Paulo Lima Junior

MÉTODOS QUANTITATIVOS DA PESQUISA EM EDUCAÇÃO:

uma introdução baseada em linguagem R

EDITORA
UnB

Paulo Lima Junior

MÉTODOS QUANTITATIVOS DA PESQUISA EM EDUCAÇÃO:

uma introdução baseada em linguagem R

EDITORA
UnB

Coordenação de produção editorial
Preparação e revisão

Equipe editorial

Marília Carolina de Moraes Florindo
Jeane Pedrozo

Editora Universidade de Brasília
Centro de Vivência, Bloco A – 2ª etapa, 1º andar
Campus Darcy Ribeiro, Asa Norte, Brasília/DF
CEP: 70910-900
Site: www.editora.unb.br
E-mail: contatoeditora@unb.br

O autor agradece o apoio do Conselho Nacional de Desenvolvimento Científico e Tecnológico – CNPq

Dados Internacionais de Catalogação na Publicação (CIP)
(Biblioteca Central da Universidade de Brasília – BCE/UnB)

L732m Lima Junior, Paulo.
Métodos quantitativos da pesquisa em educação [recurso eletrônico] : uma introdução baseada em linguagem R / Paulo Lima Junior. - Brasília : Editora Universidade de Brasília, 2023.
374 p.

Formato PDF.
Inclui bibliografia.
ISBN 978-65-5846-086-2 (e-book).

1. Estatística educacional. 2. R (Linguagem de programação de computador). 3. Pesquisa educacional. I. Título.

CDU 311.21:37

Heloiza Faustino dos Santos – Bibliotecária – CRB 1/1913

LISTA DE DIAGRAMAS

Diagrama 5.1: Representação da matriz de correlações..... **111**
Diagrama 5.2: Representação vetorial de variáveis correlacionadas **113**
Diagrama 6.1: Mosaico da tabela de contingência (escolaridade do pai contra escolaridade da mãe).....**138**
Diagrama 7.1: Representação de um modelo com preditores associados **150**
Diagrama 7.2: Poder explicativo exclusivo e comum das variáveis do modelo.....**152**
Diagrama 7.3: Repartição do poder explicativo entre variáveis associadas **154**
Diagrama 9.1: Posição da regressão logística com relação às demais versões do modelo linear generalizado **176**
Diagrama 10.5: Representação gráfica da matriz de correlações.....**209**
Diagrama 12.1: Exemplo de dendograma com dados fictícios **257**
Diagrama 12.2: Representação da matriz de dissimilaridade.....**260**
Diagrama 12.3: Dendograma de 50 estudantes obtidos aleatoriamente do Enem 2018 .....**263**
Diagrama 14.1: Decomposição das variáveis manifestas em fatores comuns e únicos.....**300**
Diagrama 14.2: Carga fatorial dos itens de motivação.....**318**
Diagrama 14.3: Relação entre fator latente e variáveis manifestas..**319**

Diagrama 14.4: Análise fatorial exploratória empregada como etapa preliminar de uma análise posterior **322**
Diagrama 14.5: Análise fatorial confirmatória integrada ao teste de modelos preditivos e associativos envolvendo fatores latentes **323**
Diagrama 15.1: Diagrama de caminho do modelo fatorial das motivações para o curso de Física **332**
Diagrama 15.2: Diagrama de caminho do modelo fatorial das motivações para o curso de Física (editado) **335**
Diagrama 15.3: Exemplo de modelo fatorial hierárquico **341**
Diagrama 15.4: Diagrama de caminho do modelo de Tinto **345**
Diagrama 15.5: Diagrama de caminho do modelo de Tinto (somente relações estruturais) **346**

LISTA DE FIGURAS

Figura 15.1: Representação de uma relação de moderação **343**
Figura 15.2: Representação de uma relação de mediação.....**343**
Figura 15.3: Representação simplificada do modelo de Tinto (1987, 2012) **344**

LISTA DE GRÁFICOS

Gráfico 3.1: Densidade de probabilidade de uma distribuição não normal **65**
Gráfico 3.2: Histograma de dados normalmente distribuídos.....**66**
Gráfico 3.3: Intervalos normais de cobertura.....**66**
Gráfico 3.4: Histograma de uma distribuição uniforme **70**
Gráfico 3.5: Histograma da soma das variáveis.....**71**
Gráfico 4.1: Distribuição-t com *gl = 30* (região significativa em destaque) **85**
Gráfico 4.2: Histograma das notas de ciências da natureza.....**89**
Gráfico 4.3: Gráfico Q-Q para distribuição normal **90**
Gráfico 4.4: Desvios até a média geral **96**
Gráfico 4.5: Desvios até a média do grupo.....**96**
Gráfico 4.6: Médias de NotaCN por tipo de escola (intervalos de confiança a 95%).....**100**
Gráficos 4.7-4.10: Diagramas diagnósticos do modelo **103**
Gráfico 5.1: Dispersão de NotaMT contra NotaCN (intervalos de confiança a 95%).....**107**
Gráficos 5.2-5.5: Diagramas de dispersão com diversas correlações**110**
Gráfico 6.6: Menor correlação significativa por tamanho da amostra (com $p<0,05$).....**114**
Gráfico 5.7: Representação dos resíduos em uma regressão linear ..**117**

Gráfico 5.8: Regressão de ciências da natureza contra matemática (Enem 2018)..... **118**
Gráficos 5.9-5.10: Comparando a qualidade do ajuste em dois casos **120**
Gráficos 5.11-5.14: Diagramas diagnósticos do modelo **122**
Gráficos 5.15-5.16: Outliers centrais e alavancados.....**125**
Gráfico 6.1: Distribuição chi-quadrado com *gl = 9* (região significativa em destaque) **135**
Gráfico 7.1: Interação de EnsMed e Cor sobre NotaCN (intervalos de confiança a 95%).....**156**
Gráfico 8.1: AIC para todos os modelos possíveis **168**
Gráfico 9.1: Função de resposta `logit` **178**
Gráfico 10.1: Exemplo de diagrama de dispersão entre dois indicadores latentes **205**
Gráfico 10.2: Exemplo de como a correlação policórica é obtida....**206**
Gráfico 10.3: *Scree plot* dos itens de discriminação escolar.....**212**
Gráfico 10.4: Cargas das variáveis nas componentes principais **216**
Gráfico 11.1: Posição de algumas capitais brasileiras **224**
Gráfico 71.2: Mapa das notas em ciências da natureza e matemática **225**
Gráfico 11.3: Mapa simétrico das escolaridades do pai e da mãe (Enem 2018).....**233**
Gráfico 11.4: Mapa simétrico da renda, escolaridade e ocupação do pai e da mãe (Enem 2018) **238**
Gráfico 11.5: Mapa simétrico com uma variável suplementar **239**
Gráfico 12.1: Representação hipotética de dois clusters no espaço (ou seja, dois agrupamentos de instâncias individuais).....**249**
Gráfico 12.2: Mapa das variáveis na análise de componentes principais **254**
Gráfico 12.3: Mapa dos indivíduos na análise de componentes principais **255**
Gráfico 12.4: Gráfico de barras das alturas em que ocorrem as junções **258**

Gráficos 12.5-12.8: Representação dos métodos das distâncias simples, completas, médias e método de Ward **261**
Gráfico 12.9: Alturas dos agrupamentos – dados Enem **262**
Gráfico 12.10: Análise de componentes principais dos indivíduos .. **264**
Gráficos 12.11-12.14: Representação dos clusters produzidos pelo método de partição k-médias **265**
Gráfico 13.1: Testes educacionais aplicados sucessivamente **281**
Gráfico 14.1: Análise paralela dos itens de motivação **307**
Gráficos 14.2-14.3: Exemplo de rotação dos fatores extraídos **310**
Gráfico 14.4: Diagrama de dispersão dos escores dos fatores latentes **316**

LISTA DE QUADROS

Quadro 1.1: Identificando as lições mais importantes em vista dos seus propósitos **28**
Quadro 1.2: Sobre a elaboração e validação de testes educacionais .. **31**
Quadro III.1: Versões do modelo linear clássico **141**
Quadro 8.1: Síntese dos critérios para comparação de modelos **166**
Quadro 10.1: Interpretação geométrica das correlações entre variáveis **195**
Quadro 10.2: Comparando as funções mais populares para análise de componentes principais **220**
Quadro 11.1: Informações disponíveis no sumário da análise de correspondência **232**
Quadro 12.1: Comparando análise de componentes principais e de correspondência múltipla com respeito às variáveis ativas (i.e. não suplementares) que podem ser incluídas **242**
Quadro 14.1: Comparando análise fatorial exploratória e de componentes principais **303**
Quadro 14.1: Comparando análise fatorial exploratória e de componentes principais **304**
Quadro 15.1: Algumas diretivas da função semPaths **333**
Quadro 15.2: Sintaxe da função `sem` **344**

LISTA DE TABELAS

Tabela 10.1: Tabulação cruzada das respostas dadas a dois itens tipo-Likert **206**
Tabela 10.2: Relação entre componentes principais e variáveis indicadoras **215**
Tabela 11.1: Exemplo de uma matriz de Burt **234**

SUMÁRIO

APRESENTAÇÃO..... **21**

PARTE I. FUNDAMENTOS

LIÇÃO 1. INTRODUÇÃO À ANÁLISE ESTATÍSTICA **26**

Escolhendo o método de análise..... **26**
Validando testes educacionais..... **28**
Design experimental e causalidade **31**
Incerteza e intervalos de confiança..... **34**
Significância estatística..... **36**
A "objetividade" da pesquisa quantitativa..... **37**
Algumas palavras sobre o programa R..... **38**
Considerações finais **40**
Atividades propostas..... **41**

LIÇÃO 2. *DATAFRAMES* **42**

Ferramentas de visualização..... **42**
Gerando fatores (variáveis categóricas)..... **44**

Gerenciando nomes das variáveis **45**
Publicando partes do *dataframe* **49**
Ordenando o *dataframe* **51**
Filtrando o *dataframe* **54**
Importar e salvar **55**
Revisando a lição **56**
Atividades propostas **57**

LIÇÃO 3. MÉDIA E DISPERSÃO **58**
Calculando médias aritméticas **58**
Calculando variâncias **61**
Calculando desvios padrão **63**
Distribuição e densidade de probabilidade **64**
Intervalos normais de cobertura **66**
Testes de significância estatística **68**
Teorema do limite central **68**
Revisando a lição **72**
Atividades propostas **72**

PARTE II. ANÁLISE BIVARIADA

LIÇÃO 4. COMPARANDO MÉDIAS **76**
Algumas palavras sobre o ***design*** experimental **76**
Importando os dados da análise **80**
Comparando a média em dois grupos **80**
Identificando diferenças casuais **81**
Teste t de Student **84**

As premissas do teste t.....**86**
A premissa de normalidade.....**88**
A premissa de homoscedasticidade **90**
Teste de Wilcoxon.....**92**
Análise de variância.....**93**
Decompondo as somas quadráticas **94**
Das somas quadráticas ao teste.....**95**
Um exemplo de ANOVA **98**
Interpretando as diferenças **99**
Testando as premissas do modelo.....**102**
Revisando a lição.....**104**
Atividades propostas.....**105**

LIÇÃO 5. CORRELAÇÃO E REGRESSÃO **106**
Definindo covariância e correlação **107**
Gráficos de variáveis correlacionadas **109**
Matriz de correlações.....**110**
Interpretação geométrica **112**
Testes de significância estatística.....**113**
Regressão linear.....**115**
Qualidade do ajuste.....**119**
Verificação do modelo **121**
Interpretando a alavancagem **124**
Uma visão panorâmica **125**
Revisando a lição.....**126**
Atividades propostas.....**126**

LIÇÃO 6. TABELAS DE CONTINGÊNCIA **127**
As quantidades observadas **130**
A distância do observado ao esperado **131**
A tabela de contingência em um comando só **135**
O diagrama mosaico **137**
Revisando a lição **139**
Atividade proposta **139**

PARTE III. MODELAGEM ESTATÍSTICA

LIÇÃO 7. MODELO LINEAR CLÁSSICO **145**
Definição, extração e interpretação do modelo **148**
Variáveis associadas e qualidade do ajuste **150**
Interações entre variáveis explicativas **154**
Sintaxe do modelo **158**
Revisando a lição **159**
Atividades propostas **160**

LIÇÃO 8. MODELOS CONCORRENTES **161**
Métricas de avaliação do ajuste **163**
Calculando todas as métricas **167**
Verossimilhança relativa **169**
Seleção ***stepwise*** de variáveis **170**
Testando as hipóteses do modelo **172**
Revisando a lição **173**
Atividade proposta **173**

LIÇÃO 9. REGRESSÃO LOGÍSTICA **174**
A função `logit` **177**
Interpretando os coeficientes da regressão logística **179**
Um exemplo prático **180**
Seleção ***stepwise*** da regressão logística **186**
Revisando a lição **188**
Atividade proposta **189**

PARTE IV. ANÁLISE EXPLORATÓRIA MULTIVARIADA

LIÇÃO 10. ANÁLISE DE COMPONENTES PRINCIPAIS **193**
O que é uma redução dimensional? **194**
Exemplos da vida e da literatura **196**
Definindo as componentes principais **198**
Um problema de diagonalização **200**
A estrutura do preconceito escolar **202**
Correlações tetracóricas e policóricas **205**
Obtendo a matriz das correlações entre itens tipo-Likert **208**
Quantas dimensões devemos reter? **210**
Investigando os autovetores **214**
Sobre as bibliotecas disponíveis **217**
Revisitando a lição **221**
Atividade proposta **222**

LIÇÃO 11. ANÁLISE DE CORRESPONDÊNCIA **223**
A matriz dos resíduos padronizados **227**
Análise de correspondência simples **229**

Análise de correspondência múltipla e conjunta **233**
Variáveis suplementares..... **239**
A qualidade do mapa **240**
Comparando PCA e MCA **241**
A análise de correspondência na sociologia da educação de Bourdieu **242**
Revisando a lição..... **246**

LIÇÃO 12. ANÁLISE DE *CLUSTER* (OU AGRUPAMENTO)..... **247**
Exemplos da literatura **249**
Realizando uma análise de componentes principais..... **253**
Interpretando um dendograma..... **256**
A matriz de dissimilaridade **258**
Realizando a análise de agrupamento hierárquico..... **260**
Algoritmo de partição (K-means)..... **264**
Uma visão geral sobre a análise exploratória multivariada **266**
Revisando a lição..... **267**

PARTE V. ANÁLISES FATORIAIS

LIÇÃO 13. TESTES EDUCACIONAIS..... **272**
Relação com a realidade **272**
Estrutura e fidedignidade **275**
Validade e validação **276**
A teoria clássica de testes **277**
A falácia da atenuação das diferenças **279**
A elaboração de testes educacionais..... **285**
Alfa de Cronbach..... **288**

Restrições do coeficiente alfa **293**
Revisando a lição **294**

LIÇÃO 14. ANÁLISE FATORIAL EXPLORATÓRIA **296**
Definições do modelo da análise fatorial **300**
O modelo matemático da análise fatorial exploratória **301**
Realizando a análise fatorial exploratória **304**
Escolhendo a matriz de covariâncias **305**
Determinando a quantidade de fatores **306**
Métodos de extração **308**
Métodos de rotação **309**
Interpretando o output da análise fatorial exploratória **311**
Extraindo os escores fatoriais **315**
Visualização e interpretação das cargas fatoriais **317**
Uma visão geral sobre a análise fatorial exploratória **318**
Revisando a lição **324**

LIÇÃO 15. ANÁLISE FATORIAL CONFIRMATÓRIA E MODELAGEM *SEM* **325**
O modelo matemático da análise fatorial confirmatória **328**
Um modelo fatorial simples **330**
Edição do diagrama de caminho **333**
Significância estatística dos parâmetros **335**
Avaliação da plausibilidade do modelo **337**
Modelo fatorial hierárquico **341**
Relações de Moderação e Mediação **342**
Modelagem de equações estruturais **344**
Considerações finais **349**

REFERÊNCIAS **351**

APÊNDICE. ÁLGEBRA MATRICIAL **363**

Matrizes especiais **364**

Operações com matrizes **366**

Autovalor e autovetor **370**

Decomposições espectrais **372**

APRESENTAÇÃO

Os métodos estatísticos representam uma ferramenta tão importante quanto negligenciada pela pesquisa educacional brasileira. Ainda que análises qualitativas sejam fundamentais para a produção de conhecimento profundo sobre as experiências concretas de ensinar e aprender, indicadores educacionais quantificáveis geralmente apresentam uma força singular para influenciar a tomada de decisão de gestores, professores e cidadãos.

