

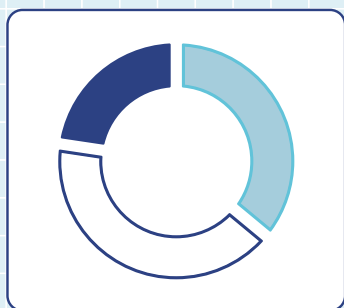
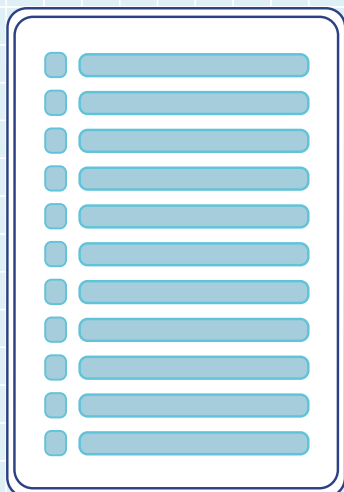
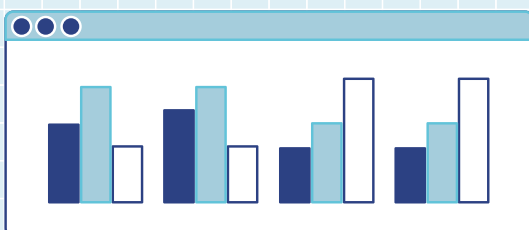
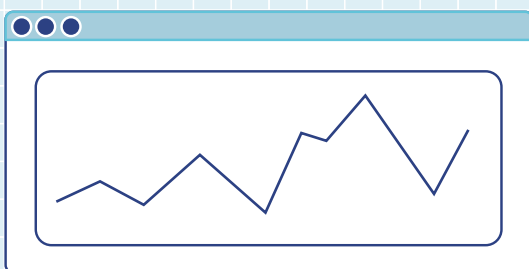
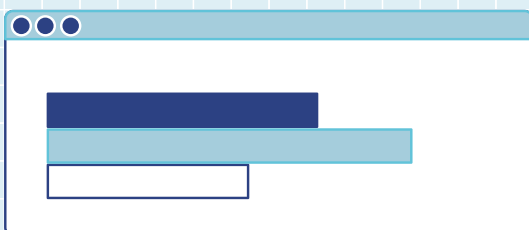
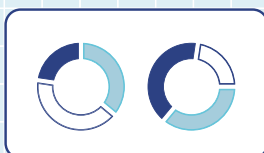
ANÁLISE DE DADOS



procedimentos exploratórios

Francisco José da Costa

Patrícia de Oliveira Campos



EU Editora
UFPB



UNIVERSIDADE FEDERAL DA PARAÍBA

Terezinha Domiciano Dantas Martins
Reitora

Mônica Nóbrega
Vice-Reitora



Editora UFPB

Geysa Flávia Câmara de Lima Nascimento
Diretora Geral da Editora UFPB

Aline dos Santos Deiró
Vice-Diretora da Editora UFPB

Rildo Coelho
Coordenador de Editoração

ANÁLISE DE DADOS_

procedimentos exploratórios

1ª Edição – 2026

Esta obra integra a publicação de livros digitais (e-books) do Edital 01/2025 da Editora UFPB, em regime de fluxo contínuo. A chamada pública visa fomentar a divulgação da produção científica, literária, artística e didático-pedagógica da comunidade acadêmica da Universidade Federal da Paraíba, contemplando obras de autoria individual, coautoria ou coletâneas.

Direitos autorais 2026 – Editora da UFPB



Esta obra é de acesso aberto. É permitida a reprodução parcial ou total desta obra, desde que citada a fonte e autoria e respeitada a Licença Creative Commons indicada.

O CONTEÚDO DESTA PUBLICAÇÃO, SEU TEOR, SUA REVISÃO E SUA NORMALIZAÇÃO SÃO DE INTEIRA RESPONSABILIDADE DOS AUTORES.

OS DIREITOS DE PROPRIEDADE DESTA EDIÇÃO SÃO RESERVADOS À:



Cidade Universitária, Campus I – Prédio da Editora Universitária, s/n
João Pessoa – PB CEP 58.051-970
Site: www.editora.ufpb.br
Instagram: @editoraufpb
E-mail: atendimento@editora.ufpb.br
Fone: (83) 3216.7147



CONSELHO EDITORIAL – Editora UFPB

Portaria Nº 202, de 08 de agosto de 2025 – Reitoria

- Geysa Flávia Câmara de Lima Nascimento** – Presidente | Editora UFPB
- Alexandre Coelho Serquiz** | Ciências da Saúde
- Daniel Araújo de Macedo** | Engenharias
- Eduardo Rodrigues Viana de Lima** | Ciências Exatas e da Natureza
- Franklin Kaic Dutra-Pereira** | Ciências Biológicas
- Gustavo Barbosa de Mesquita Batista** | Ciências Humanas
- José Ferrari Neto** | Linguística, Letras e Artes
- Maurício Rombaldi** | Ciências Humanas
- Milton César Costa Campos** | Ciências Agrárias
- Tiago Bernardon de Oliveira** | Ciências Humanas
- Márcia Félix da Silva** | Interdisciplinar
- Viviane da Costa Freitag** | Ciências Sociais Aplicadas

Catálogo na Publicação
Seção de Catalogação e Classificação

C838a	Costa, Francisco José. Análise de dados : procedimentos exploratórios [recurso eletrônico] / Francisco José da Costa, Patrícia de Oliveira Campos. – Dados eletrônicos – João Pessoa : Editora UFPB, 2026. E-book. Modo de acesso: www.ufpb.br/editora/ ISBN: 978-65-5942-330-9 1. Análise estatística. 2. Análise exploratória de dados. 3. Estatística descritiva. 4. Análise de dados quantitativos. I. Campos, Patrícia de Oliveira. II. Título.
UFPB/BC	CDU 519.23

Como citar a publicação no todo (ABNT 6023:2025):
COSTA, Francisco José da.; CAMPOS, Patrícia de Oliveira. **Análise de dados: procedimentos exploratórios**. 1. ed. João Pessoa: Editora UFPB, 2026. Disponível em: _____. Acesso em: _____.
-----.

Francisco José da Costa
Patrícia de Oliveira Campos

ANÁLISE DE DADOS: **PROCEDIMENTOS EXPLORATÓRIOS**

Editora UFPB

João Pessoa

2026

SUMÁRIO

APRESENTAÇÃO	9
PARTE I. FUNDAMENTOS GERAIS	15
CAPÍTULO 1 CONTEXTO GERAL	17
CAPÍTULO 2 FUNDAMENTOS CONCEITUAIS E OPERACIONAIS	39
CAPÍTULO 3 ANÁLISE DE DADOS E OUTROS FUNDAMENTOS	62
PARTE II. ESTATÍSTICA DESCRITIVA UNIVARIADA	94
CAPÍTULO 4 ORGANIZAÇÃO E APRESENTAÇÃO DE DADOS POR TABELAS	95
CAPÍTULO 5 REPRESENTAÇÃO GRÁFICA DE DADOS	127
CAPÍTULO 6 MEDIDAS DESCRITIVAS DE POSIÇÃO	147
CAPÍTULO 7 MEDIDAS DESCRITIVAS DE DISPERSÃO	190
CAPÍTULO 8 MEDIDAS DESCRITIVAS DE FORMATO	205

PARTE III.	
ANÁLISE BIVARIADA	218
CAPÍTULO 9	
ANÁLISE BIVARIADA ENVOLVENDO VARIÁVEIS CATEGÓRICAS	220
CAPÍTULO 10	
ANÁLISE DE DUAS VARIÁVEIS QUANTITATIVAS_	235
CAPÍTULO 11	
APLICAÇÕES COM SUPORTE DE INTELIGÊNCIA ARTIFICIAL_	264
Apêndice A	
INSTALAÇÃO DO R E RSTUDIO	275
Apêndice B	
INSTALAÇÃO DO JASP	277
Apêndice C	
INSTALAÇÃO DO VSCODE (PYTHON)	278
Apêndice D	
DADOS DA PESQUISA COM ESTUDANTES DE ADMINISTRAÇÃO E TURISMO	279
REFERÊNCIAS	283
SOBRE OS AUTORES	284

APRESENTAÇÃO_

Este texto tem por finalidade servir como referencial teórico e metodológico para estudo e consulta de estudantes, professores e pesquisadores da grande área profissional, acadêmica e disciplinar das Ciências sociais e comportamentais, com extensão de uso para áreas do conhecimento, como Administração, Economia, Contabilidade, Turismo, Psicologia, Sociologia, dentre outras.

O foco para este volume é a construção do conteúdo da disciplina estatística de Análise Exploratória de Dados (AED). Mesmo sendo possível expandir o conceito de análise exploratória para uma dimensão mais ampla da análise estatística, concentramos a exposição nas técnicas de estatística descritiva. Por esta razão, não serão analisados aqui aspectos mais específicos e detalhados da análise quantitativa que envolve conteúdos de inferência estatística, como, por exemplo, os testes estatísticos, tão amplamente usados em alguns campos da pesquisa social. Também não serão abordados aqui os conteúdos que envolvam a operacionalização multivariada de dados, embora também haja técnicas multivariadas que são caracterizadas como exploratórias.

Para esta decisão de recortes foram levados em conta os condicionantes específicos do contexto de aplicação (Ciências sociais e comportamentais), na proposta de apresentar um texto útil em termos de utilização prática e ao mesmo tempo introdutório a outros textos e aplicações mais avançados. Portanto, este texto é antes de tudo uma introdução ao conteúdo da análise exploratória de dados, mas também é um texto que antecede outros conteúdos mais específicos de análise inferencial e de análise multivariada.

A meta é não ser apenas mais um manuscrito sobre estatística aplicada. Pelo contrário, nossa intenção é ao mesmo tempo cobrir o conteúdo convencional do assunto e de suas aplicações, mas ser também uma atualização do que se tem feito ao longo das últimas décadas, em uma reflexão aplicada ao contexto de interesse. Neste sentido, dois desafios foram colocados desde o primeiro momento, e são aqui explicitados: a incorporação do conteúdo mais atual da AED, e realinhamento do conteúdo para suas aplicações em Ciências sociais e comportamentais. Comentamos a seguir cada desafio.

- O conteúdo mais atual da AED

Primeiro, o conteúdo convencional da Análise Exploratória de Dados tem uma conformação convencional, que foi considerada e incorporada aqui, e que está presente nos diversos manuais escritos sobre o tema. Mas nas últimas duas décadas ocorreram diversos avanços. Desde o primeiro grande texto de Análise Exploratória de Dados, escrito por John Tukey em 1977, até o presente, a evolução dos recursos computacionais progrediu tremendamente, de modo que, atualmente, temos disponíveis diversas ferramentas para o tratamento gráfico e para o cálculo e a operacionalização de técnicas quantitativas, mesmo para grandes quantidades de dados.

Em geral, é possível afirmar que estas ferramentas promoveram uma popularização do uso de algumas técnicas, o que é algo positivo, sem dúvidas. De fato, executivos, estudantes e pesquisadores utilizam com frequência ferramentas de softwares como o MS Excel, por exemplo, com boa fluência, e já se habituaram a analisar conjuntos de dados por meio de ferramentas gráficas e de medidas descritivas variadas.

Mas temos um limitante parcial, que é o alto custo de licenciamento de alguns softwares, como o Mplus. Apesar do seu uso já ser popularizado, os custos elevados desestimulam novos licenciamentos, ao passo que estimulam o uso de outros programas mais acessíveis, como o R e o JASP, por exemplo. Desenvolvemos o conteúdo dando uma ênfase no Excel, no R, no JASP e em Python, que são programas consistentes em termos de ferramentas para análise de dados em geral. Os procedimentos para download e instalação do R, JASP e VSCode (para códigos em Python) softwares estão no Apêndice A, Apêndice B, Apêndice C deste manuscrito. Ressaltamos que trabalharemos mais enfaticamente o software R, que é, sem dúvidas, uma opção das mais promissoras em termos de análise quantitativa de dados, tendo em vista comportar um elevado volume de técnicas e ser totalmente livre para download e utilização.

O software R tem a restrição de demandar rotinas computacionais programadas, o que requer conhecimentos de fundamentos básicos sobre o assunto (programação). Por esta razão, o JASP tem maiores vantagens, dada sua interface mais amigável. Por outro lado, na medida em que o usuário se habitua com as rotinas do R, este pacote mostra sua superioridade e suas diversas vantagens frente aos demais.

Ao lado da expansão do uso de ferramentas de AED, a tecnologia da informação também assegurou condições para o desenvolvimento de novas técnicas de análise de dados. O custo computacional e a rapidez de processamento vêm deixando de ser problema em análise de dados, o que providencia condições para experimentações e novos avanços. De fato, as ferramentas clássicas, como as medidas de tendência central (média, mediana e moda), foram submetidas a tentativas de aperfeiçoamento para o melhor entendimento da informação que geram. A média, em especial, foi a medida que passou por mais aperfeiçoamentos, havendo atualmente largo uso complementar de ferramentas como média aparada e a média winsorizada.

Adicionalmente, a partir dos anos 1980, a AED se aproximou do universo da chamada análise estatística robusta e da estatística não paramétrica, o que teve como consequência um grande avanço de seus métodos. Não se trata de substituir técnicas não robustas por outras robustas, mas em se diversificar os referenciais interpretativos, complementando resultados e permitindo um melhor entendimento da realidade a partir dos dados que temos em mãos.

Por esta realidade, trazer para a AED uma abordagem que queremos chamar de moderna implica em resgatar esses diálogos e avanços, como forma de fornecer conteúdos relevantes para estudantes, profissionais e pesquisadores que precisam utilizar as melhores técnicas e ferramentas para fundamentar seus estudos e suas decisões.

- Aplicações principais em Ciências sociais e comportamentais

O conteúdo foi pensado levando em conta o processo de formação no Brasil, que tem disciplinas de estatística geral ou aplicada em cursos de graduação e pós-graduação. Em uma primeira visualização, na formação estatística em Ciências sociais e comportamentais, os componentes temáticos seguem a seguinte organização, para o nível de graduação:

- Fundamentos estatísticos: envolve os conhecimentos gerais da teoria estatística clássica, desde os elementos conceituais até a teoria inferencial de intervalo de confiança e teste de hipótese;
- Estatística aplicada: envolve os esforços de aplicação dos conhecimentos estatísticos convencionais nos problemas específicos da área de interesse. Em geral, a formação se resume a uma apresentação da teoria

estatística geral, com exemplificação baseada em variáveis do contexto de interesse.

Para o nível de pós-graduação, os estudos de estatística têm sido restritos aos cursos de mestrado e doutorado. A organização é geralmente a seguinte:

- Métodos quantitativos aplicados: consiste nos conteúdos convencionais da teoria estatística convencional, em uma espécie de retomada do conteúdo já ministrado nas disciplinas de graduação, porém em uma perspectiva orientada à pesquisa;
- Análise multivariada: envolve os conteúdos da teoria estatística multivariada orientada a aplicações em pesquisas, como, por exemplo, as técnicas de análise de regressão múltipla, análise fatorial, análise discriminantes etc.;
- Métodos variados: consiste em disciplinas de conteúdos diversos, adotados em situações específicas, como, por exemplo, as disciplinas de modelos lineares, análise de dados categóricos, modelos de *forecasting*, análise de sobrevivência etc.

Da experiência acumulada por um dos autores ao longo de anos de ensino, pesquisa e reflexões sobre o assunto, algumas conclusões emergiram. Primeiramente, parece restar poucas dúvidas de que o modelo de formação atual é ineficiente. Na verdade, o dia a dia do processo de formação dos estudantes mostra que, em geral, as disciplinas de Estatística são grandes entraves no processo de formação, tendo em vista que, na maior parte das vezes, são disciplinas que demandam muito tempo de estudo, e que os estudantes parecem estar mais preocupados em ‘se livrar’ da disciplina do que propriamente aprender o conteúdo.

Foi isto que motivou vários autores a desenvolverem materiais com conteúdo mais aplicado e mais próximo do dia a dia dos estudantes e dos problemas que enfrentarão como profissionais ou pesquisadores. Ao mesmo tempo, os docentes parecem vir tentando adaptar o processo didático para tornar o aprendizado mais fácil, com diversas novas ferramentas de ensino sendo recorrentemente tentadas.

Mas, na verdade, nem os autores nem os docentes bem-intencionados parecem ter solucionado o problema. Em uma afirmação arriscada, e que faremos baseados em uma crença oriunda somente da experiência de um dos autores,

acreditamos que depois que estas disciplinas terminam, o conhecimento que fica na memória dos alunos é muito restrito. Mais que isto, e pensando especialmente no nível de formação de pós-graduação, acreditamos que depois que a primeira disciplina termina mais da metade dos alunos simplesmente deixa de lado aquele conhecimento, procurando, ao máximo que puderem, se distanciar das disciplinas de formação quantitativa.

É muito provável que o mesmo ocorra com a maioria dos estudantes dos cursos de graduação, que guardam nas disciplinas de estatística uma memória associada aos momentos mais difíceis da faculdade. Na carreira profissional, esta dificuldade emerge, e, não sem razão, os profissionais em prática pouco usam os recursos de estatística aprendidos durante a faculdade.

Acreditamos haver motivações variadas para um comportamento assim, mas somos otimistas, pois a razão histórica de tal dificuldade parecia ser, entre outras coisas, a forte associação da disciplina de Estatística com os conteúdos de Matemática. Naturalmente, não deixou de ser assim, e não cremos no êxito de textos que se propõem a ensinar ou servir de suporte a profissionais que procuram providenciar um conhecimento de ‘Estatística sem Matemática’. Definitivamente, não existe estatística sem matemática!

Isto não quer dizer, por outro lado, que para fazer uso fluente das ferramentas de Estatística tenhamos que conhecer toda a Matemática subjacente. Isto não seria possível, nem mesmo para os profissionais com formação completa em Estatística; mas temos alternativas. Recordamos de um professor que um dia disse que não precisamos ser engenheiros mecânicos para operar bem um veículo. De fato, somos bons motoristas conhecendo bem os processos e procedimentos de uso do veículo e aplicando recorrentemente as melhores práticas no nosso cotidiano.

A comparação tem sentido, e de modo mais evidente quando lembramos que a evolução recente da tecnologia da informação, como as ferramentas de inteligência artificial, nos possibilita operar grandes volumes de dados e gerar resultados aplicáveis aos nossos problemas sem maiores dificuldades operacionais (ou seja, sem saber toda a engenharia por detrás dos procedimentos computacionais). A nós parece que nosso problema seria outro, ou seja, estamos provavelmente mais próximos de termos uma aproximação com a Estatística sem traumas, e assim ganhar o suporte de uma disciplina que, em nosso entendimento, tem um

potencial enorme de contribuir com nosso conhecimento e com nossas demandas profissionais, estudantis e de pesquisa.

Para este livro, fizemos um ordenamento no que nos parece ser uma forma mais eficiente de organização do conhecimento de estatística aplicada. Desenvolvemos uma concepção de conteúdo que segue uma gradação próxima do que se faz atualmente, porém procurando separar os componentes em blocos especializados, e procurando trazer alguns elementos mais modernos em termos de técnicas e ferramentas de informática.

O volume que aqui apresentamos cobre uma primeira disciplina, introdutória em relação às demais, mas nem por isto menos importante. Nossa meta é contribuir com a análise de dados, em primeiro momento oferecendo conhecimentos que são indispensáveis, mas que no processo de formação e nos livros são restritos a uma unidade temática, quando isto ocorre. Esperamos que, se bem assimilado o conteúdo aqui proposto, o estudante, o profissional e o pesquisador de Ciências sociais e comportamentais possam se defrontar com uma massa de dados quantitativos e desenvolver todas as análises preliminares e exploratórias, sem adentrar nos procedimentos mais depurados de testes estatísticos e ferramentas mais sofisticadas de análise multivariada.

Exercícios

1. Consulte projetos pedagógicos de cursos de graduação em Ciências sociais e comportamentais e analise a estruturação da formação em estatística e métodos quantitativos. Analise as ementas e as bibliografias e compare diferentes cursos e instituições.
2. Repita o exercício anterior, agora no contexto de cursos de pós-graduação.
3. Discuta com seus colegas e apresente o que você imagina serem problemas aplicados em sua área que demandam conhecimento de métodos estatísticos. Tente identificar em que medida você domina os conhecimentos demandados e o que você precisaria aprender mais, e veja se o que se ensina atualmente contempla a demanda.
4. Consulte periódicos ou sites de notícias e analise as matérias de interesse de sua área de formação. Em que medida o que se expõe ali é oriundo de aplicação de técnicas estatísticas? Tente identificar o quanto você domina os conhecimentos demandados e o que você precisaria aprender mais, e veja se o que se ensina atualmente contempla a demanda.

PARTE I.

FUNDAMENTOS GERAIS_

A finalidade desta parte é apresentar ao leitor as bases conceituais e instrumentais necessárias para a análise de dados. Podemos afirmar sem maiores dúvidas que todo o conhecimento da teoria estatística tem na análise de dados sua razão de existência (Pereira, 1997). Ao contrário de alguns desenvolvimentos da teoria matemática dita pura, não seria pensável o desenvolvimento de um conhecimento puro em Estatística, até pela concepção da disciplina como uma área de Matemática aplicada, que encontra seu real sentido nos diversos contextos de aplicação, que vão desde as ciências mais exatas (como as engenharias), passando pelas ciências médicas e biológicas, chegando até a Psicologia e a Administração.

Esta parte explora os elementos conceituais preliminares, contextualizando as aplicações indicadas, partindo de um esforço de conceituação que permita ao estudante, ao professor ou ao profissional uma base de consulta e referência sempre disponível. Se analisamos dados por métodos estatísticos, o entendimento é que precisamos antes de tudo entender o contexto mais amplo da teoria estatística, sem entrar, por outro lado, em conteúdos específicos de tal teorização.

Por esta razão, temos aqui em um capítulo a apresentação do conceito de Estatística, de sua estrutura geral na sua visualização como uma disciplina científica, explicitando as competências gerais do conhecimento estatístico, os principais contextos de aplicação e a estrutura de desenvolvimento do conhecimento na área. Ainda nesse primeiro capítulo, apresento uma primeira parte dos conceitos fundamentais que serão usados nos capítulos posteriores.

No segundo capítulo indico os elementos gerais do processo estatístico, explicando, e exemplificando quatro fases fundamentais. A etapa de análise, pela especificidade e pelo foco principal deste texto, é tratada no terceiro capítulo, que abordará, além deste conteúdo sobre análise, os conteúdos de Matemática que serão utilizados nos capítulos posteriores.

Os capítulos desta parte são fundamentais como formação da base conceitual e da contextualização de todo o processo de análise de dados, porém, sem perda de continuidade, os conteúdos aqui expostos poderão ser dispensados, caso o leitor, profissional ou professor entendam que já possuem os requisitos suficientes para ir diretamente à parte II. De toda sorte, recomendo ao menos a leitura do resumo no final dos capítulos, que sintetiza tudo o que foi exposto.

CAPÍTULO 1

CONTEXTO GERAL_

Como já indicado, a finalidade deste manuscrito é contribuir para o aprendizado e o uso de ferramentas estatísticas de análise exploratória de dados. Adicionalmente, informei que o conteúdo aqui indicado é, na verdade, uma etapa que tem continuidade a partir de outros conhecimentos aplicados, como os testes estatísticos e a análise multivariada. Por esta razão, é necessário primeiramente contextualizar a AED, com a indicação dos conceitos preliminares da teoria estatística geral, até chegarmos à visualização e à justificativa dos desdobramentos que seguem.

Ao final deste capítulo, o leitor deverá estar apto a responder aos seguintes questionamentos:

- O que é Estatística? Qual seu objeto, suas principais características e principais níveis de competência?
- O que caracteriza a Estatística aplicada e a Estatística teórica? No caso da dimensão aplicada, quais os principais campos de aplicação?
- O que é uma variável? De que forma podemos caracterizar nosso interesse de análise sobre variáveis? Quais são os principais tipos de variáveis?
- Como se mensuram variáveis? Quais são os tipos possíveis de escalas de mensuração?

1.1 Conceito de Estatística

A análise de dados é uma tarefa cotidiana, usada por quase todas as pessoas nas mais diversas situações da vida. Por exemplo, quando julgamos as alternativas de compra de um calçado, comumente levantamos informações gerais sobre o produto, incluindo a marca, as especificações, as alternativas de uso, o preço etc., e seguindo paralelamente ou em momento posterior, analisamos o conjunto de dados levantados e, daí tomamos a decisão.

Situações como a citada envolvem a tomada de uma decisão, que deve ser mais ou menos cuidadosa, a depender dos investimentos envolvidos. Mas não analisamos dados somente para tomada de decisões. Analisamos dados in-

clusivo para conhecer uma determinada situação, ainda que não envolva uma decisão específica, como a que indiquei acima. Podemos, por exemplo, levantar dados em retrospectiva, de um clube de futebol, ou da evolução populacional de uma rua, por exemplo, simplesmente para conhecermos melhor o que for de nosso interesse, por razões as mais diversas.

Tiramos desta primeira discussão uma conclusão central: analisamos dados, continuamente, e por diferentes razões, e por isto, **precisamos de dados**. Como consequência, a própria existência dos dados se coloca como um requisito fundamental para o processo decisório, ou mesmo para a construção do conhecimento. Por isto, mais que precisar de dados, **precisamos produzir e acessar os dados**.

Mas a análise de dados tem uma razão fundamental, eventualmente esquecida, mas nos últimos anos reiteradamente realçada: **precisamos avaliar dados em razão, fundamentalmente, da variação inerente a cada situação**. Por exemplo, se todas as informações possíveis sobre sapatos fossem de nosso conhecimento e fossem iguais (mesmas marcas, especificações, preços...), não faríamos pesquisas nem demandaríamos análises. Mas esta não é a realidade, como nossa experiência evidencia.

E afirmo com destaque neste texto o seguinte: **a característica mais fundamental de todo o desenvolvimento de técnicas de análise de dados é justamente a variabilidade da variáveis que são de nosso interesse**. O impacto desta afirmação é maior do que pode parecer pelo primeiro enunciado, porém ao longo deste texto entenderemos a razão da força deste entendimento.

Temos, portanto, três elementos centrais para a justificativa da análise de dados: demandamos dados para decisão e conhecimento; necessitamos produzir dados; e os dados variam, ou seja, possuem variabilidade. Em consequência, temos um novo desafio, que é **analisar estes dados**, ou seja, avaliar os dados demandados e produzidos, considerando a sua natureza variável. A resposta a este desafio motivou um largo desenvolvimento teórico nos campos de metodologia de pesquisa e de métodos quantitativos. Em nossa perspectiva, interessa principalmente a discussão associada aos métodos quantitativos existentes.

Uma ciência precisaria se desenvolver com esta finalidade. Esta foi a chamada por Estatística, assim denominada em decorrência de sua associação inicial com a análise de dados dos estados nacionais. Etimologicamente, a estatística

vem do latim *status*, que significa estado, de modo que a estatística foi durante séculos, e ainda hoje guarda esta responsabilidade, a ciência do estado, por sua finalidade de captar, organizar e apresentar dados sobre os estados. As outras línguas refletem bem isto, como por exemplo, o inglês, que usa *statistics* e *state*, e o espanhol, que utiliza *estadística* e *estado*. Em ambos os casos, a semelhança da palavra reafirma sua aproximação conceitual e histórica. Naturalmente, os grandes sistemas estatísticos nacionais (liderados pelo Instituto Brasileiro de Geografia e Estatística - IBGE, por exemplo), são órgãos de estatística, porém a Estatística disciplinar, aquela que se estuda nos cursos superiores, já foi além deste conceito, convergindo para uma outra definição.

Atualmente a Estatística é entendida **como a ciência (no sentido de ramo do conhecimento organizado e com um objeto bem definido) que se preocupa com a necessidade e a produção de dados, além de sua organização e apresentação, com a finalidade de analisar a variabilidade inerente a estes dados, por meio do emprego de técnicas e ferramentas de descrição, análise de associação e previsão**. Alguns autores partem da definição de estatística já a anunciando como um ramo da matemática aplicada. Reconheço o valor desta visão pelo ordenamento intelectual que promove, porém questionamentos outros levantam dúvidas sobre o valor e a própria lógica de enquadramentos e ramificações como esta. Não entrarei no mérito do entendimento, mas realço que o mais relevante é o entendimento da Estatística como um corpo de conhecimentos próprio e bem estruturado.

1.2 Competências centrais e aplicações

O desafio imposto pela definição acima é grande o suficiente para mobilizar um grande volume de conhecimentos e técnicas, que seguem avançando em termos de inovações e usos. Embora a análise de dados seja antes de tudo uma atividade de análise, que engloba elementos como interpretação e *insights* dos sujeitos que a empreendem, quando pensamos em análise de dados é natural que utilizemos ferramentas matemáticas, ou seja, instrumentos de manipulação quantitativa dos dados (convém ressaltar que nem toda análise de dados com métodos quantitativos é entendida como parte da disciplina de Estatística; outras disciplinas, como a de Matemática aplicada fazem o mesmo, sendo mais conhecidos os métodos de Análise Numérica e os modelos e métodos de Pesquisa Operacional).

Mas a Estatística não é necessariamente matemática (embora exista uma especialidade chamada de Estatística matemática). Nos últimos anos, a formação e o próprio conceito operacional da Estatística ganharam complementações. Em uma visão contemporânea, podemos elencar quatro componentes fundamentais para um referencial de competências estatísticas, que são os seguintes:

- **Fundamentos de Matemática:** consiste nos conhecimentos de teoria matemática que é usada na fundamentação da teoria e dos métodos estatísticos. Os conteúdos mais comuns são os seguintes: cálculo diferencial e integral, álgebra linear, métodos numéricos e principalmente teoria das probabilidades;
- **Teoria estatística:** envolve os aspectos teóricos próprios da Estatística, que vão desde os métodos de coleta de dados e princípios de sua análise, chegando até a teoria mais formal da inferência estatística. Aqui se incluem a larga diversidade de métodos e técnicas aplicáveis na análise de dados, desde as técnicas de análise exploratória até as técnicas sofisticadas da análise multivariada;
- **Fundamentos de computação:** são os conhecimentos e ferramentas computacionais que atualmente revolucionaram o conhecimento e o uso das ferramentas de estatística, tendo em vista que, com a tecnologia computacional hoje disponível, é possível manipular grandes volumes de dados e executar operações matemáticas que seriam por demais trabalhosas, se não impossíveis, por métodos manuais. Os conteúdos centrais são os de informática básica (edição de textos, planilha eletrônica, Internet), teoria da programação, sistemas de bancos de dados, e, principalmente, os pacotes estatísticos;
- **Conteúdo substantivo especializado:** são os conhecimentos de áreas de aplicação que servem de base para a interpretação dos dados, como, por exemplo, os fundamentos da Teoria econômica, de Psicologia, de Marketing, de Finanças, de Engenharia, de saúde, de educação, dentre outras.

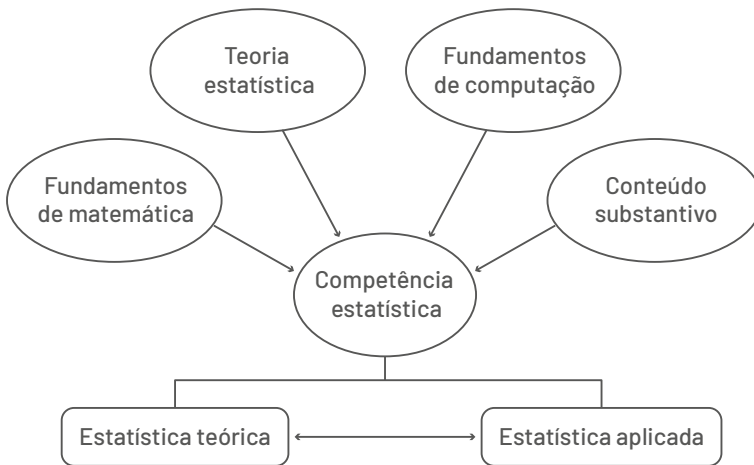
Em termos de estrutura disciplinar, as pessoas podem ser estatístico(a)s de profissão, ou podem ser usuários de ferramentas estatísticas. No primeiro caso, é necessário que façam curso superior em Estatística ou Matemática e, opcionalmente, sigam para cursos de pós-graduação. Este é o caso menos

recorrente, embora possamos dizer que a Estatística é uma das alternativas profissionais das mais promissoras para jovens em busca de inserção profissional. Os profissionais desta carreira desenvolvem com maior ênfase as habilidades de uso de ferramentas estatísticas, fundamentação matemática e uso de ferramentas computacionais, e em menor ênfase desenvolvem conhecimentos gerais algumas áreas disciplinares específicas. Aqui é que temos a **Estatística teórica** (ou estatística pura, que seria também possível).

O caso dos demais usuários já é diferente. Aqui, na verdade, temos os profissionais das mais diversas áreas que se utilizam das ferramentas e ideias centrais da teoria estatística com aplicação em seu contexto de trabalho ou de estudo. É o caso mais recorrente, e parece haver uma tendência de expansão de uso de ferramentas estatísticas em um número cada vez maior de áreas profissionais. É o contexto da chamada **Estatística aplicada**.

A Figura 1.1 ilustra a articulação do conjunto de competências apresentadas, mais os desdobramentos que a Estatística ganha como consequência de variações de uso e aplicação dessas competências.

Figura 1.1 – Formação da competência estatística



Em geral, no contexto da Estatística aplicada temos um desenvolvimento de competências associadas ao conteúdo substantivo de aplicação da teoria estatística básica (normalmente em nível superficial), e com ênfase desejavelmente maior em ferramentas computacionais. Uma ênfase reduzida é dada ao conteúdo matemático, como veremos posteriormente. Pelo estado que observamos atualmente, é possível atestar as seguintes aplicações principais:

- **Biologia e saúde:** aplicações na análise de dados biológicos e de saúde, seja na análise de seres vivos (na dita Biometria), seja na análise epidemias (na Epidemiologia), seja na análise específica de variáveis de interesse de saúde (na Estatística médica e na Análise de sobrevivência);
- **Agronomia:** aplicações nas diversas especialidades agronômicas, com ênfase especial nos métodos de teoria estatística geral, e nas especialidades da Experimentação agrônômica;
- **Engenharia (da produção):** aplicações na modelagem na análise de dados do processo de produção industrial, por meio do Controle estatístico de processos, do Controle estatístico da qualidade, e da Análise de confiabilidade;
- **Economia:** aplicações em análise de dados econômicos diversos, com foco em modelos de previsão e de análise de séries históricas de dados. A especialidade tem o nome de Econometria;
- **Ciências atuariais:** envolve a aplicação de ferramentas e conhecimentos de estatística para análise e gerenciamento de risco e previdência, incluindo os conhecimentos gerais (como Demografia e Análise de sobrevivência) e aqueles especializados desta área (análise de risco, tábuas atuariais...);
- **Psicologia:** está associada ao uso de ferramentas estatísticas aplicáveis à análise de variáveis psicológica e na produção de testes psicológicos (na especialização denominada testagem psicológica). A disciplina especializada nesta área é conhecida por Psicometria, que inclusive tem largo desenvolvimento teórico e metodológico, e cujas aplicações vão além do escopo da Psicologia;
- **Ciências sociais:** aplicações na análise dos fenômenos sociais diversos. Aqui as preocupações centrais estão na análise de adequação de ferramentas estatísticas na mensuração de variáveis sociais, ferramentas específicas da área, além dos estudos das populações (Demografia);
- **Administração:** aplicações de métodos e ferramentas estatísticas aos diversos contextos de decisão gerencial e de pesquisa acadêmica e de mercado (ver introdução).

Outras aplicações especializadas vêm emergindo, não pela recentidade de uso, mas pela progressiva especialização das ferramentas. É o caso, por exem-

plo, dos campos de Ecologia e Meio ambiente, Geografia e análise do espaço e Esportes.

A depender das áreas de aplicação, as intensidades de uso dos quatro componentes centrais variam. Assim, temos na Economia, por exemplo, uma tendência a um largo desenvolvimento de Teoria estatística, o mesmo ocorrendo com a área de Agronomia, Engenharia e (uma corrente da) Psicologia. Eu arriscaria dizer que estas áreas criam teoria estatística (em conjunto com os estatísticos de profissão), ou seja, operam com Estatística teórica, além de utilizarem a Estatística aplicada.

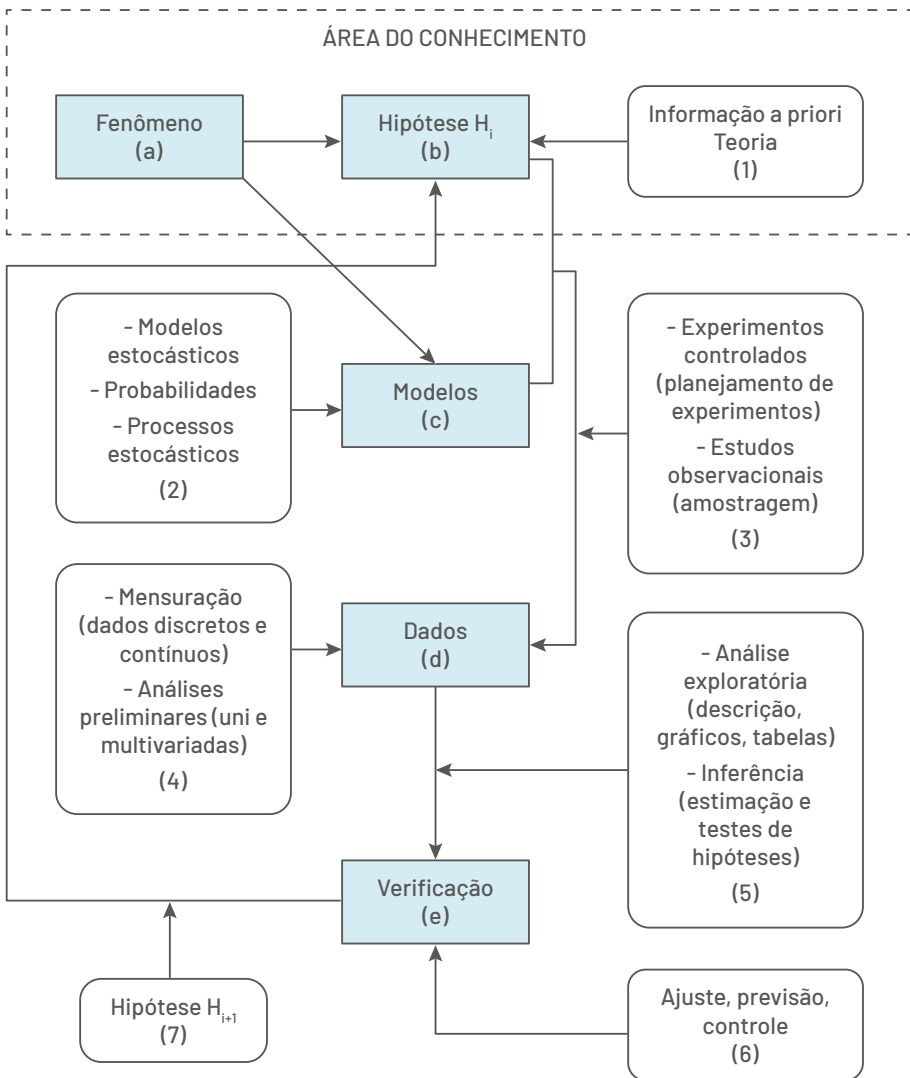
Mas se estas áreas mesclam uso e construção teórica, este já não é o caso das aplicações em Ciências sociais e Administração, por exemplo. Estas áreas, em geral, são grandes usuárias, e por isto a preocupação central está nas ferramentas fundamentais da análise de dados e em sua operacionalização por meio dos instrumentos computacionais. A base substantiva do conhecimento, ou seja, o conhecimento especializado da área de aplicação, é a referência central do uso da Estatística nestas áreas.

Como indicado anteriormente, nosso foco é justamente contribuir para o desenvolvimento de habilidades Ciências sociais e comportamentais, o que passa pela exploração do conteúdo estatístico básico (até o nível razoável para os profissionais e estudantes da área, ou seja, sem maiores desenvolvimentos especializados de Matemática e teoria estatística quantitativa), das ferramentas computacionais mais adequadas, e da aplicação dos conhecimentos especializados.

1.3 Processos centrais

A análise quantitativa de dados é uma atividade associada ao processo de geração, organização e apresentação de dados com vistas à construção de conhecimentos úteis (como indicado, tanto para a decisão como para o conhecimento sobre algo). A utilidade da análise de dados aos propósitos acadêmicos e profissionais é inegável, e não sem razão, todo o processo de preparação profissional envolve, em maior ou menor grau, a construção de competências de análise.

Figura 1.2 - Contexto geral da produção de conhecimento



Apreciamos de modo especial, e por isto aqui a retratamos, a reflexão do professor Basílio de Bragança Pereira, estatístico brasileiro que propõe o entendimento do conhecimento estatístico como a 'tecnologia da ciência' (ou seja, a tecnologia para a produção e estruturação do conhecimento especializado em torno de seu objeto específico). O processo completo está ilustrado na Figura 1.2¹.

1 PEREIRA, B. B. Estatística: a tecnologia da ciência. *Boletim da Associação Brasileira de Estatística*, ano XIII, n. 37, 2º quadrimestre, p. 27-35, 1997. Embora a proposta do autor esteja mais associada a uma perspectiva de produção de conhecimento, vemos facilmente que todos estes procedimentos em uma análise mais detalhada dos processos na esfera profissional.

A ilustração indica que, em uma dada área do conhecimento, são observados os fenômenos (a), que, em conjunto com informações *a priori* e teorias (1), geram uma afirmação que cremos, preliminarmente, ser a verdadeira, ou seja, uma hipótese, que está representada por H_i (b). Por exemplo, quando observamos um comportamento de reclamação e sinais de fadiga nos funcionários de uma organização (**fenômeno**), e em seguida confrontamos esta realidade com as teorias de gestão de pessoas e com as informações disponíveis (**informação a priori e teoria**), levantamos a hipótese de que os funcionários estão vivenciando um elevado nível de estresse no trabalho (**hipótese**).

Os fenômenos, em conjunto com modelos matemático-teóricos (2 – probabilísticos e estocásticos), gerarão modelos simplificados, e potencialmente explicativos, da realidade (c). As hipóteses geradas na área disciplinar, em conjunto com os modelos teóricos gerados, demandarão então a coleta de dados (d), que ocorrerá por meio de experimentação ou por procedimentos de amostragem (3). Seguindo o entendimento de nosso exemplo, é provável que as teorias de gestão de pessoas deem indicações de possíveis relações que expliquem a realidade de estresse, assim como podem apontar outros estudos que tenham explicado o fenômeno, porém sem a expectativa de ser um modelo determinístico, ou seja, isento de erros (**modelos probabilísticos e estocásticos**). De posse de modelos de relação entre variáveis, é possível definir então um **modelo** explicativo a ser testado. Com a modelagem matemática do fenômeno, decide-se então qual a melhor forma de levantar dados diretamente dos funcionários, seja realizando observação, aplicação de questionários, realização de entrevistas, ou mesmo realizando procedimentos experimentais.

A etapa seguinte consiste na realização dos procedimentos de mensuração de variáveis de interesse, além das avaliações preliminares dos dados (4). Temos então o conjunto de dados levantados. Em nosso exemplo, esta seria a etapa associada à definição de como as variáveis de interesse seriam aferidas, e depois de coletados os dados, estes seriam avaliadas preliminarmente.

É agora que entra em cena nosso foco principal nesta obra, que são os procedimentos de análise de dados. Na sequência indicada, se observam dois campos distintos da análise (que comentarei em breve): o primeiro é o exploratório (e descritivo); o segundo são os procedimentos de natureza inferencial e de estimação (5). Estes procedimentos de análise viabilizarão a verificação do modelo (e). No nosso exemplo, após o levantamento de dados sobre o comportamento dos funcionários, são então aplicados procedimentos de análise

para verificar se, efetivamente, o que os dados indicam converge com o que foi modelado anteriormente. Aqui, teremos uma indicação a respeito do motivo do comportamento estranho dos servidores e temos, adicionalmente, como testar se este comportamento é ou não próprio do que é tipicamente chamado de estresse.

Pela avaliação dos resultados destas análises é possível verificar se há, ou não, discrepâncias entre o modelo proposto (no passo c) e o verificado (no passo e). Caso haja variações, provavelmente são necessários procedimentos de ajustes, previsões e controles diversos (6). O resultado final de todos estes procedimentos desencadeia a construção de novas hipóteses H_{i+1} , que retornam então para a base de conhecimentos da área inicial (7), e contribuirão para a construção de outras hipóteses explicativas do fenômeno de interesse.

Segundo nosso exemplo, depois de confrontados os resultados, se necessário, são ajustados os entendimentos a respeito do comportamento dos funcionários, consolidando agora uma hipótese ‘nova’, que pode na verdade ser uma mera reafirmação do que foi pensado antes. Se for confirmado o estresse, então é hora de os executivos tomarem decisões para reduzir o nível (de estresse). Se não for, uma nova hipótese sinalizará outras potenciais explicações para o estado de ânimo observado nos funcionários, e novamente será necessário dar encaminhamento a novas especulações, modelos etc., até haver fundamentos sólidos para uma ação de intervenção dos executivos.

1.4 Requisitos conceituais de base

Para uniformização do que será exposto a partir do capítulo que segue, alguns conceitos preliminares precisam ser lançados desde agora. Do que mais utilizaremos, recorrentemente nos referiremos aos conceitos de variáveis e sua tipologia, e de escalas. Vejamos cada um destes.

1.4.1 Conceito de variável e abordagens possíveis

Os fenômenos que mais interessam à análise de dados e, como de resto, a toda a teoria estatística, são os fenômenos que apresentam variação de seu estado. Como indicado no início deste capítulo, a variação é a característica fundamental que sustenta toda a construção teórica e que demanda a formulação de métodos e técnicas de análise. Chamamos qualquer característica de interesse de um objeto que sofre variação de **variável**.

Com efeito, analisamos variáveis, e as características, sínteses, e regularidades de sua variação. A título de exemplificação, tomemos como referência o estado civil de uma pessoa. Como bem sabemos, em relação ao estado civil uma pessoa pode estar casada, solteira, viúva, divorciada, em união estável, dentre outras possibilidades. Quando analisamos o estado civil de um conjunto de 300 pessoas, é provável que todas estas configurações apareçam, inclusive algumas mais que outras, ou seja, é esperado que tenhamos pessoas casadas, solteiras etc. Além da variável em si, é relevante também a própria definição das categorias associadas.

A natureza da variação é que torna a variável em si interessante, e que motiva a necessidade de avaliação de um conjunto de pessoas. Adiante que a definição de categorias é algo que requer um cuidado especial, até pela variação possível (por exemplo, no caso do estado civil, podíamos usar apenas três: casado, solteiro, outros). O detalhamento depende do interesse da pesquisa e da relevância da informação mais ou menos detalhada. Retornaremos a esta questão ao longo dos capítulos seguintes.

Cabe observar que, normalmente, não estamos interessados somente em uma única variável. Em boa parte das vezes, estamos interessados em diversas características que permitam conhecer melhor cada variável levando em conta outras variáveis. No nosso exemplo, esperamos que as 300 pessoas pesquisadas apresentem manifestação de cada uma das alternativas de variação, porém se levamos em conta outras variáveis deste mesmo conjunto de pessoas, desconfiamos que a configuração destas categorias tem alguns determinantes. Assim, por exemplo, se no Brasil avaliamos separadamente as pessoas com faixa de idade até 20 anos, é muito provável que tenhamos um número muito maior de pessoas solteiras, um número menor de pessoas casadas, e um número raro de pessoas viúvas ou divorciadas. Por outro lado, se avaliamos um grupo de pessoas com idade acima de 50 anos é mais provável que a característica de solteiro seja mais rara, e as demais se apresentem em maior número.

Posteriormente, detalharei com maior clareza a caracterização agora apresentada, mas antecipadamente, convém deixar indicada a denominação de uso. Assim, no caso de interesse e foco restrito em uma só variável, dizemos que nossa abordagem é **univariada**. Por outro lado, quando estamos interessados no relacionamento de mais de uma variável, teremos, para o caso específico de duas variáveis (como no exemplo do parágrafo anterior), uma abordagem

dita **bivariada**; já no caso em que analisamos conjuntamente três ou mais variáveis, usamos uma abordagem chamada de **multivariada**.

1.4.2 Tipos de variáveis

Além de considerar as abordagens sobre as variáveis, convém ainda atentar para as possibilidades de indicação de suas características possíveis, e da forma como estas características podem ser anotadas. Na avaliação dos tipos possíveis de manifestação de uma variável, observamos duas direções possíveis, ou seja, podemos ter manifestações (alternativas de caracterização) que indiquem uma característica não quantitativa, e podemos ter manifestações que indiquem características tipicamente quantitativas. Vejamos dois exemplos:

- Suponhamos que uma fábrica de brinquedos produza quatro linhas de brinquedos, que chamaremos A, B, C e D. Neste caso, se temos diante de nós um conjunto de dois lotes de brinquedos e nosso interesse é classificar cada unidade, então nossa variável é o ‘tipo de brinquedo’, e as alternativas possíveis são cada um dos 4 tipos. Indicar que um brinquedo qualquer é da categoria A não indica uma quantidade; apenas o categoriza;
- Por outro lado, se neste conjunto de brinquedos estamos interessados em indicar a quantidade de produtos defeituosos por lote, então neste caso nossa variável será “número de defeitos por lote’ e atribuiremos a cada lote um número que dá expressão de uma quantidade.

No primeiro exemplo acima, indicamos as possibilidades de variação como características possíveis da variável, que são em si categorias nas quais a variável pode se manifestar. No exemplo sobre o estado civil, quando anotamos que uma pessoa é casada, estamos informando a categoria de variação do estado civil, porém não temos aí nenhuma indicação quanto à sua quantificação. Neste caso, temos o que chamamos de **variável categórica**. Assim, a **variável categórica é aquele em que sua variação é indicada por meio de categorias que representam as características que a variável deve assumir, sem referência a quantidades**. Temos, por outro lado, duas alternativas de indicação de categorias: nominal e ordinal.

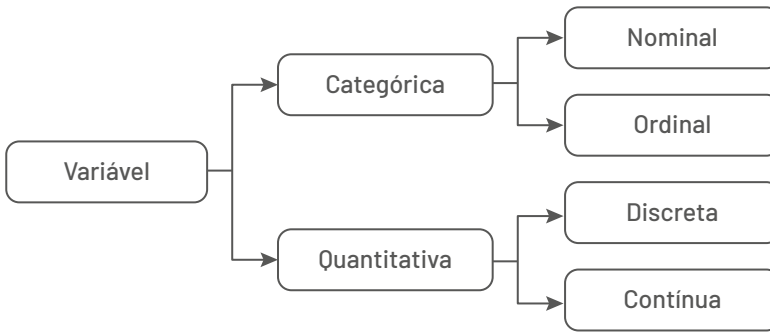
Temos uma **variável categórica nominal** quando as categorias possíveis são apenas para denominação das alternativas possíveis, de modo que a única relação que guarda com as outras categorias é diferença. Por exemplo, na variável gênero, as categorias possíveis são apenas ‘masculino’ e ‘feminino’. Entre essas duas opções, somente podemos dizer que o masculino é diferente do feminino, e vice-versa, não sendo possível estabelecer qualquer relação de ordem ou de quantidade relativa.

Por outro lado, é possível que as categorias que a variável pode assumir também estabeleçam uma ordem comparativa entre os sujeitos. Neste caso temos a chamada **variável categórica ordinal**, e podemos dar como exemplos as indicações das classes econômicas ou de níveis de instrução. No primeiro caso, se um sujeito está na classe A isto indica que está em uma posição econômica acima de um sujeito que está na classe B ou C, por exemplo; já no caso do grau de instrução, quando informamos que um sujeito possui nível básico, estamos indicando que tem uma formação menor que uma pessoa que possui nível médio ou nível superior. Na exposição do item subsequente esta classificação poderá ser mais bem compreendida.

Mas além desta classificação, é também possível adotar diferentes tipos de as **variáveis quantitativas**, ou seja, daquelas em que **as alternativas de manifestação indicam de forma explícita uma quantidade**. Para este caso, temos também dois tipos associadas às possibilidades de quantificação, que são as variáveis discretas e as variáveis contínuas. Vejamos cada uma destas.

Uma **variável quantitativa discreta** é aquela em que os possíveis valores a serem assumidos ou são finitos, ou são do tipo infinito enumerável (ou seja, possuem uma associação com o conjunto dos números naturais, embora as alternativas não precisem ser necessariamente de números naturais). Por exemplo, se estamos averiguando o grau de satisfação de um cliente com um serviço qualquer em uma escala de 7 pontos, de -3 a +3 (na sequência dos inteiros), então temos como alternativas de verificação um número finito de opções, o que torna a variável discreta. Por outro lado, se estamos averiguando a quantidade de itens com defeito em pacotes de parafusos, não temos definição de um limite estabelecido, podendo haver (teoricamente) um número infinito de itens com defeitos; neste caso temos sempre um número associado a algum número natural, o que torna a variável discreta.

Figura 1.3 – Classificação das variáveis



Já uma **variável quantitativa contínua** é aquela em que os valores possíveis que a variável pode assumir são números de uma escala contínua, incluindo qualquer valor do conjunto dos números reais. Como exemplo de variável contínua, vejamos inicialmente o peso de uma pessoa, e observemos que, a partir de zero (pois não existe uma pessoa com 0kg) até o infinito (embora não se conheçam pessoas com pesos elevados ao nível de 600kg ou 700kg, por exemplo), qualquer valor real é possível, inclusive números irracionais (mesmo que não tenhamos nenhuma aferição de peso com esta forma de apresentação, ela é teoricamente possível). Outras variáveis comumente abordadas como contínuas são as seguintes: altura das pessoas, raio de produto circular, distância percorrida, tempo, velocidade, consumo de água, consumo de energia, dentro outras.

Como forma de visualização do que foi exposto anteriormente, a figura 1.3 ilustra as classificações apresentadas. Convém observar que, do ponto de vista da operacionalização matemática e estatística, trabalhar com variáveis contínuas possui maiores vantagens, embora não pareça à primeira vista (os estudos de teoria das probabilidades e estatística matemática não deixam quaisquer dúvidas disto). Por outro lado, em certas circunstâncias operacionais, é mais conveniente usar variáveis discretas, como veremos posteriormente.

Como exemplo, temos o caso da variável idade, que, por ser referenciada em tempo, pode ser abordada de forma contínua. No entanto, o uso corrente de idade a manifesta sempre em número de anos, o que a torna uma variável discreta. Naturalmente, o estudante, o profissional ou o pesquisador deverá ter a sensibilidade para definir a forma de acesso e análise considerando os condicionantes diversos de cada situação.

1.4.3 Tipos de escalas

Chamamos de escalas ao instrumento de aferição (ou mensuração), que suporta o processo de indicação ou atribuição de símbolos (normalmente numéricos) para as categorias ou quantidades de uma variável. Trataremos deste assunto no capítulo seguinte, mas por enquanto é conveniente definir os quatro tipos genéricos clássicos de escalas, com base em Stevens (1946), que são: nominal, ordinal, intervalar e razão².

– Escala nominal

A escala nominal é aquela em que utilizamos símbolos para representar as categorias de um objeto, para não mais que a indicação do nome da característica. É o caso, por exemplo, de variáveis como gênero, que podemos aferir assim: 1 – indica masculino; 2 – indica feminino. Outras variáveis com aferição semelhante são local de origem (por exemplo, entre interior e capital), raça, cor, formação, profissão, religião... Naturalmente, na exploração de uma variável nominal usaremos sempre uma escala do tipo também nominal.

O cuidado especial que precisamos tomar na aplicação desta escala está justamente na definição das categorias, pois estas precisam ser ao mesmo tempo exaustivas (ou seja, contemplem todas as possibilidades de manifestação da variável), e exclusivas (ou seja, que não se sobreponham). Mas em geral, desde que as categorias estejam indicadas de forma cuidadosa, este tipo de escala é o mais simples de se verificar e analisar, mas nem por isto é menos relevante no universo das alternativas disponíveis de escalas.

– Escala ordinal

Além da escala nominal, temos, como vimos, aquelas variáveis cuja aferição é feita por categorias que fixam uma ordem. Nestes termos, a regra de aferição é feita pela chamada escala ordinal. Assim, por exemplo, quando avaliamos as posições nos *rankings* de eficiência de empresas, indicamos uma empresa como estando em 1º (primeiro lugar), que indica ser a mais eficiente; outra fica em 2º (segundo lugar), representando a segunda mais eficiente, e assim segue. Neste

2 Esta classificação foi proposta por Stanley Smith Stevens, na seguinte referência: STEVENS, S. S. On the theory of scales of measurement. *Science*, v. 103, n. 2684, p. 677–680, jun. 1946. Observe que os tipos de variáveis e de escalas possuem contiguidade, embora haja uma diferença central: a escala é dirigida à medição ou à aferição de uma variável.

caso, embora estejamos indicando uma ordem, cada posição é não mais que uma categoria possível nas alternativas de verificação da variável ‘posição’.

Isto fica mais claro quando observamos que a lógica de classificação não tem que seguir, *a priori*, uma lógica única da posição. Assim, no nosso exemplo usamos a referência de 1º para o melhor, porém podíamos ter usado uma inversão, de tal modo que o 1º fosse o pior, e isto não faria perder o sentido, desde que fosse indicada a convenção usada. Mais que isto, a ordem convencionalizada pode até contrariar a lógica matemática associada, pois uma posição não indica quantidade sobre outra posição (por exemplo, se a posição 1 representa o melhor e o número 2 segundo melhor, porém em termos matemáticos, o 2 é maior que 1).

Dois aspectos sobre as escalas nominal e ordinal merecem ser indicados. Primeiramente, são tipos de escalas presentes na grande maioria das pesquisas. Por esta razão, foram desenvolvidos métodos de análise e operacionalização específicos, como forma de superar as limitações próprias de sua natureza não quantitativa. Em outras palavras, mesmo sendo escalas não quantitativas (alguns preferem chamar de qualitativas; o cuidado aqui é não confundir a medida qualitativa com o método qualitativo), há um sem número de técnicas estatísticas para sua análise, desde técnicas univariadas até técnicas multivariadas sofisticadas.

Além disto, e como um meio de facilitar sua operacionalização em manuseio estatístico, é comum que as categorias sejam representadas por símbolos numéricos (guardando-se sempre a informação da convenção adotada ao longo das análises). Assim, na variável gênero, em lugar de operar com as categorias de masculino e feminino, podemos convencionalizar que a categoria masculina será indicada por 1, e a categoria feminino por 2, ou o inverso. Sempre que necessário, as convenções são resgatadas, para efeito de análise.

É evidente a associação dos tipos de escalas apresentados com os tipos de variáveis anteriormente descritos. Como é natural, também teremos tipos específicos de escalas para as variáveis quantitativas. Temos então dois tipos fundamentais que são as escalas intervalar e de razão. Vejamos cada uma delas.

– Escala intervalar

Uma escala intervalar é aquela em que adotamos, por nossa conveniência, pontos de referência a partir do qual definimos intervalos fixos de variação.

Por exemplo, considerando a temperatura (que é a quantidade de energia cinética de um determinado corpo físico), se tomamos a água como um líquido de referência, e definimos a temperatura da água no estado de fusão como 0° (na escala Celsius) e no estado de ebulição como 100° (Celsius), então estamos adotando uma escala de intervalo para mensuração de temperatura.

Naturalmente, não estamos usando uma medida que tenha referência no ponto de partida da variação de energia cinética, que seria o estado de zero energia. Embora isto fosse desejável (e há escalas para tanto), é fácil perceber que uma referência de medidas como aquela indicada (0 a 100) é mais fácil de ser usada e compreendida, pois estão baseadas em uma realidade de temperatura e em valores de referência que vivenciamos cotidianamente.

As escalas de intervalo têm largo uso em Administração, Ciência política, Educação e Psicologia, dada a natureza de suas variáveis. De fato, se estamos medindo, por exemplo, aprendizado, a melhor forma que encontramos foi definir uma variação de pontos (10, 20, 100, ou 1000, dentre outras), sendo mais comum no Brasil o intervalo de 0 a 10 (certos exames de competência, como ENADE ou teste ANPAD, usam escalas distintas, definidas em uma fórmula de padronização em Z). O mesmo acontece em aferição, por exemplo, do nível de qualidade de vida percebida por uma pessoa ou o nível de aprovação de um governo. A alternativa possível que encontramos foi definir uma variação entre dois extremos, por exemplo, de 1 a 7 ou de 1 a 10, em que o menor valor indica o nível mais baixo, e o maior valor o nível mais elevado.

Embora as escalas intervalares largamente usadas, seu uso possui algumas limitações e críticas. Ainda assim, com um pouco de flexibilidade e tolerância, a manipulação destas variáveis vem sendo realizadas sem maiores restrições em termos de ferramentas quantitativas.

– Escala de razão

Observamos que a escala intervalar têm sua definição em uma convenção de uso que não tem por referência um ponto de zero absoluto. Como vimos, o ponto zero da escala de temperatura Celsius não indica a nulidade da energia cinética do corpo. É uma referência em que a razão entre os números não possui a equivalência de razão entre quantidades nos números reais. Por exemplo, uma medida de temperatura de 20°C é duas vezes a medida de temperatura de 10°C , mas isto não significa que a temperatura (física) seja o dobro. De

fato, mudando a escala para outro parâmetro de referência (como a escala Fahrenheit), por exemplo, as duas temperaturas terão outros valores (como a fórmula de mudança de escala é $F=1,8C+32$, então $20^{\circ}\text{C}=68^{\circ}\text{F}$ e $10^{\circ}\text{C}=50^{\circ}\text{F}$). Isto indica que, se mudamos os pontos de referência, nossa interpretação da razão entre os números também muda (de fato, 68 não é o dobro de 50). Isto ocorre justamente porque o ponto de referência da escala não é o zero.

A dita escala razão é aquela em que temos a referência de medida em seu zero absoluto, a partir do qual são indicadas variações em unidades conveniadas. Esta unidade de variação é a magnitude da unidade padrão, a partir da qual outras medidas são feitas na razão do que se quer medir pela unidade de referência. Por exemplo, para medir distâncias a unidade de referência é o metro; portanto, se queremos medir qualquer distância, basta tomar o valor observado e extrair a razão com esta magnitude de referência. Se o metro ‘cabe’ 9,5 vezes dentro do objeto medido, então dizemos que o objeto mede 9,5 metros (observe que a distância tem uma medida de zero absoluto, ou seja, nenhuma distância, ou a igualdade dos pontos de partida e de chegada).

Em geral, as ciências possuem um grande volume de variáveis com mensuração por escalas razão, e quanto maior o número, melhores serão as análises, tendo em vista que as variáveis assim medidas viabilizam a aplicação de quase todas as técnicas de análise de dados existentes. Alguns exemplos de escalas razão de interesse são aquelas aplicadas em idade, tempo de serviço, número de trabalhadores, peso, altura, salário, quantidade produzida, quantidade de erros de produção, dentre outras.

Uma observação relevante acerca das quatro alternativas de variáveis é que uma variável não é mensurada somente por uma escala razão, intervalar, ordinal ou nominal. Ou seja, uma variável pode estar sendo mensurada na forma de razão, mas é possível que venha a ser mensurada na forma intervalar, ou mesmo ordinal ou nominal. Vejamos dois exemplos:

- A variável temperatura quando medida na escala Kelvin é mensurada em uma escala do tipo razão. No entanto, quando abordada na escala Celsius se torna mensurada por uma escala intervalar. Com um pouco de esforço, podemos adotar uma referência de conversão tal que uma temperatura até 10°C seja dita ‘menor’; de 10°C a 50°C seja dita ‘média’; e acima de 50°C seja dita “maior”. Isto torna a mensuração por meio de uma escala ordinal;

- Para o caso da variável idade, esta é obviamente mensurável por uma escala razão; no entanto, dada a dificuldade de acessar a idade em determinadas circunstâncias, é comum que sejam adotadas faixas de idade em intervalo fixos, do tipo até 20 anos, de 21 a 30 anos, de 31 a 40 anos... Isto converte a mensuração da idade para uma escala de intervalo, mas é possível ainda definir faixas para efeito de ordenamento de idades (escala ordinal), ou para simples denominação (escala nominal).

Este processo de conversão é possível, mas possui limitações. Assim, em geral, é possível tornar uma escala ordinal em nominal, uma escala de intervalo em ordinal e nominal, e uma escala razão em intervalo, ordinal e nominal. No entanto, não é necessariamente possível tornar uma variável de medida estritamente nominal em uma variável quantitativa de razão. Por exemplo, a variável nominal ‘gênero’ pode ser categorizada em ‘masculino’ e ‘feminino’, e, ainda que possamos adotar como referência que masculina será indicado por 0 e feminino por 1, isto não torna a variável quantitativa, simplesmente porque as duas categorias de gênero não são conversíveis em quantidades.

O Quadro 1.1 sintetiza a discussão deste subitem, relativo às quatro alternativas indicadas.

Quadro 1.1 – Síntese sobre os tipos de escalas

Escala	Sentido	Exemplo	Técnicas
Nominal	Uso de números para classificar por meio de denominação	Gênero, Estado civil, Renda	Algumas técnicas descritivas e multivariadas
Ordinal	Uso de números para classificar por meio de ordenamento	Preferências, Opiniões, Classes sociais	Algumas técnicas descritivas e multivariadas
Intervalo	Uso de números para quantificar por meio de uma convenção de extremos	Atitudes, Preferências, Intenções, Temperatura	A maioria das técnicas descritivas e multivariadas
Razão	Uso de números para quantificar por meio da razão da magnitude por uma unidade padrão	Idade, Renda, Preço, Tempo-padrão	Praticamente todas as técnicas descritivas e multivariadas

1.5 Resumo do capítulo

Neste capítulo foram destacados os seguintes pontos centrais:

- A finalidade foi contextualizar o campo da Estatística, como a ciência de referência para as técnicas, ferramentas, conceitos e modelos de análises de dados;
- Foram destacadas as propriedades fundamentais dos dados que são abordados em estatística, a saber: dados são base do processo decisório e de formação de conhecimento; necessitam serem produzidos para uso; são caracterizados pela variação que lhe é inerente. Daí emerge o desafio de construir alternativas e formas de sua análise;
- A Estatística é a área de conhecimento que se debruça sobre este problema, sendo entendida como a ciência (no sentido de ramo do conhecimento organizado e com um objeto bem definido) que se preocupa com a necessidade e a produção de dados, além de sua organização e apresentação, com a finalidade de analisar a variabilidade inerente a esses dados por meio do emprego de técnicas e ferramentas de descrição, análise de associação e previsão;
- Para o trabalho estatístico são demandadas quatro competências centrais, a saber: fundamentos de Matemática; teoria estatística; fundamentos de computação; e conteúdo substantivo especializado;
- O envolvimento com a Estatística pode ser feito por pessoas preocupadas principalmente com a teoria estatística em si, na Estatística teórica, ou podem ser profissionais de outras áreas que utilizam os instrumentos da estatística em problemas de suas áreas, por meio da Estatística aplicada. Este texto tem foco na Estatística aplicada;
- Há diversas áreas de aplicação, sendo destacadas as áreas de Biologia e Saúde, Agronomia, Engenharia, Economia, Ciências atuariais, Psicologia e Administração;
- A Estatística pode ser vista como a tecnologia da produção de conhecimento, partindo de fenômenos empíricos e de informações teóricas, definindo hipóteses, desenvolvendo modelos para avaliação dessas hipóteses, coletando dados para teste dos modelos, implementando ferramentas de análise, e julgando a adequação das hipóteses e o aperfeiçoamento dos modelos, como forma de aperfeiçoar o conhecimento;

- A Estatística está interessada em avaliar características de objetos, que, por sua variação, são chamados variáveis. Quando nosso interesse e abordagem estão sobre uma só variável, temos a análise univariada; se estamos interessados em avaliar conjuntamente duas variáveis, temos a análise bivariada; já se avaliamos simultaneamente três ou mais variáveis, temos a chamada análise multivariada;
- A depender das possibilidades de variação, as variáveis podem ser categóricas (não quantitativas) ou quantitativas. Se forem categóricas podem ser ou nominais ou ordinais, e se forem quantitativas podem ser discretas ou contínuas;
- Para aferir, ou mensurar, as variáveis, temos quatro tipos de escalas, a saber: escalas nominais, que somente denominam as características; escalas ordinais, que ordenam as categorias de respostas; escalas intervalares, que adotam um ponto de referência arbitrário e mede em intervalos fixos; escalas razão, baseadas na comparação da característica de um objeto com uma grandeza fixa pré-especificada.

Exercícios

1. Tomando por base os requisitos dos dados de interesse da análise estatística (demanda de dados para decisão e conhecimento; necessidade de produção de dados; imperativo da variabilidade), avalie de que forma estes se apresentam em situações como:
 - a) Atrasos de voos em um aeroporto
 - b) Fluxo de veículos em um terminal portuário
 - c) Duração do esforço de diagramação de um novo lançamento de uma editora
 - d) Falhas de atendimento de um call center
 - e) Acerto em uma prova de concurso
 - f) Trânsito de pessoas entre cidades de regiões metropolitanas
2. Encontre dois problemas típicos de sua realidade de trabalho ou estudo e em seguida explique a aplicação dos os passos do modelo da figura 1 na avaliação do problema (sugestão: use um exemplo sobre comportamento de compra de consumidores finais [marketing], e um exemplo de redução de lucros mensais [finanças]).

