] dx = \theta\,E[X] + \theta^2 = 2\,\theta^2$$

De volta à equação 3.21, verifica-se que Var[X]=2 θ^2 - θ^2 = θ^2. Portanto, $\sigma = \theta$ e CV_X=1.

3.6.3 – Coeficientes de Assimetria e Curtose Populacionais

O coeficiente de assimetria de uma variável aleatória X é um número adimensional definido por

$$\gamma = \frac{\mu^3}{(\sigma_X)^3} = \frac{E\left[(X-\mu_X)^3\right]}{(\sigma_X)^3} \qquad (3.23)$$

O numerador do segundo membro da equação 3.23, ou seja, o momento central de ordem 3 reflete a equivalência ou, contrariamente, a predominância dos desvios positivos ou negativos da variável aleatória X, em relação à média μ_X. Se houver equivalência, o numerador e o coeficiente de assimetria serão nulos e a função densidade de probabilidades será simétrica. Entretanto, se a cauda superior da FDP, ou seja, se os valores de X, superiores à média μ_X, estiverem muito mais dispersos do que os inferiores, os cubos dos desvios positivos irão prevalecer sobre os negativos e o coeficiente será positivo, configurando uma função densidade assimétrica positivamente. Caso contrário, teremos uma função densidade de probabilidade assimétrica negativamente. A Figura 3.11 ilustra três funções densidades de probabilidades: uma com coeficiente de assimetria nulo, uma assimétrica positivamente com γ = 1,14 e outra assimétrica negativamente com γ = -1,14.

O *coeficiente de curtose* κ de uma variável aleatória X é uma medida de quão pontiaguda é $p_X(x)$ ou $f_X(x)$. Esse coeficiente adimensional estabelece também uma medida relativa do peso das caudas superior e inferior das distribuições de probabilidade. É definido pela seguinte equação:

$$\kappa = \frac{\mu_4}{(\sigma_X)^4} = \frac{E\left[(X-\mu_X)^4\right]}{(\sigma_X)^4} \qquad (3.24)$$

Para distribuições simétricas, define-se o *coeficiente de excesso* de curtose (κ - 3) para estabelecer uma medida em relação a uma distribuição perfeitamente simétrica de referência, cujo valor de κ é 3.

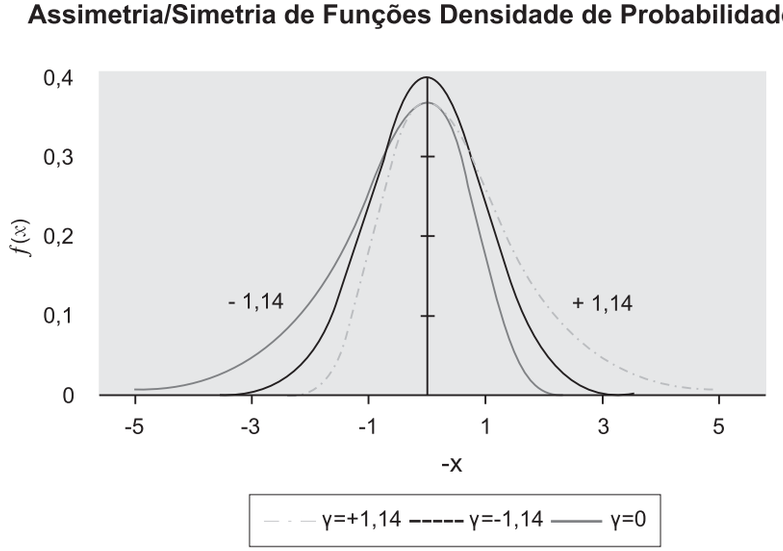


Figura 3.11 – Funções densidade de probabilidades simétricas e assimétricas

Exemplo 3.13 - Considere a variável aleatória exponencial *X*, tal como no Exemplo 3.9. Calcule os coeficientes de assimetria e curtose de *X*.
Solução: Prosseguindo com as integrações por partes efetuadas para o cálculo de $E[X]$ e de $E[X^2]$, tal como nos exemplos 3.9 e 3.12, é possível concluir que, para qualquer inteiro *k*, é válida a seguinte expressão:

$E\left[X^k\right]=\int_0^{\infty}\frac{x^k}{\theta}exp\left(-\frac{x}{\theta}\right)dx=\theta^k\Gamma(k+1)$, na qual $\Gamma(.)$ denota a função Gama (ver Anexo 4 para uma breve revisão). Se o argumento da função Gama é inteiro, é válida a propriedade $\Gamma(k+1)=k!$. Aplicando esse resultado aos momentos em relação à origem de ordens 3 e 4, segue-se que $E\left[X^3\right]=6\theta^3$ *e* $E\left[X^4\right]=24\theta^4$. Para o cálculo do coeficiente de assimetria, deve-se, de início, expandir o cubo no numerador do segundo membro da equação 3.23, para, em seguida, usar as propriedades do operador esperança matemática e obter a expressão

$\gamma=\frac{E\left[X^3\right]-3E\left[X^2\right]E[X]+2(E[X])^3}{(\sigma_X)^3}$. Substituindo os momentos já calculados, resulta $\gamma=2$. Do mesmo modo, o coeficiente de curtose pode ser expresso por $\kappa=\frac{E\left[X^4\right]-4E\left[X^3\right]E[X]+6E\left[X^2\right](E[X])^2-3(E[X])^4}{(\sigma_X)^4}$.

Com os momentos já calculados, $\kappa=9$.

3.6.4 – Função Geratriz de Momentos

O comportamento estatístico de uma variável aleatória é completamente especificado por sua função massa (ou densidade) de probabilidades, a qual, por sua vez, pode ser determinada por um certo número de momentos, suficientes para particularizar sua forma. A *função geratriz de momentos* de uma distribuição de probabilidades é uma função $\phi(t)$, *do argumento t* definido no intervalo $(-\varepsilon, \varepsilon)$ em torno de $t = 0$, que permite o cálculo alternativo de seus momentos em relação à origem, de ordem genérica $k \geq 1$. Para uma variável aleatória X, a função $\phi(t)$ é definida por

$$\phi(t) = E\left[e^{tX}\right] = \begin{cases} \sum_{todos\ x} e^{tx} p_X(x) \text{ se } X \text{ é discreta} \\ \int_{-\infty}^{\infty} e^{tx} f_X(x)dx \text{ se } X \text{ é contínua} \end{cases} \quad (3.25)$$

A função $\phi(t)$ é chamada geratriz de momentos porque sua k-ésima derivada em relação a t, calculada no ponto $t = 0$, fornece o momento μ'_k da distribuição massa (ou densidade) de probabilidades em questão.
Por exemplo, supondo que $k = 1$, tem-se

$$\phi'(t) = \frac{d}{dt} E\left[e^{tX}\right] = E\left[\frac{d\, e^{tX}}{dt}\right] = E\left[Xe^{tX}\right] \Rightarrow \phi'(t=0) = E[X] = \mu_X \quad (3.26)$$

Do mesmo modo, pode-se concluir que $\phi''(0) = E[X^2] = \mu'_2$, $\phi'''(0) = E[X^3] = \mu'_3$ e assim sucessivamente até $\phi^k(0) = E[X^k] = \mu'_k$. De fato, a expansão da função geratriz de momentos $\phi(t)$, de uma variável aleatória X, em uma série de Maclaurin (ver Anexo 4) de potências inteiras de t, produz

$$\phi(t) = E\left[e^{tX}\right] = E\left[1 + Xt + \frac{1}{2!}(Xt)^2 + \ldots\right] = 1 + \mu'_1 t + \frac{1}{2!}\mu'_2 t^2 + \ldots \quad (3.27)$$

Exemplo 3.14 – A função massa de probabilidade $p_X(x) = e^{-\nu}\frac{\nu^x}{x!}, x = 0,1,\ldots$

é conhecida como distribuição de Poisson,com parâmetro $\nu > 0$. Use a função geratriz de momentos para calcular a média e a variância de uma variável aleatória discreta de Poisson.
Solução: A equação 3.25, aplicada à FMP dada, resulta em

$$\phi(t) = E\left[e^{tX}\right] = \sum_{x=0}^{\infty} \frac{e^{tx} e^{-\nu}\, \nu^x}{x!} = e^{-\nu} \sum_{x=0}^{\infty} \frac{\left(\nu e^t\right)}{x!}$$. Usando a identidade

$\sum_{k=0}^{\infty}\frac{a^k}{k!}=e^a$, escreve-se $\phi(t)=e^{-\nu}e^{\nu\exp(t)}=\exp\left[\nu\left(e^t-1\right)\right]$. Derivando em relação a t, $\phi'(t)=\nu e^t\, exp\left[\nu\left(e^t-1\right)\right]$ e

$\phi''(t)=\left(\nu e^t\right)^2 exp\left[\nu\left(e^t-1\right)\right]+\nu e^t\, exp\left[\nu\left(e^t-1\right)\right]$. Para $t = 0$,

$E[X]=\phi'(0)=\nu$ e $E\left[X^2\right]=\phi''(0)=\nu^2+\nu$. Lembrando que $\mathrm{Var}(X)=E[X^2]-(E[X])^2$, conclui-se que $\mu_X=\mathrm{Var}(X)=\nu$.

Exemplo 3.15 – A distribuição normal é a mais conhecida e uma das mais úteis na construção do raciocínio estatístico. Sua função densidade de probabilidade é dada por $f_X(x)=\frac{1}{\sqrt{2\pi}\theta_2}\exp\left[-\frac{1}{2}\left(\frac{x-\theta_1}{\theta_2}\right)^2\right]$, na qual θ_1 e θ_2 são parâmetros que definem, respectivamente, a posição e a escala de variação da variável X, cuja amplitude é de $-\infty$ a $+\infty$. Após substituição e desenvolvimento, a função geratriz de momentos para essa distribuição pode ser expressa por

$$\phi(t)=E\left[e^{tX}\right]=\frac{1}{\sqrt{2\pi}\theta_2}\int_{-\infty}^{\infty}\exp\left[-\frac{x^2-2\theta_1x+\theta_2^2-2\theta_2^2tx}{2\theta_2^2}\right]dx \quad .$$

Calcule μ_X e $\mathrm{Var}(X)$ de uma variável Normal.

Solução: Na expressão da função $\phi(t)$, pode-se reescrever

$x^2-2\theta_1x+\theta_2^2-2\theta_2^2tx=x^2-2\left(\theta_1+\theta_2^2t\right)x+\theta_2^2$

O segundo membro não irá ser alterado pelo artifício

$\left[x-\left(\theta_1+\theta_2^2t\right)\right]^2-\left(\theta_1+\theta_2^2t\right)^2+\theta_1^2=\left[x-\left(\theta_1+\theta_2^2t\right)\right]^2-\theta_2^4t^2-2\theta_1\theta_2^2t$

De volta a $\phi(t)$, tem-se

$$\phi(t)=\exp\left[\frac{\theta_2^4t^2+2\theta_1\theta_2^2t}{2\theta_2^2}\right]\frac{1}{\sqrt{2\pi}\theta_2}\int_{-\infty}^{\infty}\exp\left[-\frac{\left[x-\left(\theta_1+\theta_2^2t\right)\right]^2}{2\theta_2^2}\right]dx$$

Agora, podemos definir uma nova variável dada por $Y=\frac{x-\left(\theta_1+\theta_2^2t\right)^2}{\theta_2^2}$

a qual também é normalmente distribuída, porém com parâmetros $\theta_1+\theta_2^2t$ e θ_2

Nesse caso, $\frac{1}{\sqrt{2\pi}\theta_2}\int_{-\infty}^{\infty}\exp\left[-\frac{\left[x-\left(\theta_1+\theta_2^2t\right)\right]^2}{2\theta_2^2}\right]dx=1$ e

$$\phi(t)=\exp\left[\frac{\theta_2^4t^2+2\theta_1\theta_2^2t}{2\theta_2^2}\right]$$

As derivadas de $\phi(t)$ são

$$\phi'(t)=\left(\theta_1+t\theta_2^2\right)exp\left[\frac{\theta_2^2t^2}{2}+\theta_1t\right] \text{ e } \phi''(t)=\left(\theta_1+t\theta_2^2\right)^2\ exp\left[\frac{\theta_2^2t^2}{2}+\theta_1t\right]+\theta_2^2\ exp\left[\frac{\theta_2^2t^2}{2}+\theta_1t\right]$$

No ponto $t = 0$,

$$\phi'(0)=\theta_1 \Rightarrow E[X]=\theta_1 \text{ e } \phi''(0)=\theta_1^2+\theta_2^2 \Rightarrow E\left[X^2\right]=\theta_1^2+\theta_2^2$$

Lembrando que $\text{Var}(X) = E[X^2] - (E[X])^2$, conclui-se que $\mu_X = \theta_1$ e $\text{Var}(X)=\sigma_X^2=\theta_2$. Em decorrência desses resultados, a função densidade da distribuição normal é geralmente expressa por:

$$f_X(x)=\frac{1}{\sqrt{2\pi}\sigma_X}\exp\left[-\frac{1}{2}\left(\frac{x-\mu_X}{\sigma_X}\right)^2\right]$$

3.7 – Distribuições de Probabilidades Conjuntas de Variáveis Aleatórias

Até esse ponto, lidamos com as principais características das distribuições de probabilidades de uma única variável aleatória. Entretanto, são diversas as ocasiões em que o interesse se volta para a descrição probabilística do comportamento conjunto de duas ou mais variáveis aleatórias. As argumentações expostas para uma única variável aleatória serão aqui estendidas apenas para o caso bivariado. Supondo, portanto, que X e Y representem duas variáveis aleatórias, define-se a *função de distribuição acumulada de probabilidades conjuntas* de tais variáveis por meio de

$$\left.\begin{matrix}F_{X,Y}(x,y)\\P_{X,Y}(x,y)\end{matrix}\right\} = \text{P}(X \le x, Y \le y) \tag{3.28}$$

É possível deduzir a distribuição que descreve o comportamento de somente uma das variáveis, a partir de $F_{X,Y}(x,y)$ ou de $P_{X,Y}(x,y)$. Com efeito, *no caso contínuo*, a distribuição acumulada de probabilidades de X é definida por

$$F_X(x) = \text{P}(X \le x) = \text{P}(X \le x, Y \le \infty) = F_{X,Y}(x,\infty) \tag{3.29}$$

Similarmente para Y,

$$F_Y(x) = \text{P}(Y \le y) = \text{P}(X \le \infty, Y \le y) = F_{X,Y}(\infty,x) \tag{3.30}$$

$F_X(x)$ e $F_Y(y)$ são denominadas *distribuições marginais* de X e Y, respectivamente.

Se as variáveis X e Y são contínuas, define-se a *função densidade de probabilidades conjuntas* pela expressão

$$f_{X,Y}(x,y) = \frac{\partial^2}{\partial x\,\partial y} F_{X,Y}(x,y) \tag{3.31}$$

A Figura 3.12 ilustra a função densidade de probabilidades conjuntas das variáveis X e Y.

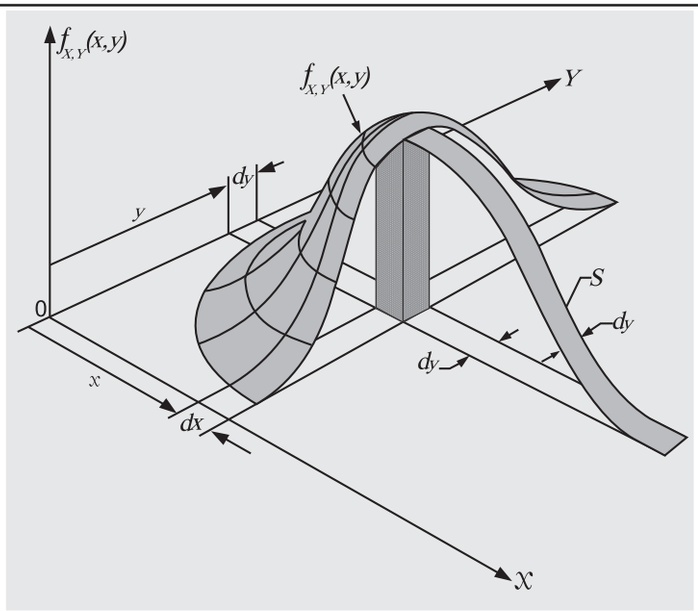


Figura 3.12 – Perspectiva de uma função densidade de probabilidade conjunta bivariada (adap. de Beckmann, 1968)

Como para qualquer função densidade de probabilidades, $f_{X,Y}(x,y)$ deve ser não negativa. Da mesma forma, o volume compreendido entre sua superfície e o plano XY deve ser igual a 1, ou seja,

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dx\,dy = 1 \tag{3.32}$$

A função *densidade marginal* de X pode ser obtida pela projeção da distribuição conjunta no plano formado pelo eixo vertical e o eixo dos X. Formalmente,

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dy \tag{3.33}$$

Do mesmo modo, a função densidade marginal de Y, ou seja, aquela que descreve apenas o comportamento isolado de Y, sem levar em conta a variação de X, pode ser deduzida da densidade conjunta por

$$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dx \tag{3.34}$$

Como decorrência, pode-se escrever

$$F_X(\infty) = \int_{-\infty}^{\infty} f_X(x)dx = 1 \text{ e } F_Y(\infty) = \int_{-\infty}^{\infty} f_Y(y)dy = 1 \tag{3.35}$$

$$F_X(x) = \int_{-\infty}^{x} f_X(x)dx = \mathrm{P}(X \le x) \text{ e } F_Y(\infty) = \int_{-\infty}^{y} f_Y(y)dy = \mathrm{P}(Y \le y) \tag{3.36}$$

Essa mesma lógica pode ser estendida para as funções massa de probabilidades, conjunta e marginais, das variáveis aleatórias discretas X e Y. Portanto, são válidas as seguintes relações:

$$P_{X,Y}(x,y) = \mathrm{P}(X \le x, Y \le y) = \sum_{x_i \le x} \sum_{y_j \le y} p_{X,Y}(x_i, y_j) \tag{3.37}$$

$$p_X(x_i) = \mathrm{P}(X = x_i) = \sum_j p_{X,Y}(x_i, y_j) \tag{3.38}$$

$$p_Y(y_j) = \mathrm{P}(Y = y_j) = \sum_i p_{X,Y}(x_i, y_j) \tag{3.39}$$

$$P_X(x) = \mathrm{P}(X \le x_i) = \sum_{x_i \le x} p_X(x_i) = \sum_{x_i \le x} \sum_j p_{X,Y}(x_i, y_j) \tag{3.40}$$

$$P_Y(y) = \mathrm{P}(Y \le y_j) = \sum_{y_j \le y} p_Y(y_j) = \sum_{y_j \le y} \sum_i p_{X,Y}(x_i, y_j) \tag{3.41}$$

Exemplo 3.16 – Suponha que

$f_{X,Y}(x,y) = 2x\exp(-x^2 - y)$ para $x \ge 0$ e $y \ge 0$.

Pergunta-se (a) se $f_{X,Y}(x,y)$ é, de fato, uma função densidade de probabilidade e (b) calcule P($X > 0{,}5$, $Y > 1$).

Solução:

(a) Como a função $f_{X,Y}(x,y)$ é sempre não negativa, resta verificar a condição imposta pela equação 3.32. Portanto,

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,Y}(x,y)dx\,dy = 2\int_0^{\infty} x\,exp(-x^2)dx \int_0^{\infty} exp(-y)dy =$$

$$= -\exp(-x^2)\Big|_0^{\infty}\left(-e^{-y}\right)\Big|_0^{\infty} = 1.$$ Logo, $f_{X,Y}(x,y)$ é uma densidade.

(b) $\mathrm{P}(X > 0{,}5,\ Y > 1) = \int_{0,5}^{\infty} 2x\exp(-x^2)dx \int_1^{\infty} \exp(-y)dy = \exp(-1{,}25) = 0{,}2865.$

A distribuição de uma das variáveis, *com restrições impostas à outra variável*, é denominada *distribuição condicional*. Para o caso de variáveis aleatórias discretas, a função massa de probabilidade de X, condicionada à ocorrência $Y = y_0$, é uma decorrência direta da definição de probabilidade condicionada, dada pela equação 3.3, ou seja,

$$p_{X|Y=y_0} = \frac{p_{X,Y}(x, y_0)}{p_Y(y_0)} \tag{3.42}$$

Para o caso de variáveis aleatórias contínuas, o conceito de distribuição condicional requer maior atenção. Para melhor explicar tal conceito, considere os eventos $x < X < x + dx$, denotado por A, e $y < Y < y + dy$, representado por B. A *função densidade de probabilidade condicional* $f_{X|Y}(x|y)$, *multiplicada por dx*, é equivalente à probabilidade condicional P($A|B$), ou seja,

$$f_{X|Y}(x|y)dx = \mathrm{P}(x < X < x + dx | y < Y < y + dy) = \mathrm{P}(A|B) \tag{3.43}$$

Note que, nesse caso, somente X é uma variável aleatória, uma vez que Y permaneceu fixa e contida no intervalo (y, y+dy), demonstrando que $f_{X|Y}(x|y)$ é unidimensional. Ora, se, por decorrência da equação 3.3, a probabilidade da ocorrência conjunta dos eventos A e B é dada por $\mathrm{P}(A \cap B) = \mathrm{P}(A|B)\mathrm{P}(B) = f_{X,Y}(x, y)dx\,dy$ e se, $\mathrm{P}(B) = \mathrm{P}(y < Y < y + dy) = f_Y(y)dy$, então, define-se a função densidade condicional $f_{X|Y}(x|y)$ por

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x, y)}{f_Y(y)} \tag{3.44}$$

sendo válidas as mesmas propriedades de qualquer função densidade de probabilidades. Usando o mesmo raciocínio anterior e o teorema da probabilidade total, é fácil demonstrar que o teorema de Bayes, quando aplicado a variáveis aleatórias contínuas, reduz-se a

$$f_{X|Y}(x|y) = \frac{f_{Y|X}(y|x)f_X(x)}{f_Y(y)} \text{ ou } f_{X|Y}(x|y) = \frac{f_{Y|X}(y|x)f_X(x)}{\int_{-\infty}^{\infty} f_{Y|X}(y|x)f_X(x)dx} \tag{3.45}$$

Com referência à Figura 3.12 e à luz das novas definições, pode-se interpretar a equação 3.44 como o quociente entre o volume do prisma $f_{X,Y}(x,y).dx.dy$, hachurado na figura, e o volume da faixa S contida pela superfície $f_{X,Y}(x,y)$ e o intervalo (y, $y + dy$). Entretanto, existe também o caso especial em que X e Y são variáveis aleatórias contínuas e que se quer conhecer a função densidade condicional de X, dado que $Y = y_0$; nesse caso, Y é um valor fixo, a faixa S passa a ser uma fatia plana da superfície de $f_{X,Y}(x,y)$ e, portanto, ter uma área e não um volume. A equação 3.44, para $Y = y_0$, pode ser reescrita como

$$f_{X|Y}(x|Y = y_0) = \frac{f_{X,Y}(x, y_0)}{f_Y(y_0)} \tag{3.46}$$

Em decorrência da equação 3.5, as variáveis aleatórias X e Y são *estatisticamente independentes* se a probabilidade de ocorrência de determinada realização de uma delas não é afetada pelo comportamento da outra, ou seja,

$$\mathrm{P}(X \le x_0, Y \le y_0) = \mathrm{P}(X \le x_0)\mathrm{P}(Y \le y_0) \tag{3.47}$$

Em termos da função acumulada de probabilidades conjuntas, as variáveis aleatórias X e Y são estatisticamente independentes se,

$$P_{X,Y}(x_0, y_0) = P_X(x_0)P_Y(y_0) \text{ ou } F_{X,Y}(x_0, y_0) = F_X(x_0)F_Y(y_0) \tag{3.48}$$

No caso de variáveis aleatórias discretas, a condição de independência reduz-se a

$$p_{X,Y}(x, y) = p_X(x)\, p_Y(y) \tag{3.49}$$

enquanto que, para variáveis aleatórias contínuas,

$$f_{X,Y}(x, y) = f_X(x)\, f_Y(y) \tag{3.50}$$

Portanto, a condição necessária e suficiente para que duas variáveis aleatórias sejam independentes é que a sua *função massa (ou densidade) de probabilidades conjuntas seja igual ao produto das funções massa (ou densidade) marginais*.

Exemplo 3.17 – Considere as funções não negativas de X e Y:
(a) $f(x, y) = 4xy, \text{com}\,(0 \le x \le 1, 0 \le y \le 1)$ e
(b) $g(x, y) = 8xy, \text{ com}\,(0 \le x \le y, 0 \le y \le 1)$. Verifique se tais funções são densidades e se X e Y são independentes.
Solução:
(a) Para que $f(x, y) = 4xy$ seja uma densidade, a condição é que

$\int_0^1\int_0^1 4xy\,dx\,dy = 1$. Portanto, $\int_0^1\int_0^1 4xy\,dx\,dy = 4\int_0^1 x\,dx\int_0^1 y\,dy = 1$ e, de fato, $f(x, y) = 4xy$ é uma densidade conjunta. Para a verificação de independência, a condição necessária e suficiente é dada pela equação 3.50, requerendo, para isso, o cálculo das marginais.

Marginal de X: . $f_X(x) = \int_0^1 f_{X,Y}(x, y)\,dy = 4\int_0^1 xy\,dy = 2x$

Marginal de *Y*: $f_Y(y) = 4\int_0^1 xy\,dx = 2y$. Portanto, como a densidade conjunta é o produto das marginais, as variáveis são independentes. (b) Procedendo da mesma forma para a função $g(x, y) = 8xy$, verifica-se que se trata de uma densidade conjunta. As marginais são $g_X(x) = 4x$ e $g_Y(y) = 4y^3$. Nesse caso, $g_{X,Y}(x, y) \neq g_X(x)\, g_Y(y)$ e, portanto, as variáveis não são independentes.

As propriedades do operador esperança matemática podem ser estendidas às funções de distribuição de probabilidades conjuntas. De fato, as equações 3.14 e 3.15, que definem as propriedades gerais do operador esperança matemática, podem ser estendidas para o caso de uma função *g*(*X*, *Y*) de duas variáveis aleatórias *X* e *Y*, por meio de

$$E[g(X,Y)] = \begin{cases} \sum_x \sum_y g(x,y)\, p_{X,Y}(x,y) \text{ para o caso discreto} \\ \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x,y)\, f_{X,Y}(x,y)\,dx\,dy \text{ para o caso contínuo} \end{cases} \quad (3.51)$$

Por meio da imposição $g(X,Y) = X^r Y^s$ na equação 3.51, é possível estender, para o caso bi-variado, a definição dos *momentos* $\mu'_{r,s}$, de ordens *r* e *s*, em relação à origem. Analogamente, fazendo $g(X,Y) = (X - \mu_X)^r (Y - \mu_Y)^s$ na equação 3.51, são definidos os *momentos centrais* $\mu_{r,s}$ de ordens *r* e *s*. É fácil verificar os seguintes casos particulares: (i) $\mu'_{1,0} = \mu_X$; (ii) $\mu'_{0,1} = \mu_Y$; (iii) $\mu_{2,0} = \mathrm{Var}[X] = \sigma_X^2$ e (iv) $\mu_{0,2} = \mathrm{Var}[Y] = \sigma_Y^2$.

O momento central $\mu_{r=1,s=1}$ recebe o nome específico de *covariância* de *X* e *Y* e fornece uma medida proporcional ao grau de associação linear entre essas variáveis. Formalmente, a covariância de *X* e *Y* é definida por

$$\mathrm{Cov}[X,Y] = \sigma_{X,Y} = E[(X - \mu_X)(Y - \mu_Y)] = E[XY] - E[X]\,E[Y] \quad (3.52)$$

Observe que se *X* e *Y* são variáveis independentes, é fácil demonstrar que *E*[*XY*]=*E*[*X*].*E*[*Y*]; nesse caso, verifica-se na equação 3.52 que, se *X* e *Y* são variáveis independentes, a covariância dessas variáveis é nula. Entretanto, *se* Cov[*X*,*Y*] = 0, *as variáveis X e Y não são necessariamente independentes*; de fato, nesse caso, não há *dependência linear* entre *X* e *Y*, embora *possa existir* dependência não linear. Como a covariância tem as unidades do produto entre as unidades de *X* e *Y*, é mais prático torná-la uma medida adimensional, dividindo-a por $\sigma_X . \sigma_Y$. A essa padronização, dá-se o nome de *coeficiente de correlação* $\rho_{X,Y}$. Portanto,

$$\rho_{X,Y} = \frac{\mathrm{Cov}[X,Y]}{\sigma_X \sigma_Y} = \frac{\sigma_{X,Y}}{\sigma_X \sigma_Y} \tag{3.53}$$

A exemplo de sua estimativa amostral $r_{X,Y}$, objeto do item 2.4.1 do capítulo 2, o coeficiente de correlação populacional é um número limitado entre –1 e 1. Novamente, se *as variáveis X e Y são independentes*, então $\rho_{X,Y} = 0$; a recíproca, entretanto, não é necessariamente verdadeira, pois X e Y podem estar associados por outra relação funcional, diferente da linear.

É importante ressaltar os seguintes resultados que decorrem da aplicação do operador esperança matemática às variáveis aleatórias X e Y

i. $E[aX + bY] = aE[X] + bE[Y]$, onde a e b são constantes.

ii. $\mathrm{Var}[aX + bY] = a^2\mathrm{Var[X]} + b^2\mathrm{Var[Y]} + 2ab\mathrm{Cov}[X,Y]$, se X e Y são dependentes.

iii. $\mathrm{Var}[aX + bY] = a^2\mathrm{Var[X]} + b^2\mathrm{Var[Y]}$, se X e Y são independentes.

iv. No caso de k variáveis aleatórias $X_1, X_2, \ldots, X_k$,
$E[a_1X_1 + a_2X_2 + \ldots + a_kX_k] = a_1E[X_1] + a_2E[X_2] + \ldots$
$\ldots + a_kE[X_k]$, onde $a_1, a_2, \ldots, a_k$ são constantes.

v. No caso de k variáveis aleatórias $X_1, X_2, \ldots, X_k$,
$\mathrm{Var}[a_1X_1 + a_2X_2 + \ldots + a_kX_k] =$
$= \sum_{i=1}^{k} a_i^2 \mathrm{Var}[X_i] + 2\sum_{i<j} a_i a_j \mathrm{Cov}[X_i, X_j]$.

vi. Para k variáveis *independentes*,
$\mathrm{Var}[a_1X_1 + a_2X_2 + \ldots + a_kX_k] = \sum_{i=1}^{k} a_i^2 \mathrm{Var}[X_i]$

Exemplo 3.18 – Considere que uma *amostra aleatória simples* de N elementos foi extraída de uma população de média μ e variância σ^2. Defina que Y represente a média aritmética dos N elementos da amostra. Calcule a média e a variância de Y.
Solução: A média aritmética pode ser expressa por $Y = \frac{X_1}{N} + \frac{X_2}{N} + \ldots + \frac{X_N}{N}$,
onde $X_1, X_2, \ldots, X_N$ representam os elementos constituintes da amostra. Como se trata de uma amostra aleatória simples, tais elementos podem ser vistos como variáveis aleatórias independentes, todas extraídas de uma

população de média μ e variância σ^2. Usando as propriedades (iv) e (vi) com $a_1 = a_2, \ldots = a_N = (1/N)$, com $E[X_1] = E[X_2] = \ldots = E[X_N] = \mu$ e $\mathrm{Var}[X_1] = \mathrm{Var}[X_2] = \ldots = \mathrm{Var}[X_N] = \sigma^2$, segue-se que $E[Y] = \frac{N\mu}{N} = \mu$ e $\mathrm{Var}[Y] = \frac{N\sigma^2}{N^2} = \frac{\sigma^2}{N}$ ou $\sigma_Y = \frac{\sigma}{\sqrt{N}}$.

Exemplo 3.19 – Demonstrar que a função geratriz de momentos conjuntos de duas variáveis aleatórias estatisticamente independentes X e Y, é igual ao produto das respectivas funções geratrizes de X e Y.
Solução: A função geratriz de momentos conjuntos de duas variáveis aleatórias X e Y é dada por $\phi_{X,Y}(t_1, t_2) = E[exp(t_1 X + t_2 Y)]$. Os momentos em relação à origem, de ordens r e s, podem ser obtidos a partir da função geratriz de momentos conjuntos, pelo cálculo de sua r-ésima derivada em relação a t_1 e da s-ésima derivada em relação a t_2, nos pontos $t_1 = t_2 = 0$. Entretanto, se as variáveis são independentes, pode-se escrever $\phi_{X,Y}(t_1, t_2) = E[\exp(t_1 X + t_2 Y)] = E[\exp(t_1 X)]E[\exp(t_2 Y)] = \phi_X(t_1)\phi_Y(t_2)$
Portanto, se duas variáveis são estatisticamente independentes, a função geratriz de momentos conjuntos é igual ao produto das funções geratrizes individuais. Inversamente, se a função geratriz de momentos conjuntos é igual ao produto das funções geratrizes individuais, então as variáveis são independentes.

De modo análogo à definição de valor esperado de uma variável aleatória X, pode-se definir também o *valor esperado condicional* de X, a partir de sua função de distribuição condicional. Com efeito, se duas variáveis aleatórias discretas X e Y, com funções massa de probabilidades conjuntas $p_{X,Y}(x, y)$ e marginais $p_X(x)$ e $p_Y(y)$, podem ser definidas as seguintes médias condicionais:

$$E[X|Y = y_0] = \sum_{\text{todos } x_i} x_i \frac{p_{X,Y}(x_i, y_0)}{p_Y(y_0)} = \sum_{\text{todos } x_i} x_i \, p_{X|Y}(x_i|y_0) \tag{3.54}$$

$$E[Y|X = x_0] = \sum_{\text{todos } y_j} y_j \frac{p_{X,Y}(x_0, y_j)}{p_X(x_0)} = \sum_{\text{todos } y_j} y_j \, p_{Y|X}(y_j|x_0) \tag{3.55}$$

Se as variáveis X e Y forem contínuas, com densidade conjunta dada por $f_{X,Y}(x, y)$ e marginais $f_X(x)$ e $f_Y(y)$, as médias condicionais são definidas como

$$E[X|Y = y_0] = \int_{-\infty}^{\infty} x \frac{f_{X,Y}(x, y_0)}{f_Y(y_0)} = \int_{-\infty}^{\infty} x \, f_{X|Y}(x|Y = y_0) \tag{3.56}$$

$$E[Y|X = x_0] = \int_{-\infty}^{\infty} y \frac{f_{X,Y}(x_0, y)}{f_X(x_0)} = \int_{-\infty}^{\infty} y \, f_{Y|X}(y|x = x_0) \tag{3.57}$$

3.8 – Distribuições de Probabilidades de Funções de Variáveis Aleatórias

Suponha que uma certa variável Y esteja associada a uma variável aleatória X, por alguma relação funcional monotônica crescente ou decrescente $Y = g(X)$, tais como $Y = \ln(X)$ ou $Y = \exp(-X)$, respectivamente, para $X > 0$. Por tratar-se de uma função de uma variável aleatória, Y também é uma variável aleatória. Uma vez conhecida a distribuição de probabilidades de X e a forma $Y = g(X)$, é possível deduzir a distribuição de Y.

Se X é uma variável aleatória discreta, com função massa de probabilidades dada por $p_X(x)$, o objetivo é deduzir a função massa de Y, ou seja $p_Y(y)$. Se a função $Y = g(X)$ é monótona *crescente* ou *decrescente*, existe uma relação biunívoca entre Y e X, sendo válido escrever que a cada $g(x) = y$ corresponde um $x = g^{-1}(y)$ e, portanto, $P(Y=y) = P[X = g^{-1}(y)]$, ou, genericamente,

$$p_Y(y) = p_X\left[g^{-1}(y)\right] \tag{3.58}$$

Se X é uma variável aleatória contínua, com funções densidade $f_X(x)$ e acumulada $F_X(x)$, considerações adicionais se fazem necessárias. De fato, o que se deseja calcular é $P(Y \leq y)$ ou $P[g(X) \leq y]$. Se a função $Y=g(X)$ é monótona *crescente*, existe uma relação biunívoca entre Y e X, sendo válido escrever que a cada $g(x) \leq y$ corresponde um $x \leq g^{-1}(y)$ e, portanto,

$$P\left(Y \leq y\right) = P\left[X \leq g^{-1}(y)\right] \text{ ou } F_Y(y) = F_X\left[g^{-1}(y)\right] \tag{3.59}$$

Inversamente, se a função $Y = g(X)$ é monótona *decrescente*, a cada $g(x) \leq y$ corresponde um $x \geq g^{-1}(y)$ e, portanto,

$$P\left(Y \leq y\right) = 1 - P\left[X \leq g^{-1}(y)\right] \text{ ou } F_Y(y) = 1 - F_X\left[g^{-1}(y)\right] \tag{3.60}$$

Em ambos os casos, a função densidade de Y é obtida pela derivação da função acumulada em relação a Y. Entretanto, como as funções densidades são sempre positivas e sua integração, no domínio completo da variável, deve ser igual a 1, é necessário tomar o valor absoluto da derivada de $g^{-1}(y)$, em relação a y. Em outros termos,

$$f_Y(y) = \frac{d}{dy} F_Y(y) = \frac{d\,F_X\left[g^{-1}(y)\right]}{dx}\left|\frac{d\left[g^{-1}(y)\right]}{dy}\right| = f_X\left[g^{-1}(y)\right]\left|\frac{d\left[g^{-1}(y)\right]}{dy}\right| = f_X\left[g^{-1}(y)\right]\left|J\right| \tag{3.61}$$

Na equação 3.61, o termo J, referente à derivada de $g^{-1}(y)$, em relação a y, é denominado *Jacobiano*.

Exemplo 3.20 (adap. de Kottegoda e Rosso, 1997) – Uma variável discreta geométrica X tem sua função massa de probabilidades dada por $p_X(x) = p(1-p)^{x-1}$, para $x = 1,2,3,...$ e $0 \le p \le 1$. Suponha que a variável X esteja associada à ocorrência no ano x, e não antes de x, de uma enchente maior ou igual à cheia de projeto de uma ensecadeira construída para proteger o canteiro de obras de uma barragem. A probabilidade de ocorrência de uma cheia maior do que a de projeto, em um ano qualquer, é p. Suponha que a ensecadeira original foi alteada e que, agora, o tempo para acontecer uma falha (em anos) passou a ser $Y = 3X$. Calcule a probabilidade do tempo para acontecer uma falha, sob o novo cenário de uma ensecadeira mais alta.
Solução: Com referência à equação 3.58, $Y = 3X \Rightarrow g^{-1}(Y) = Y/3$ e, portanto, $p_Y(y) = p(1-p)^{[(y/3)-1]}$, para $y = 3,6,9,...$ e $0 \le p \le 1$. Logo, conclui-se que as probabilidades de falha depois de 1,2,3 ... anos, antes do alteamento da ensecadeira, são equivalentes às probabilidades de falha depois de 3, 6, 9, ... anos, sob o novo cenário.

Exemplo 3.21 – Suponha que X seja uma variável Normal com parâmetros μ e σ. Defina uma nova variável $Y = \exp(X)$. Determine a função densidade de probabilidades de Y.
Solução: A distribuição Normal (ver exemplo 3.15) é ilimitada à esquerda e à direita. Quando X varia de $-\infty$ a $+\infty$, Y irá variar de 0 a ∞; portanto, a densidade de Y aplica-se apenas para $y \ge 0$. Com referência à equação 3.61, a função inversa é $x = g^{-1}(y) = ln(y)$ e, portanto, $|J| = 1/y$.
Substituindo essas funções na equação 3.61,

$$f_Y(y) = \frac{1}{y\sigma\sqrt{2\pi}} exp\left[-\frac{(ln\, y - \mu)^2}{2\sigma^2}\right], \text{ para } y \ge 0.$$ Essa distribuição é conhecida como LogNormal, a qual representa a distribuição de uma variável $Y = \exp(X)$, quando X é uma variável aleatória Normal.

A transformação dada pela equação 3.61 pode ser estendida para o caso de densidades bi-variadas. Para isso, considere a transformação de $f_{X,Y}(x,y)$ em $f_{U,V}(u,v)$, onde $U = u(X,Y)$ e $V = v(X,Y)$ são funções biunívocas continuamente diferenciáveis. Nesse caso, pode-se escrever

$$f_{U,V}(u,v) = f_{X,Y}\left[x(u,v), y(u,v)\right]|J| \qquad (3.62)$$

onde J representa o Jacobiano, calculado pelo seguinte determinante:

$$J = \begin{vmatrix} \dfrac{\partial x}{\partial u} & \dfrac{\partial x}{\partial v} \\ \dfrac{\partial y}{\partial u} & \dfrac{\partial y}{\partial v} \end{vmatrix} \tag{3.63}$$

Os limites de U e V dependem de suas relações com X e Y e devem ser cuidadosamente determinados, para cada caso particular.

Uma aplicação importante da equação 3.62 refere-se à determinação da distribuição da soma de duas variáveis aleatórias, ou seja, de $U = X + Y$, dada a densidade $f_{X,Y}(x,y)$. Nesse caso, cria-se uma variável fictícia auxiliar $V = X$, de modo a obter as seguintes funções inversas: $x\ (u,v) = v$ e $y\ (u,v) = u\text{-}v$. O Jacobiano, para esse caso, é

$$J = \begin{vmatrix} 0 & 1 \\ 1 & -1 \end{vmatrix} = -1 \tag{3.64}$$

Substituindo essas grandezas na equação 3.62,

$$f_{U,V}(u,v) = f_{X,Y}[v, u-v] \tag{3.65}$$

Entretanto, o que nos interessa é a distribuição marginal de U, a qual pode ser obtida integrando-se a densidade conjunta, dada pela equação 3.65, no domínio $[A,B]$ de definição da variável V. Portanto,

$$f_U(u) = \int_A^B f_{X,Y}(v, u-v)\,dv = \int_A^B f_{X,Y}(x, u-x)\,dx \tag{3.66}$$

Para a situação particular em que X e Y são *independentes*, $f_{X,Y}(x,y) = f_X(x).f_Y(y)$ e a equação 3.66 torna-se

$$f_U(u) = \int_A^B f_X(x) f_Y(u-x)\,dx \tag{3.67}$$

A operação contida no segundo membro da equação 3.67 é conhecida por *convolução*. Portanto, a densidade da soma de duas variáveis aleatórias independentes é igual à convolução das funções densidades dos termos em foco.

Exemplo 3.22 – A distribuição de uma variável aleatória X é dita uniforme se sua densidade é $f_X(x) = 1/a$, para $0 \le x \le a$. Suponha duas variáveis aleatórias uniformes *independentes* X e Y, ambas definidas no intervalo [0,a]. Determine a densidade de $U = X + Y$.

Solução: A aplicação da equação 3.67 a esse caso específico é simples, à exceção da definição dos limites A e B de integração. De fato, as seguintes condições devem ser obedecidas: $0 \leq (u - x) \leq a$ e $0 \leq x \leq a$. Essas inequações podem ser manipuladas e transformadas em $(u - a) \leq x \leq u$ e $0 \leq x \leq a$. Assim, os limites de integração passam a ser $A = \text{Max}(u - a, 0)$ e $B = \text{Min}(u, a)$, o que implica em duas possibilidades: $u < a$ e $u > a$. Para $u < a$, $A = 0$ e $B = u$, e a equação 3.67 torna-se

$$f_U(u) = \frac{1}{a^2}\int_0^u dx = \frac{u^2}{a^2}, \text{ para } 0 \leq u \leq a.$$ Para $u > a$, $A = (u - a)$ e $B = a$, e a equação 3.67 torna-se $$f_U(u) = \frac{1}{a^2}\int_{u-a}^{a} dx = \frac{2a-u}{a^2}, \text{ para } a \leq u \leq 2a.$$

Portanto, a densidade da soma de duas variáveis uniformes tem a forma de um triângulo isósceles.

3.9 – Distribuições Mistas

Considere que uma variável aleatória contínua X tem o seu comportamento probabilístico descrito por uma composição de m distribuições, denotadas por $f_i(x)$, ponderadas por parâmetros λ_i, com $i = 1,2,...,m$, tais que $\sum_{i=1}^{m}\lambda_i = 1$.

Nesse caso, a função densidade de probabilidades de X é do *tipo mista* e dada por

$$f_X(x) = \sum_{i=1}^{m}\lambda_i \, f_i(x) \tag{3.68}$$

A função acumulada de probabilidades é expressa por

$$F_X(x) = \int_{-\infty}^{x}\sum_{i=1}^{m} f_i(x)dx \tag{3.69}$$

Em hidrologia, as distribuições mistas encontram aplicação no estudo probabilístico de variáveis aleatórias cujas ocorrências resultam da ação de fatores causais diferentes. Por exemplo, as precipitações de curta duração, em um dado local, podem ser do tipo frontal ou do tipo convectivo, a depender do mecanismo de ascensão das massas de ar úmido. Se do tipo frontal, o comportamento probabilístico das intensidades pode ser descrito por uma densidade $f_1(x)$. Entretanto, se do tipo convectivo, as intensidades serão certamente maiores do que as primeiras e serão descritas por $f_2(x)$. Se a proporção com que ocorrem

precipitações frontais é dada por λ_1, a proporção das chuvas convectivas é $\lambda_2 = (1 - \lambda_1)$. Em seguida, o comportamento global das intensidades de precipitação de curta duração, sejam frontais ou convectivas, será dado pela composição das densidades parciais $f_1(x)$ e $f_2(x)$, ponderadas por λ_1 e λ_2, por meio das equações 3.68 e 3.69.

Exercícios

1) Os valores possíveis dos níveis d'água H (com relação ao nível médio), em cada um dos rios A e B, são: H = -3, -2, -1, 0, 1, 2, 3, 6 metros.
(a) Considere os seguintes eventos para o rio A : $A_1 = \{H_A > 0\}$, $A_2 = \{H_A = 0\}$ e $A_3 = \{H_A \leq 0\}$. Faça uma lista dos pares possíveis de eventos disjuntos entre A_1, A_2 e A_3.
(b) Em cada rio considere os seguintes eventos: nível médio: $M = \{-1 \leq H \leq 1\}$, estiagem: $E = \{H < 1\}$ e cheia: $C = \{H > 1\}$. Ordene os pares (h_A, h_B) e identifique os pontos amostrais que definem os níveis d'água em A e B, respectivamente; por exemplo, (3,-1) define a condição simultânea $h_A = 3$ e $h_B = -1$. Determine os pontos amostrais para os eventos $M_A \cap M_B$ e $(C_A \cup E_A) \cap M_B$.

2) Considere a seção de um reservatório de acumulação, ilustrada na figura a seguir, na qual o volume útil V $(0 \leq V \leq c)$ foi discretizado em volumes contidos entre os níveis w_1 e w_2, w_2 e w_3, w_3 e w_4, w_4 e c, e, respectivamente, agrupados nos eventos A_1, A_2, A_3 e A_4.

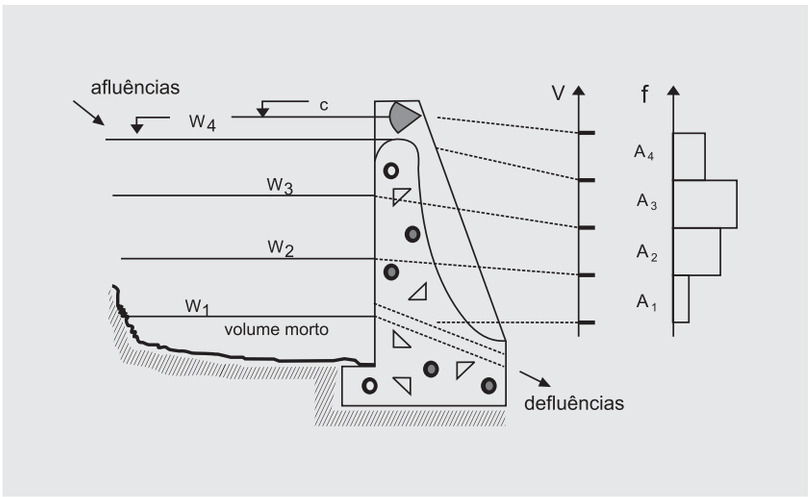


Figura 3.13 – Exercício 2

Pede-se identificar os limites do NA do reservatório para os seguintes eventos:

a) $(A_4)^c \cap (A_1)^c$

b) $(A_3 \cup A_2)^c \cap (A_1)^c$

c) $[A_4 \cup (A_1 \cup A_2)^c]^c$

d) $(A_1 \cap A_2)^c$

3) Se a ocorrência de um dia chuvoso é um evento independente com probabilidade 0,25, qual é a probabilidade

(a) de ocorrerem 4 dias chuvosos em 1 semana?

(b) dos próximos 4 dias serem chuvosos?

(c) de ocorrerem 4 dias consecutivos com chuva durante uma semana qualquer, com 3 dias sem chuva no restante da semana?

4) O rio R perto da cidade C atinge ou supera o nível de cheia, a cada ano, com probabilidade de 0,2. Algumas partes da cidade são inundadas a cada ano com probabilidade 0,1. A observação mostra que quando o rio R se encontra em níveis de enchente, a probabilidade da cidade C ser inundada aumenta para 0,2.

(a) calcule a probabilidade de ocorrer enchente ou no rio ou na cidade;

(b) calcule a probabilidade de ocorrer enchentes tanto no rio como na cidade.

5) Uma barragem de gravidade pode romper-se por escorregamento ao longo do plano de contato com as fundações (evento A) ou por rotação em torno do ponto mais baixo da face de jusante (evento B). Se (i) P(A) = 2P(B); (ii) P(A|B) = 0,8; e (iii) a probabilidade de rompimento da barragem é igual a 10^{-3}, pede-se (a) determinar a probabilidade de que o escorregamento irá ocorrer e (b) se ocorreu o rompimento da barragem, qual é a probabilidade de que ele se deveu somente ao escorregamento?

6) O rio Blackwater, cuja bacia localiza-se na área central da Inglaterra, tem sido constantemente monitorado para controle da poluição, através de 38 estações ao longo do rio. A tabela abaixo lista uma das amostras para oxigênio dissolvido (OD) e demanda bioquímica de oxigênio (DBO), ambos em mg/ℓ, para as 38 estações (adap. de Kottegoda e Rosso, 1997).

Tabela 3.1 – Exercício 6

OD	DBO	OD	DBO	OD	DBO	OD	DBO
8,15	2,27	6,74	3,83	7,28	3,22	8,46	2,82
5,45	4,41	6,9	3,74	7,44	3,17	8,54	2,79
6,05	4,03	7,05	3,66	7,59	3,13	8,62	2,76
6,49	3,75	7,19	3,58	7,73	3,08	8,69	2,73
6,11	3,37	7,55	3,16	7,85	3,04	8,76	2,7
6,46	3,23	6,92	3,43	7,97	3	9,26	2,51
6,22	3,18	7,11	3,36	8,09	2,96	9,31	2,49
6,05	4,08	7,28	3,3	8,19	2,93	9,35	2,46
6,3	4	7,44	3,24	8,29	2,89	**Média :**	**Média:**
6,53	3,92	7.6	3,19	8,38	2,86	**7,5**	**3,2**

Sabendo que as médias amostrais de OD e DBO são respectivamente 7,5 e 3,2 mg/l, definem-se os seguintes eventos: B_1= {OD ≤ 7,5 e DBO >3,2}; B_2= {OD >7,5 e DBO > 3,2}; B_3= {OD > 7,5 e DBO ≤ 3,2} e B_4= {OD ≤ 7,5, DBO ≤ 3,2}. Um evento de referência, com base em OD e DBO, pode ser aquele definido pela variação de ambas variáveis dentro do intervalo [média - desvio padrão, média + desvio padrão]. Se os d.p.'s de OD e DBO são iguais a 1,0 e 0,5 mg/l, respectivamente, o evento de referência é A = {6,5 < OD < 8,5 e 2,7 < DBO < 3,7). Pede-se:

a) fazer um diagrama de dispersão entre OD e DBO, demarcando, no gráfico, os eventos B_1, B_2, B_3, B_4 e A;

b) estimar as probabilidades dos eventos B_i pelas respectivas freqüências relativas;

c) usar o teorema da probabilidade total para calcular a probabilidade de OD e DBO situarem-se dentro dos limites do evento de referência; e

d) usar o teorema de Bayes para calcular a probabilidade de OD e DBO situarem-se nos limites definidos pelos eventos B_1 a B_4, sabendo-se que eles estão dentro da variação do evento de referência A.

7) Um rio se bifurca nos trechos A e B, imediatamente a jusante de uma instalação industrial situada às suas margens. O nível de oxigênio dissolvido nos trechos A e B é uma indicação do grau de poluição causada pelo lançamento do efluente no curso d'água. Medições realizadas ao longo de vários anos indicam que as probabilidades dos trechos A e B estarem poluidos são de 2/5 e 3/4, respectivamente. Além disso, a probabilidade de *pelo menos* um dos trechos estar poluido é 4/5.

a) Determine a probabilidade do trecho A estar poluido sabendo-se que o trecho B está poluido.

b) Determine a probabilidade do trecho B estar poluido sabendo-se que o trecho A está poluido.

8) As probabilidades de ocorrer uma altura de chuva superior a 60 mm nos meses de Janeiro, Fevereiro, ... , Dezembro são, respectivamente, 0,24; 0,31; 0,30; 0,45; 0,20; 0,10; 0,05; 0,05; 0,04; 0,06; 0,10 e 0,20. Suponha que um registro de altura mensal de chuva superior a 60 mm foi tomado ao acaso. Calcule a probabilidade de que tal registro se refira ao mês de Julho.

9) Se a função densidade de probabilidade de uma variável aleatória X é dada por $f_X(x) = c(1 - x^2), -1 \leq x \leq 1$ e c constante,
a) calcular o valor de c
b) determine a função de probabilidade acumulada de X.
c) calcule P($X \leq 0{,}75$)

10) Numa bacia hidrográfica de pequeno porte, a probabilidade de que não chova em um dia qualquer é 0,60. Dado que chove, a magnitude da precipitação é uma variável exponencialmente distribuída com $\theta = 10$ mm. Dependendo das condições antecedentes do solo, uma precipitação inferior a 20 mm pode ocasionar o transbordamento de um riacho. A probabilidade desse evento é 0,10. Se chover mais de 20 mm, a probabilidade de que o riacho transborde é 0,90. Sabendo-se que o riacho transbordou, qual é a probabilidade de que tenha ocorrido uma chuva superior a 20 mm?

11) Determine a média e a variância de uma variável aleatória geométrica cuja função massa de probabilidades é dada por

$p_X(x) = p(1-p)^{x-1}$, para $x = 1,2,3,...$ e $0 \leq p \leq 1$

12) Sob quais condições a relação P($X \leq E[X]$) = 50% é válida?

13) Demonstre que $E[X^2] \geq (E[X])^2$

14) Se X e Z são variáveis aleatórias, demonstre as seguintes relações:

(a) $E\left(\frac{X - \mu_X}{\sigma_X}\right) = 0$

(b) $Var\left(\frac{X - \mu_X}{\sigma_X}\right) = 1$

(c) $\rho_{X,Z} = \text{Cov}\left(\frac{X - \mu_X}{\sigma_X}, \frac{Z - \mu_Z}{\sigma_Z}\right)$

15) Uma amostra de 36 observações foi extraída da população de uma variável Normal X, com parâmetros $\mu_X = 4$ e $\sigma_X = 3$. Determine o valor esperado e o desvio padrão da média aritmética da amostra.

16) A função massa de probabilidades da distribuição binomial é dada por $p_X(x) = \binom{n}{x} p^x (1-p)^{n-x}$, $x = 0,1,2,\ldots$. Calcule a média e a variância da distribuição binomial de parâmetros n e p, através da função geratriz de momentos. Lembre- se, pelo binômio de Newton, que $(a+b)^n = \sum_{k=0}^{n} \binom{n}{k} a^k b^{n-k}$.

17) X e Y são duas variáveis aleatórias independentes com densidades $\lambda_1 \exp(-x\lambda_1)$ e $\lambda_2 \exp(-y\lambda_2)$ respectivamente, para $x \geq 0$ e $y \geq 0$. Pede-se:
a) determinar a função geratriz de momentos de $Z=X+Y$; e
b) determinar a média e a variância de Z a partir da função geratriz de momentos.

18) Suponha que a função densidade de probabilidade conjunta de X e Y seja dada por

$$f_{X,Y}(x,y) = \frac{exp\left(-\frac{x}{y}\right) exp(-y)}{y}; \; 0 < x < \infty, 0 < y < \infty$$

a) calcule $P(X<2|Y=3)$;
b) calcule $P(Y>3)$; e
c) determine $E[X|Y=4]$.

19) Suponha que a duração X de uma precipitação e sua intensidade Y tenham distribuição de probabilidades conjuntas, cuja função densidade é $f_{X,Y}(x,y) = [(a+cy)(b+cx)-c] exp(-ax-by-cxy)$, para $x, y \geq 0$ e parâmetros a, $b \geq 0$ e $0 \leq c \leq 1$. Suponha que os parâmetros valham a = 0,07 h^{-1}, b =1,1 h/mm e c = 0,08 mm^{-1}. Para o propósito de se projetar um sistema de drenagem, pergunta-se qual é a probabilidade de que uma precipitação que dure 6 horas vá exceder a intensidade de 3 mm/h?

20) Volte ao exercício 19 e suponha que $c = 0$. Nesse caso, demonstre que as variáveis X e Y são estatisticamente independentes.

21) Considere a função densidade de probabilidade de uma variável aleatória dada por $f_X(x) = 0,35, 0 \leq X \leq a$ Pede-se (a) expressar a densidade de $Y = \ln(X)$, com seus limites de definição e (b) elaborar um gráfico de $f_Y(y)$.

22) Uma barragem deve possuir borda livre acima do NA máximo-maximorum para a arrebentação de ondas devidas ao vento, evitando que essas sobreponham sua crista. Suponha válida a seguinte relação empírica para a altura da onda eólica (em cm):

$$Z=\frac{F}{1500d}V^2$$

onde:

V = velocidade do vento em km/h,

F = pista de vento ou "fetch" em m, e

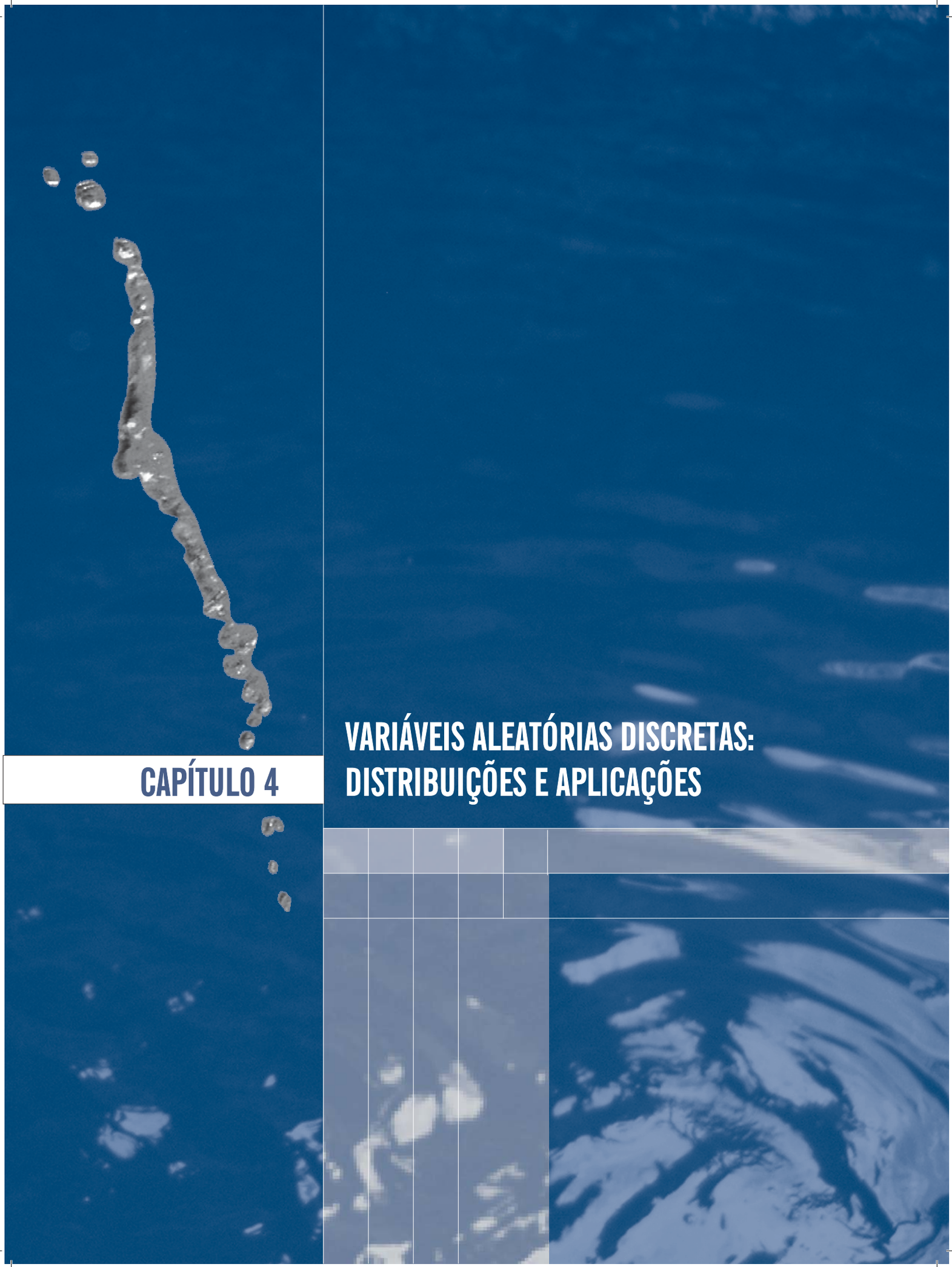
d = profundidade média do reservatório em m.

a) Se a velocidade do vento possui distribuição exponencial com média v_0, para $v \geq 0$, determine a função densidade de probabilidade de Z.

b) Se $v_0 = 30$ km/h, $F = 300$ m e $d = 10$ m, calcule P(Z >30 cm).

23) A função densidade de probabilidade da distribuição Gama (com parâmetros α e λ) é dada por $f_X(x)=\frac{\lambda^\alpha x^{\alpha-1} exp(-\lambda x)}{\Gamma(\alpha)}$, com $x,\alpha,\lambda>0$, onde $\Gamma(\alpha)=\int_0^\infty t^{\alpha-1}\exp(-t)\,dt$ [ver Anexo 4 para uma breve revisão sobre as propriedades da função $\Gamma(.)$]. Suponha que X e Y sejam variáveis aleatórias contínuas e independentes, distribuídas segundo Gama com parâmetros (α_1 e λ_1) e (α_2 e λ_2), respectivamente. Ache a expressão das funções densidade de probabilidades conjuntas e de probabilidades marginais de $U = X+Y$ e $V = X/(X+Y)$.

24) Suponha que, para as chuvas de duração igual a 2 horas, a proporção de chuvas convectivas é de 0,55, enquanto a de chuvas frontais é de 0,45. Se X denota as intensidades dessas chuvas e supondo que as de ambos os tipos são exponencialmente distribuídas com parâmetros θ =15 mm/h, para as do tipo convectivo, e θ = 8 mm/h, para as frontais, pede-se: (a) determinar e fazer um gráfico da função densidade de probabilidades das intensidades de chuva de qualquer origem e (b) calcule P(X > 25 mm/h).



CAPÍTULO 4

VARIÁVEIS ALEATÓRIAS DISCRETAS: DISTRIBUIÇÕES E APLICAÇÕES

CAPÍTULO 4 VARIÁVEIS ALEATÓRIAS DISCRETAS: DISTRIBUIÇÕES E APLICAÇÕES

No capítulo 3, foram apresentados os fundamentos da teoria de probabilidades, necessários à compreensão das variáveis aleatórias e de suas distribuições. No presente capítulo, dá-se início à formulação e à descrição dos principais *modelos de distribuição de probabilidades* capazes de sintetizar o comportamento das variáveis aleatórias hidrológicas. Um modelo de distribuição de probabilidades é uma forma matemática abstrata, a qual, por suas características intrínsecas de variabilidade e conformação, devem ser capazes de representar, de modo conciso, as variações possíveis de uma variável aleatória. Um modelo de distribuição de probabilidades também é uma *forma paramétrica*, ou seja, um modelo matemático prescrito por parâmetros, cujos valores numéricos o definem completamente e o particularizam para uma certa amostra de observações de uma variável aleatória. Uma vez estimados os valores numéricos de seus parâmetros, um modelo de distribuição de probabilidades pode constituir-se em uma síntese plausível do comportamento de uma variável aleatória e ser empregado para interpolar, ou extrapolar, probabilidades e/ou quantis não contidos na amostra de observações.

Os modelos de distribuição de probabilidades são classificados em *discretos e contínuos*, de modo consoante com as variáveis aleatórias cujo comportamento visam modelar. Uma função de distribuição discreta é aquela empregada para modelar o comportamento de uma variável aleatória cujo espaço amostral é do tipo numerável, composto por valores isolados, em geral, números inteiros. Os principais modelos de variáveis aleatórias discretas, que encontram uma ampla gama de aplicações em hidrologia, podem ser agrupados em três grandes categorias. A primeira está relacionada as variações dos chamados *processos de Bernoulli* e inclui as distribuições binomial, geométrica e binomial negativa. A segunda refere-se aos *processos de Poisson*, na qual se destaca a própria distribuição de Poisson. A terceira inclui as distribuições hipergeométrica e multinomial. A descrição de tais modelos discretos de distribuição de probabilidades é o objeto deste capítulo 4. Os principais modelos contínuos serão descritos no capítulo 5.

4.1 – Processos de Bernoulli

Considere um experimento com somente dois resultados possíveis e dicotômicos: ‘sucesso’, designado pelo símbolo S, e ‘falha’, por F. O espaço amostral desse

experimento é dado pelo conjunto {S,F}. Tal experimento é conhecido como de Bernoulli. Se a probabilidade de ocorrer um sucesso é igual a p e se associarmos a esse experimento uma variável aleatória discreta X, cujos valores possíveis são $X = 1$ para o resultado S e $X = 0$ para o resultado F, diz-se que X segue uma distribuição de Bernoulli. A correspondente função massa de probabilidades é dada por

$$p_X(x) = p^x(1-p)^{1-x}, \text{ para } x = 0,1 \text{ e } 0 \leq p \leq 1 \tag{4.1}$$

com valor esperado $E[X] = p$ e $\text{Var}[X] = p(1-p)$.

Agora, de modo mais geral, suponha que a escala de tempo de um determinado processo estocástico tenha sido discretizada em intervalos de largura definida, por exemplo, em intervalos anuais, indexados por $i = 1, 2, ...$ Suponha também que, em cada intervalo de tempo, pode ocorrer um único 'sucesso', com probabilidade p, ou uma única 'falha', com probabilidade $(1-p)$, e que essas probabilidades não são afetadas pelas ocorrências anteriores. Um processo composto por essa seqüência de repetições independentes de experimentos de Bernoulli é igualmente denominado processo de Bernoulli.

Para ilustrar a aplicação dos processos de Bernoulli em hidrologia, considere uma seção fluvial hipotética cujo nível d'água de extravasamento corresponde à vazão Q_0. As vazões médias diárias nesta seção fluvial são monitoradas por uma estação fluviométrica, cujos registros se estendem por N anos de observações e constituem a *série hidrológica completa* para esse local. Para cada ano, seleciona-se o máximo valor entre as 365 (ou 366) vazões médias diárias, o qual é um dos N elementos da *série hidrológica reduzida* de vazões médias diárias máximas anuais Q^{max}, ilustrada na Figura 4.1. Em um ano qualquer i, para $1 \leq i \leq N$, podemos definir como 'sucesso' o evento $\left\{\text{S} : Q_i^{max} > Q_0\right\}$ e como 'falha' o evento complementar $\left\{\text{F} : Q_i^{max} \leq Q_0\right\}$. Pela natureza do mecanismo de formação da cheia anual, é bastante plausível admitir a hipótese de que a probabilidade de ocorrência de um 'sucesso' (ou de uma 'falha'), em um ano qualquer, não seja afetada pelas ocorrências anteriores. Supondo que a probabilidade anual do evento $\left\{\text{S} : Q_i^{max} > Q_0\right\}$ é igual p, verifica-se, então, o preenchimento de todos os requisitos para considerar essa seqüência independente como um processo de Bernoulli.

Aos processos de Bernoulli associam-se três diferentes tipos de variáveis aleatórias discretas Y:

i. a variável é dita *binomial*, quando Y refere-se ao número de 'sucessos' em N repetições independentes;

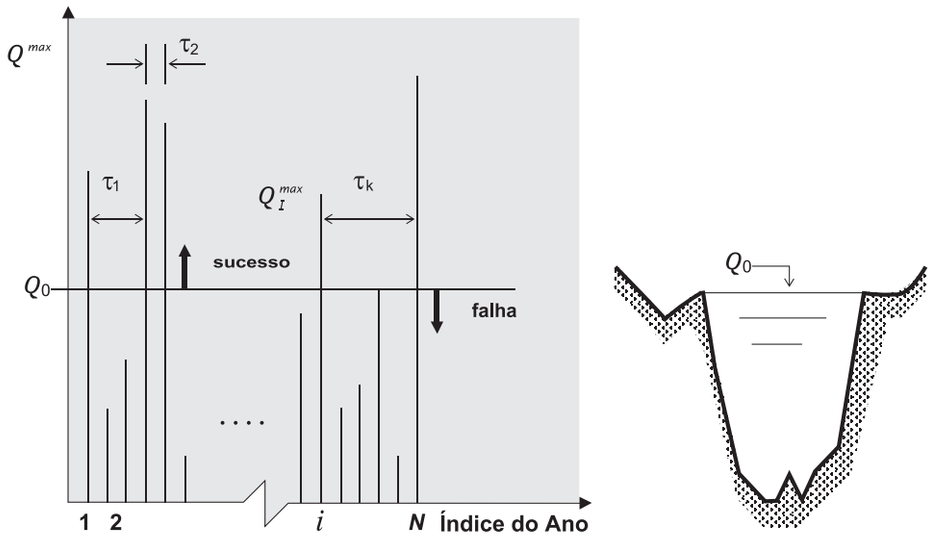


Figura 4.1 – Cheias máximas anuais como ilustração de um processo de Bernoulli

ii. a variável é denominada *geométrica*, quando *Y* refere-se ao número de repetições independentes necessárias para que um único 'sucesso' ocorra; e

iii. a variável é denominada *binomial negativa*, quando *Y* refere-se ao número de repetições independentes necessárias para que um certo número *r* de 'sucessos' ocorram.

As distribuições de probabilidades dessas três variáveis, associadas aos processos de Bernoulli, serão detalhadas a seguir.

4.1.1 – Distribuição Binomial

Considere um experimento composto por uma seqüência de N repetições independentes de um experimento de Bernoulli. Em cada um desses experimentos de Bernoulli, a probabilidade de ocorrer um 'sucesso', designado por S, é constante e igual a p, e a probabilidade de 'falha' F é dada por $(1-p)$. O espaço amostral do experimento composto contém 2^N pontos, com cada um deles correspondendo aos N pares de S's e F's. Para cada experimento isolado, a variável de Bernoulli, denotada por X, pode ter o valor $X = 1$, se o resultado for um 'sucesso', ou $X=0$, se o experimento resultar em uma 'falha'. Um ponto qualquer, tomado ao acaso no espaço amostral, poderia conter, por exemplo, a seqüência {S, F, S, S, ... , F, F}, o que faria com que $X_1 = 1$, $X_2 = 0$, $X_3 = 1$, $X_4 = 1$, ... , $X_{N-1} = 0$, $X_N = 0$. O experimento composto desse modo é caracterizado como um processo de Bernoulli.

Com base no processo de Bernoulli, tal como anteriormente descrito, considere que a variável aleatória discreta Y representa o número de 'sucessos', entre as N possibilidades. É evidente que a variável Y pode assumir os valores 0, 1, ... , N e que $Y = \sum_{i=1}^{N} X_i$. Como decorrência da hipótese de independência entre os experimentos 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 função massa de probabilidade da variável Y é dada por

$$p_Y(y) = \frac{N!}{y!(N-y)!} p^y (1-p)^{N-y} = \binom{N}{y} p^y (1-p)^{N-y},\ y = 0,1,\ldots,N \text{ e } 0 < p < 1 \quad (4.2)$$

a qual é denominada distribuição *binomial*, com parâmetros N e p. Note que a distribuição de Bernoulli é um caso particular da distribuição binomial com parâmetros $N = 1$ e p. As funções massa da distribuição binomial com parâmetros $N = 8$, $p = 0,3$, $p = 0,5$ e $p = 0,7$ estão ilustradas na Figura 4.2. Observe, nessa figura, que o valor central e a forma da função massa de probabilidades da variável aleatória binomial sofrem profundas alterações quando o valor do parâmetro p é modificado, mantendo-se N constante.

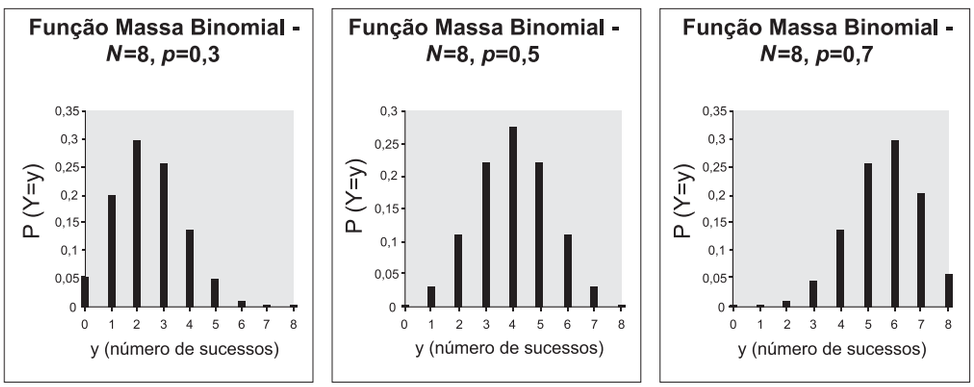


Figura 4.2 – Exemplos de funções massa de probabilidades da distribuição binomial

A função acumulada de probabilidades da distribuição binomial fornece a probabilidade de X ser menor ou igual ao argumento x e é dada por

$$F_Y(y) = \sum_{i=0}^{y} \binom{N}{i} p^i (1-p)^{N-i},\ y = 0,1,2,\ldots,N \quad (4.3)$$

O valor esperado, a variância e o coeficiente de assimetria da distribuição binomial (ver exercício 16 do capítulo 3) são dados pelas seguintes expressões:

$$E[Y] = N\,p \tag{4.4}$$

$$\mathrm{Var}[Y] = N\,p(1-p) \tag{4.5}$$

$$\gamma = \frac{1-2p}{\sqrt{N\,p(1-p)}} \tag{4.6}$$

A função massa da distribuição binomial é simétrica quando $p = 0{,}5$, assimétrica positivamente quando $p < 0{,}5$ e negativamente, em caso contrário, tal como demonstram os exemplos da Figura 4.2.

Exemplo 4.1 – Fez-se a contagem de *E. Coli* em 10 amostras de água. As contagens positivas, expressas em centenas de organismos por 100 ml de água (10^2/100ml), são 17, 21, 25, 23, 17, 26, 24, 19, 21 e 17, com média e a variância amostrais iguais a 21 e 10,6 respectivamente. Suponha que N represente o número total dos diferentes organismos presentes em cada amostra (número de 'tentativas') e que p represente a fração correspondente ao organismo *E. Coli* (probabilidade de 'sucesso'). Se X denota o número de *E. Coli* (10^2/100ml) em cada amostra, estimar P(X = 20). (adap. de Kottegoda e Rosso, 1997)

Solução: No caso presente, não conhecemos os verdadeiros valores numéricos da média e da variância populacionais. Entretanto, podemos estimá-los pelos valores amostrais, ou seja, $\hat{\mu}_Y = \bar{y}$ e $\hat{\sigma}_Y^2 = S_y^2$, onde o símbolo '^' indica 'estimativa'. Explicitando (1-p), na equação 4.5, segue-se que

$$1-p = \frac{Var[Y]}{Np} = \frac{Var[Y]}{E[Y]} \Rightarrow 1-\hat{p} = \frac{S_y^2}{\bar{y}} = \frac{10{,}6}{21} = 0{,}505 \Rightarrow \hat{p} = 0{,}495\,.$$

Como $E[Y] = Np$, pode-se estimar N como (21/0,495) = 43. Na seqüência,

$$\mathrm{P}(y = 20) = \mathrm{p}_Y(20) = \binom{43}{20} 0{,}495^{20}\ 0{,}505^{23} = 0{,}1123$$

Exemplo 4.2 - Na situação ilustrada pela Figura 4.1, suponha que N = 10 anos e que a probabilidade da vazão Q_0 ser superada em um ano qualquer é p = 0,25. Pergunta-se (a) qual é a probabilidade de que a vazão Q_0 tenha sido superada exatamente 2 vezes em 10 anos? e (b) qual é a probabilidade de que a vazão Q_0 tenha sido superada pelo menos 2 vezes em 10 anos?

Solução: É fácil verificar a completa adequação do cenário ilustrado pela Figura 4.1 a um processo de Bernoulli, bem como da variável 'número de sucessos em N anos' a uma variável binomial Y. (a) A probabilidade de que a vazão Q_0 tenha sido superada exatamente 2 vezes em 10 anos pode ser calculada diretamente pela equação 4.2, ou

$$p_Y(2) = \frac{10!}{2!8!} 0{,}25^2 \left(1-0{,}25\right)^8 = 0{,}2816.$$

(b) A probabilidade de que a vazão Q_0 tenha sido superada pelo menos 2 vezes em 10 anos é igual à probabilidade de que o evento tenha ocorrido 2, 3, 4, ... , 10 vezes, em 10 anos, ou seja, a soma dos resultados da função massa para todos esses argumentos. Entretanto, esse cálculo é equivalente ao complemento, em relação a 1, da soma das probabilidades de que o evento não tenha ocorrido ou que tenha ocorrido apenas 1 vez. Portanto, $\mathrm{P}(Y \geq 2) = 1 - \mathrm{P}(Y < 2) = 1 - p_Y(0) - p_Y(1) = 0{,}7560$.

A distribuição binomial possui a *propriedade aditiva*, ou seja, se Y_1 e Y_2 são variáveis binomiais, com parâmetros respectivamente iguais a (N_1, p) e (N_2, p), então, a variável (Y_1+Y_2) também será binomial, com parâmetros $(N_1 + N_2, p)$. Outra propriedade importante dos processos de Bernoulli, em geral, e da distribuição binomial, em particular, é que a probabilidade de qualquer combinação de 'sucessos' e 'falhas' não depende da origem, na escala de tempos, a partir da qual eles são contados. Esse fato decorre da hipótese de independência entre as ocorrências e da consideração de que a probabilidade do 'sucesso' p é constante.

4.1.2 – Distribuição Geométrica

Em um processo de Bernoulli, a *variável geométrica Y* está associada ao número de experimentos (ou tentativas) necessários para que um único 'sucesso' ocorra. Portanto, se a variável assume o valor $Y = y$, isso significa que ocorreram $(y - 1)$ 'falhas' antes da ocorrência do 'sucesso', exatamente na y-ésima tentativa. As funções massa e acumuladas da distribuição geométrica são dadas pelas seguintes equações:

$$p_Y(y) = p(1-p)^{y-1},\ y = 1,2,3,\ldots \text{ e } 0 < p < 1 \tag{4.7}$$

$$P_Y(y) = \sum_{i=0}^{y} p\,(1-p)^{i-1},\ y = 1,2,3,\ldots \tag{4.8}$$

nas quais, p, ou seja, a probabilidade de ocorrência de um 'sucesso', representa o único parâmetro da distribuição.

O valor esperado da distribuição geométrica é determinado do seguinte modo:

$$E[Y] = \sum_{y=1}^{\infty} y\,p(1-p)^{y-1} = p\sum_{y=1}^{\infty} y(1-p)^{y-1} = p\sum_{y=1}^{\infty} \frac{d}{d(1-p)}(1-p)^y = p\,\frac{d}{d(1-p)} \sum_{y=1}^{\infty} (1-p)^y \tag{4.9}$$

Na equação 4.9, a soma $\sum_{y=1}^{\infty}(1-p)^{y}$, com $0 < p < 1$, converge para $\left(\frac{1-p}{p}\right)$.

Substituindo esse termo na equação 4.9 e tomando a derivada em relação a (1-*p*), resulta que

$$E[Y] = \frac{1}{p} \tag{4.10}$$

Portanto, o valor esperado de uma variável geométrica é o inverso da probabilidade de 'sucesso' *p* de um processo de Bernoulli. A variância de uma variável geométrica pode ser obtida por artifício similar e resulta ser

$$\mathrm{Var}[Y] = \frac{1-p}{p^{2}} \tag{4.11}$$

O coeficiente de assimetria da distribuição geométrica é dado por

$$\gamma = \frac{2-p}{\sqrt{1-p}} \tag{4.12}$$

As funções massa da distribuição geométrica com parâmetros $p = 0{,}3$, $p = 0{,}5$ e $p = 0{,}7$ estão ilustradas na Figura 4.3.

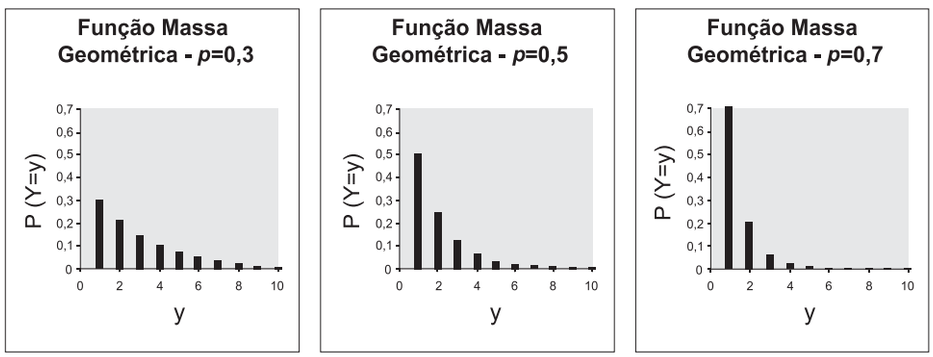


Figura 4.3 - Exemplos de funções massa de probabilidades da distribuição geométrica

Aproveitemos o cenário ilustrado pela Figura 4.1 para introduzir um conceito de extrema importância em hidrologia, que é o de *tempo de retorno*. Na Figura 4.1, considere que o número de anos entre 'sucessos' consecutivos seja denotado pela variável τ, a qual chamaremos aqui de *tempo de recorrência*. Portanto, com referência à Figura 4.1, se tomarmos a origem da escala de tempo, como o ano do primeiro 'sucesso', teríamos que aguardar $\tau_1 = 3$ anos para a recorrência do evento $\left\{S : Q_{i=4}^{max} > Q_0\right\}$. Em seguida, a partir do segundo 'sucesso', $\tau_2 = 2$ anos e assim sucessivamente até $\tau_k = 5$ anos de recorrência. Se supusermos, por exemplo,

que $N = 50$ anos e que 5 'sucessos' ocorreram durante esse período, a média aritmética dos tempos de recorrência seria $\bar{\tau} = 10$ anos, implicando que, em média, a vazão Q_0 é superada uma vez a cada 10 anos.

É evidente que a variável 'tempo de recorrência' enquadra-se completamente na definição de uma variável aleatória discreta geométrica e que, portanto, a ela podemos associar as características populacionais dadas pelas equações 4.10 a 4.12. Em particular, podemos definir o *tempo de retorno*, denotado por T e expresso em anos, como o *valor esperado* da variável geométrica 'tempo de recorrência', aqui representada por τ. Com essa definição e usando a equação 4.10, escreve-se que

$$T = E[\tau] = \frac{1}{p} \tag{4.13}$$

O tempo de retorno, portanto, não se refere a um 'tempo cronológico'. De fato, T é uma medida de tendência central dos 'tempos cronológicos', aqui denominados tempos de recorrência. Em outras palavras, o tempo de retorno T, associado a um certo evento de referência de um processo de Bernoulli indexado em anos, corresponde ao *tempo médio necessário* (em anos) *para que o evento recorra, em um ano qualquer, e é igual ao inverso da probabilidade de que tal evento de referência ocorra.*

Em hidrologia, o conceito de tempo de retorno é empregado com muita freqüência no estudo probabilístico de eventos máximos anuais, tais como enchentes ou alturas diárias de precipitação máximas anuais. Tais variáveis aleatórias são contínuas e, portanto, têm seu comportamento definido por funções densidade de

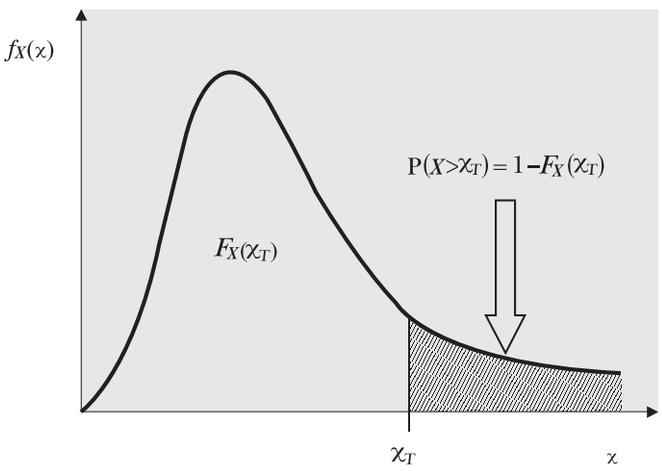


Figura 4.4 – Ilustração do conceito de tempo de retorno para eventos máximos anuais

probabilidades, tais como a ilustrada na Figura 4.4. Se, para a variável X dessa figura, definirmos um quantil de referência x_T, de modo que o 'sucesso' seja a superação de x_T, então, o tempo de retorno T, associado ao quantil de referência, corresponde ao número médio de anos necessário para que o evento $\{X > x_T\}$ recorra uma vez, em um ano qualquer. Pela equação 4.13, o tempo de retorno corresponde ao inverso de $P(X > x_T)$, indicada pela área hachurada, na Figura 4.4.

Exemplo 4.3 – Considere a situação descrita no Exemplo 3.6 do capítulo 3. Determine (a) o tempo de retorno da vazão $X = 300$ m^3/s e (b) a vazão de tempo de retorno $T = 50$ anos.
Solução: (a) A variável X, nesse caso, refere-se a vazões máximas anuais e, portanto, o tempo de retorno é igual ao inverso da probabilidade de superação. De volta ao Exemplo 3.6, já havia sido determinada que $P(X > 300) = 0{,}083$. Logo, o tempo de retorno de $X = 300$ m^3/s é $T = 1/0{,}083 = 12{,}05$ anos. (b) A vazão de tempo de retorno $T = 50$ anos encontra-se em algum ponto X_{50}, entre 300 e 400 m^3/s. Suponha que a ordenada da função densidade nesse ponto seja denotada por w. A primeira equação a ser escrita é $(400\text{-}X_{50}).w/2 = 1/50$. A segunda equação decorre da semelhança entre o triângulo formado pelas vazões 300 e 400, e a densidade no ponto 300, e o triângulo definido por X_{50}, 400 e a densidade no ponto X_{50}, ou seja, $[(400\text{-}300)/z] = [(400\text{-}X_{50})/w]$. Sabendo-se que $z = 1/600$ (ver Exemplo 3.6) e combinando as duas equações acima, resulta a seguinte equação do segundo grau: $X_{50}^2 - 800X_{50} + 157000 = 0$. Uma das raízes dessa equação é maior do que 400 m^3/s e, portanto, está fora do domínio de definição de X. A outra, resposta do problema, é $X_{50} = 351$ m^3/s.

Um importante desdobramento da noção de tempo de retorno refere-se à definição de *risco hidrológico*, tal como aplicado em projetos de estruturas hidráulicas de controle de cheias. Considerado um quantil de referência X_T, de tempo de retorno T, o risco hidrológico é definido como a probabilidade de que X_T seja igualado ou superado *pelo menos uma vez*, em um período de N anos. Em geral, o quantil de referência X_T corresponde à cheia para a qual foi projetada a estrutura hidráulica, enquanto o período de N anos corresponde à sua vida útil. Uma das possíveis deduções da expressão do risco hidrológico, aqui denotado por R, remete-nos à distribuição binomial. Com efeito, a probabilidade de que pelo menos um 'sucesso' ocorra em um período de N anos é equivalente à probabilidade do complemento, 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

$$R = P(Y \geq 1) = 1 - P(Y = 0) = 1 - \binom{N}{0} p^0 (1-p)^{N-0} \qquad (4.14)$$

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 esse resultado na equação 4.14,

$$R = 1 - \left(1 - \frac{1}{T}\right)^N \tag{4.15}$$

Se o risco hidrológico é previamente fixado, em função da importância e das dimensões da estrutura hidráulica, bem como das conseqüências de seu eventual colapso para as populações ribeirinhas ou para as comunidades localizadas a jusante de sua posição no sistema fluvial, pode-se empregar a equação 4.15 para determinar para qual tempo de retorno deve ser calculada a cheia de projeto, por exemplo, do vertedouro de uma barragem, cuja vida útil estimada é de N anos. A Figura 4.5 ilustra tal possibilidade.

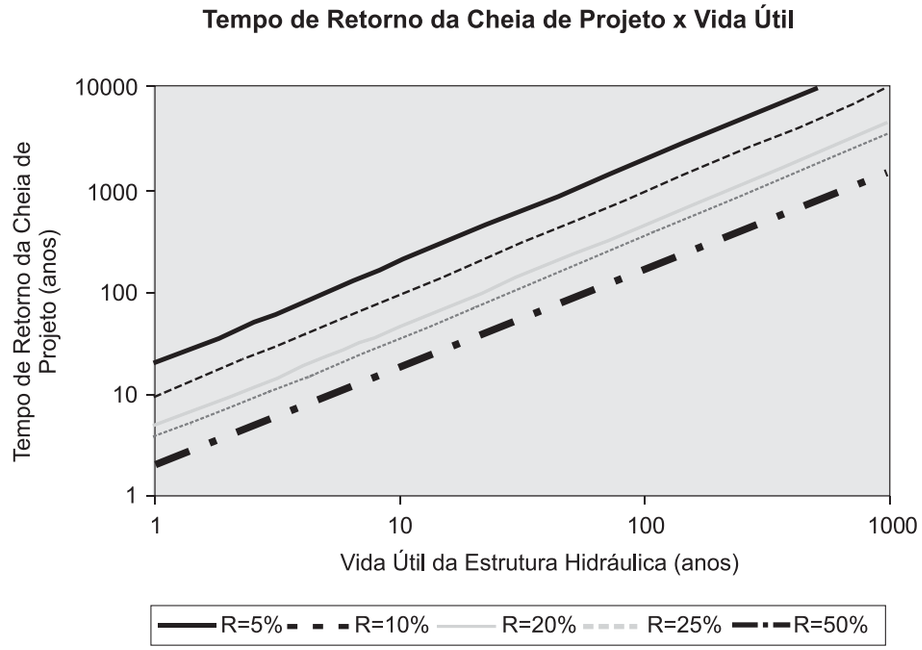


Figura 4.5 – Tempo de retorno da cheia de projeto em função do risco hidrológico e da vida útil estimada para uma estrutura hidráulica

Exemplo 4.4 – A Figura 4.6 mostra o esquema de desvio de um rio durante a construção de uma barragem. Duas ensecadeiras A e B garantem que o canteiro de obras esteja a seco durante o período de construção, enquanto o rio é desviado de seu curso natural por meio de um túnel T, escavado em rocha, pela margem fluvial direita. Suponha que o período de construção é de 5 anos e que

a empresa projetista tenha fixado o risco de 10% para que o canteiro de obras seja inundado pelo menos uma vez nesse período. Com base nesses elementos, determine para qual período de retorno deve ser calculada a cheia de projeto a ser escoada pelo túnel.

Solução:

A inversão da equação 4.15, para T, resulta em $T = \frac{1}{1-(1-R)^{1/N}}$ Com $R = 0{,}10$ e $N = 5$, na equação acima, tem se que T=47,95 anos. Portanto, nesse caso, o túnel **T** deve ter sua seção transversal dimensionada para escoar uma cheia de tempo de retorno igual a aproximadamente 50 anos.

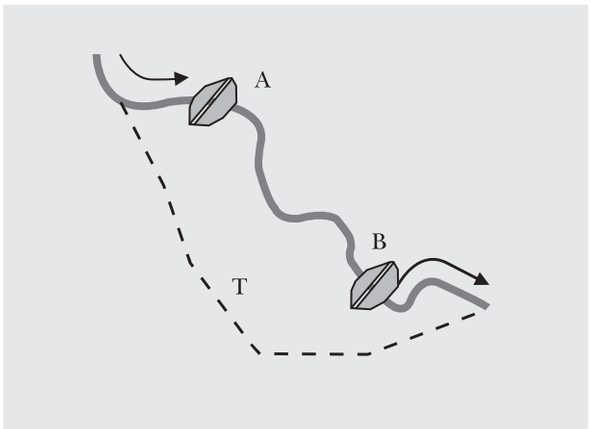


Figura 4.6 – Esquema de Desvio por Túnel

Embora o conceito de tempo de retorno esteja, geralmente, vinculado a eventos máximos anuais, ele também pode ser estendido ao estudo probabilístico de eventos mínimos anuais, tais como vazões médias mensais mínimas anuais. O processo de Bernoulli, nesse caso, é semelhante ao de máximos anuais, porém, o que determina o 'sucesso' é o fato de o evento mínimo anual encontrar-se *abaixo* de um certo valor limiar x_T. O tempo de retorno, nesse caso, passa a ser entendido como o tempo médio, em anos, para que haja a recorrência de uma *estiagem mais severa* do que a definida por x_T, ou seja, a recorrência de um novo evento $\{X < x_T\}$, em um ano qualquer. Supondo que X represente a variável aleatória contínua, característica do evento mínimo anual em questão, verifica-se que, nesse caso, o tempo de retorno T, associado ao quantil de referência, corresponde ao inverso de $P(X < x_T)$, ou seja, ao inverso da função acumulada de probabilidades $F_X(x_T)$. A Figura 4.7 ilustra a extensão do conceito de tempo de retorno aos eventos mínimos anuais, por meio de uma função densidade hipotética $f_X(x)$.

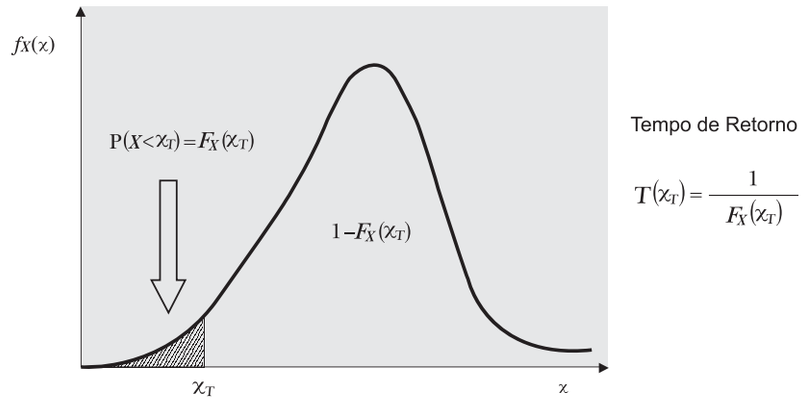


Figura 4.7 – Ilustração do conceito de tempo de retorno para eventos mínimos anuais

4.1.3 – Binomial Negativa

Em um processo de Bernoulli, a variável é denominada *binomial negativa*, quando *Y* refere-se ao número de repetições independentes necessárias para que um certo número *r* de 'sucessos' ocorram. A função massa de probabilidades de uma variável binomial negativa pode ser deduzida a partir da interseção de dois eventos independentes, a saber, o evento A de que o *r*-ésimo 'sucesso' ocorre na *y*-ésima tentativa, com $y \geq r$, e o evento *B* de que ocorrem (*r* - 1) 'sucessos' nas (*y* - 1) tentativas anteriores. O evento *A* ocorre com probabilidade *p* de 'sucesso', em uma tentativa qualquer. Por outro lado, a probabilidade do evento *B* é dada pela distribuição binomial aplicada a (*r* - 1) 'sucessos' em (*y* - 1) tentativas, ou seja, $P(B) = \binom{y-1}{r-1} p^{r-1}(1-p)^{y-r}$. Portanto, $P(A \cap B) = P(A)P(B)$ resulta em

$$p_Y(y) = \binom{y-1}{r-1} p^r (1-p)^{y-r}, \text{ com } y = r, r+1, \ldots \tag{4.16}$$

A equação 4.16 fornece a função massa de probabilidades de uma variável binomial negativa, com parâmetros *r* e *p*; alguns exemplos de funções massa de probabilidades da distribuição binomial negativa encontram-se ilustrados na Figura 4.8. Considerando que essa variável é, de fato, a soma de *r* variáveis geométricas independentes, é fácil demonstrar, pelas propriedades do operador esperança matemática, que o valor esperado e a variância da distribuição binomial negativa são dados, respectivamente, por

$$E[Y] = \frac{r}{p} \tag{4.17}$$

$$\mathrm{Var}(Y) = \frac{r(1-p)}{p^2} \tag{4.18}$$

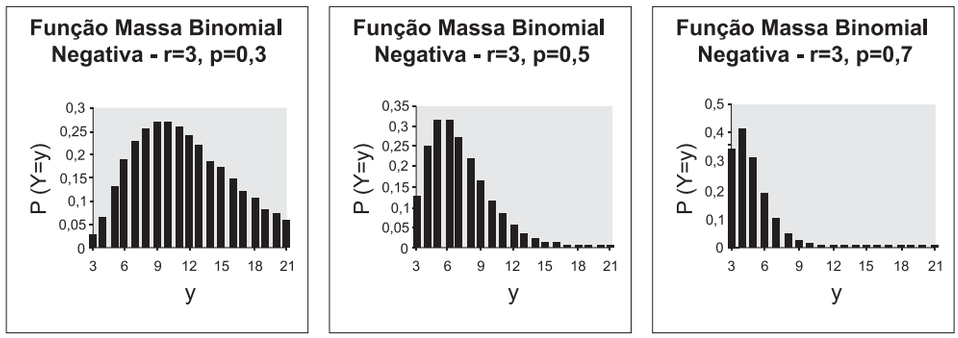


Figura 4.8 - Exemplos de funções massa de probabilidades da distribuição binomial negativa

Exemplo 4.5 – De volta à situação descrita no Exemplo 4.4, suponha que o túnel tenha sido projetado para a cheia de tempo de retorno igual a 10 anos. Pergunta-se (a) qual é a probabilidade de que a segunda inundação do canteiro de obras vá ocorrer no quarto ano de construção? e (b) qual é o risco hidrológico para essa nova situação?

Solução: (a) A probabilidade de que o canteiro de obras vá ser inundado pela segunda vez no quarto ano de construção pode ser calculada diretamente pela equação 4.16, com $r = 2$, $y = 4$ e $p = 1/T = 0{,}10$, ou seja,

$$p_Y(4) = \binom{4-1}{2-1} 0{,}1^{\,2}\ 0{,}9^{4-2} = 0{,}0243 .$$

(b) O novo risco hidrológico, com $N = 5$ e $T = 10$, decorre de aplicação direta da equação $R = 1 - \left(1 - \frac{1}{T}\right)^N = 1 - 0{,}90^5 = 0{,}41$ e, portanto, é exageradamente alto para a situação descrita.

4.2 – Processos de Poisson

Os *processos de Poisson* estão entre os mais importantes processos estocásticos. Na presente publicação, eles são abordados como um *caso limite* de um processo de Bernoulli que se desenvolve em uma escala de tempo, embora possam ser aplicados ao longo de um comprimento, ou de uma área, ou de um volume. Considere um *intervalo de tempo* de comprimento t, o qual é subdividido em N subintervalos de

comprimento t / N. Suponha que cada subintervalo é suficientemente pequeno para que a probabilidade de *mais de uma ocorrência* de um certo evento S, no tempo t/N, seja considerada desprezível, quando comparada à probabilidade p de apenas uma *única ocorrência* do evento S nesse intervalo. Considere ainda que a probabilidade p é constante para cada um dos subintervalos. Finalmente, suponha que o *número médio de ocorrências* do evento S, em um intervalo de tempo qualquer, é proporcional ao comprimento de tal intervalo e que a constante de proporcionalidade é dada por λ; sob tais condições, é possível escrever que $p = \lambda t / N$.

O número de ocorrências Y do evento S, em um tempo t, é igual ao número de subintervalos, nos quais se registrou a ocorrência de S. Se considerarmos tais subintervalos como uma seqüência de N experimentos independentes de Bernoulli, pode-se escrever

$$p_Y(y) = \binom{N}{y}\left(\frac{\lambda t}{N}\right)^y\left(1-\frac{\lambda t}{N}\right)^{N-y} \tag{4.19}$$

Se, nessa expressão, fizermos $p = \lambda t / N$ suficiente pequeno e N suficiente grande, de modo que $Np = \lambda t$, é possível demonstrar que

$$\lim_{N\to\infty}\binom{N}{y}\left(\frac{\lambda t}{N}\right)^y\left(1-\frac{\lambda t}{N}\right)^{N-y} = \frac{(\lambda t)^y}{y!}e^{-\lambda t}, \text{ para } y = 0,1,\ldots \text{ e } \lambda t > 0 \tag{4.20}$$

Fazendo $\nu = \lambda t$ na equação 4.20, chega-se à função massa de probabilidade de Poisson, dada por

$$p_Y(y) = \frac{\nu^y}{y!}e^{-\nu}, \text{ para } y = 0,1,\ldots \text{ e } \nu > 0 \tag{4.21}$$

na qual o parâmetro ν representa o *número médio de ocorrências por intervalo de tempo*. A função de probabilidades acumuladas de Poisson é dada pela seguinte expressão:

$$P_Y(y) = \sum_{i=0}^{y}\frac{\nu^i}{i!}e^{-\nu}, \text{ para } y = 0,1,\ldots \tag{4.22}$$

Conforme demonstrado no Exemplo 3.14 do capítulo 3, a média e a variância de uma variável discreta de Poisson são expressos por

$$E[Y] = \nu \text{ ou } E[Y] = \lambda t \tag{4.23}$$

$$\text{Var}[Y] = \nu \text{ ou } \text{Var}[Y] = \lambda t \tag{4.24}$$

Analogamente à determinação de $E[Y]$ e $\text{Var}[Y]$, demonstra-se que o coeficiente de assimetria da distribuição de Poisson é

$$\gamma = \frac{1}{\sqrt{\nu}} \; ou \; \gamma = \frac{1}{\sqrt{\lambda t}} \tag{4.25}$$

A Figura 4.9 fornece alguns exemplos de funções massa de probabilidades de Poisson.

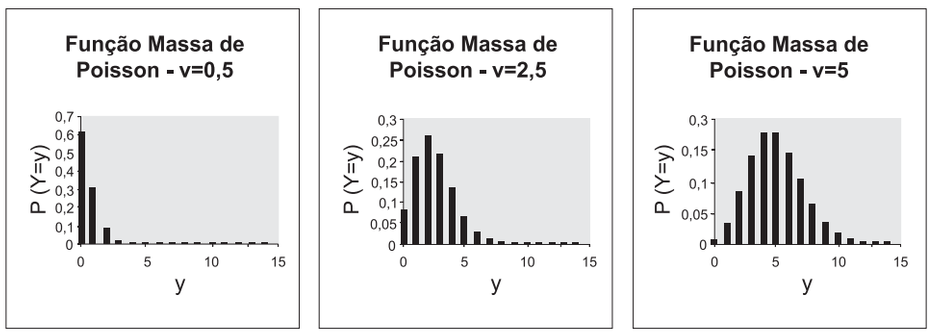


Figura 4.9 - Exemplos de funções massa de probabilidades de Poisson

O parâmetro ν representa o *número médio de ocorrências* de Poisson em um intervalo de tempo *t*; a constante de proporcionalidade λ é denominada *intensidade de Poisson* e representa a razão média de ocorrência dos eventos por intervalo de tempo. Os processos estocásticos construídos com base nas premissas mencionadas recebem o nome de processos de Poisson. Apesar de terem sido deduzidos como caso limite da distribuição binomial, os processos de Poisson referem-se a uma escala de tempo *contínua*. Se ao longo dessa escala contínua, ν e λ forem constantes, os processos de Poisson são considerados homogêneos ou estacionários; caso contrário, para os processos de Poisson não homogêneos, $\lambda(t)$ é uma função do tempo e o número médio de ocorrências ν, em um intervalo $[t_1,t_2]$, será dado pela integral de $\lambda(t)$ nesse intervalo.

Depreende-se da dedução da distribuição de Poisson que ela pode ser usada como uma aproximação da distribuição binomial, desde que *N* seja suficientemente grande e *p* suficientemente pequeno. Na prática, é possível aproximar a binomial pela distribuição de Poisson, com parâmetro $\nu = N.p$, para valores de $N > 20$ e $p < 0{,}1$. Essa aproximação apresenta a vantagem de não exigir a especificação de *N*; de fato, desde que a probabilidade de 'sucesso' *p* seja suficientemente pequena, basta prescrever o número médio de ocorrências por intervalo de tempo. A exemplo da distribuição binomial, a propriedade aditiva também se aplica à distribuição de Poisson, ou seja, se as variáveis Y_1 e Y_2 seguem a distribuição de Poisson, com seus respectivos parâmetros λ_1 e λ_2, então (Y_1+Y_2) também é uma variável de Poisson com parâmetro $\lambda_1 + \lambda_2$.

Exemplo 4.5 – Embarcações chegam a uma eclusa à razão média de 4/hora. Se a chegada de embarcações é um processo de Poisson, calcule (a) a probabilidade de que 6 barcos cheguem em 2 horas; e (b) a probabilidade de que o operador da eclusa possa se ausentar por 15 minutos sem que nenhum barco chegue nesse intervalo. (adap. de Haan, 1977)

Solução : a) $\lambda = 4 \text{ horas}^{-1}$ e $t = 2$ horas $\Rightarrow \lambda t = \nu = 8$. Portanto,

$$P(X = 6) = p_X(6) = (8)^6 \frac{e^{-8}}{6!} = 0{,}1221$$

b) Para que o operador da eclusa possa se ausentar por 15 minutos, nenhuma embarcação pode ter chegado nesse intervalo. Trata-se, portanto, de calcular a probabilidade de nenhuma embarcação haver chegado à eclusa no intervalo de 0,25 horas. Para $\lambda = 4 \text{ horas}^{-1}$ e $t = 0{,}25$ hora $\lambda t = \nu = 1$

$$P(X = 0) = p_X(0) = (1)^0 \frac{e^{-1}}{0!} = 0{,}3679$$

4.3 – Outras Distribuições de Variáveis Aleatórias Discretas

Existem outras distribuições de variáveis aleatórias discretas que não se enquadram entre aquelas apropriadas à modelação de variáveis típicas dos processos de Bernoulli e Poisson. Destacaremos aqui duas dessas distribuições: a *hipergeométrica* e a *multinomial*.

4.3.1 – Distribuição Hipergeométrica

Suponha um conjunto com N itens, dos quais A possuem um certo atributo a (por exemplo, de cor azul ou de sinal positivo ou de alta qualidade, etc.) e (N-A) possuem o atributo b (por exemplo, de cor vermelha ou de sinal negativo ou de baixa qualidade, etc.). Considere que uma amostra contendo n itens, sorteados sem reposição, será retirada do conjunto de N itens. Finalmente, considere que a variável aleatória discreta Y refere-se ao número de itens com atributo a, contidos na amostra de n itens. A probabilidade de que Y seja igual a y itens do tipo a, é dada pela distribuição hipergeométrica, cuja função massa de probabilidades, com parâmetros N, A e n, é expressa por

$$p_Y(y) = \frac{\binom{A}{y}\binom{N-A}{n-y}}{\binom{N}{n}}, \text{ com } 0 \le y \le A;\ y \le n;\ y \ge A - N + n \qquad (4.26)$$

A função acumulada de probabilidades da distribuição hipergeométrica é dada pela seguinte equação:

$$P_Y(y) = \sum_{i=0}^{y} \frac{\binom{A}{i}\binom{N-A}{n-i}}{\binom{N}{n}} \tag{4.27}$$

O denominador da equação 4.26 fornece o número total de possibilidades de se sortear uma amostra de tamanho *n*, a partir de um conjunto de *N* itens. O numerador, por outro lado, fornece o número de possibilidades de sortear amostras de *y* itens de atributo *a*, forçando os (*n*-*y*) itens restantes a terem o atributo *b*. Demonstra-se que o valor esperado e a variância de uma variável hipergeométrica são dados, respectivamente, por

$$E[Y] = \frac{n\,A}{N} \tag{4.28}$$

$$\mathrm{Var}[Y] = \frac{nA(N-A)(N-n)}{N^2(N-1)} \tag{4.29}$$

Se $n < 0{,}1N$, a variável hipergeométrica pode ser aproximada por uma distribuição binomial com parâmetros n e $p = A/N$.

Exemplo 4.6 – Suponha que durante o mês de Fevereiro de 1935, ocorreram 18 dias chuvosos em Ponte Nova do Paraopeba. Suponha também que a ocorrência de um dia chuvoso não depende de ter chovido ou não no dia anterior. Se uma amostra de 10 dias é selecionada ao acaso, pergunta-se (a) qual é a probabilidade de que 7 dias dessa amostra sejam chuvosos? e (b) qual é a probabilidade de que pelo menos 6 dias dessa amostra sejam chuvosos?

Solução:

(a) Usando-se a função massa da distribuição hipergeométrica, com $N = 28$, $A = 18$ e $n = 10$, tem-se

$$p_Y(7) = \frac{\binom{18}{7}\binom{28-18}{10-7}}{\binom{28}{10}} = 0{,}2910$$

(b) a probabilidade de que pelo menos 6 dias dessa amostra sejam chuvosos é $P(Y \geq 6) = 1\text{-}P(Y<6) = 1\text{-}P_Y(5)$, ou seja, $P(Y \geq 6) = 1\text{-}p_Y(0) + p_Y(1) - p_Y(2) - p_Y(3) - p_Y(4) - p_Y(5) = 0{,}7785$.

4.3.2 – Distribuição Multinomial

A distribuição multinomial é uma generalização da distribuição binomial, para o caso de um experimento que pode produzir *r* resultados $a_1, a_2, \ldots, a_r$, diversos e mutuamente excludentes, cada qual com sua respectiva probabilidade de ocorrência $p_1, p_2, \ldots, p_r$, de modo que $\sum p_i = 1$. As variáveis aleatórias multinomiais são denotadas por $Y_1, Y_2, \ldots, Y_r$; nessa representação, Y_i representa o número de ocorrências do resultado a_i, em uma seqüência de *N* experimentos independentes. A função massa de probabilidades conjuntas da distribuição multinomial é dada por

$$\mathrm{P}(Y_1 = y_1, Y_2 = y_2, \ldots, Y_r = y_r) = p_{Y_1, Y_2, \ldots, Y_r}(y_1, y_2, \ldots, y_r) = \frac{N!}{y_1! y_2! \ldots y_r!} p_1^{y_1} p_2^{y_2} \ldots p_r^{y_r} \quad (4.30)$$

na qual, $\sum y_i = N$ e $N, p_1, p_2, \ldots, p_r$ são parâmetros. Cada uma das variáveis Y_i possui uma distribuição marginal binomial com parâmetros *N* e p_i. A média e a variância da distribuição multinomial são dadas pelas seguintes equações:

$$E[Y_i] = N\, p_i \quad (4.31)$$

$$\mathrm{Var}[Y_i] = N\, p_i (1 - p_i) \quad (4.32)$$

Exemplo 4.7 – Em uma certa localidade, os anos são considerados pouco chuvosos (a_1), se a altura pluviométrica anual for inferior a 500 mm e moderadamente chuvosos (a_2), se a altura estiver compreendida entre 500 e 1000 mm. A análise de freqüência dos registros pluviométricos demonstrou que as probabilidades dos eventos com resultados a_1 e a_2 são, respectivamente, 0,4 e 0,5. Considerando um período de 15 anos, calcule a probabilidade de ocorrência de 3 anos pouco chuvosos e 9 anos moderadamente chuvosos.

Solução: Para completar o espaço amostral, temos que definir o terceiro evento, com resultado a_3, correspondente aos anos excepcionalmente chuvosos com alturas superiores a 1000 mm; a probabilidade desse evento é 1-0,4-0,5 = 0,1. Dos 15 anos, se 3 correspondem ao resultado a_1 e 9 ao a_2, então apenas 3 correspondem ao resultado a_3. A probabilidade pedida é dada por

$$\mathrm{P}(Y_1 = 3,\ Y_2 = 9,\ Y_3 = 3) = \mathrm{p}_{Y_1, Y_2, Y_3}(3,9,3) = \frac{15!}{3!9!3!}\, 0{,}4^3 0{,}5^9 0{,}1^3 = 0{,}0125$$

4.4 – Sumário das Características Principais das Distribuições

Apresenta-se a seguir um sumário das principais características das seis distribuições de probabilidades de variáveis aleatórias discretas, descritas no presente capítulo. Nem todas as características que constam desse sumário foram discutidas ou demonstradas no texto principal, embora os princípios para calculá-las sejam os mesmos daqueles descritos nos capítulos anteriores. Portanto, a intenção desse sumário é a de ser um item de referência para uso das distribuições de variáveis aleatórias discretas.

4.4.1 – Distribuição Binomial

Notação: $Y \sim B(N, p)$

Parâmetros: N (inteiro positivo), $0 < p < 1$

FMP: $p_Y(y) = \binom{N}{y} p^y (1-p)^{N-y}$, $y = 0, 1, \ldots, N$

Média: $E[Y] = N\,p$

Variância: $\mathrm{Var}[Y] = N\,p(1-p)$

Coeficiente de Assimetria: $\gamma = \dfrac{1-2p}{\sqrt{N\,p(1-p)}}$

Curtose: $\kappa = 3 + \dfrac{1-6\,p(1-p)}{N\,p(1-p)}$

Função Geratriz de Momentos: $\phi(t) = \left(p\,e^t + 1 - p\right)^N$

4.4.2 – Distribuição Geométrica

Notação: $Y \sim Ge(p)$

Parâmetros: p $(0 < p < 1)$

FMP: $p_Y(y) = p(1-p)^{y-1}$, $y = 1, 2, 3, \ldots$

Média: $E[Y] = \dfrac{1}{p}$

Variância: $\mathrm{Var}[Y] = \dfrac{1-p}{p^2}$

Coeficiente de Assimetria: $\gamma = \dfrac{2-p}{\sqrt{1-p}}$

Curtose: $\kappa = 3 + \dfrac{p^2 - 6p + 6}{1-p}$

Função Geratriz de Momentos: $\phi(t) = \dfrac{p\,e^t}{1-(1-p)e^t}$

4.4.3 – Distribuição Binomial Negativa

Notação: $Y \sim BN(r, p)$

Parâmetros: r e p $(0 < p < 1)$

FMP: $p_Y(y) = \binom{y-1}{r-1} p^r (1-p)^{y-r}$, $y = r, r+1, \ldots$

Média: $E[Y] = \dfrac{r}{p}$

Variância: $\mathrm{Var}(Y) = \dfrac{r(1-p)}{p^2}$

Coeficiente de Assimetria: $\gamma = \dfrac{2-p}{\sqrt{r(1-p)}}$

Curtose: $\kappa = 3 + \dfrac{p^2 - 6p + 6}{r(1-p)}$

Função Geratriz de Momentos: $\phi(t) = \left[\dfrac{p\,e^t}{1-(1-p)\,e^t}\right]^r$

4.4.4 – Distribuição de Poisson

Notação: $Y \sim P(\nu)$

Parâmetros: $\nu\,(\nu > 0)$

FMP: $p_Y(y) = \dfrac{\nu^y}{y!} e^{-\nu}$, $y = 0, 1, \ldots$

Média: $E[X] = \nu$

Variância: $\mathrm{Var}[X] = \nu$

Coeficiente de Assimetria: $\gamma = \sqrt{\frac{1}{\nu}}$

Curtose: $\kappa = 3 + \frac{1}{\nu}$

Função Geratriz de Momentos: $\phi(t) = exp\left[\nu\left(e^{t} - 1\right)\right]$

4.4.5 – Distribuição Hipergeométrica

Notação: $Y \sim H(N, A, n)$

Parâmetros: N, A e n (inteiros positivos)

FMP: $p_Y(y) = \frac{\binom{A}{y}\binom{N-A}{n-y}}{\binom{N}{n}}$, com $0 \le y \le A$; $y \le n$; $y \ge A - N + n$

Média: $E[Y] = \frac{n\,A}{N}$

Variância: $\mathrm{Var}[Y] = \frac{nA(N-A)(N-n)}{N^2(N-1)}$

Coeficiente de Assimetria: $\gamma = \frac{(N-2A)(N-2n)\sqrt{N-1}}{(N-2)\sqrt{nA(N-A)(N-n)}}$

Curtose: $\kappa = \left[\frac{N^2(N-1)}{n(N-2)(N-3)(N-n)}\right]\left[\frac{N(N+1)-6N(N-n)}{A(N-A)} + \frac{3n(N-n)(N+6)}{N^2} - 6\right]$

Função Geratriz de Momentos: sem forma analítica

4.4.6 – Distribuição Multinomial

Notação: $Y_1, Y_2, \ldots, Y_r \sim M(N, p_1, p_2, \ldots, p_r)$

Parâmetros: N, $y_1, y_2, \ldots, y_r$ (inteiros positivos) e $p_1, p_2, \ldots, p_r$ ($p_i > 0$ e $\sum p_i = 1$)

FMP: $p_{Y_1,Y_2,\ldots,Y_r}(y_1, y_2, \ldots, y_r) = \dfrac{N!}{y_1! y_2! \ldots y_r!} p_1^{y_1} p_2^{y_2} \ldots p_r^{y_r}$

Média (marginal): $E[Y_i] = N\, p_i$

Variância (marginal): $\mathrm{Var}[Y_i] = N\, p_i (1 - p_i)$

Coeficiente de Assimetria (marginal): $\gamma(Y_i) = \dfrac{1 - 2p_i}{\sqrt{N\, p_i (1 - p_i)}}$

Curtose (marginal): $\kappa(Y_i) = 3 + \dfrac{1 - 6p_i (1 - p_i)}{N p_i (1 - p_i)}$

Função Geratriz de Momentos: $\phi(t) = \left[\sum_{i=1}^{r} p_i\, e^{t_i} \right]^N$

Exercícios

1) Considere uma distribuição binomial, com $N = 20$ e $p = 0{,}1$, e sua aproximação pela distribuição de Poisson, com $\nu = 2$. Faça um gráfico com as duas funções massa de probabilidades e compare as diferenças.

2) Refaça o exercício 1, (a) com $N = 20$ e $p = 0{,}6$ e (b) com $N = 8$ e $p = 0{,}1$.

3) Suponha que as concentrações médias diárias de um certo poluente, em um determinado trecho de rio, sejam independentes entre si. Se 0,15 é a probabilidade de que a concentração do poluente exceda 6 mg/m^3 em um dia qualquer, estime: (a) a probabilidade de que a concentração exceda 6 mg/m^3 exatamente duas vezes nos próximos 3 dias e (b) a probabilidade de que a concentração exceda 6mg/m^3 no máximo duas vezes nos próximos 3 dias.

4) Se um dique marginal foi projetado para conter a cheia de 20 anos de tempo de retorno, calcule (a) a probabilidade de que a área protegida será inundada pelo menos uma vez durante os próximos 10 anos; (b) a probabilidade de que a área protegida será inundada pelo menos três vezes durante os próximos 10 anos; e (c) a probabilidade de que a área protegida será inundada não mais de três vezes durante os próximos 10 anos.

5) Suponha que a vida útil de uma bacia de detenção para controle de cheias em uma área urbana seja de 25 anos. Pergunta-se (a) qual deve ser o tempo de retorno da cheia de projeto, de modo que exista uma probabilidade de 90% de

que ela não seja superada ao longo da vida útil da bacia de detenção; e (b) qual deve ser o tempo de retorno da cheia de projeto, de modo que exista uma probabilidade de 75% de que ela seja superada no máximo uma vez ao longo da vida útil da bacia de detenção.

6) Três diques marginais foram construídos ao longo dos rios A e B, para controlar eventuais cheias na planície situada entre os dois cursos d'água, tal como mostra a Figura 4.10. Os diques foram projetados do seguinte modo: a cheia de projeto do dique 1 tem tempo de retorno 10 anos; a do dique 2, tem tempo de retorno 20 anos; e para o dique 3, $T = 25$ anos. Supondo que a ocorrência de cheias nos dois rios, assim como a ocorrência de falhas dos diques 1 e 2, são estatisticamente independentes, pede-se (a) calcular a probabilidade anual de inundação da planície, causada exclusivamente pelo rio A; (b) calcular a probabilidade anual de inundação da planície; (c) calcular a probabilidade de não ocorrer nenhuma inundação da planície, em 5 anos consecutivos; e (d) considerando um período de 5 anos consecutivos, calcular a probabilidade de que a terceira inundação da planície irá ocorrer no quinto ano. (adap. de Ang e Tang, 1975)

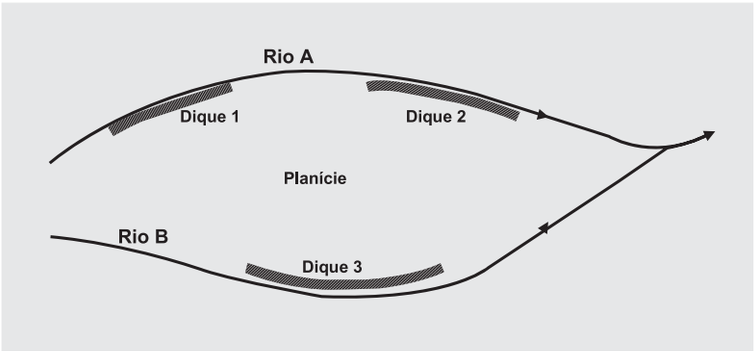


Figura 4.10 - Exercício 6

7) Considere que uma ETA recebe água bruta de um manancial de superfície, captada por uma tomada d'água simples, instalada em determinada cota. Suponha que a variável aleatória discreta X represente o número anual de dias em que o nível d'água, medido na estação fluviométrica local, é inferior à cota da tomada d'água de projeto. Com base em 20 anos de observações, determinou-se a distribuição empírica de probabilidades de X, a qual é dada pela Tabela 4.1. Supondo que o valor esperado possa ser estimado pela média das observações, ajuste uma distribuição de Poisson à variável X. Desenhe, em um único gráfico, as distribuições empírica e de Poisson, e compare os resultados. A distribuição de Poisson é um modelo adequado para a variável em questão? Calcule, pela distribuição de Poisson, a probabilidade que X esteja compreendido entre 3 e 6 dias.

Tabela 4.1 - Exercício 7

$x \rightarrow$	0	1	2	3	4	5	6	7	8	>8
P(X=x)	0,0	0,06	0,18	0,2	0,26	0,12	0,09	0,06	0,03	0,0

8) Os eventos de cheia são marcados pela rápida ascensão do hidrograma, até a vazão de pico, seguida por um período de recessão, em geral, relativamente mais lento, até uma nova ascensão do hidrograma da cheia subseqüente, e assim por diante, tal como ilustrado na Figura 4.11. Se fixarmos uma certa vazão limiar suficientemente elevada, e.g. Q_0, pode-se definir como 'excesso' (ou 'excedência') a diferença entre a vazão de pico de um hidrograma de cheia e a vazão de referência Q_0. Guardadas certas condições, admite-se, em geral, que os 'excessos' acima de Q_0, ao longo do *tempo contínuo*, são processos de Poisson; de fato, esse é o princípio de construção das chamadas 'séries de duração parcial', a serem detalhadas no capítulo 8. Nesse caso, o *número* de 'excedências', em um intervalo Δt, é uma variável aleatória discreta de Poisson, com intensidade de ocorrência λ. Entretanto, podemos estar interessados na distribuição da variável 'tempo entre ocorrências sucessivas de Poisson', e.g. t_1 na Figura 4.11; observe que, nesse caso, a variável aleatória $t \geq 0$ é contínua. De fato, a distribuição de probabilidade de t é a distribuição *exponencial*, cuja função densidade de probabilidade, em função do parâmetro λ, é $f_T(t) = \lambda.\exp(-\lambda t)$. Pede-se demonstrar tal fato, a partir unicamente da distribuição de Poisson. [sugestão: pode-se calcular inicialmente $F_T(t)$ a partir da consideração de que $P(T \leq t) = 1\text{-}P(T > t)$ e que $P(T > t)$ é equivalente a nenhuma ocorrência de Poisson no tempo t; em seguida, a derivada de $F_T(t)$, em relação a t, dá a densidade $f_T(t)$]

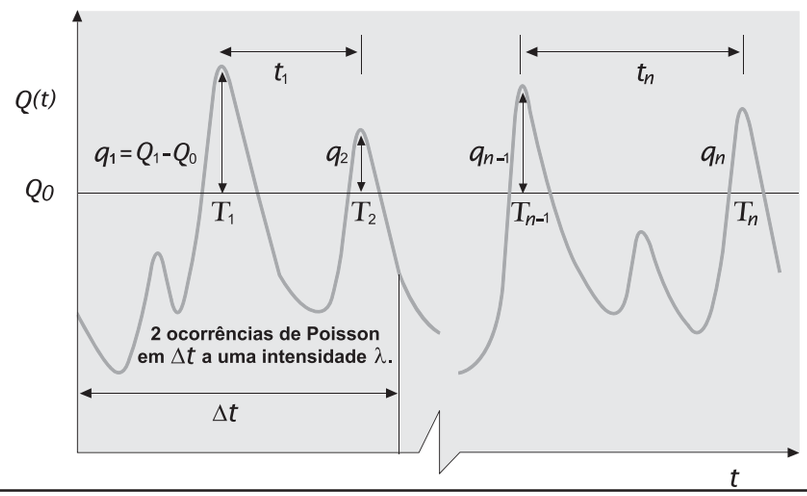


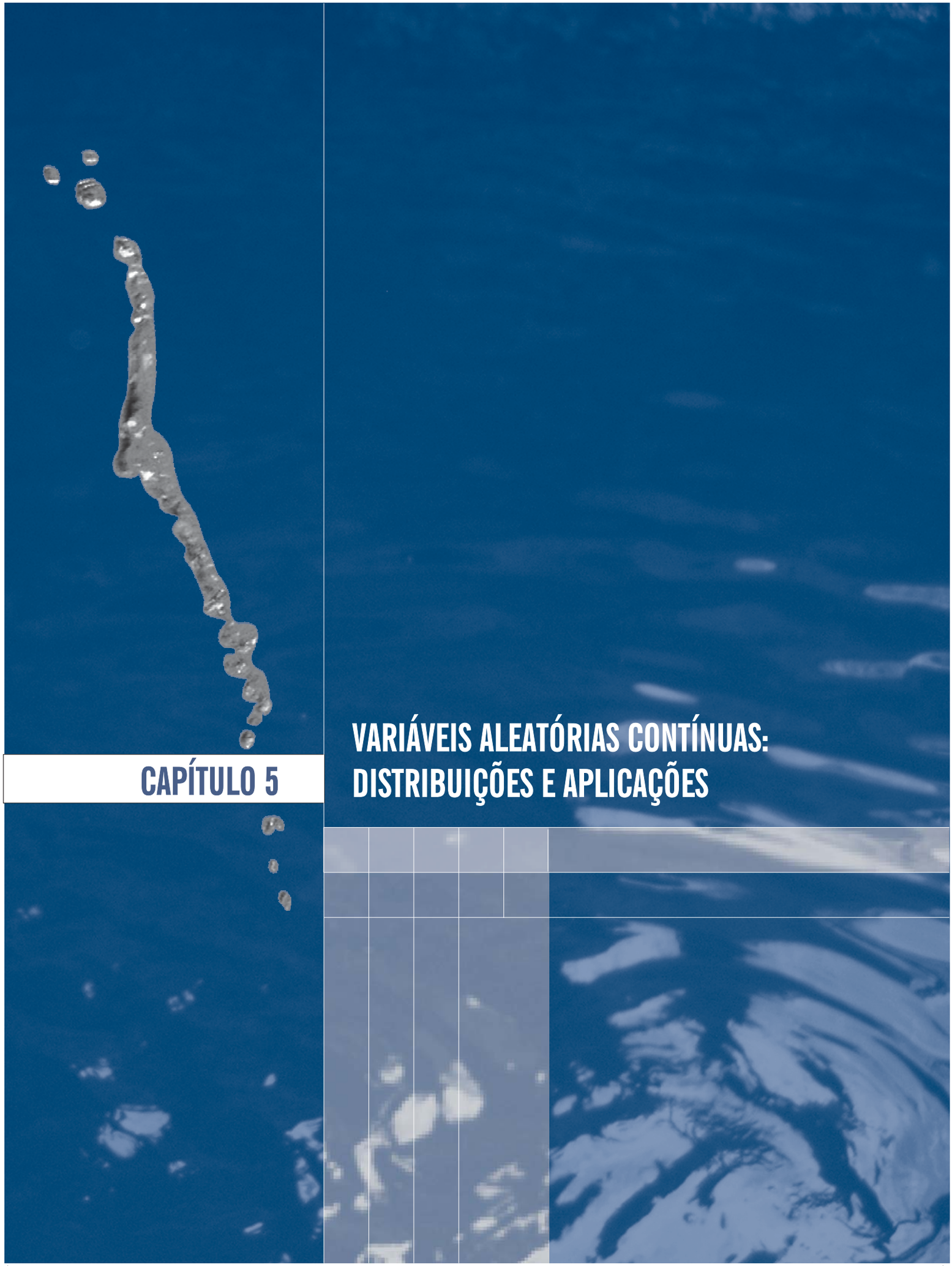
Figura 4.11 - Exercício 8

9) Com referência à Figura 4.11 e ao resultado do exercício 8, é possível deduzir a distribuição de probabilidades do tempo *t* para a *n*-ésima ocorrência de Poisson, a partir da observação de que $t = t_1 + t_2 + ... + t_n$, ou seja, que *t* é a soma de *n* variáveis exponenciais. De fato, a distribuição de *t* tem como densidade $f_T(t) = \lambda^n t^{n-1} e^{-\lambda t} / (n-1)!$; essa é a densidade Gama, para valores inteiros do parâmetro *n*. Pede-se demonstrar tal fato, a partir do resultado do exercício 8. (sugestão: use o resultado do exercício 8 e os métodos expostos no item 3.7 do capítulo 3, para encontrar a distribuição do tempo para duas ocorrências; em seguida, use a distribuição do tempo para duas ocorrências, para encontrar a distribuição do tempo para três ocorrências. Prossiga até que um padrão de repetição apareça e que o processo de indução possa ser usado para extrair a conclusão desejada).

10) Uma companhia apresentou proposta para fornecimento de ETA's compactas para abastecimento de água em área rural. Com base em experiências prévias, estima-se que 10% das ETA's compactas geralmente apresentam algum tipo de defeito de fabricação. Se a proposta consiste na entrega de 5 ETA's, determinar o número de estações a serem fabricadas tal que haja uma certeza de 95% de que nenhuma ETA defeituosa seja entregue. Supõe-se que a entrega (ou a existência de defeitos) de uma ETA seja independente da entrega (ou da ocorrência de defeitos) das demais.

11) Considere que existam 25 pequenas bacias hidrográficas, consideradas adequadas para um estudo de regionalização de vazões mínimas. O hidrólogo responsável pelo estudo desconhece o fato de que 12 dessas bacias possuem dados fluviométricos inconsistentes. Suponha que, em uma primeira fase do estudo, apenas 10 bacias serão selecionadas. Pede-se (a) calcular a probabilidade de que, entre as 10, sejam selecionadas 3 bacias com dados fluviométricos inconsistentes; (b) calcular a probabilidade de que, entre as 10, pelo menos 3 bacias, com dados fluviométricos inconsistentes, sejam selecionadas; e (c) calcular a probabilidade de que as 10 bacias selecionadas possuam dados fluviométricos inconsistentes.

12) Em uma certa localidade, a probabilidade de que qualquer dia da primeira semana de Janeiro seja chuvoso é de 0,20. Supondo tratar-se de eventos independentes, pede-se calcular (a) a probabilidade de que, em Janeiro de qualquer ano, apenas os dias 2 e 3 serão ambos chuvosos; (b) a probabilidade de se ter uma seqüência de pelo menos dois dias consecutivos com chuva, apenas no período de 4 a 7 de Janeiro de qualquer ano; (c) considerando que *C* represente o número de dias chuvosos do período de 4 dias do item (b), estabeleça a função massa de probabilidades da variável aleatória *C*; (d) $P(C > 2)$ e $P(C \geq 2)$ e (e) os primeiros três momentos centrais da variável *C*. (adap. de Shahin et al., 1993)



CAPÍTULO 5

VARIÁVEIS ALEATÓRIAS CONTÍNUAS: DISTRIBUIÇÕES E APLICAÇÕES

CAPÍTULO 5 VARIÁVEIS ALEATÓRIAS CONTÍNUAS: DISTRIBUIÇÕES E APLICAÇÕES

Os modelos de distribuição de probabilidades a serem discutidos nesse capítulo referem-se à modelação de variáveis aleatórias contínuas. Dentre tais modelos, destacaremos, aqui, aqueles que apresentam uma utilidade mais freqüente na análise de freqüência de variáveis hidrológicas, incluindo exemplos de suas respectivas aplicações. Também serão descritas distribuições de probabilidade de estatísticas amostrais que possuem utilidade na formulação e construção de intervalos de confiança e testes estatísticos de hipóteses, os quais serão abordados no capítulo 7. Daremos ênfase à descrição das principais características e às aplicações dos modelos distributivos, sem a preocupação de apresentar provas matemáticas para resultados de valores esperados e outras medidas populacionais. Ao final desse capítulo, apresenta-se também uma breve descrição da distribuição Normal bivariada, como uma ilustração dos modelos probabilísticos contínuos multivariados.

5.1 – Distribuição Uniforme

Uma variável aleatória contínua X, cujos valores possíveis x encontram-se restritos à condição $a \le x \le b$, é *distribuída uniformemente* se a probabilidade de que ela esteja compreendida em qualquer intervalo $[m, n]$, contido em $[a, b]$, for diretamente proporcional ao comprimento $(m\text{-}n)$. Se a constante de proporcionalidade for denotada por ρ, então,

$$P(m \le X \le n) = \rho(m - n) \quad \text{se } a \le m \le n \le b \tag{5.1}$$

Uma vez que $P(a \le X \le b) = 1$, é fácil verificar que $\rho = 1/(b - a)$. Portanto, para qualquer $a \le x \le b$, a função de probabilidades acumuladas da distribuição uniforme é dada por

$$F_X(x) = \frac{x - a}{b - a} \tag{5.2}$$

Se $x < a, F_X(x) = 0$ e, se $x > b$, $F_X(x) = 1$. A função densidade da distribuição uniforme decorre da diferenciação da equação 5.2 e tem a seguinte expressão:

$$f_X(x) = \frac{1}{b - a} \quad \text{se } a \le x \le b \tag{5.3}$$

A Figura 5.1 ilustra as funções densidade e de probabilidades acumulada s da distribuição uniforme.

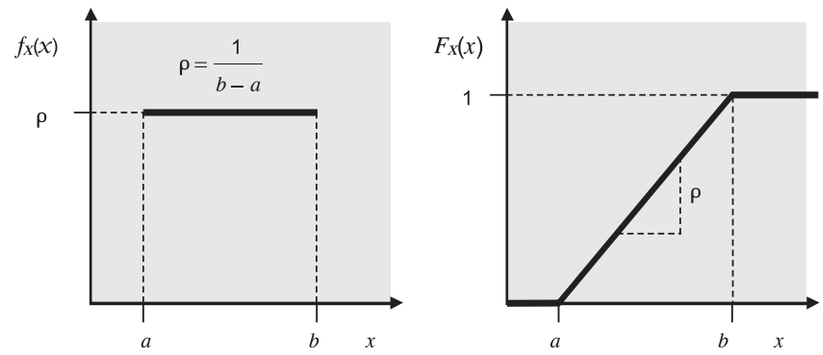


Figura 5.1 – Funções densidade e de probabilidades acumuladas da distribuição uniforme

A média e a variância de uma variável aleatória uniforme são dadas, respectivamente, por

$$E[X] = \frac{a+b}{2} \tag{5.4}$$

$$\mathrm{Var}[X] = \frac{(b-a)^2}{12} \tag{5.5}$$

Quando o intervalo de definição da variável X é fixado em [0,1], a distribuição uniforme encontra sua maior aplicação que é a de representar a distribuição de $X = F_Y(y)$, onde $F_Y(y)$ denota um *modelo distributivo qualquer* para a variável aleatória contínua Y. Com efeito, como $0 \le [\ F_Y(y) = \mathrm{P}(Y \le y)\ ] \le 1$ para qualquer distribuição de probabilidades, $X = F_Y(y)$ pode ser vista como uma variável aleatória uniforme no intervalo [0,1]. Esse fato é utilizado para *gerar números aleatórios uniformes x*, no intervalo [0,1], os quais, em seguida, podem ser empregados para obter números $y = F^{-1}_{Y=X}(y)$, distribuídos de acordo com a distribuição $F_Y(y)$, desde que a inversa dessa função exista e possa ser expressa analiticamente. A geração de números aleatórios uniformes é essencial para a simulação de um grande número de diferentes conjuntos de valores de uma variável aleatória, distribuída de acordo com uma certa função densidade de probabilidades, com o propósito de avaliar cenários estatisticamente similares aos observados. Em geral, as técnicas empregadas para gerar tais conjuntos são reunidas sob a denominação '*método da simulação de Monte Carlo*'; o leitor deve remeter-se às referências Ang e Tang (1990) e Kottegoda e Rosso (1997), para detalhes sobre o método de Monte Carlo.

Exemplo 5.1 – Denote por X a temperatura mínima diária em uma certa localidade e suponha que X varie uniformemente no intervalo de 16 a 22°C. Pede-se (a) calcular a média e a variância de X; (b) a probabilidade de X superar 18°C; e (c) dado que, em um certo dia, a temperatura já superou a marca de 18°C, calcular a probabilidade de X superar 20°C.
Solução: (a) A média e a variância decorrem de aplicação direta das equações 5.4 e 5.5, com a =16 e b = 22°C. Portanto, $E[X]$ = 19°C e Var$[X]$ = 3 (°C)^2. (b) P(X>18 °C) =1-P(X< 18 °C) =1-F_X(18) = 2/3. (c) A função densidade de X é $f_x(x) = 1/6$ para o intervalo $16 \leq X \leq 22$. Entretanto, conforme o enunciado, em um certo dia, é um fato que a temperatura já superou a marca de 18°C. Uma vez que o espaço amostral da variável já foi reduzido, pode-se redefinir a nova função densidade $f_x^R(x)=1/(22-18)=1/4$ para o intervalo $18 \leq X \leq 22$, a integral da qual deve ser igual a 1 para os novos limites. Nesse caso, P(X > 20|X > 18) = $-1\ F_x^R(20) = 1-(20-18)/(22-18)=1/2$.

5.2 – Distribuição Normal

A distribuição Normal também é conhecida como de Gauss, em referência ao emprego pioneiro dessa distribuição no tratamento dos erros aleatórios de medidas experimentais, atribuído ao matemático alemão Karl Friedrich Gauss (1777-1855). A distribuição Normal é utilizada para descrever o comportamento de uma variável aleatória que flutua de forma simétrica em torno de um valor central. Algumas de suas propriedades matemáticas, a serem discutidas no presente item, fazem do modelo Normal a distribuição apropriada à modelação de variáveis que resultam da soma de um grande número de outras variáveis independentes. Além disso, a distribuição Normal está na origem de toda a formulação teórica acerca da construção de intervalos de confiança, testes estatísticos de hipóteses, bem como da teoria de regressão e correlação.

A distribuição Normal é um modelo a dois parâmetros, cujas funções densidade e de probabilidades acumuladas são expressas, respectivamente, por

$$f_X(x) = \frac{1}{\sqrt{2\pi\theta_2^2}} exp\left[-\frac{1}{2}\left(\frac{x-\theta_1}{\theta_2}\right)^2\right] para -\infty < x < \infty \tag{5.6}$$

$$F_X(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi\theta_2^2}} exp\left[-\frac{1}{2}\left(\frac{x-\theta_1}{\theta_2}\right)^2\right] dx \tag{5.7}$$

A Figura 5.2 ilustra a forma da distribuição Normal, para o caso em que $\theta_1 = 8$ e $\theta_2 = 1$.

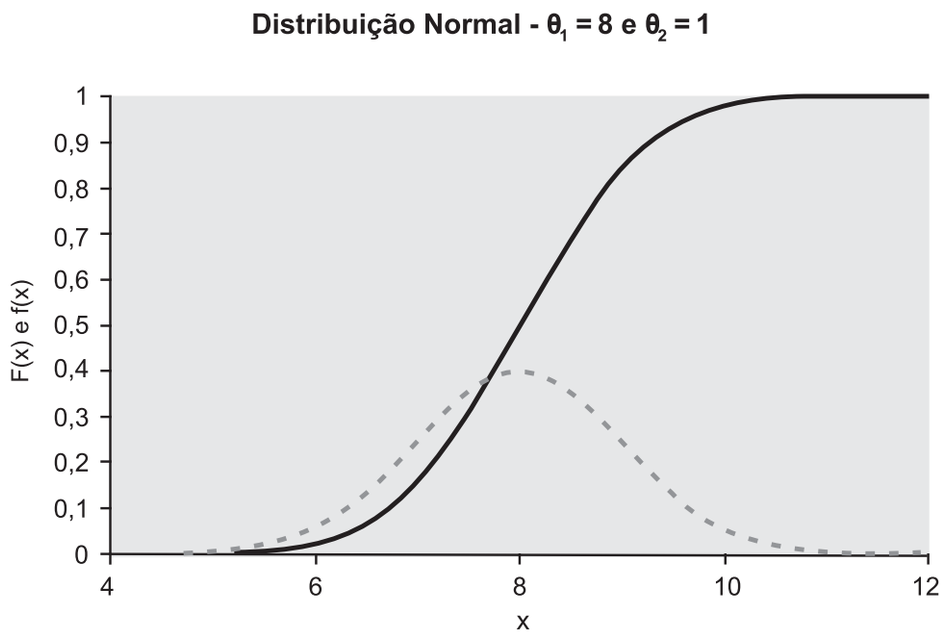


Figura 5.2 – FDP e FAP da distribuição Normal, com $\theta_1 = 8$ e $\theta_2 = 1$

O valor esperado, a variância e o coeficiente de assimetria de uma variável Normal (ver Exemplo 3.15 do capítulo 3), com parâmetros θ_1 e θ_2, são dados respectivamente por

$$E[X] = \mu = \theta_1 \tag{5.8}$$

$$\mathrm{Var}[X] = \sigma^2 = \theta_2^2 \tag{5.9}$$

$$\gamma = 0 \tag{5.10}$$

Como decorrência desses resultados, a função densidade da distribuição Normal é, em geral, escrita na forma

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma} \right)^2 \right] \text{para} -\infty < x < \infty \tag{5.11}$$

e diz-se que X é normalmente distribuída com média μ e desvio padrão σ, ou, sinteticamente, que $X \sim \mathbf{N}(\mu, \sigma)$. Portanto, a média de uma variável Normal X é igual ao parâmetro de posição, em torno do qual os valores de X se dispersam

simetricamente. O grau com que a variável *X* se dispersa em torno de μ, é dado pelo parâmetro de escala, o qual é igual ao desvio padrão σ. A Figura 5.3 exemplifica os efeitos das variações marginais dos parâmetros de posição e escala da distribuição Normal.

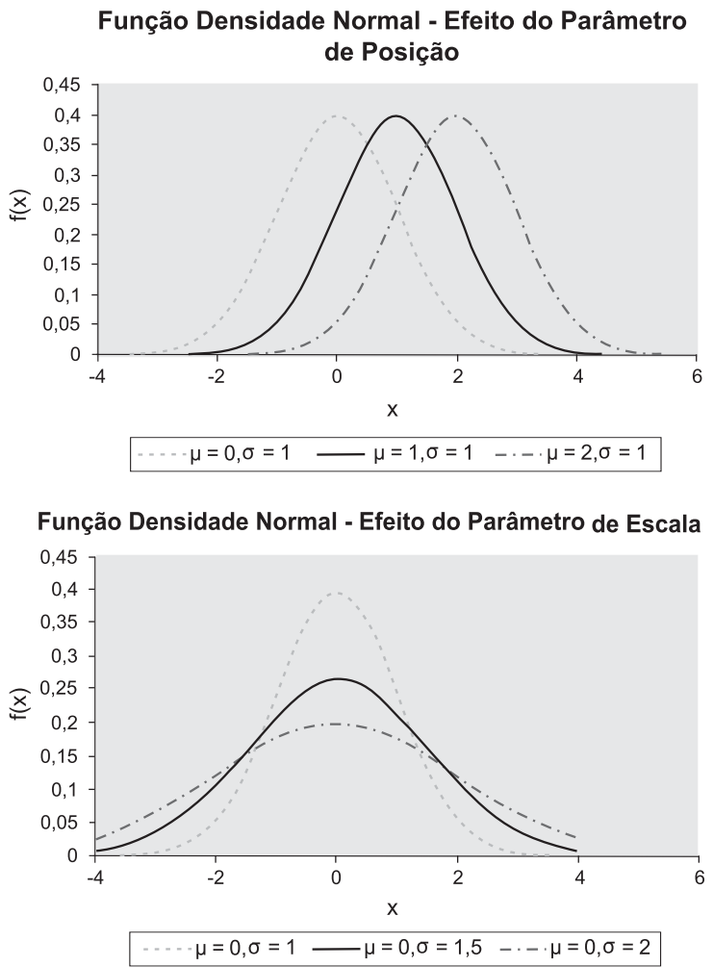


Figura 5.3 – Efeitos da variação marginal dos parâmetros de posição e escala sobre $X \sim \mathbf{N}(\mu, \sigma)$

Empregando os métodos descritos no item 3.7 do capítulo 3, é possível provar que, se $X \sim \mathbf{N}(\mu_X, \sigma_X)$, a variável aleatória $Y = aX + b$, resultante de uma combinação linear de *X*, também é normalmente distribuída com média $\mu_Y = a\mu_X + b$ e desvio padrão $\sigma_Y = a\sigma_X$, ou, sinteticamente, que $Y \sim \mathbf{N}(\mu_Y = a\mu_X + b, \sigma_Y = a\sigma_X)$. Essa propriedade da distribuição Normal, conhecida como *reprodutiva*, pode ser estendida a qualquer combinação linear de *n* variáveis aleatórias *independentes e normalmente distribuídas* $X_i, i = 1,2,...,n,$ cada qual com seus respectivos

parâmetros μ_i e σ_i. De fato, a partir da extensão do resultado obtido no Exemplo 3.19 do capítulo 3, é possível demonstrar que $Y = \sum_{i=1}^{n} a_i X_i + b$ segue uma distribuição Normal com parâmetros $\mu_Y = \sum_{i=1}^{n} a_i \mu_i + b$ e $\sigma_Y = \sqrt{\sum_{i=1}^{n} a_i \sigma_i^2}$. Como caso particular (veja Exemplo 3.18 do capítulo 3), se Y é a *média aritmética* de n variáveis normais X_i, todas com média μ_X e desvio padrão σ_X, então $Y \sim \mathbf{N}\left(\mu_X, \sigma_X / \sqrt{n}\right)$.

A FAP da distribuição Normal, dada pela equação 5.7, não tem solução analítica. Com efeito, cada par de valores específicos dos parâmetros $\theta_1 = \mu$ e $\theta_2 = \sigma$ requer uma *integração numérica* específica para a obtenção da função $F_X(x)$. Esse inconveniente pode ser superado a partir da transformação linear $Z = \frac{X - \mu}{\sigma}$ da variável Normal X, de parâmetros μ e σ. De fato, usando a propriedade reprodutiva da distribuição Normal, para o caso particular em que os coeficientes da transformação linear são $a = 1/\sigma$ e $b = -\mu/\sigma$, é fácil demonstrar que $Z \sim \mathbf{N}\left(\mu_Z = 0,\ \sigma_Z = 1\right)$. A variável Z recebe o nome de *variável Normal central reduzida* e a distribuição de probabilidades de Z é conhecida como *distribuição Normal padrão*, ou distribuição Normal em forma canônica. As funções densidade e de probabilidades acumuladas de Z são dadas, respectivamente, por

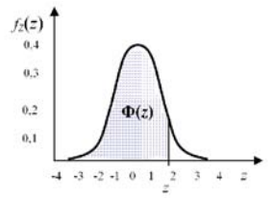
$$f_Z(z) = \frac{1}{\sqrt{2\pi}} exp\left(-\frac{z^2}{2}\right), -\infty < z < \infty \tag{5.12}$$

$$F_Z(z) = \Phi(z) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} exp\left(-\frac{z^2}{2}\right) dz \tag{5.13}$$

A função de probabilidades acumuladas da distribuição normal padrão $\Phi(z)$ pode ser obtida mediante integração numérica. Em geral, os resultados da integração numérica são dispostos em forma tabular, tal como na Tabela 5.1, na qual, aproveitando-se da simetria da distribuição, somente os valores positivos de z são mostrados. Para calcular a probabilidade $P(X \leq x)$, para $X \sim \mathbf{N}(\mu_X, \sigma_X)$, calcula-se primeiramente o valor de $z = (x - \mu_X)/\sigma_X$; em seguida, de posse do valor tabelado de $\Phi(z)$, faz-se $P(X \leq x) = \Phi(z)$. Inversamente, se o objetivo é o de calcular o quantil x, cuja probabilidade de não superação é um dado P, verifica-se, inicialmente na Tabela 5.1, a qual valor de z corresponde $\Phi(z) = P$; em seguida, acha-se o quantil $x = \mu_X + z\,\sigma_X$.

Tabela 5.1 – Função de Probabilidades Acumuladas da Distribuição Normal Padrão

$$F_Z(z)=\Phi(z)=\int_{-\infty}^{z}\frac{1}{\sqrt{2\pi}}e^{-\frac{z^2}{2}}dz$$



z	0,00	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

Exemplo 5.2 – Suponha que as vazões naturais médias anuais Q de um afluente do rio Amazonas sejam normalmente distribuídas com média de 10.000 m^3/s e desvio padrão de 5000 m^3/s. Calcule (a) $P(Q < 5000)$ e (b) a vazão média anual de tempo de retorno $T = 50$ anos.
Solução: (a) A probabilidade $P(Q < 5000)$ pode ser igualada a $P\{z < [(5000-10000) / 5000]\}$, ou seja a $\Phi(-1)$. Como a Tabela 5.1 fornece $\Phi(z)$ apenas para valores positivos de z, deve-se usar a seguinte propriedade de simetria da distribuição Normal: $\Phi(-1) = 1 - \Phi(+1) = 1 - 0{,}8413 = 0{,}1587$. (b) A definição de tempo de retorno pode ser aqui empregada, de modo idêntico ao usado para valores máximos anuais, ou seja, $T = 1/P(Q > q)$. Como $T = 50$ anos, $P(Q > q) = 1/50 = 0{,}02$ e, portanto, $\Phi(z) = 1 - 0{,}02 = 0{,}98$. Na Tabela 5.1, esse valor corresponde a $z = 2{,}054$. Logo, a vazão q de T = 50 anos corresponde ao quantil $q = 10000 + 2{,}054 \times 5000 = 20269$ m^3/s.

O exame da Tabela 5.1 demonstra que 68,26% da área da função densidade da distribuição Normal está compreendida entre os limites de 1 desvio padrão abaixo e acima da média. Do mesmo modo conclui-se que 95,44% da área corresponde ao intervalo $[\mu - 2\sigma, \mu + 2\sigma]$, enquanto 99,74% está compreendida pela área da função densidade entre os limites de $\mu - 3\sigma$ e $\mu + 3\sigma$. Embora uma variável aleatória Normal seja definida entre $-\infty$ e $+\infty$, a ínfima probabilidade de 0,0013 de um valor inferior a $(\mu - 3\sigma)$, demonstra a aplicabilidade dessa distribuição a variáveis hidrológicas não negativas, tais como precipitações e vazões. De fato, se $\mu_X > 3\sigma_X$, a chance de se obter um valor de X negativo é desprezível.