Esta obra traz um conjunto de textos de apoio para investigadores que desejem aprender ou aprimorar seu domínio das técnicas quantitativas da pesquisa educacional. Tais textos foram orientados para uma disciplina de quatro créditos oferecida pelo Programa de Pós-graduação em Educação em Ciências da Universidade de Brasília.

Assim como qualquer material que pretenda ensinar alguma coisa a alguém, este livro baseia-se em suposições sobre: *i)* o perfil do leitor, *ii)* o papel do professor; e *iii)* a relação que os dois estabelecerão com o texto.

Em primeiro lugar, é preciso deixar claro que *este livro* não foi pensando para *funcionar como uma leitura sequencial independente*. Até mesmo um leitor com bons conhecimentos de matemática e linguagem de programação poderá ter dificuldades em se aproximar deste livro com pretensão autodidata. Em situações concretas da formação de

pesquisadores, os pontos de partida e os objetivos são muito diferentes para pensarmos um livro de pós-graduação que atenda confortavelmente a todos os leitores. Por exemplo, enquanto alguns sabem programar e foram educados em todas as disciplinas fundamentais da matemática superior, outros não estudaram matemática além da escola. Além disso, em uma disciplina de métodos quantitativos da pesquisa educacional, podemos antecipar uma grande variedade de propósitos. Algumas pessoas não pretendem mais que aprender a interpretar dados representados em gráficos e tabelas. Outras precisam aprender o suficiente para comunicar o que desejam a um estatístico encarregado da análise. Pouquíssimos pesquisadores realmente almejam tornar-se programadores autônomos. Somente o professor será capaz de gerenciar, caso a caso, essa grande diversidade de propósitos de aprendizagem.

Como material de referência, este livro não foi escrito para ser lido e assimilado de forma necessariamente sequencial, mas consultado e revisitado. O espaço central da aprendizagem é o diálogo com o professor e com os colegas. Espera-se que o professor proponha exercícios de análise orientada. Também é esperado que o estudante vasculhe a internet buscando fontes alternativas para expandir ou aprofundar seus conhecimentos. Inserido em uma experiência de aprendizado “em rede”, este livro poderá funcionar como uma plataforma de partida e retorno. Todas as vezes em que o leitor se deparar com uma informação não compreendida no livro, deve procurar seu professor, seus colegas, bem como outras referências, retornando sempre ao livro.

Como plataforma de referência, os textos de apoio deste livro proporcionam uma sequencialidade mínima à experiência contínua e relativamente caótica de aprender estatística aplicada. Assim, os textos de apoio

do livro podem ajudar o leitor a manter uma visão panorâmica de como os diversos métodos se relacionam formando uma totalidade.

O(a) professor(a) mais ambicioso pode tentar percorrer todas as lições em um semestre de quinze semanas letivas. Porém, é recomendável que ele(a) elimine as lições consideradas menos relevantes, para dar mais espaço de discussão aos temas do interesse de sua turma.

Desejo a todos(as) uma ótima jornada!
Prof. Paulo Lima Junior
Estocolmo, 22 de junho de 2020.

IMPORTANTE

Para usufruir completamente deste material, será importante baixar e instalar alguns aplicativos e arquivos em seu computador. Acesse www.perspectivascriticas.com.br para ter acesso gratuito ao conteúdo complementar a este livro.

PARTE I.

FUNDAMENTOS

As três lições seguintes funcionam como um aquecimento para o curso, introduzindo algumas ideias básicas e criando familiaridade com a linguagem e o ambiente de programação do R.

- Na **lição 1**, vamos aprender alguns conceitos básicos da pesquisa quantitativa.
- Na **lição 2**, há uma introdução à linguagem R especialmente voltada ao manuseio de *dataframes* (estruturas tipicamente empregadas para armazenar os dados que desejamos analisar).
- Na **lição 3**, vamos exercitar um pouco mais nossa familiaridade com a linguagem R ao mesmo tempo em que aprendemos algumas noções básicas de probabilidade e testes estatísticos.

Após essas três lições, estaremos em condição de começar nossa incursão pelos métodos quantitativos bivariados.

LIÇÃO 1.

INTRODUÇÃO À ANÁLISE ESTATÍSTICA

Escolhendo o método de análise

Inicialmente, entender um relato de pesquisa quantitativa pode ser desafiador. Para a maioria dos leitores, expressões como “análise de variância”, “tabelas de contingência” ou “regressão multivariada” dizem nada ou muito pouco. Em geral, a escolha do método estatístico depende muito da maneira como os dados estão codificados. Como primeiro passo para a escolha do método, o analista precisa distinguir entre:

1. análises bivariadas ou multivariadas;
2. métodos associativos ou preditivos;
3. variáveis categóricas ou variáveis escalares.

Como primeira distinção, é preciso avaliar se sua análise investigará as relações entre duas ou muitas variáveis. Análises *bivariadas* são mais simples e geralmente funcionam como porta de entrada para o mundo da estatística. Porém, as pesquisas quantitativas realmente interessantes são quase sempre *multivariadas*.

A segunda distinção diz respeito ao tipo de relação que o analista supõe haver entre suas variáveis. As relações supostas podem ser de

dois tipos: *i)* associativas ou *ii)* preditivas. Voltaremos a essa distinção várias vezes ao longo do livro, mas podemos antecipar que *relações de associação* são sempre simétricas. As afirmações "X está associado a Y" e "Y está associado a X" devem ser consideradas equivalentes lógicos e, por isso, demandam métodos que respeitem essa reciprocidade ($X \leftrightarrow Y$). Os exemplos bivariados típicos são análises de correlação e tabelas de contingência.

Em contraste, *relações preditivas* (e.g., "X produz um efeito em Y", $X \rightarrow Y$) não são simétricas e, por isso, são mais convenientemente tratadas por métodos e linguagens especiais. Quando estamos investigando relações preditivas, faz sentido distinguir as variáveis em *dependentes* e *independentes*. Essa distinção só é possível quando esperamos que o comportamento de uma variável possa ser *unidirecionalmente* explicado pelo comportamento de outra. O exemplo mais clássico da pesquisa em educação é a avaliação do impacto de intervenções didáticas: métodos de ensino ou currículos diferentes (variáveis independentes) podem explicar mudanças na aprendizagem e no interesse dos estudantes (variáveis dependentes).

A terceira distinção é igualmente simples. As variáveis são chamadas *categóricas* quando apresentadas em níveis (ordenados ou não), tais como: nacionalidade (brasileiro, argentino, chileno…), grau de escolaridade (fundamental, médio, superior), faixas de renda (até um salário mínimo, de um a cinco salários mínimos…). Por outro lado, variáveis *escalares* são quantidades que, tais como posições no espaço, podem ser comparadas em intervalos. A nota dos estudantes em testes de conhecimento é um exemplo de variável escalar. Considere que Maria, João e Daniel tenham atingido 100, 90 e 80 pontos em um teste. Se o teste foi corretamente elaborado, podemos dizer que a distância de Maria a João é igual à distância de

João a Daniel, isto é, podemos interpretar os valores como posições em uma escala. Se não soubéssemos as pontuações dos estudantes, mas suas colocações em um *ranking*, não poderíamos fazer a mesma afirmação. Afinal, a distância do primeiro colocado ao segundo não é sempre igual à distância do segundo ao terceiro.

Com base em todas essas distinções, você poderá localizar, no quadro 1.1, qual lição deste livro será mais importante para seus propósitos de pesquisa. Ao tentar estudar uma lição mais avançada, em caso de dificuldade, você sempre poderá retornar às lições anteriores.

Quadro 1.1: Identificando as lições mais importantes em vista dos seus propósitos

Análise	Relação	Variáveis envolvidas	Lição
Bivariada	Preditiva	Categórica → Escalar	Análise de variância, teste t ou alternativas não paramétricas – Lição 4
		Escalar → Escalar	Regressão simples – Lição 5
	Associativa	Escalar ↔ Escalar	Análise de correlação – Lição 5
		Categórica ↔ Categórica	Tabelas de contingência – Lição 6
Multivariada	Preditiva	Todas → Escalar	Modelo linear clássico – Lições 7 e 8
		Todas → Binária	Regressão logística – Lição 9
	Associativa	Variáveis escalares	Análise de componentes principais – Lição 10
		Variáveis categóricas	Análise de correspondência – Lição 11
		Similaridades	Análise de *cluster* – Lição 12

Validando testes educacionais

É comum que os pesquisadores em educação empreguem testes (de atitude, conhecimento, valores, interesses…) para gerar as variáveis de seu estudo. Esses testes costumam ter a forma de um questionário de

perguntas objetivas, mas podem assumir formas alternativas. Por exemplo, o teste "desenhe um cientista" (Hillman *et al.*, 2014; Reinisch *et al.*, 2017) já foi muito popular nas pesquisas sobre a imagem pública da ciência com crianças pequenas. Ao solicitar que as crianças desenhem um cientista, geralmente obtemos um homem branco vestido num jaleco. Essas imagens podem ser analisadas e codificadas em variáveis categóricas ou escalares. Em seguida, podem ser inseridas em um modelo estatístico. Mesmo que não haja impedimento na forma dos testes educacionais, os mais comuns envolvem itens de múltipla escolha ou tipo-Likert (Kaya; Yager; Dogan, 2009; Mazas *et al.*, 2013). A despeito da forma de apresentação, testes podem ser chamados "instrumentos" porque pretendem qualificar ou quantificar uma variável de maneira mais ou menos confiável. Nesse sentido, eles funcionam como se fossem instrumentos de medida — tais como réguas, relógios e balanças.

Antes de tirar conclusões baseadas em dados coletados, é preciso considerar que, em alguma medida, o processo de produção das evidências pode, ele mesmo, comprometer ou impor limites à confiabilidade que razoavelmente poderá ser imputada às conclusões. Por exemplo, antes de especificar a probabilidade de uma criança desenhar um homem quando pedimos que represente um cientista, é preciso avaliar criticamente diversas questões, tais como: *i)* os critérios que nos permitem discriminar homens e mulheres em cada desenho; *ii)* a maneira como as crianças respondem à tarefa de desenhar um cientista; *iii)* a probabilidade de produzirem desenhos diferentes em testes sucessivos; *iv)* a relação entre desenhos, falas e ações das crianças nesse e em outros contextos.

Em pesquisa educacional, *validade* refere-se à qualidade das inferências, declarações ou decisões baseadas em respostas dadas a um instrumento e *validação* é o processo que pretende produzir evidências

capazes de sustentar a adequação, significância e utilidade dessas decisões e inferências (Zumbo; Chan, 2014). Embora seja comum falar em "instrumentos validados", o que validamos realmente não é o instrumento, mas o processo de produção das conclusões no qual o instrumento está inscrito. Um questionário já validado pode não produzir resultados confiáveis quando aplicado em outra época, a outra população ou com outros propósitos.

Ainda que alguns textos de nossa área tentem simplificar a discussão, a validação de um teste educacional é um terreno movediço. Não há um protocolo universal a seguir nem um conjunto de critérios consensuais para garantir a validade das declarações feitas a partir de dados coletados. A validação é um processo argumentativo que não deve estar baseado em somente uma fonte de evidência. As diretrizes contemporâneas destacam cinco fontes principais que podem sustentar uma declaração de validade. São elas: *i)* o conteúdo do teste, *ii)* o processo de resposta aos itens, *iii)* a estrutura interna do teste, *iv)* as relações com outras variáveis, e *v)* suas consequências (Zumbo; Chan, 2014).

A *estrutura interna do teste* é particularmente importante para a discussão que faremos nesse livro. Ela não é o único dos componentes acima que podem ser avaliados quantitativamente, mas costuma empregar ferramentas estatísticas específicas. Dizemos que um teste é internamente consistente quando as respostas dadas aos seus itens estão correlacionadas. Como veremos mais tarde, essa correlação dá estabilidade ao resultado do teste e contribui para que apresente aproximadamente o mesmo resultado se aplicado outra vez às mesmas pessoas sob as mesmas condições. A consistência interna de um teste educacional corresponde ao que, no vocabulário internacional de metrologia, é conhecido por fidedignidade ou reprodutibilidade (Joint Commitee for Guides in Metrology, 2008b).

Se você pretende elaborar e/ou aplicar algum tipo de teste educacional na sua pesquisa, provavelmente precisará recorrer aos métodos multivariados apresentados ao final do livro. Tais métodos são muito importantes para pensarmos a coleta de dados, a elaboração e validação dos instrumentos de pesquisa. Portanto, as lições finais são, em alguns casos, as primeiras a serem empregadas.

Quadro 1.2: Sobre a elaboração e validação de testes educacionais

Análise	Propósito	Lição
Consistência interna	Elaboração de testes educacionais	Lição 13
	Análise fatorial exploratória	Lição 14
	Análise confirmatória e modelagem SEM	Lição 15

Design experimental e causalidade

Quanto à obtenção dos dados, é usual distinguir ainda três tipos de investigação: *i)* experimental; *ii)* quasi-experimental e *iii)* observacional.

Nenhum método estatístico, sozinho, permite estabelecer *causalidade* entre dois fenômenos. Em primeiro lugar, é a "teoria" (i.e., o conjunto dos saberes disponíveis ao pesquisador no processo da pesquisa) que aponta as relações causais possíveis. Afinal, do que mais depende o resultado escolar dos estudantes? Da sua origem familiar? Dos métodos de ensino empregados? Das suas qualidades pessoais? No sentido mais flexível do termo, é a "teoria" que recomenda quais variáveis serão consideradas explicativas e quais devem ser explicadas.

Do ponto de vista da análise, chamamos *design experimental* qualquer estratégia de obtenção de dados que pretenda, rigorosamente, estabelecer uma relação causal entre variáveis explicativas (independentes) e variáveis a explicar (dependentes). Esse tipo de *design* é muito importante em ciências da saúde, cumprindo um papel também relevante em alguns

ramos da pesquisa educacional. Para construir um *design* experimental típico, o método empregado na pesquisa deve ser:

1. controlado;
2. randomizado; e
3. duplo-cego.

O *controle* de uma pesquisa educacional consiste em distribuir seus participantes em um ou mais grupos de tratamento e um grupo de controle (esse é o nome dado ao grupo no qual não será aplicado tratamento algum). O tratamento pode ser uma estratégia de ensino inovadora, um material didático especial, um currículo diferente. Em outras palavras, um tratamento é qualquer intervenção cujo efeito educacional possa ser medido por um teste educacional e comparado à experiência "tradicional" de aprendizagem no grupo de controle. Geralmente são aplicados um pré-teste e um pós-teste das variáveis que devem ser impactadas pelo tratamento (concepções, atitudes, valores...). Como o pertencimento aos grupos de tratamento ou controle pode ser representado por uma variável categórica e o resultado dos testes costuma ser uma variável escalar, é usual empregar testes t, análises de variância ou covariância para tratar esse tipo de dado.