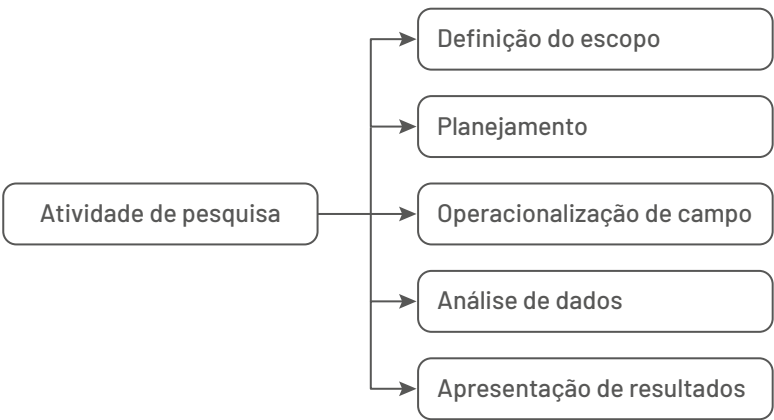
3. Avalie a demanda de conhecimentos de estatística aplicada entre as diferentes áreas (por exemplo, Saúde e Economia) a partir de livros ou programas de disciplinas disponíveis. Avalie em que medida os conteúdos se diferenciam em que medida convergem. Nos casos de convergência, avalie a variação de abordagem dada em cada contexto de aplicação.
4. Analise problemas e variáveis de seu campo de trabalho ou estudo e indique pelo menos duas possíveis variáveis de cada um dos tipos indicados (nominal, ordinal, discreta e contínua).
5. Considerando as variáveis a seguir, indique as alternativas de escalas possíveis (observe que podemos ter mais de um tipo de escalas para uma mesma variável):
 - a) Renda familiar
 - b) Conhecimento adquirido em uma disciplina
 - c) Conhecimento necessário para exercer uma profissão
 - d) Lealdade de clientes
 - e) Fervor religioso
 - f) Idade de uma máquina
 - g) Lucro diário de uma loja
 - h) Vendas de assinaturas de jornais
 - i) Acesso a um site de internet
 - j) Tempo de permanência em um site

CAPÍTULO 2

FUNDAMENTOS
CONCEITUAIS E
OPERACIONAIS_

Conforme observado no capítulo anterior, quando nos debruçamos na atividade de análise de dados, na verdade estamos operando sobre uma etapa de um processo que se inicia bem antes. Na verdade, todas as etapas desse processo desempenham um papel indispensável para a qualidade geral do processo. Assim como não temos condições de desenvolver adequadamente conhecimento sem uma análise de dados consistente, também não temos condições de desenvolver uma análise de dados consistente se as demais etapas do processo estão fragilizadas.

Figura 2.1 – Etapas da atividade de pesquisa



Neste capítulo, apresentamos uma síntese geral do detalhamento das demais etapas, ao mesmo tempo em que apresento novos conceitos fundamentais que serão usados nos capítulos seguintes. Como forma de constituir um referencial mais pragmático, a exposição que segue será exemplificada por meio

de uma pesquisa sobre avaliações de estudantes de Administração e Turismo sobre seus cursos³.

Visualizamos a construção do processo em cinco etapas, a saber: definição do escopo de pesquisa; planejamento; operacionalização de campo; análise de dados; e apresentação de resultados. A figura 2.1 ilustra esta visão, e na sequência apresento o desdobramento de cada uma das etapas.

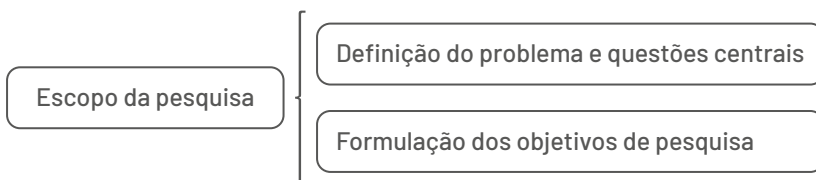
Neste capítulo, são apresentadas as três primeiras etapas, e no capítulo seguinte apresento as duas outras, bem como outros complementos. Assim, ao final deste capítulo é esperado que o leitor consiga responder às seguintes questões:

- Quais são as atividades que precedem à etapa de análise de dados? Em que medida o conteúdo destas etapas tem associação com o processo de análise em si?
- O que é um escopo de pesquisa e quais são suas principais atividades?
- Como se faz o planejamento de uma pesquisa? Quais os principais passos associados?
- O que é um trabalho de campo? De que forma se realiza e quais os condicionantes mais importantes desta etapa de pesquisa?

2.1. Definição do escopo de pesquisa

Entendemos como definição do escopo de pesquisa a realização de duas etapas fundamentais, quais sejam: a definição do problema de pesquisa e o delineamento dos objetivos. A figura 2.2 ilustra estas etapas:

Figura 2.2 – Definição do escopo da pesquisa



³ A pesquisa era parte de um projeto mais amplo que foi desenvolvido nos anos de 2007 e 2008, liderado por este autor e conduzido juntamente com estudantes do Curso de Mestrado Acadêmico em Administração da Universidade Estadual do Ceará. O que apresento aqui é apenas uma parte do que foi pesquisado, com alguns ajustes, devido à finalidade didática deste material.

É fácil entender que, quando precisamos analisar dados, isto ocorre simplesmente porque precisamos saber de algo. Em outras palavras, todo o empreendimento operacional de análise de dados atende à finalidade de busca de conhecimento sobre uma realidade ou um objeto. Entendemos que a definição da necessidade de conhecimento é na verdade a busca de responder a algum questionamento que fazemos sobre o que nos interessa.

Este questionamento é o ponto de partida de todo o processo, e é comumente citado como **‘problema de pesquisa’**. Trata-se, portanto, de uma questão que motiva o esforço empreendido. Por exemplo, quando um grande varejista não tem conhecimento da evolução dos hábitos de seus consumidores, em relação, digamos, ao consumo de massas, então levanta este questionamento genérico: “de que forma os clientes estão consumindo produtos da categoria de massas?”. É imediato que outros **questionamentos adicionais** vêm em sequência, como, por exemplo, as seguintes:

- Com que periodicidade se alimentam de massas?
- Quais as principais categorias de produtos?
- Quais as principais marcas mais consumidas?
- Quais produtos eles gostariam de ter à disposição, mas não têm ainda?

A identificação cuidadosa do problema de pesquisa é certamente a etapa mais relevante de todo o processo de produção de informações, seja para conhecimento, seja para suporte à decisão. A recomendação é que seja dado dedicado todo o tempo e o esforço necessários a esta etapa, como forma de evitar que o trabalho subsequente seja prejudicado.

A consequência mais imediata de uma construção adequada do problema de pesquisa e das questões associadas é a definição consistente dos **objetivos de pesquisa**. Não entendo conveniente fixar aqui regras para indicação e enunciação de objetivos aqui, até porque os manuais de metodologia parecem fazer isto com maior detalhamento, mas em nível genérico, recomendo que sejam indicados os objetivos em número suficiente para dar aos procedimentos subsequentes um guia de suas definições específicas, mas que não torne o esforço de pesquisa e levantamento de dados algo exagerado em termos de tempo, custos e esforço.

A título de ilustração, dou referência de um objetivo de pesquisa de uma indústria de alimentos que pretendia identificar as marcas consumidas por

peças de classes A e B, em uma verificação direta. Este objetivo praticamente impunha a necessidade de observação das geladeiras, armários e depósitos domiciliares de peças destas classes. Naturalmente, isto é possível, porém a um custo muito elevado, e, a depender da quantidade de peças a serem pesquisadas, demoraria um tempo muito grande.

- Escopo de pesquisa do exemplo ilustrativo

Em nosso exemplo de referência, nosso problema de pesquisa questionava o seguinte: como os estudantes das áreas de Administração, Turismo e Contabilidade avaliam seus cursos e suas profissões? Esta pergunta de partida induziu à construção de outros questionamentos mais específicos, a saber:

- Qual a percepção de valor dos estudantes em relação a seus cursos? Como avaliam os benefícios emocional, de reputação e de articulação social?
- Qual sua percepção sobre a reputação da profissão no universo do trabalho?
- Qual o nível de identificação pessoal dos estudantes com o curso e com a profissão?
- As percepções e o nível de identificação são uniformes entre os estudantes ou variam de acordo com as características pessoais do estudante e de seu contexto de atuação e formação?

Em qualquer contexto de pesquisa, os objetivos são sempre responder aos questionamentos indicados. No entanto, para fins de exposição e justificativa, e mesmo para um melhor ordenamento das ideias, é possível proceder aos enunciados dos objetivos. Assim, em nosso caso, e a partir destas questões foi possível desenvolver um conjunto de três objetivos, a saber:

1. Analisar, junto a uma amostra de estudantes de diferentes características pessoais e de contexto de formação, sua percepção de valor em relação à sua formação em processo.
2. Analisar, junto a estes estudantes, seu nível de identificação pessoal com a profissão a ser seguida, assim como sua percepção de reputação social da profissão.
3. Desenvolver avaliações destas dimensões (valor percebido, identificação e prestígio social) segundo as diferentes características pessoais e de contexto de formação.

É relevante ressaltar que, em geral, em processo de formação e de pesquisa acadêmica, os enunciados são praticamente obrigatórios. Já em pesquisas de interesse exploratório, gerencial ou de decisão, o grau de necessidade de

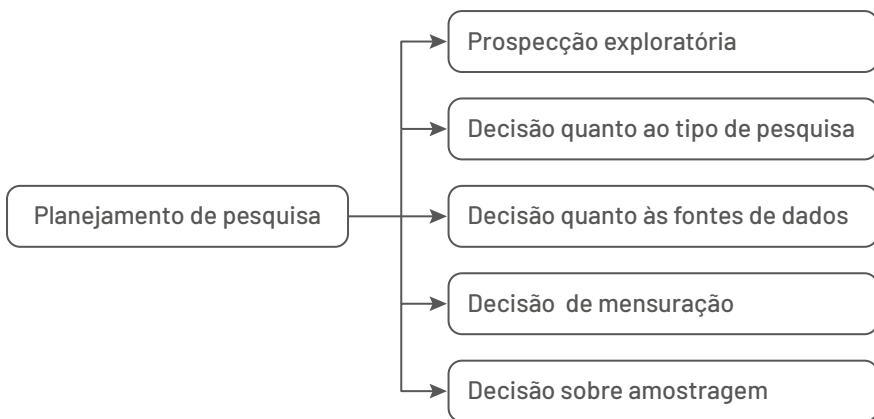
formalização de objetivos depende do nível de exigência da situação ou dos profissionais envolvidos.

No fechamento do escopo de uma pesquisa é possível que sejam também incluídos outros elementos textuais, como um texto de contextualização do assunto e a justificativa do esforço a ser empreendido. Estas demandas são mais comuns em pesquisas do tipo acadêmico. Fontes bibliográficas de metodologia da pesquisa sobre o assunto são abundantes e largamente disponíveis em livrarias, bibliotecas e mesmo na internet, e o leitor interessado pode buscar qualquer dessas fontes para mais detalhes.

2.2. Planejamento da pesquisa

Esta etapa consiste nas seguintes atividades centrais: prospecção exploratória sobre o assunto; decisão quanto ao tipo de pesquisa; decisão quanto às fontes de dados; decisões de mensuração; e decisões sobre amostragem. A figura 2.3 ilustra este conjunto de passos, que são detalhados em seguinte:

Figura 2.3 – Planejamento da pesquisa



– Prospecção exploratória

Após a definição do escopo, uma ação mais prudente, antes de qualquer empreendimento de campo, é verificar se as respostas ao problema e às questões de pesquisas já não existem ou estão disponíveis em algum lugar. Nestes termos, é conveniente que seja realizada em fontes diversas uma **prospecção exploratória**, como por exemplo, a verificação de revistas especializadas sobre o assunto, relatórios disponíveis na própria organização ou em outras

semelhantes, em sites da internet, dentre outras fontes possíveis. No caso citado no capítulo anterior, sobre o estado de ânimo dos funcionários e a possível associação com estresse no trabalho, é possível que antes de se empreender um processo sistemático de pesquisa, analise-se se outras situações semelhantes já não foram verificadas e documentadas, permitindo uma ação direta, sem maiores perdas de recursos e de tempo.

Caso a prospecção exploratória já atenda aos objetivos da pesquisa, então o processo encerra aqui. Mas este não é o caso em grande parte das vezes, ou seja, é quase sempre necessário o prosseguimento de esforços sistemáticos de pesquisa. Neste caso, a primeira decisão consiste em escolher uma ou mais alternativas, dentre os diversos encaminhamentos existentes.

– Decisão quanto ao tipo de pesquisa

Trata-se de decidir qual o **tipo de pesquisa** a ser realizado, que pode ser uma pesquisa exploratória sistemática, uma pesquisa descritiva, ou uma pesquisa experimental. Detalhadamente, temos:

- Pesquisa exploratória sistemática: esforço de pesquisa que se baseia em procedimentos sistemáticos e bem ordenados de verificação de fontes de informações variadas (ver item seguinte). A depender da necessidade do interessado e da disponibilidade de recursos, são realizados procedimentos do tipo qualitativo, que consistem em ações como a verificação sistemática do fenômeno, as entrevistas em profundidade com pessoas que podem fornecer informações, as entrevistas estruturadas em grupos (como o chamado *focus group*), e até a imersão do pesquisador no ambiente de interesse da pesquisas, no modelo chamado etnográfico (ou netnográfico, se o ambiente de observação for baseado na internet);
- Pesquisa experimental: consiste nas pesquisas em que o pesquisador manipula uma situação, na tentativa de aproximar sua manipulação (ou experimento) do que se observa na realidade. O procedimento visa fundamentalmente, analisar relações de causa e efeito, pois os resultados da situação manipulada são sempre comparados com outros resultados, e são avaliadas diferenças e especificidades. Na pesquisa experimental nem sempre são necessários muitos dados para avaliação, bastando que se tenha uma quantidade de dados suficiente para garantir a comparação das situações envolvidas na análise;

- Pesquisa de *survey*: consiste no levantamento de dados e informações sobre um conjunto razoavelmente grande de elementos, com acesso direto e sem procedimentos experimentais. Normalmente são utilizadas amostras grandes ou mesmo todos os sujeitos ou elementos da população de interesse. De posse dos dados, são aplicados métodos estatísticos de análise.

As pesquisas experimental e de *survey* caracterizam-se por serem, normalmente, etapas posteriores aos estudos exploratórios sistemáticos, e não são sempre necessárias. Ambas também requerem que o material a ser analisado possua características bem específicas, devido à necessidade de manipulação quantitativa envolvida. Esta preocupação que emerge em seguida consiste então em definir as **fontes de dados** que serão usadas nas análises.

– Decisão quanto às fontes de dados

Há uma caracterização convencional de fontes de dados, que as define como fontes primárias e fontes secundárias. As ditas **fontes secundárias** são fontes que já possuem os conjuntos de dados de interesse, que são coletados de maneiras diversas e eventualmente disponibilizados aos interessados. A depender da natureza da fonte secundária, estes dados são mais ou menos organizados, e mais ou menos confiáveis. Por exemplo, quando precisamos de dados sobre as características populacionais de um bairro para fins de decisão quanto à expansão de um supermercado ou abertura de uma nova escola, então os órgãos oficiais de planejamento e pesquisas (IBGE e institutos estaduais e municipais de planejamento) normalmente já possuem dados bem estruturados e fortemente confiáveis. Outras fontes são jornais, revistas, relatórios de empresas e órgãos públicos, sites de internet, bibliotecas, sindicatos etc.

O pesquisador deverá observar em cada situação, inclusive considerando o seu interesse de pesquisa, se os dados secundários são confiáveis ou não, e se são suficientes para o interesse de pesquisa. Caso não sejam, então é necessário empreender uma coleta direta junto às fontes que entendemos como primárias.

As **fontes primárias** são aquelas que possuem em si os dados e informações de interesse da pesquisa, e cujo acesso é vital para o desenvolvimento da análise. As fontes podem ser pessoas, mas não o são necessariamente. Por exemplo, se em uma pesquisa é necessário saber a estrutura de serviços de uma re-

gião, e as ruas são os objetivos de interesse da pesquisa, uma verificação local constitui o acesso à fonte primária de dados; o mesmo se pode dizer de uma pesquisa que tem interesse em avaliar características de um imóvel, em que informações secundárias normalmente estão disponíveis em catálogos ou na internet, mas eventualmente é recomendada uma verificação direta na fonte, ou seja, no próprio imóvel. A definição do problema e das questões de pesquisa já remeterá à natureza da fonte a ser pesquisada (o que reforça a importância de uma atenção maior nas primeiras etapas da pesquisa).

– Decisão de mensuração

Após a decisão quanto às fontes, cabe agora definir quais são exatamente os dados de interesse da pesquisa, e, em sequência, quais são as alternativas de mensuração existentes para viabilizar a etapa seguinte, além da forma de acesso, por meio da definição do instrumento de pesquisa.

Comumente, o problema de pesquisa e os questionamentos adicionais já informam quais são os principais **dados a serem coletados**, seja em pesquisas exploratórias, seja em pesquisas descritivas. Mas convém salientar que, na maioria das vezes, são acessados mais dados do que se pensa em um primeiro momento, e a indicação a partir do escopo inicial pode não ser o bastante.

Antes de definir os dados, é sempre recomendado avaliar estudos diversos já realizados, pois estes indicam as diversas alternativas de dados a serem pesquisados (observe que aqui se fortalece o valor das prospecções exploratórias iniciais e a reunião de materiais que virão a ter sua utilidade já nesta fase do estudo). A decisão sobre quais serão os dados é facilitada tremendamente por estes estudos anteriores, e sua verificação já pode inclusive indicar as alternativas de **mensuração** que são mais comumente usadas.

Por mensuração entendemos ao processo de averiguar e atribuir símbolos (normalmente números) às características de interesse de um objeto. A atribuição de símbolos é feita às características do objetivo e não ao objeto em si. Por exemplo, se queremos verificar a satisfação de uma pessoa diretamente, mensuramos sua satisfação e não a pessoa que a possui. Esta característica de interesse, quando pode ser mensurada, recebe a denominação de **construto**, e sua definição formal e detalhada é feita no momento anterior, em que se definem os dados a serem coletados.

Observe que os construtos são de dois tipos fundamentais, a depender da possibilidade de sua averiguação: se o verificamos diretamente, temos o dito **construto observável**, como é o caso, por exemplo, do peso de uma pessoa, que pode ser averiguado por meio de uma balança. Por outro lado, há construtos que não temos como fazer esta observação direta. Isto ocorre com grande frequência em pesquisas sociais, educacionais e comportamentais. No caso do campo da educação, por exemplo, como podemos medir o aprendizado de um estudante? Na verdade, ainda não temos como acessar o aprendizado verdadeiro do estudante, e o máximo que conseguimos é inferir seu grau de aprendizado pelo acerto em uma prova. Neste caso, a escala (a prova) está medindo um **construto** que chamamos **latente**. O leitor deve observar a similaridade entre o que apresentamos como construtos e como variáveis (no capítulo anterior). Na verdade, os dois conceitos têm algumas diferenças, principalmente porque o construto pode ser medido por uma ou mais variáveis. De fato, os constructos latentes normalmente não possuem uma só variável de medição, como ilustra o exemplo indicado sobre a prova, em que cada questão constitui uma variável.

Outro aspecto relevante acerca da mensuração é a averiguação sobre o que chamo de **escala de verificação**, ou seja, sobre a regra de atribuição de símbolos que usaremos. Este assunto remete à discussão sobre os sistemas de medidas, alguns dos quais altamente ordenados, como aqueles aplicados em pesquisas tecnológicas, que, inclusive, possuem a disposição largo aparato de instrumentos de aferição. Neste tipo de sistema, são fixadas unidades padrão, além das relações entre unidades, se for o caso, e as regras de conversão entre diferentes padrões de medição (por exemplo, a medição de distância tem um padrão largamente usado que é baseado no metro, e também um padrão, também muito usado, que é baseado na milha).

Como informado no capítulo anterior, em pesquisas sociais, educacionais e comportamentais, é comum o uso de escalas intervalares para medição de variáveis quantitativas. Este tipo de escala tem algumas flexibilidades na definição de seus pontos de referência na medição. Temos aqui aquelas regras de verificação de 0 a 10, classicamente usadas em provas de conhecimento, mas que podem variar de 0 a 10, de 0 a 20, dentre outras possibilidades. Há também as clássicas escalas de concordância conhecidas por “escalas de Likert”, que podem ser de 1 a 5, de 1 a 7, de 0 a 10, dentre outras possibilidades, ou as escalas de intensidade (de pouco a muito, de fraco a forte...), que podem usar as mesmas quantidades de pontos das escalas de concordância.

Estas escalas normalmente são aplicadas com a mediação de instrumentos na forma de questionários, no que convencionalmente chamamos de instrumentos do tipo ‘lápiz e papel’. Depois que a internet passou a ser mais amplamente usada em pesquisas deste tipo, vem sendo utilizada a denominação de ‘coleta online’.

Ao longo da exposição dos capítulos seguintes serão apresentados outros elementos e informações sobre escalas que poderão esclarecer melhor o que apresentei. O quadro 2.1 apresenta uma síntese do que foi apresentado acima sobre mensuração, extraído de livro publicado por este autor⁴. Um aspecto que merece ser realçado é o fato de que nem sempre a escala de verificação mais natural é aquela que pode ser aplicada. Retomando um exemplo já apresentado no capítulo anterior, suponhamos a necessidade de realização de um estudo em que precisaremos averiguar a idade dos respondentes. A escala de verificação natural seria a escala de razão, na qual o sujeito informaria sua idade em um número. No entanto, a experiência mostra que, a depender das pessoas a serem pesquisadas, parte delas não gosta de revelar a idade real, de modo que uma alternativa que já se mostrou razoável foi definir faixas de idade e pedir para a pessoa indicar a faixa de idade na qual se encontra, em uma escala que pode ser classificada como ordinal.

Quadro 2.1 - Conceitos-chave

Elemento	Definição
Mensuração	Procedimento de verificação empírica, a partir de instrumentos adequados, de uma característica bem definida e diferenciada de um objeto, com a finalidade de atribuição de números seja para definir quantidades, seja para definir classificações.
Construto	Característica de um dado objeto que apresenta variações que viabilizam quantificações ou classificações, e que são bem delimitadas em relação a outras características do mesmo objeto.
Construto latente	Tipo de construto que não pode ser mensurado diretamente, mas que apresenta manifestações mensuráveis (ex.: satisfação).
Construto observável	Tipo de construto que pode ser diretamente mensurado (ex.: altura de uma pessoa).
Escala de mensuração	O instrumento de mensuração completo, que reúne o componente físico (ou sua descrição), além do conjunto de regras de aplicação, de atribuição e de análise.
Escala de verificação	Regra de verificação direta da forma como atribuímos os números às variações e quantidades da característica mensurada no objeto.

4 Costa, F. J. (2011). Mensuração e desenvolvimento de escalas. Rio de Janeiro: LCM.

Pelo exemplo, é fácil entender por que também nesta etapa é fortemente recomendada a verificação de livros de pesquisas e outros materiais teóricos que podem ao mesmo tempo indicar outros estudos já realizados, e ainda apontar quais as alternativas de escalas de verificação mais adequadas. Somente após consolidadas as etapas anteriores (seleção das variáveis e definição de mensuração) é que se empreende o esforço de seleção ou construção do **instrumento de pesquisa**.

A este respeito, a primeira etapa é verificar se o instrumento já existe e está disponível. Caso esteja, então o trabalho do pesquisador consiste somente em efetuar os ajustes de formato e outros aspectos menos relevantes. Por outro lado, se tal instrumento não estiver ainda construído, então devemos tomar algumas decisões, sintetizadas a seguir.

Primeiramente, e a depender da natureza das variáveis a serem prospectadas, será necessário utilizar instrumentos tecnológicos ou instrumentos de questionamento direto a um respondente. Os instrumentos tecnológicos têm suas características específicas a depender dos dados a serem coletados. Por exemplo, se os dados forem referentes às categorias e marcas de produtos comercializados, os dados de registros de *check-outs* de supermercados são adequados. Já para o caso de reação neurológica a determinados fatores de motivação de uma pessoa, então instrumentos médicos e de neurociências são necessários. Em caso de demanda de instrumentos deste gênero, a recomendação é a busca de referências sobre as alternativas possíveis.

Já os instrumentos de questionamento direto são aqueles usados para levantar dados sobre pessoas ou outros objetos de interesse que requerem uma verificação pessoal (como organizações, lugares, objetos...). No primeiro caso, ou seja, quando fazemos verificação direta com pessoas, usamos o nome genérico de questionário, e quando forem outras entidades (coleta de dados de organizações, objetos...) usamos um roteiro de coleta. Tanto no caso do questionário quanto do roteiro, o cuidado deve ser principalmente na sua construção, que deve evitar sobreposições de questões, além de ser baseada em um encadeamento adequado das questões.

Especificamente sobre o questionário, o seu significado é decorrente de sua própria construção: trata-se de um conjunto de questões. Quando é usado em procedimentos exploratórios qualitativos, recebe também a denominação de roteiro de entrevista. Convém realçar um aspecto relevante da construção

do questionário, que deve guiar o esforço de sua construção: devemos sempre lembrar que o respondente está, na grande maioria das vezes, fazendo o favor de fornecer os dados demandados, e, por esta razão, não deve ser desrespeitado com questionários longos demais, complexos demais, e com questões ofensivas ou inconvenientes. A leitura cuidadosa e o pré-teste são formas de evitar este problema.

– Decisão sobre amostragem

A etapa seguinte do planejamento consiste em definir quais serão os sujeitos que fornecerão os dados de interesse. Chamamos de **universo ou população de pesquisa** todo o conjunto de objetos que possuem os dados de interesse. Por exemplo, se estamos avaliando o crescimento do Produto Interno Bruto dos países, então o universo de pesquisa serão todos os países. Também se estamos avaliando o perfil profissional de servidores públicos federais, então o universo será composto por todos os servidores públicos nesta categoria de serviço.

Observe que a definição do universo de pesquisa é uma etapa extremamente relevante, uma vez que as delimitações não são sempre muito claras. Nos exemplos acima isto fica evidenciado, e basta ver que, no caso dos países, precisamos primeiramente definir o critério de caracterização de um país; embora o reconhecimento pelas Nações Unidas seja o parâmetro aparentemente mais seguro, os interesses de pesquisa podem adotar outros critérios, e incluir países não reconhecidos por esta instituição. No caso de um universo de servidores públicos federais, é conveniente especificar, por exemplo, se os servidores são somente aqueles de carreira e com estabilidade legal, ou se incluem os servidores de empresas públicas, que são celetistas e têm uma carreira diferenciada, ou aqueles que estão somente exercendo cargos de confiança.

A quantidade de ‘elementos’ do universo de pesquisa constitui o ‘tamanho do universo’. Um aspecto importante relativo ao universo de pesquisa e seu tamanho consiste em sua natureza finita ou infinita. Nos dois exemplos indicados, independente dos critérios de caracterização, temos necessariamente um número finito de objetos de pesquisa, ou seja, o universo é de tamanho finito. Por outro lado, se nosso interesse é avaliar o volume diário de chuva de uma determinada região, em princípio, desde a primeira observação em diante, não há um limite de fim na sucessão de dias. Assim, a população neste caso é do tipo infinito.

De um modo geral, as populações finitas são as mais abordadas. No entanto, do ponto de vista do instrumental matemático, considerar uma população como infinita tem maiores vantagens operacionais, o que faz com que populações muito grandes sejam consideradas como infinitas para efeito de análise.

A abordagem de pesquisa sobre todos os elementos de uma população recebe o nome de **censo**. No entanto, nem sempre é possível abordar todos os elementos de uma população. Existem motivos que dificultam este acesso, como, por exemplo, a infinidade da população, a demora para captar todos os elementos, se possível, o elevado custo de pesquisar todos os sujeitos etc. Nestes casos, abordamos um subconjunto do universo, que recebe o nome de **amostra**, e a abordagem de pesquisa sobre uma amostra recebe o nome de **amostragem**.

A grande maioria das pesquisas e análises de dados é baseada em amostragens, inclusive porque, a depender de sua estrutura, pesquisar toda a população pode ser um trabalho muito grande e custoso sem necessidade em termos de reconhecimento da realidade, uma vez que há técnicas de análise de dados que permitem compreender bem as características da população a partir da amostra. Tendo em vista o ‘custo’ menor da pesquisa de amostragem, foram desenvolvidos estudos especializados sobre o assunto, que não caberiam no espaço aqui disponível, mas estão disponíveis em livros e sites de internet. Os aspectos de decisão mais relevantes são os seguintes:

- Quanto à forma de coleta: a amostragem pode ser de tal modo que os sujeitos da amostra sejam selecionados de forma aleatória ou probabilística, ou podem ser coletados baseados em critérios de conveniência ou acessibilidade, ou seja, de forma não aleatória ou não probabilística. Em cada caso há variações e alternativas que precisam ser levadas em conta, e os textos de metodologia apresentam detalhamentos neste sentido. Em geral, é recomendado que o procedimento se aproxime, tanto quanto possível, da amostragem probabilística, uma vez que as técnicas de análise de dados para generalização adotam por pressuposto que a seleção foi aleatória;
- Quanto ao tamanho da amostra: a amostragem é um procedimento alternativo e satisfatório para viabilizar muitas pesquisas; no entanto, o ideal de pesquisa é que sempre consigamos realizar um censo, e assim ter a informação completa sobre a população. Neste sentido, o princípio

fundamental para decisão quanto ao tamanho da amostra é simples: quanto maior, melhor, ou seja, quanto mais elementos forem pesquisados melhor será o procedimento de análise (esta regra vale inclusive para populações infinitas). Para efeito de análise e generalização, há inclusive uma formulação que permite definir o tamanho que permite fazer determinadas inferências, e os livros de pesquisa ou estatística inferencial também apresentam estas formulações⁵. A circunstância específica de cada pesquisa indicará ao pesquisador a melhor alternativa.

- Planejamento da pesquisa do exemplo ilustrativo

Na pesquisa realizada sobre as avaliações dos estudantes de administração e turismo, os procedimentos do planejamento de pesquisa foram os seguintes:

- **Prospecção exploratória:** antes de qualquer procedimento e considerando o problema e os objetivos da pesquisa, foi empreendida uma pesquisa em periódicos acadêmicos, que são as principais fontes de outros estudos já realizados com interesse convergente. Desta etapa, resultaram os elementos teóricos e inclusive indicações sobre os possíveis instrumentos a serem usados em um possível trabalho de campo;
- **Tipo de pesquisa:** considerando as evidências de outras pesquisas e as possibilidades de acesso aos dados, entendemos que a melhor opção era desenvolver uma pesquisa do tipo *survey*. Estudos anteriores e verificações exploratórias pessoais dos envolvidos na pesquisa serviram de subsídio de dados exploratórios preliminares e dispensaram uma pesquisa exploratória qualitativa;
- **Fontes de dados:** a pesquisa tinha por finalidade estudar percepções de estudante, e estes eram, naturalmente, as fontes de dados principais. Como fontes secundárias, foram verificadas também sites de faculdades para fundamentar o empreendimento de coleta e evitar custos (por exemplo, de visitar instituições nas quais não existissem os cursos de interesse);

5 Embora não seja uma informação muita difundida, o tamanho da amostra é também fortemente determinado pela natureza das técnicas de análise a serem usadas, pois, a depender da sua complexidade, algumas técnicas requerem um mínimo de dados para sua execução (como são alguns procedimentos de análise multivariada), ao passo que outras requerem somente que a quantidade de dados já viabilize a aplicação das técnicas (como é o caso dos procedimentos experimentais). Para as técnicas que usaremos aqui este requisito não surgirá, uma vez que não estamos interessados em fazer análises de inferência de amostras para populações.

- **Mensuração:** primeiramente observamos nas pesquisas anteriores, e a partir delas foi possível indicar os principais dados que favorecem a resposta ao problema de pesquisa. Além destes, a discussão entre os envolvidos possibilitou a indicação de outros dados e alternativas de mensuração. O quadro a seguir indica as principais variáveis, além da indicação dos tipos associados.

Variável	Tipo	Escala (aferição)
Curso	Categórica nominal	Nominal (com o nome do curso)
Semestre	Quantitativa discreta	Razão (número do semestre)
Instituição	Categórica nominal	Nominal (opções de pública ou privada)
Idade	Quantitativa discreta	Razão (número de anos)
Sexo	Categórica nominal	Nominal (opções de masculino e feminino)
Estado civil	Categórica nominal	Nominal (opções de casado, solteiro e outros)
Renda	Quantitativa discreta	Razão (número de salários-mínimos)
Valor funcional	Quantitativa discreta	Intervalar (escala de concordância de 1 a 7 para a afirmação: para mim tenho mais benefícios que sacrifícios neste curso)
Valor social	Quantitativa discreta	Intervalar (escala de concordância de 1 a 7 para a afirmação: para mim tenho mais benefícios que sacrifícios neste curso)
Valor emocional	Quantitativa discreta	Intervalar (escala de concordância de 1 a 7 para a afirmação: estabeleço boas relações sociais a partir do curso)
Valor de reputação	Quantitativa discreta	Intervalar (escala de concordância de 1 a 7 para a afirmação: a instituição que estudo tem boa reputação no mercado de trabalho)
Identificação com a profissão	Quantitativa discreta	Intervalar (escala de concordância de 1 a 7 para a afirmação: pessoalmente, tenho boa identificação com minha futura profissão)
Prestígio da profissão	Quantitativa discreta	Intervalar (escala de concordância de 1 a 7 para a afirmação: a minha futura profissão tem boa reputação no mercado de trabalho)

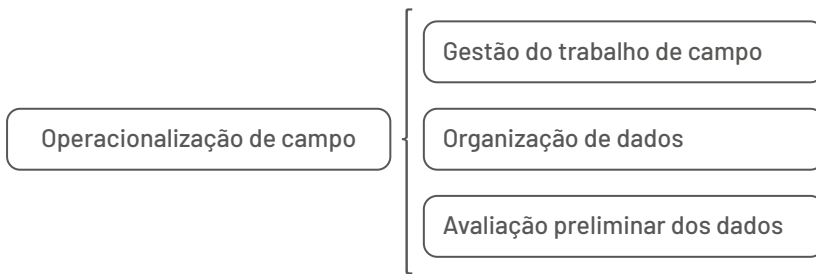
Pela finalidade da pesquisa, a melhor opção de acessar os dados era por meio do questionário estruturado, que foi desenvolvido, pré-testado e em seguida aplicado junto aos estudantes em suas próprias instituições. Para facilitar o acesso, optamos por aplicar os questionários em sala de aula, tendo em vista a disponibilidade dos estudantes e a facilidade em emitirem respostas.

- **Amostragem:** o universo de pesquisa teórico seriam todos os estudantes brasileiros dos dois cursos. Naturalmente, o acesso seria impossível, e então optamos por uma amostragem, junto a 113 estudantes de instituições superiores de ensino da cidade de Fortaleza. A forma de coleta foi por acessibilidade e conveniência.

2.3. Operacionalização de campo

A última etapa que precede a análise dos dados é o trabalho de campo, ou seja, a efetivação do processo de prospecção de dados, juntamente com a organização destes dados para a etapa de análise que segue. A figura 2.4 ilustra o conjunto de passos desta etapa, que são detalhados a seguir:

Figura 2.4 – Operacionalização do campo



– Gestão do trabalho de campo

Nesta etapa, temos o esforço empreendido para efetivamente buscar os dados onde quer que estejam. Assim, no caso de pesquisa em fontes secundárias, é necessário ir até os espaços onde esses dados estejam disponíveis, sejam bibliotecas, órgãos públicos, arquivos públicos ou de organizações privadas, ou mesmo na internet. Em geral, a etapa de campo para dados quantitativos é simplificada, comparativamente às etapas em que serão entrevistadas pessoas. Por esta razão, desenvolvo mais detalhadamente este último aspecto.

A experiência tem sinalizado que a preocupação com a captação de dados junto a pessoas envolve algumas decisões de base, a saber:

- Primeiro, o instrumento de pesquisa deve ser adaptado ao respondente, em termos de linguagem e da possibilidade de fornecimento das informações. Reiterando o que foi dito anteriormente, os instrumentos de pesquisa não podem ser ofensivos em suas perguntas, e não pode haver senso de desrespeito, como acontece em questionários longos demais;
- Quanto à forma de acesso, temos como alternativas clássicas as seguintes: envio por correio, resposta por internet, resposta por telefone, acesso direto e pessoal. Cada uma dessas opções tem suas vantagens e desvantagens, e a circunstância da pesquisa indicará a opção que consiga ao mesmo tempo viabilizar representatividade, qualidade dos dados coletados e custo;

- Preparação da equipe de coleta: consiste em selecionar adequadamente as pessoas que serão responsáveis pela coleta, inclusive no seu treinamento em relação ao instrumento e sua preparação para contingências diversas de campo;
- Estratégia de acompanhamento: sempre que possível, é relevante que seja feito um acompanhamento *in loco* do processo de coleta. Em qualquer estratégia de coleta, é possivelmente preciso que sejam realizados ajustes, como, por exemplo, em casos em que uma unidade de coleta pré-determinada não dispõe dos dados, ou quando se evidencia algum vício (como a coleta restrita a determinados grupos, quando for necessário haver variação de grupos).

Em geral, é esta etapa a que mais consome recursos, e, por esta razão, é necessário que seja construído um orçamento para o campo, com a indicação das demandas materiais e de pessoal, o que fortalece a necessidade de cuidado ao longo de todo o processo de coleta.

– Organização dos dados

Ao longo do trabalho de campo, os dados vão sendo produzidos e carecem de uma organização, que pode ser feita em momento posterior ou paralelamente ao trabalho de campo. Para pesquisa tanto qualitativas quanto quantitativas temos atualmente uma boa disponibilidade de softwares que ao mesmo tempo em que facilitam a organização, e já colocam os dados disponíveis para a etapa de análise.

Em geral, os softwares com planilhas (ou seja, com espaços pré-definidos para denominação de variáveis e recepção de dados) são os mais eficientes para organização de dados. Nas pesquisas de vocação quantitativa, os softwares com planilhas mais conhecidos são o MS Excel, da Microsoft, o SPSS (*Statistical Package for the Social Sciences*), e o JASP. Destes, o Excel, apesar de ser o de menor custo, é o que possui menos funcionalidades, embora seja um software bastante completo.

Outros softwares de análise de dados, como o R e o S-Plus, são mais limitados, não sendo recomendados para a etapa de organização de dados. Especificamente para o caso de uso do software R, que é um dos recomendados neste material, é sugerido que a organização seja feita primeiramente no MS Excel, que é compatível com o R, como veremos nos capítulos posteriores.

Vejamos agora como é o processo de preparação de dados, iniciando no Excel, em seguida no R e em Python (a exposição a seguir é parcial e direcionada ao uso que teremos neste livro; recomendo a potenciais interessados que consultem livros ou sites com maiores detalhamentos sobre os programas aqui apresentados)⁶.

a) Excel

A inserção de dados no Excel é simples, pois não requer codificação em duas telas. O procedimento consiste em definir, na primeira linha, os nomes das variáveis, e nas linhas abaixo fazer a inserção dos dados, a partir dos instrumentos coletados. O detalhe especial aqui é a possibilidade/necessidade de se definirem códigos para as respostas das variáveis, para facilitar a inserção dos dados sem ter que, necessariamente, escrever as respostas. Por exemplo, em uma indicação de faixas de renda, pode não ser interessante escrever as faixas, e somente apontar códigos relativos às respostas.

b) R

No pacote R, o procedimento mais fácil de inserção de dados consiste em organizar os dados em uma planilha de Excel e em seguida importá-los como base de dados de análise para o R. Há diversos procedimentos possíveis, e aponto a seguir aquele que mais utilizo:

- Preparação da base de dados no Excel (usando vírgula para separar decimais)
- Cópia da base de dados e colagem em um arquivo de bloco de notas
- Utilização do procedimento de importação para o R, indicado abaixo:

```
library(tcltk) # TCL/TK para abrir o bd
Endereço<-tclvalue(tkgetOpenFile(title="Abrir Banco de Dados"));
Dados<-read.table(file=Endereço, header=TRUE, dec=",");
attach(Dados)      # Fixando os dados para análise
names(Dados)       # Verificando os nomes das variáveis
```

⁶ Enfatizamos a necessidade da leitura do que segue ser feita em contato direto com os programas, para exercício paralelo à leitura.

c) Python

Seguindo a mesma recomendação para importação no R, é interessante organizar a planilha no Excel e, em seguida, importar para o ambiente de programação da linguagem Python. Para a elaboração deste manuscrito, estamos utilizando o Visual Studio Code (VS Code), que é um editor de código-fonte leve desenvolvido pela Microsoft.

- Preparação da base de dados no Excel (pode usar vírgula para separar decimais e depois indicar isto no código).
- Utilização do procedimento de importação para o VS Code, indicado abaixo, no qual importaremos a biblioteca pandas para ler o arquivo Excel:

```
import pandas as pd

df = pd.read_excel(r'COLOQUE AQUI O CAMINHO DO ARQUIVO COM BARRAS
INVERTIDAS', decimal=',')

print(df)
```

d) JASP

No JASP, é possível inserir os dados diretamente no software ao clicar no comando new data ou podemos importar uma base já preparada em outro software, como no software Excel. O JASP aceita diferentes formatos de arquivos preparados em outros programas: JASP (.jasp), arquivos de texto (.csv, .txt, .tsv), planilhas (.ods, .xls, .xlsx), SPSS (.sav, .zsav, .por), Stata (.dta) e SAS (.sas7bdat, .sas7bcat, .xpt). Para a importação, utilizaremos o seguinte caminho: clicaremos em **show main menu** (na versão que estamos trabalhando, encontra-se nas três linhas em azul no canto superior à esquerda) → **open** → **computer** → **Browse** (encontre a base de dados que quer importar e selecione a base) → **open**. Após esses passos, a base será aberta no JASP e estará pronta para ser manuseada.

e) Complementos

É comum que os dados sejam transferidos de uma planilha para outra. Assim, os dados do R podem ser facilmente transferidos para o Excel, por exemplo, utilizando o seguinte comando:

```
if (!require(openxlsx, quietly = TRUE)) install.packages("openxlsx")
library(openxlsx)
write.xlsx(dados,"NOME DO DATAFRAME.xlsx")
```

Em Python, podemos utilizar o seguinte código para salvar o banco de dados editado, transferindo do ambiente de programação para o Excel:

```
import pandas as pd
df.to_excel("NOME DO ARQUIVO.xlsx", index = False,
engine="openpyxl")

#observe que df é o nome que estabelecemos para o banco de dados.
```

Particularmente quando os dados são coletados pela internet, o programa de coleta normalmente faz a exportação de dados para alguma planilha. O programa mais usado quando escrevemos este conteúdo é o Google Forms, que exporta para uma planilha do google.sheets (que fica online), e cujos dados podem ser facilmente copiados e colados em uma planilha de Excel, o que permite a transferência para o R ou para o outro software com alguns procedimentos manuais.

– Avaliação preliminar dos dados

Depois de devidamente tabulados, os dados precisam ser avaliados em uma primeira exploração, por inspeção visual, com a finalidade de verificar o ordenamento do processo, a existência de dados faltantes, a existência de dados lançados equivocadamente por erros de digitação (por exemplo, é comum erros como lançamento de uma idade de 600 anos quando o interesse era digitar 60 anos), a existência de questionários lançados duas vezes.

Existem algumas técnicas de análise preliminar de dados que somente poderão ser mais bem compreendidas após a apresentação de outros conteúdos. Por esta razão, este tópico (avaliação preliminar) será retomado posteriormente.

- Operacionalização do campo no exemplo ilustrativo

Na pesquisa que desenvolvemos com os estudantes de Administração e Turismo, os dados foram coletados por estudantes do curso de mestrado em Administração que estavam diretamente envolvidos na pesquisa. Por esta razão, não houve necessidade de treinamento, a não ser a discussão e uniformização de entendimentos a respeito do questionário.

Como a forma de coleta foi por procedimento não probabilístico, o que dificulta a aplicação de algumas técnicas de análise, decidiu-se controlar a amostra no sentido de manter a maior heterogeneidade possível das variáveis do tipo categorias, e assim se aproximar da heterogeneidade real do universo de estudantes. Neste sentido, a organização dos dados foi feita paralelamente à coleta, e após algumas avaliações, era definida uma composição específica para outras etapas da coleta.

Os dados foram então organizados no Excel, seguindo as indicações específicas deste software. Após a finalização da tabulação, foram então procedidas todas as análises preliminares, tanto por inspeção visual quanto pela aplicação de algumas ferramentas de verificação.

O Apêndice D apresenta os dados que foram tabulados. Visualmente, já é possível observar que temos vários dados perdidos, mas, aparentemente estes estão dispersos de forma aleatória na massa de dados. O destaque está na observação 105, que apresenta diversas perdas, ou seja, o respondente deixou de responder, sistematicamente, a várias questões.

Concernente à tabulação em si, é possível observar três problemas mais evidentes, que são o respondente 4, que apresenta valor 444 na variável V3, o respondente 34, que apresenta valor 66 na variável V2, e o respondente 111, que apresenta valor 400 na variável idade. Estes casos parecem ser erros de digitação, tendo em vista que, nos dois primeiros, a escala varia de 1 a 7; no terceiro caso, é improvável que o respondente tenha idade de 400 anos. Nestes termos, seria necessário verificar novamente os questionários e fazer as correções; porém, em caso de não haver mais disponibilidade dos instrumentos respondidos, e a melhor opção parece ser a correção direta (ou seja, em lugar de 444 marcar 4, em lugar 66 marcar 6, e em lugar 400 marcar 40).

Chamou a atenção a variável Ren. a qual foi possível observar uma grande discrepância nos respondentes 45 e 79 (40 e 50 salários respectivamente). Como o número de salários não tem limite, não temos como saber se estes valores são oriundos de informação correta ou de erro de digitação, porém parece ser o caso de informação correta.

Nas demais variáveis e entradas de dados não parece haver problemas, mas é necessária uma análise mais cuidadosa por meio de ferramentas específicas para esta etapa de preparação. Isto será apresentado posteriormente, pois requer conhecimentos que serão avaliados na Parte II deste manuscrito.

2.4. Resumo

Este capítulo teve como objetivo apresentar as primeiras etapas do processo de construção de uma pesquisa, com foco nas etapas que antecedem ao processo de análise propriamente dito. Os principais pontos foram os seguintes:

- Uma análise de dados não acontece de forma isolada, sendo parte de um processo mais amplo, inclusive com atividades que a antecedem e outras que a sucedem. Em geral, este processo é sistematizado na forma de uma ação ordenada de pesquisa;
- Uma ação de pesquisa possui em geral cinco etapas fundamentais, a saber: definição do escopo de pesquisa; planejamento; operacionalização de campo; análise de dados; e apresentação de resultados;
- A primeira etapa, relativa à definição de escopo, consiste em definir o problema de pesquisa mais os questionamentos adicionais (ou seja, o que se quer saber), além da formulação dos objetivos de pesquisa, ou seja, o enunciado do que se pretende fazer ao longo da pesquisa;
- A segunda etapa é o planejamento da pesquisa, que consiste em cinco passos, quais sejam: a prospecção exploratória, que uma sondagem preliminar sobre o assunto; a decisão quanto ao tipo de pesquisa (exploratória ou descritiva), para se indicar a profundidade das análises a serem desenvolvidas; a decisão quanto às fontes de dados, ou seja, a indicação de qual lugar, contexto ou objeto dos dados de interesse deveriam ser levantados; as decisões quanto à mensuração, ou seja, a indicação de quais variáveis, de quais escalas e sobre os instrumentos que viabilizarão a coleta; e as decisões quanto à amostragem, que se refere à indicação concernente ao conjunto de objetos a serem coletados, à quantidade e forma de acesso a estes objetos;
- A terceira etapa é a operacionalização do trabalho de campo, que consiste na gestão do trabalho de prospecção de dados propriamente dito, além da organização dos dados, após a coleta ou em paralelo ao processo de coleta em si. Esta etapa termina com a análise preliminar dos dados, na intenção de tornar a planilha pronta para os procedimentos de análise, que vêm na etapa subsequente e que será explorada em outro capítulo.

- Cada uma das etapas de pesquisa apresentadas deste capítulo foi acompanhada de um exemplo concreto, de uma pesquisa sobre percepções e avaliações de estudantes de cursos de Administração e Turismo sobre sua formação e sobre suas profissões.

Exercícios

A seguir, temos um conjunto de problemas que demandam pesquisas. Para cada um deles, trace as definições do conjunto de passos apresentados neste capítulo:

1. Um executivo financeiro de uma faculdade particular de 10 mil alunos deseja construir um plano de redução da inadimplência dos estudantes, que se aproxima de 10% ao mês. Ele precisa entender mais sobre os clientes (os que pagam em dia e os inadimplentes).
2. Um estudante de um curso de Ciências sociais de 300 alunos de uma faculdade pública gostaria de desenvolver um projeto de consultoria para organizações sociais. Ele precisa conhecer a expectativa de público potencialmente atendido, as possibilidades de institucionalização e as possibilidades de articulação com outros estudantes.
3. Um líder social especula a possibilidade de se candidatar a presidente do sindicato de servidores de um município que possui cerca de 500 funcionários de carreira. Ele somente lançará sua candidatura se entender que tem chances, e precisa primeiro analisar suas possibilidades reais.
4. Um pesquisador de comportamento do consumidor acredita que a reação dos consumidores a um determinado tema de propaganda varia de acordo com a classe social e quanto ao gênero das pessoas expostas à peça de propaganda. Ele pretende buscar uma forma para confirmar ou negar sua crença.

CAPÍTULO 3

ANÁLISE DE DADOS E OUTROS FUNDAMENTOS_

Neste capítulo, é dado prosseguimento aos passos de uma pesquisa, porém chegamos ao ponto que mais interessa ao desenvolvimento deste texto, que é a análise de dados. Conforme ilustrou a Figura 2.1, teremos então a análise de dados e em seguida a apresentação da pesquisa. Como no capítulo anterior, a exposição será seguida por um exemplo de pesquisa; porém, aqui não exporemos os detalhes da análise, que pressupõe os conhecimentos que serão desenvolvidos somente na parte II.

Este capítulo encerra a parte preparatória para a apresentação da análise de dados. Por esta razão, além de finalizar o conteúdo relativo às etapas da uma pesquisa, serão apresentados também alguns fundamentos formais e operacionais de ferramentas matemáticas que serão utilizadas nos capítulos seguintes. A teoria matemática da análise exploratória de dados é de nível de ensino médio, com alguns aperfeiçoamentos, de modo que a exposição não terá maiores complicações de compreensão.

Ao final deste capítulo, o leitor terá condições de responder às seguintes questões:

- O que é análise de dados? O que é *statistical learning*? De que forma a análise de dados está relacionada com os conteúdos convencionais de teoria estatística e quando a estatística não se aplica? Quais são as principais formas de análise de dados?
- De que forma se organizam os procedimentos de análise de dados baseados em testes estatísticos? O que diferencia uma análise baseada em testes de uma que os dispensa?
- De que forma se organizam os procedimentos de análise de dados baseados em métodos multivariados?
- De que forma se organizam os procedimentos de análise de dados baseados em técnicas exploratórias? De que forma podemos visualizar uma concepção de análise de exploratórias dados?

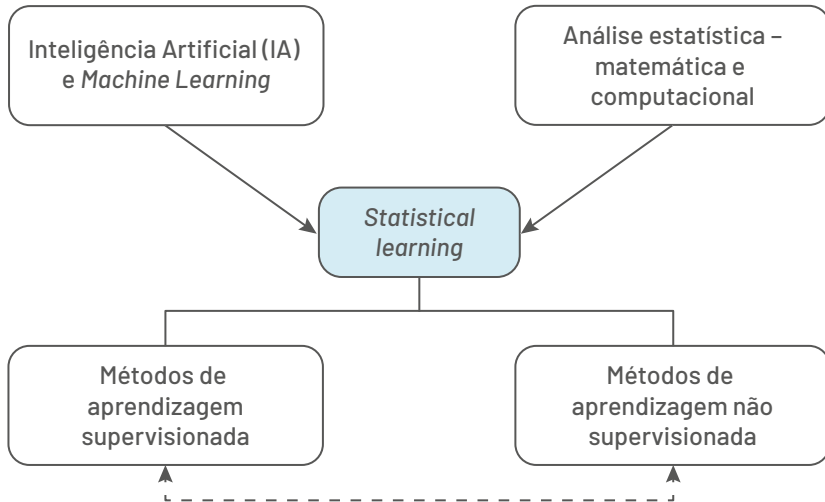
- Quais os principais conhecimentos de matemática necessários à análise exploratória? Como é possível operar e utilizar estes conhecimentos?

3.1. A análise de dados

Entendemos por análise de dados o processo sistemático de avaliação e interpretação de dados de uma pesquisa, por meio de aplicação de técnicas estatísticas direcionadas à exploração, à descrição, e ao teste de hipóteses. Conforme já realçamos nos capítulos anteriores, a análise de dados é parte indispensável do processo de pesquisa, etapa sem a qual não se completa o processo de geração de conhecimento para a finalidade determinada. Naturalmente, a análise mal construída prejudica todo o processo de formação do conhecimento, inviabilizando considerações consistentes sobre o fenômeno e sobre a base de conhecimento disciplinar.

Como vimos, e como indica a ilustração do processo de uma pesquisa, a estatística nem começa nem termina na análise de dados. As etapas anteriores e posteriores são indispensáveis ao processo geral. No entanto, se precisamos de conhecimentos adequados para definição e modelagem de problemas, de métodos consistentes de levantamento de dados nos procedimentos experimentais e de amostragem, e de consistência e validade das atividades de mensuração, tudo isto perde o valor se não tivermos uma aplicação de adequada de métodos de análise de dados.

De ponto de vista de avanços conceituais sobre análise de dados, o paradigma de *statistical learning* é o que tem regido uma reconfiguração quando escrevemos este texto. A *statistical learning* consiste em uma abordagem que integra métodos estatísticos tradicionais e métodos computacionais para análise de dados orientada a compreender e a aprender a partir dos dados. A configuração do paradigma pode ser observado na Figura 3.1.

Figura 3.1 – Configuração do campo de *statistical learning*

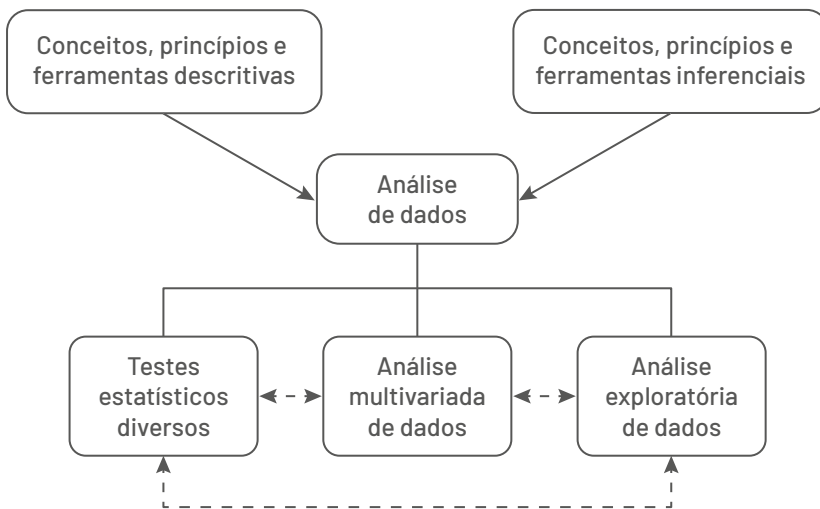
Como pode ser visualizado na Figura 3.1, a organização do conteúdo de análise de dados convencionou-se em **análises supervisionadas** (quando há o pressuposto de uma variável a ser explicada a partir de uma ou mais variáveis explicativas) e **não supervisionadas** (quando não há pressuposição de uma variável a ser explicada). As análises não supervisionadas e supervisionadas podem envolver conteúdos de estatística descritiva e inferencial.

A divisão de conteúdos da estatística advém do entendimento de que uma parte dos procedimentos estatísticos, os descritivos, se direciona somente para descrever dados, sem preocupação com a relação entre o conjunto de dados e a população da qual os dados foram extraídos (supondo-se que os dados são de uma amostra; obviamente, se estamos trabalhando com dados de todo o universo de pesquisas, somente as técnicas descritivas têm sentido, pois não há uma população para a qual se deseja inferir). Mas além de somente descrever dados, há em estatística um conjunto de ferramentas cuja finalidade é justamente viabilizar a avaliação da população de interesse a partir do que se observa na amostra.

Em geral, a visualização da análise de dados segue esta ‘divisão’ da estatística. No entanto, e considerando a realidade observada convencionalmente nos cursos e aplicações em Ciências Sociais e Comportamentais, preferimos definir outra visualização, uma que está baseada nas práticas de formação de cursos superiores.

Assim, considerando as práticas de organização do conhecimento estatístico em livros e disciplinas (assim definidos por seu encadeamento lógico e pela viabilidade de aprendizado), visualizamos três alternativas de análise, que são: a análise exploratória de dados, a análise por meio de testes estatísticos, e a análise multivariada. Naturalmente, esta divisão é uma convenção para efeitos didáticos, pois as três formas de análise se complementam. O leitor deve observar que nossa concepção é direcionada à análise de dados, não havendo proposta de alteração da divisão clássica da estatística entre descritiva e inferencial. O que propomos é não mais que uma visão para a análise de dados, que tem base justamente nos conhecimentos destas duas divisões clássicas. A Figura 3.2 ilustra esta visualização.

Figura 3.2 – Análise de dados



Naturalmente, as abordagens anteriormente apresentadas, associadas à análise de variáveis de forma univariada, bivariada e multivariada, possuem uma associação direta com estas alternativas de análise de dados aqui concebidas. De fato, com exceção da análise multivariada, temos procedimentos de análise exploratória e de testes estatísticos para cada uma destas três possibilidades de abordagens. Cada uma das partes indicadas está descrita a seguir.

3.1.1. Análise por testes estatísticos

A análise estatística por meio de testes é aquela realizada com a utilização das ferramentas da estatística inferencial (ou seja, da parte da Estatística que se preocupa em analisar os dados de uma amostra e em seguida inferir potenciais resultados que caracterizam a população da qual a amostra foi extraída), com a finalidade de analisar consistência de relações entre variáveis, para testar regularidades nos dados, e para testar possibilidade de generalização de resultados da amostra para a população. Para tanto, esta forma de análise se utiliza também de ferramentas de estatística descritiva.

Por exemplo, quando vamos avaliar o percentual de produtos com defeito em uma linha de produção de cadeiras, por hipótese temos uma população de tamanho infinito, o que requer que a análise seja feita com base em uma amostra. Assim, considerando a intenção de conhecer o percentual de cadeiras que requererão reparos antes de seguir para o canal de vendas, somente será possível termos uma estimativa, que é aquela associada ao percentual de cadeiras com defeito na amostra (uma vez que não temos como acessar todo o universo). Assim, se em amostra de 180 cadeiras observamos 9 com defeitos, então estimamos que, do total de cadeiras produzidas, aproximadamente 5% precisarão de reparos.

A análise estatística por meio de testes tem dois pré-requisitos de conhecimento centrais e associados, que são a Teoria das probabilidades e a Teoria da inferência. A primeira, a **teoria das probabilidades**, é um ramo de conhecimento altamente elaborado do ponto de vista conceitual e operacional, sendo uma especialidade considerada das mais difíceis do ponto de vista matemático. Sua finalidade é estudar os fenômenos associados ao acaso, ou seja, procura facilitar a modelagem de problemas que não têm conteúdo determinístico (fenômenos determinísticos são aqueles em que se assegura certeza de previsão, como acontece em alguns fenômenos físicos ou químicos).

Por seu objeto (ocorrências de fenômenos não determinísticos), a Teoria das probabilidades tem aplicações em jogos de sorte/azar (como as loterias, por exemplo), em situações de risco (como aqueles associados a acidentes, mortes etc., de interesse da área de Ciências atuariais), e em situações diversas das Ciências sociais, como os as pesquisas nos certames eleitorais (em que não se sabe o resultado antes da eleição, mas se procura ‘estimar’ o que deverá ocorrer), a realização de testes de conhecimento (em que há possibilidade de

o sujeito ‘chutar’ uma resposta), nas análises de decisão de consumidores ou de investidores, dentre outros.

Seguramente, os principais desdobramentos da Teoria das probabilidades vêm do estudo das funções matemáticas que associam eventos específicos de um universo de possibilidades a valores de probabilidades, no que se conhece por **variáveis aleatórias**. Por exemplo, no lançamento de uma moeda, temos duas possibilidades (cara e coroa), às quais podemos associar um valor de probabilidade entre zero e um (se não houver vício no dado, associamos 0,5 a cada, pois são igualmente prováveis).

É por meio do estudo das variáveis aleatórias que se pode compreender as regularidades de determinados fenômenos. Assim, mesmo que os eventos associados sejam não determinísticos, a ocorrência de um grande volume de repetições de um evento pode providenciar uma configuração que viabiliza a melhor compreensão do fenômeno. Por exemplo, ao se lançar um dado, não sabemos qual é o número surgirá na face superior, mas sabemos que depois de um número elevado de lançamentos, aproximadamente 16,6% do total de lançamentos deverá ser de face 1. Esta informação, para o caso, de jogos, permite prever ganhos e perdas envolvidos.

No entanto, seguramente a principal aplicação da teoria das probabilidades é justamente na modelagem dos fenômenos observados em uma amostra na tentativa de assegurar condições para o entendimento da população. Este é o caso da **Teoria inferencial**, que, na verdade, é fundamentada na Teoria das probabilidades, porém tem seu próprio corpo de conceitos, métodos e ferramentas. Vejamos dois exemplos:

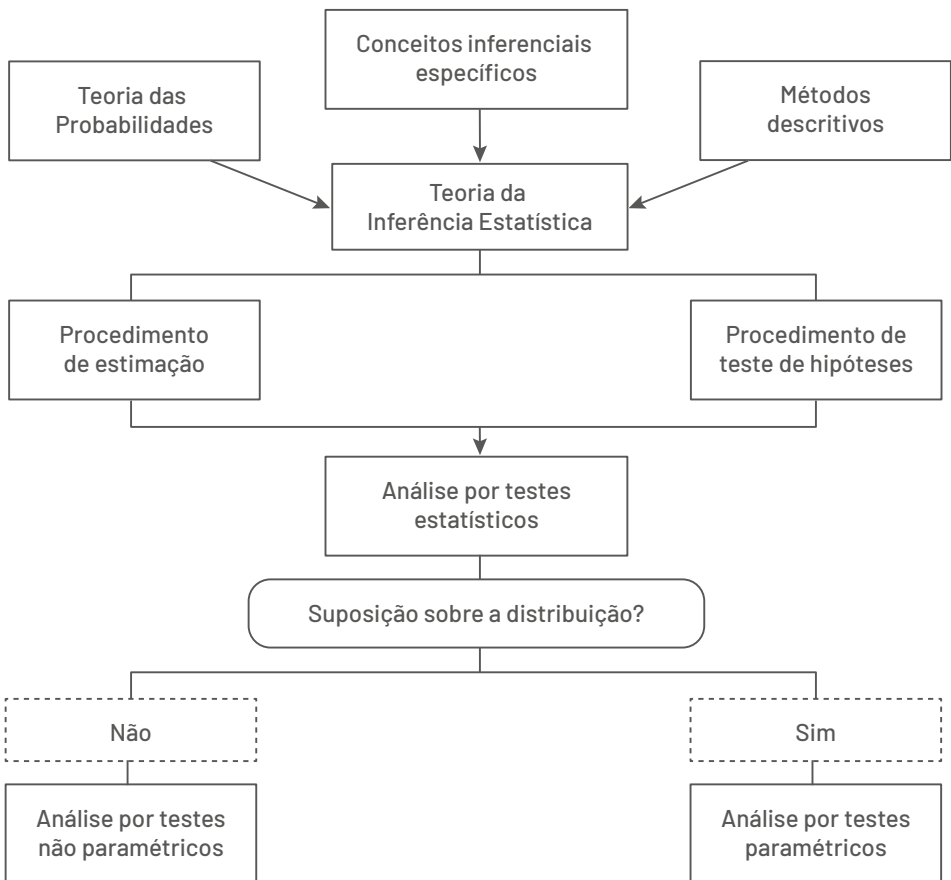
- Supondo uma população de 2 milhões de eleitores a partir da qual foi extraída uma amostra de 1000 sujeitos. Se verificarmos que um percentual de 45% de intenções de votos é declarado a um candidato A, e que 30% para um candidato B, temos ferramentas de estatística inferencial para indicar o quanto estes percentuais podem ser base de previsão de resultado real nesta população;
- Supondo que uma amostra de 500 clientes de um hotel é convidada a avaliar, separadamente, os serviços de cozinha e de bar, em uma escala de 0 a 10 pontos. Após a consolidação das notas, sendo observado que a nota de 5,5 para o bar e de 5,9 para o restaurante, temos ferramentas para avaliar em que medidas estas notas são realmente distintas (e a nota do bar é menor

que a do restaurante) ou não há diferença significativa e esta diferença observada de notas pode ser atribuída a fatores aleatórios.

Nestes casos, estas ferramentas citadas são na verdade problemas de **estimação** (em que queremos estimar alguma característica da população a partir da amostra) e de **testes de hipóteses**, uma vez que podemos adotar por hipótese um determinado entendimento, e em seguida testamos a veracidade desta hipótese. Observe que a Teoria inferencial pressupõe, além dos conhecimentos de probabilidades, seus próprios métodos, além de levar em conta os métodos de estatística descritiva.

Para qualquer dos dois casos, nos baseamos nas características da amostra e em pressupostos relacionados à variável aleatória (que, supostamente, caracteriza a variável sob análise), e em seguida fazemos uma afirmação. A Figura 3.3 ilustra a concepção por testes estatísticos.

Figura 3.3 – Análise por testes estatísticos



Na grande literatura estatística, a pressuposição definiu dois grandes vetores de desenvolvimento de técnicas e testes.

- Assim, quando é suposto que a variável sob análise é oriunda de uma variável aleatória com uma caracterização pré-estabelecida, sendo mais comum variável aleatória com caracterização especial dita ‘normal’ (nos capítulos seguintes teremos uma configuração do que seja normalidade, a partir de procedimentos descritivos), então temos os chamados **métodos paramétricos**;
- Nos casos em que não temos referenciais pré-estabelecidos para a variável aleatória subjacente à variável que foi amostrada, então temos os chamados **métodos não paramétricos**.

Os métodos mais conhecidos são os paramétricos, porém há uma séria crítica a estes métodos, pois as suposições sobre a caracterização da variável nem sempre são verdadeiras. Por outro lado, as fragilidades dos métodos paramétricos contribuíram para o crescimento de interesse pelos métodos não paramétricos. Inclusive, alguns dos principais desenvolvimentos do que chamamos de métodos de *statistical learning* são na verdade de base não paramétrica. Ao longo dos capítulos seguintes, teremos maior clareza sobre as diferenças indicadas e os impactos sobre a análise de dados.

Para finalizar esta primeira e (bastante) parcial exposição sobre os testes estatísticos, convém informar que alguns testes são, eventualmente, suficientes a uma determinada pesquisa (nos dois exemplos destacados acima isto fica mais claro), não havendo mais outras demandas de técnicas. Por outro lado, alguns testes foram desenvolvidos para serem utilizados nas ferramentas de análise exploratória e (principalmente) de análise multivariada. Especificamente na análise multivariada, é inclusive possível que uma só ferramenta envolva vários testes. Isto reitera a interdependência entre as alternativas de análise, que estão indicadas nas setas de duas pontas da Figura 3.2.

- Análise por meio de testes no exemplo ilustrativo

Em nossa pesquisa sobre percepções e avaliações dos estudantes de administração e turismo, praticamente não foram aplicadas ferramentas de testes como ferramenta restrita. Foi aplicado somente o teste de comparação de medidas das variáveis quantitativas entre os dois grupos de estudantes (em uma técnica chamada análise de variância), porém a finalidade desta aplicação foi secundária em relação aos objetivos da pesquisa. Outros testes usados foram aqueles envolvidos nas análises multivariadas que buscavam analisar as relações entre os constructos envolvidos.

3.1.2. Análise multivariada de dados

Conforme já apontado anteriormente, a análise multivariada é o tipo de análise que utiliza conteúdos teóricos e aplicados de Estatística para avaliar conjuntamente conjuntos de mais de duas variáveis (para saber mais acesse: Hair Jr et al., 2005; Lattin et al., 2011; Manly, 2008; Ferreira, 2008; Mingoti, 2005). Os dois exemplos a seguir possibilitam uma visão da aplicação deste tipo de análise:

- Quando avaliamos a disposição dos clientes de hotéis a indicarem o prestador de serviços a outras pessoas (INT), podemos avaliar se esta disposição recebe influência, simultaneamente, da satisfação com os serviços (SAT) e de sua percepção de valor (VAL) (simbolicamente: $SAT+VAL \rightarrow INT$). Por outro lado, o valor percebido é influenciado, simultaneamente, pela qualidade do atendimento (QUAT) e pela qualidade das instalações (QUAI) (temos: $QUAT+QUAI \rightarrow VAL$). Temos ao todo, um total de cinco variáveis analisadas simultaneamente
- Algumas práticas de segmentação de mercado (ou seja, reunião de conjunto de consumidores em grupos, para viabilizar uma oferta mais bem direcionada) costumam tomar por base uma só variável, normalmente demográfica, como, por exemplo, faixa de idade, faixa de renda etc. Por outro lado, é fácil entender que a verificação de diversas variáveis ao mesmo tempo parece ser mais eficiente, como seria o caso, por exemplo, de uma segmentação que levasse em conta variáveis como faixa de idade, faixa de renda, opção sexual, opção religioso e frequência de compra.

O conteúdo da disciplina de ‘Estatística multivariada’ foi, ao longo dos anos, formatando-se em torno de um conjunto específico de princípios, conceitos,

relações, e principalmente, de técnicas de análise. Evidentemente, a análise multivariada é uma extensão das técnicas de análise univariada e bivariada; no entanto, se nestes dois casos (análise univariada e bivariada) sempre foi mais fácil manusear os dados, para o caso das técnicas de análise multivariada, a situação foi diferente, em decorrência das dificuldades de operar, manualmente, conjuntos de dados e variáveis.