Tanto $\Phi(z)$, como sua inversa, podem ser muito bem aproximadas por funções de fácil implementação em códigos de programação de computadores. A aproximação mais freqüente de $\Phi(z)$, *para* $z \geq 0$, é dada pela seguinte expressão:

$$\Phi(z) \cong 1 - f\left(b_1 t + b_2 t^2 + b_3 t^3 + b_4 t^4 + b_5 t^5\right) \tag{5.14}$$

onde f denota a função densidade Normal e a variável auxiliar t é dada por

$$t = \frac{1}{1 + rz} \tag{5.15}$$

na qual $r = 0{,}2316419$. Os coeficientes b_i do argumento da função densidade são

$$
\begin{aligned}
b_1 &= 0{,}31938153 \\
b_2 &= -0{,}356563782 \\
b_3 &= 1{,}781477937 \\
b_4 &= -1{,}821255978 \\
b_5 &= 1{,}330274429
\end{aligned}
\tag{5.16}
$$

Por outro lado, a inversa $z(\Phi)$, *para* $\Phi \geq 0{,}5$, pode ser aproximada por

$$z \cong m - \frac{c_0 + c_1 m + c_2 m^2}{1 + d_1 m + d_2 m^2 + d_3 m^3} \tag{5.17}$$

onde, a variável auxiliar *m* é dada por

$$m = \sqrt{ln\left[\frac{1}{(1-\Phi)^2}\right]} \tag{5.18}$$

e os coeficientes c_i e d_i são os seguintes:

$$
\begin{aligned}
c_0 &= 2{,}515517 \\
c_1 &= 0{,}802853 \\
c_2 &= 0{,}010328 \\
d_1 &= 1{,}432788 \\
d_2 &= 0{,}189269 \\
d_3 &= 0{,}001308
\end{aligned}
\tag{5.19}
$$

Outra importante aplicação da distribuição Normal decorre do chamado *teorema do limite central*, cuja prova matemática rigorosa é atribuída ao matemático russo Aleksander Liapunov (1857-1918). De acordo com a *versão estrita* desse teorema, se S_n denota a *soma de n variáveis aleatórias* independentes e identicamente distribuídas $X_1, X_2, \ldots, X_n$, todas com média μ e desvio padrão σ, então, a variável dada pela expressão

$$Z_n = \frac{S_n - n\mu}{\sigma\sqrt{n}} \tag{5.20}$$

tende assintoticamente a uma variável Normal central reduzida, i.e., para valores de *n* suficientemente grandes, $Z_n \sim \mathbf{N}(0,1)$. Na prática, se $X_1, X_2, \ldots, X_n$ são, de fato, independentes e com distribuições idênticas, porém não exageradamente assimétricas, em geral, valores de *n* em torno de 30, e até inferiores, já são suficientes para permitir a convergência de Z_n para uma variável Normal padrão.

Como caso particular da propriedade reprodutiva da distribuição Normal, viu-se

que se Y representa a média aritmética de *n variáveis normais* X_i, todas com média μ_X e desvio padrão σ_X, então $Y \sim \mathbf{N}\left(\mu_X, \sigma_X / \sqrt{n}\right)$. A aplicação da equação 5.20 à variável Y (ver Exemplo 5.3), mostra que o teorema do limite central permite que esse mesmo resultado seja obtido, sem a restrição de que as variáveis X_i devam ser variáveis normais. A condição, nesse caso, é imposta pelo número n de componentes X_i, o qual deve ser suficientemente grande para permitir a convergência para uma distribuição Normal. Kottegoda e Rosso (1997) sugerem que se as distribuições dos componentes X_i são moderadamente não-normais, a convergência é relativamente rápida. Entretanto, se os desvios da normalidade são pronunciados, valores de n superiores a 30 podem ser necessários para garantir a convergência.

O teorema do limite central, em sua versão estrita já enunciada, tem pouca aplicação em hidrologia. De fato, é difícil admitir a noção de que uma variável hidrológica seja o resultado da soma de um grande número de variáveis independentes e identicamente distribuídas. Tomemos o exemplo da variável 'altura anual de precipitação', cujo resultado é, de fato, a soma das alturas pluviométricas diárias, medidas em uma certa localidade. Entretanto, supor que as alturas diárias de todos os dias do ano possuam a mesma distribuição de probabilidades, com a mesma média e com o mesmo desvio padrão, não é realista do ponto de vista hidrológico e, portanto, impede a aplicação da versão estrita do teorema do limite central. Por outro lado, o chamado *teorema do limite central generalizado* é flexível o bastante para permitir sua aplicação a algumas variáveis hidrológicas. De acordo com essa versão, se X_i (i=1,2,...,n) denotam variáveis independentes, cada qual com suas respectivas médias e variâncias iguais a μ_i e σ_i^2, então, a variável dada por

$$Z_n = \frac{S_n - \sum_{i=1}^{n} \mu_i}{\sqrt{\sum_{i=1}^{n} \sigma_i^2}} \tag{5.21}$$

tende a uma variável Normal padrão, quando n tende ao infinito, sob a condição de que nenhum dos componentes X_i possua um efeito dominante na soma S_n. Segundo Benjamin e Cornell (1970), Z_n tende a ser normalmente distribuída, quando n tende para o infinito, ainda que os componentes X_i não sejam coletivamente independentes entre si, porém distribuídos conjuntamente de modo que seja nula a correlação entre um componente e a grande maioria dos outros. A importância prática da versão generalizada do teorema central limite reside no fato de que, mantidas as condições gerais enunciadas, a convergência para uma distribuição Normal da soma, ou da média, de um número suficientemente grande de componentes aleatórios pode ser estabelecida sem o conhecimento exato das

distribuições marginais de X_i ou de sua distribuição conjunta.

A versão generalizada do teorema do limite central já permite alguma aplicação às variáveis hidrológicas. De volta ao exemplo da variável 'altura anual de precipitação', é plausível a suposição de que, em uma região de sazonalidade pouco marcada, não haja um efeito dominante de uma ou de algumas alturas pluviométricas, de um ou de alguns dias específicos do ano, sobre o total anual. Exceção feita à prevalência de precipitações de origem frontal, também é plausível admitir-se a hipótese de independência de, pelo menos grande parte, dos componentes X_i. Portanto, sob tais condições particulares e supondo que n =365 (ou 366) seja um número suficientemente grande para permitir a convergência, a qual, de fato, irá depender da forma das distribuições individuais dos componentes, é possível admitir-se que as alturas anuais de precipitação possam ser descritas pela distribuição Normal. Usando argumentos similares, porém ressalvando a maior dependência estatística entre os componentes X_i, é possível admitir também que as *vazões médias anuais* de bacias hidrográficas, localizadas em regiões de sazonalidade pouco marcada, possam ser modeladas por uma distribuição Normal.

Exemplo 5.3 – Deseja-se monitorar as concentrações de oxigênio dissolvido em um trecho fluvial localizado a jusante de um reservatório, cujas funções são de controlar cheias e manter calados mínimos para a navegação. O programa de monitoramento irá consistir de medições semanais sistemáticas de concentração de oxigênio dissolvido (OD) em uma seção transversal já definida. A variável aleatória 'concentração de OD', aqui denotada por X, é fisicamente limitada à esquerda pelo valor 0 e à direita pela concentração de saturação de oxigênio dissolvido (em torno de 9 mg/l), a qual depende da temperatura da água. Suponha que uma campanha de 8 medições semanais resultou em $\bar{x} = 4$ mg/l e $s_X = 2$ mg/l. À luz somente dessas informações, pergunta-se quantas medições semanais devem ser programadas para que a diferença entre a média amostral e a verdadeira média populacional de X seja no máximo de 0,5 mg/l, com uma certeza de 95%.

Solução: Contrariamente a uma variável aleatória Normal, a variável X, nesse caso, é limitada à esquerda e à direita e, em função de sua dependência da vazão no trecho fluvial, sua função densidade é provavelmente assimétrica. Suponha que X_i denote a concentração de OD na i-ésima semana do programa de n semanas de monitoramento. Dado que a seção de monitoramento encontra-se em um trecho de vazões fortemente regularizadas e que o intervalo entre as medições é semanal, é possível supor que as variáveis X_i são independentes entre si e igualmente distribuídas, todas com média μ e desvio padrão σ, mesmo que não sejam conhecidas as respectivas

distribuições marginais. Portanto, é plausível admitir que a soma e, por conseguinte, a média aritmética de *n* variáveis independentes e igualmente distribuídas (IID) tendem a ser normalmente distribuídas, quando *n* é grande o bastante para permitir tal convergência. Em outras palavras, é plausível a aplicação da versão estrita do teorema do limite central. Fazendo a soma das *n* variáveis IID $S_n = n\bar{x}$, onde $\bar{x}$ denota a media aritmética, e substituindo-a na equação 5.20, resulta que $Z_n = \frac{n\bar{x} - n\mu}{\sigma\sqrt{n}} = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \sim \mathrm{N}(0,1)$. Logo, pode-se escrever que $\mathrm{P}\left(z_{2,5\%} \le \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \le z_{97,5\%} \right) = 0{,}95$. A Tabela 5.1 fornece $z_{0,975} = 1{,}96$ e, por simetria, $z_{0,025} = -1{,}96$. Substituindo um desses valores na equação de P(.) e isolando o termo da diferença entre a média amostral e a média populacional, resulta que $\mathrm{P}\left(|\bar{x} - \mu| \le 1{,}96\,\sigma/\sqrt{n} \right) = 0{,}95$. Supondo que σ possa ser estimado por $s_X = 2$ mg/l e lembrando que $|\bar{x} - \mu| = 0{,}5$ mg/l, verifica-se que $(1{,}96 \times 2)/\sqrt{n} \ge 0{,}5$ ou que $n \ge 61{,}47$. Portanto, 62 semanas de monitoramento são minimamente necessárias para que a diferença entre a média amostral e a verdadeira média populacional de *X* seja no máximo de 0,5 mg/l, com uma certeza de 95%.

No capítulo 4, foi visto que a variável discreta binomial, representada por *X* e com parâmetro *p*, resulta da soma de *n* variáveis discretas de Bernoulli. Como conseqüência do teorema do limite central, se *n* é suficientemente grande, é possível aproximar a distribuição binomial por uma distribuição Normal. Lembrando que a média e a variância da variável binomial *X* são, respectivamente, iguais a *np* e *np*(1-*p*), verifica-se que a variável definida por

$$Z = \frac{X - np}{\sqrt{np(1-p)}} \tag{5.22}$$

tende a ser distribuída conforme uma **N**(0,1), quando *n* tende para infinito. A convergência é mais rápida para valores de *p* em torno de 0,5; para valores de *p* próximos de 0 ou 1, maiores valores de *n* são necessários.

Analogamente, pode-se aproximar uma variável de Poisson *X*, de média e variância iguais a ν, pela variável Normal padrão

$$Z = \frac{X - \nu}{\sqrt{\nu}} \tag{5.23}$$

quando $\nu > 5$. Note, entretanto, que ao aproximar uma função massa de

probabilidade de uma variável discreta por uma função densidade de uma variável contínua, deve-se proceder à chamada *correção de continuidade*. De fato, no caso discreto, quando $X = x$, a FMP é uma linha ou um ponto; a linha ou a ordenada do ponto deve ser aproximada, no caso contínuo, pela área da FAP entre (x-0,5) e (x+0,5).

5.3 – Distribuição Log-Normal

Suponha que uma certa variável contínua X resulte da *ação multiplicativa* de um grande número de componentes aleatórios independentes X_i (i = 1,2,...,n), ou seja que $X = X_1.X_2 \ldots X_n$. Nesse caso, a variável $Y = \ln(X)$[1], tal que $Y = \ln(X_1) + \ln(X_2) + \ldots + \ln(X_n)$, em decorrência do teorema do limite central, irá tender a uma variável Normal, com parâmetros μ_Y e σ_Y, quando n for suficientemente grande para permitir a convergência. Sob tais condições, diz-se que a variável X segue uma distribuição Log-Normal, com parâmetros $\mu_{\ln(X)}$ e $\sigma_{\ln(X)}$, indicando-se sinteticamente que $X \sim \mathbf{LN}(\mu_{\ln(X)}, \sigma_{\ln(X)})$. É fácil verificar, por meio da aplicação da equação 3.61 a $f_Y(y)$, que a função densidade de uma variável log-normal X é dada por

$$f_X(x) = \frac{1}{x\,\sigma_{\ln(X)}\sqrt{2\pi}}\, exp\left\{-\frac{1}{2}\left[\frac{\ln(X) - \mu_{\ln(X)}}{\sigma_{\ln(X)}}\right]\right\} \text{ para } x > 0 \qquad (5.24)$$

O cálculo de probabilidades e de funções inversas pode ser efetuado tal como demonstrado para a FAP da distribuição Normal, tomando-se $Y = \ln(X)$ [1] como variável e, em seguida, $X = \exp(Y)$ [1] para os quantis correspondentes. A Figura 5.4 exemplifica a variação da forma da densidade Log-Normal para alguns valores específicos de $\mu_{\ln(X)}$ e $\sigma_{\ln(X)}$. O valor esperado e a variância de uma variável log-normal são, respectivamente,

$$E[X] = \mu_X = exp\left[\mu_{\ln(X)} + \frac{\sigma^2_{\ln(X)}}{2}\right] \qquad (5.25)$$

$$\mathrm{Var}[X] = \sigma^2_X = \mu^2_X\left[\exp\left(\sigma^2_{\ln(X)}\right) - 1\right] \qquad (5.26)$$

Dividindo a equação da variância por μ^2_X e, em seguida, extraindo a raiz quadrada, obtém-se a seguinte expressão para o *coeficiente de variação* de uma variável log-normal:

[1] A transformação logarítmica também pode ser feita na base 10; nesse caso, como $\log_{10}(X) = 0{,}4343 \ln(X)$, a equação 5.24 deve ser multiplicada por 0,4343. Os quantis serão $x = 10^y$, ao invés de $x = \exp(y)$.

$$CV_X = \sqrt{exp\left[\sigma^2_{\ln(X)}\right] - 1} \tag{5.27}$$

O coeficiente de assimetria da distribuição log-normal é dado por

$$\gamma = 3\,CV_X + \left(CV_X\right)^3 \tag{5.28}$$

Como $CV_X > 0$, resulta que a distribuição log-normal é sempre assimetricamente positiva, com coeficiente de assimetria proporcional ao coeficiente de variação.

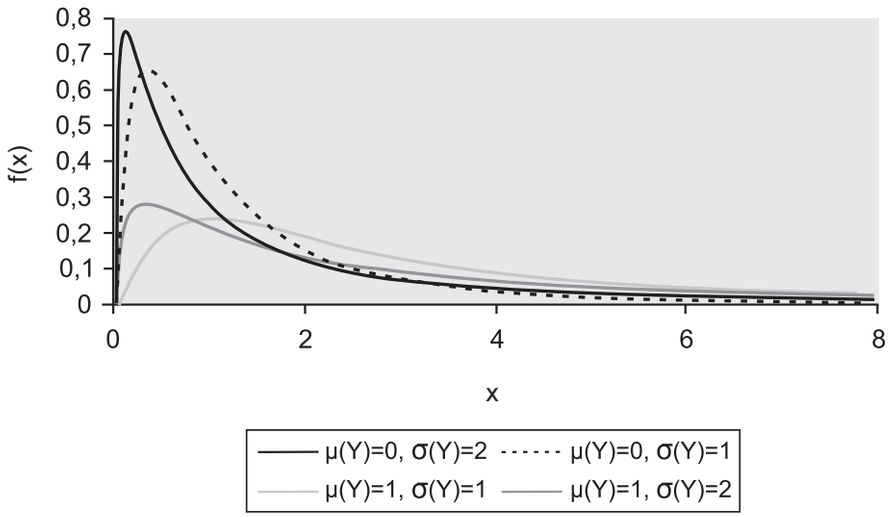


Figura 5.4 - Exemplos de Funções Densidades de Probabilidade Log-Normal

Exemplo 5.4 – Suponha que, a partir dos registros pluviométricos de uma certa localidade, é plausível a hipótese de que as alturas de precipitação do trimestre mais chuvoso são distribuídas segundo o modelo Log-Normal. A média e o desvio padrão das alturas pluviométricas trimestrais são respectivamente 600 e 150 mm. Calcule (a) a probabilidade da altura pluviométrica do trimestre mais chuvoso de um ano qualquer ficar compreendida entre 400 e 700 mm; (b) a probabilidade da altura pluviométrica do trimestre mais chuvoso de um ano qualquer ser pelo menos igual a 300 mm; e (c) a mediana das alturas pluviométricas.

Solução: (a) Denotemos a variável em questão por X. O coeficiente de variação de X é $CV = 150/600 = 0{,}25$. Com esse valor na equação 5.27, obtém-se $\sigma_{ln(X)} = 0{,}246221$. Com esse resultado e com $\mu_X = 600$ na equação 5.25, obtém-se $\mu_{\ln(X)} = 6{,}366617$. Portanto, $X \sim \mathbf{LN}\left(\mu_{\ln(X)} = 6{,}366617, \sigma_{\ln(X)} = 0{,}246221\right)$. A probabilidade pedida é

$$P(400 < X < 700) = \Phi\left(\frac{ln\,700 - 6{,}366617}{0{,}246221}\right) - \Phi\left(\frac{ln\,400 - 6{,}366617}{0{,}246221}\right) =$$

$$\Phi(0{,}7492) - \Phi(-1{,}5236) = 0{,}7093$$

Os valores de $\Phi(.)$ foram obtidos por interpolação linear entre os pontos da Tabela 5.1. (b) A probabilidade $P(X>30)=1-P(X<30)=$ $1 - \Phi\left(\frac{ln\,300 - 6{,}366617}{0{,}246221}\right) = 1 - \Phi(-2{,}69203) = 0{,}9965$

(c) Pelo fato da variável transformada $Y = \ln(X)$ ter como padrão de variação a distribuição Normal, ou seja, uma distribuição simétrica com a coincidência das medidas centrais em um único ponto, a mediana de Y é igual à média de Y, ou seja $y_{md} = 6{,}366617$. Há que se notar, entretanto, que, como a mediana de qualquer população (ou amostra) corresponde ao ponto intermediário que a divide em 50% de valores acima e abaixo, a transformação logarítmica não irá alterar a posição relativa (ou de classificação) da mediana. Daí decorre que a mediana de $\ln(X)$ é igual ao logaritmo neperiano da mediana de X, ou seja $y_{md} = \ln(x_{md})$ e, inversamente, $x_{md} = \exp(y_{md})$; observe que isso não é válido para a média ou para outras esperanças matemáticas. Portanto, a mediana das alturas pluviométricas trimestrais é $x_{md} = \exp(y_{md}) = \exp(6{,}366617) = 582{,}086$.

A distribuição Log-Normal de 3 parâmetros (LN3) é similar à distribuição já descrita, à exceção do fato de que da variável X deduz-se a quantidade a que representa um limite inferior. Nesse caso, a variável $Y = \ln(X-a)$ é distribuída de acordo com uma Normal com média μ_Y e desvio padrão σ_Y. A função densidade correspondente é

$$f_X(x) = \frac{1}{(x-a)\sigma_Y\sqrt{2\pi}} exp\left\{-\frac{1}{2}\left[\frac{ln(x-a) - \mu_Y}{\sigma_Y}\right]\right\} \tag{5.29}$$

A média e a variância da distribuição Log-Normal de 3 parâmetros são, respectivamente,

$$E[X] = a + exp\left(\mu_Y + \frac{\sigma_Y^2}{2}\right) \tag{5.30}$$

$$\mathrm{Var}[X] = \left[exp(\sigma_Y^2) - 1\right] exp\left(2\mu_Y + \sigma_Y^2\right) \tag{5.31}$$

O coeficiente de variação de uma variável LN3 é expresso por

$$CV_X = \frac{1 - \sqrt[3]{w^2}}{\sqrt[3]{w}} \tag{5.32}$$

onde w é definido pela seguinte função do coeficiente de assimetria da variável original X:

$$w = \frac{-\gamma + \sqrt{\gamma^2 + 4}}{2} \tag{5.33}$$

A proposição da distribuição log-normal justifica-se pela extensão dos princípios do teorema do limite central a uma variável que resulta da ação multiplicativa de componentes aleatórios independentes. Embora possam existir algumas evidências empíricas de que certos fenômenos hidrológicos, e suas variáveis, sejam resultantes da multiplicação de diversos fatores aleatórios [ver, por exemplo, Benjamin e Cornell (1970), Kottegoda e Rosso(1997) e Yevjevich (1972)], é controvertido preconizar o uso preferencial da distribuição log-normal, somente com base em tais argumentos. A controvérsia decorre da impossibilidade de enunciar tais fatores e compreender, com precisão, sua ação multiplicativa. Além disso, para justificar a aplicação preferencial da distribuição log-normal a variáveis hidrológicas, tais como vazões de cheia ou de estiagem, existe ainda a necessidade da verificação, quase sempre muito complexa, das condições de independência e de convergência, inerentes ao teorema do limite central. Por outro lado, o fato de que os argumentos para justificar o seu uso preferencial não são definitivos, não implica que a distribuição log-normal não seja uma forma paramétrica adequada à modelação de variáveis hidrológicas. Ao contrário, o fato da variável log-normal ser positiva, aliado à sua característica de ter como coeficiente de assimetria um valor não fixo e sempre maior do que zero, fazem da distribuição log-normal uma forma paramétrica que pode se adequar muito bem à modelação de vazões e alturas de chuva máximas (ou médias) mensais, trimestrais ou anuais.

5.4 – Distribuição Exponencial

O enunciado do exercício 8 do capítulo 4, mostra que o tempo contínuo entre duas ocorrências sucessivas de um processo de Poisson é modelado pela distribuição exponencial. Além desse fato matemático, a distribuição exponencial possui inúmeras outras aplicações em diversas áreas do conhecimento humano e, em particular, às variáveis hidrológicas. A função densidade da distribuição exponencial é expressa por

$$f_X(x) = \frac{1}{\theta} exp\left(-\frac{x}{\theta}\right) \text{ ou } f_X(x) = \lambda\, exp(-\lambda x), \text{ para } x \geq 0 \tag{5.34}$$

na qual, θ (ou $\lambda = 1/\theta$) denota o único parâmetro da distribuição. Se X~**E**(θ) ou X~**E**(λ), a função acumulada de probabilidades é dada por

$$F_X(x) = 1 - exp\left(-\frac{x}{\theta}\right) \text{ ou } F_X(x) = 1 - exp(-\lambda x) \tag{5.35}$$

O valor esperado, a variância e o coeficiente de assimetria (ver Exemplos 3.12 e 3.13 do capítulo 3) de uma variável exponencial são expressos, respectivamente, por

$$E[X] = \theta \text{ ou } E[X] = \frac{1}{\lambda} \tag{5.36}$$

$$\text{Var}[X] = \theta^2 \text{ ou Var}[X] = \frac{1}{\lambda^2} \tag{5.37}$$

$$\gamma = 2 \tag{5.38}$$

Observe que o coeficiente de assimetria da distribuição exponencial é fixo e positivo. A Figura 5.5 ilustra a FDP e a FAP dessa distribuição para $\theta = 2$ e $\theta = 4$.

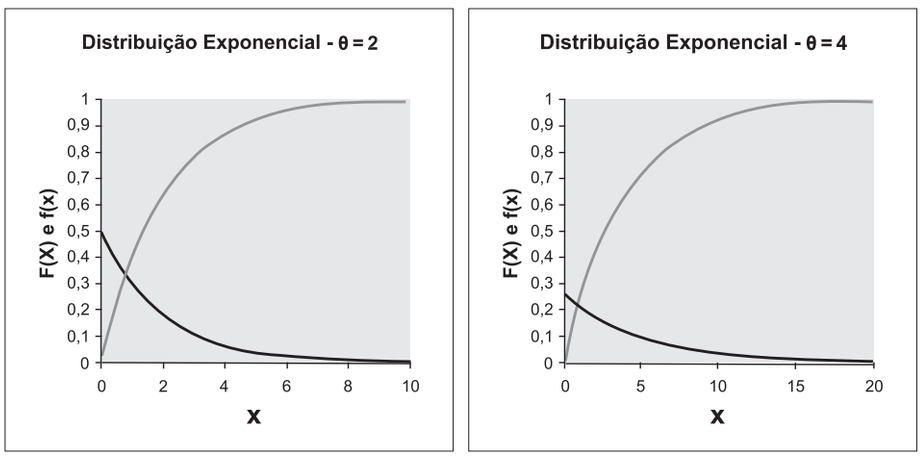


Figura 5.5 – FDP e FAP da Distribuição Exponencial para $\theta = 2$ e $\theta = 4$

Exemplo 5.5 – Com referência ao esquema de individualização de cheias, apresentado no enunciado do exercício 8 do capítulo 4, considere que, em média, ocorrem anualmente 2 cheias com vazões de pico superiores ao patamar Q_0= 60m^3/s. Considere que as 'excedências' (Q-Q_0) têm média igual a 50 m^3/s e que são exponencialmente distribuídas. Calcule a vazão de tempo de retorno T=100 anos.

Solução: Trata-se de um processo de Poisson com $\nu = \int_0^1 \lambda(t)dt = 2$,

onde os limites de integração 0 e 1 representam, respectivamente, o início e o fim do ano, λ a intensidade de Poisson e ν o número médio anual de ocorrências. Quando ocorrem, as excedências $X=(Q-Q_0)$ são distribuídas de acordo com a FAP exponencial, aqui representada por $G_X(x) = 1 - exp(-x/\theta)$, com $\theta = 50$ m³/s. Para calcular as vazões relacionadas a um certo tempo de retorno, é necessário, inicialmente, determinar a FAP das *excedências máximas anuais*, denotada por $F_{X\,max}(x)$, uma vez que $T = 1/(1 - F_{X\,max})$. Se o objetivo é calcular a probabilidade da excedência máxima anual x, é preciso raciocinar que cada uma das 1, 2, 3, ... ∞ excedências independentes, que podem ocorrer em um ano, devem ser menores ou iguais a x, uma vez que x representa o máximo anual. Logo, $F_{X\,max}(x)$ pode ser determinada, ponderando-se a probabilidade de ocorrência simultânea das n possíveis excedências independentes, ou seja $[G_X(x)]^n$, pela FMP do número anual de execedências n, o qual é distribuído segundo Poisson com parâmetro ν.

Portanto, $F_{X\,max}(x) = \sum_{n=0}^{\infty} [G_X(x)]^n \frac{\nu^n e^{-\nu}}{n!} = \sum_{n=0}^{\infty} [\nu G_X(x)]^n \frac{e^{-\nu}}{n!}$. Multiplicando e dividindo essa equação por $e^{-\nu G(x)}$, obtém-se

$$F_{X\,max}(x) = exp\{-\nu[1 - G_X(x)]\} \sum_{n=0}^{\infty} \frac{[\nu G_X(x)]^n \, exp[-\nu G_X(x)]}{n!} .$$

A somatória do segundo membro dessa equação é igual a 1 por tratar-se da soma total de uma FMP de Poisson com parâmetro $\nu G_X(x)$. Logo, chega-se a $F_{X\,max}(x) = exp\{-\nu[1 - G_X(x)]\}$, a qual é a *equação fundamental* para o cálculo de *probabilidades anuais* das séries de duração parcial com ocorrências de Poisson. No problema específico, a FAP das excedências é exponencial, ou seja, $G_X(x) = 1 - exp(-x/\theta)$, cuja substituição na equação acima resulta no chamado *modelo Poisson-Exponencial* para séries de durações parciais, ou seja, $F_{Q\,max}(q) = exp\left\{-\nu\, exp\left(-\frac{q - Q_0}{\theta}\right)\right\}$,

onde $Q_{max} = Q_0 + X$ representa a vazão máxima anual. Relembrando o fato matemático que se $a = b\,e^c \Leftrightarrow ln(a) = ln(b) + c \Leftrightarrow a = exp[ln(b) + c]$, resulta que $F_{Q\,max}(q) = exp\left\{-exp\left[-\frac{1}{\theta}(q - Q_0 - \theta\, ln\,\nu)\right]\right\}$, a qual representa a FAP da

importante *distribuição de Gumbel*, com parâmetro de escala θ e parâmetro de posição $[Q_0 + \theta\ln(\nu)]$, a ser detalhada no item 5.7 do presente capítulo. Portanto, a modelação de séries de duração parcial, com número de ocorrências distribuídas de acordo com a FMP de Poisson e excedências exponencialmente distribuídas, tem como distribuição de máximos anuais a distribuição de Gumbel. Para o problema em questão, $T = 100 \Rightarrow F_{Q\,max} = 1 - 1/100 = 0{,}99$; $\theta = 50$; $\nu = 2$ e $Q_0 = 60\text{m}^3/\text{s}$. Invertendo a FAP de Gumbel, obtém-se a função de quantis para essa distribuição, ou seja, $q(F) = Q_0 + \theta\, ln(\nu) - \theta\, ln[-ln(F)]$. Substituindo os valores, tem-se que $q(F = 0{,}98) = 289{,}8\ \text{m}^3/\text{s}$. Portanto, a vazão centenária para esse caso é 289,8 m^3/s.

5.5 – Distribuição Gama

A solução do exercício 9 do capítulo 4 mostra que a distribuição de probabilidades do tempo *t* para a *n*-ésima ocorrência de Poisson tem como função densidade $f_T(t) = \lambda^n t^{n-1} e^{-\lambda t} / (n-1)!$, a qual é denominada Gama para valores inteiros do parâmetro *n*. Nessas condições, a densidade Gama resulta da soma de *n* variáveis exponenciais independentes, cada qual com parâmetro λ ou, de modo equivalente, cada qual com parâmetro $\theta = 1/\lambda$. Em geral, o parâmetro *n* não necessita ser inteiro e, sem essa restrição, a função densidade da distribuição Gama passa a ter como expressão geral

$$f_X(x) = \frac{(x/\theta)^{\eta-1}\, exp(-x/\theta)}{\theta\,\Gamma(\eta)} \text{ para } x, \theta \text{ e } \eta > 0 \qquad (5.39)$$

na qual, θ e η representam, respectivamente, os parâmetros de escala e forma; sinteticamente, indica-se que $X \sim \mathbf{Ga}(\theta, \eta)$. Na equação 5.39, $\Gamma(\eta)$ denota o fator de normalização que obriga a área total da densidade ser igual a 1. Esse fator de normalização é expresso pela função *Gama completa* $\Gamma(.)$, do argumento η, a qual é dada por

$$\Gamma(\eta) = \int_0^\infty x^{\eta-1} e^{-x}\, dx \qquad (5.40)$$

Quando η é um número inteiro, a função Gama completa $\Gamma(\eta)$ é equivalente a $(\eta-1)!$. O leitor deve remeter-se ao Anexo 4 para uma breve revisão das propriedades matemáticas da função Gama e à referência Press et al. (1986), para a descrição de algoritmos para sua aproximação numérica. O Anexo 5 contém seus valores tabelados, para $1 \leq \eta \leq 2$; a propriedade matemática $\Gamma(\eta+1) = \eta\Gamma(\eta)$

permite a extensão dos valores tabelados para quaisquer outros valores de η. A função de probabilidades acumuladas da distribuição Gama é expressa por

$$F_X(x) = \int_0^x \frac{(x/\theta)^{\eta-1}\, exp(-x/\theta)}{\theta\Gamma(\eta)}\, dx \tag{5.41}$$

Assim como para a FAP da distribuição Normal, a integral dada pela equação 5.41 não pode ser obtida analiticamente. Portanto, o cálculo de probabilidades da distribuição Gama deve ser feito por aproximações numéricas, tais como as descritas por Press et al. (1986), ou por extensas tabelas encontradas em diversos livros-texto de estatística. Uma aproximação relativamente simples e que conduz a resultados satisfatórios, principalmente para valores elevados do parâmetro η, faz uso da variável Gama normalizada pelo parâmetro de escala; esse procedimento de aproximação da FAP da distribuição Gama encontra-se descrito a seguir. Com efeito, se X é uma variável Gama com parâmetro de escala arbitrário θ, a variável Gama padrão é dada por $\xi = x/\theta$; demonstra-se, nesse caso, que $\theta_\xi = 1$ e que o parâmetro de forma é o mesmo tanto para X, quanto para ξ. É fácil verificar que a função acumulada de probabilidade de X pode ser expressa pelo quociente

$$F_X(x) = \frac{\int_0^\xi \xi^{\eta-1} e^{-\xi} d\xi}{\int_0^\infty \xi^{\eta-1} e^{-\xi} d\xi} = \frac{\Gamma_i(\xi,\eta)}{\Gamma(\eta)} \tag{5.42}$$

entre a função Gama incompleta $\Gamma_i(\xi,\eta)$ e a função Gama completa $\Gamma(\eta)$. Maione e Moisello (2003) mostram que esse quociente pode ser *aproximado pela distribuição Normal padrão* $\Phi(u)$, calculada no ponto u, o qual é definido por

$$u = 3\sqrt{\eta}\left(\sqrt[3]{\frac{\xi}{\eta}} - 1 + \frac{1}{9\eta}\right) \tag{5.43}$$

O Exemplo 5.6 ilustra a aplicação desse procedimento para o cálculo de $F_X(x)$. O valor esperado, a variância e o coeficiente de assimetria da variável Gama são

$$E[X] = \eta\theta \tag{5.44}$$

$$\mathrm{Var}[X] = \eta\theta^2 \tag{5.45}$$

$$\gamma = \frac{2}{\sqrt{\eta}} \tag{5.46}$$

A Figura 5.6 apresenta os gráficos da função densidade Gama para alguns conjuntos de valores de θ e η. Nessa figura, observe que a função do parâmetro θ, cujas dimensões são as mesmas da variável aleatória, é a de comprimir ou estender a densidade para a esquerda ou para a direita, por meio do escalonamento dos valores de *X*. Por outro lado, a grande diversidade de formas da densidade Gama é garantida pela variação do parâmetro η. Como ilustrado na Figura 5.6, à medida que η decresce, a densidade Gama torna-se cada vez mais positivamente assimétrica. Para η =1, a densidade intercepta o eixo vertical no ponto 1/θ e configura o caso particular em que a distribuição Gama torna-se a distribuição exponencial, com parâmetro θ. Para valores crescentes do parâmetro de forma η, a função densidade Gama torna-se menos assimétrica, com o seu valor modal deslocando-se cada vez mais para a direita. Para valores muito elevados de η, a distribuição Gama aproxima-se da forma de uma distribuição Normal. Note que o parâmetro de forma η é um número adimensional.

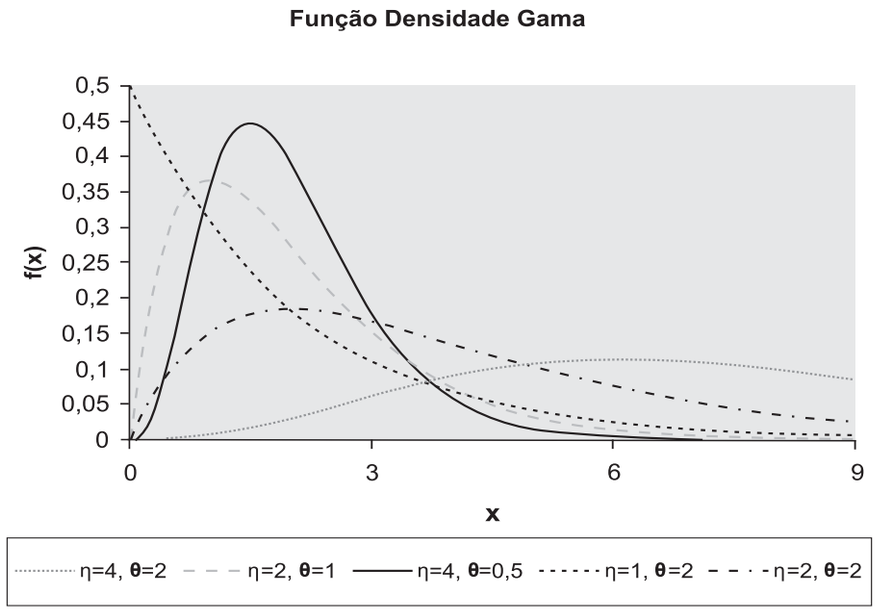


Figura 5.6 - Exemplos de Funções Densidades de Probabilidade da Distribuição Gama

A versatilidade de formas, o coeficiente de assimetria variável e positivo, aliados ao fato da variável aleatória não ser definida para valores negativos fazem da distribuição Gama um modelo probabilístico muito atraente para a representação de variáveis hidrológicas e hidrometeorológicas. Em particular, Haan (1977) destaca um grande número de aplicações bem sucedidas da distribuição Gama a alturas de precipitação de durações diárias, semanais, mensais e anuais; ressalta também uma modelação de vazões médias anuais com o emprego da distribuição Gama.

Exemplo 5.6 – Recalcule as probabilidades dos itens (a) e (b) do exemplo 5.4 para a distribuição Gama.

Solução: Inicialmente, devemos calcular os valores numéricos dos parâmetros η e θ. A combinação das equações 5.44 e 5.45 resulta em $\text{Var}[X] = E[X]\theta \Rightarrow \theta = \text{Var}[X]/E[X] = (150)^2/600 = 37{,}5$ mm. Substituindo esse valor em uma das duas equações, resulta que $\eta = 16$. (a) $P(400 < X < 700) = F_X(700) - F_X(400)$. Para calcular probabilidades da distribuição Gama, precisamos, de início, normalizar a variável dividindo o quantil pelo parâmetro de escala, ou seja, para $x = 700$, $\xi = x/\theta = 700/37{,}5 = 18{,}67$. Esse valor, levado na equação 5.43, com $\eta = 16$, resulta em $u = 0{,}7168$. A Tabela 5.1 fornece F(0,7168) = 0,7633 e, portanto, $P(X < 700) = 0{,}7633$. Procedendo do mesmo modo para $x = 400$, tem-se que $P(X < 400) = 0{,}0758$. Logo, $P(400 < X < 700) = 0{,}7633 - 0{,}0758 = 0{,}6875$. (b) A probabilidade $P(X > 30) = 1 - P(X < 30) = 1 - F_X(300)$. Para $x = 300$, $\xi = x/\theta = 300/37{,}5 = 8$. A equação 5.43, com $\eta = 16$, resulta em $u = -2{,}3926$ e, finalmente, $\Phi(-2{,}3926) = 0{,}008365$. Logo, $P(X > 30) = 1 - 0{,}00836\,5 = 0{,}9916$. Note que esses resultados não são muito diferentes daqueles obtidos no exemplo 5.4.

5.6 – Distribuição Beta

A distribuição Beta é um modelo probabilístico para uma variável aleatória contínua X, cujos valores possíveis são limitados superior e inferiormente. Na forma da distribuição Beta padronizada, a variável X é definida no intervalo [0,1]. Nesse caso, a função densidade Beta é expressa por

$$f_X(x) = \frac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1} \text{ para } 0 \le x \le 1, \alpha > 0, \beta > 0 \tag{5.47}$$

na qual, α e β são parâmetros e $B(\alpha,\beta)$ representa a função beta completa, dada por

$$B(\alpha,\beta) = \int_0^1 t^{\alpha-1}(1-t)^{\beta-1}\,dt = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} \tag{5.48}$$

De modo sintético, indica-se que $X \sim \mathbf{Be}(\alpha,\beta)$. A função de probabilidades acumuladas da distribuição Beta é

$$F_X(x) = \frac{1}{B(\alpha,\beta)} \int_0^x x^{\alpha-1}(1-x)^{\beta-1}\,dx = \frac{B_i(x,\alpha,\beta)}{B(\alpha,\beta)} \tag{5.49}$$

na qual, $B_i(x,\alpha,\beta)$ denota a função beta incompleta. Quando $\alpha=1$, a equação 5.49 pode ser resolvida analiticamente. Entretanto, para $\alpha \neq 1$, o cálculo de probabilidades da distribuição Beta exige aproximações numéricas da função $B_i(x,\alpha,\beta)$, tais como a apresentada por Press et al. (1986). A Figura 5.7 ilustra algumas formas possíveis para a função densidade Beta.

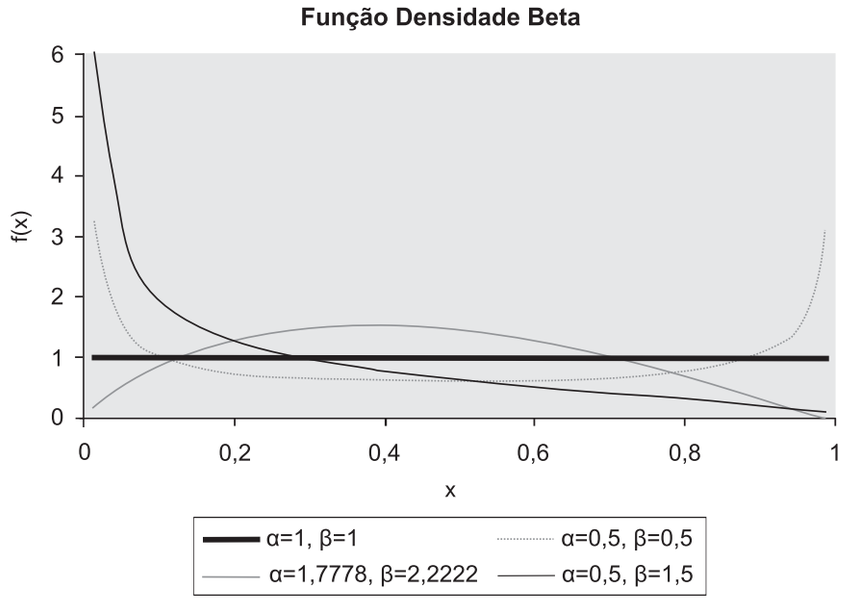


Figura 5.7 - Exemplos de Funções Densidades de Probabilidade da Distribuição Beta

A média e a variância de uma variável aleatória Beta são dadas, respectivamente, por

$$E[X] = \frac{\alpha}{\alpha+\beta} \tag{5.50}$$

$$\mathrm{Var}[X] = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} \tag{5.51}$$

Na Figura 5.7, note que a distribuição uniforme é um caso particular da distribuição Beta, para $\alpha = 1$ e $\beta = 1$. O parâmetro α controla os valores da densidade Beta em correspondência ao limite inferior da variável: se $\alpha < 1$, $f_X(x) \to \infty$, quando $x \to 0$; se $\alpha = 1$, $f_X(0) = 1/B(1,\beta)$; e, se $\alpha > 1$, $f_X(0) = 0$. Analogamente, o parâmetro β controla os valores da densidade Beta em correspondência ao limite superior. De modo geral, para valores iguais de ambos os parâmetros, a densidade Beta é simétrica; contrariamente, a distribuição Beta é assimétrica. Se ambos os

parâmetros são superiores a 1, a densidade Beta é unimodal. A variedade de formas dessa distribuição faz com que ela seja de muita utilidade para a modelação de variáveis com limites à direita e à esquerda.

Exemplo 5.6 – A concentração de oxigênio dissolvido, medida em intervalos semanais em uma seção fluvial, é uma variável X limitada à esquerda pelo valor 0 e, à direita, pela concentração de saturação, a qual depende da temperatura da água. Suponha que o limite superior seja 9 mg/l e que o valor esperado e a variância das concentrações de OD sejam, respectivamente, 4 mg/l e 4 $(mg/l)^2$. Se normalizarmos as concentrações de OD pelo limite superior, ou seja, se $Y=X/9$, é possível modelar tal variável pela distribuição Beta padronizada. Faça uso desse modelo para calcular a probabilidade da concentração de OD ser menor ou igual a 2 mg/l.
Solução: A variável transformada Y tem como média 4/9 e, como variância, 4/81. Resolvendo o sistema formado pelas equações 5.50 e 5.51, obtém-se os seguintes resultados $\alpha = 1{,}7778$ e $\beta = 2{,}2222$; observe que a densidade Beta, com esses valores numéricos dos parâmetros, encontra-se ilustrada na Figura 5.7. A probabilidade da concentração de OD ser menor ou igual a 2 mg/l é igual à probabilidade de Y ser igual ou inferior a 2/9. Para calcular $P[Y \leq (2/9)]$ por meio da equação 5.49, é necessário obter a aproximação numérica da função beta incompleta $B_i[(2/9), \alpha = 1{,}7778, \beta = 2{,}2222]$. Além do algoritmo proposto por Press et al. (1986), o programa Microsoft Excel incorpora a função estatística 'DISTBETA', a qual implementa o cálculo completo da equação 5.49. Usando essa função, $P[Y \leq (2/9)] = 0{,}1870$. Portanto, a probabilidade da concentração de OD ser menor ou igual a 2 mg/l é 0,1870.

5.7 – Distribuições de Valores Extremos

Uma categoria importante de distribuições de probabilidades provém da teoria clássica de valores extremos, cujo desenvolvimento iniciou-se com os trabalhos pioneiros do matemático Maurice Fréchet (1878-1973) e dos estatísticos Ronald Fisher (1912-1962) e Leonard Tippet (1902-), seguidos pelas contribuições devidas a Boris Gnedenko (1912-1995) e a consolidação teórica por parte de Emil Gumbel (1891-1966). Atualmente, a teoria de valores extremos é um ramo importante e ativo da estatística matemática, com desdobramentos práticos de grande relevância, principalmente, para as áreas de economia e engenharia. O

objetivo do presente item desse capítulo é o de sintetizar os fundamentos da teoria de valores extremos e suas principais aplicações em hidrologia; para o leitor interessado em aprofundar conhecimentos nesse ramo da estatística matemática, sugere-se o excelente livro escrito por Coles (2001).

5.7.1 – Distribuições Exatas de Valores Extremos

Os valores máximo e mínimo de uma amostra de tamanho N de uma variável aleatória X, cuja FAP é conhecida e dada por $F_X(x)$, também são variáveis aleatórias e possuem distribuições de probabilidades próprias, as quais estão relacionadas à distribuição da *variável original*. Na amostra aleatória simples $\{x_1, x_2, \ldots, x_N\}$, x_i denota a i-ésima das N observações da variável X. Como não é possível prever o valor de x_i antes de sua ocorrência, pode-se presumir que x_i representa o valor da variável aleatória X_i, ou, em outras palavras, que a amostra $\{x_1, x_2, \ldots, x_N\}$ é uma *realização* das N variáveis aleatórias independentes e igualmente distribuídas $\{X_1, X_2, \ldots, X_N\}$. A partir dessas considerações, a teoria de valores extremos visa determinar as distribuições de probabilidades do máximo $Y = \max\{X_1, X_2, \ldots, X_N\}$ e do mínimo $Z = \min\{X_1, X_2, \ldots, X_N\}$ de X.

A distribuição de Y pode ser deduzida do fato que, se $Y = \max\{X_1, X_2, \ldots, X_N\}$ é menor ou igual a y, então todas as variáveis aleatórias X_i também devem ser menores ou iguais a y. Como todas as variáveis X_i são independentes entre si e distribuídas conforme a função $F_X(x)$ da variável original X, a distribuição de probabilidades acumuladas de Y pode ser deduzida do seguinte modo:

$$F_Y(y) = \mathrm{P}(Y \le y) = \mathrm{P}\left[(X_1 \le y) \cap (X_2 \le y) \cap \ldots \cap (X_N \le y)\right] = \left[F_X(y)\right]^N \quad (5.52)$$

A função densidade de probabilidades de Y é, portanto,

$$f_Y(y) = \frac{d\,F_Y(y)}{dy} = N\left[F_X(y)\right]^{N-1} f_X(y) \quad (5.53)$$

A equação 5.52 indica que, para um dado y, $F_Y(y)$ decresce com N e que, portanto, ambas as funções densidade e acumulada de Y irão deslocar-se para a direita, para valores crescentes de N; tal fato é ilustrado na Figura 5.8, para o caso em que $f_X(x) = 0{,}25\,exp(-0{,}25x)$. Nessa figura, observe também que a moda, ou seja, o valor mais freqüente de Y desloca-se para a direita para N crescente e que, mesmo para valores moderados de N, tal valor já coincide com aqueles extraídos da cauda superior da densidade da variável original.

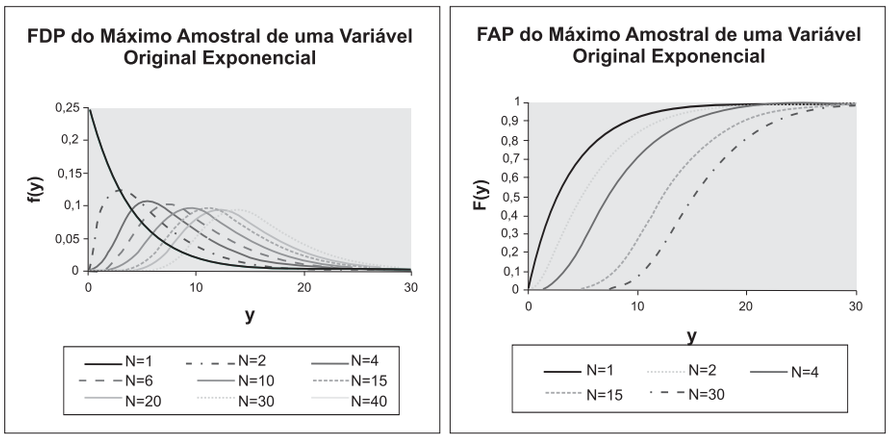


Figura 5.8 – FDP e FAP do máximo amostral de uma variável original exponencial

Empregando raciocínio idêntico, é possível deduzir as funções densidade e de probabilidades acumuladas do mínimo amostral $Z = \min\{X_1, X_2, \ldots, X_N\}$. Com efeito, a FAP de Z é dada por

$$F_Z(z) = 1 - [1 - F_X(z)]^N \tag{5.54}$$

e a função densidade por

$$f_Z(z) = N[1 - F_X(z)]^{N-1} f_X(z) \tag{5.55}$$

Contrariamente às distribuições do máximo amostral, as funções $F_Z(z)$ e $f_Z(z)$ deslocar-se-ão para a esquerda para valores crescentes de N.

As equações 5.52 a 5.55 representam as *distribuições exatas de valores extremos* de uma amostra de tamanho N, extraída da população da variável original X, da qual se conhecem integralmente as funções densidade e acumulada. Essas equações revelam que as distribuições exatas de valores extremos *dependem da distribuição* $F_X(x)$ da variável original X e também *do tamanho da amostra* N. Em geral, exceção feita a algumas distribuições simples da variável original, tais como a distribuição exponencial, as expressões analíticas de $F_Y(y)$ e $F_Z(z)$ não são de fácil obtenção ou dedução.

Exemplo 5.7 – Suponha que, em uma dada região, o tempo entre episódios de chuva seja uma variável exponencialmente distribuída, com média de 4 dias, e que seja válida a hipótese de independência entre os tempos

consecutivos que separam tais episódios. Com o fim de planejar os turnos de rega entre os meses de Abril e Junho, sob condições críticas, os irrigantes da região necessitam conhecer o máximo tempo entre episódios de chuva. Se, nesses meses, espera-se ter 16 episódios de chuva, calcule a probabilidade de que o tempo máximo entre eles seja maior do que 10 dias. (adap. de Haan, 1977)

Solução: A ocorrência de 16 episódios de chuva implica em 15 tempos separando tais eventos; para efeito da aplicação da equação 5.52, isso implica em N=15. Denotando por T_{max} a variável aleatória 'tempo máximo entre chuvas', P(T_{max}>10) = 1-P(T_{max}<10) = 1-F_{Tmax}(10). A FAP de T_{max} é

$F_{T\,max}(10) = [F_T(10)]^{15} = [1 - exp(-10/4)]^{15}$, ou seja,

$F_{T\,max}(10) = [1 - exp(-2{,}5)]^{15} = 0{,}277$. Portanto, P($T_{max}$>10) = 1-0,277 = 0,723.

Obtém-se a densidade de T_{max} pela aplicação direta da equação 5.53, ou seja, $f_{T\,max}(t_{max}) = N\left[1 - exp\left(-\frac{t_{max}}{4}\right)\right]^{N-1}\left[\frac{1}{4}exp\left(-\frac{t_{max}}{4}\right)\right]$; essa função densidade está ilustrada na Figura 5.8, para diversos valores de N.

5.7.2 – Distribuições Assintóticas de Valores Extremos

A utilidade prática do estudo estatístico de extremos é grandemente aumentada pela *teoria assintótica de valores extremais*, cujo foco principal é a determinação das formas limites de $F_Y(y)$ e $F_Z(z)$, ou de suas respectivas densidades, quando N tende ao infinito, sem o completo conhecimento da forma exata da distribuição $F_X(x)$, da variável original. De fato, freqüentemente, $F_X(x)$ não é completamente conhecida ou não pode ser analiticamente determinada, o que impede a aplicação das equações 5.52 a 5.55 e, portanto, a explicitação das distribuições exatas do máximo e do mínimo. A contribuição principal da teoria assintótica de valores extremos é demonstrar que os limites $\lim_{N\to\infty} F_Y(y)$ e $\lim_{N\to\infty} F_Z(z)$ convergem para certas formas funcionais, independentemente do conhecimento exato da distribuição $F_X(x)$ da variável original. De fato, a convergência desses limites depende fundamentalmente do *comportamento da cauda de $F_X(x)$ na direção do extremo*, ou seja, da cauda superior de $F_X(x)$, se o interesse for o máximo Y, ou da cauda inferior de $F_X(x)$, se o interesse volta-se para o mínimo Z; a parte central de $F_X(x)$ tem pouca influência sobre a convergência de $\lim_{N\to\infty} F_Y(y)$ e $\lim_{N\to\infty} F_Z(z)$.

Suponha que $\{X_1, X_2, \ldots, X_N\}$ represente um conjunto de N variáveis aleatórias independentes, com distribuição comum $F_X(x)$. Particularizando para o máximo ou mínimo anual, N pode ser interpretado como o número de observações de X, em instantes de tempo eqüidistantes entre si, ao longo de um período fixo de 1 ano. Se $Y = \max\{X_1, X_2, \ldots, X_N\}$ e $Z = \min\{X_1, X_2, \ldots, X_N\}$, tomemos as transformações lineares $Y_N = (Y - b_N)/a_N$ e $Z_N = (Z - b_N)/a_N$, onde a_N e b_N são constantes de escala e posição, respectivamente. A teoria assintótica de valores extremos demonstra que os limites $\lim_{N\to\infty} F_{Y_N}(y)$ e $\lim_{N\to\infty} F_{Z_N}(z)$ convergem, embora de modo não exaustivo, para três formas funcionais, a depender do comportamento da cauda da distribuição da variável original, na direção do extremo em questão. Gumbel (1958) classificou essas três formas assintóticas em

• Tipo I: *a forma dupla exponencial*:
(a) para máximos, $exp[-e^{-y}]$, com $-\infty < y < \infty$, ou
(b) para mínimos, $1 - exp(-e^{z})$, com $-\infty < z < \infty$, quando X é ilimitado e sua densidade decai de modo *exponencial* na direção do extremo;

• Tipo II: *a forma exponencial simples*:
(a) para máximos, $exp(-y^{-\gamma})$, se $y > 0$, e 0, se $y \leq 0$, ou
(b) para mínimos $1 - exp[-(-z)^{-\gamma}]$, se $z<0$, e 1, se $z \geq 0$, quando X é ilimitado e sua densidade decai de modo *polinomial* na direção do extremo; e

• Tipo III: *a forma exponencial com limite* superior para máximos ou inferior para mínimos:
(a) para máximos, $exp[-(-y)^{\gamma}]$, se $y < 0$, e 1, se $y \geq 0$, ou
(b) para mínimos, $1 - exp(-z^{\gamma})$, se $z > 0$, e 0, se $z \leq 0$, quando X é *limitado* na direção do extremo.

Na caracterização das formas assintóticas acima, γ denota uma constante positiva.

Tomando-se o caso de máximos apenas, a distribuição da variável original X possui uma cauda superior exponencial se ela não possui limite superior e se, para valores positivos elevados de x, as ordenadas de $f_X(x)$ e de $1\text{-}F_X(x)$ são pequenas, enquanto $f'_X(x)<0$, sendo válida a seguinte relação $\frac{f_X(x)}{1-F_X(x)} = -\frac{f'_X(x)}{f_X(x)}$. Em palavras, a distribuição da variável original tem cauda superior exponencial se $F_X(x)$, além de ilimitada superiormente, aproxima-se de 1 pelo menos tão rapidamente quanto a distribuição exponencial o faz, quando $x \to \infty$. Por outro lado, $F_X(x)$ possui uma cauda superior polinomial, também

denominada de Cauchy-Pareto, se ela não for limitada à direita e se $\lim_{x\to\infty} x^k \left[1-F_X(x)\right] = a$, onde a e k são números positivos. Em palavras, a distribuição da variável original tem cauda superior de Cauchy-Pareto se $F_X(x)$, além de ilimitada superiormente, aproxima-se de 1 menos rapidamente que a distribuição exponencial o faz, quando $x \to \infty$. Finalmente, se X é limitada superiormente pelo valor w, ou seja, se $F_X(w) = 1$, a distribuição assintótica de seu valor máximo será do tipo III. A Figura 5.9 exemplifica os 3 tipos de cauda superior de funções densidade da variável original X.

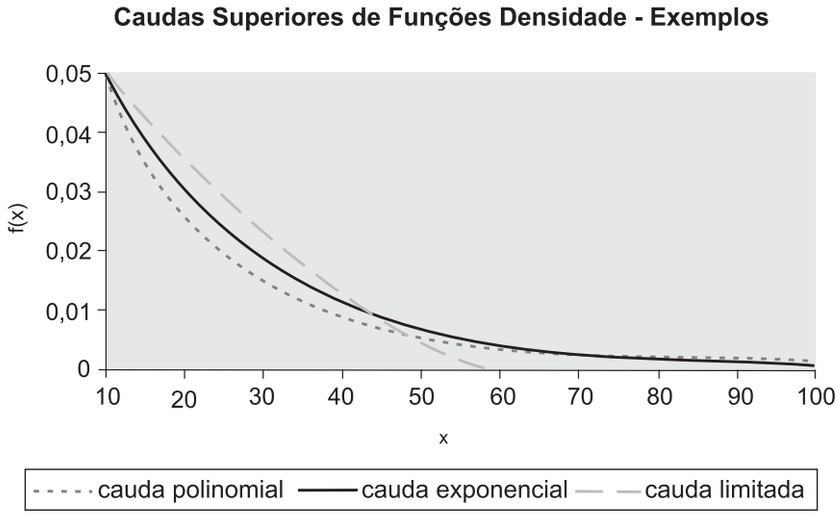


Figura 5.9 – Exemplos de caudas superiores de funções densidades de probabilidades

O comportamento da cauda da distribuição da variável original, na direção do extremo em foco, determina, portanto, para qual das três formas assintóticas a distribuição dos máximos ou dos mínimos irá convergir. No caso de máximos, a convergência será para a distribuição (a) do Tipo I, se $F_X(x)$ for, por exemplo, exponencial, ou Gama, ou Normal, ou Log-Normal, ou a própria distribuição de máximos do Tipo I; (b) do Tipo II, se $F_X(x)$ for, por exemplo, a distribuição Gama dos logaritmos da variável (Log-Gama), ou a distribuição t de Student, a ser descrita no item 5.9.2 desse capítulo, ou a própria distribuição de máximos do Tipo II; e (c) do Tipo III, se $F_X(x)$ for, por exemplo, uniforme, ou Beta, ou a própria distribuição de máximos do Tipo III. No caso de mínimos, a convergência será para a distribuição do (a) Tipo I, se $F_X(x)$ for, por exemplo, Normal, ou a própria distribuição de mínimos do Tipo I; (b) Tipo II, se $F_X(x)$ for, por exemplo, a distribuição t de Student ou a própria distribuição de mínimos do Tipo II; e (c) Tipo III, se $F_X(x)$ for, por exemplo, uniforme, ou exponencial, ou Beta, ou Log-Normal, ou Gama, ou a própria distribuição de mínimos do Tipo III.

As distribuições oriundas da teoria assintótica de valores extremos encontram numerosas aplicações às variáveis hidrológicas, embora as premissas, sobre as quais se baseiam, não se verifiquem completamente na realidade dos fenômenos do ciclo da água. De fato, as premissas fundamentais da teoria clássica de valores extremos são que as variáveis originais são independentes e igualmente distribuídas. Se contextualizarmos, por exemplo, que Y e Z referem-se, respectivamente, ao máximo e ao mínimo anual das vazões médias diárias $\{X_1, X_2, \ldots, X_{365}\}$, essas devem ser independentes entre si e devem ter uma única e idêntica distribuição de probabilidades. Se por um lado, a independência entre vazões médias diárias consecutivas é uma hipótese pouco plausível, por outro, admitir, por exemplo, que a vazão média do dia 16 de Janeiro tem a mesma distribuição, mesma média e mesma variância, da vazão do dia 19 de Agosto, é de aceitação muito difícil.

Essas contradições estão entre as diversas que, de fato, impedem a aplicação de leis dedutivas para a seleção de modelos probabilísticos de máximos e mínimos hidrológicos. Entretanto, de modo análogo à lógica de utilização de outras distribuições, o fato que suas premissas de base não encontram respaldo completo na realidade física, não implica que as distribuições de valores extremos não sejam formas paramétricas adequadas à modelação de variáveis hidrológicas. Ao contrário, as distribuições de valores extremos, ou distribuições extremais, são modelos válidos e muito empregados na prática hidrológica.

Em particular, a forma assintótica de máximos do Tipo I, também conhecida por distribuição de Gumbel de máximos, é muito utilizada na análise de freqüência de eventos hidrológicos. Em menor grau, também o é a forma assintótica de máximos do Tipo II, ou distribuição de Fréchet de máximos. A forma assintótica de máximos do tipo III, ou distribuição de Weibull de máximos, não é muito utilizada em hidrologia, principalmente, porque possui um limite à direita. Por essas razões, destacaremos aqui a descrição das distribuições Gumbel, Fréchet e do modelo geral que reúne as três formas assintóticas de máximos, a saber, a distribuição Generalizada de Valores Extremos. No que se refere aos extremos mínimos, o destaque será dado à descrição dos modelos extremais mais usados, a saber, o do Tipo I, ou distribuição de Gumbel de mínimos, e o do Tipo III, ou distribuição de Weibull de mínimos.

5.7.2.1 – Distribuição de Gumbel (Máximos)

A distribuição de valores extremos do Tipo I recebeu as seguintes outras denominações: distribuição de Gumbel, Fisher-Tippet tipo I e dupla exponencial. No caso de valores máximos, a distribuição de Gumbel refere-se à forma

assintótica limite para um conjunto de N variáveis aleatórias originais $\{X_1, X_2, \ldots, X_N\}$, independentes e igualmente distribuídas conforme um modelo $F_X(x)$, de cauda superior exponencial. A distribuição de Gumbel (máximos) é a distribuição extremal mais usada na análise de freqüência de variáveis hidrológicas, com inúmeras aplicações na determinação de relações intensidade-duração-freqüência de precipitações intensas e estudos de vazões de enchentes. A função de probabilidades acumuladas da distribuição de Gumbel é dada por

$$F_Y(y) = exp\left[-exp\left(-\frac{y-\beta}{\alpha}\right)\right] \text{ para } -\infty < y < \infty, -\infty < \beta < \infty, \alpha > 0 \tag{5.56}$$

na qual, α representa o parâmetro de escala e β o parâmetro de posição; de fato, β também é a moda de Y. A função densidade da distribuição de Gumbel é

$$f_Y(y) = \frac{1}{\alpha} exp\left[-\frac{y-\beta}{\alpha} - exp\left(-\frac{y-\beta}{\alpha}\right)\right] \tag{5.57}$$

O valor esperado, a variância e o coeficiente de assimetria de Y são, respectivamente,

$$E[Y] = \beta + 0{,}5772\alpha \tag{5.58}$$

$$\text{Var}[Y] = \sigma_Y^2 = \frac{\pi^2 \alpha^2}{6} \tag{5.59}$$

$$\gamma = 1{,}1396 \tag{5.60}$$

Observe, portanto, que a distribuição Gumbel (máximos) possui um coeficiente de assimetria positivo e constante. A Figura 5.10 ilustra a função densidade Gumbel, para alguns valores específicos dos parâmetros α e β.

A função inversa da FAP de Gumbel, ou função de quantis, é expressa por

$$y(F) = \beta - \alpha\, ln[-ln(F)] \text{ ou } y(T) = \beta - \alpha\, ln\left[-ln\left(1-\frac{1}{T}\right)\right] \tag{5.61}$$

na qual, T denota o período de retorno em anos e F representa a probabilidade anual de não superação. Na equação 5.61, substituindo-se y pelo valor esperado $E[Y]$, resulta que a média de uma variável de Gumbel corresponde ao período de retorno $T = 2{,}33$ anos. Em alguns estudos de regionalização de vazões de cheias, esse quantil, ou seja, $y\,(T = 2{,}33)$, recebe a denominação de 'cheia média anual'.

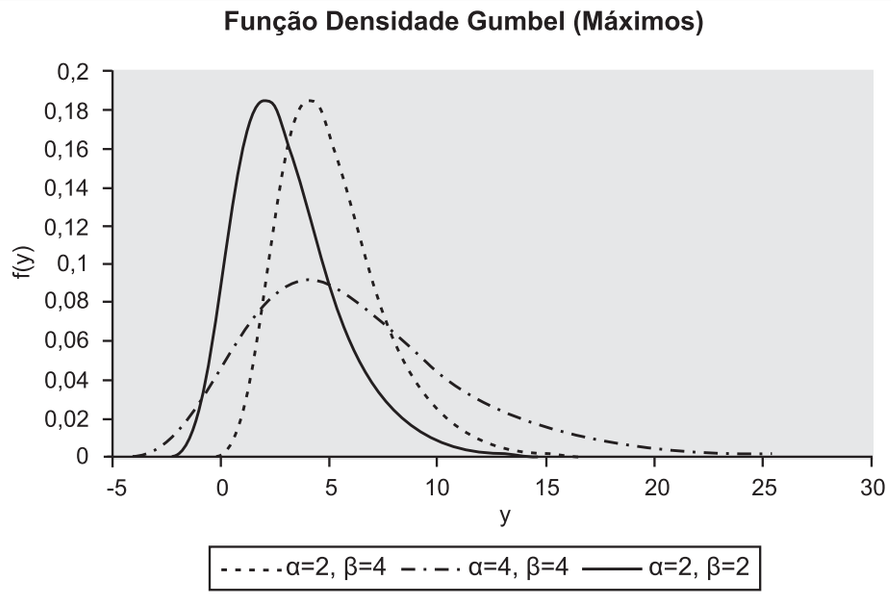


Figura 5.10 – Exemplos de funções densidades da distribuição de Gumbel (máximos)

Exemplo 5.8 – Denote por X a variável aleatória 'vazões médias diárias máximas anuais'; suponha que, em um certo local, $E[X] = 500$ m³/s e $E[X^2] = 297025$ (m³/s)². Utilize o modelo de Gumbel para calcular (a) a vazão média diária máxima anual de tempo de retorno 100 anos e (b) dado que a vazão média diária máxima anual é maior do que 600 m³/s, a probabilidade de X superar 800 m³/s.

Solução: (a) Lembrando que $\text{Var}[X]=E[X^2]-(E[X])^2$, resulta que $\text{Var}[X] = 47025$ (m³/s)². Resolvendo o sistema formado pelas equações 5.58 e 5.59, com os valores de Var[X] e $E[X]$, obtém-se $\alpha = 169{,}08$ m³/s e $\beta = 402{,}41$ m³/s. Com esses valores numéricos dos parâmetros na equação 5.61, conclui-se que vazão média diária máxima anual de tempo de retorno 100 anos é $x(100) = 1180$ m³/s. (b) Representemos o fato de que as vazões superaram 600 m³/s pelo evento A e que o evento B denote que as vazões superaram 800 m³/s. Portanto, desejamos calcular a probabilidade condicional P(B|A), a qual pode ser posta sob a forma $\text{P}(B|A) = \text{P}(B \cap A)/\text{P}(A)$. O numerador dessa última equação é equivalente a P(B), ou seja, $\text{P(B)} = 1\text{-}F_X(800) = 0{,}091$. O denominador é $\text{P(A)} = 1\text{-}F_X(600) = 0{,}267$. Logo, P(B|A) = 0,34.

5.7.2.2 – Distribuição de Fréchet (Máximos)

A distribuição de Fréchet é uma forma particular da distribuição de valores extremos do Tipo II. A distribuição de Fréchet é conhecida também pela denominação Log-Gumbel, a qual justifica-se pelo fato que, se Z~Gumbel(α, β), então $Y = \ln(Z)$~Fréchet[$1/\alpha$, exp(β)]. No caso de valores máximos, a distribuição de Fréchet refere-se à forma assintótica limite para um conjunto de N variáveis aleatórias originais $\{X_1, X_2, \ldots, X_N\}$, independentes e igualmente distribuídas conforme um modelo $F_X(x)$, de cauda superior polinomial. A distribuição foi usada pela primeira vez na análise de freqüência de vazões de enchentes por Fréchet (1927), tendo, desde então, encontrado aplicações, como distribuição extremal de eventos hidrológicos máximos.

A função de probabilidades acumuladas da distribuição de Fréchet é dada por

$$F_Y(y) = \exp\left[-\left(\frac{y_0}{y}\right)^{\theta}\right] \text{ para } y > 0,\ y_0, \theta > 0 \tag{5.62}$$

na qual, y_0 representa o parâmetro de escala e θ o parâmetro de forma. A função densidade da distribuição de Fréchet é

$$f_Y(y) = \frac{\theta}{y_0}\left(\frac{y_0}{y}\right)^{\theta+1} \exp\left[-\left(\frac{y_0}{y}\right)^{\theta}\right] \tag{5.63}$$

O valor esperado, a variância e o coeficiente de variação de Y são, respectivamente,

$$E[Y] = y_0\,\Gamma\left(1-\frac{1}{\theta}\right) \text{ para } \theta > 1 \tag{5.64}$$

$$\text{Var}[Y] = \sigma_Y^2 = y_0^2\left[\Gamma\left(1-\frac{2}{\theta}\right) - \Gamma^2\left(1-\frac{1}{\theta}\right)\right] \text{para } \theta > 2 \tag{5.65}$$

$$CV_Y = \sqrt{\frac{\Gamma(1-2/\theta)}{\Gamma^2(1-1/\theta)} - 1} \text{ para } \theta > 2 \tag{5.66}$$

Observe, portanto, que o parâmetro de forma da distribuição de Fréchet (máximos) é função unicamente do coeficiente de variação; tal fato simplifica o cálculo dos parâmetros da distribuição de Fréchet. Com efeito, se CV_Y é conhecido, a equação 5.66 pode ser resolvida para θ, por meio de iterações numéricas; em seguida, resolve-se a equação 5.64 para y_0. A Figura 5.11 ilustra a função densidade de Fréchet, para alguns valores específicos dos parâmetros y_0 e θ. A equação de quantis da distribuição de Fréchet é dada por

$$y(F) = y_0[-ln(F)]^{1/\theta} \tag{5.67}$$

ou, em termos do período de retorno T,

$$y(F) = y_0\left[ln\left(\frac{T}{T-1}\right)\right]^{-1/\theta} \tag{5.68}$$

Como mencionado anteriormente, as distribuições de Gumbel e de Fréchet são relacionadas entre si por meio da transformação logarítmica das variáveis, ou seja, se Y é uma variável de Fréchet, com parâmetros y_0 e θ, a variável ln(Y) é uma variável de Gumbel, com parâmetros $\alpha = 1/\theta$ e $\beta = \ln(y_0)$. Esse fato matemático faz com que, para um mesmo período de retorno, os quantis calculados pela distribuição de Fréchet sejam muito superiores àqueles calculados pela distribuição de Gumbel.

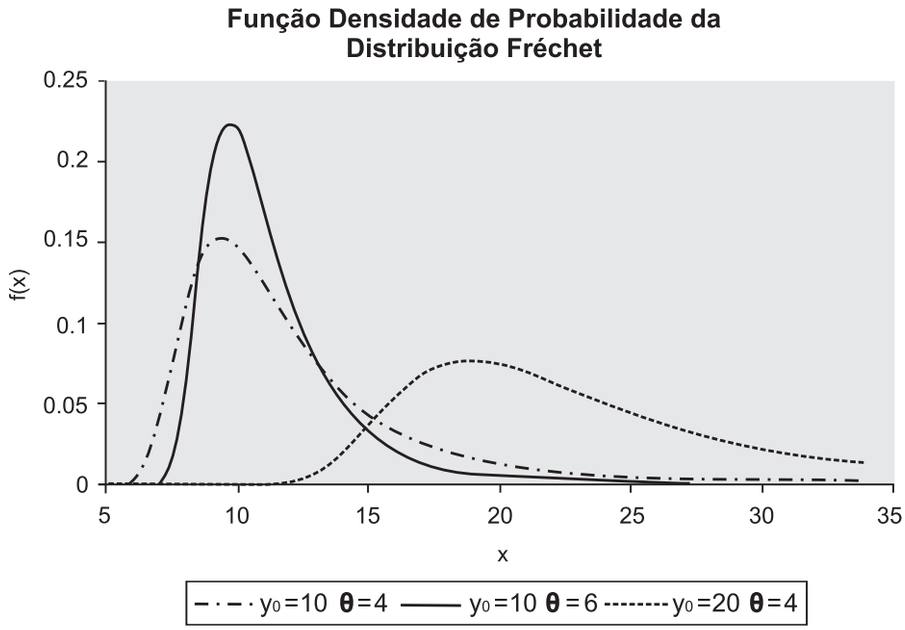


Figura 5.11 – Exemplos de funções densidades da distribuição de Fréchet (máximos)

5.7.2.3 – Distribuição Generalizada de Valores Extremos (Máximos)

A distribuição Generalizada de Valores Extremos, ou distribuição GEV da terminologia inglesa *Generalized Extreme Value*, foi introduzida por Jenkinson (1955) e incorpora as três formas assintóticas de valores extremos máximos em uma única expressão. A função de probabilidades acumuladas da distribuição GEV é dada por

$$F_Y(y) = exp\left\{-\left[1-\kappa\left(\frac{y-\beta}{\alpha}\right)\right]^{1/\kappa}\right\} \tag{5.69}$$

na qual, κ, α e β denotam, respectivamente, os parâmetros de forma, escala e posição. O valor e o sinal de κ determinam a forma assintótica de valores extremos máximos, ou seja, se $\kappa < 0$, a GEV representa a distribuição do Tipo II, definida apenas para $y > (\beta+\alpha)/\kappa$, enquanto que, se $\kappa > 0$, a GEV representa a distribuição do Tipo III, definida apenas para $y < (\beta+\alpha)/\kappa$. Se $\kappa = 0$, a GEV corresponde à distribuição de Gumbel com parâmetro de escala α e parâmetro de posição β. A função densidade da distribuição GEV é expressa por

$$f_Y(y) = \frac{1}{\alpha}\left[1-\kappa\left(\frac{y-\beta}{\alpha}\right)\right]^{1/\kappa-1} exp\left\{-\left[1-\kappa\left(\frac{y-\beta}{\alpha}\right)\right]^{1/\kappa}\right\} \tag{5.70}$$

A Figura 5.12 ilustra as três formas possíveis da distribuição GEV.

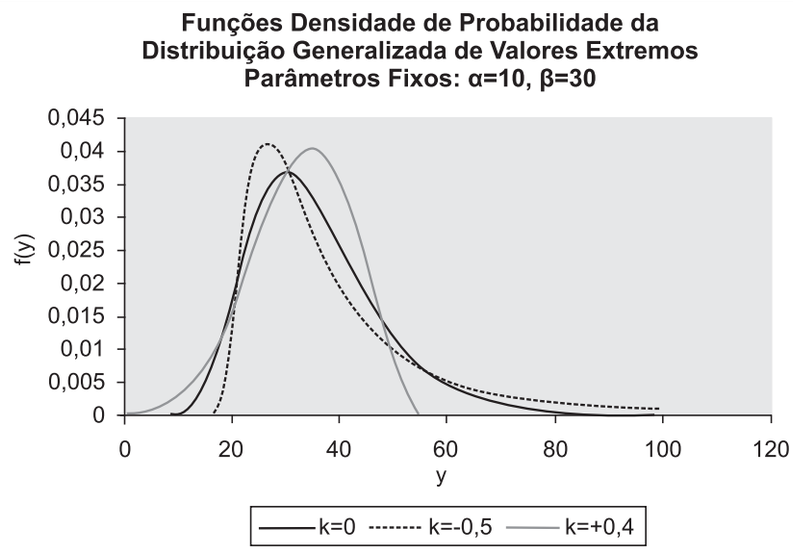


Figura 5.12 – Exemplos de funções densidades da distribuição GEV

Os momentos de ordem *r* da distribuição GEV existem apenas se $\kappa > -1/r$. Por conseguinte, a média de uma variável GEV não é definida para $\kappa < -1$, a variância não o é para $\kappa < -1/2$, enquanto o coeficiente de assimetria existe somente para $\kappa < -1/3$. Sob essas condições, a média, a variância e o coeficiente de assimetria de uma variável GEV são dados, respectivamente, por

$$E[Y] = \beta + \frac{\alpha}{\kappa}[1-\Gamma(1+\kappa)] \tag{5.71}$$

$$\mathrm{Var}[Y] = \left(\frac{\alpha}{\kappa}\right)^2 \left[\Gamma(1+2\kappa) - \Gamma^2(1+\kappa)\right] \tag{5.72}$$

$$\gamma = \langle \text{sinal de } \kappa \rangle \frac{-\Gamma(1+3\kappa) + 3\Gamma(1+\kappa)\Gamma(1+2\kappa) - 2\Gamma^3(1+\kappa)}{[\Gamma(1+2\kappa) - \Gamma^2(1+\kappa)]^{3/2}} \tag{5.73}$$

Observa-se, portanto, que o parâmetro de forma κ depende unicamente do coeficiente de assimetria γ; essa dependência unívoca é ilustrada na Figura 5.13, para $\kappa > -1/3$. Nessa figura, note que o ponto assinalado pelo símbolo +, corresponde à distribuição de Gumbel, com $\kappa = 0$ e γ=1,1396.

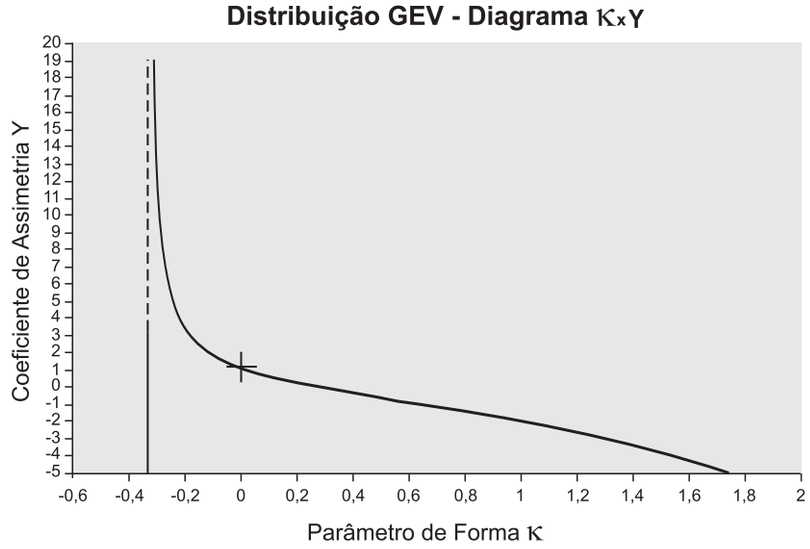


Figura 5.13 – Relação entre o parâmetro de forma e o coeficiente de assimetria de uma variável GEV, para $\kappa > -1/3$

O cálculo dos parâmetros da distribuição GEV deve começar pela equação 5.73, a qual deve ser resolvida para κ, por meio de iteração numérica ou com o auxílio do gráfico da Figura 5.13, a partir do valor do coeficiente de assimetria. Em seguida, calcula-se o valor de α, isolando-o na equação 5.72, ou seja,

$$\alpha = \sqrt{\frac{\kappa^2 \text{Var}[Y]}{\Gamma(1+2\kappa) - \Gamma^2(1+\kappa)}} \tag{5.74}$$

Finalmente, a manipulação da equação 5.71 permite o cálculo de β, ou seja,

$$\beta = E[Y] - \frac{\alpha}{\kappa}[1 - \Gamma(1+\kappa)] \tag{5.75}$$

De posse dos valores numéricos dos parâmetros, os quantis da distribuição GEV são dados por

$$x(F) = \beta + \frac{\alpha}{\kappa}\left[1 - (-\ln F)^{\kappa}\right] \tag{5.76}$$

ou, se F refere-se a probabilidades anuais de não superação e T ao período de retorno,

$$x(T) = \beta + \frac{\alpha}{\kappa}\left\{1 - \left[-\ln\left(1 - \frac{1}{T}\right)\right]^{\kappa}\right\} \tag{5.77}$$

Exemplo 5.9 – Empregue o modelo GEV para resolver o exemplo 5.8, supondo que o coeficiente de assimetria da variável X seja γ=1,40.
Solução: (a) Com γ=1,40, na Figura 5.13, obtém-se que o valor de κ que satisfaz a equação 5.73 está compreendido entre – 0,10 e 0. O *software* Microsoft Excel dispõe da função estatística LNGAMA, a qual corresponde ao logaritmo neperiano da função Gama de um certo argumento; nesse caso, a exponencial de LNGAMA(w) corresponde a Γ(w). Estabelecendo diversos valores de κ entre – 0,10 e 0 e, em seguida, usando a exponencial de LNGAMA(κ) para calcular o coeficiente de assimetria pela equação 5.73, nota-se que γ = 1,40 corresponde ao valor κ = -0,04. Com esse resultado e com Var[X] = 47025 $(m^3/s)^2$ na equação 5.74, resulta que α = 159,97. Finalmente, a equação 5.75 fornece β = 401,09. A vazão centenária é dada pela equação 5.77, ou seja, x (100) = 1209 m^3/s. (b) Representemos o fato de que as vazões superaram 600 m^3/s pelo evento A e que o evento B denote que as vazões superaram 800 m^3/s. Portanto, desejamos calcular a probabilidade condicional P(B|A), a qual pode ser posta sob a forma $P(B|A) = P(B \cap A)/P(A)$. O numerador dessa última equação é equivalente a P(B), ou seja, P(B) = 1-F_X(800) = 0,0886. O denominador é P(A) = 1-F_X(600) = 0,2571. Logo, P(B|A)= 0,345.

Exemplo 5.10 – Resolva o exemplo 5.5 para o caso em que as 'excedências' (Q-Q_0) têm média igual a 50 m^3/s, desvio padrão igual a 60 m^3/s e que são distribuídas de acordo com uma Distribuição Generalizada de Pareto.
Solução: Trata-se de um processo de Poisson com ν = 2 representando o número médio anual de ocorrências. Quando ocorrem, as excedências X = (Q-Q_0) são distribuídas de acordo com a Distribuição Generalizada de Pareto (DGP) cuja FAP é dada por $G_X(x) = 1 - \left[1 - \kappa\left(\frac{x}{\alpha}\right)\right]^{1/\kappa}$, na qual κ e α denotam, respectivamente, os parâmetros de forma e escala; para $\kappa > 0$, a variável é limitada por α/κ e para $\kappa < 0$ a variável é ilimitada à

direita, com cauda superior polinomial. Se $\kappa = 0$, a DGP torna-se a distribuição exponencial, com $G_X(x) = 1 - exp(-x/\alpha)$. A DGP recebeu essa denominação por seu emprego pioneiro em análise econômica pelo economista italiano Vilfredo Pareto (1848-1923) e, recentemente, tem sido utilizada na moderna teoria de valores extremos para caracterizar os 3 tipos de cauda superior. De fato, as 3 funções densidades ilustradas na Figura 5.9 são as formas possíveis da FDP de Pareto para κ positivo, nulo e negativo. Para uma variável de Pareto X, são válidas as seguintes relações:

$\alpha = \frac{E[X]}{2}\left[\frac{(E[X])^2}{\text{Var}[X]} + 1\right]$ e $\kappa = \frac{1}{2}\left[\frac{(E[X])^2}{\text{Var}[X]} - 1\right]$. Resolvendo essas duas equações, com $E[X] = 50$ e $\text{Var}[X] = 3600$, resulta que $\alpha = 42{,}36$ e $\kappa = -0{,}153$; portanto, temos, nesse caso, uma distribuição de 'excedências' com cauda superior ilimitada à direita e polinomial. Entretanto, tal como no exemplo 5.5, para calcular as vazões relacionadas a um certo tempo de retorno, é necessário, determinar a FAP das *excedências máximas anuais*, denotada por $F_{X\,max}(x)$, a qual, nos termos dos resultados parciais do exemplo 5.5, é dada por $F_{X\,max}(x) = exp\{-\nu[1 - G_X(x)]\}$. Se a FAP $G_X(x)$ é uma DGP, tem-se o chamado *modelo Poisson-Pareto* para séries de durações parciais, ou seja, $F_{Q\,max}(q) = exp\left\{-\nu\left[1 - \kappa\left(\frac{q - Q_0}{\alpha}\right)\right]^{1/\kappa}\right\}$, onde $Q_{max} = Q_0 + X$ representa a vazão máxima anual. Depois de simplificações semelhantes às realizadas no exemplo 5.5, resulta que $F_{Q\,max}(q) = exp\left\{-\left[1 - \kappa\left(\frac{q - \beta}{\alpha^*}\right)\right]\right\}$, a qual representa a FAP da *distribuição GEV*, com parâmetro de escala $\alpha^* = \alpha(\nu)^{-\kappa}$ e parâmetro de posição $\beta = Q_0 + (\alpha - \alpha^*)/\kappa$. Portanto, a modelação de séries de duração parcial, com número de ocorrências distribuídas de acordo com a FMP de Poisson e excedências distribuídas segundo uma DGP, tem como distribuição de máximos anuais a distribuição GEV. Para o problema em questão, $T = 100 \Rightarrow F_{Q\,max} = 1 - 1/100 = 0{,}99$; $\nu = 2$, $Q_0 = 60\,\text{m}^3/\text{s}$, $\alpha = 42{,}36$, $\kappa = -0{,}153$, $\alpha^* = 47{,}1$ e $\beta = 90{,}96$. Invertendo a FAP da GEV, obtém-se a função de quantis para essa distribuição, ou seja, $q(F) = \beta + \frac{\alpha^*}{\kappa}\{1 - [-ln(F)]^{\kappa}\}$. Substituindo os valores, tem-se que $q(F = 0{,}98) = 342{,}4\ \text{m}^3/\text{s}$. Portanto, a vazão centenária para esse caso a $342{,}4\ \text{m}^3/\text{s}$.

5.7.2.4 – Distribuição de Gumbel (Mínimos)

No caso de valores mínimos, a distribuição de Gumbel refere-se à forma assintótica limite para um conjunto de N variáveis aleatórias originais $\{X_1, X_2, \ldots, X_N\}$, independentes e igualmente distribuídas conforme um modelo $F_X(x)$ de cauda inferior exponencial. A distribuição de Gumbel (mínimos) é uma distribuição extremal bastante usada na análise de freqüência de eventos hidrológicos mínimos anuais.

A função de probabilidades acumuladas da distribuição de Gumbel (mínimos) é dada por

$$F_Z(z) = 1 - exp\left[-exp\left(\frac{z-\beta}{\alpha}\right)\right] \text{ para } -\infty < z < \infty, -\infty < \beta < \infty, \alpha > 0 \qquad (5.78)$$

na qual, α representa o parâmetro de escala e β o parâmetro de posição; de fato, β também é a moda de Z. A função densidade da distribuição de Gumbel (mínimos) é

$$f_Z(z) = \frac{1}{\alpha} exp\left[\frac{z-\beta}{\alpha} - exp\left(\frac{z-\beta}{\alpha}\right)\right] \qquad (5.79)$$

O valor esperado, a variância e o coeficiente de assimetria de Z são, respectivamente,

$$E[Z] = \beta - 0{,}5772\alpha \qquad (5.80)$$

$$\text{Var}[Z] = \sigma_Z^2 = \frac{\pi^2\alpha^2}{6} \qquad (5.81)$$

$$\gamma = -1{,}1396 \qquad (5.82)$$

Observe, portanto, que a distribuição Gumbel (mínimos) possui um coeficiente de assimetria negativo e constante. A Figura 5.14 ilustra a função densidade Gumbel, para alguns valores específicos dos parâmetros α e β.

A inversa da FAP de Gumbel (mínimos), ou função de quantis, é expressa por

$$z(F) = \beta + \alpha ln[-ln(1-F)] \text{ ou } y(T) = \beta + \alpha ln\left[-ln\left(1-\frac{1}{T}\right)\right] \qquad (5.83)$$

na qual, T denota o período de retorno em anos e F representa a probabilidade anual de não superação. Observe que, no caso de mínimos anuais, $T = 1/\text{P}(Z \le z) = 1/F_Z(z)$.

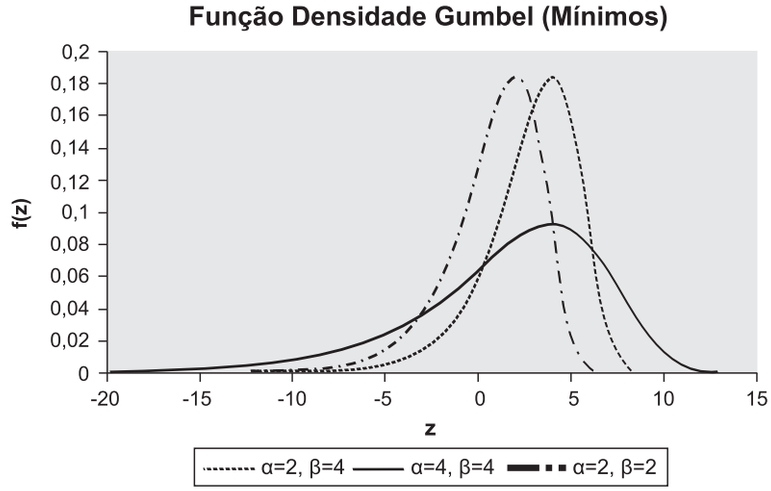


Figura 5.14 - Exemplos de funções densidades da distribuição de Gumbel (mínimos)

Exemplo 5.11 – Alguns estados brasileiros adotam como vazão de referência, para a outorga de direito de uso da água, a vazão média mínima anual de 7 dias de duração e de tempo de retorno 10 anos, geralmente representada por $Q_{7,10}$; para um dado ano de registros fluviométricos, o valor Q_7 anual corresponde à menor média de sete vazões consecutivas ocorridas naquele período. Suponha que as Q_7 anuais sejam denotadas pela variável aleatória Z e que, em um dado local, $E[Z] = 28{,}475$ m^3/s e $\sigma[Z] = 7{,}5956$ m^3/s. Calcule a vazão $Q_{7,10}$ pelo modelo de Gumbel (mínimos).

Solução: As soluções simultâneas do sistema formado pelas equações 5.80 e 5.81 resultam em $\alpha = 5{,}9223$ e $\beta = 31{,}8933$. Com esses valores e $T = 10$ anos na equação 5.83, conclui-se que a $Q_{7,10}$ pelo modelo de Gumbel (mínimos) é $z(T = 10) = 18{,}6$ m^3/s.

5.7.2.5 – Distribuição de Weibull (Mínimos)

No caso de valores mínimos, a distribuição de Weibull refere-se à forma assintótica limite para um conjunto de N variáveis aleatórias originais $\{X_1, X_2, \ldots, X_N\}$, independentes e igualmente distribuídas conforme um modelo $F_X(x)$ de cauda inferior limitada. A distribuição de extremos mínimos do Tipo III recebeu a denominação de distribuição de Weibull por ter sido usada pela primeira vez pelo engenheiro sueco Waloddi Weibull (1887-1979) na análise da resistência à fadiga de certos materiais. A constatação de que, em um cenário extremo, as vazões que escoam por uma seção fluvial são forçosamente limitadas inferiormente pelo

valor zero, faz com que a distribuição de Weibull seja uma candidata natural à modelação de eventos hidrológicos mínimos.

A função de probabilidades acumuladas da distribuição de Weibull é

$$F_Z(z) = 1 - exp\left[-\left(\frac{z}{\beta}\right)^{\alpha}\right] \text{ para } z \geq 0, \beta \geq 0 \text{ e } \alpha > 0 \tag{5.84}$$

na qual, β e α são, respectivamente, parâmetros de escala e forma; para $\alpha = 1$, a distribuição de Weibull é a exponencial com parâmetro de escala β. A função densidade da distribuição de Weibull é dada por

$$f_Z(z) = \frac{\alpha}{\beta}\left(\frac{z}{\beta}\right)^{\alpha-1} exp\left[-\left(\frac{z}{\beta}\right)^{\alpha}\right] \tag{5.85}$$

O valor esperado e a variância de uma variável de Weibull (mínimos) são dados, respectivamente, por

$$E[Z] = \beta\,\Gamma\left(1+\frac{1}{\alpha}\right) \tag{5.86}$$

$$\text{Var}[Z] = \beta^2\left[\Gamma\left(1+\frac{2}{\alpha}\right) - \Gamma^2\left(1+\frac{1}{\alpha}\right)\right] \tag{5.87}$$

Os coeficientes de variação e assimetria da distribuição de Weibull são

$$CV_Z = \frac{\sqrt{\Gamma\left(1+\frac{2}{\alpha}\right) - \Gamma^2\left(1+\frac{1}{\alpha}\right)}}{\Gamma\left(1+\frac{1}{\alpha}\right)} = \frac{\sqrt{B(\alpha) - A^2(\alpha)}}{A(\alpha)} \tag{5.88}$$

$$\gamma = \frac{\Gamma\left(1+\frac{3}{\alpha}\right) - 3\Gamma\left(1+\frac{2}{\alpha}\right)\Gamma\left(1+\frac{1}{\alpha}\right) + 2\Gamma^3\left(1+\frac{1}{\alpha}\right)}{\sqrt{\left[\Gamma\left(1+\frac{2}{\alpha}\right) - \Gamma^2\left(1+\frac{1}{\alpha}\right)\right]^3}} \tag{5.89}$$

A Figura 5.15 ilustra a função densidade da distribuição de Weibull para alguns conjuntos paramétricos específicos.

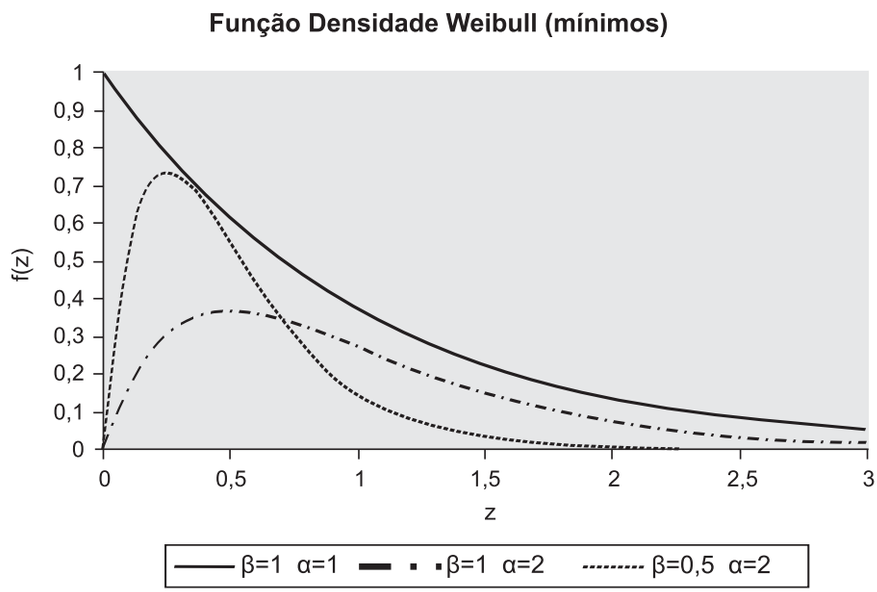


Figura 5.15 – Exemplos de funções densidades da distribuição de Weibull (mínimos)

Dados $E[Z]$ e Var$[Z]$, o cálculo dos parâmetros da distribuição de Weibull pode ser grandemente facilitado pelo tabelamento da equação 5.88, a qual relaciona o coeficiente de variação somente ao parâmetro α. A Tabela 5.2 mostra a variação de CV, $A(\alpha)$ e $B(\alpha)$ para um conjunto previamente especificado de valores possíveis do parâmetro α. Com efeito, conhecido o valor de CV, calcula-se inicialmente o parâmetro α pela Tabela 5.2 e, em seguida, determina-se o parâmetro β pela equação 5.86, ou seja,

$$\beta = \frac{E[Z]}{A(\alpha)} \tag{5.90}$$

Conhecidos os dois parâmetros, os quantis de Weibull (mínimos) podem ser calculados por

$$z(F) = \beta\,[-ln(1-F)]^{\frac{1}{\alpha}} \text{ ou } z(T) = \beta\left[-ln\left(1-\frac{1}{T}\right)\right]^{\frac{1}{\alpha}} \tag{5.91}$$

Exemplo 5.12 – Repita o exemplo 5.11 para o modelo de Weibull.
Solução: Com $E[Z]$ = 28,475 m³/s e s $[Z]$ = 7,5956 m³/s, calcula-se CV = 0,2667. Na Tabela 5.2, tomando-se a primeira linha com CV = 0,2667, obtém-se $A(\alpha)$ = 0,9093, $B(\alpha)$ = 0,8856 e α = 4,2301. Com $A(\alpha)$ = 0,9093 na equação 5.90, tem-se β =31,3153. Com os dois parâmetros na equação 5.91, conclui-se que a $Q_{7,10}$ pelo modelo de Weibull (mínimos) é z (T=10) = 18,4 m³/s.

Tabela 5.2 – Relações auxiliares para a estimativa do parâmetro de escala de Weibull

1/α	A(α)	B(α)	CV	1/α	A(α)	B(α)	CV	1/α	A(α)	B(α)	CV
0,000	1,0000	1,0000	0,0000	0,105	0,9493	0,9155	0,1259	0,210	0,9155	0,8863	0,2394
0,005	0,9971	0,9943	0,0063	0,110	0,9474	0,9131	0,1316	0,215	0,9143	0,8860	0,2446
0,010	0,9943	0,9888	0,0127	0,115	0,9454	0,9107	0,1372	0,220	0,9131	0,8858	0,2498
0,015	0,9915	0,9835	0,0190	0,120	0,9435	0,9085	0,1428	0,225	0,9119	0,8856	0,2549
0,020	0,9888	0,9784	0,0252	0,125	0,9417	0,9064	0,1483	0,230	0,9107	0,8856	0,2601
0,025	0,9861	0,9735	0,0315	0,130	0,9399	0,9044	0,1539	0,231	0,9105	0,8856	0,2611
0,030	0,9835	0,9687	0,0376	0,135	0,9381	0,9025	0,1594	0,232	0,9103	0,8856	0,2621
0,035	0,9809	0,9641	0,0438	0,140	0,9364	0,9007	0,1649	0,234	0,9098	0,8856	0,2642
0,040	0,9784	0,9597	0,0499	0,145	0,9347	0,8990	0,1703	0,235	0,9096	0,8856	0,2652
0,045	0,9759	0,9554	0,0559	0,150	0,9330	0,8974	0,1758	0,2355	0,9095	0,8856	0,2657
0,050	0,9735	0,9513	0,0619	0,155	0,9314	0,8960	0,1812	0,2360	0,9094	0,8856	0,2662
0,055	0,9711	0,9474	0,0679	0,160	0,9298	0,8946	0,1866	0,2361	0,9093	0,8856	0,2663
0,060	0,9687	0,9435	0,0739	0,165	0,9282	0,8933	0,1919	0,2362	0,9093	0,8856	0,2664
0,065	0,9664	0,9399	0,0798	0,170	0,9267	0,8922	0,1973	0,2363	0,9093	0,8856	0,2665
0,070	0,9641	0,9364	0,0857	0,175	0,9252	0,8911	0,2026	0,2364	0,9093	0,8856	0,2666
0,075	0,9619	0,9330	0,0915	0,180	0,9237	0,8901	0,2079	0,2364	0,9093	0,8856	0,2667
0,080	0,9597	0,9298	0,0973	0,185	0,9222	0,8893	0,2132	0,2364	0,9093	0,8856	0,2667
0,085	0,9575	0,9267	0,1031	0,190	0,9208	0,8885	0,2185	0,2364	0,9093	0,8856	0,2667
0,090	0,9554	0,9237	0,1088	0,195	0,9195	0,8878	0,2238	0,2364	0,9093	0,8856	0,2667
0,095	0,9533	0,9208	0,1146	0,200	0,9181	0,8872	0,2290	0,2364	0,9093	0,8856	0,2667
0,100	0,9513	0,9181	0,1203	0,205	0,9168	0,8867	0,2342	0,2364	0,9093	0,8856	0,2667

Se o limite inferior de Z é positivo e diferente de zero, a distribuição torna-se a Weibull de 3 parâmetros pela inclusão do terceiro parâmetro ξ. A função densidade e a função de probabilidades acumuladas passam a ser

$$f_Z(z) = \alpha\left(\frac{z-\xi}{\beta-\xi}\right)^{\alpha-1} exp\left[-\left(\frac{z-\xi}{\beta-\xi}\right)^{\alpha}\right] \text{ para } z > \xi, \beta \geq 0 \text{ e } \alpha > 0 \tag{5.92}$$

$$F_Z(z) = 1 - exp\left[-\left(\frac{z-\xi}{\beta-\xi}\right)^{\alpha}\right] \tag{5.93}$$

Os dois primeiros momentos dessa distribuição são

$$E[Z] = \xi + (\beta-\xi)\Gamma\left(1+\frac{1}{\alpha}\right) \tag{5.94}$$

$$\text{Var}[Z] = (\beta-\xi)^2\left[\Gamma\left(1+\frac{2}{\alpha}\right) - \Gamma^2\left(1+\frac{1}{\alpha}\right)\right] \tag{5.95}$$

os quais, de acordo com Haan (1977), podem ser postos sob as seguintes formas:

$$\beta = E[Z] + \sigma_Z C(\alpha) \tag{5.96}$$

$$\xi = \beta - \sigma_Z D(\alpha) \tag{5.97}$$

onde

$$C(\alpha) = D(\alpha)\left[1 - \Gamma\left(1 + \frac{1}{\alpha}\right)\right] \tag{5.98}$$

$$D(\alpha) = \frac{1}{\sqrt{\Gamma\left(1 + \frac{2}{\alpha}\right) - \Gamma^2\left(1 + \frac{1}{\alpha}\right)}} \tag{5.99}$$

O coeficiente de assimetria da distribuição de Weibull de 3 parâmetros continua sendo expresso pela equação 5.89, a qual é função unicamente de α. O cálculo dos parâmetros dessa distribuição é feito do seguinte modo: (i) inicialmente, com o valor do coeficiente de assimetria γ, determina-se α por meio da solução, por iterações numéricas, da equação 5.89; (ii) em seguida, $C(\alpha)$ e $D(\alpha)$ são calculados pelas equações 5.98 e 5.99; e (iii) finalmente, β e ξ são determinados pelas equações 5.96 e 5.97. Tais cálculos podem ser facilitados pela construção de uma tabela, semelhante à Tabela 5.2, relacionando o coeficiente de assimetria, o parâmetro α e as funções auxiliares $C(\alpha)$ e $D(\alpha)$.

5.8 – Distribuições de Pearson

O estatístico inglês Karl Pearson (1857-1936) propôs um sistema de distribuições de probabilidades, segundo o qual uma função densidade pode ser posta sob a forma

$$f_X(x) = exp\left[\int_{-\infty}^{x} \frac{x + a}{b_0 + b_1 x + b_2 x^2 \ldots} dt\right] \tag{5.100}$$

na qual, certos valores específicos dos coeficientes a, b_0, b_1, ... podem definir oito grandes famílias de distribuições que incluem a Normal, a Gama e a Beta. Essas famílias são comumente referidas na literatura estatística como Pearson Tipo I, Tipo II, e, assim por diante, até a Pearson Tipo VIII. De todo esse sistema de funções, as distribuições pertencentes à família Gama, ou distribuições Pearson Tipo III, estão entre aquelas que encontraram o maior número de aplicações na análise de freqüência de variáveis hidrológicas, com destaque para vazões e

precipitações máximas anuais. Em decorrência desse fato, destacaremos aqui duas distribuições do sistema Pearson de funções densidades, a saber, as distribuições Pearson Tipo III e Log-Pearson Tipo III.

5.8.1 – Distribuição Pearson Tipo III

Uma variável aleatória X possui uma distribuição de Pearson Tipo III se a variável $(X-\gamma)$ é distribuída conforme uma Gama com parâmetro de escala α e parâmetro de forma β; de fato, se o parâmetro de posição γ, da distribuição Pearson do Tipo III, for nulo, essa distribuição reduz-se a uma Gama. Por essa razão, a distribuição Pearson Tipo III também recebe o nome de Gama de 3 parâmetros. A função densidade de probabilidade de uma distribuição Pearson Tipo III é dada por

$$f_X(x) = \frac{1}{\alpha\Gamma(\beta)}\left(\frac{x-\gamma}{\alpha}\right)^{\beta-1} exp\left(-\frac{x-\gamma}{\alpha}\right) \tag{5.101}$$

A variável X é definida no intervalo $\gamma < x < \infty$. Em geral, o parâmetro de escala α pode ser positivo ou negativo. Entretanto, se $\alpha < 0$, a distribuição é limitada superiormente. A função de probabilidades acumuladas da distribuição Pearson Tipo III é expressa por

$$F_X(x) = \frac{1}{\alpha\Gamma(\beta)}\int_{\gamma}^{\infty}\left(\frac{x-\gamma}{\alpha}\right)^{\beta-1} exp\left(-\frac{x-\gamma}{\alpha}\right)dx \tag{5.102}$$

e pode ser avaliada do mesmo modo que o descrito no item 5.5, para a FAP da distribuição Gama. A Figura 5.16 ilustra alguns exemplos para a função densidade da distribuição Pearson Tipo III.

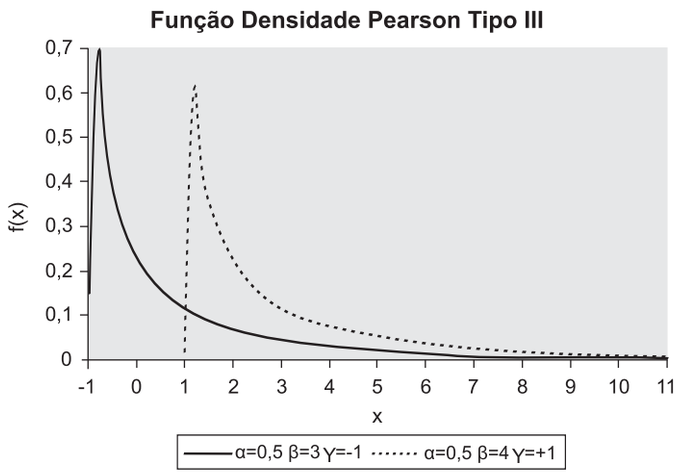


Figura 5.16 - Exemplos de funções densidades da distribuição Pearson Tipo III

A média, a variância e o coeficiente de assimetria de uma variável aleatória Pearson Tipo III são, respectivamente,

$$E[X] = \alpha\beta + \gamma \tag{5.103}$$

$$\mathrm{Var}[X] = \alpha^2\beta \tag{5.104}$$

$$\gamma = \frac{2}{\sqrt{\beta}} \tag{5.105}$$

5.8.2 – Distribuição Log-Pearson Tipo III

Se a variável ln(X), ou log(X), é distribuída segundo uma Pearson Tipo III, a distribuição da variável X é uma Log-Pearson Tipo III. A função densidade correspondente é dada por

$$f_X(x) = \frac{1}{\alpha x \Gamma(\beta)}\left[\frac{\ln(x)-\gamma}{\alpha}\right]^{\beta-1} \exp\left[-\frac{\ln(x)-\gamma}{\alpha}\right] \tag{5.106}$$

A função densidade da distribuição Log-Pearson Tipo III (LPIII) possui uma grande variedade de formas. Para a análise de freqüência de eventos hidrológicos máximos, somente as distribuições Log-Pearson Tipo III, com valores de β maiores do que 1 e valores de $1/\alpha$ maiores do que zero, são de interesse. Isso decorre do fato que valores negativos do coeficiente de assimetria implicam em $\alpha < 0$ e, por conseguinte, em um limite superior para a variável aleatória. A FAP da distribuição Log-Pearson Tipo III é dada por

$$F_X(x) = \frac{1}{\alpha\Gamma(\beta)}\int_0^x \frac{1}{x}\left[\frac{ln(x)-\gamma}{\alpha}\right]^{\beta-1} exp\left[-\frac{ln(x)-\gamma}{\alpha}\right]dx \tag{5.107}$$

Nessa equação, se $y = [ln(x) - \gamma]/\alpha$, a FAP Log-Pearson Tipo III torna-se

$$F_Y(y) = \frac{1}{\Gamma(\beta)}\int_0^y y^{\beta-1}\, exp(-y)\,dy \tag{5.108}$$

a qual pode ser avaliada pela equação 5.41, com $\theta = \alpha = 1$ e $\eta = \beta$. O valor esperado de uma variável Log-Pearson Tipo III é

$$E[X] = \frac{e^{\gamma}}{(1-\alpha)^{\beta}} \tag{5.109}$$

Os momentos de ordem superior são complexos. Bobée e Ashkar (1991) deduziram a seguinte expressão geral para os momentos, em relação à origem, de uma variável LPIII:

$$\mu_r' = \frac{e^{\gamma r}}{(1 - r\alpha)^{\beta}} \tag{5.110}$$

na qual, r denota a ordem do momento. Deve-se notar, entretanto, que, para essa distribuição, os momentos de ordem r não existem se $\alpha > 1/r$. O cálculo dos parâmetros de uma distribuição LPIII pode ser feito de dois modos: o indireto e o direto. O modo indireto, mais simples, é calcular os parâmetros da distribuição Pearson III, tal como aplicada aos logaritmos da variável X, ou seja, aplicar as equações 5.103 a 5.105 à variável transformada $Z = \ln(X)$ ou $Z = \log(X)$. O modo indireto é mais complexo e não será abordado no presente item; o leitor deve remeter-se às referências Bobée e Ashkar (1991), Kite (1977) e Rao e Hamed (2000) para detalhes com relação ao comportamento de uma variável LPIII.

O Conselho de Recursos Hídricos dos Estados Unidos da América (U.S. Water Resources Council, 1981) recomendou o uso da distribuição LPIII por parte das agências federais daquele país. Ao longo dos anos subseqüentes, tal fato tem gerado uma certa polêmica entre os especialistas da área e, conseqüentemente, produzido um volume considerável de pesquisas sobre esse modelo distributivo. Essas pesquisas abordam tópicos que vão desde os estudos comparativos entre métodos de estimação de parâmetros, quantis e intervalos de confiança, até temas relacionados à regionalização do coeficiente de assimetria, cuja determinação é essencial para o cálculo de probabilidades pela distribuição Log-Pearson Tipo III. A discussão de tais tópicos encontra-se além do escopo desta publicação e, por essa razão, o leitor interessado nesse tema, deve remeter-se, novamente, às referências Bobée e Ashkar (1991), Kite (1977) e Rao e Hamed (2000).

5.9 – Distribuições de Estatísticas Amostrais

Até aqui, estivemos tratando de distribuições de probabilidades que se prestam a representar o modo de variação de certas grandezas, formalizadas como variáveis aleatórias. As distribuições aqui descritas foram selecionadas com o intento de apresentar um elenco de modelos distributivos mais adequados à representação de variáveis hidrológicas. Existem, entretanto, outros problemas estatísticos, entre os quais destacam-se os testes de hipóteses e a construção de intervalos de confiança, que requerem distribuições de probabilidades particulares. Tais distribuições, freqüentemente denominadas distribuições de *estatísticas amostrais*,

não são utilizadas para a modelação de variáveis hidrológicas, mas são úteis na solução de outros problemas estatísticos que as concernem; esses problemas serão abordados em capítulos subseqüentes. Entre as distribuições de estatísticas amostrais, serão destacadas aqui as distribuições do Qui-Quadrado χ^2, de *t* de Student e de *F* de Snedecor.

5.9.1 – Distribuição do Qui-Quadrado χ^2

Se, para $X_i \sim \mathbf{N}(\mu, \sigma)$, $Z_i = \dfrac{X_i - \mu}{\sigma}$, com $i = 1, 2, \ldots, N$, representa um conjunto de N variáveis aleatórias independentes e identicamente distribuídas conforme uma distribuição Normal padrão, então, demonstra-se que a variável Y definida por

$$Y = \sum_{i=1}^{N} Z_i^2 \tag{5.111}$$

segue uma distribuição do χ^2, cuja função densidade de probabilidade depende apenas do parâmetro ν e tem como expressão

$$f_{\chi^2}(y) = \frac{y^{\frac{\nu}{2}-1} e^{-\frac{y}{2}}}{2^{\frac{\nu}{2}} \Gamma(\nu/2)} \text{ para } y \text{ e } \nu > 0 \tag{5.112}$$

O parâmetro ν recebe a denominação de '*número de graus de liberdade*' por mera analogia a esse conceito originário da mecânica racional, relativo ao número de movimentos possíveis de um corpo sólido. A distribuição do χ^2 é um caso especial da distribuição Gama (ver equação 5.39), com $\eta = \nu/2$ e $\theta = 2$. Por essa razão, a função de probabilidades acumuladas da distribuição do χ^2 pode ser posta nos termos da FAP da distribuição Gama (ver equação 5.42), ou seja

$$F_{\chi^2}(y) = \frac{\Gamma_i(u = y/2, \eta = \nu/2)}{\Gamma(\eta = \nu/2)} \tag{5.113}$$

e calculada como o quociente entre as funções Gama incompleta e completa, tal como ilustrado no item 5.5. O Anexo 6 desse boletim técnico apresenta uma tabela da função de probabilidades acumuladas da distribuição do χ^2, para diferentes graus de liberdade.

O valor esperado, a variância e o coeficiente de assimetria da distribuição do χ^2 são

$$E[\chi^2] = \nu \tag{5.114}$$

$$\mathrm{Var}[\chi^2] = 2\nu \tag{5.115}$$

$$\gamma = \frac{2}{\sqrt{\nu/2}} \tag{5.116}$$

A Figura 5.17 ilustra as formas possíveis da distribuição do χ^2, para alguns valores de ν.

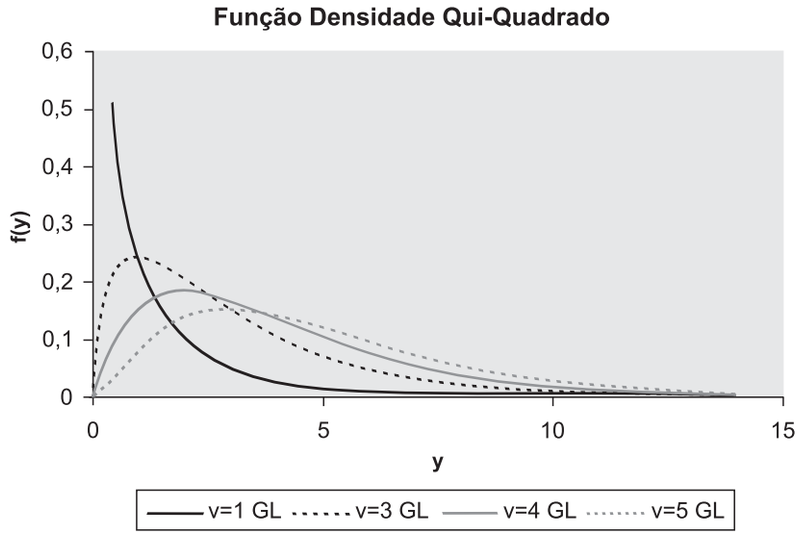


Figura 5.17 – Exemplos de funções densidades da distribuição do χ^2

Se agora, diferentemente de sua definição anterior, as variáveis Z_i forem definidas por $Z_i = \dfrac{X_i - \bar{x}}{\sigma}$, para $i = 1, 2, \ldots, N$, onde X_i representam elementos de uma amostra aleatória simples de uma população Normal, cuja média aritmética é $\bar{x}$, então, é possível demonstrar que a variável $Y = \sum_{i=1}^{N} Z_i^2$ segue uma distribuição do χ^2, com $\nu = (N\text{-}1)$ graus de liberdade. Diz-se, nesse caso, que temos um grau de liberdade a menos pelo fato da média populacional μ ter sido estimada pela média aritmética amostral $\bar{x}$. Além disso, relembrando que a variância amostral é dada por $s_X^2 = \sum_{i=1}^{N}(X_i - \bar{x}) \Big/ (N-1)$ e que $Y = \sum_{i=1}^{N} \left(X_i - \bar{x} \right)^2 \Big/ \sigma^2$, é fácil verificar que

$$Y = (N-1)\frac{s_X^2}{\sigma_X^2} \tag{5.117}$$

segue uma distribuição do χ^2, com $\nu = (N\text{-}1)$ graus de liberdade. Esse resultado será usado extensivamente na formulação e implementação de testes de hipóteses e construção de intervalos de confiança para a variância de populações Normais.

5.9.2 – Distribuição do *t* de Student

Se $U \sim \mathbf{N}(0,1)$ e $V \sim \chi^2(\sigma)$ são variáveis aleatórias independentes, então, demonstra-se que a função densidade de probabilidades da variável T, definida por $T = U\sqrt{\nu}/\sqrt{V}$, é dada por

$$f_T(t) = \frac{\Gamma[(\nu+1)/2](1+t^2/\nu)^{-(\nu+1)/2}}{\sqrt{\pi\nu}\,\Gamma(\nu/2)} \quad \text{para} -\infty < t < \infty \text{ e } \nu > 0 \tag{5.118}$$

a qual, individualiza a distribuição *t* de Student, com parâmetro ν. Essa distribuição é devida ao químico inglês William Gosset (1876-1937), que assinava seus artigos e contribuições ao conhecimento estatístico, sob o pseudônimo de Student. A função de probabilidades acumuladas é dada pela integral de $-\infty$ a t da densidade expressa pela equação 5.118 e pode ser avaliada apenas numericamente. O Anexo 7 apresenta uma tabela da FAP de Student sob a forma $\alpha = F_T(t)$, para diversos valores de α e ν.

A média e a variância de uma variável de Student são dadas, respectivamente, por

$$E[T] = 0 \tag{5.119}$$

$$\mathrm{Var}[T] = \frac{\nu}{\nu-2} \tag{5.120}$$

Trata-se de uma distribuição simétrica em relação à origem de *t* e que aproxima-se da distribuição Normal padrão, para valores elevados de ν. A Figura 5.18 apresenta os gráficos da função densidade do *t* de Student, para alguns valores do parâmetro ν.

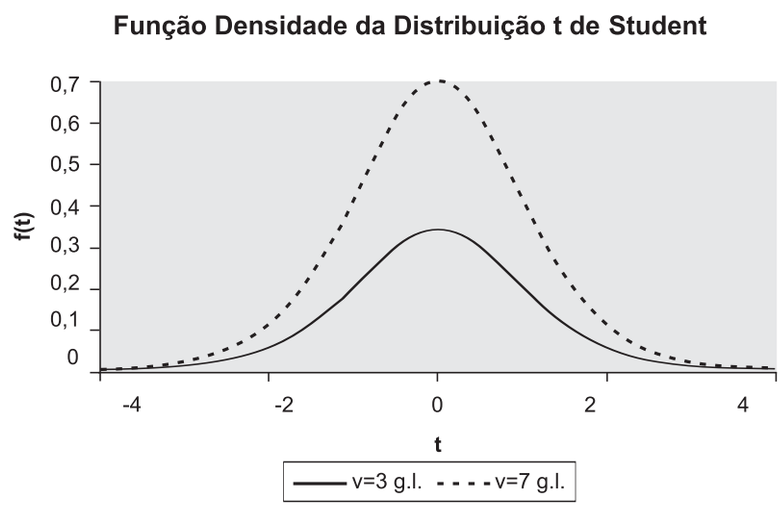


Figura 5.18 – Exemplos da função densidade *t* de Student

A distribuição *t* de Student é usada como *distribuição de amostragem da média de uma população Normal*, com variância desconhecida. De fato, se a variável *T* é expressa sob a forma

$$T = \frac{\bar{x} - \mu_X}{\sqrt{s_X^2 / N}} \tag{5.121}$$

sendo, em seguida, multiplicada e dividida por σ_X, obtém-se

$$T = \frac{\dfrac{\bar{x} - \mu_X}{\sigma_X / \sqrt{N}}}{\sqrt{s_X^2 / \sigma_X^2}} = \frac{U\sqrt{N-1}}{\sqrt{V}} \tag{5.122}$$

que corresponde à definição da variável *T*; recorde-se que $U = \left(\bar{x} - \mu_X\right) / \left(\sigma\sqrt{N}\right)$ é uma variável Normal padrão (ver exemplo 5.3) e que $V = (N-1)\, s_x^2 / \sigma_X^2$ segue uma distribuição do χ^2, com (*N*-1) graus de liberdade, conforme equação 5.117. Comparando-se a equação 5.122 com a definição da variável de Student, verifica-se, portanto, que a distribuição de amostragem da média de uma população Normal, com variância desconhecida, é a distribuição *t* de Student, com (*N*-1) graus de liberdade. Nesse caso, tem-se um grau de liberdade a menos, pelo fato da variância populacional ter sido estimada por s_X^2.

Exemplo 5.13 – De volta à solução do exemplo 5.3, constate o fato que a variância populacional da variável 'concentração de oxigênio dissolvido' foi estimada pela variância amostral e calcule a probabilidade que a amostra de 8

semanas de monitoramento produza uma média aritmética que se diferencie da verdadeira média populacional em pelo menos 0,5 mg/l.
Solução: Continua válido o raciocínio exposto na solução do exemplo 5.3, à exceção do fato que, agora, a variável $T = \frac{\bar{x} - \mu_X}{s_X / \sqrt{n}} \sim t$ de Student, com n-1 = 7 graus de liberdade. A probabilidade pedida corresponde a $P(|\bar{x} - \mu_X| > 0{,}5)$; dividindo os termos dessa inequação por $s_X / \sqrt{n}$, resulta que a probabilidade solicitada é equivalente a $P\left(|T| > \frac{0{,}5}{s_X / \sqrt{n}}\right)$ ou $P(|T| > 0{,}707)$, ou ainda $1 - P(|T| < 0{,}707)$. Para calcular probabilidades ou a função inversa da FAP de Student, pode-se fazer uso da tabela do Anexo 6 ou das funções estatísticas DISTT e INVT do *software* Microsoft Excel; em particular, para $\nu = 7$ e para $t = 0{,}707$, a função DISTT, com opção bilateral, retorna o valor 0,502. Portanto, a probabilidade que a amostra de 8 semanas de monitoramento produza uma média aritmética que se diferencie da verdadeira média populacional em pelo menos 0,5 mg/l é (1-0,502) = 0,498.

5.9.3 – Distribuição *F* de Snedecor

Se $U \sim \chi^2$ com m graus de liberdade, $V \sim \chi^2$ com n graus de liberdade e se essas variáveis são independentes, então, demonstra-se que a variável definida por

$$Y = \frac{U/m}{V/n} \tag{5.123}$$

segue a distribuição F, com parâmetros $\gamma_1 = m$ e $\gamma_2 = n$, cuja função densidade é dada por

$$f_F(f) = \frac{\Gamma[(\gamma_1 + \gamma_2)/2]}{\Gamma(\gamma_1/2)\Gamma(\gamma_2/2)} \gamma_1^{\gamma_1/2} \gamma_2^{\gamma_2/2} f^{(\gamma_1 - 2)/2} (\gamma_2 + \gamma_1 f)^{-(\gamma_1 + \gamma_2)/2} \text{ para } \gamma_1, \gamma_2, f > 0 \tag{5.124}$$

A função de probabilidades acumuladas é dada pela integral de 0 a f da densidade expressa pela equação 5.124 e pode ser avaliada apenas numericamente. O Anexo 8 apresenta uma tabela da FAP da distribuição F, para diversos valores de γ_1 e γ_2, denominados, respectivamente, graus de liberdade do numerador e do denominador. A média e a variância de uma variável aleatória F são, respectivamente,

$$E[F] = \frac{\gamma_1}{\gamma_2 - 2} \tag{5.125}$$

$$\mathrm{Var}[X] = \frac{\gamma_2^2(\gamma_1 + 2)}{\gamma_1(\gamma_2 - 2)(\gamma_2 - 4)} \tag{5.126}$$

A Figura 5.19 ilustra a função densidade *F* para alguns conjuntos paramétricos específicos.

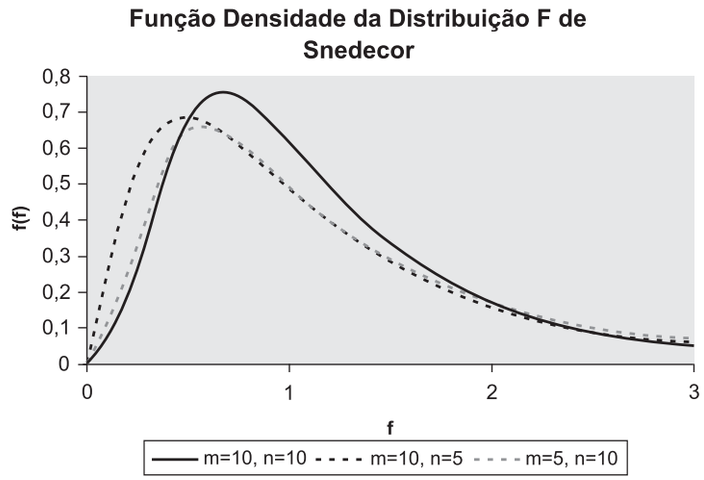


Figura 5.19 – Exemplos da função densidade *F*

Essa distribuição foi proposta pelo estatístico americano William Snedecor (1882-1974) como distribuição de amostragem do quociente entre variâncias de duas populações normais; a denominação *F* decorre de uma homenagem ao estatístico inglês Ronald Fisher. A distribuição *F* é usada para testes de hipóteses relativos à comparação de variâncias de populações normais diferentes, assim como para a análise de variância e dos resíduos de regressões.

5.10 – Distribuição Normal Bivariada

A distribuição conjunta de duas variáveis aleatórias normais é denominada distribuição Normal bivariada. Formalmente, se *X* e *Y* possuem distribuições marginais Normais, com respectivos parâmetros μ_X, σ_X, μ_Y e σ_Y, e se o coeficiente de correlação entre as variáveis é representado por ρ, a função densidade da distribuição Normal bivariada é

$$f_{X,Y}(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \times$$
$$\times exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x-\mu_X}{\sigma_X}\right)^2 - 2\rho\frac{(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y} + \left(\frac{y-\mu_Y}{\sigma_Y}\right)^2\right]\right\} \quad (5.127)$$

para $\infty < x < \infty$ e $\infty < y < \infty$. As probabilidades conjuntas P(X< x, Y< y) são dadas pela integração dupla da função densidade da distribuição Normal bivariada e requerem métodos numéricos para sua avaliação. Alguns programas de computador que implementam rotinas de integração dupla da densidade Normal bivariada estão disponíveis na Internet para *download*. A URL http://stat-athens.aueb.gr/~karlis/morematerial.html oferece uma lista de tópicos relacionados à distribuição Normal bivariada e disponibiliza para *download* o programa Bivar1b.exe, elaborado pelo Instituto Nacional de Saúde Ocupacional da Dinamarca, o qual executa o cálculo da FAP conjunta das variáveis X e Y.

A Figura 5.20 ilustra a função densidade Normal bivariada para três diferentes valores do coeficiente de correlação. Observe que, quando as variáveis X e Y são independentes, o volume da função densidade se distribui simetricamente e de modo mais disperso em torno da origem das variáveis. À medida que a dependência linear entre as variáveis cresce, os pares (x,y) e suas respectivas probabilidades de não superação, dadas pelos volumes abaixo da superfície da densidade bivariada, concentram-se ao longo da projeção da reta de dependência, no plano xy.

Usando a equação 3.34, é fácil mostrar que as distribuições marginais são as respectivas distribuições normais univariadas de X e Y. Por outro lado, as distribuições condicionais são obtidas pela aplicação da equação 3.44.

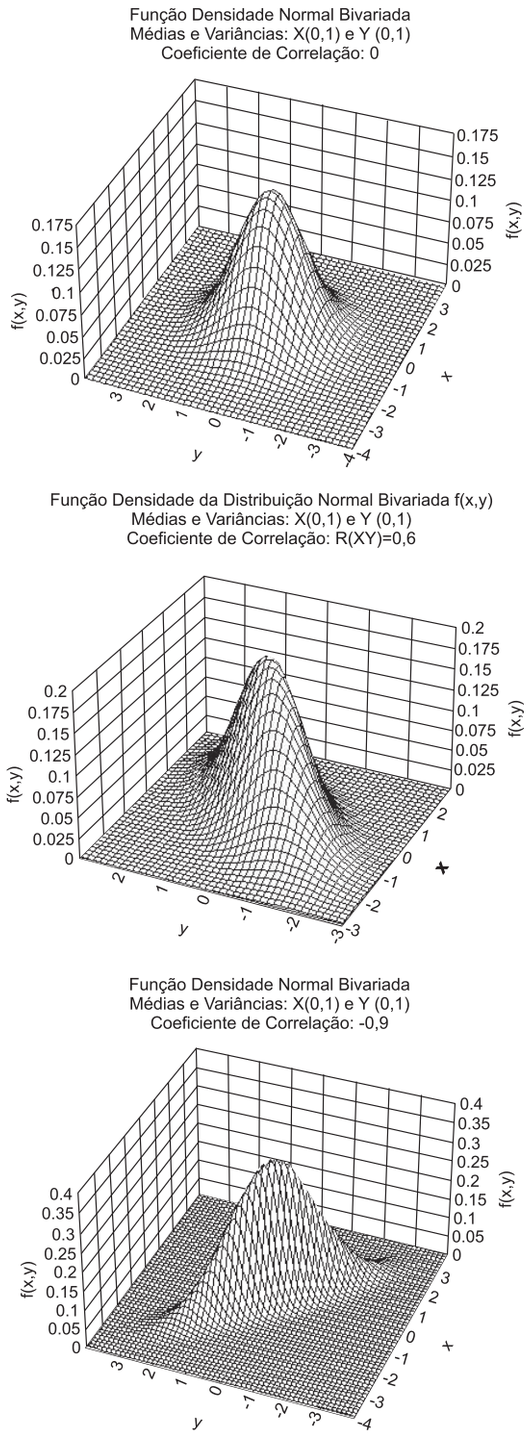


Figura 5.20 – Exemplos de funções densidades conjuntas da distribuição Normal bivariada

5.11 - Sumário das Características Principais das Distribuições

Apresenta-se a seguir um sumário das características das principais distribuições de probabilidades de variáveis aleatórias contínuas, descritas no presente capítulo. A distribuição Wakeby, de 5 parâmetros, e a mistura de duas distribuições de valores extremos TCEV ('*Two-component Extreme Value*') são exemplos de algumas outras distribuições de probabilidades, que não foram descritas nesse capítulo e que são úteis na modelação de variáveis aleatórias hidrológicas; o leitor deve remeter-se à referência Rao e Hamed (2000) para detalhes sobre a primeira e a Rossi et al. (1984) para a descrição da segunda. A exemplo do resumo das distribuições de variáveis aleatórias discretas do capítulo 4, nem todas as características que constam do sumário a seguir foram discutidas ou demonstradas no texto principal. Portanto, a intenção desse sumário é a de ser um item de referência para uso das distribuições de variáveis aleatórias contínuas.

5.11.1 – Distribuição Uniforme

Notação: $X \sim U(a,b)$

Parâmetros: a e b

FDP: $f_X(x) = \dfrac{1}{b-a}$ se $a \le x \le b$

Média: $E[X] = \dfrac{a+b}{2}$

Variância: $\mathrm{Var}[X] = \dfrac{(b-a)^2}{12}$

Coeficiente de Assimetria: $\gamma = 0$

Curtose: $\kappa = 1{,}8$

Função Geratriz de Momentos: $\phi(t) = \dfrac{e^{bt} - e^{at}}{t(b-a)}$

5.11.2 – Distribuição Normal

Notação: $X \sim N(\mu,\sigma)$

Parâmetros: μ e σ

FDP: $f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma} \right)^2 \right]$ para $-\infty < x < \infty$

Média: $E[X] = \mu$

Variância: $\text{Var}[X] = \sigma^2$

Coeficiente de Assimetria: $\gamma = 0$

Curtose: $\kappa = 3$

Função Geratriz de Momentos: $\phi(t) = exp\left(\mu t + \frac{\sigma^2 t^2}{2} \right)$

5.11.3 – Distribuição Log-Normal (2 parâmetros)

Notação: $X \sim LN(\mu_Y, \sigma_Y)$

Parâmetros: μ_Y e σ_Y, com $Y = \ln(X)$

FDP: $f_X(x) = \frac{1}{x\sigma_{\ln(X)}\sqrt{2\pi}} exp\left\{ -\frac{1}{2}\left[\frac{\ln(X) - \mu_{\ln(X)}}{\sigma_{\ln(X)}} \right] \right\}$ para $x > 0$

Média: $E[X] = \mu_X = exp\left[\mu_{\ln(X)} + \frac{\sigma^2_{\ln(X)}}{2} \right]$

Variância: $\text{Var}[X] = \sigma_X^2 = \mu_X^2 \left[exp\left(\sigma^2_{\ln(X)} \right) - 1 \right]$

Coeficiente de Variação: $CV_X = \sqrt{\exp\left[\sigma^2_{\ln(X)} \right] - 1}$

Coeficiente de Assimetria: $\gamma = 3CV_X + \left(CV_X \right)^3$

Curtose: $\kappa = 3 + \left(e^{\sigma^2_{\ln(X)}} - 1 \right) \left(e^{3\sigma^2_{\ln(X)}} + 3e^{2\sigma^2_{\ln(X)}} + 6e^{\sigma^2_{\ln(X)}} + 6 \right)$

5.11.4 – Distribuição Exponencial

Notação: $X \sim \mathbf{E}(\theta)$

Parâmetro: θ

FDP: $f_X(x) = \frac{1}{\theta} exp\left(-\frac{x}{\theta}\right), x \geq 0$

Função de Quantis: $x(F) = -\theta ln(1-F)$

Média: $E[X] = \theta$

Variância: $\mathrm{Var}[X] = \theta^2$

coeficiente de Assimetria: $\gamma = 2$

Curtose: $\kappa = 9$

Função Geratriz de Momentos: $\phi(t) = \frac{1}{1-\theta t}$ para $t < \frac{1}{\theta}$

5.11.5 – Distribuição Gama

Notação: $X \sim$ **Ga** (θ, η)

Parâmetros: θ e η

FDP: $f_X(x) = \frac{(x/\theta)^{\eta-1} exp(-x/\theta)}{\theta \Gamma(\eta)}$ para x, θ e $\eta > 0$

Média: $E[X] = \eta\theta$

Variância: $\mathrm{Var}[X] = \eta\theta^2$

Coeficiente de Assimetria: $\gamma = \frac{2}{\sqrt{\eta}}$

Curtose: $\kappa = 3 + \frac{6}{\eta}$

Função Geratriz de Momentos: $\phi(t) = \left(\frac{1}{1-\theta t}\right)^{\eta}$ para $t < \frac{1}{\theta}$

5.11.6 – Distribuição Beta

Notação: $X \sim$ **Be** (α, β)

Parâmetros: α e β

FDP: $f_X(x) = \frac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1}$ para $0 \le x \le 1, \alpha > 0, \beta > 0$ e

$B(\alpha,\beta) = \int_0^1 t^{\alpha-1}(1-t)^{\beta-1} dt$

Média: $E[X] = \frac{\alpha}{\alpha+\beta}$

Variância: $\mathrm{Var}[X] = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$

Coeficiente de Assimetria: $\gamma = \frac{2(\beta-\alpha)\sqrt{\alpha+\beta+1}}{\sqrt{\alpha\beta}(\alpha+\beta+2)}$

Curtose: $\kappa = \frac{3(\alpha+\beta+1)[2(\alpha+\beta)^2 + \alpha\beta(\alpha+\beta-6)]}{\alpha\beta(\alpha+\beta+2)(\alpha+\beta+3)}$

5.11.7 – Distribuição Gumbel (Máximos)

Notação: $Y \sim \mathbf{Gu_{max}}(\alpha, \beta)$

Parâmetros: α e β

FDP: $f_Y(y) = \frac{1}{\alpha} exp\left[-\frac{y-\beta}{\alpha} - exp\left(-\frac{y-\beta}{\alpha}\right)\right]$

Função de Quantis: $y(F) = \beta - \alpha ln[-ln(F)]$

Média: $E[Y] = \beta + 0{,}5772\alpha$

Variância: $\mathrm{Var}[Y] = \sigma_Y^2 = \frac{\pi^2\alpha^2}{6}$

Coeficiente de Assimetria: $\gamma = 1{,}1396$

Curtose: $\kappa = 5{,}4$

5.11.8 – Distribuição Generalizada de Valores Extremos (Máximos)

Notação: $Y \sim \mathbf{GEV}(\alpha, \beta, \kappa)$

Parâmetros: α, β e κ

FDP: $f_Y(y) = \frac{1}{\alpha}\left[1-\kappa\left(\frac{y-\beta}{\alpha}\right)\right]^{1/\kappa-1} exp\left\{-\left[1-\kappa\left(\frac{y-\beta}{\alpha}\right)\right]^{1/\kappa}\right\}$ se $\hat{e} \neq 0$ e

$f_Y(y) = \frac{1}{\alpha} exp\left[-\frac{y-\beta}{\alpha} - exp\left(-\frac{y-\beta}{\alpha}\right)\right]$ se $\kappa = 0$

Função de Quantis: $x(F) = \beta + \frac{\alpha}{\kappa}\left[1-(-ln\,F)^{\kappa}\right]$

Média: $E[Y] = \beta + \frac{\alpha}{\kappa}\left[1-\Gamma(1+\kappa)\right]$

Variância: $\mathrm{Var}[Y] = \left(\frac{\alpha}{\kappa}\right)^2\left[\Gamma(1+2\kappa)-\Gamma^2(1+\kappa)\right]$

Coeficiente de Assimetria: $\gamma = \langle \text{sinal de } \kappa \rangle \frac{-\Gamma(1+3\kappa)+3\Gamma(1+\kappa)\Gamma(1+2\kappa)-2\Gamma^3(1+\kappa)}{\left[\Gamma(1+2\kappa)-\Gamma^2(1+\kappa)\right]^{3/2}}$

5.11.9 – Distribuição Gumbel (Mínimos)

Notação: $Z \sim \mathbf{Gu}_{\mathbf{min}}(\alpha, \beta)$

Parâmetros: α e β

FDP: $f_Z(z) = \frac{1}{\alpha} exp\left[\frac{z-\beta}{\alpha} - exp\left(\frac{z-\beta}{\alpha}\right)\right]$

Função de Quantis: $z(F) = \beta + \alpha\, ln\left[-ln(1-F)\right]$

Média: $E[Z] = \beta - 0{,}5772\alpha$

Variância: $\mathrm{Var}[Z] = \sigma_Z^2 = \frac{\pi^2\alpha^2}{6}$

Coeficiente de Assimetria: $\gamma = -1{,}1396$

Curtose: $\kappa = 5{,}4$

5.11.10 – Distribuição Weibull (Mínimos) de 2 parâmetros

Notação: $Z \sim \mathbf{W}_{\mathbf{min}}(\alpha\,, \beta)$

Parâmetros: α e β

FDP: $f_Z(z) = \frac{\alpha}{\beta}\left(\frac{z}{\beta}\right)^{\alpha-1} exp\left[-\left(\frac{z}{\beta}\right)^{\alpha}\right]$

Função de Quantis: $z(F) = \beta\,[-ln(1-F)]^{\frac{1}{\alpha}}$

Média: $E[Z] = \beta\,\Gamma\left(1+\frac{1}{\alpha}\right)$

Variância: $\mathrm{Var}[Z] = \beta^2\left[\Gamma\left(1+\frac{2}{\alpha}\right) - \Gamma^2\left(1+\frac{1}{\alpha}\right)\right]$

Coeficiente de Assimetria: $\gamma = \frac{\Gamma\left(1+\frac{3}{\alpha}\right) - 3\Gamma\left(1+\frac{2}{\alpha}\right)\Gamma\left(1+\frac{1}{\alpha}\right) + 2\Gamma^3\left(1+\frac{1}{\alpha}\right)}{\sqrt{\left[\Gamma\left(1+\frac{2}{\alpha}\right) - \Gamma^2\left(1+\frac{1}{\alpha}\right)\right]^3}}$

5.11.11 – Distribuição Pearson Tipo III

Notação: $X \sim$ **PIII** (α, β, γ)

Parâmetros: α, β e γ

FDP: $f_X(x) = \frac{1}{\alpha\Gamma(\beta)}\left(\frac{x-\gamma}{\alpha}\right)^{\beta-1} exp\left(-\frac{x-\gamma}{\alpha}\right)$

Média: $E[X] = \alpha\beta + \gamma$

Variância: $\mathrm{Var}[X] = \alpha^2\beta$

Coeficiente de Assimetria: $\gamma = \frac{2}{\sqrt{\beta}}$

Curtose: $\kappa = 3 + \frac{6}{\sqrt{\beta}}$

5.11.12 – Distribuição do χ^2

Notação: $Y \sim \chi^2(\nu)$

Parâmetro: ν

FDP: $f_{\chi^2}(y) = \dfrac{y^{\frac{\nu}{2}-1} e^{-\frac{y}{2}}}{2^{\frac{\nu}{2}}\Gamma(\nu/2)}$ *para* y *e* $\nu > 0$

Média: $E[\chi^2] = \nu$

Variância: $\mathrm{Var}[\chi^2] = 2\nu$

Coeficiente de Assimetria: $\gamma = \dfrac{2}{\sqrt{\nu/2}}$

5.11.13 – Distribuição do *t* de Student

Notação: $T \sim \boldsymbol{t}(\nu)$

Parâmetro: ν

FDP: $f_T(t) = \dfrac{\Gamma[(\nu+1)/2](1+t^2/\nu)^{-(\nu+1)/2}}{\sqrt{\pi\nu}\Gamma(\nu/2)}$ para $-\infty < t < \infty$ e $\nu > 0$

Média: $E[T] = 0$

Variância: $\mathrm{Var}[T] = \dfrac{\nu}{\nu-2}$

Coeficiente de Assimetria: $\gamma = 0$

5.11.14 – Distribuição *F* de Snedecor

Notação: $F \sim \boldsymbol{F}(\gamma_1, \gamma_2)$

Parâmetros: γ_1 e γ_2

FDP:

$f_F(f) = \dfrac{\Gamma[(\gamma_1+\gamma_2)/2]}{\Gamma(\gamma_1/2)\Gamma(\gamma_2/2)} \gamma_1^{\gamma_1/2} \gamma_2^{\gamma_2/2} f^{(\gamma_1-2)/2} (\gamma_2 + \gamma_1 f)^{-(\gamma_1+\gamma_2)/2}$ para $\gamma_1, \gamma_2, f > 0$

Média: $E[F] = \dfrac{\gamma_1}{\gamma_2 - 2}$

Variância: $\mathrm{Var}[X] = \dfrac{\gamma_2^2(\gamma_1+2)}{\gamma_1(\gamma_2-2)(\gamma_2-4)}$

Exercícios

1) Suponha que a concentração média diária de ferro em um trecho fluvial, representada por X, varie uniformemente entre 2 e 4 mg/l. Pede-se (a) calcular a média e a variância de X; (b) a probabilidade de X superar 3,5 mg/l; e (c) dado que, em um certo dia, a concentração de ferro temperatura já superou 3mg/l, calcular P($X \geq 3{,}5$ mg/l).

2) Além das aproximações descritas no item 5.2, a integração numérica da função densidade da variável normal central reduzida pode ser realizada através de qualquer um dos métodos tradicionais de integração numérica (regra trapezoidal ou regra de Simpson). Entretanto, o cálculo numérico de integrais impróprias exige transformação de variáveis de forma a tornar finito o limite de integração. Para essa finalidade e sob a condição que a função a ser integrada decresça a zero pelo menos tão rapidamente quanto $1/x^2$, quando x tende para infinito, utiliza-se, geralmente, a seguinte identidade:

$$\int_a^b f(x)dx = \int_{\frac{1}{b}}^{\frac{1}{a}} \frac{1}{t^2} f\left(\frac{1}{t}\right)dt, \quad \text{para} \quad ab > 0 \tag{5.128}$$

Para o caso da integração de $-\infty$ até um valor positivo, a integração pode ser feita em duas etapas. Por exemplo, considere a integração

$$\int_{-\infty}^{b} f(x)dx = \int_{-\infty}^{-A} f(x)dx + \int_{-A}^{b} f(x)dx \tag{5.129}$$

onde $-A$ é um valor negativo suficientemente grande tal que a premissa de decréscimo da função seja válida. A primeira integral em 5.129 pode ser calculada através do artifício da equação 5.128 e a segunda integral através do método de Simpson, por exemplo. A seguir, você encontrará o código fonte em Fortran de um programa de computador. Refaça e/ou compile esse programa em uma linguagem de programação que você conheça, para integrar numericamente a FDP da variável normal padrão, utilizando as equações 5.128 e 5.129.

```
c Calculo da Integral da Distribuição Normal N(0,1)
c
c Esse programa calcula P(X<x), dado x, onde X é uma variável
c normal padrão, ou seja X~N(0,1). O cálculo é feito por meio da
c avaliação numérica de duas integrais : I1, de -∞ a -4, usando
c transformação de variável e I2, de -4 a x, usando a regra 1/3
c de Simpson, com um número fixo de 500 segmentos. O resultado
```

```
c final é a soma (I1+I2), multiplicado pela raiz quadrada de 1/2p.
c
      Program Normal
      external func,transf
 99   do 2 j=1,24
       write(*,*)
  2   continue
      write(*,*) 'Digite o valor de x da variavel aleatoria normal X'
      read(*,*) c
      do 3 j=1,24
       write(*,*)
  3   continue
      xl=-1./4.
            b=-4.
c definindo o limite inferior de -1/4 para a integral I1 e -4 para I2
      xh=0.0
c definindo o limite superior de 0 para a integral I1
              write(*,*) 'FUNCAO NORMAL PADRÃO DE PROBABILIDADES
ACUMULADAS'
              write(*,*) '_____'
              write(*,*)
              write(*,*) '       Resultados da Integração Numérica'
              write(*,*)
              write(*,*) '     x                P(X<x)'
              write(*,*) '_____      _____'
              write(*,*)
      call lefti(transf,xl,xh,res1)
      call righti(func,b,c,res2)
       res=(res1+res2)/sqrt(2.*3.14592654)
      write(*,'(2x,f8.3,11x,f7.3)') c,res
      write(*,*)
      write(*,*)
      write(*,*) 'Deseja executar o programa para novo x? sim=1,não=0'
      read(*,*) iq
      if(iq.eq.1) goto 99
      end
c subrotina para calcular a cauda esquerda I1
      subroutine lefti(transf,xl,xh,res1)
            nn=49
             xhl=(xh-xl)/(float(nn)+1)
           sum=transf(xl+xhl/2.)
```

```
      do 12 i=1,nn
            sum=sum+transf(xl+xhl/2.+float(i)*xhl)
 12   continue
     res1=sum*xhl
            return
            end
c subrotina para calcular a cauda direita I2
      subroutine righti(func,b,c,res2)
            N=500
            xhr=abs(c-b)/float(n)
            sume=0.
            sumo=0.
            do 14 j=1,n-1,2
            sumo=sumo+func(b+float(j)*xhr)
  14   continue
       do 16 k=2,n-2,2
            sume=sume+func(b+float(k)*xhr)
  16   continue
       res2=(c-b)*(func(b)+4.*sumo+2.*sume+func(c))/(3*float(n))
            return
            end
c funcao densidade normal
      function func(x)
            func=exp(-x*x/2.)
            return
            end
c  funcao densidade transformada
      function transf(x)
            transf=exp(-1./(2.*x*x))/(x*x)
            return
            end
```

3) Pede-se:
(a) testar o seu programa (Exercício 2), calculando $\Phi(-3,5)$, $\Phi(-1)$, $\Phi(0)$, $\Phi(1)$ e $\Phi(3,5)$;
(b) se $X \sim \mathbf{N}(300,180)$, utilize o programa para calcular $P(220 \leq X \leq 390)$
(c) se $X \sim \mathbf{N}(300,180)$, utilize o programa para calcular $P(X<450|X>390)$
(d) refaça os itens (a), (b) e (c), com a aproximação dada pela equação 5.14.

4) Resolva o exercício 7 do capítulo 4, usando a aproximação da distribuição de Poisson pela distribuição Normal.

5) Resolva os itens (a) e (b) do Exemplo 5.4, aplicando a distribuição Normal. Faça um gráfico da função densidade correspondente. Calcule o quantil de tempo de retorno 100 anos.

6) No Exemplo 5.4, suponha que o coeficiente de assimetria seja igual a 1,5. Resolva os itens (a) e (b) do Exemplo 5.4, aplicando a distribuição Log-Normal de 3 parâmetros. Faça um gráfico da função densidade correspondente. Calcule o quantil de tempo de retorno 100 anos.

7) Resolva os itens (a) e (b) do Exemplo 5.4, aplicando a distribuição Exponencial. Faça um gráfico da função densidade correspondente. Calcule o quantil de tempo de retorno 100 anos.

8) Resolva os itens (a) e (b) do Exemplo 5.4, aplicando a distribuição Gama. Faça um gráfico da função densidade correspondente. Calcule o quantil de tempo de retorno 100 anos.

9) A direção do vento em certo local é uma variável aleatória X, medida a partir do Norte, cuja média e desvio padrão são, respectivamente, 200^o e 100^o. Discuta a conveniência do modelo Beta para X. Calcule os parâmetros da distribuição Beta e a probabilidade de X estar compreendida entre 90^o e 150^o. Faça um gráfico da função densidade correspondente.

10) Resolva o Exemplo 5.7 supondo que o tempo entre episódios de chuva seja uma variável Normal, com média de 4 dias e desvio padrão de 2 dias. Elabore um único gráfico com a densidade da variável original e a densidade do tempo máximo.

11) Resolva os itens (a) e (b) do Exemplo 5.4, aplicando a distribuição Gumbel para máximos. Faça um gráfico da função densidade correspondente. Calcule o quantil de tempo de retorno 100 anos.

12) As descargas máximas anuais em uma certa seção fluvial são descritas por uma distribuição de Gumbel com parâmetros de posição $\beta = 173$ m^3/s e escala $\alpha = 47$ m^3/s. Nessa seção fluvial, a cota de extravasamento para o leito maior corresponde à descarga $Q_t = 250$ m^3/s. Sabendo-se que houve extravasamento, calcule a probabilidade da excedência sobre a vazão Q_t ser menor ou igual a 100 m^3/s.

13) O Rio Alva em Ponte de Mucela, em Portugal, apresenta um número médio de 3 excedências por ano sobre a descarga de referência de 65 m^3/s. Testes estatísticos comprovaram serem plausíveis as hipóteses nulas do número

Poissoniano de excedências, independência serial e exponencialidade da cauda superior. Se a média das excedências é de 72,9 m^3/s, estime a descarga máxima anual de tempo de retorno 500 anos.

14) Resolva o Exemplo 5.8, aplicando a distribuição de Fréchet para máximos. Faça um gráfico da função densidade correspondente.

15) Considere novamente o exercício 13 e suponha, agora, que não existem evidências de cauda superior exponencial e também que o desvio padrão das excedências é de 75 m^3/s. Estime a descarga máxima anual de tempo de retorno 500 anos.

16) Faça um único gráfico com as funções acumuladas de probabilidades da distribuição GEV, para os conjuntos paramétricos mostrados na Figura 5.12. Discuta o uso dessa distribuição para a modelação de vazões máximas anuais, quando $\kappa > 0$ e $\kappa \leq 0$.

17) A média, a variância e o coeficiente de assimetria das vazões diárias mínimas anuais em uma certa seção fluvial são 694,6 m^3/s, 26186,62 $(m^3/s)^2$ e 1,1, respectivamente. Use o modelo Gumbel (mínimos) para estimar a vazão diária mínima de tempo de retorno 25 anos.

18) Resolva o exercício 17 para o modelo Weibull (mínimos) de 2 parâmetros.

19) Organize as equações 5.98 e 5.99 em forma de tabelas e defina, a partir delas, um esquema prático para o cálculo dos parâmetros para o modelo Weibull (mínimos) de 3 parâmetros.

20) Resolva o exercício 17 para o modelo Weibull (mínimos) de 3 parâmetros.

21) Resolva os itens (a) e (b) do Exemplo 5.4, aplicando a distribuição Pearson Tipo III. Faça um gráfico da função densidade correspondente. Calcule o quantil de tempo de retorno 100 anos.

22) Resolva os itens (a) e (b) do Exemplo 5.4, aplicando a distribuição Log-Pearson Tipo III. Faça um gráfico da função densidade correspondente. Calcule o quantil de tempo de retorno 100 anos.

23) Considere uma distribuição do χ^2 com $\nu = 4$. Calcule $P(\chi^2 > 5)$.

24) As concentrações diárias de oxigênio dissolvido em uma certa seção fluvial foram medidas durante 30 dias consecutivos. A amostra produziu uma média de 2,52 mg/l e um desvio-padrão de 2,05 mg/l. Admitindo-se que se trata de uma variável normalmente distribuída, determine o valor absoluto do máximo erro de estimativa da média populacional μ, com probabilidade de 95%. Em outros termos, determine d tal que $Pr\left(\left|\overline{X}-\mu\right| \le d\right) = 0{,}95.$

25) Considere uma distribuição de F com γ_1=10 e γ_2 = 5. Calcule P($F > 2$).