A *randomização* de uma pesquisa educacional, por sua vez, consiste em atribuir os participantes aleatoriamente entre os grupos (de tratamento e controle). Em algumas pesquisas educacionais, essa atribuição pode representar um obstáculo. Se os participantes (estudantes) estão organizados em grupos (turmas) definidos pela autoridade escolar antes do início da pesquisa, é possível que essa autoridade tenha usado critérios de atribuição às turmas que comprometerão os resultados da pesquisa. Por exemplo, os estudantes podem ter sido agrupados por afinidade entre colegas, por mau comportamento, por desempenho em avaliações

anteriores. A condição de randomização requer que a atribuição de participantes aos grupos de pesquisa seja completamente aleatória.

Finalmente, é chamado *duplo-cego* o método no qual nem o "paciente" nem o "terapeuta" sabem quando está sendo aplicado tratamento ou placebo. Essa condição é facilmente satisfeita em testes de vacina, mas é pouco realizável na pesquisa educacional, pois tanto os professores quanto os estudantes costumam ser informados de que um método de ensino novo está sendo testado. Somente estudos do tipo duplo-cego seriam capazes de controlar o *efeito placebo educacional*. Nesse sentido, é possível que muitos métodos de ensino comercializados por sua eficiência tenham efeito devido às expectativas positivas que mobilizam nos participantes (não por causa do método propriamente dito). Por essas e outras razões, o *design* experimental é, ao mesmo tempo, importante, mas amplamente desacreditado na pesquisa educacional.

Denominamos *design quasi-experimental* qualquer método de pesquisa que possua algumas (mas não todas) as qualidades de um ensaio controlado, randomizado e duplo-cego. Na falta de qualquer uma dessas qualidades, será sempre possível que os efeitos observados em uma intervenção tenham sido produzidos por outra razão. Pelas particularidades da pesquisa educacional, um método pode ser considerado experimental quando controlado e randomizado (métodos educacionais do tipo duplo-cego estão geralmente fora do horizonte). Em outras palavras, o controle do efeito placebo não costuma ser feito (ou não costuma ser possível) na pesquisa em educação. Porém, mesmo adotada uma definição mais abrangente do termo, uma quantidade expressiva das pesquisas sobre o impacto de intervenções educacionais é de caráter quasi-experimental (e.g., controladas, mas não randomizadas) ou não experimental (nem controladas nem randomizadas).

Finalmente, há estudos quantitativos de outra natureza que não pretendem avaliar o "impacto" de coisa alguma. Uma grande variedade de estudos aborda temas por demais complexos para que se pretenda investigar relações causais simples. Esses podem ser considerados *observacionais*. Neles, a questão da causalidade não se coloca ou é resolvida no âmbito da teoria. São exemplos várias pesquisas de larga escala inspiradas na sociologia da educação, a análise de políticas públicas ou do funcionamento dos sistemas de ensino (Wainer; Melguizo, 2017).

Incerteza e intervalos de confiança

Na análise de dados empíricos, *todos os valores têm alguma variabilidade e não devem ser interpretados como propriedades tipo-ponto* (Buffler *et al.*, 2001; Buffler; Lubben; Ibrahim, 2009). Observe, por exemplo, sua altura. Você sabia que ela sofre variações ao longo do dia? Deitado ou pendurado numa barra, você pode ganhar facilmente mais de 1 centímetro. Portanto, qual é o sentido de declarar "eu tenho 1,79 m de altura" se esse valor está sujeito a variações? Suponha, agora, que você tenha feito um teste de QI e ele tenha resultado em 98 pontos. Qual é o sentido de declarar "meu QI é de 98 pontos" se, em caso de repetição do teste, o resultado pode ser diferente? Que sentido tem afirmar "seu QI é maior que o meu" se todos estão sujeitos à variação?

Em teoria da medição, esse problema é contornado pelo conceito de *incerteza* (Joint Commitee for Guides in Metrology, 2008a). Todos os resultados de medição devem ser obrigatoriamente acompanhados de sua respectiva incerteza. Ela permite construir *intervalos de confiança/cobertura* que provavelmente contêm outros resultados da mesma medição. Em vez de pensar a minha altura como um valor pontual (1,79 m),

vou especificá-la como um intervalo (1,79 m mais ou menos 0,02 m). Com isso, quero dizer que observações posteriores da minha altura muito provavelmente pertencerão a esse intervalo. Isso implica que uma pessoa com 1,78 m de altura talvez não seja realmente mais baixa que eu, mas uma pessoa com 1,76 m certamente será. Assim, percebemos como a incerteza é importante para fazermos comparações.

Em pesquisa educacional, o problema da medição tem os mesmos contornos. Lidamos com instrumentos que pretendem quantificar inteligência, competência, interesse, crenças… Mesmo quando a única informação que temos é um valor singular, será preciso interpretá-lo como se fosse um intervalo — i.e., se o teste educacional fosse reaplicado sob as mesmas condições, os valores medidos não seriam os mesmos. Um teste educacional pode atribuir nota a um estudante por sua competência escolar. Porém, para dizer que João é mais competente que José, não basta compararmos suas notas. É preciso levar em consideração quão dispersas seriam suas notas se a mesma avaliação fosse repetida diversas vezes sob as mesmas condições. Portanto, a incerteza é fundamental para fazemos comparações entre indivíduos e grupos.

Outra questão importante é que *todas as medidas empíricas* (em qualquer área de conhecimento) **só** *podem ser definidas dentro de um modelo teórico*. Por exemplo, se você deseja medir o diâmetro de uma bola de bilhar, sua intenção de medir já inclui uma idealização. Até que ponto essa bola de bilhar é mesmo uma esfera para que possamos nos referir ao seu diâmetro? Se um modelo teórico é condição necessária para a medição de algo tão aparentemente inconteste, o que podemos dizer das medidas educacionais, como competências, valores, interesses e atitudes? Nas ciências naturais e sociais, uma medida empírica não deve ser encarada jamais como algo que existe *em si*, mas como uma quantidade

que só é pensável assumidos certos pressupostos teóricos e idealizações (cf. Heidemann; Araujo; Veit, (2016).

Em síntese, todas as medidas empíricas carregam alguma incerteza e não devem ser encaradas como quantidades pontuais, mas como intervalos.

Significância estatística

Resultados de uma medição podem variar ao acaso. Dito isso, como podemos fazer inferências sólidas com base em evidência empírica? *Em que medida os padrões e as diferenças que observamos nos dados podem ser atribuídos ao acaso?*

Para dar conta desse problema, precisamos formular uma hipótese (ou um conjunto de hipóteses) que descreva os padrões que acreditamos observar em nossos dados. Alguns exemplos da literatura são:

- Os alunos que aprenderam ciências com métodos baseados em investigação têm desempenho *superior* àqueles que aprenderam com métodos tradicionais.
- Os estudantes mais interessados pela ciência escolar tendem a apresentar *maior desempenho* nas provas de ciência e uma *diversidade maior* de experiências científicas extraclasse.
- Os meninos são *sobrerrepresentados* entre os estudantes de mais alto de desempenho, mas também *sobrerrepresentados* entre os estudantes de mais baixo desempenho nas provas de ciências.
- O desempenho nas avaliações escolares está relacionado à origem social, mas essa relação é *mais forte* nas disciplinas de ciências da natureza e matemática que nas disciplinas de linguagens, códigos e ciências sociais.

Para cada uma dessas hipóteses, é sempre possível formular outra, segundo a qual nada está acontecendo. Por exemplo: “Os alunos que

aprenderam ciências com métodos baseados em investigação têm desempenho *igual* àqueles que aprenderam com métodos tradicionais". Afirmações desse tipo serão chamadas *hipóteses nulas ou H0*. Elas são, geralmente, uma afirmação que nós desejamos refutar. As hipóteses que afirmam algum tipo de associação entre as variáveis são chamadas "hipóteses alternativas" ou H1, H2… Em regra são essas hipóteses que nós desejamos corroborar.

Designa-se por *valor-p* a probabilidade de a hipótese nula ser verdadeira. Cabe ao analista decidir quão pequena ela deve ser. Geralmente, probabilidades inferiores a 5% (i.e., $p < 0,05$) são suficientes para que a hipótese nula seja descartada. Alguns estudos exigem que o valor-p seja inferior a 1%, ou 0,1%. Evidentemente, quando o valor-p não atinge o limite designado, a hipótese nula não chega a ser comprovada. Diz-se, então, que o estudo falhou em refutá-la.

Qualquer teste capaz de calcular o valor-p com o qual H0 é verdadeira a partir de um conjunto de dados é chamado *teste de significância estatística*. Enfim, é importante lembrar que a rejeição ou aceitação da hipótese nula é uma *decisão* do analista e, como tal, pode estar equivocada. Dizemos que o analista incorre em *Erro do tipo I* quando rejeita uma hipótese nula verdadeira. Chama-se *Erro tipo II* quando admite a hipótese nula falsa.

A "objetividade" da pesquisa quantitativa

A pesquisa educacional, ao menos no Brasil, tem uma grande resistência aos métodos quantitativos. Sob a acusação de que eles se pretendem objetivistas ou positivistas, métodos estatísticos são rechaçados antes mesmo de serem compreendidos. É usual afirmar que, no passado,

a pesquisa educacional era predominantemente quantitativa e que essa tendência foi superada. Contudo, revisões da pesquisa educacional indicam que, no Brasil, os métodos quantitativos jamais constituíram uma tradição (Gatti, 2004) e que, no exterior, jamais foram abandonados. A raridade de pesquisas quantitativas no campo educacional brasileiro contribui para reduzir nossa inserção internacional, pois uma grande parte da pesquisa educacional fora do Brasil é quantitativa. Em decorrência do rechaço aos métodos quantitativos, vai se instalando uma falta de letramento na comunidade.

De fato, quem aprendeu a fazer análises estatísticas consistentes conhece muito bem a distância que existe entre a alegada objetividade estatística e o que realmente acontece no curso da investigação. A quantidade de decisões que o analista precisa tomar é enorme e nunca se justifica somente com base em critérios estatísticos, mas demanda conhecimento teórico e intuição. Com um mínimo de honestidade intelectual, ao final da análise, é preciso reconhecer que nada está definitivamente comprovado. Será preciso ainda convencer o leitor e juntar à estatística todas as informações relevantes de seu campo de investigação. Mesmo quando nossas análises produzem convicções, elas não vêm simplesmente dos dados, tampouco podem ser generalizadas para quaisquer contextos.

Algumas palavras sobre o programa R

R é uma linguagem e um ambiente livre para computação estatística e gráfica. Ele congrega uma comunidade vasta de usuários e, por ser livre, está sempre recebendo contribuições, sempre se renovando. Não há nada em outro programa que não possa ser feito com R. Além disso, como há muitos usuários e muitas pessoas querendo ajudar, a quantidade

de tutoriais é impressionante. A qualquer momento, se você não souber como fazer alguma coisa, experimente digitar no Google ou no YouTube "How to… with R" e você encontrará exatamente a ajuda que procura (por mais específica que seja sua pergunta).

O programa básico para começar a trabalhar com o R pode ser obtido na *página do Projeto R*: https://www.r-project.org. Ao acessar o *site*, escolha um CRAN da sua preferência e baixe a versão mais atual do programa para seu sistema operacional.

O R vem com algumas funções instaladas na sua biblioteca de *base*, mas geralmente é preciso instalar novas bibliotecas para que ele aprenda novas funções. Você pode fazer isso de dentro do próprio programa digitando **`install.packages`** seguido do nome do pacote entre parênteses e aspas:

```
install.packages("Hmisc")
```

Na primeira instalação, o R pedirá que você escolha um CRAN. Após esse processo, a nova biblioteca de funções fica disponível na sua máquina e pode ser ativada pela função **`library`** a seguir (observe que o nome da biblioteca vem sem aspas agora):

```
library(Hmisc)
```

À primeira vista, o R pode ser menos atrativo porque sua interface com o usuário é bem simples, sendo necessário conhecer um pouco de linguagem de programação. Se você desejar, é possível transformar o R em um programa com interface semelhante às principais plataformas pagas (SPSS e SAS) por meio da biblioteca **`Rcmdr`**. Porém, o principal problema dessas plataformas não é o preço. Justamente pelas características da sua

interface, elas não desafiam o usuário a entender o que está ocorrendo durante a análise. Por isso, não vamos utilizar o **`Rcmdr`** aqui. Em vez dele, vamos usar o R Studio Desktop, que pode ser obtido no *site*: https://posit.co.

Dentro do R Studio, você tem muitas opções interessantes de organização do seu trabalho. Minha solução preferida é o arquivo do tipo *markdown*. Ele organiza porções de código e de texto, permitindo acrescentar explicações mais detalhadas sobre o que você está fazendo. Ele inclui um interpretador de LaTex, o que permite incluir equações matemáticas nas suas ponderações. Também permite gerar gráficos e diagramas muito mais amigáveis e seu *output* pode ser um documento *.docx, *.pdf ou *.html, facilitando o processo de produção do seu relatório de pesquisa. Para usar essa opção, dentro do R Studio, vá em `arquivo > novo > R Markdown`.

Para usufruir completamente deste livro, será importante baixar outros arquivos em seu computador. Acesse www.perspectivascriticas.com.br para ter acesso gratuito ao conteúdo que complementa este livro.

Considerações finais

Enfim, estamos começando um curso cuja proposta é predominantemente prática. Ao final, você deve ser capaz de:

1. identificar os sentidos e critérios dos métodos estatísticos a ponto de ler criticamente um relato de pesquisa empírica da educação em ciências que empregue esses métodos;
2. criar e validar instrumentos de pesquisa, coletar e analisar dados usando o R como linguagem e ambiente de programação;
3. desenvolver o senso prático da pesquisa quantitativa, discernindo quais convicções esse tipo de pesquisa é capaz de produzir em cada situação.

Atividades propostas

1. As razões que levam as pessoas a estudar (ou evitar) métodos quantitativos são as mais diversas. Quais são as suas? O que você acredita ser capaz de realizar (em seu projeto de pesquisa e em seu desenvolvimento profissional) com esses métodos de análise?
2. O que suas respostas à questão anterior revelam de suas crenças sobre a natureza da ciência e da pesquisa educacional?

LIÇÃO 2.

DATAFRAMES

O ponto de partida da maioria das análises estatísticas é um *dataframe*: uma estrutura de dados semelhante a uma planilha em que as linhas representam instâncias individuais (estudantes, professores, escolas...) e as colunas representam variáveis (i.e., informações que temos sobre as instâncias individuais). O *dataframe* é diferente de uma *matriz*, pois suas variáveis podem ser de tipos distintos (categóricas ou escalares). Na matriz, todas as entradas são necessariamente do mesmo tipo.

Para iniciar, vamos abrir um *dataframe* de *gastos educacionais* chamado "education" que já vem incluído na biblioteca `robustbase` (instale-a, se ainda não o fez). Em seguida, chame o *dataframe* com a função `data`.

```
data(education, package = "robustbase")
```

Ferramentas de visualização

O *dataframe* pode ser visualizado integralmente com a função `View`. Uma aba será aberta ao lado da janela onde você digita o código.

```
View(education)
```

Os dados de *education* podem ser interpretados da seguinte maneira:

- *State* é a sigla do estado na federação estadunidense.
- *Region* é a macrorregião à qual o estado pertence (1=Nordeste, 2=Centro-Norte, 3=Sul, 4=Oeste).
- *X1* é o número de residentes por mil em áreas urbanas em 1970.
- *X2* é a renda *per capita* em 1973.
- *X3* é o número de residentes por mil até 18 anos de idade em 1974.
- *Y* é o gasto *per capita* em educação pública no estado projetado para 1975.

Na maioria das vezes, visualizar os dados é pouco produtivo. Quando o *dataframe* tem muitas variáveis ou muitas observações, é difícil tirar qualquer conclusão de sua observação direta. Há diversas funções que nos permitem uma melhor visualização nessas condições.

