Alexandre Xavier Ywata de Carvalho
Daniel Oliveira Cajueiro
Reinaldo Soares de Camargo

INTRODUÇÃO AOS MÉTODOS ESTATÍSTICOS PARA ECONOMIA E FINANÇAS

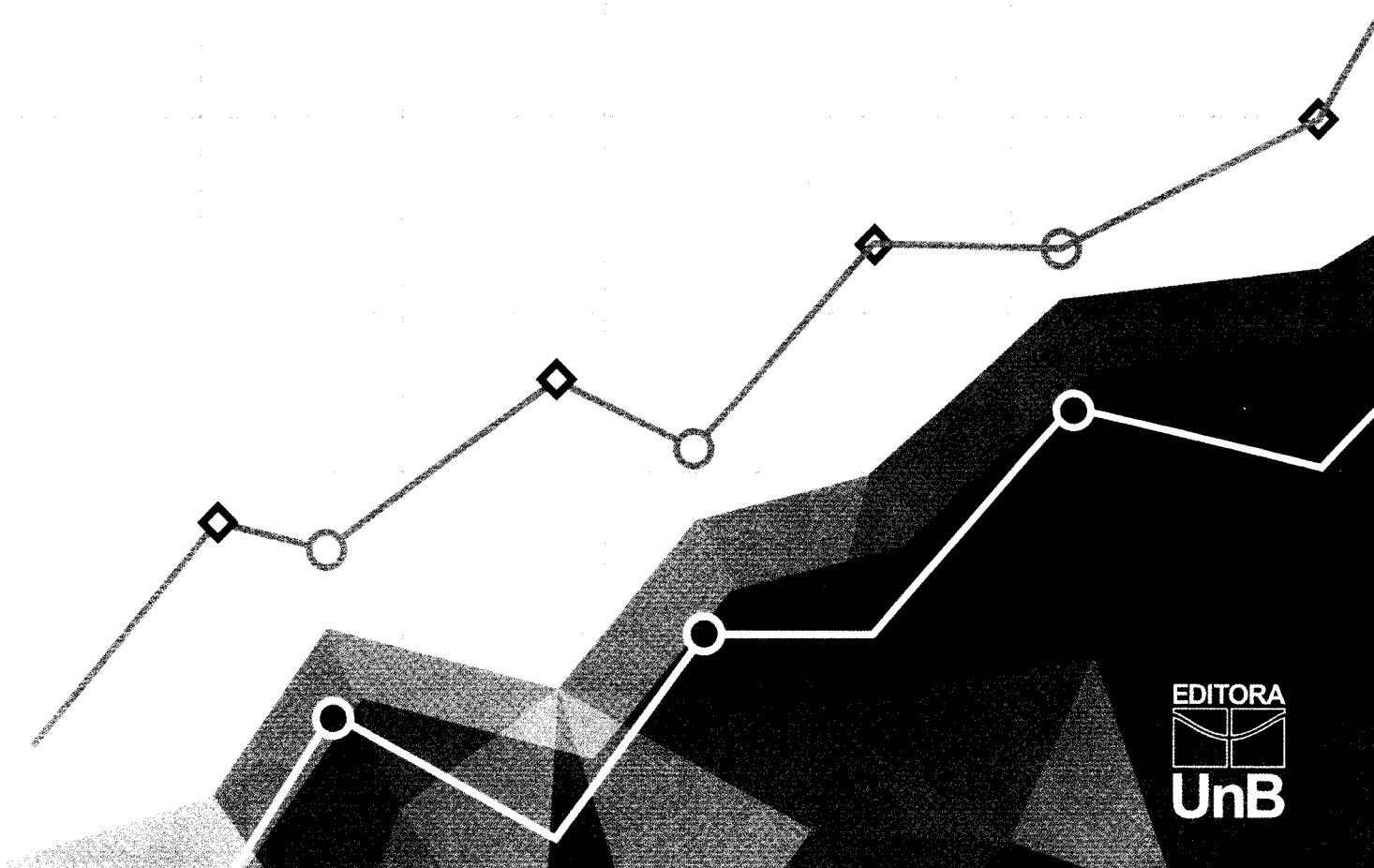
EDITORA UnB

INTRODUÇÃO AOS MÉTODOS ESTATÍSTICOS PARA ECONOMIA E FINANÇAS

Alexandre Xavier Ywata de Carvalho[1]
Daniel Oliveira Cajueiro[2]
Reinaldo Soares de Camargo[3]

INTRODUÇÃO AOS MÉTODOS ESTATÍSTICOS PARA ECONOMIA E FINANÇAS

1 Alexandre Xavier Ywata de Carvalho, Instituto de Pesquisa Econômica Aplicada - IPEA
2 Daniel Oliveira Cajueiro, Departamento de Economia da Universidade de Brasília - UnB
3 Reinaldo Soares de Camargo, Fundação dos Economiários Federais - FUNCEF



EDITORA UnB

Equipe Editorial

Gerência de produção editorial Marcus Polo Rocha Duarte
Revisão Regina Coeli marques
Editoração eletrônica Eduardo Silva de Medeiros

SCS, quadra 2, bloco C, nº 78, edifício OK,
2º andar, CEP 70302-907, Brasília, DF.
Telefone: (61) 3035-4200
Fax: (61) 3035-4230
Site: www.editora.unb.br
E mail: contatoeditora@unb.br

Ficha catalográfica elaborada pela Biblioteca Central da Universidade de Brasília

Carvalho, Alexandre Xavier Ywata de.

C331 Introdução aos métodos estatísticos para economia e finanças / Alexandre Xavier Ywata de Carvalho, Daniel Oliveira Cajueiro, Reinaldo Soares de Camargo. – Brasília : Editora Universidade de Brasília, 2015.
360 p. ; 25 cm.

ISBN 978-85-230-1148-2

1. Estatística básica. 2. Econometria. 3. Probabilidade. 4. Economia. 5. Finanças. I. Cajueiro, Daniel Oliveira. II. Camargo, Reinaldo Soares de. III. Título.

CDU 311:33

Alexandre dedica este livro a Carolina, Francisco, Mariza, Ana Catharina, Andrei e Artur.

Daniel dedica este livro às suas duas princesas Ana Patricia e Barbara, aos seus pais, às suas avós e à memória de seus avôs.

Reinaldo dedica este livro à sua mãe Maria Maura, à sua esposa Ângela e aos seus filhos Fernanda, Reinaldo Jr e William.

Sumário

1. Considerações iniciais

"We are shaped and fashioned by what we love."
Johann Wolfgang von Goethe

As últimas décadas têm testemunhado uma crescente evolução dos métodos matemáticos em geral para melhor descrição, entendimento e tomada de decisões nos mais variados campos da ciência. Entre esses métodos, a utilização de métodos estatísticos tem crescido e se popularizado cada vez mais. Áreas como engenharia, zoologia, biologia, medicina, física, dentre outras, têm se beneficiado das ferramentas originadas na estatística para tratamento de dados empíricos. Nesse contexto, este texto traz uma introdução geral a alguns dos principais conceitos em análise estatística de dados, com foco em aplicações em economia e finanças. Para finanças, especificamente, o foco principal está na aplicação de técnicas estatísticas para mensuração e gerenciamento de risco. Nesse caso, há um interesse grande nas características mais detalhadas,[1] por exemplo, do processo aleatório por trás da geração dos dados observados nas perdas monetárias por fraudes bancárias, ou por inadimplência de devedores em empréstimos financeiros. Em muitas das aplicações para dados econômicos abordadas neste texto, o interesse reside na descrição via modelos matemáticos da interrelação entre variáveis socioeconômicas.

Os modelos matemáticos intrínsecos às análises estatísticas de dados correspondem a modelos em que algumas medidas das observações não são observadas com perfeita precisão, e são sujeitas a aleatoriedades, que podem ser causadas por uma série de fatores: aleatoriedades resultantes do processo de amostragem, de erros de mensuração, de desconhecimento pleno do conjunto de informação relevante, e assim por diante. Diante disso, um elemento crucial em toda a análise estatística de dados é o conceito de variáveis aleatórias. De maneira simples e intuitiva, variáveis aleatórias são grandezas das quais não conhecemos os valores ao certo. Por exemplo, o valor da taxa SELIC[2] a ser divulgada pelo Banco Central após a próxima reunião do COPOM (Comitê de Política Monetária) é uma variável aleatória. O valor dos números sorteados na próxima rodada da Megasena é uma variável aleatória.

De um modo geral, a formulação de um modelo matemático pode ser vista como a nossa humilde e humana tentativa de tentar interpretar e/ou entender algumas características localizadas da maneira de como Deus (ou a natureza) governa o funcionamento do mundo.[3] As três Leis de Newton, por exemplo, são uma maneira de representar matematicamente princípios localizados de funcionamento do universo. Nesse contexto, a análise estatística de dados pode ser vista como uma tentativa humilde de nós, seres humanos, lermos da melhor maneira possível os sinais divinos contidos nos dados empíricos. Imagine, por exemplo, que estamos interessados em estudar o tempo decorrido entre fortes secas na região Nordeste do

[1] Conforme veremos nos próximos capítulos, essas características mais detalhadas do processo de geração dos dados aleatórios é representada na função de distribuição acumulada.

[2] A taxa SELIC é a taxa básica de juros brasileira que é utilizada como referência para a política monetária.

[3] Apesar da formação cristã dos autores, o leitor agnóstico ou adepto de outras religiões poderá interpretar essas analogias de outras maneiras, em concordância com suas próprias crenças (ou descrenças).

Brasil. Uma forma de abordar diretamente esse problema é entender como, no universo divino, funciona o processo de geração de secas no nordeste brasileiro. Para isso, podemos inicialmente formular um modelo matemático, com um componente aleatório, para explicar o tempo decorrido entre secas fortes. Em seguida, precisamos coletar uma massa de dados correspondente aos anos em que foram observadas secas na região – esses são os sinais divinos sobre a maneira pela qual Ele governa o aparecimento de secas. A esse processo de geração dos dados observados pelo pesquisador é atribuída a denominação de **processo gerador de dados** ou **PGD**.[4] Por meio de técnicas estatísticas de análise, podemos (1) estimar, da melhor maneira possível, algumas características dos modelos estocásticos que descrevem o processo de geração das secas, e (2) levantar medidas da imprecisão dessas estimativas.

Com base nessa discussão, este texto pretende:

1) Discutir o processo inicial de exploração de uma base de dados com informações disponíveis, para termos uma ideia de que método utilizar posteriormente para a análise dessas informações. Em qualquer trabalho de análise de dados, esse passo corresponde ao primeiro, a partir do qual o analista irá adquirir uma fotografia das características agregadas principais da base de dados como um todo, calculando, por exemplo, medidas de dependência entre diversas variáveis, quando for o caso. Uma vez obtida essa fotografia inicial, o próximo passo consiste na utilização de modelos matemáticos para descrever, prever e avaliar o efeito de determinadas decisões sobre as variáveis de interesse.

2) Fazer uma introdução aos modelos matemáticos comumente utilizados para a descrição de fenômenos da natureza, nos quais existe um componente (ou componentes) de aleatoriedade. Para isso, uma discussão sobre variáveis aleatórias será introduzida, cobrindo algumas das principais variáveis aleatórias utilizadas na prática. Serão abordadas variáveis tanto discretas quanto contínuas. Variáveis aleatórias discretas são utilizadas para modelar, por exemplo, a frequência de perdas operacionais por fraudes internas ou externas em um determinado período de tempo, enquanto variáveis aleatórias contínuas são utilizadas para modelar a severidade (valor monetário incorrido) dos eventos de perda, ou para modelar a renda dos domicílios em uma determinada unidade da federação no Brasil. Tanto as variáveis aleatórias discretas quanto as contínuas possuem parâmetros livres, que podem ser ajustados para melhor adaptar essas variáveis aos dados históricos dos eventos estudados. Além disso, faremos uma discussão sobre quais características das variáveis aleatórias são mais importantes na prática, e por que essas características são importantes.

3) Apresentar alguns dos principais métodos utilizados para encontrar os modelos matemáticos, dentro de famílias pré-determinadas de variáveis aleatórias, que mais se aproximam do processo gerador de dados. Esses métodos são utilizados justamente para encontrarmos os parâmetros livres das variáveis aleatórias discretas ou contínuas.[5] O primeiro método corresponde ao método de momentos, enquanto o segundo método corresponde aos estimadores de máxima verossimilhança.

[4] Em inglês, *data generating process* ou *DGP*

[5] Conforme discutiremos mais adiante neste texto, em muitos casos, não necessariamente um determinado processo estocástico pode ser modelado por uma variável puramente discreta ou puramente contínua. Pode haver processos onde a variável aleatória apresenta descontinuidade, por exemplo, no valor zero. Esses processos podem ser tratados via modelos de mistura, conforme veremos no Capítulo 7 deste livro.

4) Discutir o processo comumente conhecido como **simulações de Monte Carlo**, de acordo com o qual procura-se replicar o processo gerador de dados, para estudar o que acontece quando utilizamos determinadas técnicas estatísticas. As simulações de Monte Carlo constituem-se em uma ferramenta poderosa para entender todos os procedimentos por trás das análises estatísticas de dados.

5) Identificar e entender, a partir de simulações de Monte Carlo, as imprecisões incorridas na análise de dados estatísticos usados, por exemplo, para a estimação de características de modelos matemáticos. Nesse caso, abordaremos também a imprecisão incorrida na estimação dos parâmetros livres de variáveis aleatórias discretas e contínuas. De fato, estimadores de parâmetros livres de variáveis aleatórias são obtidos a partir de cálculos com dados aleatórios. Portanto, esses estimadores também são variáveis aleatórias, e como tal também possuem suas próprias características. Em um primeiro momento, essas características serão estudadas via simulações de Monte Carlo. Em um segundo momento, discutiremos como essas características podem ser antecipadas analiticamente, mesmo sem se recorrer a processos de simulação. Essa é uma das bases do processo conhecido como **inferência estatística**. A inferência estatística, antes de mais nada, corresponde justamente a como o analista pode antecipar o que seria observado nas simulações de Monte Carlo caso ele decidisse por assim proceder. O processo de inferência estatística parte do que denominamos de **distribuição de um estimador estatístico**.

6) Apresentar uma discussão sobre a qualidade do ajuste dos modelos matemáticos teóricos aos dados empíricos usados para a estimação desses modelos. Com base na distribuição dos estimadores estatísticos, dois procedimentos comumente utilizados são a construção de **intervalos de confiança** para os parâmetros livres estimados e os **testes de hipótese** a respeito desses parâmetros. A abordagem utilizada na discussão conta tanto com a discussão analítica desses dois procedimentos, quanto com ilustrações via simulações de Monte Carlo. A partir de então, uma discussão mais detalhada será conduzida sobre os testes de hipóteses para verificar a qualidade do ajuste dos modelos matemáticos (variáveis aleatórias utilizadas) aos dados empíricos observados. Ou seja, queremos inferir o quão próximo nosso modelo matemático está dos sinais divinos coletados.

7) Introduzir técnicas de combinação de modelos simples para a obtenção de modelos mais complexos e flexíveis. No processo de seleção de **modelos paramétricos** que melhor se adequam aos dados observados, avaliando-se esses modelos via testes de ajuste, pode-se chegar à conclusão de que não necessariamente as distribuições padrões tradicionais utilizadas são suficientes para modelar os diversos conjuntos de dados encontrados na prática. Nesse caso, pode-se recorrer a combinações de distribuições simples, chegando-se a formas funcionais bem mais flexíveis. Uma primeira classe de modelos que vem ganhando cada vez mais espaço, dada a sua flexibilidade e a sua facilidade de interpretação, são os **modelos de mistura**.[6] Os modelos de mistura podem ser construídos a partir de combinações convexas entre distribuições padrões tradicionais.

8) Considerar o processo de utilização de modelos que de fato não correspondem exatamente ao processo gerador de dados, mas que nem por isso deixam de ser úteis. De fato, a partir da discussão

[6] Em inglês, *mixture models*.

sobre a flexibilidade dos modelos de mistura para o ajuste de diversas situações encontradas na prática, iniciaremos uma discussão sobre um princípio que nós particularmente achamos um dos mais importantes na análise estatística de dados: "todo modelo estatístico está errado, mas alguns modelos são úteis".[7] Conforme discutimos anteriormente, um dos objetivos da análise estatística de dados é tentar encontrar um modelo estatístico para aproximar o processo gerador de dados (que somente Deus conhece). É difícil de acreditar que conseguiremos reproduzir exatamente o processo gerador de dados do Divino Criador. Porém, podemos tentar chegar o mais próximo possível, de forma que o nosso modelo possa satisfazer aos nossos objetivos de tomada de decisão ou os nossos objetivos de pesquisa científica. Neste texto, fazemos uma discussão intuitiva sobre os princípios de aproximação de modelos e sobre os princípios de seleção de modelos estatísticos, mostrando que, mesmo quando o modelo utilizado por nós não corresponde exatamente ao processo gerador de dados, os resultados obtidos ao final da análise podem ser tão bons quanto se estivéssemos utilizando exatamente o modelo correto. Para todos esses fatos, faremos uma discussão teórica intuitiva, e apresentaremos ilustrações a partir de simulações de Monte Carlo.

9) Introduzir a teoria para a modelagem de processos estocásticos que dependem de um vetor de várias variáveis aleatórias. Nos tópicos de (1) a (8) anteriores, são discutidos princípios básicos da análise estatística de dados, com foco em ajuste de modelos paramétricos a bases de dados com apenas uma variável de interesse (ou uma base de dados com diversas variáveis de interesse, mas sendo que cada variável é analisada separadamente). Na maioria dos projetos de análise de dados, independentemente dos objetivos e da área da ciência, o interesse reside na análise conjunta entre diversas variáveis. Nesse contexto, o interesse do analista pode ser, por exemplo, descobrir como uma determinada variável responde a alterações/estímulos em diversas outras. Por exemplo, caso o governo aumente as tarifas de importação de determinado produto, qual o impacto dessa alteração sobre o volume total importado para esse produto? O interesse de pesquisa pode ser construir um sistema dinâmico conjunto, de acordo com o qual diversas variáveis econômicas (por exemplo, produto interno bruto, investimento agregado, consumo das famílias e gasto do governo) interagem entre si, gerando um processo de retroalimentação que faz com que elas caminhem juntas ao longo do tempo. Alternativamente, o interesse do pesquisador pode estar em descobrir, a partir de diversas variáveis de comportamento humano, por exemplo, alguns fatores não observados diretamente que expliquem uma grande parcela do que é observado nas variáveis de comportamento. Em análise de risco, para uma carteira de investimento composta por vários ativos, o interesse pode ser modelar a estrutura de dependência entre esses ativos para que, por meio de simulações de Monte Carlo, possamos aproximar a distribuição agregada da carteira como um todo. Todas essas, e diversas outras modalidades de análise conjunta de variáveis aleatórias, estão diretamente ou indiretamente ligadas ao conceito de **variável aleatória multivariada** ou **vetor de variáveis aleatórias**. Esse conceito básico e os princípios matemáticos básicos por trás da modelagem de variáveis aleatórias conjuntas serão abordados também neste texto.

10) Apresentar os chamados **modelos de regressão** no contexto de variáveis aleatórias multivariadas. Nesses modelos, o interesse reside em se conhecer como uma determinada variável aleatória (ou um conjunto

[7] A versão original dessa frase "*essentially, all models are wrong, but some are useful*" foi escrita por George Edward Pelham Box em seu livro Box e Draper (1987).

de variáveis aleatórias) responde a outras variáveis ou pode ser prevista a partir dos valores de outras variáveis. Seguindo a denominação comumente encontrada na literatura, as variáveis a serem explicadas e/ou previstas são conhecidas como **variáveis dependentes**, **variáveis resposta**, **variáveis explicadas**, ou **variáveis preditas**. As variáveis que preveem e/ou explicam as variáveis preditas são conhecidas como **variáveis preditoras**, **variáveis explicativas**, **variáveis independentes** ou **covariáveis**.

Neste livro, em particular, apresentaremos uma introdução aos **modelos de regressão linear**, de acordo com os quais o valor médio de uma variável dependente está linearmente relacionado aos valores de outras variáveis aleatórias (ou variáveis constantes, no caso de experimentos controlados, por exemplo). Então, introduziremos o método de estimação chamado de mínimos quadrados cujos parâmetros são estimados a partir da minimização do erro quadrático médio produzido pela diferença entre os valores da variável dependente da amostra original e os valores da variável dependente gerada pelo modelo. Em seguida, outros modelos de regressão, conhecidos como **modelos de resposta binária**, serão introduzidos para lidar com diversas outras situações encontradas na natureza. A diferença principal entre o modelo de regressão linear tradicional e os modelos de resposta binária é que os modelos de regressão linear são mais adequados para análise de dados onde a variável resposta supõe valores contínuos, entre menos infinito e mais infinito. Obviamente, na prática, os valores possíveis para a variável resposta estarão contidos dentro de limites superiores e inferiores factíveis. No entanto, para fins de modelagem (lembrando que todo modelo é uma aproximação da realidade), para determinados conjuntos de dados, a variável resposta pode ser suposta como apresentando valores contínuos entre menos infinito e mais infinito, sendo os modelos de regressão linear a ferramenta adequada. Nos modelos de resposta binária a variável dependente supõe apenas valores 0 e 1. Isso significa que o objetivo desses modelos é tentar explicar variáveis binárias. Por exemplo, estamos interessados em tentar entender: (1) Por que, em um grupo de firmas, algumas faliram e outras não? (2) Por que, em um grupo de tomadores de empréstimos, alguns são bons pagantes e outros não? (3) Em uma população, por que alguns estão empregados e outros não? (4) Em uma cidade, por que alguns moradores são proprietários dos imóveis que residem e outros não? Como veremos também, esses modelos podem ser usados para classificar uma população em dois grupos e um tópico ativo de pesquisa atual é a busca por bons **modelos de classificação**.

11) Apresentar os conceitos básicos associados aos modelos de regressão com flexibilização na forma funcional. Conforme discutido acima, os modelos de regressão linear tradicionais e os modelos de resposta binária partem da premissa de que a média da variável resposta é uma função de uma combinação linear entre as variáveis preditoras. Essa premissa pode ser relaxada, de maneira que possamos supor formas mais flexíveis para a função que liga as variáveis preditoras à média da variável resposta. Nesse sentido, diversos outros modelos existem na literatura para permitir essa **flexibilização**. De fato, quando temos disponível uma base de dados com um número pequeno de observações, a hipótese de linearidade, mesmo não sendo totalmente verdadeira, pode ser adequada dada a insuficiência de informação disponível. Para bases de dados maiores, a informação disponível pode permitir ao analista desvendar tanto a forma funcional da relação entre as variáveis explicativas e a variável preditora, bem como os parâmetros dessa relação. Essas ideias serão brevemente consideradas.

12) Discutir, de forma natural e mais intuitiva, os diversos conceitos relativos às variáveis aleatórias e à inferência estatística. O leitor mais familiarizado com outros textos em estatística notará uma leve alteração na sequência dos tópicos cobertos. Preferimos, por exemplo, não abordar os conceitos básicos de probabilidade explicitamente, como é feito em livros de estatística introdutória. Nossa experiência como professores é que os alunos de áreas aplicadas podem entender rapidamente os conceitos de variáveis aleatórias, bem como suas características expressas na forma de função de distribuição acumulada e função densidade de probabilidade, sem recorrer a conceitos mais básicos de probabilidade. Ao longo dos capítulos, quando houver necessidade de recorrermos a tópicos específicos de probabilidade, isso será feito de forma natural no contexto de cada tópico específico. Esse é o caso, por exemplo, do conceito de probabilidades condicionais e de probabilidades conjuntas, para entender os conceitos de distribuição conjunta entre diversas variáveis aleatórias.

Uma outra alteração na sequência dos capítulos é a inclusão de várias ilustrações via simulações de Monte Carlo. A nossa intenção em usar simulações de Monte Carlo incessantemente deve-se principalmente ao fato de que, via simulações de Monte Carlo, resultados como transformações de variáveis aleatórias, a **lei dos grandes números** e o **teorema central do limite** são muito mais facilmente explicáveis. Além disso, o conceito de distribuição amostral dos estimadores torna-se extremamente mais palpável quando simulações de Monte Carlo são empregadas para ilustração. Uma outra razão para a utilização de simulações nos diversos capítulos deve-se à crescente utilização de simulações de Monte Carlo para levantamento da distribuição de variáveis aleatórias de interesse na área de gerenciamento e mensuração de risco. Por exemplo, simulações são comumente empregadas para estimar a distribuição de perdas operacionais[8] agregadas, via convolução de frequência e severidade das perdas. Em risco de mercado, simulações históricas são comumente empregadas para determinação de indicadores de risco (mais precisamente, *value-at-risk*) da carteira. Em risco de crédito, podemos usar simulações de Monte Carlo para levantar a distribuição de perdas por inadimplência (*default*) em uma carteira de empréstimos, por exemplo. Finalmente, simulações têm se tornado muito importantes nas últimas décadas devido ao surgimento e popularização de métodos de reamostragem para inferência estatística. Entre os diversos métodos que se utilizam de simulações, podemos citar: *bootstrap*, *jacknife*, *Markov Chain Monte Carlo MCMC*, *amostrador de Gibbs* e outros.

13) Apresentar aplicações dos métodos estatísticos em diversas áreas. Mesmo tendo como foco principal do texto o estudo da mensuração e do gerenciamento do risco na área de finanças e o estudo de bases socioeconômicas em economia, outros tópicos importantes relacionados com economia e finanças são introduzidos no texto, dentre eles: o problema de **apreçamento de opções**, o problema de **gerenciamento de carteira**, incluindo uma discussão sobre **carteiras eficientes**, e o **CAPM** (modelo de apreçamento de ativos).

Este texto obviamente não tem a menor intenção de ser exaustivo nos tópicos cobertos. O interesse é meramente de passar alguns dos conceitos básicos para o aluno de graduação, aluno que se prepara para a prova da ANPEC, aluno de MBA (ou especialização) ou profissional aplicado, da forma mais intuitiva

[8] Risco operacional corresponde ao risco de eventos causados por falhas humanas, de sistemas, catástrofes, fraudes e roubos etc.

possível, tentando evitar se transformar em um texto demasiadamente simplório. O leitor é fortemente encorajado a recorrer a outras referências para melhor entender alguns dos tópicos de maior interesse. Um dos nossos objetivos é que este texto sirva de subsídio para trabalhos aplicados por parte dos leitores. Nesse caso, é importante que, durante a execução de alguma pesquisa e/ou trabalho de avaliação, o leitor faça uma extensa pesquisa literária, buscando artigos e/ou relatórios que tenham empregado técnicas parecidas a problemas similares. Comumente encontramos relatórios e/ou trabalhos científicos onde os autores poderiam ter coletado mais referências bibliográficas, o que beneficiaria em muito o desenvolvimento e a qualidade do produto escrito. Por esse motivo, fortemente aconselhamos que os leitores que forem utilizar os métodos aqui descritos façam sempre uma coleta adequada de trabalhos já desenvolvidos. A internet é obviamente um local ideal para começar.

É válido ainda mencionar que dentro da estrutura do texto, diferenciamos o que chamamos de "Exemplos" e "Aplicações". Enquanto os Exemplos pretendem apenas exemplificar a teoria dada ou ser usados para introduzir uma teoria, as Aplicações pretendem conectar a teoria dada com tópicos relevantes em economia e finanças. Adicionalmente, se uma Aplicação não for lida numa primeira leitura, não haverá prejuízo para o entendimento do leitor. Também diferenciamos "Práticas" de "Exercícios". Enquanto as Práticas são exercícios curtos introduzidos logo após a teoria (ao longo do texto) para fixar ideais. Os Exercícios estão localizados no fim do capítulo e podem ser bem fáceis ou até mesmo difíceis.

Um outro ponto importante que deve ser ressaltado aqui é que seguimos a convenção da língua inglesa de usar "."como separador de decimais e "," como separador de classes numéricas (classes das unidades simples, classes dos milhares e classes dos milhões). A justificativa para esse procedimento é que praticamente todos os programas estatísticos foram desenvolvidos seguindo essa convenção e, por isso, na área de estatística essa notação é mais comum.

I

Métodos básicos de estatística

2. Medidas de descrição para bases de dados

"It is a capital mistake to theorize before one has data. Insensibly one begins to twist facts to suit theories, instead of theories to suit facts."
Arthur Conan Doyle

Imagine que o analista possua uma base de dados disponível sobre o seu problema de pesquisa ou sobre o seu processo de tomada de decisão, da qual ele precise extrair informações. Neste capítulo, apresentamos algumas das principais medidas sobre as características agregadas da base de dados. Todos essas medidas estão disponíveis em pacotes estatísticos vendidos no mercado, e portanto não temos a intenção de entrar nos detalhes das diversas estratégias de cálculo para cada uma delas. O leitor interessado pode recorrer à literatura amplamente disponível sobre introdução à estatística básica, em nível de graduação. O nosso objetivo é municiar o leitor com alguns dos conceitos fundamentais para a utilização dos diversos pacotes estatísticos e econométricos, além de entender os passos básicos na análise exploratória que antecede à utilização de modelos matemáticos, descritos nos próximos capítulos.

Inicialmente, serão discutidas as medidas básicas que descrevem a localização "média" da massa de dados sendo analisada, e descrevem a dispersão desse conjunto de observações. Medidas de dispersão, por exemplo, são muito utilizadas na análise de risco em finanças. Em seguida faremos uma discussão sobre alguns recursos gráficos comumente empregados para avaliação da "forma" da disposição dos dados, discutindo também medidas que forneçam uma ideia geral da simetria ou da assimetria da disposição dos dados, bem como da ocorrência das chamadas caudas pesadas ou caudas grossas. Essas últimas correspondem à ocorrência com maior frequência de observações extremas na base de dados. Caudas pesadas são comumente encontradas em séries históricas de variações diárias nos preços de ativos financeiros transacionados em bolsa. Finalmente, trataremos de medidas e recursos visuais para analisar a dependência entre conjuntos de dados para variáveis distintas. Nesse caso, podemos avaliar, por exemplo, qual a relação encontrada entre a renda per capita de um determinado domicílio e o seu gasto total com saúde.

2.1 Medidas básicas

Considere uma base de dados contendo uma única variável,[1] com n observações, e valores $x_1, x_2, \ldots, x_n$. Vamos a partir de agora descrever diversas maneiras de se analisar esses dados, utilizando-se medidas

[1] Ela pode conter várias variáveis, sendo que essa primeira etapa da análise será feita para cada variável individualmente.

agregadas. A primeira medida é a soma total S_n dos valores, onde

$$S_n = \sum_{i=1}^{n} x_i.$$

Uma outra medida é a **média**, ou média aritmética $\bar{x}$ dos valores da massa de dados

$$\bar{x} = \frac{1}{n} S_n = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

A média é comumente conhecida na literatura como medida de localização, pois dá ideia do valor "médio" da massa de dados. Outras medidas de localização comumente encontradas são a **mediana** e a **moda**. A mediana é um valor que divide a massa de dados em duas metades iguais. Por exemplo, imagine o conjunto de informações composto pelos valores 12.0, 1.5, -0.2, 2.5, 3.1, 2.9, 1.2, 3.6. Ordenando-se esses $n = 8$ valores, obtemos a sequência -0.2, 1.2, 1.5, 2.5, 2.9, 3.1, 3.6, 12.0. O valor que divide a massa de dados em dois conjuntos com o mesmo número de observações é o valor $(2.5 + 2.9)/2 = 2.7$. Portanto, a mediana m para os dados do exemplo é igual a $m = 2.7$. Não nos ateremos a detalhes para o cálculo da mediana em diversas situações específicas, como por exemplo, quando há valores repetidos na base de dados. A intenção aqui é somente descrever o conceito principal da medida, a qual poderá ser calculada utilizando-se programas estatísticos facilmente disponíveis no mercado, ou utilizando-se planilhas eletrônicas. Pode-se mostrar que a média é mais sensível que a mediana a observações com valores extremos na base de dados. Por exemplo, se estivermos trabalhando com a variável idade das pessoas entrevistadas, e um digitador tiver digitado o valor igual a 200 para a idade de um dos entrevistados, esse valor espúrio afetará a média, mas não afetará a mediana (ou afetará muito pouco).

A moda corresponde ao valor mais frequente em um conjunto de valores. Por exemplo, considere uma base de dados consistindo de observações do número de filhos em domícilios visitados por uma pesquisa amostral. Consideremos então os valores 2, 3, 2, 4, 1, 2, 0, 0, 3, 4, 4, 2, 5, 8, 2. Nessa massa de dados, o valor 2 filhos aparece 5 vezes, sendo o valor mais frequente. Portanto, a moda M para essa massa de dados é $M = 2$. Tanto para a moda como para a mediana, diversas metodologias de cálculo estão descritas na literatura de estatística básica. Não é nossa intenção fazer uma discussão mais detalhada sobre esses cálculos, dado que o nosso objetivo central é passar os conceitos por trás da análise estatística de dados. O leitor interessado pode recorrer a referências como Hoffman (2006).

As três medidas média, moda e mediana fornecem uma ideia geral da localização central da massa de dados. Além da localização central, é interessante conhecer também qual a dispersão da massa de dados em torno de uma medida central. Medidas de dispersão são fundamentais em finanças, quando queremos avaliar o risco de operações financeiras. Quanto maior a dispersão, em geral, mais arriscada é a operação. Uma medida comumente utilizada para a avaliação da dispersão de um conjunto de dados é a **variância**.

A variância s^2 tem expressão

$$\text{Var}_A(X) = s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left[x_i - \bar{x}\right]^2,$$

onde $\bar{x}$ é a média dos dados. Uma das desvantagens da variância como medida de dispersão é que ela não está na mesma escala da variável original. Por exemplo, caso tenhamos uma massa de dados referentes a valores perdidos em fraudes externas por um determinado banco, os valores estarão registrados em R$ ou milhares de R$. Pela fórmula para a variância acima, nota-se que a medida s^2 estará em R$ ao quadrado, ou milhões de R$ ao quadrado. Por esse motivo, pode-se utilizar uma medida alternativa, conhecida como **desvio padrão**. O desvio padrão s é dado por

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left[x_i - \bar{x}\right]^2}.$$

Ou seja, o desvio padrão é simplesmente a raiz quadrada da variância, e estará na mesma escala que a variável original (R$ ou milhares de R$, por exemplo).

Um fato a princípio curioso na fórmula para a variância e para o desvio padrão corresponde ao valor $n-1$, ao invés de n, no denominador, dado que o somatório no numerador vai de $i = 1$ até $i = n$. Na verdade, a variância s^2 calculada de acordo com a expressão acima trata-se da **variância amostral**. Ela pode ser vista como uma estimativa para a **variância populacional** σ^2, cuja expressão é dada por

$$\text{Var}_P(X) = \sigma^2 = \frac{1}{n}\sum_{i=1}^{n}\left[x_i - \bar{x}\right]^2.$$

Similarmente, a raiz quadrada da variância amostral fornece o **desvio padrão amostral**, cuja expressão é dada acima. A raiz quadrada da variância populacional fornece o **desvio padrão populacional** σ. A expressão para σ é

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left[x_i - \bar{x}\right]^2}.$$

De fato, uma forma interessante para entender a divisão por $n-1$ e não por n é baseada na definição de **graus de liberdade** que é o número de componentes livres que é usado para calcular a estatística de interesse (no nosso caso, a variância). Em particular, note que para o cálculo da variância, nós usamos os desvios da média $(x_i - \bar{x})$, para $i = 1, \cdots, n$, cuja soma é zero. Logo, embora tenhamos usado n componentes para a estimativa da variância que poderia sugerir uma divisão por n, apenas $n-1$ deles são independentes e por isso fazemos a divisão por $n-1$. Abordaremos essa questão novamente no Capítulo 3. A partir de agora, sempre que for o caso, faremos a distinção entre medidas amostrais e medidas populacionais. Entretanto, é importante notar que, quando a amostra tem tamanho n muito grande, a divisão por n ou por $n-1$ não faz muita diferença.

Da mesma maneira que a mediana divide a massa de dados em duas metades com o mesmo número de observações em cada uma, os **quartis** dividem a massa de dados em quatro partes com o mesmo número de observações. Por exemplo, considere a massa de dados com os seguintes valores: 2.2, 3.1. 1.0, -0.2, 5.2, 3.2, 7.5, -2.4. Ordenando-se esses valores, obtemos -2.4, -0.2, 1.0, 2.2, 3.1, 3.2, 5.2, 7.5. O **primeiro quartil** Q_1 é igual a $(-0.2 + 1.0)/2 = 0.4$, o **segundo quartil** Q_2 é igual a $(2.2 + 3.1)/2 = 2.65$, e o **terceiro quartil** Q_3 é igual $(3.2 + 5.2)/2 = 4.2$. Com base no primeiro e terceiro quartis, uma medida de dispersão comumente utilizada é o **intervalo interquartil** DQ. O intervalo interquartil é dado pela diferença $DQ = Q_3 - Q_1$, e é menos sensível a obervações extremas do que a variância e o desvio padrão.

Similarmente aos quartis, os **decis** dividem a massa de dados em dez grupos, cada qual contendo 10% das observações. Para um conjunto de dados $x_1, x_2, \ldots, x_n$, o **primeiro decil** $d_{10\%}$ é um valor tal que $x_i \leq d_{10\%}$ para 10% das observações. O **sexto decil** $d_{60\%}$ é tal que $x_i \leq d_{60\%}$ para 60% das observações e $x_i > d_{60\%}$ para 40% das observações. Para pequenas amostras, diversas expressões existem para cálculo dos quartis e decis. Não nos ateremos a essas expressões neste texto. O nosso objetivo aqui é somente passar uma ideia geral da interpretação dos resultados fornecidos pela maioria dos programas estatísticos e econométricos. Finalmente, os **percentis** dividem a massa de dados em 100 intervalos contendo cada qual 1% das observações na amostra. Retornaremos a uma discussão sobre percentis quando formos tratar de mensuração de risco.

Aplicação 2.1 (Problema do bar El Farol, jogo da minoria e a variância como medida de ineficiência no uso dos recursos limitados)

Arthur (1994) introduziu o agora famoso **problema do Bar El Farol** com o objetivo de fazer uma crítica à hipótese da racionalidade perfeita e dedutiva.[2,3] O problema do Bar El Farol pode ser descrito da seguinte forma: suponha que você esteja em Santa Fé, goste de música irlandesa e queira ir ao bar El Farol numa sexta feira a noite. Suponha também que, como você, existam cem amantes de música irlandesa, mas apenas 60 lugares. O show é agradável somente quando menos que 60 pessoas aparecem. O que as pessoas devem fazer?

1) Existem nesse caso apenas duas possibilidades: ir ou não ir;

2) As estratégias boas são: ficar em casa quando mais de 60 pessoas vão ao bar ou ir ao bar quando menos que 60 pessoas pretendem ir ao bar.

[2]Embora seja muito difícil de justificar em muitas situações, a hipótese da racionalidade dedutiva tem sido já há muito tempo a hipótese padrão em teoria econômica. Basicamente, essa hipótese se baseia no princípio de que cada agente em um jogo sabe o que é melhor para si próprio, dado que todos os outros agentes são tão inteligentes quanto eles mesmos (por hipótese) na escolha das melhores ações. Existem pelo menos dois motivos para que essa hipótese não seja válida em situações pelo menos um pouco complicadas. O primeiro motivo é que a racionalidade humana é limitada e, por isso, os agentes reduzem a complexidade de suas decisões, visto que o cérebro humano tem habilidade computacional limitada e decisões perfeitamente racionais não são factíveis na prática. O segundo motivo é que sob situações mais complicadas, um agente não pode supor que os outros agentes estão "jogando" sob perfeita racionalidade e então ele é forçado a chutar o seu comportamento.

[3]Uma hipótese alternativa à hipótese da racionalidade dedutiva é a hipótese da racionalidade indutiva. A racionalidade indutiva é baseada no princípio de que agentes possuem um número limitado de estratégias e ao invés de decidirem os méritos da estratégia antes do "jogo", os agentes as avaliam depois de cada rodada de acordo com o desempenho de cada uma e ajustam suas decisões de acordo com essa medida de desempenho.

3) As estratégias ruins são: ir ao bar quando mais que 60 pessoas também decidiram ir ao bar e ficar em casa quando menos que 60 pessoas decidiram ir ao bar.

Uma vez que as pessoas não se comunicam entre si, elas podem decidir aparecer aleatoriamente no bar ou usar algum tipo de previsão. Alguns exemplos de preditores do número de pessoas que irão ao bar nessa semana são:

1) O mesmo da última semana;

2) A média das quatro últimas semanas.

3) A tendência das últimas oito semanas, limitada entre zero e cem.

Será que algum desses modelos é um bom modelo? De fato, não existe modelo de previsão correto. Pois se todos usarem o mesmo modelo de previsão, ou todos irão aparecer no bar ou todos ficarão em casa. Dessa forma, uma vez que não existe um modelo óbvio de previsão do comportamento dos outros agentes e vários modelos são defensáveis, é difícil encontrar uma solução dedutiva para esse problema.

Com o objetivo de estudar outras dimensões desse problema não consideradas no artigo original (ARTHUR, 1994), uma versão computacional simplificada do problema do Bar El Farol conhecida como **jogo da minoria** foi introduzida por Challet e Zhang (1997). Uma versão simplificada do problema foi necessária, pois o problema original sofre do mal da dimensionalidade. Por exemplo, suponha que indivíduos que estão decidindo se vão ao bar ou não tomam suas decisões baseados no número de pessoas que apareceram nas últimas M semanas, onde M é o tamanho da memória do agente. Se existem N agentes decidindo se vão ao bar ou não, então são possíveis $N+1$ valores em cada semana. Isso gera $(N+1)^M$ combinações possíveis de informação sobre o passado. Se as estratégias de previsão são baseadas nessa informação, então existem $(N+1)^{(N+1)^M}$ estratégias possíveis. A simplificação proposta por Challet e Zhang (1997) supõe que existe uma população com N (número ímpar para não haver empates) indivíduos interessada em ir ao bar com $(N-1)/2$ lugares e sugere que ao invés de prever o número de indivíduos que vai ao bar, pode-se prever apenas se o indivíduo deveria ter ido ao bar (1) ou não (-1). Dessa forma, é apenas necessário guardar se o indivíduo deveria ter ido ao bar ou não. Com essas simplificações o número de estratégias se reduz para 2^{2^N} e o problema se resume à escolha do lado da minoria.

Resumidamente, o jogo da minoria pode ser descrito da seguinte forma: em um dado instante de tempo, um agente que pertence a uma população de tamanho N escolhe entre duas ações opostas: $a = \pm 1$. A dificuldade é que cada agente não sabe o que os outros irão escolher. Uma vez que os recursos são limitados, o objetivo de cada agente é escolher o lado dividido pela minoria da população. O agente escolhe sua próxima ação baseado em uma estratégia, que é um mapeamento que define a ação a ser tomada em função da informação global, que é a sequencia dos últimos M resultados do jogo. Os livros de estratégias são aleatoriamente escolhidos para cada agente antes do início do jogo. Portanto, não existe uma solução ótima para o problema, isto é, os agentes não sabem qual é a melhor estratégia a ser escolhida no jogo. Cada agente tem um número fixo de estratégias s que não mudam no tempo. Para expressar o fato de que

os agentes têm diferentes crenças sobre o ambiente, as estratégias diferem de agente para agente. Em cada partida do jogo, os agentes usam suas estratégias mais pontuadas, que são aquelas que foram mais bem sucedidas na escolha do lado da minoria nas partidas anteriores do jogo.

O jogo da minoria, por ser um dos sistemas complexos mais simples, introduzidos com o objetivo de estudar a dinâmica e o comportamento coletivo de populações de agentes que competem por recursos limitados, tem sido útil, por exemplo, para entender mercados financeiros (CHALLET; MARSILI; ZHANG, 2000, 2001; JEFFERIES; HART; HUI, 2001), o conceito de inteligência (WAKELING; BAK, 2001; MELLO; CAJUEIRO, 2008), efeito manada (CAJUEIRO; CAMARGO, 2006; MELLO et al., 2010), questões ambientais (BOSCHETTI, 2007) e o mercado de leilões de carros usados (LUSTOSA, 2008; LUSTOSA; CAJUEIRO, 2010). Uma revisão da teoria e de vários desses resultados pode ser encontrada em Johnson, Jefferies e Hui (2003), Coolen (2005) e Challet, Marsili e Zhang (2005).

Uma propriedade importante do jogo da minoria é que os agentes se organizam em torno do ótimo, isto é, em média, a soma do número de agentes que vai ao bar é exatamente igual à capacidade do bar (CHALLET; MARSILI; OTTINO, 2004).

Embora os jogos da minoria tenham sido estudados em várias situações diferentes e considerados em várias aplicações, uma das propriedades mais interessantes, e que provavelmente foi responsável pela popularização dos jogos da minoria (SAVIT; MANUCA; RIOLO, 1999), é baseada no cálculo da variância s^2 do número de pessoas que decidem ir ao bar em um determinado instante – uma medida global de eficiência do sistema. Note que, como vimos acima, um propriedade do jogo da minoria é a organização dos agentes em torno do ótimo. Portanto, quanto menor a variância em torno dessa média, melhor estarão sendo utilizados os recursos do sistema. Dessa forma, se plotarmos a razão s^2/N versus $\alpha = 2^M/N$, pode-se concluir:

1) Para valores pequenos de $\alpha = 2^M/N$, o jogo é menos eficiente que se os agentes estivessem tomando decisões totalmente aleatórias. Nesse caso, a memória de um agente é tão pequena que ele não é capaz de cooperar com os outros agentes, pois não consegue entender o que está ocorrendo no jogo.

2) Para valores grandes de $\alpha = 2^M/N$, o desempenho dos agentes converge para a decisão aleatória. Nesse caso, o número de estratégias possíveis é tão grande que é difícil para um agente prever o que um outro agente vai jogar.

3) Existe um valor crítico $\alpha = \alpha_c$, onde os recursos do jogo são usados da melhor forma possível, isto é, a razão s^2/N é a mínima possível. A região de baixos valores de M é caracterizada pelo decréscimo de s^2/N, com o aumento de $\alpha = 2^M/N$, e a região de altos valores de M é caracterizada pelo aumento de s^2/N com o aumento de $\alpha = 2^M/N$. No ponto crítico, a cooperação entre agentes é clara e o sistema é muito mais eficiente do que o caso aleatório descrito acima.

4) Excluindo o caso trivial onde cada agente tem somente uma estratégia, mudar o número de estratégias disponíveis para cada agente não altera qualitativamente o comportamento do jogo da minoria.

A Figura 2.1 apresenta o padrão descrito acima para $N = 101$. $s = 2$ e o tempo total de simulação de cada jogo $T = 10000$. A linha constante nessa figura é a variância para o caso em que os agentes tomam sempre decisões aleatórias.

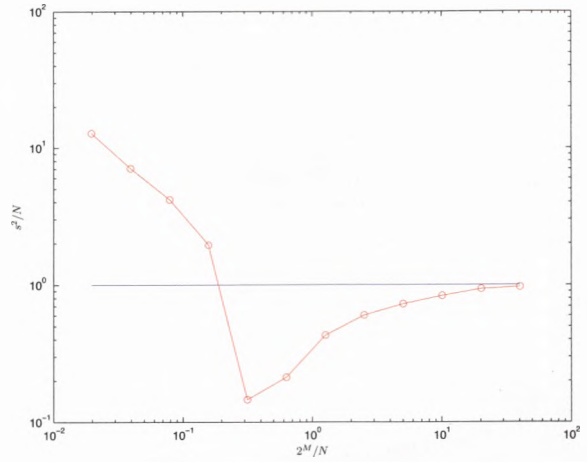


Figura 2.1: Variância versus memória no jogo da minoria.

2.2 Histogramas, assimetria e curtose

As medidas numéricas vistas acima fornecem uma descrição agregada da massa de dados sendo analisada. É interessante também termos descrições visuais dos dados a partir de recursos gráficos. Nesse sentido, uma ferramenta muito utilizada na prática são os **histogramas**. A partir de um conjunto de dados com valores $x_1, x_2, x_3, \ldots, x_n$, o **histograma de frequências absolutas**, com K colunas, pode ser construído com os passos a seguir.

(i) Seja $x_{(1)}$ o valor mínimo das observações e $x_{(n)}$ o valor máximo.

(ii) Dividimos o intervalo $[x_{(1)}, x_{(n)}]$ em K subintervalos de igual comprimento, divididos pelos pontos de corte $c_1, c_2, \ldots, c_{K-1}$. O subintervalo I_1 é igual a $[x_{(1)}, c_1]$, o subintervalo I_K é dado por $(c_{K-1}, x_{(n)}]$, e os demais subintervalos I_i são dados por $I_i = (c_{i-1}, c_i]$, para $i = 2, 3, \ldots, K-1$.

(iii) Para cada subintervalo I_k, $k = 1, \ldots, K$, contamos o número h_k de valores x_i contidos em I_k. Portanto, $\sum_{k=1}^{K} h_k = n$.

(iv) O histograma de frequências absolutas é dado pelo gráfico (geralmente gráfico de barras), onde no eixo vertical, apresentam-se os valores h_k, e no eixo horizontal, os pontos médios dos intervalos I_k.

O **histograma de frequências relativas** é similar ao histograma de frequências absolutas, onde, ao invés de usarmos h_k igual à contagem bruta do número de observações em cada subintervalo, h_k é dado pelo percentual de observações em cada subintervalo I_k. Portanto, para o histograma de frequências relativas,

temos $\sum_{k=1}^{K} h_k = 100\%$, ou $\sum_{k=1}^{K} h_k = 1.0$, caso não queiramos trabalhar com percentuais. O gráfico superior da Figura 2.2 apresenta o histograma de frequências absolutas para uma base de dados hipotética, contendo $n = 2000$ observações, e com $K = 30$ subintervalos. O gráfico inferior dessa figura apresenta o histograma de frequências relativas para a mesma base. Observe que a forma dos dois histogramas é exatamente a mesma, havendo mudança apenas nos valores no eixo vertical.

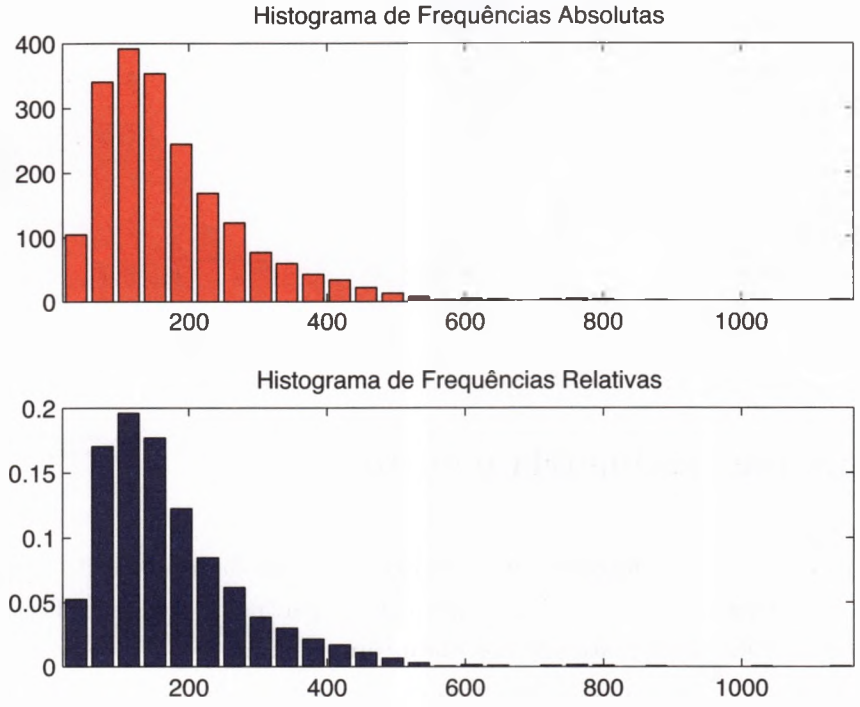


Figura 2.2: Histogramas de frequências absolutas e relativas para uma base de dados hipotética.

Os histogramas fornecem uma visualização da distribuição de dados como um todo, a partir da qual podemos inferir, por exemplo, se a distribuição é **simétrica** – ou seja, se os valores estão igualmente distribuídos em torno da média dos dados. Quando a distribuição das observações é simétrica, isso irá se refletir na simetria do histograma dos dados. Além disso, para distribuições simétricas, a média é igual à mediana. No gráfico (A) da Figura 2.3 a seguir, está ilustrado um histograma de frequências relativas para uma massa de dados com disposição simétrica. O gráfico (B) apresenta um histograma para uma massa de dados com **assimetria à direita**. O gráfico (C) apresenta um histograma com **assimetria à esquerda**.

A Figura 2.3 apresenta também três histogramas para ilustrar a ocorrência de assimetria e múltiplas modas em um mesmo conjunto de dados. No gráfico (D), notamos a presença de duas modas para a disposição dos dados, havendo uma simetria no histograma. Há uma moda no valor $x = 3.0$ e outra moda no valor $x = 6.0$. No gráfico (E), notamos novamente a presença de duas modas nos mesmos valores, mas agora há uma assimetria à direita. O gráfico (F) apresenta um histograma com duas modas e uma

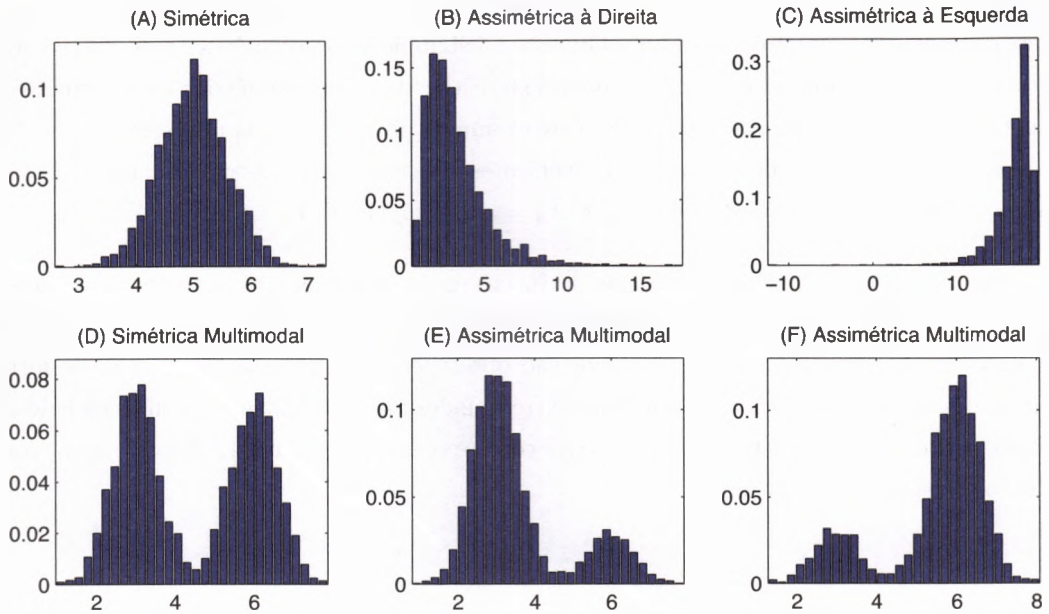


Figura 2.3: Histogramas para ilustrar assimetria e multimodalidade: (A) histograma de uma massa de dados com distribuição simétrica; (B) distribuição assimétrica à direita; (C) distribuição assimétrica à esquerda; (D) distribuição simétrica com duas modas; (E) distribuição assimétrica à direita, com duas modas; (F) distribuição assimétrica à esquerda com duas modas.

assimetria à esquerda. Conforme veremos na Seção 7.4, distribuições multimodais podem ser o resultado de um massa de dados geradas a partir de modelos de mistura, podendo representar duas subpopulações.

A assimetria pode ser capturada numericamente por uma medida conhecida como **coeficiente de assimetria**[4] CA. O coeficiente de assimetria populacional tem expressão

$$CA_P = \frac{\frac{1}{n}\sum_{i=1}^{n}\left[x_i - \bar{x}\right]^3}{\left[\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left[x_i - \bar{x}\right]^2}\right]^3},$$

enquanto o coeficiente de assimetria amostral tem expressão, para $n \geq 3$,

$$CA_A = \frac{\sqrt{(n-1)n}}{n-2} \times \frac{\frac{1}{n}\sum_{i=1}^{n}\left[x_i - \bar{x}\right]^3}{\left[\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left[x_i - \bar{x}\right]^2}\right]^3}.$$

A fração $\frac{\sqrt{(n-1)n}}{n-2}$ do lado direito da expressão acima corresponde ao que chamamos de termo para **correção de viés** do estimador do coeficiente de assimetria. Esse termo de correção não é definido para $n < 3$. O problema de viés dos estimadores será abordado em mais detalhes via simulações de Monte Carlo

[4]Em inglês, o termo utilizado para coeficiente de assimetria é *skewness*.

no Capítulo 6. Para uma massa de dados disponível para análise, quando a distribuição for simétrica, o coeficiente de assimetria será próximo a zero. Quando a distribuição for assimétrica à direita, o coeficiente de assimetria será maior do que zero, enquanto quando a distribuição for assimétrica à esquerda, o coeficiente de assimetria apresentará sinal negativo. Para os exemplos de histogramas apresentados na Figura 2.3, gerados a partir de bases de dados fictícias, os coeficientes de assimetria amostrais são: (A) $CA_A = -0.0698$, (B) $CA_A = 2.2367$, (C) $CA_A = -2.6963$, (D) $CA_A = -0.0016$, (E) $CA_A = 1.0537$, (F) $CA_A = -1.1140$.

Na análise de dados para séries históricas de variações de preços de ativos financeiros transacionados em bolsa, observa-se que as distribuições apresentam o que os analistas chamam de **caudas pesadas** ou **caudas grossas**. Isso porque valores extremos são observados com grande frequência, de forma que os histogramas dessas séries históricas apresentam extremidades mais evidentes em ambos os lados. Esse fato é conhecido na literatura estatística e econométrica como **excesso de curtose**.[5] A expressão para a medida de curtose populacional é dada por

$$K_P = \frac{\frac{1}{n}\sum_{i=1}^{n}\left[x_i - \bar{x}\right]^4}{\left[\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left[x_i - \bar{x}\right]^2}\right]^4},$$

enquanto a expressão para a curtose amostral para $n \geq 4$ é dada por

$$K_A = 3 + \left[(n+1)K_P - 3(n-1)\right] \times \frac{n-1}{(n-2)(n-3)}.$$

Note que a expressão para a curtose amostral apresenta também uma correção de viés, com base na expressão para a curtose populacional. Essa correção não é definida para $n < 4$.

Para ilustrar o excesso de curtose, a Figura 2.4 a seguir apresenta histogramas de quatro bases de dados diferentes, onde observamos diferentes configurações para as caudas da distribuição. No gráfico (A) da Figura 2.4, temos uma distribuição onde as caudas são leves. Conforme veremos no Capítulo 3, esse histograma corresponde ao que chamamos de distribuição de uma variável aleatória normal. O gráfico (B) ilustra uma situação onde não há caudas, já que os valores são limitados entre 5.0 e 15.0. Esse histograma corresponde à distribuição de uma variável aleatória que chamamos de uniforme, conforme veremos também no Capítulo 3. O gráfico (C) ilustra uma situação onde as caudas são pesadas, enquanto o gráfico (D) ilustra uma situação onde a caudas são muito pesadas. Note os valores para a curtose nos quatro casos. Quanto mais pesadas as caudas, maior a curtose.

[5] Em inglês, *kurtosis excess*.

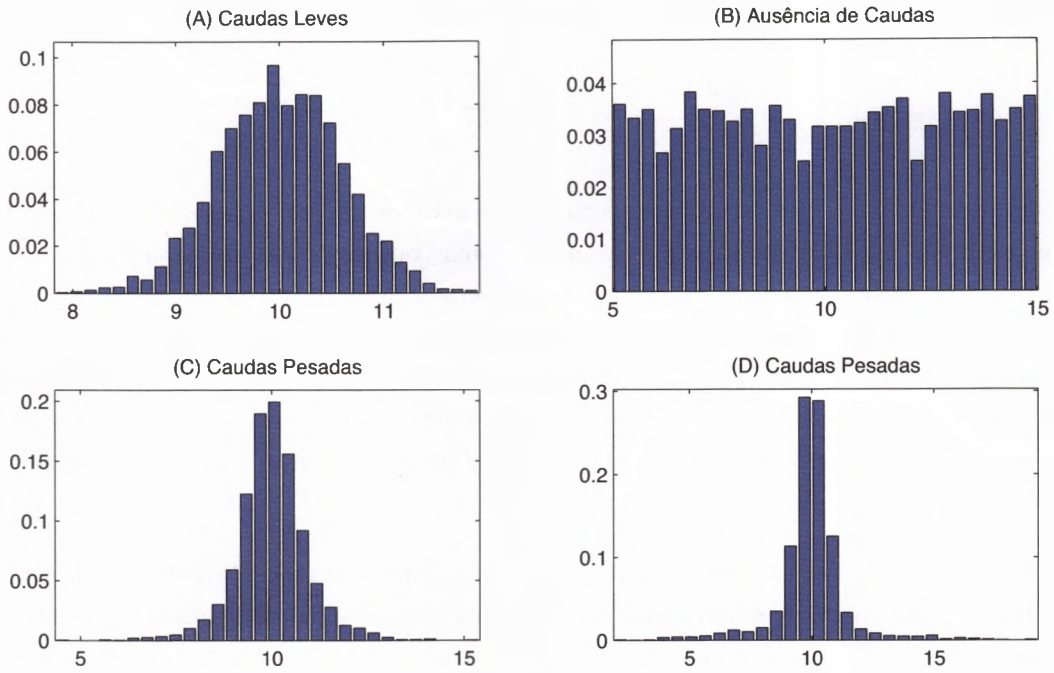


Figura 2.4: Histogramas para ilustrar excesso de curtose: (A) histograma de uma massa de dados com caudas leves, $K_A = 3.0087$; (B) distribuição sem caudas, $K_A = 1.7661$; (C) distribuição com caudas pesadas, $K_A = 5.9135$; (D) distribuição com caudas muito pesada, $K_A = 10.7750$.

2.3 Medidas de relação entre variáveis

Nas seções anteriores, discutimos algumas medidas básicas para caracterizar observações para uma determinada variável independentemente de outras variáveis disponíveis em um base de informações. No entanto, na maioria dos casos, o interesse reside justamente em conseguirmos identificar e modelar relações entre as observações coletadas para diferentes variáveis. Nesta seção, apresentaremos então uma série de medidas e ferramentas gráficas para avaliação dessas interrelações. Na discussão que se segue, suporemos que temos disponível uma massa de dados referente a n unidades observacionais, que podem ser domicílios, famílias, indivíduos entrevistados, plantas industriais, dias de pregão etc. Para cada uma dessas n unidades, suporemos, sem perda de generalidade, que possuímos informações referentes a duas variáveis, X e Y. Sejam então $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, os pares de valores referentes a cada uma das n unidades observacionais.

A primeira medida abordada nesta seção é conhecida como **covariância**, e tem expressão similar à expressão para a variância. A expressão para a covariância populacional entre as variáveis X e Y é dada por

$$\mathrm{Cov}_P(X, Y) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}),$$

enquanto a expressão para a covariância amostral é

$$\text{Cov}_A(X,Y) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}).$$

Note que, quando $X = Y$, a covariância corresponde à variância para a variável aleatória X (ou Y). Um dos problemas com a covariância é que, quando temos variáveis aleatórias com escalas diferentes, as covariâncias não são comparáveis. Por exemplo, se estivermos medindo a dependência entre duas variáveis aleatórias, sendo a primeira o volume total diário transacionado na Bovespa, e a segunda o índice ibovespa no fechamento do dia, somente pelo fato de alterarmos a unidade do volume total transacionado de R$ milhões para R$ bilhões, a covariância também se altera. Portanto, se alguém nos informar que a covariância entre a variável X e Y é igual a R$ 200 milhões, não será possível inferir se a relação de dependência entre as duas variáveis é alta ou baixa.

Para contornar esse problema de escala, podemos utilizar uma outra medida muito comum na literatura, conhecida como coeficiente de **correlação**, ou coeficiente de correlação de Pearson, entre duas variáveis. A correlação, representada pela letra grega ρ, é calculada a partir da razão entre a covariância e o produto dos desvios padrões de cada variável individualmente. Portanto, na sua versão populacional, o coeficiente de correlação tem expressão

$$\rho = \frac{\text{Cov}_P(X,Y)}{\sqrt{\text{Var}_P(X)\text{Var}_P(Y)}} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

onde $\bar{x}$ e $\bar{y}$ correspondem às médias dos valores para X e Y respectivamente. A versão amostral para o coeficiente de correlação é dada por

$$r = \frac{\text{Cov}_A(X,Y)}{\sqrt{\text{Var}_A(X)\text{Var}_A(Y)}} = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2}}.$$

O coeficiente de correlação fornece uma medida da relação linear entre duas variáveis. Quando a variável Y é uma função linear positiva da variável X (e vice-versa), o coeficiente de correlação é igual a 1.0, sendo esse o maior valor possível para o coeficiente de correlação, tanto amostral quanto populacional.[6] Quando a variável Y é uma função linear negativa da variável X, o coeficiente de correlação assume o valor -1.0, sendo esse o menor valor possível para o coeficiente de correlação, tanto amostral quanto populacional. Quando não há relação linear alguma entre as variáveis, o coeficiente de correlação é zero. Na prática, raramente encontram-se valores para o coeficiente de correlação iguais a um desses extremos.[7] Quanto mais próximo a 1.0 (-1.0) for o coeficiente, mais forte a relação linear positiva (negativa) entre as duas variáveis analisadas.

[6] Entende-se como relação linear entre as variáveis Y e X quando $Y = a + bX$, onde a e b são coeficientes constantes reais. Quando $b < 0$, a relação é negativa, e quando $b > 0$, a relação é positiva.

[7] Em análises de séries temporais, por exemplo, valores de correlação muito próximos a 1.0 podem ser indicativos de relações espúrias entre as duas variáveis sendo analisadas.

Note que, dado que o coeficiente de correlação necessariamente se localiza entre -1.0 e 1.0, ele não sofre o problema de escala que ocorre para a covariância.

Para ilustrar essa medida de dependência, considere a Figura 2.5 a seguir. Nessa figura, apresentamos gráficos em duas dimensões, plotando os valores para variável X no eixo horizontal e para a variável Y no eixo vertical. Esses gráficos são conhecidos como *gráficos de dispersão*, e fornecem uma visão bem clara da relação existente entre essas duas variáveis. O gráfico (A) apresenta uma situação onde não existe dependência entre os valores observados para a variável X e os valores observados para a variável Y. Nesse caso, o coeficiente de correlação amostral de Pearson resultou igual a -0.0107 (o que é muito próximo a zero). Os gráficos (B) e (C) apresentam exemplos de bases de dados onde existe uma correlação linear positiva entre as variáveis X e Y. Os gráficos (D) e (E) trazem ilustrações de conjuntos de dados onde existe uma correlação negativa. Finalmente, o gráfico (F) ilustra uma das desvantagens da utilização do coeficiente de correlação como medida de relação entre duas variáveis. De fato, o gráfico mostra uma clara relação de dependência quadrática, do tipo $Y = X^2$. No entanto, o coeficiente de correlação amostral calculado foi igual a $r = -0.0102$, o que é mais próximo a zero do que o coeficiente para o caso (A), onde não há relação alguma. Esse exemplo ilustra a adequação do coeficiente de correlação para capturar relações puramente lineares.

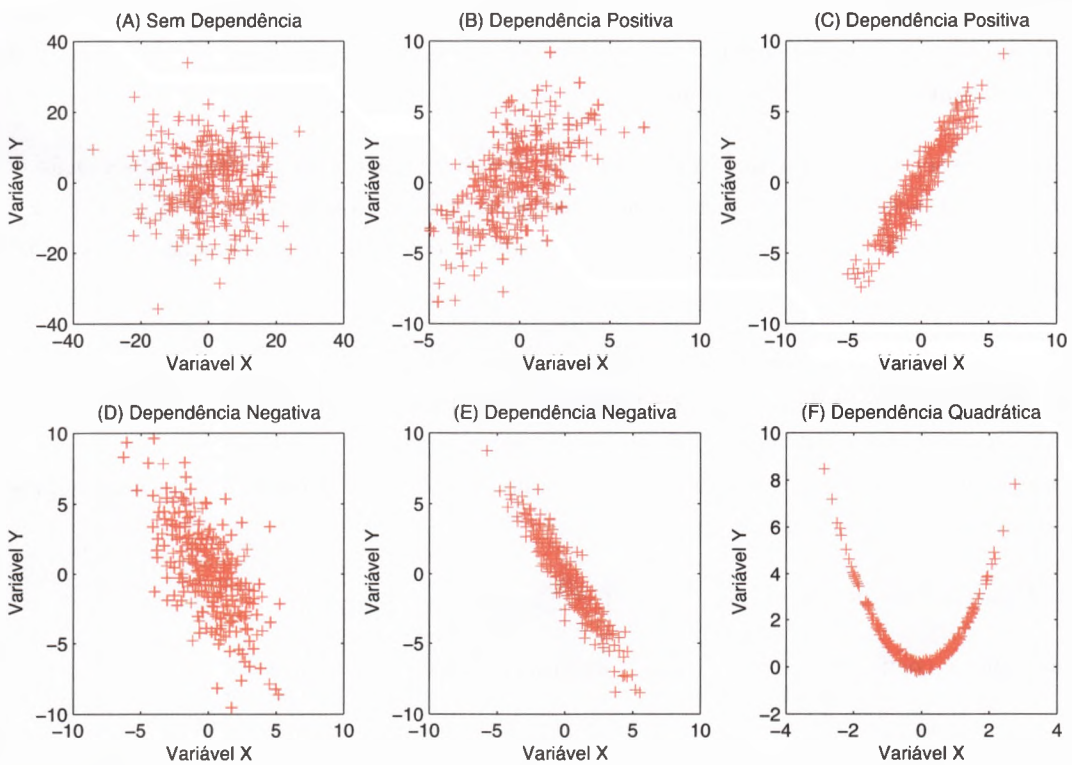


Figura 2.5: Histogramas para ilustrar o coeficiente de correlação entre duas variáveis: (A) Sem dependência entre as duas variáveis, $r = -0.0107$; (B) Baixa dependência positiva, $r = 0.6292$; (C) Alta dependência positiva, $r = 0.9504$; (D) Baixa dependência negativa, $r = -00.6250$; (E) Alta dependência negativa, $r = -0.9588$; (F) Dependência quadrática, $r = -0.0102$.

Outras medidas comumente encontradas para descrever a relação entre duas variáveis são a medida conhecida como **correlação de Spearman** (*Spearman rank correlation*) e o **coeficiente de Kendall Tau**. Essas duas medidas são denominadas medidas não paramétricas de associação, e supõe-se que o processo gerador de dados para ambas as variáveis é contínuo, de forma que a probabilidade de obtermos um mesmo valor na amostra, para uma mesma variável aleatória, é nula (probabilidade de empates[8] é nula). Essas medidas de associação são importantes, por exemplo, quando estamos procurando relacionar duas variáveis correspondentes a escalas ordinais, onde os valores reais não têm muito significado, mas o que importa é justamente a ordem desses valores. Conforme veremos mais adiante, essas duas medidas de associação são insensíveis às grandezas das observações, valendo apenas o ordenamento delas.

Para o coeficiente de correlação de Spearman, imagine que tenhamos uma amostra de n pares de observações, para duas variáveis aleatórias X e Y. Portanto, temos um conjunto de n pares (x_1, y_1), (x_2, y_2), (x_3, y_3), ... , (x_n, y_n). Um exemplo é que temos n observações de uma amostra de domicílios, onde a variável X corresponde à renda per capita do domicílio e a variável Y corresponde ao número de horas que a televisão passa ligada. Para cada domicílio, temos uma observação para a variável X ligada a uma observação para a variável Y. A partir dessa amostra de pares observacionais, seguimos os passos:

(1) Para cada observação x_i da variável aleatória X, determinamos o rank u_i dessa observação dentro do conjunto de observações de X. Ou seja, se x_i for o menor valor dentro dos valores x_1, x_2, ... , x_n, então $u_i = 1$; se x_i for o maior valor, então $u_i = n$; se x_i for o k-ésimo maior valor, então $u_i = k$. Portanto, as variáveis u_i assumem valores entre 1 e n.

(2) Para cada observação y_i da variável aleatória Y, determinamos o rank v_i dessa observação dentro do conjunto de observações de Y, da mesma maneira que no caso da variável aleatória X. Com isso, temos agora um conjunto de pares (u_i, v_i), $i = 1, ..., n$, onde u_i é o rank da observação x_i e v_i é o rank da observação y_i.

(3) Para cada par (u_i, v_i), determinamos a diferença entre os ranks, $d_i = u_i - v_i$. Note que, caso haja total concordância na sequência de observações para as duas variáveis X e Y, então os valores d_i serão todos nulos. Caso haja total discordância na sequência de valores, teremos $u_1 = 1$ e $v_1 = n$, $u_2 = 2$ e $v_2 = n - 1$, $u_3 = 3$ e $v_3 = n - 2$ e assim por diante. Portanto, uma ideia lógica é utilizar alguma função da sequência d_i para descrever a relação entre as duas variáveis.

(4) Calculamos a soma dos quadrados $\sum_{i=1}^{n} d_i^2 = \sum_{i=1}^{n} (u_i - v_i)^2$.

(5) Finalmente, o coeficiente de correlação de Spearman é dado pela expressão

$$R = 1 - \frac{6 \sum_{i=1}^{n} d_i^2}{n(n^2 - 1)}. \tag{2.1}$$

[8] Em inglês, *ties*.

Pode-se mostrar que, quando há total desacordo entre os ranks das observações das duas variáveis X e Y, ou seja, quando $u_1 = 1$ e $v_1 = n$, $u_2 = 2$ e $v_2 = n - 1$, $u_3 = 3$ e $v_3 = n - 2$ e assim por diante, temos que

$$\sum_{i=1}^{n} d_i^2 = \frac{n(n^2-1)}{3}.$$

Portanto, a medida relativa de "desassociação"pode ser escrita por

$$\frac{\sum_{i=1}^{n} d_i^2}{\frac{n(n^2-1)}{3}} = \frac{3\sum_{i=1}^{n} d_i^2}{n(n^2-1)}.$$

Essa medida acima será 0 quando os pares de observações estiverem em total acordo e 1 quando houver total discordância. Por outro lado, o nosso objetivo é criar uma medida de concordância, o que pode ser obtido pela expressão dada pelo coeficiente de Spearman na Eq. (2.1). Pode-se mostrar que, quando há total desacordo entre as duas amostras, $R = -1$. Quando há total acordo entre as duas amostras, $R = 1$. Na prática, o valor de R encontrado ficará entre -1 e 1.

Presença de empates. A expressão acima parte do pressuposto de que não há empates entre duas observações dentro da amostra para uma mesma variável. Portanto, supusemos que não encontramos dois valores $x_i = 2.3$ e $x_j = 2.3$, $i \neq j$, por exemplo. Quando isso acontece e a proporção de empates na amostra não é muito grande, podemos atribuir a essas observações um *rank* de empate, e proceder com a Eq. (2.1). Para essas duas observações hipotéticas, podemos associar o *rank* de empate $u_i = u_j = 12.5$, por exemplo, ao invés de atribuir $u_i = 12$ e $u_j = 13$, ou vice-versa. No entanto, quando a proporção de empates é significativa, podemos utilizar uma correção para a correlação de Spearman, com base na expressão

$$R = \frac{n(n^2-1) - \left[6\sum_{i=1}^{n} d_i^2\right] - 6(u' + v')}{\sqrt{n(n^2-1) - 12u'}\sqrt{n(n^2-1) - 12v'}}, \tag{2.2}$$

onde $u' = (\sum u^3 - \sum u)/12$ para u o número de observações, na amostra para X, empatadas para um determinado *rank*. Os somatórios $\sum u^3$ e $\sum u$ são efetuados para todos os *ranks* para os quais há empate. Similarmente, $v' = (\sum v^3 - \sum v)/12$, onde v é o número de observações, na amostra para a variável Y, empatadas para um determinado *rank*.

Para a **estatística de Kendall Tau**, considere o mesmo cenário de dados do coeficiente de correlação de Spearman. Portanto, temos um conjunto de n pares (x_1, y_1), (x_2, y_2), (x_3, y_3), ... , (x_n, y_n). Novamente, calculamos os *ranks* $u_1, \ldots, u_n$ para as observações da variável X, e os *ranks* $v_1, \ldots, v_n$ para as observações da variável Y. Temos então os pares de *ranks* (u_1, v_1), (u_2, v_2), (u_3, v_3), ... , (u_n, v_n). Os passos a seguir descrevem o processo de cálculo da estatística de Kendall Tau.

(1) Com base nos pares de *ranks* (u_1, v_1), (u_2, v_2), (u_3, v_3), ... , (u_n, v_n), ordene os pares, de acordo com os valores da variável u_i, em ordem crescente, obtendo a sequência de pares $(u_{(1)}, v_{(1)})$, $(u_{(2)}, v_{(2)})$, $(u_{(3)}, v_{(3)})$, ... , $(u_{(n)}, v_{(n)})$. Um exemplo é a sequência ordenada, abaixo, pelo *rank* da variável X, com $n = 6$:

Rank de X: 1, 2, 3, 4, 5, 6
Rank de Y: 2, 4, 1, 3, 6, 5

Caso houvesse uma associação perfeita entre as variáveis X e Y, a ordenação do *rank* da variável Y também seria a ordenação natural 1, 2, 3, 4, 5, 6.

(2) Consideremos então todos os pares de *ranks* da variável Y. Como são n observações, têm-se ao todo $n(n-1)/2$ pares. Se esses pares apareçam na ordem natural, de acordo com a sequência acima, atribuímos uma nota 1; casos os pares apareceram na ordem inversa, atribuímos uma nota -1. No exemplo acima, os pares estão apresentados na Tabela 2.1. Ao todo são 4 notas -1 e 11 notas +1. Caso houvesse total acordo entre as variáveis X e Y, o total de notas -1 seria zero, enquanto o total de notas +1 seria igual a $n(n-1)/2$. Por outro lado, se houvesse total desacordo entre as duas variáveis, o total de notas +1 seria nulo, e total de notas -1 seria $n(n-1)/2$. Portanto, a razão (notas positivas / total de notas) fornece um indicador para o grau de associação positiva entre as variáveis X e Y, enquanto a razão (notas negativas / total de notas) fornece um indicador para o grau de associação negativa.

Tabela 2.1: Pares para estatística de Kendall-Tau.

2, 4: nota = 1	4, 1: nota = -1	1, 6: nota = 1
2, 1: nota = -1	4, 3: nota = -1	1, 5: nota = 1
2, 3: nota = 1	4, 6: nota = 1	3, 6: nota = 1
2, 6: nota = 1	4, 5: nota = 1	3, 5: nota = 1
2, 5: nota = 1	1, 3: nota = 1	6, 5: nota = -1

(3) Seja então U o total de pares com nota positiva e V o total de pares com nota negativa. Uma medida de total acordo poderia ser escrita como

$$\frac{U}{\frac{n(n-1)}{2}} = \frac{2U}{n(n-1)},$$

enquanto uma medida de total desacordo poderia ser escrita como

$$\frac{V}{\frac{n(n-1)}{2}} = \frac{2V}{n(n-1)}.$$

No entanto, queremos uma medida que seja -1 quando houver total desacordo e +1 quando houver total acordo.

(4) Essa medida, conhecida como coeficiente ou estatística de Kendall Tau, é dada pela expressão

$$T = \left[1 - \frac{4V}{n(n-1)}\right] = \left[\frac{4U}{n(n-1)} - 1\right]. \tag{2.3}$$

Funções para cálculo do coeficiente de correlação, da estatística de Kendall Tau e do coeficiente correlação de Spearman estão amplamente disponíveis na maioria dos softwares estatísticos, muitos deles livremente disponibilizados na internet.[9] Além das três medidas de associação apresentadas neste capítulo, diversas outras estão disponíveis na literatura. A nossa intenção foi descrever algumas das mais conhecidas, além de despertar o leitor para a importância de estatísticas não paramétricas, como é o caso do coeficiente de Spearman e da estatística de Kendall Tau. As funções disponíveis nos softwares tratam inclusive de problemas de empates, que podem complicar a situação das estatísticas não paramétricas, conforme visto na Eq. (2.2).

2.4 Exercícios

Exercício 2.1 Para o coeficiente de correlação de Spearman, uma grandeza importante é a soma do quadrado do desacordo entre pares $\sum_{i=1}^{n}(u_i - v_i)^2$. Mostre que, quando há total desacordo entre os pares de observações para as variáveis X e Y, temos

$$\begin{aligned}\sum_{i=1}^{n}(u_i - v_i)^2 &= (n-1)^2 + [(n-1) - 2]^2 + \cdots + [2 - (n-1)]^2 + (1-n)^2 \\ &= 2[(n-1)^2 + (n-3)^2 + \ldots] = \frac{n(n^2-1)}{3}.\end{aligned}$$

Exercício 2.2 Uma medida comumente utilizada para a descrição da volatilidade de um ativo transacionado na bolsa de valores é a variância dessa variável. Seja p_t o preço de fechamento de um determinado papel (por exemplo, valor da ação ordinária da Petrobrás) no dia t. Esses valores podem ser baixados diretamente do *website* (por exemplo, vide *site* do Yahoo; http://br.finance.yahoo.com). O retorno de um dia para o outro é dado pela variação percentual $s_t = \frac{p_t - p_{t-1}}{p_{t-1}}$. Alternativamente, pode-se optar pelo log-preço retorno $r_t = \log(p_t/p_{t-1})$. Para uma série histórica de retornos diários r_t, $t = 1, \ldots, T$, podemos calcular a variância da série, utilizando-se a expressão

$$\sigma^2 = \frac{1}{T}\sum_{t=1}^{T}\left[r_t - \bar{r}\right]^2.$$

Utilizamos a expressão para a variância populacional (dividindo por T ao invés de $T-1$), mas na prática qualquer uma das duas forneceria resultado bem similar, dado que estamos supondo que o tamanho da

[9] É o caso do software estatístico R. Vide www.r-project.org.

amostra T é suficientemente grande (séries de retornos diários são bastante grandes; para cinco anos de dados históricos, por exemplo, temos em torno de 1.250 observações). Um dos problemas em se utilizar uma expressão dessa natureza é que a volatilidade, expressa pela grandeza σ^2, pode estar se alterando ao longo do tempo.[10] Para contornar tais situações, muitos analistas utilizam métodos conhecidos com médias móveis e médias móveis exponencialmente ponderadas - EWMA.[11]

A expressão para a variância σ_t^2 em médias móveis para o período t é dada por

$$\sigma_{t,K}^2 = \frac{1}{K} \sum_{i=t-K+1}^{t} \left[r_i - \bar{r}_{t,K}\right]^2,$$

onde K é uma janela de períodos para a média móvel. Podemos escolher, por exemplo, $K = 21$ (janela de um mês, considerando-se dias úteis apenas) ou $K = 252$ (janela de um ano). A média $\bar{r}_{t,K}$ também é uma média móvel, com expressão

$$\bar{r}_{t,K} = \frac{1}{K} \sum_{i=t-K+1}^{t} r_i.$$

Portanto, a variância em médias móveis captura a volatilidade dos últimos K dias. Variando-se o t ao longo do tempo, podemos obter uma série histórica para a variância, identificando períodos de maior e menor instabilidade no mercado de ações. Note que, quanto maior o valor da janela K, maior a suavização na série $\bar{r}_{t,K}$ e na série $\sigma_{t,K}^2$.

A média móvel exponencialmente ponderada tem uma expressão diferente da média móvel tradicional, mas se presta a um papel similar: identificar alterações ao longo do tempo na média e na variância da série histórica. Para a média, utilizando EWMA, temos a expressão

$$\bar{r}_{t,\lambda} = \lambda \times \bar{r}_{t-1,\lambda} + (1 - \lambda) \times r_t,$$

onde λ um número real pertencente ao intervalo $(0, 1]$. Com base em um valor inicial

$$\bar{r}_{1,\lambda} = r_1$$

no instante de início da análise $t = 1$,[12] podemos preceder recursivamente construindo toda a sequência de médias móveis $\bar{r}_{t,\lambda}$, $t = 2, \ldots, T$. Note que o valor de λ regula o grau de suavização na média móvel. Quanto mais próximo de zero o valor de λ, maior o peso da última observação r_t, implicando em uma

[10] De fato, em períodos de turbulência no mercado financeiro, a volatilidade tende a aumentar. Em períodos de calmaria, a volatilidade tende a se reduzir. Para esses tipos de situação, uma alternativa é utilizar modelos contendo heteroscedasticidade autorregressiva condicional (exemplo, modelos ARCH e GARCH, e suas extensões).

[11] Em inglês, *exponentially weighted moving average*.

[12] Alternativamente, o analista pode iniciar o processo recursivo de suavização exponencial utilizando uma média móvel, com K períodos iniciais.

menor suavização da série de média estimada $\bar{r}_{1,\lambda}$. Para a variância, a expressão via EWMA será

$$\sigma^2_{t,\lambda} = \lambda \times \sigma^2_{t-1,\lambda} + (1-\lambda) \times [r_t - \bar{r}_{t,\lambda}]^2.$$

Com base nos conceitos acima, responda às questões a seguir.

(1) Baixe uma série de observações para os últimos cinco anos de cinco papeis na bolsa de valores de São Paulo (vide http://br.finance.yahoo.com). Para cada uma dessas ações, determine uma série histórica de pelo menos quatro anos, via EWMA e via MA, para a variância do log-retorno das séries. Utilize janelas $K = 21$ e $K = 126$, e utilize parâmetros de suavização $\lambda = 0.9$ e $\lambda = 0.5$.

(2) O que acontece com os gráficos de volatilidade quando o K aumenta?

(3) O que acontece com os gráficos de volatilidade quando o parâmetro λ aumenta?

(4) Repita o Exercício 2.2 acima para o índice Ibovespa. Considerando-se as séries – o índice Ibovespa e as cinco séries do Exercício –, qual delas tem comportamento menos sensível a turbulências de mercado? É possível detectar comportamentos similares entre as volatilidades das cinco séries históricas? Para isso, determine as três medidas de associação entre as seis séries históricas (correlação, correlação de Spearman e coeficiente de Kendall Tau).

(5) Para os cinco papeis e para o Ibovespa, calcule a curtose e o coeficiente de assimetria (pode utilizar todos os cinco anos de dados diretamente). Qual das seis séries apresenta caudas mais pesadas? Qual das seis séries apresenta maior assimetria? Qual o sinal da assimetria? Explique os resultados encontrados.

(6) Caso utilizássemos uma distribuição normal para prever a probabilidade de retornos negativos muito baixos para qualquer uma das séries estudadas, você acha que o risco estaria sendo bem estimado, subestimado ou superestimado? Explique a sua resposta.

Exercício 2.3 Proponha versões EWMA para a curtose e para o coeficiente de assimetria. Pode considerar diretamente as versões populacionais desses dois coeficientes. Aplique as expressões sugeridas para essas duas medidas às seis séries históricas do exercício anterior.

Dica: Lembre-se de que, na versão original para a curtose, calculamos a média dos resíduos à quarta e dividimos essa média pela variância ao quadrado. Portanto, para termos um versão EWMA para a curtose, precisamos ter uma fórmula recursiva para o numerador (média dos resíduos à quarta), e dividir essa fórmula recursiva pela variância EWMA, elevada ao quadrado. A fórmula recursiva para o numerador $V_{t,\lambda}$ é dada por

$$V_{t,\lambda} = \lambda \times V_{t-1,\lambda} + (1-\lambda) \times [r_t - \bar{r}_{t,\lambda}]^4.$$

O mesmo princípio se aplica à versão EWMA para o coeficiente de assimetria. A fórmula recursiva para o numerador, equivalentemente ao caso da curtose, é dada por

$$U_{t,\lambda} = \lambda \times U_{t-1,\lambda} + (1 - \lambda) \times [r_t - \bar{r}_{t,\lambda}]^3.$$

Para obter a versão EWMA do coeficiente de assimetria, basta dividir o numerador $U_{t,\lambda}$ pelo desvio padrão (versão EWMA) ao cubo.

Exercício 2.4 Um dos problemas comumente encontrados em análise de séries temporais é a presença de componentes sazonais em séries mensais ou trimestrais, por exemplo. O nível de emprego aumenta nos últimos meses do ano, o consumo de cerveja aumenta no primeiro trimestre, o consumo de energia elétrica aumenta nos meses de verão, etc. Para melhor visualizar o comportamento tendencial de uma série, os analistas em geral procuram dessazonalizá-la antes de continuar com o estudo. Portanto, técnicas de dessazonalização são muito importantes em análises de séries temporais. Hoje em dia, estão disponíveis na literatura uma grande quantidade de algoritmos para dessazonalização, alguns dos quais são abordados em cursos de pós-graduação em Economia e Estatística (GHYSELS; OSBORN, 2001).

A utilização de médias móveis fornece uma maneira simples de dessazonalização. Por exemplo, para dados mensais, podemos utilizar médias móveis do tipo

$$\bar{y}_{t,K} = \frac{1}{K} \sum_{i=t-K+1}^{t} y_i,$$

com $K = 12$. Para séries trimestrais, basta utilizar $K = 4$. Não necessariamente a dessazonalização por médias móveis trará resultados satisfatórios. Dependendo do processo gerador de dados da série histórica, outros métodos podem ser mais adequados. Em todo caso, as médias móveis têm a grande vantagem de serem de simples aplicação e bastante intuitivas.

Com base na discussão acima, responda às questões abaixo:

(1) Baixe uma série de nível de emprego de *sites* especializados (vide www.ipeadata.gov.br, por exemplo), e aplique o método de médias móveis para dessazonalizá-la.

(2) Aplique o mesmo procedimento do item (1) para uma série de PIB brasileiro (nesse caso, a série é trimestral).

(3) Nos dois itens acima, plote a série histórica original e a série histórica dessazonalidada, ambas em uma mesma figura, para melhor observar o processo de dessasonalização.

Observação: Cuidado para não baixar séries históricas já dessazonalizadas. Atentar para a descrição das séries adquiridas.

Exercício 2.5 Mostre que o coeficiente de correlação de Spearman assume valores no intervalo fechado $[-1, 1]$.

Exercício 2.6 Mostre que a estatística de Kendall Tau assume valores no intervalo fechado $[-1, 1]$.

3. Variáveis aleatórias e modelos estocásticos

"If you apply reason and logic to this career of mine,
you're not going to get very far.
You simply won't.
The journey has been incredible from its beginning.
So much of life, it seems to me,
is determined by pure randomness."
Sidney Poitier

Conforme veremos ao longo deste livro, a abordagem utilizada aqui é fortemente baseada em modelos estatísticos e processos estocásticos. Por esse motivo, neste capítulo fazemos uma revisão dos principais tópicos, em probabilidade e estatística, necessários para o entendimento desses modelos descritos mais adiante ao longo do texto. Na próxima seção, introduzimos o conceito de variáveis aleatórias e descrevemos os principais componentes usados na sua caracterização. Nas Seções 3.2 e 3.3, apresentamos algumas das principais variáveis aleatórias comumente utilizadas na prática. Ao estudar outros livros de econometria e estatística tradicionais, o leitor notará que a maioria dessas variáveis aleatórias não são tão importantes na análise de modelos de regressão mais comuns, onde a variável resposta é suposta como tendo distribuição normal. No entanto, nas últimas décadas, com a popularização da análise de dados microeconométricos, a utilização de modelos com outros tipos de distribuições para modelar a variável resposta tem crescido bastante. Por exemplo, na análise de modelos de sobrevivência, onde o interesse reside em estudar como o tempo de sobrevivência de uma empresa depende das suas características, diversas outras variáveis aleatórias contínuas, com espaço amostral em $[0, +\infty)$, são muito relevantes. Alternativamente, em modelos para dados de contagem, onde a variável resposta corresponde ao número de crimes cometidos em um determinado perímetro urbano ou um determinado município, ou corresponde ao número de eventos de perda operacional em uma determinada agência, a distribuição de Poisson e a distribuição binomial negativa são bastante utilizadas.

Conforme veremos ao longo deste capítulo, todas as variáveis, discretas ou contínuas, possuem diversos indicadores para caracterizá-las. Os mais importantes desses indicadores são a função de densidade de probabilidade (ou função de frequência, para as variáveis discretas) e a função de distribuição acumulada. Os modelos apresentados, baseados nessas distribuições, são conhecidos como **modelos paramétricos**,[1] pois supomos que o processo gerador de dados é totalmente conhecido pelo analista, a menos de um conjunto

[1] Normalmente, a literatura divide os modelos estatísticos em três classes: modelos paramétricos, modelos semiparamétricos e modelos não paramétricos. Em modelos não paramétricos um número finito de parâmetros não é suficiente para conhecer o processo gerador de dados. Modelos semiparamétricos são uma situação intermediária. Embora o modelo tenha uma forma funcional como no caso dos modelos paramétricos e um número finito de parâmetros é suficiente para especifica-lo, o processo gerador de dados não é totalmente conhecido, pois usualmente esses modelos têm uma forma funcional maleável que é capaz de aproximar uma grande classe de sistemas. Um exemplo de modelos semiparamétricos é uma **rede neural**.

finito de parâmetros (por exemplo, a média e variância para o caso da distribuição normal). No Capítulo 5 descrevemos os principais métodos de estimação dos parâmetros das diversas distribuições utilizadas. No Capítulo 7 apresentamos alguns procedimentos comumente utilizados para seleção de modelos estatísticos, além de apresentar algumas técnicas estatísticas e procedimentos para avaliação da adequação dos modelos selecionados.

Além da conceituação de variáveis aleatórias e da introdução dos conceitos de função de densidade de probabilidade, função de frequência e função de distribuição acumulada, este capítulo traz a definição de momentos de uma variável aleatória. A média, por exemplo, é um momento da variável aleatória (conhecido como momento de primeira ordem). A variância corresponde ao momento (centrado) de segunda ordem. Algumas desigualdades importantes são também apresentadas, como é o caso da desigualdade de Cauchy-Schwarz, pela qual é possível demonstrar que o coeficiente de correlação tem sempre valor absoluto menor ou igual a 1. Finalmente, apresentamos um resultado extremamente importante que é o Teorema da Transformação de Variáveis Contínuas. Entre outras utilidades, por esse teorema podemos mostrar que o quadrado de uma distribuição normal padronizada possui distribuição qui-quadrada, com um grau de liberdade. Esse resultado será estendido quando estudarmos variáveis aleatórias multivariadas, mais adiante neste texto.

3.1 Variáveis aleatórias

Variáveis aleatórias são o principal componente dos modelos estatísticos e econométricos, sendo importantes, por exemplo, para modelos de risco operacional, de risco de mercado e de risco de crédito. A apresentação aqui tem o objetivo de ser bem intuitiva, sem a preocupação com tecnicalidades probabilísticas. Esperamos, contudo, que a nossa abordagem possa transmitir ao leitor as principais ideias envolvidas na modelagem de processos estocásticos, via distribuições paramétricas. O leitor interessado pode consultar, por exemplo, referências de teoria da medida e probabilidade ou estatística-matemática, tais como Bartle (1966), Fernandez (1973), Billingsley (1995), Durret (1996), Roussas (1997), Bickel e Doksum (2000), Casella e Berger (2001), Grimmett e Stirzaker (2001), Shao (2003), Rosenthal (2006) e Bierens (2004).

De maneira simples e intuitiva, variáveis aleatórias são grandezas das quais não conhecemos os valores ao certo. Por exemplo, o valor da taxa SELIC a ser divulgada pelo Banco Central após a próxima reunião do COPOM é uma variável aleatória. Outros exemplos são: (1) o crescimento do PIB Brasileiro ao final do ano corrente; (2) o valor da taxa de câmbio R$/US$; (3) o valor do déficit público daqui a dois anos; (4) o número de fraudes no autoatendimento de um determinado banco no próximo mês; (5) a proporção de empresas inadimplentes ao final do semestre; (6) o número de eleitores que irão votar a favor da reeleição; (7) a receita bruta total no próximo ano da empresa A, etc.

Conforme fica claro pelos exemplos acima (poderíamos passar todo o nosso curso pensando em uma infinidade de outros exemplos), essa entidade chamada variável aleatória é bem mais comum na nossa vida cotidiana do que aparenta. Por isso, espertamente, os matemáticos e estatísticos resolveram criar todo um arcabouço teórico e computacional para que essa infinidade de problemas cotidianos possam ser resolvidos de maneira sistemática. É justamente esse arcabouço que iremos cobrir nas próximas seções, com uma ênfase óbvia em modelos aplicados a economia e finanças.

A principal ideia no tratamento de variáveis aleatórias é definir alguma medida para o grau de conhecimento, ou de ignorância, que temos a respeito dos possíveis valores que a variável aleatória irá assumir. Por exemplo, é pouco provável que o crescimento do PIB Brasileiro ao final do ano corrente seja de 25%, ou de -15%. Certamente existe um conjunto ou intervalo de valores mais plausíveis para o valor do crescimento do PIB no ano corrente assumir: talvez algo em torno de 3%, variando, com algum grau de incerteza, entre 1% e 5%. Da mesma maneira, é pouco provável que a proporção de empresas inadimplentes ao final do semestre seja de 95%.

Antes de prosseguir com a definição de possíveis medidas para a incerteza de variáveis aleatórias, precisamos introduzir duas definições que estarão nitidamente ligadas aos vários modelos econométricos descritos nos próximos capítulos. As variáveis aleatórias podem ser divididas em dois grandes grupos:[2] **variáveis aleatórias discretas** e **variáveis aleatórias contínuas**. Variáveis aleatórias discretas são aquelas cujos valores possíveis pertencem a um conjunto enumerável (ou seja, discreto). Exemplos de variáveis discretas são o número de empresas inadimplentes ao final de seis meses, o número de fraudes no autoatendimento no próximo mês, o número de clientes na fila de uma agência em determinado momento do dia, etc. Obviamente, não podemos ter 3.6 empresas inadimplentes, ou 12.4 clientes na fila. Os conjuntos de valores possíveis nos três casos é o conjunto de números inteiros não negativos $\mathbb{N} = \{0, 1, 2, 3, \dots\}$. Um outro exemplo muito importante de variável aleatória discreta é a variável de Bernoulli, a qual pode assumir apenas valor 0 ou valor 1. Uma aplicação dessa variável são modelos de *default* ou não *default*, em risco de crédito. Podemos associar valor 1 a uma empresa inadimplente e valor 0 quando a empresa não estiver inadimplente. Abordaremos a variável aleatória de Bernoulli em mais detalhes na Seção 3.2.

Variáveis aleatórias contínuas são aquelas que podem assumir qualquer valor em um intervalo do conjunto de números reais. Vamos aproveitar a oportunidade e introduzir a notação. Representaremos variáveis aleatórias por letras maiúsculas do tipo X, Y, Z etc. O conjunto de valores possíveis de uma variável aleatória será representado pela letra $\mathbb{X}$. Esse conjunto de valores possíveis é conhecido como **espaço amostral**. Um exemplo de variável aleatória contínua, comumente utilizado em risco de crédito, é a proporção Z do valor não sacado em uma linha de crédito, que será sacada caso a empresa enfrente estresse financeiro. O espaço amostral nesse caso é o intervalo $[0, 1]$ de valores entre 0 e 1. A ideia de espaço amostral para essa variável aleatória pode ser melhor representada pela notação $Z \in \mathbb{X} = [0, 1]$, onde se lê que Z tem valor no conjunto $\mathbb{X}$, o qual corresponde ao intervalo $[0, 1]$.

[2] Variáveis discretas e contínuas não exaurem todas as possibilidades para variáveis aleatórias. Podemos ter, por exemplo, variáveis aleatórias mistas entre contínuas e discretas. Entretanto, em termos de exposição dos conceitos básicos de variáveis aleatórias, vamos supor a dicotomia básica entre variáveis aleatórias discretas e variáveis aleatórias contínuas.

Especificamente para risco operacional, uma variável aleatória contínua extremamente importante é o valor monetário total Y incorrido em um evento de perda específico. Nesse caso, é difícil imaginar um evento de perda negativa. Também é meio implausível acreditar que o valor total incorrido em um determinado evento de perda seja o valor do PIB brasileiro. Mesmo assim, o arcabouço de variáveis aleatórias é usado primariamente para construir representações (ou seja, aproximações) úteis, e não necessariamente totalmente exatas, da realidade. Nesse sentido, apesar de perdas operacionais do tamanho do PIB brasileiro serem pouco defensáveis, por preferência à simplicidade das derivações analíticas, suporemos que Y pode assumir qualquer valor não negativo no conjunto dos números reais $\Re^+ = [0, \infty)$. Portanto, podemos escrever $Y \in \mathbb{X} = [0, \infty)$.

3.1.1 Caracterização de variáveis aleatórias

Depois de introduzir o conceito de variáveis aleatórias discretas e contínuas, podemos prosseguir com a construção de medidas para o nosso conhecimento, ou ignorância, dos valores que uma variável aleatória pode assumir. A principal medida nesse caso é a **função de frequência**, para variáveis aleatórias discretas, e a **função de densidade de probabilidade**, para variáveis aleatórias contínuas. Devido à sua maior simplicidade de interpretação, iniciaremos a nossa discussão pelas variáveis aleatórias discretas. Geralmente, utilizamos o símbolo $f(x)$ para representar funções densidade e funções de frequência.

Funções de densidade de probabilidade e funções de frequência de variáveis aleatórias

A **função de frequência** de uma variável aleatória discreta X é uma função $f(x)$ que associa, a cada valor x possível de X assumir, o valor da probabilidade de ocorrência desse valor. Ou seja,

$$f(x) = \text{Prob}[X = x], \quad \text{para } x \in \mathbb{X}. \tag{3.1}$$

Observe que utilizamos uma letra maiúscula X para representar a variável aleatória e uma letra minúscula x para representar um valor específico que uma variável aleatória pode assumir. O conjunto $\mathbb{X}$ corresponde ao conjunto de todos os valores possíveis para X. Essa notação é comumente utilizada em textos de estatística, e será empregada de agora em diante neste documento. A função de frequência é extremamente importante porque ela caracteriza completamente a **distribuição** da variável aleatória X. Conhecendo a função de frequência, caracteriza-se todo o comportamento aleatório de X. A partir da função de frequência é possível por exemplo derivar o que chamamos **momentos** de uma variável aleatória ou de uma distribuição.[3]

No caso de variáveis aleatórias contínuas, a função de frequência é substituída pela **função de densidade de probabilidade (fdp)**. A função de densidade $f(\cdot)$ não indica diretamente a probabilidade

[3] A partir de agora, usaremos 'distribuição' ou 'variável aleatória' indiscriminadamente, como é geralmente usado na literatura estatística.

de ocorrência de um determinado valor, mas sim a probabilidade de ocorrência de valores dentro de um intervalo. Dessa forma, a probabilidade da variável aleatória Y assumir valores entre a e b, com $a < b$, é dada por

$$\text{Prob}[a < Y < b] = \int_{y=a}^{y=b} f(y)dy, \tag{3.2}$$

onde $\int_{y=a}^{y=b}$ indica integral entre a e b.

Um outro componente extremamente importante na caracterização de variáveis aleatórias é a **função de distribuição acumulada** $F(x)$. A função de distribuição acumulada tem expressão

$$F(x) = \text{Prob}[X \leq x],$$

e essa expressão vale tanto para variáveis aleatórias discretas, quanto para variáveis aleatórias contínuas. A função de distribuição acumulada pode ser obtida a partir da função de densidade, no caso contínuo, via expressão

$$F(x) = \int_{u=-\infty}^{x} f(u)du, \text{ para } x \in \Re.$$

Similarmente, a função de densidade pode ser obtida a partir da função de distribuição acumulada via primeira derivada

$$f(x) = \frac{\partial F}{\partial u}(u)\Big|_{u=x}.$$

Portanto, existe uma equivalência entre função de distribuição acumulada e função de densidade, de forma que qualquer uma das duas pode caracterizar completamente a variável aleatória contínua correspondente. Os mesmos argumentos valem para o caso discreto, onde a função de distribuição acumulada pode ser obtida a partir da função de frequência, utilizando-se a expressão

$$F(x) = \sum_{u=0}^{[x]} f(u), \text{ para } x \in \Re,$$

onde $[x]$ corresponde à parte inteira do número real x.

Exemplo 3.1 (Função de frequência) Considere uma variável aleatória discreta X, com espaço amostral $\mathbb{X} = \{0, 1, 2, 3\}$, e função de frequência $f(0) = f(1) = f(2) = 0.20$. Para escrevermos a função de distribuição acumulada para a variável aleatória X, precisamos notar que a função $F(x)$ deve ser definida

em todo intervalo $(-\infty, +\infty)$, $\mathrm{P}[X \leq x] = F(x)$ para todo $x \in (-\infty, +\infty)$. Portanto, temos

$$\begin{aligned} F(x) &= 0, \text{ para } x < 0, \\ F(x) &= 0.2, \text{ para } x \in [0, 1), \\ F(x) &= 0.4, \text{ para } x \in [1, 2), \\ F(x) &= 0.6, \text{ para } x \in [2, 3), \\ F(x) &= 1.0, \text{ para } x >= 3. \end{aligned}$$

Exemplo 3.2 (Função de densidade) Considere uma variável aleatória com função de densidade $f(x) = e^{-x}$, para $x > 0$ e $f(x) = 0$, para $x <= 0$. Podemos mostrar que a função de distribuição acumulada $F(x)$ é dada por

$$\begin{aligned} F(x) &= 1 - e^{-x}, \text{ para } x > 0, \\ &= 0, \text{ caso contrário.} \end{aligned} \tag{3.3}$$

Proposição 3.1 Toda função de frequência ou função de densidade de probabilidade $f(x)$ satisfaz

(i) $f(x) \geq 0, \forall x \in \Re$,

(ii) $\sum_{x=0}^{\infty} f(x) = 1$, para funções de frequência,

(iii) $\int_{-\infty}^{\infty} f(x)dx = 1$, para funções de densidade.

Prática 3.1 Seja Y uma variável aleatória com função de densidade

$$\begin{aligned} f(x) &= Ae^{-0.1x}, \text{ para } x > 0, \\ &= 0, \text{ caso contrário.} \end{aligned} \tag{3.4}$$

Determine o valor de A para que $f(y)$ seja uma função de densidade.

Nota 3.1 Em geral, para variáveis contínuas,

$\mathrm{P}[a < X \leq b] = \mathrm{P}[a \leq X \leq b] = \mathrm{P}[a \leq X < b] = \mathrm{P}[a < X < b] = \int_{x=a}^{x=b} f(x)dx$.

Observe que, para duas constantes a e b, com $a < b$,

$$\int_{-\infty}^{b} f(x)dx = \int_{-\infty}^{a} f(x)dx + \int_{a}^{b} f(x)dx,$$

e, portanto, $F(b) = F(a) + \mathrm{P}[a < X \leq b]$, resultando

$$\mathrm{P}[a < X \leq b] = F(b) - F(a).$$

A interpretação da função de densidade de probabilidade para variáveis aleatórias contínuas é intuitiva no sentido de que a probabilidade de a variável X assumir um valor entre os números a e b, com $a < b$, é dada pela área abaixo da função $f(x)$ e acima do eixo horizontal, delimitada pelos pontos a e b. Da mesma forma, seja A um conjunto formado pela união finita de vários intervalos disjuntos (interseção vazia), então a probabilidade de x assumir um valor no conjunto A é dada pela área abaixo da função $f(x)$, acima do eixo horizontal, delimitada pelos intervalos que compõem A.

Imagine agora que o conjunto B é formado apenas por pontos isolados na reta $\Re$. O valor da probabilidade $\mathrm{P}[X \in B]$ é dada pela área acima do conjunto B, delimitada por cima pela função $f(x)$. Ora, dado que o conjunto B é formado apenas por pontos isolados, a área acima do conjunto B será a área de algumas retas; ou seja, a área acima do conjunto B será nula, e portanto $\mathrm{P}[X \in B] = 0$. Em particular, se X é uma variável contínua, a probabilidade $\mathrm{P}[X = a] = 0$, para qualquer ponto $a \in \Re$. O conjunto B é denominado um **conjunto de medida nula**.[4] Isso explica as desigualdades do tipo $\mathrm{P}[a < X \leq b] = \mathrm{P}[a \leq X \leq b]$ na Nota 3.1. Um outro conceito importante é a diferença entre um conjunto de medida nula e um conjunto de eventos impossíveis. O fato de termos $\mathrm{P}[X = a] = 0$ não significa que X nunca possa assumir o valor a.

Proposição 3.2 Toda função de distribuição acumulada $F(x)$ satisfaz

(i) $F(x) \geq 0$,

(ii) $F(x) \to 1$, quando $x \to \infty$,

(iii) $F(x) \to 0$, quando $x \to -\infty$,

(iv) $F(x)$ é uma função monotônica crescente (não estritamente crescente),

(v) $F(x)$ é contínua pela direita.

Valor esperado e momentos de uma distribuição

Com base nas funções de densidade ou nas funções de frequência, podemos obter uma grandeza extremamente importante na caracterização de variáveis aleatórias. Essa grandeza é conhecida como **valor esperado** de uma variável aleatória X, e tem representação $\mathrm{E}[X]$. Intuitivamente, o valor esperado corresponde ao valor que observamos em média para a variável aleatória X quando fazemos sucessivas observações dessa variável. Para a variável aleatória "nota de um aluno na rede pública em matemática",

[4] Exemplos de conjuntos de medida nula são os **conjuntos contáveis** ou **enumeráveis**. Um conjunto B é dito enumerável ou contável quando ele é um conjunto finito, ou quando existe uma bijeção $h : \mathbb{N} \longrightarrow B$, com $B = \{b_n : n \in \mathbb{N}\}$, sendo $b_n = h(n)$, para todo $n \in \mathbb{N}$. Um exemplo de conjunto contável é o conjunto de números racionais $\mathbb{Q}$.

um valor esperado igual a 6.2 significa que, em média, as notas observadas para vários alunos observados em uma amostra é igual a 6.2. A expressão para o valor esperado $\mathrm{E}[X]$ é dada por

(i) $\mathrm{E[X]} = \sum_{x=0}^{\infty} xf(x)$, no caso discreto, supondo que o espaço amostral da variável aleatória X é $\mathbb{X} = \{0, 1, 2, 3, 4, \ldots\}$,

(ii) $\mathrm{E[X]} = \int_{x=-\infty}^{\infty} xf(x)dx$, no caso contínuo.

No caso mais geral, seja $g(\cdot)$ uma função qualquer.[5] Podemos escrever o valor esperado de $g(X)$ como

(i) $\mathrm{E[g(X)]} = \sum_{x=0}^{\infty} g(x)f(x)$, no caso discreto, supondo que o espaço amostral da variável aleatória X é $\mathbb{X} = \{0, 1, 2, 3, 4, \ldots\}$,

(ii) $\mathrm{E[g(X)]} = \int_{x=-\infty}^{\infty} g(x)f(x)dx$, no caso contínuo.

A partir da definição de valor esperado, podemos calcular diversas medidas para caracterizar uma variável aleatória. Essas medidas são conhecidas como **momentos** de uma variável aleatória. Os principais momentos são a **média** e a **variância**. A média é comumente representada pela letra μ ou pelo símbolo $\mathrm{E}[X]$ e indica o valor médio da variável aleatória. O segundo momento é a variância, comumente representada por σ^2 ou por $\mathrm{Var}[X]$, e indica a dispersão da variável aleatória em torno do valor médio. Supondo que o espaço amostral da variável aleatória discreta X é $\mathbb{X} = \{0, 1, 2, 3, 4, \ldots\}$, as expressões para a média e para a variância são

$$\mathrm{E}[X] = \mu = \sum_{x=0}^{\infty} xf(x) = 0 \times f(0) + 1 \times f(1) + 2 \times f(2) + \ldots$$

$$\mathrm{Var}[X] = \sigma^2 = \sum_{x=0}^{\infty} (x-\mu)^2 f(x) = (0-\mu)^2 \times f(0) + (1-\mu)^2 \times f(1) + (2-\mu)^2 \times f(2) + \ldots$$

A partir da variância, podemos encontrar o desvio padrão σ, que corresponde simplesmente à raiz quadrada de σ^2. Portanto, assim como a variância, o desvio padrão também dá indicação de dispersão da distribuição da variável aleatória em torno do ponto médio μ. Os outros dois momentos comumente encontrados são o coeficiente de assimetria *assy* e a curtose *kurt*. Como vimos no Capítulo 2, o primeiro indica o quão assimétrica é a distribuição da variável aleatória, enquanto o segundo indica o quão comum são valores extremos (ou seja, o quão frequente a variável aleatória assume valores altos). As expressões para a curtose e para o coeficiente de assimetria são dadas por

$$\begin{aligned} \text{assy} &= \frac{1}{\sigma^3}\mathrm{E}[(X-\mu)^3] = \frac{1}{\sigma^3}\sum_{x=0}^{\infty}(x-\mu)^3 f(x) \\ \text{kurt} &= \frac{1}{\sigma^4}\mathrm{E}[(X-\mu)^4] = \frac{1}{\sigma^4}\sum_{x=0}^{\infty}(x-\mu)^4 f(x) \end{aligned} \tag{3.5}$$

Valores positivos para o coeficiente de assimetria indicam que a distribuição é mais esticada para a direita, enquanto que valores negativos indicam que a distribuição é mais esticada para a esquerda. Um coeficiente de assimetria igual a zero indica que a distribuição é simétrica (vide discussão no Capítulo 2).

Para distribuições contínuas, as expressões para os momentos são análogas ao caso discreto, trocando-se os operadores de somatório por operadores de integração. As fórmulas para as quatro primeiras principais medidas de caracterização, no caso de variáveis aleatórias contínuas, são

$$\begin{aligned} \mathrm{E}[Y] &= \mu = \int_{y=-\infty}^{y=+\infty} y f(y) dy \\ \mathrm{Var}[Y] &= \mathrm{E}[(Y-\mu)^2] = \sigma^2 = \int_{y=-\infty}^{y=+\infty} (y-\mu)^2 f(y) dy \end{aligned} \tag{3.6}$$

para a média e a variância, e

$$\begin{aligned} \mathrm{assy} &= \frac{1}{\sigma^3}\mathrm{E}[(Y-\mu)^3] = \frac{1}{\sigma^3}\int_{y=-\infty}^{y=+\infty} (y-\mu)^3 f(y) dy \\ \mathrm{kurt} &= \frac{1}{\sigma^4}\mathrm{E}[(Y-\mu)^4] = \frac{1}{\sigma^4}\int_{y=-\infty}^{y=+\infty} (y-\mu)^4 f(y) dy, \end{aligned} \tag{3.7}$$

para o coeficiente de assimetria e para a curtose.

A variância é também denominada **segundo momento centrado**, uma vez que ela corresponde ao valor esperado do quadrado da variável aleatória menos a sua média. O segundo momento não centrado será simplesmente

$$\mathrm{E}[Y^2] = \int_{y=-\infty}^{y=+\infty} y^2 f(y) dy. \tag{3.8}$$

Analogamente, o K-ésimo momento não centrado é dado por

$$\mathrm{E}[Y^K] = \int_{y=-\infty}^{y=+\infty} y^K f(y) dy, \tag{3.9}$$

enquanto o K-ésimo momento centrado pode ser escrito como

$$\mathrm{E}[(Y-\mu)^K] = \int_{y=-\infty}^{y=+\infty} (y-\mu)^K f(y) dy, \tag{3.10}$$

onde μ é o valor esperado $\mathrm{E}[Y]$. Para uma variável aleatória discreta X, as expressões para os momentos centrados e não centrados de ordem K são

$$\begin{aligned} \mathrm{E}[(X-\mu)^K] &= \sum_{x=0}^{\infty}(x-\mu)^K f(x) \\ \mathrm{E}[X^K] &= \sum_{x=0}^{\infty} x^K f(x), \end{aligned} \tag{3.11}$$

onde $\mu = \sum_{x=0}^{\infty} x f(x)$ é o primeiro momento.

Proposição 3.3 Propriedades do valor esperado (ou **expectância**):

(i) $\mathrm{E}[a] = a$; ou seja, o valor esperado de uma constante $a \in \Re$ é igual à constante.

(ii) $\mathrm{E}[bX] = b\mathrm{E}[X]$, onde b é uma constante e X é uma variável aleatória (discreta ou contínua); ou seja, o valor esperado do produto de uma constante vezes uma variável aleatória é igual à constante vezes o valor esperado da variável aleatória.

(iii) $\mathrm{E}[X+Y] = \mathrm{E}[X] + \mathrm{E}[Y]$, onde X e Y são duas variáveis aleatórias (discretas ou contínuas); ou seja, o valor esperado da soma de duas ou mais variáveis aleatórias é a soma dos valores esperados.

(iv) $\mathrm{E}[aX + bY + c] = a\mathrm{E}[X] + b\mathrm{E}[Y] + c$.

(v) Se $X \geq 0$ sempre, então $\mathrm{E}[X] \geq 0$.

(vi) Se a densidade de X é simétrica em torno de um valor η, então $\mathrm{E}[X] = \eta$.

Prática 3.2 Mostre as propriedades (i) a (v) do valor esperado apresentadas na Proposição 3.3 usando propriedades do somatório ou da integral.

Nota 3.2 Existe uma relação entre os momentos centrados e os momentos não centrados. Para a variância, por exemplo, temos $\mathrm{E}[(Y-\mu)^2] = \mathrm{E}[Y^2 - 2Y\mu + \mu^2] = \mathrm{E}[Y^2] - \mathrm{E}[2Y\mu] + \mathrm{E}[\mu^2]$. Mas lembremos que o valor esperado de uma variável aleatória vezes uma constante é igual à constante vezes o valor esperado da variável aleatória; portanto, $\mathrm{E}[2Y\mu] = 2\mu\mathrm{E}[Y]$. Além disso, o valor esperado de uma constante é a própria constante; portanto, $\mathrm{E}[\mu^2] = \mu^2$. Finalmente, uma vez que $\mathrm{E}[Y] = \mu$, obtemos a relação

$$\mathrm{E}[(Y-\mu)^2] = \mathrm{E}[Y^2] - \mathrm{E}[Y]^2. \tag{3.12}$$

Ou seja, $\mathrm{Var}(Y) = \mathrm{E}[Y^2] - \mathrm{E}[Y]^2$ como mostra o item (i) da Proposição 3.4 abaixo.

Prática 3.3 Considere a variável aleatória X, com função de densidade

$$f(x) = B(1-x), \text{ para } 0 < x < 3,$$
$$= 0, \text{ caso contrário.}$$

Determine o valor da constante B para que a função acima seja uma função de densidade. Para o valor de B calculado, determine:

(a) $\mathrm{E}[X^2 + 5]$,

(b) $\mathrm{E}[3X^2 + 6X]$.

Discutiremos a seguir algumas das principais propriedades para a variância populacional. Diferentemente do valor esperado $\mathrm{E}[\cdot]$, a variância da soma de variáveis aleatórias não necessariamente é igual à soma das variâncias. No entanto, isso acontece quando as variáveis aleatórias são **não correlacionadas**; nesse caso, não há relação linear alguma entre elas. Um caso particular de variáveis não correlacionadas acontece quando elas são **independentes** entre si. Intuitivamente, duas variáveis aleatórias são independentes quando, conhecendo-se o valor da primeira variável, não adiciona informação alguma sobre o valor da segunda. Variáveis aleatórias independentes são necessariamente não correlacionadas. O contrário, porém, não é necessariamente verdade: variáveis não correlacionadas não necessariamente são independentes. Esses conceitos serão vistos em mais detalhes no Capítulo 4.

Via de regra, a chamada **covariância** entre duas variáveis aleatórias tem expressão

$$\mathrm{Cov}[X, Y] = \mathrm{E}\big[(X - \mu_X)(Y - \mu_Y)\big],$$

onde $\mu_X = \mathrm{E}[X]$ e $\mu_Y = \mathrm{E}[Y]$. A covariância mede a relação linear entre duas variáveis aleatórias. O problema da covariância, conforme fórmula acima, é que o valor resultante é de difícil interpretação, conforme vimos no Capítulo 2. Uma solução é calcular a **correlação** ρ entre as duas variáveis aleatórias, onde

$$\rho = \frac{\mathrm{Cov}[X, Y]}{\sqrt{\mathrm{Var}[X]\mathrm{Var}[Y]}}. \tag{3.13}$$

Pode-se mostrar que a correlação necessariamente possui valor absoluto menor ou igual a 1, que é uma consequência da desigualdade de Cauchy-Schwarz que será apresentada na Seção 3.4.5. Quando $\rho = -1$, existe uma dependência linear negativa exata entre as variáveis X e Y; ou seja, $Y = aX + b$, com $a < 0$. Quando $\rho = 1$, há uma dependência linear positiva exata, com $Y = aX + b$, com $a > 0$. Quando não há dependência linear alguma entre X e Y, tem-se $\rho = 0$. Em geral, tem-se ρ assumindo algum valor intermediário entre -1 e 1. Quanto mais próximo ρ estiver de 1, maior a relação linear positiva. Similarmente, quando mais próximo ρ estiver de -1, maior a relação linear negativa.

Proposição 3.4 (Propriedades da variância):

(i) $\text{Var}[X] = \text{E}[X^2] - \mu^2$.

(ii) $\text{Var}[a] = 0$, onde a é uma constante real.

(iii) $\text{Var}[aX + b] = a^2\ \text{Var}[X]$.

(iv) Sejam X e Y duas variáveis aleatórias reais (discretas ou contínuas). Então,

$$\text{Var}[aX + bY] = a^2\text{Var}[X] + b^2\text{Var}[Y] + 2ab\text{Cov}[X, Y].$$

Em particular, se $a = b = 1$, temos

$$\text{Var}[X + Y] = \text{Var}[X] + \text{Var}[Y] + 2\text{Cov}[X, Y].$$

(v) Seja X_i, $i = 1, \ldots, n$, uma sequência de variáveis aleatórias reais (discretas ou contínuas) e seja a_1, ..., a_n, uma sequência de números reais. Então,

$$\text{Var}\left[\sum_{i=1}^{n} a_i X_i\right] = \sum_{i=1}^{n} a_i^2 \text{Var}[X_i] + 2\sum_{i<j} a_i a_j \text{Cov}[X_i, X_j].$$

(vi) Em particular, se as variáveis aleatórias $X_1, \ldots, X_n$ forem não correlacionadas, ou seja $\text{Cov}[X_i, X_j] = \text{Corr}[X_i, X_j] = 0$ para todos os pares i, j, $i \neq j$, então

$$\text{Var}\left[\sum_{i=1}^{n} a_i X_i\right] = \sum_{i=1}^{n} a_i^2 \text{Var}[X_i].$$

Quando $a_i = 1$, para todo i, então

$$\text{Var}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \text{Var}[X_i].$$

É válido comentar que na Proposição 3.4 (vi), a variância da soma de variáveis não correlacionadas é igual à soma das variâncias. Essa característica é particularmente importante quando as observações coletadas em um amostra X_1, ... , X_n, são ditas independentes. Intuitivamente, as observações são independentes quando os valores obtidos nas primeiras observações coletadas não fornecem informação alguma sobre os valores das observações coletadas posteriormente. Um exemplo desse tipo de situação é quando nós temos o processo de retirada de números de dentro de uma urna, sendo que, após cada retirada, o número sorteado é reinserido na urna. Esse tipo de situação especificamente é conhecido como amostragem aleatória simples com reposição. Uma outra situação onde aproximadamente obtemos independência é quando o número n de observações coletadas da urna é muito pequeno em relação ao número N de números existentes dentro dela,

e o processo de amostragem é sem reposição (ou seja, os números não são reinseridos na urna após o sorteio). O conceito de independência é fundamental, conforme veremos no Capítulo 5, quando trataremos de alguns métodos comumente utilizados para estimação dos parâmetros livres das distribuições paramétricas.

Prática 3.4 Mostre as propriedades (i) a (iv) da variância apresentadas na Proposição 3.4 usando propriedades do somatório ou da integral.

Nota 3.3 Nos parágrafos acima, demos a definição de momentos baseados na definição de valor esperado. É importante ter em mente que nem sempre todos os momentos existem para uma determinada variável aleatória. Pode acontecer também que o momento não centrado de ordem 3 exista, e os demais momentos de ordem maior não existam. Nesse sentido, dizemos que o momento não centrado de ordem r de uma variável aleatória X existe quando

$$\mathrm{E}\big[|X|^r\big] < \infty.$$

Note o símbolo de valor absoluto na expressão do valor esperado acima. Portanto, dizemos que a média (ou momento de primeira ordem) de uma variável aleatória X existe quando

$$\mathrm{E}\big[|X|\big] < \infty.$$

De acordo com o Exercício 3.21, podemos mostrar que, se o momento não centrado de ordem r existe, então todos os momentos não centrados de ordem q também existem, para todo q tal que $1 \leq q \leq r$. Além disso, se o momento centrado ou não centrado, de ordem r, existe então todos os momentos centrados e não centrados de ordem q, com $1 \leq q \leq r$, também existem. Isso pode ser provado utilizando-se a desigualdade de Hölder que será apresentada na Seção 3.4.4.

Exemplo 3.3 (Variável aleatória discreta sem nenhum momento) Considere a variável aleatória discreta definida $X \in \{1, 2, 3, 4, \dots\}$, com função de frequência $f(x) = A/x^2$. Temos que encontrar o valor de A para que a função $f(x)$ seja de fato uma função de frequência. Nesse caso,

$$\frac{A}{1} + \frac{A}{2^2} + \frac{A}{3^2} + \cdots = A \sum_{x=1}^{\infty} \frac{1}{x^2} = 1.$$

Para resolver esse somatório, podemos recorrer à função zeta de Riemann $\zeta(s)$, definida como

$$\zeta(s) = \sum_{x=1}^{\infty} \frac{1}{x^s}.$$

Para $s \in \Re$, a função $\zeta(s)$ é definida (menor que ∞) para $s > 1$. Caso contrário, a série não converge e o resultado é ∞. Para resolver o problema da variável aleatória X acima, basta calcular o valor de $\zeta(2)$. Esse

valor é amplamente disponível na literatura e tem-se $\zeta(2) = \pi^2/6$, que é aproximadamente igual a 1.645. Portanto, para a função $f(x)$ ser uma função de frequência, basta fazer $A = 6/\pi^2$.

Vamos agora calcular o primeiro momento (ou valor esperado) da variável aleatória X. Pela definição de valor esperado, temos

$$\mathrm{E}[X] = \sum_{x=1}^{\infty} xf(x) = \sum_{x=1}^{\infty} Ax\frac{1}{x^2} = \sum_{x=1}^{\infty} A\frac{1}{x} = A\sum_{x=1}^{\infty} \frac{1}{x}.$$

O somatório acima corresponde a $A\zeta(1)$, que é igual a ∞, dado que $\zeta(s) = \infty$ para todo $s <= 1$, $s \in \Re$. Portanto, para a variável aleatória X descrita acima, não existe momento de ordem 1. O leitor poderá detectar facilmente que também não existem os momentos de ordem maior do que 1.

Exemplo 3.4 (Variável aleatória contínua sem nenhum momento) A variável aleatória de Cauchy tem função de densidade de probabilidade $f(x)$ dada por

$$f(x) = \frac{1}{\pi(1+x^2)}, \quad \text{para } x \in \Re.$$

Pode-se mostrar que $f(x)$ é uma função de densidade, no sentido de que sua integral é igual a 1. Além disso,

$$\int_{x=0}^{\infty} xf(x)dx = -\int_{x=-\infty}^{0} xf(x)dx,$$

já que a distribuição de Cauchy é simétrica em torno do ponto zero. Portanto,

$$\mathrm{E}[X] = \int_{x=0}^{\infty} xf(x)dx + \int_{x=-\infty}^{0} xf(x)dx = 0.$$

Se olharmos especificamente para o valor de $\mathrm{E}[x]$, concluiremos que o valor esperado (ou primeiro momento) da distribuição de Cauchy existe e é nulo. No entanto, pode-se mostrar que

$$\int_{x=0}^{\infty} xf(x)dx = \infty,$$

resultando em

$$\mathrm{E}[|X|] = \int_{x=0}^{\infty} xf(x)dx - \int_{x=-\infty}^{0} xf(x)dx = 2\infty,$$

abusando da notação da última igualdade acima. Portanto, $\mathrm{E}[|X|] = \infty$ para uma variável aleatória de Cauchy e dizemos que o primeiro momento não existe. Pode-se mostrar que todos os momentos de ordem maior (centrados ou não) também não existem.

3.1.2 Momentos populacionais versus momentos amostrais

Os quatro momentos descritos anteriormente (média, variância, coeficiente de assimetria e curtose) são conhecidos como **momentos populacionais**, pois eles são derivados diretamente das fórmulas para a distribuição das variáveis aleatórias. Em muito casos, é importante calcular também equivalentes amostrais para os momentos populacionais. Esses equivalentes amostrais são conhecidos como **momentos amostrais**. Suponha que temos disponível uma série de n valores, ou seja, uma amostra $X_1, X_2, X_3, \ldots, X_n$, com n suficientemente grande. Os momentos amostrais são calculados diretamente a partir desses valores amostrais, e têm expressão

$$\begin{aligned} \bar{X} = \hat{\mu} &= \frac{1}{n}\sum_{i=0}^{n} X_i \\ s^2 &= \frac{1}{n}\sum_{i=0}^{n}(X_i - \bar{X})^2, \end{aligned} \tag{3.14}$$

para a média $\hat{\mu}$ (ou $\bar{X}$) e a variância s^2 amostrais, e

$$\begin{aligned} \widehat{\text{assy}} &= \frac{1}{ns^3}\sum_{i=0}^{n}(X_i - \bar{X})^3 \\ \widehat{\text{kurt}} &= \frac{1}{ns^4}\sum_{i=0}^{n}(X_i - \bar{X})^4, \end{aligned} \tag{3.15}$$

para o coeficiente de assimetria e para a curtose amostrais. O desvio padrão amostral s é igual à raiz quadrada da variância amostral s^2.

Em muitos livros de estatística, quando o tamanho da amostra n não é grande o suficiente, as Eqs. (3.14) e (3.15) apresentam algum fator de correção para o viés de estimação dos momentos (STEVENSON, 1997), conforme discussão no Capítulo 2. Por exemplo, a fórmula para a variância amostral não viesada é dada por

$$s^2 = \frac{1}{n-1}\sum_{i=0}^{n}(X_i - \bar{X})^2, \tag{3.16}$$

com o denominador $n-1$ ao invés de n. Porém, para problemas típicos em microeconometria, com base de dados governamentais (PNAD, CENSO, POF etc.), por exemplo, o número de observações nas amostras é grande o suficiente, de forma que utilizar n ou $n-1$ no denominador não faz muita diferença. Dessa forma, as expressões apresentadas nas Eqs. (3.14) e (3.15) como estão já são adequados para os nossos objetivos.

Pode-se mostrar, tanto analiticamente quanto via simulações de Monte Carlo (conforme discutiremos na Seção 5.4), que os momentos amostrais aproximam-se dos momentos populacionais, e essa aproximação torna-se mais precisa à medida que o número de observações n na amostra aumenta. Esse fato é importante

para a construção de um dos métodos mais utilizados na estimação de parâmetros livres de variáveis aleatórias comumente conhecidas. Justamente por utilizar a equivalência entre momentos amostrais e momentos populacionais, esse método de estimação é conhecido como **método de momentos** e será apresentado no Capítulo 5.

3.2 Principais variáveis aleatórias discretas

Nesta seção, faremos uma revisão de algumas das principais variáveis discretas, comumente utilizadas em problemas aplicados. Juntamente com as variáveis aleatórias, apresentaremos as suas principais propriedades em termos de momentos. Todas as variáveis aleatórias discretas discutidas nesta seção serão representadas pela letra maiúscula X.