26) Considere a função densidade Normal bivariada, com parâmetros μ_X= 2, σ_X= 2, μ_Y= 1, σ_Y= 0,5 e ρ = 0,7. Expresse a função densidade condicional $f_{Y|X}(y|x=3)$.. Calcule a probabilidade P($Y < 3|X = 3$).

27) *O problema da agulha de Buffon*. Suponha que uma agulha é lançada aleatoriamente sobre um plano contendo linhas paralelas e separadas por uma distância fixa L, entendendo-se por agulha um segmento de reta de comprimento $l \le L$. O problema de Buffon é calcular a probabilidade de que a agulha intercepte uma das linhas paralelas. Para solucioná-lo, suponha que ξ_1 represente o ângulo entre a agulha e a direção das linha paralelas e que ξ_2 seja a distância entre a extremidade inferior da agulha e a linha mais próxima acima desse ponto (Figura 5.21a). As condições do experimento são tais que a variável aleatória ξ_1 é distribuída uniformemente no intervalo [0,π] e ξ_2 também uniformemente no intervalo[0,L]. Supondo que essas duas variáveis sejam independentes, a densidade conjunta de ambas é dada por $p(x_1, x_2) = \frac{1}{\pi L},\ 0 \le x_1 \le \pi,\ 0 \le x_2 \le L$. O evento A, correspondente ao fato da agulha interceptar a linha, ocorre se e somente se $\xi_2 \le l\text{sen}(\xi_1)$, ou seja se o ponto ($\xi_1$, ξ_2) se localizar na região B, a parte não hachurada da Figura 5.21b. Logo,

$$P\{(\xi_1, \xi_2) \in B\} = \iint_B \frac{dx_1 dx_2}{\pi L} = \frac{2l}{\pi L},\ onde\ l\int_0^{\pi} sen x_1 dx_1 = 2l$$ é a área de B. A premissa de independência entre as duas variáveis pode ser testada experimentalmente. De fato, se uma agulha é lançada n vezes e se A ocorre n_A vezes, então, $\frac{n_A}{n} \approx \frac{2l}{\pi L}$ para um valor elevado do número de lançamentos n. Nesse caso portanto, a quantidade $\frac{2l\ n}{L\ n_A}$ deve ser uma boa aproximação do número π=3,1415... . Você poderá simular o experimento de Buffon através do aplicativo de domínio público *Buffon*, disponível para *download* a partir da URL

http://www.efg2.com/Lab/Mathematics/Buffon.htm. Execute o programa para diversos valores crescentes de n, obtendo as respectivas aproximações de π e faça um gráfico dos seus resultados

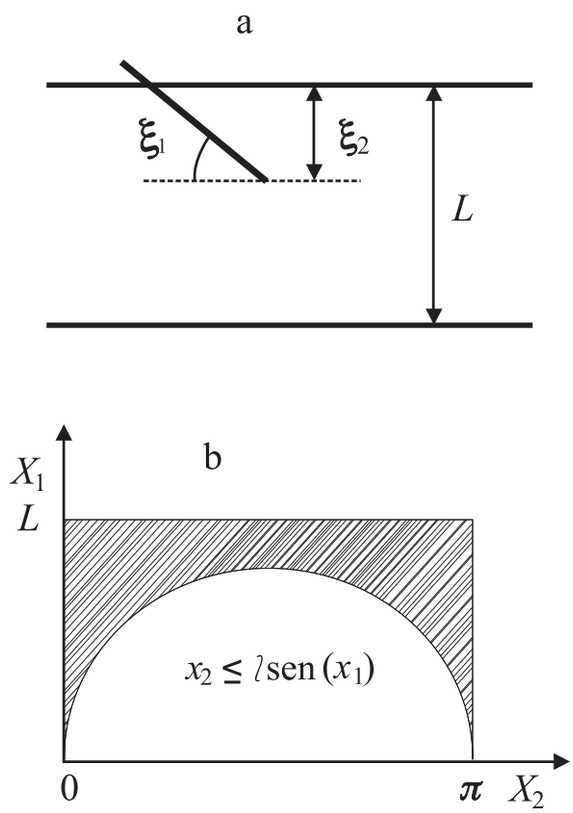
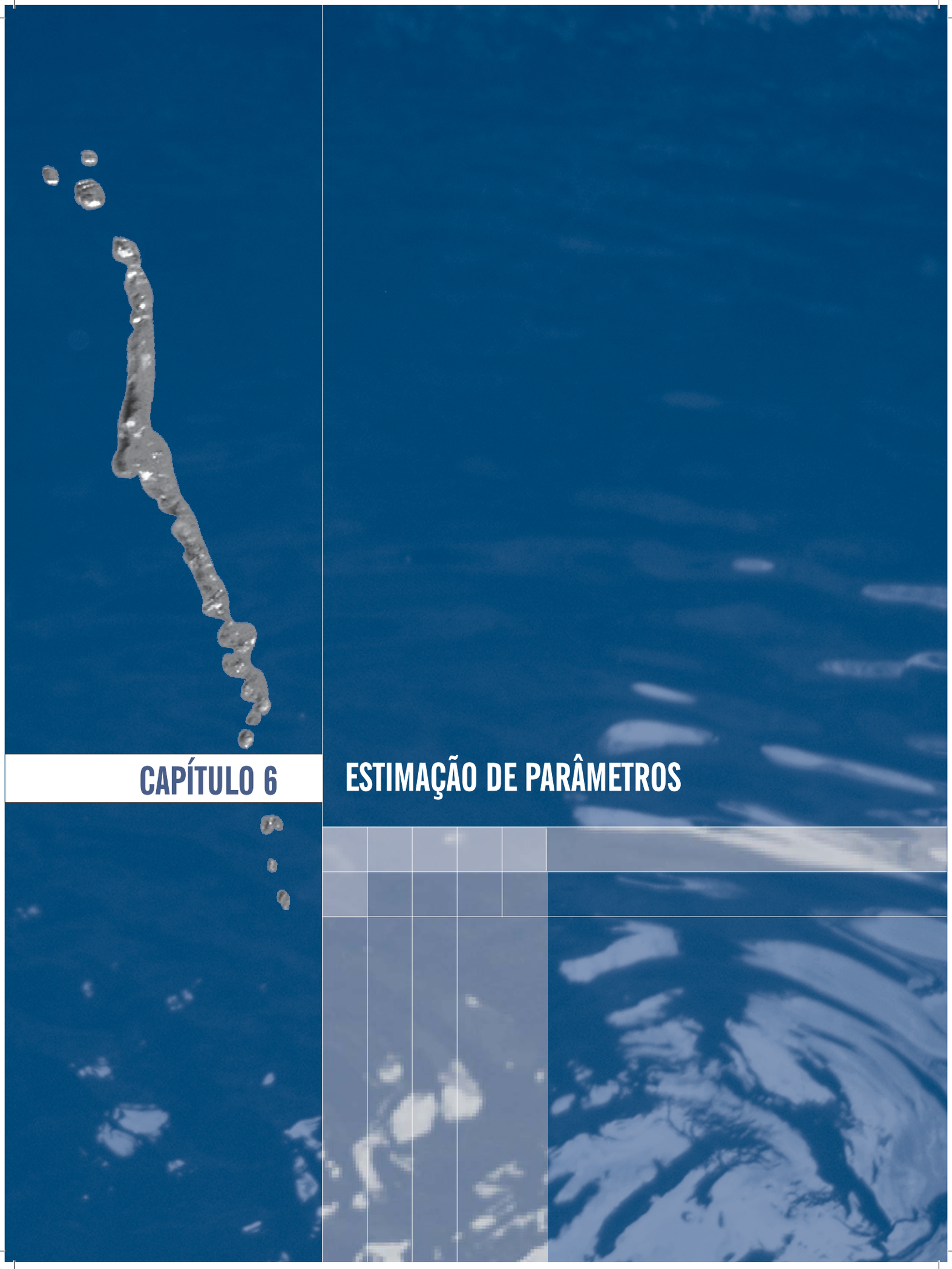


Figura 5.21 – Ilustração do problema da agulha de Buffon



CAPÍTULO 6

ESTIMAÇÃO DE PARÂMETROS

CAPÍTULO 6 ESTIMAÇÃO DE PARÂMETROS

Nos capítulos precedentes, foram estabelecidas as bases do cálculo de probabilidades para variáveis aleatórias discretas e contínuas. Uma vez conhecido (ou presumido) o modelo distributivo de uma variável aleatória e uma vez determinados os valores numéricos dos parâmetros que o definem, podemos calcular as probabilidades associadas a quaisquer eventos definidos pelos valores da variável em questão. Entretanto, conforme a discussão do item 1.4 do capítulo 1, o modelo distributivo e os verdadeiros valores numéricos de seus parâmetros seriam conhecidos apenas se toda a população tivesse sido amostrada, o que, na prática, pelo menos no tocante às variáveis hidrológicas, seria impossível. Assim, de posse apenas de uma amostra finita de observações de uma variável aleatória, devemos extrair conclusões (i) quanto ao modelo distributivo da população que contém a amostra e (ii) quanto às estimativas dos valores numéricos dos parâmetros que descrevem o modelo distributivo.

As técnicas de extração da informação probabilística e de obtenção das estimativas dos parâmetros a partir de uma amostra de observações, podem ser englobadas nos métodos da *inferência estatística*. Em termos gerais, esses são métodos que fazem a associação entre a realidade física de um conjunto de observações e a concepção abstrata de um modelo probabilístico prescrito para uma variável aleatória. De fato, a população é um termo conceitual porque consiste de um conjunto de elementos possivelmente observáveis, mas que não existem no sentido físico. Por outro lado, a amostra é constituída por um conjunto de N observações reais $\{x_1, x_2, ..., x_N\}$, que se supõem terem sido extraídas da população. As observações $\{x_1, x_2, ..., x_N\}$ representam os fatos concretos, a partir dos quais, são obtidas as estimativas de características populacionais, tais como valor esperado, variância e coeficiente de assimetria, assim como as inferências sobre a respectiva distribuição de probabilidades e seus parâmetros. A Figura 6.1 apresenta uma ilustração do raciocínio subjacente a esses métodos de inferência estatística. Nessa figura, a população, associada a um certo fenômeno hipotético, foi mapeada por uma variável aleatória contínua X, cuja função densidade de probabilidade foi prescrita como $f_X(x)$, definida por parâmetros $\theta_1, \theta_2, \ldots, \theta_k$; em alguns casos, a forma de $f_X(x)$ pode ser deduzida seja das características físicas do fenômeno em questão, seja do cotejo com as estatísticas amostrais. Entretanto, mesmo que $f_X(x)$ tenha sido corretamente prescrita, as estimativas $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_k$, dos parâmetros $\theta_1, \theta_2, \ldots, \theta_k$, devem ser necessariamente obtidas das observações amostrais.

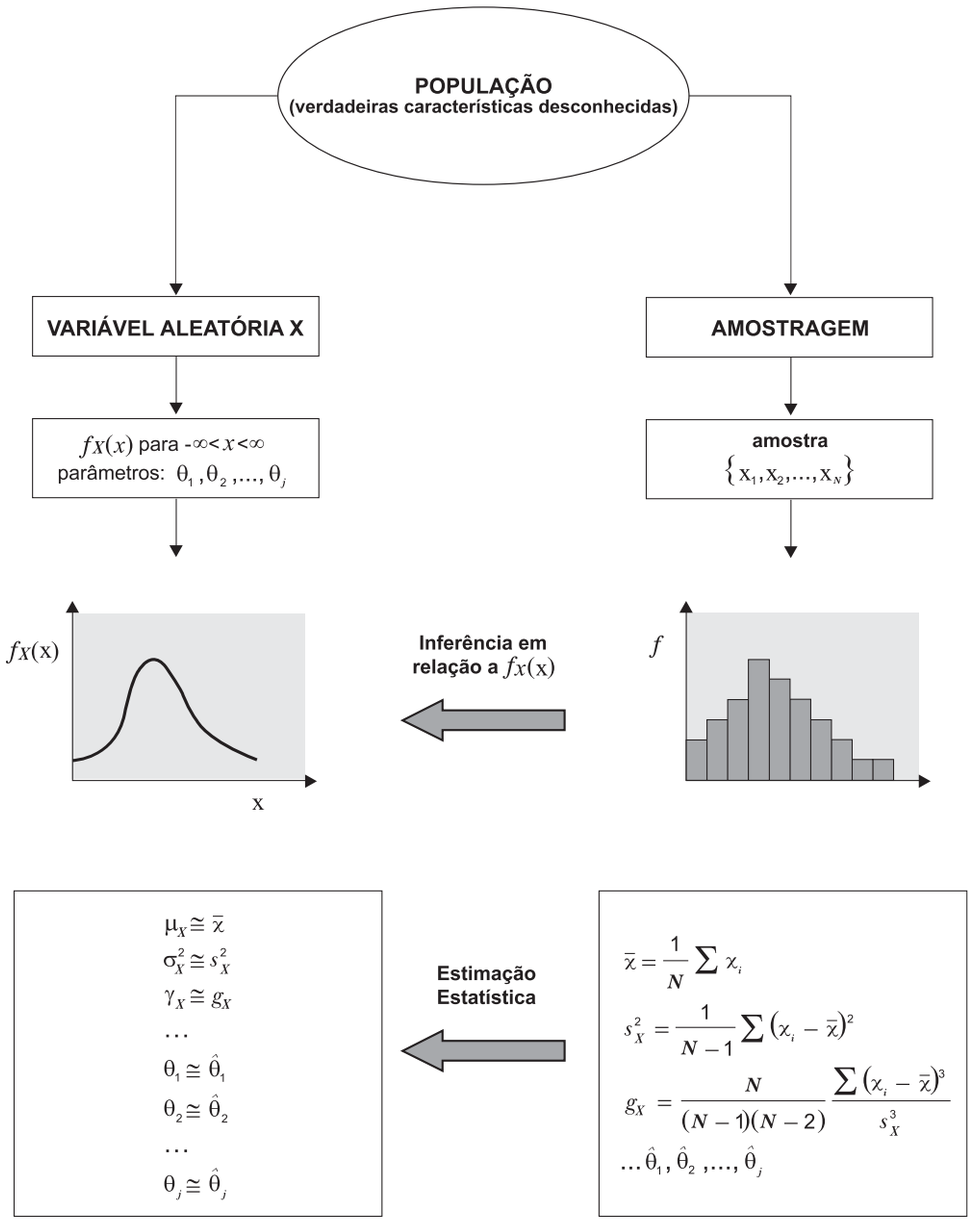


Figura 6.1 – Amostragem e inferência estatística

O problema, anteriormente descrito, é denominado *estimação de parâmetros*; o termo 'estimação' é aqui usado livremente, para significar o ato de produzir estimativas de parâmetros populacionais, a partir de uma amostra. Dentre os métodos clássicos da inferência estatística, existem dois caminhos possíveis para se obter estimativas de parâmetros: a *estimação pontual* e a *estimação por intervalos*. A estimação pontual refere-se à atribuição de um único valor numérico a um certo parâmetro populacional, a partir de estatísticas amostrais. A estimação por intervalos utiliza as informações contidas na amostra, para estabelecer uma afirmação quanto à probabilidade, ou grau de confiança, com que um certo intervalo de valores irá conter o verdadeiro valor do parâmetro populacional. Nos itens que se seguem, iremos estabelecer as bases para a estimação pontual e por intervalos, com maior ênfase, entretanto, sobre a primeira, por ser de uso mais freqüente para os propósitos da hidrologia estatística.

6.1 – Preliminares sobre a Estimação Pontual de Parâmetros

Como mencionado, o ponto de partida para a estimação de parâmetros é uma amostra de tamanho *N*, constituída pelos elementos $\{x_1, x_2, ..., x_N\}$. Esses representam as realizações das variáveis aleatórias $\{X_1, X_2, ..., X_N\}$. Para que a amostra seja considerada aleatória simples, ou simplesmente uma AAS, as variáveis $\{X_1, X_2, ..., X_N\}$ devem ser independentes e identicamente distribuídas, ou seja, variáveis IID. Em termos formais, se a densidade comum às variáveis $\{X_1, X_2, ..., X_N\}$ é $f_X(x)$, a função densidade conjunta da AAS é dada por $f_{X_1, X_2, ..., X_N}(x_1, x_2, ..., x_N) = f_X(x_1) f_X(x_2) ... f_X(x_N)$. Nessa expressão, uma vez especificada a distribuição $f_X(x)$, a qual é completamente definida por valores, ainda desconhecidos, dos parâmetros $\theta_1, \theta_2, ..., \theta_k$, toda a informação está contida na AAS $\{x_1, x_2, ..., x_N\}$.

Suponha, por facilidade, que há um único parâmetro θ a ser estimado a partir da AAS $\{x_1, x_2, ..., x_N\}$. Se toda a informação está ali contida, a estimativa de θ deve ser, necessariamente, uma função $g(x_1, x_2, ..., x_N)$ das observações. Como os elementos $\{x_1, x_2, ..., x_N\}$ são as realizações das variáveis aleatórias $\{X_1, X_2, ..., X_N\}$, podemos interpretar a função $g(x_1, x_2, ..., x_N)$ como uma realização da variável aleatória $g(X_1, X_2, ..., X_N)$. Se essa função é a utilizada para a estimação do parâmetro θ de $f_X(x)$, então, é forçosa a distinção entre o *estimador* de θ, representado por Θ, ou $\hat{\underline{\theta}}$, e a *estimativa* de θ, denotada por $\hat{\theta}$. De fato, a estimativa $\hat{\theta} = g(x_1, x_2, ..., x_N)$ é simplesmente um número, ou seja, uma *realização* do estimador $\Theta = \hat{\underline{\theta}} = g(X_1, X_2, ..., X_N)$. Esse, por sua vez, é uma variável aleatória, cujas propriedades podem ser estudadas pela teoria de probabilidades. Nesse contexto, é inapropriado levantar a questão se uma *estimativa* é melhor ou pior do que outra estimativa. Entretanto, é absolutamente legítimo e relevante perguntar como se

comparam, por exemplo, o estimador $\Theta_1 = \underline{\hat{\theta}}_1 = g_1(X_1, X_2, ..., X_N)$ com seu competidor $\Theta_2 = \underline{\hat{\theta}}_2 = g_2(X_1, X_2, ..., X_N)$. A resposta a essa questão está relacionada às propriedades dos estimadores.

Primeiramente, é indesejável que um procedimento de estimação, materializado por um certo estimador, produza estimativas que, em seu conjunto, sejam sistematicamente maiores ou menores do que o verdadeiro valor do parâmetro. Com efeito, o que se deseja é que a *média das estimativas* seja igual ao valor populacional do parâmetro. Formalmente, um estimador pontual $\underline{\hat{\theta}}$ é dito um *estimador sem viés* (ou não enviesado) do parâmetro populacional θ se

$$E[\underline{\hat{\theta}}] = \theta \tag{6.1}$$

Caso o estimador seja enviesado, o *viés*, ou erro sistemático, é dado pela diferença $E[\underline{\hat{\theta}}] - \theta$. Muitos estimadores são enviesados, mas possuem outras propriedades desejáveis.

Exemplo 6.1 – Demonstre que a média aritmética e a variância de uma amostra são estimadores não enviesados de μ e σ^2.
Solução: Considere uma amostra $\{x_1, x_2, ..., x_N\}$, de tamanho *N*. O estimador da média populacional é $\underline{\hat{\theta}} = \overline{X} = \frac{1}{N}(X_1 + X_2 + ... + X_N)$. Nesse caso, a equação 6.1 fornece $E[\overline{X}] = \frac{1}{N}\{E[X_1] + E[X_2] + ... + E[X_N]\}$, ou seja, $E[\overline{X}] = \frac{1}{N} N\mu = \mu$.

Para a variância, $\underline{\hat{\theta}} = S^2 = \frac{1}{(N-1)}\sum_{i=1}^{N}(X_i - \overline{X})^2$. A aplicação da equação 6.1, nesse caso, resulta em

$$E[S^2] = \frac{1}{N-1} E\left[\sum_{i=1}^{N}(X_i - \overline{X})^2\right] = \frac{1}{N-1} E\left[\sum_{i=1}^{N}(X_i - \mu)^2 - N(\overline{X} - \mu)^2\right].$$

Recordando que o valor esperado de uma soma é a soma dos valores esperados,

$$E[S^2] = \frac{1}{N-1}\left\{\sum_{i=1}^{N} E[(X_i - \mu)^2] - NE[(\overline{X} - \mu)^2]\right\}.$$ Nessa última equação, o primeiro valor esperado, entre chaves, é a variância de *X*, ou seja, σ^2, enquanto o segundo representa a variância de $\overline{X}$, igual a σ^2/N. Logo, $E[S^2] = \frac{1}{N-1}\left(N\sigma^2 - N\frac{\sigma^2}{N}\right) = \sigma^2$. Portanto, a média aritmética e a variância de uma amostra são, de fato, estimadores não enviesados de μ e σ^2

A segunda propriedade desejável dos estimadores é a consistência. Um estimador $\hat{\underline{\theta}}$ é considerado um *estimador consistente* de θ, se, para qualquer número positivo ε,

$$\lim_{N\to\infty} P\left[\left|\hat{\underline{\theta}} - \theta\right| \le \varepsilon\right] = 1 \qquad (6.2)$$

Em alguns casos, um estimador não enviesado pode não ser consistente. Essa situação é ilustrada pelo exemplo 6.2, a seguir.

Exemplo 6.2 – Considere os estimadores $\hat{\underline{\theta}}_1 = \frac{1}{N}\sum_{i=1}^{N}(X_i - \overline{X})^2$ e $\hat{\underline{\theta}}_2 = \frac{1}{(N-1)}\sum_{i=1}^{N}(X_i - \overline{X})^2$ da variância σ^2 de uma população. No exemplo 6.2 demonstrou-se que $\hat{\underline{\theta}}_2$ é um estimador sem viés de σ^2. Usando o mesmo raciocínio do exemplo 6.2, pode-se mostrar que $E[\hat{\underline{\theta}}_1] = \frac{N-1}{N}\sigma^2 \neq \sigma^2$ e que, portanto, $\hat{\underline{\theta}}_1$ é um estimador enviesado de σ^2. Entretanto, Kottegoda e Rosso (1997) afirmam que, apesar de enviesado, $\hat{\underline{\theta}}_1$ é um estimador consistente de σ^2, ao contrário de $\hat{\underline{\theta}}_2$. Pelo fato do atributo de inconsistência ter conseqüências menos severas do que o enviesamento, a prática usual é empregar $\hat{\underline{\theta}}_2$ como o estimador da variância populacional σ^2.

A terceira propriedade desejável dos estimadores é a *eficiência*. Um estimador *não enviesado* é considerado o mais eficiente entre todos os outros estimadores não-enviesados, se sua variância, denotada por $\mathrm{Var}\left[\hat{\underline{\theta}}\right]$, é menor ou igual à variância de qualquer outro estimador não-enviesado de θ.

Finalmente, a quarta propriedade desejável de um estimador é a suficiência. Um estimador $\hat{\underline{\theta}}$ é considerado um *estimador suficiente* de θ, se ele usa, ao máximo, toda a informação sobre θ, contida na amostra $\{x_1, x_2, \ldots x_N\}$, de modo que nenhuma outra informação pode ser adicionada por qualquer outro estimador. Essa e as propriedades de não-enviesamento, consistência e eficiência, são os fundamentos que guiam a seleção dos estimadores mais apropriados. Um tratamento rigoroso das propriedades dos estimadores pode ser encontrado em livros de estatística matemática, como, particularmente, os escritos por Cramér (1946) e Rao (1973).

Conforme menção anterior, uma vez escolhida a distribuição a ser ajustada aos dados amostrais, seus parâmetros devem ser estimados por algum procedimento da estatística

matemática, para que, em seguida, as estimativas paramétricas possam ser usadas para o cálculo de probabilidades e quantis. Há uma variedade de métodos de estimação de parâmetros, entre os quais destacam-se: (i) o método dos momentos; (ii) o método da máxima verossimilhança; (iii) o método dos momentos-L; (iv) o método da máxima entropia; (v) o método dos mínimos quadrados; (vi) o método generalizado dos momentos; e (vii) o método dos momentos mistos. Desses, consideraremos aqui os três primeiros, a saber: os métodos dos momentos (MOM), de máxima verossimilhança (MVS) e dos momentos-L (MML).

O método da máxima verossimilhança (MVS) é considerado o método de estimação mais eficiente porque produz os estimadores de menor variância. Entretanto, para alguns casos, a maior eficiência do método MVS é apenas assintótica, o que faz com que sua aplicação a amostras de pequeno tamanho produza estimadores de qualidade comparável ou inferior a outros métodos. Os estimadores de MVS são consistentes, suficientes e assintóticamente sem viés. Para amostras finitas, entretanto, os estimadores de MVS podem ser enviesados, embora o viés possa ser corrigido. O método MVS exige um maior esforço computacional, pelo fato de envolver soluções numéricas de sistemas de equações, freqüentemente, não lineares e implícitas.

O método dos momentos (MOM) é método de estimação mais simples. Entretanto, os estimadores desse método são, em geral, de qualidade inferior e menos eficientes do que os estimadores de MVS, particularmente para distribuições de três ou mais parâmetros. Cabe salientar, no entanto, que, para as pequenas amostras, freqüentes em hidrologia, os estimadores MOM podem ter atributos comparáveis ou até mesmo superiores aos de outros estimadores.

O método dos momentos-L (MML) produz estimadores de parâmetros comparáveis, em qualidade, àqueles produzidas pelo método da MVS, com a vantagem de exigirem um menor esforço computacional para a solução de sistemas de equações menos complexas. Para amostras pequenas, os estimadores MML são, com alguma freqüência, mais acurados do que os de MVS. Na seqüência, detalharemos os princípios de cada um dos três métodos, apresentando exemplos de suas respectivas aplicações.

6.2 – Método dos Momentos (MOM)

O método dos momentos consiste em igualar os momentos amostrais aos populacionais. O resultado dessa operação produzirá as estimativas dos parâmetros da distribuição de probabilidades em questão. Formalmente, sejam

$y_1, y_2, y_3, \ldots, y_N$ as observações constituintes de uma AAS retirada de uma população de uma variável aleatória distribuída conforme $f_Y\left(y;\theta_1,\theta_2,\ldots,\theta_k\right)$ de k parâmetros. Se μ_j e m_j representam, respectivamente, os momentos populacionais e amostrais, o sistema de equações fundamental do método dos momentos é

$$\mu_j\left(\hat{\theta}_1,\hat{\theta}_2,\ldots,\hat{\theta}_k\right) = m_j \text{ com } j=1,2,\ldots,k \tag{6.3}$$

As soluções $\hat{\theta}_1$, $\hat{\theta}_2$..., $\hat{\theta}_k$ desse sistema de k equações e k incógnitas serão as estimativas dos parâmetros θ_j pelo método dos momentos.

Exemplo 6.3 - Seja $Y_1, Y_2, Y_3, \ldots, Y_n$ uma AAS retirada da população de uma variável aleatória Y, cuja função densidade de probabilidade, a um único parâmetro θ, é $f_Y(y;\theta)=(\theta+1)y^{\theta}$ para $0 \le y \le 1$. Pede-se: (a) determinar o estimador de θ pelo método dos momentos; e (b) supondo que a AAS de Y seja constituída pelos seguintes elementos {0,2; 0,9; 0,05; 0,47; 0,56; 0,8; 0,35}, calcular a estimativa de θ pelo método dos momentos e a probabilidade de Y ser maior do que 0,8.

Solução: (a) Método dos momentos: $\mu_1 = m_1$. Momento populacional:

$\mu_1 = E(Y) = \int_0^1 y(\theta+1)y^{\theta}dy = \frac{\theta+1}{\theta+2}$. Momento Amostral: $m_1 = \frac{1}{n}\sum_{i=1}^{n} Y_i = \overline{Y}$

Logo, $\frac{\hat{\theta}+1}{\hat{\theta}+2} = \overline{Y} \Rightarrow \hat{\theta} = \frac{2\overline{Y}-1}{1-\overline{Y}}$. Esse é o estimador de θ pelo método dos momentos. (b) A AAS {0,2; 0,9; 0,05; 0,47; 0,56; 0,8; 0,35} produz $\overline{y} = 0{,}4757$. O estimador de θ, determinado no item (a), fornece a estimativa $\hat{\theta} = \frac{2\times 0{,}4757 - 1}{1 - 0{,}4757} = -0{,}0926$. A FAP é

$F_Y(y) = \int_0^y (\theta+1)y^{\theta}dy = y^{\theta} - 1$. Com $\hat{\theta} = -0{,}0926$,

$P(Y > 0{,}8) = 1 - F_Y(8) = 1 - 0{,}8248 = 0{,}1752$.

Exemplo 6.4 - Use o método dos momentos para ajustar uma distribuição Binomial com $n = 4$ aos dados abaixo. Calcule $P(X \ge 1)$. Lembre-se que $E(X) = n.p$, p = probabilidade de "sucesso" e n = nº de tentativas independentes de um processo de Bernoulli.

Valor de X (Nº de "sucessos")	0	1	2	3	4
Nº de observações para o valor dado de X	10	40	60	50	16

Solução: A distribuição Binomial é definida pelos parâmetros *n* e *p*. No caso presente, o parâmetro *n* foi especificado em 4, restando, portanto, estimar *p*. O método dos momentos impõe a condição $\mu_1 = m_1$, a qual, no caso presente, se particulariza para $n\hat{p} = \overline{X}$, ou seja, $\hat{p} = \overline{X}/4$ Esse é o estimador de *p*, pelo método dos momentos. A estimativa de *p* exige o cálculo da média aritmética $\overline{x}$, a qual, para a AAS em questão, é dada por $\overline{x} = (0\times10 + 1\times40 + 2\times60 + 3\times50 + 4\times16)/176 = 2{,}125$; portanto, $\hat{p} = 0{,}53125$. Finalmente,

$$P(X\geq 1)=1\text{-}P(X=0)= 1-\binom{4}{0}\times 0{,}53125^{0}\times(1-0{,}53125)^{4} = 0{,}9517 \quad .$$

Exemplo 6.5 – O Anexo 3 apresenta as alturas diárias máximas anuais, observadas na estação pluviométrica de Ponte Nova do Paraopeba, entre os anos hidrológicos de 1940/41 a 1999/2000, com algumas falhas no período. Para essa amostra, foram calculadas as seguintes estatísticas: $\overline{x} = 82{,}267\ mm, s_X = 22{,}759\ mm, s_X^2 = 517{,}988\ mm^2$ e $g = 0{,}7623$.
Pede-se: (a) determinar os estimadores MOM para os parâmetros da distribuição de Gumbel (máximos); (b) as estimativas MOM para os parâmetros da distribuição de Gumbel; (c) calcular a probabilidade da altura diária máxima anual superar 150 mm, em um ano qualquer; e (d) calcular a altura diária máxima anual de tempo de retorno igual a 100 anos.
Solução: (a) Suponha que $X \sim \mathbf{Gu}_{\mathbf{max}}(\alpha,\beta)$. Nesse caso, temos dois parâmetros a estimar e, portanto, são necessários os dois primeiros momentos: a média e a variância de *X*, quais sejam, $E[X] = \beta + 0{,}5772\alpha$

e $Var[X] = \sigma_X^2 = \dfrac{\pi^2\alpha^2}{6}$. Substituindo nessas equações os momentos populacionais pelos amostrais e resolvendo para α e β, temos, como resultado, os estimadores MOM da distribuição de Gumbel (máximos), a saber: $\hat{\alpha} = S_X / 1{,}283$ e $\hat{\beta} = \overline{X} - 0{,}45S_X$. (b) As estimativas MOM de α e β decorrem da substituição de $\overline{X}$ e S_X pelas correspondentes estatísticas amostrais $\overline{x} = 82{,}267\ e\ s_X = 22{,}759$ Resultados: $\hat{\alpha} = 17{,}739$ e $\hat{\beta} = 72{,}025$. (c) A probabilidade pedida é

$$1 - F_X(150) = 1 - exp\left[-exp\left(-\frac{150-\hat{\beta}}{\hat{\alpha}}\right)\right] = 0{,}0123 \quad .$$ (d) A equação das estimativas de quantis para *T*=100 anos é

$$\hat{x}(T=100) = \hat{\beta} - \hat{\alpha}\, ln\left[-ln\left(1-\frac{1}{100}\right)\right] = 153{,}63 \text{ mm.}$$

Exemplo 6.6 – Repita o exemplo 6.5 para a distribuição GEV.

Solução: (a) Suponha que $X \sim$ **GEV**(α, β, κ). Nesse caso, temos três parâmetros a estimar e, portanto, são necessários os três primeiros momentos: a média, a variância e o coeficiente de assimetria de X, dados respectivamente pelas equações 5.71, 5.72 e 5.73 do capítulo 5. Conforme mencionado no capítulo 5, o cálculo dos parâmetros da distribuição GEV deve começar pela equação 5.73, a qual deve ser resolvida para κ, por meio de iteração numérica ou com o auxílio do gráfico da Figura 5.13, a partir do valor do coeficiente de assimetria. Uma forma alternativa para o cálculo de κ é o uso de equações de regressão de κ x γ, tais como as seguintes, sugeridas por Rao e Hamed (2000):

para $1{,}1396 < \gamma < 10$ (Extremos Tipo 2 ou Fréchet):

$$\kappa = 0{,}2858221 - 0{,}357983\gamma + 0{,}116659\gamma^2 - 0{,}022725\gamma^3 + 0{,}002604\gamma^4 - 0{,}000161\gamma^5 + 0{,}000004\gamma^6,$$

para $-2 < \gamma < 1{,}1396$ (Extremos Tipo 3 ou Weibull):

$$\kappa = 0{,}277648 - 0{,}322016\gamma + 0{,}060278\gamma^2 + 0{,}016759\gamma^3 - 0{,}005873\gamma^4 - 0{,}00244\gamma^5 - 0{,}00005\gamma^6 \quad ,$$

e para $-10 < \gamma < 0$ (Extremos Tipo 3 ou Weibull):

$$\kappa = -0{,}50405 - 0{,}00861\gamma + 0{,}015497\gamma^2 + 0{,}005613\gamma^3 + 0{,}00087\gamma^4 + 0{,}000065\gamma^5$$

No caso presente, com $\hat{\gamma} = 0{,}7623$, a segunda equação é a indicada e compõe a primeira peça $\underline{\hat{\kappa}}$ da resolução dos estimadores MOM da distribuição GEV. Em seguida, conforme a seqüência apresentada no capítulo 5, temos os seguintes estimadores MOM:

$$\underline{\hat{\alpha}} = \sqrt{\frac{\underline{\hat{\kappa}}^2 S_X^2}{\Gamma(1+2\underline{\hat{\kappa}}) - \Gamma^2(1+\underline{\hat{\kappa}})}} \quad \text{e} \quad \underline{\hat{\beta}} = \overline{X} - \frac{\underline{\hat{\alpha}}}{\underline{\hat{\kappa}}}\left[1 - \Gamma(1+\underline{\hat{\kappa}})\right].$$

(b) As estimativas MOM de α, β e κ decorrem da substituição de $\overline{X}$, S_X e $\hat{\gamma}$ pelas correspondentes estatísticas amostrais $\overline{x} = 82{,}267$, $s_X = 22{,}759$ e $g = 0{,}7623$, na seqüência acima descrita. Resultados: $\hat{\kappa} = 0{,}072$, $\hat{\alpha} = 19{,}323$ e $\hat{\beta} = 72{,}405$.

(c) $$1 - F_X(150) = 1 - exp\left\{-\left[1 - \hat{\kappa}\left(\frac{150-\hat{\beta}}{\hat{\alpha}}\right)\right]^{1/\hat{\kappa}}\right\} = 0{,}0087\,.$$

(d) A equação das estimativas de quantis para T=100 anos é

$$x(T) = \hat{\beta} + \frac{\hat{\alpha}}{\hat{\kappa}}\left\{1 - \left[-ln\left(1 - \frac{1}{T}\right)\right]^{\hat{\kappa}}\right\} = 148{,}07 \ \text{mm}.$$

6.3 – Método da Máxima Verossimilhança (MVS)

O método da máxima verossimilhança consiste basicamente em maximizar uma função dos parâmetros da distribuição, conhecida como *função de verossimilhança*. O equacionamento para a condição de máximo resulta em um sistema de igual número de equações e incógnitas, cujas soluções produzem os estimadores de máxima verossimilhança.

Considere que $y_1, y_2, y_3, \ldots, y_N$ representem as observações constituintes de uma AAS retirada de uma população de uma variável aleatória distribuída conforme a densidade $f_Y(y;\theta_1,\theta_2,\ldots,\theta_k)$ de k parâmetros. A função densidade conjunta da AAS, constituída por Y_1, Y_2, Y_3, ... , Y_N, é dada por $f_{Y_1,Y_2,\ldots,Y_N}(y_1,y_2,\ldots,y_N) = f_Y(y_1) f_Y(y_2) \ldots f_Y(y_N)$. Essa densidade conjunta é proporcional à probabilidade de que a AAS tenha sido extraída da população, definida por $f_Y(y;\theta_1,\theta_2,\ldots,\theta_k)$, sendo conhecida por função de verossimilhança. Portanto, em termos formais, a função de verossimilhança é dada por

$$L(\theta_1,\theta_2,\ldots,\theta_k) = \prod_{i=1}^{N} f_Y(y_i;\theta_1,\theta_2,\ldots,\theta_k) \tag{6.4}$$

Essa é uma função dos parâmetros θ_j, exclusivamente. Os valores θ_j que maximizam essa função são aqueles que também maximizam a probabilidade de que aquela AAS específica, constituída por Y_1, Y_2, Y_3, ... , Y_N, tenha sido sorteada da população, tal como definida pela densidade prescrita. A busca da condição de máximo para a função de verossimilhança resulta no seguinte sistema de k equações e k incógnitas:

$$\frac{\partial L(\theta_1,\theta_2,\ldots,\theta_k)}{\partial \theta_j} = 0;\ j = 1,2,\ldots,k \tag{6.5}$$

As soluções desse sistema de equações são os estimadores $\hat{\theta}_j$ de máxima verossimilhança. É freqüente o emprego da função logaritmo de verossimilhança ln [L (θ)], em substituição à função de verossimilhança propriamente dita, para facilitar a construção do sistema de equações 6.5. Isso se justifica pelo fato da função logaritmo ser contínua, monótona e crescente, e, portanto, maximizar o logaritmo da função é o mesmo que maximizar a função.

Exemplo 6.7 - Seja $y_1, y_2, y_3, \ldots, y_N$ uma AAS retirada da população de uma variável aleatória discreta Y, distribuída segundo uma distribuição de Poisson, com parâmetro λ. Determine o estimador de λ pelo método da máxima verossimilhança.

Solução: A função massa de Poisson é $p_Y(y;\lambda)=\frac{\lambda^y \exp(-\lambda)}{y!}; y=0,1,2,..$ e a respectiva função de verossimilhança é

$$L(\lambda;Y_1,Y_2,..,Y_N)=\prod_{i=1}^{N}\frac{\lambda^{Y_i}exp(-\lambda)}{Y_i!}=\frac{\lambda^{\sum_{i=1}^{N}Y_i}exp(-N\lambda)}{\prod_{i=1}^{N}Y_i!}$$

A pesquisa do valor de λ que maximiza essa função pode ser grandemente facilitada por sua substituição pela função log de verossimilhança, ou seja, por

$$ln[L(\lambda;Y_1,Y_2,..,Y_N)]=-N\lambda+ln(\lambda)\sum_{i=1}^{N}Y_i-ln\left(\prod_{i=1}^{N}Y_i!\right)$$

Tomando a derivada dessa função em relação a λ, resulta em $\frac{d\,ln\left[L(\lambda;Y_1,Y_2,..,Y_N)\right]}{d\lambda}=-N+\frac{1}{\lambda}\sum_{i=1}^{N}Y_i$. Igualando essa derivada a zero, resulta o estimador de MVS de λ, ou seja, $\underline{\hat{\lambda}}=\frac{1}{N}\sum_{i=1}^{n}Y_i \; ou \; \underline{\hat{\lambda}}=\overline{Y}$.

Exemplo 6.8 – Repita o exemplo 6.5, usando o método da máxima verossimilhança.

Solução: (a) A função de verossimilhança de uma amostra de tamanho *N*, extraída de uma população Gumbel (máximos), é $L(\alpha,\beta)=\frac{1}{\alpha^N}exp\left[-\sum_{i=1}^{N}\left(\frac{Y_i-\beta}{\alpha}\right)-\sum_{i=1}^{N}exp\left(-\frac{Y_i-\beta}{\alpha}\right)\right]$. Analogamente ao exemplo anterior, a função log de verossimilhança é $ln[L(\alpha,\beta)]=-N\,ln(\alpha)-\frac{1}{\alpha}\sum_{i=1}^{N}(Y_i-\beta)-\sum_{i=1}^{N}exp\left(-\frac{Y_i-\beta}{\alpha}\right)$. Derivando essa função em relação a α e β, e igualando ambas derivadas a zero, resulta o seguinte sistema de equações:

$$\frac{\partial}{\partial\alpha}ln[L(\alpha,\beta)]=-\frac{N}{\alpha}+\frac{1}{\alpha^2}\sum_{i=1}^{N}(Y_i-\beta)-\frac{1}{\alpha^2}\sum_{i=1}^{N}(Y_i-\beta)exp\left(-\frac{Y_i-\beta}{\alpha}\right)=0 \qquad (I)$$

$$\frac{\partial}{\partial\beta}ln[L(\alpha,\beta)]=\frac{N}{\alpha}-\frac{1}{\alpha}\sum_{i=1}^{N}exp\left(-\frac{Y_i-\beta}{\alpha}\right)=0 \qquad (II)$$

Rao e Hamed (2000) sugerem o procedimento, descrito a seguir, para a solução do sistema de equações acima. Primeiramente, deduzindo da

equação (II) que $exp\left(\frac{\beta}{\alpha}\right)=\frac{N}{\sum_{i=1}^{N} exp\left(-Y_i/\alpha\right)}$, substituindo na equação (I) e simplificando, resulta a seguinte equação:

$$F(\alpha)=\sum_{i=1}^{N} Y_i\, exp\left(-\frac{Y_i}{\alpha}\right)-\left(\frac{1}{N}\sum_{i=1}^{N} Y_i-\alpha\right)\sum_{i=1}^{N} exp\left(-\frac{Y_i}{\alpha}\right)=0$$

. Essa equação, embora função apenas de α, não tem solução analítica. Para resolvê-la, recorre-se ao método iterativo de Newton, no qual, dado um valor inicial para α, o valor da iteração seguinte é atualizado pela expressão $\alpha_{j+1}=\alpha_j-F(\alpha_j)/F'(\alpha_j)$. Nessa equação, F' representa a derivada de F, em relação a α, ou seja,

$$F'(\alpha)=\frac{1}{\alpha^2}\sum_{i=1}^{N} Y_i^2\, exp\left(-\frac{Y_i}{\alpha}\right)+\sum_{i=1}^{N} exp\left(-\frac{Y_i}{\alpha}\right)+\frac{1}{\alpha}\sum_{i=1}^{N} Y_i\, exp\left(-\frac{Y_i}{\alpha}\right).$$

As iterações terminam quando $F(\alpha)$ está suficientemente próximo de zero, obtendo-se assim o estimador $\underline{\hat{\alpha}}$. Em seguida, o estimador $\underline{\hat{\beta}}$ é obtido a partir da equação

$$\underline{\hat{\beta}}=\underline{\hat{\alpha}}\; ln\left[\frac{N}{\sum_{i=1}^{N} exp\left(-Y_i/\alpha\right)}\right]$$

. Esses são os estimadores de MVS da distribuição Gumbel (máximos). (b) As estimativas de MVS de α e β decorrem da substituição das somatórias envolvidas no cálculo dos estimadores pelos seus respectivos valores amostrais. O *software* ALEA, desenvolvido pelo Departamento de Engenharia Hidráulica e Recursos Hídricos da Escola de Engenharia da UFMG, possui uma rotina que implementa o procedimento de Rao e Hamed (2000) para uma dada amostra, assim como outras rotinas para o cálculo de estimativas MOM e de MVS para diversas distribuições de probabilidades. O programa executável e um manual do usuário de *software* ALEA podem ser *downloaded* a partir da URL http://www.ehr.ufmg.br. As estimativas de MVS, calculadas pelo *software* ALEA, para a amostra de alturas diárias máximas anuais de Ponte Nova do Paraopeba são $\hat{\alpha}=19{,}4$ e $\hat{\beta}=71{,}7$.

c) A probabilidade pedida é

$$1-F_X(150)=1-exp\left[-exp\left(-\frac{150-\hat{\beta}}{\hat{\alpha}}\right)\right]=0{,}0175\,.$$

(d) A equação das estimativas de quantis para T=100 anos é

$$\hat{x}(T=100)=\hat{\beta}-\hat{\alpha}\,ln\left[-ln\left(1-\frac{1}{100}\right)\right]=160{,}94\,\text{mm}.$$

6.4 – Método dos Momentos-L (MML)

Greenwood et al. (1979) introduziram os momentos ponderados por probabilidades (MPP), os quais são definidos pela seguinte expressão geral:

$$M_{p,r,s}=E\left[X^{p}\left[F_{X}(x)\right]^{r}\left[1-F_{X}(x)\right]^{s}\right]=\int_{0}^{1}\left[x(F)\right]^{p}F^{r}\left(1-F\right)^{s}dF \tag{6.6}$$

onde $x(F)$ denota a função de quantis, e p, r e s representam números reais. Quando r e s são nulos e p é um número não negativo, os MPP's $M_{p,0,0}$ são iguais aos momentos convencionais μ'_{p} de ordem p, em relação à origem. Em particular, os MPP's $M_{1,0,s}$ e $M_{1,r,0}$ são os de utilidade mais freqüente na caracterização de distribuições de probabilidades e especificados por

$$M_{1,0,s}=\alpha_{s}=\int_{0}^{1}x(F)\left(1-F\right)^{s}dF \tag{6.7}$$

$$M_{1,r,0}=\beta_{r}=\int_{0}^{1}x(F)F^{r}dF \tag{6.8}$$

Hosking (1986) demonstrou que α_s e β_r, como funções lineares de x, possuem a generalidade suficiente para a estimação de parâmetros de distribuições de probabilidades, além de estarem menos sujeitos a flutuações amostrais e, portanto, serem mais *robustos* do que os correspondentes momentos convencionais. Para uma amostra $x_1 \le x_2 \le ... \le x_N$, ordenada de modo crescente, as estimativas não-enviesadas de α_s e β_r podem ser calculadas pelas seguintes expressões:

$$a_{s}=\hat{\alpha}_{s}=\frac{1}{N}\sum_{i=1}^{N}\frac{\binom{N-i}{s}}{\binom{N-1}{s}}x_{i} \tag{6.9}$$

$$b_{r}=\hat{\beta}_{r}=\frac{1}{N}\sum_{i=1}^{N}\frac{\binom{i-i}{r}}{\binom{N-1}{r}}x_{i} \tag{6.10}$$

Os MPP's α_s e β_r, assim como suas correspondentes estimativas amostrais a_s e b_r, estão relacionados entre si pelas expressões

$$\alpha_s = \sum_{i=1}^{s} \binom{s}{i} (-1)^i \beta_i \text{ ou } \beta_r = \sum_{i=1}^{r} \binom{r}{i} (-1)^i \alpha_i \qquad (6.11)$$

Exemplo 6.9 – Dadas as descargas médias anuais (m³/s), observadas no Rio Paraopeba em Ponte Nova do Paraopeba, listadas na Tabela 6.1 para os anos civis de 1990 a 1999, calcule as estimativas de α_s e β_r, ($r,s \leq 3$).

Tabela 6.1 – Vazões Médias Anuais (m³/s) do Rio Paraopeba em Ponte Nova do Paraopeba

1	2	3	4	5	6	7	8
Ano Civil	Vazão Q Anual (m³s)	Ordem i	Q_i ordenadas	$\binom{N-i}{0} Q_i$	$\binom{N-i}{1} Q_i$	$\binom{N-i}{2} Q_i$	$\binom{N-i}{3} Q_i$
1990	53,1	1	53,1	53,1	477,9	1911,6	4460,4
1991	112,1	2	57,3	57,3	458,4	1604,4	3208,8
1992	110,8	3	63,6	63,6	445,2	1335,6	2226
1993	82,2	4	80,9	80,9	485,4	1213,5	1618
1994	88,1	5	82,2	82,2	411	822	822
1995	80,9	6	88,1	88,1	352,4	528,6	352,4
1996	89,8	7	89,8	89,8	269,4	269,4	89,8
1997	114,9	8	110,8	110,8	221,6	110,8	-
1998	63,6	9	112,2	112,2	112,2	-	-
1999	57,3	10	114,9	114,9	-	-	-

Solução: A Tabela 6.1 apresenta alguns cálculos necessários á aplicação da equação 6.9, para s = 0,1,2 e 3. O valor de a_0 é obtido pela divisão da soma dos 10 itens da coluna 5 por $N\binom{N-1}{0} = 10$, o que resulta em a_0= 85,29; observe que a_0 é, de fato, *equivalente à média aritmética* da amostra. Cálculos semelhantes com as colunas 6 a 8, conduzem aos resultados a_1 = 35,923, a_2 = 21,655 e a_3 = 15,211. Os valores de b_r podem ser calculados pela equação 6.10 ou deduzidos de a_s, a partir da expressão 6.11. Nesse último caso, para $r,s \leq 3$, é fácil verificar que

$\alpha_0 = \beta_0$ ou $\beta_0 = \alpha_0$

$\alpha_1 = \beta_0 - \beta_1$ ou $\beta_1 = \alpha_0 - \alpha_1$

$\alpha_2 = \beta_0 - 2\beta_1 + \beta_2$ ou $\beta_2 = \alpha_0 - 2\alpha_1 + \alpha_2$

$\alpha_3 = \beta_0 - 3\beta_1 + 3\beta_2 - \beta_3$ ou $\beta_3 = \alpha_0 - 3\alpha_1 + 3\alpha_2 - \alpha_3$

Nas equações acima, substituindo os MPP's pelas suas estimativas e com os valores anteriormente calculados, obtém-se b_0= 85,29, b_1= 49,362, b_2= 35,090 e b_3= 27,261.

Os MPP's α_s e β_r, embora passíveis de serem usados na estimação de parâmetros, não são de fácil interpretação como descritores de forma das distribuições de probabilidades. Tendo em vista tal fato, Hosking (1990) introduziu o conceito de *momentos-L*, os quais são grandezas diretamente interpretáveis como descritores de escala e forma das distribuições de probabilidades. Os momentos-L de ordem r, denotados por λ_r, são combinações lineares dos MPP's α_s e β_r e formalmente definidos por

$$\lambda_r = (-1)^{r-1} \sum_{k=0}^{r-1} p_{r-1,k}\, \alpha_k = \sum_{k=0}^{r-1} p_{r-1,k}\, \beta_k \qquad (6.12)$$

onde $p_{r-1,k} = (-1)^{r-k-1} \binom{r-1}{k} \binom{r+k-1}{k}$. A aplicação da equação 6.12 para os momentos-L, de ordem inferior a 4, resulta em

$$\lambda_1 = \alpha_0 = \beta_0 \qquad (6.13)$$

$$\lambda_2 = \alpha_0 - 2\alpha_1 = 2\beta_1 - \beta_0 \qquad (6.14)$$

$$\lambda_3 = \alpha_0 - 6\alpha_1 + 6\alpha_2 = 6\beta_2 - 6\beta_1 + \beta_0 \qquad (6.15)$$

$$\lambda_3 = \alpha_0 - 12\alpha_1 + 30\alpha_2 - 20\alpha_3 = 20\beta_3 - 30\beta_2 + 12\beta_1 - \beta_0 \qquad (6.16)$$

Os momentos-L amostrais são denotados por l_r e são calculados pela substituição de α_s e β_r, nas equações 6.13 a 6.16, pelas suas estimativas a_s e b_r.

O momento-L λ_1 é equivalente à média e, portanto, uma medida populacional de posição. Para ordens superiores a 1, os *quocientes de momentos-L* são particularmente úteis na descrição da escala e forma das distribuições de probabilidades. Como medida equivalente ao coeficiente de variação convencional, define-se o coeficiente τ, dado por

$$\tau = \frac{\lambda_2}{\lambda_1} \qquad (6.17)$$

o qual pode ser interpretado como uma medida populacional de dispersão ou de escala. Analogamente aos coeficientes de assimetria e curtose convencionais, podem ser definidos os coeficientes τ_3 e τ_4, dados, respectivamente, por

$$\tau_3 = \frac{\lambda_3}{\lambda_2} \tag{6.18}$$

$$\tau_4 = \frac{\lambda_4}{\lambda_2} \tag{6.19}$$

Os quocientes de momentos-L amostrais, cujas notações são t, t_3 e t_4, são calculados pela substituição de λ_r, nas equações 6.17 a 6.19, por suas estimativas l_r. Em relação aos momentos convencionais, os momentos-L apresentam diversas vantagens, entre as quais destacam-se os limites de variação de τ, τ_3 e τ_4. De fato, se X é uma variável aleatória não negativa, demonstra-se que $0 < \tau < 1$. Quanto a τ_3 e τ_4, é um fato matemático que esses coeficientes estão compreendidos no intervalo [-1,+1], em oposição aos seus correspondentes convencionais que podem assumir valores arbitrariamente mais elevados. Outras vantagens dos momentos-L, em relação aos momentos convencionais, são discutidas por Vogel e Fennessey (1993).

O método dos momentos-L (MML), para a estimação de parâmetros de distribuições de probabilidades é semelhante ao método dos momentos convencionais. De fato, tal como exemplifica a Tabela 6.2, os momentos-L e seus quocientes, a saber λ_1, λ_2, τ, τ_3 e τ_4 podem ser postos como funções dos parâmetros das distribuições de probabilidades e vice-versa. O método MML de estimação de parâmetros consiste em igualar os momentos-L populacionais aos momentos-L amostrais. O resultado dessa operação produzirá as estimativas dos parâmetros da distribuição de probabilidades em questão. Formalmente, sejam $y_1, y_2, y_3, \ldots, y_N$ as observações constituintes de uma AAS retirada de uma população de uma variável aleatória distribuída conforme $f_Y(y;\theta_1,\theta_2,\ldots,\theta_k)$ de k parâmetros. Se $[\lambda_1, \lambda_2, \tau_j]$ e $[l_1, l_2, t_j]$ representam, respectivamente, os momentos-L (e seus quocientes) populacionais e amostrais, o sistema de equações fundamental do método dos momentos-L é

$$\begin{aligned} &\lambda_i(\theta_1,\theta_2,\ldots,\theta_k) = l_i \text{ com } i = 1,2 \\ &\tau_j(\theta_1,\theta_2,\ldots,\theta_k) = t_j \text{ com } j = 3,4,\ldots,k-2 \end{aligned} \tag{6.20}$$

As soluções $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_k$ desse sistema de k equações e k incógnitas serão as estimativas dos parâmetros θ_j pelo método MML.

Tabela 6.2 – Momentos-L e seus quocientes para algumas distribuições de probabilidades (adap. de Stedinger et al., 1993)

Distribuição	Parâmetros	λ_1	λ_2	τ_3	τ_4
Uniforme	a, b	$\frac{a+b}{2}$	$\frac{b-a}{6}$	0	0
Exponencial	θ	θ	$\frac{\theta}{2}$	$\frac{1}{3}$	$\frac{1}{6}$
Normal	μ,σ	μ	$\frac{\sigma}{\sqrt{\pi}}$	0	0,1226
Gumbel	α,β	$\beta + 0{,}5772\alpha$	$\alpha\, ln(2)$	0,1699	0,1504

Exemplo 6.10 – Encontre as estimativas MML dos parâmetros da distribuição de Gumbel para os dados do exemplo 6.9.
Solução: Os resultados do exemplo 6.9 mostram que as estimativas MPP de β_r são b_0= 85,29, b_1= 49,362, b_2=35,090 e b_3= 27,261. Temos dois parâmetros a estimar e, portanto, precisamos apenas dos dois primeiros momentos-L, a saber λ_1 e λ_2. As estimativas desses podem ser obtidas pelas equações 6.13 e 6.14: $l_1 = b_0 = 85{,}29$ e $l_2 = 2b_1 - b_0 = 2 \times 49{,}362 - 85{,}29 = 13{,}434$. Com a relação entre λ_2 e α da distribuição de Gumbel (Tabela 6.2), segue-se que $\hat{\alpha} = l_2 / ln(2) = 19{,}381$. Em seguida, tem-se que $\hat{\beta} = l_1 - 0{,}5772\hat{\alpha} = 74{,}103$.

6.5 – Estimação por Intervalos

Uma estimativa pontual de um parâmetro de uma distribuição de probabilidades, tal como apresentado nos itens anteriores, é um número que se encontra na vizinhança do verdadeiro e desconhecido valor populacional do parâmetro. A questão do erro presente na estimação pontual de parâmetros, devido à variabilidade inerente às amostras aleatórias que lhe deram origem, nos remete à construção dos chamados *intervalos de confiança*. De fato, um estimador pontual de um parâmetro θ é uma estatística $\underline{\hat{\theta}}$, a qual, por ser uma função de uma variável aleatória *X*, é também uma variável aleatória e possui, ela mesma, uma densidade de probabilidades $f_{\underline{\hat{\theta}}}(\hat{\theta})$. É bem verdade que, se $\underline{\hat{\theta}}$ é uma variável aleatória contínua, então $P(\underline{\hat{\theta}} = \theta) = 0$, o que tornaria inócuo um tal equacionamento, na forma de igualdade. Entretanto, se construirmos as variáveis aleatórias *I*, correspondente a limite *inferior*, e *S*, correspondente a limite *superior*, ambas em função da variável $\underline{\hat{\theta}}$, é possível estabelecer a seguinte afirmação probabilística:

$$P(I \le \theta \le S) = 1 - \alpha \qquad (6.21)$$

na qual θ denota o valor populacional do parâmetro e (1- α) representa o *nível de confiança*. Como θ é um parâmetro e não uma variável aleatória, deve-se ter cuidado com a interpretação da equação 6.21. Seria incorreto interpretá-la como se fosse de (1-α) a probabilidade do parâmetro θ estar contido entre os limites do intervalo. Precisamente porque θ não é uma variável aleatória, a equação 6.21 deve ser corretamente interpretada da seguinte forma: a probabilidade do intervalo [*I*, *S*] conter o verdadeiro valor populacional do parâmetro θ é igual a (1- α).

Para melhor clarear a afirmação dada pela equação 6.21, considere que queiramos estimar a média μ de uma população qualquer, cujo desvio padrão populacional é conhecido e igual a σ, e que, para tal, usaremos a média aritmética $\overline{X}$ de uma amostra de tamanho *N*, suficientemente grande. Da solução do exemplo 5.3 e, portanto, do teorema do limite central, sabe-se que a variável $\left(\frac{\overline{X}-\mu}{\sigma/\sqrt{N}}\right) \sim N(0,1)$. Logo, pode-se escrever, para o exemplo em questão, que $P\left(-1{,}96 < \frac{\overline{X}-\mu}{\sigma/\sqrt{N}} < +1{,}96\right) = 0{,}95$. Para transformar essa expressão em uma afirmação semelhante àquela dada pela equação 6.21, é necessário isolar o parâmetro μ no centro da desigualdade, entre parênteses, ou seja, $P\left(\overline{X} - 1{,}96\frac{\sigma}{\sqrt{N}} < \mu < \overline{X} + 1{,}96\frac{\sigma}{\sqrt{N}}\right) = 0{,}95$. Essa expressão deve ser interpretada do seguinte modo: se construíssemos uma grande quantidade de intervalos $\left[\overline{X} - 1{,}96\sigma/\sqrt{N}\ ,\ \overline{X} + 1{,}96\sigma/\sqrt{N}\right]$, a partir de amostras de tamanho *N*, 95% desses intervalos conteriam o parâmetro μ e 5% deles não o conteriam. A Figura 6.2 ilustra o raciocínio, acima exposto, que é, de fato, a essência da estimação por intervalos. Observe que, nessa figura, todos os *k* intervalos, construídos a partir das *k* amostras de tamanho *N*, têm a mesma largura, mas são posicionados de modo diferente, em relação ao parâmetro μ. Se uma amostra específica produzir os limites [*i*, *s*], esses serão realizações das variáveis *I* e *S*, e, pelo exposto, terão uma chance de 95% de conter μ.

O raciocínio exposto nos parágrafos anteriores pode ser generalizado para a construção de intervalos de confiança para um parâmetro θ, de uma distribuição de probabilidades qualquer, a partir de uma amostra $y_1, y_2, y_3, \ldots, y_N$, extraída da população correspondente. Esse procedimento geral consta das seguintes etapas:

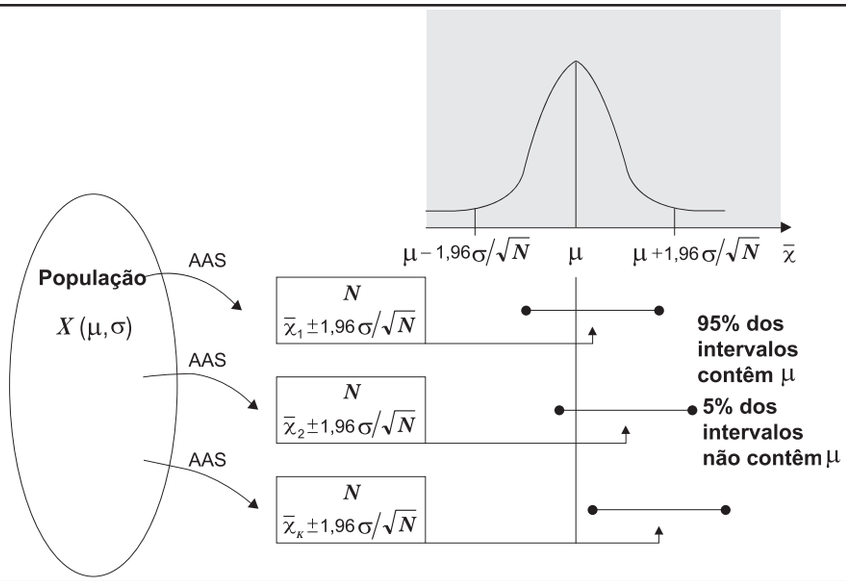


Figura 6.2 – Ilustração de um intervalo de confiança para μ, com σ conhecido e (1-α)=0,95 (adap. de Bussab e Morettin, 2002)

- selecione uma *função-pivô* $V = v(\theta, Y_1, Y_2, ..., Y_N)$, do parâmetro θ e das variáveis Y_1, Y_2, … ,Y_N, cuja densidade de probabilidades $g_V(v)$ tenha unicamente θ como parâmetro desconhecido;

- determine as constantes v_1 e v_2, tais que $P(v_1 < V < v_2) = 1 - \alpha$ ou que $P(V < v_1) = \alpha/2$ e $P(V > v_2) = \alpha/2$;

- usando as regras da álgebra, reescreva a desigualdade $v_1 < V < v_2$, de modo que o parâmetro θ fique isolado, em seu centro, e que se possa escrever que $P(I < \theta < S) = 1 - \alpha$;

- considere a amostra propriamente dita, substituindo as variáveis Y_1, Y_2, …, Y_N pelas observações $y_1, y_2, y_3, ..., y_N$, e calcule as realizações *i* e *s*, das variáveis aleatórias *I* e *S*; e

- o intervalo com confiança 100(1-α)%, para o parâmetro θ, é [*i*, *s*].

A maior dificuldade desse procedimento geral é a seleção de uma função-pivô adequada, o que nem sempre é possível. Entretanto, para alguns casos práticos importantes, a função-pivô e sua respectiva função densidade de probabilidades podem ser adequadamente obtidas. Esses casos práticos estão sumariados na Tabela 6.3.

Tabela 6.3 – Algumas funções-pivô para a construção de intervalos de confiança (IC), a partir de uma amostra de tamanho *N*

População	IC para o parâmetro:	Atributo do segundo parâmetro	Função-pivô *V*	Distribuição de *V*
Normal	μ	σ^2 conhecido	$\frac{\overline{Y}-\mu}{\sigma/\sqrt{N}}$	**N**(0,1)
Normal	μ	σ^2 desconhecido	$\frac{\overline{Y}-\mu}{S/\sqrt{N}}$	$t_{(n-1)}$
Normal	σ^2	μ conhecido	$\sum_{i=1}^{N}\left(\frac{Y_i-\mu}{\sigma}\right)^2$	χ^2_N
Normal	σ^2	μ desconhecido	$(N-1)\frac{S^2}{\sigma^2}$	$\chi^2_{(N-1)}$
Exponencial	θ	-	$\frac{2N\overline{Y}}{\theta}$	χ^2_{2N}

Exemplo 6.11 – Suponha que o consumo diário de água de uma comunidade seja uma variável Normal X e que uma amostra de 30 observações produziu $\overline{x} = 50\,\mathrm{m}^3$ e $s_X^2 = 256\,\mathrm{m}^6$. Pede-se (a) construir um IC para a média populacional μ, a um nível $100(1-\alpha) = 95\%$ e (b) construir um IC para a variância populacional σ^2, a um nível $100(1-\alpha) = 95\%$.
Solução: (a) Pela Tabela 6.3, a função-pivô, para esse caso é $V = \frac{\overline{X}-\mu}{S/\sqrt{N}}$, a qual segue uma distribuição ***t*** de Student, com $\nu = 30-1 = 29$ graus de liberdade. Com o objetivo de estabelecer a afirmação $P(v_1 < V < v_2) = 0{,}95$, verifica-se na tabela de ***t*** de Student, do Anexo 7, que $-v_1 = v_2 = \left|t_{0{,}025,\,29}\right| = 2{,}045$; observe que a distribuição ***t*** de Student é simétrica e que, portanto, os quantis correspondentes a $\alpha/2 = 0{,}025$ e $\alpha/2 = 0{,}975$ são simétricos em relação à média 0. Logo, $P\left(-2{,}045 < \frac{\overline{X}-\mu}{S/\sqrt{30}} < 2{,}045\right) = 0{,}95$. Manipulando essa desigualdade de tal modo que a média populacional μ reste isolada no centro da inequação, segue-se que $P\left(\overline{X} - 2{,}045\frac{S}{\sqrt{30}} < \mu < \overline{X} + 2{,}045\frac{S}{\sqrt{30}}\right) = 0{,}95$. Substituindo $\overline{X}$ e S pelas suas respectivas realizações $\overline{x} = 50\,\mathrm{m}^3$ e $s = \sqrt{256} = 16\,\mathrm{m}^3$, o IC a 95% para μ é [44,03; 55,97]. (b) Pela Tabela 6.2, a função-pivô, para esse caso é $(N-1)\frac{S^2}{\sigma^2}$, cuja distribuição é $\chi^2_{N-1=29}$. Para estabelecer

$P(v_1 < V < v_2) = 0{,}95$, verifica-se na tabela do Anexo 6 que v_1=16,047, para $\alpha/2 = 0{,}025$ e 29 graus de liberdade, e que v_2=45,722, para $\alpha/2 = 0{,}975$ e 29 graus de liberdade; observe que, no caso da distribuição do χ^2, não há simetria para os quantis. Logo, $P\left(16{,}047 < (30-1)\frac{S^2}{\sigma^2} < 45{,}722\right) = 0{,}95$. Manipulando essa desigualdade de modo semelhante ao feito no item (a), segue-se que $P\left(\frac{29S^2}{45{,}722} < \sigma^2 < \frac{29S^2}{16{,}047}\right) = 0{,}95$. Substituindo S^2 por sua realização $s^2 = 256$, segue-se que o IC para a variância populacional σ^2 é [162,37;462,64]. Nesse último caso, se 100(1- α) fosse alterado para 90%, o IC seria [174,45; 419,24] e, portanto, mais estreito, porém com menor nível de confiança.

A construção de intervalos de confiança para a média e a variância de uma população Normal é facilitada pela possibilidade de dedução de suas respectivas *distribuições exatas de amostragem*, tais como as distribuições do *t* de Student e do χ^2. De fato, as distribuições exatas de amostragem podem ser obtidas em forma explícita, quando a variável aleatória *X* segue distribuições de probabilidades que gozam da propriedade aditiva, tais como a Normal, a Gama, a Binomial e de Poisson. Para outras variáveis aleatórias, é quase sempre impossível determinar, de forma explícita, as distribuições exatas de amostragem de funções de momentos, tais como os coeficientes de assimetria e curtose, ou de um estimador $\hat{\theta}$ de um parâmetro populacional θ. Para esses casos, duas alternativas para a determinação das distribuições de amostragem são possíveis: os métodos que envolvem a simulação de Monte Carlo e os métodos assintóticos. Em ambas alternativas, os resultados são aproximados e, em muitos casos, os únicos disponíveis em problemas de inferência estatística.

Os métodos assintóticos, mais freqüentes para a solução desses problemas de inferência estatística, produzem resultados que são válidos quando os tamanhos das amostras tendem ao infinito. Obviamente, na prática, uma dada amostra é finita, sendo natural que se considere a questão de qual deve ser o seu tamanho para que as aproximações sejam razoáveis. Embora não haja respostas concisas e totalmente satisfatórias para questões como essa, é uma recomendação muito freqüente em livros de inferência estatística, que uma amostra é de tamanho 'suficientemente grande', quando $N > 50$ ou, pelo menos, quando $N > 30$. Cramér (1946) demonstrou que, sob condições gerais e para grandes valores de

N, as distribuições de amostragem de características, tais como funções de momentos e estimadores genéricos $\hat{\underline{\theta}}$, convergem assintoticamente para uma distribuição Normal de média igual ao valor populacional em questão, e de variância que pode ser escrita sob a forma c/N, onde *c* depende da característica estudada e do método de estimação. Hosking (1986) estendeu tais resultados para os estimadores de MPP's, momentos-L e suas respectivas funções, sob a condição que a distribuição da variável aleatória original tenha variância finita. Uma vez obtidas a média e o desvio padrão da distribuição Normal assintótica de um parâmetro genérico $\hat{\underline{\theta}}$, pode-se construir intervalos de confiança aproximados para θ, tais como os previamente exemplificados.

Como anteriormente mencionado, o fator *c* da variância, da distribuição Normal assintótica, depende do método de estimação. Se, por exemplo, o estimador $\hat{\underline{\theta}}$ é de máxima verossimilhança e se a distribuição tem um único parâmetro θ, prova-se que $1/c = E\left[\left\{\partial ln\left[f_X(x;\theta)\right] / \partial\theta\right\}^2\right]$. Entretanto, se a distribuição tem mais de um parâmetro, o cálculo do fator *c* da variância, da distribuição Normal assintótica, é relativamente mais complexo, pela necessária inclusão da dependência entre os estimadores de parâmetros. O método de estimação também afeta a eficiência assintótica dos estimadores, sendo um fato matemático que os estimadores MOM são assintoticamente menos eficientes do que os estimadores MVS. O leitor interessado em detalhes sobre essas questões deve remeter-se às referências Cramér (1946) e Rao (1973), para considerações teóricas, e Kaczmarek (1977) e Kite (1977), para exemplos e aplicações em hidrologia e meteorologia. O item seguinte, relativo à construção de intervalos de confiança para quantis, apresenta alguns resultados que são pertinentes às questões associadas aos erros inerentes aos estimadores de parâmetros.

6.6 – Intervalos de Confiança para Quantis

Uma vez estimados os parâmetros de uma distribuição de probabilidades $F_X(x)$, o interesse volta-se para um dos mais importantes objetivos da hidrologia estatística, que é o de estimar o quantil X_F, correspondente à probabilidade de não superação *F*, ou X_T, correspondente ao período de retorno *T*. O quantil X_F pode ser estimado pela função inversa de *F*, aqui denotada por $\phi(F)$, ou, em outros termos, $x_F = \hat{X}_F = \phi(F)$, ou ainda, $x_T = \hat{X}_T = \phi(T)$. É evidente que um estimador pontual, como $\hat{X}_T$, contém erros que são inerentes às incertezas presentes na estimação das características e parâmetros populacionais, a partir de amostras de tamanho *N*. Uma medida freqüentemente usada para quantificar a variabilidade presente em $\hat{X}_T$, e, portanto, indicar a confiabilidade das estimativas de quantis

de variáveis hidrológicas, é dada pelo chamado *erro padrão da estimativa*, denotado por S_T e definido por

$$S_T = \sqrt{E\left[\left\{\hat{X}_T - E\left[\hat{X}_T\right]\right\}^2\right]} \tag{6.22}$$

Deve-se ressaltar que 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 erro devido à seleção de uma distribuição de probabilidades inadequada. Logo, supondo que a distribuição $F_X(x)$ tenha sido corretamente especificada, o erro padrão da estimativa deverá subentender os erros presentes nas estimativas dos parâmetros de $F_X(x)$. Conseqüentemente, os métodos de estimação mais usuais, a saber, os métodos MOM, MVS e MML, produzirão diferentes erros-padrão da estimativa, sendo que o de maior eficiência, do ponto de vista estatístico, é aquele que resultar no menor valor para S_T.

A teoria assintótica de distribuições de amostragem demonstra que a distribuição de $\hat{X}_T$ é assintoticamente Normal, com média X_T e desvio-padrão S_T, quando $N \to \infty$. Como decorrência desse resultado, pode-se construir intervalos de confiança aproximados, a um nível 100(1-α)%, cujos limites são expressos por

$$\hat{X}_T \pm \left| z_{\alpha/2} \right| S_T \tag{6.23}$$

onde $z_{\alpha/2}$ representa a variável Normal padrão, de probabilidade de não superação igual a $\alpha/2$. Aplicando as propriedades do operador esperança matemática à equação 6.22, é possível demonstrar que, para uma distribuição de probabilidades genérica $F_X(x; \alpha,\beta)$, de 2 parâmetros quaisquer α e β, o quadrado do erro padrão da estimativa pode ser expresso por

$$S_T^2 = \left(\frac{\partial x}{\partial \alpha}\right)^2 \mathrm{Var}(\hat{\alpha}) + \left(\frac{\partial x}{\partial \beta}\right)^2 \mathrm{Var}(\hat{\beta}) + 2\left(\frac{\partial x}{\partial \alpha}\right)\left(\frac{\partial x}{\partial \beta}\right)\mathrm{Cov}(\hat{\alpha},\hat{\beta}) \tag{6.24}$$

Analogamente para uma distribuição $F_X(x; \alpha,\beta,\gamma)$, de 3 parâmetros quaisquer α, β e γ, prova-se que

$$\begin{aligned} S_T^2 = {} & \left(\frac{\partial x}{\partial \alpha}\right)^2 \mathrm{Var}(\hat{\alpha}) + \left(\frac{\partial x}{\partial \beta}\right)^2 \mathrm{Var}(\hat{\beta}) + \left(\frac{\partial x}{\partial \gamma}\right)^2 \mathrm{Var}(\hat{\gamma}) + \\ & + 2\left(\frac{\partial x}{\partial \alpha}\right)\left(\frac{\partial x}{\partial \beta}\right)\mathrm{Cov}(\hat{\alpha},\hat{\beta}) + 2\left(\frac{\partial x}{\partial \alpha}\right)\left(\frac{\partial x}{\partial \gamma}\right)\mathrm{Cov}(\hat{\alpha},\hat{\gamma}) + 2\left(\frac{\partial x}{\partial \beta}\right)\left(\frac{\partial x}{\partial \gamma}\right)\mathrm{Cov}(\hat{\beta},\hat{\gamma}) \end{aligned} \tag{6.25}$$

Nas equações 6.24 e 6.25, as derivadas parciais são calculadas pela relação $x_T = \hat{X}_T = \phi(T)$ e, portanto, dependem da expressão analítica da função inversa

da distribuição de probabilidades $F_X(x)$. Por outro lado, as variâncias e as covariâncias dos parâmetros dependem se o método de estimação é o dos momentos (MOM), o de máxima verossimilhança (MVS) ou dos momentos-L (MML). Examinaremos, a seguir, o caso mais geral de uma distribuição de 3 parâmetros, considerando cada um desses métodos de estimação.

6.6.1 – Intervalos de Confiança para Estimadores MOM de Quantis

Se o método para a estimação dos parâmetros α, β e γ, de $F_X(x; \alpha,\beta,\gamma)$, é o dos *momentos*, as respectivas variâncias e as covariâncias são calculadas a partir das relações entre os parâmetros e os momentos populacionais μ_1' (ou μ_X), μ_2 (ou σ_X^2) e μ_3 (ou $\gamma_X\sigma_X^3$), os quais são estimados pelos momentos amostrais m_1' (ou $\bar{x}$), m_2 (ou s_X^2) e m_3 (ou $g_X s_X^3$), com γ_X e g_X representando, respectivamente, os coeficientes de assimetria populacional e amostral de X. Pelo método dos momentos, portanto, o quantil $\hat{X}_T$ é uma função dos momentos amostrais m_1', m_2 e m_3, ou seja $\hat{X}_T = f(m_1', m_2 \text{ e } m_3)$, para um dado tempo de retorno T. Em decorrência dessa particularidade do método dos momentos, Kite (1977) reapresenta a equação 6.25, da seguinte forma:

$$\begin{aligned} S_T^2 = {} & \left(\frac{\partial \hat{X}_T}{\partial m_1'}\right)^2 \mathrm{Var}(m_1') + \left(\frac{\partial \hat{X}_T}{\partial m_2}\right)^2 \mathrm{Var}(m_2) + \left(\frac{\partial \hat{X}_T}{\partial m_3}\right)^2 \mathrm{Var}(m_3) + \\ & + 2\left(\frac{\partial \hat{X}_T}{\partial m_1'}\right)\left(\frac{\partial \hat{X}_T}{\partial m_2}\right)\mathrm{Cov}(m_1', m_2) + 2\left(\frac{\partial \hat{X}_T}{\partial m_1'}\right)\left(\frac{\partial \hat{X}_T}{\partial m_3}\right)\mathrm{Cov}(m_1', m_3) + \\ & + 2\left(\frac{\partial \hat{X}_T}{\partial m_3}\right)\left(\frac{\partial \hat{X}_T}{\partial m_2}\right)\mathrm{Cov}(m_3, m_2) \end{aligned} \tag{6.26}$$

onde as derivadas parciais devem ser obtidas das relações entre o quantil $\hat{X}_T$ e m_1', m_2 e m_3, tal como usadas em sua estimação. Ainda segundo Kite (1977), as variâncias e covariâncias de m_1', m_2 e m_3 são dadas por expressões que dependem dos *parâmetros populacionais* μ_2 a μ_6. São elas:

$$\mathrm{Var}(m_1') = \frac{\mu_2}{N} \tag{6.27}$$

$$\mathrm{Var}(m_2) = \frac{\mu_4 - \mu_2^2}{N} \tag{6.28}$$

$$\mathrm{Var}(m_3) = \frac{\mu_6 - \mu_3^2 - 6\mu_4\mu_2 + 9\mu_2^3}{N} \tag{6.29}$$

$$\mathrm{Cov}\left(m_1', m_2\right) = \frac{\mu_3}{N} \tag{6.30}$$

$$\mathrm{Cov}\left(m_1', m_3\right) = \frac{\mu_4 - 3\mu_2^2}{N} \tag{6.31}$$

$$\mathrm{Cov}\left(m_2, m_3\right) = \frac{\mu_5 - 4\mu_3\mu_2}{N} \tag{6.32}$$

Kite (1977) propõe que a solução da equação 6.26 seja facilitada pela expressão do quantil X_T como uma função dos dois primeiros momentos populacionais e do chamado fator de freqüência K_T, esse, por sua vez, dependente do tempo de retorno T e dos parâmetros da distribuição $F_X(x)$. Portanto, usando o fator de freqüência, dado pela expressão

$$K_T = \frac{X_T - \mu_1'}{\sqrt{\mu_2}} \tag{6.33}$$

e manipulando as equações 6.26 a 6.32, Kite (1977) propõe, finalmente, a seguinte equação para o cálculo de S_T^2 para estimadores MOM:

$$S_T^2 = \frac{\mu_2}{N}\left\{1 + K_T\gamma_1 + \frac{K_T^2}{4}\left(\gamma_2 - 1\right) + \frac{\partial K_T}{\partial \gamma_1}\left[2\gamma_2 - 3\gamma_1^2 - 6 + K_T\left(\gamma_3 - 6\gamma_1\frac{\gamma_2}{4} - 10\frac{\gamma_1}{4}\right)\right]\right\} +$$
$$+ \frac{\mu_2}{N}\left[\left(\frac{\partial K_T}{\partial \gamma_1}\right)^2\left(\gamma_4 - 3\gamma_3\gamma_1 - 6\gamma_2 - 9\gamma_1^2\frac{\gamma_2}{4} + 35\frac{\gamma_1^2}{4} + 9\right)\right] \tag{6.34}$$

onde,

$$\gamma_1 = \gamma_X = \frac{\mu_3}{\mu_2^{3/2}} \qquad \text{(coeficiente de assimetria populacional)} \tag{6.35}$$

$$\gamma_2 = \kappa = \frac{\mu_4}{\mu_2^2} \qquad \text{(coeficiente de curtose populacional)} \tag{6.36}$$

$$\gamma_3 = \frac{\mu_5}{\mu_2^{5/2}} \tag{6.37}$$

$$\gamma_4 = \frac{\mu_6}{\mu_2^3} \tag{6.38}$$

Observe, entretanto, que, para uma *distribuição de dois parâmetros*, o fator de freqüência K_T não depende do momento de ordem 3 e, portanto, as derivadas parciais presentes na equação 6.33 são nulas. Nesse caso, a equação 6.34 reduz-se a

$$S_T^2 = \frac{\mu_2}{N}\left\{1 + K_T\gamma_1 + \frac{K_T^2}{4}\left(\gamma_2 - 1\right)\right\} \tag{6.39}$$

Finalmente, o cálculo de intervalos de confiança do quantil X_T, estimado pelo método dos momentos a partir de uma amostra de tamanho N, é feito, inicialmente, pela substituição de $\gamma_1, \gamma_2, \gamma_3, \gamma_4, K_T$ e $\partial K_T/\partial\gamma_1$, na equação 6.34, pelos valores populacionais da distribuição de probabilidades em questão, e μ_2, por sua estimativa amostral. Em seguida, toma-se a raiz quadrada de S_T^2 e aplica-se a equação 6.23 para um nível de confiança previamente especificado 100(1-α) %. O exemplo 6.12, a seguir, ilustra o procedimento para a distribuição Gumbel, de 2 parâmetros. Outros exemplos e aplicações podem ser encontrados nas referências Kite (1977) e Rao e Hamed (2000).

Exemplo 6.12 – De posse dos resultados e estimativas MOM do exemplo 6.5, estime o intervalo de confiança, ao nível de 95%, para o quantil de tempo de retorno 100 anos.

Solução: Sabe-se que que $X \sim \mathbf{Gu_{max}}$ ($\hat{\alpha} = 17{,}739$, $\hat{\beta} = 72{,}025$) e que $N = 55$ (ver Anexo 3). A distribuição de Gumbel, com coeficientes de assimetria e curtose populacionais fixos e iguais a γ_1=1,1396 e γ_2=5,4, é de dois parâmetros, sendo válida, portanto, a equação 6.39. Substituindo as equações válidas para essa distribuição, a saber, de momentos $\mu_1' = \beta + 0{,}5772\alpha$ e $\mu_2 = \frac{\pi^2\alpha^2}{6}$, e a de quantis $X_T = \beta - \alpha\, ln\left[-ln\left(1-\frac{1}{T}\right)\right]$, na equação 6.33, é fácil verificar que $K_T = -0{,}45 - 0{,}7797\, ln\left[-ln\left(1-1/T\right)\right]$ e que para T=100, K_T=3,1367. De volta à equação 6.39, substituindo K_T, γ_1=1,1396, γ_2=5,4 e $\hat{\mu}_2 = \frac{\pi^2\hat{\alpha}^2}{6} = 517{,}6173$, resulta que $\hat{S}^2_{T=100} = 144{,}908$ e, portanto, $\hat{S}_{T=100} = 12{,}038$. Com esse último valor, com o quantil estimado $x_{T=100}$=153,16 e com $z_{0{,}025}$=-1,96 na equação 6.23, conclui-se que os limites do intervalo de confiança, a 95%, são [130,036; 177,224]. De acordo com o exposto e com o método MOM de estimação, esses limites contêm o verdadeiro quantil populacional, de tempo de retorno igual a 100 anos, com 95% de confiança.

6.6.2 – Intervalos de Confiança para Estimadores MVS de Quantis

Se o método para a estimação dos parâmetros α, β e γ, de $F_X(x; \alpha,\beta,\gamma)$, é o da *máxima verossimilhança*, as derivadas parciais, presentes nas equações 6.24 e 6.25, são calculadas pela relação $x_T = \hat{X}_T = \phi(T)$ e, portanto, dependem da expressão analítica da função inversa da distribuição de probabilidades $F_X(x)$. Por outro lado, segundo Kite (1977) e Rao e Hamed (2000), as variâncias e as covariâncias dos parâmetros são os elementos da seguinte matriz simétrica, denominada *matriz de covariância*:

$$I = \begin{bmatrix} \mathrm{Var}(\hat{\alpha}) & \mathrm{Cov}(\hat{\alpha},\hat{\beta}) & \mathrm{Cov}(\hat{\alpha},\hat{\gamma}) \\ & \mathrm{Var}(\hat{\beta}) & \mathrm{Cov}(\hat{\beta},\hat{\gamma}) \\ & & \mathrm{Var}(\hat{\gamma}) \end{bmatrix} \tag{6.40}$$

a qual, é dada pela inversa da seguinte outra matriz:

$$M = \begin{bmatrix} -\dfrac{\partial^2 ln\,L}{\partial\alpha^2} & -\dfrac{\partial^2 ln\,L}{\partial\alpha\partial\beta} & -\dfrac{\partial^2 ln\,L}{\partial\alpha\partial\gamma} \\ & -\dfrac{\partial^2 ln\,L}{\partial\beta^2} & -\dfrac{\partial^2 ln\,L}{\partial\beta\partial\gamma} \\ & & -\dfrac{\partial^2 ln\,L}{\partial\gamma^2} \end{bmatrix} \tag{6.41}$$

onde L representa a função de verossimilhança. Se D denota o determinante da matriz M, então, a variância de $\hat{\alpha}$, por exemplo, pode ser calculada pela divisão por D, do determinante da matriz restante, ao serem eliminadas a primeira linha e a primeira coluna de M. Em outros termos, a variância de $\hat{\alpha}$ é dada por

$$\mathrm{Var}(\hat{\alpha}) = \frac{\dfrac{\partial^2 ln\,L}{\partial\beta^2}\cdot\dfrac{\partial^2 ln\,L}{\partial\gamma^2} - \left(\dfrac{\partial^2 ln\,L}{\partial\beta\partial\gamma}\right)^2}{D} \tag{6.42}$$

Depois de calculados os elementos da matriz I, volta-se à equação 6.25 e estima-se S_T^2. Em seguida, toma-se a raiz quadrada de S_T^2 e aplica-se a equação 6.23 para um nível de confiança previamente especificado 100(1-α)%. O exemplo 6.13, a seguir, ilustra o procedimento para a distribuição Gumbel, de 2 parâmetros. Outros exemplos e aplicações podem ser encontrados nas referências Kite (1977) e Rao e Hamed (2000).

Exemplo 6.13 – De posse dos resultados e estimativas MVS do exemplo 6.8, estime o intervalo de confiança, ao nível de 95%, para o quantil de tempo de retorno 100 anos.

Solução: A função *ln L* da distribuição de Gumbel é

$ln[L(\alpha,\beta)] = -N\,ln(\alpha) - \frac{1}{\alpha}\sum_{i=1}^{N}(Y_i-\beta) - \sum_{i=1}^{N} exp\left(-\frac{Y_i-\beta}{\alpha}\right)$. Kimball (1949), citado por Kite (1977), apresenta as seguintes expressões aproximadas para as derivadas parciais de segunda ordem:

$\frac{\partial^2 ln L}{\partial\alpha^2} = -\frac{1{,}8237N}{\alpha^2}$; $\frac{\partial^2 ln L}{\partial\beta^2} = -\frac{N}{\alpha^2}$ e $\frac{\partial^2 ln L}{\partial\alpha\partial\beta} = \frac{0{,}4228N}{\alpha^2}$, as quais compõem os elementos da matriz M, que no caso presente tem dimensões 2×2. Invertendo-se a matriz M, conforme procedimento descrito no texto, tem-se finalmente, os elementos da matriz I, a saber, $\mathrm{Var}(\hat{\alpha}) = 0{,}6079\frac{\alpha^2}{N}$, $\mathrm{Var}(\hat{\beta}) = 1{,}1087\frac{\alpha^2}{N}$ e $\mathrm{Cov}(\hat{\alpha},\hat{\beta}) = 0{,}2570\frac{\alpha^2}{N}$. Uma vez que a função de quantis da distribuição de Gumbel é $Y_T = \beta - \alpha\, ln\left[-ln\left(1-\frac{1}{T}\right)\right]$, as derivadas parciais, presentes na equação 6.24 são as seguintes:

$\frac{\partial Y_T}{\partial\alpha} = -ln\left[-ln\left(1-\frac{1}{T}\right)\right] = W$ e $\frac{\partial Y_T}{\partial\beta} = 1$. Substituindo, na equação 6.24, as variâncias, covariâncias e derivadas parciais, tal como calculadas, resulta que a variância dos quantis de MVS de Gumbel é

$S_T^2 = \frac{\alpha^2}{N}(1{,}1087 + 0{,}5140W + 0{,}6079W^2)$. Para a amostra de N=55, em questão, os resultados da estimação MVS do exemplo 6.8 são $\hat{\alpha} = 19{,}4$ e $\hat{\beta} = 71{,}7$. Com esses resultados e W = 4,60, para T = 100, conclui-se que $S_T^2 = 130{,}787$ e, portanto, $S_T = 11{,}436$. Comparando esse resultado com o obtido no exemplo 6.12, verifica-se que os estimadores MVS produzem quantis de menor variância e, portanto, mais confiáveis, do que os estimadores MOM. Com o valor calculado para S_T, com o quantil estimado $x_{T=100}$=160,94 e com $z_{0,025}$= -1,96 na equação 6.23, conclui-se que os limites do intervalo de confiança, a 95%, são [138,530; 183,350]. De acordo com o exposto e com o método MVS de estimação, esses limites contêm o verdadeiro quantil populacional, de tempo de retorno igual a 100 anos, com 95% de confiança.

6.6.3 – Intervalos de Confiança para Estimadores MML de Quantis

Se o método para a estimação dos parâmetros α, β e γ, de $F_X(x; \alpha,\beta,\gamma)$, é o dos *momentos-L*, as derivadas parciais, presentes nas equações 6.24 e 6.25, são calculadas pela relação $x_T = \hat{X}_T = \phi(T)$ e, portanto, dependem da expressão analítica da função inversa da distribuição de probabilidades $F_X(x)$. Por outro lado, as variâncias e as covariâncias dos parâmetros são os elementos da matriz de covariância, idêntica à expressa pela equação 6.40. Seus elementos, porém, são calculados pela matriz de covariância dos MPP's α_r e β_r, para r =1, 2 e 3. Hosking (1986) demonstrou que o vetor $b = \left(b_1, b_2, b_3 \right)^T$ é assintoticamente distribuído segundo uma Normal multivariada, com médias $\beta = \left(\beta_1, \beta_2, \beta_3 \right)^T$ e matriz de covariância $\boldsymbol{V}/N$. As expressões para avaliar a matriz $\boldsymbol{V}$ e, na seqüência, o erro padrão da estimativa S_T são bastante complexas e encontram-se disponíveis, em Hosking (1986), para algumas distribuições notáveis.

Exemplo 6.14 – De posse dos resultados e estimativas MML do exemplo 6.10, estime o intervalo de confiança, ao nível de 95%, para o quantil de tempo de retorno 100 anos.

Solução: Hosking (1986) apresenta as seguintes expressões para as variâncias e covariâncias dos estimadores MML, para os parâmetros α e β da distribuição de Gumbel: $\mathrm{Var}(\hat{\alpha}) = 0{,}8046\dfrac{\alpha^2}{N}$, $\mathrm{Var}(\hat{\beta}) = 1{,}1128\dfrac{\alpha^2}{N}$ e $\mathrm{Cov}(\hat{\alpha}, \hat{\beta}) = 0{,}2287\dfrac{\alpha^2}{N}$. As derivadas parciais, presentes na equação 6.24 são as seguintes: $\dfrac{\partial Y_T}{\partial \alpha} = -ln\left[-ln\left(1 - \dfrac{1}{T} \right) \right] = W$ e $\dfrac{\partial Y_T}{\partial \beta} = 1$. Substituindo, na equação 6.24, as variâncias, covariâncias e derivadas parciais, tal como calculadas, resulta que a variância dos quantis de MVS de Gumbel é $S_T^2 = \dfrac{\alpha^2}{N}(1{,}1128 + 0{,}4574W + 0{,}8046W^2)$. Para a amostra de N =10, em questão, os resultados da estimação MML do exemplo 6.8 são $\hat{\alpha} = 19{,}381$ e $\hat{\beta} = 74{,}103$. Com esses resultados e W = 4,60, para T=100, conclui-se que $S_T^2 = 760{,}39$ e, portanto, $S_T = 27{,}58$. Observe que, para uma amostra pequena de apenas 10 observações, S_T^2 é relativamente muito maior do que nos exemplos anteriores. O quantil de 100 anos é $\hat{y}(T = 100) = \hat{\beta} - \hat{\alpha}\, ln\left[-ln\left(1 - \dfrac{1}{100} \right) \right] = 163{,}26$. Com o valor calculado para S_T, com o quantil estimado e com $z_{0{,}025}$= -1,96 na equação 6.23, conclui-se que os limites do intervalo de confiança, a 95%, são [109,21; 217,31].

6.7 – Sumário dos Estimadores Pontuais

Apresenta-se a seguir um sumário das equações de estimativas de parâmetros, pelos métodos MOM e MVS, para algumas distribuições de probabilidades, organizadas em ordem alfabética. As soluções para estimadores MOM e MVS, de grande parte das distribuições de variáveis aleatórias contínuas, listadas a seguir, encontram-se implementadas no *software* ALEA, cujos programa executável e manual do usuário estão disponíveis na URL http://www.ehr.ufmg.br. Em todos os casos, as equações baseiam-se em uma amostra $\{x_1, x_2, \ldots, x_N\}$, de tamanho N. Em alguns casos, apresenta-se também um sumário das equações das estimativas pelo método MML.

6.7.1 – Distribuição de Bernoulli

Método MOM: $\hat{p} = \overline{x}$

Método MVS: $\hat{p} = \overline{x}$

Método MML: $\hat{p} = l_1$

6.7.2 – Distribuição Beta

Método MOM:

$\hat{\alpha}$ e $\hat{\beta}$ são as soluções do sistema

$$\overline{x} = \frac{\alpha}{\alpha + \beta} \quad \text{e}$$

$$s_X^2 = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

Método MVS:

$\hat{\alpha}$ e $\hat{\beta}$ são as soluções do sistema

$$\frac{\partial}{\partial\alpha}[ln\,\Gamma(\alpha) - ln\,\Gamma(\alpha+\beta)] = \frac{1}{N}\sum_{i=1}^{N} ln(x_i)$$

$$\frac{\partial}{\partial\beta}[ln\,\Gamma(\beta) - ln\,\Gamma(\alpha+\beta)] = \frac{1}{N}\sum_{i=1}^{N} ln(1-x_i)$$

6.7.3 – Distribuição Binomial

Suponha que o número de tentativas independentes de Bernoulli seja conhecido e igual a *m*.

Método MOM: $\hat{p} = \bar{x}/m$

Método MVS: $\hat{p} = \bar{x}/m$

Método MML: $\hat{p} = l_1/m$

6.7.4 – Distribuição Exponencial

Método MOM: $\hat{\theta} = \bar{x}$

Método MVS: $\hat{\theta} = \bar{x}$

Método MML: $\hat{\theta} = l_1$

6.7.5 – Distribuição Gama

Método MOM: $\hat{\theta} = \dfrac{s_X^2}{\bar{x}}$

$$\hat{\eta} = \frac{\bar{x}^2}{s_X^2}$$

Método MVS:

$\hat{\eta}$ é a solução da equação $\quad ln\,\eta - \dfrac{\partial}{\partial \eta} ln\,\Gamma(\eta) = ln\,\bar{x} - \dfrac{1}{N}\sum_{i=1}^{N} ln\,x_i \quad$ (A)

Depois de resolver (A), $\hat{\theta} = \bar{x}/\hat{\eta}$.

A solução da equação (A) pode ser aproximada por:

$$\hat{\eta} = \frac{0{,}5 + 0{,}1649y - 0{,}0544y^2}{y} \quad \text{se } 0 \le y \le 0{,}5772, \text{ ou}$$

$$\hat{\eta} = \frac{8{,}899 + 9{,}060y - 0{,}9775y^2}{y\left(17{,}7973 + 11{,}9685y + y^2\right)} \quad \text{se } 0{,}5772 < y \le 17$$

onde $y = ln\,\overline{x} - \frac{1}{N}\sum_{i=1}^{N} ln\,x_i$

Método MML:

$\hat{\eta}$ é a solução (método de Newton) da equação $\frac{l_2}{l_1} = \frac{\Gamma(\eta + 0{,}5)}{\sqrt{\pi}\,\Gamma(\eta + 1)}$ (B)

Depois de resolver (B), $\hat{\theta} = l_1 / \hat{\eta}$.

6.7.6 – Distribuição Geométrica

Método MOM: $\hat{p} = 1/\overline{x}$

Método MVS: $\hat{p} = 1/\overline{x}$

Método MML: $\hat{p} = 1/l_1$

6.7.7 – Distribuição Generalizada de Valores Extremos (GEV)

Método MOM:

Alternativa 1:

resolver para κ, a equação 5.73, do capítulo 5, substituindo γ pelo coeficiente de assimetria amostral g_X. A solução é iterativa, pelo método de Newton.

Alternativa 2:

para coeficientes de assimetria amostrais $1{,}1396 < g_X < 10$ ($g = g_X$):

$\hat{\kappa} = 0{,}2858221 - 0{,}357983g + 0{,}116659g^2 - 0{,}022725g^3 + 0{,}002604g^4 - 0{,}000161g^5 + 0{,}000004g^6$

para coeficientes de assimetria amostrais $-2 < g_X < 1{,}1396$ ($g_X = g$):

$\hat{\kappa} = 0{,}277648 - 0{,}322016g + 0{,}060278g^2 + 0{,}016759g^3 - 0{,}005873g^4 - 0{,}00244g^5 - 0{,}00005g^6$

para coeficientes de assimetria amostrais $-10 < g_X < 0$ ($g = g_X$):

$\hat{\kappa} = -0{,}50405 - 0{,}00861g + 0{,}015497g^2 + 0{,}005613g^3 + 0{,}00087g^4 + 0{,}000065g^5$

Em seguida, $\hat{\alpha} = \frac{s_X \hat{\kappa}}{\sqrt{\Gamma(1 + 2\hat{\kappa}) - \Gamma^2(1 + \hat{\kappa})}}$ e $\hat{\beta} = \overline{x} - \frac{\hat{\alpha}}{\hat{\kappa}}[1 - \Gamma(1 + \hat{\kappa})]$

Método MVS:

$\hat{\alpha}, \hat{\beta}, \hat{\kappa}$ são as soluções simultâneas (método de Newton) do seguinte sistema:

$$\frac{1}{\alpha}\left[\sum_{i=1}^{N} exp(-y_i - \kappa y_i) - (1-\kappa)\sum_{i=1}^{N} exp(\kappa y_i)\right] = 0 \quad \text{(C)}$$

$$\frac{1}{\kappa\alpha}\left[\sum_{i=1}^{N} exp(-y_i - \kappa y_i) - (1-\kappa)\sum_{i=1}^{N} exp(\kappa y_i) + N - \sum_{i=1}^{N} exp(-y_i)\right] = 0 \quad \text{(D)}$$

$$\frac{1}{\kappa^2}\left[\sum_{i=1}^{N} exp(-y_i - \kappa y_i) - (1-\kappa)\sum_{i=1}^{N} exp(\kappa y_i) + N - \sum_{i=1}^{N} exp(-y_i)\right] +$$

$$+\frac{1}{\kappa}\left[-\sum_{i=1}^{N} y_i + \sum_{i=1}^{N} y_i\, exp(y_i) + N\right] = 0 \quad \text{(E)}$$

onde $y_i = \frac{1}{\kappa} ln\left[1 - \kappa\left(\frac{x_i - \beta}{\alpha}\right)\right]$. A resolução desse sistema é complexa; sugere-se as referências Prescott e Walden (1983) e Hosking (1985) para algoritmo de resolução.

Método MML:

$$\hat{\kappa} = 7{,}8590C + 2{,}9554C^2 \text{, onde } C = 2/(3 + t_3) - ln\,2/ln\,3$$

$$\hat{\alpha} = \frac{l_2 \hat{\kappa}}{\Gamma(1+\hat{\kappa})\left(1 - 2^{-\hat{\kappa}}\right)}$$

$$\hat{\beta} = l_1 - \frac{\hat{\alpha}}{\hat{\kappa}}[1 - \Gamma(1+\hat{\kappa})]$$

6.7.8 – Distribuição Gumbel (máximos)

Método MOM:

$$\hat{\alpha} = 0{,}7797\, s_X$$
$$\hat{\beta} = \bar{x} - 0{,}45 s_X$$

Método MVS:

$\hat{\alpha}$ e $\hat{\beta}$ são as soluções do seguinte sistema de equações:

$$\frac{\partial}{\partial\alpha} ln\left[L(\alpha,\beta)\right] = -\frac{N}{\alpha} + \frac{1}{\alpha^2}\sum_{i=1}^{N}(x_i - \beta) - \frac{1}{\alpha^2}\sum_{i=1}^{N}(x_i - \beta) exp\left(-\frac{x_i - \beta}{\alpha}\right) = 0 \qquad \text{(F)}$$

$$\frac{\partial}{\partial\beta} ln\left[L(\alpha,\beta)\right] = \frac{N}{\alpha} - \frac{1}{\alpha}\sum_{i=1}^{N} exp\left(-\frac{x_i - \beta}{\alpha}\right) = 0 \qquad \text{(G)}$$

Manipulando-se ambas equações, chega-se a

$$F(\alpha) = \sum_{i=1}^{N} x_i \, exp\left(-\frac{x_i}{\alpha}\right) - \left(\frac{1}{N}\sum_{i=1}^{N} x_i - \alpha\right)\sum_{i=1}^{N} exp\left(-\frac{x_i}{\alpha}\right) = 0 \qquad \text{(H)}$$

A solução de (H), pelo método de Newton, fornece $\hat{\alpha}$.

Em seguida, $\hat{\beta} = \hat{\alpha}\, ln\left[\dfrac{N}{\sum_{i=1}^{N} exp(-x_i/\alpha)}\right]$.

Método MML:

$$\hat{\alpha} = \frac{l_2}{ln\,2}$$

$$\hat{\beta} = l_1 - 0{,}5772\hat{\alpha}$$

6.7.9 – Distribuição Gumbel (mínimos)

Método MOM:

$$\hat{\alpha} = 0{,}7797\, s_X$$

$$\hat{\beta} = \hat{x} + 0{,}45 s_X$$

Método MML:

$$\hat{\alpha} = \frac{l_2}{ln\,2}$$

$$\hat{\beta} = l_1 + 0{,}5772\hat{\alpha}$$

6.7.10 – Distribuição Log-Normal

Método MOM:

$$\hat{\sigma}_Y = \sqrt{ln\left(CV_X^2 + 1\right)}$$

$$\hat{\mu}_Y = ln\,\bar{x} - \frac{\hat{\sigma}_Y^2}{2} \qquad \text{com } Y = \ln X$$

Método MVS:

$$\hat{\mu}_Y = \bar{y}$$
$$\hat{\sigma}_Y = s_Y$$

Método MML:

$$\hat{\sigma}_Y = 2\,erf^{-1}(t)$$
$$\hat{\mu}_Y = ln\,l_1 - \frac{\hat{\sigma}_Y^2}{2}$$

onde $erf\,(w) = \frac{2}{\sqrt{\pi}} \int_0^w e^{-u^2}\,du$. A inversa $erf^{-1}(t)$ é igual a $u/\sqrt{2}$, com u representando a variável Normal padrão correspondente $\Phi[(t+1)/2]$.

6.7.11 – Distribuição Log-Pearson Tipo III

Método MOM:

Lembrando que $\mu'_r = \frac{exp(\gamma r)}{(1 - r\alpha)^{\beta}}$ são estimados por m'_r, $\hat{\alpha}$, $\hat{\beta}$, $\hat{\gamma}$ são as soluções de:

$$ln\,m'_1 = \gamma - \beta\,ln(1-\alpha)$$
$$ln\,m'_2 = 2\gamma - \beta\,ln(1-2\alpha)$$
$$ln\,m'_3 = 3\gamma - \beta\,ln(1-3\alpha)$$

Para a solução desse sistema, Kite (1977) sugere:

- defina $B = \frac{ln\,m'_3 - 3\,ln\,m'_1}{ln\,m'_2 - 2\,ln\,m'_1}$, $A = \frac{1}{\alpha} - 3$ e $C = \frac{1}{B-3}$
- para $3,5 < B < 6$, $A = -0,23019 + 1,65262C + 0,20911C^2 - 0,04557C^3$
- para $3,0 < B \leq 3,5$, $A = -0,47157 + 1,99955C$

- $\hat{\alpha} = \dfrac{1}{A+3}$
- $\hat{\beta} = \dfrac{ln\, m_2^{'} - 2\, ln\, m_1^{'}}{ln(1-\hat{\alpha})^2 - ln(1-2\hat{\alpha})}$
- $\hat{\gamma} = ln\, m_1^{'} + \hat{\beta}\, ln(1-\hat{\alpha})$

Método MVS:

$\hat{\alpha}, \hat{\beta}, \hat{\gamma}$ são as soluções (método de Newton) do seguinte sistema:

$$\sum_{i=1}^{N} \left(ln\, x_i - \gamma \right) = N\alpha\beta$$

$$N\,\Psi(\beta) = \sum_{i=1}^{N} ln\left[\left(ln\, x_i - \gamma \right) / \alpha \right]$$

$$N = \alpha \left(\beta - 1 \right) \sum_{i=1}^{N} \frac{1}{ln\, x_i - \gamma}$$

onde $\Psi(\beta) = \dfrac{\Gamma'(\beta)}{\Gamma(\beta)}$, a qual, conforme Abramowitz e Stegun (1965), pode ser aproximada por

$$\Psi(\beta) \cong ln\,\beta - \frac{1}{2\beta} - \frac{1}{12\beta^2} + \frac{1}{120\beta^4} - \frac{1}{252\beta^6} + \frac{1}{240\beta^8} - \frac{1}{132\beta^{10}} \cdot$$

Método MML:

As estimativas pelo método MML podem ser obtidas por procedimento idêntico ao ilustrado para a distribuição Pearson Tipo III, com a transformação z_i=ln(x_i).

6.7.12 – Distribuição Normal

Método MOM:

$$\hat{\mu}_X = \bar{x}$$
$$\hat{\sigma}_X = s_X$$

Método MVS:

$$\hat{\mu}_X = \bar{x}$$
$$\hat{\sigma}_X = s_X$$

Método MML:

$$\hat{\mu}_X = l_1$$
$$\hat{\sigma}_X = \sqrt{\pi} l_2$$

6.7.13 – Distribuição Pearson Tipo III

Método MOM:

$$\hat{\beta} = \left(\frac{2}{g_X} \right)^2$$

$$\hat{\alpha} = \sqrt{\frac{s_X^2}{\hat{\beta}}}$$

$$\hat{\gamma} = \bar{x} - \sqrt{s_X^2 \hat{\beta}}$$

Método MVS:

$\hat{\alpha}, \hat{\beta}, \hat{\gamma}$ são as soluções (método de Newton) do seguinte sistema:

$$\sum_{i=1}^{N} \left(x_i - \gamma \right) = N\alpha\beta$$

$$N\,\Psi(\beta) = \sum_{i=1}^{N} ln\left[\left(ln\, x_i - \gamma \right) / \alpha \right]$$

$$N = \alpha \left(\beta - 1 \right) \sum_{i=1}^{N} \frac{1}{ln\, x_i - \gamma}$$

onde $\Psi(\beta) = \dfrac{\Gamma'(\beta)}{\Gamma(\beta)}$ (ver distribuição Log-Pearson Tipo III).

Método MML:

Para $t_3 \geq 1/3$ e com $t_m = 1 - t_3$, $\hat{\beta} = \dfrac{0{,}36067 t_m - 0{,}5967 t_m^2 + 0{,}25361 t_m^3}{1 - 2{,}78861 t_m + 2{,}56096 t_m^2 - 0{,}77045 t_m^3}$

Para $t_3 < 1/3$ e com $t_m = 3\pi t_3^2$, $\hat{\beta} = \dfrac{1 + 0{,}2906 t_m}{t_m + 0{,}1882 t_m^2 + 0{,}0442 t_m^3}$

$$\hat{\alpha} = \sqrt{\pi} l_2 \frac{\Gamma\left(\hat{\beta} \right)}{\Gamma\left(\hat{\beta} + 0{,}5 \right)}$$

$$\hat{\gamma} = l_1 - \hat{\alpha}\hat{\beta}$$

6.7.14 – Distribuição de Poisson

Método MOM:

$$\hat{\nu} = \overline{x}$$

Método MVS:

$$\hat{\nu} = \overline{x}$$

6.7.15 – Distribuição Uniforme

Método MOM:

$$\hat{a} = \overline{x} - \sqrt{3} s_x$$
$$\hat{b} = \overline{x} + \sqrt{3} s_x$$

Método MVS:

$$\hat{a} = Min(x_i)$$
$$\hat{b} = Max(x_i)$$

Método MML:

$\hat{a}$ e $\hat{b}$ são as soluções de $l_1 = (a+b)/2$ e $l_2 = (b-a)/6$.

6.7.16 – Distribuição Weibull (mínimos)

Método MOM:

$\hat{\alpha}$ e $\hat{\beta}$ são as soluções do seguinte sistema de equações:

$$\overline{x} = \beta \Gamma\left(1 + \frac{1}{\alpha}\right)$$

$$s_X^2 = \beta^2 \left[\Gamma\left(1 + \frac{2}{\alpha}\right) - \Gamma^2\left(1 + \frac{1}{\alpha}\right) \right]$$

`

(Ver item 5.7.2.5 do capítulo 5).

Método MVS:

$\hat{\alpha}$ e $\hat{\beta}$ são as soluções (método de Newton) do seguinte sistema de equações:

$$\beta^{-\alpha} = \frac{N}{\sum_{i=1}^{N} x_i^{\alpha}}$$

$$\alpha = \frac{N}{\beta^{-\alpha}\sum_{i=1}^{N} x_i^{\alpha}\, ln(x_i) - \sum_{i=1}^{N} ln(x_i)}$$

Exercícios

1) Dada a função densidade $f_X(x) = \frac{x^{\theta-1} exp(-x)}{\Gamma(\theta)}, x > 0, \theta > 0,$ determine o valor de c, tal que cX seja um estimador não-enviesado de θ. Recorde-se da seguinte propriedade da função gama: $\Gamma(\theta + 1) = \theta\,\Gamma(\theta)$.

2) Suponha que $\{Y_1, Y_2, \ldots, Y_N\}$ seja uma AAS de uma FDP cuja média é μ. Pergunta-se sob quais condições $W = \sum_{i=1}^{N} a_i Y_i$ é um estimador não-enviesado de μ.

3) Considere que X_1 e X_2 seja uma AAS de tamanho 2 de uma distribuição exponencial. Se $Y = \sqrt{X_1 X_2}$ representa a média geométrica de X_1 e X_2, prove que $W = \pi/4Y$ é um estimador não-enviesado de θ.

4) A distribuição Exponencial de 2 parâmetros é definida pela função densidade $f_X(x) = \frac{1}{\theta} exp\left(-\frac{x-\xi}{\theta}\right), x \geq \xi$, onde ξ denota o parâmetro de posição. Determine os estimadores de ξ e θ, pelos métodos MOM e MVS.

5) Suponha que W_1 e W_2 denotem dois estimadores não-enviesados de um certo parâmetro θ, com variâncias respectivamente iguais a Var(W_1) e Var(W_2). O estimador W_1 é dito *mais eficiente* do que W_2 se Var(W_1) < Var(W_2). Além disso, a *eficiência relativa* de W_1, em relação a W_2 é definida pela razão Var(W_2)/ Var(W_1). Considere que X_1, X_2 e X_3 seja uma AAS de tamanho 3 de uma distribuição exponencial de parâmetro θ. Calcule a eficiência relativa de $W_1 = \frac{X_1 + 2X_2 + X_3}{4}$, em relação a $W_2 = \overline{X}$.

6) Conforme menção anterior, um estimador $W_N = h(X_1, X_2, ..., X_N)$ é considerado *consistente*, para θ, se ele converge, em probabilidade, para θ. Em outros termos, se, para quaisquer ε, $\delta > 0$, existir um $n(\varepsilon, \delta)$ tal que

$P(|W_N - \theta| < \varepsilon) > 1 - \delta$ para $N > n(\varepsilon, \delta)$. Suponha que $\{X_1, X_2, \ldots, X_N\}$ seja uma AAS da FDP $f_X(x;\theta) = 1/\theta$ para $0 < y < \theta$ e que $W_N = X_{max}$. É possível demonstrar que W_N é um estimador enviesado de θ, embora possa ser consistente. A questão da consistência passa a ser posta na existência (ou não) de $n(\varepsilon,\delta)$, suficientemente grande, para que $\mathbf{P}(|W_N - \theta| < \varepsilon) > 1 - \delta$ *para* $N > n(\varepsilon,\delta)$. Mostre que W_N é um estimador consistente de θ. Para resolver esse exercício, recorde-se que a FDP *exata* do máximo de uma AAS pode ser obtida pelos métodos descritos no item 5.7.1, do capítulo 5. No caso presente, pode-se mostrar que $f_{W_N}(w_N) = \frac{N(w_N)^{N-1}}{\theta^N}$.

7) Conforme menção anterior, um estimador $W = h(X_1, X_2, \ldots, X_N)$ é considerado *suficiente*, para θ, se, para todo θ e para quaisquer valores amostrais, a FDP de $(X_1, X_2, \ldots, X_N)$, condicionada a *w*, não depende de θ. Mais precisamente, W é suficiente se $\frac{f_{X_1}(x_1).f_{X_2}(x_2)\ldots f_{X_N}(x_N)}{f_W(w)}$ não depende de θ. Considere o estimador W_N, descrito no exercício 6. Demonstre que W_N é um estimador suficiente.

8) O Anexo 2 apresenta as vazões médias diárias máximas anuais da estação fluviométrica do Rio Paraopeba em Ponte Nova do Paraopeba (código 40800001), para os anos hidrológicos de 1938-39 a 1998-99. Use os métodos descritos nesse capítulo para calcular (a) as estimativas dos parâmetros da distribuição Gama, pelos métodos MOM, MVS e MML; (b) a probabilidade da vazão média diária máxima anual superar 1000 m^3/s, em um ano qualquer, usando as estimativas de parâmetros obtidas pelos três métodos; (c) o quantil de tempo de retorno igual a 100 anos, usando as estimativas de parâmetros obtidas pelos três métodos; e (d) compare os resultados obtidos em (b) e (c).

9) Repita o exercício 8 para a distribuição Exponencial.

10) Repita o exercício 8 para a distribuição GEV.

11) Repita o exercício 8 para a distribuição Gumbel (máximos).

12) Repita o exercício 8 para a distribuição Log-Normal.

13) Repita o exercício 8 para a distribuição Log-Pearson Tipo III.

14) Repita o exercício 8 para a distribuição Pearson Tipo III.

15) Os dados da tabela abaixo correspondem aos números de Manning *n,* determinados experimentalmente por Haan (1965), para tubos plásticos.

0,0092	0,0085	0,0083	0,0091
0,0078	0,0084	0,0091	0,0088
0,0086	0,0090	0,0089	0,0093
0,0081	0,0092	0,0085	0,0090
0,0085	0,0088	0,0088	0,0093

Suponha que essa amostra tenha sido extraída de uma população Normal, de parâmetros μ e σ. Pede-se: (a) construir um intervalo de confiança para a média μ, a um nível $100(1-\alpha) = 95\%$; e (b) construir um intervalo de confiança para a variância σ^2, a um nível $100(1-\alpha) = 95\%$.

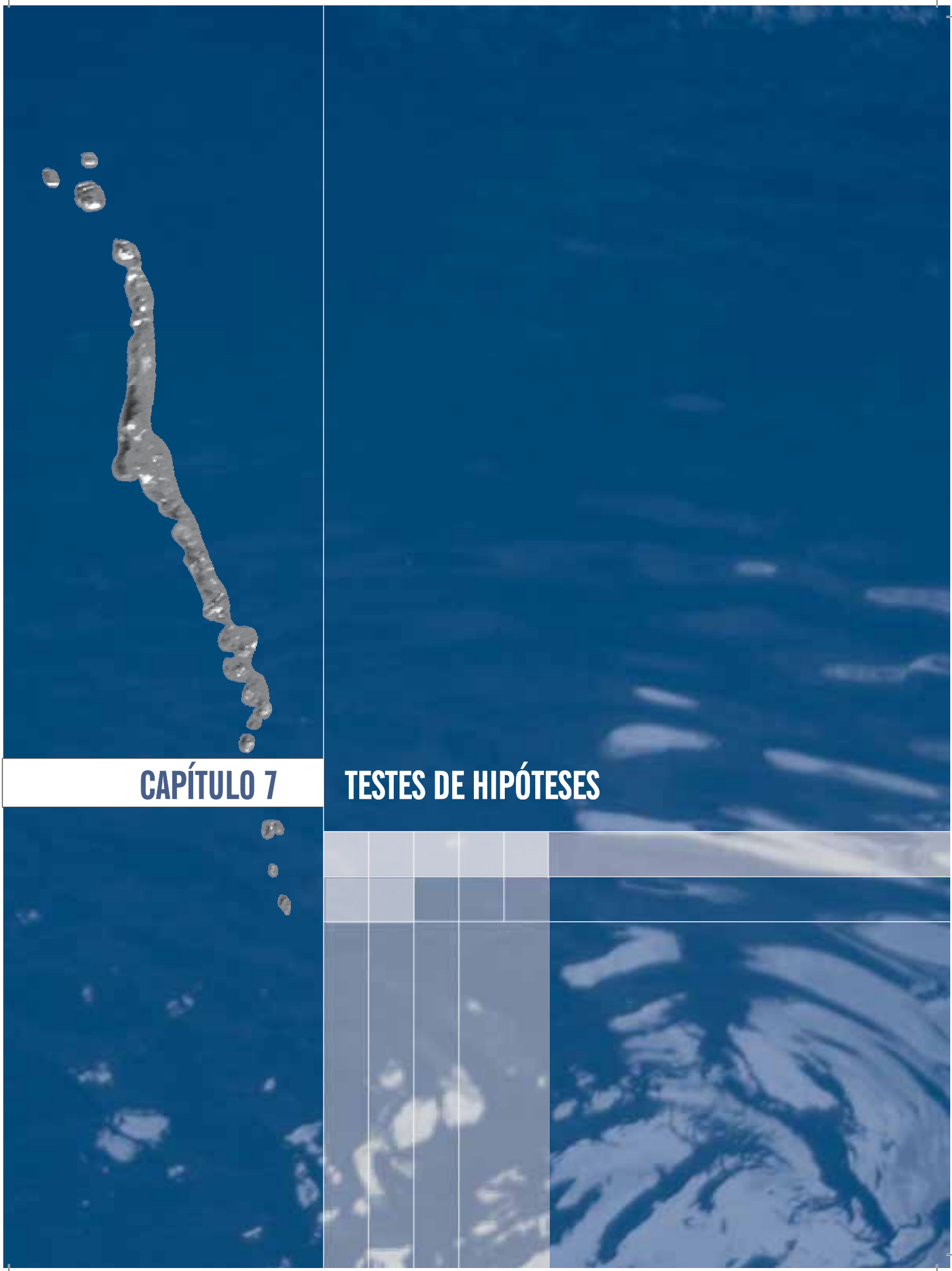
16) Repita o exercício 15, para um nível de confiança de 90%. Interprete as diferenças.

17) Suponha que, no item (a) do exercício 15, a variância populacional fosse conhecida e igual à estimativa obtida por meio da amostra. Sob essa condição, refaça o item (a) do exercício 15 e interprete as diferenças nos resultados.

18) Suponha que, no item (b) do exercício 15, a média populacional fosse conhecida e igual à estimativa obtida por meio da amostra. Sob essa condição, refaça o item (b) do exercício 15 e interprete as diferenças nos resultados.

19) De volta às vazões médias diárias máximas anuais do Rio Paraopeba em Ponte Nova do Paraopeba (Anexo 2), construa os intervalos de confiança, a um nível 95%, para os quantis de Gumbel, estimados pelos métodos MOM, MVS e MML, para os tempos de retorno iguais a 2, 50, 100 e 500 anos. Decida qual é o método de estimação mais eficiente. Interprete os resultados obtidos, do ponto de vista da variação do tempo de retorno.

20) A confiabilidade dos estimadores MOM, MVS e MML de parâmetros e quantis das distribuições de probabilidades mais usadas em hidrologia tem sido objeto de numerosos estudos. Esses estudos levam em consideração as principais propriedades dos estimadores e muitos deles, permitem a comparação entre os estimadores MOM, MVS e MML, de parâmetros e quantis. As referências Rao e Hamed (2000), Kite (1977) e Hosking (1986) fazem uma síntese dos principais resultados obtidos nesses estudos. Pede-se ao leitor recorrer a essas referências e preparar um sumário comparado das principais características dos estimadores MOM, MVS e MML, para as distribuições Exponencial, Gumbel (máximos), GEV, Gama, Pearson Tipo III, Log-Pearson Tipo III, Normal e Log-Normal.



CAPÍTULO 7

TESTES DE HIPÓTESES

CAPÍTULO 7 TESTES DE HIPÓTESES

Além dos métodos de estimação de parâmetros e de construção de intervalos de confiança, os *testes de hipóteses* são procedimentos usuais da inferência estatística, úteis na tomada de decisões que concernem à forma, ou ao valor de um certo parâmetro, de uma distribuição de probabilidades, da qual se conhece apenas uma amostra de observações. Tais testes envolvem a formulação de uma hipótese, na forma de uma declaração conjectural sobre o comportamento probabilístico da população. Essa hipótese pode se materializar, por exemplo, em uma premissa, formulada *a priori*, a respeito de um certo parâmetro populacional de uma variável aleatória. Não rejeitar ou rejeitar uma tal hipótese irá depender do confronto entre a conjectura e a realidade física, essa concretizada pelas observações que compõem a amostra. A *rejeição* da hipótese implica na necessidade de eventual revisão da conjectura inicial, em decorrência de seu desacordo com a realidade imposta pelos dados amostrais. Por outro lado, a *não rejeição* da hipótese significa que, com base nos dados amostrais, não há elementos suficientes para descartar a plausibilidade da premissa inicial sobre o comportamento da variável aleatória; observe que 'não rejeitar' não significa 'aceitar' a hipótese.

Por tratar-se de uma inferência a respeito de uma variável aleatória, a decisão de não rejeitar (ou de rejeitar) uma hipótese, é tomada com base em uma certa probabilidade ou *nível de significância* α. Pode-se, por exemplo, não rejeitar a hipótese de que houve um decréscimo *significativo* da vazão média dos últimos trinta anos, em uma certa seção fluvial. Contrariamente, a eventual variação da vazão média do período, pode ser uma mera decorrência das flutuações amostrais, sem conseqüências para a média populacional em questão; nesse caso, a variação é dita *não significativa*. A especificação prévia de um nível de significância α, cumpre o papel de remover o grau de subjetividade associado à tomada de decisão intrínseca a um teste de hipótese. De fato, para um mesmo nível de significância, dois analistas diferentes, ao realizarem o teste de uma certa hipótese, sob condições idênticas, tomariam uma única e igual decisão. O nível de significância α de um teste de hipótese é complementar à probabilidade $(1-\alpha)$ com que um certo intervalo de confiança $[I, S]$ contém o valor populacional de um parâmetro θ. De fato, o intervalo $[I, S]$ estabelece os limites de variação da chamada *estatística de teste*, dentro dos quais a hipótese sobre θ não pode ser rejeitada. Contrariamente, se os valores da estatística de teste localizarem-se fora dos limites impostos por $[I, S]$, a hipótese sobre θ deve ser rejeitada, a um nível de significância α. Portanto, segundo essa interpretação, a construção de intervalos de confiança representa a operação inversa à de testar uma certa hipótese sobre um parâmetro populacional θ.

Em essência, testar uma hipótese é recolher evidências nos dados amostrais, que justifiquem a rejeição ou a não rejeição de uma certa afirmação (i) sobre um parâmetro populacional ou (ii) sobre a forma de um modelo distributivo, tendo-se em conta as probabilidades de serem tomadas decisões incorretas. Os testes de hipóteses podem ser classificados em *paramétricos* ou *não paramétricos*. Eles são ditos paramétricos se os dados amostrais, por premissa, tiverem sido extraídos de uma população Normal ou de qualquer outra população, cujo modelo distributivo é conhecido ou previamente especificado. Ao contrário, os testes não paramétricos não necessitam da especificação prévia do modelo distributivo da população, da qual foram extraídos os dados amostrais. De fato, em geral, os testes não paramétricos não são formulados com base nas observações amostrais, propriamente ditas, e, sim, em algumas de suas características ou atributos, tais como, ordens de classificação ou número de diferenças positivas ou negativas entre dados. Do ponto de vista da hipótese a ser testada, os testes mais freqüentes são aqueles que se referem a afirmações sobre um parâmetro populacional. Quando a hipótese, a ser testada, diz respeito à forma do modelo distributivo da população de onde a amostra foi extraída, os testes são denominados de aderência.

No presente capítulo, abordaremos, nos itens iniciais, as linhas gerais, segundo as quais os testes de hipóteses são construídos. Em seguida, ilustraremos esses procedimentos gerais, com a formulação dos testes de hipóteses paramétricos mais conhecidos, para populações normais. Na seqüência, descreveremos a lógica inerente aos testes não paramétricos, concentrando-nos naqueles de maior aplicação às variáveis hidrológicas. Nos itens finais, abordaremos os testes de aderência, enfatizando os testes do Qui-Quadrado, de Kolmogorov-Smirnov, de Anderson-Darling e de Filliben, bem como o teste de Grubbs e Beck, para a detecção de pontos amostrais atípicos, os quais são de grande utilidade na análise de freqüência de variáveis hidrológicas.

7.1 – Os Elementos de um Teste de Hipótese

Os procedimentos gerais para a realização de um teste de hipótese são:

• Formule a hipótese a ser testada, denotando-a por H_0 e denominando-a hipótese nula. Essa pode ser, por exemplo, a declaração conjectural de que não houve, nos últimos trinta anos, uma alteração da vazão média anual μ_1, de uma certa seção fluvial, quando comparada à média μ_0, do período anterior. Se a hipótese nula é verdadeira, qualquer diferença entre as médias populacionais μ_1 e μ_0 é devida meramente a flutuações das amostras extraídas de uma única população. A hipótese nula é expressa por H_0: $\mu_1 - \mu_0 = 0$.

• Formule a hipótese alternativa e denote-a por H_1. De acordo com o exemplo da etapa anterior, a hipótese alternativa, e contrária a H_0, é expressa por H_1: $\mu_1 - \mu_0 \neq 0$.

• Especifique uma estatística de teste *T*, que esteja em acordo com as hipóteses nula e alternativa, anteriormente formuladas. No exemplo em foco, a estatística de teste deve ter como base a diferença $T = \overline{X}_1 - \overline{X}_0$, entre as médias observadas nos períodos correspondentes às médias populacionais a serem testadas.

• Especifique a distribuição de amostragem da estatística de teste, de acordo com a hipótese nula, bem como com a distribuição de probabilidades da população de onde as observações foram extraídas. No exemplo em foco, caso as vazões médias anuais tenham sido extraídas de uma população Normal, sabe-se que é possível deduzir a distribuição de amostragem da estatística de teste *T*.

• Especifique a região de rejeição R, ou região crítica R, para a estatística de teste. A especificação da região crítica depende da definição prévia do nível de significância α, o qual, conforme menção anterior, cumpre o papel de remover o grau de subjetividade associado à tomada de decisão. No exemplo em foco, o nível de significância 100α poderia ser arbitrado, por exemplo, em 5 %, o que resultaria na fixação dos limites $[T_{0,025}, T_{0,975}]$, respectivamente abaixo e acima dos quais inicia-se a região de rejeição R.

• Verifique se a estatística de teste $\hat{T}$, estimada a partir das observações amostrais, está dentro ou fora dos limites estabelecidos para a região de rejeição R. No exemplo, se $\hat{T} < T_{0,025}$, ou se $\hat{T} > T_{0,975}$, a hipótese nula H_0 deve ser rejeitada; nesse caso, interpreta-se que a diferença $\mu_1 - \mu_0$ é significativa, a um nível $\alpha = 0{,}05$. Caso contrário, se $\hat{T}$ estiver dentro dos limites $[T_{0,025}, T_{0,975}]$, a decisão é a de não rejeitar a hipótese H_0, implicando que não há diferença significativa entre as médias populacionais μ_1 e μ_0.

Nos procedimentos gerais, anteriormente delineados, o exemplo citado refere-se diferenças positivas ou negativas entre μ_1 e μ_0, o que implica que a região crítica R estende-se pelas duas caudas da distribuição de amostragem da estatística de teste *T*. Nesse caso, diz-se que o teste é bilateral. Se a hipótese nula tivesse sido formulada de modo diferente, tal como H_0: $\mu_1 > \mu_0$ ou H_0: $\mu_1 < \mu_0$, o teste seria unilateral porque a região crítica se estenderia por apenas uma das caudas da distribuição de amostragem da estatística de teste.

Depreende-se, dos procedimentos gerais, que há uma relação estreita entre os testes de hipóteses e a construção de intervalos de confiança. Para melhor esclarecer esse fato, considere a hipótese nula H_0: $\mu = \mu_0$, a respeito da média de uma população Normal de variância conhecida e igual a σ^2. Sob essas circunstâncias, sabe-se que, para uma amostra de tamanho *N*, a estatística de teste é

$T = \left(\overline{X} - \mu_0\right) / \sigma / \sqrt{N}$ e que a distribuição de probabilidades dessa estatística de teste é a Normal padrão. Nesse caso, se fixarmos o nível de significância em $\alpha = 0{,}05$, o teste bilateral estaria definido para a região crítica abaixo de $T_{\alpha/2=0,025} = z_{0,025} = -1{,}96$ e acima de $T_{1-\alpha/2=0,975} = z_{0,975} = +1{,}96$. Se, a esse nível de significância, H_0 não foi rejeitada, verifica-se que tal decisão teve como argumento os fatos que $\hat{T} > T_{0,025}$ ou $\hat{T} < T_{0,975}$, os quais são equivalentes a $\overline{X} > \mu_0 - 1{,}96\sigma/\sqrt{N}$ ou $\overline{X} < \mu_0 + 1{,}96\sigma/\sqrt{N}$. Manipulando essas desigualdades, é possível colocá-las sob a forma $\overline{X} - 1{,}96\sigma/\sqrt{N} < \mu_0 < \overline{X} + 1{,}96\sigma/\sqrt{N}$, a qual é a expressão do intervalo a $100\,(1-\alpha) = 95\%$ de confiança para a média μ_0. Por meio desse exemplo, verifica-se a estreita ligação, no sentido matemático, entre a construção de intervalos de confiança e os testes de hipóteses. A despeito dessa ligação, entretanto, as duas técnicas servem a propósitos diferentes: enquanto o intervalo de confiança estabelece o quão acurado é o conhecimento de μ, o teste de hipótese indica se é plausível assumir o valor μ_0 para μ.

De acordo com o exposto, a rejeição da hipótese nula acontece quando a estimativa da estatística de teste encontrar-se dentro da região crítica. A decisão de rejeitar a hipótese nula é o mesmo que declarar que a estatística de teste é estatisticamente significativa. Em outros termos, no contexto de H_0: $\mu = \mu_0$ e de $\alpha = 0{,}05$, se as diferenças observadas ocorrem, de modo aleatório, em menos de 5 de 100 testes idênticos, então, os resultados são considerados estatisticamente significativos e a hipótese nula deve ser rejeitada. Por outro lado, a falta de evidências empíricas para rejeitar a hipótese nula, não implica na imediata aceitação de H_0 e, sim, em sua eventual reformulação, seguida de verificações suplementares.

Supondo que a hipótese nula é, de fato, *verdadeira*, a probabilidade de que H_0 seja rejeitada é dada por

$$\mathrm{P}\left(T \in R \middle| H_0 \text{ verdadeira}\right) = \mathrm{P}\left(T \in R \middle| H_0\right) = \alpha \qquad (7.1)$$

É evidente que se uma hipótese verdadeira é rejeitada, tomou-se uma decisão incorreta. O erro decorrente dessa decisão é denominado *erro do tipo I*. Da equação 7.1, resulta que a probabilidade de ocorrer o erro do tipo I é expressa por

$$\mathrm{P}\left(\text{Erro do Tipo I}\right) = \mathrm{P}\left(T \in R \middle| H_0\right) = \alpha \qquad (7.2)$$

Na ausência de erro, ou seja, se uma hipótese verdadeira H_0 não é rejeitada, a probabilidade dessa decisão é complementar à probabilidade do erro do tipo I. Em termos formais,

$$P(T \notin R | H_0) = 1 - \alpha \tag{7.3}$$

Contrariamente, não rejeitar a hipótese nula quando ela é, de fato, *falsa*, é outra possível decisão incorreta. O erro decorrente dessa decisão é denominado *erro do tipo II*. A probabilidade de ocorrer o erro do tipo II é expressa por

$$P(\text{Erro do Tipo II}) = P(T \notin R | H_1) = \beta \tag{7.4}$$

Na ausência de erro, ou seja, se uma hipótese falsa H_0 é rejeitada, a probabilidade dessa decisão é complementar à probabilidade do erro do tipo II. Em termos formais,

$$P(T \in R | H_1) = 1 - \beta \tag{7.5}$$

A probabilidade complementar a β, expressa pela equação 7.5, é denominada *poder do teste* e, como se verá mais adiante, é um importante critério de comparação entre diferentes testes de hipóteses.

Os erros dos tipos I e II estão fortemente relacionados. Para demonstrar essa relação, considere que a Figura 7.1 ilustra um teste unilateral de uma hipótese nula $H_0: \mu = \mu_0$ contra a hipótese alternativa $H_1: \mu = \mu_1$, onde μ representa a média de uma população Normal e $\mu_1 > \mu_0$.

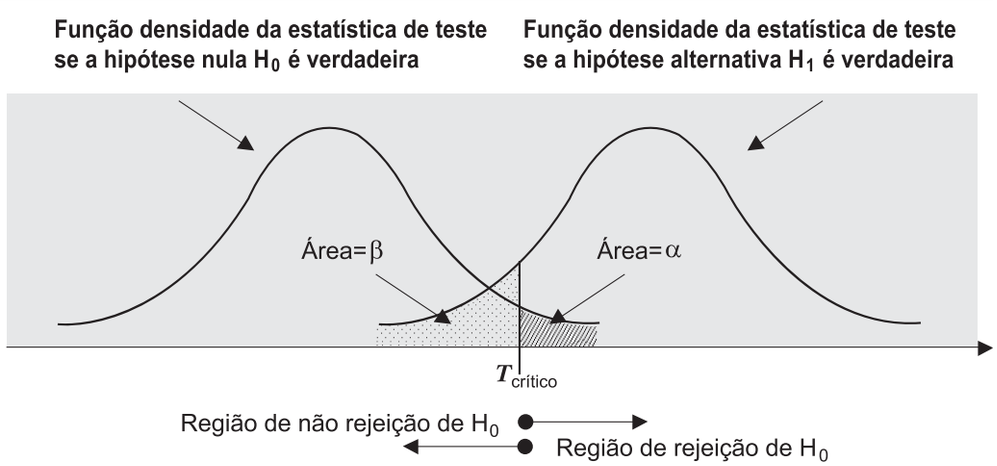


Figura 7.1 – Ilustração dos erros dos tipos I e II em um teste de hipótese unilateral

Se a estatística de teste T é superior ao valor $T_{crítico}$, a hipótese nula é rejeitada, a um nível de significância α. Nesse caso, supondo que H_0 é verdadeira, a decisão de rejeitá-la é incorreta e a probabilidade de se cometer esse erro é α. Contrariamente, se a estatística de teste T é inferior ao valor $T_{crítico}$, a hipótese

nula não é rejeitada, a um nível de significância α. Nesse outro caso e supondo que, desta feita, H_1 é verdadeira, a decisão de não rejeitar a hipótese falsa H_0 também é incorreta e a probabilidade de se cometer esse erro é β. Pela ilustração da Figura 7.1, é evidente que a diminuição de α irá levar o valor de $T_{crítico}$ mais para a direita, causando um aumento de β. Conclui-se, portanto, que diminuir a probabilidade de se cometer um erro do tipo I provoca o aumento da probabilidade de se cometer o erro do tipo II. A situação inversa é igualmente verdadeira.

É óbvio que, ao realizar um teste de hipótese, não se quer tomar uma decisão incorreta e que, portanto, a situação desejável é a de minimizar as probabilidades de se cometer erros de ambos os tipos. Entretanto, em função da dependência entre α e β, ilustrada pela Figura 7.1, bem como das diferentes características dos erros dos tipos I e II, é forçosa uma solução de compromisso no planejamento das regras de decisão de um teste de hipóteses. Em geral, essa solução de compromisso passa pela prescrição prévia de um determinado nível de significância α, tal que ele seja suficientemente pequeno para que β encontre-se em uma faixa aceitável de variação. Essa estratégia de ação advém do fato que, em geral, é possível prescrever antecipadamente o nível α , enquanto tal possibilidade não existe para a probabilidade β. Essa afirmação é justificada pela constatação de que a hipótese alternativa é mais genérica do que a hipótese nula; por exemplo, a hipótese alternativa $H_1: \mu_1 - \mu_0 \neq 0$ compreende a união de diversas outras hipóteses alternativas (e.g.: $H_1: \mu_1 - \mu_0 < 0$ ou $H_1: \mu_1 - \mu_0 > 0$), enquanto a hipótese nula $H_0: \mu_1 - \mu_0 = 0$ é completamente definida. Em outras palavras, enquanto α depende apenas da hipótese nula, β irá depender de qual das hipóteses alternativas é de fato verdadeira, o que, evidentemente, não se sabe *a priori*. Na prática, é considerado razoável prescrever, antecipadamente, o nível de significância α em 0,05, o que implica em uma média de 5 rejeições incorretas de H_0, em 100 decisões possíveis. Se as conseqüências de um erro do tipo I forem muito graves, pode-se escolher um nível de significância ainda menor, como $\alpha = 0{,}01$ ou $\alpha = 0{,}001$.

Embora β dependa de qual hipótese alternativa H_1 é, de fato, verdadeira e, portanto, não possa ser antecipadamente prescrito, é útil o estudo do comportamento de β, sob diferentes possibilidades para H_1. Essa investigação é feita por meio da quantidade 1-β, a qual, conforme menção anterior, é denominada poder do teste. Na Figura 7.1, o poder do teste, para a hipótese alternativa específica $H_1: \mu = \mu_1$, pode ser visualizado pela área da função densidade da estatística de teste, sob H_1, à direita da abscissa $T_{crítico}$. Para outra hipótese alternativa, por exemplo $H_1: \mu = \mu_2$, é claro que o poder do teste teria outro valor. As relações entre β, ou (1-β), e uma seqüência contínua de hipóteses alternativas específicas, definem, respectivamente, a *curva característica operacional*, ou a *função poder de teste*, as quais permitem distinguir e comparar testes diferentes.

Para exemplificar a construção da curva característica operacional e da função poder do teste, considere o seguinte teste bilateral da média de uma população Normal de parâmetros μ e α: H_0: $\mu = \mu_0$ contra o conjunto de hipóteses alternativas H_1: $\mu \neq \mu_0$. Mais uma vez, a estatística de teste, nesse caso, é $T = \left(\overline{X} - \mu_0\right) / \sigma / \sqrt{N}$, a qual segue uma distribuição N(0,1). O numerador da estatística de teste pode ser alterado para expressar deslocamentos $\mu_0 + k$, em relação a μ_0, onde k denota uma constante positiva ou negativa. Desse modo, com $T = k\sqrt{N}/\sigma$, o teste refere-se a H_0: $\mu = \mu_0$ contra um conjunto de deslocamentos padronizados $k\sqrt{N}/\sigma$, em relação a zero, ou, equivalentemente, contra um conjunto de deslocamentos $\mu_0 + k$, em relação a μ_0. O erro do tipo II corresponde a não rejeitar H_0, quando H_1 é verdadeira, o que irá acontecer quando a estatística de teste T satisfizer $-z_{\alpha/2} \leq T \leq +z_{\alpha/2}$, onde $-z_{\alpha/2}$ e $z_{\alpha/2}$ representam os limites de definição da região crítica. A probabilidade β de se cometer o erro do tipo II pode ser escrita como $\beta = \Phi\left(z_{\alpha/2} - k\sqrt{N}/\sigma\right) - \Phi\left(-z_{\alpha/2} - k\sqrt{N}/\sigma\right)$, onde $\Phi(.)$ denota a FAP da distribuição Normal padrão. Portanto, percebe-se que β depende de α, de N e das diferentes hipóteses alternativas dadas por k/σ. Essa dependência pode ser expressa graficamente por meio da curva característica operacional, ilustrada na Figura 7.2, para $\alpha = 0{,}10$ ($z_{0{,}05} = 1{,}645$), amostras de tamanho N variável entre 1 e 50, e $k/\sigma = 0{,}25$, 0,50, 0,75 e 1.

O exame da curva característica operacional mostra que, para uma amostra de tamanho N fixo, a probabilidade de se cometer o erro do tipo II decresce, quando

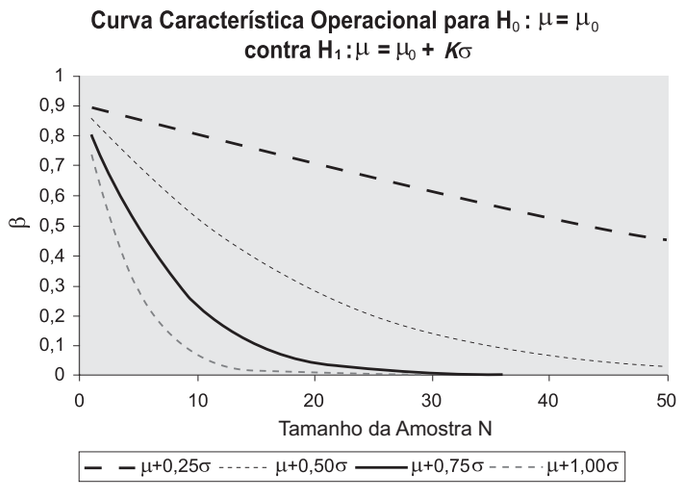


Figura 7.2 – Exemplos da curva característica operacional de um teste de hipóteses

k/σ aumenta. Isso equivale a dizer que pequenas diferenças na média são mais difíceis de detectar, o que conduz a maiores probabilidades de se tomar a decisão incorreta de não rejeitar uma falsa hipótese nula. Observa-se também um decréscimo de β, com o aumento de *N*, demonstrando a menor probabilidade de se cometer um erro do tipo II, quando o teste tem, como base, amostras de maior tamanho.

A função poder de teste é dada pelo complemento de β, em relação a 1, e encontra-se ilustrada na Figura 7.3, para o exemplo em foco. O poder do teste, conforme definição anterior, representa a probabilidade de se tomar a decisão correta de rejeitar uma falsa hipótese nula, em favor de uma hipótese alternativa. A Figura 7.3 mostra que, para amostras de mesmo tamanho, a probabilidade de *não* se cometer o erro do tipo II cresce, quando k/σ aumenta. Do mesmo modo, o poder do teste aumenta quando o tamanho da amostra cresce.

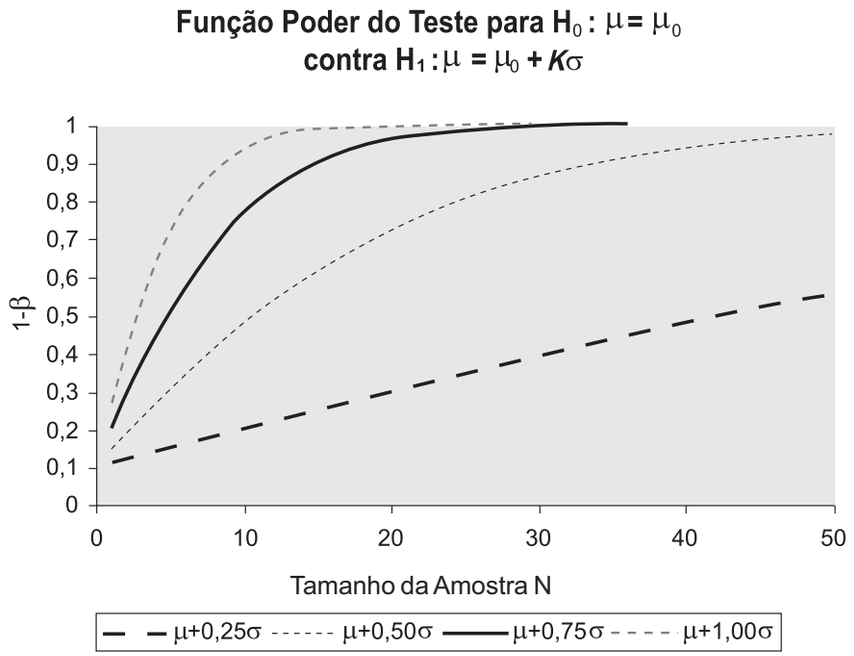


Figura 7.3 - Exemplos de função poder de um teste de hipóteses

As Figuras 7.2 e 7.3 mostram que se, por exemplo, desejarmos manter simultaneamente as respectivas probabilidades de se cometer os erros dos tipos I e II em $100\alpha = 10\%$ e $100\beta = 20\%$, e se estivermos testando a hipótese nula H_0: $\mu = \mu_0$, contra a hipótese alternativa H_1: $\mu = \mu_0 + 0{,}5\sigma$, necessitaríamos de uma amostra de tamanho pelo menos igual a 26. Para esse exemplo, se uma amostra

de tamanho 26 não estiver disponível ou se a obtenção de observações adicionais for excessivamente onerosa, o analista deve buscar alguma solução de compromisso entre a confiabilidade do teste, imposta pela especificação de α e β, e a disponibilidade e/ou ônus de amostragem suplementar. No restante desse capítulo, nos restringiremos a testes de hipóteses que têm como base uma amostra de tamanho fixo, sob a especificação prévia de um nível de significância usual, digamos $100\alpha = 5\%$ ou 10%, aceitando implicitamente os níveis de β decorrentes dessa decisão.

7.2 – Alguns Testes Paramétricos Usuais para Populações Normais

Grande parte da construção matemática em torno dos testes paramétricos de hipóteses refere-se a populações normais. Essa constatação pode ser explicada, primeiramente, pela possibilidade de dedução das distribuições de amostragem de variáveis normais, e, em segundo lugar, pela ampla extensão de aplicações do teorema do limite central. O que se apresenta a seguir é uma descrição dos principais testes paramétricos para populações normais, incluindo as premissas e as estatísticas de teste, sob as quais são construídos. Para que tais testes produzam resultados rigorosos, as premissas devem ser rigorosamente observadas. Em alguns casos práticos e como decorrência do teorema do limite central, pode-se cogitar a extensão desses testes paramétricos para populações não-normais. Deve-se ressalvar, entretanto, que os resultados dessa extensão serão apenas aproximados. Em geral, o grau de aproximação, nesses casos, é dado pela diferença entre o verdadeiro nível de significância do teste, o qual, pode ser avaliado por meio de simulações de Monte Carlo, e o nível previamente estabelecido.

7.2.1 – Testes Paramétricos sobre a Média de uma Única População Normal

A premissa básica dos testes, descritos a seguir, é a de que as variáveis aleatórias independentes $\{X_1, X_2, \ldots, X_N\}$, componentes de uma certa amostra aleatória simples, foram todas extraídas de uma única população normal, de média μ desconhecida. O conhecimento ou o desconhecimento da variância populacional σ^2 determina a estatística de teste a ser usada.

- H_0: $\mu = \mu_1$ contra H_1: $\mu = \mu_2$. Atributo de σ^2: conhecida.

Estatística de teste: $Z = \dfrac{\overline{X} - \mu_1}{\sigma / \sqrt{N}}$

Distribuição de probabilidades da estatística de teste: Normal N(0,1)

Tipo de Teste: unilateral a um nível de significância α
Decisão:

Se $\mu_1 > \mu_2$, rejeitar H_0 se $\dfrac{\overline{X} - \mu_1}{\sigma/\sqrt{N}} < -z_{1-\alpha}$

Se $\mu_1 < \mu_2$, rejeitar H_0 se $\dfrac{\overline{X} - \mu_1}{\sigma/\sqrt{N}} > +z_{1-\alpha}$

• H_0: $\mu = \mu_1$ contra H_1: $\mu = \mu_2$. Atributo de σ^2: desconhecida e estimada por s_X^2.

Estatística de teste: $T = \dfrac{\overline{X} - \mu_1}{s_X/\sqrt{N}}$

Distribuição de probabilidades da estatística de teste: t de Student com $\nu = N\text{-}1$ ou t_{N-1}
Tipo de Teste: unilateral a um nível de significância α
Decisão:

Se $\mu_1 > \mu_2$, rejeitar H_0 se $\dfrac{\overline{X} - \mu_1}{\sigma/\sqrt{N}} < -t_{1-\alpha,\nu=N-1}$

Se $\mu_1 < \mu_2$, rejeitar H_0 se $\dfrac{\overline{X} - \mu_1}{\sigma/\sqrt{N}} > +t_{1-\alpha,\nu=N-1}$

• H_0: $\mu = \mu_0$ contra H_1: $\mu \neq \mu_0$. Atributo de σ^2: conhecida.

Estatística de teste: $Z = \dfrac{\overline{X} - \mu_0}{\sigma/\sqrt{N}}$

Distribuição de probabilidades da estatística de teste: Normal N(0,1)
Tipo de Teste: bilateral a um nível de significância α
Decisão:

Rejeitar H_0 se $\left|\dfrac{\overline{X} - \mu_0}{\sigma/\sqrt{N}}\right| > z_{1-\alpha/2}$

• H_0: $\mu = \mu_0$ contra H_1: $\mu \neq \mu_0$. Atributo de σ^2: desconhecida e estimada por s_X^2.

Estatística de teste: $T = \dfrac{\overline{X} - \mu_0}{s_X/\sqrt{N}}$

Distribuição de probabilidades da estatística de teste: *t* de Student com $\nu = N\text{-}1$ ou t_{N-1}
Tipo de Teste: bilateral a um nível de significância α
Decisão:

Rejeitar H_0 se $\left|\frac{\overline{X} - \mu_0}{s_X / \sqrt{N}}\right| > t_{1-\alpha/2,\nu=N-1}$

Exemplo 7.1 – Considere as vazões médias do mês de Julho do Rio Paraopeba em Ponte Nova do Paraopeba, listadas no Anexo 1, para o período de 1938 a 1999. Teste a hipótese de que a média populacional do mês de Julho é 47,65 m^3/s, a um nível de significância $100\alpha = 5\%$.
Solução: A premissa básica é a que as vazões médias do mês de Julho, em Ponte Nova do Paraopeba, seguem uma distribuição Normal. A amostra de 62 observações fornece $\overline{X} = 44{,}526$ e $s_X = 12{,}406\ m^3/s$, não havendo nenhuma informação adicional sobre a variância populacional. Nesse caso, a hipótese nula é H_0: $\mu = 47{,}65$ contra a hipótese alternativa H_1: $\mu \neq 47{,}65$. Trata-se, portanto, de um teste bilateral ao nível $100\ \alpha = 5\%$, com a estatística de teste dada por $T = \frac{\overline{X} - 47{,}65}{s_X / \sqrt{N}}$, a qual possui uma distribuição *t* de Student com 61 graus de liberdade. Substituindo os valores amostrais, resulta que o valor absoluto da estimativa de *T* é igual a 1,9828. A tabela de *t* de Student, do Anexo 7, fornece $t_{0,975,\nu=61} = 1{,}9996$. Como 1,9828 < 1,9996, a hipótese H_0 não deve ser rejeitada, em favor de H_1. Em outras palavras, com base na amostra disponível, não há evidências de que a média populacional difira significativamente de 47,65 m^3/s, ou seja, que a diferença existente entre a média amostral $\overline{X} = 44{,}526$ e a média hipotética $\mu = 47{,}65$ deve-se unicamente a flutuações aleatórias das observações.