Além disto, sabemos que, no desenvolvimento teórico para as técnicas univariadas e bivariadas, o ferramental matemático é muito mais simples que o ferramental necessário para o desenvolvimento e a análise de métodos com muitas variáveis, que requer, sempre, suporte de teoria matemática de Álgebra linear (ou matricial) e de Cálculo numérico. Em ambos os casos (Álgebra matricial e Cálculo numérico), resolução de problemas práticos não é viável sem o suporte de instrumentos computacionais.

Por esta razão, entendemos que foi o desenvolvimento da tecnologia da informação que viabilizou um aumento do interesse pelas técnicas multivariadas de análise. Atualmente, os softwares estatísticos, como o JASP e o R, trazem um razoável número de técnicas multivariadas prontas para serem aplicadas pelo esforço de alguns comandos no computador. Adicionalmente, dada a forma de desenvolvimento do software R (em que pesquisadores desenvolvem rotinas e as disponibilizam na internet, cabendo apenas alguns procedimentos de instalação para quem possui o software em seu computador), as principais novidades que vem sendo desenvolvidas em todo o mundo são rapidamente disponibilizadas.

Além das sofisticadas técnicas matemática, a análise multivariada tem na teoria inferencial uma base de sustentação indispensável. Mesmo que haja técnicas multivariadas que não utilizam testes, grande parte daquelas técnicas mais aplicadas em Administração, Educação ou Psicologia, por exemplo, somente se desenvolveu a partir dos testes. Na maioria das aplicações atuais, os programas de computadores já exibem nos resultados todos os testes envolvidos.

Em síntese, podemos dizer que, atualmente, as técnicas multivariadas são já parte corrente dos processos de análise de dados. Como bem sabemos, o desenvolvimento de procedimentos de análise restrito ao manuseio de softwares não desenvolve efetivamente a competência estatística do usuário. De fato, e como indicamos no Capítulo 1, para usuários de Estatística aplicada,

além do domínio de instrumentos computacionais, é necessário também o conhecimento de base da teoria estatística subjacente, obviamente em menor grau que o sujeito envolvido com Estatística teórica. Por esta razão, foram publicados vários livros sobre o assunto com uma visão mais explicativa, e menos orientada às formulações de Estatística teórica e Matemática⁷.

Cabe observar que alguns autores preferem utilizar a expressão '**Análise multivariada**' para o uso recorrente que se faz, em oposição ao conteúdo estatístico e matemático, chamado de '**Estatística multivariada**', e que é a disciplina teórica que fundamenta e fornece as técnicas de análise. Embora não seja uma regra de uso, acreditamos que esta denominação tem a vantagem de indicar as especificidades e especialidades de desenvolvimento teórico e de sua aplicação.

A visão convencional da análise multivariada a dimensionou em duas correntes centrais: as técnicas de análise de dependência, e as técnicas com interdependência. As **técnicas de análise de dependência**, como o nome sugere, são aquelas que têm por finalidade analisar conjuntos de variáveis em que se supõe que uma (ou mais) é (são) influenciada(s) por outra(s) variável(eis). O exemplo sobre clientes de hotel, apresentado no início do item, ilustra este tipo de análise.

A mais conhecida é análise de regressão múltipla, em que, por hipótese, uma determinada variável é influenciada por um conjunto de pelo menos duas outras variáveis (a primeira é a dependente, e as demais são as independentes). A análise de regressão é provavelmente a técnica mais estudada em estatística aplicada, e possui um número bastante grande de modelos, ramificações e principalmente de aplicações.

No entanto, a análise de regressão tem a limitação de supor uma variável como predita por outras, apenas. Porém bem sabemos que na realidade corrente, uma variável é ao mesmo tempo influenciada por um conjunto de variáveis

7 O principal texto de análise multivariada nesta orientação, em língua portuguesa, é o seguinte. HAIR, J. F. Jr.; BLACK, W. C.; BABIN, B. J.; ANDERSON, R. E. *Análise multivariada de dados*. Porto Alegre: Bookman, 2005. Outros semelhantes são: LATTIN, J.; CARROLL, J. D.; GREEN, P. E. *Análise de dados multivariados*. São Paulo: Cengage, 2011; MANLY, B. J. F. *Métodos estatísticos multivariados: uma introdução*. Porto Alegre: Bookmann, 2008. De uma orientação mais matemática e estatística (teórica), temos os seguintes: FERREIRA, D. F. *Estatística multivariada*. Lavras: Ed. UFLA, 2008; MINGOTI, S. A. *Análise de dados através de métodos de estatística multivariada: uma abordagem aplicada*. Belo Horizonte: Editora UFMG, 2005. Além destes, temos os textos com foco em algumas técnicas, como são aqueles de Econometria, por exemplo.

e também exerce influência sobre outras variáveis. Por esta razão, foram desenvolvidos os métodos multivariados de modelos de equações simultâneas (principalmente em Econometria) e de modelagem de equações estruturais (com aplicações em áreas que utilizam construtos latentes mensurados por mais de uma variável).

É possível dizer que estas são as principais técnicas de análise de dependência, porém a literatura e o uso corrente trazem ainda outras técnicas relevantes, como a análise de correlação canônica, a análise multivariada de variância, ou a análise discriminante (o espaço disponível não permite maiores explicações sobre estas técnicas, porém as indicações da bibliografia dão bons caminhos).

Sobre as **técnicas de análise de interdependência**, estas são entendidas como aquelas em que não se supõe que haja entre as variáveis envolvidas na análise qualquer dependência ou independência. O exemplo sobre segmentação de mercado, o segundo apresentado no início deste item, ilustra este tipo de análise.

Dentre estas, as mais conhecidas e mais utilizadas são: a análise fatorial (que avalia um conjunto de variáveis e propõe a reunião deste conjunto em fatores que podem as representar; por esta razão, a análise fatorial é também entendida como uma técnica de redução de variáveis); e a análise de agrupamentos ou de *cluster*, uma técnica que, a partir do conjunto de variáveis, define uma regra de proximidade entre sujeitos avaliados e em seguida propõe uma reunião dos sujeitos mais próximos em grupos.

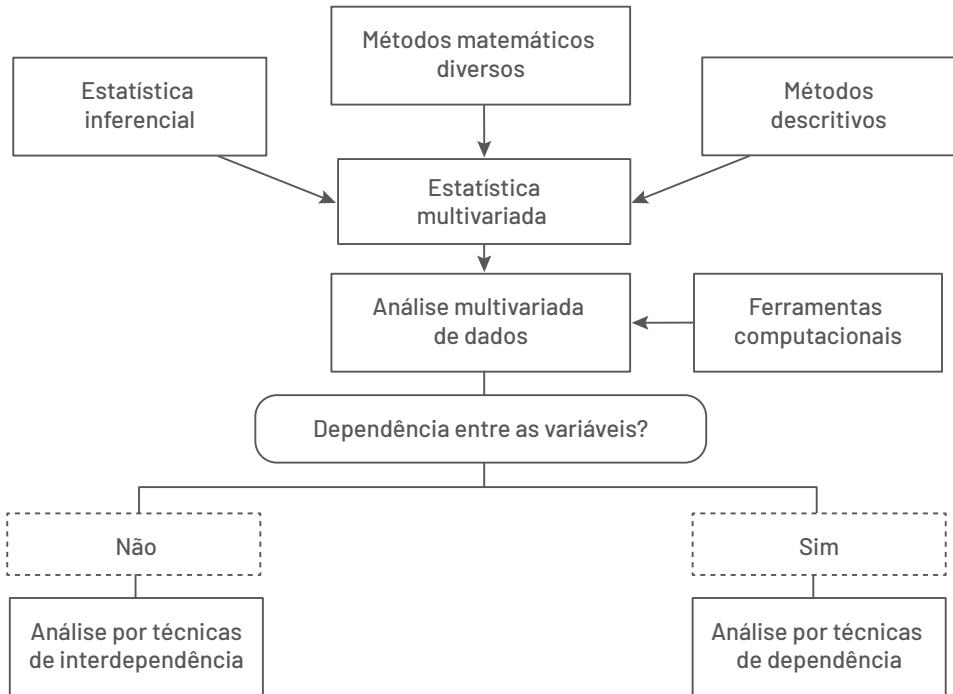
Em razão da fundamentação na teoria inferencial, as técnicas multivariadas também são, parte delas, desenvolvidas sob algumas hipóteses ou suposições. Assim, a técnica de análise de regressão clássica, por exemplo, adota como pressuposto que o erro do modelo tem a característica de uma variável aleatória definida por ‘normal’. Em grande parte das vezes, esta suposição não se verifica, o que se torna um fator de dificuldades para a aplicação desta ferramenta (e de todas as demais que são desenvolvidas de forma associada) para a análise de dados.

As suposições na verdade dependem das técnicas, de modo que cada técnica tem suas próprias suposições e, inclusive, alternativas para correção de situações em que as suposições não se verificam. Diferente do caso dos testes, não temos um uso recorrente de denominações como ‘análise multivariada

paramétrica’ e uma ‘não paramétrica’, embora tenhamos, no nível de técnicas, a possibilidade de se usar esta denominação.

Por exemplo, no caso da regressão, como temos citado, temos, de fato, a regressão paramétrica e a não paramétrica, mas o mesmo não ocorre, por exemplo, com a análise de agrupamentos. A Figura 3.4 apresenta uma ilustração visual do que comentamos neste item.

Figura 3.4 – Análise multivariada de dados



- Análise multivariada no exemplo ilustrativo

Pela finalidade da pesquisa, foram usadas técnicas tanto de dependência quanto de interdependência. No caso da técnica de dependência, utilizamos a análise de regressão para testar se as dimensões de valor percebido pelos estudantes eram de fato influenciadoras da percepção geral de valor. Adicionalmente, utilizamos análise de cluster para agrupar os pesquisados em três blocos, com avaliação a partir do conjunto de variáveis associadas a valor, mais as variáveis de reputação e identificação profissional.

3.1.3. Etapa de análise exploratória de dados – AED

A análise exploratória de dados envolve todo o conjunto de ferramentas de exploração dos dados para decisão ou conhecimento. Em geral, AED não pressupõe a aplicação de testes para confirmar ou refutar hipóteses previamente elaboradas, embora não seja ‘proibida’ sua aplicação.

Para alcançar sua finalidade, a análise exploratória tem primeiramente à disposição todo o conjunto de conceitos, ferramentas e técnicas de análise da **análise descritiva**. Por exemplo, se temos um conjunto de 300 clientes e queremos saber quantos deles estão vinculados a uma determinada classe econômica, basta calcular o percentual de pessoas, deste conjunto de 300, que está com renda familiar dentro de uma faixa de classificação previamente determinada. Este percentual permite uma visão agregada e exploratória do conjunto de pessoas, em especial se acreditamos que haja pessoas de diferentes faixas. Vejamos outro exemplo: imaginemos que uma organização tem um conjunto de funcionários com diferentes idades e deseja desenvolver uma política de formação continuada por idade. Ora, é muito provável que em 30 funcionários tenhamos diversas idades distintas; no entanto, se nosso foco é o conjunto de funcionários, então, em lugar de tomarmos referência de cada pessoa, seria preferível tomarmos a média de idade deste conjunto de servidores. A média é, portanto, uma medida exploratória que representa e sintetiza o conjunto de dados indicadores das idades dos servidores.

Adicionalmente, a AED está preocupada em fornecer *insights* sobre os dados, especialmente se pouco sabemos a seu respeito. Assim, um conjunto de ferramentas de AED eficiente para esta finalidade consiste na **análise gráfica** de conjuntos de dados. As ferramentas gráficas, além de servirem para visualização da estrutura e da organização dos dados, permitem ainda uma primeira visão sobre o formato da distribuição dos dados, ou seja, é possível perceber se há alguma regularidade ou aderência da forma de distribuição dos dados levantados em associação com outros formatos de distribuição previamente conhecidos.

Temos ainda a **análise de associação** entre variáveis, que se faz por meio de técnicas específicas para os diferentes tipos de dados e variáveis (análise de associação também pode ser feita por meio de testes, mas, na esfera exploratória, o uso de testes é raro). As finalidades possíveis são duas: (1) simplesmente explorar possibilidade de relações, para verificar se emergem dos dados poten-

ciais hipóteses a serem testadas posteriormente; ou (2) visualizar previamente as indicações de consistência de hipóteses anteriormente lançadas, porém sem a finalidade de indicação de sua confirmação ou refutação.

Em nível de finalidade, além das que estão indicadas logo acima, temos também a preocupação em AED de verificar se as condições para aplicação de técnicas inferenciais de testes e análise multivariada se confirmam em um conjunto de dados. Por exemplo, indicamos anteriormente que a técnica de análise de regressão pressupõe que as variáveis envolvidas seguem a distribuição normal. Obviamente, se isto não se verifica, as técnicas de análise de regressão gerarão resultados não confiáveis, de modo que se faz necessário verificar antecipadamente se estas condições se verificam. Para este tipo de procedimento, na análise preliminar de suposições como estas, é comum a AED utilizar testes estatísticos.

Mais que simplesmente indicar a adequação dos dados a uma determinada ferramenta estatística, a avaliação exaustiva de conjuntos de dados por AED permite inclusive a indicação de possíveis técnicas a serem aplicadas. De fato, se em uma análise de previsão (que é normalmente realizada por métodos inferenciais) verificamos pelos procedimentos de AED que a análise de regressão do modelo não normal não se aplica, então é possível que esta mesma análise sinalize quais outras ferramentas de previsão seriam aplicáveis, considerando a estrutura e a organização dos dados.

A AED pode sinalizar ainda a necessidade de mais dados, ou de uma reorganização dos dados disponíveis. Por exemplo, se pretendemos desenvolver uma análise comparativa entre homens e mulheres sobre seus hábitos de investimento de recursos financeiros, e em uma primeira coleta de dados descobrimos que temos 90% do total de homens e o restante de mulheres. Neste caso, salvo uma situação em que a amostra for realmente muito grande, não teremos condições efetivas de informar resultados comparativos simplesmente porque temos uma grande discrepância entre os números de mulheres e homens.

Em síntese, podemos definir como objetivos desta etapa da análise de dados os seguintes:

- Apresentar uma descrição sintética do conjunto de dados, por meio de técnicas de estatística descritiva;
- Apresentar uma organização tabular e visual do conjunto de dados, por meio de técnicas gráficas;

- Analisar relações entre diferentes variáveis, por meio de técnicas de análise associativa;
- Fornecer indicações e *insights* sobre estrutura de dados e relações possíveis entre variáveis, sugerindo potenciais hipóteses para testes posteriores;
- Verificar pressupostos para as etapas posteriores de análise inferencial de testes de hipóteses e aplicações de ferramentas multivariadas;
- Indicar potenciais alternativas de ferramentas de análise de análise complementar;
- Apresentar potenciais demandas de ajustes nos dados ou de novos procedimentos de coleta de dados.

Do ponto de vista do ordenamento da análise de dados, a AED é uma etapa preliminar em relação às análises por testes e multivariada, mas isto não implica uma importância menor. Até pelo contrário, pois não teria muito sentido irmos diretamente a essas análises sem antes ‘explorar’ os dados disponíveis. Além disto, e a depender da finalidade da análise, a etapa exploratória pode já ser suficiente, não havendo a necessidade de procedimentos adicionais de análise.

- AED no exemplo ilustrativo

Na pesquisa desenvolvida, foram usadas praticamente todas as técnicas de análise exploratória disponíveis, desde o primeiro momento de limpeza dos dados na planilha, até a verificação dos requisitos de utilização da técnica de análise de regressão. Também foram usadas técnicas de estatística descritiva, em todas as variáveis, de tal modo que naquelas da natureza categórica foram utilizadas técnicas de descrição de percentuais, e naquelas de natureza quantitativa foram utilizadas medidas descritivas apropriadas. Foram utilizadas ferramentas gráficas, porém em menor intensidade.

3.1.4. Uma concepção para a AED

Como a finalidade aqui é discorrer sobre análise exploratória, e considerando a necessidade de um ordenamento do que será exposto, seguimos o ordenamento temático indicado no Quadro 3.1.

Quadro 3.1 – Temas de análise exploratória de dados

Estatística descritiva (Parte 2)
Envolve o conteúdo de análise relacionado a: <ul style="list-style-type: none">Exposição de dados, que envolve a representação tabular e gráfica de dadosMedidas estatísticas de localização, incluindo as medidas clássicas e as modernasMedidas estatísticas de escala, incluindo também as medidas clássicas e as modernasMedidas estatísticas de formato, envolvendo as medidas de assimetria e curtose
Análise bivariada (parte 3)
Envolve o conteúdo de análise de associação entre: <ul style="list-style-type: none">Duas variáveis quantitativas, incluindo análise de correlação e de regressão bivariadaDuas variáveis qualitativas, incluindo tabelas cruzadas e os testes de associação e contingênciaUma variável quantitativa e uma categórica, em uma introdução geral à análise de variância

Este ordenamento segue próximo do que entendemos ser um caminho mais eficiente. Uma abordagem nesta ordem seguramente viabiliza uma visão preliminar bastante completa sobre os dados de uma pesquisa. No entanto, este ordenamento não é rígido, e as etapas não são pré-requisitos uma da outra. Cada atividade de pesquisa indicará a necessidade de quais conteúdos.

3.2. Apresentação de resultados de pesquisa

A etapa final da pesquisa consiste na apresentação de seus resultados, incluindo desde a descrição do escopo até os resultados do trabalho, juntamente com comentários e análises. Esta exposição pode alcançar diferentes níveis de formalidade, que pode ir desde um relato oral, que se aplica a pesquisas exploratórias de realização mais rápida, até a escrita de manuscrito completo, inclusive na forma de livro ou de trabalhos finais de cursos de mestrado e principalmente doutorado. Em uma tentativa de organização das possibilidades, temos o seguinte:

- Relatório escrito estritamente descritivo: relato impresso ou em arquivo de computador direcionado a subsidiar decisões organizacionais, sem preocupação com revisão teórica. A extensão do relatório depende da profundidade da pesquisa, porém, por demandar tempo dos decisores, não pode ser muito extensa a ponto de requerer muito tempo de leitura;
- Relatório detalhado com revisão teórica: consiste na apresentação da pesquisa justamente com todo o detalhamento de sua construção, e com a análise dos dados fundamenta em uma revisão teórica.
 - Quando são apresentados em uma versão completa e com todos os detalhes, constituem os relatórios de pesquisa monográfica, como dissertações e teses, que podem chegar a centenas de páginas;
 - Quando são apresentados em versão reduzida em termos de detalhamento, constituem os artigos científicos, que não costumam ser de menos de 5 nem de mais de 30 páginas.
- Apresentação: consiste na apresentação dos pontos principais do relatório em lâminas de apresentação. A principal forma hoje é aquela feita em *slides* de computador (na maioria das vezes em número pequeno de *slides*), podendo também serem construídos pôsteres (um só pôster para uma pesquisa). Esta forma pode ser a única alternativa de exibição dos resultados, se a pesquisa não requerer um texto adicional, ou pode ser complemento dos dois demais tipos acima apresentados. Na maioria das vezes, a apresentação requer os slides escritos mais a apresentação oral (o que não é o caso dos dois tipos anteriores).

Cada situação de pesquisa indicará as especificidades da apresentação a ser feita. Assim, para o caso dos relatórios gerenciais há liberdade de decisão para o pesquisador, salvo definições da própria organização. Já nos relatórios científicos, que envolvem revisão teórica, há regras de apresentação (como as regras da Associação Brasileira de Normas Técnicas – ABNT) que as instituições de pesquisa costumam adotar. Sobre as apresentações, também aqui há uma grande flexibilidade e abertura para inovações.

- Apresentação de resultados no exemplo ilustrativo

A pesquisa realizada tinha por finalidade ser uma pesquisa científica, de modo que a apresentação do final foi feita na forma de artigos científicos. Por esta razão, os detalhes de formato eram definidos ao mesmo tempo pelas normas da ABNT e por outras definições específicas dos contextos nos quais tentamos a publicação. Também foram feitas apresentações com slides do programa PowerPoint, para complementar as exposições orais em eventos científicos em que o artigo foi aceito para apresentação.

3.3. Fundamentos matemáticos

Conforme apresentado na abertura do capítulo, este item tem por finalidade elucidar os principais fundamentos matemáticos que são usados em análise exploratória de dados e nas análises subsequentes. Os conteúdos principais são dois, a saber: operações de adição e produto de conjuntos de dados; logaritmos. O primeiro conteúdo é mais extensamente apresentado, por seu maior uso.

3.3.1. Soma e produto de conjuntos de dados

Na representação de uma dada variável, é comum em Estatística a utilização de uma denominação de variáveis que simplifique a notação e o manuseio. Desta forma, se temos, por exemplo, a variável peso, com medidas em quilogramas, e temos um total de 10 indivíduos que tomamos seus pesos, denominaremos a variável por uma letra maiúscula, normalmente X , Y , dentre outras. Cada observação será representada por uma letra minúscula indexada, em que o índice remeterá ao indivíduo. Assim, teremos $x_1, x_2, x_3, \dots, x_{10}$ ou $y_1, y_2, y_3, \dots, y_{10}$. Isto requer, naturalmente, que cada indivíduo seja previamente associado a um número, que irá ao índice das observações da variável.

No manuseio conjunto mais de uma variável, é comum utilizarmos uma mesma letra maiúscula, também indexada, porém aqui o índice representa cada variável. Assim, variáveis como peso, altura e idade de indivíduos podem ser representados por x_1 (peso), x_2 (altura) e x_3 (idade). Se tivermos os mesmos 10 indivíduos, estes serão indicados como um segundo número no índice. Assim, teremos os indivíduos da primeira variável representados assim os pesos $x_{1,1}, x_{1,2}, x_{1,3}, \dots, x_{1,10}$, e nas alturas e nas idades assim: $x_{2,1}, x_{2,2}, x_{2,3}, \dots, x_{2,10}$ e $x_{3,1}, x_{3,2}, x_{3,3}, \dots, x_{3,10}$.

Na manipulação destas representações também usamos símbolos específicos, sendo os mais comuns os de somatório e o de produtório. O primeiro é representado pela letra grega ‘sigma’ maiúscula, associada aos elementos de uma dada variável. Assim, se temos uma variável X composta pelas observações $x_1, x_2, x_3, \dots, x_n$, teremos a soma destas dos elementos assim representada:

$$\sum_{i=1}^n x_i = x_1 + x_2 + x_3 + \dots + x_n$$

- Lê-se desta forma: o somatório de x_i , com i variando de 1 até n (os pontos de partida são convencionais, ou seja, i poderia variar de 0 a 1, de $(n-2)$ a $(n+6)$ etc.). A variação do i segue sempre o ordenamento crescente de números inteiros.

Já o produtório é representado pela letra ‘pi’ maiúscula. No caso da variável X acima citada, a representação do produto dos seus elementos é feita assim:

$$\prod_{i=1}^n x_i = x_1 \cdot x_2 \cdot x_3 \cdot \dots \cdot x_n$$

- Lê-se desta forma: o produtório de x_i , com i variando de 1 até n . A regra de variação do i é a mesma que se usa no somatório.

Tanto o somatório quanto o produtório são muito usados em Estatística. No entanto, em procedimentos exploratórios, o operador mais usado é o somatório (o operador de produtório é amplamente usado em inferência estatística). Tendo em vista o uso que faremos daqui em diante, convém apresentar algumas propriedades deste operador, a seguir indicadas.

P1 – Somatório da constante: o somatório de uma constante, com índice variando de m a n , é dado por $(n-m+1)$ vezes a constante, ou seja,

$$\sum_{i=m}^n a = (n - m + 1) \cdot a$$

De fato, temos que:

$$\sum_{i=m}^n a = \underbrace{a + a + a \dots + a}_{(n-m+1) \text{ vezes}} = (n - m + 1)a$$

- Exemplo:

$$\sum_{i=1}^{10} 3 = \underbrace{3 + 3 + 3 \dots + 3}_{(10-1+1) \text{ vezes}} = (10 - 1 + 1) \cdot 3 = 30$$

P2 – Somatório do produto por constante: o somatório do produto dos elementos de uma variável por uma constante é o produto da constante pelo somatório dos elementos da variável, ou seja:

$$\sum_{i=m}^n ax_i = a \sum_{i=m}^n x_i$$

Com efeito, é verdade que

$$\begin{aligned} \sum_{i=m}^n ax_i &= ax_1 + ax_2 + ax_3 + \dots + ax_n \\ &= a(x_1 + x_2 + x_3 + \dots + x_n) \\ &= a \sum_{i=m}^n x_i \end{aligned}$$

- Exemplo: Se $X = \{x_2 = 10, x_3 = 15, x_4 = 40, x_5 = 35\}$, teremos que:

$$\gg \sum_{i=2}^5 6x_i = 6 \cdot 10 + 6 \cdot 15 + 6 \cdot 40 + 6 \cdot 35 = 60 + 90 + 240 + 210 = 600$$

$$\gg \sum_{i=2}^5 x_i = 10 + 15 + 40 + 35 = 100$$

$$\gg \text{Portanto, } \sum_{i=2}^5 6x_i = 600 = 6 \cdot \sum_{i=2}^5 x_i$$

P3 – Somatório da soma: o somatório da soma de duas variáveis é a soma dos somatórios destas variáveis, ou seja:

$$\sum_{i=m}^n (x_i + y_i) = \sum_{i=m}^n x_i + \sum_{i=m}^n y_i$$

De fato, teremos que:

$$\begin{aligned}
 \sum_{i=m}^n (x_i + y_i) &= x_m + y_m + x_{m+1} + y_{m+1} + x_{m+2} + y_{m+2} + \dots + x_n + y_n = \\
 &= x_m + x_{m+1} + x_{m+2} + \dots + x_n + y_m + y_{m+1} + y_{m+2} + \dots + y_n = \\
 &= (x_m + x_{m+1} + x_{m+2} + \dots + x_n) + (y_m + y_{m+1} + y_{m+2} + \dots + y_n) \\
 &= \sum_{i=m}^n x_i + \sum_{i=m}^n y_i
 \end{aligned}$$

- Exemplo: se temos duas variáveis X e Y tais que:

$$X = \{x_0 = 2, x_1 = 1,5, x_2 = 3,5, x_3 = 9\} \text{ e}$$

$$Y = \{y_0 = 1, y_1 = 3, y_2 = 10, y_3 = 30\}, \text{ então,}$$

$$\gg \sum_{i=0}^3 (x_i + y_i) = (2 + 1) + (1,5 + 3) + (3,5 + 10) + (9 + 30) = 60$$

$$\gg \sum_{i=0}^3 x_i = 2 + 1,5 + 3,5 + 9 = 16$$

$$\gg \sum_{i=0}^3 y_i = 1 + 3 + 10 + 30 = 44$$

$$\gg \sum_{i=0}^3 x_i + \sum_{i=0}^3 y_i = 16 + 44 = 60$$

$$\gg \text{Ou seja, } \sum_{i=0}^3 (x_i + y_i) = \sum_{i=0}^3 x_i + \sum_{i=0}^3 y_i, \text{ conforme esperado.}$$

A propriedade P3 tem duas extensões imediatas, que não serão demonstradas nem exemplificadas, quais sejam:

$$\bullet \sum_{i=m}^n (x_i - y_i) = \sum_{i=m}^n x_i - \sum_{i=m}^n y_i$$

$$\bullet \sum_{i=m}^n (x_{1i} + x_{2i} + \dots + x_{ki}) = \sum_{i=m}^n x_{1i} + \sum_{i=m}^n x_{2i} + \dots + \sum_{i=m}^n x_{ki}$$

P4 – Somatório ao quadrado: o quadrado de somatório dos elementos de uma variável é igual ao somatório dos quadrados dos elementos adicionado ao dobro do produto dos elementos distintos, tomados dois a dois, ou seja,

$$\left(\sum_{i=m}^n x_i \right)^2 = \sum_{i=m}^n x_i^2 + 2 \sum_{\substack{i=m \\ i \neq j}}^n x_i x_j$$

Observe agora que $\left(\sum_{i=m}^n x_i \right)^2 = (x_m + x_{m+1} + \dots + x_n)^2$. Desenvolvendo esta soma, temos:

$$\begin{aligned} \left(\sum_{i=m}^n x_i \right)^2 &= x_m^2 + x_{m+1}^2 + \dots + x_n^2 + 2x_m x_{m+1} + 2x_m x_{m+2} + \dots + 2x_k x_l + \dots + 2x_{n-1} x_n = \\ &= (x_m^2 + x_{m+1}^2 + \dots + x_n^2) + 2(x_m x_{m+1} + x_m x_{m+2} + \dots + x_k x_l + \dots + x_{n-1} x_n) = \\ &= \sum_{i=m}^n x_i^2 + 2 \sum_{\substack{i=m \\ i \neq j}}^n x_i x_j \end{aligned}$$

- Exemplo: tomando $Y = \{y_0 = 1, y_1 = 3, y_2 = 10, y_3 = 30\}$, teremos

$$\gg \sum_{i=0}^3 y_i = 1 + 3 + 10 + 30 = 44 \Rightarrow \left(\sum_{i=m}^n y_i \right)^2 = 44^2 = 1936$$

$$\gg \sum_{i=0}^{3n} y_i^2 = 1^2 + 3^2 + 10^2 + 30^2 = 1 + 9 + 100 + 900 = 1010$$

$$\gg \sum_{\substack{i=0 \\ i \neq j}}^3 y_i y_j = 1.3 + 1.10 + 1.30 + 3.10 + 3.30 + 10.30 = 463$$

$$\gg \sum_{i=0}^3 y_i^2 + 2 \sum_{\substack{i=0 \\ i \neq j}}^3 y_i y_j = 1010 + 2.463 = 1936$$

O Quadro 3.2 apresenta uma síntese das quatro propriedades apresentadas. Este conjunto de resultados será utilizado nos conteúdos posteriores, em especial a partir do capítulo sobre medidas descritivas de posição. Como forma de exercitar, e desde agora já apresentar resultados que serão apresentados nestes capítulos, temos a seguir um conjunto de três exercícios resolvidos.

Quadro 3.2 – Resumo das propriedades

P1 – Somatório da constante	P3 – Somatório do produto por constante
$\sum_{i=m}^n a = (n - m + 1) \cdot a$	$\sum_{i=m}^n ax_i = a \sum_{i=m}^n x_i$
P3 – Somatório da soma	P4 – Somatório ao quadrado
$\sum_{i=m}^n (x_i + y_i) = \sum_{i=m}^n x_i + \sum_{i=m}^n y_i$	$\left(\sum_{i=m}^n x_i \right)^2 = \sum_{i=m}^n x_i^2 + 2 \sum_{\substack{i=m \\ i \neq j}}^n x_i x_j$

Exercícios resolvidos

1. Seja $a = \sum_{i=m}^n \frac{x_i}{(n - m + 1)}$. Usando as propriedades indicadas, mostre que:

$$\sum_{i=m}^n (x_i - a) = 0$$

Solução:

- Observemos inicialmente que, da primeira expressão, tomaremos a como uma constante.
- Veja ainda que: $a = \sum_{i=m}^n \frac{x_i}{(n - m + 1)} = \frac{1}{(n - m + 1)} \sum_{i=m}^n x_i$, pois $[1/(n - m + 1)]$ é constante em relação ao somatório.
- Portanto, teremos que: $\sum_{i=m}^n x_i = a(n - m + 1)$ (I)
- Mas sabemos que: $\sum_{i=m}^n a = (n - m + 1)a$ (II)
- E que: $\sum_{i=m}^n (x_i - a) = \sum_{i=m}^n x_i - \sum_{i=m}^n a$ (III).
- Logo, substituindo as expressões I e II no segundo membro de III, teremos que:

$$\sum_{i=1}^n (x_i - a) = a(n - m + 1) - a(n - m + 1) \Rightarrow \sum_{i=1}^n (x_i - a) = 0$$

2. Novamente, se $a = \sum_{i=m}^n \frac{x_i}{(n-m+1)}$, mostre que:

$$\sum_{i=m}^n \frac{(x_i - a)^2}{(n-m+1)} = \sum_{i=m}^n \frac{x_i^2}{(n-m+1)} - a^2$$

Solução:

- Vejamos inicialmente que:

$$\begin{aligned} \sum_{i=m}^n \frac{(x_i - a)^2}{(n-m+1)} &= \frac{1}{(n-m+1)} \sum_{i=m}^n (x_i - a)^2 = \frac{1}{(n-m+1)} \sum_{i=m}^n (x_i^2 - 2x_i a + a^2) = \\ &= \frac{1}{(n-m+1)} \left(\sum_{i=m}^n x_i^2 - \sum_{i=m}^n 2x_i a + \sum_{i=m}^n a^2 \right) = \\ &= \frac{1}{(n-m+1)} \left(\sum_{i=m}^n x_i^2 - 2a \sum_{i=m}^n x_i + \sum_{i=m}^n a^2 \right) \quad (\text{I}) \end{aligned}$$

- Mas do exercício 1 vimos que

$$a = \sum_{i=m}^n \frac{x_i}{(n-m+1)} \Rightarrow \sum_{i=m}^n x_i = a(n-m+1) \quad (\text{II})$$

- E sabemos que

$$\sum_{i=m}^n a^2 = a^2(n-m+1) \quad (\text{III})$$

- Substituindo II e III em I, teremos que:

$$\begin{aligned} \sum_{i=m}^n \frac{(x_i - a)^2}{(n-m+1)} &= \frac{1}{(n-m+1)} \left(\sum_{i=m}^n x_i^2 - 2aa(n-m+1) + (n-m+1)a^2 \right) = \\ &= \frac{1}{(n-m+1)} \left(\sum_{i=m}^n x_i^2 - 2a^2(n-m+1) + a^2(n-m+1) \right) = \\ &= \frac{1}{(n-m+1)} \left(\sum_{i=m}^n x_i^2 - a^2(n-m+1) \right) = \left(\sum_{i=m}^n \frac{x_i^2}{(n-m+1)} - \frac{a^2(n-m+1)}{(n-m+1)} \right) \end{aligned}$$

$$\text{Portanto, } \sum_{i=m}^n \frac{(x_i - a)^2}{(n-m+1)} = \sum_{i=m}^n \frac{x_i^2}{(n-m+1)} - a^2$$

3. Sejam agora $a = \sum_{i=m}^n \frac{x_i}{(n-m+1)}$ e $b = \sum_{i=m}^n \frac{y_i}{(n-m+1)}$. Mostre que:

$$\sum_{i=m}^n \frac{(x_i - a) \cdot (y_i - b)}{(n-m+1)} = \sum_{i=m}^n \frac{x_i \cdot y_i}{(n-m+1)} - ab$$

Solução:

- Vejamos inicialmente que:

$$\begin{aligned} \sum_{i=m}^n \frac{(x_i - a) \cdot (y_i - b)}{(n-m+1)} &= \frac{1}{(n-m+1)} \sum_{i=m}^n (x_i - a) \cdot (y_i - b) = \\ &= \frac{1}{(n-m+1)} \left(\sum_{i=m}^n (x_i y_i - x_i b - a y_i + ab) \right) = \\ &= \frac{1}{(n-m+1)} \left(\sum_{i=m}^n x_i y_i - \sum_{i=m}^n x_i b - \sum_{i=m}^n a y_i + \sum_{i=m}^n ab \right) \\ &= \frac{1}{(n-m+1)} \left(\sum_{i=m}^n x_i y_i - b \sum_{i=m}^n x_i - a \sum_{i=m}^n y_i + (n-m+1)ab \right) \quad (\text{I}) \end{aligned}$$

- Mas de $a = \sum_{i=m}^n \frac{x_i}{(n-m+1)}$ e de $b = \sum_{i=m}^n \frac{y_i}{(n-m+1)}$, teremos que:

$$\sum_{i=m}^n x_i = a(n-m+1) \quad (\text{II})$$

$$\sum_{i=m}^n y_i = b(n-m+1) \quad (\text{III})$$

- Substituindo II e III em I, teremos que:

$$\begin{aligned} \sum_{i=m}^n \frac{(x_i - a) \cdot (y_i - b)}{(n-m+1)} &= \frac{1}{(n-m+1)} \left(\sum_{i=m}^n x_i y_i - b \sum_{i=m}^n x_i - a \sum_{i=m}^n y_i + (n-m+1)ab \right) \\ &= \frac{1}{(n-m+1)} \left(\sum_{i=m}^n x_i y_i - ba(n-m+1) - ab(n-m+1) + (n-m+1)ab \right) = \\ &= \frac{1}{(n-m+1)} \left(\sum_{i=m}^n x_i y_i - ba(n-m+1) \right) = \sum_{i=m}^n \frac{x_i y_i}{(n-m+1)} - \frac{ba(n-m+1)}{(n-m+1)} \end{aligned}$$

Portanto, resolvendo a última expressão, chegamos ao resultado indicado, ou seja:

$$\sum_{i=m}^n \frac{(x_i - a) \cdot (y_i - b)}{(n - m + 1)} = \sum_{i=m}^n x_i y_i - ba$$

3.3.2. Logaritmo

Além dos conteúdos de somatório e produtório, temos ainda o uso de propriedades da chamada função logaritmo, que viabiliza transformação de operações de produto em soma e vice-versa. O conteúdo é extenso e não temos como avançar mais longamente aqui, porém não precisaremos de mais que a definição e algumas propriedades centrais.

- Definição: Dados dois números reais x e y , tais que ambos sejam maiores que zero e que x seja diferente de um, dizemos que z é o logaritmo de y na base x se x elevado a z for igual a y . Simbolicamente, temos a seguinte representação:

$$\log_x y = z \Leftrightarrow x^z = y, \text{ onde } x, y \in \mathbb{R}, 0 < x \neq 1, y > 0$$

Usamos as seguintes denominações: x – base do logaritmo; y – logaritmando; z – logaritmo. Vejamos alguns exemplos:

- $\log_{10} 100 = 2$, pois $10^2 = 100$
- $\log_{30} 30 = 1$, pois $30^1 = 30$
- $\log_{1000} 1 = 0$, pois $1000^0 = 1$
- $\log_{0,5} 64 = -6$, pois $0,5^{-6} = \left(\frac{1}{2}\right)^{-6} = (2^{-1})^{-6} = 2^6 = 64$
- $\log_{\pi} \pi = 1$, pois $\pi^1 = \pi$

Vejamos algumas propriedades relevantes:

P1 – Um valor real x elevado ao logaritmo de outro valor y na base x é sempre igual ao logaritmando. Ou seja, é válida a relação: $x^{\log_x y} = y$.

Com efeito, é evidente que $\log_x y = \log_x y$. A relação da propriedade é imediata.

- Exemplos:

» Vimos que $\log_{10} 100 = 2$. Portanto, $10^{\log_{10} 100} = 10^2 = 100$

» Vejamos agora que

$$125^{\log_5 30} = (125)^{\log_5 30} = (5^3)^{\log_5 30} = (5^{\log_5 30})^3 = 30^3 = 27000$$

P2 – Dois logaritmos de mesma base são iguais se, somente se, os logaritmandos são iguais, ou seja, $\log_x a = \log_x b \Leftrightarrow a = b$.

Observe que: $\log_x a = \log_x b \Leftrightarrow x^{\log_x b} = a$. Mas, pela propriedade anterior, teremos que $x^{\log_x b} = b \Leftrightarrow a = b$

P3 – O logaritmo do produto em uma determinada base é igual à soma dos logaritmos dos fatores na mesma base, ou seja,

$$\log_x ab = \log_x a + \log_x b, \text{ onde } x, a, b \in \mathbb{R}, 0 < x \neq 1, a > 0, b > 0$$

- Com efeito, sejam: $\log_x a = y$, $\log_x b = z$, e $\log_x ab = h$. Teremos então que:
- De $\log_x a = y \Leftrightarrow x^y = a$, e de $\log_x b = z \Leftrightarrow x^z = b$, vem que $ab = x^y x^z = x^{y+z}$.
- Mas de $\log_x ab = h \Leftrightarrow x^h = ab \Leftrightarrow ab = x^h$.
- Portanto, é imediato que $x^{y+z} = x^h$, ou seja, $h = y + z$
- Consequentemente, $\log_x ab = \log_x a + \log_x b$
- Exemplo:

» Sabemos que $\log_2 512 = 9$.

» Mas como $512 = 4 \times 128$, teremos $\log_2 512 = \log_2 4 \cdot 128 = \log_2 4 + \log_2 128 = 2 + 7 = 9$

Esta propriedade é, seguramente, a mais relevante, pois é a que transforma um produto em uma soma. Pelo que está colocado, é imediato que valem também as seguintes relações:

- $\log_x \frac{a}{b} = \log_x a - \log_x b$
- $\log_x y_1 y_2 y_3 \dots y_n = \log_x y_1 + \log_x y_2 + \log_x y_3 + \dots + \log_x y_n$

Esta última propriedade é a mais relevante, e será usada posteriormente. Uma forma alternativa de reescrita é a seguinte:

$$\log_x \prod_{i=1}^n y_i = \sum_{i=1}^n \log_x y_i$$

Para finalizar, convém atentar para as possibilidades associadas à base. Com efeito, uma vez que a base atenda aos requisitos (maior que zero e diferente de 1), é possível operar qualquer logaritmo. No entanto, para facilitar a operação e a preparação de tabelas para suporte aos cálculos, duas bases tiveram maior uso, que foram as bases 10, e o número irracional ‘e’, o número de Euler, que tem valor aproximado para 2,7182, com quatro casas decimais.

A maioria dos desenvolvimentos teóricos de Estatística utiliza a base e , tendo em vista um conjunto de propriedades matemáticas relevantes. Por esta razão, o logaritmo nesta base pode ser indicado sem realçar a base, e pode ainda ser anotado como $\ln$, de logaritmo natural. Assim, teremos: $\log_e a = \log a = \ln a$

Exercícios resolvidos

1. Mostre que sendo a uma constante, é verdade que: $\log_x a^n = n \log_x a$

Solução:

- Da definição de produtório, é fácil deduzir que, se a é uma constante,

$$\prod_{i=1}^n a = \underbrace{a \cdot a \cdot a \cdot \dots \cdot a}_{n \text{ vezes}} = a^n$$

- Portanto, será válido que: $\log_x \prod_{i=1}^n a = \log_x a^n = \sum_{i=1}^n \log_x a$. Mas, como $\log_x a$ é constante em relação ao somatório, teremos que:

$$\log_x a^n = n \log_x a$$

2. Transforme a expressão a seguir em somas e subtrações, usando logaritmo natural:

$$a = \frac{\sqrt[n]{x} \cdot e^y}{z^m} \cdot \prod_{i=1}^j h$$

Solução:

- Aplicando as propriedades estudadas, teremos que:

$$\begin{aligned}
 a &= \frac{\sqrt[n]{x} \cdot e^y}{z^m} \cdot \prod_{i=1}^j h \Rightarrow \ln a = \ln \frac{\sqrt[n]{x} \cdot e^y}{z^m} \cdot \prod_{i=1}^j h = \ln \frac{\sqrt[n]{x} \cdot e^y}{z^m} + \ln \prod_{i=1}^j h = \\
 &= \ln \sqrt[n]{x} \cdot e^y + \ln \frac{1}{z^m} + \ln \prod_{i=1}^j h = \ln \sqrt[n]{x} + \ln e^y + \ln z^{-m} + \ln \prod_{i=1}^j h = \\
 &= \ln x^{\frac{1}{n}} + \ln e^y + \ln z^{-m} + \ln \prod_{i=1}^j h = \frac{1}{n} \ln x + y \ln e - m \ln z + j \ln h = \\
 &= \frac{1}{n} \ln x + y - m \ln z + j \ln h
 \end{aligned}$$

Chegamos então à seguinte relação:

$$a = \frac{\sqrt[n]{x} \cdot e^y}{z^m} \cdot \prod_{i=1}^j h \Leftrightarrow \ln a = y + \frac{\ln x}{n} - m \ln z + j \ln h$$

3.4. Resumo

A finalidade do capítulo foi complementar o processo de pesquisa, com as etapas de análise de dados e de apresentação dos resultados. Também foram desenvolvidos os conhecimentos matemáticos que servirão de apoio aos capítulos seguintes. Os principais pontos foram os seguintes:

- A análise de dados foi apresentada como sendo o processo sistemático de avaliação e interpretação de dados de uma pesquisa, por meio de aplicação de técnicas estatísticas direcionadas à exploração, à descrição, e ao teste de hipóteses;
- O paradigma de análise de dados corrente é o de *statistical learning* em que se aplicam métodos estatísticos tradicionais e computacionais na análise de dados, que se divide em conteúdos de análise supervisionada e não supervisionada;
- A análise de dados está baseada em conceitos, princípios e ferramentas de estatística descritiva e inferencial, e temos três formas de exploração

mais recorrentes: análise por meio de testes, análise multivariada, e análise exploratória de dados;

- A análise estatística por meio de testes é aquela realizada com a utilização das ferramentas da Estatística inferencial, com a finalidade de analisar consistência de relações entre variáveis, para testar regularidades nos dados, e para testar possibilidade de generalização de resultados da amostra para a população;
- Foi indicado que análise por meio de testes pode ser de dois tipos centrais, que são a análise paramétrica e a análise não paramétrica. Adicionalmente, vimos que embora uma análise por testes possa ser suficiente para determinados problemas de pesquisa, os testes são usados tanto em ferramentas multivariadas quanto exploratórias;
- A análise multivariada é o tipo de análise que utiliza conteúdos teóricos e aplicados de estatística para avaliar conjuntamente conjuntos de mais de duas variáveis. Foi ressaltado que a análise multivariada teve historicamente dificuldades de uso devido à complexidade das ferramentas matemáticas usadas, porém os avanços da computação viabilizaram um extenso uso de suas técnicas;
- Foi observado também que a análise multivariada pode envolver relações de dependência ou de independência entre as variáveis, e que se utilizam ou não testes estatísticos;
- A análise exploratória de dados envolve todo o conjunto de ferramentas de exploração dos dados de interesse de decisão ou de conhecimento, utilizando ferramentas de estatística descritiva, de análise visual, e de alguns testes e ferramentas multivariadas;
- A AED tem ainda como objetivos: analisar relações entre diferentes variáveis, fornecer indicações e *insights* sobre estrutura de dados; verificar pressupostos para as etapas posteriores de análise; indicar potenciais alternativas de ferramentas de análise; e apresentar potenciais demandas de ajustes nos dados ou de novos procedimentos de coleta de dados;
- Os métodos de apresentação de resultados de uma pesquisa são vários, e dependem da finalidade da pesquisa. Os principais são: relatório escrito estritamente descritivo; relatório detalhado com revisão; apresentação por slides ou pôsteres;

- Para a finalidade de estudo neste manuscrito, as principais ferramentas matemáticas de base são: somatório, produtório e logaritmos. Foram apresentados os principais conceitos, propriedades e foram resolvidos alguns exercícios.

Exercícios

1. Considere os quatro exercícios do capítulo anterior, e apresente como você acredita que se deve proceder em relação as duas demais etapas que foram apresentadas neste capítulo.
2. Demonstre as extensões indicadas na Propriedade P3 do somatório.
3. Mostre que $\sum_{i=m}^n (a \pm x_i) = (n - m + 1)a \pm \sum_{i=m}^n x_i$, onde a é uma constante.
4. Use um exemplo para mostrar que $\sum_{i=m}^n (x_i \cdot y_i) \neq \left(\sum_{i=m}^n x_i\right) \left(\sum_{i=m}^n y_i\right)$, ou seja que o somatório dos produtos não é necessariamente igual ao produto dos somatórios.
5. Demonstre que $\sum_{i=m}^n (x_{i+1} - x_i) = (x_{n+1} - x_m)$, a chamada propriedade telescópica. Exemplifique.
6. Sejam agora $a = \sum_{i=m}^n \frac{x_i}{(n - m + 1)}$ e $b = \sum_{i=m}^n \frac{y_i}{(n - m + 1)}$. Mostre que:

$$\sum_{i=m}^n (x_i - a) \cdot (y_i - b) = \sum_{i=m}^n \frac{x_i \cdot y_i}{(n - m + 1)} - ab$$
7. Demonstre as duas consequências da propriedade P3 dos logaritmos.
8. Transforme a expressão a seguir em somas e subtrações, usando logaritmo natural:

$$a = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-\frac{1}{2} \cdot \left(\frac{x_i - \mu}{\sigma}\right)^2}$$

PARTE II. ESTATÍSTICA DESCRITIVA UNIVARIADA_

Nesta parte, apresentamos os conteúdos da estatística descritiva univariada, que são os principais conteúdos da análise exploratória de dados. Seguimos o ordenamento convencional dos temas de estatística descritiva, que são os seguintes: organização e apresentação tabular e gráfica de dados; medidas estatísticas de posição; medidas estatísticas de dispersão; e, medidas estatísticas de formato. Estes conteúdos definem os capítulos desta parte, da seguinte forma.

Primeiramente, no Capítulo 4, temos o conteúdo relacionado à organização de dados nas formas de tabelas, sendo apresentados os principais conceitos e o detalhamento sobre os métodos de construção de tabelas. Os conteúdos deste capítulo serão parciais, pois serão complementados nos capítulos posteriores na Parte III.

No Capítulo 5, temos o complemento do capítulo 4, com a apresentação do conteúdo relacionado à análise e à construção de gráficos. É destacado o valor dos gráficos pela possibilidade de viabilizar uma análise visual. A finalidade de análise ficou restrita à análise univariada e com apenas parte dos gráficos, porém nos capítulos posteriores outros gráficos serão apresentados, na medida em que os conteúdos necessários à sua compreensão forem apresentados.

No Capítulo 6, iniciamos a análise de dados por meio de medidas descritivas de posição. A finalidade é desenvolver conhecimentos sobre as principais ferramentas de análise, enfatizando as ferramentas de uso ampliado nos últimos anos, e desenvolvendo as competências de extração por meio de softwares.

O Capítulo 7 continua a exposição sobre as medidas descritivas, agora com foco na descrição da dispersão dos dados. Este conhecimento complementa os conhecimentos sobre as medidas de posição, permitindo uma caracterização muito mais completa do conjunto de dados. Em diversas aplicações, as medidas de posição e de dispersão já chegam a serem suficientes para efeito de análise.

Por fim, no Capítulo 8, temos a exposição das principais medidas de posição, que complementam as demais medidas, e dão a base de referência completa de descrição de uma variável.

CAPÍTULO 4

ORGANIZAÇÃO E APRESENTAÇÃO DE DADOS POR TABELAS

Neste capítulo, expomos os tópicos relativos à organização e à apresentação de dados na forma de tabelas. A finalidade é desenvolver competências para a utilização desta primeira forma de exploração dos dados. O que será exposto aqui é apenas o conteúdo mais fundamental sobre o tema, pois uma parte relevante relativa à análise de tabelas cruzadas somente poderá ser desenvolvida após a compreensão dos conteúdos a serem explorados em capítulos posteriores.

Na organização do capítulo, apresentamos inicialmente os conteúdos conceituais relativos às tabelas, e em seguida indico os elementos centrais que qualquer tabela precisa possuir. Na sequência, apresentamos os procedimentos de construção de uma tabela, considerando os diferentes tipos de variáveis (discretas e quantitativas). Ao final teremos o resumo do capítulo e alguns exercícios de fixação.

Para construirmos um referencial mais pragmático, novamente aqui serão usados os dados do Apêndice D, relativos à pesquisa junto a estudantes de cursos de Administração e Turismo a respeito de sua visão sobre o curso. Ao final deste capítulo, o leitor deverá estar apto a responder às seguintes questões:

- Como organizar dados em uma tabela? Quais os detalhes formais e as decisões mais relevantes na construção de uma tabela?
- Como construir tabelas a partir de softwares estatísticos?
- Qual a melhor forma de construir uma tabela para dados quantitativos?
- Como utilizar tabelas na análise preliminar de planilhas de dados?

4.1. Organização tabular de dados

A tabela é comumente a primeira forma usada na organização de conjuntos de dados oriundos de esforços de pesquisa. Podemos definir a tabela como **uma matriz de linhas e colunas que define células que contêm informações que sintetizam dados de forma não discursiva, sendo os números referência central de informação** (deste modo, se as fontes centrais de informações de uma matriz são textos, caracterizamos a planilha como um ‘quadro’).

Atualmente, temos uma facilidade maior para a construção de tabelas, tendo em vista a disponibilidade de rotinas bastante completas nos softwares de análise de dados. No processo de organização de dados em tabelas por meio de planilhas computacionais, na verdade passamos de uma forma tabular detalhada para uma forma tabular agregada, uma vez que uma planilha contendo conjuntos de dados é na verdade uma tabela em que todos os dados estão ‘totalmente’ detalhados. Na forma agregada, fazemos um esforço que visa, acima de tudo, providenciar um conhecimento mais profundo sobre os dados, que, no detalhamento da planilha completa, não se verifica facilmente. Podemos dizer então que, no processo de construção de tabelas, ‘trocamos’ detalhamento por facilidade de entendimento, ou seja, perdemos detalhes individuais, mas isto nos permite uma visão agregada do conjunto dados.

Há um conjunto de elementos conceituais, formais e procedimentais relativos à construção e uma tabela. Para facilitar a compreensão sobre o assunto, optamos por discorrer sobre estes aspectos em dois contextos: formal e procedimental (os aspectos conceituais aparecerão ao longo da exposição).

4.2. Elementos centrais de uma tabela

Entendemos por elementos gerais àquilo que precisa necessariamente estar em uma tabela, e que são definidos por ‘normas’ de construção. No Brasil, várias instituições utilizam como referência as normas definidas pelo IBGE⁸, embora tais normas sejam específicas para regular somente as tabelas produzidas por este órgão público. Embora estas regras sejam uma base de referência relevante, aqui não seguiremos regras específicas nem do IBGE nem de qualquer instituição.

8 A própria Associação Brasileira de Normas Técnicas, a ABNT, que fixa normas para diversas atividades no Brasil, prescreve que se usem as regras do IBGE para tabelas.

Em situações de prática, o acadêmico, o pesquisador ou o estudante devem atentar para o contexto de sua análise para saber se deve ou não seguir determinadas regras de apresentação. Em geral, em caso de uma pesquisa de interesse decisório, há certa flexibilidade na construção da tabela; já no caso de pesquisas acadêmicas, é comum as instituições de pesquisa (como as universidades, por exemplo) fixarem regras específicas de apresentação.

Como vimos acima, uma tabela é uma matriz com linhas e colunas; no entanto, a simples disposição de uma matriz pode não ser suficientemente informativa, e, por esta razão, a representação de dados em tabelas normalmente requer pelo menos mais dois elementos adicionais, que são seu título e as notas. Vejamos cada um destes elementos:

- **Título:** texto sintético, que enuncia o conteúdo da tabela e dá outras informações relevantes. A depender do objeto da análise, o título pode conter um nome para a tabela, além da indicação de seu conteúdo central, e de dados referentes a local e período de referência dos dados. A regra básica é que o título seja breve, claro e bem explicativo do conteúdo;
- **Notas:** texto sintético, que complementa as informações contidas na tabela, apresentando informações como a fonte dos dados, as explicações sobre siglas usadas, dentre outros. As notas são associadas às chamadas feitas no corpo da tabela, na forma de asteriscos, letras ou números sobrescritos. A regra central é a mesma aplicada ao título, ou seja, as notas, para serem bons complementos do conteúdo da tabela, precisam ser sucintas, claras e bem explicativas.

Na tabela em si, temos três conjuntos de elementos específicos, que são: o cabeçalho, a coluna indicadora e as células de dados. Vejamos cada um destes elementos:

- **Cabeçalho:** é conjunto informações colocadas na primeira linha da tabela para informar o conteúdo de cada coluna. O cabeçalho pode eventualmente ocupar mais de uma linha, quando uma principal é subdividida em outras;
- **Coluna indicadora:** é conjunto de informações que vêm na primeira coluna da tabela, para informar o conteúdo de cada linha;

- Células de dados: são os dados numéricos que aparecem no cruzamento das colunas e linhas, que têm seu significado associado ao conteúdo do cabeçalho e da coluna indicadora simultaneamente.

O cabeçalho é bastante simples, e contém na maioria das vezes somente a indicação de frequência (contagem) e percentuais, além do conteúdo da primeira célula, que remete ao conteúdo da coluna indicadora. Casos especiais podem dar ao cabeçalho mais elementos, como será observado no item 4.3.

A coluna indicadora tem sua construção dependente da natureza da variável, considerando sua classificação em categórica e quantitativa. Este assunto será discutido no item 4.4, porém desde já convém realçar que, no caso das variáveis categóricas, cada célula da coluna indicadora será ocupada por uma categoria (eventualmente as categorias podem ser reunidas, a critério da conveniência do pesquisador), e no caso das variáveis quantitativas, as células serão ocupadas por valores ou por intervalos de valores (ver item 4.5).

Quanto às células, são os locais onde ficam os dados numéricos. Um cuidado especial deve ser dado para a primeira célula superior e à esquerda, que pertence ao mesmo tempo ao cabeçalho e à coluna indicadora. Esta célula refere-se sempre ao que vem logo abaixo, ou seja, às informações da coluna indicadora. Comumente usamos uma denominação específica nesta célula, dependendo da natureza das informações indicadoras. Assim, por exemplo, se estamos avaliando frequências de gênero, a primeira célula pode receber a denominação de ‘Gênero’ ou ‘Sexo’, ou simplesmente ‘Categorias’, ou ‘Opções’. Ao longo da exposição outros casos aparecerão e possibilitarão um melhor entendimento das alternativas existentes.

Figura 4.1 – Exemplo de tabela

Título →

Cabeçalho →

Coluna indicadora →

Tabela 1 – Religiões dos trabalhadores do Brasil		
Categorias	Frequência	Percentual
Religião 1	f_1	p_1^a
Religião 2	f_2	p_2
...	...	...
Total	Soma ^b	100

←

Células de dados

Fonte: IBGE.

^a Números em percentual; ^b Soma = $f_1 + f_2 + \dots$

← Notas

A Figura 4.1 apresenta um modelo de tabela que ilustra os elementos acima apresentados. Cabe observar, primeiro, que é possível que algumas bordas de linhas ou de colunas sejam propositadamente retiradas, por razões estéticas; na figura, retirei somente as linhas laterais, mas seria possível também retirar outras, a critério do pesquisador ou das normas que estão sendo seguidas. Em segundo lugar, é usual deixar a última linha contendo somatório dos valores de cada coluna, sendo também possível utilizar a última coluna contendo o somatório dos valores por linha, de modo que a última célula abaixo e à direita apresente o somatório total dos dados.

4.3. Opções de preenchimento de tabelas

Em uma tabela, a quantidade de linhas e colunas depende do interesse de pesquisa em termos de desdobramento de informações. Para facilitar o entendimento desta afirmação, tomemos por base a tabela da Figura 4.1 e observamos primeiramente a coluna indicadora, que contém as alternativas de religião, ou seja, as categorias selecionadas para uma variável categórica sob análise, e cada célula está indicando uma religião diferente. Assim, teremos tantas linhas quantas forem as religiões registradas na pesquisa. Somente para os casos em que temos uma grande quantidade de categorias ou de faixas nas linhas da coluna indicadora com frequência pequena, é que poderão ser reunidas as linhas em nomes mais genéricos, que evitem a apresentação de uma tabela com muitas linhas de conteúdo pouco informativo (como ‘outras’, ‘demais’...).

Na mesma figura, temos na segunda coluna as frequências de pessoas de cada religião, que devem ser contadas a partir da massa de dados disponibilizada (representaremos cada frequência pela letra f indexada, conforme indicado $(f_1, f_2, \dots)$). Além da coluna de frequência, é comum termos ainda em uma terceira coluna a apresentação dos percentuais que cada quantidade representa em relação ao total (representaremos cada percentual pela letra p indexada, conforme indicado $(p_1, p_2, \dots)$).

Embora seja possível indicar somente a quantidade expressa nas frequências, é sempre recomendada a apresentação de percentuais, pois estes permitem uma visualização comparativa em uma base fixa, que é 100%. Além dos percentuais, em determinados tratamentos é mais conveniente trabalhar com os valores somente da razão da quantidade pelo total, sem apresentação na

forma de percentuais (por exemplo, 0,10 em lugar de 10%). Cada circunstância indicará a opção mais apropriada.

A indicação de percentuais se faz da seguinte forma: tomamos o valor da frequência f_i , o dividimos pela soma de todas as frequências, e em seguida multiplicamos o valor por 100. Simbolicamente, a formulação é a seguinte:

$$p_i = \frac{f_i}{\sum_{i=1}^n f_i} 100$$

- Lê-se assim: o percentual da categoria i é a frequência correspondente dividida pela soma de todas as frequências, tudo isto multiplicado por 100.

Caso o interesse seja de apresentação de razões e não de percentuais, a fórmula se altera somente pela saída da multiplicação por 100. É óbvio que a soma dos percentuais precisa ser 100 (e das razões precisa ser 1), mas é comum que em algumas situações e softwares haja erros de arredondamento, o que requer cuidados do pesquisador antes de apresentar os resultados.

O tratamento e a apresentação de frequências e percentuais depende da quantidade de variáveis que são apresentadas simultaneamente na tabela. Quando temos uma situação como a que está acima exemplificada, nossa tabela será chamada de '**tabela simples**', ou 'tabela de uma entrada', pois, embora tenhamos duas colunas com números, uma delas é uma reescrita da outra. Neste caso, a tabela é dirigida à análise unidimensional. Quando temos duas variáveis em uma mesma tabela, teremos uma 'tabela de entrada dupla', que é dirigida à análise bidimensional. Vejamos como dispor as informações em cada uma delas, e os detalhes de implementação nos softwares R, JASP e em Python.

– Tabela simples

Em tabelas de uma entrada, é recomendado que nas células de informações sejam apresentados as frequências e os percentuais em colunas separadas. Outra opção, menos é que se coloquem os percentuais unidos às frequências. Vejamos um exemplo.

Exemplo 1: Nos dados referentes à pesquisa sobre avaliações de alunos de Administração e de Turismo sobre seus cursos e profissões, uma tabela inicial que podemos extrair é aquela que indica as frequências de pesquisados por cursos. Temos abaixo a apresentação em duas opções de formato: a tabela da

esquerda contendo os percentuais em colunas separadas e a da direita com os percentuais junto com as frequências.

Tabela 1 – Cursos de origem dos alunos			Tabela 1 – Cursos de origem dos alunos	
Curso	Frequência	Percentual	Curso	Frequência (%)
Administração	60	53,1	Administração	60 (53,1%)
Turismo	53	46,9	Turismo	53 (46,9%)
Total	113	100,0	Total	113 (100%)
Fonte: dados da pesquisa.			Fonte: dados da pesquisa.	

Os percentuais (e as frequências) podem ser apresentados em seus valores em si, mas podem ainda ser apresentados em versões acumuladas, em novas colunas. Temos então duas opções: acumulada ‘abaixo de’, e ‘acima de’. Vejamos as duas abaixo:

- Na formação da acumulação ‘abaixo de’, apresentamos progressivamente os valores que se acumulam ao longo das linhas, permitindo verificar quantidades e percentuais acumulados em até cada linha;
- Já a formação da acumulação ‘acima de’ contém as informações das quantidades ou percentuais que vão se reduzindo ao longo das linhas da tabela. O exemplo seguinte ilustra as duas situações.

Exemplo 2: nos dados da pesquisa com estudantes, verifiquemos abaixo a disposição dos estudantes por semestre do curso. Além das duas colunas acima indicadas (frequência e percentuais) temos os percentuais ‘acima de’ e ‘abaixo de’ em mais duas colunas (como ilustração de mais uma opção de edição, nesta tabela não temos as listas internas, salvo que fecha a última linha).

Semestre	Freq.	Perc.	Perc. ‘abaixo de’	Perc. ‘acima de’
1	23	20,5	20,5	100
2	5	4,5	25,0	79,5
3	17	15,2	40,2	75,0
4	14	12,5	52,7	59,8
5	12	10,7	63,4	47,3
6	11	9,8	73,2	36,6
7	25	22,3	95,5	26,8
8	5	4,5	100,0	4,5
Total	112*	100,0	-	-

* Valor total com 1 dado perdido dos 113 originais

Das formações acumuladas, a mais comumente usada é a acumulada ‘abaixo de’, tanto em análise exploratória quanto nos estudos de Teoria das probabilidades e Métodos não paramétricos. Por esta razão, a expressão ‘distribuição acumulada’ é usada frequentemente em lugar da expressão completa ‘distribuição acumulada abaixo de’. Cabe ainda ressaltar que a indicação das frequências acumuladas é mais apropriadamente aplicada para o caso de ordenamento de variáveis quantitativas, como será exposto no item 4.4, embora nos demais casos seja também possível desenvolver análises com este tipo de exposição.

– Tabela simples no R

Seguindo o exemplo 2 acima, faremos no R uma tabela simples da distribuição dos alunos por semestre com indicação de frequência, percentual absoluto e percentual acumulado ‘abaixo de’. Para tanto, deve-se:

- Importar os dados do Apêndice D para o R, conforme procedimento indicado no item b do tópico 2.3.
- Implementar o código abaixo.

```
tabela = data.frame(
  Frequência = as.vector(table(Sem)),
  Percentual_absoluto = round(as.vector(prop.table(table(Sem)) *
100), 2),
  Percentual_acumulado = round(cumsum(prop.table(table(Sem)) *
100), 2)); tabela
```

Pode-se, ainda, solicitar a alteração da saída da tabela para que os dados fiquem em ordem crescente ou decrescente. Para tanto, utiliza-se o comando “order” e altera o dataframe ‘tabela’ criado para que as colunas sejam ordenadas de acordo com a disposição crescente ou decrescente da coluna frequência.

```
tabela = tabela[order(tabela$Frequência),]; tabela #crescente

tabela = tabela[order(-tabela$Frequência),]; tabela #decrescente
```

– Tabela simples no JASP

Para a construção de tabela simples no JASP, também importaremos a nossa base de dados do Apêndice D, conforme indicado no item d do tópico 2.3. Isto feito, clique em: *descriptives* → *descriptive statistics* → arraste a variável “Sem” para o campo de *variables* → clique em *Tables* → *Frequency tables*.

Note que as frequências não estarão ordenadas. Para ordenar, temos que utilizar o R Console disponível no JASP. Para acessar o R Console, clique no ícone de  (em nossa versão, fica à direita na parte superior em azul) e selecione R console. Após isto, o módulo aparecerá na parte superior do ambiente do JASP. Clique, portanto, em R console e utilize os seguintes códigos para que possamos ordenar as frequências de forma crescente ou decrescente:

```
tabela = data.frame(
  Frequência = as.vector(table(data$Sem)),
  Percentual_absoluto = round(as.vector(prop.table(table(data$Sem))
* 100), 2),
  Percentual_acumulado = round(cumsum(prop.table(table(data$Sem)) *
100), 2)
)
head(tabela) #exibir

tabela$Frequência = as.numeric(tabela$Frequência)#para deixar a
frequência como numérica e evitar erros

tabela = tabela[order(-tabela$Frequência), ] #ordem decrescente
head(tabela)#exibir

tabela = tabela[order(tabela$Frequência), ] #ordem crescente
head(tabela)#exibir
```

– Tabela simples em Python

Segue-se a mesma lógica de código realizada na aplicação em R. Portanto:

- Importar os dados do Apêndice D para o VSCode, conforme procedimento indicado no item c do tópico 2.3.
- Implementar o código abaixo.

```
tabela = pd.DataFrame({
    "Frequência": df['Sem'].value_counts(),
    "Percentual_absoluto": (df['Sem'].value_counts() / df['Sem'].
value_counts().sum() * 100).round(2),
})
print(tabela)
```

O processo de ordenação dos resultados de acordo com a ordem crescente ou decrescente da frequência também é possível. Modifica-se a tabela com a função `.sort_values()`, conforme exposto abaixo:

```
tabela_crescente = tabela.sort_
values(by="Frequência",ascending=True) #crescente
print(tabela_crescente)

tabela_decrescente = tabela.sort_
values(by="Frequência",ascending=False) #decrescente
print(tabela_decrescente)
```

– Tabela de dupla entrada

A tabela de dupla entrada é aquela na qual são apresentadas simultaneamente as informações de duas variáveis, a partir do cruzamento das observações (por isto esta tabela é também chamada de ‘tabela cruzada’). Aqui, teremos nas linhas as frequências da primeira variável e nas colunas as frequências da segunda.

Como elemento de maior distinção em relação à tabela simples, cabe observar que o cabeçalho se altera, uma vez que passa a ser a representação de uma nova variável, o que requer ao menos uma linha adicional, que normalmente vem logo acima das categorias da variável. Assim, as colunas passam a apresentar não somente valores de frequências e percentuais gerais, mas também os valores para cada categoria da nova variável.

Devido ao fato de termos duas variáveis, a análise se torna bidimensional. Por esta razão, apresentarei elementos gerais aqui, porém o conteúdo será retomado, com enfoque complementar, na parte III deste manuscrito. Vejamos um exemplo:

Exemplo 3: Retomando as duas tabelas anteriormente apresentadas, vejamos agora a tabela em que as duas variáveis são apresentadas simultaneamente.

Curso	Qual o semestre (aproximado)?								Total
	1	2	3	4	5	6	7	8	
Administração	14	2	10	9	6	6	10	3	60
Turismo	9	3	7	5	6	5	15	2	52
Total	23	5	17	14	12	11	25	5	112

Na tabela bidimensional, temos dois espaços que totalizam as frequências, quais sejam: a última coluna, que totaliza as linhas; e a última linha, que totaliza as colunas (a última célula contém somatório geral). Os valores destas linhas indicam a ‘distribuição marginal’ das respectivas variáveis. Assim, a última coluna representa a distribuição marginal da variável cujas categorias estão nas linhas, e a última linha representa a distribuição marginal da variável cujas categorias estão indicadas nas colunas. Embora seja possível não apresentar os totais, em tabelas cruzadas a apresentação é sempre recomendada.