3.2.1 Variável aleatória de Bernoulli

A variável aleatória de Bernoulli é extremamente simples, apresentando somente dois valores possíveis: 0 ou 1. Portanto, o espaço amostral $\mathbb{X} = \{0, 1\}$. Essa variável possui apenas um parâmetro livre p, que indica a probabilidade de a variável aleatória assumir valor 1. Dessa forma,

$$\begin{aligned} &\text{Prob}\{X = 1\} = p, \\ &\text{Prob}\{X = 0\} = 1 - p. \end{aligned} \tag{3.17}$$

A variável aleatória de Bernoulli é normalmente utilizada para modelar processos de resposta binária, como por exemplo *default* e não *default*, vota ou não pela reeleição etc. A média μ e a variância σ^2 têm expressões

$$\begin{aligned} &\mu = 1 \times p + 0 \times (1 - p) = p \text{ e} \\ &\sigma^2 = p \times (1 - p). \end{aligned} \tag{3.18}$$

Prática 3.5 Plote o gráfico da função de distribuição acumulada $F(x)$ para a variável de Bernoulli.

Prática 3.6 Explicite o cálculo da variância da variável aleatória de Bernoulli apresentada na Eq. (3.18).

3.2.2 Variável aleatória binomial

A variável aleatória binomial é comumente utilizada para modelar a contagem do número de 1's em N eventos independentes de Bernoulli. Por exemplo, ao abordar $N = 112$ pessoas aleatoriamente na rua, a variável aleatória X pode ser utilizada para modelar quantos deles vão votar para reeleição. Essa variável aleatória possui 2 parâmetros: o número N de eventos de Bernoulli, e a probabilidade p de tirar 1 em cada

evento de Bernoulli. No nosso exemplo de reeleição, p seria a probabilidade de cada um dos 112 indivíduos votar a favor da reeleição.

O espaço amostral da variável aleatória binomial é $\mathbb{X} = \{0, 1, 2, \ldots, N\}$ e a função de frequência $f(x)$ é dada por

$$f(x) = \text{Prob}\{X = k\} = \binom{N}{k} p^k (1-p)^{N-k}, \text{ para } x = 0, 1, 2, \ldots, N, \tag{3.19}$$

onde

$$\begin{aligned} &\binom{N}{k} = \frac{N!}{k!(N-k)!}, \quad k! = k \times (k-1) \times (k-2) \times \cdots \times 3 \times 2 \times 1, \\ &(N-k)! = (N-k) \times (N-k-1) \times \cdots \times 2 \times 1 \text{ e } N! = N \times (N-1) \times \cdots \times 2 \times 1. \end{aligned} \tag{3.20}$$

A variável aleatória binomial tem média μ e variância σ^2 dados por

$$\begin{aligned} \mu &= N \times p \text{ e} \\ \sigma^2 &= N \times p \times (1-p). \end{aligned} \tag{3.21}$$

A Figura 3.1 apresenta a função de frequência para diferentes valores de N e p. Observe que quando N aumenta ou quando p aproxima-se de 0.5, a função de frequência possui gráfico com formato assemelhando-se a um sino. Na Seção 3.3.1, apresentaremos a variável aleatória normal, que possui exatamente o formato de sino ao qual a distribuição binomial se aproxima quando o tamanho de tentativas N aumenta.

Aplicação 3.1 (Contrato de opções e apreçamento usando o modelo binomial)

Um contrato de opção é um derivativo[6] que dá o direito (mas não a obrigação) a uma das partes interessadas (que pagou por esse direito a outra parte interessada) de comprar (ou vender) um ativo por um preço especificado numa data futura (preço de exercício) até o vencimento (data a partir da qual a opção não pode mais ser exercida).

Contratos de opções podem ser caracterizados pelo tipo de operação e, nesse caso, os tipos mais simples são:

1) Opção de compra[7] é uma opção para comprar um ativo especificado (ativo objeto) a um preço fixo.

2) Opção de venda[8] é uma opção para vender um ativo especificado (ativo objeto) a um preço fixo.

[6] Um contrato derivativo é um contrato cujo preço deriva de um ativo subjacente. Os exemplos mais comuns de contratos derivativos são futuros, opções e swaps. Vide, por exemplo, Hull (1997).

[7] Em inglês, *call*.

[8] Em inglês, *put*.

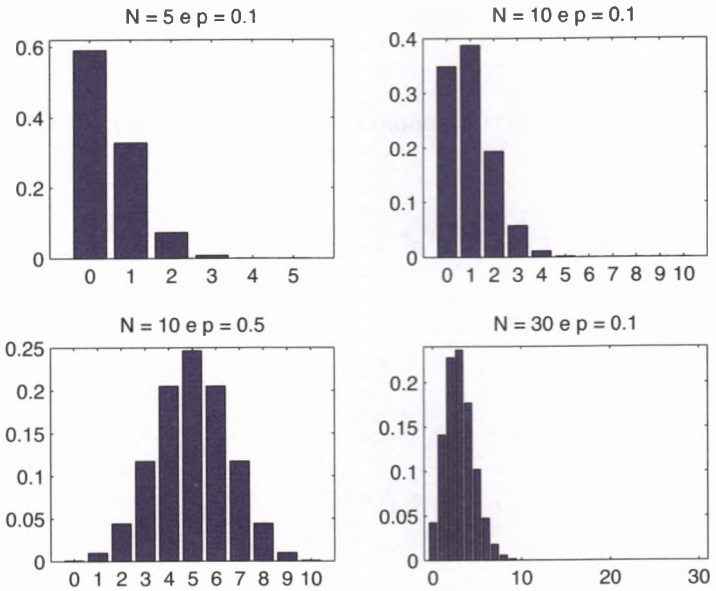


Figura 3.1: Função de frequência para a variável aleatória binomial.

Contratos de opções também podem ser caracterizados pela especificação do período de exercício e, nesse caso, os tipos mais comuns são:

1) Opção Europeia é um contrato de opção que só pode ser exercido apenas em uma data fixa específica no futuro.

2) Opção Americana é um contrato de opção que pode ser exercido em qualquer instante até a data de vencimento.

É válido comentar que existe uma grande variedade de contratos de opções especificados de várias formas e disponíveis para vários tipos de mercados. Uma apresentação didática desses itens pode ser encontrada, por exemplo, em Hull (1997) e Wilmott, Howison e Dewynne (1995). Adicionalmente, uma aplicação importante da teoria de apreçamento de contratos de opções é o apreçamento de opções reais que podem ser usadas para flexibilizar a avaliação de projetos (COPELAND; ANTIKAROV, 2002; KOLLER; MURRIN; COPELAND, 2001; DIXIT; PINDYCK, 1994).

O problema em que estamos interessados aqui é como apreçar contratos de opções Europeias.[9] Desejamos responder à seguinte pergunta: quanto devemos pagar hoje por um contrato de opção para ter direito à compra (venda) de um ativo que vale hoje S, numa data futura, por um preço X?

Sabemos que uma opção de compra numa determinada data t tem valor dado por

$$C_t = \max(S_t - X, 0)$$

[9] A metodologia considerada aqui pode ser facilmente adaptada para o cálculo de preços de opções Americanas.

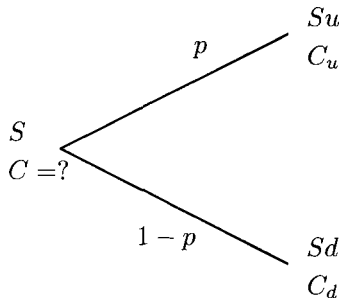


Figura 3.2: Árvore binomial com um período de tempo.

e uma opção de venda numa determinada data t tem valor dado por

$$P_t = \max(X - S_t, 0)$$

onde C_t é o preço de uma opção de compra no instante t, P_t é o preço de uma opção de venda no instante t, X é o preço de exercício e S_t é o preço do ativo no instante t. Note que o cálculo desses valores é bem intuitivo. Considere, por exemplo, o caso da opção de compra. Na data t iremos pagar X por S_t. Logo, se $S_t - X > 0$, C_t deve ser justamente igual a esse valor. Por outro lado, se $S_t - X < 0$, esse contrato de opção não vale nada, pois um contrato de opção é um direito, e não um dever. Dessa forma, não faz sentido pagar por um ativo mais do que ele vale no mercado.

O modelo binomial introduzido por Cox, Ross e Rubinstein (1979) é a forma mais simples e flexível para avaliar contratos de opções. O método é baseado na ideia de construir uma árvore chamada de binomial que pressupõe a ausência de arbitragem[10] e que deve acompanhar as trajetórias do preço.

Considere a árvore binomial apresentada na Figura 3.2 que contém apenas um período. Nessa árvore, estamos interessados em calcular o valor de uma opção de compra C[11] no instante inicial a partir das estimativas que temos do valor do ativo no próximo período. Supondo que preço do ativo hoje é S, no próximo período será Su, se o valor do ativo subir, ou Sd, se o valor do ativo cair, $u > 1$ e $d < 1$. Note que de posse dos valores de subida Su e de que queda Sd do ativo, podemos calcular imediatamente o preço do contrato de opção C_u no caso de subida do preço do ativo, e C_d no caso de queda no preço do ativo. De fato, se estamos calculando o preço de uma opção de compra e o ativo subiu para Su, então o preço da opção no próximo período será $C_u = \max(Su - X, 0)$. Se o preço do ativo caiu para Sd, então $C_d = \max(Sd - X, 0)$.

[10] Arbitragem pode ser definida como a compra e venda simultânea de um ativo com o objetivo de obter lucro pela diferença de preço nos dois mercados. Formalmente, é uma carteira que paga *payoff* positivo com preço zero (ou negativo) ou uma carteira que paga payoff positivo (ou zero) com preço negativo (LEROY; WERNER, 2001).

[11] No caso de calcular o preço de uma opção de venda P, a ideia é análoga.

Para usarmos explicitamente a hipótese de não arbitragem, vamos construir uma carteira livre de risco com valor Π, formada pela compra de Δ ativos e pela venda de um contrato de opção, como apresentado abaixo:

$$\Pi = \Delta S - C.$$

Para essa carteira ser livre de risco, precisamos que independentemente do estado da natureza- um estado com um aumento do valor do ativo para Su ou a redução do valor do ativo para Sd- ela tenha o mesmo valor. Então, para isso ocorrer, precisamos fazer

$$\Delta Su - C_u = \Delta Sd - C_d,$$

que resolvendo para o valor de Δ, encontramos

$$\Delta = \frac{C_u - C_d}{Su - Sd}.$$

Portanto, para não haver arbitragem,[12] devemos ter

$$\Delta S - C = \frac{\Delta Su - C_u}{1+r} = \frac{\Delta Sd - C_d}{1+r},$$

onde r é a taxa livre de risco. Substituindo Δ na equação anterior, chegamos a

$$C = \frac{qC_u + (1-q)C_d}{1+r}$$

onde

$$q = \frac{(1+r) - d}{u - d}.$$

O preço da opção C hoje calculado acima tem uma interpretação bastante interessante. Primeiro, note que $d < 1 + r < u$, pois em caso contrário, teríamos o ativo livre de risco sempre melhor ou sempre pior que o ativo arriscado, o que faria que, em equilíbrio, um dos dois títulos sumisse do mercado. Por exemplo, se $d > 1 + r$, nunca ninguém teria interesse em investir no ativo livre de risco, pois em qualquer situação o ativo arriscado estaria dando retorno maior que esse ativo livre de risco.[13] Portanto, q tem sabor de probabilidade, pois $0 < q < 1$. Dessa forma, o que esse modelo nos diz é que o valor da opção hoje C é o

[12] Note que uma carteira livre de risco deve capitalizar à taxa livre de risco.
[13] Formalmente, isso é uma arbitragem.

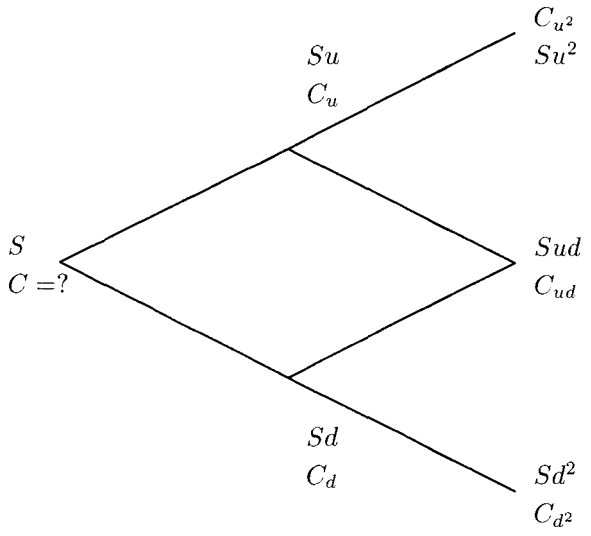


Figura 3.3: Árvore binomial com dois períodos.

valor esperado descontado dos valores futuros possíveis do contrato de opção num mundo neutro ao risco (independentemente do estado, a carteira com valor Π que construímos possui o mesmo valor), onde q é a probabilidade neutra ao risco. A probabilidade p do mundo real não tem nenhuma influência no cálculo do preço da opção.

Vamos agora estender a metodologia apresentada acima para vários períodos. Considere, inicialmente, o caso de dois períodos apresentado na Figura 3.3. Logo, usando os mesmos argumentos já apresentados acima, o preço da opção deve ser calculado de trás para frente. Dessa forma, primeiro calculamos C_{u^2}, C_{ud} e C_{d^2} (que são os preços da opção nos instantes terminais) em função do preço do ativo S, dos valores u e d, da taxa livre de risco r e do preço de exercício X. Depois usamos C_{u^2}, C_{ud} e C_{d^2} para calcular o valor de C_u e C_d (que são os preços das opções nos instantes intermediários), conforme

$$C_u = \frac{qC_{u^2} + (1-q)C_{ud}}{1+r}$$

e

$$C_d = \frac{qC_{ud} + (1-q)C_{d^2}}{1+r}.$$

Finalmente, utilizamos C_u e C_d para calcular o valor de C como em

$$C = \frac{qC_u + (1-q)C_d}{1+r}$$

Podemos estender essa mesma metodologia para T períodos. Considerando que n é o número de movimentos ascendentes do ativo objeto e $ud = 1$, depois de muita álgebra, o valor de um contrato de opção de compra Europeia é dado por

$$\begin{aligned} C &= \frac{1}{(1+r)^T}\left[\sum_{n=0}^{T}\frac{T!}{(T-n)!n!}q^n(1-q)^{T-n}\max(u^n d^{T-n}S - X, 0)\right] \\ &= S\left[\sum_{n=\overline{n}}^{T}\frac{T!}{(T-n)!n!}q^n(1-q)^{T-n}\frac{u^n d^{T-n}}{(1+r)^T}\right] - \frac{X}{(1+r)^T}\left[\sum_{n=\overline{n}}^{T}\frac{T!}{(T-n)!n!}q^n(1-q)^{T-n}\right] \\ &= S\left[\sum_{n=\overline{n}}^{T}\frac{T!}{(T-n)!n!}q'^n(1-q')^{T-n}\right] - \frac{X}{(1+r)^T}\left[\sum_{n=\overline{n}}^{T}\frac{T!}{(T-n)!n!}q^n(1-q)^{T-n}\right] \end{aligned}$$

onde $\overline{n}$ é o número de vezes que a opção zera nos períodos finais, que é o menor inteiro não negativo maior que $\ln(X/Sd^T)/\ln(u/d)$[14], $q = \frac{(1+r)-d}{u-d}$ e $q' = \frac{u}{1+r}q$. A segunda igualdade na expressão acima vem do fato de que os vários termos nulos na expressão que envolve a função max são explicitamente retirados, usando a definição de $\overline{n}$. O valor de um contrato de uma opção Europeia de venda pode ser similarmente derivado substituindo o *payoff* da opção de compra pelo *payoff* da opção de venda.

Se usarmos a notação convencional na literatura de finanças,[15] podemos reescrever a fórmula acima para o preço de um contrato de opção de compra Europeia como

$$C = SBin(n \geq \overline{n}|T, q') - \frac{X}{(1+r)^T}Bin(n \geq \overline{n}|T, q),$$

onde $Bin(\cdot)$ é a função de distribuição binomial acumulada. Sabendo que a distribuição binomial converge para a distribuição normal (que será apresentada na Seção 3.3.1), pode-se mostrar que esse modelo converge para a famosa equação de Black-Scholes (BLACK; SCHOLES, 1973) quando $T \to \infty$, $qT \to \infty$ e $q'T \to \infty$. Essa derivação pode ser encontrada em Cox, Ross e Rubinstein (1979) ou em Copeland e Antikarov (2002). É válido comentar que a equação de Black-Scholes é uma das ideias seminais de finanças que foi responsável por dar o prêmio Nobel de Economia a Scholes em 1997 (Black já tinha falecido dois anos antes).

Nessa aplicação, utilizamos a restrição $ud = 1$ com o objetivo de chegar a uma fórmula fechada para o valor do contrato de uma opção Europeia de compra. No caso geral, essa conta pode ser feita computacionalmente e essa restrição não é necessária.[16] De fato, uma das vantagens dessa metodologia é a

[14] $\ln(X/Sd^T)/\ln(u/d)$ é o n que resolve $Su^n d^{T-n} - X = 0$.
[15] Vide, por exemplo, Copeland e Antikarov (2002).
[16] Em geral, como já foi explicado, o que necessitamos é que $d < 1 + r < u$.

flexibilidade, pois podemos usá-la para apreçar diversos tipos de contratos de opção. Uma referência bem interessante nesse contexto é Wilmott, Howison e Dewynne (1995).

Prática 3.7 Detalhe os cálculos necessários para se chegar a Eq. (3.22).

3.2.3 Variável aleatória de Poisson

A variável aleatória de Poisson pode ser utilizada para modelar processos de contagem, onde o valor resultante pode ser 0, 1, 2, etc. O espaço amostral nesse caso é $\mathbb{X} = \{0, 1, 2, \dots\}$ e a função de frequência é dada por

$$f(x) = \frac{e^{-\lambda}\lambda^x}{x!}, \quad \text{para } x = 0, 1, 2, \dots. \tag{3.22}$$

A variável de Poisson tem um único parâmetro $\lambda \in (0, \infty)$, que assume valores estritamente positivos. A Figura 3.4 apresenta a função de frequência para diferentes valores de λ. Note que, quanto maior o valor de λ, mais a distribuição assume a forma de sino, também observada no caso da variável aleatória binomial.

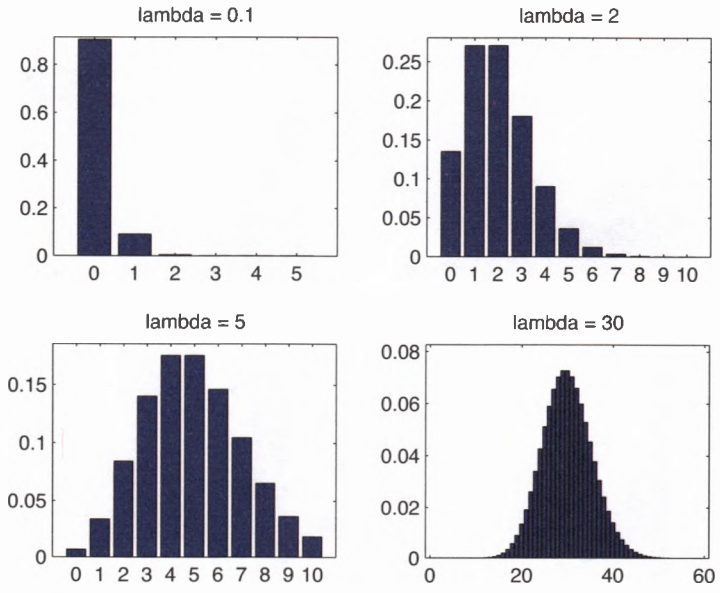


Figura 3.4: Função de frequência para a variável aleatória de Poisson.

Finalmente, a partir da função de frequência apresentada na Eq. (3.22), podemos chegar à média μ e à variância σ^2 da variável aleatória de Poisson

$$\mu = \sigma^2 = \lambda. \tag{3.23}$$

Prática 3.8 Para uma variável aleatória de Poisson, mostre que:
(1) a função $f(x)$ é uma função de frequência,
(2) a média é igual a λ.

3.2.4 Variável aleatória geométrica

A variável aleatória geométrica tem espaço amostral $\mathbb{X} = \{0, 1, 2, 3, \ldots\}$ e, por isso, também é uma candidata na modelagem de processos de frequência de perdas operacionais, por exemplo. Ela possui apenas um parâmetro livre p, que tem de estar localizado no intervalo entre 0 e 1. Os Exercícios 3.14 e 3.15 apresentam possíveis definições para a variável aleatória geométrica.

A função de frequência da variável aleatória geométrica é dada por

$$f(x) = p(1-p)^x, \quad \text{para } x \in \{0, 1, 2, 3, \ldots\}. \tag{3.24}$$

A variável aleatória geométrica possui média $\mu = (1-p)/p$ e variância $\sigma^2 = (1-p)/p^2$. O gráfico da função de frequência da variável aleatória geométrica está mostrado na Figura 3.5 a seguir.

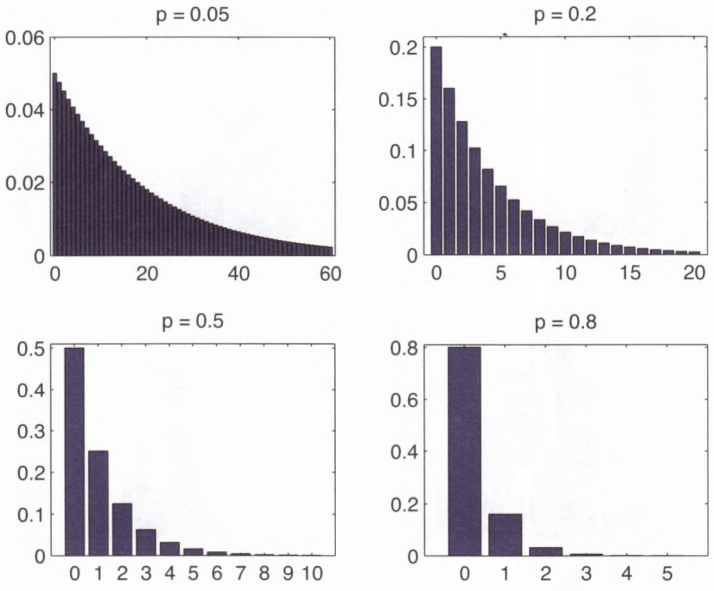


Figura 3.5: Função de frequência para a variável aleatória geométrica.

Prática 3.9 Para uma variável aleatória geométrica, mostre que:
(1) a função $f(x)$ é uma função de frequência,
(2) a média é igual a $(1-p)/p$.

3.2.5 Variável aleatória binomial negativa

A variável aleatória binomial negativa tem dois parâmetros livres $r \in (0, \infty)$ e $p \in (0, 1)$, e espaço amostral $\mathbb{X} = \{0, 1, 2, 3, \dots\}$. A função de frequência da variável aleatória binomial negativa é dada por

$$f(x) = \frac{\Gamma(r+x)}{\Gamma(r)\Gamma(x+1)} p^r (1-p)^x, \text{ para } x \in \{0, 1, 2, 3, \dots\}. \tag{3.25}$$

onde $\Gamma(\cdot)$ é chamada função gamma, discutida no Exercício 3.1. A variável aleatória binomial negativa possui média $\mu = r(1-p)/p$ e variância $\sigma^2 = r(1-p)/p^2$. O gráfico da função de frequência da variável aleatória binomial negativa está mostrado na Figura 3.6 abaixo. A partir das expressões para as funções de frequência da distribuição geométrica e da distribuição binomial negativa, notamos que a distribuição geométrica é um caso particular da distribuição binomial negativa quando o parâmetro r é igual a 1.

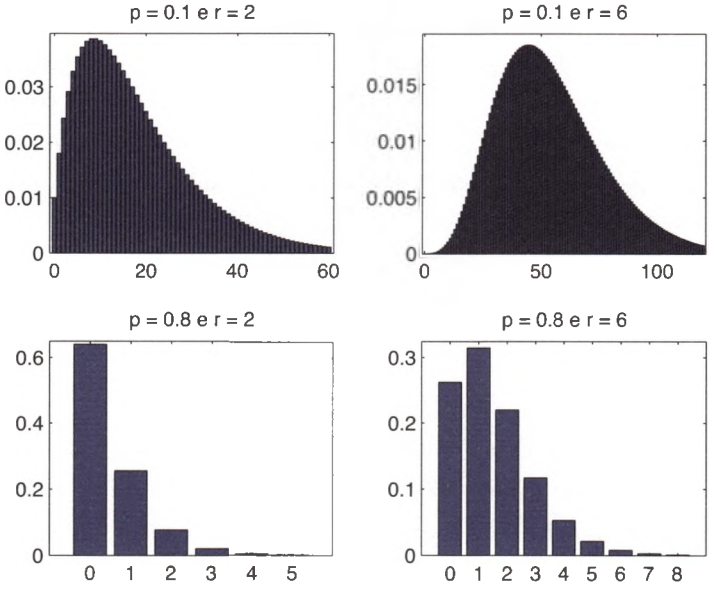


Figura 3.6: Função de frequência para a variável aleatória binomial negativa.

Intuitivamente, imagine um processo de jogar uma moeda, onde a probabilidade de tirar cara é igual a p. Em uma sequência de sucessivas jogadas, a variável binomial negativa modela o número X de jogadas antes de aparecer a r-ésima cara. Portanto, a variável geométrica modela o número X de jogadas antes de aparecer a primeira cara.

3.3 Principais variáveis aleatórias contínuas

Nesta seção faremos uma breve revisão das principais variáveis aleatórias contínuas, comumente utilizadas em modelos aplicados em estatística e econometria. Novamente, não há preocupação aqui em descrever os detalhes técnicos envolvidos nas diversas distribuições. O leitor mais interessado pode recorrer a referências bibliográficas comumente encontradas na literatura introdutória de estatística e probabilidade. Todas as variáveis aleatórias contínuas especificadas aqui serão representadas pela letra maiúscula Y.

3.3.1 Variável aleatória normal

Muito provavelmente a variável aleatória mais comum de ser encontrada é a variável aleatória normal. Conforme sugerido nas Figuras 3.1 e 3.4, a função de densidade da variável aleatória normal tem forma de sino e é o limite da distribuição binomial com N tendendo a infinito, ou da distribuição de Poisson quando λ aumenta indefinidamente. O espaço amostral da variável aleatória normal é dado por $\mathbb{X} = \Re = (-\infty, +\infty)$, e, portanto, ela pode assumir qualquer valor na reta real, tanto positivo quanto negativo. Essa é uma das razões pelas quais a variável aleatória normal é mais comumente utilizada em risco de mercado, e não tão utilizada em risco operacional ou risco de crédito. A distribuição normal também é extremamente importante na literatura de probabilidade e estatística, devido ao resultado conhecido como **teorema central do limite**, que é apresentado no Capítulo 6. Além disso, a distribuição normal é muito utilizada na discussão sobre modelos de regressão linear que serão apresentados no Capítulo 8.

A função de densidade $f(y)$ da distribuição normal é dada por

$$f(y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(y-\mu)^2}, \quad \text{para} \quad y \in (-\infty, +\infty). \tag{3.26}$$

A distribuição normal tem justamente dois parâmetros livres: a média μ e a variância σ^2. O gráfico de $f(y)$ está representado na Figura 3.7, para diferentes valores de μ e σ^2. Note que a dispersão da distribuição aumenta quando σ^2 aumenta, o que está de acordo com a própria definição de variância.

Similarmente a diversas variáveis aleatórias contínuas comumente conhecidas, a função de distribuição acumulada $F(y)$ para a distribuição normal não possui forma explícita, diferentemente dos casos, por exemplo, da variável aleatória exponencial negativa ou da variável aleatória uniforme, que veremos nas próximas seções. Isso ocorre porque não é possível encontrar explicitamente a integral da função de densidade $f(y)$ dada acima. Dada a importância da distribuição normal, mesmo não conhecendo a expressão para a função $F(y)$, é possível obter probabilidades do tipo $\mathrm{P}[Y < 2.0]$, por exemplo, para qualquer distribuição normal (com parâmetros μ e σ^2 gerais). Isso é possível graças à tabela existente, e comumente divulgada, nos livros de econometria e estatística, para algumas probabilidades básicas do tipo $\mathrm{P}[Z < z]$, onde Z é uma variável aleatória com distribuição conhecida como distribuição **normal padronizada**, e

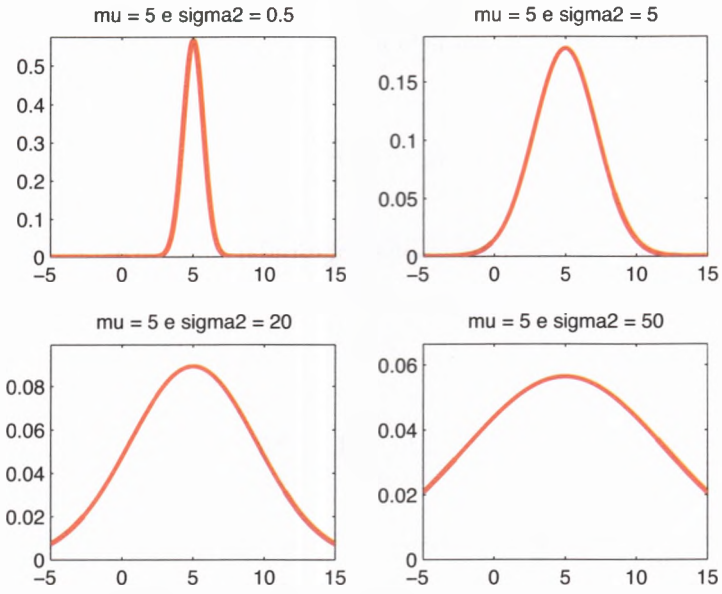


Figura 3.7: Função de densidade para a variável aleatória normal.

z é um valor real qualquer. A distribuição normal padronizada é uma distribuição com parâmetros $\mu = 0$ e $\sigma^2 = 1$. Portanto, problemas gerais envolvendo probabilidades do tipo $\mathrm{P}[Z < z]$ podem ser facilmente resolvidos utilizando-se a tabela da distribuição normal padronizada.

Mas como resolver questões envolvendo probabilidades do tipo $\mathrm{P}[Y < y]$ para variáveis aleatórias X com distribuições normais mais gerais? Felizmente, podemos recorrer a um fato importante no caso de variáveis aleatórias normais. Seja Y uma variável aleatória normal qualquer, com parâmetros μ e σ^2. Pode-se mostrar que a nova variável

$$Z = \frac{Y - \mu}{\sigma}$$

tem distribuição normal padronizada (para detalhes, vide Exemplo 4.17). Portanto, para encontrar $\mathrm{P}[Y < y]$, basta fazer

$$\mathrm{P}[Y < y] = \mathrm{P}\left[\frac{Y - \mu}{\sigma} < \frac{y - \mu}{\sigma}\right] = \mathrm{P}\left[Z < \frac{y - \mu}{\sigma}\right],$$

e a última probabilidade pode ser obtida com base na tabela de probabilidades para a distribuição normal padronizada.

Aplicação 3.2 (Teoria média-variância e CAPM) Nessa aplicação, vamos conhecer as contribuições seminais em finanças que permitiram considerar risco de forma quantitativa na alocação de ativos em uma carteira (MARKOWITZ, 1952) e no apreçamento de ativos (SHARPE, 1963, 1964). Ainda é válido comentar que essas contribuições (MARKOWITZ, 1952; SHARPE, 1963, 1964) foram responsáveis pelos prêmios nobeis recebidos por Harry Max Markowitz e William Forsyth Sharpe em 1990.

Considere que você investiu no ativo i a um preço p_i e depois de um período de tempo você recebeu o *payoff* x_i. Então, o **retorno** que você recebeu pelo investimento nesse ativo é

$$R_i = \frac{x_i - p_i}{p_i}.$$

Vamos considerar agora que estamos interessados em estudar uma carteira formada por dois ativos i e j com retornos R_i e R_j. Então, o retorno da carteira pode ser expresso como a soma ponderada de duas variáveis aleatórias

$$R_c = \omega R_i + (1 - \omega) R_j,$$

onde ω é a proporção investida no primeiro ativo.

Portanto, de acordo com a Proposição 3.3, o valor esperado dessa carteira é dado por

$$\mu_{R_c}(\omega) = \mathrm{E}[R_c] = \mathrm{E}[\omega R_i + (1 - \omega) R_j] = \omega \mu_{R_i} + (1 - \omega) \mu_{R_j} \tag{3.27}$$

onde μ_{R_i} é o valor esperado do retorno R_i e μ_{R_j} é o valor esperado do retorno R_j.

Adicionalmente, de acordo com a definição de variância em Eq. (3.6) e a Proposição 3.4, a variância da carteira pode ser expressa como

$$\begin{aligned} \sigma^2_{R_c}(\omega) &= \mathrm{E}[R_c - \mathrm{E}[R_c]]^2 = \mathrm{E}[\omega R_i + (1 - \omega) R_j - \mathrm{E}[\omega R_i + (1 - \omega) R_j]]^2 \\ &= \omega^2 \sigma^2_{R_i} + (1 - \omega)^2 \sigma^2_{R_j} + 2\omega(1 - \omega) \sigma_{R_i R_j}, \end{aligned}$$

onde $\sigma^2_{R_i}$ é a variância de R_i, $\sigma^2_{R_j}$ é a variância de R_j e $\sigma_{R_i R_j}$ é a covariância entre R_i e R_j.

Utilizando a definição do coeficiente de correlação apresentada na Eq. (3.13) entre duas variáveis aleatórias $\rho_{R_i R_j} = \frac{\sigma_{R_i R_j}}{\sigma_{R_i} \sigma_{R_j}}$, podemos calcular a variância da carteira usando

$$\sigma^2_{R_c}(\omega) = \omega^2 \sigma^2_{R_i} + (1 - \omega)^2 \sigma^2_{R_j} + 2\omega(1 - \omega) \rho_{R_i R_j} \sigma_{R_i} \sigma_{R_j}.$$

e o desvio padrão da carteira por

$$\sigma_{R_c}(\omega) = \sqrt{\omega^2 \sigma^2_{R_i} + (1 - \omega)^2 \sigma^2_{R_j} + 2\omega(1 - \omega) \rho_{R_i R_j} \sigma_{R_i} \sigma_{R_j}}. \tag{3.28}$$

Note que esse resultado é bem interessante. Variando o valor de ω nas Eqs. (3.27) e (3.28), podemos construir uma curva com todas as oportunidades de investimento disponíveis para uma carteira formada por dois ativos i e j. Essas oportunidades estão sendo caracterizadas por seu valor esperado (retorno) e por seu desvio padrão (seu risco). Na Figura 3.8, as curvas do valor esperado $\mu_{R_c}(\omega)$ e do desvio padrão $\sigma_{R_c}(\omega)$ da carteira formada por dois ativos i e j são apresentadas. Na Figura 3.9, apresentamos o gráfico $\mathrm{E}[R_c]$ versus σ_{R_c}. O procedimento para construir esse gráfico é escolher ω, calcular os valores correspondentes de $\mu_{R_c}(\omega)$ e $\sigma_{R_c}(\omega)$ e colocar no gráfico esses valores correspondentes. Por isso, para cada σ_{R_c} aparecem dois valores de μ_{R_c}. Para construir as Figuras 3.8 e 3.9, foram usados 100 retornos i e j gerados independentemente, usando $R_i \sim \text{Normal}[0.25, 1]$ e $R_j \sim \text{Normal}[0.15, 0.81]$. Note que na Figura 3.8 o valor esperado aumenta com ω, pois o retorno esperado de i é maior que o retorno esperado de j, isto é, $\mu_{R_i} > \mu_{R_j}$, e aumentando ω estamos aumentando a quantidade investida em i. Adicionalmente, ainda nessa figura, o desvio padrão é mínimo quando a carteira é formada igualmente pelos dois ativos – isso ocorre pois esses retornos foram gerados independentemente.[17]

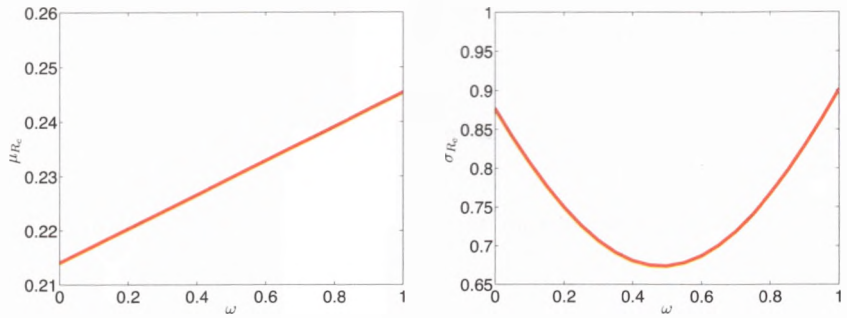


Figura 3.8: Valor esperado da carteira e desvio padrão da carteira em função de ω.

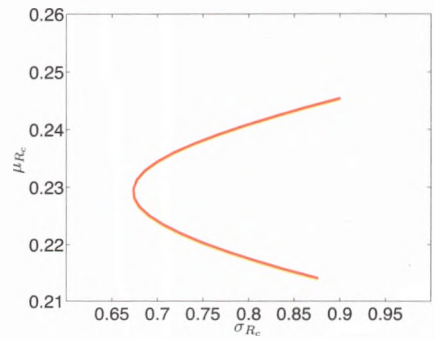


Figura 3.9: Oportunidades de investimento em uma carteira formada por 2 ativos.

Em geral, os investidores gostam de retornos altos e riscos baixos. Portanto, no caso de construção de carteiras formadas por dois ativos, os investidores nunca escolheriam as carteiras que geraram a parte inferior da curva apresentada na Figura 3.9. O valor ω_{Min} onde o desvio padrão da carteira é mínimo pode

[17] Ainda é válido notar uma diferença entre os valores das médias e variâncias usados para o processo gerador e aqueles apresentados, por exemplo, na Figura 3.9. Essa diferença reduz quando o número de retornos gerados aumenta.

ser encontrado facilmente derivando a Eq. (3.28) em relação a ω e igualando essa derivada a zero. Fazendo isso, encontramos

$$\omega_{Min} = \frac{\sigma_{R_j}^2 - \rho_{R_i R_j}\sigma_{R_i}\sigma_{R_j}}{\sigma_{R_i}^2 + \sigma_{R_j}^2 - 2\rho_{R_i R_j}\sigma_{R_i}\sigma_{R_j}}.$$

Vamos agora considerar três casos particulares. Primeiro considere que exista correlação perfeita positiva entre os ativos i e j, isto é, $\rho_{R_i R_j} = 1$, então

$$\mu_{R_c}(\omega) = \omega\mu_{R_i} + (1-\omega)\mu_{R_j},$$

$$\sigma_{R_c}^2(\omega) = \omega^2\sigma_{R_i}^2 + (1-\omega)^2\sigma_{R_j}^2 + 2\omega(1-\omega)\sigma_{R_i}\sigma_{R_j} = (\omega\sigma_{R_i} + (1-\omega)\sigma_{R_j})^2,$$

ou seja,

$$\sigma_{R_c}(\omega) = \omega\sigma_{R_i} + (1-\omega)\sigma_{R_j}$$

Note que se isolarmos o valor ω como função de σ_{R_c}, σ_{R_i} e σ_{R_j}, então teremos uma relação linear entre μ_{R_c} e σ_{R_c}.

Agora considere que exista correlação perfeita negativa entre os ativos i e j, isto é, $\rho_{R_i R_j} = -1$, então

$$\mu_{R_c}(\omega) = \omega\mu_{R_i} + (1-\omega)\mu_{R_j}$$

$$\sigma_{R_c}^2(\omega) = \omega^2\sigma_{R_i}^2 + (1-\omega)^2\sigma_{R_j}^2 - 2\omega(1-\omega)\sigma_{R_i}\sigma_{R_j} = (\omega\sigma_{R_i} - (1-\omega)\sigma_{R_j})^2,$$

ou seja,

$$\sigma_{R_c}(\omega) = \pm(\omega\sigma_{R_i} - (1-\omega)\sigma_{R_j}).$$

Note que aqui, se isolarmos o valor ω como função de σ_{R_c}, σ_{R_i} e σ_{R_j}, então teremos duas relações lineares entre $\mathrm{E}[R_c]$ e σ_{R_c}.

Os dois casos particulares $\rho_{R_i R_j} = 1$ e $\rho_{R_i R_j} = -1$ são casos limites para a curva apresentada na Figura 3.9. Como mostramos na Figura 3.10, quando $\rho_{R_i R_j} \to 1$, então a curva apresentada na Figura 3.9 converge para a reta $\mu_{R_c}(\omega) = \omega\mu_{R_i} + (1-\omega)\mu_{R_j}$ com $\omega = \frac{\sigma_{R_c} - \sigma_{R_j}}{\sigma_{R_i} - \sigma_{R_j}}$. Por outro lado, quando $\rho_{R_i R_j} \to -1$, então a

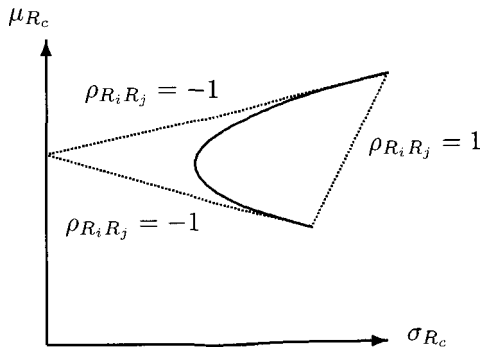


Figura 3.10: Casos limites para a curva de oportunidades de investimentos para uma carteira formada por dois ativos arriscados.

curva apresentada na Figura 3.9 converge para as retas $\mu_{R_c}(\omega) = \omega\mu_{R_i} + (1-\omega)\mu_{R_j}$ com $\omega = \frac{\sigma_{R_c}+\sigma_{R_j}}{\sigma_{R_i}+\sigma_{R_j}} \in [\omega_{Min}, 1]$ para uma reta e $\omega = \frac{\sigma_{R_j}-\sigma_{R_c}}{\sigma_{R_i}+\sigma_{R_j}} \in [0, \omega_{Min}]$ para a outra reta.

Finalmente, considere no último caso que um dos ativos é livre de risco R_f e, portanto, fazendo $\sigma^2_{R_f} = 0$ e $\rho_{R_fR_i} = 0$ e usando as Eqs. (3.27) e (3.28), chegamos a

$$\mu_{R_c}(\omega) = \omega\mu_{R_i} + (1-\omega)R_f,$$

e

$$\sigma_{R_c}(\omega) = \omega^2\sigma^2_{R_i}.$$

Portanto, o conjunto de oportunidades nesse caso é dado por

$$\mu_{R_c} = R_f + \frac{\mu_{R_i} - R_f}{\sigma(R_i)}\sigma_{R_c}.$$

e apresentado na Figura 3.11.

Nesse exemplo, quando decidimos trabalhar com apenas dois ativos simplificamos drasticamente o nosso problema de escolha. Entretanto, pode-se mostrar (MERTON, 1972; COSTA; ASSUNçAO, 2005; MELE, 2007) que qualquer carteira formada apenas por ativos arriscados terá uma curva de oportunidades de investimentos com forma similar àquela apresentada na Figura 3.9.

Portanto, se traçarmos uma curva que representa as oportunidades de investimentos de uma carteira formada por vários ativos arriscados e um ativo livre de risco, então essa curva terá a forma apresentada na Figura 3.12. Note que essa curva é a união de duas curvas, a curva apresentada na Figura 3.9 e a curva

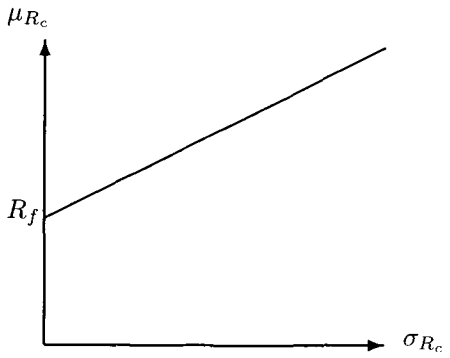


Figura 3.11: Oportunidades de investimento para uma carteira formada por um ativo arriscado e um ativo livre de risco.

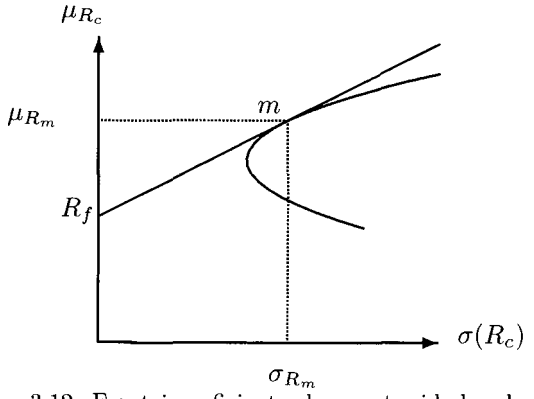


Figura 3.12: Fronteira eficiente de oportunidades de investimentos para um ativo livre de risco e vários ativos arriscados

apresentada na Figura 3.11. Nessa curva podemos identificar a **fronteira eficiente**, que é o conjunto de todas as carteiras eficientes, onde uma **carteira eficiente** é uma carteira que tem a menor variância dentre todas aquelas carteiras que têm o mesmo retorno que ela. A fronteira eficiente é formada por carteiras com o nível máximo de diversificação, pois para conseguir alcançar um mesmo retorno com uma menor variância, é necessário que as variações idiossincráticas do conjunto de ativos que formam a carteira sejam canceladas ao máximo. A fronteira eficiente está representada pela reta que passa pelos pontos $(0, R_f)$ e $(\mu_{R_m}, \sigma_{R_m})$. Supondo que os investidores são aversos ao risco, todos os investidores vão escolher carteiras formadas pelo ativo livre de risco e o ativo m. O ativo m é chamado de carteira de mercado. Esse ativo é caracterizado pelo ponto em que a reta que sai do ativo livre de risco tangencia a curva que representa as carteiras formadas pelos ativos arriscados. Note que, dado que o retorno desejado foi escolhido, nenhuma carteira tem menor variância que essa carteira formada pelo ativo m e o ativo livre de risco.

Considere agora uma carteira consistindo de uma porcentagem ω investida em um ativo arriscado i e uma porcentagem $(1-\omega)$ investida na carteira de mercado m. Então, de acordo com as Eqs. (3.27) e (3.28), essa carteira terá valor esperado e desvio dados da seguinte forma:

$$\mu_{R_c}(\omega) = \omega\mu_{R_i} + (1-\omega)\mu_{R_m}$$

$$\sigma_{R_c}(\omega) = [\omega^2\sigma^2_{R_i} + (1-\omega)^2\sigma^2_{R_m} + 2\omega(1-\omega)\sigma_{R_iR_m}]^{\frac{1}{2}}$$

Derivando essas equações em relação a ω, chega-se a

$$\frac{d\mu_{R_c}(\omega)}{d\omega} = \mu_{R_i} - \mu_{R_m}$$

$$\begin{aligned}\frac{d\sigma_{R_c}(\omega)}{d\omega} = \frac{1}{2}&\left[\omega^2\sigma^2_{R_i} + (1-\omega)^2\sigma^2_{R_m} + 2\omega(1-\omega)\sigma_{R_iR_m}\right]^{-\frac{1}{2}} \\ &\times \left[2\omega\sigma^2_{R_i} - 2\sigma^2_{R_m} + 2\omega\sigma^2_{R_m} + 2\sigma_{R_iR_m} - 4\omega\sigma_{R_iR_m}\right].\end{aligned}$$

No ponto m, temos $\omega = 0$, e, portanto,

$$\left.\frac{d\mu_{R_c}(\omega)}{d\omega}\right|_{\omega=0} = \mu_{R_i} - \mu_{R_m}$$

$$\left.\frac{d\sigma_{R_c}(\omega)}{d\omega}\right|_{\omega=0} = \frac{1}{2}(\sigma^2_{R_m})^{-\frac{1}{2}}\left[-2\sigma^2_{R_m} + 2\sigma_{R_iR_m}\right] = \frac{\sigma_{R_iR_m} - \sigma^2_{R_m}}{\sigma_{R_m}}.$$

Logo, a inclinação da curva μ_{R_c} versus σ_{R_c} no ponto m é dada por[18]

$$\frac{\left.\frac{d\mu_{R_c}(\omega)}{d\omega}\right|_{\omega=0}}{\left.\frac{d\sigma_{R_c}}{d\omega}\right|_{\omega=0}} = \frac{\mu_{R_i} - \mu_{R_m}}{\frac{\sigma_{R_iR_m} - \sigma^2_{R_m}}{\sigma_{R_m}}}.$$

Igualando essa expressão à inclinação da fronteira eficiente chegamos a

$$\frac{\mu_{R_m} - R_f}{\sigma_{R_m}} = \frac{\mu_{R_i} - \mu_{R_m}}{\frac{\sigma_{R_iR_m} - \sigma^2_{R_m}}{\sigma_{R_m}}}.$$

Resolvendo para μ_{R_i}, chega-se ao CAPM (*capital asset pricing model*)

$$\mu_{R_i} = R_f + [\mu_{R_m} - R_f]\frac{\sigma_{R_iR_m}}{\sigma^2_{R_m}}. \tag{3.29}$$

A quantidade $\frac{\sigma_{R_iR_m}}{\sigma^2_{R_m}}$ é conhecida como β_i e representa o **risco não diversificável** (risco sistemático) de uma carteira no mercado em equilíbrio. O termo risco não diversificável vem do fato de que as carteiras eficientes são maximamente diversificadas. Portanto, o único risco que sobra é o não diversificável. A

[18] Esse resultado é encontrado aplicando-se a regra da cadeia e o teorema da função inversa. Esses teoremas são clássicos em cálculo e podem ser encontrados, por exemplo, em Simon e Blume (2004).

variável $\frac{\mu_{R_m}-R_f}{\sigma_{R_m}}$ é chamada de **preço ao risco de mercado**,[19] que é uma medida do retorno extra que os investidores exigem por aceitarem mais risco. Na Figura 3.13 é apresentada a linha de mercado de capitais que é a forma gráfica do modelo do CAPM apresentado na Eq. (3.29). O CAPM mostra que existe uma relação unívoca entre risco e retorno. Se um investidor deseja receber um determinado retorno médio por seu investimento, então, para que isso ocorra, ele deve aceitar uma parcela específica de risco dada pela linha de mercado de capitais. Se um investidor não aceita que o risco de sua carteira seja maior que um determinado valor, então, de acordo com a linha de mercado de capitais, ele também estará restringindo o retorno médio de sua carteira. O CAPM é um dos modelos mais importantes e a base da teoria de apreçamento de ativos em finanças, pois a partir das relações de retornos esperados apresentadas acima, podemos calcular preços. Além disso, ele é muito útil para calcular taxas de descontos que poderão ser usadas, por exemplo, em métodos de valoração baseados no cálculo do valor presente líquido (KOLLER; MURRIN; COPELAND, 2001).

Gostaríamos ainda de comentar que algumas hipóteses precisaram ser feitas para que a matemática desenvolvida acima seja válida. Primeiro, supusemos que nesse mercado, os investidores são indivíduos aversos ao risco que maximizam a utilidade esperada de sua riqueza (vide Seção 4.6.2). Note que se essa hipótese não fosse válida, nem todos gostariam de investir na fronteira eficiente. Segundo, uma outra hipótese importante é que existe um ativo livre de risco que os investidores podem tomar emprestado ou emprestar sem restrições nessa taxa. De fato, embora haja algumas aproximações, na prática, sabemos que não existem ativos livres de risco. Entretanto, a hipótese que cerne sobre a existência de um ativo livre de risco pode ser relaxada (BLACK, 1972). Terceiro, supusemos também que todos os investidores acessam a mesma informação, pois em caso contrário nem todos poderiam saber sobre a fronteira eficiente. Quarto, todos os investidores são tomadores de preço, isto é, suas ações não afetam o mercado. Quinto, todos os ativos podem ser comprados ou vendidos em qualquer quantidade. Note que, no mundo real, ativos são negociados em lotes. Sexto, supusemos que as únicas características relevantes dos ativos são seus valores médios, suas variâncias e a covariância entre eles. Supusemos também que esses valores são constantes ao longo do tempo. Na prática, isso não é verdade, mas essa hipótese pode ser considerada uma primeira aproximação. Sétimo, não consideramos impostos ou custos de transação.

Ainda é válido comentar que a apresentação da teoria média-variância usada aqui usou basicamente noções de estatística e cálculo como em, por exemplo, Copeland e Weston (1992), Securato (1996), Cuthbertson e Nitzsche (2005), Costa e Assunçao (2005) e Mele (2007). A versão mais moderna dessa teoria é construída em espaços vetoriais com produto interno (LUENBERGER, 1969). Embora essa versão fuja dos objetivos deste livro, o leitor mais interessado pode buscar essa apresentação em LeRoy e Werner (2001) e Cochrane (2005).

[19] Do inglês, *market price of risk*.

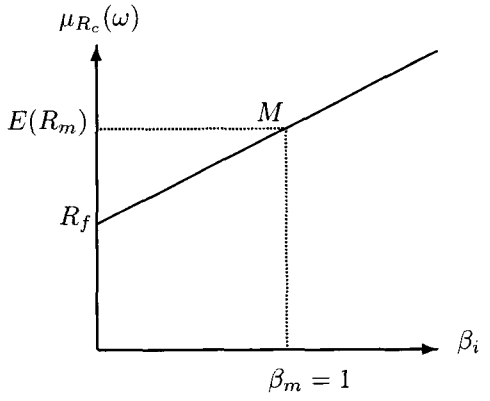


Figura 3.13: Linha de mercado de capitais.

3.3.2 Variável aleatória exponencial negativa

A variável aleatória exponencial negativa apresenta espaço amostral $\mathbb{X} = (0, \infty)$ e tem apenas um parâmetro livre $\lambda \in (0, \infty)$. A função de densidade[20] $f(y)$ é dada por

$$\begin{aligned} f(y) &= \lambda e^{-\lambda y}, \text{ para } y \in (0, \infty), \\ &= 0, \text{ para } y \in (-\infty, 0]. \end{aligned} \tag{3.30}$$

A distribuição exponencial negativa tem média $\mu = 1/\lambda$ e variância $\sigma^2 = 1/\lambda^2$. A Figura 3.14 apresenta os gráficos da função de densidade para a variável aleatória exponencial negativa para diferentes valores de λ. Alguns autores utilizam uma formulação um pouco diferente para variável aleatória exponencial negativa, onde, ao invés de utilizarmos um parâmetro λ conforme Eq. (3.30), utiliza-se um parâmetro m, com $m = 1/\lambda$ e $m \in (0, +\infty)$. Com essa nova parametrização, a função de densidade torna-se

$$f(y) = \frac{1}{m} e^{-y/m}, \text{ para } y \in (0, \infty). \tag{3.31}$$

Obviamente, as funções de densidade apresentadas nas Eqs. (3.30) e (3.31) são completamente equivalentes. A vantagem em se utilizar especificamente a parametrização apresentada na Eq. (3.31) é que o parâmetro m corresponde exatamente ao valor esperado $\mathrm{E}[y]$.

Nota 3.4 Na especificação acima para a função de densidade para a variável aleatória exponencial negativa, escrevemos $f(y) = \lambda e^{-\lambda y}$, para $y \in (0, \infty)$, e $f(y) = 0$, para $y \in (-\infty, 0]$. Alguns autores escrevem uma especificação ligeiramente diferente: $f(y) = \lambda e^{-\lambda y}$, para $y \in [0, \infty)$, ou seja, a função assume valor não nulo no ponto $y = 0$. Então estamos tratando de variáveis aleatórias diferentes? Na verdade,

[20] A expressão abaixo apresenta o valor de $f(y)$ positivo para $y > 0$ e 0 caso contrário. Na definição da função de densidade para as próximas variáveis aleatórias contínuas, sempre que for indicado o valor da função $f(y)$ apenas para um subconjunto da reta real $\Re$, isso significa que o valor de $f(y)$ será nulo fora desse subconjunto.

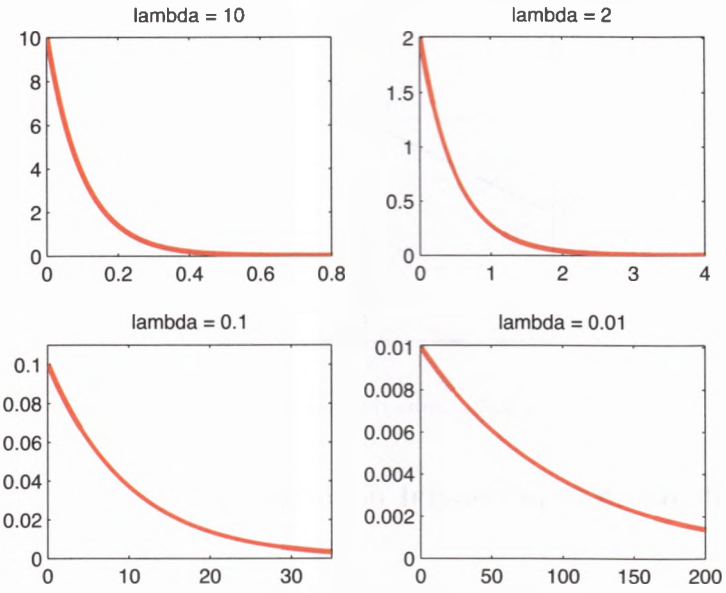


Figura 3.14: Função de densidade para a variável aleatória exponencial negativa.

não. Dizemos que duas variáveis aleatórias contínuas possuem a mesma distribuição quando elas possuem a mesma função de densidade de probabilidade para todo o conjunto $\Re$, com exceção de um conjunto de medida nula. Portanto, as duas especificações para a distribuição exponencial negativa diferem apenas no conjunto $\{0\}$, que tem medida nula, e concluímos que ambas as especificações referem-se à mesma distribuição.

3.3.3 Variável aleatória gamma

Uma outra distribuição também comumente utilizada em risco operacional e outras aplicações em finanças e economia é a variável aleatória gamma. Ela possui dois parâmetros livres: α e β, ambos estritamente positivos, e tem espaço amostral $\mathbb{X} = (0, \infty)$, e, portanto, similarmente à distribuição exponencial negativa, também só pode assumir valores positivos. A distribuição gamma tem função de densidade

$$f(y) = \frac{1}{\beta^{\alpha}\Gamma(\alpha)} y^{\alpha-1} e^{-y/\beta}, \quad \text{para } y \in (0, \infty), \tag{3.32}$$

média $\mu = \alpha\beta$ e variância $\sigma^2 = \alpha\beta^2$. A função $\Gamma(\cdot)$ é conhecida como função gamma, sendo comumente encontrada em qualquer programa matemático ou qualquer livro introdutório de cálculo (veja também Exercício 3.1).

Observe que a distribuição exponencial negativa é um caso particular da distribuição gamma, quando $\alpha = 1$ e $\beta = 1/\lambda$. Esse fato pode ser observado também diretamente da Figura 3.15, que apresenta o gráfico da função de densidade $f(y)$ para diferentes valores de α e β.

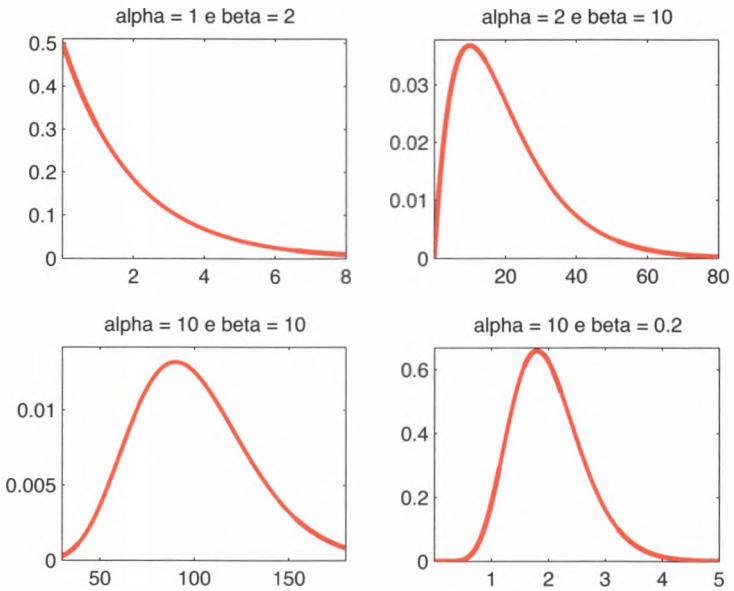


Figura 3.15: Função de densidade para a variável aleatória gamma.

3.3.4 Variável aleatória de Weibull

A variável aleatória de Weibull, assim como a variável gamma e a exponencial negativa, assume somente valores positivos, sendo $\mathbb{X} = (0, \infty)$. Ela possui dois parâmetros livres α e β, e tem função de densidade $f(y)$ dada por

$$f(y) = \beta\alpha^{-\beta}y^{\beta-1}e^{-\left[\frac{y}{\alpha}\right]^{\beta}}, \quad \text{para } y \in (0, \infty). \tag{3.33}$$

Os principais momentos da variável aleatória de Weibull têm expressões

$$\begin{aligned} \mathrm{E}[Y] &= \mu = \alpha\left[\Gamma(1+\beta^{-1})\right] \\ \mathrm{Var}[Y] &= \sigma^2 = \alpha^2\left[\Gamma(1+2\beta^{-1}) - \Gamma(1+\beta^{-1})^2\right]. \end{aligned} \tag{3.34}$$

Note que a variável aleatória exponencial negativa também é um caso particular da variável aleatória de Weibull: basta fazer $\beta = 1$ e $\alpha = 1/\lambda$. A Figura 3.16 apresenta os gráficos da função de densidade $f(y)$ da distribuição de Weibull para diferentes valores de α e β. Assim como no caso da variável gamma, a existência de dois parâmetros livres possibilita à função de densidade da variável de Weibull assumir curvas de formatos variados. Podemos notar que a cauda direita é relativamente leve, o que sugere a inadequação da variável de Weibull para modelar processos de perda onde eventos de alto valor monetário ocorrem com uma razoável frequência. Esse pouco peso na cauda direita da distribuição de Weibull pode ser explicado pelo expoente β no termo $\frac{y}{\alpha}$ na função de densidade apresentada na Eq. (3.33). Esse exponente faz com que o termo $e^{-\left[\frac{y}{\alpha}\right]^{\beta}}$ decaia rapidamente, fazendo com que a função de densidade $f(y)$ se aproxime mais rapidamente do zero quando y aumenta. Finalmente, a partir da distribuição de Weibull, podemos derivar

a distribuição de Gumbel (vide Seção 3.3.7), muito apropriada para dados onde há uma alta ocorrência de valores extremos.

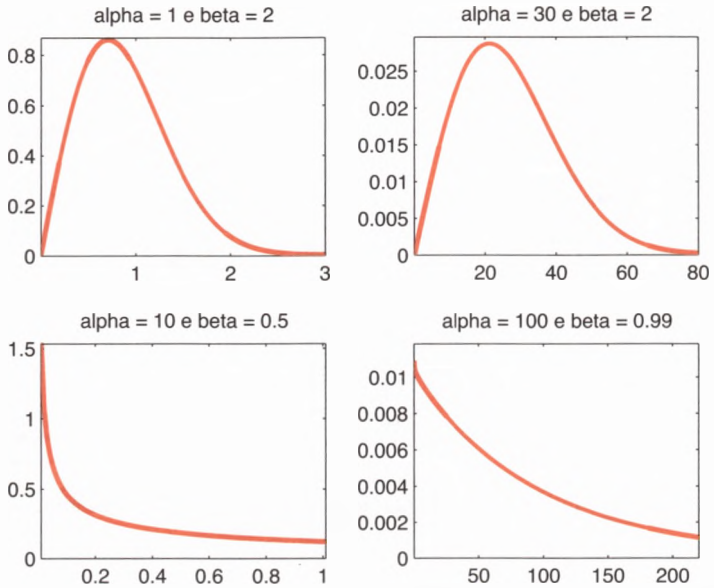


Figura 3.16: Função de densidade para a variável aleatória de Weibull.

3.3.5 Variável aleatória lognormal

A variável aleatória lognormal, como o nome já indica, é derivada da distribuição normal, vista anteriormente. De fato, seja W uma variável aleatória normal, com média μ e variância σ^2, a variável aleatória $Y = e^W$ terá distribuição lognormal com parâmetros μ e σ^2, e função de densidade

$$f(y) = \frac{1}{y\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(\log y - \mu)^2}, \quad \text{para } y \in (0, \infty). \tag{3.35}$$

Quando apresentarmos o teorema para transformação de variáveis aleatórias contínuas na Seção 3.5, ficará claro como obter a função de densidade da distribuição lognormal, a partir da função de densidade da variável aleatória normal. A Figura 3.17 apresenta os gráficos da função de densidade da distribuição lognormal para diferentes valores de μ e σ^2. Apesar de a variável lognormal ter parâmetros μ e σ^2, esses não são os valores da sua média e da sua variância. De fato, a média e a variância no caso da variável lognormal têm expressões

$$\begin{aligned} \mathrm{E}[Y] &= e^{\mu + \frac{\sigma^2}{2}}, \\ \mathrm{Var}[Y] &= \left[e^{2\mu+2\sigma^2} - e^{2\mu+\sigma^2}\right] = e^{2\mu+\sigma^2}\left[e^{\sigma^2} - 1\right]. \end{aligned} \tag{3.36}$$

Conforme veremos na próxima seção, o fato de a variável lognormal ser derivada diretamente da distribuição normal permite a estimação dos parâmetros μ e σ^2 de maneira bem simples.

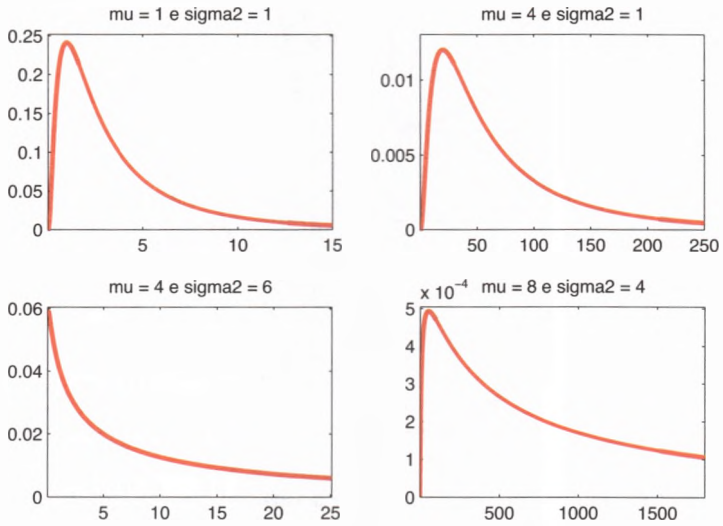


Figura 3.17: Função de densidade para a variável aleatória lognormal.

3.3.6 Variável aleatória de Rayleigh

A variável aleatória de Rayleigh, similarmente às variáveis aleatórias exponencial negativa, de Weibull, lognormal e gamma, também pode ser utilizada para modelar a severidade das perdas operacionais, apresentando espaço amostral $\mathbb{X} = (0, \infty)$. Essa variável aleatória possui apenas um parâmetro livre β. A função de densidade tem expressão

$$f(y) = \frac{y}{\beta^2} e^{\frac{-y^2}{2\beta^2}}, \text{ para } y \in (0, \infty). \tag{3.37}$$

A média e a variância da variável aleatória de Rayleigh são dadas por

$$\begin{aligned} \mathrm{E}[Y] &= \mu = \beta\sqrt{\frac{\pi}{2}}, \\ \mathrm{Var}[Y] &= \sigma^2 = \frac{4-\pi}{2}\beta^2. \end{aligned} \tag{3.38}$$

A Figura 3.18 apresenta a gráfico da função de densidade da variável aleatória de Rayleigh para diferentes valores de β.

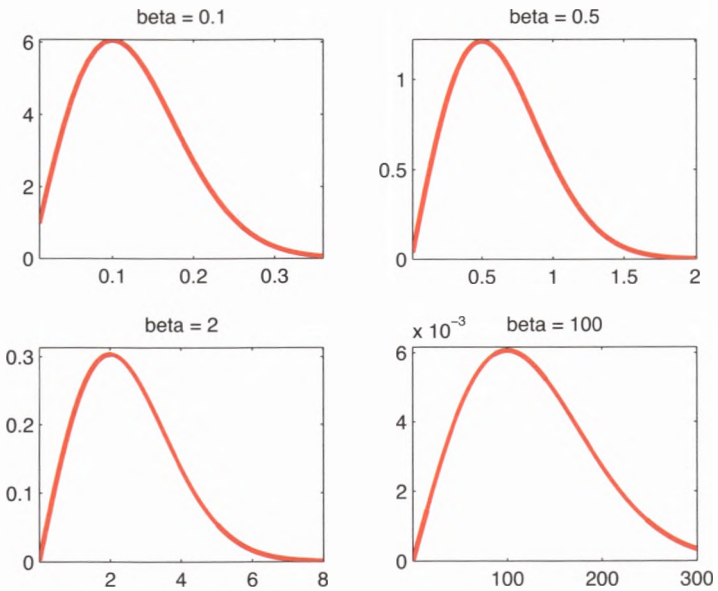


Figura 3.18: Função de densidade para a variável aleatória de Rayleigh.

3.3.7 Variável aleatória de valores extremos

A variável aleatória de valores extremos é também conhecida como variável aleatória de Fisher-Tippett, ou variável aleatória log-Weibull, e pode ser utilizada para modelar processos onde a ocorrência de perdas operacionais com altos valores monetários aconteça com alta probabilidade. Ela pode ser derivada a partir da variável aleatória de Weibull, daí o nome log-Weibull. O espaço amostral novamente é $\mathbb{X} = (0, \infty)$, e a função de densidade tem expressão

$$f(y) = \frac{e^{(\beta-y)/\alpha - e^{(\beta-y)/\alpha}}}{\alpha}, \quad \text{para } y \in (0, \infty). \tag{3.39}$$

A distribuição de valores extremos possui média e variância

$$\begin{aligned} \mathrm{E}[Y] &= \mu = \beta + \alpha\gamma \\ \mathrm{Var}[Y] &= \sigma^2 = \frac{1}{6}\pi^2\alpha^2, \end{aligned} \tag{3.40}$$

onde γ corresponde à constante de Euler-Mascheroni. A Figura 3.19 apresenta o gráfico da função de densidade para diferentes valores de α e β. Observe na figura que a distribuição de valores extremos apresenta, como já esperado, uma cauda à direita bem pesada, indicando que eventos com altos valores de perdas acontecem com mais alta frequência. Conforme veremos mais adiante neste livro, a distribuição de valores extremos pode ser combinada com outras distribuições, de forma a dar mais flexibilidade à modelagem de processos de perdas em geral. Esses modelos combinados são conhecidos como modelos de

mistura ou *mixture models*. Quando $\beta = 0$ e $\alpha = 1$, a distribuição de valores extremos é também conhecida como distribuição de Gumbel.

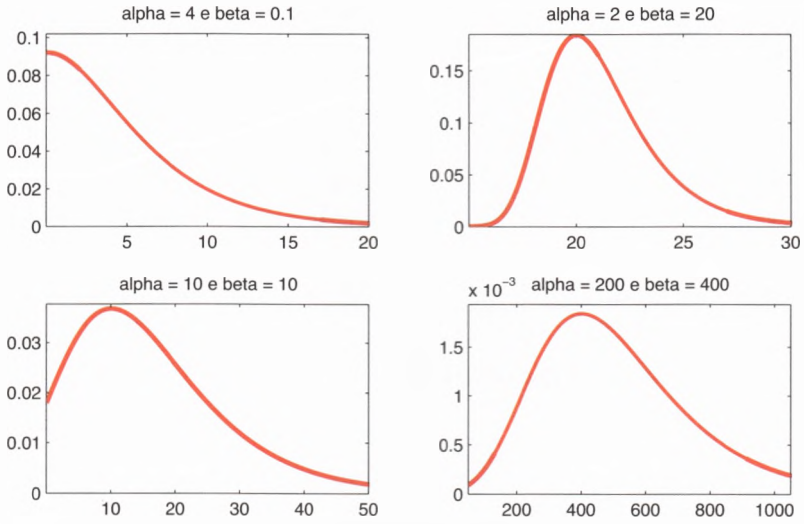


Figura 3.19: Função de densidade para a distribuição de valores extremos.

3.3.8 Variável aleatória de Pareto

A variável aleatória de Pareto tem espaço amostral $\mathbb{X} = (0, \infty)$, e a sua função de densidade tem expressão

$$f(y) = \frac{\alpha\theta}{(y+\theta)^{1+\alpha}}, \quad \text{para} \quad y \in (0, \infty), \tag{3.41}$$

onde α e θ são os dois parâmetros livres. A Figura 3.20 apresenta o gráfico da função de densidade de Pareto para diferentes valores de α e θ.

3.3.9 Variável aleatória qui

A variável aleatória qui tem espaço amostral $\mathbb{X} = (0, \infty)$, e deriva da distribuição qui-quadrada, a qual possui função de densidade

$$f(x) = \frac{x^{(\nu-2)/2} e^{-x/2}}{2^{\nu/2}\Gamma(\nu/2)}, \quad x \in (0, +\infty). \tag{3.42}$$

O parâmetro ν é conhecido como número de graus de liberdade da distribuição qui-quadrada. Pode-se mostrar que a distribuição qui-quadrada é um caso particular da distribuição gamma. De fato, comparando-se a função de densidade apresentada na Eq. (3.42) com a função de densidade para

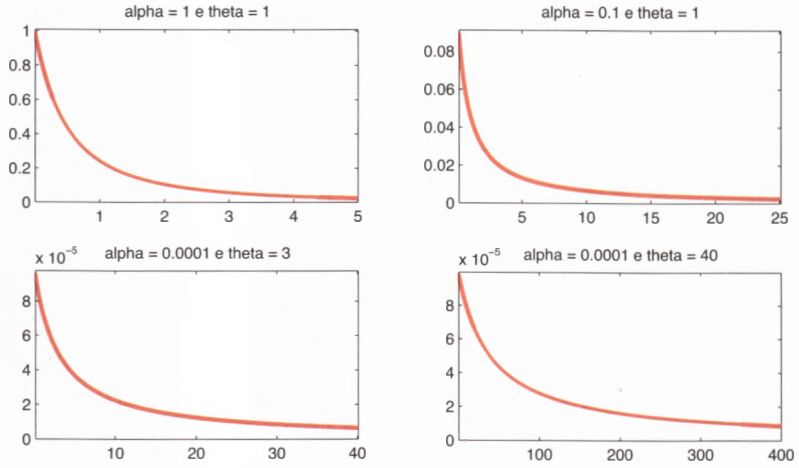


Figura 3.20: Função densidade para a distribuição de Pareto.

a distribuição gamma apresentada na Eq. (3.32), pode-se notar que uma distribuição qui-quadrada corresponde a uma distribuição gamma com parâmetros $\alpha = \nu/2$ e $\beta = 2$.

A distribuição qui pode ser obtida da distribuição qui-quadrada da seguinte forma: seja X uma variável aleatória com distribuição qui-quadrada com parâmetro ν. Seja Y a variável aleatória obtida por meio da expressão $Y = X^{1/2}$. Utilizando-se a fórmula para transformação de variáveis aleatórias (vista mais adiante, neste capítulo), pode-se mostrar que a variável aleatória qui tem função de densidade

$$f(y) = \frac{y^{\nu-1}e^{-y^2/2}}{2^{\nu/2-1}\Gamma(\nu/2)}, \quad y \in (0, +\infty), \tag{3.43}$$

onde o parâmetro $\nu \in (0, \infty)$. A variável aleatória qui possui média e variância

$$\begin{aligned} \mathrm{E}[Y] &= \mu = \frac{\sqrt{2}\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)}, \\ \mathrm{Var}[Y] &= \sigma^2 = \frac{\nu\Gamma\left(\frac{\nu}{2}\right)^2 - 2\Gamma\left(\frac{\nu}{2}+\frac{1}{2}\right)^2}{\Gamma\left(\frac{\nu}{2}\right)^2}. \end{aligned} \tag{3.44}$$

A Figura 3.21 apresenta o gráfico da função de densidade para a distribuição qui para diferentes valores de ν.

Apesar de as expressões para a função de densidade e para os momentos nas variáveis aleatórias discretas e contínuas acima parecerem relativamente simples, a utilização dessas distribuições na prática não é necessariamente uma tarefa fácil. Em alguns casos, os procedimentos numéricos envolvidos podem se tornar demasiadamente complicados e demorados. Uma sugestão é focar especificamente na implementação de distribuições que envolvam procedimentos numéricos mais diretos e confiáveis. Por outro lado, a partir

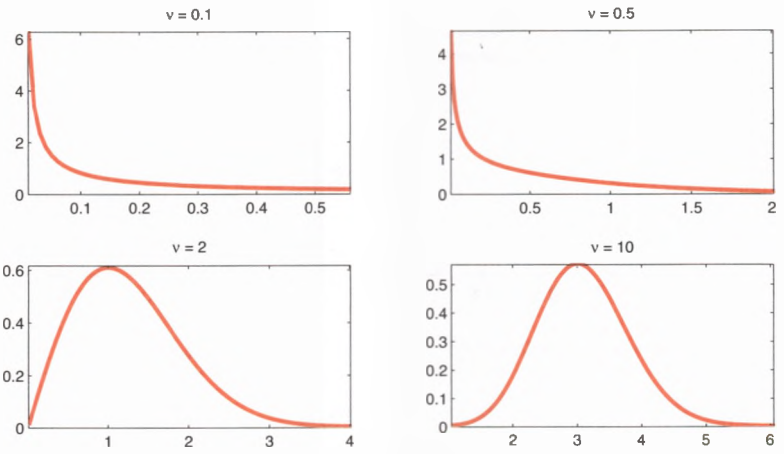


Figura 3.21: Função de densidade para a distribuição qui (raiz quadrada da distribuição qui-quadrada).

da utilização de distribuições simples, é possível combiná-las de forma intuitiva e direta, resultando em modelos muito mais flexíveis e robustos, conforme será discutido mais adiante na Seção 7.4.

3.3.10 Variável aleatória beta

A variável aleatória beta assume valores entre 0 e 1, de forma que $\mathbb{X} = (0, 1)$. Por esse motivo, ela pode ser utilizada, em geral, para modelar variáveis aleatórias que representam taxas, como por exemplo a perda em caso de inadimplência (*loss given default*), comumente encontrada em análise de risco de crédito. A função de densidade $f(y)$ nesse caso é dada por

$$f(y) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} y^{\alpha-1}(1-y)^{\beta-1}, \quad \text{para } y \in (0,1). \tag{3.45}$$

Para ilustrar a variável aleatória beta, a Figura 3.22 apresenta os gráficos de $f(y)$ para diferentes valores de α e β. Observe a grande diversidade de formatos que podem ser obtidos a partir dos diferentes valores de α e β. Quando $\alpha = \beta = 1$, a distribuição beta transforma-se numa variável aleatória uniforme contínua (intuitivamente falando, todos os valores no intervalo (0,1) têm a mesma probabilidade de ocorrência), que será apresentada no Exercício 3.2. Quando α e β são ambos menores do que um, a distribuição beta apresenta um formato de "rede", onde as valores próximos a 0 ou a 1 têm alta probabilidade de ocorrência, quando comparados aos valores mais ao centro. Quando α e β são ambos maiores do que 1, a função de densidade apresenta um formato similar aos observados no caso da distribuição gamma, lognormal ou Weibull, por exemplo.

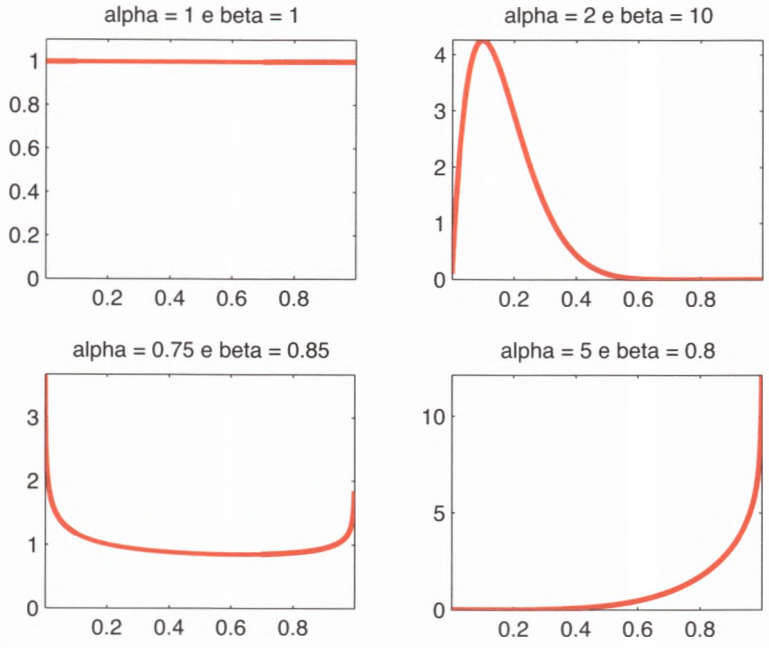


Figura 3.22: Função de densidade para a variável aleatória beta.

A média e a variância de uma variável aleatória beta, com parâmetros livres α e β, são dadas por

$$\begin{aligned} \mu &= \frac{\alpha}{\alpha+\beta}, \\ \sigma^2 &= \frac{\alpha\beta}{(\alpha+\beta+1)(\alpha+\beta)^2}. \end{aligned} \tag{3.46}$$

A variável aleatória beta está relacionada à chamada função beta $B(\cdot,\cdot)$, que tem expressão

$$B(u,v) = \frac{\Gamma(u)\Gamma(v)}{\Gamma(u+v)} = \int_{y=0}^{y=1} y^{u-1}(1-y)^{v-1}dy,$$

onde $\Gamma(\cdot)$ é a função gamma. A expressão acima pode ser utilizada para mostrar que a função de densidade de uma variável aleatória beta integra para um.

Aplicação 3.3 (Redes complexas, distribuição dos graus e "assortatividade")

Até a década de 90, para estudar sistemas estruturados em forma de redes eram tradicionalmente usados modelos baseados em redes regulares e modelos baseados em redes aleatórias. Uma **rede regular** é uma rede em que cada nó da rede tem exatamente o mesmo número de vizinhos. Por outro lado, uma **rede aleatória** (ERDÓS; RÉNYI, 1960) é uma rede em que cada nó tem uma probabilidade constante p de ser conectado com um outro nó pertencente à rede. Uma forma usual de construção de uma rede aleatória é a partir do seguinte algoritmo:

1) Comece com uma rede com n nós;

2) Conecte cada nó com probabilidade p.

De fato, diz-se que enquanto redes regulares são sistemas que apresentam homogeneidade determinística, redes aleatórias possuem homogeneidade estocástica. Na Figura 3.23, nós apresentamos um exemplo de cada uma dessas redes com 20 nós. A rede regular foi construída de forma que cada nó possua 6 vizinhos e a rede aleatória foi construída com probabilidade $p = 0.2$.

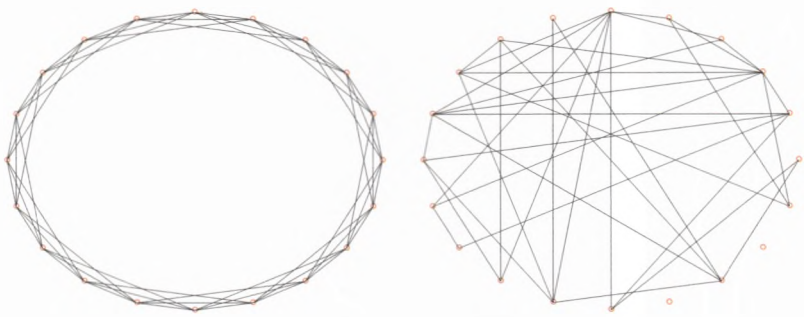


Figura 3.23: Exemplos de uma rede regular e de uma rede aleatória.

No final de década de 90 alguns cientistas do campo da física estatística começaram a tentar entender mais cuidadosamente dados empíricos de redes reais que não podiam ser modelados utilizando a homogeneidade apresentada nas redes regulares e redes aleatórias. Essas redes reais tinham topologias fortemente heterogêneas com padrões não triviais e ficaram conhecidas como **redes complexas**. Uma classe bem ampla de modelos e técnicas foram utilizadas para caracterizar essas redes. Uma revisão desses modelos e técnicas pode ser encontrada em Albert e Barabasi (2002), Boccaletti et al. (2006), Costa et al. (2007) e Newman (2010). Redes complexas têm sido usadas para entender sistemas importantes em economia e finanças, tais como o mercado bancário (BOSS et al., 2004; WAN; CHEN; LIU, 2006; SORAMAKI et al., 2007; CAJUEIRO; TABAK, 2008; IORI et al., 2008; CAJUEIRO; TABAK; ANDRADE, 2009) e a rede de comércio internacional (SERRANO; BOGUNÁ, 2003; SERRANO; BOGUNÁ; VESPIGNANI, 2007; HIDALGO et al., 2007). A teoria econômica também tem sido útil para entender a formação de redes complexas (CAJUEIRO, 2005; JACKSON; ROGERS, 2005; CARVALHO; IORI, 2008).

Um exemplo de uma rede social complexa muito estudada é o "clube de karatê de Zachary" (ZACHARY, 1977). Essa rede representa as relações sociais de um clube de karatê que foi observado no período de três anos, de 1970 a 1972. Durante o período de observação, o clube mantinha entre 50 e 100 membros e suas atividades incluíam, além das aulas usuais de karatê, festas, bailes e banquetes. A organização política do clube era informal e a maioria das decisões era tomada por meio de consenso. Um outro fato relevante sobre esse clube é que, no início do estudo, havia um conflito entre o professor de karatê Sr. Hi e o presidente do clube Sr. John, por causa dos preços das aulas de karatê. Dessa forma, durante esse tempo, o clube ficou dividido em torno dessa questão. Nessa rede, uma conexão é colocada entre dois indivíduos se eles consistentemente interagiam fora do clube. Uma matriz de adjacência de uma rede não direcionada[21] é

[21] Redes em geral podem ser direcionadas ou não direcionadas.

uma matriz quadrada simétrica de ordem n com diagonal nula, formada por 0s e 1s, onde existe um 1 quando existe uma conexão entre os dois membros da rede e 0 em caso contrário. A Tabela 3.1 apresenta a matriz de adjacência da rede do clube de karatê de Zachary, apenas para os nós que possuem pelo menos uma conexão, onde o nó 1 é o Sr. Hi e nó 34 é o Sr. John.

Tabela 3.1: Matriz de adjacência do "clube de karatê de Zachary".

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34
1	0	1	1	1	1	1	1	1	1	0	1	1	1	1	0	0	0	1	0	1	0	1	0	0	0	0	0	0	0	0	0	1	0	0
2	1	0	1	1	1	0	0	1	1	0	0	0	0	1	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	1	0	0	0
3	1	1	0	1	0	0	0	1	1	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
4	1	1	1	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
5	1	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	1	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	1	1	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9	1	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
11	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
12	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
13	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
14	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
15	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
16	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
17	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
18	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
20	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
21	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
22	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
23	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
24	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
25	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
26	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
27	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
28	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
29	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
30	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
31	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
32	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
33	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
34	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

A rede do clube de karatê também é apresentada na Figura 3.24.

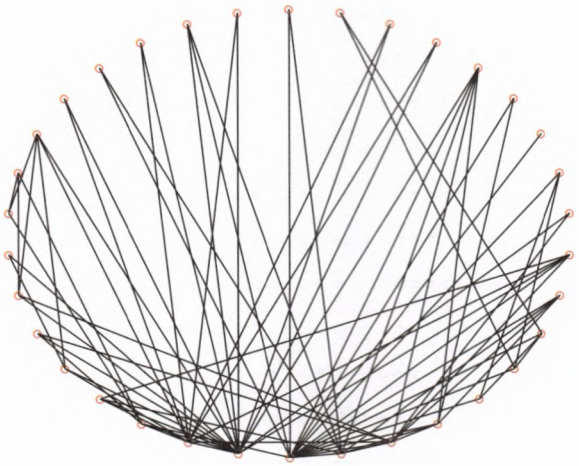


Figura 3.24: Rede do clube de karatê.

Nesse exemplo, vamos estudar apenas duas formas de caracterizar essa rede que estão relacionadas com os graus dos nós:

1) Distribuição dos graus;

2) Coeficiente de "assortatividade".

O grau de um nó é o número de vizinhos que um nó possui. Na rede regular apresentada na Figura 3.23, o grau de cada nó por definição é igual a 6. Na rede aleatória apresentada acima com probabilidade de conexão p, o grau do nó i por definição segue uma distribuição binomial (apresentada na Seção 3.2.2) com parâmetros $n-1$ e p

$$p(k_i = k) = \binom{n-1}{k} p^k (1-p)^{n-1-k}.$$

Essa probabilidade representa o número de formas em que k arestas podem ser sorteadas a partir de um nó i.

Na Figura 3.25 apresentamos a distribuição real dos graus para a rede do clube de Karatê. Esse histograma sugere que essa distribuição tem um comportamento possivelmente similar à distribuição de Pareto[22] (vide Seção 3.3.8) e possui o comportamento bem distinto da distribuição binomial.

Diz-se que uma rede apresenta a propriedade de "assortatividade" se os nós mais conectados de uma rede estiverem conectados a outros nós que têm muitas conexões. Diz-se que uma rede apresenta a propriedade de dissortatividade se os nós mais conectados de uma rede estiverem conectados aos nós menos conectados

[22]Muitas redes complexas possuem distribuição de graus bem representada por uma lei de potência que é uma família de distribuições cuja a distribuição de Pareto é um caso particular.

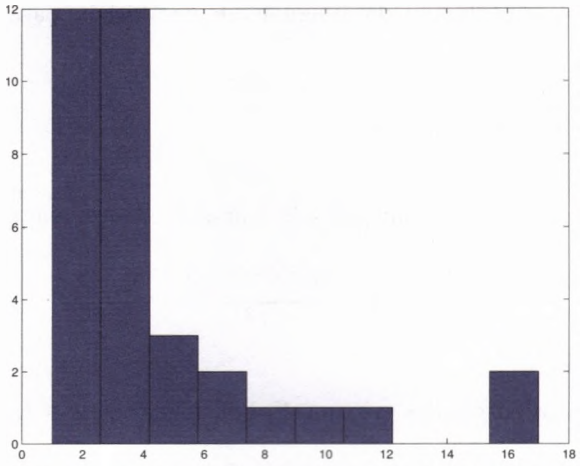


Figura 3.25: Frequencia dos graus da rede de Karatê.

de uma rede. Uma forma de medir essa propriedade, como mostrado em Newman (2002), é calcular o coeficiente de "assortatividade" r, que é equivalente a calcular o coeficiente de correlação de Pearson dos graus dos nós que estão nas pontas de cada aresta. Obviamente, de acordo com a definição do coeficiente de correlação de Pearson, se uma rede possui $r < 0$, ela é dita ser dissortativa. Se uma rede possui $r > 0$, então ela é dita ser "assortativa". Para o caso da rede regular, o coeficiente de "assortatividade" não está definido, pois para o cálculo desse coeficiente, precisa-se dividir pela variância dos graus dos nós, que nesse caso é nula. Para o caso da rede aleatória, o coeficiente de "assortatividade" é, por definição, nulo. E para a rede do clube de karatê, o coeficiente de "assortatividade" é igual a $r = -0.47$. É válido comentar que essa rede é um caso curioso, pois a maioria das redes sociais apresentam coeficientes de "assortatividade" positivos (NEWMAN, 2002). Ou seja, os indivíduos mais populares estão conectados a outros também populares.

3.4 Desigualdades importantes

Nesta seção apresentaremos algumas desigualdades importantes para trabalharmos com os momentos de variáveis aleatórias. As primeiras duas desigualdades abaixo, a desigualdade de Markov e a desigualdade de Chebichev, possibilitam a construção de intervalos de probabilidade para variáveis aleatórias de forma genérica, sejam elas discretas, contínuas ou mistas.

3.4.1 Desigualdade de Markov

Seja $X \in \Re$ uma variável aleatória com média μ e sejam c e r constantes reais quaisquer. Então temos

$$\mathrm{P}\big[|X - \mu| \geq c\big] \leq \frac{\mathrm{E}\big[|X - \mu|^r\big]}{c^r}. \tag{3.47}$$

A desigualdade abaixo é um caso particular da desigualdade de Markov, para $r = 2$.

3.4.2 Desigualdade de Chebishev

Seja $X \in \Re$ uma variável aleatória com média μ e seja c uma constante real qualquer. Então temos

$$\mathrm{P}\big[|X - \mu| \geq c\big] \leq \frac{\mathrm{E}\big[(X-\mu)^2\big]}{c^2} = \frac{\mathrm{Var}[X]}{c^2}. \tag{3.48}$$

Supondo que todos os momentos envolvidos existam. Em particular, se $c = K\sigma$, onde $\sigma = \sqrt{\mathrm{Var}[X]}$, então

$$\mathrm{P}\big[|X - \mu| \geq K\sigma\big] \leq \frac{\sigma^2}{K^2\sigma^2} = \frac{1}{K^2}, \text{ ou}$$
$$\mathrm{P}\big[|X - \mu| < K\sigma\big] \geq 1 - \frac{1}{K^2}.$$

Exemplo 3.5 Seja X uma variável aleatória qualquer, com média $\mu = 2$ e variância $\sigma^2 = 9$. Então

$$\mathrm{P}[|X - \mu| < 6] \geq 1 - 1/4 = 0.75.$$

Ou seja, a probabilidade de uma variável aleatória ter o seu valor ocorrendo dentro de uma faixa de dois desvios padrões, para cada lado, em torno da média é maior ou igual a 75%.

3.4.3 Desigualdade de Jensen

Seja $X \in \Re$ uma variável aleatória qualquer (discreta ou contínua), e seja $g(\cdot)$ uma função convexa, no sentido de que

$$\lambda g(x) + (1-\lambda)g(y) \geq g(\lambda x + (1-\lambda)y),$$

para todo $\lambda \in (0,1)$ e para todo x e $y \in \Re$. Então

$$\mathrm{E}[g(X)] \geq g(\mathrm{E}[X]), \tag{3.49}$$

dado que ambos os valores esperados $\mathrm{E}[|X|]$ e $\mathrm{E}[|g(X)|] < \infty$. Similarmente, se $g(\cdot)$ for côncava, com

$$\lambda g(x) + (1-\lambda)g(y) \leq g(\lambda x + (1-\lambda)y),$$

para todo $\lambda \in (0,1)$ e para todo x e $y \in \Re$. Então

$$\mathrm{E}[g(X)] \leq g(\mathrm{E}[X]). \tag{3.50}$$

Em particular, suponha que a função $g(\cdot)$ é côncava em $\Re$ e existe um intervalo (a, b) onde $g(\cdot)$ é estritamente côncava, ou seja $\lambda g(x) + (1 - \lambda)g(y) < g(\lambda x + (1 - \lambda)y)$ para todo $\lambda \in (0, 1)$ e para todo x e $y \in (a, b)$, $a < b$. Além disso, suponha que a função de densidade de probabilidade para X seja estritamente positiva nesse intervalo; ou seja, $f(x) > 0$ para todo $x \in (a, b)$. Então, temos a desigualdade estrita

$$\mathrm{E}[g(X)] < g(\mathrm{E}[X]).$$

A versão da desigualdade de Jensen vista nesta seção é uma versão para variáveis aleatórias univariadas. Para variáveis aleatórias multivariadas (envolvendo mais de uma variável aleatória), veremos a versão multivariada no Capítulo 4. Neste capítulo, veremos que a desigualdade de Jensen é fundamental para justificar a escolha de funções de utilidade de indivíduos risco-aversos (vide Seção 4.6.2).

3.4.4 Desigualdade de Hölder

Sejam p e q dois números reais $\in [1, \infty]$, com $1/p + 1/q = 1$. Então

$$\mathrm{E}[XY] \leq ||X||_p ||Y||_q, \tag{3.51}$$

onde $||X||_r = (E[|X|]^r)^{1/r}$ para todo $r \in [1, \infty)$, e $||X||_\infty = \inf\{M : \mathrm{P}[|X| > M]) = 0\}$.

3.4.5 Desigualdade de Cauchy-Schwarz

O caso especial onde $p = q = 2$ na desigualdade de Hölder é conhecido como desigualdade de Cauchy-Schwarz:

$$\mathrm{E}\big[|XY|\big] \leq \Big(\mathrm{E}\big[X^2\big]\mathrm{E}\big[Y^2\big]\Big)^{1/2}. \tag{3.52}$$

3.5 Transformações de variáveis aleatórias contínuas

Nesta seção trataremos de um tema bastante importante na análise de variáveis aleatórias, que é o problema de transformação de variáveis aleatórias. Para ilustrar esse tópico, apresentaremos a seguir três exemplos comumente conhecidos na literatura de variáveis transformadas. No primeiro caso, apresentamos a transformação quadrática da variável aleatória normal padronizada. No segundo exemplo, mostramos a soma de três variáveis aleatórias de Poisson, com parâmetros diferentes. Finalmente, no terceiro exemplo, apresentamos uma transformação correspondente à razão entre duas variáveis aleatórias: a primeira é uma variável aleatória normal padronizada; a segunda é a raiz quadrada de uma variável aleatória gamma.

Exemplo 3.6 (Variável aleatória gamma a partir de uma variável aleatória normal) Seja X uma variável aleatória com distribuição normal com média $\mu = 0$ e variância $\sigma^2 = 1$, ou seja, X é uma normal padronizada. Consideremos agora a variável aleatória $Y = X^2$. Para simular uma amostra aleatória de 100.000 observações para a variável Y, basta proceder como se segue:

(1) Simule aleatoriamente uma observação x_1 a partir de uma normal padronizada (utilizando algum pseudo-gerador de números aleatórios).

(2) Calcule $y_1 = x_1^2$. Portanto y_1 corresponde à primeira observação simulada para a variável aleatória Y.

(3) Repita os passos (1) e (2) mais 99.999 vezes, obtendo ao final uma amostra de 100.000 observações para a variável $Y = X^2$.

O processo acima corresponde ao que chamamos de simulação de Monte Carlo, que será abordada com mais detalhes nos Capítulos 5 e 6. O gráfico superior da Figura 3.26 apresenta o histograma de frequências relativas para a amostra de 100.000 observações simuladas para a variável Y, com base no processo acima.

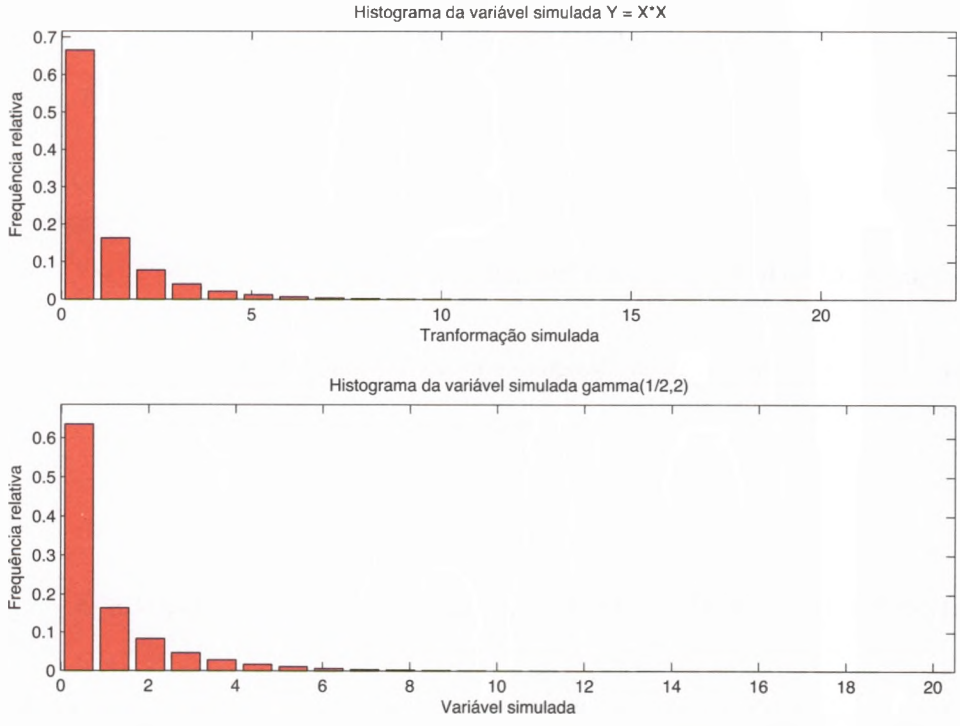


Figura 3.26: Exemplo de transformação quadrática para a variável normal padronizada.

Para comparação, usaremos uma variável aleatória Z, com distribuição gamma, com parâmetros $\alpha = 1/2$ e $\beta = 2$. Simulamos então 100.000 observações z_i com base na variável aleatória especificada. O histograma de frequências relativas está apresentado no gráfico inferior da Figura 3.26. Comparando-se os dois gráficos, observamos que os dois histogramas são praticamente os mesmos. Caso calculemos diversas medidas de

descrição das 100.000 observações simuladas para ambas as bases, notaremos que a média, mediana, desvio padrão, curtose, coeficiente de assimetria, etc. são praticamente os mesmos em ambos os casos. Isso nos sugere que os dois histogramas foram gerados pela mesma distribuição; ou seja, os dois histogramas nos sugerem que o quadrado de uma variável aleatória com distribuição normal padronizada tem distribuição gamma, com parâmetros $\alpha = 1/2$ e $\beta = 2$.

Exemplo 3.7 (Soma de três variáveis aleatórias de Poisson) Consideremos agora a variável aleatória discreta $Y = X_1 + X_2 + X_3$, onde X_i tem distribuição de Poisson com parâmetro λ_i. Suporemos que $\lambda_1 = 2.1$, $\lambda_2 = 3.5$ e $\lambda_3 = 5.3$. Adicionalmente, suporemos que essas três variáveis aleatórias são independentes, no sentido de que o conhecimento do valor de duas delas não adiciona informação alguma sobre o valor da terceira. Novamente, vamos efetuar uma simulação de Monte Carlo com 100.000 observações. Em cada repetição, simulamos uma observação de Poisson para cada uma das três distribuições específicas, obtendo $x_{1,i}, x_{2,i}$ e $x_{3,i}$, e em seguida $y_i = x_{1,i} + x_{2,i} + x_{3,i}$, para $i = 1, \ldots, 100.000$. O histograma para a amostra de 100.000 observações para a variável aleatória Y está apresentado no gráfico superior da Figura 3.27.

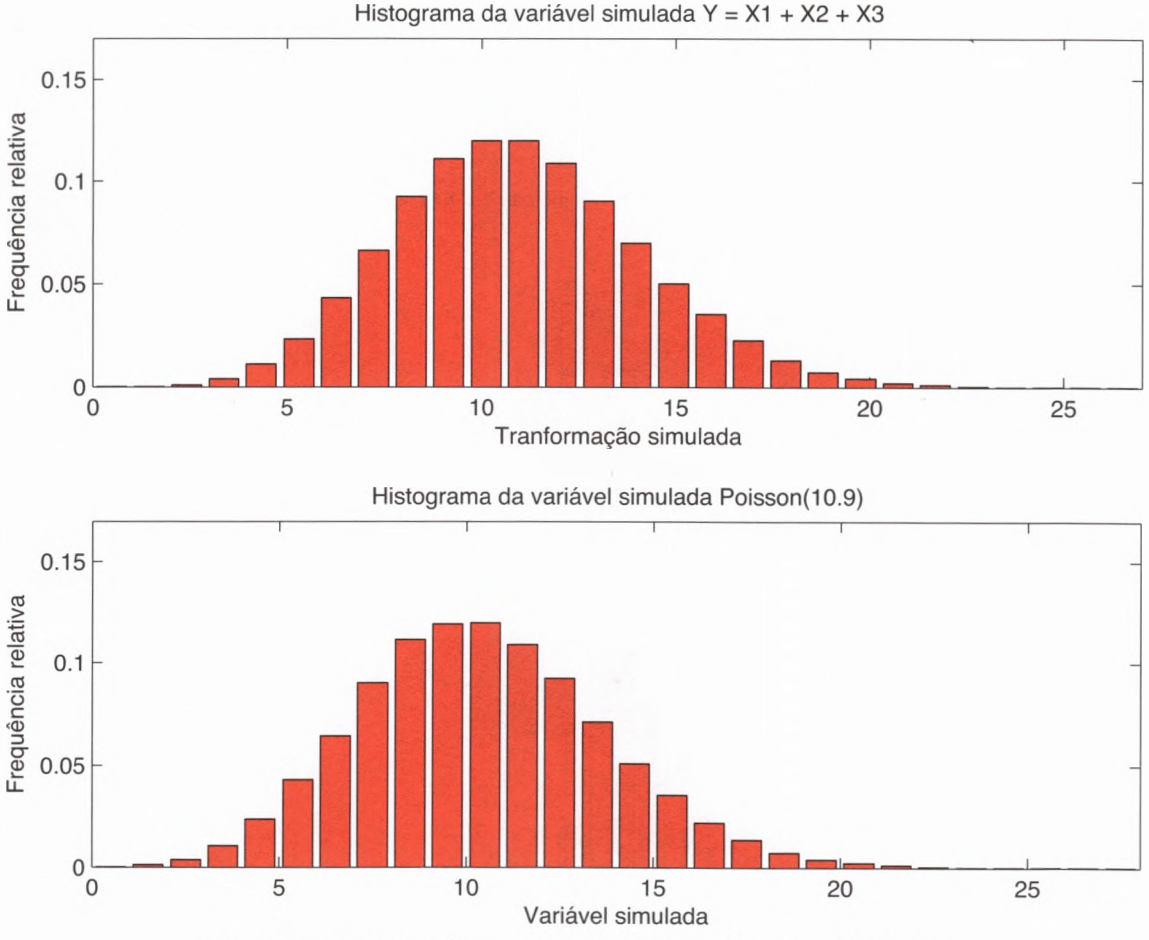


Figura 3.27: Exemplo de soma de três variáveis aleatórias de Poisson.

Para comparação, simulamos uma amostra de 100.000 observações de uma única variável aleatória Z, com distribuição de Poisson com $\lambda = \lambda_1 + \lambda_2 + \lambda_3 = 10.9$. O histograma para a amostra gerada está apresentado no gráfico inferior da Figura 3.27. Novamente, note que os dois histogramas são praticamente idênticos, sugerindo que as variáveis foram geradas da mesma distribuição, apesar de supostamente isso não ter acontecido. Isso nos leva a crer que a variável aleatória dada pela soma de várias variáveis com distribuição de Poisson é também uma variável aleatória de Poisson, com parâmetro λ igual à soma dos parâmetros λs das distribuições individuais que compõem o somatório.

Exemplo 3.8 (Variável aleatória t-Student a partir da combinação de uma variável aleatória gamma e uma variável aleatória normal) Seja X_1 uma variável aleatória com distribuição normal padronizada e seja X_2 uma variável aleatória com distribuição gamma, com parâmetros $\alpha = 8/2$ e $\beta = 2$. Ambas as distribuições são independentes por hipótese. Seja então Y uma variável aleatória construída a partir da razão $Y = X_1/\sqrt{X_2/8}$. Procedemos com simulações de Monte Carlo análogas àquelas feitas nos exemplos anteriores, gerando uma amostra com 100.000 observações; ou seja, geramos observações $x_{1,i}$ e $x_{2,i}$ para as variáveis X_1 e X_2, e depois calculamos a razão $y_i = x_{1,i}/\sqrt{x_{2,i}}$. O histograma dessas 100.000 observações para Y está apresentado no gráfico superior da Figura 3.28.

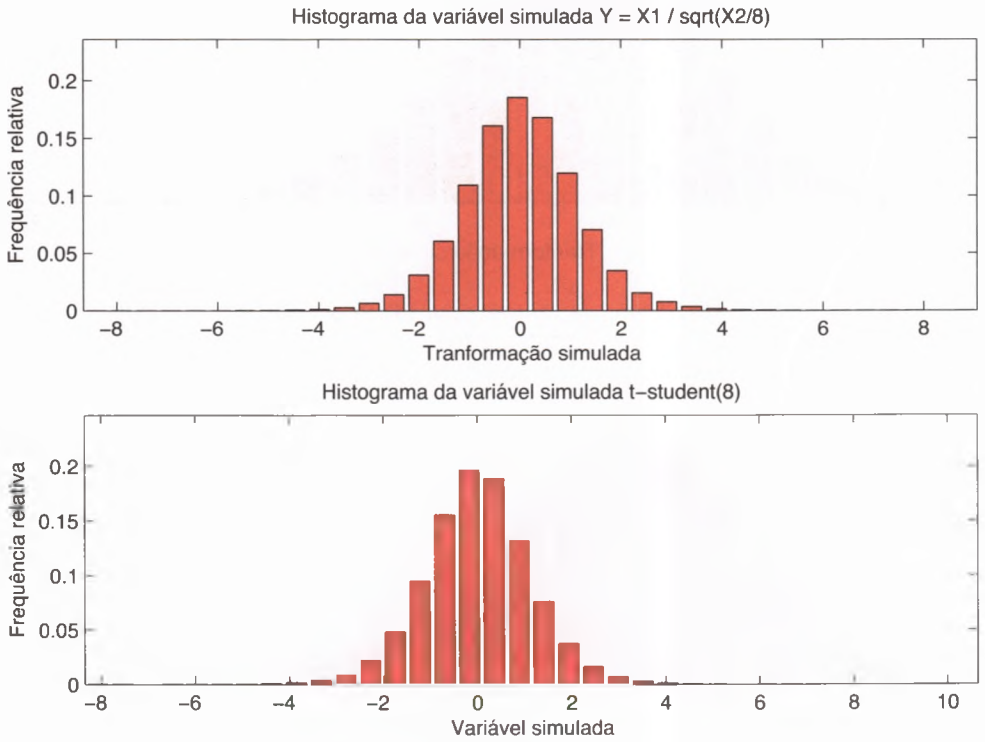


Figura 3.28: Exemplo de transformação quadrática para a variável normal padronizada.

Para comparação, geramos 100.000 observações para uma variável aleatória Z com distribuição t-Student. Essa variável aleatória possui função de densidade de probabilidade dada por

$$f(z) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{(\pi\nu)^{\frac{1}{2}}\Gamma\left(\frac{\nu}{2}\right)\left(1+\frac{z^2}{\nu}\right)^{\frac{\nu+1}{2}}}, \quad \text{para } z \in \Re. \tag{3.53}$$

O parâmetro ν é conhecido como número de graus de liberdade. Devido à sua simetria, a distribuição t-Student possui média $\mathrm{E}[Z] = 0$. Ela é utilizada em risco de mercado ou operacional para modelar processos de caudas pesadas. O parâmetro ν regula o peso nas caudas. Quanto menor o valor de ν, mais pesadas são as caudas da distribuição. Quando ν aumenta, a distribuição t-Student aproxima-se cada vez mais de uma distribuição normal padronizada. De fato, no limite, quando $\nu \to \infty$, a distribuição t-Student converge para uma distribuição normal padronizada.

A distribuição t-Student desempenhará um papel fundamental mais adiante, quando trabalharmos com inferência estatística nos Capítulos 6 e 8. Geramos então 100.000 observações de uma variável aleatória t-Student com número de graus de liberdade $\nu = 8$. O histograma das 100.000 observações geradas está apresentado no gráfico inferior da Figura 3.28. Novamente, observe que, apesar de as duas amostras terem sido geradas de maneiras diferentes, utilizando distribuições diferentes, os histogramas sugerem que estamos trabalhando com a mesma variável aleatória. Portanto, a Figura 3.28 nos leva a crer que o quociente entre uma variável aleatória normal padronizada e a raiz quadrada de uma variável aleatória qui-quadrada,[23] dividida pelo seu número de graus de liberdade, é igual a uma variável t-Student, com o número de graus de liberdade igual ao número de graus de liberdade da distribuição qui-quadrada.

Os três exemplos acima sugerem que transformações de variáveis aleatórias com distribuições conhecidas podem gerar outras variáveis aleatórias, cujas distribuições também são comumente conhecidas (ou pelo menos podem ser escritas analiticamente), mesmo que aparentemente essas outras distribuições pertençam a outras famílias. Esta seção trata exatamente de como podemos obter a função de densidade de probabilidade de uma variável aleatória contínua, construída a partir de uma transformação de outras variáveis aleatórias com funções de densidade de probabilidade conhecidas.

Teorema 3.1 (Transformação de variáveis aleatórias contínuas) Seja X uma variável aleatória contínua com função de densidade de probabilidade $f_X(x)$, e seja $Y = g(X)$, para uma função $g(\cdot)$. Seja $\mathbb{X}$ o espaço amostral[24] para a variável aleatória X. Vamos supor que exista uma partição A_0, A_1, A_2, ..., A_k, do conjunto $\mathbb{X}$, tal que $\mathrm{P}[X \in A_0] = 0$ e $f_X(x)$ é contínua em cada A_i, $i = 1, 2, \ldots, K$. Além disso, vamos supor que existam funções $g_1(x)$, $g_2(x)$, ..., $g_K(x)$, definidas em A_1, A_2, ..., A_K, respectivamente, satisfazendo

[23] Lembre-se da discussão sobre variável aleatória com distribuição qui, que uma variável aleatória qui-quadrada, com ν graus de liberdade, corresponde a uma distribuição gamma, com parâmetros $\alpha = \nu/2$ e $\beta = 2$.

[24] Por exemplo, se X é uma variável aleatória normal, o conjunto $\mathbb{X}$ é o conjunto $(-\infty, +\infty)$.

(i) $g(x) = g_i(x)$, para todo $x \in A_i$, $i = 1, \ldots, K$,
(ii) $g_i(x)$ é monotônica (crescente ou decrescente) em A_i,
(iii) o conjunto $\mathbb{Y} = \mathbb{Y}_i = \{y : y = g_i(x) \text{ para algum } x \in A_i\}$ é o mesmo para todo $i = 1, \ldots, K$,
(iv) $g_i^{-1}(y)$ possui derivadas contínuas em todo $\mathbb{Y}$, para todo $i = 1, \ldots, K$.

Então,

$$\begin{aligned} f_Y(y) &= \sum_{i=1}^{K} f_X\left(g_i^{-1}(y)\right)\left|\frac{d}{dy}g_i^{-1}(y)\right|, \text{ se } y \in \mathbb{Y}, \\ &= 0, \text{ caso contrário.} \end{aligned} \tag{3.54}$$

Exemplo 3.9 (Variável aleatória qui-quadrada com um grau de liberdade a partir de uma variável aleatória normal) Consideremos novamente o exemplo onde X tem distribuição normal padronizada e $Y = X^2$. Queremos encontrar a distribuição de Y. Podemos então aplicar o Teorema 3.1, com $g(x) = x^2$. O espaço amostral $\mathbb{X}$ é a reta real $\Re$, onde a função $g(x)$ é não monotônica; $g(x)$ é decrescente em $(-\infty, 0)$ e crescente em $[0, +\infty)$. Podemos então dividir o conjunto $\mathbb{X}$ em três subconjuntos: $A_1 = (-\infty, 0)$, $A_2 = (0, +\infty)$ e $A_0 = \{0\}$. O conjunto $\{0\}$ possui apenas um ponto, e, portanto, é um conjunto de medida nula, de forma que $\mathrm{P}[X \in A_0] = 0$. O conjunto $\mathbb{Y}$ é igual a $(0, \infty)$, e satisfaz a condição (iii) do Teorema 3.1, para transformação de variáveis aleatórias contínuas.

Para aplicar a Eq. (3.54), precisamos calcular as duas parcelas do somatório (para A_1 e A_2), com

$$\begin{aligned} g_1(x) &= x^2, \quad g_1^{-1}(y) = -\sqrt{y}, \\ g_2(x) &= x^2, \quad g_2^{-1}(y) = \sqrt{y}. \end{aligned} \tag{3.55}$$

A função de densidade de probabilidade de X é

$$f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}, \quad -\infty < x < \infty.$$

A função de densidade de probabilidade para Y, utilizando-se o teorema, é dada por

$$f_Y(y) = \frac{1}{\sqrt{2\pi}} e^{-(-\sqrt{y})^2/2}\left|-\frac{1}{2\sqrt{y}}\right| + \frac{1}{\sqrt{2\pi}} e^{-(\sqrt{y})^2/2}\left|\frac{1}{2\sqrt{y}}\right|,$$

$$f_Y(y) = \frac{1}{\sqrt{2\pi}}\frac{1}{\sqrt{y}} e^{-y/2}, \quad 0 < y < \infty.$$

Comparando-se a função de densidade obtida acima à função de densidade para uma distribuição qui-quadrada, notamos que essa função de densidade acima corresponde a uma distribuição qui-quadrada com um grau de liberdade, que é o mesmo que uma distribuição gamma com parâmetros $\alpha = 1/2$ e $\beta = 2$.

O Teorema 3.1 cobre os casos de variáveis aleatórias contínuas, onde as funções $g(\cdot)$'s têm apenas um argumento escalar. Os exercícios ao final do capítulo apresentam uma variedade de exemplos adicionais para fixação dos conceitos. No entanto, esse teorema não cobre a situação referente à soma de várias variáveis de Poisson ou a situação do quociente entre duas variáveis aleatórias diferentes. Esses casos serão cobertos quando tratarmos de distribuição conjuntas no Capítulo 4.

3.6 Exercícios

Nesta seção, apresentamos alguns exercícios de fixação. O leitor é encorajado a recorrer a outras referências caso haja necessidade. A maioria dos exercícios envolve princípios básicos de integração e de manipulação algébrica.

Exercício 3.1 Para a variável aleatória gamma

$$f(y) = \frac{1}{\Gamma(\alpha)\beta^{\alpha}} y^{\alpha-1} e^{-y/\beta}, \ y > 0,$$

a Weibull

$$f(y) = \beta\lambda y^{\beta-1} e^{-\lambda y^{\beta}}, \ y > 0,$$

a exponencial negativa

$$f(y) = \lambda e^{-\lambda y}, \ y > 0,$$

a beta

$$f(y) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}, \ y > 0,$$

e a normal

$$f(y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(y-\mu)^2}, \ y \in \Re,$$

mostre que $f(y)$ é uma função de densidade.

Dica: (1) Utilize a definição da função gamma nos casos acima em que ela aparece:

$$\Gamma(a) = \int_0^{\infty} x^{a-1} e^{-x} dx,$$

e as propriedades

$$\Gamma(z) = (z-1)!, \text{ quando } z \text{ é um número inteiro}$$
$$\Gamma(a+1) = a\Gamma(a)$$
$$\Gamma(1/2) = \sqrt{\pi}.$$

Dica: (2) Para a variável aleatória beta, utilize a definição de função beta vista na Seção 3.3.10.

Exercício 3.2 Para cada variável aleatória a seguir, determine o valor esperado (primeiro momento), o segundo momento e o segundo momento centrado (variância):
(1) N(μ, σ^2);
(2) Exponencial negativa, com parâmetro λ;
(3) Gamma(α, β);
(4) Qui-quadrada com r graus de liberdade;
(5) Beta.
(6) Uniforme $U(a, b)$. Nesse caso, a função de densidade de probabilidade é dada por

$$\begin{aligned} f(y) &= \frac{1}{b-a}, \text{ se } y \in [a, b], \\ &= 0, \text{ caso contrário.} \end{aligned}$$

Exercício 3.3 Para as variáveis aleatórias no Exercício 3.2, determine a função geratriz de momentos, definida como

$$E[e^{sY}] = \int_{y=-\infty}^{\infty} e^{sy} f(y) dy,$$

onde s é um número real, em uma vizinhança em torno do ponto 0.

Exercício 3.4 Seja X uma variável aleatória com função de densidade

$$\begin{aligned} f(x) &= 2(1-x), \text{ para } 0 < x < 1 \\ &= 0, \text{ caso contrário.} \end{aligned} \tag{3.56}$$

Determine

(1) E$[X]$;
(2) Variância de X;
(3) Segundo momento não centrado E$[X^2]$.

Exercício 3.5 Seja Y uma variável aleatória com distribuição exponencial negativa, com média $\mu = 10$. Determine as probabilidades a seguir:
(1) Prob[Y $>$ 10];
(2) Prob[2 $<$ Y $<$ 15];
(3) Prob[Y $<$ 2 ou Y $>$ 20].

Exercício 3.6 Escreva o terceiro momento centrado E$[(X-\mu)^3]$ como uma função dos três primeiros momentos não centrados.

Exercício 3.7 Seja X uma variável aleatória com distribuição binomial, com parâmetros n (número de tentativas) e p (probabilidade de sucesso). Determine

(1) O valor esperado $\mathrm{E}[X]$;

(2) A variância $\mathrm{E}[(X-\mu)^2]$.

Exercício 3.8 Para uma variável aleatória de Poisson, mostre que a função

$$f(x) = \frac{e^{-\lambda}\lambda^x}{x!}$$

é uma função de frequência.

Exercício 3.9 Seja X uma variável aleatória com função de densidade

$$\begin{aligned} f(x) &= \frac{3}{8}x^2, \text{ para } 0 < x < 2 \\ &= 0, \text{ caso contrário.} \end{aligned} \tag{3.57}$$

Determine as probabilidades

(1) $\mathrm{P}[0 < \mathrm{X} < 1]$

(2) $\mathrm{P}[1 < \mathrm{X} < 2]$

(3) $\mathrm{P}[0 < \mathrm{X} < 1/2 \text{ ou } 1 < \mathrm{X} < 2]$

Exercício 3.10 Seja X uma variável aleatória com função de densidade

$$\begin{aligned} f(x) &= \frac{A}{x^3}, \text{ se } 1 < x < \infty \\ &= 0, \text{ caso contrário.} \end{aligned} \tag{3.58}$$

Determine

(1) O valor de A para que a função $f(\cdot)$ seja uma função de densidade;

(2) a função de distribuição acumulada $F(x)$ de X.

Exercício 3.11 Escreva as funções de distribuição acumulada $F(x)$ para as seguintes variáveis aleatórias

(1) Exponencial negativa;

(2) Weibull.

Exercício 3.12 Seja X uma variável aleatória com distribuição de Poisson, com parâmetro $\lambda = 10$. Encontre

(1) $\mathrm{P}[3 < \mathrm{X} < 11]$;

(2) A variância de X;

(3) A função geratriz de momentos (vide Exercício 3.3).

Exercício 3.13 Considere uma variável aleatória X de Poisson, com média λ. Calcule:
(1) O segundo momento não-centrado $E[X^2]$;
(2) O terceiro momento não-centrado $E[X^3]$;
(3) A variância de X.

Exercício 3.14 A variável aleatória geométrica modela o número de insucessos até a ocorrência do primeiro sucesso, em sucessivos experimentos independentes de Bernoulli, com probabilidade de sucesso $p \in (0,1)$. Para a variável geométrica, responda:
(1) Mostre que a função de frequência é dada por

$$f(x) = p \times (1-p)^x, \quad x = 0, 1, 2, ...$$

(2) Calcule a média da variável aleatória geométrica;
(3) Calcule a variância da variável aleatória geométrica.

Exercício 3.15 Alguns autores usam a variável aleatória geométrica para modelar o número de eventos de Bernoulli (incluindo insucessos e o sucesso) até se obter o primeiro sucesso. Essa definição é ligeiramente diferente da definição no Exercício 3.14. Responda novamente os itens de (1) a (3) no Exercício 3.14, considerando essa nova definição.

Exercício 3.16 A variável aleatória binomial negativa modela o número de insucessos, em sucessivos experimentos de Bernoulli, antes de obtermos o r-ésimo sucesso, onde r é um número inteiro positivo. A probabilidade de sucesso em cada experimento de Bernoulli é $p \in (0,1)$. Responda:
(1) Mostre que a função de frequência é dada por

$$f(x) = \binom{r+x-1}{x} \times p^r \times (1-p)^x, \quad x = 0, 1, 2, ...$$

(2) Calcule a média da variável binomial negativa;
(3) Calcula a variância da variável binomial negativa;
(4) Mostre que a variável aleatória geométrica é um caso particular da variável aleatória binomial negativa.

Exercício 3.17 Considere uma variável aleatória X com distribuição normal com média $\mu = 10$ e variância $\sigma^2 = 16$. Responda:
(1) Qual a probabilidade de X ser menor que 8?
(2) Qual a probabilidade de X estar entre 5 e 12?
(3) Ache o valor y para o qual $X < y$ com probabilidade de 5%.

Exercício 3.18 Considere uma variável aleatória Z normal padronizada. Determine:
(1) Qual o valor $t_{95\%}$ para o qual $\text{Prob}[|Z| < t_{95\%}] = 0.95$?
(2) Qual o valor $t_{99\%}$ para o qual $\text{Prob}[|Z| < t_{99\%}] = 0.99$?
(3) Qual o valor $t_{90\%}$ para o qual $\text{Prob}[|Z| < t_{90\%}] = 0.90$?

Exercício 3.19 Utilizando a desigualdade de Cauchy-Schwarz, mostre que o coeficiente de correlação está sempre entre -1 e 1.

Exercício 3.20 Seja X uma variável aleatória discreta em $\{1, 2, 3, \dots\}$, com função de frequência $f(x) = K/x^3$.
(1) Encontre o valor de K para que o somatório de $f(x)$ seja igual a um.
(2) Qual o valor do primeiro momento de X, caso ele exista?
(3) Determine o valor de r, para o qual os momentos de ordem r e acima não mais existem.

Exercício 3.21 Mostre que, se uma variável aleatória X tem momento não centrado de ordem r, então ela também tem todos os momentos, centrados e não centrados, de ordem q, com $1 \leq q \leq r$.

Dica: Use a desigualdade de Holder.

Exercício 3.22 Seja X uma variável aleatória normal, com média μ e variância σ^2. Responda:
(1) Seja $Y = \exp(X)$. Qual a distribuição de Y?
(2) Seja $W = a + Xb$. Qual a distribuição de W?
(3) Mostre que a variável $Z = \left[(X - \mu)/\sigma\right]$ tem distribuição normal padronizada.

Exercício 3.23 Seja X uma variável aleatória exponencial negativa com parâmetro λ. Seja $Y = a + bX$. Qual a distribuição de Y?

Exercício 3.24 Seja X uma variável aleatória com distribuição qui, com parâmetro ν. Qual a distribuição de $Y = X^2$?

Exercício 3.25 Seja X uma variável aleatória com distribuição gamma, com parâmetros α e β. Qual a distribuição da variável aleatória $Y = a + bX$?

4. Distribuições conjuntas

"The more things are forbidden, the more popular they become."
Mark Twain

No Capítulo 3, estudamos os casos onde a variável aleatória de interesse X, por exemplo, tinha apenas uma dimensão; ou seja, $X \in \Re$. Neste capítulo, vamos estender o conceito de variável aleatória para o caso onde temos várias variáveis de interesse, e queremos estudar não somente cada variável individualmente, mas também todas as variáveis conjuntamente. Por exemplo, o interesse de pesquisa pode estar justamente focado na interrelação entre as variáveis aleatórias, ou como uma determinada variável pode ser prevista a partir de valores conhecidos das demais variáveis. Da mesma maneira que nos capítulos anteriores, iremos definir algumas características que nos ajudarão a melhor descrever cada variável aleatória ou a relação entre elas.

Seja então X uma vetor de variáveis aleatórias pertencente ao $\Re^K$. Ou seja, X é um vetor composto por K variáveis aleatórias reais individuais, $X = [X_1, \ldots, X_K]^T$. Portanto, trabalharemos neste livro com vetores aleatórios, de dimensão $K \times 1$; ou seja, trabalharemos com vetores coluna. O símbolo $(\cdot)^T$ corresponde à matriz transposta ou ao vetor transposto do argumento da operação de transposição. Por exemplo, se A é uma matriz $m \times n$, então A^T corresponde à matriz transposta, com dimensão $n \times m$. Se v é um vetor linha (dimensão $1 \times n$), então v^T é o vetor coluna transposto, com dimensão $n \times 1$. Alguns autores utilizam também o símbolo A' ao invés de A^T. Neste livro, usaremos A' ou A^T indistintamente. O vetor X pode ser denominado também uma variável aleatória multivariada.

Da mesma forma que no Capítulo 3, o nosso objetivo aqui é fornecer a intuição por detrás dos modelos probabilísticos multivariados e não nos entreter com aspectos técnicos. O leitor interessado nas provas das proposições desse capítulo pode consultar as mesmas referências apresentadas na Seção 3.1, que versa sobre variáveis aleatórias.

4.1 Funções de distribuição conjunta

A **função de distribuição acumulada conjunta** de uma variável aleatória multivariada $X \in \Re^K$ é dada por

$$F(x_1, x_2, \ldots, x_K) = \text{Prob}\big[X_1 \leq x_1, X_2 \leq x_2, \ldots, X_K \leq x_K\big], \tag{4.1}$$

para pontos $x_1, \ldots, x_K$ pertencentes cada qual ao conjunto $\Re$. Portanto, a função de distribuição acumulada conjunta $F(x_1, \ldots, x_K)$ é uma função de $\Re^K$ em $\Re$. Note que, quando o vetor X possui apenas um componente, a função $F(\cdot)$ transforma-se na função de distribuição acumulada vista no Capítulo 3. Da

mesma maneira que para variáveis aleatórias univariadas, a função de distribuição acumulada tem a mesma expressão, independentemente de os componentes da variável aleatória serem discretos ou contínuos. A Eq. (4.1) vale para todo tipo de variável aleatória conjunta.

Para variáveis aleatórias discretas, pode-se estender a definição de função de frequência no caso univariado para a função de frequência no caso multivariado que chamamos de **função de frequência conjunta**. A função de frequência para uma variável aleatória $X \in \Re^K$ é dada por

$$f(x_1, x_2, \ldots, x_K) = \text{Prob}\big[X_1 = x_1, X_2 = x_2, \ldots, X_K = x_K\big], \tag{4.2}$$

onde a vírgula dentro da probabilidade acima significa que todos as igualdades devem acontecer ao mesmo tempo.

Exemplo 4.1 (Função de frequência conjunta e função de distribuição acumulada conjunta de uma vetor aleatório discreto bivariado) Considere um vetor discreto bivariado, composto pelas variáveis individuais X e Y. A função de frequência conjunta tem a seguinte especificação

$$\begin{aligned} f_{X,Y}(1,1) &= 0.20, \quad f_{X,Y}(1,2) = 0.05, \\ f_{X,Y}(2,1) &= 0.15, \quad f_{X,Y}(2,2) = 0.15, \\ f_{X,Y}(3,1) &= 0.25, \quad f_{X,Y}(3,2) = 0.20. \end{aligned}$$

O espaço amostral, que corresponde ao conjunto de todos os valores possíveis de ocorrerem, é um conjunto de pares ordenados, e é dado por $\mathbb{X} = \{(1,1), (1,2), (2,1), (2,2), (3,1), (3,2)\}$. Note que $\mathbb{X} \subset \Re^2$. Podemos então calcular a função de distribuição acumulada para os pontos do espaço amostra. Aplicando a definição de acordo com a Eq. (4.1), obtém-se

$$\begin{aligned} F_{X,Y}(1,1) &= 0.20, \quad F_{X,Y}(1,2) = 0.25, \\ F_{X,Y}(2,1) &= 0.35, \quad F_{X,Y}(2,2) = 0.55, \\ F_{X,Y}(3,1) &= 0.60, \quad F_{X,Y}(3,2) = 1.00. \end{aligned}$$

Para pontos (x, y) não contidos no espaço amostra $\mathbb{X}$, o valor da função de distribuição acumulada $F(\cdot)$ também é definido. Basta utilizar a definição na Eq. (4.1) para concluir que $F_{X,Y}(1.2, 2.149) = \text{Prob}\big[X \leq 1.2,\ Y \leq 2.149\big] = 0.25$. Analogamente, $F_{X,Y}(5, 2.2) = 1$, e $F_{X,Y}(x, y) = 0.35$ para qualquer ponto (x, y) com $x \in [2, 3)$ e $y \in [1, 2)$.