Exemplo 7.2 – Repita o exemplo 7.1, supondo que a variância populacional σ^2 seja conhecida e igual a 153,9183 $(m^3/s)^2$.
Solução: A premissa básica continua sendo a de que as vazões médias do mês de Julho, em Ponte Nova do Paraopeba, seguem uma distribuição Normal. O fato de que a variância populacional é conhecida altera a estatística de teste. Nesse caso, trata-se de um teste bilateral ao nível $100\ \alpha = 5\%$, com a estatística de teste dada por $Z = \frac{\overline{X} - 47{,}65}{\sigma / \sqrt{N}}$, a qual possui uma distribuição N(0,1). Substituindo os valores amostrais, resulta que o valor absoluto da estimativa de Z é igual a 1,9828. A tabela 5.1, do capítulo 5, fornece $z_{0,975} = 1{,}96$.

Como 1,9828 >1,96, a hipótese H_0 deve ser rejeitada, em favor de H_1. Portanto, sob as condições estipuladas para esse caso, é significativa a diferença entre a média amostral $\overline{X} = 44{,}526$ e a média hipotética $\mu = 47{,}65$.

7.2.2 – Testes Paramétricos sobre as Médias de Duas Populações Normais

A premissa básica dos testes, descritos a seguir, é a de que as variáveis aleatórias independentes $\{X_1, X_2, \ldots, X_N\}$ e $\{Y_1, Y_2, \ldots, Y_M\}$, componentes de duas amostras aleatórias simples de tamanhos iguais a N e M, foram extraídas de duas populações normais, de respectivas médias μ_X e μ_Y desconhecidas. O conhecimento ou o desconhecimento das variâncias populacionais σ_X^2 e σ_Y^2, assim como a condição de igualdade entre elas, determinam a estatística de teste a ser usada. Os testes descritos a seguir são tomados como bilaterais, podendo ser transformados em unilaterais pela modificação de H_1 e de α.

• $H_0: \mu_X - \mu_Y = \delta$ contra $H_1: \mu_X - \mu_Y \neq \delta$.
Atributos de σ_X^2 e σ_Y^2: conhecidas

Estatística de teste: $Z = \dfrac{(\overline{X} - \overline{Y}) - \delta}{\sqrt{\dfrac{\sigma_X^2}{N} + \dfrac{\sigma_Y^2}{M}}}$

Distribuição de probabilidades da estatística de teste: Normal N(0,1)
Tipo de Teste: bilateral a um nível de significância α
Decisão:

Rejeitar H_0 se $\left| \dfrac{(\overline{X} - \overline{Y}) - \delta}{\sqrt{\dfrac{\sigma_X^2}{N} + \dfrac{\sigma_Y^2}{M}}} \right| > z_{1-\frac{\alpha}{2}}$

• $H_0: \mu_X - \mu_Y = \delta$ contra $H_1: \mu_X - \mu_Y \neq \delta$.
Atributos de σ_X^2 e σ_Y^2: supostamente iguais, mas desconhecidas.
Estimadas por s_X^2 e s_Y^2.

Estatística de teste: $T = \dfrac{(\overline{X} - \overline{Y}) - \delta}{\sqrt{(N-1)s_X^2 + (M-1)s_Y^2}} \sqrt{\dfrac{NM(N+M-2)}{N+M}}$

Distribuição de probabilidades da estatística de teste: *t* de Student com $\nu = N + M - 2$

Tipo de Teste: bilateral a um nível de significância α
Decisão:
Rejeitar H_0 se

$$\left| \frac{(\overline{X}-\overline{Y})-\delta}{\sqrt{(N-1)s_X^2+(M-1)s_Y^2}} \sqrt{\frac{NM(N+M-2)}{N+M}} \right| > t_{1-\frac{\alpha}{2},\nu=N+M-2}$$

• $H_0: \mu_X - \mu_Y = \delta$ contra $H_1: \mu_X - \mu_Y \neq \delta$.
Atributos de σ_X^2 e σ_Y^2: supostas desiguais, mas desconhecidas.
Estimadas por s_X^2 e s_Y^2.

Estatística de teste: $T = \dfrac{(\overline{X}-\overline{Y})-\delta}{\sqrt{(s_X^2/N)+(s_Y^2/M)}}$

Distribuição de probabilidades da estatística de teste: segundo Casella e Berger (1990), a distribuição de *T* pode ser *aproximada* por uma distribuição *t* de Student com $\nu = \dfrac{[(s_X^2/N)+(s_Y^2/M)]^2}{\left[\dfrac{(s_X^2/N)^2}{N-1}+\dfrac{(s_Y^2/M)^2}{M-1}\right]}$

Tipo de Teste: bilateral a um nível de significância α
Decisão:

Rejeitar H_0 se $\left| \dfrac{(\overline{X}-\overline{Y})-\delta}{\sqrt{(s_X^2/N)+(s_Y^2/M)}} \right| > t_{1-\frac{\alpha}{2},\nu}$

Exemplo 7.3 – Considere as vazões médias do mês de Julho do Rio Paraopeba em Ponte Nova do Paraopeba, listadas no Anexo 1, separando-as em duas amostras iguais de mesmo tamanho: a amostra denotada por *X*, para o período de 1938 a 1968, e a amostra *Y*, para o período de 1969 a 1999. Teste a hipótese de que, considerados os períodos de 1938-1968 e de 1969-1999, as médias populacionais do mês de Julho não sofreram alterações importantes, a um nível de significância $100\alpha = 5\%$.
Solução: A premissa básica é a que, considerados os períodos de 1938-1968 e de 1969-1999, as vazões médias do mês de Julho, em Ponte Nova do Paraopeba, seguem duas distribuições normais de médias μ_X e μ_Y, com variâncias σ_X e σ_Y supostamente desiguais e desconhecidas. A amostra de 31 observações, para o período de 1938 a 1968, fornece $\overline{X} = 45{,}08$ e $s_X = 11{,}505\ m^3/s$, enquanto, para o período restante, esses valores resultam ser $\overline{Y} = 43{,}97$ e $s_y = 13{,}415\ m^3/s$. Nesse caso, a

hipótese nula é H_0: $\mu_X - \mu_Y = \delta = 0$ contra a hipótese alternativa H_1: $\mu_X - \mu_Y = \delta \neq 0$. Como as variâncias são supostamente desiguais e devem ser estimadas pelas variâncias amostrais, a estatística de teste é

$T = \dfrac{(\overline{X} - \overline{Y})}{\sqrt{((s_X^2/31)) + (s_Y^2/31)}}$, a distribuição de probabilidades da qual pode ser aproximada por uma *t* de Student com $\nu = \dfrac{[(s_X^2/31) + (s_Y^2/31)]^2}{\left[\dfrac{(s_X^2/31)^2}{30} + \dfrac{(s_Y^2/31)^2}{30}\right]} = 58$ graus de liberdade. Substituindo os valores amostrais, resulta que o valor absoluto da estimativa de *T* é igual a 0,3476. A tabela de *t* de Student, do Anexo 7, fornece $t_{0,975,\nu=58} = 2{,}00$. Como 0,3476 < 2,00, a hipótese H_0 não deve ser rejeitada, em favor de H_1. Em outras palavras, com base nas amostras disponíveis, não há evidências de que as médias populacionais, dos períodos considerados, difiram significativamente entre si, ao nível de $100\alpha = 5\%$.

7.2.3 – Testes Paramétricos sobre a Variância de uma Única População Normal

A premissa básica dos testes, descritos a seguir, é a de que as variáveis aleatórias independentes $\{X_1, X_2, \ldots, X_N\}$, componentes de uma certa amostra aleatória simples, foram todas extraídas de uma única população normal, de variância σ^2 desconhecida. O conhecimento ou o desconhecimento da média populacional μ determina a estatística de teste a ser usada. Os testes são tomados como bilaterais, podendo ser transformados em unilaterais pela modificação de H_1 e de α.

• H_0: $\sigma^2 = \sigma_0^2$ contra H_1: $\sigma^2 \neq \sigma_0^2$.
Atributo de μ: conhecida.

Estatística de teste: $Q = \dfrac{\sum_{i=1}^{N}(X_i - \mu)^2}{\sigma_0^2} = N\dfrac{s_x^2}{\sigma_0^2}$

Distribuição de probabilidades da estatística de teste: χ^2 com $\nu = N$, ou χ_N^2
Tipo de Teste: bilateral a um nível de significância α
Decisão:

Rejeitar H_0 se $N\dfrac{s_x^2}{\sigma_0^2} < \chi^2_{\frac{\alpha}{2},N}$ ou se $N\dfrac{s_x^2}{\sigma_0^2} > \chi^2_{1-\frac{\alpha}{2},N}$

• H_0: $\sigma^2 = \sigma_0^2$ contra H_1: $\sigma^2 \neq \sigma_0^2$.

Atributo de μ: desconhecida, estimada por $\overline{X}$.

Estatística de teste: $K = \dfrac{\sum_{i=1}^{N}(X_i - \overline{X})^2}{\sigma_0^2} = (N-1)\dfrac{s_x^2}{\sigma_0^2}$

Distribuição de probabilidades da estatística de teste: χ^2 com $\nu = N$ -1, ou χ^2_{N-1}

Tipo de Teste: bilateral a um nível de significância

Decisão:

Rejeitar H_0 se $(N-1)\dfrac{s_x^2}{\sigma_0^2} < \chi^2_{\frac{\alpha}{2},N-1}$ ou se $(N\text{-}1)\dfrac{s_x^2}{\sigma_0^2} > \chi^2_{1-\frac{\alpha}{2},N-1}$

Exemplo 7.4 – Considere novamente as vazões médias do mês de Julho do Rio Paraopeba em Ponte Nova do Paraopeba, listadas no Anexo 1, para o período de 1938 a 1999. Teste a hipótese nula de que a variância populacional σ_0^2, das vazões médias do mês de Julho, é de 150 $(m^3/s)^2$ contra a hipótese alternativa H_1: $\sigma_0^2 > 150$ $(m^3/s)^2$, a um nível de significância $100\alpha = 5\%$.

Solução: Novamente, a premissa básica é a que as vazões médias do mês de Julho, em Ponte Nova do Paraopeba, seguem uma distribuição Normal. A amostra de 62 observações fornece $\overline{X} = 44{,}526\ e\ s_X = 12{,}406\ m^3/s$, não havendo nenhuma informação adicional sobre a média populacional. Nesse caso, a hipótese nula é H_0: $\sigma_0^2 = 150$ contra a hipótese alternativa H_1: $\sigma_0^2 > 150$. Trata-se, portanto, de um teste unilateral ao nível $100\alpha = 5\%$, com a estatística de teste dada por $K = (N-1)\dfrac{s_x^2}{\sigma_0^2}$, a qual possui uma distribuição χ^2 com 61 graus de liberdade. Substituindo os valores amostrais, resulta que o valor de K é igual a 62,593. A tabela de χ^2, do Anexo 6, fornece $\chi^2_{0,95,61} = 80{,}232$. Como 62,593 < 80,232, a hipótese H_0 não deve ser rejeitada, em favor de H_1. Em outras palavras, com base na amostra disponível, não há evidências de que a variância populacional supere significativamente o valor de 150 $(m^3/s)^2$, ou seja, que a diferença existente entre a variância amostral $s_x^2 = 153{,}918$ e a variância $\sigma_0^2 = 150$ deve-se unicamente a flutuações aleatórias das observações.

7.2.4 – Testes Paramétricos sobre as Variâncias de Duas Populações Normais

A premissa básica dos testes, descritos a seguir, é a de que as variáveis aleatórias independentes $\{X_1, X_2, \ldots, X_N\}$ e $\{Y_1, Y_2, \ldots, Y_M\}$, componentes de duas amostras aleatórias simples de tamanhos iguais a N e M, foram extraídas de duas populações normais, de respectivas variâncias σ_X^2 e σ_Y^2 desconhecidas. O conhecimento ou o desconhecimento das médias populacionais μ_X e μ_Y determina a estatística de teste a ser usada. Os testes são tomados como bilaterais, podendo ser transformados em unilaterais pela modificação de H_1 e de α.

• $H_0: \dfrac{\sigma_X^2}{\sigma_Y^2} = 1$ contra $H_1: \dfrac{\sigma_X^2}{\sigma_Y^2} \neq 1$

Atributos de μ_X e μ_Y: conhecidas

Estatística de teste: $\varphi = \dfrac{s_X^2 / \sigma_X^2}{s_Y^2 / \sigma_Y^2}$

Distribuição de probabilidades da estatística de teste: F de Snedecor com $\nu_1 = N$ e $\nu_2 = M$, ou $F_{N,M}$

Tipo de Teste: bilateral a um nível de significância α

Decisão:

Rejeitar H_0 se $\varphi < F_{N,M,\alpha/2}$ ou se $\varphi > F_{N,M,1-\alpha/2}$

• $H_0: \dfrac{\sigma_X^2}{\sigma_Y^2} = 1$ contra $H_1: \dfrac{\sigma_X^2}{\sigma_Y^2} \neq 1$

Atributos de μ_X e μ_Y: desconhecidas, estimadas por $\overline{X}$ e $\overline{Y}$

Estatística de teste: $f = \dfrac{s_X^2 / \sigma_X^2}{s_Y^2 / \sigma_Y^2}$

Distribuição de probabilidades da estatística de teste: F de Snedecor com $\nu_1 = N - 1$ e $\nu_2 = M - 1$, ou $F_{N\text{-}1,M\text{-}1}$

Tipo de Teste: bilateral a um nível de significância α

Decisão:

Rejeitar H_0 se $f < F_{N\text{-}1,M\text{-}1,\alpha/2}$ ou se $f > F_{N\text{-}1,M\text{-}1,1\text{-}\alpha/2}$

Exemplo 7.5 – Um certo constituinte de um efluente foi analisado 7 e 9 vezes por meio dos procedimentos X e Y, respectivamente. Os resultados das análises apresentaram os seguintes desvios-padrão: $s_X = 1{,}9$ *e* $s_Y = 0{,}8$ mg/l. Teste a hipótese de que o procedimento *Y* é mais preciso do que o procedimento *X*, ao nível de significância $100\alpha = 5\%$. (adap. de Kottegoda e Rosso, 1997)

Solução: Supondo tratarem-se de duas populações normais, a hipótese nula a ser testada é $H_0: \frac{\sigma_X^2}{\sigma_Y^2} = 1$ contra a hipótese alternativa $H_1: \frac{\sigma_X^2}{\sigma_Y^2} > 1$ ou $\sigma_X^2 > \sigma_Y^2$. Trata-se, portanto, de um teste unilateral com $\alpha = 0{,}05$. A estatística de teste é $f = \frac{s_X^2 / \sigma_X^2}{s_Y^2 / \sigma_Y^2}$, a qual segue uma distribuição F de Snedecor com v = 7-1 = 6 e v_2 = 9-1 = 8 graus de liberdade para o numerador e denominador, respectivamente. Substituindo os valores amostrais, resulta que f=5,64. Da tabela de F, do Anexo 8, lê-se que $F_{6,8,0,05}$ = 3,58. Como 5,64 > 3,58, a decisão é de rejeitar a hipótese nula em favor da hipótese alternativa, ao nível de significância $\alpha = 0{,}05$. Em outras palavras, conclui-se que a variância dos resultados do procedimento Y é menor do que a de seu concorrente, tratando-se, portanto, de um método de análise mais preciso.

7.3 – Alguns Testes Não-Paramétricos Usuais em Hidrologia

Os testes paramétricos de hipóteses, anteriormente descritos, requerem que a distribuição da variável aleatória, ou das variáveis aleatórias de origem, seja a distribuição Normal. De fato, se a distribuição dos dados originais é a Normal, é possível deduzir as distribuições das estatísticas de testes, em razão, principalmente, da propriedade reprodutiva das variáveis Gaussianas e do teorema do limite central. Entretanto, se a distribuição dos dados originais não é Gaussiana, o uso das distribuições das estatísticas de testes conhecidas terá como conseqüência, a violação do nível de significância previamente estabelecido. Por exemplo, se T denota a estatística de teste $T = \left(\overline{X} - \mu_0\right) / s_X / \sqrt{N}$ para uma variável aleatória X, cujo comportamento se afasta da normalidade, a verdadeira probabilidade de se cometer o erro do tipo I não será necessariamente igual ao nível nominal α. Em outros termos, nesse caso, pode-se escrever que

$$\int_{-\infty}^{-t_{\alpha/2}} f_T\left(t|H_0\right)dt + \int_{t_{\alpha/2}}^{\infty} f_T\left(t|H_0\right)dt \neq \alpha \qquad (7.6)$$

onde f_T(t) é a função densidade desconhecida de $T = \left(\overline{X} - \mu_0\right) / s_X / \sqrt{N}$, sob a premissa que X não seja normalmente distribuída.

A estatística matemática apresenta duas soluções possíveis para o problema identificado pela equação 7.6. A primeira solução diz respeito à tentativa de mostrar, via simulações de Monte Carlo, que, mesmo que uma certa variável aleatória X não seja Gaussiana e que, portanto, a estatística de teste T não tenha uma distribuição de probabilidades conhecida, a verdadeira densidade $f_T\left(t|H_0\right)$ irá se comportar, em muitos casos, de modo suficientemente similar à distribuição usual, caso X fosse, de fato, normal. Por exemplo, Larsen e Marx (1986) mostram alguns exemplos nos quais, se a distribuição de X não é exageradamente assimétrica ou se o tamanho da amostra não é excessivamente pequeno, a distribuição t de Student pode aproximar satisfatoriamente a distribuição $f_T\left(t|H_0\right)$, para testes de hipóteses relativas à média populacional de X. Nesses casos, afirma-se que o teste de t de Student é *robusto*, em relação a desvios moderados da normalidade. Dadas as características marcadamente assimétricas das distribuições de probabilidades de grande parte das variáveis hidrológicas, essa primeira possível solução, para o problema identificado pela equação 7.6, tem aplicações muito limitadas na hidrologia estatística.

A segunda solução possível, para o problema posto pela equação 7.6, é a de substituir as estatísticas de teste convencionais por outras, cujas distribuições de probabilidades permanecem invariáveis, sob a veracidade da hipótese H_0 e a despeito das características distributivas populacionais da variável aleatória de origem X. Os procedimentos de inferência estatística e, particularmente, os testes de hipóteses, que possuem tais propriedades, são denominados *não-paramétricos*. Os procedimentos gerais para a construção de testes paramétricos de hipóteses, alinhavados no item 7.1, permanecem os mesmos para os testes não-paramétricos. A diferença fundamental entre eles é que os testes não-paramétricos são formulados com base em estatísticas invariáveis com a distribuição dos dados originais. De fato, as estatísticas de testes não-paramétricos, em sua grande maioria, baseiam-se em características que podem ser deduzidas dos dados amostrais, mas que não os incluem diretamente em seu cálculo. São exemplos dessas características: o número de diferenças positivas (ou negativas) entre uma certa mediana hipotética e os valores amostrais, ou o coeficiente de correlação entre as ordens de classificação dos elementos de duas amostras, ou, ainda, o número de inflexões dos valores amostrais ao longo de uma seqüência de índices de tempo, entre outras.

A variedade e o número de testes não-paramétricos têm crescido enormemente desde que foram introduzidos, no início da década de 1940. Não se tem aqui o objetivo de abordar a formulação e a construção dos inúmeros testes não-paramétricos de hipóteses; o leitor interessado nesses detalhes deve remeter-se a textos especializados, tais como Siegel (1956), Gibbons (1971) e Hollander e

Wolfe (1973). O que se segue é uma descrição, acompanhada de exemplos de aplicação, dos principais testes não-paramétricos de hipóteses empregados em hidrologia. Os testes aqui descritos têm, como objeto principal, a verificação das hipóteses fundamentais da análise de freqüência de uma variável hidrológica. De fato, a premissa de base para a aplicação dos métodos estatísticos a um conjunto de observações de uma variável hidrológica, é que se trata de uma amostra aleatória simples, extraída de uma população única, cuja função de distribuição de probabilidades não é conhecida *a priori*. Nessa premissa de base estão implícitas as hipóteses de *aleatoriedade*, *independência*, *homogeneidade* e *estacionariedade*, as quais, pelas características distributivas das variáveis hidrológicas e pelo tamanho típico de suas amostras, podem ser testadas apenas com o emprego dos testes não-paramétricos. Os testes, apresentados a seguir, estão entre os procedimentos não-paramétricos de maior utilidade na hidrologia estatística.

7.3.1 – Teste da Hipótese de Aleatoriedade

No contexto dos fenômenos do ciclo da água, o termo 'aleatoriedade' significa, essencialmente, que as flutuações de uma certa variável hidrológica decorrem de causas naturais. Nesse sentido, as vazões de um curso d'água regularizadas pela operação de um reservatório, a montante, constituiriam um exemplo de uma série não-aleatória. A aleatoriedade de uma série hidrológica não pode ser demonstrada, mas pode ser rejeitada pela presença de uma estrutura ou por alguma intervenção de natureza não-aleatória. NERC (1975) sugere que a rejeição/não-rejeição da hipótese de aleatoriedade de uma série hidrológica possa ser decidida por meio do *teste não-paramétrico do número de inflexões*. Particularizando para uma série de vazões máximas anuais Q_t, ou seja, considerando um gráfico entre Q_t versus o ano de ocorrência t, uma inflexão poderia ser tanto um pico, quanto um 'vale', nesse diagrama. Um número excessivamente pequeno, ou excessivamente grande, de inflexões é um indicador de não-aleatoriedade. Por outro lado, se uma amostra de N observações é aleatória, pode-se mostrar que o valor esperado do número de inflexões, denotado por p, é dado por

$$E[p] = \frac{2(N-2)}{3} \tag{7.7}$$

com variância que pode ser aproximada por

$$\mathrm{Var}[p] = \frac{16N-29}{90} \tag{7.8}$$

Para amostras de tamanho N >30, é possível provar que a variável p segue aproximadamente uma distribuição Normal. Portanto, se a hipótese nula é H_0: (a amostra é aleatória), a estatística do teste não-paramétrico pode ser formulada como

$$T = \frac{p - E[p]}{\sqrt{\mathrm{Var}[p]}} \tag{7.9}$$

onde p representa o número de picos, e/ou de 'vales', no gráfico da variável aleatória, ao longo do tempo. Por tratar-se de um teste bilateral, a um nível de significância α, a decisão deve ser a de rejeitar a hipótese nula se $|T| > z_{1-\alpha/2}$.

7.3.2 – Teste da Hipótese de Independência

O termo 'independência' significa, essencialmente, que nenhuma observação presente na amostra pode influenciar a ocorrência, ou a não ocorrência, de qualquer outra observação seguinte. Mesmo que uma série seja considerada aleatória, as observações que a constituem podem não ser independentes. No contexto de variáveis hidrológicas, os armazenamentos naturais de uma bacia hidrográfica, por exemplo, podem determinar a ocorrência de vazões de maior porte, na seqüência de vazões elevadas, ou, contrariamente, de vazões de menor porte, na seqüência de vazões reduzidas. A dependência, nesse caso, varia com o intervalo de tempo que separa as observações consecutivas da série hidrológica: forte, para vazões médias diárias, e fraca ou nenhuma, para vazões médias (ou máximas, ou mínimas) anuais. A rejeição ou não-rejeição da hipótese de independência de uma série hidrológica é freqüentemente decidida por meio do *teste não-paramétrico* proposto por Wald e Wolfowitz (1943), o qual encontra-se descrito a seguir.

Dada uma amostra $\{X_1, X_2, \ldots, X_N\}$, de tamanho N, e as diferenças $\{X'_1, X'_2, \ldots, X'_N\}$, entre as observações X_i e a média amostral $\overline{X}$, a estatística do teste de Wald-Wolfowitz é dada por

$$R = \sum_{i=1}^{N-1} X'_i X'_{i+1} + X'_1 X'_N \tag{7.10}$$

Sob a hipótese de que as observações são independentes, pode-se demonstrar que a estatística R segue uma distribuição Normal de média igual a

$$E[R] = -\frac{s_2}{N-1} \tag{7.11}$$

e variância dada por

$$\mathrm{Var}[R] = \frac{s_2^2 - s_4}{N-1} + \frac{s_2^2 - 2s_4}{(N-1)(N-2)} - \frac{s_2^2}{(N-1)^2} \tag{7.12}$$

onde *r* denota a ordem dos momentos amostrais em relação à origem, $s_r = Nm_r'$ e $m_r' = \sum_{i=1}^{N} (X_i')^r \Big/ N$. Portanto, se a hipótese nula é H_0: (os elementos da amostra são independentes), a estatística do teste não-paramétrico de Wald-Wolfowitz pode ser formulada como

$$T = \frac{R - E[R]}{\sqrt{\mathrm{Var}[R]}} \tag{7.13}$$

a qual segue uma distribuição Normal padrão. Por tratar-se de um teste bilateral, a um nível de significância α, a decisão deve ser a de rejeitar a hipótese nula se $|T| > z_{1-\alpha/2}$.

7.3.3 – Teste da Hipótese de Homogeneidade

O termo "homogeneidade" implica que todos os elementos de uma certa amostra provêm de uma única e idêntica população. Em uma série de vazões máximas anuais, por exemplo, muitos de seus valores podem decorrer de enchentes provocadas por precipitações ordinárias ou comuns, enquanto outros advêm de precipitações extraordinariamente elevadas, resultantes de condições hidrometeorológicas muito especiais, como por exemplo, a ocorrência do fenômeno El Niño. Nesse exemplo, temos duas populações de enchentes, diferenciadas pelo seu mecanismo de formação, e, certamente, a série hidrológica deveria ser considerada heterogênea. Entretanto, as amostras hidrológicas, geralmente de tamanhos pequenos, tornam difícil a detecção da heterogeneidade eventualmente presente na série completa. Em geral, é mais fácil identificar heterogeneidades em séries de valores médios ou totais anuais, do que em séries de valores extremos, tomados em intervalos de tempo mais curtos. A rejeição ou não-rejeição da hipótese de homogeneidade de uma série hidrológica é freqüentemente decidida por meio do *teste não-paramétrico* proposto por Mann e Whitney (1947), o qual encontra-se descrito a seguir.

Dada uma amostra $\{X_1, X_2, \ldots, X_N\}$, de tamanho *N*, separe-a em duas sub-amostras $\{X_1, X_2, \ldots, X_{N_1}\}$, de tamanho N_1, e $\{X_{N_1+1}, X_{N_1+2}, \ldots, X_N\}$, de tamanho N_2, de modo que $N_1+N_2=N$ e que N_1 e N_2 sejam aproximadamente iguais, com $N_1 \leq N_2$. Em seguida, classifique, em ordem crescente, a amostra completa de tamanho *N*, indicando a ordem de classificação *m* de cada observação e se ela provem da

primeira ou da segunda sub-amostra. A idéia intuitiva do teste de Mann-Whitney é se as duas sub-amostras não forem homogêneas, os elementos da primeira apresentarão ordens de classificação consistentemente mais baixas (ou mais altas), em relação às ordens de classificação correspondentes à segunda sub-amostra. A estatística do teste *V* de Mann-Whitney é dada pelo *menor valor* entre as quantidades

$$V_1 = N_1 N_2 + \frac{N_1(N_1+1)}{2} - R_1 \tag{7.14}$$

$$V_2 = N_1 N_2 - V_1 \tag{7.15}$$

onde R_1 denota a *soma das ordens de classificação dos elementos da primeira sub-amostra*. Se $N_1, N_2 > 20$, e sob a hipótese de que se trata de uma amostra homogênea, demonstra-se que *V* segue uma distribuição Normal de média igual a

$$E[V] = \frac{N_1 N_2}{2} \tag{7.16}$$

e variância dada por

$$\mathrm{Var}[V] = \frac{N_1 N_2 (N_1 + N_2 + 1)}{12} \tag{7.17}$$

Portanto, se a hipótese nula é H_0: (a amostra é homogênea), a estatística do teste não-paramétrico de Mann-Whitney pode ser formulada como

$$T = \frac{V - E[V]}{\sqrt{\mathrm{Var}[V]}} \tag{7.18}$$

a qual segue uma distribuição Normal padrão. Por tratar-se de um teste bilateral, a um nível de significância α, a decisão deve ser a de rejeitar a hipótese nula se $|T| > z_{1-\alpha/2}$

7.3.4 – Teste da Hipótese de Estacionariedade

O termo "estacionariedade" refere-se ao fato que, excluídas as flutuações aleatórias, as observações amostrais são invariantes, com relação à cronologia de suas ocorrências. Os tipos de não-estacionariedades incluem tendências, 'saltos' e ciclos, ao longo do tempo. Em um contexto hidrológico, os "saltos" estão relacionados a alterações bruscas em uma bacia ou trecho fluvial, tais como, por exemplo, a construção de barragens. Os ciclos, por sua vez, podem estar relacionados a flutuações climáticas de longo período, sendo de difícil detecção. As tendências temporais, em geral, estão associadas a alterações graduais que se processam na bacia, tais como, por exemplo, uma evolução temporal lenta da

urbanização de uma certa área geográfica. Uma tendência temporal, eventualmente presente em uma serie hidrológica X_t, ao longo do tempo t, pode ser detectada pela correlação entre a série e o índice de tempo. Essa é a idéia essencial do *teste não-paramétrico de Spearman*, conforme descrito por NERC (1975), cuja base é o coeficiente de correlação entre as ordens de classificação m_t, da seqüência X_t, e os índices de tempo T_t, esses iguais a 1, 2, ... , N.

A estatística do teste de Spearman tem, como base, o seguinte coeficiente:

$$r_S = 1 - \frac{6\sum_{t=1}^{N}(m_t - T_t)^2}{N^3 - N} \tag{7.19}$$

Para N >10 e sob a hipótese nula de que não há correlação entre m_t e T_t, demonstra-se que a distribuição de r_S pode ser aproximada por uma Normal de média igual a

$$E[r_S] = 0 \tag{7.20}$$

e variância dada por

$$\mathrm{Var}[r_S] = \frac{1}{N-1} \tag{7.21}$$

Portanto, se a hipótese nula é H_0: (a amostra não apresenta tendência temporal), a estatística do teste não-paramétrico de Spearman pode ser formulada como

$$T = \frac{r_S}{\sqrt{\mathrm{Var}[r_S]}} \tag{7.22}$$

a qual segue uma distribuição Normal padrão. Por tratar-se de um teste bilateral, a um nível de significância α, a decisão deve ser a de rejeitar a hipótese nula se $|T| > z_{1-\alpha/2}$.

Exemplo 7.6 – Considere a série de vazões médias anuais do Rio Paraopeba em Ponte Nova do Paraopeba, listadas na Tabela 7.1, e teste as hipóteses de (a) aleatoriedade, (b) independência, (c) homogeneidade e (d) estacionariedade, a um nível de significância $\alpha = 0{,}05$.

Solução: (a) Teste da hipótese de aleatoriedade. A variação temporal da vazões é mostrada na Figura 7.4. Nessa figura, observa-se que o número total de inflexões importantes é $p = 34$. Para $N = 62$, as equações 7.7 e 7.8 resultam em $E[p] = 40$ e $\mathrm{Var}[p] = 10{,}7$. Com esses valores, a estatística de teste, da equação 7.9, é T = -1,8340. Para o nível de significância $\alpha = 0{,}05$, $z_{0{,}975} = 1{,}96$. Como $|T| < z_{0{,}975}$, a decisão é a de não rejeitar a hipótese H_0 de que as observações são aleatórias.

(b) Teste da hipótese de independência. A sexta coluna da tabela 7.1 apresenta as diferenças entre as vazões médias anuais e o valor médio global de 86,105 m^3/s. São esses os valores necessários para o cálculo da estatística do teste de Wald-Wolfowitz, pela equação 7.10. O resultado desse cálculo é R = 8253,759. As diferenças, listadas na tabela 7.1, também fornecem os valores s_2 = 38003,47 e s_4 = 87362890,7, cuja substituição nas equações 7.11 e 7.12 resultam em $E[R]$ = -623,01 e Var[R] = 22203003,87. Com esses valores, a estatística de teste, da equação 7.13, é T = 1,8839. Para o nível de significância $\alpha = 0,05$, $z_{0,975} = 1,96$. Como $|T| < z_{0,975}$, a decisão é a de não rejeitar a hipótese H_0 de que as observações são independentes.

(c) Teste da hipótese de homogeneidade. A quarta coluna da tabela 7.1 apresenta as ordens de classificação das vazões médias anuais, denotadas por m_t. São esses os valores necessários para o cálculo da estatística do teste de Mann-Whitney, pelas equações 7.14 e 7.15, lembrando que a soma das ordens de classificação da primeira sub-amostra de 31 elementos, também anotada na tabela 7.1, é R_1 = 1004. A estatística de teste é o menor valor entre V_1 e V_2, ou seja V = 453. A substituição de R_1 e V nas equações 7.16 e 7.17 resulta em $E[V]$ = 480,5 e Var[V] = 71,0299. Com esses valores, a estatística de teste, da equação 7.18, é T = 0,3872. Para o nível de significância $\alpha = 0,05$, $z_{0,975} = 1,96$. Como $|T| < z_{0,975}$, a decisão é a de não rejeitar a hipótese H_0 de que as observações são homogêneas.

(d) Teste da hipótese de estacionariedade. A quarta coluna da tabela 7.1 apresenta as ordens de classificação das vazões médias anuais e a segunda coluna lista o índice de tempo cronológico T_t. São esses os valores necessários para o cálculo da estatística do teste de Spearman, pela equação 7.19. A estatística de teste calculada é r_S = -0,07618. As equações 7.20 e 7.21 resultam em $E[r_S] = 0$ e Var[r_S] = 0,0164. Com esses valores, a estatística de teste, da equação 7.22, é T = 0,5949. Para o nível de significância $\alpha = 0,05$, $z_{0,975} = 1,96$. Como $|T| < z_{0,975}$, a decisão é a de não rejeitar a hipótese H_0 de que as observações são estacionárias.

Tabela 7.1 – Vazões médias anuais do Rio Paraopeba em Ponte Nova do Paraopeba (m^3/s) e grandezas auxiliares para a realização dos testes de hipóteses de Wald-Wolfowitz, Mann-Whitney e Spearman.

Ano Civil	T_t	X_t	m_t	Sub-Amostra	$X_i^{'} = X_t - \overline{X}$	X_t classificados
1938	1	104,3	51	1	18,20	43,6
1939	2	97,9	45	1	11,80	46,8
1940	3	89,2	38	1	3,10	49,4
1941	4	92,7	40	1	6,60	50,1
1942	5	98	46	1	11,90	53,1
1943	6	141,7	60	1	55,60	57

Tabela 7.1 – Continuação

Ano Civil	T_t	X_t	m_t	Sub-Amostra	$X_i^{'} = X_t - \overline{X}$	X_t classificados
1944	7	81,1	30	1	-5,00	57,3
1945	8	97,3	43	1	11,20	59,9
1946	9	72	20	1	-14,10	60,6
1947	10	93,9	41	1	7,80	61,2
1948	11	83,8	33	1	-2,30	62,6
1949	12	122,8	58	1	36,70	63,6
1950	13	87,6	36	1	1,50	64,2
1951	14	101	50	1	14,90	66,8
1952	15	97,8	44	1	11,70	67,2
1953	16	59,9	8	1	-26,20	68,2
1954	17	49,4	3	1	-36,70	68,7
1955	18	57	6	1	-29,10	69,3
1956	19	68,2	16	1	-17,90	71,6
1957	20	83,2	32	1	-2,90	72
1958	21	60,6	9	1	-25,50	72,4
1959	22	50,1	4	1	-36,00	74,8
1960	23	68,7	17	1	-17,40	76,4
1961	24	117,1	56	1	31,00	77,6
1962	25	80,2	28	1	-5,90	78
1963	26	43,6	1	1	-42,50	78,9
1964	27	66,8	14	1	-19,30	79
1965	28	118,4	57	1	32,30	80,2
1966	29	110,4	52	1	24,30	80,9
1967	30	99,1	47	1	13,00	81,1
1968	31	71,6	19	1	-14,50	82,2
			Soma=1004			
1969	32	62,6	11	2	-23,50	83,2
1970	33	61,2	10	2	-24,90	83,8
1971	34	46,8	2	2	-39,30	85,1
1972	35	79	27	2	-7,10	87,4
1973	36	96,3	42	2	10,20	87,6
1974	37	77,6	24	2	-8,50	88,1
1975	38	69,3	18	2	-16,80	89,2
1976	39	67,2	15	2	-18,90	89,8
1977	40	72,4	21	2	-13,70	92,7
1978	41	78	25	2	-8,10	93,9
1979	42	141,8	61	2	55,70	96,3
1980	43	100,7	49	2	14,60	97,3
1981	44	87,4	35	2	1,30	97,8
1982	45	100,2	48	2	14,10	97,9
1983	46	166,9	62	2	80,80	98
1984	47	74,8	22	2	-11,30	99,1
1985	48	133,4	59	2	47,30	100,2
1986	49	85,1	34	2	-1,00	100,7
1987	50	78,9	26	2	-7,10	101
1988	51	76,4	23	2	-9,70	104,3
1989	52	64,2	13	2	-21,90	110,4
1990	53	53,1	5	2	-33,00	110,8
1991	54	112,2	54	2	26,10	112,2
1992	55	110,8	53	2	24,70	114,9
1993	56	82,2	31	2	-3,90	117,1
1994	57	88,1	37	2	2,00	118,4
1995	58	80,9	29	2	-5,20	122,8
1996	59	89,8	39	2	3,70	133,4
1997	60	114,9	55	2	28,80	141,7
1998	61	63,6	12	2	-22,50	141,8
1999	62	57,3	7	2	-28,80	166,9
			Soma=949			

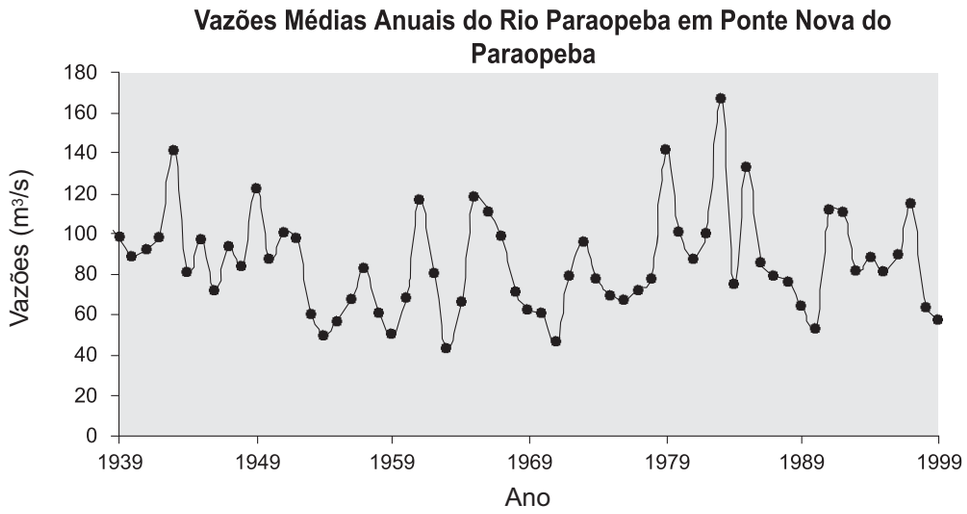


Figura 7.4 - Variação temporal das vazões médias anuais do Rio Paraopeba em Ponte Nova do Paraopeba

7.4 – Alguns Testes de Aderência Usuais em Hidrologia

Nos itens anteriores, foram descritos alguns testes de hipóteses referentes aos parâmetros de uma certa população ou referentes a atributos necessários a uma amostra aleatória simples. Outra classe importante de testes de hipóteses refere-se à *verificação da forma* de uma distribuição de probabilidades. Essa classe é constituída pelos chamados *testes de aderência*, por meio dos quais, verifica-se a eventual adequação entre as probabilidades ou freqüências, tal como calculadas por um certo modelo distributivo hipotético, e as correspondentes freqüências com que, determinados valores amostrais são observados. Os testes de aderência permitem, por exemplo, verificar se uma variável aleatória discreta segue uma distribuição de Poisson ou se uma variável aleatória contínua é distribuída segundo um modelo de Gumbel.

No contexto das variáveis aleatórias hidrológicas, é muito freqüente a situação em que não se conhece *a priori* a distribuição de probabilidades que descreve a população da qual se extraiu um certo conjunto de observações. Nessas circunstâncias, a seleção das distribuições de probabilidades aptas à modelação de uma determinada variável hidrológica é realizada com base (i) nas características físicas do fenômeno em foco; (ii) em possíveis deduções teóricas quanto às propriedades distributivas da variável em questão; e (iii) na aderência da distribuição

proposta à distribuição empírica dos valores amostrais. No que concerne ao item (i), a dinâmica do mecanismo de formação de cheias, por exemplo, é um fator que indica que as distribuições positivamente assimétricas sejam mais aptas à modelação de vazões máximas anuais, enquanto que a capacidade máxima de dissolução de um gás, em um meio líquido, é outro fato que determina que as distribuições limitadas, à esquerda e à direita, sejam mais adequadas à descrição do comportamento probabilístico das concentrações de oxigênio dissolvido, em um trecho fluvial.

No que se refere ao item (ii), é possível conceber algumas poucas variáveis hidrológicas, tais como as alturas anuais de precipitação em locais de sazonalidade pouco marcada, como decorrentes da aplicação do teorema do limite central às alturas pluviométricas diárias. Entretanto, para a grande maioria das variáveis aleatórias hidrológicas, é patente a inexistência de leis dedutivas teóricas que amparem a escolha do modelo que descreve o seu comportamento probabilístico.

Em relação ao item (iii), embora não se prestem à seleção de uma dentre várias distribuições, os testes de aderência são instrumentos da estatística matemática que auxiliam a tomada de decisão quanto à adequação, ou inadequação, de um certo modelo distributivo a uma dada amostra. Os principais testes de aderência, empregados na hidrologia estatística, são o do Qui-Quadrado, o de Kolmogorov-Smirnov, o de Anderson-Darling e o de Filliben. A descrição e aplicação de tais testes são objetos dos itens que se seguem.

7.4.1 – O Teste de Aderência do Qui-Quadrado (χ^2)

Considere que A_1, A_2, ... , A_r representem um conjunto de eventos mútua e coletivamente disjuntos, de modo que o espaço amostral seja definido pela união desses eventos. Considere também a hipótese nula H_0: $P(A_i) = p_i$, para i = 1, 2, ..., r, tal que $\sum_{i=1}^{r} p_i = 1$. Sob tais condições, suponha que, de um certo número de experimentos N, as freqüências absolutas dos elementos pertencentes aos eventos $A_1, A_2, \ldots, A_r$ sejam dadas, respectivamente, pelas quantidades $\rho_1, \rho_2, \ldots, \rho_r$. Se a hipótese nula é verdadeira, a distribuição conjunta das variáveis $\rho_1, \rho_2, \ldots, \rho_r$ é a multinomial (ver item 4.3.1, do capítulo 4), cuja função massa é dada por

$$P\left(\rho_1 = O_1, \rho_2 = O_2, \ldots, \rho_r = O_r \middle| H_0\right) = \frac{N!}{O_1! O_2! \ldots O_r!} p_1^{O_1} p_2^{O_2} \ldots p_r^{O_r} \qquad (7.23)$$

onde $\sum_{i=1}^{r} O_i = N$

Em seguida, considere a seguinte estatística :

$$\chi^2 = \sum_{i=1}^{r} \frac{(O_i - Np_i)^2}{O_i} = \sum_{i=1}^{r} \frac{(O_i - E_i)^2}{O_i} \tag{7.24}$$

formada pelas realizações O_i, das variáveis ρ_i, e pelos seus respectivos valores esperados $E_i = E[\rho_i]$, os quais, *sob a veracidade da hipótese nula*, são iguais a Np_i. A estatística χ^2 expressa, portanto, a soma das diferenças quadráticas entre as realizações das variáveis aleatórias ρ_i e suas respectivas médias populacionais.

No item 5.9.1, do capítulo 5, viu-se que a soma das diferenças quadráticas entre N variáveis normais e independentes, e sua média comum μ, possui uma distribuição do χ^2 com $\nu = N$ graus de liberdade. Embora seja evidente a semelhança entre a definição da variável da distribuição do Qui-Quadrado com a estatística χ^2, a equação 7.24 contem a soma de r variáveis não necessariamente independentes ou normais. Entretanto, é possível demonstrar que, quando N tende para o infinito, a estatística χ^2, tal como expressa pela equação 7.24, segue uma distribuição do Qui-Quadrado, com $\nu = (r\text{-}1)$ graus de liberdade. Em outros termos,

$$\lim_{N\to\infty} \mathrm{P}\left(\chi^2 < x \middle| \mathrm{H}_0\right) = \int_0^x \frac{x^{(r-3)/2} e^{-x/2}}{2^{(r-1)/2}\,\Gamma[(r-1)/2]}\,dx \tag{7.25}$$

Para grandes valores de N, pode-se, portanto, empregar esse resultado para testar a hipótese nula H_0 de que as freqüências relativas esperadas de ρ_i sejam dadas por Np_i, com p_i calculadas pela distribuição de probabilidades proposta. Um valor elevado da estatística de teste revela grandes diferenças entre as freqüências observadas e esperadas, sendo um indicador da pouca aderência da distribuição especificada, sob H_0, à amostra.

Observe que a distribuição limite da estatística de teste, dada pela equação 7.25, não depende de p_i, contido em H_0. De fato, a distribuição limite de χ^2 depende apenas do número de partições r do espaço amostral. Isso faz com que o teste possa ser aplicado para diferentes hipóteses nulas, desde que r seja adequadamente especificado. Na prática, o teste de aderência do χ^2 fornece resultados satisfatórios para $N > 50$ e para $Np_i \geq 5$, com $i = 1, 2, \ldots, r$. Se as probabilidades p_i forem calculadas a partir de uma distribuição de k parâmetros, estimados pelas observações amostrais, perde-se k graus de liberdade adicionais. Em outras palavras, o parâmetro ν, da distribuição da estatística de teste χ^2, será $\nu = (r - k\text{-}1)$. Os exemplos 7.7 e 7.8 ilustram a aplicação do teste de aderência do χ^2 para variáveis aleatórias discretas e contínuas.

Exemplo 7.7 - Considere que uma ETA recebe água bruta de um manancial de superfície, captada por uma tomada d'água simples, instalada em determinada cota. Suponha que a variável aleatória discreta X represente o número anual de dias em que o nível d'água, medido na estação fluviométrica local, é inferior à cota da tomada d'água de projeto. Com base em 50 anos de observações, determinou-se a distribuição empírica das freqüências de X, a qual é dada pela Tabela 7.2. Use o método dos momentos para ajustar uma distribuição de Poisson à variável X, calcule as freqüências esperadas por esse modelo e teste sua aderência aos dados empíricos, a um nível de significância $\alpha = 0{,}05$.

Tabela 7.2 - Número anual de dias em que o nível d'água é inferior à cota da tomada d'água de projeto

$x_i \longrightarrow$	0	1	2	3	4	5	6	7	8	>8
$f(x_i)$	0,0	0,06	0,18	0,2	0,26	0,12	0,09	0,06	0,03	0,0

Solução: A função massa de Poisson é $p_X(x) = \frac{\nu^x}{x!} e^{-\nu}$, para $x = 0, 1, \ldots$ e $\nu > 0$, com valor esperado $E[X] = \nu$. A média amostral pode ser calculada pela ponderação de x por suas freqüências observadas e resulta ser $\bar{x} = 3{,}86$. Portanto, pelo método dos momentos, a estimativa do parâmetro $\hat{\nu} \in 3{,}86$. Os valores E_i, da Tabela 7.3 representam as freqüências esperadas, tal como calculadas pela produto de $N = 50$, pela função massa de Poisson.

Tabela 7.3 – Freqüências observadas e empíricas.

x_i	$O_i = 50\, f(x_i)$	$E_i = 50 p_X(x_i)$	$O_i - E_i$	$\frac{(O_i - E_i)^2}{E_i}$
0	0	1,0534	-1,0534	1,0534
1	3	4,0661	-1,0661	0,2795
2	9	7,8476	1,1524	0,1692
3	10	10,0973	-0,0973	0,0009
4	13	9,7439	3,2561	1,0880
5	6	7,5223	-1,5223	0,3080
6	4,5	4,8393	-0,3393	0,0238
7	3	2,6685	0,3315	0,0412
8	1,5	1,2876	0,2124	0,0350
>8	0	0,8740	-0,8740	0,8740
Soma	**50**	**50**	———	3,8733

A Tabela 7.3 também mostra os outros elementos necessários para o cálculo da estatística de teste χ^2, quais sejam, as diferenças simples e quadráticas padronizadas, entre as freqüências empíricas e esperadas pelo modelo de Poisson. A soma da última coluna da tabela fornece o valor da estatística de teste $\chi^2 = 3{,}8733$. O número total de partições do espaço amostral, nesse caso, é $r = 10$. Como foi estimado um parâmetro a partir da amostra, $k = 1$, o que resulta em $\nu = (r - k - 1) = 8$ graus de liberdade para a distribuição da estatística de teste. Trata-se de *um teste unilateral*, no qual, a região crítica, para $\alpha = 0{,}05$, é definida por $\chi^2_{0,95,\,\nu=8} = 15{,}5$ (Anexo 6). Como $\chi^2 < \chi^2_{0,95,\,\nu=8}$, a decisão é a de não rejeitar a hipótese H_0 de que o comportamento probabilístico da variável aleatória, em questão, possa ser modelado pela distribuição de Poisson. Nesse exemplo, embora $N = 50$, algumas freqüências esperadas pelo modelo de Poisson foram inferiores a 5, o que pode vir a comprometer o poder do teste de aderência. Essa situação pode ser resolvida satisfatoriamente pela agregação de algumas partições do espaço amostral; por exemplo, as freqüências esperadas para $x = 0$ e $x = 1$ podem ser agrupadas em uma nova partição, correspondente a $x \leq 1$, cuja nova freqüência seria 5,1195. Da mesma forma, outras partições poderiam ser agrupadas para constituir a nova classe $x \geq 6$. Evidentemente, esse rearranjo das partições implicaria em novos valores de r e da estatística de teste χ^2.

Exemplo 7.8 – Considere as vazões médias anuais do Rio Paraopeba em Ponte Nova do Paraopeba, listadas na Tabela 7.1, e faça um teste de aderência da distribuição Normal a esses dados, por meio do teste do χ^2, a um nível de significância $\alpha = 0{,}05$.

Solução: No caso de variáveis aleatórias contínuas, as partições do espaço amostral são feitas por meio da divisão em classes, com o cálculo das freqüências observadas e esperadas, dentro dos limites dos intervalos de classe. Para a amostra em questão, já foram mostradas anteriormente a tabela de freqüências absolutas e o histograma correspondente, com 7 classes de largura fixa. A tabela de freqüências absolutas, anteriormente elaborada, é a Tabela 2.3, do capítulo 2. Entretanto, nessa tabela, observa-se uma freqüência muito pequena em alguns intervalos, o que força o rearranjo das classes, cujas larguras não necessitam ser fixas. A Tabela 7.4 mostra esse rearranjo e outros elementos auxiliares para a construção da estatística de teste.

Tabela 7.4 – Freqüências observadas e empíricas.

Classe	Intervalos	O_i	E_i	O_i-E_i	$\frac{(O_i - E_i)^2}{E_i}$
1	(0,60]	8	9,1468	-1,1468	0,1438
2	(60,70]	10	6,9179	3,0822	1,3732
3	(70,90]	21	18,7621	2,2379	0,2669
4	(90,105]	12	13,2355	-1,2355	0,1153
5	(105,120]	6	8,5117	-2,5117	0,7412
6	(120,200]	5	5,4085	-0,4085	0,0309
Soma		62	61,9824	——	2,6712

Com $r = 6$ classes de largura variável, as freqüências observadas O_i variam em torno de valores aceitáveis. Para o cálculo das freqüências esperadas pela distribuição Normal, é preciso estimar os seus parâmetros μ e σ. A amostra fornece a média $\bar{x} = 86{,}105$ e o desvio-padrão $s_X = 24{,}960$, os quais, pelo método dos momentos, resultam nas estimativas $\hat{\mu} = 86{,}105$ e $\hat{\sigma} = 24{,}960$. Desse modo, a freqüência relativa esperada na classe i é dada por $p_i = \Phi[(LS - \hat{\mu})/\hat{\sigma}] - \Phi[(LI - \hat{\mu})/\hat{\sigma}]$, onde LS e LI representam, respectivamente, os limites superior e inferior de cada classe, e $\Phi(.)$ denota a FAP da distribuição Normal de parâmetros $\hat{\mu}$ *e* $\hat{\sigma}$. A freqüência absoluta E_i, da classe i, é dada pelo produto de p_i pelo tamanho da amostra $N = 62$. Em seguida, calcula-se as diferenças simples e quadráticas padronizadas, entre as freqüências empíricas e esperadas pelo modelo Normal. A soma da última coluna da Tabela 7.4 fornece o valor da estatística de teste $\chi^2 = 2{,}6712$. O número total de partições do espaço amostral, nesse caso, é $r = 6$. Como foram estimados dois parâmetros a partir da amostra, $k = 2$, o que resulta em $\nu = (r - k - 1) = 3$ graus de liberdade para a distribuição da estatística de teste. Trata-se de *um teste unilateral*, no qual, a região crítica, para $\alpha = 0{,}05$, é definida por $\chi^2_{0,95,\nu=3} = 7{,}81$ (Anexo 6). Como $\chi^2 < \chi^2_{0,95,\nu=3}$, a decisão é a de não rejeitar a hipótese H_0 de que o comportamento probabilístico da variável aleatória, em questão, possa ser modelado pela distribuição Normal.

7.4.2 – O Teste de Aderência de Kolmogorov-Smirnov (KS)

O teste de aderência de Kolmogorov-Smirnov (KS) é um teste não paramétrico, cuja estatística de teste tem como base a *diferença máxima* entre as funções de probabilidades acumuladas, empírica e teórica, de variáveis aleatórias contínuas. O teste não é aplicável a variáveis aleatórias discretas.

Considere que X represente uma variável aleatória contínua, de cuja população extraiu-se a amostra $\{X_1, X_2, \ldots, X_N\}$. A hipótese nula a ser testada é H_0: $P(X < x) = F_X(x)$, onde $F_X(x)$ é suposta *completamente conhecida*, ou seja, seus parâmetros não são estimados a partir da amostra. Para implementar o teste KS, inicialmente, classifique os elementos da amostra $\{X_1, X_2, \ldots, X_N\}$ em ordem crescente, de modo a constituir a seqüência $\{x_{(1)}, x_{(2)}, \ldots, x_{(m)}, \ldots x_{(N)}\}$, na qual $1 \leq m \leq N$ denota a ordem de classificação. Para cada elemento $x_{(m)}$, a distribuição empírica $F_N(x_m)$ é calculada pela proporção de valores amostrais que não excedem $x_{(m)}$, ou seja,

$$F_N(x_m) = \frac{m}{N} \tag{7.26}$$

Em seguida, calcule as probabilidades teóricas, segundo $F_X(x)$, tendo como argumento os valores $x_{(m)}$. A estatística do teste KS é dada por

$$D_N = \sup_{-\infty < x < \infty} \left| F_N(x) - F_X(x) \right| \tag{7.27}$$

e corresponde, portanto, à maior diferença entre as probabilidades empírica e teórica.

Se H_0 é verdadeira e quando $N \to \infty$, a estatística D_N irá tender a 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. Smirnov (1948) determinou a distribuição limite da variável aleatória $\sqrt{N} D_N$, a qual, sob a premissa de veracidade da hipótese H_0, é expressa por

$$\lim_{N \to \infty} P\left(\sqrt{N} D_N \leq z\right) = \frac{\sqrt{2\pi}}{z} \sum_{k=1}^{\infty} exp\left[-\frac{(2k-1)^2 \pi^2}{8 z^2} \right] \tag{7.28}$$

Portanto, para amostras de tamanho superior a 40, os valores críticos da estatística de teste D_N serão $1{,}3581/\sqrt{N}$, para o nível de significância α = 0,05, e $1{,}6276/\sqrt{N}$, para $\alpha = 0{,}01$; esses valores resultam da soma dos cinco primeiros termos da somatória da equação 7.28, e permanecem praticamente inalterados a partir do sexto termo. Para amostras de tamanho inferior a 40, os valores críticos de D_N devem ser obtidos da Tabela 7.5. O exemplo 7.9 ilustra a aplicação do teste de aderência de Kolmogorov-Smirnov.

A construção da estatística do teste KS parte da premissa que $F_X(x)$ é completamente conhecida e, portanto, que seus parâmetros são especificados e, portanto, não são estimados a partir da amostra. Entretanto, quando as estimativas dos parâmetros são obtidas dos elementos da amostra, simulações de Monte

Carlo demonstram que o teste KS é conservador quanto à magnitude do erro do tipo I, podendo ocorrer rejeições indevidas da hipótese nula. Com o objetivo de corrigir tal situação, Crutcher (1975) *apud* Haan (1977), apresenta novas tabelas de valores críticos da estatística $D_{N,\alpha}$ para amostras de tamanhos variáveis, considerando, sob H_0, as distribuições exponencial, gama, normal e Gumbel.

Tabela 7.5 – Valores críticos da estatística $D_{N,\alpha}$ do teste de aderência KS

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	$1{,}22/\sqrt{N}$	$1{,}36/\sqrt{N}$	$1{,}52/\sqrt{N}$	$1{,}63/\sqrt{N}$

Exemplo 7.9 – Refaça o exemplo 7.8, com o teste de aderência de Kolmogorov-Smirnov.

Solução: A última coluna da Tabela 7.1 apresenta as vazões médias anuais do Rio Paraopeba, em Ponte Nova do Paraopeba, já classificadas em ordem crescente. As freqüências empíricas correspondentes às vazões classificadas podem ser calculadas pela equação 7.26. As freqüências teóricas, correspondentes à distribuição Normal, podem ser calculadas por $F_X(X_t^*) = \Phi[(X_t^* - \mu)/\sigma]$, onde X_t^* representa a vazão classificada, μ e σ denotam os parâmetros populacionais, supostamente iguais a $\overline{X} = 86{,}105$ e $s_X = 24{,}960$, respectivamente, e $\Phi(.)$ é a FAP da distribuição Normal. A Figura 7.5 apresenta o gráfico das freqüências empíricas e teóricas, versus as vazões médias anuais classificadas.

No gráfico da Figura 7.5, também está indicada a máxima diferença absoluta entre as freqüências empíricas e teóricas, a qual foi calculada em $D_N^{calc} = 0{,}08179$. Consultando a Tabela 7.5, para $\alpha = 0{,}05$ (teste unilateral) e

N=62, vê-se que o valor crítico da estatística de teste é $D_{N,0,05} = 1{,}36/\sqrt{N} = 1{,}36/\sqrt{62} = 0{,}1727$, o qual define o limite inferior da região de rejeição da hipótese nula. Portanto, como $D_N^{calc} < D_{n,0,05}$, a decisão é a de não rejeitar a hipótese H_0 de que o comportamento probabilístico da variável aleatória, em questão, possa ser modelado pela distribuição Normal.

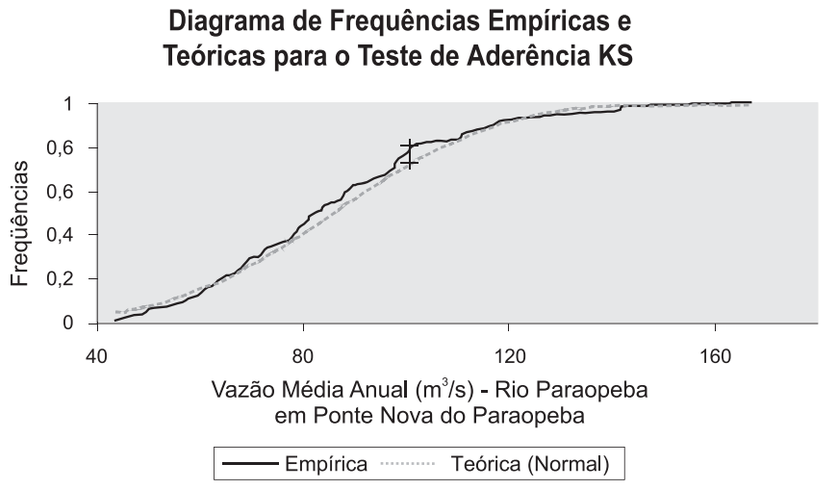


Figura 7.5 – Freqüências empíricas e teóricas para o teste de aderência de Kolmogorov-Smirnov

7.4.3 – O Teste de Aderência de Anderson-Darling (AD)

O poder dos testes de aderência do Qui-Quadrado e de Kolmogorov-Smirnov, de discriminar entre hipóteses falsas e verdadeiras, é bastante diminuído nas caudas inferior e superior, tanto em função do reduzido número de observações amostrais, quanto em decorrência dos maiores erros de estimação, nessas partições do espaço amostral. Alternativamente, o teste de aderência de Anderson-Darling é um teste não-paramétrico que procura ponderar mais fortemente as caudas das distribuições, nas quais, as maiores (ou as menores) observações da amostra podem alterar sobremaneira a qualidade do ajuste. O teste de aderência de Anderson-Darling, tal como o de Kolmogorov-Smirnov, baseia-se na diferença entre as funções de probabilidades acumuladas, empírica, $F_N(x)$, e teórica, $F_X(x)$, de variáveis aleatórias contínuas. Entretanto, o teste AD dá mais peso às caudas, por meio da divisão das diferenças entre $F_N(x)$ e $F_X(x)$ por $\sqrt{F_X(x)[1-F_X(x)]}$. Desse modo, a estatística do teste de Anderson-Darling torna-se

$$A^2 = \int_{-\infty}^{\infty} \frac{[F_N(x) - F_X(x)]^2}{F_X(x)[1-F_X(x)]} f_X(x)\,dx \qquad (7.29)$$

onde $f_X(x)$ é a função densidade, segundo a hipótese nula. Anderson e Darling (1954) demonstraram que a equação 7.29 é equivalente a

$$A^2 = -N - \sum_{i=1}^{N} \frac{(2i-1)\left\{ln\, F_X\left(x_{(i)}\right) + ln\left[1 - F_X\left(x_{(N-i+1)}\right)\right]\right\}}{N} \tag{7.30}$$

onde $\{x_{(1)}, x_{(2)}, \ldots, x_{(m)}, \ldots x_{(N)}\}$ representam as observações ordenadas em modo crescente.

Se a estatística A^2 resulta ser um valor elevado, as distribuições empírica, $F_N(x)$, e teórica, $F_X(x)$, diferem muito entre si e, em conseqüência, a hipótese nula deve ser rejeitada. A distribuição de probabilidades da estatística do teste AD depende da distribuição de probabilidades hipotética $F_X(x)$. Se a distribuição de probabilidades, sob H_0, é a Normal ou a Log-Normal, os valores críticos de A^2 são os apresentados na Tabela 7.6.

Tabela 7.6 - Valores críticos da estatística A^2_α do teste de aderência AD, se a distribuição hipotética é Normal ou Log-Normal (Fonte: D'Agostino e Stephens, 1986).

α	0,1	0,05	0,025	0,01
$A^2_{crit,\alpha}$	0,631	0,752	0,873	1,035

Para esse caso, a estatística de teste, calculada pela equação 7.30, deve ser multiplicada pelo fator de correção $\left(1 + 0{,}75/N + 2{,}25/N^2\right)$.

Se a distribuição de probabilidades, sob H_0, é a Weibull de dois parâmetros, para mínimos, ou a Gumbel, para máximos, os valores críticos de A^2 são os apresentados na Tabela 7.7.

Tabela 7.7 - Valores críticos da estatística A^2_α do teste de aderência AD, se a distribuição hipotética é Weibull (mínimos, 2p) ou Gumbel (máximos) (Fonte: D'Agostino e Stephens, 1986).

α	0,1	0,05	0,025	0,01
$A^2_{crit,\alpha}$	0,637	0,757	0,877	1,038

Para esse caso, a estatística de teste, calculada pela equação 7.30, deve ser multiplicada pelo fator de correção $\left(1 + 0{,}2/\sqrt{N}\right)$. A Tabela 7.7 pode ser usada também para a distribuição exponencial.

Não existem tabulações dos valores críticos da estatística A^2_a, para outras distribuições de probabilidades passíveis de serem incluídas na hipótese H_0. Para essas, as alternativas são (i) utilizar os outros testes de aderência ou (ii) obter

resultados aproximados e independentes de N, com os valores da Tabela 7.6. O exemplo 7.10 ilustra a aplicação do teste de Anderson-Darling para as vazões médias anuais do Rio Paraopeba em Ponte Nova do Paraopeba.

Exemplo 7.10 - Refaça o exemplo 7.8, com o teste de aderência de Anderson-Darling.

Solução: A Tabela 7.8 apresenta um resumo dos cálculos necessários para a determinação da estatística A^2. Na Tabela 7.8, a distribuição hipotética $F_X(x)$ é a Normal, calculada por $\Phi[(x-86{,}105)/24{,}960]$. A estatística de teste A^2, sem correção, é calculada por

$A^2 = -N - \sum_{i=1}^{N} S_i / N = -62 - (-3876{,}63)/62 = 0{,}5262$. Nesse caso, o fator de correção é $(1+0{,}75/N+2{,}25/N^2) = 1{,}0127$. Logo, a estatística de teste, já corrigida, é A^2=0,5329. Consultando a Tabela 7.6, para $\alpha = 0{,}05$ (teste unilateral), vê-se que o valor crítico da estatística de teste é $A^2_{crit,0,05} = 0{,}752$, o qual define o limite inferior da região de rejeição da hipótese nula. Portanto, como $A^2 < A^2_{crit,0,05}$, a decisão é a de não rejeitar a hipótese H_0 de que o comportamento probabilístico da variável aleatória, em questão, possa ser modelado pela distribuição Normal.

Tabela 7.8 – Cálculo da estatística do teste de aderência AD – Vazões médias anuais em Ponte Nova do Paraopeba

Ano Civil	X_t	i	$x_{(i)}$	$x_{(N-i+1)}$	$w=F_X(x_{(i)})$	ln$w$	$t = 1- F_X(x_{(N-i+1)})$	ln$t$	S_i *
1938	104,3	1	43,6	166,9	0,0443	-3,1170	0,0006	-7,4118	-10,5288
1939	97,9	2	46,8	141,8	0,0577	-2,8532	0,0128	-4,3561	-21,6279
1940	89,2	3	49,4	141,7	0,0707	-2,6492	0,0130	-4,3458	-34,9750
1941	92,7	4	50,1	133,4	0,0746	-2,5959	0,0291	-3,5385	-42,9406
1942	98	5	53,1	122,8	0,0930	-2,3748	0,0708	-2,6485	-45,2095
1943	141,7	6	57	118,4	0,1218	-2,1054	0,0979	-2,3243	-48,7266
1944	81,1	7	57,3	117,1	0,1242	-2,0855	0,1072	-2,2335	-56,1469
1945	97,3	8	59,9	114,9	0,1469	-1,9181	0,1243	-2,0849	-60,0445
1946	72	9	60,6	112,2	0,1534	-1,8745	0,1479	-1,9112	-64,3572
1947	93,9	10	61,2	110,8	0,1592	-1,8377	0,1612	-1,8249	-69,5882
1948	83,8	11	62,6	110,4	0,1732	-1,7535	0,1652	-1,8007	-74,6371
1949	122,8	12	63,6	104,3	0,1836	-1,6949	0,2330	-1,4567	-72,4854
1950	87,6	13	64,2	101	0,1901	-1,6603	0,2753	-1,2898	-73,7520
1951	101	14	66,8	100,7	0,2196	-1,5158	0,2794	-1,2752	-75,3583
1952	97,8	15	67,2	100,2	0,2244	-1,4943	0,2861	-1,2513	-79,6224
1953	59,9	16	68,2	99,1	0,2366	-1,4415	0,3013	-1,1996	-81,8735
1954	49,4	17	68,7	98	0,2428	-1,4155	0,3168	-1,1494	-84,6409
1955	57	18	69,3	97,9	0,2504	-1,3848	0,3183	-1,1449	-88,5369
1956	68,2	19	71,6	97,8	0,2806	-1,2709	0,3197	-1,1404	-89,2179
1957	83,2	20	72	97,3	0,2860	-1,2518	0,3269	-1,1181	-92,4257
1958	60,6	21	72,4	96,3	0,2915	-1,2328	0,3415	-1,0745	-94,5990
1959	50,1	22	74,8	93,9	0,3253	-1,1230	0,3774	-0,9744	-90,1897
1960	68,7	23	76,4	92,7	0,3487	-1,0535	0,3958	-0,9268	-89,1165
1961	117,1	24	77,6	89,8	0,3666	-1,0034	0,4412	-0,8184	-85,6203
1962	80,2	25	78	89,2	0,3727	-0,9870	0,4507	-0,7970	-87,4178
1963	43,6	26	78,9	88,1	0,3864	-0,9508	0,4681	-0,7590	-87,1999
1964	66,8	27	79	87,6	0,3880	-0,9469	0,4761	-0,7421	-89,5146
1965	118,4	28	80,2	87,4	0,4065	-0,9002	0,4793	-0,7354	-89,9578
1966	110,4	29	80,9	85,1	0,4174	-0,8737	0,5161	-0,6615	-87,5081
1967	99,1	30	81,1	83,8	0,4205	-0,8662	0,5368	-0,6221	-87,8140
1968	71,6	31	82,2	83,2	0,4378	-0,8259	0,5463	-0,6045	-87,2570

Tabela 7.8 – Continuação

Ano Civil	X_t	i	$x_{(i)}$	$x_{(N-i+1)}$	$w=F_X(x_{(i)})$	ln$w$	$t = 1- F_X(x_{(N+i+1)})$	ln$t$	S_i *
1969	62,6	32	83,2	82,2	0,4537	-0,7904	0,5622	-0,5760	-86,0798
1970	61,2	33	83,8	81,1	0,4632	-0,7696	0,5795	-0,5457	-85,4897
1971	46,8	34	85,1	80,9	0,4839	-0,7258	0,5826	-0,5403	-84,8258
1972	79	35	87,4	80,2	0,5207	-0,6526	0,5935	-0,5217	-81,0272
1973	96,3	36	87,6	79	0,5239	-0,6465	0,6120	-0,4909	-80,7582
1974	77,6	37	88,1	78,9	0,5319	-0,6314	0,6136	-0,4884	-81,7478
1975	69,3	38	89,2	78	0,5493	-0,5990	0,6273	-0,4663	-79,9020
1976	67,2	39	89,8	77,6	0,5588	-0,5819	0,6334	-0,4567	-79,9734
1977	72,4	40	92,7	76,4	0,6042	-0,5039	0,6513	-0,4288	-73,6792
1978	78	41	93,9	74,8	0,6226	-0,4739	0,6747	-0,3935	-70,2554
1979	141,8	42	96,3	72,4	0,6585	-0,4177	0,7085	-0,3446	-63,2723
1980	100,7	43	97,3	72	0,6731	-0,3958	0,7140	-0,3369	-62,2813
1981	87,4	44	97,8	71,6	0,6803	-0,3852	0,7194	-0,3293	-62,1633
1982	100,2	45	97,9	69,3	0,6817	-0,3831	0,7496	-0,2882	-59,7466
1983	166,9	46	98	68,7	0,6832	-0,3810	0,7572	-0,2781	-59,9829
1984	74,8	47	99,1	68,2	0,6987	-0,3585	0,7634	-0,2699	-58,4502
1985	133,4	48	100,2	67,2	0,7139	-0,3371	0,7756	-0,2541	-56,1626
1986	85,1	49	100,7	66,8	0,7206	-0,3276	0,7804	-0,2480	-55,8340
1987	78,9	50	101	64,2	0,7247	-0,3220	0,8099	-0,2108	-52,7537
1988	76,4	51	104,3	63,6	0,7670	-0,2653	0,8164	-0,2029	-47,2844
1989	64,2	52	110,4	62,6	0,8348	-0,1805	0,8268	-0,1902	-38,1828
1990	53,1	53	110,8	61,2	0,8388	-0,1758	0,8408	-0,1734	-36,6677
1991	112,2	54	112,2	60,6	0,8521	-0,1601	0,8466	-0,1666	-34,9479
1992	110,8	55	114,9	59,9	0,8757	-0,1328	0,8531	-0,1589	-31,7866
1993	82,2	56	117,1	57,3	0,8928	-0,1133	0,8758	-0,1327	-27,3070
1994	88,1	57	118,4	57	0,9021	-0,1030	0,8782	-0,1299	-26,3125
1995	80,9	58	122,8	53,1	0,9292	-0,0734	0,9070	-0,0976	-19,6692
1996	89,8	59	133,4	50,1	0,9709	-0,0295	0,9254	-0,0775	-12,5184
1997	114,9	60	141,7	49,4	0,9870	-0,0130	0,9293	-0,0733	-10,2788
1998	63,6	61	141,8	46,8	0,9872	-0,0129	0,9423	-0,0594	-8,7484
1999	57,3	62	166,9	43,6	0,9994	-0,0006	0,9557	-0,0453	-5,6465
Soma	-	-	-	-	-	-	-	-	**-3876,63**

* $S_i = (2i-1)\{ln\, w + ln\, t\}$

7.4.4 – O Teste de Aderência de Filliben

O teste de aderência de Filliben foi introduzido por Filliben (1975), como um teste de verificação da hipótese nula de normalidade. Posteriormente, o teste de Filliben foi adaptado, para contemplar diversas outras distribuições de probabilidades, sob H_0. Dada uma amostra $\{X_1, X_2, \ldots, X_N\}$, de uma variável aleatória X, e posta a hipótese nula de que a amostra foi extraída de uma população cuja distribuição de probabilidades é $F_X(x)$, a estatística do teste de aderência de Filliben é construída com base no coeficiente de correlação linear r, entre as observações ordenadas em modo crescente $\{x_{(1)}, x_{(2)}, \ldots, x_{(i)}, \ldots x_{(N)}\}$ e os quantis teóricos $\{w_1, w_2, \ldots, w_i, \ldots w_N\}$, os quais são calculados por $w_i = F_X^{-1}(1-q_i)$, onde q_i representa a probabilidade empírica correspondente à ordem de classificação i. Formalmente, a estatística do teste de Filliben é expressa por

$$r = \frac{\sum_{i=1}^{N}\left(x_{(i)} - \overline{x}\right)\left(w_i - \overline{w}\right)}{\sqrt{\sum_{i=1}^{N}\left(x_{(i)} - \overline{x}\right)^2 \sum_{i=1}^{N}\left(w_i - \overline{w}\right)^2}} \qquad (7.31)$$

onde $\overline{x} = \sum_{i=1}^{N} x_{(i)} \Big/ N$ e $\overline{w} = \sum_{i=1}^{N} w_i \Big/ N$.

A idéia essencial do teste de aderência de Filliben é que a eventual existência de uma forte associação linear entre $x_{(i)}$ e w_i, é um indicador de que as observações podem, de fato, ter sido extraídas de uma população cuja distribuição de probabilidades é $F_X(x)$. Portanto, a hipótese nula é H_0: $r = 1$, a qual deve ser testada contra a hipótese alternativa H_1: $r < 1$, tratando-se de um teste unilateral. Nesse caso, a região de rejeição de H_0, a um nível se significância α, é formada pelos valores de r inferiores ao valor crítico $r_{crit,}$ dado pela distribuição de probabilidades da estatística de teste. Assim, se $r < r_{crit,}$, a hipótese nula deve ser rejeitada em favor de H_1.

Na construção da estatística de teste, expressa pela equação 7.31, é implícita a especificação de $F_X(x)$, na forma de $w_i = F_X^{-1}(1 - q_i)$. As probabilidades empíricas q_i, correspondentes às ordens de classificação i, são também denominadas posições de plotagem e variam em conformidade à especificação de $F_X(x)$. Em geral, cada uma das diferentes fórmulas para a posição de plotagem q_i procura obter quantis quase não-enviesados, em relação a cada uma das diferentes distribuições de probabilidade $F_X(x)$. A maioria dessas fórmulas pode ser expressa pela seguinte expressão geral:

$$q_i = \frac{i - a}{N + 1 - 2a} \tag{7.32}$$

onde a varia conforme a especificação de $F_X(x)$. A Tabela 7.9 apresenta um sumário das diferentes fórmulas para a posição de plotagem, indicando também os valores de a correspondentes, bem como as principais motivações de sua proposição, em conformidade com a especificação de $F_X(x)$.

Tabela 7.9 – Fórmulas para o cálculo da posição de plotagem q_i

Denominação	Fórmula	a	Motivação
Weibull	$q_i = \frac{i}{N+1}$	0	Probabilidades de superação não-enviesadas para todas as distribuições.
Blom	$q_i = \frac{i - 3/8}{N + 1/4}$	0,375	Quantis não-enviesados para a distribuição Normal.
Cunnane	$q_i = \frac{i - 0{,}40}{N + 0{,}2}$	0,40	Quantis aproximadamente não-enviesados para quase todas as distribuições.
Gringorten	$q_i = \frac{i - 0{,}44}{N + 0{,}12}$	0,44	Otimizada para a distribuição de Gumbel.

Fonte: adaptada de tabela original de Stedinger et al. (1993).

Uma vez que os quantis w_i variam conforme $F_X(x)$, é evidente que a distribuição de probabilidades da estatística do teste também irá variar, de acordo com a especificação da distribuição $F_X(x)$, sob a hipótese H_0. A Tabela 7.10 apresenta os valores críticos $r_{crit,\,\alpha}$ para o caso em que $F_X(x)$ é especificada como a distribuição Normal, com as probabilidades empíricas q_i calculadas pela fórmula de Blom. Os valores da Tabela 7.10 permanecem válidos para os logaritmos de uma variável Log-Normal.

Tabela 7.10 – Valores críticos $r_{crit,\alpha}$ para a distribuição Normal, com $a = 0{,}375$ na equação 7.32.

N	$\alpha = 0{,}10$	$\alpha = 0{,}05$	$\alpha = 0{,}01$
10	0,9347	0,9180	0,8804
15	0,9506	0,9383	0,9110
20	0,9600	0,9503	0,9290
30	0,9707	0,9639	0,9490
40	0,9767	0,9715	0,9597
50	0,9807	0,9764	0,9664
60	0,9835	0,9799	0,9710
75	0,9865	0,9835	0,9757
100	0,9893	0,9870	0,9812

Fonte: adaptada de tabela original de Stedinger et al. (1993).

A Tabela 7.11 apresenta os valores críticos $r_{crit,\alpha}$ para o caso em que $F_X(x)$ é especificada como a distribuição de Gumbel, para máximos, com as probabilidades empíricas q_i calculadas pela fórmula de Gringorten. Os valores da Tabela 7.11 permanecem válidos para o caso em que $F_X(x)$ é especificada como a distribuição de Weibull de 2 parâmetros.

Tabela 7.11 – Valores críticos $r_{crit,\alpha}$ para a distribuição Gumbel, com $a = 0{,}44$ na equação 7.32.

N	$\alpha = 0{,}10$	$\alpha = 0{,}05$	$\alpha = 0{,}01$
10	0,9260	0,9084	0,8630
20	0,9517	0,9390	0,9060
30	0,9622	0,9526	0,9191
40	0,9689	0,9594	0,9286
50	0,9729	0,9646	0,9389
60	0,9760	0,9685	0,9467
70	0,9787	0,9720	0,9506
80	0,9804	0,9747	0,9525
100	0,9831	0,9779	0,9596

Fonte: adaptada de tabela original de Stedinger et al. (1993).

A Tabela 7.12 apresenta os valores críticos $r_{crit,\alpha}$ para o caso em que $F_X(x)$ é especificada como a distribuição Generalizada de Valores Extremos - GEV, com as probabilidades empíricas q_i calculadas pela fórmula de Cunnane. Os valores críticos da Tabela 7.12 foram obtidos por Chowdhury et al. (1991), mediante simulações de amostras de diferentes tamanhos, extraídas da população de uma variável aleatória GEV, com parâmetro de forma especificado por κ.

Tabela 7.12 – Valores críticos $r_{crit,\alpha}$ para a distribuição GEV, com $a=0,40$ na equação 7.32

α	N	$\kappa=-0,30$	$\kappa=-0,20$	$\kappa=-0,10$	$\kappa=0$	$\kappa=0,10$	$\kappa=0,20$
0,01	5	0,777	0,791	0,805	0,817	0,823	0,825
0,01	10	0,836	0,845	0,856	0,866	0,876	0,882
0,01	20	0,839	0,855	0,878	0,903	0,923	0,932
0,01	30	0,834	0,858	0,89	0,92	0,942	0,953
0,01	50	0,825	0,859	0,902	0,939	0,961	0,97
0,01	100	0,815	0,866	0,92	0,959	0,978	0,985
0,05	5	0,853	0,863	0,869	0,874	0,877	0,88
0,05	10	0,881	0,89	0,9	0,909	0,916	0,92
0,05	20	0,898	0,912	0,926	0,938	0,948	0,953
0,05	30	0,903	0,92	0,937	0,952	0,961	0,967
0,05	50	0,908	0,929	0,95	0,965	0,974	0,979
0,05	100	0,914	0,94	0,963	0,978	0,985	0,989
0,10	5	0,888	0,892	0,896	0,899	0,901	0,903
0,10	10	0,904	0,912	0,92	0,927	0,932	0,936
0,10	20	0,92	0,932	0,943	0,952	0,958	0,962
0,10	30	0,928	0,941	0,953	0,962	0,969	0,973
0,10	50	0,935	0,95	0,963	0,973	0,979	0,982
0,10	100	0,944	0,961	0,974	0,983	0,988	0,991

Vogel e McMartin (1991) empregaram simulações de Monte Carlo para encontrar os valores críticos $r_{crit,\alpha}$ válidos para variáveis aleatórias distribuídas segundo o modelo Pearson Tipo III. De acordo com esses autores, o valor crítico da estatística do teste de Filliben, a um nível de significância $\alpha = 0,05$, pode ser aproximado pela seguinte expressão:

$$r_{crit,\alpha=0,05} \cong exp\left(3,77 - 0,0290\gamma^2 - 0,000670N\right)N^{0,105\gamma-0,758} \qquad (7.33)$$

onde γ denota o coeficiente de assimetria populacional da distribuição de Pearson Tipo III, com posição de plotagem calculada pela fórmula de Blom ($a = 0,375$). A equação 7.33 é válida para $|\gamma| \leq 5$, podendo ser empregada também para os logaritmos de variáveis aleatórias distribuídas segundo um modelo Log-Pearson Tipo III.

Exemplo 7.11 - Refaça o exemplo 7.8, com o teste de aderência de Filliben.
Solução: A Tabela 7.8 apresenta os quantis observados $x_{(i)}$, já ordenados em modo crescente. Os quantis teóricos da distribuição Normal, de média 86,105 e desvio-padrão 24,960, devem ser calculados pela função inversa $\Phi^{-1}(q_i)$, onde q_i denota a probabilidade empírica, tal como calculada pela fórmula de Blom, para a ordem de classificação *i*. Para exemplificar esse cálculo, considere que $i = 1$, resultando, portanto, em $q_1 = 0{,}01004$, para $N = 62$ e $a = 0{,}375$ na equação 7.32. A inversa $\Phi^{-1}(q_i)$ pode ser facilmente calculada pela função estatística INV.NORM do software Microsoft Excel, com argumentos q_i, μ e σ; para o exemplo $q_1 = 0{,}01004$, com $\mu = 86{,}105$ e $\sigma = 24{,}960$, a função INV.NORM retorna o valor $w_1 = 28{,}0769$. Esse cálculo deve ser efetuado para todas as ordens de classificação até $i = N = 62$. A Figura 7.6 apresenta o gráfico entre os quantis teóricos w_i e os observados $x_{(i)}$, assim como a linha correspondente à associação linear entre ambos.

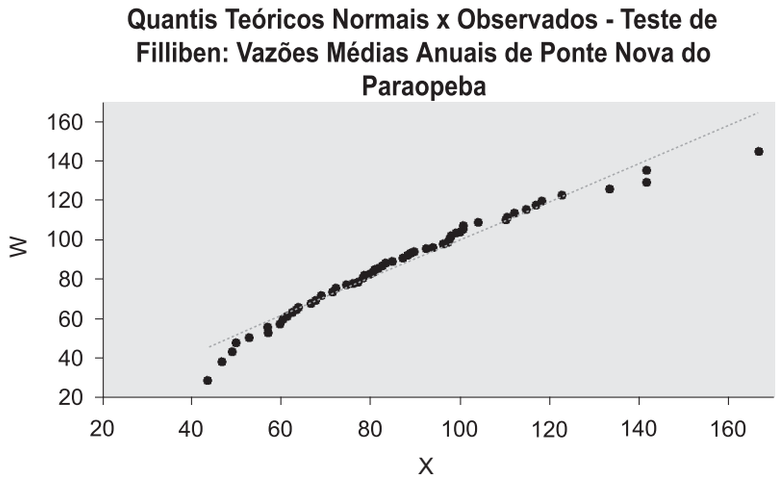


Figura 7.6 – Associação entre os quantis teóricos Normais e os observados no Rio Paraopeba em Ponte Nova do Paraopeba

Em seguida aos cálculos mencionados anteriormente, determina-se a estatística do teste de Filliben pela aplicação da equação 7.31, cujo resultado é $r = 0{,}9798$. Consultando a Tabela 7.10, para $\alpha = 0{,}05$ (teste unilateral), e usando interpolação linear entre os valores de *N* iguais a 60 e 75, vê-se que o valor crítico da estatística de teste é $r_{crit,0,05} = 0{,}9803$, o qual define o limite superior da região de rejeição da hipótese nula. Portanto, como $r < r_{crit,0,05}$, a decisão é a de rejeitar a hipótese H_0 de que o comportamento probabilístico da variável aleatória, em questão, possa ser modelado pela distribuição Normal.

7.4.5 – Comentários a Respeito dos Testes de Aderência

Em geral, os testes de aderência são deficientes em discernir as diferenças entre as freqüências teóricas e empíricas (e/ou quantis teóricos e empíricos), nas caudas inferior e superior das distribuições em análise. No contexto da análise de freqüência de variáveis hidrológicas, essa limitação dos testes de aderência é crítica, uma vez que as amostras são de tamanho relativamente pequeno e que, em geral, o interesse é o de inferir sobre o comportamento da variável aleatória, justamente, nas caudas de sua distribuição de probabilidades. Por exemplo, o teste do χ^2, quando aplicado a variáveis aleatórias contínuas, está sujeito à prescrição de classes, cujo número e amplitude podem interferir profundamente na decisão do teste. No caso do teste de Kolmogorov-Smirnov, a mera observação de sua tabela de valores críticos (Tabela 7.5), revela o conservadorismo do teste no que se refere à decisão de rejeição da hipótese nula. O teste de Anderson-Darling, apesar de constituir uma interessante alternativa aos testes KS e do χ^2, apresenta a limitação de que a distribuição de sua estatística de teste é conhecida apenas para algumas distribuições hipotéticas $F_X(x)$. O teste de Filliben, como alternativa restante, apresenta, como principais vantagens, a simplicidade de construção de sua estatística de teste e algumas comparações favoráveis de seu poder de teste, em relação aos demais, tais como aquelas apontadas por Chowdhury et al. (1991) e Vogel e McMartin (1991). Entretanto, esses mesmos autores demonstram o baixo poder do teste de Filliben, quando se trata de análise de freqüência local, com base em amostras de tamanho relativamente pequeno.

Os testes de aderência, como quaisquer testes de hipóteses, têm o objetivo de verificar se há uma diferença estatisticamente significativa entre as observações e as supostas realizações, caso essas proviessem de uma população hipotética incluída em H_0. Portanto, a eventual decisão de não rejeitar a hipótese nula, a um nível de significância previamente estabelecido, não implica em estabelecer a idéia de que os dados foram, de fato, amostrados a partir da população hipotética. Essa é, por princípio, desconhecida e pode ser uma, entre tantas outras populações incluídas na hipótese alternativa H_1. Por outro lado, as estatísticas dos testes de aderência têm distribuições de probabilidades e, portanto, valores críticos que dependem da distribuição $F_X(x)$, sob H_0, assim como, implicitamente, de suas estimativas paramétricas e dos respectivos erros de estimativa. Com essas considerações em mente, é possível concluir que os resultados de diferentes testes de aderência *não são comparáveis entre si* e, portanto, *não se prestam à seleção do modelo distributivo mais adequado* para uma certa amostra de observações.

7.5 – Teste para Detecção e Identificação de Pontos Atípicos (*outliers*)

Em uma certa amostra de observações, um elemento ou ponto amostral é considerado atípico, ou um *outlier*, do ponto de vista estatístico, quando ele se desvia *significativamente* do conjunto dos outros pontos. Esse desvio pode ter origem em erros de medição ou de processamento, mas também pode ser o produto de causas naturais indeterminadas. Em qualquer caso, a presença de pontos atípicos em uma dada amostra, pode afetar drasticamente o ajuste de uma certa distribuição de probabilidades àqueles dados. No item 2.1.4 do capítulo 2, foi descrito um procedimento de identificação de pontos atípicos, por meio dos quartis amostrais e da amplitude inter-quartis. Este procedimento, embora bastante útil, é meramente exploratório e não constitui, do ponto de vista estatístico, um teste de hipótese, com um nível de significância previamente estabelecido.

Entre os diversos testes de hipóteses para detecção e identificação de pontos atípicos, o teste de Grubbs e Beck, descrito por Grubbs (1950, 1969) e estendido por Grubbs e Beck (1972), encontra-se entre os mais freqüentemente empregados. De acordo com esse teste, as quantidades x_S e x_I definem, respectivamente, os limites superior e inferior, acima e abaixo dos quais, os pontos atípicos, eventualmente presentes em uma amostra, são detectados e identificados. Essas quantidades são definidas pelas seguintes expressões:

$$x_S = exp\left(\overline{x} + k_{N,\alpha} s_X\right) \tag{7.34}$$

e

$$x_I = exp\left(\overline{x} - k_{N,\alpha} s_X\right) \tag{7.35}$$

onde $\overline{x}$ e s_X representam, respectivamente, a média aritmética e o desvio-padrão de uma amostra de tamanho N, de uma variável aleatória X, e $k_{N,\alpha}$ denota o valor crítico da estatística de Grubbs e Beck, para um nível de significância α. Para $100\alpha = 10\%$, Pilon et al. (1985) propõem a seguinte aproximação para o valor crítico da estatística de Grubbs e Beck:

$$k_{N,\alpha=0,10} = -3{,}62201 + 6{,}28446\, N^{1/4} - 2{,}49835\, N^{1/2} + 0{,}491436\, N^{3/4} - 0{,}037911\, N \tag{7.36}$$

De acordo com o teste de Grubbs e Beck, a um nível $\alpha = 0{,}10$ e $k_{N,0,10}$ dado pela equação 7.36, as observações eventualmente superiores a x_S, e/ou inferiores a x_I, estariam se desviando significativamente do conjunto dos dados e deveriam ser consideradas como *outliers*.

Uma vez detectados e identificados os pontos atípicos presentes em uma amostra, a decisão de mantê-los ou expurgá-los da análise de freqüência é matéria de investigação suplementar. Se o exame detalhado de uma certa observação atípica for conclusivo, quanto a caracterizá-la como uma medição incorreta ou sujeita a erros de processamento, ela deve ser certamente expurgada da análise. Entretanto, se a observação atípica resultar de causas naturais, tais como a manifestação de fenômenos extraordinários e diferenciados, em relação ao conjunto dos outros pontos amostrais, a melhor decisão é certamente a de manter os *outliers* na análise de freqüência, buscando encontrar o modelo probabilístico, ou os modelos probabilísticos, que melhor descrevam aquele comportamento observado.

Exercícios

1) Considere o teste da hipótese nula H_0: $p = 0,5$, contra H_1: $p > 0,5$, onde p representa a probabilidade de sucesso em 18 tentativas independentes de um processo de Bernoulli. A decisão é arbitrada como a de rejeitar a hipótese nula, caso a variável aleatória discreta Y, dada pelo número de sucessos em 18 tentativas, seja maior ou igual a 13. Calcule a função poder do teste, denotada por $[1-\beta(p)]$, e ilustre-a graficamente, para diferentes valores de $p > 0,5$.

2) Repita o exercício 1, para a hipótese alternativa H_1: $p \neq 0,5$.

3) Considere as vazões médias anuais do Rio Paraopeba em Ponte Nova do Paraopeba, listadas na Tabela 7.1. Suponha que essa amostra tenha sido extraída de uma população Normal, de desvio-padrão populacional conhecido e igual a σ = 24,960 m^3/s. Teste a hipótese H_0: μ_1= 85 m^3/s, contra a alternativa H_1:μ_2= 90 m^3/s, para $\alpha = 0,05$.

4) Repita o exercício 3, supondo que, desta feita, o desvio-padrão populacional não é conhecido.

5) Refaça o exercício 3 para a hipótese alternativa H_1 : $\mu_1 \neq$ 85 m^3/s.

6) Repita o exercício 5, supondo que, desta feita, o desvio-padrão populacional não é conhecido.

7) Considerando, novamente, as vazões médias anuais do Rio Paraopeba em Ponte Nova do Paraopeba, listadas na Tabela 7.1, separe-as em duas amostras de igual tamanho, uma para o período de 1938 a 1968, e a outra para o período de 1969 a 1999. Supondo tratarem-se de variáveis normais, teste a hipótese de

que, considerados os períodos de 1938-1968 e de 1969-1999, as médias populacionais correspondentes não diferem entre si, em *mais* de 5 m^3/s, para $\alpha = 0,05$.

8) De volta às vazões médias anuais do Rio Paraopeba em Ponte Nova do Paraopeba, listadas na Tabela 7.1, suponha que essa amostra tenha sido extraída de uma população Normal, de média populacional conhecida e igual a $\mu = 86,105$ m^3/s. Teste a hipótese H_0: $\sigma_1 = 25$ m^3/s, contra a alternativa H_1:= $\sigma_1 < 25$ m^3/s, para $\alpha = 0,05$.

9) Repita o exercício 8, supondo que, desta feita, a média populacional não é conhecida.

10) Considerando, novamente, as vazões médias anuais do Rio Paraopeba em Ponte Nova do Paraopeba, listadas na Tabela 7.1, separe-as em duas amostras de igual tamanho, uma para o período de 1938 a 1968, e a outra para o período de 1969 a 1999. Supondo tratarem-se de variáveis normais, teste a hipótese de que, considerados os períodos de 1938-1968 e de 1969-1999, as variâncias populacionais correspondentes não diferem entre si, para $\alpha = 0,05$.

11) Repita o exercício 10, considerando que a hipótese nula, desta feita, é a de que a variância do período de 1938 a 1968, é 10% maior do que a correspondente ao período de 1969-1999.

12) Considere a amostra de alturas diárias de precipitação máxima anual da estação pluviométrica de Ponte Nova do Paraopeba, listadas no Anexo 3. Teste a hipótese nula de que as observações são aleatórias, para $\alpha = 0,05$.

13) Com os dados do exercício 12, teste a hipótese nula de que as observações são independentes, para $\alpha = 0,05$.

14) Com os dados do exercício 12, teste a hipótese nula de que as observações são homogêneas, para $\alpha = 0,05$.

15) Com os dados do exercício 12, teste a hipótese nula de que as observações são estacionárias, para $\alpha = 0,05$.

16) Fez-se a contagem de *E. Coli* em 10 amostras de água. As contagens positivas, expressas em centenas de organismos por 100 ml de água (10^2/100ml), são 17, 21, 25, 23, 17, 26, 24, 19, 21 e 17, com média e a variância amostrais iguais a 21 e 10,6 respectivamente. Suponha que *N* represente o número total dos diferentes

organismos presentes em cada amostra e que *p* represente a fração correspondente ao organismo *E. Coli*. Ajuste uma distribuição Binomial à variável *Y*=centenas de organismos *E. Coli* por 100 ml de água. Verifique a aderência da distribuição Binomial aos dados amostrais, por meio do teste do χ^2, a um nível de significância $\alpha = 0,10$.

17) Para os dados do exercício 12, teste a aderência da distribuição Log-Normal, de 2 parâmetros, por meio do teste do χ^2, a um nível de significância $\alpha = 0,05$.

18) Para os dados do exercício 12, teste a aderência da distribuição Gumbel (máximos), por meio do teste do χ^2, a um nível de significância $\alpha = 0,05$.

19) Para os dados do exercício 12, teste a aderência da distribuição GEV, por meio do teste do χ^2, a um nível de significância $\alpha = 0,05$.

20) Para os dados do exercício 12, teste a aderência da distribuição Exponencial, por meio do teste do χ^2, a um nível de significância $\alpha = 0,05$.

21) Para os dados do exercício 12, teste a aderência da distribuição Pearson Tipo III, por meio do teste do χ^2, a um nível de significância $\alpha = 0,05$.

22) Para os dados do exercício 12, teste a aderência da distribuição Log-Pearson Tipo III, por meio do teste do χ^2, a um nível de significância $\alpha = 0,05$.

23) Para os dados do exercício 12, teste a aderência da distribuição Log-Normal, de 2 parâmetros, por meio do teste de Kolmogorov-Smirnov, a um nível de significância $\alpha = 0,05$.

24) Para os dados do exercício 12, teste a aderência da distribuição Gumbel (máximos), por meio do teste de Kolmogorov-Smirnov, a um nível de significância $\alpha = 0,05$.

25) Para os dados do exercício 12, teste a aderência da distribuição GEV, por meio do teste de Kolmogorov-Smirnov, a um nível de significância $\alpha = 0,05$.

26) Para os dados do exercício 12, teste a aderência da distribuição Exponencial, por meio do teste de Kolmogorov-Smirnov, a um nível de significância $\alpha = 0,05$.

27) Para os dados do exercício 12, teste a aderência da distribuição Pearson Tipo III, por meio do teste de Kolmogorov-Smirnov, a um nível de significância $\alpha = 0,05$.

28) Para os dados do exercício 12, teste a aderência da distribuição Log-Pearson Tipo III, por meio do teste de Kolmogorov-Smirnov, a um nível de significância $\alpha = 0,05$.

29) Para os dados do exercício 12, teste a aderência da distribuição Log-Normal, de 2 parâmetros, por meio do teste de Anderson-Darling, a um nível de significância $\alpha = 0,05$.

30) Para os dados do exercício 12, teste a aderência da distribuição Gumbel (máximos), por meio do teste de Anderson-Darling, a um nível de significância $\alpha = 0,05$.

31) Para os dados do exercício 12, teste a aderência da distribuição Exponencial, por meio do teste de Anderson-Darling, a um nível de significância $\alpha = 0,05$.

32) Para os dados do exercício 12, teste a aderência da distribuição Log-Normal, de 2 parâmetros, por meio do teste de Filliben, a um nível de significância $\alpha = 0,05$.

33) Para os dados do exercício 12, teste a aderência da distribuição Gumbel (máximos), por meio do teste de Filliben, a um nível de significância $\alpha = 0,05$.

34) Para os dados do exercício 12, teste a aderência da distribuição GEV, por meio do teste de Filliben, a um nível de significância $\alpha = 0,05$.

35) Para os dados do exercício 12, teste a aderência da distribuição Pearson Tipo III, por meio do teste de Filliben, a um nível de significância $\alpha = 0,05$.

36) Para os dados do exercício 12, teste a aderência da distribuição Log-Pearson Tipo III, por meio do teste de Filliben, a um nível de significância $\alpha = 0,05$.

37) Para os dados do exercício 12, use o teste de Grubbs e Beck, com $\alpha = 0,10$, para detectar e identificar a presença de pontos atípicos. Compare os resultados com aqueles encontrados por meio do critério da amplitude inter-quartis. Lembre-se que, segundo tal critério, é considerado um ponto atípico superior todo elemento da amostra superior a $(Q_3+1,5AIQ)$ e, analogamente, um ponto atípico inferior é todo e qualquer elemento menor do que $(Q_1-1,5AIQ)$, onde Q_1 e Q_3 representam, respectivamente, o primeiro e o terceiro quartis, e $AIQ= Q_3- Q_1$.

CAPÍTULO 8

ANÁLISE LOCAL DE FREQUÊNCIA DE VARIÁVEIS HIDROLÓGICAS

CAPÍTULO 8 ANÁLISE LOCAL DE FREQÜÊNCIA DE VARIÁVEIS HIDROLÓGICAS

Os sistemas hidrológicos podem ser impactados por eventos extremos, tais como tempestades, grandes cheias e secas. A magnitude de um evento extremo é inversamente relacionada à sua freqüência de ocorrência, uma vez que os eventos muito severos ocorrem com menor freqüência do que os mais moderados. O objetivo da análise de freqüência das variáveis hidrológicas é relacionar a magnitude dos eventos com sua freqüência de ocorrência por meio do uso de uma distribuição de probabilidade. Os resultados da análise de freqüência são necessários para a solução de vários problemas de engenharia, tais como, por exemplo, os projetos de vertedores de barragens, pontes, bueiros e estruturas de controle de cheias, além de problemas que envolvem a estimativa de algum valor característico, tal como a vazão mínima com 7 dias de duração e 10 anos de tempo de retorno.

Do ponto de vista da extensão espacial das informações envolvidas, a análise de freqüência pode ser classificada em local ou regional. Na análise de freqüência local, a definição dos quantis de interesse, ou seja, dos valores da variável hidrológica associados a certas probabilidades de excedência, é realizada utilizando uma única série de registros hidrométricos ou hidrometeorológicos, observados em certa estação fluviométrica, ou pluviométrica, ou climatológica. No caso da análise regional, são utilizados os dados de várias estações de uma região geográfica. Na análise regional, as informações podem ser agrupadas em subconjuntos que apresentam semelhanças fisiográficas, climáticas e/ou estatísticas, as quais justificam a transferência de informações de um local para outro, dentro daquele contexto geográfico. Atualmente, a utilização da análise de freqüência regional não se restringe apenas à estimativa de variáveis hidrológicas em locais que não possuem uma coleta sistemática de informações, mas também para aumentar a confiabilidade das estimativas dos parâmetros de uma distribuição de probabilidades, para identificar a ausência de postos de observação em partes de uma região, bem como para verificar a consistência das séries hidrológicas. A análise de freqüência regional será abordada, em maior profundidade, no capitulo 10.

A análise de freqüência, tanto local como regional, pode ser realizada a partir das chamadas séries de duração anual ou de duração parcial. As séries de duração anual, ou séries anuais, são formadas por um único valor para cada ano de observações, tendo como referência temporal o ano hidrológico ou o ano civil, a depender da variável hidrológica sob análise. As séries de duração parcial consistem

das observações independentes de magnitude superior (ou inferior) a certo valor limiar de referência. Por exemplo, em uma determinada estação fluviométrica, as vazões de pico, extraídas de hidrogramas de cheia convenientemente isolados e independentes entre si, e que superaram um valor limiar especificado, irão formar uma série de duração parcial, a qual pode ter um número de elementos superior ou inferior ao da série anual, a depender da especificação do valor limiar. Existem relações importantes entre a distribuição de probabilidade para máximos anuais e a freqüência de eventos em uma série de duração parcial, as quais serão examinadas ao final do presente capítulo.

As séries constituídas para a análise de freqüência devem ser representativas da variável em questão, não apresentando erros de observação ocasionais e/ou sistemáticos, além de possuir um número suficiente de elementos que permita realizar extrapolações confiáveis. Além disso, na análise de freqüência, é necessário que os dados sejam homogêneos e independentes. A condição de homogeneidade pretende assegurar que todas as observações tenham sido extraídas de uma única população. Para o caso de análise de vazões, por exemplo, pretende-se assegurar que o uso e a ocupação da bacia não tenham sido modificados ou, ainda, que não tenham sido implantadas estruturas hidráulicas que hajam modificado o escoamento natural nos cursos d'água. Por outro lado, a condição de independência procura assegurar que não exista dependência serial entre os elementos que constituem a série. Para efeito de ilustração, considere o caso de uma bacia hipotética, situada na região sudeste do Brasil, sobre a qual abateu-se uma precipitação duradoura, que resultou da formação de uma zona de convergência do Atlântico Sul (ZCAS), dando origem a dois ou mais eventos de cheia. Nesse caso, para garantir a hipótese de independência, apenas a vazão de pico da enchente de maior magnitude deve ser representada na série.

A análise de freqüência pode ser realizada de modo analítico, caso se admita que uma função paramétrica descreva o comportamento probabilístico da variável hidrológica. A análise de freqüência também pode ser efetuada de modo empírico. Nesse último caso, o analista grafa as observações ordenadas contra uma escala de probabilidades e utiliza seu melhor julgamento para determinar a associação entre as magnitudes de ocorrências passadas, ou eventos hipotéticos, e os respectivos tempos de retorno. Na análise de freqüência analítica de variáveis hidrológicas, além dos problemas afetos à inferência estatística, surge ainda a questão de identificação do modelo distributivo a ser empregado. Em algumas aplicações da estatística, nas quais é possível conhecer *a priori* a distribuição populacional da variável aleatória sob análise, o problema se restringe à estimação dos parâmetros populacionais a partir dos dados amostrais. Porém, em se tratando de variáveis hidrológicas, para as quais não se conhece *a priori* a distribuição

populacional, é ainda necessário especificar um certo modelo distributivo, o qual deve ser capaz de descrever o comportamento probabilístico da variável analisada. De fato, várias distribuições têm sido propostas para a modelagem estatística das variáveis hidrológicas, não havendo, todavia, uma distribuição específica consensual que seja capaz de, sob quaisquer condições, descrever o comportamento da variável em foco. Em suma, em uma análise de freqüência típica, o analista procura selecionar, dentre as diversas distribuições candidatas, aquela que parece ser a mais capaz, por um lado, de sintetizar as principais características estatísticas amostrais, e, por outro, de predizer quantis hipotéticos com confiabilidade razoável. De modo resumido, as etapas para análise de freqüência local são as seguintes:

- Optar pela utilização de séries anuais ou séries de duração parcial.
- Avaliar os dados das séries, quanto aos atributos de homogeneidade, independência e representatividade.
- Propor uma ou algumas distribuições teóricas de probabilidade, com a estimativa de seus respectivos parâmetros, quantis e intervalos de confiança, seguida da verificação de aderência à distribuição empírica.
- Realizar a identificação e tratamento de eventuais pontos atípicos, com possível repetição de algumas etapas precedentes.
- Selecionar o modelo distributivo mais apropriado.

Os procedimentos de realização da análise local de freqüência de variáveis hidrológicas serão analisados em detalhes nos itens que se seguem. Inicialmente serão descritos os métodos para a construção de papeis de probabilidade, bem como algumas técnicas utilizadas na estimação das probabilidades de eventos observados, as quais são etapas importantes da análise de freqüência empírica.

8.1 – Análise de Freqüência com Gráficos de Probabilidade

A análise de freqüência hidrológica local pode ser realizada com ou sem a hipótese de que os dados amostrados sejam oriundos da população de uma determinada distribuição de probabilidades. Em não se admitindo tal hipótese, a análise de freqüência se restringe a grafar, ou plotar, os pares constituídos pelas freqüências empíricas e pelas observações devidamente ordenadas.

Nessa análise gráfica, a associação das observações ordenadas às respectivas probabilidades empíricas de excedência, ou aos respectivos tempos de retorno, apresenta consideráveis incertezas que dependem, principalmente, do tamanho e representatividade da amostra. Há ainda a incerteza posta pela questão de extrapolação para tempos de retorno muito superiores ao número de anos de

observações amostrais. Essas incertezas podem ser parcialmente reduzidas (a) pela construção dos chamados papeis "de probabilidades" e/ou (b) pela definição de critérios para estimar as probabilidades empíricas associadas às observações amostrais. Esses tópicos são os objetos dos itens que se seguem.

8.1.1 – Construção de Papéis de Probabilidade

Os gráficos podem ser traçados em escalas aritméticas ou em escalas transformadas. De modo geral, quando se plota, em escala aritmética uma certa função acumulada de probabilidades, $F_X(x)$, versus o valor da variável aleatória X, têm-se um gráfico não linear, tal como exemplificado na Figura 8.1.

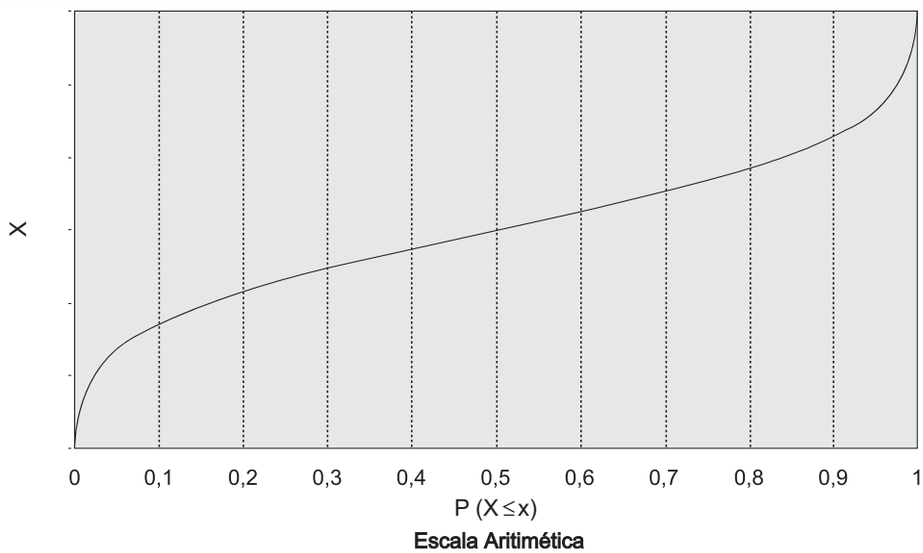


Figura 8.1 – Distribuição Normal em escala aritmética

Os papeis de probabilidade são gráficos para plotagem de observações amostrais e suas respectivas probabilidades empíricas, cujas escalas são previamente transformadas de modo a linearizar a relação entre $F_X(x)$ (ou $[1-F_X(x)]$) e X, tal como ilustrado na Figura 8.2. A escala apropriada para a linearização de uma certa função acumulada de probabilidades $F_X(x)$, descrita por não mais de dois parâmetros, é geralmente construída por meio da *variável padrão ou variável reduzida da distribuição*. A verificação visual de linearidade de um conjunto de dados amostrais, plotados em um papel de probabilidades, pode ser empregada para aceitar ou rejeitar, ainda que empiricamente, a hipótese de aderência a um certo modelo de distribuição de probabilidades.

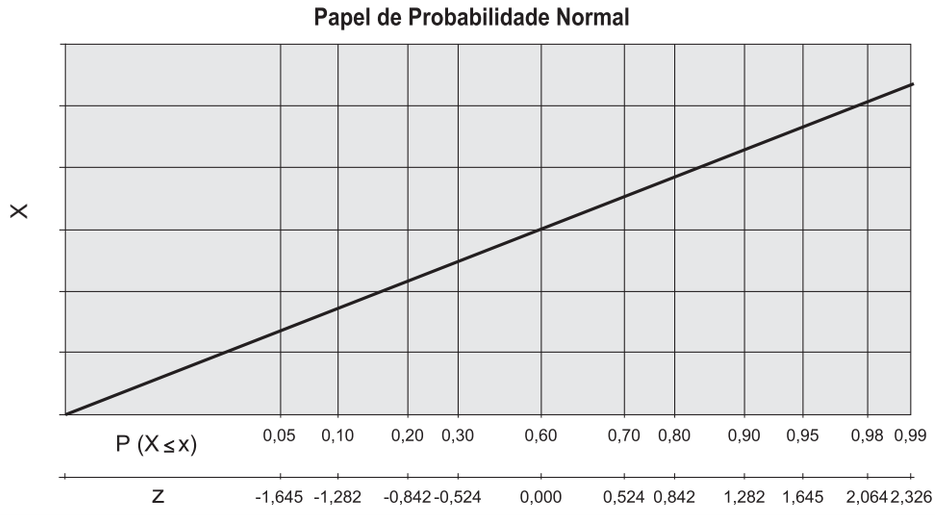


Figura 8.2 – Distribuição Normal no papel de probabilidade Normal

Exemplo 8.1 - Construir o papel de probabilidades da distribuição Normal. Solução: O papel de probabilidade Normal, ou Gaussiano, é construído com base na *distribuição normal padrão* (ver capítulo 5; equações 5.12 e 5.13, e a Tabela 5.1). O eixo das ordenadas, em escala aritmética, representa o valor da variável *X*, como está ilustrado na Figura 8.2. O eixo das abscissas é composto por duas escalas paralelas, uma em escala aritmética que representa os valores da *variável normal central reduzida*, *Z*, enquanto que a outra escala mostra os valores da *distribuição normal padrão*, $\Phi(Z)$, correspondentes aos valores de *Z*, tal como está apresentado na Figura 8.2. Como foi visto no capítulo 5, quando uma variável é normalmente distribuída, o quantil é calculado pela relação $x = \mu_X + Z.\sigma_X$, a qual é a equação de uma reta, onde σ_X é o coeficiente angular e μ_X é o coeficiente linear. A Tabela 8.1 apresenta alguns valores de *Z* e $\Phi(Z)$ da Tabela 5.1, para a construção do eixo das abscissas. Portanto, para construir um gráfico de probabilidades Normal basta plotar a *variável normal central reduzida*, $Z = [(x - \mu)/\sigma]$, a qual está associada a uma probabilidade de não excedência da *distribuição normal padrão* [$\Phi(Z)$], versus *x*. Em geral, omite-se a escala da *variável normal central reduzida*, *Z*.

Tabela 8.1 - Valores de *Z* e $\Phi(Z)$ para construção do papel normal

Z	-1,645	-1,282	-0,842	-0,524	0,000	0,524	0,842	1,282	1,645	2,054	2,326
$P(X \le x) = \Phi(Z)$	0,05	0,10	0,20	0,30	0,50	0,70	0,80	0,90	0,95	0,98	0,99

Exemplo 8.2 - Construir o papel de probabilidades da distribuição exponencial

Solução: A construção de um papel de probabilidade passa pela definição de uma *variável padrão* apropriada para uma função acumulada de probabilidades, que permita a linearização do gráfico $F_X(x)$ versus x. No caso da distribuição exponencial temos que a FAP é dada por:

$$F_X(x) = 1 - exp(-\lambda.x) \tag{8.1}$$

A equação anterior pode ser linearizada por anamorfose logarítmica, ou seja,

$$exp(-\lambda.x) = 1 - F_X(x) \tag{8.2}$$

$$ln[exp(-\lambda.x)] = ln[1 - F_X(x)] \tag{8.3}$$

$$-\lambda.x = ln[1 - F_X(x)] \tag{8.4}$$

$$x = \frac{1}{\lambda}\{-ln[1 - F_X(x)]\}. \tag{8.5}$$

Assim, para uma variável aleatória exponencialmente distribuída, plotando $\{-ln[1 - F_X(x)]\}$, a variável padrão, versus x, obtém-se uma reta com coeficiente angular igual a $1/\lambda$. De forma que, no papel da distribuição exponencial, $\{-ln[1 - F_X(x)]\}$ é grafado nas abscissas e x nas ordenadas, como ilustrado na Figura 8.3.

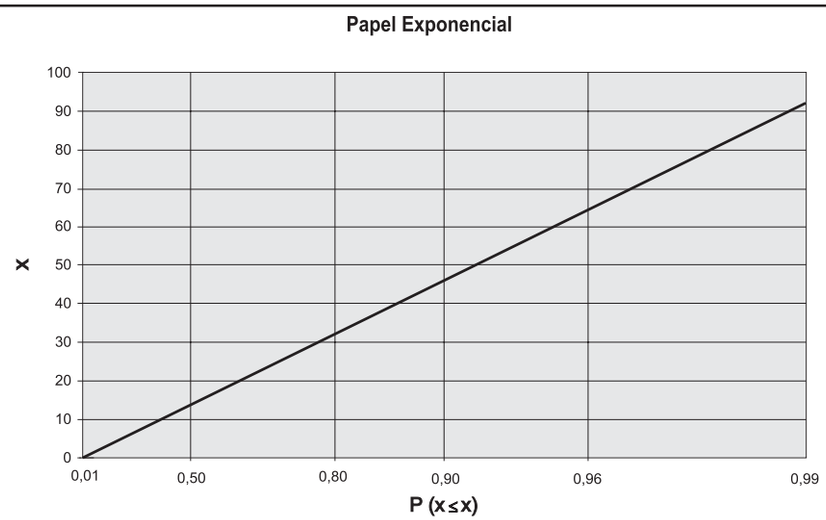


Figura 8.3 – Papel de probabilidade Exponencial

8.1.2 – Posição de Plotagem

Conforme definição anterior, um gráfico de probabilidade associa as magnitudes das observações ordenadas às suas respectivas probabilidades empíricas. No caso de eventos máximos, a estimação da probabilidade empírica de excedência, associada a um certo ponto, é geralmente referida como a determinação da posição de plotagem, a qual pode ser expressa como uma fração entre 0 e 1, ou como uma porcentagem entre 0 e 100. No caso da população, a determinação da posição de plotagem é meramente um problema de determinação da "fração" da população cujos valores são maiores ou iguais ao valor em questão. Assim, para uma variável aleatória de máximos, o menor valor da população terá uma posição de plotagem igual a 1 (um) e o maior valor terá uma posição de plotagem igual a 0 (zero). A definição das posições de plotagem para dados amostrais não é tão direta como no caso populacional, uma vez que nunca haverá certeza de que a amostra contém o maior e/ou o menor valor da população. Assim, para dados amostrais, as posições de plotagem 0 e/ou 1 devem ser evitadas, à exceção dos casos em que existirem informações definitivas acerca dos limites populacionais. Em síntese, para o caso de máximos, uma fórmula para a estimação da posição de plotagem deve especificar a freqüência com que um, entre n valores ordenados de modo decrescente, será igualado ou superado.

A estimação da posição de plotagem de dados hidrológicos requer observações individuais independentes entre si e representativas da população. Gumbel (1958) estabeleceu os seguintes critérios para definição das fórmulas para estimativa das posições de plotagem:

- A posição de plotagem deve ser tal que todas as observações possam ser plotadas;
- A posição de plotagem deve estar compreendida entre $(i\text{-}1)/n$ e i/n , onde i denota a ordem de classificação de uma amostra ordenada de tamanho n;
- No caso de séries anuais, o tempo de retorno de um valor maior ou igual à maior observação (ou menor ou igual à menor observação) deve convergir para n, para valores elevados de n.
- As observações devem ser igualmente espaçadas na escala de freqüências;
- A posição de plotagem deve ser intuitiva, analiticamente simples e fácil de usar.

Várias fórmulas de posição de plotagem têm sido apresentadas na literatura, as quais, geralmente, produzem valores similares no centro da distribuição, mas variam consideravelmente nas caudas. Algumas dessas fórmulas e seus atributos de aplicação encontram-se apresentados na Tabela 8.2.

As expressões apresentadas na Tabela 8.2 são casos particulares da fórmula mais geral introduzida por Cunnane (1978):

$$q_i = \frac{i - a}{n + 1 - 2.a} \tag{8.6}$$

onde a é uma constante que pode ter diferentes valores em conformidade com as hipóteses distributivas. Se $a = 0$, obtém-se a fórmula de Weibull; se $a = 0{,}44$, a fórmula de Gringorten; se $a = 0{,}375$, a fórmula de Blom; se $a = 0{,}50$, a fórmula de Hazen e se $a = 0{,}40$, a fórmula de Cunnane.

De uma forma geral, as séries hidrológicas de valores máximos ou médios anuais são ordenadas de forma decrescente, o que faz com que a posição de plotagem represente a probabilidade da variável X ser maior ou igual a um certo quantil x, ou seja, $P(X \geq x)$.

Entretanto, quando os valores de uma série são ordenados de forma crescente, como na análise de valores mínimos anuais, a posição de plotagem denota a probabilidade de não-excedência, ou seja, a probabilidade da variável X ser menor ou igual a x, ou seja, $P(X \leq x)$.

Tabela 8.2 – Fórmulas para estimativa das posições de plotagem

Fórmula	Autor	Atributos de aplicação
$q_i = \frac{i}{n+1}$	Weibull	Probabilidades de excedência não enviesadas para todas as distribuições
$q_i = \frac{i-0{,}44}{n+0{,}12}$	Gringorten	Usada para quantis das distribuições de Gumbel e GEV
$q_i = \frac{i-0{,}375}{n+0{,}25}$	Blom	Quantis não enviesados para as distribuições Normal e Log-Normal
$q_i = \frac{i-0{,}5}{n}$	Hazen	Usada para quantis da distribuição Gama de 3 parâmetros
$q_i = \frac{i-0{,}40}{n+0{,}20}$	Cunnane	Quantis aproximadamente não enviesados para todas as distribuições

i é posição na amostra ordenada e n é o tamanho da amostra

A estimativa do conjunto das posições de plotagem dos eventos observados, chamada de distribuição empírica, permite a elaboração de um gráfico de probabilidades em conformidade com as seguintes etapas:

a) classificação dos dados em ordem decrescente (análise de máximos) ou crescente (análise de mínimos);
b) cálculo da posição de plotagem por uma das fórmulas apresentadas na Tabela 8.2;

c) seleção do tipo de gráfico, em escala aritmética ou papel de probabilidades apropriado; e
d) plotagem dos pares $[q_i, x_i]$, formando o gráfico da distribuição empírica.

Quando são plotadas as distribuições empíricas dos dados hidrológicos, freqüentemente, um ou dois eventos extremos da amostra parecem ter comportamento atípico em relação aos outros pontos amostrais, como pode ser visto na Figura 8.4. Nessa figura, estão plotadas as alturas diárias de chuva máximas anuais, por ano hidrológico, da estação pluviográfica de Caeté (MG), código 01943010. Foram utilizados 47 máximos anuais na montagem da série (41/42 a 99/2000), sendo que o maior valor é de 210,2mm, registrado em 15/02/1978, e a segunda maior precipitação diária é de 147,1mm. Por meio da fórmula de Gringorten, o tempo de retorno empírico para a precipitação de 210,2mm é de 84 anos, o qual foi estimado pelo inverso da probabilidade de excedência de excedência com $i = 1$ e $n = 47$. Entretanto, observa-se no gráfico da Figura 8.4 que este evento deveria estar associado a um tempo de retorno maior, caso fosse mantida a tendência do alinhamento dos dados amostrais. Trata-se, nesse exemplo específico, de uma observação atípica em relação àquele conjunto particular de observações amostrais. Esse comportamento atípico pode decorrer de diversas razões, entre as quais, podem ser citadas a eventual existência de erros grosseiros de medição ou, ainda, a associação de uma probabilidade empírica incorreta àquela observação específica, como resultado do pequeno tamanho da amostra. Nesse último caso, supondo que a série de Caeté tivesse, digamos, 150 anos de observações e que, ainda assim, a altura de chuva de 210,2 mm continuasse sendo o maior valor amostral, o seu tempo de retorno seria de 268 anos, pela fórmula de Gringorten. Esse exemplo hipotético demonstra a incerteza intrínseca à associação de tempos de retorno empíricos às observações amostrais. O tratamento desses *outliers* é uma questão não resolvida e controvertida, sendo freqüente a ocorrência de tais observações em amostras de dados hidrológicos. De fato, como foi visto anteriormente, a probabilidade de um evento de T anos de tempo de retorno, ocorrer pelo menos uma vez em um período de n anos, é calculada pela relação $1-\left(1-\frac{1}{T}\right)^n$. Assim, a probabilidade de ocorrer pelo menos um evento de 100 anos de tempo de retorno T, durante um período de observação de 30 anos, é de 0,26 ou 26%.

Ao se grafar a distribuição empírica, em certo um papel de probabilidades, os pares $[q_i, x_i]$ podem apresentar uma tendência a se alinharem ao longo de uma reta, a qual pode ser parcimoniosamente extrapolada para tempos de retorno

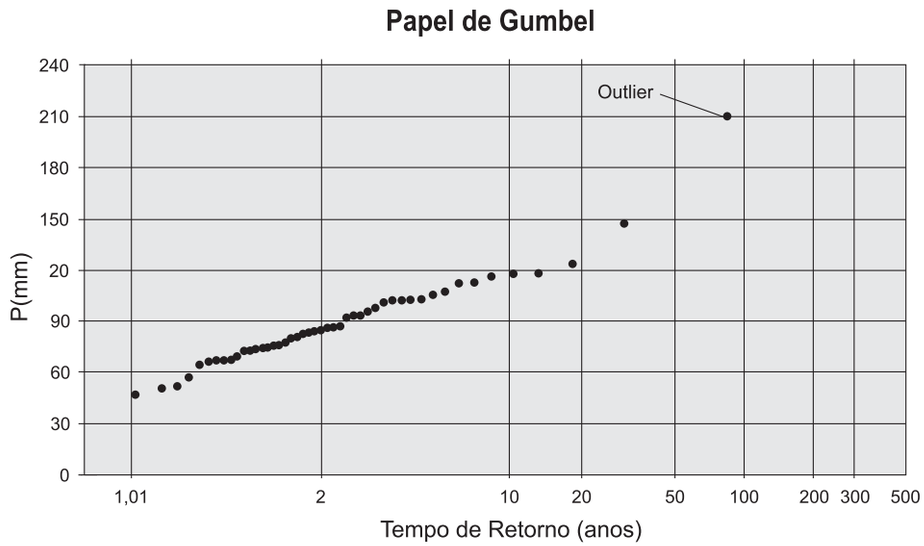


Figura 8.4 – Série com presença de pontos atípicos

superiores àquele associado à maior observação. Todavia, tal situação é pouco freqüente na análise de freqüência de séries hidrológicas. Em geral, os pares $[q_i, x_i]$ apresentam curvaturas e irregularidades que se devem a fatores diversos, entre os quais, os mais importantes são (a) a inadequação do modelo distributivo implicitamente definido pela seleção do papel de probabilidades e (b) problemas de amostragem decorrentes das séries hidrológicas de curta extensão. Tais dificuldades impõem um limite ao uso da análise de freqüência com gráficos de probabilidade, principalmente, quando os quantis de tempos de retorno muito elevados são o principal interesse.

8.1.3 – Posição de Plotagem de Eventos Históricos

As observações sistemáticas de níveis d'água e as medições de vazões nos rios brasileiros tiveram início nos princípios do século XX, com as décadas de 40, 50, 60 e 70 sendo os períodos de maior expansão da rede fluviométrica. Atualmente, em algumas regiões do país, existe um número razoável de séries fluviométricas cujas extensões variam de 30 a 60 anos de observações. Todavia, em alguns locais, é possível obter informações sobre eventos históricos de cheias que ocorreram anteriormente ao início do programa de coleta sistemática de dados hidrológicos. Essas informações podem ser incorporadas à análise de freqüência e obtidas por meio de pesquisas em arquivos públicos e particulares, os quais

guardam documentos de instituições ligadas aos recursos hídricos ou que sofreram as conseqüências das cheias, tais como, institutos históricos e geográficos, museus, empresas relacionadas ao projeto, construção e operação de sistemas de transporte ferroviário, rodoviário e fluvial; arquivos particulares e públicos com fotos e filmes de enchentes; arquivos de jornais e revistas locais, regionais e nacionais; registros paroquiais, entre outras fontes. Informações sobre grandes cheias ocorridas no passado longínquo também podem ser obtidas por meio do uso dos chamados métodos paleohidrológicos. Em síntese, esses métodos fazem uso de técnicas de datação para reconstituir a cronologia de grandes cheias, ocorridas em passado longínquo, ao longo de certo trecho fluvial, a partir das evidências de sua passagem, tais como depósitos de sedimentos e outras marcas deixadas nas seções transversais próximas.

A incorporação de dados históricos nas estimativas de freqüência de vazões de enchentes tem sido objeto de considerável debate na literatura especializada (Hirsch, 1987; Hosking e Wallis, 1986; Sutcliffe, 1987). Do mesmo modo, a utilização dos métodos paleohidrológicos também recebe grande atenção, principalmente nos Estados Unidos (Baker, 1987; Stedinger *et al.*, 1993).

Uma das questões ligadas à utilização de informações sobre eventos históricos está relacionada à estimativa de suas respectivas posições de plotagem. Essa questão pode ser ilustrada pelo diagrama da Figura 8.5. Nessa figura, *h* representa o número de anos do período histórico e *s* denota o período de coleta sistemática de dados, enquanto que *e* indica o número de vazões extremas observadas no período sistemático e *e'* refere-se ao número de eventos extremos do período histórico. O limite Q_o, indicado por uma linha tracejada na Figura 8.5, refere-se à vazão acima da qual as cheias foram consideradas extremas. De acordo com Bayliss e Reed (2001), o limiar Q_o pode ser definido pelos registros históricos e corresponde a um nível de referencia acima do qual as vazões extremas foram percebidas. Ainda segundo Bayliss e Reed (2001), em algumas situações, o limite é determinado por uma cheia extrema recente, sendo razoável supor que o limite seja pouco inferior a esse evento de grandes proporções. Para Hirsch (1987), o limite Q_0 pode ser estabelecido pela vazão que produz destruição e sérios prejuízos econômicos.

Quando as informações históricas são incorporadas como no formato da Figura 8.5, Hirsch (1987), Hirsch e Stedinger (1987) e Salas *et al.* (1994) sugerem a utilização das seguintes fórmulas para cálculo da posição de plotagem:

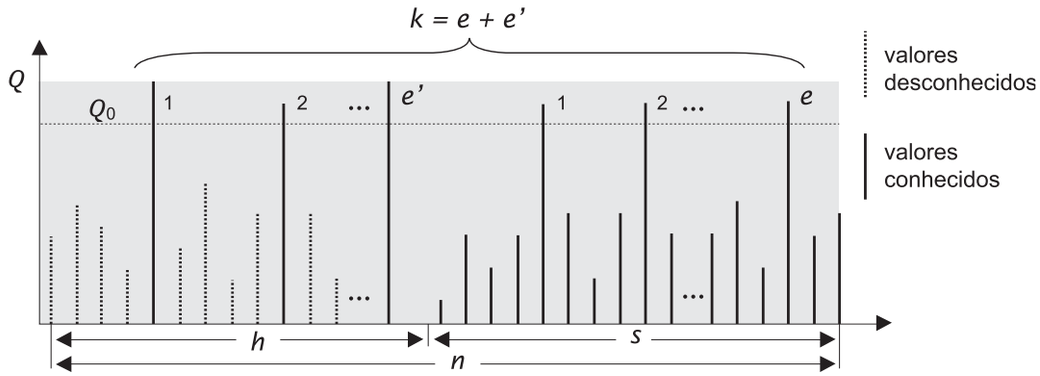


Figura 8.5 – Registros sistemáticos e informações históricas - Modificado de Bayliss e Reed (2001)

$$\begin{cases} q_i = \dfrac{i-a}{k+1-2a}.\dfrac{k}{n}, & i = 1,\cdots,k \quad (a) \\ q_i = \dfrac{k}{n} + \dfrac{n-k}{n}.\dfrac{i-k-a}{s-e+1-2a}, & i = k+1,\cdots,k+s-e \quad (b) \end{cases} \tag{8.7}$$

onde a é a constante de posição de plotagem de Cunnane; n é o número de anos resultante da união das séries de dados sistemáticos e informações históricas, ou seja, $n = h + s$; k representa o número total de vazões extremas, ou seja, o número de vazões que superam o valor limite Q_0 no período combinado, $k = e + e'$.

As equações, contidas no sistema 8.7, permitem a plotagem do gráfico de probabilidade, ou seja, a probabilidade anual de excedência versus a magnitude das vazões. A equação 8.7*a* é aplicada a todas as vazões que estão acima da vazão limite. Em outras palavras, ela é utilizada para todas as vazões das séries histórica e sistemática que estão acima do limite Q_0. A equação 8.7*b* é aplicada às vazões da série sistemática abaixo do limite.

A publicação britânica *Flood Estimation Handbook*, mencionada por Bayliss e Reed (2001), sugere que as maiores vazões da série combinada de informações históricas e dados sistemáticos sejam plotadas, por meio da utilização da fórmula de Gringorten. Isso difere da recomendação da utilização da equação 8.7*a*, com $a = 0{,}44$, apesar das diferenças serem pequenas. As diferenças são significativas somente quando k é muito pequeno em comparação a n, como por exemplo, quando o limite Q_0 é tão alto que poucas vazões, por século, são plotadas.

A vantagem do uso das equações do sistema 8.7 é que elas permitem que os dados sistemáticos abaixo do limite sejam plotados de um modo consistente e compatível aos outros dados. A seguir, apresenta-se um exemplo para ilustrar a utilização de informações históricas na análise de freqüência.

Exemplo 8.3 (Modificado de Bayliss e Reed, 2001) - O rio Avon, em Evesham Worcestershire, na Inglaterra, com uma área de drenagem de 2200 km², é monitorado sistematicamente desde 1937. Por meio de pesquisas em jornais, publicações técnicas e arquivos do Severn River Authority, Bayliss e Reed (2001) selecionaram, a partir de 1822, 15 eventos históricos que foram superiores a 265 m³/s. Plotar em um mesmo papel de probabilidades as distribuições empíricas das séries sistemática (1937-1998) e combinada (1822-1998).

Solução: No caso do período sistemático, os eventos máximos por ano hidrológico foram ordenados de forma decrescente e associados às suas respectivas posições de plotagem, por meio da fórmula de Gringorten. A série sistemática ordenada, as posições de plotagem e os respectivos períodos de retorno calculados estão na Tabela 8.3.

A série combinada, formada pela série sistemática acrescida dos 15 eventos históricos, também foi ordenada de forma decrescente. As posições de plotagem foram calculadas com as equações do sistema 8.7. A equação 8.7*a* foi aplicada a todos os eventos superiores a 265 m³/s, que é o limite definido por Bayliss e Reed (2001) para incorporação de vazões históricas extremas. A equação 8.7*b* definiu a posição de plotagem das vazões da série sistemática inferiores ao limite de 265 m³/s. Os parâmetros das equações são: n = 177 anos (1998 – 1882 +1); h = 115 anos ; s = 62 anos; k = 19 (eventos superiores a 265 m³/s); e = 4 (eventos do período sistemático superiores a 265 m³/s); e' = 15 (eventos do período histórico superiores a 265 m³/s) e a = 0,44.

Os resultados com as posições de plotagem e tempos de retorno calculados também estão apresentados na Tabela 8.3. As distribuições empíricas das séries sistemática e combinada foram grafadas em um papel de probabilidades de Gumbel, conforme ilustrado na Figura 8.6. Por essa figura observa-se o efeito da inclusão da informação histórica. Além das mudanças óbvias nos pontos acima da vazão limite, outro efeito perceptível ocorre sobre os pontos da série combinada logo abaixo do limite. Os tempos de retorno desses pontos são levemente menores do que os calculados sem a utilização da informação histórica.

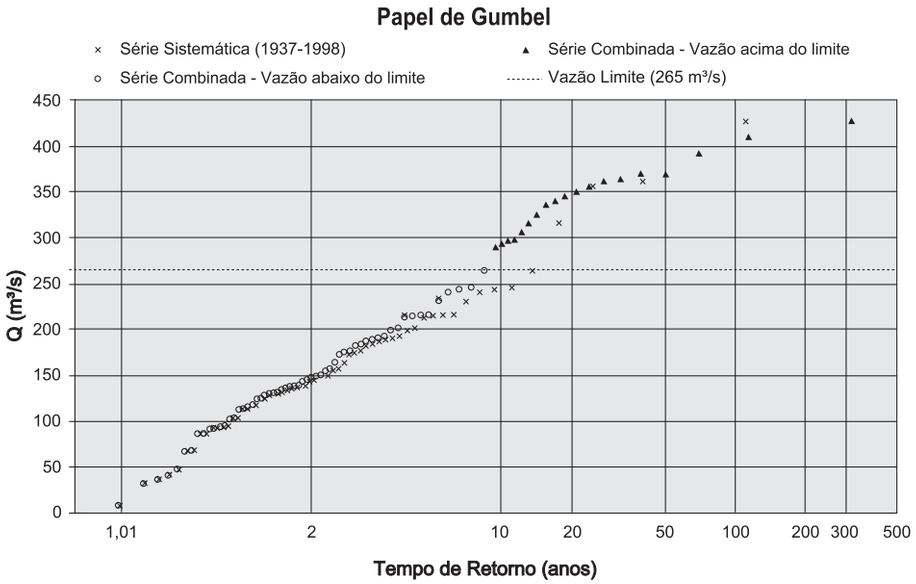


Figura 8.6 – Distribuições empíricas sistemática e combinada

Tabela 8.3 – Cálculo das posições de plotagem das séries sistemática e combinada

Série Sistemática (1937-1998)					Série Combinada (1822-1998)				
Ordem	AH	Q(m³/s)	Gring.	T(anos)	Ordem	AH	Q(m³/s)	Eq. (8.7)	T(anos)
1	1997	427,000	0,0090148	110,93	1	1997	427	0,003144	318,07
2	1967	361,909	0,0251127	39,82	2	1900*	410	0,008758	114,18
3	1946	356,187	0,0412106	24,27	3	1848*	392	0,014373	69,58
4	1939	316,213	0,0573084	17,45	4	1852*	370	0,019987	50,03
5	1981	264,091	0,0734063	13,62	5	1829*	370	0,025601	39,06
6	1959	245,633	0,0895042	11,17	6	1882*	364	0,031215	32,04
7	1958	243,687	0,1056021	9,47	7	1967	362	0,03683	27,15
8	1938	240,382	0,1216999	8,22	8	1946	356	0,042444	23,56
9	1979	230,596	0,1377978	7,26	9	1923*	350	0,048058	20,81
10	1980	215,716	0,1538957	6,50	10	1875*	345	0,053672	18,63
11	1960	215,279	0,1699936	5,88	11	1931*	340	0,059287	16,87
12	1978	214,387	0,1860914	5,37	12	1888*	336	0,064901	15,41
13	1992	212,600	0,2021893	4,95	13	1874*	325	0,070515	14,18
14	1942	201,259	0,2182872	4,58	14	1939	316	0,076129	13,14
15	1968	198,944	0,2343851	4,27	15	1935*	306	0,081744	12,23
16	1987	192,414	0,2504829	3,99	16	1932*	298	0,087358	11,45
17	1954	190,617	0,2665808	3,75	17	1878*	296	0,092972	10,76
18	1971	188,904	0,2826787	3,54	18	1885*	293	0,098586	10,14
19	1940	187,123	0,2987766	3,35	19	1895*	290	0,104201	9,60
20	1941	183,657	0,3148744	3,18	20	1981	264,091	0,115946	8,62
21	1950	181,934	0,3309723	3,02	21	1959	245,633	0,131304	7,62
22	1976	176,653	0,3470702	2,88	22	1958	243,687	0,146663	6,82
23	1984	174,533	0,3631681	2,75	23	1938	240,382	0,162022	6,17
24	1974	172,612	0,3792659	2,64	24	1979	230,596	0,177381	5,64
25	1989	163,307	0,3953638	2,53	25	1980	215,716	0,19274	5,19
26	1970	157,400	0,4114617	2,43	26	1960	215,279	0,208099	4,81
27	1982	155,035	0,4275596	2,34	27	1978	214,387	0,223457	4,48
28	1998	149,700	0,4436574	2,25	28	1992	212,600	0,238816	4,19
29	1949	148,908	0,4597553	2,18	29	1942	201,259	0,254175	3,93

* Eventos históricos

Tabela 8.3 – Continuação

Série Sistemática (1937-1998)					Série Combinada (1822-1998)				
Ordem	AH	Q(m³/s)	Gring.	T(anos)	Ordem	AH	Q(m³/s)	Eq.(8.7)	T(anos)
30	1965	148,443	0,4758532	2,10	30	1968	198,944	0,269534	3,71
31	1985	145,447	0,4919511	2,03	31	1987	192,414	0,284893	3,51
32	1993	143,400	0,5080489	1,97	32	1954	190,617	0,300252	3,33
33	1991	138,800	0,5241468	1,91	33	1971	188,904	0,31561	3,17
34	1956	138,782	0,5402447	1,85	34	1940	187,123	0,330969	3,02
35	1957	137,556	0,5563426	1,80	35	1941	183,657	0,346328	2,89
36	1973	135,722	0,5724404	1,75	36	1950	181,934	0,361687	2,76
37	1990	134,179	0,5885383	1,70	37	1976	176,653	0,377046	2,65
38	1966	131,490	0,6046362	1,65	38	1984	174,533	0,392405	2,55
39	1952	130,432	0,6207341	1,61	39	1974	172,612	0,407763	2,45
40	1951	130,432	0,6368319	1,57	40	1989	163,307	0,423122	2,36
41	1986	128,578	0,6529298	1,53	41	1970	157,400	0,438481	2,28
42	1994	124,300	0,6690277	1,49	42	1982	155,035	0,45384	2,20
43	1977	123,646	0,6851256	1,46	43	1998	149,700	0,469199	2,13
44	1963	117,402	0,7012234	1,43	44	1949	148,908	0,484558	2,06
45	1988	115,592	0,7173213	1,39	45	1965	148,443	0,499916	2,00
46	1995	113,900	0,7334192	1,36	46	1985	145,447	0,515275	1,94
47	1972	112,565	0,7495171	1,33	47	1993	143,400	0,530634	1,88
48	1944	103,298	0,7656149	1,31	48	1991	138,800	0,545993	1,83
49	1983	102,542	0,7817128	1,28	49	1956	138,782	0,561352	1,78
50	1969	94,897	0,7978107	1,25	50	1957	137,556	0,576711	1,73
51	1955	93,851	0,8139086	1,23	51	1973	135,722	0,592069	1,69
52	1961	92,290	0,8300064	1,20	52	1990	134,179	0,607428	1,65
53	1948	91,377	0,8461043	1,18	53	1966	131,490	0,622787	1,61
54	1953	86,275	0,8622022	1,16	54	1952	130,432	0,638146	1,57
55	1945	86,275	0,8783001	1,14	55	1951	130,432	0,653505	1,53
56	1962	67,913	0,8943979	1,12	56	1986	128,578	0,668864	1,50
57	1947	67,110	0,9104958	1,10	57	1994	124,300	0,684222	1,46
58	1937	47,021	0,9265937	1,08	58	1977	123,646	0,699581	1,43
59	1964	41,032	0,9426916	1,06	59	1963	117,402	0,71494	1,40
60	1975	35,937	0,9587894	1,04	60	1988	115,592	0,730299	1,37
61	1996	31,880	0,9748873	1,03	61	1995	113,900	0,745658	1,34
62	1943			1,01	62	1972	112,565	0,761017	1,31
					63	1944	103,298	0,776375	1,29
					64	1983	102,542	0,791734	1,26
					65	1969	94,897	0,807093	1,24
					66	1955	93,851	0,822452	1,22
					67	1961	92,290	0,837811	1,19
					68	1948	91,377	0,85317	1,17
					69	1953	86,275	0,868528	1,15
					70	1945	86,275	0,883887	1,13
					71	1962	67,913	0,899246	1,11
					72	1947	67,110	0,914605	1,09
					73	1937	47,021	0,929964	1,08
					74	1964	41,032	0,945323	1,06
					75	1975	35,937	0,960681	1,04
					76	1996	31,880	0,97604	1,02
					77	1943	7,574	0,991399	1,01

* Eventos históricos

8.2 – Análise de Freqüência Analítica

A análise convencional de freqüência de realizações de uma variável aleatória, da qual se conhece uma amostra e a distribuição de probabilidades da população de onde a amostra foi retirada, consiste em estimar os parâmetros populacionais a partir dos dados observados e, em seguida, estimar os quantis para a probabilidade desejada. No caso de eventos máximos (e/ou mínimos) de variáveis hidrológicas, a distribuição de probabilidades da população não é conhecida e tem-se somente uma amostra de dados observados. Esse fato complicador leva à proposição de modelos probabilísticos, ou sejam funções paramétricas de probabilidade, as quais, em função de suas características de assimetria e da eventual existência de limites superiores (e/ou inferiores) no domínio de definição da variável aleatória, se atribuem propriedades de modelarem os fenômenos hidrológicos. Muitas distribuições têm sido propostas para a modelação estatística dos valores máximos anuais de variáveis hidrológicas ou hidrometeorológicas, mas não há uma distribuição específica consensual que seja capaz de, sob quaisquer condições, descrever o comportamento da variável em foco. Portanto, em uma análise típica, cabe ao analista selecionar, entre as diversas distribuições candidatas, aquela que parece mais apropriada à modelação dos dados amostrais. Os procedimentos típicos de uma análise de freqüência local de séries anuais estão descritos nos próximos subitens.

a) Avaliação dos dados amostrais

A qualidade e a aplicabilidade da análise de freqüência dependem diretamente dos dados utilizados para estimação de seus parâmetros. Desse modo, é um fato reconhecido que, por mais sofisticado que seja, a qualidade de um modelo estocástico jamais superará a dos dados disponíveis para a estimação de seus parâmetros. Nesse sentido, cabe ao hidrólogo julgar a qualidade dos registros hidrológicos disponíveis para dar prosseguimento à analise de freqüência.

É um pressuposto da análise de freqüência convencional que a amostra de dados disponível seja uma entre um número infinito de outras amostras possíveis, as quais representariam realizações, com igual chance de sorteio, de uma única população. Também são pressupostos da análise de freqüência convencional que os dados hidrológicos devem satisfazer as condições de independência, estacionariedade e representatividade. De modo sintético, pode-se dizer que os eventos são considerados independentes quando não há correlação entre os valores da série. Sendo assim, a independência significa a inexistência de correlação entre um registro de um dado ano e o registro posterior (ou anterior), considerados todos os anos disponíveis. Por outro lado, uma série de dados hidrológicos é dita

estacionária quando não ocorrem modificações nas características estatísticas de sua série ao longo do tempo. A análise de freqüência de séries hidrológicas não estacionárias e, por conseguinte, a estimação de parâmetros e quantis com tendências ou variações temporais são objetos de investigações recentes [e.g: Cox et al. (2002) e Clarke (2002)] e não serão aqui consideradas. Em termos da análise de freqüência convencional, dados não estacionários devem ser analisados em sub-séries homogêneas ou ajustados de modo a corrigir as heterogeneidades encontradas. As causas principais de possíveis não-estacionariedades em uma série hidrológica ou hidrometeorológica são: a relocação das estações de observação, a construção de barragens a montante, a urbanização ou o desmatamento das bacias, as eventuais modificações do leito fluvial, a ocorrência de cheias catastróficas, além, evidentemente, de mudanças climáticas.

A confiabilidade das estimativas dos parâmetros de uma dada distribuição de probabilidade está intrinsecamente ligada ao tamanho da amostra e à sua representatividade. Os dados da amostra devem ser representativos da variabilidade inerente ao processo natural ou experimento em foco. Em se tratando de variáveis hidrológicas ou hidrometeorológicas, uma amostra, obtida ao longo de um período predominantemente seco (ou úmido), irá certamente distorcer os resultados da análise, produzindo, em conseqüência, estimativas tendenciosas dos parâmetros populacionais. Por outro lado, uma amostra de dados possui propriedades estatísticas apenas similares às da população; elas serão idênticas se e somente se toda a população tiver sido amostrada. Yevjevich (1972) resume a questão afirmando que tanto a presença de erros sistemáticos em uma amostra, os quais podem ser provenientes de problemas de processamento e medição, de heterogeneidades e falta de representatividade, quanto os erros aleatórios, esses inerentes às naturais flutuações amostrais em torno de valores populacionais, podem produzir grandes incertezas quanto às estimativas de parâmetros estatísticos, realizadas a partir de amostras de tamanho relativamente pequeno. De qualquer modo, é um pressuposto básico dos métodos de inferência estatística a inexistência de erros sistemáticos, atribuindo somente às flutuações amostrais as diferenças entre estimativas e valores populacionais.

Benson (1960), utilizou uma série sintética de 1000 anos de vazões máximas e demonstrou que para se estimar uma cheia de 50 anos são necessárias amostras de pelo menos 39 anos, para que as estimativas ficassem na faixa de 24% do valor correto, em 95% dos casos. Caso a confiança de acerto decresça para 80%, o período mínimo de dados necessário seria de 15 anos. É freqüente encontrar na literatura referências à consideração que de uma série de máximos anuais de n valores pode-se estimar, com alguma confiabilidade, quantis com tempos de retorno de até $2n$. Watt et al. (1988), editor do guia "*Hydrology of*

Floods in Canada - A Guide to Planning and Design", preparado para o Conselho Nacional de Pesquisas do Canadá, relacionam o tamanho da amostra ao tipo de abordagem a ser tomada pela análise de freqüência de vazões máximas. Nesse guia, a análise de freqüência local de vazões máximas anuais é recomendada apenas para as amostras com mais de 10 anos de dados e para estimativas de quantis com tempos de retorno no máximo menores do que quatro vezes o tamanho da série. Apesar de existirem outras formas de avaliar qualitativamente a aplicabilidade da análise de freqüência, não se pode negar a importância do tamanho da amostra como uma forma de avaliação qualitativa dos estimadores amostrais e quantis, uma vez que a variância de todos eles é inversamente proporcional a alguma potência do tamanho da amostra.

Testes estatísticos paramétricos e não paramétricos podem ser usados como ferramentas auxiliares na identificação da presença de dependência e heterogeneidade serial. Os testes paramétricos são fundamentados em suposições distributivas mais severas do que as exigidas por testes não paramétricos similares. Geralmente, em sua formulação, os testes paramétricos são baseados na suposição de uma distribuição de probabilidades específica para os dados amostrais. Conforme visto no capítulo 7, os testes não paramétricos, também chamados de "testes livres de distribuição", não exigem a premissa de uma distribuição de probabilidade específica e têm suas estatísticas de decisão construídas com base em características indiretas dos dados originais. Portanto, tendo como motivação não assumir *a priori* compromissos com as características distributivas populacionais durante a etapa de verificação de dados amostrais, é claramente recomendável o uso dos testes não paramétricos, discutidos no capítulo 7, para a identificação da eventual presença de heterogeneidade e dependência serial na amostra. Cabe esclarecer, entretanto, que, embora os testes estatísticos sejam válidos para pequenas amostras e sob situações diversas, eles devem ser vistos apenas como indicadores, pois não são constituem, por si, argumentos suficientemente fortes para se abandonar uma amostra caso indiquem, por exemplo, a presença de dependência serial entre seus dados. Nesses casos, deve-se procurar uma evidência física que justifique o resultado do teste.

Ainda na etapa de verificação inicial de dados, deve-se lembrar que alguns cuidados devem ser tomados durante a seleção dos eventos de modo a assegurar a independência serial da amostra. Em regiões com sazonalidade muito acentuada, a seleção de eventos para compor uma dada amostra deve ser feita de forma diferenciada para vazões máximas e mínimas anuais; por exemplo, no estado de Minas Gerais, como de resto em grande parte da região sudeste do Brasil, a estação chuvosa vai de Outubro a Março, com grande possibilidade de ocorrência de eventos máximos em Dezembro. Neste caso, as vazões máximas anuais devem

ser individualizadas por ano hidrológico, o qual corresponde a um período fixo de 12 meses, a começar no início do período chuvoso (Outubro) e terminar no final da estação seca (Setembro). Mesmo em regiões com sazonalidade não tão evidente como o sudeste brasileiro, tais como o sul de Santa Catarina e grande parte do Rio Grande do Sul, o ano hidrológico de Maio a Abril deve ser empregado para a seleção de eventos. Por outro lado, no caso da seleção da amostra de vazões mínimas anuais, a abordagem anterior merece restrições, já que uma estiagem prolongada pode fazer com que valores dependentes sejam escolhidos. Neste caso, os períodos anuais devem ser limitados pelos meses mais chuvosos.

b) Definição da distribuição de probabilidade, estimação de seus parâmetros e a verificação de aderência à distribuição empírica.

Existe um conjunto não muito extenso de funções de distribuição de probabilidades que podem ser empregadas para a modelação de eventos máximos anuais de variáveis hidrológicas e hidrometeorológicas. Dentro desse conjunto, pode-se distinguir as distribuições oriundas da teoria clássica de valores, quais sejam as distribuições Gumbel, Fréchet, Weibull e a Generalizada de Valores Extremos (GEV), e aquelas ditas não-extremais, entre as quais as de maior uso são: as distribuições Exponencial e sua forma mais geral que é a Generalizada de Pareto, Pearson III, Log-Pearson III e Log-Normal de 2 parâmetros. Embora a adequação destas distribuições candidatas dependa de critérios variados, incluindo alguns de caráter subjetivo, talvez o atributo mais desejável seja a capacidade dessas distribuições de reproduzir algumas características amostrais relevantes. Apresenta-se, a seguir, as principais considerações a levar em conta quando da seleção de um modelo probabilístico local.