Concernente aos valores em percentual, temos aqui três opções para cada célula: o percentual em relação à variável representada nas linhas, o percentual em relação à variável apresentada nas colunas, e o percentual em relação ao total de elementos. O interesse de pesquisa indicará se será ou não necessário apresentar nas tabelas os valores percentuais, e, se necessário, quais deles apresentar e de que forma. O maior cuidado deve ser sempre com relação à disposição visual dos dados, devendo-se sempre buscar uma tabela que seja ‘limpa’ visualmente. Vejamos mais um exemplo:

Exemplo 4: a partir dos dados da planilha do Apêndice D, apresentamos a tabela de dupla entrada das variáveis curso e gênero, conforme indicado abaixo. Decidimos apresentar todos os percentuais possíveis, com a adição de mais três linhas, uma para cada categoria de curso, e acrescentamos uma coluna adicional com a indicação das medidas extraídas.

Curso	Medidas	Gênero		Total
		Masculino	Feminino	
Administração	Frequência	29	28	57
	% no curso	50,9	49,1	100,0
	% do gênero	69,0	41,2	51,8
	% do total	26,4	25,5	51,8
Turismo	Frequência	13	40	53
	% no curso	24,5	75,5	100,0
	% do gênero	31,0	58,8	48,2
	% do total	11,8	36,4	48,2
Total	Frequência	42	68	110
	% no curso	38,2	61,8	100,0
	% do gênero	100,0	100,0	100,0
	% do total	38,2	61,8	100,0

- Tabela de dupla entrada no R

Para uma tabela de dupla entrada apenas com uma informação, por exemplo, frequência, utiliza-se o comando `table(var1, var2)`, em que `var1` é a variável de interesse 1 (por exemplo, curso) e `var2` é variável de interesse 2 (por exemplo, sexo). Para uma tabela de dupla entrada com informações de frequência e percentual, deve-se implementar o código abaixo. Utilizaremos a base de dados do Apêndice D e as variáveis `Cur` (curso) e `Sex` (sexo).

```
tabela_dupla = cbind(
  Frequência = table(Cur,Sex),
  Percentual = round(prop.table(table(Cur,Sex))*100, 2)); tabela_
dupla
```

#para ordenar a apresentação das colunas para que as informações de frequência e percentual de cada sexo fiquem vizinhas, pode-se implementar o código abaixo.

```
tabela_dupla = data.frame(
  Frequência_Feminino = table(Cur,Sex)[, "Fem."],
  Percentual_Feminino = round(prop.table(table(Cur,Sex)[,
"Fem."])*100, 2),
  Frequência_Masculino = table(Cur,Sex)[, "Masc."],
  Percentual_Masculino = round(prop.table(table(Cur,Sex)[,
"Masc."])*100, 2)); tabela_dupla
```

A situação *default* do software é aquela em que a apresentação segue o ordenamento crescente da codificação adotada na construção da variável que estiver na primeira coluna (como no caso da tabela simples). Mas é aqui possível ordenar em ordem crescente ou decrescente pelo mesmo critério (ordem de construção da variável), a partir do seguinte comando:

```
tabela_dupla = tabela_dupla[order(tabela_dupla$Frequência_
Feminino),]; tabela_dupla #Crescente por frequência feminino

tabela_dupla = tabela_dupla[order(-tabela_dupla$Frequência_
Feminino),]; tabela_dupla #Decrescente por frequência feminino

tabela_dupla = tabela_dupla[order(tabela_dupla$Frequência_
Masculino),]; tabela_dupla #Crescente por frequência masculino

tabela_dupla = tabela_dupla[order(-tabela_dupla$Frequência_
Masculino),]; tabela_dupla #Decrescente por frequência masculino
```

– Tabela de dupla entrada no JASP

Para a construção de tabela de dupla entrada no JASP, utilizaremos a função Tabela de Contingências, clicando em: frequencys → contingencies table → arraste a variável “Cur” para o campo de rows → arraste a variável “Sex” para o campo de columns.

Para ordenar as frequências em ordem crescente ou decrescentes, podemos ir em options e selecionar ascending para crescentes e descending para decrescente. Também podemos optar por exibir percentuais e mostrar os totais no campo cells.

– Tabela de dupla entrada em Python

Utilizando o pacote pandas, vamos criar uma tabela de dupla entrada para o exemplo que estamos seguindo. A função central para criar a tabela de dupla entrada é `pd.crosstab(df['var1'],df['var2'])`. O código completo pode ser consultado abaixo:

```
tabela_dupla = pd.concat([pd.crosstab(df['Cur'], df['Sex']),
pd.crosstab(df['Cur'], df['Sex'], normalize='all').mul(100).
round(2)],Keys=["Frequência","Percentual"])

#para mudar o percentual da parte inferior para a parte superior,
pode-se adicionar a função axis=1:

tabela_dupla = pd.concat([
pd.crosstab(df['Cur'], df['Sex']),
pd.crosstab(df['Cur'], df['Sex'], normalize='all').mul(100).
round(2)],
axis=1, keys=["Frequência", "Percentual"]
)
print(tabela_dupla)
```

4.4. Construção de tabelas segundo os tipos de variáveis

Conforme comentado no item 4.2, a construção da coluna indicadora dependerá da natureza da variável, conforme sejam categóricas ou quantitativas. Para as variáveis categóricas, independentemente de serem nominais ou ordinais, a sua apresentação se fará pela simples indicação das categorias, uma em

cada linha, seguida pelas respectivas frequências e percentuais (se for o caso). Nas situações em que se achar conveniente, é possível que se reúnam mais de uma categoria em uma só linha; cada pesquisa indicará a necessidade ou não deste procedimento (todos os exemplos apresentados acima são de variáveis categóricas, e ilustram esta primeira descrição).

Já no caso das variáveis quantitativas, a forma de apresentação varia de acordo com o subtipo de variável. Assim, para o caso das variáveis do tipo discreta, se a quantidade de observações possíveis não for muito extensa, é possível que cada valor ocupe uma linha da coluna indicadora. No exemplo a seguir ilustramos esta possibilidade de apresentação.

Exemplo 5: Na tabela abaixo, temos a apresentação da variável ‘Prestígio percebido na profissão’, segundo resposta dos estudantes, na escala quantitativa discreta de 7 pontos. Temos em cada uma das opções de resposta (de 1 a 7) valores de frequências e de percentuais considerando os 3 dados perdidos, percentuais sem considerar os dados perdidos, e os percentuais acumulados (referentes somente aos percentuais válidos).

Prestígio percebido na profissão				
	Frequência	Percentual	Percentual válido	Percentual cumulativo
1	22	19,5	20,0	20,0
2	20	17,7	18,2	38,2
3	14	12,4	12,7	50,9
4	10	8,8	9,1	60,0
5	17	15,0	15,5	75,5
6	16	14,2	14,5	90,0
7	11	9,7	10,0	100,0
Total	110	97,3	100,0	
Ausentes	3	2,7		
Total	113	100,0		

Especificamente para este exemplo, é possível apresentar a tabela com todas as respostas possíveis em razão de termos somente 7 categorias. No entanto, o mesmo não poderia ser dito da variável renda, que apresenta várias alternativas possíveis, algumas das quais com frequências muito pequenas. Nestas situações, a melhor opção são as tabelas com intervalos, detalhadas no subitem seguinte.

4.5. Tabelas para dados quantitativos

Nas ocasiões em que temos variáveis com grande número de valores distintos, sejam discretos ou contínuos, a melhor forma é trabalhar com intervalos numéricos⁹, na forma de faixas que contêm um número maior de valores, o que evita problemas de espaço, e providencia um melhor ordenamento das informações para efeito de análise dos dados.

A lógica é bastante simples: reunimos todos os valores em um conjunto de intervalos numéricos, intervalos estes que ao mesmo tempo nem se sobreponham, e que contemplem todos os valores existentes (ou seja, definimos intervalos de números reais que tenham interseção vazia, e cuja união seja igual a todo o conjunto de valores).

O maior problema é justamente definir a forma para criação destes intervalos. Temos alguns procedimentos, que expomos a seguir em 8 passos, mas que não são definitivos nem suficientes, e ao final comentaremos as razões. Para facilitar a exposição, optamos por ilustrar cada passo por meio de um conjunto de dados. Temos então os valores indicados na Tabela 4.1, e suponhamos que sejam oriundos de uma aferição do proprietário de um restaurante com *self-service*, que decidiu verificar em cada dia da semana os pesos (em gramas) dos pratos consumidos, tomando 10 pessoas por dia, ou seja, 70 pessoas ao total.

Tabela 4.1 – Dados brutos de pesos de comida, por dia da semana

Seg.	Ter.	Qua.	Qui.	Sex.	Sab.	Dom.
632,3	834,1	455,5	605,1	221,7	294,5	350,4
1108,7	837,2	975,2	1112,5	1005,9	465,0	417,5
842,4	825,3	490,3	267,9	677,0	431,3	192,1
646,8	559,6	740,1	727,3	1010,9	665,3	460,4
904,0	1026,3	839,9	618,1	360,5	390,7	328,3
1025,6	695,2	273,0	970,9	1112,3	454,6	316,7
1190,9	736,6	292,9	1076,5	1029,8	187,9	406,5
850,9	1091,0	763,8	736,7	682,6	114,3	321,5
1017,2	875,6	856,7	1028,0	269,6	203,3	306,2
768,9	679,7	454,4	1152,1	746,6	278,9	290,9

9 É comum em vários livros de análise exploratória e estatística descritiva a denominação ‘classe’ na indicação dos intervalos (recomendo verificar especialmente no seguinte texto: TOLEDO, G. L.; OVALLE, I. *Estatística básica*. 2. ed. São Paulo: Atlas, 1995). Aqui serão utilizadas quaisquer das palavras, entendidas como sinônimas.

1º Definição da amplitude dos dados

Primeiramente, precisamos identificar toda a extensão dos dados, e encontrar seu ‘comprimento’, que chamaremos de ‘amplitude total’ indicaremos por A_t . Para tanto, o procedimento é simplesmente calcular a diferença entre o maior valor observado, chamado de ‘limite superior’, e indicado por L_{sup} , e o menor valor observado, chamado de ‘limite inferior’, e indicado por L_{inf} . Simbolicamente, calculamos o valor: $A_t=L_{sup}-L_{inf}$.

Para o nosso conjunto de dados da tabela 4.1, como temos um conjunto razoavelmente grande de dados, parece ser mais apropriado inicialmente dispor esses dados em ordem crescente, o que permitirá a visualização imediata do valor máximo e do valor mínimo. Temos o ordenamento apresentado na Tabela 4.2.

Tabela 4.2 – Dados em ordem crescente (por coluna)

114,3	292,9	417,5	618,1	736,7	850,9	1026,3
187,9	294,5	431,3	632,3	740,1	856,7	1028,0
192,1	306,2	454,4	646,8	746,6	875,6	1029,8
203,3	316,7	454,6	665,3	763,8	904,0	1076,5
221,7	321,5	455,5	677,0	768,9	970,9	1091,0
267,9	328,3	460,4	679,7	825,3	975,2	1108,7
269,6	350,4	465,0	682,6	834,1	1005,9	1112,3
273,0	360,5	490,3	695,2	837,2	1010,9	1112,5
278,9	390,7	559,6	727,3	839,9	1017,2	1152,1
290,9	406,5	605,1	736,6	842,4	1025,6	1190,9

É imediato que $L_{inf}=114,3$, $L_{sup}=1190,9$, e, portanto, $A_t=1190,9-114,3=1076,6$.

2º Definição da quantidade de classes

De posse da amplitude total, nosso próximo passo consiste em definir o número de intervalos em que queremos subdividi-la. Esta etapa pode ser bastante simples se temos poucos dados ou poucas alternativas de valores (como no caso discreto), pois bastaria apontar a quantidade que for mais conveniente ao interesse de pesquisa¹⁰. No entanto, para grandes volumes de dados, e se forem com valores muito variados, a tarefa é mais delicada.

10 O caso acima ilustra a situação em que temos um número razoavelmente grande de dados, mas os valores somente podem ser de 1 a 7, ou seja, são muitos dados, mas são poucos valores distintos.

Para este último caso (muitos dados com muitos valores distintos), já foram feitas várias tentativas para definir uma regra para indicação do melhor número de intervalos, não tendo havido ainda possibilidade de universalizar uma alternativa de organização como a melhor de todas. Das principais propostas, as que nosso conhecimento até aqui permite averiguar são as seguintes (indicaremos sempre k como o número de intervalos, e n como o número de elementos da amostra):

- Regra de Sturges¹¹: $k = \lceil 1 + 3,3\log_{10}n \rceil$
- Regra da maior potência de dois: k é o maior inteiro tal que $2^k \leq n$.
- Regra da raiz quadrada: $k = \lceil n^{1/2} \rceil$

É razoável esperar que estas regras apresentem números de intervalos próximos entre si, e para número pequeno de observações, isto ocorre. No entanto, para grandes volumes de dados as regras geram indicações muito discrepantes entre si, o que dificulta a decisão sobre qual regra seguir. Mas ainda assim há ocasiões em que nenhuma destas regras apresenta um número de intervalos que seja adequado para a pesquisa. Nestes termos, é sempre bom avaliar a construção ao final, e aplicar uma regra de bom senso, mesmo que a justificativa seja mais frágil em termos formais (ver passo 8º).

De nosso exemplo, com base nos dados já informados desde a tabela 4.1 podemos observar que o total de observações é 70, ou seja, $n=70$. Temos então as seguintes indicações:

- Regra de Sturges: $k = \lceil 1 + 3,3\log_{10}70 \rceil = \lceil 1 + 3,3 \cdot 1,845 \rceil$, ou seja, $k = \lceil 7,11 \rceil = 8$;
- Regra da maior potência de dois: k é o maior inteiro tal que $2^k \leq 70$, ou seja, $k = 6$, pois $2^6 = 64 \leq 70$ (veja que $2^7 = 128 \geq 70$);
- Regra da raiz quadrada: $k = \lceil 70^{1/2} \rceil = \lceil \sqrt{70} \rceil$, ou seja, $k = \lceil 8,36 \rceil = 9$.

Temos, portanto, três possibilidades distintas para o número de intervalos, que são 6, 8 e 9 (além da definição do bom senso). Para este exemplo, adotemos o número gerado pela regra de Sturges, que está entre as duas demais.

11 A notação $\lceil \cdot \rceil$ representa a função maior inteiro, ou seja, qualquer que seja o valor calculado dentro da função, o resultado será sempre o inteiro superior.

3º Determinação da amplitude dos intervalos

Depois de definido o número de intervalos, agora precisaremos informar qual será o tamanho de cada intervalo. Assim, de posse das informações já extraídas, dividiremos a amplitude total (A_t) em k intervalos de igual tamanho. Como temos uma amplitude fixa¹², esta será representada somente por h , e a obteremos dividindo-se a amplitude total pelo número de intervalos, ou seja:

$$h = \frac{A_t}{k}$$

Em nosso exemplo, teremos então o seguinte: $h = 1076,6/8$, ou seja, $h = 134,6$, com arredondamento para uma casa decimal.

4º Construção dos intervalos

O quarto passo consiste em tomar o conjunto de dados, e levando em conta as decisões dos passos anteriores, determinar os valores de referência de cada uma das classes (adotaremos a notação C_i para cada classe). Iniciamos sempre pela primeira classe (C_1), e para sua construção, tomamos primeiramente o limite inferior do conjunto de dados de dados (L_{inf} ; para uniformização da apresentação, também denominaremos o limite inferior por l_1), e o adicionamos o valor da amplitude do intervalo (h), definindo assim um limite superior, que indicaremos por L_1 .

Para a definição final do intervalo, caberá a indicação a respeito da pertinência dos limites de classes, ou seja, precisaremos saber exatamente o que estará dentro do intervalo e o que está fora. Não há regra definitiva, porém, na ausência de uma justificativa de melhor adequação, a opção preferencial é colocar o limite inferior como pertencendo ao intervalo, e o limite superior como fora do intervalo. Por este entendimento, teremos o primeiro intervalo assim determinado: $C_1 = [l_1; L_1) = [L_{inf}; L_{inf} + h)^{13}$.

O segundo intervalo (C_2) será então determinado tendo como seu limite inferior (l_2) o valor que foi o limite superior do primeiro intervalo

12 Para o caso de admitirmos intervalos com diferentes tamanhos, não temos uma regra específica a seguir. Intervalos com amplitudes distintas são possíveis, e não inviabilizam outras análises relevantes, porém, para um ordenamento, é de bom senso que se evitem intervalos assim.

13 Esta é a notação matemática, ou seja, o colchete indica que o valor ao seu lado está contido no intervalo, e o parêntese indica que o valor ao seu lado é o limite, mas não pertence ao intervalo.

(ou seja, $l_2 = L_1$), e como o limite superior (L_2) o valor do limite inferior somado à amplitude do intervalo (ou seja, $L_2 = l_2 + h$). Assim, teremos $C_2 = [l_2; L_2) = [l_2 = L_1; L_2 = l_2 + h)$.

É fácil generalizar agora a construção de qualquer intervalo, seguindo a lógica apresentada até aqui. Temos então que, um intervalo genérico C_i será dado da seguinte forma:

$$\begin{aligned} C_1 &= [l_1; L_1) = [l_1 = L_{inf}; L_1 = l_1 + h) \\ C_2 &= [l_2; L_2) = [l_2 = L_1; L_2 = l_2 + h) \\ &\vdots \\ C_i &= [l_i; L_i) = [l_i = L_{i-1}; L_i = l_i + h) \\ &\vdots \\ C_k &= [l_k; L_k) = [l_k = L_{k-1}; L_i = L_{sup}) \end{aligned}$$

É necessário um cuidado especial com o limite superior do último intervalo $C_k = [l_k; L_k) = [l_k = L_{k-1}; L_i = L_{sup})$, pois procedimentos de arredondamento podem fazer o limite calculado não coincidir com o limite superior. Assim, se arredondamos a amplitude do intervalo para baixo, o limite superior ficará abaixo do limite final, e se arredondamos para baixo, o limite superior ficará acima do limite final. Por esta razão, se houver necessidade de arredondamento na amplitude, então é recomendado que seja sempre para cima.

Em nosso exemplo, seguimos os procedimentos indicados, e tivemos então os intervalos assim definidos:

$$\begin{aligned} C_1 &= [l_1; L_1) = [114,3; 114,3 + 134,6) = [114,3; 248,9); \\ C_2 &= [l_2; L_2) = [248,9; 248,9 + 134,6) = [248,9; 383,5); \\ C_3 &= [l_3; L_3) = [383,5; 383,5 + 134,6) = [383,5; 518,1); \\ C_4 &= [l_4; L_4) = [518,1; 518,1 + 134,6) = [518,1; 652,7); \\ C_5 &= [l_5; L_5) = [652,7; 652,7 + 134,6) = [652,7; 787,3); \\ C_6 &= [l_6; L_6) = [787,3; 787,3 + 134,6) = [787,3; 921,9); \\ C_7 &= [l_7; L_7) = [921,9; 921,9 + 134,6) = [921,9; 1056,5); \\ C_8 &= [l_8; L_8) = [1056,5; 1056,5 + 134,6) = [1056,5; 1191,1). \end{aligned}$$

Observemos que, como arredondamos o valor da amplitude do intervalo para cima (ver passo anterior), o último limite, que foi 1191,1, é maior (por uma diferença de 0,2) que o limite superior do conjunto de dados, que é 1190,9. Como o maior valor está contido no intervalo, não temos problemas, e podemos então seguir a análise sem outros ajustes e revisões.

5º Disposição dos recortes na tabela

Após a definição dos intervalos, a etapa seguinte consiste em dispor os intervalos na tabela. Para esta etapa apontamos cada intervalo em cada linha da coluna indicadora, da mesma forma que dispomos cada categoria na construção de tabelas para variáveis categóricas. Para a denominação da coluna indicadora, na primeira célula (acima e à esquerda), são recomendadas as denominações ‘faixas’, ‘classes’ ou simplesmente ‘intervalos’.

A principal decisão aqui concerne à representação do intervalo, e dentre as opções existentes temos as seguintes (ver figura 4.2):

- Primeiro, podemos adotar a forma convencional de representação de intervalos matemáticos, como construído acima (ou seja, com colchetes e parênteses);
- Em segundo lugar, temos a opção de usar o símbolo ‘⊢’, em que do lado da haste vertical indica o valor com intervalo fechado, e o lado oposto indica que ali o intervalo é aberto;
- Por fim, temos a opção de indicação literal, com algo como ‘Desde x até y’, ou ‘De x até y’, dentre outras, com o cuidado de que fiquem claros os limites inferiores e superiores.

Figura 4.2 – Opções de coluna indicadora

Opção 1	Opção 2	Opção 3
[a; b]	$a \vdash b$	De ‘a’ até ‘b’
[b; c]	$b \vdash c$	De ‘b’ até ‘c’
...	...	...

Conforme observaremos posteriormente, há circunstâncias em que os intervalos não são definidos de forma completa (ou seja, com a especificação de um limite inferior e um superior). Para estes casos, a terceira opção se mostrará

mais adequada, como veremos posteriormente. Para nosso exemplo, qualquer das opções é adequada, porém optamos pela primeira opção (ver passo 7).

6º Reescrita da variável e contagem

A etapa posterior à decisão acima seria a contagem e em seguida a alocação dos valores de frequência dentro de cada intervalo. No entanto, havendo disponibilidade de uma planilha de computador, a forma mais adequada consiste em redefinir a variável, segundo os intervalos adotados. O procedimento é o seguinte:

- 1. Primeiramente, inserimos todo o conjunto de dados em uma só coluna da planilha;
- 2. Depois, dispomos o conjunto de dados em ordem crescente (não algo de fato necessário nos softwares, mas facilita a visualização da variável);
- 3. Em uma coluna paralela, redefinimos a primeira variável segundo as faixas determinadas anteriormente.

O procedimento de reescrita da variável pode ser facilmente realizado no software Excel, com a recodificação da variável. No R também é possível fazer a recodificação. Em nosso exemplo, o procedimento acima foi seguido, e considerando as duas possibilidades de software, primeiramente fizemos a recodificação no MS Excel. Nestes termos, os 70 dados foram organizados na coluna A da planilha, e depois colocados em ordem crescente. Em seguida, marcamos a célula B1 e desenvolvemos o seguinte algoritmo genérico¹⁴:

1	Se A1<248,9 Então B1=1
2	Senão se A1<383,5 Então B1=2
3	Senão se A1<518,1 Então B1=3
4	Senão se A1<652,7 Então B1=4
5	Senão se A1<787,3 Então B1=5

14 No MS Excel a rotina é anotada assim: =SE(A1<248,9;1; SE(A1<383,5;2; SE(A1<518,1;3; SE(A1<652,7;4; SE(A1<787,3;5; SE(A1<921,9;6; SE(A1<1056,5;7; SE(A1<1191,1;8;0))))))

6	Senão se $A1 < 921,9$ Então $B1=6$
7	Senão se $A1 < 1056,5$ Então $B1=7$
8	Senão se $A1 < 1191,1$ Então $B1=8$

Este procedimento retornou para a célula B1 o número 1. Em seguida, bastou ‘puxar’, no Excel, a rotina para baixo até cobrir as 70 observações (puxando a indicação do canto direito inferior da célula), o que gerou a nova variável agora categorizada de 1 a 8. Cada nova categoria representa um dos intervalos, na ordem da numeração (ou seja, 1 para C_1 , ..., 8 para C_8). Para facilitar ainda mais a extração dos percentuais, foi possível, ainda no Excel, acrescentar uma terceira coluna identificando cada um dos códigos com os respectivos.

No que concerne ao procedimento no R, levemos os dados para o software, conforme indicado no ponto b do tópico 2.3. O dataframe é composto de apenas uma coluna intitulada ‘pesos’. Definiremos os intervalos como segue abaixo.

```
categorias = cut(dados$Pesos,
breaks = c(-Inf, 248.9, 383.5, 518.1, 652.7, 787.3, 921.9, 1056.5, 1191.1, Inf),
labels= c(1, 2, 3, 4, 5, 6, 7, 8, 9),
right=FALSE); categorias

#Para incluir a nova variável de classes no dataframe
dados=as.data.frame(cbind(dados,categorias))
```

Após este procedimento surgirá no dataframe (comando: *View(dados)*) uma nova variável na coluna paralela à variável inicial, contendo os códigos criados. Assim como no caso do MS Excel, aparecerá na nova variável apenas o código numérico, e, se queremos ordenar os dados segundo os intervalos definidos, basta então rodar o código acima excluindo a linha *labels*.

7º Construção final da tabela

Após este conjunto de procedimentos, teremos então a variável original reconstruída e codificada nos mesmos moldes das variáveis categóricas ou discretas. Nestes termos, basta então proceder à extração das medidas de inte-

resse (frequências, percentuais, percentuais acumulados...). De nosso exemplo, optamos por proceder à extração diretamente do R, seguindo os caminhos anteriormente já apresentados para construção de tabelas simples. A tabela já editada está exibida na Tabela 4.3.

Tabela 4.3 – Primeira tabela de intervalos

Intervalos	Frequência	Percentuais	Perc. Acumu.
[114,3; 248,9)	5	7,1	7,1
[248,9; 383,5)	13	18,6	25,7
[383,5; 518,1)	10	14,3	40,0
[518,1; 652,7)	5	7,1	47,1
[652,7; 787,3)	12	17,1	64,3
[787,3; 921,9)	9	12,9	77,1
[921,9; 1056,5)	9	12,9	90,0
[1056,5; 1191,1)	7	10,0	100,0
Total	70	100,0	-

8º Análise da adequação final

Conforme indicado, não temos regras fixas para a construção de uma tabela de frequência (salvo em casos de regras associadas a alguma instituição na (ou para) qual a pesquisa se realiza). Deste modo, ao final do sétimo passo, é conveniente avaliar a adequação da tabela criada, e, se necessário (e possível), empreender novos ajustes, de modo a atender aos requisitos centrais de adequação para efeito de análise.

Em nosso exemplo, é possível observar que tivemos faixas com frequência baixas (o máximo que temos é 12 observações na faixa). Isto parece indicar que, neste caso, o total de 8 intervalos pareceu excessivo, sendo possível, **por uma regra somente de bom senso**, reduzir o número de faixas. Não havendo qualquer determinação fixa neste sentido, decidimos então redefinir a tabela para quatro faixas, o que facilita todo o processo de reconstrução. De fato, agregando em quatro novas faixas basta recodificar cada duas faixas consecutivas para um novo intervalo, mantendo assim amplitudes iguais nos novos intervalos. Isto pode ser feito, inclusive, diretamente na tabela já extraída, ou, como indicado, recodificando os valores no MS Excel.

Tabela 4.4 – Segunda tabela de intervalos

Intervalos	Frequência	Percentuais	Perc. Acumu.
[114,3; 383,5)	18	25,7	25,7
[383,5; 652,7)	15	21,4	47,1
[652,7; 921,9)	21	30,0	77,1
[921,9; 1191,1)	16	22,9	100,0
Total	70	100,0	-

A Tabela 4.4 pode ser então exibida. Ao que parece, a nova distribuição está mais bem organizada para efeito de análise, comparativamente ao que se observou na primeira tabela gerada. No exemplo dado, chegamos a um bom resultado após duas tentativas apenas, porém, ao longo da construção da tabela para intervalos, é possível que várias tentativas sejam feitas, de modo a alcançar a melhor forma de disposição dos dados para avaliação. Para sintetizar os passos, temos um resumo no Quadro 4.1.

Quadro 4.1 – Sínteses dos procedimentos de construção de tabelas

Passo	Procedimento
1	Definição da amplitude: fixação da extensão que separa o valor máximo e o valor mínimo do conjunto de dados, pela diferença entre os limites superior e inferior
2	Definição da quantidade de classes: decisão sobre o número de intervalos que a extensão identificada será subdividida, segundo métodos específicos
3	Determinação da amplitude dos intervalos: identificação da extensão de cada intervalo, que se obtém dividindo a amplitude pelo número de intervalos
4	Construção dos intervalos: procedimento de fixação da extensão de cada intervalo, com procedimento pela soma sucessiva da amplitude do intervalo
5	Disposição dos recortes na tabela: decisão quanto à disposição de cada intervalo na tabela
6	Reescrita da variável e contagem: procedimento de recodificação da variável para geração da tabela
7	Construção final da tabela: extração, por meio do software estatístico selecionado, as frequências de cada classe, mais os percentuais
8	Análise da adequação final: avaliação do resultado do projeto e da adequação da tabela gerada aos propósitos da pesquisa

Relativo a este conjunto de procedimentos, entendemos que ainda mais quatro observações são necessárias, antes de seguirmos para o resumo do capítulo e para os exercícios:

- Variáveis quantitativas discretas: a construção de faixas para indicação de frequência, na forma aqui apresentada, pode não ter resultados adequados em variáveis discretas, pois os cálculos podem gerar valores que não têm referência no conjunto de dados. São eventualmente necessários arredondamentos que, se não forem bem avaliados, podem complicar os resultados. Por esta razão o método é sempre preciso uma avaliação cuidadosa dos resultados das tabelas nestes tipos de variáveis, como proposta no oitavo passo;
- Amostras pequenas: pequenas amostras podem não se prestar aos procedimentos aqui indicados, pois a organização de poucos dados em faixas pode algumas vezes perder o sentido. A própria aplicação dos procedimentos de definição do número de intervalos gera indicações sem qualquer sentido. Por exemplo, em uma amostra com 3 pessoas, a regra de Sturges indica a necessidade de 3 classes (ou seja, podemos ter uma observação apenas por classe, o que não é de fato nada informativo). Para 10 pessoas, a mesma regra indica 5 classes, e as demais 4 (regra da raiz quadrada) e 3 (maior potência de 2); em qualquer dos dados, o nível de organização contribui pouco para análise. Não temos referências para indicação de um mínimo adequado para a construção de tabelas, mas a aplicação de regras como as apontadas começam a ter sentido para amostras de ao menos 50 elementos. A melhor opção aqui é a regra de bom senso;
- Poucos dados distintos: o procedimento aplicado pode ser problemático para o caso de termos poucos dados distintos, mesmo em amostras grandes. Isto pode fazer com que alguns dos intervalos gerados fiquem com um nível muito elevado de frequência e percentual, e os demais números muito pequenos. Nestes termos, além de precisarmos de uma amostra razoavelmente grande, precisamos também que os dados da amostra não se repitam com grande frequência;
- Os limites dos intervalos: em situações nas quais temos dados com grandes discrepâncias de valores, é possível que os intervalos gerados não organizem adequadamente os dados, pois os dados discrepantes ‘forçam’ a

construção de intervalos que praticamente ficam sem dados. Neste caso, temos a opção de, nos extremos inferior e superior, abrir totalmente os intervalos (teoricamente, para menos infinito no limite inferior, e para mais infinito no limite superior). Para este caso, utilizamos na coluna indicadora as expressões escritas adequadas (para o limite inferior, algo como ‘até x ’, ou ‘abaixo de x ’, e para o limite superior, algo como ‘acima de x ’, ou ‘mais que x ’). Naturalmente, este procedimento facilita a compreensão, mas a custo do comprometimento de outras atividades de análise, como veremos nos capítulos posteriores.

Cabe então reiterar a recomendação anterior, que realça a relevância do bom senso na tomada de decisão quanto às tabelas. A seguir, temos um exercício resolvido que ilustra parte do que foi anteriormente comentado, e principalmente as últimas observações.

Exercício resolvido

1. Considerando os dados da variável renda da planilha do Apêndice D, aplique os passos indicados para construção da tabela de frequência.

Resolução: seguindo os passos, temos:

1º Definição da amplitude: na planilha, o limite inferior é 1 e o superior é 50. Logo, a amplitude é $50-1=49$

2º Quantidade de classes: temos 113 respondentes. Logo, teremos:

- Pela regra de Sturges: $k = \lceil 1 + 3,3\log_{10}113 \rceil = \lceil 7,78 \rceil$, ou seja, $k = 8$
- Pela regra da maior potência de dois: o maior k inteiro tal que $2^k \leq 113$ é 6, pois $2^6 = 64 \leq 113$ (veja que $2^7 = 128 \geq 113$).
- Pela regra da raiz quadrada: $k = \lceil 113^{1/2} \rceil = \lceil 10,63 \rceil$, ou seja, $k = 11$.

Por conveniência, adotemos aqui $k = 6$, pela regra da maior potência de dois.

3º Definição da amplitude: basta agora dividir 49 por 6, que gera 8,2, com uma casa decimal de arredondamento.

4º e 5º Construção dos intervalos e apresentação da tabela: faremos os procedimentos conjunto, construindo os intervalos na própria tabela. Seguindo o processo no R, decidimos extrair as frequências, os percentuais, e os percentuais acumulados; logo, com a coluna indicadora e o cabeçalho, teremos uma tabela de 8 linhas e 4 colunas. Para a coluna indicadora adotemos o símbolo para definição dos intervalos. A tabela e os intervalos construídos segundo a recomendação indicada virão após o passo seguinte.

6º Reescrita da variável: isto pode ser procedido pelo MS Excel ou pelo R. Aqui a variável foi redefinida no Excel, e os dados foram importados para o R, a partir da qual foi extraída a contagem, no passo seguinte.

7º Construção da tabela: já no R os dados foram extraídos, e editados na tabela a seguir:

Faixas	Frequências	Percentuais	Acumulado
1 – 9,2	107	94,7	94,7
9,2 – 17,4	4	3,5	98,2
17,4 – 25,6	0	0,0	98,2
25,6 – 33,8	0	0,0	98,2
33,8 – 42,0	1	0,9	99,1
42,0 – 50,2	1	0,9	100,0
Total	113	100,0	-

8º Avaliação dos resultados: é evidente que esta tabela não descreve adequadamente os resultados, pois praticamente todos os valores (98,2%) ficaram nas duas primeiras classes.

Assim, reavaliando a planilha, observamos que o problema maior está nas duas últimas observações, que destoam dos demais valores (40 e 50). Adicionalmente, verificamos que a variável é discreta, e que a maioria dos valores está concentrada entre 1 e 12. Assim, utilizando somente uma lógica de bom senso, parece mais razoável dividir os intervalos de 2 e 2, em ordem crescente, e o último intervalo fica aberto. Temos então a seguinte tabela, já ‘pulando’ o conjunto de passos anteriores.

Faixas	Frequências	Percentuais	Acumulado
1 ou 2	18	15,9	15,9
3 ou 4	38	33,6	49,6
5 ou 6	15	13,3	62,8
7 ou 8	27	23,9	86,7
9 ou 10	11	9,7	96,5
11 ou mais	4	3,5	100,0
Total	113	100,0	-

Obviamente, esta tabela é muito mais informativa, pois dá a indicação mais clara da distribuição verificada na amostra. Outras melhorias seriam possíveis, como, por exemplo, reunir as últimas duas faixas, em uma só, indicada por ‘9 ou mais’, dentre outras possibilidades.

4.6. Uso de tabelas na análise exploratória preliminar

No Capítulo 2, indicamos a necessidade de realização de uma etapa preparatória preliminar da planilha para efeito de análise, que é requisito indispensável para ‘limpar’ os dados de erros e problemas que podem vir a comprometer a análise adequada dos dados. A extração de tabelas pode ser de grande utilidade nesta etapa, em conjunto com a análise gráfica, que veremos no capítulo seguinte.

As saídas dos softwares também dão a indicação de como os dados se distribuem. As tabelas permitem a avaliação de três características relevantes dos dados, que são: a indicação a respeito da concentração dos dados; a indicação de como os dados se dispersam entre as diferentes possibilidades de manifestação; e a informação a respeito do formato da distribuição.

Sobre o primeiro aspecto, é fácil observar nas indicações de frequências e percentuais em que categorias, números ou intervalos os dados estão mais concentrados. Por exemplo, na tabela do exemplo 5 é possível verificar que, na variável sob análise, os dados estão mais concentrados nos números 1 e 2, embora não seja uma concentração muito intensa, comparativamente aos demais casos.

O fato de não termos no exemplo indicado (exemplo 5) uma concentração maior, proporcionalmente aos demais números, dá a indicação de que os dados da variável possuem um grau razoavelmente grande de dispersão, ou seja,

há uma maior heterogeneidade das posições dos respondentes. A tabela bem construída indica, adicionalmente, os problemas potenciais com dados que são tão dispersos em relação aos demais que podem ser entendidos como discrepantes. No exemplo 5 não temos casos assim, porém no exemplo do exercício resolvido é possível verificar na primeira tabela gerada dois dados que praticamente se isolam dos demais. A consequência desta ocorrência é evidente na própria tabela, uma vez que estes dois valores geraram uma amplitude muito grande, e na geração dos intervalos para inserção de dados tivemos dois intervalos sem nenhuma observação, e dois intervalos com apenas uma observação cada. Isto sugere que a análise dos dados provavelmente será melhor se estas duas observações forem analisadas em separado.

Um terceiro aspecto que a tabela permite verificar é o formato, embora haja maiores limitações neste aspecto. De fato, verificando as frequências ou percentuais na perspectiva de uma curva, é possível verificar se a curva está concentrada mais no meio, caindo para cima ou para baixo (que constituiria algo esperado em várias situações em que valores intermediários concentram mais observações – ver tabela 4.4 para uma aproximação), ou se a concentração é maior em algum dos dois extremos (ou em ambos) (como seria a primeira extração da tabela do exercício resolvido).

4.7. Resumo

A finalidade do capítulo foi desenvolver competências para a análise e construção de tabelas de dados. Os principais pontos do capítulo foram os seguintes:

- Uma tabela é uma matriz de linhas e colunas que contém informações que sintetizam dados de forma não discursiva, por meio de números que são a fonte principal de informação;
- Embora não haja uma regra universal para construção de tabelas, são desejáveis alguns elementos, a saber: seu título, seu conteúdo propriamente dito, e as notas explicativas a respeito de detalhes do conteúdo;
- No corpo da tabela em si, existem três elementos centrais, a saber: o cabeçalho, logo na primeira linha e explicando as colunas; a coluna indicadora, que é a primeira de todas, e informa sobre o conteúdo das linhas; e as células, que são as unidades de espaço que contêm as informações sobre os dados;

- A tabela pode ser do tipo simples, quando expõe resultados de apenas uma variável, e pode ser de dupla entrada, quando apresenta resultados de duas variáveis;
- No preenchimento da tabela, temos que considerar os tipos de variáveis. Assim, para as variáveis categóricas e algumas das quantitativas discretas o procedimento consiste em extrair as frequências mais os respectivos percentuais;
- Para algumas variáveis quantitativas discretas e para as variáveis contínuas (caso haja muitas observações de valores distintos), devemos organizar as tabelas por meio da agregação dos valores em intervalos de frequências, e em seguida extrair as tabelas correspondentes;
- Há um conjunto de passos e técnicas para construção de tabelas baseadas em intervalos numéricos, porém, salvo no caso de aderência a algum conjunto de regras especificado, os resultados deverão ser julgados à luz do bom senso e do propósito de pesquisa;
- A construção de tabelas permite, além da análise dos dados propriamente dita, a avaliação preliminar dos dados, com a indicação preliminar sobre a concentração dos dados, sobre sua dispersão, e sobre o formato da distribuição.

Exercícios

1. Utilizando o R, construa as tabelas das variáveis do Apêndice D (das que não foram apresentados nos exemplos).
2. Considere as notas de 30 alunos de uma disciplina de Administração de marketing, indicadas abaixo. Utilize o passo a passo indicado no item 4.5 e construa uma tabela correspondente aos dados. Analise exploratoriamente a tabela, sabendo que, nesta faculdade, alunos com nota abaixo de 7 são reprovados, e alunos com nota acima de 9,5 recebem certificado de distinção. Se for necessário, faça ajustes para melhorar a exposição dos resultados.

4,2	7,8	8,7	9,2	6,4
8,7	7,5	7,5	9,6	7,2
7,8	8,1	9,1	7,8	9,1
7,2	6,2	7,0	7,4	3,2
10,0	6,0	9,2	5,6	6,5
7,5	6,7	7,0	9,8	9,0

3. Um gerente de recursos humanos decidiu testar os efeitos de uma promessa de benefício salarial sobre a performance de trabalhadores do setor de conferência de peças de uma linha de produção de confecções. Considerando as metas semanais fixas e uniformes para todos os funcionários, foi prometida uma gratificação proporcional ao aumento da produção em relação à meta.

Os resultados para os 45 funcionários estão indicados abaixo. Utilize o passo a passo indicado no item 4.5 e construa uma tabela correspondente aos dados. Analise exploratoriamente a tabela, sabendo que os funcionários mais produtivos podem receber uma promoção de carreira, e os que são muito pouco motivados podem ser transferidos para outro setor ou demitidos.

0,22	0,14	0,32	0,99	0,24	0,61	0,61	0,71	0,94
0,18	0,28	0,77	0,83	0,92	0,33	0,16	0,94	0,24
0,40	0,03	0,63	0,18	0,54	0,81	0,96	0,60	0,24
0,53	0,03	0,78	0,44	0,83	0,08	0,10	0,53	0,42
0,58	0,02	0,59	0,39	0,01	0,02	0,16	0,72	0,26

CAPÍTULO 5

REPRESENTAÇÃO GRÁFICA DE DADOS

Neste capítulo, temos o complemento do que foi exposto no capítulo anterior, com a apresentação das alternativas de organização e representação gráfica de dados. A finalidade é desenvolver competências para o entendimento e a construção de diferentes formas de visualização de conjuntos de dados. Nossa análise ainda terá foco univariado, de modo que o que apresento é apenas uma parte (e pequena) do conteúdo sobre o tema, pois temos ainda as alternativas de representação gráfica conjunta de duas ou de até mais variáveis.

Na organização do capítulo, apresento no primeiro momento os conteúdos conceituais, e em seguida a indicação dos principais elementos de um gráfico. No terceiro, foco está sobre as diferentes alternativas de gráficos, e, ao final, enfoco a utilização de gráficos na análise exploratória preliminar de dados. Novamente aqui utilizo os dados do Apêndice D como suporte nas exemplificações.

Ao final deste capítulo o leitor deverá estar apto a responder às seguintes questões:

- O que é um gráfico? Quais os elementos fundamentais da construção de um gráfico para a análise univariada?
- Como construir gráficos? Quais as alternativas existentes e como os pacotes estatísticos informatizados contribuem para sua construção?
- Como um gráfico pode ser utilizado na análise exploratória preliminar de uma planilha de dados?

5.1. Organização gráfica de dados

Entenderemos por gráficos a disposição dos dados agregados na forma de figuras, sejam estas as figuras geométricas convencionais (como linhas, pontos, e figuras diversas), sejam ilustrações do tipo cartográficas (como mapas).

Há um número muito grande de possibilidades de gráficos nos softwares estatísticos, e além destes diversos outros são possíveis, a depender da criatividade e da disposição dos pesquisadores e interessados sobre no assunto. Por esta razão, foi preciso selecionar o conteúdo sobre o assunto para ser exposto aqui no alinhamento do que interesse a este material. Para este capítulo, nossa exposição estará sobre os conteúdos gerais sobre gráficos, e sobre a exploração dos gráficos de análise univariada. Outros gráficos serão expostos ao longo dos próximos capítulos.

5.2. Elementos gerais

Da mesma forma que na disposição de dados em tabelas, os gráficos também possuem um conjunto de elementos ‘formais’, e de normas de construção. As regras adotadas por instituições específicas variam, de modo que também aqui não tive a preocupação de seguir normas assim determinadas. O pesquisador deverá, portanto, atentar para as possíveis regras que devem ser seguidas nas pesquisas que forem empreender.

Independente de normas específicas, temos em gráficos um conjunto de elementos fundamentais que estão presentes em qualquer enquadramento de normas proposto. Assim, conforme indicado, o gráfico é uma representação visual, que tem na figura de exposição seu núcleo central de informação. No entanto, assim como na tabela, os gráficos também possuem um título que o identifica, e algumas notas que especificam informações relevantes a respeito do conteúdo exposto no gráfico.

A diferença central em relação à tabela diz respeito às notas. Como vimos, na tabela as notas vêm abaixo e fora do corpo da tabela; já nos gráficos, as notas podem vir em qualquer outro local, inclusive acima ou dentro das próprias figuras representadas, ou seja, as notas, como legendas, fontes, indicações de valores etc. podem ser parte integral da própria figura do gráfico.

Na figura em si, além dos elementos de informação de notas explicativas, as opções são as mais variadas. No entanto, na grande maioria das opções gráficas temos os seguintes elementos centrais:

- Eixos: são as linhas ‘horizontal’ e ‘vertical’ que representam pelo menos duas informações centrais, que são as categorias, números ou faixas numéricas de cada variável, e os valores ou percentuais (em representações para e dimensões, acrescenta-se mais um eixo);

- Elementos indicadores: são as figuras ou representações que aparecem ao lado (ou acima ou abaixo) do gráfico são usadas para indicar as categorias, valores ou faixas numéricas de cada uma das variáveis;
- Códigos ou legendas: são as indicações dos significados de indicações específicas do gráfico, como cores, siglas, símbolos etc.;
- Números: são as indicações numéricas de frequências ou percentuais relativos às variáveis.

Há gráficos que permitem visualizar bem conjuntos com três ou mais variáveis, porém são muito mais comuns as representações bivariada e univariada. Para este capítulo, tratarei de alguns gráficos de análise univariada, e na parte 3 veremos alguns gráficos de análise bivariada. Em qualquer das situações, há um componente estético a ser levado em conta, além das disponibilidades computacionais envolvidas. Por resta razão, na representação gráfica costumamos ter maior flexibilidade com relação às normas e componentes específicos. No item seguinte e nos próximos capítulos teremos ilustrações que reafirmam este entendimento.

5.3. Construindo gráficos

Para este manuscrito utilizamos os recursos gráficos dos softwares (enfatizamos especialmente o R, que, para análise univariada, tem ferramentas gráficas úteis especificamente para variáveis categóricas, e outras para variáveis quantitativas). Para o primeiro caso, apresentaremos os gráficos de barras, de linha e de pizza; já para as variáveis quantitativas apresentaremos o histograma e os gráficos de ‘ramo e folha’. Nos subitens seguintes, apresentamos os tipos juntamente com as rotinas dos softwares mais exemplos ilustrativos, principalmente aqueles baseados na planilha do Apêndice D.

5.3.1. Gráficos para variáveis categóricas

Como já indicado, apresentarei aqui os gráficos de barras, de linhas e de pizza. Embora sejam gráficos mais usados em variáveis categóricas, cada gráfico pode, eventualmente, ser utilizado também em variáveis quantitativas. Vejamos cada um deles.

– Gráfico de barras

Consiste na representação visual orientada em que as categorias das variáveis (e eventualmente valores ou intervalos) são representadas na forma de retângulos, que se dispõem em um dos eixos do gráfico, e sua área corresponde à medida de frequência ou de percentuais (em ambos os casos tanto de valores individuais quanto acumulados).

No R, pode-se gerar o gráfico de barras utilizando o comando *barplot()*. Ver exemplo abaixo que foi feito utilizando os dados do Apêndice D:

```
library(openxlsx) #chamando a biblioteca para ler a base de dados
dados=read.xlsx("dados.xlsx")
attach(dados)

barplot(Sem) #comando para gerar o gráfico de barras do semestre
dos alunos
```

O comando acima resultará em um gráfico em que o eixo y terá os semestres e as frequências ficarão no eixo x. Consideramos mais interessante trabalhar com os percentuais, inverter a ordem dos eixos para que o gráfico fique legível e incluir rótulos. Portanto, utilizaremos o seguinte comando:

```
if (!require(ggplot2, quietly = TRUE)) install.packages("ggplot2")
library(ggplot2) #instalar pacote

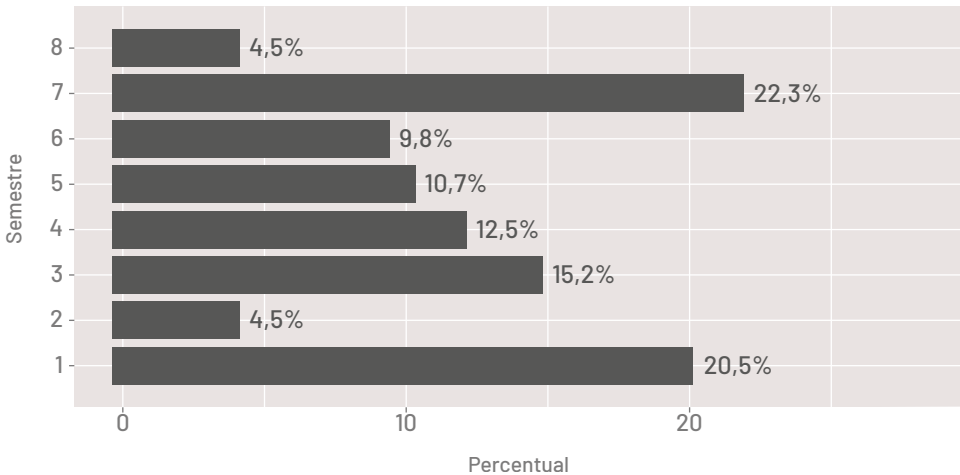
df_sem = as.data.frame(table(dados$Sem))
colnames(df_sem) = c("Semestre", "Quantidade") #separar o semestre
como dataframe

df_sem$Percentual = (df_sem$Quantidade/sum(df_sem$Quantidade))*100
#Percentuais

plot=ggplot(df_sem,aes(Percentual,Semestre)) +
  geom_bar(stat = "identity") +
  geom_text(aes(label = paste0(round(Percentual, 1), "%")),
            hjust = -0.2, size = 4) +
  xlim(0, max(df_sem$Percentual) + 5)
plot
```

Na tela de *output*, o gráfico ficará em uma apresentação padrão em termos de fonte das letras e de cores. Mas há diversas possibilidades de edição, que se procede por meio da inclusão de argumentos. Tomando por base a planilha do Apêndice D, fizemos o gráfico de barras para a variável ‘semestre do curso’, com apresentação dos valores em percentuais absolutos, conforme indicado no Gráfico 5.1.

Gráfico 5.1 – Gráfico de barras relativo ao semestre dos estudantes



No JASP, iremos em *descriptives* → *descriptives statistics* → arraste “Sem” para o campo *variables*. Em seguida, vamos na função *basic plots* → *distribution plots*.

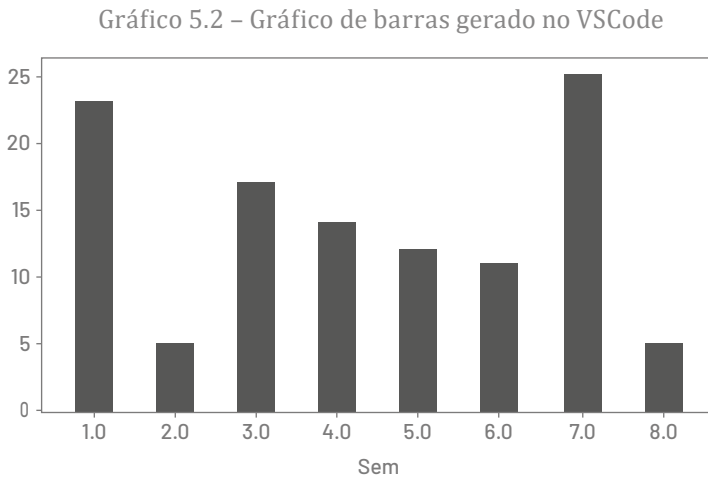
Em Python, pode-se gerar o gráfico de barras utilizando uma função do pacote *matplotlib*. Para tanto, instale o pacote digitando no terminal *pip install matplotlib* e, na sequência, insira os comandos abaixo:

```
import pandas as pd
import matplotlib.pyplot as plt

df = pd.read_excel(CAMINHO DO ARQUIVO COM BARRAS INVERTIDAS',
decimal=',')
print(df)

#Gráfico de barras
plt.figure(figsize=(8,6))
df['Sem'].value_counts().sort_index().plot(kind='bar',
color='steelblue')
plt.xticks(rotation=0)
plt.show()
```

Este é o gráfico que deverá ser exibido (ver Gráfico 5.2):



Podemos também incluir rótulos nas barras. Para tanto, incluiremos no comando mais duas linhas de código. Segue abaixo o comando completo:

```
import pandas as pd
import matplotlib.pyplot as plt

df = pd.read_excel(CAMINHO DO ARQUIVO COM BARRA INVERTIDA',
decimal=',')
print(df)

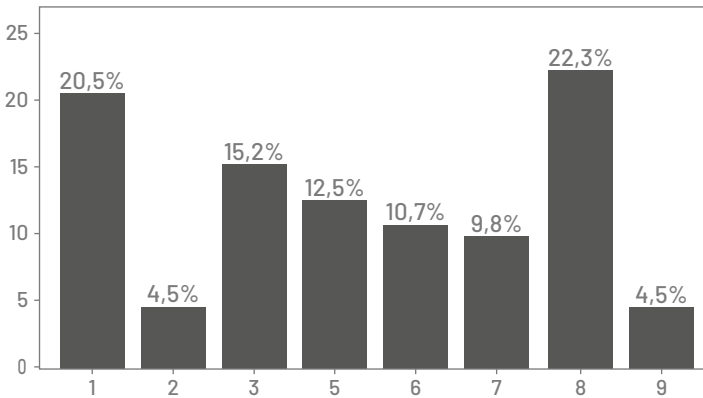
contagem = df['Sem'].value_counts(normalize=True).sort_index()*100

#Gráfico de barras
plt.figure(figsize=(8,6))
barras = plt.bar(contagem.index, contagem.values,
color='steelblue')

#Rótulos
for bar, valor in zip(barras, contagem.values):
    plt.text(bar.get_x() + bar.get_width() / 2, bar.get_height() +
1, f"{valor:.1f}%", ha='center', fontsize=12)
plt.xticks(rotation=0)
plt.ylim(0, 30)
plt.show()
```

O Gráfico 5.3 ilustra qual deve ser o *output*.

Gráfico 5.3 – Gráfico de barras com rótulos gerado no VSCode



No software Excel (tomando por base a versão 2010), o procedimento consiste em selecionar a variável em uma coluna ou linha, e no menu ‘inserir’ marcar a opção ‘Barras’, na qual serão indicadas várias opções.

É fácil ver, pelos gráficos anteriormente mostrados, que temos uma visualização de como se distribuem os dados, em um complemento ao que foi apresentado na tabela. O gráfico e a respectiva tabela podem ou não ser apresentados em conjunto, a depender da circunstância da pesquisa. Se vierem juntos, teremos, além da indicação de valores, mais informações que permitem compreender melhor os dados, o que é a finalidade da análise exploratória de dados.

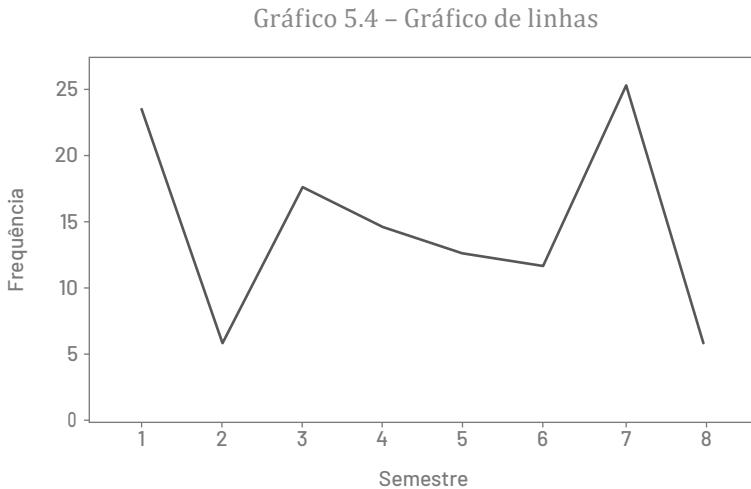
– Gráfico de linhas

O gráfico de linhas é aquele em que primeiramente são marcados os pontos no plano de eixos, indicando cada categoria da variável e as respectivas quantidades, e em seguida estes pontos são ligados por segmentos de reta. O gráfico de linhas é semelhante ao gráfico de barras, ou seja, as categorias (e eventualmente valores ou intervalos) são representadas no eixo horizontal, e as quantidades (frequências ou percentuais) são representadas no eixo vertical; no entanto, a única figura geométrica que aparece é a linha que liga os pontos do gráfico.

Assim como no caso dos gráficos de barras, é possível gear o gráfico de linhas no R e no VSCode. No R, utiliza-se a função `plot` e especifica-se o tipo de gráfico, como demonstrado a seguir:

```
plot(table(Sem), type='l', xlab="Semestre", ylab="Frequência")
```

O Gráfico 5.4 ilustra o *output*.



No JASP, iremos em *descriptives* → *plot builder* → arraste “Sem” para o campo “X-Axis Variable”. Em seguida, vamos na função *amounts* (*count and sum*) → *count line*.

Em Python, continuaremos utilizando o pacote *matplotlib* e incluiremos os seguintes comandos, que deverão resultar em um gráfico semelhante ao 5.4 demonstrado anteriormente.

```
frequencias = df['Sem'].value_counts().sort_index()

plt.plot(frequencias.index, frequencias.values, marker='o',
color='black')
plt.show()
```

Deixamos ao leitor o desafio de verificar os procedimentos para o Excel. Particularmente neste último, temos uma grande variedade de opções de edição que, em grande medida, torna o Excel preferível ao R ou ao VSCode.

O leitor pode notar que o gráfico de barras e o gráfico de linhas parecem ter o mesmo apelo visual, o que é natural. Na verdade, qualquer um poderia facilmente gerar o outro com alguns manuseios de desenho geométrico. Cabe ao pesquisador então decidir qual dos dois é mais apropriado para sua pesquisa.

- Gráfico de pizza

Consiste na representação visual de conjuntos de dados por meio de um círculo, que é então dividido em partes a partir do seu centro. Cada parte do círculo representa uma categoria (ou quantidade ou faixa) e suas áreas são proporcionais às respectivas frequências ou percentuais absolutos (embora não seja impossível a construção, não são comuns os gráficos de pizza para quantidades acumuladas).

No R, para a construção do gráfico a partir da planilha podemos utilizar os comandos detalhados abaixo. Utilizaremos para este exemplo a variável categórica *sexo*.

```
pie(table(Sex), labels = paste(names(table(Sex)), "\n",
round(table(Sex) / sum(table(Sex)) * 100, 1), "%"))
```

A saída do *output* será um gráfico nas especificações indicadas. Também aqui, fica como exercício a verificação das opções da construção do gráfico no Excel.

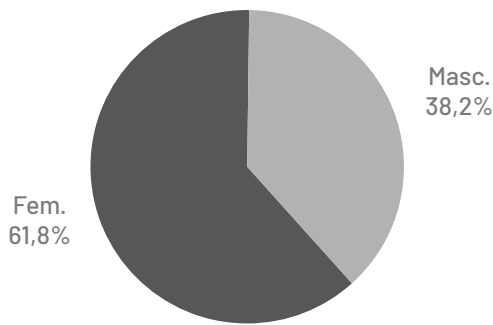
No JASP, iremos em *descriptives* → *descriptives statistics* → arraste “Sex” para o campo *variables*. Em seguida, vamos na função *basic plots* → *pie charts*.

Em Python, podemos utilizar o seguinte comando:

```
plt.pie(df['Sex'].value_counts(), labels=[f'{label}\n{percent:.1f}%\' for label, percent in zip(df['Sex'].value_counts().index, df['Sex'].value_counts() / df['Sex'].value_counts().sum() * 100)], startangle=90)
plt.show()
```

Em seguida, e a critério do pesquisador, podem ser procedidas as edições que se fizerem necessárias e convenientes. No Gráfico 5.5, temos um exemplo para a variável *sexo*, já após a edição realizada. A visualização da figura em conjunto com os valores percentuais (que optamos por acrescentar) permite uma compreensão mais clara dos dados que devem vir apresentados na tabela.

Gráfico 5.5 – gráfico de pizza para a variável sexo



A decisão pela escolha de um ou outro destes gráficos cabe somente ao pesquisador, que pode, ou não, ter condicionamentos de regras específicas de sua instituição de trabalho ou estudo. O treinamento na produção de gráficos também contribui para decisão de qual tipo de gráfico e qual a edição são mais apropriados para a análise que se estiver realizando. Mais relevante do que o gráfico em si é sua adequação para compreensão dos dados, principal razão do desenvolvimento de tantas propostas.

5.3.2. Gráficos para variáveis quantitativas

Para variáveis do tipo quantitativo, os principais gráficos são o histograma e o gráfico de ramo e folha, independente da variável ser discreta ou contínua. A seguir apresentamos cada um destes, seguindo o mesmo procedimento do exemplo anterior, ou seja, apresento os conceitos e elementos gerais, mais as rotinas dos pacotes estatísticos e os exemplos associados.

– O histograma

O histograma é, essencialmente, um gráfico de colunas verticais, desenvolvido nos mesmos moldes do gráfico de barras anteriormente apresentado, porém mantendo na base dos retângulos os intervalos da distribuição de frequência, conforme apresentado na construção de tabelas. Por esta razão, e diferente dos gráficos de barras, o eixo horizontal é sempre numérico, e se constrói na disposição convencional dos números reais (ou seja, em ordem crescente da esquerda para a direita).

Para sua construção, o histograma deve apresentar as bases dos retângulos com extensão proporcional ao tamanho relativo dos intervalos. Assim, nos

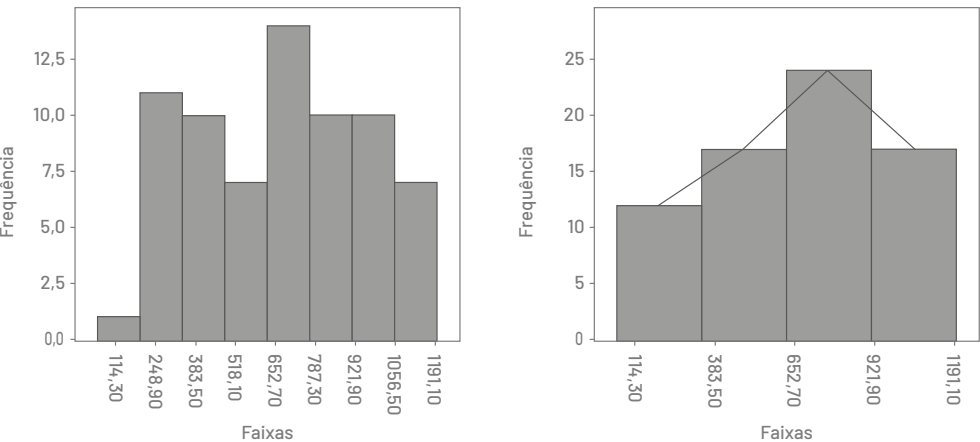
casos em que os intervalos têm tamanhos iguais, os gráficos são praticamente idênticos ao gráfico de barras.

Para variáveis quantitativas contínuas, o histograma convencional é, sem dúvidas, a melhor forma de visualização de dados. No entanto, para os histogramas de variáveis discretas, algumas alternativas derivadas da edição do próprio histograma parecem mais interessantes, como verificaremos nos exemplos.

No R, o histograma é facilmente desenhado pela função `hist()`. No JASP, o caminho inicia em `descriptives` → `plot builder` → arraste, por exemplo, “Ida” para o campo “X-Axis Variable”. Em seguida, vamos na função `distributions` → `histogram`. E, em Python, usamos a função `plt.hist()`. No Excel, é possível construir o histograma a partir da função ‘análise de dados’, marcando a opção histograma (o software gera primeiramente uma tabela, e em seguida basta marcar a tabela e solicitar a apresentação do gráfico de colunas).

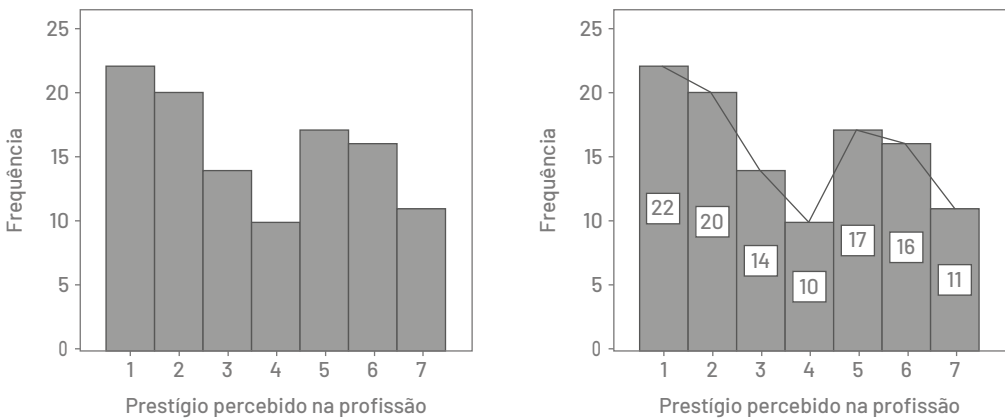
No Gráfico 5.6, temos dois exemplos para variável contínua, com dados relativos à variável ‘peso dos alimentos’, que foi apresentada na discussão sobre a construção de tabelas. Na coluna da esquerda, temos o gráfico com 8 intervalos, e na coluna da direita temos o gráfico com 4 intervalos. Apenas como ilustração das possibilidades de representação, temos também no gráfico da direita a indicação da linha sobre os pontos médios dos retângulos, ou seja, temos o gráfico de linhas sobreposto ao histograma.

Gráfico 5.6 – Construções para a variável ‘peso de alimentos’



O procedimento de edição que segue a cada tipo de gráfico é o mesmo. Fica então a critério do pesquisador definir o melhor caminho. No Gráfico 5.7 temos o histograma para uma variável quantitativa discreta, que é a percepção de prestígio na profissão do Apêndice D. O gráfico da esquerda foi construído a partir do caminho de histogramas e o segundo pelo caminho do código de barras. Observemos que, como temos poucos números possíveis de resposta (de 1 a 7), então cada barra corresponderá a um destes números, não havendo necessidade de outros ajustes nem da aplicação dos passos para a construção de tabelas de frequências aplicadas a variáveis contínuas.

Gráfico 5.7 – Construção para a variável discreta ‘prestígio percebido’



Em geral, as vantagens do histograma, para efeito de análise de dados, são as seguintes:

- Permite visualizar a amplitude dos dados, na medida em que indica o maior e o menor dos valores de dados;
- Indica o nível de concentração dos dados, seus valores mais e os menos frequentes e aqueles que estão no seus ‘entornos’;
- Indica o nível de dispersão do conjunto de dados, na medida em que mostra se os dados estão mais concentrados em torno de um valor específico, ou se estão distribuídos ao longo do eixo que contém os intervalos e valores numéricos;
- Possibilita identificar os dados que estão mais discrepantes em relação aos demais, e as lacunas que separa estes dados daqueles ‘regulares’ em relação ao conjunto;

- Dá a primeira indicação do formato da distribuição de dados, permitindo constatar:
 - » Se verticalmente os dados estão distribuídos de forma achatada ou pontiaguda;
 - » Se horizontalmente estão distribuídos de maneira uniforme, ou na forma de sino (e se estiverem na forma de sino, se este é simétrico para a esquerda e para a direita, ou se há alguma assimetria).

- O gráfico de ramos e folhas

O histograma reflete visualmente o resultado do processo de agregação de dados que vimos na construção de tabelas com intervalos de frequências. Conforme indicado, a construção dos intervalos tem um conjunto de vantagens decorrentes da possibilidade de uma reorganização dos dados, porém tem a desvantagem de perder informação, em razão da agregação dos dados nas faixas delimitadas. Isto é uma desvantagem que alcança o histograma. Mas há outras alternativas de gráficos que viabilizam uma visualização mais detalhada dos dados, principalmente nas análises em conjunto com outras variáveis.

Especificamente para aferição univariada, que é o que fazemos neste capítulo, temos o chamado gráfico de ‘ramos e folhas’, uma opção de análise que permite ao mesmo tempo visualizar o conjunto de dados e ainda acessar algumas de suas especificidades. A construção do gráfico consiste, fundamentalmente, em separar os dados numéricos duas partes, uma que é comum a um subconjunto dos dados, que constituirá o ramo, e a outra que é específica de cada dado, que são as folhas.

Por exemplo, se temos o número 143, podemos adotar o algarismo 1 como o ramo, e o algarismo 43 como a folha. Se tivermos 432 e 439, podemos ter 4 como os ramos dos dois números, e 32 e 39 como folhas, ou então 43 como ramo e 2 e 9 como folhas. Obviamente, precisamos ter números de pelo menos dois dígitos (ainda que algum destes seja parte decimal). Sempre tomando como referência a ordem dos algarismos nos números da esquerda para direita, nossa decisão central será definir qual ou quais algarismos serão fixados nos ramos. Em seguida, os ramos são dispostos, ordenadamente, um abaixo do outro, e ao seu lado (normalmente à direita, mas não necessariamente) são posicionadas as folhas, tantas quantas existam e independente de repetição.

Na medida em que vão sendo desenhados/escritos os ramos e as folhas, a figura vai emergindo, definindo o formato da distribuição dos dados, de forma assemelhada ao histograma, porém com a indicação numérica, e não de uma figura geométrica. Por isto, dizemos que o gráfico de ramos e folhas mantém informações sobre os dados, além de manter a possibilidade de visualização.

A indicação de quais são os dados dos ramos depende da quantidade de dados e das magnitudes presentes. Assim, por exemplo, se temos muitos dados com valores em centenas (entre 100 e 999), então usamos ou os algarismos de centenas ou de dezenas e centenas em conjunto. Vejamos um exemplo:

Exemplo: sejam os dados no conjunto {123, 200, 454, 456, 323, 595, 127, 543, 987, 390, 312, 999, 500, 465, 900, 334, 540, 410, 456 701, 686, 699, 743}. Observando a sequência, vemos que apenas os algarismos das centenas se repetem, não se repetindo com regularidade os algarismos de dezenas. Portanto, podemos tomar os algarismos das centenas e fixar como ramos, e em seguida colocar ao lado de cada ramo os complementos dos números, em ordem crescente. Teríamos o resultado indicado abaixo. É evidente que este gráfico permite boa indicação de como estão os dados, podendo-se verificar que há uma concentração maior em torno dos ramos 3, 4 e 5, principalmente a partir de 312 até o valor de 595:

Ramos	Folhas
1	23 27
2	00
3	12 23 34 90
4	10 54 56 56 65
5	00, 40 43 95
6	68 99
7	01 43
8	
9	00 99

Consideremos agora o conjunto {221, 273, 245, 256, 234, 271, 272, 280 272, 272, 285, 286, 267, 266, 265, 270, 278, 279, 280, 286, 278, 289, 299, 298, 291}. É fácil ver que o algarismo 2 da casa das centenas é fixado, e variam os algarismos das dezenas e unidades. Para este caso, parece mais apropriado orga-

nizar os ramos a partir da centena e, dentro de cada centena, das dezenas; as folhas são melhor organizadas em torno das unidades, como indicado abaixo. Novamente aqui, é fácil ‘ver’ no gráfico que os dados estão mais concentrados na parte de cima, principalmente a partir de 265 até o valor de 299.