A função **head**, que mostra as primeiras linhas do *dataframe*, pode ser editada para mostrar uma quantidade qualquer de linhas, mas sua quantidade padrão é *n = 6*.

```
head(education)

##   State Region  X1   X2  X3   Y
## 1    ME      1 508 3944 325 235
## 2    NH      1 564 4578 323 231
## 3    VT      1 322 4011 328 270
## 4    MA      1 846 5233 305 261
## 5    RI      1 871 4780 303 300
## 6    CT      1 774 5889 307 317
```

Outra forma de visualizar a estrutura do *dataframe* é usar a função **str**, que lista as variáveis e fornece algumas informações.

```
str(education)

## 'data.frame':    50 obs. of  6 variables:
##  $ State : Factor w/ 50 levels "AK","AL","AR",..: 21 30 46 19 39 7 34 31 38 35 ...
##  $ Region: int  1 1 1 1 1 1 1 1 2 ...
##  $ X1    : int  508 564 322 846 871 774 856 889 715 753 ...
##  $ X2    : int  3944 4578 4011 5233 4780 5889 5663 5759 4894 5012 ...
##  $ X3    : int  325 323 328 305 303 307 301 310 300 324 ...
##  $ Y     : int  235 231 270 261 300 317 387 285 300 221 ...
```

O *dataframe* apresentado é composto por 50 observações de 6 variáveis. No R, *variáveis categóricas são chamadas fatores*. Como é possível perceber, embora as variáveis *State* e *Region* sejam categóricas, somente a primeira está representada por um fator, enquanto a segunda é equivocadamente representada por uma variável escalar do tipo inteira (*int*). Isso precisa ser corrigido para não gerar inconsistências no futuro.

Gerando fatores (variáveis categóricas)

Transformar uma variável em fator é muito simples. Basta usar a função **`factor`** e chamar novamente **`str`** para ver como a estrutura mudou:

```
education$Region = factor(education$Region)
str(education)

## 'data.frame':    50 obs. of  6 variables:
##  $ State : Factor w/ 50 levels "AK","AL","AR",..: 21 30 46 19 39 7 34 31 38 35 ...
##  $ Region: Factor w/ 4 levels "1","2","3","4": 1 1 1 1 1 1 1 1 2 ...
##  $ X1    : int  508 564 322 846 871 774 856 889 715 753 ...
##  $ X2    : int  3944 4578 4011 5233 4780 5889 5663 5759 4894 5012 ...
##  $ X3    : int  325 323 328 305 303 307 301 310 300 324 ...
##  $ Y     : int  235 231 270 261 300 317 387 285 300 221 ...
```

A partir de agora, a coluna *Region* do *dataframe education* tornou-se um fator. Observe que, para chamar a coluna de um *dataframe*, utilizamos um cifrão (e.g., *education$Region*). Para atribuir um valor a uma variável do R, basta usar um sinal de igual. Ele atribui à variável à esquerda o resultado da operação feita à direita. No caso, a função *factor(education$Region)* gerou um fator a partir de *education$Region*. O sinal de igual armazena o resultado dessa operação na mesma variável *education$Region*, sobrescrevendo a informação que ali estava.

No entanto, os níveis do fator que criamos ainda não podem ser interpretados. Precisamos atribuir a eles rótulos que sejamos capazes de ler e entender. Para transformar os níveis de um fator em outros níveis, usamos a função **`mapvalues`** da biblioteca **`plyr`**. Se você ainda não instalou essa biblioteca, entre no comando **`install.packages("plyr")`**. Caso já a tenha instalado, basta acionar a biblioteca usando a função **`library`** da maneira como é indicado a seguir.

```
library(plyr)
education$Region = mapvalues(education$Region,
                        from = c("1", "2", "3", "4"),
                        to = c("Nordeste", "Centro-Norte", "Sul", "Oeste"))
```

Gerenciando nomes das variáveis

No topo de cada coluna do *dataframe*, está o nome da variável. Para facilitar a análise, é bom que esses nomes sejam significativos. Eles não devem conter espaço nem caracteres especiais, pois é possível chamar as variáveis pelos seus nomes no código. A função que permite ver e manipular o nome das variáveis é chamada **`colnames`**.

```
colnames(education)

## [1] “State”  “Region” “X1”     “X2”     “X3”     “Y”
```

Vamos mudar os nomes das colunas de maneira que sejam mais significativas:

```
colnames(education) = c(“Estado”, “Regiao”, “Resid”, “Renda”, “Resid18”, “Invest”)
str(education)

## ‘data.frame’:    50 obs. of  6 variables:
##  $ Estado : Factor w/ 50 levels “AK”,”AL”,”AR”,..: 21 30 46 19 39 7 34 31 38 35 ...
##  $ Regiao : Factor w/ 4 levels “Nordeste”,”Centro-Norte”,..: 1 1 1 1 1 1 1 1 1 2 ...
##  $ Resid  : int  508 564 322 846 871 774 856 889 715 753 ...
##  $ Renda  : int  3944 4578 4011 5233 4780 5889 5663 5759 4894 5012 ...
##  $ Resid18: int  325 323 328 305 303 307 301 310 300 324 ...
##  $ Invest : int  235 231 270 261 300 317 387 285 300 221 ...
```

O que está ocorrendo? Observe que há uma atribuição, um sinal de igual. Ou seja, o resultado da operação à direita é atribuído aos nomes das colunas (à esquerda).

Mas o que temos à direita? Saiba que *a função c cria vetores* de qualquer tipo. Vetores são estruturas de dados simples, uma sequência ordenada de informações *do mesmo tipo*. Pode ser uma sequência de números inteiros, números racionais, caracteres… No caso anterior, o vetor é uma sequência de palavras. As aspas indicam ao R que os caracteres ali dentro não devem ser lidos como variáveis ou comandos, mas como simples sequência de caracteres.

A função **`str`** mostra que as alterações no *dataframe* foram feitas com sucesso. Porém, ainda temos um problema. O nome da variável não define precisamente seu significado. Seria preciso acrescentar um rótulo a cada uma delas. Nos rótulos, nós podemos usar espaços e caracteres

especiais, colocando uma descrição mais completa do sentido de cada variável. Isso pode ser feito usando a função **`label`** da biblioteca **`Hmisc`**.

```
library(Hmisc)

label(education$Estado)  = “sigla do estado na federação estadunidense”
label(education$Regiao)  = “macrorregião à qual o estado pertence”
label(education$Resid)   = “residentes por mil em áreas urbanas em 1970”
label(education$Renda)   = “renda per capita em 1973”
label(education$Resid18) = “residentes por mil até 18 anos de idade em 1974”
label(education$Invest)  = “gasto per capita em educação pública”
```

Agora que os dados estão todos organizados, é possível obter diversas estatísticas descritivas por meio da função **`describe`**, também da biblioteca **`Hmisc`**.

```
describe(education)

## education
##
##  6  Variables      50  Observations
## ---------------------------------------------------------
## Estado : sigla do estado na federação estadunidense
##        n  missing distinct
##       50        0       50
##
## lowest : AK AL AR AZ CA, highest: VT WA WI WV WY
## ---------------------------------------------------------
## Regiao : macrorregião à qual o estado pertence
##        n  missing distinct
##       50        0        4
##
## Value        Nordeste Centro-Norte       Sul        Oeste
## Frequency           9           12        16           13
## Proportion       0.18         0.24      0.32         0.26
## ---------------------------------------------------------
```

```
## Resid : residentes por mil em áreas urbanas em 1970
##      n  missing distinct  Info   Mean    Gmd    .05    .10
##     50        0       49     1  657.8  167.7  443.9  449.6
##    .25      .50      .75      .90      .95
##  546.8    662.5    782.2    832.5    864.2
##
## lowest : 322 390 443 445 446, highest: 846 856 871 889 909
## ------------------------------------------------------------
## Renda : renda per capita em 1973
##      n  missing distinct  Info   Mean    Gmd    .05    .10
##     50        0       50     1   4675  747.3   3742   3824
##    .25      .50      .75      .90      .95
##   4137     4706     5054     5565     5712
##
## lowest : 3448 3680 3724 3764 3817, highest: 5613 5663 5753 5759 5889
## ------------------------------------------------------------
## Resid18 : residentes por mil até 18 anos de idade em 1974
##      n  missing distinct  Info   Mean    Gmd    .05    .10
##     50        0       36 0.999  325.7  21.05  301.9  304.9
##    .25      .50      .75   .90      .95
##  310.8    324.5    333.0 345.1    362.4
##
## lowest : 287 300 301 303 304, highest: 355 358 366 378 386
## ------------------------------------------------------------
## Invest : gasto per capita em educação pública
##      n  missing distinct  Info   Mean    Gmd    .05    .10
##     50        0       43     1  284.6  64.55  214.4  220.5
##    .25      .50      .75   .90    .95
##  234.2    269.5    316.8 342.2  378.5
##
## lowest : 208 212 214 215 216, highest: 344 378 379 387 546
## ------------------------------------------------------------
```

O que temos aqui já é muito mais que diversas análises costumam oferecer: um conjunto de estatísticas descritivas dos dados obtidos. Para as variáveis categóricas (aqui chamadas fatores), há uma análise de frequência. Para variáveis escalares, a função que executamos publicará informações como média, maiores e menores valores e percentis.

Publicando partes do *dataframe*

Já aprendemos que, usando o cifrão, é possível chamar uma variável do *dataframe* pelo nome. Porém, há uma forma mais versátil de publicar partes de um *dataframe*. Assim como matrizes, seus elementos podem ser localizados por um par de coordenadas entre colchetes que significam [linha, coluna]. Portanto, informação localizada na terceira linha, quarta coluna pode ser publicada desta maneira:

```
education[3, 4]

## renda per capita em 1973
## [1] 4011
```

Se quisermos publicar a terceira linha (ou a quarta coluna) na sua integridade, basta deixar em branco o espaço do outro índice:

```
education[3, ]

##   Estado   Regiao Resid Renda Resid18 Invest
## 3     VT Nordeste   322  4011     328    270

education[, 4]

## renda per capita em 1973
##  [1] 3944 4578 4011 5233 4780 5889 5663 5759 4894 5012 4908 5753 5439 4634 4921
## [16] 4869 4672 4782 4296 4827 5057 5540 5331 4715 3828 4120 3817 4243 4647 3967
## [31] 3946 3724 3448 3680 3825 4189 4336 4418 4323 4813 5046 3764 4504 4005 5560
## [46] 4989 4697 5438 5309 5613
```

Se quisermos mostrar todo o *dataframe* com exceção de uma linha (ou uma coluna), basta colocar um sinal de subtração. A seguir, publicamos as primeiras seis linhas do *dataframe* após remover sua quarta coluna (renda *per capita*):

```
head(education[, -4])

##   Estado   Regiao Resid Resid18 Invest
## 1     ME Nordeste   508     325    235
## 2     NH Nordeste   564     323    231
## 3     VT Nordeste   322     328    270
## 4     MA Nordeste   846     305    261
## 5     RI Nordeste   871     303    300
## 6     CT Nordeste   774     307    317
```

Se desejarmos publicar ou subtrair mais de uma coluna (ou linha) ao mesmo tempo, é preciso usar a função **c**, que concatena uma sequência de elementos em um vetor. Se tentarmos inserir uma sequência de valores separados por vírgula no interior dos colchetes, o R produzirá um erro, pois só pode haver uma vírgula na chamada de um *dataframe*:

```
head(education[,-c(3:5)])

##   Estado   Regiao Invest
## 1     ME Nordeste    235
## 2     NH Nordeste    231
## 3     VT Nordeste    270
## 4     MA Nordeste    261
## 5     RI Nordeste    300
## 6     CT Nordeste    317
```

Observe que o sinal de dois pontos entre dois números **`(3:5)`** é interpretado pelo R como uma sequência inteira que vai do primeiro ao último: **`(3,4,5)`**.

A chamada numérica que aprendemos nesta seção é fundamental para manipular, ordenar e filtrar o *dataframe*.

Ordenando o *dataframe*

Para ordenar um *dataframe*, usa-se a função **`order`**, que funciona de maneira bastante intuitiva. Considere, por exemplo, que desejemos ordenar as unidades da federação estadunidense em função do seu investimento *per capita* em educação pública. Nesse caso, entramos com essa variável na função **`order`** para ver o que ocorre:

```
order(education$Invest)

##  [1] 32 31 25 33 30 10 34 19  2 17 16 27 36  1 29 35 26 18 28  4 24 11 20 39 37
## [26]  3  8 45  5  9 38 41 12 49 46 44 47  6 42 40 23 43 48 21 14 22 15 13  7 50
```

A função **`order`** recebe um vetor e publica outro. Após ordenar o vetor de entrada, a função publica os endereços originais dos itens em ordem crescente. Ou seja, na 32ª linha do *dataframe*, está o estado que menos investe em educação pública. O segundo que menos investe está na 31ª linha. O terceiro, na 25ª linha… O estado da 50ª linha apresenta o maior investimento público *per capita* em educação.

Para reconstruir a lista de valores ordenados, bastaria chamar os itens na ordem publicada anteriormente:

```
education$Invest[32]
education$Invest[31]
education$Invest[25]
```

Porém, isso é o mesmo que inserir a função **`order(education$Invest)`** nos colchetes de chamada do *dataframe*:

```
head(education[order(education$Invest), -c(3:5)], n = 10)
```

```
##    Estado       Regiao Invest
## 32     AL          Sul    208
## 31     TN          Sul    212
## 25     WV          Sul    214
## 33     MS          Sul    215
## 30     DY          Sul    216
## 10     OH Centro-Norte    221
## 34     AR          Sul    221
## 19     SD Centro-Norte    230
## 2      NH     Nordeste    231
## 17     MO Centro-Norte    231
```

Como é possível perceber, o *dataframe* foi ordenado com sucesso. Os estados com menor investimento público *per capita* em educação estão nas regiões Sul e Centro-Norte dos EUA. Se quisermos colocar a planilha em ordem decrescente, basta lançar um sinal negativo no interior da função **`order`**:

```
head(education[order(-education$Invest), -c(3:5)], n = 10)
```

```
##    Estado       Regiao Invest
## 50     HI        Oeste    546
## 7      NY     Nordeste    387
## 13     MI Centro-Norte    379
## 15     MN Centro-Norte    378
## 22     DE          Sul    344
## 14     WI Centro-Norte    342
## 21     KS Centro-Norte    337
## 43     AZ        Oeste    332
## 48     CA        Oeste    332
## 23     MD          Sul    330
```

É importante observar que nenhum dos comandos que abordamos nesta seção (ou na seção anterior) altera o *dataframe* original. Eles são maneiras de publicar um *dataframe* filtrando-o ou ordenando-o, mas sempre preservando sua informação original. Por exemplo, o comando **`education[, -4]`** chama o *dataframe education* omitindo a quarta coluna, mas sem apagá-la da memória. Da mesma maneira, o comando **`education[order(-education$Invest), ]`** mostra o *dataframe* ordenado sem salvar o ordenamento. Se você chamar o *dataframe* em seguida, perceberá que ele permanece em sua desordem original.

Se você quiser salvar o *dataframe* após ordená-lo e/ou eliminar algumas colunas, será preciso fazer uma atribuição (usando o sinal de igualdade):

```
education = education[order(education$Invest),]
```

Essa atribuição sobrescreverá o *dataframe* original. Por isso, deve ser empregada com muita cautela. Para evitar surpresas desagradáveis, eu prefiro concentrar todas as atribuições que efetivamente alteram o conteúdo do *dataframe* no início do código. Uma vez que o *dataframe* tenha atingido a forma desejada, sigo programando sem alterá-lo jamais. Se for muito necessário salvar uma versão reordenada do *dataframe* no meio do código, recomendo criar outro *dataframe* com os dados ordenados:

```
ordenado = education[order(education$Invest), ]
```

Isso manterá o código mais legível, organizado e consistente.