No que concerne às distribuições limitadas à direita, é um fato que algumas quantidades físicas possuem limites superiores inerentemente definidos; é o caso, por exemplo, da concentração de oxigênio dissolvido em um corpo d'água, limitado fisicamente em um valor entre 9 a 10 mg/l, a depender da temperatura ambiente. Outras quantidades podem igualmente possuir um limite superior, muito embora, tal limite possa não ser conhecido *a priori*, fato decorrente da insuficiente compreensão e/ou quantificação de todos os processos físicos causais envolvidos. A esse respeito, é conhecida a controvérsia quanto à existência da Precipitação Máxima Provável (PMP), originalmente formulada como um limite superior de produção de precipitação pelo ar atmosférico. Admitindo-se que esse limite exista de fato, é consensual que sua determinação fica comprometida pela insuficiente quantificação da variabilidade espaço-temporal das variáveis que lhe dão origem. Entretanto, pode-se conjecturar que seria fisicamente impossível a ocorrência de uma vazão, digamos de 100.000 m^3/s, em uma pequena bacia hidrográfica, por

exemplo, da ordem de 100 km^2 de área de drenagem. Por essa razão, alguns pesquisadores, como Boughton (1980) e Laursen (1983), recomendam que somente distribuições limitadas superiormente devem ser usadas para modelar variáveis com essas características. Hosking e Wallis (1997) consideram errônea essa recomendação e sustentam que, se o objetivo da análise de freqüência é o de estimar o quantil de tempo de retorno de 100 anos, é irrelevante considerar como "fisicamente impossível" a ocorrência do quantil de 100.000 anos. Acrescentam que impor um limite superior ao modelo probabilístico pode comprometer a obtenção de boas estimativas de quantis para os tempos de retorno que realmente interessam. Hosking e Wallis (1997) concluem afirmando que, ao se empregar uma distribuição ilimitada superiormente, as premissas implícitas são (i) que o limite superior não é conhecido e nem pode ser estimado com a precisão necessária e (ii) que no intervalo de tempos de retorno de interesse do estudo, a distribuição de probabilidades da população pode ser melhor aproximada por uma função ilimitada do que por uma que possua um limite superior. Evidentemente, quando existem evidências empíricas que a distribuição populacional possui um limite superior, ela deve ser aproximada por uma distribuição limitada superiormente. Seria o caso, por exemplo, do ajuste da distribuição Generalizada de Valores Extremos a uma certa amostra, cuja tendência de possuir um limite superior estaria refletida na estimativa de um valor positivo para o parâmetro de forma κ.

O chamado "peso" da cauda superior de uma função distribuição de probabilidades determina a intensidade com que os quantis aumentam, à medida que os tempos de retorno tendem para valores muito elevados. Em outras palavras, o peso da cauda superior é proporcional às probabilidades de excedência associadas a quantis elevados e é reflexo da intensidade com que a função densidade $f_X(x)$ decresce quando x tende para valores muito elevados. Os pesos das caudas superiores de algumas das principais funções de distribuição de probabilidades encontram-se relativizados na Tabela 8.4.

Para a maioria das aplicações envolvendo variáveis hidrológicas/ hidrometeorológicas, a correta prescrição da cauda superior de uma distribuição de probabilidades é de importância fundamental e, em muitos casos, representa a motivação primeira da análise de freqüência. Entretanto, os tamanhos das amostras disponíveis para essas aplicações são invariavelmente insuficientes para se determinar, com alguma precisão, a forma da cauda superior do modelo probabilístico. Segundo Hosking e Wallis (1997), não havendo razões suficientes para se recomendar o emprego exclusivo de somente um tipo de cauda superior, é aconselhável utilizar um grande conjunto de distribuições candidatas cujos pesos de suas caudas superiores se estendam por um amplo espectro.

Tabela 8.4 – Pesos das caudas superiores de algumas distribuições de probabilidade

Cauda Superior	Forma de $f(x)$ para valores elevados de x	Distribuição
Pesada	x^{-A}	Generalizada de valores extremos, generalizada de Pareto e Logística generalizada com parâmetro de forma $k<0$.
↑	$x^{-A\ln x}$	Lognormal
	$\exp(-x^{A})$ $0 < A < 1$	Weibull com parâmetro de forma $\lambda<1$.
	$x^{A}\exp(-Bx)$	Pearson tipo III com assimetria positiva.
	$\exp(-x)$	Exponencial, Gumbel.
↓	$\exp(-x^{A})$ $A > 1$	Weibull com parâmetro de forma $\lambda<1$.
Leve	Existe um Limite superior	Generalizada de valores extremos, generalizada de Pareto e Logística generalizada com parâmetro de forma $k > 0$; e Pearson tipo III com assimetria negativa.

A e *B* representam constantes positivas. (adap. de Hosking e Wallis, 1997, p. 75)

Considerações semelhantes às anteriores se aplicam à cauda inferior, ou seja, é necessário utilizar um conjunto razoável de distribuições candidatas cujos pesos de suas caudas inferiores se estendam por um amplo espectro. Entretanto, se o interesse do estudo encontra-se centrado em se prescrever a melhor aproximação da cauda superior, a forma da cauda inferior é irrelevante. Em alguns casos, conforme enfatizado no relatório "*Estimating Probabilities of Extreme Floods, Methods and Recommended Research*" do *National Research Council* (*NRC*, 1987), a presença de "*outliers*" baixos em uma dada amostra pode inclusive vir a comprometer a correta estimação das características da cauda superior.

As reflexões sobre o limite superior também são aplicáveis ao limite inferior. Contudo, diferentemente do limite superior, o inferior é, em geral, conhecido ou pode ser igualado a zero; algumas distribuições, como a Generalizada de Pareto, permitem, com facilidade, o ajuste do parâmetro de posição, quando se conhece ou se prescreve o limite inferior. Hosking e Wallis (1997) ressaltam, entretanto, que, em diversos casos, a prescrição de limite inferior nulo é inútil e que melhores resultados podem ser obtidos sem nenhuma prescrição *a priori*. Exemplificam afirmando que os totais anuais de precipitação em regiões úmidas, apesar de números positivos, são muito superiores a zero; para esse exemplo, uma distribuição de probabilidades realista deve ter um limite inferior muito maior do que zero.

As distribuições oriundas da teoria clássica de valores extremos (Gumbe1, 1958), quais sejam os modelos Gumbel, Fréchet e Weibull, são as únicas para as quais existem justificativas teóricas para seu emprego na modelação de valores máximos (ou mínimos) de dados empíricos. Por exemplo, o modelo de valores extremos do tipo I para máximos (EV1 ou Gumbel) é a distribuição assintótica do maior valor de uma seqüência ilimitada de variáveis aleatórias independentes e igualmente distribuidas (*iid*), a distribuição das quais possui uma cauda superior do tipo exponencial. Analogamente, a distribuição do tipo II para máximos (EV2 ou Fréchet) relaciona-se a variáveis *iid* com cauda superior do tipo polinomial, enquanto a distribuição do tipo III (EV3 ou Weibull) refere-se a variáveis *iid* que possuem um limite superior finito. Sob as premissas da teoria de valores extremos, por exemplo, a distribuição de probabilidades das vazões médias diárias máximas anuais de uma certa bacia hidrográfica, depende da distribuição inicial única dos valores diários considerados independentes. A maior objeção ao uso das distribuições oriundas da teoria de valores extremos em hidrologia refere-se à premissa de variáveis iniciais *iid*, a qual muito dificilmente é satisfeita por variáveis hidrológicas ou hidrometeorológicas. A esse respeito, transcreve-se o seguinte comentário escrito por Perichi e Rodríguez-Iturbe (1985, p. 515) :

"*Presumir que duas vazões médias diárias, observadas digamos no dia 15 de maio e em 20 de Dezembro, são variáveis aleatórias identicamente distribuídas, é uma clara violação da realidade hidrológica. Essa premissa 'regulariza' as distribuições históricas iniciais afirmando não só que elas são do mesmo tipo, mas também que elas possuem os mesmos parâmetros (e.g. média e variância) para qualquer dia do ano. Sob essa premissa, não se pode admitir o fato que se uma mesma vazão média diária foi observada em dois dias diferentes, é mais provável que aquele que possui a maior variância produzirá cheias maiores do que aquele de menor variância. A realidade hidrológica é que a combinação da média e da variância de um dado mês faz com que alguns meses do ano sejam mais suscetíveis à ocorrência de cheias do que outros."*

Além dessas considerações, a seqüência de variáveis hidrológicas/ hidrometeorológicas, amostradas em intervalos horários (ou diários) ao longo de um ano, pode apresentar correlação serial significativa e o número anual de intervalos de tempo não é uma garantia de convergência para alguma das três formas assintóticas extremais.

As características probabilísticas de um fenômeno aleatório não são facilmente definidas, tanto que a dedução teórica do modelo probabilístico necessário para definir tais características não é uma tarefa simples. Sob certas circunstâncias, as

bases ou propriedades do fenômeno físico em análise podem sugerir a forma da distribuição a ser utilizada. Por exemplo, se o processo é composto pelo somatório de muitos efeitos individuais, como no caso da precipitação anual que é a soma das precipitações diárias, a distribuição Normal pode ser utilizada com base no *teorema do limite central*. Além disso, o fato que variáveis hidrológicas e hidrometeorológicas dificilmente satisfazem as premissas da teoria clássica de valores extremos, pode vir a justificar o uso de distribuições não-extremais, tais como a Log-Normal, na análise local de freqüência de eventos máximos anuais. Chow (1954) apresenta a seguinte justificativa para o emprego da distribuição Log-Normal: os fatores causais de várias variáveis hidrológicas agem de forma multiplicativa, ao invés de aditiva, e a soma dos logaritmos desses fatores, em conseqüência do *teorema central limite* da teoria de probabilidades, tende a ser normalmente distribuída. Stedinger et al. (1993) afirmam que algumas variáveis como a diluição, por exemplo, podem resultar do produto de fatores causais. Entretanto, para o caso de enchentes ou precipitações máximas, a interpretação dessa ação multiplicativa não é evidente.

As objeções anteriores referem-se às justificativas teóricas inerentes à distribuição Log-Normal, bem como às distribuições oriundas da teoria clássica de valores extremos. Entretanto, conforme menção anterior, tais objeções não têm a finalidade de excluir os modelos citados do elenco de distribuições candidatas à modelação de variáveis hidrológicas e hidrometeorológicas. No contexto da análise de freqüência local de variáveis hidrológicas, elas devem ser consideradas candidatas como quaisquer outras distribuições e, portanto, devem ser discriminadas de acordo com outros critérios, tais como suas *medidas de aderência* aos dados amostrais.

Com relação ao número de parâmetros desconhecidos de uma distribuição de probabilidades, Hosking e Wallis (1997) afirmam que as distribuições de dois parâmetros produzem estimativas precisas de quantis quando as características distributivas populacionais a elas se assemelham. Entretanto, quando isso não ocorre, pode-se produzir estimativas tendenciosas dos quantis. A busca de um modelo probabilístico mais geral e flexível levou as agências do governo norte-americano a preconizarem o uso da distribuição Log-Pearson do tipo III para a análise local de freqüência de cheias máximas anuais em projetos com participação federal. O modelo Log-Pearson III é uma distribuição de três parâmetros, resultante da transformação logarítmica de variáveis aleatórias distribuídas de acordo com Gama ou Pearson do tipo III. Embora os seus três parâmetros confiram flexibilidade de forma a essa distribuição, a sua estimação, com base exclusiva em dados locais, é uma fonte de controvérsias. Bobée (1975) reporta situações em que a simples alteração do método de inferência estatística faz com que o parâmetro de forma dessa distribuição passe de negativo a positivo, o que a torna limitada

superiormente ou inferiormente de acordo com o sinal do parâmetro. São essas características indesejáveis da distribuição Log-Pearson do tipo III que levaram, por exemplo, Reich (1977) a argumentar contra a sua utilização na análise de freqüência local de vazões máximas anuais. No contexto da análise regional, Hosking e Wallis (1997) observam que, obedecido o preceito da *parcimônia estatística*, recomenda-se o uso de distribuições de mais de dois parâmetros por produzirem estimativas menos tendenciosas dos quantis nas caudas superior e inferior. No contexto da análise local, entretanto, resta apenas o preceito da parcimônia de parâmetros na especificação da função de distribuição de probabilidades.

As considerações anteriores, revelando a inexistência de leis dedutivas para a seleção de uma distribuição de probabilidades ou de uma família de distribuições para a análise de freqüência de eventos hidrológicos máximos anuais, remetem o analista a critérios variados e de algum modo subjetivos, entre os quais aqueles relacionados à capacidade descritiva dos modelos propostos. Alguns especialistas utilizam, como um possível critério de escolha, a comparação entre o coeficiente de assimetria amostral e o valor de assimetria teórico esperado para uma determinada distribuição de probabilidade. Por exemplo, enquanto estimativas amostrais do coeficiente de assimetria amostral próximas de zero podem sugerir a distribuição Normal como candidata à modelação estatística, amostras com assimetrias próximas a 1,14 ou -1,14 indicam a prescrição de uma distribuição de Gumbel. A utilização deste critério está sujeita à precisão da estimativa do coeficiente de assimetria, a qual cresce com o aumento do tamanho da amostra, e serve apenas como um indicador de ajuste, tornando necessário o emprego de outros critérios, tais como indicadores de aderência, para selecionar uma distribuição probabilidades apropriada.

Apesar de ser um procedimento subjetivo, o exame visual do ajuste entre as distribuições de probabilidades candidatas e os dados observados pode ser útil na seleção da distribuição de probabilidades apropriada. Para isto, os dados observados são ordenados de forma decrescente, para análise de máximos, e plotados em um papel de probabilidade específico para cada distribuição. A tendência linear dos pontos plotados em papel de probabilidade apropriado é um indício de que a amostra pode ter sido extraída daquela população. Por exemplo, uma tendência linear em um papel de probabilidade normal é uma evidência que os dados amostrais possam ter sido sorteados de uma distribuição normal. No caso de distribuições de 3 parâmetros, o exame visual ainda pode ser realizado nos papéis de probabilidade mais comuns, tais como exponencial ou normal. Entretanto, neste caso, serão observadas tendências curvilíneas e não mais lineares. Embora útil, o exame visual dos dados é adequado para amostras de grandes

tamanhos, uma vez que, amostras pequenas são muito mais sensíveis à presença de erros de amostragem ou de imprecisão na estimação da posição de plotagem, os quais podem tornar a análise visual pouco informativa ou, mesmo, pouco confiável.

A definição de um modelo distributivo que descreva as características probabilísticas de um fenômeno hidrológico é um problema complexo e passa também pela estimação dos seus parâmetros. Conforme exposição anterior, as distribuições freqüentemente utilizadas em hidrologia apresentam um número de parâmetros bastante variado. Apesar dos modelos de 3 parâmetros apresentarem maior flexibilidade de forma, de modo geral, quando se dispõe de amostras curtas (50 valores ou menos), é aconselhável que se investigue primeiramente apenas as funções que estão definidas por um ou dois parâmetros, pois a qualidade da estimativa é proporcional ao tamanho e à representatividade da amostra. O cálculo dos parâmetros pode ser realizado por vários métodos, mas os mais empregados são o método dos momentos, o método da máxima verossimilhança e o método dos momentos-L, que foram descritos e discutidos no capítulo 6.

Além das considerações anteriores sobre a definição do modelo que descreve o comportamento probabilístico de uma variável hidrológica, outro aspecto importante é a verificação do ajuste ou aderência da distribuição teórica à curva da distribuição empírica. Essa verificação de ajuste ou aderência pode ser realizada aplicando alguns testes, como por exemplo, os testes do Qui-Quadrado, de Anderson-Darling, de Kolmogorov-Smirnov, de Filliben e do teste visual, os quais foram descritos no item 7.4 do capítulo 7.

c) Identificação e tratamento de pontos atípicos

O ajuste entre a distribuição empírica e teórica pode ser comprometido pela presença de *outliers* como foi examinado no item 7.5 do capítulo 7. Esses eventos atípicos podem ser identificados por diferentes métodos. Nesta publicação destacamos os procedimentos dos quartis amostrais e da amplitude inter-quartis, descritos no item 2.1.4 do capítulo 2, e o teste de Grubbs e Beck (1972) exposto no item 7.5 do capítulo 7. Entretanto, caso ocorra a identificação de pontos atípicos, a exclusão desses *outliers* da análise de freqüência é uma decisão que exige cuidados, conforme discutido no item 7.5 do capítulo 7.

8.3 – Análise de Freqüência Utilizando o Fator de Freqüência

De acordo com Chow (1964), um quantil de uma variável hidrológica pode ser representado pela média μ_x , somada a um desvio ΔX, da seguinte forma:

$$X=\mu_X+\Delta X \tag{8.8}$$

O termo ΔX depende da dispersão característica da distribuição de X, do tempo de retorno e de outros parâmetros do modelo probabilístico. Ainda segundo Chow (1964), o termo ΔX pode ser assumido igual ao produto do desvio padrão σ por um fator de freqüência k_T, ou seja, $\Delta X = \sigma k_T$. O fator de freqüência é uma função do tempo de retorno e da distribuição de probabilidades empregada na análise. Desse modo, a equação 8.8 pode ser reescrita como

$$X = \mu_x + \sigma k_T \tag{8.9}$$

Substituindo pelas estimativas amostrais tem-se

$$x_T = \bar{x} + s k_T \tag{8.10}$$

onde, x_T é estimativa do quantil associado ao tempo de retorno T; $\bar{x}$ é a média amostral; s é o desvio-padrão amostral e k_T é o fator de freqüência associado ao modelo probabilístico e ao tempo de retorno T. A aplicação do método dos fatores de freqüência aos modelos distributivos mais usuais é objeto dos itens que se seguem.

8.3.1 – Distribuição Normal

Quando uma variável é normalmente distribuída, o quantil é calculado pela relação $x = \mu_X + Z.\sigma_X$, onde e σ_X e μ_X são, respectivamente, o desvio padrão e a média da variável e Z é a *variável normal central reduzida*. Por analogia, conclui-se que o fator de freqüência da distribuição normal, k_T, é igual à *variável normal central reduzida Z*. Os valores de Z podem ser obtidos nas Tabela 5.1 e 8.1 ou aproximados pela seguinte equação apresentada por Kite (1977):

Para $0 \le P(X \le x) \le 0{,}5$:

$$Z \approx -\left(W - \frac{C_0 + C_1 W + C_2 W^2}{1 + d_1 W + d_2 W^2 + d_3 W^3} \right), \text{ com } \quad W = \sqrt{ln\left(\frac{1}{P(X \le x)^2} \right)} \tag{8.11}$$

para $P(X \le x) > 0{,}5$

$$Z \approx \left(W - \frac{C_0 + C_1 W + C_2 W^2}{1 + d_1 W + d_2 W^2 + d_3 W^3} \right), \text{com} \quad W = \sqrt{ln\left(\frac{1}{(1 - P(X \leq x))^2} \right)} \quad (8.12)$$

Nos dois casos, os valores das constantes são: $C_0 = 2{,}515517$; $C_1 = 0{,}802853$; $C_2 = 0{,}010328$; $d_1 = 1{,}432788$; $d_2 = 0{,}189269$ e $d_3 = 0{,}001308$.

8.3.2 – Distribuição Log-Normal

Como para a distribuição Log-Normal, os logaritmos neperianos dos elementos da amostra devem ser ajustados a uma distribuição normal, tem-se que o fator de freqüência k_T também deve ser igual à *variável normal central reduzida Z* e que, portanto, a equação 8.10 pode ser reescrita da seguinte forma:

$$x_T = \exp(\bar{x}_{\ln x} + s_{\ln x} k_T) \quad (8.13)$$

onde x_T, é a estimativa do quantil associado ao tempo de retorno T; $\bar{x}_{lnx}$ é a média dos logaritmos de X; s_{lnx} é o desvio-padrão dos logaritmos de X e k_T é o fator de freqüência, igual à *variável normal central reduzida Z*.

8.3.3 – Distribuição Log-Pearson Tipo III

Kite (1977) apresenta a seguinte equação para a estimação dos quantis da distribuição Log-Pearson tipo III por meio do fator de freqüência:

$$Y_T = \ln x_T = \bar{y} + k_T s_Y \quad (8.14)$$

na qual, x_T é a estimativa do quantil associado ao tempo de retorno T, $\bar{y}$ é a média dos logaritmos neperianos de X; s_Y é o desvio padrão dos logaritmos de X; e k_T é o fator de freqüência, o qual pode ser obtido por meio de tabelas apropriadas (United States Water Resources Council, Guidelines for Determining Flood Frequency – Bulletin 17-B, U. S. Government Printing Ofiice, Washington, 1982) ou aproximado pela transformação de Wilson-Hilferty, dada por

$$k_T \approx Z + (Z^2 - 1).\frac{\gamma_Y}{6} + \frac{1}{3}(Z^3 - 6.Z).\left(\frac{\gamma_Y}{6}\right)^2 - (Z^2 - 1).\left(\frac{\gamma_Y}{6}\right)^3 + Z.\left(\frac{\gamma_Y}{6}\right)^4 + \frac{1}{3}\left(\frac{\gamma_Y}{6}\right)^5 \quad (8.15)$$

$$\text{para} \begin{cases} 0{,}01 \le \dfrac{1}{T} \le 0{,}99 \\ |\gamma_Y| < 2 \end{cases}$$

onde o coeficiente de assimetria γ_Y pode ser estimado por

$$g_Y = \frac{n\sum_{1}^{n}\left(\ln x - \overline{Y}\right)^3}{(n-2).\left[\sum_{1}^{n}\left(\ln x - \overline{Y}\right)^2\right]^{3/2}} \tag{8.16}$$

Z é a *variável normal central reduzida.*

8.3.4 – Distribuição de Gumbel

A FAP da distribuição de Gumbel para máximos é representada pela equação 5.56 e a sua inversa pode ser escrita da seguinte forma:

$$x(T) = \beta - \alpha \ln\left(-\ln\left(1 - \frac{1}{T}\right)\right) \tag{8.17}$$

onde β é parâmetro de posição; α é o parâmetro de escala e *T* é o tempo de retorno em anos. Estimando os parâmetros da distribuição pelo método dos momentos obtêm-se:

$$\hat{\beta} = \overline{X} - 0{,}45 s_X \tag{8.18}$$

$$\hat{\alpha} = \frac{s_X}{1{,}283} \tag{8.19}$$

nas quais $\overline{X}$ e s_x denotam a média e o desvio padrão amostrais.

Substituindo os parâmetros estimados pelas equações 8.18 e 8.19 em 8.17, e fazendo algumas simplificações encontra-se:

$$x(T) = \overline{X} + \left[-0{,}45 - \frac{1}{1{,}283}\ln\left(-\ln\left(1 - \frac{1}{T}\right)\right)\right] s_X \tag{8.20}$$

Comparando as equações 8.20 e 8.10, conclui-se que o fator de freqüência da distribuição de Gumbel pode ser expresso pela seguinte equação:

$$k_T = -\left[0{,}45 + \frac{1}{1{,}283}\ln\left(-\ln\left(1-\frac{1}{T}\right)\right)\right] \tag{8.21}$$

De acordo com Kite (1977), o fator de forma da distribuição de Gumbel também pode ser calculado considerando o tamanho das amostras disponíveis com a estimativa dos quantis através da equação:

$$x_T = \bar{x} + sk_T(n) \tag{8.22}$$

na qual x_T é a estimativa do quantil associado ao tempo de retorno T; $\bar{x}$ é a média amostral; s é o desvio-padrão amostral e $k_T(n)$ é fator de freqüência em função do tamanho da amostra.

O fator de freqüência, $k_T(n)$, pode ser obtido em tabelas (Haan, 1979 e Kite, 1977) ou calculado pela seguinte equação:

$$k_T(n) = \frac{Y_T - \mu_{Y_i}}{\sigma_{Y_i}} \tag{8.23}$$

na qual Y_T é a variável reduzida de Gumbel, associada a tempo de retorno T, calculada por

$$Y_T = -ln\left\{-ln\left[1-\frac{1}{T}\right]\right\} \tag{8.24}$$

Denota-se por por $\hat{\mu}_{Y_i}$ a média dos $Y_i(n)$, enquanto o desvio padrão é representado por σ_{Y_i}, o qual pode ser estimado por

$$\hat{\sigma}_{Y_i} = \sqrt{\frac{n\sum Y_i^2(n) - \left(\sum Y_i(n)\right)^2}{n^2}} \tag{8.25}$$

Nessa equação, $Y_i(n) = -ln\{-ln[F(x)]\}$ é a variável reduzida de Gumbel calculada para cada posição i de uma amostra ordenada de tamanho n. Admitindo que a posição de plotagem é calculada pela fórmula de Weilbull, obtém-se, então, a seguinte equação:

$$Y_i(n) = -ln\left(-ln\left(1-\frac{i}{n+1}\right)\right) \tag{8.26}$$

na qual i é a ordem de classificação do elemento amostral e n é o tamanho da amostra.

Os fatores de freqüência calculados com a equação 8.21 correspondem aos resultados assintóticos, resultantes da utilização da equação 8.23, quando o tamanho da amostra tende para infinito ($n \to \infty$).

8.3.5 – Distribuição Weibull (mínimos)

A estimativa dos quantis da distribuição de Weibull para análise de mínimos também pode ser realizada por meio da equação 8.10. Nesse caso, o fator de freqüência, k_T, segundo Kite (1977), é dado por:

$$k_T = A(\lambda) + B(\lambda)\left\{\left[-ln\left(1-\frac{1}{T}\right)\right]^{1/\lambda} - 1\right\} \tag{8.27}$$

na qual,

$$A(\lambda) = \left[1 - \Gamma\left(1+\frac{1}{\lambda}\right)\right]B(\lambda) \tag{8.28}$$

$$B(\lambda) = \left[\Gamma\left(1+\frac{2}{\lambda}\right) - \Gamma^2\left(1+\frac{1}{\lambda}\right)\right]^{-1/2} \tag{8.29}$$

$$\lambda = \frac{1}{H_0 + H_1\gamma + H_2\gamma^2 + H_3\gamma^3 + H_4\gamma^4} \quad \text{para} \quad -1{,}0 \le \gamma \le 2 \tag{8.30}$$

Nas equações acima, $\Gamma(.)$ denota a função Gama; H_0 = 0,2777757913; H_1 = 0,3132617714; H_2 = 0,0575670910; H_3 = -0,0013038566; H_4 = -0,0081523408 e γ é o coeficiente de assimetria estimado pela equação:

$$\hat{\gamma} = \frac{n\sum_1^n \left(x - \bar{x}\right)^3}{(n-2)\left[\sum_1^n \left(x-\bar{x}\right)^2\right]^{3/2}} \tag{8.31}$$

Exemplo 8.4 – Calcular o fator de freqüência da distribuição Gumbel, $k_T(n)$, referente ao tempo de retorno de 50 anos para uma amostra de 10 elementos. Solução: A primeira etapa consiste em calcular a variável reduzida de Gumbel, $Y_i(n)$, para cada posição *i* através da equação 8.26. A Tabela 8.5 apresenta os resultados. Em seguida é estimada a média dos valores de $Y_i(n)$ e o desvio padrão pela equação 8.25. Os resultados estão na Tabela 8.5.

A variável reduzida de Gumbel para o tempo de retorno de 50 anos é calculada pela equação 8.24:

$$Y_T = -ln\left\{-ln\left[1-\frac{1}{T}\right]\right\} = -ln\left\{-ln\left[1-\frac{1}{50}\right]\right\} = 3{,}9019 \qquad (8.32)$$

Tabela 8.5 – Cálculo dos $Y_m(n)$

i	$\frac{i}{n+1}$	$Y_i(n)$	*i*	$\frac{i}{n+1}$	$Y_i(n)$
1	0,090909	2,350619	8	0,727273	-0,26181
2	0,181818	1,60609	9	0,818182	-0,53342
3	0,272727	1,144278	10	0,909091	-0,87459
4	0,363636	0,794106		$\hat{\mu}_{Y_i}$	0,4952
5	0,454545	0,500651		$\hat{\sigma}_{Y_i}$	0,9496
6	0,545455	0,237677			
7	0,636364	-0,01153			

Como, $\hat{\mu}_Y$ = 0,4952 e $\hat{\sigma}_Y$ = 0,9496, o fator de freqüência pode ser calculado pela equação 8.23, de forma que:

$$k_{50}(10) = \frac{Y_{50} - \hat{\mu}_Y}{\hat{\sigma}_Y} = \frac{3{,}9019 - 0{,}4952}{0{,}9496} = 3{,}5874 \qquad (8.33)$$

Assim, o fator de freqüência da distribuição de Gumbel para o tempo de retorno de 50 anos e uma amostra de 10 elementos é igual a 3,5874.

Exemplo 8.5 – Admitindo que uma série de vazões mínimas com 7 dias de duração apresenta um coeficiente de assimetria de -0,10, calcular o fator de freqüência da distribuição de Weibull para os tempos de retorno de 2, 5, 10, 20, 50 e 100 anos.

Solução: Utilizando o valor do coeficiente de assimetria γ = - 0,10, calcula-se o parâmetro λ pela equação 8.30. O valor de λ é igual a 4,048160583. O parâmetro permite que se calcule $B(\lambda)$ pela equação 8.29 e em seguida

$A(\lambda)$ através da equação 8.28, a saber $B(\lambda)$ = 3,972674215 e $A(\lambda)$ = 0,369376575.
Com esses valores é possível estimar o fator de freqüência pela equação 8.27,

$$k_T = 0{,}369376575 + 3{,}972674215.\left\{\left[-ln\left(1-\frac{1}{T}\right)\right]^{1/4{,}048160583} - 1\right\} \quad (8.34)$$

Substituindo os tempos de retorno na equação 8.34, calcula-se os valores dos fatores de freqüência. Neste exemplo, tem-se:

***T* (anos)**	**2**	**5**	**10**	**20**	**50**	**100**
k_T	0,0255	-0,8607	-1,3247	-1,6959	-2,0881	-2,3281

8.4 – Intervalo de Confiança para os Quantis

Os intervalos de confiança para os quantis estimados podem ser definidos a partir da equação 6.23, como foi detalhado no item 6.6 do capitulo 6. Naquele item foi visto que, assintoticamente (para grandes valores de *n*), os estimadores de quantis $\hat{x}_T$ são normalmente distribuídos. Sendo assim, com base na equação 6.23, o intervalo de confiança aproximado para um quantil $\hat{x}_T$ a um nível de confiança 100 (1-α)% é definido por:

$$\hat{x}_T - Z_{1-\frac{\alpha}{2}} s_T \le \hat{x}_T \le \hat{x}_T + Z_{1-\frac{\alpha}{2}} s_T \quad (8.35)$$

onde $Z_{1-\frac{\alpha}{2}}$ é a *variável normal central reduzida* associada à probabilidade (1-α/2) e s_T é o erro-padrão da estimativa de $\hat{x}_T$, o qual varia com o modelo distributivo em análise.

No capítulo 6, analisou-se a definição dos intervalos de confiança quando os parâmetros das distribuições foram estimados pelos métodos dos momentos, da máxima verossimilhança e dos momentos-L. Para ilustrar a definição de intervalos de confiança aproximados de quantis, apresenta-se a seguir as expressões para os erros-padrão para algumas distribuições, cujos parâmetros foram estimados pelo método dos momentos.

• Normal

$$s_T = \sqrt{\frac{s_X^2}{n}\left(1 + \frac{Z^2_{1-\frac{1}{T}}}{2}\right)} \tag{8.36}$$

na qual T é o tempo de retorno; s_X é o desvio padrão amostral e Z é a *variável normal central reduzida*.

• Log-Normal

$$s_T = \sqrt{exp\left[\left(\overline{Y} + Z_{1-\frac{1}{T}} s_Y\right) \pm Z_{1-\frac{\alpha}{2}} \sqrt{\frac{s_Y^2}{n}\left(1 + \frac{1}{2} Z^2_{1-\frac{1}{T}}\right)}\right]} \tag{8.37}$$

na qual $Y = ln(X)$; s_Y é o desvio padrão dos logaritmos dos dados observados; T é o tempo de retorno e Z é a *variável normal central reduzida*.

• Log-Pearson Tipo III

Segundo Kite (1977), o erro-padrão para a distribuição Log-Pearson Tipo III pode ser estimado, no espaço logarítmico, a partir da seguinte equação:

$$s_{T,Y} = \delta \sqrt{\frac{s_Y^2}{n}} \tag{8.38}$$

na qual $s_{T,Y}$ é o erro-padrão dos logaritmos dos eventos observados; n é o tamanho da amostra; s_Y é o desvio padrão dos logaritmos dos dados observados e δ pode ser obtido a partir da Tabela 8.6, em dependência do tempo de retorno e do coeficiente de assimetria dos logaritmos dos dados amostrais.

O erro padrão pode ser convertido para o espaço aritmético por meio da relação:

$$s_T = \frac{\hat{x}_T\left(e^{s_{T,Y}} - e^{-s_{T,Y}}\right)}{2{,}0} \tag{8.39}$$

Tabela 8.6 – Parâmetro δ para estimativa do erro padrão da Log-Pearson Tipo III

	Tempo de retorno (anos)					
Assimetria	**2**	**5**	**10**	**20**	**50**	**100**
0,00	1,0801	1,1698	1,3748	1,6845	2,1988	2,6363
0,10	1,0808	1,2006	1,4367	1,7810	2,3425	2,8168
0,20	1,0830	1,2309	1,4989	1,8815	2,4986	3,0175
0,30	1,0866	1,2609	1,5610	1,9852	2,6656	3,2365
0,40	1,0918	1,2905	1,6227	2,0915	2,8423	3,4724
0,50	1,0987	1,3199	1,6838	2,1998	3,0277	3,7238
0,60	1,1073	1,3492	1,7441	2,3094	3,2209	3,9895
0,70	1,1179	1,3785	1,8032	2,4198	3,4208	4,2684
0,80	1,1304	1,4082	1,8609	2,5303	3,6266	4,5595
0,90	1,1449	1,4385	1,9170	2,6403	3,8374	4,8618
1,00	1,1614	1,4699	1,9714	2,7492	4,0522	5,1741
1,10	1,1799	1,5030	2,0240	2,8564	4,2699	5,4952
1,20	1,2003	1,5382	2,0747	2,9613	4,4896	5,8240
1,30	1,2223	1,5764	2,1237	3,0631	4,7100	6,1592
1,40	1,2457	1,6181	2,1711	3,1615	4,9301	6,4992
1,50	1,2701	1,6643	2,2173	3,2557	5,1486	6,8427
1,60	1,2952	1,7157	2,2627	3,3455	5,3644	7,1881
1,70	1,3204	1,7732	2,3081	3,4303	5,5761	7,5339
1,80	1,3452	1,8374	2,3541	3,5100	5,7827	7,8783
1,90	1,3690	1,9091	2,4018	3,5844	5,9829	8,2196
2,00	1,3913	1,9888	2,4525	3,6536	6,1755	8,5562

• Weibull

Segundo Kite (1977), o erro-padrão para a distribuição de Weibull, para mínimos, pode ser estimado por:

$$s_T = \delta_w \sqrt{\frac{s_X^2}{n}} \tag{8.40}$$

na qual, s_X é o desvio padrão amostral; n é o tamanho da amostra e δ_w pode ser obtido a partir da Tabela 8.7, na dependência do tempo de retorno e do coeficiente de assimetria amostral.

Tabela 8.7 – Parâmetro δ_w para estimativa do erro padrão da distribuição de Weibull (mínimos)

	Tempo de retorno (anos)					
Assimetria	2	5	10	20	50	100
-0,80	0,9265	1,3665	1,8116	2,2267	2,6325	2,7650
-0,70	0,9743	1,3556	1,7517	2,1869	2,7877	3,2475
-0,60	1,0242	1,3492	1,6940	2,1413	2,8843	3,5450
-0,50	1,0710	1,3434	1,6356	2,0820	2,9084	3,6757
-0,40	1,0954	1,3259	1,5738	1,9846	2,7731	3,5067
-0,30	1,0886	1,2934	1,5063	1,8351	2,4456	3,0047
-0,20	1,0952	1,2624	1,4374	1,7320	2,2300	2,7011
-0,10	1,1065	1,2282	1,3709	1,6181	2,0938	2,5248
0,00	1,1157	1,1916	1,3042	1,5255	1,9631	2,3559
0,10	1,1244	1,1532	1,2374	1,4371	1,8437	2,2043
0,20	1,1318	1,1130	1,1711	1,3529	1,7336	2,0658
0,30	1,1394	1,0712	1,1078	1,2814	1,6496	1,9627
0,40	1,1460	1,0281	1,0467	1,2172	1,5775	1,8740
0,50	1,1517	0,9839	0,9905	1,1653	1,5236	1,8065
0,60	1,1567	0,9392	0,9414	1,1287	1,4905	1,7623
0,70	1,1605	0,8943	0,8981	1,1014	1,4661	1,7262
0,80	1,1636	0,8500	0,8646	1,0895	1,4583	1,7074
0,90	1,1657	0,8072	0,8422	1,0914	1,4630	1,7006
1,00	1,1671	0,7669	0,8319	1,1064	1,4788	1,7047
1,10	1,1678	0,7303	0,8348	1,1338	1,5049	1,7189
1,20	1,1681	0,6988	0,8507	1,1719	1,5394	1,7413
1,30	1,1680	0,6739	0,8792	1,2196	1,5815	1,7715
1,40	1,1676	0,6569	0,9196	1,2745	1,6291	1,8075
1,50	1,1669	0,6488	0,9673	1,3354	1,6816	1,8488
1,60	1,1658	0,6494	1,0218	1,3987	1,7355	1,8921
1,70	1,1643	0,6585	1,0807	1,4638	1,7908	1,9376
1,80	1,1622	0,6742	1,1406	1,5274	1,8446	1,9823
1,90	1,1596	0,6940	1,1987	1,5877	1,8952	2,0247
2,00	1,1544	0,7148	1,2523	1,6421	1,9405	2,0628

Exemplo 8.6 – Realizar uma análise de freqüência com os dados de vazões diárias máximas anuais do rio Paraopeba em Ponte Nova do Paraopeba apresentados no Anexo 2. Considerar como candidatas as distribuições Log-Normal, Gumbel, Exponencial, Pearson III, Log-Pearson III e Generalizada de eventos extremos (GEV).

Solução: A primeira etapa consiste no cálculo das estatísticas e os momentos-L da série. Os resultados estão apresentados na Tabela 8.8.

Tabela 8.8 – Estatísticas de série de vazões diárias máximas anuais de Ponte Nova do Paraopeba

	Estatísticas descritivas	Estatísticas dos Logaritmos	Momentos-L	
Tamanho da Amostra	57	57	*l*1	534,2
Valor Máximo	1017	6,9250	*l*2	99,63
Valor Mínimo	246	5,5050	t3	0,1288
Média	534,2	6,2270	t4	0,1070
Desvio-Padrão	176,0	0,3320		
Coeficiente de Assimetria	0,6040	-0,0972		

O segundo passo consiste em aplicar alguns testes para verificar as hipóteses de independência e homogeneidade da série. Nesse exemplo a independência foi verificada com o teste não-paramétrico proposto por Wald e Wolfowitz (1943), descrito no item 7.3.2, e a homogeneidade pelo teste de Mann e Whitney (1947), descrito no item 7.3.3. A série pode ser considerada independente e homogênea a um nível de significância de 5%.

A terceira etapa é a verificação da presença de eventos atípicos na amostra. Nesse caso foi aplicado teste de Grubbs e Beck (1972), descrito no item 7.5. Na série analisada, a um nível de significância de 10%, não foi observada a presença de *outliers*.

Após a análise inicial dos dados, são calculados os parâmetros das distribuições candidatas. Nesse exemplo os parâmetros foram calculados pelo método dos momentos-L apresentado no capítulo 6. Os resultados estão na Tabela 8.9.

Tabela 8.9 – Parâmetros das distribuições candidatas

Distribuição	Posição (ξ)	Escala (α)	Forma (κ)
Log-Normal	6,2274	0,3382	
Gumbel	451,2123	143,7298	
Exponencial	334,9236	199,2519	
Pearson-III	534,1754	180,0157	0,7854
Log-Pearson-III	6,2274	0,3383	-0,1226
GEV	455,6143	152,0965	0,0650

Definidos os parâmetros das distribuições, é possível calcular os quantis associados a diferentes tempos de retorno a partir das inversas das FAP's dos modelos candidatos. As distribuições candidatas foram detalhadas nos capítulos 5 e 6. Na Tabela 8.10 são apresentadas algumas funções inversas das distribuições candidatas.

Tabela 8.10 – Funções Inversas da FAP de algumas distribuições

Distribuição	Inversa x(T)	Observações
Log-Normal	$x(T) = exp(\xi + \alpha . Z_T)$	Z_T é a variável normal central reduzida associada à probabilidade (1-1/T)
Gumbel	$x(T) = \xi - \alpha . ln\left(-ln\left(-\frac{1}{T}\right)\right)$	
Exponencial	$x(T) = \xi - \alpha . ln\left(\frac{1}{T}\right)$	
GEV	$x(T) = \xi + \alpha . \left\{\frac{1-\left[-ln\left(1-\frac{1}{T}\right)\right]^k}{k}\right\}$	para $k \neq 0$

Os parâmetros de posição μ, escala σ e forma γ da distribuição Pearson Tipo III podem ser calculados com as equações

$$\mu = \lambda_{1,}\ \sigma = \frac{\lambda_2 \pi^{1/2} c^{1/2} \Gamma(c)}{\Gamma\left(c + \frac{1}{2}\right)} \quad \text{e} \quad \gamma = 2c^{-1/2} sinal(\tau_3)\ .$$

A variável c é estimada considerando duas situações. A primeira, se $0 < |\tau_3| < 1/3$, nesse caso adotar $z = 3\pi\tau_3^2$ e aplicar a equação $c \approx \frac{1 + 0{,}2906z}{z + 0{,}1882z^2 + 0{,}0442z^3}$. A segunda, se $1/3 \leq |\tau_3| < 1$, nessa situação adota-se $z = 1 - |\tau_3|$ e emprega-se a equação $c \approx \frac{0{,}36067z - 0{,}59567z^2 + 0{,}25361z^3}{1 - 0{,}78861z + 2{,}56096z^2 - 0{,}77045z^3}$.

A distribuição Pearson Tipo III com parâmetros de posição μ, escala σ e forma γ, apresenta algumas relações importantes com as distribuições Gama e Normal, as quais facilitam a estimação dos quantis. Quando o parâmetro de forma γ é positivo, a Pearson-III está associada à distribuição Gama. Se o parâmetro de forma γ é negativo, a Pearson-III está associada à distribuição Gama refletida. E, quando o parâmetro de forma γ é igual a zero, a Pearson-III está relacionada à distribuição Normal. Considerando que uma variável X segue uma distribuição Pearson tipo III, com parâmetros de posição μ, escala σ e forma γ, a relação entre esses parâmetros e os das distribuições Gama e Normal são as seguintes:

• Se $\gamma > 0$, então $X - \mu + \frac{2.\sigma}{\gamma}$ segue uma distribuição Gama com

parâmetros $\alpha = \frac{4}{\gamma^2}$ e $\beta = \frac{\sigma\gamma}{2}$. Desse modo, os quantis da Pearson-III com parâmetro de forma positivo podem ser calculados pela equação:

$$x(T) = \mu - \frac{2\sigma}{\gamma} + G^{-1}\left(1 - \frac{1}{T}, \alpha, \beta\right) \quad (8.41)$$

onde T é o tempo de retorno e $G^{-1}()$ é a inversa da distribuição Gama com parâmetros α e β.

• Se $\gamma < 0$, então $-X + \mu - \frac{2.\sigma}{\gamma}$ segue uma distribuição Gama com parâmetros $\alpha = \frac{4}{\gamma^2}$ e $\beta = \left|\frac{\sigma\gamma}{2}\right|$. Desse modo, os quantis da Pearson-III com parâmetro de forma negativo podem ser calculados pela equação:

$$x(T) = \mu - \frac{2\sigma}{\gamma} - G^{-1}\left(\frac{1}{T}, \alpha, \beta\right) \quad (8.42)$$

onde T é o tempo de retorno e $G^{-1}()$ é a inversa da distribuição Gama com parâmetros α e β.

• Se $\gamma = 0$, então X segue uma distribuição Normal com parâmetros μ e σ. Assim, os quantis da Pearson-III com parâmetro de forma nulo podem ser calculados pela equação:

$$x(T) = \mu + \sigma Z_T \quad (8.43)$$

onde T é o tempo de retorno e Z_T é a variável normal central reduzida associada uma probabilidade (1 - 1/T). Recorde que, no programa Microsoft EXCEL, a inversa da distribuição Gama com parâmetros α e β pode ser calculada com a função INVGAMA() e a variável normal central reduzida com a função INV.NORMP().

Quando uma variável X segue a distribuição Log-Pearson tipo III, é um fato matemático que a variável transformada $Y = ln(X)$ distribui-se de acordo com a Pearson tipo III. Assim, os parâmetros podem ser calculados por meio dos logaritmos dos valores observados e os quantis são estimados por meio das seguintes equações:

• Para $\gamma_{lnX} > 0$

$$x(T) = exp\left\{\mu_{lnX} - \frac{2\sigma_{lnX}}{\gamma_{lnX}} + G^{-1}\left(1 - \frac{1}{T}, \alpha, \beta\right)\right\} \quad (8.44)$$

• Para $\gamma_{lnX} < 0$

$$x(T) = exp\left\{\mu_{lnX} - \frac{2\sigma_{lnX}}{\gamma_{lnX}} - G^{-1}\left(\frac{1}{T}, \alpha, \beta\right)\right\} \quad (8.45)$$

• Para $\gamma_{lnX} = 0$

$$x(T) = exp(\mu_{lnX} + \sigma_{lnX} Z_T) \quad (8.46)$$

Os quantis das distribuições candidatas foram estimados por meio das funções inversas anteriormente apresentadas, pela substiuição das estatísticas populacionais pelas amostrais. Os resultados estão apresentados na Tabela 8.11.

Tabela 8.11 – Quantis calculados para o exemplo 8.1 (m³/s)

	T (anos)						
Distribuição	**2**	**5**	**10**	**50**	**100**	**200**	**1000**
Log-Normal	506,4	673,2	781,2	1014,3	1112,2	1210,1	1440,0
Gumbel	503,9	666,8	774,7	1012,0	1112,4	1212,4	1444,0
Exponential	473,0	655,6	793,7	1114,4	1252,5	1390,6	1711,3
Pearson-III	510,8	674,8	774,7	974,5	1052,8	1128,1	1294,4
Log-Pearson-III	510,0	674,5	777,7	992,1	1079,1	1164,4	1358,5
GEV	510,7	673,0	774,0	979,8	1060,4	1137,1	1302,1

Antes de se iniciar a verificação do ajuste entre as distribuições teóricas e a empírica, a escolha do modelo probabilístico mais adequado pode ser feita por meio da análise dos parâmetros das distribuições candidatas e das estatísticas amostrais. Nesse exemplo, observa-se que o parâmetro de forma da GEV é positivo (ver Tabela 8.9), indicando uma distribuição com limite superior, o que, para alguns especialistas, conforme discussão no presente capitulo, não é adequado para a análise de máximos. Outra distribuição candidata que pode ser excluída da análise é a Log-Pearson tipo III, pois o coeficiente de assimetria no espaço logaritmo é negativo, indicando que esta distribuição também apresenta um limite superior. Considerando esses critérios, restam como candidatas as distribuições Log-Normal, Gumbel, Exponencial e Pearson-III.

Após essa seleção inicial, o próximo passo na escolha da distribuição teórica que melhor se ajustou à distribuição empírica é a verificação do ajuste por meio de testes de aderência e análise visual dos gráficos de probabilidades. Os testes de aderência foram descritos no item 7.4. Neste exemplo, foi aplicado o teste de Filliben, no qual as probabilidades empíricas para a verificação das distribuições Log-Normal e Pearson foram calculadas por meio da fórmula de posição de plotagem de Blom; para as distribuições de Gumbel e Exponencial, foi utilizada a fórmula de Gringorten. Os resultados obtidos do teste de Filliben estão na Tabela 8.12.

Tabela 8.12 – Resultados do teste de Filliben

Distribuição	$r_{crit,\alpha}$	r	Situação
Log-Normal (α = 10%)	0,9835	0,9952	Aceita
Gumbel (α = 10%)	0,9760	0,9919	Aceita
Exponencial (α = 10%)	0,9716	0,9616	Rejeitada
Pearson-III (α = 5%)	0,9860	0,9958	Aceita

As Figuras 8.7 e 8.8 permitem a verificação visual do ajuste entre as distribuições empíricas e teóricas. As probabilidades empíricas foram calculadas com ordenamento decrescente da amostra e a utilização das fórmulas de Blom e Gringorten, com os resultados na Tabela 8.13.

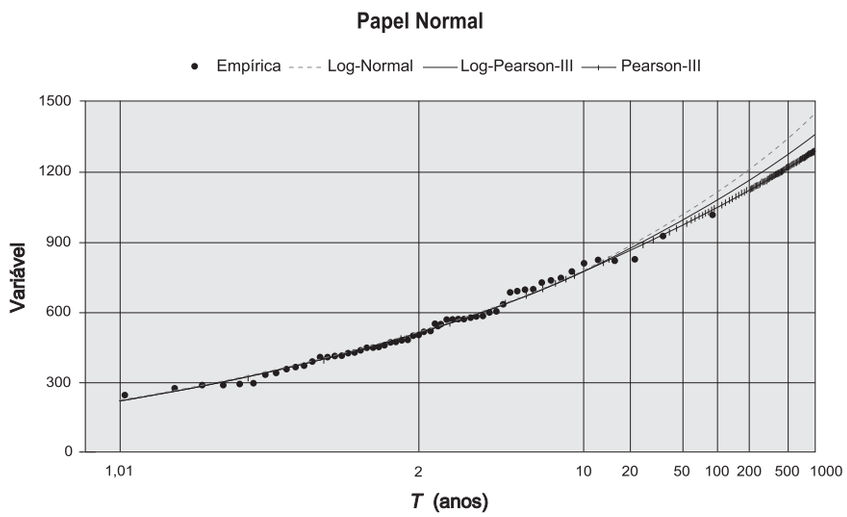


Figura 8.7 – Ajuste das distribuições Log-Normal, Pearson-III e Log-Pearson III

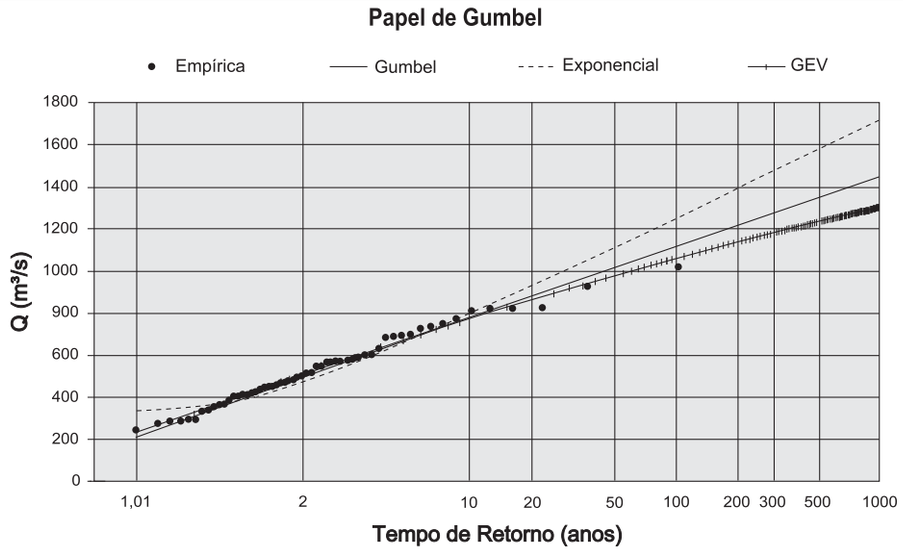


Figura 8.8 – Ajuste das distribuições de Gumbel, Exponencial e GEV

Após a aplicação do teste de Filliben e a verificação do ajuste visual restaram três distribuições candidatas, Log-Normal, Gumbel e Pearson-III. Qualquer um desses modelos pode ser adotado como a distribuição das vazões máximas anuais no rio Paraopeba em Ponte Nova do Paraopeba, ou seja, a partir desse ponto a escolha do modelo incorpora critérios muito subjetivos.

Nesse caso, por se tratar de uma análise de máximos, se o interesse é por tempos de retorno inferiores a 1000 anos, a escolha da distribuição de Gumbel ou da Log-Normal seria praticamente indiferente.

Tabela 8.13 – Probabilidades empíricas

i	AH	Q(m³/s)	Blom	*T*	Gring	*T*	*i*	AH	Q(m³/s)	Blom	*T*	Gring	*T*
1	84/85	1017	0,0109	91,6	0,0098	102,0	30	54/55	498	0,5175	1,93	0,5175	1,93
2	90/91	927	0,0284	35,2	0,0273	36,6	31	89/90	481	0,5349	1,87	0,5350	1,87
3	91/92	827	0,0459	21,8	0,0448	22,3	32	68/69	478	0,5524	1,81	0,5525	1,81
4	60/61	822	0,0633	15,8	0,0623	16,0	33	40/41	472	0,5699	1,75	0,5700	1,75
5	78/79	822	0,0808	12,4	0,0798	12,5	34	55/56	470	0,5873	1,70	0,5875	1,70
6	48/49	810	0,0983	10,2	0,0973	10,3	35	41/42	458	0,6048	1,65	0,6050	1,65
7	56/57	774	0,1157	8,6	0,1148	8,7	36	67/68	450	0,6223	1,61	0,6225	1,61
8	63/64	748	0,1332	7,5	0,1324	7,6	37	73/74	449	0,6397	1,56	0,6401	1,56
9	77/78	736	0,1507	6,6	0,1499	6,7	38	59/60	448	0,6572	1,52	0,6576	1,52
10	65/66	726	0,1681	5,9	0,1674	6,0	39	85/86	437	0,6747	1,48	0,6751	1,48
11	82/83	698	0,1856	5,4	0,1849	5,4	40	98/99	427	0,6921	1,44	0,6926	1,44
12	95/96	695	0,2031	4,9	0,2024	4,9	41	92/93	424	0,7096	1,41	0,7101	1,41
13	50/51	690	0,2205	4,5	0,2199	4,5	42	39/40	414	0,7271	1,38	0,7276	1,37
14	42/43	684	0,2380	4,2	0,2374	4,2	43	61/62	414	0,7445	1,34	0,7451	1,34
15	94/95	633	0,2555	3,9	0,2549	3,9	44	43/44	408	0,7620	1,31	0,7626	1,31

Tabela 8.13 – Continuação

i	AH	Q(m³/s)	Blom	*T*	Gring	*T*	*i*	AH	Q(m³/s)	Blom	*T*	Gring	*T*
16	93/94	603	0,2729	3,7	0,2724	3,7	45	58/59	408	0,7795	1,28	0,7801	1,28
17	87/88	601	0,2904	3,4	0,2899	3,4	46	57/58	388	0,7969	1,25	0,7976	1,25
18	83/84	585	0,3079	3,2	0,3074	3,3	47	44/45	371	0,8144	1,23	0,8151	1,23
19	66/67	580	0,3253	3,1	0,3249	3,1	48	49/50	366	0,8319	1,20	0,8326	1,20
20	38/39	576	0,3428	2,9	0,3424	2,9	49	74/75	357	0,8493	1,18	0,8501	1,18
21	46/47	570	0,3603	2,8	0,3599	2,8	50	69/70	340	0,8668	1,15	0,8676	1,15
22	51/52	570	0,3777	2,6	0,3775	2,6	51	45/46	333	0,8843	1,13	0,8852	1,13
23	64/65	570	0,3952	2,5	0,3950	2,5	52	97/98	296	0,9017	1,11	0,9027	1,11
24	71/72	568	0,4127	2,4	0,4125	2,4	53	53/54	295	0,9192	1,09	0,9202	1,09
25	79/80	550	0,4301	2,3	0,4300	2,3	54	52/53	288	0,9367	1,07	0,9377	1,07
26	86/87	549	0,4476	2,2	0,4475	2,2	55	88/89	288	0,9541	1,05	0,9552	1,05
27	72/73	520	0,4651	2,2	0,4650	2,2	56	75/76	276	0,9716	1,03	0,9727	1,03
28	62/63	515	0,4825	2,1	0,4825	2,1	57	70/71	246	0,9891	1,01	0,9902	1,01
29	47/48	502	0,5000	2,0	0,5000	2,0							

Como visto no exemplo 8.6, a seleção do modelo probabilístico que melhor se ajusta aos dados amostrais não é uma tarefa fácil, o que obriga o analista a fazer uso de uma combinação de critérios objetivos e subjetivos. A subjetividade presente no processo de escolha do modelo pode gerar soluções diferenciadas para uma mesma série hidrológica dependendo dos critérios aplicados pelo analista.

De qualquer forma é importante ressaltar que devido ao pequeno tamanho das amostras disponíveis é impossível comprovar que o modelo selecionado representa a verdadeira distribuição populacional.

Dentre as ferramentas disponíveis para a análise de freqüência local, os sistemas especialistas computacionais, que emulam os princípios de raciocínio de um especialista humano ao selecionar uma distribuição de probabilidades, têm-se mostrado muito úteis. Um exemplo desse tipo de sistema é o SEAF (Sistema Especialista para Análise de Freqüência local de eventos máximos anuais), disponível a partir da URL http://www.ehr.ufmg.br/downloads.php.

Exemplo 8.7 – No Anexo 2, encontram-se os dados de vazões mínimas, para diversas durações, da estação fluviométrica de Ponte Nova do Paraopeba, código 40800001. Ajustar as distribuições de Gumbel (de mínimos) e Weibull (de mínimos) às vazões mínimas com duração de 3 dias. Qual distribuição apresenta o melhor ajuste ?
Solução: Inicialmente são calculadas as estatísticas da série: n = 59; $\overline{X} = 27{,}778\,\text{m}^3/\text{s}$; $s = 7{,}683\,\text{m}^3/\text{s}$; $g =$ 0,04706 , esse calculado pela equação 8.31. Para ajustar a distribuição de Weibull pelo método do fator de

freqüência, é necessário estimar os parâmetros da equação 8.27 por meio das equações 8.28 a 8.30. Os resultados obtidos foram: $\lambda = 3{,}417092$; $B(\lambda) = 3{,}441457$ e $A(\lambda) = 0{,}34891$. Substituindo esses valores na equação 8.27, referente à estimativa do fator de freqüência e, em seguida, aplicando-a na equação geral de freqüência (equação 8.10), foram calculadas as vazões mínimas associadas a diferentes tempos de retorno, conforme apresentado na Tabela 8.14. Estimando os parâmetros da distribuição de Gumbel para mínimos (ver item 6.7.9), obtém-se $\hat{\alpha} = 5{,}990612$ e $\hat{\beta} = 31{,}23533$. Os quantis da distribuição de Gumbel para mínimos são estimados pela equação:

$$x(T) = \beta + \alpha\, ln\left(-ln\left(1 - 1/T\right)\right) \tag{8.47}$$

onde β e α são os parâmetros de posição e escala respectivamente, e T é tempo de retorno. A Tabela 8.14 apresenta os quantis calculados pela equação 8.47.

Tabela 8.14 – Quantis das distribuições de Weibull e Gumbel

	T (anos)	2	5	10	15	25	50
Weibull	k_t	-0,00111	-0,8738	-1,31127	-1,51884	-1,74288	-1,99398
	Q_T (m³/s)	27,77	21,06	17,70	16,11	14,39	12,46
Gumbel	Q_T (m³/s)	29,04	22,25	17,75	15,22	12,07	7,86

A Figura 8.9 apresenta as distribuições empírica e teóricas, grafadas em um papel de probabilidade de Gumbel. A série foi ordenada de forma crescente e a posição de plotagem da distribuição empírica foi calculada por meio da fórmula de Gringorten, conforme Tabela 8.15.

Analisando a Figura 8.9, percebe-se visualmente que a distribuição de Weibull se ajustou melhor à distribuição empírica. Em algumas análises de vazões mínimas, a primeira tentativa de ajuste entre a distribuição empírica e a teórica não apresenta resultados satisfatórios. Uma das causas pode ser a presença de valores altos na amostra que não permitem o ajuste adequado. Como esses valores estão, em geral, fora da parte de maior interesse da análise, novas tentativas podem ser realizadas para tentar melhorar o ajuste, retirando da série alguns valores elevados e refazendo os cálculos.

Tabela 8.15 – Distribuição empírica das vazões mínimas de Ponte Nova de Paraopeba com 3 dias de duração

m	Ano	Q-3 dias	PP	T (anos)	*m*	Ano	Q-3 dias	PP	T (anos)
1	1999	11,97	0,009472	105,6	31	1961	27,50	0,516915	1,93
2	1971	12,80	0,026387	37,9	32	1968	27,50	0,533829	1,87
3	1955	15,20	0,043302	23,1	33	1980	28,53	0,550744	1,82
4	1998	15,80	0,060217	16,6	34	1984	28,77	0,567659	1,76
5	1954	17,90	0,077131	13,0	35	1958	28,97	0,584574	1,71
6	1963	17,90	0,094046	10,6	36	1993	29,03	0,601488	1,66
7	1964	18,13	0,110961	9,0	37	1997	29,73	0,618403	1,62
8	1959	19,33	0,127876	7,8	38	1940	29,90	0,635318	1,57
9	1990	20,50	0,14479	6,9	39	1944	30,30	0,652233	1,53
10	1960	20,80	0,161705	6,2	40	1951	30,70	0,669147	1,49
11	1969	21,20	0,17862	5,6	41	1973	30,70	0,686062	1,46
12	1956	21,47	0,195535	5,1	42	1991	31,90	0,702977	1,42
13	1975	21,70	0,212449	4,7	43	1966	32,00	0,719892	1,39
14	1995	21,87	0,229364	4,4	44	1952	32,20	0,736806	1,36
15	1988	22,70	0,246279	4,1	45	1946	32,70	0,753721	1,33
16	1976	23,93	0,263194	3,8	46	1950	34,00	0,770636	1,30
17	1972	24,00	0,280108	3,6	47	1982	34,57	0,787551	1,27
18	1953	24,03	0,297023	3,4	48	1939	34,83	0,804465	1,24
19	1978	24,17	0,313938	3,2	49	1965	34,83	0,82138	1,22
20	1974	24,40	0,330853	3,0	50	1949	35,37	0,838295	1,19
21	1948	24,67	0,347767	2,9	51	1992	35,37	0,85521	1,17
22	1957	24,67	0,364682	2,7	52	1942	37,00	0,872124	1,15
23	1989	24,90	0,381597	2,6	53	1941	37,33	0,889039	1,12
24	1987	24,93	0,398512	2,5	54	1979	37,60	0,905954	1,10
25	1994	25,07	0,415426	2,4	55	1945	38,27	0,922869	1,08
26	1986	25,17	0,432341	2,3	56	1947	38,60	0,939783	1,06
27	1970	25,40	0,449256	2,2	57	1938	42,13	0,956698	1,05
28	1962	25,77	0,466171	2,1	58	1985	44,00	0,973613	1,03
29	1967	26,97	0,483085	2,1	59	1943	50,00	0,990528	1,01
30	1996	27,23	0,5	2,0					

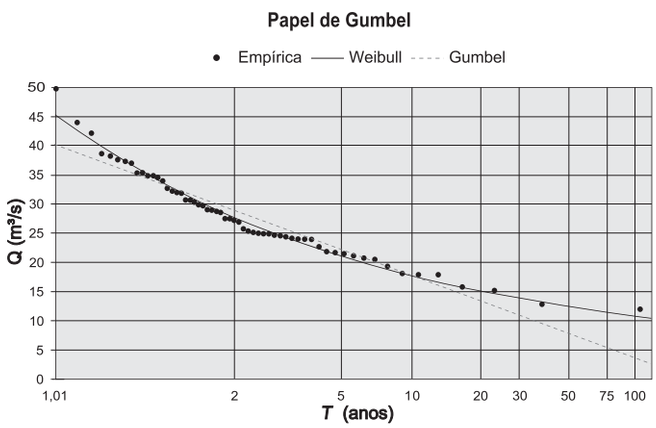


Figura 8.9 – Distribuições ajustadas às vazões mínimas de Ponte Nova de Paraopeba com 3 dias de duração

8.5 – Análise de Freqüência de Séries de DuraçãoParcial

Conforme menção anterior, a modelação probabilística das variáveis hidrológicas pode ser realizada por meio de duas abordagens gerais. A primeira, utilizando as séries de máximos anuais, as quais consideram apenas o maior evento em cada ano hidrológico, e a segunda, empregando as séries de duração parcial (SDP), também denominadas de séries de picos acima de um limiar (POT, da expressão inglesa *Peaks over Threshold*), as quais incluem todos os picos que superaram certo valor de referência ou limiar. A principal objeção à utilização de séries de máximos anuais está relacionada ao fato de se empregar somente o maior evento de cada ano hidrológico, não considerando que o segundo maior evento de um ano pode ser superior aos picos de outros anos, circunstância que é comum em regiões mais secas. A análise com séries de duração parcial evita este tipo de problema, pois considera todos os picos independentes que superam um limite especificado. Entretanto, o uso das séries de duração parcial apresenta a dificuldade adicional de se definir os critérios para identificar somente aqueles eventos superiores ao limite estabelecido que sejam independentes, com a garantia de que não sejam utilizados duas ou mais ocorrências que tenham, como origem, o mesmo mecanismo ou evento causal, conforme comentário no inicio deste capítulo. Essa dificuldade adicional torna a análise com séries de duração parcial um pouco mais trabalhosa.

Como a série de duração parcial é formada por eventos que superaram um valor limiar, a modelação probabilística para esse tipo de série requer respostas para duas questões importantes. A primeira refere-se à decisão de qual deve ser o modelo que melhor representa a freqüência ou a taxa anual de excedências dos eventos maiores que o limiar estipulado, ou seja, qual é a distribuição que descreve o número médio anual de eventos que superaram o valor de referência. A segunda refere-se à decisão de qual deve ser o modelo distributivo das magnitudes das excedências acima do valor limiar. Em geral, a distribuição de Poisson é freqüentemente usada para modelar a taxa de excedências dos eventos, enquanto a distribuição exponencial é empregada para descrever a magnitude dos picos excedentes sobre o limiar estabelecido (Stedinger et al. 1993).

Uma vez que, em geral, o interesse volta-se para o intervalo de tempo anual, é necessário calcular as estimativas das probabilidades anuais de excedência a partir das séries de duração parcial. Supondo que o número médio anual de eventos maiores que um limiar estabelecido u, seja um estimador da taxa de excedências ν, da distribuição de Poisson, é possível demonstrar, conforme dedução apresentada no Anexo 9, que a relação entre a função acumulada de probabilidades para máximos anuais $F_a(x)$, a razão de ocorrência dos eventos acima do limite

estipulado ν, e a distribuição acumulada da série de duração parcial $H_u(x)$ é dada pela seguinte equação:

$$F_a(x) = exp\{-\nu[1 - H_u(x)]\} \tag{8.48}$$

Uma vez que a probabilidade de excedência anual é dada por [1- $F_a(x)$], a equação 8.48 pode ser alterada para:

$$1 - F_a(x) = 1 - exp\{-\nu[1 - H_u(x)]\} \tag{8.49}$$

Como a probabilidade de excedência anual, [1- $F_a(x)$], é igual a $1/T_a$, onde T_a é o período de retorno anual, e a correspondente probabilidade de excedência para um valor x em uma série de duração parcial, [1- $H_u(x)$], pode ser representada por q_i, verifica-se que a equação 8.49 pode ser transformada em:

$$\frac{1}{T_a} = 1 - exp\{-\nu q_i\} \tag{8.50}$$

Segundo Stedinger et al. (1993), o tempo de retorno da série parcial T_p é expresso pela relação

$$T_P = \frac{1}{\nu q_i} \tag{8.51}$$

Substituindo esse resultado na equação 8.50, obtém-se

$$\frac{1}{T_a} = 1 - exp\left\{-\frac{1}{T_p}\right\} \tag{8.52}$$

Após algumas transformações da equação 8.52 obtêm-se as seguintes relações:

$$T_a = \frac{1}{1 - exp\left(-\frac{1}{T_p}\right)} \tag{8.53}$$

ou

$$T_p = \frac{1}{ln(T_a) - ln(T_a - 1)} \tag{8.54}$$

A relação entre as funções acumuladas de probabilidades de séries anual e parcial, representada pela equação 8.48, está intrinsecamente relacionada à taxa média dos eventos excedentes ν, ou seja ao número médio anual de eventos a ser especificado. Como mencionado no Anexo 9, a experiência de alguns estudos anteriores indica que especificar o valor de $\hat{\nu}$ entre 2 e 3, parece trazer vantagens para o uso das séries de duração parcial, facilitando, desse modo, a garantia de

independência serial dos eventos selecionados. Além disso, outro aspecto importante na aplicação do modelo expresso pela equação 8.48, é a verificação da adequação da distribuição de Poisson às taxas de excedência dos eventos ν. Uma das maneiras de se verificar esta condição é por meio de um teste proposto por Cunnane (1979), o qual se fundamenta na aproximação da distribuição de Poisson pela distribuição Normal. Esse teste encontra-se descrito em detalhes no Anexo 9.

Exemplo 8.8 – Partindo da equação 8.48, deduzir o modelo Poisson-Pareto. Nessa situação, a taxa de excedências é poissoniana e as magnitudes dos eventos que superam o limite estabelecido seguem a distribuição de Generalizada de Pareto. (Ver exemplos 5.5 e 5.10)
Solução: A FAP da distribuição Generalizada de Pareto é dada por:

$$H(x) = 1 - exp(-y) \tag{8.55}$$

$$\text{com } y = \begin{cases} -\dfrac{Ln\left[1 - \dfrac{k(x-\xi)}{\alpha}\right]}{k} & k \neq 0 \\ \dfrac{(x-\xi)}{\alpha} & k = 0 \end{cases}$$

onde ξ é o parâmetro de posição, α é o parâmetro de escala e κ é o parâmetro de forma. Os limites de variação de x são: para $k > 0$ $\xi \leq x \leq \xi + \dfrac{\alpha}{k}$; e para $k \leq 0$ $\xi \leq x < \infty$
Para facilitar a dedução do modelo Poisson-Pareto, as representações de $F_a(x)$ e $H_u(x)$ foram trocadas por $F(x)$ e $H(x)$, respectivamente. Assim, a equação 8.48 foi reescrita como:

$$F(x) = exp\{-\nu[1 - H(x)]\} \tag{8.56}$$

Desenvolvendo a equação 8.56 obtém-se:

$$ln(F(x)) = -\nu[1 - H(x))]$$

$$\frac{ln(F(x))}{\nu} = H(x) - 1$$

$$H(x) = 1 + \frac{1}{\nu}\ln[F(x)] \tag{8.57}$$

Igualando as equações 8.55 e 8.57, tem-se o desenvolvimento

$$1-exp(-y)=1+\frac{1}{\nu}ln[F(x)]$$

$$-exp(-y)=\frac{1}{\nu}ln[F(x)]$$

$$exp(-y)=-\frac{1}{\nu}ln[F(x)]$$

$$exp(-y)=-ln[F(x)]^{1/\nu}$$

$$-y=ln\left\{-ln[F(x)]^{\frac{1}{\nu}}\right\}$$

$$y=-ln\left\{-ln[F(x)]^{\frac{1}{\nu}}\right\} \qquad (8.58)$$

• Para $\kappa = 0$, na distribuição Generalizada de Pareto $y=\frac{(x-\xi)}{\alpha}$.

Substituindo *y* na equação 8.58, segue-se que

$$\frac{(x-\xi)}{\alpha}=-ln\left\{-ln[F(x)]^{\frac{1}{\nu}}\right\}$$

e

$$x=\xi-\alpha\,ln\left\{-ln[F(x)]^{\frac{1}{\nu}}\right\} \qquad (8.59)$$

na qual $F(x)=1-\frac{1}{T(anos)}$

Na equação 8.59, tem-se $-ln[F(x)]^{\frac{1}{\nu}}=-\frac{1}{\nu}ln[F(x)]$ e desenvolvendo a equação 8.59, os quantis também são dados por

$$x=\xi-\alpha\,ln\left\{-\frac{1}{\nu}ln[F(x)]\right\} \qquad \text{, como } ln(ab)=ln(a)+ln(b)$$

$$x=\xi-\alpha\left\{ln\left(\frac{1}{\nu}\right)+ln(-Ln[F(x)])\right\} \qquad ln\left(\frac{a}{b}\right)=ln(a)-ln(b)$$

$$x=\xi-\alpha\{ln(1)-ln(\nu)+ln(-ln[F(x)])\} \qquad ln(1)=0$$

$$x=\xi-\alpha\{-ln(\nu)+ln(-Ln[F(x)])\} \qquad \text{ou}$$

$$x=\xi+\alpha\{ln(\nu)-ln(-ln[F(x)])\} \qquad (8.60)$$

• Para $\kappa \neq 0$, na distribuição Generalizada de Pareto $y = -\frac{Ln\left[1-\frac{k(x-\xi)}{\alpha}\right]}{k}$. Substituindo y na equação 8.58, segue-se que

$$-\frac{ln\left[1-\frac{k(x-\xi)}{\alpha}\right]}{k} = -ln\left\{-ln[F(x)]^{\frac{1}{\nu}}\right\}$$

$$ln\left[1-\frac{k(x-\xi)}{\alpha}\right] = k\,ln\left\{-ln[F(x)]^{\frac{1}{\nu}}\right\}$$

$$ln\left[1-\frac{k(x-\xi)}{\alpha}\right] = ln\left\{-ln[F(x)]^{\frac{1}{\nu}}\right\}^{k}$$

$$1-\frac{k(x-\xi)}{\alpha} = \left\{-ln[F(x)]^{\frac{1}{\nu}}\right\}^{k}$$

$$\frac{k(x-\xi)}{\alpha} = 1-\left\{-ln[F(x)]^{\frac{1}{\nu}}\right\}^{k}$$

$$x = \xi + \frac{\alpha}{k}\left\{1-\left[-ln[F(x)]^{\frac{1}{\nu}}\right]^{k}\right\} \quad \text{ou}$$

$$x = \xi + \frac{\alpha}{k}\left\{1-\left[-\frac{ln[F(x)]}{\nu}\right]^{k}\right\} \qquad (8.61)$$

na qual $F(x) = 1 - \frac{1}{T(anos)}$

Em resumo, pode-se dizer que, conhecendo-se a taxa de excedência ν e os parâmetros da distribuição Generalizada de Pareto, esses estimados a partir das excedências sobre o limiar estabelecido *u*, os quantis anuais podem ser calculados por meio das equações 8.60 ou 8.61, conforme o caso.

Exemplo 8.9 – Ajustar o modelo Poisson-Pareto aos dados de uma série de duração parcial de precipitação com duas horas de duração da estação pluviográfica de Entre Rios de Minas, código 02044007. O período de dados disponíveis é de 13 anos hidrológicos (73/74 a 85/86); o valor limiar estabelecido para definição da série é 39 mm e a taxa de excedência ν, é igual a 2.

Solução: A primeira etapa consiste em verificar se as taxas de excedências anuais seguem um modelo poissoniano. Esta verificação é realizada com o teste de Cunnane (1979), que está descrito no Anexo 9. Inicialmente é feita a contagem do número de eventos por ano que superam o limite estabelecido. Esses valores permitem a estimativa da estatística do teste de Cunnane, equação A9.11 do Anexo 9. O número de excedências e a estatística de Cunnane estão na Tabela 8.16.

Tabela 8.16 – Contagem das excedências anuais

AH	73/74	74/75	75/76	76/77	77/78	78/79	79/80	80/81	81/82	82/83	83/84	84/85	85/86	
m	1	3	1	2	2	3	2	4	2	2	0	3	1	Soma
γ	0,5	0,5	0,5	0	0	0,5	0	2	0	0	2	0,5	0,5	7

O valor da estatística do teste deve ser comparado ao quantil $\chi^2_{1-\alpha,\eta}$ da distribuição do Qui-Quadrado, com 12 graus de liberdade (*n*-1), e nível de significância α de 5%. Analisando o Anexo 6, verifica-se que $\chi^2_{0,95;12}$ é igual a 21. Como a estatística de Cunnane, $\gamma = 7$, é menor que o quantil $\chi^2_{0,95;12}$ da distribuição Qui-Quadrado, aceita-se, a um nível de significância de 5%, a hipótese de que as excedências anuais ocorrem segundo um modelo poissoniano.

Em seguida, pode-se calcular a distribuição empírica por meio da estimativa da posição de plotagem e dos tempos de retorno da série parcial e o seu correspondente anual. O cálculo da posição de plotagem foi realizado com a fórmula de Gringorten, $q_i = (i - 0{,}44)/(n + 0{,}12)$; o tempo de retorno parcial foi estimado com equação 8.51, $T_p = 1/\nu q_i$,e o seu correspondente anual com a equação 8.53, $T_a = 1/\{1 - exp(-1/T_p)\}$. A série parcial e os resultados de cálculo estão na Tabela 8.17.

A FAP da distribuição Generalizada de Pareto está apresentada no exemplo 8.8, equação 8.55. A estimativa dos parâmetros pelo método dos momentos é realizada do modo descrito a seguir.
ξ é o parâmetro de posição e nesse caso é igual ao limite estabelecido, ou seja, 39 mm, α é o parâmetro de escala, estimado por

$$\hat{\alpha} = \frac{\overline{X}}{2}\left(\frac{\overline{X}^2}{S_X^2} + 1\right) \quad \text{com} \quad X = x_i - \xi \tag{8.62}$$

e *k* é o parâmetro de forma, cuja estimativa é dada por

$$\hat{k} = \frac{1}{2}\left(\frac{\overline{X}^{\,2}}{S_X^2} - 1 \right) \quad \text{com } X = x_i - \xi \tag{8.63}$$

onde $\overline{X}$ e S_X são a média e o desvio padrão amostrais da variável $X = x - \xi$

A média e o desvio padrão da variável $X = x - \xi$ são, respectivamente, 10,57692mm e 11,06318mm. Os parâmetros estimados são $\hat{\alpha} = 10{,}12226$ e $\hat{k} = -0{,}04299$. Como o parâmetro de forma, *k*, é negativo, o cálculo dos quantis anuais é realizado pela equação 8.61, ou seja,

$$x(F) = 39 + \frac{10{,}12226}{-0{,}04299}\left\{ 1 - \left[-\frac{Ln(F(x))}{2} \right]^{-0{,}04299} \right\} \tag{8.64}$$

na qual $F(x) = 1 - \frac{1}{T(anos)}$

Tabela 8.17 – Cálculo da distribuição empírica do exemplo 8.9

i	AH	P (mm)	X -ξ	q_e	T_p	T_a
1	77/78	80	41	0,0214	23,3214	23,8250
2	84/85	73,4	34,4	0,0597	8,3718	8,8817
3	80/81	64,1	25,1	0,0980	5,1016	5,6179
4	84/85	63,1	24,1	0,1363	3,6685	4,1912
5	78/79	61,1	22,1	0,1746	2,8640	3,3931
6	81/82	57,2	18,2	0,2129	2,3489	2,8843
7	77/78	55,6	16,6	0,2511	1,9909	2,5325
8	82/83	53,1	14,1	0,2894	1,7275	2,2755
9	73/74	51,1	12,1	0,3277	1,5257	2,0799
10	84/85	48,6	9,6	0,3660	1,3661	1,9266
11	79/80	48,4	9,4	0,4043	1,2367	1,8034
12	81/82	48,3	9,3	0,4426	1,1298	1,7026
13	75/76	47,4	8,4	0,4809	1,0398	1,6187
14	76/77	44,5	5,5	0,5191	0,9631	1,5481
15	80/81	44,3	5,3	0,5574	0,8970	1,4880
16	80/81	43,6	4,6	0,5957	0,8393	1,4363
17	82/83	43,3	4,3	0,6340	0,7886	1,3916
18	85/86	41,2	2,2	0,6723	0,7437	1,3525
19	78/79	41	2	0,7106	0,7037	1,3183
20	80/81	40,8	1,8	0,7489	0,6677	1,2881
21	74/75	40,5	1,5	0,7871	0,6352	1,2613
22	78/79	40,2	1,2	0,8254	0,6058	1,2375
23	74/75	40	1	0,8637	0,5789	1,2162
24	74/75	39,6	0,6	0,9020	0,5543	1,1971
25	76/77	39,4	0,4	0,9403	0,5318	1,1799
26	79/80	39,2	0,2	0,9786	0,5110	1,1645

n = 26; ν = 2; *u* = ξ = *39 mm* e o número de anos igual a 13

A equação 8.64 permite que se calcule os quantis anuais associados a diferentes tempos de retorno. Alguns quantis estão apresentados na Tabela 8.18. A Figura 8.10 apresenta os quantis anuais do modelo Poisson-Pareto, calculados pela equação 8.64, e os quantis empíricos, apresentados da Tabela 8.17, ambos grafados em um papel de probabilidade de Gumbel.

Tabela 8.18 – Quantis anuais - Modelo Poisson-Pareto

T (anos)	2	5	10	20	30	50	75	100
Quantis (mm)	50,0	62,3	70,8	79,2	84,1	90,4	95,5	99,2

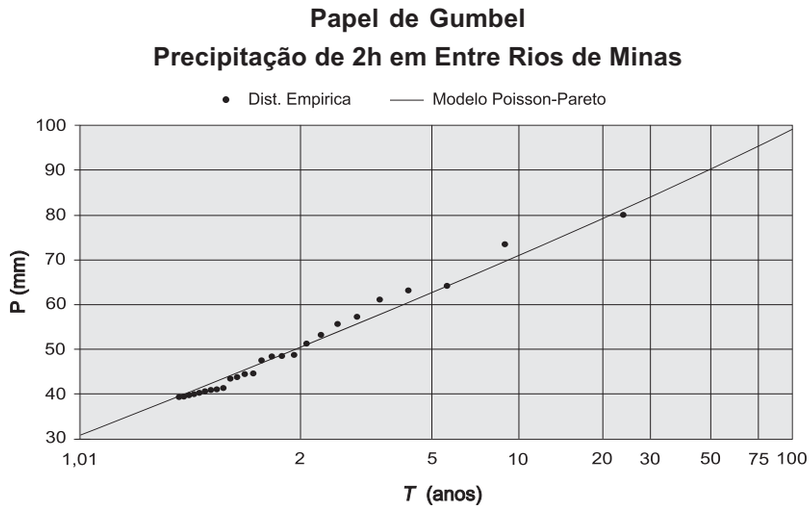


Figura 8.10 – Ajuste do modelo Poisson-Pareto à distribuição empírica

Exercícios

1) Construir os papeis de probabilidade das seguintes distribuições:
a) Log-Normal de 2 parâmetros
b) Gumbel

2) Grafar os dados de vazões médias anuais do rio Paraopeba em Ponte Nova do Paraopeba, apresentados na Tabela 7.1, nos papeis de probabilidade Normal e Log-Normal. Qual dessas distribuições parece se ajustar melhor aos dados?

3) Grafar os dados de vazões médias diárias máximas anuais do rio Paraopeba em Ponte Nova do Paraopeba, apresentados no Anexo 2, utilizando todas as fórmulas de posição de plotagem da Tabela 7.19 no mesmo papel de probabilidade. Avaliar as diferenças entre os resultados.

4) Os dados da Tabela 8.19 referem-se às descargas médias diárias máximas anuais (m^3/s) observadas no rio Hercílio em Ibirama, Sta. Catarina. A área de drenagem é de 3314 km^2. Utilize o papel de probabilidades de Gumbel, construído para o exercício 1, e plote as vazões versus posição de plotagem, utilizando a fórmula de Gringorten [(*m*-0,44)/(*n*+0,12)].

Tabela 8.19 – Dados do exercício 4

Ano	Q (m^3/s)	Ano	Q (m^3/s)	Ano	Q (m^3/s)	Ano	Q (m^3/s)	Ano	Q (m^3/s)
1935	1342	1945	474	1955	969	1965	708	1975	1406
1936	625	1946	763	1956	566	1966	998	1976	801
1937	619	1947	592	1957	1300	1967	477	1977	741
1938	797	1948	981	1958	526	1968	298	1978	1002
1939	1250	1949	438	1959	520	1969	872	1979	1090
1940	271	1950	281	1960	487	1970	483	1980	faltoso
1941	263	1951	556	1961	897	1971	1040	1981	589
1942	566	1952	393	1962	582	1972	1010	1982	490
1943	649	1953	726	1963	510	1973	1240	1983	2475
1944	236	1954	897	1964	faltoso	1974	697	1984	2125

5) Ajuste a distribuição de Gumbel à amostra do rio Hercílio em Ibirama, calcule os quantis para T = 5, 25, 50, 100 e 500 anos pelo método dos fatores de freqüência e plote a reta obtida no gráfico do exercício 1. Calcule e plote também os intervalos de confiança a um nível $100(1-\alpha) = 95\%$ para os quantis estimados.

6) Ajuste uma distribuição log-Pearson III à amostra do rio Hercílio em Ibirama, calcule os quantis e os intervalos de confiança a 95%, correspondentes a T = 5, 25, 50, 100 e 500 anos, usando o método dos fatores de freqüência.

7) O programa ALEA, disponível para download a partir da URL http://www.ehr.ufmg.br, contém rotinas para ao ajuste da distribuição GEV, pelos métodos dos momentos e da máxima verossimilhança. Utilize o programa ALEA para ajustar a distribuição Generalizada Valores Extremos à amostra do rio Hercílio em Ibirama, pelo método da máxima verossimilhança. Calcule os quantis para T = 5, 25, 50, 100 e 500 anos pela expressão da função inversa da GEV. Plote os quantis da GEV no gráfico do exercício 5 e comente sobre a influência do sinal do parâmetro de forma nos resultados obtidos.

8) A Tabela 8.20 apresenta os dados de precipitação máxima diária de Caeté, código 01943010. Realizar uma análise analítica de freqüência de máximos, calculando os parâmetros das distribuições candidatas pelo método dos momentos-L.

Tabela 8.20 – Dados do exercício 8

AH	P (mm)	AH	P (mm)	AH	P (mm)	AH	P (mm)
41/42	72,8	53/54	87	77/78	210,2	89/90	97,7
42/43	69,4	54/55	112,8	78/79	92,1	90/91	116,2
43/44	77,8	56/57	80,1	79/80	86,5	91/92	100,9
44/45	74,2	58/59	95,7	80/81	86,3	92/93	66,2
45/46	102,2	59/60	102,3	81/82	123,6	93/94	84,2
46/47	93,4	60/61	105,5	82/83	84,6	94/95	93,4
47/48	75	64/65	75,9	83/84	64,6	95/96	147,1
48/49	117,4	66/67	112,7	84/85	80,7	96/97	118,2
49/50	47,2	67/68	50,7	85/86	73	97/98	67,5
50/51	67,4	69/70	82,8	86/87	83,4	98/99	107,3
51/52	76	70/71	52	87/88	73,6	99/00	102,8
52/53	102,6	76/77	66,9	88/89	57,2		

9) Repetir o exercício 8 realizando a análise de freqüência com métodos do fator de freqüência utilizando as mesmas distribuições candidatas. Comparar com os resultados do exercício 8.

10) A série utilizada nos exercício 8 apresenta um *outlier* de 210,2mm. Qual é a probabilidade desse evento atípico ocorrer em período de 50 anos, admitindo que as precipitações máximas diárias de Caeté seguem a distribuição ajustada no exercício 8.

11) O Anexo 2 apresenta os dados de vazões mínimas com duração de 7 dias da estação fluviométrica de Ponte Nova do Paraopeba, código 40800001. Utilizando esses dados, ajustar as seguintes distribuições:
a) Gumbel para mínimos com os parâmetros calculados pelo método dos momentos-L
b) Weibull (2P) com os parâmetros calculados pelo método dos momentos-L
Para a solução deste exercício, destaca-se que Stedinger et al. (1993) indicam que, caso uma variável aleatória X se ajuste à distribuição de Weibull, então a variável $Y = -ln[X]$ se distribui segundo a distribuição de Gumbel. Assim, os procedimentos de estimativa dos parâmetros e os testes de ajuste disponíveis para a distribuição de Gumbel podem ser utilizados para a distribuição de Weibull. Desse modo, $+ln[X]$ possui uma média $\lambda_{1,(\ln X)}$ e o momento-L $\lambda_{2,(\ln X)}$, os parâmetros de ajuste da distribuição de Weibull (2P) para a variável X são os seguintes:

$$k = \frac{ln(2)}{\lambda_{2,(ln\,X)}} \tag{8.65}$$

$$\alpha = exp\left(\lambda_{1,(ln\,X)} + \frac{0{,}5772}{k}\right) \tag{8.66}$$

Para efetuar o ajuste da distribuição de Weibull, é necessário calcular os logarítmos naturais dos valores das séries. Em seguida, são calculados os momentos-L. Os valores dos momentos-L permitem o cálculo dos parâmetros da distribuição de Weibull através das equações 8.65 e 8.66.

12) Grafar as distribuições empírica e teóricas ajustadas no exercício 11 em um papel de probabilidades de Gumbel utilizando a fórmula de posição de plotagem de Gringorten.

13) Considerando o exercício 11, calcular a probabilidade de ocorrência de vazões mínimas com duração de 7 dias inferiores a $Q_{7,10}$ durante um período de 5 anos.

14) Montar uma tabela que contenha os valores do fator de freqüência da distribuição de Weibull, para mínimos, em função da assimetria amostral e do tempo de retorno.

15) Montar uma tabela que contenha os valores do fator de freqüência da distribuição de Gumbel (máximos), em função do tempo de retorno e do tamanho da amostra.

16) Ajustar o modelo Poisson-Pareto à série de duração parcial de precipitações, com duração de 3h, na estação de Pium-í, código 02045012, apresentada na Tabela 8.21. Essa série se refere ao período de anos (75/76 a 85/86), com taxa média de excedências de 2 eventos por ano para um limite de 44,5mm.

Tabela 8.21 – Dados do exercício 16

AH	P (mm)	AH	P (mm)	AH	P (mm)	AH	P (mm)
75/76	70,2	78/79	47,6	81/82	53	83/84	46,6
75/76	50	79/80	49,8	82/83	47,9	84/85	72,2
76/77	47,2	79/80	46	82/83	59,4	84/85	46,4
77/78	52	79/80	46,8	82/83	50,2	85/86	48,4
77/78	47,6	80/81	50,6	82/83	53,4		
77/78	47,4	81/82	44,1	82/83	59,4		

17) O Rio Alva em Ponte de Mucela, em Portugal, apresenta um número médio de 3 excedências por ano sobre a descarga de referência de 65 m^3/s. Testes estatísticos comprovaram serem plausíveis as hipóteses nulas do número Poissoniano de excedências, independência serial e exponencialidade da cauda superior, a um nível de significância de 5%. Se a média das excedências é de 72,9 m^3/s, estime a descarga máxima anual de tempo de retorno 500 anos.

18) A Tabela 8.22 apresenta as 205 maiores enchentes ao longo dos 72 anos contínuos (1896-1967) de registros fluviométricos do Rio Greenbrier em Alderson (West Virginia, EUA) e que excederam 17000 cfs.

a) Escolha o maior valor possível para o número médio anual de cheias ($\hat{\Lambda}$), tal que as excedências possam ser modeladas por um processo de Poisson. Verifique a conveniência de sua escolha através do teste da hipótese Poissoniana pela estatística

$\gamma = \sum_{k=1}^{N} \left(\frac{m_k - \hat{\Lambda}}{\hat{\Lambda}} \right)^2$ lembrando que essa segue uma distribuição do Qui-Quadrado

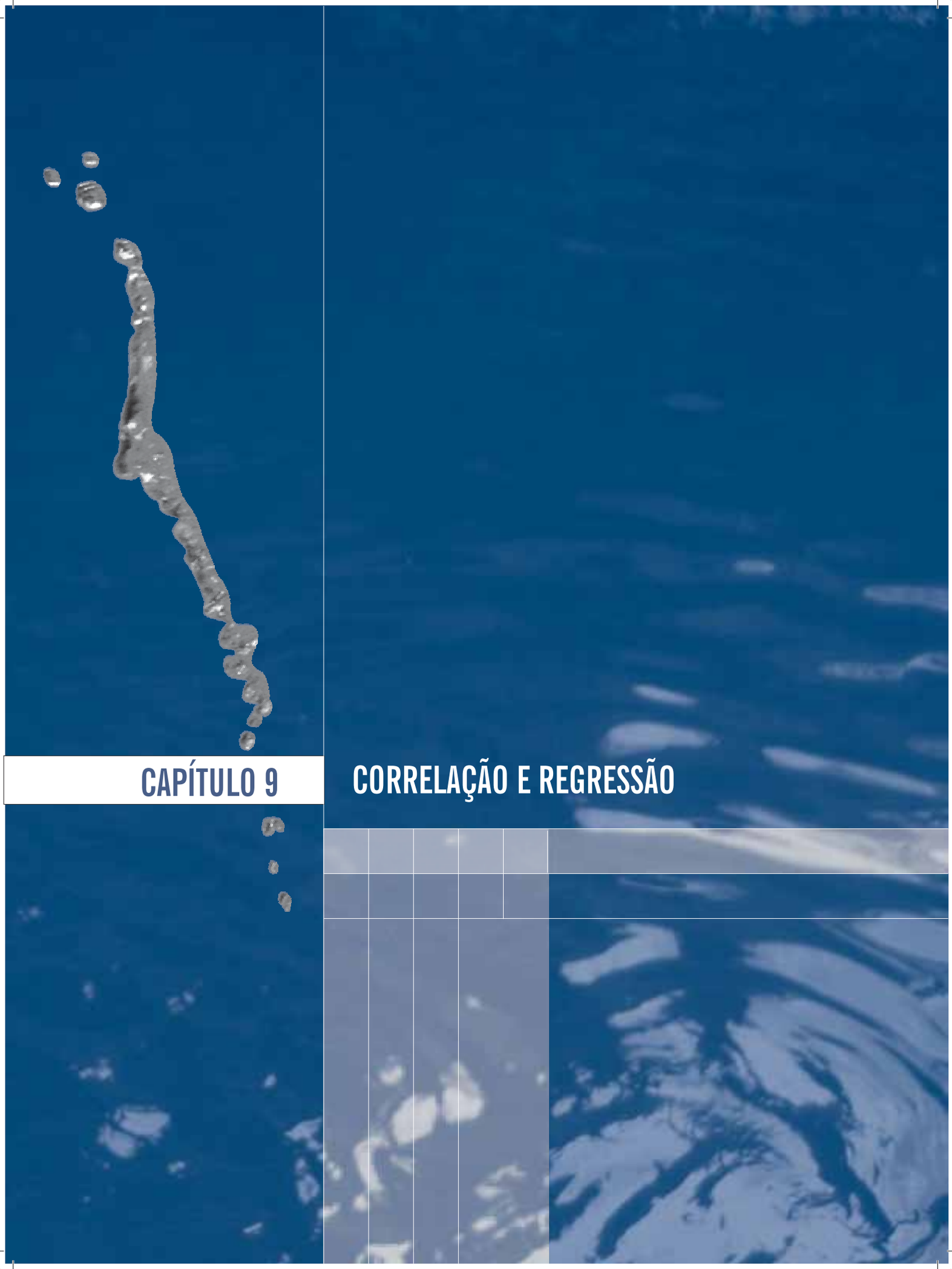
com (N-1) graus de liberdade, onde N indica o número de anos de registros, e que o número de excedências que ocorrem no ano k é representado por m_k,.

b) Depois de escolhido o maior valor possível para o número médio anual de enchentes, modele-as através da distribuição generalizada de Pareto dada por $H(x) = 1 - exp(-y)$ onde e $y = -\frac{ln\left[1 - \frac{k}{\alpha}(x - \xi)\right]}{k}$ para $k \neq 0$, $y = \frac{x - \xi}{\alpha}$ para $k = 0$ e ξ, α e k são, respectivamente, os parâmetros de posição, escala e forma. Lembre-se que a distribuição generalizada de Pareto é ilimitada superiormente para $k \leq 0$ e possui limite superior para $k > 0$. Observe que, quando $k = 0$, ela se reduz à distribuição exponencial com parâmetros ξ e α.

c) Calcule os quantis de cheias anuais para diversos tempos de retorno (2 a 1000 anos) invertendo a expressão da função de distribuição acumulada de probabilidades anuais do modelo Poisson-Pareto dada por $F(x) = exp\{-\Lambda[1 - H(x)]\}$. Plote os quantis com o tempo de retorno (T em coordenadas logarítmicas).

Tabela 8.22 – Vazões do rio Greenbrier em Alderson (West Virginia, EUA) superiores a 17.000 cfs

Ano	Q (cfs)	Ano	Q (cfs)	Ano	Q (cfs)	Ano	Q (cfs)
1896	28800	1915	34000	1935	20800	1954	29700
1897	27600		40800	1936	19400		18800
	54000	1916	27200		20800	1955	32000
	40900		24400		27100		28000
1898	17100	1917	17300		58600		44400
	18600		43000		28300		26200
	52500		28000	1937	21200	1956	18200
1899	25300	1918	17900		22300	1957	23900
	20000		77500		36600		28900
	23800		24000		26400		22000
	48900	1919	28600	1938	21200	1958	21800
1900	17100		24800		32800		23900
1901	56800		49000		22300		22200
	21100	1920	38000	1939	40200		17500
	20400		20700		41600		26700
	19300		33500		21200	1959	17200
	20000	1922	21500		17200		23900
1902	36700		20100		19400	1960	17800
	43800		22200	1940	29900		35500
1903	25300	1923	19500		21500		32500
	29600	1924	26500		19400	1961	25000
	33500		20400		18700		21800
	34400		36200	1942	35300		31400
	48900		17900	1943	33600		17200
1904	25700	1926	20700		17200	1962	34700
	25700		17600		36200		20100
1905	29600	1927	17900		21200		21500
	37600		24000	1944	25200		17800
1906	18200		40200		17200		23200
	26000		18800	1945	17900		35500
1907	17500		19500		19000	1963	22700
	52500	1928	18000	1946	43600		34800
1908	17800	1929	22800	1947	20000		47200
	23000		32700		24400		26100
	31500		23800	1948	35200		30400
	52500		20000		23500	1964	19100
	26800	1930	36600		40300		39600
	27600	1932	50100	1949	18500		22800
	31500		17600		37100	1965	22000
1909	20000		31500		26300		28400
1910	45900		27500		23200		19800
1911	43800		21900	1950	31500		18600
	20000	1933	26400	1951	25600	1966	26400
	23800	1934	32300		27800	1967	54500
	18900		20500		26700		39900
	18900		27900		18500		20900
	35500	1935	19400		19800		
	27200		49600		29300		
	20000		22300	1952	17800		
	21100		17900		19100		
1913	21800		24800		27600		
	64000		20100	1953	47100		
	20000		24800		20100		



CAPÍTULO 9

CORRELAÇÃO E REGRESSÃO

CAPÍTULO 9 CORRELAÇÃO E REGRESSÃO

Existe um conjunto de métodos estatísticos que visam estudar a associação entre duas ou mais variáveis aleatórias. Dentre tais métodos, a teoria da regressão e correlação ocupa um lugar de destaque por ser o de uso mais difundido. Neste capítulo serão abordados os fundamentos dos métodos estatísticos da correlação e regressão, com vistas à sua aplicação em hidrologia. O objetivo deste capítulo é o de apresentar os conceitos básicos que permitam ao leitor realizar estudos de correlação e regressão linear entre duas ou mais variáveis aleatórias hidrológicas.

Na engenharia de recursos hídricos, algumas questões referem-se ao conhecimento da associação e do grau de associação entre duas ou mais variáveis, como por exemplo, as relações (i) entre as intensidades, as durações e as freqüências das precipitações intensas (ii) entre as vazões médias anuais e as áreas de drenagem ou (iii) entre as alturas anuais de precipitação e as altitudes dos postos pluviométricos. Nesses estudos, o primeiro objetivo é o de analisar o comportamento simultâneo das variáveis, tomadas duas a duas, verificando se a variação positiva (ou negativa) de uma delas está associada a uma variação positiva (ou negativa) da outra, ou mesmo, se não há nenhuma forma de dependência entre elas. Nesse sentido, uma primeira abordagem exploratória é a elaboração de um diagrama de dispersão entre as observações simultâneas das variáveis. O diagrama de dispersão permite visualizar o grau de associação entre as variáveis e a tendência de variação conjunta que apresentam. A Figura 9.1 apresenta alguns exemplos de variação conjunta entre duas variáveis.

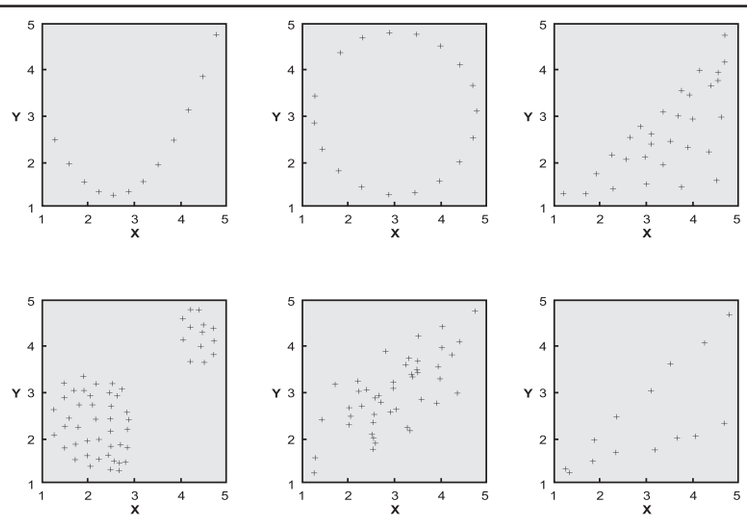


Figura 9.1 – Exemplos de relacionamento (Adaptado de Helsel e Hirsh, 1992)

A medida da variação conjunta das variáveis ou co-variação observada em um diagrama de dispersão é a correlação entre as duas variáveis. Essa medida é realizada numericamente por meio dos coeficientes de correlação que representam o grau de associação entre duas variáveis continuas. As medidas genéricas de correlação, freqüentemente são designadas por ρ, são adimensionais e variam entre -1 e +1. No caso de $\rho = 0$, não existe correlação entre as duas variáveis. Quando $\rho > 0$, a correlação é positiva e uma variável aumenta quando a outra cresce. A correlação é negativa, $\rho < 0$,quando as variáveis variam em direções opostas.

A correlação é chamada de monotônica se uma das variáveis aumenta ou diminui sistematicamente quando a outra decresce, com associações que podem ter forma linear ou não linear. A Figura 9.2 apresenta exemplos de correlações monotônicas não lineares e não monotônicas.

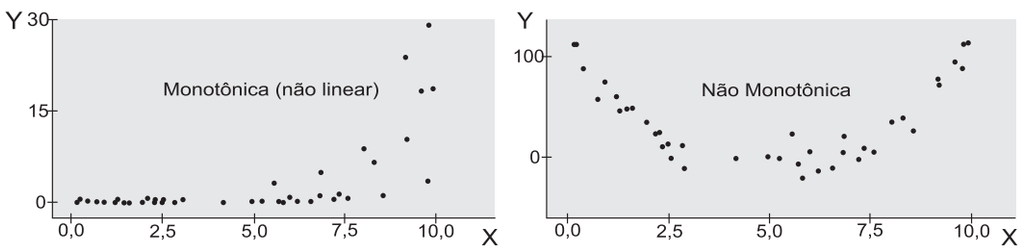


Figura 9.2 – Exemplos de correlações (Adaptado de Helsel e Hirsh, 1992)

É importante salientar que variáveis altamente correlacionadas *não* apresentam necessariamente qualquer relação de causa e efeito. A correlação representa simplesmente a tendência que as variáveis apresentam quanto à sua variação conjunta. Assim, a medida da correlação não indica necessariamente que há evidências de relações causais entre duas variáveis. As evidências de relações causais devem ser obtidas a partir do conhecimento dos processos envolvidos. Obviamente haverá casos em que uma variável está na origem da outra, tais como aqueles que associam a precipitação e o escoamento superficial em uma dada bacia. Entretanto, existirão situações em que as variáveis apresentam a mesma causa, como, por exemplo, a eventual forte correlação entre as vazões médias mensais de duas bacias vizinhas não significa que a mudança da vazão de uma delas é causada pela alteração da outra; certamente, as alterações são causadas por fatores comuns às duas bacias.

9.1 – Coeficiente de Correlação Linear de Pearson

Duas variáveis apresentam uma correlação linear quando os pontos do diagrama de dispersão se aproximam de uma reta. Essa correlação pode ser positiva (para valores crescentes de *X*, há uma tendência a valores também crescentes de *Y*) ou negativa (para valores crescentes de *X*, a tendência é observarem-se valores decrescentes de *Y*). As correlações lineares positivas e negativas encontram-se ilustradas na Figura 9.3.

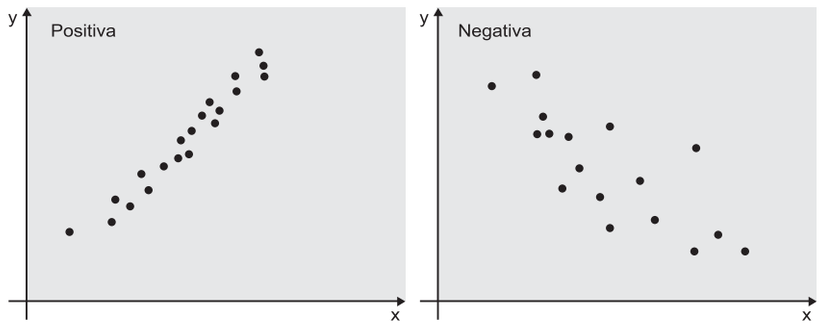


Figura 9.3 – Correlações Lineares Positivas e Negativas

O coeficiente de correlação linear, também chamado de covariância normalizada e representado por ρ, é expresso por:

$$\rho_{X,Y} = \frac{\sigma_{X,Y}}{\sigma_X \sigma_Y} \tag{9.1}$$

onde, $\sigma_{X,Y}$ é a covariância entre as variáveis *X* e *Y*; σ_X e σ_Y são os desvios-padrão das variáveis *X* e *Y*, respectivamente.

Quando duas variáveis, *X* e *Y*, são estatisticamente independentes, o coeficiente de correlação linear é igual a zero, $\rho = 0$. Entretanto a recíproca não é verdadeira, ou seja, se o coeficiente de correlação linear é igual a zero, $\rho = 0$, isso não significa que as variáveis são independentes. Trata-se de uma decorrência do fato de que o coeficiente de correlação linear, ρ, é uma medida da dependência linear entre as variáveis *X* e *Y*, e, em algumas situações, *X* e *Y* podem apresentar dependência funcional não linear.

A covariância entre duas variáveispode ser estimada pela equação 9.2 e representa uma medida possível do grau e do sinal da correlação.

$$s_{X,Y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1} \tag{9.2}$$

onde, s_{xy} é a covariância amostral entre as variáveis X e Y; $\bar{x}$ e $\bar{y}$ são as médias aritméticas de cada uma das variáveis; n é o tamanho da amostra; x_i e y_i são as observações simultâneas das variáveis.

Entretanto, admitindo-se que a distribuição conjunta das variáveis X e Y é uma distribuição normal bivariada, torna-se conveniente utilizar, como medida da correlação, o chamado coeficiente de correlação linear de Pearson cujo estimador é apresentado a seguir:

$$r = \frac{s_{X,Y}}{s_X s_Y} \tag{9.3}$$

Na equação 9.3, r é coeficiente de correlação linear ($-1 \leq r \geq 1$), s_{XY} é covariância entre as variáveis, s_X e s_Y são os desvios-padrão das amostras calculados pelas equações:

$$s_X = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} \tag{9.4}$$

$$s_Y = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}} \tag{9.5}$$

O coeficiente de correlação linear de Pearson é adimensional e varia entre -1 e +1, o que não ocorre com a covariância. Assim, as unidades adotadas pelas variáveis não afetam o valor do coeficiente de correlação. Caso os dados se alinhem perfeitamente ao longo de uma reta com declividade positiva teremos a correlação linear positiva perfeita com o coeficiente de Pearson igual a 1. A correlação linear negativa perfeita ocorre quando os dados se alinham perfeitamente ao longo de uma reta com declividade negativa e o coeficiente de correlação de Pearson é igual a -1. O significado de valores intermediários é facilmente percebido. A Figura 9.4 apresenta alguns diagramas de dispersão com os respectivos valores do coeficiente de correlação.

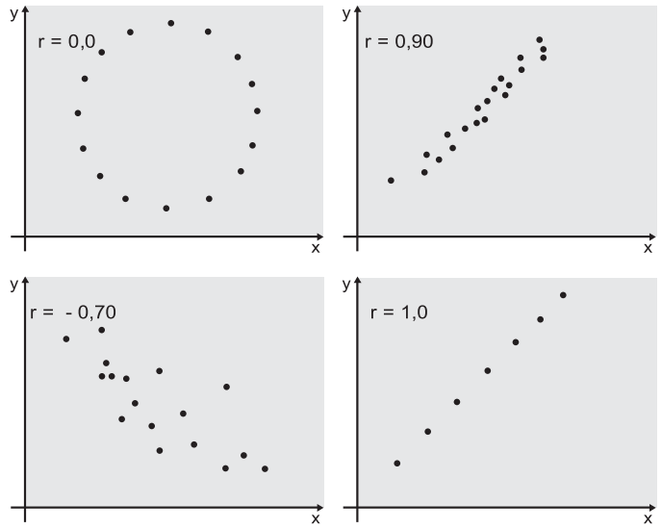


Figura 9.4 – Exemplos de coeficientes de correlação

Ressalta-se, novamente, que um valor do coeficiente de correlação alto, embora estatisticamente significativo, não implica necessariamente numa relação de causa e efeito, mas, simplesmente indica a tendência que aquelas variáveis apresentam quanto à sua variação conjunta.

Outro cuidado que se deve tomar na análise de duas variáveis é com a ocorrência de correlações espúrias, ou seja, qualquer correlação aparente entre duas variáveis que não são correlacionadas de fato. As causas mais freqüentes da ocorrência dessas correlações são: a distribuição não equilibrada dos dados, como está apresentada na Figura 9.5; a relação entre quocientes de variáveis que apresentam o mesmo denominador, ilustrado na Figura 9.6, e a relação de variáveis que foram multiplicadas por uma delas, tal como mostrado na Figura 9.7.

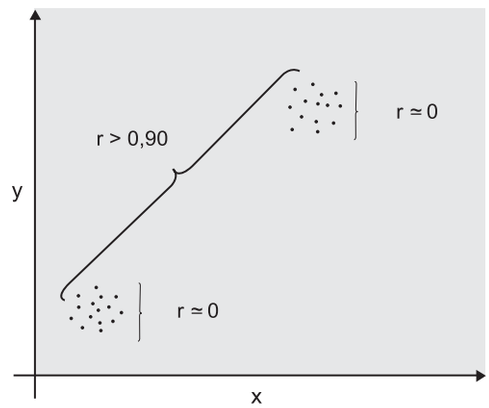


Figura 9.5 – Distribuição não equilibrada dos dados

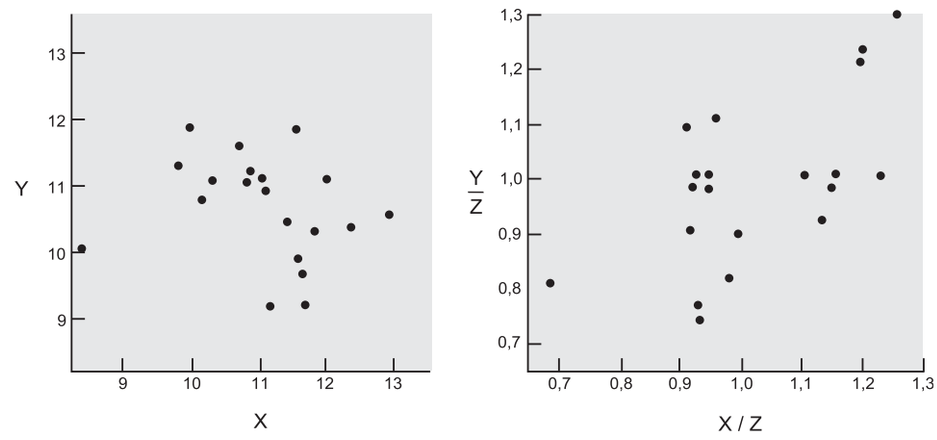


Figura 9.6 – Correlação entre quocientes de variáveis

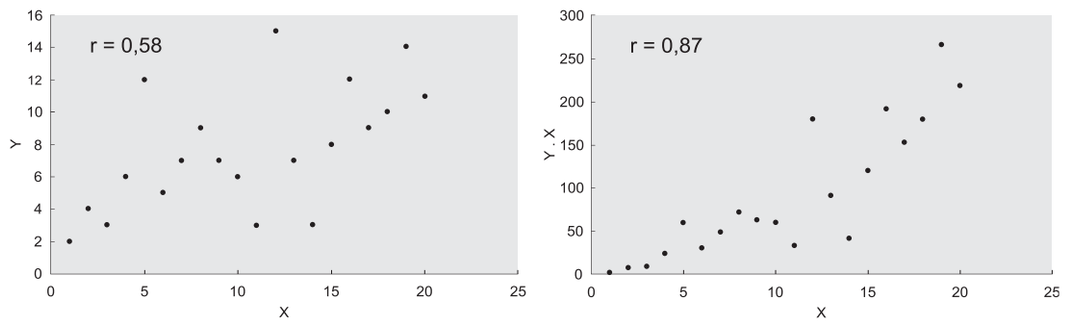


Figura 9.7 – Correlação entre produto de variáveis

9.1.1 – Testes de Hipóteses sobre o Coeficiente de Correlação

É possível testar a hipótese de que o coeficiente de correlação linear é igual a zero, ou seja:

$H_0 : \rho = 0$

$H_1 : \rho \neq 0$

Como decorrência de algumas hipóteses distributivas, a estatística apropriada para esse teste é a seguinte:

$$t_0 = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \qquad (9.6)$$

onde, t_0 é a estatística do teste; n é o tamanho da amostra e r é a estimativa do coeficiente de correlação linear.

A estatística do teste, t_0 , segue uma distribuição t de Student com $(n-2)$ graus de liberdade, sob a plausibilidade da hipótese nula H_0 : $\rho = 0$. A hipótese nula é rejeitada se:

$$|t_0| > t_{\alpha/2,n-2} \tag{9.7}$$

onde, $t_{\alpha/2,n-2}$ é o valor crítico para a estatística do teste bilateral para um nível de significância α, com $(n-2)$ graus de liberdade.

Testar hipóteses para o coeficiente de correlação, ρ_0 , diferente de zero, conforme apresentado a seguir, é um pouco mais complicado.

$$H_0 : \rho = \rho_0$$
$$H_1 : \rho \neq \rho_0$$

Segundo Montgomery e Peck (1992), para amostras de tamanho razoável ($n \geq 25$), a estatística:

$$Z = arctan\, h(r) = \frac{1}{2} ln\left(\frac{1+r}{1-r}\right) \tag{9.8}$$

é aproximadamente normalmente distribuída com média

$$\mu_Z = arctan\, h(\rho) = \frac{1}{2} ln\left(\frac{1+\rho}{1-\rho}\right) \tag{9.9}$$

e variância

$$\sigma_Z^2 = (n-3)^{-1} \tag{9.10}$$

Para testar a hipótese nula, $\rho = \rho_0$, pode ser calculada a estatística

$$Z_0 = [arctan\, h(r) - arctan\, h(\rho_0)](n-3)^{1/2} \tag{9.11}$$

A hipótese nula será rejeitada se:

$$|Z_0| > Z_{\alpha/2} \tag{9.12}$$

onde, $Z_{\alpha/2}$ é o valor crítico para a estatística do teste bilateral, a qual é dada pela

variável central reduzida da distribuição normal padrão associada a um nível de significância α.

Segundo os mesmos autores, também é possível construir um intervalo de confiança, 100(1-α), para ρ utilizando a transformação obtida pela equação (9.8). Nesse caso, o intervalo de confiança é dado por

$$tanh\left[arctan\,h(r) - \frac{Z_{\alpha/2}}{\sqrt{n-3}} \right] \leq \rho \leq tanh\left[arctan\,h(r) + \frac{Z_{\alpha/2}}{\sqrt{n-3}} \right] \qquad (9.13)$$

onde *r* é o coeficiente de correlação estimado, $Z_{\alpha/2}$ é o quantil da distribuição normal padronizada com um nível de significância α, *n* é tamanho da amostra e

$$tanh(u) = \frac{\left(e^{u} - e^{-u}\right)}{\left(e^{u} + e^{-u}\right)} \qquad (9.14)$$

9.2 – Regressão Linear Simples

Muitas vezes, a simples visualização do diagrama de dispersão sugere a existência de uma relação funcional entre as duas variáveis. Essa observação introduz o problema de se determinar uma função que exprima esse relacionamento. A análise de regressão é uma técnica estatística cujo escopo é investigar e modelar a relação entre variáveis.

Considerando que exista um relacionamento funcional entre os valores *Y* e *X*, responsável pelo aspecto do diagrama, essa função deverá explicar parcela significativa da variação de *Y* com *X*. Contudo, uma parcela da variação permanece inexplicada e deve ser atribuída ao acaso. Colocando em outros termos, admite-se a existência de uma função que explica, em termos médios, a variação de uma das variáveis com a variação da outra. Freqüentemente, os pontos observados apresentarão uma variação em torno da linha da função de regressão, devido à existência de uma variação aleatória adicional denominada de *variação residual*. Portanto, essa equação de regressão fornece o valor médio de uma das variáveis em função da outra. Obviamente, caso se suponha conhecida a forma do modelo de regressão, a análise será facilitada. O problema, então, estará restrito à estimação dos parâmetros do modelo de regressão. Esse caso ocorrerá se existirem razões teóricas que permitam saber previamente que modelo rege a associação entre as variáveis. Geralmente, a forma da linha de regressão fica aparente na própria análise do diagrama de dispersão.

Admitindo ser uma reta a linha teórica de regressão, a função entre X e Y é a seguinte:

$$Y = \alpha + \beta X + e \tag{9.15}$$

onde, Y é a variável dependente, X é a variável independente, α *e* β são os coeficientes do modelo e e denota os erros ou resíduos da regressão.

Os coeficientes α *e* β da reta teórica são estimados através dos dados observados fornecidos pela amostra, obtendo uma reta estimativa na forma

$$\hat{y}_i = a + bx_i \tag{9.16}$$

onde a é a estimativa do coeficiente α ($\hat{\alpha} = a$); b é a estimativa de β ($\hat{\beta} = b$); $\hat{y}_i$ é o valor estimado da variável dependente e x_i é o valor observado da variável independente.

Existem vários métodos para a obtenção da reta desejada. O mais simples de todos, que podemos chamar de "método do ajuste visual", consiste simplesmente em traçar diretamente a reta, com auxílio de uma régua, no diagrama de dispersão, procurando fazer, da melhor forma possível, com que essa reta passe por entre os pontos. Entretanto, esse procedimento subjetivo, somente será razoável se a correlação linear for muito forte.

Um dos procedimentos objetivos mais adequados é a aplicação do método dos *mínimos quadrados*, segundo o qual a reta a ser adotada deverá ser aquela que torna mínima a soma dos quadrados dos erros ou resíduos da regressão.

9.2.1 – Método dos Mínimos Quadrados

O objetivo do método dos mínimos quadrados é encontrar a função de regressão que minimize a soma das distâncias entre a função ajustada e os pontos observados como apresentado na Figura 9.8. Adotando um modelo linear como da equação 9.15, os coeficientes α *e* β da reta teórica podem ser estimados através dos pontos experimentais fornecidos pela amostra, obtendo uma reta estimativa na forma da equação 9.16.

A distância, e_i, entre o valor observado e o valor estimado pela reta de regressão é dado por:

$$e_i = y_i - \hat{y}_i \tag{9.17}$$

onde y_i é o valor observado da variável dependente e $\hat{y}_i$ é o valor estimado da variável dependente.

Substituindo na equação 9.17 o valor estimado pela equação 9.16, $\hat{y}_i$, obtém-se:

$$e_i = y_i - a - bx_i \tag{9.18}$$

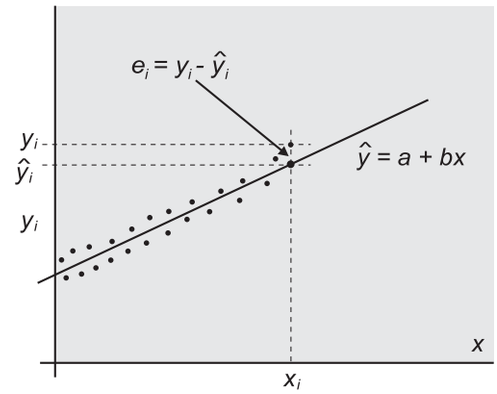


Figura 9.8 – Linha de Regressão

O método dos mínimos quadrados consiste em minimizar o somatório dos quadrados dos desvios entre o valor observado y_i e o valor estimado $\hat{y}_i$. Para o ponto indexado por *i*, o desvio quadrático é dado por

$$e_i^2 = \left(y_i - a - bx_i\right)^2 = y_i^2 - 2y_i a - 2y_i bx_i + a^2 + 2abx_i + b^2 x_i^2 \tag{9.19}$$

Para todos os *n* elementos da amostra,

$$Z = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} y_i^2 - 2a\sum_{i=1}^{n} y_i - 2b\sum_{i=1}^{n} x_i y_i + na^2 + 2ab\sum_{i=1}^{n} x_i + b^2 \sum_{i=1}^{n} x_i^2 \tag{9.20}$$

Como $Z = f\left(a,b\right)$, os valores de *a* e *b* que minimizam a equação acima são aqueles obtidos calculando-se as derivadas parciais, em relação a *a* e *b*, e igualando-as a zero,

$$\text{Mínimo de } Z \begin{cases} \dfrac{\partial Z}{\partial a} = 0 \\ \dfrac{\partial Z}{\partial b} = 0 \end{cases} \tag{9.21}$$

Calculando as derivadas para 9.20, obtém-se o seguinte sistema de equações

$$\begin{cases} \dfrac{\partial Z}{\partial a} = -2\sum_{i=1}^{n} y_i + 2na + 2b\sum_{i=1}^{n} x_i = 0 \\ \dfrac{\partial Z}{\partial b} = -2\sum_{i=1}^{n} x_i y_i + 2a\sum_{i=1}^{n} x_i + 2b\sum_{i=1}^{n} x_i^2 = 0 \end{cases} \tag{9.22}$$

Multiplicando as equações do sistema acima por (-1/2) encontra-se as equações normais da regressão linear simples:

$$\begin{cases} \sum_{i=1}^{n} y_i - na - b\sum_{i=1}^{n} x_i = 0 \\ \sum_{i=1}^{n} x_i y_i - a\sum_{i=1}^{n} x_i - b\sum_{i=1}^{n} x_i^2 = 0 \end{cases} \tag{9.23}$$

A resolução do sistema de equações normais permite a estimativa dos parâmetros do modelo de regressão linear simples a partir dos dados amostrais:

$$a = \frac{\sum_{i=1}^{n} y_i}{n} - b\frac{\sum_{i=1}^{n} x_i}{n} = \bar{y} - b\bar{x} \tag{9.24}$$

$$b = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} \tag{9.25}$$

9.3 – Coeficiente de Determinação

Após a estimativa dos coeficientes da reta de regressão, é necessário verificar se os dados amostrais são descritos pelo modelo da equação 9.16 e, além disso, determinar a parcela da variabilidade amostral que foi, de fato, explicada pela reta de regressão. Essas questões podem ser analisadas considerando a Figura 9.9, a qual possibilita a dedução da seguinte relação simples:

$$y_i = (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y}) + \bar{y} \tag{9.26}$$

A partir dessa equação, é possível demonstrar que

$$\sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \tag{9.27}$$

O primeiro membro da equação 9.27 pode ser interpretado como proporcional à variância total de *Y*, enquanto o segundo membro reflete a soma de termos

proporcionais às suas variâncias residual e explicada pelo modelo de regressão. Em outros termos,

$$SQT = SQ\,Res + SQ\,Reg \tag{9.28}$$

onde SQT é a soma quadrática total; $SQ\,Res$ é soma dos quadrados dos resíduos e $SQ\,Reg$ é a soma dos quadrados devidos à regressão.

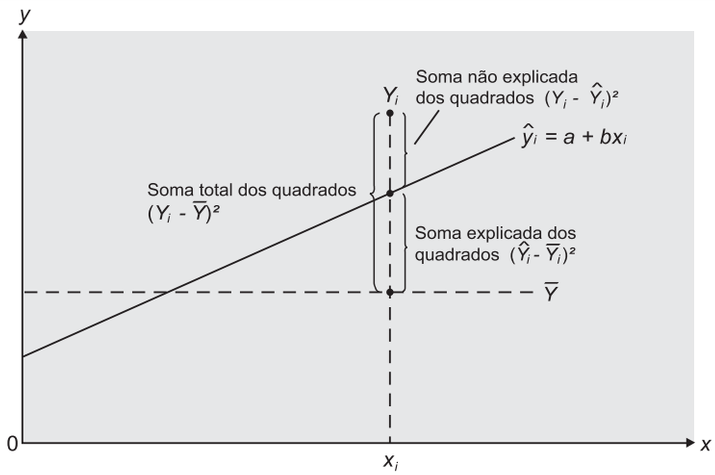


Figura 9.9 – Componentes de *Y*

O coeficiente de determinação é dado pela relação entre a soma dos quadrados devidos à regressão ($SQ\,Reg$) e a soma total dos quadrados (SQT), ou seja

$$r^2 = \frac{Variância\quad Explicada}{Variância\quad Total} = \frac{SQ\,Reg}{SQT} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{9.29}$$

onde r^2 é o coeficiente de determinação $\left(0 \le r^2 \le 1\right)$, y_i é o valor observado da variável dependente, $\hat{y}_i$ é o valor estimado da variável dependente e $\bar{y}$ é a média da variável dependente.

O coeficiente de determinação é sempre positivo e deve ser interpretado como a proporção da variância total da variável dependente *Y* que é explicada pelo modelo de regressão e que também pode ser estimado por:

$$r^2 = b^2 \frac{s_X^2}{s_Y^2} \tag{9.30}$$

onde s_X^2 é a variância amostral de X; s_Y^2 é a variância amostral de Y e b é o coeficiente angular da reta de regressão calculado pela equação 9.25.

O coeficiente de correlação amostral, r, está relacionado ao coeficiente de determinação, r^2, através da seguinte equação:

$$r = \pm\sqrt{r^2} \tag{9.31}$$

onde o sinal de r é o mesmo do de b.

9.4 – Hipóteses Básicas da Análise de Regressão Linear Simples (RLS)

As principais hipóteses da análise de regressão linear simples são a linearidade, a normalidade e a homoscedasticidade dos resíduos. A hipótese de linearidade define que a relação entre as variáveis analisadas deve ser linear, enquanto que o pressuposto de normalidade estabelece que os valores de Y são normalmente distribuídos para cada valor de X, conforme ilustrado na Figura 9.10.

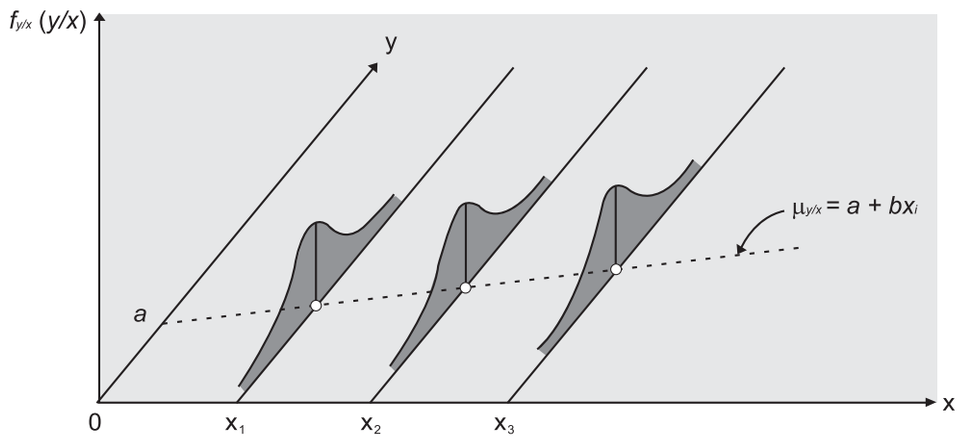


Figura 9.10 – Hipótese de normalidade

A hipótese de homoscedasticidade estabelece que os resíduos ou erros e_i, $e_i = y_i - (\alpha + \beta x_i)$, são realizações de uma variável aleatória independente e normalmente distribuída, com média zero e variância constante σ_e^2. A hipótese de homoscedasticidade dos resíduos implica nas seguintes afirmações:

a) O valor esperado da variável erro e_i é igual a zero, $E(e_i) = 0$
b) A correlação entre e_i e e_j com $i \neq j$ é é igual a zero

c) Como $Var(e_i) = Var(e_j)$, para $i \neq j$, a $Var(e_i)$ não varia com x_i, ou seja, a variância dos resíduos é constante.

9.4.1 – Erro Padrão da Estimativa

O modelo de regressão linear simples será perfeito se todos os pontos da amostra utilizados na estimativa dos parâmetros estiverem sobre a reta ajustada. Entretanto, a ocorrência de um modelo perfeito dificilmente será observada. A regressão linear simples possibilita uma estimativa aproximada de um valor de *Y* para um dado valor de *X*. Sendo assim, é importante uma medida da variabilidade dos pontos amostrais acima e abaixo da reta de regressão, tal como a dispersão esquematicamente ilustrada na Figura 9.8. Intrinsecamente ao processo de estimação dos parâmetros da reta de regressão, foi feita a premissa de que os erros são realizações de uma variável aleatória independente e normalmente distribuída com média zero, ou seja, $E(e_i) = 0$, e variância σ_e^2. Como $E(e_i) = 0$, a variância dos erros ou resíduos e_i será:

$$Var(e_i) = \sigma_e^2 = E(e_i^2) - E^2(e_i) = E(e_i^2) \tag{9.32}$$

Uma estimativa não enviesada da variância dos resíduos em torno da reta de regressão pode ser obtida por:

$$\hat{\sigma}_e^2 = s_e^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-2} = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n-2} \tag{9.33}$$

A raiz quadrada da variância dos resíduos e_i é chamada de erro padrão da estimativa, σ_e, e mede a dispersão dos resíduos em torno da reta de regressão. O erro padrão da estimativa pode ser estimado por

$$\hat{\sigma}_e = s_e = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n-2}} \tag{9.34}$$

9.5 – Teste de Hipóteses e Intervalos de Confiança para os Coeficientes da RLS

Devido à variabilidade amostral, a reta de regressão obtida da amostra extraída da população é uma das muitas retas possíveis. Os valores calculados para a e b

são estimativas pontuais dos parâmetros populacionais α e β. As retas da população e da amostra são paralelas quando $b = \beta$ e terão apenas um ponto necessariamente coincidente, a saber, a média da amostra x e a média da amostra y, quando $b \neq \beta$.

Os intervalos de confiança para os coeficientes α e β da reta de regressão são estimados por

$$a - t_{1-\frac{\alpha}{2},n-2} s_a \leq \alpha \leq a + t_{1-\frac{\alpha}{2},n-2} s_a \tag{9.35}$$

$$b - t_{1-\frac{\alpha}{2},n-2} s_b \leq \beta \leq b + t_{1-\frac{\alpha}{2},n-2} s_b \tag{9.36}$$

onde $t_{1-\frac{\alpha}{2},n-2}$ é valor do t de Student para (1- α/2) e $(n-2)$ graus de liberdade; a e b são os estimadores dos parâmetros da reta de regressão; s_a é o desvio-padrão da estimativa do parâmetro a e indica quão afastado o parâmetro estimado está do parâmetro populacional. A equação utilizada para o cálculo de s_a é dada por:

$$s_a = \sqrt{s_e^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right)} \tag{9.37}$$

s_b é desvio-padrão da estimativa de b, calculado por:

$$s_b = \sqrt{\frac{s_e^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}} \tag{9.38}$$

no cálculo de s_a e s_b tem-se:

$$s_e^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-2} \tag{9.39}$$

onde $e_i = y_i - \hat{y}_i$; n é o tamanho da amostra; $\bar{x}$ é a média da variável independente; e x_i é o valor observado da variável independente.

9.5.1 – Intervalos de Confiança para a Linha de Regressão Linear Simples

A reta obtida por mínimos quadrados é uma estimativa da função de regressão dada pela equação 9.15. De forma que, para um valor fixo x', o $\hat{y}'$ calculado pela relação $a + bx'$, corresponde a uma estimativa do valor que seria obtido pelo modelo de regressão linear, $y = \alpha + \beta x'$.

A construção de um intervalo de confiança para $\alpha + \beta x'$ pode se basear em sua estimativa, $\hat{y}'$. Considerando um valor x' que não foi utilizado no cálculo dos parâmetros da reta de regressão, demonstra-se que:

$$\mu(\hat{y}') = \alpha + \beta x' \tag{9.40}$$

$$\hat{\sigma}^2(\hat{y}') = \hat{\sigma}_e^2 \left[\frac{1}{n} + \frac{(x' - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right] \tag{9.41}$$

O intervalo de confiança para a reta de regressão é dado por:

$$\hat{y}' \pm t_{1-\frac{\alpha}{2},n-2} s_e \sqrt{\frac{1}{n} + \frac{(x' - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}} \tag{9.42}$$

onde $\hat{y}' = a + bx'$, $t_{1-\frac{\alpha}{2},n-2}$ é valor do t de Student, para (1-α/2) e $(n-2)$ graus de liberdade; e s_e é calculado pela equação 9.34.

Analisando a equação 9.42, observa-se que a amplitude do intervalo será mínima quando x' for igual ao valor médio da amostra utilizada na definição da equação de regressão. Além disso, percebe-se que quanto mais distante x' estiver da média mais amplo será o intervalo. O limite inferior e superior do intervalo de confiança define a região de confiança em torno da reta de regressão, ou seja, tem-se um nível de confiança, 1 - α, de que a reta teórica, $y = \alpha + \beta x$, estará contida dentro dessa região. A Figura 9.11 ilustra a região de confiança em torno da reta de regressão.

9.5.2 – Intervalos de Confiança para um Valor Previsto pela RLS

Também é interessante estimar um intervalo com nível de confiança 1 - α, no qual estará contido um valor previsto de *y*, calculado para um certo valor especificado de *x'*. Os intervalos de confiança para um valor da variável dependente a ser previsto, $\hat{y}'$,utilizando um valor x' , são estimados por:

$$\hat{y}' - t_{1-\frac{\alpha}{2},n-2}\, s_e \sqrt{1+\frac{1}{n}+\frac{(x'-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}} \le \hat{y}' \le \hat{y}' + t_{1-\frac{\alpha}{2},n-2}\, s_e \sqrt{1+\frac{1}{n}+\frac{(x'-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}} \quad (9.43)$$

onde $\hat{y}' = a + bx'$, $t_{1-\frac{\alpha}{2},n-2}$ é valor do *t* de Student para (1- α/2) e $(n-2)$ graus; e s_e é calculado pela equação 9.34.

Variando *x'* na equação 9.43 obtêm-se a região de previsão para *y'*. Comparando as equações 9.42 e 9.43 verifica-se que o intervalo de confiança para um valor previsto é mais amplo que o estimado para a reta de regressão, como pode ser visualizado na Figura 9.11.

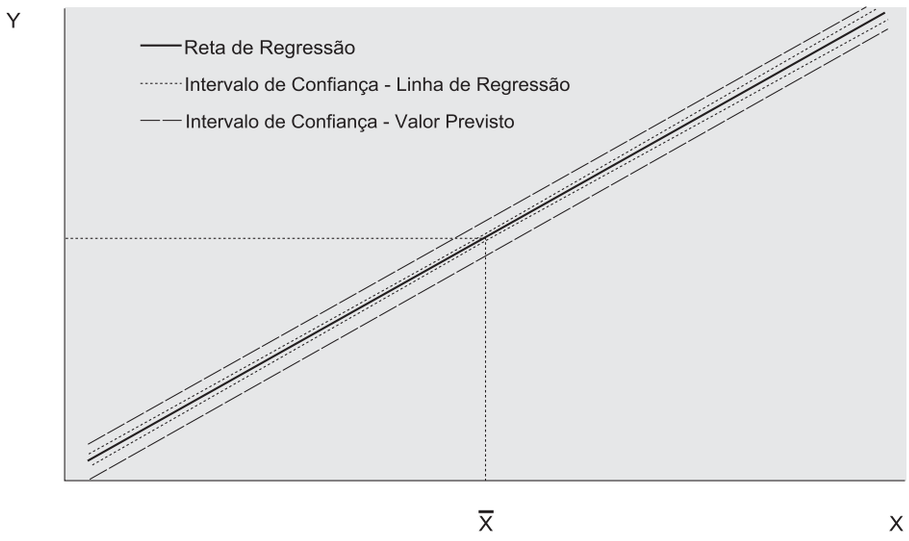


Figura 9.11 – Intervalos e Confiança

9.6 – Avaliação da Regressão Linear Simples

A análise de regressão é uma das técnicas mais úteis na hidrologia, mas exige certo cuidado na sua aplicação. Inicialmente devem ser verificadas as hipóteses da regressão, ou seja, avaliar a linearidade entre as variáveis X e Y, a independência dos resíduos e se estes seguem uma distribuição normal com média zero e variância constante σ_e^2.

A linearidade pode ser avaliada por meio do gráfico de dispersão entre as variáveis X e Y e pelo exame do valor da estimativa do coeficiente de correlação de Pearson. A existência de relação linear entre as variáveis X e Y também pode ser avaliada a partir de um teste de hipótese sobre o coeficiente angular β da equação 9.15. As hipóteses nula e alternativa podem ser expressas da seguinte forma:

$H_0 : \beta = 0$ (não existe relação linear)
$H_0 : \beta \neq 0$ (existe relação linear)

A estatística do teste, t, é igual a diferença entre a inclinação estimada a partir dos dados amostrais, b, e a inclinação da população, β, dividida pelo erro padrão da inclinação, s_b, calculado pela equação 9.38, ou seja,

$$t = \frac{b - \beta}{s_b} \tag{9.44}$$

No caso da plausibilidade da hipótese nula, $H_0 : \beta = 0$, obtém-se

$$t = \frac{b}{s_b} \tag{9.45}$$

A hipótese nula, H_0, é rejeitada se $|t| > t_{1-\alpha/2,n-2}$, onde $t_{1-\frac{\alpha}{2},n-2}$ é valor do t de Student para um nível de significância α (teste bilateral) e $(n-2)$ graus de liberdade.

Outra maneira de se avaliar a existência de uma relação linear entre as variáveis é realizada a partir do intervalo de confiança do parâmetro β, cuja estimativa foi detalhada no item 9.5. O teste consiste em verificar se o valor zero está contido dentro do intervalo de confiança de β. Se o valor zero estiver contido dentro do intervalo de confiança, não existe relação linear entre as variáveis.

A independência dos resíduos pode ser verificada com gráficos dos resíduos em relação à variável prevista, Y. A Figura 9.12 ilustra duas situações: uma onde se

verifica a independência dos resíduos e a outra na qual se observa a ocorrência de dependência.

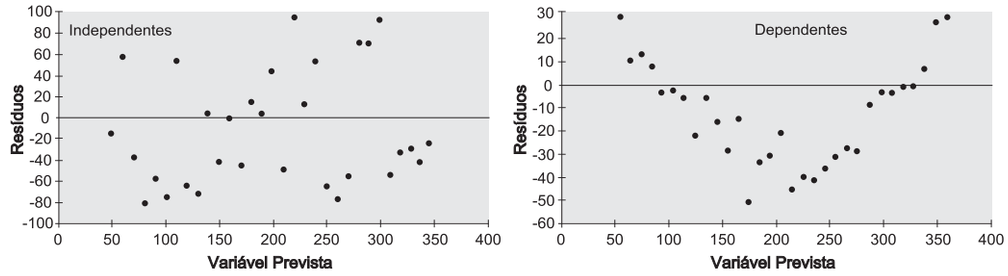


Figura 9.12 – Verificação da independência

Os métodos de análise de freqüência, descritos no capitulo 8, assim como a elaboração de gráficos de probabilidade Normal dos resíduos possibilitam a verificação da hipótese de normalidade. Contudo, para amostras pequenas, as definições sobre a normalidade dos resíduos geralmente não são conclusivas.

No caso da homoscedasticidade, a hipótese de média nula para os resíduos é garantida por construção. Entretanto, a hipótese de variância constante, σ_e^2, deve ser verificada por meio de análise gráfica entre os resíduos e a variável dependente X. A Figura 9.13 apresenta situações de verificação e violação de variância constante.

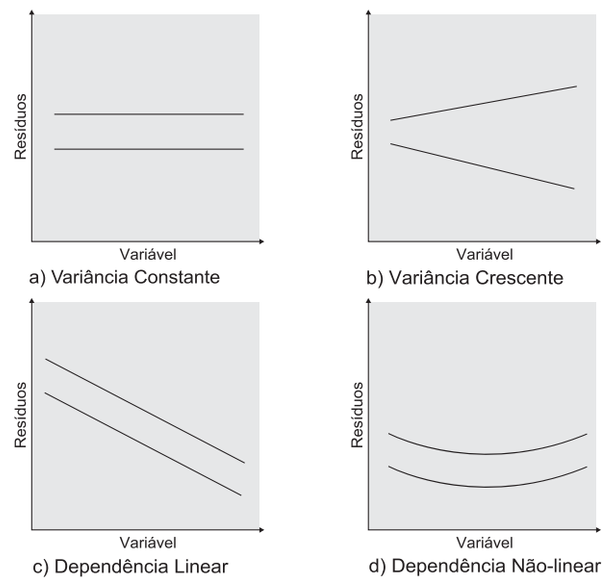


Figura 9.13 – Verificação da variância dos resíduos

Uma medida da qualidade da regressão pode ser obtida pela comparação do erro padrão da estimativa, s_e, com o desvio padrão da variável dependente Y, s_Y. Ambos, s_Y e s_e, apresentam as mesmas unidades e são, portanto, diretamente comparáveis, embora s_e tenha apenas n - 2 graus de liberdade e s_Y tenha n - 1. Caso a equação de regressão se ajuste bem aos dados amostrais, o erro padrão da estimativa se aproxima de zero. Entretanto, se o erro padrão da estimativa tiver valor próximo do desvio padrão de Y, o ajuste entre os dados amostrais e a equação de regressão será muito ruim. Assim, o erro padrão da estimativa deve ser comparado em seus extremos, a saber, zero e s_Y. Além disso, deve ser avaliado o coeficiente de determinação r^2, que expressa a proporção da variância total da variável dependente Y que é explicada pela equação de regressão.

Outro aspecto importante no uso de modelos de regressão é a sua extrapolação. De uma forma geral, não é recomendada a extrapolação da equação de regressão para além dos limites dos dados amostrais utilizados na estimativa dos parâmetros do modelo de regressão linear. O desestímulo à extrapolação apresenta basicamente dois motivos. O primeiro está associado ao fato do intervalo de confiança sobre a linha de regressão alargar, à medida que os valores da variável independente X se afastam da média, como pode ser visto na Figura 9.11. A outra razão é que a relação entre as variáveis X e Y pode não ser linear para valores que extrapolam os dados utilizados na regressão, como ilustrado na Figura 9.14.

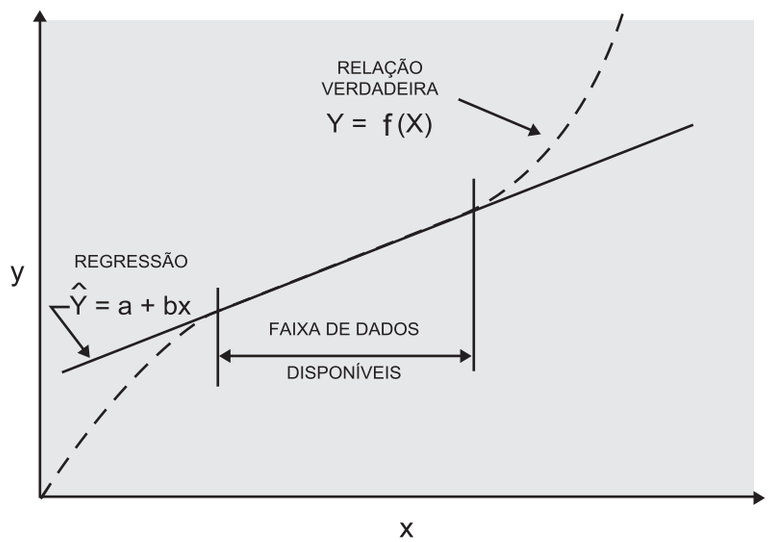


Figura 9.14 – Extrapolação do modelo de regressão

9.7 – Regressão Não-Linear com Funções Linearizáveis

Algumas funções podem ser linearizadas mediante o uso de transformações adequadas permitindo a aplicação da regressão linear simples. Um exemplo pode ser a função potencial a seguir:

$$y = ax^b \tag{9.46}$$

Realizando a anamorfose logarítmica dessa função, obtém-se:

$$ln\, y = ln\left(ax^b\right) \tag{9.47}$$

$$ln\, y = ln\, a + ln\left(x^b\right) \tag{9.48}$$

$$ln\, y = ln\, a + b\, ln\, x \tag{9.49}$$

Alterando as variáveis de forma que $z = ln\, y$,$k = ln\, a$ e $v = ln\, x$, a equação 9.49 se transforma na equação da reta:

$$z = k + bv \tag{9.50}$$

Trabalhando com as variáveis transformadas $z = ln\, y$ e $v = ln\, x$, é possível estimar os parâmetros *k* e *b* com as equações 9.24 e 9.25, respectivamente. Calculando o antilogaritmo de *k* estima-se o parâmetro *a* da equação 9.46.

De forma análoga, a função $y = ab^x$ pode ser resolvida utilizando as variáveis *x* e a transformada $ln\, y$. Existem muitas outras funções linearizáveis, como por exemplo, $y = \left(a + b.x\right)^{-2}$, que estão listadas no Anexo 10. Porém, como o processo de linearização pode envolver a transformação da variável dependente *Y*, em alguns casos as hipóteses da regressão podem não ser atendidas, após a modificação, prejudicando a aplicação dos testes estatísticos descritos anteriormente.

Exemplo 9.1 – Na Tabela 9.1 estão apresentados os valores médios de vazões máximas anuais e as respectivas áreas de drenagem de 22 estações fluviométricas que compõem uma região homogênea de um estudo de regionalização de vazões máximas da bacia do alto São Francisco no qual foi aplicado o método *index-flood*, ou cheia-índice, a ser descrito no capítulo 10. Nesse estudo as médias das vazões máximas anuais foram utilizadas como fator de adimensionalização das séries. Estabelecer uma regressão entre as médias das vazões máximas anuais e as áreas de drenagem, de

forma a permitir a estimativa da cheia-índice (ou *index-flood*) em locais que não possuam estações fluviométricas.

Tabela 9.1 – Área de drenagem e médias das vazões máximas anuais

Est.	1	2	3	4	5	6	7	8	9	10	11
Área (Km²)	269,1	481,3	1195,8	1055,0	1801,7	1725,7	1930,5	2000,2	1558,0	2504,1	5426,3
Q (m³/s)	31,2	49,7	100,2	109,7	154,3	172,8	199,1	202,2	207,2	263,8	483,8
ln A	5,59508	6,17649	7,08657	6,96130	7,49649	7,45339	7,56553	7,60100	7,35116	7,82568	8,59901
ln Q	3,44074	3,90560	4,60707	4,69784	5,03857	5,15190	5,29376	5,30906	5,33364	5,57500	6,18161
Est.	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**
Área (Km²)	7378,3	9939,4	8734,0	8085,6	8986,9	11302,2	10711,6	13881,8	14180,1	16721,9	26553,0
Q (m³/s)	539,4	671,4	690,1	694,0	742,8	753,5	823,3	889,4	1032,4	1336,9	1964,8
ln A	8,90630	9,20426	9,07498	8,99784	9,10352	9,33275	9,27908	9,53833	9,55959	9,72447	10,18690
ln Q	6,29038	6,50941	6,53685	6,54241	6,61043	6,62469	6,71336	6,79050	6,93964	7,19810	7,58312

Solução: Inicialmente é elaborado um diagrama de dispersão, conforme está apresentado na Figura 9.15.

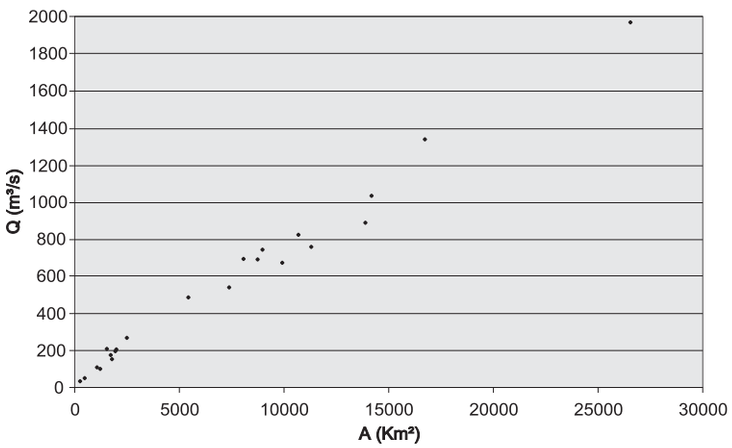


Figura 9.15 – Diagrama de dispersão

Analisando esse diagrama, percebe-se que a relação entre as variáveis área de drenagem e média da vazão máxima anual pode ser expressa por uma função potencial como a equação 9.46, ou seja,

$$Q = kA^b \tag{9.51}$$

Os parâmetros *k* e *b* podem ser estimados por meio da regressão linear simples, após a linearização da equação 9.51. A linearização é realizada

por anamorfose logarítmica como apresentado a seguir:

$$ln\,Q = ln\,k + b\,ln\,A \qquad (9.52)$$

Assim, para concretização da regressão linear simples é necessário calcular os logaritmos da área de drenagem e das médias das vazões máximas anuais, como apresentado na Tabela 9.1. A linearidade entre as variáveis, em coordenadas logarítmicas, pode ser visualizada na Figura 9.16.

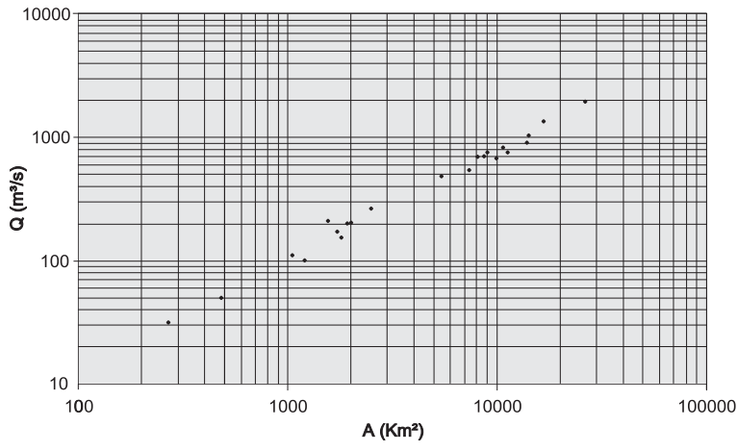


Figura 9.16 – Linearidade entre as variáveis

Utilizando as equações 9.24 e 9.25 e os logaritmos da Tabela 9.1, calcula-se os parâmetros da equação 9.52, b = 0,8751 e $a = ln(k)$ = - 1,4062 . A equação 9.52 é reescrita da seguinte forma:

$$ln\,Q = -1{,}4062 + 0{,}8751.ln\,A \qquad (9.53)$$

A equação 9.53 permite a estimativa de $ln\,Q$ em função do logaritmo da área de drenagem. O ajuste entre os logaritmos das médias das vazões máximas anuais e a reta de regressão da equação 9.53 está apresentado na Figura 9.17

As diferenças ou os resíduos entre os valores observados e os calculados pela reta de regressão estão na Tabela 9.2.

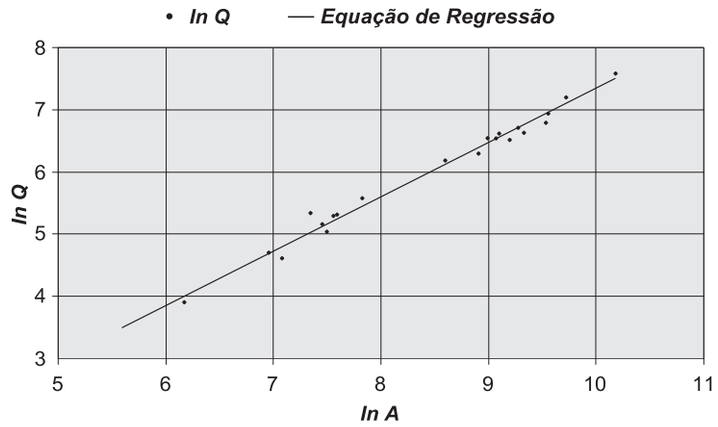


Figura 9.17 – Ajuste entre as observações e a reta de regressão

Tabela 9.2 – Resíduos

Est.	1	2	3	4	5	6	7	8	9	10	11
ln Q	3,4407	3,9056	4,6071	4,6978	5,0386	5,1519	5,2938	5,3091	5,3336	5,5750	6,1816
Previsto	3,4900	3,9988	4,7952	4,6856	5,1540	5,1162	5,2144	5,2454	5,0268	5,4420	6,1188
Res.	-0,0493	-0,0932	-0,1882	0,0122	-0,1154	0,0357	0,0794	0,0636	0,3069	0,1330	0,0628
Est.	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**
ln Q	6,2904	6,5094	6,5369	6,5424	6,6104	6,6247	6,7134	6,7905	6,9396	7,1981	7,5831
Previsto	6,3877	6,6484	6,5353	6,4678	6,5603	6,7609	6,7139	6,9408	6,9594	7,1037	7,5083
Res.	-0,0973	-0,1390	0,0016	0,0746	0,0502	-0,1362	-0,0005	-0,1503	-0,0197	0,0944	0,0748

Os valores observados e os calculados com a equação de regressão permitem a estimativa dos termos da equação 9.27, ou seja, os somatórios dos quadrados total, dos resíduos e os devidos à regressão. Os valores desses somatórios estão apresentados na Tabela 9.3.