Ramos	Folhas
21	1
22	
23	4
24	5
25	6 6
26	5 6 7
27	1 2 2 2 3 0 8 8 9
28	0 0 5 6 6 9
29	1 8 9

Da mesma forma, para dados com maior frequência em dezenas, milhares etc., a disposição dos valores é que indicará a forma de construção do gráfico. Em casos de termos muitos dados, é possível que sejam definidas regras com número fixo de ramos, para evitar um comprimento muito longo da dimensão vertical do gráfico (ou seja, para evitar uma grande ‘profundidade’ do gráfico). Neste caso, os números são organizados de acordo com os algoritmos computacionais, que não cabe aqui sua explicitação.

A falta de uma regra específica de definição, e o fato de eventualmente termos muitos dados em conjunto, pode dificultar o processo de construção do gráfico de ramos e folha. Para nosso caso, tanto a linguagem Python quanto o R dispõem de algoritmos próprios que permitem o desenho dos gráficos, o que dispensa maiores preocupações quanto às decisões indicadas. No R, por exemplo, utiliza-se o comando `stem()` e indica-se entre parênteses o vetor de dados. O mesmo procedimento é realizado no JASP. Ou seja, habilitamos o R console e utilizamos o comando `stem()`, indicando entre parênteses o vetor de dados. Já em Python, podemos utilizar a biblioteca `stemgraphic` e utilizar a função `stemgraphic.stem_graphic()`.

Na tela de ‘output’ é apresentado o gráfico de ramos e folhas. Dependendo do número e da estrutura dos dados, o algoritmo pode fazer alguns ajustes automáticos, n,o sentido de aperfeiçoar a apresentação do gráfico. No Gráfico 5.8 temos um gráfico de ramos e folhas da variável ‘peso de alimentos’, que temos utilizado desde a apresentação da tabela de intervalos.

Gráfico 5.8 – Ramo e folha da variável peso de alimentos

Frequência	Ramo	Folha
3	0.	111
16	0.	2222222233333333
10	0.	4444444445
16	0.	6666666667777777
11	0.	88888888999
14	1.	0000000011111

Largura do ramo: 1000,00
Cada folha: 1 case(s)

No gráfico indicado, cabem algumas explicações. Primeiramente, tendo em vista que tínhamos valores concentrados em 3 e 4 algarismos, o algoritmo do software adotou a referência do último dividido por 1000, e em seguida considerou somente os valores resultantes da primeira casa decimal (ou seja, desconsiderou todos os valores numéricos originais de dezenas, unidades, e todas as casas decimais). Assim, o valor de peso 114,3 (que era o menor valor) quando dividido por 1000 ficou 0,1133, e o algoritmo considerou somente 0,1. Nestes termos, foram definidos 6 ramos, considerando os algarismos das unidades para as folhas que foram assim reunidas: 0,0 e 0,1; 0,2 e 0,3; 0,4 e 0,5; 0,6 e 0,7; 0,8 e 0,9; 1,0 e 1,1 (ou seja, tivemos 6 ramos para o conjunto de 70 dados). Observe que o gráfico gerado traz também a frequência de cada um dos ramos, que complementa a análise. Evidentemente, é sempre necessário que o pesquisador compreenda as especificações dos resultados gerados, e avalie a adequação do resultado ao seu interesse de pesquisa.

Em geral, as vantagens do gráfico de ramos e folhas são as mesmas que temos no histograma, ou seja, este gráfico:

- Permite visualizar a amplitude dos dados;
- Indica o nível de concentração dos dados;

- Indica o nível de dispersão do conjunto de dados;
- Possibilita identificar os dados que estão mais discrepantes e as ‘falhas’ da distribuição;
- Dá a primeira indicação do formato da distribuição tanto horizontal quanto verticalmente.

5.4. Gráficos para análise exploratória preliminar de dados

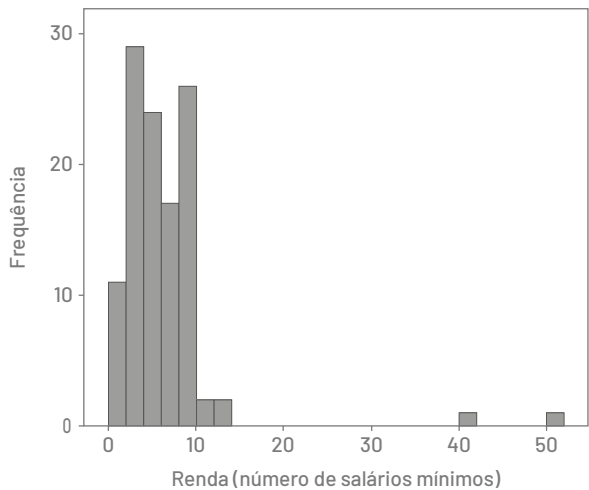
Conforme indicado na Parte I, a primeira tarefa para uma análise consistente de dados consiste em organizar o conjunto de dados em uma planilha e ‘limpar’ desta planilha todos os potenciais problemas. Na análise preliminar indicada no Capítulo 2 nosso foco foi sobre os problemas de tabulação e de dados faltantes. A análise por tabelas permite ainda verificar a existência de valores extremos e de formato. Já a análise gráfica permite ao pesquisador ver cada um destes detalhes.

Embora a análise exploratória preliminar seja tanto para as variáveis categóricas quanto para as quantitativas, a maior utilidade é na análise das variáveis quantitativas. Para este caso, por meio da análise gráfica temos a possibilidade de identificar o formato da distribuição dos dados, uma informação que tem importância central nas técnicas de análise bivariada e multivariada. Adicionalmente, é possível apontar, mesmo sem aplicação de qualquer outro critério de análise, se os dados estão ou não de acordo com as expectativas pré-estabelecidas.

A aplicação da análise gráfica tem uma regra central: utilizar o máximo possível de gráficos, desde que cabido para os tipos de variáveis sob análise. Vejamos como isto ocorre na variável ‘renda’. Como se trata de uma variável quantitativa, os gráficos que aplicaremos são o histograma e o de ramos e folhas.

Tomemos primeiramente o histograma, que está indicado no Gráfico 5.9. É fácil observar que o gráfico aparenta uma forma de sino, porém tem dois casos extremos à direita que comprometem sua simetria. Uma primeira sinalização é a de que estes dois valores podem prejudicar qualquer análise posterior, pois enviesam a estrutura de dados. Uma recomendação possível é que sua análise seja feita em paralelo, separada da análise do conjunto de dados.

Gráfico 5.9 – Histograma da variável renda



Observemos agora o Gráfico 5.10, de ramos e folhas. Como os valores eram discretos, alguns com um algarismo e outros com dois, os softwares podem acrescentar o decimal nulo (,0), o que permite a construção dos gráficos. Assim, as 11 observações com 1 salário são indicadas como 1,0, de modo que o 1 será o ramo, e os 11 zeros serão as folhas. Para evitar uma extensão despropositada de 50 linhas, a maioria das quais sem qualquer observação, ao final o software indica os dois valores maiores (40 e 50 salários). Observamos ainda aqui a semelhança do formato com o histograma, e, além disto, é reiterada a conclusão de que os dois valores maiores estão totalmente fora dos padrões dos demais, e requerem uma análise paralela.

Gráfico 5.10 – Ramos e folhas da variável renda

Frequência	Ramos	Folhas
11	1,	0000000000
7	2,	0000000
22	3,	00000000000000000000
16	4,	0000000000000000
8	5,	00000000
7	6,	0000000
10	7,	000000000
17	8,	0000000000000000
9	9,	000000000
2	10,	00

0	11,	
2	12,	00
2		Extremos (maiores ou iguais que 40)

Os dois gráficos permitem ainda observar o seguinte:

- Quanto à concentração, vemos que os valores estão bem concentrados entre 1 e 10 salários (o que parece uma indicação razoável para a realidade geral das pessoas), permitindo identificar a maior concentração nas proximidades de 3 e de 8 salários;
- Quanto à dispersão, é evidente que, entre 1 e 10 salários, temos uma dispersão razoavelmente grande, o que indica a heterogeneidade de valores salariais (neste intervalo);
- Quanto ao formato, na disposição do histograma, principalmente, permite observar uma aproximação com o formato de sino, mas com assimetria à direita. Buscando uma perspectiva vertical para as folhas do segundo gráfico, reiteramos esta observação, com a indicação adicional de que o formato de sino é menos pronunciado do que o histograma indica.

5.5. Resumo

Este capítulo teve por finalidade discorrer sobre a análise e a construção de gráficos, em complemento ao estudo realizado sobre tabelas. Os principais pontos do capítulo foram os seguintes:

- Entenderemos por gráficos a disposição dos dados agregados na forma de figuras, sejam figuras geométricas convencionais, sejam ilustrações do tipo cartográficas;
- Da mesma forma que na disposição de dados em tabelas, os gráficos também possuem um conjunto de elementos centrais, que são: o título; a figura central; e algumas notas que especificam informações relevantes;
- Além destes elementos, o gráfico apresenta ainda mais alguns elementos, a saber: os eixos; as figuras indicadoras; os elementos indicadores e os números representativos de frequências ou percentuais;
- Há gráficos que permitem visualizar bem conjuntos com três ou mais variáveis, porém são muito mais comuns as representações bivariadas e

univariada. Para este capítulo foram tratados os gráficos para avaliação univariada, tanto para variáveis categóricas quanto quantitativas;

- Para as variáveis categóricas, os gráficos apresentados foram os seguintes: barras, linhas e pizza. Eventualmente, estes gráficos também são úteis para variáveis quantitativas;
- Para as variáveis quantitativas foram apresentados o histograma e o gráfico de ramos e folhas. Ambos permitem a visualização da concentração da variável, além da dispersão de seus dados e do formato da distribuição;
- Por meio destas alternativas de uso, os gráficos viabilizam tanto a análise de resultados das variáveis no interesse da pesquisa, assim como a análise preliminar dos dados.

Exercícios

1. Construa em Python, no JASP e R os gráficos a partir dos dados consolidados nas tabelas do capítulo 4.
2. Construa em Python, no JASP e R os gráficos apropriados para cada uma das variáveis do Apêndice D, e, nas variáveis quantitativas, desenvolva a análise exploratória das saídas do software.
3. Construa os gráficos correspondentes aos exercícios 2 e 3 do capítulo 4 e desenvolva a análise exploratória.

CAPÍTULO 6

MEDIDAS DESCRITIVAS DE POSIÇÃO

Nos capítulos anteriores tratamos da apresentação de dados, porém nos preocupamos apenas com contagens e porcentagens. Neste e nos próximos capítulos nossa preocupação é diferenciada, pois a intenção é indicar medidas que descrevem os dados.

No primeiro momento, trabalharemos as medidas que indicam posições. Fundamentalmente, são 3 medidas de centro clássicas mais relevantes (média aritmética, mediana e moda), e um conjunto de outras medidas posição (os diferentes tipos de médias e os quantis). Deste conjunto, a medida de maior utilização na teoria estatística e na análise de dados em geral é a média, e nos últimos anos vem ganhando maior relevância a mediana e os quantis. Estes serão, portanto, os tópicos de maior desenvolvimento e exemplificação. Ao final do capítulo, e juntamente com o resumo, temos uma tabela de notação de todas as siglas utilizadas.

Ao final deste capítulo o leitor deverá estar apto a responder às seguintes questões:

- O que são medidas de posição e o que as caracterizam?
- O que é a média de um conjunto de dados? Quais as variações existentes e como estas variações são aplicáveis na análise de dados?
- O que é uma mediana e uma moda de uma variável? Quais suas aplicações na análise de dados?
- O que são os quantis? De que forma o conhecimento dos quantis pode ser útil à análise exploratória de dados?

6.1. Definição de medidas de posição

Classicamente, as medidas apresentadas nos manuais de estatística básica são a média, a mediana e a moda, e a denominação utilizada é de ‘medidas de tendência central’. Mais recentemente, vêm sendo estudadas outras medidas,

o que indica necessidade de um primeiro entendimento do que sejam medidas de posição.

Imaginemos uma série de dados de uma variável, e suponhamos o seu posicionamento em uma reta numerada, com os valores dos dados correspondendo a pontos da reta. Quando tomamos ao longo do segmento de reta que contém o conjunto de dados uma medida qualquer que represente alguma característica do conjunto de dados, então estaremos tomando uma medida relativa àquela posição de referência. Por exemplo, se tomamos o ponto do segmento que separa os 10% primeiros valores dos demais 90%, estamos tomando uma medida da posição do 10º percentil. Podemos então tomar uma infinidade de medidas, inclusive o valor mínimo e o valor máximo do conjunto de dados, o valor mais frequente, o valor que seja uma composição do conjunto de dados, dentre outras.

Como as possibilidades são inúmeras, definimos mais formalmente uma medida de posição como aquela que, relativa a um dado conjunto de dados, possui as seguintes características:

- A medida está entre os valores mínimo e máximo (inclusive) do conjunto de dados;
- A multiplicação de um valor constante por todo o conjunto de dados implica na multiplicação do mesmo valor à medida de posição original.

Como indicado, as principais medidas que possuem estas características são a média aritmética, a mediana e a moda, que são medidas de tendência central. Esta denominação vem do fato de os valores resultantes estarem, normalmente, em torno do ponto central do conjunto de dados.

Desde agora é relevante fixar a potencial diferença na captação das medidas, de acordo com a natureza dos dados coletados. Assim, se o conjunto de observações se refere a todo o universo relativo à variável, teremos a **medida populacional**; se temos, por outro lado, um conjunto de valores na forma de uma amostra, teremos então a **medida amostral**. É comum que nosso interesse seja na avaliação da população, e para tanto utilizamos uma medida amostral para estimar a medida populacional.

Temos aqui uma questão relevante, concernente a populações infinitas, ou seja, com aquelas populações em que não temos um número definido de valores. Embora pareça, em uma primeira visualização, uma situação complicada

para se definir uma medida de resumo, na verdade não há complicação alguma quando o tratamento desses conceitos é feito a partir da Teoria das probabilidades (que não será tratada nesse manuscrito). **Aqui nos limitaremos, portanto, ao tratamento particular em que as variáveis representam amostras ou populações finitas.**

6.2. Média

A média é, sem dúvidas, a medida estatística mais conhecida e a mais intuitiva de todas. Não temos somente um tipo de média; pelo contrário, há diversos tipos, dos quais, correntemente, são abordados os seguintes: média aritmética simples e suas derivações; média ponderada; média geométrica; média quadrática; média harmônica e média generalizada.

6.2.1. Média aritmética

Definimos a média de um conjunto de dados como a soma dos valores dos dados dividida pelo total de observações realizadas (Toledo & Ovalle, 1995). Se os dados representam todo o universo, temos então a média aritmética populacional; se forem dados de uma amostra, temos então a média aritmética amostral. Simbolicamente, temos uma representação formal (e quase universal) para cada média, conforme indica o Quadro 6.1.

Assim, para efeito de análise de dados, esta representação costuma ser flexibilizada, cabendo ao pesquisador indicar, quando for o caso, qual média está sendo verificada. Para evitar diferentes usos ao longo do texto, aqui utilizaremos a representação ‘ $Me(X)$ ’ para a média e N para o número de elementos de uma variável genérica $X = \{x_1, x_2, \dots, x_i, \dots, x_N\}$. Sempre que necessário, em cada situação será sinalizada se a média é amostral ou populacional¹⁵.

¹⁵ Na verdade, a grande maioria das verificações de média é feita sobre amostras, tendo em vista a dificuldade eventual (ou mesmo a impossibilidade) de acesso à média da população de algumas variáveis (em especial as variáveis com universo infinito).

Quadro 6.1 – Representações da média aritmética

Média populacional	Média amostral
$\mu = \frac{\sum_{i=1}^N x_i}{N}$ <p>Onde μ (letra grega ‘mi’) representa a média populacional, x_i representa cada valor observado na população, e N representa a quantidade de elementos do universo¹⁶.</p>	$\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$ <p>Onde $\bar{x}$ (x barra) representa a média amostral, representa cada valor observado na amostra e n representa a quantidade de elementos da amostra.</p>

Vejamos então um exemplo:

Exemplo 1: com o intuito de verificar a necessidade de aquisição de um gerador de energia para um condomínio de 280 apartamentos, o síndico tomou uma amostra de 20 apartamentos e verificou seu consumo em quilowatts de um mês. Os resultados estão a seguir indicados.

Consumo={99, 131, 106, 84, 119, 147, 105, 140, 97, 139, 257, 202, 283, 277, 134, 240, 229, 130, 249, 192}

Nestes termos, a média será dada assim:

$$\text{Me}(\text{Consumo}) = \frac{99 + 131 + 106 + \dots + 130 + 249 + 191}{20} \Rightarrow \text{Me}(\text{Consumo}) = 168$$

Ou seja, por este cálculo, as residências consomem, em média, aproximadamente 168 kWt de energia elétrica mensalmente.

A média amostral é usada tanto para fins de análise de conjuntos de dados, como, e principalmente, para ser uma estimativa de qual seja a média populacional. Por exemplo, tomando o valor médio do exemplo 1 como uma estimativa da média dos 280 apartamentos, o decisor tem a possibilidade de desenvolver seu planejamento de forma mais consistente. O exemplo a seguir complementa esta ideia.

Exemplo 2: o gerente de uma fábrica de telha de cerâmica estuda a possibilidade de aquisição de uma máquina de produção, supondo que haverá menos perdas e retrabalho com as telhas produzidas. O vendedor informou que há

16 Por convenção, doravante tomaremos o índice da representação dos dados iniciando sempre em 1 e terminando no número total de observações.

uma máquina que garante aproveitamento de 98% (0,98) das telhas processadas, e, para comparação, o gerente resolveu verificar exploratoriamente a produção de 500 telhas na linha de produção.

Definindo para cada telha sem erro a indicação 1 e para a telha com defeito a indicação 0, verificou na linha que 477 telhas receberam indicação 1 e as demais 0. Neste caso, a média vem da divisão do valor observado de 1's e 0's pelo total de 500 telhas. Teremos então:

$$\text{Me} \left(\begin{array}{c} \text{telha} \end{array} \right) = \frac{\overbrace{1+1+\dots+1}^{477 \text{ vezes}} + \overbrace{0+0+\dots+0}^{23 \text{ vezes}}}{500} \Rightarrow \text{Me} \left(\begin{array}{c} \text{telha} \end{array} \right) = \frac{477}{500} = 0,954$$

Neste segundo exemplo, temos um caso de média menor que 1, que, comparando com um total de 100 telhas, indica que 95,4 telhas estarão sem defeito, ou seja, a eficiência estimada, nesta amostra, é de que o trabalho manual garante aproveitamento de 95,4%, abaixo do que está indicado para a máquina. Neste tipo de situação temos o percentual como representação da média da amostra, com a adoção da convenção apresentada (1 para ausência de defeito e 0 para defeito). Duas observações são destacadas:

- Primeiro, é fácil ver a associação direta entre percentuais e médias. Em verdade, os percentuais das categorias de uma variável são igualmente médias, em relação às demais categorias da mesma variável;
- A linha de produção em uma fábrica produz, teoricamente, uma quantidade infinita de itens, pois supomos que não haja tempo previsto para cessar a produção. Nestes termos, a média de defeitos de produção de um determinado produto é sempre estimada a partir da média amostral, pois nunca teremos acesso ao número infinito de itens produzidos.

Nos *softwares* estatísticos, a média é uma medida que pode ser verificada por diversos procedimentos e caminhos distintos. Apresento aqui os dois caminhos que geram diretamente a média, e nos tópicos seguintes veremos onde a média também pode aparecer em outras análises.

- No Excel (2010), o procedimento consiste em marcar uma célula e escrever o comando '=média()', e entre os parênteses selecionar a linha ou a coluna de dados.

- No R, primeiramente é preciso escrever o vetor contendo os dados observados. Sendo 'x' o vetor, basta escrever `mean(x)`, e imediatamente abaixo surgirá o valor da média.
- No JASP, o caminho é `descriptives` → `descriptive statistics` → arrasta a variável de interesse para o campo `variables` → no campo `statistics`, clica em `mean`.
- Em Python, podemos utilizar a função `df['VARIÁVEL DE INTERESSE'].mean()` (observe que o `df` é o nome que escolhemos para o banco de dados, utilize o nome que definiu para o seu banco de dados), e imediatamente surgirá o valor da média.

Vejam um exemplo no R, para a variável 'Consumo', apresentada no exemplo 1. Para escrever o vetor, precisamos colocar os valores entre parênteses, que vêm junto com a letra `c` (da seguinte forma `c(...)`). Teremos, com a extração direta do R, o seguinte:

```
>Consumo=c(99, 131, 106, 84, 119, 147, 105, 140, 97, 139, 257, 202,
283, 277, 134, 240, 229, 130, 249, 192)
> mean(Consumo)
[1] 168
```

Vejam agora o exemplo em Python com a mesma variável utilizada acima. Criaremos um vetor chamado `consumo` e, na sequência, extrairemos a média:

```
import pandas as pd
dados = pd.DataFrame ({'consumo': [99, 131, 106, 84, 119, 147, 105,
140, 97, 139, 257, 202, 283, 277, 134, 240, 229, 130, 249, 192]})
print(dados['consumo'].mean())
```

Temos duas variações da média aritmética, que são a média ponderada e a média para dados agrupados. Vejam a seguir os detalhes.

6.2.1.1. Média ponderada

A média ponderada é aquela em que os valores dos dados da variável recebem pesos especiais, a depender de critérios preestabelecidos. Neste caso, a média é definida como o somatório do produto dos dados pelos respectivos pesos, dividido pelo somatório dos pesos. Simbolicamente, representaremos a média de uma variável $X = \{x_1, x_2, \dots, x_i, \dots, x_N\}$ ponderada pelos pesos $P = \{p_1, p_2, \dots, p_i, \dots, p_N\}$ por $Mp(X)$ e a calcularemos da seguinte forma:

$$Mp(X) = \frac{\sum_{i=1}^N x_i p_i}{\sum_{i=1}^n p_i}$$

O exemplo a seguir permitirá visualizar de forma mais clara esta fórmula.

Exemplo 3: na seleção dos cursos de uma instituição, é comum as diferentes áreas acadêmicas definirem pesos diferenciados para as disciplinas. Assim, a gestão acadêmica de uma universidade decidiu que os alunos que fizeram prova do Exame Nacional do Ensino Médio que estivessem concorrendo ao curso de Engenharia de Energia teriam pesos 2 para as notas das provas de ‘ciências da natureza e suas tecnologias’ e ‘matemática e suas tecnologias’ e peso 1 nas provas de ‘ciências humanas e suas tecnologias’, ‘redação’ e de ‘linguagens, códigos e suas tecnologias’.

Já para o curso de Administração, os pesos foram os seguintes: 0,5 para ‘ciências da natureza e suas tecnologias’, 1 para ‘matemática e suas tecnologia’, 1,5 para ‘ciências humanas e suas tecnologias’ e 2 para as provas de ‘redação’ e de ‘linguagens, códigos e suas tecnologias’. A seguir temos as notas de quatro alunos, dois de cada curso:

Aluno	Curso	Pontos				
		Ci. Nat.	Matem.	Ci. Hum.	Lingua.	Redação
Aluno 1	Engenharia	640	730	800	500	840
Aluno 2	Engenharia	730	540	840	790	650
Aluno 3	Administração	450	690	890	880	670
Aluno 4	Administração	340	530	800	940	970

Para cada aluno devem ser calculadas as médias aritmética simples e ponderada. Ilustremos o cálculo para o Aluno 1 e para os demais, os resultados já calculados estão indicados na tabela em seguida. Teremos então:

- Média aritmética simples:

$$Me(\text{Aluno 1}) = \frac{640 + 730 + 800 + 500 + 840}{5} \Rightarrow Me = 702$$

- Média ponderada:

$$Mp(\text{Aluno 1}) = \frac{2 \times 640 + 2 \times 730 + 1 \times 800 + 1 \times 500 + 1 \times 840}{2 + 2 + 1 + 1 + 1} \Rightarrow Mp \cong 697$$

Aluno	Curso	Pontos					Me	Mp
		Ci. Nat.	Matem.	Ci. Hum.	Lingua.	Redação		
Aluno 1	Engenharia	640	730	800	500	840	702	697
Aluno 2	Engenharia	730	540	840	790	650	710	689
Aluno 3	Administração	450	690	890	880	670	716	764
Aluno 4	Administração	340	530	800	940	970	716	818

Os resultados das duas últimas colunas mostram a reconfiguração que o peso aplicado sobre as diferentes matérias gera no resultado. Veja que, em Engenharia, o Aluno 1 tem média aritmética menor que a média do Aluno 2; no entanto, sua média ponderada é maior em relação ao Aluno 2. No caso dos Alunos 3 e 4, temos um empate na média aritmética simples, porém na média ponderada o Aluno 4 teve média significativamente mais elevada. Nos dois cursos, as variações ocorrem devido ao fato de os alunos com maior média ponderada possuírem notas mais elevadas nas matérias que têm maior peso relativo.

É fácil ver que a média aritmética simples é uma média ponderada, em que todos os valores possuem peso de ponderação igual a 1. Nos softwares, o procedimento de extração consiste em programar a média indicando os valores e a multiplicação pelos respectivos pesos. Por esta razão, é mais comum o uso do Ms Excel para ponderação, tendo em vista a melhor forma de operacionalizar a replicação de cálculos. Não exibiremos aqui o procedimento de programação, ficando recomendada a verificação como exercício.

6.2.1.2. Média de valores tabulados

Nos dois primeiros casos tivemos a extração da média a partir do conjunto detalhado de valores. Mas é possível ainda extrair as médias dos valores dispostos na forma de tabelas. Para tanto, nosso procedimento é semelhante ao que foi aplicado na extração de média ponderada, com a utilização dos valores de frequências como os pesos da ponderação. Assim, para o caso de tabulação de variáveis discretas sem intervalos, o procedimento consiste em multiplicar cada valor pela respectiva frequência, depois somar todos os valores e em seguida dividir o resultado pela soma das frequências. Tomando então uma variável com valores $X = \{x_1, x_2, \dots, x_i, \dots, x_k\}$ com as respectivas frequências $F = \{f_1, f_2, \dots, f_i, \dots, f_k\}$. Representando a média de valores tabulados por Mt , calcularemos da seguinte forma:

$$Mt(X) = \frac{\sum_{i=1}^k x_i f_i}{\sum_{i=1}^n f_i}$$

A seguir temos um exemplo.

Exemplo 4: No exemplo 5 do capítulo 5 mostramos os dados tabulados da variável ‘Prestígio percebido na profissão’. Aqui, os valores dos escores estão indicados na coluna indicadora, e os valores de frequência estão na segunda coluna. Assim, criamos mais uma coluna para conter o produto entre as frequências. Na última linha, temos os somatórios da frequência e do produto, aí colocados para facilitar os cálculos.

Nível de prestígio	Frequência (f_i)	Produto ($x_i f_i$)
1	22	22
2	20	40
3	14	42
4	10	40
5	17	85
6	16	96
7	11	77
Total		
	$\sum_{i=1}^n f_i = 110$	$\sum_{i=1}^n x_i f_i = 402$

Portanto, a média será dada por:

$$Mt\left(\text{Nível de prestígio}\right) = \frac{\sum_{i=1}^7 x_i f_i}{\sum_{i=1}^7 f_i} = \frac{402}{110} \Rightarrow Mt \cong 3,7$$

Quando temos dados tabulados com intervalos, não podemos aplicar o mesmo procedimento acima, tendo em vista que os valores de referência () não estão especificados como um valor determinado. Para efeito de cálculo, uma alternativa seria adotar qualquer valor dentro de cada intervalo como sua representação e então aplicar o procedimento. Convencionalmente, é mais comum utilizar o valor médio do intervalo, que se obtém somando os extremos e dividindo por 2. Este valor médio será então o valor de referência, a partir do qual será aplicada a fórmula acima. Vejamos um exemplo.

Exemplo 5: Ao longo do item 4.5 foram desenvolvidas duas tabelas de intervalo. Do conjunto de dados indicados na Tabela 4.2, a média aritmética é 651,1. Calculemos então a média para os valores tabulados, inicialmente conforme o resultado da tabela 4.3.

Intervalos de peso	Ponto médio (x_i)	Frequência (f_i)	Produto ($x_i f_i$)
[114,3; 248,9]	181,6	5	908
[248,9; 383,5]	316,2	13	4110,6
[383,5; 518,1]	450,8	10	4508
[518,1; 652,7]	585,4	5	2927
[652,7; 787,3]	720,0	12	8640
[787,3; 921,9]	854,6	9	7691,4
[921,9; 1056,5]	989,2	9	8902,8
[1056,5; 1191,1]	1123,8	7	7866,6
Total	-	$\sum_{i=1}^n f_i = 70$	$\sum_{i=1}^n x_i f_i = 45554,4$

Temos, portanto, uma média de:

$$Mt\left(\text{Peso}\right) = \frac{\sum_{i=1}^8 x_i f_i}{\sum_{i=1}^8 f_i} = \frac{45554,4}{70} \Rightarrow Mt \cong 650,8$$

Observemos que o valor difere muito pouco em relação ao valor dos dados detalhados (651,1). Vejamos agora os resultados para os mesmos dados, organizados na tabela 4.4, e reorganizados a seguir.

Intervalos de peso	Ponto médio (x_i)	Frequência (f_i)	Produto ($x_i f_i$)
[114,3; 383,5]	248,9	18	4480,2
[383,5; 652,7]	518,1	15	7771,5
[652,7; 921,9]	787,3	21	16533,3
[921,9; 1191,1]	1056,05	16	16896,8
Total	-	$\sum_{i=1}^n f_i = 70$	$\sum_{i=1}^n x_i f_i = 45681,8$

Por este procedimento, a média será:

$$Mt(\text{Peso}) = \frac{\sum_{i=1}^8 x_i f_i}{\sum_{i=1}^8 f_i} = \frac{45681,8}{70} \Rightarrow Mt \cong 652,6$$

Novamente aqui, a média resultante é bastante próxima da média do conjunto de valores detalhados. Nos dois casos, temos pequenas diferenças, que podem ser mais ou menos relevantes, dependendo do interesse de pesquisa.

Assim como o procedimento de extração da média ponderada, nos softwares é necessária a realização de procedimentos de programação para cálculo da média de dados tabulados. Novamente aqui, o MsExcel tem melhores funcionalidades para operacionalização, e fica como exercício o treino neste tipo de tratamento de dados.

6.2.2. Propriedades da média

Para qualquer destes modelos de extração da média, teremos um conjunto de regularidades que constituem propriedades matemáticas, que inclusive justificam o uso tão recorrente da média. As quatro principais propriedades estão a seguir indicadas, e sua demonstração remete diretamente às propriedades de somatório apresentadas no Capítulo 3. A demonstração das propriedades fica como exercício e aqui me limitarei a exemplificar cada uma delas.

P1 – A média é maior ou igual ao valor mínimo e menor ou igual ao valor máximo do conjunto de dados. Simbolicamente, teremos:

$x_{(1)} \leq Me(X) \leq x_{(n)}$, onde $x_{(1)} = \text{Min}(X)$ e $x_{(n)} = \text{Max}(X)$, sendo X a variável.

- Exemplo: Se $X = \{x_1 = 8, x_2 = 10, x_3 = 12, x_4 = 30\}$, teremos:

» A média de X , representada por $Me(X)$, vem de:

$$Me(X) = \frac{8 + 10 + 12 + 30}{4} = 15;$$

» Veja que $8 < 15 < 30$

P2 – A soma ou a subtração de um valor constante a todos os dados implica na soma ou na subtração deste mesmo valor à média.

- Exemplo: Se $X = \{x_1 = 8, x_2 = 10, x_3 = 12, x_4 = 30\} \Rightarrow Me(X) = 15$
Teremos:

» $X + 5 = \{x_1 = 8 + 5 = 13, x_2 = 10 + 5 = 15, x_3 = 12 + 5 = 17, x_4 = 30 + 5 = 35\};$

» A média de $X + 5$, representada por $Me(X + 5)$, vem de:

$$Me(X + 5) = \frac{13 + 15 + 17 + 35}{4} = 20;$$

» Ou seja, $Me(X + 5) = Me(X) + 5$

P3 – A multiplicação ou a divisão de um valor constante a todos os dados implica na multiplicação ou na divisão deste mesmo valor à média.

- Exemplo: dos mesmos dados do exemplo anterior, teremos

$$5X = \{x_1 = 5 \cdot 8 = 40, x_2 = 5 \cdot 10 = 50, x_3 = 5 \cdot 12 = 60, x_4 = 5 \cdot 30 = 150\};$$

» A média de $5X$, representada por $Me(5X)$, vem de:

$$Me(5X) = \frac{40 + 50 + 60 + 150}{4} = 75;$$

» Ou seja, $Me(5X) = 5Me(X)$

P4 – O somatório dos desvios das observações em relação à média é nulo, ou

$$\text{seja, } \sum_{i=1}^n dme_i = 0, \text{ onde } dme_i = \left(x_i - Me(X) \right)$$

• *Exemplo: no exemplo de P1, temos que*

» $D = \{d_1 = 8 - 15 = -7, d_2 = 10 - 15 = -5, d_3 = 12 - 15 = -3, d_4 = 30 - 15 = 15\}$

» A média de D , vem de: $Me(D) = \frac{-7 - 5 - 3 + 15}{4} = 0$.

P5 – A soma dos quadrados dos desvios das observações em relação à média é mínima (ou seja, é menor que a soma dos quadrados dos desvios em relação a qualquer outro valor diferentes da média). Simbolicamente teremos:

$$\sum_{i=1}^n d_i^2 = \sum_{i=1}^n (x_i - Me(X))^2 < \sum_{i=1}^n (x_i - c)^2, \text{ para todo } c \neq Me, c \text{ um número real.}$$

• *Exemplo: tomando os valores de referência que temos usado nos últimos exemplos, vimos que a média foi de 15. Verifiquemos então os desvios em relação aos valores 10 e 20, conforme indicado na tabela abaixo.*

Valores	$(x_i - Me)$	$(x_i - Me)^2$	$(x_i - 10)$	$(x_i - 10)^2$	$(x_i - 20)$	$(x_i - 20)^2$
8	-7	49	-4	16	-12	144
10	-5	25	-2	4	-10	100
12	-3	9	0	0	-8	64
30	15	225	18	324	10	100
Soma	0	308	12	344	-20	408

Ou seja, $(x_i - Me)^2 < (x_i - 10)^2$, e $(x_i - Me)^2 < (x_i - 20)^2$

Observemos que as duas propriedades P1 e P3 asseguram as condições para que a média aritmética seja considerada uma medida de posição. Em geral, a média é uma indicação dos valores típicos de um conjunto de dados, que é usado como uma estimativa para caracterização aproximada do conjunto de elementos. Por exemplo, se temos 20 pessoas com diferentes alturas, e temos uma média de 1,70m, podemos tomar esta altura como indicação de uma altura dos indivíduos da amostra, ou seja, 1,70m seria a altura típica do conjunto. Naturalmente, espera-se que haja pessoas com alturas maiores e menores que 1,70 m, mas temos neste valor uma ideia preliminar da altura das pessoas.

No entanto, a média aritmética é suscetível a grandes diferenças de valores, um problema comum e que a prejudica severamente enquanto medida de análise de dados. Um exemplo ilustra este problema em potencial.

Exemplo 6: O gerente de Recursos Humanos de uma construtora decide criar um projeto de endomarketing associado ao envolvimento familiar dos mestres de obras, e, para tanto, realiza um levantamento sobre o número de filhos dos seus 11 mestres. A tabela a seguir indica os resultados.

Número de filhos	Frequência
0	1
1	3
2	3
3	2
4	1
23	1
Total	11

A média dos dados tabulados vem de:

$$Mt(\text{Filhos}) = \frac{0 \times 1 + 1 \times 3 + 2 \times 3 + 3 \times 2 + 4 \times 1 + 23 \times 1}{11} = 3,8 \cong 4$$

Neste exemplo, temos uma média de aproximadamente 4 filhos, o que parece ser um número razoável. No entanto, a inspeção visual da tabela mostra que, na verdade, das 11 pessoas temos somente duas com 4 filhos ou mais. É evidente a concentração em torno das pessoas com 1, 2 ou 3 filhos. Ou seja, neste caso a média se distancia dos valores típicos de forma ‘artificial’, e 4 filhos não é uma indicação do conjunto.

Verificando os dados novamente vemos que, na verdade, este valor de média foi fortemente influenciado pela pessoa que tem 23 filhos, algo que, na verdade, destoa completamente dos demais. Este valor é atípico, e constitui o que chamamos mais correntemente de *outlier* (palavra usada em inglês sem tradução definitiva para o português, e que vem mantendo seu uso no Brasil).

Os problemas gerados pelos *outliers* são evidentes, principalmente quando estamos tratando de médias amostrais para avaliação das médias populacionais (de fato, um valor extremo em uma amostra pode ser determinante para que a média amostral difira muito da média populacional). Por esta razão, foram desenvolvidos métodos alternativos de extração da média aritmética que minimizam problemas como este indicado acima. As duas principais alter-

nativas são a média aparada (ou truncada) e a média winsorizada, que estão apresentadas no subitem seguinte.

6.2.3. Média aparada

A média aparada é a média aritmética resultante da exclusão de um determinado percentual de valores (as aparas) nas fronteiras inferior e superior do conjunto de dados. Este procedimento retira, portanto, potenciais valores que potencialmente viciam a média dos dados.

No exemplo 6, acima se extraíssemos os dois valores dos extremos, ou seja, o sujeito sem filho e o sujeito com 23 filhos, a nova média, representada por Mtr , seria: $Mtr(\text{Filhos}) = (1x3 + 2x3 + 3x2 + 4x1)/9 \cong 2$, ou seja, teríamos um valor de média que é evidentemente muito mais próximo dos valores típicos do conjunto de dados.

Neste caso excluímos os dois valores, porém poderíamos haver excluídos 4 dos valores, ou até mais que isto. Esta é a decisão central da extração da média aparada. Não temos referência com uma justificativa que não aquela de bom senso. Há autores que sustentam que um valor de 10% acima e abaixo, outros sugerem 5%, 20% e assim sucessivamente. O processo numérico para um conjunto de dados $X = \{x_1, x_2, \dots, x_N\}$ é o seguinte:

1. Seleção do percentual ‘p’ de ‘truncagem’ da variável;
2. Aplicação do número ‘t’ de itens a serem recortados, pela seguinte fórmula:
 $t = \lfloor Np/100 \rfloor$, em que N representa o número de elementos da variável¹⁷.
3. Ordenamento da variável do menor para o maior.
4. Cálculo da média truncada, pela fórmula a seguir:

$$Mtr(X) = \frac{x_{(t+1)} + x_{(t+2)} + \dots + x_{(N-1-t)} + x_{(N-t)}}{N - 2t} = \frac{\sum_{i=t+1}^{N-t} x_i}{N - 2t}$$

Evidentemente, a média aritmética simples é uma média aparada com 0% de apara. Vejamos agora um exemplo de aplicação deste conjunto de procedimentos:

¹⁷ Aqui, o símbolo $\lfloor \cdot \rfloor$ representa o menor inteiro abaixo do valor encontrado. Por este procedimento, é possível que o número de dados recortados não seja equivalente ao percentual de truncagem definido, porém, para grandes quantidades de dados, a diferença tende a ser muito pequena.

Exemplo 7: Uma pesquisa foi realizada em um bairro com a finalidade de verificar a possibilidade de o poder público criar uma biblioteca no lugar. Para tanto, foi levantado o número de livros disponíveis em 25 residências, tendo-se verificado os seguintes resultados.

7	8	14	6	7
0	16	13	19	17
0	13	7	14	250
5	2	9	4	1300
17	2	9	3	19

Este conjunto de dados tem média aritmética igual a 70,4, o que é, claramente, um número que não representa bem o conjunto das 25 observações ou a população de origem dessa amostra. Realizemos então a extração da média, aplicando os passos indicados.

1. Na definição do percentual, tomemos duas alternativas: 10% e 20%.
2. Para os números de dados a serem recortados, teremos:

$$t = \left\lfloor \frac{25 \times 10}{100} \right\rfloor = \lfloor 2,5 \rfloor = 2, \text{ e } t = \left\lfloor \frac{25 \times 20}{100} \right\rfloor = \lfloor 5 \rfloor = 5$$

3. Reordenando os dados, teremos:

$x_{(1)}$	$x_{(2)}$	$x_{(3)}$	$x_{(4)}$	$x_{(5)}$	$x_{(6)}$	$x_{(7)}$	$x_{(8)}$	$x_{(9)}$	$x_{(10)}$	$x_{(11)}$	$x_{(12)}$	$x_{(13)}$
0	0	2	2	3	4	5	6	7	7	7	8	9
$x_{(14)}$	$x_{(15)}$	$x_{(16)}$	$x_{(17)}$	$x_{(18)}$	$x_{(19)}$	$x_{(20)}$	$x_{(21)}$	$x_{(22)}$	$x_{(23)}$	$x_{(24)}$	$x_{(25)}$	
9	13	13	14	14	16	17	17	19	19	250	1300	

4. Para a média, teremos então, nos dois casos, o seguinte:

- Truncarem de 10%:

$$Mtr(\text{Livros}) = \frac{x_{(3)} + x_{(4)} + \dots + x_{(22)} + x_{(23)}}{25 - 4} = \frac{2 + 2 + 3 + \dots + 19 + 19}{21} \cong 10$$

- Truncarem de 20%:

$$Mtr(\text{Livros}) = \frac{x_{(6)} + x_{(7)} + \dots + x_{(19)} + x_{(20)}}{25 - 10} = \frac{4 + 5 + 3 + \dots + 16 + 17}{15} \cong 10$$

No exemplo, os dois resultados foram aproximadamente iguais, que, verificando os dados, parece ser um valor de média que representa bem o conjunto dos dados e parece ser um valor de referência melhor para a população do bairro. A média anterior foi fortemente afetada pelos dois valores extremos de

250 e 1300 livros. Estes dois casos são atípicos, e provavelmente requereriam uma análise separada dos demais.

O exemplo permite verificar, portanto, que o procedimento de aparar a média traz a vantagem de livrá-la do problema de valores extremos. Em outras palavras, dizemos que o procedimento é ‘robusto’ para o cálculo da média, uma vez que o valor calculado não é afetado por valores que podem enviesar o resultado.

O procedimento manual é bastante trabalhoso, como vimos. No entanto, isto não é problema, tendo em vista que os softwares já possuem rotinas específicas para os cálculos. Temos os seguintes procedimentos:

- No Excel, utilizamos o comando ‘=MÉDIA.INTERNA(1; 2)’, em que na parte 1 são indicados os valores, e na parte 2 é indicado o percentual de aparar;
- No R, é possível calcular a média aparada para qualquer percentual. Assim, indicando por ‘x’ o vetor da variável, e y o percentual (dividido por 100), basta escrever na tela: `mean(x, tr=y)` ou `mean(x, y)`.
- No JASP, utilizaremos o R console e o código: `mean(x, trim=y)`.
- Em Python, é possível calcular a média aparada para qualquer percentual. Podemos utilizar a biblioteca `scipy.stats` e a função `trim_mean(x, y)` em que ‘x’ é o vetor da variável e y o percentual (dividido por 100).

Exercitando no R a extração das médias dos dados do exemplo 7 (aritmética simples, aparada com 10% e aparada com 20%), teremos o seguinte:

```
> Livros=c(7, 8, 14, 6, 7, 0, 16, 13, 19, 17, 0, 13, 7, 14, 250, 5,
2, 9, 4, 1300, 17, 2, 9, 3, 19)
> mean(Livros)
[1] 70.44

> mean(Livros, 0.1);
[1] 10.04762

> mean(Livros, 0.2)
[1] 9.933333
```

Fica como atividade exercitar no JASP e em Python. Os exemplos apresentados deixam claro o quanto valores extremos podem afetar a média, mas, no estágio de exposição que estamos, ainda não é possível compreender o problema que dados extremos podem provocar na análise de dados. As consequências são mais

pronunciadas nas técnicas mais sofisticadas (como as multivariadas). Os riscos são tão grandes que, nos últimos anos, os procedimentos robustos vêm, progressivamente, ganhando espaço na análise de dados. A média aparada é, portanto, um conceito introdutório que encontrará aplicações posteriormente, como veremos.

6.2.4. Média winsorizada

A necessidade de construção de alternativas robustas para o cálculo de medidas estatísticas conduziu ao desenvolvimento e à ampla utilização da média aparada. No subitem anterior foi possível verificar o potencial desta técnica para análise de dados, ainda em nível preliminar.

Uma alternativa que vem progredindo em uso nos últimos anos é a média winsorizada, que é semelhante à média aparada, mas tem uma diferença central. Aqui, em lugar de excluir os valores dos extremos, nos percentuais indicados, nosso procedimento consiste em substituir os valores que seriam excluídos pelo último valor vizinho do que seria excluído, acima e abaixo.

No caso do exemplo 6, que tínhamos 11 observações, na média aparada extraída retiramos o menor e o maior valores. Na média winsorizada, o menor valor, no caso 0 (zero), é substituído pelo valor 1 (que passa a ter frequência 4 e não 3), e o maior valor, que é 23, é substituído pelo valor 4 (que passa a ter frequência 2 e não mais 1). Teremos então a nova média, representada por Mw , assim calculada (observe que o número de observações não muda): $Mw(\text{Filhos}) = (1 \times 4 + 2 \times 3 + 3 \times 2 + 4 \times 2) / 11 = 2,2 \cong 2$. Ou seja, um valor muito próximo do valor da média aparada. O procedimento de extração da média winsorizada de um conjunto de dados $X = \{x_1, x_2, \dots, x_N\}$ é o seguinte:

1. Seleção do percentual 'p' de 'winsorização' da variável;
2. Aplicação do número 'w' de itens a serem substituídos, pela seguinte fórmula:
 $w = \lfloor Np/100 \rfloor$, em que N representa o número de elementos da variável;
3. Ordenamento da variável do menor para o maior;
4. Cálculo da média winsorizada, pela fórmula a seguir:

$$\begin{aligned}
 Mw(X) &= \frac{(w+1) \cdot x_{(w+1)} + x_{(w+2)} \dots + x_{(N-1-w)} + (w+1) \cdot x_{(N-w)}}{N} = \\
 &= \frac{(w+1) [x_{(w+1)} + x_{(N-w)}] + \sum_{i=w+2}^{N-w-1} x_i}{N}
 \end{aligned}$$

Temos a seguir um exemplo de aplicação deste conjunto de procedimentos:

Exemplo 8: um canal de televisão precisa saber o tempo médio de uma partida de vôlei para decidir sobre a compra dos direitos de transmissão e o encaixe na grade de programação. Foi feito um levantamento do tempo em minutos de 16 partidas, tendo-se verificado os seguintes resultados:

60	120	97	58	96
109	99	130	116	99
97	105	122	167	107
124	127	94	49	104

O tempo médio desta amostra é de 104 minutos, mas é evidente que temos 3 valores que destoam dos demais, a saber: 49, 60 e 167 minutos. Vejamos então a média winsorizada, considerando os passos indicados.

1. Para este exemplo, tomemos como percentuais de winsorização 5% e 20%;
2. Os números de itens de cada caso são:

$$w = \lfloor 20 \times 5 / 100 \rfloor = \lfloor 1 \rfloor = 1, \text{ e } w = \lfloor 20 \times 20 / 100 \rfloor = \lfloor 4 \rfloor = 4$$

3. Ordenamento da variável do menor para o maior.
4. Reordenando os dados, teremos:

$x_{(1)}$	$x_{(2)}$	$x_{(3)}$	$x_{(4)}$	$x_{(5)}$	$x_{(6)}$	$x_{(7)}$	$x_{(8)}$	$x_{(9)}$	$x_{(10)}$
49	58	60	94	96	97	97	99	99	104
$x_{(11)}$	$x_{(12)}$	$x_{(13)}$	$x_{(14)}$	$x_{(15)}$	$x_{(16)}$	$x_{(17)}$	$x_{(18)}$	$x_{(19)}$	$x_{(20)}$
105	107	109	116	120	122	124	127	130	167

5. As médias virão de:

- Winsorização de 5%:

$$\begin{aligned} Mw(\text{Tempo}) &= \frac{(1+1) \cdot x_{(2)} + x_{(3)} + \dots + x_{(18)} + (1+1) \cdot x_{(19)}}{20} = \\ &= \frac{2.58 + 60 + \dots + 127 + 2.130}{20} \cong 102,6 \end{aligned}$$

- Winsorização de 20%:

$$\begin{aligned} Mw(\text{Tempo}) &= \frac{(4+1) \cdot x_{(5)} + x_{(3)} + \dots + x_{(15)} + (4+1) \cdot x_{(16)}}{20} = \\ &= \frac{5.96 + 97 + \dots + 120 + 5.122}{20} \cong 107,2 \end{aligned}$$

No primeiro caso, a média aparada foi de 103,5, e no segundo, a média aparada foi de 105,9 minutos. Os resultados do exemplo mostram que nem sempre há alterações substanciais nas médias, mesmo com a existência de valores discrepantes (veja que as médias são todas próximas, mesmo das médias aparadas). No entanto, o efeito dos valores extremos não vem somente sobre as médias, mas podem vir na aplicação de outras técnicas.

Com relação aos softwares, a situação da média winsorizada é diferente. De fato, o Excel dispõe de uma rotina pré-estabelecida para seu cálculo. Já no R temos duas opções, uma mais breve, baseada no download de uma biblioteca já estabelecida, e uma mais trabalhosa, mas sempre disponível quando não houver forma de baixar a biblioteca. Abaixo apresentamos as duas opções.

Quadro 6.1 – Programa do R para média winsorizada

```
#Opção 1
x=c(...)           #Vetor de dados
install.packages("psych") #Instala o pacote de base
library(psych)      #Habilita a biblioteca
w=...               #Define o valor de winsorização
winsor.mean(x, w/100) #Calcula a média winsorizada

#Opção 2
x=c(...)           #Vetor de dados
w=...               #Define o valor de winsorização
y<-sort(x)          #Ordenando x
n<-length(x)        #Fixando o tamanho de x
ibot<-floor(w*n)+1   #Quantidade inferior
itop<-n-ibot+1       #Quantidade superior
xbot<-y[ibot]         #Valor de substituição inferior
xtop<-y[itop]         #Valor de substituição superior
y<-ifelse(y<=xbot,xbot,y) #Substituição inferior
y<-ifelse(y>=xtop,xtop,y) #Substituição superior
wm=mean(y); wm       #Cálculo da média
```

Exercitando no R (ou no JASP, visto que os códigos são os mesmos ao utilizar o R console), tomando por base o exemplo dos tempos das partidas de vôlei, calculemos, ordenadamente, a média aritmética, e as médias aparadas a 5% e a 20%, e em seguida, calculemos a médias winsorizadas nestes mesmos percentuais. Teremos:

```
>x=c(49, 58, 60, 94, 96, 97, 97, 99, 99, 104, 105, 107, 109, 116,
120, 122, 124, 127, 130, 157)
# Média aritmética simples
> mean(x)
[1] 103.5
# Média aparada a 5%
> mean(x,0.05)
[1] 103.5556
# Média aparada a 20%
> mean(x,0.2)
[1] 105.9167

#Programa para o cálculo da média winsorizada a 5%
> x=c(49, 58, 60, 94, 96, 97, 97, 99, 99, 104, 105, 107, 109, 116,
120, 122, 124, 127, 130, 157)
> tr=0.05
> y<-sort(x)                                #Ordenando x
> n<-length(x)                              #Fixando o tamanho de x
> ibot<-floor(tr*n)+1                       #Quantidade inferior
> itop<-n-ibot+1                             #Quantidade superior
> xbot<-y[ibot]                             #Valor de substituição inferior
> xtop<-y[itop]                             #Valor de substituição superior
> y<-ifelse(y<=xbot,xbot,y) #Substituição inferior
> y<-ifelse(y>=xtop,xtop,y) #Substituição inferior
> wm=mean(y); wm                            #Cálculo da média
[1] 102.6

#Programa para o cálculo da média winsorizada a 20%
> x=c(49, 58, 60, 94, 96, 97, 97, 99, 99, 104, 105, 107, 109, 116,
120, 122, 124, 127, 130, 157)
> tr=0.20
```

```

> y<-sort(x)                                #Ordenando x
> n<-length(x)                             #Fixando o tamanho de x
> ibot<-floor(tr*n)+1                       #Quantidade inferior
> itop<-n-ibot+1                           #Quantidade superior
> xbot<-y[ibot]                             #Valor de substituição inferior
> xtop<-y[itop]                             #Valor de substituição superior
> y<-ifelse(y<=xbot,xbot,y) #Substituição inferior
> y<-ifelse(y>=xtop,xtop,y) #Substituição inferior
> wm=mean(y); wm                           #Cálculo da média
[1] 107.15

```

Para exercitar em Python com o mesmo vetor de dados do exemplo acima, utilizaremos a biblioteca `scipy.stats.mstats` e a função `winsorize`, como segue abaixo.

```

import numpy as np
from scipy.stats.mstats import winsorize

dados = np.array([49,58,60,94,96,97,99,99,104,105,107,109,116,
120,122,124,127,130,157])

#Para 5% de winsorização
winsor_mean=np.mean(winsorize(dados,limits=(0.05,0.05)))
print(winsor_mean)

#Para 20% de winsorização
winsor_mean=np.mean(winsorize(dados,limits=(0.2,0.2)))
print(winsor_mean)

```

O exemplo acima ilustra uma situação em que, provavelmente, a média não representa uma boa medida para tomada de decisão. Na verdade, as programações de televisão possuem mais ou menor rigidez, a depender da natureza do canal. Nos canais abertos, por exemplo, a preferência é evidente por programas atividades com tempo bem delimitado, como futebol, fórmula 1, lutas, dentre outros. No caso de esportes como o vôlei, o basquete ou o tênis, em que não é fácil prever quanto tempo a partida durará, parece mais interessante saber o tempo mínimo que a partida ocupará na grade, ou então o tempo máximo.

Diversos problemas práticos envolvem outras medidas que não a média, e a seguir apresentamos um conjunto destas medidas, a começar pela mediana.

6.3. Mediana

Do ponto de vista da estatística descritiva para variáveis finitas ou para amostras, a mediana é o valor que divide estes dados em duas partes iguais em número de elementos. Em outras palavras, a mediana é o valor da variável que divide o número total de elementos (afora o próprio dado da mediana), de tal modo que metade tem valor menor ou igual e a outra metade tem valor maior ou igual ao valor da mediana. Por exemplo, nos valores 1, 10, e 56, o valor 10 é a mediana, pois do que sobra (dois dados), um é menor (1) e o outro é maior (56) que 10.

O procedimento de determinação da mediana para dados não tabulados é bastante simples, bastando ordenar o conjunto de valores e em seguida identificar o valor central. Sendo n o número de elementos da variável $X = \{x_{(1)}, x_{(2)}, \dots, x_{(N)}\}$, e Md a mediana, temos duas alternativas de verificação:

- Se o número de elementos for ímpar, basta tomar somar 1 ao número de elementos e dividir por 2. O número observado será o índice da mediana. Formalmente, teremos:

$$Md(X) = x_{(m)}, \text{ tal que } m = \frac{n+1}{2}$$

- Se o número de elementos for par, então teoricamente, a mediana pode ser representada por qualquer número entre os dois valores intermediários. Para efeito de cálculo, é usual tomar a média aritmética entre estes dois valores. Formalmente, teremos:

$$Md(X) = \frac{x_{(m)} + x_{(m+1)}}{2}, \text{ tal que } m = \frac{n}{2}$$

Vejamos então um exemplo:

Exemplo 9: foram identificadas notas (na escala de 0 a 100) dos alunos da disciplina de Relações de trabalho, representadas no seguinte conjunto $\text{Notas}=\{83, 67, 98, 75, 61, 49, 83, 70, 87, 85, 60\}$.

Para o cálculo da mediana, primeiramente precisamos ordenar os elementos, como feito a seguir.

$x_{(1)}$	$x_{(2)}$	$x_{(3)}$	$x_{(4)}$	$x_{(5)}$	$x_{(6)}$	$x_{(7)}$	$x_{(8)}$	$x_{(9)}$	$x_{(10)}$	$x_{(11)}$
60	61	67	69	70	75	83	83	85	87	98

Como temos um número ímpar de valores (11), então a mediana será o valor indexado em 6 (ou seja, $(11+1)/2$). Portanto, a mediana ($Md(Notas)$) é 75.

Exemplo 10: foi levantado o número de horas diante da internet de 8 jovens, tendo-se observado o seguinte conjunto $Horas=\{1; 0; 2,5; 0; 0,5; 2; 9; 3,5, \}$.

Reordenando os dados para a extração da mediana, temos a tabela seguinte:

$x_{(1)}$	$x_{(2)}$	$x_{(3)}$	$x_{(4)}$	$x_{(5)}$	$x_{(6)}$	$x_{(7)}$	$x_{(8)}$
0	0	0,5	1	2	2,5	9	10

Como o número de elementos é par (8 elementos), então a mediana seria a média aritmética dos dois valores do meio, ou seja $x_{(4)}$ e $x_{(5)}$, ou seja, $Md(horas) = 1,5$.

O cálculo da mediana nos softwares é bastante simples, como descrito a seguir:

- No Excel, o procedimento consiste em escrever o comando ‘=med()’, e indicar entre os parênteses a série de dados de interesse;
- No R, primeiramente é preciso escrever o vetor contendo os dados observados. Sendo ‘x’ o vetor, basta escrever median(x), e imediatamente abaixo surgirá o valor da média.
- No JASP, o caminho é descriptives → descriptive statistics → arrasta a variável de interesse para o campo variables → no campo statistics, clica em median.
- Em Python, primeiramente é preciso importar a biblioteca statistics e definir a lista contendo os dados observados. Na sequência, pode-se utilizar a função statistics.median(x) em que x é a lista de dados.

Além destes procedimentos de cálculo da mediana, temos ainda a possibilidade de sua averiguação em dados tabulados. Neste caso, temos um conjunto de passos que dependem da forma de tabulação. Assim, para o caso de tabelas de números discretos e sem intervalos, a mediana pode ser calculada de maneira direta a partir da frequência acumulada. Vejamos um exemplo ilustrativo:

Exemplo 11: na planilha do Apêndice D, tomemos a tabela de frequências da variável ‘Identificação com a profissão’, que está apresentada abaixo.

Valores	Frequência	Percentuais	Acumulada (%)
1	1	0,9	0,9
2	2	1,8	2,7
3	4	3,6	6,3
4	18	16,1	22,3
5	53	47,3	69,6
6	30	26,8	96,4
7	4	3,6	100,0
Total	112	100,0	–

É evidente que, por esta tabela, o valor mediano será 5, pois até 5 temos 22,3% do total de observações, e logo em seguida temos o valor 5, que se repete até completar 69,6% do total; portanto, a metade dos valores estará ao longo do valor 5.

Um problema maior emerge quando temos valores tabulados na forma de intervalos. Neste, com a coluna de valores acumulados é possível indicar diretamente a classe que contém a mediana, mas não um valor mediano específico. Neste caso, é necessário estimar em seguida qual valor representará a mediana dentro da classe definida. Precisaremos, portanto, da frequência acumulada, e não apenas da percentagem acumulada. Temos uma formulação específica que ilustrarei por meio de um exemplo.

Exemplo 12: Tomemos o conjunto de dados do item 4.5 do capítulo 4. Os valores detalhados possuem mediana igual a 678,4 (e média 651,1). Na primeira tabela gerada, com 8 intervalos (reproduzida abaixo), é fácil ver que o intervalo que contém a mediana é [652,7; 787,3), que acumula entre 47,1% e 64,3% do total de observações.

Intervalos	Frequência	Freq. Acumu	Percentuais	Perc. Acumu.
[114,3; 248,9]	5	5	7,1	7,1
[248,9; 383,5]	13	18	18,6	25,7
[383,5; 518,1]	10	28	14,3	40,0
[518,1; 652,7]	5	33	7,1	47,1
[652,7; 787,3]	12	45	17,1	64,3
[787,3; 921,9]	9	54	12,9	77,1
[921,9; 1056,5]	9	63	12,9	90,0
[1056,5; 1191,1]	7	70	10,0	100,0
Total	70	-	100,0	-

Nosso problema agora é definir um valor adequado dentro deste intervalo para representar a mediana. Uma forma de bom senso seria encontrar a faixa proporcional do limite inferior até o ponto mediano. No entanto, não temos um ponto mediano bem definido neste caso, pois temos um número par de elementos (70). Seguimos então os procedimentos abaixo:

- Primeiro, fixamos o valor $n/2$ como a posição do elemento mediano, independente de termos um número par ou ímpar de elementos. Assim, entenderemos o elemento mediano ($E_{(md)}$) como o elemento 35 ($70/2$);
- Definiremos o valor $S_{(md)}$ que somaremos ao valor mínimo $l_{(i)}$ da classe mediana (em nosso caso, $l_{(i)} = 652,7$), ou seja, $Md = l_{(i)} + S_{(md)}$. Precisamos calcular $S_{(md)}$;
- Para tanto, usamos a seguinte regra:
 - » Tomamos a diferença da frequência acumulada até a classe mediana ($F_{(md)}$) pela frequência acumulada até a classe anterior ($F_{(a)}$) e definimos a frequência da classe mediana dada por $f_{(md)} = F_{(md)} - F_{(a)}$. Em nosso caso, $f_{(md)} = 12$. Ou seja, do limite inferior até o limite superior da classe mediana temos 12 valores;
 - » Tomamos em seguida a diferença entre o do elemento mediano ($E_{(me)}$) e a frequência acumulada até a classe anterior ($F_{(a)}$), e a representaremos por $f_{(E_{md})} = E_{(md)} - F_{(a)}$. Em nosso caso, $f_{(E_{md})} = 35 - 33 = 2$;
 - » O raciocínio e o seguinte: na o valor da amplitude total do intervalo ($c = 134,6$) está distribuído em 12 partes (que supomos iguais), de 33 até 45.

Com isto, podemos identificar o valor que chega de 33 até 35, considerando a proporção dentro do intervalo;

- » Simbolicamente, esta afirmação é equivalente a seguinte: $(f_{(md)} = F_{(md)} - F_{(a)})$ está para c assim como $(f_{(E_{md})} = E_{(md)} - F_{(a)})$ está para $S_{(md)}$. No nosso exemplo, 12 está para 134,6 assim como 2 está para $S_{(md)}$;

- » Pela regra de três:

$$\frac{F_{(md)} - F_{(a)}}{c} = \frac{E_{(md)} - F_{(a)}}{S_{(md)}} \Rightarrow S_{(md)} = c \left(\frac{E_{(md)} - F_{(a)}}{F_{(md)} - F_{(a)}} \right),$$

$$\text{ou ainda, } S_{(md)} = c \left(\frac{E_{(md)} - F_{(a)}}{f_{(md)}} \right).$$

Portanto, no exemplo, $S_{(md)} = 134,6 \cdot \left(\frac{2}{12} \right) = 22,43$.

- Portanto, usamos a seguinte regra: $Md = l_{(i)} + c \left(\frac{E_{(md)} - F_{(a)}}{f_{(md)}} \right)$, ou seja, $Md = 652,7 + 22,4 \Rightarrow Md = 675,1$.

É fácil ver que o valor mediano calculado nos dados tabulados é bem próximo do valor mediano identificado nos dados detalhados (que foi de 678,4). Podemos fixar a fórmula gerada no exemplo, como indicado a seguir:

$$Md = l_{(i)} + c \left(\frac{E_{(md)} - F_{(a)}}{f_{(md)}} \right)$$

- onde Md é a mediana, $l_{(i)}$ é o limite inferior da classe mediana, c é sua amplitude, $E_{(md)}$ é o elemento mediano, $F_{(a)}$ é a frequência acumulada até a classe anterior à classe mediana e $f_{(md)}$ é a frequência da classe mediana.

Esta fórmula facilita sobremaneira o cálculo da mediana em situações como a indicada, e o resultado demonstra que não há uma grande perda em relação ao valor calculado nos dados detalhados. Aplicando a fórmula à segunda tabela de intervalos do item 4.5, que ficou definida com 4 intervalos, o valor calculado será 678,3, um valor bastante próximo dos dois valores anteriormente indicados (valor real e valor na tabela de 8 intervalos).

O leitor deve ter observado nos exemplos que a média se aproxima do valor mediano. De fato, no conjunto de dados do exemplo 9, a média aritmética foi de 76,2, valor bem próximo da mediana, que foi de 75. Em dados ‘bem comportados’, ou seja, sem valores extremos, as duas medidas são convergentes.

De fato, do exemplo 8 vimos que a média foi de 108, e a mediana será 104,5 e no exemplo 12 a mediana foi de 675,1 (na primeira extração) e a média 656,9.

No entanto, para dados com valores extremos, média e mediana costumam ser bem distintas. De fato, no exemplo 10 tivemos uma mediana de 1,5, mas a média é de 3,1, bastante acima da mediana. Isto ocorre porque os dois maiores valores daquela série são claramente discrepantes em relação aos demais. No exemplo 7, em que também verificamos valores extremos, a mediana foi de 9, e a média de 70,4.

Cabe observar por outro lado, que, no exemplo 7, a média aparada foi de 10, já bem mais próxima do valor da mediana. Em verdade, isto ocorre porque a mediana, assim como a média aparada e a média winsorizada, é uma medida de tendência central robusta a valores extremos, que não se afeta com valores extremos¹⁸.

O problema em potencial da mediana é justamente o fato de ser um caso extremo de média aparada, que desconsidera muitos valores. De fato, de uma certa quantidade de dados, apenas um ou dois são utilizados na computação da medida, e simplesmente se desconsideram todos os demais. Assim, mesmo sendo uma medida com a vantagem de ser robusta, a mediana é mais frágil em termos de uso diante das outras medidas igualmente robustas anteriormente apresentadas (médias aparada e winsorizada). A situação de uso indicará ao analista a melhor alternativa de medida descritiva, mas é sempre recomendado que se analisem todos estes valores.

É fácil ver que, assim como a média, a mediana é uma medida de posição, tendo em vista que, necessariamente, fica entres os valores máximo e mínimo do conjunto de dados, e que a multiplicação ou a divisão dos dados por uma constante implica em igual operação sobre a medida. Como informado, assim como a média, a mediana é uma medida de tendência central. Uma terceira medida é a moda, sobre a qual tratarei rapidamente no subitem seguinte.

18 Os leitores podem observar que a mediana é justamente um caso extremo de média aparada (com aparadas de 50%).

6.4. Moda

Em estatística de variáveis finitas ou de amostras, a moda (Mo) é definida como o valor mais frequente de um conjunto de dados, ou seja, a moda é o valor que aparece mais vezes, comparativamente aos demais (os leitores podem perceber a associação conceitual da ideia corrente de moda, que se refere ao que mais se usa em um determinado momento; no entanto, não há razões para crer que um conceito [a moda estatística] tem associação mais intensa com o outro [algo em uso corrente]).

Quando temos somente um valor mais frequente, este é chamado de único valor modal de uma variável unimodal. Se tivermos dois valores igualmente mais frequentes, teremos dois valores modais de uma variável dita bimodal. Quando temos mais de dois valores, dizemos que nossa variável é multimodal. Se não tivermos qualquer valor com frequência maior que os demais, dizemos que a amostra é amodal.

O procedimento para determinação da moda de valores detalhados é bastante simples, bastando ordenar os dados e verificar de forma direta o(s) valor(es) mais frequente(s). Retomando os exemplos 9 e 1 anteriormente fica fácil perceber o procedimento. De fato, no exemplo 9 vemos que o valor 83 é a moda, pois se repete duas vezes; já no exemplo 1, temos uma amostra amodal, pois todos os valores aparecem apenas uma vez.

Nos softwares de referência, temos os seguintes caminhos:

- No Excel, o procedimento consiste em escrever o comando `=modo()`, e indicar entre os parênteses a série de dados de interesse;
- Diferente da média e da mediana, a moda não possui uma rotina implantada no R, de forma que é necessário desenvolver um pequeno algoritmo. O primeiro procedimento, depois de gerado o vetor 'x' com os valores da variável, é dispor os dados em uma tabela, escrevendo `table(x)` (a rigor, este comando não é necessário, mas sua utilização permite verificar o resultado em uma tabela). Será então gerada uma tabela em que estão indicadas as frequências de cada valor. Em seguida, é solicitada a indicação do(s) valor(es) mais frequentes, escrevendo o seguinte comando: `moda=subset(table(x),table(x)==max(table(x))); moda`.

- No JASP, o caminho é *descriptives* → *descriptive statistics* → arrasta a variável de interesse para o campo *variables* → no campo *statistics*, clica em *mode*.
- Em Python, o caminho é mais simples do que no R porque há funções implementadas para calcular a moda. Um primeiro caminho é utilizar a biblioteca *statistics*. A função, nesse caso, é *statistics.mode(x)* em que *x* é a lista de dados; porém, esse procedimento pode reportar erros se os dados forem bimodais. Um segundo caminho é utilizar a biblioteca *statistics.multimode(x)*. Em ambos os casos, primeiramente, é preciso importar as bibliotecas e definir a lista contendo os dados observados.