Filtrando o *dataframe*

A realização de filtros em um *dataframe* é igualmente intuitiva. Filtros permitem selecionar linhas e/ou colunas com base em alguma regra lógica. Dessa forma, em primeiro lugar, é preciso entender como funcionam os testes lógicos no R. Assim como em outras linguagens de programação, testes lógicos são operações que podem resultar em dois valores: TRUE ou FALSE. Há dois tipos de operadores que podemos empregar para gerar esses testes. Os *operadores relacionais* estabelecem relações entre variáveis (geralmente numéricas) e são os seguintes:

- < Menor que;
- > Maior que;
- <= Menor ou igual;
- >= Maior ou igual;
- == Igual;
- != Diferente;
- %in% Pertence.

Dos operadores apresentados, os quatro primeiros estabelecem relações entre números. Os demais, entre quaisquer variáveis. O último operador é um teste de pertinência. Ele compara um vetor a outro, retornando TRUE para cada elemento do primeiro vetor que estiver contido no segundo.

Os *operadores lógicos* estabelecem relações entre proposições lógicas e são os seguintes:

- ! Negação lógica;
- & Disjunção E;
- | Conjunção OU;

Suponha que estejamos interessados em localizar os estados dos EUA com investimento educacional superior à média dos que se situem

na região Sul ou Oeste. Para essa busca, formulamos um teste lógico e o inserimos na chamada das linhas entre os colchetes do *dataframe*:

```
teste = education$Regiao %in% c("Sul", "Oeste") &
             education$Invest > mean(education$Invest)
head(education[teste, ], n = 10)

##    Estado Regiao Resid Renda Resid18 Invest
## 22     DE    Sul   722  5540     328    344
## 23     MD    Sul   766  5331     323    330
## 38     MT  Oeste   534  4418     335    302
## 40     WY  Oeste   605  4813     331    323
## 41     CO  Oeste   785  5046     324    304
## 42     NM  Oeste   698  3764     366    317
## 43     AZ  Oeste   796  4504     340    332
## 44     UT  Oeste   804  4005     378    315
## 45     NV  Oeste   809  5560     330    291
## 46     WA  Oeste   726  4989     313    312
```

Ao inserir um teste lógico na chamada de linhas de um *dataframe*, serão publicadas somente as linhas para as quais o teste é verdadeiro. As linhas em que o teste resulta falso serão omitidas. Para que essa operação funcione da maneira como descrevemos aqui, é desejável que a quantidade de saídas no teste lógico seja igual à quantidade de linhas do *dataframe*.

Importar e salvar

Na maioria das situações de pesquisa, é preciso importar um *dataframe* de uma planilha do Excel. Nesse caso, empregamos a biblioteca **`readxl`**.

```
library(readxl)

setwd("C:/nomediretorio/")
dataset = read_excel("nomearquivo.xls")
```

O comando **setwd** estabelece o diretório de trabalho. Uma vez empregado esse comando, todos os arquivos serão salvos e lidos no mesmo diretório. Geralmente o R assume que o diretório de trabalho é o mesmo em que está salvo o código que você está executando. Ao terminar sua análise, sobretudo se estiver trabalhando com muitos dados, pode ser útil salvar seu *dataframe* em um arquivo binário, que poderá ser carregado rapidamente no dia seguinte:

```
save(education, file = "nomearquivo.dat")
load("nomearquivo.dat")
```

Com o comando **load**, as variáveis importadas do arquivo sobrescreverão as variáveis eventualmente ativas no ambiente global do R. É preciso, portanto, estar atento ao carregar arquivos binários para o ambiente de programação.

Revisando a lição

Nesta lição, você aprendeu a:

1. visualizar um *dataframe* com as funções **view**, **head** e **str**;
2. gerar e recodificar variáveis categóricas com as funções **factor** e **mapvalues**(plyr);
3. gerenciar nomes das variáveis com **colnames**, **label**(Hmisc);
4. gerar um relatório de estatísticas descritivas com **describe**(Hmisc);
5. chamar partes de um *dataframe* usando a notação matricial **DataFrame[i,j]**;
6. ordenar um *dataframe*, inserindo a função **order** na sua chamada matricial;

7. filtrar um *dataframe*, inserindo **`testes lógicos`** na sua chamada matricial;
8. importar e salvar um *dataframe* usando as funções **`save`** e **`load`**.

Atividades propostas

1. Busque uma planilha com dados educacionais que seja do seu interesse. Pode ser uma planilha de avaliações dos alunos de uma turma ou uma escola. Podem ser dados que você já coletou. Como alternativa, no *site* do Inep, você pode encontrar dados educacionais individualizados (eles são chamados *microdados*) de abrangência nacional. Após salvar sua planilha em um formato adequado (.*xls ou* .csv), experimente:
    - carregá-la para o ambiente do R;
    - visualizar a planilha;
    - gerenciar rótulos e nomes das variáveis;
    - gerar um relatório de estatísticas descritivas;
    - filtrar e ordenar a planilha;
    - salvá-la como um arquivo binário `*.dat` para uso posterior.

LIÇÃO 3.

MÉDIA E DISPERSÃO

As duas informações mais úteis tanto para descrever quanto para fazer inferências e comparações sobre um conjunto de dados escalares são a *média* e a *dispersão* desse conjunto de dados. Neste capítulo, aprenderemos a gerar dados aleatórios com propriedades específicas, obtendo suas médias e variâncias, criando gráficos que nos permitam visualizar como esses dados estão distribuídos.

Calculando médias aritméticas

Considere que se pretenda calcular a média de uma série de observações:

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Podemos usar a função **c** para concatenar um vetor numérico dos valores x_i:

```
x = c(0, 2, 4, 8, 12, 16, 24, 28, 32)
```

A média aritmética dos valores x_i pode ser obtida somando-os com a função **sum** e dividindo o resultado pela quantidade de parcelas dessa soma, obtida pela função **length**:

```
sum(x)/length(x)

## [1] 14
```

Também é possível calcular a média dos valores no vetor **x** pela função **mean**:

```
mean(x)

## [1] 14
```

Para tornar o exercício um pouco mais interessante, vamos usar uma função capaz de gerar uma sequência aleatória de valores segundo uma distribuição específica. A função **rnorm(n, mean, sd)** gera uma amostra aleatória de **n** valores consistentes com uma distribuição normal de média **mean** e desvio padrão **sd**. Se demandarmos uma distribuição com média igual a 500 e desvio padrão igual a 100, os resultados poderão ser interpretados como se fossem notas do Enem.

```
x = rnorm(n = 5, mean = 500, sd = 100)
```

Com a função **round**, arredondamos os valores para não apresentarem nenhum dígito decimal. Vamos ver como fica:

```
x = round(x, digits = 0)
x

## [1] 526 460 587 417 475
```

Experimente, agora, calcular a média dos valores aleatórios gerados:

```
mean(x)

## [1] 493
```

Como é possível perceber, a média calculada não coincide rigorosamente com a média que fixamos na função **rnorm**. Isso ocorre por um *efeito de amostragem*. Você pode imaginar que a função **rnorm** está "sorteando" valores de uma população de dados que tem média igual a 500. A média dos valores sorteados flutua, portanto, em torno da média populacional (ora acima, ora abaixo). Você pode executar o código várias vezes para perceber como a média da amostra flutua aleatoriamente em torno de 500.

Outra questão que você deve ser capaz de perceber agora é que, se aumentarmos o tamanho **n** da amostra, a flutuação da média será menor. Ou seja, os valores das médias amostrais tendem a ser mais próximos da média populacional quando **n** aumenta. Você pode executar o código anterior várias vezes, variando **n**, para perceber que a média da amostra flutua menos quando a amostra é maior.

De fato, quando pretendemos fazer inferência sobre o comportamento médio de uma população, *o tamanho da amostra é fundamental*. Em amostras maiores, a flutuação estatística da média será menor, tornando-nos capazes de fazer inferências mais precisas. As flutuações são efeitos aleatórios que acrescentam "ruído" aos dados. A repetição de observações e o aumento do tamanho da amostra permitem reduzir o efeito desse ruído sobre a média.

Calculando variâncias

Há diversas medidas de tendência central (e.g., moda, mediana, média geométrica), mas a média aritmética é a mais empregada. Ao lado dela, a *variância amostral* é a medida de dispersão mais usual. Ela corresponde à *soma dos desvios quadráticos* dividida pela respectiva quantidade de *graus de liberdade gl*:

$$s_x^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{gl}$$

Em primeiro lugar, um *desvio* de uma observação é sua distância à média aritmética. Ele é calculado com uma simples operação de subtração:

$$x_i - \overline{x}$$

Como a média é uma medida de tendência central, o desvio será negativo para algumas observações e positivo para as outras de tal maneira que a soma dos desvios seja sempre igual a zero. No nosso exemplo, podemos calcular os desvios fazendo:

```
x - mean(x)

## [1]  33 -33  94 -76 -18
```

A *soma dos desvios quadráticos* corresponde a elevar cada um dos desvios ao quadrado e somar o resultado.

$$\sum_{i=1}^{n}(x_i - \overline{x})^2$$

Isso pode ser feito de forma simples:

```
(x - mean(x))^2

## [1] 1089 1089 8836 5776  324

sum((x - mean(x))^2)

## [1] 17114
```

Ao elevarmos os desvios ao quadrado, duas coisas ocorrem:

1. todos os valores negativos ficam positivos;
2. os valores mais distantes da média tornam-se mais relevantes.

A soma dos desvios quadráticos deve ser tão maior quanto mais dispersos forem os dados. Por isso, estamos construindo aqui uma *medida de dispersão*. Porém, como essa soma tem n termos, ela não dependerá somente da dispersão dos dados, mas do tamanho da amostra. Para resolver isso, precisamos dividir a soma por uma quantidade que dependa do número de elementos na amostra. Essa é a chamada *quantidade de graus de liberdade* (Crawley, 2005).

A noção de *grau de liberdade* leva em consideração duas informações:

1. o tamanho *n* da amostra; e
2. a quantidade de parâmetros *p* que precisam ser estimados a partir da amostra.

Como é possível perceber, o cálculo da média é um passo necessário para obtermos a variância. Há casos em que nosso modelo analítico requer agregar a amostra em grupos diferentes (grupos de tratamento e controle, por exemplo), avaliando a média das medidas em cada grupo e comparando-as. Nessas situações, a quantidade p de parâmetros a serem

estimados será maior que 1. Porém, nas análises de amostra simples, como a analisada nesta lição, temos p=1.

A quantidade de graus de liberdade *gl* pode ser definida como a diferença entre a quantidade de observações *n* e a quantidade de parâmetros a estimar *p* (Crawley, 2005):

$$gl = n - p$$

Enfim, a *variância* pode ser calculada pela expressão a seguir:

```
sum((x - mean(x))^2)/(length(x) - 1)

## [1] 4278.5
```

Ou, diretamente, pela função **`var`**:

```
var(x)

## [1] 4278.5
```

Calculando desvios padrão

O principal componente das variâncias é um somatório de desvios quadráticos. Portanto, é fácil somar variâncias umas às outras. Em vista de poderem ser somadas facilmente (além de outras propriedades), elas são medidas de dispersão muito convenientes para gerar algoritmos mais elaborados. Por outro lado, uma limitação das variâncias é que, por serem quadráticas, elas estão fora de escala. Por exemplo, se a variável escalar x fosse medida em metros, a variância seria medida em metros ao quadrado.

Para retornar à escala de x, calculamos a raiz quadrada (**sqrt**) da variância:

$$s_x = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n - p}}$$

Essa quantidade é definida como *desvio padrão* e pode ser calculada da seguinte maneira:

```
sqrt(sum((x - mean(x))^2)/(length(x) - 1))

## [1] 65.41024
```

Ou, diretamente, pela função **sd** (ing.: *standard deviation*):

```
sd(x)

## [1] 65.41024
```

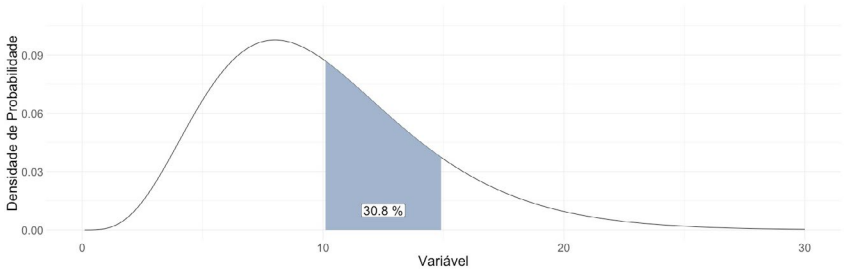
Desvios padrão não têm as propriedades aditivas da variância (para somá-los será necessário elevá-los ao quadrado), mas são mais fáceis de interpretar.

Distribuição e densidade de probabilidade

A probabilidade de ocorrência de um evento pode ser representada de forma equivalente por diversas funções. A mais empregada com variáveis escalares é a *densidade de probabilidade*, uma função cuja integral (i.e., a área sob a curva) em um dado intervalo reproduz a probabilidade de ocorrência da variável naquele intervalo. Quanto maior for a área, maior

será a probabilidade de ocorrência. Evidentemente, a área total sob uma curva de densidade de probabilidade deve ser igual a 100%. O gráfico 3.1 ilustra essa questão ao integrar uma densidade de probabilidade não normal no intervalo de x = 10 a x = 15:

Gráfico 3.1: Densidade de probabilidade de uma distribuição não normal



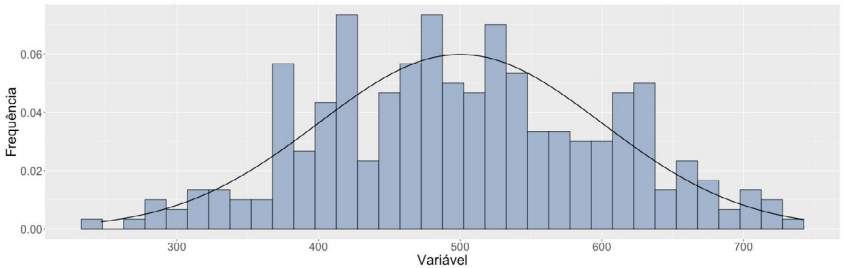
Para podemos gerar representações gráficas de dados normalmente distribuídos, vamos trabalhar com uma amostra mais numerosa a partir de agora (n = 300). Em seguida, geramos um histograma com a função **`hist`**:

```
y = rnorm(n = 300, mean = 500, sd = 100)
hist(y)
```

O histograma, representado no gráfico 3.2, separa a variável escalar analisada em algumas faixas de valores, informando a frequência relativa dessa variável em cada faixa. É possível perceber duas coisas:

1. os valores mais prováveis estão em torno de 500, a média da população de dados de onde a amostra foi “sorteada”;
2. a probabilidade em cada faixa diminui rapidamente quando nos afastamos do centro da distribuição.
3. o gráfico de barras lembra uma distribuição em formato de sino semelhante (mas não idêntica) à distribuição normal teórica.

Gráfico 3.2: Histograma de dados normalmente distribuídos



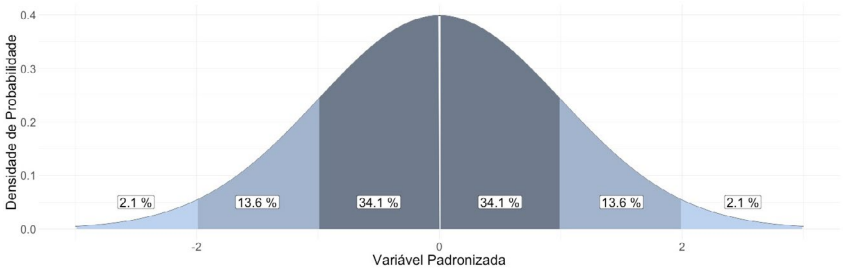
A saber, *quanto maior for o tamanho da amostra, mais fielmente ela representará a distribuição normal teórica*. Você pode verificar essa afirmação aumentando o tamanho da amostra e percebendo como o histograma tende ao formato de sino característico da distribuição normal.