Tabela 9.3 – Somatórios dos Quadrados

	Graus de Liberdade	Somatórios dos Quadrados
Regressão	1	24,7726
Resíduo	20	0,2803
Total	21	25,0529

O coeficiente de determinação r^2 é calculado através da equação 9.29.

$$r^2 = \frac{SQ\,Reg}{SQT} = \frac{24{,}7726}{25{,}0529} = 0{,}989 \qquad (9.54)$$

O coeficiente de correlação, r, é igual a 0,994.

Após o cálculo dos parâmetros e dos resíduos é possível verificar as hipóteses da regressão. A seguir é verificada a hipótese de homoscedasticidade dos resíduos. Avaliando a Figura 9.18 observa-se que os resíduos parecem ser independentes e que a variância pode ser considerada aproximadamente constante.

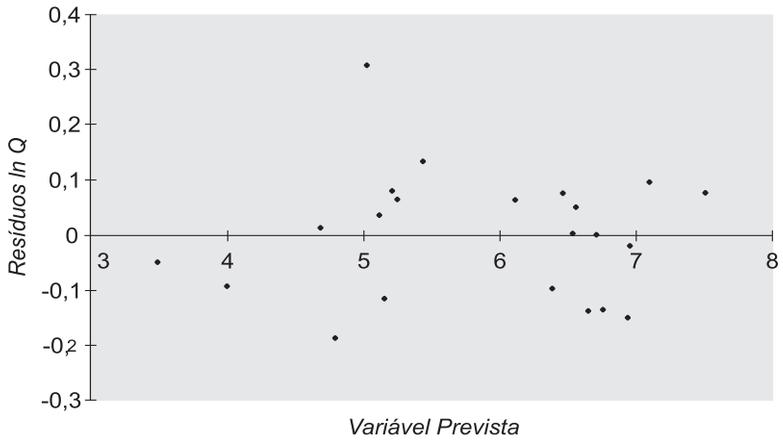


Figura 9.18 – Resíduos

Como o somatório dos resíduos é igual a zero, a sua média também é igual a zero. A raiz quadrada da variância dos resíduos ou o erro padrão da estimativa é calculado pela equação 9.34.

$$\hat{\sigma}_e = s_e = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2}} = \sqrt{\frac{SQ\,Res}{n-2}} = \sqrt{\frac{0{,}2803}{20}} = 0{,}1184 \qquad (9.55)$$

A Figura 9.19 apresenta o ajuste entre os resíduos e uma distribuição normal de média zero e desvio padrão igual a 0,1184.

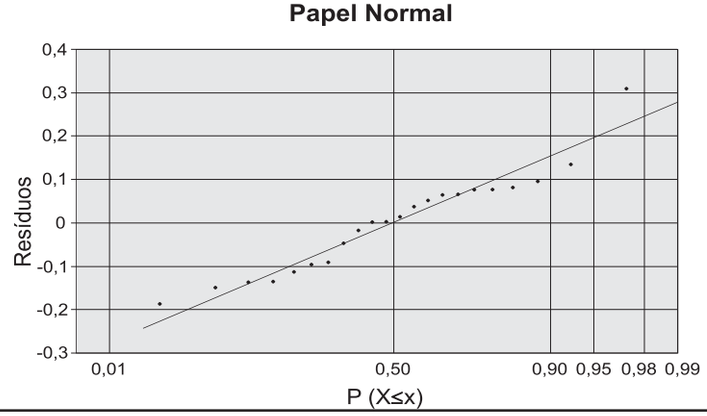


Figura 9.19 – Ajuste dos resíduos à distribuição normal

Os intervalos de confiança para os coeficientes α e β da reta de regressão são estimados com as equações 9.35 e 9.36. Adotando um nível de significância de 5% obtém-se:

$$-1{,}77045 \le \alpha \le -0{,}04196 \text{ e } 0{,}83168 \le \beta \le 0{,}91851$$

No calculo dos limites desses intervalos foram utilizadas os seguintes valores:

$$t_{1-\frac{\alpha}{2},n-2} = t_{0{,}975,21} = 2{,}086$$

$$s_a = \sqrt{s_e^2\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)} = 0{,}1746 \text{ e } s_b = \sqrt{\frac{s_e^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}} = 0{,}0208$$

A relação linear entre as variáveis *ln Q* e *ln A* também pode ser avaliada através de um teste de hipótese com o coeficiente angular da reta de regressão, como descrito no item 9.5. Neste exemplo, a estatística do teste é dada por:

$$t = \frac{b - \beta}{s_b} = \frac{0{,}8751 - 0}{0{,}0208} = 42{,}072 \tag{9.56}$$

Como $|t| > t_{1-\alpha/2,n-2}$, pois $t_{0{,}975,21} = 2{,}086$, a hipótese nula, $\beta = 0$, é rejeitada a um nível de significância de 5%, ou seja, a relação entre as variáveis pode ser considerada linear com uma confiança de 95%.

As etapas anteriores descreveram a regressão linear simples das variáveis transformadas, entretanto, para estimativa do fator "index-flood" utiliza-se a equação na forma potencial como descrito acima. Assim, o parâmetro *k* da equação 9.51 é definido da seguinte forma:

$$k = exp(a) = exp(-1{,}4062) = 0{,}2451 \tag{9.57}$$

A equação 9.51 é reescrita como:

$$Q = kA^b = 0{,}2451A^{0{,}8751} \tag{9.58}$$

Finalmente é realizada uma comparação entre os valores observados e os estimados com a equação 9.58 como está apresentado na Tabela 9.4 e Figura 9.20.

Tabela 9.4 – Desvios Percentuais (DP)

n	1	2	3	4	5	6	7	8	9	10	11
Qobs (m³/s)	31,2	49,7	100,2	109,7	154,3	172,8	199,1	202,2	207,2	263,8	483,8
Qcalc (m³/s)	32,8	54,5	120,9	108,4	173,1	166,7	183,9	189,7	152,4	230,9	454,3
DP (%)	5,1	9,8	20,7	-1,2	12,2	-3,5	-7,6	-6,2	-26,4	-12,5	-6,1
n	12	13	14	15	16	17	18	19	20	21	22
Qobs (m³/s)	539,4	671,4	690,1	694,0	742,8	753,5	823,3	889,4	1032,4	1336,9	1964,8
Qcalc (m³/s)	594,5	771,6	689,0	644,1	706,5	863,4	823,8	1033,6	1053,0	1216,4	1823,2
DP (%)	10,2	14,9	-0,2	-7,2	-4,9	14,6	0,1	16,2	2,0	-9,0	-7,2

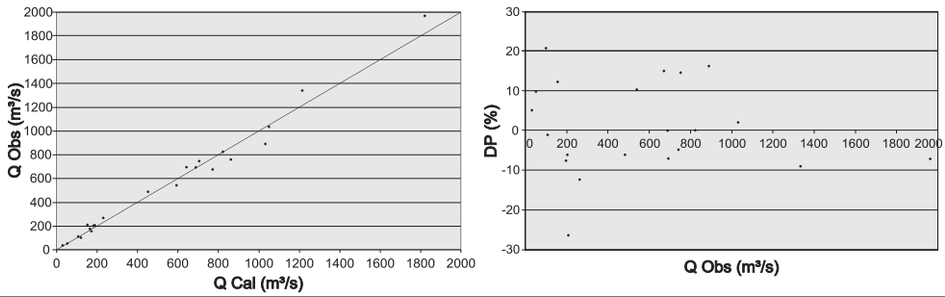


Figura 9.20 – Vazões calculadas versus observadas e desvio percentual

9.8 – Regressão Linear Múltipla

Na regressão múltipla estuda-se o comportamento de uma variável dependente Y em função de duas ou mais variáveis independentes X_i. Se a variável Y variar linearmente com as variáveis X_i, pode-se adotar um modelo geral com a seguinte forma:

$$Y = \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_P X_P \tag{9.59}$$

onde Y é a variável dependente ou prevista; $X_1, X_2, \cdots, X_P$ são as variáveis independentes ou explicativas e $\beta_1, \beta_2, \cdots, \beta_P$ são os coeficientes de regressão.

A partir de um conjunto de *n* valores da variável *Y*, associados às *n* observações correspondentes das *P* variáveis independentes, e utilizando a equação 9.59, pode-se escrever

$$\begin{cases} Y_1 = \beta_1 X_{1,1} + \beta_2 X_{1,2} + \cdots + \beta_P X_{1,P} \\ Y_2 = \beta_1 X_{2,1} + \beta_2 X_{2,2} + \cdots + \beta_P X_{2,P} \\ \vdots \\ Y_n = \beta_1 X_{n,1} + \beta_2 X_{n,2} + \cdots + \beta_P X_{n,P} \end{cases} \tag{9.60}$$

no qual Y_i é o i-ésimo valor da variável dependente e $X_{i,j}$ é a i-ésima observação da j-ésima variável independente. O sistema de equações 9.60 pode ser representado na forma de matriz:

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} X_{1,1} & X_{1,2} & \cdots & X_{1,P} \\ X_{2,1} & X_{2,2} & \cdots & X_{2,P} \\ \vdots & \vdots & \cdots & \vdots \\ X_{n,1} & X_{n,2} & \cdots & X_{n,P} \end{bmatrix} \cdot \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{bmatrix} \tag{9.61}$$

ou em notação matricial,

$$[Y] = [X] \cdot [\beta] \tag{9.62}$$

onde $[Y]$ é um vetor (n x 1) das observações da variável dependente; $[X]$ é uma matriz (n x P) com as n observações de cada uma das P variáveis independentes, e $[\beta]$ é um vetor (P x 1) com os parâmetros desconhecidos. A equação 9.62 terá um termo de intercepto, β_1, se $X_{i,1} = 1$; doravante, no presente texto, adota-se a condição de $X_{i,1} = 1$ para i de 1 até n.

De maneira análoga à regressão linear simples, os coeficientes desconhecidos β_i podem ser estimados pela minimização do somatório dos erros quadráticos, $\sum_{i=1}^{n} e_i^2$, onde,

$$e_i = Y_i - \hat{Y}_i = Y_i - \sum_{i=1}^{P} \hat{\beta}_j X_{i,j} \tag{9.63}$$

Em representação matricial,

$$\sum e_i^2 = [e]^T [e] = \left([Y] - [X\hat{\beta}]\right)^T \left([Y] - [X\hat{\beta}]\right) \tag{9.64}$$

Diferenciando a equação 9.64, em relação a $\hat{\beta}$, e igualando a derivada parcial a zero, obtém-se o sistema

$$[X]^T [Y] = [X]^T [X\hat{\beta}] \tag{9.65}$$

que representa as equações normais de regressão. As soluções da equação 9.65 são encontradas pela multiplicação dois termos da equação por $\left([X]^T[X]\right)^{-1}$.

Desse modo, o vetor $[\hat{\beta}]$ pode ser estimado por:

$$[\hat{\beta}]=([X]^T[X])^{-1}[X]^{-1}[Y] \tag{9.66}$$

De maneira semelhante à regressão simples, o somatório total dos quadrados pode ser apresentado em três parcelas:

$$\sum Y_i^2 = n\overline{Y}^2 + \sum\left(Y_i - \hat{Y}_i\right)^2 + \sum\left(\hat{Y} - \overline{Y}\right)^2 \tag{9.67}$$

ou, em notação matricial, como:

$$[Y]^T[Y] = n\overline{Y}^2 + \left([\hat{\beta}]^T[X]^T[Y] - n.\overline{Y}^2\right) + \left([Y]^T[Y] - [\hat{\beta}]^T[X]^T[Y]\right) \tag{9.68}$$

Freqüentemente, essas parcelas dos somatórios dos quadrados são apresentadas na forma de uma tabela de análise de variância (ANOVA), tal como a ilustrada na Tabela 9.5. O quadrado médio, na Tabela 9.5, resulta da divisão do somatório dos quadrados pelo respectivo número de graus de liberdade.

Tabela 9.5 – Tabela ANOVA da regressão múltipla

Fonte	Graus de liberdade	Somatório dos quadrados	Quadrado médio
Regressão	P	$SQ\,Reg = [\hat{\beta}]^T[X]^T[Y] - n\overline{Y}^2$	$QM\,Reg = \frac{SQ\,Reg}{P}$
Resíduos	$n - P$ -1	$SQ\,Res = [Y]^T[Y] - [\hat{\beta}]^T[X]^T[Y]$	$QM\,Res = \frac{SQ\,Res}{n-P-1}$
Total	n - 1	$SQT = [Y]^T[Y] - n\overline{Y}^2$	

O coeficiente de determinação múltipla R^2 é definido pela seguinte relação:

$$R^2 = \frac{SQ\,Reg}{SQT} = \frac{[\hat{\beta}]^T[X]^T[Y] - n\overline{Y}^2}{[Y]^T[Y] - n\overline{Y}^2} \tag{9.69}$$

O coeficiente de determinação múltipla varia entre 0 a 1 e expressa a proporção da variância que é explicada pelo modelo de regressão. O coeficiente de correlação múltipla é calculado pela extração da raiz quadrada da equação 9.69.

Uma estimativa não enviesada da variância dos erros, $Var(\varepsilon)$ ou σ_e^2, é dada por s_e^2 que é calculada pelo quadrado médio dos resíduos, conforme está apresentado a seguir.

$$s_e^2 = QM\ Res = \frac{SQ\ Res}{n-P-1} = \frac{[Y]^T[Y] - \left[\hat{\beta}\right]^T[X]^T[Y]}{n-P-1} \tag{9.70}$$

O erro padrão da equação de regressão linear múltipla, σ_e, é estimado por s_e, o qual é calculado pela raiz quadrada da equação 9.70.

9.8.1 – Teste da Significância da Equação de Regressão Linear Múltipla

A existência de uma relação significativa entre a variável dependente e as variáveis independentes ou explicativas, pode ser avaliada pelo seguinte teste de hipóteses:

$H_0 : \beta_1 = \beta_2 = \cdots = \beta_n = 0$ (a relação entre as variáveis não é linear)
H_1: pelo menos um $\beta_i \neq 0$

Esse teste é conhecido como 'teste do *F* total', o qual é utilizado para testar a razão entre duas variâncias e, assim, pode ser empregado para verificar a hipótese nula. A estatística do teste é a relação entre a variância decorrente da regressão linear múltipla e variância dos resíduos, ou seja,

$$F = \frac{QM\ Reg}{QM\ Res} \tag{9.71}$$

Os quadrados médios da regressão e dos resíduos ($QM\ Reg$ e $QM\ Res$) podem ser calculados pelas equações apresentadas na Tabela 9.5. A hipótese nula será aceita se

$$F < F(\alpha, P, n-p-1) \tag{9.72}$$

onde α é o nível de significância, P e $n-P-1$ são os graus de liberdade da distribuição *F* de Snedecor, sendo que *P* é o número de variáveis independentes.

9.8.2 – Teste de Partes de um Modelo de Regressão Linear Múltipla

A contribuição de uma variável explicativa ao modelo de regressão múltipla pode ser determinada pelo critério do chamado 'teste do *F* parcial'. De acordo com esse critério, avalia-se a contribuição de uma variável explicativa para a soma dos quadrados devido a regressão, depois que todas as outras variáveis independentes foram incluídas no modelo. Sendo assim, a contribuição de uma variável X_k para a soma dos quadrados da regressão, $SQ\ Reg(X_k)$, considerando que as outras

variáveis estão incluídas, é estimada pela diferença dada por

$SQ\,Reg(X_k)$ = $SQ\,Reg$ (todas as variáveis com X_k) - $SQ\,Reg$ (todas as variáveis sem X_k) (9.73)

A verificação se a inclusão de uma variável X_k melhora significativamente o modelo de regressão é realizada por meio de um teste com as seguintes hipóteses nula e alternativa:

H_0 : a variável X_k não melhora significativamente o modelo
H_1 : a variável X_k melhora significativamente o modelo

A estatística do teste é dada pela relação entre a contribuição da variável X_k à soma dos quadrados devido a regressão, $SQ\,Reg(X_k)$, calculada pela equação 9.73, e a variância dos resíduos considerando o modelo com todas as variáveis inclusive X_k, que é estimada pelo quadrado médio dos resíduos apresentado na Tabela 9.5. Formalmente,

$$F_P = \frac{SQ\,Reg(X_k)}{QM\,Res} \tag{9.74}$$

A hipótese nula deve ser rejeitada se a estatística F_P for maior que o valor critico da distribuição F de Snedecor, com 1 e $n-P-1$ graus de liberdade, e nível de significância α, onde n é o tamanho da amostra e P é o número de variáveis explicativas incluindo X_k, ou seja, rejeita-se H_0 se

$$F_P > F(\alpha,1,n-p-1) \tag{9.75}$$

9.8.3 – Coeficiente de Determinação Parcial

O coeficiente de determinação múltipla, R^2, avalia a proporção da variância da variável dependente Y que é explicada pelas variáveis independentes X_i. Todavia, também é importante avaliar a contribuição de cada variável explicativa em relação ao modelo de regressão múltipla. A proporção da variância da variável dependente Y que é explicada por uma variável independente X_k, enquanto se mantém constante as outras variáveis explicativas, é estimada pelo coeficiente de regressão parcial $R^2_{Yk(P-k)}$. Para um modelo de regressão múltipla com P variáveis explicativas, o coeficiente de determinação parcial para a k-ésima variável é dado por:

$$R^2_{Yk(P-k)} = \frac{SQ\,Reg(X_k)}{SQT - SQ\,Reg + SQ\,Reg(X_k)} \tag{9.76}$$

onde SQT é a soma dos quadrados total, $SQ\,Re\,g$ é a soma dos quadrados da regressão com todas as variáveis inclusive X_k, ambos calculados pelas fórmulas apresentadas na Tabela 9.5, e $SQ\,Re\,g(X_k)$ é a contribuição da variável X_k para a soma dos quadrados da regressão estimada pela equação 9.73.

9.8.4 – Inferências sobre os Coeficientes da Regressão Linear Múltipla

Nesse item também serão admitidas as hipóteses que os resíduos ou erros e_i são independentes e normalmente distribuídos com média zero e variância σ_e^2. A variância de $\hat{\beta}_i$ é estimada pela seguinte relação:

$$\hat{V}ar(\hat{\beta}_i) = \hat{\sigma}^2_{\hat{\beta}_i} = S^2_{\hat{\beta}_i} = C^{-1}_{ii}\hat{\sigma}^2_e \tag{9.77}$$

onde C^{-1}_{ii} é o i-ésimo elemento da diagonal de $\left[X^T X\right]^{-1}$ e $\hat{\sigma}^2_e$ á estimativa de variância dos erros e_i.

Se o modelo estiver correto, então $\hat{\beta}_i / S_{\hat{\beta}_i}$ é distribuído conforme *t* de Student, com $n - P - 1$ graus de liberdade, onde $s_{\hat{\beta}_i}$ é uma estimativa de $\sigma_{\hat{\beta}_i}$ calculada por:

$$s_{\hat{\beta}_i} = \sqrt{C^{-1}_{ii} s^2_e} \tag{9.78}$$

s^2_e é uma estimativa da variância dos resíduos e_i, tal como calculada pela equação 9.70.

Um teste de hipótese para verificar se $\beta_i = \beta_0$, onde β_0 é um valor constante conhecido, pode ser implementado com as seguintes hipóteses nula e alternativa:

$$H_0 : \beta_i = \beta_0$$
$$H_1 : \beta_i \neq \beta_0$$

Para tais hipóteses, a estatística do teste é calculada pela relação:

$$t = \frac{\hat{\beta}_i - \beta_0}{s_{\hat{\beta}_i}} \tag{9.79}$$

A hipótese nula H_0 deve ser rejeitada se

$$|t| > t_{1-\alpha/2,\,n-P-1} \tag{9.80}$$

onde α é o nível de significância (teste bilateral), *n* é tamanho da amostra e *P* é número de variáveis independentes do modelo.

Um teste para a hipótese nula, $H_0 : \beta_i = 0$, e hipótese alternativa, $H_1 : \beta_i \neq 0$, é equivalente a testar a significância da i-ésima variável independente na explicação da variância da variável dependente. A estatística do teste é calculada pela equação 9.79 considerando $\beta_0 = 0$ e a verificação da hipótese é realizada com a equação 9.80. Caso a hipótese nula seja aceita, $\beta_i = 0$, sendo recomendável que a i-ésima variável explicativa seja retirada do modelo.

Verifica-se facilmente que a estatística do teste *F* parcial, equação 9.74, e a estatística *t*, equação 9.79, apresentam a seguinte relação:

$$F_{1,gl} = t_{gl}^2 \tag{9.81}$$

onde *gl* é são os graus de liberdade.

Os intervalos de confiança para os coeficientes da regressão, β_i , são dados por:

$$\hat{\beta}_i \pm t_{1-\frac{\alpha}{2},n-P-1} s_{\hat{\beta}_i} \tag{9.82}$$

9.8.5 – Intervalos de Confiança da Regressão Linear Múltipla

Os limites de confiança de Y_h, onde $Y_h = [X_h][\hat{\beta}]$, são definidos a partir da variância de $\hat{Y}_h$. Neste caso, $\hat{Y}_h$ é uma estimativa de Y (um escalar), no ponto $[X_h]$ (um vetor 1 x *P*) no espaço *P* dimensional e $[\hat{\beta}]$ é um vetor contendo as estimativas de β . A variância de $\hat{Y}_h$ é calculada por:

$$Var(\hat{Y}_h) = \sigma_e^2 [X_h][X^T X]^{-1} [X_h]^T \tag{9.83}$$

onde σ_e^2 é a estimado por s_e^2 através da equação 9.70.

Os limites de confiança de $\hat{Y}_h$ são estabelecidos por:

$$[X_h][\hat{\beta}] \pm t_{1-\frac{\alpha}{2},n-P-1} \sqrt{Var(\hat{Y}_h)} \tag{9.84}$$

Os intervalos de confiança de um valor individual previsto $\hat{Y}_h$ são estimados pela equação a seguir:

$$[X_h][\hat{\beta}] \pm t_{1-\frac{\alpha}{2},n-P-1} \sqrt{Var_i(\hat{Y}_h)} \tag{9.85}$$

onde $Var_i(\hat{Y}_h)$ é a variância de um valor individual previsto de *Y* calculado com

$[X_h]$, sendo estimada por:

$$\hat{V}ar_i\left(\hat{Y}_h\right) = \hat{\sigma}_e^2\left(1 + [X_h][X^T X]^{-1}[X_h]^T\right) \tag{9.86}$$

9.8.6 – Transformações de um Modelo de Regressão Múltipla

Em alguns casos, a violação do pressuposto de homoscedasticidade dos resíduos pode ser superada, por meio da transformação da variável dependente, das variáveis explicativas ou de ambas. Além disso, a transformação de variáveis pode permitir a linearização de uma relação não linear. De uma forma geral, a modificação das variáveis para alcançar os critérios de homoscedasticidade não é uma tarefa fácil. As transformações mais utilizadas são a de raiz quadrada, a logarítmica e a recíproca, conforme apresentado a seguir:

$$Y = \beta_0 + \beta_1\sqrt{X_1} + \beta_2\sqrt{X_2} + \cdots + \varepsilon \tag{9.87}$$

$$Y = \beta_0 + \beta_1 \ln X_1 + \beta_2 \ln X_2 + \cdots + \varepsilon \tag{9.88}$$

$$Y = \beta_0 + \beta_1 \frac{1}{X_1} + \beta_2 \frac{1}{X_1} + \cdots + \varepsilon \tag{9.89}$$

As transformações de modelos não lineares podem ser obtidas por meio de anamorfose logarítmica, tal como exemplificado a seguir.

Modelo multiplicativo do tipo

$$Y = \beta_0 X_1^{\beta_1} X_2^{\beta_2} \varepsilon \tag{9.90}$$

Após a transformação obtêm-se:

$$\ln Y = \ln \beta_0 + \beta_1 \ln X_1 + \beta_2 \ln X_2 + \ln \varepsilon \tag{9.91}$$

No caso de um modelo exponencial

$$Y = e^{(\beta_0 + \beta_1 X_1 + \beta_2 X_2)}\varepsilon \tag{9.92}$$

A transformação logarítmica resulta em:

$$\ln Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ln \varepsilon \tag{9.93}$$

9.8.7 – Comentários Sobre a Regressão Múltipla

Em situações onde as variáveis explicativas são fortemente correlacionadas podem ocorrer problemas na regressão múltipla. Variáveis colineares não fornecem novas informações, dificultando a interpretação dos coeficientes obtidos na regressão, pois em alguns casos o sinal do coeficiente de regressão pode ser o oposto do esperado. Por isso é fortemente recomendável a montagem de uma matriz de coeficientes de correlação simples entre as variáveis explicativas para verificar a existência de uma possível colinearidade entre essas variáveis. Um modo expedito de evitar a colinearidade é a eliminação de uma, entre cada conjunto de duas variáveis explicativas que apresentarem coeficientes de correlação superiores a 0,85. Desse modo, espera-se que as variáveis mantidas no modelo de regressão contribuam significativamente para explicar a variabilidade de *Y*.

O número de observações disponíveis para a análise de regressão deve ser no mínimo 3 a 4 vezes maior que o número de coeficientes da equação regressão que serão estimados. Esta regra procura evitar um falso ajuste causado pelas oscilações que podem ocorrer nas variáveis independentes e que são de difícil detecção nas amostras muito pequenas.

Existem alguns procedimentos que facilitam a elaboração dos modelos de regressão múltipla, do ponto de vista da seleção de variáveis explicativas. Dentre os vários métodos podem ser destacado o de todas as equações possíveis e o da regressão passo a passo.

As diferentes combinações das variáveis independentes permitem a construção de vários modelos de regressão. Caso as equações de regressão tenham um intercepto, β_1, podem ser definidos 2^{P-1} modelos, onde P é o número de variáveis independentes. A definição pelo melhor modelo está associada à análise de cada um separadamente.

A regressão passo a passo consiste na incorporação ao modelo de uma variável, a cada vez, com o objetivo de explicar a maior parte da variância que ainda não foi explicada pelo modelo. Esse método inicia-se com a variável independente que apresenta o maior coeficiente de correlação simples com a variável dependente. Em seguida, é acrescentada uma variável independente à equação, a cada passo, com a avaliação da significância do modelo elaborado e de suas variáveis explicativas, por meio do teste do *F* parcial. Se a contribuição de uma das variáveis explicativas não for considerada significativa, ela é retirada do modelo.

A definição sobre qual a melhor equação de regressão a ser adotada envolve

certa subjetividade. Entretanto, a avaliação da equação de regressão pode ser realizada objetivamente a partir das considerações descritas a seguir. O erro padrão da estimativa deve ser inferior ao desvio padrão da variável independente, $0 \leq S_e \leq S_Y$, pelos mesmos motivos apontados para a regressão linear simples. O coeficiente de determinação deve se aproximar de 1, pois quanto maior o valor desse coeficiente, maior será a proporção da variância explicada pelo modelo. Os testes *F* total, *F* parcial e o teste *t* dos coeficientes da regressão devem ser aplicados para avaliar a significância de cada preditor e do modelo. O sinal do coeficiente de correlação entre uma variável explicativa (X_i) e a variável dependente (Y) deve ser o mesmo do coeficiente da regressão associado a essa variável independente. Os resíduos devem ser examinados através de gráficos com as variáveis independentes e dependentes, para identificar deficiências na equação de regressão e conferir as hipóteses da regressão. E finalmente, comparar os valores previstos com a equação de regressão e dados observados.

Uma maneira de se avaliar os resultados da equação de regressão é verificar a capacidade do modelo prever a variável dependente a partir de observações das variáveis explicativas que não foram utilizadas na estimativa dos coeficientes da regressão. Obviamente, para se fazer essa avaliação é necessário que os dados observados sejam separados aleatoriamente em dois grupos, um para estimar os coeficientes da regressão e o outro para verificar o modelo. Entretanto, na maioria dos casos, o número reduzido de observações não permite esse procedimento.

Exemplo 9.2 – Em um estudo de regionalização de vazões mínimas com 7 dias de duração na bacia do rio Paraopeba, no qual foi aplicado o método *index-flood*, definiu-se uma região homogênea com 15 estações fluviométricas. Nesse estudo as médias das vazões mínimas anuais com 7 dias de duração foram utilizadas como fator de adimensionalização das séries. Defina um modelo de regressão que permita a estimativa do fator *index-flood* em locais que não possuam estações fluviométricas utilizando como prováveis variáveis explicativas as apresentadas na Tabela 9.6.

Tabela 9.6 – Vazões mínimas, área de drenagem, declividade e densidade de drenagem

Estação	1	2	3	4	5	6	7	8
Qmin méd (m³/s)	2,6	1,49	1,43	3,44	1,37	2,53	15,12	16,21
Área (Km²)	461	291	244	579	293	486	2465	2760
I equiv (m/km)	2,69	3,94	7,20	3,18	2,44	1,25	1,81	1,59
DD (Junções/Km²)	0,098	0,079	0,119	0,102	0,123	0,136	0,121	0,137
Estação	**9**	**10**	**11**	**12**	**13**	**14**	**15**	
Qmin méd (m³/s)	21,16	30,26	28,53	1,33	0,43	39,12	45	
Área (Km²)	3939	5414	5680	273	84	8734	10192	
I equiv (m/km)	1,21	1,08	1,00	4,52	10,27	0,66	0,60	
DD (Junções/Km²)	0,134	0,018	0,141	0,064	0,131	0,143	0,133	

Solução: Inicialmente avalia-se a existência de colinearidade entre as variáveis explicativas através da matriz de correlações como apresentado a seguir.

Tabela 9.7 – Matriz de correlações

	Qmin méd (m³/s)	Área (Km²)	I equiv (m/km)	DD (Junções/Km²)
Qmin méd (m³/s)	1			
Área (Km²)	0,992	1		
I equiv (m/km)	-0,625	-0,594	1	
DD (Junções/Km²)	0,141	0,186	-0,049	1

Analisando a Tabela 9.7 observa-se que não existe colinearidade entre as variáveis independentes e que aparentemente as médias das vazões mínimas com 7 dias de duração apresentam uma forte relação linear com a área de drenagem. Assim, para verificar a linearidade entre as variáveis e a possível ocorrência de correlações espúrias foram elaborados os diagramas de dispersão da Figura 9.21.

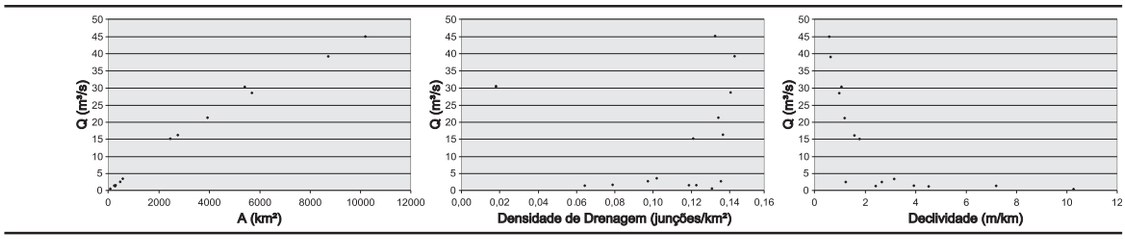


Figura 9.21 – Diagramas de dispersão

Os resultados da Tabela 9.7 e os gráficos da Figura 9.21 indicam que no modelo de regressão a ser adotado terá obrigatoriamente como uma das variáveis explicativas a área de drenagem. Sendo assim, o problema se restringe a avaliar se a inclusão de novas variáveis trará melhora significativa aos resultados do modelo. O modelo de regressão adotado será do tipo multiplicativo como apresentado a seguir:

$$Q = \beta_0 A^{\beta_1} X_2^{\beta_2} X_3^{\beta_3} \tag{9.94}$$

Após a transformação logarítmica obtêm-se:

$$ln\,Q = ln\,\beta_0 + \beta_1\, ln\,A + \beta_2\, ln\,X_2 + \beta_3\, ln\,X_3 \tag{9.95}$$

Assim, para calcular os parâmetros da equação 9.95 é necessário calcular os logaritmos das variáveis independentes e dependentes conforme está apresentado na Tabela 9.8

Tabela 9.8 – Logaritmos das variáveis

Estação	1	2	3	4	5	6	7	8
Qmin méd (m³/s)	0,9555	0,3988	0,3577	1,2355	0,3148	0,9282	2,7160	2,7856
Área (Km²)	6,1343	5,6737	5,4972	6,3604	5,6812	6,1870	7,8100	7,9230
I equiv (m/km)	0,9895	1,3712	1,9741	1,1569	0,8920	0,2231	0,5933	0,4637
DD (Junções/Km2)	-2,3276	-2,5382	-2,1299	-2,2829	-2,0977	-1,9974	-2,1095	-1,9908
Estação	9	10	11	12	13	14	15	
Qmin méd (m³/s)	3,0521	3,4098	3,3510	0,2852	-0,8440	3,6666	3,8067	
Área (Km²)	8,2787	8,5968	8,6448	5,6095	4,4296	9,0750	9,2293	
I equiv (m/km)	0,1906	0,0770	0,0000	1,5085	2,3292	-0,4155	-0,5108	
DD (Junções/Km²)	-2,0077	-4,0118	-1,9614	-2,7423	-2,0317	-1,9465	-2,0207	

A definição sobre quais serão as variáveis explicativas que comporão o modelo de estimativa das vazões mínimas é realizada através da análise das equações de regressão que contenham as seguintes variáveis independentes: somente a área de drenagem (QA); a área de drenagem e a declividade (QAI); a área de drenagem e densidade de drenagem (QADD); e área de drenagem, a declividade e a densidade de drenagem (QAIDD). A avaliação da inclusão de uma nova variável ao modelo QA é realizada através do teste da significância da equação de regressão linear múltipla e do teste de partes de um modelo de regressão linear múltipla.

Inicialmente analisa-se o modelo que utiliza somente a área de drenagem como variável independente, ou seja,

$$Q = \beta_0 A^{\beta_1} \tag{9.96}$$

$$ln\,Q = ln\,\beta_0 + \beta_1\, ln\,A \tag{9.97}$$

A Tabela 9.9 apresenta os somatórios dos quadrados e a estatística *F* do teste de significância da equação de regressão na forma de uma tabela ANOVA.

Tabela 9.9 – ANOVA modelo QA

	gl	*SQ*	*MQ*	*F*
Regressão	1	33,04321	33,04321	2915,798
Resíduo	13	0,147322	0,011332	
Total	14	33,19053		

O modelo QA é considerado significativo, pois a hipótese nula do teste, $\beta_1 = 0$, é rejeitada uma vez que:

$$(F = 2916) > [F(0{,}05;1;13) = 4{,}67] \tag{9.98}$$

Os parâmetros do modelo QA, o coeficiente de determinação e o erro padrão estão na Tabela 9.12. A inclusão da declividade como mais uma variável explicativa no modelo da equação 9.96 resulta em:

$$Q = \beta_0 A^{\beta_1} I^{\beta_2} \tag{9.99}$$

$$ln\,Q = ln\,\beta_0 + \beta_1\, ln\,A + \beta_2\, ln\,I \tag{9.100}$$

Os parâmetros do modelo QAI, o coeficiente de determinação e o erro padrão estão na Tabela 9.12. A estatística *F* do teste de significância da equação de regressão e os somatórios dos quadrados do modelo QAI estão na Tabela 9.10.

Tabela 9.10 – ANOVA modelo QAI

	gl	*SQ*	*MQ*	*F*
Regressão	2	33,07298	16,53649	1688,119
Resíduo	12	0,11755	0,009796	
Total	14	33,19053		

O modelo QAI também é considerado significativo pois a estatística do teste é maior que o valor de referência para um nível de significância de 5%, ou seja, $(F = 1688) > [F(0{,}05;2;12) = 3{,}89]$. A contribuição da variável declividade para a soma dos quadrados da regressão, $SQ\,Reg(X_I)$, considerando que a variável área de drenagem já está incluída, é estimada pela equação 9.73.

$$SQ\,Reg(X_I) = 33{,}07 - 33{,}04 = 0{,}03$$

A estatística do teste de partes de um modelo de regressão linear múltipla é calculada pela equação 9.74. Sendo assim,

$$F_P = \frac{SQ\,Reg(X_I)}{MQ\,Res} = \frac{0{,}03}{0{,}0098} = 3{,}04$$

Como $(F_P = 3{,}04) < [F(0{,}05;1;12) = 4{,}75]$, a inclusão da variável declividade não melhora significativamente o modelo quando se considera um nível de significância de 5%.

Acrescentando a densidade de drenagem como mais uma variável explicativa no modelo da equação 9.96 obtêm-se:

$$Q = \beta_0 A^{\beta_1} DD^{\beta_2} \tag{9.101}$$

$$ln\,Q = ln\,\beta_0 + \beta_1\, ln\, A + \beta_2\, ln\, DD \tag{9.102}$$

Os parâmetros do modelo QADD, o coeficiente de determinação e o erro padrão estão na Tabela 9.12. A estatística *F* do teste de significância da equação de regressão e os somatórios dos quadrados do modelo QADD estão na Tabela 9.11.

Tabela 9.11 – ANOVA modelo QADD

	gl	*SQ*	*MQ*	*F*
Regressão	2	33,04797	16,52399	1390,935
Resíduo	12	0,142557	0,01188	
Total	14	33,19053		

O teste da significância da equação de Regressão Linear Múltipla indicou que o modelo QADD pode ser considerado significativo para um nível de significância de 5%, uma vez que $\left(F = 1390{,}9\right) > \left[F\left(0{,}05;2;12\right) = 3{,}89\right]$.

A contribuição da variável densidade de drenagem para a soma dos quadrados da regressão, $SQ\,Re\,g\left(X_{DD}\right)$, considerando que a variável área de drenagem já está incluída, é estimada pela equação 9.73.

$$SQ\,Re\,g\left(X_{DD}\right) = 33{,}048 - 33{,}043 = 0{,}005$$

A estatística do teste de partes de um modelo de regressão linear múltipla é calculada pela equação 9.74. Sendo assim,

$$F_P = \frac{SQ\,Re\,g\left(X_1\right)}{MQ\,Res} = \frac{0{,}005}{0{,}01188} = 0{,}40$$

A inclusão da variável densidade de drenagem não melhora significativamente o modelo quando se considera um nível de significância de 5%, pois $\left(F_P = 0{,}40\right) < \left[F\left(0{,}05;1;12\right) = 4{,}75\right]$.

Acrescentando a densidade de drenagem como mais uma variável explicativa no modelo da equação 9.99 obtêm-se:

$$Q = \beta_0 . A^{\beta_1} . I^{\beta_2} DD^{\beta_3} \quad (9.103)$$

$$ln\, Q = ln\, \beta_0 + \beta_1\, ln\, A + \beta_2\, ln\, I + \beta_3\, ln\, DD \quad (9.104)$$

Os parâmetros do modelo QAIDD, o coeficiente de determinação e o erro padrão estão na Tabela 9.12. Entretanto, como a inclusão das variáveis declividade e densidade de drenagem mostrou-se não significativa, não é necessário avaliar o modelo a três variáveis explicativas, uma vez que teríamos um modelo significativo, mas com excesso de variáveis explicativas que não contribuem significativamente para a explicação da variância total da vazão mínima com 7 dias de duração.

Tabela 9.12 – Parâmetros dos modelos

Modelo	$ln\ (\beta_0)$	(β_1)	(β_2)	(β_3)	r^2	Erro Padrão
QA	-5,1696	0,9889			0,9956	0,1065
QAI	-5,7309	1,0551	0,1344		0,9965	0,0990
QADD	-5,24512	0,9884	-0,0348		0,9957	0,1090
QAIDD	-5,7579	1,05224	0,12930	- 0,0223	0,9965	0,1025

Analisando os resultados anteriores verifica-se que a inclusão das variáveis declividade e densidade de drenagem não traz ganhos significativos ao modelo de estimativa das vazões mínimas médias com 7 dias de duração. Dessa forma, o melhor modelo é o que adota somente a área de drenagem como variável explicativa, ou seja, a equação 9.97. A partir do comportamento dos resíduos na Figura 9.22 verifica-se que os resíduos são independentes e que a variância pode ser considerada aproximadamente constante. A Figura 9.22 apresenta o ajuste entre os resíduos e uma distribuição normal de média zero e desvio padrão igual a 0,1065.

A análise de regressão foi realizada com dados transformados, sendo assim, é necessário realizar a operação de inversão do parâmetro $ln(\beta_0)$ para definir o modelo na forma da equação 9.96.

$$\beta_0 = exp[ln(\beta_0)] = exp(-5{,}1696) = 0{,}00569$$

$$Q = 0{,}00596 A^{0{,}9889}$$

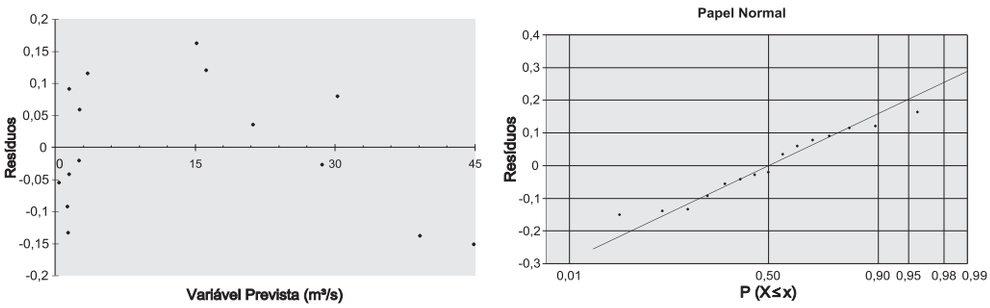


Figura 9.22 – Resíduos

Exercícios

1 – Deduzir a equação 9.28

2 – Mostrar que a correlação entre a variável independente, Y, e a sua estimativa, $\hat{Y}$, é equivalente ao coeficiente de correlação da regressão simples.

3 – A Tabela 9.13 apresenta os valores da área de drenagem e a vazão média de longo termo de 22 estações fluvioméricas da bacia do alto rio São Francisco. Estime a equação de regressão linear considerando a área de drenagem (km²) como a variável independente.
a) Verificar se os desvios atendem a hipótese de homoscedasticidade
b) Calcular o erro padrão e o coeficiente de determinação
c) Plotar os intervalos de confiança de 95% da linha de regressão e do valor previsto.

Tabela 9.13 – Áreas de drenagem e vazões médias de longo termo – Exercício 3

Estação	Área (km²)	Q_{mlt} (m³/s)	Estação	Área (km²)	Q_{mlt} (m³/s)	Estação	Área (km²)	Q_{mlt} (m³/s)
1	83,9	1,32	9	1206,9	19,3	17	5680,4	85,7
2	188,3	2,29	10	1743,5	34,2	18	8734	128
3	279,4	4,24	11	2242,4	40,9	19	10191,5	152
4	481,3	7,34	12	3727,4	65,3	20	13881,8	224
5	675,7	8,17	13	4142,9	75,0	21	14180,1	241
6	769,7	8,49	14	4874,2	77,2	22	29366,18	455
7	875,8	18,9	15	5235	77,5			
8	964,2	18,3	16	5414,2	86,8			

4 – (Adaptado de Haan,1979) Estime a equação de regressão do exercício 3 considerando a vazão média de longo termo como variável independente.
a) O modelo obtido concorda com o estimado no exercício anterior
b) Os modelos deveriam concordar? Por quê?

5 – Utilizando os dados da Tabela 9.13, estime a equação de regressão considerando uma relação potencial entre a vazão média de longo termo e a área de drenagem, ou seja, $Q = kA^C$. Compare os resultados do modelo com os obtidos no exercício 3.

6 – Em muitos casos é mais conveniente utilizar um modelo de regressão do tipo $Y = ax$, ou seja, a reta de regressão passa pela origem e o parâmetro b é igual a zero.
a) Deduza a equação normal para essa situação
b) Calcule a reta de regressão passando pela origem para os dados do exercício 3.

7) Deduzir as equações normais para o seguinte modelo parabólico $Q = a + bH + cH^2$, no qual Q denota as descargas e H os níveis d'água em uma estação fluviométrica.

8) A Tabela 9.14 apresenta uma lista de medições de descargas realizadas em um posto fluviométrico.

Tabela 9.14 – Lista de medições de descargas do exercício 8

H (m)	Q (m³/s)	H (m)	Q (m³/s)	H (m)	Q (m³/s)	H (m)	Q (m³/s)
0,0	20	1,91	170	4,73	990	8,21	2540
0,8	40	2,36	240	4,87	990	8,84	2840
1,19	90	2,70	300	5,84	1260	9,64	3320
1,56	120	4,07	680	7,19	1920	——	——

a) Faça um gráfico dos pontos cota-descarga com H em ordenadas e Q em abcissas.
b) Estime a relação cota-descarga (curva chave), usando os seguintes modelos de regressão:
- $Q = a + bH + cH^2$
- $Q = a\left(H - h_0\right)^n$ onde h_0 representa a cota para a vazão nula.

c) Desenhe no gráfico do item (a) as duas curvas ajustadas. Decida qual é o melhor modelo de regressão a partir da comparação da variância residual, dada pela fórmula $S_{res}^2 = \dfrac{\sum_{i=1}^{n}\left(Q_i^{obs} - Q_i^{est}\right)^2}{n-k-1}$, onde n é o tamanho da amostra, k é o número de variáveis explicativas e os índices *obs* e *est* referem-se aos valores observados e estimados, respectivamente.
d) Uma ponte será construída nesse local, o qual situa-se a cerca de 500 m a jusante de uma barragem. O tabuleiro dessa ponte deverá ter uma altura suficientemente grande para permitir a passagem da descarga de projeto do

vertedor da barragem que é de 5200 m^3/s. Determine a cota altimétrica mínima do tabuleiro da ponte, sabendo que o RN-2, de cota arbitrária 5,673 m em relação ao zero da régua, possui cota altimétrica 731,229 m.

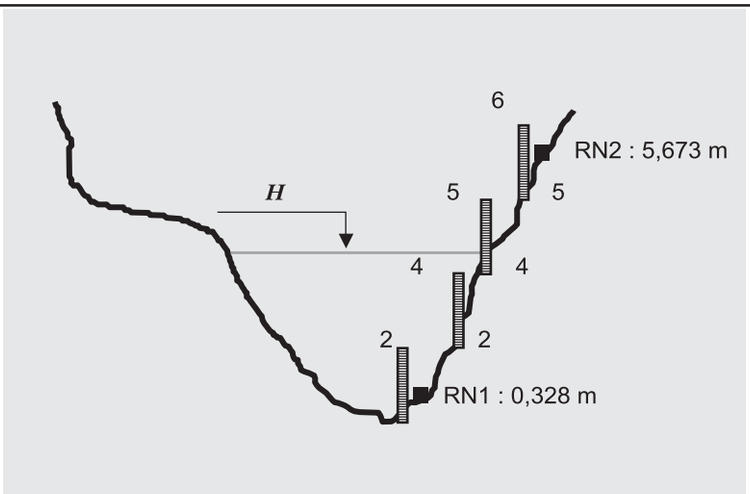


Figura 9.23 – Exercício 8

9 – A curva de dupla massa é muito utilizada em engenharia de recursos hídricos para detectar problemas na consistência de dados pluviométricos. Essa curva permite a comparação gráfica entre os valores acumulados das precipitações anuais (ou mensais) observadas na estação em análise e os valores acumulados das precipitações anuais (ou mensais) regionais, que são estimadas como as médias aritméticas de várias estações vizinhas. A Tabela 9.15 apresenta os totais anuais de uma estação em análise e da média regional. Grafe a precipitação acumulada regional no eixo das abscissas e a precipitação acumulada da estação em análise no eixo das ordenadas.

a) A partir de que ano parece haver uma mudança na inclinação da curva de dupla massa?

b) Calcule as inclinações das retas de regressão considerando dois cenários distintos. O primeiro, com os dados anteriores a aparente mudança de inclinação e o outro utilizando os dados posteriores a essa alteração.

c) Testar a hipótese das inclinações serem significativamente diferentes.

Tabela 9.15 – Dados do exercício 9

Ano	1960	1961	1962	1963	1964	1965	1966	1967	1968	1969	1970
Analisada (mm)	1700	1300	2100	1900	1800	1200	1450	1250	1710	1700	1400
Média Regional (mm)	1067	857	1440	1393	1233	980	1177	1043	1490	1450	1200

10 – Em um estudo de regionalização de vazões máximas, no qual foi aplicado o método *index-flood*, definiu-se uma região homogênea com 13 estações

fluviométricas. Nesse estudo as médias das vazões máximas foram utilizadas como fator de adimensionalização das séries. Defina um modelo de regressão que permita a estimativa do fator *index-flood* em locais que não possuam estações fluviométricas utilizando como possíveis variáveis explicativas as apresentadas na Tabela 9.16. Calcular o erro padrão e plotar os intervalos de confiança de 90% do plano de regressão e do valor previsto.

Tabela 9.16 – Dados do exercício 10

Estações	Q_{max} médio	Área (Km2)	P médio (m)	I equiv (m/km)	L (km)L (km)
1	12,6	83,9	1,436	10,27	18
2	29,8	188,3	1,460	3,1	26,4
3	30,4	244	1,466	7,2	18,3
4	35,5	273	1,531	4,52	40
5	31,5	291,1	1,462	3,94	32,7
6	64,7	461,4	1,400	2,69	52
7	86,9	486,4	1,369	1,25	47,3
8	78,2	578,5	1,464	3,18	41,6
9	74,5	675,2	1,485	2,96	53,8
10	241,6	2465,1	1,409	1,81	88,9
11	437,1	3939,2	1,422	1,21	187,4
12	541,7	5414,2	1,448	1,08	218,2
13	534,2	5680,4	1,449	1	236,33

CAPÍTULO 10

Análise Regional de Freqüência de Variáveis Hidrológicas

CAPÍTULO 10 ANÁLISE REGIONAL DE FREQÜÊNCIA DE VARIÁVEIS HIDROLÓGICAS

A análise local de freqüência de variáveis hidrológicas dispõe de um conjunto de técnicas de inferência estatística e de modelos probabilísticos, as quais têm sido objeto de freqüente investigação, visando principalmente a obtenção de estimativas cada vez mais eficientes e confiáveis. Entretanto, a inexistência de amostras suficientemente longas impõe um limite superior ao grau de sofisticação estatística a ser empregado na análise local de freqüência. Por isso, nesse contexto, o Conselho Nacional de Pesquisas dos Estados Unidos (NRC,1987) propôs a 'substituição do tempo pelo espaço', a qual se dá pela análise regional de freqüência de um conjunto de informações hidrológicas, obtidas em diferentes locais, de modo a compensar as amostras individuais de tamanho relativamente curto. Nesse sentido, a análise regional de freqüência representa uma alternativa que procura compensar a insuficiente caracterização temporal do comportamento de eventos extremos por uma coerente caracterização espacial da variável hidrológica em questão. Em linhas gerais, a análise regional de freqüência utiliza um grande conjunto de dados espacialmente disseminados de certa variável, como por exemplo vazões e precipitações, observados em pontos distintos de uma região considerada homogênea, do ponto de vista estatístico ou dos processos físicos em foco, para estimar os quantis associados a diferentes probabilidades de excedência, para um certo local dentro dessa região. A análise de freqüência regional pode ser usada para aumentar a confiabilidade dos quantis estimados para um ponto já monitorado, bem como para estimar os quantis em locais que não possuem uma coleta sistemática de informações. Em geral, essa última aplicação da análise de freqüência regional é a mais comum.

Os princípios da análise regional de freqüência há muito são conhecidos e empregados em diversas metodologias de uso corrente em hidrologia [ver por exemplo Dalrymple (1960), NERC (1975), Eletrobrás (1985) e Tucci (2002)]. Dentre os vários procedimentos, neste texto serão destacados os (i) métodos que regionalizam os quantis associados a um risco previamente especificado, (ii) métodos que regionalizam os parâmetros das distribuições de probabilidades, e (iii) métodos que regionalizam uma curva de quantis adimensionais, geralmente denominados de métodos da cheia-índice ou métodos *index-flood*. Em particular, o método *index-flood* com momentos-L, sistematizado em Hosking e Wallis (1997), será objeto de detalhe em um dos itens deste capítulo.

A seguir, serão abordados, de início, os procedimentos para identificação de regiões homogêneas, com uma descrição pormenorizada da análise de aglomerados (ou

de *clusters*) e, na seqüência, os principais métodos de análise de freqüência regional.

10.1 – Regiões Homogêneas

Dentre as etapas que compõem a análise regional de freqüência de variáveis aleatórias, a identificação e a delimitação de regiões homogêneas é considerada a mais difícil e mais sujeita a subjetividades. Uma região é homogênea se existem evidências suficientes de que as diferentes amostras do grupo possuem a mesma distribuição de freqüências, a menos, é claro, do fator de escala local. Potter (1987) considera que essa etapa é crucial por exigir do analista, e da metodologia empregada, a capacidade de discernir se observações anômalas, eventualmente existentes em uma ou mais amostras do grupo, devem-se a diferenças populacionais em relação ao modelo probabilístico proposto ou a meras flutuações amostrais. Embora diversas técnicas tenham sido propostas para a identificação e delimitação de regiões homogêneas, nenhuma delas constitui um critério estritamente objetivo ou uma solução consensual para o problema. De fato, Bobée e Rasmussen (1995) reconhecem que, por si, a análise regional de freqüência e, em particular a delimitação de regiões homogêneas, são construídas com base em premissas difíceis de serem tratadas com rigor matemático. Concluem enfatizando que esse fato deve ser visto como um desafio a ser vencido por futuras investigações pertinentes à área de análise de freqüência.

Uma primeira fonte de controvérsias quanto à correta abordagem para a identificação de regiões homogêneas diz respeito ao tipo de dado local a ser utilizado. Faz-se distinção entre *estatísticas locais* e *características locais*. As estatísticas locais referem-se, por exemplo, a estimadores das medidas de dispersão e assimetria, tais como os coeficientes de variação e de assimetria, calculados diretamente a partir dos dados objetos da análise regional de freqüência. Por outro lado, as características locais são, em princípio, quantidades previamente conhecidas e não dedutíveis, ou estimadas, a partir das amostras pontuais. Como exemplos de características locais para o caso de variáveis hidrológicas ou hidrometeorológicas, podem ser citadas a latitude, a longitude, a altitude e outras propriedades relacionadas a um certo local específico. Podem ser incluídas também outras características indiretamente relacionadas à amostra, tais como a altura média de precipitação anual, o mês mais freqüente de ocorrência de cheias ou o volume médio anual do escoamento-base. Alguns autores, nominalmente Wiltshire (1986), Burn (1989) e Pearson (1991), propuseram técnicas que fazem uso somente das estatísticas locais para definir regiões homogêneas de vazões de enchentes na Inglaterra, Estados Unidos e Nova Zelândia, respectivamente.

Contrariamente, Hosking e Wallis (1997) recomendam que a identificação de regiões homogêneas se faça em duas etapas consecutivas: a primeira, consistindo de uma delimitação preliminar baseada unicamente nas *características* locais, e a segunda, consistindo de um teste estatístico, construído com base somente nas *estatísticas* locais, cujo objetivo é o de verificação dos resultados preliminarmente obtidos.

Os diversos métodos e técnicas de agrupamento de locais similares em regiões homogêneas podem ser categorizados como se segue.

• *Conveniência Geográfica*
Dentro dessa categoria, encontram-se todas as experiências de identificação de regiões homogêneas que se baseiam no agrupamento subjetivo e/ou conveniente dos postos de observação, geralmente contíguos, em áreas administrativas ou em zonas previamente definidas segundo limites arbitrários. Dentre os inúmeros trabalhos que fizeram uso da conveniência geográfica, podem ser citados as regionalizações de vazões de enchentes das Ilhas Britânicas (*NERC*, 1975) e da Austrália (Institution of Engineers Australia, 1987).

• *Agrupamento Subjetivo*
Nessa categoria, a delimitação subjetiva das regiões homogêneas é feita por agrupamento dos postos de observação em conformidade à similaridade de algumas características locais, tais como classificação climática, relevo ou conformação das isoietas anuais. Schaefer (1990), por exemplo, utilizou alturas similares de precipitação anual para delimitar regiões homogêneas de chuvas máximas anuais no estado americano de Washington. Da mesma forma, Pinto e Naghettini (1999) combinaram as conformações de relevo, clima e isoietas anuais, para a delimitação preliminar de regiões homogêneas de alturas diárias de chuva máximas anuais na bacia do Alto Rio São Francisco. Embora um grau considerável de subjetividade esteja presente nessas experiências, os seus resultados podem ser objetivamente verificados através do teste estatístico da medida de heterogeneidade, a ser descrito no item 10.3.2.1.

• *Agrupamento Objetivo*
Nesse caso, as regiões são formadas pelo agrupamento dos postos de observação em um ou mais conjuntos de modo que uma dada estatística não exceda um valor limiar previamente selecionado. Esse valor limiar é arbitrado de forma a minimizar critérios variados de heterogeneidade. Por exemplo, Wiltshire (1985) utilizou como critério a razão de verossimilhança e, posteriormente, Wiltshire (1986) e Pearson (1991) empregaram as variabilidades intra-grupos de estatísticas locais, tais como os coeficientes de variação e assimetria. Na seqüência, os grupos são subdivididos

iterativamente até que se satisfaça o critério de homogeneidade proposto. Hosking e Wallis (1997) apontam como uma desvantagem dessa técnica o fato de que as iterações sucessivas de reagrupamento dos postos de observação nem sempre conduzem a uma solução final otimizada. Apontam também para o fato que as estatísticas intra-grupos empregadas podem ser influenciadas, em grau indeterminado, pela eventual existência de dependência estatística entre as amostras consideradas.

• *Análise de Aglomerados ou Análise de Clusters*
Trata-se de um método usual de análise estatística multivariada, no qual associa-se a cada posto um vetor de dados contendo as características e/ou estatísticas locais. Em seguida, os postos são agrupados e reagrupados de forma que seja possível identificar a maior ou menor similaridade entre os seus vetores de dados. Hosking e Wallis (1997) citam diversos estudos (Burn, 1989 e Guttman, 1993, entre outros), nos quais a análise de *clusters* foi empregada com sucesso para a regionalização de freqüências de precipitação, vazões de enchentes e outras variáveis. Esses autores consideram a análise de *clusters* como o método mais prático, porém ainda sujeito a subjetividades, para a identificação preliminar de regiões homogêneas. Por constituir-se em um método preferencial, apresenta-se no item 10.1.1 uma descrição da técnica de análise de *clusters* e recomendações para o seu emprego na identificação preliminar de regiões homogêneas.

• *Outros Métodos*
Além dos mencionados anteriormente, outros métodos têm sido empregados para a identificação e delimitação de regiões homogêneas. No contexto de variáveis hidrológicas/ hidrometeorológicas, podem ser citados os seguintes exemplos: (a) análise de resíduos de regressão (Tasker, 1982), (b) análise de componentes principais (Nathan e McMahon, 1990), (c) análise fatorial (White, 1975), (d) correlação canônica (Cavadias, 1990), (e) análise discriminante (Waylen e Woo, 1984) e (f) análise de formas das funções densidades de probabilidade (Gingras e Adamowski, 1993). Da mesma forma que os anteriores, esses métodos também apresentam elementos subjetivos e limitações.

10.1.1 – Noções sobre Análise de *Clusters*

O termo análise de *clusters* foi empregado pela primeira vez por Tryon (1939) e engloba um grande número de diferentes algoritmos de classificação em grupos, ou taxonomias, estruturalmente similares. Essencialmente, a análise de *clusters* é a aglomeração seqüencial de indivíduos a grupos cada vez maiores, de acordo com algum critério, distância ou medida de dissimilaridade. Um indivíduo pode

ter diversos atributos ou características, as quais são quantificadas e representadas pelo vetor de dados locais $\{Z_1, Z_2, ..., Z_p\}$. As medidas ou distâncias de dissimilaridade entre dois indivíduos devem ser representativas da variação mútua das características locais em um espaço *p*-dimensional. A medida mais usada é a *distância Euclidiana generalizada*, a qual é simplesmente a distância geométrica tomada em um espaço de *p* dimensões. Por exemplo, a distância Euclidiana entre dois indivíduos *i* e *j* é dada por

$$d_{ij} = \sqrt{\sum_{k=1}^{p} \left(Z_{ik} - Z_{jk}\right)^2} \qquad (10.1)$$

Para efeito de entendimento da lógica inerente à análise de *clusters*, tomemos um de seus métodos de aglomeração mais simples que é conhecido como o do *vizinho mais próximo*. A aglomeração em *clusters* inicia-se pelo cálculo das distâncias *d* entre um certo indivíduo e todos os outros do grupo, para cada um deles. Inicialmente, existem tantos grupos quanto numerosos forem os indivíduos. O primeiro *cluster* se forma com o par de indivíduos mais próximos (ou de menor distância Euclidiana); se a distância para outros indivíduos for a mesma da anterior, estes também farão parte do *cluster*. Em seguida, forma-se o *cluster* seguinte com o par (ou grupo, ou *cluster*) de menor distância Euclidiana e assim sucessivamente até que, ao final, todos os indivíduos estejam todos aglomerados. Considere o exemplo hipotético da Figura 10.1, no qual 10 indivíduos, assinalados em abscissas, tiveram suas distâncias Euclidianas calculadas e grafadas em ordenadas, de acordo com certo número de atributos. Se somente dois *clusters* forem considerados, o primeiro seria formado pelo indivíduo 1 e o segundo pelos nove indivíduos restantes. Na seqüência, o segundo *cluster* poderia ser dividido em dois: um formado pelos indivíduos 8, 9 e 10, enquanto o outro o seria pelos indivíduos restantes; dessa forma, teríamos um total de três *clusters*. Se agora seis *clusters* são necessários, então os indivíduos 1 a 4 formariam quatro *clusters* e os seis indivíduos remanescentes se agrupariam tal como se apresenta no *dendograma d*a Figura 10.1, ou seja, um grupo é formado pelos indivíduos 5, 6 e 7, enquanto os indivíduos 8, 9 e 10 formam o outro grupo. Dessa maneira, pode-se ler em ordenadas a distância em que os indivíduos se aglomeram para formar um *cluster* e pode-se, através das distintas ramificações do dendograma, interpretar a estrutura de similaridade dos dados.

Inicialmente, quando cada indivíduo constitui o seu próprio *cluster*, as distâncias entre indivíduos são definidas por *d*, tal como calculado pela equação 10.1. Entretanto, a partir do momento em que vários indivíduos formam um ou mais *clusters*, a questão é de como serão determinadas as distâncias de dissimilaridade entre esses novos *clusters*. Em outras palavras, faz-se necessária uma *regra de aglomeração* para definir quando dois *clusters* são suficientemente similares para

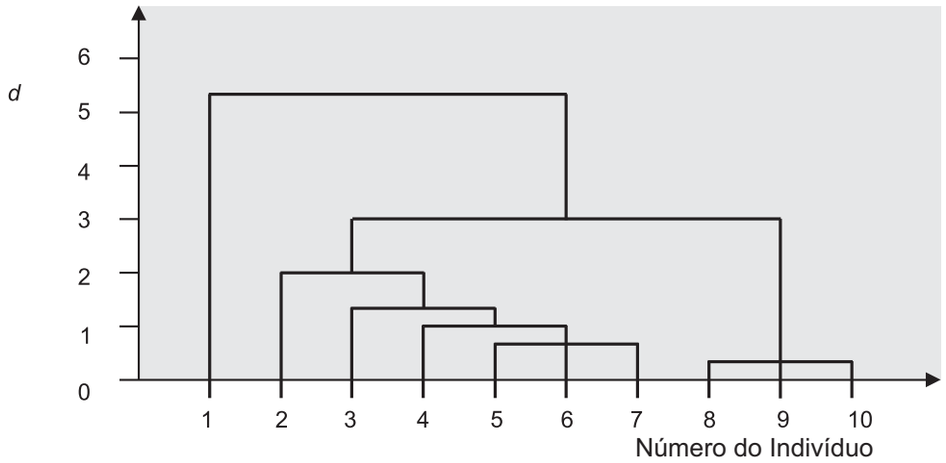


Figura 10.1 – Dendograma hipotético - 10 indivíduos (adap. de Kottegoda e Rosso, 1997)

se juntarem. Uma das várias possibilidades para se definir essa regra foi exemplificada na Figura 10.1; nesse caso, usou-se o critério do *vizinho mais próximo* segundo o qual, a distância entre dois *clusters* é determinada pela distância entre os seus dois respectivos indivíduos que mais se aproximam. Uma possível desvantagem desse critério é a de que ele pode conduzir à formação de extensos *clusters* que se aglomeram meramente porque contem indivíduos próximos.

Um método alternativo e muito utilizado como regra de aglomeração é o descrito por Ward (1963). Em linhas gerais, o método de Ward emprega a análise de variância para determinar as distâncias entre *clusters* e, a cada nova iteração, aglomerá-los de forma a minimizar a soma dos quadrados de quaisquer pares de dois *clusters* hipotéticos. O método de Ward é considerado como eficiente e, em geral, tende a produzir *clusters* pouco extensos e de igual número de indivíduos. Outro método muito empregado é o devido a Hartigan (1975) e conhecido como o das *K-médias* (*K-means clustering*). O princípio desse método é o de que o analista *a priori* pode ter indícios ou hipóteses relativas ao número correto de *clusters* a ser considerado. Dessa forma, o método das *K*-médias irá produzir *K clusters*, os quais deverão ser os mais distintos entre si. Para fazê-lo, o método começa com a formação de *K clusters* iniciais, cujos membros são escolhidos aleatoriamente entre os indivíduos a serem agrupados. Em seguida, os indivíduos são movidos iterativamente de um *cluster* para outro de forma a (1) minimizar a variabilidade intra-*cluster* e (2) maximizar a variabilidade entre os *clusters*. Essa lógica é análoga a se proceder a uma análise de variância ao revés, no sentido que, ao testar a hipótese nula de que as médias grupais são diferentes entre si, a análise de variância confronta a variabilidade entre-grupos com a variabilidade intra-grupos. Em geral, os resultados do método das *K*-médias devem ser

examinados de forma a se avaliar quão distintas são as médias dos *K clusters* obtidos.

Quando aplicada à identificação preliminar de regiões homogêneas para estudos regionais de freqüência de variáveis hidrológicas/hidrometeorológicas, a análise de *clusters* requer algumas considerações específicas. Hosking e Wallis (1997) recomendam atenção para os seguintes pontos :

1. Muitos algoritmos para a aglomeração em *clusters* utilizam o recíproco da distância Euclidiana como medida de similaridade. Nesse caso, é usual padronizar os elementos do vetor das características, dividindo-os pela sua amplitude, ou pelo seu desvio-padrão, de forma que passem a ter variabilidades de ordem de grandeza similares. Essa padronização implica em atribuir ponderações iguais às diferentes características locais, o que pode ocultar a maior ou menor influência relativa de uma delas na forma da curva regional de freqüências. Pode-se compensar essa deficiência pela atribuição direta de diferentes ponderações às características locais consideradas.

2. Alguns métodos, como o das *K*-médias por exemplo, requerem a definição prévia do número de *clusters* a se considerar. É certo, entretanto, que, objetivamente, não se tem *a priori* o número "correto" de *clusters*. Na prática, deve-se buscar um equilíbrio entre regiões demasiadamente grandes ou demasiadamente pequenas, com muitos ou poucos postos de observação. Para as metodologias de análise regional de freqüências que utilizam o princípio da cheia-índice (ou *index-flood*), existe muito pouca vantagem em se empregar regiões muito extensas. Segundo Hosking e Wallis (1997), ganha-se pouca precisão nas estimativas de quantis, ao se usar mais de 20 postos em uma região. Portanto, não há razão premente para se juntar regiões extensas cujas estimativas das distribuições de freqüências são similares.

3. Os resultados da análise de *clusters* devem ser considerados como preliminares. Em geral, são necessários ajustes, muitas vezes subjetivos, cuja finalidade é a de tornar fisicamente coerente a delimitação das regiões, assim como a de reduzir a *medida de heterogeneidade* a ser descrita no item 10.3.2.1. Os ajustes mencionados podem ser obtidos pelas seguintes providências :

- mover um ou mais postos de uma região para outra;
- desconsiderar ou remover um ou mais postos;
- subdividir uma região;
- abandonar uma região e re-alocar os seus postos para outras regiões;
- combinar uma região com outra, ou com outras;

• combinar duas ou mais regiões e redefini-las; e
• obter mais dados e redefinir as regiões.

10.2 – Métodos de Regionalização

10.2.1 – Método de Regionalização dos Quantis Associados a um Risco Especificado

Nesse método, a primeira etapa consiste de uma análise de freqüência local para cada amostra de observações hidrológicas/hidrometeorológicas, de modo a estimar os quantis da variável hidrológica, associados a períodos de retorno previamente especificados, em cada uma das estações de coleta de dados. Em seguida, uma vez fixado um certo período de retorno T, procura-se, por meio de análise de regressão, estabelecer uma relação entre os quantis Q_{Tj} estimados nas diversas estações $j = 1, 2, \ldots, N$, de uma região geográfica, e suas respectivas características fisiográficas e/ou climatológicas. Observe que, nesse caso, não é necessário o ajuste de uma mesma função de distribuição de probabilidades para as amostras provenientes das diferentes estações de coleta de dados, dentro da área em estudo. Portanto, segundo esse método, a partir de algumas características mensuráveis dos locais (ou das bacias) desprovidas de observações hidrológicas/ hidrometeorológicas, pode-se estimar o quantil associado a um determinado tempo de retorno, por meio de um modelo de regressão ajustado aos quantis, tais como estimados localmente a partir das amostras existentes, e as correspondentes características fisiográficas e/ou climatológicas dos locais ou bacias de monitoramento. Apresenta-se, a seguir, uma síntese da seqüencia das etapas necessárias à aplicação desse método:

a) Análise de freqüência local das séries disponíveis na área em estudo;
b) Definição dos tempos de retorno de interesse para regionalização;
c) Definição de uma relação entre os quantis estimados no item (a) e as grandezas fisiográficas e/ou climatológicas dos locais ou bacias monitoradas, para um tempo de retorno fixado, ou seja, $Q_T = f$(características fisiográficas e/ou climatológicas); e
d) Estimação de quantis em locais ou bacias não monitoradas, pela aplicação da equação determinada no item (c), a partir da mensuração das características fisiográficas e/ou climatológicas do local ou bacia de interesse.

Exemplo 10.1 – Apresenta-se, no Anexo 11, os menores valores anuais das vazões médias de 7 dias de duração observadas nas 11 estações da bacia do rio Paraopeba, listadas na Tabela 10.1 e localizadas no mapa da

Figura 10.2. Pede-se realizar um estudo de regionalização das vazões mínimas anuais médias de 7 dias de duração e 10 anos de tempo de retorno, empregando o método de regionalização dos quantis de risco especificado.

Tabela 10.1 – Características fisiográficas das estações do exemplo 10.1

Código	Estação	Rio	Área Km²	$P_{médio}$ (m)	I_{equiv} (m/Km)	L (Km)	Junções/Km²
40549998	São Brás do Suaçui Montante	Paraopeba	461,4	1,400	2,69	52	0,098
40573000	Joaquim Murtinho	Bananeiras	291,1	1,462	3,94	32,7	0,079
40577000	Ponte Jubileu	Soledade	244	1,466	7,20	18,3	0,119
40579995	Congonhas Linigrafo	Maranhão	578,5	1,464	3,18	41,6	0,102
40680000	Entre Rios de Minas	Brumado	486	1,369	1,25	47,3	0,136
40710000	Belo Vale	Paraopeba	2760,1	1,408	1,59	118,9	0,137
40740000	Alberto Flores	Paraopeba	3939,2	1,422	1,21	187,4	0,134
40800001	Ponte Nova do Paraopeba	Paraopeba	5680,4	1,449	1,00	236,33	0,141
40818000	Juatuba	Serra Azul	273	1,531	4,52	40	0,066
40850000	Ponte da Taquara	Paraopeba	8734	1,434	0,66	346,3	0,143
40865001	Porto do Mesquita (CEMIG)	Paraopeba	10192	1,414	0,60	419,83	0,133

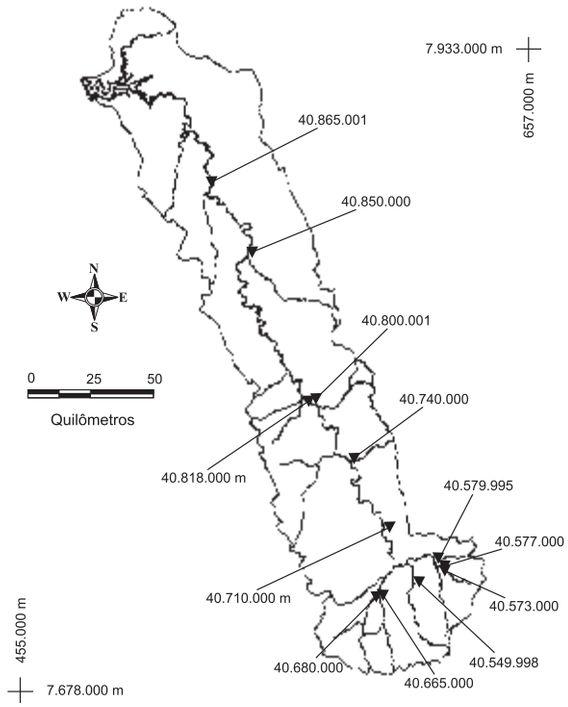


Figura 10.2 – Localização das estações da bacia do rio Paraopeba

Solução: Quando se aplica o método de regionalização dos quantis associados a um risco especificado, faz-se, inicialmente, uma análise de freqüência dos dados de cada uma das estações dentro da área em estudo. Como mencionado anteriormente, não é necessário que se ajuste uma mesma distribuição de freqüência aos dados das estações. Neste exemplo, foram

verificados os ajustes das distribuições de Gumbel e Weibul, para mínimos, aos dados das 11 estações da bacia do rio Paraopeba, de acordo com os procedimentos descritos no capítulo 8. A distribuição que melhor se ajustou aos de todas as estações foi a de Weibull, cuja FAP é apresentada a seguir:

$$F(x) = 1 - \exp(-y) \quad \text{com} \quad y = \left(\frac{x - \varepsilon}{\beta - \varepsilon} \right)^{\alpha} \quad \text{e} \quad \varepsilon \neq 0 \tag{10.2}$$

Nesse caso, a função de quantis é

$$x(F) = \varepsilon + \left\{ (\beta - \varepsilon) \left[-ln\left(1 - \frac{1}{T} \right) \right]^{1/\alpha} \right\} \tag{10.3}$$

Segundo Kite (1977), o parâmetro α pode ser estimado por:

$$\hat{\alpha} = \frac{1}{C_0 + C_1\gamma + C_2\gamma^2 + C_3\gamma^3 + C_4\gamma^4} \tag{10.4}$$

onde o coeficiente de assimetria, g, é dado por:

$$g = \frac{n\sum_{1}^{n} \left(X - \overline{X} \right)^3}{(n-2)\left[\sum \left(X - \overline{X} \right)^2 \right]^{3/2}} \tag{10.5}$$

e deve estar compreendido no intervalo $-1{,}02 \leq g \leq 2{,}0$. Segundo Kite (1977), os coeficientes da equação 10.4 são dados por

C_0	C_1	C_2	C_3	C_4
0,2777757913	0,3132617714	0,0575670910	-0,0013038566	-0,0081523408

Em seguida, o parâmetro β pode ser estimado por:

$$\hat{\beta} = \overline{X} + S_X A(\alpha) \tag{10.6}$$

na qual, $\overline{X}$ denota a média amostral, S_X representa o desvio-padrão amostral e

$$A(\alpha)=\left[1-\Gamma\left(1+\frac{1}{\alpha}\right)\right]B(\alpha) \tag{10.7}$$

$$B(\alpha)=\left[\Gamma\left(1+\frac{2}{\alpha}\right)-\Gamma^2\left(1+\frac{1}{\alpha}\right)\right]^{-1/2} \tag{10.8}$$

Finalmente, o parâmetro ε é estimado por:

$$\hat{\varepsilon}=\hat{\beta}-S_X B(\alpha) \tag{10.9}$$

Os parâmetros da distribuição de Weibull, calculados com as equações acima, e as vazões mínimas com 10 anos de tempo de retorno estimadas com a equação 10.3 das 11 estações estão apresentadas na Tabela 10.2.

Tabela 10.2 – Parâmetros da distribuição de Weibull e a $Q_{7,10}$

	40549998	40573000	40577000	40579995	40680000	40710000	40740000	40800001	40818000	40850000	40865001
α	3,2926	3,1938	3,6569	3,3357	3,9337	3,4417	4,1393	3,3822	3,5736	3,5654	2,6903
β	2,7396	1,6172	1,5392	4,0013	2,3535	17,4100	21,6187	31,3039	1,4530	42,5004	44,3578
ε	0,8542	0,6035	0,3837	0,9098	0,3668	5,5824	7,1853	4,0250	0,2456	8,4619	21,5915
$Q_{7,10}(m^3/s)$	1,806	1,105	1,008	2,484	1,488	11,733	15,566	18,049	0,889	26,569	31,455

Após a análise de freqüência e a definição dos tempos de retorno que serão regionalizados, a próxima etapa é a definição de uma relação para cada tempo de retorno, entre os quantis e as grandezas fisiográficas e climatológicas que permitam a explicação da variável de interesse. Nesse exemplo, de acordo com o enunciado, será estabelecida a relação entre a $Q_{7,10}$, da Tabela 10.2, e as características fisiográficas apresentadas na Tabela 10.1. Apresenta-se, a seguir, os resultados obtidos por meio da aplicação dos procedimentos de cálculo para a definição de um modelo de regressão múltipla, tal como detalhados no capítulo 9. Inicialmente foi calculada a matriz de correlação simples entre a variável prevista, $Q_{7,10}$, e os possíveis preditores, Tabela 10.1, cujos resultados encontram-se na Tabela 10.3.

Tabela 10.3 – Matriz de correlações

	Área Km²	$P_{médio}$ (m)	I_{equiv} (m/Km)	L (Km)	Junções (Km²)	$Q_{7,10}(M^3/S)$
Área (Km²)	1					
P médio (m)	-0,22716	1				
I equiv (m/km)	-0,675	0,600687	1			
L (km)	0,997753	-0,24112	-0,69617	1		
Junções/Km2	0,624234	-0,65707	-0,61808	0,609301	1	
$Q_{7,10}$ (m³/s)	0,993399	-0,25256	-0,69904	0,992116	0,65669	1

Em seguida, foram testados diversos modelos potenciais, combinando as variáveis preditoras apresentadas na Tabela 10.3. Para tanto, foi necessário fazer a transformação logarítmica das variáveis. Como as variáveis Área

(km²) e L (km) apresentam alta correlação entre si, elas não foram utilizadas conjuntamente nos modelos testados. Ao final da análise, estabeleceu-se a seguinte relação:

$$Q_{7,10}\left(m^3/s\right) = 0{,}0047\,A^{0{,}9629}\left(km^2\right) \qquad (10.10)$$

a qual é valida para o intervalo $244km^2 \le A\left(km^2\right) \le 10.192km^2$. A Figura 10.3 apresenta a equação ajustada e os intervalos de confiança a 95% da reta de regressão e do valor previsto, no espaço logarítmico.

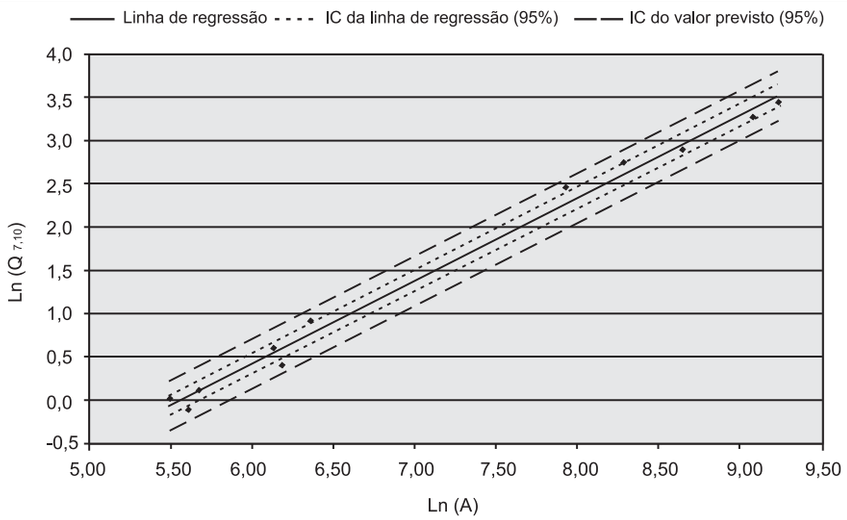


Figura 10.3 – Linha de regressão e os intervalos de confiança para o exemplo 10.1

10.2.2 – Métodos que Regionalizam os Parâmetros da Distribuição de Probabilidades

Para a aplicação desses métodos, o pressuposto é o de que uma mesma função de distribuição de probabilidades seja válida para todas as estações de coleta de dados, localizadas em uma região considerada homogênea, do ponto de vista da variável a ser regionalizada, havendo, portanto, a necessidade de delimitação desta área geográfica. Como conseqüência dessa premissa, os dados de cada estação devem ser ajustados a uma função de distribuição de probabilidades previamente definida para a região homogênea. Uma forma de se avaliar se a distribuição de probabilidades pode ser usada em toda a região é a de analisar o comportamento das distribuições empíricas das diferentes estações, adimensionalizadas pelas respectivas médias amostrais locais, em um único papel de probabilidades. O uso de papéis de probabilidades e o cálculo de distribuições empíricas foram detalhados no capitulo 8.

A variabilidade espacial, ao longo da região em estudo, pode ser sintetizada por meio de um estudo de regressão entre os *i*-ésimos parâmetros, θ_{ij}, que definem a distribuição de probabilidades em cada estação (*j*) e as grandezas fisiográficas e/ou climatológicas locais. Desse modo, é possível definir os parâmetros da distribuição de probabilidades ajustada para a região, em qualquer ponto, a partir das características fisiográficas e/ou climatológicas locais. Em síntese, as etapas seqüenciais para a execução desse método de regionalização são as seguintes:

a) Definição da região homogênea;
b) Definição da distribuição de probabilidades a ser ajustada às diferentes amostras localizadas no interior da região homogênea;
c) Estimação dos parâmetros da distribuição para cada série da região homogênea;
d) Definição de uma relação entre os *i*-ésimos parâmetros, θ_{ij}, que definem a distribuição de probabilidades em cada estação (*j*) e as grandezas fisiográficas e/ou climatológicas locais, e.g. $\theta_{ij} = f$ (características fisiográficas e/ou climatológicas);
e) Estimação dos quantis para um certo local de interesse, a partir da distribuição de probabilidades adotada para a região, utilizando os parâmetros estimados pela relação estabelecida no item (d).

Exemplo 10.2 – Apresenta-se, no Anexo 12, os valores das vazões médias diárias máximas anuais de 07 estações fluviométricas da bacia do rio Paraopeba, listadas na Tabela 10.4 e localizadas no mapa da Figura 10.2. Pede-se realizar um estudo de regionalização das vazões médias diárias máximas anuais, pelo o método dos parâmetros regionais da distribuição de probabilidades.

Tabela 10.4 – Estações para regionalização de vazões diárias máximas anuais

Código	Estação	Rio	Área (Km²)	$P_{médio}$ (m)	I_{equiv} (m/Km)	L (Km)	Junções/Km²
40549998	São Brás do Suaçui Montante	Paraopeba	461,4	1,400	2,69	52	0,098
40573000	Joaquim Murtinho	Bananeiras	291,1	1,462	3,94	32,7	0,079
40577000	Ponte Jubileu	Soledade	244	1,466	7,2	18,3	0,119
40579995	Congonhas Linigrafo	Maranhão	578,5	1,464	3,18	41,6	0,102
40665000	Usina João Ribeiro	Camapuã	293,3	1,373	2,44	45,7	0,123
40710000	Belo Vale	Paraopeba	2760,1	1,408	1,59	118,9	0,137
40740000	Alberto Flores	Paraopeba	3939,2	1,422	1,21	187,4	0,134

A primeira etapa consiste em verificar se as estações da Tabela 10.4 formam uma região homogênea, ou seja, se os dados fluviométricos podem ser ajustados a uma mesma distribuição de probabilidade. Considerando a precipitação média sobre a área de drenagem (Tabela 10.4), como uma

característica local, verifica-se que é plausível supor que a região seja considerada homogênea. Os valores de coeficiente de variação das séries, conforme os resultados das estatísticas locais da Tabela 10.5, também são indicadores positivos da premissa de que a região é homogênea. Além disso, avaliou-se o comportamento das distribuições empíricas adimensionalisadas pelas médias amostrais locas, em um papel de probabilidades. Na Figura 10.4, pode-se verificar o alinhamento das distribuições empíricas adimensionais, grafadas em papel de Gumbel, utilizando a fórmula de Gringorten para cálculo da posição de plotagem.

Tabela 10.5 – Estatísticas locais das amostras do exemplo 10.2

Estações	40549998	40573000	40577000	40579995	40665000	40710000	40740000
Média (m³/s)	60,9	31,5	29,7	78,2	30,0	351,6	437,1
DP (m³/s)	24,0	10,6	9,2	35,7	10,3	149,0	202,8
CV	0,39	0,34	0,31	0,46	0,34	0,42	0,46

Após a definição da região homogênea, efetua-se a seleção da distribuição de probabilidades. Nesse caso, foram testadas somente distribuições de dois parâmetros. Seguindo os procedimentos preconizados no capítulo 8, selecionou-se a distribuição de Gumbel.

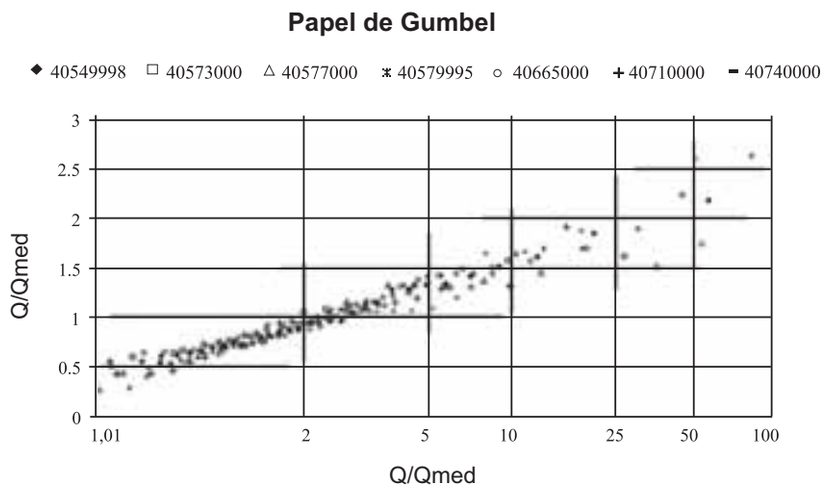


Figura 10.4 – Distribuições empíricas adimensionais

A Tabela 10.6 apresenta os parâmetros de posição (α) e de escala (β) da distribuição de Gumbel, estimados pelo método dos momentos, a partir das observações das sete estações da região homogênea.

Tabela 10.6 – Parâmetros da distribuição de Gumbel

	Estação	Rio	α	β
40549998	São Brás do Suaçui Montante	Paraopeba	18,69	50,07
40573000	Joaquim Murtinho	Bananeiras	8,24	26,71
40577000	Ponte Jubileu	Soledade	7,21	25,53
40579995	Congonhas Linigrafo	Maranhão	27,83	62,13
40665000	Usina João Ribeiro	Camapuã	8,05	25,31
40710000	Belo Vale	Paraopeba	116,15	284,61
40740000	Alberto Flores	Paraopeba	158,13	345,81

A próxima etapa do método de regionalização dos parâmetros da distribuição de probabilidades consiste na definição da variabilidade espacial por meio de um estudo de regressão entre os parâmetros, θ_{ij}, que definem a distribuição de probabilidades em cada estação (j) e as grandezas fisiográficas e/ou climatológicas da região. Assim, foi feita uma análise regressão entre os parâmetros da distribuição de Gumbel, da Tabela 10.6, e as características fisiográficas da Tabela 10.4. Apresenta-se na Tabela 10.7 a matriz de correlação simples entre os parâmetros e as variáveis preditoras.

Tabela 10.7 – Matriz de correlações, exemplo 10.2

	Área Km²	$P_{médio}$ (m)	I_{equiv} (m/Km)	L (Km)	Junções (Km²)	α	β
Área (Km²)	1,000						
P médio (m)	-0,202	1,000					
I equiv (m/km)	-0,635	0,627	1,000				
L (km)	0,984	-0,306	-0,715	1,000			
Junções/Km²	0,688	-0,437	-0,323	0,651	1,000		
α	0,999	-0,193	-0,643	0,978	0,687	1,000	
β	0,995	-0,209	-0,641	0,967	0,699	0,997	1,000

Diversos modelos lineares e potenciais foram aqui testados, a partir de combinações das variáveis preditoras apresentadas na Tabela 10.4. Em alguns casos, foi necessário fazer a transformação logarítmica das variáveis. Em decorrência da alta correlação entre as variáveis explicativas Área (km²) e L (km), elas não foram utilizadas conjuntamente nos modelos testados. Ao final da análise, foram estabelecidas s as seguintes relações:

$$\hat{\alpha} = 0{,}0408 A(km^2) \tag{10.11}$$

$$\hat{\beta} = 0{,}1050 A^{0{,}9896}(km^2) \tag{10.12}$$

onde A é área de drenagem em km². A Figura 10.5 apresenta as equações ajustadas e os intervalos de confiança a 95% da reta de regressão e do valor previsto.

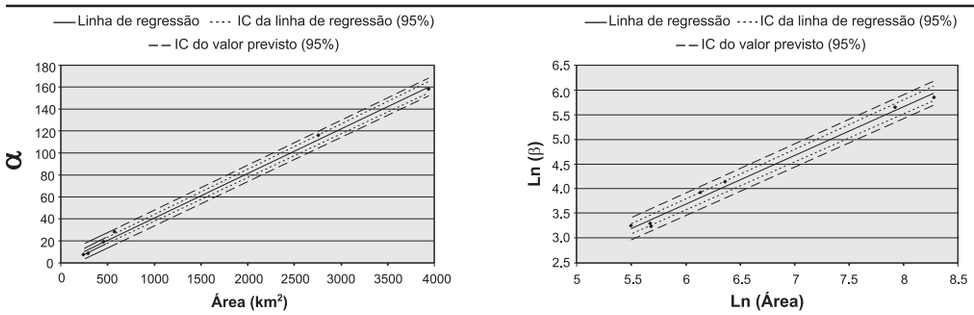


Figura 10.5 – Linhas de regressão e intervalos de confiança, exemplo 10.2

Com as equações 10.11, 10.12 e a função inversa da distribuição de Gumbel, obtém-se, então, a seguinte função de quantis regional:

$$x(T) = \hat{\beta} - \hat{\alpha}\left[ln\left(-ln\left(1 - \frac{1}{T} \right) \right) \right] =$$

$$= \left(0{,}1050\, A^{0{,}9896}\right) - \left(0{,}0408\, A\right)\left[ln\left(-ln\left(1 - \frac{1}{T} \right) \right) \right] \qquad (10.13)$$

válida para $244km^2 \leq A(km^2) \leq 3940km^2$, onde A é área de drenagem em km^2. A equação 10.13 permite a estimativa de vazões médias diárias máximas associadas a diferentes tempos de retorno, em locais da região homogênea que não possuem coleta sistemática de informações, apenas a partir de suas respectivas áreas de drenagem.

10.2.3 – Método *Index-Flood* ou da Cheia-Índice

O termo *index-flood* (cheia-índice) foi introduzido por Dalrymple (1960), dentro de um contexto de regionalização de vazões de cheia. Trata-se de um expediente para adimensionalizar quaisquer dados obtidos em pontos distintos de uma região considerada homogênea, com a finalidade de utilizá-los como um conjunto amostral único. Apesar de fazer referência a cheias, o método e o termo *index-flood* têm uso consagrado em estudos de regionalização de freqüência de qualquer tipo de variável.

Seja o caso de se regionalizar as freqüências de uma variável genérica X, cuja variabilidade espaço-temporal foi amostrada em N locais, estações ou *postos* de observação, de uma certa área geográfica. As observações indexadas por i, tomadas nos postos indexados por j, formam amostras de tamanho variável n_j e são denotadas por $X_{i,j}$, $i = 1, ..., n_j$; $j = 1, ..., N$. Se F, $0 < F < 1$, representa a distribuição de freqüências da variável X no posto j, então, a função de *quantis*

nesse local é simbolizada por $X_j(F)$. A hipótese básica do método *index-flood* é a de que os postos formam uma região *homogênea*, ou seja, as distribuições de freqüências nos *N* pontos são idênticas, a menos de um *fator de escala local* denominado *index-flood* ou fator de adimensionalização. Formalmente,

$$X_j(F) = \mu_j \; x(F), \quad j = 1, \ldots, N \tag{10.14}$$

onde μ_j é o *index-flood*, ou fator de adimensionalização do local *j*, e $x(F)$ representa a *curva regional de quantis adimensionais*, algumas vezes denominada *curva regional de crescimento*, comum a todos os postos.

O fator de escala μ_j pode ser estimado por qualquer medida de posição ou tendência central da amostra de observações $\{X_{1,j}, X_{2,j}, ..., X_{nj,j}\}$. Os *dados adimensionais padronizados* $x_{i,j} = X_{i,j} / \hat{\mu}_j, i = 1, \ldots, n_j; j = 1, \ldots, N$ formam a base para se estimar a curva regional de quantis adimensionais $x(F)$. A curva de freqüência regional pode ser paramétrica, ou seja, obtida pelo ajuste de uma distribuição de probabilidades aos dados adimensionais regionais, ou não paramétrica. A curva regional não paramétrica é definida a partir das curvas empíricas das estações da mesma região homogênea, grafadas em papel de probabilidade. A curva regional não paramétrica é traçada a sentimento, de modo que ela seja próxima da mediana das curvas empíricas individuais da região homogênea.

As premissas inerentes ao método *index-flood* são:

a) as observações em um posto qualquer são identicamente distribuídas;
b) as observações em um posto qualquer não apresentam dependência estatística serial;
c) as observações em diferentes postos são estatisticamente independentes;
d) as distribuições de freqüência em diferentes postos são idênticas, a menos de um fator de escala; e
e) a forma matemática da curva regional de quantis adimensionalizados pode ser corretamente especificada.

Segundo Hosking e Wallis (1997), as premissas (a) e (b) são plausíveis para diversos tipos de variáveis, principalmente aquelas relacionadas a máximos anuais. Entretanto, é improvável que as três últimas premissas possam ser completamente verificadas por dados hidrológicos, meteorológicos ou ambientais. Sabe-se, por exemplo, que precipitações frontais ou estiagens severas são eventos que afetam extensas áreas. Como essas áreas podem conter vários postos de observação da variável em questão, é provável que as amostras, coletadas em pontos distintos,

apresentem, entre si, um grau de correlação significativo. Ainda segundo Hosking e Wallis (1997), na prática, as premissas (d) e (e) jamais são verificadas com exatidão. Apesar dessas restrições, esses autores sugerem que as premissas do método *index-flood* podem ser *razoavelmente aproximadas* tanto pela escolha criteriosa dos postos componentes de uma região, como também pela seleção apropriada de uma função de distribuição de freqüências que apresente consistência com os dados amostrais.

De forma esquemática, as etapas necessárias para aplicação do método *index-flood* são as seguintes:

A) Análise Regional de Consistência de Dados

A primeira etapa da análise regional de freqüência de variáveis aleatórias é certificar-se (i) que os dados coletados em qualquer dos postos de observação estão isentos de erros grosseiros e (ii) que todos os dados individuais provêm de uma mesma distribuição de freqüências.

No caso de dados hidrológicos ou hidrometeorológicos, os erros grosseiros devem-se principalmente a leitura, transcrição ou processamento incorretos. Esses erros são muito freqüentes nas leituras linimétricas e pluviométricas, nas quais a intervenção humana é mais presente e, em conseqüência, a probabilidade de erro é maior. Em alguns casos, a identificação e eliminação dos erros grosseiros presentes nas séries hidrológicas/hidrometeorológicas não são tarefas de fácil execução.

Quando são alteradas as circunstâncias (localização, regime, equipamento de medição) sob as quais os dados são coletados, as séries hidrológicas/ hidrometeorológicas podem vir a apresentar tendências e não-estacionariedade. Nesses casos, a distribuição de freqüências dos dados coletados passa a não ser constante no tempo e a série hidrológica/hidrometorológica, como uma amostra única, não pode ser considerada *homogênea* e nem utilizada para a inferência estatística. São exemplos pertinentes : (a) a relocação de um posto pluviométrico para local com características de vento muito diferentes daquelas apresentadas na instalação de origem; (b) a alteração do regime hidrológico causada pela implantação de reservatório de acumulação a montante de um posto fluviométrico; e (c) a utilização de equipamentos não aferidos, defeituosos ou incompatíveis com a sistemática padrão de coleta de dados primários.

As técnicas mais usuais para a identificação de erros e heterogeneidades nas séries hidrológicas/hidrometeorológicas são :

• comparação de cotagramas e/ou fluviogramas de postos fluviométricos próximos;
• comparação entre totais mensais de precipitação entre postos pluviométricos próximos ou entre um posto e a média de postos vizinhos;
• curvas de dupla acumulação de séries mensais/anuais do posto em questão e do "padrão regional", esse tomado como a média de vários postos das proximidades; e
• testes estatísticos convencionais para verificação de independência, homogeneidade e pontos atípicos (Spearman, Mann-Whitney, Grubbs-Beck, entre outros)

B) Organização e adimensionalização das séries

Essa etapa consiste na montagem das séries com a variável a ser regionalizada, seguida, quando necessário, pelo preenchimento de eventuais falhas. Em seguida, cada elemento, X_{ij} , das séries, onde *i* é o número de ordem do elemento na estação (*j*), é adimensionalizado através da relação entre o elemento e o fator de adimensionalização, μ_j , da estação (*j*), formando, dessa maneira, a série de elementos adimensionais X_{ij}/μ_j. Na proposta inicial de Dalrymple (1960), as séries utilizadas devem ter períodos comuns de dados. Todavia, alguns autores, como Hosking e Wallis (1997), defendem a opinião de que se as séries são homogêneas e representativas da variável em análise, não é necessário o uso de períodos comuns.

C) Definição das curvas empíricas de freqüência de cada estação hidrometeorológica

As curvas empíricas individuais são delineadas por meio de plotagem, em papel de probabilidades, dos valores das séries adimensionalizadas e das posições de plotagem a eles associadas. O uso de papeis de probabilidades e o cálculo de distribuições empíricas estão detalhados no capitulo 8. No trabalho de Dalrymple (1960) e nos estudos de NERC (1975), foi utilizado o papel de probabilidade de Gumbel.

D) Definição das regiões homogêneas e das curvas de freqüência regional

A definição de regiões homogêneas foi anteriormente discutida no item 10.1, onde se fez distinção entre estatísticas locais e características locais, na identificação dessas regiões. Dentre os métodos de estatísticas locais, um dos procedimentos utilizados é a verificação da similaridade da "tendência" das curvas de freqüência individuais. Desse modo, um grupo de curvas com a mesma "tendência", dentro de uma região com características locais semelhantes, forma uma região homogênea.

Como mencionado anteriormente, a curva de freqüência regional pode ser paramétrica ou não paramétrica. A curva regional não paramétrica pode ser definida a partir das curvas empíricas das estações da mesma região homogênea, grafadas no papel de probabilidade. Essa é traçada 'a sentimento', de modo que a curva regional aproxime-se da mediana das curvas empíricas individuais da região homogênea. Tucci (2002) apresenta também o ajuste gráfico aos pontos médios: os pontos médios são determinados pela média aritmética dos valores adimensionais, $X_{i,j}/\hat{\mu}_j$, localizados em intervalos iguais pré-estabelecidos da variável reduzida utilizada para construir o gráfico de probabilidades. Por exemplo, se for utilizado o papel de probabilidades de Gumbel, a variável reduzida é calculada por $y = -ln(-ln(1-1/T))$ e os intervalos podem ser -3,5 a –3,0; -3,0 a –2,5; -2,5 a –2,0; .... 4,0. Ressalve-se, entretanto, que, como o traçado da curva regional é efetuado 'a sentimento', sua extrapolação para tempos de retorno maiores é subjetiva e problemática.

No caso de se desejar estabelecer uma curva regional paramétrica, os *dados adimensionais padronizados* $x_{i,j} = X_{i,j}/\hat{\mu}_j$, $i = 1, \ldots, n_j$; $j = 1, \ldots, N$ formam a base para se estimar a curva regional de quantis adimensionais $x(F)$. A forma genérica de $x(F)$ é conhecida, a menos dos p parâmetros θ_1, ..., θ_p que são próprios da distribuição F e, em geral, são funções das características populacionais de posição central, dispersão e assimetria. Hosking e Wallis (1997) propõem que os parâmetros da curva regional de quantis adimensionais, denotada por $x(F;\ \theta_1,\ \ldots,\ \theta_p)$, sejam obtidos pela ponderação dos parâmetros locais $\hat{\theta}_k^{(j)}$, $k = 1, \ldots, p$, estimados para cada posto j, pelos respectivos tamanhos das amostras. Portanto, a estimativa do parâmetro regional θ_k^R é dada pela média ponderada dos parâmetros da distribuição adotada para a região homogênea, os quais são calculados considerando as séries de valores adimensionais de cada estação da região. As médias são ponderadas pelo tamanho das séries, n_j, que formam a região homogênea, ou seja

$$\hat{\theta}_k^R = \frac{\sum_{j=1}^{N} n_j \hat{\theta}_k^{(j)}}{\sum_{j=1}^{N} n_j} \qquad (10.15)$$

O cálculo dos parâmetros regionais da distribuição adotada para a região homogênea permite a estimativa da curva regional de quantis adimensionais $\hat{x}(F) = x\left(F;\hat{\theta}_1^R,\ldots,\hat{\theta}_p^R\right)$. Salienta-se que, a escolha da distribuição regional também é balizada pelas mesmas considerações feitas em relação à análise de freqüência local apresentada no capitulo 8.

E) Análise de regressão

A análise de regressão objetiva explicar a variação espacial do fator de adimensionalização, μ_j, de cada estação (*j*), a partir das características da bacia, tais como, áreas de drenagem, precipitação anual, declividade do canal principal, entre outras, ou seja,

$$\hat{\mu}_j = f \text{ (características da bacia)} \qquad (10.16)$$

Os modelos de regressão mais freqüentes são o potencial, o exponencial e o logarítmico, com alguma preferência pelo primeiro. Independentemente do tipo de função empregada, o modelo ideal é aquele com o menor número de variáveis explicativas e que apresenta pequeno erro padrão de estimativa com alto coeficiente de determinação.

F) Estimação de um evento associado a um período de retorno qualquer

Estima-se o quantil adimensional associado a um período de retorno, $(X/\mu)_T$, a partir da curva adimensional regional. Em seguida, estima-se o fator de adimensionalização, $\hat{\mu}_j$, por meio da equação de regressão válida para qualquer local da região homogênea, e calcula-se o evento, X_T, para o período de retorno, *T*, através da seguinte equação:

$$X_T = (X/\mu)_T \hat{\mu}_j \qquad (10.17)$$

Exemplo 10.3 – No Anexo 12, estão apresentados os valores das vazões médias diárias máximas anuais de 07 estações da bacia do rio Paraopeba, localizadas no mapa da Figura 10.2 e listadas na Tabela 10.4 do exemplo 10.2. Pede-se realizar um estudo de regionalização das vazões máximas aplicando o método *index-flood* ou da cheia-índice.

Após análise de consistência dos dados, as séries são organizadas e adimensionalizadas. Neste exemplo, o fator de adimensionalização adotado foi a média das séries. Em seguida, são definidas as curvas empíricas de freqüência das séries adimensionalizadas, as quais devem ser grafadas por meio de plotagem, em papel de probabilidades, dos valores das séries adimensionalizadas e das correspondentes posições de plotagem. A Figura 10.4, do exemplo 10.2, apresenta as distribuições empíricas das 7 estações da região homogênea grafadas em papel de Gumbel, empregando a fórmula

de Gringorten para cálculo da posição de plotagem. Conforme análise, no exemplo 10.2, as precipitações médias sobre as áreas de drenagem das estações, Tabela 10.4, as estatísticas locais da Tabela 10.5 e o alinhamento das distribuições empíricas adimensionais em papel de Gumbel, Figura 10.4, são indicadores de que as estações formam uma região homogênea.

A definição da curva regional pode ser realizada traçando, 'a sentimento', a curva que se aproxima da mediana entre as curvas empíricas individuais da região homogênea, ou, ainda, pela estimação dos parâmetros da curva regional paramétrica. Os parâmetros regionais θ_k^R podem ser estimados pelas médias dos parâmetros da distribuição adotada para a região homogênea, ponderadas pelos respectivos comprimentos das séries de valores adimensionais de cada estação da região, ou seja, pelo tamanho das séries, n_j, conforme equação 10.15. Avaliando somente a possibilidade de ajuste de distribuições de 2 parâmetros, o exemplo 10.2 mostrou que a distribuição de Gumbel pode ser uma candidata a distribuição regional. Assim, usando os parâmetros estimados a partir das séries adimensionais, foram calculados os parâmetros da distribuição regional de Gumbel, por meio da equação 10.15, os quais estão apresentados na Tabela 10.8.

Tabela 10.8 – Parâmetros das distribuições de Gumbel adimensionais, exemplo 10.3

Estações	Média	Desvio Padrão	N	α	β
40549998	1	0,394	32	0,307	0,823
40573000	1	0,336	15	0,262	0,849
40577000	1	0,311	20	0,243	0,860
40579995	1	0,456	47	0,356	0,795
40665000	1	0,345	30	0,269	0,845
40710000	1	0,424	25	0,330	0,809
40740000	1	0,464	28	0,362	0,791
Parâmetros Regionais				0,314	0,819

Em decorrência, a função inversa da distribuição de Gumbel regional é a seguinte:

$$\frac{Q_{max}}{Q_{med-max}}(T) = 0{,}819 - 0{,}314\left[ln\left(-ln\left(1 - \frac{1}{T} \right) \right) \right] \qquad (10.18)$$

A Tabela 10.9 apresenta os quantis adimensionais associados a diferentes tempos de retorno e a Figura 10.6 mostra as posições relativas da distribuição regional, em meio às distribuições empíricas adimensionais individuais.

Tabela 10.9 – Quantis regionais adimensionais

T (anos)	1,01	2	5	10	20	25	50	75	100
Quantil Regional	0,339	0,934	1,289	1,525	1,751	1,822	2,043	2,171	2,262

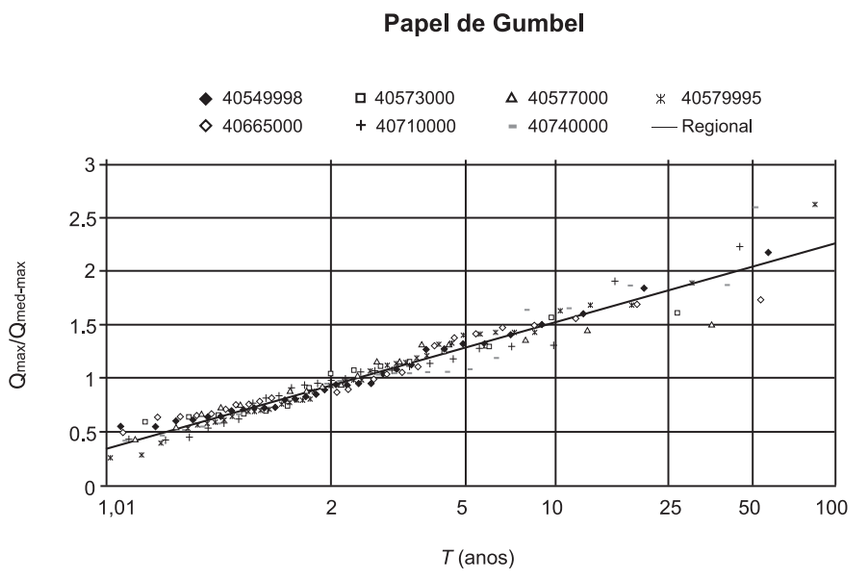


Figura 10.6 – Distribuição regional adimensional

Finalmente, foi feito o estudo de regressão entre os fatores de adimensionalização, neste caso as médias das séries, $Q_{med\text{-}max}$, apresentadas na Tabela 10.5, com as características da bacia (Tabela 10.4). Ao final da análise, foi adotado o seguinte modelo potencial:

$$Q_{med-max} = 0{,}1098\,A^{1{,}0125}\left(km^2\right) \tag{10.19}$$

válida para $244km^2 \le A\left(km^2\right) \le 3940km^2$

O cálculo do fator de adimensionalização, $Q_{med\text{-}max}$, permite a estimação de quantis associados a diferentes tempos de retorno para locais não monitorados, situados dentro da região homogênea, por meio da seguinte equação:

$$Q_{max}(T) = Q_{med-max}\left\{0{,}819 - 0{,}314\left[ln\left(-ln\left(1-\frac{1}{T}\right)\right)\right]\right\} \tag{10.20}$$

Substituindo a equação 10.19 na 10.20, obtém-se uma relação que permite a estimação direta de quantis associados a diferentes tempos de retorno para pontos não monitorados, localizados dentro da região homogênea, ou seja,

$$Q_{max}(T) = \left(0{,}1098 A^{1{,}0125}(km^2)\right)\left\{0{,}819 - 0{,}314\left[ln\left(-ln\left(1-\frac{1}{T}\right)\right)\right]\right\} \quad (10.21)$$

Assim, se o projeto de uma determinada estrutura hidráulica, localizada em algum curso d'água da região homogênea, exigir a estimativa da vazão de cheia com tempo de retorno de 100 anos, basta substituir o valor da área de drenagem na equação 10.21. Supondo, por exemplo, que a área de drenagem correspondente é de 450 km², a estimativa da vazão média diária máxima anual, com 100 anos de tempo de retorno, é

$$Q_{max}(100) = \left[0{,}1098\left(450^{1{,}0125}\right)\right]\left\{0{,}819 - 0{,}314\left[ln\left(-ln\left(1-\frac{1}{100}\right)\right)\right]\right\} \quad (10.22)$$

$Q_{max}(100) = 120{,}6$ m³/s

10.3 – Regionalização *Index-Flood* Utilizando Momentos-L

As subjetividades presentes em algumas etapas das metodologias existentes, bem como o aparecimento de novas técnicas de inferência estatística, como os momentos ponderados por probabilidades (MPP), apresentados por Greenwood et al. (1979), motivaram os pesquisadores J. R. M. Hosking, do centro de investigações Thomas J. Watson da IBM, e J. R. Wallis, da Universidade Yale, a proporem um conjunto unificado de procedimentos para a análise regional de freqüência de diversos tipos de variáveis, com destaque para as hidrológicas, meteorológicas e ambientais. Em sua revisão sobre os avanços recentes da pesquisa na área de análise de freqüência, Bobée e Rasmussen (1995) consideram a contribuição de Hosking e Wallis como a mais relevante para a obtenção de melhores estimativas das probabilidades de eventos raros.

Em linhas gerais, a metodologia descrita por Hosking e Wallis (1997) baseia-se nos princípios do *index-flood*, ou "cheia-índice", tal como enunciados por Dalrymple (1960), e utiliza os momentos-L, quantidades deduzidas dos momentos ponderados por probabilidades (ver capítulo 6), não só para estimar parâmetros e quantis da distribuição regional de probabilidade, como também para construir estatísticas capazes de tornar menos subjetivas algumas etapas da análise regional de freqüência. Nesse capítulo, os itens subseqüentes procuram apresentar uma visão das etapas da metodologia descrita por Hosking e Wallis (1997).

A metodologia de Hosking e Wallis (1997) fundamenta-se tanto nos princípios do método *index-flood*, enunciados no item 10.2.3, como também em algumas

estatísticas construídas a partir dos momentos-L. Estes últimos foram formalmente definidos no item 6.4 do capítulo 6. Essas estatísticas, a serem detalhadas nos próximos subitens, constituem instrumentos valiosos para diminuir o grau de subjetividade presente nas quatro etapas usuais da análise regional de freqüência. Essas etapas encontram-se sumariadas a seguir.

Etapa 1 : Análise Regional de Consistência de Dados

Essa etapa refere-se à detecção e eliminação de erros grosseiros e/ou sistemáticos eventualmente existentes nas amostras individuais dos vários postos de observação. Além das técnicas usuais de análise de consistência, como as curvas de dupla acumulação, por exemplo, Hosking e Wallis (1997) sugerem o uso de uma estatística auxiliar, denominada *medida de discordância* (ver item 10.3.1.1), a qual fundamenta-se na comparação das características estatísticas do conjunto de postos com as apresentadas pela amostra individual em questão.

Etapa 2 : Identificação de Regiões Homogêneas

Conforme definição anterior, uma região homogênea consiste de um agrupamento de postos de observação, cujas curvas de quantis adimensionalizados podem ser aproximadas por uma única curva regional. Para determinar a correta divisão dos postos em regiões homogêneas, Hosking e Wallis (1997) sugerem o emprego da técnica de análise de *clusters*. De acordo com essa técnica, os postos são agrupados em regiões consonantes com a variabilidade espacial de algumas características locais, as quais devem ser selecionadas entre aquelas que supostamente podem ter influência sobre as realizações da variável a ser regionalizada. Depois dos postos terem sido convenientemente agrupados em regiões, Hosking e Wallis (1997) sugerem a *medida de heterogeneidade* para testar a correção dos agrupamentos efetuados. Essa medida baseia-se na comparação da variabilidade grupal das características estatísticas dos postos de observação com a variabilidade esperada dessas mesmas características em uma região homogênea. O teste da medida de heterogeneidade será abordado no item 10.3.2.1.

Etapa 3 : Seleção da Função Regional de Distribuição de Probabilidades

Depois dos erros grosseiros e sistemáticos terem sido eliminados das amostras individuais e das regiões homogêneas haverem sido identificadas, a etapa seguinte é a correta prescrição do modelo probabilístico. Para a seleção da função regional de distribuição de probabilidades entre diversos modelos candidatos, Hosking e Wallis (1997) sugerem o emprego do teste da *medida de aderência* (ver item

10.3.3.2). Esse teste é construído de modo a poder comparar algumas características estatísticas regionais com aquelas que se espera obter de uma amostra aleatória simples retirada de uma população, cujas propriedades distributivas são as mesmas do modelo candidato.

Etapa 4 : Estimativa dos Parâmetros e Quantis da Função Regional de Distribuição de Probabilidades

Identificado o modelo probabilístico regional, representado por $\hat{x}(F) = x\left(F; \theta_1^R, \ldots, \hat{\theta}_p^R\right)$, os parâmetros locais $\hat{\theta}_k^{(j)}$, $k = 1, \ldots, p$ são estimados separadamente para cada posto *j* e, em seguida, ponderados, conforme equação 10.15, para produzir a curva regional de quantis adimensionais. Hosking e Wallis (1997) também sugerem que os parâmetros da distribuição regional adotada sejam calculados a partir das estimativas adimensionais regionais dos momentos-L e razões-L, as quais são obtidas a partir das médias ponderadas dos momentos-L e razões-L amostrais das estações da região homogênea. Conforme será descrito no item 10.3.4, as ponderações são feitas pelos tamanhos das amostras.

Hosking e Wallis (1997) codificaram um conjunto de rotinas, em linguagem Fortran-77, para automatização das quatro etapas da metodologia proposta para análise regional de freqüência. Esse conjunto de rotinas encontra-se disponibilizado ao público no repositório de programas *StatLib*, acessível via Internet através da URL *http://lib.stat.cmu.edu/general/lmoments*.

10.3.1 – Análise Regional de Consistência de Dados

Além das técnicas de uso corrente em hidrologia para consistência de dados, Hosking e Wallis (1997) sugerem também a comparação entre os quocientes de momentos-L amostrais calculados para os diferentes postos de observação. Segundo esses autores, os quocientes de momentos-L amostrais são capazes de refletir erros, pontos atípicos e heterogeneidades eventualmente presentes em uma série de observações. Isso pode ser efetuado por meio de uma estatística-síntese, a qual representa a *medida da discordância* entre os quocientes de momentos-L amostrais de um dado local e a média dos quocientes de momentos-L dos vários postos da região.

10.3.1.1 – A Medida de Discordância

10.3.1.1.1 – Descrição

Em um grupo de amostras, a *medida de discordância* tem por objetivo identificar aquelas que apresentam características estatísticas muito discrepantes das grupais. A medida de discordância é expressa como uma estatística única envolvendo as estimativas dos principais quocientes de momentos-L, a saber, o CV-L (ou τ), a Assimetria-L (ou τ_3) e a Curtose-L (ou τ_4). Em um espaço tridimensional de variação desses quocientes de momentos-L, a idéia é assinalar como discordantes as amostras cujos valores $\{\hat{\tau}, \hat{\tau}_3, \hat{\tau}_4\}$, representados por um ponto no espaço, se afastam 'demasiadamente' do núcleo de concentração das amostras do grupo. Para melhor visualização do significado dessa estatística, considere o plano definido pelos limites de variação das estimativas do CV-L e da Assimetria-L para diversos postos de observação de uma região geográfica (Figura 10.7). Nessa figura, as médias grupais encontram-se no ponto assinalado pelo símbolo + , em torno do qual se constroem elipses concêntricas cujos eixos maiores e menores são funções da matriz de covariância amostral dos quocientes de momentos-L. Os pontos considerados discordantes são aqueles que se encontram fora da área definida pela elipse mais externa.

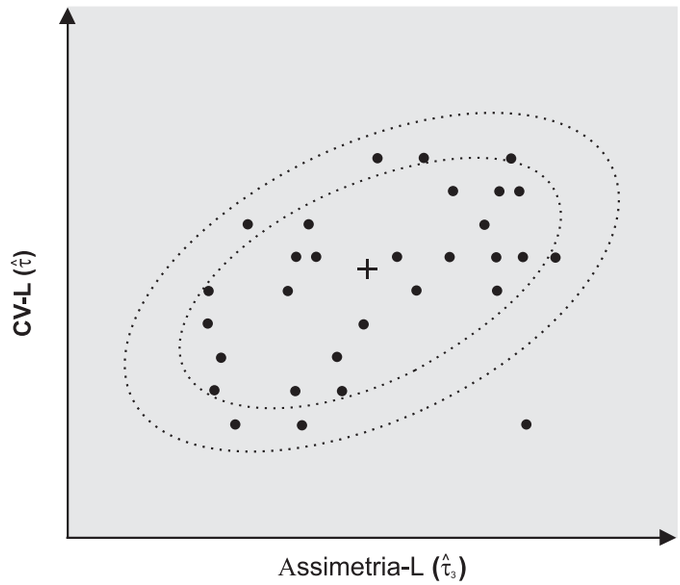


Figura 10.7 – Descrição esquemática da medida de discordância

10.3.1.1.2 – Definição Formal

Os quocientes de momentos-L de um local *j*, a saber, o CV-L, a assimetria-L e a curtose-L, são considerados como um ponto em um espaço tridimensional. Em termos formais, considere que u_j representa um vetor (3x1) contendo esses quocientes de momentos-L, dado por :

$$u_j = \left(t^j \;\; t_3{}^j \;\; t_4{}^j\right)^T \tag{10.23}$$

onde t, t_3 e t_4 denotam CV-L, assimetria-L e curtose-L, respectivamente, e o símbolo T indica matriz transposta. Seja $\overline{u}$ um vetor (3x1), contendo a média grupal ou regional dos quocientes de momentos-L, tomada como a média aritmética simples de u_i para todos os postos estudados, ou seja

$$\overline{u} = \frac{\sum_{i=1}^{N} u_i}{N} = \left(t^R \;\; t_3^R \;\; t_4^R\right)^T \tag{10.24}$$

onde N representa o número de postos de observação do grupo ou região $\boldsymbol{R}$ em questão. Dada a matriz de covariância amostral S, definida por

$$S = (N-1)^{-1} \sum_{i=1}^{N} \left(u_i - \overline{u}\right)\left(u_i - \overline{u}\right)^T \tag{10.25}$$

Hosking e Wallis (1995) definem a medida de discordância D_j, para o local j pela expressão

$$D_j = \frac{N}{3(N-1)} (u_i - \overline{u})^T S^{-1} (u_i - \overline{u}) \tag{10.26}$$

Em trabalhos anteriores, Hosking e Wallis (1993) sugeriram o valor limite $D_j = 3$ como critério para decidir se a amostra é discordante das características grupais. Por exemplo, quando certa amostra produz $D_j \geq 3$, isso significa que ela pode conter erros grosseiros e/ou sistemáticos, ou mesmo pontos atípicos, que a tornam discordantes ou discrepantes das demais do grupo de amostras. Posteriormente, Hosking e Wallis (1995) apresentaram novos valores críticos para D_j, para grupos ou regiões com menos de 15 postos de observação. Esses valores críticos para D_j encontram-se listados Tabela 10.10.

Tabela 10.10 – Valores críticos da medida de discordância - D_j

Nº de postos da região	D_{jcrit}	Nº de postos da região	D_{jcrit}
5	1,333	11	2,632
6	1,648	12	2,757
7	1,917	13	2,869
8	2,140	14	2,971
9	2,329	≥ 15	3
10	2,491		

Fonte: Hosking e Wallis (1995)

De acordo com Hosking e Wallis (1995), para grupos ou regiões com número muito reduzido de postos de observação, a estatística D_j não é informativa. Por exemplo, para $N \leq 3$, a matriz de covariância S é singular e o valor de D_j não pode ser calculado. Para $N = 4$, $D_j = 1$ e, para $N = 5$ ou $N = 6$, os valores de D_j, como indicados na Tabela 10.10, são bastante próximos do limite algébrico da estatística, definido por $D_j \leq (N-1)/3$. Em conseqüência, os autores sugerem o uso da medida de discordância D_j somente para $N \geq 7$.

10.3.1.1.3 – Discussão

Hosking e Wallis (1997) fazem as seguintes recomendações para o uso da medida de discordância D_j:

a) A análise regional de consistência de dados inicia-se com o cálculo das D_j's individuais de todos os postos de uma grande região geográfica, sem considerações preliminares relativas à homogeneidade regional. Aqueles postos assinalados como discordantes devem ser submetidos a uma cuidadosa análise individual (testes estatísticos, curva de dupla acumulação, comparação com postos vizinhos), visando a identificação/eliminação de eventuais inconsistências em seus dados.
b) Em seguida, quando a homogeneidade regional já houver sido definida, as medidas de discordância devem ser recalculadas, desta feita com os postos devidamente agrupados em suas respectivas regiões homogêneas. Se um certo posto se apresentar discordante em uma região, deve ser considerada a possibilidade de sua transferência para outra.
c) Ao longo de toda a análise regional de consistência de dados, deve-se ter em conta que os quocientes de momentos-L amostrais podem apresentar diferenças naturalmente possíveis, mesmo entre postos similares do ponto de vista dos processos físicos em questão. Hosking e Wallis (1997) exemplificam que um evento extremo, porém localizado, pode ter afetado somente alguns postos em uma região. Entretanto, se é provável que um evento como este pode afetar qualquer posto da região, então a providência mais sensata seria a de tratar todo o grupo de

postos como uma única região homogênea, mesmo que alguns possam apresentar medidas de discordância superiores aos valores limites estabelecidos.

10.3.2 – Identificação e Delimitação de Regiões Homogêneas

A identificação e delimitação de regiões homogêneas podem ser realizadas a partir das características locais e estatísticas locais, conforme enunciado no item 10.1. Hosking e Wallis (1997) recomendam que os procedimentos para identificação de regiões homogêneas, baseados em estatísticas locais, sejam utilizados para confirmar a delimitação realizada previamente com as características locais. Dentre os métodos de estatísticas locais, esses autores propõem um teste estatístico, materializado pela *medida de heterogeneidade*, e construído com base nos quocientes de momentos-L amostrais. A descrição da medida de heterogeneidade é objeto do item que se segue.

10.3.2.1 – A Medida de Heterogeneidade Regional

10.3.2.1.1 – Descrição

Como princípio, em uma região homogênea, todos os indivíduos possuem os mesmos quocientes de momentos-L populacionais. Entretanto, as suas estimativas, quais sejam os quocientes de momentos-L calculados a partir das amostras, apresentaram diferenças devidas às flutuações amostrais. Portanto, para um certo conjunto de postos, é natural questionar se a dispersão entre seus quocientes de momentos-L amostrais é maior do que aquela que se esperaria encontrar em uma região homogênea. Essencialmente, é essa a lógica empregada para a construção da medida de heterogeneidade regional.

Pode-se visualizar o significado da medida de heterogeneidade através de diagramas de quocientes de momentos-L, como o da Figura 10.8. Embora outras estatísticas também possam ser usadas, no exemplo hipotético da Figura 10.8, encontram-se grafados o CV-L e a Assimetria-L amostrais de um lado, enquanto que, do outro, estão os seus correspondentes, obtidos a partir de simulações de amostras de mesmo tamanho das originais localizadas, por hipótese, em uma região homogênea. Em diagramas como esses, uma região possivelmente heterogênea mostraria, por exemplo, que os CV-L´s amostrais são mais dispersos do que aqueles obtidos por simulação. Em termos quantitativos, essa idéia básica pode ser traduzida pela diferença relativa centrada entre as dispersões observada e simulada, ou seja, pela razão

$$\frac{\text{(dispersão observada) - (média das simulações)}}{\text{desvio padrão das simulações}}$$

Para tornar possível o cálculo das estatísticas simuladas para a região homogênea, é necessário especificar uma função de distribuição de probabilidades para a população de onde serão extraídas as amostras. Hosking e Wallis (1997) recomendam o emprego da distribuição Kapa de 4 parâmetros, a ser formalmente definida no item 10.3.2.1.2, e justificam que essa recomendação prende-se à preocupação de não assumir *a priori* nenhum comprometimento com distribuições de 2 e/ou 3 parâmetros. Os momentos-L da distribuição Kapa populacional devem reproduzir as médias grupais dos quocientes CV-L, Assimetria-L e Curtose-L, calculados para os dados observados.

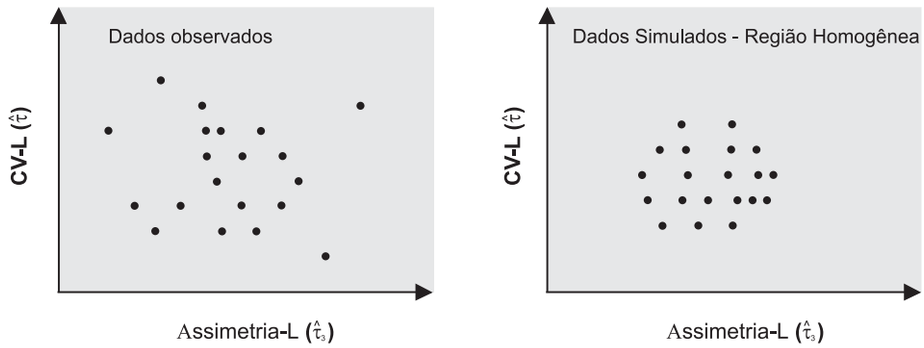


Figura 10.8 – Descrição esquemática do significado de heterogeneidade regional

10.3.2.1.2 – Definição Formal

Considere que uma dada região contenha *N* postos de observação, cada um deles indexado por *j*, com amostra de tamanho n_j e quocientes de momentos-L amostrais representados por $t^j, t_3^j \text{ e } t_4^j$. Considere também que t^R, t_3^R e t_4^R denotam respectivamente as médias regionais dos quocientes CV-L, Assimetria-L e Curtose-L, ponderados, de forma análoga à especificada pela equação 10.15, pelos tamanhos das amostras individuais. Hosking e Wallis (1997) recomendam que a *medida de heterogeneidade*, denotada por *H*, baseie-se preferencialmente no cálculo da dispersão de *t*, ou seja, o CV-L para as regiões proposta e simulada. Inicialmente, efetua-se o cálculo do desvio padrão ponderado *V* dos CV-L´s das amostras observadas, por meio da seguinte expressão :

$$V = \left[\frac{\sum_{j=1}^{N} n_j \left(t^j - t^R \right)^2}{\sum_{j=1}^{N} n_j} \right]^{\frac{1}{2}} \tag{10.27}$$

Em seguida, para a simulação da região homogênea, Hosking e Wallis (1997) sugerem, conforme menção anterior, a utilização da distribuição Kapa de quatro parâmetros. Essa distribuição é definida pelos parâmetros ξ, α, k e h e inclui, como casos particulares, as distribuições Logística, Generalizada de Valores Extremos e Generalizada de Pareto, sendo, portanto, teoricamente capaz de representar variáveis hidrológicas e hidrometeorológicas. As funções densidade, acumulada de probabilidades e de quantis da distribuição Kapa são dadas respectivamente por

$$f(x) = \frac{1}{\alpha} \left[i - \frac{k(x-\xi)}{\alpha} \right]^{\frac{1}{k-1}} \left[F(x) \right]^{1-h} \tag{10.28}$$

$$F(x) = \left\{ 1 - h \left[1 - \frac{k(x-\xi)}{\alpha} \right]^{\frac{1}{k}} \right\}^{\frac{1}{h}} \tag{10.29}$$

$$x(F) = \xi + \frac{\alpha}{k} \left[1 - \left(\frac{1 - F^h}{h} \right)^k \right] \tag{10.30}$$

Se $k > 0$, x tem um limite superior em $\xi + \alpha/k$; se $k \leq 0$, x é ilimitado superiormente; x tem um limite inferior em $\xi + \alpha\left(1 - h^{-k}\right)/k$ se $h > 0$, em $\xi + \alpha/k$ se $h \leq 0$ e $k < 0$, e em -∞ se $h \leq 0$ e $k \geq 0$. Os momentos-L da distribuição Kapa são definidos para $h \geq 0$ e $k > -1$ ou para $h < 0$ e $-1 < k < -1/h$, e dados pelas seguintes expressões :

$$\lambda_1 = \xi + \frac{\alpha(1 - g_1)}{k} \tag{10.31}$$

$$\lambda_2 = \frac{\alpha(g_1 - g_2)}{k} \tag{10.32}$$

$$\tau_3 = \frac{(-g_1 + 3g_2 - 2g_3)}{g_1 - g_2} \tag{10.33}$$

$$\tau_4 = \frac{(-g_1 + 6g_2 - 10g_3 + 5g_4)}{g_1 - g_2} \tag{10.34}$$

onde

$$g_r = \begin{cases} \dfrac{r\Gamma(1+k)\Gamma\left(\dfrac{r}{h}\right)}{h^{1+k}\Gamma\left(1+k+\dfrac{r}{h}\right)} \text{ se } h > 0 \\ \dfrac{r\Gamma(1+k)\Gamma\left(-k-\dfrac{r}{h}\right)}{(-h)^{1+k}\Gamma\left(1-\dfrac{r}{h}\right)} \text{ se } h < 0 \end{cases} \tag{10.35}$$

e $\Gamma(.)$ representa a função gama, tal como anteriormente definida.

Os parâmetros da população Kapa são estimados de modo a reproduzir os quocientes de momentos-L regionais $\{1, t^R, t_3^R, t_4^R\}$. Com os parâmetros populacionais, são simuladas N_{SIM} regiões homogêneas, sem correlação cruzada e/ou serial, contendo N amostras individuais, cada qual com n_i valores da variável normalizada. Em seguida, as estatísticas V_j (j=1, 2, ... , N_{SIM}) são calculadas para todas as simulações de regiões homogêneas, por meio da equação 10.27. A sugestão é a que se faça o número de simulações, N_{SIM}, igual a 500.

A média aritmética das estatísticas V_j, calculadas para cada simulação, fornecerá a dispersão média esperada para a região homogênea, ou seja,

$$\mu_V = \frac{\sum_{j=1}^{N_{SIM}} V_j}{N_{SIM}} \tag{10.36}$$

A medida de heterogeneidade H estabelece uma comparação entre a dispersão observada e a dispersão simulada. Formalmente,

$$H = \frac{(V - \mu_V)}{\sigma_V} \tag{10.37}$$

onde V é a estatística calculada por meio da equação 10.27 utilizando os dados observados na região supostamente homogênea, μ_v é a média aritmética das estatísticas V_j calculada para cada simulação e σ_V é o desvio padrão entre os N_{SIM} valores da medida de dispersão V_j, ou seja,

$$\sigma_V = \sqrt{\frac{\sum_{j=1}^{N_{SIM}} \left(V_j - \mu_V\right)^2}{N_{SIM} - 1}} \tag{10.38}$$

De acordo com o teste de significância, proposto por Hosking e Wallis (1997), se $H < 1$, considera-se a região como "aceitavelmente homogênea", se $1 \leq H < 2$, a região é "possivelmente heterogênea" e, finalmente, se $H \geq 2$, a região deve ser classificada como "definitivamente heterogênea".

10.3.2.1.3 – Discussão

Conforme menção anterior, alguns ajustes subjetivos, tais como a remoção ou o reagrupamento de postos de uma ou mais regiões, podem se tornar necessários para fazer com que a medida de heterogeneidade se amolde aos limites propostos. Entretanto, é possível que, em alguns casos, a heterogeneidade aparente seja devida à presença de um pequeno número de postos 'atípicos' na região. Uma alternativa é a de reagrupá-los em outra região na qual sejam 'mais típicos', muito embora não exista nenhuma razão física evidente de que esse pequeno grupo de postos tenha comportamento distinto do restante dos postos da região de origem. Hosking e Wallis (1997) argumentam que, nesses casos, as razões de natureza física devem ter precedência sobre os de natureza estatística e recomendam a alternativa de manter o grupo de postos 'atípicos', na região originalmente proposta. Hosking e Wallis (1997) continuam a argumentação tomando, como exemplo, a situação em que uma certa combinação de eventos meteorológicos extremos seja passível de ocorrer em qualquer ponto de uma região, mas que, de fato, ela tenha sido registrada em somente alguns de seus postos, durante o período disponível de observações. Os verdadeiros benefícios potenciais da regionalização poderiam ser atingidos em situações como a exemplificada, na qual o conhecimento dos mecanismos físicos associados à ocorrência de eventos extremos permite agrupar todos os postos em uma única região homogênea. Para esse exemplo, os dados locais encontram-se indevidamente influenciados pela *presença ou ausência* de eventos raros e a curva regional de freqüências, construída como a média das

curvas individuais, constitui certamente o melhor instrumento para se estimar os riscos de futuras ocorrências dessa natureza.

A medida de heterogeneidade é construída como um teste de significância da hipótese nula de que a região é homogênea. Entretanto, Hosking e Wallis (1997) argumentam que não se deve interpretá-lo rigorosamente como tal, porque um teste de homogeneidade exato só seria válido sob as premissas que os dados não possuem correlações cruzada e/ou serial e que a função Kapa representa a verdadeira distribuição regional. Mesmo se fosse possível construir um rigoroso teste de significância, ele teria utilidade duvidosa pois, na prática, mesmo uma região moderadamente heterogênea pode produzir melhores estimativas de quantis do que aquelas produzidas pela exclusiva análise de dados locais.

Os critérios $H = 1$ e $H = 2$, embora arbitrários, representam indicadores úteis. Se a medida de heterogeneidade fosse interpretada como um teste de significância e supondo que a estatística H possuísse uma distribuição Normal, o critério de rejeição da hipótese nula de homogeneidade, ao nível $\alpha = 10\%$, seria $H = 1{,}28$. Nesse contexto, o critério arbitrário de $H = 1$ pode parecer muito rigoroso. Entretanto, conforme argumentação anterior, não se quer interpretar a medida H como um teste de significância exato. A partir de resultados de simulação, Hosking e Wallis (1997) demonstraram que, em média, $H \approx 1$ para uma região suficientemente heterogênea, na qual as estimativas de quantis são 20 a 40% menos precisas do que as obtidas para uma região homogênea. Assim sendo, o limite $H = 1$ é visto como o ponto a partir do qual a redefinição da região pode apresentar vantagens. Analogamente, o limite $H = 2$ é visto como o ponto a partir do qual redefinir a região é definitivamente vantajoso.

Em alguns casos, H pode apresentar valores negativos. Eles indicam que há menos dispersão entre os valores amostrais de CV-L do que se esperaria de uma região homogênea com *distribuições individuais de freqüência independentes*. A causa mais provável para esses valores negativos é a presença de correlação positiva entre os dados dos diferentes postos. Se valores muito negativos, como $H < -2$, são observados durante a regionalização, isso pode ser uma indicação de que há muita correlação cruzada entre as distribuições individuais de freqüência ou de que há uma regularidade excessiva dos valores amostrais de CV-L. Para esses casos, Hosking e Wallis (1997) recomendam reexaminar os dados de forma mais cuidadosa.

10.3.3 – Seleção da Distribuição Regional de Freqüência

10.3.3.1 – Seleção das Distribuições Candidatas – Propriedades Gerais

Existem diversas famílias de distribuições de probabilidade que podem ser consideradas candidatas a modelar um conjunto de dados regionais. A sua adequação como distribuições candidatas depende de sua capacidade de reproduzir algumas características amostrais relevantes. Em geral, a seleção da 'melhor' distribuição de probabilidade baseia-se na qualidade e consistência de seu ajuste aos dados disponíveis. Entretanto, o objetivo da análise regional de freqüência não é o de ajustar uma distribuição a uma amostra em particular. De fato, o que se objetiva é a obtenção de estimativas de quantis de uma distribuição de probabilidades da qual se espera serem extraídos futuros valores amostrais. Em outras palavras, o que se preconiza é a seleção, entre diversas candidatas, da distribuição mais *robusta*, ou seja daquela que seja a mais capaz de produzir boas estimativas de quantis, mesmo que os valores por ela previstos possam ter sido extraídos de uma distribuição diferente da que foi ajustada. Todas as considerações feitas no capítulo 8 para a seleção de uma distribuição de probabilidades para a análise de freqüência local são validas para a análise regional. Todavia, no contexto de regionalização, Hosking e Wallis (1997) observam que a grande vantagem potencial da análise regional de freqüência é justamente a de poder estimar as distribuições de mais de dois parâmetros de forma mais confiável do que o seria a partir de uma única amostra local. Seguem adiante afirmando que, uma vez obedecido o preceito da *parcimônia estatística*, recomenda-se o uso de distribuições de mais de dois parâmetros por produzirem estimativas menos enviesadas de quantis nas caudas superior e inferior. Concluem dizendo que para as aplicações da análise regional de freqüência, as distribuições de três a cinco parâmetros são mais apropriadas.

Existem diversos testes de aderência de uma distribuição aos dados amostrais que são passíveis de serem adaptados ao contexto da análise regional de freqüência. Nesse contexto, os seguintes exemplos podem ser citados: gráficos quantil-quantil, testes do Qui-Quadrado, de Kolmogorov-Smirnov e Filliben, bem como diagramas de momentos ou de quocientes de momentos-L. Hosking e Wallis (1997) consideram uma escolha natural tomar como base para um teste de aderência as médias regionais de estatísticas de momentos-L, como por exemplo a Assimetria-L e a Curtose-L, e compará-las às características teóricas das diferentes distribuições candidatas. Essa é a idéia básica da *medida de aderência Z* a ser descrita no item que se segue.

10.3.3.2 – A Medida de Aderência

10.3.3.2.1 – Descrição

Em uma região homogênea, os quocientes de momentos-L individuais flutuam em torno de suas médias regionais. Na maioria dos casos, as distribuições de probabilidade, candidatas a modelar o comportamento da variável em estudo, possuem parâmetros de posição e escala que reproduzem a média e o CV-L regionais. Portanto, a aderência de uma certa distribuição aos dados regionais deve se basear necessariamente em momentos-L de ordem superior; Hosking e Wallis (1997) consideram suficientes a Assimetria-L e a Curtose-L. Logo, pode-se julgar a aderência pelo grau com que uma certa distribuição aproxima as médias regionais de Assimetria-L e Curtose-L. Por exemplo, suponha que a distribuição candidata é a Generalizada de Valores Extremos (GEV) de três parâmetros. Quando ajustada aos dados da região pelo método dos momentos-L, essa distribuição irá reproduzir a média regional de Assimetria-L. Portanto, pode-se julgar o grau de ajuste pela diferença entre a Curtose-L τ_4^{GEV} da distribuição e a média regional correspondente t_4^R, tal como esquematizado na Figura 10.9. Contudo, essa diferença deve levar em conta a variabilidade amostral de t_4^R. Essa pode ser quantificada através de σ_4, ou seja o desvio-padrão de t_4^R, o qual é obtido por simulação de um grande número de regiões homogêneas, todas extraídas de uma população de valores distribuídos conforme uma GEV, contendo os mesmos indivíduos e tamanhos de amostras dos dados observados. Nesse caso, portanto, a *medida de aderência* da distribuição GEV pode ser calculada como

$$Z^{GEV} = \frac{\left(t_4^R - \tau_4^{GEV}\right)}{\sigma_4} \tag{10.39}$$

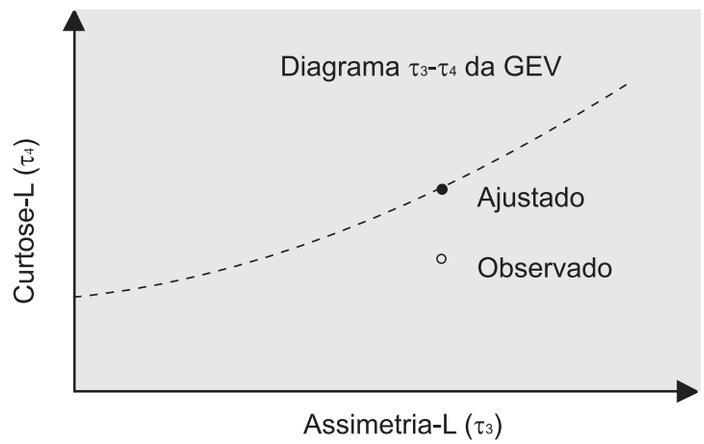


Figura 10.9 – Descrição esquemática da medida de aderência Z

Hosking e Wallis (1997) reportam as seguintes dificuldades relacionadas ao procedimento de cálculo da medida de aderência tal como anteriormente descrito:

• Para obter os valores corretos de σ_4, é necessário um conjunto de simulações específico para cada distribuição candidata. Entretanto, na prática, Hosking e Wallis (1997) consideram que é suficiente supor que σ_4 tem o mesmo valor para todas as distribuições candidatas de três parâmetros. Justificam afirmando que, como todas as distribuições ajustadas têm a mesma Assimetria-L, é razoável supor que elas também se assemelham com relação a outras características. Assim sendo, também é razoável supor que uma distribuição Kapa de quatro parâmetros, ajustada aos dados regionais, terá um valor de σ_4 próximo ao das distribuições candidatas. Portanto, σ_4 pode ser obtido a partir da simulação de um grande número de regiões homogêneas extraídas de uma população Kapa. Para esse objetivo, podem ser empregadas as mesmas simulações usadas no cálculo da medida de heterogeneidade, conforme descrito no item 10.3.2.1.2.

• As estatísticas aqui mencionadas pressupõem a inexistência de qualquer viés no cálculo dos momentos-L amostrais. Hosking e Wallis (1997) observam que essa suposição é válida para t_3 mas não o é para t_4, sob as condições de amostras de pequeno tamanho ($n_i \leq 20$) ou de populações de grande assimetria ($\tau_3 \geq 0{,}4$). A solução desse problema é feita por uma correção de viés para t_4. Essa correção, denotada por B_4, pode ser calculada através dos mesmos resultados de simulação usados para se calcular σ_4.

• A medida de aderência *Z* refere-se a distribuições candidatas de três parâmetros. Embora seja possível construir procedimento semelhante para as distribuições de dois parâmetros, elas possuem valores populacionais fixos de τ_3 e τ_4 e, em conseqüência, tornam problemática a estimação de σ_4. Apesar de haverem sugerido algumas adaptações plausíveis, Hosking e Wallis (1997) desaconselham o uso da medida de aderência para distribuições de apenas dois parâmetros.

10.3.3.2.2 – Definição Formal

Considere que uma dada região contenha *N* postos de observação, cada um deles indexado por *j*, com amostra de tamanho n_j e quocientes de momentos-L amostrais representados por $t^j, t_3^j \; e \; t_4^j$. Considere também que t^R, t_3^R e t_4^R denotam respectivamente as médias regionais dos quocientes CV-L, Assimetria-L e Curtose-L, ponderados, de forma análoga à especificada pela equação 10.15, pelos tamanhos das amostras individuais.

Considere também um conjunto de distribuições candidatas de três parâmetros. Hosking e Wallis (1997) propõem o seguinte conjunto de distribuições candidatas: Logística Generalizada - LG, Generalizada de Valores Extremos - GEV, Generalizada de Pareto - GP, Lognormal – LN3 e Pearson do tipo III - P3. Em seguida, cada distribuição candidata deve ter seus parâmetros ajustados ao grupo de quocientes de momentos-L regionais $\{1, t^R, t_3^R, t_4^R\}$. Denota-se por τ_4^{DIST} a Curtose-L da distribuição ajustada, onde *DIST* poderá ser qualquer uma das distribuições (e.g. LG, GEV, LN3).

Na seqüência, deve-se ajustar a distribuição Kapa ao grupo de quocientes de momentos-L regionais e proceder à simulação de um grande número, N_{SIM} de regiões homogêneas, cada qual tendo a Kapa como distribuição de freqüência. Essa simulação deverá ser efetuada exatamente da mesma forma como a apresentada para o cálculo da medida de heterogeneidade (ver item 10.3.2.1). Em seguida, calculam-se as médias regionais t_3^m e t_4^m da Assimetria-L e Curtose-L da $m^{\text{ésima}}$ região simulada. O viés de t_4^R é dado por

$$B_4 = \frac{\sum_{m=1}^{N_{SIM}} \left(t_4^m - t_4^R \right)}{N_{SIM}} \tag{10.40}$$

enquanto o desvio padrão de t_4^R o é pela expressão

$$\sigma_4 = \sqrt{\frac{\sum_{m=1}^{N_{SIM}} \left(t_4^m - t_4^R \right)^2 - N_{SIM} B_4^2}{N_{SIM} - 1}} \tag{10.41}$$

A medida de aderência *Z* de cada distribuição candidata, pode ser calculada pela equação

$$Z^{DIST} = \frac{\tau_4^{DIST} - t_4^R + B_4}{\sigma_4} \tag{10.42}$$

A hipótese de um ajuste adequado é mais verdadeira quanto mais próxima de zero for a medida de aderência. Nesse contexto, Hosking e Wallis (1997) sugerem como critério razoável o limite $\left| Z^{DIST} \right| \leq 1{,}64$.

10.3.3.2.3 – Discussão

A estatística *Z* é especificada sob a forma de um teste de significância. Segundo

Hosking e Wallis (1997), Z possui distribuição que se aproxima da Normal padrão, sob as premissas de que a região é perfeitamente homogênea e de que não há correlação cruzada entre os seus indivíduos. Se a distribuição de Z é de fato a Normal, o critério $|Z^{DIST}| \leq 1{,}64$ corresponde à aceitação da hipótese de que os dados provêm da distribuição candidata, com um nível de significância de 10%. Entretanto, as premissas, necessárias para aproximar a distribuição de Z pela Normal padrão, dificilmente são completamente satisfeitas na prática. Assim sendo, o critério $|Z^{DIST}| \leq 1{,}64$ é simplesmente um indicador de boa aderência e não uma estatística de teste formal. Hosking e Wallis (1997) relatam que o critério $|Z^{DIST}| \leq 1{,}64$ é particularmente inconsistente se os dados apresentarem correlação serial e/ou correlação cruzada. Tanto uma quanto a outra tendem a fazer aumentar a variabilidade de t_4^R. Nesse caso, como não há correlação para as regiões simuladas de população Kapa, a estimativa de σ_4 resulta ser excessivamente pequena e a estatística Z excessivamente grande, conduzindo a uma falsa indicação de falta de aderência.

Se, ao se aplicar o teste da medida de aderência a uma região homogênea, resultar que várias distribuições são consideradas candidatas, Hosking e Wallis (1997) recomendam o exame das curvas de quantis adimensionais. Se essas fornecerem resultados aproximadamente iguais, qualquer uma das distribuições candidatas pode ser selecionada. Entretanto, se os resultados diferem significativamente, a escolha deve tender para o modelo probabilístico que apresentar maior *robustez*. Nesses casos, ao invés de um modelo probabilístico de três parâmetros, recomenda-se a seleção da distribuição Kapa de quatro parâmetros ou da Wakeby de cinco parâmetros, as quais são mais robustas à incorreta especificação da curva regional de freqüência. A mesma recomendação se aplica aos casos em que nenhuma das distribuições de três parâmetros atendeu ao critério $|Z^{DIST}| \leq 1{,}64$ ou aos casos de regiões "possivelmente heterogêneas" ou "definitivamente heterogêneas".

Além da verificação da medida de aderência Z, recomenda-se grafar as médias regionais da Assimetria-L e Curtose-L $\{t_3^R, t_4^R\}$ em um diagrama de quocientes de momentos-L, tal como o da Figura 10.10. Hosking e Wallis (1993) sugerem que, se o ponto $\{t_3^R, t_4^R\}$ se localizar acima da curva da distribuição Logística Generalizada, nenhuma distribuição de dois ou três parâmetros se ajustará aos dados, devendo possivelmente se adotar uma distribuição Kapa de quatro parâmetros ou Wakeby de cinco parâmetros. Finalmente, ao se analisar uma grande área geográfica, sujeita à divisão em várias regiões homogêneas, a especificação da distribuição de freqüência de uma região pode afetar a das outras. Se uma determinada distribuição se ajusta bem aos dados da maioria das regiões, é de bom senso utilizá-la para todas, muito embora ela possa não ser a distribuição que particularmente melhor se ajusta aos dados de uma ou de algumas das regiões.

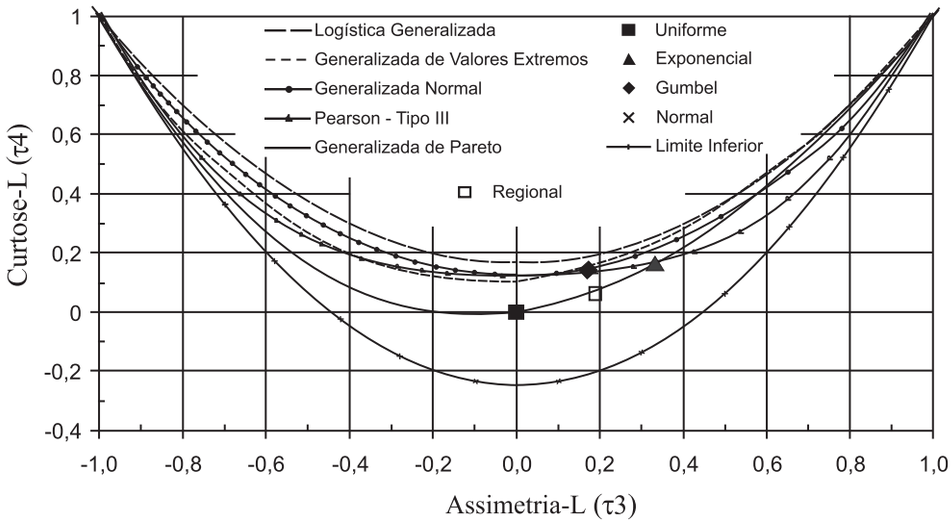


Figura 10.10 – Diagrama assimetria-L x Curtose-L

10.3.4 – Estimação da Distribuição Regional de Freqüência

10.3.4.1 – Justificativas

Depois que os dados dos diferentes postos da área em estudo foram submetidos às etapas descritas nos itens anteriores, tem-se como resultado a partição da área em regiões aproximadamente homogêneas, nas quais as distribuições de freqüência de seus indivíduos são idênticas, a menos de um fator de escala local, e podem ser modeladas por uma única distribuição de probabilidades regional, selecionada entre diversas funções candidatas. Essa relação entre as distribuições de freqüência dos diversos locais representa a própria justificativa para a análise regional de freqüência, permitindo a obtenção de melhores estimativas de parâmetros e quantis a partir da combinação de dados espacialmente disseminados.