Observe que no Exemplo 4.1 utilizamos a notação $f_{X,Y}$ e $F_{X,Y}$ para deixar claro que as funções frequência conjunta e distribuição acumulada conjunta referem-se à variável aleatória bivariada, composta pelas variáveis aleatórias individuais X e Y. Da mesma maneira que no caso de variáveis aleatórias univariadas, a

função de distribuição acumulada e a função de frequência satisfazem determinadas propriedades, conforme especificado a seguir.

Proposição 4.1 (Propriedades da função de distribuição acumulada conjunta) Seja $F(\cdot)$ uma função $F : \Re^K \mapsto \Re$ qualquer; ou seja, $F(\cdot)$ é uma função definida no $\Re^K$, e assume valores reais. F será uma função de distribuição acumulada conjunta se e somente se

(i) $F(x_1, \cdots, x_K) = 1$, quando $x_i \to +\infty$, para todo $i = 1, \ldots, K$.
(ii) $F(x_1, \ldots, x_{i-1}, x_i, x_{i+1}, \ldots, x_K) = 0$, quando $x_i \to -\infty$, para qualquer $i = 1, \ldots, K$. Portanto, fixando todos os pontos x_j's, exceto x_i, o qual se supõe convergindo para $-\infty$, o valor da função $F(\cdot)$ converge para o valor 0.
(iii) $F(x_1, \ldots, x_K) \in [0, 1]$, para todo $[x_1, \ldots, x_K]^T \in \Re^K$.
(iv) $F(x_1, \ldots, x_K)$ é contínua à direita. Portanto,

$$\lim_{\Delta \to 0^+} F(x_1, \cdots, x_{i-1}, x_i + \Delta, x_{i+1}, \ldots, x_K) = F(x_1, \cdots, x_{i-1}, x_i, x_{i+1}, \ldots, x_K),$$

para todo $i = 1, \ldots, K$.

As propriedades acima valem tanto para variáveis aleatórias discretas quanto variáveis aleatórias contínuas. As propriedades abaixo correspondem especificamente ao caso de variáveis aleatórias discretas. O caso contínuo será visto mais adiante.

Proposição 4.2 (Propriedades da função de frequência conjunta) Seja $f(\cdot)$ uma função $f : \Re^K \mapsto \Re$ qualquer; ou seja, f é uma função definida no $\Re^K$, e assume valores reais. f será uma função de frequência conjunta se e somente se

(i) $\sum_{[x_1, \ldots, x_K]^T \in \mathbb{X}} f(x_1, \ldots, x_K) = 1$.
(ii) $f(x_1, \ldots, x_K) \geq 0$, para todo $[x_1, \ldots, x_K]^T \in \Re^K$.

O somatório na propriedade (i) é calculado para todos os pontos no espaço amostra $\mathbb{X}$. Observe que as funções $f_{X,Y}$ e $F_{X,Y}$ no Exemplo 4.1 acima satisfazem às propriedades para a função de frequência conjunta e para a função de distribuição acumulada conjunta.

A função de frequência conjunta e a função de distribuição acumulada conjunta fornecem, entre outras coisas, a forma como os componentes individuais se interrelacionam. No entanto, o pesquisador pode estar interessado em extrair as características específicas de um dos componentes individualmente, sem se preocupar com os demais componentes. Por exemplo, no Exemplo 4.1, imagine que o nosso interesse seja saber qual a probabilidade de o componente individual X assumir valor 2; ou seja, gostaríamos de saber o valor de $\text{Prob}[X = 2]$. Como encontrar esse valor específico para a variável X se apenas temos, de acordo com o exemplo, a função de frequência para as duas variáveis X e Y conjuntamente? Para responder a essa pergunta, temos que encontrar uma função conhecida com **função de frequência marginal** $f_X(x)$ para a variável aleatória X.

Exemplo 4.2 (Continuação do Exemplo 4.1 - Função de frequência marginal de um vetor aleatório discreto bivariado) Para o Exemplo 4.1, a função de frequência marginal pode ser obtida diretamente da função de frequência conjunta, de acordo com as seguintes expressões:

$$f_X(1) = \text{Prob}[X = 1] = \text{Prob}[X = 1,\ Y = 1] + \text{Prob}[X = 1,\ Y = 2] = 0.20 + 0.05 = 0.25,$$
$$f_X(2) = \text{Prob}[X = 2] = \text{Prob}[X = 2,\ Y = 1] + \text{Prob}[X = 2,\ Y = 2] = 0.15 + 0.15 = 0.30,$$
$$f_X(3) = \text{Prob}[X = 3] = \text{Prob}[X = 3,\ Y = 1] + \text{Prob}[X = 3,\ Y = 2] = 0.25 + 0.20 = 0.45.$$

Portanto,

$$f_X(1) = f_{X,Y}(1,1) + f_{X,Y}(1,2) = \sum_{y \in \mathbb{X}_Y} f_{X,Y}(1,y),$$
$$f_X(2) = f_{X,Y}(2,1) + f_{X,Y}(2,2) = \sum_{y \in \mathbb{X}_Y} f_{X,Y}(2,y),$$
$$f_X(3) = f_{X,Y}(3,1) + f_{X,Y}(3,2) = \sum_{y \in \mathbb{X}_Y} f_{X,Y}(3,y),$$

onde $\mathbb{X}_Y = \{1,2\}$ é o conjunto de valores possíveis para a variável aleatória Y individualmente. Analogamente, a função de frequência marginal $f_Y(y)$ para a variável aleatória Y é dada por

$$f_Y(1) = f_{X,Y}(1,1) + f_{X,Y}(2,1) + f_{X,Y}(3,1) = \sum_{x \in \mathbb{X}_X} f_{X,Y}(x,1) = 0.60,$$
$$f_Y(2) = f_{X,Y}(1,2) + f_{X,Y}(2,2) + f_{X,Y}(3,2) = \sum_{x \in \mathbb{X}_X} f_{X,Y}(x,2) = 0.40,$$

onde $\mathbb{X}_X = \{1,2,3\}$ é o conjunto de valores possíveis para a variável aleatória X individualmente. Observe as notações $f_Y(y)$ e $f_X(x)$ para diferenciar as funções de frequência marginais da função de frequência conjunta $f_{X,Y}(x,y)$. Os conjuntos $\mathbb{X}_X$ e $\mathbb{X}_Y$ são os espaços amostrais das distribuições marginais para as variáveis aleatórias X e Y individualmente. As funções de frequência marginais são funções de frequência tradicionais, no sentido visto no Capítulo 3. Portanto, os seus valores possuem soma 1 e elas são sempre não negativas.

De maneira geral, seja $X = [X_1, \ldots, X_K]^T \in \Re^K$ um vetor aleatório multivariado, com função de frequência conjunta $f_{X_1,\ldots,X_K}(x_1 \ldots, x_K)$. Para uma variável aleatória específica X_i, com $i \in \{1, \ldots, K\}$, a função de frequência marginal $f_{X_i}(x_i)$ pode ser obtida por meio da expressão

$$f_{X_i}(x_i) = \sum_{x_1 \in \mathbb{X}_{X_1}, \ldots, x_{i-1} \in \mathbb{X}_{X_{i-1}}, x_{i+1} \in \mathbb{X}_{X_{i+1}}, \ldots, x_K \in \mathbb{X}_{X_K}} f_{X_1,\ldots,X_K}(x_1, \ldots, x_K), \tag{4.3}$$

onde $\mathbb{X}_{X_j}$ é o conjunto de valores possíveis (ou o espaço amostral) para a variável aleatória individual X_j. A distribuição de frequência marginal não necessariamente corresponde a uma função de frequência de uma variável univaridada. De fato, para um conjunto de $K > 2$ variáveis aleatórias, compondo o vetor aleatório X, podemos estar interessados na função de frequência marginal para o vetor de variáveis aleatórias composto apenas pelas variáveis aleatórias X_1 e X_2. Nesse caso, para encontrar a função de frequência marginal $f_{X_1,X_2}(x_1, x_2)$ (que também é uma função de frequência conjunta) a partir da função de frequência conjunta $f_{X_1,\ldots,X_K}(x_1, \ldots, x_K)$, basta fazer, para cada par de valores (x_1, x_2), o somatório sobre todos os valores possíveis para as demais variáveis que compõem o vetor X, similarmente ao que é feito na Eq. (4.3).

Repetindo o mesmo procedimento usado para calcular a função de frequência marginal a partir da função de frequência conjunta, podemos calcular a **função de distribuição acumulada marginal** a partir da função de distribuição acumulada:

Exemplo 4.3 (Continuação do Exemplo 4.1 – Função de distribuição acumulada marginal de um vetor aleatório discreto bivariado). Para o Exemplo 4.1 acima, vamos encontrar as funções de distribuição acumulada marginais $F_X(x)$ e $F_Y(y)$ para as variáveis aleatórias X e Y. Para isso, basta proceder tratando as variáveis X e Y individualmente, como se elas não pertencessem a um vetor de variáveis aleatórias. Temos então

$$
\begin{aligned}
F_X(x) &= 0, \text{ para } x < 1, \\
F_X(x) &= 0.25, \text{ para } x \in [1, 2), \\
F_X(x) &= 0.55, \text{ para } x \in [2, 3), \\
F_X(x) &= 1.00, \text{ para } x \geq 3.
\end{aligned}
$$

Para a variável aleatória Y, a função de distribuição acumulada marginal é dada por

$$
\begin{aligned}
F_Y(y) &= 0, \text{ para } y < 1, \\
F_Y(y) &= 0.60, \text{ para } y \in [1, 2), \\
F_Y(y) &= 1.00, \text{ para } y \geq 2.
\end{aligned}
$$

A função de distribuição acumulada marginal é uma função de distribuição acumulada, valendo todas as propriedades vistas no Capítulo 3.

Para variáveis aleatórias contínuas multivariadas, similarmente ao caso de variáveis contínuas univariadas, a função de frequência é substituída pela **função de densidade de probabilidade conjunta** $f_{X_1,X_2,\ldots,X_K}(x_1, \ldots, x_K)$. Essa função não corresponde à probabilidade de um vetor aleatório contínuo assumir um determinado valor, mas ela pode ser utilizada para calcularmos a probabilidade de o vetor aleatório assumir valores em um determinado intervalo. A função de distribuição acumulada conjunta pode

ser obtida a partir da função de densidade de probabilidade conjunta, via integração

$$F_{X_1,\ldots,X_K}(x_1,\ldots,x_K) = \int_{u_1=-\infty}^{x_1} \cdots \int_{u_K=-\infty}^{x_K} f_{X_1,\ldots,X_K}(u_1,\ldots,u_K)du_K \ldots du_1. \tag{4.4}$$

Portanto, a expressão acima estende a relação entre a função de densidade e a função de distribuição acumulada para o caso multivariado; ao invés de uma integração simples, temos que efetuar uma integração de volume. A Figura 4.1 ilustra uma função de densidade conjunta, para o caso de um vetor de duas variáveis aleatórias (vetor aleatório bivariado). Essa função de densidade conjunta corresponde a uma variável aleatória normal bivariada, conforme discussão mais adiante. O exemplo abaixo ilustra numericamente a função de densidade conjunta, para o caso de um caso simples de um vetor aleatório contínuo bivariado.

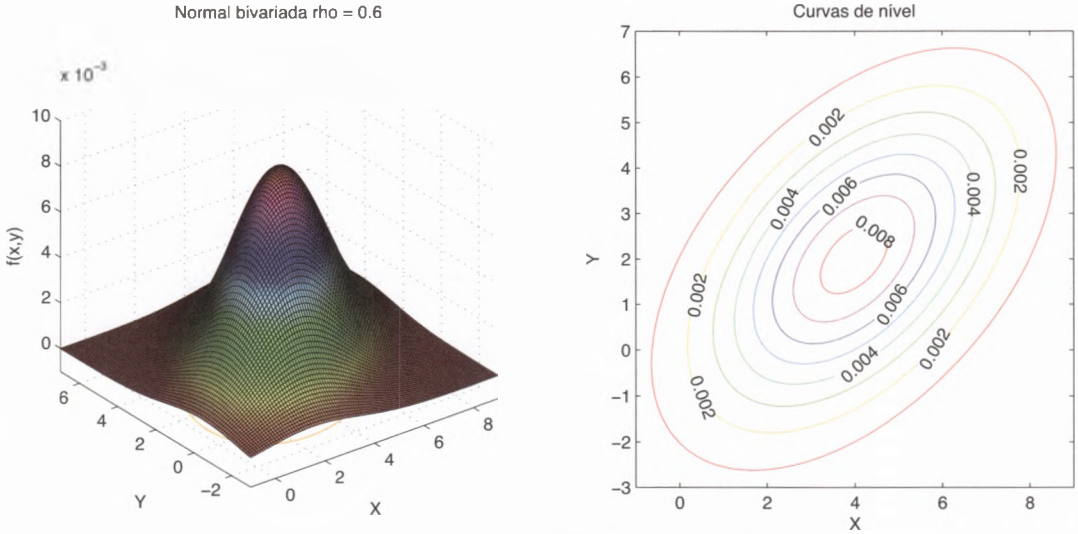


Figura 4.1: Função de densidade para um vetor normal bivariado, com coeficiente de correlação $\rho = 0.6$.

Exemplo 4.4 (Função de densidade de probabilidade conjunta e função de distribuição acumulada conjunta de um vetor aleatório contínuo bivariado) Seja um vetor aleatório bivariado, formado pelos componentes individuais X e Y, com função de densidade de probabilidade conjunta

$$f_{X,Y}(x,y) = \frac{3}{2}(x + y - 2xy^2), \text{ quando } 0 < x < 1 \text{ e } 0 < y < 1,$$
$$f_{X,Y}(x,y) = 0, \text{ caso contrário.}$$

Podemos calcular a função de distribuição acumulada por meio da integral dupla

$$F_{X,Y}(x,y) = \int_{u=-\infty}^{x} \int_{v=-\infty}^{y} f_{X,Y}(u,v)dvdu.$$

Para $x \leq 0$ ou $y \leq 0$, temos $F_{X,Y}(x,y) = 0$. Para $x \geq 1$ e $y \geq 1$, a função de distribuição acumulada assume valor 1. Quando $0 < x < 1$ ou $0 < y < 1$, podemos integrar a função de densidade de probabilidade

conjunta obtendo

$$F_{X,Y}(x,y) = \int_{u=-\infty}^{x} \int_{v=-\infty}^{y} f_{X,Y}(u,v)dvdu = \frac{3}{4}xy(x + y - \frac{2}{3}xy^2).$$

Proposição 4.3 (Propriedades da função de densidade de probabilidade conjunta) Seja $f(\cdot)$ uma função $f : \Re^K \mapsto \Re$ qualquer; ou seja, f é uma função definida no $\Re^K$, e assume valores reais. f será uma função de densidade de probabilidade conjunta se e somente se

(i) $\int_{u_1=-\infty}^{+\infty} \cdots \int_{u_K=-\infty}^{+\infty} f_{X_1,\ldots,X_K}(u_1,\ldots,u_K)du_K \ldots du_1 = 1.$
(ii) $f(x_1,\ldots,x_K) \geq 0$, para todo $[x_1,\ldots,x_K]^T \in \Re^K$.

Observe que a função $f_{X,Y}$ no Exemplo 4.4 acima satisfaz às propriedades para a função de densidade de probabilidade conjunta.

Da mesma maneira que no caso das variáveis aleatórias discretas conjuntas, o pesquisador pode estar interessado em estudar o comportamento de uma das variáveis individualmente no vetor de variáveis aleatórias por meio da **função de densidade de probabilidade marginal** ou da **função de distribuição acumulada marginal**, como mostra o exemplo abaixo:

Exemplo 4.5 (Continuação do Exemplo 4.4 – Funções de densidade de probabilidade marginal e de distribuição acumulada marginal de um vetor aleatório contínuo bivariado) Por exemplo, no caso do vetor bivariado do Exemplo 4.4, supondo que queremos estudar o comportamento da variável aleatória X individualmente, precisamos calcular a função de densidade de probabilidade marginal para a variável aleatória X. Essa variável pode ser calculada integrando-se a dimensão y na função $f_{X,Y}(x,y)$ ao longo de toda a reta real. Portanto, a função de densidade de probabilidade marginal para X tem expressão

$$f_X(x) = \int_{y=-\infty}^{\infty} f_{X,Y}(x,y)dy = \frac{3}{4} + \frac{x}{2}, \text{ para } 0 < x < 1,$$
$$f_X(x) = 0, \text{ caso contrário.}$$

Analogamente, a função de densidade marginal para a variável Y tem expressão

$$f_Y(y) = \int_{x=-\infty}^{\infty} f_{X,Y}(x,y)dx = \frac{3}{4} + \frac{3}{2}y - \frac{3}{2}y^2, \text{ para } 0 < y < 1,$$
$$f_Y(y) = 0, \text{ caso contrário.}$$

Finalmente, assim como no caso discreto, podemos calcular a função de distribuição acumulada marginal para o caso contínuo. Nesse caso, fazemos $F_X = F_{X,Y}(x,1) = \frac{3}{4}x(1 + x - \frac{2}{3}x)$ e $F_Y = F_{X,Y}(1,y) = \frac{3}{4}y(1 + y - \frac{2}{3}y^2)$.

4.2 Distribuições condicionais

Em aplicações de econometria e estatística, é muito comum o pesquisador estar interessado em como uma determinada variável se comporta quando se alteram os valores de outras variáveis no sistema. O interesse pode estar em entender como valores de variáveis preditoras podem adicionar informações sobre o valor de uma variável de interesse desconhecida. Por exemplo, imagine um vetor de variáveis aleatórias composto por três variáveis individuais: renda do domicílio, escolaridade do chefe de família, idade do chefe de família. O interesse pode estar em conseguir identificar qual o valor médio de renda domiciliar para uma família que tem chefe com nível superior e idade acima dos 50 anos. Ou seja, condicionado em famílias com chefe acima de cinquenta anos e com nível superior, qual a renda média desses domicílios. Apesar de esse exemplo ser bastante intuitivo para a maioria dos pesquisadores em ciências sociais, ele está mais relacionado aos conceitos estatísticos de distribuição conjunta do que imaginamos. De fato, quando falamos em valores de uma determinada variável "dado"os valores de variáveis preditoras, estamos implicitamente nos referindo à ideia de **distribuições condicionais**.

Considere uma variável aleatória contínua multivariada, com elementos X e Y, e com função de densidade conjunta $f_{X,Y}(x,y)$. A **função de densidade condicional** de Y, condicional a um valor x de X, é dada por

$$f_{Y/X}(y/x) = \frac{f_{X,Y}(x,y)}{f_X(x)}, \tag{4.5}$$

supondo que $f_X(x) > 0$, onde $f_X(x)$ é a função de densidade marginal de X. Observe a notação $f_{Y/X}(y/x)$ para denotar condicionalidade. Note que a função de densidade condicional é também uma função de densidade, no sentido de que

(i) $f_{Y/X}(y/x) \geq 0$,
(ii) $\int_{y=-\infty}^{+\infty} f_{Y/X}(y/x)dy = \frac{\int_{y=-\infty}^{+\infty} f_{Y,X}(y,x)dy}{f_X(x)} = \frac{f_X(x)}{f_X(x)} = 1.$

Exemplo 4.6 (Continuação do Exemplo 4.4 - Função de densidade condicional de um vetor aleatório contínuo bivariado) Para o Exemplo 4.4 acima, podemos calcular a função de densidade condicional $f_{Y/X}(y/x)$ para todos os valores de x em $(0,1)$ (para valores de x fora desse intervalo, $f_X(x)$ assume valor zero, e não podemos definir a função de densidade condicional para Y dado X). De fato,

$$f_{Y/X}(y/x) = \frac{\frac{3}{2}(x + y - 2xy^2)}{\frac{3}{4} + \frac{x}{2}}, \text{ para } x \in (0,1),\ y\ (0,1).$$

Para $x \in (0,1)$, e $y \geq 1$ ou $y \leq 0$, $f_{Y/X}(y/x) = 0$. Analogamente, podemos calcular a função de densidade condicional de X dado valores para Y.

$$f_{X/Y}(x/y) = \frac{\frac{3}{2}(x + y - 2xy^2)}{\frac{3}{4} + \frac{3}{2}y - \frac{3}{2}y^2}, \text{ para } x \in (0,1),\ y\ (0,1).$$

Para $y \in (0,1)$, e $x \geq 1$ ou $x \leq 0$, $f_{X/Y}(x/y) = 0$. Quando $y \geq 1$ or $y \leq 0$ a função de densidade condicional $f_{X/Y}(x/y)$ não está definida, já que $f_Y(y) = 0$ quando y assume valores fora do intervalo aberto $(0,1)$.

Para variáveis discretas, podemos definir a **função de frequência condicional**, similarmente ao que foi feito no caso contínuo. Seja uma variável aleatória discreta bivariada com componentes X e Y. Gostaríamos de estudar o comportamento da variável discreta Y condicionada a determinados valores x da variável X. A função de frequência condicional vai ter expressão

$$f_{Y/X}(y/x) = \frac{f_{X,Y}(x,y)}{f_X(x)}, \tag{4.6}$$

supondo que $f_X(x) > 0$.

Exemplo 4.7 (Continuação do Exemplo 4.1 – Função de frequência condicional) No Exemplo 4.1, a função de frequência de Y condicionada a valores de X pode ser calculada da forma

$$\begin{aligned} f_{Y/X}(1/1) &= 0.20/0.25, \quad f_{Y/X}(2/1) = 0.05/0.25, \\ f_{Y/X}(1/2) &= 0.15/0.30, \quad f_{Y/X}(2/2) = 0.15/0.30, \\ f_{Y/X}(1/3) &= 0.25/0.45, \quad f_{Y/X}(2/3) = 0.20/0.45. \end{aligned}$$

Para valores de x fora do espaço amostral $\mathbb{X}_X = \{1,2,3\}$, a função $f_{Y/X}(y/x)$ não é definida, já que a função de frequência marginal de X possui valor nulo. A função de frequência condicional da variável aleatória X condicionada a valores da variável aleatória Y é dada pelas expressões

$$\begin{aligned} f_{X/Y}(1/1) &= 0.20/0.60, \quad f_{X/Y}(1/2) = 0.05/0.40, \\ f_{X/Y}(2/1) &= 0.15/0.60, \quad f_{X/Y}(2/2) = 0.15/0.40, \\ f_{X/Y}(3/1) &= 0.25/0.60, \quad f_{X/Y}(3/2) = 0.20/0.40. \end{aligned}$$

Para valores de y fora do espaço amostral $\mathbb{X}_Y = \{1,2\}$, a função $f_{X/Y}(x/y)$ não é definida, já que a função de frequência marginal de Y possui valor nulo.

A função de frequência condicional é uma função de frequência, conforme vimos no Capítulo 3. De fato, a função de frequência condicional tem soma 1 ao longo de todos os elementos do espaço amostral e é sempre não negativa.

Aplicação 4.1 (Teoria estatística da decisão) Considere que um investidor esteja interessado em investir uma parte não aplicada de seus fundos no mercado e precise decidir entre três possíveis ações a_1, a_2 e a_3, conforme apresentado na Tabela 4.1.

Tabela 4.1: Ações disponíveis para o investidor.

a_1	Investir na bolsa de valores (mercado de alto risco)
a_2	Investir em títulos da previdência (mercado de médio risco)
a_3	Investir em títulos do governo (mercado de baixo risco)

Obviamente a sua predisposição de investir em um desses mercados depende de como estará o mercado nos próximos períodos. Com o objetivo de simplificar a sua decisão, ele considera que existem dois estados possíveis y_1 e y_2 apresentados na Tabela 4.2.

Tabela 4.2: Estados possíveis do mercado.

Estado	Descrição	Probabilidade $P(y)$
y_1	Alta (mercado em alta)	0.52
y_2	baixa (mercado em baixa)	0.48

Esse investidor também associa uma perda $l(y, a)$ a cada uma de suas ações. Essa perda quantifica a escolha de uma ação em um determinado estado em termos de custos associados a essa decisão. Conforme apresentado na Tabela 4.3, assume-se que a perda associada a situação mais favorável é nula (CHERNOFF; MOSES, 1986) e todas as outras perdas são padronizadas em relação a esse valor.

Tabela 4.3: Perdas $l(y, a)$ associadas a cada estado y e a cada ação a.

	a_1 (bolsa de valores)	a_2 (previdência)	a_3 (títulos do governo)
y_1 (alta)	0	1	2
y_2 (baixa)	3	2	1

Infelizmente, não é possível prever o estado do próximo período. Entretanto, tendo como base notícias de jornais e *blogs* da área de economia e finanças, ele pode fazer uma estimativa desse estado. Ele considera três estimativas possíveis, apresentadas na Tabela 4.4.

Infelizmente, as estimativas dos estados não é perfeita e essa imperfeição é quantificada utilizando as probabilidades condicionais $f_{X/Y}(x/y)$ apresentadas na Tabela 4.5.

Uma vez que o estado y não é conhecido no momento da tomada de decisão, o processo de tomada decisão é feito tomando como base os eventos x. Nesse arcabouço, uma **estratégia pura** é um mapa que associa a cada evento x uma possível ação, isto é, uma estratégia pura especifica exatamente como o tomador de decisão irá proceder em caso de um evento x ocorrer. Como o número de eventos x é finito, o número de estratégias puras disponíveis também é finito e, precisamente, nesse caso é igual a $3^3 = 27$, como mostra a Tabela 4.6.

A perda esperada $L(y, s)$ associada a cada estratégia pura s em cada estado y pode ser calculada usando

$$L(y, s) = \sum_x l(y, s_x) f_{X/Y}(x/y),$$

Tabela 4.4: Possíveis estimativas dos estados.

x_1	Positiva (O mercado estará em alta)
x_2	Neutra (Não posso dizer nada)
x_3	Negativa (O mercado estará em baixa)

Tabela 4.5: Probabilidades condicionais $f_{X/Y}(x/y)$ para cada evento $X = x$ dado cada estado $Y = y$.

	x_1 (positiva)	x_2 (neutra)	x_3 (negativa)
y_1 (alta)	0.5	0.4	0.1
y_2 (baixa)	0.1	0.2	0.7

onde s_x é a ação usada pela estratégia s quando o evento x ocorre. Tabela 4.7 apresenta as perdas esperadas de cada estratégia em cada estado y.

Cada estratégia pura é representada na Figura 4.2 usando um sinal + a partir de seus valores de $L(y_1, s)$ (abcissa) e $L(y_2, s)$ (ordenada). Nessa figura, também podemos observar as perdas associadas a cada **estratégia mista**. Uma estratégia mista[1] associa cada estratégia pura a uma probabilidade. Dessa forma, cada estratégia mista supõe que o tomador de decisão vai selecionar uma estratégia pura aleatoriamente. O número de estratégias mistas é infinito e forma um conjunto convexo que é delimitado nessa figura.

Estratégias interessantes são aquelas conhecidas como admissíveis. Uma estratégia é dita ser **admissível** se não houver nenhuma outra estratégia (pura ou mista) disponível que possua perda menor que a dela em todos os estados. De acordo com a Figura 4.2 , as estratégias admissíveis são s_1, s_2, s_3, s_6, s_9, s_{18}, s_{27} e também as estratégias mistas que pertencem às linhas que ligam essas estratégias. Para se escolher entre essas chamadas estratégias admissíveis, precisa-se de algum critério adicional. Aqui, consideraremos apenas o critério de Bayes. Nesse critério, escolhemos a estratégia que minimiza

[1] Matematicamente, uma estratégia mista é uma combinação linear convexa de estratégias puras.

Tabela 4.6: Lista das estratégias puras disponíveis.

	s_1	s_2	s_3	s_4	s_5	s_6	s_7	s_8	s_9
x_1	a_1	a_1	a_1	a_1	a_1	a_1	a_1	a_1	a_1
x_2	a_1	a_1	a_1	a_2	a_2	a_2	a_3	a_3	a_3
x_3	a_1	a_2	a_3	a_1	a_2	a_3	a_1	a_2	a_3
	s_{10}	s_{11}	s_{12}	s_{13}	s_{14}	s_{15}	s_{16}	s_{17}	s_{18}
x_1	a_2	a_2	a_2	a_2	a_2	a_2	a_2	a_2	a_2
x_2	a_1	a_1	a_1	a_2	a_2	a_2	a_3	a_3	a_3
x_3	a_1	a_2	a_3	a_1	a_2	a_3	a_1	a_2	a_3
	s_{19}	s_{20}	s_{21}	s_{22}	s_{23}	s_{24}	s_{25}	s_{26}	s_{27}
x_1	a_3	a_3	a_3	a_3	a_3	a_3	a_3	a_3	a_3
x_2	a_1	a_1	a_1	a_2	a_2	a_2	a_3	a_3	a_3
x_3	a_1	a_2	a_3	a_1	a_2	a_3	a_1	a_2	a_3

Tabela 4.7: Perda esperadas $L(y,s)$ associadas a cada estratégia pura s em cada estado y.

	s_1	s_2	s_3	s_4	s_5	s_6	s_7	s_8	s_9
y_1	0	0.1	0.2	0.4	0.5	0.6	0.8	0.9	1
y_2	3	2.3	1.6	2.8	2.1	1.4	2.6	1.9	1.2
	s_{10}	s_{11}	s_{12}	s_{13}	s_{14}	s_{15}	s_{16}	s_{17}	s_{18}
y_1	0.5	0.6	0.7	0.9	1	1.1	1.3	1.4	1.5
y_2	2.9	2.2	1.5	2.7	2	1.3	2.5	1.8	1.1
	s_{19}	s_{20}	s_{21}	s_{22}	s_{23}	s_{24}	s_{25}	s_{26}	s_{27}
y_1	1	1.1	1.2	1.4	1.5	1.6	1.8	1.9	2
y_2	2.8	2.1	1.4	2.6	1.9	1.2	2.4	1.7	1

$$\mathcal{L}(s) = \sum_y P(y)L(y,s).$$

A solução s que minimiza esse critério usualmente é uma estratégia pura admissível que é representada por um vértice na Figura 4.2. Entretanto, se duas estratégias admissíveis s' e s'' satisfizerem esse critério com $\mathcal{L}(s') = \mathcal{L}(s'')$, todas as estratégias mistas formadas a partir dessas estratégias também serão soluções para o problema, pois a combinação convexa delas possui a mesma perda. Na Figura 4.2, a solução pode ser encontrada graficamente deslocando-se a reta $P(y_1)L(y_1,s)+P(y_2)L(y_2,s) =$ constante até encontrar o primeiro ponto de soluções admissíveis. A reta solução é $P(y_1)L(y_1,s)+P(y_2)L(y_2,s) = 0.872$ implicando, que a estratégia s_3 é a melhor segundo esse critério. Essa estratégia sugere que "se invista na bolsa de valores" em caso de haver uma estimativa positiva ou neutra do mercado e "se invista em títulos do governo" em caso de haver uma estimativa negativa do mercado.

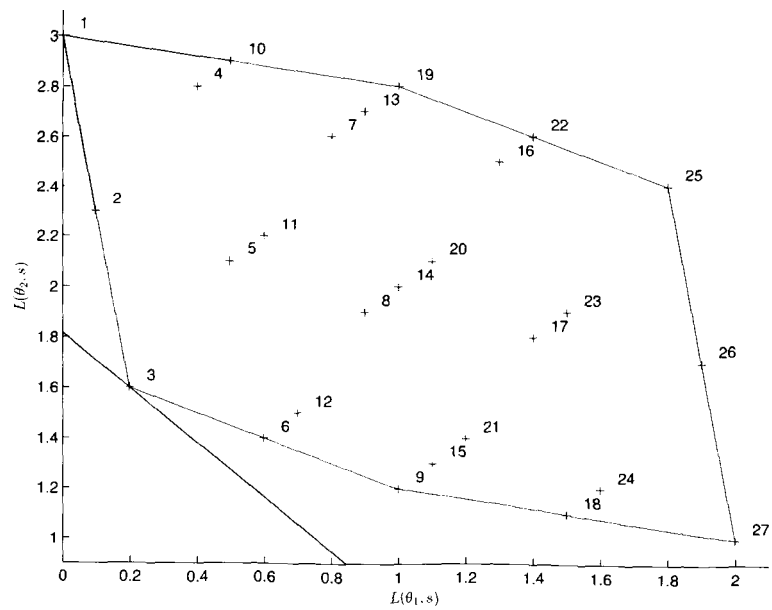


Figura 4.2: Perdas esperadas e solução do problema do investidor.

É válido mencionar que o critério de Bayes é apenas um dos critérios possíveis para resolver esse problema. A apresentação usada aqui é próxima daquela considerada em Chernoff e Moses (1986), onde outros critérios também são apresentados para lidar com o problema de tomada de decisão. Outras referências interessantes sobre o tema que podem ser acessadas pelo leitor interessado são Weiss (1961), Halter e Dean (1971), Beckman e Neto (1980) e Clemen (1996).

Utilizando o conceito de distribuição condicional, vamos agora introduzir o teorema de Bayes, para o caso bivariado: sejam X e Y duas variáveis aleatórias com espaços amostrais respectivamente iguais a $\mathbb{X}$ e $\mathbb{Y}$. Considere também as definições das funções de densidades condicionais $f_{Y/X}(y/x)$ e $f_{X/Y}(x/y)$ dadas respectivamente por

$$f_{Y/X}(y/x) = \frac{f_{X,Y}(x,y)}{f_X(x)}$$

e

$$f_{X/Y}(x/y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}.$$

Note também que podemos escrever

$$f_{X,Y}(x,y) = f_{X/Y}(x/y)f_Y(y) = f_{Y/X}(y/x)f_X(x),$$

$$f_X(x) = \int_{\mathbb{Y}} f_{X,Y}(x,y)dy = \int_{\mathbb{Y}} f_{X/Y}(x/y)f_Y(y)dy$$

e

$$f_Y(y) = \int_{\mathbb{X}} f_{X,Y}(x,y)dx = \int_{\mathbb{X}} f_{Y/X}(y/x)f_X(x)dx.$$

Então, usando essas manipulações algébricas, podemos enunciar o seguinte teorema:

Teorema 4.1 (Bayes) Sejam X e Y duas variáveis aleatórias com espaços amostrais respectivamente iguais a $\mathbb{X}$ e $\mathbb{Y}$. Então, podemos escrever

$$f_{Y/X}(y/x) = \frac{f_{X/Y}(x/y)f_Y(y)}{\int_{\mathbb{Y}} f_{X,Y}(x,y)dy} = \frac{f_{X/Y}(x/y)f_Y(y)}{\int_{\mathbb{Y}} f_{X/Y}(x/y)f_Y(y)dy}.$$

Nota 4.1 Utilizando manipulações algébricas análogas, a versão discreta desse teorema pode ser escrita da seguinte forma:

$$f_{Y/X}(y/x) = \frac{f_{X/Y}(x/y)f_Y(y)}{\sum_{\mathbb{Y}} f_{X,Y}(x,y)} = \frac{f_{X/Y}(x/y)f_Y(y)}{\sum_{\mathbb{Y}} f_{X/Y}(x/y)f_Y(y)}.$$

Exemplo 4.8 (Continuação do Exemplo 4.4 – Teorema de Bayes) Note que podemos calcular novamente as funções de densidade condicional $f_{Y/X}(y/x)$ e $f_{X/Y}(x/y)$ do Exemplo 4.6 utilizando o Teorema de Bayes:

$$f_{Y/X}(y/x) = \frac{f_{X/Y}(x/y)f_Y(y)}{\int_{\mathbb{Y}} f_{X/Y}(x/y)f_Y(y)dy} = \frac{\frac{\frac{3}{2}(x+y-2xy^2)}{\frac{3}{4}+\frac{3}{2}y-\frac{3}{2}y^2} \times \left(\frac{3}{4}+\frac{3}{2}y-\frac{3}{2}y^2\right)}{\int_0^1 \frac{\frac{3}{2}(x+y-2xy^2)}{\frac{3}{4}+\frac{3}{2}y-\frac{3}{2}y^2} \times \left(\frac{3}{4}+\frac{3}{2}y-\frac{3}{2}y^2\right) dy} = \frac{\frac{3}{2}(x+y-2xy^2)}{\frac{3}{4}+\frac{x}{2}}.$$

$$f_{X/Y}(x/y) = \frac{f_{Y/X}(y/x)f_X(x)}{\int_{\mathbb{X}} f_{Y/X}(y/x)f_X(x)dx} = \frac{\frac{\frac{3}{2}(x+y-2xy^2)}{\frac{3}{4}+\frac{x}{2}} \times \left(\frac{3}{4}+\frac{x}{2}\right)}{\int_0^1 \frac{\frac{3}{2}(x+y-2xy^2)}{\frac{3}{4}+\frac{x}{2}} \times \left(\frac{3}{4}+\frac{x}{2}\right) dx} = \frac{\frac{3}{2}(x+y-2xy^2)}{\frac{3}{4}+\frac{3}{2}y-\frac{3}{2}y^2}.$$

Para o caso mais geral, onde temos variáveis aleatórias multivariadas, discretas ou contínuas, com mais de dois elementos, podemos ter a função de densidade condicional ou a função de frequência condicional com versões multivariadas. Isso ocorre quando estamos interessados em modelar, por exemplo, a distribuição conjunta da renda domiciliar e da escolaridade do chefe de família, condicionadas a famílias com chefe entre 25 e 40 anos. Nesse caso, a distribuição das primeiras duas variáveis está condicionada a valores da terceira variável. Vamos então escrever a expressão mais geral para funções de frequência condicional e de densidade condicional.

Seja $X = \left[X_1, \ldots, X_K\right]^T$ um vetor aleatório, contínuo ou discreto[2], com K componentes. Vamos supor, sem perda de generalidade, que estamos interessados no comportamento dos primeiros 2 componentes do vetor X, condicionados a valores dos demais componentes. Portanto, a função de densidade conjunta condicional (ou a função de frequência conjunta condicional), do vetor $\left[X_1, X_2\right]^T$ tem expressão

$$f_{X_1,X_2/X_3,\ldots,X_K}(x_1, x_2/x_3, \ldots, x_K) = \frac{f_{X_1,X_2,X_3,\ldots,X_K}(x_1, x_2, x_3, \ldots, x_K)}{f_{X_3,\ldots,X_K}(x_3, \ldots, x_K)}, \tag{4.7}$$

onde $f_{X_3,\ldots,X_K}(x_3, \ldots, x_K)$ é a função de densidade (ou de frequência) marginal conjunta para o vetor de variáveis $\left[X_3, \ldots, X_K\right]^T$. A função condicional é definida $f_{X_1,X_2/X_3,\ldots,X_K}(x_1, x_2/x_3, \ldots, x_K)$ nos pontos $\left[x_3, \ldots, x_K\right]^T$ onde $f_{X_3,\ldots,X_K}(x_3, \ldots, x_K) > 0$. A expressão para a função de densidade ou frequência marginal é dada por

$$f_{X_3,\ldots,X_K}(x_3, \ldots, x_K) = \int_{x_1=-\infty}^{\infty} \int_{x_2=-\infty}^{\infty} f_{X_1,X_2,X_3,\ldots,X_K}(x_1, x_2, x_3, \ldots, x_K)dx_2dx_1,$$

conforme visto anteriormente. Expressões similares podem ser aplicadas para encontrar a função de distribuição acumulada condicional, tanto no caso contínuo quanto no caso discreto.

[2]Em muitas aplicações, o vetor de variáveis aleatórias pode conter tanto variáveis discretas como contínuas. Para facilitar as explicações neste capítulo, estamos considerando que um vetor aleatório tem todos os seus componentes contínuos ou todos os seus componentes discretos. No entanto, os resultados apresentados aqui podem ser facilmente estendidos para situações onde os vetores aleatórios possuem componentes discretos e contínuos ao mesmo tempo.

Vamos finalizar essa seção considerando a versão multidimensional do Teorema de Bayes (apresentado no Teorema 4.1). Seja $Z = [Z_1, \cdots, Z_K]$ um vetor aleatório dividido em duas partes $X = [X_1, \cdots, X_m]$ e $Y = [Y_1, \cdots, Y_n]$, onde $n + m = K$, cada variável X_i está definida no espaço amostral $\mathbb{X}_i$, $i = 1, \cdots, m$ e cada variável Y_i está definida no espaço amostral $\mathbb{Y}_i$, $i = 1, \cdots, n$. Da mesma forma que procedemos para enunciar o Teorema 4.1, considere as definições de probabilidades condicionais $f_{Y/X}(y/x)$ e $f_{X/Y}(x/y)$ dadas respectivamente por

$$f_{Y/X}(y/x) = \frac{f_Z(z)}{f_X(x)} = \frac{f_{X,Y}(x,y)}{f_X(x)}$$

e

$$f_{X/Y}(x/y) = \frac{f_Z(z)}{f_Y(y)} = \frac{f_{X,Y}(x,y)}{f_Y(y)}$$

Note também que podemos escrever

$$f_Z(z) = f_{X,Y}(x,y) = f_{X/Y}(x/y) f_Y(y) = f_{Y/X}(y/x) f_X(x),$$

$$\begin{aligned} f_X(x) &= \int_{\mathbb{Y}} f_{X,Y}(x,y)dy = \int_{\mathbb{Y}_1} \cdots \int_{\mathbb{Y}_n} f_{X_1 \cdots, X_m, Y_1, \cdots, Y_n}(x_1, \cdots, x_m, y_1, \cdots, y_n)dy_1 \cdots dy_n \\ &= \int_{\mathbb{Y}_1} \cdots \int_{\mathbb{Y}_n} f_{X_1, \cdots, X_m/Y_1, \cdots, Y_n}(x_1, \cdots, x_m/y_1, \cdots, y_n) f_{Y_1, \cdots, Y_n}(y_1, \cdots, y_n)dy_1 \cdots dy_n \end{aligned}$$

e

$$\begin{aligned} f_Y(y) &= \int_{\mathbb{X}} f_{X,Y}(x,y)dx = \int_{\mathbb{X}_1} \cdots \int_{\mathbb{X}_m} f_{X_1 \cdots, X_m, Y_1, \cdots, Y_n}(x_1, \cdots, x_m, y_1, \cdots, y_n)dx_1 \cdots dx_m \\ &= \int_{\mathbb{X}_1} \cdots \int_{\mathbb{X}_m} f_{Y_1, \cdots, Y_n/X_1, \cdots, X_m}(y_1, \cdots, y_n/x_1, \cdots, x_m) f_{X_1, \cdots, X_m}(x_1, \cdots, x_m)dx_1 \cdots dx_m. \end{aligned}$$

Usando essas manipulações algébricas, podemos enunciar o seguinte teorema:

Teorema 4.2 (Bayes) Seja $Z = [Z_1, \cdots, Z_K]$ um vetor aleatório dividido em duas partes $X = [X_1, \cdots, X_m]$ e $Y = [Y_1, \cdots, Y_n]$, onde $n + m = K$, cada variável X_i está definida no espaço amostral $\mathbb{X}_i$, $i = 1, \cdots, m$ e cada variável Y_i está definida no espaço amostral $\mathbb{Y}_i$, $i = 1, \cdots, n$. Então, podemos escrever

$$f_{Y/X}(y/x) = \frac{f_{X/Y}(x/y) f_Y(y)}{\int_{\mathbb{Y}} f_{X,Y}(x,y)dy} = \frac{f_{X/Y}(x/y) f_Y(y)}{\int_{\mathbb{Y}} f_{X/Y}(x/y) f_Y(y)dy}.$$

Na Seção 5.7 aplicaremos o Teorema 4.2 em vários casos particulares.

Utilizando manipulações algébricas análogas, a versão discreta desse teorema pode ser escrita analogamente à versão discreta apresentada na Nota 4.1.

4.3 Momentos de variáveis aleatórias multivariadas

Nesta seção estenderemos o conceito de momentos de variáveis aleatórias para o caso multivariado. Da mesma maneira que vimos funções densidade e frequência conjuntas, marginais e condicionais, os momentos também podem ser definidos de acordo com esses três tipos de funções de densidade ou de frequência. Inicialmente, apresentaremos as definições para momentos com base nas distribuições conjuntas. Em seguida apresentaremos como os momentos podem ser calculados para as distribuições marginais. Finalmente, apresentaremos as definições de momentos condicionais, que são extremamente importantes, por exemplo, quando estudarmos modelos de regressão em geral. De fato, nos modelos de regressão, em muitos casos, os pesquisadores estão interessados em como determinados momentos condicionais de uma determinada variável aleatória respondem a valores de variáveis aleatórias preditoras.

Seja uma variável aleatória X multivariada contínua com K componentes, e função de densidade conjunta $f_{X_1,\ldots,X_K}(x_1,\ldots,x_K)$. Seja uma função $h(\cdot)$ qualquer, com $h : \Re^K \mapsto \Re$. Ou seja, a função $h(\cdot)$ é definida no $\Re^K$ e assume valores[3] em $\Re$. O valor esperado, ou a expectância, ou o momento da função $h(X_1,\ldots,X_K)$ é dado por

$$\mathrm{E}\big[h(X_1,\ldots,X_K)\big] = \int_{x_1=-\infty}^{\infty} \cdots \int_{x_K=-\infty}^{\infty} h(x_1,\ldots,x_K) f_{X_1,\ldots,X_K}(x_1,\ldots,x_K) dx_K \ldots dx_1. \tag{4.8}$$

Para variáveis aleatórias discretas multivariadas, com função de frequência conjunta $f_{X_1,\ldots,X_K}(x_1,\ldots,x_K)$, o valor esperado da função $h(X_1,\ldots,X_K)$ é dado por

$$\mathrm{E}\big[h(X_1,\ldots,X_K)\big] = \sum_{x_1 \in \mathbb{X}_{X_1}} \cdots \sum_{x_K \in \mathbb{X}_{X_K}} h(x_1,\ldots,x_K) f_{X_1,\ldots,X_K}(x_1,\ldots,x_K), \tag{4.9}$$

onde $\mathbb{X}_{X_i}$ é o conjunto de valores possíveis, ou espaço amostral, para a variável individual X_i, $i = 1,\ldots,K$. Note que o somatório do valor esperado para variáveis discretas é efetuado para todos os valores possíveis do vetor aleatório.

Exemplo 4.9 (Continuação do Exemplo 4.1 – Cálculo do valor esperado). Vamos agora retornar ao Exemplo 4.1, para um vetor bivariado discreto composto pelos componentes X e Y. Consideremos a função

[3]Podemos ter o caso mais geral, onde $h : \Re^K \mapsto \Re^M$, ou seja, $h(\cdot)$ assume valores no conjunto $\Re^M$. Para simplificar a discussão nesta seção, vamos supor que a função $h(\cdot)$ assume apenas valores reais.

$h(X,Y) = (X+Y)^2$. Portanto, o valor esperado de $h(X,Y)$ é dado pelo somatório

$$\begin{aligned}\mathrm{E}\big[h(X,Y)\big] &= (1+1)^2 f_{X,Y}(1,1) + (1+2)^2 f_{X,Y}(1,2) + \cdots + (3+2)^2 f_{X,Y}(3,2) \\ &= 2^2 \times 0.20 + 3^2 \times 0.05 + \cdots + 5^2 \times 0.20 = 14.\end{aligned}$$

Considere agora que a função $h(X,Y) = X^3$; ou seja, a função $h(X,Y)$ depende apenas do valor do primeiro elemento X. A utilização da fórmula para o valor esperado de $h(X,Y)$ continua a mesma. De fato, podemos calcular a expectância da nova função usando

$$\begin{aligned}\mathrm{E}\big[h(X,Y)\big] &= 1^3 f_{X,Y}(1,1) + 1^3 f_{X,Y}(1,2) + \cdots + 3^3 f_{X,Y}(3,2) \\ &= 1 \times 0.20 + 1 \times 0.05 + \cdots + 27 \times 0.20 = 14.8.\end{aligned}$$

Exemplo 4.10 (Continuação do Exemplo 4.4 – Cálculo do valor esperado). Voltemos agora ao exemplo de uma variável bivariada contínua, conforme visto no enunciado do Exemplo 4.4. Considere a função $h(X,Y) = (X-Y)^2$. Portanto, o valor esperado para a função $h(X,Y)$ terá expressão

$$\begin{aligned}\mathrm{E}\big[h(X,Y)\big] &= \int_{x=0}^{1}\int_{y=0}^{1} (x-y)^2 f_{X,Y}(x,y)dxdy \\ &= \int_{x=0}^{1}\int_{y=0}^{1} (x-y)^2 \frac{3}{2}(x+y-2xy^2)dxdy = \frac{1}{5}.\end{aligned}$$

Imaginemos agora uma função $h(X,Y) = \sqrt{Y}$; ou seja, o valor de $h(\cdot)$ depende apenas do segundo componente do vetor bivariado $\big[X,Y\big]^T$. O valor esperado é calculado como

$$\begin{aligned}\mathrm{E}\big[h(X,Y)\big] &= \int_{x=0}^{1}\int_{y=0}^{1} y^{1/2} f_{X,Y}(x,y)dxdy \\ &= \int_{x=0}^{1}\int_{y=0}^{1} y^{1/2} \frac{3}{2}(x+y-2xy^2)dxdy = \frac{47}{70}.\end{aligned}$$

Com base nos exemplos acima, podemos agora tentar calcular o valor esperado de X individualmente, por exemplo, a partir da função de densidade conjunta ou função de frequência conjunta das variáveis X e Y. Para isso, podemos proceder de duas maneiras: (1) podemos simplesmente calcular o valor esperado de X fazendo $h(X,Y) = X$, e procedendo da mesma maneira que procedemos nas segundas versões da função $h(X,Y)$ para os Exemplos 4.9 e 4.10 (variáveis discretas e variáveis contínuas, respectivamente); (2) podemos calcular a função de densidade de probabilidade marginal $f_X(x)$ para a variável X, a partir da função de densidade conjunta, e em seguida utilizá-la para calcular o valor esperado, usando a expressão

já conhecida para variáveis univariadas

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

Pode-se mostrar que ambos os procedimentos irão fornecer os mesmos valores para os momentos individuais das variáveis aleatórias que compõem o vetor aleatório. De fato,

$$\int_{x=-\infty}^{+\infty} x f_X(x) dx = \int_{x=0}^{1} x \left[\frac{x}{2} + \frac{3}{4}\right] dx = \frac{13}{24},$$

$$\int_{x=0}^{1}\int_{y=0}^{1} x f_{X,Y}(x,y) dx dy = \int_{x=0}^{1}\int_{y=0}^{1} x \left[\frac{3}{2}(x + y - 2xy^2)\right] dx dy = \frac{13}{24}.$$

Para calcular momentos de ordem maior do que 1, como é o caso da variância, pode-se também utilizar qualquer um desses dois procedimentos e os resultados obtidos serão exatamente os mesmos. De fato, tudo isso pode ser provado matematicamente.

Para um vetor bivariado $[X, Y]^T$, um momento muito importante é conhecido como a **covariância** entre duas variáveis aleatórias. A covariância possui expressão

$$\mathrm{Cov}[X, Y] = \mathrm{E}\Big[(X - \mathrm{E}[X])(Y - \mathrm{E}[Y])\Big], \tag{4.10}$$

onde $\mathrm{E}[X]$ é o valor esperado da variável X e $\mathrm{E}[Y]$ é o valor esperado da variável aleatória Y. Observe que

$$\begin{aligned}
\mathrm{E}\Big[(X - \mathrm{E}[X])(Y - \mathrm{E}[Y])\Big] &= \mathrm{E}\Big[XY - \mathrm{E}[X]Y - \mathrm{E}[Y]X + \mathrm{E}[X]\mathrm{E}[Y]\Big] \\
&= \mathrm{E}\Big[XY\Big] - \mathrm{E}\Big[\mathrm{E}[X]Y\Big] - \mathrm{E}\Big[\mathrm{E}[Y]X\Big] + \mathrm{E}\Big[\mathrm{E}[X]\mathrm{E}[Y]\Big] \\
&= \mathrm{E}[XY] - \mathrm{E}[X]\mathrm{E}[Y] - \mathrm{E}[X]\mathrm{E}[Y] + \mathrm{E}[X]\mathrm{E}[Y] \\
&= \mathrm{E}[XY] - \mathrm{E}[X]\mathrm{E}[Y].
\end{aligned}$$

Quando $X = Y$, a covariância transforma-se na expressão da variância de X. A partir da covariância, podemos definir o **coeficiente de correlação** entre X e Y, expresso por

$$\rho(X, Y) = \mathrm{Corr}[X, Y] = \frac{\mathrm{Cov}[X, Y]}{\sqrt{\mathrm{Var}[X]\mathrm{Var}[Y]}}, \tag{4.11}$$

onde $\mathrm{Var}[X]$ e $\mathrm{Var}[Y]$ são as variâncias de X e Y respectivamente. Conforme já mencionado na Seção 2.3, pode-se mostrar que

$$-1 \leq \rho(X, Y) \leq 1, \tag{4.12}$$

para quaisquer variáveis aleatórias X e Y. Em particular, quando $X = aY + b$, a e b constantes reais, com $a \neq 0$, ou seja, quando X e Y tem uma relação linear perfeita, $|\rho(X,Y)| = 1$. Quando $a < 0$, $\rho(X,Y) = -1$, e quando $a > 0$, $\rho(X,Y) = 1$.

Exemplo 4.11 (Continuação do Exemplo 4.1 – Cálculo do coeficiente de correlação). Para a variável discreta bivariada no Exemplo 4.1, podemos calcular a covariância entre X e Y e o coeficiente de correlação. Os valores esperados para X e Y são

$$\mathrm{E}[X] = 2.2 \text{ e } \mathrm{E}[Y] = 1.4.$$

A covariância é dada por

$$\begin{aligned}\mathrm{E}\big[(X-2.2)(Y-1.4)\big] =& (1-2.2)(1-1.4)f_{X,Y}(1,1) + (1-2.2)(2-1.4)f_{X,Y}(1,2) \\ &+ \cdots + (3-2.2)(2-1.4)f_{X,Y}(3,2) = 0.07,\end{aligned}$$

enquanto as variâncias têm valores

$$\mathrm{Var}[X] = 0.66 \text{ e } \mathrm{Var}[Y] = 0.24.$$

Finalmente, o coeficiente de correlação tem valor

$$\rho(X,Y) = \frac{0.07}{\sqrt{0.24 \times 0.66}} = 0.175881618.$$

Exemplo 4.12 (Continuação do Exemplo 4.4 – Cálculo do coeficiente de correlação). Para a variável contínua bivariada do Exemplo 4.4, os valores esperados são

$$\mathrm{E}[X] = \frac{13}{24} \text{ e } \mathrm{E}[Y] = \frac{1}{2}.$$

enquanto as variâncias têm valores

$$\mathrm{Var}[X] = \frac{47}{576} \text{ e } \mathrm{Var}[Y] = \frac{3}{40}.$$

A covariância e coeficiente de correlação entre X e Y possuem valores

$$\mathrm{E}\big[(X-\mathrm{E}[X])(Y-\mathrm{E}[Y])\big] = -\frac{1}{48} \text{ e } \rho(X,Y) = \frac{-\frac{1}{48}}{\sqrt{\frac{47}{576} \times \frac{3}{40}}} = -0.2663118206.$$

Considere agora a soma de duas variáveis aleatórias X e Y, ponderadas pelas constantes a e b. Como já vimos na Proposição 3.4 (v), A variância de $aX + bY$ tem expressão

$$\text{Var}[aX + bY] = a^2\text{Var}[X] + b^2\text{Var}[Y] + 2ab\text{Cov}[X, Y].$$

No caso mais geral, temos que a soma ponderada de variáveis aleatórias $a_1X_1 + a_2X_2 + \cdots + a_nX_n$ tem variância (vide Exercícios 4.1 e 4.2)

$$\text{Var}[a_1X_1 + a_2X_2 + \cdots + a_nX_n] = \sum_{i=1}^{n} a_i^2\text{Var}[X_i] + 2\sum\sum_{i<j} a_ia_j\text{Cov}[X_i, X_j], \tag{4.13}$$

onde a_i, $i = 1, \ldots, n$, são constantes reais. Conforme vimos na Aplicação 3.2, a Eq. (4.13) é particularmente importante em administração de carteiras, onde uma carteira de ações, por exemplo, pode ser modelada como uma soma de variáveis aleatórias $a_1X_1 + a_2X_2 + \cdots + a_nX_n$, onde as constantes a_i, $i = 1, \ldots, n$, correspondem às posições da carteira em cada ação X_i.

A consequência imediata da Eq. (4.13) é que, quando a variáveis X_1, ..., X_n são não correlacionadas entre si, ou seja, quando $\rho(X_i, X_j) = 0$, para todo $i, j = 1, \ldots, n$, com $i \neq j$, então a variância da soma é igual à soma das variâncias. Portanto, quando as variáveis são não correlacionadas, e para a_1, ..., a_n constantes reais, temos, de acordo com a Proposição 3.4 (vi),

$$\text{Var}\big[a_1X_1 + \cdots + a_nX_n\big] = \sum_{i=1}^{n} a_i^2\text{Var}\big[X_i\big],$$

Quando $a_i = 1$, para todo $i = 1, \ldots, n$, a expressão acima torna-se $\text{Var}\big[X_1 + \cdots + X_n\big] = \sum_{i=1}^{n}\text{Var}\big[X_i\big]$.

4.3.1 Matriz de variância-covariância

Conforme vimos acima, a covariância é definida entre duas variáveis aleatórias. No entanto, em muitas aplicações, principalmente devido à importância da distribuição normal multivariada, que será discutida na Seção 4.5, estamos interessados não na covariância de um par de variáveis aleatórias especificamente, mas nas covariâncias para os diversos pares de uma sequência de variáveis aleatórias $X = [X_1, \ldots, X_n]^T$. Estamos interessados nas covariâncias $\text{Cov}[X_1, X_2]$, $\text{Cov}[X_1, X_3]$, ..., $\text{Cov}[X_{n-1}, X_n]$. Uma forma sucinta de representar toda a estrutura de variâncias e covariâncias para as variáveis em um vetor aleatório $X = [X_1, \ldots, X_n]^T$ é por meio da **matriz de variância-covariância**.

A matriz de variância-covariância, representada pela letra maiúscula Σ, contém na sua diagonal principal a variância de cada elemento individual X_i, $i = 1, \ldots, n$. Portanto, o elemento $\Sigma_{i,i}$, ou seja, o i-ésimo elemento da diagonal principal corresponde à variância $\text{Var}[X_i]$. Os elementos fora da diagonal principal

são as covariâncias entre dois elementos individuais X_i e X_j. Portanto, $\text{Cov}[X_i, X_j] = \Sigma_{i,j}$. Temos então

$$\Sigma = \begin{bmatrix} \text{Var}[X_1] & \text{Cov}[X_1, X_2] & \dots & \text{Cov}[X_1, X_n] \\ \text{Cov}[X_2, X_1] & \text{Var}[X_2] & \dots & \text{Cov}[X_2, X_n] \\ \dots & \dots & \dots & \dots \\ \text{Cov}[X_n, X_1] & \text{Cov}[X_n, X_2] & \dots & \text{Var}[X_n] \end{bmatrix}. \tag{4.14}$$

Além da matriz de variância-covariância, podemos escrever a matriz de correlações

$$R = \begin{bmatrix} 1 & \rho(X_1, X_2) & \dots & \rho(X_1, X_n) \\ \rho(X_2, X_1) & 1 & \dots & \rho(X_2, X_n) \\ \dots & \dots & \dots & \dots \\ \rho(X_n, X_1) & \rho(X_n, X_2) & \dots & 1 \end{bmatrix},$$

onde os elementos na diagonal principal, que correspondem à correlação de cada elemento com ele mesmo, são sempre 1. Os elementos fora da diagonal principal. $R_{i,j}$, $i \neq j$, correspondem às correlações entre X_i e X_j. Dado que $\text{Cov}[X_i, X_j] = \text{Cov}[X_j, X_i]$, tanto a matriz de variância-covariância quanto a matriz de correlações são simétricas. Pela definição de correlações, a partir da definição de covariância, a matriz de variância-covariância pode ser reescrita como

$$\Sigma = \begin{bmatrix} \text{Var}[X_1] & \sqrt{\text{Var}[X_1]\text{Var}[X_2]}\rho(X_1, X_2) & \dots & \sqrt{\text{Var}[X_1]\text{Var}[X_n]}\rho(X_1, X_n) \\ \sqrt{\text{Var}[X_2]\text{Var}[X_1]}\rho(X_2, X_1) & \text{Var}[X_2] & \dots & \sqrt{\text{Var}[X_2]\text{Var}[X_n]}\rho(X_2, X_n) \\ \dots & \dots & \dots & \dots \\ \sqrt{\text{Var}[X_n]\text{Var}[X_1]}\rho(X_n, X_1) & \sqrt{\text{Var}[X_n]\text{Var}[X_2]}\rho(X_n, X_2) & \dots & \text{Var}[X_n] \end{bmatrix}.$$

Muitos autores, representam a correlação entre as variáveis X_i e X_j por $\rho_{i,j}$, as variâncias $\text{Var}[X_i]$ por σ_i^2, e os desvios-padrões $\sqrt{\text{Var}[X_i]}$ por σ_i. Portanto, a matriz de variância-covariância pode ser reescrita de forma sucinta como

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_1\sigma_2\rho_{1,2} & \dots & \sigma_1\sigma_n\rho_{1,n} \\ \sigma_2\sigma_1\rho_{2,1} & \sigma_2^2 & \dots & \sigma_2\sigma_n\rho_{2,n} \\ \dots & \dots & \dots & \dots \\ \sigma_n\sigma_1\rho_{n,1} & \sigma_n\sigma_2\rho_{n,2} & \dots & \sigma_n^2 \end{bmatrix}.$$

Seja V uma matriz diagonal (elementos fora da diagonal principal são todos nulos), cujo elemento $V_{i,i}$ (i-ésimo elemento da diagonal principal) é igual à variância $\text{Var}[X_i]$. Pode-se mostrar que (vide Exercício 4.4) que

$$\Sigma = V^{1/2} \times R \times V^{1/2}.$$

A matriz de variância-covariância e a matriz de correlações serão bastantes utilizadas ao longo deste livro, principalmente quando trabalharmos com aproximações de variáveis aleatórias via distribuição normal multivariada, e quando trabalharmos com cópulas (que têm sido muito utilizadas para gerenciamente de risco).

Proposição 4.4 (Resultados importantes sobre o valor esperado de vetores aleatórios e a matriz de variância-covariância) Sejam X e Y vetores aleatórios com dimensão $n \times 1$; ou seja, tanto X quanto Y possuem n componentes aleatórios individuais cada um. Seja a um vetor constante, com dimensão $m \times 1$. Sejam A e B matrizes constantes com dimensão $m \times n$ (podendo m ser igual a n ou não). Então

(i) $\mathrm{E}[X+Y] = \mathrm{E}[X] + \mathrm{E}[Y]$.
(ii) $\mathrm{E}[A \times X + a] = A \times \mathrm{E}[X] + a$.
(iii) $\mathrm{Var}[A \times X + a] = A \times \mathrm{Var}[X] \times A'$. Note que $A \times X + a$ tem dimensão $m \times 1$. A matrix $\mathrm{Var}[A \times X + a]$ é a matriz de variância-covariância do vetor aleatório resultante da operação $A \times X + a$; essa matriz de variância-covariância tem dimensão $m \times m$.

4.3.2 Momentos condicionais

Apresentaremos agora uma discussão sobre momentos condicionais, que são importantes, por exemplo, quando estudarmos modelos de regressão em geral. De fato, nos modelos de regressão, em muitos casos, os pesquisadores estão interessados em como determinados momentos condicionais de uma determinada variável aleatória respondem a valores de variáveis aleatórias preditoras. Por exemplo, em modelos de regressão linear, a variável resposta é muitas vezes suposta como tendo uma distribuição normal, com valor esperado como função de variáveis preditoras. Nesse caso, estamos interessados no primeiro momento condicional (valor esperado condicional) da variável resposta em relação às variáveis independentes ou variáveis preditoras.

Considere uma variável aleatória bivariada contínua, composta pelos elementos X e Y. Na Seção 4.2, apresentamos os conceitos de distribuições condicionais. Considere então a densidade condicional de X dado Y $f_{X/Y}(x/y)$. O primeiro momento condicional ou **valor esperado condicional** ou expectância condicional de X dado $Y = y$ é dado por

$$\mathrm{E}[X/Y = y] = \int_{x=-\infty}^{\infty} x f_{X/Y}(x/y) dx, \tag{4.15}$$

para qualquer y com $f_Y(y) > 0$. Portanto, o momento condicional é simplesmente um momento no sentido que vimos no Capítulo 3, onde a função de densidade corresponde à função de densidade condicional. Para o caso onde X e Y são componentes de um vetor aleatório multivariado discreto, a expectância condicional

de X dado $Y = y$ pode ser escrita como

$$\mathrm{E}[X/Y = y] = \sum_{x \in \mathbb{X}_X} x f_{X/Y}(x/y), \tag{4.16}$$

para qualquer y com $f_Y(y) > 0$.

Para o caso mais geral, considere um vetor aleatório multivariado $[X_1, \ldots, X_K]^T \in \Re^K$, e uma função multivariada $h : \Re^K \mapsto \Re$. O valor esperado condicional de $h(\cdot)$ condicionado a determinados valores para as variáveis $X_3, \ldots, X_K$ (sem perda de generalidade) será

$$\begin{aligned}\mathrm{E}[h(X_1, \ldots, X_K)/X_3 = x_3, \ldots, X_K = x_K] = \\ = \int_{x_1=-\infty}^{\infty} \int_{x_2=-\infty}^{\infty} h(x_1, \ldots, x_K) f_{X_1, X_2/X_3, \ldots, X_K}(x_1, x_2/x_3, \ldots, x_K) dx_2 dx_1,\end{aligned}$$

para o caso contínuo. A versão dessa expressão para o caso discreto é a seguinte

$$\begin{aligned}\mathrm{E}[h(X_1, \ldots, X_K)/X_3 = x_3, \ldots, X_K = x_K] = \\ = \sum_{x_1 \in \mathbb{X}_{X_1}} \sum_{x_2 \in \mathbb{X}_{X_2}} h(x_1, \ldots, x_K) f_{X_1, X_2/X_3, \ldots, X_K}(x_1, x_2/x_3, \ldots, x_K).\end{aligned}$$

Em ambos os casos, estamos supondo que, nos pontos $x_3, \ldots, x_K$, $f_{X_3, \ldots, X_K}(x_3, \ldots, x_K)$ é maior do que zero.

Note que os valores esperados acima são necessariamente constantes reais, dado que todos os valores condicionantes $x_3, \ldots, x_K$ são conhecidos. Podemos considerar que os valores para $X_3, \ldots, X_K$ são desconhecidos, e são as variáveis aleatórias individuais para cada um dos componentes. Nesse caso, as expectâncias estão condicionadas ao valores aleatórios e, consequentemente, também são variáveis aleatórias. Portanto, temos

$$\begin{aligned}\mathrm{E}[h(X_1, \ldots, X_K)/X_3, \ldots, X_K] = \\ = \int_{x_1=-\infty}^{\infty} \int_{x_2=-\infty}^{\infty} h(x_1, \ldots, x_K) f_{X_1, X_2/X_3, \ldots, X_K}(x_1, x_2/X_3, \ldots, X_K) dx_2 dx_1 = g(X_3, \ldots, X_K),\end{aligned}$$

onde $g : \Re^{K-2} \mapsto \Re$ é uma função que dependerá da distribuição conjunta das variáveis $X_1, \ldots, X_K$. Para o caso discreto, a expressão correspondente é

$$\begin{aligned}\mathrm{E}[h(X_1, \ldots, X_K)/X_3, \ldots, X_K] = \\ = \sum_{x_1 \in \mathbb{X}_{X_1}} \sum_{x_2 \in \mathbb{X}_{X_2}} h(x_1, \ldots, x_K) f_{X_1, X_2/X_3, \ldots, X_K}(x_1, x_2/X_3, \ldots, X_K) = g(X_3, \ldots, X_K).\end{aligned}$$

Tanto no caso discreto quanto no contínuo, a função $g(\cdot)$ possui argumentos que são variáveis aleatórias; portanto, $g(X_3, \ldots, X_K)$ também é uma variável aleatória.

Proposição 4.5 (Lei das expectâncias iteradas) Seja uma variável aleatória bivariada, com elementos X e Y. Se $g : \Re^2 \mapsto \Re$ uma função bivariada, assumindo valores reais. Independente de X e Y serem discretas ou contínuas, sempre temos

(i) Se $g(x, y)$ pode ser escrita como $g_1(x) \times g_2(y)$, então $\mathrm{E}\big[g(X,Y)/Y\big] = g_2(Y)\mathrm{E}\big[g_1(X)/Y\big]$.
(ii) $\mathrm{E}\big[g(X,Y)\big] = \mathrm{E}\big[\mathrm{E}[g(X,Y)/Y]\big]$
(iii) $\mathrm{E}\big[g(X,Y)/Y\big] = \mathrm{E}\big[\mathrm{E}[g(X,Y)/X,Y]/Y\big]$
(iv) $\mathrm{Var}\big[Y\big] = \mathrm{Var}\big[\mathrm{E}[Y/X]\big] + \mathrm{E}\big[\mathrm{Var}[Y/X]\big]$.

A **Lei das Expectâncias Iteradas** é muito útil em várias situações como, por exemplo, quando estivermos estudando regressão linear no Capítulo 8.

4.4 Independência de variáveis aleatórias

Nesta seção trataremos de um dos conceitos mais importantes em análise estatística: o conceito de independência de variáveis aleatórias. Esse conceito será bastante utilizado principalmente no Capítulo 5, quando trabalharemos com simulações de Monte Carlo (onde as observações simuladas serão consideradas independentes) e trabalharemos com estimação via máxima verossimilhança. Também usaremos esse conceito no Capítulo 6 para discutir a questão da amostragem aleatória simples com ou sem reposição. Intuitivamente, variáveis aleatórias (ou observações) independentes são aquelas para as quais, quando conhecemos os valores para um subconjunto delas, isso não acrescenta informação alguma sobre os valores das demais variáveis. A seguir, abordamos esse conceito mais formalmente.

Considere um vetor de variáveis aleatórias (contínuas ou discretas) $X_1, \ldots, X_n$. Dizemos que essas n variáveis são **variáveis aleatórias independentes** se e somente se

$$F_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = F_{X_1}(x_2) \ldots F_{X_n}(x_n), \tag{4.17}$$

para todos os valores de $x_1, \ldots, x_n$. A função $F_{X_1,\ldots,X_n}(x_1, \ldots, x_n)$ corresponde à função de distribuição acumulada conjunta entre as variáveis $X_1, \ldots, X_n$, e $F_{X_i}(x_i)$ é a função de distribuição acumulada marginal para a variável aleatória individual X_i, $i = 1, \ldots, n$.

No caso de variáveis aleatórias discretas, podemos caracterizar independência a partir da função de frequência conjunta $f_{X_1,\ldots,X_n}(x_1, \ldots, x_n)$. Portanto, as variáveis aleatórias discretas $X_1, \ldots, X_n$ são

independentes se e somente se

$$f_{X_1,\ldots,X_n}(x_1,\ldots,x_n) = f_{X_1}(x_2)\ldots f_{X_n}(x_n), \tag{4.18}$$

onde $f_{X_i}(x_i)$ é a função de frequência marginal para a variável aleatória individual X_i, $i = 1,\ldots,n$. No caso de variáveis aleatórias contínuas, onde existe a função de densidade de probabilidade conjunta, as variáveis $X_1,\ldots,X_n$ são independentes se e somente se

$$f_{X_1,\ldots,X_n}(x_1,\ldots,x_n) = f_{X_1}(x_2)\ldots f_{X_n}(x_n), \tag{4.19}$$

onde $f_{X_i}(x_i)$ é a função de densidade de probabilidade marginal para a variável aleatória individual X_i, $i = 1,\ldots,n$. Pode-se mostrar que, no caso de variáveis discretas, as Eqs. (4.17) e (4.18) são equivalentes. Similarmente, no caso de variáveis contínuas, as Eqs. (4.17) e (4.19) são equivalentes.[4] As Eqs. (4.18) e (4.19) são as justificativas para os procedimentos de máxima verossimilhança que serão apresentados no Capítulo 5. Ressaltamos que as técnicas de máxima verossimilhança não são aplicadas apenas a problemas onde a amostra possui observações independentes. Em situações de séries temporais, ou de dados espaciais, por exemplo, onde a hipótese de independência não é mais válida, máxima verossimilhança é um método de estimação muito utilizado.

Da mesma maneira que tratamos de independência entre variáveis aleatórias individuais, cada qual pertencente ao conjunto $\Re$, também podemos tratar independência entre vetores de variáveis aleatórias. Por exemplo, considere um vetor de variáveis aleatórias $[X_1,\ldots,X_n,Y_1,\ldots,Y_m]^T \in \Re^{m+n}$, e sejam $X = [X_1,\ldots,X_n]^T$ e $Y = [Y_1,\ldots,Y_m]^T$ dois subvetores. Sejam

$$\begin{aligned} f_{X,Y}(x,y) &= f_{X_1,\ldots,X_n,Y_1,\ldots,Y_m}(x_1,\ldots,x_n,y_1,\ldots,y_m), \\ f_X(x) &= f_{X_1,\ldots,X_n}(x_1,\ldots,x_n), \\ f_Y(y) &= f_{Y_1,\ldots,Y_m}(y_1,\ldots,y_m), \end{aligned}$$

as funções de densidade de probabilidade conjunta e marginais respectivamente. No caso de variáveis discretas, o resultado é exatamente o mesmo, só que nesse caso devemos trabalhar com funções de frequência e não de densidade. Dizemos que os subvetores X e Y são independentes se e somente se

$$f_{X,Y}(x,y) = f_X(x)f_Y(y).$$

[4]De fato, para as variáveis aleatórias serem independentes, a igualdade na Eq. (4.19) deve ser verdadeira para todos os pontos em $\Re^n$, exceto em conjuntos de medida nula.

Note que, quando há independência entre os vetores X e Y (que podem ser vetores compostos por apenas um elemento), então, para a função de densidade (ou de frequência) condicional, vale

$$f_{X/Y}(x/y) = \frac{f_{X,Y}(x,y)}{f_Y(y)} = \frac{f_X(x)f_Y(y)}{f_Y(y)} = f_X(x). \tag{4.20}$$

A expressão acima tem uma implicação intuitiva importante. O fato de trabalharmos com densidades ou funções de frequência condicionais é que, em muitos casos, quando sabemos os valores da variável condicionante Y, por exemplo, isso nos traz uma nova função de densidade para X; ou seja, passamos de $f_X(x)$ original para $f_{X/Y}(x/y)$ condicional. Portanto, em muitos casos, o conhecimento sobre valores de Y nos trazem informações sobre os valores de X. No entanto, quando X e Y são independentes, o que acontece é que a densidade (ou função de frequência) marginal não se altera quando condicionamos a valores da variável Y; portanto, o conhecimento dos valores de Y não nos trazem informação alguma sobre os valores de X. A Eq. (4.20) nos passa então a formalização matemática embasando essa intuição. Devido ao fato de que a função de densidade (ou de frequência) condicional é igual à função de densidade marginal, os momentos condicionais também são iguais aos momentos marginais. De fato, seja $h : \Re^n \mapsto \Re$ uma função qualquer. Então

$$\mathrm{E}\big[h(X)/Y\big] = \mathrm{E}\big[h(X)\big].$$

Proposição 4.6 (Resultados importantes sobre independência entre variáveis aleatórias). Sejam X e Y dois vetores de variáveis aleatórias, com $X \in \Re^n$ e $Y \in \Re^m$. Seja $f_{X,Y}(x,y)$ a função de densidade ou de frequência conjunta entre todos $m+n$ elementos do vetor $[X^T, Y^T]^T$.

(i) Se existem funções $g_1(x)$ e $g_2(y)$, tais que podemos escrever $f_{X,Y}(x,y) = g_1(x)g_2(y)$, então os vetores X e Y são independentes entre si, e além disso, o vetor X tem função de densidade ou de frequência marginal

$$f_X(x) = K_1 g_1(x),$$

e o vetor Y tem função de densidade ou de frequência marginal

$$f_Y(y) = K_2 g_2(y),$$

para algum par de constantes K_1 e K_2. As constantes K_1 e K_2 são escolhidas de maneira que as funções marginais $f_X(x)$ e $f_Y(y)$ integrem (ou somem) para um. Portanto, no caso variáveis aleatórias contínuas, temos

$$K_1 = \frac{1}{\int_{x\in\Re^n} g_1(x)dx},$$
$$K_2 = \frac{1}{\int_{y\in\Re^m} g_2(y)dy}.$$

(ii) Se X e Y são vetores aleatórios independentes, então

$$\mathrm{E}\big[h_1(X)h_2(Y)\big] = \mathrm{E}\big[h_1(X)\big]\mathrm{E}\big[h_2(Y)\big],$$

para funções $h_1 : \Re^n \mapsto \Re$ e $h_2 : \Re^m \mapsto \Re$, tais que os valores esperados existam.

Prática 4.1 Considere duas variáveis aleatórias independentes X e Y, onde X tem uma distribuição lognormal com parâmetro μ e σ e Y tem uma distribuição gamma, com parâmetros α e β. Determine:

(i) a função de densidade conjunta para X e Y,
(ii) a função de densidade marginal de X,
(iii) a função de densidade marginal de Y,
(iv) a função de densidade condicional de X dado Y,
(v) a função de densidade condicional de Y dado X.

De acordo com a definição acima de independência, vimos que, para diversas variáveis aleatórias serem independentes, é preciso que a função de densidade de probabilidade, ou a função de frequência conjunta, ou a função de distribuição acumulada conjunta, seja igual ao produto das funções marginais. No entanto, temos um outro conceito, conhecido com **independência dois a dois**,[5] que é um conceito mais fraco que independência conforme vimos anteriormente, onde basta que os pares de variáveis aleatórias sejam independentes entre si. Portanto, dizemos que a sequência de variáveis aleatórias $X_1, \ldots, X_n$ é independente dois a dois se o somente se

$$f_{X_i,X_j}(x_i, x_j) = f_{X_i}(x_i) f_{X_j}(x_j),$$

para todo par de variáveis X_i e X_j, com $j \neq i$, $i, j = 1, \ldots, n$. O conceito de independência dois a dois é mais fraco que independência entre todas as variáveis. Portanto, se a sequência de variáveis $X_1, \ldots, X_n$ é independente, então essa sequência também é independente dois a dois; no entanto, o contrário não é verdade. Independência dois a dois não implica independência. Um outro fato importante em relação à independência dois a dois é que, se a sequência $X_1, \ldots, X_n$ for independente dois a dois, então as correlações $\rho(X_i, X_j)$ são todas nulas para todo par i, j, com $i \neq j$, e $i, j = 1, 2, \ldots, n$. Conforme vimos na Proposição 3.4 e revisitamos nesse capítulo, quando as correlações entre as variáveis aleatórias são todas nulas, a variância da soma é igual à soma das variâncias. Portanto, a variância da soma de uma sequência de variáveis aleatórias independentes é igual à soma das variâncias.

Exemplo 4.13 (Independência dois a dois) Considere X e Y duas variáveis aleatórias discretas com funções de frequência definidas da seguinte forma:

$$f_X(-1) = 1/2, \; f_X(1) = 1/2$$

[5] Em inglês, *pairwise independence*.

e

$$f_Y(-1) = 1/2,\ f_Y(1) = 1/2.$$

Defina $Z = XY$. Note que

$$f_Z(1) = f_X(1)f_Y(1) + f_X(-1)f_Y(-1) = 1/2$$

e

$$f_Z(-1) = f_X(1)f_Y(-1) + f_X(-1)f_Y(1) = 1/2.$$

É óbvio que X e Y são independentes. Note também que os pares X e Z e Y e Z também são independentes, pois

$$f_{X,Z}(1,1) = 1/4 = f_X(1)f_Z(1),$$
$$f_{X,Z}(1,-1) = 1/4 = f_X(1)f_Z(-1),$$
$$f_{X,Z}(-1,1) = 1/4 = f_X(-1)f_Z(1),$$
$$f_{X,Z}(-1,-1) = 1/4 = f_X(-1)f_Z(-1)$$

e

$$f_{Y,Z}(1,1) = 1/4 = f_Y(1)f_Z(1),$$
$$f_{Y,Z}(1,-1) = 1/4 = f_Y(1)f_Z(-1),$$
$$f_{Y,Z}(-1,1) = 1/4 = f_Y(-1)f_Z(1),$$
$$f_{Y,Z}(-1,-1) = 1/4 = f_Y(-1)f_Z(-1).$$

Entretanto, obviamente X, Y, Z não são independentes, pois $XYZ = 1$ para qualquer valor de X, Y e Z.

Uma versão análoga do Exemplo 4.13 para o caso contínuo pode ser encontrada em Romano e Siegel (1986).

4.5 Distribuição normal multivariada

Vamos agora estudar a distribuição multivariada que é provavelmente a mais importante em estatística e econometria: a **distribuição normal multivariada**. Sejam $Z_1, \ldots, Z_n$ uma sequência independente e identicamente distribuída (todos os componentes da sequência possuem a mesma distribuição), cada

qual com distribuição normal padronizada (ou seja, com média 0 e variância 1). Portanto, o vetor $Z = [Z_1, \ldots, Z_n]^T$ possui função de densidade conjunta

$$f_Z(z) = \frac{1}{(2\pi)^{n/2}} e^{-\frac{1}{2}z'z},$$

onde $z' = z^T$ corresponde à transposta do vetor z. Portanto, como z é um vetor coluna, o vetor z' será um vetor linha. O produto $z'z$ é um escalar, e é uma forma sucinta de escrevermos $z_1^2 + \cdots + z_n^2$. Seja A uma matriz $n \times n$, não singular, e seja b um vetor coluna com dimensão $n \times 1$. Então, o vetor aleatório $X = AZ + b$ tem distribuição normal multivariada com função de densidade conjunta

$$f_X(x) = \frac{1}{(2\pi)^{n/2}\left|\Sigma\right|^{1/2}} e^{-\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)}, \text{ para } x \in \Re^n, \tag{4.21}$$

onde $\mu = b$ e $\Sigma = AA'$, e $\left|\Sigma\right|$ corresponde ao determinante da matriz Σ. Dizemos que X tem distribuição normal multivariada com média μ e variância (ou variância-covariância) Σ. A notação nesse caso é $X \sim \mathrm{N}(\mu, \Sigma)$.

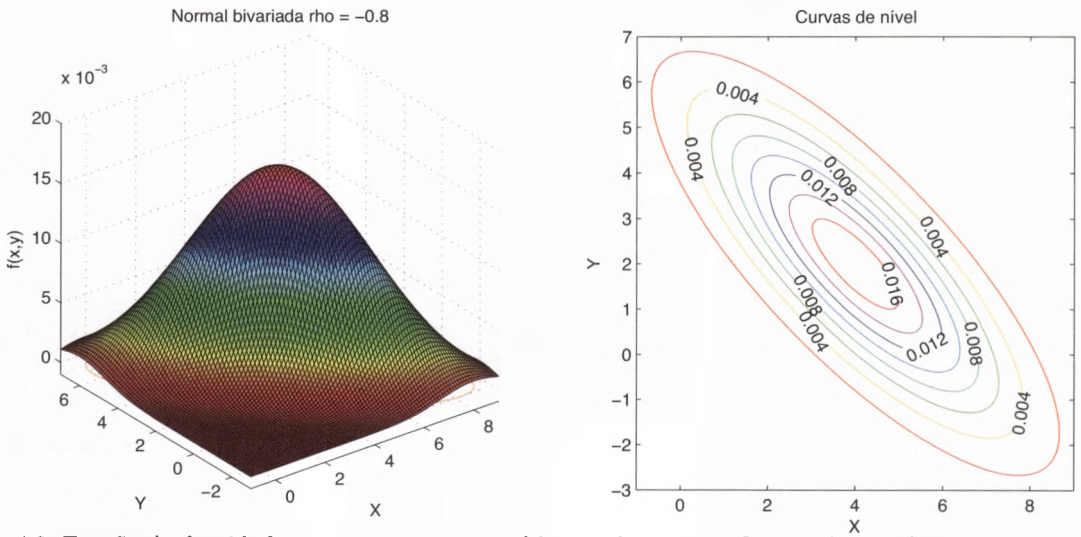


Figura 4.3: Função de densidade para um vetor normal bivariado, com coeficiente de correlação $\rho = -0.8$.

Pode-se mostrar que, se $X \sim \mathrm{N}(\mu, \Sigma)$ e D é uma matriz constante, com dimensão $m \times n$, então a variável $Y = D \times X$ tem distribuição normal multivariada com média $D\mu$ e variância $D\Sigma D'$. Uma aplicação direta desse resultado é a distribuição da média de variáveis aleatórias normais. Seja $X_1, \ldots, X_n$, uma sequência de variáveis aleatórias normais independentes e identicamente distribuídas, cada qual com variância σ^2 e média individual δ. O vetor $X = [X_1, \ldots, X_n]^T$ tem distribuição normal multivariada com média $\mu = [\delta, \ldots, \delta]^T$ e variância Σ, que é uma matriz diagonal, onde todos os elementos da diagonal principal são iguais a σ^2.

Pode-se mostrar que a variável aleatória S, que corresponde à média da sequência, $X_1, \ldots, X_n$, ou seja,

$$S = \frac{X_1 + \cdots + X_n}{n},$$

tem distribuição normal univariada, com média δ e variância σ^2/n.

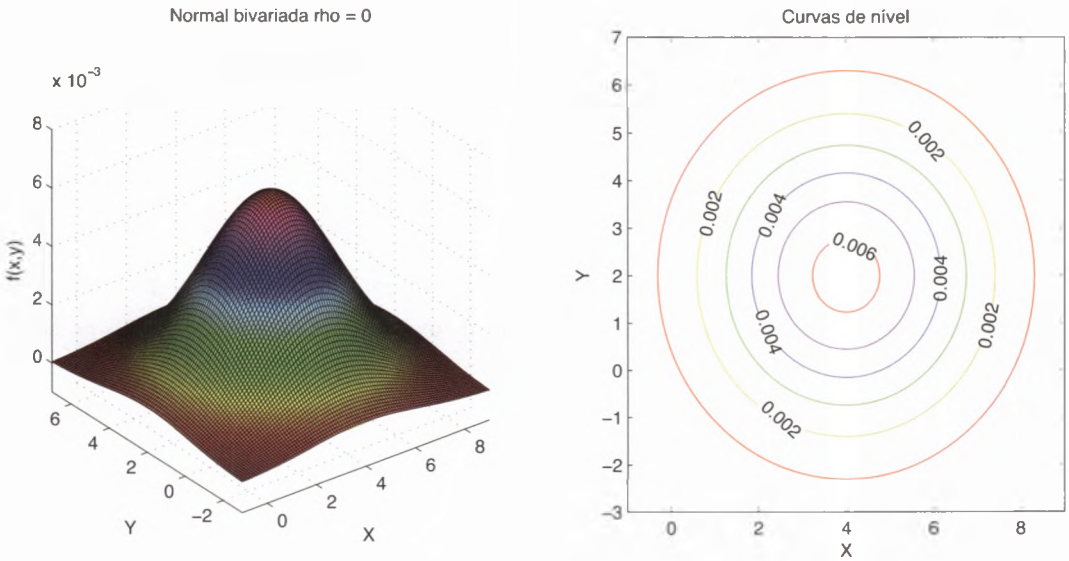


Figura 4.4: Função de densidade para um vetor normal bivariado, com coeficiente de correlação $\rho = 0$; ou seja, as duas variáveis aleatórias são independentes.

É importante ressaltar que se duas variáveis X e Y escalares são não correlacionadas, isso não significa que elas sejam independentes. Em geral, correlação nula não implica em independência. No entanto, no caso em que X e Y tenham uma distribuição normal bivariada, se a correlação entre elas for zero, então elas também são independentes. No caso mais geral, se a sequência de variáveis aleatórias $X_1, X_2, \ldots, X_n$ tiver distribuição normal multivariada, e a matriz de variância-covariância entre elas for diagonal (ou seja, todos os elementos fora da diagonal principal forem nulos), então os componentes $X_1, \ldots, X_n$ são independentes.

As Figuras 4.1, 4.3 e 4.4 apresentam as superfícies correspondentes às funções densidade de probabilidade de distribuições normais bivariadas para dois componentes X e Y, com coeficientes de correlação igual a 0.6, -0.8 e 0. No último caso, correlação nula implica duas normais independentes. Note que, a depender do coeficiente de correlação, observam-se a superfícies mais elipsoidais. O eixo maior da elipse será crescente ou decrescente, de acordo com o sinal da correlação. Para correlações positivas, o eixo maior da elipse é crescente, enquanto para correlações negativas o eixo maior da elipse é decrescente. A Figura 4.5 apresenta amostras geradas para diferentes valores de correlação. Note que o formato dos gráficos de dispersão das amostras geradas está bem em acordo com as curvas de nível das funções de densidade de probabilidade conjunta, nas Figuras 4.1, 4.3 e 4.4.

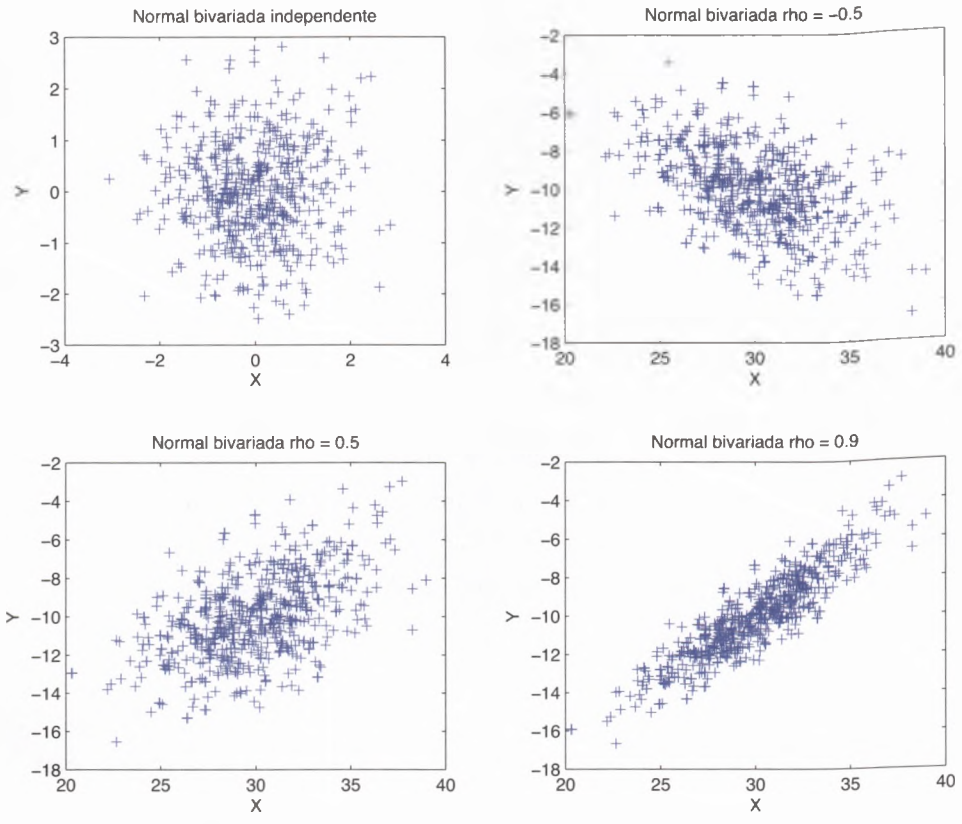


Figura 4.5: Amostras geradas para uma distribuição normal bivariada, com diferentes coeficientes de correlação ρ.

Proposição 4.7 (Alguns fatos estilizados a respeito da interrelação entre duas variáveis aleatórias) Sejam X e Y duas variáveis aleatórias, então:

(i) Se $\text{Cov}[X, Y] = 0$ (ou seja, $\rho(X, Y) = 0$), então X e Y são ditos não correlacionados.
(ii) Se X e Y são independentes, então $\text{Cov}[X, Y] = \rho(X, Y) = 0$.
(iii) Se X e Y são não correlacionados, isso não significa que eles sejam independentes.
(iv) Se X e Y são não correlacionados e $[X, Y]^T$ tem distribuição normal bivariada, então X e Y são independentes.

Exemplo 4.14 (Variáveis aleatórias não correlacionadas, mas não independentes) Seja X uma variável aleatória contínua com distribuição uniforme no intervalo $[-1, 1]$. Logo, a função de densidade de probabilidade

$$f_X(x) = \frac{1}{2}, \text{ quando } -1 < x < 1$$
$$f_X(x) = 0, \text{ caso contrário.}$$

Note que $E[X] = \int_{-1}^{1} x\frac{1}{2}dx = 0$.

Seja $Y = X^2$ uma variável aleatória contínua. Então, $Cov(X,Y) = E[XY] - E[X]E[Y] = E[X^3] = \int_{-1}^{1} x^3\frac{1}{2}dx = 0$ e, portanto, X e Y são não correlacionadas. Considerando a Proposição 4.6, defina $h_1(X) = X^2$ e $h_2(Y) = Y$ e note que $E[h_1(X)h_2(Y)] = E[X^4] = E[Y^2] = var(Y) + E[Y]^2$. Então, note que $E[h_1(X)h_2(Y)] \neq E[h_1(X)]E[h_2(Y)] = E[X^2]E[X^2] = E[X^2]^2 = E[Y]^2$. Logo, X e Y não são independentes.

4.6 Resultados adicionais

Apresentamos agora alguns resultados adicionais a respeito de variáveis aleatórias multivariadas. Inicialmente, discutiremos um resultado muito importante em estatística e econometria, conhecido como **lei dos grandes números**, que está diretamente ligado a convergência em probabilidade. Em seguida, apresentaremos uma versão multivariada da desigualdade de Jensen, apresentada no Capítulo 3 para o caso univariado.

4.6.1 Lei dos grandes números

Considere uma sequência de variáveis aleatórias Y_1, Y_2, ..., Y_n, ... Dizemos que a sequência $\{Y_n\}$ **converge em probabilidade** para a variável aleatória Y, quando, para todo $\epsilon > 0$, temos que

$$\lim_{n\to\infty} \text{Prob}\left[\left|Y_n - Y\right| \geq \epsilon\right] = 0.$$

Em geral, Y é uma constante fixa a. Quando a sequência $\{Y_n\}$ converge em probabilidade para Y ou para a, escrevemos

$$Y_n \xrightarrow{\text{P}} Y,$$

ou, no caso de uma constante,

$$Y_n \xrightarrow{\text{P}} a.$$

Vamos agora apresentar um resultado extremamente importante em teoria assintótica, muito utilizada para caracterizar os estimadores estatísticos. O resultado abaixo é uma versão simplificada do teorema conhecido como **lei fraca dos grandes números**. Existem muitas versões desse resultado, para os mais variados casos. Por exemplo, existem versões específicas para tratamento de convergência de médias onde as variáveis na média são temporalmente dependentes (WHITE, 2000). Essas versões são importantes quando estamos tratando das características dos estimadores de máxima verossimilhança, por exemplo, aplicados a dados de séries temporais.

Teorema 4.3 (Lei fraca dos grandes números) Seja $Y_n = \frac{X_1+\cdots+X_n}{n}$, onde $X_1, \ldots, X_n$ é uma sequência de variáveis indepedentes e identicamente distribuídas, cada qual com média μ e variância σ^2 (não necessariamente as variáveis X_i precisam ter distribuição normal). Essa sequência de variáveis X_i pode ser uma sequência de variáveis aleatórias discretas ou contínuas. Então

$$Y_n \xrightarrow{\text{P}} \mu.$$

Prova: Para demonstrar tal resultado, note que Y_n tem média μ e variância σ^2/n. De fato,

$$\text{E}\big[Y_n\big] = \text{E}\Big[\frac{X_1 + \cdots + X_n}{n}\Big] = \frac{\text{E}[X_1] + \cdots + \text{E}[X_n]}{n} = \frac{\mu + \cdots + \mu}{n} = \mu.$$

Para a variância, vale

$$\text{Var}\big[Y_n\big] = \text{Var}\Big[\frac{X_1 + \cdots + X_n}{n}\Big] = \frac{\text{Var}[X_1] + \cdots + \text{Var}[X_n]}{n^2} = \frac{\sigma^2 + \cdots + \sigma^2}{n^2} = \frac{\sigma^2}{n}.$$

Na última derivação acima, usamos o fato de a variância da soma de variáveis aleatórias independentes ser igual à soma das variâncias. Pela desigualdade de Chebishev (estudada na Seção 3.4.2) , temos que

$$\text{Prob}\big[\big|Y_n - \mu\big| \geq \epsilon\big] \leq \frac{\text{E}\big[\big|Y_n - \mu\big|^2\big]}{\epsilon^2} = \frac{\sigma^2}{n\epsilon^2} \to 0,$$

quando $n \to \infty$, e o portanto a sequência $\{Y_n\}$ obedece à definição de convergência em probabilidade para a constante μ, conforme queríamos provar.

4.6.2 Desigualdade de Jensen

Seja $X \in \Re^n$ uma variável aleatória multivariada qualquer (discreta ou contínua), e seja $g : \Re^n \mapsto \Re$ uma função convexa, no sentido de que

$$\lambda g(x) + (1-\lambda)g(y) \geq g(\lambda x + (1-\lambda)y),$$

para todo $\lambda \in (0,1)$ e para todo x e $y \in \Re^n$. Então

$$\text{E}\big[g(X)\big] \geq g(\text{E}\big[X\big]),$$

dado que ambos os valores esperados $\mathrm{E}[|X|]$ e $\mathrm{E}[|g(X)|] < \infty$. Analogamente, se $g : \Re^n \mapsto \Re$ for côncava, com

$$\lambda g(x) + (1-\lambda)g(y) \leq g(\lambda x + (1-\lambda)y),$$

para todo $\lambda \in (0,1)$ e para todo x e $y \in \Re^n$, então

$$\mathrm{E}[g(X)] \leq g(\mathrm{E}[X]).$$

Em particular, suponha que a função $g : \Re^n \mapsto \Re$ é côncava em $\Re^n$ e existe uma bola aberta $\mathbb{B} \in \Re^n$ onde $g(\cdot)$ é estritamente côncava, ou seja $\lambda g(x) + (1-\lambda)g(y) < g(\lambda x + (1-\lambda)y)$ para todo $\lambda \in (0,1)$ e para todo x e $y \in \mathbb{B}$. Além disso, suponha que a função de densidade de probabilidade para X seja estritamente positiva nesse conjunto aberto; ou seja, $f(x) > 0$ para todo $x \in \mathbb{B}$. Então, temos a desigualdade estrita

$$\mathrm{E}[g(X)] < g(\mathrm{E}[X]).$$

A recíproca é totalmente verdadeira para o caso de $g(\cdot)$ ser convexa.

Prática 4.2 Sejam X e Y duas variáveis aleatórias, com distribuição normal bivariada. Seja $g(x,y) = (x+y)^2$, para qualquer valor real de x e y. Mostre que $\mathrm{E}\big[g(X,Y)\big] > g\big(\mathrm{E}[X], \mathrm{E}[Y]\big)$.

Aplicação 4.2 (Aversão ao risco) Em teoria econômica, uma função de utilidade mapeia uma cesta $x = (x_1, \cdots, x_k)$ de k mercadorias $x_i, i = 1, \cdots, k$ em um número $u(x) = u(x_1, \cdots, x_k)$ que mede a satisfação ou utilidade dessa cesta de mercadorias. Dessa forma, supõe-se que agentes (consumidores) tomam decisões que maximizam as suas funções de utilidade.

Um agente é dito ser averso ao risco se prefere a expectativa de qualquer plano de consumo ao próprio plano de consumo, isto é,

$$E[u(x)] \leq u(E[x]).$$

Em outras palavras, isso significa que um agente prefere cestas certas (possivelmente menores) do que cestas arriscadas (possivelmente) maiores.

Logo, uma consequência direta da desigualdade de Jensen é que funções de utilidade de agentes aversos ao risco são côncavas. Discussões mais detalhadas sobre esse tema podem ser encontradas em LeRoy e Werner (2001) ou Mas-Colell, Whinston e Green (1995).

4.6.3 Função geratriz de momentos

Nesta seção, trataremos de um tópico muito importante para a caracterização de variáveis aleatórias: a **função geratriz de momentos**. Dentre as várias utilidades da função geratriz de momentos, podemos citar a possibilidade de identificar rapidamente a distribuição de somas de variáveis aleatórias independentes, conforme veremos nos exemplos a seguir. Inicialmente, consideraremos a função geratriz de momentos para uma variável univariada; em seguida, estenderemos esse conceito para variáveis multivariadas.

Considere o valor esperado a seguir, para uma variável aleatória contínua $X \in \Re$,

$$\mathrm{E}\left[e^{t_0|X|}\right] = \int_{x=-\infty}^{\infty} e^{t_0|X|} f(x)dx.$$

Se $\mathrm{E}\left[e^{t_0|X|}\right] < \infty$ para algum valor $t_0 > 0$, então a função geratriz de momentos de X, $M_X(t) = \mathrm{E}\left[e^{tX}\right]$ existe para todo t, com $|t| < t_0$. Pode-se mostrar que, se a matriz geratriz de momentos existe, então todos os momentos $\mathrm{E}\left[X^k\right]$ existem, para $k = 1, 2, 3, \ldots$. Além disso,

$$M_X(t) = \sum_{k=0}^{\infty} \frac{\mathrm{E}\left[X^k\right]}{k!},$$

para todo t, com $|t| < t_0$. Consequentemente,

$$\mathrm{E}\left[X^k\right] = \left.\frac{d^k M_X(t)}{dt^k}\right|_{t=0}.$$

Uma outra transformação importante para a caracterização de variáveis aleatórias é a **função característica**, definida por

$$\psi_X(t) = \mathrm{E}\left[e^{itX}\right] = \mathrm{E}\left[\cos tx + i \sin tx\right],$$

onde $i = \sqrt{-1}$ é o número imaginário. Pode-se mostrar que a função característica existe para qualquer valor de $t \in \Re$.

Para variáveis aleatórias discretas univariadas, as expressões para a função geratriz de momentos e para a função característica são

$$\mathrm{E}\left[e^{tX}\right] = \sum_{x \in \mathbb{X}} e^{tX} f(x),$$

$$\psi_X(t) = \mathrm{E}\left[e^{itX}\right] = \mathrm{E}\left[\cos tx + i \sin tx\right] = \sum_{x \in \mathbb{X}} \left[\cos tx + i \sin tx\right] f(x),$$

onde $\mathbb{X}$ é o espaço amostral para a variável aleatória discreta X. Por exemplo, se X é uma variável aleatória binomial, com parâmetros n e p, então $\mathbb{X} = \{0, 1, 2, \ldots, n\}$. Da mesma forma que no caso contínuo, para que a função geratriz de momentos exista, é necessário que o valor esperado $\mathrm{E}\left[e^{t_0|X|}\right] < \infty$, para algum valor $t_0 > 0$.

Teorema 4.4 (Propriedade da função geratriz de momentos) Sejam a e b constantes reais. Se a função geratriz de momentos $M_X(t)$ existe para a variável aleatória X, então

$$M_{a+bX}(t) = e^{at} M_X(bt).$$

Prática 4.3 Use a definição de função geratriz de momentos e prove o Teorema 4.4.

Exemplo 4.15 (Função geratriz de momentos para uma variável aleatória com distribuição binomial) Seja X uma variável aleatória com distribuição binomial, com parâmetros n e p. A função geratriz de momentos tem expressão

$$M_X(t) = \sum_{x=0}^{n} e^{tx} \binom{n}{x} p^x (1-p)^{n-x} = \sum_{x=0}^{n} \binom{n}{x} (pe^t)^x (1-p)^{n-x} = \left[pe^t + (1-p)\right]^n.$$

Note que, no caso da variável aleatória binomial, a função geratriz de momentos existe para qualquer valor $t > 0$. Então,

$$\mathrm{E}[X] = \left.\frac{dM_X(t)}{dt}\right|_{t=0} = n\left[pe^t + (1-p)\right]^{n-1} pe^t \Big|_{t=0} = np.$$

Para o segundo momento não-centrado, temos

$$\begin{aligned}
\mathrm{E}[X^2] &= \left.\frac{d^2 M_X(t)}{dt^2}\right|_{t=0} \\
&= n(n-1)\left[pe^t + (1-p)\right]^{n-2} p^2 e^{2t} \Big|_{t=0} + n\left[pe^t + (1-p)\right]^{n-1} pe^t \Big|_{t=0} \\
&= n(n-1)p^2 + np \\
&= n^2 p^2 + np - np^2.
\end{aligned}$$

Portanto,

$$\mathrm{Var} = \mathrm{E}[X^2] - \mathrm{E}[X]^2 = np - np^2 = np(1-p).$$

Exemplo 4.16 (Função geratriz de momentos para uma variável aleatória com distribuição gamma) Seja X uma variável aleatória com distribuição gamma, com parâmetros α e β. A função de densidade nesse caso é dada por

$$f(x) = \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-x/\beta}, \text{ para } x > 0.$$

A função geratriz de momento é dada pelo valor esperado

$$\begin{aligned} M_X(t) &= \int_{x=0}^{\infty} e^{xt} f(x) dx \\ &= \int_{x=0}^{\infty} e^{xt} \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-x/\beta} dx \\ &= \int_{x=0}^{\infty} \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-x(1/\beta - t)} dx \\ &= \frac{1}{\beta^\alpha \Gamma(\alpha)} \frac{\Gamma(\alpha)}{\left(\frac{1}{\beta} - t\right)^\alpha} = \frac{1}{(1-t\beta)^\alpha}. \end{aligned}$$

A integral acima existe para qualquer valor t, com $|t| < 1/\beta$.

Exemplo 4.17 (Função geratriz de momentos para uma variável aleatória com distribuição normal) Seja X uma variável aleatória com distribuição normal, com média μ e variância σ^2. A função de densidade nesse caso é dada por

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}, \text{ para } x \in (-\infty, +\infty).$$

A função geratriz de momentos é dada pelo valor esperado

$$\begin{aligned} M_X(t) &= \int_{-\infty}^{\infty} e^{xt} f(x) dx \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} e^{xt} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} dx \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} e^{(y+\mu)t} e^{-\frac{1}{2\sigma^2}y^2} dy \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} e^{\mu t} \int_{-\infty}^{\infty} e^{yt} e^{-\frac{1}{2\sigma^2}y^2} dy \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} e^{\mu t} \int_{-\infty}^{\infty} e^{-\frac{1}{2\sigma^2}(y^2 - 2\sigma^2 yt)} dy \\ &= e^{t\mu} e^{t^2\sigma^2/2}, \end{aligned}$$

onde, para encontrarmos o resultado acima, fizemos a mudança de variável $x = y + \mu$ e completamos o quadrado multiplicando e dividindo simultaneamente por $e^{t^2\sigma^2/2}$. Note que a função geratriz de momentos existe para qualquer valor $t > 0$.

De fato, se fizermos $\mu = 0$ e $\sigma^2 = 1$, temos que

$$M_X(t) = e^{t^2/2}.$$

Se X é uma variável aleatória normal com média μ e variância σ^2 e fizermos uma conta equivalente para a variável aleatória $(X-\mu)/\sigma$, como em $M_{(X-\mu)/\sigma}(t) = \int_{-\infty}^{\infty} e^{t(x-\mu)/\sigma} f(x)dx = e^{t^2/2}$, chegamos ao mesmo resultado. Adicionalmente, sabendo que $M_X(t) = e^{t^2/2}$ para uma variável aleatória com distribuição normal com média 0 e variância 1 e utilizarmos o Teorema 4.4, chegamos a conclusão que $M_{\mu+\sigma X}(t) = e^{t\mu}e^{t^2\sigma^2/2}$.

Portanto, podemos concluir que:

1) Se uma variável aleatória X tem distribuição normal com média μ e variância σ^2, então a variável aleatória $Z = (X-\mu)/\sigma$ tem distribuição normal com média 0 e variância 1.

2) Se uma variável aleatória X tem distribuição normal com média 0 e variância 1, então a variável aleatória $Z = \mu + \sigma X$ tem distribuição normal com média μ e variância σ^2.

Nota 4.2 Considere duas funções densidade de probabilidade conforme expressões abaixo.

$$f_1(x) = \frac{1}{\sqrt{2\pi}x} e^{(\log x)^2/2}, \quad \text{para } x \in (0,\infty),$$
$$f_2(x) = f_1(x)\big[1 + \sin(2\pi \log x)\big], \quad \text{para } x \in (0,\infty).$$

Se X_1 tem função de densidade de probabilidade $f_1(x)$ e X_2 tem função de densidade $f_2(x)$, então, pode-se mostrar que

$$\mathrm{E}\big[X_1^r\big] = \mathrm{E}\big[X_2^r\big], \quad \text{para todo } r = 0,1,2,3,\ldots.$$

Portanto, os momentos não-centrados de todas as ordens são iguais, apesar de as funções densidades serem diferentes. Concluímos então que, mesmo que duas variáveis aleatórias tenham todos os seus momentos iguais, não necessariamente elas possuem as mesmas distribuições. O teorema a seguir incorre no fato de que quando duas variáveis aleatórias possuem a mesma função geratriz de momentos, então elas necessariamente possuem a mesma distribuição.

Teorema 4.5 (Variáveis aleatórias com mesmas funções geratrizes de momentos) Considere duas variáveis aleatórias (discretas ou contínuas) X e Y, tais que, para algum valor $\epsilon > 0$, $M_X(t) = M_Y(t)$ para todo $t \in (-\epsilon, +\epsilon)$. Então, X e Y possuem a mesma distribuição.

Teorema 4.6 (Variáveis aleatórias com todos momentos iguais) Sejam X e Y duas variáveis aleatórias (discretas ou contínuas), com funções distribuições acumuladas $F_X(x)$ e $F_Y(y)$. Supõe-se que todos os momentos de X e Y existam. Se $|X| \leq K_X$ e $|Y| \leq K_Y$ para dois valores constantes K_X e K_Y (ou seja, X e Y possuem espaços amostrais limitados), então

$$F_X(u) = F_Y(u), \quad \text{para todo } u \in \Re,$$

se e somente se

$$\mathrm{E}\big[X_1^r\big] = \mathrm{E}\big[X_2^r\big], \quad \text{para todo } r = 0,1,2,3,\ldots.$$

Portanto, se X e Y são variáveis aleatórias limitadas e possuem todos os momentos iguais, então elas também possuem a mesma distribuição.

Vamos agora estender a definição de função geratriz de momentos para o caso multivariado. Seja então X um vetor aleatório com componentes individuais $X_1, \ldots, X_m$. A função geratriz de momentos conjunta tem expressão

$$\begin{aligned} M(t_1, \ldots, t_m) &= \mathrm{E}\left[e^{t_1 x_1 + \cdots + t_m x_m}\right] \\ &= \int_{x_m=-\infty}^{\infty} \cdots \int_{x_1=-\infty}^{\infty} e^{t_1 x_1 + \cdots + t_m x_m} f_{X_1,\ldots,X_m}(x_1, \ldots, x_m) dx_1 \ldots dx_m. \end{aligned}$$

Se $X_1, \ldots, X_m$ são variáveis aleatórias independentes, então

$$\begin{aligned} M(t_1, \ldots, t_m) &= \int_{x_m=-\infty}^{\infty} \cdots \int_{x_1=-\infty}^{\infty} e^{t_1 x_1} \ldots e^{t_m x_m} f_{X_1}(x_1) \ldots f_{X_m}(x_m) dx_1 \ldots dx_m \\ &= M_{X_1}(t_1) \ldots M_{X_m}(t_m). \end{aligned}$$

Para variáveis aleatórias discretas, a expressão para função geratriz de momentos conjunta passa a ser

$$\begin{aligned} M(t_1, \ldots, t_m) &= \mathrm{E}\left[e^{t_1 x_1 + \cdots + t_m x_m}\right] \\ &= \sum_{x_1 \in \mathbb{X}_1} \cdots \sum_{x_m \in \mathbb{X}_m} e^{t_1 x_1 + \cdots + t_m x_m} f_{X_1,\ldots,X_m}(x_1, \ldots, x_m), \end{aligned}$$

onde $\mathbb{X}_i$ é o espaço amostral para a variável aleatória X_i, $i = 1, \ldots, m$. Pode-se mostrar que, também para variáveis aleatórias discretas, quando elas são independentes, tem-se

$$M(t_1, \ldots, t_m) = M_{X_1}(t_1) \ldots M_{X_m}(t_m).$$

Exemplo 4.18 (Função geratriz de momentos para uma soma de variáveis aleatórias de Poisson) No Exemplo 3.7, vimos que a soma de três variáveis aleatórias de Poisson supostamente também tem uma distribuição de Poisson, cujo parâmetro λ corresponde à soma dos parâmetros λ's individuais de cada parcela compondo o somatório. Vamos agora provar, utilizando a conceito de função geratriz de momento, que essa suspeita de fato é verdadeira. Consideremos então m variáveis aleatórias X_j de Poisson, independentes, cada qual um parâmetro livre λ_j, $j = 1, \ldots, m$. Não necessariamente todos os λ_j são iguais.

Seja $Y = \sum_{j=1}^{m} X_j$. Então, a função geratriz de momentos de Y é dada por

$$\begin{aligned} M_Y(t) &= \mathrm{E}\left[e^{t(x_1+\cdots+x_m)}\right] \\ &= \int_{x_m=-\infty}^{\infty} \cdots \int_{x_1=-\infty}^{\infty} e^{t(x_1+\cdots+x_m)} f_{X_1,\ldots,X_m}(x_1,\ldots,x_m)dx_1\ldots dx_m \\ &= \int_{x_m=-\infty}^{\infty} \cdots \int_{x_1=-\infty}^{\infty} e^{tx_1}\ldots e^{tx_m} f_{X_1}(x_1)\ldots f_{X_m}(x_m)dx_1\ldots dx_m \\ &= M_{X_1}(t)\ldots M_{X_m}(t). \end{aligned}$$

Vamos agora calcular a função geratriz de momentos para uma variável aleatória X de Poisson individualmente.

$$\begin{aligned} M_X(t) &= \sum_{x=0}^{\infty} e^{tx}\frac{e^{-\lambda}\lambda^x}{x!} \\ &= \sum_{x=0}^{\infty} \frac{e^{-\lambda}(\lambda e^t)^x}{x!} = e^{-\lambda}\sum_{x=0}^{\infty}\frac{(\lambda e^t)^x}{x!} \\ &= e^{-\lambda}e^{\lambda e^t} = e^{\lambda(e^t-1)}, \quad \text{para todo } t \in \Re, \end{aligned}$$

onde a penúltima igualdade acima vem da expansão em série de Taylor da função exponencial.

Portanto, a função geratriz de momentos para a soma $Y = X_1 + \cdots + X_m$ tem expressão

$$\begin{aligned} M_Y(t) &= M_{X_1}(t)\ldots M_{X_m}(t) \\ &= e^{\lambda_1(e^t-1)}\ldots e^{\lambda_m(e^t-1)} \\ &= e^{(\lambda_1+\cdots+\lambda_m)(e^t-1)} = e^{\lambda_Y(e^t-1)}, \end{aligned}$$

onde $\lambda_Y = \lambda_1 + \cdots + \lambda_m$. Note que a função geratriz de momentos $M_Y(t) = e^{\lambda_Y(e^t-1)}$ é a mesma função geratriz de momentos de uma variável aleatória de Poisson, com parâmetro livre λ_Y. Portanto, de acordo com o Teorema 4.5, a soma Y possui distribuição de Poisson, com parâmetro livre $\lambda_1 + \cdots + \lambda_m$, conforme queríamos mostrar.

Exemplo 4.19 (Função geratriz de momentos para a soma de variáveis aleatórias normais independentes) Sejam X_i variáveis aleatórias normais independentes com média μ_i e variância σ_i^2, para $i = 1, \cdots, m$ e $Y = \sum_{i=1}^{m} X_i$, então, usando a independência das variáveis X_i e os resultados parciais dos Exemplos 4.17 e 4.18, chegamos a

$$M_Y(t) = \mathrm{E}\left[e^{t(x_1+\cdots+x_m)}\right] = \prod_{i=1}^{m} e^{t\mu_i}e^{t^2\sigma_i^2/2} = e^{t\sum_{i=1}^{m}\mu_i}e^{t^2\sum_{i=1}^{m}\sigma_i^2/2}.$$

Portanto, comparando com a função geratriz de momentos de uma variável aleatória normal com média μ e variância σ^2 apresentada no Exemplo 4.17, a variável aleatória $Y = \sum_{i=1}^{m} X_i$ é uma variável aleatória normal com média $\mu = \sum_{i=1}^{m} \mu_i$ e variância $\sigma^2 = \sum_{i=1}^{m} \sigma_i^2$.

Exemplo 4.20 (Função geratriz de momentos para uma soma de variáveis aleatórias com distribuição exponencial negativa) Considere agora uma sequência independente e identicamente distribuída de variáveis aleatórias, $X_1, \ldots, X_m$, com distribuição exponencial negativa. Todas as n variáveis possuem o mesmo parâmetro λ. Considere a soma $Y = \sum_{i=1}^{n} X_i$. Conforme vimos no Exemplo 4.18, devido ao fato de a soma envolver variáveis aleatórias independentes, a função geratriz de momentos de Y é o produto das funções geratrizes de momentos para cada variável individualmente. Portanto,

$$M_Y(t) = M_{X_1}(t) \ldots M_{X_m}(t) = \prod_{i=1}^{n} \frac{1}{1 - (t/\lambda)}.$$

Para chegar à função geratriz de momentos $M_X(t) = \frac{1}{1-(t/\lambda)}$ para cada variável X_i individualmente, pode-se utilizar diretamente a função geratriz de momentos para a variável aleatória gamma (vide Exemplo 4.16), juntamente com o fato de que uma variável aleatória exponencial negativa corresponde a uma variável gamma, com parâmetros $\alpha = 1$ e $\beta = 1/\lambda$ (vide Seção 3.3.2). Portanto, para a variável soma Y,

$$M_Y(t) = \frac{1}{(1 - (t/\lambda))^n}.$$

Note que, a função geratriz de momentos $M_Y(t)$ corresponde à uma função geratriz de momentos para uma variável aleatória gamma, com parâmetros $\alpha = n$ e $\beta = 1/\lambda$. Portanto, de acordo com o Teorema 4.5, concluímos que Y possui distribuição gamma. No Exercício 4.9 no final deste capítulo, o leitor poderá verificar, utilizando a função geratriz de momentos, que a soma de variáveis aleatórias independentes com distribuição qui-quadrada também possui uma distribuição qui-quadrada; o número de graus de liberdade da soma é igual à soma dos números de graus de liberdade de cada qui-quadrada compondo a soma.

Transformações de variáveis aleatórias

No Capítulo 3, discutimos o teorema de transformações de variáveis aleatórias contínuas no caso univariado. Nesta seção, esse teorema será estendido para o tratamento de variáveis multivariadas. Consideremos então as variáveis aleatórias $X_1, \ldots, X_m$, e seja $f_{X_1,\ldots,X_m}(x_1, \ldots, x_m)$ a função de densidade conjunta. Considere as variáveis aletórias $Y_1, \ldots, Y_m$, obtidas a partir de transformações das variáveis $X_1, \ldots, X_m$. Portanto, podemos escrever

$$
\begin{aligned}
Y_1 &= h_1(X_1, \ldots, X_m) &&= h_1(X), \\
Y_2 &= h_2(X_1, \ldots, X_m) &&= h_2(X), \\
&\cdots \\
Y_m &= h_m(X_1, \ldots, X_m) &&= h_m(X),
\end{aligned}
$$

onde $X = \begin{bmatrix} X_1 & X_2 & \ldots & X_m \end{bmatrix}'$ é o vetor de variáveis aleatórias, e $h_i(\cdot)$, $i = 1, \ldots, m$ são funções definidas em $\Re^m$. Podemos reescrever as expressões acima da forma

$$
Y = h(X),
$$

onde $Y = \begin{bmatrix} Y_1 & Y_2 & \ldots & Y_m \end{bmatrix}'$ e $h(X) = \begin{bmatrix} h_1(X) & h_2(X) & \ldots & h_m(X) \end{bmatrix}'$ são vetores coluna de dimensão $m \times 1$. O nosso objetivo é encontrar a função de densidade de probabilidade conjunta para o vetor Y. O teorema a seguir apresenta um resultado importante para resolver esse problema em uma grande variedade de situações. Esse teorema irá utilizar o jacobiano $J_h(x)$ da função $h(\cdot)$ e o jacobiano $J_{h^{-1}}(y)$ da função inversa $g(\cdot) = h^{-1}(\cdot)$, onde

$$
J_h(x) = \begin{bmatrix}
\frac{\partial h_1(x)}{\partial x_1} & \frac{\partial h_2(x)}{\partial x_1} & \cdots & \frac{\partial h_m(x)}{\partial x_1} \\
\frac{\partial h_1(x)}{\partial x_2} & \frac{\partial h_2(x)}{\partial x_2} & \cdots & \frac{\partial h_m(x)}{\partial x_2} \\
\cdots & \cdots & \cdots & \cdots \\
\frac{\partial h_1(x)}{\partial x_m} & \frac{\partial h_2(x)}{\partial x_m} & \cdots & \frac{\partial h_m(x)}{\partial x_m}
\end{bmatrix},
$$

$$
J_{h^{-1}}(y) = \begin{bmatrix}
\frac{\partial g_1(y)}{\partial y_1} & \frac{\partial g_2(y)}{\partial y_1} & \cdots & \frac{\partial g_m(y)}{\partial y_1} \\
\frac{\partial g_1(y)}{\partial y_2} & \frac{\partial g_2(y)}{\partial y_2} & \cdots & \frac{\partial g_m(y)}{\partial y_2} \\
\cdots & \cdots & \cdots & \cdots \\
\frac{\partial g_1(y)}{\partial y_m} & \frac{\partial g_2(y)}{\partial y_m} & \cdots & \frac{\partial g_m(y)}{\partial y_m}
\end{bmatrix}.
$$

Teorema 4.7 (Transformação de variáveis aleatórias contínuas multivariadas) Considere o problema de encontrar a função de densidade de probabilidade conjunta para a transformação $Y = h(X)$, conforme discutido acima. Seja A um conjunto aberto em $\Re^m$, tal que $\text{Prob}[x \in A] = 1$. Vamos supor que as condições seguintes são satisfeitas

(i) A função $h(\cdot)$, $h : A \mapsto h(A)$, com $h(A) \in \Re^m$, é bijetora (relação de um para um entre o vetor Y e o vetor X);
(ii) A função $h(\cdot)$ tem primeiras derivadas parciais contínuas;
(iii) A matriz jacobiana $J_h(x)$ da função $h(\cdot)$ tem determinante $|J_h(x)|$ maior que zero, para todo $x \in A$, e a matriz jacobiana $J_{h^{-1}}(y)$ da função inversa $h^{-1}(\cdot)$ tem determinante $|J_{h^{-1}}(y)|$ maior que zero, para

todo $y \in h(A)$.[6]

Então, a função de densidade de probabilidade conjunta do vetor Y é dada por

$$f_{Y_1,\ldots,Y_m}(y) = f_{X_1,\ldots,X_m}(h^{-1}(y))\left|J_{h^{-1}}(y)\right|, \tag{4.22}$$

onde $y = [y_1, \ldots, y_m]'$.

Exemplo 4.21 (Variável aleatória lognormal a partir de uma variável aleatória normal) Seja X uma variável aleatória com distribuição normal, com média μ e variância σ^2. Seja $Y = e^X$. Para encontrar a distribuição de Y, façamos $A = (-\infty, \infty)$, $h(A) = (0, \infty)$ e $g(y) = h^{-1}(y) = \log y$. Portanto,

$$J_{h^{-1}}(y) = \frac{1}{y},$$

e obtemos, usando Eq. 4.22,

$$\begin{aligned} f_Y(y) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big[-\frac{1}{2\sigma^2}(\log y - \mu)^2\Big]\frac{1}{y} \\ &= \frac{1}{y\sqrt{2\pi\sigma^2}} \exp\Big[-\frac{1}{2\sigma^2}(\log y - \mu)^2\Big], \quad \text{para } y > 0. \end{aligned}$$

Note que Y tem uma distribuição lognormal, com parâmetros μ e σ^2.

O problema apresentado no Exemplo 4.21, por ser um caso univariado, também poderia ter sido resolvido usando o Teorema 3.1.

Exemplo 4.22 (Transformação de um vetor aleatório bivariado) Seja $X = (X_1, X_2)$, com $f_X(x_1, x_2) = 4x_1x_2$, para $x_1 \in (0,1)$ e $x_2 \in (0,1)$ (note que X_1 e X_2 são independentes). Seja $Y_1 = X_1$ e $Y_2 = X_1X_2$. Temos então $A = \{(x_1, x_2) \in \Re^2 : 0 < x_1 < 1,\ 0 < x_2 < 1\}$, $h_1(x_1, x_2) = x_1$ e $h_2(x_1, x_2) = x_1x_2$. Portanto, $x_1 = y_1 = g_1(y)$, $x_2 = y_2/y_1 = g_2(y)$ e $x = g(y) = h^{-1}(y)$, onde $y = (y_1, y_2)$ e $x = (x_1, x_2)$. Para encontrar a imagem $h(A)$ da função $h(\cdot)$, temos

$$\begin{aligned} h(A) &= \{(y_1, y_2) \in \Re^2 : 0 < y_1 < 1,\ 0 < y_2/y_1 < 1\} \\ &= \{(y_1, y_2) \in \Re^2 : 0 < y_2 < y_1 < 1\}. \end{aligned}$$

[6] Por simplicidade, enunciamos esse teorema com a condição de que os valores dos determinantes sejam maiores que zero para todos os valores $x in A$ e $y \in h(A)$. De fato, os determinantes podem ser iguais a zero em subconjuntos de A e $h(A)$, desde que esses subconjuntos tenham probabilidade zero de acontecer.

A matriz jacobiana $J_{h^{-1}}(y)$ tem expressão

$$J_{h^{-1}}(y) = \begin{bmatrix} 1 & -y_2/y_1^2 \\ 0 & 1/y_1 \end{bmatrix},$$

e, portanto, $|J_{h^{-1}}(y)| = 1/y_1$. Logo, usando Eq. 4.22, a função de densidade conjunta $f_Y(y)$ de Y é dada por

$$f_Y(y) = 4y_1 \frac{y_2}{y_1}\frac{1}{y_1} = \frac{4y_2}{y_1}, \quad \text{para } 0 < y_2 < y_1 < 1.$$

Uma maneira simples de checar se a nova expressão é de fato uma função de densidade conjunta, basta integrar ao longo de todo o conjunto $\Re^2$, e verificar se a integral é igual a um.

Exemplo 4.23 (Variável aleatória com distribuição t-Student a partir de uma variável aleatória com distribuição normal e variável aleatória com distribuição qui-quadrada) Considere duas variáveis aleatórias X_1 e X_2 independentes, onde X_1 tem distribuição normal padrão e X_2 tem distribuição qui-quadrada com q graus de liberdade. Então sabemos que a função de densidade de probabilidade para a variável aleatória X_1 é dada por

$$f_{X_1}(x_1) = \frac{1}{\sqrt{2\pi}} e^{-x_1^2/2}, \quad \text{para } -\infty < x_1 < \infty$$

e a função de densidade de probabilidade de X_2 é dada por

$$f_{X_2}(x_2) = \frac{x_2^{(q-2)/2} e^{-x_2/2}}{2^{q/2}\Gamma(q/2)}, \quad \text{para } x_2 > 0.$$

Portanto, a função de densidade conjunta de $X = (X_1, X_2)$, devido à hipótese de independência, é o produto das funções densidades individuais,

$$f_{X_1,X_2}(x_1, x_2) = \frac{1}{\sqrt{2\pi}} e^{-x_1^2/2} \frac{1}{\Gamma(q/2)2^{q/2}} x_2^{(q-2)/2} e^{-x_2/2}$$

para $-\infty < x_1 < \infty$ e $x_2 > 0$. Considere agora as variáveis $Y_1 = \frac{X_1}{\sqrt{X_2/q}}$ e $Y_2 = X_2$. O conjunto A é dado por $A = (-\infty, \infty) \times (0, \infty)$, enquanto o conjunto imagem $h(A) = (-\infty, \infty) \times (0, \infty)$.[7] Invertendo a relação entre X e Y, obtemos

$$x_1 = g_1(y_1, y_2) = y_1\sqrt{y_2/q},$$
$$x_2 = g_2(y_1, y_2) = y_2.$$

[7] O conjunto $B_1 \times B_2$ corresponde ao produto cartesiano entre os conjuntos B_1 e B_2.

A matriz jacobiana da função $h(\cdot)$ inversa tem expressão

$$J_{h^{-1}}(y) = \begin{bmatrix} \sqrt{y_2/q} & \frac{y_1}{2\sqrt{qy_2}} \\ 0 & 1 \end{bmatrix},$$

cujo determinante é dado por $|J_{h^{-1}}(y)| = \sqrt{y_2/q}$. Utilizando o Teorema 4.7, chegamos à função de densidade de probabilidade conjunta para o vetor $Y = (Y_1, Y_2)$,

$$f_{Y_1,Y_2}(y_1, y_2) = f_{X_1,X_2}(y_1\sqrt{x_2/q}, y_2)\sqrt{y_2/q} = \frac{1}{\sqrt{2\pi}} \frac{1}{2^{q/2}\Gamma(q/2)} e^{-\frac{1}{2q}y_1^2 y_2} y_2^{(q-2)/2} e^{-y_2/2}\sqrt{y_2/q}.$$

Para encontrar a função de densidade de probabilidade marginal para a variável Y_1, basta integrar a função de densidade conjunta $f_Y(y)$ em y_2,

$$f_{Y_1}(y_1) = \int_{y_2=0}^{\infty} f_Y(y_1, y_2) dy_2.$$

Essa integral pode ser obtida diretamente, utilizando-se a definição da função gamma. Obtém-se então a distribuição de t-Student com q graus de liberdade dada pela Eq. (3.53).

Dessa forma, a luz do Teorema 4.7, o Exemplo 4.23 mostra como obter uma variável com distribuição t-Student a partir de duas variáveis independentes, uma variável com distribuição normal e uma outra distribuição qui-quadrada. Como foi já foi dito na Seção 3.5, variáveis com distribuição t-Student são muito úteis em modelos de regressão linear e, como veremos no Capítulo 8, o resultado apresentado no Exemplo 4.23 será essencial.