Intervalos normais de cobertura

Quando estamos lidando com uma variável *normalmente distribuída* (i.e., que varia aleatoriamente conforme uma distribuição normal), é possível construir intervalos cuja cobertura aproximada é conhecida (cf. gráfico 3.3):

- o intervalo de $\overline{x} \pm 1s_x$ encerra 68,2% dos dados;
- o intervalo de $\overline{x} \pm 2s_x$ encerra 95,0% dos dados;
- o intervalo de $\overline{x} \pm 3s_x$ encerra 99,7% dos dados.

Gráfico 3.3: Intervalos normais de cobertura



Essas proporções podem ser evocadas, por exemplo, para interpretar os resultados de grandes avaliações escolares (sob a hipótese de que elas sejam normalmente distribuídas). Afinal, o que significa obter 600 pontos em uma prova? Sem informações acessórias, é impossível avaliar se 600 pontos representam sucesso ou fracasso. Porém, se as pontuações têm média 500, desvio padrão 100 e podem ser consideradas normalmente distribuídas, é possível estimar que um estudante com 600 pontos tenha apenas 16% dos concorrentes à sua frente. Da mesma maneira, um estudante que tenha obtido 700 pontos na avaliação está entre os 2,5% de maior desempenho.

Para facilitar esse tipo de interpretação, é usual realizarmos uma *padronização dos dados*. A propósito, padronizar é o mesmo que dividir os desvios das medidas ($x_i - \overline{x}$) pelo seu respectivo desvio padrão s_x:

$$z_i = \frac{x_i - \overline{x}}{s_x}$$

Em consequência, variáveis padronizadas z_i têm média zero e desvio padrão unitário. Valores negativos são inferiores à média, valores positivos, superiores. A variável padronizada z_i informa imediatamente sua distância à média em unidades de desvios padrão. A padronização de uma variável pode ser realizada pela definição apresentada anteriormente:

```
z = (x - mean(x))/sd(x)
```

Ou empregando a função **`scale`**:

```
z = scale(x)
```

Testes de significância estatística

Na lição 1, aprendemos que, quando o valor-p (i.e., a probabilidade de a hipótese nula ser verdadeira) for inferior a 5%, é usual considerar refutada a hipótese H0. Mas como funcionam os testes de significância estatística? Como esse valor-p é calculado?

Na maioria dos casos do nosso interesse, os testes de significância estatística consistem em:

1. **determinar a estatística do teste**, uma quantidade que possa ser avaliada a partir dos dados (por exemplo, a diferença padronizada entre as médias de duas amostras);
2. **estabelecer a distribuição teórica** que a estatística do teste deve satisfazer quando a hipótese nula for verdadeira;
3. posicionar a estatística na distribuição teórica e a calcular a área da cauda da distribuição, que informa a probabilidade de obter um valor igual ou superior ao obtido para a estatística do teste.

O **valor-p** é a área da cauda da distribuição.

Portanto, onde for válida a distribuição normal, *afastamentos iguais ou superiores a dois desvios padrão serão considerados estatisticamente significativos*, pois as áreas das duas caudas somadas correspondem exatamente a 5% da distribuição. Para a maioria dos propósitos, dois desvios padrão podem ser considerados uma distância grande ao centro da distribuição.

Teorema do limite central

Tudo o que fizemos até agora, nesta lição, foi trabalhar com dados gerados artificialmente, supondo uma distribuição normal. Mas por que

essa distribuição é tão importante? Em que situações reais é possível assumir que estamos diante de dados normalmente distribuídos?

Para atacar essa questão, recorremos ao *teorema do limite central*. Segundo esse teorema, *quando somamos um conjunto suficientemente numeroso de variáveis aleatórias descorrelacionadas e finitas, o resultado tende a uma variável normalmente distribuída* (Vuolo, 1996). Vamos verificar isso computacionalmente.

Em primeiro lugar, vamos criar um *dataframe* com 50 variáveis e 500 observações — todas geradas aleatoriamente a partir de uma *distribuição uniforme*. Para isso, usamos a função **`runif`** a seguir:

```
df = as.data.frame(matrix(runif(25000), nrow = 500, ncol = 50))
```

A função **`runif`** gera números aleatórios com distribuição uniforme entre 0 e 1. Ou seja, diferente do que ocorre na distribuição normal, todos os valores entre 0 e 1 são equiprováveis na distribuição uniforme. Além disso, a função **`matrix`** empregada anteriormente gera uma matriz com 500 linhas e 50 colunas. A função **`as.data.frame`** transforma a matriz em um *dataframe*. O resultado é atribuído à estrutura de dados **`df`**.

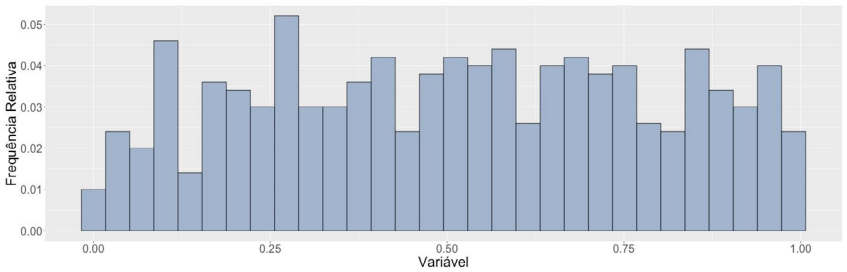
Observe que, quando chamamos uma função dentro de outra, o interpretador avalia as funções de dentro para fora. Em outras palavras, um comando do tipo **`A(B(x))`** consistente em fazer primeiro **`B(x)`** e levar o resultado na função **`A`**. No nosso exemplo, **`as.data.frame(matrix(runif()))`** calcula primeiro **`runif`**, leva o resultado em **`matrix`** e, finalmente, leva o resultado de **`matrix`** em **`as.data.frame`**. Essa diferença entre a ordem de execução e a ordem em que as funções aparecem na linha de comando pode tornar o código difícil de compreender.

O operador **pipe** %>% da biblioteca **magrittr** (instale-a, se ainda não o fez) permite contornar esse problema. Com ele, podemos escrever:

```
library(magrittr)

df = runif(25000) %>%
          matrix(nrow = 500, ncol = 50) %>%
          as.data.frame()
```

A seguir, um histograma da primeira coluna desse *dataframe* mostra como sua distribuição não é nada parecida com uma distribuição normal:

Gráfico 3.4: Histograma de uma distribuição uniforme



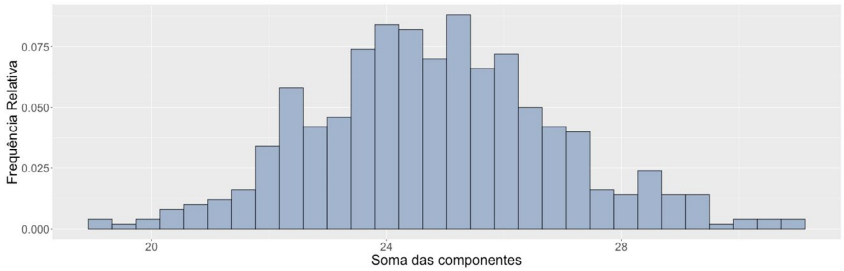
Se o *teorema do limite central* estiver correto, ao somar todas as colunas de **df**, obteremos uma variável com distribuição aproximadamente normal. Isso pode ser feito por meio da função **rowSums**, que soma os valores em cada linha de um *dataframe*. O histograma dos dados somados foi gerado e pode ser visto no gráfico 3.5. Conforme esperávamos, obtivemos uma distribuição em formato de boca de sino, que lembra uma distribuição normal.

Assim, uma consequência direta do *teorema do limite central* é que *variáveis escalares obtidas como uma composição de muitas observações não correlacionadas têm distribuição aproximadamente normal* (Vuolo, 1996).

De fato, uma grande parte das variáveis escalares com as quais trabalhamos são produzidas como uma composição de medidas mais ou menos independentes. O desempenho em uma prova de conhecimentos resulta de responder a uma bateria de inúmeras questões. Escalas de atitude geralmente são construídas como um conjunto de itens tipo-Likert. Assim, o resultado de um teste educacional (de conhecimentos, atitudes, valores, interesses…) tem grande chance de seguir aproximadamente uma distribuição normal.

```
hist(rowSums(df))
```

Gráfico 3.5: Histograma da soma das variáveis



A maioria dos métodos estatísticos presume que nossas variáveis, de alguma maneira, sejam normalmente distribuídas. Pelo *teorema do limite central*, esse pressuposto de normalidade não é absurdo, mas usualmente satisfeito dentro de certos limites práticos (Vuolo, 1996).

Por tudo isso, a distribuição normal é muito importante para a análise estatística.

Revisando a lição

Nessa lição, você aprendeu a:

1. gerar uma sequência normalmente ou uniformemente distribuída de valores aleatórios por meio das funções **`rnorm`** e **`runif`**, respectivamente;
2. calcular e interpretar média, variância e desvio padrão de uma série de dados com suas respectivas funções **`- mean`**, **`var`**, **`sd`**;
3. visualizar uma distribuição de probabilidades por meio de um histograma (**`hist`**);
4. *interpretar variáveis padronizadas* e normalmente distribuídas, conhecendo os intervalos da distribuição normal;
5. reconhecer que uma variável composta por uma soma de diversas variáveis independentes tende a ser normalmente distribuída pelo *teorema central do limite*.

Atividades propostas

1. A normalidade da distribuição obtida pela soma de variáveis depende fundamentalmente da quantidade de somas realizadas. Se somarmos muitas variáveis, é esperado obtermos uma distribuição aproximadamente normal. Crie um *dataframe* com *k* variáveis aleatórias de distribuição não normal como fizemos nesta lição e experimente somar as linhas desse *dataframe* variando sua quantidade de colunas. Quantas variáveis são necessárias para obtermos uma distribuição aproximadamente normal?
2. Nesta lição, utilizamos a função **`hist`** para gerar histogramas. Ao fazer **`help(hist)`**, leia a documentação dessa função. Observe que

é possível inserir título do gráfico e nomes dos eixos, definir escalas e pontos de quebra do histograma. Experimente gerar seus gráficos controlando esses parâmetros.

3. Nesta lição, utilizamos a função **`rnorm`** para gerar dados normalmente distribuídos. Todas as principais distribuições estatísticas estão implementadas no R por meio de funções semelhantes, sejam elas: **`rchisq`** (para a distribuição chi-quadrado), **rf** (para a distribuição F) e **`runif`** (para a distribuição uniforme). Após inspecionar a documentação dessas funções, gere valores aleatórios segundo essas distribuições e represente-os em histogramas.

PARTE II.

ANÁLISE BIVARIADA

Diversos problemas simples da pesquisa em educação consistem em testar a associação entre duas variáveis diferentes. Podemos estar interessados em saber se o interesse científico está relacionado ao gênero, se a aquisição de conhecimento melhora quando os estudantes são apresentados a um currículo, um método, um material de ensino diferente, se a escolha por carreiras científicas tem relação com a origem social, se a relação com o conhecimento está associada a crenças religiosas… A lista não tem fim.

Nas três lições seguintes, vamos aprender a produzir gráficos e testes estatísticos simples que permitam avaliar *a relação entre duas variáveis*. Como as variáveis podem estar codificadas em escala ou categoria (e isso precisa ser levado em consideração na escolha dos testes), estas lições ficaram organizadas da seguinte maneira:

- **lição 4** - ferramentas para investigar o poder preditivo de uma variável categórica (e.g., sexo, renda familiar) sobre *uma variável escalar* (e.g., concepções, atitudes, desempenho).
- **lição 5** - ferramentas para investigar as relações associativas ou preditivas entre *duas variáveis escalares* (e.g., se as notas em avaliações de ciências da natureza e matemática estão associadas ou se é

possível predizer o desempenho em ciências da natureza a partir do desempenho em matemática).

- **lição 6** - ferramentas para investigar a associação entre *duas variáveis categóricas* (e.g., se o grau de escolaridade do pai está associado ao grau de escolaridade da mãe).

Nessas três lições, vamos carregar um *dataframe* de dados do Enem de 2018. Trata-se de uma planilha com variáveis selecionadas e recodificadas. A planilha original era muito maior. Foi filtrada (eliminando dados faltantes) e 500 mil casos foram selecionados aleatoriamente. As variáveis desse *dataframe* já receberam nomes e rótulos, tornando-se autoexplicativas. O arquivo pode ser encontrado e baixado no endereço eletrônico www.perspectivascriticas.com.br.

LIÇÃO 4.

COMPARANDO MÉDIAS

A comparação entre médias é a principal porta de entrada para os métodos estatísticos da pesquisa educacional. Talvez isso ocorra porque o chamado *design* experimental, mesmo pouco empregado na prática, protagoniza alguns debates sobre os métodos quantitativos da pesquisa em educação. Nesse tipo de desenho de pesquisa, uma comparação entre médias costuma ser suficiente para produzir resultados consistentes.

Há, também, outras situações simples e não experimentais que podem ser resolvidas por uma comparação entre médias. Por exemplo, podemos estar interessados em medir as diferenças de desempenho ou de interesse por ciências entre meninos e meninas, pardos e brancos, estudantes de baixa e alta renda, cristãos e agnósticos. Todos esses objetivos de pesquisa podem ser abordados comparando o valor médio de uma medida educacional em dois ou mais grupos.

Algumas palavras sobre o *design* experimental

Suponha, por exemplo, que desejemos avaliar o efeito de um material didático, de uma metodologia de ensino, de uma visita ao museu

ou de um desenho curricular inovador sobre o desempenho, o interesse, as crenças, a motivação dos estudantes. Por analogia à pesquisa em saúde, é usual chamar de "tratamento" qualquer intervenção cujo efeito educacional desejamos medir. Uma maneira relativamente simples e tradicional de avaliar o efeito de um "tratamento" consiste em:

1. *quantificar o output educacional* que desejamos monitorar (e.g., desempenho, interesse, crenças, motivação), garantindo a validade dessa quantificação;
2. *distribuir aleatoriamente* os participantes da pesquisa em dois grupos;
3. *aplicar o "tratamento"* a um grupo, negando-o ao outro, que será chamado "grupo de controle";
4. *comparar a variação média* do *output* educacional entre os grupos por meio de um teste estatístico adequado.

Embora raramente seja empregado, o desenho investigativo anterior funciona como uma porta de entrada ou um modelo de referência para pensarmos procedimentos de pesquisa em educação mais viáveis ou mais interessantes. Da maneira como está descrito, ele pode ser considerado um exemplo de *desenho/design experimental* (vide lição 1), pois há uma distribuição aleatória dos participantes nas turmas e há um grupo de controle bem definido. Em outras palavras, trata-se de um método randomizado e controlado. Para contornar o problema do efeito placebo educacional, seria necessário que o método fosse duplo-cego. Porém, esse tipo de investigação é um pouco difícil de imaginar na pesquisa educacional. Assim, a randomização e presença do grupo de controle costumam ser suficientes para configurar um *design* experimental. Por outro lado, se o controle e a randomização forem perdidos, o desenho perde a força de criar um vínculo causal entre o tratamento e seu respectivo efeito educacional.