Diversos métodos podem ser utilizados para se ajustar uma distribuição de probabilidades aos dados de uma região homogênea. Para descrevê-los, considere, inicialmente, uma certa variável aleatória X, cuja variabilidade foi amostrada em N locais ou postos de observação, situados em uma região homogênea. As observações, tomadas nos postos indexados por j, formam amostras de tamanho variável n_j e são denotadas por $X_{i,j}$, $i = 1,\ldots,n_j$; $j = 1,\ldots,N$. Se F, $0 < F < 1$, representa a distribuição de freqüências da variável X no posto j, então, a função de *quantis* nesse local é simbolizada por $X_j(F)$. Por definição, em uma região

homogênea, as distribuições de freqüências nos N pontos são idênticas, à exceção de um *fator de escala local* μ_j, o *index-flood*, ou seja,

$$X_j(F) = \mu_j \; x(F), \quad j = 1, \ldots, N \tag{10.43}$$

Se $\hat{\mu}_j$ denota a estimativa do fator de escala no local j, pode-se representar os dados adimensionais padronizados por $x_{i,j} = X_{i,j} / \hat{\mu}_j, i = 1, \ldots, n_j; j = 1, \ldots, N$. O método mais simples e antigo para se combinar os dados locais, com o objetivo de se estimar os parâmetros e quantis da distribuição regional, é conhecido como o da *estação-ano*. Esse método simplesmente agrupa todos os dados adimensionais padronizados em única amostra, considerada aleatória simples, a qual é em seguida usada para se ajustar a distribuição regional. Hosking e Wallis (1997) consideram que, na atualidade, esse método é raramente empregado principalmente porque não é correto tratar os dados adimensionais padronizados como uma *amostra aleatória simples*, ou seja uma realização de variáveis aleatórias independentes e igualmente distribuídas. De fato, como os fatores de escala locais $\hat{\mu}_j$ são, em geral, estimativas obtidas a partir de amostras de diferentes tamanhos, os dados adimensionais padronizados dos diversos postos considerados não serão igualmente distribuídos.

Em outro extremo, encontra-se o *método de estimação através do máximo da função de verossimilhança*, tal como aplicado aos N fatores de escala locais μ_j e aos p parâmetros de $x(F; \theta_1, ..., \theta_p)$, contidos na equação 10.43. O modelo estatístico procura encontrar, em geral de forma iterativa, as $N+p$ soluções de um sistema de $N+p$ equações que visam maximizar a função de verossimilhança [ver, por exemplo, Buishand (1989)]. Esse método pode ser usado também para situações em que os fatores de escala são considerados parâmetros dependentes de informações covariadas, ou seja, $\mu_j = h(z_j, \omega)$, onde z_j representa um vetor de características ou informações covariadas no local j, h uma função matemática convenientemente escolhida e ω um vetor de parâmetros a serem estimados. Exemplos de utilização dessa abordagem podem ser encontrados nos trabalhos de Smith (1989) e de Naghettini et al. (1996).

O método *index-flood* utiliza as estatísticas características dos dados locais para obter as estimativas regionais, ponderando-as através da equação

$$\hat{\lambda}_k^R = \frac{\sum_{j=1}^{N} n_j \hat{\lambda}_k^{(j)}}{\sum_{j=1}^{N} n_j} \tag{10.44}$$

onde $\hat{\lambda}_k^R$ denota a estimativa regional e $\hat{\lambda}_k^{(j)}, k = 1, \ldots, p$ representam as estatísticas locais. Se essas têm como base os quocientes de momentos-L, Hosking e Wallis (1997) definem a metodologia de estimação como a do *algoritmo dos momentos-L regionais*. Apesar de reconhecerem não haver nenhuma superioridade teórica da metodologia proposta, em relação à do máximo de verossimilhança, justificam o seu emprego pela maior simplicidade de cálculo. O algoritmo dos momentos-L regionais será descrito nos itens que se seguem, tomando como premissa a inexistência de correlação cruzada entre as observações dos diferentes indivíduos de uma região homogênea ou de correlação serial entre as observações de um dado local.

10.3.4.2 – O Algoritmo dos Momentos-L Regionais

10.3.4.2.1 – Descrição

O objetivo é o de ajustar uma única distribuição de freqüência aos dados adimensionais padronizados, observados em diferentes locais de uma região considerada aproximadamente homogênea. O ajuste é efetuado através do *método dos momentos-L*, o qual consiste em igualar os momentos-L populacionais da distribuição em questão aos respectivos momentos-L amostrais. De forma mais conveniente, os *quocientes* de momentos-L locais são ponderados pelos seus respectivos tamanhos de amostra, de forma a produzir as estimativas regionais dos *quocientes* de momentos-L, as quais são, em seguida, empregadas para a inferência estatística. Se o *index-flood* é representado pela média da distribuição local de freqüências, cuja estimativa é dada pela média amostral dos dados individuais, então a média dos dados adimensionais padronizados, bem como da ponderação regional, é 1. Isso faz com que os quocientes de momentos-L amostrais t e t_r, para $r \geq 3$, sejam os mesmos, não importando se foram calculados a partir dos dados originais $\{X_{i,j}\}$ ou pelos dados adimensionais padronizados $\{x_{i,j}\}$.

10.3.4.2.2 – Definição Formal

Considere que uma dada região contenha N postos de observação, cada um deles indexado por j, com amostra de tamanho n_j e quocientes de momentos-L amostrais representados por $t^j, t_3^j, t_4^j, \ldots$. Considere também que $t^R, t_3^R, t_4^R, \ldots$ denotam as médias regionais dos quocientes de momentos-L ponderados, de forma análoga à especificada pela equação 10.44, pelos tamanhos das amostras individuais. Conforme justificativa anterior, a média regional é 1, ou seja $\ell_1^R = 1$.

Efetua-se o ajuste da distribuição regional, igualando-se os seus quocientes de momentos-L populacionais $\lambda_1, \tau, \tau_3, \tau_4, \ldots$ às médias regionais $1, t^R, t_3^R, t_4^R, \ldots$. Se a distribuição F a ser ajustada, é definida por p parâmetros θ_k, $k = 1, \ldots, p$, resultará um sistema de p equações e p incógnitas, cujas soluções serão as estimativas $\hat{\theta}_k, k = 1, \ldots, p$. Com essas, pode-se obter a estimativa da curva regional de quantis adimensionais $\hat{x}(F) = x\left(F; \hat{\theta}_1, \ldots, \hat{\theta}_p\right)$. Inversamente, as estimativas dos quantis para o posto j são obtidas pelo produto de $\hat{x}(F)$ por $\hat{\mu}_j$, ou seja

$$\hat{X}_j(F) = \ell_1^j \, \hat{x}(F) \tag{10.45}$$

10.3.4.2.3 – Momentos-L Amostrais

A estimação dos MPP's e momentos-L, a partir de uma amostra finita de tamanho n, inicia-se com a ordenação de seus elementos constituintes em ordem crescente, ou seja $x_{1:n} \leq x_{2:n} \leq \cdots \leq x_{n:n}$. Um estimador não-enviesado do MPP β_r pode ser escrito como

$$b_r = \hat{\beta}_r = \frac{1}{n} \sum_{j=r+1}^{n} \frac{(j-1)(j-2)\ldots(j-r)}{(n-1)(n-2)\ldots(n-r)} x_{j:n} \tag{10.46}$$

Dessa forma, os estimadores de $\beta_r, r \leq 2$, são dados por

$$b_0 = \frac{1}{n} \sum_{j=1}^{n} x_{j:n} \tag{10.47}$$

$$b_1 = \frac{1}{n} \sum_{j=2}^{n} \frac{(j-1)}{(n-1)} x_{j:n} \tag{10.48}$$

$$b_2 = \frac{1}{n} \sum_{j=3}^{n} \frac{(j-1)(j-2)}{(n-1)(n-2)} x_{j:n} \tag{10.49}$$

Outra forma de se estimar o MPP β_r é por meio do uso de estimadores por posição de plotagem, os quais foram introduzidos por Landewehr et al. (1979) e podem ser escritos como

$$\hat{\beta}_r = n^{-1} \sum_{i=1}^{n} \left(p_{j:n}\right)^r x_{j:n} \tag{10.50}$$

As estimativas das posições de plotagem são realizadas com a equação $p_{j:n} = (j + \gamma)/(n + \delta)$, onde $\delta > \gamma > -1$. Em particular, adota-se

$p_{j:n} = (j - 0{,}35)/n$ em estudos que utilizam as distribuições de Wakeby, a Generalizada de Eventos Extremos e a Generalizada de Pareto.

Sendo assim, os estimadores por posição de plotagem de β_r, $r \leq 2$, são dados por

$$b_0 = \frac{1}{n}\sum_{j=1}^{n} x_{j:n} \tag{10.51}$$

$$b_1 = \frac{1}{n}\sum_{j=1}^{n}\left(\frac{j-0{,}35}{n}\right)x_{j:n} \tag{10.52}$$

$$b_2 = \frac{1}{n}\sum_{j=1}^{n}\left(\frac{j-0{,}35}{n}\right)^2 x_{j:n} \tag{10.53}$$

Segundo Hosking (1995), os estimadores por posição de plotagem dos momentos-L e razões-L apresentam algumas desvantagens, quando comparado aos estimadores não enviesados. Para uso geral, devem ser utilizados os estimadores não enviesados. Os estimadores por posição de plotagem podem ser adequados para a estimativa dos quantis extremos da cauda superior nas análises de freqüência regional.

Os estimadores não-enviesados de λ_r são os momentos-L amostrais, esses definidos pelas seguintes expressões :

$$\ell_1 = b_0 \tag{10.54}$$

$$\ell_2 = 2b_1 - b_0 \tag{10.55}$$

$$\ell_3 = 6b_2 - 6b_1 + b_0 \tag{10.56}$$

$$\ell_4 = 20b_3 - 30b_2 + 12b_1 - b_0 \tag{10.57}$$

$$\ell_{r+1} = \sum_{k=0}^{r} l_{r,k}^{*} b_k \,; \quad r = 0,1,\cdots,n-1 \tag{10.58}$$

Na equação 10.58, os coeficientes $l_{r,k}^{*}$ são definidos por

$$l_{r,k}^{*} = (-1)^{r-k}\binom{r}{k}\binom{r+k}{k} = \frac{(-1)^{r-k}(r+k)!}{(k!)^2(r-k)!} \tag{10.59}$$

Da mesma forma, os quocientes de momentos-L amostrais são dados por

$$t_r = \frac{\ell_r}{\ell_2}; \quad r \geq 3 \tag{10.60}$$

enquanto o CV-L amostral calcula-se através de

$$t = \frac{\ell_2}{\ell_1} \tag{10.61}$$

Os estimadores de τ_r, fornecidos pelas equações 10.60 e 10.61, são muito pouco enviesados quando calculados para amostras de tamanho moderado a grande. *Hosking* (1990, p. 116) utilizou a teoria assintótica para calcular o viés para amostras grandes. Para a distribuição Gumbel, por exemplo, o viés assintótico de t_3 é $0{,}19n^{-1}$, enquanto o de t_4, para a distribuição Normal, é $0{,}03n^{-1}$, onde n representa o tamanho da amostra. Para amostras de pequeno tamanho, o viés pode ser avaliado por simulação. Segundo *Hosking & Wallis* (1997, p. 28) e para uma gama variada de distribuições, o viés de t pode ser considerado desprezível para $n \geq 20$. Ainda segundo esses autores, mesmo em se tratando de amostras de tamanho em torno de 20, o viés de t_3 e o viés de t_4 são considerados relativamente pequenos e definitivamente menores do que os produzidos por estimadores convencionais de assimetria e curtose.

10.3.4.2.4 – Discussão

Os resultados obtidos por qualquer análise estatística possuem uma incerteza inerente, a qual pode ser avaliada por métodos tradicionais como, por exemplo, a construção de intervalos de confiança para as estimativas de parâmetros e quantis. Em geral, a construção de intervalos de confiança pressupõe que todas as premissas do modelo estatístico empregado sejam satisfeitas, o que, em termos da análise regional de freqüência, equivale a dizer que as seguintes hipóteses tenham que ser rigorosamente verdadeiras : (a) a região é exatamente homogênea, (b) o modelo probabilístico foi especificado com exatidão e (c) não há correlação cruzada ou serial entre as observações. Por essa razão, Hosking e Wallis (1997) consideram que, no contexto da análise regional de freqüência, a construção de intervalos de confiança para parâmetros e quantis é de utilidade limitada. Como alternativa, propõem uma abordagem de avaliação da precisão das estimativas de quantis, com base em simulação de Monte Carlo, na qual leva-se em consideração a possibilidade de heterogeneidade regional e existência de correlação cruzada e/ou serial, bem como da incorreta especificação do modelo probabilístico regional. A descrição do experimento de Monte Carlo encontra-se fora do escopo do presente capítulo. A seguir, estão transcritas as principais conclusões do estudo levado a termo por Hosking e Wallis (1997).

• Mesmo em regiões com grau moderado de heterogeneidade, presença de correlação cruzada e incorreta especificação do modelo probabilístico regional, os resultados da análise regional de freqüência são mais confiáveis do que os obtidos pela análise local.
• A regionalização é particularmente útil para a estimação de quantis muito altos ou baixos, respectivamente, das caudas superior e inferior das distribuições de freqüência.
• Em se tratando de regiões heterogêneas com um grande número de postos (N), os erros das estimativas de quantis e da curva regional de quantis adimensionalizados decrescem lentamente em função de N. Como conclusão, pode-se afirmar que, em geral, o ganho em precisão é pequeno em regiões com mais de 20 postos.
• As amostras maiores fazem com que a análise regional de freqüência seja de menor utilidade, relativamente à análise local. Entretanto, as amostras maiores facilitam a identificação de heterogeneidade regional. Como conclusão, pode-se afirmar que, em geral, quando os tamanhos das amostras são grandes, as regiões devem conter poucos postos.
• Não se recomenda o uso de distribuições de dois parâmetros para a análise regional de freqüência. Preconiza-se o seu emprego somente se o analista está completamente seguro de que a Assimetria-L e a Curtose-L da distribuição são precisamente reproduzidas pelas estimativas amostrais. Caso contrário, as estimativas de quantis estarão fortemente enviesadas.
• Os erros provenientes da incorreta especificação da distribuição de freqüência são importantes somente para quantis muito altos ou baixos, respectivamente, das caudas superior e inferior. Por exemplo, para a cauda superior, ocorrem erros significativos somente para $F > 0{,}99$.
• Certas distribuições robustas, como a Kapa e Wakeby, produzem estimativas de quantis razoavelmente precisas para uma ampla variedade de distribuições locais.
• A heterogeneidade regional introduz um viés nas estimativas de quantis dos postos considerados atípicos, em relação à região como um todo.
• A dependência estatística entre os postos aumenta a variabilidade das estimativas de quantis, mas tem pouca influência sobre o viés. Um pequeno grau de correlação cruzada não invalida os resultados da estimação regional.
• Para quantis extremos ($F \geq 0{,}999$), a vantagem da análise regional sobre a local é muito maior. Para quantis dessa ordem de grandeza, a heterogeneidade é menos importante como fonte de erros, ao passo que a incorreta especificação do modelo probabilístico é mais significativa.

Exemplo 10.4 – No Anexo 12, estão apresentadas as vazões médias diárias máximas anuais de 07 estações da bacia do rio Paraopeba, localizadas no mapa da Figura 10.2 e listadas na Tabela 10.4 do exemplo 10.2. Pede-se realizar um estudo de regionalização das vazões máximas anuais aplicando o método *index-flood*, ou da cheia-índice, com momentos-L.
Solução: Para resolver esse exemplo, foram utilizadas as rotinas em linguagem Fortran-77, desenvolvidas por J. R. M. Hosking, e disponibilizadas para *download* nos endereços http://lib.stat.cmu.edu/general/lmoments e http://www.research.ibm.com/people/h/hosking/lmoments.html#papers1.

A primeira etapa dessa metodologia se refere à análise regional de consistência de dados que se baseia nas técnicas usuais de análise de consistência e no uso da estatística auxiliar de medida de discordância (D_j), descrita no item 10.3.1.1. A análise de consistência desses dados está descrita em Pinto e Alves (2001) e os valores de medida de discordância, considerando que as 7 estações formam uma região homogênea, estão apresentados na Tabela 10.11. Os resultados da medida de discordância mostram que as amostras não apresentam características estatísticas muito discrepantes das grupais, uma vez que os valores de D_j são inferiores a 1,917.

Tabela 10.11 – Medidas de discordância

Estação	40549998	40573000	40577000	40579995	40665000	40710000	40740000
Medida de discordância	0,59	0,64	1,45	1,67	0,75	0,8	1,11

Na segunda etapa, é realizada a identificação de regiões homogêneas. Nos exemplos 10.2 e 10.3, a identificação de uma única região homogênea foi realizada considerando as características físicas, as estatísticas locais e o comportamento, em papel de probabilidades, das curvas de freqüência empíricas das séries adimensionalizadas. As rotinas permitem o cálculo da medida de heterogeneidade (H), descrita no item 10.3.2.1, para verificar a hipótese de homogeneidade da região anteriormente definida. De acordo com o teste de significância, proposto por Hosking e Wallis (1997), a região pode ser considerada como "aceitavelmente homogênea", pois a medida de heterogeneidade calculada é igual a -0,42, ou seja, de valor absoluto inferior a 1.

A seleção da função regional de distribuição de probabilidades corresponde à terceira etapa da metodologia. Novamente, a seleção foi efetuada com as rotinas Fortran-77 já mencionadas. Essas fazem o ajuste das seguintes distribuições de três parâmetros: Logística Generalizada (LG), Generalizada de Valores Extremos (GEV), Log-Normal (LN-3P) ou Generalizada Normal,

Pearson tipo III (P-III) e Generalizada de Pareto (GP), estimando os seus parâmetros a partir dos momentos-L regionais, além de aplicar o teste de aderência, detalhado no item 10.3.3.2, para verificar o ajuste entre a distribuição candidata e os dados regionais. Os resultados do teste de aderência, apresentados na Tabela 10.12, demonstram que as distribuições Generalizada de Valores Extremos (GEV), Log-Normal (LN-3P) ou Generalizada Normal e a Pearson tipo III (P-III) podem ser adotadas na região.

Tabela 10.12 – Resultados dos testes de aderência (Z)

	Distribuições				
	LG	GEV	LN-3P	P-III	GP
Região	1,69	0,44*	0,21*	-0,31*	-2,36

* $Z \leq 1{,}64$

Além dos resultados do teste de aderência, a definição das distribuições regionais pôde ser corroborada pelo posicionamento dos valores regionais no diagrama Assimetria-L x Curtose-L. Os momentos ponderados por probabilidade, β_r, de cada estação, foram calculados utilizando os estimadores não enviesados, equações 10.47, 10.48 e 10.49. As estimativas dos MPP β_r permitiram o cálculo dos momentos-L, equações 10.54, 10.55, 10.56 e 10.57, e das razões-L, equações 10.60 e 10.61. As razões-L das estações e as regionais encontram-se na Tabela 10.13, enquanto a Figura 10.11 apresenta o diagrama Assimetria-L x Curtose-L.

Tabela 10.13 – Valores das Razões-L e dos Momentos-L

Estações	l_1	CV-L (t_2)	Assimetria L (t_3)	Curtose L (t_4)
40549998	1	0,2147	0,268	0,1297
40573000	1	0,1952	0,1389	-0,0006
40577000	1	0,1823	0,0134	0,0222
40579995	1	0,2489	0,1752	0,1479
40665000	1	0,1926	0,2268	0,0843
40710000	1	0,2284	0,1414	0,2304
40740000	1	0,2352	0,2706	0,3001
Valores Regionais	1	0,2194	0,1882	0,1433

Analisando os resultados dos testes de aderência e o diagrama Assimetria-L x Curtose-L, as seguintes três distribuições podem ser adotadas como modelos distributivos regionais: a Generalizada de Valores Extremos (GEV), a Log-Normal (LN-3P) ou a Generalizada Normal e a Pearson tipo III (P-III).

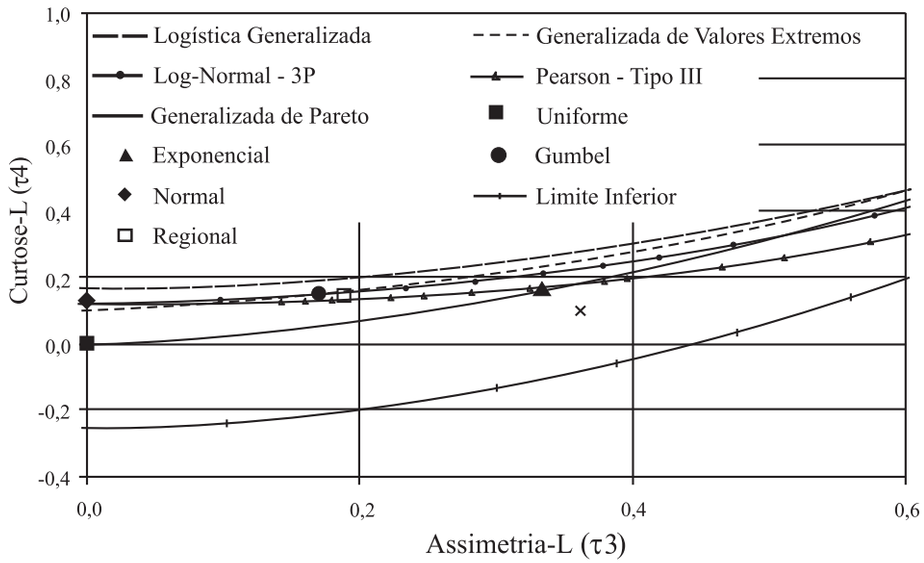


Figura 10.11 – Diagrama assimetria-L x Curtose-L, exemplo 10.4

As funções densidade de probabilidade e de distribuição acumulada da Log-Normal (3P), da GEV e da Pearson tipo III (P-III), além de suas respectivas funções inversas, tal como parametrizadas por Hosking (1997), encontram-se transcritas a seguir.

• Log-Normal (3P)

$$f(x)=\frac{e^{ky-y^2/2}}{\alpha\sqrt{2\pi}} \qquad y=\begin{cases}-k^{-1}\,ln\{1-k(x-\xi)/\alpha\} & k\neq 0\\ (x-\xi)/\alpha & k=0\end{cases} \qquad (10.62)$$

$$F(x)=\Phi(y)$$

na qual, Φ denota a função de distribuição acumulada Normal padrão.
$x(F)$ não apresenta forma analítica explícita.
Parâmetros : ξ (Posição), α (Escala) e k (forma)
Os limites da função são:
Para, $k > 0$: $-\infty < x \le \xi + \alpha/k$; $k = 0$: $-\infty < x < \infty$; $k < 0$: $\xi + \alpha/k \le x < \infty$
Os parâmetros podem ser estimados pelas seguintes equações:

$$k \approx -\tau_3 \frac{E_0 + E_1\tau_3^2 + E_2\tau_3^4 + E_3\tau_3^6}{1 + F_1\tau_3^2 + F_2\tau_3^4 + F_3\tau_3^6} \quad \text{para } |\tau_3| \le 0{,}94 \qquad (10.63)$$

E_0	E_1	E_2	E_3	F_1	F_2	F_3
2,0466534	-3,6544371	1,8396733	-0,20360244	-2,0182173	1,2420401	-0,21741801

$$\alpha = \frac{\lambda_2 k e^{-\frac{k^2}{2}}}{1 - 2\Phi\left(-k/\sqrt{2}\right)} \tag{10.64}$$

$$\xi = \lambda_1 - \frac{\alpha}{k}\left(1 - e^{\frac{k^2}{2}}\right) \tag{10.65}$$

Nesta parametrização, a distribuição Log-Normal é a distribuição de uma variável aleatória X que está relacionada a uma variável aleatória Z de distribuição Normal padrão, pela seguinte equação:

$$X = \begin{cases} \xi + \alpha(1 - e^{-kZ})/k & k \neq 0 \\ \xi + \alpha Z & k = 0 \end{cases} \tag{10.66}$$

Z é variável normal central reduzida cujos valores podem ser obtidos nas Tabela 5.1 e 8.1 ou aproximados pelas equações 8.11 e 8.12.

• GEV

$$f_X(x) = \frac{1}{\alpha} exp\left[-(1-k)y - exp(-y)\right] \tag{10.67}$$

Para $k = 0$, $\quad y = \dfrac{x - \xi}{\alpha}$

Para $k \neq 0$ $\quad y = -\dfrac{1}{k} ln\left[1 - \dfrac{(x-\xi)k}{\alpha}\right]$.

Os limites da função são:

Para $k < 0$: $\xi + \dfrac{\alpha}{k} \leq x \leq \infty$, para $k = 0$: $-\infty \leq x \leq \infty$ e para $k > 0$: $-\infty < x \leq \xi + \dfrac{\alpha}{k}$

$$F_X(x) = \exp[-\exp(-y)] \tag{10.68}$$

$$x(F) = \xi - \alpha \ln[-\ln(F)], k = 0 \tag{10.69}$$

$$x(F) = \xi + \frac{\alpha}{k}\left\{1 - \left[-ln(F)\right]^k\right\};\ k \neq 0 \qquad (10.70)$$

Onde k, α e ξ são os parâmetros de forma, escala e posição, respectivamente. A estimação dos parâmetros pelos momentos-L pode ser efetuada por meio das seguintes equações:

$$\hat{k} \approx 7{,}8590c + 2{,}9554c^2, \text{ para } -0{,}5 \leq \tau_3 \leq 0{,}5 \qquad (10.71)$$

Sendo

$$c = \frac{2}{3+\tau_3} - \frac{ln(2)}{ln(3)} = \frac{2\lambda_2}{\lambda_3 + 3\lambda_2} - \frac{ln(2)}{ln(3)} = \frac{(2\beta_1 - \beta_0)}{(3\beta_2 - \beta_0)} - \frac{ln(2)}{ln(3)} \qquad (10.72)$$

$$\hat{\alpha} = \frac{\hat{k}\lambda_2}{\left(1 - 2^{-k}\right)\Gamma\left(1+\hat{k}\right)} \qquad (10.73)$$

$$\hat{\xi} = \lambda_1 - \frac{\hat{\alpha}}{\hat{k}}\left[1 - \Gamma\left(1+\hat{k}\right)\right] \qquad (10.74)$$

• A distribuição Pearson Tipo III, com parâmetros de posição, escala e forma, foi detalhada no exemplo 8.6 do capítulo 8.

Os momentos-L e as razões-L regionais, apresentados na Tabela 10.13, foram utilizados para estimar os parâmetros das três distribuições. Os parâmetros da GEV foram estimados com as equações 10.72, 10.73 e 10.74; os da Log-Normal (3P) com as equações 10.63, 10.64 e 10.65; e os da Pearson III com as equações apresentadas no exemplo 8.6 do capítulo 8. As estimativas dos parâmetros encontram-se na Tabela 10.14.

Tabela 10.14 – Parâmetros das distribuiçoes regionais

Distribuição	Posição	Escala	Forma
Generalizada de Eventos Extremos – GEV	0,813	0,308	-0,028
Log-Normal (3P) – LN-3P	0,926	0,365	-0,388
Pearson Tipo III – PIII	1	0,405	1,14

Após a estimação dos parâmetros das 3 distribuições, foram calculados os quantis regionais adimensionalizados associados a vários períodos de retorno, de acordo com a equação 10.70 para a GEV, da equação 10.66 para a LN-3P e da equação 8.41 para a PIII. Os resultados obtidos estão apresentados na Tabela 10.15.

Tabela 10.15 – Quantis regionais adimensionais

	Tempo de retorno (anos)					
Distribuição	1,01	2,00	10	20	100	1000
Generalizada de Eventos Extremos	0,353	0,927	1,529	1,768	2,327	3,163
Log-Normal (3P)	0,367	0,926	1,533	1,767	2,307	3,108
Pearson Tipo III	0,397	0,925	1,543	1,769	2,260	2,915

A quarta etapa refere-se à estimação de parâmetros e quantis da função regional de distribuição de probabilidades. A distribuição regional adotada é a Generalizada de Eventos Extremos (GEV), uma vez que os quantis adimensionais são um pouco maiores à medida que o tempo de retorno aumenta. Os parâmetros e os quantis constam das Tabelas 10.14 e 10.15, respectivamente. A estimação dos quantis adimensionais regionais associados a diferentes tempos de retorno pode ser efetuada por meio da equação:

$$x(F) = \xi + \frac{\alpha}{k}\left\{1 - \left[-ln(F)\right]^{k}\right\} = 0{,}813 - \frac{0{,}308}{0{,}028}\left\{1 - \left[-ln(F)\right]^{-0{,}028}\right\} \quad (10.75)$$

No método *index-flood*, o cálculo de quantis, associados a diferentes tempos de retorno, é realizado por meio da equação 10.43. Assim, em locais situados na região homogênea e que não sejam monitorados sistematicamente, é necessário estimar o fator de adimensionalização. A última etapa, portanto, corresponde à regressão entre os fatores de adimensionalização, nesse caso as médias amostrais, $Q_{med\text{-}max}$, apresentadas na Tabela 10.5, e as correspondentes características das bacias, da Tabela 10.4. Como foram utilizados os mesmos dados do exemplo 10.3, adota-se, aqui, a mesma equação de regressão estabelecida naquele exemplo, a saber,

$$Q_{med-max} = 0{,}1098\,A^{1{,}0125}\left(km^2\right) \quad (10.76)$$

válida para $244km^2 \le A\left(km^2\right) \le 3940km^2$

Substituindo as equações 10.75 e 10.76 na equação 10.43, é possível estimar diretamente os quantis associados a diferentes tempos de retorno, para locais não monitorados sistematicamente e situados dentro da região homogênea, por meio da seguinte equação:

$$Q_{max}(T) = \left\{0{,}1098\,A^{1{,}0125}\right\}\left\{0{,}813 - \frac{0{,}308}{0{,}028}\left\{1 - \left[-ln\left(1 - \frac{1}{T}\right)\right]^{-0{,}028}\right\}\right\} \quad (10.77)$$

De volta ao exemplo hipotético, para uma área de drenagem de 450 km², dentro da região homogênea, a vazão média máxima diária com 100 anos de tempo de retorno será:

$$Q_{max}(100) = \left\{0{,}1098\left[450^{1{,}0125}\right]\right\}\left\{0{,}813 - \frac{0{,}308}{0{,}028}\left\{1 - \left[-ln\left(1 - \frac{1}{100}\right)\right]^{-0{,}028}\right\}\right\} \quad (10.78)$$

Q_{max} (100) = 124,1 m³/s

Exemplo 10.5 - Aplicar o método *index-flood*, com momentos-L, para regionalizar as vazões mínimas da bacia do alto rio das Velhas. O Anexo 13 apresenta os dados de vazões mínimas de 5 estações para 4 durações diferentes, 1 dia, 3 dias, 5 dias e 7 dias. A Tabela 10.16 apresenta algumas informações sobre as estações e o mapa da Figura 10.12 mostra a localização dos postos.

Tabela 10.16 – Estações para regionalização de vazões mínimas

Código	Estação	Rio	Área Km²	$P_{médio}$ (m)	L (Km)	DD (Junções/Km²)	I_{equiv} (m/Km)
41151000	Faz. Água Limpa	Velhas	174,6	1,498	26,15	0,115	8,59
41180000	Itabirito-Linígrafo	Itabirito	330	1,518	47,7	0,252	5,25
41199998	Honório Bicalho – Mont.	Velhas	1698	1,535	90,3	0,212	2,56
41260000	Pinhões	Velhas	3727	1,475	156,8	0,204	1,42
41340000	Ponte Raul Soares	Velhas	4874	1,458	200,3	0,209	1,13

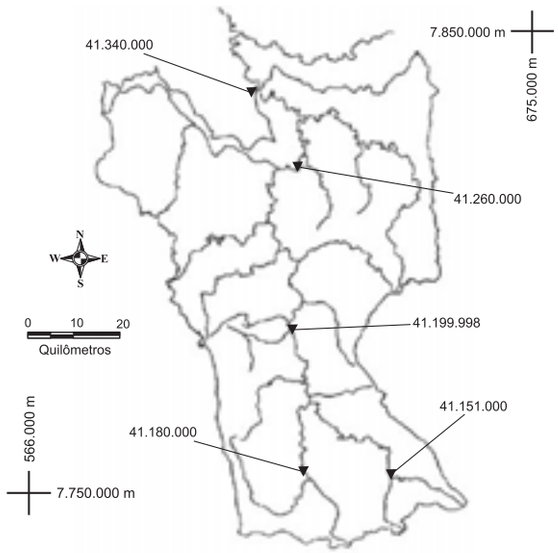


Figura 10.12 – Localização das estações da bacia do rio das Velhas

Solução: A primeira etapa consistiu na definição da região homogênea. Como o número de estações é reduzido, optou-se por verificar inicialmente a homogeneidade da região formada pelas 5 estações. Essa análise foi realizada grafando as curvas empíricas adimensionais para verificação do comportamento das curvas em um papel de probabilidades. O fator de adimensionalização utilizado foi a média de cada série. A posição de plotagem de Weibull foi utilizada para cálculo da freqüência empírica. Os resultados

para cada uma das durações mostraram que a região formada pelas 5 estações pode ser considerada homogênea. A Figura 10.13 ilustra o comportamento das distribuições empíricas para as vazões mínimas com duração de 7 dias.

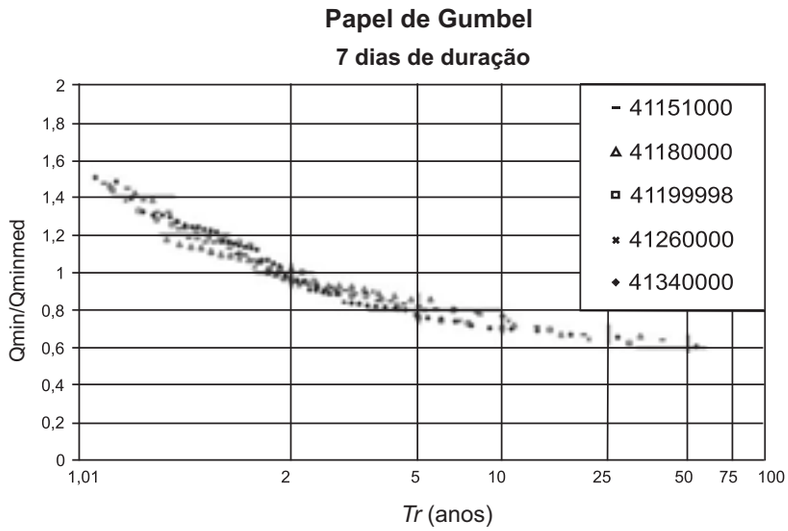


Figura 10.13 – Distribuições empíricas com 7 dias de duração, exemplo 10.5

Nesse exemplo, em que é efetuada a regionalização de vazões mínimas, não é possível utilizar as rotinas descritas por Hosking (1991), uma vez que esses algoritmos ajustam distribuições adequadas para a análise de vazões máximas. Sendo assim, a seleção das distribuições de freqüências regionais foi efetuada a partir da verificação do ajuste das distribuições de Gumbel e Weibull (2 parâmetros) para mínimos.

Grafando em um mesmo papel de probabilidades de Gumbel as distribuições empíricas adimensionais para todas as durações, constatou-se que essas apresentavam a mesma tendência, sem dispersões significativas. A Figura 10.14 ilustra as distribuições empíricas de 4 diferentes durações, para a estação de Honório Bicalho, código 41199998. A constatação do comportamento similar das distribuições empíricas adimensionais de diferentes durações permitiu a utilização das séries de 7 dias de duração para a verificação do ajuste das distribuições de Gumbel e Weibull. A verificação consistiu no ajuste dessas distribuições a cada uma das séries com duração de 7 dias, e posterior aplicação do teste de Filliben (ver capítulo 7 e Stedinger et al., 1993) e verificação visual do ajuste. A distribuição de Weibull foi aprovada em todas séries pelo teste de Filliben, para um nível de

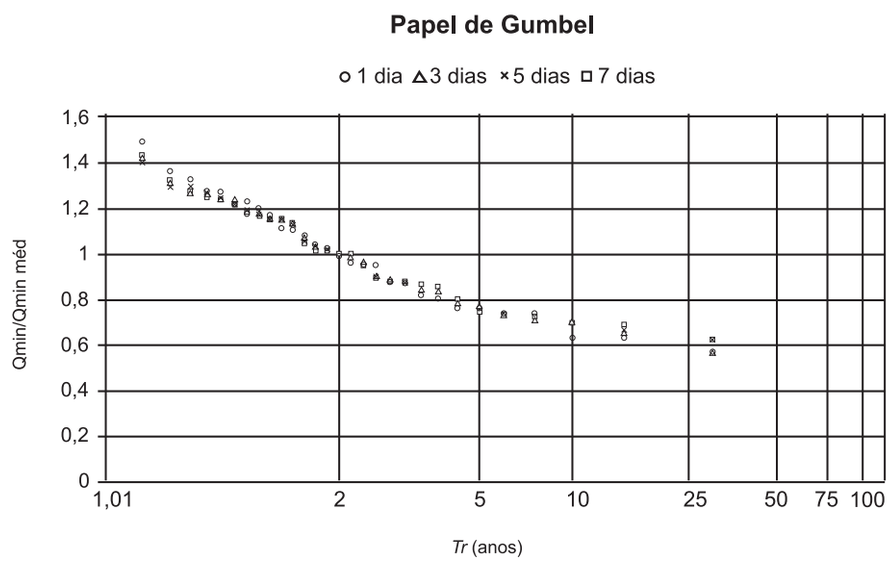


Figura 10.14 – Distribuições empíricas de Honório Bicalho, exemplo 10.5

significância de 5%, e apresentou um ajuste visual bem melhor do que o da distribuição de Gumbel. Dessa maneira, optou-se por ajustar a distribuição de Weibull (2P) às distribuições empíricas regionais adimensionais. As funções densidade de probabilidade, de distribuição acumulada e inversa de Weibull (2P) são as seguintes:

$$f(x) = \left(\frac{k}{\alpha}\right)\left(\frac{x}{\alpha}\right)^{k-1} \exp\left[-\left(\frac{x}{\alpha}\right)^{k}\right] \tag{10.79}$$

$$F(x) = 1 - \exp\left[-\left(\frac{x}{\alpha}\right)^{k}\right] \tag{10.80}$$

$$x(F) = \alpha[-ln(F)]^{1/k} \tag{10.81}$$

Definida para $x > 0$ e $\alpha, k > 0$, onde α e k são os parâmetros de escala e de forma. Segundo Stedinger et al. (1993) existe uma importante relação entre a distribuição de Weibull e a de Gumbel. Se uma variável aleatória X possui distribuição de Weibull, então a variável $Y = -ln[X]$ será distribuída conforme um modelo de Gumbel. Os métodos de estimação de parâmetros e os testes de aderência disponíveis para a distribuição de Gumbel podem ser utilizados também para a distribuição de Weibull. Assim, se $+ln[X]$ possui média $\lambda_{1,(lnX)}$ e momento-L de ordem 2 dado por $\lambda_{2,(lnX)}$, então os parâmetros da distribuição de Weibull (2P), para a variável X, obedecem às seguintes relações:

$$k = \frac{ln(2)}{\lambda_{2,(lnX)}} \tag{10.82}$$

$$\alpha = exp\left(\lambda_{1,(lnX)} + \frac{0{,}5772}{k}\right) \tag{10.83}$$

Assim, após a definição da região homogênea, e com o objetivo de efetuar o ajuste da distribuição de Weibull, foi necessário calcular os logaritmos naturais dos valores das séries adimensionalizadas, de forma a permitir a estimação de parâmetros como acima descrito. Em seguida, foram calculados os momentos-L e as razões-L individuais e regionais dos logaritmos das séries adimensionalizadas. Os momentos ponderados por probabilidade, β_r ,de cada estação, foram calculados utilizando os estimadores não enviesados, equações 10.47, 10.48 e 10.49. As estimativas dos MPP β_r permitiram o cálculo dos momentos-L, equações 10.54, 10.55, 10.56 e 10.57, e das razões-L, equações 10.60 e 10.61. Os valores regionais foram calculados com a equação 10.44. Os resultados estão na Tabela 10.17.

Tabela 10.17 – Momentos-L e Razões-L, exemplo 10.5

Duração: 1 Dia						
Código	**N**	**L1**	**L2**	**T3**	**T4**	**T5**
41151000	39	-0,02105	0,1188	0,0155	0,1277	-0,0537
41180000	32	-0,01954	0,1133	-0,0395	0,2044	-0,025
41199998	29	-0,03049	0,1481	-0,0889	0,0553	-0,0039
41260000	20	-0,02673	0,1397	-0,0606	0,0339	-0,0641
41340000	53	-0,03026	0,1444	-0,027	0,0549	0,0094
Regional		-0,02583	0,132953	-0,03399	0,096604	-0,02191
Duração: 3 Dias						
Código	**N**	**L1**	**L2**	**T3**	**T4**	**T5**
41151000	39	-0,02138	0,1202	0,0264	0,1073	-0,0365
41180000	32	-0,01807	0,1092	-0,0142	0,1791	0,0003
41199998	29	-0,02583	0,1361	-0,1119	0,0496	-0,0166
41260000	20	-0,02506	0,1354	-0,0562	0,0381	-0,0611
41340000	53	-0,02737	0,1375	-0,0273	0,0267	0,0191
Regional		-0,02377	0,127888	-0,03029	0,078216	-0,01217
Duração: 5 Dias						
Código	**N**	**L1**	**L2**	**T3**	**T4**	**T5**
41151000	39	-0,02126	0,1199	0,0359	0,1038	-0,0444
41180000	32	-0,0174	0,1072	-0,0069	0,177	-0,0074
41199998	29	-0,02381	0,1309	-0,0878	0,0231	-0,0042
41260000	20	-0,02132	0,124	-0,0186	0,0636	-0,0901
41340000	53	-0,02702	0,1368	-0,0257	0,0202	0,0045
Regional		-0,02275	0,125046	-0,01792	0,073553	-0,02112
Duração: 7 Dias						
Código	**N**	**L1**	**L2**	**T3**	**T4**	**T5**
41151000	39	-0,02096	0,1191	0,0336	0,1036	-0,0418
41180000	32	-0,01765	0,1079	-0,0101	0,1806	-0,0022
41199998	29	-0,02363	0,1301	-0,0775	0,0343	0,0084
41260000	20	-0,02076	0,1224	-0,0214	0,0769	-0,0946
41340000	53	-0,02727	0,1375	-0,0289	0,0109	-0,0017
Regional		-0,02271	0,124891	-0,01861	0,07474	-0,01988

Os valores dos momentos-L regionais permitiram a estimação dos parâmetros da distribuição de Weibull, por meio das equações 10.82 e 10.83. Os parâmetros regionais calculados encontram-se na Tabela 10.18.

Tabela 10.18 – Parâmetros da distribuição de Weibull

	Duração			
Parâmetros	**1 Dia**	**3 Dias**	**5 Dias**	**7 Dias**
Forma (k)	5,2135	5,4200	5,5431	5,5500
Escala (α)	1,0886	1,0862	1,0848	1,0847

Com as estimativas dos parâmetros da Tabela 10.18 e com a equação 10.81, foi possível calcular os quantis regionais adimensionais apresentados na Tabela 10.19. O ajuste das distribuições regionais e empíricas pode ser visualizado na Figura 10.15.

Tabela 10.19 – Quantis regionais adimensionais

	T (anos)							
Duração	1,01	2	5	10	20	25	50	100
1 Dia	1,460	1,015	0,816	0,707	0,616	0,589	0,515	0,450
3 Dias	1,440	1,015	0,824	0,717	0,628	0,602	0,529	0,465
5 Dias	1,429	1,015	0,828	0,723	0,635	0,609	0,537	0,473
7 Dias	1,429	1,015	0,828	0,723	0,635	0,610	0,537	0,474

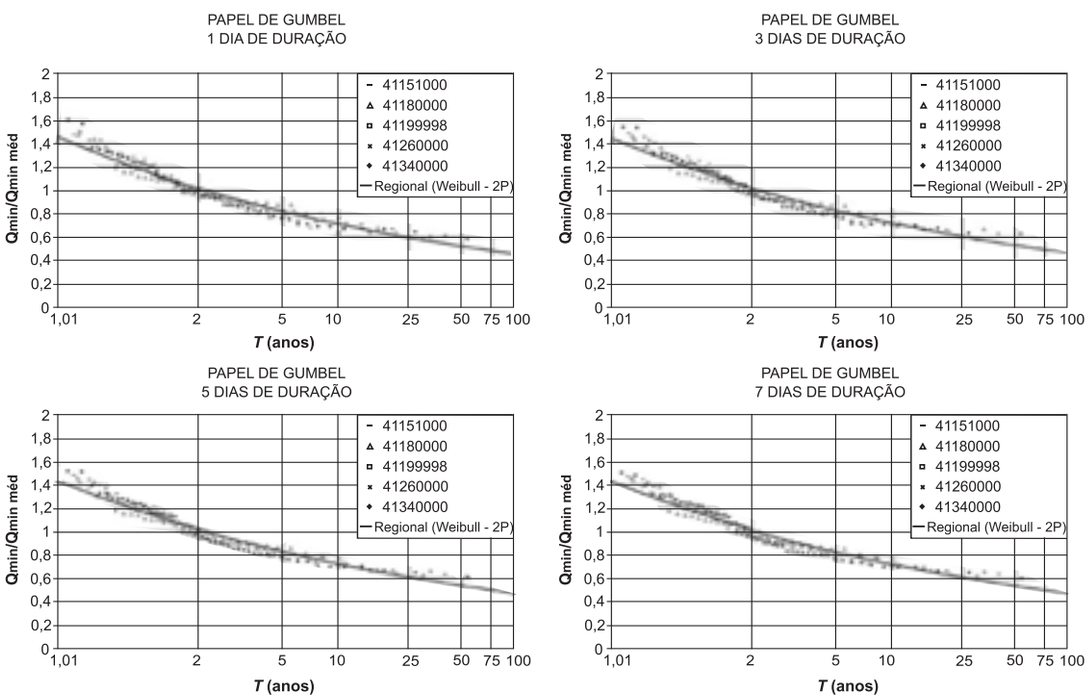


Figura 10.15 – Ajuste das distribuições empíricas e regionais, exemplo 10.5

No método *index-flood*, a estimação de quantis associados a diferentes tempos de retorno é realizada por meio da equação 10.43. Analogamente aos outros exemplos, em locais não monitorados e situados dentro da região homogênea, é necessário estimar o fator de adimensionalização. Assim, a última etapa corresponde à regressão entre os fatores de adimensionalização, nesse caso as médias amostrais, $Q_{med\text{-}min}$, apresentadas na Tabela 10.20, e as características das bacias, dadas na Tabela 10.16.

Tabela 10.20 – Vazões médias das séries de mínimas (m³/s)

			Duração			
Código	Estação	Rio	1 Dia	3 Dias	5 Dias	7 Dias
41151000	Faz. Água Limpa	Velhas	1,479	1,501	1,515	1,527
41180000	Itabirito-Linígrafo	Itabirito	4,030	4,060	4,090	4,11
41199998	Honório Bicalho – Mont.	Velhas	13,16	13,93	14,26	14,46
41260000	Pinhões	Velhas	25,0	25,6	26,3	26,73
41340000	Ponte Raul Soares	Velhas	25,6	27,0	27,73	28,27

Durante a análise de regressão, verificou-se que as informações da estação de Fazenda Água Limpa, código 41151000, introduziam distorções nos resultados da região onde estava inserida, tendo sido, então, eliminada do processo. Ao final da análise, foi adotado o seguinte modelo potencial:

$$Q_{min-med-D} = 0{,}0585 D^{0{,}0357} A^{0{,}7273} \qquad (A \geq 330 \text{ Km}^2) \qquad (10.84)$$

onde $Q_{min\text{-}med\text{-}D}$ é a média das vazões mínimas anuais com duração D em (m³/s), D é a duração em dias e A a área de drenagem em km². Substituindo as equações 10.81 e 10.84 na equação 10.43, é possível estimar diretamente os quantis associados a diferentes tempos de retorno, para pontos não monitorados e dentro da região homogênea, por meio da seguinte equação:

$$Q_{min-D,T} = \left\{0{,}0585 D^{0{,}0357} A^{0{,}7273}\right\} \left\{\alpha \left[-ln\left(\frac{1}{T}\right)\right]^{1/k}\right\} \qquad (10.85)$$

onde $Q_{min\text{-}D,T}$ é a vazão mínima com duração D, associada ao tempo de retorno T anos, e α e k são os parâmetros de escala e de forma da distribuição de Weibull para diferentes durações, apresentados na Tabela 10.18.

Exemplo 10.6 – A Tabela 10.21 apresenta a lista de 8 estações pluviográficas localizadas na região da Serra dos Órgãos, no Estado do Rio de Janeiro, tal como está ilustrado na Figura 10.16. Utilizar os dados das séries de duração parcial (2.*n*) destas estações, apresentados no Anexo 14 e aplicar o método *index-flood*, com as estatísticas-L, para estabelecer as relações intensidade-duração-frequência regionais para durações de 2, 3, 4, 8, 14 e 24 horas.

Tabela 10.21 – Estações pluviográficas

Código	Estações	Ent.	*N* (anos)	2.*n*	Precipitação Média anual (mm)	Altitude (m)
02243235	Andorinhas	SERLA	21	42	2462	79,97
02242092	Apolinário	SERLA	20	40	2869	719,20
02242096	Faz. Sto. Amaro	SERLA	21	42	2619	211,89
02242070	Nova Friburgo	INMET	19	38	1390	842,38
02243188	Petrópolis	INMET	6	12	1939	895,00
02242098	Posto Garrafão	SERLA	21	42	2953	641,54
02242093	Quizanga	SERLA	21	42	1839	13,96
02243151	Teresópolis-PN	INMET	8	16	2550	959,30

Fonte: Davis e Naghettini (2001)

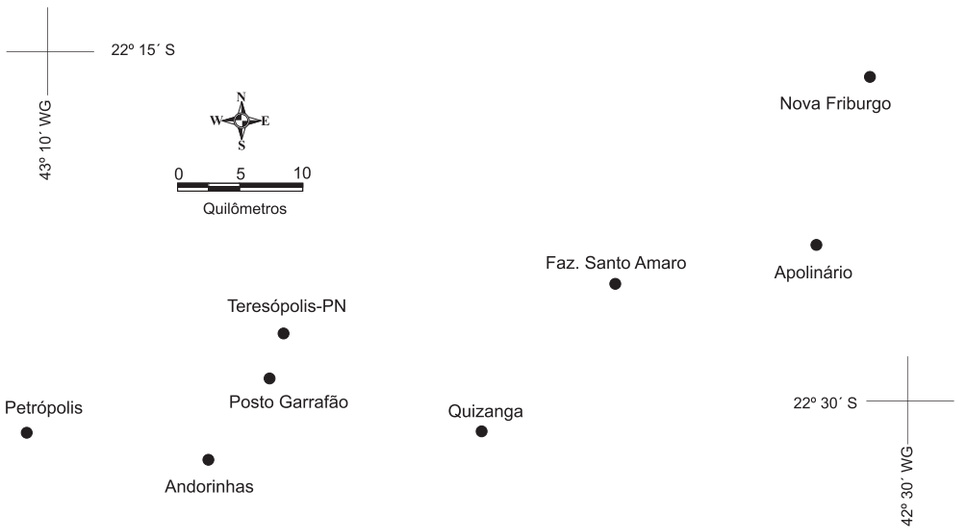


Figura 10.16 – Localização das estações do exemplo 10.6

O emprego do método *index-flood* na análise regional das relações intensidade-duração-freqüência de precipitações implica no uso da seguinte equação:

$$\hat{i}_{T,D,j} = \bar{i}_D \mu_{D,T} \tag{10.86}$$

onde $\hat{i}_{T,D,j}$ é a estimativa da intensidade da chuva (mm/h), de duração D (horas), no local j, associada ao tempo de retorno T (anos); $\bar{i}_D$ é o fator de adimensionalização das precipitações intensas (mm/h), de duração D, cuja estimativa em locais sem dados pluviográficos é efetuada por meio de uma análise de regressão entre os fatores de adimensionalização e as características fisiográficas e climáticas da região; e $\mu_{D,T}$ representa o quantil adimensional, de validade regional, associado à duração D e ao tempo de retorno T (anos).

Para a concretização da análise regional das relações IDF, utilizando séries de duração parcial, é necessário realizar basicamente dois estudos: o primeiro, de estimativa dos quantis adimensionais regionais, ou seja, a análise de freqüência regional das séries de intensidade de precipitação, de duração *D*, adimensionalizadas por um fator; e o segundo, que corresponde à análise de regressão entre os fatores de adimensionalização e as características fisiográficas e climáticas da região.

A análise de freqüência, com séries de duração parcial (SDP), pressupõe respostas a duas questões importantes: (i) qual é o melhor modelo distributivo discreto para o número de excedências dos eventos maiores que um limite previamente estipulado? e (ii) qual é o modelo distributivo contínuo para as magnitudes das excedências? O Anexo 9 apresenta os fundamentos teóricos da relação entre as distribuições de probabilidade dos máximos anuais e das excedências que compõem a SDP. Essa relação pode ser sintetizada por meio da seguinte equação:

$$F_a(x) = exp\{-\nu[1 - H_u(x)]\} \qquad (10.87)$$

onde ν indica a *intensidade média anual de ocorrências*, $H_u(x)$ denota a função de distribuição que está associada aos eventos que superaram o valor limiar *u* e pode ser prescrita pelo modelo paramétrico que melhor se ajustar aos dados amostrais, e $F_a(x)$ representa a distribuição de máximos anuais. A intensidade ou taxa anual de ocorrências pode ser estimada pelo número médio anual de eventos que superam o valor limiar *u*; por exemplo, se houverem *n* anos de registros e forem selecionados os 2*n* maiores valores de *X*, a estimativa de ν é 2. De acordo com a construção teórica, descrita no Anexo 9, a equação 10.87 pressupõe que as ocorrências superiores ao valor limiar *u* sejam independentes entre si e que o número dessas excedências, em um dado intervalo de tempo, seja uma variável de Poisson (ver Anexo 9 para detalhes sobre essas duas condicionantes). Como retratado no anexo citado, a verificação da hipótese de que as ocorrências são oriundas de um processo de Poisson pode ser efetuada pela aplicação do teste de Cunnane (1979). Assim, a primeira etapa do trabalho consiste em verificar se o número de excedências em relação a um valor limiar é uma variável de Poisson. Neste caso, aplicou-se o teste proposto por Cunnane (1979), a um nível de significância 2,5%, aos dados das séries das estações da Tabela 10.21 e constatou-se que não existem razões para descartar a hipótese de que o número de excedências é uma variável de Poisson. A seguir, foi realizada a definição das regiões homogêneas. Como o número

de estações é reduzido, optou-se por verificar inicialmente a homogeneidade da região formada pelas 8 estações. Essa análise foi realizada grafando as curvas empíricas adimensionais no papel de Gumbel. O fator de adimensionalização utilizado foi a média de cada série histórica. A posição de plotagem de Weibull foi utilizada para cálculo da freqüência empírica. Os resultados para cada uma das durações mostraram que a região formada pelas 8 estações pode ser considerada homogênea. A Figura 10.17 ilustra o comportamento das distribuições empíricas adimensionais com duração de 24 horas.

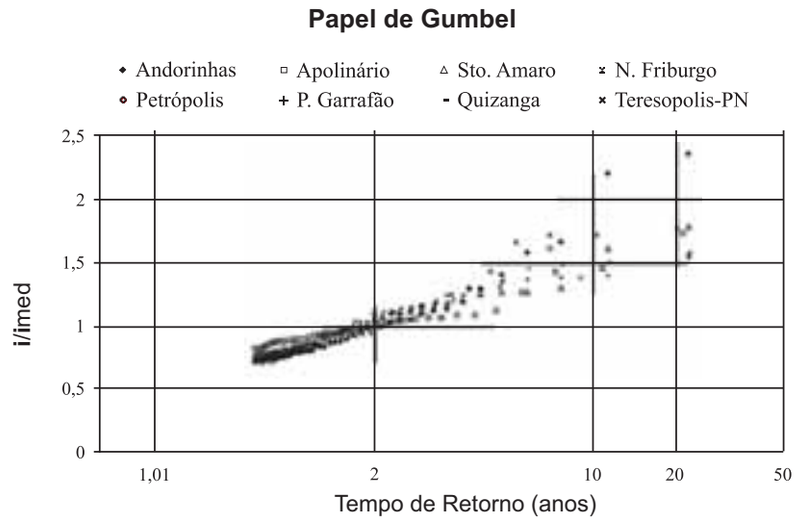


Figura 10.17 – Distribuições empíricas adimensionais com duração de 24 horas, exemplo 10.6

Além da análise gráfica, foram calculadas as medidas de heterogeneidade para cada duração, utilizando as rotinas mencionadas anteriormente. Os valores das medidas de heterogeneidade foram calculados com base no CV-L, na Assimetria-L e na Curtose-L, e estão apresentados na Tabela 10.22. Os resultados mostram que as medidas de heterogeneidade calculadas a partir do CV-L, da Assimetria-L e da Curtose-L indicam que a região pode ser considerada 'aceitavelmente homogênea', uma vez que $H < 1$.

Tabela 10.22 – Resultados da medida de heterogeneidade, exemplo 10.6

	2 Horas	3 Horas	4 Horas	8 Horas	14 Horas	24 Horas
$H\ (t^R)$	3,20**	1,20*	0,49	0,59	0,35	0,65
$H\ (t/t_3)$	1,01*	-1,24	-0,96	-0,15	0,37	0,24
$H\ (t_3/t_4)$	-0,05	-1,46	-1,64	-0,31	-0,18	0,40

* Possivelmente heterogênea

** Definitivamente heterogênea

Após a definição das regiões homogêneas, inicia-se a seleção das distribuições de freqüências regionais para as séries parciais utilizando as rotinas Fortran-77 mencionadas anteriormente. Essas fazem o ajuste das distribuições Logística Generalizada, Generalizada de Valores Extremos, Generalizada de Pareto, Generalizada Normal e Pearson tipo III, estimando os seus parâmetros a partir dos momentos-L regionais das séries parciais, além de aplicar o teste de aderência para verificar o ajuste entre a distribuição candidata e os dados regionais. Os momentos ponderados por probabilidade, β_r, de cada estação, foram calculados utilizando os estimadores por posição de plotagem, dados pelas equações 10.51, 10.52 e 10.53. As estimativas dos MPP's β_r permitiram o cálculo dos momentos-L, por meio das equações 10.54, 10.55, 10.56 e 10.57, e das razões-L, pelas equações 10.60 e 10.61. Os momentos-L e as razões-L amostrais estão apresentados em forma de tabela, no Anexo 14. A Tabela 10.23 mostra os valores das razões-L e dos momentos-L regionais para cada duração, obtidos a partir das séries parciais adimensionalizadas.

Tabela 10.23 – Valores regionais das Razões-L e dos Momentos-L, exemplo 10.6

Duração	*II*	CV - L (t_2)	Assimetria - L (t_3)	Curtose - L (t_4)
2 Horas	1	0,1222	0,2360	0,1534
3 Horas	1	0,1230	0,2838	0,2073
4 Horas	1	0,1281	0,2973	0,2187
8 Horas	1	0,1365	0,3441	0,2150
14 Horas	1	0,1404	0,3441	0,2101
24 horas	1	0,1357	0,3132	0,1894

Os valores das razões-L e dos momentos-L regionais, apresentados na Tabela 10.23, permitem a estimação dos parâmetros das distribuições ajustadas. A Tabela 10.24 apresenta os resultados dos testes de aderência, os quais, segundo os critérios de Hosking e Wallis (1993), são considerado adequados quando $|Z| \leq 1{,}64$.

Tabela 10.24 – Resultados dos testes de aderência (Z)

Distribuição	2 Horas	3 Horas	4 Horas	8 Horas	14 Horas	24 Horas
Generalizada Logística	2,11	0,5*	0,28*	1,08*	1,26*	1,59*
Generalizada de Valores Extremos	0,87*	-0,34*	-0,48*	0,46*	0,63*	0,81*
Generalizada Normal	0,38*	-0,91*	-1,07*	-0,28*	-0,12*	0,1*
Pearson Tipo III	-0,52*	-1,9	-2,09	-1,54*	-1,4*	-1,1*
Generalizada de Pareto	-2,13	-2,56	-2,54	-1,41*	-1,27*	-1,38*

* $|Z| \leq 1{,}64$

Analisando os resultados da Tabela 10.24, verifica-se que, segundo os critérios de Hosking e Wallis (1993), as distribuições Generalizada de Valores Extremos e Generalizada Normal ajustaram-se às distribuições empíricas, para todas as durações. Como, neste exemplo, foram utilizadas

as séries com valores de intensidade de precipitação superiores a determinados limites, é razoável que se ajuste a distribuição Generalizada de Valores Extremos para todas as durações. Como critério suplementar, verificou-se no diagrama Curtose-L x Assimetria-L o posicionamento das razões-L regionais, conforme ilustrado pela Figura 10.18.

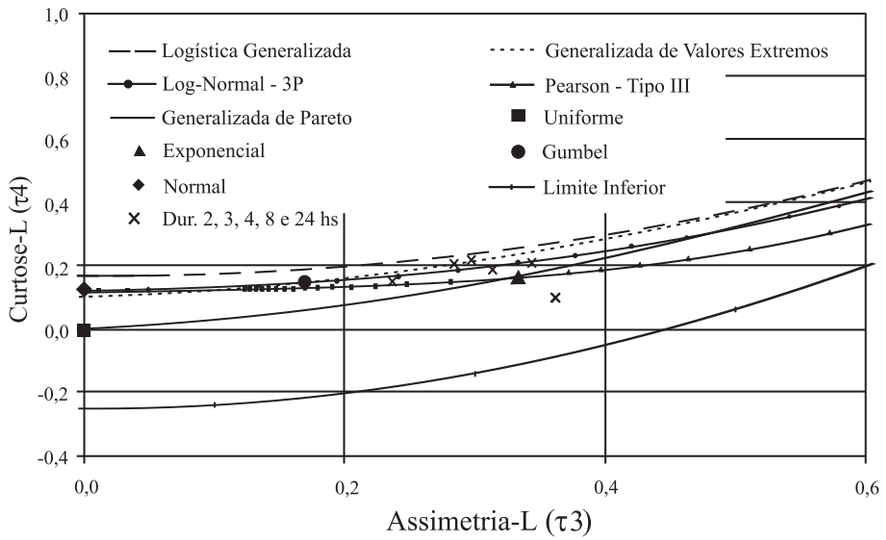


Figura 10.18 – Diagrama Curtose-L x Assimetria-L, exemplo 10.6

As razões-L e os momentos-L regionais da Tabela 10.23 foram empregados para calcular os parâmetros da GEV regional adimensional por meio das equações 10.72, 10.73 e 10.74. Os valores estimados dos parâmetros da GEV regional para cada duração estão na Tabela 10.25.

Tabela 10.25 – Parâmetros da Distribuição Generalizada de Valores Extremos regional

Duração	Posição (ξ)	Escala (α)	Forma (k)
2 Horas	0,891	0,159	-0,101
3 Horas	0,885	0,148	-0,171
4 Horas	0,879	0,150	-0,190
8 Horas	0,867	0,146	-0,255
14 Horas	0,863	0,150	-0,255
24 horas	0,870	0,154	-0,212

Os parâmetros da GEV apresentados na Tabela 10.25 foram estimados utilizando séries de duração parcial, de modo que, para o cálculo de quantis associados a probabilidades anuais (ou tempos de retorno), deve-se aplicar o modelo Poisson-GEV, cuja dedução está no Anexo 9. Em resumo, conhecendo-se a taxa de excedência ν, no caso $\nu = 2$, e os parâmetros da

distribuição Generalizada de Valores Extremos, esses estimados a partir das excedências sobre o limiar estabelecido *u*, os quantis anuais podem ser calculados por meio da equação A9.33, reescrita a seguir:

$$x(F) = \xi + \frac{\alpha}{k}\left\{1 - \left[-Ln\left(\frac{\nu + Ln(F(x))}{\nu}\right)\right]^k\right\} \quad \text{se} \quad k \neq 0 \qquad (10.88)$$

onde $F(x) = 1 - \frac{1}{T(anos)}$; k, α e ξ são, respectivamente, os parâmetros de forma, escala e posição da distribuição Generalizada de Valores Extremos, estimados a partir das excedências sobre o limiar estabelecido *u* e que constam da Tabela 10.25. A partir dos parâmetros da distribuição GEV, listados na Tabela 10.25, e aplicando a equação 10.88, com $\nu = 2$, obtém-se, para cada duração, os quantis regionais adimensionais anuais apresentados na Tabela 10.26.

Tabela 10.26 – Quantis regionais adimensionalizados, $\mu_{D,T}$

	Duração					
T (anos)	2 Horas	3 Horas	4 Horas	8 Horas	14 Horas	24 Horas
2	1,033	1,021	1,018	1,006	1,007	1,015
5	1,270	1,265	1,274	1,281	1,290	1,287
10	1,431	1,443	1,463	1,499	1,514	1,493
20	1,593	1,633	1,669	1,748	1,769	1,721
25	1,647	1,697	1,740	1,836	1,860	1,800
50	1,820	1,913	1,977	2,141	2,174	2,069
75	1,926	2,050	2,130	2,345	2,383	2,244
100	2,004	2,153	2,246	2,503	2,546	2,377

A próxima etapa consistiu na análise de regressão entre os fatores de adimensionalização e as características físicas e climáticas da bacia. As intensidades médias das séries de duração parcial, usadas como fator de adimensionalização, e as variáveis de características físicas e climáticas empregadas na análise de regressão estão apresentados na Tabela 10.27.

Tabela 10.27 – Fatores de adimensionalização e variáveis explicativas, exemplo 10.6

Código	Estações	2H (mm/h)	3H (mm/h)	4H (mm/h)	8H (mm/h)	14H (mm/h)	24H (mm/h)	Precipitação Média Anual (mm)	Altitude (m)
02243235	Andorinhas	39,91	30,14	24,37	14,2	8,9	5,75	2462	79,97
02242092	Apolinário	31,75	23,75	19,01	11,38	7,42	5,17	2869	719,20
02242096	Faz. Sto. Amaro	34,02	25,42	20,44	11,82	7,61	5,23	2619	211,89
02242070	Nova Friburgo	21,52	15,42	12,51	7,15	4,5	2,73	1390	842,38
02243188	Petrópolis	29,46	21,9	17,75	10,86	7,01	4,36	1939	895,00
02242098	Posto Garrafão	39,74	29,96	24,14	14,11	9	5,95	2953	641,54
02242093	Quizanga	34,25	24,59	19,81	11,15	6,95	4,42	1839	13,96
02243151	Teresópolis-PN	20,34	15,41	12,45	7,32	4,49	2,76	2550	959,30

Na análise de regressão, foram testados modelos lineares e potenciais, o que obrigou a transformação logarítmica das variáveis da Tabela 10.27. Durante a análise de regressão verificou-se que as informações da estação

de Teresópolis-PN, código 02243151, introduziam distorções nos resultados da região onde estava inserida, tendo sido excluída por essa razão. Ao final da análise, foi adotado o seguinte modelo potencial:

$$\bar{i}_D = 0{,}241 D^{-0{,}78} PMA^{0{,}708} \qquad (2\text{h} \geq D \leq 24\text{h}) \qquad (10.89)$$

onde $\bar{i}_D$ é o fator de adimensionalização das precipitações intensas (mm/h) de duração D, D é a duração em horas e PMA é precipitação média anual em mm.

As relações IDF podem ser estabelecidas para qualquer local dentro da região homogênea utilizando a equação 10.86, onde o quantil adimensional, $\mu_{D,T}$, de validade regional, associado a duração D e ao tempo de retorno T (anos), pode ser obtido na Tabela 10.26 ou calculado através da equação 10.88; e $\bar{i}_D$, que é o fator de adimensionalização das precipitações intensas de duração D, é estimado com a equação 10.89.

De modo análogo, substituindo as equações 10.88 e 10.89 na equação 10.86, obtém-se a seguinte equação que também permite definir as relações IDF em locais dentro da região homogênea:

$$\hat{i}_{T,D,j} = \left(0{,}241 D^{-0{,}78} PMA^{0{,}708}\right)\left\{\xi + \frac{\alpha}{k}\left\{1 - \left[-Ln\left(\frac{\nu + Ln\left(F(x)\right)}{\nu}\right)\right]^k\right\}\right\} \qquad (10.90)$$

Nessa equação, $\hat{i}_{T,D,j}$ é a estimativa da intensidade da chuva (mm/h), de duração D (horas), no local j, associada ao tempo de retorno T (anos); D é a duração em horas; PMA é precipitação média anual no local j em mm;

$F(x) = 1 - \frac{1}{T(anos)}$; $\nu = 2$; k, α e ξ são, respectivamente, os parâmetros de forma, escala e posição da distribuição Generalizada de Valores Extremos, regional e adimensional, estimados a partir das excedências sobre o limiar estabelecido u, conforme valores da Tabela 10.25.

Exercícios

1 – Aplicar o método de regionalização de quantis associados a diferentes riscos aos dados de vazões médias diárias máximas anuais de algumas estações da bacia

do rio Paraopeba que constam do Anexo 12. Regionalizar os quantis associados aos tempos de retorno de 5, 10, 25, 50 e 100 anos.

2 – A resolução Nº 394 da ANEEL, de 04 de dezembro de 1998, define: "Artigo 2º Os empreendimentos hidrelétricos com potência superior a 1.000 kw e igual ou inferior a 30.000 kw, com área total de reservatório igual ou inferior a 3,0 km^2, serão considerados como aproveitamentos com características de pequenas centrais hidrelétricas. Parágrafo único: "a área do reservatório é delimitada pela cota d'água associada à vazão de cheia com tempo de recorrência de 100 anos." Suponhamos que um empreendedor necessite da estimativa da vazão de cheia com tempo de retorno de 100 anos para definir a área do reservatório e conseqüentemente estabelecer se o aproveitamento que pretende construir terá as características de pequena central hidrelétrica. Comparar os resultados de estimativa da vazão de cheia com tempo de retorno de 100 anos, utilizando as regionalizações do exercício 1 e dos exemplos 10.2, 10.3 e 10.4. O futuro empreendimento estará localizado no rio Maranhão, próximo à cidade mineira de Congonhas, na bacia do rio Paraopeba. A área de drenagem até o ponto de instalação do empreendimento é de 300 km^2.

3 – Regionalizar as vazões mínimas anuais médias de 7 dias de duração das estações da bacia do rio Paraopeba. Os dados constam do Anexo 11. Aplicar o método de regionalização *index-flood* com e sem a utilização dos momentos-L.

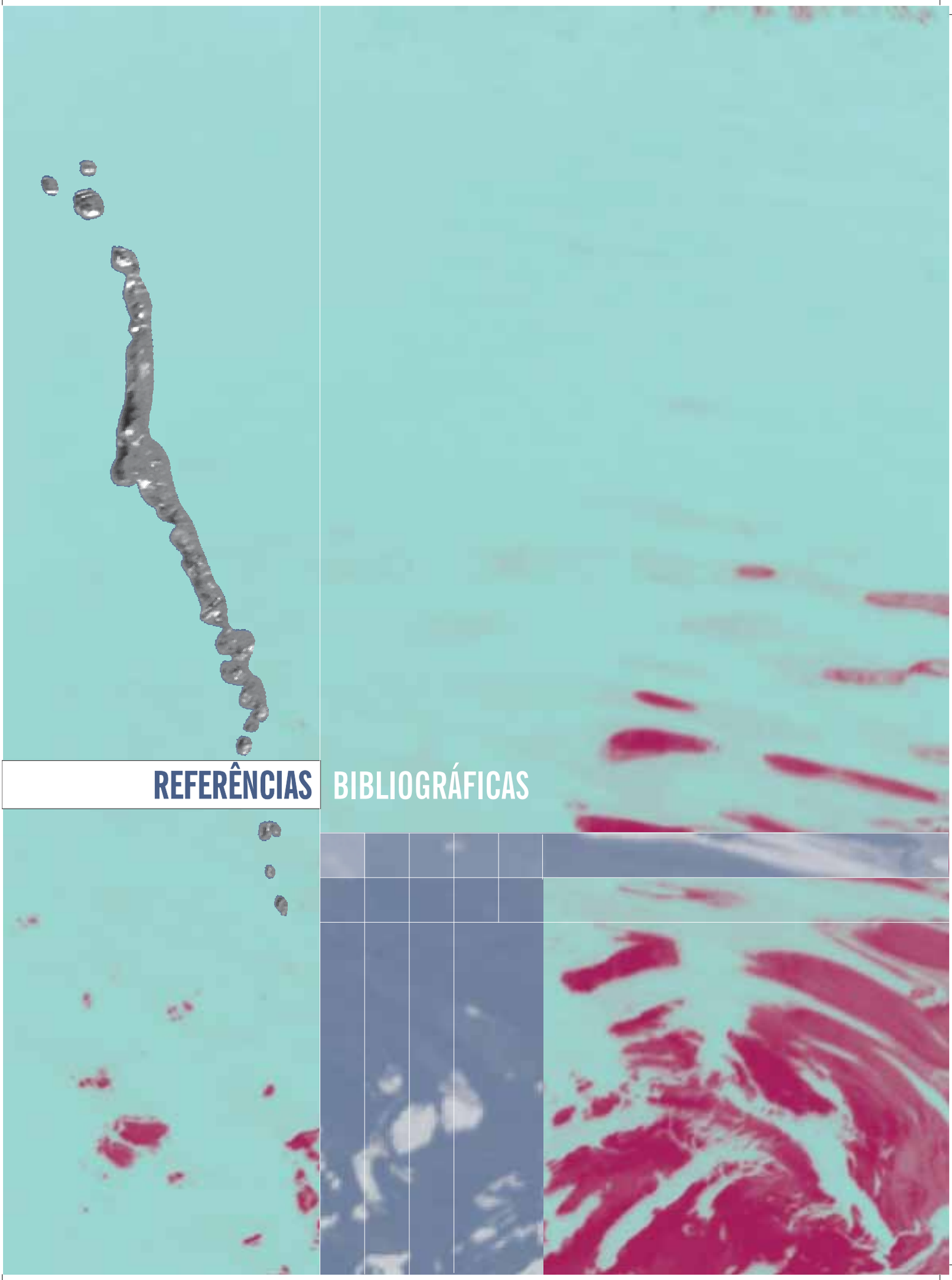
4 – Um empreendedor deseja estimar a vazão de referência para cálculo das disponibilidades hídricas, ou seja, a vazão mínima com 7 dias de duração e 10 anos de tempo de retorno, no rio Paraopeba, a jusante da localidade mineira de Belo Vale, para solicitar a outorga de derivação consuntiva dos recursos hídricos na seção do rio. A área de drenagem até o ponto de derivação é de 2900 km^2. Comparar as vazões de referência estimadas com as regionalizações do exercício 3 e do exemplo 10.1.

5 – O Anexo 15 apresenta os dados de precipitações diárias máximas anuais de 92 estações pluviométricas da bacia do Alto São Francisco. No anexo mencionado também estão disponíveis a listagem das estações com coordenadas geográficas e altitude, além dos mapas de localização e isoietas. Utilizando as informações do Anexo 15 e outras originárias de pesquisa individual, regionalizar as precipitações diárias máximas anuais da bacia do Alto São Francisco aplicando os seguintes métodos:

a) Regionalização de eventos associados aos tempos de retorno de 5, 10, 25 e 50 anos.
b) Regionalização dos parâmetros de uma distribuição de probabilidades.
c) Método *index-flood*.
d) Método *index-flood* empregando os momentos-L e as estatísticas-L

6 – Empregar o método de regionalização *index-flood* utilizando os momentos-L e as estatísticas-L para calcular as probabilidades das precipitações anuais de 19 estações da Área de Proteção Ambiental da região metropolitana de Belo Horizonte ocorrerem em determinadas classes (<1200mm; 1200-1500mm; 1500-1800mm; 1800-2200mm; >2200mm). Os resultados dessa estimativa podem ser encarados como padrões climatológicos das estações analisadas. O Anexo 16 apresenta informações e os dados de precipitações anuais das 19 estações pluviométricas da APA SUL-RMBH.

7 – A Tabela 10.21 apresenta a lista de 8 estações pluviográficas localizadas na região da Serra dos Órgãos no Estado do Rio de Janeiro, como ilustrado na Figura 10.16. Utilizar os dados das séries de duração parcial (2.*n*) destas estações apresentadas no Anexo 14 e aplicar o método *index-flood*, com estatísticas-L, para estabelecer as relações intensidade-duração-frequência regionais para durações de 5, 10, 15, 20, 30, 45 minutos e 1 hora.



REFERÊNCIAS BIBLIOGRÁFICAS

REFERÊNCIAS BIBLIOGRÁFICAS

ABRAMOWITZ, M. e STEGUN, I. A. *Handbook of Mathematical Functions*. New York: Dover, 1965.

ANDERSON, T. W. e DARLING, D. A. A test of goodness of fit. *Journal of the American Statistical Association*, 49, pp. 756-769, 1954.

ANG, A. H-S e TANG, W. H. *Probability Concepts in Engineering Planning and Design – Volume I Basic Principles*. New York: John Wiley & Sons, 1975.

ANG, A. H-S e TANG, W. H. *Probability Concepts in Engineering Planning and Design – Volume II Decicion, Risk, and Reliability*. Copyright Ang & Tang, 1990.

BAKER, V. R. Paleoflood Hydrology and extraordinary flood events. *Journal of Hydrology,* 96, p. 79-99, 1987.

BAYLISS, A. C. e REED, D. W. *The use of historical data in flood frequency estimation.* Report to Ministry of Agriculture, Fisheries and Food. Centre for Ecology and Hydrology (www.ceh.ac.uk). United Kingdom. Mar., 2001.

BECKMAN, P. *Elements of Applied Probability Theory*. New York: Harcourt, Brace and World, Inc., 1968.

BENJAMIN, J. R. e CORNELL, C. A. P*robability, Statistics, and Decision for Civil Engineers*. New York: McGraw-Hill, Inc., 1970.

BENSON, M. A. *Evolution of the methods for evaluating the occurrence of floods*. USGS, Water Resources Paper 1580-A, 1960.

BOBÉE, B., The log Pearson type 3 distribution and its application in hydrology. *Water Resources Research*, v.11, n.5, p. 681-689, 1975.

BOBÉE, B. e ASHKAR, F. *The Gamma Family and Derived distributions Applied in Hydrology*. Littleton (CO): Water Resources Publications, 1991.

BOBÉE, B. e RASMUSSEN, P. *Recent advances in flood frequency analysis*. U.S. National Report to IUGG, 1991-1994, R*ev. Geophysics*, v. 33 Suppl. (http://earth.agu.org/revgeophys/bobee01/bobee01.htm), 1995.

BOUGHTON, W. C. A frequency distribution for annual floods. *Water Resources Research*, v.16, p. 347-354, 1980.

BUISHAND, T.A., Statistics of extremes in climatology. *Statistica Neerlandica*, v. 43, p. 1-30, 1989.

BURN, D. H., Cluster analysis as applied to regional flood frequency. *Journal of Water Resources Planning and Management*, V. 115, p. 567-582, 1989.

BUSSAB, W. O. e MORETTIN, P. A. *Estatística Básica*. São Paulo: Editora Saraiva, 2002.

CASELLA, G. e BERGER, R. *Statistical Inference*. Belmont (CA): Duxbury Press, 1990.

CAVADIAS, G. S., The canonical correlation approach to regional flood estimation. In *Regionalization in Hydrology*, IAHS Publication 191, Ed. BERAN, M.; BRILLY, M.; BECKER, A. e BONACCI, O. IAHS, Wallingford, Reino Unido p. 171-178, 1990.

CHOW, V. T., The log probability law and its engineering applications. *Proceedings ASCE*, V. 80(536), p. 1-25, 1954.

CHOW, V. T. Section 8-I. Statistical and probability analysis of hydrologic data. Part I – Frequency Analysis. In: *Handbook of Applied Hydrology*. McGraw-Hill. USA. 1964

CHOWDHURY J. U., STEDINGER, J. R. e LU, L-H. Goodness of fit tests for regional generalized extreme value flood distributions. *Water Resources Research*, 27(7), pp. 1765-1776, 1991.

CLARKE, R. T. Estimating trends in data from the Weibull and a generalized extreme value distribution. *Water Resources Research*, v.38, n. 6, p. 25.1-25.10, 2002.

CORREIA, F. N. *Métodos de Análise e Determinação de Caudais de Cheia*, tese de concurso para Investigador Auxiliar do LNEC, Laboratório Nacional de Engenharia Civil, Lisboa, 380 pp., 1983.

COX, D. R.; ISHAM, V. S.. e NORTHROP, P. J. *Floods: some probabilistic and statistical approaches*. Research Report 224, University College London, Londres, 2002.

CRAMÉR, H. *Mathematical Methods of Statistics*. Princeton: Princeton University Press, 1946.

CRAMÉR, H. e LEADBETTER, M. R. *Stationary and related stochastic processes*, John Wiley, New York, 1967.

CRUTCHER, H. L. A note on the possible misuse of the Kolmogorov-Smirnov test. *Journal of Applied Meteorology*, 14, pp. 1600-1603, 1975.

CUNNANE, C. A particular comparison of annual maximum and partial duration series methods of flood frequency prediction, *Journal of Hydrology*, v.18, p. 257-271, 1973.

CUNNANE, C. Unbiased plotting positions – a review. *Journal of Hydrology,* 37, p. 205-222. 1978.

CUNNANE, C., A note on the Poisson assumption in partial duration series models, *Water Resources Research*, v.15, n. 2, p. 489-494, 1979.

D'AGOSTINO, R. B. e STEPHENS, M. *Goodness-of-fit Techniques*. New York: Marcel Dekker, 1986.

DALRYMPLE, T., *Flood-frequency analyses, Manual of Hydrology: Part.3. Flood-flow Techniques*, Geological Survey Water Supply Paper 1543-A, U.S. Government Printing Office, Washington, D.C., 80p., 1960.

DAVIS, E. G. e NAGHETTINI, M. C, *Estudo de Chuvas Intensas no Estado do Rio de Janeiro.* Belo Horizonte: CPRM, 2001.

ELETROBRÁS. *Metodologia para Regionalização de Vazões*, Vol. 1, Eletrobrás, DPE, Departamento de Recursos Energéticos, Rio de Janeiro, 203 pp.,1985.

FILLIBEN J. J. The probability plot correlation coefficient test for normality. *Technometrics*, 17(1), pp. 111-117, 1975.

FRÉCHET, M. Sur la loi de probabilité de l'écart maximum. *Annales de la Societé Polonaise de Mathématique*, 6, pp. 93-117, 1927.

GIBBONS, J. D., *Nonparametric Statistical Inference*. New York; McGraw-Hill, 1971.

GINGRAS, D. e ADAMOWSKI, K. Homogeneous region delineation based on annual flood generation mechanisms. *Hydrological Sciences Journal*, V. 38, n. 1, p. 103-121, 1993.

GREENWOOD, J.A.; LANDWEHR, J. M.; MATALAS, N. C. e WALLIS, J. R. Probability weighted moments: definition and relation to parameters expressible in inverse form. *Water Resources Research*, v.15, n.5, p.1049-1054, 1979.

GRUBBS, F. E. Sample criteria for testing outlying observations. *Annals of Mathematical Statistics*, 21(1), pp. 27-58, 1950.

GRUBBS, F. E. Procedures for detecting outlying observations in samples. *Technometrics*, 11(1), pp. 1-21, 1969.

GRUBBS F. E. e BECK G. Extension of sample sizes and percentage points for significance tests of outlying observations. *Technometrics*, 14(4), pp. 847-854, 1972.

GUMBEL E. J. *Statistics of Extremes*. New York: Columbia University Press, 1958.

GUPTA, V. K.; DUCKSTEIN, L. e PEEBLES, R. W. On the joint distribution of the largest flood and its occurrence time. *Water Resources Research*, v.12, n.2, p. 295-304, 1976.

GUTTMAN, N. B., The use of L-moments in the determination of regional precipitation climates. *Journal of Climate*, V. 6, p. 2309-2325, 1993.

HAAN, C. T. *Point of Impending Sediment Deposition for Open Channel Flow in a Circular Conduit*. Dissertação de mestrado, Purdue University, 1965.

HAAN, C. T. *Statistical Methods in Hydrology*. Ames (IA): The Iowa University Press, 1977.

HARTIGAN, J. A., *Clustering Algorithms*, Wiley, New York, 1975, *apud* Statsoft Inc., Electronic Statistics Textbook, Statsoft, Tulsa, OK, Estados Unidos (http://www.statsoft.com/textbook/stathome.html), 1997.

HELSEL, D. R. e HIRSCH R. M. *Statistical Methods in Water Resources*. Amsterdam: Elsevier, 1992.

HIRSH, R. M. Probability plotting position formulas for flood records with historical information. *Journal of Hydrology,* 96, p. 185-199. 1987.

HIRSH, R. M e STEDINGER, J. R. Plotting position for historical floods and their precision. *Water Resources Research*, v.23, p. 715-727, 1987.

HOLLANDER, M. e WOLFE, D. A. *Nonparametric Statistical Methods*. New York: John Wiley & Sons, 1973.

HOSKING J. R. M. The theory of probability weighted moments. *Research Report RC 12210*. Yorktown Heights (NY): IBM Research, 1986.

HOSKING J. R. M. L-Moments: analysis and estimation of distributions using linear combinations of order statistics. *Journal of the Royal Statistical Society*, B, 52(2), pp. 105-124, 1990.

HOSKING, J. R. M. Fortran routines for use with the method of L-moments - Version 2. In: *IBM Research Report*, New York, IBM Research Division, RC 17097, 117p., Ago., 1991.

HOSKING, J. R. M. e WALLIS, J. R. Paleoflood hydrology and flood frequency analysis. *Water Resources Research*, 22, p. 543-550, 1986a.

HOSKING, J. R. M. e WALLIS, J. R. The value of historical data in flood frequency analysis. *Water Resources Research*, 22, p. 1606-1612, 1986b.

HOSKING, J.R.M. e WALLIS, J. R. Some statistics useful in regional frequency analysis. *Water Resources Research*, v.29, n.1, p.271-281, 1993.

HOSKING, J.R.M. e WALLIS, J. R. Correction to "some statistics useful in regional frequency analysis". *Water Resources Research*. v.31, n.1, p.251, 1995a.

HOSKING, J.R.M. e WALLIS, J. R.. A comparison of unbiased and plotting-position estimators of *L*-moments. *Water Resources Research*, **31**, 2019-2025, 1995b.

HOSKING, J. R. M. e WALLIS, J. R *Regional Frequency Analysis - An Approach Based on L-Moments*, 224 p. Cambridge University Press, Cambridge, Reino Unido, 1997.

INSTITUTION OF ENGINEERS AUSTRALIA. *Australian rainfall and runoff : a guide to flood estimation.* V. 1, Institution of Engineers Australia, Canberra, Austrália, 374pp., 1987.

KACZMAREK, Z. *Statistical Methods in Hydrology and Meteorology*. Report TT 76-54040. Springfield (VA): National Technical Information Service, 1977.

KITE, G. W. *Frequency and Risk Analysis in Hydrology*. Fort Collins (CO): Water Resources Publications, 1977.

KOTTEGODA, N. T. e ROSSO, R. *Statistics, Probability, and Reliability for Civil and Environmental Engineers*. New York: McGraw-Hill, 1997.

LANDWEHR, J. M.; MATALAS, N. C. e WALLIS, J. R. Estimation of parameters and quantiles of Wakeby distributions. *Water Resources Research*, v. 15, p. 1361-1379, 1979.

LANGBEIN, W. B. Annual Floods and Partial-Duration Floods Series. In: *Transactions American Geophysical Union*, vol. 30, N. 6, Dec., 1949.

LARSEN, R. J. e MARX, M. L. *An Introduction to Mathematical Statistics and its Applications*. Englewood Cliffs (NJ): Prentice-Hall, 1986.

LAURSEN, E. M. Comment on "Paleohydrology of southwestern Texas" por KOCHEL, R. C ; BAKER, V. R.; PATTON, P. C. *Water Resources Research*, v.19, p.1339, 1983.

LEADBETTER, M. R.; LINDGREN, G. e ROOTZÉN, H. *Extremes and related properties of random sequences and processes*, Springer-Verlag, New York, 335 pp., 1983.

MADSEN, H., ROSBJERG, D. e HARREMOES, P. Application of the partial duration series approach in the analysis of extreme rainfalls in extreme hydrological events: precipitation, floods and droughts. *Proceedings of the Yokohama Symposium*, I.A.S.H. Publication 213, p.257-266, 1993.

MAIONE, U. e MOISELLO, U. *Elementi di Statistica per l'Idrologia*. Pavia (Itália): La Goliardica Pavese, 2003.

MANN, H. B. e WHITNEY, D. R. On the test of whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics*, 18, pp. 50-60, 1947.

MONTGOMERY, D. C. e PECK, E. A. *Introduction to Linear Regression Analysis*. John Wiley, New York, NY, USA, 504p., 1992.

NAGHETTINI, M.; POTTER K. W. e ILLANGASEKARE, T.. Estimating the upper-tail of flood-peak frequency distributions using hydrometeorological information. *Water Resources Research*, v.32, n.6, p.1729-1740, 1996.

NATHAN. R. J. e MCMAHON, T. Identification of homogeneous regions for the purpose of regionalization. *Journal of Hydrology*, V. 121, p. 217-238, 1990.

NERC. *Flood Studies Report, Vol. 1*. London: National Environmental Research Council, 1975.

NORTH, M., Time-dependent stochastic model of floods. *Journal of Hydraulics Division*, ASCE, V. 106, n. 05, p. 717-731, 1980.

NRC. *Estimating Probabilities of Extreme Floods*. National Research Council, National Academy Press, Washington, 141 pp., 1987.

PEARSON, C. P.. Regional flood frequency for small New Zealand basins 2 : flood frequency groups. *Journal of Hydrology* (Nova Zelândia), V. 30, p. 53-64, 1991.

PERICHI, L. R. e RODRÍGUEZ-ITURBE, I. On the statistical analysis of floods, in *A Celebration of Statistics*, ed. A C. Atkinson & S. E. Fienberg, Springer-Verlag, New York, p. 511-541, 1985.

PILON, P. J., CONDIE, R. e HARVEY, K. D. *Consolidated Frequency Analysis Package – Users Manual for Version 1*. Ottawa: Water Resources Branch, Inland Waters Directorate, 1985.

PINTO, E. J. A. e ALVES, M. S. *Regionalização de vazões das sub-bacias 40 e 41*. Belo Horizonte. ANEEL/CPRM. CD-ROM, dez.,2001.

PINTO E. J. A. e NAGHETTINI, M. Definição de Regiões Homogêneas e Regionalização de Freqüência das Precipitações Diárias Máximas Anuais da Bacia do Alto Rio São Francisco, *Anais do 13º Simpósio Brasileiro de Recursos Hídricos* (CD-ROM), Belo Horizonte, 1999.

POTTER, K. W. Research on flood frequency analysis: 1983-1986. *Rev. Geophys*., V. 26, n. 3, p. 113-118, 1987.

PRESS, W., TEUKOLSKY, S. A., VETTERLING, W. T. e FLANNERY, B. P. *Numerical Recipes in Fortran 77 – The Art of Scientific Computing*. Cambridge: Cambridge University Press, 1986.

RAO, C. R. *Linear Statistical Inference and its Applications*. New York: John Wiley & Sons, 1973.

RAO, A. R. e HAMED, K. H. *Flood Frequency Analysis*. Boca Raton (FL): CRC Press, 2000.

REICH, B. M., Lysenkoism in U. S. flood determinations, *AGU Surface Runoff Committee – Session on flood frequency methods*, San Francisco, CA, 13 pp., 1977.

ROSBJERG, D. Estimation in partial duration series with independent and dependent peak values, *Journal of Hydrology*, V. 76, p. 183-195, 1984.

ROSBJERG, D. e MADSEN, H. On the choice of threshold level in partial duration series, *Proceedings of the Nordic Hydrological Conference, Alta (Noruega)*, NHP Report 30, pp. 604-615, 1992.

ROSSI, F. M., FIORENTINO, M. e VERSACE, P. Two component extreme value distribution for flood frequency analysis, *Water Resources Research*, 20(7), 1984.

SALAS, J. D.; WOLD, E. E. e JARRETT, R. D. Determination of flood characteristics using systematic, historical and paleoflood data. *In: Coping with floods* (eds. ROSSI, G.; HARMONCIOGLU, N.; YEVJEVICH, V.), Kluwer, Dordrecht, p. 111-134, 1994.

SCHAEFER, M.C., Regional analysis of precipitation annual maxima in Washington State. *Water Resources Research*, v.26, n.1, p.119-131, 1990.

SHAHIN, M., VAN OORSCHOT, H. J. L. e DE LANGE, S. J. *Statistical Analysis in Water Resources Engineering*. Rotterdam: A. A. Balkema, 1993.

SIEGEL, S. *Nonparametric Statistics for the Behavioral Sciences*. New York: McGraw-Hill, 1956.

SINGER, J. M. e ANDRADE, D. F. Regression models in connection with random digits. *Biometrics*, 53, pp. 729-735, 1997.

SMIRNOV, N. Table for estimating the goodness of fit of empirical distributions. *Annals of Mathematical Statistics*, 19, pp. 279-281, 1948.

SMITH, R. L. Threshold models for sample extremes, in *Statistical Extremes and Applications*, ed. J. Tiago de Oliveira, 621-638, D. Reidel, Hingham, Ma., EUA, 1984.

SMITH, J. A.. Regional flood frequency analysis using extreme order statistics of the annual peak record. *Water Resources Research*, v.25, n.2, p. 311-317, 1989.

STEDINGER, J. R., VOGEL, R. M. e FOUFOULA-GEORGIOU, E. Frequency Analysis of Extreme Events, capítulo 18 in *Handbook of Hydrology*, MAIDMENT, D. R. (ed.), New York: McGraw-Hill, 1993.

STURGES, H. A. The choice of a class interval. *Journal of the American Statistical Association*, 21, pp. 65-66, 1926.

SUTCLIFFE, J. V. The use of historical records in flood frequency analysis. *Journal of Hydrology,* 96, p. 159-171, 1987.

TAESOMBUT, V. e YEVJEVICH, V. *Use of partial flood series for estimating distributions of maximum annual flood peak*, Hydrology Paper 82, Colorado State University, Fort Collins, CO, EUA, 1978.

TASKER, G. D., Simplified testing of hydrologic regression regions. *Journal of Hydraulics Division*, ASCE, V. 108, n. 10, p. 1218-1222, 1982.

TODOROVIC, P. Stochastic models of floods. *Water Resources Research*, v. 14, n.2, p. 345-356, 1978.

TODOROVIC, P. e ZELENHASIC, E. A stochastic model for flood analysis. *Water Resources Research*, v. 6, n.6, p. 411-424, 1970.

TRYON R.C., *Cluster Analysis*, Edwards Brothers, Ann Arbor, MI, EUA,1939, *apud* Statsoft Inc., Electronic Statistics Textbook, Statsoft, Tulsa, OK, Estados Unidos (http://www.statsoft.com/textbook/stathome.html), 1997.

TUCCI, C.E. *Regionalização de vazões*. Porto Alegre, UFRGS/IPH , 2002.

TUKEY, J. W. *Exploratory Data Analysis*. Reading (MA): Addison Wesley, 1977.

U. S. WATER RESOURCES COUNCIL. *Guidelines for Determining Flood Flow Frequency – Bulletin 17B*. Washington (DC): US WRC, 1981.

VAN MONTFORT, M. A. J. e WITTER, J. V. The generalized Pareto distribution applied to rainfall depths. *Hydrological Sciences Journal*, v.31, n.2, p.151-162, 1986.

VOGEL, R. M. e FENNESSEY, N. M. L-moment diagrams should replace product-moment diagrams. *Water Resources Research*, 29(6), pp. 1745-1752, 1993.

VOGEL, R. M. e MCMARTIN, D. E. Probability plot goodness-of-fit and skewness estimation procedures for the Pearson type III distribution. *Water Resources Research*, 27(12), pp. 3149-3158, 1991.

WALD, A. e WOLFOWITZ, J. An exact test for randomness in the non-parametric case based on serial correlation. *Annals of Mathematical Statistics*, 14, pp. 378-388, 1943.

WARD J. H.. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, V. 58, p. 236, 1963, *apud* Statsoft Inc., Electronic Statistics Textbook, Statsoft, Tulsa, OK, Estados Unidos (http://www.statsoft.com/textbook/stathome.html), 1997.

WATT, W. E.; LATHEM, K. W.; NEILL, C. R.; RICHARDS, T. L. e ROUSSELE, J. *The Hydrology of Floods in Canada: A Guide to planning and Design*. National Research Council of Canada. 1988.

WAYLEN P. R. e WOO, M. K.. Regionalization and prediction of floods in the Fraser river catchment, *Water Resources Bulletin*, V. 20, n. 6, p. 941-949, 1984.

WHITE, E. L., Factor analysis of drainage basin properties : classification of flood behavior in terms of basin geomorphology. *Water Resources Bulletin*, V. 11, n. 4, p. 676-687, 1975.

WILTSHIRE, S. E., Grouping basins for regional flood frequency analysis. *Hydrological Sciences Journal*, V. 30, n. 1, p. 151-159, 1985.

WILTSHIRE, S. E., Identification of homogeneous regions for flood frequency analysis. *Journal of Hydrology*, V. 84, p. 287-302, 1986.

YEVJEVICH, V. M. *Probability and Statistics in Hydrology*. Fort Collins (CO): Water Resources Publications, 1972.

YEVJEVICH, V. M. Section 8-II Statistical and probability analysis of hydrological data. Part II Regression and correlation analysis. In: CHOW, V. T. *Handbook of applied hydrology*. Ed. McGraw-Hill. 1964.

ANEXOS

HIDROLOGIA ESTATÍSTICA

ANEXO 1A

Vazões médias mensais e anuais (m³/s) do Rio Paraopeba em Ponte Nova do Paraopeba (código 40800001) – Redução por ano civil.

Ano	Jan	Fev	Mar	Abr	Mai	Jun	Jul	Ago	Set	Out	Nov	Dez	Média
1938	175	161	128	96,1	88,6	66,8	56,4	57,5	54,3	71,2	102	195	104,3
1939	230	219	118	104	76,4	62,3	57,7	50,5	50,1	61,3	50	94,9	97,9
1940	109	136	161	77,9	61,6	51	41,1	33,9	35,8	51,6	142	169	89,2
1941	192	112	122	120	69,2	60,3	57,2	45,1	51,5	58,6	65,5	159	92,7
1942	199	115	149	88,2	71,6	58,5	51,2	42,6	44,6	67,4	106	183	98
1943	351	228	323	122	93,2	81,1	66,2	65,1	57,9	67,5	66,7	179	141,7
1944	107	185	132	89,9	70,7	59,5	50,4	43,2	35,7	44,3	60,5	94,5	81,1
1945	168	163	128	120	79,4	69,8	58,7	47,2	44,9	54,9	77,1	156	97,3
1946	144	81,9	105	93,2	65,5	53,1	44,4	37,3	36,1	52,1	70,8	81,1	72
1947	133	123	230	95,4	70,4	59,4	54,4	49,6	61,1	54,6	64,3	131	93,9
1948	106	124	129	68,6	50,9	45,6	37,9	31,1	27,8	33,3	92,6	259	83,8
1949	254	396	165	115	81,8	71,6	57,2	47,7	40,6	52,4	61,7	130	122,8
1950	155	161	107	81,3	64,1	54,2	46,3	39,7	39,2	46,6	117	140	87,6
1951	134	212	209	159	90,5	73,4	61,3	52,8	45,9	46,7	39,2	88,7	101
1952	161	234	183	112	72,9	67,2	51,8	43,4	43,9	41,6	67,5	95,8	97,8
1953	59,8	91,7	81,3	90,8	50,5	41	34,1	29,1	30,3	37,4	66,4	106	59,9
1954	64,4	91	53,8	71,8	45,3	33,1	26,6	21,6	19,1	28,3	69,6	68	49,4
1955	138	77,7	58	55,8	37,4	35,7	25,3	20,6	16,8	33,2	49,2	136	57
1956	120	70	102	49,5	45,9	45,9	35,1	30,4	30,2	24,8	36,8	228	68,2
1957	128	124	115	110	75,8	55	45,8	38	46,2	36	85,1	140	83,2
1958	89,3	107	75,6	63,1	56,4	43,1	45,4	33,5	39,3	60,1	51	63,3	60,6
1959	78,6	55,8	127	53,8	37,6	31,5	26,8	23,9	20,8	34	54,8	57,1	50,1
1960	127	118	142	61,7	50,5	41,6	35,1	27,6	25,7	34	41,9	119	68,7
1961	303	331	236	116	87,4	64,4	52,8	43,8	35,2	31,9	47,5	55,8	117,1
1962	146	165	87,6	58,2	47,7	42	33,9	28,5	33,3	43,9	66,7	210	80,2
1963	109	98,9	60,4	40,1	32,5	29,1	26,8	24,4	19,8	25,6	35,6	21,1	43,6
1964	120	179	72,2	46,5	39,1	30,9	33,8	26,1	20,5	44,1	70,8	119	66,8
1965	199	295	223	106	86,3	62,7	54,3	49,6	40,5	85,3	103	116	118,4
1966	326	210	139	89,8	68,9	56	48,5	39,2	35	59,3	118	135	110,4
1967	222	232	139	88	64,5	52,5	45,7	36,5	31,1	42,2	109	127	99,1

ANEXO 1A

Vazões médias mensais e anuais (m^3/s) do Rio Paraopeba em Ponte Nova do Paraopeba (código 40800001) – Redução por ano civil.

Ano	Jan	Fev	Mar	Abr	Mai	Jun	Jul	Ago	Set	Out	Nov	Dez	Média
1968	132	111	104	71,3	47,5	40,9	35,2	32,9	36,8	50,8	54,2	143	71,6
1969	90,4	73	62,9	46,3	35,6	37,6	30,3	27,3	24,8	51,2	115	157	62,6
1970	157	96,9	64,6	58,5	41,2	34,1	31,7	28,9	35,2	53,5	77,8	54,7	61,2
1971	36,9	30,4	37,4	26,4	19,5	25,6	17,5	14,1	19,9	36,4	111	186	46,8
1972	84,3	123	138	80	52,6	41,5	43,6	32,8	30,5	54,2	119	148	79
1973	192	160	159	90	67,3	55,3	48,8	39	34,5	52	91,4	166	96,3
1974	181	103	127	93,4	62	53,5	46,1	37,9	28,4	45,9	43,8	109	77,6
1975	168	130	69	66,4	49,1	35,2	41,1	29,5	23,6	39,5	97,3	82,5	69,3
1976	57,6	74	62,8	43,3	36,1	30,3	36,3	32,6	63,4	77,8	109	183	67,2
1977	170	135	98	80,5	52	44,8	37	32,4	38	28,8	62,3	88,9	72,4
1978	208	99,6	76,3	71,6	61,6	60,7	50,5	38,3	36,2	40,1	87,6	105	78
1979	239	478	198	115	88,8	72,9	63,4	55,1	59	48,3	93,7	190	141,8
1980	294	148	89,3	113	68,9	64,7	53,8	43,5	39,2	39,2	69,8	185	100,7
1981	181	105	98,1	66,8	55,9	54,4	41,2	38,2	33,7	64,4	142	169	87,4
1982	236	108	188	120	81,6	65,2	53,5	44,9	36,9	68,9	63,3	136	100,2
1983	323	264	238	189	123	121	89,2	70,3	84,8	107	140	253	166,9
1984	131	87,9	75,4	77	56,7	45,4	40,4	42,3	58,6	46,6	77,8	159	74,8
1985	311	224	279	148	102	79,8	66,6	56,2	55,1	63,1	92,2	124	133,4
1986	205	143	106	71,2	63,1	51,2	51,1	51,6	36,3	29,4	41,7	172	85,1
1987	140	104	132	89,2	66,4	55,3	43,8	36,6	42,7	37	49,4	151	78,9
1988	121	187	121	82,4	58,3	48,3	38,9	33,4	26,3	52,8	59,4	87,7	76,4
1989	75,3	91,2	105	48	36,9	40	36	35,5	32,3	53,6	72,6	144	64,2
1990	111	66	76,1	63,2	52,2	35	35,1	31,5	36,5	29,6	40,8	59,7	53,1
1991	332	226	171	129	81	58,8	48,4	39,3	41,5	60,3	68,2	90,8	112,2
1992	293	207	110	92,1	79,4	52,5	45	39,6	56,4	57,3	137	160	110,8
1993	171	145	124	108	65,1	59,5	44,5	38,7	38,3	66,4	54,1	72,1	82,2
1994	259	93,2	145	89,4	76,3	58	49,7	39,3	30,7	34,3	42	140	88,1
1995	86,1	152	118	82,4	61,7	47,4	39,6	29,8	28,3	61,9	71	192	80,9
1996	218	105	103	70,9	55,7	43,7	37,6	33,1	42,5	37,9	154	176	89,8
1997	456	139	158	112	77,8	71,6	50,7	40,2	45,4	48,9	53,9	125	114,9
1998	149	129	80,2	55,3	41,6	38,1	26,6	29	19	36,3	68,4	91	63,6
1999	115	53,8	159	56,4	35,8	29,1	25,2	19,1	18,6	22,9	61,7	91,2	57,3
Média	173,7	149,8	128,6	86,3	62,7	52,4	44,5	37,9	37,8	48,7	77,6	133	86,1

ANEXO 1B

Vazões médias mensais e anuais (m^3/s) do Rio Paraopeba em Ponte Nova do Paraopeba (código 40800001) – Redução por ano hidrológico (Outubro a Setembro)

Ano Inicial	Ano Final	Out	Nov	Dez	Jan	Fev	Mar	Abr	Mai	Jun	Jul	Ago	Set	Média
1938	1939	71,2	102	195	230	219	118	104	76,4	62,3	57,7	50,5	50,1	111,4
1939	1940	61,3	50	94,9	109	136	161	77,9	61,6	51	41,1	33,9	35,8	76,1
1940	1941	51,6	142	169	192	112	122	120	69,2	60,3	57,2	45,1	51,5	99,3
1941	1942	58,6	65,5	159	199	115	149	88,2	71,6	58,5	51,2	42,6	44,6	91,9
1942	1943	67,4	106	183	351	228	323	122	93,2	81,1	66,2	65,1	57,9	145,3
1943	1944	67,5	66,7	179	107	185	132	89,9	70,7	59,5	50,4	43,2	35,7	90,6
1944	1945	44,3	60,5	94,5	168	163	128	120	79,4	69,8	58,7	47,2	44,9	89,9
1945	1946	54,9	77,1	156	144	81,9	105	93,2	65,5	53,1	44,4	37,3	36,1	79,0
1946	1947	52,1	70,8	81,1	133	123	230	95,4	70,4	59,4	54,4	49,6	61,1	90,0
1947	1948	54,6	64,3	131	106	124	129	68,6	50,9	45,6	37,9	31,1	27,8	72,6
1948	1949	33,3	92,6	259	254	396	165	115	81,8	71,6	57,2	47,7	40,6	134,5
1949	1950	52,4	61,7	130	155	161	107	81,3	64,1	54,2	46,3	39,7	39,2	82,7
1950	1951	46,6	117	140	134	212	209	159	90,5	73,4	61,3	52,8	45,9	111,8
1951	1952	46,7	39,2	88,7	161	234	183	112	72,9	67,2	51,8	43,4	43,9	95,3
1952	1953	41,6	67,5	95,8	59,8	91,7	81,3	90,8	50,5	41	34,1	29,1	30,3	59,5
1953	1954	37,4	66,4	106	64,4	91	53,8	71,8	45,3	33,1	26,6	21,6	19,1	53,0
1954	1955	28,3	69,6	68	138	77,7	58	55,8	37,4	35,7	25,3	20,6	16,8	52,6
1955	1956	33,2	49,2	136	120	70	102	49,5	45,9	45,9	35,1	30,4	30,2	62,3
1956	1957	24,8	36,8	228	128	124	115	110	75,8	55	45,8	38	46,2	85,6
1957	1958	36	85,1	140	89,3	107	75,6	63,1	56,4	43,1	45,4	33,5	39,3	67,8
1958	1959	60,1	51	63,3	78,6	55,8	127	53,8	37,6	31,5	26,8	23,9	20,8	52,5
1959	1960	34	54,8	57,1	127	118	142	61,7	50,5	41,6	35,1	27,6	25,7	64,6
1960	1961	34	41,9	119	303	331	236	116	87,4	64,4	52,8	43,8	35,2	122,0
1961	1962	31,9	47,5	55,8	146	165	87,6	58,2	47,7	42	33,9	28,5	33,3	64,8
1962	1963	43,9	66,7	210	109	98,9	60,4	40,1	32,5	29,1	26,8	24,4	19,8	63,5
1963	1964	25,6	35,6	21,1	120	179	72,2	46,5	39,1	30,9	33,8	26,1	20,5	54,2
1964	1965	44,1	70,8	119	199	295	223	106	86,3	62,7	54,3	49,6	40,5	112,5
1965	1966	85,3	103	116	326	210	139	89,8	68,9	56	48,5	39,2	35	109,7
1966	1967	59,3	118	135	222	232	139	88	64,5	52,5	45,7	36,5	31,1	102,0

ANEXO 1B

Vazões médias mensais e anuais (m^3/s) do Rio Paraopeba em Ponte Nova do Paraopeba (código 40800001) – Redução por ano hidrológico (Outubro a Setembro)

Ano Inicial	Ano Final	Out	Nov	Dez	Jan	Fev	Mar	Abr	Mai	Jun	Jul	Ago	Set	Média
1967	1968	42,2	109	127	132	111	104	71,3	47,5	40,9	35,2	32,9	36,8	74,2
1968	1969	50,8	54,2	143	90,4	73	62,9	46,3	35,6	37,6	30,3	27,3	24,8	56,4
1969	1970	51,2	115	157	157	96,9	64,6	58,5	41,2	34,1	31,7	28,9	35,2	72,6
1970	1971	53,5	77,8	54,7	36,9	30,4	37,4	26,4	19,5	25,6	17,5	14,1	19,9	34,5
1971	1972	36,4	111	186	84,3	123	138	80	52,6	41,5	43,6	32,8	30,5	80,0
1972	1973	54,2	119	148	192	160	159	90	67,3	55,3	48,8	39	34,5	97,3
1973	1974	52	91,4	166	181	103	127	93,4	62	53,5	46,1	37,9	28,4	86,8
1974	1975	45,9	43,8	109	168	130	69	66,4	49,1	35,2	41,1	29,5	23,6	67,6
1975	1976	39,5	97,3	82,5	57,6	74	62,8	43,3	36,1	30,3	36,3	32,6	63,4	54,6
1976	1977	77,8	109	183	170	135	98	80,5	52	44,8	37	32,4	38	88,1
1977	1978	28,8	62,3	88,9	208	99,6	76,3	71,6	61,6	60,7	50,5	38,3	36,2	73,6
1978	1979	40,1	87,6	105	239	478	198	115	88,8	72,9	63,4	55,1	59	133,5
1979	1980	48,3	93,7	190	294	148	89,3	113	68,9	64,7	53,8	43,5	39,2	103,9
1980	1981	39,2	69,8	185	181	105	98,1	66,8	55,9	54,4	41,2	38,2	33,7	80,7
1981	1982	64,4	142	169	236	108	188	120	81,6	65,2	53,5	44,9	36,9	109,1
1982	1983	68,9	63,3	136	323	264	238	189	123	121	89,2	70,3	84,8	147,5
1983	1984	107	140	253	131	87,9	75,4	77	56,7	45,4	40,4	42,3	58,6	92,9
1984	1985	46,6	77,8	159	311	224	279	148	102	79,8	66,6	56,2	55,1	133,8
1985	1986	63,1	92,2	124	205	143	106	71,2	63,1	51,2	51,1	51,6	36,3	88,2
1986	1987	29,4	41,7	172	140	104	132	89,2	66,4	55,3	43,8	36,6	42,7	79,4
1987	1988	37	49,4	151	121	187	121	82,4	58,3	48,3	38,9	33,4	26,3	79,5
1988	1989	52,8	59,4	87,7	75,3	91,2	105	48	36,9	40	36	35,5	32,3	58,3
1989	1990	53,6	72,6	144	111	66	76,1	63,2	52,2	35	35,1	31,5	36,5	64,7
1990	1991	29,6	40,8	59,7	332	226	171	129	81	58,8	48,4	39,3	41,5	104,8
1991	1992	60,3	68,2	90,8	293	207	110	92,1	79,4	52,5	45	39,6	56,4	99,5
1992	1993	57,3	137	160	171	145	124	108	65,1	59,5	44,5	38,7	38,3	95,7
1993	1994	66,4	54,1	72,1	259	93,2	145	89,4	76,3	58	49,7	39,3	30,7	86,1
1994	1995	34,3	42	140	86,1	152	118	82,4	61,7	47,4	39,6	29,8	28,3	71,8
1995	1996	61,9	71	192	218	105	103	70,9	55,7	43,7	37,6	33,1	42,5	86,2
1996	1997	37,9	154	176	456	139	158	112	77,8	71,6	50,7	40,2	45,4	126,6
1997	1998	48,9	53,9	125	149	129	80,2	55,3	41,6	38,1	26,6	29	19	66,3
1998	1999	36,3	68,4	91	115	53,8	159	56,4	35,8	29,1	25,2	19,1	18,6	59,0
1999	2000	22,9	61,7	91,2	173,7	149,8	128,6	86,3	62,7	52,4	44,5	37,9	37,8	79,1

ANEXO 2

Vazões Médias Diárias Máximas Anuais (m³/s) do Rio Paraopeba em Ponte Nova do Paraopeba (código 40800001) - redução por ano hidrológico (Outubro a Setembro)

Ano Hidrológico	Vazão Máxima	Ano Hidrológico	Vazão Máxima	Ano Hidrológico	Vazão Máxima	Ano Hidrológico	Vazão Máxima
38/39	576,0	53/54	295,0	68/69	478,0	86/87	549,0
39/40	414,0	54/55	498,0	69/70	340,0	87/88	601,0
40/41	472,0	55/56	470,0	70/71	246,0	88/89	288,0
41/42	458,0	56/57	774,0	71/72	568,0	89/90	481,0
42/43	684,0	57/58	388,0	72/73	520,0	90/91	927,0
43/44	408,0	58/59	408,0	73/74	449,0	91/92	827,0
44/45	371,0	59/60	448,0	74/75	357,0	92/93	424,0
45/46	333,0	60/61	822,0	75/76	276,0	93/94	603,0
46/47	570,0	61/62	414,0	77/78	736,0	94/95	633,0
47/48	502,0	62/63	515,0	78/79	822,0	95/96	695,0
48/49	810,0	63/64	748,0	79/80	550,0	97/98	296,0
49/50	366,0	64/65	570,0	82/83	698,0	98/99	427,0
50/51	690,0	65/66	726,0	83/84	585,0	-	-
51/52	570,0	66/67	580,0	84/85	1017,0	-	-
52/53	288,0	67/68	450,0	85/86	437,0	-	-

Vazões Mínimas Anuais (m³/s) para diferentes durações do Rio Paropeba em Ponte Nova do Paraopeba (código 40800001)

Ano	1 Dia	2 Dias	5 Dias	7 Dias	Ano	1 Dia	2 Dias	5 Dias	7 Dias	Ano	1 Dia	2 Dias	5 Dias	7 Dias
1938	41,20	42,13	44,70	44,81	1958	28,20	28,97	29,42	30,04	1979	37,60	37,60	38,24	38,89
1939	34,50	34,83	35,10	35,21	1959	19,00	19,33	19,40	19,60	1980	28,00	28,53	28,92	28,87
1940	29,90	29,90	29,90	29,90	1960	20,80	20,80	21,20	21,46	1982	34,30	34,57	34,78	35,10
1941	36,40	37,33	38,28	39,11	1961	27,50	27,50	27,94	28,87	1984	28,00	28,77	29,22	29,41
1942	36,40	37,00	37,12	37,46	1962	25,30	25,77	26,14	26,53	1985	43,40	44,00	44,80	45,40
1943	48,00	50,00	51,40	52,14	1963	17,90	17,90	17,90	18,04	1986	22,70	25,17	25,70	26,14
1944	30,30	30,30	30,30	30,30	1964	18,00	18,13	18,16	18,31	1987	24,20	24,93	25,70	26,24
1945	37,30	38,27	39,04	39,79	1965	34,30	34,83	34,94	35,10	1988	22,00	22,70	22,98	23,43
1946	32,70	32,70	32,88	33,09	1966	32,00	32,00	32,00	32,10	1989	24,90	24,90	25,06	25,57
1947	38,30	38,60	39,24	39,93	1967	26,70	26,97	27,18	27,27	1990	19,80	20,50	21,10	21,36
1948	24,00	24,67	25,20	25,14	1968	27,50	27,50	28,10	28,36	1991	31,90	31,90	32,22	32,36
1949	35,10	35,37	35,74	36,01	1969	21,20	21,20	21,44	21,70	1992	33,50	35,37	36,74	37,23
1950	33,80	34,00	34,16	34,33	1970	25,40	25,40	25,64	25,84	1993	28,80	29,03	29,84	30,13
1951	30,70	30,70	31,02	31,16	1971	12,80	12,80	12,80	12,86	1994	24,00	25,07	25,60	25,83
1952	31,20	32,20	32,26	33,07	1972	24,00	24,00	24,40	24,76	1995	21,60	21,87	22,24	22,51
1953	23,40	24,03	24,96	25,90	1973	30,70	30,70	31,02	31,16	1996	25,60	27,23	27,88	28,29
1954	17,90	17,90	18,10	18,11	1974	24,10	24,40	24,64	24,74	1997	29,20	29,73	30,44	30,97
1955	15,20	15,20	15,20	15,27	1975	21,70	21,70	21,70	21,70	1998	15,20	15,80	16,14	16,29
1956	20,80	21,47	21,70	21,73	1976	23,40	23,93	24,04	24,29	1999	11,80	11,97	12,42	12,77
1957	24,00	24,67	25,50	26,07	1978	23,40	24,17	24,62	24,93					

ANEXO 3

Alturas de precipitação diária máximas anuais (mm) observadas na estação pluviométrica de Ponte Nova do Paraopeba (código 19440004) - redução por ano hidrológico (Outubro a Setembro)

Ano Hidrológico	Altura Diária Máxima	Ano Hidrológico	Altura Diária Máxima	Ano Hidrológico	Altura Diária Máxima	Ano Hidrológico	Altura Diária Máxima
41/42	68,8	56/57	69,3	71/72	70,3	86/87	109
42/43	-	57/58	54,3	72/73	81,3	87/88	88
43/44	-	58/59	36	73/74	85,3	88/89	99,6
44/45	67,3	59/60	64,2	74/75	58,4	89/90	74
45/46	-	60/61	83,4	75/76	66,3	90/91	94
46/47	70,2	61/62	64,2	76/77	91,3	91/92	99,2
47/48	113,2	62/63	76,4	77/78	72,8	92/93	101,6
48/49	79,2	63/64	159,4	78/79	100	93/94	76,6
49/50	61,2	64/65	62,1	79/80	78,4	94/95	84,8
50/51	66,4	65/66	78,3	80/81	61,8	95/96	114,4
51/52	65,1	66/67	74,3	81/82	83,4	96/97	-
52/53	115	67/68	41	82/83	93,4	97/98	95,8
53/54	67,3	68/69	101,6	83/84	99	98/99	65,4
54/55	102,2	69/70	85,6	84/85	133	99/00	114,8
55/56	54,4	70/71	51,4	85/86	101	-	-

ANEXO 4 Matemática: alguns tópicos importantes

A4.1 – Contagem

Em certas situações, o cálculo de probabilidades exige a contagem do número possível de modos de se selecionar uma amostra de *k* itens, de um conjunto de *n* elementos passíveis de serem sorteados. A especificação do número de tais possibilidades pode ser facilitada com o emprego de algumas definições e fórmulas da análise combinatória.

A seleção, ou a amostragem, dos *k* itens pode ser realizada *com reposição*, quando cada item escolhido pode ser novamente sorteado, ou *sem reposição*, em caso contrário. Além disso, a *ordem* com que os distintos itens são sorteados pode, ou não, ser um fator importante. Como resultando, os seguintes quatro tipos de amostragem são possíveis: com ordem e com reposição; com ordem e sem reposição; sem ordem e com reposição; e sem ordem e sem reposição.

No caso de amostragem com ordem e com reposição, o primeiro item deve ser sorteado das *n* possibilidades que constituem a população. Em seguida, o primeiro item sorteado retorna à população e, tal como anteriormente, o segundo sorteio é feito de um universo de *n* itens. Prosseguindo com esse mesmo raciocínio, verifica-se que o número de possibilidades de se realizar o sorteio de *k* itens de *n* possíveis, com ordem e com reposição, é n^k.

Se o primeiro item sorteado não retornar à população para o próximo sorteio, o número de possibilidades para o segundo item é (*n*-1). O terceiro será sorteado em meio a (*n*-2) possibilidades, o quarto entre (*n*-3) e assim, sucessivamente, até o *k*-ésimo item. Portanto, o número de possibilidades de se realizar o sorteio de *k* itens de *n* possíveis, com ordem e sem reposição, é $n(n-1)(n-2)...(n-k+1)$. Essa expressão é equivalente à formula do número de *arranjos* da análise combinatória, ou seja,

$$A_{n,k} = \frac{n!}{(n-k)!} \qquad \text{(A4.1)}$$

Quando a ordem do sorteio não é importante, a amostragem sem reposição é semelhante ao caso ordenado à exceção do fato que os itens sorteados podem ser arranjados em *k*! modos diferentes. Em outras palavras, para calcular o número

de possibilidades para a amostragem sem ordem e sem reposição, é necessário deduzir da expressão A4.1 os $k!$ sorteios que irá conter os mesmos elementos. Portanto, o número de possibilidades de se realizar o sorteio de k itens de n possíveis, sem ordem e sem reposição, é $A_{n,k}/k!$. Essa expressão é equivalente à formula do número de *combinações* da análise combinatória, ou seja,

$$C_{n,k} = \binom{n}{k} = \frac{n!}{(n-k)!\,k!} \tag{A4.2}$$

Finalmente, quando a ordem do sorteio não é importante, mas a amostragem é realizada com reposição, o número de possibilidades é equivalente à da seleção, sem ordem e sem reposição, de k itens entre (n+k-1) possíveis. Em outras palavras, tudo se passa como se a população sofresse o acréscimo de (k-1) itens adicionais. Portanto, o número de possibilidades de se realizar o sorteio de k itens de n possíveis, sem ordem e com reposição, é dado por

$$C_{n+k-1,k} = \binom{n+k-1}{k} = \frac{(n+k-1)!}{(n-1)!\,k!} \tag{A4.3}$$

O operador fatorial, presente nas diversas equações da análise combinatória, pode ser aproximado pela fórmula de Stirling, a qual é expressa por

$$n! \cong \frac{\sqrt{2\pi}\, n^{n+1/2}}{e^n} \tag{A4.4}$$

Haan (1977) aponta que o erro de aproximação pela fórmula de Stirling é inferior a 1%, para n = 10, e decresce, quando n aumenta.

A4.2 – A Série de MacLaurin

Se uma função $f(x)$ possui derivadas contínuas até a ordem (n+1), então esta função pode ser expandida do seguinte modo:

$$f(x) = f(a) + f'(a)(x-a) + \frac{f''(a)(x-a)^2}{2!} + \cdots \frac{f^{(n)}(a)(x-a)^n}{n!} + R_n \tag{A4.5}$$

onde R_n denota o *resto*, após a expansão de (n+1) termos, sendo expresso por

$$R_n = \int_a^x f^{(n+1)}(w)\frac{(x-w)}{n!}dw = \frac{f^{(n+1)}(\tau)(x-a)^{n+1}}{(n+1)!} \qquad a < \tau < x \tag{A4.6}$$

Se a expansão dada pela equação A4.5 converge, dentro de um certo intervalo de variação de x, ou seja, se $\lim_{n\to\infty} R_n = 0$, ela é denominada a *série de Taylor* de $f(x)$, em torno de a.

Se $a = 0$, a expansão é denominada *série de MacLaurin*, sendo formalmente expressa por

$$f(x) = f(0) + f'(0)x + \frac{f''(0)}{2!}x^2 + \cdots \tag{A4.7}$$

A série de MacLaurin é, portanto, um tipo de expansão em série, na qual todos os termos são potências inteiras não-negativas da variável em questão. Apresenta-se, a seguir, alguns exemplos de expansão de funções simples por meio da série de MacLaurin:

$$cos(x) = 1 - \frac{x^2}{2} + \frac{x^4}{24} - \frac{x^6}{720} - \cdots \qquad -\infty < x < \infty \tag{A4.8}$$

$$e^x = 1 + x + \frac{x^2}{2} + \frac{x^3}{6} + \frac{x^4}{24} + \cdots \qquad -\infty < x < \infty \tag{A4.9}$$

$$ln(1+x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \cdots \qquad -1 < x < 1 \tag{A4.10}$$

$$\frac{1}{1-x} = 1 + x + x^2 + x^3 + x^4 + \cdots \qquad -1 < x < 1 \tag{A4.11}$$

A4.3 – A Função Gama

A função Gama $\Gamma(z)$ é uma extensão do conceito de fatorial para números não inteiros. $\Gamma(z)$ é definida, para qualquer valor real $z > 0$, pela integral

$$\Gamma(z) = \int_0^{\infty} x^{z-1} e^{-x} dx \qquad \text{(A4.12)}$$

A função Gama é contínua e possui derivadas contínuas para qualquer ordem. Quando z tende a 0 ou $+\infty$, $\Gamma(z)$ tende a $+\infty$. Por integração por partes, é possível demonstrar a seguinte propriedade da função Gama:

$$\Gamma(z+1) = z\,\Gamma(z) \qquad \text{(A4.13)}$$

Se z é igual a um inteiro positivo n e uma vez que $\Gamma(1) = 1$, o uso repetido da propriedade expressa por A4.13 conduz a

$$\Gamma(n+1) = n! \qquad \text{(A4.14)}$$

Alguns valores notáveis da função Gama são: $\Gamma(2) = \Gamma(1) = 1$ e $\Gamma(0{,}5) = \sqrt{\pi}$ e $\Gamma(0{,}5) = \sqrt{\pi}$.

A função pode ser aproximada por diversas expressões. Uma das mais eficientes, com erros da ordem de 2×10^{-10}, é a aproximação de Lanczos, a qual é dada por

$$\Gamma(z) = \left[\frac{\sqrt{2\pi}}{z} \left(p_0 + \sum_{i=1}^{6} \frac{p_i}{z+i} \right) \right] (z+5{,}5)^{z+0{,}5}\, e^{-(z+5{,}5)} \qquad \text{(A4.15)}$$

com

$$p_0 = 1{,}000000000190015$$
$$p_1 = 76{,}18009172947146$$
$$p_2 = -86{,}50532032941677$$
$$p_3 = 24{,}01409824083091$$
$$p_4 = -1{,}231739572450155$$
$$p_5 = 1{,}208650973866179 \times 10^{-3}$$
$$p_6 = -5{,}395239384953 \times 10^{-6}$$

A4.4 – A Função Beta

A função Beta, denotada por $B(z,w)$, para quaisquer números reais positivos z e w, é definida pela integral

$$B(z,w) = \int_0^1 x^{z-1}(1-x)^{w-1}\,dx \tag{A4.16}$$

Cramér (1946) demonstrou a seguinte importante relação entre as funções Beta e Gama:

$$B(z,w) = \frac{\Gamma(z)\Gamma(w)}{\Gamma(z+w)} \tag{A4.17}$$

A partir dessa relação e da aproximação de Lanczos, dada pela equação A4.15, torna-se possível avaliar a função Beta, para quaisquer números reais z e w.

ANEXO 5

Função Gama $\Gamma(t)$

$$\Gamma(t) = \int_0^{\infty} e^{-x} x^{t-1} dx$$

t	$\Gamma(t)$	t	$\Gamma(t)$	t	$\Gamma(t)$	t	$\Gamma(t)$
1,00	1,00000	1,25	0,90640	1,50	0,88623	1,75	0,91906
1,01	0,99433	1,26	0,90440	1,51	0,88659	1,76	0,92137
1,02	0,98884	1,27	0,90250	1,52	0,88704	1,77	0,92376
1,03	0,98355	1,28	0,90072	1,53	0,88757	1,78	0,92623
1,04	0,97844	1,29	0,89904	1,54	0,88818	1,79	0,92877
1,05	0,97350	1,30	0,89747	1,55	0,88887	1,80	0,93138
1,06	0,96874	1,31	0,89600	1,56	0,88964	1,81	0,93408
1,07	0,96415	1,32	0,89464	1,57	0,89049	1,82	0,93685
1,08	0,95973	1,33	0,89338	1,58	0,89142	1,83	0,93969
1,09	0,95546	1,34	0,89222	1,59	0,89243	1,84	0,94261
1,10	0,95135	1,35	0,89115	1,60	0,89352	1,85	0,94561
1,11	0,94739	1,36	0,89018	1,61	0,89468	1,86	0,94869
1,12	0,94359	1,37	0,89931	1,62	0,89592	1,87	0,95184
1,13	0,93993	1,38	0,88854	1,63	0,89724	1,88	0,95507
1,14	0,93642	1,39	0,88785	1,64	0,89864	1,89	0,95838
1,15	0,93304	1,40	0,88726	1,65	0,90012	1,90	0,96177
1,16	0,92980	1,41	0,88676	1,66	0,90167	1,91	0,96523
1,17	0,92670	1,42	0,88636	1,67	0,90330	1,92	0,96878
1,18	0,92373	1,43	0,88604	1,68	0,90500	1,93	0,97240
1,19	0,92088	1,44	0,88580	1,69	0,90678	1,94	0,97610
1,20	0,91817	1,45	0,88565	1,70	0,90864	1,95	0,97988
1,21	0,91558	1,46	0,88560	1,71	0,91057	1,96	0,98374
1,22	0,91311	1,47	0,88563	1,72	0,91258	1,97	0,98768
1,23	0,91075	1,48	0,88575	1,73	0,91466	1,98	0,99171
1,24	0,90852	1,49	0,88595	1,74	0,91683	1,99	0,99581

Observações:

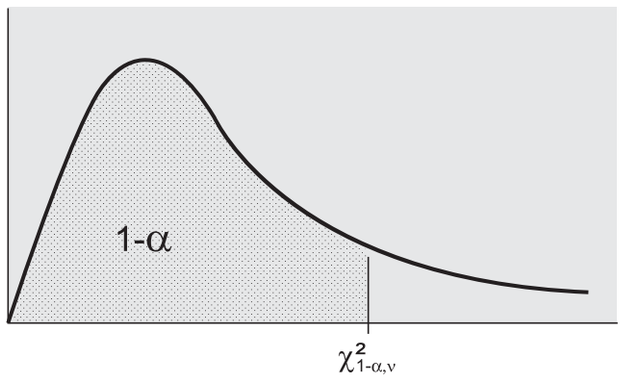
• Para outros valores de *t*, usar a propriedade $\Gamma(t+1) = t\,\Gamma(t)$

• Para valores positivos elevados de *t*, pode-se usar a aproximação de Stirling:

$$\Gamma(t) \approx t^t e^{-t} \sqrt{\frac{2\pi}{t}} \left(1 + \frac{1}{12t} + \frac{1}{288t^2} - \frac{139}{51840t^3} - \frac{571}{2488320t^4} + \ldots\right)$$

ANEXO 6

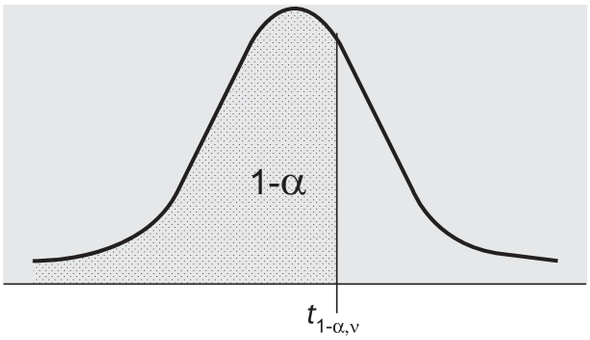
Quantis $\chi^2_{1-\alpha,\nu}$ da distribuição do Qui-Quadrado, com ν graus de liberdade



ν	$\chi^2_{0,995,\nu}$	$\chi^2_{0,99,\nu}$	$\chi^2_{0,975,\nu}$	$\chi^2_{0,95,\nu}$	$\chi^2_{0,90,\nu}$	$\chi^2_{0,10,\nu}$	$\chi^2_{0,05,\nu}$	$\chi^2_{0,025,\nu}$	$\chi^2_{0,01,\nu}$	$\chi^2_{0,005,\nu}$
1	7,88	6,63	5,02	3,84	2,71	0,0158	0,0039	0,0010	0,0002	0,0000
2	10,6	9,21	7,38	5,99	4,61	0,211	0,103	0,0506	0,0201	0,0100
3	12,8	11,3	9,35	7,81	6,25	0,584	0,352	0,216	0,115	0,072
4	14,9	13,3	11,1	9,49	7,78	1,06	0,711	0,484	0,297	0,207
5	16,7	15,1	12,8	11,1	9,24	1,61	1,15	0,831	0,554	0,412
6	18,5	16,8	14,4	12,6	10,6	2,20	1,64	1,24	0,872	0,676
7	20,3	18,5	16,0	14,1	12,0	2,83	2,17	1,69	1,24	0,989
8	22,0	20,1	17,5	15,5	13,4	3,49	2,73	2,18	1,65	1,34
9	23,6	21,7	19,0	16,9	14,7	4,17	3,33	2,70	2,09	1,73
10	25,2	23,2	20,5	18,3	16,0	4,87	3,94	3,25	2,56	2,16
11	26,8	24,7	21,9	19,7	17,3	5,58	4,57	3,82	3,05	2,60
12	28,3	26,2	23,3	21,0	18,5	6,30	5,23	4,40	3,57	3,07
13	29,8	27,7	24,7	22,4	19,8	7,04	5,89	5,01	4,11	3,57
14	31,3	29,1	26,1	23,7	21,1	7,79	6,57	5,63	4,66	4,07
15	32,8	30,6	27,5	25,0	22,3	8,55	7,26	6,26	5,23	4,60
16	34,3	32,0	28,8	26,3	23,5	9,31	7,96	6,91	5,81	5,14
17	35,7	33,4	30,2	27,6	24,8	10,1	8,67	7,56	6,41	5,70
18	37,2	34,8	31,5	28,9	26,0	10,9	9,39	8,23	7,01	6,26
19	38,6	36,2	32,9	30,1	27,2	11,7	10,1	8,91	7,63	6,84
20	40,0	37,6	34,2	31,4	28,4	12,4	10,9	9,59	8,26	7,43
21	41,4	38,9	35,5	32,7	29,6	13,2	11,6	10,3	8,90	8,03
22	42,8	40,3	36,8	33,9	30,8	14,0	12,3	11,0	9,54	8,64
23	44,2	41,6	38,1	35,2	32,0	14,8	13,1	11,7	10,2	9,26
24	45,6	43,0	39,4	36,4	33,2	15,7	13,8	12,4	10,9	9,89
25	46,9	44,3	40,6	37,7	34,4	16,5	14,6	13,1	11,5	10,5
26	48,3	45,6	41,9	38,9	35,6	17,3	15,4	13,8	12,2	11,2
27	49,6	47,0	43,2	40,1	36,7	18,1	16,2	14,6	12,9	11,8
28	51,0	48,3	44,5	41,3	37,9	18,9	16,9	15,3	13,6	12,5
29	52,3	49,6	45,7	42,6	39,1	19,8	17,7	16,0	14,3	13,1
30	53,7	50,9	47,0	43,8	40,3	20,6	18,5	16,8	15,0	13,8
40	66,8	63,7	59,3	55,8	51,8	29,1	26,5	24,4	22,2	20,7
50	79,5	76,2	71,4	67,5	63,2	37,7	34,8	32,4	29,7	28,0
60	92,0	88,4	83,3	79,1	74,4	46,5	43,2	40,5	37,5	35,5
70	104,2	100,4	95,0	90,5	85,5	55,3	51,7	48,8	45,4	43,3
80	116,3	112,3	106,6	101,9	96,6	64,3	60,4	57,2	53,5	51,2
90	128,3	124,1	118,1	113,1	107,6	73,3	69,1	65,6	61,8	59,2
100	140,2	135,8	129,6	124,3	118,5	82,4	77,9	74,2	70,1	67,3

ANEXO 7

Quantis $t_{1-\alpha,\nu}$ da distribuição de t de Student, com ν graus de liberdade



ν	$t_{0,995,\nu}$	$t_{0,99,\nu}$	$t_{0,975,\nu}$	$t_{0,95,\nu}$	$t_{0,90,\nu}$	$t_{0,80,\nu}$	$t_{0,75,\nu}$	$\chi^2 t_{0,70,\nu}$	$t_{0,60,\nu}$	$t_{0,55,\nu}$
1	63,66	31,82	12,71	6,31	3,08	1,376	1,000	0,727	0,325	0,158
2	9,92	6,96	4,30	2,92	1,89	1,061	0,816	0,617	0,289	0,142
3	5,84	4,54	3,18	2,35	1,64	0,978	0,765	0,584	0,277	0,137
4	4,60	3,75	2,78	2,13	1,53	0,941	0,741	0,569	0,271	0,134
5	4,03	3,36	2,57	2,02	1,48	0,920	0,727	0,559	0,267	0,132
6	3,71	3,14	2,45	1,94	1,44	0,906	0,718	0,553	0,265	0,131
7	3,50	3,00	2,36	1,90	1,42	0,896	0,711	0,549	0,263	0,130
8	3,36	2,90	2,31	1,86	1,40	0,889	0,706	0,546	0,262	0,130
9	3,25	2,82	2,26	1,83	1,38	0,883	0,703	0,543	0,261	0,129
10	3,17	2,76	2,23	1,81	1,37	0,879	0,700	0,542	0,260	0,129
11	3,11	2,72	2,20	1,80	1,36	0,876	0,697	0,540	0,260	0,129
12	3,06	2,68	2,18	1,78	1,36	0,873	0,695	0,539	0,259	0,128
13	3,01	2,65	2,16	1,77	1,35	0,870	0,694	0,538	0,259	0,128
14	2,98	2,62	2,14	1,76	1,34	0,868	0,692	0,537	0,258	0,128
15	2,95	2,60	2,13	1,75	1,34	0,866	0,691	0,536	0,258	0,128
16	2,92	2,58	2,12	1,75	1,34	0,865	0,690	0,535	0,258	0,128
17	2,90	2,57	2,11	1,74	1,33	0,863	0,689	0,534	0,257	0,128
18	2,88	2,55	2,10	1,73	1,33	0,862	0,688	0,534	0,257	0,127
19	2,86	2,54	2,09	1,73	1,33	0,861	0,688	0,533	0,257	0,127
20	2,84	2,53	2,09	1,72	1,32	0,860	0,687	0,533	0,257	0,127
21	2,83	2,52	2,08	1,72	1,32	0,859	0,686	0,532	0,257	0,127
22	2,82	2,51	2,07	1,72	1,32	0,858	0,686	0,532	0,256	0,127
23	2,81	2,50	2,07	1,71	1,32	0,858	0,685	0,532	0,256	0,127
24	2,80	2,49	2,06	1,71	1,32	0,857	0,685	0,531	0,256	0,127
25	2,79	2,48	2,06	1,71	1,32	0,856	0,684	0,531	0,256	0,127
26	2,78	2,48	2,06	1,71	1,32	0,856	0,684	0,531	0,256	0,127
27	2,77	2,47	2,05	1,70	1,31	0,855	0,684	0,531	0,256	0,127
28	2,76	2,47	2,05	1,70	1,31	0,855	0,683	0,530	0,256	0,127
29	2,76	2,46	2,04	1,70	1,31	0,854	0,683	0,530	0,256	0,127
30	2,75	2,46	2,04	1,70	1,31	0,854	0,683	0,530	0,256	0,127
40	2,70	2,42	2,02	1,68	1,30	0,851	0,681	0,529	0,255	0,126
60	2,66	2,39	2,00	1,67	1,30	0,848	0,679	0,527	0,254	0,126
120	2,62	2,36	1,98	1,66	1,29	0,845	0,677	0,526	0,254	0,126
∞	2,58	2,33	1,96	1,645	1,28	0,842	0,674	0,524	0,253	0,126

ANEXO 8

Função F de probabilidades acumuladas, com $\gamma_1 = m$ (g.l. do numerador) e $\gamma_2 = n$ (g.l. do denominador)

1 - α	n	m = 1	m = 2	m = 3	m = 4	m = 5	m = 6	m = 7	m = 8	m = 9	m = 10	m = 12	m = 15	m = 20	m = 30	m = 60	m = 120	m = ∞
0,9	1	39,86	49,50	53,59	55,83	57,24	58,20	58,91	59,44	59,86	60,19	60,71	61,22	61,74	62,26	62,79	63,06	63,32
0,95	1	161,45	199,50	215,71	224,58	230,16	233,99	236,77	238,88	240,54	241,88	243,90	245,95	248,02	250,10	252,20	253,25	254,29
0,975	1	647,79	799,48	864,15	899,60	921,83	937,11	948,20	956,64	963,28	968,63	976,72	984,87	993,08	1001	1010	1014	1018
0,99	1	4052	4999	5404	5624	5764	5859	5928	5981	6022	6056	6107	6157	6209	6260	6313	6340	6366
0,995	1	16212	19997	21614	22501	23056	23440	23715	23924	24091	24222	24427	24632	24837	25041	25254	25358	25462
0,9	2	8,53	9,00	9,16	9,24	9,29	9,33	9,35	9,37	9,38	9,39	9,41	9,42	9,44	9,46	9,47	9,48	9,49
0,95	2	18,51	19,00	19,16	19,25	19,30	19,33	19,35	19,37	19,38	19,40	19,41	19,43	19,45	19,46	19,48	19,49	19,50
0,975	2	38,51	39,00	39,17	39,25	39,30	39,33	39,36	39,37	39,39	39,40	39,41	39,43	39,45	39,46	39,48	39,49	39,50
0,99	2	98,50	99,00	99,16	99,25	99,30	99,33	99,36	99,38	99,39	99,40	99,42	99,43	99,45	99,47	99,48	99,49	99,50
0,995	2	198,50	199,01	199,16	199,24	199,30	199,33	199,36	199,38	199,39	199,39	199,42	199,43	199,45	199,48	199,48	199,49	199,51
0,9	3	5,54	5,46	5,39	5,34	5,31	5,28	5,27	5,25	5,24	5,23	5,22	5,20	5,18	5,17	5,15	5,14	5,13
0,95	3	10,13	9,55	9,28	9,12	9,01	8,94	8,89	8,85	8,81	8,79	8,74	8,70	8,66	8,62	8,57	8,55	8,53
0,975	3	17,44	16,04	15,44	15,10	14,88	14,73	14,62	14,54	14,47	14,42	14,34	14,25	14,17	14,08	13,99	13,95	13,90
0,99	3	34,12	30,82	29,46	28,71	28,24	27,91	27,67	27,49	27,34	27,23	27,05	26,87	26,69	26,50	26,32	26,22	26,13
0,995	3	55,55	49,80	47,47	46,20	45,39	44,84	44,43	44,13	43,88	43,68	43,39	43,08	42,78	42,47	42,15	41,99	41,83
0,9	4	4,54	4,32	4,19	4,11	4,05	4,01	3,98	3,95	3,94	3,92	3,90	3,87	3,84	3,82	3,79	3,78	3,76
0,95	4	7,71	6,94	6,59	6,39	6,26	6,16	6,09	6,04	6,00	5,96	5,91	5,86	5,80	5,75	5,69	5,66	5,63
0,975	4	12,22	10,65	9,98	9,60	9,36	9,20	9,07	8,98	8,90	8,84	8,75	8,66	8,56	8,46	8,36	8,31	8,26
0,99	4	21,20	18,00	16,69	15,98	15,52	15,21	14,98	14,80	14,66	14,55	14,37	14,20	14,02	13,84	13,65	13,56	13,47
0,995	4	31,33	26,28	24,26	23,15	22,46	21,98	21,62	21,35	21,14	20,97	20,70	20,44	20,17	19,89	19,61	19,47	19,33
0,9	5	4,06	3,78	3,62	3,52	3,45	3,40	3,37	3,34	3,32	3,30	3,27	3,24	3,21	3,17	3,14	3,12	3,11
0,95	5	6,61	5,79	5,41	5,19	5,05	4,95	4,88	4,82	4,77	4,74	4,68	4,62	4,56	4,50	4,43	4,40	4,37
0,975	5	10,01	8,43	7,76	7,39	7,15	6,98	6,85	6,76	6,68	6,62	6,52	6,43	6,33	6,23	6,12	6,07	6,02
0,99	5	16,26	13,27	12,06	11,39	10,97	10,67	10,46	10,29	10,16	10,05	9,89	9,72	9,55	9,38	9,20	9,11	9,02
0,995	5	22,78	18,31	16,53	15,56	14,94	14,51	14,20	13,96	13,77	13,62	13,38	13,15	12,90	12,66	12,40	12,27	12,15
0,9	6	3,78	3,46	3,29	3,18	3,11	3,05	3,01	2,98	2,96	2,94	2,90	2,87	2,84	2,80	2,76	2,74	2,72
0,95	6	5,99	5,14	4,76	4,53	4,39	4,28	4,21	4,15	4,10	4,06	4,00	3,94	3,87	3,81	3,74	3,70	3,67
0,975	6	8,81	7,26	6,60	6,23	5,99	5,82	5,70	5,60	5,52	5,46	5,37	5,27	5,17	5,07	4,96	4,90	4,85
0,99	6	13,75	10,92	9,78	9,15	8,75	8,47	8,26	8,10	7,98	7,87	7,72	7,56	7,40	7,23	7,06	6,97	6,88

ANEXO 8

Função F de probabilidades acumuladas, com $\gamma_1 = m$ (g.l. do numerador) e $\gamma_2 = n$ (g.l. do denominador)

1 - α	n	m = 1	m = 2	m = 3	m = 4	m = 5	m = 6	m = 7	m = 8	m = 9	m = 10	m = 12	m = 15	m = 20	m = 30	m = 60	m = 120	m = ∞
0,995	6	18,63	14,54	12,92	12,03	11,46	11,07	10,79	10,57	10,39	10,25	10,03	9,81	9,59	9,36	9,12	9,00	8,88
0,9	7	3,59	3,26	3,07	2,96	2,88	2,83	2,78	2,75	2,72	2,70	2,67	2,63	2,59	2,56	2,51	2,49	2,47
0,95	7	5,59	4,74	4,35	4,12	3,97	3,87	3,79	3,73	3,68	3,64	3,57	3,51	3,44	3,38	3,30	3,27	3,23
0,975	7	8,07	6,54	5,89	5,52	5,29	5,12	4,99	4,90	4,82	4,76	4,67	4,57	4,47	4,36	4,25	4,20	4,14
0,99	7	12,25	9,55	8,45	7,85	7,46	7,19	6,99	6,84	6,72	6,62	6,47	6,31	6,16	5,99	5,82	5,74	5,65
0,995	7	16,24	12,40	10,88	10,05	9,52	9,16	8,89	8,68	8,51	8,38	8,18	7,97	7,75	7,53	7,31	7,19	7,08
0,9	8	3,46	3,11	2,92	2,81	2,73	2,67	2,62	2,59	2,56	2,54	2,50	2,46	2,42	2,38	2,34	2,32	2,29
0,95	8	5,32	4,46	4,07	3,84	3,69	3,58	3,50	3,44	3,39	3,35	3,28	3,22	3,15	3,08	3,01	2,97	2,93
0,975	8	7,57	6,06	5,42	5,05	4,82	4,65	4,53	4,43	4,36	4,30	4,20	4,10	4,00	3,89	3,78	3,73	3,67
0,99	8	11,26	8,65	7,59	7,01	6,63	6,37	6,18	6,03	5,91	5,81	5,67	5,52	5,36	5,20	5,03	4,95	4,86
0,995	8	14,69	11,04	9,60	8,81	8,30	7,95	7,69	7,50	7,34	7,21	7,01	6,81	6,61	6,40	6,18	6,06	5,95
0,9	9	3,36	3,01	2,81	2,69	2,61	2,55	2,51	2,47	2,44	2,42	2,38	2,34	2,30	2,25	2,21	2,18	2,16
0,95	9	5,12	4,26	3,86	3,63	3,48	3,37	3,29	3,23	3,18	3,14	3,07	3,01	2,94	2,86	2,79	2,75	2,71
0,975	9	7,21	5,71	5,08	4,72	4,48	4,32	4,20	4,10	4,03	3,96	3,87	3,77	3,67	3,56	3,45	3,39	3,33
0,99	9	10,56	8,02	6,99	6,42	6,06	5,80	5,61	5,47	5,35	5,26	5,11	4,96	4,81	4,65	4,48	4,40	4,31
0,995	9	13,61	10,11	8,72	7,96	7,47	7,13	6,88	6,69	6,54	6,42	6,23	6,03	5,83	5,62	5,41	5,30	5,19
0,9	10	3,29	2,92	2,73	2,61	2,52	2,46	2,41	2,38	2,35	2,32	2,28	2,24	2,20	2,16	2,11	2,08	2,06
0,95	10	4,96	4,10	3,71	3,48	3,33	3,22	3,14	3,07	3,02	2,98	2,91	2,85	2,77	2,70	2,62	2,58	2,54
0,975	10	6,94	5,46	4,83	4,47	4,24	4,07	3,95	3,85	3,78	3,72	3,62	3,52	3,42	3,31	3,20	3,14	3,08
0,99	10	10,04	7,56	6,55	5,99	5,64	5,39	5,20	5,06	4,94	4,85	4,71	4,56	4,41	4,25	4,08	4,00	3,91
0,995	10	12,83	9,43	8,08	7,34	6,87	6,54	6,30	6,12	5,97	5,85	5,66	5,47	5,27	5,07	4,86	4,75	4,64
0,9	12	3,18	2,81	2,61	2,48	2,39	2,33	2,28	2,24	2,21	2,19	2,15	2,10	2,06	2,01	1,96	1,93	1,90
0,95	12	4,75	3,89	3,49	3,26	3,11	3,00	2,91	2,85	2,80	2,75	2,69	2,62	2,54	2,47	2,38	2,34	2,30
0,975	12	6,55	5,10	4,47	4,12	3,89	3,73	3,61	3,51	3,44	3,37	3,28	3,18	3,07	2,96	2,85	2,79	2,73
0,99	12	9,33	6,93	5,95	5,41	5,06	4,82	4,64	4,50	4,39	4,30	4,16	4,01	3,86	3,70	3,54	3,45	3,36
0,995	12	11,75	8,51	7,23	6,52	6,07	5,76	5,52	5,35	5,20	5,09	4,91	4,72	4,53	4,33	4,12	4,01	3,91
0,9	15	3,07	2,70	2,49	2,36	2,27	2,21	2,16	2,12	2,09	2,06	2,02	1,97	1,92	1,87	1,82	1,79	1,76
0,95	15	4,54	3,68	3,29	3,06	2,90	2,79	2,71	2,64	2,59	2,54	2,48	2,40	2,33	2,25	2,16	2,11	2,07
0,975	15	6,20	4,77	4,15	3,80	3,58	3,41	3,29	3,20	3,12	3,06	2,96	2,86	2,76	2,64	2,52	2,46	2,40
0,99	15	8,68	6,36	5,42	4,89	4,56	4,32	4,14	4,00	3,89	3,80	3,67	3,52	3,37	3,21	3,05	2,96	2,87
0,995	15	10,80	7,70	6,48	5,80	5,37	5,07	4,85	4,67	4,54	4,42	4,25	4,07	3,88	3,69	3,48	3,37	3,26

Função F de probabilidades acumuladas, com $\gamma_1 = m$ (g.l. do numerador) e $\gamma_2 = n$ (g.l. do denominador)

1 - α	n	m = 1	m = 2	m = 3	m = 4	m = 5	m = 6	m = 7	m = 8	m = 9	m = 10	m = 12	m = 15	m = 20	m = 30	m = 60	m = 120	m = ∞
0,9	20	2,97	2,59	2,38	2,25	2,16	2,09	2,04	2,00	1,96	1,94	1,89	1,84	1,79	1,74	1,68	1,64	1,61
0,95	20	4,35	3,49	3,10	2,87	2,71	2,60	2,51	2,45	2,39	2,35	2,28	2,20	2,12	2,04	1,95	1,90	1,84
0,975	20	5,87	4,46	3,86	3,51	3,29	3,13	3,01	2,91	2,84	2,77	2,68	2,57	2,46	2,35	2,22	2,16	2,09
0,99	20	8,10	5,85	4,94	4,43	4,10	3,87	3,70	3,56	3,46	3,37	3,23	3,09	2,94	2,78	2,61	2,52	2,42
0,995	20	9,94	6,99	5,82	5,17	4,76	4,47	4,26	4,09	3,96	3,85	3,68	3,50	3,32	3,12	2,92	2,81	2,69
0,9	30	2,88	2,49	2,28	2,14	2,05	1,98	1,93	1,88	1,85	1,82	1,77	1,72	1,67	1,61	1,54	1,50	1,46
0,95	30	4,17	3,32	2,92	2,69	2,53	2,42	2,33	2,27	2,21	2,16	2,09	2,01	1,93	1,84	1,74	1,68	1,62
0,975	30	5,57	4,18	3,59	3,25	3,03	2,87	2,75	2,65	2,57	2,51	2,41	2,31	2,20	2,07	1,94	1,87	1,79
0,99	30	7,56	5,39	4,51	4,02	3,70	3,47	3,30	3,17	3,07	2,98	2,84	2,70	2,55	2,39	2,21	2,11	2,01
0,995	30	9,18	6,35	5,24	4,62	4,23	3,95	3,74	3,58	3,45	3,34	3,18	3,01	2,82	2,63	2,42	2,30	2,18
0,9	60	2,79	2,39	2,18	2,04	1,95	1,87	1,82	1,77	1,74	1,71	1,66	1,60	1,54	1,48	1,40	1,35	1,29
0,95	60	4,00	3,15	2,76	2,53	2,37	2,25	2,17	2,10	2,04	1,99	1,92	1,84	1,75	1,65	1,53	1,47	1,39
0,975	60	5,29	3,93	3,34	3,01	2,79	2,63	2,51	2,41	2,33	2,27	2,17	2,06	1,94	1,82	1,67	1,58	1,48
0,99	60	7,08	4,98	4,13	3,65	3,34	3,12	2,95	2,82	2,72	2,63	2,50	2,35	2,20	2,03	1,84	1,73	1,60
0,995	60	8,49	5,79	4,73	4,14	3,76	3,49	3,29	3,13	3,01	2,90	2,74	2,57	2,39	2,19	1,96	1,83	1,69
0,9	120	2,75	2,35	2,13	1,99	1,90	1,82	1,77	1,72	1,68	1,65	1,60	1,55	1,48	1,41	1,32	1,26	1,19
0,95	120	3,92	3,07	2,68	2,45	2,29	2,18	2,09	2,02	1,96	1,91	1,83	1,75	1,66	1,55	1,43	1,35	1,26
0,975	120	5,15	3,80	3,23	2,89	2,67	2,52	2,39	2,30	2,22	2,16	2,05	1,94	1,82	1,69	1,53	1,43	1,31
0,99	120	6,85	4,79	3,95	3,48	3,17	2,96	2,79	2,66	2,56	2,47	2,34	2,19	2,03	1,86	1,66	1,53	1,38
0,995	120	8,18	5,54	4,50	3,92	3,55	3,28	3,09	2,93	2,81	2,71	2,54	2,37	2,19	1,98	1,75	1,61	1,44
0,9	∞	2,71	2,30	2,08	1,95	1,85	1,78	1,72	1,67	1,63	1,60	1,55	1,49	1,42	1,34	1,24	1,17	1,03
0,95	∞	3,84	3,00	2,61	2,37	2,22	2,10	2,01	1,94	1,88	1,83	1,75	1,67	1,57	1,46	1,32	1,22	1,05
0,975	∞	5,03	3,69	3,12	2,79	2,57	2,41	2,29	2,19	2,12	2,05	1,95	1,84	1,71	1,57	1,39	1,27	1,06
0,99	∞	6,64	4,61	3,79	3,32	3,02	2,81	2,64	2,51	2,41	2,32	2,19	2,04	1,88	1,70	1,48	1,33	1,07
0,995	∞	7,89	5,30	4,28	3,72	3,35	3,10	2,90	2,75	2,63	2,52	2,36	2,19	2,00	1,79	1,54	1,37	1,08

ANEXO 9

Modelos de séries de duração parcial

O conteúdo desse anexo é baseado no trabalho de Davis e Naghettini (2001).