Exemplo 4.24 (Variável aleatória com distribuição F a partir de duas variáveis aleatórias com distribuição qui-quadrada) Considere duas variáveis aleatórias X_1 e X_2 independentes, com distribuição qui-quadrada. A primeira distribuição possui r graus de liberdade e a segunda possui q graus de liberdade. A função de densidade de probabilidade para uma variável aleatória qui-quadrada com r graus de liberdade e dada por

$$f_X(x) = \frac{x^{(r-2)/2}e^{-x/2}}{2^{r/2}\Gamma(r/2)}, \quad \text{para } x > 0.$$

A função de densidade conjunta de $X = (X_1, X_2)$, devido à hipótese de independência, é o produto das funções densidades individuais,

$$f_X(x_1, x_2) = \frac{x_1^{(r-2)/2} x_2^{(q-2)/2} e^{-(x_1+x_2)/2}}{2^{(r+q)/2}\Gamma(q/2)\Gamma(r/2)},$$

para $x_1, x_2 > 0$. Considere agora as variáveis $Y_1 = \frac{X_1/r}{X_2/q}$ e $Y_2 = X_2$. O conjunto A é dado por $A = (0,\infty) \times (0,\infty)$, enquanto o conjunto imagem $h(A) = (0,\infty) \times (0,\infty)$. Invertendo a relação entre X e Y, obtemos

$$x_1 = g_1(y_1, y_2) = \frac{r y_1 y_2}{q},$$

$$x_2 = y_2.$$

A matriz jacobiana da função $h(\cdot)$ inversa tem expressão

$$J_{h^{-1}}(y) = \begin{bmatrix} \frac{r y_2}{q} & \frac{r y_1}{q} \\ 0 & 1 \end{bmatrix},$$

e o determinante é dado por $|J_{h^{-1}}(y)| = r y_2/q$. Utilizando o Teorema 4.7, chegamos à função de densidade de probabilidade conjunta para o vetor $Y = (Y_1, Y_2)$,

$$\begin{aligned} f_Y(y_1, y_2) &= \left(\frac{r}{q}\right)^{(r-2)/2} \left(\frac{y_1^{(r-2)/2} y_2^{(r+q-4)/2} e^{-(1+r y_1/q) y_2/2}}{\Gamma(\frac{r}{2})\Gamma(\frac{q}{2}) 2^{(r+q)/2}} \right) \left(\frac{r y_2}{q}\right) \\ &= \left(\frac{r}{q}\right)^{r/2} \left(\frac{y_1^{(r-2)/2} y_2^{(r+q)/2-1} e^{-(1+r y_1/q) y_2/2}}{\Gamma(\frac{r}{2})\Gamma(\frac{q}{2}) 2^{(r+q)/2}} \right), \quad \text{para } y_1, y_2 \in (0,\infty). \end{aligned}$$

Para encontrar a função de densidade de probabilidade marginal para a variável Y_1, basta integrar a função de densidade conjunta $f_Y(y)$ em y_2,

$$f_{Y_1}(y_1) = \int_{y_2=0}^{\infty} f_Y(y_1, y_2) dy_2.$$

Essa integral pode ser obtida diretamente, utilizando-se a definição da função gamma. Obtém-se então

$$f_{Y_1}(y_1) = \left(\frac{r}{q}\right)^{r/2} \frac{y_1^{(r-2)/2}}{\Gamma(\frac{r}{2})\Gamma(\frac{q}{2})} \frac{\Gamma(\frac{r+q}{2})}{\left(\frac{r}{q} y_1 + 1\right)^{(r+q)/2}}, \quad \text{para } y_1 > 0. \tag{4.23}$$

A função de densidade de probabilidade na Eq. (4.23) corresponde à função de densidade de uma variável aleatória com distribuição F, com r graus de liberdade no numerador e q graus de liberdade no denominador. Portanto, analogamente ao Exemplo 4.23, as derivações apresentadas no Exemplo 4.24 nos levam a concluir que o quociente entre um primeira variável aleatória qui-quadrada (dividida pelo seu número de graus de liberdade) e uma segunda variável aleatória qui-quadrada (também dividida pelo seu número de graus de liberdade) resulta em uma variável aleatória F. Esse resultado vale quando as duas variáveis aleatórias qui-quadradas são independentes.

Justamente pelo fato de a variável aleatória F resultar do quociente entre variáveis aleatórias qui-quadradas, ela é muito utilizada em modelos lineares, tanto em estatística quanto em econometria, como veremos no Capítulo 8. Similarmente ao caso da variável aleatória normal e da variável aleatória t-Student, o percentis da variável aleatória F, para diferentes combinações de r e q, estão amplamente disponíveis em livros de estatística e em softwares comerciais, como, por exemplo, a planilha Excel.

Para uma variável aleatória X com distribuição F, com r graus de liberdade no numerador e q graus de liberdade no denominador, pode-se mostrar que o seu valor esperado é

$$\mathrm{E}[X] = \frac{q}{q-2},$$

e sua variância é dada por

$$\mathrm{Var}[X] = \frac{2q^2(r+q-2)}{r(q-2)^2(q-4)}.$$

Além desses dois resultados interessantes apresentados nos Exemplos 4.23 e 4.24, pode-se mostrar também que a raiz quadrada de uma variável aleatória F, com 1 grau de liberdade no numerador, e q graus de liberdade no denominador, tem distribuição t-Student com q graus de liberdade (vide Exercício 4.13).

4.7 Estrutura de dependência via cópulas

Em termos práticos, quando duas ou mais variáveis aleatórias possuem distribuições marginais normais, os pesquisadores e analistas, na área de mensuração e gerenciamento de risco, por exemplo, utilizam-se de distribuições normais multivariadas para lidar com a modelagem da estrutura de dependência entre as variáveis aleatórias marginais. Por exemplo, em análise de risco de mercado, muitas vezes o analista está interessado na estrutura de dependência entre os retornos de diferentes papeis ou diferentes segmentos da carteira de investimentos. Em risco operacional, as distribuições de perdas agregadas são modeladas para diferentes segmentos de perdas (por exemplo, linhas de negócio ou eventos de perda). O interesse seguinte pode residir em como agregar essas perdas por segmento em uma distribuição de perdas totais para toda a instituição, por exemplo. Nesse caso, dado que as distribuições de perdas são conceitualmente positivas (ou não negativas), a ideia de modelar a estrutura de dependência entre as perdas de cada segmento individualmente utilizando-se uma distribuição normal multivariada não é mais adequada.

Portanto, surge a necessidade de trabalharmos com ferramentas diferentes para compôr a estrutura de dependência entre variáveis aleatórias não normais. Nas últimas décadas, tem crescido bastante a utilização de modelos de cópulas para modelagem da dependência entre variáveis aleatórias em finanças. O leitor pode recorrer, por exemplo, a Nelsen (1998), McNeil, Frey e Embrechts (2005) e Fredheim (2008) para maiores detalhes. A partir de agora, faremos uma breve introdução aos modelos de cópulas, e como elas podem ser aplicadas no cálculo da distribuição agregada de perdas operacionais individuais, por exemplo, a partir

das distribuições de perdas operacionais de cada segmento de perda. As cópulas na verdade estendem a ideia de correlação entre variáveis aleatórias normais, para o caso onde as distribuições marginais não são normais.

Em termos gerais imagine que tenhamos n variáveis aleatórias $X_1, \ldots, X_n$. Em muitas situações, estamos interessados em encontrar a função de distribuição conjunta $F(x_1, \ldots, x_n)$ a partir das distribuições marginais $F_{X_1}(x_1), \ldots, F_{X_n}(x_n)$, que são supostas não normais. Podemos definir formalmente a função **cópula** C como uma função de distribuição acumulada conjunta de forma que

1. $C : [0,1]^n \mapsto [0,1]$;

2. C é crescente e contínua em todo $[0,1]^n$;

3. C possui funções distribuições marginais C_k, tais que $C_k(u_k) = C(1, \ldots, 1, u_k, 1, \ldots, 1) = u_k$, para todo $u_k \in [0,1]$;

De fato, grande parte do sucesso dos modelos de cópulas se deve ao teorema a seguir:

Teorema 4.8 (Sklar) Considere uma função de distribuição acumulada $F_{X_1,\cdots,X_K}(x_1, \cdots, x_K)$ com marginais $F_{X_1}(x_1), \cdots F_{X_K}(x_K)$ contínuas. Então existe uma única cópula C tal que

$$F(x_1, \cdots, x_d) = C(F_{X_1}(x_1), \cdots, F_{X_K}(x_K)). \tag{4.24}$$

Em termos práticos, o Teorema 4.8 nos diz que toda distribuição acumulada conjunta pode ser expressa a partir de suas distribuições marginais e de uma cópula que as une. De fato, em todas aplicações de cópulas, deve-se partir do princípio que tanto as distribuições marginais são conhecidas, como também a cópula que as une.

Apresentamos agora um corolário muito útil do Teorema 4.8 e que permite a construção de cópulas:

Corolário 4.1 A única cópula do Teorema 4.8 é dada por

$$C(u_1, \cdots, u_K) = F(F_{X_1}^{-1}(u_1), \cdots, F_{X_K}^{-1}(u_k)), \tag{4.25}$$

onde u_k, $k = 1, \cdots, K$ estão definidos em $[0,1]$ que é a imagem de cada marginal F_{X_k}, $k = 1, \cdots, K$.

Exemplo 4.25 (Existência de cópula) Considere que duas variáveis aleatórias possuem distribuição conjunta dada por

$$F_{X_1,X_2}(x_1, x_2) = \max\left(1 - \sum_{k=1}^{2} \exp(-\lambda_k x_k), 0\right)$$

com marginais dadas por

$$F_{X_k}(x_k) = 1 - \exp(-\lambda_k x_k),\ k = 1, 2.$$

Note que

$$F_{X_k}^{-1}(u_k) = -\frac{\ln(1-u_k)}{\lambda_k},\ k = 1, 2.$$

Usando Eq. (4.24), sabemos que

$$C(F_{X_1}(x_1), F_{X_2}(x_2)) = C(1 - \exp(-\lambda_1 x_1), 1 - \exp(-\lambda_2 x_2)) = \max\left(1 - \sum_{k=1}^{2} \exp(-\lambda_k x_k), 0\right)$$

Por outro lado, usando Eq. (4.25), temos

$$C(u_1, u_2) = F(F_{X_1}^{-1}(u_1), F_{X_2}^{-1}(u_2)) = F(-\ln(1-u_1)/\lambda_1, -\ln(1-u_2)/\lambda_2)$$

Finalmente, utilizando a definição de F_{X_1,X_2}, chegamos a

$$C(u_1, u_2) = \max\Big(u_1 + u_2 - 1, 0\Big)$$

.

que é um exemplo de cópula. Essa cópula é conhecida como **cópula limitante inferior de Fréchet-Hoeffding**[8].

Assim como a cópula apresentada no Exemplo 4.25, as cópulas $C(u_1, \ldots, u_n) = u_1 \ldots u_n$, conhecida como **cópula da independência**, e $C(u_1, \ldots, u_n) = \min(u_1, \ldots, u_n)$, conhecida como **cópula limitante superior de Fréchet-Hoeffding**, são modelos de cópulas simples, pois elas já especificam a estrutura de dependência entre as variáveis aleatórias especificadas por suas distribuições marginais. Em particular, pode-se mostrar que cópula limitante inferior de Fréchet-Hoeffding especifica uma correlação perfeitamente negativa entre as variáveis aleatórias especificadas pelas distribuições marginais, a cópula da independência, como o próprio nome sugere, especifica independência entre as variáveis aleatórias e a cópula limitante superior de Fréchet-Hoeffding especifica uma estrutura de correlação perfeitamente positiva.

[8] É válido notar que, em geral, essa função não define uma cópula para mais de duas dimensões. Para detalhes, vide Joe (1997).

Em situações práticas, prefere-se cópulas em que a estrutura de dependência seja estimada. Por exemplo, nas aplicações em análise de risco, utilizam-se funções cópulas com formas paramétricas conhecidas, como por exemplo, a cópula normal, a cópula t-Student, as cópulas de Gumbel δ e α, ou a cópula de Frank. No contexto bivariado, a cópula normal é dada por

$$C(u_1, u_2) = \int_{-\infty}^{\Phi^{-1}(u_1)} \int_{-\infty}^{\Phi^{-1}(u_2)} \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left[-\frac{v_1^2 - 2\rho v_1 v_2 + v_2^2}{2(1-\rho^2)}\right] dv_1 dv_2, \tag{4.26}$$

onde $\Phi(\cdot)$ é a função de distribuição acumulada de uma variável aleatória normal padronizada e $\Phi^{-1}(\cdot)$ é a correspondente função de distribuição acumulada inversa. O escalar $\rho \in (0,1)$ é o coeficiente de correlação linear entre u_1 e u_2.

A cópula t-Student com ν graus de liberdade para o caso bivariado é definida por

$$C(u_1, u_2) = \int_{-\infty}^{t_\nu^{-1}(u_1)} \int_{-\infty}^{t_\nu^{-1}(u_2)} \frac{1}{2\pi\sqrt{1-\rho^2}} \left[1 + \frac{v_1^2 - 2\rho v_1 v_2 + v_2^2}{2(1-\rho^2)}\right]^{\frac{\nu+2}{2}} dv_1 dv_2, \tag{4.27}$$

onde $t_\nu(\cdot)$ é função de distribuição acumulada para distribuição t-Student com ν graus de liberdade e $t_\nu^{-1}(\cdot)$ é a correspondente função de distribuição acumulada inversa.

A cópula de Gumbel δ tem expressão

$$C(u_1, u_2) = \exp\left[-\left[(-\log u_1)^\delta + (-\log u_2)^\delta\right]^{1/\delta}\right], \tag{4.28}$$

onde δ é um parâmetro que regula o grau de dependência, onde $\delta \in [1, \infty)$. Quando δ aumenta, o grau de dependência positiva aumenta, sendo que $\delta = 1$ corresponde a variáveis aleatórias independentes. A cópula de Gumbel α é definida por

$$C(u_1, u_2) = u_1 u_2 \exp\left[\frac{\alpha(\log u_1)(\log u_2)}{\log(u_1 u_2)}\right], \tag{4.29}$$

onde α é o parâmetro de dependência entre as variáveis aleatórias X_1 e X_2, com $\alpha \in [0,1]$. Quando α aumenta em direção a $\alpha = 1$, o grau de dependência positiva aumenta. Quando α diminui para zero, as variáveis aleatórias tornam-se mais independentes, e $\alpha = 0$ corresponde à hipótese de independência entre X_1 e X_2. Finalmente, a cópula de Frank tem expressão

$$C(u_1, u_2) = \frac{1}{\delta}\left[\log\left[1 - \exp\delta - \left[1 - \exp(\delta u_1)\right]\left[1 - \exp(\delta u_2)\right]\right] - \log\left[1 - \exp\delta\right]\right], \tag{4.30}$$

onde δ é o parâmetro que regula a dependência entre as variáveis X_1 e X_2, com $\delta \in \Re$.

As cópulas descritas acima podem ser utilizadas para compôr qualquer par de variáveis aleatórias. Por exemplo, podemos utilizar tanto a cópula normal quanto a cópula t-Student ou a cópula de Gumbel para encontrar a distribuição conjunta entre uma distribuição marginal lognormal e uma distribuição marginal Weibull. Um caso especial ocorre quando, tanto as distribuições marginais são normais, quanto a cópula para agregação. Nesse caso, a distribuição conjunta é uma distribuição normal multivariada.

Especificamente para as cópulas normais e t-Student, é possível trabalhar com versões multivariadas, possibilitando a agregação de perdas operacionais, por exemplo, para vários segmentos de perdas (células da matriz linhas de negócios por evento de perda) simultaneamente. As demais cópulas citadas acima não apresentam versões multivariadas, mas apenas versões bivariadas, implicando na necessidade de agregar as perdas operacionais aos pares, para finalmente chegar na perda agregada final. Por exemplo, imagine que queiramos somar as perdas operacionais S_1, S_2, ..., S_n, dos segmentos $k = 1, 2, 3, \ldots, n$. Se quisermos utilizar a cópula de Gumbel, por exemplo, teremos que agregar primeiramente os segmentos 1 e 2, obtendo a perdas $S_{1,2}$. Em seguida, podemos agregar os segmentos $S_{1,2}$ e S_3, obtendo $S_{1,2,3}$, e assim por diante. Se quisermos agregar os n segmentos de perdas via cópulas normais ou cópulas t-Student, é possível obter $S_{1,2,3,\ldots,n}$ diretamente, a partir dos n segmentos de perdas.

A cópula multivariada normal possui função cópula

$$C(u_1, \ldots, u_n) = \Phi_\Sigma\Big(\Phi^{-1}(u_1), \ldots, \Phi^{-1}(u_n)\Big), \tag{4.31}$$

onde Σ é uma matriz de correlações, que indica o grau de dependência entre cada um dos n pares de segmentos de perda sendo agregados. A função Φ refere-se a uma distribuição normal multivariada com média zero, todas as variâncias iguais a 1, e matriz de variância-covariância igual Σ (lembrando que, quando as variâncias são iguais a 1, a matriz de variância-covariância é igual à matriz de correlações). Conforme visto acima, a função Φ^{-1} é a função de distribuição acumulada inversa de uma variável aleatória normal padronizada (média zero e variância unitária). Quando $n = 2$, a cópula na Eq. (4.31) transforma-se na Eq. (4.26).

Para descrever a cópula t-Student, considere inicialmente um vetor X com distribuição t-Student n-variada, com ν graus de liberdade, média μ (se $\nu > 1$) e matriz de variância-covariância $\frac{\nu}{\nu-2}\Sigma$ (para $\nu > 2$).[9] O vetor aleatório X pode ser representado como

$$X = \mu + \frac{\sqrt{\nu}}{\sqrt{S}} Z,$$

onde μ é um vetor de dimensão $n \times 1$, a variável aleatória univariada S tem distribuição χ^2_ν (distribuição qui-quadrada, com ν graus de liberdade), e o vetor aleatório Z tem distribuição normal multivariada $N(\mathbf{0}, \Sigma)$. O vetor Z e a variável S são independentes. A cópula t-Student multivariada pode ser

[9] Quando $\nu \leq 2$, a matriz de variância-covariância não existe.

analiticamente representada como

$$C^t_{\nu,R} = t^n_{\nu,R}\Big(t^{-1}_\nu(u_1), \ldots, t^{-1}_\nu(u_n)\Big),$$

onde R é uma matriz de correlação $n \times n$. Uma maneira de definir R, por exemplo, é fazer cada elemento $R_{i,j} = \Sigma_{i,j}/\sqrt{\Sigma_{i,i}\Sigma_{j,j}}$, para $i, j \in \{1, 2, \ldots, n\}$, e Σ uma matriz de variância-covariância $n \times n$. A função $t^n_{\nu,R}$ corresponde à distribuição acumulada de uma variável aleatória multivariada t-Student, com ν graus de liberdade, e matriz de variância-covariância igual $\frac{\nu}{\nu-2}R$. Conforme definido acima, $t_\nu(\cdot)$ e t^{-1}_ν denotam a função de distribuição acumulada e a função de distribuição acumulada inversa de uma variável aleatória t-Student univariada com ν graus de liberdade.

Em geral as cópulas t-Student, de Gumbel δ e de Gumbel α implicam em uma maior dependência entre valores nos extremos (caudas) das distribuições. Isso a princípio pode sugerir a superioridade dessas duas cópulas em relação à cópula normal quando trabalhamos com análise de risco em geral, onde os eventos extremos são os mais importantes de serem modelados. A cópula de Frank aparenta ser mais apropriada quando a estrutura de dependência é simétrica (com dependência positiva quando $\delta < 0$, e dependência negativa quando $\delta > 0$). Isso sugere que cópula de Frank não seria tão apropriada em modelos de risco operacional, haja visto estarmos trabalhando com distribuições em geral bastante assimétricas. No entanto, a superioridade de uma determinada cópula em relação às demais vai depender do real processo gerador de dados. Na prática, especificamente para risco operacional, devido ao fato de não termos dados disponíveis suficientes para testar estatisticamente a superioridade de uma determinada cópula, deixamos ao analista a possibilidade de escolher entre uma das cópulas, por experiência, da mesma maneira que a determinação dos valores das correlações. Para risco de mercado, onde há informações disponíveis suficientes para estimar os parâmetros de cada tipo paramétrico de cópula, é possível testar qual a família paramétrica de cópulas que mais se adequa aos dados observados.

Para aplicações em risco operacional, a utilização das expressões acima para a determinação da distribuição conjunta a partir das distribuições marginais e das funções de cópula pode não ser possível analiticamente. Isso acontece basicamente pelo fato de não termos disponível a forma analítica para a distribuição de perdas agregadas de cada segmento de perda operacional específico. Felizmente, podemos recorrer novamente a simulações de Monte Carlo para encontrar (ou pelo menos estimar) a distribuição conjunta entre as variáveis dependentes.

Os passos a seguir esboçam o processo de simulações para gerar uma amostra de tamanho L de vetores n-dimensionais a partir da distribuição conjunta entre $X_1, \ldots, X_n$, com base em uma função cópula normal, com matriz de correlação Σ.

(1) Seja Σ a matriz de correlação de dimensão $n \times n$, onde os elementos da diagonal principal são 1 e os elementos fora da diagonal principal correspondem às respectivas correlações (o elemento $\Sigma_{i,j}$ é a correlação entre as variáveis X_i e X_j). Encontre a decomposição de Cholesky da matrix Σ, ou seja, encontre uma matriz quadrada P tal que $PP' = \Sigma$.

(2) Gere n valores $z_1, \ldots, z_n$ de variáveis aleatórias normais padronizadas independentes.

(3) Encontre o vetor $[v_1 \ \ldots \ v_n]' = P \times [z_1 \ \ldots \ z_n]'$. Nesse caso, o vetor $[v_1, \ldots, v_n]'$ corresponde a uma retirada aleatória de uma distribuição normal multivariada com matriz de covariância Σ.

(4) Calcule os valores $u_i = \Phi^{-1}(v_i)$, para $i = 1, \ldots, n$.

(5) Finalmente, sejam $F_{X_1}(\cdot), \ldots, F_{X_n}(\cdot)$ as funções distribuição acumulada de cada uma das variáveis X_1, $\ldots$, X_n. Encontre então os valores $x_1 = F_{X_1}^{-1}(u_1), \ldots, x_n = F_{X_n}^{-1}(u_n)$. O vetor $[x_1, \ldots, x_n]'$ corresponde a uma retirada aleatória com distribuições marginais $F_{X_1}(\cdot), \ldots, F_{X_n}(\cdot)$ e estrutura de dependência dada pela cópula normal. Os valores $x_1, \ldots, x_n$ são justamente os valores que queríamos gerar.

(6) Repetimos os passos (2) a (5) um número $L - 1$ de vezes, e encontramos L retiradas de vetores n-dimensionais com estrutura de dependência especificada pela cópula normal, e com distribuições marginais $F_{X_1}(\cdot), \ldots, F_{X_n}(\cdot)$.

Nas aplicações em risco operacional, as variáveis aleatórias $X_1, X_2, \ldots, X_n$ correspondem às perdas em cada segmento de perda específico (células da matriz linha de negócios versus evento de perda). Nesse caso, não necessariamente estamos interessados em toda a distribuição conjunta entre $X_1, \ldots, X_n$, mas somente na soma $S = X_1 + \cdots + X_n$, onde S corresponderia à perda operacional total da instituição. Para encontrar a distribuição da soma S, basta incluir dois passos adicionais no esquema de simulação anterior:

(7) Para cada um dos L vetores gerados no passo (6), encontre a soma $s = x_1 + \cdots + x_n$, obtendo L valores para caracterizar a distribuição da soma S.

(8) A partir dos L valores gerados para S, podemos estimar a perda média $E[S]$, os percentis de S, e portanto os diversos VaR's etc.

A utilização de cópulas para análise de risco em geral tem crescido bastante nos últimos anos. A rapidez com que novos instrumentos financeiros são criados, com grau de complexidade cada vez maior, gera uma série de novas oportunidades de aplicações estatísticas e econométricas a dados reais. O leitor interessado pode recorrer à bibliografia dada neste livro, como também aos diversos artigos disponíveis em periódicos acadêmicos ou disponíveis livremente na internet.

4.8 Exercícios

Exercício 4.1 Sejam X e Y variáveis aleatórias quaisquer, sejam a e b duas constantes reais. Mostre que a variância da soma ponderada $aX + bY$ tem expressão

$$\text{Var}[aX + bY] = a^2\text{Var}[X] + b^2\text{Var}[Y] + 2ab\text{Cov}[X, Y].$$

Exercício 4.2 Estenda o resultado do exercício anterior para o caso mais geral, onde temos n variáveis aleatórias e n constantes reais, mostrando que

$$\text{Var}[a_1X_1 + a_2X_2 + \cdots + a_nX_n] = \sum_{i=1}^{n} a_i^2 \text{Var}[X_i] + 2\sum\sum_{i<j} a_i a_j \text{Cov}[X_i, X_j]. \tag{4.32}$$

Exercício 4.3 Sejam X e Y duas variáveis aleatórias independentes, cada qual com distribuição exponencial negativa com parâmetro $\lambda = 5$. Ou seja, para ambas a função de densidade de probabilidade tem expressão

$$f(x) = \frac{1}{5}e^{-x/5}, \text{ para } x > 0.$$

Seja a variável aleatória $W = X + XY$. Determine a variância $\text{Var}\big[\text{E}[W/Y]\big]$.

Exercício 4.4 Seja $X_1, \ldots, X_n$ uma sequência de variáveis aleatórias. Seja V uma matriz diagonal, cujo i-ésimo elemento da diagonal principal, $V_{i,i}$ é igual à variância de X_i, $i = 1, \ldots, n$. Seja R a matriz de correlações da sequência e seja Σ a matriz de variância-covariância. Mostre que

$$\Sigma = V^{1/2} \times R \times V^{1/2}.$$

Exercício 4.5 Sejam X, Y e Z três variáveis aleatórias, com distribuição conjunta qualquer. Seja $g(x, y, z) = X^2Y^2Z^2$, para qualquer valor real de x, y e z.

(i) Mostre que $\text{E}\big[g(X, Y, Z)\big] \geq g\big(\text{E}[X], \text{E}[Y], \text{E}[Z]\big)$;
(ii) O que acontece com o sinal da desigualdade do último item anterior para uma nova função $g(\cdot)$ definida por $g(x, y, z) = X^{1/3}Y^{1/3}Z^{1/3}$.

Exercício 4.6 Sejam X e Y duas variáveis aleatórias contínuas, com distribuição bivariada, com função de densidade de probabilidade conjunta

$$f_{X,Y}(x, y) = A(x^2 + y^2 + 2xy^{1/3}), \text{ para } 0 < x < 3,\ 0 < y < 2.$$

(1) Determine o valor da constante A para que a função $f_{X,Y}(x, y)$ seja uma função de densidade conjunta.
(2) Escreva a função de distribuição acumulada conjunta $F_{X,Y}(x, y)$ para X e Y.
(3) Determine a covariância de entre X e Y.
(4) Determine a função de densidade marginal de X.
(5) Determine a função de densidade marginal de Y.

(6) Determine a função de densidade condicional de X, dado $Y = 1$.
(7) Determine a função de densidade condicional de X, dado $Y = 1.5$.
(8) Determine a função de densidade condicional de X, dado Y, como uma função de Y.
(9) Determine a função de densidade condicional de Y, dado $X = 1$.
(10) Determine a função de densidade condicional de Y, dado $X = 2$.
(11) Determine a função de densidade condicional de Y, dado X, como uma função de X.
(12) Determine o valor esperado de $X \times Y$.
(13) Determine a variância de X.
(14) Determina a variância de Y.
(15) Determine o coeficiente de correlação entre X e Y.
(16) Determine o momento condicional $\mathrm{E}[X/Y = 0.5]$.
(17) Determine a variância condicional $\mathrm{Var}[Y/X = 1.5]$.

Exercício 4.7 Sejam X e Y duas variáveis aleatórias discretas, com distribuição bivariada, com função de frequência de probabilidade conjunta

$$f_{X,Y}(x,y) = A\Big[\frac{1}{x^2} + \frac{1}{x}\frac{1}{y}\Big], \text{ para } x \in \{1,2,3,4\},\ y \in \{2,3,4\}.$$

(1) Determine o valor da constante A para que a função $f_{X,Y}(x,y)$ seja uma função de frequência conjunta.
(2) Escreva a função de distribuição acumulada conjunta $F_{X,Y}(x,y)$ para X e Y.
(3) Determine a covariância de entre X e Y.
(4) Determine a função de densidade marginal de X.
(5) Determine a função de densidade marginal de Y.
(6) Determine a função de densidade condicional de X, dado $Y = 2$.
(7) Determine a função de densidade condicional de X, dado $Y = 3$.
(8) Determine a função de densidade condicional de X, dado Y, como uma função de Y.
(9) Determine a função de densidade condicional de Y, dado $X = 1$.
(10) Determine a função de densidade condicional de Y, dado $X = 2$.
(11) Determine a função de densidade condicional de Y, dado X, como uma função de X.
(12) Determine o valor esperado de $X \times Y$.
(13) Determine a variância de X.
(14) Determina a variância de Y.
(15) Determine o coeficiente de correlação entre X e Y.
(16) Determine o momento condicional $\mathrm{E}[X/Y = 4]$.
(17) Determine a variância condicional $\mathrm{Var}[Y/X = 4]$.

Exercício 4.8 Sejam $X_1, \ldots, X_n$ uma sequência de variáveis aleatórias independentes e identicamente distribuídas, onde X_i tem função de densidade $f(x)$, $i = 1, \ldots, n$. Considere uma função multivariada $g : \Re^n \to \Re$, onde $g(\cdot)$ pode ser escrita como $g(x_1, \ldots, x_n) = h_1(x_1) \times h_2(x_2) \times \cdots \times h_n(x_n)$. Mostre que $E[g(X_1, \ldots, X_n)] = E[h_1(X_1)] \times E[h_2(X_2)] \times \cdots \times E[h_n(X_n)]$.

Exercício 4.9 Mostre que:
(1) A soma de m variáveis aleatórias binomiais Bin(n, p) independentes também é uma variável aleatória binomial. Nesse caso, determine os dois parâmetros da variável aleatória correspondente à soma.
(2) A soma de n variáveis aleatórias qui-quadradas independentes, cada qual com r graus de liberdade, também é uma variável qui-quadrada. Nesse caso, determine o número de graus de liberdade da variável resultante da soma.

Exercício 4.10 Seja Z_1 uma variável aleatória normal padronizada. Seja $Y_1 = X_1^2$ a variável aleatória correspondente ao quadrado da normal padronizada. Pode-se mostrar que Y_1 tem distribuição qui-quadrada com um grau de liberdade. Agora considere uma sequência de variáveis normais padronizadas $Z_1, \ldots, Z_n$ independentes. Responda:
(1) Escreva a função de densidade conjunta de $Z_1, \ldots, Z_n$;
(2) Seja $W = Z_1^2 + Z_2^2 + \cdots + Z_n^2$. Qual a distribuição de W?
(3) Escreva a média e a variância de W como uma função de n.
(4) Para $n = 5$, qual a probabilidade de $W < 4$?
(5) Para $n = 3$, qual a probabilidade de W estar entre 2 e 5?
(6) Para $n = 10$, qual o valor x para o qual Prob$[W > x] = 0.05$?
(7) Para $n = 15$, qual o valor x para o qual Prob$[W > x] = 0.01$?

Dica: Para responder aos ítens (3) a (6), você pode utilizar tanto uma tabela comumente encontrada no Apêndice de livros de estatística introdutória, como também alguma função apropriada no Excel.

Exercício 4.11 Considere uma variável aleatória normal bivariada, com componentes X e Y. A média dessa variável aleatória é o vetor $\mu = [3, 5]'$. A variância de X é igual a $\sigma_X^2 = 9$, a variância de Y é igual a $\sigma_Y^2 = 16$ e o coeficiente de correlação é igual a 0.5. Responda:

(1) Escreva a matriz de variância-covariância da distribuição normal multivariada.
(2) Escreva a função de densidade marginal para X.
(3) Escreva a função de densidade marginal para Y.
(4) Escreva a função de densidade condicional para X, dado que $Y = 4$. A distribuição condicional de X nesse caso é normal? Com que média? Com que variância?
(5) Escreva a função de densidade condicional para Y, dado que $X = 2.5$. A distribuição condicional de Y nesse caso é normal? Com que média? Com que variância?

Exercício 4.12 Seja X uma variável aleatória, com distribuição F com r graus de liberdade no numerador e q graus de liberdade no denominador. A função de densidade de probabilidade de X é dada por

$$f(x) = \left(\frac{r}{q}\right)^{r/2} \frac{x^{(r-2)/2}}{\Gamma(\frac{r}{2})\Gamma(\frac{q}{2})} \frac{\Gamma(\frac{r+q}{2})}{\left(\frac{r}{q}x + 1\right)^{(r+q)/2}}, \quad \text{para } x > 0. \tag{4.33}$$

Responda:

(i) Escreva a função de densidade de probabilidade para uma variável aleatória Y, com distribuição F, com um grau de liberdade no denominador e q graus de liberdade no denominador.
(ii) Seja $W = \sqrt{Y}$. Mostre que W tem distribuição t-Student com q graus de liberdade.

Exercício 4.13 Para uma variável aleatória X com função de densidade na Eq. (4.33), responda:

(i) Mostre que $\mathrm{E}[X] = \frac{q}{q-2}$.
(ii) Mostre que a variância de X é dada por

$$\mathrm{Var}[X] = \frac{2q^2(r+q-2)}{r(q-2)^2(q-4)}.$$

5. Métodos de estimação de parâmetros

"When you have eliminated the impossible whatever remains, however improbable, must be the truth."
Arthur Conan Doyle

No Capítulo 3, apresentamos diversas distribuições discretas e contínuas comumente encontradas em aplicações em economia e finanças. A conceituação de variáveis aleatórias apresentadas, bem como as suas características (valores esperados, funções de densidade de probabilidade, funções de frequência, funções distribuições acumuladas), pode ser vista como a formatação de uma caixa de ferramentas, no caso de modelos matemáticos estocásticos, para serem usados em problemas práticos de análise de dados reais. Neste capítulo, damos um passo adicional, lidando com métodos que podem ser usados para a estimação dos parâmetros livres desses modelos.

Para contextualizar de forma mais geral o processo de estimação de parâmetros e ajuste de distribuições, vamos fazer uma discussão agora sobre o processo de modelagem de dados, em problemas estatísticos e econométricos. Na prática, o processo de modelagem aplicada de um processo estocástico funciona seguindo-se os passos a seguir:

(1) Primeiramente, temos que definir com clareza quais são os nossos objetivos de pesquisa. Essa tarefa, aparentemente simples, constitui-se na etapa mais importante do processo de modelagem. Muito comumente, encontram-se trabalhos onde não há clareza nos objetivos, ou aparentemente os pesquisadores se perderam em relação aos objetivos iniciais e os resultados obtidos ao final do trabalho. Além disso, a depender dos objetivos, não necessariamente as ferramentas estatísticas a serem utilizadas serão ferramentas avançadas. Pode acontecer que o problema está tão bem montado e a base de dados possui características tais que uma simples comparação entre médias pode fornecer a resposta ao nosso questionamento de pesquisa.

(2) A partir do objetivo de pesquisa, é preciso entender quais são os requisitos em termos de bases de dados necessárias. Essa etapa requer uma garimpagem das bases disponíveis, com uma descrição detalhada do processo de seleção das unidades observacionais na base, e com uma descrição das variáveis coletadas. Em muitos casos, os dados a serem estudados estão disponíveis em sistemas públicos de informações, fornecidos por instituições como o IBGE, Ipeadata, Banco Central etc. Outra fonte importante de informações são as bases internas das próprias instituições, privadas ou públicas, com interesse no estudo. Um exemplo são as bases de dados de clientes. Finalmente, quando as bases de dados disponíveis não contiverem todas as

informações necessárias, e quando houver orçamento para tal, pode-se recorrer à coleta de bases adicionais, para complementar as informações.

Em geral, os passos (1) e (2) acima não necessariamente acontecem na ordem em que foram apresentados. Em muitos casos, excelentes pesquisas são planejadas e implementadas a partir do conhecimento do pesquisador de que certo conjunto de informações está disponível. Dessa forma, quando o IBGE divulga uma determinada pesquisa, com uma variável que antes não existia em outras bases, pesquisadores automaticamente vislumbram uma série de perguntas que podem ser respondidas com base nessas novas informações. Diversos pesquisadores realizam trabalhos especificamente para montagem de séries históricas confiáveis, de variáveis macroeconômicas relevantes. Um exemplo de trabalhos nesse sentido são as séries divulgadas no Ipeadata (www.ipeadata.gov.br). Em muitos casos, notas metodológicas são divulgadas especificamente para divulgar a metodologia empregada nessas montagens.

(3) Uma vez tendo em mãos as bases de informações a serem estudadas, para responder aos objetivos do estudo, o próximo passo é fazer uma análise exploratória dos dados disponíveis. Nesse caso, análises gráficas, tabelas de frequências, histogramas, gráficos de dispersão, gráficos de séries históricas, medidas do tipo média, mediana, coeficiente de assimetria, desvio padrão, percentis etc. são úteis. Na verdade, o leitor mais interessado na exploração de bases de dados pode recorrer a uma vasta literatura sobre análise exploratória de dados[1] (HOAGLIN; MOSTELLER; TUKEY, 1983, 1985; TUKEY, 1977; VELLEMAN; HOAGLIN, 1981; LEINHARDT; LEINHARDT, 1980).

(4) A análise exploratória de dados pode nos ajudar na escolha de que tipo de metodologia utilizar para responder à pergunta em questão. Além disso, é extremamente importante fazer uma vasta pesquisa bibliográfica sobre trabalhos já feitos, na mesma área de estudo ou em áreas correlatas, onde os autores se depararam com o mesmo tipo de questionamento de pesquisa. Na maioria das vezes, a solução para que metodologia empregar no nosso problema de pesquisa estará documentada em referências amplamente disponíveis. Obviamente, fica a cargo do pesquisador aplicar, quando possível, metodologias alternativas e/ou mais modernas. Em um conjunto muito pequeno de situações aplicadas em economia e finanças, o pesquisador irá se deparar com situações de estudo para as quais não haja metodologias disponíveis ou trabalhos similares documentados. Em todo caso, ao final do processo de exploração dos dados e do processo de coleta de referências bibliográficas, o pesquisador terá uma boa ideia de que modelo estocástico ele utilizará para resolver o seu problema de pesquisa.

(5) Com base nos dados disponíveis, no questionamento e no objetivo do estudo, nas referências bibliográficas encontradas e na análise exploratória sobre as características particulares dos dados, o próximo passo é definir qual o modelo estocástico será utilizado. Essa etapa é menos simples do que parece. A depender das habilidades do pesquisador, dos softwares disponíveis, e das características da base de dados, levantadas na análise exploratória, o pesquisador pode ficar tentado a utilizar ferramentas muito sofisticadas. Além disso, criou-se uma verdadeira indústria de se reproduzir trabalhos acadêmicos sofisticados, utilizando-se técnicas recentes e/ou aplicando-se a bases de dados nacionais, com o intuito

[1] Do inglês, *exploratory data analysis* (EDA).

apenas de ter mais uma publicação acadêmica. No entanto, é importante ter em mente que não necessariamente a utilização de técnicas mais sofisticadas é o mais apropriado. Quando tivermos alguma ferramenta na cabeça para ser utilizada, a pergunta que devemos fazer é a seguinte: "o que a utilização dessa ferramenta mais sofisticada, vis-a-vis uma técnica mais simples e intuitiva, me compra?". Muitas vezes, criatividade e a escolha correta de uma boa base de dados são muito mais importantes do que uma metodologia super sofisticada e moderna.

Independentemente de como é feita a escolha da ferramenta de análise sofisticada ou não a ser utilizada, esse constitui-se em um passo importante no processo de estudo de dados. Um exemplo interessante na área de finanças é a aplicação de técnicas estatísticas para modelagem de risco operacional, que corresponde ao risco de perdas financeiras em instituições financeiras devido a falhas humanas, falhas de sistema, roubos, fraudes internas e externas etc. Uma das características principais nesse processo é a presença de dois tipos de processos estocásticos: um processo discreto, contabilizando pela frequência de perdas operacionais em cada período (exemplo, dias, semanas, meses) de tempo, e um processo contínuo positivo contabilizando pela severidade (valor monetário) das perdas. Um procedimento comum para modelar tal situação é o procedimento baseado em modelos de perda,[2] onde a frequência e a severidade dos eventos de perdas são modelados separadamente, e em seguida a convolução entre esses dois processos é efetuada, utilizando-se, por exemplo, simulações de Monte Carlo. Uma vez escolhido o modelo matemático estocástico para modelar os dados disponíveis, parte-se finalmente para o foco deste capítulo, que é o processo de ajuste dos modelos estocásticos escolhidos aos dados disponíveis e explorados.

(6) Imagine que o modelo matemático escolhido para modelar, por exemplo, o tempo de permanência na situação de desemprego, seja uma variável aleatória exponencial negativa. Para o nosso estudo, temos disponível uma base de dados contendo uma amostra de pessoas que estiveram desempregadas, e que já encontraram outra ocupação. Queremos determinar a distribuição, dentro da família de distribuições exponenciais, que melhor se aproxima dos dados coletados. Nesse caso, a variável de interesse é o tempo de permanência na situação de desemprego.[3] O foco deste capítulo e justamente encontrar estimativas para os valores dos parâmetros de modelos matemáticos paramétricos, com base em uma amostra de observações disponíveis.

(7) Uma vez estimados os parâmetros do modelo paramétrico escolhido (ou da cesta de modelos paramétricos escolhidos), resta saber se o modelo escolhido está realmente de acordo com os dados utilizados para as estimações. Nesse sentido, diversos testes de ajustes e critérios gráficos podem ser empregados para investigar o quão ajustados os modelos estão aos dados. Via de regra, todo modelo está errado, e desejamos encontrar um modelo que seja interpretável ao ponto de podermos utilizá-lo para tirar as nossas conclusões no estudo. De fato, para um estatístico com várias ferramentas de cheque de ajuste, é muito provável que ele sempre consiga rejeitar a maioria dos modelos estimados. Mas isso não significa que os modelos, mesmo não

[2]Em inglês, *loss models*.

[3]Na prática, esse tipo de problema é tratado com uma metodologia ligeiramente diferente, dado que pode haver indivíduos na amostra que ainda estão em situação de desemprego, por exemplo. Nesse caso, dizemos que temos dados censurados (à direita), pois não observamos ao certo o valor de permanência para situações de desemprego para esses indivíduos. O tratamento estatístico para essas situações se dá pela utilização de modelos conhecidos como modelos de sobrevivência.

passando em todos os critérios de ajuste, devam ser desprezados. Cabe ao pesquisador ter o discernimento de saber o quanto ele vai aceitar em termos de falta de ajuste. O mais importante é ser honesto consigo mesmo, de forma que as conclusões tiradas sejam eticamente aceitáveis. Alguns dos critérios formais e gráficos para checar o ajuste de modelos serão vistos nos Capítulos 7, 8 e 9.

(8) Se após os testes de ajustes o modelo paramétrico (ou a família de modelos paramétricos) estimado não for aceitável, pode-se escolher outros modelos que eventualmente possam apresentar um ajuste melhor. Uma outra solução é encontrar um conjunto adicional de dados para ajudar no estudo. Portanto, o processo de modelagem de dados nos passos de (1) a (7) pode acontecer iterativamente até termos um modelo propriamente ajustado aos dados de interesse, para tirarmos as nossas conclusões de pesquisa.

No exemplo de modelagem de perdas operacionais, para calcular a distribuição de perdas agregadas totais da instituição, é necessário efetuar simulações de Monte Carlo. Para isso, é necessário encontrar, para cada tipo de perda, qual a distribuição que melhor se aplica ao modelo de frequência e ao modelo de severidade das perdas. Para atingir tal objetivo, é preciso descobrir qual o valor mais adequado para os parâmetros livres em cada variável aleatória estudada nas Seções 3.2 e 3.3. Por exemplo, conforme vimos na Figura 3.14, a depender do valor do parâmetro livre λ para a distribuição exponencial negativa, a função densidade ficará mais deslocada para a direita ou para a esquerda. Quando a função está deslocada para a direita, significa que estamos observando valores de perda em média maiores do que se a densidade estivesse mais deslocada para esquerda. Analogamente, para a variável aleatória lognormal, gostaríamos de encontrar os valores dos parâmetros livres μ e σ^2 que mais se aproximam do conjunto de dados analisado.

Neste capítulo apresentaremos os principais métodos de estimação de parâmetros livres de modelos estocásticos. Apesar de os métodos apresentados aqui serem aplicados especificamente a algumas das distribuições cobertas neste documento, esses métodos podem facilmente ser aplicados aos mais variados tipos de modelos estatísticos. Conforme ficará claro nas descrições a seguir, esses métodos apresentam uma sistemática suficientemente geral. Para maiores detalhes em diversos métodos de estimação de parâmetros, vide Bickel e Doksum (2000), Tanner (1996), White (1996) White (2000), Hansen (1982a), Hansen e Singleton (1982b) e Gourieroux e Jasiak (2001).

Como pode ser notado, o foco desse livro segue nitidamente a chamada abordagem de **inferência frequentista**, de acordo com a qual os parâmetros das distribuições são variáveis fixas, com valores desconhecidos, para os quais o objetivo do analista é justamente obter uma estimativa para esses valores. Uma abordagem alternativa aos métodos frequentistas é conhecida como **inferência Bayesiana**. Na abordagem Bayesiana, os parâmetros das distribuições são tratados como variáveis aleatórias. Então, partindo-se de uma distribuição inicial para os parâmetros do modelo e observando-se uma amostra de dados, a inferência Bayesiana permite que a distribuição dos parâmetros seja atualizada, combinando-se o conhecimento inicial e a informação contida na amostra.

Considerando o foco frequentista desse material, inicialmente, será descrito o **método de momentos**. Esse procedimento apresenta a grande vantagem de ser extremamente simples e intuitivo. Entretanto, o

método de momentos, em muitos casos, não é o método ótimo em termos de utilização de toda a informação disponível nos dados, para encontrar os valores dos parâmetros livres das diversas distribuições. O segundo método abordado será o **método de máxima verossimilhança** (MV). Diferentemente do método de momentos, o método de máxima verossimilhança nem sempre é tão intuitivo e simples, mas em geral é o melhor que se pode fazer para aproveitar a informação existente no conjunto de dados. Dizemos, nesse caso, que o método de máxima verossimilhança é **estatisticamente eficiente**. Para maiores detalhes, vide Casella e Berger (2001) e Bickel e Doksum (2000), por exemplo.

Depois de apresentar essas técnicas de inferência frequentista, concluímos o capítulo com uma pequena introdução à técnica de inferência Bayesiana. É válido mencionar, entretanto, que a área de inferência Bayesiana é muito extensa e não é o objetivo deste livro cobrí-la de forma exaustiva. O leitor interessado pode recorrer a Gelman et al. (1995), Neopolitan (2004), Koch (2007), Greenberg (2013) e Tanner (1996) para maiores detalhes.

5.1 Estimação via método de momentos

O **método dos momentos** baseia-se na ideia de encontrar parâmetros de modelos igualando os momentos populacionais (geralmente a média e a variância) aos momentos amostrais. Para melhor apresentar o método de momentos, considere inicialmente o problema de estimar o parâmetro λ da distribuição de Poisson, a um conjunto de dados de frequência diária de perdas, como mostra o Exemplo 5.1.

Exemplo 5.1 (Variável aleatória de Poisson) Seja $X_1, X_2, \ldots, X_n$, um conjunto correspondente ao número de perdas operacionais em n dias, na nossa amostra para modelagem da frequência das perdas operacionais. Nesse caso, podemos ter por exemplo $X_1 = 3$, $X_2 = 4$, $X_3 = 10$, $\ldots$, $X_n = 8$. Queremos encontrar o valor λ da distribuição de Poisson que melhor se adequa a esse conjunto de dados. Para isso, lembremos que as médias e as variâncias populacionais no caso da variável aleatória de Poisson são dadas por

$$\mathrm{E}[X] = \mu = \sum_{x=0}^{x=\infty} x f(x) = \sum_{x=0}^{x=\infty} x \frac{e^{-\lambda}\lambda^x}{x!} = \lambda, \tag{5.1}$$

onde $f(x)$ é a função de frequência. Conforme discutimos anteriormente no Capítulo 3, intuitivamente a média populacional indica simplesmente o valor médio que a variável aleatória de Poisson vai apresentar quando o experimento é repetido um número infinito de vezes. Esse valor médio na amostra de dados $X_1, X_2, \ldots, X_n$ é simplesmente a média amostral $\hat{\mu}$ ou $\bar{X}$

$$\hat{\mu} = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{X_1 + X_2 + X_3 + \cdots + X_n}{n}. \tag{5.2}$$

Obviamente, dado um conjunto de dados $X_1, X_2, \ldots, X_n$, o cálculo da média $\bar{X}$ é bastante simples. O princípio básico do método de momentos consiste em igualar o momento populacional ao momento amostral. Nesse caso, podemos simplesmente fazer

$$\begin{aligned} &\text{Média populacional} = \mathrm{E}[X] = \lambda \text{ (no caso da distribuição de Poisson)} \\ &\text{Média amostral} = \bar{X} \text{ (calculado diretamente da amostra)} \\ &\text{Igualando as duas médias: } \lambda = \bar{X}. \end{aligned} \tag{5.3}$$

Portanto, de acordo com o método de momentos, o valor de λ que mais se adequa à amostra $X_1, X_2, \ldots, X_N$ é simplesmente a média amostral $\bar{X}$. Esse valor que mais se adequa denominamos **estimador** do parâmetro λ via método de momentos. A partir de agora, todo estimador de um determinado parâmetro será representado pela letra do parâmetro com um acento circunflexo. A média amostral $\hat{\mu}$ é simplesmente o estimador da média populacional μ. Portanto, o estimador do parâmetro livre λ será representado por $\hat{\lambda}$, e no caso da variável aleatória exponencial, temos $\hat{\lambda} = \bar{X}$, para o estimador de momentos utilizando-se o momento de primeira ordem.

É importante frisar que o estimador de momentos baseado na média amostral para o parâmetro λ não é único. De fato, lembremos agora que a variância populacional para uma variável aleatória de Poisson é dada por $\mathrm{Var}[X] = \lambda$. Ou seja, a média e a variância de uma variável aleatória de Poisson são iguais. A variância amostral para a amostra disponível é dada por

$$\text{Variância amostral } = \frac{1}{n}\sum_{i=1}^{n}\left[X_i - \hat{\mu}\right]^2.$$

Igualando-se a variância populacional à variância amostral, obtêm-se um outro estimador para λ, com base no método de momentos. Esse segundo estimador tem expressão

$$\hat{\lambda} = \frac{1}{n}\sum_{i=1}^{n}\left[X_i - \hat{\mu}\right]^2.$$

Analisando o Exemplo 5.1, duas perguntas aparecem a partir da observação de que podem ser definidos mais de um estimador, somente com base no método de momentos:

(i) Qual dos dois estimadores utilizar na prática?

(ii) Se utilizarmos os dois estimadores de momentos (o baseado na média e o outro baseado na variância), e as estimativas obtidas para λ forem diferentes, o que isso quer dizer?

Para responder à primeira pergunta, podemos pensar em dois critérios. O primeiro critério está ligado ao viés do estimador. Em termos intuitivos, o viés do estimador tem a seguinte interpretação. Imagine

que outros pesquisadores coletassem amostras diferentes, todas com tamanho n, para o mesmo processo de frequência de perdas operacionais. Além disso, imagine que cada um desses pesquisadores aplicasse o estimador de método de momentos da média populacional, para encontrar o valor de λ. Como as amostras utilizadas por cada pesquisador são diferentes, os valores encontrados para $\hat{\lambda}$ também são diferentes. Além disso, é muito improvável que algum desses pesquisadores acerte o valor correto do parâmetro λ (conhecido apenas por Deus). Portanto, cada pesquisador está encontrando um valor diferente para λ. O viés do estimador corresponde a se, na média, os diversos pesquisadores estão atingindo o valor correto. Se na média as estimativas estiverem corretas dizemos que o estimador é **não viesado**. Caso as estimativas não acertem em média o valor correto do parâmetro, então dizemos que o estimador é **viesado**. Portanto, o primeiro critério que gostaríamos de ter em relação a um estimador é que ele seja não viesado, ou que pelo menos o viés seja pequeno, e decaia quando o tamanho da amostra aumenta.

Para entender o segundo critério para escolha do estimador, consideremos novamente o exemplo de vários pesquisadores, cada qual com diferentes amostras do mesmo tamanho n, utilizando o estimador de método de momentos com base na média populacional. Podemos calcular a dispersão (por alguma medida vista no Capítulo 2, por exemplo) dos valores obtidos para $\hat{\lambda}$ em torno do valor correto λ. Idealmente, gostaríamos de ter um estimador que tenha menor dispersão ao redor da média. Portanto, o segundo critério para escolha de um estimador está relacionado à dispersão do estimador. Via de regra, gostaríamos de ter um estimador que seja não viesado e que tenha a menor variância possível. Normalmente dizemos que o estimador que possui a menor variância possível é dito ser **eficiente**.

A segunda pergunta, em relação ao que fazer quando os resultados fornecidos por diferentes estimadores de momento, para a mesma amostra, forem diferentes, pode ser vista de duas óticas. O primeiro fator para explicar a diferença entre as duas estimativas é a imprecisão amostral. Pode acontecer que os estimadores estejam fornecendo estimativas diferentes simplesmente devido ao erro amostral. O outro motivo para essa possível discrepância é o erro de especificação; ou seja, a distribuição dos dados pode não ser implicitamente uma distribuição de Poisson. Nesse caso, o pesquisador deve tentar outras distribuições que melhor se adequem aos dados coletados.

Exemplo 5.2 (Variável aleatória exponencial negativa) Imagine agora que temos uma sequência de dados $Y_1, Y_2, Y_3, \ldots, Y_m$ correspondente aos valores monetários de m eventos de perdas operacionais, que podem ou não ter acontecido no mesmo dia. Podemos ter, por exemplo, $Y_1 = \text{R\$ } 3500.00$, $Y_2 = \text{R\$ } 10600.00$, $Y_3 = \text{R\$ } 4100.00$, $\ldots$, $Y_m = \text{R\$ } 1630.00$. O nosso objetivo é encontrar os valor do parâmetro λ que mais se adequa a esse conjunto de dados, para um variável aleatória exponencial negativa.

Para a distribuição exponencial negativa, recordemos que a média populacional $\mathrm{E}[X]$ é dada pela integral

$$\mathrm{E}[X] = \int_{y=0}^{\infty} x f(x) dx = \int_{y=0}^{\infty} x\lambda e^{-\lambda x} dx = \frac{1}{\lambda}. \tag{5.4}$$

Podemos então empregar o mesmo procedimento visto acima no Exemplo 5.1, e fazer

$$\begin{aligned}
&\text{Média populacional} = \mathrm{E}[X] = \frac{1}{\lambda} \text{ (no caso da distribuição da exponencial negativa)} \\
&\text{Média amostral} = \bar{X} \text{ (calculado diretamente da amostra)} \\
&\text{Igualando as duas médias: } \frac{1}{\lambda} = \bar{X} \Rightarrow \hat{\lambda} = \frac{1}{\bar{X}},
\end{aligned} \tag{5.5}$$

e concluímos que o estimador de método de momentos do parâmetro λ na distribuição exponencial negativa é simplesmente $\hat{\lambda} = 1/\bar{X}$.

Prática 5.1 Encontre um estimador de método de momentos para o parâmetro λ de uma variável aleatória exponencial negativa com base na variância populacional.

Exemplo 5.3 (Variável aleatória lognormal) No caso da variável aleatória lognormal, temos dois parâmetros livres para serem estimados. Conforme veremos na discussão de escolha de modelos estatísticos, o fato de termos dois parâmetros livres nos possibilita uma maior adaptabilidade aos dados, pois temos mais "eixos livres"para ajustar. Por causa disso, para estimar ambos os parâmetros será preciso utilizar duas igualdades de momentos, ao invés de apenas uma.

Lembremos que, na discussão sobre a distribuição lognormal, partimos de uma variável aleatória normal W com média μ e variância σ^2. Nesse caso, para a nova variável aleatória definida por $Y = e^W$, onde e é o número neperiano, temos que Y tem uma distribuição lognormal com parâmetros μ e σ^2. Imagine agora que temos uma amostra de severidades de perdas operacionais $Y_1, Y_2, Y_3, \ldots, Y_m$. Supondo que esses dados tenham sido gerados a partir de uma distribuição lognormal com parâmetros μ e σ^2 (desconhecidos), o nosso objetivo é encontrar os valores para esses parâmetros, de forma a obter o ajuste mais adequado aos dados. Para isso, podemos então fazer o caminho inverso: se $Y = e^W$, onde $W \sim \mathrm{N}(\mu, \sigma^2)$, então $W = \log Y$ tem distribuição normal com média μ e variância σ^2. A notação $W \sim \mathrm{N}(\mu, \sigma^2)$ pode ser lida como: "a variável W tem distribuição normal com média μ e variância σ^2".

Considere a amostra transformada $W_1 = \log Y_1, W_2 = \log Y_2, W_3 = \log Y_3, \ldots, W_m = \log Y_m$. De acordo com a discussão anterior, a sequência $W_1, W_2, \ldots, W_m$ corresponde então a uma amostra de observações normais, com média μ e variância σ^2. Aplicando a ideia do método de momentos à amostra $W_1, W_2, \ldots, W_m$, temos para a média μ

$$\begin{aligned}
&\text{Média populacional} = \mathrm{E}[X] = \mu \text{ (no caso da distribuição normal)} \\
&\text{Média amostral} = \bar{W} \text{ (calculado diretamente da amostra)} \\
&\text{Igualando as duas médias: } \hat{\mu} = \bar{W} \Rightarrow \hat{\mu} = \bar{W} = \frac{1}{m}\sum_{i=1}^{m} \log Y_i,
\end{aligned} \tag{5.6}$$

e para a variância σ^2

$$\begin{aligned}
&\text{Variância populacional} = \text{Var}[X] = \sigma^2 \text{ (no caso da distribuição normal)} \\
&\text{Variância amostral} = s^2 = \frac{1}{m}\sum_{i=1}^{m}(W_i - \bar{W})^2 \text{ (calculado diretamente da amostra)} \\
&\text{Igualando as duas variâncias: } \hat{\sigma^2} = s^2 = \frac{1}{m}\sum_{i=1}^{m}(W_i - \bar{W})^2.
\end{aligned} \tag{5.7}$$

Lembrando que $W_i = \log Y_i$, para $i = 1, 2, \ldots, m$, temos finalmente, os estimadores de método de momentos dos parâmetros μ e σ^2 referentes à variável aleatória lognormal

$$\begin{aligned}
\hat{\mu} &= \frac{1}{m}\sum_{i=1}^{m}\log Y_i = \overline{\log Y_i}, \\
\hat{\sigma^2} &= \frac{1}{m}\sum_{i=1}^{m}(\log Y_i - \overline{\log Y_i})^2.
\end{aligned} \tag{5.8}$$

Exemplo 5.4 (Variável aleatória gamma) No caso da variável aleatória gamma, com parâmetros α e β e função densidade de probabilidade

$$f(y) = \frac{1}{\beta^\alpha \Gamma(\alpha)} y^{\alpha-1} e^{-y/\beta}, \tag{5.9}$$

a média e a variância populacionais são dadas por $\mu = \alpha\beta$ e $\sigma^2 = \alpha\beta^2$. Para uma amostra de valores de severidades de perdas $Y_1, Y_2, \ldots, Y_m$, para utilizarmos o método de momentos, devemos igualar os momentos populacionais μ e σ^2 aos momentos amostrais $\bar{Y}$ e s^2. Note que, devido ao fato de termos dois parâmetros livres α e β, devemos utilizar dois momentos amostrais. Dessa forma,

$$\begin{aligned}
\text{E}[Y] &= \alpha\beta = \bar{Y} = \frac{1}{m}\sum_{i=1}^{m} Y_i, \\
\text{Var}[Y] &= \alpha\beta^2 = s^2 = \frac{1}{m}\sum_{i=1}^{m}(Y_i - \bar{Y})^2,
\end{aligned} \tag{5.10}$$

e resolvendo a Eq. (5.10), obtemos os estimadores de método de momentos para α e β

$$\hat{\alpha} = \frac{\left[\frac{1}{m}\sum_{i=1}^{m} Y_i\right]^2}{\frac{1}{m}\sum_{i=1}^{m}(Y_i - \bar{Y})^2}, \qquad \hat{\beta} = \frac{\frac{1}{m}\sum_{i=1}^{m}(Y_i - \bar{Y})^2}{\frac{1}{m}\sum_{i=1}^{m} Y_i}. \tag{5.11}$$

Exemplo 5.5 (Variável aleatória beta) Para a variável aleatória beta, com parâmetros α e β, e função densidade

$$f(y) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} y^{\alpha-1}(1-y)^{\beta-1}, \text{ para } y \in (0,1), \tag{5.12}$$

a média e a variância são dadas por

$$\mu = \frac{\alpha}{\alpha+\beta}, \qquad \sigma^2 = \frac{\alpha\beta}{(\alpha+\beta+1)(\alpha+\beta)^2}. \tag{5.13}$$

Para utilizar o método de momentos, dada uma amostra $Y_1, Y_2, \ldots, Y_m$, precisamos de duas igualdades entre momentos populacionais e amostrais, já que temos dois parâmetros livres α e β. Procedendo da mesma maneira que no caso da distribuição gamma, podemos fazer

$$\begin{aligned} \mathrm{E}[Y] &= \frac{\alpha}{\alpha+\beta} = \bar{Y} = \frac{1}{m}\sum_{i=1}^{m} Y_i, \\ \mathrm{Var}[Y] &= \frac{\alpha\beta}{(\alpha+\beta+1)(\alpha+\beta)^2} = s^2 = \frac{1}{m}\sum_{i=1}^{m}(Y_i - \bar{Y})^2. \end{aligned} \tag{5.14}$$

Resolvendo as duas equações acima em α e β, encontramos os estimadores de método de momentos $\hat{\alpha}$ e $\hat{\beta}$, como função da média e da variância amostrais $\bar{Y}$ e s^2.

$$\begin{aligned} \hat{\alpha} &= \frac{\bar{Y}^2 - \bar{Y}^3 - \bar{Y}s^2}{s^2}, \\ \hat{\beta} &= \hat{\alpha}\frac{1-\bar{Y}}{\bar{Y}}. \end{aligned} \tag{5.15}$$

No Exemplo 5.5 fica claro que nem sempre os estimadores de método de momentos apresentam expressões triviais. Mesmo assim, o método de momentos se constitui em um método muito utilizado em aplicações, sendo bem mais simples e intuitivo do que o método de máxima verossimilhança que será apresentado a seguir na Seção 5.2. O método de momentos é o percursor do muito utilizado atualmente **método de momentos generalizado**, ou GMM (*generalized method of moments*) introduzido por Hansen (1982a). Uma das vantagens do método de momentos e do método de momentos generalizados, em relação ao método de máxima verossimilhança, é que o método de momentos e o GMM não precisam da especificação da função densidade completa, como ocorre no fato da estimação via máxima verossimilhança. O método de momentos e o GMM precisam apenas do que chamamos de **condições de momento** que aparecem nos exemplos dessa seção. Quando informações sobre a função densidade estão disponíveis, o método de máxima verossimilhança utiliza melhor a informação contida na amostra, no sentido de que o estimador de máxima verossimilhança tem melhor relação de compromisso entre viés e variância da distribuição dos estimadores.

5.2 Estimação via máxima verossimilhança

Nesta seção descrevemos um dos métodos mais populares para estimação de parâmetros em modelos estatísticos. Apesar de o método de máxima verossimilhança não apresentar a mesma simplicidade que o método de momentos, ao utilizar o método de máxima verossimilhança, estamos em geral utilizando o máximo de informação disponível nos dados para estimar os parâmetros das distribuições de interesse. Isso permite ao método de máxima verossimilhança obter estimativas mais precisas, com menor variância (intuitivamente com uma menor margem de erro), do que o método de momentos. No jargão estatístico, dizemos que os estimadores de máxima verossimilhança são **eficientes**.

Para ilustrar o método de estimação via máxima verossimilhança, considere uma amostra de frequência de perdas diárias $X_1, X_2, \ldots, X_n$, para um determinado tipo de evento de perda. O nosso objetivo é estimar o parâmetro (ou parâmetros) de forma que a distribuição considerada melhor se ajuste ao conjunto de dados. Nos exemplos que se seguem, vamos então aplicar esse método de estimação para algumas das distribuições de frequência descritas nas Seções 3.2 e 3.3.

Exemplo 5.6 (Variável aleatória de Poisson) Para a variável aleatória de Poisson, lembremos que a função de frequência é dada por

$$f(x) = \frac{e^{-\lambda}\lambda^x}{x!}, \text{ para } x = 0, 1, 2, \ldots. \tag{5.16}$$

A expressão acima fornece a função densidade para cada uma das n observações de frequências de perdas em $X_1, X_2, \ldots, X_n$. Supondo que as observações nessa amostra são independentes umas das outras, podemos escrever a função de frequência conjunta entre as n observações como o produto das n funções densidades para cada observação individualmente. Dessa forma, a função de frequência conjunta vai ser

$$f(x_1, x_2, x_2, \ldots, x_n) = \frac{e^{-\lambda}\lambda^{x_1}}{x_1!} \times \frac{e^{-\lambda}\lambda^{x_2}}{x_2!} \times \frac{e^{-\lambda}\lambda^{x_3}}{x_3!} \times \cdots \times \frac{e^{-\lambda}\lambda^{x_n}}{x_n!}. \tag{5.17}$$

Da mesma maneira que a soma de uma sequência $a_1 + a_2 + a_3 + \cdots + a_n$ pode ser representada pelo símbolo de somatório $\sum_{i=1}^{n} a_i$, o produto da sequência $a_1 \times a_2 \times a_3 \times \cdots \times a_n$ pode ser representado pelo símbolo de produtório $\prod_{i=1}^{n} a_i$. Portanto, podemos reescrever a Eq. (5.17) como

$$f(x_1, x_2, x_2, \ldots, x_n) = \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{x_i}}{x_i!}, \tag{5.18}$$

onde $x_1, x_2, \ldots, x_n$ são os valores numéricos correspondentes às variáveis aleatórias $X_1, X_2, \ldots, X_n$ na amostra. O princípio básico do método de máxima verossimilhança consiste em encontrar o valor $\hat{\lambda}$ para o parâmetro λ que maximize a função $f(x_1, x_2, x_2, \ldots, x_n)$ apresentada na Eq. (5.18). Nesse caso, podemos escrever a função $f(x_1, x_2, x_2, \ldots, x_n)$ explicitamente como uma função de λ. Essa função escrita explicitamente como função do parâmetro (ou dos parâmetros) a ser estimado é conhecida como **função de**

verossimilhança e é representada pelo símbolo $L(\lambda)$. Em inglês, a função de verossimilhança é denominada *likelihood function*, o que explica a notação $L(\lambda)$.

Em forma mais compacta e matematicamente mais elegante, podemos dizer que o estimador de máxima verossimilhança $\hat{\lambda}$ do parâmetro λ, supondo que a amostra de frequências independentes de perdas $X_1, X_2, \ldots, X_n$ advém de uma variável aleatória de Poisson, é o valor $\hat{\lambda}$ que maximiza a função de verossimilhança

$$L(\lambda) = \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{x_i}}{x_i!}, \tag{5.19}$$

ou seja,

$$\hat{\lambda} = \arg\max_{\lambda} L(\lambda) = \arg\max_{\lambda} \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{x_i}}{x_i!}, \tag{5.20}$$

Equivalentemente, devido ao fato de a função logaritmica ser estritamente crescente, o valor $\hat{\lambda}$ que maximiza $L(\lambda)$ é o mesmo valor que maximiza o logaritmo da função de verossimilhança $\log L(\lambda)$. Além disso, o logaritmo da função de verossimilhança é bem mais tratável tanto analiticamente quanto computacionalmente. Por esse motivo, essa função é denominada na literatura de **função de log-verossimilhança**. Especificamente para variável aleatória de Poisson, a função de log-verossimilhança pode ser escrita como

$$\log L(\lambda) = \log\left[\prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{x_i}}{x_i!}\right] = -n\lambda + \log\lambda \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} \log x_i! \tag{5.21}$$

Na passagem acima, utilizamos o fato de que o logaritmo do produto é a soma dos logaritmos. O estimador de máxima verossimilhança então será o valor de $\hat{\lambda}$ que maximiza a função $\log L(\lambda)$ acima.

Especificamente para variável aleatória de Poisson, maximizar a função $\log L(\lambda)$ acima é uma tarefa relativamente simples, com base nos procedimentos básicos de cálculo. De fato, derivando a função $\log L(\lambda)$ em λ e igualando a zero, obtemos uma forma explicita para o estimador de máxima verossilhança $\hat{\lambda}$:

$$\hat{\lambda} = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X}. \tag{5.22}$$

O gráfico superior na Figura 5.1 apresenta o gráfico da função de log-verossimilhança para uma amostra ajustada a uma distribuição de Poisson discutida no Exemplo 5.6. Observe que a função de log-verossimilhança apresenta apenas um ponto de máximo local. Isso é característico da variável aleatória de Poisson e de algumas outras distribuições pertencentes a uma família de variáveis aleatórias conhecida como distribuições da família exponencial. Para variáveis aleatórias um pouco complexas, a função de

verossimilhança pode apresentar mais de um máximo local, o que pode incorrer em algumas dificuldades (felizmente sanáveis) nos processos de estimação.

O leitor deve ter percebido que o estimador de máxima verossimilhança de λ coincide com o estimador de método de momentos de λ. Esse fato acontece em alguns casos apenas, como no caso da variável aleatória de Poisson. Para a maioria das distribuições, os estimadores de máxima verossimilhança e de método de momentos são diferentes algebricamente. Porém, quando o tamanho da amostra vai para infinito, esses dois estimadores passam a coincidir numericamente.

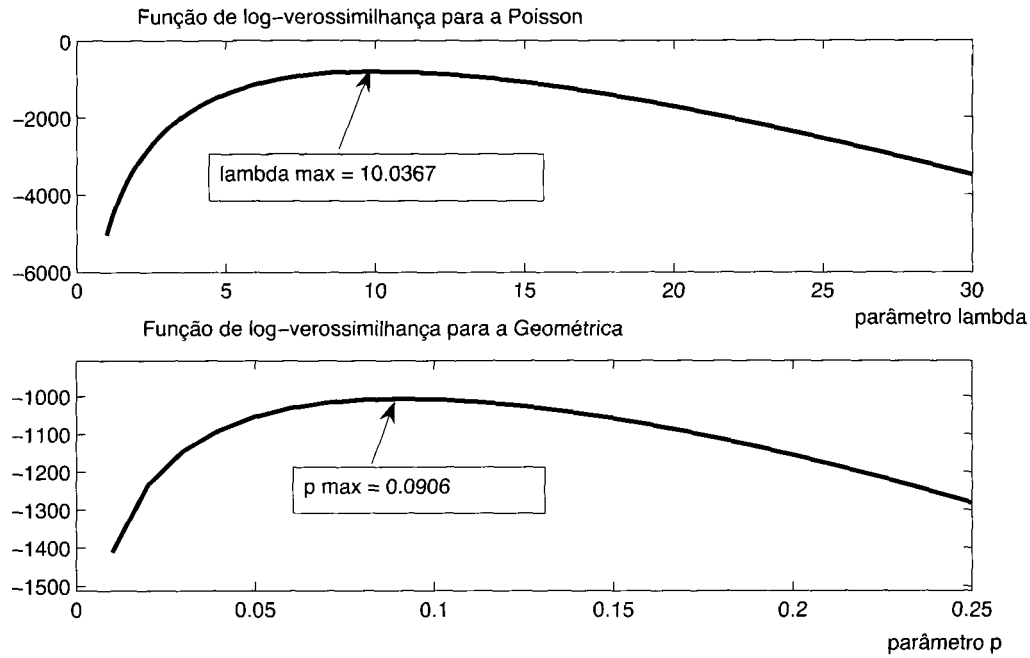


Figura 5.1: Função de log-verossimilhança para a variável aleatória de Poisson e para a variável aleatória geométrica.

Exemplo 5.7 (Variável aleatória geométrica) No caso de uma variável aleatória geométrica, o procedimento usado para obter uma estimativa do parâmetro livre p usando o estimador de máxima verossimilhança é bem parecido com o empregado no caso da variável aleatória de Poisson. Inicialmente, encontramos a função de densidade conjunta para as n observações da amostra $X_1, X_2, \ldots, X_n$, supondo independência entre elas. Por causa da independência, a função de densidade conjunta $f(x_1, x_2, \ldots, x_n)$ será dada pelo produto de cada função densidade individualmente. Dessa forma,

$$f(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} f(x_i) = \prod_{i=1}^{n} p(1-p)^{x_i}. \tag{5.23}$$

A partir da função densidade conjunta, podemos derivar a função de log-verossimilhança, que será maximizada para encontrarmos a estimativa de máxima verossimilhança $\hat{p}$. A partir da expressão anterior

para a função de densidade conjunta, a função de log-verossimilhança tem expressão

$$\log L(p) = n \log p + \log(1-p) \sum_{i=1}^{n} x_i. \tag{5.24}$$

Observe que escrevemos a função de log-verossimilhança $\log L(p)$ como uma função explícita do parâmetro livre p. O gráfico inferior na Figura 5.1 apresenta a função de log-verossimilhança para uma amostra hipotética de frequências diárias de perdas operacionais. Observe que esta função apresenta um único máximo local, que corresponde ao estimador de máxima verossimilhança de p. Similarmente ao caso da estimação do parâmetro λ para a variável aleatória de Poisson, o máximo da função $\log L(p)$ pode ser encontrado analiticamente, bastando derivar a função $\log L(p)$ e igualar a derivada a zero. A solução explícita então para a função de verossimilhança é simplesmente

$$\hat{p} = \frac{1}{1 + \frac{1}{n}\sum_{i=1}^{n} X_i} = \frac{1}{1+\bar{X}}. \tag{5.25}$$

Novamente, como vimos no Exemplo 5.7, a forma da função de log-verossimilhança permite que encontremos uma fórmula explícita para o estimador de máxima verossimilhança do parâmetro livre procurado. No entanto, na maioria dos problemas, não é possível encontrar o valor do parâmetro que maximiza a função de máxima verossimilhança explicitamente, e temos que recorrer a métodos numéricos de maximização. No próximo exemplo, nos deparamos com esse problema, onde a função de log-verossimilhança agora passa a depender de dois parâmetros livres, ao invés de apenas um.

Exemplo 5.8 (Variável aleatória binomial negativa) A função densidade da variável aleatória binomial negativa tem expressão

$$f(x) = \frac{\Gamma(r+x)}{\Gamma(r)\Gamma(x+1)} p^r (1-p)^x, \text{ para } x \in \{0,1,2,3,\ldots\}, \tag{5.26}$$

e, portanto, a função de log-verossimilhança é dada por:

$$\log L(r,p) = -n \log \Gamma(r) + rn \log p + \sum_{i=1}^{n} \log \Gamma(r+x_i) - \sum_{i=1}^{n} \log \Gamma(1+x_i) + \log(1-p) \sum_{i=1}^{n} x_i$$

Observe agora que a função de log-verossimilhança é função dos dois parâmetros livres r e p. Para encontrar o estimador de máxima verossimilhança desses dois parâmetros, temos que encontrar os valores que maximizam a função $\log L(r,p)$. Diferentemente das distribuições de Poisson e geométrica, para a distribuição binomial negativa, não é possível escrever explicitamente expressões para os estimadores de máxima verossimilhança $\hat{r}$ e $\hat{p}$, derivando-se a função $\log L(r,p)$ e igualando as derivadas a zero. Nesse

caso, os valores de r e p que maximizam a função $\log L(r.p)$ devem ser encontrados numericamente, via algoritmos computacionais. Isso é justamente o que a maioria dos programas estatísticos fazem.

Exemplo 5.9 (Variável aleatória gamma) Imagine agora que queremos ajustar uma variável aleatória gamma a um conjunto de dados de severidade (valores monetários) dos eventos de perdas operacionais. Nesse caso, o nosso objetivo é encontrar os valores de α e β na função densidade

$$f(y) = \frac{1}{\Gamma(\alpha)\beta^{\alpha}} y^{\alpha-1} e^{-y/\beta}. \text{ para } y \in (0, \infty), \tag{5.27}$$

que melhor se adequam à massa de dados disponíveis $Y_1, Y_2, Y_3, \ldots, Y_m$. Similarmente aos três exemplos anteriores, a função densidade conjunta, supondo que os eventos são independentes, pode ser escrita como

$$f(y_1, y_2, \ldots, y_m) = \prod_{i=1}^{m} f(y_i) = \prod_{i=1}^{m} \frac{1}{\Gamma(\alpha)\beta^{\alpha}} y_i^{\alpha-1} e^{-y_i/\beta}. \tag{5.28}$$

A partir da função de densidade conjunta, podemos escrever a função de log-verossimilhança $\log L(\alpha, \beta)$, que pode ser obtida diretamente aplicando a transformação logarítmica à função de densidade conjunta acima.

$$\log L(\alpha, \beta) = -m\alpha \log \beta - m \log \Gamma(\alpha) + (\alpha - 1) \sum_{i=1}^{m} \log y_i - \frac{1}{\beta} \sum_{i=1}^{m} y_i. \tag{5.29}$$

Para encontrar os estimadores de máxima verossimilhança de α e β, temos que encontrar os valores de α e β que maximizam a função $\log L(\alpha, \beta)$. Derivando e igualando as duas derivadas a zero não fornecem expressão fechadas para as estimativas $\hat{\alpha}$ e $\hat{\beta}$, diferentemente do que acontece no caso da distribuição de Poisson e da distribuição geométrica. Similarmente à distribuição binomial negativa, dada uma amostra $Y_1, Y_2, \ldots, Y_m$, temos que empregar métodos numéricos para encontrar os pontos de máximo da função de log-verossimilhança. Esses métodos de maximização estão disponíveis na maioria dos programas estatísticos e matemáticos (como o Matlab e o Gauss), e em planilhas eletrônicas, como Excel.

5.3 Distribuição dos estimadores, viés e consistência

Um dos problemas principais na estimação de parâmetros em modelos estatísticos é a preocupação sobre o que de fato estamos obtendo com o estimador. Por exemplo, gostaríamos de saber em média qual o valor obtido para a estimativa do parâmetro, a partir de uma amostra de tamanho n. Idealmente, queremos que o estimador utilizado para estimar um parâmetro μ, por exemplo, resulte em média no valor de μ populacional. Quando isso acontece, dizemos que o estimador é **não viesado**.[4]

[4]Em inglês, esses estimadores são denominados *unbiased estimators*.

De forma mais geral, seja $\theta \in \Re$ um parâmetro de um modelo paramétrico o qual queremos estimar. Para esse parâmetro, construímos um estimador $\hat{\theta}$, via métodos dos momentos ou via máxima verossimilhança, por exemplo. Dizemos que o estimador $\hat{\theta}$ é não viesado para o parâmetro θ, quando $\mathrm{E}[\hat{\theta}] = \theta$; ou seja, o valor esperado do estimador $\hat{\theta}$ é igual ao parâmetro de interesse.

Um ponto importante nessa discussão é o fato de que um estimador é uma função da amostra, que são variáveis aleatórias. Portanto, o estimador também é uma variável aleatória, e consequentemente também tem uma função distribuição acumulada, podendo ter também uma função densidade de probabilidade. Além disso, podemos calcular os momentos para o estimador, da mesma forma como é feito para qualquer outra variável aleatória. Os próximos exemplos ilustram melhor esses conceitos.

Exemplo 5.10 (Variável aleatória de Poisson) Conforme vimos acima, para uma amostra $X_1, X_2, \ldots, X_n$, de observações independentes de uma variável aleatória de Poisson, com parâmetro λ, um estimador de método de momentos para λ é dado pela média

$$\hat{\lambda} = \frac{X_1 + X_2 + \cdots + X_n}{n}.$$

Note que, dado que $X_1, \ldots, X_n$ são variáveis aleatórias, a média aritmética também é uma variável aleatória. O valor esperado do estimador $\hat{\lambda}$ pode ser derivado utilizando as propriedades do valor esperado, conforme visto no capítulo anterior. Portanto,

$$\begin{aligned}\mathrm{E}[\hat{\lambda}] &= \mathrm{E}\left[\frac{X_1 + X_2 + \cdots + X_n}{n}\right] = \mathrm{E}\left[\frac{X_1}{n} + \cdots + \frac{X_n}{n}\right] \\ &= \mathrm{E}\left[\frac{X_1}{n}\right] + \cdots + \mathrm{E}\left[\frac{X_n}{n}\right] = \frac{\mathrm{E}[X_1] + \cdots + \mathrm{E}[X_n]}{n}.\end{aligned}$$

No entanto, é importante lembrar que $E[X_i] = \lambda$, dado que cada variável X_i tem uma distribuição de Poisson com parâmetro λ. Portanto,

$$\mathrm{E}[\hat{\lambda}] = \frac{\lambda + \cdots + \lambda}{n} = \frac{n\lambda}{n} = \lambda,$$

e pela definição acima o estimador $\hat{\lambda}$ é não viesado.

Exemplo 5.11 (Variável aleatória exponencial negativa) No caso da variável aleatória exponencial negativa, com parâmetro λ, de acordo com a parametrização acima, o estimador de método de momentos é dado por

$$\hat{\lambda} = \frac{1}{\bar{X}} = \frac{n}{X_1 + \cdots + X_n}.$$

O estimador acima coincide com o estimador de máxima verossimilhança. Mostraremos que $\mathrm{E}[\hat{\lambda}] \neq \lambda$, e o estimador é viesado. De fato,

$$\frac{1}{\hat{\lambda}} = \frac{X_1 + \cdots + X_n}{n},$$

e, portanto,

$$\mathrm{E}\left[\frac{1}{\hat{\lambda}}\right] = \mathrm{E}\left[\frac{X_1 + \cdots + X_n}{n}\right] = \frac{1}{\lambda}.$$

Na última igualdade acima, utilizamos o fato de que o valor esperado de uma variável aleatória exponencial negativa é igual a $1/\lambda$, e seguimos os mesmos argumentos na derivação do exemplo anterior. A ausência de viés para o estimador $\hat{\lambda}$ apareceria caso tivéssemos

$$\mathrm{E}\left[\frac{1}{\hat{\lambda}}\right] = \frac{1}{E[\hat{\lambda}]} = \frac{1}{\lambda} \Rightarrow E[\hat{\lambda}] = \lambda.$$

Acontece que a igualdade $\mathrm{E}\left[\frac{1}{\hat{\lambda}}\right] = \frac{1}{E[\hat{\lambda}]}$ não é verdadeira, e portanto $E[\hat{\lambda}] \neq \lambda$.

Exemplo 5.12 (Variável aleatória lognormal) Para a variável aleatória lognormal, o estimador de momentos para o parâmetro μ, com base em uma amostra $X_1, \ldots, X_n$, é dado pela expressão

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} \log X_i.$$

Portanto,

$$\mathrm{E}[\hat{\mu}] = \mathrm{E}\left[\frac{1}{n} \sum_{i=1}^{n} \log X_i\right] = \frac{\mathrm{E}[\log X_1] + \cdots + \mathrm{E}[\log X_n]}{n}.$$

Lembrando que, se X tem distribuição lognormal com parâmetros μ e σ, então a variável $W = \log X$ tem distribuição normal com média μ e variância σ^2. Portanto, $\mathrm{E}[\log X_1] = \ldots = \mathrm{E}[\log X_n] = \mu$, levando-nos a concluir que $\mathrm{E}[\hat{\mu}] = \mu$, e o estimador de $\hat{\mu}$ é não viesado.

Conforme vimos acima, não necessariamente os estimadores utilizados na prática são não viesados. Então por que utilizá-los, se podemos estar incorrendo em estimativas "enganosas"? Acontece que, apesar de alguns estimadores serem viesados, eles apresentam uma característica que ainda os tornam atraentes. Esses estimadores são conhecidos como estimadores **consistentes**. O conceito de estimadores consistentes ficará mais claro quando estudarmos as simulações de Monte Carlo para descrever as características das distribuições dos estimadores. Intuitivamente, um estimador é dito consistente quando o viés converge para zero quando o tamanho da amostra tende para infinito.

5.4 Simulações de Monte Carlo

Para entender melhor o conceito de viés do estimador e o conceito de distribuição dos estimadores, vamos apresentar alguns exercícios de simulações de Monte Carlo. Simulações de Monte Carlo são um instrumento bastante utilizado em diversas áreas para a avaliação de cenários onde um componente estocástico (aleatório) está presente. Por exemplo, em avaliações de viabilidade de projetos, onde os fluxos futuros de receitas e dispêndios são incertos, por meio da inserção de aleatoriedade nesses componentes de receitas e despesas, é possível simular quais seriam os valores para indicadores de viabilidade econômico-financeira (*payback time*, valor presente líquido, taxa de desconto etc.) para o projeto, para diferentes cenários aleatórios das variáveis incertas. Para análise estatística, o componente estocástico é a amostra coletada.

Neste texto, simulações de Monte Carlo serão utilizadas para transmitir a intuição por trás das distribuições dos estimadores. Para isso, iremos "brincar" de ser Deus, onde geraremos as observações a serem coletadas nas amostras pelos seres mortais, e investigaremos como os mortais interpretam os nossos sinais divinos, por meio das ferramentas estatísticas. Para isso, antes de mais nada, precisamos definir qual o **processo gerador de dados** que iremos utilizar para gerar as observações. No nosso primeiro exemplo, iremos supor que o processo gerador de dados é uma distribuição exponencial negativa, com parâmetro real $\lambda = 3$. Obviamente esse valor real é conhecido apenas por nós, seres divinos. Aos humanos cabem coletar os valores que iremos gerar, a partir do processo gerador de dados escolhido, e tentar inferir qual o processo gerador de dados divino.

Vamos então gerar uma amostra aleatória de $n = 5$, observações a partir da nossa variável aleatória exponencial negativa, com $\lambda = 3$. Os valores gerados são: $x_1 = 0,6076$, $x_2 = 0,0496$, $x_3 = 0,0978$, $x_4 = 0,0648$, $x_5 = 1,3505$. Para essa amostra, podemos utilizar o estimador de máxima verossimilhança para λ, obtendo $\hat{\lambda} = 1/\bar{x} = 2,3001$. Obviamente, esse valor obtido pelos humanos, com base na amostra de $n = 5$ unidades está bem abaixo do valor real de $\lambda = 3$. Imagine agora que outro grupo de pesquisadores coletou outra amostra de $n = 5$ unidades, geradas a partir do mesmo processo gerador de dados (ou seja, da mesma distribuição exponencial negativa). Para essa nova amostra, os pesquisadores obtiveram uma estimativa $\hat{\lambda} = 3,8921$, que está acima do valor real $\lambda = 3$. Um terceiro grupo de pesquisadores coletou uma outra amostra de $n = 5$ unidades e chegou a um valor estimado igual a $\hat{\lambda} = 3.2123$, que está mais próximo do valor real. Imagine agora que, ao invés de 3 grupos de pesquisadores, 100.000 ou 1.000.000 de grupos de pesquisadores coletassem, cada qual, uma amostra diferente de $n = 5$ unidades, independentemente. Gostaríamos de saber qual o comportamento dos 100.000 ou 1.000.000 valores estimados $\hat{\lambda}$ por cada grupo. Justamente o comportamento desses valores encontrados para a estimativa é que nos fornecerão uma ideia da **distribuição do estimador**. Já que estamos tratando de viés dos estimadores nesta seção, a primeira pergunta é "em média, o estimador de máxima verossimilhança, para o parâmetro da distribuição exponencial negativa, é igual ao valor real do parâmetro?". Analiticamente, vimos anteriormente que isso não é o caso, havendo um viés de estimação no caso do parâmetro da variável aleatória exponencial negativa. Vamos então checar esse fato via simulações de Monte Carlo.

Iremos então simular 100.000 amostras, cada qual com $n = 5$ unidades, a partir de uma variável aleatória exponencial negativa, com parâmetro $\lambda = 3$. Para cada uma dessas amostras, iremos estimar o parâmetro λ via máxima verossimilhança, com $\hat{\lambda} = 1/\bar{x}$. O gráfico superior esquerdo da Figura 5.2 apresenta o histograma para as 100.000 estimativas do parâmetro λ. Esse histograma fornece claramente o formato da função densidade de probabilidade do estimador de máxima verossimilhança $\hat{\lambda}$. Lembremos que o estimador de máxima verossimilhança é uma função da amostra, e portanto também é uma variável aleatória. A média para os 100.000 valores estimados $\hat{\lambda}$ é igual a 3.7403, que é maior do que 3 (valor real do parâmetro λ). Portanto, claramente esse exemplo de simulações ilustra o viés que existe no estimador de máxima verossimilhança para λ. Além disso, o desvio padrão dos 100.000 valores calculados, que nos fornecem uma ideia da imprecisão das estimativas resultou igual a 2.1343.

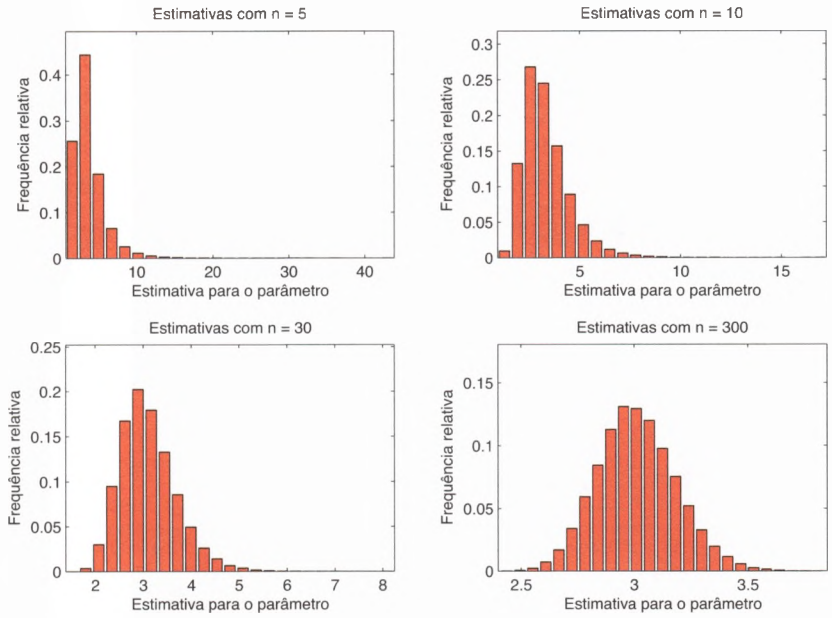


Figura 5.2: Simulações de Monte Carlo para MV do parâmetro de uma variável aleatória exponencial negativa.

Vamos investigar agora o que acontece com a distribuição do estimador de máxima verossimilhança quando as amostras geradas e utilizadas para estimação do parâmetro λ aumentam. Inicialmente, vamos utilizar amostras de tamanho $n = 10$, ao invés de $n = 5$. O histograma dos 100.000 valores estimados $\hat{\lambda}$ com base nessas novas amostras gerados de tamanho $n = 10$ está apresentado no gráfico superior direito da Figura 5.2. A média para essas novas estimativas agora é 3.3321; ainda diferente do valor real, mas já com um viés menor. O desvio padrão das 100.000 estimativas agora é igual a 1.1756; havendo, portanto, uma redução tanto no viés quanto na imprecisão.

Os gráficos esquerdo e direito inferiores da Figura 5.2 apresentam os histogramas para simulações com amostras de tamanhos $n = 30$ e $n = 300$ respectivamente. As novas médias para as 100.000 estimativas são 3.1006, 3.0106, e os desvios padrões são 0.5834, 0.1744. Portanto, o aumento da amostra tem de fato um impacto de redução sobre o viés da estimativa, como também sobre a imprecisão dos estimadores. Além disso, note no formato dos quatro histogramas apresentados Figura 5.2 que, à medida que a amostra

aumenta, as funções densidade de probabilidade da variável aleatória $\hat{\lambda}$ se aproxima mais e mais da função densidade de uma variável aleatória normal. Esse resultado observado nos histogramas é explicado pelo **teorema central do limite**, que será discutido no Capítulo 6. Para $n = 300$, o valor do viés é igual a $3.0106 - 3 = 0.0106$, já bastante próximo de zero. Portanto, as simulações de Monte Carlo também ilustram a consistência do estimador de máxima verossimilhança, que é um estimador viesado e consistente no caso do parâmetro da variável aleatória exponencial negativa.

Para ilustrar a distribuição do estimador de máxima verossimilhança em um caso onde ele é não viesado, a Figura 5.3 apresenta os histogramas para 100.000 estimativas para amostras aleatórias geradas a partir de uma variável aleatória de Poisson, com parâmetro $\lambda = 2$. Ou seja, o processo gerador de dados agora não é mais uma exponencial negativa, mas sim uma variável aleatória de Poisson. O estimador de máxima verossimilhança nesse caso é $\hat{\lambda} = \bar{x}$, que mostramos ser não viesado, de acordo com a subseção anterior. As médias das 100.000 estimativas são 1.9974, 1.9983, 1.9989, 1.9999, para $n = 5, 10, 30$ e 300 respectivamente, enquanto os desvios padrões são 0.6343, 0.4466, 0.2583, 0.0819. Novamente, notamos que (1) o desvio padrão (ou seja, a imprecisão) das estimativas reduz-se à media em que a amostra aumenta; (2) à medida que o tamanho da amostra aumenta, a função densidade de probabilidade da variável aleatória $\hat{\lambda}$ aproxima-se mais da função densidade de probabilidade de uma variável aleatória normal. Além disso, os viéses calculados foram iguais a 0, para todos os tamanhos de amostra, ilustrando que o estimador de máxima verossimilhança no caso da variável aleatória de Poisson é não viesado.

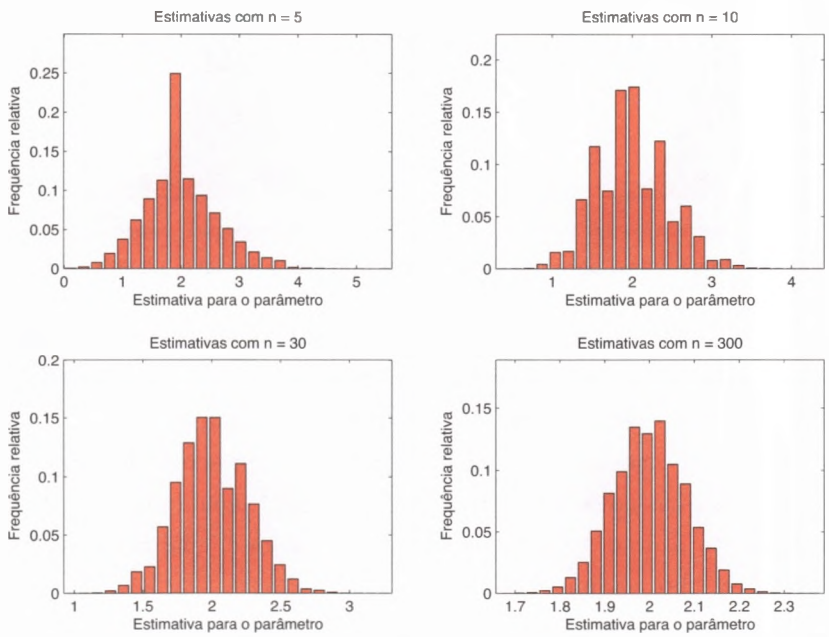


Figura 5.3: Simulações de Monte Carlo para o MV do parâmetro de uma variável aleatória de Poisson.

Finalmente, vamos apresentar os resultados de simulações de Monte Carlo onde o processo gerador de dados não é mais uma variável aleatória com apenas um parâmetro livre. Vamos considerar agora que as amostras geradas obedecem a uma variável aleatória gamma, com parâmetros $\alpha = 4$ e $\beta = 6$. Novamente

iremos gerar amostras de tamanhos $n = 5, 10, 30$ e 300, e para cada amostra gerada iremos estimar os parâmetros α e β via máxima verossimilhança. Os histogramas para as estimativas para o parâmetro α estão apresentadas na Figura 5.4, enquanto os histogramas para as estimativas para o parâmetro β estão apresentadas na Figura 5.5. As médias das estimativas para o parâmetro α são 9.6569, 5.5662, 4.4180, 4.0373, enquanto os desvios padrões das estimativas são 17.1019, 3.3447, 1.2165, 0.3220, respectivamente para os tamanhos de amostra $n = 5, 10, 30$ e 300. As médias das estimativas para o parâmetro β são 4.8076, 5.4084, 5.8179, 5.9830, enquanto os desvios padrões das estimativas são 3.4731, 2.6112, 1.5373, 0.4999.

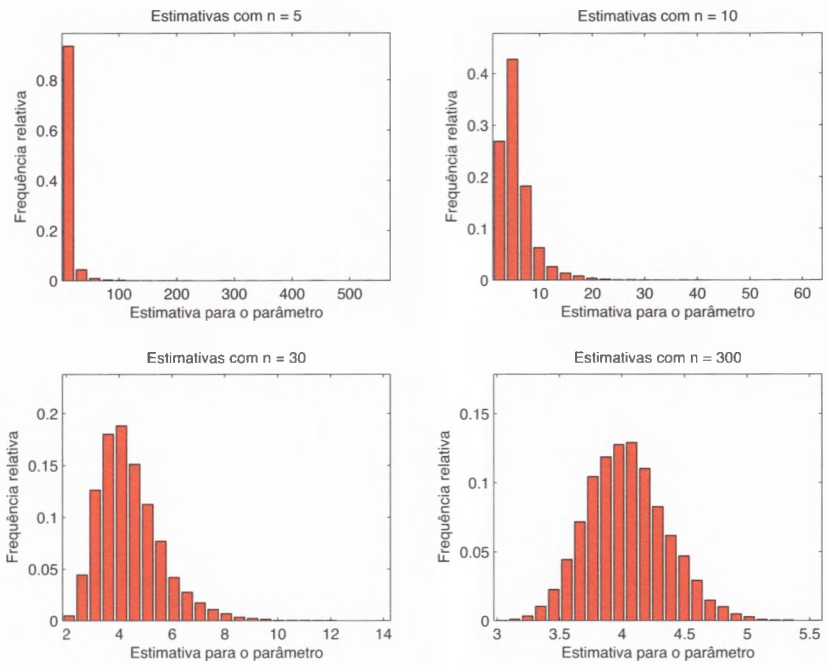


Figura 5.4: Simulações de Monte Carlo para MV do parâmetro α de uma variável aleatória gamma.

Novamente, a partir das simulações de Monte Carlo podemos perceber que: (1) os estimadores para os parâmetros α e β são viesados; (2) o viés das estimativas converge para zero quando a amostra tende para infinito, ou seja, o estimador é consistente; (3) a imprecisão das estimativas (desvio padrão) decresce quando a amostra aumenta; (4) as funções densidade de ambos os estimadores ($\hat{\alpha}$ e $\hat{\beta}$) se aproximam da função densidade de uma variável aleatória normal. No entanto, quando o tamanho da amostra aumenta, o efeito sobre a distribuição das estimativas dos parâmetros α e β vai além da aproximação da normal para cada parâmetro individualmente. De fato, a distribuição conjunta das estimativas converge para uma distribuição normal multivariada. Esse fato analítico é suportado pelos gráficos de dispersão na Figura 5.6, onde as estimativas para o parâmetro β são comparadas às estimativas para o parâmetro α. Note que, existe de fato uma relação negativa entre as estimativas $\hat{\alpha}$ e $\hat{\beta}$. Ou seja, quando a amostra é coletada é tal que a estimativa para o parâmetro α está acima do seu valor real, a estimativa para o parâmetro β fica,

em média, abaixo do seu valor real. A recíproca também é verdadeira. Além disso, a relação negativa entre os estimadores $\hat{\alpha}$ e $\hat{\beta}$ torna-se mais linear quando a amostra aumenta.[5]

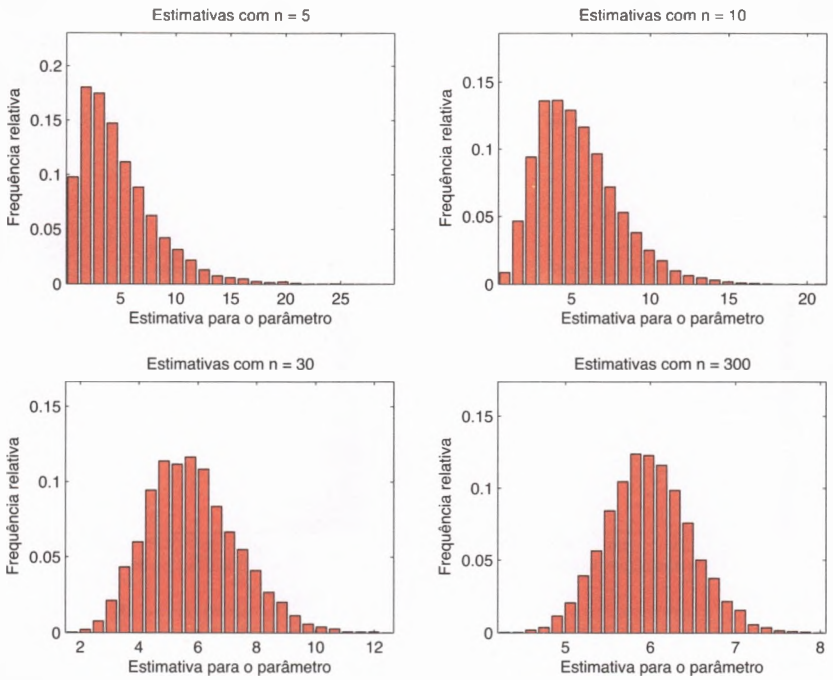


Figura 5.5: Simulações de Monte Carlo para MV do parâmetro β de uma variável aleatória gamma.

5.5 Imprecisão das estimativas

Nas seções anteriores, analisamos o problema de viés e imprecisão dos estimadores. Também estudamos o efeito do tamanho da amostra sobre a distribuição dos estimadores; quando n aumenta, a distribuição dos estimadores tende para uma distribuição normal. Na Seção 5.3, ilustramos por meio de exemplos como detectar, quando possível, a presença ou não de viés nos estimadores. Mostramos que, mesmo sem efetuar simulações de Monte Carlo, é possível mostrar que o estimador de máxima verossimilhança para o parâmetro de uma variável aleatória de Poisson é não viesado; foi possível mostrar que o estimador para o parâmetro de uma variável aleatória exponencial negativa é viesado. A pergunta a ser abordada nesta seção é a seguinte: "é possível inferir a dispersão, ou imprecisão dos estimadores dos parâmetros, analiticamente, sem efetuar simulações de Monte Carlo?". A resposta é sim. Vamos apresentar algumas técnicas comumente utilizadas para encontrar a dispersão, medida pelo desvio padrão (ou variância), de um estimador. Inicialmente, trataremos do cálculo da imprecisão de estimadores de máxima verossimilhança.

[5]Essa relação negativa é uma consequência do termo $\alpha \log \beta$ que aparece na função de máxima verossimilhança para esse caso que está apresentado no Exemplo 6.11.

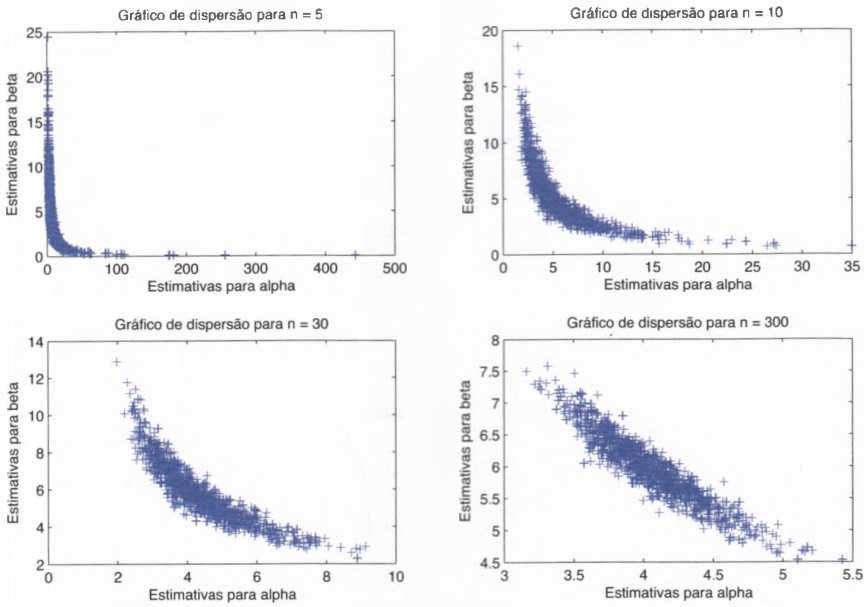


Figura 5.6: Gráficos de dispersão para as estimativas via MV dos parâmetros α e β de uma variável aleatória gamma.

Considere novamente um processo gerador de dados a partir de uma variável aleatória de Poisson, com parâmetro λ qualquer. Vimos na Seção 5.2, no Exemplo 5.6, que a função de log-verossimilhança para uma amostra de tamanho n é dada por

$$l(\lambda) = \log L(\lambda) = \log\left[\prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{x_i}}{x_i!}\right] = -n\lambda + \log\lambda \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} \log x_i!, \tag{5.30}$$

onde $a!$ corresponde ao fatorial do número inteiro a. Imagine que temos uma amostra disponível de tamanho $n = 5$, gerada a partir de um processo gerador de dados com $\lambda = 2$, conforme simulações apresentadas na Seção 5.4. Para essa amostra, a função de log-verossimilhança está apresentada no gráfico superior esquerdo da Figura 5.7.

Os demais gráficos da Figura 5.7 apresentam as funções de log-verossimilhança para amostras geradas com $n = 10$, $n = 30$ e $n = 300$. Note que o formato da função é bastante parecido para os quatro valores de n. A diferença é que a função $l(\lambda)$ torna-se mais "afiada"à medida que a amostra aumenta. De fato, note que para $n = 5$, a diferença absoluta entre $l(1)$ e $l(2)$ é de 5 unidades, enquanto que para $n = 300$, essa diferença passa a ser de 200 unidades. Quanto mais afiada a função de log-verossimilhança, mais preciso é o estimador de máxima verossimilhança, dado que o máximo da função fica mais claro. Temos, então, uma primeira indicação de que uma forma de medir a precisão das estimativas venha de uma medida de quão afiada seja a função de log-verossimilhança. Essa medida é dada justamente pela concavidade da função de log-verossimilhança, que pode ser quantificada pela segunda derivada desta função no ponto de máximo.

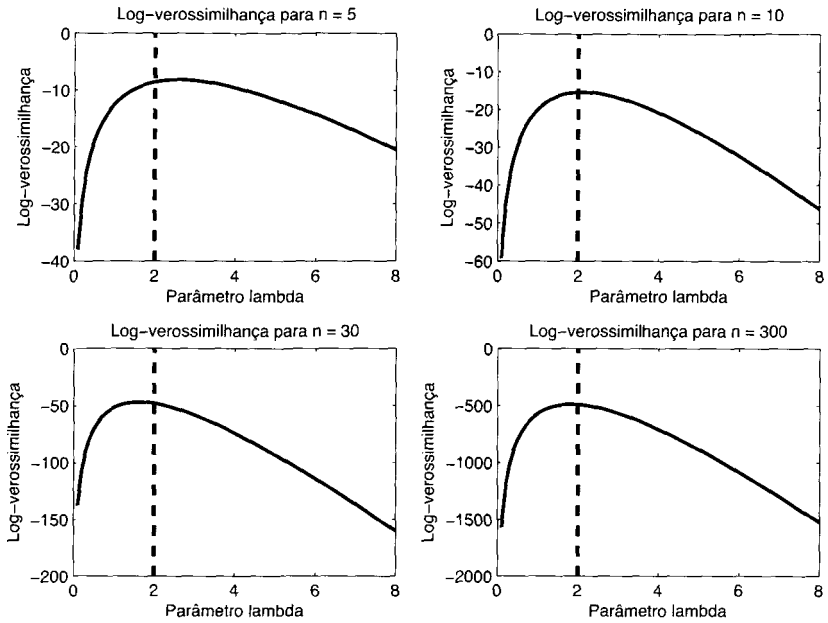


Figura 5.7: Função de log-verossimilhança $l(\lambda)$ para amostras aleatórias simuladas para uma variável aleatória de Poisson.

Com base na Eq. (5.30), a primeira derivada é dada por

$$\frac{d}{d\lambda} l(\lambda) = -n + \frac{1}{\lambda} \sum_{i=1}^{n} x_i,$$

e a segunda derivada tem expressão

$$\frac{d^2}{d\lambda^2} l(\lambda) = -\frac{1}{\lambda^2} \sum_{i=1}^{n} x_i.$$

No ponto de máximo da função $l(\lambda)$, o parâmetro será $\lambda = \bar{x}$, que é justamente o estimador de máxima verossimilhança, pela própria definição desse estimador. Portanto, o valor da segunda derivada no ponto de máximo é dado por

$$\frac{d^2}{d\lambda^2} l(\hat{\lambda}) = -\frac{1}{\bar{x}^2} \sum_{i=1}^{n} x_i = -\frac{1}{\bar{x}^2} n\bar{x} = -\frac{n}{\bar{x}}, \tag{5.31}$$

já que $\sum_{i=1}^{n} x_i = n\bar{x}$. Nos quatro gráficos apresentados na Figura 5.7, os valores para a segunda derivada calculada no ponto de máximo, calculados com base na Eq. (5.31), são -1.9231, -3.5714, -15.7895, -141.5094, para $n = 5, 10, 30$ e 300 respectivamente, indicando que a concavidade da função log-verossimilhança aumenta com o tamanho n da amostra.

Voltemos às simulações de Monte Carlo da Seção 5.4, para estudar o viés do estimador de máxima verossimilhança, no caso da variável aleatória de Poisson. Para cada amostra simulada, estimamos a

estimativa $\hat{\lambda}$ para o parâmetro λ. Vamos acrescentar, para cada amostra gerada, o cálculo da concavidade dada pela segunda derivada da função de log-verossimilhança, no máximo estimado, de acordo com a Eq. (5.31). Vamos calcular a média desses valores de concavidade para as 100.000 amostras geradas, para os diferentes tamanho n de amostra. As médias calculadas para as 100.000 amostras são $\frac{d^2}{d\lambda^2}l(\hat{\lambda}) =$ -2.8246, -5.2773, -15.2525, -150.2370, para $n = 5, 10, 30$ e 300 respectivamente. Para essas médias obtidas, calculemos as medidas $\sqrt{-1/-2.8246} = 0.5950$, $\sqrt{-1/-5.2773} = 0.4353$, $\sqrt{-1/-15.2525} = 0.2561$, $\sqrt{-1/-150.2370} = 0.0816$. Quando comparamos esses valores 0.5950, 0.4353, 0.2561, 0.0816 aos desvios padrões 0.6343, 0.4466, 0.2583, 0.0819, apresentados na Seção 5.4, notamos a quase perfeita similaridade. Isso nos sugere que a variância das estimativas para o parâmetro λ via máxima verossimilhança possa ser calculada simplesmente pela expressão

$$\text{Var}[\hat{\lambda}] = -\frac{1}{\text{E}\left[\frac{d^2}{d\lambda^2}l(\lambda_0)\right]},$$

enquanto o desvio padrão é dado pela expressão

$$\text{Desvio padrão}[\hat{\lambda}] = \sqrt{-\frac{1}{\text{E}\left[\frac{d^2}{d\lambda^2}l(\lambda_0)\right]}},$$

onde λ_0 representa o valor verdadeiro do parâmetro, que no caso das simulações acima tem valor $\lambda_0 = 2$. Nas simulações, o valor esperado $\text{E}[\cdot]$ foi estimado via média dos valores calculados para as 100.000 simulações.

A grandeza $\text{I}(\lambda_0) = \text{E}\left[-\frac{d^2}{d\lambda^2}l(\lambda_0)\right]$ é conhecida como **coeficiente de informação de Fisher**. Pode-se mostrar analiticamente que de fato a variância da distribuição do estimador λ é dada pela expressão $\text{Var}[\hat{\lambda}] = 1/\text{I}(\lambda_0)$, enquanto o desvio padrão é simplesmente $1/\sqrt{\text{I}(\lambda_0)}$. Podemos então utilizar essas expressões para calcular analiticamente o desvio padrão para os exemplos de estimação via máxima verossimilhança para a variável aleatória exponencial negativa e para a variável de Poisson, conforme exemplos a seguir:

Exemplo 5.13 (Cálculo do desvio padrão do estimador do parâmetro livre da distribuição de Poisson) Para a variável aleatória de Poisson, no ponto $\lambda = \lambda_0$, onde λ_0 é o valor verdadeiro do parâmetro, temos a expressão

$$\frac{d^2}{d\lambda^2}l(\lambda_0) = -\frac{1}{\lambda_0}\sum_{i=1}^{n} x_i,$$

e, portanto,

$$\text{E}\left[-\frac{d^2}{d\lambda^2}l(\lambda_0)\right] = \text{E}\left[\frac{1}{\lambda_0{}^2}\sum_{i=1}^{n} x_i\right] = \frac{1}{\lambda_0{}^2}\sum_{i=1}^{n}\text{E}[x_i] = \frac{1}{\lambda_0{}^2}\sum_{i=1}^{n}\lambda_0 = \frac{n}{\lambda_0},$$

usando a definição do coeficiente de informação de Fisher, chegamos no valor para o desvio padrão do estimador de máxima verossimilhança

$$\text{Desvio padrão}[\hat{\lambda}] = \sqrt{\frac{\lambda_0}{n}}.$$

Note que, para $\lambda_0 = 2$, e $n = 5, 10, 30, 300$, temos $\sqrt{\frac{\lambda_0}{n}} = 0.6325, 0.4472, 0.2582$ e 0.0816, que são justamente os valores encontrados nas simulações da Seção 5.4 para o desvio padrão das distribuições dos estimadores $\hat{\lambda}$.

Exemplo 5.14 (Cálculo do desvio padrão do estimador do parâmetro livre da distribuição exponencial negativa) Para a distribuição exponencial negativa, a função de log-verossimilhança, com base em uma amostra de tamanho n, tem expressão

$$l(\lambda) = n \log \lambda - \lambda \sum_{i=1}^{n} x_i.$$

A primeira derivada em um ponto qualquer λ é dada por

$$\frac{d}{d\lambda} l(\lambda) = \frac{n}{\lambda} - \sum_{i=1}^{n} x_i,$$

e a segunda derivada tem expressão

$$\frac{d^2}{d\lambda^2} l(\lambda) = -\frac{n}{\lambda^2},$$

e tem-se o coeficiente de informação de Fisher dado por

$$\mathrm{E}\Big[-\frac{d^2}{d\lambda^2} l(\lambda_0)\Big] = \frac{n}{\lambda_0^2},$$

O desvio padrão da distribuição para o estimador $\hat{\lambda}$ tem valor $\lambda_0/\sqrt{n}$. Nas simulações da Seção 5.4, o exemplo com a variável exponencial negativa supôs que $\lambda_0 = 3$. Com base nesse valor e nas amostras de tamanho $n = 5, 10, 30, 300$, o desvio padrão da distribuição do estimador assume valores 1.3416, 0.9487, 0.5477 e 0.1732. O leitor pode observar que esses valores coincidem com os valores apresentados para os desvios padrões nas simulações na Seção 5.4. Na próxima seção, estenderemos os resultados do cálculo da imprecisão das estimativas via máxima verossimilhança para distribuições onde há mais de um parâmetro livre.

5.6 Estimação via máxima verossimilhança no caso geral

Nas seções anteriores, apresentamos diversos exemplos e simulações de Monte Carlo para ilustrar a utilização do estimador de máxima verossimilhança. Nesses exemplos, abordamos alguns pontos da implementação desse estimador para encontrar os parâmetros de distribuições paramétricas simples. Analisando os exemplos apresentados na seção anterior, podemos visualizar um padrão geral de procedimento, que corresponde justamente ao processo de estimação via máxima verossimilhança amplamente coberto na literatura estatística. Nesta seção, discutimos as linhas gerais dos estimadores de máxima verossimilhança.

Considere então uma variável aleatória Y_i, com função densidade $f(y_i;\theta)$, onde θ é um vetor de k parâmetros livres, que devem ser estimados a partir de uma amostra disponível $y_1, \ldots, y_n$, de tamanho n. Supondo que as n observações na amostra são independentes, a função de verossimilhança pode ser escrita como

$$L(\theta) = \prod_{i=1}^{n} f(y_i;\theta). \tag{5.32}$$

O estimador de máxima verossimilhança do vetor de parâmetros θ corresponde ao vetor que maximiza a função de verossimilhança $L(\theta)$ a partir da amostra $y_1, \ldots, y_n$. Por razões computacionais e analíticas, ao invés de maximizarmos diretamente a função $L(\theta)$, é preferível maximizar o logaritmo da função $L(\theta)$, que é conhecido como função de log-verossimilhança (*log-likelihood function*). A função de log-verossimilhança pode então ser escrita como

$$l(\theta) = \sum_{i=1}^{n} \log f(y_i;\theta). \tag{5.33}$$

Portanto, podemos escrever o estimador de máxima verossimilhança $\hat{\theta}$ como

$$\hat{\theta} = \arg\max_{\theta \in \Theta} \sum_{i=1}^{n} \log f(y_i;\theta), \tag{5.34}$$

onde Θ é o espaço de parâmetros. O procedimento de máxima verossimilhança é utilizado tanto em modelos mais simples, conforme visto nesse capítulo, quanto em modelos mais complexos, como é o caso dos modelos de distribuições combinadas (*mixture models*) e os modelos de distribuições intervalares que serão apresentados no Capítulo 7.

Além da estimativa pontual dos parâmetros no vetor $\hat{\theta}$, é interessante também ter uma ideia da imprecisão decorrente dessa estimação. A ideia é que, quanto maior o tamanho da amostra, mais informações temos sobre o parâmetro que estamos querendo estimar e obviamente menor o grau de imprecisão. Em todo o caso, é importante termos uma indicação numérica para essa imprecisão. Felizmente, a literatura estatística já apresenta vários resultados consolidados sobre como calcular esse grau de incerteza na estimativa paramétrica obtida, conforme visto na seção anterior, para distribuições com apenas um

parâmetro livre. De fato, dado que o vetor estimado $\hat{\theta}$ é uma função de variáveis aleatórias $y_1, \ldots, y_n$, tem-se que $\hat{\theta}$ também é uma variável aleatória. Portanto, para $\hat{\theta}$ vale toda a discussão sobre variáveis aleatórias, apresentada ao longo Capítulo 3. A pergunta então é: "dado que $\hat{\theta}$ é uma variável aleatória, qual a distribuição de $\hat{\theta}$? Além disso, já que estamos interessados na incerteza do estimador de máxima verossimilhança, qual a variância (ou o desvio padrão) da distribuição de $\hat{\theta}$?"Essas perguntas não possuem respostas simples na grande maioria das funções $f(y_i; \theta)$ comumentes encontradas na prática. Felizmente, alguns resultados existem para as situações onde o tamanho da amostra n é suficientemente grande. Esses resultados são conhecidos como **resultados assintóticos**, ou resultados advindos da **teoria assintótica**. Na prática, mesmo com amostras com tamanhos não muito grandes, podemos utilizar os resultados assintóticos, supondo que esses resultados são boas aproximações para as distribuições reais, para as amostras disponíveis.[6]

Pode-se mostrar que, quando n é grande,[7] a variável aleatória $\hat{\theta}$ é consistente e tem distribuição que se aproxima de uma distribuição normal multivariada discutida na Seção 4.5, e essa aproximação é tão melhor, quanto maior for o tamanho n da amostra. Além disso, a matriz de variância-covariância Σ dessa distribuição normal multivariada pode ser aproximada por

$$\Sigma \approx I(\theta_0)^{-1}, \tag{5.35}$$

onde $I(\theta_0)$ é a **matriz de informação de Fisher**.

Essa matriz é dada por

$$\begin{aligned} I(\theta_0) &= -\mathrm{E}\left[\frac{\partial^2}{\partial\theta\partial\theta'} l(\theta)\Big|_{\theta=\theta_0}\right] \\ &= -\mathrm{E}\begin{bmatrix} \frac{\partial^2}{\partial\theta_1^2} l(\theta)\Big|_{\theta=\theta_0} & \frac{\partial^2}{\partial\theta_1\partial\theta_2} l(\theta)\Big|_{\theta=\theta_0} & \cdots & \frac{\partial^2}{\partial\theta_1\partial\theta_k} l(\theta)\Big|_{\theta=\theta_0} \\ \frac{\partial^2}{\partial\theta_2\partial\theta_1} l(\theta)\Big|_{\theta=\theta_0} & \frac{\partial^2}{\partial\theta_2^2} l(\theta)\Big|_{\theta=\theta_0} & \cdots & \frac{\partial^2}{\partial\theta_2\partial\theta_k} l(\theta)\Big|_{\theta=\theta_0} \\ \cdots & \cdots & \cdots & \cdots \\ \frac{\partial^2}{\partial\theta_k\partial\theta_1} l(\theta)\Big|_{\theta=\theta_0} & \frac{\partial^2}{\partial\theta_k\partial\theta_2} l(\theta)\Big|_{\theta=\theta_0} & \cdots & \frac{\partial^2}{\partial\theta_k^2} l(\theta)\Big|_{\theta=\theta_0} \end{bmatrix}, \end{aligned} \tag{5.36}$$

onde, como vimos, $l(\theta)$ corresponde à função de log-verossimilhança na Eq. (5.33) e o parâmetro θ_0 corresponde ao valor verdadeiro do vetor de parâmetros que estamos querendo estimar. No caso onde θ contém um único parâmetro livre, como é o caso das distribuições exponencial negativa, de Poisson, geométrica, de Rayleigh, a matriz de informação de Fisher coincide com o coeficiente de informação de

[6] Uma série de resultados existem para melhorar essas aproximações para amostras pequenas. Essas aproximações são conhecidas com aproximações de ordem maior (*higher order approximations*). Uma outra alternativa é a utilização de métodos de reamostragem (*bootstrapping methods*).

[7] Na verdade, esses resultados de aproximação se utilizam do conceito de limites quando n tende para o infinito (escrevemos $n \to \infty$). A ideia é que, quanto maior o valor de n, melhores são as aproximações utilizadas. Os resultados gerais de aproximações de variáveis aleatórias são foco da área de teoria assintótica.

Fisher, visto na seção anterior. Logo, a matriz de informação de Fisher é a generalização do coeficiente de informação de Fisher para o caso onde θ possui mais de uma dimensão.

Nota 5.1 Note que a matriz de informação de Fisher (assim como o coeficiente de informação de Fisher) depende dos valores verdadeiros dos parâmetros que apenas a natureza conhece. Uma vez que em geral desejamos estimar a matriz de variância-covariância, o procedimento usual é substituir os valores verdadeiros dos parâmetros pelos seus valores estimados. Esse procedimento é bem justificado pelo fato que podemos mostrar que os parâmetros estimados convergem em probabilidade para seus valores reais. Então, o procedimento pode ser resumido da seguinte forma: (1) Encontramos os parâmetros que maximizam a função de log-verossimilhança; (2) Calculamos a matriz de informação de Fisher, como se tivéssemos os parâmetros verdadeiros, como apresentada na Eq. (5.36); (3) Substituímos os parâmetros verdadeiros pelos parâmetros estimados; (4) Invertemos essa matriz para encontrarmos a matriz de variância-covariância; (5) Finalmente, a estimativa para o desvio padrão da variável aleatória $\hat{\theta}_i$, onde $\hat{\theta}_i$ é o estimador para o i-ésimo parâmetro livre no vetor θ, é dada por $\sqrt{\widehat{\Sigma}_{i,i}}$. Essa estimativa para o desvio padrão de um estimador é conhecida como **erro padrão** do estimador. Diversos softwares estatísticos e econométricos fornecem, juntamente com as estimativa pontuais dos parâmetros estimados via máxima verossimilhança, os erros padrões dessas estimativas. Com isso, o usuário pode ter uma ideia da precisão nas estimações.

Uma situação prática que ocorre é quando o cômputo analítico da matriz de informação de Fisher não é simples e é necessário recorrer a aproximações. Uma forma comumente utilizada para estimar a matriz $I(\theta)$ é por meio do estimador (sem a utilização do valor esperado das segundas derivadas)

$$\begin{aligned}\hat{I}(\hat{\theta}) &= -\frac{\partial^2}{\partial\theta\partial\theta'}l(\theta)\Big|_{\theta=\hat{\theta}} \\ &= -\begin{bmatrix} \frac{\partial^2}{\partial\theta_1^2}l(\theta)\Big|_{\theta=\hat{\theta}} & \frac{\partial^2}{\partial\theta_1\partial\theta_2}l(\theta)\Big|_{\theta=\hat{\theta}} & \cdots & \frac{\partial^2}{\partial\theta_1\partial\theta_k}l(\theta)\Big|_{\theta=\hat{\theta}} \\ \frac{\partial^2}{\partial\theta_2\partial\theta_1}l(\theta)\Big|_{\theta=\hat{\theta}} & \frac{\partial^2}{\partial\theta_2^2}l(\theta)\Big|_{\theta=\hat{\theta}} & \cdots & \frac{\partial^2}{\partial\theta_2\partial\theta_k}l(\theta)\Big|_{\theta=\hat{\theta}} \\ \cdots & \cdots & \cdots & \cdots \\ \frac{\partial^2}{\partial\theta_k\partial\theta_1}l(\theta)\Big|_{\theta=\hat{\theta}} & \frac{\partial^2}{\partial\theta_k\partial\theta_2}l(\theta)\Big|_{\theta=\hat{\theta}} & \cdots & \frac{\partial^2}{\partial\theta_k^2}l(\theta)\Big|_{\theta=\hat{\theta}} \end{bmatrix}.\end{aligned}$$

Portanto, uma vez estimado o vetor de parâmetros θ, obtendo-se $\hat{\theta}$, o próximo passo é calcular as segundas derivadas parciais da função de log-verossimilhança em relação aos parâmetros livres do modelo. Em modelos paramétricos simples (o que é o caso da normal, Poisson, gamma etc.) essas derivadas podem ser facilmente escritas em forma fechada, de forma que o estimador $\hat{I}(\hat{\theta})$ na expressão acima pode ser facilmente computado, como apresentado na Nota 5.1. Porém, em modelos mais complexos, como é o caso dos modelos de combinação de distribuições, o cálculo das segundas derivadas, apesar de permitir a obtenção de expressões fechadas, podem envolver operações analíticas tediosas. Para evitar esse problema, uma solução simples é a utilização do cálculo das segundas derivadas numericamente que serão calculadas usando os valores dos parâmetros estimados. A utilização de técnicas numéricas tem a vantagem de ser facilmente aplicáveis para a maioria dos problemas.

Exemplo 5.15 (Variável aleatória normal) Considere uma amostra aleatória $X_1, \ldots, X_n$, com observações independentes. Cada observação tem distribuição normal, com média μ e variância σ^2. A função densidade de probabilidade de cada observação individualmente é dada por

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}[x-\mu]^2}, \quad \text{para } x \in \Re.$$

A função de log-verossimilhança tem expressão

$$l(\theta) = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}[x_i - \mu]^2,$$

onde $\theta = [\mu\ \sigma]'$. Diferenciando em relação a μ e igualando a zero, obtemos

$$\begin{aligned}\frac{\partial l(\theta)}{\partial \mu} = \frac{2}{2\sigma^2}\sum_{i=1}^{n}[x_i - \mu] = 0 \ &\Rightarrow \Big[\sum_{i=1}^{n} x_i\Big] - n\mu = 0 \\ &\Rightarrow \hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}.\end{aligned}$$

Diferenciando em relação a σ^2 e igualando a zero,

$$\begin{aligned}\frac{\partial l(\theta)}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}\sum_{i=1}^{n}[x_i - \mu]^2 = 0 \ &\Rightarrow -\frac{n}{2} + \frac{1}{2\sigma^2}\sum_{i=1}^{n}[x_i - \mu]^2 = 0 \\ &\Rightarrow \sigma^2 = \frac{1}{n}\sum_{i=1}^{n}[x_i - \mu]^2 \\ &\Rightarrow \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}[x_i - \bar{x}]^2.\end{aligned}$$

Para as derivadas parciais de segunda ordem, temos

$$\begin{aligned}\frac{\partial l^2(\theta)}{\partial \mu^2} &= -\frac{n}{\sigma^2}, \\ \frac{\partial l^2(\theta)}{\partial (\sigma^2)^2} &= \frac{n}{2(\sigma^2)^2} - \frac{1}{(\sigma^2)^3}\sum_{i=1}^{n}[x_i - \mu]^2, \\ \frac{\partial l^2(\theta)}{\partial \mu \partial \sigma^2} &= \frac{\partial l^2(\theta)}{\partial \sigma^2 \partial \mu} = -\frac{1}{(\sigma^2)^2}\sum_{i=1}^{n}[x_i - \mu].\end{aligned} \tag{5.37}$$

Precisamos agora determinar o valor das derivadas parciais na Eq. (5.37) no ponto $\theta_0 = \left[\mu_0 \;\; \sigma_0^2\right]'$. Portanto,

$$\left.\frac{\partial l^2(\theta)}{\partial \mu^2}\right|_{\theta=\theta_0} = -\frac{n}{\sigma_0^2},$$

$$\left.\frac{\partial l^2(\theta)}{\partial (\sigma^2)^2}\right|_{\theta=\theta_0} = \frac{n}{2(\sigma_0{}^2)^2} - \frac{1}{(\sigma_0{}^2)^3}\sum_{i=1}^{n}[x_i - \mu_0]^2,$$

$$\left.\frac{\partial l^2(\theta)}{\partial \mu \partial \sigma^2}\right|_{\theta=\theta_0} = \left.\frac{\partial l^2(\theta)}{\partial \sigma^2 \partial \mu}\right|_{\theta=\theta_0} = -\frac{1}{(\sigma_0{}^2)^2}\sum_{i=1}^{n}[x_i - \mu_0].$$

A matriz de informação de Fisher $I(\theta_0)$ nesse caso será

$$\begin{aligned} I(\theta_0) &= \mathrm{E}\left[\begin{array}{cc} \frac{n}{\sigma_0^2} & \frac{1}{(\sigma_0{}^2)^2}\sum_{i=1}^{n}[X_i - \mu_0] \\ \frac{1}{(\sigma_0{}^2)^2}\sum_{i=1}^{n}[X_i - \mu_0] & \frac{1}{(\sigma_0{}^2)^3}\sum_{i=1}^{n}[X_i - \mu_0]^2 - \frac{n}{2(\sigma_0{}^2)^2} \end{array}\right] \\ &= \left[\begin{array}{cc} \frac{n}{\sigma_0^2} & \frac{1}{(\sigma_0{}^2)^2}\sum_{i=1}^{n}[\mathrm{E}[X_i] - \mu_0] \\ \frac{1}{(\sigma_0{}^2)^2}\sum_{i=1}^{n}[\mathrm{E}[X_i] - \mu_0] & \frac{1}{(\sigma_0{}^2)^3}\sum_{i=1}^{n}\mathrm{E}[(X_i - \mu_0)^2] - \frac{n}{2(\sigma_0{}^2)^2} \end{array}\right] \\ &= \left[\begin{array}{cc} \frac{n}{\sigma_0^2} & 0 \\ 0 & \frac{n}{2(\sigma_0{}^2)^2} \end{array}\right]. \end{aligned}$$

Na última passagem acima, utilizamos o fato de que $\mathrm{E}[X_i] = \mu_0$ e o fato de que $\mathrm{E}[(X_i - \mu_0)^2] = \sigma_0^2$. A inversa da matriz de informação de Fisher nos fornece a matriz de variância-covariância $\Sigma_{\hat{\theta}}$ do estimador $\hat{\theta}$ de máxima verossimilhança. Portanto,

$$\Sigma_{\hat{\theta}} = I(\theta_0)^{-1} = \left[\begin{array}{cc} \frac{\sigma_0^2}{n} & 0 \\ 0 & \frac{2(\sigma_0{}^2)^2}{n} \end{array}\right].$$

Note que a matriz de variância-covariância $\Sigma_{\hat{\theta}}$ é diagonal, e portanto os estimadores $\hat{\mu}$ e $\hat{\sigma}^2$ são não correlacionados entre si. Isso não é verdade para a maioria dos casos encontrados na prática. Em geral, os elementos fora da diagonal principal da matriz de variância-covariância não são nulos, havendo portanto uma correlação entre os estimadores. Essa situação aparece explicitamente no Exemplo 6.11 para uma amostra aleatória com observações seguindo uma distribuição gamma, onde os estimadores são negativamente correlacionados, mas já foi descoberta nesse capítulo por meio de simulações de Monte Carlo na Figura 5.6.