Cumpre lembrar que há, na literatura, ótimas razões para desconfiar da validade ou da viabilidade das pesquisas que pretendem empregar um desenho experimental como esse:

- muitos pesquisadores quantificam *outputs* educacionais sem investigar a validade dessa quantificação, comprometendo os resultados da análise (vide lição 13);
- dificilmente o pesquisador tem controle sobre a alocação dos estudantes nas turmas e a falta de aleatoriedade desqualifica o desenho experimental;
- quando desenvolvemos um "tratamento" em que realmente acreditamos, costuma ser um pouco sofrido ou tumultuado negá-lo a alguns alunos com o propósito de constituir um grupo de controle;
- a própria necessidade de quantificar o *output* educacional em uma escala fidedigna limita a capacidade de apreender a experiência real dos estudantes, restringindo as conclusões que razoavelmente podem ser tiradas da pesquisa;[1]
- geralmente, os atores educacionais (professores e estudantes) sabem o que está sendo testado e, sem dúvida, sua atitude positiva ou negativa frente ao teste participa do resultado final (vide, por exemplo, o "efeito placebo" e a preferência por testes "duplo-cego" na pesquisa em saúde);

[1] Ainda que seja procedente, essa crítica costuma ser lida de forma equivocada. De fato, afirmações semelhantes costumam ser formuladas de maneira muito ingênua pelos investigadores que transformam a pesquisa qualitativa em uma bandeira pessoal, dando a entender que os métodos qualitativos conseguem fazer aquilo de que as quantificações jamais seriam capazes: *captar o sentido último da experiência humana*. Afinal, todos os métodos de pesquisa (quantitativos ou não) capturam a experiência de aprendizagem dentro de limites específicos. Ninguém está imune ao realismo ingênuo por meio do qual alguns pesquisadores escorregam do modelo da realidade para a realidade do modelo. Ninguém.

- resultados de pesquisa que adotam esse tipo de desenho correm o risco de serem considerados muito simplórios ou óbvios com relação ao esforço necessário para produzi-los.

Tácita ou explicitamente, muitos desenhos de pesquisa (quantitativos ou não) surgem como resposta aos problemas apontados no *design experimental* já descrito. Como posição pessoal, eu não considero desenhos experimentais preferíveis frente aos não experimentais e, por isso, sinto-me confortável em trabalhar com dados do tipo *survey* (i.e., levantamento) nas minhas pesquisas. Na verdade, questões que eu considero particularmente relevantes — desigualdades sociais e variações individuais frente à escola — não fazem muito sentido em desenhos de tipo experimental. Por outro lado, preciso reconhecer que esse tipo de desenho de pesquisa funciona, mesmo que negativamente, como um modelo de referência para aprendermos a argumentar a partir de evidência.

A minha geração cresceu orgulhosa de ter abandonado as famigeradas pesquisas do tipo pré-teste e pós-teste (cf. Lima Junior; Ostermann; Rezende, 2013). Contudo, nós sabíamos do que estávamos falando! Hoje, eu vejo muitos pesquisadores em educação que nunca tiveram real contato com esse tipo de abordagem e, sem saber realmente o que estão criticando, recuperam em suas análises o mesmo realismo ingênuo e os mesmos equívocos epistêmicos que costumam ser imputados aos defensores do *design* experimental. De fato, na formação do pesquisador, é fundamental conhecermos muito bem aquilo que desejamos não fazer justamente para não correr o risco de fazê-lo sem perceber! Nesse sentido, o desenho experimental participa negativamente de várias metodologias de pesquisa.

Importando os dados da análise

O Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep) disponibiliza livremente os dados desagregados e não identificados da maioria das suas pesquisas. Qualquer pessoa pode baixar os microdados disponíveis em: inep.gov.br/microdados. Para facilitar nosso trabalho, baixamos do Enem 2018. Esses dados foram importados para um *dataframe* do R e filtrados (algumas variáveis e células em branco foram eliminadas). Atribuímos nomes e rótulos às variáveis. Finalmente, os dados foram reordenados aleatoriamente e salvos no arquivo "`ENEM_exemplo.dat`" prontos para uso. Obtenha o resultado no *site* www.perspectivascriticas.com.br e importe dessa maneira:

```
library(Hmisc)
load("ENEM_exemplo.dat")
```

Os primeiros métodos que vamos apresentar nesta lição são muito simples e, em sua maioria, adequados a *amostras consideradas pequenas* (i.e., n < 30), embora possam ser empregados em amostras maiores. Para tentar reproduzir o cenário mais próximo ao usual quando do emprego desses testes em turmas de estudantes, proponho guardar no *dataframe* Enem somente as primeiras 50 linhas.

```
ENEM = ENEM[c(1:50), ]
```

Comparando a média em dois grupos

A maneira mais simples de comparar a média de uma variável escalar em dois grupos é publicar uma tabela. O método mais versátil capaz de

gerar uma tabela com quaisquer estatísticas descritivas emprega funções da biblioteca **`dplyr`** (instale-a, se ainda não o fez):

```
library(dplyr)

ENEM %>%
    group_by(Sexo) %>%
    summarise(mCN = mean(NotaCN),
              mCH = mean(NotaCH),
              mLC = mean(NotaLC),
              mMT = mean(NotaMT))

## # A tibble: 2 x 9
##   Sexo         mCN  mCH  mLC  mMT
## 1 “Masculino”  515  579  524  565
## 2 “Feminino “  489  580  538  545
```

No comando anterior, levamos o *dataframe* na função **`group_by`**, que o agrupa por uma variável categórica escolhida pelo usuário (no caso, `Sexo`). O resultado é inserido na função **`summarise`**, que calcula uma lista de estatísticas descritivas definidas pelo usuário.

Como é possível perceber, a pontuação média dos meninos no componente de ciências da natureza é maior que a das meninas. Porém, *é necessário realizar um conjunto de testes estatísticos para especificar com que probabilidade essa diferença que estamos observando é casual.* Em outras palavras, precisamos avaliar se essa diferença não pode ser atribuída ao acaso.

Identificando diferenças casuais

Primeiramente, por que razão precisamos considerar a possibilidade de que as diferenças observadas sejam devidas ao acaso? Não bastaria ler

a tabela anterior e perceber que os meninos têm um desempenho superior e, com isso, dar a questão por encerrada? Bem, em lições anteriores, aprendemos que quantidades avaliadas empiricamente não devem ser encaradas como entidades tipo-ponto, mas como intervalos (cf. lição 1). Em outras palavras, há sempre um conjunto de valores que, com boas razões, poderiam ser atribuídos àquilo que estamos medindo (Lima Junior; Da Silveira, 2011).

A partir de uma amostra de 50 observações, obtivemos uma pontuação média de 515 pontos em ciências da natureza, para os meninos, e 489 pontos, para as meninas. Porém, se tomarmos outros 50 estudantes aleatoriamente, essas médias provavelmente sofreriam uma variação expressiva devido ao processo de amostragem (cf. lição 3).

```
load(“ENEM_exemplo.dat”)
ENEM = ENEM[c(51:100), ]
```

Assim, a média dos meninos na prova de ciências da natureza não deve ser interpretada como um valor pontualmente definido (no caso, 515 pontos), mas como um intervalo de valores que razoavelmente podem ser atribuídos a essa média. A saber, a dispersão dos valores razoavelmente imputáveis à média de uma série de observações pode ser estimada por meio do *desvio padrão da média* (cf. lição 3):

$$s_{\bar{x}} = \frac{s_x}{\sqrt{n}}$$

Em palavras, o desvio padrão da média pode ser obtido pelo desvio padrão amostral dividido pela raiz quadrada da quantidade de elementos presentes na amostra. Com a função **`summarise`**, podemos gerar uma tabela com todas essas informações:

```
ENEM %>%
    group_by(Sexo) %>%
    summarise(mCN  = mean(NotaCN),
              sdCH = sd(NotaCN)/sqrt(n()),
              n    = n())

## # A tibble: 2 x 4
##   Sexo         mCN  sdCN     n
## 1 “Masculino”  515  22.7    18
## 2 “Feminino “  489  13.3    32
```

Se a média for normalmente distribuída, o intervalo definido pela média mais ou menos duas vezes o seu desvio padrão deve abranger 95% dos valores imputáveis a essa média (cf. lição 3):

$$\overline{x} \pm 2s_{\overline{x}}$$

Dito isso, podemos observar que a média do desempenho das meninas pertence ao intervalo da média dos meninos (de 468,6 a 560,4 pontos). Isso indica que, muito provavelmente, a diferença de desempenho entre meninos e meninas poderá ser atribuída ao acaso. Observe também que, como a largura do intervalo depende do inverso da raiz de *n*, o aumento do tamanho da amostra reduz a variabilidade da média (cf. lição 3), tornando nosso teste estatístico com maior capacidade de identificar diferenças, caso elas existam.

Essa mesma situação ocorre em outras pesquisas educacionais: os participantes compõem uma amostra de uma população mais abrangente. Quando os participantes são selecionados ao acaso, temos sempre alguma chance de que as diferenças entre as médias sejam devidas ao processo de amostragem e, portanto, não representem diferenças reais entre as populações das quais os grupos foram tomados. Observando esse exemplo, o leitor deve ser capaz de perceber que muitas pesquisas educacionais

encarregadas de vender soluções didáticas mais eficientes, mais motivadoras, mais desejáveis... não sobreviveriam ao mais simples teste estatístico.

É justamente para lidar com esses efeitos ilusórios em razão da amostragem que empregamos testes de significância estatística para a comparação de médias. A saber, *o principal objetivo de um teste de comparação entre médias é avaliar em que medida a diferença entre as médias pode ser atribuída ao acaso*.

Teste t de Student

O principal teste para comparar a média de uma variável escalar em dois grupos é o chamado *teste t de Student*. Para obter estatística do teste t, calculamos a diferença entre as médias de uma variável nos dois grupos e dividimos o resultado pelo desvio padrão dessa diferença.

$$t = \frac{\overline{x}_A - \overline{x}_B}{s_{\overline{x}_A - \overline{x}_B}}$$

O cálculo do denominador pode variar um pouco conforme a situação, mas a ideia fundamental do teste t é simples: ele avalia a *diferença padronizada das médias*. Outros testes serão construídos da mesma maneira, pois, como vimos anteriormente (cf. lição 3), a padronização ajuda a discernir quando as diferenças são significativas e quando são casuais.

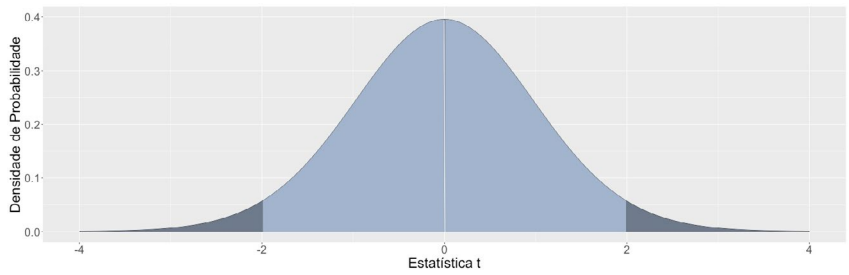
Dependendo do processo de medição, é possível que as observações de x estejam correlacionadas. Por exemplo, quando comparamos testes educacionais administrados às mesmas pessoas em dois momentos diferentes, é esperado que haja uma correlação entre as observações. Nesses casos, dizemos que as *observações são pareadas* (i.e., organizadas

em pares de medidas). Em outros casos, quando comparamos resultados de testes educacionais aplicados uma só vez a grupos diferentes (e.g., meninos e meninas), não faz sentido esperar essa correlação porque os próprios dados não podem ser organizados aos pares.

O teste t emprega uma distribuição estatística especial, denominada *distribuição-t*, capaz de produzir resultados consistentes mesmo em amostras muito pequenas (n < 10). O uso de amostras tão pequenas em pesquisa educacional é desaconselhável por uma questão de representatividade. Contudo, em algumas ciências experimentais, pode ser suficiente realizar três ou cinco repetições de uma medição. Em pesquisa educacional, as menores amostras costumam ser do tamanho de uma turma de 30 estudantes, mas estudos de padrão internacional tipicamente contam com alguns milhares de participantes. Em todos esses casos (*n > 30*), já não há diferença expressiva entre a distribuição-t e uma distribuição normal. Portanto, t maior ou igual a 2 geralmente indica uma diferença estatisticamente significativa entre as médias que estamos comparando.

Como é possível perceber (gráfico 4.1), a densidade de probabilidade de uma distribuição-*t* é semelhante à curva normal:

Gráfico 4.1: Distribuição-t com *gl = 30* (região significativa em destaque)



A saber, o teste t pode ser evocado com a função **`t.test`**:

```
Grupo1 = ENEM$NotaCN[ENEM$Sexo==”Masculino”]
Grupo2 = ENEM$NotaCN[ENEM$Sexo==”Feminino “]

t.test(Grupo1, Grupo2)

##
##  Welch Two Sample t-test
##
## data:  Grupo1 and Grupo2
## t = 0.99536, df = 28.805, p-value = 0.3279
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -27.59047  79.87589
## sample estimates:
## mean of x mean of y
##  515.0833  488.9406
```

Conforme esperado, a média da nota em ciências da natureza no grupo dos meninos é de 515 pontos, contra 488 pontos no grupo das meninas. Porém, com valor-p igual a 0,3279, o teste falhou em refutar a hipótese nula. Lembre-se que o valor-p é definido como a probabilidade de a hipótese nula ser verdadeira. Em outras palavras, a diferença observada entre as médias de meninos e meninas na nossa amostra pode ser atribuída ao acaso.

As premissas do teste t

Testes estatísticos costumam estar baseados em *premissas* que precisam ser satisfeitas pelos dados para que as inferências sejam consistentes. Antes (ou depois) de testar a significância estatística de uma diferença entre médias, é preciso avaliar se as premissas do teste foram respeitadas.

Com relação a isso, é sempre útil distinguir duas classes de testes estatísticos (Crawley, 2005):

5. *testes paramétricos* são aqueles que, como ponto de partida, assumem pressupostos restritivos com relação aos parâmetros da distribuição e *exigem a realização de testes auxiliares*.
6. *testes não paramétricos* não partem de pressupostos restritivos, mas são geralmente *menos poderosos* (i.e., têm menor chance de detectar diferenças quando elas existem).

Quando os pressupostos dos testes paramétricos são desrespeitados, recomenda-se escolher uma alternativa não paramétrica. Porém, em várias situações, optamos por manter os testes paramétricos mesmo quando suas premissas são levemente violadas. De fato, os testes paramétricos mais usuais podem ser considerados *robustos*, ou seja, eles são capazes de produzir resultados confiáveis mesmo com suas premissas levemente violadas. Além disso, eles costumam ser mais *poderosos*, ou seja, eles têm maior capacidade de detectar diferenças onde elas existem. Robustos e poderosos, os testes paramétricos costumam ser a escolha preferida dos analistas. De qualquer maneira, *jamais podemos deixar de testar as premissas de um teste*.

O teste t está baseado em *duas premissas*:

- *normalidade*, i.e., a variável deve ser normalmente distribuída em sua população;
- *homoscedasticidade*, ou "igualdade de variâncias", supõe que as variâncias dos grupos sejam iguais em sua população.

A premissa de normalidade

Os dados disponíveis podem não ser rigorosamente normais, mas seu desvio da normalidade deve ser pequeno ou indetectável. A propósito, os testes de normalidade mais clássicos seguem a lógica comum aos demais testes de significância estatística: eles informam, no *valor-p* (cf. lição 1), a probabilidade com que a hipótese nula pode ser considerada verdadeira. Nos testes de normalidade, a hipótese nula geralmente afirma que *os dados são normalmente distribuídos em sua população*. Portanto, para prosseguir com o teste t, precisamos que a hipótese de normalidade não seja refutada.

O teste de normalidade mais tradicional é o de *Shapiro-Wilk* (Crawley, 2005) e está implementado na função **`shapiro.test`**:

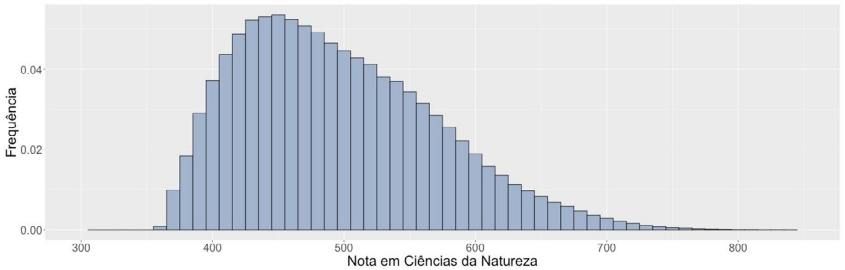
```
shapiro.test(ENEM$NotaCN)

##
##  Shapiro-Wilk normality test
##
## data:  ENEM$NotaCN
## W = 0.95918, p-value = 0.08213
```

Para os nossos dados, obtivemos um valor-p igual a 0,08. Portanto, declaramos que o teste falhou em refutar a hipótese de normalidade. Ou seja, as inferências feitas com o teste t não estão comprometidas.