A9.1 - Formulação teórica

Diversas variáveis hidrológicas/hidrometeorológicas variam no tempo de forma a constituir períodos de curta duração em que seus valores são muito elevados em relação à média, separados por períodos de valores inferiores à média ou mesmo nulos. Esse fato confere a essas variáveis a configuração característica de uma sucessão de *excedências*, em relação a um certo valor limiar de referência, a magnitude e número das quais são naturalmente aleatórios e passíveis de serem modeladas por um *processo estocástico bivariado*. Para maior clareza, considere que a Figura A9.1 representa um trecho da variação temporal de uma variável hidrológica X, ao longo do qual são identificadas todas as ocorrências superiores a um certo valor limiar u. Dessa forma, a i-ésima ocorrência de X superior a u terá o seu *valor máximo* denotado por X_i, resultado da soma de u e da excedência z_i, enquanto o tempo a ela associado será representado por T_i. Essa representação constitui o processo estocástico bivariado $\{T_i, X_i; i = 1,2,...\}$ a modelação do qual tem sido objeto de diversos estudos e investigações, entre as quais podem ser citadas as referências clássicas de Todorovic e Zelenhasic (1970), Gupta et al. (1976), Todorovic (1978) e North (1980). Outras referências importantes são os trabalhos de Taesombut e Yevjevich (1978), Smith (1984), Rosbjerg (1984) e Van Montfort e Witter (1986).

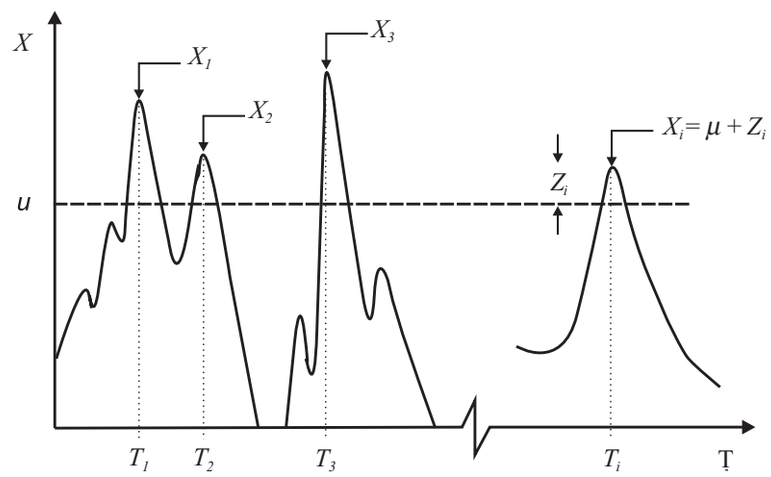


Figura A9.1 – Representação gráfica de processos estocásticos bivariados

Sob condições gerais, os eventos $\{T_i, X_i;\ i = 1,2,...\}$ podem ser representados pela classe de processos estocásticos compostos e não-homogêneos de Poisson. Para isso, dois requisitos são necessários :

1. O número $N_{\alpha\beta}$ de excedências em um intervalo de tempo $[\alpha,\beta]$ é uma variável aleatória discreta, cuja função massa de probabilidades é a de Poisson com intensidade ou taxa de ocorrência $\lambda(t)$ dependente do tempo. Logo, por definição,

$$P\left(N_{\alpha\beta} = n\right) = \frac{\left[\int_{\alpha}^{\beta} \lambda(t)dt\right]^{n} \exp\left[-\int_{\alpha}^{\beta} \lambda(t)dt\right]}{n!} \tag{A9.1}$$

2. $\{X_i\}$ é uma seqüência de variáveis aleatórias mutuamente independentes com distribuição de probabilidades dependente do tempo de ocorrência T_i.

Suponha que o intervalo $[\alpha,\beta]$ possa ser dividido em k_0 subintervalos, dentro de cada qual a distribuição de $\{X_i\}$ não dependa do tempo. Denotando o número de ocorrências dentro do j-ésimo subintervalo por N_j e o máximo de X correspondente por M_j, pode-se escrever

$$P\left(M_j \le x\right) = P(N_j = 0) + \sum_{n=1}^{\infty} P\left[\bigcap_{i=1}^{n}\left(X_{i,j} \le x\right) \bigcap \left(N_j = n\right)\right] \tag{A9.2}$$

onde $X_{i,j}$ denota a i-ésima ocorrência superior ao valor limiar u, dentro do j-ésimo subintervalo, e $\cap$ representa a simultaneidade ou interseção dos eventos indicados. Pela condição de independência mútua, imposta pelo requisito 2, segue-se que

$$P\left(M_j \le x\right) = P\left(N_j = 0\right) + \sum_{n=1}^{\infty} P\left(N_j = n\right)\left[H_{u,j}(x)\right]^{n} \tag{A9.3}$$

Nessa equação, $H_{u,j}$ representa a função de distribuição de probabilidades das ocorrências de X que excedem u, dentro do j-ésimo subintervalo. Substituindo a equação A9.1 na expressão A9.3, segue-se que

$$P\left(M_j \le x\right) = exp\left\{-\left[1 - H_{u,j}(x)\right] \int_{j} \lambda(t)dt\right\} \tag{A9.4}$$

Conforme North (1980), pode-se deduzir a distribuição do máximo $M_{\alpha\beta}$ ao longo

do intervalo [α,β], da seguinte forma:

$$P\left(M_{\alpha\beta} \le x\right) = P\left(\bigcap_{j=1}^{k_0} M_j \le x\right) \tag{A9.5}$$

ou, pela condição expressa pelo requisito 2,

$$P\left(M_{\alpha\beta} \le x\right) = \prod_{j=1}^{k_0} P\left(M_j \le x\right) \tag{A9.6}$$

onde o símbolo Π indica o produto das probabilidades indicadas. Combinando as equações A9.6 e A9.4, resulta que

$$P\left(M_{\alpha\beta} \le x\right) = exp\left\{-\sum_{j=1}^{k_0}\left[1 - H_{u,j}(x)\right]\int_j \lambda(t)\,dt\right\} \tag{A9.7}$$

Quando $k_0 \to \infty$, a equação A9.7 torna-se

$$P\left(M_{\alpha\beta} \le x\right) = exp\left\{-\int_{\alpha}^{\beta}\left[1 - H_u\left(x/t\right)\right]\lambda(t)\,dt\right\} \tag{A9.8}$$

Essa equação permite o cálculo da probabilidade do máximo $M_{\alpha\beta}$ dentro de qualquer intervalo de tempo [α,β]. Em geral, como o interesse se volta para a obtenção da distribuição dos *máximos anuais* $F_a(x)$, faz-se com que os limites $\alpha = 0$ e $\beta = 1$ representem respectivamente o início e o fim do ano, e a equação A9.8 torna-se

$$F_a(x) = exp\left\{-\int_0^1 \lambda(t)\left[1 - H_u\left(x/t\right)\right]dt\right\} \tag{A9.9}$$

Nessa equação, a distribuição de probabilidades das ocorrências de *Y* que excedem o valor limiar *u*, representada por $H_u(x/t)$, depende do tempo. Em geral, os diversos estudos e aplicações das séries de duração parcial sugerem não haver evidências empíricas suficientemente fortes para rejeitar a hipótese de que a distribuição $H_u(x/t)$ não depende do tempo. Se essa dependência não é

considerada, a equação A9.9 pode ser muito simplificada e a distribuição dos máximos anuais passa a ser

$$F_a(x) = exp\left\{-\left[1 - H_u(x)\right] \int_0^1 \lambda(t)\, dt\right\} = exp\{-\nu\left[1 - H_u(x)\right]\} \qquad \text{(A9.10)}$$

onde ν indica a *intensidade anual de ocorrências*. A equação A9.10 é a base para o emprego de séries de duração parcial e requer a estimação de ν e da função de distribuição $H_u(x)$. A intensidade ou taxa anual de ocorrências pode ser estimada pelo número médio anual de eventos que superam o valor limiar u; por exemplo, se houverem n anos de registros e forem selecionados os $2n$ maiores valores de X, a estimativa de ν é 2. A função de distribuição $H_u(x)$ está associada aos eventos que superaram o valor limiar u e pode ser prescrita pelo modelo paramétrico que melhor se ajustar aos dados amostrais.

A9.2 Condicionantes

Conforme sua construção teórica, descrita no item A9.1, a equação A9.10 pressupõe que as ocorrências superiores ao valor limiar u sejam independentes entre si e que o número dessas excedências seja uma variável de Poisson. Essas são condicionantes fundamentais para a correta modelação de séries de duração parcial e serão objeto de discussão nos sub-itens que se seguem.

A9.2.1 Independência Serial

A independência serial das ocorrências superiores ao valor limiar u é um pressuposto importante e sua confirmação empírica deve anteceder o uso do modelo estocástico bivariado, desenvolvido no item A9.1. Entretanto, algumas características próprias dos processos hidrológicos/hidrometeorológicos, bem como diversos estudos empíricos, indicam certas condições gerais sob as quais a hipótese de independência pode ser aceita. Embora não se possa estabelecer regras gerais, em se tratando de hidrogramas de cheia, os eventos devem ser selecionados de forma que estejam separados por um período de recessão suficientemente grande para que sejam considerados oriundos de episódios de chuva distintos. Da mesma forma, a seleção de eventos chuvosos deve ser condicionada à existência de um período significativo sem precipitação; no caso de chuvas intensas, por exemplo, é usual selecionar eventos separados por um

mínimo de 6 horas sem precipitação. Por tratarem-se de processos estocásticos contínuos, é de se esperar que a dependência serial contida nas séries hidrológicas/ hidrometeorológicas de duração parcial irá decrescer com o aumento do valor limiar *u* ou, contrariamente, irá crescer com o acréscimo da intensidade anual *v*. De fato, um valor limiar suficientemente elevado, faz com que o número de excedências se torne relativamente pequeno, enquanto o período entre os eventos que se torna relativamente grande; em consequência, as excedências tendem a se tornar independentes entre si. Taesombut e Yevjevich (1978) estudaram a variação do coeficiente de correlação serial de primeira ordem com o valor médio do número de excedências $\hat{\nu}$ para as vazões observadas em 17 estações fluviométricas dos Estados Unidos; concluíram que este coeficiente cresce com $\hat{\nu}$, mantendo-se dentro do limite de tolerância de 95% para $\hat{\nu} \leq 4{,}5$. Conclusões semelhantes foram obtidas por Madsen et al. (1993) a partir de séries de duração parcial de precipitação, observadas em diversas estações pluviométricas da Dinamarca.

A9.2.2 Distribuição de Freqüência do Número de Excedências

Para as variáveis hidrológicas/hidrometeorológicas, a premissa de que o número de excedências em relação a um valor limiar é uma variável de Poisson tem justificativas empíricas e teóricas. Do ponto de vista empírico, são inúmeros os estudos e aplicações em que essa premissa se verifica para valores limiares elevados [e.g. : Todorovic (1978), Taesombut e Yevjevich (1978), Correia (1983), Rosbjerg e Madsen (1992) e Madsen et al. (1993)]. As justificativas teóricas de se usar um processo de Poisson para modelar excedências mutuamente independentes provêm dos trabalhos de Cramér e Leadbetter (1967) e Leadbetter et al. (1983). Em particular, Cramér e Leadbetter (1967, p. 256) demonstraram que se um processo estocástico é Gaussiano, então, sob condições gerais, pode-se afirmar que o número de excedências em relação a um valor limiar *u* converge para um processo de Poisson, quando *u* tende para o infinito. Em relação a esse estudo, Todorovic (1978) argumenta que não há razão para presumir que esta conclusão estaria incorreta se o processo não for Gaussiano. Posteriormente, Leadbetter et al. (1983, p. 282) demonstraram que as excedências de alguns outros processos não Gaussianos também convergem para um processo de Poisson quando *u* aumenta.

Apesar das justificativas téoricas mencionadas, resta, do ponto de vista prático, perguntar quão elevado deve ser o valor limiar para que as excedências possam ser consideradas independentes e aproximadas por um processo de Poisson. Langbein (1949, p. 879) propôs o critério prático de se escolher o valor limiar de modo que, em média, não mais de duas ou três excedências anuais sejam selecionadas; em outras palavras, $\hat{\nu} \leq 3$. Por outro lado, Taesombut e Yevjevich (1978) concluíram pela aceitação da hipótese de Poisson nos casos em que a

relação entre a média e a variância de X é aproximadamente igual a um. Outros resultados obtidos por Taesombut e Yevjevich (1978) mostram também que, quando comparadas às séries de máximos anuais, as de duração parcial conduzem a menores erros de estimação de quantis de Gumbel *apenas* quando $\nu \geq 1{,}65$; concluem pela recomendação das séries de duração parcial para um número médio anual de excedências igual ou superior a 1,95. Cunnane (1973), por sua vez, já recomenda sem reservas o uso das séries de duração parcial, principalmente para amostras com menos de dez anos de registros. Apesar da dificuldade de se propor um critério geral, a experiência indica que especificar $\hat{\nu}$ entre 2 e 3 parece ser suficiente para auferir as vantagens de uso das séries de duração parcial e, ao mesmo tempo, garantir a independência serial dos eventos selecionados e, em muitos casos, a hipótese de Poisson. Entretanto, tal recomendação deve ser sempre sujeita a teste estatístico para verificar a sua adequação. O teste apropriado para se averiguar a veracidade da hipótese de Poisson foi primeiramente formulado por Cunnane (1979) e baseia-se na aproximação da distribuição de Poisson pela distribuição Normal. Considera-se que o número de excedências que ocorrem no ano k, denotado por m_k, segue uma distribuição Normal com média $\hat{\nu}$ e desvio padrão $\hat{\nu}$. Nessas condições, pode-se afirmar que a estatística

$$\gamma = \sum_{k=1}^{N} \left(\frac{m_k - \hat{\nu}}{\hat{\nu}} \right)^2 \qquad \text{(A9.11)}$$

segue uma distribuição do Qui-Quadrado com (N-1) graus de liberdade (η), onde N indica o número de anos de registros. Esse teste é considerado válido para os valores de $\hat{\nu}$ correntemente empregados e para tamanhos de amostra superiores a cinco. Deste modo, a hipótese de que as ocorrências são oriundas de um evento poissoniano é rejeitada, para um nível de significância α, se:

$$\gamma = \sum_{k=1}^{N} \left(\frac{m_k - \hat{\nu}}{\hat{\nu}} \right)^2 > \chi^2_{1-\alpha,\eta} \qquad \text{(A9.12)}$$

A.9.3 Modelo Poisson-GEV

Admitindo que a distribuição da série de duração parcial associada aos eventos que superaram o valor limiar u é a Generalizada de Eventos Extremos (GEV), pode ser deduzido o modelo Poisson-GEV a partir da equação A9.10 e reapresentada na seguinte forma:

$$F_a(x) = exp\{-\nu[1 - H_u(x)]\} \tag{A9.13}$$

onde $F_a(x)$ é a distribuição dos *máximos anuais* ; ν é a intensidade anual de ocorrências e $H_u(x)$ é a distribuição de probabilidades da série de duração parcial associada aos eventos que superaram o valor limiar *u*.

Desenvolvendo a equação A9.13 temos:

$$ln[F_a(x)] = -\nu[1 - H_u(x)] \tag{A9.14}$$

$$\frac{ln[F_a(x)]}{\nu} = H_u(x) - 1 \tag{A9.15}$$

$$H_u(x) = 1 + \frac{1}{\nu} ln[F_a(x)] \tag{A9.16}$$

Emprega-se, neste caso, distribuição Generalizada de Eventos Extremos (GEV):

$$H_u(x) = exp[-exp(-y)] \tag{A9.17}$$

$$y = \begin{cases} -\dfrac{ln\left[1 - \dfrac{k(x-\xi)}{\alpha}\right]}{k} & k \neq 0 \\ \dfrac{(x-\xi)}{\alpha} & k = 0 \end{cases} \tag{A9.18}$$

onde α é o parâmetro de escala; k é o parâmetro de forma e ξ é o parâmetro de posição.

Limites: para $k < 0$ $\xi + \dfrac{\alpha}{k} \leq x \leq \infty$, para $k > 0$ $-\infty \leq x \leq \xi + \dfrac{\alpha}{k}$ e para $k = 0$ $-\infty \leq x < \infty$

Igualando as equações A9.16 e A9.17 temos:

$$exp[-exp(-y)] = 1 + \frac{1}{\nu} ln[F_a(x)] \tag{A9.19}$$

$$[-exp(-y)] = ln\left\{1 + \frac{1}{\nu} ln[F_a(x)]\right\} \tag{A9.20}$$

$$exp(-y) = -ln\left\{1 + \frac{1}{\nu} ln[F_a(x)]\right\} \tag{A9.21}$$

$$-y = ln\left\{-ln\left[1+\frac{1}{\nu}ln[F_a(x)]\right]\right\} \quad \text{(A9.22)}$$

$$y = -ln\left\{-ln\left[1+\frac{1}{\nu}ln[F_a(x)]\right]\right\} \quad \text{(A9.23)}$$

Para $k = 0$, na distribuição GEV, $y = \frac{(x-\xi)}{\alpha}$. Substituindo y na equação A9.23:

$$\frac{(x-\xi)}{\alpha} = -ln\left\{-ln\left[1+\frac{1}{\nu}ln[F_a(x)]\right]\right\} \quad \text{(A9.24)}$$

$$x = \xi - \alpha\, ln\left\{-ln\left[1+\frac{1}{\nu}ln[F_a(x)]\right]\right\} \qquad \text{ou} \quad \text{(A9.25)}$$

$$x = \xi - \alpha\, ln\left\{-ln\left[\frac{\nu+ln[F_a(x)]}{\nu}\right]\right\} \quad \text{(A9.26)}$$

onde $F_a(x) = 1 - \frac{1}{T(anos)}$

Para $k \neq 0$, na distribuição GEV, $y = -\frac{ln\left[1-\frac{k(x-\xi)}{\alpha}\right]}{k}$. Substituindo y na equação A9.23:

$$-\frac{ln\left[1-\frac{k(x-\xi)}{\alpha}\right]}{k} = -ln\left\{-ln\left[1+\frac{1}{\nu}ln[F_a(x)]\right]\right\} \quad \text{(A9.27)}$$

$$ln\left[1-\frac{k(x-\xi)}{\alpha}\right] = k\, ln\left\{-ln\left[1+\frac{1}{\nu}ln[F_a(x)]\right]\right\} \quad \text{(A9.28)}$$

$$ln\left[1-\frac{k(x-\xi)}{\alpha}\right] = ln\left\{-ln\left[1+\frac{1}{\nu}ln[F_a(x)]\right]\right\}^k \quad \text{(A9.29)}$$

$$1-\frac{k(x-\xi)}{\alpha} = \left\{-ln\left[1+\frac{1}{\nu}ln[F_a(x)]\right]\right\}^k \quad \text{(A9.30)}$$

$$\frac{k(x-\xi)}{\alpha} = 1 - \left\{ -ln\left[1 + \frac{1}{\nu} ln[F_a(x)] \right] \right\}^k \quad \text{(A9.31)}$$

$$x = \xi + \frac{\alpha}{k} \left\{ 1 - \left\{ -ln\left[1 + \frac{1}{\nu} ln[F_a(x)] \right] \right\}^k \right\} \quad \text{ou} \quad \text{(A9.32)}$$

$$x = \xi + \frac{\alpha}{k} \left\{ 1 - \left\{ -ln\left[\frac{\nu + ln[F_a(x)]}{\nu} \right] \right\}^k \right\} \quad \text{(A9.33)}$$

onde $F_a(x) = 1 - \frac{1}{T(anos)}$

ANEXO 10

Transformações para linearização de diferentes tipos de funções

	Tipo de Função	Coordenadas		Equação na forma linear
		Abscissa	Ordenada	
1	$y = a + bx$	x	y	$[y] = a + b[x]$
2	$y = be^{ax}$	x	$\log y$	$[\log y] = \log b + (a \log c)[x]$
3	$y = ax^b$	$\log x$	$\log y$	$[\log y] = \log a + b[\log x]$
4	$y = a_0 + a_1x + a_2x^2$	$x - x_0$	$\frac{y - y_0}{x - x_0}$	$\left[\frac{y - y_0}{x - x_0}\right] = a_1 + 2a_1x_0 + a_2\,[(x - x_0)]$
5	$y = a + b/x$	$1/x$	y	$[y] = a + b[1/x]$
6	$y = x/(a + bx)$	x	x/y	$[x/y] = a + b[x]$
7	$y = a/(b + cx)$	x	$1/y$	$[1/y] = (b/a) + (c/a)\,[x]$
8	$y = c + be^{ax}$	x	$\log \frac{\Delta y}{\Delta x}$	$\left[\log \frac{dy}{dx} = \log(ab) + (a \log e)[x]\right]$
9	$y = c + ax^b$	$\log x$	$\log \frac{\Delta y}{\Delta x}$	$\left[\log \frac{dy}{dx} = \log(ab) + (b-1)[\log x]\right]$
10	$y = c + \frac{b}{x - a}$	$x - x_0$	$\frac{x - x_0}{y - y_0}$	$\left[\frac{x - x_0}{y - y_0}\right] = -\frac{a - x}{c - y_0} + \frac{1}{c - y_0}[x - x_0]$
11	$y = c + \frac{x}{a + bx}$	x	$\frac{x - x_0}{y - y_0}$	$\left[\frac{x - x_0}{y - y_0}\right] = (a + bx_0) + \frac{b(a + bx_0)}{a}[x]$
12	$y = d + cx + be^{ax}$	x	$\log \frac{\Delta^2 y}{\Delta x^2}$	$\lfloor y - be^{ax} \rfloor = d + c[x]$ $\left[\log \frac{d^2y}{dx^2} = \log(a^2b) + (a \log e)[x]\right]$
13	$y = dc^xb^m$, onde $m = a^x$	x	$\log \frac{\Delta^2 (\log y)}{\Delta x^2}$	$\left[\frac{\log d^2(\log y)}{dx^2}\right] = \log\left[\frac{(\log b)(\log a)^2}{(\log e)^2}\right] + (\log a)[x]$
14	$y = de^{cx} + be^{ax}$	$\frac{y_{k+1}}{y_k}$	$\frac{y_{k+2}}{y_k}$	$\lfloor ye^{-cx} \rfloor = d + b\lfloor e^{(a-c)x} \rfloor$ $\lfloor \log y - a^x \log b \rfloor = \log d + (\log e)[x]$ $\left[\frac{y_{k+2}}{y_k}\right] = -e^{(a+c)\Delta x} + \left(e^{a\Delta x} + e^{c\Delta x}\right)\left[\frac{y_{k+1}}{y_k}\right]$
15	$y = e^{ax}(d \cos bx + c \operatorname{sen} bx)$	$\frac{y_{k+1}}{y_k}$	$\frac{y_{k+2}}{y_k}$	$\left[\frac{y_{k+2}}{y_k}\right] = -e^{2a\Delta x} = \left(2e^{a\Delta x} \cos b\Delta x\right)\left[\frac{y_{k+1}}{y_k}\right]$ $\left[\frac{yc^{-ax}}{\cos bx}\right] = d + c[\tan bx]$

OBS: Nas equações 14 e 15, y_k, y_{k+1} e y_{k+2} são valores consecutivos para um incremento Δx.
Fonte: Yevjevich (1964), pág. 8-49.

ANEXO 11

Vazões mínimas anuais (m³/s) com 7 dias de duração de algumas estações da bacia do rio Paraopeba

N	Ano	40549998	Ano	40573000	Ano	40577000	Ano	40579995	Ano	40680000	Ano	40710000	Ano	40740000	Ano	40800001	Ano	40818000	Ano	40850000	Ano	40865001
1	1957	2,171	1945	1,877	1943	1,966	1939	3,930	1940	2,461	1966	18,071	1967	21,600	1938	44,814	1944	1,830	1968	42,17	1978	39,23
2	1958	2,576	1946	1,741	1944	1,601	1940	3,581	1942	2,847	1967	16,771	1968	22,000	1939	35,214	1945	1,716	1969	29,67	1980	41,33
3	1959	1,974	1947	1,991	1945	1,820	1941	5,160	1946	2,687	1968	15,800	1969	16,343	1940	29,900	1946	1,450	1970	36,39	1981	39,96
4	1960	2,271	1949	2,080	1946	1,450	1942	3,886	1948	1,960	1969	12,657	1970	18,100	1941	39,114	1947	1,553	1971	16,00	1982	59,97
5	1961	3,147	1950	1,580	1947	1,807	1944	5,950	1949	2,443	1970	13,757	1971	10,771	1942	37,457	1948	1,280	1972	34,14	1984	42,10
6	1962	2,414	1951	1,510	1948	1,453	1945	4,980	1950	2,250	1971	8,257	1972	19,500	1943	52,143	1949	1,884	1973	43,14	1986	36,60
7	1963	1,704	1952	1,880	1949	1,924	1946	4,054	1954	1,979	1972	14,886	1973	23,886	1944	30,300	1950	1,630	1974	32,44	1987	37,20
8	1965	2,770	1953	1,963	1950	1,600	1947	5,220	1955	1,451	1973	17,686	1976	18,171	1945	39,786	1951	1,463	1975	30,51	1988	36,70
9	1966	3,089	1954	1,310	1951	1,276	1948	2,966	1956	2,104	1974	14,557	1977	19,171	1946	33,086	1952	1,400	1976	31,81	1989	37,50
10	1967	2,751	1955	1,110	1952	1,700	1949	4,500	1958	2,344	1975	12,343	1978	20,314	1947	39,929	1954	0,973	1977	25,57	1990	30,41
11	1968	2,657	1956	1,431	1953	1,649	1950	3,800	1959	1,670	1976	15,129	1980	23,214	1948	25,143	1955	0,805	1978	33,01	1991	48,34
12	1969	2,633	1957	1,237	1954	1,160	1951	3,710	1960	2,130	1977	15,271	1981	23,271	1949	36,014	1956	1,089	1979	61,20	1992	54,63
13	1972	1,854	1958	1,484	1955	0,817	1952	3,089	1962	1,871	1978	15,514	1982	27,243	1950	34,329	1957	1,744	1980	41,61	1993	47,31
14	1973	3,180	1959	1,110	1956	1,250	1953	4,000	1964	1,460	1979	21,800	1984	23,914	1951	31,157	1958	1,559	1981	41,00	1994	34,40
15	1974	2,120	1960	1,230	1957	1,276	1956	2,650	1965	2,714	1980	17,200	1986	21,557	1952	33,072	1959	1,190	1982	46,60		
16	1975	2,000	1961	1,360	1958	1,643	1957	3,190	1966	2,350	1981	20,700	1987	21,071	1953	25,900	1960	0,912	1984	43,59		
17	1976	2,480	1962	1,346	1959	1,250	1958	3,179	1969	1,884	1982	23,300	1988	18,414	1954	18,114	1961	1,080	1985	56,06		
18	1977	2,259	1963	1,160	1960	1,289	1961	3,659	1970	2,397	1984	16,986	1993	25,129	1955	15,271	1962	1,000	1986	34,24		
19	1978	2,194	1964	1,260	1961	1,276	1962	3,620	1971	1,000	1985	21,671	1994	18,814	1956	21,729	1963	0,944	1987	41,34		
20	1981	3,500	1965	1,566	1962	0,969	1963	2,300	1972	2,680	1986	14,257	1995	17,114	1957	26,072	1964	0,944	1988	34,56		
21	1982	3,370			1963	0,978	1964	2,340	1973	2,683	1987	15,400	1996	21,300	1958	30,043	1965	1,550	1989	39,63		
22	1984	2,190			1964	1,080	1965	3,474	1974	2,060	1988	13,943	1997	22,386	1959	19,600			1990	35,63		
23	1985	3,770			1965	1,560	1966	4,580	1975	2,131	1989	16,200	1998	18,929	1960	21,457			1991	43,97		
24	1986	2,221					1968	3,244	1976	2,327	1990	12,929	1999	14,829	1961	28,872			1992	51,54		
25	1987	2,666					1972	3,907	1977	2,764	1992	20,286			1962	26,529			1994	41,79		
26	1988	2,367					1973	5,217	1978	2,909					1963	18,043			1995	42,24		
27	1989	3,009					1974	4,390	1981	2,780					1964	18,314			1996	46,50		
28	1990	2,046					1975	3,639	1982	2,641					1965	35,100			1997	51,34		
29	1992	3,609					1976	4,583	1984	1,891					1966	32,100			1998	26,71		
30	1993	3,227					1977	3,233	1985	2,781					1967	27,272						

ANEXO 11

Vazões mínimas anuais (m^3/s) com 7 dias de duração de algumas estações da bacia do rio Paraopeba

N	Ano	40549998	Ano	40573000	Ano	40577000	Ano	40579995	Ano	40680000	Ano	40710000	Ano	40740000	Ano	40800001	Ano	40818000	Ano	40850000	Ano	40865001
31	1994	1,887					1978	2,859	1986	1,140					1968	28,357						
32	1995	1,840					1979	3,271	1988	1,454					1969	21,700						
33	1996	2,406					1980	2,830	1989	2,679					1970	25,843						
34	1997	3,004					1981	2,080	1992	2,553					1971	12,857						
35	1999	1,727					1982	3,170	1995	1,466					1972	24,757						
36							1983	4,520	1996	2,037					1973	31,157						
37							1984	2,279	1997	1,654					1974	24,743						
38							1985	2,967	1999	1,669					1975	21,700						
39															1976	24,286						
40															1978	24,929						
41															1979	38,886						
42															1980	28,871						
43															1982	35,100						
44															1984	29,414						
45															1985	45,400						
46															1986	26,143						
47															1987	26,243						
48															1988	23,429						
49															1989	25,571						
50															1990	21,357						
51															1991	32,357						
52															1992	37,229						
53															1993	30,129						
54															1994	25,829						
55															1995	22,514						
56															1996	28,286						
57															1997	30,971						
58															1998	16,286						
59															1999	12,771						
60																						

ANEXO 12

Vazões médias diárias máximas anuais (m^3/s) de algumas estações da bacia do rio Paraopeba

AH	40549998	AH	40573000	AH	40577000	AH	40579995	AH	40665000	AH	40710000	AH	40740000
56/57	97,2	46/47	33	42/43	43	38/39	93,6	38/39	22,6	65/66	457	67/68	315
57/58	42,2	49/50	28,8	43/44	22,2	39/40	85,6	39/40	19,3	66/67	350	68/69	356
58/59	44,1	51/52	39,3	44/45	34,4	40/41	112	43/44	24,5	67/68	220	69/70	255
59/60	51,7	52/53	34,9	45/46	19,9	41/42	48,4	44/45	31,5	68/69	268	70/71	182
60/61	80,3	53/54	20	46/47	26,5	42/43	103	45/46	23,5	69/70	190	71/72	474
61/62	56,6	55/56	22	47/48	30,9	43/44	62,9	46/47	24,2	70/71	147	72/73	410
62/63	44	56/57	49,3	48/49	44,8	44/45	76,7	47/48	22,1	71/72	378	73/74	351
65/66	81	57/58	20,5	49/50	21,8	45/46	41,2	48/49	30,7	72/73	330	76/77	456
66/67	77,8	58/59	23,5	51/52	34,1	46/47	54,1	49/50	19,7	73/74	295	77/78	723
67/68	58	59/60	21	52/53	28,2	47/48	62,3	50/51	26,2	74/75	207	78/79	457
68/69	56,9	60/61	50,7	53/54	12,9	48/49	104	51/52	25,7	75/76	150	79/80	460
72/73	91,7	61/62	40,6	54/55	39,3	49/50	46	52/53	19,5	76/77	350	80/81	432
73/74	53,3	62/63	18,5	55/56	16,5	50/51	206	53/54	24,4	77/78	670	81/82	519
74/75	48	63/64	36,1	57/58	22,5	51/52	110	54/55	25,7	78/79	403	82/83	443
75/76	33,1	64/65	33,7	58/59	26,6	52/53	89	55/56	24,9	79/80	336	83/84	387
76/77	65,6			59/60	40,4	55/56	41,2	56/57	27	80/81	385	84/85	816
77/78	112			60/61	39,3	56/57	112	64/65	21	81/82	460	85/86	345
78/79	132			61/62	23,8	57/58	44,1	65/66	18,9	82/83	451	86/87	423
82/83	68,6			63/64	39,3	61/62	75,6	66/67	14,7	83/84	374	87/88	455
83/84	48,5			64/65	27,4	62/63	52,8	72/73	44,4	84/85	785	88/89	222
84/85	50,4					63/64	72,9	73/74	41	85/86	287	90/91	715
85/86	38,3					64/65	68,2	74/75	29,1	86/87	322	92/93	300
86/87	36,4					65/66	112	75/76	33,1	87/88	418	93/94	336
87/88	77					72/73	77	76/77	50,7	88/89	161	94/95	461
88/89	37,2					73/74	111	77/78	44,7	89/90	397	95/96	372
89/90	44,6					74/75	45,5	78/79	39			96/97	1133
92/93	57,9					75/76	30,8	79/80	46,6			97/98	205
93/94	33					76/77	55,8	80/81	42,4			98/99	235
94/95	63,2					77/78	148	83/84	29,5				
95/96	42,4					78/79	128	84/85	52				

ANEXO 12

Vazões médias diárias máximas anuais (m^3/s) de algumas estações da bacia do rio Paraopeba

AH	40549998	AH	40573000	AH	40577000	AH	40579995	AH	40665000	AH	40710000	AH	40740000
96/97	85,7					79/80	59,4						
98/99	39					80/81	50,9						
						81/82	94,6						
						82/83	132						
						83/84	88,2						
						84/85	132						
						87/88	54,7						
						88/89	22,1						
						89/90	63,4						
						90/91	19,8						
						91/92	74,6						
						92/93	57,2						
						93/94	55,1						
						94/95	60,3						
						97/98	66,5						
						98/99	84,2						
						99/00	90,2						

ANEXO 13

Vazões mínimas anuais (m^3/s) com durações de 1, 3, 5 e 7 dias de duração de algumas estações da bacia do rio das Velhas

41151000 – Faz Água Limpa					41180000 – Itabirito Linígrafo				
Ano	1 Dia	3 Dias	5 Dias	7 Dias	Ano	1 Dia	3 Dias	5 Dias	7 Dias
1957	1,330	1,363	1,390	1,409	1967	3,700	3,700	3,740	3,797
1958	1,600	1,600	1,612	1,626	1968	4,650	4,650	4,694	4,713
1959	1,230	1,230	1,230	1,230	1969	3,600	3,600	3,640	3,657
1960	1,180	1,180	1,190	1,201	1970	3,750	3,750	3,750	3,750
1961	1,380	1,413	1,420	1,423	1972	2,700	2,700	2,712	2,717
1962	1,180	1,197	1,210	1,230	1973	4,480	4,480	4,480	4,480
1964	0,978	0,995	1,009	1,015	1974	4,390	4,390	4,390	4,390
1965	1,220	1,220	1,240	1,249	1975	4,640	4,640	4,640	4,640
1966	1,450	1,490	1,486	1,510	1976	3,830	3,830	3,830	3,830
1969	1,120	1,153	1,160	1,163	1977	3,860	3,860	3,886	3,916
1970	1,390	1,410	1,426	1,441	1978	4,080	4,130	4,170	4,187
1971	1,220	1,220	1,264	1,270	1979	5,800	5,800	5,800	5,824
1973	1,760	1,760	1,760	1,760	1980	3,680	3,760	3,850	3,944
1974	1,520	1,520	1,520	1,520	1981	4,440	4,487	4,524	4,560
1975	1,310	1,310	1,340	1,386	1982	3,200	3,273	3,310	3,310
1976	1,040	1,057	1,070	1,083	1983	5,520	5,520	5,618	5,684
1977	1,350	1,350	1,350	1,390	1984	4,010	4,010	4,010	4,010
1978	1,320	1,320	1,340	1,349	1985	5,940	5,940	5,940	5,989
1979	1,980	1,980	1,994	2,010	1986	3,020	3,020	3,206	3,230
1980	1,790	1,877	1,994	2,010	1987	4,230	4,280	4,308	4,340
1981	1,730	1,730	1,754	1,764	1988	4,500	4,593	4,668	4,660
1982	2,110	2,180	2,180	2,200	1989	3,600	3,600	3,600	3,600
1983	2,110	2,110	2,110	2,110	1990	2,880	2,913	2,962	2,954
1984	1,610	1,630	1,646	1,670	1991	3,530	3,530	3,530	3,530
1985	2,180	2,227	2,236	2,240	1992	4,380	4,380	4,482	4,500
1986	1,790	1,790	1,790	1,790	1993	4,810	4,810	4,810	4,830
1987	1,440	1,440	1,474	1,496	1994	4,060	4,083	4,102	4,110
1988	1,730	1,813	1,860	1,877	1995	4,350	4,350	4,396	4,426
1989	1,380	1,380	1,380	1,380	1996	3,810	3,810	3,810	3,810
1990	1,280	1,280	1,280	1,294	1997	3,630	3,690	3,794	3,810
1991	1,730	1,750	1,766	1,801	1998	3,520	3,520	3,520	3,520
1992	1,610	1,730	1,754	1,764	1999	2,420	2,703	2,760	2,760
1993	1,670	1,710	1,718	1,721					
1994	1,360	1,470	1,470	1,470					
1995	1,290	1,290	1,290	1,290					
1996	1,420	1,420	1,420	1,430					
1997	1,650	1,650	1,650	1,661					
1998	1,350	1,350	1,350	1,350					
1999	0,898	0,939	0,960	0,977					

Vazões mínimas anuais (m^3/s) com durações de 1, 3, 5 e 7 dias de duração de algumas estações da bacia do rio das Velhas

41199998 – Honório Bicalho					41260000 - Pinhões				
Ano	1 Dia	3 Dias	5 Dias	7 Dias	Ano	1 Dia	3 Dias	5 Dias	7 Dias
1971	7,500	7,950	8,860	8,979	1980	27,80	27,97	28,12	29,43
1972	13,100	13,800	14,260	14,529	1981	24,50	24,67	24,70	24,80
1973	10,100	11,700	12,240	12,771	1982	30,50	31,30	31,68	31,76
1974	9,620	10,240	10,468	10,491	1983	33,40	33,80	34,24	34,60
1975	8,280	9,183	9,554	9,981	1984	24,50	25,37	25,98	26,56
1976	8,280	9,963	10,654	10,710	1985	33,90	34,87	36,02	36,83
1977	10,600	10,867	10,840	10,900	1986	22,00	22,83	23,84	24,03
1978	11,300	12,467	12,520	12,671	1987	16,70	17,60	19,56	19,83
1979	14,700	15,467	16,160	16,686	1988	24,00	24,33	25,04	25,64
1980	14,200	16,567	17,040	17,086	1989	19,80	20,90	21,62	22,14
1981	12,700	14,533	14,720	14,714	1990	21,70	22,37	23,90	24,64
1982	15,800	16,200	16,380	16,457	1991	32,20	32,93	32,96	33,06
1983	14,600	14,933	15,120	15,200	1992	31,70	31,87	32,10	32,36
1984	12,500	13,500	13,740	13,743	1993	25,40	26,77	27,40	27,47
1985	19,700	19,900	20,060	20,743	1994	16,90	17,47	20,14	21,39
1986	18,000	18,400	18,480	18,429	1995	16,00	16,57	16,94	17,17
1987	11,500	11,833	12,100	12,357	1996	23,20	24,03	24,40	24,74
1988	16,800	17,700	17,980	18,100	1997	32,90	33,03	33,54	34,04
1989	12,500	12,667	12,900	12,929	1998	23,70	23,87	24,08	24,34
1990	13,500	13,900	14,180	14,457	1999	19,40	19,53	19,74	19,83
1991	16,200	17,300	17,420	17,629					
1992	16,200	16,600	16,780	16,929					
1993	16,800	17,333	17,760	17,871					
1994	15,400	16,100	16,580	16,700					
1995	10,800	12,133	12,400	12,586					
1996	13,700	14,367	14,500	14,714					
1997	17,500	17,667	18,500	19,100					
1998	10,100	11,033	11,320	11,643					
1999	9,620	9,780	10,008	10,106					

Vazões mínimas anuais (m^3/s) com durações de 1, 3, 5 e 7 dias de duração de algumas estações da bacia do rio das Velhas

41340000 – Ponte Raul Soares

Ano	1 Dia	3 Dias	5 Dias	7 Dias	Ano	1 Dia	3 Dias	5 Dias	7 Dias
1938	33,70	35,20	36,52	36,70	1990	30,80	30,97	31,42	31,77
1939	28,30	29,87	31,04	31,60	1991	34,50	34,87	35,16	35,76
1940	22,10	25,00	26,18	26,93	1992	41,10	41,50	41,70	41,77
1941	28,30	31,87	33,62	34,91	1993	32,90	33,63	34,20	34,46
1942	24,20	29,87	31,28	32,74	1994	31,90	32,93	34,10	34,61
1944	30,60	33,10	34,08	34,90	1995	23,80	24,73	24,92	25,21
1947	32,50	36,97	38,38	39,27	1996	27,70	28,93	29,58	30,07
1950	27,10	30,47	31,02	32,97	1997	34,00	35,43	36,58	37,24
1951	18,00	19,13	19,36	20,23	1998	28,00	28,33	28,80	29,14
1952	29,40	31,43	32,84	34,19	1999	20,90	21,57	22,36	23,14
1953	25,40	28,53	29,12	30,07					
1954	17,10	20,60	22,16	22,59					
1955	14,80	16,97	17,94	18,43					
1956	17,50	19,03	20,78	20,81					
1957	31,50	31,50	31,78	31,87					
1958	31,90	33,10	35,00	36,47					
1959	17,50	18,17	18,58	18,86					
1960	21,00	21,93	22,32	23,03					
1961	25,40	28,47	29,08	29,51					
1962	18,00	19,23	20,60	21,31					
1963	17,10	18,83	19,18	19,70					
1964	19,40	21,03	21,04	21,19					
1965	34,90	35,17	35,26	35,23					
1967	22,40	22,67	22,98	23,26					
1968	22,00	23,13	23,54	23,59					
1969	19,10	19,23	19,58	19,80					
1970	24,60	24,73	24,94	25,03					
1971	16,30	16,83	16,94	17,04					
1972	21,30	21,53	21,76	21,80					
1973	19,90	22,97	23,30	23,60					
1974	21,80	21,97	22,50	22,80					
1975	24,40	24,57	24,60	24,97					
1976	16,90	18,73	19,26	19,44					
1977	24,00	24,83	25,20	25,64					
1978	22,00	22,67	24,40	25,43					
1979	33,40	34,13	35,40	36,69					
1981	31,20	31,37	31,84	32,20					
1984	26,90	29,57	31,64	32,37					
1985	40,20	41,40	41,88	42,36					
1986	25,70	26,70	27,00	27,00					
1987	22,30	23,13	23,48	23,43					
1988	25,70	26,20	26,70	27,30					
1989	25,70	26,20	27,00	27,73					

ANEXO 14

Séries de intensidade de precipitação (mm/h) de Andorinhas, código 02243235

N	5 min	10 min	15 min	30 min	45 min	1 H	2 H	3 H	4 H	8 H	14 H	24 H
1	240,00	168,00	148,60	123,00	111,60	100,15	69,40	53,70	49,58	32,89	22,01	13,54
2	223,20	167,40	144,80	112,10	108,67	86,20	63,85	51,68	47,10	32,74	21,56	12,67
3	216,00	162,30	136,20	111,80	100,20	84,70	60,10	47,90	41,05	25,51	15,89	9,51
4	186,00	150,00	133,20	110,90	99,47	84,10	49,95	42,60	31,95	22,21	13,21	9,03
5	186,00	149,10	126,00	110,70	91,73	77,25	49,65	39,93	31,46	20,11	12,79	8,12
6	172,80	144,30	125,80	109,80	90,00	75,20	49,40	38,83	30,60	18,47	12,63	7,46
7	169,80	141,30	123,20	97,40	83,33	74,95	47,13	38,05	30,10	17,91	10,78	6,92
8	168,00	138,30	121,60	96,00	82,53	72,50	46,78	36,07	29,73	17,58	10,28	6,55
9	161,40	136,80	121,00	95,80	81,53	71,50	45,75	35,08	29,71	17,23	10,07	6,45
10	158,40	132,90	119,20	94,50	80,07	71,40	45,70	34,80	28,89	16,19	9,87	6,30
11	158,40	131,10	118,80	94,30	77,87	67,10	45,45	34,38	28,05	16,11	9,43	6,07
12	156,00	130,50	118,80	92,90	76,93	66,95	45,18	33,30	27,78	15,86	9,21	6,00
13	153,60	129,90	118,40	91,00	75,33	65,50	45,15	32,80	27,18	15,61	9,13	5,93
14	151,20	129,60	117,00	90,80	75,27	62,80	45,10	32,00	25,08	15,23	8,92	5,79
15	142,80	125,40	115,20	90,20	75,13	62,80	44,10	31,23	24,51	13,76	8,71	5,76
16	141,60	123,00	115,20	88,70	74,53	62,55	43,58	30,50	24,23	13,56	8,69	5,72
17	141,60	121,50	113,60	88,60	73,33	62,50	41,60	30,30	23,99	13,54	8,53	5,51
18	140,40	121,50	113,20	87,10	73,07	61,80	40,83	30,28	23,55	13,01	8,31	5,37
19	138,60	120,60	112,80	86,80	71,87	61,65	39,85	30,18	23,38	12,68	7,91	5,20
20	138,00	120,60	112,00	86,60	70,73	59,00	38,75	29,72	22,89	12,16	7,87	5,08
21	136,80	119,70	112,00	85,00	70,00	58,10	37,80	29,33	22,73	12,11	7,51	5,04
22	133,20	118,80	111,60	82,80	69,87	57,75	37,78	28,95	21,96	12,06	7,46	5,02
23	132,00	118,80	109,80	82,00	68,27	57,50	37,50	27,35	21,91	11,96	7,36	4,83
24	129,60	117,60	108,20	80,80	67,20	56,90	36,35	27,07	21,90	11,45	7,26	4,83
25	127,80	117,00	108,00	80,60	67,07	56,50	35,75	26,78	21,16	11,36	7,22	4,83
26	127,20	117,00	108,00	80,40	66,60	55,90	35,70	26,25	20,54	11,33	7,03	4,61
27	127,20	116,40	106,00	80,20	66,53	55,75	35,60	26,23	20,33	11,09	6,96	4,61
28	127,20	115,50	105,80	79,60	65,67	55,50	35,33	26,05	20,10	10,85	6,87	4,59
29	126,60	115,20	105,80	79,60	65,60	54,30	35,33	25,53	20,03	10,69	6,82	4,56
30	125,40	115,20	105,60	78,70	65,40	53,85	34,00	25,03	19,55	10,68	6,76	4,53
31	124,80	114,60	104,40	78,60	64,60	53,70	32,58	24,42	19,34	10,61	6,54	4,41
32	124,80	113,40	104,00	78,40	62,87	53,25	32,25	24,38	19,26	10,41	6,52	4,38
33	123,60	113,10	104,00	78,20	62,60	52,85	31,53	23,85	19,04	10,34	6,52	4,35
34	123,60	112,80	103,20	78,00	62,33	52,40	31,45	22,98	18,89	10,18	6,52	4,33
35	123,00	112,80	103,20	76,70	62,00	51,45	29,45	22,82	18,84	10,06	6,52	4,27
36	122,40	112,80	102,80	76,00	60,80	50,90	29,30	21,33	17,44	9,98	6,49	4,26
37	122,40	112,50	102,00	75,60	60,53	50,80	29,15	21,15	17,44	9,88	6,49	4,25
38	122,40	112,50	99,60	75,10	60,13	50,40	28,68	21,12	16,86	9,87	6,46	4,23
39	121,20	112,20	99,40	74,70	60,00	50,30	28,65	21,05	16,69	9,86	6,40	4,22
40	121,20	112,20	99,00	72,90	59,80	49,85	28,38	20,60	16,43	9,81	6,21	4,21
41	120,00	112,20	98,00	72,40	59,53	49,75	28,28	20,33	16,24	9,80	6,14	4,19
42	120,00	112,20	97,40	72,20	59,20	48,90	28,25	19,77	16,04	9,78	6,12	4,14

Séries de intensidade de precipitação (mm/h) de Apolinário, código 02242092

N	5 min	10 min	15 min	30 min	45 min	1 H	2 H	3 H	4 H	8 H	14 H	24 H
1	193,20	160,80	131,20	100,40	90,33	77,30	47,00	41,27	39,39	21,65	12,67	8,92
2	192,00	154,20	128,60	93,80	83,73	76,60	44,45	40,68	32,15	16,50	10,56	7,54
3	171,60	136,20	122,00	92,00	73,60	68,20	41,33	31,08	24,60	16,20	10,01	7,39
4	170,40	122,40	120,40	88,40	68,27	61,00	40,50	29,83	23,33	15,18	9,65	6,57
5	170,40	121,20	116,00	81,20	68,13	60,60	39,90	29,30	22,99	13,90	9,26	5,81
6	165,60	121,20	114,80	81,10	67,73	60,50	39,08	27,35	22,98	13,64	8,68	5,65
7	140,40	120,60	114,00	79,40	66,13	59,95	38,65	27,28	22,39	13,09	8,67	5,64
8	132,00	119,70	113,20	78,60	66,13	58,90	37,48	27,15	21,51	12,76	8,66	5,53
9	131,40	119,10	112,80	78,40	63,73	53,75	37,20	26,83	21,31	12,30	8,55	5,52
10	129,60	118,80	108,80	77,60	62,47	53,30	35,48	26,80	20,83	12,19	8,31	5,48
11	129,60	118,80	107,60	76,70	60,67	53,20	34,58	26,45	20,61	12,17	8,14	5,40
12	129,60	118,50	107,20	76,00	60,27	53,10	34,30	26,32	20,18	12,06	7,81	5,39
13	129,00	117,60	106,60	76,00	59,67	51,95	33,93	24,45	20,10	11,80	7,80	5,30
14	128,40	117,60	106,20	74,40	58,20	50,70	33,10	24,40	18,86	11,76	7,61	5,30
15	126,00	116,10	105,60	73,80	57,93	50,60	32,33	23,67	18,75	11,19	7,45	5,09
16	124,80	115,20	104,80	73,20	57,87	49,45	32,30	23,45	18,59	11,10	7,42	5,06
17	124,80	115,20	104,20	73,20	57,80	49,10	32,10	23,43	18,15	11,01	7,20	5,06
18	123,60	115,20	103,60	72,80	57,47	48,55	32,05	22,98	17,91	10,86	7,05	4,99
19	122,40	114,60	103,20	72,70	57,07	47,70	30,55	22,63	17,78	10,69	6,95	4,95
20	121,20	114,60	102,40	71,40	55,67	47,55	30,50	21,97	17,75	10,68	6,93	4,88
21	120,00	114,00	100,40	71,00	54,53	47,15	29,60	21,90	17,21	10,46	6,87	4,85
22	120,00	114,00	99,00	70,60	54,53	47,00	29,48	21,87	17,18	10,37	6,86	4,84
23	120,00	114,00	98,20	70,40	54,40	45,80	28,73	21,43	17,14	10,33	6,74	4,83
24	120,00	113,70	96,80	69,80	53,87	45,75	28,70	21,07	17,13	10,26	6,70	4,79
25	120,00	113,40	96,00	69,70	53,07	45,50	28,65	21,03	16,59	10,11	6,62	4,73
26	120,00	113,40	95,80	68,80	52,47	45,10	28,38	20,72	16,49	10,06	6,52	4,66
27	120,00	112,80	95,00	67,20	52,40	44,90	27,80	20,60	16,40	10,04	6,45	4,66
28	120,00	112,80	94,60	66,20	52,27	44,10	27,65	20,35	16,15	9,93	6,44	4,64
29	119,40	112,80	93,60	65,80	52,07	44,00	27,55	20,32	16,09	9,89	6,40	4,61
30	118,80	111,90	93,60	65,10	51,73	44,00	26,60	20,12	16,03	9,79	6,33	4,60
31	118,80	111,60	92,00	64,50	51,47	43,95	26,58	19,98	15,99	9,59	6,30	4,58
32	118,80	111,60	91,20	63,60	50,53	43,90	26,25	19,73	15,93	9,51	6,29	4,57
33	118,80	111,60	90,80	63,20	50,40	42,70	26,20	19,67	15,75	9,44	6,28	4,55
34	117,60	111,60	90,40	62,90	50,27	42,30	26,20	19,57	15,40	9,41	6,17	4,51
35	117,60	111,60	89,40	62,60	49,60	42,15	26,20	19,35	15,26	9,41	6,16	4,49
36	117,60	110,40	88,40	62,60	49,47	42,00	26,00	19,33	15,21	9,40	6,09	4,39
37	117,60	109,80	87,40	62,60	49,07	41,60	25,88	19,23	15,08	9,23	6,03	4,32
38	117,60	108,90	87,00	61,80	48,80	41,55	25,75	19,07	15,08	9,21	6,02	4,28
39	117,60	108,00	85,60	61,40	48,53	41,50	25,50	18,92	15,01	9,10	6,02	4,26
40	117,60	108,00	85,60	61,00	47,87	41,20	25,40	18,60	14,95	9,06	5,97	4,26

Séries de intensidade de precipitação (mm/h) de Faz. Santo Amaro, código 02242096

N	5 min	10 min	15 min	30 min	45 min	1 H	2 H	3 H	4 H	8 H	14 H	24 H
1	191,40	154,50	138,80	111,60	98,93	83,60	43,28	37,90	33,84	26,10	15,29	9,31
2	141,60	123,60	114,40	111,30	96,87	76,30	42,23	29,35	27,41	18,09	11,76	8,39
3	130,80	115,20	103,60	111,20	91,20	73,35	41,93	27,97	26,16	16,70	11,30	6,87
4	123,00	110,10	96,00	110,00	87,20	72,60	41,43	25,80	26,14	16,06	10,64	6,68
5	170,40	151,20	135,40	104,30	87,00	70,25	40,93	25,17	24,65	15,89	10,42	6,67
6	141,60	119,10	114,20	100,80	86,73	69,30	39,10	24,57	24,61	13,99	9,93	6,63
7	130,20	114,90	103,20	99,60	80,93	67,70	38,60	23,68	23,69	13,91	9,18	6,33
8	121,20	109,50	94,80	96,80	78,87	66,30	38,28	23,25	23,18	13,20	8,84	6,22
9	168,00	149,10	135,20	96,60	78,80	66,20	37,95	22,75	22,63	12,38	8,35	5,99
10	139,20	118,80	113,20	95,40	77,80	66,00	37,45	22,55	22,26	12,16	8,15	5,93
11	129,60	114,60	101,40	94,70	76,80	63,10	37,35	21,28	21,70	12,00	7,85	5,91
12	121,20	109,20	94,40	94,40	76,53	62,10	37,20	34,57	21,19	11,98	7,85	5,87
13	166,80	139,80	132,40	91,20	76,00	61,95	36,98	28,57	21,08	11,95	7,55	5,84
14	139,20	118,80	112,80	88,50	74,40	61,35	35,20	27,57	20,00	11,59	7,14	5,65
15	129,00	114,60	99,00	87,70	73,27	60,75	35,13	25,62	19,96	11,26	7,13	5,57
16	120,00	107,70	91,80	86,40	72,27	60,05	34,95	24,92	19,95	11,16	7,09	5,29
17	165,60	135,60	127,20	83,80	71,93	58,80	34,78	24,43	19,65	11,06	7,00	5,25
18	133,20	118,20	112,00	83,60	67,47	57,85	34,63	23,62	19,65	11,04	6,99	5,17
19	127,20	114,00	98,80	83,50	66,80	57,60	34,63	23,02	19,64	11,01	6,96	5,14
20	120,00	107,40	90,80	82,10	66,40	56,15	34,60	22,68	19,56	10,85	6,96	5,10
21	153,60	134,40	120,00	79,80	65,53	55,80	34,38	22,47	19,50	10,74	6,95	5,02
22	133,20	117,90	112,00	77,80	65,13	53,40	33,90	21,13	19,21	10,70	6,89	4,99
23	125,40	113,40	97,60	77,60	65,00	52,50	33,55	33,80	19,14	10,67	6,88	4,95
24	120,00	107,40	89,60	77,20	64,07	50,35	33,23	28,17	19,09	10,63	6,86	4,85
25	151,20	133,20	119,60	76,80	63,87	50,05	33,08	26,33	19,00	10,63	6,79	4,60
26	133,20	117,60	108,40	75,80	61,93	49,75	33,03	25,52	18,93	10,58	6,78	4,46
27	125,40	112,80	97,00	75,30	61,80	49,75	31,90	24,90	18,80	10,44	6,75	4,40
28	120,00	107,40	89,60	74,80	60,53	49,00	31,45	23,93	18,75	10,39	6,63	4,37
29	144,00	131,40	118,80	73,00	60,27	48,85	31,20	23,53	18,63	10,26	6,63	4,37
30	132,00	117,00	108,00	73,00	59,47	48,80	30,90	22,95	18,50	10,24	6,56	4,30
31	124,80	112,80	96,80	72,40	58,67	48,75	30,33	22,60	18,45	10,23	6,51	4,28
32	119,40	106,50	87,20	71,60	57,40	48,60	30,30	22,33	18,30	10,14	6,48	4,27
33	144,00	129,60	117,60	70,90	56,73	48,55	29,70	31,25	18,10	10,11	6,46	4,21
34	132,00	116,70	107,40	70,80	55,93	48,50	29,58	28,10	17,94	10,10	6,40	4,16
35	123,60	112,20	96,80	70,80	55,53	48,15	29,08	26,10	17,78	10,01	6,37	4,16
36	118,80	106,20	87,20	70,40	55,47	47,90	29,03	25,33	17,64	9,95	6,34	4,15
37	142,80	126,90	116,20	70,20	55,40	47,25	28,28	24,90	17,54	9,89	6,29	4,12
38	132,00	116,10	104,80	70,00	55,33	47,00	28,18	23,92	17,45	9,81	6,24	4,11
39	123,00	110,40	96,00	69,60	54,93	46,80	27,95	23,50	17,40	9,78	6,20	4,10
40	118,80	105,90	86,80	69,40	54,93	46,40	27,80	22,92	17,36	9,69	6,13	4,05
41	142,80	124,20	115,60	68,90	54,93	45,55	27,80	22,57	17,19	9,64	6,11	4,04
42	132,00	115,20	104,80	68,60	54,60	45,45	27,75	21,92	16,96	9,64	6,08	4,01

Séries de intensidade de precipitação (mm/h) de Nova Friburgo, código 02242070

N	5 min	10 min	15 min	30 min	45 min	1 H	2 H	3 H	4 H	8 H	14 H	24 H
1	232,80	178,20	159,20	120,40	96,53	81,60	42,30	28,30	22,88	12,34	8,27	4,83
2	225,60	177,60	122,80	87,00	67,87	55,10	34,65	23,77	21,23	10,64	7,55	4,68
3	195,60	132,60	120,40	81,80	64,80	53,40	29,85	23,40	18,20	10,64	7,15	4,68
4	150,60	124,80	120,00	78,80	64,40	51,40	29,80	21,60	16,78	10,61	7,15	4,52
5	129,60	124,20	117,60	77,00	62,53	51,20	27,85	19,90	15,93	10,54	6,66	3,93
6	126,00	120,60	111,60	75,20	61,87	51,00	27,28	18,57	14,96	9,61	6,06	3,54
7	126,00	118,20	108,80	73,80	58,67	47,90	26,80	18,32	14,20	9,46	5,59	3,44
8	126,00	114,60	88,80	70,20	53,73	45,30	26,65	18,17	14,20	9,44	5,46	3,28
9	124,80	114,00	85,60	67,00	52,00	42,60	26,40	17,77	14,20	8,66	5,41	3,15
10	124,80	112,80	85,20	63,40	51,87	42,60	25,70	17,15	13,93	7,83	5,24	3,05
11	124,80	111,00	84,00	63,20	51,20	40,90	24,13	16,17	13,65	7,73	4,75	2,95
12	124,80	106,80	82,80	63,00	50,67	39,80	23,93	16,10	13,45	7,53	4,46	2,85
13	124,80	104,40	82,40	61,80	48,60	39,10	23,05	15,95	13,33	7,49	4,39	2,73
14	123,60	100,80	82,00	60,20	47,47	36,70	21,10	15,47	12,88	7,29	4,32	2,60
15	123,60	99,60	81,20	60,20	46,53	36,50	20,55	15,07	12,78	6,96	4,31	2,57
16	123,60	96,60	80,80	59,40	45,33	36,40	19,83	14,83	12,35	6,66	4,28	2,53
17	122,40	96,60	79,20	58,40	44,27	36,00	19,50	14,60	12,26	6,59	4,14	2,52
18	122,40	93,00	79,20	57,80	43,47	35,90	19,15	14,37	12,08	6,53	4,07	2,50
19	121,20	92,40	78,80	56,80	43,47	35,30	19,10	14,27	11,98	6,44	3,98	2,46
20	121,20	91,20	78,80	56,20	43,07	34,90	18,98	14,27	11,74	6,39	3,81	2,36
21	121,20	89,40	78,40	54,40	42,67	33,60	18,50	13,93	11,68	6,33	3,79	2,33
22	121,20	88,80	77,20	54,40	42,40	33,30	18,30	13,70	11,68	6,11	3,75	2,32
23	120,00	88,80	77,20	53,60	42,00	33,30	17,95	13,60	11,48	6,08	3,74	2,24
24	120,00	88,20	76,80	53,40	41,60	32,90	17,83	13,57	11,43	5,99	3,71	2,22
25	120,00	88,20	76,60	52,80	41,20	32,70	17,65	13,53	11,00	5,97	3,68	2,22
26	120,00	87,60	75,20	52,40	40,13	32,70	17,50	13,25	10,63	5,88	3,66	2,20
27	120,00	83,70	74,00	52,40	39,87	32,70	17,50	13,17	10,28	5,85	3,65	2,19
28	118,80	82,80	73,20	51,80	39,73	32,60	17,50	13,03	10,28	5,74	3,60	2,15
29	118,80	82,20	72,00	50,80	38,93	32,20	17,40	12,83	10,01	5,71	3,54	2,14
30	118,80	82,20	71,20	50,00	38,40	31,90	17,30	12,68	9,78	5,68	3,49	2,14
31	118,80	82,20	70,00	49,40	38,00	31,60	17,08	12,48	9,63	5,61	3,46	2,13
32	117,60	81,00	68,40	48,40	37,60	31,20	17,03	11,97	9,60	5,53	3,44	2,13
33	117,60	81,00	68,40	48,20	37,60	31,20	16,95	11,90	9,50	5,46	3,43	2,13
34	117,60	80,40	66,80	47,80	37,47	31,05	16,65	11,70	9,43	5,40	3,43	2,04
35	116,40	79,80	65,60	47,60	37,33	30,30	16,60	11,70	9,09	5,33	3,42	2,02
36	115,20	78,60	65,20	47,40	36,53	29,65	16,55	11,68	9,08	5,30	3,41	2,00
37	115,20	78,00	64,80	47,00	36,13	29,30	16,55	11,63	8,98	5,27	3,39	2,00
38	112,80	77,40	64,80	43,80	36,00	29,10	16,50	11,47	8,98	5,21	3,34	1,98

Séries de intensidade de precipitação (mm/h) de Posto Garrafão, código 02242098

N	5 min	10 min	15 min	30 min	45 min	1 H	2 H	3 H	4 H	8 H	14 H	24 H
1	192,00	156,00	137,60	114,00	96,67	85,60	57,15	45,37	38,43	25,10	15,80	9,24
2	189,60	148,20	132,00	112,40	95,87	82,40	53,40	42,77	36,68	20,55	13,97	8,36
3	183,60	147,60	128,00	110,00	93,33	82,00	52,50	40,20	35,63	20,02	13,19	8,28
4	182,40	141,00	128,00	108,40	92,93	79,80	52,15	39,67	32,68	19,88	12,78	8,18
5	182,40	133,80	126,40	107,00	90,67	78,40	51,00	37,63	32,05	18,50	12,32	8,07
6	177,60	132,00	123,20	105,20	89,60	78,10	50,00	37,33	28,85	18,40	11,78	7,70
7	176,40	128,40	121,20	98,80	86,53	74,00	48,85	37,10	28,70	18,39	11,74	7,45
8	165,60	125,40	120,00	97,60	83,87	70,70	46,45	36,73	28,65	17,75	10,79	7,44
9	164,40	124,80	120,00	96,00	80,00	70,70	44,40	36,20	28,00	17,41	10,78	7,03
10	158,40	124,80	118,40	93,40	79,73	70,60	44,10	33,17	27,33	16,26	10,71	6,87
11	154,80	123,60	116,80	93,20	78,40	69,90	43,50	32,37	26,80	15,70	10,56	6,85
12	154,80	123,60	116,00	89,60	78,40	68,10	43,30	31,20	26,03	15,55	10,21	6,58
13	151,20	120,00	115,20	88,40	77,87	66,10	43,25	30,93	25,00	15,13	9,64	6,54
14	148,80	120,00	115,20	87,80	77,47	66,00	42,83	30,33	24,93	14,88	9,51	6,40
15	146,40	118,80	114,80	87,40	76,80	64,80	42,05	30,05	24,63	14,48	9,21	6,29
16	144,00	118,80	114,00	87,40	74,13	64,00	40,85	29,93	24,33	14,05	9,06	6,21
17	144,00	118,80	113,20	86,30	73,47	63,80	40,55	29,62	23,63	14,03	9,03	6,20
18	138,00	118,20	111,60	86,00	72,00	61,60	40,40	29,30	22,83	13,64	8,99	5,98
19	128,40	117,60	111,20	86,00	70,67	60,90	40,10	28,87	22,80	13,36	8,52	5,91
20	127,20	117,60	110,40	84,80	67,60	60,70	40,00	28,80	22,80	13,18	8,35	5,56
21	124,80	117,60	109,60	84,40	67,20	58,30	39,30	28,57	22,73	13,15	8,27	5,51
22	124,80	117,60	109,60	83,60	66,13	58,00	37,75	28,57	22,64	13,01	8,26	5,38
23	124,80	117,60	108,40	81,60	66,00	57,70	37,25	28,27	22,28	12,79	8,13	5,33
24	124,80	117,60	108,40	81,20	64,80	57,60	36,90	28,10	21,65	12,73	8,03	5,20
25	123,60	117,00	106,80	78,60	64,67	57,20	36,35	27,87	21,58	12,39	8,01	5,13
26	122,40	116,40	103,60	78,40	64,53	55,60	36,05	27,80	21,43	12,23	7,99	5,12
27	122,40	115,80	102,40	78,00	64,53	54,70	36,05	27,32	21,20	12,16	7,65	5,12
28	122,40	115,20	100,80	77,80	64,00	54,50	35,20	27,30	21,20	11,91	7,63	5,04
29	120,00	115,20	100,80	77,60	63,20	54,40	34,80	26,87	21,16	11,70	7,59	5,00
30	120,00	115,20	100,00	76,80	62,80	54,00	34,10	26,73	20,90	11,64	7,46	4,94
31	120,00	115,20	98,80	76,80	62,53	52,70	33,40	25,53	20,83	11,59	7,40	4,88
32	120,00	114,00	98,40	75,80	62,53	52,40	33,35	25,47	20,76	11,55	7,38	4,84
33	120,00	114,00	97,60	75,60	62,27	51,90	33,13	25,20	20,73	11,37	7,18	4,83
34	120,00	114,00	97,60	75,40	62,00	51,60	32,95	25,17	20,50	11,20	6,92	4,83
35	120,00	113,40	97,20	75,20	61,87	51,00	32,95	24,73	20,50	11,14	6,74	4,82
36	120,00	113,40	96,80	74,80	61,47	50,70	32,75	24,30	20,40	11,11	6,66	4,77
37	120,00	112,80	96,80	73,60	61,20	50,60	32,25	24,30	19,70	10,96	6,66	4,74
38	120,00	112,20	95,20	73,40	61,07	50,50	32,10	24,20	18,89	10,83	6,66	4,73
39	118,80	112,20	91,20	73,40	60,80	50,40	31,80	24,10	18,88	10,71	6,66	4,72
40	118,80	111,60	91,20	73,40	60,80	50,20	31,55	24,07	18,88	10,69	6,62	4,68
41	118,80	111,60	91,20	73,40	60,67	50,20	31,20	23,80	18,23	10,69	6,60	4,68
42	117,60	110,40	90,00	72,80	59,47	49,90	30,90	22,68	18,08	10,66	6,55	4,65

Séries de intensidade de precipitação (mm/h) de Quizanga, código 02242093

N	5 min	10 min	15 min	30 min	45 min	1 H	2 H	3 H	4 H	8 H	14 H	24 H
1	226,80	173,40	166,00	120,20	102,20	87,35	56,88	38,30	28,89	17,78	11,72	6,95
2	196,20	170,10	151,20	104,90	91,93	83,40	49,35	36,38	28,81	15,38	11,43	6,67
3	181,80	153,60	139,80	103,10	85,07	70,70	47,70	33,78	28,45	15,07	9,03	6,58
4	178,20	131,40	121,60	99,40	81,67	68,40	47,60	33,28	26,24	15,07	8,95	6,44
5	162,60	123,60	121,00	98,20	79,13	67,50	45,60	32,98	25,91	15,06	8,95	6,06
6	160,80	122,10	119,80	97,70	79,07	66,80	42,25	31,17	25,58	14,80	8,79	5,74
7	151,80	121,50	119,20	97,20	77,87	66,10	41,38	30,37	24,96	13,06	8,70	5,52
8	151,20	121,20	118,80	92,30	73,33	62,65	41,30	28,60	24,74	12,78	8,70	5,22
9	148,80	120,30	116,80	91,20	73,00	62,55	39,53	27,85	23,21	12,58	8,46	5,17
10	142,20	119,40	113,60	87,60	72,87	61,50	38,73	27,63	22,79	12,48	8,18	5,13
11	141,60	119,40	112,20	84,30	72,27	60,30	38,48	27,12	21,51	12,37	7,97	5,08
12	140,40	119,40	112,00	83,90	71,07	59,65	38,13	27,07	20,74	11,91	7,47	5,07
13	134,40	118,20	111,60	81,20	69,20	59,40	36,70	26,55	20,70	11,76	7,44	4,94
14	132,00	117,60	111,60	81,10	68,80	59,25	36,50	26,48	20,39	11,47	7,32	4,94
15	132,00	117,00	110,80	80,60	67,40	58,80	35,98	26,42	20,01	11,26	7,14	4,83
16	130,80	115,20	109,80	80,20	65,60	56,75	35,88	26,27	20,00	11,23	7,07	4,44
17	130,20	115,20	106,00	79,50	65,40	56,50	32,55	25,67	19,95	11,06	7,02	4,39
18	124,80	114,90	104,80	78,90	65,20	53,25	32,05	24,02	19,86	10,85	6,86	4,36
19	124,20	114,30	102,60	78,80	63,20	53,00	31,68	23,50	19,60	10,71	6,68	4,27
20	123,60	114,30	102,00	78,40	63,07	52,90	31,58	23,37	19,40	10,67	6,62	4,17
21	123,60	113,70	101,20	77,80	62,93	50,00	31,48	22,90	19,25	10,65	6,45	4,16
22	123,00	113,10	101,20	77,20	62,53	49,95	30,95	22,58	19,09	10,38	6,44	4,14
23	123,00	112,20	100,40	75,40	62,27	49,30	30,65	22,32	18,73	10,34	6,39	4,12
24	122,40	112,20	99,00	74,90	61,33	49,10	30,53	22,13	18,50	10,28	6,28	4,05
25	121,80	111,60	99,00	73,70	58,27	48,65	30,40	21,87	18,43	10,17	6,25	3,90
26	121,20	111,60	99,00	73,70	58,13	47,70	30,38	21,40	17,89	10,01	5,98	3,84
27	120,60	111,30	98,60	73,10	56,27	47,60	30,15	21,37	17,75	9,94	5,93	3,78
28	120,00	110,40	98,40	72,80	56,27	47,45	29,75	21,02	17,65	9,82	5,93	3,76
29	120,00	109,80	97,00	72,80	56,07	47,40	29,63	20,93	17,05	9,74	5,74	3,75
30	120,00	108,60	96,80	72,60	55,73	46,40	29,50	20,67	16,80	9,63	5,72	3,73
31	119,40	108,60	94,00	72,00	55,60	46,25	29,40	20,47	16,29	9,60	5,71	3,48
32	119,40	108,30	94,00	71,30	54,33	45,60	29,25	20,33	16,28	9,45	5,69	3,48
33	119,40	107,70	93,20	71,20	54,00	45,20	28,95	20,23	16,15	9,44	5,69	3,46
34	117,60	106,80	92,40	71,00	53,93	44,85	28,38	20,20	16,06	9,41	5,63	3,42
35	117,60	106,80	91,20	70,40	53,93	44,75	28,28	20,13	15,99	9,36	5,59	3,39
36	116,40	106,80	90,20	70,10	53,80	44,50	28,23	19,87	15,79	9,28	5,58	3,39
37	116,40	106,20	90,20	69,20	53,73	44,20	27,85	19,83	15,61	9,21	5,50	3,36
38	116,40	105,60	90,00	69,20	53,27	44,15	27,73	19,60	15,55	9,06	5,43	3,34
39	116,40	105,00	89,60	68,80	52,80	44,15	27,28	19,52	15,53	8,93	5,38	3,32
40	116,40	104,40	89,00	68,60	52,53	44,00	26,90	19,50	15,40	8,77	5,35	3,29
41	116,40	104,40	88,80	67,70	52,20	43,90	26,48	19,47	15,23	8,76	5,31	3,28
42	115,80	104,10	88,60	67,60	52,07	43,75	26,45	19,43	15,09	8,59	5,27	3,25

Séries de intensidade de precipitação (mm/h) de Petrópolis, código 02243188

N	5 min	10 min	15 min	30 min	45 min	1 H	2 H	3 H	4 H	8 H	14 H	24 H
1	186,00	151,20	139,60	114,40	95,47	80,80	45,90	31,55	23,70	15,78	9,59	7,00
2	126,00	117,00	114,80	99,60	81,87	52,40	44,10	29,92	22,59	13,34	9,09	5,69
3	115,20	115,20	108,00	88,00	71,73	34,50	35,08	25,20	21,35	12,31	8,13	5,09
4	115,20	108,00	100,00	78,40	64,93	65,60	34,95	23,48	20,23	11,85	7,59	4,49
5	114,00	100,20	84,40	76,40	64,13	51,00	28,60	23,17	17,63	11,34	6,77	4,33
6	114,00	99,60	82,80	71,00	63,20	33,40	28,55	20,33	16,95	11,25	6,65	4,28
7	112,80	97,80	82,80	60,80	49,87	64,70	27,00	20,00	16,50	10,83	6,63	3,95
8	111,60	92,40	79,20	58,00	43,20	38,90	25,40	19,57	16,18	9,70	6,53	3,81
9	108,00	91,80	76,40	52,20	41,20	33,00	22,00	19,53	15,63	8,89	6,48	3,78
10	105,60	87,60	67,20	52,20	40,80	53,10	21,10	18,77	15,33	8,81	6,27	3,66
11	104,40	81,00	66,00	51,60	38,13	34,70	20,50	16,47	13,53	8,26	5,32	3,12
12	97,20	79,20	64,40	49,40	35,87	31,20	20,35	14,83	13,33	7,97	5,04	3,09

Séries de intensidade de precipitação (mm/h) de Teresópolis - PN, código 02243151

N	5 min	10 min	15 min	30 min	45 min	1 H	2 H	3 H	4 H	8 H	14 H	24 H
1	142,80	135,60	102,40	79,40	71,33	63,60	40,28	28,93	21,79	10,95	6,53	3,85
2	115,20	108,00	95,60	63,80	49,87	38,60	35,80	25,37	19,88	10,26	6,26	3,65
3	114,00	105,60	82,80	61,20	43,87	36,00	25,15	23,53	19,33	10,03	5,73	3,34
4	112,80	96,00	80,00	50,40	43,60	35,60	23,00	16,13	12,93	9,66	5,52	3,22
5	112,80	85,80	71,20	49,80	38,67	33,50	19,90	15,27	12,55	7,89	5,16	3,08
6	110,40	79,20	70,40	49,60	36,40	33,30	19,35	14,37	12,35	7,69	4,81	3,04
7	106,80	78,00	68,80	49,20	36,00	29,80	18,65	13,67	11,88	6,54	4,44	2,87
8	105,60	77,40	68,40	45,40	35,47	27,90	18,45	13,63	10,90	6,39	4,03	2,72
9	104,40	75,60	67,60	43,60	33,20	26,80	18,35	13,43	10,33	6,30	3,94	2,66
10	99,60	75,60	66,40	43,40	32,93	26,70	16,50	13,10	10,20	6,06	3,87	2,35
11	98,40	75,00	66,00	42,80	31,73	26,20	16,20	13,10	10,05	6,04	3,83	2,33
12	96,00	72,60	65,20	41,00	30,40	24,90	15,88	12,40	9,85	6,03	3,66	2,30
13	96,00	70,20	64,80	40,40	30,13	24,40	15,45	11,57	9,73	5,96	3,60	2,26
14	96,00	66,00	64,80	40,40	29,60	24,20	15,00	11,23	9,15	5,85	3,59	2,24
15	94,80	64,80	64,00	39,80	29,33	24,00	13,85	11,03	9,13	5,85	3,46	2,18
16	92,40	64,20	62,40	39,80	29,20	23,80	13,55	9,87	9,08	5,65	3,46	2,13

Momentos-L e Razões-L das séries de intensidade de precipitação

2 HORAS

Estações	n	l_1	t_2	t_3	t_4
Andorinhas	42	39,91	0,1371	0,1864	0,1535
Apolinário	40	31,75	0,1056	0,2369	0,1221
Faz. Santo Amaro	42	34,02	0,0809	0,0876	0,1283
Nova Friburgo	38	21,52	0,1436	0,3665	0,1699
Petrópolis	12	29,46	0,1823	0,2311	0,1631
Posto Garrafão	42	39,74	0,1054	0,1896	0,1144
Quizanga	42	34,25	0,1174	0,3021	0,1704
Teresópolis-PN	16	20,34	0,1947	0,3962	0,3094
Regional		1	0,1222	0,2360	0,1534

3 HORAS

Estações	n	l_1	t_2	t_3	t_4
Andorinhas	42	30,14	0,1529	0,2282	0,1819
Apolinário	40	23,75	0,1158	0,3295	0,2199
Faz. Santo Amaro	42	25,42	0,0789	0,31	0,2697
Nova Friburgo	38	15,42	0,1345	0,3178	0,2201
Petrópolis	12	21,9	0,1458	0,1798	0,2632
Posto Garrafão	42	29,96	0,1072	0,245	0,1788
Quizanga	42	24,59	0,1171	0,2717	0,1255
Teresópolis-PN	16	15,41	0,1904	0,3776	0,2965
Regional		1	0,1230	0,2838	0,2073

4 HORAS

Estações	n	l_1	t_2	t_3	t_4
Andorinhas	42	24,37	0,1649	0,2972	0,215
Apolinário	40	19,01	0,1232	0,3793	0,2733
Faz. Santo Amaro	42	20,44	0,0888	0,3401	0,2662
Nova Friburgo	38	12,51	0,1401	0,2617	0,2131
Petrópolis	12	17,75	0,1289	0,1498	0,2158
Posto Garrafão	42	24,14	0,1161	0,2899	0,2058
Quizanga	42	19,81	0,1170	0,2245	0,1345
Teresópolis-PN	16	12,45	0,1779	0,3865	0,2381
Regional		1	0,1281	0,2973	0,2187

8 HORAS

Estações	n	l_1	t_2	t_3	t_4
Andorinhas	42	14,2	0,1939	0,4171	0,2306
Apolinário	40	11,38	0,1136	0,3614	0,2602
Faz. Santo Amaro	42	11,82	0,1159	0,4749	0,3523
Nova Friburgo	38	7,15	0,1486	0,3106	0,1341
Petrópolis	12	10,86	0,1377	0,145	0,2343
Posto Garrafão	42	14,11	0,1332	0,2878	0,1451
Quizanga	42	11,15	0,1100	0,274	0,1793
Teresópolis-PN	16	7,32	0,1462	0,3259	0,1562
Regional		1	0,1365	0,3441	0,2150

14 HORAS

Estações	n	l_1	t_2	t_3	t_4
Andorinhas	42	8,9	0,1914	0,4593	0,2833
Apolinário	40	7,42	0,1084	0,3172	0,197
Faz. Santo Amaro	42	7,61	0,1208	0,4477	0,307
Nova Friburgo	38	4,5	0,1563	0,3884	0,1803
Petrópolis	12	7,01	0,1273	0,1566	0,3011
Posto Garrafão	42	9	0,1423	0,2647	0,1437
Quizanga	42	6,95	0,1265	0,2772	0,1448
Teresópolis-PN	16	4,49	0,1410	0,2569	0,1448
Regional		1	0,1404	0,3441	0,2101

24 HORAS

Estações	n	l_1	t_2	t_3	t_4
Andorinhas	42	5,75	0,1742	0,4459	0,2781
Apolinário	40	5,17	0,0942	0,3603	0,314
Faz. Santo Amaro	42	5,23	0,1260	0,2723	0,1501
Nova Friburgo	38	2,73	0,1594	0,3778	0,1875
Petrópolis	12	4,36	0,1549	0,2509	0,2992
Posto Garrafão	42	5,95	0,1224	0,2627	0,0872
Quizanga	42	4,42	0,1360	0,2356	0,1101
Teresópolis-PN	16	2,76	0,1267	0,1846	0,1468
Regional		1	0,1357	0,3132	0,1894

ANEXO 15

Listagem das estações do Alto São Francisco com coordenadas geográficas e altitude

	Código	Estação	Município	Latitude	Longitude	Altitude (m)
1	01844000	CURVELO	Curvelo	18º45'58" S	44º25'34" WG	608
2	01844001	SANTO HIPÓLITO	Santo Hipólito	18º18'00" S	44º13'22" WG	510
3	01844010	PONTE DO LICÍNIO (JUSANTE)	Presidente Juscelino	18º40'22" S	44º11'28" WG	560
4	01845002	FAZENDA SÃO FÉLIX	São Gonçalo do Abaeté	18º27'52" S	45º38'48" WG	760
5	01845004	LAGOA DO GOUVEIA	Tiros	18º50'29" S	45º51'05" WG	1035
6	01845008	MORAVANIA	Morada Nova de Minas	18º40' S	45º21' WG	600
7	01845009	TRÊS MARIAS	Três Marias	18º10' S	45º18' WG	570
8	01845010	VILA CANASTRÃO	Tiros	18º34' S	45º43' WG	835
9	01845011	SÃO GONÇALO DO ABAETÉ	São Gonçalo do Abaeté	18º21' S	45º50' WG	800
10	01845012	ANDREQUICÉ	Três Marias	18º17' S	45º00' WG	830
11	01845013	S. GONÇALO DO ABAETÉ	São Gonçalo do Abaeté	18º20'37" S	45º50'12" WG	836
12	01845014	TIROS	Tiros	18º59'59" S	45º57'58" WG	1030
13	01845026	FAZENDA DAS PEDRAS	Três Marias	18º00' S	45º06' WG	600
14	01846003	MAJOR PORTO	Patos de Minas	18º42'25" S	46º02'13" WG	672
15	01943000	MIN. MORRO VELHO	Nova Lima	19º58'45" S	43º51'00" WG	770
16	01943006	SABARÁ	Sabará	19º53'35" S	43º48'54" WG	720
17	01943009	VESPASIANO	Vespasiano	19º41'14" S	43º55'15" WG	676
18	01943010	CAETÉ	Caeté	19º54'00" S	43º40'03" WG	840
19	01943011	INSTITUTO AGRONÔMICO	Belo Horizonte	19º55' S	43º54' WG	850
20	01943012	LAGOA SANTA	Lagoa Santa	19º38' S	43º54' WG	777
21	01943013	CARLOS PRATES	Belo Horizonte	19º54'43" S	43º57'28" WG	915
22	01943022	CAIXA DE AREIA	Belo Horizonte	19º56'42" S	43º54'45" WG	950
23	01943053	AVENIDA DO CONTORNO	Belo Horizonte	19º56'04" S	43º57'07" WG	915
24	01944000	PRUDENTE DE MORAIS - A	Prudente de Morais	19º29'01" S	44º10'14" WG	732
25	01944003	MATEUS LEME	Mateus Leme	19º59'18" S	44º25'48" WG	836
26	01944004	PONTE NOVA DO PARAOPEBA	Juatuba	19º57'20" S	44º18'24" WG	721
27	01944005	BETIM	Betim	19º58'17" S	44º12'06" WG	832
28	01944007	FAZENDA ESCOLA FLORESTAL	Florestal	19º52'47" S	44º25'18" WG	745
29	01944009	PEDRO LEOPOLDO	Pedro Leopoldo	19º38'04" S	44º03'12" WG	698
30	01944010	HORTO FLORESTAL (Paraopeba)	Paraopeba	19º16'05" S	44º24'06" WG	733
31	01944011	JAGUARUNA (Onça do Pitangui)	Onça de Pitangui	19º43'37" S	44º48'24" WG	685

Listagem das estações do Alto São Francisco com coordenadas geográficas e altitude

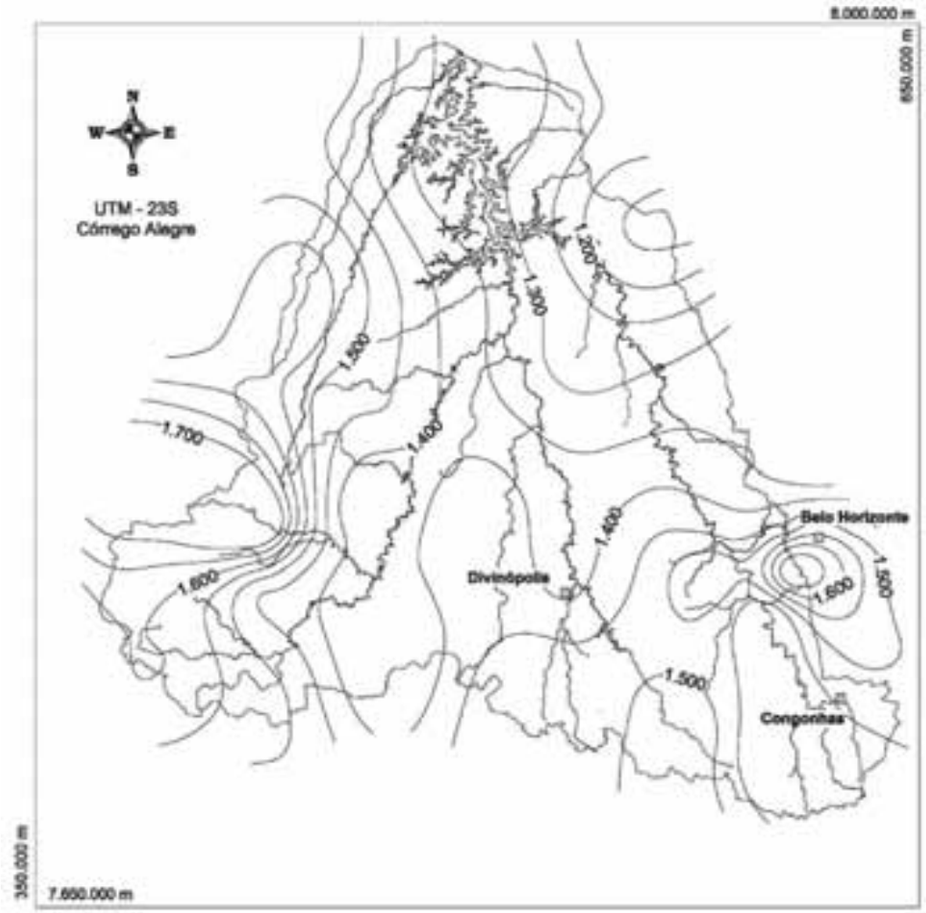
	Código	Estação	Município	Latitude	Longitude	Altitude (m)
32	01944016	SETE LAGOAS	Sete Lagoas	19º28'01" S	44º15'02" WG	780
33	01944018	CAETANÓPOLIS	Caetanópolis	19º17'33" S	44º24'40" WG	738
34	01944019	FÁBRICA TECIDOS S. ANTÔNIO	Sete Lagoas	19º28'03" S	44º14'14" WG	751
35	01944021	VELHO DA TAIPA	Conceição do Pará	19º41'46" S	44º55'46" WG	585
36	01944023	COMPANHIA INDUSTRIAL B.H..	Pedro Leopoldo	19º36'53" S	44º02'31" WG	720
37	01944024	FAZENDA VARGEM BONITA	Jequitibá	19º14'14" S	44º07'23" WG	636
38	01944027	JUATUBA	Mateus Leme	19º57'20" S	44º20'04" WG	728
39	01944031	PONTE DA TAQUARA	Paraopeba	19º25'23" S	44º32'54" WG	624
40	01944032	PITANGUI	Pitangui	19º41'04" S	44º52'44" WG	696
41	01944040	POMPÉU VELHO	Pompéu	19º16' S	44º49' WG	650
42	01944049	PAPAGAIOS	Papagaios	19º25'42" S	44º43'11" WG	703
43	01944060	PORTO MESQUITA	Pompéu	19º10' S	44º40' WG	670
44	01945000	ARAÚJOS	Araújos	19º56'54" S	45º10'01" WG	813
45	01945002	BARRA DO FUNCHAL	Serra da Saudade	19º23'41" S	45º53'04" WG	720
46	01945004	ESTAÇÃO ÁLVARO DA SILVEIRA	Bom Despacho	19º45'06" S	45º07'01" WG	648
47	01945008	BOM DESPACHO	Bom Despacho	19º44'33" S	45º15'18" WG	750
48	01945013	MATUTINA	Matutina	19º14' S	45º58' WG	1100
49	01945014	ENGENHO RIBEIRO	Bom Despacho	19º41' S	45º23' WG	650
50	01945015	FAZENDA NOVO HORIZONTE	Córrego Danta	19º43' S	45º56' WG	1050
51	01945016	FAZENDA DA CURVA	Luz	19º58' S	45º35' WG	650
52	01945017	PORTO PARÁ	Pompéu	19º18' S	45º05' WG	600
53	01945019	DORES DO INDAIÁ	Dores do Indaiá	19º28'07" S	45º36'06" WG	692
54	01945035	ABAETÉ	Abaeté	19º09'47" S	45º26'33" WG	565
55	01946000	TAPIRAÍ	Tapiraí	19º52'46" S	46º01'58" WG	670
56	01946004	IBIÁ	Ibiá	19º28'32" S	46º32'33" WG	855
57	01946007	FAZENDA SÃO MATEUS	Ibiá	19º31'03" S	46º34'22" WG	870
58	01946009	SÃO GOTARDO	São Gotardo	19º18'55" S	46º02'40" WG	1100
59	01946010	PRATINHA	Pratinha	19º45'05" S	46º22'43" WG	1150
60	01946011	TAPIRA	Tapira	19º55'37" S	46º49'31" WG	1120
61	02043002	LAGOA GRANDE	Nova Lima	20º10'45" S	43º56'34" WG	1350
62	02043004	RIO DO PEIXE	Nova Lima	20º08'16" S	43º53'33" WG	1097

Listagem das estações do Alto São Francisco com coordenadas geográficas e altitude

	Código	Estação	Município	Latitude	Longitude	Altitude (m)
63	02043013	CONGONHAS	Congonhas	20º31'19" S	43º49'48" WG	871
64	02043016	RIO ACIMA	Rio Acima	20º05'15" S	43º47'16" WG	730
65	02043018	CARANDAÍ	Carandaí	20º57'21" S	43º48'03" WG	1056
66	02043042	REPRESA DAS CODORNAS	Nova Lima	20º09'53" S	43º53'31" WG	1200
67	02044002	ITAÚNA	Itaúna	20º04'17" S	44º34'13" WG	859
68	02044003	CARMO CAJURU	Carmo do Cajuru	20º11'32" S	44º47'37" WG	746
69	02044005	CARMO DA MATA	Carmo da Mata	20º33'28" S	44º52'03" WG	846
70	02044006	DIVINÓPOLIS	Divinópolis	20º08'13" S	44º53'31" WG	672
71	02044007	ENTRE RIOS DE MINAS	Entre Rios de Minas	20º39'40" S	44º04'14" WG	885
72	02044008	MELO FRANCO	Brumadinho	20º11'52" S	44º07'15" WG	761
73	02044009	FAZENDA CAMPO GRANDE	Passa Tempo	20º37'31" S	44º26'00" WG	915
74	02044012	IBIRITÉ	Ibirité	20º02'34" S	44º02'36" WG	1073
75	02044016	FAZENDA BENEDITO CHAVES	Itatiaiuçu	20º10'09" S	44º30'54" WG	944
76	02044019	FAZENDA VISTA ALEGRE	Mateus Leme	20º03'05" S	44º27'06" WG	913
77	02044021	ALTO DA BOA VISTA	Mateus Leme	20º06'20" S	44º24'04" WG	905
78	02044024	FAZENDA CURRALINHO	Igarapé	20º00'27" S	44º19'52" WG	754
79	02044026	FAZENDA COQUEIROS	Itaúna	20º07'47" S	44º28'28" WG	975
80	02044036	ITAGUARA	Itaguara	20º24' S	44º28' WG	840
81	02044040	USINA JOÃO RIBEIRO	Entre Rios de Minas	20º38'07" S	44º02'56" WG	850
82	02044046	BONFIM	Bonfim	20º20' S	44º15' WG	952
83	02045001	BAMBUÍ	Bambuí	20º01'16" S	45º57'58" WG	654
84	02045002	IGUATAMA	Iguatama	20º10'44" S	45º42'01" WG	606
85	02045005	LAMOUNIER	Itapecerica	20º28'20" S	45º02'10" WG	738
86	02045010	ARCOS (COPASA)	Arcos	20º17'41" S	45º32'34" WG	791
87	02045011	LAGOA DA PRATA	Lagoa da Prata	20º02'12" S	45º32'07" WG	658
88	02045012	PIUMHI	Piumhi	20º27'31" S	45º56'38" WG	806
89	02045013	STº ANTONIO DO MONTE	Stº Antonio do Monte	20º05'04" S	45º17'48" WG	950
90	02045015	FAZENDA OLOS D'ÁGUA	Pimenta	20º26' S	45º50' WG	810
91	02046007	FAZENDA AJUDAS	Bambuí	20º06'06" S	46º03'18" WG	705
92	02046009	DELFINÓPOLIS	Delfinópolis	20º20'50" S	46º50'46" WG	680

Localização das estações do Alto São Francisco

Isoietas de precipitação média anual do Alto São Francisco (mm)



Precipitações diárias máximas anuais (mm) de 92 estações pluviométricas da bacia do Alto São Francisco

	Estação	N	40/41	41/42	42/43	43/44	44/45	45/46	46/47	47/48	48/49	49/50	50/51	51/52	52/53	53/54	54/55	55/56	56/57	57/58	58/59	59/60	60/61	61/62	62/63	63/64	64/65	65/66	66/67
1	01844000	10		71,4	83,8		101,4	71,8	60,9	85		50	121,2																
2	01844001	22		63	63,3	167,4	58,2				110												92,3		70,8		78,2	63,8	58,2
3	01844010	10																											
4	01845002	11																											
5	01845004	10																											
6	01845008	14																											
7	01845009	17																											
8	01845010	16																											
9	01845011	18																											
10	01845012	18																											
11	01845013	15																					110,4						
12	01845014	15																					127,8	94,6	63,2				66,5
13	01845026	13																											
14	01846003	19																											
15	01943000	39		59,2	144		71,1	60,9	79,2	103,6	126	54,6	96	68,1	100,8	101,1	80,8		81,3	90,4	56,1	84,8	87,4		73,9	88,4	76,7	76,7	
16	01943006	26		69			70,8	58,8	114	57,4	132,4		110	100	75	118		49		68			72	80	100				
17	01943009	31		62,6	61	86	108,6	54,4	90	79				81,4					83,4	63	54,6	86	70	75,8			79	84	
18	01943010	37		72,8	69,4	77,8	74,2	102,2	93,4	75	117,4	47,2	67,4	76	102,6	87	112,8		80,1		95,7	102,3	105,5				75,9		
19	01943011	24					72,2	63,2	66,2	73,5		78	90	80	69,8	71,8	88,2	56	116,2	56,8	71,8	76	109	133	94	108,4	80	86,8	87,5
20	01943012	20		45	71,4	77	71,6			79	81,5	103	86,4			66,5	64,5	49,6	72	72,5	40,5		94,8		97,7	137,5	78		
21	01943013	13		65	69	94,4	139	56,6	149	85,4		74,6	118,6	89,4	63,8	75	71,4												
22	01943022	16			86																				61	116			

Precipitações diárias máximas anuais (mm) de 92 estações pluviométricas da bacia do Alto São Francisco

	Estação	68/69	68/69	69/70	70/71	71/72	72/73	73/74	74/75	75/76	76/77	77/78	78/79	79/80	80/81	81/82	82/83	83/84	84/85	85/86	86/87	87/88	88/89	89/90	90/91	91/92	92/93	93/94
1	01844000				60,2		56,3																					
2	01844001		62		66,8	57,2	78,4		51,4	53,3					82		123	43,3	59,9	74,6					60			
3	01844010									60,6	86,4	84,2	101	70	39,8	104					67	46,2				121,6		
4	01845002		72,8	135,8	72,6	80,0	90,2	89,0	87,2		71,0				90,4		109,0									59,0		
5	01845004									65,0	68,2	81,0				99,0		93,0	75,3				75,4	78,2	72,1	91,1		
6	01845008									73,2		101,2	90,8	80,2	127,2	140,0	104,8		151,4		76,7	66,0	70,3	72,2	65,1	84,3		
7	01845009										80,5	83,2	85,6	85,0	92,6	94,6	87,0	71,4	60,0	82,0	65,6	104,0	53,4	103,0	52,4	78,6	79	
8	01845010									98,4		125,0	77,8	128,0	80,4	84,2	170,0	67,0	40,3	97,0	57,6	89,6	70,9	87,9	105,7	80,6		
9	01845011									61,2	75,0	75,2	124,0	166,0	81,8	83,2	92,6	98,8	77,0	103,0	73,6	73,0	80,0	58,0	75,1	95,0	78,4	
10	01845012									62,4	92,4	114,8	109,8	105,8	75,6	66,8	104,6	122,4	70,0	74,0	79,6	73,2	69,8	79,4	82,2	95,2	61,8	
11	01845013				50						70,2	107,8	126,8	88,4	80	88	123	79	78	95	78					92,3	83,2	
12	01845014		72,6						62,2					86,9		95,1	132,2	83,3			72,1		76	102	78,6		63,2	
13	01845026														62,3	71,8	104,4	79,6	94,6	68,4	61,2	86,4	63,4	112,0	48,8	78,0	82,8	
14	01846003		63,7		40,5		81,6	87,4			54,2	146,6	83	75		98	96	83	75	67	70	71,2	62,2	71		95	75,1	
15	01943000	61,7	104,4		55,1	74,4	78,3		72	52		119,1	130			140,4	99,1	70,3	110	100		76,2	51	75				
16	01943006				70		75		58,4		70,8	100	72,6	70,6	70,4		87,4			80,4	64,2							
17	01943009	54				65,2	128,9		61,2	32,8		210,8	102,6	77,9		98,2			114	62,5	60				60,8	62	88	
18	01943010	50,7			52						66,9	210,2	92,1	86,5	86,3	123,6	84,6	64,6	80,7	73	83,4	73,6	57,2	97,7	116,2	100,9		
19	01943011	66,4	90																									
20	01943012	108,5	70																									
21	01943013																											
22	01943022						80		68,1		67,3		97,5	78,2	83				138,4		112,4	79,2	71	110	124		147,3	

Precipitações diárias máximas anuais (mm) de 92 estações pluviométricas da bacia do Alto São Francisco

	Estação	N	40/41	41/42	42/43	43/44	44/45	45/46	46/47	47/48	48/49	49/50	50/51	51/52	52/53	53/54	54/55	55/56	56/57	57/58	58/59	59/60	60/61	61/62	62/63	63/64	64/65	65/66	66/67
23	01943053	15							161,5	94,4	77	68,5	122		78,8	111,8	85,4	74,4	98,7			69,8	79,8		74		76,7	116,9	
24	01944000	33		55,4	75,2		72,9	65,6	88,7	87,7	81,9	105,9	156,8	68,8	47,6	83,1	101,5	71,7	96	69	56,4	71	83,9	76,4	96,6		102,8	79,2	62,5
25	01944003	34		48,1	65,4	94,8	64,2	50,8	79,8	122,8	73,8	62,6	120,8	60,4	128,4	80,4	93,8	113,2	122,8	73,4	83,2	73,4	60,2	52,4	85,4	110,2	60,8	80,2	48,0
26	01944004	48		68,8			67,3		70,2	113,2	79,2	61,2	66,4	65,1	115,0	67,3	102,2	54,4	69,3	54,3	36,0	64,2	83,4	64,2	76,4	159,4	62,1	78,3	74,3
27	01944005	15					64,2	94,0		104,0	85,0	54,0		84,0	68,1		119,0		83,0	51,0			87,0						
28	01944007	34		72,0	84,3	78,2								58,8				65,6	60,4	71,5	61,6	130,4	61,8	80,0		153,2	64,4	79,8	70,2
29	01944009	26		72,2	57,2	76	54	64,2	70							72						55	74		80			75	102,2
30	01944010	27		108,6																72,4	71,6	136,8		72,0			67,0	95,0	55,0
31	01944011	30						74,6					100,2		64,0						55,2		66,2	73,1	75,0	104,0	75,2	58,1	88,8
32	01944016	22										120,5	149,8	68,1			70,1	68,8	99,6	72,5	64,5	98,5	86,4	111,6	82		105	103,5	
33	01944018	18																	71,7	54,5		116,3	72,5	51,6		78,2	76,8		
34	01944019	12																		64,1	63,4	75,8	84,6	91,6	96,3	79,2			77,8
35	01944021	22																								71,0	84,8		70,2
36	01944023	10																						90	90,4	68		70	82
37	01944024	21																						80,2	48,2		57,2		55,3
38	01944027	18																											
39	01944031	11																											
40	01944032	15																											
41	01944040	17																											
42	01944049	12																											
43	01944060	16																											
44	01945000	24			74,0	105,6		88,0		50,8	100,5		101,8	51,2	32,1	80,6	74,0	112,0	74,0	100,0	97,0	39,0	69,0		62,2	72,2			60,0
45	01945002	26		59,1					130,4	60,4	48,0	48,2	88,2	118,2	76,3	77,0				66,4									

Precipitações diárias máximas anuais (mm) de 92 estações pluviométricas da bacia do Alto São Francisco

	Estação	68/69	68/69	69/70	70/71	71/72	72/73	73/74	74/75	75/76	76/77	77/78	78/79	79/80	80/81	81/82	82/83	83/84	84/85	85/86	86/87	87/88	88/89	89/90	90/91	91/92	92/93	93/94
23	01943053																											
24	01944000	60,4	88,2		77,8	99,9	70,6	93,9	76,6	55,4	152,9																	
25	01944003	58,8	110,4	83,2	78,0	98,6	86,0	61,1	72,0																			
26	01944004	41,0	101,6	85,6	51,4	70,3	81,3	85,3	58,4	66,3		72,8	100,0	78,4	61,8	83,4	93,4	99,0	133,0	101,0	109,0	88,0	99,6	74,0	94,0	99,2	101,6	
27	01944005				55,0	54,6		51,4	40,2																			
28	01944007	69,6	60,0	59,8	82,6	79,0	94,0	80,0	68,0	104,6	79,2	86,4			78,0			69,0		60,0	86,0	166,0	53,3	105,1		78,8		
29	01944009	65	90				60,6		68		134,6	45,2		50,1	69	39,2	74,2			62,2	49,5	67,5	62					
30	01944010	57,3	58,5	103,0	80,8	129,2	47,8	71,8	74,0	58,4	97,6	178,0	146,0	84,2	57,3			66,6			55,4	75,2	65,0	67,6				
31	01944011	58,3	91,2	107,4				51,8		88,0	72,5	55,6	65,5	101,3	65,3	69,3	82,8		90,2	58,2	57,3	122,3	74,3	100,3		84,5		
32	01944016	76,7				97,2	75,7	84,8	90,8	60,4	148,5	146,7																
33	01944018			85,8	69,2	136,2	83,8	74,5	85,3	60,8	64,5	157,5	147,3	75,2														
34	01944019	80,3	66,3			95	75																					
35	01944021	57,1	107,0	114,8	157,0				51,8	75,0		55,0	123,7		60,2	56,0		72,1	94,3	76,3	59,2	126,3	56,8	71,6	64,6	87,6		
36	01944023	91,6						68	77	45,5		151																
37	01944024	58,3			73,5	102,5	67,8					125			60	123		52	102,6	57,6	49,2	67,6	55,2	70	200	69	63,2	
38	01944027					69,6	93,0	115,6	57,0	90,4	63,3	70,3	108,0	91,3	86,6	102,8	84,7	93,4	140,6	85,4	76,8	124,7		95,5				
39	01944031							103,8	71,2	76,0	89,4			90,4	57,2	83,8		75,6	94,0	54,8			50,8					
40	01944032								66,5		76,8	54,0	130,0	116,0	73,8	91,8	83,0	61,0			109,9	139,9	71,7	67,9	70,8	91,8		
41	01944040									68,0	139,2	67,2	114,8	50,0	80,2	80,3	112,4	64,6	93,0	74,0	40,0	72,0	47,0	60,0	87,0	110,0		
42	01944049									60,1	153,2	111,0	92,2	75,0	54,1	97,0	53,0			97,6		61,6	100,6	69,8				
43	01944060									49,6	70,2	90,8	80,4	85,0	70,2		75,0	63,0	82,0	75,0	66,2	60,0	111,0	69,3	85,0	75,0		
44	01945000			42,5				78,2	61,6	85,4	70,4																	
45	01945002					75,0	69,0	80,0		56,0	74,2	70,4	80,0	88,0	68,0			93,0	62,2	46,0		95,0	47,0	66,2		78,0		

ANEXO 15

Precipitações diárias máximas anuais (mm) de 92 estações pluviométricas da bacia do Alto São Francisco

	Estação	N	40/41	41/42	42/43	43/44	44/45	45/46	46/47	47/48	48/49	49/50	50/51	51/52	52/53	53/54	54/55	55/56	56/57	57/58	58/59	59/60	60/61	61/62	62/63	63/64	64/65	65/66	66/67
46	01945004	15		69,0	52,0	57,2		78,3	98,0				105,4								84,0	50,0							
47	01945008	11																											
48	01945013	17																											
49	01945014	16																											
50	01945015	16																											
51	01945016	15																											
52	01945017	18																											
53	01945019	21																			69,2	72,2		58,2	96,2	126,4	78,8		
54	01945035	10																											
55	01946000	33		75,0	56,0		73,0		144,0	60,0	80,1	96,0	76,1	93,5	58,0	70,2	62,2					80,0		79,1		63,0	59,1		
56	01946004	34							140,4	78	52	77	64	88	60		52	73,2	82,4			53,2			105,2		67,6	78	67
57	01946007	23																											
58	01946009	17																											
59	01946010	14																											
60	01946011	14																											
61	02043002	27			79		130	98,5		96,5	165,1			71,9	90,2	55,9				92,5	82,6	64,8				108,2	62,2		
62	02043004	27					76,7	100,6	78,5	103,9	111,8	66,5	100,6	66,3	104,6	86,1	159	97,5				73,7		79,5	80,8	92,5			
63	02043013	31		51,6	90,2	49,2	45,4	47,0			108,2	56,8	102,2	55,2	96,6	52,8	110,4	68,2	71,2	60,2	65,8		97,6	57,2	71,4	52,6	102,4	79,6	
64	02043016	10				76,4	62,4			92	97	89,4	84	55,2		58,2	81,2	48											
65	02043018	30				66,4				67,4					50,6	75,8		75,6	74,4	50,6				58	70,2	55,2			
66	02043042	11																											
67	02044002	24			51,0		80,2		90,0																				

Precipitações diárias máximas anuais (mm) de 92 estações pluviométricas da bacia do Alto São Francisco

	Estação	68/69	68/69	69/70	70/71	71/72	72/73	73/74	74/75	75/76	76/77	77/78	78/79	79/80	80/81	81/82	82/83	83/84	84/85	85/86	86/87	87/88	88/89	89/90	90/91	91/92	92/93	93/94	94/95
46	01945004							72,4			84,2	74,8						68,2				67,0			68,8	76,0			
47	01945008									67,8	77,2	68,8	83,2					118,0	85,8	74,0	113,6	105,6	75,0		98,0				
48	01945013									74,4	73,2	81,0	67,0	76,0	90,0	88,3	128,0	74,6	62,4	67,6		67,0	70,0	83,4	89,6	130,0	68,6		
49	01945014										55,6	106,2	100,7	87,0	66,4	78,0	100,2	113,8	76,6	99,8	64,6	73,6	58,4	87,8	71,2	91,4			
50	01945015									71,9		63,0	72,4	66,0	71,0	96,0	135,0	114,0	130,0	99,0	67,2	68,6	56,6	110,8	99,0	87,6			
51	01945016										68,4	91,6	67,8	74,0	66,0	96,2	143,2	46,0	76,2		98,6	60,0	78,0	86,2	195,2	81,2			
52	01945017						66,0	66,0	79,2	64,8	60,2	73,2	80,2	56,8	62,6	98,2	146,6	60,6	71,6	80,6	65,0		54,5	93,0	69,1				
53	01945019	129,1	110,2	104,6					62,1		64,2	119,2		94,0		74,0			100,9	68,0	76,9		56,4	89,4	70,2	116,0			
54	01945035									108	87,8	79,2		58,7	82,2		96,2		90			86,1					78,4	87,3	
55	01946000		88,4	84,5	77,3	70,2	72,4	47,3	63,2	126,1	86,4	93,0	61,3	69,4	73,4		81,0	70,1	104,0	73,0									
56	01946004	59,4	77	94		93,8			70,2	71	108,2	66			59	58	74	91		125,7	120			110	71,7	124	75	71,4	
57	01946007	104,8	70	62,8	60	71,2	88,6	46,6		61,3		64	85,9	86,2	71,2	53,2	47,8	95,4	78	57,3	122,3			75,9	68,2	37	56,5	66	
58	01946009									67	65,5	103,2	82,6	79,7		95,8	100,3	80	78	72,3			50,5	87	81,5	113	80	95,5	71,3
59	01946010										70		67	80		95	125	65	62,5	67,3	52			108	136	100	78,3	105,5	
60	01946011									91	83	69	96,4		76,2	81,6	56,8	67	68,2					77	72,8	68,4	72,4	70	
61	02043002				63	54,1	74,3		64,9			105,2	79,2	87,1	71,6	78			76,2	56,8	72,1			65,1	79,4				
62	02043004							65,2	52,6		92	125		96,3	85,9	90				75,4			58,2		67,3	120,2			
63	02043013	85,4		67,6	97,4	68,2	66,8	69,8			88,6				54,2							79,0							
64	02043016																												
65	2043018	51,8	121,8	95	73,2		57	61,4	70	80		125,1	109,5	60	57,5	107	95		65	78,2	66	77,2		89					
66	02043042										68,4	97		92,2	70,1	63	76,2				82,3	108,4		76,4	100	130,1			
67	02044002	70,0		59,0	115,0	89,0		70,4	80,3	60,3	71,4		113,0	90,4	80,2	79,4	85,4	60,4	92,4	48,1	85,3	75,4		87,2	70,4	80,4			

Precipitações diárias máximas anuais (mm) de 92 estações pluviométricas da bacia do Alto São Francisco

	Estação	N	40/41	41/42	42/43	43/44	44/45	45/46	46/47	47/48	48/49	49/50	50/51	51/52	52/53	53/54	54/55	55/56	56/57	57/58	58/59	59/60	60/61	61/62	62/63	63/64	64/65	65/66	66/67
68	02044019	18		74,8	86,2	70,6	82,2	54,8	76,6	64,4	94,2	106,2	66,6	71,8		72,4					53,0	73,2	129,0	71,0	73,1	91,2	73,2	62,1	91,1
69	02044021	18		95,0	65,0	59,0		69,0	79,3	89,0		100,2		73,1		86,0			82,1				91,1		60,1	117,1			74,0
70	02044024	15		78,4	58,6	92,6	63,3		107,8		57,2	83,1											149,3	57,4		81,9			
71	02044026	17		75,0		105,3	99,0	62,0	62,0	90,0	94,0	70,0			89,0	75,0						95,0	115,0	97,0			80,0		
72	02044036	16			68,0	75,4	78,0	75,0					87,3	94,0	43,0	86,3				68,2				55,0				62,2	
73	02044040	11							65,0		118,2					74,8	105,6	80,0			93,2	77,0		63,0	87,2				94,0
74	02044046	14						62,5	65,0	66,4		62,5	71,4		64,0				64,2	107,4		74,0		87,8	80,8	100,0	86,8		77,4
75	02045001	29																											
76	02045002	31																											
77	02045005	44																											
78	02045010	12																											
79	02045011	17																											
80	02045012	12																											
81	02045013	16																											
82	02045015	16																											
83	02046007	38			62,6	69,2	140,4	103,4	82,2		63,8				61,4											99,8		79,2	73,8
84	02046009	22		61,0	61,4				142,0	70,0		62,4	88,0			55,0	50,0	103,0	71,6	53,4	65,6					74,4			
85	02044003	38		62,2	116,4	79,2	78,2	64,8	75,2	74,8	166,8		125,0	68,6	78,6	93,0	58,2	170,0	80,2	93,6	70,2	68,4	83,2		68,6	125,2	98,0		
86	02044005	22																											
87	02044006	31																											
88	02044007	30																											
89	02044008	21																											
90	02044009	30																											
91	02044012	33		50,7	82,4	73,0	110,0	76,0	105,6	81,3	57,2	86,4	69,0	90,5		71,5	45,4		54,1	69,0									
92	02044016	13		83,6	61,4	88,4	54,4	90,2	73,2			55,4	93,2	60,4	108	74	66,2	94	82	74	56	73	99	75	57,6	58,2	84,2		

Precipitações diárias máximas anuais (mm) de 92 estações pluviométricas da bacia do Alto São Francisco

	Estação	68/69	68/69	69/70	70/71	71/72	72/73	73/74	74/75	75/76	76/77	77/78	78/79	79/80	80/81	81/82	82/83	83/84	84/85	85/86	86/87	87/88	88/89	89/90	90/91	91/92	92/93	93/94	94/95
68	02044003	78,5		91,2		91,1		59,3	102,3	84,0	56,0	81,2	133,0	77,2	79,0	91,2	72,0	64,2	72,0	79,4						80,8			
69	02044005	83,1	133,2	82,0		69,0	100,5	62,1	70,0	96,0																			
70	02044006			100,0	72,2	75,0	85,0	64,0	97,3	97,1	67,0	56,0	118,1	62,4	70,1	78,0	73,4	69,3	79,0	63,0	92,0	69,0			121,2	52,3			
71	02044007			65,0		81,8	66,8		43,7	52,1	79,2	107,2	72,4	64,0		75,3	108,3	76,9	69,0	61,1	70,0	67,0							
72	02044008						92,4	78,2		84,2	78,3		65,1	95,3	82,5		66,3	59,9	74,3										
73	02044009	77,0	76,0	78,0	82,0	80,2	80,6	58,0				91,0	76,0	66,6	146,0		79,0		95,0	82,4	120,0	74,0	60,0	74,0	74,0	113,0			
74	02044012	59,2			57,0	79,3		66,2	46,1		117,0		105,0	144,4	94,6	96,2	90,4	107,8	117,4	61,2		75,4	79,8	85,4	129,4	71,6			
75	02044016									60,6	53,2	56,0	94,8	55,2	61,6	85,5	97,4			81,4	75,2	59,6	89,6			79,2			
76	02044019				89,2	63,0	100,0	62,0	50,2	70,2	75,2	52,0	64,2	50,0		90,1		82,0	87,2	78,0	95,0	83,0	45,0	97,0				
77	02044021						76,5	105,0	62,6	80,5	64,0	71,6	94,2	65,0	80,2	81,1	111,8	108,6	87,8	107,0	90,0		80,4	106,4	112,8				
78	02044024							95,0	58,0	75,0	60,6	80,0	75,0		74,0	64,2	95,6		113,6	61,2	109,0		70,4	93,8		87,4			
79	02044026								66,2		51,6	89,6	99,0	65,6		64,6	69,4	110,0	85,1	95,2	80,4	101,4	72,4	89,2	69,0	80,2	83,2		
80	02044036									59,0		71,6	82,2	69,0	93,0	149,0	59,0	52,0	147,0	60,4	72,0	110,0	80,0	59,0	130,0	52,0			
81	02044040								65,6	53,0	87,0	123,0	80,2		90,0							72,0	50,5	88,0	97,0	65,0			
82	02044046											74,2	63,4	58,0	69,0	128,4	77,6	73,8	97,4	49,4	85,0	56,0	63,2		75,0	82,6			
83	02045001	54,0	112,2	56,0	50,0	59,2	73,2	61,0		135,0	88,0	71,6	50,0		51,0	95,0	200,0	146,0	75,0	60,2		75,3	72,0						
84	02045002							56,4	69,2	104,2	65,0	74,2	53,8	66,4	46,8	92,2	165,0	63,0	84,6	50,4	95,2	94,8	72,0		100,5	134,3			
85	02045005	97,2	155,0	100,2	56,0			57,0	67,0	82,6	94,6	76,2	104,0	86,2		117,4	139,0	83,4	116,8	53,8	81,6	92,4	60,6	115,0	57,0	117,4			
86	02045010									96,8	86,5	84,3	61,0			88,0	109,0	67,1	146,2	86,4	108,0	54,1	65,0						
87	02045011									94,2	79,1	87,3	71,3	75,9	74,2	97,3	123,4	62,5	69,3	63,5	90,3	63,2		64,5			97,3	112,9	78,3
88	02045012										74,9		70,8	61,8	50,0	117,5			123,4	50,4	87,0	64,0	112,0	140,0		84,0			
89	02045013									81,0	57,0	68,2		96,2	54,0	96,0	102,2	50,0	82,0	107,2	123,0	85,0	72,4	63,0	102,0	73,0			
90	02045015									62,2	93,0	69,8	84,6	65,0	57,0	87,0	81,2	75,0	100,1	45,0	94,0		59,0	31,0	89,3	142,0			
91	02046007	85,0	91,5	76,9		69,2	80,1	52,0	79,0	91,0	57,1	74,2	102,2	70,2	52,1	64,3		70,2	75,3	70,2	127,2	71,2	38,0	67,2	110,0	81,3			
92	02046009																												

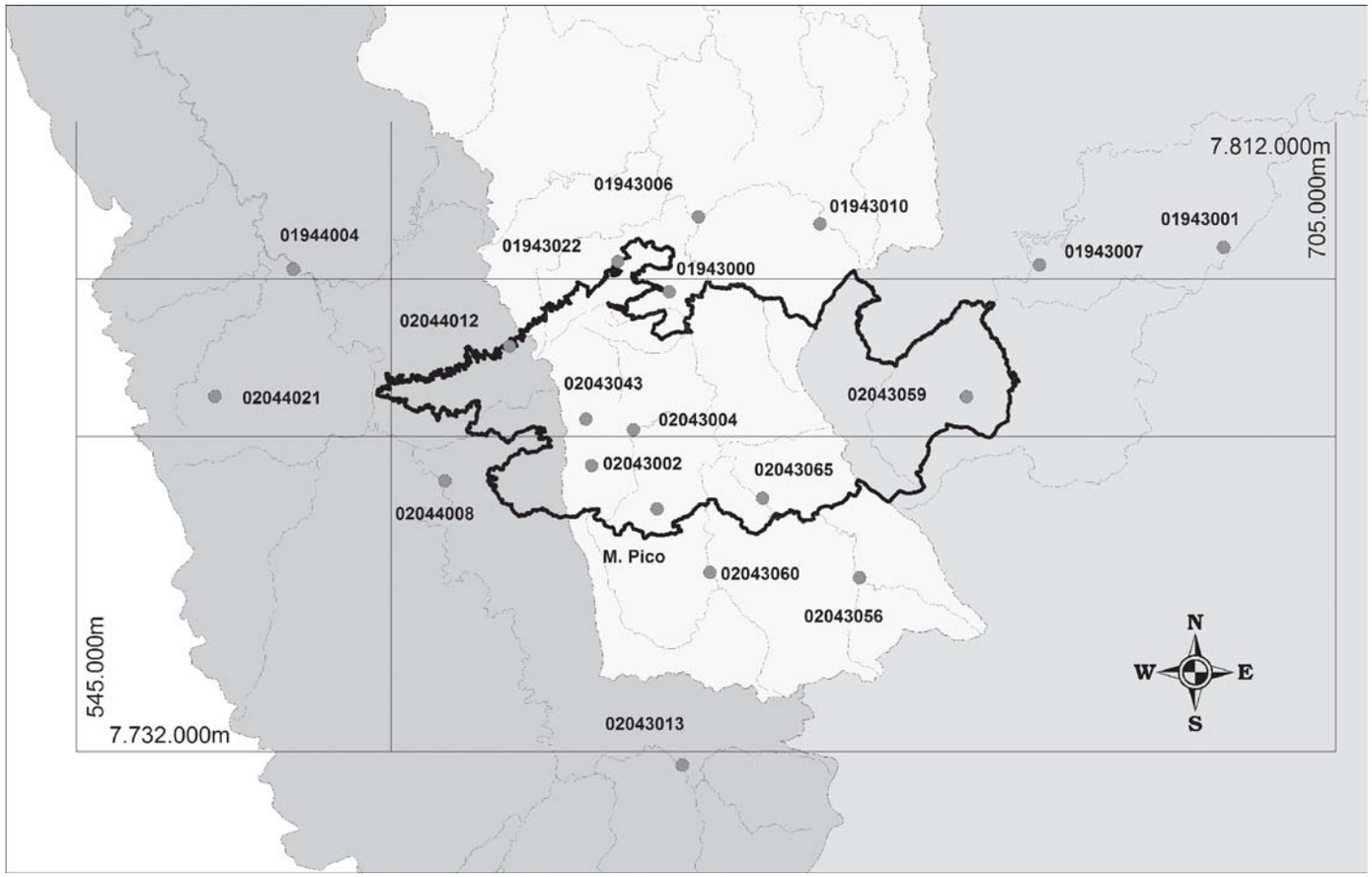
ANEXO 15

ANEXO 16

Estações pluviométricas da APA SUL-RMBH

Código	Nome	SB	Entidade	Latitude	Longitude	Altitude (m)	N
01943000	MIN. MORRO VELHO	41	ANA	19º58'45" S	43º51'00"WG	770 m	59
01943001	RIO PIRACICABA	56	ANA	19º55'22" S	43º10'40" WG	750 m	58
01943006	SABARÁ	41	ANA	19º53'35" S	43º48'54" WG	720 m	56
01943007	SANTA BÁRBARA	56	ANA	19º56'43" S	43º24'04" WG	748 m	59
01943010	CAETÉ	41	ANA	19º54'00" S	43º40'03" WG	950 m	60
01943022	CAIXA DE AREIA	41	ANA	19º56'42" S	43º54'45" WG	1.034 m	55
01944004	Pte. NOVA DO PARAOPEBA	40	ANA	19º57'20" S	44º18'24" WG	721 m	60
02043002	LAGOA GRANDE	41	ANA	20º10'45" S	43º56'34" WG	1.350 m	59
02043004	RIO DO PEIXE	41	ANA	20º08'16" S	43º53'33" WG	1097 m	56
02043013	CONGONHAS	40	ANA	20º31'19" S	43º49'48" WG	871 m	57
02043043	REPRESA DO MIGUELÃO	41	ANA	20º07'34" S	43º57'01" WG	1152 m	25
02043056	FAZENDA ÁGUA LIMPA	41	ANA	20º18'19" S	43º36'59" WG	965 m	17
02043059	COLÉGIO CARAÇA	56	ANA	20º05'49" S	43º29'17" WG	1297 m	41
02043060	ITABIRITO	41	ANA	20º18'04" S	43º47'53" WG	877 m	17
02043065	UHE RIO DE PEDRAS	41	CEMIG	20º12'55" S	43º44'05" WG		10
	MINA DO PICO	41	MBR	20º13'43" S	43º51'47" WG	1500 m	11
02044008	MELO FRANCO	40	ANA	20º11'52" S	44º07'15" WG	761 m	60
02044012	IBIRITÉ	40	ANA	20º02'34" S	44º02'36" WG	1073 m	56
02044021	ALTO BOA VISTA	40	ANA	20°06'07" S	44°24'04" WG	905 m	29

Localização das estações pluviométricas da APA SUL-RMBH



Precipitações anuais (mm) das 19 estações pluviométricas da APA SUL-RMBH

AI	AF	01943000	AI	AF	01943001	AI	AF	01943006	AI	AF	01943007	AI	AF	01943010	AI	AF	1943022	AI	AF	1944004
1941	1942	1414,1	1941	1942	1594,8	1941	1942	1238,7	1942	1943	1987,9	1941	1942	1342,4	1941	1942	1563,8	1941	1942	1248,5
1942	1943	1856	1942	1943	1887,2	1942	1943	1783,5	1943	1944	1466,2	1942	1943	1735,6	1942	1943	2046,6	1942	1943	1319,3
1943	1944	1417,7	1943	1944	1429,4	1943	1944	1366,5	1944	1945	1946,4	1943	1944	1338,4	1943	1944	1639,6	1943	1944	1191,3
1944	1945	1717	1944	1945	1748,4	1944	1945	1738,2	1945	1946	1548,2	1944	1945	1650,2	1944	1945	2161	1944	1945	1440
1945	1946	1545,4	1945	1946	1569,4	1945	1946	1389,7	1946	1947	1441,8	1945	1946	1672,1	1945	1946	1904,1	1945	1946	1251,3
1946	1947	1506,4	1946	1947	1527,7	1946	1947	1388,2	1947	1948	1365,3	1946	1947	1374,1	1946	1947	1659,9	1946	1947	1507
1947	1948	1500,3	1947	1948	1479,7	1947	1948	1305,5	1948	1949	2192,7	1947	1948	1384,8	1947	1948	1750,2	1947	1948	1363,3
1948	1949	2404,1	1948	1949	2212,2	1948	1949	2285,5	1949	1950	1401,5	1948	1949	1988,8	1948	1949	2570,8	1948	1949	1814,1
1949	1950	1300	1949	1950	1367,8	1949	1950	1117,6	1950	1951	1683,1	1949	1950	1223,8	1950	1951	1872,9	1949	1950	1321,9
1950	1951	1688,9	1950	1951	1588,5	1950	1951	1534,2	1951	1952	1492,3	1950	1951	1501	1951	1952	1695,8	1950	1951	1337,8
1951	1952	1402,9	1951	1952	1771,4	1951	1952	1639	1952	1953	1355,2	1951	1952	1367,4	1952	1953	1382,9	1951	1952	1326,7
1952	1953	1236,4	1952	1953	1263	1952	1953	1289	1953	1954	1171	1952	1953	1349,4	1953	1954	1456,4	1952	1953	1300,7
1953	1954	1233,2	1953	1954	858,8	1953	1954	1328,2	1954	1955	1158,3	1953	1954	1097,1	1954	1955	1462,1	1953	1954	1138
1954	1955	1444,6	1954	1955	966,1	1954	1955	1377,7	1955	1956	1001,7	1954	1955	1166,5	1955	1956	1609,4	1954	1955	1121
1955	1956	1557	1955	1956	984,1	1955	1956	1367,3	1956	1957	1476,3	1955	1956	1210,1	1956	1957	1895,6	1955	1956	1453,6
1956	1957	1732,4	1956	1957	1576,3	1956	1957	1716,4	1957	1958	1246,2	1956	1957	1456,4	1957	1958	1705,4	1956	1957	1648,1
1957	1958	1662,3	1957	1958	1116	1957	1958	1601,8	1958	1959	923,8	1957	1958	1504,9	1958	1959	1118,6	1957	1958	1294,3
1958	1959	920,3	1958	1959	881,6	1958	1959	913,1	1959	1960	1675,6	1958	1959	926,4	1961	1962	1849,2	1958	1959	882,8
1959	1960	1528,3	1959	1960	1528,8	1959	1960	1122,4	1960	1961	1571,6	1959	1960	1437,1	1962	1963	1176	1959	1960	1600,8
1960	1961	1696,4	1960	1961	1462,9	1960	1961	1729,8	1961	1962	1298,6	1960	1961	1452,3	1963	1964	1400,8	1960	1961	1487,2
1961	1962	1346,4	1961	1962	1153,8	1961	1962	1463,3	1962	1963	1082	1961	1962	1252,2	1964	1965	2289,6	1961	1962	1347,1
1962	1963	1050,8	1962	1963	1237,4	1962	1963	1219	1963	1964	1239,3	1962	1963	1125,5	1965	1966	1742,9	1962	1963	1249,8
1963	1964	1365,4	1963	1964	1140,1	1963	1964	1226	1964	1965	1948,2	1963	1964	1153,5	1966	1967	1552,4	1963	1964	1297,6
1964	1965	1932,1	1964	1965	1790,2	1964	1965	2034,1	1965	1966	1773,8	1964	1965	1841,2	1967	1968	1405,8	1964	1965	1673
1965	1966	1654,1	1965	1966	1441,7	1965	1966	1599,6	1966	1967	1492,7	1965	1966	1507,8	1968	1969	1652,5	1965	1966	1452,3
1966	1967	1311,7	1966	1967	1164,9	1966	1967	1164,5	1967	1968	1137,7	1966	1967	1286,7	1971	1972	1637,9	1966	1967	1169,4

ANEXO 16

Precipitações anuais (mm) das 19 estações pluviométricas da APA SUL-RMBH

AI	AF	01943000	AI	AF	01943001	AI	AF	01943006	AI	AF	01943007	AI	AF	01943010	AI	AF	1943022	AI	AF	1944004
1967	1968	1170,8	1967	1968	1260,7	1967	1968	1252,9	1968	1969	1520,5	1967	1968	1116	1972	1973	1651	1967	1968	1189,1
1968	1969	1472,2	1968	1969	1288,2	1968	1969	1387,9	1969	1970	1489	1968	1969	1303,7	1973	1974	1632,4	1968	1969	1219,6
1969	1970	1486,6	1969	1970	1459,4	1969	1970	1193,3	1970	1971	1145,3	1969	1970	1038,7	1974	1975	1401,2	1969	1970	1306
1970	1971	998,2	1970	1971	1036,5	1970	1971	987,9	1971	1972	1627	1970	1971	871,7	1975	1976	1321,7	1970	1971	1012,7
1971	1972	1548,7	1971	1972	1487,9	1971	1972	1424,6	1972	1973	1547,7	1971	1972	1325,5	1976	1977	1697,7	1971	1972	1530,8
1972	1973	1477,2	1972	1973	1619,7	1972	1973	1610,4	1973	1974	1599,3	1972	1973	1619,8	1977	1978	1779,9	1972	1973	1486,9
1973	1974	1478,3	1973	1974	1454,3	1973	1974	1404,1	1974	1975	1275,2	1973	1974	1414,7	1978	1979	2442,9	1973	1974	1395,2
1974	1975	1441,1	1974	1975	1338,7	1974	1975	1508,2	1975	1976	1092,1	1974	1975	1198,2	1979	1980	1668,8	1974	1975	1089,9
1975	1976	1231,7	1975	1976	1250,9	1975	1976	1256,7	1976	1977	1674,9	1975	1976	1058,1	1980	1981	1549,6	1975	1976	1310,9
1976	1977	1770,4	1976	1977	1353,7	1976	1977	1485,3	1977	1978	1573,7	1976	1977	1454,3	1981	1982	2149,5	1976	1977	1291,1
1977	1978	1708	1977	1978	1414	1977	1978	1464,4	1978	1979	2003,2	1977	1978	1406,9	1982	1983	2272,2	1977	1978	1272,6
1978	1979	2219,5	1980	1981	1204,1	1978	1979	2052,5	1979	1980	1450,4	1978	1979	1981,2	1983	1984	1399,9	1978	1979	2027,2
1979	1980	1624,7	1981	1982	2127	1979	1980	1660	1980	1981	1081,2	1979	1980	1467	1984	1985	2596,8	1979	1980	1696,6
1980	1981	1423,6	1982	1983	1725,1	1980	1981	1343,4	1981	1982	1726,7	1980	1981	1500,9	1985	1986	1380,9	1980	1981	1341,2
1981	1982	1912,3	1983	1984	1214,1	1981	1982	1662,2	1982	1983	1624,8	1981	1982	1781,9	1986	1987	1647,5	1981	1982	1764,4
1982	1983	2078,9	1984	1985	2126,7	1982	1983	1827,4	1983	1984	1203	1982	1983	1746,4	1987	1988	1808,7	1982	1983	1785,8
1983	1984	1303,3	1985	1986	1295,3	1983	1984	1220,2	1984	1985	1808,5	1983	1984	1311,5	1988	1989	1302,7	1983	1984	1728,3
1984	1985	2383,6	1986	1987	1335,1	1984	1985	2063	1985	1986	1224,6	1984	1985	1963,9	1989	1990	1777,2	1984	1985	1879,5
1985	1986	1496,3	1987	1988	1131	1985	1986	1273,3	1986	1987	1002,1	1985	1986	1389,6	1990	1991	1902,1	1985	1986	1429,1
1986	1987	1442,9	1988	1989	1315,7	1986	1987	1235,2	1987	1988	1526,7	1986	1987	1255	1991	1992	1852	1986	1987	1411,5
1987	1988	1575,4	1989	1990	1162,1	1987	1988	1563,5	1988	1989	921,9	1987	1988	1550,7	1992	1993	1917	1987	1988	1606,3
1988	1989	1292,2	1990	1991	1177	1988	1989	1309,9	1989	1990	1149,5	1988	1989	1287,2	1993	1994	1428,1	1988	1989	1289,6
1989	1990	1447	1991	1992	1489	1989	1990	1312,9	1990	1991	1164,3	1989	1990	1484,9	1994	1995	1410,8	1989	1990	1450,8
1991	1992	1616,7	1992	1993	1644,4	1990	1991	1537,9	1991	1992	1274,4	1990	1991	1531,7	1995	1996	1998,8	1990	1991	1446,7
1992	1993	1624,7	1993	1994	1405,2	1991	1992	1590,3	1992	1993	1073,7	1991	1992	1565,8	1996	1997	1940,7	1991	1992	1580,7
1993	1994	1418,6	1994	1995	1193,8	1992	1993	1212,5	1993	1994	1196,9	1992	1993	1202,2	1997	1998	1474	1992	1993	1642,2

Precipitações anuais (mm) das 19 estações pluviométricas da APA SUL-RMBH

AI	AF	01943000	AI	AF	01943001	AI	AF	01943006	AI	AF	01943007	AI	AF	01943010	AI	AF	1943022	AI	AF	1944004
1994	1995	1294,8	1995	1996	1376,8	1995	1996	1531	1994	1995	891,3	1993	1994	1198,9	1998	1999	1537,3	1993	1994	1340,6
1995	1996	1689	1996	1997	1832,6	1998	1999	1222,8	1995	1996	1330,3	1994	1995	1331,5	1999	2000	1777,6	1994	1995	1359,4
1996	1997	1706,2	1997	1998	1688,9	1999	2000	1822	1996	1997	2207,1	1995	1996	1677	2000	2001	1286	1995	1996	1503,4
1997	1998	1096,1	1998	1999	1247,3	2000	2001	1176,8	1997	1998	1365,9	1996	1997	1736,3				1996	1997	1926,5
1998	1999	1342,2	1999	2000	1656				1998	1999	1295,7	1997	1998	1116,8				1997	1998	1235,7
1999	2000	1698,9	2000	2001	1205,5				1999	2000	2246,9	1998	1999	1193,6				1998	1999	1163,3
2000	2001	1047,6							2000	2001	1208	1999	2000	1593,1				1999	2000	1489,9
												2000	2001	1046,6				2000	2001	1091,8
		01943000			**01943001**			**01943006**			**01943007**			**01943010**			**1943022**			**1944004**
Média		1524,4			1418,3			1455,9			1433,9			1401,1			1712,9			1408,9
Mínimo		920,3			858,8			913,1			891,3			871,7			1118,6			882,8
Máximo		2404,1			2212,2			2285,5			2246,9			1988,8			2596,8			2027,2

Precipitações anuais (mm) das 19 estações pluviométricas da APA SUL-RMBH

AI	AF	02043002	AI	AF	02043004	AI	AF	02043013	AI	AF	02043043	AI	AF	02043056	AI	AF	2043059	AI	AF	2043060
1942	1943	1875,8	1941	1942	1443	1941	1942	1348,2	1976	1977	1710,5	1984	1985	1822,4	1942	1943	2475,1	1984	1985	1780
1943	1944	1492,3	1942	1943	1905,9	1942	1943	1522,6	1977	1978	1563	1985	1986	1699,6	1943	1944	1789,3	1985	1986	1370,6
1944	1945	1819	1943	1944	1416,3	1943	1944	1125,5	1978	1979	2182,8	1986	1987	1415,8	1944	1945	2240,5	1986	1987	1431,6
1945	1946	1457,8	1944	1945	1807,1	1944	1945	1415,5	1979	1980	1774,1	1987	1988	1434	1945	1946	2222,7	1987	1988	1228,1
1946	1947	1379,9	1945	1946	1613	1945	1946	1261,4	1980	1981	1555,2	1988	1989	1087,9	1946	1947	1731	1988	1989	1295,6
1947	1948	1370,8	1946	1947	1582,8	1946	1947	1504,5	1981	1982	1669,4	1989	1990	1265,8	1947	1948	1846,8	1989	1990	1417,4
1948	1949	2235,5	1947	1948	1375,4	1947	1948	1154,1	1982	1983	2238,4	1990	1991	1651	1948	1949	2294,6	1990	1991	1781,5
1949	1950	1390,5	1948	1949	2108,1	1948	1949	1609,9	1983	1984	1290,3	1991	1992	1585,2	1949	1950	1760,6	1991	1992	1707,9
1950	1951	1714,6	1949	1950	1439,1	1949	1950	1510	1984	1985	1984	1992	1993	1420,6	1950	1951	1839,8	1992	1993	1548,3
1951	1952	1378,7	1950	1951	1741,8	1950	1951	1638,5	1985	1986	1695,2	1993	1994	1279,2	1951	1952	1800,4	1993	1994	1413,9
1952	1953	1446,6	1951	1952	1438,1	1951	1952	1388,4	1986	1987	1623,4	1994	1995	1072,3	1952	1953	1527,9	1994	1995	1254,6
1953	1954	1295,5	1952	1953	1465,6	1952	1953	1208,6	1987	1988	1697,9	1995	1996	1247,2	1953	1954	1543,1	1995	1996	1506,6
1954	1955	1417,5	1953	1954	1369,2	1953	1954	1014,9	1988	1989	1397,1	1996	1997	1521,5	1954	1955	1536,1	1996	1997	1637,2
1955	1956	1394	1954	1955	1613,6	1954	1955	1311,7	1989	1990	1565,4	1997	1998	1072,3	1955	1956	1795,3	1997	1998	1279,8
1956	1957	1663,2	1955	1956	1482,1	1955	1956	1390,9	1990	1991	1882,7	1998	1999	1026,4	1956	1957	2463,3	1998	1999	1131,4
1957	1958	1600,4	1956	1957	1764,1	1956	1957	1571,3	1991	1992	1751,3	1999	2000	1461,9	1957	1958	1894,6	1999	2000	1537,8
1958	1959	1089,5	1959	1960	1636,1	1957	1958	1567,7	1992	1993	1771,8	2000	2001	1043,9	1958	1959	1321,8	2000	2001	1170,3
1959	1960	1532,9	1960	1961	2086,6	1958	1959	1017,6	1993	1994	1547,1				1959	1960	2070,8			
1960	1961	1986,1	1961	1962	1529,9	1959	1960	1366,6	1994	1995	1552,8				1960	1961	2154,3			
1961	1962	1418,6	1962	1963	1056	1960	1961	1586,6	1995	1996	1711,4				1961	1962	1530,9			
1962	1963	1123,9	1963	1964	1583,1	1961	1962	1316,1	1996	1997	2104,9				1962	1963	1593,7			
1963	1964	1372,6	1964	1965	2054,5	1962	1963	1145,7	1997	1998	1194,1				1963	1964	1748,7			
1964	1965	2066,7	1965	1966	1559,6	1963	1964	1256	1998	1999	1307,1				1983	1984	1773,6			
1965	1966	1524,7	1966	1967	1423,8	1964	1965	2068	1999	2000	1600,4				1985	1986	1785,5			
1966	1967	1451,9	1967	1968	1445,7	1965	1966	1622,7	2000	2001	1297,1				1986	1987	1506,7			
1967	1968	1379,3	1968	1969	1536,8	1966	1967	1548,8							1987	1988	2292,6			
1968	1969	1238,9	1969	1970	1510,5	1967	1968	1406,3							1988	1989	1604			
1969	1970	1389,2	1972	1973	1854,7	1968	1969	1076,9							1989	1990	2169,8			
1970	1971	1108,6	1973	1974	1497,4	1969	1970	1363,2							1990	1991	2473,5			
1971	1972	1596,3	1974	1975	1350,2	1970	1971	1153,7							1991	1992	2196,2			
1972	1973	1673,9	1975	1976	1351,1	1971	1972	1558							1992	1993	2102,3			

Precipitações anuais (mm) das 19 estações pluviométricas da APA SUL-RMBH

AI	AF	02043002	AI	AF	02043004	AI	AF	02043013	AI	AF	02043043	AI	AF	02043056	AI	AF	2043059	AI	AF	2043060
1973	1974	1451	1976	1977	1618,1	1972	1973	1642,8							1993	1994	1693,2			
1974	1975	1317	1977	1978	1517,8	1973	1974	1241,4							1994	1995	1591,1			
1975	1976	1364,2	1978	1979	2274,3	1974	1975	1253,4							1995	1996	1851,2			
1976	1977	1676,1	1979	1980	1674,2	1975	1976	1416,1							1996	1997	2524,1			
1977	1978	1487,2	1980	1981	1476	1976	1977	1537,3							1997	1998	1663,5			
1978	1979	1971,3	1981	1982	1637,2	1977	1978	1392,3							1998	1999	1401,4			
1979	1980	1628	1982	1983	1957,8	1978	1979	1798,5							1999	2000	2119,3			
1980	1981	1380	1983	1984	1196,3	1979	1980	1356,8							2000	2001	1420,6			
1981	1982	1735	1984	1985	2081	1980	1981	1366,6												
1982	1983	2185,2	1985	1986	1540,3	1981	1982	1431,2												
1983	1984	1324,8	1986	1987	1494,7	1982	1983	2096,7												
1984	1985	2133,5	1987	1988	1436	1983	1984	1073,7												
1985	1986	1546,6	1988	1989	1377,5	1984	1985	1779,5												
1986	1987	1575,4	1989	1990	1429,7	1985	1986	1480,9												
1987	1988	1615,4	1990	1991	1782,2	1986	1987	1326,7												
1988	1989	1373,3	1991	1992	1622	1987	1988	1278,5												
1989	1990	1596,4	1992	1993	1477,4	1988	1989	926,4												
1990	1991	1874,5	1993	1994	1454,6	1989	1990	1239,2												
1991	1992	1521,7	1994	1995	1553,6	1990	1991	1497,9												
1992	1993	1574	1995	1996	1773,3	1991	1992	1279,3												
1993	1994	1457,5	1996	1997	2000,6	1992	1993	1251,1												
1994	1995	1526,2	1997	1998	1290,7	1996	1997	1731,3												
1995	1996	1856,5	1998	1999	1178,8	1997	1998	1196,7												
1996	1997	2142,2	1999	2000	1662	1998	1999	1131,3												
1997	1998	1428,4	2000	2001	1153	1999	2000	1496,4												
1998	1999	1274,9				2000	2001	1143,5												
1999	2000	1654,5																		
2000	2001	1396,7																		
		02043002			**02043004**			**02043013**			**02043043**			**02043056**			**02043059**			**02043060**
Média		1558,5			1583,9			1392,9			1666,7			1359,2			1880,8			1440,7
Mínimo		1089,5			1056,0			926,4			1194,1			1026,4			1321,8			1131,4
Máximo		2235,5			2274,3			2096,7			2238,4			1822,4			2524,1			1781,5

Precipitações anuais (mm) das 19 estações pluviométricas da APA SUL-RMBH

AI	AF	02043065	AI	AF	M. Pico	AI	AF	02044008	AI	AF	2044012	AI	AF	02044021
1986	1987	1205,9	1990	1991	1905,6	1941	1942	1301,6	1945	1946	1648,8	1972	1973	1794
1987	1988	1286,3	1991	1992	1716,6	1942	1943	1585,4	1946	1947	1667	1973	1974	1688,1
1988	1989	1233,1	1992	1993	1645,6	1943	1944	1155,5	1947	1948	1662,7	1974	1975	1133,2
1989	1990	1336,9	1993	1994	1465	1944	1945	1465,1	1948	1949	2452,8	1975	1976	1548,9
1990	1991	1754,7	1994	1995	1356,5	1945	1946	1382,8	1949	1950	1651,3	1976	1977	1232,4
1991	1992	1256,6	1995	1996	1429,1	1946	1947	1346	1950	1951	1726,2	1977	1978	1525,2
1992	1993	1596,1	1996	1997	2111,8	1947	1948	1094,2	1951	1952	1642,1	1978	1979	1952,5
1993	1994	1417,4	1997	1998	1259,4	1948	1949	1699,4	1952	1953	1548,2	1979	1980	1649,5
1994	1995	1229,9	1998	1999	1094,8	1949	1950	1265,2	1953	1954	1551,2	1980	1981	1301,4
1995	1996	1541,4	1999	2000	1759	1950	1951	1394	1954	1955	1546,4	1981	1982	1746,3
			2000	2001	1429,2	1951	1952	1165,9	1955	1956	1650,5	1982	1983	2175,7
						1952	1953	1079,9	1956	1957	1985,1	1983	1984	1595,3
						1953	1954	1267,4	1957	1958	1793	1984	1985	1874,7
						1954	1955	1091,7	1958	1959	1103	1985	1986	1597,6
						1955	1956	1165,8	1959	1960	1843,8	1986	1987	1428
						1956	1957	1299,6	1960	1961	2038,3	1987	1988	1562,1
						1957	1958	1519,8	1961	1962	1570,8	1988	1989	1233,1
						1958	1959	939,8	1962	1963	1511,3	1989	1990	1406,1
						1959	1960	1235	1963	1964	1532,6	1990	1991	1589,4
						1960	1961	1655,5	1964	1965	2451,1	1991	1992	1782,7
						1961	1962	1199,7	1965	1966	1929,2	1992	1993	1700,6
						1962	1963	1085,6	1966	1967	1552,3	1993	1994	1478,4
						1963	1964	1139,9	1967	1968	1537,4	1994	1995	1508,2
						1964	1965	1700,1	1968	1969	1376,4	1995	1996	1651,8
						1965	1966	1325,6	1969	1970	1614,1	1996	1997	2146,3
						1966	1967	1230,9	1970	1971	1280,6	1997	1998	1183,9
						1967	1968	1466,5	1971	1972	1814,5	1998	1999	1398,1
						1968	1969	1027,4	1972	1973	1808	1999	2000	1608,2
						1969	1970	1155,4	1973	1974	1753,4	2000	2001	1192,6
						1970	1971	1121	1974	1975	1417,1			
						1971	1972	1479,9	1975	1976	1626,1			
						1972	1973	1560,1	1976	1977	1825,2			
						1973	1974	1463,7	1977	1978	1699,7			
						1974	1975	1244,5	1978	1979	2158,1			
						1975	1976	1307,3	1979	1980	1945,1			
						1976	1977	1292,8	1980	1981	1645,1			
						1977	1978	1323,4	1981	1982	1869,2			
						1978	1979	1896,5	1982	1983	2345,1			
						1979	1980	1457,6	1983	1984	1397,7			
						1980	1981	1315	1984	1985	2399,4			
						1981	1982	1741	1985	1986	1411,2			
						1982	1983	1826,3	1986	1987	1756,2			
						1983	1984	1187,3	1987	1988	1758,4			
						1984	1985	1655,8	1988	1989	1434,2			
						1985	1986	1379,5	1989	1990	1864,4			
						1986	1987	1389,4	1990	1991	2020,7			
						1987	1988	1223,6	1991	1992	1611,9			
						1988	1989	1047,8	1992	1993	1736			

Precipitações anuais (mm) das 19 estações pluviométricas da APA SUL-RMBH

AI	AF	02043065	AI	AF	M. Pico	AI	AF	02044008	AI	AF	2044012	AI	AF	02044021
						1989	1990	1285,1	1993	1994	1692			
						1990	1991	1370,1	1994	1995	1704,4			
						1991	1992	1224,5	1995	1996	1894,1			
						1992	1993	1233	1996	1997	2311,4			
						1993	1994	1326,6	1997	1998	1468			
						1994	1995	1220,7	1998	1999	1591,4			
						1995	1996	1480,8	1999	2000	1672			
						1996	1997	1790,6	2000	2001	1202,8			
						1997	1998	1256,2						
						1998	1999	1279,4						
						1999	2000	1349,3						
						2000	2001	1029,2						
Média		**02043065**			**M. Pico**			**02044008**			**02044012**			**02044021**
Mínimo		1385,8			1561,1			1336,6			1726,8			1575,3
Máximo		1205,9			1094,8			939,8			1103,0			1133,2
		1754,7			2111,8			1896,5			2452,8			2175,7