Esta seção apresentou o tratamento da imprecisão das estimativas para o caso dos estimadores de máxima verossimilhança. Uma especial atenção foi dada a esse tipo de estimador, pois máxima verossimilhança vai ser o principal método de estimação nos capítulos que se seguem. Para estimadores via método de momentos, existem tratamentos similares para calcular a variância e o desvio padrão das estimativas. Esses métodos não serão cobertos neste livro. O leitor interessado pode recorrer a referências como Bickel e Doksum (2000).

5.7 Inferência e atualização Bayesiana

Nessa seção, vamos fazer uma pequena introdução às técnicas de inferência Bayesiana. Como já dissemos anteriormente, na abordagem Bayesiana, os parâmetros das distribuições são tratados como variáveis aleatórias, da mesma forma que as observações na amostra. Nesse caso, parte-se de uma distribuição inicial, conhecida como **distribuição a priori**, para os parâmetros do modelo. Uma vez observada uma amostra de dados, a inferência Bayesiana permite que a distribuição dos parâmetros seja atualizada, combinando-se o conhecimento inicial, resumido na distribuição a priori, e as informações na amostra. A distribuição atualizada dos parâmetros é conhecida como **distribuição a posteriori**. Esse processo de atualização é conhecido na literatura como **atualização Bayesiana**.[8]

Inicialmente, para exemplificar o processo de atualização Bayesiana, considere uma variável aleatória X, com distribuição binomial, com número de tentativas n conhecido e probabilidade de sucesso p, que precisa ser estimada. A ideia da atualização Bayesiana é inicialmente assumir uma forma paramétrica para a suposição de qual possa ser esse parâmetro desconhecido p. Essa suposição inicial pode vir de um processo de estimação anterior ou de informações de especialistas. Uma forma paramétrica muito utilizada neste caso é supor que p tenha, em princípio, uma distribuição beta, com parâmetros α e β.

Conforme visto na Seção 3.3.10, a variável aleatória beta assume valores entre 0 e 1, de forma que o seu espaço amostral é dado por $\mathbb{X} = (0, 1)$. Por esse motivo, ela pode ser utilizada, em geral, para modelar variáveis aleatórias que representam taxa, como por exemplo a perda em caso de inadimplência (*loss given default*), comumente encontrada em análise de risco de crédito (BIS, 2004).

Devido ao fato de a variável aleatória beta assumir valores entre 0 e 1, é adequado dizer que o parâmetro desconhecido p, de uma variável aleatória binomial, tem distribuição beta, em princípio.[9] Usamos a notação $p \sim \text{Beta}(\alpha, \beta)$, e dizemos que beta é a distribuição *a priori* para o parâmetro desconhecido p. Essa distribuição *a priori* pode ter vindo de um processo de estimação anterior ou de uma compilação de informações de especialistas.

Suponha, por um instante, que a informação que os analistas têm a respeito do parâmetro p possa ser resumida por uma distribuição, *a priori*, $Beta(\alpha, \beta)$, para a qual os parâmetros α e β tenham valores 2.0 e 10.0 respectivamente. Utilizando-se a fórmula para valor esperado de uma variável aleatória beta, $E[p] = \alpha/(\alpha + \beta)$, isso significa que os analistas acreditam que o parâmetro p localiza-se em torno do valor $p = 0.17$. Por outro lado, se os analistas escolherem parâmetros α e β iguais a 5.0 e 0.8 respectivamente, isso implica que os analistas acreditam que o parâmetro p localiza-se em torno do valor 0.86.

Dada informação *a priori* a respeito do parâmetro p, à medida que formos observar valores para variável aleatória X (esta com distribuição binomial), poderemos atualizar o nosso conhecimento a respeito da

[8] Do inglês, *Bayesian update*.

[9] Lembre-se que o parâmetro p é uma probabilidade, e, portanto, tem que assumir valores no intervalo $(0, 1)$.

probabilidade de sucesso p. Nesse caso, é interessante ter um método para combinar a informação *a priori* com as informações a partir dos dados.

Seja $x_1, x_2, \ldots, x_m$ o conjunto de m valores observados para variável aleatória X. Imagine, por exemplo, que X seja o número de unidades vendidas de um determinado produto por semana, sendo que n é o número total de unidades disponibilizados a cada semana. Então $x_1, x_2, \ldots, x_m$ podem ser as quantidades de unidades vendidas para m semanas consecutivas, por exemplo. Nesse caso, a probabilidade de observar esses valores, dado um determinado valor para o parâmetro p, é igual a

$$\text{Prob}[x_1, x_2, \ldots, x_m|p] = \prod_{i=1}^{m} \binom{n}{x_i} p^{x_i}(1-p)^{n-x_i}, \tag{5.38}$$

onde estamos supondo independência entre os m valores observados.[10] Sabemos, *a priori*, que p tem distribuição Beta(α, β), com função densidade $f(p)$ de acordo com (3.45). Podemos então recorrer ao Teorema de Bayes (Teorema 4.2), para atualizar o nosso conhecimento a respeito de p, com base nos dados observados. Para maiores detalhes, vide Tanner (1996) ou Gelman et al. (1995). Em linhas gerais, a nova distribuição $f(p|x_1, x_2, \ldots, x_m)$ para o parâmetro p, de acordo com o Teorema de Bayes, será dada por

$$\begin{aligned} f(p|x_1, x_2, \ldots, x_m) &= \frac{\text{Prob}[x_1, x_2, \ldots, x_m|p] \times f(p)}{\int_{p\in(0,1))} \Big[\text{Prob}[x_1, x_2, \ldots, x_m|p] \times f(p)\Big] dp} \\ &= \left[\prod_{i=1}^{m} \binom{n}{x_i} p^{x_i}(1-p)^{n-x_i}\right] \times \left[\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1}\right] \times \frac{1}{C}, \end{aligned} \tag{5.39}$$

onde p tem uma distribuição $Beta(\alpha, \beta)$, C é uma constante para fazer com que a densidade $f(p|x_1, \ldots, x_m)$ tenha integral igual a 1. Note que o denominador no quociente do lado direto da Eq. (5.39) independe do parâmetro de interesse p. Para simplificar as equações, sem prejuízo das derivações, é muito comum na literatura de inferência Bayesiana reescrever essa equação da seguinte forma

$$f(p|x_1, x_2, \ldots, x_m) \propto \left[\prod_{i=1}^{m} \binom{n}{x_i} p^{x_i}(1-p)^{n-x_i}\right] \times \left[\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1}\right]. \tag{5.40}$$

O símbolo "$\propto$" significa justamente que o lado direito da equação é igual ao lado esquerdo vezes uma contante, que é de certa forma 'irrelevante' para as análises. Em geral, o termo de proporcionalidade, que está ausente explicitamente nas equações de atualizações Bayesianas, poderia ser recuperado por meio do cálculo de integrais para que as funções de densidade e as funções de frequência, respectivamente, integrassem e somassem para um. Em geral, do ponto de vista prático, tanto analiticamente quanto computacionalmente, desconhecer os valores do termo de proporcionalidade é de fato irrelevante e as metodologias existentes não necessitam desses valores explicitamente.

[10] Vamos supor que não há correlação temporal entre o número de unidades vendidas em semanas consecutivas.

Rearranjando os termos na Eq. (5.39), podemos reescrever

$$f(p|x_1, x_2, \ldots, x_m) = \left[\frac{\Gamma(mn+\alpha+\beta)}{\Gamma(\sum_{i=1}^m x_i+\alpha)\Gamma(mn-\sum_{i=1}^m x_i+\beta)}\right] p^{\left[\sum_{i=1}^m x_i+\alpha-1\right]}(1-p)^{\left[mn-\sum_{i=1}^m x_i+\beta-1\right]}. \tag{5.41}$$

O termo C está contabilizado na equação acima, de forma que a integral de $f(p|x_1, x_2, \ldots, x_m)$ para $p \in (0,1)$ é igual a um. A Eq. (5.41) ilustra um fato muito importante: observe que a nova distribuição para o parâmetro p, depois que combinamos a informação *a priori* com a informação a partir dos dados $x_1, \ldots, x_m$, é também uma distribuição beta, mas agora temos novos valores para os parâmetros dessa distribuição. Os novos parâmetros são $\left[\sum_{i=1}^m x_i + \alpha\right]$ e $\left[mn - \sum_{i=1}^m x_i + \beta\right]$. Portanto, escrevemos $p \sim \text{Beta}([\sum_{i=1}^m x_i + \alpha], [mn - \sum_{i=1}^m x_i + \beta])$, e a nova distribuição é conhecida como distribuição *a posteriori*. Portanto, depois de combinarmos a informação *a priori* com a informação a partir dos dados $x_1, \ldots, x_m$, obtemos não somente uma estimativa pontual para o parâmetro p, mas também toda uma distribuição (*a posteriori*), que expressa o grau de incerteza a respeito da nossa estimativa sobre p.

5.7.1 Média e moda da distribuição *a posteriori*

A partir da distribuição *a posteriori*, usando a Eq. (3.46), podemos obter o valor esperado para o parâmetro p (lembrando que agora é ele considerado uma variável aleatória), condicionando-se aos dados observados na amostra $x_1, \ldots, x_m$,

$$\text{E}[p/x_1, \ldots, x_m] = \frac{\left[\sum_{i=1}^m x_i + \alpha\right]}{\left[\sum_{i=1}^m x_i + \alpha\right] + \left[mn - \sum_{i=1}^m x_i + \beta\right]} = \frac{\sum_{i=1}^m x_i + \alpha}{mn + \alpha + \beta}. \tag{5.42}$$

Essa média da distribuição *a posteriori* pode ser considerada como um estimador pontual para o parâmetro p. A Eq. (5.42) pode ser reescrita como

$$\text{E}[p/x_1, \ldots, x_m] = (1-w) \times \frac{\alpha}{\alpha+\beta} + w \times \frac{\sum_{i=1}^m x_i}{mn}, \tag{5.43}$$

onde $w = (mn)/(\alpha+\beta+mn)$. Portanto, nota-se que a média da distribuição *a posteriori* para o parâmetro p pode ser escrita como uma média ponderada entre a média da distribuição *a priori*, dada pela expressão $\alpha/(\alpha+\beta)$, e a estimativa para o parâmetro p, caso utilizássemos um método de estimação via máxima verossimilhança ou método de momentos.[11] O estimador de máxima verossimilhança de p é dado justamente pelo quociente

$$\hat{p}_{\text{MV}} = \frac{\sum_{i=1}^m x_i}{mn}. \tag{5.44}$$

[11] O estimador de máxima verossimilhança coincide com o estimador de método de momentos nesse caso.

O peso w da parcela correspondente ao estimador de máxima verossimilhança $\hat{p}_{\text{MV}}$ tende para 1 quando mn tende para o infinito. Portanto, à medida que o tamanho da amostra m aumenta, a parcela w se aproxima de 1, e a média da distribuição *a posteriori* se aproxima do estimador de máxima verossimilhança. Portanto, à medida que o tamanho da amostra aumenta, o estimador Bayesiano, dado pelo valor esperado da distribuição *a posteriori*, converge para o estimador de máxima verossimilhança.

Por outro lado, observe que o peso $(1 - w)$, para a parcela correspondente à média $\alpha(\alpha + \beta)$ da distribuição *a priori*, aumenta quando a soma $\alpha + \beta$ também aumentam. Para uma variável aleatória p com distribuição $Beta(\alpha, \beta)$, usando a Eq. (3.46), a variância é dada pela equação

$$\text{Var}[p] = \frac{\alpha\beta}{(\alpha + \beta + 1)(\alpha + \beta)^2}. \tag{5.45}$$

Portanto, quando a soma $(\alpha + \beta)$ aumenta a variância *a priori* de p cai. Temos então alguns fatos interessantes nesse exemplo, que ocorrem de forma similar em muitas das aplicações da abordagem Bayesiana. Em primeiro lugar, o estimador Bayesiano (no exemplo até agora, dado pela média da distribuição *a posteriori*) é uma combinação entre a estimativa a partir puramente da informação *a priori* (no exemplo, essa estimativa seria a razão $\alpha/(\alpha + \beta)$) e a estimativa via máxima verossimilhança. Em segundo lugar, a importância da estimativa via máxima verossimilhança aumenta à medida que o tamanho da amostra também aumenta. Finalmente, a importância da informação *a priori* é maior quando a imprecisão nessa informação é menor. No nosso exemplo, a imprecisão na informação *a priori* para o parâmetro p é dado justamente pela variância da distribuição *a priori*, de acordo com a Eq. (5.45).

Uma alternativa para estimação pontual do parâmetro p, a partir da distribuição *a posteriori*, é a utilização da moda da função densidade de probabilidade $f(p/x_1, \ldots, x_m)$. Nesse caso, busca-se o valor mais frequente para p que é aquele valor de p que maximiza a função densidade correspondente à variável aleatória beta. Derivando-se a função $f(p/x_1, \ldots, x_m)$ e igualando-se a zero, obtém-se o ponto de máximo

$$p_{\text{máx}} = \frac{\left[\sum_{i=1}^{m} x_i + \alpha\right] - 1}{\left[\sum_{i=1}^{m} x_i + \alpha\right] + \left[mn - \sum_{i=1}^{m} x_i + \beta\right] - 2}. \tag{5.46}$$

A Eq. (5.46) pode ser reescrita como

$$\text{E}[p/x_1, \ldots, x_m] = (1 - w) \times \frac{\alpha - 1}{\alpha + \beta - 2} + w \times \frac{\sum_{i=1}^{m} x_i}{mn}, \tag{5.47}$$

onde novamente $w = (mn)/(\alpha + \beta + mn)$. Portanto, podemos observar que a moda da distribuição *a posteriori*, similarmente ao que acontece no caso da média dessa distribuição, é uma média ponderada entre a moda da distribuição *a priori* e o estimador de máxima verossimilhança $\hat{p}_{\text{MV}}$ em (5.44). A moda da distribuição *a priori* nesse caso é igual a $(\alpha - 1)/(\alpha + \beta - 2)$. Portanto, se utilizarmos tanto a média quanto a moda da distribuição *a posteriori* como estimador pontual para o valor de p, as conclusões

gerais sobre a importância de cada parcela na média ponderada são as mesmas. A importância da parcela correspondente ao estimador de máxima verossimilhança aumenta quando o tamanho da amostra aumenta. A importância da média ou moda da distribuição *a priori* aumenta quando reduzimos a imprecisão (variância) na informação *a priori*.

O exemplo anterior ilustra bem o processo de atualização Bayesiana, num modelo que considera a distribuição binomial (para os dados observados) e a distribuição beta (para o parâmetro desconhecido). Conforme vimos, os parâmetros para a distribuição *a posteriori* combinam a informação *a priori* com a informação a partir dos dados. Observamos também que, tanto a distribuição *a priori* quanto a distribuição *a posteriori* possuem distribuições pertencentes à mesma família (no caso, à família de distribuições beta), mas com diferentes valores para os parâmetros. Devido ao fato de as distribuições *a posteriori* e *a priori* pertencerem à mesma família, o par binomial × beta é conhecido como **distribuições conjugadas**. Outros exemplos de distribuições conjugadas são os pares Poisson × gamma, normal (variância conhecida) × normal, normal (média conhecida) × normal-inversa.[12]

Infelizmente, modelos envolvendo distribuições conjugadas são mais a exceção do que a regra. Em geral, situações onde a distribuição *a priori* e a distribuição *a posteriori* não pertencem à mesma família são comuns, o que torna as derivação analíticas bem mais complicadas. Por exemplo, nem sempre é possível encontrar uma forma conhecida para a distribuição *a posteriori*, conforme especificado na Eq. (5.41). De fato, em muitas das aplicações mais interessantes, não é possível escrever explicitamente expressões matemáticas para as distribuições *a posteriori*. Felizmente, a literatura nas últimas três décadas avançou muito no sentido de utilização de procedimentos computacionais intensivos para cálculo de medidas (média e demais momentos, quantis etc.) das distribuições *a posteriori*, sem a necessidade de se conhecerem as formas funcionais dessas distribuições. O leitor pode recorrer a Tanner (1996) para uma discussão sobre o assunto.

5.7.2 Geração de amostras aleatórias incorporando incerteza dos parâmetros

Em muitas aplicações, pode ser de interesse do analista simular valores para a variável observada a partir da distribuição para os dados. Por exemplo, em um modelo de frequências de perdas operacionais, em um determinado mês, o analista pode estar interessado em gerar 10.000 amostras aleatórias do número de assaltos a agências bancárias. Vide Glasserman (2004) para diversos exemplos de simulações com aplicações ao setor financeiro. Vamos supor que o processo aleatório do qual resulta o número de agências assaltadas pode ser descrito por uma distribuição binomial, similarmente ao modelo visto no caso anterior.

Um procedimento comumente empregado é inicialmente estimar o parâmetro desconhecido p para a probabilidade de ocorrência do assalto. Essa estimação pode ocorrer utilizando-se o método de máxima

[12]Com exceção da distribuição normal-inversa, todas as outras distribuições foram introduzidas nas seções 3.2 e 3.3. A distribuição normal-inversa (também conhecida como distribuição de Wald), diferentemente do que o nome poderia sugerir, não descreve a distribuição de uma variável aleatória cuja o recíproco de sua distribuição possui distribuição normal (vide Exercício 5.9).

verossimilhança (que coincide com o método de momentos), ou utilizando-se estimadores a partir da média ou da moda da distribuição *a posteriori*. Conforme vimos no exemplo anterior, para amostras grandes o suficiente, esses três estimadores assumem valores muito próximos. Seja então $\hat{p}$ a estimativa pontual para o parâmetro p, com base em um desses métodos de estimação.

Para gerar então 10.000 mil amostras aleatórias da variável aleatória X, que corresponde ao número de assaltos a agências, geramos retiradas de uma distribuição binomial com parâmetros n (estamos supondo n conhecido) e $\hat{p}$. Implicitamente, estamos desconsiderando toda a incerteza incorrida na estimação de p, e supondo que a estimativa $\hat{p}$ é suficientemente precisa.

Uma das vantagens dos procedimentos Bayesianos é que eles nos fornecem uma maneira simples e direta de simular valores para a variável aleatória X, incorporando-se as imprecisões na estimação do parâmetro p, utilizando-se assim toda a informação contida na distribuição *a posteriori* $f(p|x_1, \ldots, x_m)$. Para isso, procedemos com um esquema de retiradas aleatórias em dois estágios.

Para gerar uma amostra de $M = 10.000$ elementos, a partir de uma variável binomial $\text{Bin}(n, p)$, onde p tem uma distribuição beta de acordo com (5.41), podemos proceder com os passos a seguir:

1. Gerar um número aleatório p^1 a partir da distribuição *a posteriori* $\text{Beta}([\sum_{i=1}^{m} x_i + \alpha], [mn - \sum_{i=1}^{m} x_i + \beta])$.

2. Gerar um número aleatório x^1 a partir da distribuição binomial $\text{Bin}(n, p^1)$.

3. Repetir os passos 1 e 2 um número de $M - 1$ de vezes, e obter dessa forma uma sequência de valores para o número de perdas operacionais $x^1, x^2, \ldots, x^M$.

No processo de geração de números aleatórios acima, estamos automaticamente contabilizando para a incerteza na estimativa do parâmetro p, quando simulamos as amostras para X. Portanto, toda a informação *a posteriori* em $f(p|x_1, \ldots, x_m)$ está sendo utilizada na geração das amostras da distribuição binomial.

5.7.3 Contextualização geral

De maneira mais geral, para representar a abordagem de atualização Bayesiana, considere o problema de estimar um parâmetro desconhecido θ, que pode ser um escalar (como no exemplo acima) ou um vetor de parâmetros (conforme veremos mais adiante), a partir de uma amostra de observações sobre a variável ou as variáveis aleatórias de interesse. Considere uma função densidade de probabilidade *a priori* $p(\theta)$, e seja $L(\theta/y_1, \ldots, y_m)$ a função de verossimilhança de θ condicionada à amostra $y_1, \ldots, y_m$. Os pontos y_i, $i = 1, \ldots, m$, na amostra, também podem ser escalares ou vetores. Nesse contexto, a função densidade de probabilidade *a posteriori* é dada por

$$p(\theta/y_1, \ldots, y_m) = c \times p(\theta) \times L(\theta/y_1, \ldots, y_m), \tag{5.48}$$

onde

$$c = \frac{1}{\int_{\theta \in \Theta}[p(\theta) \times L(\theta/y_1, \ldots, y_m)]d\theta}, \tag{5.49}$$

e Θ é o conjunto de valores possíveis para θ. A Eq. (5.49) corresponde a situações nas quais θ assume valores contínuos. Para valores discretos, a Eq. (5.49) é substituída por

$$c = \frac{1}{\sum_{\theta \in \Theta}[p(\theta) \times L(\theta/y_1, \ldots, y_m)]}. \tag{5.50}$$

Observe que no lado direito da Eq. (5.48) aparece o termo de verossimilhança $L(\theta/y_1, \ldots, y_m)$, no qual o parâmetro θ está condicionado aos dados observados $y_1, \ldots, y_m$. Isso está diretamente de acordo com o Teorema de Bayes (Teorema 4.2), uma vez que algebricamente a função $L(\theta/y_1, \ldots, y_m)$ é igual à função densidade de probabilidade conjunta $f(y_1, \ldots, y_m/\theta)$. No caso de $y_1, \ldots, y_m$ serem variáveis aletórias discretas, $L(\theta/y_1, \ldots, y_m)$ é igual à função frequência conjunta.

5.7.4 População normal com média desconhecida e variância conhecida

Dando continuidade à nossa discussão sobre atualização Bayesiana, considere uma amostra $y_1, \ldots, y_m$ de observações independentes e identicamente distribuídas, de uma distribuição normal com média desconhecida μ (que corresponde ao parâmetro θ na formulação geral acima) e variância σ^2 conhecida. Nesse caso, a função de verossimilhança (algebricamente igual à função densidade conjunta de $y_1, \ldots, y_m$, dado μ), é igual a

$$\begin{aligned} L(\mu/y_1, \ldots, y_m) &= \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2}\frac{(y_i - \mu)^2}{\sigma^2}\right] \\ &= \frac{1}{(2\pi\sigma^2)^{(m/2)}} \exp\left[-\frac{1}{2}\sum_{i=1}^{m}\frac{(y_i - \mu)^2}{\sigma^2}\right]. \end{aligned} \tag{5.51}$$

Dado que σ^2 é conhecido, o termo $\frac{1}{(2\pi\sigma^2)^{(m/2)}}$ não altera o formato da distribuição de verossimilhança nessa formulação. Por esse motivo, iremos omitir esse termo das derivações. Além disso, podemos escrever o termo

$$\begin{aligned} \sum_{i=1}^{m}(y_i - \mu)^2 &= \sum_{i=1}^{m}[(y_i - \bar{y}) + (\bar{y} - \mu)^2] \\ &= \sum_{i=1}^{m}(y_i - \mu)^2 + 2\sum_{i=1}^{m}(y_i - \bar{y}) \times (\bar{y} - \mu) + \sum_{i=1}^{m}(\bar{y} - \mu)^2 \\ &= \sum_{i=1}^{m}(y_i - \mu)^2 + \sum_{i=1}^{m}(\bar{y} - \mu)^2, \end{aligned} \tag{5.52}$$

uma vez que

$$2\sum_{i=1}^{m}(y_i - \bar{y}) \times (\bar{y} - \mu) = 2(\bar{y} - \mu) \times \sum_{i=1}^{m}(y_i - \bar{y}) = 0,$$

onde $\bar{y}$ é a média amostral das observações $y_1, \ldots, y_m$, e $\sum_{i=1}^{m} y_i = m \times \bar{y}$. Portanto, temos

$$\exp\left[-\frac{1}{2}\sum_{i=1}^{m}\frac{(y_i - \mu)^2}{\sigma^2}\right] = \exp\left[-\frac{1}{2}\sum_{i=1}^{m}\frac{(y_i - \bar{y})^2}{\sigma^2}\right] \times \exp\left[-\frac{1}{2}\sum_{i=1}^{m}\frac{(\bar{y} - \mu)^2}{\sigma^2}\right]. \tag{5.53}$$

O termo

$$\exp\left[-\frac{1}{2}\sum_{i=1}^{m}\frac{(y_i - \bar{y})^2}{\sigma^2}\right]$$

é uma contante, uma vez que no conceito de função de verossimilhança $L(\mu/y_1, \ldots, y_m)$, condicionamos o parâmetro aos dados observados e $\overline{y} = \sum_{i=1}^{m} \frac{y_i}{m}$. Portanto, podemos escrever a função $L(\mu/y_1, \ldots, y_m)$ como

$$\begin{aligned} L(\mu/y_1, \ldots, y_m) =& C \times \exp\left[-\frac{1}{2}\sum_{i=1}^{m}\frac{(\bar{y} - \mu)^2}{\sigma^2}\right] \\ =& C \times \exp\left[-\frac{m}{2}\frac{(\bar{y} - \mu)^2}{\sigma^2}\right] \propto \exp\left[-\frac{m}{2}\frac{(\bar{y} - \mu)^2}{\sigma^2}\right], \end{aligned} \tag{5.54}$$

onde C é uma constante independente do parâmetro $\theta = \mu$.

Consideremos agora uma função densidade de probabilidade *a priori* normal $N(\mu_0, \sigma_0^2)$ para o parâmetro desconhecido μ, onde μ_0 e σ_0^2 são conhecidos. Portanto, a função densidade *a priori* tem a seguinte fórmula

$$p(\mu) = \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left[\frac{1}{2\sigma_0^2}(\mu - \mu_0)^2\right]. \tag{5.55}$$

onde os parâmetros da distribuição *a priori* μ_0 e σ_0 são chamados de **hiperparâmetros**.

Seguindo a ideia da atualização Bayesiana, a função densidade de probabilidade *a posteriori* é dada então por

$$p(\mu/y_1, \ldots, y_m) \propto \frac{1}{2\pi\sigma_0^2} \exp\left[\frac{1}{2\sigma_0^2}(\mu - \mu_0)^2\right] \times \exp\left[-\frac{m}{2}\frac{(\bar{y} - \mu)^2}{\sigma^2}\right]. \tag{5.56}$$

A partir da Eq. (5.56), podemos mostrar que a função densidade de probabilidade *a posteriori* $p(\mu/y_1, \ldots, y_m)$ corresponde a uma distribuição normal, com média *a posteriori*

$$\mathrm{E}[\mu/y_1, \ldots, y_m] = \mu_0 \times \left[\frac{1/\sigma_0^2}{1/\sigma_0^2 + m/\sigma^2}\right] + \bar{y} \times \left[\frac{m/\sigma^2}{1/\sigma_0^2 + m/\sigma^2}\right], \tag{5.57}$$

e com variância *a posteriori*

$$\mathrm{Var}[\mu/y_1, \ldots, y_m] = \left[\frac{1}{\sigma_0^2} + \frac{m}{\sigma^2}\right]^{-1}. \tag{5.58}$$

Note que a média *a posteriori* é uma média ponderada entre a média da distribuição *a priori* μ_0 e a média amostral $\bar{y}$. Da mesma forma que no exemplo anterior, quando o tamanho da amostra m aumenta, o peso da média amostral também aumenta. Quando a variância populacional σ^2 diminui, o peso da média amostral $\bar{y}$ também aumenta. Quando a variância da distribuição *a priori* diminui, o peso da média *a priori* μ_0 aumenta.

Ainda de acordo com as expressões (5.57) e (5.58), notamos que quando $\sigma_0 \to \infty$, a média e a variância *a posteriori* convergem para $\mathrm{E}[\mu/y_1, \ldots, y_m] \to \bar{y} = \mu_{\mathrm{MV}}$ e $\mathrm{Var}[\mu/y_1, \ldots, y_m] \to \frac{\sigma^2}{m} = \sigma^2_{\mathrm{MV}}$, onde o subscrito MV indica que essas estimativas são aquelas dadas pelo estimadores de máxima verossimilhança.

Portanto, se considerarmos uma distribuição *a priori* com variância infinita (ou seja, sem informação relevante a respeito do parâmetro μ), então a média e a variância da distribuição *a posteriori* seriam as mesmas, caso considerássemos um estimador baseado apenas nos dados, como o estimador de máxima verossimilhança.

Esse caso no qual a variância da distribuição *a priori* é infinita ilustra situações para as quais o analista não possui informações minimamente precisas sobre o valor do parâmetro desconhecido, ou situações relacionadas a avaliação de políticas públicas, por exemplo, nas quais diferentes indivíduos possuem diferentes crenças sobre o parâmetro. Nessas situações, é importante que a informação *a priori* não afete os resultados obtidos a partir da amostra. Distribuições *a priori* que não afetam os resultados finais da análise deixando que apenas a amostra influencie na distribuição *a posteriori*, são chamadas distribuições *a priori* **não informativas**.

Consideremos novamente o exemplo no qual a amostra, de observações independentes e identicamente distribuídas $y_1, \ldots, y_m$, vem de uma população com distribuição normal com média desconhecida μ e variância conhecida σ^2. Vamos considerar agora uma função densidade *a priori* da forma $p(\mu) = k$, $\forall \mu \in \Re$, onde k é um valor contante conhecido. Observe que essa função densidade não integra para um, e por isso é chamada função densidade **imprópria**. Não necessariamente uma função densidade de probabilidade *a priori* imprópria incorre em uma função densidade de probabilidade *a posteriori* também imprópria, conforme veremos abaixo.

Para essa nova função densidade de probabilidade *a priori*, a função de densidade de probabilidade *a posteriori* é dada por

$$p(\mu/y_1,\ldots,y_m) = k \times C \times \exp\left[-\frac{1}{2}\sum_{i=1}^{m}\frac{(\bar{y}-\mu)^2}{\sigma^2}\right] = C' \times \exp\left[-\frac{1}{2}\frac{(\bar{y}-\mu)^2}{\sigma^2/m}\right], \tag{5.59}$$

onde C' é uma constante que independe do parâmetro desconhecido μ, e cujo valor é tal que o lado direito da Eq. (5.59) tem integral igual a um. Seguindo a notação na literatura de inferência Bayesiana, podemos escrever a Eq. (5.59) da forma

$$p(\mu/y_1,\ldots,y_m) \propto \exp\left[-\frac{1}{2}\frac{(\bar{y}-\mu)^2}{\sigma^2/m}\right], \tag{5.60}$$

cuja leitura é justamente de que o lado esquerdo da equação acima é igual ao lado direito multiplicado por uma constante que independe do vetor θ de parâmetros desconhecidos. Conclui-se então que a distribuição *a posteriori* de μ corresponde a uma distribuição normal com média $\bar{y}$ e variância σ^2/m. Observe, portanto, que toda a informação relevante para estimar o parâmetro μ vem da amostra $y_1,\ldots,y_m$. A função densidade de probabilidade $p(\mu) = k$ é uma função de densidade *a priori* não informativa. Apesar de a distribuição *a priori* ser imprópria (não integra para um), a distribuição *a posteriori* integra para um e, portanto, é chamada função densidade de probabilidade **própria**. Finalmente, note que a distribuição *a posteriori* com a *priori* $p(\mu) = k$ é igual à função *a posteriori* que resulta do limite $\sigma_0^2 \to \infty$ quando a função *a priori* tem distribuição normal com média μ_0 e variância σ_0^2.

5.7.5 População normal com média conhecida e variância desconhecida

Considere agora uma amostra de observações independentes e identicamente distribuídas $y_1,\ldots,y_m$, com distribuição normal com média μ conhecida e variância $\theta = \sigma^2$ desconhecida. O objetivo da inferência nesse caso é estimar o valor de σ^2. A função de verossimilhança nesse caso é dada por

$$L(\sigma^2/y_1,\ldots,y_m) = \frac{1}{\sqrt{2\pi}}(\sigma^2)^{-m/2}\exp\left[\frac{-ms^2}{2\sigma^2}\right] \propto (\sigma^2)^{-m/2}\exp\left[\frac{-ms^2}{2\sigma^2}\right], \tag{5.61}$$

com $s^2 = \sum_{i=1}^{m}(y_i-\mu)^2/m$. A função densidade de probabilidade conjugada correspondente é uma função densidade gamma-inversa,[13] com parâmetros α_0 e β_0, e equação

$$\begin{aligned} p(\sigma^2) &= \frac{\beta_0^{\alpha_0}}{\Gamma(\alpha_0)}(\sigma^2)^{-(\alpha_0+1)}\exp\left[-\beta_0/\sigma^2\right] \\ &\propto (\sigma^2)^{-(\alpha_0+1)}\exp\left[-\beta_0/\sigma^2\right]. \end{aligned} \tag{5.62}$$

[13]Uma variável aleatória é dita ter distribuição gamma-inversa se a sua recíproca tem distribuição gamma, que foi apresentada na Seção 3.3.

Combinando a função de verossimilhança e a função de densidade *a priori*, obtemos a função de densidade *a posteriori*

$$\begin{aligned} p(\sigma^2/y_1,\ldots,y_m) &\propto (\sigma^2)^{-m/2}\exp\left[\frac{-ms^2}{2\sigma^2}\right] \times (\sigma^2)^{-(\alpha_0+1)}\exp\left[-\beta_0/\sigma^2\right] \\ &\propto (\sigma^2)^{-(\alpha_0+1+m/2)} \times \exp\left[-\frac{ms^2+2\beta_0}{2}\times\frac{1}{\sigma^2}\right] \\ &\propto (\sigma^2)^{-(\alpha+1)}\exp\left[-\beta/\sigma^2\right]. \end{aligned} \tag{5.63}$$

Portanto, a função densidade de probabilidade *a posteriori* para o parâmetro σ^2 é uma distribuição gamma-inversa, com parâmetros atualizados para α e β, onde $\alpha = \alpha_0 + m/2$ e

$$\beta = \frac{ms^2+2\beta_0}{2}.$$

O fato de a função densidade de probabilidade *a posteriori* ser uma uma gamma-inversa confirma o fato de que essa distribuição é a conjugada para o modelo normal com média conhecida e variância desconhecida.

Uma forma conveniente de abordar o mesmo problema é considerar uma parametrização alternativa para a função densidade *a priori*. Consideremos a parametrização para a densidade *a priori*, fazendo $\alpha_0 = \nu_0/2$ e $\beta_0 = \sigma_0^2\nu_0/2$, onde ν_0 e σ_0^2 são os dois novos hiperparâmetros. A função densidade em (5.62) pode ser reescrita como

$$\begin{aligned} p(\sigma^2) &= \frac{(\nu_0/2)^{(\nu_0/2)}}{\Gamma(\nu_0/2)}(\sigma_0^2)^{(\nu_0/2)}(\sigma^2)^{-(\nu_0/2+1)}\exp\left[-\frac{\nu_0\sigma_0^2}{2\sigma^2}\right] \\ &\propto \left[\frac{\sigma_0^2}{\sigma^2}\right]^{(\nu_0/2+1)} \times \exp\left[-\frac{\nu_0\sigma_0^2}{2\sigma^2}\right]. \end{aligned} \tag{5.64}$$

A Eq. (5.64) corresponde à função densidade de probabilidade de uma variável aleatória qui-quadrada-inversa[14] com parâmetro de escala (*scaled inverse-*χ^2). Pode-se mostrar que uma variável aleatória W possui distribuição qui-quadrada-inversa com parâmetro escala σ_0^2, se e somente se $W = \sigma_0^2\nu_0/V$, onde V é uma variável aleatória com distribuição qui-quadrada com ν_0 graus de liberdade. Para a função densidade de probabilidade *a priori* considerada, escreve-se $\sigma^2 \sim$ Inv-χ^2 (ν_0, σ_0^2). Para essa parametrização da função de densidade *a priori*, a função de densidade *a posteriori* passa a ter expressão

[14] Uma variável aleatória é dita ter distribuição qui-quadrada-inversa se a sua recíproca tem distribuição qui-quadrada, que foi apresentada na Seção 3.3.

equivalente

$$
\begin{aligned}
p(\sigma^2/y_1,\ldots,y_m) &\propto (\sigma^2)^{-m/2}\exp\left[\frac{-ms^2}{2\sigma^2}\right]\times\left[\frac{\sigma_0^2}{\sigma^2}\right]^{(\nu_0/2+1)}\times\exp\left[-\frac{\nu_0\sigma_0^2}{2\sigma^2}\right] \\
&\propto (\sigma^2)^{-((m+\nu_0)/2+1))}\times\exp\left[-\frac{\nu_0\sigma_0^2+ms^2}{2}\times\frac{1}{\sigma^2}\right],
\end{aligned}
\tag{5.65}
$$

e podemos concluir que a distribuição *a posteriori* corresponde a uma distribuição qui-quadrada-inversa

$$
\sigma^2/y_1,\ldots,y_m \sim \text{Inv-}\chi^2\left(\nu_0+m,\frac{\nu_0\sigma_0^2+ms^2}{\nu_0+m}\right).
$$

O parâmetro de escala $\frac{\nu_0\sigma_0^2+ms^2}{\nu_0+m}$ da distribuição *a posteriori* é uma média ponderada entre o parâmetro de escala σ_0^2 da distribuição *a priori* e a variância amostral s^2. O número de graus de liberdade da distribuição *a posteriori* é igual à soma do número de graus de liberdade da distribuição *a priori* e do número de observações m na amostra. Note que, dependendo da parametrização que utilizamos para a função densidade *a priori*, a interpretação dos resultados pode ser bem mais direta.

A informação na função densidade de probabilidade *a priori* pode ser vista então como a informação em um amostra prévia de ν_0 observações, com variância amostral igual a σ_0^2. Essa forma de interpretar a distribuição *a priori* nos sugere que, caso o parâmetro ν_0 seja muito pequeno quando comparado ao tamanho da amostra m, a informação *a priori* torna-se irrelevante. No limite, quando $\nu_0 \to 0$, temos que a distribuição *a posteriori* corresponde a uma qui-quadrada-inversa com parâmetros m e s^2.

Consideremos agora uma função densidade *a priori* da forma $p(\sigma^2) \propto 1/\sigma^2$, $\forall\sigma \in (0,+\infty)$. Essa função densidade *a priori* é imprópria, uma vez que ela não integra para um. A função densidade de probabilidade *a posteriori* é dada por

$$
\begin{aligned}
p(\sigma^2/y_1,\ldots,y_m) &\propto (\sigma^2)^{-m/2}\exp\left[\frac{-ms^2}{2\sigma^2}\right]\times\frac{1}{\sigma^2} \\
&\propto (\sigma^2)^{-(m/2+1)}\exp\left[\frac{-ms^2}{2\sigma^2}\right].
\end{aligned}
\tag{5.66}
$$

A partir da expressão para a função densidade de probabilidade de uma distribuição qui-quadrada-inversa com parâmetro de escala, é possível concluir que, de acordo com a Eq. (5.66), para a distribuição *a priori* $p(\sigma^2) \propto 1/\sigma^2$, o parâmetro desconhecido σ^2 tem distribuição *a posteriori* $\sigma^2/y_1,\ldots,y_m \sim \text{Inv-}\chi^2\ (m,s^2)$, que corresponde justamente ao caso limite acima no qual σ^2 tem distribuição *a priori* $\text{Inv-}\chi^2(\nu_0,\sigma_0^2)$ e $\nu_0 \to 0$. Portanto, a distribuição *a priori* $\sigma \propto 1/\sigma^2$ corresponde a uma distribuição *a priori* não informativa para o parâmetro σ^2. Novamente, temos um exemplo no qual a distribuição *a priori* é imprópria, mas a distribuição *a posteriori* é própria.

5.7.6 Modelos com vários parâmetros desconhecidos

Nos exemplos anteriores, consideraram-se modelos estocásticos, para os quais havia apenas um parâmetro desconhecido. No primeiro caso, o parâmetro desconhecido era a probabilidade de sucesso para uma variável aleatória binomial com n tentativas (n conhecido). Nos exemplos seguintes, consideremos amostras de populações com distribuições normais, para as quais a média ou a variância populacionais eram desconhecidas. Discutiremos agora uma situação na qual temos dois parâmetros desconhecidos, de forma que precisaremos de uma distribuição *a priori* bivariada.

Consideremos novamente a situação na qual temos uma amostra de observações $y_1, \ldots, y_m$ independentes e identicamente distribuídas, com distribuição normal, com média μ e variância σ^2, desta vez ambos parâmetros desconhecidos. Precisamos então especificar a distribuição *a priori*. A forma mais simples de se proceder nesse caso é supor que os parâmetros desconhecidos são independentes *a priori*, de forma que a distribuição *a priori* multivariada é simplesmente o produto das distribuições marginais de cada parâmetro individualmente. Podemos considerar, como distribuição *a priori* para o parâmetro μ, uma distribuição normal, com média μ_0 e variância τ_0^2. Podemos considerar uma distribuição *a priori* pertencente à família qui-quadrada-inversa com parâmetros ν_0 e σ_0^2. Portanto, a distribuição *a priori* para o vetor de parâmetros $\theta = (\mu, \sigma^2)$ é igual a

$$\begin{aligned} p(\mu, \sigma^2) = & \frac{1}{\sqrt{2\pi\tau_0^2}} \exp\left[\frac{1}{2\tau_0^2}(\mu - \mu_0)^2\right] \\ & \times \frac{(\nu_0/2)^{(\nu_0/2)}}{\Gamma(\nu_0/2)} (\sigma_0^2)^{(\nu_0/2)} (\sigma^2)^{-(\nu_0/2+1)} \exp\left[-\frac{\nu_0\sigma_0^2}{2\sigma^2}\right]. \end{aligned} \tag{5.67}$$

Supondo que a função de verossimilhança tem distribuição normal com parâmetros μ and σ^2, a função de densidade de probabilidade *a posteriori* é dada por

$$\begin{aligned} p(\mu, \sigma^2 / y_1, \ldots, y_m) \propto & \frac{1}{\sqrt{\tau_0^2}} \exp\left[\frac{1}{2\tau_0^2}(\mu - \mu_0)^2\right] \\ & \times \frac{(\nu_0/2)^{(\nu_0/2)}}{\Gamma(\nu_0/2)} (\sigma_0^2)^{(\nu_0/2)} (\sigma^2)^{-(\nu_0/2+1)} \exp\left[-\frac{\nu_0\sigma_0^2}{2\sigma^2}\right] \\ & \times \frac{1}{\sqrt{\sigma^2}} \exp\left[\frac{1}{2\sigma^2}\sum_{i=1}^{m}(y_i - \mu)^2\right]. \end{aligned} \tag{5.68}$$

Para a distribuição *a priori* multivariada acima, podemos mostrar que a distribuição *a posteriori* para o parâmetro μ, condicionada ao parâmetro σ^2 e à amostra $y_1, \ldots, y_m$, corresponde a uma distribuição normal, com média *a posteriori*

$$\mu_m = \frac{\frac{1}{\tau_0^2}\mu_0 + \frac{m}{\sigma^2}\bar{y}}{\frac{1}{\tau_0^2} + \frac{m}{\sigma^2}},$$

e variância *a posteriori*

$$\sigma_m^2 = \frac{1}{\frac{1}{\tau_0^2} + \frac{m}{\sigma^2}}.$$

Observe que tanto μ_m quanto σ_m^2 dependem do outro parâmetro desconhecido σ^2. Escrevemos então $\mu/\sigma^2, y_1, \ldots, y_m \sim \mathrm{N}(\mu_m, \sigma_m^2)$. Adicionalmente, para encontrar as funções densidades marginais $p(\mu/y_1, \ldots, y_m)$ e $p(\sigma^2/y_1, \ldots, y_m)$, é necessário integrar a função de densidade de probabilidade *a posteriori* multivariada apresentada na Eq. (5.68) para todos os valores possíveis do outro parâmetro. Portanto, a função densidade de probabilidade marginal $p(\mu/y_1, \ldots, y_m)$ é dada por

$$p(\mu/y_1, \ldots, y_m) = \int_{\sigma^2=0}^{\infty} p(\mu, \sigma^2/y_1, \ldots, y_m) d\sigma^2, \ \forall \mu \in \Re,$$

enquanto a função densidade de probabilidade marginal *a posteriori* para σ^2 é dada por

$$p(\sigma^2/y_1, \ldots, y_m) = \int_{\mu=-\infty}^{\infty} p(\mu, \sigma^2/y_1, \ldots, y_m) d\mu, \ \forall \sigma^2 > 0.$$

Na prática, a obtenção das funções de densidade marginais pode ser feita de forma analítica ou de forma numérica. Em problemas mais complexos, e em geral mais interessantes em termos de aplicações, nem sempre é de interesse do pesquisador encontrar explicitamente formas fechadas para as distribuições marginais *a posteriori*. Ao invés disso, é possível obter todas as medidas de interesse por meio de simulações de Monte Carlo. Para maiores detalhes sobre métodos computacionais aplicados a problemas de inferência Bayesiana, o leitor pode recorrer a Tanner (1996).

5.8 Exercícios

Exercício 5.1 Seja $X_1, X_2, \ldots, X_n$ uma sequência de variáveis aleatórias independentes e identicamente distribuídas, cada qual com distribuição binomial com parâmetros n e p, onde conhecemos n, mas não conhecemos p.[15] Responda:

(1) Encontre pelo menos dois estimadores diferentes para p, baseados no método de momentos;

(2) Escreva a função densidade conjunta de $X_1, X_2, \ldots, X_n$;

(3) Qual o estimador de máxima verossimilhança de p?

(4) O estimador de máxima verossimilhança obtido em (3) é viesado?

(5) Qual a variância e o desvio padrão do estimador obtido no item (3)?

Exercício 5.2 Seja $X_1, X_2, \ldots, X_n$ uma sequência de variáveis aleatórias independentes e identicamente distribuídas, cada qual com distribuição geométrica, com parâmetro p. Responda:

[15] Na prática, isso é o que normalmente acontece. O caso onde n é desconhecido e tem que ser estimado da amostra é bem mais complicado e não será coberto neste livro.

(1) Encontre pelo menos dois estimadores diferentes para p, baseados no método de momentos;
(2) Escreva a função densidade conjunta de $X_1, X_2, \ldots, X_n$;
(3) Qual o estimador de máxima verossimilhança de p?
(4) Qual a variância e o desvio padrão do estimador obtido no item (3)?

Exercício 5.3 Seja $X_1, X_2, \ldots, X_n$ uma sequência de variáveis aleatórias independentes e identicamente distribuídas, cada qual com distribuição de Rayleigh, com parâmetro β. Responda:
(1) Encontre pelo menos dois estimadores diferentes para β, baseados no método de momentos;
(2) Escreva a função densidade conjunta de $X_1, X_2, \ldots, X_n$;
(3) Qual o estimador de máxima verossimilhança de β?
(4) Qual a variância e o desvio padrão do estimador obtido no item (3)?

Exercício 5.4 Seja $X_1, X_2, \ldots, X_n$ uma sequência de variáveis aleatórias independentes e identicamente distribuídas, cada qual com distribuição qui, com parâmetro ν. Responda:
(1) Encontre pelo menos dois estimadores diferentes para ν, baseados no método de momentos;
(2) Escreva a função densidade conjunta de $X_1, X_2, \ldots, X_n$;
(3) Qual o estimador de máxima verossimilhança de ν?
(4) Qual a variância e o desvio padrão do estimador obtido no item (3)?

Exercício 5.5 Seja $X_1, X_2, \ldots, X_n$ uma sequência de variáveis aleatórias independentes e identicamente distribuídas, cada qual com distribuição uniforme $\mathrm{U}(0, \theta)$, onde θ é um parâmetro desconhecido. Então, a função densidade de X_i, $i = 1, 2, \ldots, n$, é dada por

$$f(x) = \frac{1}{\theta} \times \mathrm{I}_{\{x \in (0,\theta)\}},$$

onde $\mathrm{I}_{\{x \in (0,\theta)\}}$ é uma função indicadora, que vale 1 caso a condição entre chaves seja satisfeita, e 0, caso contrário. Ou seja, $\mathrm{I}_{\{x \in (0,\theta)\}}$ vale 1 quando $x \in (0, \theta)$ e 0, caso contrário. Dada a amostra $X_1, \ldots, X_n$, responda:
(1) Escreva a função de verossimilhança $L(\theta)$;
(2) Desenhe o gráfico da função de verossimilhança;
(3) Determine o estimador de máxima verossimilhança $\hat{\theta}_{MV}$ de θ;
(4) Determine o valor esperado de $\hat{\theta}_{MV}$;
(5) Determine o viés do estimador $\hat{\theta}_{MV}$.

Exercício 5.6 Seja $X_1, X_2, \ldots, X_n$ uma sequência de variáveis aleatórias independentes e identicamente distribuídas, cada qual com distribuição gamma, com parâmetros α e β. Responda:
(1) Encontre um estimador de método de momentos para α e para β, a partir da amostra.
(2) Escreva a função de densidade conjunta para $X_1, X_2, \ldots, X_n$.
(3) Escreva a função de log-verossimilhança como função dos parâmetros desconhecidos α e β.
(4) É possível escrever os estimadores de máxima verossimilhança para α e β como função explícita da amostra?

(5) Caso não seja possível escrever explicitamente os estimadores de máxima verossimilhança para α e β, qual seria uma estratégia computacional para achar os valores desses estimadores?

Exercício 5.7 Mostre que, a partir da Eq. (5.39), podemos obter a função densidade de probabilidade *a posteriori* na Eq. (5.41).

Exercício 5.8 Mostre que, a partir da Eq. (5.65), podemos obter a distribuição qui-quadrada-inversa

$$\sigma^2/y_1,\ldots,y_m \sim \text{Inv-}\chi^2\left(\nu_0+m, \frac{\nu_0\sigma_0^2+ms^2}{\nu_0+m}\right).$$

Exercício 5.9 Uma variável aleatória Y com espaço amostral $\mathbb{X} = (0,\infty)$ é dita ter distribuição normal-inversa se possui função de densidade

$$f(y) = \left(\frac{\lambda}{2\pi y^3}\right)^{1/2} \exp\left(\frac{-\lambda(y-\mu)^2}{2\mu^2 y}\right),$$

onde μ e λ são dois parâmetros estritamente positivos que representam respectivamente a média da variável aleatória e a forma da distribuição.

(1) Mostre que se Y possui uma distribuição normal-inversa, então kY também possui, onde k é uma constante positiva.

(2) Mostre que se Y_i, $i = 1,\cdots,n$, são independentes e possuem distribuição normal-inversa, então $\sum_i^n Y_i$ também possui.

6. Intervalos de confiança e testes de hipóteses

"To succeed in life, you need two things: ignorance and confidence."
Mark Twain

No Capítulo 5 descrevemos o processo de estimação de parâmetros para distribuições paramétricas, utilizando-se notadamente do método de momentos e do método de máxima verossimilhança. Para esses métodos, iniciamos a discussão sobre distribuição dos estimadores, já que eles também são variáveis aleatórias, uma vez que são funções de variáveis aleatórias. Um estimador para o parâmetro de uma variável aleatória de Poisson corresponde simplesmente à média amostral dos valores observados na amostra. Um estimador do parâmetro de uma variável aleatória exponencial negativa é igual ao inverso da média amostral. Para cada um desses estimadores, é importante levantar a distribuição dos estimadores, para termos uma ideia, por exemplo, do grau de imprecisão na estimação dos parâmetros. Vimos, por exemplo, que quanto maior o tamanho da amostra, mais precisos são os estimadores.

Neste capítulo iremos explorar em mais detalhes conceitos que são relevantes para o entendimento da distribuição dos estimadores. Recorreremos, sempre que possível, a simulações de Monte Carlo para tornar a discussão mais intuitiva. A partir do conceito de distribuições dos estimadores, discutiremos dois tópicos extremamente importantes para inferência estatística: **intervalos de confiança** e **testes de hipótese**. O primeiro está mais diretamente ligado a medidas de imprecisão das estimativas. O segundo tópico está ligado à procura de evidência, nos dados da amostra, para suportar ou contrariar uma suposição ou hipótese teórica. Ressaltamos que em momento algum os testes de hipótese comprovam ou descartam uma teoria com 100% de certeza. Os testes de hipótese na verdade servem como mais uma evidência para ajudar o pesquisador, analista ou planejador na tomada de decisões. Por exemplo, imagine que o governo tenha implantado uma determinada política pública, sobre a qual gostaríamos de aferir sua efetividade. A pergunta de interesse é "será que o dinheiro investido nessa política pública conseguiu atingir aos objetivos esperados?". Responder a perguntas dessa natureza raramente é possível com base em um único estudo. Idealmente, devem ser feitos diversos estudos, de preferência por instituições sérias independentes. Após diversos debates com base nesses estudos, pode-se chegar a conclusões mais robustas a respeito dos resultados da política pública de interesse nos estudos.

6.1 Introdução ao processo de inferência estatística

Nesta seção, faremos nossa primeira incursão a respeito de um tópico extremamente importante para o entendimento do que chamamos estimadores estatísticos e ao que chamamos de distribuições amostrais. Escolhemos trabalhar com simulações de Monte Carlo, com o objetivo de descrever os princípios básicos relevantes para o entendimento de medidas populacionais e medidas amostrais, bem como os conceitos básicos por trás de amostras e populações. Inicialmente, iremos trabalhar com o que chamamos conceitualmente de **populações finitas**. Mais adiante, apresentaremos uma discussão sobre o que chamamos de **populações infinitas**, as quais também possuem uma enorme importância em termos práticos. No momento, vamos nos ater às populações finitas. que são conjuntos finitos de unidades observacionais, para os quais estamos interessados em algumas das suas características. Um exemplo de uma população finita é o conjunto de todos os domicílios na região metropolitana de São Paulo – as unidades observacionais nesse caso são os domicílios, que existem em um número finito e bem definido.[1] Vamos supor que as variáveis que queremos estudar são a renda *per capita* de cada um desses domicílios, o nível de escolaridade do chefe de cada um desses domicílios, o número de filhos em cada um desses domicílios, e a idade do chefe de família. Portanto, temos quatro variáveis de interesse.

6.1.1 Amostragem aleatória simples

De acordo com a Pesquisa Nacional de Amostragem Domiciliar (PNAD) de 2006, do Instituto Brasileiro de Geografia e Estatística, o número de domicílios na Região Metropolitana de São Paulo é da ordem de $N = 4.5$ milhões. Note que estamos utilizando letra maiúscula N para o número de unidades observacionais na população. Vamos supor que um pesquisador pretenda pôr seus entrevistadores na rua para visitar esses domicílios e coletar informações a respeito de cada um deles. Esses entrevistadores irão coletar informações sobre quatro variáveis de interesse:

(1) renda *per capita*,
(2) nível de escolaridade do chefe do domicílio,
(3) número de filhos,
(4) idade do chefe do domicílio.

Como estamos tratando de análise estatística de dados, iremos supor, por simplicidade e para fins didáticos, que a escolha dos domicílios entrevistados será feita de forma aleatória, via **amostragem aleatória simples**, com reposição. Não nos ateremos aos detalhes por trás dos métodos de amostragem, já que esse não é o interesse nesta publicação. O leitor interessado pode recorrer a referências como Cochran (1977), Bussab e Morettin (2002) e Lohr (2002) e Bolfarine e Bussab (2005).

[1] Número esse que não necessariamente é conhecido por nós com total certeza – mas com certeza é conhecido por Deus.

No processo de amostragem aleatória simples, em um primeiro momento são listados todos os domicílios (ou **unidades observacionais**) na população. Em seguida, sorteamos um número n de domicílios para fazerem parte da amostra de entrevistados. O processo de amostragem aleatória simples pode ser intuitivamente entendido como um processo onde os identificadores de todos os N domicílios da população são postos em N pequenos papéis depositados em uma jarra. Retiramos então n desses papéis aleatoriamente. No processo de amostragem aleatória simples **sem reposição**, uma vez que um determinado papel (domicílio) é sorteado, ele não é posto de volta à jarra, enquanto no processo de amostragem aleatória simples **com reposição**, uma vez sorteado e anotado o identificador do domicílio para compor a amostra, ele é reinserido na jarra, antes da retirada do novo papel sorteado. Obviamente, do ponto de vista prático, a amostragem sem reposição faz muito mais sentido, uma vez que não poderemos sortear um mesmo domicílio duas ou mais vezes, o que pode acontecer na amostragem com reposição. No entanto, o tratamento matemático do processo de amostragem aleatória simples com reposição é mais simples, no sentido de que as n retiradas são independentes. Ou seja, os domicílios sorteados nas primeiras retiradas não afetam os domicílios retirados nas retiradas posteriores. Para a amostragem sem reposição, caso retiremos os domicílios de maior renda no início do sorteio, a probabilidade de retirarmos domicílios com renda menor nas retiradas posteriores será maior. Ou seja, na amostragem sem reposição, as retiradas e os valores coletados para as variáveis de interesse, não serão independentes.

Mas por que a independência nas retiradas (ou valores observados para as variáveis) é tão importante? Em primeiro lugar, conforme já comentamos acima, independência traz uma série de simplificações no tratamento matemático dos estimadores. Obviamente dado que a amostragem aleatória sem reposição faz muito mais sentido em termos práticos, já existe uma vasta literatura e desenvolvimento científico para o tratamento analítico de amostragem sem reposição. E então vem o segundo e principal motivo de estarmos focando no processo de amostragem com reposição, e a consequente independência das observações: no processo de modelagem matemática de processos estocásticos, cuja discussão iniciamos no Capítulo 3, independência é crucial para o entendimento de toda a teoria de estimação apresentada no Capítulo 5. Como o nosso foco é justamente prover o leitor com conceitos e ferramentas básicas para a modelagem de processos estocásticos, optamos por focar a discussão em observações independentes. Além disso, ao nosso ver, uma das dificuldades no entendimento dos conceitos básicos de estatística é a passagem da análise de populações finitas para populações infinitas. Nesse sentido, na nossa opinião, se a hipótese de independência das observações for considerada desde o início da discussão, essa passagem ficará mais didática.

6.1.2 Medidas populacionais e medidas amostrais

A Figura 6.1 a seguir apresenta o histograma para a distribuição das quatro variáveis de interesse, com base na massa de dados para toda a população. Obviamente, essa é uma massa de dados populacionais hipotética, com o objetivo meramente ilustrativo. Além disso, a menos que seja feito um censo[2] de toda a população

[2] O que está sujeito a erros conhecidos como não-amostrais, que advém, entre outras coisas, do processo de coleta, que nunca é perfeito.

da região metropolitana de São Paulo, somente Deus conhece as características populacionais com certeza. A Tabela 6.1 apresenta algumas das características populacionais, para essa população hipotética.

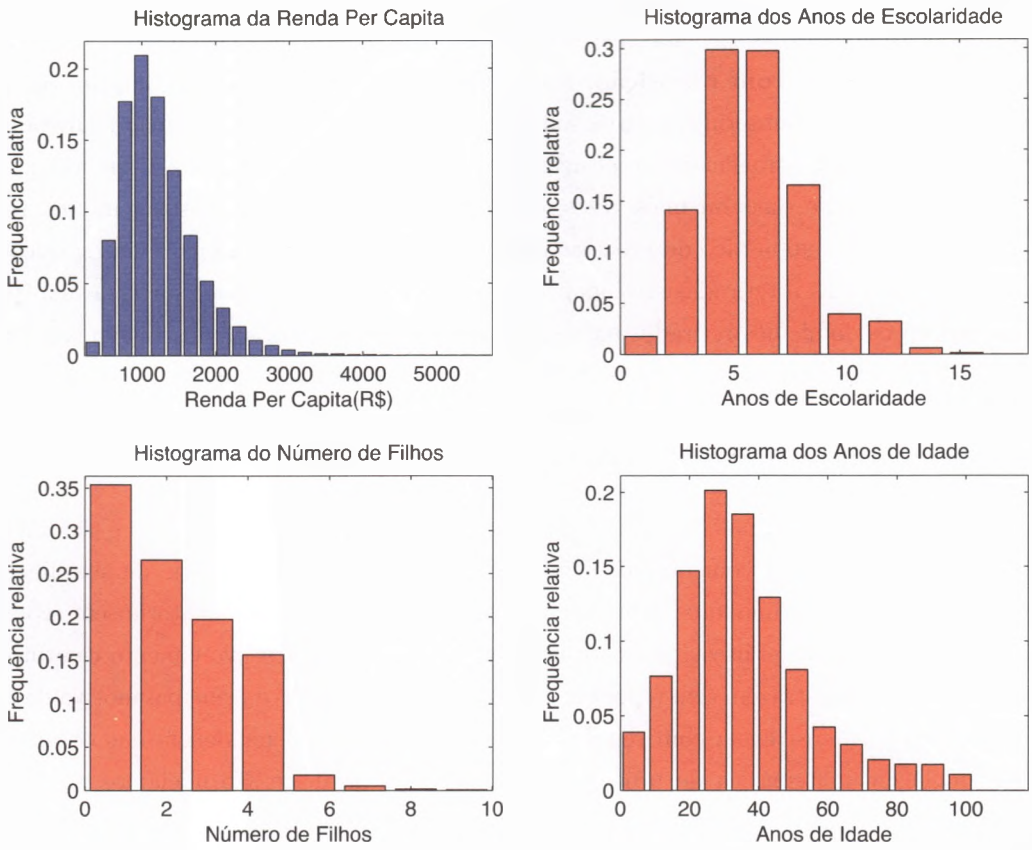


Figura 6.1: Histogramas para quatro variáveis populacionais.

A Tabela 6.1 apresenta também as medidas amostrais para uma amostra aleatória simples, com reposição (e portanto, com observações independentes), com $n = 20$ unidades amostradas. Observe que os valores coletados para a amostra não correspondem exatamente aos valores populacionais. É bom ter em mente que os valores populacionais somente são conhecidos aqui neste exemplo porque estamos brincando de ser Deus. Na prática, nós, meros mortais, conhecemos apenas as colunas 6 a 9 da Tabela 6.1. Portanto, toda a informação que temos a respeito da população está nessas colunas. Aqui entra o objetivo principal de todo o processo de inferência estatística: entender o quão perto nossas estimativas amostrais estão das medidas populacionais que queremos conhecer, mesmo estando sujeitos à imprecisão incorrida por termos disponível apenas uma amostra de observações. Nos próximos parágrafos, ilustraremos por meio de simulações de Monte Carlo, o que chamamos de distribuição amostral dos estimadores estatísticos.

Tabela 6.1: Parâmetros populacionais e amostrais para uma amostra aleatória, com reposição, composta por $n = 20$ domicílios.

	População				Amostra			
Indicadores estatísticos	Renda (R$ mil)	Anos de escola	Número de filhos	Anos de idade	Renda (R$ mil)	Anos de escola	Número de filhos	Anos de idade
N. observações	45 mil	45 mil	45 mil	45 mil	20	20	20	20
Média	1.224	5.922	2.209	36.01	1.378	5.900	2.350	38.80
Variância	0.261	5.919	2.208	377.51	0.322	5.190	2.828	273.36
Desvio padrão	0.511	2.433	1.486	19.43	0.567	2.278	1.682	16.53
Curtose	6.205	3.129	3.472	3.93	2.648	3.945	3.748	3.36
Assimetria	1.322	0.410	0.680	0.93	0.750	-0.257	0.765	0.90
Mediana	1.129	6.000	2.000	33.00	1.249	6.000	2.000	34.00
1o. quartil	0.862	4.000	1.000	23.00	0.958	5.000	1.000	28.50
3o. quartil	1.477	7.000	3.000	45.00	1.677	7.000	3.500	48.50
Δ Interquartil	0.615	3.000	2.000	22.00	0.719	2.000	2.500	20.00

6.2 Simulações de Monte Carlo

Na Seção 5.4, consideramos uma amostra aleatória, com reposição, de uma população hipotética. Para essa população hipotética, com base na amostra coletada, calculamos uma série de indicadores estatísticos amostrais, e comparamos os valores amostrais aos valores populacionais. Obviamente, conforme comentamos anteriormente, as medidas populacionais são conhecidas com certeza apenas por Deus. O que enxergamos de fato são os indicadores amostrais.

De acordo com os números apresentados na amostra da Tabela 6.1, comparando-os aos valores populacionais, observamos que os valores estimados estão bem próximos aos valores populacionais, mesmo sendo as estatísticas amostrais calculadas utilizando-se uma amostra aleatória. Mesmo esses valores sendo próximos aos populacionais, eles não são exatamente os mesmos, havendo uma pequena diferença devido ao acaso. A Tabela 6.2 apresenta os valores para as medidas amostrais, especificamente para a variável renda (em R$ milhares), considerando-se quatro outras amostras, independentes entre si, e independentes da amostra da Tabela 6.1, coletadas para a mesma população hipotética da Figura 6.1. Todas as amostras possuem o mesmo tamanho $n = 20$. Observe que, os valores amostrais, apesar de mais ou menos próximos aos valores populacionais, apresentam claramente uma certa variação em torno dos valores populacionais. Em nenhuma das amostras, os valores amostrais são exatamente iguais aos valores populacionais.

A partir desse simples experimento, diversas perguntas podem ser feitas. Essas perguntas são fundamentais para o processo de inferência estatística. Algumas dessas perguntas já foram abordadas, de alguma forma, no Capítulo 5, quando discutimos a estimação de parâmetros via máxima verossimilhança e via método de momentos.

1. Apesar de haver uma certa dispersão dos valores amostrais em torno dos valores populacionais, em média a estimativa amostral é igual ao valor populacional?

Tabela 6.2: Parâmetros populacionais e amostrais da variável renda para outras quatro amostras aleatórias, com reposição, composta por $n = 20$ domicílios.

Indicadores estatísticos para a renda (R$ milhares)	População	Amostras 2	3	4	5
N. observações	45 mil	20	20	20	20
Média	1.224	1.214	1.325	1.191	1.069
Variância	0.261	0.413	0.292	0.149	0.142
Desvio padrão	0.511	0.643	0.541	0.386	0.377
Curtose	6.205	8.899	2.966	2.534	5.049
Assimetria	1.322	2.318	0.944	0.546	1.312
Mediana	1.129	1.069	1.204	1.158	1.035
1o. quartil	0.862	0.844	0.928	0.880	0.834
3o. quartil	1.477	1.335	1.521	1.377	1.195
Δ Interquartil	0.615	0.491	0.593	0.498	0.362

2. Os indicadores amostrais são calculados a partir de amostras aleatórias, e, portanto, esses também são variáveis aleatórias. Nesse caso, qual a distribuição dos indicadores (ou estimadores) amostrais?

3. O que acontece com a distribuição do estimadores amostrais, quando o número de observações na amostra aumenta?

4. Existe alguma forma de medir a precisão das estimativas, também a partir da amostra disponível?

5. Como podemos utilizar as informações na amostra para testar hipóteses de pesquisa?

6. O que acontece com a distribuição dos estimadores, caso as hipóteses que estamos fazendo a respeito do processo estocástico gerador das amostras não forem totalmente corretas?

7. Existem estimadores melhores do que os que estamos utilizando?

8. Como definimos critérios para escolha de estimadores?

Neste Capítulo, mostraremos como as simulações de Monte Carlo podem ser utilizadas para responder parte dessas perguntas. Ao longo deste texto, discutiremos os demais questionamentos. Uma das grandes benesses da análise estatística de dados é que, apesar de as simulações de Monte Carlo– que constituem um processo computacionalmente intensivo, fornecerem ilustrações para os problemas de inferência estatística, mostraremos que é possível antecipar matematicamente quais seriam os resultados das simulações. De fato, no Capítulo 5 discutimos alguns resultados a respeito da distribuição dos estimadores de máxima verossimilhança. Naquele Capítulo, discutimos a questão da consistência dos estimadores de máxima verossimilhança, de forma que, mesmo quando há viés nos estimadores, esse viés vai para zero quando o tamanho da amostra tende para infinito. Sempre que possível, estaremos fazendo o paralelo entre teoria de inferência estatística, ilustrações via simulações de Monte Carlo e exemplos práticos.

6.3 Distribuições dos estimadores e imprecisão das estimações

As colunas nas Tabelas 6.1 e 6.2 nos fornecem uma primeira ideia de como se comportam, estocasticamente falando, os estimadores das diversas medidas populacionais. De fato, em primeiro lugar observamos que, para as cinco amostras selecionadas (uma na Tabela 6.1 e quatro na Tabela 6.2), os valores estimados estão próximos dos valores populacionais. Além disso, os valores estimados encontram-se dispersos em torno das medidas populacionais. Imagine agora que se, ao invés de cinco amostras, tivéssemos milhares ou milhões. Para cada uma dessas 100 mil ou 1 milhão de amostras. teríamos uma estimativa, por exemplo, para a curtose da renda populacional. Com base nos cem mil ou um milhão de estimativas aleatórias geradas para a curtose, podemos ter uma ideia muito clara do comportamento do estimador de curtose. Comportamento, nesse caso, pode ser entendido como a distribuição da variável aleatória correspondente ao estimador da curtose populacional.

A ideia de estudar as características dos estimadores a partir de um número muito grande de diferentes e independentes amostras aleatórias é justamente a base das simulações de Monte Carlo. Para cada problema de interesse, simulamos uma quantidade grande de amostras, tendo o cuidado que, em todas as amostras geradas em um mesmo experimento, as condições de seleção das amostras sejam exatamente as mesmas. Por exemplo, precisamos utilizar o mesmo número de observações amostradas em cada amostra. Os passos nas simulações de Monte Carlo são descritos a seguir:

(i) Defina a população de interesse, bem como os parâmetros populacionais que se deseja inferir.

(ii) Defina um processo gerador das amostras, com as quais serão calculadas as estimativas para os parâmetros populacionais de interesse.

(iii) Defina as expressões e/ou os procedimentos para cálculo das estimativas dos parâmetros com base nas amostras geradas. No caso de estimações dos parâmetros de uma variável aleatória gamma, por exemplo, podemos utilizar o procedimento de maximização da função de log-verossimilhança ou o procedimento de método de momentos.

(iv) Gere uma amostra aleatória, com n observações, utilizando a definição do item (ii).

(v) Calcule a estimativa para o parâmetro (ou os parâmetros) de interesse com base nas definições do item (iii).

(vi) Guarde os resultados das estimativas do item (v).

(vii) Replique os passos (iv) a (vi) um número grande $M - 1$ de vezes (por exemplo, $M = 10,000$ ou $M = 1,000,000$), totalizando M valores estimados.

(viii) Com base nos M valores armazenados, podemos então estudar as características da distribuição dos estimadores dos parâmetros de interesse.

correspondente à estimação da média populacional. De acordo com a Tabela 6.1, as médias de renda (em R$ milhares), anos de escolaridade, número de filhos e idade para a população são 1.224, 5.922, 2.209 e 36.01 respectivamente. Inicialmente, vamos considerar uma amostra aleatória de tamanho $n = 5$; ou seja, uma amostra com poucas observações. As amostras foram geradas por meio de amostragem aleatória simples. Foram geradas $M = 100,000$ amostras. Para cada amostra $A_m = \{x_{m,1}, \ldots, x_{m,n}\}$ gerada, $m = 1, \ldots, M$, a estimativa para a média é dada por $\bar{x}_m = \frac{\sum_{i=1}^{n} x_{m,i}}{n}$. Ao final do processo de simulação, temos $M = 100,000$ valores para $\bar{x}_m$. A Figura 6.2 abaixo apresenta os histogramas desses $M = 100,000$ valores gerados para cada uma das quatro variáveis populacionais de interesse.

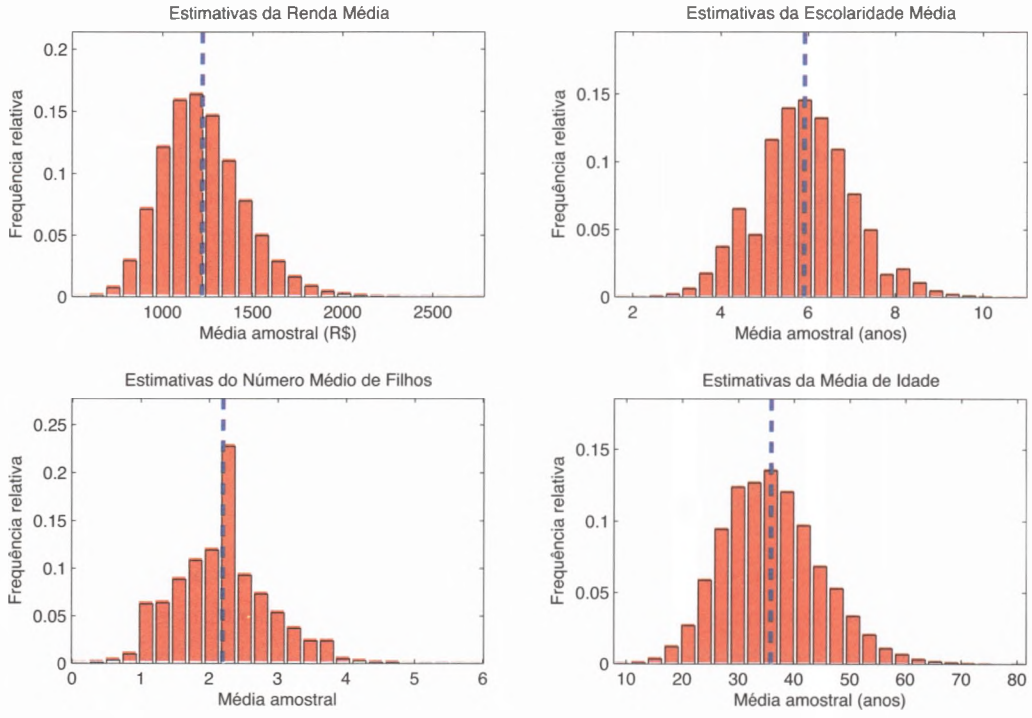


Figura 6.2: Histogramas para as médias amostrais com base nas simulações de Monte Carlo, para as quatro variáveis populacionais (amostras de tamanho $n = 5$).

Observe que, para as quatro variáveis, conforme era de se esperar, de acordo com os números apresentados nas Tabelas 6.1 e 6.2, os valores estimados estão dispersos em torno dos valores populacionais (demarcados pelas linhas verticais tracejadas na Figura 6.2). Para a variável renda, o histograma da distribuição das estimativas apresenta uma assimetria para a direita. De fato, quando calculamos o coeficiente de assimetria para as M estimativas da média de renda, obtemos um valor igual a 0.602. A Tabela 6.3 a seguir apresenta as estatísticas descritivas para os M valores obtidos, para cada uma das variáveis, com base em uma amostra de $n = 5$ observações. As médias amostrais (estimadores para as médias populacionais) estão bem próximas aos valores populacionais. Isso indica que o estimador "acerta" em média

o valor que desejamos estimar. Quando isso acontece. dizemos que o estimador é **não viesado**,[3] conforme discussão no Capítulo 5.

Tabela 6.3: Parâmetros populacionais e amostrais da variável renda para outras quatro amostras aleatórias, com reposição, composta por $n = 5$ domicílios.

Indicadores estatísticos	Amostras do experimento de Monte Carlo			
para os estimadores da média	Renda	Escolaridade	Num. Filhos	Idade
N. de simulações	100 mil	100 mil	100 mil	100 mil
Média	1.225	5.92	2.21	35.99
Variância	52.424	1.19	0.44	75.20
Desvio padrão	229	1.09	0.66	8.67
Curtose	3.675	3.003	3.07	3.17
Assimetria	0.602	0.175	0.296	0.40
Mediana	1.204	5.8	2.2	35.40
1o. quartil	1.063	5.2	1.8	28.8
3o. quartil	1.362	6.6	2.6	41.6
Δ Interquartil	299	1.4	0.8	11.8
Média populacional	1.224	5.922	2.209	36.01

O primeiro quartil da distribuição do estimador, de acordo com a Tabela 6.3, é igual a R$ 1.063. Isso indica que, em média, 25% das amostras de 5 observações escolhidas aleatoriamente, para a população em questão, incorrerão em estimativas com valores menores do que R$ 1.063. Analogamente, em média, 25% das amostras de 5 observações, para a população em questão, incorrerão em estimativas com valores maiores do que R$ 1.362 (terceiro quartil). Ou seja, mesmo sabendo que em média o nosso estimador é não viesado, isso não significa que a amostra que temos em mãos está nos fornecendo uma estimativa razoável: podemos estar com uma estimativa menor do que R$ 1.063, por exemplo, incorrendo em uma subestimativa de aproximadamente R$ 160. Observe os indicadores de dispersão, dados pelo intervalo interquartil, pelo desvio padrão e pela variância, da distribuição do estimador da média para as quatro variáveis socioeconômicas. Conforme veremos mais adiante, quando aumentamos o tamanho da amostra n, esses indicadores tornam-se menores indicando distribuições amostrais menos dispersas em torno da média.

Vamos agora estudar o que acontece com a distribuição dos estimadores da média, quando aumentamos o tamanho da amostra para $n = 20$ ou para $n = 500$, por exemplo. Repetimos então o experimento de Monte Carlo, dessa vez considerando $n = 20$ e $n = 500$, ao invés de $n = 5$. Os histogramas correspondentes estão apresentamos nas Figuras 6.3 e 6.4. As estatísticas descritivas para as médias com base nas amostras simuladas também estão apresentadas nas Tabelas 6.4 e 6.5. Similarmente aos histogramas apresentados na Figura 6.2, para amostras de tamanho $n = 20$ e $n = 500$, as médias amostrais em média acertam os valores corretos populacionais. No entanto, quando comparamos os resultados para diferentes tamanhos de amostras, notamos que:

(i) quanto maior o tamanho da amostra, a dispersão dos estimadores amostrais ao redor da média populacional diminui. Por exemplo, para a variável renda *per capita*, para $n = 5$, o intervalo interquartil é igual a R$ 299; para $n = 20$ e $n = 500$, o intervalo interquartil passa a ser R$ 154 e R$ 31 respectivamente.

[3]Em inglês, *unbiased*.

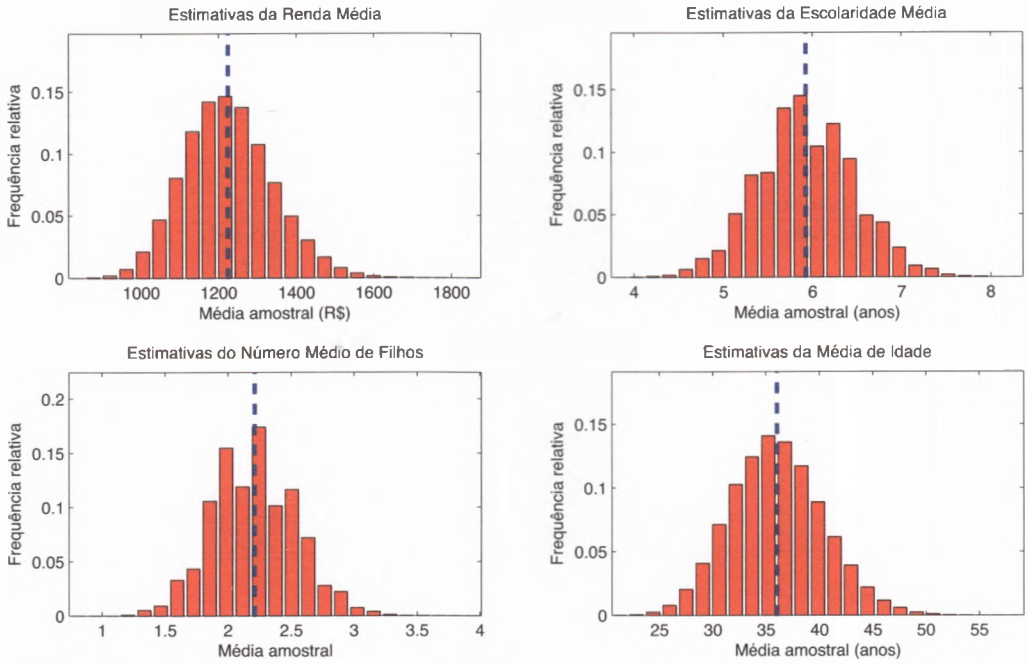


Figura 6.3: Histogramas para as médias amostrais com base nas simulações de Monte Carlo, para as quatro variáveis populacionais (amostras de tamanho $n = 20$).

Isso indica que, em 75% das amostra coletadas de tamanho $n = 500$, o valor da estimativa da média estará entre R$ 1.208 e R$ 1.239, implicando em um erro de estimação de no máximo R$ 15.

(ii) quanto maior o tamanho da amostra, menor a assimetria da distribuição amostral. Para a variável número de filhos, por exemplo, o coeficiente de assimetria passa de 0.3 para 0.16, e depois para 0.025, quando o tamanho da amostra assume valores $n = 5$, $n = 20$ e $n = 500$ respectivamente. Portanto, quando o tamanho da amostra aumenta, a distribuição amostral torna-se mais simétrica.

(iii) quanto maior o tamanho da amostra, a curtose da distribuição amostral aproxima-se do valor 3.0. Para a variável renda *per capita*, a curtose passou de 3.675 para 3.14, e depois para 3.00, quando o tamanho da amostra assume valores $n = 5$, $n = 20$ e $n = 500$.

(iv) quanto maior o tamanho da amostra, o histograma da distribuição amostral aproxima-se de um histograma da distribuição normal. Esse fato está consistente com a observação de que o coeficiente de assimetria aproxima-se de 0.0 e que a curtose aproxima-se de 3.0. Esses valores são exatamente os valores do coeficiente de assimetria e da curtose para uma distribuição normal.

(v) a variância da distribuição amostral decresce de forma inversamente proporcional ao tamanho da amostral. Por exemplo, para $n = 5$, a variância da renda per capita é igual a 52.424, enquanto para $n = 20$, essa variância passa a ser 13.019. Note que 52.424 / 13.019 = 4.02, que é aproximadamente a 20 / 5. Analogamente, quando $n = 500$ a variância é igual a 526, e 13.019 / 526 = 24.8, que é aproximadamente a 500 / 20. Se fizermos contas parecidas para todas as demais variáveis analisadas neste exemplo, notaremos

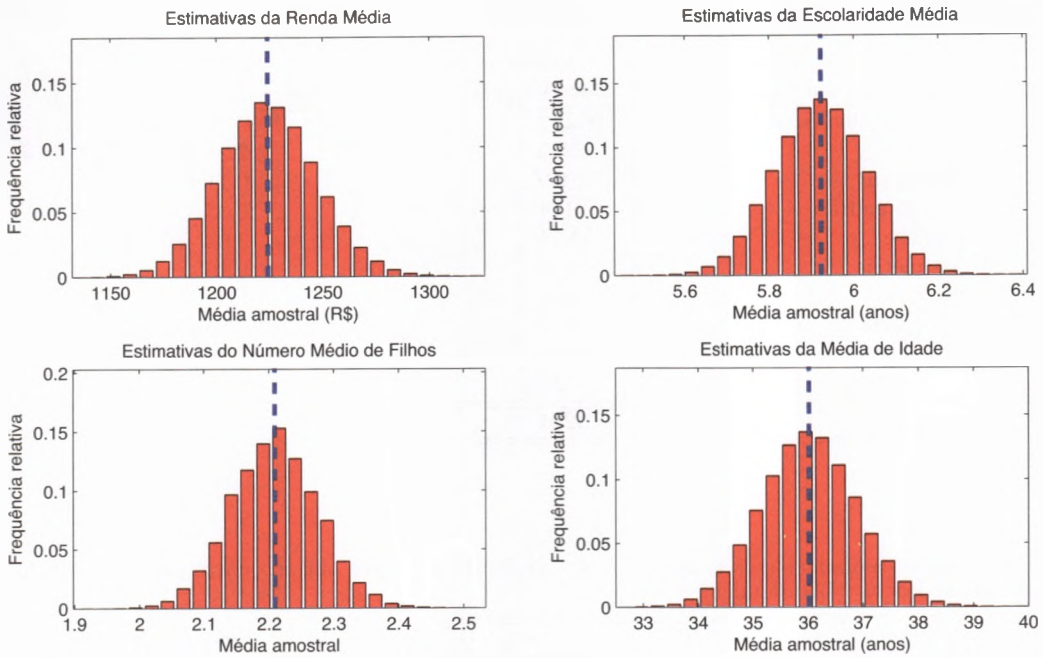


Figura 6.4: Histogramas para as médias amostrais com base nas simulações de Monte Carlo, para as quatro variáveis populacionais (amostras de tamanho $n = 500$).

a mesma relação entre variância e tamanho da amostra. De maneira geral, os resultados das simulações de Monte Carlo sugerem uma regra para a variância da forma

$$\text{Variância para amostra de tamanho } n = \frac{\text{Variância básica}}{n}. \tag{6.1}$$

Lembrando que o desvio padrão é a raiz quadrada da variância, uma forma similar pode ser escrita para o desvio padrão

$$\text{Desvio padrão para amostra de tamanho } n = \frac{\text{Desvio padrão básico}}{\sqrt{n}}. \tag{6.2}$$

Os cinco fatos observados acima podem ser analiticamente demonstrados utilizando-se ferramentas de processos estocásticos. A observação (i) acima está ligada ao teorema conhecido com **lei dos grandes números**, visto no Capítulo 4, enquanto os demais quatro itens estão relacionados ao teorema conhecido com **teorema central do limite**. Esses resultados são fundamentais em tudo que veremos neste texto, além de em tudo que o leitor usará em termos de econometria e estatística. Resultados como esses permitem inferir o que encontraremos nas simulações de Monte Carlo, em relação às características das distribuições amostrais dos estimadores, sem necessariamente efetuarmos tais simulações.

Especificamente para o estimador da média populacional, note que, para cada observação independente X_i na amostra, temos $\mathrm{E}[X_i] = \mu$, onde μ é a média populacional verdadeira (no caso, $\mu =$ R\$ 1.224). Ou seja, para cada observação retirada da população, em média o valor dessa observação é igual à média

Tabela 6.4: Parâmetros populacionais e amostrais da variável renda para outras quatro amostras aleatórias, com reposição, composta por $n = 20$ domicílios.

Indicadores estatísticos para os estimadores da média	Amostras do experimento de Monte Carlo			
	Renda	Escolaridade	Num. Filhos	Idade
N. de simulações	100 mil	100 mil	100 mil	100 mil
Média	1.225	5.93	2.21	35.99
Variância	13.019	0.296	0.11	18.97
Desvio padrão	114	0.544	0.33	4.36
Curtose	3.14	3.007	3.02	3.01
Assimetria	0.29	0.096	0.16	0.19
Mediana	1.219	5.9	2.2	35.85
1o. quartil	1.145	5.55	2	32.95
3o. quartil	1.296	6.3	2.45	39.00
Δ Interquartil	154	0.75	0.45	5.95
Média populacional	1.224	5.922	2.209	36.01

Tabela 6.5: Parâmetros populacionais e amostrais da variável renda para outras quatro amostras aleatórias, com reposição, composta por $n = 500$ domicílios.

Indicadores estatísticos para os estimadores da média	Amostras do experimento de Monte Carlo			
	Renda	Escolaridade	Num. Filhos	Idade
N. de simulações	100 mil	100 mil	100 mil	100 mil
Média	1.224	5.92	2.209	36.01
Variância	526	0.01	0.004	0.756
Desvio padrão	22.93	0.11	0.067	0.869
Curtose	3.00	3.01	3.01	3.006
Assimetria	0.06	0.02	0.025	0.027
Mediana	1.224	5.92	2.208	36.004
1o. quartil	1.208	5.85	2.164	35.416
3o. quartil	1.239	5.99	2.254	36.594
Δ Interquartil	31	0.146	0.09	1.178
Média populacional	1.224	5.922	2.209	36.01

populacional. De fato, a população completa possui valores de renda *per capita* $x_1, x_2, \ldots, x_N$, onde N é igual a 4.5 milhões de unidades na população completa. Seja X_i uma variável aleatória, que corresponde ao i-ésimo valor retirado em uma amostragem aleatória simples (com reposição). A variável X_i possui uma distribuição discreta, onde

$$\text{Prob}[X_i = x] = f(x) = \frac{1}{N},$$

onde x é um valor qualquer de renda na população $\{x_1, x_2, \ldots, x_N\}$. Note que a probabilidade de retirar cada um dos N valores da população é o mesmo; portanto, cada probabilidade tem valor igual a $1/N$. Portanto, o valor esperado de X_i, pela própria definição de valor esperado, conforme visto no Capítulo 3, é dado por

$$\text{E}[X_i] = \sum_{k=1}^{N} x_k f(x_k) = \sum_{k=1}^{N} x_k \frac{1}{N} = \frac{1}{N} \sum_{k=1}^{N} x_k,$$

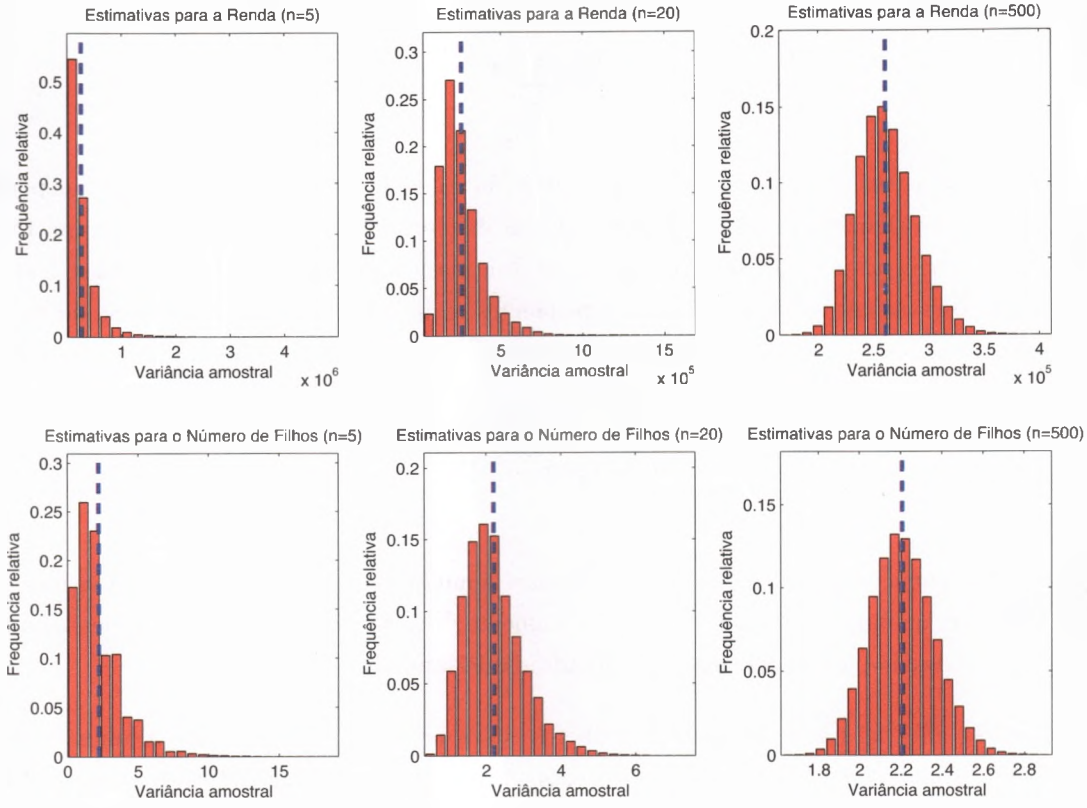


Figura 6.5: Histogramas para as variâncias amostrais com base nas simulações de Monte Carlo, para as variáveis número de filhos e renda per capita, com base em amostras de tamanhos $n = 5$, $n = 20$ e $n = 500$.

que é justamente a média populacional μ. Analogamente, a variância de X_i é calculada pela expressão para a variância populacional, conforme vimos no Capítulo 3. Dessa forma,

$$\mathrm{Var}[X_i] = \sum_{k=1}^{N}(x_k - \mu)^2 f(x_k) = \sum_{k=1}^{N}(x_k - \mu)^2 \frac{1}{N} = \frac{1}{N}\sum_{k=1}^{N}(x_k - \mu)^2,$$

que é justamente igual à variância populacional σ^2. Vamos agora ao estimador da média populacional, dado por $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i$. Gostaríamos então de estudar analiticamente a distribuição de $\hat{\mu}$. Já temos uma boa descrição da distribuição de $\hat{\mu}$ via simulações de Monte Carlo, conforme apresentado nas Figuras 6.2 a 6.4 e nas Tabelas 6.1 a 6.5. O primeiro fato que podemos investigar analiticamente é o viés do estimador da média; ou seja, em média, o estimador $\hat{\mu}$ está atingindo o valor verdadeiro do parâmetro μ? De fato,

$$\mathrm{E}[\hat{\mu}] = \mathrm{E}\Big[\frac{1}{n}\sum_{i=1}^{n} X_i\Big] = \frac{1}{n}\mathrm{E}\Big[\sum_{i=1}^{n} X_i\Big] = \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}[X_i] = \frac{n\mu}{n} = \mu.$$

Portanto, o estimador da média $\hat{\mu}$ é não-viesado. Quanto à variância do estimador $\hat{\mu}$, observe que

$$\text{Var}[\hat{\mu}] = \text{Var}\Big[\frac{1}{n}\sum_{i=1}^{n} X_i\Big] = \frac{1}{n^2}\text{Var}\Big[\sum_{i=1}^{n} X_i\Big] = \frac{1}{n^2}\sum_{i=1}^{n}\text{Var}[X_i] = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}.$$

Na derivação acima, utilizamos o fato de que a variância da soma de variáveis aleatórias independentes é igual à soma das variâncias, já que as retiradas X_i são independentes entre si, como visto na Proposição 3.4. A independência nesse caso deve-se ao fato de as retiradas ocorrerem com reposição. Caso não houvesse reposição, as observações não mais seriam independentes, e na derivação para a variância do estimador $\hat{\mu}$ teríamos que considerar também as covariâncias entre as retiradas. Para o desvio padrão do estimador $\hat{\mu}$, basta tirar a raiz quadrada da variância, obtendo-se

$$\text{desvio padrão}[\hat{\mu}] = \frac{\sigma}{\sqrt{n}}.$$

As duas últimas equações justificam as observações nas simulações de Monte Carlo, descritas nas Eqs. (6.1) e (6.2). Portanto, quando o tamanho da amostra aumenta, a variância cai proporcionalmente ao tamanho da amostra, e o desvio-padrão cai com a raiz quadrada do tamanho da amostra.

De acordo com a lei fraca dos grandes números, vista na Seção 4.6.1, quando a amostra vai para o infinito, a probabilidade de a média amostral se afastar da média populacional vai para zero. Portanto, a lei fraca dos grandes números justifica a consistência do estimador da média $\hat{\mu}$. Mais formalmente,

$$\hat{\mu} \xrightarrow{\text{P}} \mu,$$

ou seja, o estimador $\hat{\mu}$ converge em probabilidade para o parâmetro populacional μ. Conforme vimos nos exemplos para as médias estimadas para as quatro variáveis populacionais de interesse, quando o tamanho da amostra vai para o infinito, o histograma dos valores obtidos para $\hat{\mu}$ se aproxima cada vez mais de uma distribuição normal. Esse fato não é mera coincidência. Podemos então enunciar o **teorema central do limite** para a média amostral. Antes disso, porém, iremos discutir outro tipo de convergência para variáveis aleatórias (até agora, vimos o conceito de convergência em probabilidade).

Definição 6.1 (Convergência em distribuição ou em lei) Considere uma sequência de variáveis aleatórias $Y_1, Y_2, \ldots, Y_n, \ldots$. Dizemos que a sequência $\{Y_n\}$ converge em distribuição ou em lei para a variável aleatória Y, quando

$$F_{Y_n}(z) \to F_Y(z),$$

para todo ponto z no qual a função distribuição acumulada $F_Y(z)$ de Y é contínua.[4] As funções $F_{Y_1}(\cdot)$, $F_{Y_2}(\cdot)$, ..., $F_{Y_n}(\cdot)$, ... correspondem à sequência de funções distribuições acumuladas para as variáveis

[4] A função F_Y é contínua no ponto z se e somente se $\text{Prob}[Z = z] = 0$.

aleatórias $Y_1, Y_2, \ldots, Y_n, \ldots$. Quando a sequência $\{Y_n\}$ converge em distribuição ou em lei para a variável aleatória Y, escrevemos

$$Y_n \xrightarrow{\text{L}} Y.$$

A seguir, apresentaremos alguns resultados básicos relacionando convergência em probabilidade e convergência em distribuição. Os resultados abaixo. quando não especificado em contrário, aplicam-se tanto a variáveis aleatórias discretas quanto a variáveis aleatórias contínuas.

Proposição 6.1 (*Convergência em probabilidade implica em convergência em lei*) Se $Y_n \xrightarrow{\text{P}} Y$, então $Y_n \xrightarrow{\text{L}} Y$.

A Proposição 6.1 mostra que se uma sequência de variáveis aleatórias converge em probabilidade, então essa sequência também converge em distribuição. No caso geral, a contrapartida da Proposição 6.1 não ocorre, como mostra o Exemplo 6.1. Entretanto, se uma sequência de variáveis aleatórias converge em lei para uma constante, essa sequência também converge em probabilidade. Note que $Y = y_0$ possui uma distribuição degenerada na Proposição 6.2, visto que essa variável é uma constante.

Proposição 6.2 Se $Y = y_0$, ou seja, se Y é uma constante, então, se $Y_n \xrightarrow{\text{L}} Y$, a sequência $Y_n \xrightarrow{\text{P}} Y$.

Exemplo 6.1 (Convergência em distribuição não implica em convergência em probabilidade) Seja X_n uma sequencia variáveis aleatórias discretas com funções de frequência

$$f_{X_n}(1) = 1/2 + 1/n \text{ e } f_{X_n}(-1) = 1/2 - 1/n,$$

seja X um variável aleatória discreta com funções de frequência

$$f_X(1) = 1/2 \text{ e } f_X(-1) = 1/2,$$

e seja $Y = -X$ uma variável aleatória discreta que possui funções de frequência iguais a de X dadas por

$$f_Y(1) = 1/2 \text{ e } f_Y(-1) = 1/2.$$

Usando a Proposição 6.4, note que $X_n \xrightarrow{\text{L}} X$ e $X_n \xrightarrow{\text{L}} Y$. Note também que $X_n \xrightarrow{\text{P}} X$, mas X_n não converge em probabilidade para Y, pois $|X_n - Y| = 2$ em qualquer situação.

Proposição 6.3 Se $Y_n \xrightarrow{\text{L}} Y$, e a função $g(\cdot)$ é contínua, então $g(Y_n) \xrightarrow{\text{L}} g(Y)$.

Nesse caso, a Proposição 6.3 mostra que a convergência em lei continua válida se a sequência de variáveis aleatórias sofrer uma transformação contínua.

Adicionalmente, os dois teoremas e o corolário abaixo são referidos a Slutsky e aparecem em várias situações da teoria de probabilidade.

Teorema 6.1 (Slutsky) Se $Y_n \xrightarrow{\text{P}} y_0$ (uma constante) e a função $g(\cdot)$ é contínua em y_0, então $g(Y_n) \xrightarrow{\text{P}} g(y_0)$.

Teorema 6.2 (Slutsky) Se $Y_n \xrightarrow{\text{L}} Y$ e $U_n \xrightarrow{\text{P}} u_0$ (uma constante), então

(a) $Y_n + U_n \xrightarrow{\text{L}} Y + u_0$,
(b) $U_n Y_n \xrightarrow{\text{L}} u_0 Y$.

Corolário 6.1 (Slutsky) Seja $a_1, a_2, \ldots, a_n, \ldots$ uma sequência de constantes tendendo ao infinito, seja b um número fixo, e seja $Y_1, Y_2, \ldots, Y_n, \ldots$ uma sequência de variáveis aleatórias tal que,

$$a_n\big[Y_n - b\big] \xrightarrow{\text{L}} X,$$

para uma variável aleatória X. Seja $g(\cdot)$ uma função diferenciável, e que tenha primeira derivada $g'(\cdot)$ seja contínua no ponto b. Então,

$$a_n\big[g(Y_n) - g(b)\big] \xrightarrow{\text{L}} g'(b)X.$$

Proposição 6.4 Vamos supor que a sequência de variáveis aleatórias $\{Y_n\}$ seja composta por variáveis discretas, assumindo valores apenas no espaço amostral enumerável $\mathbb{X} = \{x_1, x_2, x_3, \ldots\}$. Sejam $f_{Y_1}(y)$, $f_{Y_2}(y), \ldots, f_{Y_n}(y), \ldots$, a sequência de funções de frequências para a sequência de variáveis aleatórias $\{Y_n\}$. Se $f_{Y_n}(y) \to f_Y(y)$ para todo $y \in \mathbb{X}$, onde $f_Y(y)$ é a função de frequência de Y, então $Y_n \xrightarrow{\text{L}} Y$.

Proposição 6.5 Vamos supor que a sequência de variáveis aleatórias $\{Y_n\}$ seja composta por variáveis contínuas estritamente. Sejam $f_{Y_1}(y)$, $f_{Y_2}(y)$, $\ldots$, $f_{Y_n}(y)$, $\ldots$, a sequência de funções de densidade para a sequência de variáveis aleatórias $\{Y_n\}$. Se $f_{Y_n}(y) \to f_Y(y)$ para todo $y \in \Re$, exceto em um conjunto de medida nula, onde $f_Y(y)$ é a função de densidade de Y, então $Y_n \xrightarrow{\text{L}} Y$.

Proposição 6.6 Se $Y_n \xrightarrow{\text{L}} Y$ e a função de distribuição acumulada $F_Y(y)$ é contínua, então $F_{Y_n}(y) \to F_Y(y)$ uniformemente em y.

Os resultados apresentados acima são muito utilizados quando queremos estudar as características de estimadores comumente encontrados em econometria e estatística. Apesar de eles terem um teor puramente para variáveis aleatórias univariadas, versões multivariadas estão disponíveis na literatura. Podemos então enunciar o **teorema central do limite**.

Teorema 6.3 (Central do limite). Seja $X_1, \ldots, X_n$ uma amostra aleatória independente e identicamente distribuída, com $\mathrm{E}[X_i] = \mu$ e $\mathrm{Var}[X_i] = \sigma^2$. Seja $\bar{X}$ a média amostral $\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$. Então, quando $n \to \infty$, temos

$$\sqrt{n}\left[\frac{\bar{X} - \mu}{\sigma}\right] \xrightarrow{\mathrm{L}} Z,$$

onde Z tem distribuição normal padronizada (média zero e variância um).

De acordo com o teorema central do limite, quando a amostra é grande o suficiente ($n = 500$, por exemplo, de acordo com as simulações apresentadas anteriormente), temos a aproximação

$$\sqrt{n}\left[\frac{\bar{X} - \mu}{\sigma}\right] \approx Z \Rightarrow \bar{X} \approx \frac{\sigma}{n} Z + \mu.$$

Lembrando do Exemplo 4.17 que, se Z tem distribuição normal padronizada, então $bZ + a$ tem distribuição normal com média a e variância b^2, concluímos que

$$\bar{X} \approx W, \tag{6.3}$$

onde W tem distribuição normal, com média μ e variância $\frac{\sigma^2}{n}$. A Eq. (6.3) explica a forma normal dos histogramas vistos nas Figuras 6.2 a 6.4.

6.4 Testes de hipóteses

Um dos principais tópicos em inferência estatística são os chamados **testes de hipóteses**. Conforme o próprio nome já especifica, a ideia dos testes de hipóteses é verificar, por meio dos dados, a validade ou não de um determinado conjunto de pressupostos. O primeiro passo nos testes de hipóteses é transformar o conjunto de pressupostos a serem testados em um conjunto de restrições a respeito dos parâmetros livres do modelo estatístico. A partir de então, é preciso construir uma **estatística teste**, que é calculada a partir dos dados amostrais. A estatística teste também é uma variável aleatória, e como tal possui propriedades probabilísticas, tendo por exemplo uma função densidade de probabilidade. A partir da distribuição da estatística teste, devem-se ser especificadas regras de aceitação ou rejeição dos pressupostos. Essas regras de aceitação ou rejeição levam em conta valores de corte: para valores abaixo (ou acima, a depender da situação) dos valores de corte, os pressupostos são aceitos ou rejeitados.

Testes de hipóteses são bastante utilizados na maioria das aplicações estatísticas. A discussão nesta seção tem por objetivo descrever a intuição geral da utilização dos testes de hipótese e como eles podem ser empregados em diferentes contextos de estimação. Inicialmente, trataremos a aplicação de testes de

hipóteses para testar valores da média populacional. Em seguida, iremos tratar de testes de hipóteses no contexto de estimação via máxima verossimilhança. Para máxima verossimilhança especificamente, discutiremos as três categorias de testes de hipótese: testes de Wald; testes de razão de verossimilhança; testes de multiplicadores de Lagrange. Esses testes também serão importantes no contexto de modelos de regressão apresentados nos Capítulos 8 e 9.

6.4.1 Testes de hipóteses para a média populacional

Para exemplificar melhor os conceitos que dão suporte aos testes de hipóteses, vamos considerar inicialmente um exemplo bastante simplificado, onde a amostra selecionada é composta de observações, cada qual obedecendo a uma distribuição normal, com média μ desconhecida e variância σ^2 conhecida (o que, na prática, dificilmente será o caso, mas utilizaremos a variância conhecida por enquanto por motivos puramente didáticos). A partir desse exemplo simplificado, relaxaremos diversas das hipóteses iniciais, mostrando que os conceitos gerais não se alteram.

População normal com variância conhecida

Seja $X_1, X_2, \ldots, X_n$ uma amostra aleatória, com observações independentes e identicamente distribuídas, cada qual com distribuição normal com média μ desconhecida e variância σ^2 conhecida. O objetivo da análise estatística é fazer inferência a respeito do parâmetro μ. O estimador da média populacional μ é dado pela média populacional $\hat{\mu} = \bar{X}$. O estimador $\hat{\mu}$ tem distribuição normal, com média

$$\mathrm{E}[\hat{\mu}] = \frac{\mathrm{E}[X_1 + \cdots + X_n]}{n} = \frac{\sum_{i=1} n\mathrm{E}[X_i]}{n} = \frac{n\mu}{n} = \mu,$$

e, portanto, o estimador $\hat{\mu}$ é não viesado. A variância de $\hat{\mu}$ tem expressão

$$\mathrm{Var}[\hat{\mu}] = \frac{\mathrm{Var}[X_1 + \cdots + X_n]}{n^2} = \frac{\sum_{i=1} n\mathrm{Var}[X_i]}{n^2} = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}.$$

Note que, na derivação acima, utilizamos o fato de que a variância da soma de variáveis aleatórias independentes é igual à soma das variâncias de cada variável aleatória individualmente, conforme visto na Proposição 3.4. O fato de que o estimador $\hat{\mu}$ tem distribuição normal vem do fato de que a soma de variáveis aleatórias normais também tem distribuição normal (vide Exemplo 4.19). Utilizaremos a notação $\sigma^2_{\hat{\mu}}$ para se referir à variância do estimador $\hat{\mu}$ da média populacional, e a notação $\sigma_{\hat{\mu}}$ para se referir ao desvio padrão do estimador $\hat{\mu}$.

Exemplo 6.2 (Avaliação de programa de governo) Vamos supor que a amostra $X_1, X_2, \ldots, X_n$ corresponde a valores coletados sobre as notas de matemática na rede pública de ensino médio em Salvador.

A meta da Secretaria de Educação do Estado da Bahia é atingir uma nota média acima de 70% para um exame padronizado. Depois de dois anos de um programa de melhoria no ensino público, a Secretaria pretende testar se de fato a média de ensino evoluiu acima dos 70% almejados. Para fazer essa avaliação, a Secretaria tem duas alternativas:

1) Aplicar o exame a absolutamente todos os alunos da rede fundamental de ensino, em todas as escolas da região metropolitana de Salvador, tendo então uma espécie de censo.

2) Aplicar o exame a apenas uma amostra aleatória de alunos e efetuar uma inferência estatística.

A vantagem da primeira alternativa é que não estaríamos sujeitos à variabilidade implícita ao processo amostral da segunda alternativa. Aplicar o exame a todos os alunos do universo objetivado (região metropolitana de Salvador) necessitaria de uma estrutura muito mais cara em termos de infraestrutura. Seria necessário um controle sobre o processo de aplicação das provas, que poderia incorrer nos chamados **erros não amostrais**. O grande problema dos erros não amostrais é a dificuldade para corrigi-los. Por exemplo, pode ser que um determinado diretor de uma escola, para fazer a sua escola parecer melhor do que as outros, burlasse o processo de coleta, comprometendo as conclusões gerais do estudo. Por esses motivos, ao invés de um recenseamento exaustivo, sujeito a inúmeros erros não amostrais e incorrendo em gastos bem maiores, talvez seja mais interessante ter um processo amostral, controlado, a custo mais baixo, onde os erros possíveis são os erros amostrais, que podem ser tratados via procedimentos amplamente disponíveis na literatura estatística. Esses procedimentos serão discutidos nesta seção.

A Secretaria de Educação decide então por fazer uma coleta de informações via amostra de alunos que prestarão o exame. Uma vez selecionada a amostra de n alunos, as provas são aplicadas, em ambiente controlado, e são então coletadas as observações $X_1, X_2, \ldots, X_n$. Estima-se então a média populacional de notas para toda a população de alunos do ensino médio da rede pública na região metropolitana de Salvador. O estimador da média populacional é dado pela média amostral $\hat{\mu}$. Supondo que as notas dos alunos nas provas têm distribuição normal, com média μ desconhecida e variância σ^2 conhecida, vamos agora discutir um pouco o processo de inferência estatística e testes de hipóteses.

Vamos imaginar, primeiramente, que a nota média da amostra foi igual $\hat{\mu} = 32.6\%$. Com uma nota dessa ordem, é muito provável que o ensino médio em Salvador esteja muito aquém da meta de 70%. A questão é a seguinte: será que essa diferença de 32.6% para 70% não se deve meramente ao fato de que escolhemos uma amostra aleatória de n alunos não muito preparados? Nesse caso, condenar o programa de melhorias da Secretaria com base nessas notas médias seria totalmente injusto. Imagine que a nota média da amostral foi igual a $\hat{\mu} = 83.6\%$. Nesse caso, a nota média é muito superior à meta de 70%, mas essa média amostral pode decorrer meramente de termos selecionado uma amostra com n alunos muito bons, que não são representativos da população de aluno da rede de ensino público médio. Nesse outro extremo, estaríamos possivelmente mantendo, ou até mesmo aumentando, os gastos para um programa de políticas públicas educacionais que não está gerando os resultados desejados. Esses cenários se agravam quando as notas médias obtidas na amostra são $\hat{\mu} = 72.3\%$ ou $\hat{\mu} = 68.2\%$. No primeiro caso, a nota média está acima

da meta, e no segundo caso a nota média está abaixo; porém, em ambos os casos, pode haver uma grande possibilidade de que essas notas (acima ou abaixo) sejam exclusivamente efeito da amostra coletada, e não reflitam a realidade. Como então resolver esse impasse?

Felizmente, os estatísticos desenvolveram uma metodologia para fornecer subsídios para os tomadores de decisões, seja no setor público, seja no setor privado, com base em resultados amostrais. Imagine que o concenso entre os educadores e o público em geral é que, em média, a população de estudantes do ensino público médio não possua nota acima dos 70%. Ou seja, a política educacional de melhorias não cumpriu os objetivos previstos. Nesse sentido, dizemos que a hipótese básica, ou **hipótese nula**, é que a média da população μ é menor ou igual a 70%, sem ser maior do que isso. Por outro lado, a hipótese dos defensores da política pública é que as notas em média são maiores do que 70%. Esse segunda hipótese é conhecida como **hipótese alternativa**; matematicamente, a hipótese alternativa pode ser escrita como $\mu > 70\%$.

Vamos supor que o resultado da pesquisa foi uma nota média amostral de $\hat{\mu} = 72.3\%$. Os indivíduos contrários à manutenção da política educacional existente acham que essa diferença de 2.3% acima da meta deve-se exclusivamente ao erro amostral; ou seja, foram coletados alunos, em média, melhores do que os alunos em geral do ensino público, e que, portanto, esses alunos não estariam refletindo a realidade da população como um todo. Aqui entra o cerne da metodologia de testes de hipóteses. Os indivíduos a favor da política educacional existente argumentam então da seguinte maneira:

(i) Vamos supor que, na melhor das hipóteses, a nota média populacional seja $\mu = 70\%$, e que, portanto, a política educacional não conseguiu cumprir o seu objetivo de superar a meta. Adicionalmente, conhece-se (conforme suposto no início desta seção) a variância populacional igual a $\sigma^2 = 1\% = 0.01$. O desvio padrão populacional é igual então a $\sigma = 10\%$. De acordo com a discussão acima, o estimador $\hat{\mu}$ tem distribuição normal, com média $\mu = 70\%$ (supondo-se o melhor caso, de acordo com a hipótese nula), e desvio padrão $\sigma/\sqrt{n}$. Para uma amostra de $n = 100$ indivíduos, o desvio padrão da distribuição do estimador $\hat{\mu}$ será igual a $10\%/\sqrt{100} = 1\%$. Portanto, supondo-se a hipótese nula (programa não foi bem sucedido), o estimador $\hat{\mu}$ tem distribuição normal com média 70% e desvio padrão igual a 1%. Sob essas condições, qual a probabilidade então de a amostra resultar em uma média amostral igual a $\hat{\mu} = 72.3\%$?

(ii) Antes de responder à pergunta acima, voltemos ao fato de que o estimador $\hat{\mu}$ tem distribuição normal com média $\mu = 70\%$ e desvio padrão $\sigma = 1\%$. Com base nesse valores, vimos no Capítulo 3, que a variável aleatória Z, dada por

$$Z = \frac{\hat{\mu} - \mu}{\sigma_{\hat{\mu}}}$$

tem distribuição normal, com média zero e variância um (normal padronizada). Para uma distribuição normal padronizada, qual então o valor de corte $c_{5\%}$ para o qual é possível obter um valor acima desse valor de corte com probabilidade igual a 5%? Checando na tabela da distribuição normal padronizada, nota-se que o valor de corte $c_{5\%}$ é igual a 1.645. Similarmente, o valor de corte $c_{1\%}$, para o qual é possível obter um valor acima dele com probabilidade igual a 1%, é igual a $c_{1\%} = 2.326$. Portanto, obter valores

para Z acima de 1.645 é um acontecimento raro; enquanto obter um valor de Z acima de 2.326 é um evento raríssimo (acontece com probabilidade igual a 1%).

(iii) Podemos então estabelecer uma regra de procedimento. Essa regra consiste no seguinte: com base no valor obtido para $\hat{\mu}$ na amostra, calculamos o valor da estatística Z. Caso esse valor seja maior do que um valor de corte pré-definido, suporemos que a hipótese nula está errada, e a hipótese alternativa está correta. Caso contrário, caso o valor da estatística Z seja menor do que o valor de corte, suporemos que a hipótese nula está correta e a alternativa está errada. Qual valor de corte definir? Geralmente, assume-se valores de corte para probabilidades de 10%, 5% ou 1%. Se a probabilidade escolhida for igual a 1%, o valor de corte será, conforme visto acima, igual a $c_{1\%} = 2.326$. A probabilidade escolhida é chamada **nível de significância** do teste de hipótese. A estatística Z é conhecida como **estatística teste**.

(iv) Toda regra de decisão tem um lado positivo e um lado negativo. O lado positivo é que, uma vez estabelecida a regra, fica claro como proceder. No caso de regras estatísticas para testes de hipóteses, dado que ela tem sido utilizada há várias décadas, em diversas áreas da ciência, as decisões tomadas de acordo com elas são amplamente divulgáveis. O lado negativo é que, ao tomarmos uma decisão com base na regra, poderemos estar incorrendo em erros de decisão. Esses erros podem ser divididos em dois tipos básicos no caso de regras de decisão em testes de hipóteses: aceitar a hipótese nula, quando ela na verdade está errada (no exemplo, esse seria de achar que o programa educacional não é eficiente quando na verdade ele está funcionando); ou recusar a hipótese nula, quando na verdade ela está correta (no exemplo, concluir que o programa educacional é eficaz, quando na verdade ele não é). Esses dois tipos de erros possuem uma terminologia amplamente conhecida em estatística e econometria:

Erro do tipo I. Acontece quando rejeitamos a hipótese nula, quando na verdade ela está correta.

Erro do tipo II. Acontece quando aceitamos a hipótese nula, quando na verdade ela está errada.

Quando a hipótese nula está correta, a estatística teste Z tem distribuição normal com média $\mu = 70\%$ e desvio padrão $\sigma = 1\%$. Somente rejeitaremos a hipótese nula quando a estatística teste, calculada com base na amostra, apresentar valor maior do que o valor de corte $c_{5\%}$, por exemplo. Porém, pela própria definição do valor de corte, a estatística teste assume valor maior do que $c_{5\%}$ com probabilidade igual a 5%, que é o nível de significância. Portanto, o nível de significância pode ser interpretado como a probabilidade de erro do tipo I, de acordo com a regra do teste de hipótese.

(v) Pois então vamos aplicar a regra de decisão ao exemplo do estudo sobre a eficácia da política educacional em superar a meta estabelecida. Vamos considerar um nível de significância de 5%. O valor da média amostral obtida da amostral foi igual a $\hat{\mu} = 72.3\%$ de rendimento nas provas. Qual o valor da estatística teste correspondente? Basta então aplicar a fórmula da transformação

$$z = \frac{72.3\% - 70\%}{1\%} = \frac{2.3\%}{1\%} = 2.3.$$

Note que o valor $z = 2.3$ é adimensional. Para o nível de significância de 5%, o valor de corte é igual a $c_{5\%} = 1.645$. Portanto, de acordo com a nossa regra de teste de hipóteses, como o valor da estatística teste z é maior do que o valor de corte, rejeitamos a hipótese nula e aceitamos a hipótese alternativa de que o programa educacional foi de fato efetivo em superar a meta de 70% de aproveitamento nos exames do estudo. O valor z da estatística teste não é maior do que o valor crítico $c_{1\%}$, e, portanto, não é possível rejeitar a hipótese nula a um nível de significância de 1%.

Os passos descritos no Exemplo 6.2 para testar a hipótese estatisticamente de que o programa de governo era efetivo tem três componentes fundamentais, que serão o alicerce de todos os testes de hipóteses apresentados nos exemplos neste capítulo.

(A) **A estatística teste**. No exemplo acima, a estatística teste é dada pelo quociente

$$Z = \frac{\hat{\mu} - \mu_0}{\sigma_{\hat{\mu}}}.$$

O símbolo μ_0 foi usado nessa equação para deixar claro que esse é o valor de μ supondo-se a hipótese nula.

(B) **A distribuição da estatística teste sob a hipótese nula**. Para o exemplo acima, a estatística teste Z tem distribuição normal padronizada

$$Z \sim \text{Normal}[0, 1].$$

(C) **As regras de rejeição ou aceitação da hipótese nula**. Essas regras de aceitação ou rejeição da hipótese nula em detrimento da hipótese alternativa irão depender do nível de significância,[5] que é a probabilidade de rejeitar a hipótese nula quando ela for verdadeira. No exemplo acima, para um nível de significância de 5%, a regra era: rejeitar a hipótese nula se o valor z da estatística teste fosse maior do que o valor crítico $c_{5\%} = 1.645$. Uma regra similar pode ser utilizada para um nível de significância de 1%. De fato, normalmente usamos valores pequenos, pois queremos evitar com alta probabilidade rejeitar a hipótese nula quando ela for verdadeira.

Com base na estatística teste e na distribuição da estatística teste sob a hipótese nula, podemos proceder com os mesmos passos estabelecidos acima, em relação à regra de rejeição ou aceitação da hipótese nula. De fato, conforme veremos ao longo deste capítulo, o que muda de situação para situação são somente a estatística teste e sua distribuição correspondente.

Exemplo 6.3 (Avaliação da redução de perdas da instituição financeira) Iremos agora considerar o caso de uma instituição financeira que está tentando reduzir o montante total perdido com perdas operacionais por falhas humanas nas suas agências. O valor médio reportado para as perdas em todas as agências há dois anos foi de R$ 250 mil por agência, ao ano. Por conta desse grande montante, a instituição financeira

[5] O nível de significância é representado em muitos livres de estatística e econometria como α.

resolveu executar um grande projeto para reduzir essas perdas. O objetivo do projeto é reduzir a média de perdas para menos de R$ 150 mil por agência, por ano. Após dois anos de implantação, o setor de avaliação da instituição está incumbido de avaliar se essa meta foi atingida; ou seja, se em média as perdas nas agências são menores do que R$ 150 mil. Uma amostra aleatória de $n = 64$ agências foi visitada, e suas perdas foram medidas com precisão durante um ano. A instituição gostaria então de avaliar, por meio da média amostral, se o impacto do projeto foi de fato válido, considerando-se todo o universo de agências do conglomerado.

Da mesma maneira que no caso anterior, vamos supor que as perdas das agências seguem uma distribuição normal, com média μ desconhecida (que queremos estimar e testar) e com variância conhecida (apenas para fins didáticos) $\sigma^2 = (\text{R\$ } 20.000)^2$; ou seja, o desvio padrão das perdas na população é de σ = R$ 20 mil. Temos uma amostra de tamanho $n = 64$, constituída pelas observações $X_1, X_2, \ldots, X_n$. O estimador da média de perdas na população de agências é dada por $\hat{\mu} = \bar{X}$. Conforme já visto anteriormente, o estimador $\hat{\mu}$ tem distribuição normal com média $\hat{\mu}$ (o estimador da média é não viesado) e com desvio padrão

$$\sigma_{\hat{\mu}} = \frac{\sigma}{\sqrt{n}} = \frac{20.000}{8} = \text{R\$2.500}.$$

Queremos testar se a média populacional é de fato menor do que a meta de R$ 150 mil. Depois de coletada a amostra, e calculada a média, obteve-se $\hat{\mu}$ = R$ 146 mil. Portanto, um valor menor do que a meta R$ 150 mil, mas que pode ser decorrente puramente de termos selecionado um conjunto de agências adequadas na amostra aleatória. Temos que testar então se essa diferença deve-se a um acontecimento aleatório ou corresponde de fato à eficácia do programa de melhorias. Para isso, novamente, vamos executar os procedimentos de testes de hipóteses. Inicialmente, precisamos de uma estatística teste, que nesse caso vai ser exatamente a estatística Z vista acima,

$$Z = \frac{\hat{\mu} - \mu_0}{\sigma_{\hat{\mu}}} \sim \text{Normal}[0, 1].$$

Neste exemplo, diferentemente do Exemplo 6.2, a hipótese nula é de que a média de perdas por agência é maior ou igual a R$ 150 mil por ano, enquanto a hipótese alternativa é de que a média de perdas é menor do que R$ 150 mil. Introduzimos agora mais notações para a hipótese nula e para a hipótese alternativa. A hipótese nula é representada pela notação H_0, enquanto a hipótese alternativa é representada pela notação H_A. Representamos então

$$H_0 : \ \mu \geq \mu_0$$
$$H_A : \ \mu < \mu_0.$$

onde μ_0 = R$ 150 mil. Tanto o teste de hipótese para o Exemplo 6.2 quanto o teste para o exemplo atual são chamados de **testes unicaudais**, pois as hipóteses testadas utilizam probabilidades em uma das

caudas da distribuição da estatística teste. No Exemplo 6.4, veremos uma situação onde teremos um **teste bicaudal**, pois utilizaremos ambas as caudas da distribuição da estatística teste.

Até agora conseguimos escrever a estatística teste Z, identificar a distribuição da estatística teste, e conseguimos escrever a hipótese nula e a hipótese alternativa. Temos agora que especificar a regra de aceitação ou rejeição da hipótese nula. Para isso, temos que estabelecer um valor de corte tal que, se o valor da estatística teste calculada a partir da amostra for menor do que o valor de corte, rejeitamos a hipótese nula; caso contrário, aceitamos a hipótese nula (diferentemente do exemplo anterior, onde a regra estabelecia que a rejeição da hipótese nula ocorreria quando o valor da estatística teste, calculado a partir da amostra, fosse maior do que o valor de corte). Novamente, o valor de corte virá do valor definido para o nível de significância do teste estatístico. Para um nível de significância de 5%, temos que determinar um valor de corte $c_{5\%}$ tal que a probabilidade da estatística teste (supondo que a hipótese nula seja verdadeira) assumir a um valor menor que $c_{5\%}$ seja igual a 5%. A partir da tabela da distribuição normal padronizada, o valor de corte é igual a $c_{5\%} = -1.645$. Para um nível de significância de 1%, o valor de corte é igual a $c_{1\%} = -2.326$. Note que os valores de corte para a regra de rejeição ou aceitação da hipótese nula neste exemplo são diferentes dos valores de corte no exemplo anterior. Se o nível de significância escolhido for igual a 5%, então a probabilidade de erro tipo I (rejeitar a hipótese nula quando ela é verdadeira) será igual a 5%. Similarmente, se o nível de significância selecionado for 1%, então a probabilidade de erro tipo I será de 1%. Para um nível de significância de 5%, o intervalo $(-\infty, -1.645)$ é conhecido como **região de rejeição**, pois essa é a região para a qual rejeitaremos a hipótese nula, caso a estatística teste caia dentro dela. Similarmente, para um nível de significância de 1%, o intervalo $(-\infty, -2.326)$ é a correspondente região de rejeição.

Vamos então aplicar a regra selecionada aos dados coletados na amostra de 64 agências. O valor da estatística teste será

$$z = \frac{\hat{\mu} - \mu_0}{\sigma_{\hat{\mu}}} = \frac{\text{R\$146.000} - \text{R\$150.000}}{\text{R\$2.500}} = -\frac{4.000}{2.500} = -1.6.$$

Portanto, o valor da estatística teste com base na amostra coletada é igual a -1.6, que não é menor do que $c_{5\%}$ nem do que $c_{1\%}$. Portanto, de acordo com a nossa regra de rejeição ou aceitação da hipótese nula, não rejeitamos a hipótese nula a um nível de significância de 5% nem de 1%.

Exemplo 6.4 (Controle de qualidade de uma fábrica) Nos dois exemplos anteriores, trabalhamos com exemplos de testes de hipóteses monocaudais. Neste exemplo, trataremos de uma situação de teste de hipótese bicaudal. Imaginemos agora que o controle de qualidade de uma fábrica está tentando reduzir a falha média nas peças produzidas para 1.2 milímetros, já que esse é o permitido pela associação de produtores em acordo com a instituição de fiscalização. Para o gerente de processos da fábrica, o ideal é que a falha média fique exatamente em 1.2 mm; uma falha média maior do que essa pode incorrer em rejeição do lote produzido, enquanto uma falha média menor do que 1.2 mm pode significar muito tempo perdido na inspeção, incorrendo em custos desnecessários. Para chegar a uma falha média de 1.2 mm, um

processo novo de fabricação foi introduzido, e agora precisa ser testado continuamente. A cada mês, uma amostra de $n = 400$ peças é selecionada, e as medidas das falhas são tiradas e anotadas em um banco de dados. Os analistas da empresa realizam então um teste de hipótese para checar se a falha média da população de todas as peças produzidas tem ou não uma falha média exatamente igual a 1.2 mm, conforme desejado pela gerência.

Sejam então $X_1, X_2, \ldots, X_n$ as n observações coletadas para a amostra. Vamos supor que a população de todas as peças da fábrica possui média populacional desconhecida μ (a qual queremos testar) para as falhas das peças, e desvio padrão conhecido $\sigma = 1.5$ mm. A amostra $X_1, X_2, \ldots, X_n$ é composta de observações independentes e identicamente distribuídas (iid), todas com distribuição normal. O estimador para a média populacional μ é o velho estimador $\hat{\mu}$, com base na média amostral $\bar{X}$. O valor do estimador $\hat{\mu}$ a partir da amostra foi calculado em $\hat{\mu} = 1.117$ mm. Queremos testar se o valor da falha média em toda a população é diferente de 1.2, não podendo ser nem maior nem menor do que 1.2. Portanto, formulamos o teste de hipótese de acordo com as hipóteses nula e alternativa a seguir

$$H_0 : \mu = \mu_0$$
$$H_A : \mu \neq \mu_0,$$

onde $\mu_0 = 1.2$. Temos novamente que especificar a estatística teste, e a sua respectiva distribuição. Em seguida, iremos estabelecer uma regra de rejeição ou aceitação da hipótese nula, com base na estatística teste calculada a partir dos dados da amostra. Finalmente, iremos aplicar a regra estabelecida para checar se, no caso da amostra coletada, rejeitamos ou aceitamos a hipótese nula.

A estatística teste novamente será a estatística Z, dada por

$$Z = \frac{\hat{\mu} - \mu_0}{\sigma_{\hat{\mu}}} \sim \text{Normal}[0, 1].$$

Novamente, supondo que a hipótese nula está correta, a estatística teste tem distribuição normal padronizada. Para a nossa regra de rejeição ou aceitação, como estamos querendo testar se o valor de μ é ou não igual a $\mu_0 = 1.2$ mm, temos que estabelecer uma região de rejeição tanto quando Z for menor de que um determinado valor crítico, quanto quando Z for maior do que um outro valor crítico. Ou seja, quando Z for muito grande ou Z for muito pequeno, teremos evidências suficientes para rejeitar a hipótese nula, e concluir que $\mu \neq 1.2$ mm. Mas como escolher esses valores críticos para cima e para baixo. A resposta a essa pergunta vem novamente da determinação do nível de significância do teste, que corresponde também à probabilidade de erro do tipo I (rejeitar a hipótese nula quando ela está correta). Para um nível de significância de 5%, a regra é de que iremos rejeitar H_0 quando $Z > 1.96$ ou quando $Z < -1.96$. Mas por que 1.96 como extremos? Note que, para uma distribuição normal padronizada, a probabilidade de ela assumir valores menores do que -1.96 ou maiores do que 1.96 é igual a 5%, de acordo com a tabela específica. Portanto, quando a hipótese nula é de fato verdadeira, a probabilidade de rejeitá-la

é igual a 5%, de acordo com a regra estabelecida. A região de rejeição nesse caso é composta pelos dois intervalos $(-\infty, -1.96) \cup (1.96, +\infty)$. Para um nível de significância de 1%, os valores críticos são -2.58 e 2.58; ou seja, rejeitamos a hipótese nula quando $Z < -2.58$ ou quando $Z > 2.58$. Nesse caso, a região de rejeição é a união dos intervalos $(-\infty, -2.58)) \cup (2.58, +\infty)$.

Finalmente, com base nos valores da amostra, podemos calcular o valor da estatística teste Z e aplicar as regras de rejeição ou aceitação da hipótese nula. Com base no valor de $\hat{\mu} = 1.117$ para o estimador da média populacional, o valor da estatística Z será

$$Z = \frac{\hat{\mu} - \mu_0}{\sigma_{\hat{\mu}}} = \frac{1.117 - 1.2}{\frac{1.5}{\sqrt{400}}} = \frac{-0.083}{0.075} = -1.107.$$

Portanto, o valor da estatística Z, igual a -1.107 cai fora da região de rejeição para a hipótese nula, tanto para um nível de significância de 1% quanto para um nível de significância de 5%. Aceitamos, então, a hipótese nula de que a média populacional μ é igual ao valor alvo de 1.2 mm.

População normal com variância desconhecida

Nos exemplos anteriores, supusemos que as populações que geravam as amostras tinham distribuições normais, com médias desconhecidas (para as quais efetuamos testes de hipóteses) e variâncias conhecidas. Essa simplificação para a variância foi meramente um artifício didático para facilitar a discussão. Na prática, raramente iremos nos confrontar com situações onde a variância populacional é conhecida. Nesta seção, a suposição de variância conhecida será relaxada, e iremos ter que estimá-la também a partir da amostra. Conforme veremos, os procedimentos para efetuarmos testes de hipóteses a respeito da média populacional μ são muito similares aos procedimentos nos três exemplos anteriores. A única diferença é que a distribuição da estatística teste não mais possui distribuição normal, e sim distribuição t-Student. Por conta disso, as regras de rejeição ou aceitação da hipótese nula utilizarão valores críticos diferentes dos utilizados acima. Ao longo desta seção, revisitaremos os 3 exemplos da seção anterior, considerando agora que as variâncias populacionais são desconhecidas.