Vejamos um exercício. Retomando a variável *tempo*, a qual utilizamos na exemplificação da média winsorizada. Teremos, no R, os detalhes a seguir, que mostram, ao final, que a variável é bimodal, com os valores modais 97 e 99, cada um aparecendo duas vezes:

```
> Tempo=c(49, 58, 60, 94, 96, 97, 97, 99, 99, 104, 105, 107, 109,
116, 120, 122, 124, 127, 130, 157)
# Fórmula da moda
> moda=subset(table(Tempo),table(Tempo)==max(table(Tempo))); moda
Tempo
97 99
2 2
```

Em Python, teremos:

```
import statistics
dados = [49, 49, 58, 60, 94, 96, 97, 97, 99, 99, 104, 105, 107,
109, 116, 120, 122, 124, 127, 130, 157]
statistics.multimode(dados)
```

Temos fórmulas especiais para dados tabulados, que não creio valer a pena detalhar aqui. Na verdade, a moda, embora seja uma medida relevante em alguns tratamentos estatísticos, é em geral uma medida de uso menor em Ciências sociais e comportamentais, comparativamente às demais. De fato, é fácil ver que a moda não é exatamente uma medida de tendência central, salvo em alguns tipos de variáveis. Podemos até mesmo nem ter uma moda, e se a temos, é possível que esteja em qualquer posição no elenco de dados.

Deixo ao leitor como exercício a verificação das alternativas de cálculo da moda, e sua aplicação, quando cabido, aos exemplos aqui apresentados. Tendo em vista nosso interesse em providenciar um referencial para análise exploratória de dados, creio ser pertinente agora averiguar os quantis, como faço no item seguinte.

6.5. Os quantis

Os quantis são medidas estatísticas de posição que, em boa medida, generalizam a mediana. Vimos que a mediana divide o restante dos dados ao meio, ou seja, a mediana é maior ou igual que 50% dos demais dados, e menor ou igual que 50% do restante dos dados. O quantil de ordem p de uma variável X , representado genericamente por $Q_{(p)}(X)$ com $0 < p < 1$, é definido como o valor de um conjunto de dados que é maior ou igual que 100p% dos demais elementos, e menor ou igual que 100(1-p)% dos demais elementos. Evidentemente, a mediana é justamente o quantil de ordem 50, ou seja $Md(X) = Q_{(0,50)}(X)$.

A formulação desta definição permite irmos além da mediana e definirmos tantos quantis quantos forem de nosso interesse, entre 0 e 1. A literatura corrente costuma enfatizar 3 categorias de quantis, que são os percentis, os decis, e os quartis. Estas medidas são utilizadas em várias técnicas estatísticas, e por isto as apresentaremos aqui.

– Percentis

Os percentis são os valores que dividem o conjunto de dados em 100 partes iguais, de forma ordenada crescente. Assim, o primeiro percentil ($Q_{(0,01)}$) é o valor que é maior ou igual que 1% dos demais valores da amostra, e menor ou igual que 99% dos demais valores; o segundo percentil ($Q_{(0,02)}$) é maior ou igual que 2% do restante dos valores, e menor ou igual que 98% dos demais valores, e assim sucessivamente (a mediana é igual ao 50º percentil, ou seja, $Q_{(0,50)} = Md$).

A determinação dos percentis pode ser deduzida de forma semelhante à dedução que foi feita para a mediana. Aqui faremos somente com base nos valores detalhados, e deixo como exercício a indicação da forma como os percentis podem ser deduzidos em dados tabulados.

No caso da mediana, vimos que sua determinação depende da quantidade (par ou ímpar) de elementos da amostra. Sigamos um raciocínio semelhante para o caso dos percentis. Como vimos, para o número ímpar de elementos, no caso da mediana o número de elementos adicionado de 1 era dividido por 2; portanto, no caso dos percentis é esperado que a divisão seja por 100, de modo que o primeiro percentil será então o elemento que tiver índice igual ao resultado da referida divisão. Por exemplo, se temos 299 elementos ($n=299$), então fazemos $(299+1)/100=3$, e o primeiro percentil será o terceiro elemento.

Assim, dos 298 elementos restantes, 2 (0,68%, aproximadamente 1%) serão menores ou iguais ao primeiro percentil, e 296 (99,32%, aproximadamente 99%) serão maiores ou iguais ao valor observado¹⁹. Para o segundo percentil, o procedimento consiste então em multiplicar a divisão anterior por 2, ou seja, fazemos $2 \times (299+1)/100=6$, de modo que o elemento com índice 6 será o segundo percentil.

Seguindo este procedimento, para calcular o percentil $Q_{(k)}$, fazemos então $i = k(n + 1)$, e o valor observado corresponderá ao índice do valor procurado. No entanto, nem sempre esta divisão gera um valor inteiro. Por exemplo, se temos 125 elementos, para a identificação do índice do 36º percentil ($p_{(0,36)}$), aplicando a fórmula teremos $i = 0,36 \times (125 + 1) = 45,36$, ou seja, um valor que não é índice de qualquer dos elementos (pois os índices são sempre valores inteiros).

Neste caso, uma alternativa lógica seria arredondar o índice para 45 ou 46, e daí teríamos que o percentil procurado seria o elemento com índice 45 ou 46 (ou seja, $Q_{(0,36)} = x_{(45)}$ ou $Q_{(0,36)} = x_{(46)}$). Alternativamente, seria possível interpolar o valor decimal em relação aos valores que correspondem aos índices inteiros imediatamente inferior e superior (como fizemos no cálculo da mediana de dados tabulados). Como o valor decimal está mais próximo de 45 do que de 46, uma ponderação possível seria multiplicar o decimal pela observação da posição acima, e multiplicar a diferença para 1 (ou seja, $1-0,26=0,64$) pelo valor inferior. De outra forma, no conjunto de 125 elementos, o 36º percentil virá da multiplicação de 0,36 pelo 46º elemento somada à multiplicação de 0,64 pelo 45º elemento. Simbolicamente, temos: se $n = 125$, então $Q_{(0,36)} = 0,64x_{(45)} + 0,36x_{(46)}$.

19 Os leitores podem observar de forma evidente que o procedimento parece apresentar alguma dificuldade. Na verdade, para o cálculo de quantis distintos da mediana é bastante provável que isto ocorra, de modo que alguns arredondamentos precisarão ser tolerados.

Em uma generalização, podemos enunciar a seguinte regra para identificação de percentis:

- Primeiramente, identificamos o percentil de interesse $Q_{(k)}$, fazendo $i = k(n + 1)$;
- Se i é inteiro, então o percentil $Q_{(k)}$ é dado por $Q_{(k)} = x_{(i)}$;
- Se i não é inteiro, seja $\lfloor i \rfloor$ a parte inteira de i e $(i - \lfloor i \rfloor)$ a parte decimal; então, o percentil será dado por:

$$Q_{(k)} = (1 - (i - \lfloor i \rfloor))x_{(\lfloor i \rfloor)} + (i - \lfloor i \rfloor)x_{(\lfloor i \rfloor + 1)}.$$

O leitor deve observar que a fórmula final acima generaliza o cálculo, pois se não houver parte decimal, ou seja, se i for inteiro, então $i = \lfloor i \rfloor$, de modo que $i - \lfloor i \rfloor = 0$. Aplicando este resultado na fórmula, teremos: $Q_{(k)} = (1 - 0)x_{(\lfloor i \rfloor)} + 0x_{(\lfloor i \rfloor + 1)} \Rightarrow Q_{(k)} = x_{(\lfloor i \rfloor)} = x_{(i)}$. Esta fórmula, que foi deduzida para um número ímpar de elementos, também pode ser válida para um número par de elementos, ou seja, esta passa a ser a fórmula geral para o cálculo de percentis.

Quando aplicamos a fórmula para o cálculo da mediana é fácil compreender porque seu funcionamento pode ser generalizado. De fato, seja n o número de elementos de uma amostra qualquer. A mediana é o percentil que ocupa a 50% posição, ou seja, $Md = Q_{(0,50)}$. Portanto, $i = 50(n + 1)/100 = (n + 1)/2$. Teremos:

- Se n é ímpar (por exemplo, 299), então esta razão gera exatamente o índice elemento mediano da amostra, dado por $Q_{(0,50)} = x_{(i)}$ (que é o elemento 150 de 299 elementos);
- Se n é par (por exemplo, 300), então o número gerado terá uma parte inteira de $n/2$ e uma parte decimal de 0,5 (de 300, teríamos 150,5); então, neste caso, o elemento mediano será $Q_{(50)} = 0,5x_{(n/2)} + 0,5x_{((n/2)+1)}$, ou, de outra forma, $Q_{(0,50)} = (x_{(n/2)} + x_{((n/2)+1)})/2$.

Nos dois casos, o resultado gerado é exatamente o mesmo que indicamos na definição do procedimento operacional de cálculo da mediana, no item 6.3 acima. A compreensão do conceito e dos procedimentos de cálculo de percentis praticamente dispensa maiores detalhamentos dos decis e dos quartis, que nas mais são que medidas quantis particulares, como observaremos a seguir.

- Decis

Os decis são os valores que dividem os conjuntos de dados em 10 partes iguais em número de elementos (na amostra ordenada). Portanto, o 1º decil ($Q_{(0,1)}$) é o valor que é maior ou igual que 10%, e menor ou igual que 90% do restante dos elementos; o 2º decil ($Q_{(0,2)}$) é maior ou igual que 20%, e menor ou igual que 80% do restante dos elementos; e assim sucessivamente, até o 9º decil ($Q_{(0,9)}$), que é maior ou igual que 90%, e menor ou igual 10% dos demais elementos.

É imediato concluir que o 1º decil é igual ao 10º percentil $Q_{(0,1)} = Q_{(0,10)}$, que o 2º decil é igual ao 20º percentil $Q_{(0,2)} = Q_{(0,20)}$, e assim por diante. Deste modo, para a extração dos decis basta usar o mesmo procedimento de extração utilizado no percentil correspondente.

- Quartis

Os quartis são os valores que dividem os dados da amostra em 4 partes, de forma ordenada. Temos 3 quartis: o primeiro ($Q_{(0,25)}$) dos quais é maior ou igual que 25%, e menor ou igual que 75% dos demais valores; o segundo quartil ($Q_{(0,50)}$) se confunde com a mediana; e o terceiro quartil ($Q_{(0,75)}$) é maior ou igual que 75%, e menor ou igual que 75% dos valores restantes.

Por esta relação, o cálculo dos quartis pode ser facilmente executado a partir do cálculo dos percentis correspondentes, seguindo o método que apresentei anteriormente. Alternativamente, é possível determinar os quartis a partir de cálculos consecutivos de medianas. De fato, desde que encontrada a mediana de um conjunto de dados, o primeiro quartil será então a mediana dos valores que estão abaixo da mediana do conjunto completo, e o terceiro quartil será a mediana dos valores que estão acima da mediana.

Vejamos agora um exemplo no qual aplicaremos os conceitos acima indicados (recomendo ao leitor dar atenção ao exemplo, pois este será objeto de análise nos exercícios).

Exemplo 13: a gerência de um frigorífico analisa a possibilidade de mudar de segmento, considerando a opção de vender somente para clientes institucionais e reduzir a venda para clientes residenciais. Para verificar a lucratividade dos clientes residenciais, o gerente levantou uma amostra de 112 tickets de um total de cerca de 650 registros de uma semana. Os dados estão indicados

abaixo. Analisemos então todo o conjunto de medidas que foram estudadas até aqui: média, mediana, moda e quantis.

32	77	27	62	24	77	39	96	36	82	34	79	28	41
21	65	35	84	28	89	33	112	37	72	20	57	23	114
34	58	26	72	31	46	31	124	26	50	28	76	67	30
22	77	25	65	26	63	25	99	26	113	40	25	73	112
20	83	40	80	28	55	33	112	25	113	70	21	62	23
38	48	40	56	21	59	24	93	29	99	102	35	63	15
21	70	27	82	21	82	29	123	35	100	36	125	20	118
40	75	40	80	25	47	30	113	22	95	36	90	24	90

a) A média do conjunto de dados foi de 55,33, o que se obtém pela soma dos valores e sua divisão por 112. A média aparada com 10% é de 52,16, e com 20% é 49,82. Já a média winsorizada com 10% é 54,19, e com 20% é 51,66. A proximidade dos valores de média sugere que potenciais dados discrepantes não estão afetando a média.

b) Para a mediana, é necessário ordenar os elementos, com os resultados indicados na tabela abaixo.

$x_{(1)}$	$x_{(2)}$	$x_{(3)}$	$x_{(4)}$	$x_{(5)}$	$x_{(6)}$	$x_{(7)}$	$x_{(8)}$	$x_{(9)}$	$x_{(10)}$	$x_{(11)}$	$x_{(12)}$	$x_{(13)}$	$x_{(14)}$
15	20	20	20	21	21	21	21	21	22	22	23	23	24
$x_{(15)}$	$x_{(16)}$	$x_{(17)}$	$x_{(18)}$	$x_{(19)}$	$x_{(20)}$	$x_{(21)}$	$x_{(22)}$	$x_{(23)}$	$x_{(24)}$	$x_{(25)}$	$x_{(26)}$	$x_{(27)}$	$x_{(28)}$
24	24	25	25	25	25	25	26	26	26	26	27	27	28
$x_{(29)}$	$x_{(30)}$	$x_{(31)}$	$x_{(32)}$	$x_{(33)}$	$x_{(34)}$	$x_{(35)}$	$x_{(36)}$	$x_{(37)}$	$x_{(38)}$	$x_{(39)}$	$x_{(40)}$	$x_{(41)}$	$x_{(42)}$
28	28	28	29	29	30	30	31	31	32	33	33	34	34
$x_{(43)}$	$x_{(44)}$	$x_{(45)}$	$x_{(46)}$	$x_{(47)}$	$x_{(48)}$	$x_{(49)}$	$x_{(50)}$	$x_{(51)}$	$x_{(52)}$	$x_{(53)}$	$x_{(54)}$	$x_{(55)}$	$x_{(56)}$
35	35	35	36	36	36	37	38	39	40	40	40	40	40
$x_{(57)}$	$x_{(58)}$	$x_{(59)}$	$x_{(60)}$	$x_{(61)}$	$x_{(62)}$	$x_{(63)}$	$x_{(64)}$	$x_{(65)}$	$x_{(66)}$	$x_{(67)}$	$x_{(68)}$	$x_{(69)}$	$x_{(70)}$
41	46	47	48	50	55	56	57	58	59	62	62	63	63
$x_{(71)}$	$x_{(72)}$	$x_{(73)}$	$x_{(74)}$	$x_{(75)}$	$x_{(76)}$	$x_{(77)}$	$x_{(78)}$	$x_{(79)}$	$x_{(80)}$	$x_{(81)}$	$x_{(82)}$	$x_{(83)}$	$x_{(84)}$
65	65	67	70	70	72	72	73	75	76	77	77	77	79
$x_{(85)}$	$x_{(86)}$	$x_{(87)}$	$x_{(88)}$	$x_{(89)}$	$x_{(90)}$	$x_{(91)}$	$x_{(92)}$	$x_{(93)}$	$x_{(94)}$	$x_{(95)}$	$x_{(96)}$	$x_{(97)}$	$x_{(98)}$
80	80	82	82	82	83	84	89	90	90	93	95	96	99
$x_{(99)}$	$x_{(100)}$	$x_{(101)}$	$x_{(102)}$	$x_{(103)}$	$x_{(104)}$	$x_{(105)}$	$x_{(106)}$	$x_{(107)}$	$x_{(108)}$	$x_{(109)}$	$x_{(110)}$	$x_{(111)}$	$x_{(112)}$
99	100	102	112	112	112	113	113	113	114	118	123	124	125

Como temos 112 elementos, ou seja, um número par de elementos, então a mediana será a metade da soma do elemento $x_{(56)}$ com o elemento $x_{(57)}$. Na tabela, é fácil que ver que a mediana será, portanto, $(40+41)/2=40,5$.

c) A moda da distribuição é o elemento mais frequente, que na tabela é fácil observar que temos 3 valores modais, que são 21, 25 e 40.

d) Para o cálculo dos quantis, façamos somente de alguns. Tomemos o 5º e o 95º percentis, o 2º e o 8º decis, o 1º e o 3º quartis (aplicaremos sempre a fórmula dos percentis). Teremos:

- 5% percentil: $i = 0,05(112 + 1) = 5,56$.
Portanto, $[i] = 5$ e $i - [i] = 0,56$, de modo que,
 $Q_{(0,05)} = (1 - 0,56)x_{(5)} + 0,56x_{(6)} = 0,44.21 + 0,56.21 \Rightarrow Q_{(0,05)} = 21$;
- 95% percentil: $i = 95(112 + 1)/100 = 107,35$. Portanto, $[i] = 107$ e $i - [i] = 0,35$, de modo que, $Q_{(0,95)} = (1 - 0,35)x_{(107)} + 0,35x_{(108)} = 0,65.113 + 0,35.114 \Rightarrow Q_{(0,95)} = 113,35$
- 2º decil: $i = 20(112 + 1)/100 = 22,6$.
Logo, $[i] = 22$ e $i - [i] = 0,6$, de modo que,
 $Q_{(0,2)} = (1 - 0,6)x_{(22)} + 0,6x_{(23)} = 0,4.26 + 0,6.26 \Rightarrow Q_{(0,2)} = 26$;
- 8º decil: $i = 80(112 + 1)/100 = 90,4$.
Logo, $[i] = 90$ e $i - [i] = 0,4$, de modo que,
 $Q_{(0,8)} = (1 - 0,4)x_{(90)} + 0,4x_{(91)} = 0,6.83 + 0,4.84 \Rightarrow Q_{(0,8)} = 83,4$;
- 1º quartil: $i = 25(112 + 1)/100 = 28,25$.
Portanto, $[i] = 28$ e $i - [i] = 0,25$, de modo que,
 $Q_{(0,25)} = (1 - 0,25)x_{(28)} + 0,25x_{(29)} = 0,75.28 + 0,25.28 \Rightarrow Q_{(0,25)} = 28$;
- 3º quartil: $i = 75(112 + 1)/100 = 84,75$.
Portanto, $[i] = 84$ e $i - [i] = 0,75$, de modo que,
 $Q_{(0,75)} = (1 - 0,75)x_{(84)} + 0,25x_{(29)} = 0,25.79 + 0,75.80 \Rightarrow Q_{(0,75)} = 79,75$.

Para realizar a extração no R, a função é a seguinte: dado um vetor de dados 'x', aplicamos o comando `quantile(x, p)` em que 'p' é o valor do percentil dividido por 100. Por exemplo, para encontrar o 27º percentil, basta fazer no R `quantile(x, 0.27)`, e em seguida surgirá o valor correspondente. No JASP, o caminho é `descriptives` → `descriptive statistics` → arrasta a variável de interesse para o campo `variables` → no campo `statistics`, clica em `percentiles` e insere, por exemplo, o valor 27.

Em Python, pode-se utilizar a função `percentile` da biblioteca `numpy`, a saber: `numpy.percentile(x, p)` em que ‘p’ é o valor do percentil dividido por 100.

Uma relevante função do R, que podemos agora compreender seu significado, é a extração dos ‘cinco números’ que descrevem um conjunto de dados. Tratam-se, ordenadamente, do valor mínimo ($x_{(1)}$), do primeiro quartil ($Q_{(0,25)}$), da mediana (Md), do terceiro quartil ($Q_{(0,75)}$) e do valor máximo ($x_{(n)}$). O procedimento é o seguinte: dado um vetor de dados ‘x’, aplicamos o comando `fivenum(x)`.

Alternativamente, é possível solicitar no R mais um detalhamento, com a inclusão da média além dos 5 números acima indicados. O procedimento é o seguinte: sendo ‘x’ o vetor de dados, aplicamos o comando `summary(x)`, e em seguida aparecerão os valores demandados. Aplicando este procedimento no conjunto de dados do exemplo, teremos no R o seguinte:

```
>x=c(15, 20, 20, 20, 21, 21, 21, 21, 21, 22, 22, 23, 23, 24, 24,
24, 25, 25, 25, 25, 25, 26, 26, 26, 26, 27, 27, 28, 28, 28, 28, 29,
29, 30, 30, 31, 31, 32, 33, 33, 34, 34, 35, 35, 35, 36, 36, 36, 37,
38, 39, 40, 40, 40, 40, 40, 41, 46, 47, 48, 50, 55, 56, 57, 58,
59, 62, 62, 63, 63, 65, 65, 67, 70, 70, 72, 72, 73, 75, 76, 77, 77,
77, 79, 80, 80, 82, 82, 82, 83, 84, 89, 90, 90, 93, 95, 96, 99, 99,
100, 102, 112, 112, 112, 113, 113, 113, 114, 118, 123, 124, 125)
# x é o conjunto de dados da variável ticket médio
> fivenum(x)
[1] 15.0 28.0 40.5 79.5 125.0
# Ordenadamente, temos: o primeiro valor é 15.0,
# o primeiro quartil é 28.0, a mediana é 40,5, #
# o terceiro quartil é 79,5 e o valor máximo é 125.0

> summary(x)
   Min.    1st Qu.    Median     Mean    3rd Qu.     Max.
 15.00    28.00    40.50    55.33    79.25   125.00
# Mostra o sumário de 5 números mais a média do conjunto de dados.
```

Estes cinco valores dão uma primeira ideia sobre a tendência central (pela mediana), da dispersão dos dados em relação a este valor (pelos quartis) e da amplitude geral dos dados (pelos valores máximo e mínimo). Por esta razão, em procedimentos exploratórios, os ‘cinco números’ já dão uma boa caracterização do conjunto de dados.

6.6. Outras medidas

As possibilidades de desenvolvimento de novas medidas são tantas e tão variadas que não seria possível apresentá-las aqui. Apresento somente as 4 principais médias, que são a média quadrática, a média geométrica, a média harmônica e a média generalizada.

– Média quadrática (M_q)

É a raiz quadrática da média dos quadrados dos valores da variável, ou seja, se $X = \{x_1, x_2, \dots, x_N\}$

$$M_q(X) = \sqrt{\frac{\sum_1^N x_i^2}{N}}, \text{ ou ainda, } M_q(X) = \sqrt{\frac{x_1^2 + x_2^2 + \dots + x_N^2}{N}}$$

No R, o procedimento de extração da média quadrática de um vetor 'x' é dado por `mq=sqrt(mean(x^2));mq`, ou ainda por `mq=mean(x^2)^0.5;mq`. Em Python, pode-se utilizar a biblioteca `numpy` e a função `np.sqrt(np.mean(x**2))`.

– Média geométrica (M_g)

É a N-ésima raiz do produto do conjunto de elementos de uma variável. Assim, se $X = \{x_1, x_2, \dots, x_N\}$, então

$$M_g(X) = \sqrt[N]{\prod_1^N x_i}, \text{ ou ainda, } M_g(X) = \sqrt[N]{x_1 \cdot x_2 \cdot \dots \cdot x_N}$$

No R ou JASP, para a extração da média geométrica de um vetor 'x', primeiramente informamos o número 'N' de elementos da amostra, fazendo `n=length(x)`, e em seguida calculamos a média geométrica pelo comando `mg=prod(x)^(1/n);mg`. Alternativamente, é possível encontrar a média geométrica pelo seguinte comando `mg=exp(mean(log(x)));mg`. Em Python, os comandos são equivalentes aos do R. Utilizaremos a biblioteca `numpy` e informamos o número 'N' de elementos da amostra, fazendo `n=len(x)`, e em seguida calculamos a média `mg=np.prod(x)**(1/n); print(mg)`. Alternativamente, temos mais duas possibilidades: (1) `mg = np.exp(np.mean(np.log(x))); print(mg)` ou (2) `from scipy.stats import gmean; mg=gmean(x); print(mg)`.

– Média harmônica (Mh)

A média harmônica é o inverso multiplicativo da média dos valores inversos do conjunto de dados da amostra. Desta forma, se $X = \{x_1, x_2, \dots, x_N\}$, então teremos

$$Mh(X) = \frac{N}{\sum_{i=1}^N \frac{1}{x_i}}, \text{ ou ainda, } M_q(X) = \frac{N}{\frac{1}{x_1} + \frac{1}{x_2} + \dots + \frac{1}{x_N}}$$

Para o software R ou JASP, no procedimento de extração da média harmônica de uma variável 'x' fazemos, primeiramente, a fixação do número de elementos da amostra, pelo comando `n=length(x)`, e em seguida calculamos a média harmônica fazendo `mh=n/sum(1/x)`; `mh`. Alternativamente, bastaria seguir diretamente para o comando `mh=1/mean(1/x)`; `mh`. O mesmo procedimento pode ser executado em Python. Podemos definir o número de elementos pelo comando `n=len(x)`, e em seguida calculamos a média harmônica, importando o `numpy` previamente, fazendo `mh=n/np.sum(1/x)`; `print(mh)`.

– Média generalizada (M)

A média generalizada de ordem p é definida como a p -ésima raiz da média aritmética das potências em 'p' dos valores da variável. Assim, se $X = \{x_1, x_2, \dots, x_N\}$, então teremos:

$$M(X) = \sqrt[p]{\frac{\sum_{i=1}^N x_i^p}{N}}, \text{ ou ainda, } M(X) = \sqrt[p]{\frac{x_1^p + x_2^p + \dots + x_N^p}{N}}.$$

É fácil ver que a média generalizada de fato generaliza as médias aritmética simples, quadrática e harmônica. De fato:

- Se $p = -1$, então $M(X) = Mh(X)$;
- Se $p = 1$, então $M(X) = Me(X)$;
- Se $p = 2$, então $M(X) = Mq(X)$;

No R ou JASP, o procedimento para extração da média generalizada é semelhante ao que foi apresentado para o cálculo da média quadrática, ou seja, dado um vetor 'x', definimos o fator 'p', e em seguida damos o comando `m=mean(x^p)^(1/p)`; `m`. Em Python, o procedimento é espelhado, teremos: dado um vetor 'x', definimos o fator 'p', e em seguida executamos `m=np.mean(x**p)**(1/p)`; `print(m)`.

6.7. Resumo

Neste capítulo foram apresentadas diversas medidas descritivas de posição de um conjunto de dados, com a finalidade de permitir um melhor entendimento a respeito da variável. Os destaques do capítulo foram os seguintes:

- Uma medida de posição é aquela que possui as seguintes características: a medida está entre o valor mínimo e o valor máximo do conjunto de dados; a multiplicação de um valor constante aplicada a todo o conjunto de dados implica na multiplicação do mesmo valor à medida de posição original;
- As principais medidas de posição são a média, a mediana e a moda, e a denominação utilizada é de ‘medidas de tendência central’;
- A média aritmética de um conjunto de dados é a soma dos valores dos dados dividida pelo total de observações realizadas. Se os dados representam todo o universo, temos então a média aritmética populacional; se forem dados de uma amostra, temos então a média aritmética amostral. Temos duas variações da média aritmética, que são a média ponderada e a média para dados agrupados;
- A média aritmética é muito sensível a *outliers*, e, por esta razão, foram desenvolvidos métodos alternativos de extração da média aritmética que minimizam problemas como este indicado acima, sendo os principais a média aparada (ou truncada) e a média winsorizada;
- A mediana de um conjunto de dados é o valor que divide estes dados em duas partes iguais em número de elementos. O problema em potencial da mediana é justamente o fato de ser um caso de média aparada extremo, que desconsidera muitos valores. Por esta razão, tem sido mais recorrente o uso das médias truncada e winsorizada como alternativa para o problema de *outliers*;
- Em estatística, a moda é definida como o valor mais frequente de um conjunto de dados, ou seja, a moda é o valor que aparece mais vezes, comparativamente aos demais;
- Os quantis são medidas estatísticas de posição que generalizam a mediana, e que dividem os dados em tantas partes quantas forem de interesse. São mais comuns as divisões em 100 partes (para os percentis), em 10 partes (para os decis), e em 4 partes (para os quartis);

- Ao final, foram apresentados mais 4 tipos de médias, que foram a média quadrática, a média geométrica, a média harmônica e a média generalizada, esta última que pode se converter em outras médias a partir da manipulação do parâmetro;
- Em cada uma das medidas analisadas foram dados exemplos, e apresentadas as rotinas de extração do Excel, do R e em Python;
- A seguir temos a notações utilizadas em cada medida:

Símbolo	Significado
Me	Média aritmética simples
Mp	Média aritmética ponderada
Mt	Média de dados tabulados
Mtr	Média aparada
Mw	Média winsorizada
Mq	Média quadrática
Mg	Média geométrica
Mh	Média harmônica
M	Média generalizada
Md	Mediana
Mo	Moda
$Q_{(p)}$	Quantil de ordem p

Exercícios

1. Demonstre as propriedades matemáticas da média aritmética.
2. Construa as tabelas e os gráficos apropriados para os dados dos exemplos 1, 4, 7 e 10. Em seguida, calcule as médias, medianas e modas a partir dos dados tabulados, e compare os resultados com as medidas extraídas pelos dados detalhados.
3. Considere os dados do exemplo 1, e calcule, usando o software mais apropriado:
 - a) A média aparada com 5, 10 , 15 e 20%.
 - b) A média winsorizada com 5, 10 , 15 e 20%.
 - c) A mediana e a moda

- d) Todos os decis e todos os quartis
 - e) As médias quadrática, geométrica e harmônica.
 - f) Compare todos os resultados e comente-os.
4. Considere os dados da variável do exemplo 10 e, usando o software mais apropriado:
- a) Calcule as médias aritmética, quadrática, geométrica e harmônica.
 - b) Use o comando da média generalizada de ordem 'p' e verifique as relações indicadas (para $p=-1$, $p=1$ e $p=2$).
 - c) Desenvolva simulações da média generalizada para valores decrescentes (ou seja, faça $p \rightarrow -\infty$), e observe que há convergência o valor mínimo da série de dados (ou seja, $\lim_{p \rightarrow -\infty} M(X) = x_{(1)}$).
 - d) Desenvolva simulações da média generalizada para valores crescentes (ou seja, faça $p \rightarrow +\infty$), e observe que há convergência o valor máximo da série de dados (ou seja, $\lim_{p \rightarrow +\infty} M(X) = x_{(n)}$).
5. Considere novamente os dados do exemplo 13.
- a) Sabendo que o lucro líquido sobre o ticket médio é de 7%, aponte uma estimativa do lucro total de uma semana e de um mês, tomando por base a média aritmética simples, e as médias aparada e winsorizada (10 e 15%) (ou seja, são 3 estimativas para a semana e para o mês).
 - b) Na situação de decisão em que o gestor do frigorífico se encontra, foi sugerido que deixasse de atender cerca de 40% dos clientes para poder atender aos potenciais clientes institucionais. Neste sentido, indique a nova estimativa de lucro semanal e mensal da conta de clientes residenciais.
 - c) No levantamento de interesse de churrascarias e restaurantes, o gerente espera fechar contrato de fornecimento para cerca de 30 empresas, cada um com potencialidade de render por semana cerca de R\$ 600,00, porém a margem cai para 3%. Calcule então o lucro esperado semanal e mensal de clientes institucionais.
 - d) Você foi convidado(a) a opinar sobre a decisão, levando em conta o lucro e possíveis questões operacionais do atendimento de cada tipo de cliente. O que recomenda? Fundamente suas recomendações.

6. Um estudante de comportamento organizacional empreendeu uma pesquisa para avaliar o grau de comprometimento dos trabalhadores com a organização, e fez um levantamento de uma amostra de 54 pessoas, em uma escala de phrase completion de 0 (nenhum comprometimento) a 10 (comprometimento muito forte). Os resultados estão indicados abaixo, já em ordem crescente (da esquerda para direita).

0	1	2	2	3	3	4	4	4
4	4	4	5	5	6	6	6	6
6	7	7	7	7	7	7	7	7
8	8	8	8	8	8	8	8	8
8	8	8	9	9	9	9	9	9
9	9	9	9	10	10	10	10	10

- a) Construa uma tabela de frequências dos dados da pesquisa.
- b) Construa o gráfico correspondente.
- c) Calcule a média aritmética com base nos dados detalhados e tabulados e compare os resultados.
- d) Calcule as médias aparada e winsorizada a 10 e a 20%. Compare os resultados com a média e avalie potenciais efeitos de outliers.
- e) Calcule os 3 quartis com base nos dados tabulados e detalhados.
- f) Calcule as médias quadrática, geométrica e harmônica. Compare os resultados com a média aritmética.

CAPÍTULO 7

MEDIDAS DESCRITIVAS DE DISPERSÃO

No capítulo 6, vimos as medidas de posição, e foi fácil notar que, se as medidas de posição dão uma ideia do valor típico de um conjunto de dados, isto não é suficiente para termos uma real noção de como os dados se caracterizam. De fato, como vimos nos diferentes exemplos, a maioria dos dados é diferente da medida de posição, ou seja, há desvios em relação ao valor típico, mesmo no caso da moda.

Neste capítulo, nosso foco será justamente na análise das variações que existem no conjunto de dados. Assim, além das medidas de posição, encontraremos as medidas de variação, que acrescentarão mais informações sobre os dados. No primeiro item, trataremos das medidas absolutas, com destaque para o desvio padrão e o desvio mediano absoluto, e no segundo item trataremos das medidas relativas. Em todos os casos, serão apresentados os procedimentos de extração das medidas nos softwares. Ao final, apresentaremos um gráfico especial e de grande relevância na análise exploratória de dados.

Ao final deste capítulo o leitor deverá estar apto a responder às seguintes questões:

- O que são medidas de dispersão?
- Quais as principais medidas de dispersão e como é possível obtê-las a partir de um conjunto de dados?
- De que forma o conhecimento das diferentes medidas será útil à análise exploratória de dados?

7.1. Principais medidas de dispersão absoluta

As principais medidas de dispersão absoluta são a amplitude geral, a amplitude interquartil, a amplitude semi-interquartil, o desvio médio (da média e da mediana), o desvio mediano absoluto, o desvio padrão, a variância.

- Amplitude total

A amplitude total, representada por At , é a diferença entre o valor máximo e o valor mínimo do conjunto de dados. Assim, dado um conjunto de dados $X = \{x_1, x_2, \dots, x_N\}$, então teremos:

$$At(X) = x_{(N)} - x_{(1)}$$

No software R, a amplitude pode ser obtida da seguinte maneira: dado um vetor 'x', a amplitude é 'at' vem de $at = \max(x) - \min(x)$; `at`. No JASP, o caminho é `descriptives` → `descriptive statistics` → arrasta a variável de interesse para o campo `variables` → no campo `statistics`, clica em `range`. Em Python, podemos obter da seguinte maneira: `at = np.max(x) - np.min(x); print(at)`.

- Amplitude interquartil

É fácil perceber que, assim como a média aritmética, a amplitude é uma medida fortemente influenciada por potenciais *outliers*. Assim, um conceito adicional que 'apara' os valores extremos pode ser obtido pela diferença entre quantis equidistantes em relação à mediana. Dentre as medidas possíveis, a mais comumente usada é a diferença entre o terceiro e o primeiro quartis. Esta diferença corresponde à amplitude dos valores que comportam 50% dos valores do conjunto de dados. Simbolicamente, dado um conjunto de dados $X = \{x_1, x_2, \dots, x_N\}$, então a amplitude interquartil (AIQ) de X será dada por:

$$AIQ(X) = Q_{(0,75)}(X) - Q_{(0,25)}(X)$$

No software R ou JASP, a amplitude interquartil pode ser obtida da seguinte maneira: dado um vetor 'x', a amplitude é 'aiq' vem de $aiq = \text{quantile}(x, 0.75) - \text{quantile}(x, 0.25)$; `aiq`. A amplitude semi-interquartil nada mais é que a metade da amplitude interquartil. Em Python, podemos executar da seguinte maneira: (1) amplitude interquartil: `aiq = np.percentile(x, 75) - np.percentile(x, 25); print(aiq)`. (2) amplitude semi-interquartil: `asiq = aiq/2; print(asiq)`.

- O desvio médio

O desvio médio é a média dos módulos dos desvios dos valores em relação à média aritmética (neste caso, temos o desvio médio para a média) ou à mediana (neste caso, temos o desvio médio para a mediana; não se calcula

o desvio médio para a moda). Simbolicamente, para um conjunto de dados $X = \{x_1, x_2, \dots, x_N\}$, teremos:

- O desvio médio para a média (DMe): para cada x_i , teremos o desvio em relação à média dado por $dme_i = (x_i - Me(X))$, e o módulo dado por $|dme_i| = |x_i - Me(X)|$. Portanto, teremos:

$$DMe(X) = \frac{\sum_1^N |dme_i|}{N}, \text{ ou ainda, } DMe(X) = \frac{\sum_1^N |x_i - Me(X)|}{N}$$

- O desvio médio para a mediana (DMd): para cada x_i , teremos o desvio em relação à mediana dado por $dmd_i = (x_i - Md(X))$, e o módulo dado por $|dmd_i| = |x_i - Md(X)|$. Portanto, teremos:

$$DMd(X) = \frac{\sum_1^N |dmd_i|}{N}, \text{ ou ainda, } DMd(X) = \frac{\sum_1^N |x_i - Md(X)|}{N}$$

Para o caso do desvio médio, a utilização do módulo dos desvios (e não dos desvios em si) deve-se ao fato de os desvios em relação à média terem soma zero (demonstre isto).

O software Excel tem a opção de cálculo do desvio médio em relação à média a partir do comando ‘=des.médio()’, com a indicação entre parênteses do conjunto de valores de interesse. No R ou no JASP, os procedimentos de extração das duas modalidades de desvio médio são bastante similares. Assim, dado um vetor de dados ‘x’, o desvio médio em relação à média (dme) pode ser obtido assim: `dme=mean(abs(x-mean(x)));dme`. Já para o desvio médio em relação à mediana (dmd), basta fazer: `dme=mean(abs(x-median(x)));dme`. Em Python, temos as seguintes possibilidades de cálculo: (1) para o desvio médio em relação à média, `dme = np.mean(np.abs(x - np.mean(x))); print(dme)`; (2) para o desvio médio em relação à mediana, `dmd = np.mean(np.abs(x - np.median(x))); print(dmd)`.

– O desvio absoluto mediano

O desvio mediano absoluto, representado por ‘DAMd’ é a mediana dos módulos dos desvios dos valores em relação à mediana. Simbolicamente, para um conjunto de dados $X = \{x_1, x_2, \dots, x_N\}$, seja $DMD = \{|x_1 - Md(X)|, \dots, |x_N - Md(X)|\}$. Então teremos:

$$DAMd(X) = Md(DMD)$$

O desvio mediano absoluto é uma medida de variação de uso menor, comparativamente ao desvio mediano absoluto normalizado, representado por MADN. Não cabe agora explicar as razões, mas por definição teremos:

$$MADN(X) = \frac{DAMd(X)}{0,6745}$$

No R, o procedimento mais fácil e já programado é a extração do desvio mediano absoluto normalizado. Assim, dado um vetor 'x' de dados, o desvio mediano absoluto normalizado é calculado fazendo `mad(x)`. Para o cálculo especificamente do desvio mediano absoluto temos duas opções: primeiro, multiplicamos o desvio mediano absoluto normalizado por 0,6745, ou seja, `damd=0.6745*mad(x)`; `damd`. Alternativamente, é possível fazer o cálculo direto, assim: `damd=median(abs(x-median(x)))`; `damd`.

No JASP, o caminho é `descriptives` → `descriptive statistics` → arrasta a variável de interesse para o campo `variables` → no campo `statistics`, clica em `MAD` (desvio mediano absoluto).

Em Python, seguiremos a mesma lógica aplicada ao código em R. Primeiramente, no entanto, teremos que importar a biblioteca `numpy` por meio do comando `import numpy as np`. Na sequência, iremos calcular desvio mediano absoluto normalizado: `def mad(x): return 1.4826*np.median(np.abs(x-np.median(x)))`; `print(mad(x))`. Para o desvio mediano absoluto, tem-se: `def damd(x): return 0.6745*mad(x)`; `print(damd(x))`.

– O desvio padrão

O desvio padrão é, de longe, a medida de dispersão mais utilizada, principalmente nas técnicas de análise de dados clássicas e paramétricas. Sua definição é menos intuitiva, mas tem todo sentido quando analisamos situações práticas. Por definição, o desvio padrão, representado por Dp , é a média quadrática dos desvios em relação à média do conjunto de dados. Sendo o conjunto de dados $X = \{x_1, x_2, \dots, x_N\}$, então,

$$Dp(X) = \sqrt{\frac{\sum_1^N (x_i - Me(X))^2}{N}}$$

Há um ajuste de correção desta fórmula, que agora não temos condições de explicitar as razões formais, que faz com que a fórmula, no trabalho com

amostras de tamanho n , da raiz tenha sua divisão por $(n-1)$ e não por n , como parece ser mais lógico. Do ponto de vista de avaliação da medida para efeito de análise exploratória, é indiferente usar uma ou outra fórmula. No entanto, para efeito de operacionalização de técnicas estatísticas mais sofisticadas, é necessário utilizar a fórmula com a divisão por $(n-1)$. Este é o procedimento utilizado nos softwares.

No Excel, o desvio padrão é calculado pelo comando `=desv.pad()`, com a indicação entre parênteses do conjunto de valores de interesse. No R, o desvio padrão de uma variável representada em um vetor `x` é calculado assim: `sd(x)`. No JASP, o caminho é `descriptives` → `descriptive statistics` → arrasta a variável de interesse para o campo `variables` → no campo `statistics`, clica em `Std. deviation` (desvio médio em relação à média). Em Python, o desvio padrão de uma variável pode ser calculado utilizando a função `statistics.stdev(x)` do pacote `statistics`.

- A variância

A variância de um conjunto de dados é a média dos quadrados dos desvios em relação à média. Obviamente, a variância é o quadrado do desvio padrão. Para seu cálculo, no conjunto de dados $X = \{x_1, x_2, \dots, x_N\}$, teremos

$$Var(X) = \frac{\sum_1^N (x_i - Me(X))^2}{N}$$

No R, a variância de um vetor `x` pode ser obtida de duas formas: de maneira direta, basta fazer `var(x)`; a partir do desvio padrão, basta fazer `v=sd(x)^2`;v. No JASP, o caminho é `descriptives` → `descriptive statistics` → arrasta a variável de interesse para o campo `variables` → no campo `statistics`, clica em `variance`. Em Python, podemos utilizar o pacote `statistics` e a função `statistics.variance(x)`.

A variância é uma medida que, para análise exploratória de dados, parece ter pouca utilidade, mas, em outras análises, esta medida ganha uma relevância muito maior do que pode parecer. Da mesma forma que ocorreu com a média aritmética e com a amplitude, a variância e o desvio padrão são medidas sensíveis a valores extremos. Por esta razão, foram derivadas medidas robustas de variância (e, por consequência, de desvio padrão), pela variância aparada e pela variância winsorizada. Vejamos os detalhes.

– Variância aparada

A variância aparada ou truncada, representada por Vtr , é a média aritmética dos quadrados dos desvios dos valores do subconjunto da variável inicial em relação à média aparada, após a exclusão de um determinado percentual de valores (a aparta) nas fronteiras inferior e superior dos dados. Para seu cálculo, considerando o conjunto de dados $X = \{x_1, x_2, \dots, x_N\}$, com aparta de t elementos, onde $t = \lfloor Np/100 \rfloor$, sendo p o percentual de truncagem, teremos

$$Vtr(X) = \frac{\sum_{i=t+1}^{N-t} (x_i - Mtr(X))^2}{N - 2t} \Rightarrow$$

$$Vtr(X) = \frac{(x_{(t+1)} - Mtr(X))^2 + \dots + (x_{(N-1-t)} - Mtr(X))^2 + (x_{(N-t)} - Mtr(X))^2}{N - 2t}$$

Por este procedimento, são excluídos os valores extremos que podem inflacionar a variância e o desvio padrão. Isto faz da variância aparada uma medida robusta de dispersão.

No R ou JASP, temos algumas opções de cálculo da variância aparada e por consequência, do desvio padrão aparado. Vejamos duas opções (ver Quadro 7.1), considerando um vetor de dados 'x':

Quadro 7.1 – Programa do R para variância aparada

```
#Opção 1
x=c(...)                                #Vetor de dados
install.packages("chemometrics")        #Instala o pacote de base
library(chemometrics)                   #Habilita a biblioteca
tr=...                                  #Define o valor de aparta
sdtr=sd_trim(x, tr);sdtr                #Calcula o desvio aparado

#Opção 2
x=c(...)                                #Vetor de dados
tr=...                                  #Define o valor de aparta

#Um novo vetor y é derivado pela aplicação das aparas em x
y=subset(x, x>quantile(x, (t/100) & x<quantile(x, (100-t)/100))
sdtr=sd(y);sdtr                         #Calcula o desvio aparado
vtr=var(y);vtr                          #Calcula a variância aparada
```

Vejamos, agora, o caminho para calcular a variância aparada em Python (Quadro 7.2):

Quadro 7.2 – Programa em Python para variância aparada

```
import numpy as np

def trimmed_variance(x, tr):
    lower_quantile = np.percentile(x, tr)
    upper_quantile = np.percentile(x, 100-tr)
    trimmed_x = x[(x > lower_quantile) & (x < upper_quantile)]
    return np.var(trimmed_x, ddof=1)
print(trimmed_variance(x, tr))
```

– Variância winsorizada

A variância winsorizada, representada por V_w , é a variância da variável winsorizada, ou seja, aquela em que seus valores são definidos pela substituição de um determinado percentual de valores nas fronteiras inferior e superior dos dados pelos últimos valores vizinho do que seria excluído. Para seu cálculo, considerando o conjunto de dados $X = \{x_1, x_2, \dots, x_N\}$, com n elementos, onde $w = \lfloor Np/100 \rfloor$, sendo p o percentual de winsorização, teremos

$$V_w(X) = \frac{(w+1) \left[(x_{(w+1)} - Mw(X))^2 + (x_{(N-w)} - Mw(X))^2 \right] + \sum_{i=w+2}^{N-w-1} (x_i - Mw(X))^2}{N}, \text{ ou ainda,}$$

$$V_w(X) = \frac{(w+1) \left[(x_{(w+1)} - Mw(X))^2 \right] + (x_{(w+2)} - Mw(X))^2 + \dots + (x_{(N-w-1)} - Mw(X))^2 + (w+1) \left[(x_{(N-w)} - Mw(X))^2 \right]}{N}$$

No R ou JASP, assim como no caso da variância aparada, não existe também uma forma fixada de cálculo da variância winsorizada, sendo necessário programar o cálculo. Duas alternativas possíveis de extração, dado um vetor de dados 'x', são as seguintes (Quadro 7.3):

Quadro 7.3 – Programa do R para variância winsorizada

```
#Opção 1
x=c(...)           #Vetor de dados
install.packages("psych") #Instala o pacote de base
library(psych)      #Habilita a biblioteca
w=...              #Define o valor de winsorização
winsor.var(x, w)    #Calcula a variância winsorizada
winsor.sd(x, w)     #Calcula o desvio winsorizado
```

```

#Opção 2
x=c(...)           #Vetor de valores
w=...              #Nível de winsorização
y<-sort(x)         #Ordenando x
n<-length(x)       #Fixando o tamanho de x
ibot<-floor(w*n)+1  #Quantidade inferior
itop<-n-ibot+1      #Quantidade superior
xbot<-y[ibot]       #Valor de substituição inferior
xtop<-y[itop]       #Valor de substituição superior
y<-ifelse(y<=xbot,xbot,y) #Substituição inferior
y<-ifelse(y>=xtop,xtop,y) #Substituição inferior
vw=var(y);vw        #Calcula a variância winsorizada
sdw=sd(y);sdw       #Calcula o desvio winsorizado

#O leitor deve observar uma pequena diferença nas duas
#opções decorrentes de uma ‘correção’ implementada no pacote
#‘psych’. Recomendo particularmente a Opção 2.

```

Vejam, agora, o caminho para calcular a variância winsorizada em Python (Quadro 7.4):

Quadro 7.4 – Programa em Python para variância winsorizada

```

import numpy as np

x=np.array([seus dados])
w=coeficiente de winsorização

def winsorized_variance(x, w):
    x_sorted = np.sort(x)
    n=len(x)
    ibot=int(np.floor(w*n))
    itop=n-ibot-1
    xbot=x_sorted[ibot]
    xtop=x_sorted[itop]
    x_winsorized=np.clip(x, xbot, xtop)
    return np.var(x_winsorized, ddof = 1)
print(winsorized_variance(x,w))

def winsorized_sd(x,w):
    return np.sqrt(winsorized_variance(x,w))
print(winsorized_sd(x,w))

```

7.2. Principais medidas de dispersão relativa

As principais medidas de dispersão relativa são a variância relativa (à média), o coeficiente de variação de Pearson, o desvio quartil reduzido, o coeficiente quartílico de variação.

– Variância relativa e coeficiente de variação de Pearson

A variância relativa ($CVar$) consiste na razão entre a variância calculada e o quadrado da média aritmética. Assim, se $X = \{x_1, x_2, \dots, x_N\}$, então:

$$CVar(X) = \frac{Var(X)}{Me(X)^2}$$

Já o coeficiente de variação de Pearson (CV) é a raiz quadrada da variância relativa, ou seja, é a razão entre o desvio padrão e a média aritmética do conjunto de dados. No mesmo conjunto de dados acima, teremos:

$$CV(X) = \frac{Dp(X)}{Me(X)}$$

No R, o procedimento de cálculo é a simples extração das respectivas razões. Fica como exercício a aplicação no software R, em Python e nos demais softwares.

O leitor deve observar que a variância relativa e o coeficiente de variação podem ter formas alternativas, como as razões entre a média aparada com o desvio padrão aparado, ou a variância winsorizada e o quadrado da média winsorizada. Fica como exercício a verificação destas e de outras alternativas.

– Desvio quartil reduzido e coeficiente quartílico

O desvio quartil reduzido, representado por DQr , é a razão entre o desvio semi-interquartil e a mediana. Ou seja, dado o conjunto de dados $X = \{x_1, x_2, \dots, x_N\}$, então:

$$DQr(X) = \frac{\frac{Q_{(0,75)}(X) - Q_{(0,25)}(X)}{2}}{Md(X)}, \text{ ou ainda,}$$

$$DQr(X) = \frac{Q_{(0,75)}(X) - Q_{(0,25)}(X)}{2Md(X)}$$

O coeficiente quartílico de variação, indicado por CVq , é a razão entre o desvio interquartil e a soma do primeiro e do terceiro quartis. No mesmo conjunto de dados X , teremos:

$$CVq(X) = \frac{Q_{(0,75)}(X) - Q_{(0,25)}(X)}{Q_{(0,75)}(X) + Q_{(0,25)}(X)}$$

Os procedimentos no R, JASP e em Python consistem em escrever estas fórmulas, seguindo os comandos anteriormente indicados. A construção dos procedimentos fica como exercício.

7.3. O gráfico *boxplot*

De posse dos conhecimentos acumulados nos capítulos 6 e 7, poderemos agora construir um novo gráfico de análise univariada, que tem as medidas dos quantis e dos desvios indicados como base. Trata-se do *boxplot* ou ‘gráfico de caixas’, cuja construção veremos a seguir.

- Primeiramente, tomamos como referência de análise os três quartis, ou seja, o primeiro e o terceiro quartis mais a mediana, e os plotamos no eixo cartesiano das ordenadas;
- Traçamos em cada ponto segmentos de reta congruentes (a largura fica a critério do pesquisador);
- Os extremos à direita e à esquerda dos segmentos traçados são ligados de modo a formar um retângulo, cuja base inferior é o primeiro quartil e a superior é o terceiro quartil (portanto, altura do retângulo será o desvio interquartil);
- Acima do terceiro quartil será demarcada uma linha definida horizontal de largura igual ou menor que aquela construída sobre os quartis, distando do quartil uma vez e meia o intervalo interquartil. Esta linha constitui o limite superior (LS);
- Com a mesma característica, será uma linha abaixo, distando do primeiro quartil uma vez e meia o intervalo interquartil. Esta linha constitui o limite inferior (LI).

Veja que, por estas delimitações, teremos que $LS = Q_{(0,75)} + 1,5(Q_{(0,75)} - Q_{(0,25)})$, e ainda $LI = Q_{(0,25)} - 1,5(Q_{(0,75)} - Q_{(0,25)})$. A Figura 7.1 apresenta uma visão vertical do gráfico de caixas.

Figura 7.1 – Visão horizontal do *boxplot*

No R, a construção do *boxplot* é direta, ou seja, dado um vetor de dados ‘x’, basta escrever `boxplot(x)`, e o software abrirá uma tela contendo o gráfico. No JASP, iremos em *descriptives* → *plot builder* → arraste as variáveis de interesse para os campos “X AND Y-Axis Variable”. Em seguida, vamos na função *distributions* (*histogram*, *boxplot*, *violin*) → *Boxplot*. Em Python, podemos utilizar a biblioteca *matplotlib* e os seguintes comandos: `import matplotlib.pyplot as plt; dados = [seu conjunto de dados]; plt.boxplot(dados); plt.show()`.

A relevância central do *boxplot* para a análise exploratória de dados está justamente na possibilidade que o gráfico dá para visualização de valores extremos. Valores acima do limite superior, e abaixo do limite inferior constituirão *outliers*. Assim, os valores que estiverem dentro de tais limites são considerados valores regulares, e os que estiverem fora desses limites serão marcados com asterisco, o que sinalizará que se trata de um valor extremo.

7.4. Identificação de extremos

A forma de identificação usada no *boxplot* é na verdade somente uma das alternativas que temos agora disponíveis para verificação de dados extremos. O conhecimento acumulado até aqui permite então analisarmos outras duas principais opções de identificação univariada, que são a padronização em Z e a padronização em relação ao desvio mediano normalizado. Vejamos cada uma das opções.

– Padronização em Z

A padronização em Z é um procedimento de larga utilização na teoria estatística e consiste em associar para cada valor do conjunto de dados um novo valor que é igual ao seu desvio em relação à média dividido pelo desvio padrão. Ou seja, dado um conjunto de observações $X = \{x_1, x_2, \dots, x_N\}$, então, para cada observação x_i teremos um valor padronizado z_i correspondente, assim calculado:

$$z_i = \frac{x_i - Me(X)}{Dp(X)}$$

É fácil provar que a média do conjunto de valores padronizados é igual a 0 e seu desvio padrão é igual a 1. Portanto, se um dado valor padronizado for igual a 1,5, isto significa que o valor está 1,5 desvios padrão acima da média. Nestes termos, a decisão central consiste em definir uma distância (em desvios padrão) em relação à média que seja considerada regular, a partir da qual qualquer outro valor possa ser considerado um valor extremo.

Há diversas possibilidades, e inclusive razões teóricas que justificam determinados valores. Sem maiores justificativas aqui, podemos definir a distância 2 como um primeiro ponto de corte, e, se tivermos muitos dados (acima de 200), podemos adotar a distância entre 3 e 3,5. Ou seja, consideraremos, em um primeiro momento, que um valor de z_i maior que 2 ou menor que -2 é um *outlier*, mas, se tivermos muitos dados, podemos ser mais tolerantes, e adotar um valor de z_i entre -3 e 3 , ou entre $-3,5$ e $3,5$.

– Padronização em relação ao desvio mediano normalizado

Esta padronização consiste na associação de cada valor do conjunto de dados com a divisão de seu desvio em relação à mediana pelo desvio mediano absoluto normalizado do conjunto de dados. Ou seja, dado um conjunto de dados $X = \{x_1, x_2, \dots, x_N\}$, então, para cada observação x_i teremos um valor padronizado zmd_i correspondente, assim calculado:

$$zmd_i = \frac{x_i - Md(X)}{MADN(X)}$$

Assim como no caso da padronização em Z , precisaremos definir um valor a partir do qual saberemos se o dado correspondente é ou não um *outlier*. Temos alternativas de referência cuja justificativa não temos como justificar agora, mas, para efeito de análise, podemos usar os mesmos referenciais da padronização em Z , ou seja, em geral, temos sinalização de *outlier* se os valores corresponderem a padronizações tais que $zmd_i < -2$ ou $zmd_i > 2$. Para amostras grandes, podemos tomar como *outliers* os valores cuja padronização gerem zmd_i tais que $zmd_i < -3$ ou $zmd_i > 3$, ou ainda $zmd_i < -3,5$ ou $zmd_i > 3,5$.

7.5. Resumo

Neste capítulo, apresentamos as diversas medidas descritivas de dispersão de dados, como medidas complementares das medidas de posição apresentadas no capítulo anterior. Os destaques do capítulo foram os seguintes:

- Uma medida de dispersão é aquela em que indica a variação dos valores de um conjunto de dados em relação a uma medida de posição. Normalmente temos medidas de dispersão relativas à média e à mediana;
- Há dois tipos principais de medidas de dispersão: as medidas de dispersão absoluta e as medidas relativas. Algumas dessas medidas são robustas em relação à presença de *outliers*;
- As principais medidas de dispersão absoluta são a amplitude total (geral e interquartil), o desvio semi-interquartil, o desvio médio e o mediano, além do desvio padrão e da variância;
- A amplitude total, o desvio médio, o desvio padrão e a variância são medidas não robustas à presença de *outliers*, e por isto sua análise é melhor quando complementada pelas demais medidas, que são robustas. Em particular para o desvio padrão e a variância, temos ainda a opção das medidas truncadas e winsorizadas;
- As principais medidas de dispersão relativa são o coeficiente e variância relativa, o coeficiente de variação, o desvio quartil reduzido e coeficiente quartílico;
- A construção conceitual deste capítulo e do anterior permite a construção do gráfico de caixas, ou boxplot, que, além de ser um gráfico relevante para interpretação dos dados, dá a indicação de valores extremos;
- Em complemento, foram ainda avaliadas duas estratégias de identificação de valores extremos, uma baseada na padronização em Z, e outra baseada no desvio mediano normalizado;
- Em cada medida, foram indicados os procedimentos de acesso nos pacotes estatísticos de referência.

Exercícios

1. Demonstre que a média dos desvios de um conjunto de dados $X = \{x_1, x_2, \dots, x_N\}$ em relação à sua média é igual a zero, ou seja,

$$D(X) = \frac{\sum_1^n dme_i}{n} = 0, \text{ onde } dme_i = (x_i - Me(X))$$

2. Demonstre que o desvio padrão de dados não constantes é menor que o desvio médio, ou seja, se $X = \{x_1, x_2, \dots, x_N\}$, então $Dp(X) < Dm(X)$.

3. Demonstre que a padronização em Z de uma variável $X = \{x_1, x_2, \dots, x_N\}$ gera uma nova variável $Z_X = \{z_1, z_2, \dots, z_N\}$, na correspondência ordenada dos índices, de tal modo que a média de Z_X é 0 e sua variância é 1, ou seja, se $z(x) = \frac{x - Me(X)}{Dp(X)}$, então $Me(Z_X) = 0$, $Var(Z_X) = 1$.

4. Considere os dados dos exemplos 1, 4, 7 e 10 do capítulo 6, e calcule, usando preferencialmente o R e a linguagem Python, todas as medidas de dispersão absoluta e relativa. Construa nestes mesmos exemplos o gráfico boxplot e verifique os potenciais outliers usando o gráfico e as medidas de detecção indicadas.

5. Um professor de Psicologia resolveu investigar a relação entre tempo de estudo e desempenho dos alunos de suas 3 disciplinas. Para tanto, levantou mediante um questionário a quantidade de tempo, em horas, dedicados às disciplinas. A amostra de tempo, conforme declarado pelos alunos, está indicada abaixo:

1	11	2	3	0	3	0	1	3	3	3	1	2
3	3	2	0	3	0	1	0	1	1	0	2	5
3	1	8	0	0	1	3	2	3	0	3	0	3
5	2	3	4	3	5	3	4	5	8	2	12	1

Com base nestes dados, realize os procedimentos abaixo:

- a) Organize os dados por meio de uma tabela, com frequências organizadas por intervalos.
- b) Faça a apresentação gráfica dos dados por meio do histograma.
- c) Extraia as medidas descritivas de tendências central, mais os quartis e os decis.

- d) Extraia as médias geométrica, quadrática, harmônica e generalizada de ordem 3 (cúbica) e 4 (biquadrática). Comente os resultados.
- e) Extraia as medidas de dispersão observadas absolutas e relativas.
- f) Construa o boxplot e avalie os potenciais outliers, considerando o gráfico e das demais alternativas de detecção.

O professor quer fixar uma medida de horas que possibilite a premiação de 5% dos alunos mais empenhados. Indique qual deve ser o ponto de corte.

CAPÍTULO 8

MEDIDAS DESCRITIVAS DE FORMATO_

Nos dois últimos capítulos, tratamos de duas categorias principais de medidas de descrição de dados, que são as medidas de posição e as medidas de dispersão. Ao passo que as primeiras apresentam sínteses dos dados que permitem avaliá-los pontualmente, as medidas de dispersão indicam o quanto os dados estão dispersos particularmente em relação às medidas de posição central. Mas estas medidas não são suficientes para a construção de uma visão mais informativa dos dados.

Com efeito, também é relevante saber como os dados se distribuem na variável, ou o padrão que desenvolvem quando são dispostos na forma de um histograma. Esta verificação é feita pelas ditas ‘medidas de formato’, que indicam a forma como o conjunto de dados se distribui levando em conta as perspectivas horizontal e vertical.

Temos então dois interesses de análise: na perspectiva horizontal, interessa saber se a distribuição está centrada de forma simétrica, ou se está mais dispersa para algum dos lados; já na perspectiva vertical, interessa saber se a distribuição está ‘centrada’, ‘achatada’ ou ‘pontiaguda’. As medidas para análise horizontal são medidas de **‘assimetria’**, e as medidas da análise vertical são as medidas de **‘curtose’**.

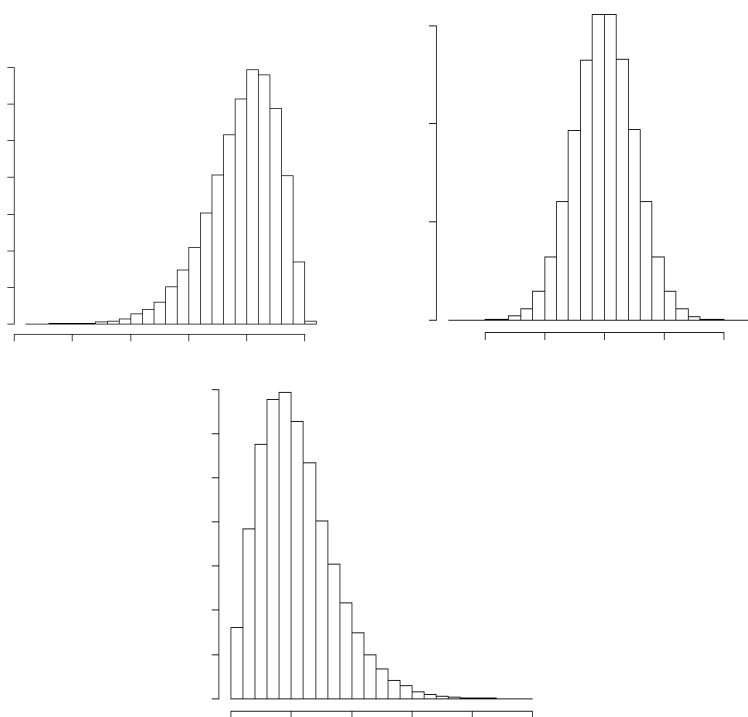
Ao final deste capítulo o leitor deverá estar apto a responder às seguintes questões:

- O que são medidas de formato?
- Quais as principais medidas de assimetria e de curtose, e como é possível obtê-las a partir de um conjunto de dados?
- Como essas medidas são úteis à análise exploratória de dados?

8.1. Principais medidas de assimetria

Uma variável é simétrica quando suas medidas de tendência central (média, mediana e moda) coincidem e quando a dispersão em relação a estas medidas segue o mesmo padrão. Visualmente, o histograma dá uma primeira ideia da simetria. Na Figura 8.1, vemos que, na primeira coluna, o histograma mostra uma variável concentrada à direita, caindo de forma suave à esquerda. Neste caso, dizemos que a distribuição é **assimétrica à esquerda**. A variável da terceira coluna tem comportamento diferente, apresentando ‘**assimetria à direita**’. Já a variável do meio está centrada, com comportamentos à esquerda e à direita semelhantes; dizemos que esta variável é **simétrica**.

Figura 8.1 – Ilustração da ideia de simetria



Observemos que, quando a distribuição é assimétrica à esquerda, a média ficará abaixo da moda (que é o valor mais frequente) e da mediana. Já no caso da assimetria à direita, a média será maior que a moda e que a mediana. Por outro lado, quando uma variável se distribui de forma simétrica, suas medidas de tendência central (média, mediana e moda) coincidem.

Estas características são regularidades que servem inclusive para se verificar quando uma variável é, ou não, simétrica, e caso seja assimétrica, se à

esquerda ou à direita. Em particular, quando estamos trabalhando com amostras de uma população que não sabemos o formato, se verificamos que a moda, a média e a mediana possuem valores próximos, temos uma sinalização forte de simetria, e caso contrário, a evidência é de assimetria na população de onde a amostra foi extraída. Caso contrário, podemos verificar se a média é menor (maior) que a mediana para avaliar se a variável tem distribuição assimétrica à esquerda (direita) (veja que não devemos fazer esta verificação pela moda, pois a variável pode ser multimodal).

Em termos de medidas, temos várias alternativas de verificação de formato, que exibo a seguir com breves comentários.

– Primeiro coeficiente de assimetria de Pearson

Consiste em uma medida baseada na relação entre a média, a mediana e o desvio padrão. Seja então uma variável $X = \{x_1, x_2, \dots, x_N\}$, tal que $Me(X)$, $Mo(X)$ e $Dp(X)$ são, respectivamente, a média, a moda e o desvio padrão de X . Então a assimetria (*skewness* em inglês), representada por $Sk(X)$, será dada por:

$$Sk(X) = \frac{Me(X) - Mo(X)}{Dp(X)}$$

Os valores de referência para interpretação dessa medida são os seguintes: se a medida for zero, a variável é simétrica, e se for diferente de zero, teremos assimetria (positiva à esquerda ou à direita a depender de o sinal ser negativo ou positivo). Esta medida é obviamente frágil, pois para ser única é requerido que tenhamos uma só moda. Isto, bem sabemos, não ocorre sempre.

– Segundo coeficiente de assimetria de Pearson

Consiste em uma medida baseada na relação entre a média, a mediana e o desvio padrão. Seja então uma variável $X = \{x_1, x_2, \dots, x_N\}$, tal que $Me(X)$, $Md(X)$ e $Dp(X)$ são, respectivamente, a média, a mediana e o desvio padrão de X . Então a assimetria, representada por $Sk(X)$, será dada por:

$$Sk(X) = \frac{3(Me(X) - Md(X))}{Dp(X)}$$

Os valores de referência para interpretação dessa medida são as seguintes.

- Se $Sk(X) = 0$, a variável é simétrica;
- Se $Sk(X) < 0$, a variável é assimétrica à esquerda, e se $Sk(X) > 0$ a assimetria é à direita;

Esta medida é sempre implementável no pacote R, da seguinte forma: dado um vetor 'x', a assimetria 'sk' vem de: $sk=3*(\text{mean}(x)-\text{median}(x))/\text{sd}(x)$; sk. Em Python, podemos utilizar a mesma lógica, com os seguintes comandos: `import numpy as np; df.skewness(x): return 3*(np.mean(x) - np.median(x))/np.std(x, ddof=1); print(skewness(x))`.

Embora a medida seja superior à primeira, pois qualquer variável (na verdade quase todas as variáveis) possui sua média, sua mediana e seu desvio padrão, temos a fragilidade de utilizarmos aqui duas medidas não robustas a valores extremos, que são a média e do desvio padrão.

- Coeficiente momento de assimetria

Tomando as mesmas referências de variáveis e medidas acima descritas, temos uma terceira medida, mais utilizada que as demais, possui a seguinte construção:

$$Sk(X) = \frac{\sum_{i=1}^N \frac{(x_i - Me(X))^3}{N}}{Dp(X)^3} = \frac{1}{N} \sum_{i=1}^N \left[\frac{(x_i - Me(X))}{Dp(X)} \right]^3$$

Os valores de referência para interpretação dessa medida são as seguintes:

- Se $0 \leq |Sk(X)| < 0,2$, a variável é dada por simétrica;
- Se $0,2 \leq |Sk(X)| \leq 1$, a variável é fracamente assimétrica, à esquerda se $Sk(X) < 0$, e à direita se $Sk(X) > 0$;
- Se $|Sk(X)| > 1$, a variável é fortemente assimétrica, à esquerda se $Sk(X) < 0$, e à direita se $Sk(X) > 0$.

Assim como o segundo coeficiente de assimetria de Pearson, esta medida é sempre implementável no pacote R, da seguinte forma: dado um vetor 'x', a assimetria 'sk' vem de $sk=(\text{sum}((x-\text{mean}(x))^3)/\text{length}(x))/\text{sd}(x)^3$;sk; ou ainda, $sk=\text{sum}(((x-\text{mean}(x))/\text{sd}(x))^3)/\text{length}(x)$;sk. No JASP, em *descriptives statistics*, há a opção de *skewness* para verificar a assimetria. Também é imple-

mentável em Python, da seguinte forma: `import numpy as np; def sk(x): return np.sum([(x-np.mean(x))/ np.std(x))**3] / len(x); print(sk(x))`.

Há bibliotecas implementadas no R que permitem a extração direta desse coeficiente. A principal é a biblioteca ‘moments’, e fica como exercício para o leitor acessar esta biblioteca e avaliar como é feito o procedimento de extração da medida.

Temos também aqui o problema da não robustez das duas medidas envolvidas (desvio padrão e média), o que torna esta medida de assimetria não robusta. Por esta razão, foram definidas algumas medidas de assimetria baseadas em medidas robustas a *outliers*, e baseadas em quantis. A seguir temos os dois apresentados abaixo.