Exemplo 6.5 (Continuação do Exemplo 6.2 – Agora supondo a variância desconhecida) Para o Exemplo 6.2, vamos supor agora que a variância é desconhecida e tem que ser estimada a partir dos dados. O estimador não viesado para a variância é dado pela fórmula

$$\hat{\sigma}^2 = \frac{1}{n-1} \sum_{i=1}^{n} [X_i - \bar{X}]^2. \tag{6.4}$$

Com base na amostra $X_1, X_2, \ldots, X_n$ coletada para as notas dos estudantes na amostra, calculou-se o valor da estimativa $\hat{\sigma}^2 = 2.21\% = 0.0221$, e a média amostral continua sendo $\hat{\mu} = 72.3\%$. O procedimento

para testar a hipótese nula versus a hipótese alternativa continua sendo o mesmo para o caso de variância populacional conhecida. A estatística teste agora passa a ser dada pela expressão

$$Z = \frac{\hat{\mu} - \mu_0}{\hat{\sigma}_{\hat{\mu}}}.$$

onde $\hat{\sigma}_{\hat{\mu}}$ corresponde à estimativa do desvio padrão da distribuição do estimador $\hat{\mu}$. Quando conhecíamos a variância populacional σ^2, o denominador na fórmula para estatística teste Z era dada por

$$\sigma_{\hat{\mu}} = \frac{\sigma}{\sqrt{n}}.$$

onde σ era a variância populacional, que supúnhamos ser conhecida. No caso de a variância populacional ser desconhecida, como é o caso neste exemplo, substituímos a expressão acima pela estimativa

$$\hat{\sigma}_{\hat{\mu}} = \frac{\sqrt{\hat{\sigma}^2}}{\sqrt{n}},$$

onde $\hat{\sigma}^2$ é dado pela Eq. (6.4). Portanto, substituímos a variância populacional pela sua estimativa, e daí calculamos o desvio padrão para o estimador $\hat{\mu}$. A distribuição da estatística teste não é mais uma normal padronizada. No caso anterior, quando a variância populacional era conhecida, não tínhamos que levar em conta a imprecisão na estimação da própria variância. No caso atual, onde a variância é desconhecida, a distribuição para Z é influenciada pela imprecisão[6] na estimação da variância $\hat{\sigma}^2_{\hat{\mu}}$, e consequentemente do desvio padrão $\hat{\sigma}_{\hat{\mu}}$. Pode-se mostrar que,[7] quando a amostra é composta de observações $X_1, \ldots, X_n$, independentes e identicamente distribuídas, com distribuição normal, com média μ e variância σ^2, a estatística teste Z tem distribuição t-Student com $n-1$ graus de liberdade. Conforme vimos na Seção 3.5, se uma variável aleatória W tem distribuição t-Student, ela tem função de densidade

$$f(w) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{(\pi\nu)^{\frac{1}{2}}\Gamma\left(\frac{\nu}{2}\right)\left(1 + \frac{w^2}{\nu}\right)^{\frac{\nu+1}{2}}}, \quad \text{para } w \in \Re. \tag{6.5}$$

O parâmetro ν é conhecido como número de graus de liberdade. Devido à sua simetria, a distribuição t-Student possui média $\mathrm{E}[W] = 0$. No caso da estatística teste Z, para o nosso teste de hipótese sobre a média μ, ela tem distribuição t-Student com $\nu = n - 1$. A forma da função de densidade de probabilidade da distribuição t-Student se assemelha à forma da função de densidade de uma variável aleatória normal

[6] A Figura 6.5 apresenta histogramas ilustrando a imprecisão na estimação da variância populacional.

[7] Note que queremos mostrar que $\frac{\hat{\mu}-\mu_0}{\hat{\sigma}/\sqrt{n}}$ tem distribuição t-Student com $n-1$ graus de liberdade. Esse resultado pode ser encontrado a partir dos seguintes passos: (1) Primeiro precisamos mostrar que a média $\hat{\mu}$ e a variância $\hat{\sigma}^2$ são variáveis aleatórias independentes. (2) Já sabemos da seção anterior que $\sqrt{n}(\hat{\mu} - \mu_0)/\sigma$ é uma variável aleatória normal padrão. (3) Precisamos mostrar que $\hat{\sigma}/\sigma$ é uma variável aleatória $\sqrt{Y/(n-1)}$, onde Y tem distribuição qui-quadrada com $n-1$ graus de liberdade. (4) A razão entre $\sqrt{n}(\hat{\mu} - \mu_0)/\sigma$ e $\hat{\sigma}/\sigma$ tem distribuição t-Student, como vimos no Exemplo 4.23. Todos essas etapas estão disponíveis por exemplo em Bierens (2004).

padronizada; a diferença principal está nas caudas, uma vez que a distribuição t-Student possui caudas mais pesadas do que a distribuição normal padronizada. À medida que o número de graus de liberdade aumenta, a função de densidade de uma variável aleatória t-Student se aproxima da função de densidade de uma variável aleatória normal padronizada. De fato, quando $\nu \to \infty$, a densidade de probabilidade da t-Student converge para a densidade de probabilidade de uma normal padronizada.

Diante da discussão acima, podemos então resumir a estatística teste Z e sua distribuição, usando a expressão compacta

$$Z = \frac{\hat{\mu} - \mu_0}{\hat{\sigma}_{\hat{\mu}}} \sim t_{n-1}.$$

Sob essa nova distribuição para a estatística teste, o que muda então na especificação das regras de rejeição ou aceitação da hipótese nula, em detrimento da hipótese alternativa? A diferença em relação à situação onde a variância era conhecida está na determinação dos valores de corte ou **valores críticos** para as regras dos testes de hipóteses. Para um nível de confiança de 5%, temos que encontrar um valor crítico $c_{5\%}$, para o qual a probabilidade de Z ser maior que ele seja igual a 5% (supondo que a hipótese nula seja verdadeira). De acordo com uma tabela da distribuição t-Student, com $n - 1 = 99$ graus de liberdade, o valor crítico para 5% é igual a $c_{5\%} = 1.66$, e o valor crítico para 1% é igual a $c_{1\%} = 2.36$.

Vamos então aplicar a regra de rejeição ou aceitação da hipótese nula aos valores numéricos encontrados a partir da amostra. A estimativa para o desvio padrão da distribuição do estimador $\hat{\mu}$ é igual a

$$\hat{\sigma}_{\hat{\mu}} = \sqrt{\frac{\hat{\sigma}^2}{n}} = \sqrt{\frac{0.0221}{100}} = 0.0149.$$

A estatística teste Z tem valor

$$Z = \frac{\hat{\mu} - \mu_0}{\hat{\sigma}_{\hat{\mu}}} = \frac{0.023}{0.0149} = 1.5471.$$

Portanto, dado que 1.5471 é menor do que o valor crítico $c_{5\%}$ e o valor crítico $c_{1\%}$, não podemos rejeitar a hipótese nula de que as notas médias da população no exame padronizado são menores ou iguais a 70% de aproveitamento. Portanto, a amostra coletada não fornece evidências suficientes para suportar o novo programa de melhorias educacionais no ensino médio público na região metropolitana de Salvador.

Exemplo 6.6 (Continuação do Exemplo 6.3 – Agora supondo variância desconhecida) Revisitando o Exemplo 6.3, vamos agora supor que a variância populacional σ^2 é desconhecida, e que temos que estimá-la a partir de uma amostra de tamanho n. Utilizando o estimador não viesado apresentado na Eq. (6.4), encontramos uma estimativa $\hat{\sigma}^2 = (\text{R\$ } 14.310)^2$; portanto, a estimativa para o desvio padrão populacional é $\hat{\sigma} = \text{R\$ } 14.310$. A estimativa para a média populacional continua sendo $\hat{\mu} = \text{R\$ } 146$ mil. A partir da estimativa para o desvio padrão populacional, podemos estimar o desvio padrão do estimador $\hat{\mu}$, que é dado por

$$\hat{\sigma}_{\hat{\mu}} = \sqrt{\frac{\hat{\sigma}^2}{n}} = \frac{14.310}{8} = R\$1.788,8.$$

A estatística teste Z terá valor

$$Z = \frac{\hat{\mu} - \mu_0}{\hat{\sigma}_{\hat{\mu}}} = \frac{146.000 - 150.000}{1.788.8} = -2.2362.$$

Supondo que a hipótese nula é verdadeira, sabemos que a estatística teste Z tem distribuição t-Student com $n - 1 = 63$ graus de liberdade. Os valores críticos para níveis de significância de 5% e 1% são $c_{5\%}$ = -1.67 e $c_{1\%}$ = -2.39, respectivamente. A estatística teste Z é menor do que o valor crítico para um nível de 5%, mas não é menor do que o valor crítico para um nível de 1%. Portanto, podemos rejeitar a hipótese nula se o nível considerado for 5%, mas não podemos rejeitar a hipótese nula para um nível de significância de 1%.

Exemplo 6.7 (Continuação do Exemplo 6.4 – Agora supondo a variância desconhecida) Para o Exemplo 6.4, consideraremos agora que a variância populacional é desconhecida, e foi estimada, utilizando-se a Eq. (6.4) em $\hat{\sigma}^2 = (1.37 \text{ mm})^2$, e, portanto, a estimativa para o desvio padrão populacional é $\hat{\sigma}^2 = 1.37$ mm. A estatística teste possui valor

$$Z = \frac{\hat{\mu} - \mu_0}{\hat{\sigma}_{\hat{\mu}}} = \frac{1.117 - 1.2}{\frac{1.37}{20}} = -1.2117.$$

Vamos agora especificar as regras de rejeição ou aceitação da hipótese nula. Para uma distribuição t-Student com $n-1 = 399$ graus de liberdade, os valores críticos, para um teste bicaudal, com nível de significância de 5%, são -1.97 e 1.97. Para um teste bicaudal, com nível de significância de 1%, os valores críticos são -2.59 e 2.59. Portanto, para um nível de significância de 1%, rejeitamos a hipótese nula quando $Z < -2.59$ ou $Z > 2.59$ (equivalentemente, podemos estabelecer a regra como: rejeitamos a hipótese nula quando $|Z| > 2.59$). Adicionalmente, para um nível de significância de 5%, rejeitamos a hipótese nula quando $Z < -1.97$ ou $Z > 1.97$ (equivalentemente, podemos estabelecer a regra como: rejeitamos a hipótese nula quando $|Z| > 1.97$). O valor da estatística teste Z = -1.2117 cai fora das regiões de rejeição para os dois níveis de significância, e não é possível rejeitar a hipótese nula.

População qualquer com variância desconhecida

Nas últimas duas seções, trabalhamos com a hipótese de que as amostras coletadas vinham de uma população com distribuição normal, com média μ e variância σ^2. Inicialmente, supusemos que a variância populacional era conhecida, e depois relaxamos essa hipótese para o caso onde tínhamos que estimar essa variância a partir dos dados amostrais. Nesta seção relaxaremos mais uma hipótese, em relação ao modelo inicial: a hipótese de normalidade. Portanto, a amostra $X_1, \ldots, X_n$ não mais vai ser suposta como gerada a partir de uma população com distribuição normal. Na verdade, não faremos hipótese alguma a respeito da distribuição da população; nem mesmo se a população é discreta ou contínua. A única hipótese de fato que será suposta aqui é que a distribuição populacional possui segundo momento finito; ou seja, possui

variância (conforme vimos no Capítulo 3, se uma variável aleatória possui segundo momento finito, ela também possui primeiro momento finito).

A principal consequência de relaxarmos a hipótese de a distribuição da população ser normal é que agora o estimador da média $\hat{\mu}$ não é mais uma média de variáveis aleatórias normais, e, portanto, não possui distribuição normal. O que fazer então, já que não conhecemos a distribuição exata do estimador $\hat{\mu}$? Como efetuar testes de hipóteses a respeito do parâmetro populacional μ? Felizmente, graças ao teorema central do limite (Teorema 6.3), vimos que a distribuição da média amostral converge para uma distribuição normal, à medida que a amostra aumenta. Podemos então nos basear nesse resultado para construir testes de hipóteses, no caso mais geral onde a distribuição da população não é mais normal. Para isso, vamos utilizar o resultado abaixo.

Nota 6.1 Seja $X_1, \ldots, X_n$ uma amostra aleatória, de observações independentes e identicamente distribuídas. A distribuição das variáveis aleatórias é arbitrária, valendo apenas que $\text{Var}[X_i] = \sigma^2 < +\infty$ para todos os $i = 1, 2, \ldots, n$. Portanto, a população tem segundo momento finito. A média populacional é dada por μ. Seja $\hat{\mu} = \bar{X}$ o estimador da média μ. Conforme vimos anteriormente, a média amostral é um estimador não viesado para a média populacional, uma vez que $\text{E}[\hat{\mu}] = \mu$. Pode-se mostrar que, quando n tende para o infinito, temos

$$\frac{\hat{\mu} - \mu}{\frac{\hat{\sigma}}{\sqrt{n}}} \xrightarrow{\text{L}} \text{Normal}[0, 1], \tag{6.6}$$

onde $\hat{\sigma}$ é o estimador do desvio padrão populacional σ; ou seja,

$$\hat{\sigma} = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} [X_i - \bar{X}]^2}.$$

Se utilizarmos o denominador n ao invés de $n - 1$ na fórmula para o estimador da variância (e do desvio padrão), a aproximação assintótica na Eq. (6.6) não se altera. A grandeza $\hat{\sigma}/\sqrt{n}$ corresponde ao estimador do desvio padrão especificamente para a distribuição do estimador $\hat{\mu}$. Portanto,

$$\hat{\sigma}_{\hat{\mu}} = \frac{\hat{\sigma}}{\sqrt{n}} = \frac{1}{\sqrt{n}} \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} [X_i - \bar{X}]^2}.$$

Exemplo 6.8 (Continuação do Exemplo 6.3 – Sem a hipótese de normalidade e supondo a variância desconhecida) Consideremos novamente o Exemplo 6.3, onde agora não podemos supor a hipótese de normalidade das observações amostrais $X_1, \ldots, X_n$. A variância populacional desconhecida foi estimada em $\hat{\sigma}^2 = (\text{R\$ } 14.310)^2$. A média amostral é igual a $\hat{\mu} = \text{R\$ } 146$ mil. Queremos testar a hipótese nula $H_0 : \mu \geq \mu_0$, contra a hipótese alternativa $H_A : \mu < \mu_0$, onde $\mu_0 = \text{R\$ } 150$ mil. Para isso, precisamos definir as regras de rejeição ou aceitação da hipótese nula, com base na estatística teste e na distribuição da estatística teste. Como definir então a estatística teste e sua distribuição, se não conhecemos a distribuição das observações da amostra? A resposta a essa pergunta é utilizar a aproximação dada na Eq. (6.6).

Portanto, a estatística teste é simplesmente

$$Z = \frac{\hat{\mu} - \mu_0}{\hat{\sigma}_{\hat{\mu}}}.$$

Se a hipótese nula estiver correta, a estatística teste tem distribuição aproximadamente igual à distribuição normal padronizada, e essa aproximação torna-se mais adequada à medida que o tamanho da amostra aumenta. Com base nesse fato, podemos utilizar **valores críticos aproximados** a partir da tabela da distribuição normal padronizada. Portanto, para um nível de significância de 5%, o valor crítico será $c_{5\%}$ = -1.645; para um nível de significância de 1%, o valor crítico será $c_{1\%}$ = -2.326. O valor da estatística teste com base nos dados obtidos na amostra de $n = 64$ observações é igual a

$$z = -2.2362,$$

conforme cálculos efetuados na seção anterior, para o Exemplo 6.3. Dessa forma, de acordo com os valores críticos aproximados, podemos rejeitar a hipótese nula com um nível de significância de 5%, mas não podemos rejeitar a hipótese nula para um nível de significância de 1%.

Portanto, quando não conhecemos a distribuição das observações coletadas na amostra aleatória, podemos proceder com os testes de hipótese, utilizando como estatística teste a estatística Z, e como distribuição da estatística teste a distribuição normal padronizada. A distribuição normal padronizada é uma aproximação assintótica, e torna-se mais adequada à medida que a o tamanho da amostra n aumenta.[8] Para n com valor baixo, não necessariamente a aproximação normal é adequada. Quando isso acontece, pode-se recorrer a aproximações de ordens maiores para refinar os valores críticos aproximados utilizados nos testes. Essas aproximações não estão no escopo deste livro, mas podem ser encontradas em Severini (2001).

6.4.2 Testes de hipóteses e estimação via máxima verossimilhança

Vamos agora discutir um tópico extremamente importante em econometria e estatística: testes de hipótese no contexto de estimação via máxima verossimilhança. Para isso, iremos utilizar exemplos simples, considerando-se tanto variáveis aleatórias discretas, quanto variáveis aleatórias contínuas. Um fato importante para os testes de hipótese nesse contexto é que a distribuição exata da estatística teste raramente é conhecida, e temos que recorrer a aproximações assintóticas. Normalmente, essas aproximações utilizam a distribuição normal padronizada ou a distribuição qui-quadrada. Por meio dos exemplos simples, estudaremos os três tipos básicos de testes: testes de Wald; testes de razão de verossimilhança; testes dos multiplicadores Lagrange.

[8] Por exemplo, a distribuição t-Student se aproxima da distribuição normal para n próximo de 30.

Testando parâmetros individualmente

Nessa subseção começamos com um objetivo mais simples apresentando exemplos onde o intuito é testar os parâmetros estimados individualmente (um a um). Na próxima subseção, consideramos o caso onde existe uma função que define o teste dos parâmetros de interesse e esses parâmetros podem ser testados conjuntamente.

Exemplo 6.9 (Testando hipóteses em uma amostra de variáveis aleatórias coletada de uma população com distribuição exponencial negativa) Considere uma amostra aleatória $X_1, \ldots, X_n$, com observações independentes e identicamente distribuídas, extraídas a partir de uma população com distribuição exponencial negativa, com parâmetro livre λ. Conforme vimos na Seção 3.3.2, a função densidade de probabilidade de cada observação X_i é dada por

$$f(x) = \lambda e^{-\lambda x}, \text{ para } x > 0.$$

O estimador de máxima verossimilhança de λ é simplesmente

$$\hat{\lambda} = 1/\bar{X}. \tag{6.7}$$

Vimos no Exemplo 5.11 que esse estimador é viesado, mas mostramos na Seção 5.4 que ele é consistente.

Na seção anterior, estávamos interessados em testar hipóteses a respeito da média populacional μ. Neste exemplo, iremos testar hipóteses a respeito do parâmetro λ. Da mesma maneira que no caso dos testes de hipótese para a média populacional μ, podemos ter três tipos de hipótese nula para o parâmetro λ. O teste bicaudal consiste em testar

$$H_0 : \lambda = \lambda_0 \text{ versus } H_A : \lambda \neq \lambda_0.$$

Os testes monocaudais consistem em testar

$$H_0 : \lambda \geq \lambda_0 \text{ versus } H_A : \lambda < \lambda_0,$$

ou

$$H_0 : \lambda \leq \lambda_0 \text{ versus } H_A : \lambda > \lambda_0.$$

Vamos então descrever os procedimentos para testar esses tipos de testes de hipóteses a respeito do parâmetro λ. Primeiramente, precisamos definir uma estatística teste, e sua correspondente distribuição

(exata ou aproximada). A estatística teste que iremos utilizar é a estatística Z, com expressão

$$Z = \frac{\hat{\lambda} - \lambda_0}{\hat{\sigma}_{\hat{\lambda}}}. \tag{6.8}$$

onde $\hat{\sigma}_{\hat{\lambda}}$ é o estimador do desvio padrão do estimador $\hat{\lambda}$. Para encontrar o estimador do desvio padrão, precisamos encontrar a variância do estimador de máxima verossimilhança. Conforme vimos no Capítulo 5, para encontrar a variância do estimador de máxima verossimilhança, é preciso encontrar a matriz hessiana da função de log-verossimilhança $l(\theta)$. No caso do modelo para uma distribuição exponencial negativa, onde há apenas um parâmetro livre, a matriz hessiana corresponde simplesmente à segunda derivada da função de log-verossimilhança. A partir da função densidade de probabilidade para o modelo exponencial negativo, podemos escrever a função de log-verossimilhança para uma amostra $X_1, \ldots, X_n$ independente e identicamente distribuída. Dessa forma, como vimos no Exemplo 5.14,

$$\log L(\theta) = l(\lambda) = \sum_{i=1}^{n} \log \lambda - \lambda \sum_{i=1}^{n} X_i = n \log \lambda - \lambda \sum_{i=1}^{n} X_i.$$

A primeira derivada será

$$\frac{d}{d\lambda} l(\lambda) = \frac{n}{\lambda} - \sum_{i=1}^{n} X_i,$$

enquanto a segunda derivada tem expressão

$$\frac{d^2}{d\lambda^2} l(\lambda) = -\frac{n}{\lambda^2}.$$

O **coeficiente de informação de Fisher**, que corresponde ao negativo do valor esperado da segunda derivada da função de log-verossimilhança, tem fórmula

$$I(\lambda) = -\mathrm{E}\left[\frac{d^2}{d\lambda^2} l(\lambda)\right] = \frac{n}{\lambda^2}.$$

Finalmente, a variância do estimador de máxima verossimilhança $\hat{\lambda}$ é igual ao inverso do coeficiente de informação de Fisher,

$$\mathrm{Var}\left[\hat{\lambda}\right] = \frac{1}{I(\lambda)} = \frac{\lambda^2}{n}.$$

Note que a variância do estimador de máxima verossimilhança $\hat{\lambda}$ depende do parâmetro verdadeiro λ, que é desconhecido. No entanto, como vimos na Nota 5.1, para estimar a variância $\mathrm{Var}\left[\hat{\lambda}\right]$, podemos utilizar o

valor estimado $\hat{\lambda}$ no lugar de λ na expressão acima. Portanto, o estimador para a variância de $\hat{\lambda}$ é dado por

$$\hat{\sigma}^2_{\hat{\lambda}} = \frac{\hat{\lambda}^2}{n},$$

e o desvio padrão tem estimativa

$$\hat{\sigma}_{\hat{\lambda}} = \frac{\hat{\lambda}}{\sqrt{n}}.$$

Podemos então reescrever a Eq. (6.8) para a estatística teste Z, obtendo

$$Z = \frac{\hat{\lambda} - \lambda_0}{\sigma_{\hat{\lambda}}} = \frac{\hat{\lambda} - \lambda_0}{\frac{\hat{\lambda}}{\sqrt{n}}}.$$

Pode-se mostrar que a estatística Z converge para uma distribuição normal padronizada, supondo que a hipótese nula seja verdadeira; ou seja, $Z \xrightarrow{\text{L}} \text{Normal}[0,1]$ quando $n \to \infty$. A distribuição exata da variável aleatória $\hat{\lambda}$ não é trivial de ser encontrada; no entanto, podemos utilizar a aproximação assintótica para aproximar a distribuição da estatística teste. Finalmente, a partir da estatística teste e da aproximação da estatística teste, precisamos definir as regras de rejeição ou aceitação da hipótese nula.

Para um teste bicaudal, com hipótese nula H_0: $\lambda = \lambda_0$, em detrimento da hipótese alternativa H_A: $\lambda \neq \lambda_0$, a regra de rejeição da hipótese nula é dada por:

(i) para um nível de significância de 1%, rejeitamos a hipótese nula quando o valor absoluto de Z for maior do que o valor crítico $c_{1\%} = 2.576$.

(ii) para um nível de significância de 5%, rejeitamos a hipótese nula quando o valor absoluto de Z for maior do que o valor crítico $c_{5\%} = 1.960$.

Para um teste *monocaudal*, com hipótese nula H_0: $\lambda \geq \lambda_0$, em detrimento da hipótese alternativa H_A: $\lambda < \lambda_0$, a regra de rejeição da hipótese é dada por:

(i) para um nível de significância de 1%, rejeitamos a hipótese nula quando o valor de Z for menor do que o valor crítico $c_{1\%} = -2.326$.

(ii) para um nível de significância de 5%, rejeitamos a hipótese nula quando o valor de Z for menor do que o valor crítico $c_{5\%} = -1.645$.

Finalmente, para um teste monocaudal, com hipótese nula H_0: $\lambda \leq \lambda_0$, em detrimento da hipótese alternativa H_A: $\lambda > \lambda_0$, a regra de rejeição da hipótese é dada por:

(i) para um nível de significância de 1%, rejeitamos a hipótese nula quando o valor de Z for maior do que o valor crítico $c_{1\%} = 2.326$.

(ii) para um nível de significância de 5%, rejeitamos a hipótese nula quando o valor de Z for maior do que o valor crítico $c_{5\%} = 1.645$.

Para outros níveis de significância, regras similares podem ser escritas, com base nos valores da tabela da normal padronizada. Alguns programas estatísticos e econométricos utilizam também valores críticos para o nível de significância de 10%. Alternativamente, esses programas apresentam o p-valor, conforme será discutido na Seção 6.4.3.

Exemplo 6.10 (Testando hipóteses em uma amostra de variáveis aleatórias coletada de uma população com distribuição geométrica) Consideremos agora uma amostra aleatória, com observações discretas, $X_1, \ldots, X_n$, coletadas de uma população com distribuição geométrica. Como vimos na Seção 3.2.4, a função de frequência para uma variável aleatória geométrica tem expressão

$$f(x) = p(1-p)^x, \text{ para } x \in \{0, 1, 2, \ldots\}.$$

A variável aleatória geométrica possui apenas um parâmetro livre p, que varia no intervalo $(0, 1)$. Para os valores $X_1, \ldots, X_n$ da amostra aleatória independente e identicamente distribuída, como vimos no Exemplo 5.7, a função de log-verossimilhança é dada por

$$\log L(p) = l(p) = n \log p + \log(1-p) \sum_{i=1}^{n} X_i.$$

A primeira derivada da função $l(p)$ é igual a

$$\frac{d}{dp} l(p) = n\frac{1}{p} - \frac{1}{1-p} \sum_{i=1}^{n} X_i.$$

Igualando a primeira derivada a zero e isolando o parâmetro p, obtemos o estimador de máxima verossimilhança para o parâmetro livre p. Esse estimador tem expressão

$$\hat{p} = \frac{n}{n + \sum_{i=1}^{n} X_i} = \frac{1}{1 + \bar{X}}. \tag{6.9}$$

Vamos agora encontrar a estimativa para a variância, e, consequentemente, o desvio padrão, do estimador $\hat{p}$. A segunda derivada da função $l(p)$ é igual a

$$\frac{d^2}{dp^2} l(p) = -\frac{n}{p^2} - \frac{\sum_{i=1}^{n} X_i}{(1-p)^2}.$$

O coeficiente de informação de Fisher é igual a

$$I(p) = -\mathrm{E}\Big[-\frac{n}{p^2} - \frac{\sum_{i=1}^{n} X_i}{(1-p)^2}\Big] = \frac{n}{p^2} + \frac{\sum_{i=1}^{n} \mathrm{E}[X_i]}{(1-p)^2} = \frac{n}{p^2} + \frac{\sum_{i=1}^{n}(1-p)/p}{(1-p)^2} = \frac{n}{p^2(1-p)}.$$

Portanto, a variância e o desvio padrão do estimador $\hat{p}$ têm expressões

$$\sigma^2_{\hat{p}} = \frac{1}{I(p)} = \frac{p^2(1-p)}{n},$$
$$\sigma_{\hat{p}} = \frac{p\sqrt{1-p}}{\sqrt{n}}.$$

Note que a variância $\sigma^2_{\hat{p}}$ depende do parâmetro desconhecido p. Para estimar a variância $\sigma^2_{\hat{p}}$, pode-se substituir o valor de $\hat{p}$ no lugar de p. Portanto,

$$\hat{\sigma}^2_{\hat{p}} = \frac{\hat{p}^2(1-\hat{p})}{n},$$
$$\hat{\sigma}_{\hat{p}} = \frac{\hat{p}\sqrt{1-\hat{p}}}{\sqrt{n}}. \tag{6.10}$$

Finalmente, podemos construir a estatística teste para os testes de hipótese por meio da expressão

$$Z = \frac{\hat{p} - p_0}{\sigma_{\hat{p}}} = \frac{\hat{p} - p_0}{\frac{\hat{p}\sqrt{1-\hat{p}}}{\sqrt{n}}}.$$

Podemos utilizar Z para testar hipóteses bicaudais ou monocaudais a respeito do parâmetro p, da mesma maneira que no caso do parâmetro λ para a variável aleatória exponencial negativa no Exemplo 6.9. Quando a hipótese nula é verdadeira, a estatística Z converge em distribuição para uma variável aleatória normal padronizada.

Note que os procedimentos utilizados para os testes de hipóteses no contexto de estimação via máxima verossimilhança são os mesmos, tanto para variáveis aleatórias discretas como para variáveis aleatórias contínuas. Essa é uma das vantagens da utilização de máxima verossimilhança para estimação dos parâmetros livres: a replicabilidade dos procedimentos para diferentes situações. Conforme veremos nos Capítulos 8 e 9, quando tratarmos de modelos de regressão, a utilização de máxima verossimilhança para estimação dos coeficientes segue basicamente os mesmos passos utilizados para a estimação de parâmetros livres nos modelos mais simples abordados até agora.

Exemplo 6.11 (Testando hipóteses em uma amostra de variáveis aleatórias coletada de uma população com distribuição gamma) Os Exemplos 6.9 e 6.10 ilustram a aplicação de testes de hipóteses para testar pressupostos a respeito do único parâmetro livre. Neste exemplo, estenderemos os procedimentos para um caso onde há mais de um parâmetro livre. Nesse caso, ao invés de trabalharmos com o coeficiente

de informação de Fisher, teremos que calcular a **matriz de informação de Fisher**. Essa matriz será calculada como função da matriz hessiana da função de log-verossimilhança.

Consideremos então uma amostra aleatória $X_1, \ldots, X_n$, com observações independentes e identicamente distribuídas, coletadas de uma população com distribuição gamma, com parâmetros livres α e β. Conforme discutimos no Capítulo 5, não é possível escrever fórmulas fechadas para os estimadores dos parâmetros α e β, da mesma forma que foi feito para o caso das variáveis aleatórias geométrica e exponencial negativa. Portanto, para a variável aleatória gamma, não é possível escrever equações similares às Eqs. (6.7) e (6.9). Os estimadores de máxima verossimilhança para os parâmetros α e β devem ser calculados numericamente por meio de algoritmos de maximização não linear. Os softwares estatísticos e econométricos trazem essas rotinas de otimização implicitamente.

Para uma variável gamma, com parâmetros α e β, a função densidade de probabilidade possui expressão

$$f(y) = \frac{1}{\beta^{\alpha}\Gamma(\alpha)} y^{\alpha-1} e^{-y/\beta}, \quad \text{para } y \in (0, \infty).$$

Para uma amostra aleatória de tamanho n, $X_1, \ldots, X_n$, com observações independentes e identicamente distribuídas, e valores observados $x_1, \ldots, x_n$, a função de log-verossimilhança tem fórmula

$$\log L(\alpha, \beta) = l(\alpha, \beta) = -n \log \Gamma(\alpha) - n\alpha \log \beta + (\alpha - 1) \sum_{i=1}^{n} \log X_i - \frac{1}{\beta} \sum_{i=1}^{n} X_i.$$

As derivadas parciais de primeira ordem de $l(\alpha, \beta)$ são

$$\frac{\partial}{\partial \alpha} l(\alpha, \beta) = -n\Psi(\alpha) - n \log \beta + \sum_{i=1}^{n} \log X_i,$$

$$\frac{\partial}{\partial \beta} l(\alpha, \beta) = -\frac{n\alpha}{\beta} + \frac{1}{\beta^2} \sum_{i=1}^{n} X_i.$$

A função $\Psi(\alpha)$, também conhecida como função digamma, corresponde à primeira derivada da função $\log \Gamma(\alpha)$ com respeito a α. A derivada da função digamma é a função trigamma $\Psi'(\alpha)$, que corresponde à segunda derivada da função $\log \Gamma(\alpha)$ com respeito a α. Da mesma maneira que no caso da função gamma, para as funções digamma e trigamma, tabelas numéricas estão disponíveis na literatura. Além disso, diversos softwares matemáticos e estatísticos possuem funções implícitas que fornecem os valores para essas funções.

As derivadas parciais de segunda ordem de $l(\alpha, \beta)$ são

$$\begin{aligned}
\frac{\partial^2}{\partial \alpha^2} l(\alpha, \beta) &= -n\Psi'(\alpha), \\
\frac{\partial^2}{\partial \beta^2} l(\alpha, \beta) &= \frac{n\alpha}{\beta^2} - \frac{2}{\beta^3} \sum_{i=1}^{n} X_i, \\
\frac{\partial^2}{\partial \beta \partial \alpha} l(\alpha, \beta) &= \frac{\partial^2}{\partial \alpha \partial \beta} l(\alpha, \beta) = -\frac{n}{\beta}.
\end{aligned}$$

As derivadas $\frac{\partial^2}{\partial \alpha^2} l(\alpha, \beta)$, $\frac{\partial^2}{\partial \beta \partial \alpha} l(\alpha, \beta)$ e $\frac{\partial^2}{\partial \alpha \partial \beta} l(\alpha, \beta)$ são funções apenas dos parâmetros α e β; portanto, os valores esperados dessas derivadas são triviais. Para a derivada $\frac{\partial^2}{\partial \beta^2} l(\alpha, \beta)$, o valor esperado será

$$\begin{aligned}
\mathrm{E}\left[\frac{\partial^2}{\partial \beta^2} l(\alpha, \beta)\right] &= \mathrm{E}\left[\frac{n\alpha}{\beta^2} - \frac{2}{\beta^3} \sum_{i=1}^{n} X_i\right] = \frac{n\alpha}{\beta^2} - \frac{2}{\beta^3} \sum_{i=1}^{n} \mathrm{E}[X_i] \\
&= \frac{n\alpha}{\beta^2} - \frac{2}{\beta^3} \sum_{i=1}^{n} \alpha\beta = \frac{n\alpha}{\beta^2} - \frac{2n\alpha\beta}{\beta^3} \\
&= \frac{n\alpha}{\beta^2} - \frac{2n\alpha}{\beta^2} = -\frac{n\alpha}{\beta^2}.
\end{aligned}$$

Com base na Eq. (5.36), no Capítulo 5, a **matriz de informação de Fisher** $I(\theta) = I(\alpha, \beta)$ corresponde ao negativo do valor esperado da matriz hessiana, composta pelas segundas derivadas parciais da função de log-verossimilhança $l(\theta) = l(\alpha, \beta)$. O vetor de parâmetros θ corresponde a um vetor bidimensional, composto pelos componentes α e β; ou seja, $\theta = [\alpha, \ \beta]'$. Portanto,

$$\begin{aligned}
I(\theta) = I(\alpha, \beta) &= -\mathrm{E}\left[\frac{\partial^2}{\partial \theta \partial \theta'} l(\theta)\right] \\
&= -\mathrm{E}\begin{bmatrix} \frac{\partial^2}{\partial \theta_1^2} l(\theta) & \frac{\partial^2}{\partial \theta_1 \partial \theta_2} l(\theta) \\ \frac{\partial^2}{\partial \theta_2 \partial \theta_1} l(\theta) & \frac{\partial^2}{\partial \theta_2^2} l(\theta) \end{bmatrix} = -\mathrm{E}\begin{bmatrix} \frac{\partial^2}{\partial \alpha^2} l(\alpha, \beta) & \frac{\partial^2}{\partial \alpha \partial \beta} l(\alpha, \beta) \\ \frac{\partial^2}{\partial \beta \partial \alpha} l(\alpha, \beta) & \frac{\partial^2}{\partial \beta^2} l(\alpha, \beta) \end{bmatrix} \\
&= -\mathrm{E}\begin{bmatrix} -n\Psi'(\alpha) & -\frac{n}{\beta} \\ -\frac{n}{\beta} & \left[\frac{n\alpha}{\beta^2} - \frac{2}{\beta^3} \sum_{i=1}^{n} X_i\right] \end{bmatrix} = -\begin{bmatrix} \mathrm{E}\left[-n\Psi'(\alpha)\right] & \mathrm{E}\left[-\frac{n}{\beta}\right] \\ \mathrm{E}\left[-\frac{n}{\beta}\right] & \mathrm{E}\left[\frac{n\alpha}{\beta^2} - \frac{2}{\beta^3} \sum_{i=1}^{n} X_i\right] \end{bmatrix},
\end{aligned}$$

e chegamos à expressão para a matriz de informação de Fisher

$$I(\theta) = I(\alpha, \beta) = -\mathrm{E}\left[\frac{\partial^2}{\partial \theta \partial \theta'} l(\theta)\right] = \begin{bmatrix} n\Psi'(\alpha) & \frac{n}{\beta} \\ \frac{n}{\beta} & \frac{n\alpha}{\beta^2} \end{bmatrix}. \tag{6.11}$$

Note que a matriz $I(\theta)$ na Eq. (6.11) depende dos parâmetros livres (verdadeiros) α e β, que são desconhecidos. Usando a Nota 5.1, podemos estimar a matriz $I(\theta) = I(\alpha, \beta)$ substituindo os valores

estimados para α e β pelos valores desconhecidos, obtendo

$$I(\hat{\theta}) = I(\hat{\alpha}, \hat{\beta}) = \begin{bmatrix} n\Psi'(\hat{\alpha}) & \frac{n}{\hat{\beta}} \\ \frac{n}{\hat{\beta}} & \frac{n\hat{\alpha}}{\hat{\beta}^2} \end{bmatrix}. \tag{6.12}$$

A matriz $I(\hat{\theta})$ na Eq. (6.12) é denominada **matriz de informação de Fisher esperada**. Lembre-se que, para obter $I(\hat{\theta})$, foi necessário calcular o valor esperado da matriz hessiana, eliminando assim as observações $X_1, \ldots, X_n$ da expressão para $I(\theta) = I(\alpha, \beta)$. Alternativamente, é possível estimar a matriz de informação de Fisher utilizando a **matriz de informação de Fisher observada** $\hat{I}(\hat{\theta}) = \hat{I}(\hat{\alpha}, \hat{\beta})$, dada por

$$\hat{I}(\hat{\theta}) = \hat{I}(\hat{\alpha}, \hat{\beta}) = \begin{bmatrix} -n\Psi'(\hat{\alpha}) & -\frac{n}{\hat{\beta}} \\ -\frac{n}{\hat{\beta}} & \left[\frac{n\hat{\alpha}}{\hat{\beta}^2} - \frac{2}{\hat{\beta}^3}\sum_{i=1}^{n} X_i\right] \end{bmatrix}. \tag{6.13}$$

Note que a matriz observada $\hat{I}(\hat{\theta})$ (ao contrário da matriz esperada $I(\hat{\theta})$) possui valores de $X_1, \ldots, X_n$ na sua expressão (mais precisamente, no segundo termo da sua diagonal principal). A vantagem de se utilizar a matriz de informação observada, ao invés da matriz de informação esperada, é que a primeira não exige que calculemos os valores esperados das segundas derivadas da função de log-verossimilhança $l(\theta)$. Para modelos mais complexos, calcular a matriz de informação esperada pode ser complicado, e o analista pode preferir utilizar diretamente a matriz observada. Os softwares estatísticos e econométricos, comumente disponíveis, possuem rotinas implícitas que já calculam as matrizes de informação automaticamente para modelos estocásticos mais comuns. Alguns softwares permitem ao usuário escolher entre a matriz de informação esperada e a matriz de informação observada. De qualquer forma, caso o analista queira utilizar algum modelo específico, não existente no software disponível, ele pode empregar procedimentos similares aos procedimentos acima para estimar a matriz de informação de Fisher, e consequentemente chegar a estimativas para a variância do estimador de máxima verossimilhança.

Como discutido na Seção 5.6, para estimar a matriz de variância-covariância do estimador de máxima verossimilhança $\hat{\theta} = [\alpha, \beta]'$, basta inverter a matriz de informação de Fisher, observada ou esperada. Se o analista escolher utilizar a matriz de informação de Fisher esperada, por exemplo, a estimativa para a matriz de variância-covariância $\Sigma_{\hat{\theta}}$ do estimador de máxima verossimilhança será

$$\hat{\Sigma}_{\hat{\theta}} = I(\hat{\theta})^{-1} = I(\hat{\alpha}, \hat{\beta})^{-1} = \begin{bmatrix} n\Psi'(\hat{\alpha}) & \frac{n}{\hat{\beta}} \\ \frac{n}{\hat{\beta}} & \frac{n\hat{\alpha}}{\hat{\beta}^2} \end{bmatrix}^{-1} = \frac{1}{\det I(\hat{\theta})} \begin{bmatrix} \frac{n\hat{\alpha}}{\hat{\beta}^2} & -\frac{n}{\hat{\beta}} \\ -\frac{n}{\hat{\beta}} & n\Psi'(\hat{\alpha}) \end{bmatrix},$$

onde $\det I(\hat{\theta})$ corresponde ao determinante da matriz $I(\hat{\theta})$,

$$\det I(\hat{\theta}) = n\Psi'(\hat{\alpha})\frac{n\hat{\alpha}}{\hat{\beta}^2} - \frac{n^2}{\hat{\beta}^2} = \frac{n^2}{\hat{\beta}^2}\left[\hat{\alpha}\Psi'(\hat{\alpha}) - 1\right].$$

Obtemos finalmente

$$\hat{\Sigma}_{\hat{\theta}} = \frac{1}{n} \begin{bmatrix} \frac{\hat{\alpha}}{[\hat{\alpha}\Psi'(\hat{\alpha})-1]} & -\frac{\hat{\beta}}{[\hat{\alpha}\Psi'(\hat{\alpha})-1]} \\ -\frac{\hat{\beta}}{[\hat{\alpha}\Psi'(\hat{\alpha})-1]} & \frac{\hat{\beta}^2\Psi'(\hat{\alpha})}{[\hat{\alpha}\Psi'(\hat{\alpha})-1]} \end{bmatrix}.$$

Para testar hipóteses específicas sobre os parâmetros α e β individualmente, podemos calcular estatísticas Z específicas para cada parâmetro sendo testado. Por exemplo, se quizermos testar a hipótese de $\alpha = \alpha_0$, podemos utilizar a estatística teste

$$Z = \frac{\hat{\alpha} - \alpha_0}{\hat{\sigma}_{\hat{\alpha}}}.$$

A estimativa $\hat{\sigma}_{\hat{\alpha}}$ pode ser obtida diretamente da matriz de variância-covariância $\hat{\Sigma}_{\hat{\theta}}$. De fato, a estimativa $\hat{\sigma}^2_{\hat{\alpha}}$ para a variância do estimador de máxima verossimilhança $\hat{\alpha}$ para o parâmetro α corresponde ao primeiro elemento da diagonal principal da matriz $\hat{\Sigma}_{\hat{\theta}}$. Portanto,

$$\hat{\sigma}^2_{\hat{\alpha}} = \frac{\hat{\alpha}}{n[\hat{\alpha}\Psi'(\hat{\alpha}) - 1]},$$

e a estatística Z pode ser reescrita como

$$Z = \frac{\hat{\alpha} - \alpha_0}{\hat{\sigma}_{\hat{\alpha}}} = \frac{\hat{\alpha} - \alpha_0}{\sqrt{\frac{\hat{\alpha}}{n[\hat{\alpha}\Psi'(\hat{\alpha})-1]}}}.$$

Pode-se mostrar que, quando $n \to \infty$, essa estatística Z converge em distribuição para uma variável aleatória normal padronizada. Portanto, novamente podemos utilizar a tabela da normal padronizada para encontrar os valores críticos para as regras de rejeição ou aceitação da hipótese nula. Por exemplo, para um teste bicaudal, com H_0: $\alpha = \alpha_0$, rejeitamos H_0 quando $|Z| > 1.960$, para um nível de significância de 5%. Para um nível de significância de 1%, rejeitamos H_0 quando $|Z| > 2.576$.

Regras similares podem ser utilizadas para testar hipóteses a respeito do parâmetro β individualmente. Para isso, basta utilizar o segundo componente da matriz de variância covariância estimada $\hat{\Sigma}_{\hat{\theta}}$ como estimativa para a variância do estimador de máxima verossimilhança $\hat{\beta}$. Para estatística Z, no caso do parâmetro β, a aproximação normal padronizada também pode ser utilizada para obtermos os valores críticos aproximados. Na próxima seção, discutiremos mecanismos para testar hipóteses conjuntas a respeito dos parâmetros α e β simultaneamente.

Testando parâmetros conjuntamente

Até agora apresentamos técnicas para testar parâmetros individualmente em um modelo estatístico. Iniciamos nossa discussão com testes de hipótese para a média especificamente de distribuições normais e não normais. Em seguida apresentamos uma discussão para testes de hipóteses no contexto de estimação via máxima verossimilhança. Vamos agora continuar a nossa discussão apresentando algumas alternativas para testar hipóteses mais gerais, onde são feitas suposições conjuntas a respeito dos parâmetros de um modelo estatístico. Esses procedimentos são bastante utilizados em econometria, como veremos nos Capítulos 8 e 9.

Exemplo 6.12 (Continuação do Exemplo 6.11 – Testando parâmetros conjuntamente em uma amostra de variáveis aleatórias coletada de uma população com distribuição gamma) Para exemplificar a utilização de testes conjuntos dos parâmetros de um modelo estatístico, vamos supor o contexto de uma amostra aleatória $X_1, \ldots, X_n$, coletada de uma população com distribuição gamma, com parâmetros desconhecidos α e β, conforme apresentado no Exemplo 6.11 acima. Vamos supor que, ao invés de querermos testar hipóteses a respeito dos parâmetros α e β individualmente, estamos interessados em testar as hipóteses $\alpha = \alpha_0$ e $\beta = \beta_0$ conjuntamente. Nesse caso, se rejeitarmos a hipótese nula, pode ter acontecido que $\alpha \neq \alpha_0$ ou $\beta \neq \beta_0$. Portanto, representamos a hipótese nula como H_0: $\alpha = \alpha_0$ e $\beta = \beta_0$, e a hipótese alternativa como H_A: $\alpha \neq \alpha_0$ ou $\beta \neq \beta_0$.

Similarmente a todos os testes de hipóteses discutidos até agora, precisamos de três ítens fundamentais:

(A) uma estatística teste;
(B) a distribuição, ou uma aproximação da distribuição, para a estatística teste;
(C) regras de aceitação ou rejeição da hipótese nula.

Vamos então apresentar a seguir um resultado geral para testes de hipóteses, conjuntos ou não. Esse procedimento é conhecido como **teste de Wald**.

Para uma distribuição gamma, com parâmetros livres α e β, vamos considerar uma amostra aleatória $X_1, \ldots, X_n$, com observações independentes e identicamente distribuídas. Seja uma função $h : \Re^2 \mapsto \Re^r$, onde $r = 1$ ou $r = 2$. Vamos escrever então a hipótese nula como $h(\theta) = 0$, onde $\theta = [\alpha, \beta]'$ é o vetor de parâmetros livres. Portanto, no caso mais geral, temos

$$\begin{aligned} H_0 &: \ h(\theta) = 0 \\ H_A &: \ h(\theta) \neq 0. \end{aligned} \tag{6.14}$$

Por exemplo, podemos escolher

$$h(\theta) = \begin{bmatrix} \alpha - \alpha_0 \\ \beta - \beta_0 \end{bmatrix},$$

e a hipótese nula $h(\theta) = 0$ corresponde às duas hipóteses conjuntas $\alpha = \alpha_0$ e $\beta = \beta_0$. Uma outra função $h(\cdot)$ possível é $h(\theta) = \alpha - \alpha_0$. Nesse caso, $H_0 : h(\theta) = 0$ corresponde a $\alpha = \alpha_0$. Portanto, utilizando o formato $h(\theta) = 0$ para representar a hipótese nula, conseguimos representar uma série de situações. A estatística teste no teste de Wald tem expressão

$$W = h(\hat{\theta})' \left[\left[\frac{\partial h(\theta)}{\partial \theta'} \Big|_{\theta=\hat{\theta}} \right] \Sigma_{\hat{\theta}} \left[\frac{\partial h(\theta)'}{\partial \theta} \Big|_{\theta=\hat{\theta}} \right] \right]^{-1} h(\hat{\theta}). \tag{6.15}$$

Na expressão acima, o termo $\left[\frac{\partial h(\theta)}{\partial \theta'} \Big|_{\theta=\hat{\theta}} \right]$, pré-multiplicando a matriz de covariância $\Sigma_{\hat{\theta}}$, corresponde à matriz de derivadas parciais

$$\frac{\partial h(\theta)}{\partial \theta'} \Big|_{\theta=\hat{\theta}} = \begin{bmatrix} \frac{\partial h_1(\hat{\theta})}{\partial \theta_1} & \frac{\partial h_1(\hat{\theta})}{\partial \theta_2} & \cdots & \frac{\partial h_1(\hat{\theta})}{\partial \theta_K} \\ \frac{\partial h_2(\hat{\theta})}{\partial \theta_1} & \frac{\partial h_2(\hat{\theta})}{\partial \theta_2} & \cdots & \frac{\partial h_2(\hat{\theta})}{\partial \theta_K} \\ \cdots & \cdots & \cdots & \cdots \\ \frac{\partial h_r(\hat{\theta})}{\partial \theta_1} & \frac{\partial h_r(\hat{\theta})}{\partial \theta_2} & \cdots & \frac{\partial h_r(\hat{\theta})}{\partial \theta_K} \end{bmatrix},$$

onde $\frac{\partial h_i(\hat{\theta})}{\partial \theta_j}$ é a derivada parcial do i-ésimo componente do vetor $h(\theta)$ em relação ao j-ésimo componente do vetor θ, avaliada no ponto $\theta = \hat{\theta}$. Portanto, a matriz $\left[\frac{\partial h(\theta)}{\partial \theta'} \Big|_{\theta=\hat{\theta}} \right]$ tem dimensão $r \times K$, onde r é a dimensão do vetor $h(\theta)$ e K é o número de parâmetros individuais livres no vetor θ. Similarmente, a matrix $\left[\frac{\partial h(\theta)'}{\partial \theta} \Big|_{\theta=\hat{\theta}} \right]$ é a transposta da matriz anterior; ou seja,

$$\frac{\partial h(\theta)'}{\partial \theta} \Big|_{\theta=\hat{\theta}} = \begin{bmatrix} \frac{\partial h_1(\hat{\theta})}{\partial \theta_1} & \frac{\partial h_2(\hat{\theta})}{\partial \theta_1} & \cdots & \frac{\partial h_r(\hat{\theta})}{\partial \theta_1} \\ \frac{\partial h_1(\hat{\theta})}{\partial \theta_2} & \frac{\partial h_2(\hat{\theta})}{\partial \theta_2} & \cdots & \frac{\partial h_r(\hat{\theta})}{\partial \theta_2} \\ \cdots & \cdots & \cdots & \cdots \\ \frac{\partial h_1(\hat{\theta})}{\partial \theta_K} & \frac{\partial h_2(\hat{\theta})}{\partial \theta_K} & \cdots & \frac{\partial h_r(\hat{\theta})}{\partial \theta_K} \end{bmatrix}.$$

Uma vez definida a estatística teste, precisamos definir a distribuição da estatística teste. Pode-se mostrar que, quando o número de observações n na amostra tende para o infinito, a estatística W converge para uma variável aleatória com distribuição qui-quadrada com r graus de liberdade. Portanto, o número de graus de liberdade da distribuição qui-quadrada limite é igual à dimensão do vetor $h(\theta)$, ou, similarmente, ao número de restrições conjuntas sendo testadas. Por exemplo, no caso da distribuição gamma, onde estamos testando as duas restrições conjuntamente $\alpha = \alpha_0$ e $\beta = \beta_0$, o número de restrições é igual a dois. Portanto, o número de graus de liberdade r da distribuição qui-quadrada limite também será dois. Matematicamente, podemos escrever

$$W \xrightarrow{L} \chi^2_r. \tag{6.16}$$

Uma segunda classe de procedimentos para testar hipóteses conjuntas para restrições nos parâmetros é conhecida como **testes de razão de verossimilhança**[9] (AMEMIYA, 1985; BIERENS, 2004). A estatística teste nesse caso é dada por

$$\text{LRT} = -2\log\left[\frac{\max_{h(\theta)=0} L(\theta)}{\max L(\theta)}\right] = 2\left[\log L(\hat{\theta}) - \log L(\hat{\theta}^*)\right], \tag{6.17}$$

onde $\hat{\theta}^*$ corresponde à estimativa do parâmetro θ sujeita à restrição $h(\theta) = 0$. O valor de $\hat{\theta}^*$ pode ser calculado por meio de uma maximização da função $\log L(\theta)$ diretamente, sujeita à restrição $h(\theta) = 0$. Alternativamente, o valor de $\hat{\theta}^*$ pode ser obtido por meio de uma reparametrização do modelo estatístico. Nesse caso, seja $q = K - r$, onde K é o número original de parâmetros livres no vetor de parâmetros θ e r é o número de restrições. Portanto, q é o número de parâmetros livres restantes no modelo restrito.[10] Vamos supor que podemos reescrever $\theta = g(\eta)$, onde η é um novo vetor de parâmetros livres, com dimensão q, e $g(\cdot)$ é uma função $\Re^q \mapsto \Re^K$, tal que $h(g(\eta)) = 0$ para todo η no espaço paramétrico Θ_η correspondente. Portanto, podemos encontrar o valor de $\hat{\theta}^*$ maximizando irrestritamente a função de log-verossimilhança no parâmetro η. Ou seja, primeiro encontramos $\hat{\eta}$, onde

$$\hat{\eta} = \max_{\eta \in \Theta_\eta} \log L(g(\eta)).$$

Em seguida, encontramos $\hat{\theta}^* = h(\hat{\eta})$.

Pode-se mostrar que o valor da função de log-verossimilhança no ponto $\hat{\theta}^*$ é menor do que o valor da função de log-verossimilhança no ponto $\hat{\theta}$. Isso porque o valor de $\log L(\hat{\theta})$ corresponde ao valor máximo quando efetuamos uma maximização irrestrita, enquanto o valor de $\log L(\hat{\theta}^*)$ corresponde ao valor máximo quando efetuamos uma maximização restrita. Portanto, o valor da estatística LRT é sempre não negativo. Pode-se mostrar que, quando o tamanho da amostra vai para o infinito, a estatística teste LRT converge em distribuição para uma variável aleatória com distribuição qui-quadrada, também com r graus de liberdade. Em forma matemática, temos a distribuição assimptótica para a estatística teste

$$\text{LRT} \xrightarrow{L} \chi^2_r. \tag{6.18}$$

Finalmente, a terceira classe de procedimentos para testar hipóteses conjuntas no contexto de estimação via máxima verossimilhança é o teste de escore de Rao[11] (AMEMIYA, 1985), também conhecido como **teste dos multiplicadores de Lagrange**.[12] A estatística teste para o teste dos multiplicadores de Lagrange é

[9] Do inglês, *likelihood ratio tests* (LRT).

[10] Pode acontecer de $q = 0$. De fato, no caso de uma população com distribuição gamma, com parâmetros α e β, e restrições $\alpha = \alpha_0$ e $\beta = \beta_0$, temos $r = K = 2$ e $q = 0$.

[11] Do inglês, *Rao's score test.*

[12] Do inglês, *Lagrange multiplier test.*

dada por

$$\text{LM} = -\left[\left.\frac{\partial \log L(\theta)}{\partial \theta'}\right|_{\theta=\hat{\theta}^*}\right]\left[\left.\frac{\partial^2 \log L(\theta)}{\partial \theta \partial \theta'}\right|_{\theta=\hat{\theta}^*}\right]^{-1}\left[\left.\frac{\partial \log L(\theta)}{\partial \theta}\right|_{\theta=\hat{\theta}^*}\right]. \tag{6.19}$$

Analogamente aos testes de razão de verossimilhança e de Wald, a estatística teste LM também converge em distribuição para uma variável aleatória com distribuição qui-quadrada, com r graus de liberdade,

$$\text{LM} \xrightarrow{L} \chi^2_r. \tag{6.20}$$

Note que, para calcular o valor da estatística teste LM para o teste dos multiplicadores de Lagrange, basta estimar $\hat{\theta}^*$, que corresponde ao valor estimado para θ sujeito à restrição $h(\theta) = 0$. Para o teste de Wald, a estatística teste LRT depende tanto do valor de $\hat{\theta}$, estimado sem levar em conta a restrição $h(\theta) = 0$, quanto do valor de $\hat{\theta}^*$. Finalmente, a estatística teste para o teste de Wald depende apenas da estimativa $\hat{\theta}$ irrestrita.

Para os três procedimentos para testar hipóteses conjuntas, no contexto de estimação via máxima verossimilhança, as estatísticas testes convergem em distribuição para a mesma distribuição. Portanto, as regras de rejeição ou aceitação da hipótese nula, em detrimento da hipótese alternativa, são as mesmas. De fato, para uma distribuição qui-quadrada com um grau de liberdade ($r = 1$, havendo, portanto, apenas uma restrição sendo testada), o valor crítico para um nível de significância de $\alpha = 10\%$ é igual a $c_{10\%} = 2.7055$. Para $\alpha = 5\%$, o valor crítico é $c_{5\%} = 3.8415$, e para $\alpha = 1\%$, o valor crítico é $c_{1\%} = 6.6349$. Portanto, seja lá qual dos três testes acima estivermos utilizando, rejeitamos a hipótese nula $h(\theta) = 0$ quando a estatística teste for maior do que $c_{10\%}$, $c_{5\%}$ ou $c_{1\%}$, a depender do nível de significância α que estivermos considerando. Para um teste de hipóteses conjuntas, com $r = 3$, os valores críticos serão os percentis de uma variável aleatória qui-quadrada com 3 graus de liberdade. Nesse caso, os valores críticos serão $c_{10\%} = 6.2514$, $c_{5\%} = 7.8147$ e $c_{1\%} = 11.3449$.

6.4.3 P-valores

Até o momento neste capítulo, as regras de rejeição ou aceitação das hipóteses nulas sendo testadas, em detrimento das hipóteses alternativas, estão baseadas na comparação entre as estatísticas testes e os valores críticos. Esses valores críticos são definidos a partir dos níveis de significância considerados, do fato de os testes serem bi-caudais ou unicaudais, e da distribuição da estatística teste (mesmo que essa distribuição corresponda ao caso limite, onde n tende para o infinito). Portanto, para que o analista possa concluir que a hipótese nula é rejeitada ou aceita no teste de hipótese, ele terá que ter em mãos tanto a estatística teste quando os valores críticos para comparação. No entanto, existe uma maneira mais direta de o analista checar se a hipótese nula é rejeitada ou não, observando uma medida apenas. Essa medida, conhecida como **p-valor**, é apresentada na grande maioria dos programas estatísticos e econométricos, e permite que o analista possa concluir de imediato se a hipótese nula é rejeitada ou não. De forma simples, o p-valor

pode ser definido como a probabilidade de se obter uma estatística teste igual ou mais extrema que aquela observada em uma amostra considerando que a hipótese nula é verdadeira.

Para apresentar o cálculo por trás do p-valor, vamos revisitar alguns dos exemplos anteriores. Para os Exemplos 6.2, 6.3 e 6.4, onde estamos testando hipóteses a respeito do valor da média populacional μ, a estatística teste é dada por

$$Z = \frac{\hat{\mu} - \mu_0}{\hat{\sigma}_{\hat{\mu}}},$$

onde $\hat{\sigma}_{\hat{\mu}}$ é a estimativa do desvio padrão do estimador $\hat{\mu}$. Nesse caso, estamos supondo que a variância σ^2 da população é desconhecida e precisa ser estimada a partir da amostra disponível. Quando a população possui distribuição normal, a estatística teste Z possui distribuição t-Student, com $n-1$ graus de liberdade (n é o tamanho da amostra). Para o Exemplo 6.2, onde a hipótese nula é dada por $H_0: \mu \leq \mu_0$ e a hipótese alternativa é $H_A: \mu > \mu_0$, o p-valor é dado por

$$\text{p-valor} = 1 - F_{t_{(n-1)}}(z),$$

onde $F_{t_{(n-1)}}$ é a função distribuição acumulada para uma variável aleatória t-Student com $n-1$ graus de liberdade, e z corresponde ao valor da estatística teste Z, calculado a partir da amostra. Para o Exemplo 6.2, onde $z = 1.5471$, e $n = 100$, o valor da função de distribuição acumulada é 0.9375, e o p-valor é igual a 0.0625. Finalmente, para checar se a hipótese nula é rejeitada ou não, basta comparar o p-valor aos níveis de significância. Portanto, como o p-valor é menor do que 0.10, podemos rejeitar a hipótese nula para um nível de significância $\alpha = 10\%$. No entanto, 0.0625 é maior do que 0.05 e 0.01; portanto, não podemos rejeitar a hipótese nula para níveis de significância de 1% e 5%.

Para o Exemplo 6.3, onde a hipótese nula H_0: $\mu \geq \mu_0$ e a hipótese alternativa é H_A: $\mu < \mu_0$, o p-valor é calculado a partir da expressão

$$\text{p-valor} = F_{t_{(n-1)}}(z).$$

Para $n = 64$ e z = -2.2362, o p-valor é dado por 0.0144. Portanto, o p-valor é menor do que 5%, e é possível rejeitar a hipótese nula para $\alpha = 5\%$ ou 10%.

Para o Exemplo 6.4, temos um teste bicaudal, onde a hipótese nula é H_0: $\mu = \mu_0$ e a hipótese alternativa é H_A: $\mu \neq \mu_0$. Nesse caso, o p-valor é calculado utilizando-se a expressão

$$\text{p-valor} = 2 \times \left[1 - F_{t_{(n-1)}}(|z|)\right],$$

onde $|z|$ é o valor absoluto de z. Para $z = -1.2117$ e $n = 400$, o valor da função distribuição acumulada no ponto $|z| = 1.2117$ é $F_{t_{(399)}}(1.2117) = 0.8868$, e o p-valor é igual a 0.2263. Portanto, não é possível rejeitar a hipótese nula para níveis de significância α iguais a 10%, 5% ou 1%.

Quando a população não possui distribuição normal, a estatística teste Z não mais possui distribuição exata t-Student. No entanto, em uma grande quantidade de aplicações, a estatística teste converge para uma variável aleatória com distribuição normal padronizada. Nesse caso, o analista pode utilizar valores críticos aproximados, para especificar as regras de rejeição ou aceitação da hipótese nula. Da mesma forma, a distribuição normal padronizada pode ser utilizada também para se obter p-valores aproximados. Para o Exemplo 6.2, 6.3 e 6.4, respectivamente, os p-valores, com base na aproximação normal padronizada, são dados pelas expressões

$$\begin{aligned} &\text{p-valor} = 1 - \Phi(z), \ \text{ para } H_0: \ \mu \leq \mu_0, \\ &\text{p-valor} = \Phi(z), \ \text{ para } H_0: \ \mu \geq \mu_0, \\ &\text{p-valor} = 2 \times \left[1 - \Phi(|z|)\right], \ \text{ para } H_0: \ \mu = \mu_0, \end{aligned} \tag{6.21}$$

onde $\Phi(\cdot)$ é a função de distribuição acumulada para uma variável aleatória normal padronizada. Para os Exemplos 6.2, 6.3 e 6.4, quando não é possível supor normalidade das observações, podemos utilizar a distribuição normal padronizada para aproximar a estatística teste Z. Os p-valores serão então 0.0609, 0.0127 e 0.2256, para os Exemplos 6.2, 6.3 e 6.4 respectivamente. Note que as conclusões a respeito de rejeição ou não da hipótese nula, *vis-a-vis* a hipótese alternativa, não se alteram quando utilizamos a aproximação normal no lugar da distribuição t-Student. No entanto, os p-valores calculados com a aproximação normal são menores do que os p-valores com base na distribuição t-Student. Isso já era de se esperar, pois a distribuição t-Student possui caudas mais pesadas do que a distribuição normal.

No contexto de estimação via máxima verossimilhança, quando estamos testando parâmetros individualmente e a estatística teste utilizada é a estatística Z, conforme Exemplos 6.9 e 6.10, o cálculo do p-valor é efetuado utilizando expressões similares às fórmulas apresentadas na Eq. (8.28). Isso porque os p-valores também são calculados com base na aproximação assimptótica normal padronizada. Para hipóteses nulas, conjuntas ou não, testadas com base nos testes de Wald, dos multiplicadores de Lagrange ou da razão de verossimilhança, as estatísticas testes convergem não mais para uma variável aleatória com distribuição normal padronizada. A distribuição limite para esses três testes é a distribuição qui-quadrada, com número de graus de liberdade igual ao número de restrições. Portanto, o cálculo do p-valor deve ser efetuado com base na distribuição qui-quadrada correspondente. Portanto, para a hipótese nula $H_0: \ h(\theta) = 0$,

$$\begin{aligned} &\text{p-valor} = 1 - F_{\chi^2_r}(W), \ \text{ para o teste de Wald,} \\ &\text{p-valor} = 1 - F_{\chi^2_r}(\text{LRT}), \ \text{ para o teste da razão de verossimilhança,} \\ &\text{p-valor} = 1 - F_{\chi^2_r}(\text{LM}), \ \text{ para o teste dos multiplicadores de Lagrange,} \end{aligned} \tag{6.22}$$

onde $F_{\chi^2_r}(\cdot)$ é a função de distribuição acumulada para uma variável aleatória qui-quadrada com r graus de liberdade.

6.5 Intervalos de confiança

Na seção anterior, consideramos o problema de testar hipóteses a respeito de parâmetros populacionais a partir de valores observados na amostra. Nesta seção, iremos discutir um outro tópico importante na análise de inferência estatística: os **intervalos de confiança**. Os intervalos de confiança fornecem uma ideia da imprecisão (ou precisão) nas estimativas dos parâmetros populacionais a partir dos dados da amostra. É comum encontrar, em divulgações de pesquisas eleitorais, que um determinado candidato tem 46% das intenções de votos, com margem de erro de 2%. Essa margem de erro corresponde justamente aos intervalos de confiança, cuja ideia geral iremos apresentar nesta seção. Para tornar a discussão mais didática, novamente iremos recorrer a alguns dos exemplos apresentados na seção anterior.

Intervalos de confiança para a média populacional

Dando continuidade aos Exemplos 6.2 e 6.5, mostraremos como podemos calcular intervalos de confiança para a média estimada.

Exemplo 6.13 (Continuação do Exemplo 6.2 – Construção de intervalos de confiança) Consideremos novamente o Exemplo 6.2, onde queremos estimar a média populacional μ, correspondente à nota média em matemática para os alunos da rede pública de ensino médio, na região metropolitana de Salvador. Para estimar essa média populacional, obteve-se uma amostra de tamanho $n = 100$. A estimativa pontual obtida foi $\hat{\mu} = 72.3\%$. Mas será que é possível construir um intervalo, ao redor da estimativa pontual, que contenha, com algum grau de segurança, o valor verdadeiro da média populacional? A resposta é sim. Podemos construir os chamados **intervalos de confiança**, que contêm o valor verdadeiro do parâmetro populacional real (no caso, a média populacional μ) com um certo grau de confiança. Esse grau de confiança é representado pela **probabilidade de cobertura** do intervalo de confiança. Por costumes tribais, as probabilidades de cobertura comumente utilizadas são 90%, 95% e 99%. Obviamente, outras probabilidades de cobertura podem ser utilizadas, mas em geral utilizam-se esses três valores.

Vamos supor inicialmente que a população possui distribuição normal, com variância desconhecida σ^2. A estimativa para a variância populacional é dada por $\hat{\sigma}^2 = 2.21\%$. O intervalo de confiança nesse caso pode ser construído utilizando-se a expressão

$$\text{IC}_{95\%} = \left[\hat{\mu} - t_{(n-1),97.5\%} \times \hat{\sigma}_{\hat{\mu}},\ \hat{\mu} + t_{(n-1),97.5\%} \times \hat{\sigma}_{\hat{\mu}}\right],$$

onde $t_{(n-1),97.5\%}$ é o valor do percentil de 97.5% para uma distribuição t-Student com $n - 1$ graus de liberdade. Por exemplo, para $n = 100$, $t_{(n-1),97.5\%} = 1.9842$. A medida $\hat{\sigma}_{\hat{\mu}}$ corresponde à estimativa do

$$\hat{\sigma}_{\hat{\mu}} = \sqrt{\frac{1}{n(n-1)} \sum_{i=1}^{n} [X_i - \bar{X}]^2}.$$

Similarmente, os intervalos de confiança para probabilidades de cobertura de 90% e 99% são

$$\text{IC}_{90\%} = \left[\hat{\mu} - t_{(n-1),95\%} \times \hat{\sigma}_{\hat{\mu}},\ \hat{\mu} + t_{(n-1),95\%} \times \hat{\sigma}_{\hat{\mu}}\right],$$
$$\text{IC}_{99\%} = \left[\hat{\mu} - t_{(n-1),99.5\%} \times \hat{\sigma}_{\hat{\mu}},\ \hat{\mu} + t_{(n-1),99.5\%} \times \hat{\sigma}_{\hat{\mu}}\right].$$

Para o caso onde não é possível supor que a população possui distribuição normal, podemos utilizar a aproximação normal padronizada, supondo que n é grande o suficiente, tal que essa aproximação seja válida. Nesse caso, ao invés de utilizar os percentis a partir de uma distribuição t-Student, podemos utilizar os valores críticos com base em uma distribuição normal padronizada. Portanto, os intervalos de confiança para probabilidades de cobertura de 90%, 95% e 99% são

$$\text{IC}_{90\%} = \left[\hat{\mu} - \Phi^{-1}(0.95) \times \hat{\sigma}_{\hat{\mu}},\ \hat{\mu} + \Phi^{-1}(0.95) \times \hat{\sigma}_{\hat{\mu}}\right],$$
$$\text{IC}_{95\%} = \left[\hat{\mu} - \Phi^{-1}(0.975) \times \hat{\sigma}_{\hat{\mu}},\ \hat{\mu} + \Phi^{-1}(0.975) \times \hat{\sigma}_{\hat{\mu}}\right],$$
$$\text{IC}_{99\%} = \left[\hat{\mu} - \Phi^{-1}(0.995) \times \hat{\sigma}_{\hat{\mu}},\ \hat{\mu} + \Phi^{-1}(0.995) \times \hat{\sigma}_{\hat{\mu}}\right],$$

onde $\Phi^{-1}(\cdot)$ corresponde à função distribuição acumulada inversa. Portanto, $\Phi^{-1}(0.95) = 1.6449$, $\Phi^{-1}(0.975) = 1.9600$ e $\Phi^{-1}(0.995) = 2.5758$.

Estimação via máxima verossimilhança

Para vetores parâmetros estimados via máxima verossimilhança, a construção de intervalos de confiança, para cada parâmetro θ_i individualmente, segue a mesma lógica dos intervalos de confiança para a média populacional. Supondo que o tamanho n da amostra é suficientemente grande, de tal forma que a aproximação normal padronizada é válida para fins práticos, a expressão geral para o intervalo de confiança de 95%, por exemplo, é dada por

$$\text{IC}_{95\%} = \left[\hat{\theta}_i - \Phi^{-1}(0.975) \times \hat{\sigma}_{\hat{\theta}_i},\ \hat{\theta}_i + \Phi^{-1}(0.975) \times \hat{\sigma}_{\hat{\theta}_i}\right], \tag{6.23}$$

onde $\hat{\theta}_i$ é o i-ésimo componente do vetor estimado de parâmetros livres $\hat{\theta}$, e $\hat{\sigma}_{\hat{\theta}_i}$ corresponde à estimativa do desvio padrão[13] do estimador $\hat{\theta}_i$. A estimativa $\hat{\sigma}_{\hat{\theta}_i}$ pode ser obtida a partir da matriz de variância-covariância $\hat{\Sigma}_{\hat{\theta}}$ para o estimador $\hat{\theta}$ de máxima verossimilhança, conforme vimos na seção anterior. Nesse caso, podemos

[13] Lembrando que a estimativa do desvio padrão de um estimador é também conhecida como erro padrão.

fazer

$$\hat{\sigma}_{\hat{\theta}_i} = \hat{\Sigma}_{\hat{\theta}}^{i,i}.$$

onde $\hat{\Sigma}_{\hat{\theta}}^{i,i}$ é o i-ésimo elemento da diagonal principal da matriz $\hat{\Sigma}_{\hat{\theta}}$.

No caso de construção de intervalos de confiança para parâmetros estritamente positivos, como é o caso, por exemplo, do parâmetro λ de uma variável aleatória de Poisson, ou dos parâmetros α e β de uma variável aleatória gamma, intervalos de confiança construídos diretamente a partir da Eq. (6.23) podem ter limites inferiores negativos, o que faz pouco sentido. Uma solução é empregar transformações dos parâmetros dos modelos, tal que os novos parâmetros variem em toda a reta real. Intervalos de confiança são então calculados para esses novos parâmetros. Finalmente, intervalos de confiança para os parâmetros originais podem ser obtidos a partir de transformações dos intervalos de confiança para os parâmetros transformados. Apresentamos essa metodologia no exemplo abaixo.

Exemplo 6.14 (Construindo intervalos de confiança em uma amostra de variáveis aleatórias coletada de uma população com distribuição de Poisson) Suponhamos que queremos construir um intervalo de confiança para o parâmetro λ de uma variável aleatória de Poisson (vide Seção 3.2.3). Ao invés de estimar λ diretamente, podemos fazer uma transformação do tipo $\lambda = e^{\nu}$, onde $\nu \in \Re$. Note que, qualquer que seja o valor de ν, o parâmetro λ será sempre positivo, conforme queremos. O novo modelo estatístico terá função densidade de probabilidade com expressão

$$f(x) = \frac{e^{-e^{\nu}}(e^{\nu})^x}{x!} = \frac{e^{-e^{\nu}}e^{\nu x}}{x!}, \text{ para } x = 0, 1, 2, \ldots.$$

Com base na função densidade com base no novo parâmetro ν, no lugar de λ, podemos proceder normalmente com todos os passos na obtenção da estimativa de máxima verossimilhança $\hat{\nu}$ para ν. Com base no coeficiente de informação de Fisher, obtemos a estimativa $\hat{\sigma}_{\hat{\nu}}^2$ para a variância do estimador $\hat{\nu}$. O intervalo de confiança com probabilidade de cobertura de 95% para ν pode ser escrito como

$$\text{IC}_{95\%} = \left[\hat{\nu} - \Phi^{-1}(0.975) \times \hat{\sigma}_{\hat{\nu}},\ \hat{\nu} + \Phi^{-1}(0.975) \times \hat{\sigma}_{\hat{\nu}}\right].$$

Finalmente, a estimativa para o parâmetro original λ é dada por $\hat{\lambda} = e^{\hat{\nu}}$, e o intervalo de confiança com probabilidade de cobertura de 95% para o parâmetro λ pode ser construído com a expressão

$$\text{IC}_{95\%} = \left[e^{\hat{\nu} - \Phi^{-1}(0.975) \times \hat{\sigma}_{\hat{\nu}}},\ e^{\hat{\nu} + \Phi^{-1}(0.975) \times \hat{\sigma}_{\hat{\nu}}}\right].$$

Note que o intervalo de confiança acima para o parâmetro λ tem limites inferior e superior necessariamente positivos. Não temos mais um intervalo simétrico em torno da estimativa pontual $\hat{\lambda}$, como seria o caso, se utilizássemos a Eq. (6.23).

Finalmente, para ilustrar a discussão sobre intervalos de confiança, endereçando ao mesmo tempo a aproximação normal para a distribuição dos estimadores de máxima verossimilhança, apresentamos agora algumas simulações de Monte Carlo. Os exemplos considerados aqui referem-se a modelos paramétricos simples, estimados via máxima verossimilhança, onde as estimativas das variâncias dos estimadores de máxima verossimilhança são obtidos a partir da matriz de informação de Fisher observada. Inicialmente, consideraremos uma variável aleatória de Poisson. Em seguida, apresentaremos os resultados para variáveis aleatórias gamma, Weibull e lognormal.

Inicialmente, geramos uma amostra de tamanho $n = 5$, com observações independentes e identicamente distribuídas, de uma população com distribuição de Poisson, com parâmetro $\lambda = 2.5$. O nosso objetivo é construir um intervalo de confiança para o parâmetro λ, com probabilidade de cobertura de 90%. Após estimarmos o parâmetro λ, via máxima verossimilhança, obtivemos o intervalo de confiança $[2.0072, 4.8602]$. Repetimos esse mesmo exercício, simulando uma segunda amostra, com $n = 5$, e obtivemos um novo intervalo de confiança $[1.3848, 3.8885]$. Note que ambos os intervalos de confiança contêm o parâmetro verdadeiro $\lambda = 2.5$. Podemos então repetir esses passos, gerando amostras de tamanho $n = 5$, estimando o parâmetro λ, e calculando o intervalo de confiança com probabilidade de cobertura igual a 90%, um grande número de vezes. Replicamos esses passos 20 mil vezes, obtendo uma ideia da probabilidade real de cobertura do intervalo de confiança calculado. Os nosso objetivos com esse experimento de Monte Carlo são os seguintes:

(1) Mostrar que os intervalos de confiança também são variáveis aleatórias;
(2) Avaliar a aproximação normal para o cálculo dos intervalos de confiança;
(3) Avaliar o efeito do aumento do tamanho n das amostras sobre a eficácia da aproximação normal para o intervalo de confiança.

A Tabela 6.6 apresenta o percentual de vezes, nas 20 mil repetições nas simulações de Monte Carlo, em que o intervalo de confiança calculado continha o parâmetro verdadeiro (a ser estimado). Por exemplo, para o parâmetro λ da variável de Poisson, para amostras de tamanho $n = 5$, o intervalo de confiança com 90% de probabilidade de cobertura contiveram o parâmetro $\lambda = 2.5$ verdadeiro em 93.44% das replicações. Para $n = 10$, esse número baixou para 91.11%, e para $n = 100$, a probabilidade real de cobertura ficou em 89.84%. Lembremos que o valor nominal da probabilidade de cobertura é de 90%. Portanto, um maior tamanho da amostra implica em probabilidade de cobertura real mais próxima da probabilidade de cobertura nominal; isso indica que a aproximação normal torna-se cada vez mais apropriada à medida que o tamanho n da amostra aumenta (como já era de se esperar).

As Figuras 6.6 e 6.7 apresentam os histogramas para os limites inferior e superior calculados, para o parâmetro λ. Nesses histogramas, note que há limites inferiores maiores do que $\lambda = 2.5$ e limites superiores menores do que $\lambda = 2.5$. Quando uma dessas situações acontece, o intervalo de confiança correspondente não contém o parâmetro λ verdadeiro. Comparando os histogramas para $n = 5$ e $n = 100$, note que os

Tabela 6.6: Comparação entre a probabilidade real de cobertura e a probabilidade nominal de cobertura para parâmetros estimados via máxima verossimilhança.

Distribuição utilizada	Tamanho da amostra	Probabilidade de cobertura 90%	95%	99%
Poisson (Parâmetro λ)	$n = 5$	93.44%	96.70%	99.30%
	$n = 10$	91.11%	95.74%	99.46%
	$n = 20$	91.18%	95.05%	99.10%
	$n = 100$	89.84%	95.09%	98.90%
Gamma (Parâmetro α)	$n = 5$	78.23%	84.89%	92.35%
	$n = 10$	84.23%	90.41%	96.21%
	$n = 20$	87.26%	92.80%	97.78%
	$n = 100$	89.52%	94.95%	98.90%
Weibull (Parâmetro β)	$n = 5$	78.82%	84.96%	92.56%
	$n = 10$	84.44%	90.36%	96.37%
	$n = 20$	87.40%	92.12%	97.72%
	$n = 100$	89.75%	94.79%	98.66%
lognormal (Parâmetro σ)	$n = 5$	89.96%	95.23%	99.14%
	$n = 10$	90.11%	95.12%	98.96%
	$n = 20$	90.13%	94.98%	99.02%
	$n = 100$	89.98%	95.06%	99.08%

intervalos de confiança com $n = 100$ são bem mais estreitos em média do que os intervalos de confiança para $n = 5$. Além disso, a dispersão dos intervalos de confiança também é menor quando o tamanho n das amostras aumenta.

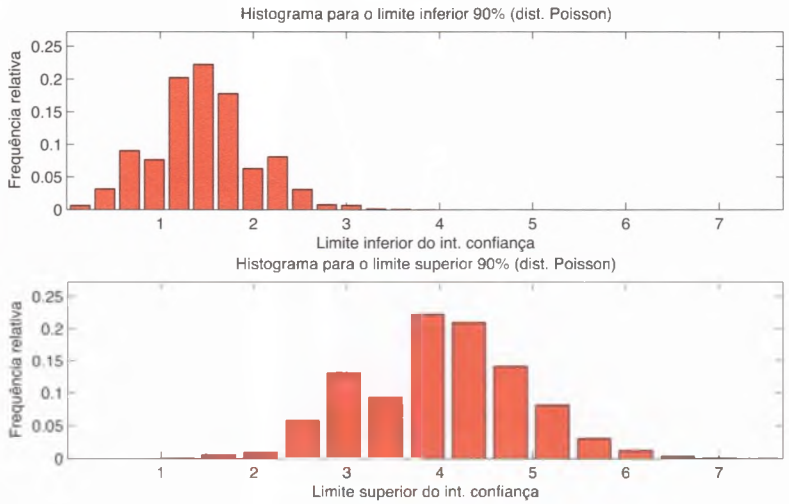


Figura 6.6: Histogramas dos limites superior e inferior para intervalos de confiança, com nível de cobertura de 90%, para o parâmetro λ de uma variável aleatória de Poisson. Os intervalos foram calculados com base em amostras aleatórias de tamanho $n = 5$.

Repetimos o mesmo experimento de Monte Carlo, agora considerando amostras geradas a partir de uma distribuição gamma, com parâmetros verdadeiros $\alpha = 1.4$ e $\beta = 10.6$. A Tabela 6.6 apresenta os resultados das probabilidades de cobertura reais, para intervalos de confiança calculados a partir da estimação via máxima verossimilhança. Para simplificar a apresentação dos resultados, mostramos na

tabela apenas os valores correspondentes aos intervalos de confiança para o parâmetro α. Os resultados correspondentes ao parâmetro β foram bastante similares. Note que para $n = 5$, as probabilidades reais de cobertura dos intervalos de confiança estão significativamente distantes das probabilidades nominais. Para uma probabilidade nominal de cobertura de 90%, a probabilidade real, de acordo com as simulações, foi de 78.23%. Isso indica que a aproximação normal, para $n = 5$, não parece ser muito razoável. À medida que o tamanho da amostra aumenta, a probabilidade real de cobertura aproxima-se da probabilidade nominal.

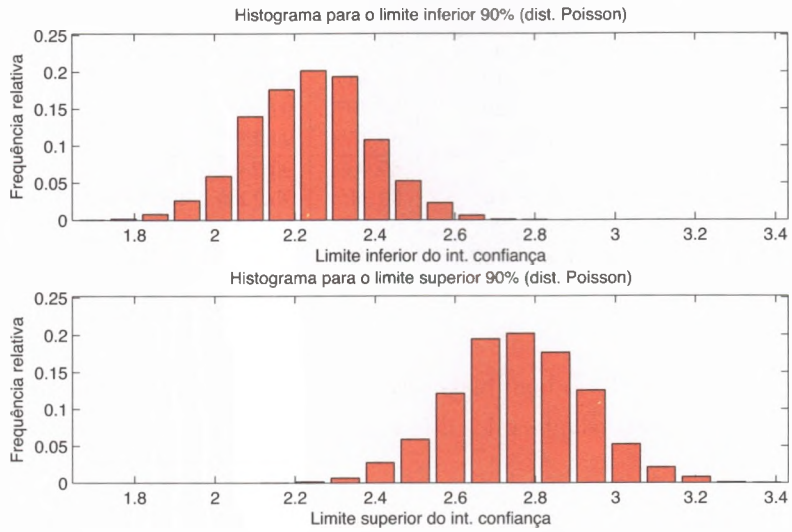


Figura 6.7: Histogramas dos limites superior e inferior para intervalos de confiança, com nível de cobertura de 90%, para o parâmetro λ de uma variável aleatória de Poisson. Os intervalos foram calculados com base em amostras aleatórias de tamanho $n = 100$.

Finalmente, replicamos o mesmo experimento de Monte Carlo para amostras aleatórias, independentes e identicamente distribuídas, geradas a partir de populações com distribuições de Weibull e lognormal. Em ambos os casos, utilizamos estimação via máxima verossimilhança, calculando em seguida os intervalos de confiança com probabilidades nominais de cobertura de 90%, 95% e 99%. Para a distribuição de Weibull, os valores verdadeiros para os parâmetros foram $\alpha = 3000$ e $\beta = 1.2$. Para a distribuição lognormal, os valores escolhidos para os parâmetros foram $\mu = 5.3$ e $\sigma = 2.5$. O número de replicações em cada experimento foi igual a 20 mil. Novamente, para facilitar a apresentação dos resultados, a tabela 6.6 mostra apenas os resultados para o parâmetro β, no caso da distribuição de Weibull, e σ, no caso da distribuição lognormal. No caso da variável aleatória de Weibull, para $n = 5$, os valores reais para a probabilidade de cobertura estão distantes dos valores nominais. À medida que o tamanho da amostra aumenta, a aproximação normal passa a ser mais razoável, implicando em intervalos de confiança mais apropriados para as probabilidades nominais de cobertura almejados. Já no caso da distribuição lognormal, mesmo para amostras com poucas observações, a aproximação normal pareceu bastante apropriada, no sentido de que os intervalos de confiança apresentaram probabilidades de cobertura reais muito próximas aos valores nominais.

O experimento de Monte Carlo apresentado nesta seção corrobora o fato de que os intervalos de confiança, com base nos estimadores de máxima verossimilhança, muitas vezes baseiam-se em aproximações para a distribuição dos estimadores dos parâmetros livres. Essas aproximações tornam-se melhores à medida que o tamanho da amostra aumenta. No entanto, mesmo para amostras pequenas, há maneiras de melhorar os intervalos de confiança, para que eles apresentem probabilidades reais de cobertura mais próximas das probabilidades nominais. Nesse caso, utilizam-se aproximações de maiores ordens. Esse tópico foge ao escopo deste livro, e o leitor interessado pode recorrer a referências como Severini (2001). Em todo caso, a maioria dos programas estatísticos utilizam as aproximações de primeira ordem, que correspondem às aproximações normais estudadas neste capítulo.

Outro fato importante de ser mencionado a partir dos experimentos aqui apresentados é a dependência da validade da aproximação de primeira ordem em relação ao tipo de modelo estatístico sendo estimado. Para o parâmetro da variável de Poisson, mesmo para $n = 5$, a aproximação normal pareceu bastante adequada. Para modelos mais complexos, como é o caso da distribuição gamma, os intervalos de confiança, calculados com pequenas amostras, não foram tão apropriados. Portanto, a partir de que tamanho n de amostras a aproximação normal se torna apropriada dependerá bastante do tipo de modelo estocástico sendo estimado. O número de parâmetros não necessariamente é o único indicador da qualidade da aproximação normal: note que, para a variável lognormal, que contém dois parâmetros, mesmo para $n = 5$, a aproximação normal mostrou-se bastante adequada, quando o objetivo era a construção de intervalos de confiança. Simulações de Monte Carlo são sempre uma boa maneira de estudar a validade das aproximações.

6.6 Exercícios

Exercício 6.1 Considere uma amostra aleatória, com observações independentes e identicamente distribuídas, com tamanho $n = 11$ e valores

$$1.32, 0.11, 2.1, 3.14, 4.91, 2.82, 1.64, 7.12, 5.12, 5.02, 0.87.$$

Supondo que a população é normal, com média μ e variância σ^2 desconhecidas, responda aos ítens abaixo.

(i) Teste a hipótese nula $H_0 : \mu = 4.1$, para níveis de significância de 1%, 5% e 10%.
(ii) Teste a hipótese nula $H_0 : \mu \leq 3.50$, para níveis de significância de 1%, 5% e 10%.
(iii) Teste a hipótese nula $H_0 : \mu \geq 5.50$, para níveis de significância de 1%, 5% e 10%.

Exercício 6.2 Para as observações no Exercício 6.1, determine os p-valores para os testes de hipóteses dos ítens (i), (ii) e (iii).

Exercício 6.3 Para as observações no Exercício 6.1, encontre intervalos de confiança, considerando-se probabilidades de cobertura de 90%, 95% e 99%.

Exercício 6.4 Repita o Exercício 6.2, supondo agora que a população não mais segue uma distribuição normal. Você poderá utilizar a aproximação normal padronizada.

Exercício 6.5 Repita o Exercício 6.3, supondo agora que a população não mais segue uma distribuição normal. Você poderá utilizar a aproximação normal padronizada.

7. Testes de ajuste, seleção e combinações de distribuições

"All models are wrong, but some models are useful."
George E. P. Box

Nos capítulos anteriores, descrevemos um conjunto de distribuições discretas e contínuas, e introduzimos dois métodos básicos para estimar os parâmetros livres dessas distribuições a partir de uma base de dados disponível. As variáveis aleatórias discretas podem ser utilizadas para modelar o comportamento da frequência de ocorrência de eventos de perdas operacionais em um determinado período, por exemplo. As variáveis aleatórias contínuas, por sua vez, podem ser utilizadas para modelar o comportamento das séries de severidades das perdas operacionais.

Na prática, para uma determinada amostra de frequência de perdas, não sabemos *a priori* qual a melhor distribuição de frequência a ser utilizada. A princípio, qualquer uma das distribuições discretas apresentadas neste documento podem ser usadas. Além disso, a literatura tem várias outras distribuições que poderiam servir aos nossos propósitos de modelagem. Portanto, existe a necessidade de utilizarmos algum critério de seleção para encontrarmos uma distribuição (ou um conjunto de distribuições) mais adequada. Esse assunto será tratado mais adiante na Seção 7.1.

Apesar de a Seção 7.1 apresentar critérios básicos para escolher o melhor, ou os melhores modelos, dentro de uma cesta de modelos disponíveis, em termos práticos, não necessariamente a melhor distribuição disponível apresenta um padrão aproximativo razoável dos dados que queremos modelar. Pode acontecer que várias distribuições, no conjunto disponível, sejam suficientemente próximas dos dados a serem modelados. Nas Seções 7.2 e 7.3, descreveremos métodos para avaliar o quão próximos, ou adequados, os modelos escolhidos estão dos dados disponíveis. Esses testes são conhecidos como testes de ajustes de modelos ou testes de ajustes das distribuições. Além dos testes de ajustes, apresentaremos alguns critérios gráficos que ajudarão os analistas na detecção de problemas na especificação dos modelos estatísticos. Inicialmente, discutiremos os testes de ajustes para distribuições contínuas, apresentando os critérios gráficos via distribuição empírica, QQ-plot e PP-plot, e o teste de Kolmogorov-Smirnov. Em seguida, trataremos dos testes de ajuste para distribuições discretas, onde será apresentado o teste qui-quadrado.

Em muitas situações práticas, pode acontecer de os modelos paramétricos disponíveis, seja para dados discretos, seja para dados contínuos, não satisfazerem aos testes de ajustes discutidos nas Seções 7.2 e 7.3. Nesse caso, o analista pode recorrer a outros softwares, que tenham disponíveis mais tipos de modelos estocásticos. Alternativamente, o analista pode tentar programar esses outros modelos. Mesmo assim, depois de tentar utilizar outras distribuições discretas ou contínuas, não necessariamente alguma dessas

irá satisfazer os critérios de ajustes. O que fazer nesse caso então? A sugestão apresentada na Seção 7.4 é utilizar combinações de modelos simples para compôr modelos estocásticos com mais poder de capturar uma variedade maior de especificidades dos dados empíricos. Neste capítulo, discutimos duas classes de modelos, tanto para dados contínuos quanto para dados discretos, que podem ser utilizados para prover mais flexibilidade aos analistas quantitativos. Com essas duas classes adicionais, será possível modelar praticamente todas as bases de frequências e severidades de perdas operacionais, por exemplo. A primeira classe corresponde à **mistura de distribuições** ou **distribuições combinadas**.[1] A segunda classe de modelos corresponde ao que chamamos de **distribuições por subintervalos**, pois ela consiste em dividir o espaço amostral em subintervalos, e ajustar uma distribuição específica a cada um desses subintervalos.

7.1 Critérios para seleção de modelos

Nos Capítulos 3 e 5, descrevemos um conjunto de distribuições discretas e contínuas, e introduzimos o método de momentos e o método de máxima verossimilhança para estimador os parâmetros livres dessas distribuições a partir de uma base de dados disponíveis. Na prática, para uma determinada amostra de dados discretos, não sabemos *a priori* qual a melhor distribuição a ser utilizada. A princípio, qualquer uma das distribuições discretas apresentadas neste documento pode ser usada. Além disso, a literatura tem várias outras distribuições que poderiam servir aos nossos propósitos de modelagem. Portanto, existe a necessidade de utilizarmos algum critério de seleção para encontrarmos uma distribuição (ou um conjunto de distribuições) mais adequada.

A literatura de seleção de modelos estatísticos é bem ampla, sendo que os critérios de seleção de modelos mais utilizados são o **critério de seleção de Akaike**, cuja sigla é **AIC**, e o **critério de seleção Bayesiano**, com sigla **BIC**.[2] Uma análise aprofundada desses dois critérios é dada por Burnham e Anderson (1998).

Os critérios AIC e BIC são derivados diretamente dos estimadores de máxima verossimilhança e podem ser calculados da seguinte forma: seja então uma amostra aleatória $X_1, \ldots, X_n$, para a qual iremos estimar os parâmetros livres de um modelo estocástico para variáveis discretas. No caso da distribuição de Poisson, usando as expressões do Exemplo 5.6, as expressões para o AIC e o BIC são

$$\begin{aligned} AIC &= -2\log L(\hat{\lambda}) + 2K = \left[n\hat{\lambda} - \log\hat{\lambda}\sum_{i=1}^{n} x_i + \sum_{i=1}^{n}\log x_i!\right] + 2\times 1 \\ BIC &= -2\log L(\hat{\lambda}) + K\log n = \left[n\hat{\lambda} - \log\hat{\lambda}\sum_{i=1}^{n} x_i + \sum_{i=1}^{n}\log x_i!\right] + 1\times\log n, \end{aligned} \tag{7.1}$$

[1] Essa classe de modelos é comumente conhecida como *mixture models*.

[2] Em inglês, o AIC é conhecido como *Akaike Information Criterion* e o BIC é conhecido como *Bayesian Information Criterion*.

onde K é o número de parâmetros livres; no caso da distribuição de Poisson temos $K = 1$. Note que a primeira parcela nos dois critérios acima é a mesma. e corresponde a menos duas vezes a função log-verossimilhança avaliada no ponto $\hat{\lambda}$ (estimativa de máxima verossimilhança de λ). A segunda parcela nas fórmulas acima é o que diferencia os dois critérios de seleção. O AIC tem segunda parcela igual a $2K$ e o BIC tem a segunda parcela igual a $K \log n$. Essa segunda parcela consiste em um fator para penalizar a inclusão de mais parâmetros livres no modelo. Em geral, se for possível ajustar modelos com menos parâmetros livres, melhores são os resultados finais, e tanto $2K$ no AIC quanto $K \log n$ no BIC têm o papel de forçar a escolha de modelos com menor número de parâmetros. Quando n for maior ou igual a 8, o que certamente é o caso em bases de dados utilizadas em risco operacional, a penalização dada pelo BIC é maior do que a penalização dada pelo AIC, pois $\log 8 > 2$.

Dada uma base de dados de frequências $X_1, X_2, \ldots, X_n$, depois de estimarmos a estimativa de máxima verossimilhança $\hat{\lambda}$ e calcularmos os critérios AIC e BIC, podemos proceder da mesma maneira com as demais distribuições de frequências disponíveis na nossa lista de variáveis aleatórias discretas disponíveis. Podemos então estimar a distribuição geométrica, por exemplo, obtendo a estimativa $\hat{p}$ do parâmetro livre. Em seguida, para compararmos o ajuste da distribuição geométrica com o ajuste da ditribuição de Poisson, usando os resultados do Exemplo 5.7, calculamos os critérios AIC e BIC

$$\begin{aligned} AIC &= -2 \log L(\hat{p}) + 2K = \left[-n \log \hat{p} - \log(1 - \hat{p}) \sum_{i=1}^{n} x_i \right] + 2 \times 1 \\ BIC &= -2 \log L(\hat{p}) + K \log n = \left[-n \log \hat{p} - \log(1 - \hat{p}) \sum_{i=1}^{n} x_i \right] + 1 \times \log n. \end{aligned} \tag{7.2}$$

Da mesma maneira, usando os resultados do Exemplo 5.8, estimamos os parâmetros r e p para a distribuição binomial negativa e obtemos as estimativas de máxima verossimilhança $\hat{r}$ e $\hat{p}$. Em seguida, calculamos os critérios AIC e BIC para essa distribuição

$$\begin{aligned} AIC &= -2 \log L(\hat{r}, \hat{p}) + 2K = \left[n \log \Gamma(\hat{r}) - \hat{r} n \log \hat{p} - \sum_{i=1}^{n} \log \Gamma(\hat{r} + x_i) \right. \\ &\left. + \sum_{i=1}^{n} \log \Gamma(1 + x_i) - \log(1 - \hat{p}) \sum_{i=1}^{n} x_i \right] + 2 \times 2 \\ BIC &= -2 \log L(\hat{r}, \hat{p}) + K \log n = \left[n \log \Gamma(\hat{r}) - \hat{r} n \log \hat{p} - \sum_{i=1}^{n} \log \Gamma(\hat{r} + x_i) \right. \\ &\left. + \sum_{i=1}^{n} \log \Gamma(1 + x_i) - \log(1 - \hat{p}) \sum_{i=1}^{n} x_i \right] + 2 \times \log n. \end{aligned} \tag{7.3}$$

Note agora que o número de parâmetros livres disponíveis K é igual a 2. Procedemos então dessa maneira com todas as distribuições discretas disponíveis na nossa lista. A distribuição utilizada será justamente aquela (ou aquelas) que apresentar o menor AIC ou o menor BIC. Em geral, os dois critérios costumam concordar, de forma que uma distribuição escolhida de acordo com o AIC também será escolhida de

acordo com o BIC. Quando há discordância entre ambos os critérios, o BIC seleciona distribuições mais parcimoniosas, ou seja, com menor número de parâmetros livres. Isso é decorrência da maior penalidade $K \log n$ dada pelo BIC ao uso de modelos com mais parâmetros livres.

O procedimento para selecionar a distribuição de severidade mais adequada é completamente análogo ao empregado para selecionar a distribuição de frequência mais apropriada. Seja então $Y_1, \ldots, Y_n$ uma amostra aleatória de observações contínuas. Para cada distribuição contínua na nossa lista de distribuições disponíveis, estimamos os parâmetros livres e em seguida calculamos os critérios AIC e BIC. No caso da variável aleatória gamma, por exemplo, usando os resultados do Exemplo 5.9, os critérios AIC e BIC são

$$\begin{aligned} AIC &= -2 \log L(\hat{\alpha}, \hat{\beta}) + 2K = \Big[n\hat{\alpha} \log \hat{\beta} + n \log \Gamma(\hat{\alpha}) \\ &- (\hat{\alpha} - 1) \sum_{i=1}^{n} \log y_i + \frac{1}{\hat{\beta}} \sum_{i=1}^{n} y_i \Big] + 2 \times 2 \\ BIC &= -2 \log L(\hat{\alpha}, \hat{\beta}) + K \log n = \Big[n\hat{\alpha} \log \hat{\beta} + n \log \Gamma(\hat{\alpha}) \\ &- (\hat{\alpha} - 1) \sum_{i=1}^{n} \log y_i + \frac{1}{\hat{\beta}} \sum_{i=1}^{n} y_i \Big] + 2 \times \log n. \end{aligned} \tag{7.4}$$

Uma vez estimados os parâmetros livres e calculados os critérios AIC e BIC para cada variável aleatória contínua disponível para a análise, podemos finalmente escolher a variável (ou variáveis) aleatória que apresenta o menor AIC ou BIC. Em geral, é razoável escolher mais de uma variável aleatória contínua e mais de uma variável aleatória discreta para proceder com os outros passos da análise. Pode acontecer de uma variável aleatória ficar um pouco atrás, em termos de AIC e BIC, da variável com menores critérios de seleção, e apresentar melhor performance nos testes de ajuste (apresentados na Seção 7.2). É importante manter em mente que não existe um único critério estatístico que seja soberano: é aconselhável sempre combinar os resultados de várias abordagens de seleção e avaliação de modelos, de forma a termos resultados mais defensáveis.[3] Some-se a isso a avaliação subjetiva dos analistas especializados e a intuição fornecida pela teoria econômica, que sempre deve ser inserida também na análise.

Uma vez estimados os parâmetros das variáveis aleatórias contínuas e discretas, e selecionados os modelos mais adequados de acordo com os critérios de AIC e BIC, podemos utilizar critérios para avaliar se os modelos (distribuições de frequência e severidade) escolhidos de fato se adequam às amostras analisadas. Para isso, podemos utilizar diversos procedimentos estatísticos, como por exemplo o **teste de Kolmogorov-Smirnov**, **o teste de Anderson-Darling** e o **teste de Crámer-Von Mises** para variáveis aleatórias contínuas. Para variáveis aleatórias discretas, o teste mais comum é o teste qui-quadrado.

[3]Outra forma muito comum usada para a avaliação de modelos é a chamada validação cruzada (*cross validation*). Nessa metodologia se particiona a amostra em subconjuntos e, enquanto um dos subconjuntos é usado para fazer a análise estatística desejada, os outros subconjuntos são usados para validar a análise estatística.

7.2 Avaliação do ajuste de distribuições contínuas

Nesta seção, discutiremos alguns critérios gráficos e numéricos para avaliar o ajuste dos modelos estatísticos para variáveis aleatórias contínuas. Inicialmente faremos uma discussão sobre o conceito de função distribuição empírica $\hat{F}(x)$. A ideia de utilizar a distribuição empírica vale especialmente no caso de variáveis aleatórias contínuas, quando não é possível estimar com grande precisão a função densidade. Em termos práticos, é muito mais conveniente estimar a função distribuição empírica $\hat{F}(x)$ e compará-la com a função distribuição teórica $F(x;\hat{\theta})$, onde $\hat{\theta}$ é o vetor de parâmetros estimados, via máxima verossimilhança ou via método de momentos, por exemplo.

A função distribuição empírica $\hat{F}(x)$ é construída a partir de uma amostra disponível S para uma variável contínua (por exemplo, severidade das perdas operacionais), com $S = \{X_1, X_2, \ldots, X_n\}$, de tamanho n. A expressão para a função distribuição empírica é

$$\hat{F}(x) = \frac{1}{n}\sum_{i=1}^{n} I_{(-\infty,x]}(X_i), \tag{7.5}$$

onde $I_A(x)$ é uma função indicadora, com $I_A(x) = 1$ quando $x \in A$, e $I_A(x) = 0$ caso contrário (A é um intervalo da reta $\Re$). Dessa forma, $\hat{F}(x)$ é a proporção de valores na amostra S que são menores ou igual a x. Pode-se mostrar que $\hat{F}(x)$ é um estimador para a probabilidade Prob$[X \leq x] = F(x)$. Por conveniência nos cálculos, pode-se utilizar a sugestão em Cruz (2005), fazendo-se uma pequena modificação do estimador apresentado na Eq. (7.5). O novo estimador, utilizado em diversos aplicativos de análise de risco, por exemplo, é dado por

$$\hat{F}(x) = \frac{1}{n}\left[-0.5 + \sum_{i=1}^{n} I_{(-\infty,x]}(X_i)\right] = \frac{1}{n}\sum_{i=1}^{n} I_{(-\infty,x]}(X_i) - \frac{1}{2n}. \tag{7.6}$$

Note que a diferença entre as Eqs. (7.5) e (7.6) é simplesmente o fator aditivo $-\frac{1}{2n}$, que converge para zero quando o tamanho da amostra n aumenta. A vantagem de utilizar essa modificação ficará mais clara quando falarmos dos gráficos de QQ-plot.

Para exemplificar a utilização da distribuição empírica na detecção de problemas de ajuste, a Figura 7.1 apresenta o gráfico da distribuição empírica $\hat{F}(x)$ versus a função distribuição teórica $F(x;\theta)$, baseada em uma variável aleatória lognormal. Os dados foram gerados exatamente a partir de uma variável aleatória lognormal, com parâmetros $\mu = 1.0$ e $\sigma = 0.8$. Portanto, estamos supondo um modelo teórico que coincide com os dados reais, de forma que os gráficos teórico $F(x)$ e empírico $\hat{F}(x)$ deveriam coincidir. De fato, isso é observado na Figura 7.1. Utilizamos quatro tamanhos de amostras diferentes ($n = 40$, $n = 100$, $n = 200$, $n = 400$), para mostrar que quanto maior a amostra, quando o modelo está corretamente especificado, as curvas teóricas e empíricas se aproximam cada vez mais.

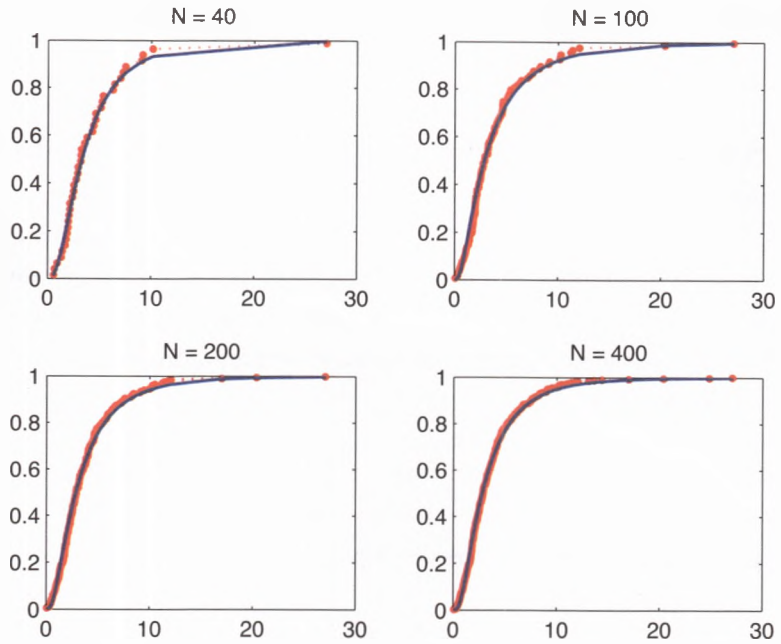


Figura 7.1: Função distribuição empírica versus função distribuição de um modelo teórico baseado em uma variável aleatória lognormal. Os dados foram gerados exatamente a partir de uma variável aleatória lognormal, com parâmetros $\mu = 1.0$ e $\sigma = 0.8$.

A Figura 7.2 apresenta o gráfico da distribuição empírica $\hat{F}(x)$ (gerada utilizando uma distribuição lognormal) *versus* a função distribuição teórica $F(x;\theta)$, baseada em uma variável aleatória de Rayleigh. Os dados foram gerados novamente a partir de uma variável aleatória lognormal, com parâmetros $\mu = 1.0$ e $\sigma = 0.8$. Portanto, nesse caso estamos tentando ajustar uma distribuição de Rayleigh a dados lognormais, e a análise gráfica deveria detectar esse problema de má especificação do modelo teórico. De fato, as curvas na Figura 7.2 mostram a diferença entre as duas curvas (teórica $F(x)$ e empírica $\hat{F}(x)$), e essa diferença não desaparece quando o tamanho da amostra N aumenta.

O segundo gráfico muito utilizado na análise de ajuste de distribuições advém de uma outra forma de comparar a distribuição empírica com a distribuição teórica. De fato, podemos fazer um gráfico da função distribuição empírica $\hat{F}(x)$ *versus* a função distribuição teórica $F(x)$, conforme apresentado nas Figuras 7.3 e 7.4 abaixo. Na Figura 7.3, apresentamos o gráfico de dispersão entre a função distribuição teórico no eixo horizontal e a função distribuição empírica no eixo vertical. A curva sólida corresponde ao que deveria ser a curva pontilhada caso o modelo teórico estivesse perfeitamente adequado aos dados. Nesse caso, estamos gerando os dados com uma distribuição lognormal, e estimando exatamente um modelo teórico lognormal, o que implica que as curvas sólidas e pontilhadas deveriam estar coincidentes. De fato, observe que a curva empírica (curva pontilhada) e a curva teórica (curva sólida) se aproximam cada vez mais, à medida que o tamanho n da amostra aumenta. Esse tipo de gráfico é comumente conhecido como **PP-plot**, ou **probability-probability plot**.

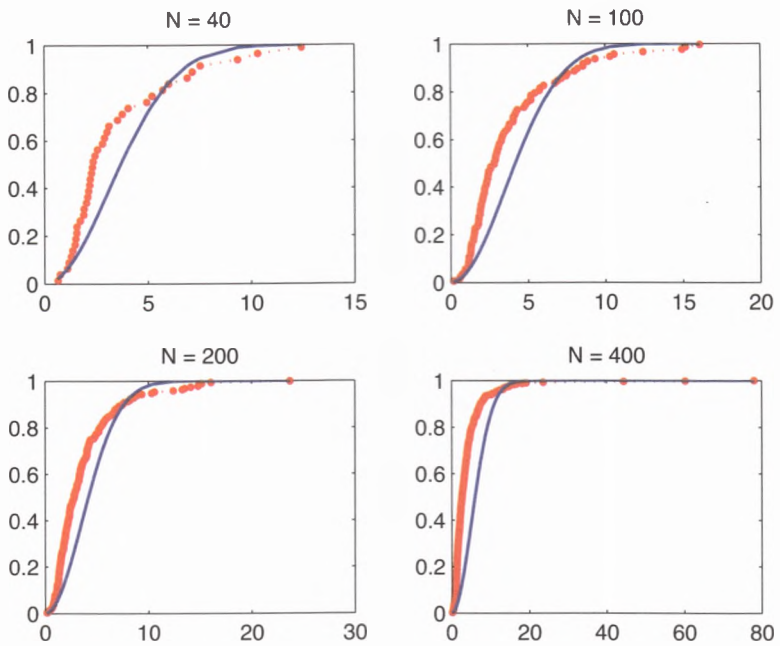


Figura 7.2: Função distribuição empírica *versus* função distribuição de um modelo teórico baseado em uma variável aleatória de Rayleigh. Os dados foram gerados a partir de uma variável aleatória lognormal, com parâmetros $\mu = 1.0$ e $\sigma = 0.8$.

Na Figura 7.4, apresentamos o PP-plot, dessa vez comparando os dados reais gerados novamente a partir de uma variável aleatória lognormal, com o modelo teórico, supondo-se uma distribuição de Rayleigh. Observe o nítido descasamento entre as curvas empírica e teórica, que não desaparece à medida que o tamanho n da amostra aumenta. Portanto, o PP-plot também pode se constituir em uma ferramenta muito útil para detecção de problemas na especificação da distribuição teórica utilizada para modelar os dados reais disponíveis.

Além de comparar as funções distribuição empírica e teórica, uma outra forma também muito eficaz de detectar problemas no ajuste dos modelos paramétricos utilizados é via comparação dos quantis. Nesse caso, a ideia é comparar os quantis da distribuição teórica com os quantis da distribuição empírica. A partir dessa comparação, construímos um tipo de gráfico comumente conhecido como **QQ-plot**, ou **quantil-quantil plot**. As Figuras 7.5 e 7.6 apresentam exemplos de QQ-plots, para modelos teóricos com distribuições lognormal e de Rayleigh respectivamente.

Similarmente ao que fizemos no caso do PP-plot, no QQ-plot devemos encontrar valores teóricos para serem comparados a valores empíricos. O quantis empíricos nesse caso serão simplesmente os valores ordenados, em ordem crescente, na amostra S. Portanto, os quantis empíricos são dados pela sequência crescente $X_{(1)}, X_{(2)}, X_{(3)}, \ldots, X_{(n)}$, onde $X_{(i)}$ corresponde ao i-ésimo maior elemento na amostra S. Obviamente, $X_{(1)}$ é o valor mínimo na amostra, enquanto $X_{(n)}$ é o valor máximo na amostra. Os quantis

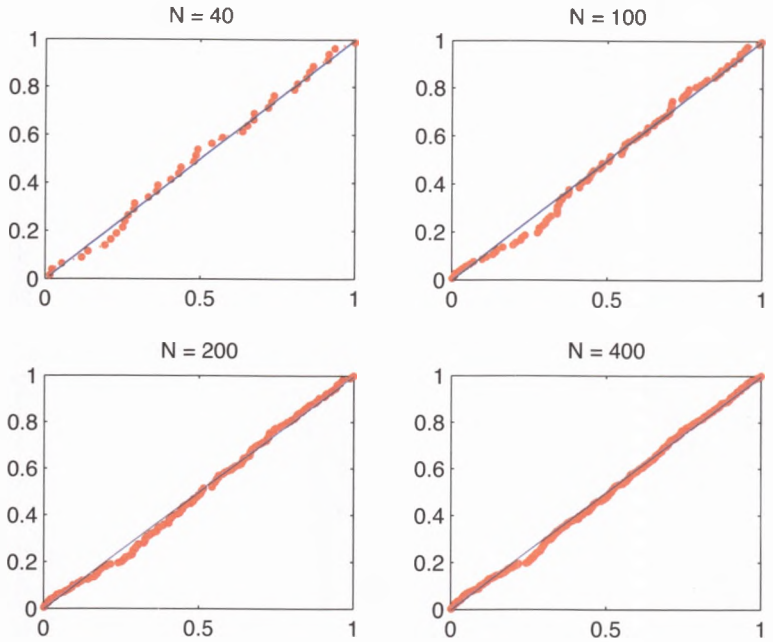


Figura 7.3: PP-plot um modelo teórico baseado em uma variável aleatória lognormal. Os dados foram gerados a partir de uma variável aleatória lognormal, com parâmetros $\mu = 1.0$ e $\sigma = 0.8$.

teóricos são obtidos a partir da expressão

$$q_i = F^{-1}(\hat{F}(X_{(i)}); \hat{\theta}), \quad \text{para } i = 1, 2, \ldots, n,$$

onde

$$\hat{F}(X_{(i)}) = \frac{1}{n} \sum_{i=1}^{n} I_{(-\infty, X_{(i)}]}(X_i) - \frac{1}{2n} = \frac{i - 0.5}{n}.$$

e $F^{-1}(u; \hat{\theta})$ corresponde ao quantil de $(100 \times u)$ % da distribuição teórica $F(x; \hat{\theta})$, onde $\hat{\theta}$ é o parâmetro estimado, utilizando-se máxima verossimilhança ou estimador via método de momentos. O QQ-plot consiste então em um gráfico de dispersão dos quantis empíricos (realmente observados) no eixo horizontal *versus* os quantis teóricos no eixo vertical (previstos pelo modelo teórico paramétrico). Esse gráfico de dispersão corresponde à curva pontilhada na Figura 7.5. Essa figura apresenta o gráfico QQ-plot para um modelo teórico lognormal, ajustado a dados de fato gerados a partir de uma variável aleatória lognormal. Portanto, o modelo está corretamente especificado. Para fins de comparação, a curva sólida seria o a curva ideal, quando o modelo está corretamente especificado. Observe que, devido ao fato de o modelo teórico coincidir com o modelo real (do qual os dados foram gerados), a curva empírica, pontilhada, está muito próxima da curva sólida (para comparação), e essa proximidade aumenta, à medida que aumentamos o tamanho n da amostra.

A Figura 7.6 apresenta a comparação entre o modelo teórico, baseado em uma variável aleatória de Rayleigh, e o modelo real, onde os dados foram gerados a partir de uma variável aleatória lognormal.

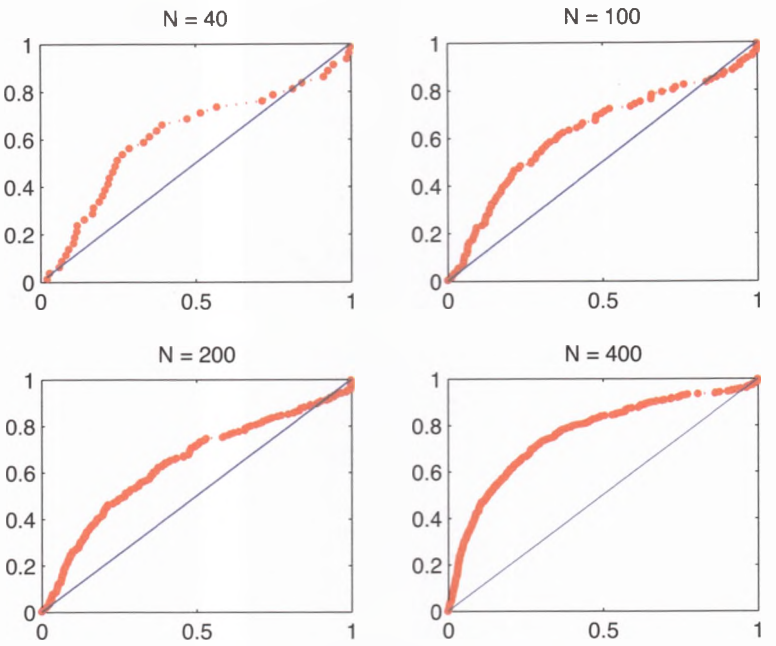


Figura 7.4: PP-plot um modelo teórico baseado em uma variável aleatória de Rayleigh. Os dados foram gerados a partir de uma variável aleatória lognormal, com parâmetros $\mu = 1.0$ e $\sigma = 0.8$.

Portanto, o modelo teórico suposto está errado. De fato, isso é sugerido pelo QQ-plot, dado que a curva empírica (pontilhada) está bem afastada da curva ideal (sólida). Um observação importante a ser feita aqui é que o QQ-plot constitui-se em uma ferramenta muito conveniente para a análise de como o modelo teórico utilizado ajusta-se nas caudas. Esse fato é especialmente importante em análise de risco, já que as caudas representam justamente os eventos mais indesejáveis. Especificamente na Figura 7.6, observe que os quantis empíricos, que correspondem justamente aos valores observados, estão à direita do que é previsto pelo modelo teórico. Isso indica que os dados apresentam valores extremos com maior frequência do que é previsto pelo modelo teórico. Portanto, caso usássemos a distribuição de Rayleigh para modelar esses eventos com valores altos, estaríamos incorrendo em uma subestimação do risco real ao qual a instituição financeira está exposta.

Para alguns modelos paramétricos específicos, como é o caso dos modelos de distribuições combinadas, a serem vistos na Seção 7.4, o cálculo da função distribuição inversa $F^{-1}(\hat{F}(X_{(i)}); \hat{\theta})$, necessária para a obtenção do QQ-plot, pode se constituir em uma tarefa demorada, e computacionalmente demandante. Especificamente para modelos estocásticos para variáveis contínuas em risco (como por exemplo, modelos para a severidade das perdas em risco operacional), existe uma preocupação especial em utilizar o QQ-plot para identificar problemas de ajuste na cauda (valores de perdas muito altos) da distribuição. Por esse motivo, em algumas situações, pode ser conveniente o cálculo do QQ-plot especificamente para uma parcela das observações (por exemplo, 10% ou 5%) na amostra, sendo essa parcela referente aos maiores valores de ocorrências observadas na base de dados. Com isso, seria necessário calcular a função distribuição inversa

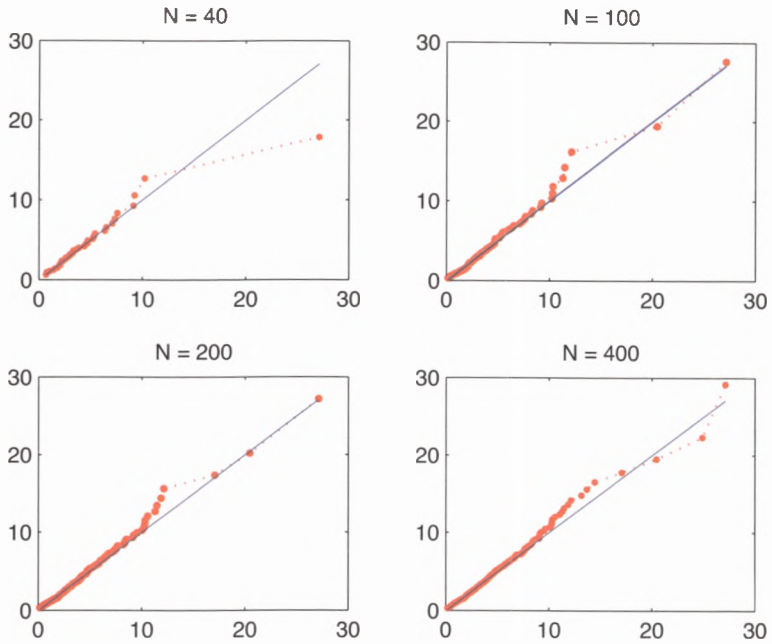


Figura 7.5: QQ-plot um modelo teórico baseado em uma variável aleatória lognormal. Os dados foram gerados a partir de uma variável aleatória lognormal, com parâmetros $\mu = 1.0$ e $\sigma = 0.8$.

$F^{-1}(\hat{F}(X_{(i)}); \hat{\theta})$ apenas para os 5% ou 10% maiores valores na amostra. Isso pode reduzir bastante o esforço computacional para processamento dos modelos e para processamento dos critérios de ajustes.

Finalmente, além dos critérios gráficos baseados na distribuição empírica (*versus* a distribuição acumulada teórica), no QQ-plot e no PP-plot, o analista pode utilizar testes estatíticos formais, como o **teste de Kolmogorov-Smirnov** (GIBBONS, 1992),[4] comumente representado como teste KS. A estatística teste nesse caso é dada por

$$\text{KS} = \sup_x \left| F(x; \hat{\theta}) - \frac{1}{n} \sum_{i=1}^{n} I_{(-\infty, X_{(i)}]}(x) \right|, \tag{7.7}$$

onde $F(x; \hat{\theta})$ corresponde à distribuição teórica e $\hat{\theta}$ é o vetor de parâmetros do modelo paramétrico, estimado a partir dos dados históricos. Quanto maior o valor da estatística KS, mais longe o modelo teórico estimado estará dos dados observados. Note que o segundo termo dentro do operador de valor absoluto $|\cdot|$ corresponde à distribuição empírica, com base nos dados observados.

A estatística KS nos fornece então o primeiro componente para realizar o teste de hipótese de ajuste dos modelos. O segundo componente é a distribuição da estatística teste, e o terceiro é a regra de rejeição da hipótese nula. Como estamos testando a qualidade do ajuste do modelo teórico aos dados, a hipótese

[4] Uma aplicação muito útil e recentemente popularizada da estatística de Kolmogorov-Smirnov é testar se variáveis aleatórias pertencentes a uma determinada amostra de dados usualmente advinda de um sistema com características complexas (por exemplo, mercados financeiros) têm distribuição dada por uma lei de potência (um exemplo de uma lei de potência é a distribuição de Pareto apresentada na Seção 3.3.8). Para detalhes, consultar (CLAUSET; SHALIZI; NEWMAN, 2009).

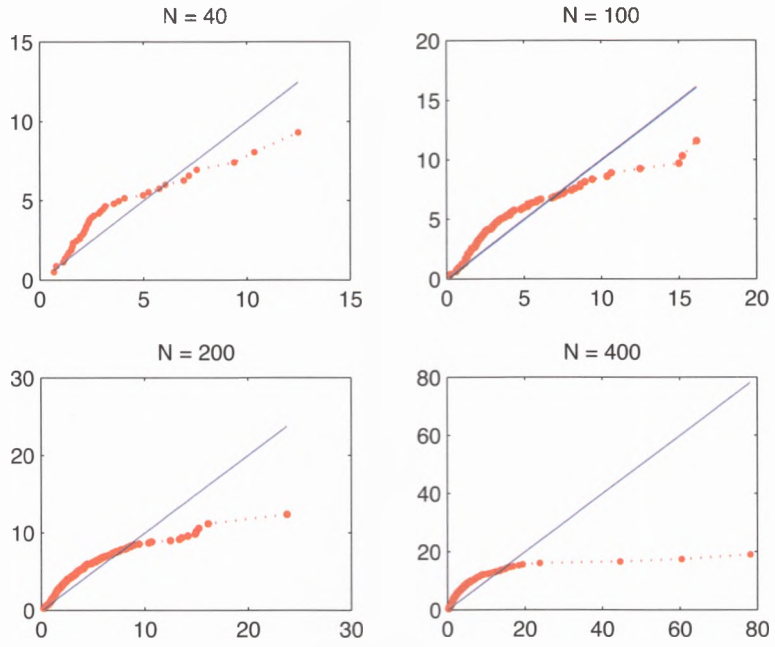


Figura 7.6: QQ-plot um modelo teórico baseado em uma variável aleatória de Rayleigh. Os dados foram gerados a partir de uma variável aleatória lognormal, com parâmetros $\mu = 1.0$ e $\sigma = 0.8$.

nula é H_0 : "o modelo teórico está bem ajustado aos dados", enquanto a hipótese alternativa é H_A : "o modelo teórico não está bem ajustado aos dados". A distribuição da estatística teste não corresponde a distribuições conhecidas, como é o caso da normal padronizada ou da distribuição qui-quadrada. Para a estatística KS, os valores críticos v_c da distribuição da estatísica teste são obtidos via simulações de Monte Carlo, e estão tabelados nas devidas referências. O leitor pode encontrar tais tabelas em Gibbons (1992). Para os valores críticos v_c tabelados, a regra de rejeição então será: rejeitar a hipótese nula (de que o modelo teórico é adequado) caso KS $> v_c$; caso contrário, podemos aceitar a hipótese nula, aceitando o modelo teórico testado. Como é de costume para os testes de hipóteses, os valores críticos tabelados dependem necessariamente de um nível de confiabilidade, conhecido como nível do teste α. As tabelas disponíveis de valores críticos já levam isso em consideração e apresentam valores críticos diferentes para diferentes níveis do teste. Os níveis geralmente apresentados são $\alpha = 1\%$, $\alpha = 5\%$, $\alpha = 10\%$. Quando o tamanho da amostra n é maior do que 40, podemos utilizar os valores críticos aproximados, com base na aproximação assimptótica. Para níveis de confiança $\alpha = 20\%$, 10%, 5%, 2% e 1%, os valores críticos aproximados são dados na tabela 7.1 a seguir. De acordo com essa tabela, para $n = 100$, o valor crítico para o teste KS, com $\alpha = 1\%$, é igual a $1.63/\sqrt{100} = 0.163$.

Um segundo teste de ajuste comumente empregado na literatura estatística é o **teste de Crámer-von-Mises**, que também baseia-se em uma medida de discrepância entre a distribuição acumulada teórica $F(x)$ e a distribuição empírica. As hipóteses nula e alternativa testadas são as mesmas

Tabela 7.1: Valores críticos para o teste de Kolmogorov-Smirnov, utilizando a aproximação assimptótica.

Nível de significância (α)	Valores críticos assimptóticos
20%	$1.07/\sqrt{n}$
10%	$1.22/\sqrt{n}$
5%	$1.36/\sqrt{n}$
2%	$1.52/\sqrt{n}$
1%	$1.63/\sqrt{n}$

do testes de Kolmogorov-Smirnov. A estatística teste tem expressão

$$W^2 = n\int_{-\infty}^{\infty}\Big[F(x;\hat{\theta}) - \frac{1}{n}\sum_{i=1}^{n} I_{(-\infty,X_{(i)}]}(x)\Big]^2 f(x;\hat{\theta})dx, \tag{7.8}$$

onde $f(x;\hat{\theta})$ é a função densidade de probabilidade do modelo teórico, com $\hat{\theta}$ o vetor de parâmetros estimados a partir da base de dados. Caso estejamos tratando de observações puramente positivas, como é o caso de modelos de perdas operacionais, a Eq. (7.8) transforma-se em

$$W^2 = n\int_{0}^{\infty}\Big[F(x;\hat{\theta}) - \frac{1}{n}\sum_{i=1}^{n} I_{(-\infty,X_{(i)}]}(x)\Big]^2 f(x;\hat{\theta})dx. \tag{7.9}$$

Seja $x_{(1)},\ldots,x_{(n)}$ os valores observados, na amostra aleatória, em ordem crescente. Então, pode-se mostrar que W^2 pode ser calculada por meio da expressão

$$W^2 = \frac{1}{12n} + \sum_{i=1}^{n}\left[\frac{2i-1}{2n} - F(x_{(i)};\hat{\theta})\right]^2.$$

Com base em Tiku (1965), é possível construir valores críticos aproximados para o teste de Crámer-von-Mises. Esses valores críticos baseiam-se na aproximação da estatística W^2 por uma distribuição qui-quadrada centrada. De fato, a estatística W^2 pode ser aproximada pela distribuição da variável aleatória $a + b\chi^2$, onde χ^2 é uma distribuição qui-quadrada com r graus de liberdade. A expressões para a, b e r são

$$\begin{aligned} a &= \frac{336n^2 - 959n + 609}{210(32n^2 - 61n + 30)}, \\ b &= \frac{32n^2 - 61n + 30}{84n(4n-3)}, \\ r &= \frac{98}{5}\frac{n(4n-3)^3}{(32n^2 - 61n + 30)^2}, \end{aligned}$$

onde n é o tamanho da amostra. Portanto, para um nível de significância de 1% por exemplo, e um tamanho de amostra $n = 200$, temos os valores $a = 0.049762$, $b = 0.094688$, $c = 1.234689$. O valor crítico a

1% para uma variável aleatória qui-quadrada com 1 grau de liberdade é 6.634897. Portanto, o valor crítico aproximado, para o teste de Crámer-von-Mises é igual a $v_c = a + b \times 6.634897 = 0.678005$.

Finalmente, o terceiro teste abordado nesta seção é o **teste de Anderson-Darling** (ANDERSON; DARLING, 1952). As hipóteses nula e alternativa são as mesmas hipóteses consideradas para o teste de Kolmogorov-Smirnov e para o teste de Crámer-von-Mises. Para valores ordenados $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$, de uma amostra de tamanho n, a estatística teste A^2 tem expressão

$$A^2 = -n - S, \tag{7.10}$$

onde S é dada por

$$S = \sum_{i=1}^{n} \frac{2i-1}{n} \left[\log F(y_{(i)}; \hat{\theta}) + \log(1 - F(y_{(n+1-i)}; \hat{\theta}))\right].$$

Os valores críticos dependem de que distribuição teórica $F(y) = F(y; \theta)$ está sendo testada. A maioria dos softwares estatísticos contêm valores críticos específicos para a distribuição teórica pertencente a famílias de distribuições específicas (normal, lognormal, exponencial, Weibull, logística, distribuição de valores extremos do tipo I). Para maiores detalhes, vide Stephens (1974), Stephens (1976), Stephens (1977), Stephens (1977b) e Stephens (1979).

7.3 Avaliação do ajuste de distribuições discretas

Na seção anterior, descrevemos critérios gráficos e estatísticos para avaliar o ajuste de modelos para variáveis aleatórias contínuas. Nesta seção, apresentaremos o tratamento para avaliação de ajuste dos modelos para variáveis aleatórias discretas. A abordagem inicialmente apresentada no caso discreto é um pouco mais simples que no caso de variáveis aleatórias contínuas. Basicamente, dividimos o intervalo de possíveis valores para a variável discreta Y_i, e comparamos a frequência relativa esperada (a partir do modelo teórico) e a frequência relativa observada (a partir da amostra observada). Essa comparação pode ser efetuada tanto de forma gráfica, quanto de forma analítica por meio de estatística teste.

Seja então Y_i, $i = 1, \ldots, n$, uma sequência de n valores observados para a variável discreta a ser modelada (por exemplo, o número de perdas operacionais que ocorrem por dia ou por semana). Sejam $Y_{\max}$ e $Y_{\min}$ os valores máximo e mínimo na sequência Y_i, $i = 1, \ldots, n$. Considere uma sequência estritamente crescente $(Y_{\min} - 1) = c_0, c_1, c_2, \ldots, c_m = Y_{\max}$ de valores reais que dividem o intervalo $[Y_{\min}, Y_{\max}]$ em m sub-intervalos. A frequência *relativa* esperada $r_{(c_k, c_{k+1}]}$ de observações no intervalo $(c_k, c_{k+1}]$ é dada por

$$r_{(c_k, c_{k+1}]} = F(c_{k+1}; \hat{\theta}) - F(c_k; \hat{\theta}),$$

onde $F(y;\hat{\theta})$ é a função distribuição acumulada do modelo paramétrico para a distribuição discreta, com vetor de parâmetros $\hat{\theta}$ estimado a partir dos dados históricos Y_i, $i = 1, \ldots, n$. Note que $Y_{\min} \in (c_0, c_1]$. Portanto, a frequência absoluta esperada $e_{(c_k, c_{k+1}]}$, que corresponde ao número esperado de observações dentro do intervalo $(c_k, c_{k+1}]$, é obtida por

$$e_{(c_k, c_{k+1}]} = n \times (F(c_{k+1};\hat{\theta}) - F(c_k;\hat{\theta})),$$

onde n é o tamanho da amostra. A frequência absoluta observada $o_{(c_k, c_{k+1}]}$ para o número de observações no intervalo $(c_k, c_{k+1}]$ é igual a

$$o_{(c_k, c_{k+1}]} = \sum_{i=1}^{n} I_{(c_k, c_{k+1}]}(Y_i),$$

onde, conforme visto acima, $I_{(c_k, c_{k+1}]}(Y_i) = 1$ quando $Y_i \in (c_k, c_{k+1}]$, e $I_{(c_k, c_{k+1}]}(Y_i) = 0$ caso contrário. Portanto, uma maneira de estudar a adequação do modelo paramétrico teórico para a variável aleatória discreta é simplesmente comparar as sequências $o_{(c_k, c_{k+1}]}$, $k = 0, \ldots, m-1$, e $e_{(c_k, c_{k+1}]}$, $k = 0, \ldots, m-1$. Espera-se que, quando o modelo teórico estimado estiver descrevendo bem os dados observados, as duas sequências de dados estejam bastante próximas.

O teste qui-quadrado é uma maneira formal de testar a proximidade entre essas duas sequências (GIBBONS, 1992). A definição da estatística teste χ nesse caso é

$$\chi = \sum_{k=0}^{m-1} \frac{\left[o_{(c_k, c_{k+1}]} - e_{(c_k, c_{k+1}]}\right]^2}{e_{(c_k, c_{k+1}]}}.$$

Sob a hipótese nula de ajuste adequado do modelo paramétrico teórico, a estatística χ possui distribuição χ^2_{m-1} assintoticamente,[5] onde χ^2_{m-1} é a conhecida distribuição qui-quadrada com $m-1$ graus de liberdade. Pode se mostrar que a variável aleatória χ^2_{m-1} corresponde a uma soma do quadrado de $m-1$ variáveis aleatórias normais padronizadas independentes. Com base na distribuição assintótica χ^2_{m-1}, é possível encontrar valores críticos para testar a hipótese nula H_0 de que o modelo paramétrico se ajuste bem aos dados observados empiricamente. A hipótese alternativa H_A corresponde à hipótese de que o modelo paramétrico não está devidamente ajustado. Os valores críticos assimptóticos podem ser obtidos a partir de uma tabela para a variável qui-quadrada.

Finalmente, além do critério estatístico via teste qui-quadrado descrito acima, a avaliação do modelo discreto pode levar em consideração critérios gráficos. Um tipo de gráfico simples, que pode ser utilizado

[5] Assintoticamente nesse caso implica que, quanto maior a amostra n, mais próxima a distribuição da variável aleatória χ vai estar de uma variável aleatória χ^2_{m-1}. Em uma análise mais rigorosa, alguns cuidados têm que ser tomados pois, dado que os limites c_0 e c_m dependem da amostra disponível Y_i, $i = 1, \ldots, n$, esses limites também são variáveis aleatórias. Portanto, pode acontecer de a distribuição assintótica da estatística χ ser mais complexa do que esperamos em princípio. Em todo caso, a abordagem proposta com base no teste qui-quadrado segue a sugestão em Gibbons (1992), e pode ser usada para auxiliar na avaliação do ajuste dos modelos de frequência. Além da avaliação dos modelos utilizando o teste qui-quadrado, é interessante também utilizar as indicações gráficas conforme descrito nesta seção.

na avaliação do ajuste dos modelos para variáveis aleatórias discretas, baseia-se na comparação entre as frequências relativas esperadas $r_{(c_k,c_{k+1}]}$, $k = 0, \ldots, m-1$. e as frequências relativas observadas $\hat{r}_{(c_k,c_{k+1}]}$, dadas por

$$\hat{r}_{(c_k,c_{k+1}]} = \frac{1}{n} \times o_{(c_k,c_{k+1}]},$$

onde $o_{(c_k,c_{k+1}]}$ é a frequência absoluta observada. Portanto, o analista pode comparar, graficamente, essas duas sequências de frequências relativas, o que permite identificar, por exemplo, em que intervalo $(c_k, c_{k+1}]$ do conjunto de dados há um maior descasamento entre o modelo paramétrico teórico e os dados observados.

Além da estatística teste χ acima, para o teste qui-quadrado (também conhecido como teste qui-quadrado de Pearson), outras estatísticas testes podem ser empregadas para verificar a adequação do modelo paramétrico teórico aos dados empíricos. Para mais detalhes, vide Steele, Chaseling e Hurst (2005). A primeira estatística teste alternativa é a estatística discreta de Cramér-von-Mises, com expressão

$$W^2 = \frac{1}{n} \sum_{k=0}^{m-1} Z_k^2 p_k, \tag{7.11}$$

onde p_k é a probabilidade de ocorrência de um determinado valor no intervalo $(c_k, c_{k+1}]$, e

$$Z_k = \sum_{j=0}^{k} \left[o_{(c_j,c_{j+1}]} - e_{(c_j,c_{j+1}]} \right], \tag{7.12}$$

para $k = 0, 1, \ldots, m-1$. A segunda estatística teste alternativa é a estatística discreta de Anderson-Darling, com expressão

$$A^2 = \frac{1}{n} \sum_{k=0}^{m-1} \frac{Z_k^2 p_k}{H_k(1 - H_k)},$$

onde $H_k = \sum_{j=0}^{k} e_{(c_j,c_{j+1}]}$, para $k = 0, 1, \ldots, m-1$. Finalmente, temos a estatística discreta de Watson, com expressão

$$U^2 = \frac{1}{n} \sum_{k=0}^{m-1} \left[Z_k - \bar{Z} \right]^2, \tag{7.13}$$

onde $\bar{Z} = \sum_{k=0}^{m-1} Z_k p_k$. Para as três estatísticas nas Eqs. (7.11), (7.12) e (7.13), os valores críticos precisam ser obtidos via simulações de Monte Carlo, e dependerão da distribuição teórica sendo testada. Uma alternativa é a utilização de *bootstrap* para obtenção dos valores críticos. Para uma discussão mais aprofundada, vide Stute, Mateiga e Quindimil (2007) e Meintanis e Swanepoel (2007).

Depois de empregados os testes formais, e os critérios gráficos, para avaliar a qualidade dos modelos estatísticos paramétricos aos dados empíricos, pode acontecer de o pesquisador não encontrar ajuste adequado para os modelos disponíveis no software que está sendo empregado. Uma primeira alternativa

a ser seguida seria tentar ajustar outras distribuições descritas na literatura. No entanto, pode acontecer de que, mesmo depois de testar outros modelos paramétricos, o pesquisador ainda assim não encontre o ajuste desejado. Uma outra alternativa é empregar modelos que combinem os modelos paramétricos mais simples. Conforme veremos na próxima seção, pode-se combinar por exemplo uma variável aleatória de Poisson com uma variável aleatória binomial negativa, obtendo uma flexibilidade maior, o que permitirá ao analista ajustar dados empíricos com características das mais variadas. Combinando-se distribuições, é possível por exemplo capturar a presença de duas ou mais modas nos dados observados. Alternativamente, o analista pode estar interessado em capturar observações extremas, que podem obedecer a um processo estocástico diferente das observações menores. Essas diferentes situações serão discutidas na Seção 7.4.

7.4 Combinação de modelos

As distribuições apresentadas no Capítulo 3 constituem a base das variáveis aleatórias para a modelagem de variáveis aleatórias discretas e contínuas. Dadas as diversas distribuições apresentadas, os analistas de risco de uma instituição bancária, por exemplo, dispõem de uma variedade razoável de opções para ajustar os modelos teóricos aos dados. No entanto, para alguns conjuntos de dados específicos (tanto discretos, quanto contínuos), nem sempre será possível ajustar adequadamente alguma das distribuições padrões apresentadas nas Seções 3.2 e 3.3, de acordo com os testes de ajustes apresentados na seção anterior. A depender da complexidade das distribuições do dados, alternativas mais flexíveis deverão ser utilizadas.

Nesta seção, discutiremos duas classes de modelos, tanto para dados discretos quanto para dados contínuos, que podem ser utilizadas para prover mais flexibilidade aos analistas quantitativos e pesquisadores. Com essas duas classes adicionais, será possível modelar praticamente todas as bases de dados encontradas na prática. A primeira classe corresponde à mistura de distribuições ou distribuições combinadas[6]. A segunda classe de modelos corresponde ao que chamamos de distribuições por subintervalo, pois ela consiste em dividir o intervalo dos dados em subintervalos, e ajustar uma distribuição específica a cada um desses subintervalo.

7.4.1 Mistura de distribuições

Inicialmente, descreveremos a classe de **modelo de mistura**[7] ou **distribuições combinadas**. As distribuições combinadas são construídas pela combinação linear de duas ou mais distribuições padrões, apresentadas previamente nas Seções 3.2 e 3.3. Para simplificar a discussão, inicialmente utilizaremos combinações de apenas duas variáveis aleatórias, que podem pertencer inclusive à mesma família.[8] Além disso, para combinações de mais de duas distribuições padrões, o modelo a ser estimado apresenta um grau

[6] Essa classe de modelos é comumente conhecida como *mixture models*.
[7] Do inglês, *mixture models*.
[8] Podemos compôr, por exemplo, uma lognormal com uma gamma, ou duas lognormais, com parâmetros diferentes.

maior de complexidade, o que incorre em maior dificuldade computacional na estimativa dos parâmetros. Para maiores detalhes sobre misturas de distribuições. vide McLachlan e Peel (2000).

Conforme comentamos anteriormente, uma das medidas mais importantes na caracterização de variáveis aleatórias é a função densidade (ou função de frequência, no caso de variáveis aleatórias discretas). A partir dela, é possível obter todas as demais caracterizações da distribuição, como a média, a variância, a curtose, o coeficiente de assimetria etc. No caso de distribuições combinadas, a função densidade (ou função frequência) da distribuição combinada tem a expressão a seguir,

$$f_c(x;\theta) = p_1 \times f_1(x;\theta_1) + p_2 \times f_2(x;\theta_2), \tag{7.14}$$

onde $f_1(x;\theta_1)$ e $f_2(x;\theta_2)$ são as funções densidade (ou funções frequências) para as duas distribuições padrões sendo combinadas, θ_1 e θ_2 são os conjuntos de parâmetros para cada uma das distribuições padrões, p_1 e p_2 são os pesos respectivamente, com $p_1+p_2 = 1$ e $p_1, p_2 \in (0,1)$. A função $f_c(x;\theta)$ é a função densidade (ou frequência) da distribuição combinada resultante. O parâmetro θ corresponde à união de θ_1, θ_2 e p_1 (e consequentemente p_2).

Similarmente à função densidade (ou frequência), a função distribuição acumulada $F_c(x;\theta)$ da distribuição combinada tem expressão

$$F_c(x;\theta) = p_1 \times F_1(x;\theta_1) + p_2 \times F_2(x;\theta_2), \tag{7.15}$$

onde $F_1(x;\theta_1)$ e $F_2(x;\theta_2)$ são as funções distribuição acumulada de cada variável aleatória combinada.

Em modelos para variáveis aleatórias contínuas, podemos combinar uma distribuição exponencial com uma distribuição gamma (vide Figura 7.7), ou uma distribuição lognormal com uma distribuição gamma, como no Exemplo 7.1.

Exemplo 7.1 (Combinação entre lognormal e gamma) Na combinação entre lognormal e gamma, a função densidade tem expressão

$$f_c(x;\theta) = p_1 \times \left[\frac{1}{x\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(\log x - \mu)^2}\right] + p_2 \times \left[\frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-x/\beta}\right], \tag{7.16}$$

para $x \in (0,\infty)$. Nesse caso, temos $\theta_1 = [\mu\ \sigma^2]'$, $\theta_2 = [\alpha\ \beta]'$ e $\theta = [\mu\ \sigma^2\ \alpha\ \beta\ p_1]'$.

Exemplo 7.2 (Combinação de duas variáveis Weibull) Podemos combinar duas distribuições da mesma família, como por exemplo, duas variáveis aleatórias de Weibull, onde a função densidade resultante será

$$f_c(x;\theta) = p_1 \times \left[\beta_1 \alpha_1^{-\beta_1} y^{\beta_1 - 1} e^{-\left[\frac{y}{\alpha_1}\right]^{\beta_2}}\right] + p_2 \times \left[\beta_2 \alpha_2^{-\beta_2} y^{\beta_2 - 1} e^{-\left[\frac{y}{\alpha_2}\right]^{\beta_2}}\right], \tag{7.17}$$

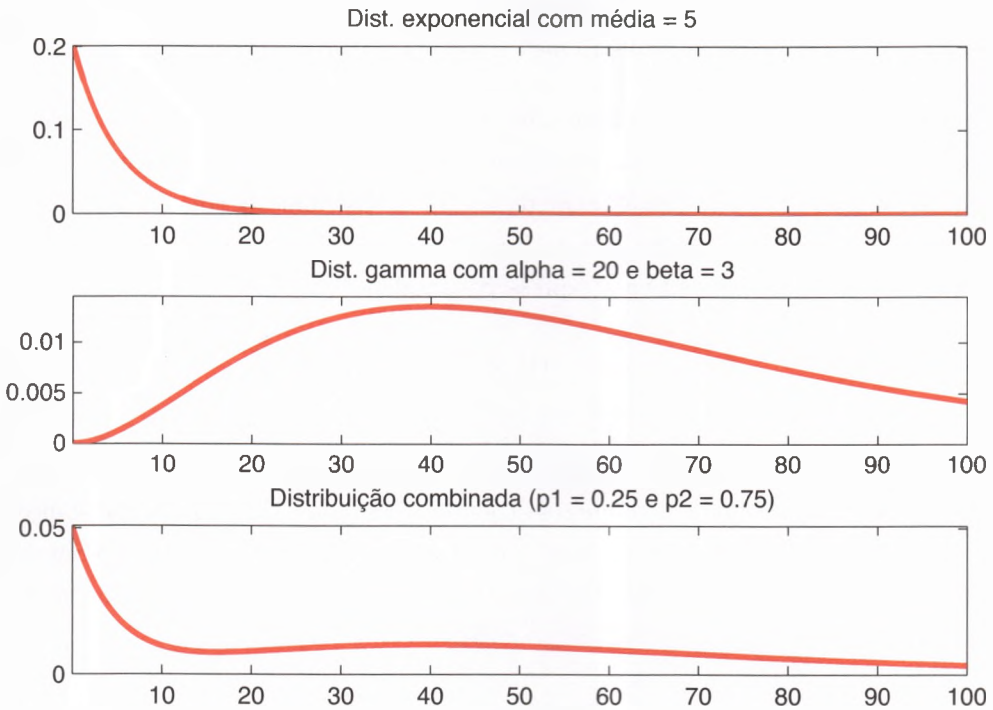


Figura 7.7: Função densidade para a combinação de uma distribuição exponencial com média 5.0 e uma distribuição gamma com parâmetros $\alpha = 20$ e $\beta = 3.0$. O peso para a distribuição exponencial é igual a 0.25 e para a distribuição gamma é igual a 0.75.

para $x \in (0, \infty)$. Nesse caso, temos $\theta_1 = [\alpha_1 \ \beta_1]'$, $\theta_2 = [\alpha_2 \ \beta_2]'$ e $\theta = [\alpha_1 \ \beta_1 \ \alpha_2 \ \beta_2 \ p_1]'$. Para evitar problemas de identificação de parâmetros, supusemos que $\alpha_1 \neq \alpha_2$ e/ou $\beta_1 \neq \beta_2$. Não pode ocorrer $\alpha_1 = \alpha_2$ e $\beta_1 = \beta_2$ ao mesmo tempo, pois nesse caso o modelo resultante corresponde simplesmente a uma única variável aleatória de Weibull.

Exemplo 7.3 (Combinação de binomial negativa e Poisson) Para variáveis aleatórias discretas, utilizadas para modelagem da frequência das perdas operacionais, o procedimento é completamente análogo. Podemos, por exemplo, combinar uma variável aleatória Binomial negativa com uma variável aleatória de Poisson, obtendo uma função de frequência

$$f_c(x;\theta) = p_1 \times \left[\frac{\Gamma(r+x)}{\Gamma(r)\Gamma(x+1)} p^r (1-p)^x\right] + p_2 \times \left[\frac{e^{-\lambda}\lambda^x}{x!}\right], \tag{7.18}$$

para $x \in \{0, 1, 2, \dots\}$. Para os parâmetros do modelo, temos $\theta_1 = [r \ p]'$, $\theta_2 = [\lambda]$ e $\theta = [r \ p \ \lambda \ p_1]'$.

Exemplo 7.4 (Combinação de duas variáveis Poisson) Podemos também combinar duas distribuições da mesma família, como por exemplo duas variáveis aleatórias de Poisson

$$f_c(x;\theta) = p_1 \times \left[\frac{e^{-\lambda_1}\lambda_1^x}{x!}\right] + p_2 \times \left[\frac{e^{-\lambda_2}\lambda_2^x}{x!}\right], \tag{7.19}$$

para $x \in \{0, 1, 2, \dots\}$, e nesse caso $\theta_1 = [\lambda_1]'$, $\theta_2 = [\lambda_2]$ e $\theta = [\lambda_1\ \lambda_2\ p_1]'$.

A partir da função densidade, no caso de variáveis contínuas, ou da função de frequência, no caso de variáveis aleatórias discretas, é possível obter também os momentos (média, variância, curtose, coeficiente de assimetria etc.) das distribuições combinadas. Por exemplo, a média da distribuição combinada pode ser obtida simplesmente como uma combinação linear também das médias de cada distribuição padrão incluída na combinação. Portanto

$$\mu_c = p_1 \times \mu_1 + p_2 \times \mu_2, \tag{7.20}$$

onde μ_1, μ_2 e μ_c são as médias da primeira distribuição padrão, da segunda distribuição padrão e da distribuição combinada respectivamente.

No modelo apresentado no Exemplo 7.1 onde combinamos lognormal e gamma, a média da distribuição combinada será

$$\mu_c = p_1 \times \left[e^{\mu + \frac{\sigma^2}{2}}\right] + p_2 \times [\alpha\beta], \tag{7.21}$$

enquanto no modelo apresentado no Exemplo 7.3 onde combinamos uma variável aleatória binomial negativa e uma variável aleatória de Poisson, a média da distribuição combinada será

$$\mu_c = p_1 \times \left[\frac{r(1-p)}{p}\right] + p_2 \times [\lambda]. \tag{7.22}$$

Similarmente, os modelos com duas variáveis aleatórias de Weibull ou duas variáveis aleatórias de Poisson, respectivamente apresentados nos Exemplos 7.2 e 7.4, têm médias

$$\mu_c = p_1 \times \left[\alpha_1\left[\Gamma(1+\beta_1^{-1})\right]\right] + p_2 \times \left[\alpha_2\left[\Gamma(1+\beta_2^{-1})\right]\right], \tag{7.23}$$

$$\mu_c = p_1 \times [\lambda_1] + p_2 \times [\lambda_2], \tag{7.24}$$

respectivamente.

Conforme vimos nos diversos exercícios de simulações de Monte Carlo apresentados ao longo deste livro, em muitas situações é necessário gerar números aleatórios a partir das distribuições ajustadas aos dados reais. Em risco operacional, por exemplo, é necessário gerar números aleatórios, tanto para os modelos teóricos de frequência das perdas, quanto para os modelo de severidade. A partir dessas simulações, obtém-se estimativas para a distribuição de perdas operacionais agregadas, durante o período de um ano, por exemplo. No caso da geração de números aleatórios para distribuições combinadas, podemos utilizar um processo de geração de número aleatórios em dois estágios. Para entender esse processo de geração, utilizaremos o exemplo onde combinamos uma variável aleatória lognormal com uma variável aleatória gamma. Os passos para gerar um número aleatório X dessa combinação são:

1. Gerar um número aleatório a partir de uma variável aleatória Z, com distribuição de Bernoulli com probabilidade de sucesso p_1, onde p_1 é o peso da primeira distribuição sendo combinada (neste exemplo, a primeira distribuição é a lognormal).

2. Caso o número resultante da geração no item 1 for igual a um, escolhemos a primeira distribuição (no caso, lognormal) na combinação para ir para o segundo estágio da geração. Caso o número resultante da geração no item 1 for igual a 0, escolhemos a segunda distribuição (no caso, a gamma) para ir para o segundo estágio da geração.

3. Finalmente, geramos um número aleatório a partir da distribuição selecionada no item 2 acima. Portanto, caso $Z = 1$, o número X será gerado a partir de uma variável aleatória lognormal. Caso $Z = 0$, o número X será gerado a partir de uma variável aleatória gamma.

Para estimar o vetor de parâmetros θ da distribuição combinada, utilizamos o procedimento de máxima verossimilhança. Nesse caso, suponha que tenhamos disponível uma amostra X_1, X_2, X_3, ..., X_n, de tamanho n. A amostra $X_1, X_2, X_3, \ldots, X_n$ pode corresponder a uma sequência de números de perdas por dia, por mês, por semana etc., e nesse caso utilizamos modelos combinados de variáveis aleatórias discretas, ou pode corresponder uma sequência de severidades de perdas, onde utilizamos combinados de variáveis aleatórias contínuas. Conforme discutido na Seção 5.2, o estimador de máxima verossimilhança $\hat{\theta}$ de θ será o vetor de parâmetros que maximiza a função de log-verossimilhança

$$\log L(\theta_1, \theta_2, p_1) = \sum_{i=1}^{n} \log \Big[p_1 \times f_1(x; \theta_1) + (1 - p_1) \times f_2(x; \theta_2) \Big], \tag{7.25}$$

e o procedimento é o mesmo tanto no caso de modelos para variáveis discretas, quanto no caso de modelos para variáveis contínuas.

A maximização da função de log-verossimilhança na expressão acima não possui solução analítica e o usuário tem que recorrer a métodos numéricos para encontrar as estimativas dos parâmetros nessa maximização. Para modelos de mistura em geral, o processo de maximização da função de log-verossimilhança pode ser efetuado utilizando-se o algoritmo EM (*expectation-maximization*). Esse algoritmo tem a vantagem de dividir o problema de maximização em todo o espaço paramétrico (que pode ter grande dimensão, a depender do número de componentes misturados) em diversos problemas de maximização separados - cada uma dessas maximizações separadas é efetuada em um conjunto de parâmetros com dimensões bem menores. Essa separação do problema de maximização em problemas menores, no caso de um modelo com muitos parâmetros desconhecidos, permite uma maior estabilidade numérica ao algoritmo EM, quando comparado a métodos de otimização do tipo Newton-Raphson ou BFGS (Broyden-Fletcher-Goldfarb-Shanno), por exemplo. O algoritmo EM é mais lento do que algoritmos mais gerais, o que pode ser um problema quando estamos trabalhando com bases de dados com muitas observações.

Pela *própria natureza do algoritmo* EM, ele é especialmente vantajoso quando estamos estimando modelos de mistura com muitos componentes combinados. Em diversos sistemas estatísticos encontrados no mercado, os modelos de mistura combinam duas distribuições apenas, cada qual contendo no máximo dois parâmetros. Portanto, os modelos de mistura possuem no máximo cinco parâmetros livres – dois parâmetros para cada distribuição e o peso do primeiro componente.[9] Em testes realizados pelos autores para testar a performance de diversos algoritmos na estimação dos diversos modelos de mistura, os algoritmos que apresentaram melhor performance foram o BFGS e o algoritmo Simplex (Nelder-Mead).

Nota 7.1 (Escolha dos parâmetros iniciais) Uma das particularidades do problema de maximização da função de log-verossimilhança nos modelos de mistura é que a função a ser maximizada pode ter vários máximos locais. Isso implica que os algoritmos tradicionais de maximização não-linear (EM, Newton-Raphson, BFGS, Simplex) podem ser sensíveis aos valores iniciais utilizados na inicializações desses algoritmos. Portanto, deve-se ter um certo cuidado ao escolher os parâmetros iniciais utilizados nos algoritmos de maximização para encontrar as estimativas dos parâmetros no modelo de distribuições combinadas. Podemos considerar, por exemplo, a escolha de três alternativas para a especificação dos valores iniciais nos algoritmos de maximização. Para ilustrar as três configurações, suponha que estejamos estimando um modelo de mistura onde a primeira distribuição seja uma log-normal (com parâmetros μ e σ) e a segunda distribuição seja uma Weibull (com parâmetros α e β). Além dos parâmetros μ, σ, α e β das distribuições misturadas, é preciso estimar também o peso $p \in (0,1)$ da primeira distribuição. Portanto, ao todo, o algoritmo de maximização tem que encontrar os valores desses cinco parâmetros, que maximizam a função de log-verossimilhança – para isso, precisamos escolher valores de inicialização para esses cinco parâmetros. Seja S a amostra contendo os dados utilizados para estimação dos parâmetros (conforme veremos na nota 7.2 mais adiante, S pode ser na verdade uma subamostra da amostra total dos dados).

Configuração 1. De acordo com a configuração 1, a partir de toda a amostra S, estime os parâmetros μ e σ para a distribuição log-normal individualmente e estime os parâmetros α e β para a distribuição de Weibull, também individualmente. Os valores iniciais escolhidos no algoritmo de maximização serão justamente esses valores estimados para cada distribuição individualmente. Para o peso p, escolha $p = 0.5$.

Configuração 2. Na configuração 2, repartimos a amostra S em duas subamostras S_1 e S_2 com o mesmo número de observações. A primeira subamostra S_1 contém os 50% menores valores de S, enquanto S_2 possui os 50% maiores valores. Estima-se então uma distribuição log-normal individualmente utilizando S_1, obtendo-se μ e σ. Estima-se uma distribuição de Weibull individualmente utilizando S_2, obtendo-se α e β. Esses valores de μ, σ, α e β como valores iniciais no algoritmo de maximização. Para o peso p, escolhe-se $p = 0.5$.

Configuração 3. Finalmente, na configuração 3 para os valores iniciais no algoritmo de estimação dos modelos de mistura, a distribuição log-normal (primeira distribuição) individual seria estimada para a

[9] O peso da segunda distribuição é obviamente um menos o peso do primeiro componente.

amostra S_2 (vide configuração 2 acima) e a distribuição de Weibull (segunda distribuição) individual seria estimada para a amostra S_1. Novamente, o peso p terá valor inicial $p = 0.5$.

Em alguns sistemas comerciais disponíveis no mercado, o usuário pode indicar que configuração de parâmetros iniciais ele deseja utilizar (configuração 1, 2 ou 3).

Nota 7.2 (Utilização de subamostras) Mesmo escolhendo algoritmos que permitam maior rapidez nas estimações, ainda assim o processo de maximização da função de log-verossimilhança pode ser lento, quando a estimação estiver sendo efetuada sobre bases de dados com muitas observações. Para alguns tipos de fraudes bancárias, por exemplo, a base de dados de cinco anos de perda pode conter mais de 500 mil observações de severidade. A partir das propriedades assintóticas dos estimadores de máxima verossimilhança para modelos paramétricos em geral (sob certas condições de regularidade), as estimativas dos parâmetros convergem para os verdadeiros valores[10] desses parâmetros à medida que o número de observações na amostra vai para infinito. Portanto, é de se esperar que as estimativas para os parâmetros obtidas com 500 mil observações sejam bem parecidas com as estimativas para os parâmetros obtidas com uma amostra representativa com 100 mil observações, por exemplo. Baseando-se nesse fato, o pesquisador pode escolher trabalhar com subamostras, ao invés de trabalhar com a base de dados completas, nas estimações dos parâmetros. Esse artifício pode acelerar sensivelmente as estimações.

A partir da amostra total disponível na base de dados, pode-se selecionar uma subamostra aleatória dessa amostra total, e realizar as estimativas dos parâmetros com base nessa subamostra. Com isso, obtêm-se boas estimativas para os parâmetros no modelo de mistura, com uma velocidade de processamento bem maior. Alguns softwares permitem ainda realizar amostragem estratificada das observações da base de dados total. Seja f a fração amostral na subamostra: se $f = 0.10$, então a subamostra contém 10% dos dados originais. É possível que a base de dados contenha as observações ordenadas por data de coleta da informação, por exemplo. Por conta desse ordenamento dos dados, é possível que haja pequenas diferenças entre as observações no início da amostra e as observações no final dela. Portanto, pode haver "estratos"de informações ao longo da base de dados, e o processo de seleção da subamostra pode ser refinado por meio da utilização de uma amostragem estratificada. Suponhamos então que a amostra esteja disponível para cinco anos de dados, e a base total contenha um milhão de observações. Seja s o número de estratos no processo de escolha aleatória da subamostra. Se $s = 5$, então o processo de amostragem consiste nos passos a seguir.

[10]Os parâmetros estimados convergem para os parâmetros verdadeiros no caso de estarmos supondo que o os dados observados pertencem ao modelo paramétrico estimado; ou seja, no caso de o modelo estar corretamente especificado. Porém, em geral, o modelo paramétrico estimado é apenas uma aproximação para o processo gerador dos dados (modelo real) e, portanto, o modelo paramétrico estaria mal especificado. Felizmente, pode-se provar, dadas algumas condições de regularidade, que o parâmetro estimado converge, quando o número de observações na amostra vai para infinito, para o valor do parâmetro que melhor aproxima o processo gerador dos dados, de acordo com alguma medida de distância (ou pseudodistância) entre funções. No caso específico de estimação via máxima verossimilhança, e a pseudodistância que está sendo minimizada é a pseudodistância de Kullback-Leibler. Portanto, mesmo quando o modelo paramétrico está erradamente especificado, as estimativas obtidas via máxima verossimilhança convergem para um valor do parâmetro que melhor aproxima o processo gerador dos dados reais. Para mais detalhes, vide White (1996) e Burnham e Anderson (1998).

(1) Dividir o total de um milhão de observações na base total em cinco subgrupos de 200 mil observações cada. Dado que a base está ordenada por data da coleta da informação, o primeiro subgrupo possui as primeiras 200 mil observações (possivelmente ocorridas no primeiro ano de coleta dos dados), enquanto o quinto subgrupo possui as últimas 200 mil observações (possivelmente ocorridas no último ano de coleta dos dados).

(2) Em cada um dos cinco subgrupos, selecionam-se aleatoriamente subamostras de $f \times 200.000 = 20.000$ observações.

(3) Finalmente, a subamostra total (de 100.000 observações) utilizada nas estimações consiste na união das cinco subamostras de 20.000 observações obtidas em cada um dos cinco estratos. A vantagem de utilizar o esquema de amostragem estratificada é que isso garante que a subamostra de dados utilizada nas estimações conterão observações representando mais uniformemente todo o período de coleta dos dados históricos (cinco anos, por exemplo).

Para ilustrar a validade da utilização de modelos de distribuições combinadas para modelar dados arbitrários, simulamos uma base de dados que corresponde a uma combinação entre uma variável aleatória de Rayleigh e uma variável aleatória exponencial. Em seguida, ajustamos um modelo combinado, que corresponde a uma combinação entre uma variável aleatória lognormal e uma variável aleatória gamma. Portanto, estamos utilizando um modelo mal especificado (lognormal e gamma) para representar os dados reais (exponencial e Rayleigh). Depois de ajustar os parâmetros via máxima verossimilhança, plotamos os gráficos para checar a qualidade do ajuste.

As Figuras 7.8 e 7.9 apresentam os gráficos correspondentes ao ajuste do modelo combinado. No gráfico superior da Figura 7.8, temos o histograma dos dados originais (exponencial e Rayleigh), e no gráfico inferior da mesma figura temos o histograma gerado a partir do modelo teórico ajustado (gamma e lognormal). Observe que tanto os dados originais, quanto os dados simulados a partir da distribuição combinada apresentam histogramas bem parecidos, apesar da distribuição combinada que está sendo ajustada aos dados não coincidir com a distribuição combinada original. A Figura 7.9 apresenta os gráficos da distribuição empírica, do PP-plot e do QQ-plot para avaliar a qualidade do ajuste da distribuição combinada. Os gráficos corroboram a hipótese de que a utilização das distribuições combinadas apresenta um bom ajuste aos dados originais.

No caso mais geral, para modelos combinados com J distribuições individuais, a expressão geral para a função densidade de probabilidade (ou função frequência) é dada por

$$f_c(x;\theta) = \sum_{j=1}^{J} p_j f_j(x;\theta_j),$$

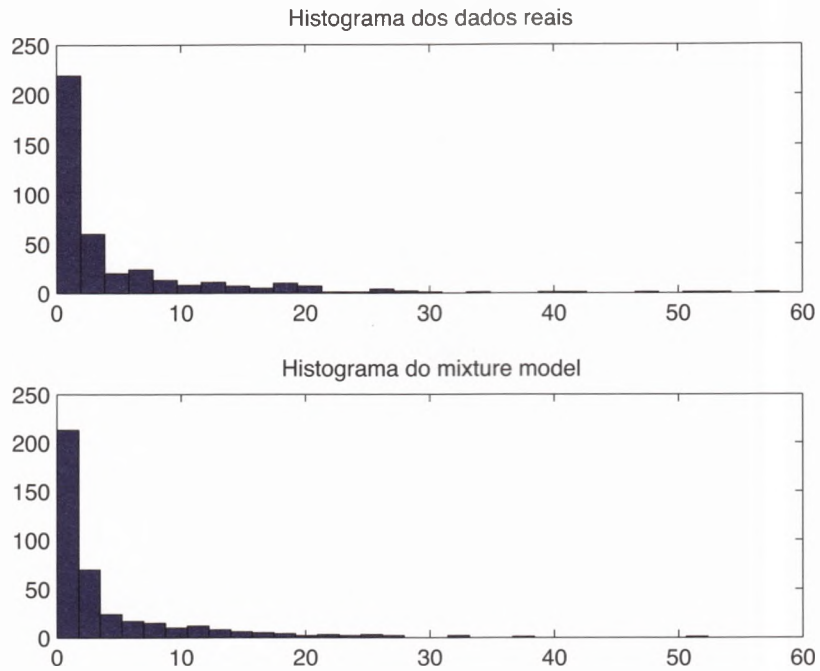


Figura 7.8: Histograma dos dados reais e dos modelo de distribuições combinadas.

onde p_j, $j = 1, \ldots, J$, correspondem aos pesos de cada densidade individual $f_j(x; \theta_j)$, com $\sum_{i=1}^{J} p_j = 1$, e $p_j \in (0, 1)$. O grande vetor de parâmetros θ corresponde à união de todos os parâmetros θ_j individuais mais os pesos $p_1, \ldots, p_{J-1}$. O processo de geração de um número aleatório X a partir de um modelo de mistura com J componentes é totalmente análogo ao processo de dois estágios visto acima. A única diferença é que, com $J > 2$, no primeiro passo, ao invés de gerar um número Z a partir de uma variável aleatória de Bernoulli, o valor de Z é gerado a partir de uma variável aleatória multinomial, com J valores possíveis, onde a probabilidade de tirar cada um dos J valores é exatamente p_j. Se $Z = k$, com $k \in \{1, \ldots, J\}$, então geramos finalmente uma observação aleatória a partir da distribuição cuja função densidade (ou função frequência) é igual a $f_j(x; \theta_j)$. A estimação dos parâmetros é também efetuada via máxima verossimilhança, com utilização do algoritmo EM. A função distribuição acumulada possui expressão

$$F_c(x; \theta) = \sum_{j=1}^{J} p_j F_j(x; \theta_j).$$

Conforme vimos acima, a utilização de modelos de mistura permite a composição de modelos bastante robustos e flexíveis tanto para variáveis aleatórias discretas, quanto para variáveis aleatórias contínuas. Alguns softwares disponíveis no mercado possuem um conjunto de funções específicas para estimação e testes desses tipos de modelos. Além de modelar processos estocásticos puramente contínuos ou puramente discretos, os modelos combinados podem ser empregados também para modelar processos mistos. Por exemplo, imagine que queremos modelar observações de renda domiciliar em uma determinada região no Brasil. É possível que a base de dados contenha um grande percentual de domicílios com renda nula. Um

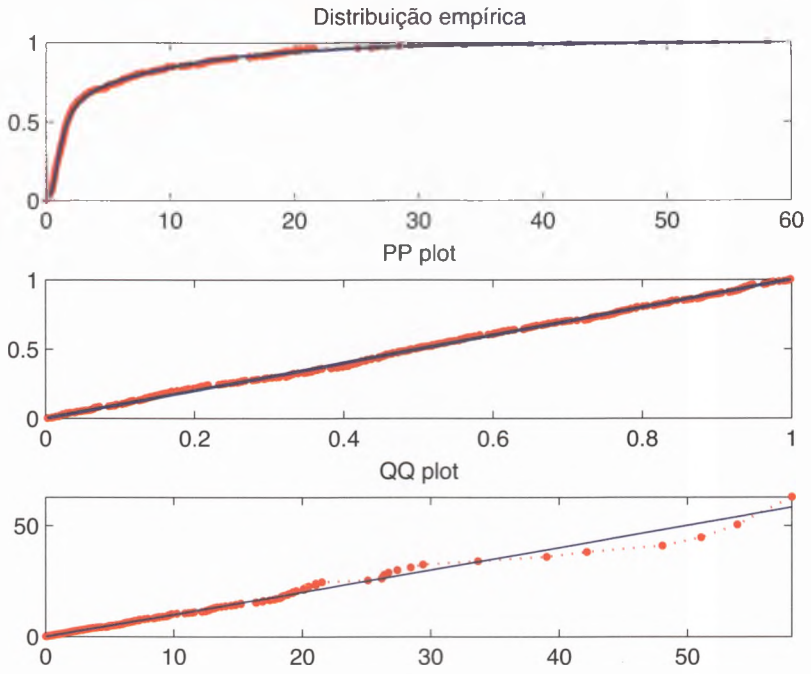


Figura 7.9: Gráficos para análise da qualidade do ajuste da distribuição combinada.

modelo para modelar tal base de dados pode ser construído por meio da combinação de uma distribuição concentrada no ponto zero e um modelo para variáveis contínuas.

7.4.2 Distribuições por subintervalos

Os modelos de distribuição combinada vistos na seção anterior supõem que ambas as distribuição incluídas na combinação atuam em todo o espaço amostral $\mathbb{X}$ dos dados.[11] Nesta seção, apresentamos outra maneira de combinar distribuições, que consiste em dividir o intervalo dos dados em subintervalos e ajustar uma distribuição simples (vista nas Seções 3.2 e 3.3) para cada um desses subintervalos. Por exemplo, no caso de distribuições para modelos de perdas monetárias, podemos ajustar uma distribuição gamma, ou uma distribuição de valores extremos, para valores até R$ 50,000.00 e uma distribuição de valores extremos para valores acima de R$ 50,000.00.

Para facilitar a descrição dos modelos por subintervalos, vamos considerar um modelo onde dividimos o espaço amostral $\mathbb{X}$ em 3 subintervalos. Com isso, é possível ilustrar os passos computacionais referentes a subintervalos no início, no meio e no final do espaço amostral. Com base na discussão para três intervalos, é possível estender os modelos por subintervalos para tratar um número maior de subintervalos. Obviamente, quanto maior o número de subintervalos utilizados, maior deverá ser a base de dados para haver graus de

[11] No caso de modelos para severidade de perdas operacionais, por exemplo, o espaço amostral $\mathbb{X}$ é igual ao lado positivo da reta real, ou seja, $\mathbb{X} = \Re_+^*$. Para modelos de frequência de perdas, o espaço amostral $\mathbb{X}$ é igual ao conjunto $\{0, 1, 2, 3, \dots\}$.

liberdade suficientes para estimar todos os parâmetros envolvidos. Para modelos para variáveis aleatórias contínuas, a função densidade pode ser escrita como

$$\begin{aligned} f_p(x;\theta) = &\frac{p_1}{A_1} f_1(x;\theta_1) I_{(0,c_1]}(x) \\ &+ \frac{p_2}{A_2} f_2(x-c_1;\theta_2) I_{(c_1,c_2]}(x) \\ &+ \frac{p_3}{A_3} f_3(x-c_2;\theta_3) I_{(c_2,+\infty)}(x), \end{aligned} \tag{7.26}$$

onde θ_i corresponde ao vetor de parâmetros da função densidade $f_i(x;\theta_i)$, ajustada ao subintervalo i, $i = 1, 2, 3$. Os valores c_1 e c_2 correspondem ao dois cortes que delimitam o três subintervalos. Alguns sistemas automaticamente escolhem cortes iguais aos percentis da distribuição da amostra, de forma que cada subintervalo escolhido tenha o mesmo número de observações na amostra. O usuário pode posteriormente alterar esses valores de corte, de forma a obter melhores ajustes. As funções $I_{(0,c_1]}(x)$, $I_{(c_1,c_2]}(x)$ e $I_{(c_2,+\infty)}(x)$ são funções indicadoras, com $I_V(x) = 1$ se x pertence ao conjunto V, e $I_V(x) = 0$ caso contrário. Os valores p_i, $i = 1, 2, 3$, correspondem aos pesos de cada intervalo, com $p_1 + p_2 + p_3 = 1$ e $p_i \in (0, 1)$. Os valores A_i, $i = 1, 2, 3$, têm expressões

$$\begin{aligned} A_1 &= \int_{u=0}^{u=c_1} f_1(u;\theta_1)du, \\ A_2 &= \int_{u=0}^{u=c_2-c_1} f_2(u;\theta_2)du, \\ A_3 &= \int_{u=0}^{u=+\infty} f_3(u;\theta_3)du. \end{aligned} \tag{7.27}$$

A inclusão das constantes A_1, A_2, A_3 garantem que a função densidade $f_p(x;\theta)$ do modelo por subintervalos tenha integral igual a 1 em todo o espaço amostral $\Re_+^*$. O vetor total de parâmetros θ do modelo por subintervalos é a união de todos os parâmetros para cada modelo separadamente; ou seja, $\theta = [\theta_1' \ \theta_2' \ \theta_3' \ c_1 \ c_2 \ p_1 \ p_2]'$ (o peso p_3 é consequência direta dos demais dois pesos, com $p_3 = 1 - p_2 - p_1$).

No caso de variáveis aleatórias discretas, o tratamento é completamente análogo. Nesse caso, temos que substituir as funções densidade na Eq. (7.26) pelas funções frequência

$$\begin{aligned} f_p(x;\theta) = &\frac{p_1}{A_1} f_1(x;\theta_1) I_{[0,c_1]}(x) \\ &+ \frac{p_2}{A_2} f_2(x-c_1;\theta_2) I_{(c_1,c_2]}(x) \\ &+ \frac{p_3}{A_3} f_5(x-c_2;\theta_3) I_{(c_2,+\infty)}(x). \end{aligned} \tag{7.28}$$

Note que o primeiro intervalo $[0, c_1]$, correspondente à função $I_{[0,c_1]}(x)$, é fechado em ambos os extremos, de forma a conter o valor zero. As constantes A_1, A_2 e A_3 no caso discreto, passam a ter expressões

$$\begin{aligned} A_1 &= \sum_{u \leq c_1} f_1(u;\theta_1), \\ A_2 &= \sum_{0 < u \leq c_2 - c_1} f_i(u;\theta_2), \\ A_3 &= \sum_{u > 0} f_3(u;\theta_3). \end{aligned} \tag{7.29}$$

A partir de função densidade do modelo por subintervalos apresentada na Eq. (7.26), podemos escrever a expressão para a função distribuição acumulada $F_p(x;\theta)$

$$\begin{aligned} F_p(x;\theta) = &\frac{p_1}{A_1} F_1(x;\theta_1) I_{(0,c_1]}(x) \\ &+ \left[p_1 + \frac{p_2}{A_2} F_2(x - c_1;\theta_2)\right] I_{(0,c_2-c_1]}(x) \\ &+ \left[p_1 + p_2 + \frac{p_3}{A_3} F_3(x - c_2;\theta_3)\right] I_{[0,+\infty)}(x). \end{aligned} \tag{7.30}$$

Conforme discutido para o caso dos modelos de mistura de distribuições, na seção anterior, um dos objetivos de se ajustar modelos estatísticos teóricos aos dados empíricos é poder posteriormente gerar números aleatórios, a partir do modelo ajustado, para utilizá-los nas simulações de Monte Carlo. A geração de um número aleatório a partir do modelo por subintervalos, da mesma maneira que no caso de mistura de distribuições, também é efetuada em dois estágios. Os passos para a geração de um número aleatório X a partir da distribuição por subintervalos são:

1. Gerar um número aleatório a partir de uma variável aleatória multinomial $Z \in \{1, 2, 3\}$, com probabilidade $\text{Prob}[Z = i] = p_i$, $i = 1, 2, 3$, onde p_i é o peso da distribuição no intervalo i.

2. O segundo estágio da geração do número aleatório dependerá do número gerado para Z. Caso $Z = 1$, no segundo estágio geraremos um número aleatório W a partir da distribuição no primeiro subintervalo. Note que essa distribuição deverá ser truncada, de forma que o número gerado se mantenha dentro do subintervalo que está sendo considerado. Analogamente, caso $Z = 3$, no segundo estágio geraremos um número aleatório W a partir da distribuição no terceiro subintervalo.

3. Finalmente, o número final X será uma função do número W, de forma que $X = W$ se $Z = 1$, e $X = W + c_{i-1}$, para $Z = i$, $i = 2$ ou $i = 3$.

Para a estimação dos parâmetros θ_i, $i = 1, 2, 3$, de cada subintervalo, considere a amostra disponível $S = \{X_1, X_2, \ldots, X_n\}$, de tamanho n. Dividimos então essa amostra em 3 subamostras S_i, $i = 1, 2, 3$, de

forma que a subamostra S_1 contenha os valores em $S \cap (0, c_1]$,[12] S_2 contenha os valores em $S \cap (c_1, c_2]$, e S_3 contenha os valores em $(c_2, +\infty)$. Finalmente, o vetor de parâmetros θ_i, da distribuição do subintervalo i, é estimada com base na amostra S_i, $i = 1, 2, 3$, utilizando o estimador de máxima verossimilhança.

Na estimação via máxima verossimilhança para modelos por subintervalos, duas estratégias podem ser empregadas. Vamos imaginar que queremos estimar a distribuição do segundo intervalo, e nesse caso utilizaremos a subamostra S_2, contendo os valores em $S \cap (c_1, c_2]$. Para isso, pode-se estimar o vetor de parâmetros θ_2 maximizando-se a função de log-verossimilhança tradicional, supondo-se que os dados seguem um modelo paramétrico simples (sem se preocupar com fato de os dados serem de fato truncados, restritos ao subintervalo considerado). Nesse caso, tem-se a estimativa do parâmetro θ_2,

$$\hat{\theta}_{2\text{MV}} = \arg\max_{\theta_2 \in \Theta} \sum_{x_i \in S_2} \log f(x_i - c_1; \theta_2). \tag{7.31}$$

Observe a notação $\sum_{x_i \in S_2}$ para especificar que entram no somatório apenas as observações x_i que pertencem à subamostra S_2. Alternativamente, dada a natureza de distribuição truncada típica dos modelos por subintervalos, pode-se estimar o vetor de parâmetros θ_2 a partir da maximização da função de log-verossimilhança truncada

$$\hat{\theta}_{2\text{trunc}} = \arg\max_{\theta_2 \in \Theta} \sum_{x_i \in S_2} \log \left[\frac{f(x_i - c_1; \theta_2)}{F(c_2 - c_1; \theta_2)} \right], \tag{7.32}$$

onde $F(c_2 - c_1; \theta_2)$ é a função distribuição acumulada no comprimento do intervalo $c_2 - c_1$. Para o primeiro intervalo, o estimador de máxima verossimilhança truncada para o parâmetro θ_1 tem expressão

$$\hat{\theta}_{1\text{trunc}} = \arg\max_{\theta_1 \in \Theta} \sum_{x_i \in S_1} \log \left[\frac{f(x_i; \theta_1)}{F(c_1; \theta_1)} \right]. \tag{7.33}$$

Finalmente, para o terceiro subintervalo, o estimador de máxima verossimilhança truncada é dado por

$$\hat{\theta}_{3\text{trunc}} = \arg\max_{\theta_3 \in \Theta} \sum_{x_i \in S_3} \log \left[\frac{f(x_i - c_2; \theta_1)}{F(+\infty - c_2; \theta_3)} \right], \tag{7.34}$$

abusando da notação quando escrevemos $F(+\infty - c_2; \theta_3)$. Obviamente, $F(+\infty - c_2; \theta_3) = F(+\infty; \theta_3) = 1$, e concluímos que o estimador de máxima verossimilhança truncada é exatamente o estimador de máxima verossimilhança tradicional, aplicado aos valores $x_i - c_2$, para todo $x_i \in S_3$. Portanto, $\hat{\theta}_{3\text{MV}} = \hat{\theta}_{3\text{trunc}}$.

Alguns sistemas permitem ao usuário optar por qual função de log-verossimilhança (truncada ou não) será maximizada, para obtenção dos vetores de parâmetros livres em cada subintervalo. Com isso, o usuário

[12] Esses intervalos correspondem aos modelos para variáveis aleatórias contínuas. Para os modelos para variáveis aleatórias discretas, os intervalos têm que ser ligeiramente modificados.

tem mais uma alternativa para escolher o modelo que mais se adequa aos dados observados. Para selecionar o modelo, estimado por um dos dois métodos de log-verossimilhança, o analista pode recorrer aos diversos métodos de testes de ajustes. Via de regra, os modelos por subintervalos constituem-se em mais uma ferramenta poderosa para modelagem de observações discretas e contínuas em diversas aplicações, dada a sua grande flexibilidade em capturar diferentes características dos dados.

7.5 Exercícios

Exercício 7.1 Seja uma X uma variável aleatória exponencial negativa, com função densidade

$$f(y) = \lambda e^{-\lambda y}, \quad \text{para } y \in (0, \infty).$$

Determine a função distribuição acumulada inversa $x = F^{-1}(u)$ para a distribuição exponencial negativa, onde $u \in (0, 1)$. A função distribuição acumulada inversa é tal que, caso se $x = F^{-1}(u)$, então $u = F(x)$.

Exercício 7.2 Seja X uma variável aleatória de Weibull, com função densidade

$$f(y) = \beta \alpha^{-\beta} y^{\beta-1} e^{-(y/\alpha)^{\beta}}, \quad \text{para } y \in (0, \infty).$$

Determine a função distribuição acumulada inversa $x = F^{-1}(u)$ para a distribuição de Weibull, onde $u \in (0, 1)$.

Exercício 7.3 Considere uma amostra aleatória com valores observados $x_1, \ldots, x_n$. As observações são independentes e identicamente distribuídas, com distribuição beta, com parâmetros livres α e β.

(i) Escreva a função de log-verossimilhança para a amostra considerada.

(ii) Escreva a expressão para o critério AIC.

(iii) Escreva a expressão para o critério BIC.

Exercício 7.4 Considere uma amostra aleatória com valores observados $x_1, \ldots, x_n$. As observações são independentes e identicamente distribuídas, com distribuição de Weibull, com parâmetros livres α e β.

(i) Escreva a função de log-verossimilhança para a amostra considerada.

(ii) Escreva a expressão para o critério AIC.

(iii) Escreva a expressão para o critério BIC.

II

Modelos de regressão

8. Modelos de regressão linear

"In inner-city, low-income communities of color, there's such a high correlation in terms of educational quality and success."
Bill Gates

Neste capítulo introduziremos o **modelo de regressão linear**, que é possivelmente o modelo paramétrico mais simples que pode ser usado para relacionar um grupo de variáveis aleatórias. O modelo de regressão linear, além de ser o modelo estatístico mais utilizado em economia e em finanças, é a base para o entendimento de vários outros modelos paramétricos que relacionam variáveis aleatórias tais como modelos de séries temporais (HAMILTON, 1994; ENDERS, 2003; MORETTIN; TOLOI, 2006; BUENO, 2008), modelos de painel de dados (HSIAO, 2003; BALTAGI, 2008) e modelos de regressão binária e classificação que serão discutidos no Capítulo 9.

Embora o modelo de regressão linear seja muito usado e existam inúmeras aplicações dele, provavelmente uma das aplicações mais conhecidas desse modelo em economia e finanças é o teste do modelo de equilíbrio de mercado chamado CAPM (*Capital Asset Pricing Model*), apresentado na Aplicação 3.2. Como vimos, de forma muito simples, esse modelo descreve a relação que existe entre o retorno de um ativo financeiro e o seu risco, ou seja,

$$E[R_j] = R_f + \beta_j(E[R_m] - R_f), \tag{8.1}$$

onde $E[R_j]$ é o retorno esperado do ativo j, R_f é o retorno do ativo livre de risco (por exemplo, o retorno de um título do tesouro americano), $E[R_m]$ é o retorno esperado de mercado (por exemplo, o retorno do SP500 ou do Bovespa) e β_j é uma medida de risco conhecida como beta do CAPM, risco sistêmico ou risco não diversificável, que é o risco que não pode ser eliminado do ativo j mesmo em uma carteira bem diversificada. Note que na equação acima se $E[R_j] = E[R_f]$ então $\beta_j = 0$, se $E[R_j] = E[R_m]$ então $\beta_j = 1$. Então para ativos onde $0 < \beta_j < 1$, $R_f < E[R_j] < E[R_m]$ e para ativos onde $\beta_j > 1$, $E[R_j] > E[R_m]$. Com o objetivo de testar empiricamente a Eq. (8.1), ela pode ser reescrita como

$$E[R_{j,i}] = R_f + \beta_j(E[R_{m,i}] - R_f) + u_i,\ i = 1, \cdots, T, \tag{8.2}$$

ou ainda como

$$E[Z_{j,i}] = \alpha_j + \beta_j E[Z_{m,i}] + u_i,\ i = 1, \cdots, T, \tag{8.3}$$

onde T é o tamanho da amostra, $Z_{j,i} = R_{m,i} - R_f$ é definido como o excesso de retorno e u_i é uma variável aleatória que representa o erro do modelo, para todo $i = 1, \cdots, T$. Diz-se que o CAPM representa bem

o ativo j no mercado se $\alpha_j = 0$. O modelo apresentado acima na Eq. (8.3) é chamado de **modelo de regressão linear simples**, visto que ele além do termo constante tem apenas um regressor, Z_m.

Exemplo 8.1 (CAPM) Como aprendemos na Seção 5.4, usando simulações Monte Carlo, vamos gerar uma amostra de retornos para um ativo financeiro j supondo que o ativo j satisfaz o CAPM apresentado na Eq. (8.2). Para esse fim, vamos supor que $T = 250$, que é aproximadamente o número de dias úteis de um ano, que os retornos diários[1] são dados por $R_f = 2 \times 10^{-4}$, $R_m \sim \text{Normal}[\mu_m, \sigma_m^2]$, onde $\mu_m = 0.54$ e $\sigma_j^2 = 2.6 \times 10^{-4}$ e $u \sim \text{Normal}[\mu_u, \sigma_u^2]$, onde $\mu_u = 0$ e $\sigma_u^2 = 1 \times 10^{-5}$ e finalmente que $\beta_j = 0.8$. Essas variáveis geradas são apresentadas na Figura 8.1. Note que nessa figura é explícito o caráter linear da relação entre as variáveis y_i e $Z_{m,i}$ para $i = 1, \cdots, T$.

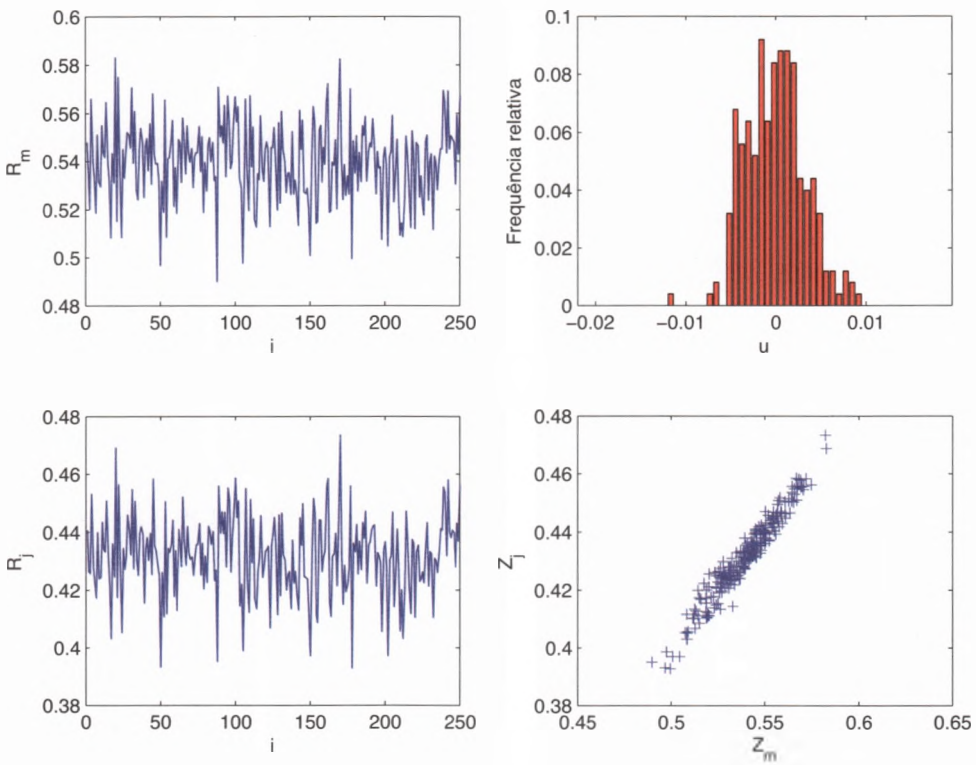


Figura 8.1: Retornos simulados de um ativo financeiro usando o CAPM.

Uma extensão da relação estatística dada pela Eq. (8.2), que não está diretamente relacionada com as ideias de equilíbrio de mercado apresentadas no modelo do CAPM descrito pela Eq. (8.1), é supor que o retorno de um ativo j depende de vários fatores, como por exemplo em

$$E[R_{j,i}] = \alpha_i + \beta_{j,1} F_{1,i} + \beta_{j,2} F_{2,i} + \cdots + \beta_{j,L} F_{L,i} + u_i, \; i = 1, \cdots, T, \tag{8.4}$$

[1] Para fins didáticos supusemos que os retornos financeiros de mercado são variáveis aleatórias normais. Na verdade, nem sempre isso ocorre (MANTEGNA; STANLEY, 1999; LO; MACKINLAY, 2001).

fator. Diferentemente, do modelo apresentado na Eq. (8.3), esse modelo, chamado de modelo de vários fatores, possui mais do que um regressor, além do termo constante, e por isso é chamado de **modelo regressão linear múltipla**. Para mais detalhes sobre modelos econométricos aplicados a finanças, o leitor interessado deve consultar Campbell, Lo e MacKinlay (1996), Gourieroux e Jasiak (2001), Cochrane (2005), Cuthbertson e Nitzsche (2005) e Morettin (2008).

Obviamente vários outros trabalhos têm tratado do modelo de regressão linear e esse capítulo sem dúvida foi influenciado por alguns deles como Amemiya (1985), Ruud (2000), Hayashi (2000) e Davidson e MacKinnon (2004). A maior diferença desse capítulo para essas referências é que enquanto nesse capítulo a apresentação da teoria é preferencialmente intuitiva, nessas outras referências a apresentação é preferencialmente formal.

Este capítulo continua da seguinte forma: na Seção 8.1 explicitamos as hipóteses básicas do modelo de regressão linear e as implicações dessas hipóteses. Na Seção 8.2 discutimos a estimação do modelo de regressão linear usando três métodos diferentes: o método de mínimos quadrados, o método de momentos já introduzido na Seção 5.1 e o método de estimação via máxima verossimilhança, já apresentado na Seção 5.2. Além disso, enquanto na Seção 8.2.1 discutimos como formular o teste de hipóteses para o modelo de regressão linear utilizando as propriedades do estimador de mínimos quadrados e a hipótese de normalidade, na Seção 8.2.3 apresentamos o teste de hipóteses para o modelo de regressão linear no contexto do estimador de máxima verossimilhança e supondo que estamos trabalhando com amostras grandes, como já feito na Seção 6.4.2.

8.1 Hipóteses do modelo de regressão linear

Nesta seção explicitamos as hipóteses que estão por trás do modelo de regressão linear. A primeira hipótese dessa seção concerne a respeito da linearidade em relação às variáveis do modelo.

Hipótese 8.1 (Linearidade) Seja y_i a i-ésima observação da chamada variável dependente (ou resposta, ou explicada, ou predita) e $x_i = [x_{i1}, x_{i2}, \cdots, x_{iK}]'$ a i-ésima observação dos K regressores (ou variáveis independentes, explicativas, preditoras ou covariáveis).[2] Então y_i está relacionado com x_i de acordo com

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_K x_{iK} + u_i \quad (i = 1, 2, \cdots n) \tag{8.5}$$

onde n é o número de observações e u_i é o resíduo ou erro do modelo de regressão linear.

[2] Aqui nesse texto seguindo a literatura mais moderna de econometria que inclui o Amemiya (1985), Ruud (2000), Hayashi (2000) e Wooldridge (2001), considera-se que tanto as variáveis dependentes como as variáveis independentes são variáveis aleatórias – fato que está de acordo com a situação mais usual em economia, onde a maioria das bases de dados são construídas a partir de situações reais e não de experimentos controlados.

Neste capítulo, vamos trabalhar principalmente com a notação matricial da Eq. (8.5), muito comum em todos os textos modernos de econometria:

$$y = X\beta + u, \tag{8.6}$$

onde y é o vetor coluna de observações de ordem K, X é a matriz de dados de ordem $n \times K$, β é o vetor coluna de regressores de ordem K e u é o vetor coluna de resíduos. Assim sendo,

$$y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \quad X = \begin{bmatrix} x_1' \\ \vdots \\ x_n' \end{bmatrix} = \begin{bmatrix} x_{11} & \cdots & x_{1K} \\ \vdots & \cdots & \vdots \\ x_{n1} & \cdots & x_{nK} \end{bmatrix} \quad \beta = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_n \end{bmatrix} \text{ e } u = \begin{bmatrix} u_1 \\ \vdots \\ u_n \end{bmatrix}. \tag{8.7}$$

Exemplo 8.2 (Modelos lineares de apreçamento de ativos) Como vimos, os modelos de apreçamento de ativos apresentados nas Eqs. (8.3) e (8.4) satisfazem a hipótese de linearidade. Na Eq. (8.3), $K = 2$. O primeiro regressor é $x_{i1} = 1$ e $x_{i2} = Z_{m,i}$, para $i = 1, \cdots, T$. Na Eq. (8.4), temos $x_{i1} = 1$ e $x_{i,l+1} = F_{l,i}$, para $l = 1, \cdots, L$, $K = L$ e $i = 1, \cdots, T$.

Existem outras inúmeras possibilidades de aplicações do modelo de regressão linear:

Exemplo 8.3 (Modelos de regressão linear com termos polinomiais) Outra forma do modelo de regressão linear é o modelo de regressão polinomial que pode ser expresso da seguinte forma

$$y_i = \alpha_0 + \alpha_1 x_i^1 + \alpha_2 x_i^2 + \cdots + \alpha_m x_i^m + u_i \quad (i = 1, 2, \cdots n). \tag{8.8}$$

Note que nesse caso, o primeiro regressor da regressão apresentada na Eq. (8.5) é dado por uma constante e os outros regressores são as potências de x_i, para $i = 1, \cdots, n$.

Embora o modelo apresentado na Eq. (8.8) tenha perdido caráter linear em relação ao x, ele continua sendo um modelo de regressão linear, pois ele é linear em relação aos regressores x^j, para $j = 1, \cdots, m$.

Exemplo 8.4 (Estimação da função de produção de Cobb-Douglas) Considere uma função de produção de Cobb-Douglas dada por

$$y_i = \alpha_0 z_{i1}^{\alpha_1} \cdots z_{im}^{\alpha_m} \tag{8.9}$$

Aplicando a função log a ambos os lados dessa equação, podemos escrever essa equação num formato linear dado por

$$\log y_i = \log \alpha_0 + \alpha_1 \log z_{i1} + \cdots + \alpha_n \log z_{im} \tag{8.10}$$

Portanto, a função de produção de Cobb-Douglas pode ser estimada como um modelo linear a partir da seguinte equação

$$\log y_i = \log \alpha_0 + \alpha_1 \log z_{i1} + \cdots + \alpha_m \log z_{im} + u_i. \tag{8.11}$$

Nesse caso, os regressores são 1 e os termos $(\log z_{ij})$. para $j = 1, \cdots, m$.

A próxima hipótese conhecida como exogeneidade diz que a melhor previsão que podemos fazer do erro u usando a matriz de dados é supor que o erro é nulo.

Hipótese 8.2 (Exogeneidade) $E[u_i/X] = 0 \quad (i = 1.2, \cdots, n)$.

Note que se a Hipótese 8.2 não fosse válida, isto é, $E[u/X] \neq 0$, isso significa que poderíamos ainda reduzir o erro da regressão usando informação disponível nos dados e, portanto, essa não seria uma especificação interessante.

Utilizando a Hipótese 8.2, a lei das expectâncias iteradas[3] (vide Proposição 4.5) e calculando o valor esperado condicional de y_i de acordo com a Eq. (8.5) em relação a $x_{i1}, x_{i2}, \cdots, x_{iK}$, chegamos a

$$E[y_i/x_{i1}, x_{i2}, \cdots, x_{iK}] = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_K x_{iK}. \tag{8.12}$$

Portanto, o modelo de regressão linear fornece o valor esperado condicional de y_i em relação a $x_{i1}, x_{i2}, \cdots, x_{iK}$. Adicionalmente, se calcularmos a derivada parcial de $E[y_i/x_{i1}, x_{i2}, \cdots, x_{iK}]$ em relação a x_{ij}, para $1 \leq j \leq K$ chegamos a

$$\frac{\partial E[y_i/x_{i1}, x_{i2}, \cdots, x_{iK}]}{\partial x_{ij}} = \beta_j. \tag{8.13}$$

Desse modo, cada coeficiente da regressão β_j, para $1 \leq j \leq K$, mede a taxa de variação do valor esperado $E[y_i/x_{i1}, x_{i2}, \cdots, x_{iK}]$ em relação a x_{ij}.

Nota 8.1 (Variáveis *dummy*) Na discussão feita no início desta seção basicamente consideramos variáveis quantitativas como a variável risco no modelo do CAPM, os fatores de produção na função de produção de Cobb-Douglas (horas de trabalho, capital) que são variáveis contínuas. Modelos de regressão linear podem

[3]Note que a informação gerada pelo conjunto $\{x_{i1}, \cdots, x_{ik}\}$ é menor do que a informação gerada por X. Dessa forma, vale a aplicação dessa proposição.

e usualmente incluem variáveis qualitativas chamadas de variáveis *dummy* que assumem valores 0 e 1. Por convenção, um elemento da população ter o valor de uma variável qualitativa igual a 0 significa que esse elemento da população não possui uma determinada característica. Um elemento da população ter o valor de variável qualitativa igual a 1 significa que ele possui uma determinada característica.

Considere um modelo de regressão linear onde o K-ésimo regressor é uma variável qualitativa

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_{K-1} x_{i\,K-1} + \beta_K x_{iK} + u_i \quad (i = 1, 2, \cdots n).$$

Supondo que a Hipótese 8.2 é válida, então

$$\beta_k = E[y_i / \{x_{i1}, \cdots, x_{i\,K-1}\}, x_{iK} = 1] - E[y_i / \{x_{i1}, \cdots, x_{i\,K-1}\}, x_{iK} = 0],$$

ou seja, a variável qualitativa x_k gerará uma variação esperada no valor do intercepto entre aqueles que possuem ou não a característica.

Exemplo 8.5 (Regressão com variáveis *dummy*) Imagine um modelo linear aplicado a uma determinada população que relacione o salário de uma pessoa na população com o seu tempo de estudo, com a sua experiência e também com o seu gênero (sexo) como em

$$y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + u_i, \tag{8.14}$$

onde x_{i2} =tempo de estudo, x_{i3} =experiência e x_{i4} =gênero.

Em geral, resultados empíricos, pelos menos usando dados do século passado, implicam que homens com as mesmas características que mulheres têm um salário significativamente maior. Entretanto, é possível que hoje, em várias culturas, como a maior inserção da mulher no mercado de trabalho, esse fato tenha sido alterado.

Usando a Hipótese 8.2, também pode-se mostrar usando a lei das expectâncias iteradas (vide Seção 4.5) que $E[u_i] = 0$, isto é, o termo de erro tem valor esperado zero e o modelo de regressão linear não tem viés. Adicionalmente, também usando a Hipótese 8.2 e lei das expectâncias iteradas, pode-se mostrar que $\text{Cov}(x_{jk}, u_i) = 0$, isto é, que a correlação entre o termo de erro e qualquer um dos regressores é zero. Essa hipótese em geral não é válida quando o modelo de regressão linear não está corretamente especificado.

Prática 8.1 Use a Hipótese 8.2 e a lei das expectâncias iteradas e mostre esses dois resultados.

Exemplo 8.6 (O problema de omissão de variáveis) Considere que o processo usado pela natureza para gerar os dados foi baseado no modelo $y = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + u$, onde $u \sim \text{Normal}(0, \sigma_u^2)$, $0 < \text{cov}(x_1, x_2) < 1$,

e, por algum motivo, um econometrista acredita que o modelo correto que relaciona y com x_1 é dado por $y = \beta_0 + \beta_1 x_1 + v$. Note que uma parcela do termo não modelado $\beta_2 x_2$ estará presente no erro e, portanto, no segundo modelo dada a sua má especificação, $cov(v/x_1) \neq 0$, isto é, a Hipótese 8.2 não é satisfeita. No Exemplo 8.19 no final desse capítulo mostramos como podemos detectar esse problema.

Exemplo 8.7 (O problema de má especificação funcional) Considere que o processo usado pela natureza para gerar os dados foi baseado no modelo $y = \alpha_0 + \alpha_1 x^2 + u$, onde $u \sim \text{Normal}(0, \sigma_u^2)$, e, por algum motivo, um econometrista acredita que o modelo correto que relaciona y com x_1 é dado por $y = \beta_0 + \beta_1 x + v$. Note que o segundo modelo nunca conseguirá aproximar "corretamente" o primeiro modelo e, portanto, v será uma função de x obviamente correlacionada com x. Dessa forma, esse modelo também não satisfará a Hipótese 8.2. No Exemplo 8.20 no final desse capítulo mostramos como podemos detectar esse problema.

Embora esse texto não trate a questão da **endogeneidade** (o oposto de exogeneidade) que ocorre quando existe correlação entre um regressor e o termo de erro, a econometria tem encontrado formas para lidar com esse problema. Para detalhes consultar, por exemplo, Ruud (2000), Hayashi (2000) e White (2000).

Uma outra hipótese importante que é necessária para a estimação do modelo de regressão linear é conhecida como ausência de multicolinearidade perfeita entre os dados da matriz X.

Hipótese 8.3 (Ausência de multicolinearidade perfeita) O posto da matriz de dados X de ordem $n \times K$ é K com probabilidade 1.

Basicamente, o que essa hipótese nos diz é o necessário para que o modelo de regressão linear esteja corretamente especificado. Se o posto dessa matriz não é K (se ocorre **multicolinearidade perfeita**), isso significa que pelo menos um dos regressores não está trazendo nenhuma informação nova ao modelo. Considere por um único momento que a matriz de dados X possui posto $K-1$. Então, como estudamos no curso de algebra linear, existe uma coluna na matriz de dados X que é linear dependente de todas as outras $K-1$ colunas da matriz de dados. Dessa forma, essa coluna de dados pode ser escrita como um combinação linear de todas as outras colunas e por isso essa coluna de dados não traz nenhuma nova informação nova para o modelo de regressão linear. Por exemplo, considere que estamos estudando um modelo de regressão linear para explicar o preço de um imóvel. Em geral, uma variável importante que pode ser utilizada para explicar o preço de um imóvel é o tamanho do imóvel, além de outras variáveis tais como localização, idade etc. Então, um dado corretor associa a variável tamanho do imóvel a dois regressores. Um dos regressores é área do imóvel em metros quadrados e o outro regressor é área do imóvel em centímetros quadrados. Note que o segundo regressor não traz nenhuma informação nova, pois ele é apenas um múltiplo do primeiro e ele obviamente deve ser retirado da regressão. A Hipótese 8.3 justamente exclui esses casos. De fato, essa hipótese é fácil de ser detectada na prática, pois a maioria dos softwares econométricos explicitamente indicam esse problema quando ele existe.

A hipótese abaixo ainda tem o objetivo de descrever o tipo de erro que estamos considerando nesse modelo. Basicamente ela supõe que todos os erros possuem a mesma variância (uma hipótese que conhecemos como homocedasticidade) e são independentes.

Hipótese 8.4 (Erro com distribuição esférica) $E[uu'/X] = \sigma^2 I_n$, i.e.,

a) (Homocedasticidade) $E[u_i^2/X] = \sigma^2 > 0$;

b) (Erros não correlacionados) $E[u_i u_j/X] = 0 \quad (i, j = 1, 2, \cdots, n; i \neq j)$

É válido comentar que o termo **erro com distribuição esférica** foi cunhado para descrever matrizes de variância-covariância de variáveis aleatórias que possuem a mesma variância e a covariância entre elas nula, como é a situação apresentada na Hipótese 8.4. Esse termo pode ser justificado pelo fato que matrizes com essas características não se alteram quando sofrem rotações (que podem ser representadas por transformação ortogonais em espaços vetoriais lineares), assim como a esfera. No caso geral onde as variáveis aleatórias não apresentam variância constante e a covariância entre elas é diferente de zero, os erros são ditos possuírem distribuições elípticas. Mais detalhes dessas ideias podem ser encontrados em Ruud (2000).

Usando de definição de covariância e as Hipóteses 8.2 e 8.4 pode-se mostrar que $\text{cov}(u_i, u_j) = 0, \quad i \neq j$. É válido comentar que a Hipótese 8.4 é apenas simplificadora e pode ser relaxada em várias direções, se o processo de estimação dos coeficientes for alterado adequadamente. A situação mais simples é considerar que a variância do erro depende do valor dos regressores – esses modelos são chamados de modelos heterocedásticos – para detalhes ver, por exemplo, Amemiya (1985), Ruud (2000), Hayashi (2000) e Wooldridge (2001). Uma outra situação é considerar que amostras que são coletadas em locais próximos possuem comportamento similar – esses modelos são chamados de modelos de econometria espacial e podem ser encontrado em Anselin (1988) e LeSage e Pace (2009).

8.2 Estimação do modelo de regressão linear

Nesta seção apresentamos as três principais metodologias usadas para estimar o modelo de regressão linear. Veremos que embora conceitualmente as metodologias sejam bem diferentes, os valores encontrados para o vetor de coeficientes β são os mesmos.

8.2.1 Estimação usando o método dos mínimos quadrados

Nesta seção consideraremos o **método dos mínimos quadrados**. Esse método é muito interessante não só pela simplicidade e pela interpretação geométrica, mas também porque esse método pode ser entendido

para várias situações muito mais gerais.[4] De fato, como veremos ainda no fim dessa seção, esse método é baseado na ideia de projetar um vetor de um espaço vetorial em um subespaço, como considerado no teorema da Projeção (LUENBERGER, 1969).

Derivação do estimador de mínimos quadrados

A ideia por trás do método dos mínimos quadrados é minimizar a Soma dos Quadrados dos Resíduos (ou erros) (SQR),[5] isto é, buscar um estimador para o vetor de coeficientes do modelo de regressão linear de forma que

$$\hat{\beta} = \text{argmin}_{\beta} SQR(\beta), \tag{8.15}$$

onde $SQR(\beta) = \sum_{i=1}^{n}(y_i - x_i'\beta)^2 = (y - X\beta)'(y - X\beta)$.

Calculando o quadrado na Eq. (8.15) e usando as Hipótese 8.1, podemos escrever SQR como

$$SQR(\beta) = y'y - 2y'X\beta + \beta'X'X\beta.$$

Calculando a derivada parcial em relação β (vide Exercício 8.1), e igualando a zero, obtemos

$$\frac{\partial SQR(\beta)}{\partial \beta} = -2X'y + 2X'X\beta = 0.$$

Usando a Hipótese 8.3, podemos mostrar que a inversa $X'X$ existe e aplicando a inversa na equação acima concluímos que o estimador de mínimos quadrados é dado por

$$\hat{\beta} = (X'X)^{-1}X'y. \tag{8.16}$$

Prática 8.2 Considere o modelo de regressão linear simples e encontre o estimador de mínimos quadrados dos dois parâmetros desse modelo, minimizando a soma dos quadrados dos erros (não use a Eq. (8.16)).

[4] Por exemplo, os mínimos quadrados são comumente usados para estimação de parâmetros em espaços de funções. Para detalhes, ver Luenberger (1969).

[5] Veremos que essa é uma forma interessante de estimar os coeficientes do modelos de regressão linear, pois apresenta várias propriedades interessantes, mas será que essa é a única forma? De forma mais explícita, por que minimizar uma função quadrática e não outra função similar? Além dessa função apresentar propriedades algébricas interessantes tais como ser positiva para qualquer valor do erro, ser contínua e possuir derivadas contínuas, como veremos nessa seção, o estimador de mínimos quadrados busca os parâmetros estimados $\hat{\beta}$ (coeficientes lineares do modelo de regressão linear) que garantem que $E[\hat{y}/X] = X\hat{\beta}$, isto é, possibilitam exprimir a média condicional de y em relação a X. Por exemplo, se ao invés de minimizarmos a soma dos quadrados dos erros, minimizássemos a soma dos valores absolutos do erro, o modelos de regressão linear não estaria mais exprimindo y como média condicional de X e sim y como quantil condicional de X (KOENKER, 2005).

Exemplo 8.8 (Exemplo numérico de regressão linear múltipla) Considere o modelo de regressão linear dado por

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + u_i \tag{8.17}$$

para $i = 1, \cdots, n$, onde $\beta_0 = 1$, $\beta_1 = -2$ e $\beta_2 = 3$. Vamos novamente usar simulações Monte Carlo e gerar os dados do modelo usando $x_1 \sim \text{Normal}[\mu_{x_1}, \sigma^2_{x_1}]$, onde $\mu_{x_1} = 1$ e $\sigma^2_{x_1} = 0.01$, $x_2 \sim \text{Normal}[\mu_{x_2}, \sigma^2_{x_2}]$, onde $\mu_{x_2} = 2$ e $\sigma^2_{x_2} = 0.001$, $u \sim \text{Normal}[\mu, \sigma^2]$, onde $\mu = 0$ e $\sigma^2 = 0.0001$ e $n = 10$.

Para um determinado conjunto de dados gerados

$$\begin{aligned} X' &= \begin{bmatrix} 1.0000 & 1.0000 & 1.0000 & 1.0000 & 1.0000 & 1.0000 & 1.0000 & 1.0000 & 1.0000 & 1.0000 \\ 0.9943 & 1.0395 & 0.9959 & 1.0885 & 0.8100 & 1.0158 & 0.9536 & 0.9423 & 1.2114 & 0.8732 \\ 1.9819 & 2.0202 & 1.9857 & 1.9793 & 1.9787 & 2.0216 & 2.0772 & 1.9885 & 2.0136 & 2.0012 \end{bmatrix} \\ y' &= \begin{bmatrix} 4.9549 & 4.9780 & 4.9690 & 4.7449 & 5.3063 & 5.0279 & 5.3188 & 5.0672 & 4.6103 & 5.2718 \end{bmatrix}, \end{aligned}$$

encontramos

$$\begin{aligned} X'X &= \begin{bmatrix} 10.0000 & 9.9246 & 20.0479 \\ 9.9246 & 9.9612 & 19.9004 \\ 20.0479 & 19.9004 & 40.2002 \end{bmatrix}, (X'X)^{-1} = \begin{bmatrix} 482.3177 & -0.9970 & -240.0391 \\ -0.9970 & 9.1020 & -4.0086 \\ -240.0391 & -4.0086 & 121.7171 \end{bmatrix}, \\ X'y &= \begin{bmatrix} 50.2491 \\ 49.6557 \\ 100.7568 \end{bmatrix} \text{ e } \hat{\beta} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = (X'X)^{-1}X'y = \begin{bmatrix} 0.9492 \\ -2.0236 \\ 3.0348 \end{bmatrix} \end{aligned}$$

É muito comum em livros elementares de econometria exprimir primeiro resultados para um modelo de regressão linear simples e depois estender para o caso de regressão linear múltipla. No exemplo que segue faremos o contrário e apresentamos como calcular o estimador usualmente encontrado em livros introdutórios de estatística a partir da Eq. (8.16).

Exemplo 8.9 (Estimador de mínimos quadrados para o caso da regressão linear simples com intercepto) Considere, por exemplo, que temos um modelo de regressão linear simples dado por

$$y_i = \beta_0 + \beta_1 x_{i1} + u_i \tag{8.18}$$

para $i = 1, \cdots, n$.

Para calcular β_0 e β_1, usaremos a Eq. (8.16) como ponto de partida e calcularemos os termos dessa equação fazendo

$$X = \begin{bmatrix} 1 & x_{11} \\ 1 & x_{21} \\ \vdots & \vdots \\ 1 & x_{n1} \end{bmatrix} \quad y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}.$$

Usando a definição de X e y, começamos então calculando $X'X$ (vide Exercício 8.3)

$$X'X = \begin{bmatrix} n & \sum_{i=1}^n x_{i1} \\ \sum_{i=1}^n x_{i1} & \sum_{i=1}^n x_{i1}^2 \end{bmatrix}$$

e sua inversa (vide Exercício 8.4)

$$\begin{aligned} (X'X)^{-1} &= \frac{1}{det(X'X)} \begin{bmatrix} \sum_{i=1}^n x_{i1}^2 & -\sum_{i=1}^n x_{i1} \\ -\sum_{i=1}^n x_{i1} & n \end{bmatrix} \\ &= \frac{1}{n\sum_{i=1}^n x_{i1}^2 - (\sum_{i=1}^n x_{i1})^2} \begin{bmatrix} \sum_{i=1}^n x_{i1}^2 & -\sum_{i=1}^n x_{i1} \\ -\sum_{i=1}^n x_{i1} & n \end{bmatrix}. \end{aligned}$$

Usando a definição de X e y, calculamos o termo $X'y$ encontrando

$$X'y = \begin{bmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_{i1} y_i \end{bmatrix}$$

Então, usando todos esses cálculos, encontramos β dado por

$$\begin{aligned} \hat{\beta} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix} &= \frac{1}{n\sum_{i=1}^n x_{i1}^2 - (\sum_{i=1}^n x_{i1})^2} \begin{bmatrix} \sum_{i=1}^n x_{i1}^2 \sum_{i=1}^n y_i - \sum_{i=1}^n x_{i1} \sum_{i=1}^n x_{i1} y_i \\ -\sum_{i=1}^n x_{i1} \sum_{i=1}^n y_i + n\sum_{i=1}^n x_{i1} y_i \end{bmatrix} \\ &= \begin{bmatrix} \overline{y} - \beta_1 \overline{x} \\ \frac{n\sum_{i=1}^n x_{i1} y_i - \sum_{i=1}^n x_{i1} \sum_{i=1}^n y_i}{n\sum_{i=1}^n x_{i1}^2 - (\sum_{i=1}^n x_{i1})^2} \end{bmatrix} \end{aligned}$$

onde $\overline{x} = \sum_i^n x_{i1}/n$ e $\overline{y} = \sum_i^n y_i/n$.

Prática 8.3 Detalhe os cálculos feitos implicitamente no Exemplo 8.9.

Prática 8.4 Utilizando a Eq. (8.16), calcule agora o estimador de mínimos quadrados para o caso de regressão linear simples sem intercepto.

Propriedades amostrais do estimador de mínimos quadrados

Nesta seção discutiremos as propriedades amostrais do estimador de mínimos quadrados. O termo "amostral" aqui se refere ao fato que essas propriedades são válidas mesmo em amostras pequenas. Defina o valor estimado de y como $\hat{y} = X\hat{\beta}$. Então, o vetor de resíduos dado por $\hat{u} = y - \hat{y} = y - X\hat{\beta}$ é uma estimativa do termo de erro u no modelo de regressão linear apresentado na Eq. (8.5). Uma vez que nunca teremos acesso aos verdadeiros valores de β e consequentemente ao valor de u, o vetor de resíduos estimados é uma variável importante para se analisar a qualidade de um modelo de regressão linear. Define-se a estimativa do método de mínimos quadrados para a variância do termo de erro u como[6]

$$\hat{\sigma}^2 = \frac{SQR}{n-K}, \tag{8.19}$$

onde, como vimos implicitamente na derivação do estimador de mínimos quadrados, $SQR = \hat{u}'\hat{u}$.

Outro vetor de erros importante é o vetor de erros que mede o erro entre a estimativa de mínimos quadrados dos coeficientes de regressão linear e os verdadeiros coeficiente dado por $\hat{\beta} - \beta$. Embora na prática nunca teremos acesso ao valor real dessa variável, fazendo algumas manipulações algébricas podemos mostrar que o erro de estimação é dado por[7]

$$\hat{\beta} - \beta = (X'X)^{-1}X'u. \tag{8.20}$$

Exemplo 8.10 (Continuação do Exemplo 8.8 – Estimativa da variância σ^2) Utilizando a Eq. (8.19), estimamos $SQR = 6.2986 \times 10^{-4}$. Portanto, $\hat{\sigma}^2 = \frac{SQR}{n-K} = 6.2986 \times 10^{-4}/(10-3) = 8.9980e \times 10^{-}5$.

As proposições abaixo apresentam as propriedades do estimador dos mínimos quadrados em amostras finitas:

Proposição 8.1 (O estimador de mínimos quadrados é não viesado) Sob as Hipóteses 8.1, 8.2 e 8.3, $E[\hat{\beta}/X] = \beta$.

A Proposição 8.1 afirma que o estimador de mínimos quadrados é não viesado. Embora para outras classes de estimadores nem seja sempre possível provar essa propriedade, essa é uma propriedade muito importante que se espera de um estimador.[8]

[6] Note que a divisão por $n-K$ segue o mesmo propósito da divisão por $n-1$ no Capítulo 2, quando introduzimos o conceito de variância.

[7] Vide Exercício 8.11.

[8] Em geral, diz-se que para um estimador ser útil, ele pelo menos deve ser consistente, isto é, não viesado em amostras muito grandes, como vimos na Seção 5.3.

Proposição 8.2 (Estimativa da variância do estimador de mínimos quadrados) Sob as Hipóteses 8.1, 8.2, 8.3 e 8.4, $\text{var}(\hat{\beta}/X) = \sigma^2(X'X)^{-1}$.

A Proposição 8.2 calcula explicitamente o valor da variância de $\hat{\beta}$, que será muito útil para a especificação dos testes de hipóteses que serão apresentados a seguir.

Proposição 8.3 (Teorema de Gauss-Markov) Sob as Hipótese 8.1, 8.2, 8.3 e 8.4, o estimador de mínimos quadrados é eficiente[9] na classe de estimadores lineares em y.

A Proposição 8.3, também conhecida como **teorema de Gauss-Markov** nos diz que, para o modelo de regressão linear, não é possível encontrar um estimador linear não viesado com menor variância.

Proposição 8.4 (O estimador de mínimos quadrados da variância é não viesado) Sob as Hipóteses 8.1, 8.2, 8.3 e 8.4 e $n > K$, $E[\hat{\sigma}^2/X] = \sigma^2$.

A Proposição 8.4 diz que a estimativa de mínimos quadrados da variância do termo de erro também é não viesada.

A prova das proposições acima podem ser encontradas em Amemiya (1985), Ruud (2000), Hayashi (2000) para o caso de regressão linear múltipla[10] e em Gujarati (2000) para o caso de regressão linear simples.

Notando que o estimador de mínimos quadrados é calculado apenas por um método de "força bruta" (minimização da soma dos quadrados dos erros) é surpreendente que esse estimador possua propriedades tão interessantes como aquelas apresentadas nas Proposições 8.1 a 8.4. Mais que isso, essas propriedades são válidas para qualquer distribuição dos resíduos, desde que sejam satisfeitas as hipóteses apresentadas na Seção 8.1.

Exemplo 8.11 (Continuação do Exemplo 8.8 – Variância da estimativa de mínimos quadrados de β) De acordo com a Proposição 8.2, a variância de β pode ser calculada usando $\text{var}(\hat{\beta}/X) = \sigma^2(X'X)^{-1}$. Uma vez que não temos σ^2, um procedimento usual é substituir esse valor por sua estimativa $\hat{\sigma}^2$. O valor dessa variância é então chamada variância da estimativa de mínimos quadrados de β. Então usando o valor de $\hat{\sigma}^2$ calculada no Exemplo 8.10, chegamos a

$$\widehat{\text{var}}(\hat{\beta}/X) = \hat{\sigma}^2(X'X)^{-1} = \begin{bmatrix} 0.0434 & -0.0001 & -0.0216 \\ -0.0001 & 0.0008 & -0.0004 \\ -0.0216 & -0.0004 & 0.0110 \end{bmatrix}. \tag{8.21}$$

[9]Isso significa que se $\hat{\tilde{\beta}}$ é outro estimador linear não viesado de β então $\text{var}(\hat{\tilde{\beta}}/X) > \text{var}(\hat{\beta}/X)$. Lembre que discutimos ideias de eficiência de estimadores nas Seções 5.1 e 5.2.

[10]Veja também Exercício 8.13 no fim desse capítulo.

Então a $\widehat{\text{var}}(\hat{\beta}_1/X) = \widehat{\text{var}}(\hat{\beta}/X)_{11} = 0.0434$, $\widehat{\text{var}}(\hat{\beta}_2/X) = \widehat{\text{var}}(\hat{\beta}/X)_{22} = 0.0008$ e $\widehat{\text{var}}(\hat{\beta}_3/X) = \widehat{\text{var}}(\hat{\beta}/X)_{33} = 0.0110$.

Exemplo 8.12 (Continuação do Exemplo 8.8 - simulações de Monte Carlo para ilustrar o fato que os estimadores $\hat{\beta}$ e $\hat{\sigma}^2$ são não viesados) Como vimos na Seção 5.4, podemos utilizar simulações Monte Carlo para ilustrar a habilidade de estimadores. Então, utilizando o modelo do Exemplo 8.8, fizemos 10,000 simulações de amostras de tamanho 10 e encontramos a média das amostras de β_1 igual a 0.9984, a média das amostras de β_2 igual a -1.9997, a média das amostras de β_3 igual a 3.0006 e a média das amostras de $\hat{\sigma}^2$ igual a 9.9789×10^{-5}. A Figura 8.2 apresenta os histogramas dos 10,000 valores estimados dessas variáveis.

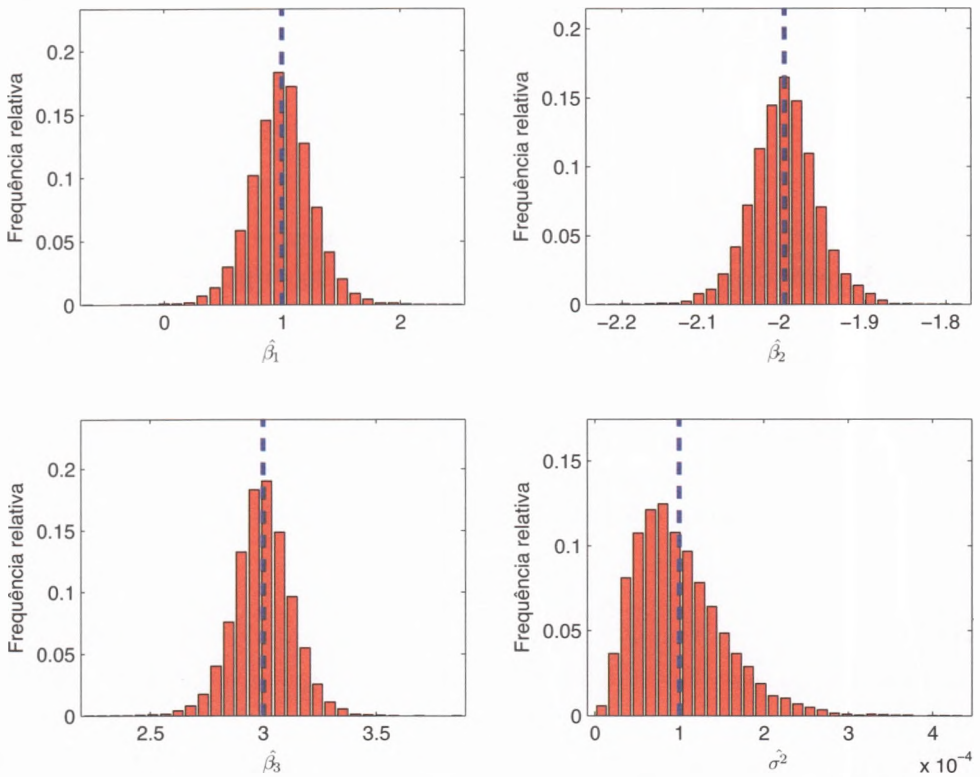


Figura 8.2: Histogramas dos 10,000 valores estimados dos parâmetros $\hat{\beta}$ e $\hat{\sigma}^2$.

Interpretação geométrica do estimador de mínimos quadrados

O estimador de mínimos quadrados possui uma interpretação geométrica muito esclarecedora. Sabemos que o valor estimado de y é dado por $\hat{y} = X\hat{\beta}$. Podemos reescrever essa equação da seguinte forma

$$\begin{bmatrix} \hat{y}_1 \\ \vdots \\ \hat{y}_n \end{bmatrix} = \begin{bmatrix} x_{11} & \cdots & x_{1K} \\ \vdots & \ldots & \vdots \\ x_{n1} & \cdots & x_{nK} \end{bmatrix} \begin{bmatrix} \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_K \end{bmatrix} = \begin{bmatrix} \hat{\beta}_1 x_{11} + & \cdots & + \hat{\beta}_K x_{1K} \\ \vdots & \ldots & \vdots \\ \hat{\beta}_1 x_{n1} + & \cdots & + \hat{\beta}_K x_{nK} \end{bmatrix}$$

$$= \hat{\beta}_1 \begin{bmatrix} x_{11} \\ \vdots \\ x_{n1} \end{bmatrix} + \cdots + \hat{\beta}_K \begin{bmatrix} x_{1K} \\ \vdots \\ x_{nK} \end{bmatrix}. \tag{8.22}$$

Portanto, é fácil verificar que $\hat{y}$ é escrito por uma combinação linear das colunas da matriz X.

Considere agora a definição do vetor de resíduos dado por $\hat{u} = y - \hat{y} = y - X\hat{\beta}$. Substituindo o valor de $\hat{\beta}$ nessa equação, chegamos a $\hat{u} = y - X(X'X)^{-1}X'y$. Multiplicando ambos os lados dessa equação por X', chegamos a $X'\hat{u} = X'y - X'X(X'X)^{-1}X'y = 0$. Note que $X'\hat{u}$ pode ser reescrito por meio de K equações da seguinte forma

$$\begin{array}{c} \sum_{i=1}^{n} x_{i1}\hat{u}_i = 0 \\ \sum_{i=1}^{n} x_{i2}\hat{u}_i = 0 \\ \vdots \\ \sum_{i=1}^{n} x_{iK}\hat{u}_i = 0. \end{array} \tag{8.23}$$

Lembrando do curso de algebra linear, se o produto interno[11] entre os vetores $\begin{bmatrix} \hat{u}_1 \\ \vdots \\ \hat{u}_n \end{bmatrix}$ e $\begin{bmatrix} x_{1j} \\ \vdots \\ x_{nj} \end{bmatrix}$ para $j = 1, \cdots, K$ é nulo (como mostrado acima), então eles são ortogonais.

Dessa forma, geometricamente, o estimador de mínimos quadrados estima os coeficientes do modelo de regressão linear fazendo uma projeção do vetor y, dada por $\hat{y} = X\hat{\beta}$, no subespaço gerado pelos vetores $\begin{bmatrix} x_{1j} \\ \vdots \\ x_{nj} \end{bmatrix}$, $j = 1, \cdots, K$, que são as colunas da matriz X, de forma que o erro de estimação seja ortogonal a esse subespaço. Fazendo isso, esse estimador garante que $E[\hat{y}/X] = X\hat{\beta}$, visto que $\hat{u}$ é ortogonal a (independente de) X.

[11] Em matemática básica usualmente chamamos esse produto de produto escalar.

Teste de hipóteses supondo normalidade do termo de erro

Até agora no estudo do estimador de mínimos quadrados não precisamos fazer nenhuma hipótese explícita a respeito da distribuição do erro. Entretanto, nesse momento, com o objetivo de introduzir o teste de hipóteses para o estimador de mínimos quadrados em problemas de amostras finitas, precisaremos introduzir a hipótese abaixo que concerne a respeito da distribuição do termo de erro.

Hipótese 8.5 (Normalidade do termo de erro) A distribuição do termo de erro u condicional a matriz de dados X é uma distribuição normal conjunta.

No teorema que segue apresentamos o teste t que é muito útil para testar hipóteses a respeito dos coeficientes do modelo de regressão linear.

Teorema 8.1 (Teste t) Suponha que as Hipóteses 8.1, 8.2, 8.3, 8.4 8.5 são válidas, então sob a hipótese nula $\mathrm{H}_0 : \beta_k = \beta_k^0$, a razão T definida como

$$T_k = \frac{\hat{\beta}_k - \beta_k^0}{\sqrt{\hat{\sigma}^2 (X'X)^{-1}_{kk}}}, \tag{8.24}$$

associada a cada parâmetro β_k, tem distribuição t-Student com $n - K$ graus de liberdade.

Apesar de não provar o Teorema 8.1 formalmente aqui, podemos fazer alguns esclarecimentos sobre esse resultado. Como vimos na Proposição 8.1, $E[\hat{\beta}/X] = \beta$. Portanto, usando esse resultado e o fato que sob a hipótese nula β^0 é o valor verdadeiro e portanto constante, temos que $E[\hat{\beta} - \beta^0/X] = 0$. Sabemos da Proposição 8.2 que $\text{var}(\hat{\beta}/X) = \sigma^2(X'X)^{-1}$. Dessa forma, usando a Hipótese 8.5, podemos concluir que $\hat{\beta}_k - \beta_k^0/X$ tem distribuição normal com média 0 e variância $\sigma^2(X'X)^{-1}$ (vide Eq. (8.20) e Exemplo 4.17). Então a razão

$$z_k = \frac{\hat{\beta}_k - \beta_k^0}{\sqrt{\sigma^2 (X'X)^{-1}_{kk}}} \tag{8.25}$$

tem distribuição normal com média 0 e variância 1. Infelizmente, a razão z_k não pode ser usada para fazer o teste de hipóteses visto que a variância σ^2 não é conhecida. Enfrentamos uma situação parecida na Seção 6.4 quando introduzimos testes de hipótese para a média populacional com variância desconhecida. Como lá, temos uma estimativa para σ^2 dada por $\hat{\sigma}^2 = \frac{SQR}{n-K}$ que podemos usar no lugar de σ^2. Então a ideia é substituir o valor real de σ^2 por sua estimativa, mas quando fazemos isso a distribuição da razão z_k não é mais normal pois $\hat{\sigma}^2$ não é constante e sim uma variável aleatória. Dessa forma, o que o Teorema 8.1 mostra é que a razão

$$T_k = \frac{z_k}{\sqrt{\hat{\sigma}^2/\sigma^2}} = \frac{z_k}{\sqrt{\frac{SQR/\sigma^2}{n-K}}}$$

tem uma distribuição t-Student com $n-k$ graus de liberdade (vide Exemplo 4.23), pois z_k tem distribuição normal, SQR/σ^2 tem uma distribuição qui-quadrada com $n-K$ graus de liberdade (vide Exercício 8.14) e z_k e SQR/σ^2 são independentes.

Nota 8.2 (Cálculo do p-valor no teste t) Como vimos na Seção 6.4.3, podemos calcular o p-valor como

$$\begin{aligned} p &= 1 - F_{t_{n-K}}(t_k) \text{ para } H_0 : \beta_k \leq \beta_k^0, \\ p &= F_{t_{n-K}}(t_k) \text{ para } H_0 : \beta_k \geq \beta_k^0, \\ p &= 2 \times (1 - F_{t_{n-K}}(|t_k|)) \text{ para } H_0 : \beta_k = \beta_k^0, \end{aligned} \tag{8.26}$$

onde $F_{t_{n-K}}$ é a função de distribuição acumulada para uma variável aleatória t-Student com $n-K$ graus de liberdade e t_k corresponde ao valor da estatística teste.

A única diferença aqui em relação a Seção 6.4.3 é que o número de graus de liberdade depende do número de regressores do modelo.

Nota 8.3 (Cálculo do intervalo de confiança) Usando o Teorema 8.1, o intervalo de confiança para $\hat{\beta}_k$ pode ser calculado da mesma forma que na Seção 6.5, ou seja,

$$\text{IC}_{\alpha\%} = \left[\hat{\beta}_k - t_{(n-K),(1-\alpha/2)\%} \times \sqrt{\text{var}(\hat{\beta}/X)_{kk}},\ \hat{\beta}_k + t_{(n-K),(1-\alpha/2)\%} \times \sqrt{\text{var}(\hat{\beta}/X)_{kk}}\right],$$

onde $1-\alpha$ é a probabilidade de cobertura.

Exemplo 8.13 (Continuação do Exemplo 8.8 – Teste de hipóteses usando o teste t) Usando o Teorema 8.1, agora podemos levar adiante testes de hipótese, seguindo a mesma metodologia apresentada na Seção 6.4, sobre o coeficiente de regressão linear $\hat{\beta}$. Por exemplo, suponha que estamos interessados em testar se o coeficiente relacionado com a variável x_2 é maior ou igual que -1 – ou seja, a hipótese nula é $\beta_2 \geq -1$ e a hipótese alternativa é $\beta_2 < -1$. Então seguindo os passos introduzidos na Seção 6.4, o primeiro passo é calcular a estatística teste que é dada pela Eq. (8.24). Logo, usando o valor estimado de β_2 no Exemplo 8.8, $\hat{\beta}_2 = -2.0236$, o valor estimado para $\text{var}(\hat{\beta}/X)_{22}^{-1} = 0.0008$ no Exemplo 8.11, e o menor valor de β_2 para que a hipótese nula seja válida identificamos $\beta_k^0 = -1$. Substituindo esses na Eq. (8.24), chegamos a

$$t_2 = \frac{-2.0236 - (-1)}{\sqrt{0.008}} = -35.7668.$$

Então o p-valor dado por $p = F_{t_{10-3}}(t_2) = 1.7339 \times 10^{-9} \approx 0$ rejeitando a hipótese nula.

Podemos usar também o Teorema 8.1 para calcular o intervalo de confiança para o coeficiente β_2. Seguindo a Nota 8.3, o intervalo de confiança desse parâmetro ao nível de 95% é dado por

$$\text{IC}_{95\%} = [\hat{\beta}_2 - t_{(n-K),97.5\%} \times \sqrt{\text{var}(\hat{\beta}/X)_{22}},\ \hat{\beta}_2 + t_{(n-K),97.5\%} \times \sqrt{\text{var}(\hat{\beta}/X)_{22}}] = [-2.0470, -2.0001].$$

Outra ferramenta muito útil para testar hipóteses no modelo de regressão linear é conhecido como teste F, pois ele diferentemente do teste t permite testar hipóteses conjuntas de vários coeficientes de regressão linear. No teorema que segue apresentamos esse resultado.

Teorema 8.2 (Teste-F) Suponha as Hipóteses 8.1, 8.2, 8.3, 8.4 8.5, então sob a hipótese nula $\text{H}_0 : R\beta = r$, onde R é uma matriz $l(r) \times K$ com $\text{rank}(R) = l(r)$ e $l(r)$ denota o número de linhas do vetor r, a razão F definida como

$$\begin{aligned} F &= \frac{(R\hat{\beta} - r)'[R(X'X)^{-1}R']^{-1}(R\hat{\beta} - r)/l(r)}{\hat{\sigma}^2} \\ &= (R\hat{\beta} - r)'[R\widehat{\text{var}}(\hat{\beta}/X)R']^{-1}(R\hat{\beta} - r)/l(r) \end{aligned} \tag{8.27}$$

é distribuída como $F(l(r), n - K)$.

Note que a razão F pode ser reescrita como

$$F = \frac{(R\hat{\beta} - r)'[\sigma^2 R(X'X)^{-1}R']^{-1}(R\hat{\beta} - r)/l(r)}{\frac{SQR/\sigma^2}{n-K}}.$$

Dessa forma, a ideia por trás do Teorema 8.2 fica simples. Explicando de forma resumida, ele mostra que as variáveis aleatórias no numerador $(R\hat{\beta} - r)'[\sigma^2 R(X'X)^{-1}R']^{-1}(R\hat{\beta} - r)/l(r)$ e no denominador SQR/σ^2 possuem distribuição qui-quadrada com graus de liberdade respectivamente iguais a l(r) e $n - K$ (vide Exercícios 8.14 e 8.16) e que a razão entre essas variáveis aleatórias divididas pelos seus graus de liberdade possuem distribuição F com $l(r)$ graus de liberdade no numerador e $n - K$ graus de liberdade no denominador (vide Exemplo 4.24).

Nota 8.4 (Cálculo do p-valor no teste F) Para o caso do teste F apenas faz sentido testar a hipótese nula H_0 ou a sua rejeição. Nesse caso, o p-valor é calculado como

$$p = (1 - F_{F_{l(r),n-K}}(f)) \text{ para } H_0 \text{ ser verdadeira.} \tag{8.28}$$

onde $F_{F_{l(r),n-K}}$ é a função de distribuição acumulada para uma variável aleatória F com com $l(r)$ graus de liberdade no numerador e $n - K$ graus de liberdade no denominador e f corresponde ao valor da estatística teste.

A vantagem do teste F é que ele pode assumir várias formas úteis, como mostra o exemplo a seguir:

Exemplo 8.14 (Continuação do Exemplo 8.8 – Exemplo prático do teste F) Uma aplicação útil do Teorema 8.2 é testar se todos os parâmetros (com exceção da constante) de uma regressão linear são conjuntamente nulos. Então retornando ao Exemplo 8.8. nosso objetivo é testar a hipótese nula H_0 que $\beta_1 = 0$ e $\beta_2 = 0$. Então, fazendo $R = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$ e r é igual ao vetor coluna nulo de ordem 2, achamos $f = 2.7128\text{e} \times 10^3$ e o p-valor dado por $p = 1 - F_{F_{2,n-3}}(f) = 7.6793 \times 10^{-11}$ rejeitando a hipótese de que os dois regressores podem ser conjuntamente nulos.

Vamos agora testar se os coeficientes desse modelo de regressão linear satisfazem $-\beta_0 + \beta_1 + \beta_2 = 0$. Então, devemos fazendo $R = [\ -1 \quad 1 \quad 1\ ]$ e $r = 0$, achamos $f = 0.0393$ e o p-valor dado por $p = 1 - F_{F_{1,n-3}}(f) = 0.8484$ não sendo possível rejeitar a hipótese nula para os níveis de significância usuais de 10%, 5% e 1%.

Exemplo 8.15 (Uma outra forma do teste F apresentado no Teorema 8.2) O teste F que normalmente aparece em livros básicos de econometria é dado por

$$F = \frac{(SQR(\tilde{\beta}) - SQR(\beta))/l(r)}{SQR(\beta)/(n-K)}, \tag{8.29}$$

onde $\tilde{\beta}$ é o estimador de mínimos quadrados para o modelo restrito, $SQR(\tilde{\beta})$ é a soma dos quadrados dos erros para o modelo restrito e $SQR(\beta)$ é a soma dos quadrados dos erros para o modelos irrestrito.

De acordo com a Eq. (8.19), sabemos que $\hat{\sigma}^2 = SQR/(n-K)$. Logo, a única coisa que precisa ser mostrada é que $(R\hat{\beta} - r)'[R(X'X)^{-1}R']^{-1}(R\hat{\beta} - r)/l(r) = (SQR(\tilde{\beta}) - SQR)/l(r)$. Para conseguir a identidade que desejamos mostrar, precisamos calcular o valor de $\tilde{\beta}$ que é o coeficiente do modelo de regressão linear para o caso com restrição. Então precisamos resolver um problema similar aquele apresentado na Eq. (8.15), ou seja,

$$\tilde{\beta} = \text{argmin}_{\beta} \frac{1}{2} SQR(\beta) \text{ s.a. } R\tilde{\beta} = r.$$

Para resolvermos o problema acima, visto que é um problema de otimização estática restrito, precisamos construir o Lagrangiano associado a esse problema que é dado por

$$\mathcal{L} = \frac{1}{2}(y - X'\tilde{\beta})'(y - X'\tilde{\beta}) + \delta'(R\tilde{\beta} - r), \tag{8.30}$$

onde δ é o multiplicador de Lagrange associado à restrição $R\tilde{\beta} = r$. Derivando o Lagrangiano em relação a $\tilde{\beta}$ e δ chegamos as seguintes condições de primeira ordem dadas por

$$X'y - (X'X)\tilde{\beta} = R'\delta \tag{8.31}$$

$$R\hat{\beta} = r. \tag{8.32}$$

Multiplicando a Eq. (8.31) por $R(X'X)^{-1}$, chegamos a conclusão que

$$\delta = (R(X'X)^{-1}R')^{-1}(R\hat{\beta} - r),$$

usando $\hat{\beta} = (X'X)^{-1}X'y$ e sabendo que $R(X'X)^{-1}R'$ possui inversa.

Substituindo o valor de δ na Eq. (8.31) chegamos ao valor de $\tilde{\beta}$

$$\tilde{\beta} = \hat{\beta} - (X'X)^{-1}R'((R(X'X)^{-1}R')^{-1}(R\hat{\beta} - r)). \tag{8.33}$$

Definindo $\tilde{u} = y - X\tilde{\beta}$ e fazendo $SQR(\tilde{\beta}) = \tilde{u}'\tilde{u} = (y - X\hat{\beta}) + X(\hat{\beta} - \tilde{\beta})$ chegamos a

$$SQR(\tilde{\beta}) = (y - X\hat{\beta})'(y - X\hat{\beta}) + (\hat{\beta} - \tilde{\beta})'X'X(\hat{\beta} - \tilde{\beta}), \tag{8.34}$$

visto que $X'(y - X\hat{\beta}) = 0$.

Portanto, substituindo o valor de $\tilde{\beta}$ dado pela Eq. (8.33) na Eq. (8.34) chegamos a

$$SQR(\tilde{\beta}) - SQR(\hat{\beta}) = (R\hat{\beta} - r)'[R(X'X)^{-1}R']^{-1}(R\hat{\beta} - r).$$

Prática 8.5 Detalhe as contas do Exemplo 8.15.

8.2.2 Estimação usando o método de momentos

É possível encontrar um estimador com as mesmas propriedades que o apresentado na Seção 8.2.1 usando o método de momentos apresentado na Seção 5.1.

Derivação do estimador baseado no método dos momentos

Como já vimos na Seção 5.1 o método de momentos baseia-se num procedimento que iguala os momentos populacionais aos momentos amostrais. Dessa forma, usando a Hipótese 8.2 o estimador deverá resolver

$$\begin{aligned} \sum_{i=1}^{n} x_{i1}(y_i - x_i'\beta) &= 0 \\ \sum_{i=1}^{n} x_{i2}(y_i - x_i'\beta) &= 0 \\ &\vdots \\ \sum_{i=1}^{n} x_{iK}(y_i - x_i'\beta) &= 0 \end{aligned} \tag{8.35}$$

que representam $E[x_{jk}u_i/X] = 0$.

Pode-se mostrar que esse sistema de equações pode ser reescrito de forma a recuperar o estimador de mínimos quadrados apresentado na Eq. (8.16). A propósito, note que esse estimador explicitamente recupera a Eq. (8.23).

8.2.3 Estimação usando máxima verossimilhança

Como vimos na Seção 6.4.2 para proceder com o método de estimação via máxima verossimilhança, precisamos estabelecer uma distribuição para o termo de erro. Então nesta seção, além das Hipóteses 8.1 (o modelo é linear), 8.2 (exogeneidade do termo de erro), 8.3 (ausência de multicolinearidade perfeita) e 8.4 (erro com distribuição esférica) usadas para a estimação do modelo de regressão linear usando o método de mínimos quadrados e métodos de momentos, consideramos a Hipótese 8.5 que requer a normalidade do termo de erro $u = y - X\beta$.

Derivação do estimador de máxima verossimilhança

Dessa forma, supondo que o modelo é linear (Hipótese 8.1), a distribuição do erro é esférica (Hipótese 8.4) e o termo de erro é normal $y/X \sim N(x\beta, \sigma^2 I)$ (Hipótese 8.5), então:

$$f(y/X) = (2\pi\sigma^2)^{-n/2} \exp\left\{-\left(\frac{1}{2\sigma^2}(y - X\beta)'(y - X\beta)\right)\right\}. \tag{8.36}$$

Portanto,

$$\log L(\beta, \sigma^2/X) = l(\beta, \sigma^2/X) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}(y - X\beta)'(y - X\beta). \tag{8.37}$$

Como na Seção 6.4.2 para encontrar os estimadores que maximizam a função de máxima verossimilhança, procedemos calculando a primeira derivada em relação aos parâmetros (o coeficiente da regressão β e a variância σ^2) do modelo de regressão linear

$$\frac{\partial \log L(\theta)}{\partial \theta} = \begin{bmatrix} \frac{\partial l(\beta,\sigma^2)}{\partial \beta} \\ \frac{\partial l(\beta,\sigma^2)}{\partial \sigma^2} \end{bmatrix} = \begin{bmatrix} \frac{1}{\sigma^2} X'(y - X\beta) \\ -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}(y - X\beta)'(y - X\beta) \end{bmatrix}. \tag{8.38}$$

Então, igualando a primeira equação da matriz apresentada na Eq. (8.38) a zero e supondo que $\hat{\sigma}^2$, o valor estimado de σ^2, é diferente de zero, encontramos exatamente o estimador $\hat{\beta} = (X'X)^{-1}X'y$ já apresentado na Eq. (8.16).

Igualando a segunda equação da matriz apresentada na Eq. (8.38) a zero e usando a definição de SQR, encontramos que o estimador de máxima verossimilhança $\hat{\sigma}^2$ de σ^2 dado por

$$\hat{\sigma}^2 = \frac{1}{n}(y - X\beta)'(y - X\beta) = \frac{1}{n} SQR(\beta). \tag{8.39}$$

Não é surpreendente o resultado de que os estimadores de mínimos quadrados e de máxima verossimilhança de β são os mesmos. Se considerarmos o procedimento conhecido como **máxima-verossimilhança concentrada** onde primeiro encontramos o valor de um dos estimadores e depois substituímos o valor desse estimador na função de máxima verossimilhança para encontrar o valor dos outros estimadores, podemos enxergar explicitamente esse resultado. Então, substituindo o valor de $\hat{\sigma}^2$ na função de máxima verossimilhança, chegamos à

$$\log L(\beta/X) = l(\beta/X) = -\frac{n}{2}\log(2\pi) - \frac{n}{2} - \frac{n}{2}\log\left(\frac{1}{n} SQR(\beta)\right), \tag{8.40}$$

que mostra que maximizar a máxima verossimilhança é equivalente a minimizar $SQR(\beta)$. Portanto, o estimador de máxima verossimilhança e o estimador de mínimos quadrados é o mesmo.

Vale a pena comentar também que os estimadores de máxima verossimilhança e de mínimos quadrados de σ^2 são diferentes. Lembrando da Proposição 8.4, esse resultado implica que o estimador de máxima verossimilhança de σ^2 é viesado. Entretanto, a única diferença entre esses dois estimadores é o denominador onde no caso do estimador de mínimos quadrados é dado por $n - K$ e no caso de máxima verossimilhança é dado por n. Portanto, a diferença entre esses dois estimadores reduz quando n aumenta, e quando n é muito grande praticamente não há diferença entre os valores desses dois estimadores. Dessa forma, para grandes amostras, o estimador de máxima verossimilhança de σ^2 é também não viesado. Uma vez que essa propriedade só é válida para grandes amostras, o estimador de máxima verossimilhança de σ^2, como vimos na Seção 5.3, é **consistente**.

Como vimos na Seção 6.4.2, para levar adiante o teste de hipóteses no contexto de estimação via máxima verossimilhança, precisamos calcular a **matriz de informação de Fisher**. Perceba que a matriz de informação de Fisher adaptada para o problema de estimação de parâmetros do modelo de regressão linear é dada por

$$I(\theta) = I(\beta, \sigma^2/X) = -E\left[\frac{\partial^2}{\partial\theta\partial\theta'} l(\beta, \sigma^2)/X\right] = -E\left[\begin{array}{cc} \frac{\partial^2 l(\beta,\sigma^2)}{\partial\beta\partial\beta'} & \frac{\partial^2 l(\beta,\sigma^2)}{\partial\beta\partial\sigma^2} \\ \frac{\partial^2 l(\beta,\sigma^2)}{\partial\sigma^2\partial\beta} & \frac{\partial^2 l(\beta,\sigma^2)}{\partial^2\sigma^2} \end{array} \Big/ X\right]. \tag{8.41}$$

Utilizando a Eq. (8.38), podemos calcular as derivadas segundas como

$$\frac{\partial^2 l(\beta, \sigma^2)}{\partial\beta\partial\beta'} = -\frac{1}{\sigma^2} X'X$$

$$\frac{\partial^2 l(\beta, \sigma^2)}{\partial^2(\sigma^2)^2} = \frac{n}{2\sigma^4} - \frac{1}{\sigma^6}(y - X\beta)'(y - X\beta)$$

$$\frac{\partial^2 l(\beta, \sigma^2)}{\partial\beta\partial\sigma^2} = -\frac{1}{\sigma^4} X'(y - X\beta).$$

Sabendo que a matriz de informação de Fisher é calculada usando os parâmetros verdadeiros, isto é, a expressão $y - X\beta = u$ será usada para simplificar o valor esperado das derivadas acima. Então, calculando o valor esperado das expressões acima condicional ao valor de X (a matriz de dados) chegamos a

$$E\left[\frac{\partial^2 l(\beta, \sigma^2)}{\partial\beta\partial\beta'}/X\right] = -\frac{1}{\sigma^2} X'X$$

$$E\left[\frac{\partial^2 l(\beta, \sigma^2)}{\partial^2(\sigma^2)^2}/X\right] = \frac{n}{2\sigma^4} - \frac{1}{\sigma^6} E[u'u/X] = \frac{n}{2\sigma^4} - \frac{n\sigma^2}{\sigma^6} = -\frac{n}{2\sigma^4}$$

$$E\left[\frac{\partial^2 l(\beta, \sigma^2)}{\partial\beta\partial\sigma^2}/X\right] = -\frac{1}{\sigma^4} X'E[u/X] = 0.$$

Finalmente, utilizamos os valores negativos dos valores esperados dessas derivadas para encontrar a matriz de Fisher

$$I(\beta, \sigma^2/X) = \begin{bmatrix} \frac{1}{\sigma^2}X'X & 0 \\ 0 & \frac{n}{2\sigma^4} \end{bmatrix} \tag{8.42}$$

e a sua inversa dada por

$$I^{-1}(\beta, \sigma^2/X) = \begin{bmatrix} \sigma^2 \frac{1}{X'X} & 0 \\ 0 & \frac{2\sigma^4}{n} \end{bmatrix}. \tag{8.43}$$

Sabemos da Seção 6.4.2 que a inversa da matriz de Fisher é igual a matriz de variância-covariância dos parâmetros estimados $\hat{\beta}$ e $\hat{\sigma}^2$. Entretanto, uma vez que não sabemos os parâmetros reais, para calcular a matriz de Fisher (matriz de variância-covariância), conforme Nota 5.1, utilizaremos as estimativas $\hat{\beta}$ e $\hat{\sigma}^2$.

Na Seção 6.4.2, também vimos que no contexto de estimação via máxima verossimilhança, podemos testar uma série de hipóteses usando uma função $h(\theta) = 0$. Aqui, estamos interessados num caso particular da função dada por $h(\theta) = [R_{l(r)\times K},\ 0_{l(r)\times 1}]\theta - r$, onde $\theta = [\beta'\ \sigma^2]'$, que serve para testar a hipótese nula $H_0 : R\beta = r$ como visto no Teorema 8.2. Note que em geral não estamos interessados em testar hipóteses a respeito de σ^2, por isso não fazemos restrições em relação a esse parâmetro. Como discutimos na Seção 6.4.2, no contexto de estimação utilizando o método de máxima verossimilhança, podemos testar hipóteses usando 3 testes diferentes: teste de Wald, teste de razão de máxima verossimilhança e teste dos multiplicadores de Lagrange. Além disso, vimos também que esses testes têm distribuição qui-quadrada com $l(r)$ graus de liberdade. Nessa seção basicamente o que fazemos é derivar os testes estudados na Seção 6.4.2, para o caso de testes de hipóteses em modelos de regressão linear.

Para proceder com o **teste de Wald** precisamos calcular os valores das derivadas parciais de h em relação a β e a σ^2

$$\frac{\partial h(\theta)}{\partial \theta} = \begin{pmatrix} R_{l(r)\times K} & | & 0_{l(r)\times 1} \end{pmatrix}. \tag{8.44}$$

Utilizando a Eq. (6.15), chegamos a estatística de Wald para o caso do modelo de regressão linear e h restrita ao caso de testes de hipóteses lineares dada por

$$\begin{aligned} W &= n(R\hat{\beta} - r)' \left[\begin{pmatrix} R & 0 \end{pmatrix} n \begin{bmatrix} \hat{\sigma}^2 \frac{1}{X'X} & 0 \\ 0 & \frac{2\hat{\sigma}^4}{n} \end{bmatrix} \begin{pmatrix} R' \\ 0 \end{pmatrix} \right]^{-1} (R\hat{\beta} - r) && (8.45) \\ &= n(R\hat{\beta} - r)'[Rn\hat{\sigma}^2(X'X)^{-1}R']^{-1}(R\hat{\beta} - r) && (8.46) \\ &= \frac{n(R\hat{\beta} - r)'[R(X'X)^{-1}R']^{-1}(R\hat{\beta} - r)}{SQR(\hat{\beta})}. && (8.47) \end{aligned}$$

Usando o fato apresentado no Exemplo 8.15 que $SQR(\tilde{\beta}) - SQR(\hat{\beta}) = (R\hat{\beta} - r)'[R(X'X)^{-1}R']^{-1}(R\hat{\beta} - r)$, onde $\tilde{\beta}$ que corresponde ao valor estimado para β sujeito à restrição $h(\beta) = 0$, chega-se a

$$W = n\frac{SQR(\tilde{\beta}) - SQR(\hat{\beta})}{SQR(\hat{\beta})}, \tag{8.48}$$

que é a estatística de Wald para o caso do modelo de regressão linear com teste de hipótese restrito ao caso linear.

Prosseguindo com a derivação dos testes de hipóteses no contexto de estimação via máxima verossimilhança como na Seção 6.4.2, agora vamos derivar a estatística conhecida como **razão de máxima verossimilhança**. Usando a Eq. (8.40) e calculando a diferença entre $\log L(\hat{\beta})$ e $\log L(\tilde{\beta})$ chegamos a

$$\begin{aligned} \log L(\hat{\beta}) - \log L(\tilde{\beta}) &= -\frac{n}{2}\log\left(\frac{1}{n}SQR(\hat{\beta})\right) + \frac{n}{2}\log\left(\frac{1}{n}SQR(\tilde{\beta})\right) && (8.49) \\ &= \frac{n}{2}\log\left(\frac{SQR(\tilde{\beta})}{SQR(\hat{\beta})}\right). && (8.50) \end{aligned}$$

Usando a definição do LRT dada pela Eq. (6.17) chegamos a

$$LRT = 2[\log L(\hat{\beta}) - \log L(\tilde{\beta})] = n\log\left(\frac{SQR(\tilde{\beta})}{SQR(\hat{\beta})}\right). \tag{8.51}$$

Finalmente, agora vamos derivar a última estatística dessa lista conhecida como **teste dos multiplicadores de Lagrange** discutida na Seção 6.4.2.

Usando a Eq. (8.38) e o fato que

$$\tilde{\sigma}^2 = \frac{1}{n}SQR(\tilde{\beta}),$$

chega-se

$$\frac{\partial \log L(\tilde{\theta})}{\partial \theta} = \frac{n}{SQR(\tilde{\beta})}\begin{bmatrix} X'(y - X\tilde{\beta}) \\ 0 \end{bmatrix}.$$

Portanto, usando a definição do teste de multiplicadores de Lagrange dado pela Eq. (6.19), chegamos ao teste de multiplicadores de Lagrange para o caso do modelo de regressão linear com testes de hipóteses lineares

$$\begin{aligned} LM &= \frac{1}{n}\left(\frac{n}{SQR(\tilde{\beta})}\begin{bmatrix} (y-X\tilde{\beta})'X & 0 \end{bmatrix} n \begin{bmatrix} \tilde{\sigma}^2\frac{1}{X'X} & 0 \\ 0 & \frac{2\tilde{\sigma}^4}{n} \end{bmatrix} \frac{n}{SQR(\tilde{\beta})}\begin{bmatrix} X'(y-X\tilde{\beta}) \\ 0 \end{bmatrix}\right) \\ &= \frac{n}{SQR(\tilde{\beta})}\left(\left[(y-X\tilde{\beta})'X\right](X'X)^{-1}\left[X'(y-X\tilde{\beta})\right]\right) \\ &= \frac{n}{SQR(\tilde{\beta})}\left((y-X\tilde{\beta})'P(y-X\tilde{\beta})\right), \end{aligned} \tag{8.52}$$

onde $P = X(X'X)^{-1}X'$ é conhecida como **matriz de projeção** (vide Exercício 8.6).

Usando o fato que[12] $(y-X\tilde{\beta})'P(y-X\tilde{\beta}) = SQR(\tilde{\beta}) - SQR(\hat{\beta})$, chega-se a

$$LM = n\frac{SQR(\tilde{\beta}) - SQR(\hat{\beta})}{SQR(\tilde{\beta})}. \tag{8.53}$$

8.3 Análise da qualidade do modelo de regressão linear estimado

Nesta seção apresentamos algumas linhas gerais de como avaliar a qualidade do modelo estimado. A primeira questão é testar se o modelo satisfaz as hipóteses propostas. A segunda é como lidar com previsão fora da amostra. A terceira é como avaliar a qualidade do modelo de regressão linear. Essa seção segue em parte o espírito de Harrell (2001).

8.3.1 As hipóteses do modelo são válidas?

Na Seção 8.2.1 fizemos várias hipóteses para o bom funcionamento do modelo de regressão linear. Como podemos verificar se essas hipóteses realmente são verdadeiras?

O primeiro método que podemos usar para avaliar a qualidade do modelo estimado é investigar graficamente os resíduos do modelo de regressão linear. Por exemplo, podemos plotar graficamente $\hat{u}$ ou $\hat{u}^2$ *versus* cada um dos regressores e também versus $\hat{y}$. Nesses gráficos devemos verificar se a variância do resíduo cresce ou decresce de acordo com algumas dessas variáveis ou se existe correlação entre os resíduos e essas variáveis. Se os dados usados forem dados de serem temporais, é válido também plotar $\hat{u}_i$ versus $\hat{u}_{i-1}$ e verificar graficamente se existe correlação entre os resíduos. Se algum desses problemas for detectado, isso

[12] Esse valor pode ser encontrado rearrumando os resultados do Exemplo 8.15.

significa que a formulação do modelo precisa ser revista ou talvez métodos de estimação robustos, tais como mínimos quadrados ponderados, devam ser considerados. Adicionalmente, podemos verificar se a hipótese de linearidade é válida. Finalmente, se a hipótese de normalidade for considerada como na Seção 8.2.1, então devemos plotar também o o gráfico QQ-plot, apresentado na Seção 7.2, para verificar a validade dessa hipótese. Existem duas vantagens nessa metodologia. A primeira é a simplicidade. A segunda é que ela pode dar dicas de onde se deve corrigir o modelo. Os Exemplos 8.16 a 8.21 ilustram essa metodologia.

Exemplo 8.16 (Continuação do Exemplo 8.8 – Modelo corretamente especificado) Usando o mesmo gerador de dados do Exemplo 8.8, geramos $n = 50$ amostras e estimamos exatamente o modelo utilizado para gerar os dados utilizando o método dos mínimos quadrados. Então na Figura 8.3 plotamos os gráficos de $\hat{u} \times x_1$, $\hat{u} \times x_2$, $\hat{u} \times y$ e finalmente o gráfico QQ-plot de $\hat{u}$ *versus* a distribuição normal. Como vemos nessa figura, as hipóteses parecem ser válidas e o modelo corretamente especificado.

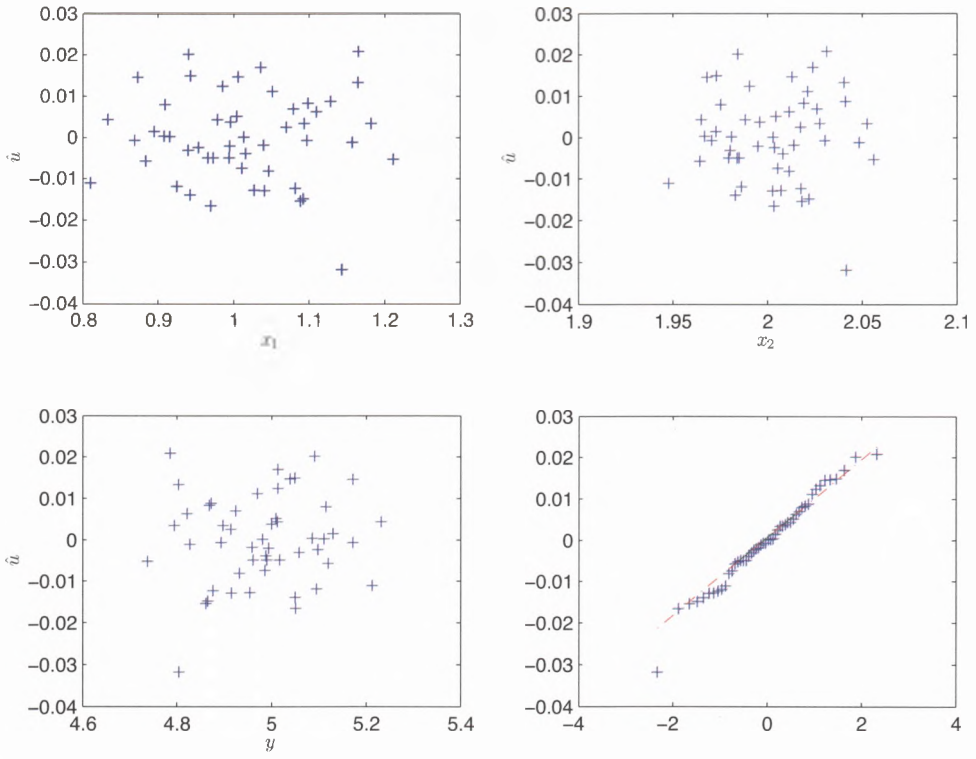


Figura 8.3: Modelo corretamente especificado.

Exemplo 8.17 (Resíduo com heterocedasticidade – variância do ruído crescendo linearmente com um regressor) Nesse exemplo, variamos um pouco o gerador de dados utilizado no Exemplo 8.8. Utilizamos o mesmo modelo linear

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + u_i$$

para $i = 1, \cdots, n$, $\beta_0 = 1$, $\beta_1 = -2$ e $\beta_2 = 3$. Mas geramos os dados do modelo usando $x_1 \sim \text{Normal}[\mu_{x_1}, \sigma^2_{x_1}]$, onde $\mu_{x_1} = 1$ e $\sigma^2_{x_1} = 0.1$, $x_2 \sim \text{Normal}[\mu_{x_2}, \sigma^2_{x_2}]$, onde $\mu_{x_2} = 2$ e $\sigma^2_{x_2} = 0.001$, $u \sim \text{Normal}[\mu, x_1\sigma^2]$, onde $\mu = 0$ e $\sigma^2 = 0.0001$ e $n = 50$.