– Medida robusta de assimetria de Bowley

Seja agora uma variável $X = \{x_1, x_2, \dots, x_N\}$, tal que $Q_{(0,25)}(X)$, $Q_{(0,50)}(X) = Md(X)$, e $Q_{(0,75)}(X)$ são, respectivamente, o primeiro quartil, a mediana e o terceiro quartil. Teremos então como medida de assimetria a seguinte:

$$Sk(X) = \frac{Q_{(0,75)}(X) + Q_{(0,25)}(X) - 2Q_{(0,50)}(X)}{Q_{(0,75)}(X) - Q_{(0,25)}(X)}$$

Os valores de referência para interpretação dessa medida são as seguintes:

- Se $Sk(X) = 0$, a variável é dada por simétrica;
- Se $Sk(X) < 0$, a variável é assimétrica à esquerda, e se $Sk(X) > 0$, a assimetria é à direita;

O procedimento de implementação desta medida no R (ou no JASP por meio do R console) é imediato: dado um vetor ‘x’, a assimetria ‘sk’ vem de: `sk=(-quantile(x,0.75)+quantile(x,0.25)-2*quantile(x,0.50))/(quantile(x,0.75)-quantile(x,0.25))`; `sk`. Em Python, temos a seguinte rotina: `import numpy as np; def sk_quantis(x): q75, q25, q50 = np.percentile(x, [75,25,50]); sk_quantis=(q75+-q25-2*q50)/(q75-q25); return sk_quantis; print(sk_quantis(x))`.

O procedimento é evidentemente robusto, mas tem a desvantagem de ser baseado em três medidas distantes de pontos extremos e não captam efeito de caldas da distribuição abaixo do primeiro e acima do terceiro quartis. Uma forma complementar de solucionar esta fragilidade está indicada abaixo.

- Medida robusta de assimetria de Kelley

Tomemos então uma variável $X = \{x_1, x_2, \dots, x_N\}$, tal que $Q_{(0,10)}(X)$, $Q_{(0,50)}(X) = Md(X)$, e $Q_{(0,90)}(X)$ são, respectivamente, o primeiro decil, a mediana e o nono decil. Teremos então como medida de assimetria a seguinte:

$$Sk(X) = \frac{Q_{(0,90)}(X) + Q_{(0,10)}(X) - 2Q_{(0,50)}(X)}{Q_{(0,90)}(X) - Q_{(0,10)}(X)}$$

A interpretação é a mesma que se aplica na assimetria de Bowley. O procedimento de implementação desta medida no R (ou no JASP por meio do R console) é imediato: dado um vetor 'x', a assimetria 'sk' vem de $sk = (\text{quantile}(x, 0.90) + \text{quantile}(x, 0.10) - 2 * \text{quantile}(x, 0.50)) / (\text{quantile}(x, 0.90) - \text{quantile}(x, 0.10))$; sk. Em Python, seguiremos a mesma lógica, qual seja: `import numpy as np; def sk_bowley(x): q90, q10, q50 = np.percentile(x, [90, 10, 50]); sk_bowley = (q90 + q10 - 2 * q50) / (q90 - q10); return sk_bowley; print(sk_bowley(x))`.

O leitor deve observar que a assimetria de Kelley é baseada em uma maior cobertura do conjunto de valores da variável. Outras configurações são evidentemente possíveis, tomando, por exemplo, outros quantis equidistantes da mediana e substituindo na fórmula acima. Veja então a seguinte construção:

Se $Q_{(p)}(X)$ é o quantil de ordem p da variável X , para $p > 0,5$. Então $Q_{(1-p)}(X)$ é o quantil simétrico em relação à mediana (quantil $Q_{(50)}(X)$). Podemos então definir uma medida de 'assimetria baseada em quantis', da seguinte fórmula:

$$Sk(X) = \frac{Q_{(p)}(X) + Q_{(1-p)}(X) - 2Q_{(0,5)}(X)}{Q_{(p)}(X) - Q_{(1-p)}(X)}, \quad p > 0,50.$$

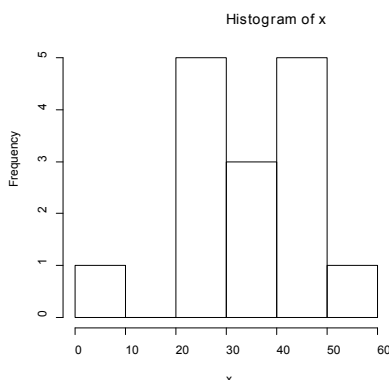
O leitor pode ver facilmente que as medidas de Bowley e Kelley são na verdade casos particulares desta construção (no primeiro tomando $p=0,75$, e no segundo com $p=0,90$). A interpretação é a mesma dos dois casos anteriores, e a implementação no R em Python fica como exercício. Vejamos agora um exemplo com todas as medidas.

Exemplo: O responsável pelo setor de planejamento de obras resolveu analisar a necessidade de ampliação de um estacionamento, e resolveu verificar em diferentes momentos do dia a quantidade de carros. A verificação foi feita durante uma semana, com uma contagem de vagas sobrando de manhã, de

tarde e de noite. Os resultados já na forma do R estão indicados abaixo (fica a implementação em Python e no JASP como exercício).

```
> x=c(28, 28, 24, 24, 41, 31, 44, 53, 41, 36, 33, 44, 28, 45, 3)
> #Segundo coeficiente de Pearson
> sk=3*(mean(x)-median(x))/sd(x); sk
[1] 0.1319102
>
> #Momento de assimetria
> sk=(sum((x-mean(x))^3)/length(x))/sd(x)^3;sk
[1] -0.6985153
>
> #Medida de Bowley
>sk=(quantile(x,0.75)+quantile(x,0.25)-2*quantile(x,0.50))/
+ (quantile(x,0.75)-quantile(x, 0.25)); sk
      75%
0.3103448
>
> #Medida de Kelley
>sk=(quantile(x,0.90)+quantile(x,0.10)-2*quantile(x,0.50))/
+ (quantile(x,0.90)-quantile(x,0.10)); sk
      90%
0.1262136
>
> #Coeficiente quantil para p=0.95 e 0,80
> p1=0.95; p2=0.80
>sk=(quantile(x,p1)+quantile(x,1-p1)-2*quantile(x,0.50))/
+ (quantile(x,p1)-quantile(x,1-p1)); sk
      95%
-0.03030303
>
>sk=(quantile(x,p2)+quantile(x,1-p2)-2*quantile(x,0.50))/
+ (quantile(x,p2)-quantile(x,1-p2)); sk
      80%
0.3095238
```

Em geral, observamos valores sempre pequenos, e oscilando entre negativo e positivo. Isto sugere que, possivelmente, a distribuição do número de vagas de carros nos diferentes turnos é simétrica. Para uma análise visual, extraímos o histograma, que está exposto abaixo, e que reafirma esta impressão, ou seja, a variável parece simétrica em formato horizontal.

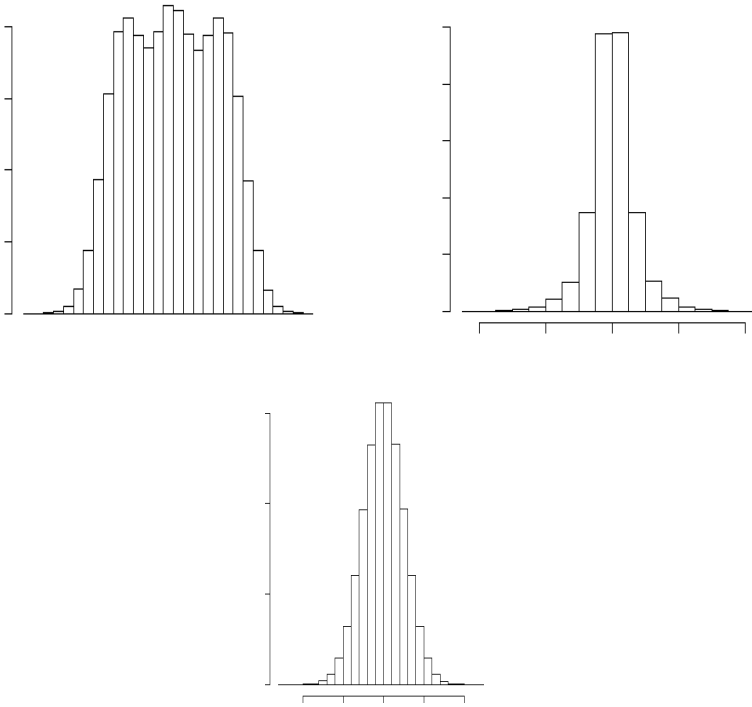


8.2. Principais medidas de curtose

Como informado, a curtose concerne à análise da forma vertical, e avalia se uma distribuição é pontiaguda, achatada ou se fica em uma posição intermediária. Visualmente, o histograma dá uma primeira ideia da curtose, conforme indicamos na Figura 8.2. Pelo que vemos, na primeira coluna, o histograma mostra uma variável achatada, ao passo que a terceira coluna mostra uma variável pontiaguda; já a variável do meio está em um nível de achatamento intermediário.

Formalmente, dizemos que no primeiro caso temos uma variável com distribuição platicúrtica, no segundo caso uma distribuição mesocúrtica, e no terceiro caso temos uma distribuição leptocúrtica. Em especial, quando temos uma distribuição mesocúrtica simétrica de uma variável ilimitada e contínua, dizemos que esta variável tem distribuição normal.

Figura 8.2 – Ilustração da ideia de curtose



- Coeficiente momento de curtose

Esta medida é a equivalente ao momento de assimétrica, e se baseia na média e no desvio padrão. Seja então uma variável $X = \{x_1, x_2, \dots, x_N\}$, tal que $Me(X)$ e $Dp(X)$ são, respectivamente, a média e o desvio padrão de X . Então a curtose, representada por $Kurt(X)$, será dada por:

$$Kurt(X) = \frac{\sum_{i=1}^N \frac{(x_i - Me(X))^4}{N}}{Dp(X)^4} = \frac{1}{N} \sum_{i=1}^M \left[\frac{(x_i - Me(X))}{Dp(X)} \right]^4$$

Os valores de referência para interpretação dessa medida são as seguintes:

- Se $Kurt = 3$, a variável é dada por mesocúrtica;
- Se $Kurt < 3$, a distribuição é platicúrtica, e se $Kurt > 3$ a distribuição é leptocúrtica.

Assim como o segundo coeficiente momento de assimetria, esta medida é sempre implementável no pacote R, da seguinte forma: dado um vetor 'x', a curtose 'kurt' vem de `kurt=(sum((x-mean(x))^4)/length(x))/sd(x)^4;kurt`.

No JASP, em *descriptives statistics*, há a opção de *kurtosis* para verificar a assimetria. Em Python, `import numpy as np; def kurt(x): return np.sum((x-np.mean(x))**4)/len(x)/(np.std(x)**4); print(kurt(x))`.

Também aqui temos o problema da não robustez das duas medidas envolvidas (desvio padrão e média), de modo que esta medida de curtose é não robusta. A medida a seguir resolve este problema.

- Coeficiente centílico (quantílico) de curtose

Uma medida robusta pode ser derivada de alguns quantis. Seja então uma variável $X = \{x_1, x_2, \dots, x_N\}$, tal que $Q_{(p)}(X)$ é o quantil de ordem p ($0 < p < 1$) de X . Então a curtose, representada por $Kurt(X)$, vem de:

$$Kurt(X) = \frac{Q_{(0,75)}(X) - Q_{(0,25)}(X)}{2(Q_{(0,90)}(X) - Q_{(0,10)}(X))}$$

Os valores de referência para interpretação dessa medida são os seguintes:

- Se $Kurt = 0,263$, a variável é dada por mesocúrtica;
- Se $Kurt < 0,263$, a distribuição é platicúrtica, e se $Kurt > 0,263$ a distribuição é leptocúrtica.

A implementação dessa medida no R (ou no JASP por meio do R console) é direta: dado um vetor 'x', a curtose 'kurt' vem de `kurt=(quantile(x, 0.75)-quantile(x, 0.25))/(2*(quantile(x, 0.90)-quantile(x,0.10)))`; `kurt`. Em Python, podemos implementar da seguinte maneira: `import numpy as np; def kurt(x): q75, q25, q90, q10 = np.percentile(x, [75, 25,90,10]); kurt = (q75-q25)/(2*(q90-q10)); return kurt; print(kurt(x))`.

Veamos agora um exemplo, em continuidade ao que foi mostrado acima para a assimetria no R (fica a realização em Python como exercício).

Exemplo: ainda referente aos dados de vagas de estacionamento, decidimos verificar a curtose, segundo as duas medidas expostas.

```
> x=c(28, 28, 24, 24, 41, 31, 44, 53, 41, 36, 33, 44, 28, 45, 3)
>
> #Coeficiente momento de curtose
> kurt=(sum((x-mean(x))^4)/length(x))/sd(x)^4;kurt
[1] 3.325342
>
> #Coeficiente centílico (quantílico) de curtose
> kurt=(quantile(x, 0.75)-quantile(x, 0.25))/
+(2*(quantile(x, 0.90)-quantile(x,0.10)));kurt
75%
0.3519417
```

Pelos resultados, a evidência é de que a distribuição é um pouco pontiaguda, com as medidas sinalizando uma distribuição leptocúrtica. A verificação do histograma permite uma reafirmação dessa sinalização. No entanto, a distância entre os valores observados e os referenciais de comparação é suficientemente pequena para se supor que a distribuição é mesocúrtica.

Considerando os resultados de assimetria e de curtose, temos uma impressão de que a distribuição das vagas livres no estacionamento se aproxima de uma distribuição normal (veja bem: temos uma aproximação, pois como a variável é discreta e limitada, não há como afirmar que a distribuição seja normal).

8.4. Resumo

Neste capítulo foram apresentados os conceitos e as principais medidas descritivas de formato de uma variável. Os destaques foram os seguintes:

- Há duas categorias de medidas de formato: as de formato horizontal e as de formato vertical;
- As medidas de formato horizontal têm como referência a posição simétrica da variável em relação às medidas de posição central. As medidas são as de assimetria;
- Temos medidas de assimetria baseadas em média e desvio padrão (assimetrias de Pearson e coeficiente momento de assimetria), que não são,

portanto, robustas à presença de *outliers*. Temos ainda medidas robustas e baseadas somente nos quantis (assimetrias de Bowley e de Kelley);

- As medidas de formato vertical têm como referência a posição intermediária (mesocúrtica) frente a uma distribuição achatada (platicúrtica) ou pontiaguda (leptocúrtica);
- Diferente da assimetria, que possui várias medidas, para a curtose apresentamos duas medidas apenas, uma baseada na média (coeficiente momento de curtose) e outra em quantis (coeficiente centílico de curtose);

Exercícios

1. Pesquise para implementação no R, JASP e em Python os procedimentos de extração de assimetria e curtose, e em seguida implemente com os dados do exemplo dado no capítulo.
2. Considere as variáveis quantitativas da planilha do Apêndice D (idade, renda, as quatro variáveis de valor (V1 a V4), além das variáveis identidade e prestígio percebido). Em cada uma delas, extraia todas as medidas de assimetria e curtose (não precisa excluir os outliers) e comente os resultados
3. Um pesquisador acredita que as alturas de estudantes brasileiros se aproximam de um padrão de normalidade. Para avaliar esta crença, um estudante curioso levantou as alturas de 20 de seus colegas de classe, e registrou os seguintes valores (exponho no formato de um vetor do R, em centímetros):

`h=c(150, 161, 178, 153, 178, 172, 178, 174, 164, 149, 159, 166, 163, 159, 142, 179, 201, 167, 160, 290)`

- a) Calcule as médias aritmética, aparada e winsorizada (10 e 20%) e apresente uma estimativa da altura na população, comparando os resultados dos três tipos de média
- b) Tabule os dados em 5 classes, e calcule a média pelos dados tabulados.
- c) Calcule as seguintes medidas de dispersão: intervalo interquartil, desvio médio, desvio padrão, desvio padrão truncado e winsorizado (10%), e desvio absoluto mediano.
- d) Calcule todas as medidas de assimetria e curtose.

- e) Uma pesquisa afirmava que a altura média do estudante era de 165cm, com desvio aproximado de 10cm, mas com distribuição fortemente assimétrica à direita. Considerando que os valores registrados podem ter problema com outliers, avalie esta afirmação e comente os resultados.
4. Faça uma pesquisa com uma amostra das alturas (mínimo 15) de seus colegas de classe e em seguida faça no pacote R e em Python, todos os procedimentos da questão 3.

PARTE III.

ANÁLISE BIVARIADA_

Chegamos à terceira e última parte deste manuscrito, abordando agora a construção das análises conjuntas de duas variáveis. Conforme já indicado nos capítulos anteriores, já temos técnicas para análises univariadas de variáveis tanto qualitativas quanto quantitativas. Uma etapa subsequente às análises univariadas é a análise bivariada, e posteriormente a análise multivariada.

Considerando então as duas alternativas de variáveis (categóricas e quantitativas), podemos já inferir que as análises bivariadas devem levar em conta ao menos três diferentes configurações. Para termos uma organização de capítulos que corresponda ao volume de conteúdos teóricos sobre o assunto, optei por apresentar primeiro as análises que envolvem variáveis categóricas e em seguida as análises envolvendo somente variáveis quantitativas.

Na primeira parte, materializada no capítulo 9, temos então dois blocos centrais, o primeiro envolvendo análises de duas variáveis categóricas, e o segundo envolvendo análises com variáveis categóricas e quantitativas. No primeiro bloco, optamos por uma análise inicialmente concentrada na construção das tabelas cruzadas (ou de contingência), para em seguida analisarmos os resultados por meio de algumas medidas de associação. Já no segundo bloco, seguimos um procedimento distinto, com a indicação da alternativa de avaliação conjunta baseada em medidas comparativas da variável quantitativa nas diferentes categorias da variável categórica, e depois foi feita a exposição das alternativas gráficas.

O último capítulo foi concentrado somente na análise de duas variáveis quantitativas. Em geral, as técnicas para análises desse tipo são abundantes, e tive que selecionar as principais para avaliação exploratória de um conjunto de dados. Nesse sentido, iniciamos então pela análise comparativa de medidas descritivas, para em seguida apresentar as análises gráficas, que se baseiam principalmente no gráfico de dispersão. Na sequência, apresento as opções de análise baseadas em medidas de associação, com a indicação de várias medidas possíveis, e finalizo com a exposição dos conteúdos de análise de regressão

simples, que é útil quando a associação entre as medidas é oriunda de alguma forma de influência ou previsão de uma variável sobre a outra.

Relativo à exposição dos procedimentos computacionais, nesta terceira parte o leitor vai observar que muitas verificações foram deixadas como exercício. A razão dessa decisão é bem fundamentada: está pressuposto que o leitor, ao chegar nesta parte do livro, já está bem ambientado nos pacotes estatísticos e já deve ter entendido que os caminhos a seguir podem ser consultados e construídos a critério do pesquisador. A partir do amadurecimento do usuário, outros caminhos são possíveis, e a experiência de seguir estes caminhos de forma mais autônoma já é uma competência a ser construída. Em grande medida, os desafios colocados impõem o exercício da autonomia do usuário e do pesquisador.

CAPÍTULO 9

ANÁLISE BIVARIADA ENVOLVENDO VARIÁVEIS CATEGÓRICAS_

Até o capítulo 8, tivemos a possibilidade de conhecer e desenvolver avaliações de procedimentos e medidas de descrição de uma variável. Mas na grande maioria das análises e aplicações, as avaliações univariadas são limitadas em sua utilidade, mesmo que possamos ter trabalhos inteiros muito bem construídos somente com este tipo de análise. Considerando, por outro lado, a perspectiva da análise exploratória de dados, um estudo baseado em uma variável pode se complementar com um ganho bastante positivo quando aplicamos procedimentos complementares, analisando conjuntamente duas ou mais variáveis.

Neste capítulo, iniciamos a exposição dos conteúdos convencionais da análise conjunta de duas variáveis (como informado anteriormente, é possível que uma análise multivariada se aplique à finalidade de análise exploratória; porém, optamos por não fazer exposições desse tipo aqui, por conta, principalmente, da demanda de conhecimentos especializados que estão além do escopo deste livro).

Como indicado na introdução da terceira parte, em geral, tais análises são definidas segundo o tipo de variáveis envolvidas. Temos, então, as análises de duas variáveis categóricas, aquelas com uma variável quantitativa e outra categórica, e as análise envolvendo duas variáveis quantitativas. Neste capítulo, abordamos as duas primeiras avaliações, e, no capítulo 10, apresentamos a terceira forma e análise. Veremos que, em geral, muitas das medidas anteriormente apresentadas são aqui resgatadas. Adicionalmente, há vários procedimentos gráficos que podem ser aplicados à análise bivariada, que apresentaremos quando forem oportunos.

Ao final do capítulo o estudante deve ser capaz de responder às seguintes questões:

- Como analisar duas variáveis categóricas?

- Como analisar conjuntamente uma variável quantitativa e uma categórica?
- De que forma essas análises podem contribuir para a análise exploratória de dados?

9.1. Análise de duas variáveis categóricas

Quando consideramos duas variáveis categóricas que reflitam informações e variações dos mesmos sujeitos (por exemplo, para cada pessoa conferimos o estado civil e a religião), o procedimento mais comum consiste em efetuar um cruzamento das categorias, com a contagem das quantidades em uma tabela de dupla entrada. O procedimento consiste na contagem de vezes em que duas categorias aparecem conjuntamente no conjunto de elementos avaliados.

Tomemos o exemplo dos dados em Apêndice D, nos quais verificamos diversas variáveis categóricas. Suponhamos que nosso interesse seja avaliar como os gêneros se distribuem em geral e pelos dois cursos. O procedimento consiste, portanto, em tomar o curso de Administração e fazer a contagem de homens e mulheres, e em seguida repetir o mesmo procedimento para o curso de Turismo (obviamente, poderíamos tomar inicialmente algum dos gêneros e fazer a contagem nos cursos de Administração e de Turismo). Teremos então um resultado como indicado na Tabela 9.1.

Tabela 9.1 – Cruzamento de curso e gênero

Curso	Gênero		Total
	Feminino	Masculino	
Administração	28	29	57
Turismo	40	13	53
Total	68	42	110

Por este resultado temos então uma visualização do que configura a análise como sendo bivariada: aqui analisamos conjuntamente as duas variáveis, e temos pela tabela uma informação que remete, em termos de análise, às duas variáveis. O leitor atento poderá já ter percebido que esta análise poderia englobar ainda mais variáveis. Isto certamente é possível, porém pode fazer crescer o número detalhes e se perder a referência de avaliação agregada dos dados. Cada pesquisa, por seus propósitos, indicará se tal incremento de variáveis na análise pode ser interessante ou não.

– Dados complementares em percentuais

Em cada célula da tabela cruzada temos um quantitativo que informa o quanto de uma variável se observa na outra, em termos absolutos. A análise pode ser enriquecida com valores relativos, na forma de percentuais. Neste caso, temos então três tipos de percentuais a serem extraídos. Tomemos por exemplo, ainda na Tabela 9.1, a célula que indica o número de mulheres no curso de Administração, que totaliza 28. Para este valor, temos três possíveis referências de porcentagem:

- Temos a indicação do percentual em relação ao gênero, que, neste caso, sinaliza 28 mulheres em um total de 57 pessoas do curso de Administração, o que corresponde a 49,1%;
- Temos ainda a indicação de percentual em relação ao curso, ou seja, temos 28 mulheres no curso de Administração, de um total de 68 mulheres nos dois cursos, o que corresponde a 41,2%;
- E, por fim, temos o quanto este total de 28 mulheres representa do total de 110 pessoas pesquisadas, o que corresponde a um total de 25,5%.

Por este resultado, e replicando para as demais células da tabela, temos para cada célula quatro informações. Os resultados ficam conforme indicado na Tabela 9.2. Ganhamos em conhecimento, mas uma visualização permite ver o risco de termos um excesso de informação, o que pode mais atrapalhar o entendimento do que contribuir para uma melhor análise dos dados.

Por esta razão, é usual o analista de dados indicar apenas uma das referências de análise, a critério de seu interesse e dos objetivos definidos para cada estudo. Certamente, uma das alternativas mais utilizadas e com maior valor para entendimento do conjunto de dados é a exposição dos resultados em relação ao total. Neste caso, as células referentes aos totais terão sempre somatório de 100%, e indicam o que chamamos de ‘distribuições marginais’ das variáveis. Por exemplo, a distribuição marginal da variável curso está na última coluna, com a indicação de uma distribuição de respostas das 110 pessoas em 51,8% no curso de Administração e de 48,2% no curso de Turismo.

Tabela 9.2 – Detalhamento em percentuais

Curso	Referência	Gênero		Total
		Feminino	Masculino	
Administração	Contagem	28	29	57
	% dentro de gênero	49,1%	50,9%	100,0%
	% dentro de curso	41,2%	69,0%	51,8%
	% do total	25,5%	26,4%	51,8%
Turismo	Contagem	40	13	53
	% dentro de gênero	75,5%	24,5%	100,0%
	% dentro de curso	58,8%	31,0%	48,2%
	% do total	36,4%	11,8%	48,2%
Total	Contagem	68	42	110
	% dentro de gênero	61,8%	38,2%	100,0%
	% dentro de curso	100,0%	100,0%	100,0%
	% do total	61,8%	38,2%	100,0%

- Sinalização de associação

Além da análise percentual, uma relevante informação que uma tabela cruzada nos dá concerne à indicação ou à sinalização de associação entre duas variáveis. Cabe uma rápida digressão sobre o conceito de associação e a questão da influência ou dependência.

Dizemos que duas variáveis estão associadas quando uma delas possui sua variação em paralelo com a outra. Por exemplo, tomemos as variáveis gênero (que varia entre masculino e feminino) e a variável ‘acompanhar resultados de futebol’ (que pode ser sim ou não). Neste caso, sabemos que a frequência (percentual) de sim para acompanhar resultados de jogos é maior nos homens que nas mulheres. Este resultado decorre de uma associação entre as duas variáveis. Veja bem: a associação não significa necessariamente influência ou predição de uma pela outra. Na verdade, a influência é uma realidade ou expectativa que explica a associação, mas o fato de haver associação não implica em haver influência. A influência normalmente é justificada por construções teóricas que a justificam, e os dados a refletem.

De posse dessas considerações (no capítulo seguinte retomo a questão), farei inicialmente uma visualização da análise, e posteriormente indicarei medi-

das de associação. Vejamos então um raciocínio ainda com base na Tabela 9.2, para em seguida apresentarmos o referencial de interpretação. Consideremos então os estudantes do curso de Administração distribuídos entre os gêneros feminino e masculino.

Pelos resultados marginais, temos a constatação de que a quantidade de alunos de Administração representa 51,8% do total de alunos pesquisados. Se supomos que não há associação entre curso e gênero, devemos ter um percentual de mulheres no curso de Administração também de 51,8% em relação ao total de mulheres pesquisadas; para os homens também seriam necessários 51,8% no curso de administração. Computando em números, seriam esperados, pelos percentuais dos totais, em torno de 35 mulheres em Administração (51,8% do total de 68 mulheres) e 22 homens (51,8% do total de 42 homens)

No entanto, não é isto que observamos. De fato, esperávamos um percentual de 51,8% de mulheres, mas observamos na verdade 41,2%; já nos homens observamos 69,0% em lugar dos 51,8% esperados. Em valores absolutos, se não houvesse associação, esperaríamos 35 mulheres, mas encontramos 28, e esperaríamos 22 homens, mas encontramos 29. A conclusão é que há mais homens que o esperado e há menos mulheres que o esperado.

Dito de outra forma, a quantidade de mulheres é desproporcionalmente menor que o total esperado caso as variáveis não tivessem associação, ao passo que nos homens observamos uma variação inversa, com um total de homens desproporcionalmente maior que o esperado. Em geral, a sinalização que temos é que o curso de Administração atrai mais homens que mulheres (relativamente aos totais de homens e mulheres pesquisados).

O procedimento realizado foi restrito a um exemplo, e precisamos fazer uma generalização para termos um referencial de identificação da quantidade esperada para o caso de independência entre as duas variáveis. Façamos então este procedimento para a tabela de uma variável 1 com l categorias cruzadas com uma variável 2 de k categorias. Na Tabela 9.3, temos uma indicação de contagens genéricas de quantidades, em que indicamos as quantidades O_{ij} e E_{ij} como sendo os valores observado e esperado da categoria i ($i=1, 2, \dots, k$) da variável 1 e na categoria j da variável 2 ($j=1, 2, \dots, l$). Observe que tomamos o somatório em relação ao valor observado porque é o que temos de informação empírica, porém é fácil inferir que o somatório dos valores observados é igual ao somatório dos valores esperados, ou seja, $\sum_{i=1}^l O_{1i} = \sum_{i=1}^l E_{1i}$.

Tabela 9.3 – Contagens genéricas

Variável 1	Variável 2				Total
	Categoria 1	Categoria 2	...	Categoria k	
Categoria 1	O_{11} E_{11}	O_{12} E_{12}	...	O_{1k} E_{1k}	$\sum_{j=1}^k O_{1j}$
Categoria 2	O_{21} E_{21}	O_{22} E_{22}	...	O_{2k} E_{2k}	$\sum_{j=1}^k O_{2j}$
...	...	...	...	...	...
Categoria l	O_{l1} E_{l1}	O_{l2} E_{l2}	...	O_{lk} E_{lk}	$\sum_{j=1}^k O_{lj}$
Total	$\sum_{i=1}^l O_{i1}$	$\sum_{i=1}^l O_{i2}$	...	$\sum_{i=1}^l O_{ik}$	$N = \sum_{i=1}^l \sum_{j=1}^k O_{ij}$

A lógica da argumentação de independência é a seguinte: se as duas variáveis não tiverem associação, a razão do valor esperado nas categorias i e j das variáveis 1 e 2 pelo total da categoria j da variável 2 (ou seja, o percentual da categoria j dentro da variável 2) deverá ser igual à razão do valor total da categoria i da variável 1 em relação ao total geral (ou seja, o percentual total da categoria i da variável 1 em relação ao total geral). Na linguagem da regra de três, diremos então que: E_{ij} está para $\sum_{i=1}^l O_{ij}$ assim como $\sum_{j=1}^k O_{ij}$ está para $\sum_{i=1}^l \sum_{j=1}^k O_{ij}$, de modo que,

$$\frac{E_{ij}}{\sum_{i=1}^l O_{ij}} = \frac{\sum_{j=1}^k O_{ij}}{\sum_{i=1}^l \sum_{j=1}^k O_{ij}} \Rightarrow E_{ij} = \frac{\left(\sum_{j=1}^k O_{ij}\right) \left(\sum_{i=1}^l O_{ij}\right)}{\sum_{i=1}^l \sum_{j=1}^k O_{ij}}$$

Por este procedimento, temos como calcular o valor esperado de uma determinada célula cruzada simplesmente multiplicando as respectivas quantidades marginais e dividindo pelo total. O valor gerado provavelmente vai se diferenciar do valor observado, sendo possível, inclusive, que surjam valores decimais. A regra preliminar de avaliação, para o momento, é a seguinte: se os valores esperados não são muito discrepantes do valor observado, dizemos que há sinais de ‘não associação’; se as distâncias são grandes, daí concluímos que as duas variáveis possuem associação.

Dado que estamos procedendo à análise exploratória, podemos fazer uma análise apenas pela argumentação em relação aos valores observados e esperado. Em estatística inferencial, por outro lado, não há muito sentido em falar em algo como ‘muito discrepante’ sem termos uma referência quantitativa para informar e algo é ou não muito discrepante. Pelo escopo deste manuscrito não temos como avançar para fazer compreender o que seria esta referência quantitativa, mas posteriormente apresentarei medidas que ‘sinalizam’ se as diferenças são grandes ou pequenas.

Vejamos então o nosso exemplo (ver Tabela 9.4).

- Para a célula de estudantes de Administração do gênero feminino, teremos um total esperado de 68 (total de mulheres) multiplicado por 57 (total de Administração), dividido por 110, o que resulta em um total esperado (supondo independência) de 35,2 pessoas (não nos preocupamos, agora, com o sentido do decimal);
- Já para a célula de estudantes de Administração do gênero masculino, teremos um total esperado de 42 (total de homens) multiplicado por 57 (total de Administração), dividido por 110, o que resulta em um total esperado de 21,8;
- Para a célula de estudantes de Turismo do gênero feminino, teremos um total esperado de 68 (total de mulheres) multiplicado por 53 (total de Turismo), dividido por 110, o que resulta em um total esperado de 32,8;
- Por fim, para a célula de estudantes de Turismo do gênero masculino, teremos um total esperado de 42 (total de homens) multiplicado por 53 (total de Turismo), dividido por 110, o que resulta em um total esperado de 20,2.

Tabela 9.4 – Tabela completa

Curso	Referência	Gênero		Total
		Feminino	Masculino	
Administração	Observado	28	29	57
	Esperado	35,2	21,8	
Turismo	Observado	40	13	53
	Esperado	32,8	20,2	
Total	-	68	42	110

O resultado está na Tabela 9.4, e a indicação é de que há sinalizações de associação entre as variáveis, tendo em vista que há uma razoável discrepância entre todos os valores esperados e observados. Pelo que parece, as mulheres se concentram mais no curso de Turismo e menos no curso de Administração, o inverso ocorrendo com os homens.

Em relação à extração de tabelas cruzadas e de informações complementares nos pacotes estatísticos, no pacote R, o procedimento para o cruzamento é simples, e consiste apenas em dar o comando `table()` e informar quais são as variáveis a serem cruzadas. Já a extração de demais valores é mais complicada por envolver procedimentos complementares. Por exemplo, a extração dos valores esperados pode ser feita da seguinte forma: `esperado=chisq.test()$exp; esperado`, o que é, obviamente, muito mais complicado que um comando direto. A extração de percentuais dentro de cada célula fica como exercício.

No JASP, deve-se ir em *frequency* → *contingency tables* → arrastar as variáveis de interesse. A tabela de contingência será apresentada. Em Python, podemos utilizar a biblioteca *pandas* e a função `tabela = pd.crosstab(df['A'], df['B'])`; `print(tabela)` para cruzamentos de variáveis. A extração dos valores esperados, pode ser realizada da seguinte forma: `from scipy.stats import chi2_contingency; _, _, _, esperado = chi2_contingency(tabela); print(esperado)`.

Por fim, no Excel, o procedimento mais fácil consiste em utilizar as chamadas tabelas dinâmicas. O procedimento consiste em selecionar os dados (incluindo os cabeçalhos) e em seguida, na aba 'inserir', marcar o comando 'Tabela dinâmica'. Será aberta então uma nova planilha para a indicação de quais são as linhas, as colunas e as contagens de referência. A depender do interesse, a tabela gerada deverá ser copiada para outra planilha para edições complementares e cálculos de medidas. O manuseio ficará como exercício.

– Medidas de associação

Pelo tratamento dado logo acima, indicamos o procedimento exploratório de avaliação de associação, mas sem indicar qualquer medida geral de sinalização dessa associação. Apresento aqui algumas alternativas de verificação exploratória, realçando que, em análises inferenciais e baseadas em testes, temos diversas alternativas consistentes de verificação em que estas aqui observadas acabam perdendo seu uso.

Quando verificamos a derivação das medidas de valor esperado de uma célula, e quando confrontamos o resultado com o que foi observado, nossa análise torna-se uma avaliação célula a célula. É possível, por outro lado, definirmos uma medida agregada, que capte as variações por todas as células. A medida mais comumente extraída é a chamada ‘estatística de qui-quadrado’, representada por χ^2 . Vejamos sua extração.

Dada uma tabela de l linhas e k colunas, em cada célula ij ($i=1, \dots, l; j=1, \dots, k$), teremos um quantitativo observado O_{ij} e um valor esperado E_{ij} . A estatística será definida pela soma, no total de células, da divisão do quadrado da diferença entre o quantitativo observado e o esperado pelo valor esperado, ou seja,

$$\chi^2 = \sum_{i=1}^l \sum_{j=1}^k \frac{(O_{ij} - E_{ij})^2}{E_{ij}},$$

$$\text{ou ainda } \chi^2 = \sum_{t=11}^{kl} \frac{(O_t - E_t)^2}{E_t}, \quad t = ij \left(i = 1, \dots, l; j = 1, \dots, k \right)$$

No pacote R, a extração dessa medida pode ser executada utilizando a seguinte rotina para duas variáveis cruzadas a e b : `qui_qua=chisq.test(a, b)$statistic; qui_qua`. No JASP, seguimos o mesmo procedimento apresentado anteriormente para o acesso à tabela. Ou seja, deve-se ir em `frequency` → `contingency tables` → arrastar as variáveis de interesse. A tabela de contingência será apresentada, bem como os resultados do teste de qui-quadrado. Em Python, podemos utilizar: `from scipy.stats import chi2_contingency; qui_qua,_,_,_ = chi2_contingency(x); print(qui_qua)`. Fica como exercício a verificação de opções de extração no Excel.

É evidente que esta estatística por si não diz muita coisa, tendo em vista que seu valor pode crescer tanto em função de grandes discrepâncias entre os valores observado e esperado, quanto pelo número de elementos avaliados e pelo número de células. Teríamos algumas opções interessantes para verificação, como, por exemplo, tomar o qui-quadrado dividido pelo total de células (que se origina da multiplicação do número de linhas pelo número de colunas), ou então pelo total de observações.

Por ideias neste sentido é que foram desenvolvidas algumas medidas gerais, que guardam a intenção de servirem de referencial fixo e de fácil interpretação. Uma boa medida neste sentido seria uma em que tivéssemos uma escala de in-

tensidade de associação, desde uma associação inexistente até uma explicação total de uma variável pela outra. Os teóricos têm dado preferência a medidas assim a partir de uma escala de 0 a 1, em que 0 indica total inexistência de associação, ao passo que 1 indicaria total associação. As principais medidas já desenvolvidas foram 3, que estão abaixo indicadas.

a) Coeficiente de contingência de Pearson

A primeira referência de análise de associação é o coeficiente de contingência de Pearson, representado por CP , que está indicado na formulação abaixo:

$$CP = \sqrt{\frac{\chi^2}{N + \chi^2}} = \sqrt{\frac{\chi^2/N}{1 + \chi^2/N}}, \text{ em que } N \text{ é o total de observações.}$$

É evidente que, quando não houver associação, a diferença entre valores esperados e observados será 0, gerando um qui-quadrado também de 0, e um valor de CP nulo. O problema, por outro lado, é que o valor superior nunca será igual a 1, e o máximo que poderá atingir depende do número de células. Portanto, este coeficiente não é uma medida adequada, ou ao menos não deve ser apresentado de forma isolada como referencial de associação.

Do ponto de vista de extração das medidas, podemos fazer o cálculo a partir do R ou em Python. Deixamos como exercício a verificação dos procedimentos.

b) Coeficiente de Cramer

Uma alternativa de verificação de associação, ainda com base no Coeficiente de Cramer, representado por CC , que está indicado abaixo:

$$CC = \sqrt{\frac{\chi^2/N}{\min\{l - 1; k - 1\}}}, \text{ em que } l \text{ é o número de linhas e } k \text{ o de colunas.}$$

Também aqui, quando não houver associação o valor de CC será nulo. O problema, por outro lado, é que C pode ser maior que 1, embora fique em torno de 1 quando a associação for forte. No entanto, a falta de um limitante superior torna a medida deficiente. Uma variação desse coeficiente consiste em tomar no denominador da expressão o produto em lugar do mínimo. Este é o Coeficiente de Tschuprov, representado por CT e tem a seguinte formulação:

$$CT = \sqrt{\frac{\chi^2/N}{(l - 1)(k - 1)}}$$

Além dessas medidas, existem ainda outras, que deixo como exercício a pesquisa e a avaliação. Em geral, o interesse do pesquisador vai indicar qual a medida mais adequada, porém a recomendação é de que sejam extraídas tantas quantas possíveis que e que gerem conteúdo informativo para se conhecer melhor os dados e para fundamentar análise posteriores.

9.2. Análise de uma variável quantitativa e uma categórica

Na análise de uma variável quantitativa conjuntamente com uma variável categórica, temos, para efeito de análise exploratória, duas alternativas, que são a comparação de medidas e a análise gráfica. A seguir temos as duas opções detalhadas e exemplificadas.

- Comparação de medidas

A análise de comparação de medidas consiste em apresentar as medidas descritivas da variável quantitativa nos diferentes grupos da variável qualitativa. Não há uma definição de qual ou quais medidas comparar, e a seleção dentre as diversas medidas disponíveis depende somente do interesse do pesquisador.

Tomemos o exemplo das variáveis curso e idade, indicadas no Apêndice D. Veja que o curso é variável categórica com possibilidades de Administração e Turismo, e a variável idade é uma quantitativa discreta (Tabela 9.5). Para análise nos grupos, foram então tomadas as medidas de posição (média e quartis), dispersão (desvio padrão) e formato (assimetria e curtose de Pearson).

Tabela 9.5 – Comparação de idades por curso

Variáveis	Dados		Média	Desvio	Assimetria	Curtose	Quartis		
	Válido	Missing					1	2	3
Administração	59	1	31,37	10,26	1,87	5,30	24,30	30,43	35,94
Turismo	52	1	35,58	51,97	7,02	50,16	23,33	29,00	34,00

Os resultados estão indicados na Tabela 9.5, e vemos que a maioria dos respondentes foi de Administração (59; 52 foram de Turismo), e em cada curso tivemos um dado perdido. Pelo que vemos, a média de idades é maior no curso de Turismo, e também são maiores nesse curso as medidas de desvio padrão, assimetria e curtose. Os resultados dos quartis, por outro lado, revelam que, ao longo destas três posições, ocorre o contrário, com as medidas do curso de Turismo sendo menores (entre, aproximadamente, 1 e 2 anos) que as medidas em Administração.

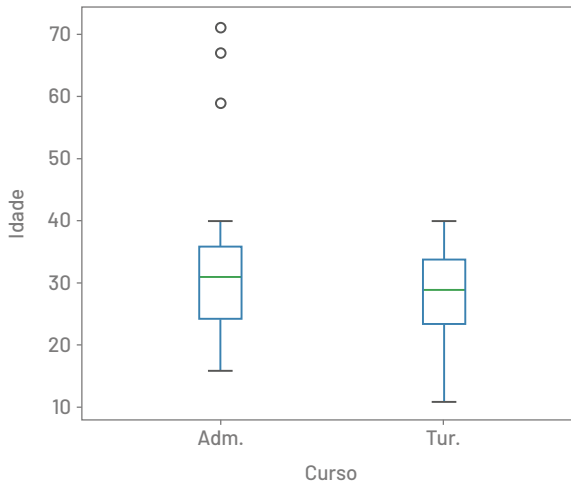
Este resultado sugere que, possivelmente, há algum problema nos dados pela presença de algum valor extremo. A verificação da planilha permite notar que, de fato, a observação 111 está associada a uma idade de 400 anos, o que certamente é resultado de um erro em que a idade correta seria, possivelmente, 40 anos. Como as medidas em que houve grande discrepância são baseadas na média, e como a média é afetada por *outliers*, é provável que este valor extremo esteja afetando a diferença de medidas. Fica como exercício, portanto, a extração de medidas robustas para medidas de dispersão e formato (ver exercícios).

- Análise gráfica

A análise por meio de gráficos segue a mesma lógica da análise por meio de medidas, ou seja, o procedimento consiste em selecionar os gráficos a serem aplicados na variável quantitativa e apresentá-los em cada categoria, construindo uma mesma figura que permite verificar o formato e as demais informações por grupo.

Para uma exemplificação, apresentamos o *boxplot* da variável idade em cada um dos dois cursos (Gráfico 9.1), com extração em Python (aqui já excluímos o valor extremo de 400 do curso de Turismo). Pelo que vemos, temos a sinalização de três valores extremos nas idades dos alunos de Administração (nas observações que são as pessoas com 59, 67 e 71 anos de idade), e em geral, as medidas de Administração são um pouco maiores que as medidas de Turismo.

Gráfico 9.1 – Comparação por gráficos



A extração de gráficos por grupos é facilitada especialmente no R, ao passo que no Excel o procedimento é mais trabalhoso (por exemplo, na versão última do software quando escrevo estas páginas, é necessário primeiramente extrair as medidas de referência do *boxplot* para em seguida se construir o gráfico).

Especificamente para o *boxplot*, no R, o procedimento básico é simples, e requer apenas que seja dada a referência da variável categórica junto com a variável quantitativa. Sendo, por exemplo, a a variável quantitativa e b a variável qualitativa, o comando é `boxplot(a~b)`. No JASP, segue-se o mesmo procedimento indicado anteriormente para análise univariada. Ou seja, deve-se ir em *descriptives* → *plot builder* → arrastar as variáveis de interesse → *distributions*(*histogram*, *boxplot*, *violin*) → *boxplot*. Em Python, o procedimento é feito utilizando a biblioteca *matplotlib*, sendo interessante, preliminarmente, separar as variáveis de interesse e, na sequência, gerar o gráfico, como segue.

```
import pandas as pd
import matplotlib.pyplot as plt

df = pd.read_excel('COLOQUE AQUI O CAMINHO DO SEU ARQUIVO COM
BARRAS INVERTIDAS')

df_interesse = df[df['Cur'].isin(['Adm.', 'Tur.'])]
df_interesse=df_interesse.dropna(subset=['Ida']) #remover ausentes

df_interesse.boxplot(column='Ida', by='Cur', grid=False)
plt.show()
```

9.3. Resumo

Neste capítulo, foram apresentados os procedimentos de avaliação exploratória de dados de duas variáveis, tomadas conjuntamente, quando uma das variáveis é de natureza categórica. Os destaques foram os seguintes:

- Quando avaliamos conjuntamente duas variáveis categóricas, o procedimento mais comum consiste em fazer o cruzamento dos dados, em uma tabela de duas entradas;
- A partir dos dados da tabela e das distribuições marginais, é possível avaliar se as duas variáveis possuem alguma associação ou não. Tal avaliação pode ser feita pelo cálculo de valores esperados e observados e por sua comparação direta;
- No entanto, o procedimento de avaliação direta pode ser complementado a partir de medidas globais de associação, a partir dos vários coeficientes de contingência existentes;
- Quando temos uma avaliação conjunta de uma variável categórica com uma variável quantitativa, o procedimento mais recorrente é avaliar o comportamento das medidas descritivas da variável quantitativa nas diferentes categorias da variável categórica;
- Para este tipo de avaliação, temos ainda a possibilidade de comparação de gráficos da variável quantitativa nas diferentes categorias. Temos várias alternativas de análise gráfica, e mostramos especificamente a comparação dos *boxplots*.

Exercícios

1. Aplique as três fórmulas de coeficiente de contingência nos dados do exemplo 9.1 (relativo às variáveis ‘curso’ e ‘gênero’), e avalie a seguinte hipótese: ‘o curso de Turismo concentra mais mulheres, em relação aos homens, que o curso de Administração’.
2. Faça uma sequência dos procedimentos do pacote R, JASP e em Python para as análises expostas no capítulo para duas variáveis categóricas. Recomendação: procure métodos de extração das medidas a partir do comando `chisq.test`.

3. Faça uma sequência de procedimentos do pacote R e em Python para análise de variáveis categórica e quantitativa.
4. Considerando os resultados da tabela 9.5, extraia as mesmas medidas com a correção do outliers (de 400 para 40 ou excluindo o valor). Em seguida, desenvolva uma tabela semelhante com medidas de dispersão e formato que sejam robustas à presença de outliers. Compare as três tabelas e comente os resultados.

CAPÍTULO 10

ANÁLISE DE DUAS VARIÁVEIS QUANTITATIVAS_

Neste capítulo, finalizamos a exposição dos conteúdos de análises exploratória bivariada, com os conteúdos associados a duas variáveis quantitativas. Há uma grande variedade de técnicas estatísticas direcionadas para análises de variáveis quantitativas, tanto contínuas quanto discretas e vamos expor aqui somente uma parte das opções disponíveis, mas que já são suficientes para uma primeira avaliação dos dados e para construção de conhecimento sobre a realidade analisada.

Temos também aqui três alternativas consistentes de avaliação, uma baseada em medidas descritivas, uma gráfica e outra baseada em medidas de associação. Ao final deste capítulo, o leitor deve ser capaz de responder às seguintes questões:

- Quais técnicas podem ser aplicadas para analisar conjuntamente duas variáveis quantitativas?
- Qual a diferença entre correlação e regressão e como se procedem suas análises?
- De que forma essas análises podem contribuir para a análise exploratória de dados?

10.1. Análise simultânea de medidas

A análise simultânea de medidas descritivas é o procedimento mais simples, e consiste, basicamente, em tomar as medidas de interesse em duas variáveis e posicioná-las conjuntamente de modo a permitir uma comparação. Para este procedimento não há especificação de quais medidas a serem tomadas, ficando a critério do pesquisador definir se seu interesse é comparar medidas de posição, de dispersão ou de forma.

Como exemplo, temos na Tabela 10.1 a extração de medidas de média aritmética simples, desvio padrão, assimetria, curtose e quartis para as variáveis identificação e reputação percebida, dos dados em anexo. Temos na variável reputação medidas de posição sempre menores que as correspondentes na variável identificação. Também verificamos que o desvio padrão de reputação é maior, e que esta variável está centrada horizontalmente (próxima da simetria), e tem um achatamento em seu formato vertical. Já a variável identificação, além de possuir um maior nível de concentração de respostas em torno da média (como mostra o desvio padrão), possui uma assimetria leve à esquerda, e tem formato vertical bastante pontiagudo. As medidas dos quartis mostram que os escores atribuídos à identificação são sistematicamente mais elevados que aqueles atribuídos a reputação.

Tabela 10.1 – Comparação de medidas

Variáveis	Dados		Média	Desvio	Assimetria	Curtose	Quartis		
	Válido	Missing					1	2	3
Reputação	110	3	3,65	2,05	0,17	-1,35	2,00	3,00	5,25
Identificação	112	1	5,02	1,01	-0,99	2,35	5,00	5,00	6,00

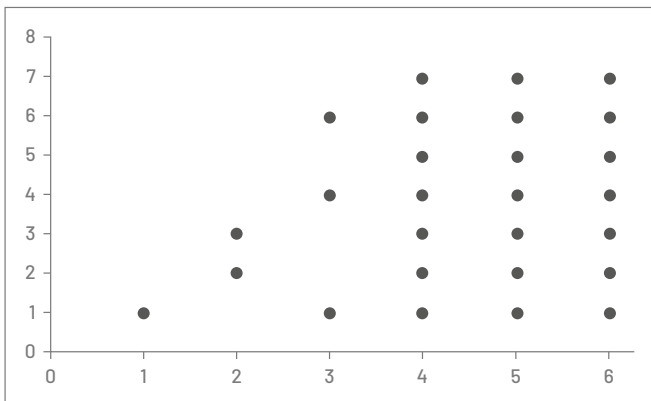
Esta primeira análise tem sentido bivariado porque a discussão é comparativa, mas não está centrada na avaliação simultânea das duas variáveis. Tal alternativa pode ser feita a partir de gráficos de exibição conjunta de ambas as variáveis ou por meio de medidas de associação.

10.2. Análise por meio gráfico

Na avaliação de duas variáveis quantitativas pareadas por meio gráfico, temos como principal forma alternativa a extração do chamado gráfico de dispersão, ou o *scatterplot*. A lógica é simples: tomamos duas variáveis pareadas, digamos $X = \{x_1, x_2, \dots, x_N\}$ e $Y = \{y_1, y_2, \dots, y_N\}$, e para cada unidade de resposta definimos pares ordenados $(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)$. Assim, o gráfico de dispersão consiste em plotar em um eixo de coordenadas cartesianas esses pontos, formando uma nuvem de pontos que dá uma sinalização da forma como as duas variáveis estão relacionadas.

Tomando como exemplo os dados do Apêndice D, foi feito no Excel gráfico de dispersão (Gráfico 9.2) das variáveis identidade (eixo das abscissas) e reputação percebida (eixo das ordenadas). Pelo exposto, é possível observar que temos uma tendência de crescimento conjunto, ou seja, temos uma sinalização de que quando o nível de identidade com a profissão cresce, o nível de percepção de reputação também cresce. Mas ao que parece, a sinalização de variação conjunta é fraca.

Gráfico 9.2 – Gráfico de dispersão



A extração do gráfico de dispersão é possível em todos os pacotes de referência, com os seguintes procedimentos:

- No Excel, o procedimento consiste em selecionar as duas variáveis e ir na aba 'inserir' e indicar a opção 'Dispersão', na qual poderão ser selecionados gráficos somente com a indicação de pontos (como o gráfico acima) ou com ligações entre os pontos;
- No pacote R, basta tomar as duas variáveis (digamos, a e b) e colocar entre parênteses no comando `plot(a, b)`. A primeira variável ficará no eixo das abscissas e a segunda no eixo das ordenadas. Há diversas formas de edição do gráfico, mediante comandos que podem ser facilmente verificados na internet;
- No JASP, o caminho é `descriptives` → `plot builder` → arrasta as variáveis de interesse → `individual data points`.
- Em Python, podemos utilizar a biblioteca `matplotlib` e a função `plt.scatter(a,b)`. A primeira variável ficará no eixo das abscissas e a segunda no eixo das ordenadas. Há diversas formas de edição do gráfico, mediante comandos que podem ser facilmente verificados na internet.

De todos os softwares, o Excel é certamente o programa com melhores funcionalidades para obtenção de gráficos, pois são disponibilizadas diversas opções de edição de eixos, formatos dos pontos, fontes de letras etc., além de o gráfico ser sincronizado com os dados, de modo que alterações e correções nos dados já repercutem ao mesmo tempo no gráfico.

10.3. Análise por medidas

Com relação às medidas de avaliação conjunta, temos diversas opções, das quais apresentamos aqui as quatro principais, que são a covariância, a correlação produto-momento de Pearson, a correlação por postos de Spearman e a correlação winsorizada.

a) Covariância

Dadas duas variáveis finitas ou duas amostras pareadas $X = \{x_1, x_2, \dots, x_N\}$ e $Y = \{y_1, y_2, \dots, y_N\}$, em que $Me(X)$ e $Me(Y)$ representam, ordenadamente, as médias aritméticas simples de X e de Y . Definimos por covariância das duas variáveis, representada por $Cov(X, Y)$, a seguinte expressão:

$$Cov(X, Y) = \frac{1}{N} \sum_{i=1}^N (x_i - Me(X))(y_i - Me(Y))$$

A medida de covariância pode ser extraída de qualquer par de variáveis quantitativas, e pode assumir qualquer valor no conjunto dos números reais. Para sua interpretação, temos o seguinte: se a covariância é nula, isto indica que as duas variáveis não possuem qualquer variação conjunta, ou seja, não possuem associação. Caso seja diferente de zero, se a covariância for positiva, a indicação é de que as duas variáveis variam conjuntamente de forma positiva, de modo que quando uma aumenta a outra tende a aumentar junto, e quando uma diminui a outra tende a seguir o mesmo comportamento; para o caso de covariância negativa, a indicação é de variação inversa, ou seja, quando uma variável aumenta (diminui) a outra diminui (aumenta) e vice versa.

A interpretação da covariância é dificultada justamente por não haver um limite de referência superior ou inferior. Isto torna a medida pouco usada como referência para análise exploratória, embora seja uma medida das mais relevantes em análises multivariadas. Por esta razão, foram então definidas

medidas adicionais, baseadas na covariância, mas com limites superiores e inferiores. São as medidas de correlação, descritas a seguir.

b) Correlação produto-momento de Pearson

Dadas duas variáveis finitas ou duas amostras pareadas $X = \{x_1, x_2, \dots, x_N\}$ e $Y = \{y_1, y_2, \dots, y_N\}$, em que $Me(X)$, $Me(Y)$, $Dp(X)$ e $Dp(Y)$ representam, ordenadamente, as médias aritméticas simples de X e de Y e seus respectivos desvios padrão. Definimos por correlação produto-momento de Pearson das duas variáveis, representada por $Cor(X, Y)$, a seguinte expressão:

$$Cor(X, Y) = \frac{\sum_{i=1}^N (x_i - Me(X))(y_i - Me(Y))}{\sqrt{\sum_{i=1}^N (x_i - Me(X))^2 \sum_{i=1}^N (y_i - Me(Y))^2}} = \frac{Cov(X, Y)}{Dp(X)Dp(Y)}$$

Esta medida de correlação, por ser baseada na covariância, tem o mesmo referencial de avaliação. Ou seja, se for nula indica que não existe associação entre as variáveis, e se for positiva, indica que as duas variáveis variam conjuntamente no mesmo sentido, ao passo que se for negativa, indica o contrário, ou seja, variação inversa. Mas a correlação tem como grande vantagem em relação à covariância o fato de sua variação ser entre -1 e +1, e mais ainda: quando a correlação se aproxima destes extremos a relação entre as duas variáveis vai se apresentando como cada vez mais linear, de tal modo que, quando as medidas atingem este extremo, a relação entre as duas variáveis é de linearidade perfeita. Isto torna a medida de correlação como sendo aquela universalmente utilizada em estudos exploratórios.

Mas é fácil perceber que a correlação (assim como a covariância) tem o problema de ser baseada na média aritmética, e, portanto, é uma medida não robusta à presença de valores extremos. Adicionalmente, dado o fato de a correlação captar variação conjunta linear, esta deixa de ser uma medida interessante quando duas variáveis são associadas de forma não linear. Para ilustrar estas questões, apresento dois exemplos.

Exemplo 1: Seja $X = \{-30, -10, -5, 10, 20, 33\}$ uma variável de medidas de temperatura, e seja uma variável Y que é igual ao valor de X adicionado a 5, ou seja, $Y = \{-25, -5, 0, 15, 25, 38\}$. É evidente que uma variável é completamente definida (linearmente) pela outra, e, portanto, sua correlação é igual a 1. No momento da computação dos dados, o observador se equivocou na variável Y e em lugar de 15 anotou 155, fazendo $Y = \{-25, -5, 0, 155, 25, 38\}$. Do R, temos os resultados

abaixo, que mostram que o erro de digitação gerou um *outlier* na variável Y e a correlação que seria 1 ficou em 0,48, ou seja, a presença do *outlier* depreciou fortemente a correlação:

```
> X=c(-30, -10, -5, 10, 20, 33)
> Y=c(-25, -5, 0, 15, 25, 38)
> Y_erro=c(-25, -5, 0, 155, 25, 38)
> cor(Y, Y); cor(X, Y_erro)
[1] 1
[1] 0.484502
```

No JASP, vamos em new → data e criaremos duas colunas X e Y com os dados do exemplo. Na sequência, iremos em descriptive statistics → selecionaremos as variáveis de interesse → classical → correlation. O resultado da correlação aparecerá (0,485).

Em Python, utilizaremos o código a seguir:

```
x = [-30,-10, -5, 10, 20, 33]
y = [-25, -5, 0, 155, 25, 38]
corr = pd.Series(x).corr(pd.Series(y))
print(corr)
```

Exemplo 2: Seja novamente $X=\{-30, -10, -5, 10, 20, 33\}$ e seja uma variável Y que se forma a partir do quadrado das medidas de X, ou seja, $Y=\{900, 100, 25, 100, 400, 1089\}$. Ou seja, temos a relação funcional perfeita tal que $Y=f(X)=X^2$, mas uma relação quadrática e não linear. Neste caso, a correlação entre as duas variáveis é 0,17, um valor já próximo de zero, que sinaliza um nível de associação (linear) baixo. O procedimento do R está indicado abaixo:

```
> X=c(-30, -10, -5, 10, 20, 33)
> Y=X^2
> cor(X, Y)
[1] 0.1740339
```

Os procedimentos no JASP e em Python são os mesmos descritos anteriormente, alterando, neste caso, os dados da variável “Y”.

c) Correlação por postos de Spearman

A correlação por postos de Spearman é uma versão da correlação de Pearson, porém aplicada aos postos das duas variáveis sob análise. Antes de indicarmos o procedimento de cálculo, vejamos inicialmente o conceito de postos.

Em estatística, quando temos uma variável qualquer ordenada de forma crescente $X = \{x_{(1)}, x_{(2)}, \dots, x_{(N-1)}, x_{(N)}\}$, dizemos que o valor menor recebe o posto 1, o segundo valor menor recebe o posto 2, e assim sucessivamente até o valor maior, que recebe o posto N. Teremos uma correspondência direta entre a variável X e seus postos $R(X)$ assim definida.

$$X = \{x_{(1)}, x_{(2)}, \dots, x_{(N-1)}, x_{(N)}\} \Leftrightarrow R(X) = \{1, 2, \dots, N-1, N\}$$

Para que cada número tenha seu respectivo posto, quando há números iguais, os postos que seriam correspondentes aos números são somados e a soma é dividida pelo total de repetidos, com o resultado da divisão sendo atribuída a cada um dos valores iguais. Por exemplo, para os números $\{30, 10, 4, 29\}$, temos os seguintes postos $\{4, 2, 1, 3\}$; já para $\{22, 6, 6, 11\}$, temos os dois menores números (que corresponderiam aos postos 1 e 2) repetidos, e, portanto, teremos para estes valores a atribuição do valor 1,5 $((1+2)/2)$, de modo que os postos atribuídos são $\{4, 1,5, 1,5, 3\}$.

De posse deste conhecimento, é possível então calcular a correlação dos postos de Spearman de duas variáveis finitas ou duas amostras $X = \{x_1, x_2, \dots, x_N\}$ e $Y = \{y_1, y_2, \dots, y_N\}$, seguindo o procedimento abaixo:

1. Primeiramente, ordenamos os valores de X do menor para o maior, mantendo a correspondência com a variável Y ;
2. Depois transformamos a variável X em uma variável auxiliar $R(X)$, com a seguinte construção:

$$X = \{x_{(1)}, x_{(2)}, \dots, x_{(i)}, \dots, x_{(N-1)}, x_{(N)}\} \Leftrightarrow R(X) = \{1, 2, \dots, r(x_{(i)}), \dots, N-1, N\}$$

3. Em seguida, ordenamos os valores de Y do menor para o maior valor, e também aqui transformamos a variável Y em uma variável auxiliar Z , com a seguinte construção:

$$Y = \{y_{(1)}, y_{(2)}, \dots, y_{(N-1)}, y_{(N)}\} \Leftrightarrow R(Y) = \{1, 2, \dots, r(y_{(i)}), \dots, N-1, N\}$$

4. Por fim, calculamos a correlação de Pearson entre as variáveis transformadas dos postos, que corresponderá à correlação de Spearman de X e Y , representada por $r\acute{o}(X, Y)$

$$r\acute{o}(X, Y) = Cor(R(X), R(Y)) = \frac{\sum_{i=1}^N (r(x_i) - Me(R(X)))(r(y_i) - Me(R(Y)))}{\sqrt{\sum_{i=1}^N (r(x_i) - Me(R(X)))^2 \sum_{i=1}^N (r(y_i) - Me(R(Y)))^2}}$$

A fórmula mais conhecida da correlação de Spearman é diferente desta que apresentei acima, mas derivada dela (por enquanto não se faz pertinente mostrar como uma se converte na outra, mas faço esta demonstração no exercício resolvido). A fórmula mais recorrente é a seguinte:

$$r\acute{o}(X, Y) = 1 - \frac{6 \sum_{i=1}^N d_i^2}{N(N^2 - 1)}, \text{ em que } d_i = r(x_i) - r(y_i),$$

Vejamos então um exemplo completo, a partir do pacote R. Sejam os resultados de uma pesquisa com seis estudantes para analisar a associação entre suas notas em Estatística e a quantidade de exercícios resolvidos. As notas, na escala de 0 a 100, foram: 65, 80, 33, 40, 59, 92; a quantidades de exercícios resolvidos foram: 55, 200, 45, 45, 90, 250. Temos os seguintes resultados (veja que temos, nas duas formas de cálculo, uma pequena diferença na medida, o que se deve a procedimentos de arredondamento):

```

> not=c(65, 80, 33, 40, 59, 92)      #Notas
> exe=c(55, 200, 45, 45, 90, 250)   #Número de exercícios
> p_not=rank(not); p_exe=rank(exe)   #Atribuição de postos
> cbind(not, exe, p_not, p_exe)      #Tabela completa
      not exe p_not p_exe
[1,]  65  55     4   3.0
[2,]  80 200     5   5.0
[3,]  33  45     1   1.5
[4,]  40  45     2   1.5
[5,]  59  90     3   4.0
[6,]  92 250     6   6.0
>
> di=p_not-p_exe;                  #Diferença de postos
> N=length(di)                    #Tamanho da amostra
> Pearson=cor(not, exe)           #Correlação de Pearson
> Spearman=cor(p_not, p_exe)      #Correlação de Spearman 1
> Spearman2=1-(6*sum(di^2))/(N*(N^2-1)) #Correlação de Spearman 2
> Pearson; Spearman1; Spearman2    #Medidas
[1] 0.8985957
[1] 0.9276337
[1] 0.9285714

```

O procedimento de extração do R pode ser tanto este que expusemos no exemplo, como pode ser de forma direta, a partir da seguinte formulação: `cor.test(Y, X, method="spearman")`; o valor da correlação está indicado por ‘rho’.

No JASP, criaremos o banco de dados com as informações em `new` → `data` e criaremos duas colunas “notas” e “exe” com as informações do exemplo. Na sequência, iremos em `regression` → `classical` → `correlation` → arrastaremos as variáveis de interesse para o campo “Variable” → `Spearman’s rho`. O resultado da correlação aparecerá (0,928).

Em Python, podemos seguir a mesma lógica implementada em R. O procedimento pode ser o que segue abaixo.

```

import numpy as np
import pandas as pd
from scipy.stats import pearsonr, spearmanr

notas=np.array([65,80,33,40,59,92])
exercicios=np.array([55,200,45,45,90,250])

p_not=pd.Series(notas).rank()
p_exe=pd.Series(exercicios).rank()

df=pd.DataFrame({'not': notas, 'exe': exercicios, 'p_not':p_not,
                 'p_exe': p_exe})
print(df) #tabela

di=p_not-p_exe
person_corr,_=pearsonr(notas, exercicios)
spearman_corr,_=spearmanr(notas,exercicios)

N=len(di)
Spearman_corr_manual = 1(6*np.sum(di**2))/(N*(N**2-1))

print(person_corr)
print(spearman_corr)
print(spearman_corr_manual)

```

A correlação de postos de Spearman tem a vantagem de ser robusta à presença de valores extremos, e, para o caso de relações ‘bem comportadas’, dá também uma indicação de associação não linear. Mas esta medida tem a desvantagem de perder o referencial de intensidade dos valores envolvidos, pois todas as medidas, independente de suas magnitudes, são simplesmente substituídas por seus postos. Como alternativa robusta de verificação de associação, ainda mantendo parte dos valores de referência, temos a variância winsorizada, apresentada a seguir.

d) Correlação winsorizada

Uma alternativa interessante para potenciais problemas oriundos da presença de outliers consiste na chamada ‘correlação winsorizada’. Trata-se, na verdade de um procedimento semelhante ao que foi adotado para a média e a variância winsorizadas, ou seja, adotamos um percentual de valores do con-

junto de dados que estejam nos extremos para serem substituídos pelo valor vizinho. Isto mantém a quantidade de observações e viabiliza a extração de resultados utilizando a mesma formulação da correlação de Pearson. A diferença em relação às outras medidas está justamente no fato de aqui utilizarmos o procedimento sobre duas variáveis, em separado.

O procedimento se torna mais fácil se transformamos as duas variáveis originais em duas outras auxiliares, conforme o passo a passo a seguir (consideremos as variáveis genericamente como X e Y):

1. Selecionamos do percentual 'p' de 'winsorização' das duas variáveis;
2. Definimos do número 'w' de itens a serem substituídos, pela seguinte fórmula: $w = \lfloor Np/100 \rfloor$ (N representa o número de elementos da variável);
3. Primeiramente, ordenamos os valores de X do menor para o maior, mantendo a correspondência com a variável Y ;
4. E depois transformamos a variável X em uma variável auxiliar W , com a seguinte construção:

$$X = \{x_{(1)}, x_{(2)}, \dots, x_{(N-1)}, x_{(N)}\} \Rightarrow$$

$$W = \{x_{(w+1)}, x_{(w+1)}, \dots, x_{(w+1)}, x_{(w+2)}, \dots, x_{(N-w-1)}, x_{(N-w)} \dots, x_{(N-w)}, x_{(N-w)}\} \Rightarrow$$

$$W = \{w_{(1)}, w_{(2)}, \dots, w_{(N)}\}$$

5. Em seguida, ordenamos os valores de Y do menor para o maior, mantendo a correspondência original com X , e novamente transformamos a variável Y em uma variável auxiliar Z , com a seguinte construção:

$$Y = \{y_{(1)}, y_{(2)}, \dots, y_{(N-1)}, y_{(N)}\} \Rightarrow$$

$$Z = \{y_{(w+1)}, y_{(w+1)}, \dots, y_{(w+1)}, y_{(w+2)}, \dots, y_{(N-w-1)}, y_{(N-w)} \dots, y_{(N-w)}, y_{(N-w)}\} \Rightarrow$$

$$Z = \{z_{(1)}, z_{(2)}, \dots, z_{(N)}\}$$

6. Por fim, calculamos a correlação winsorizada de X e Y , representada por $Corw$, a partir da correlação produto-momento de Pearson de W e Z , ou seja,

$$Corw(X, Y) = Cor(W, Z) \frac{\sum_{i=1}^N (w_i - Me(W))(z_i - Me(Z))}{\sqrt{\sum_{i=1}^N (w_i - Me(W))^2 \sum_{i=1}^N (z_i - Me(Z))^2}}$$

O procedimento de cálculo da variância winsorizada pode ser feito diretamente no pacote R (no JASP, segue-se o mesmo procedimento), com a seguinte construção (Quadro 10.1):

Quadro 10.1 – Procedimento do R para correlação winsorizada

```
#Opção 1
x=c(...);y=c(...)           #Vetor de dados
install.packages("PairedData") #Instala o pacote de base
library(PairedData)          #Habilita a biblioteca
w=...                         #Define o valor de winsorização
winsor.cor.test(x, y, w/100)  #Cálculo da correlação
```

E em Python, com a seguinte construção (Quadro 10.2):

Quadro 10.2 – Procedimento em Python para correlação winsorizada

```
#Opção 1
Import numpy as np
Import scipy.stats as stats

def winsorize_dados(dados, w = ...) #w é o percentual de
winsorização

x = np.array([....]); y = np.array([....])

x_winsorized=winsorized_dados(x,0.10)
y_winsorized=winsorize_dados(y,0.10)

pearson_corr, _=stats.pearsonr(x_winsorized, y_winsorized)
spearman_corr, _=stats.spearmanr(x_winsorized, y_winsorized)
print(pearson_corr); print(spearman_corr)
```

Temos ainda a opção de criar o algoritmo direto como segunda opção de cálculo, semelhante aos que foram mostrados nos procedimentos de média e variância winsorizadas. Deixamos este desafio para o leitor interessado em um maior aprofundamento nos procedimentos computacionais do R e em Python.