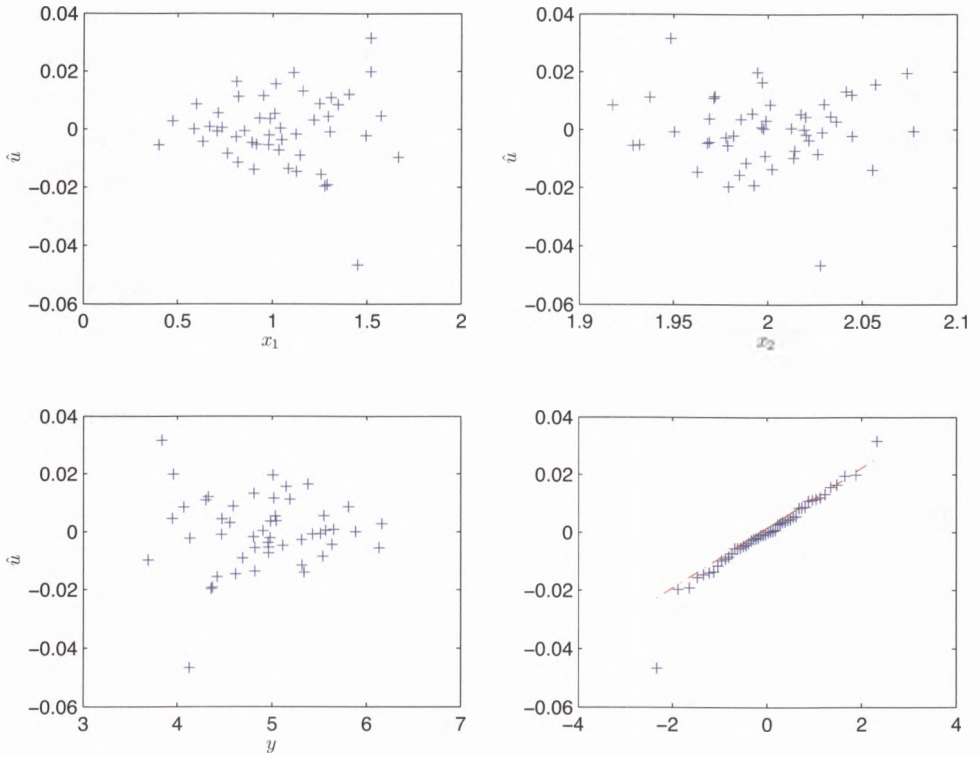


Figura 8.4: Resíduo com heterocedasticidade com a variância crescendo linearmente com o regressor x_1.

Estimamos o modelo acima e na Figura 8.4 plotamos os gráficos de $\hat{u} \times x_1$, $\hat{u} \times x_2$, $\hat{u} \times y$ e, finalmente, o gráfico QQ-plot de $\hat{u}$ *versus* a distribuição normal. Como podemos ver nesse modelo apresentado na Figura 8.4, a variância de $\hat{u}$ cresce com x_1. Além disso, como o coeficiente β_1 é negativo, vemos também que a variância do resíduo decresce em relação a y.

Exemplo 8.18 (Resíduo com heterocedasticidade – variância do ruído crescendo quadraticamente com um regressor) Nesse exemplo, variamos um pouco o gerador de dados utilizado no Exemplo 8.8. Utilizamos o mesmo modelo linear

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + u_i$$

para $i = 1, \cdots, n$, $\beta_0 = 1$, $\beta_1 = -2$ e $\beta_2 = 3$. Mas geramos os dados do modelo usando $x_1 \sim \text{Normal}[\mu_{x_1}, \sigma^2_{x_1}]$, onde $\mu_{x_1} = 1$ e $\sigma^2_{x_1} = 0.1$, $x_2 \sim \text{Normal}[\mu_{x_2}, \sigma^2_{x_2}]$, onde $\mu_{x_2} = 2$ e $\sigma^2_{x_2} = 0.001$, $u \sim \text{Normal}[\mu, x_1^2\sigma^2]$, onde $\mu = 0$ e $\sigma^2 = 0.0001$ e $n = 50$.

Estimamos o modelo acima e na Figura 8.5 plotamos os gráficos de $\hat{u} \times x_1$, $\hat{u} \times x_2$, $\hat{u} \times y$ e finalmente o gráfico QQ-plot de $\hat{u}$ versus a distribuição normal. Como podemos ver nesse modelo apresentado na Figura 8.5, os resultados do Exemplo 8.17 são amplificados. Como no Exemplo 8.17, a variância de $\hat{u}$ cresce com x_1. Além disso, como o coeficiente β_1 é negativo, vemos também que a variância do resíduo decresce em relação a y. Nota-se também nessa figura que o efeito da heterocedasticidade é tão forte que no gráfico

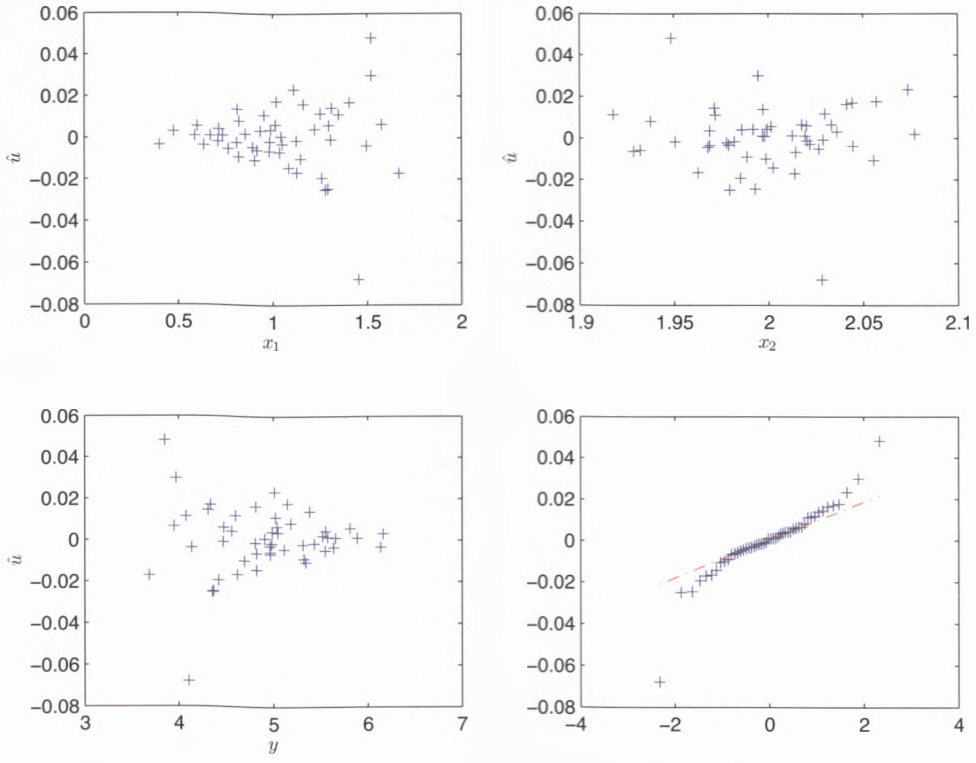


Figura 8.5: Resíduo com heterocedasticidade com a variância crescendo quadraticamente com o regressor x_1.

QQ-plot começa a haver um afastamento entre a distribuição normal e a distribuição empírica de $\hat{u}$ nos extremos.

Exemplo 8.19 (Continuação do Exemplo 8.8 – Omissão de variáveis) Usando o mesmo gerador de dados do Exemplo 8.8, geramos $n = 50$ amostras. A diferença aqui é que estimamos o modelo

$$y_i = \beta_0 + \beta_1 x_{i1} + u_i.$$

Então na Figura 8.3 plotamos os gráficos de $\hat{u} \times x_1$, $\hat{u} \times x_2$, $\hat{u} \times y$ e, finalmente, o gráfico QQ-plot de $\hat{u}$ *versus* a distribuição normal. É notável como o regressor x_2 agora aparece no erro.

Exemplo 8.20 (Erro de especificação funcional) Nesse exemplo, variamos um pouco o gerador de dados utilizado no Exemplo 8.8. Utilizamos o modelo quadrático

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}^2 + u_i$$

para $i = 1, \cdots, n$, $\beta_0 = 1$, $\beta_1 = -2$ e $\beta_2 = 3$ usando $x_1 \sim \text{Normal}[\mu_{x_1}, \sigma_{x_1}^2]$, onde $\mu_{x_1} = 1$ e $\sigma_{x_1}^2 = 0.1$, $x_2 \sim \text{Normal}[\mu_{x_2}, \sigma_{x_2}^2]$, onde $\mu_{x_2} = 2$ e $\sigma_{x_2}^2 = 0.01$, $u \sim \text{Normal}[\mu, \sigma^2]$, onde $\mu = 0$ e $\sigma^2 = 0.0001$ e $n = 50$.

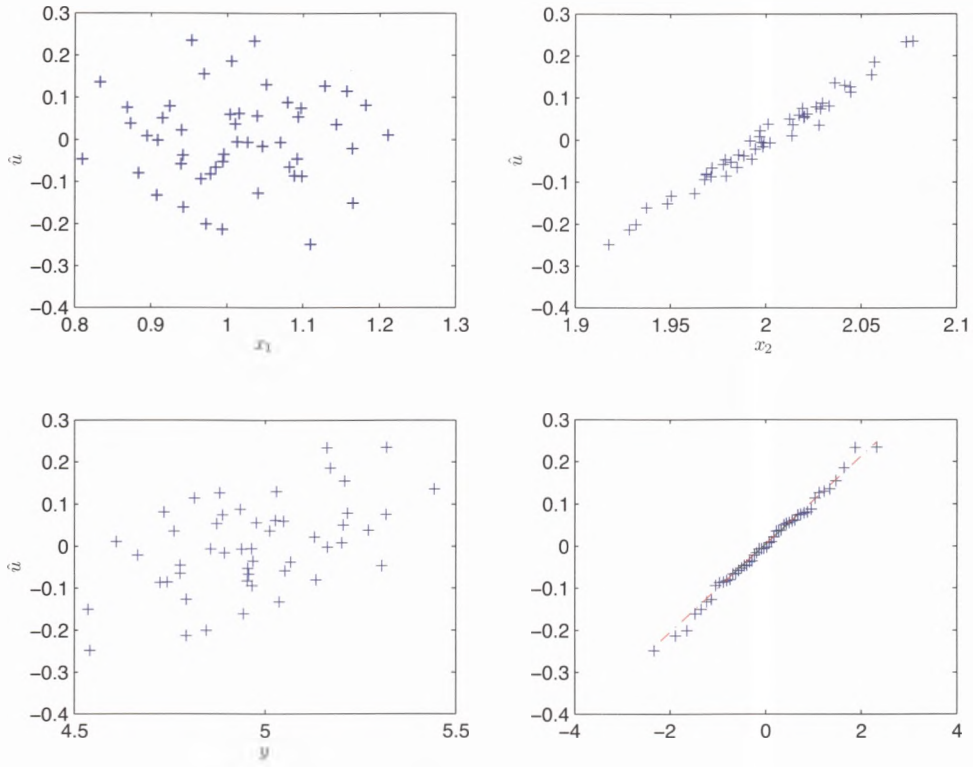


Figura 8.6: Omissão do regressor x_2.

Entretanto, cometemos um erro de especificação estimando o modelo linear

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + u_i.$$

Na Figura 8.7 plotamos os gráficos de $\hat{u} \times x_1$, $\hat{u} \times x_2$, $\hat{u} \times y$ e, finalmente, o gráfico QQ-plot de $\hat{u}$ *versus* a distribuição normal. Como podemos ver nessa figura aparece explicitamente uma relação quadrática entre o termo de erro $\hat{u}$ e o regressor x_2. Esse efeito aparece também no gráfico $\hat{u}$ versus y e no gráfico QQ-plot.

Exemplo 8.21 (Distribuição não normal) Nesse exemplo, variamos um pouco o gerador de dados utilizado no exemplo 8.8 modificando apenas a distribuição do termo de resíduo u com distribuição t-Student com um grau de liberdade e $n = 50$.

Podemos notar três fenômenos interessantes na Figura 8.8. O primeiro refere-se aos gráficos $\hat{u} \times x_1$ e $\hat{u} \times x_2$ onde existe maior concentração de pontos nos extremos visto que a distribuição t-Student tem causas mais pesadas que a distribuição normal. O segundo é uma relação crescente entre $\hat{u}$ e y. Isso ocorre pois uma vez que x_1 e x_2 têm distribuição normal e, portanto, ficam mais concentrados em torno da média, o valor de y será maior quando u for maior e será menor quando u for menor. E o terceiro refere-se ao gráfico QQ-plot da distribuição empírica de $\hat{u}$ versus a distribuição normal onde podemos notar um afastamento entre as duas distribuições principalmente nos extremos.

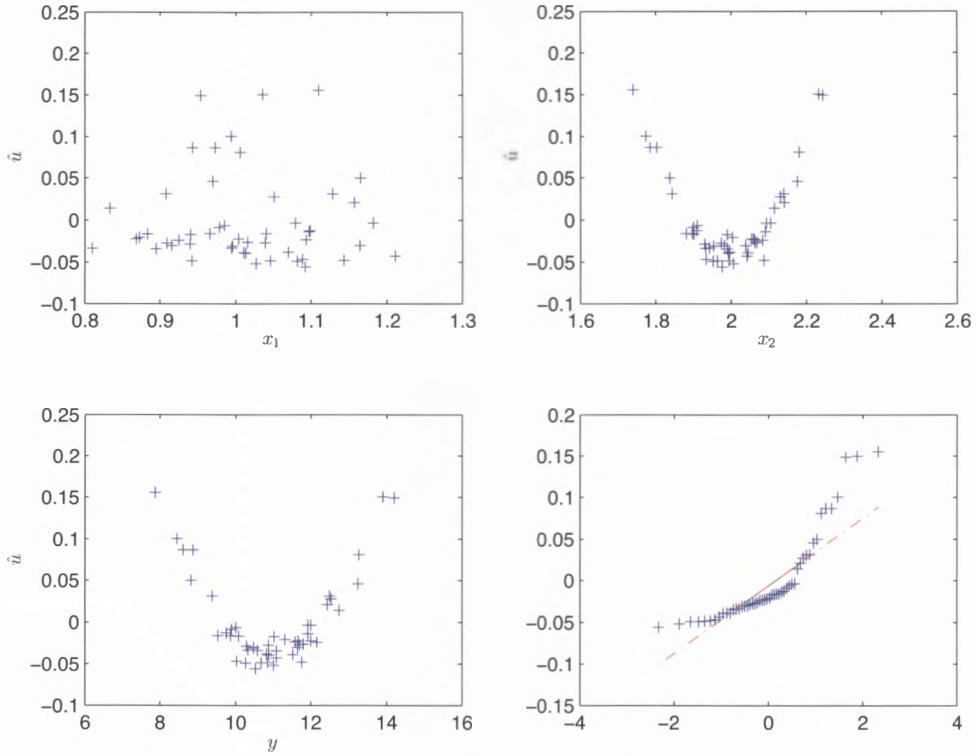


Figura 8.7: Erro de especificação.

Em caso de encontrar algum ou alguns dos fenômenos indesejados acima, sugere-se a implementação de testes estatísticos mais precisos. Por exemplo, podemos fazer o teste de correlação de Spearman estudado no Capítulo 2 para verificar a existência de heterocedasticidade. Para testar a presença de autocorrelação (no caso de dados de séries temporais), podemos rodar uma regressão de $\hat{u}_i$ contra seus valores defasados e testar a hipótese de que todos os coeficientes da regressão dos termos defasado são nulos. Finalmente, para testar se os resíduos possuem distribuição normal, podemos fazer o teste de Jarque-Bera (válido apenas para grandes amostras). Detalhes nesses procedimentos podem ser encontrados por exemplo em Ruud (2000), Gujarati (2000), Wooldridge (2001) and Wooldridge (2003).

8.3.2 O modelo é capaz de fazer previsões fora da amostra usada para a estimação?

Uma aplicação muito útil de modelos econométricos é fazer previsão de uma determinada variável dependente y usando dados de variáveis dependentes que não foram usados para construir a amostra. Nesse contexto, uma pergunta importante é se a previsão feita, utilizando dados não pertencentes a amostra usada para estimação, é confiável.

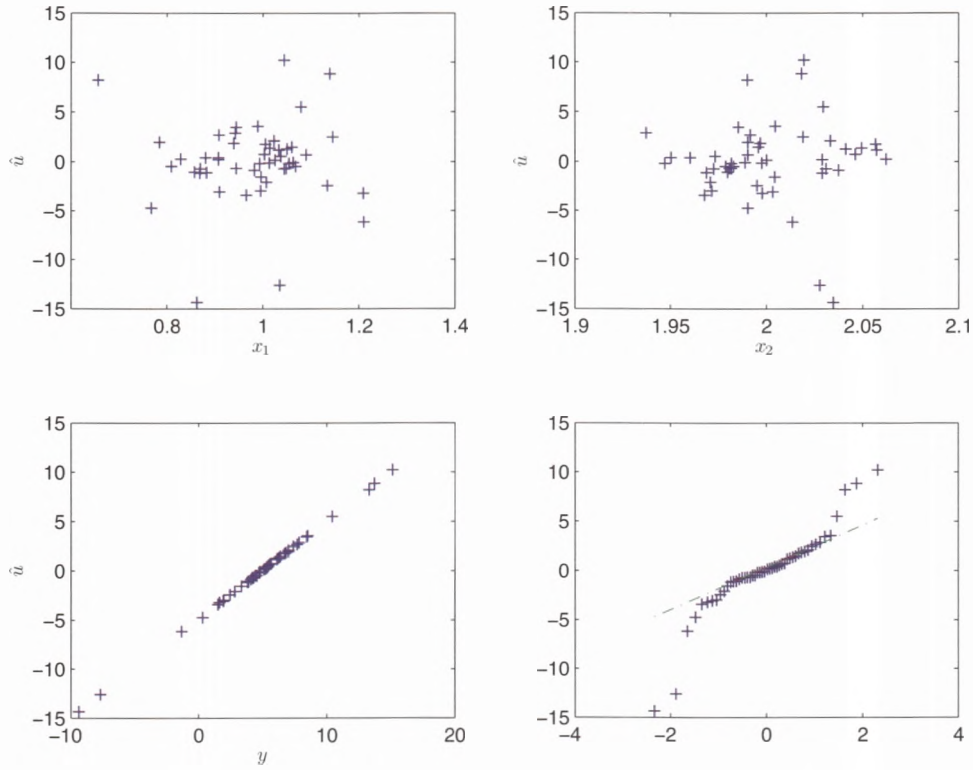


Figura 8.8: Distribuição não normal.

Uma técnica usada para testar habilidade de previsão fora da amostra é chamada de validação cruzada, e baseia-se em dividir a amostra original (aleatoriamente) em duas partes. Enquanto a primeira parte da amostra é usada para estimar o modelo, a segunda parte da amostra é usada para testar o modelo estimado.

Usando esse procedimento, pode-se responder a pergunta colocada acima com segurança. Em geral, diz que um modelo que não responde bem a previsão fora da amostra "decorou"[13] os dados da amostra usada para estimação. Em modelos lineares a causa mais comum para esse fenômeno é um número exagerado de regressores. Obviamente uma solução desse problema é eliminar alguns dos regressores. A dificuldade é como escolher os regressores a serem eliminados. Entretanto, o bom senso pode ajudar. A primeira solução é buscar suporte na literatura econômica na escolha dos regressores a serem eliminados. Uma outra solução é eliminar variáveis cujas distribuições sejam muito estreitas. Finalmente, uma solução estatística para o problema é a técnica da análise multivariada, conhecida como análise dos componentes principais, que transforma um conjunto de variáveis correlacionadas em um conjunto de variáveis não correlacionadas (ANDERSON, 2003).

[13]Em inglês, esse fenômeno é chamado de *overfitting*.

8.3.3 O Modelo responde de forma desejada ao esperado pela teoria?

Uma outra questão prática relevante na estimação de modelos lineares é se o modelo estimado responde corretamente à teoria. Por exemplo, se no caso do CAPM encontrarmos um β negativo, isso significa que algo não está funcionando bem no modelo. Então, ou o modelo não é um bom modelo, ou a amostra não é uma boa amostra, ou existe um outro problema na especificação do modelo que está causando esse problema. Por exemplo, a omissão de variáveis relevantes no modelo pode causar esse problema. Então avaliar se o modelo responde bem à teoria, e se não responde, avaliar as causas é fundamental.

8.3.4 Qualidade do ajuste e os coeficientes de determinação

Uma questão que em geral economistas estão interessados é saber qual a habilidade que o conjunto de variáveis independentes tem para explicar a variável dependente. Considerando que se um dos regressores é constante e, portanto (vide Exercício 8.19),

$$SQT = SQE + SQR, \tag{8.54}$$

esse papel pode ser desempenhado pelo chamado coeficiente de determinação R^2 que é definido por

$$R^2 = \frac{SQE}{SQT} = 1 - \frac{SQR}{SQT}, \tag{8.55}$$

onde $SQE = \sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2$, $SQT = \sum_{i=1}^{n}(y_i - \overline{y})^2$ e $\overline{y} = \sum_{i=1}^{n} \frac{y_i}{n}$. Note que o R^2 nada mais é do que a razão da variação da variável independente, que é explicada pelo conjunto de variáveis independentes em relação a variação total (explicada e não explicada).

Nota 8.5 Uma propriedade importante do R^2 trivialmente verificada é que quando um dos regressores do modelo de regressão linear é uma constante então $0 \leq R^2 \leq 1$.[14] Logo, o R^2 igual a 1 significa perfeito ajuste.

Uma restrição à habilidade do R^2 na verificação do ajuste de um modelo de regressão linear é que quando uma nova variável independente é adicionada a regressão, o R^2 nunca decresce e usualmente cresce. Isso ocorre pois o SQR nunca aumenta quando um novo regressor é adicionado ao modelo. Uma extensão do R^2 conhecida como R^2_{ajustado} é uma forma de tentar lidar com essa restrição. Com o objetivo de entender o R^2_{ajustado}, reescreva o R^2 como

[14] Se o modelo de regressão não incluir uma constante como regressor, então o R^2 pode ser negativo.

$$R^2 = 1 - \frac{SQR/n}{SQT/n}.$$

Note que SQR/n e SQT/n são respectivamente estimativas viesadas da variância amostral de $\hat{u}$ e y. O R^2_{ajustado} é então definido a partir da mesma ideia usada para definir o R^2, substituindo as variâncias de por seus valores não viesados como em

$$R^2_{\text{ajustado}} = 1 - \frac{SQR/(n-K)}{SQT/(n-1)}.$$

Um procedimento usualmente comum é usar o R^2 ou R^2_{ajustado} para a decidir se um novo regressor deve ser incluído ou não no modelo. Esse procedimento não parece ser o mais interessante, visto que a decisão de um novo regressor entrar ou não no modelo deve ser feita usando teoria econômica. De fato, de acordo com Exercício 8.20, o R^2 pode ser visto como um passo intermediário para o cálculo da estatística F. Finalmente, o R^2 mede apenas um aspecto da habilidade preditiva de um modelo que é o ajuste. Outros aspectos não considerados são habilidade de previsão fora da amostra onde o modelo é aplicado à uma amostra onde não foi estimado como discutido na Seção 8.3.2 e que o modelo responda de acordo com o esperado pela teoria como discutido na Seção 8.3.3.

Exemplo 8.22 (Continuação do Exemplo 8.8 - R^2 e R^2_{ajustado}) Para a regressão linear discutida no Exemplo 8.8, encontramos $R^2 = 0.9987$ e $R^2_{\text{ajustado}} = 0.9983$.

8.4 Exercícios

Nota 8.6 Devido a relação íntima que existe entre modelos de regressão linear e álgebra linear e matricial, vários dos exercícios desse capítulo exigem o conhecimento desses tópicos. Então sugere-se ao leitor desavisado a consulta de referências tais como Franklin (1968), Eves (2008), Lima (1995) e Abadir e Magnus (2005).

Exercício 8.1 Calcule a derivada de $SQR(\beta) = y'y - 2y'X\beta + \beta'X'X\beta$ em relação a β.

Dica: Note que $SQR(\beta) = y'y - 2y'X\beta + \beta'X'X\beta$ é uma função $SQR : \Re^K \mapsto \Re$. Então a forma mais simples de fazer essa conta é explicitar os vetores e matrizes da equação da SQR em função dos elementos, depois derivar a função real SQR em relação a cada β_k, $k = 1, \cdots, K$, e finalmente compactar os resultados no formato matricial para encontrar $\frac{\partial SQR(\beta)}{\partial \beta} = -2X'y + 2X'X\beta$.

Exercício 8.2 Considere as duas regressões lineares, uma restrita dada por $y = X\beta + u$ e outra irrestrita $y = X\beta + Z\delta + u$. Mostre que se os regressores representados por X e Z são independentes dois a dois, satisfazendo $X'Z = 0$, então as estimativas de β nas duas regressões são as mesmas.

Dica: Use a definição de $\hat{\beta}$ apresentado na Eq. (8.16).

Exercício 8.3 Mostre que os produtos matriciais abaixo satisfazem as igualdades:

1)

$$\begin{bmatrix} 1 & 1 & \cdots & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} = n$$

2)

$$\begin{bmatrix} 1 & 1 & \cdots & 1 \end{bmatrix} \begin{bmatrix} x_{11} \\ x_{21} \\ \vdots \\ x_{n1} \end{bmatrix} = \sum_{i=1}^{n} x_{i1}$$

3)

$$\begin{bmatrix} x_{11} & x_{21} & \cdots & x_{n1} \end{bmatrix} \begin{bmatrix} x_{11} \\ x_{21} \\ \vdots \\ x_{n1} \end{bmatrix} = \sum_{i=1}^{n} x_{i1}^2.$$

Exercício 8.4 Calcule a inversa de

$$X'X = \begin{bmatrix} n & \sum_{i=1}^{n} x_{i1} \\ \sum_{i=1}^{n} x_{i1} & \sum_{i=1}^{n} x_{i1}^2 \end{bmatrix}.$$

Exercício 8.5 Mostre as seguintes propriedades de matrizes idempotentes:

1) Uma matriz idempotente e diagonal têm todos os elementos iguais a 0 ou iguais a 1.

2) Todos autovalores de uma matriz idempotente são iguais a 0 ou iguais a 1.

3) Uma matriz simétrica com autovalores iguais a 0 ou iguais a 1 é idempotente.

Dica: Uma matriz A é idempotente se satisfaz $A^2 = A$. Uma matriz A é simétrica se satisfaz $A' = A$.

Exercício 8.6 Podemos escrever $y = X\hat{\beta} + \hat{u} = Py + My$, onde as matrizes $P = X(X'X)^{-1}X'$ e $M = I_n - P$ são chamadas respectivamente de **Matriz de Projeção** e **Matriz Aniquiladora**. De fato, uma vez que $\hat{u}$ é ortogonal a X (vide Eq. (8.23)), y pode ser decomposto em duas componentes: uma que é combinação linear das colunas de X e outra que é ortogonal a X.

Mostre que essas matrizes satisfazem as seguintes propriedades:

1) $PX = X$

2) $MX = 0$ (que justifica o termo Matriz Aniquiladora)

3) As matrizes P e M são idempotentes e simétricas.

Dica: O traço de uma matriz quadrada é a soma dos elementos da diagonal principal.

Exercício 8.7 Mostre as seguintes propriedades de matrizes positivas definidas:

1) Toda matriz positiva definida tem autovalores positivos.

2) Toda matriz positiva definida tem determinante positivo.

3) Toda matriz positiva definida tem inversa.

Dica: Diz-se que A é uma matriz positiva definida se para todo vetor $v \neq 0$, $v'Av > 0$. (1) Use a definição de autovalor e a definição de matriz positiva definida. (2) Use o fato que o determinante de uma matriz é o produto dos autovalores. (3) Use o fato que o determinante é positivo.

Exercício 8.8 Mostre que uma matriz simétrica é positiva definida se e somente se todos os autovalores são positivos.

Dica: Para provar que uma matriz simétrica com autovalores positivos é positiva definida, escreva a diagonalização canônica, lembrando que uma matriz simétrica pode ser diagonalizável por meio de uma matriz ortogonal, onde uma matriz A é ortogonal quando $A'A = I$. Para provar o converso, use as definições de autovalor, autovetor e matriz positiva definida.

Exercício 8.9 Mostre que $X'X$, onde X é uma matriz de ordem $n \times K$, tem posto K.

Dica: Mostre que essa matriz é positiva definida.

Exercício 8.10 Mostre que se uma matriz X de ordem $n \times K$ tem posto K então $X'X$ possui inversa.

Dica: Trivial dos Exercícios 8.7 e 8.9.

Exercício 8.11 Mostre que o erro de estimação $\hat{\beta} - \beta = (X'X)^{-1}X'u$.

Exercício 8.12 Mostre que para duas matrizes quadradas A e B:

1) traço$(A + B)$=traço(A)+traço(B).

2) traço(cA)=c traço(A), onde c é um escalar.

3) traço(AA')=traço($A'A$)

Exercício 8.13 Prove as Proposições 8.1, 8.2, 8.3 e 8.4.

Dicas: (1) Proposição 8.1: Use Eq. (8.20) e a Hipótese 8.2. (2) Proposição 8.2: Use Eq. (8.20). (3) Proposição 8.3: Construa um estimador genérico linear em y , use a Proposição 8.1 e calcule a variância desse estimador mostrando que ela é igual a variância de $\hat{\beta}$ na Proposição 8.2 adicionada de um termo positivo. (4) Proposição 8.4 Mostre que $\hat{u}'\hat{u} = u'Mu$, onde M é apresentada no Exercício 8.6. Mostre também que $E[u'Mu/X] = \sigma^2$traço(M) e calcule o valor desse traço usando Exercício 8.12.

Exercício 8.14 Mostre que $\hat{\sigma}^2 = \frac{SQR}{n-K}$ tem distribuição qui-quadrada com $n - K$ graus de liberdade.

Dica: Reescreva SQR usando a matriz aniquiladora.

Exercício 8.15 Mostrar que $R(X'X)^{-1}R'$ possui inversa.

Dica: (1) Note que $(X'X)^{-1}$ é simétrica. (2) Lembre que os autovalores da matriz inversa, são os inversos dos autovalores da matriz original. (3) Logo, usando o Exercício 8.8, $(X'X)^{-1}$ é também positiva definida. (4) Uma matriz positiva definida A pode ser fatorada como $A = BB'$.

Exercício 8.16 Mostre que a variável aleatória $(R\hat{\beta} - r)'[R(X'X)^{-1}R']^{-1}(R\hat{\beta} - r)$ tem distribuição qui-quadrada com $l(r)$ graus de liberdade.

Exercício 8.17 Considere o modelo de regressão linear dado por $y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$. Mostre como você faria para testar a hipótese $\beta_1 + \beta_2 = c$ nas duas formas do teste F – uma apresentada pelo Teorema 8.2 e a outra apresentada no Exemplo 8.15, onde c é uma constante.

Exercício 8.18 Considere o modelo de regressão linear dado por $y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$. Transforme essa regressão de forma que a hipótese $\beta_1 + \beta_2 = c$, onde c é uma constante, possa ser escrita como uma restrição de igualdade envolvendo apenas um parâmetro da regressão. Estenda essa ideia para o caso onde temos restrições do tipo $R\beta = r$.

Exercício 8.19 Mostrar que $SQT = SQE + SQR$.

Dica: Use a definição de $\hat{u}_i$ e chegue a $\sum_{i=1}^{n}(y_i - \overline{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2 + \sum_{i=1}^{n}\hat{u}_i^2 + 2\sum_{i=1}^{n}(\hat{y}_i - \overline{y})\hat{u}_i$, onde o último termo é nulo devido a Eq. (8.23).

Exercício 8.20 Mostre que para restrições de exclusão o teste F pode ser escrito como

$$F = \frac{(R_\beta^2 - R_{\tilde{\beta}}^2)/l(r)}{(1 - R_\beta^2)/(n - K)}.$$

9. Regressão com resposta binária e modelos de classificação

"Still, I have no love for the cloth.
Just as cotton,
which is in itself the most harmless substance in the world,
becomes dangerous on being dipped into nitric acid,
so the mildest of mortals is to be feared if he is once soaked in sectarian religion."
Arthur Conan Doyle

9.1 Introdução

No Capítulo 8 introduzimos o modelo de regressão linear, que é o modelo econométrico mais popular para relacionar uma variável dependente com um conjunto de variáveis independentes. Embora esse modelo seja útil para modelar vários tipos de situações, ele, por exemplo, não é adequado para todas as situações práticas, como por exemplo aquelas que trataremos nesse capítulo. Considere que desejamos explicar um conjunto de dados onde a variável dependente assume apenas valores 0 ou 1. Por exemplo, admita que temos uma amostra de dados sobre firmas ($i = 1 \cdots n$) onde uma parte dessas firmas quebraram ($y_i = 1$) e uma outra parte continua normalmente em operação ($y_i = 0$) e queremos utilizar variáveis específicas das firmas relacionadas com o setor de operação da firma (construção civil, indústria etc), estrutura de capital (relação entre passivos e patrimônio líquido), estrutura de ativos (relação entre capital de giro e ativo total, relação entre ativos geradores de renda e não geradores de renda etc.) e habilidade para a geração de caixa na firma (relação entre lucro líquido e patrimônio líquido, relação entre lucro operacional e receita líquida etc.), para explicar a quebra de firmas (DUFFIE; SINGLETON, 2003; BARTH, 2004). Uma abordagem para esse problema pode ser feita tentando explicar a probabilidade de uma firma quebrar usando covariáveis de firmas como

$$p(x_i) = P(y_i = 1/x_i). \tag{9.1}$$

Implicitamente, estamos supondo que a variável aleatória y_i, para a firma i, tem distribuição de Bernoulli, com probabilidade de sucesso $P(y_i = 1/x_i)$, que depende das características da firma i, características essas que estão representadas numericamente nos componentes do vetor x_i. Uma alternativa simples seria estimar um modelo de regressão linear usando $p(x_i)$ como variável dependente e usar as variáveis $x_{i1}, x_{i2}, \cdots, x_{iK}$ como variáveis independentes, obtendo

$$p(x_i) = P(y_i = 1/x_i) = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_K x_{iK} + u_i. \tag{9.2}$$

Embora esse modelo, conhecido como **modelo de probabilidade linear**, seja algumas vezes usado empiricamente,[1] ele não é adequado, pois depois de estimado[2] não há nada que garanta que o lado direito da Eq. (9.2) esteja limitado ao intervalo unitário $[0, 1]$, que é faixa possível de valores para uma variável que representa uma medida de probabilidade. Uma variação interessante desse modelo, onde se restringe os valores do modelo de probabilidade linear ao intervalo $[0, 1]$ utilizando uma função de distribuição acumulada uniforme, é discutida em Ruud (2000).

Neste capítulo, com o intuito de estudar o problema descrito pela Eq. (9.1), teremos o primeiro contato com uma classe mais ampla de modelos que aquela discutida no Capítulo 8, que inclui os modelos conhecidos como **modelos lineares generalizados**.[3] Mais especificamente, nesse capítulo, introduziremos dois **modelos de resposta binária** conhecidos como **logit** e **probit**, que são os modelos mais comuns utilizados para lidar com o problema descrito pela Eq. (9.1).

Outra questão interessante que será discutida nesse capítulo é como usar os modelos logit e probit para fazer classificação entre dois tipos de objetos em uma determinada amostra. Por exemplo, suponha que desejamos saber numa amostra em firmas quais estão em má situação financeira e quais firmas que estão operando normalmente. Esses modelos podem também ser aplicados para responder esse tipo de pergunta. Ainda é válido comentar que algumas referências sobre o tema de modelos de resposta binária influenciaram parcialmente este capítulo tais como Amemiya (1985), Ruud (2000), Wooldridge (2001), Wooldridge (2003) e Davidson e MacKinnon (2004).

Este capítulo é dividido em duas seções. Na Seção 9.2, introduziremos a hipóteses que estão por de trás dos modelos de resposta binária e depois usaremos essas hipóteses para delinear uma rota para estimação e teste de hipóteses nesses modelos. No fim dessa seção discutiremos ainda a qualidade do modelo estimado. Na Seção 9.3, discutiremos a aplicação desses modelos em classificação. Finalizaremos esse capítulo com a Seção 9.4 que enumera algumas possíveis extensões dos modelos apresentados aqui.

9.2 Modelos com resposta binária

O objetivo dos modelos de regressão estudados nesta seção é explicar a probabilidade de ocorrer $y_i = 1$ ou $y_i = 0$ usando variáveis explicativas como apresentado na Eq. (9.1). Como estudamos na Seção 3.1, uma vez que y_i assume valores 0 e 1, ele é uma variável aleatória de Bernoulli. De fato, a diferença daqui para

[1] Embora esse modelo não seja adequado, muitas vezes ele é uma boa aproximação para o problema.

[2] É válido ressaltar que embora esse modelo seja linear e satisfaça a Hipótese 8.1, pode-se mostrar que ele não satisfaz a Hipótese 8.4. Dessa forma, embora o método dos mínimos quadrados gere estimadores dos coeficientes de regressão linear não viesados, esses estimadores são não eficientes. Uma técnica que pode ser usada para estimar esses coeficientes é a técnica conhecida como mínimos quadrados ponderados. Para detalhes, consultar Wooldridge (2001).

[3] Modelos lineares generalizados são extensões dos modelos lineares que são especificados usando duas funções. A primeira função é uma função conexão (*link*) que descreve como a média do modelo depende do modelo linear, isto é, $g(\mu_i) = x_i\beta$, μ_i é a média. A segunda função chamada de função variância especifica como a variável dependente depende da média $\text{var}(\mu)$, isto é, $\text{var}(y) = \phi\text{var}(\mu)$.

a apresentação na Seção 3.1 é que y_i está condicionado a x_i. Portanto, tanto a média quanto a variância de y_i irão depender do conjunto de características em x_i.

9.2.1 Hipóteses dos modelos de resposta binária

Nesta seção explicitamos as hipóteses que estão por trás dos modelos de resposta binária. A primeira hipótese dessa seção restringe à classe de modelos que iremos estudar nesse capítulo.

Hipótese 9.1 (Forma funcional) Seja y_i a i-ésima observação da chamada variável dependente e $x_i = [x_{i1}, x_{i2}, \cdots, x_{iK}]'$ a i-ésima observação dos K regressores. Então

$$p(x_i) = P(y_i = 1/x_i) = F(x_i'\beta), \tag{9.3}$$

onde F é uma função de distribuição acumulada e, portanto, a fuução $F(\cdot)$ está restrita ao intervalo $0 < F(z) < 1$, $\forall z \in \Re$.

Nota 9.1 Em geral, nos modelos de resposta binária, consideramos que o primeiro elemento do vetor x_i, o termo x_{i1}, é uma constante igual a 1. Isso significa que o intercepto está sendo incluído no lado direito da regressão.

Neste capítulo, estudaremos duas formas funcionais para a função F, que estão explicitadas nos exemplos abaixo. O primeiro trata dos modelos probit, e o segundo trata dos modelos logit.

Exemplo 9.1 (Probit) O modelo probit é um caso especial do modelo apresentado na Eq. (9.3) onde

$$F(z) = \Phi(z) = \int_{-\infty}^{z} \phi(v)dv, \tag{9.4}$$

onde $\phi(v)$ é a distribuição normal padronizada (vide Seção 3.3). O modelo probit tem sido muito utilizado em estudos empíricos em Economia. O motivo é que, em modelos microfundamentados em teoria econômica, supõe-se a existência de variáveis latentes (não observadas) u_i, que possuem distribuição normal. Quando u_i está acima de um valor de corte c, observa-se $y_i = 1$; caso contrário, observa-se $y_i = 0$. Supõe-se adicionalmente que variável u_i é uma função linear (no mesmo estilo da regressão linear discutida no capítulo anterior), de um conjunto de covariáveis no vetor x_i. Pode-se mostrar que essas suposições conduzem a um modelo probit para a variável y_i, como função das covariáveis x_i (vide discussão mais adiante).

Exemplo 9.2 (Logit) O modelo logit é um caso especial do modelo descrito pela Eq. (9.3) onde $\phi(v)$ é a distribuição logística

$$F(z) = \Lambda(z) = \frac{\exp(z)}{1 + \exp(z)}. \tag{9.5}$$

Na Figura 9.1 comparamos as funções de distribuição normal acumulada usada no Exemplo 9.1 com a distribuição logística acumulada utilizada no Exemplo 9.2. Embora o comportamento qualitativo das duas distribuições seja o mesmo, é notável a diferença do comportamento das caudas dessas distribuições, onde as caudas da distribuição logística convergem mais lentamente para zero. De fato, para o intervalo $0.1 \leq z \leq 0.9$, as funções logit e probit possuem uma relação praticamente linear. Por essa razão, normalmente é difícil discriminar entre esses dois tipos de especificações, com base em medidas de qualidade do ajuste (CHAMBERS; COX, 1967). Usando a Eq. (9.3), sabendo que y_i, para $i = 1, \cdots, n$, é uma variável aleatória

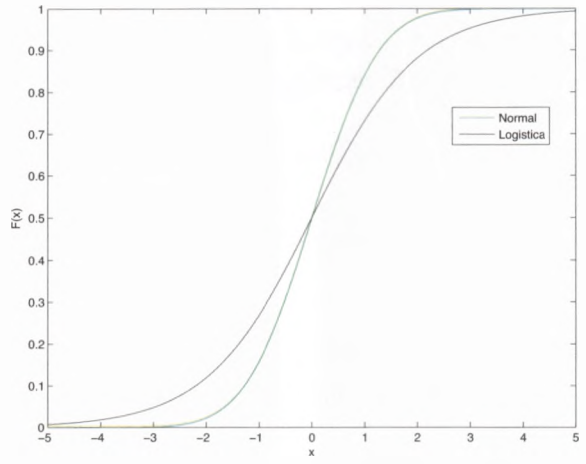


Figura 9.1: Comparação das funções de distribuição acumulada normal e logistica.

de Bernoulli, então

$$\mathrm{E}\big[y_i/x_i\big] = P(y_i = 1/x_i) \times 1 + P(y_i = 0/x_i) \times 0 = p(x_i) = F(x_i'\beta). \tag{9.6}$$

Considere que x_{ij} é uma variável contínua. Se calcularmos a derivada parcial de $p(x_i)$ em relação a x_{ij}, para $1 \leq j \leq K$, usando a Eq. (9.3), chegamos a

$$\frac{\partial p(x_i)}{\partial x_{ij}} = f(x_i'\beta)\beta_j, \tag{9.7}$$

onde $f(z) = \frac{dF(z)}{dz}$. Portanto, o efeito parcial de x_{ij} depende de $f(x_i'\beta)$. Se F é uma função estritamente crescente como nos casos dos Exemplos 9.1 e 9.2, então $f(z) > 0, \forall z$. Portanto, o sinal do efeito de x_{ij} em $p(x_i)$ é dado pelo sinal de β_j. De fato, podemos verificar também que os efeitos relativos não dependem de x_i. Para duas variáveis contínuas x_j e x_k, a razão entre os efeitos parciais é constante e dado pela razão entre os coeficientes correspondentes.

$$\frac{\partial p(x_i)/\partial x_{ij}}{\partial p(x_i)/\partial x_{ik}} = \frac{\beta_j}{\beta_k}. \tag{9.8}$$

Nota 9.2 (Interpretação usando variáveis não observadas) Em geral, é comum motivar modelos de escolha binária dados pela Eq. (9.3) como a observação parcial de uma variável não observada y^* dada por

$$y_i^* = x_i'\beta + u_i, \quad y_i = 1 \text{ quando } y_i^* > 0, \tag{9.9}$$

onde u_i é uma variável independente de X com distribuição acumulada F que deve ser contínua e simétrica em torno do ponto 0 como àquelas apresentadas nos Exemplos 9.1 e 9.2. Portanto,

$$P(y_i = 1/x_i) = P(y_i^* > 0/x_i) = P(u_i > -x_i'\beta/x_i) = 1 - F(-x_i'\beta) = F(x_i'\beta), \tag{9.10}$$

que resulta na Eq. (9.3).

Nota 9.3 (Medindo o efeito parcial de variáveis *dummy*) Considere um modelo de resposta binária, onde o K-ésimo regressor é uma variável *dummy*

$$p(x_i) = F(\beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_{K-1} x_{i\,K-1} + \beta_K x_{iK} + u_i) \quad \text{para } i = 1, 2, \cdots, n, \tag{9.11}$$

O efeito parcial de x_{iK} na probabilidade $p(x_i)$ pode ser medido fazendo $p(x_i/x_{iK} = 1) - p(x_i/x_{iK} = 0)$.

Reapresentamos a seguir a hipótese de não-multicolinearidade perfeita entre as variáveis do lado direito da regressão. Essa hipótese foi vista para o caso de modelos lineares (Capítulo 8), e é importante também para os modelos de resposta binária discutidos neste capítulo.

Hipótese 9.2 (Ausência de multicolinearidade perfeita) O posto da matriz de dados X de ordem $n \times K$ é K com probabilidade 1.

Uma outra hipótese importante para a estimação de modelos de resposta binária é similar em ideia à Hipótese 8.4 necessária para o estudo do modelo de regressão linear. Como veremos na Seção 9.2.2 essa hipótese será fundamental para a construção da função de máxima verossimilhança que será usada para a estimação dos modelos de resposta binária.

Hipótese 9.3 (Independência entre as observações) As observações aleatórias y_i, para $i = 1, \cdots, n$, são independentes.[4]

9.2.2 Estimação dos modelos de resposta binária

A estimação dos modelos de resposta binária estudados neste capítulo pode ser feita usando o método de máxima verossimilhança já estudado nas Seções 6.4.2 e 8.2.3. Usando o fato que y_i é uma variável aleatória

[4] As observações não são identicamente distribuídas, pois as médias de y_i não são as mesmas (variam com as covariáveis).

de Bernoulli, com média condicional ao vetor x_i, e que a Eq. (9.3) é válida, então a função densidade de y_i dado x_i é dada por

$$f(y_i/x_i) = F(x_i'\beta)^{y_i}(1 - F(x_i'\beta))^{1-y_i}, \ \forall \ y_i \in \{0, 1\}. \tag{9.12}$$

Supondo que a Hipótese 9.3 é válida, então a função de máxima verossimilhança é dada por

$$L(\beta/X) = \prod_{i=1}^{n} f(y_i/x_i) = \prod_{i=1}^{n} [F(x_i'\beta)^{y_i}(1 - F(x_i'\beta))^{1-y_i}], \tag{9.13}$$

e finalmente a função de log verossimilhança tem expressão

$$\log L(\beta/X) = \sum_{i=1}^{n} \log f(y_i/x_i) = \sum_{i=1}^{n} [y_i \log F(x_i'\beta) + (1 - y_i) \log(1 - F(x_i'\beta))]. \tag{9.14}$$

Exemplo 9.3 (Continuação do Exemplo 9.1 – Função de log verossimilhança para o modelo probit) Substituindo o valor de $F(\cdot)$ dado pela Eq. (9.4) na Eq. (9.14), chegamos à função de log verossimilhança para o modelo probit

$$\log L(\beta/X) = \sum_{i=1}^{n} [y_i \log \Phi(x_i'\beta) + (1 - y_i) \log(1 - \Phi(x_i'\beta))]. \tag{9.15}$$

Exemplo 9.4 (Continuação do Exemplo 9.2 – Função de log verossimilhança para o modelo logit) Substituindo o valor de $F(\cdot)$ dado pela Eq. (9.5) na Eq. (9.14), chegamos à função de log verossimilhança para o modelo logit

$$\log L(\beta/X) = l(\beta/X) = \sum_{i=1}^{n} [y_i \log \Lambda(x_i'\beta) + (1 - y_i) \log(1 - \Lambda(x_i'\beta))], \ \forall y_i \in \{0, 1\}. \tag{9.16}$$

Calculando as derivadas parciais de $l(\beta/X)$ em relação a β, chegamos a

$$\frac{\partial l(\beta/X)}{\partial \beta} = \sum_{i=1}^{n} \left\{ y_i \left[\frac{f(x_i'\beta)x_i}{F(x_i'\beta)} \right] - (1 - y_i) \left[\frac{f(x_i'\beta)x_i}{1 - F(x_i'\beta)} \right] \right\} = \sum_{i=1}^{n} \frac{f(x_i'\beta)x_i(y_i - F(x_i'\beta))}{F(x_i'\beta)(1 - F(x_i'\beta))}. \tag{9.17}$$

O estimador de máxima verossimilhança é a solução do sistema de equações não-lineares

$$\sum_{i=1}^{n} \frac{f(x_i'\hat{\beta})x_i(y_i - F(x_i'\hat{\beta}))}{F(x_i'\hat{\beta})(1 - F(x_i'\hat{\beta}))} = 0. \tag{9.18}$$

Pode-se mostrar que, quando a Hipótese 9.2 é válida, a solução desse problema é única para o caso dos modelos probit e logit, visto que $l(\beta/X)$ é concava em β.[5] Infelizmente, devido à não linearidade do problema, não é possível encontrar uma solução analítica para o estimador de máxima verossimilhança de β, tanto para o modelo probit, quanto para o modelo logit, nas Eqs. (9.15) e (9.16). Dessa forma, precisamos buscar uma solução numérica para esse problema. De fato, a boa notícia é que a maioria dos softwares econométricos livres ou comerciais já fazem isso.

Exemplo 9.5 (Exemplo numérico de modelos de resposta binária) Neste exemplo, mais uma vez utilizaremos simulações Monte Carlo para gerar os dados que serão usados para estimar um modelo de resposta binária. Diferentemente da geração de dados no modelo de regressão linear (vide Exemplo 8.8), aqui precisaremos tomar alguns cuidados especiais. A sequência básica para podermos gerar os dados desse modelo é a seguinte:

1) Criar n valores de um vetor de ordem K de regressores com uma distribuição pré-especificada.

2) Para cada um desses n valores, calcular $F(x_i'\beta)$ usando os valores reais para β (lembre-se que estamos fazendo o papel da natureza), e a fórmula para distribuição acumulada $F(\cdot)$ desejada (no caso, normal ou logística).

3) Sortear n valores de uma variável aleatória uniforme w_i, $i = 1, \cdots, n$, no intervalo $[0, 1]$.

4) Gerar os n valores de y_i fazendo o seguinte teste: se $w_i \in [0, F(x_i'\beta)]$ então $y_i = 1$; caso contrário, $y_i = 0$.

Vamos então construir um modelo de resposta binária com três regressores. O primeiro regressor será $x_{i1} = 1$, para todo $i = 1, \ldots, n$. O segundo regressor será uma variável aleatória $x_{i2} \sim \text{Normal}(\mu_{x_2}, \sigma^2_{x_2})$ onde $\mu_{x_2} = 1$ e $\sigma^2_{x_2} = 0.36$ e o terceiro regressor $x_{i3} \sim \text{Normal}(\mu_{x_3}, \sigma^2_{x_3})$ onde $\mu_{x_3} = 0$ e $\sigma^2_{x_3} = 0.64$, com $\beta_1 = 1$, $\beta_2 = -1$ e $\beta_3 = 1$.

Note que a escolha acima foi cuidadosa para conseguirmos um modelo bem balanceado. Considerando que a média de $x_i'\beta$ é zero, se a variância de $x_i'\beta$ for muito pequena, todos os valores ficarão em torno $F(x_i'\beta) = 0.5$. Se a variância de $x_i'\beta$ for muito grande, então todos os pontos gerados estarão no extremo da distribuição. Então, fizemos média de $x_i'\beta$ igual a 0 e variância de $x_i'\beta$ igual a 1. Vamos supor que o modelo que gera os dados é o probit. No caso de utilizarmos um modelo logit, todos os passos são completamente similares. Geramos uma amostra com 30 observações, apresentadas na tabela 9.1. Maximizando-se numericamente a Eq. (9.4), encontramos $\hat{\beta}_1 = 1.0058$, $\hat{\beta}_2 = -0.7761$ e $\hat{\beta}_3 = 1.1000$.

9.2.3 Teste de hipóteses nos modelos de resposta binária

Da mesma forma que para o modelo de regressão linear (vide Seção 8.2.3), quando estimamos os parâmetros usando o método de máxima verossimilhança, para lidar com testes de hipóteses no modelo de resposta

[5] Veja Exercício 9.3, o teorema de Pratt (1981) e também a discussão apresentada em Ruud (2000) ou Amemiya (1985).

Tabela 9.1: Dados simulados para a regressão probit

x_1	x_2	x_3	$F(x_i'\beta)$	w	y
1	0.9657	-0.4591	0.3355	0.2909	1
1	0.876	0.3164	0.6702	0.0484	1
1	1.3828	-0.2762	0.2549	0.0395	1
1	0.9754	-0.3609	0.3683	0.5046	0
1	1.2136	0.7084	0.6896	0.3671	1
1	0.6069	-1.265	0.1916	0.9235	0
1	-0.1401	-0.5381	0.7264	0.5968	1
1	0.4034	0.1266	0.7652	0.8085	0
1	1.4103	-0.4291	0.2006	0.9253	0
1	0.7218	1.9531	0.9872	0.3628	1
1	0.6683	-0.4619	0.4482	0.215	1
1	0.7826	-1.1082	0.1865	0.136	1
1	2.2683	0.3435	0.1775	0.4923	0
1	0.5399	-1.0142	0.2898	0.7836	0
1	1.0226	1.1726	0.8749	0.1714	1
1	1.1626	0.0556	0.4574	0.187	1
1	0.1819	0.8806	0.9553	0.6035	1
1	-0.5576	0.627	0.9855	0.8332	1
1	0.6422	-0.0757	0.6111	0.8088	0
1	2.1958	0.0866	0.1337	0.2878	0
1	1.2708	-0.6903	0.1683	0.0822	1
1	0.8722	-0.7896	0.2541	0.6477	0
1	1.2938	-0.4827	0.2187	0.3266	0
1	0.581	-0.2088	0.5833	0.6118	0
1	0.7974	-0.8146	0.2703	0.3336	0
1	0.7444	1.3183	0.9423	0.4368	1
1	0.8979	0.9489	0.8534	0.1081	1
1	0.5505	0.669	0.8683	0.3113	1
1	0.2186	0.7388	0.9358	0.3851	1
1	0.8632	-1.2637	0.1299	0.6798	0

binária, precisamos calcular a matriz de informação de Fisher. Utilizando a Eq. (9.17) e calculando a segunda derivada em relação a β chegamos a

$$\frac{\partial^2 \log L(\beta/X)}{\partial \beta^2} = \sum_{i=1}^{n} \left\{ -\frac{f(x_i'\beta)^2 x_i x_i'}{F(x_i'\beta)(1-F(x_i'\beta))} + (y_i - F(x_i'\beta)g(x_i'\beta)) \right\}, \tag{9.19}$$

onde $g(x_i'\beta)$ é uma função de $x_i'\beta$.

Lembrando que a matriz de informação de Fisher $I(\beta/X) = -E\left[\frac{\partial^2}{\partial\beta\partial\beta'} l(\beta)/X\right]$ é calculada para o β verdadeiro, trocando de ordem o valor esperado e o somatório e notando que $E[(y_i - F(x_i'\beta)g(x_i'\beta)/x_i] = 0$, chegamos a

$$I(\beta/X) = \sum_{i=1}^{n} \frac{f(x_i'\beta)^2 x_i x_i'}{F(x_i'\beta)(1-F(x_i'\beta))}, \tag{9.20}$$

cuja inversa é dada por

$$I^{-1}(\beta/X) = \left\{ \sum_{i=1}^{n} \frac{f(x_i'\beta)^2 x_i x_i'}{F(x_i'\beta)(1-F(x_i'\beta))} \right\}^{-1}. \tag{9.21}$$

Note que, para as funções de distribuição acumulada usuais, a Hipótese 9.2 é suficiente para garantir que essa inversa exista e, nesse caso, ela é positiva definida.[6] A inversa da matriz de Fisher, similarmente às Seções 6.4.2 e 8.2.3, tem um papel fundamental, pois ela é igual à matriz de variância-covariância dos parâmetros estimados $\hat{\beta}$. Por não dispormos dos valores reais dos parâmetros do modelo, conforme Nota 5.1, substituímos os valores reais desses parâmetros por seus valores estimados, o que é razoável visto que as estimativas de máxima verossimilhança são consistentes.

Nas Seções 6.4.2 e 8.2.3, vimos também que, para fazer testes de hipóteses no contexto de estimação usando máxima verossimilhança, podemos usar uma das três possibilidades: teste de Wald, teste de razão de verossimilhança e teste dos multiplicadores de Lagrange. Embora todos esses três testes tenham distribuição qui-quadrada com r graus de liberdade (r é o número de restrições), existem algumas diferenças entre eles. Conforme discutido nas Seções 6.4.2 e 8.2.3, para se usar o teste de Wald, é necessário estimar apenas o modelo irrestrito. Para o teste de multiplicadores de Lagrange, precisa-se estimar apenas o modelo restrito. Finalmente, para o teste de razão de verossimilhança, precisa-se estimar os dois modelos (restrito e irrestrito). Dessa forma, dependendo das restrições impostas, pode ser mais simples usar um teste ou outro. Por exemplo, se as restrições forem apenas restrições de exclusão, nesse caso o teste dos multiplicadores de Lagrange pode ser mais simples, pois o número de parâmetros é menor. Se as restrições forem complicadas, pode ser mais simples utilizar o teste de Wald.

Exemplo 9.6 (Continuação do Exemplo 9.5 – Teste conjunto de hipóteses em modelos de resposta binária) Neste exemplo vamos testar a hipótese nula de que os coeficientes β_2 e β_3 do modelo apresentado no Exemplo

[6] Usando a Hipótese 9.2 e algumas condições de regularidade, a garantia da existência da inversa e concavidade da função de log-verossimilhança é apresentada em Amemiya (1985) para os casos dos modelos logit e probit.

9.5 são conjuntamente nulos. O primeiro passo é calcular a matriz de variância-covariância dada pela Eq. (9.21). A matriz resultante é dada por

$$I^{-1} = \begin{bmatrix} 0.3141 & -0.2419 & 0.0675 \\ -0.2419 & 0.2504 & -0.0286 \\ 0.0675 & -0.0286 & 0.1796 \end{bmatrix}$$

Fazendo $h(\hat{\beta}) = \begin{bmatrix} \hat{\beta}_2 \\ \hat{\beta}_3 \end{bmatrix}$, e utilizando a estatística de Wald apresentada na Eq. (6.15), chegamos à estatística teste $W = 8.2027$, e ao p-valor $p = 1 - F_{\chi^2_2}(W) = 0.0166$. Portanto, a hipótese nula é rejeitada ao nível de significância de 5%.

A mesma hipótese nula conjunta pode ser testada calculando-se a estatística teste de razão de verossimilhança, apresentada na Eq. (6.17). Estimando-se os parâmetros dos modelos restrito e irretrito, e calculando-se a estatística LRT, obtem-se $LRT = 11.6712$. Calculando o p-valor $p = 1 - F_{\chi^2_2}(LRT) = 0.0029$, rejeitamos a hipótese nula aos níveis usuais de 1% ou 5%. Finalmente, calculando-se a estatística dos multiplicadores de Lagrange apresentada na Eq. (6.19), obtém-se $LM = 9.5534$, com p-valor correspondente $p = 1 - F_{\chi^2_2}(LM) = 0.0084$. Nesse caso, a hipótese nula também é rejeitada aos níveis de significância de 1% e 5%.

9.2.4 Qualidade do modelo de resposta binária

As hipóteses do modelo são válidas?

Diferentemente do modelo de regressão linear, para o qual as hipóteses apresentadas na Seção 8.1 são mais restritivas (por exemplo, as hipóteses sobre a estrutura do erro), as hipóteses dos modelos de resposta binária são mais simples. Basicamente temos três hipóteses: (1) uma que versa sobre a linearidade do modelo e sua correta especificação (Hipótese 9.1), (2) outra mais técnica sobre a matriz de dados (Hipótese 9.2) e (3) finalmente uma que exige independência das observações (Hipótese 9.3).

Em geral, para se verificar a Hipótese 9.1 sobre a correta especificação do modelo, podemos proceder de forma similar ao discutido na Seção 8.3. Uma opção é rodar regressões auxiliares e testar conjuntamente a hipótese que os coeficientes dos termos não lineares (ou de interação entre dois regressores) são nulos ou testar a ausência de novos regressores. Uma outra forma é proceder ao estudo gráfico do comportamento do ruído. No entanto, se, por exemplo, plotarmos explicitamente o ruído $\hat{u}_i = y_i - E[y_i/x_i] = y_i - F(x_i'\hat{\beta})$ versus $F(x_i'\hat{\beta})$ conseguiremos extrair muito pouca informação. Isso ocorre porque os y_is são discretos e consequentemente os resíduos (erros) do modelo também. Uma forma para lidar com esse problema é proceder de forma similar a Cleveland (1979) e Gelman (2007) e plotar gráficos com ruídos suavizados, como nos Exemplos 9.7 a 9.10.

Como foi dito acima, a Hipótese 9.2 é mais técnica e se ela não for válida, poderá haver problemas na inversão da matriz de Fisher dada pela Eq. (9.20). Na prática, se a matriz de informação de Fisher tiver inversa numericamente, tem-se indícios de que a hipótese de não-multicolinearidade perfeita está sendo satisfeita (vide Exercício 9.1).

Exemplo 9.7 (Continuação do Exemplo 9.5 – Modelo corretamente especificado) Usando o mesmo gerador de dados do Exemplo 9.5, geramos uma amostra com 500 observações e estimamos o mesmo modelo utilizado para gerar os dados. A Figura 9.2 apresenta um gráfico de $\hat{u}$ versus $F(x_i'\hat{\beta})$, no qual é fácil verificar a estrutura discreta dos dados. A disposição de dados como apresentada nessa figura ocorre também em modelos com especificação incorreta, dificultando o uso dessa figura para analisar se o modelo foi especificado corretamente. Ao invés de usar explicitamente essa figura, analisaremos uma variação desse gráfico onde suavizamos o erro $\hat{u}$, como mostra a Figura 9.3.

No primeiro gráfico, dividimos a amostra com 500 elementos em 25 categorias diferentes (cada uma com 20 elementos), de acordo com o tamanho de $F(x_i'\hat{\beta})$. Por exemplo, a primeira categoria contem os 20 elementos com menores valores de $F(x_i'\hat{\beta})$, a segunda categoria contem os 20 elementos seguintes que possuem os menores elementos de $F(x_i'\hat{\beta})$, mas que não eram pequenos o suficiente para entrar na primeira categoria e assim por diante. Finalmente, a vigésima-quinta categoria contem os elementos que possuem os maiores $F(x_i'\hat{\beta})$. Então calculamos a média de $\hat{u}_i$ e $F(x_i'\hat{\beta})$ em cada categoria e colocamos no gráfico esses valores. No segundo gráfico, o procedimento é o mesmo, mas a categorização é feita a partir do tamanho da variável x_2 e o terceiro gráfico é feita a partir do tamanho da variável x_3. O quarto gráfico é o QQ plot da distribuição dos valores médios de $\hat{u}_i$ para cada categoria construída a partir do tamanho de u_i versus a distribuição normal. Nos gráficos apresentados nessa figura, os ruídos suavizados (médios) parecem bem comportados e embora não termos feito essa hipótese explicitamente, a distribuição dos ruídos suavizados é muito próxima da distribuição normal.

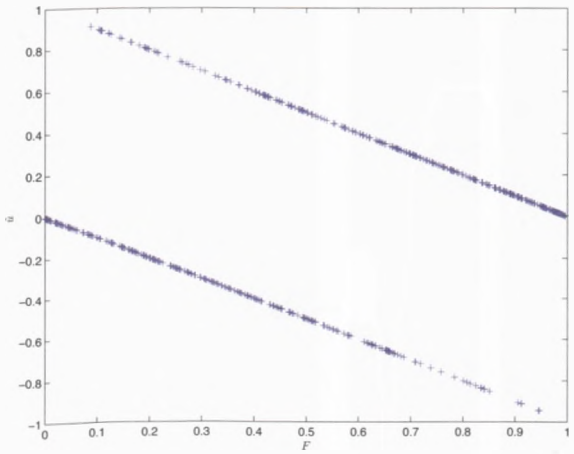


Figura 9.2: Estrutura discreta do gráfico $\hat{u} \times F(x_i'\hat{\beta})$.

Exemplo 9.8 (Continuação do Exemplo 9.5 – Omissão de variáveis) Usando o mesmo gerador de dados do Exemplo 9.5, geramos uma amostra com 500 observações. A diferença aqui é que estimamos um modelo

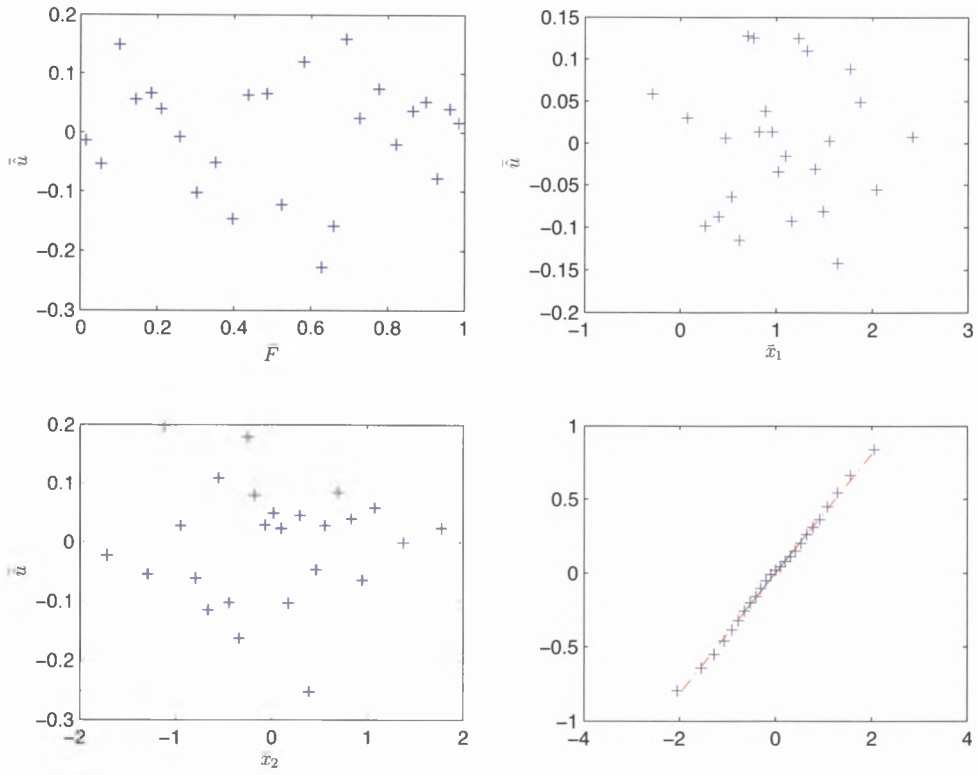


Figura 9.3: Análise do ruído suavizado em um modelo corretamente especificado.

que omite a variável x_2. Na Figura 9.4 apresentamos os mesmos gráficos já discutidos no Exemplo 9.7. No gráfico onde é apresentado o erro suavizado versus a variável x_2, é clara a relação linear entre essas duas variáveis. Note também que o gráfico QQ-plot também é afetado.

Exemplo 9.9 (Continuação do Exemplo 9.5 – Erro de especificação com padrão não-linear) Nesse exemplo, geramos uma amostra com 500 observações de forma similar ao Exemplo 9.5. A única diferença aqui é que o segundo regressor x_2 tenha sido substituído pelo regressor x_2^2. Entretanto, embora tenhamos feito essa modificação no modelo gerador de dados, estimamos o modelo como no Exemplo 9.5, considerando uma estrutura linear de todos os regressores. Na Figura 9.5 apresentamos os mesmos gráficos já discutidos no Exemplo 9.7 para esse caso. Conforme observado nessa figura, no gráfico onde é apresentado o erro suavizado *versus* a variável x_2, é clara a relação não-linear entre essas duas variáveis. Note também que o gráfico QQ-plot também é afetado, aparecendo nesse gráfico um padrão não-linear.

Exemplo 9.10 (Continuação do Exemplo 9.5 – Geração de dados usando o logit e estimação usando o probit)

Neste exemplo, geramos uma amostra com 500 observações de forma similar ao Exemplo 9.5. A única diferença é que usamos o modelo logit para gerar os dados. Entretanto, mesmo assim, estimamos o modelo probit. Na Figura 9.6 apresentamos os mesmos gráficos já discutidos no Exemplo 9.7 para esse caso. Como vemos nessa figura, podemos identificar a diferença entre a distribuição do erro suavizado e a distribuição

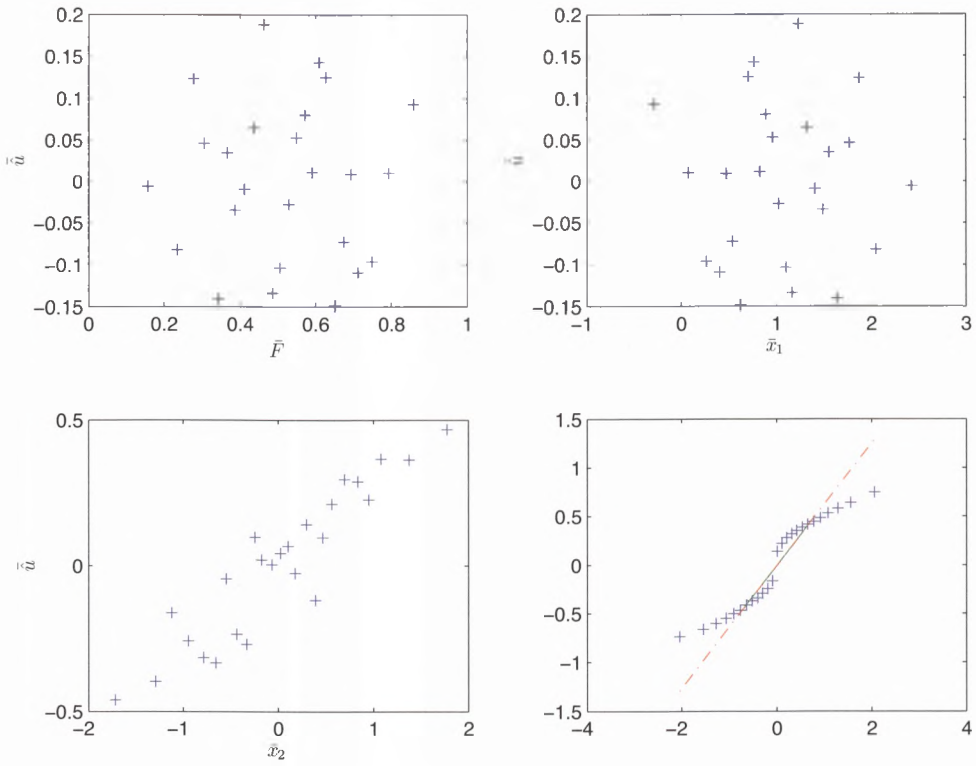


Figura 9.4: Análise do ruído suavizado em um modelo em que se omite o regressor x_2.

normal. Note que a maior diferença ocorre nas caudas onde justamente ocorre a maior diferença entre as distribuições logística e normal.

Medidas de qualidade do ajuste do modelo

Na seção anterior, apresentaram-se alguns procedimentos, utilizando-se recursos gráficos e com base nos resíduos do modelo de regressão, para avaliação da adequação dos modelos de resposta binária aos dados observados na amostra. A literatura estatística apresenta também algumas medidas numéricas para complementar a avaliação da qualidade do ajuste do modelo aos dados. Uma das medidas mais comuns é o chamado R^2_{pseudo}, proposto por McFadden (1974), que pode ser calculado utilizando-se a expressão

$$R^2_{\text{pseudo}} = 1 - \frac{l(\hat{\beta}/X)}{l(\tilde{\beta}_{\text{intercepto}}/X)}, \tag{9.22}$$

onde $l(\hat{\beta}/X)$ é função de log-verossimilhança do modelo estimado e $l(\tilde{\beta}_{\text{intercepto}}/X)$ é a função de log-verossimilhança de um modelo que contém apenas o intercepto. Se as variáveis dependentes utilizadas no modelo não tiverem poder explicativo algum, então $\frac{l(\hat{\beta}/X)}{l(\tilde{\beta}_{\text{intercepto}}/X)} = 1$, implicando que o R^2_{pseudo} será zero. Caso contrário, $\frac{l(\hat{\beta}/X)}{l(\tilde{\beta}_{\text{intercepto}}/X)} < 1$, pois $l(\hat{\beta}/X)$ e $l(\tilde{\beta}_{\text{intercepto}}/X)$ são ambos negativos (uma vez que

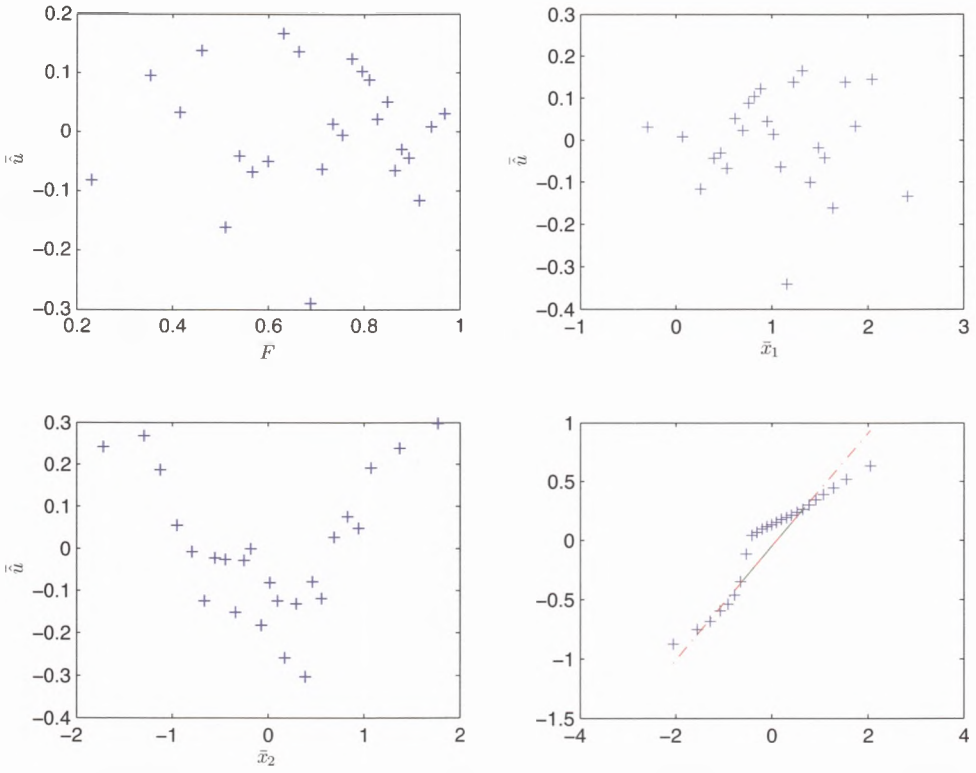


Figura 9.5: Análise do ruído suavizado em um modelo com erro de especificação, o segundo regressor apresenta um padrão quadrático.

$0 \leq F(\cdot) \leq 1$), e $l(\tilde{\beta}/X) \geq l(\tilde{\beta}_{\text{intercepto}}/X)$. Essa última desigualdade deve-se ao fato de que, em um problema de otimização restrita, o máximo encontrado é sempre menor ou igual ao máximo encontrado em problema de maximização irrestrita; portanto, $l(\tilde{\beta}_{\text{intercepto}}/X)$ será mais negativo que $l(\tilde{\beta}/X)$.

Exemplo 9.11 (Continuação do Exemplo 9.5 – Qualidade do ajuste) Calculando o R^2_{pseudo} utilizando a Eq. (9.2.4) e os dados do Exemplo 9.5, chegamos ao $R^2_{\text{pseudo}} = 1 - \frac{-14.6914}{-20.5270} = 0.2843$.

Da mesma forma que no caso do R^2 discutido para os modelos lineares de regressão, para os modelos de resposta binária, quando adiciona-se uma variável independente a mais no lado direito da regressão binária, o R^2_{pseudo} necessariamente irá apresentar um valor maior ou igual ao valor do R^2_{pseudo} anterior. A explicação para isso é novamente o fato de que, em uma maximização irrestrita, o valor final obtido para a função objetivo será necessariamente superior ou igual ao valor obtido quando impõem-se restrições à maximização. Diante desse problema de aumento artificial da estatística R^2-pseudo, pode-se utilizar alternativamente o $R^2_{\text{pseudo ajustado}}$, que possui expressão

$$R^2_{\text{pseudo ajustado}} = 1 - \frac{l(\tilde{\beta}/X) - (K-1)}{l(\tilde{\beta}_{\text{intercepto}}/X)},$$

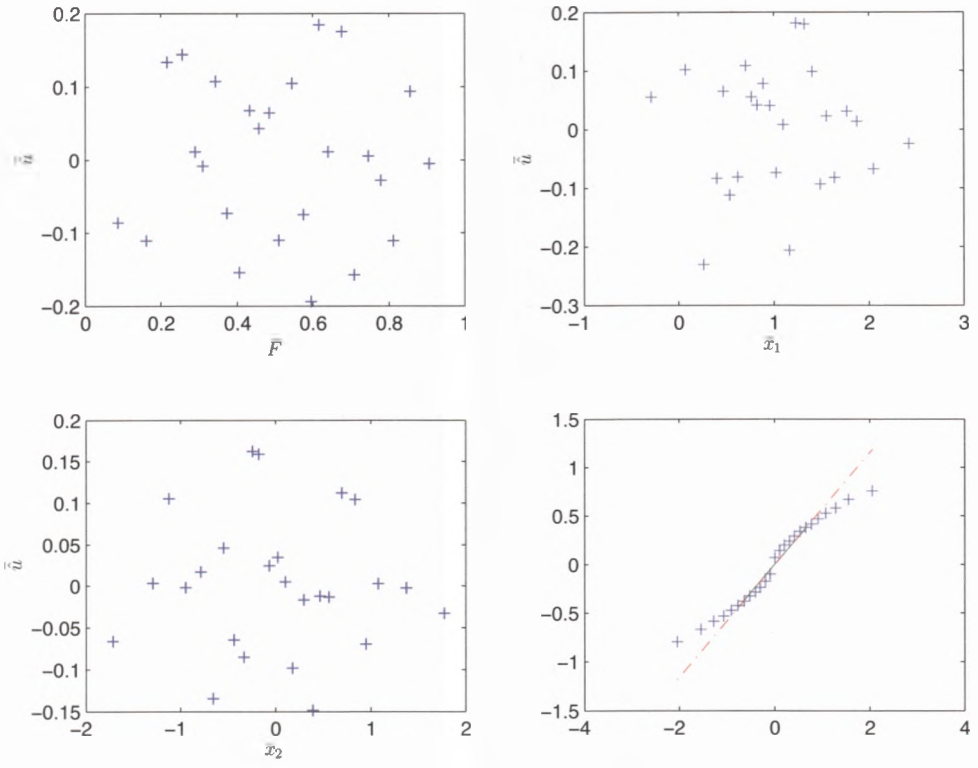


Figura 9.6: Análise do ruído suavizado em um modelo com erro de especificação onde os dados são gerados utilizando o logit e o modelo estimado como um probit.

onde K é o número de preditores no lado direito da equação (incluindo o intercepto) do modelo de regressão binária. Portanto, o termo $K - 1$ corresponde ao número de preditores, além do intercepto, na equação. Quanto maior o número de preditores adicionados, menor o numerador na expressão para $R^2_{\text{pseudo ajustado}}$, penalizando a adição indevida de variáveis independentes.

Uma outra forma de verificar a qualidade do modelo é usando a chamada **porcentagem predita correta**.[7] Defina p_c como uma probabilidade de corte (esse valor é usualmente $\frac{1}{2}$, mas dependendo do modelo estimado esse valor pode ser escolhido diferente). Para todas as observações i, $1 \leq i \leq n$, na amostra, calculamos o número de observações n_1 para as quais acontece $y_i = 1$ e $F(x_i'\hat{\beta}) > p_c$ simultaneamente, e a quantidade n_0 de observações na amostra em que ocorre $y_i = 0$ e $F(x_i'\hat{\beta}) < p_c$. Intuitivamente, a quantidade n_1 corresponde ao número de observações em que o modelo previu corretamente a ocorrência de sucesso ($y_i = 1$), enquanto n_0 corresponde ao número de vezes em que o modelo previu corretamente a ocorrência de não-sucesso ($y_i = 0$). Então,

$$\text{porcentagem predita correta} = \frac{n_1 + n_0}{n}.$$

Conforme ressaltado em Wooldridge (2001), se a amostra não for balanceada, por exemplo tiver 70% de um 1s e apenas 30% de 0s, fica mais fácil prever 1 do que 0, fazendo com que essa medida seja pouco informativa.

[7] Em inglês, *percent correctly predicted*.

Tabela 9.2: Previsão usando o modelo do Exemplo 9.5.

x_1	x_2	x_3	$F(x_i'\beta)$	y	$F(x_i'\hat{\beta})$	$\hat{y}$	Acertou
1	0.9657	-0.4591	0.3355	1	0.4018	0	0
1	0.876	0.3164	0.6702	1	0.7498	1	1
1	1.3828	-0.2762	0.2549	1	0.3552	0	0
1	0.9754	-0.3609	0.3683	0	0.4411	0	1
1	1.2136	0.7084	0.6896	1	0.8004	1	1
1	0.6069	-1.265	0.1916	0	0.1958	0	1
1	-0.1401	-0.5381	0.7264	1	0.6994	1	1
1	0.4034	0.1266	0.7652	0	0.7973	1	0
1	1.4103	-0.4291	0.2006	0	0.2875	0	1
1	0.7218	1.9531	0.9872	1	0.9953	1	1
1	0.6683	-0.4619	0.4482	1	0.4916	0	0
1	0.7826	-1.1082	0.1865	1	0.2059	0	0
1	2.2683	0.3435	0.1775	0	0.3531	0	1
1	0.5399	-1.0142	0.2898	0	0.2985	0	1
1	1.0226	1.1726	0.8749	1	0.9334	1	1
1	1.1626	0.0556	0.4574	1	0.5654	1	1
1	0.1819	0.8806	0.9553	1	0.9666	1	1
1	-0.5576	0.627	0.9855	1	0.9833	1	1
1	0.6422	-0.0757	0.6111	0	0.6642	1	0
1	2.1958	0.0866	0.1337	0	0.2732	0	1
1	1.2708	-0.6903	0.1683	1	0.2297	0	0
1	0.8722	-0.7896	0.2541	0	0.2947	0	1
1	1.2938	-0.4827	0.2187	0	0.2983	0	1
1	0.581	-0.2088	0.5833	0	0.6275	1	0
1	0.7974	-0.8146	0.2703	0	0.3053	0	1
1	0.7444	1.3183	0.9423	1	0.9698	1	1
1	0.8979	0.9489	0.8534	1	0.9119	1	1
1	0.5505	0.669	0.8683	1	0.9056	1	1
1	0.2186	0.7388	0.9358	1	0.9504	1	1
1	0.8632	-1.2637	0.1299	0	0.1459	0	1

Uma forma de resolver esse problema é utilizar a **porcentagem predita correta ponderada** apresentada em Wooldridge (2003) e dada por

$$p_{\%} = (1 - p_1) \times q_0 + p_1 \times q_1, \tag{9.23}$$

onde p_1 é a proporção de 1s na amostra, $q_0 = n_0/n$ é a porcentagem corretamente predita do resultado $y = 0$ e $q_1 = n_1/n$ é a porcentagem corretamente predita do resultado $y = 1$. Note que se a amostra for balanceada, esse valor se reduz a porcentagem predita correta.

Exemplo 9.12 (Continuação do Exemplo 9.5 – Qualidade de previsão) Podemos também analisar a qualidade de previsão do modelo estimado no Exemplo 9.5. A tabela 9.2 compara os valores de $F(x_i'\beta)$ com $F(x_i'\hat{\beta})$, y_i com $\hat{y}_i$ e os acertos do modelo. Utilizando essa tabela, podemos calcular a porcentagem predita correta ponderada, que nesse caso é igual a $p_{\%} = 0.7333$.

9.3 Classificação usando modelos de resposta binária

Uma aplicação muito útil dos modelos deste capítulo é classificação; ou seja, podemos usar as probabilidades preditas de nosso modelo de resposta binária para classificar elementos de uma amostra como um evento ou não. De fato, quando pensamos em problemas de classificação, consideramos em aplicar o nosso modelo em elementos que não foram estimados na nossa amostra. Por exemplo, considere um banco que dispõe de uma amostra bem grande de pessoas e deseja saber se pode emprestar dinheiro para essas pessoas ou não. O custo de fazer uma avaliação cuidadosa de cada pessoa é muito alto (pois envolve muitas horas) e por isso não é possível fazer essa avaliação com cada uma delas. Entretanto, é possível dizer para uma pequena parcela dessa amostra se vale ou não vale a pena investir nessas pessoas. O problema dessa seção então é: como utilizar essa pequena amostra factível de pessoas para se ter informação sobre todo o grupo?

9.3.1 Descrição do algoritmo para classificação

Considere que o objetivo é avaliar uma amostra muito grande A e dizer se cada elemento dessa amostra é um evento ou não. Um algoritmo factível para resolver esse problema é o seguinte:

1. Escolha aleatoriamente uma pequena parcela B da amostra A.

2. Avalie se cada elemento de B é um evento ou não.

3. Rode o modelos de resposta binária para uma parte dos elementos de B, que chamaremos de C, usando como variáveis explicativas as variáveis que foram relevantes para você tomar a decisão no passo 2.

4. Faça uma escolha da probabilidade de corte p_c. A probabilidade de corte p_c será usada para discriminar os elementos da amostra $A - B$ como evento ou não evento. Note que essa etapa não é fácil, mas fundamental. Um procedimento é construir uma tabela variando p_c de valores bem baixos a valores bem altos e ver o grau de acerto que ela tem para os elementos de B que não foram usados para estimar o modelo de resposta binária (a amostra $B - C$).[8] Escolha p_c de forma que você consiga o maior índice de acerto na amostra que não foi usada para estimação.

5. Utilize seu modelo de escolha discreta para classificar sua amostra $A - B$, entre eventos e não eventos usando p_c como probabilidade de corte.

Exemplo 9.13 (Continuação do Exemplo 9.5 - Classificação) Podemos utilizar o processo gerador de dados do Exemplo 9.5 para exemplificar a metodologia de classificação discutida acima. Para isso, vamos construir 3 amostras de 30 pontos. A primeira amostra é exatamente aquela usada no Exemplo 9.5 e as outras duas amostras usam o mesmo processo gerador de dados e são apresentadas na tabela 9.3. No

[8] A vantagem de usar uma amostra não usada para estimação para a escolha da probabilidade de corte é tentar reduzir possíveis efeitos de memorização de dados. Veja uma discussão sobre esse problema na Seção 8.3.2.

Tabela 9.3: As duas amostras adicionais utilizadas no Exemplo 9.13.

Amostra 2				Amostra 3			
x_1	x_2	x_3	y	x_1	x_2	x_3	y
1	1.0811	-0.0576	0	1	1.4768	0.7518	1
1	0.9694	1.0289	1	1	1.2894	-1.0471	0
1	1.842	0.4655	0	1	1.3618	-0.0996	1
1	1.9897	-1.303	0	1	1.8596	0.7038	1
1	2.259	-0.0482	0	1	-1.0576	-0.2428	1
1	-0.3569	-0.2452	1	1	2.056	-1.5147	0
1	0.4567	-0.2101	1	1	0.4477	-1.2542	0
1	1.4885	0.3234	0	1	1.1142	-0.841	1
1	0.2906	-0.9998	0	1	0.94	0.1175	0
1	1.2143	1.8669	1	1	1.4903	-0.3779	0
1	1.8139	-0.1178	1	1	0.2217	0.7864	1
1	0.4662	1.023	1	1	0.4564	0.6487	0
1	0.6576	-1.5816	0	1	1.3074	1.0472	1
1	2.006	0.5596	0	1	1.522	0.3721	0
1	0.9812	0.1152	1	1	0.9735	-0.7159	0
1	0.001	0.9159	1	1	0.7796	-1.2416	0
1	1.2027	1.4575	1	1	-0.3954	-0.2368	1
1	1.6288	0.0423	0	1	0.752	0.554	1
1	1.0335	1.4442	1	1	1.8782	0.4822	0
1	1.7222	-0.2174	0	1	1.5404	1.868	1
1	-0.2904	-0.2418	1	1	0.8581	0.3623	1
1	1.5603	0.5058	0	1	1.3743	0.3854	0
1	1.0905	1.2555	1	1	0.3553	-1.1288	0
1	1.8469	-0.3392	0	1	0.6153	-0.6154	0
1	0.3032	-0.7868	0	1	0.8534	0.4356	1
1	0.7226	0.7772	1	1	-0.0035	-0.0026	1
1	1.5453	-0.2384	0	1	1.1488	0.2038	0
1	1.0218	0.4388	1	1	1.0051	0.1078	1
1	1.2272	-0.6765	0	1	0.7449	0.0624	1
1	1.234	-0.0328	1	1	0.4475	-1.24	0

Exemplo 9.5, estimamos os parâmetros do modelo usando a primeira amostra. Agora, vamos utilizar a segunda amostra para a escolha da probabilidade de corte. A tabela 9.4 nos dá o subsídio para a escolha desse parâmetro. Nessa tabela, calculamos a porcentagem predita ponderada dada pela Eq. (9.23), em função da probabilidade de corte usada para classificar os pontos da segunda amostra. Note que de acordo com essa $p_c = 0.6$ é o ponto de melhor escolha para a probabilidade de corte. Então, agora vamos usar esse valor para classificar a amostra 3. Na Figura 9.7, apresentamos o resultado dessa classificação para a amostra 3. Os pontos em vermelho são os pontos que possuem $y = 1$ nessa amostra e os pontos em azul são os pontos que possuem $y = 0$ nessa amostra. A reta apresentada nessa figura é $x_i'\hat{\beta} = \Phi^{-1}(p_c)$. Os pontos acima dessa reta serão classificados como 1 e abaixo dessa reta serão classificados como 0. Os pontos dessa figura foram classificados corretamente em 80% das vezes com $p_{\%} = 0.8$.

Tabela 9.4: Porcentagem predita ponderada em função da probabilidade de corte para a amostra 2.

p_c	q_0	q_1	$p_{\%}$
0.1	0.0667	1	0.5333
0.2	0.1333	1	0.5667
0.3	0.4	0.9333	0.6667
0.4	0.5333	0.9333	0.7333
0.5	0.6667	0.9333	0.8
0.6	0.9333	0.8667	0.9
0.7	1	0.7333	0.8667
0.8	1	0.6667	0.8333
0.9	1	0.5333	0.7667
1	1	0	0.5

No exemplo acima, um ponto importante é que não foram considerados pesos diferentes para erros e acertos no problema de classificação. Na prática, o analista pode preferir escolher penalidades diferentes para diferentes tipos de erros. Por exemplo, imagine novamente o problema de classificar clientes que estão pleiteando um empréstimo bancário. Um modelo via regressão logit ou probit irá então classificar o cliente como potencial inadimplente ou não. Os dois tipos de erros que podem ser incorridos no problema de classificação são: (a) classificar um bom pagador como potencial inadimplente; (b) classificar um mau pagador como cliente sem potenciais problemas. Uma análise mais rebuscada pode levar à conclusão de que o erro de classificação do tipo (a) é cinco vezes mais custoso para a instituição do que erro de classificação do tipo (b). Portanto, em uma análise mais criteriosa do problema de classificação, essas penalidades assimétricas devem ser levadas em consideração.[9]

9.3.2 Interpretação geométrica do problema de classificação perfeita

Considere que exista um vetor β para um conjunto de dados X e y que no modelo de resposta binária dado pela Eq. (9.3) satisfaça ao problema de classificação perfeita

$$\begin{aligned} y_i &= 0 \quad \text{quando} \quad x_i'\beta < 0 \\ y_i &= 1 \quad \text{quando} \quad x_i'\beta > 0. \end{aligned} \tag{9.24}$$

Quando a Eq. (9.24) é válida, dizemos que existe uma **separação completa** dos dados. A interpretação geométrica desse problema é bem simples. No espaço de dimensão K gerado pelas colunas da matriz X, o vetor β define um hiperplano que separa as observações $y_i = 1$ das observações de $y_i = 0$.

Considere agora o problema de estimação usando a função de log-verossimilhança dada pela Eq. (9.14) e os dados que satisfazem o problema dado de acordo com a Eq. (9.24). Note que o argumento da função

[9] Obviamente, a escolha das penalidades assimétricas pode ser uma tarefa não trivial, e pode levar a erros adicionais. Esses erros podem fazer com que, no geral, o processo de classificação se torne menos acurado do que se penalidades simétricas fossem supostas.

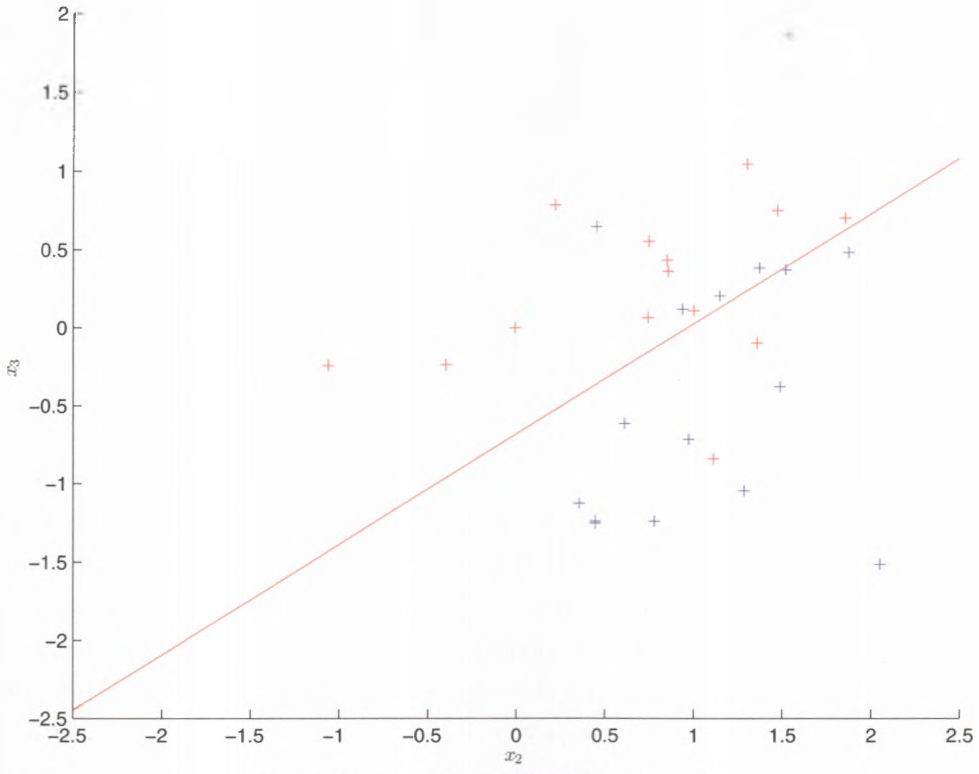


Figura 9.7: Classificação da terceira amostra.

log na função de log-verossimilhança é sempre entre 0 e 1, pois $0 \leq F \leq 1$. Portanto, a função de log-verossimilhança para essa classe de modelos é sempre negativa ou zero. Então, uma vez que queremos maximizar essa função, o valor ótimo dela é $l(\beta/X) = 0$. Portanto, para isso ocorrer, deveremos ter para todo $y_i = 1$, $F(x_i'\beta) = 1$ e para todo $y_i = 0$, $F(x_i'\beta) = 0$. Finalmente, se os dados satisfazem a Eq. (9.24), para $d \to \infty$, se $\hat{\beta} = d\beta$, então esse modelo sempre classificará perfeitamente os y_is. Na prática, devido a dificuldades que os algoritmos numéricos enfrentam quando $\hat{\beta} \to \infty$, então sempre $l(\hat{\beta}/X) < 0$.

Exemplo 9.14 (Classificação perfeita) Neste exemplo, usaremos novamente Monte Carlo e geraremos os dados de forma similar ao Exemplo 9.5. A diferença aqui será que ao invés do passo 4, usaremos o passo 4':

4') Gerar os n valores de y_i fazendo o seguinte teste: Se $x_i'\beta > 0$ então $y_i = 1$. Em caso contrário, $y_i = 0$, $i = 1, \cdots, n$.

Usando o passo 4' forçamos que os dados gerados por esse modelo satisfazem a Eq. (9.24). Aqui também geramos os regressores usando as mesmas distribuições apresentadas no Exemplo 9.5 e os mesmos valores para o vetor β. Estimando esse modelo, encontramos $\hat{\beta}_1 = 3.6562$, $\hat{\beta}_2 = -3.9880$ e $\hat{\beta}_3 = 3.8476$ e $l(\hat{\beta}/X) = -1.3504$. Agora, apresentamos uma figura similar à Figura 9.8 usando aqui $p_c = 0.5$. Note que nesse problema de classificação diferentemente do problema apresentado Exemplo 9.13, pudemos classificar a amostra perfeitamente. Adicionalmente, perceba também que se variarmos um pouco a posição e a

inclinação da reta que separa os pontos dessa figura, mesmo assim ainda conseguiremos classificar os pontos perfeitamente, mostrando que vários valores para os elementos do vetor $\hat{\beta}$ são possíveis. Finalmente, note que para conseguir classificação perfeita, não foi necessário chegar em $\hat{\beta} \to \infty$ e, por isso, também não conseguimos alcançar $l(\hat{\beta}/X) = 0$.

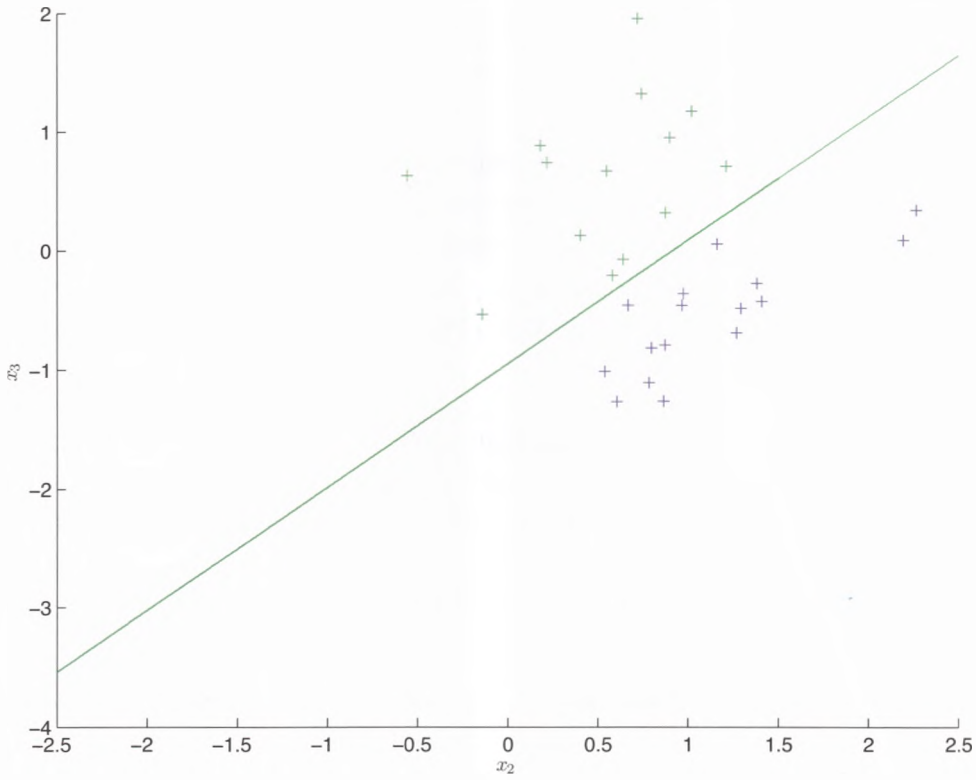


Figura 9.8: Classificação perfeita de uma amostra que satisfaz a Eq. (9.24).

9.4 Leituras adicionais

Neste capítulo, discutimos uma classe de modelos bastante utilizada na prática: a regressão para dados binários, utilizando funções logit e probit. Entre as vantagens dessa modelagem, incluem-se a facilidade de interpretação dos parâmetros do modelo, além da facilidade de estimação dos coeficientes, a possibilidade de utilização de testes de hipóteses e intervalos de confiança, e as análises de gráficas e estatísticas dos resíduos, para identificação de problemas de especificação. Esses assuntos foram cobertos nas seções anteriores.

No entanto, existe uma vasta literatura sobre classificação e modelos de regressão com dados de resposta binária, que podem estar ou não ligados aos modelos probit e logit. Muitos desses modelos têm por objetivo acrescentar flexibilidade à forma funcional que associa as variáveis independentes à probabilidade de sucesso $y_i = 1$. Mais especificamente, os modelos probit e logit assumem uma forma funcional do tipo

$$P(y_i = 1/x_i) = F(x_i'\beta),$$

onde $F(\cdot)$ é uma função distribuição acumulada. O argumento da função $F(\cdot)$ é uma função linear das covariáveis x_i. Alternativamente, poderíamos assumir formas paramétricas não-lineares, nas quais as covariáveis x_i aparececem de forma não-linear no argumento da função distribuição acumulada $F(\cdot)$. Por exemplo, pode-se assumir a forma funcional, em uma regressão logit, do tipo

$$P(y_i = 1/x_i) = \frac{e^{\beta_1 + \beta_2 x_{2,i}^{\beta_3} + \sin(x_{3,i}/\beta_4)}}{1 + e^{\beta_1 + \beta_2 x_{2,i}^{\beta_3} + \sin(x_{3,i}/\beta_4)}}. \tag{9.25}$$

Note que fizemos explicitamente $x_{1,i} = 1$. Os parâmetros β_1, β_2, β_3 e β_4 no modelo apresentado na Eq. (9.25) podem ser estimados via máxima verossimilhança, similarmente ao caso linear. A matriz de variância-covariância também pode ser estimada a partir da inversa da matriz de informação de Fisher observada. A partir da matriz de variância-covariância, pode-se proceder com testes de hipóteses e intervalos de confiança. Para maiores detalhes, vide Seber e Wild (2003), Gallant (1987), Ritz e Streibig (2008), Souza (1998).

Quando não se dispõem de bases de dados com muitas observações, supor um model restritivo da forma linear, do tipo $\beta' x_i$, pode ser o mais apropriado, uma vez que a informação disponível na amostra seja suficiente para estimação dos parâmetros β, mas não seja disponível para permitir inferência sobre formas funcionais mais gerais, do tipo

$$P(y_i = 1/x_i) = F(h(x_i; \beta)),$$

onde $h(x_i; \beta)$ é uma função de forma desconhecida, e precisa ser estimada a partir dos dados. O vetor β nesse caso, pode ser um vetor de dimensão finita ou infinita. Diversas técnicas existem na literatura para estimar a função $h(\cdot;\cdot)$ de forma flexível. Uma dessas técnicas, que se tornou muito comum em aplicações, são os **modelos aditivos generalizados.**[10] No caso de modelos aditivos generalizados, para regressões com resposta binária, a regressão tem forma geral

$$P(y_i = 1/x_i) = F(h_1(x_{1,i}; \theta_1) + h_2(x_{2,i}; \theta_2) + \cdots + h_K(x_{K,i}; \theta_K)), \tag{9.26}$$

onde $h_1(x_{1,i}; \theta_1)$, $h_2(x_{2,i}; \theta_2)$, ..., $h_K(x_{K,i}; \theta_K)$, são funções flexíveis das covariáveis $x_{1,i}$, $x_{2,i}$, ..., $x_{K,i}$, individualmente. Os vetores θ_1, ..., θ_K, são vetores de parâmetros para cada função $h_j(\cdot;\cdot)$, sendo que θ_j, $j = 1, \ldots, K$, pode ter dimensão finita ou infinita. O termo modelos aditivos generalizados vem do fato da aditividade das funções $h_1(x_{1,i}; \theta_1)$, $h_2(x_{2,i}; \theta_2)$, ..., $h_K(x_{K,i}; \theta_K)$, não havendo interação entre as covariáveis $x_{1,i}$, $x_{2,i}$, ..., $x_{K,i}$. Em formas menos restritivas em relação à aditividade, poderíamos assumir por exemplo,

$$P(y_i = 1/x_i) = F(h_1(x_{1,i}, x_{2,i}; \theta_1) + h_3(x_{3,i}; \theta_3) + \cdots + h_K(x_{K,i}; \theta_K)).$$

[10] Em inglês, *generalized additive models* (GAM's).

Note a possibilidade de interação entre as covariáveis $x_{1,i}$ e $x_{2,i}$. Finalmente, modelos não-paramétricos aditivos, da forma apresentada na Eq. (9.26), podem ser combinados com formas paramétricas lineares, gerando formas mistas, do tipo

$$P(y_i = 1/x_i, z_i) = F(\beta' x_i + h(z_i; \theta)). \tag{9.27}$$

O termo $\beta' x_i$ é parámetrico (e linear) no vetor de covariáveis x_i. O termo $h(z_i; \theta)$ é não-paramétrico no vetor de covariáveis z_i. Portanto, o vetor total de covariáveis é $[x_i' \ \ z_i']'$. No caso de um modelo logit, com a estrutura apresentada na Eq. (9.27), a expressão é dada por

$$P(y_i = 1/x_i, z_i) = \frac{e^{\beta' x_i + h(z_i;\theta)}}{1 + e^{\beta' x_i + h(z_i;\theta)}}.$$

Em todo caso, havendo termos paramétricos ou não-paramétricos, lineares ou não-lineares, para o papel das covariáveis na probabilidade de $y_i = 1$, os modelos discutidos acima são todos extensões, de alguma forma, dos modelos de reposta binária estudados neste capítulo. No entanto, a literatura de classificação contém inúmeros outras técnicas que não estão relacionadas aos modelos probit e logit ou similares. Incluem-se, nessa lista, os modelos de **classificação Bayesiana**, **análise discriminante linear**, **redes neurais** e **máquinas de vetor suporte**.[11] Detalhes desses métodos assim como a sua implementação podem ser encontrados em Hastie, Tibshirani e Friedman (2001), Alpaydin (2009) e Segaran (2008).

9.5 Exercícios

Exercício 9.1 Seja $A = [a_{ij}]$ uma matriz de ordem $m \times n$, onde $n < m$. Defina $A_i' = [a_{i1} \cdots a_{ik}]$. Mostre que se posto de A for menor que n, a matriz $\sum_{i=1}^{m} c_i A_j A_j'$ não possui inversa.

Dica: Use a definição de dependência linear.

Exercício 9.2 Mostre que a combinação linear de funções côncavas definidas num subconjunto convexo $U \subset \Re^n$ também é uma função côncava.

Dica: Use a definição de função côncava.

Exercício 9.3 Suponha que a Hipótese 9.2 é válida e mostre que a função de log-verossimilhança é estritamente côncava em β para os casos dos modelos probit e logit.

Dica: A ideia geral é mostrar que o termo $\log f(y_i/x_i) = y_i \log F(x_i'\beta) + (1 - y_i)\log(1 - F(x_i'\beta))$ é estritamente côncavo quando a Hipótese 9.2 é válida e usar o Exercício 9.2 para mostrar que a soma também é estritamente côncava. Para mostrar que o termo acima é estritamente côncavo, precisamos

[11] Em inglês, *vector support machine*.

mostrar que a matriz de segundas derivadas do termo acima é negativa definida. Para entender detalhes dessas ideias, vide Simon e Blume (2004). Para simplificar, considere separadamente os dois casos possíveis para os termos acima $y_i = 1$ e $y_i = 0$ e também use $F(-x) = 1 - F(x)$ que é uma propriedade válida para os casos das distribuições normal e logística.

Exercício 9.4 Considere um modelo de resposta binária da forma $P(y_i = 1/x_i) = F(\beta_0 + \beta_1 x_i)$ onde x_i é uma variável binária que assume valores no conjunto $\{0, 1\}$. Suponha que usando uma amostra de tamanho n e com média amostral de y igual a $\overline{y}$, você estimou esse modelo e encontrou as estimativas de β_0 e β_1 respectivamente iguais a $\hat{\beta}_0$ e $\hat{\beta}_1$.

1) Calcule o número de observações na amostra para que $x_i = 0$ em função da distribuição acumulada F, do tamanho da amostra n, da média amostral de y e das estimativas $\hat{\beta}_0$ e $\hat{\beta}_1$.

2) Assuma uma forma funcional específica para F e valores específicos para n, $\overline{y}$, $\hat{\beta}_0$ e $\hat{\beta}_1$ e calcule as variâncias de $\hat{\beta}_0$ e $\hat{\beta}_1$.

Dica: (1) Use a Eq. (9.14). (2) Use o resultado da parte (1) e use a Eq. (9.21).

Referências

ABADIR, K. M.; MAGNUS, J. R. *Matrix algebra.* Cambridge: Cambridge University Press, 2005.

ALBERT, R.; BARABASI, A. L. Statistical mechanics of complex networks. *Reviews of Modern Physics*, v. 74, p. 47–97, 2002.

ALPAYDIN, E. *Introduction to machine learning.* Cambridge: MIT Press, 2009.

AMEMIYA, T. *Advanced econometrics.* Cambridge: Harvard University Press, 1985.

ANDERSON, T. W. *An Introduction to Multivariate Statistical Analysis.* New York: Wiley-Interscience, 2003.

ANDERSON, T. W.; DARLING, D. A. Asymptotic theory of certain goodness-of-fit criteria based on stochastic processes. *Annals of Mathematical Statistics*, v. 23, p. 193–212, 1952.

ANSELIN, L. *Spatial econometrics: methods and models.* Dordrecht: Kluwer Academic Publishers, 1988.

ARTHUR, W. B. Inductive reasoning and bounded rationality. *American Economic Review*, v. 84, p. 406, 1994.

BALTAGI, B. H. *Econometric analysis of panel data.* New York: John Wiley and Sons, 2008.

BARTH, N. L. *Inadimplência: construção de modelos de previsão.* São Paulo: Nobel, 2004.

BARTLE, R. G. *The elements of integration.* New York: John Wiley and Sons, 1966.

BECKMAN, O. R.; NETO, P. L. O. C. *Análise Estatística da decisão.* São Paulo: Edgard Blucher, 1980.

BICKEL, P.; DOKSUM, K. *Mathematical statistics: basic ideas and selected topics.* Upper Saddle River: Prentice Hall, 2000.

BIERENS, H. J. *Introduction to the mathematical and statistical foundations of econometrics.* Cambridge: Cambridge University Press, 2004.

BILLINGSLEY, P. *Probability and measure.* New York: Wiley Inter-Science, 1995.

BIS. *Comitê da Basiléia sobre supervisão bancária. Convergência internacional de mensuração de capital e padrões de capital.* São Paulo: FEBRABAN., 2004.

BLACK, F. Capital market equilibrium with restricted borrowing. *Journal of Business*, v. 45, p. 444–455, 1972.

BLACK, F.; SCHOLES, M. The pricing of options and corporate liabilities. *Journal of Political Economy*, v. 81, p. 637–654, 1973.

BOCCALETTI, S. et al. Complex networks: structure and dynamics. *Physics Reports*, v. 424, p. 175–308, 2006.

BOLFARINE, H.; BUSSAB, W. O. *Elementos de amostragem.* São Paulo: Edgard Blucher, 2005.

BOSCHETTI, F. Improving resource exploitation via collective intelligence by assessing agents' impact on the community outcome. *Ecological Economics*, v. 63, p. 533–562, 2007.

BOSS, M. et al. Network topology of the interbank market. *Quantitative Finance*, v. 4, p. 677–684, 2004.

BOX, G. E. P.; DRAPER, N. *Empirical Model-Building and Response Surfaces.* New York: John Wiley and Sons, 1987.

BUENO, R. L. S. *Econometria de séries temporais.* São Paulo: Cengage, 2008.

BURNHAM, K. P.; ANDERSON, D. R. *Model selection and inference. A Practical information-theoretic approach.* New York: Springer, 1998.

BUSSAB, W. O.; MORETTIN, P. A. *Estatística básica.* São Paulo: Saraiva, 2002.

CAJUEIRO, D. O. Agent preferences and the topology of networks. *Physical Review E*, v. 72, p. 047104, 2005.

CAJUEIRO, D. O.; CAMARGO, R. S. D. Minority game with local interactions due to the presence of herding behavior. *Physics Letters A*, v. 355, p. 280–284, 2006.

CAJUEIRO, D. O.; TABAK, B. M. The role of banks in the brazilian interbank market: Does bank type matter? *Physica A*, v. 387, p. 6825–6836, 2008.

CAJUEIRO, D. O.; TABAK, B. M.; ANDRADE, R. F. S. Fluctuations in interbank network dynamics. *Physical Review E*, v. 79, p. 037101, 2009.

CAMPBELL, J. Y.; LO, A. W.; MACKINLAY, A. C. *The econometrics of financial markets.* Princeton: Princeton University Press, 1996.

CARVALHO, R.; IORI, G. Socio-economic networks with long-range interactions. *Physical Review E*, v. 78, p. 016110, 2008.

CASELLA, G.; BERGER, R. *Statistical inference.* Belmont: Duxbury Press, 2001.

CHALLET, D.; MARSILI, M.; OTTINO, G. Shedding light on El Farol. *Physica A*, v. 332, p. 469–482, 2004.

CHALLET, D.; MARSILI, M.; ZHANG, Y. C. Modeling market mechanism with minority game. *Physica A*, v. 276, p. 284–315, 2000.

CHALLET, D.; MARSILI, M.; ZHANG, Y. C. Minority games and stylized facts. *Physica A*, v. 299, p. 228–233, 2001.

CHALLET, D.; MARSILI, M.; ZHANG, Y. C. *Minority games.* New York: Oxford University Press, 2005.

CHALLET, D.; ZHANG, Y. C. Emergence of cooperation and organization in an evolutionary game. *Physica A*, v. 246, p. 407–418, 1997.

CHAMBERS, E. A.; COX, D. R. Discrimination between alternative binary response models. *Biometrika*, v. 54, p. 573–578, 1967.

CHERNOFF, H.; MOSES, L. E. *Elementary decision theory.* Mineola: Dover, 1986.

CLAUSET, A.; SHALIZI, C. R.; NEWMAN, M. E. J. Power-law distributions in empirical data. *SIAM Review*, v. 51, p. 661–703, 2009.

CLEMEN, R. T. *Making hard decisions: an introduction to decision analysis.* Belmont: Duxbury Press, 1996.

CLEVELAND, W. S. Robust locally weighted regression and smoothing scaterplots. *Journal of the American Statistical Association*, v. 74, p. 829–836, 1979.

COCHRAN, W. G. *Sampling techniques.* New York: John Wiley & Sons, 1977.

COCHRANE, J. H. *Asset pricing.* Princeton: Princeton University Press, 2005.

COOLEN, A. C. C. *The mathematical theory of minority games.* New York: Oxford University Press, 2005.

COPELAND, T. E.; ANTIKAROV, V. *Opções reais.* Rio de Janeiro: Campus, 2002.

COPELAND, T. E.; WESTON, J. F. *Financial theory and corporate policy.* Reading: Addison-Wesley Publishing Company, 1992.

COSTA, L. D. et al. Characterization of complex networks: a survey of measurements. *Advances in Physics*, v. 56, p. 167–242, 2007.

COSTA, O. L. V.; ASSUNçAO, H. G. V. *Análise de risco e retorno em investimentos financeiros.* São Paulo: Manole, 2005.

COX, J.; ROSS, S.; RUBINSTEIN, M. Option pricing: a simplified approach. *Journal of Financial Economics*, v. 7, p. 229–264, 1979.

CRUZ, M. *Modelagem, avaliação e proteção para risco Operacional.* Rio de Janeiro: Editora Teatral, 2005.

CUTHBERTSON, K.; NITZSCHE, D. *Quantitative financial economics: stocks, bonds and foreign exchange.* New York: John Wiley and Sons, 2005.

DAVIDSON, R.; MACKINNON, J. G. *Econometric theory and methods.* New York: Oxford University Press, 2004.

DIXIT, A. K.; PINDYCK, R. S. *Investment under uncertainty.* Princeton: Princeton University Press, 1994.

DUFFIE, D.; SINGLETON, K. J. *Credit Risk: Pricing, Measurement, and Management.* Princeton: Princeton University Press, 2003.

DURRET, R. *Probability: theory and examples.* Belmont: Duxbury Press, 1996.

ENDERS, W. *Applied time series econometrics.* New York: John Wiley and Sons, 2003.

ERDÓS, P.; RÉNYI, A. On the evolution of random graphs. *Bulletin of the International Statistical Institute*, v. 38, p. 343–347, 1960.

EVES, H. *Elementary matrix theory.* Mineola: Dover, 2008.

FERNANDEZ, P. J. *Introdução a teoria das probabilidades.* Brasília: Universidade de Brasília, 1973.

FRANKLIN, J. N. *Matrix algebra.* Mineola: Dover, 1968.

FREDHEIM, M. *Copula methods in finance.* Saarbrücken: VDM Verlag, 2008.

GALLANT, A. R. *Nonlinear Statistical Models.* New York: John Wiley and Sons, 1987.

GELMAN, A. *Data analysis using regression and multilevel/hierarchical models.* Cambridge: Cambridge University Press, 2007.

GELMAN, A. et al. *Bayesian data analysis.* Boca Raton: Chapman & Hall/CRC, 1995.

GHYSELS, E.; OSBORN, D. R. *The Econometric Analysis of Seasonal Time Series.* Cambridge: Cambridge University Press, 2001.

GIBBONS, J. D. *Nonparametric statistics: an introduction.* Newbury Park: Sage Publications, 1992.

GLASSERMAN, P. *Monte carlo methods in financial engineering.* New York: Springer, 2004.

GOURIEROUX, C.; JASIAK, J. *Financial econometrics.* Princeton: Princeton University Press, 2001.

GREENBERG, E. *Introduction to Bayesian Econometrics.* Cambridge: Cambridge University Press, 2013.

GRIMMETT, G.; STIRZAKER, D. *Probability and random processes.* New York: Oxford University Press, 2001.

GUJARATI, D. N. *Econometria Básica.* São Paulo: Makron Books, 2000.

HALTER, A. N.; DEAN, G. W. *Decisions under uncertainty.* Cincinnati: South-Western Publishing, 1971.

HAMILTON, J. D. *Time series analysis.* Princeton: Princenton University Press, 1994.

HANSEN, L. P. Large sample properties of generalized method of moments estimators. *Econometrica*, v. 50, p. 1029–1054, 1982a.

HANSEN, L. P.; SINGLETON, K. J. Generalized instrumental variables estimation of nonlinear rational expectations models. *Econometrica*, v. 50, p. 1269–1286, 1982b.

HARRELL, F. E. *Regression modeling strategies.* New York: Springer, 2001.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, F. *The elements of statistical learning. Data mining, inference, and prediction.* New York: Springer, 2001.

HAYASHI, F. *Econometrics.* Princeton: Princeton University Press, 2000.

HIDALGO, C. A. et al. The product space conditions the development of nations. *Science*, v. 27, p. 482–487, 2007.

HOAGLIN, D. C.; MOSTELLER, F.; TUKEY, J. W. (Ed.). *Understanding robust and exploratory data analysis.* New York: Wiley-Interscience, 1983.

HOAGLIN, D. C.; MOSTELLER, F.; TUKEY, J. W. (Ed.). *Managing operational risk.* New York: John Wiley and Sons, 1985.

HOFFMAN, R. *Estatística para economistas.* São Paulo: Thomson Pioneira, 2006.

HSIAO, C. (Ed.). *Analysis of panel data.* Cambridge: Cambridge University Press, 2003.

HULL, J. C. *Opções, futuros e outros derivativos.* São Paulo: Bolsa de Mercadorias e Futuros, 1997.

IORI, G. et al. A network analysis of the italian overnight money market. *Journal of Economic Dynamics and Control*, v. 32, p. 259–279, 2008.

JACKSON, M. O.; ROGERS, B. W. The economics of small worlds. *Journal of the European Economic Association*, v. 3, p. 617–627, 2005.

JEFFERIES, P.; HART, M. L.; HUI, P. M. From market games to real-world markets. *European Physical Journal B*, v. 20, p. 493–501, 2001.

JOE, H. *Multivariate models and dependence concepts.* New York: Chapman & Hall/CRC, 1997.

JOHNSON, N. F.; JEFFERIES, P.; HUI, P. M. *Financial market complexity.* New York: Oxford University Press, 2003.

KOCH, K. R. *Introduction to Bayesian Statistics.* New York: Springer, 2007.

KOENKER, R. *Quantile regression.* Cambridge: Cambridge University Press, 2005.

KOLLER, T.; MURRIN, J.; COPELAND, T. *Avaliação de empresas - valuation: Calculando e gerenciando o valor de empresas.* São Paulo: Makron, 2001.

LEINHARDT, G.; LEINHARDT, S. Exploratory data analysis: new tools for the analysis of empirical data. *Review of Research in Education*, v. 8, p. 85–157, 1980.

LEROY, S. F.; WERNER, J. *Principles of financial economics.* Cambridge: Cambridge University Press, 2001.

LESAGE, J.; PACE, R. K. *Introduction to spatial econometrics.* Boca Raton: Chapman & Hall/CRC, 2009.

LIMA, E. L. *Álgebra linear.* Rio de Janeiro: IMPA, 1995.

LO, A. W.; MACKINLAY, A. C. *A non-Random walk down wall street.* Princeton: Princeton University Press, 2001.

LOHR, S. L. *Sampling: design and analysis.* Belmont: Duxbury Press, 2002.

LUENBERGER, D. G. *Optimization by vector space methods.* New York: John Wiley and Sons, 1969.

LUSTOSA, B. C. *Jogos da minoria com informação incompleta. Você sabe com foi a noite no El Farol?* Tese (Dissertação de mestrado) — Universidade Católica de Brasília, 2008.

LUSTOSA, B. C.; CAJUEIRO, D. O. Constrained information minority game: How was the night at el farol? *Physica A*, p. 1230–1238, 2010.

MANTEGNA, R. N.; STANLEY, H. E. *An introduction to econophysics: Correlations and Complexity in Finance.* Cambridge: Cambridge University Press, 1999.

MARKOWITZ, H. M. Portfolio selection. *Journal of Finance*, VII, p. 77–91, 1952.

MAS-COLELL, A.; WHINSTON, M. D.; GREEN, J. R. *Microeconomic theory.* New York: Oxford University Press, 1995.

MCFADDEN, D. L. Frontiers in econometrics. In: _____. San Diego: Academic Press, 1974. cap. Conditional logit analysis of qualitatice choice behavior, p. 105–142.

MCLACHLAN, G.; PEEL, D. *Finite mixture models.* New York: Wiley-Interscience, 2000.

MCNEIL, A. J.; FREY, R.; EMBRECHTS, P. *Quantitative Risk Management: Concepts, Techniques, and Tools.* Princeton: Princeton University Press, 2005.

MEINTANIS, S.; SWANEPOEL, J. Bootstrap goodness-of-fit tests with estimated parameters based on empirical transformations. *Statistics and Probability Letters*, v. 77, p. 1004–1013, 2007.

MELE, A. *Lecture notes in financial economics.* London: London School of Economics and Political Science, 2007.

MELLO, B. A.; CAJUEIRO, D. O. Minority games, diversity, cooperativity and the concept of inteligence. *Physica A*, v. 387, p. 557–566, 2008.

MELLO, B. A. et al. Network evolution based on minority game with herding behavior. *European Physical Journal B*, v. 76, p. 147–156, 2010.

MERTON, R. An analytic derivation of the efficient set. *Journal of Financial and Quantitative Analysis*, v. 10, p. 1851–1872, 1972.

MORETTIN, P. A. *Econometria financeira.* São Paulo: Edgard Blucher, 2008.

MORETTIN, P. A.; TOLOI, C. M. C. *Análise de séries temporais.* São Paulo: Edgard Blucher, 2006.

NELSEN, R. *An introduction to copulas.* New York: Springer, 1998.

NEOPOLITAN, R. E. *Learning Bayesian networks.* Upper Saddle River: Prentice Hall, 2004.

NEWMAN, M. E. J. Assortative mixing in networks. *Physical Review Letters*, v. 89, p. 208701, 2002.

NEWMAN, M. E. J. *Networks: an introduction.* New York: Oxford University Press, 2010.

PRATT, J. W. Concavity of the log likelihood function. *Journal of the American Statistical Association*, v. 76, n. 373, p. 103–106, 1981.

RITZ, C.; STREIBIG, J. C. *Nonlinear regression with R.* New York: Springer, 2008.

ROMANO, J. P.; SIEGEL, A. F. *Counterexamples in Probability And Statistics.* Boca Raton: Chapman and Hall, 1986.

ROSENTHAL, J. S. *A first look at rigorous probability theory.* Danvers: World Scientific, 2006.

ROUSSAS, G. G. *A course in mathematical statistics.* San Diego: Academic Press, 1997.

RUUD, P. A. *An introduction to classical econometric theory.* New York: Oxford University Press, 2000.

SAVIT, R.; MANUCA, R.; RIOLO, R. Adaptive competition, market efficiency and phase transitions. *Physical Review Letters*, v. 82, p. 2203–2206, 1999.

SEBER, G. A. F.; WILD, C. J. *Nonlinear Regression.* New York: Wiley-Interscience, 2003.

SECURATO, J. R. *Decisões financeiras em condições de risco.* São Paulo: Atlas, 1996.

SEGARAN, T. *Programando a inteligência coletiva.* Rio de Janeiro: Alta Books, 2008.

SERRANO, M. A.; BOGUNÁ, M. Topology of the world trade web. *Physical Review E*, v. 68, p. 015101, 2003.

SERRANO, M. A.; BOGUNÁ, M.; VESPIGNANI, A. Patterns of dominant flows in the world trade web. *Journal of Economic Interaction and Coordination*, v. 2, p. 111–124, 2007.

SEVERINI, T. A. *Likelihood methods in statistics.* New York: Oxford University Press, 2001.

SHAO, J. *Mathematical statistics.* New York: Springer, 2003.

SHARPE, W. F. A simplified model for portfolio analysis. *Management Science*, v. 9, p. 277–293, 1963.

SHARPE, W. F. Capital asset prices – a theory of market equilibrium under conditions of risk. *Journal of Finance*, XIX, p. 425–442, 1964.

SIMON, C. P.; BLUME, L. *Matemática para economistas.* Porto Alegre: Bookman, 2004.

SORAMAKI, K. et al. The topology of interbank payment flows. *Physica A*, v. 379, p. 317–333, 2007.

SOUZA, G. S. *Introdução aos modelos de regressão linear e não-linear.* Brasília: Embrapa, 1998.

STEELE, M.; CHASELING, J.; HURST, C. Simulated power of the discrete cramer-von mises goodness-of-fit tests. In: *Annals of MODSIM 2005 International Congress on Modelling and Simulation. Advances and Applications for Management and Decision Making.* Melbourne, Australia.: [s.n.], 2005.

STEPHENS, M. A. Edf statistics for goodness of fit and some comparisons. *Journal of the American Statistical Association*, v. 69, p. 730–737, 1974.

STEPHENS, M. A. Asymptotic results for goodness-of-fit statistics with unknown parameters. *Annals of Statistics*, v. 4, p. 357–369, 1976.

STEPHENS, M. A. Goodness of fit for the extreme value distribution. *Biometrika*, v. 64, p. 583–588, 1977.

STEPHENS, M. A. Goodness of fit with special reference to test for exponentiality. *Technical Report of the Department of Statistics, Stanford University*, v. 262, 1977b.

STEPHENS, M. A. Tests of fit for the logistic distribution based on the empirical distribution function. *Biometrika*, v. 66, p. 591–595, 1979.

STEVENSON, W. A. *Estatística Aplicada a Administração*. São Paulo: Harbra, 1997.

STUTE, W.; MATEIGA, W. G.; QUINDIMIL, M. P. Bootstrap based goodness-of-fit tests. *Metrika*, v. 40, p. 243–256, 2007.

TANNER, M. *Tools for statistical inference. Methods for the exploration of posterior distributions and likelihood functions*. New York: Springer, 1996.

TIKU, M. L. Laguerre series forms of non-central chi-square and f distributions. *Biometrika*, v. 52, p. 415–427, 1965.

TUKEY, J. W. *Exploratory data analysis*. Reading: Addison-Wesley, 1977.

VELLEMAN, P. F.; HOAGLIN, D. C. *Applications, basics and computing of exploratory data analysis*. Belmont: Duxbury Press, 1981.

WAKELING, J.; BAK, P. Intelligent systems in the context of surrounding enviroment. *Phys. Rev. E*, v. 64, p. 051920–051928, 2001.

WAN, Y. S.; CHEN, Z.; LIU, Z. R. Modeling the two-power-law degree distribution of banking networks. *Dynamics of Continuous Discrete and Impulsive Systems – Series B – Applications and Algorithms*, v. 13, p. 441–449, 2006.

WEISS, L. *Statistical decision theory*. New York: McGraw-Hill, 1961.

WHITE, H. *Estimation, inference and specification analysis*. Cambridge: Cambridge University Press, 1996.

WHITE, H. *Asymptotic theory for econometricians*. San Diego: Academic Press, 2000.

WILMOTT, P.; HOWISON, S.; DEWYNNE, J. *The mathematics of financial derivatives*. Cambridge: Cambridge University Press, 1995.

WOOLDRIDGE, J. M. *Econometric Analysis of Cross Section and Panel Data*. Cambridge: MIT Press, 2001.

WOOLDRIDGE, J. M. *Introductory econometrics*. Mason: Thomson, 2003.

ZACHARY, W. W. An information flow model for conflict and fission in small groups. *Journal of Antropological Research*, v. 33, p. 452–473, 1977.

Índice Remissivo

Nas últimas décadas, tem crescido muito a utilização de métodos estatísticos nas diferentes áreas em ciências sociais. No caso de economia e finanças, praticamente todos os estudos empíricos baseiam-se em métodos estatísticos, desde os mais simples aos mais sofisticados. Em muitas situações, as aplicações são agrupadas em algumas categorias, que têm se tornado cada vez mais populares. Como exemplo, temos toda uma grande área de pesquisa referente à avaliação quantitativa de políticas públicas. Outra área de pesquisa que tem tido cada vez mais relevância congrega os estudos de avaliação de impacto regulatório (AIR). Finalmente, não se pode deixar de mencionar toda uma tradição de aplicações de ferramental estatístico em finanças empíricas.

Este livro traz uma introdução a vários dos principais conceitos necessários para o entendimento e a utilização de técnicas estatísticas em aplicações em economia e finanças. Discutem-se, por exemplo, conceitos de variáveis aleatórias e distribuições de probabilidade. Uma grande ênfase é dada ao problema de inferência estatística, em que se busca estimar os parâmetros desconhecidos de modelos probabilísticos. O problema de regressão linear e de regressão com variáveis resposta binárias também é abordado. A discussão é exposta de forma intuitiva, com a utilização de simulações de Monte Carlo para elucidar o conceito de distribuição dos estimadores – que é a base para os métodos inferenciais. Embora vise a uma discussão intuitiva, o livro é escrito em linguagem matemática, possibilitando a transição suave entre a exposição dos principais conceitos e a formalização analítica encontrada em livros mais avançados.

ISBN 978-85-230-1148-2
9 788523 011482