Tomando novamente o exemplo acima, e aplicando diretamente no R o procedimento de extração da correlação winsorizada de 20%, teremos o resultado abaixo. A implementação em Python fica como desafio para o leitor. É fácil ver

que a presença do *outlier*, decorrente do erro de digitação, gera uma modificação na correlação (a correlação foi de 1 para 0,91), porém o efeito é muito menor que aquele que se observou na correlação não winsorizada.

```
> X=c(-30, -10, -5, 10, 20, 33)
> Y=c(-25, -5, 0, 15, 25, 38)
> Y_erro=c(-25, -5, 0, 155, 25, 38)
> winsor.cor.test(X, Y)
      winsorized correlation, trim=0.2

data:  Y and Y
t = 134217728, df = 2, p-value < 2.2e-16
alternative hypothesis: true (winsorized) correlation is not equal
to 0
sample estimates:
cor
  1

> winsor.cor.test(X, Y_erro)
      winsorized correlation, trim=0.2

data:  X and Y_erro
t = 4.4359, df = 2, p-value = 0.04725
alternative hypothesis: true (winsorized) correlation is not equal
to 0
sample estimates:
cor
```

- Um comentário sobre as medidas de correlação

O conjunto de medidas apresentado para correlações não dá conta dos diversos problemas e verificações de associação possíveis quantitativas. Na verdade, existem diversas alternativas de medição de correlação. As que apresento são as mais recorrentes e certamente são suficientes para a grande maioria dos problemas em que estas medidas podem ser úteis.

Em uma perspectiva de análise exploratória de dados, e seguindo a orientação de referência de que devemos explorar os dados de tantas formas quanto possível e relevantes, o entendimento aqui defendido é de que as medidas devem ser todas extraídas para uma avaliação comparativa. Em geral, um

comentário baseado em três medidas é muito mais seguro do que um em que utilizamos somente uma medida.

Assim como indiquei na avaliação de tabelas, as medidas de correlação são indicadores de associação, e não indicam, necessariamente, que uma variável influencia, condiciona ou prediz a outra. Ou seja, *correlação é uma medida de associação e não de influência*. Jamais poderemos dizer que uma variável A influencia uma variável B somente porque há uma correlação forte entre A e B.

Há sentido sim em falar que a correlação capta e indica a intensidade de influência de uma variável sobre outra, desde que há uma justificativa para tal influência. A correlação evidencia e reflete, mas não define influência. Em outras palavras, influência, condicionamento ou predição são características dos construtos que estão medidos por meio ‘das variáveis’, e suas medidas, quando extraídas em conjunto, captam a influência pressuposta. A correlação pode ocorrer devido ao fato de uma variável influenciar ou predizer outra, mas o fato de uma correlação ser não nula não indica, por si, influência.

Para guardar uma forma de avaliação das opções associadas à correlação, deixo o registro de três opções quando a correlação de X e Y é não nula: ou X influencia Y, ou Y influencia X, ou há uma terceira variável Z que influencia X e Y, ou, por fim, a correlação existe por motivos fortuitos e eventuais. Afora esta última opção, podemos simbolicamente representar as outras da seguinte forma:

$$Cor(X, Y) \text{ é não nula} \rightarrow \text{ou } (X \rightarrow Y) \text{ ou } (Y \rightarrow X) \text{ ou } \exists Z / (Z \rightarrow X_i) \text{ e } (Z \rightarrow X_j)$$

Particularmente quando uma variável influencia a outra, em lugar de utilizar análise de correlação optamos por utilizar análise de regressão, que detalhamos a seguir.

10.4. Análise de regressão simples

A análise de correlação é uma técnica que, em geral, antecede às análises posteriores. A correlação de Pearson, em particular, é base de referência para diversas técnicas relevantes em análise de dados. No entanto, a correlação é limitada em explicar predição, quando temos uma associação dessa natureza (em que uma variável prediz ou influencia a outra). A análise de regressão linear simples vem resolver este problema.

A lógica da técnica é a seguinte: quando temos duas variáveis sobre as quais já sabemos ou esperamos que haja uma relação de predição ou de influência, então podemos tentar encontrar uma fórmula matemática que permita informar o valor de uma variável dado o valor da outra variável. Essas relações funcionais são diversas (podem assumir forma quadrática, exponencial, logarítmica...), mas a formulação mais simples, e que resolve grande parte dos problemas práticos, é aquela em que admitimos a forma de equação linear.

Conforme se verifica na teorização matemática clássica, em uma relação linear entre duas variáveis, a variável dependente é definida como a soma de uma constante a com um coeficiente não nulo b multiplicado pela segunda variável independente. Ou seja, $Y = f(X) = a + bX$, ou ainda $Y = \beta_0 + \beta_1 X$, em uma formulação de uso mais comum em estatística. Ainda de acordo com a teorização matemática, a fórmula indicada, quando levada a um eixo cartesiano ortogonal, gera pares ordenados perfeitamente alinhados.

Quando visualizamos os dados de duas variáveis plotados em um *scatterplot* é possível, por vezes, notar uma tendência de linearidade na nuvem de pontos, o que faz supor que, possivelmente, há uma explicação linear de uma variável pela outra, havendo, por outro lado, alguma perturbação na previsão perfeita.

Pelo *scatterplot* é possível imaginar que, de fato, uma reta explicaria bem a relação entre as duas variáveis. É possível visualizar que, na medida em que a quantidade de horas de estudo cresce, há uma tendência de as notas crescerem. Obviamente, isto não ocorre em uma relação linear perfeita. Por exemplo, para as pessoas que declararam duas horas de estudo, temos notas oscilando entre 28 e 40; ou seja, na tentativa de prever a nota pela quantidade de horas de estudo devemos assumir que, para cada hora, temos um valor predito de notas, mas com um erro de previsão.

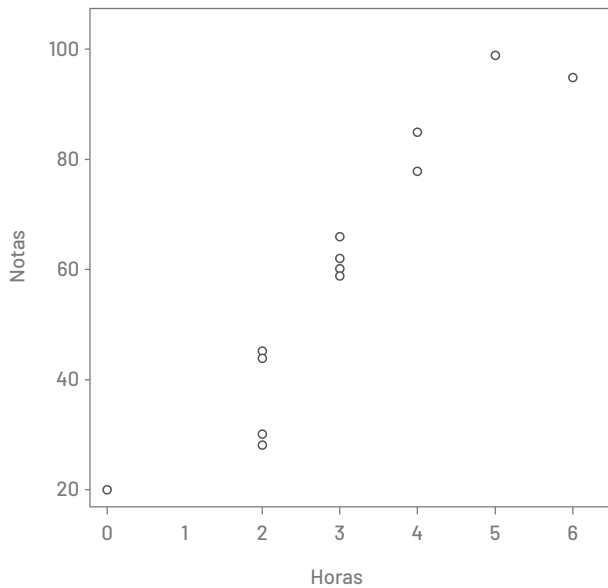
Exemplo: Em uma sala de aula de Estatística, o professor decidiu avaliar o efeito do tempo de estudo da matéria sobre as notas dos alunos. Em sua crença, quanto mais o aluno estuda maiores são suas notas. Para tanto, tomou para os 15 alunos da turma seu tempo declarado de estudo semanal da disciplina fora de sala, e suas notas. Os resultados, já no formato do R, estão indicados no quadro da esquerda, e à direita temos o *scatterplot* das duas variáveis:

```
> Horas=c(3, 2, 3, 4, 5, 0, 4, 6, 2, 4, 2, 3, 2, 3, 3)
```

```
> Notas=c(60, 30, 62, 78, 99, 20, 85, 95, 45, 78, 44, 60, 28, 66, 59)
```

```
> cbind(Horas, Notas)
```

	Horas	Notas
[1,]	3	60
[2,]	2	30
[3,]	3	62
[4,]	4	78
[5,]	5	99
[6,]	0	20
[7,]	4	85
[8,]	6	95
[9,]	2	45
[10,]	4	78
[11,]	2	44
[12,]	3	60
[13,]	2	28
[14,]	3	66
[15,]	3	59



Mas no conjunto, é de bom senso supor que, se tivermos uma relação funcional em uma fórmula, mesmo que nossas previsões tragam um erro associado, podemos ter uma análise aprimorada. Nosso grande desafio será encontrar esta fórmula, e fazemos isto ‘estimando’ uma reta a partir dos dados que temos. Isto é perfeitamente possível, desde que lancemos mão de alguns procedimentos. Vejamos então a lógica do processo.

- Derivação da equação de regressão

Para que tenhamos uma forma de ‘estimação’ da reta de regressão, primeiramente observemos que, para cada valor da variável dependente y_i teremos um valor estimado, que representaremos por $\widehat{y}_i$; teremos então $\widehat{y}_i = \beta_0 + \beta_1 x_i$. Chamemos a diferença entre o valor observado e o estimado de e_i . Teremos então:

$$e_i = y_i - \widehat{y}_i, \text{ ou ainda, } e_i = y_i - \beta_0 - \beta_1 x_i$$

Para tornar o valor com base em e_i sempre positivo (não é conveniente trabalhar com valores negativos aqui), podemos tomar o quadrado de e_i , ou seja, $e_i^2 = (y_i - \beta_0 - \beta_1 x_i)^2$. Nestes termos, podemos tomar uma medida do valor somado dos erros quadráticos, o que permitirá então que estimemos os parâmetros β_0 e β_1 . Representaremos os estimadores por $\widehat{\beta}_0$ e $\widehat{\beta}_1$. Teremos então

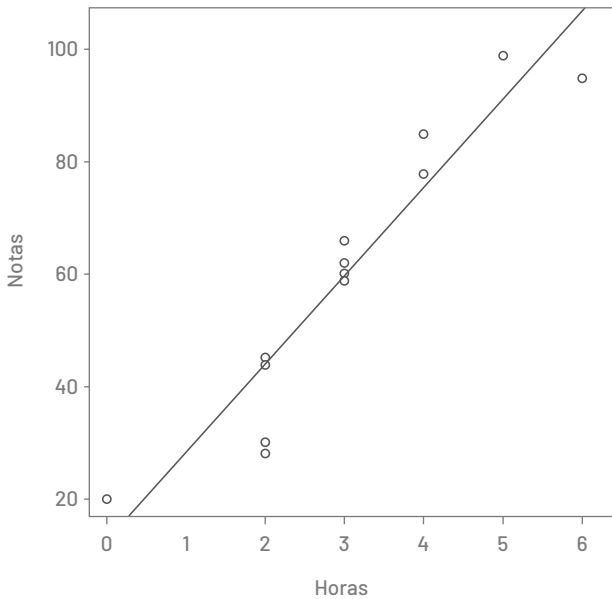
$$\sum_{i=1}^N e_i^2 = \sum_{i=1}^N (y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i)^2$$

Dentre as formas que temos para gerar valores dos estimadores, optamos por escolher uma minimização dos erros quadráticos. Tal procedimento é feito pela extração da derivada da função acima em relação aos estimadores e em seguida igualando a derivada a zero. Isto gerará um sistema de equações que, quando resolvido, terá como resultados os seguintes:

$$\begin{aligned} \widehat{\beta}_0 &= Me(Y) - \widehat{\beta}_1 Me(X) \\ \widehat{\beta}_1 &= \frac{\sum_{i=1}^N (x_i - Me(X))(y_i - Me(Y))}{\sum_{i=1}^N (x_i - Me(X))^2} = \frac{Cov(X, Y)}{Var(X)} \end{aligned}$$

O procedimento de cálculo para chegar às fórmulas acima está como exercício resolvido. Pelo que observamos, a estimação dos parâmetros pela minimização dos quadrados dos erros entre os valores observados e os estimados da variável dependente tem a grande vantagem de gerar estimadores baseados somente em medidas conhecidas: média, variância e covariância.

Gráfico 10.2 – Reta de regressão



Efetuada os cálculos completos do exemplo acima, encontraremos a seguinte formulação: $\text{Notas} = 11,13 + 15,77 \cdot \text{Horas}$. O Gráfico 10.2 mostra ao conjunto de dados juntamente com a reta associada a esta fórmula. Pelo que mostra o gráfico, a reta passa ao longo dos pontos, indicando uma boa proximidade entre os pontos observados e os valores que a reta estima.

– O coeficiente de determinação

Uma questão que imediatamente emerge concerne a quão bem esta reta/fórmula prevê os resultados observados. De fato, se neste Gráfico 10.2 observamos uma reta bem próxima dos valores observados, que já parecem bem alinhados; o mesmo não parece ser o caso do gráfico 10.1. Certamente, naquele conjunto de dados uma reta não uniria tão bem os pontos como ocorre na reta do último exemplo. Para dar uma solução para esta questão, foi desenvolvido o conceito de ‘coeficiente de determinação’, que tem a seguinte construção, sempre em relação à variável dependente:

- Primeiramente, tomamos a soma dos quadrados das diferenças entre os ‘valores observados’ e sua média, e damos o nome se ‘soma dos quadrados totais’;

- Em segundo lugar, tomamos a soma dos quadrados das diferenças entre os ‘valores estimados’ e a média da variável resposta, e damos o nome se ‘soma dos quadrados explicados’;
- Por fim, definimos o coeficiente de determinação, representado por R^2 , como a razão entre a soma dos quadrados explicados e a soma dos quadrados totais. Simbolicamente, teremos:

$$\gg \text{ Soma dos quadrados totais: } SQT = \sum_{i=1}^N (y_i - Me(Y))^2$$

$$\gg \text{ Soma dos quadrados explicados: } SQE = \sum_{i=1}^N (\widehat{y}_i - Me(Y))^2$$

$$\gg \text{ Coeficiente de determinação: } R^2 = \frac{SQE}{SQT} = \frac{\sum_{i=1}^N (\widehat{y}_i - Me(Y))^2}{\sum_{i=1}^N (y_i - Me(Y))^2}$$

O coeficiente de determinação é um valor de entre 0 e 1, e o nível de explicação é tão melhor quanto mais este se aproxima de 1. No caso do exemplo acima, aplicando os cálculos encontramos um valor de 0,89, o que é um valor já elevado, e indica que, efetivamente, a reta de regressão estimada explica muito bem os dados observados (o que é esperado, uma vez que, como mostra o gráfico, a reta se aproxima bem dos pontos do *scatterplot*).

Do ponto de vista computacional, os três softwares de referência permitem proceder à regressão linear simples. Para o Excel, deixamos como exercício a identificação dos caminhos. Para o R, damos um detalhamento maior, para o que podemos extrair desde agora (normalmente os programas trazem muito mais detalhes do que estes que apresentamos, inclusive com resultados inferenciais que não nos interessam agora). Assim, quando temos uma variável independente X que prediz uma variável dependente Y, no R fazemos `lm(Y~X)`, que teremos, como primeira saída, os estimadores dos parâmetros. Quando indicamos o comando `summary(lm(Y~X))`, teremos também o coeficiente de determinação, anotado como Multiple R-squared.

Abaixo temos os resultados para os dados do exemplo acima, e estão sublinhados os resultados que nos interessam. Veja que o coeficiente que representa o termo constante está indicado como sendo o intercepto da reta, e o termo que multiplica a variável horas está ao lado do nome da variável (para os propósitos deste livro, não nos interessam as demais medidas que o R gera).

```
> summary(lm(Notas~Horas))
Call:
lm(formula = Notas ~ Horas)

Residuals:
    Min       1Q   Median       3Q      Max
-15.7742  -0.1613   1.2258   4.5645   9.6774

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  12.226      5.091    2.401  0.032 *
Horas        15.774      1.512   10.430 1.1e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.135 on 13 degrees of freedom
Multiple R-squared:  0.8933,    Adjusted R-squared:  0.8851
F-statistic: 108.8 on 1 and 13 DF,  p-value: 1.101e-07
```

No JASP, criaremos o banco de dados com as informações em new → data e criaremos duas colunas “notas” e “horas” com as informações do exemplo. Na sequência, iremos em regression → classical → linear regression → arrastaremos a variável “notas” para o campo dependent variable e a variável “horas” para o campo “covariates”. O resultado da regressão aparecerá similar ao exposto acima.

Em Python, podemos importar a biblioteca statsmodels e utilizar a seguinte rotina:

```
import numpy as np
import statsmodels.api as sm

X = sm.add_constant(x) #variável independente
model = sm.OLS(y, X).fit() #y = variável dependente
print(model.summary())
```

Na sequência, temos dois exercícios resolvidos, relativos aos procedimentos de manuseio matemático que optamos por não exibir anteriormente.

Exercícios resolvidos

1. Mostre que as fórmulas de correlação de Spearman são idênticas, ou seja,

$$r_s(X, Y) = \frac{\sum_{i=1}^N (r(x_i) - Me(R(X)))(r(y_i) - Me(R(Y)))}{\sqrt{\sum_{i=1}^N (r(x_i) - Me(R(X)))^2 \sum_{i=1}^N (r(y_i) - Me(R(Y)))^2}} = 1 - \frac{6 \sum_{i=1}^N d_i^2}{N(N^2 - 1)}$$

em que $d_i = r(x_i) - r(y_i)$,

Partirei da primeira fórmula para chegar na segunda. Para tanto, o desenvolvimento que segue está dividido em duas partes, a primeira concentrada na expressão do denominador, e a segunda concentrada na expressão do numerador.

Parte 1 – Foco no denominador

Considerando a primeira fórmula, e dado que as duas variáveis são na verdade semelhantes quanto aos seus valores (ambas têm valores de 1 até N), é fácil ver que as médias são iguais, ou seja, $Me(R(X)) = Me(R(Y))$. Adicionalmente, como ambas são duas progressões aritméticas de razão 1, com mínimo 1 e máximo N, a média é calculada a partir da expressão a seguir:

$$Me(R(X)) = Me(R(Y)) = \frac{\sum_{i=1}^N r(y_i)}{N} = \frac{\frac{N(N+1)}{2}}{N} \implies Me(R(X)) = \frac{(N+1)}{2} \quad (\text{I})$$

Também é fácil compreender que, no denominador da correlação, as duas expressões de somatório geram o mesmo valor, pois são fórmulas associadas, na verdade, aos mesmos valores (que são os postos atribuídos às duas variáveis). Portanto, podemos reescrever a expressão em função dos postos de X ou de Y. Teremos:

$$\begin{aligned} & \sqrt{\sum_{i=1}^N (r(x_i) - Me(R(X)))^2 \sum_{i=1}^N (r(y_i) - Me(R(Y)))^2} = \\ & \sqrt{\left(\sum_{i=1}^N (r(x_i) - Me(R(X)))^2 \right)^2} = \\ & \sum_{i=1}^N (r(x_i) - Me(R(X)))^2 = \sum_{i=1}^N \left(r(x_i) - \frac{(N+1)}{2} \right)^2 \quad (\text{I}) \end{aligned}$$

O desenvolvimento do somatório acima torna a expressão mais simples, com a seguinte configuração (este resultado já foi mostrado anteriormente):

$$\sum_{i=1}^N \left(r(x_i) - \frac{(N+1)}{2} \right)^2 = \sum_{i=1}^N r(x_i)^2 - N \left(\frac{(N+1)}{2} \right)^2 \quad (\text{II})$$

Podemos desenvolver ainda a primeira parcela do segundo membro de II. Vemos inicialmente que,

$$\sum_{i=1}^N r(x_i)^2 = 1^2 + 2^2 + \dots + N^2$$

Para encontrar uma expressão melhor que esta, temos que seguir um atalho, tomando como referência a soma dos cubos. Em geral $(a+b)^3 = a^3 + 3a^2b + 3ab^2 + b^3$; em particular $(i+1)^3 = i^3 + 3i^2 + 3i + 1$. Podemos variar i , da seguinte forma

$$\text{Para } i = 1: (1+1)^3 = 1^3 + 3 \cdot 1^2 + 3 \cdot 1 + 1$$

$$\text{Para } i = 2: (2+1)^3 = 2^3 + 3 \cdot 2^2 + 3 \cdot 2 + 1$$

$$\text{Para } i = 3: (3+1)^3 = 3^3 + 3 \cdot 3^2 + 3 \cdot 3 + 1$$

...

$$\text{Para } i = N-1: ((N-1)+1)^3 = (N-1)^3 + 3 \cdot (N-1)^2 + 3 \cdot (N-1) + 1$$

$$\text{Para } i = N: (N+1)^3 = N^3 + 3 \cdot N^2 + 3 \cdot N + 1$$

Veja que o primeiro membro de cada linha é igual à primeira parcela da linha seguinte $((1+1)^3 = 2^3, \dots, ((N-1)+1)^3 = N^3)$. Desta forma, se somarmos todas as expressões em seus respectivos membros, teremos o seguinte:

$$\begin{aligned} \sum_{i=1}^{N-1} (i+1)^3 + (N+1)^3 &= 1^3 + \sum_{i=1}^{N-1} (i+1)^3 + 3 \sum_{i=1}^N i^2 + 3 \sum_{i=1}^N i + N \cdot 1 \Rightarrow \\ &\Rightarrow (N+1)^3 = 3 \sum_{i=1}^N i^2 + 3N \frac{(N+1)}{2} + N+1 \Rightarrow \\ &\Rightarrow 3 \sum_{i=1}^N i^2 = (N+1)^3 - 3N \frac{(N+1)}{2} - (N+1) \Rightarrow \\ &\Rightarrow 6 \sum_{i=1}^N i^2 = 2(N+1)^3 - 3N(N+1) - 2(N+1) = (N+1) [2(N+1)^2 - 3N - 2] \Rightarrow \end{aligned}$$

$$\Rightarrow 6 \sum_{i=1}^N i^2 = (N+1)[2N^2 + 4N + 2 - 3N - 2] = (N+1)[2N^2 + N]$$

$$\sum_{i=1}^N i^2 = \frac{N(N+1)(2N+1)}{6} \quad (\text{III})$$

Das expressões II e III, e tomando , teremos no denominador a seguinte expressão:

$$\begin{aligned} & \sum_{i=1}^N r(x_i)^2 - N \left(\frac{(N+1)}{2} \right)^2 = \frac{N(N+1)(2N+1)}{6} - N \left(\frac{(N+1)}{2} \right)^2 = \\ &= \frac{N(N+1)(2N+1)}{6} - \frac{N(N+1)^2}{4} = \frac{2N(N+1)(2N+1) - 3N(N+1)^2}{12} = \\ &= \frac{N(N+1)[2(2N+1) - 3(N+1)]}{12} = \frac{N(N+1)[4N+2-3N-3]}{12} \Rightarrow \\ &\Rightarrow \sum_{i=1}^N r(x_i)^2 - N \left(\frac{(N+1)}{2} \right)^2 = \frac{N(N+1)(N-1)}{12} = \frac{N(N^2-1)}{12} \end{aligned}$$

Ou seja,

$$\sqrt{\sum_{i=1}^N (r(x_i) - Me(R(X)))^2 \sum_{i=1}^N (r(y_i) - Me(R(Y)))^2} = \frac{N(N^2-1)}{12} \quad (\text{IV})$$

Parte 2 – Foco no numerador

Relativo ao numerador, podemos também desenvolver a expressão, conforme indicado abaixo:

$$\begin{aligned} & \sum_{i=1}^N (r(x_i) - Me(R(X)))(r(y_i) - Me(R(Y))) = \\ &= \sum_{i=1}^N \left(r(x_i) - \frac{(N+1)}{2} \right) \left(r(y_i) - \frac{(N+1)}{2} \right) = \\ &= \sum_{i=1}^N \left(r(x_i)r(y_i) - r(x_i)\frac{(N+1)}{2} - r(y_i)\frac{(N+1)}{2} + \left(\frac{(N+1)}{2} \right)^2 \right) = \\ &= \sum_{i=1}^N \left(r(x_i)r(y_i) - (r(y_i) + r(x_i))\frac{(N+1)}{2} + \left(\frac{(N+1)}{2} \right)^2 \right) = \\ &= \sum_{i=1}^N r(x_i)r(y_i) - \frac{(N+1)}{2} \sum_{i=1}^N (r(y_i) + r(x_i)) + \sum_{i=1}^N \left(\frac{(N+1)}{2} \right)^2 = \end{aligned}$$

$$\begin{aligned}
&= \sum_{i=1}^N r(x_i)r(y_i) - \frac{(N+1)}{2} \left(\sum_{i=1}^N r(y_i) + \sum_{i=1}^N r(x_i) \right) + N \left(\frac{(N+1)}{2} \right)^2 = \\
&= \sum_{i=1}^N r(x_i)r(y_i) - \frac{(N+1)}{2} \left(N \frac{(N+1)}{2} + N \frac{(N+1)}{2} \right) + N \left(\frac{(N+1)}{2} \right)^2 = \\
&= \sum_{i=1}^N r(x_i)r(y_i) - \frac{(N+1)}{2} \left(2N \frac{(N+1)}{2} \right) + N \left(\frac{(N+1)}{2} \right)^2 = \\
&= \sum_{i=1}^N r(x_i)r(y_i) - 2N \left(\frac{(N+1)}{2} \right)^2 + N \left(\frac{(N+1)}{2} \right)^2 = \\
&= \sum_{i=1}^N r(x_i)r(y_i) - N \left(\frac{(N+1)}{2} \right)^2 \Rightarrow \\
&\Rightarrow \sum_{i=1}^N (r(x_i) - Me(R(X)))(r(y_i) - Me(R(Y))) = \sum_{i=1}^N r(x_i)r(y_i) - \frac{N(N+1)^2}{4} \quad (V)
\end{aligned}$$

Precisamos agora de uma expressão para o primeiro membro do somatório acima. Lembremos que, para quaisquer valores a e b , $(a - b)^2 = a^2 + b^2 - 2ab$, de modo que $ab = \frac{a^2}{2} + \frac{b^2}{2} - \frac{(a - b)^2}{2}$. Portanto,

$$\begin{aligned}
\sum_{i=1}^N r(x_i)r(y_i) &= \sum_{i=1}^N \left(\frac{r(x_i)^2}{2} + \frac{r(y_i)^2}{2} - \frac{(r(x_i) - r(y_i))^2}{2} \right) = \\
&= \sum_{i=1}^N \frac{r(x_i)^2}{2} + \sum_{i=1}^N \frac{r(y_i)^2}{2} - \sum_{i=1}^N \frac{(r(x_i) - r(y_i))^2}{2} = \\
&= \frac{1}{2} \left(\sum_{i=1}^N r(x_i)^2 + \sum_{i=1}^N r(y_i)^2 \right) - \sum_{i=1}^N \frac{(r(x_i) - r(y_i))^2}{2} = \\
&= \sum_{i=1}^N r(x_i)^2 - \sum_{i=1}^N \frac{(r(x_i) - r(y_i))^2}{2}, \text{ pois } \sum_{i=1}^N r(x_i)^2 = \sum_{i=1}^N r(y_i)^2
\end{aligned}$$

Retomando o resultado de III, teremos que,

$$\sum_{i=1}^N r(x_i)r(y_i) = \frac{N(N+1)(2N+1)}{6} - \sum_{i=1}^N \frac{(r(x_i) - r(y_i))^2}{2}$$

Se $d_i = r(x_i) - r(y_i)$, chegaremos à seguinte expressão:

$$\sum_{i=1}^N r(x_i)r(y_i) = \frac{N(N+1)(2N+1)}{6} - \frac{1}{2} \sum_{i=1}^N d_i^2 \quad (VI)$$

Tomando por referência as expressões V e VI, teremos a seguinte configuração:

$$\begin{aligned} \sum_{i=1}^N (r(x_i) - Me(R(X)))(r(y_i) - Me(R(Y))) &= \\ &= \frac{N(N+1)(2N+1)}{6} - \frac{1}{2} \sum_{i=1}^N d_i^2 - \frac{N(N+1)^2}{4} \end{aligned}$$

Da expressão IV, já sabemos que a primeira e a última parcelas do segundo membro simplificam-se para , de modo que teremos:

$$\sum_{i=1}^N (r(x_i) - Me(R(X)))(r(y_i) - Me(R(Y))) = \frac{N(N^2 - 1)}{12} - \frac{1}{2} \sum_{i=1}^N d_i^2$$

- Retomando a fórmula conjunta

$$\begin{aligned} r\acute{o}(X, Y) &= \frac{\sum_{i=1}^N (r(x_i) - Me(R(X)))(r(y_i) - Me(R(Y)))}{\sqrt{\sum_{i=1}^N (r(x_i) - Me(R(X)))^2 \sum_{i=1}^N (r(y_i) - Me(R(Y)))^2}} = \\ &= \frac{\frac{N(N^2-1)}{12} - \frac{1}{2} \sum_{i=1}^N d_i^2}{\frac{N(N^2-1)}{12}} = \frac{12}{N(N^2 - 1)} \left(\frac{N(N^2 - 1)}{12} - \frac{1}{2} \sum_{i=1}^N d_i^2 \right) \Rightarrow \\ &\Rightarrow r\acute{o}(X, Y) = 1 - \frac{6 \sum_{i=1}^N d_i^2}{N(N^2 - 1)} \end{aligned}$$

2. Deduza, a partir das fórmulas do somatório do quadrado do erro, as fórmulas dos estimadores da equação de regressão linear simples.

Como vimos, a fórmula do somatório dos quadrados dos erros é a seguinte:

$$\sum_{i=1}^N e_1^2 = \sum_{i=1}^N (y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i)^2$$

O procedimento consiste em encontrar os valores de $\widehat{\beta}_0$ e $\widehat{\beta}_1$ que minimizem a expressão. Como indica a teoria matemática da derivada, os valores dos estimadores são identificados quando igualamos as derivadas parciais zero, ou seja:

$$\begin{aligned} \frac{\partial}{\partial \widehat{\beta}_0} \left(\sum_{i=1}^N e_1^2 \right) &= \frac{\partial}{\partial \widehat{\beta}_0} \left(\sum_{i=1}^N (y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i)^2 \right) = 0 \\ \frac{\partial}{\partial \widehat{\beta}_1} \left(\sum_{i=1}^N e_1^2 \right) &= \frac{\partial}{\partial \widehat{\beta}_1} \left(\sum_{i=1}^N (y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i)^2 \right) = 0 \end{aligned}$$

Da primeira expressão, teremos:

$$\begin{aligned}
 -2 \sum_{i=1}^N (y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i) &= 0 \Rightarrow \\
 \Rightarrow \sum_{i=1}^N y_i - \sum_{i=1}^N \widehat{\beta}_0 - \sum_{i=1}^N \widehat{\beta}_1 x_i &= 0 \Rightarrow N \widehat{\beta}_0 = \sum_{i=1}^N y_i - \widehat{\beta}_1 \sum_{i=1}^N x_i \Rightarrow \\
 \Rightarrow \widehat{\beta}_0 &= \frac{\sum_{i=1}^N y_i}{N} - \widehat{\beta}_1 \frac{\sum_{i=1}^N x_i}{N} \Rightarrow \widehat{\beta}_0 = Me(Y) - \widehat{\beta}_1 Me(X) \quad (\text{I})
 \end{aligned}$$

Da segunda expressão, teremos:

$$\begin{aligned}
 \frac{\partial}{\partial \widehat{\beta}_1} \left(\sum_{i=1}^N (y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i)^2 \right) &\Rightarrow \\
 \Rightarrow \frac{\partial}{\partial \widehat{\beta}_1} \left(\sum_{i=1}^N (y_i - (Me(Y) - \widehat{\beta}_1 Me(X)) - \widehat{\beta}_1 x_i)^2 \right) &= 0 \Rightarrow \\
 \Rightarrow \frac{\partial}{\partial \widehat{\beta}_1} \left(\sum_{i=1}^N (y_i - Me(Y) - \widehat{\beta}_1 (x_i - Me(X)))^2 \right) &= 0 \Rightarrow \\
 \Rightarrow -2 \left(\sum_{i=1}^N ((y_i - Me(Y)) - \widehat{\beta}_1 (x_i - Me(X))) (x_i - Me(X)) \right) &= 0 \Rightarrow \\
 \Rightarrow \sum_{i=1}^N ((y_i - Me(Y))(x_i - Me(X))) - \widehat{\beta}_1 \sum_{i=1}^N (x_i - Me(X))(x_i - Me(X)) &= 0 \Rightarrow \\
 \Rightarrow \widehat{\beta}_1 \sum_{i=1}^N (x_i - Me(X))^2 = \sum_{i=1}^N ((y_i - Me(Y))(x_i - Me(X))) &\Rightarrow \\
 \widehat{\beta}_1 = \frac{\sum_{i=1}^N (x_i - Me(X))(y_i - Me(Y))}{\sum_{i=1}^N (x_i - Me(X))^2} = \frac{\frac{\sum_{i=1}^N (x_i - Me(X))(y_i - Me(Y))}{N}}{\frac{\sum_{i=1}^N (x_i - Me(X))^2}{N}} &\Rightarrow \\
 \Rightarrow \widehat{\beta}_1 = \frac{Cov(X, Y)}{Var(X)}
 \end{aligned}$$

10.5. Resumo

Neste capítulo, desenvolvemos o conteúdo da análise bivariada considerando somente variáveis quantitativas. Os principais pontos de destaque foram os seguintes:

- A análise de duas variáveis quantitativas pode ser feita por meio de comparação de medidas descritivas, procedimentos gráficos, análise de medidas e análise de regressão;

- A análise comparativa de medidas consiste em extrair em cada variável as medidas de interesse do pesquisador e apresentá-las em conjunto, para viabilizar uma comparação;
- A análise gráfica consiste em plotar os pares ordenados das medidas das variáveis em um eixo cartesiano ortogonal, formando o que chamamos de gráfico de dispersão ou *scatterplot*;
- A análise por medidas é baseada na covariância e nas diversas alternativa de coeficiente de correlação existentes.
- O coeficiente mais utilizado é o coeficiente de Pearson, porém esta medida não é robusta à presença de outliers;
- Como alternativas para medidas de correlação, temos o coeficiente de correlação de Spearman, que é baseada em postos, além do coeficiente de correlação winsorizado, que segue a mesma lógica do procedimento de winsorização anteriormente mostrado;
- A correlação é uma medida de associação entre variáveis, calculada a partir de dados disponíveis. Por esta razão, a correlação não pode ser tomada como indicadora de influência ou predição. A influência, o condicionamento ou a predição são características dos construtos que geram as variáveis, e a correlação reflete tal condição e sua intensidade;
- Para o caso de haver influência, condicionamento ou predição, a técnica recomendada é a análise de correlação simples, que consiste em extrair a partir dos dados uma fórmula de uma reta que possa ser utilizada para explicar como uma variável prevê a outra;
- O nível de explicação de uma variável pela outra, quando extraída por regressão, é medida pelo coeficiente de determinação que varia entre 0 e 1, quanto maior for o coeficiente maior é a explicação da variável predita pela preditora;
- As formulações envolvendo correlação e regressão são possíveis de serem extraída a partir de alguns cálculos matemáticos, que foram efetuados e mostrados nos exercícios resolvidos.

Exercícios

1. Considere os dados do Apêndice D, e considerando especificamente as variáveis de valor, faça um procedimento de agregação para geração de uma ‘medida geral de valor’. Para tanto, em cada respondente extraia a média dos escores. Na variável resultante, efetue os seguintes procedimentos, global e por curso (são três procedimentos):
 - a) Calcule as médias aritmética, aparada e winsorizada (10%) e apresente uma estimativa da medida global de percepção de valor da população.
 - b) Calcule as seguintes medidas de dispersão: intervalo interquartil, desvio médio, desvio padrão, desvio padrão truncado e winsorizado (10%), e desvio absoluto mediano.
 - c) Calcule todas as medidas de assimetria e curtose.
2. Utilizando as medidas de correlação estudadas, faça sua aplicação nas variáveis ‘valor percebido geral’, identificação e prestígio percebido, segundo os dados em Apêndice D. Coloque as medidas em conjunto, na forma de uma matriz de correlações, e compare os resultados dos três tipos de correlação verificados.
3. Pesquise sobre a correlação de Kendall e verifique os métodos de extração dessas medidas nos programas computacionais de uso. Em seguida, faça a extração dessa medida para as variáveis da questão anterior e compare com os resultados obtidos na questão anterior.
4. Considerando os resultados das duas questões anteriores, analise as seguintes hipóteses:

Hipótese 1: ‘O nível de identificação dos estudantes de Administração e Turismo com seu curso tem forte associação com a reputação percebida na profissão’.

Hipótese 2: ‘O nível de valor percebido geral dos estudantes de Administração e Turismo com seu curso tem forte associação com a reputação percebida na profissão’.

Hipótese 3: ‘O nível de valor percebido geral dos estudantes de Administração e Turismo com seu curso tem forte associação com a identificação do estudante com seu curso.’

5. Utilizando a regressão simples no pacote R, no JASP e em Python, estime a reputação percebida na profissão do aluno a partir de sua percepção global de valor, e em seguida estime o nível de identificação do aluno com o curso a partir do valor percebido.
6. Efetue, para o exemplo dado sobre a relação entre horas de estudo e notas, a extração completa dos estimadores de regressão e do coeficiente de determinação, aplicando as fórmulas.
7. Tomemos por definição da ‘soma dos quadrados dos resíduos’ de regressão (SQR) a soma dos quadrados das diferenças entre os valores observado e estimado da variável dependente. Prove então que:

$$\text{Se } SQR = \sum_{i=1}^N (y_i - \widehat{y}_i)^2, \text{ então } R^2 = 1 - \frac{SQR}{SQT}$$

8. Utilizando os resultados das fórmulas de regressão, prove que:

$$R^2 = \frac{\widehat{\beta}_1 \sum_{i=1}^N (x_i - Me(X))y_i}{\sum_{i=1}^N (y_i - Me(Y))^2}$$

CAPÍTULO 11

APLICAÇÕES COM SUPORTE DE INTELIGÊNCIA ARTIFICIAL_

Em complemento ao conteúdo apresentado, tecemos algumas considerações neste capítulo acerca do uso de ferramentas de inteligência artificial (IA) como auxílio no manuseio estatístico de dados. É certo que o uso de tais ferramentas não dispensa a supervisão humana, e a eficácia de tal supervisão só poderá se concretizar quando há pleno domínio dos conhecimentos anteriormente expostos. Isto posto, seguiremos na compreensão de como a IA pode contribuir nos processos de análise exploratória de dados.

11.1 Princípios básicos para aplicação

Em geral, as interações com ferramentas de inteligência artificial popularizadas, como ChatGPT, Gemini, Grok entre outras em amplo uso a partir da segunda metade da década de 2020, ocorrem por meio de *prompts*. *Prompts* são definidos como um conjunto de instruções indicadas pelos usuários para que a ferramenta operacionalize a demanda.

Portanto, a capacidade de fornecer instruções claras, precisas e orientadas ao objetivo é que permitem ao usuário obter bons resultados. Nesse processo, os verbos utilizados desempenham um papel fundamental para direcionar se a tarefa para a ferramenta será uma apresentação básica ou resgate de memória (exemplos de verbos seriam ‘enumere’, ‘descreva’, ‘relate’), compreensão de um cenário e aplicação (exemplos de verbos seriam ‘classifique’, ‘diferencie’ e ‘exemplifique’), ou uma maior demanda de raciocínio, por vezes, vinculada a processos inferenciais (exemplos de verbos são ‘proponha’, ‘elabore’ e ‘analise’).

Como exemplo, tomando como referência a nossa base de dados do Apêndice D, poderemos realizar o upload dos dados na ferramenta e solicitar para ‘relatar’ as variáveis categóricas e quantitativas como exemplo de apresentação básica.

Prompt simples

Considerando a base de dados enviada, relate as variáveis categóricas e quantitativas.

Output

Variáveis categóricas (qualitativas)

Cur (curso: Adm., Tur.)

Ins (tipo de instituição: Púb., Part.)

Sex (sexo)

E_Cv (estado civil: Solt., Cas.)

Variáveis quantitativas (numéricas)

Num (identificador)

Sem (semestre)

Ida (idade)

Ren (renda)

V1, V2, V3, V4 (itens/variáveis métricas)

ID (escala/indicador)

PR (escala/indicador)

Observação

Num é apenas identificador (não analítica).

Há valores inválidos/outliers (ex.: 444, 400, 66) e ausentes que precisam tratamento antes da análise.

Temos agora que definir se o resultado gerado está correto. Observe que é necessário compreender os conceitos de cada tipo de variável para ter a competência necessária que permite avaliar o *output*. No caso acima, a análise do *output* fica como exercício.

Ainda, poderíamos solicitar para ‘indicar’ quais respondentes possuem valores ausentes, quais são as variáveis que possuem tais valores ausentes, se há valores extremos (*outliers*), além de sugestões de tratamento para ambos os casos. Neste *prompt*, temos uma mesclagem de verbos que demandam a compreensão dos dados e proposições.

Prompt simples

indique quais respondentes possuem valores ausentes, quais são as variáveis que possuem valores ausentes, se há valores extremos (outliers), e sugestões de tratamento para ambos os casos.

Output

Respondentes com valores ausentes (*missing*)

4 (Sex)
 7 (Ida)
 8 (Ren)
 9 (Ins)
 10 (V3)
 11 (Sex)
 16 (Sex, E_Cv, V2, V4)
 21 (PR)
 34 (V3)
 54 (PR)
 105 (Sem, Ins, Ida, Ren, V1, V2, V4, PR)

Variáveis com valores ausentes

Sex
 Ins
 Sem
 Ida
 Ren
 V2
 V3
 V4
 PR

Valores extremos / inconsistentes (outliers evidentes)

V2 = 444 (respondente 4)
 V2 = 66 (respondente 34)
 Ida = 400 (respondente 111)
 Ida = 11 (respondente 79 – potencial erro dependendo do público)
 Ren = 40 e 50 (respondentes 45 e 79 – verificar escala)
 Ren = 11 (respondente 19 – fora do padrão predominante)

Sugestões de tratamento

1. Valores ausentes

Poucos casos por variável → imputação simples:

Númericas: média ou mediana (preferível mediana)

Catégoricas: moda

Caso com muitos missing (ex.: respondente 105) → excluir da base

Para análises mais robustas → imputação múltipla ou KNN

2. Outliers / erros de digitação

Valores claramente impossíveis (444, 66, 400):

→ corrigir (se houver fonte original) ou tratar como missing

Valores suspeitos (Ren alta, idade muito baixa):

→ validar consistência com contexto

Alternativas de tratamento:

Winsorização (limitar extremos)

Transformação (log) se mantidos

Exclusão pontual se distorcerem análises

3. Recomendação geral

Padronizar escala (ex.: V1-V4 parecem 1-7)

Criar regra: qualquer valor fora da escala → missing

Documentar todas as decisões de limpeza para reprodutibilidade

Novamente, temos um output que precisa ser avaliado para decidir qual o melhor caminho para o tratamento das anomalias. Deve-se notar que os *prompts* fornecidos são simples e diretos. Contudo, todos possuem pelo menos um elemento central que é o verbo. *Prompts* que não indicam o que a ferramenta deve realizar claramente não terão resultados satisfatórios. Além disso, melhores serão os resultados à medida que mais contexto, definições e direcionamentos forem apresentados, como discutido no tópico a seguir.

11.2 Engenharia de prompt

A engenharia de *prompt* envolve um conjunto de princípios que possibilitam obter resultados mais acurados das ferramentas de inteligência artificial. A acurácia aqui se refere ao quanto o *output* é adequado ao seu objetivo. *Prompts* imprecisos e sem contexto resultam em resultados genéricos e sem detalhes. Isto não significa dizer que o *prompt* deve ser complexo, a objetividade

e o direcionamento claro da ação é que são os elementos-chave. Deve-se, portanto, considerar critérios de necessidade e suficiência dos termos inseridos para que não haja uma elevação desnecessária do nível de complexidade do processamento.

Informações básicas que devem conter no seu prompt são as seguintes:

1. **Papel:** configurar o papel que o modelo de IA assumirá (por exemplo, você é um estatístico, especialista em análise de dados).
2. **Definição da tarefa:** indicar claramente o que a ferramenta precisa fazer, incluindo os verbos adequados (indique, compare, analise, proponha etc.).
3. **Propósito:** apresentar o objetivo final com a tarefa acima apontada (por exemplo, o objetivo da tarefa é ter um sumário final com análises estatísticas e recomendações gerenciais a partir das análises).
4. **Contexto:** elucidar a origem da base de dados e suas especificidades, como indicar o público que compõe a amostra, definições de variáveis e o que significam, como os dados foram coletados entre outros elementos.
5. **Estrutura de resposta:** direcionar o modelo de resposta esperado (por exemplo, as respostas devem estar em um arquivo word com, no máximo, 1500 palavras).

Com a estrutura acima, você obterá respostas mais detalhadas e com maior nível de acurácia. Tomemos como exemplo a necessidade de tratar os valores ausentes e extremos da nossa base de dados do Apêndice D. Podemos solicitar à ferramenta para elaborar um script em R com esses procedimentos. Vejamos a seguir.

Prompt estruturado

Você é um estatístico expert em operacionalizações com bases de dados e programação em R.

Você deve elaborar um script em R para tratar valores ausentes e extremos da base de dados que vou enviar. O script será processado no RStudio.

O objetivo final é ter uma nova base de dados com os valores ausentes e extremos tratados.

A base de dados está salva no meu computador. Portanto, será necessário importar para o RStudio. O nome da base é “dados”, está em xlsx e salva nesta pasta: “C:\Users\aluno\ Área de Trabalho\ Estatística”. Os dados são de estudantes de graduação. Os alunos são dos cursos de turismo e administração. As variáveis representam = Num: código de identificação; Cur: curso; Sem: semestre; Ins: tipos de instituição de ensino; Ida: idade; Sex: sexo; E_Cv: estado civil; Ren: renda; V1: foco; V2: perseverança; V3: avaliação dos docentes; V4: avaliação do suporte; ID: identificação com o curso; PR: recomendação do curso. Os dados são primários.

O script deve ser todo comentado e o mais simples possível e deve, no fim, permitir salvar a base de dados ajustada em xlsx.

Todos os pontos da nossa estrutura estão atendidos no prompt acima. Envie junto ao *script* a base de dados. A solicitação de comentários ao longo do script serve para você aprender o que cada função e argumento do script realiza. Dessa forma, poderá também validar se o caminho realizado pela ferramenta, de fato, atende ao seu objetivo. A seguir temos o output.

```
#=====
# TRATAMENTO DE DADOS: AUSENTES E OUTLIERS
#=====

#-----
# 1. IMPORTAR BASE DE DADOS
#-----

# Instalar pacote apenas se necessário
if(!require(openxlsx)) install.packages("openxlsx")

# Carregar pacote
library(openxlsx)

# Importar base
dados <- read.xlsx("C:/Users/aluno/Área de Trabalho/Estatística/
dados.xlsx")

# Visualizar estrutura
str(dados)
```

```

#-----
# 2. AJUSTES INICIAIS
#-----

# Renomear variável com ponto
names(dados)[names(dados) == "Ren."] <- "Ren"

#-----
# 3. IDENTIFICAR AUSENTES
#-----

# Contagem de NA por variável
colSums(is.na(dados))

#-----
# 4. TRATAR VALORES AUSENTES
#-----

# Função para substituir NA pela mediana (variáveis numéricas)
trata_na_num <- function(x){
  x[is.na(x)] <- median(x, na.rm = TRUE)
  return(x)
}

# Função para substituir NA pela moda (categóricas)
moda <- function(x){
  ux <- unique(x)
  ux[which.max(tabulate(match(x, ux)))]
}

trata_na_cat <- function(x){
  x[is.na(x)] <- moda(x)
  return(x)
}

```

```

# Aplicar tratamento
dados$Ida <- trata_na_num(dados$Ida)
dados$Ren <- trata_na_num(dados$Ren)
dados$V1 <- trata_na_num(dados$V1)
dados$V2 <- trata_na_num(dados$V2)
dados$V3 <- trata_na_num(dados$V3)
dados$V4 <- trata_na_num(dados$V4)
dados$ID <- trata_na_num(dados$ID)
dados$PR <- trata_na_num(dados$PR)

dados$Cur <- trata_na_cat(dados$Cur)
dados$Ins <- trata_na_cat(dados$Ins)
dados$Sex <- trata_na_cat(dados$Sex)
dados$E_Cv <- trata_na_cat(dados$E_Cv)

#-----
# 5. IDENTIFICAR OUTLIERS
#-----# Função para detectar limites pelo IQR
limites <- function(x){
  Q1 <- quantile(x, 0.25)
  Q3 <- quantile(x, 0.75)
  IQR <- Q3 - Q1
  LI <- Q1 - 1.5*IQR
  LS <- Q3 + 1.5*IQR
  return(list(LI=LI, LS=LS))
}

#-----
# 6. TRATAR OUTLIERS
#-----

# Função para substituir outliers pelos limites
trata_outliers <- function(x){
  lim <- limites(x)
  x[x < lim$LI] <- lim$LI
  x[x > lim$LS] <- lim$LS
  return(x)
}

```

```

# Aplicar nas variáveis numéricas
dados$Ida <- trata_outliers(dados$Ida)
dados$Ren <- trata_outliers(dados$Ren)
dados$V1 <- trata_outliers(dados$V1)
dados$V2 <- trata_outliers(dados$V2)
dados$V3 <- trata_outliers(dados$V3)
dados$V4 <- trata_outliers(dados$V4)
dados$ID <- trata_outliers(dados$ID)
dados$PR <- trata_outliers(dados$PR)

#-----
# 7. TRATAMENTO ESPECÍFICO (ERROS EVIDENTES)
#-----

# Corrigir valores absurdos manualmente (ex: 444, 66, 400, 11 em
# escala 1-7)

dados$V3[dados$V3 > 7] <- median(dados$V3[dados$V3 <= 7])
dados$V2[dados$V2 > 7] <- median(dados$V2[dados$V2 <= 7])
dados$V1[dados$V1 > 7] <- median(dados$V1[dados$V1 <= 7])
dados$Ida[dados$Ida > 100] <- median(dados$Ida[dados$Ida <= 100])

#-----
# 8. BASE FINAL
#-----

# Verificar novamente
summary(dados)

#-----
# 9. SALVAR BASE TRATADA
#-----

write.xlsx(dados, "C:/Users/aluno/Área de Trabalho/Estatística/
dados_tratados.xlsx")

#=====
# FIM
#=====

```

Como podemos observar no *output*, as especificações do *prompt* permitiram ter um *script* totalmente alinhado ao objetivo e contexto de aplicação. Fica como exercício a solicitação do mesmo procedimento utilizando a linguagem Python.

11.3 Outras considerações

Vimos algumas exemplificações de limpeza e pré-processamento do banco de dados. No entanto, as ferramentas de IA podem auxiliar em uma série de outras etapas, como na sugestão de relações entre as variáveis para estabelecer hipóteses, apoio em transformações de variáveis, sugestões de gráficos adequados para cada tipo de variável, explicações de ferramentas estatísticas, validação e interpretação de resultados, correção de potenciais erros em scripts, apoio na elaboração de relatórios estatísticos, sugestões de ações gerenciais a partir dos resultados, integração de mais de uma base de dados (até em extensões diferentes) etc. Em todas essas etapas, a validação humana é indispensável.

11.4. Resumo

Neste capítulo, introduzimos a utilização de ferramenta de inteligência artificial para análise exploratória de dados. Os destaques foram os seguintes:

- *Prompts* são um conjunto de instruções para a ferramenta de IA seguir.
- Os verbos são fundamentais para que a ferramenta de IA saiba o que especificamente deve realizar e devem ser alinhados ao seu propósito final.
- *Prompts* devem ser minimamente estruturados, indicando papel, definição da tarefa, propósito, contexto, estrutura de resposta.
- Deve-se utilizar de objetividade para não tornar o prompt complexo de forma desnecessária.
- O uso de ferramentas de IA não dispensa o conhecimento e supervisão humana. O papel delas é acelerar processos de análises e resultados.

Exercícios

Em todos os exercícios que seguem, utilize os princípios de engenharia de prompt apresentados.

1. Considerando os dados do Apêndice D, solicite à IA sugestões de técnicas de análise exploratória de dados que poderiam ser utilizadas na base de dados.
2. Avalie as sugestões e estabelece aquelas que são coerentes com o conteúdo apresentado neste manuscrito.
3. Solicite à IA scripts em R, Python e o passo a passo para realizar no JASP as sugestões indicadas.
4. Avalie os comandos e implemente-os.

Apêndice A

INSTALAÇÃO DO R E RSTUDIO

Etapa 1 – Instalar o software R (base para o RStudio)

1. O primeiro passo a instalação do software R é acessar o endereço eletrônico a seguir: <https://cran.fiocruz.br/>
2. Ao acessar o site, selecione o download que seja compatível com o sistema operacional da sua máquina.
3. Para instalação no Windows, você será direcionado para uma nova página. Nela, clique em “base” para instalar o software pela primeira vez.
4. Na sequência, clique em Download R-4.3.2 for Windows para iniciar o download.
5. Ao clicar no executável baixado, uma janela de permissão será aberta; clique em “permitir” para continuar.
6. Em seguida, uma aba será exibida para que você selecione o idioma de sua preferência. Escolha o idioma desejado e clique em “OK”.
7. Você será solicitado a aceitar a licença de uso do software exibida na tela. Após ler as informações e concordar com os termos, clique em “Avançar”.
8. Escolha a pasta do seu computador para instalar o R. Recomenda-se aceitar a configuração padrão sugerida pelo instalador do software. Na sequência, clique em “Avançar”.
9. Selecione os componentes a serem instalados. Sugerimos a seleção de todos os componentes para instalação.
10. No passo seguinte, selecione uma das opções de inicialização. Recomenda-se selecionar a opção “Não (aceitar padrão)”.
11. A etapa seguinte consiste em escolher o local de atalho para o R. Se preferir que o atalho seja criado apenas no pasta sugerida “Menu Iniciar”, apenas clique em “Avançar”.

12. Por fim, deverá indicar as tarefas adicionais que deseja executar durante a instalação do R. Sugerimos aceitar as opções selecionadas por padrão pelo instalador e clicar em “Avançar”.
13. Para finalizar, clique em “Concluir”.

Etapa 2 – Instalar o software RStudio (recomendado para novos usuários)

1. O primeiro passo para iniciar a instalação do software RStudio é acessar o seguinte endereço eletrônico: <https://posit.co/download/rstudio-desktop/>
2. Ao acessar o site, selecione download RStudio desktop no item 2: Install RStudio.
3. Para iniciar a instalação, clique com o botão direito do mouse ou touchpad no aplicativo baixado e selecione executar como administrador.
4. Na sequência, clique em “próximo”.
5. Na próxima janela, escolha o local de armazenamento do software. Geralmente, os programas ficam instalados no Disco Local C:\. Assim, você pode apenas clicar em “próximo” novamente.
6. Na janela seguinte, você pode clicar em “instalar”.
7. A instalação iniciará, conforme a imagem abaixo. Você precisa apenas aguardar o processo ser completado.
8. Para finalizar, clique em “Concluir”.

Apêndice B

INSTALAÇÃO DO JASP

1. O primeiro passo é acessar o seguinte link: <http://jasp-stats.org/> e clicar em “Download”, que fica na parte superior.
2. Na sequência, clique em “instalar” na versão que for compatível com o sistema operacional da sua máquina. O download iniciará automaticamente.
3. Após baixado, dê dois clique para abrir. Aceite os termos e clique em instalar.
4. Por fim, clique em “concluir”.

Apêndice C

INSTALAÇÃO DO VSCODE (PYTHON)

1. Acesse o site <https://www.python.org/> e clique em downloads e, em seguida, download Python.
2. Abra o arquivo baixado, selecione o local de instalação e clique em “Install now”. Isto feito, o Python será baixado.
3. Em seguida, acesse o site <https://code.visualstudio.com/> e clique em “download”.
4. Feito o download, abra o arquivo, aceite os termos e clique em próximo.
5. Na sequência, selecione o local de instalação, atalho e tarefas adicionais (que desejar) e clique em instalar.
6. Isto feito, clique em concluir.

Apêndice D

DADOS DA PESQUISA
COM ESTUDANTES
DE ADMINISTRAÇÃO
E TURISMO

Num	Cur	Sem	Ins	Ida	Sex	E_Cv	Ren.	V1	V2	V3	V4	ID	PR
1	Adm.	1	Púb.	16	Masc.	Solt.	2	7	7	7	6	7	5
2	Adm.	3	Part.	36	Masc.	Solt.	2	6	4	6	7	7	7
3	Adm.	3	Púb.	71	Masc.	Solt.	2	5	6	7	6	7	5
4	Adm.	1	Part.	36		Solt.	3	7	7	444	6	7	5
5	Adm.	5	Púb.	24	Fem.	Solt.	3	6	7	7	6	6	2
6	Adm.	1	Púb.	39	Fem.	Solt.	3	6	6	7	6	6	1
7	Adm.	6	Part.		Fem.	Solt.	3	5	6	3	5	6	5
8	Adm.	4	Part.	25	Fem.	Solt.	3		5	6	6	6	5
9	Adm.	2		21	Fem.	Solt.	3	7	7	7	7	6	6
10	Adm.	1	Púb.	39	Fem.	Solt.	5	6	6	6		6	3
11	Adm.	1	Púb.	28		Solt.	5	7	7	7	7	6	4
12	Adm.	1	Part.	36	Fem.	Solt.	6	6	7	7	5	6	7
13	Adm.	4	Part.	18	Fem.	Solt.	7	6	7	6	5	6	1
14	Adm.	7	Púb.	31	Masc.	Cas.	8	5	6	6	4	6	2
15	Adm.	3	Púb.	32	Fem.	Solt.	8	5	6	7	6	6	3
16	Adm.	7	Púb.	32			8	6		7	6		7
17	Adm.	4	Part.	26	Fem.	Solt.	9	7	7	7	5	6	6
18	Adm.	4	Part.	34	Masc.	Solt.	12	6	6	6	6	6	1
19	Adm.	8	Púb.	39	Fem.	Solt.	1	11	6	6	5	6	6
20	Adm.	3	Part.	35	Fem.	Solt.	2	6	6	7	6	6	1
21	Adm.	1	Part.	30	Masc.	Solt.	2	6	6	6	4	6	
22	Adm.	6	Part.	28	Masc.	Solt.	3	4	5	6	6	6	1
23	Adm.	4	Part.	38	Masc.	Solt.	4	6	6	7	5	6	1
24	Adm.	1	Púb.	23	Masc.	Solt.	5	6	6	6	5	6	5
25	Adm.	2	Púb.	20	Masc.	Solt.	8	5	5	6	7	6	4

26	Adm.	4	Part.	25	Masc.	Solt.	8	5	7	6	6	5	2
27	Adm.	3	Part.	39	Fem.	Solt.	10	6	6	7	5	5	2
28	Adm.	1	Part.	31	Fem.	Solt.	1	5	5	6	5	5	3
29	Adm.	8	Púb.	22	Masc.	Solt.	2	4	3	5	5	5	2
30	Adm.	5	Púb.	20	Masc.	Solt.	3	5	5	5	5	5	4
31	Adm.	5	Púb.	27	Masc.	Solt.	4	5	5	6	3	5	5
32	Adm.	4	Part.	21	Fem.	Cas.	4	6	6	6	5	5	3
33	Adm.	1	Púb.	28	Masc.	Solt.	6	5	6	6	5	5	7
34	Adm.	7	Púb.	28	Fem.	Solt.	7	6	66		6	5	5
35	Adm.	7	Part.	30	Masc.	Solt.	8	5	5	6	6	5	2
36	Adm.	6	Part.	19	Masc.	Cas.	10	4	5	5	4	5	2
37	Adm.	7	Part.	31	Masc.	Solt.	3	5	5	6	5	5	6
38	Adm.	7	Part.	23	Fem.	Cas.	3	5	5	5	5	5	5
39	Adm.	3	Part.	23	Fem.	Solt.	4	6	4	6	5	5	6
40	Adm.	4	Part.	67	Fem.	Solt.	4	6	3	6	6	5	2
41	Adm.	1	Púb.	40	Fem.	Solt.	7	5	6	6	5	5	5
42	Adm.	3	Púb.	37	Masc.	Cas.	8	4	5	5	4	5	1
43	Adm.	5	Part.	29	Masc.	Solt.	8	6	5	5	3	5	7
44	Adm.	3	Part.	32	Fem.	Solt.	9	4	5	6	5	5	5
45	Adm.	8	Part.	31	Masc.	Solt.	40	6	6	6	6	5	4
46	Adm.	1	Part.	34	Masc.	Solt.	1	5	5	6	5	5	3
47	Adm.	6	Part.	38	Fem.	Solt.	2	5	5	5	5	5	1
48	Adm.	6	Part.	21	Fem.	Solt.	3	4	4	4	4	5	1
49	Adm.	1	Part.	59	Masc.	Solt.	3	6	6	6	4	4	4
50	Adm.	7	Part.	24	Masc.	Solt.	4	4	5	4	4	4	6
51	Adm.	7	Part.	35	Fem.	Cas.	4	5	5	5	5	4	2
52	Adm.	6	Púb.	25	Fem.	Solt.	5	5	5	7	5	4	6
53	Adm.	4	Part.	23	Masc.	Solt.	6	5	4	5	5	4	4
54	Adm.	5	Púb.	36	Masc.	Solt.	6	5	4	4	6	4	
55	Adm.	1	Part.	28	Masc.	Solt.	6	4	4	4	4	4	1
56	Adm.	5	Púb.	38	Fem.	Solt.	7	7	6	7	6	4	7
57	Adm.	3	Púb.	31	Masc.	Solt.	7	5	6	6	5	4	2
58	Adm.	7	Part.	36	Masc.	Solt.	8	6	5	6	6	4	5
59	Adm.	3	Part.	35	Fem.	Solt.	8	5	5	5	5	3	6
60	Adm.	7	Part.	28	Fem.	Solt.	9	6	7	7	5	2	2
61	Tur.	1	Part.	34	Fem.	Solt.	9	6	6	6	6	6	3
62	Tur.	6	Part.	36	Fem.	Solt.	1	5	7	7	6	6	6

63	Tur.	3		22	Masc.	Cas.	1	7	6	7	7	6	1
64	Tur.	1	Púb.	31	Masc.	Solt.	1	4	6	6	6	6	4
65	Tur.	7	Part.	21	Fem.	Solt.	3	5	4	5	5	6	4
66	Tur.	6		30	Fem.	Solt.	3	6	6	7	6	6	5
67	Tur.	7	Púb.	31	Masc.	Solt.	3	7	7	7	6	6	7
68	Tur.	8	Part.	29	Masc.	Cas.	4	6	6	6	7	6	1
69	Tur.	4	Púb.	29	Fem.	Solt.	4	6	6	6	6	6	3
70	Tur.	1	Part.	28	Masc.	Solt.	5	5	6	6	5	6	1
71	Tur.	1	Part.	30	Masc.	Solt.	7	6	7	6	5	5	1
72	Tur.	7	Púb.	19	Fem.	Cas.	9	6	7	7	6	5	2
73	Tur.	5	Púb.	40	Fem.	Solt.	1	6	7	6	6	5	5
74	Tur.	7	Púb.	18	Fem.	Solt.	1	7	6	7	7	5	6
75	Tur.	7	Púb.	26	Fem.	Solt.	3	7	6	6	6	5	1
76	Tur.	4	Part.	39	Fem.	Solt.	9	7	6	6	6	5	6
77	Tur.	7	Púb.	25	Fem.	Solt.	9	7	6	6	5	5	6
78	Tur.	3	Púb.	25	Fem.	Solt.	12	6	6	6	5	5	4
79	Tur.	5	Púb.	11	Fem.	Solt.	50	5	6	6	6	5	1
80	Tur.	7	Part.	30	Masc.	Solt.	1	7	6	6	6	5	6
81	Tur.	3	Púb.	19	Fem.	Solt.	1	6	6	5	5	5	3
82	Tur.	3	Part.	34	Fem.	Cas.	3	5	5	6	5	5	1
83	Tur.	3	Púb.	39	Fem.	Solt.	3	7	6	6	6	5	2
84	Tur.	4	Part.	36	Fem.	Solt.	3	6	5	5	5	5	6
85	Tur.	7	Púb.	39	Fem.	Solt.	3	6	6	7	6	5	7
86	Tur.	7	Part.	24	Fem.	Solt.	4	6	6	6	6	5	5
87	Tur.	7	Púb.	34	Fem.	Solt.	7	6	6	6	5	5	3
88	Tur.	1	Part.	33	Fem.	Solt.	7	4	6	6	5	5	2
89	Tur.	7	Púb.	27	Fem.	Cas.	9	6	5	5	5	5	6
90	Tur.	7	Part.	30	Fem.	Solt.	1	6	5	5	5	5	3
91	Tur.	6	Part.	28	Fem.	Solt.	3	5	5	6	6	5	7
92	Tur.	7	Part.	21	Masc.	Solt.	3	6	6	6	6	5	1
93	Tur.	1	Púb.	30	Fem.	Solt.	4	6	6	6	4	5	5
94	Tur.	2	Part.	21	Fem.	Cas.	4	5	6	6	6	5	2
95	Tur.	3	Púb.	34	Masc.	Solt.	4	5	6	5	5	5	7
96	Tur.	6	Part.	24	Fem.	Solt.	6	6	5	6	6	5	2
97	Tur.	1	Part.	36	Fem.	Solt.	7	5	5	5	5	5	5
98	Tur.	1	Púb.	33	Fem.	Solt.	8	6	6	6	5	5	2
99	Tur.	4	Púb.	22	Fem.	Solt.	8	5	5	5	5	5	3

100	Tur.	2	Part.	37	Masc.	Solt.	4	3	4	5	5	5	2
101	Tur.	7	Part.	40	Fem.	Cas.	4	4	4	4	3	4	1
102	Tur.	3	Púb.	25	Fem.	Solt.	4	4	7	7	6	4	3
103	Tur.	8	Púb.	23	Fem.	Solt.	5	6	5	5	5	4	7
104	Tur.	5	Part.	18	Fem.	Solt.	5	5	4	5	5	4	6
105	Tur.				Fem.	Solt.	5			6		4	
106	Tur.	1	Part.	22	Fem.	Solt.	6	5	5	6	6	4	3
107	Tur.	5	Part.	19	Fem.	Cas.	7	5	3	6	4	4	2
108	Tur.	2	Púb.	25	Masc.	Cas.	8	2	3	2	5	4	2
109	Tur.	5	Part.	27	Masc.	Solt.	8	6	5	5	4	3	1
110	Tur.	5	Part.	32	Fem.	Solt.	8	4	3	5	4	3	1
111	Tur.	7	Púb.	400	Fem.	Solt.	8	4	3	3	3	3	4
112	Tur.	4	Part.	28	Fem.	Solt.	8	4	3	4	4	2	3
113	Tur.	6	Púb.	36	Masc.	Solt.	9	5	6	6	6	1	1

REFERÊNCIAS

- Pereira, B. B. (1997). Estatística: a tecnologia da ciência. **Boletim da Associação Brasileira de Estatística**, ano XIII, n. 37, 2º quadrimestre, p. 27-35.
- Stevens, S. S. (1946). On the theory of scales of measurement. **Science**, 103(2684), 677-680.
- Hair, J. F. Jr., Black, W. C., Babin, B. J., & Anderson, R. E. (2005). **Análise multivariada de dados**. Porto Alegre: Bookman.
- Lattin, J., Carroll, J. D., & Green, P. E. (2011). **Análise de dados multivariados**. São Paulo: Cengage.
- Manly, B. J. F. (2008). **Métodos estatísticos multivariados**: uma introdução. Porto Alegre: Bookmann.
- Ferreira, D. F. (2008). **Estatística multivariada**. Lavras: Ed. UFLA.
- Mingoti, S. A. (2005). **Análise de dados através de métodos de estatística multivariada**: uma abordagem aplicada. Belo. Horizonte: Editora UFMG.
- Toledo, G. L., & Ovalle, I. (1995). **Estatística básica**. 2. ed. São Paulo: Atlas.

SOBRE OS AUTORES

Autor: Francisco José da Costa

Minicurrículo: Possui graduação em Administração (UECE, 2000) e em Estatística (UFPB, 2014), mestrado em Administração (UECE, 2003) e doutorado em Administração de Empresas (FGV-SP, 2007). É professor titular da Universidade Federal da Paraíba, com atuação na graduação e na pós-graduação em Administração. Exerceu funções de gestão acadêmica como Coordenador do Programa de Pós-Graduação em Administração (PPGA/UFPB) e Assessor para Análise de Dados e Avaliação Institucional na Pró-Reitoria de Graduação (PRG) e na Pró-Reitoria de Planejamento e Desenvolvimento Institucional (PROPLAN). É autor de mais de 50 artigos e três livros, além de ter orientado mais de 40 dissertações e teses. Desenvolve atividades de ensino e pesquisa nas áreas de Métodos Quantitativos Aplicados, Gestão e Análise de Desempenho e Educação em Administração.

Vínculo institucional: Universidade Federal da Paraíba

Autora: Patrícia de Oliveira Campos

Minicurrículo: Doutoranda em Administração pela Universidade Federal da Paraíba (UFPB) com ênfase em Gestão Analítica e Teoria da Mensuração. É membro do Núcleo de Análise de Dados e Desempenho ($A^2\Delta$ /UFPB) e do Laboratório de Estudos em Modelagem Aplicada (LEMA/UFPB). É autora de mais de 25 artigos publicados em periódicos nacionais e internacionais com manipulação estatística de dados.

Vínculo institucional: Universidade Federal da Paraíba



Título ANÁLISE DE DADOS: PROCEDIMENTOS EXPLORATÓRIOS

Autores Francisco José da Costa
Patrícia de Oliveira Campos

Projeto de diagramação e Capa Ana Gabriella Carvalho

Imagem da capa Adaptada de Magnific.com

Formato/ Tamanho e-book (PDF – 16x25,5 cm)/ 12MB

Tipografia Cambria e Google Sans Code

Número de páginas 286