Por outro lado, nós podemos investigar a distribuição de notas na população de origem e perceber que ela não é normal, mas bastante assimétrica. De fato, em quase todas as avaliações de matemática e ciências da natureza, as pontuações dos estudantes são distribuídas como renda: poucos com muito e muitos com pouco. No caso do Enem,

isso pode ser verificado gerando um histograma sem restringir nosso *dataframe* às 50 primeiras linhas (gráfico 4.2).

Gráfico 4.2: Histograma das notas de ciências da natureza



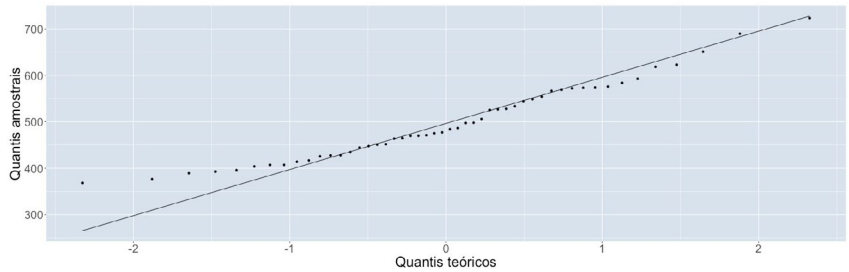
Assim, sendo a população de dados evidentemente não normal, por que o teste de Shapiro-Wilk falhou em detectar o desvio da normalidade? Bem, o segredo está no tamanho da amostra. Com amostras grandes, até mesmo as menores diferenças são detectáveis. De fato, se tentarmos medir o diâmetro de um tubo de PVC com um micrômetro, detectaremos diferenças ainda que pequeníssimas e pouco relevantes. Por outro lado, trabalhar com amostras muito pequenas é análogo a tentar medir o diâmetro de um fio de cabelo com uma régua: o resultado é tipicamente inconclusivo.

Para os testes de normalidade, isto é particularmente verdadeiro: a hipótese de normalidade é quase sempre refutada em amostras grandes. Sem haver normalidade da distribuição, ficamos impedidos de tirar conclusões a partir do teste t. Para contornar esse problema, a maioria dos analistas prefere avaliar a normalidade dos dados por meio de ferramentas menos sensíveis ao tamanho da amostra. Quando um desvio da normalidade é identificado, se ele não for considerado muito intenso, podemos nos apoiar na robustez dos testes paramétricos e não descartar o teste t.

A alternativa gráfica mais empregada para avaliar a normalidade de uma amostra é o *gráfico Q-Q*, i.e., "gráfico quantil-quantil" (Crawley, 2005). Ele plota, quantil a quantil, a distribuição da amostra contra uma distribuição normal teórica. Se os dados amostrais forem normalmente distribuídos, os pontos do gráfico Q-Q devem ficar alinhados. Figuras curvas em forma de S ou de U indicam desvios da normalidade. No R, esse gráfico pode ser gerado com as funções **`qqnorm`** e **`qqline`** (gráfico 4.3):

```
qqnorm(ENEM$NotaCN)
qqline(ENEM$NotaCN)
```

Gráfico 4.3: Gráfico Q-Q para distribuição normal



Observe que quase todos os pontos no gráfico 4.3 estão próximos à reta, exceto aqueles mais à esquerda da distribuição, sugerindo uma assimetria (cf. gráfico 4.2). Cabe ao analista, com base em sua experiência, avaliar se esse desvio não compromete os resultados do teste t.

A premissa de homoscedasticidade

Além da normalidade, o teste t presume que os dois grupos de dados cujas médias estão sendo comparadas vêm de populações com variâncias iguais. A igualdade de variâncias é chamada homoscedasticidade.

O *teste F de Fischer* para comparação de variâncias pode ser empregado nesse caso. Ele consiste em calcular a variância nas duas amostras, dividindo a maior pela menor.

$$F = \frac{s^2_{x_A}}{s^2_{x_B}} > 1$$

Se a variável *x* for normalmente distribuída (já testamos essa hipótese no passo anterior) e sua variância for igual nos dois grupos, a estatística acima segue uma *distribuição-F*. Portanto, o teste consiste em saber o quanto o valor observado para a razão das variâncias se afasta do centro da distribuição. Esse teste pode ser chamado pela função **`var.test`**:

```
var.test(Grupo1, Grupo2)

##
##  F test to compare two variances
##
## data:  Grupo1 and Grupo2
## F = 1.6398, num df = 17, denom df = 31, p-value = 0.2261
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.7324216 4.0904351
## sample estimates:
## ratio of variances
##           1.639819
```

A hipótese nula é sempre uma hipótese de igualdade. Portanto, no teste F, ela é exatamente a hipótese de variâncias iguais que desejamos corroborar. Observe que o valor-p ficou em 0,22. Portanto, *o teste*

falhou em refutar a hipótese nula e a premissa de homoscedasticidade não foi violada. Mais uma vez, esse resultado pode ser atribuído também ao tamanho da amostra. Com poucos milhares de observações, o teste provavelmente resultaria estatisticamente significativo.

Teste de Wilcoxon

Os testes paramétricos geralmente têm alternativas não paramétricas que podem ser evocadas quando as premissas do teste são violadas. Uma alternativa não paramétrica ao teste t é o chamado teste de Wilcoxon. Ele é baseado em uma *soma de rankings*. As duas amostras de dados são ordenadas em um *ranking*. Para cada grupo do teste, é calculada a soma dos *rankings*. O menor valor é comparado com as chamadas tabelas de Wilcoxon. Todo esse procedimento está incorporado à função **`wilcox.test`**:

```
wilcox.test(Grupo1, Grupo2)

##
##  Wilcoxon rank sum test
##
## data:  Grupo1 and Grupo2
## W = 328, p-value = 0.4284
## alternative hypothesis: true location shift is not equal to 0
```

Conforme esperado (com p = 0,42), o teste de Wilcoxon falhou em refutar a hipótese nula. Esse resultado era esperado porque o teste t é mais poderoso (ele tem mais chance de detectar diferenças quando elas existem). Se o teste t não detectou diferença estatisticamente significativa entre meninos e meninas, o teste de Wilcoxon dificilmente o fará.

Análise de variância

O teste t permite a comparação, em média, dos valores assumidos por uma variável escalar em dois grupos (definidos por uma variável categórica de dois níveis). Porém, o que podemos fazer quando nossa variável categórica tem mais níveis e há vários grupos que precisam ser comparados simultaneamente? Por exemplo, você pode querer comparar os efeitos de diversas intervenções educacionais diferentes com um grupo de controle. Talvez você queira saber os efeitos da renda familiar (medida em diversos níveis) sobre o desempenho escolar. Nesses casos, a variável categórica do modelo apresenta mais de dois níveis.

Como aprendemos a comparar médias duas a duas, é tentador pensar a execução de vários testes t. Porém, há um problema grave aí. Todas as vezes que fazemos um teste, assumimos o risco de refutar a hipótese nula quando ela é verdadeira ou de reter a hipótese nula quando ela é falsa. Quando nossa análise compreende a realização de muitos testes independentes, o risco de cometermos pelo menos um erro desses aumenta progressivamente. Por isso, quando a situação requer fazer mais comparações, não fazemos vários testes, mas buscamos um teste adequado para avaliar as associações das variáveis em seu conjunto.

A análise de variância (ANOVA) é uma técnica intimamente relacionada ao teste t e pode ser aplicada a todas as situações em que desejamos investigar a associação de uma variável escalar com uma ou mais variáveis categóricas (com dois ou mais níveis cada uma). A forma de argumentar da análise de variância pode parecer um pouco surpreendente no começo. Em vez de comparar as médias diretamente, comparamos variâncias totais, entre grupos, intragrupos… Porém, uma vez compreendido o método, é possível demonstrar que o teste t não é mais que um caso particular da análise de variância.

Decompondo as somas quadráticas

Variâncias são compostas fundamentalmente por *somas quadráticas*. A análise de variância está baseada na ideia de que é possível decompor tais somas em parcelas. Considere agora que y seja a variável escalar (e.g., nota em ciências da natureza) cujos valores médios, em hipótese, devem estar relacionados a uma ou mais variáveis categóricas (e.g., renda familiar, sexo, cor, escolaridade dos pais…). *Nosso objetivo, portanto, é construir um modelo capaz de fazer boas predições sobre y em vista do pertencimento aos grupos definidos pelas variáveis categóricas*. Por exemplo, qual é o valor esperado (médio) da nota em ciências da natureza entre meninas de baixa renda, ou entre jovens negros filhos de pais altamente escolarizados? É possível que as diferenças entre esses valores médios sejam devidas ao acaso?

Designaremos com um acento circunflexo a melhor estimativa de y dadas as categorias do modelo. A diferença entre os valores *y* efetivamente observados e as predições *y-chapéu* são chamadas *erro* (ou resíduo) e serão representadas pela letra grega *epsilon*:

$$y = \hat{y} + \epsilon$$

Usando essa notação, podemos identificar três tipos de somas quadráticas (Crawley, 2005):

- soma quadrática total (ing., *total sum of squares* – TSS) é a soma dos desvios quadráticos dos valores observados de *y* com relação à sua média geral;

$$TSS = \sum_{i=1}^{n} (y_i - \overline{y})^2$$

- soma quadrática explicada (ing., *explained sum of squares* – ESS), também chamada soma quadrática dos efeitos, dos tratamentos ou soma quadrática *entre grupos* (ing., *between groups*), é a soma do quadrado da diferença entre as predições do modelo e a média geral.

$$ESS = \sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2$$

- *soma quadrática residual* (*residual sum of squares* – RSS), também chamada soma quadrática dos desvios ou soma quadrática *intragrupos* (ing. *within groups*), é a soma do quadrado da diferença entre os valores observados e os valores preditos (i.e., os erros ou resíduos):

$$RSS = \sum_{i=1}^{n} (\epsilon_i)^2$$

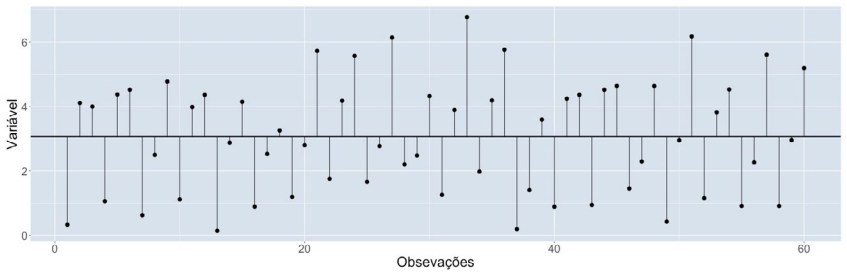
Feitas essas definições, é relativamente simples demonstrar que a soma quadrática total *TSS* pode ser decomposta em soma quadrática explicada *ESS* e soma quadrática residual *RSS* da seguinte maneira:

$$TSS = ESS + RSS$$

Das somas quadráticas ao teste

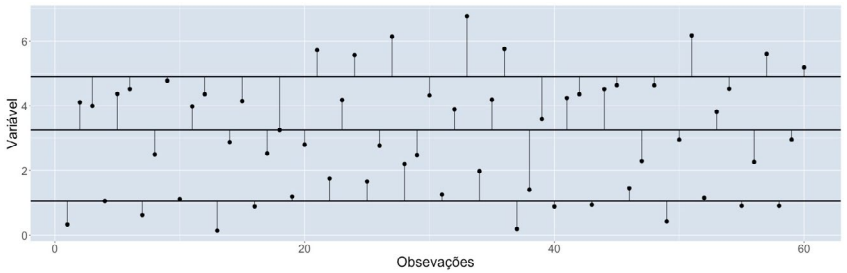
Para dar um sentido prático a essas somas quadráticas, vamos tentar visualizá-las em um gráfico. A seguir (gráfico 4.4), estão os valores assumidos pela variável escalar y. As linhas verticais indicam os desvios dos valores efetivamente observados até a média geral. Portanto, a soma do quadrado desses desvios produz TSS.

Gráfico 4.4: Desvios até a média geral



Como é possível perceber, os valores observados *y* são bastante dispersos. Por outro lado, acreditamos que eles estejam associados a grupos definidos por uma variável categórica de três níveis. Podemos, portanto, estimar a média de *y* em cada um desses grupos. Em seguida (gráfico 4.5), marcamos a distância entre os valores efetivamente observados e suas respectivas médias. A soma do quadrado dessas diferenças produz RSS.

Gráfico 4.5: Desvios até a média do grupo



Como é possível perceber, a soma residual RSS é visivelmente inferior à soma total TSS. Portanto, a soma quadrática explicada ESS deve ser grande, sugerindo que as categorias do modelo tenham um poder preditivo relevante sobre y. Se nosso modelo permitisse determinar os valores assumidos por y sem nenhum resíduo, teríamos RSS = 0 e ESS = TSS. Por outro

lado, se o modelo não fizer boas predições, então a soma residual RSS será aproximadamente igual a TSS (e a soma explicada será próxima de zero).

Para perceber como essas somas quadráticas se relacionam ao teste t de Student, considere o quadrado da estatística do teste:

$$t^2 = \frac{\left(\overline{y}_A - \overline{y}_B\right)^2}{s^2_{\overline{y}_A - \overline{y}_B}}$$

Como é possível perceber, o numerador dessa fração é o *quadrado de uma diferença entre grupos*. Portanto, deve estar relacionado à *soma quadrática explicada ESS*. Enquanto isso, o denominador pode ser obtido por uma composição de *variâncias avaliadas dentro dos grupos*. Portanto, deve estar relacionada à *soma quadrática residual RSS*. Não proponho fazer a demonstração aqui, mas cumpre saber que *o quadrado da estatística do teste t pode ser interpretado como uma razão entre variâncias*.

Definimos a *estatística-F* da ANOVA como a razão entre a variância explicada e a variância residual. Sob a hipótese nula, essa razão deve seguir a mesma *distribuição-F* que empregamos no teste de variância de Fischer. Lembre-se que, para obter uma variância, é preciso dividir a soma quadrática pela sua quantidade de graus de liberdade:

$$F = \frac{ESS \big/ gl_E}{RSS \big/ gl_R}$$

A saber, o valor de F, nos casos em que a variável categórica tem somente dois níveis, é *rigorosamente igual ao quadrado de t*:

$$F = t^2$$

Um exemplo de ANOVA

Considere que desejamos avaliar os efeitos do tipo de escola onde o estudante cursou o ensino médio (pública ou privada) sobre seu desempenho no componente de ciências da natureza. Como o sistema público atende a uma quantidade muito maior de cidadãos que o sistema privado, precisamos de muitas observações para que haja pelo menos 30 casos em cada nível da nossa variável explicativa:

```
load("ENEM_exemplo.dat")
ENEM = ENEM[c(1:400), ]
table(ENEM$EnsMed)

##
##      EscPub EscPub/Priv     EscPriv
##         301          33          66
```

Feito isso, a análise de variância pode ser chamada diretamente pela função **aov**. Observe que a expressão **ENEM$NotaCN ~ ENEM$EnsMed** informa que estamos tentando predizer os valores de **NotaCN** a partir da variável (categórica) **EnsMed**. O resultado da análise de variância pode ser chamado pela função **summary** conforme vemos a seguir:

```
fit = aov(ENEM$NotaCN ~ ENEM$EnsMed)
summary(fit)

##              Df  Sum Sq Mean Sq F value Pr(>F)
## ENEM$EnsMed   2  479520  239760      50 <2e-16 ***
## Residuals   397 1903772    4795
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Na *tabela da análise de variância* anterior, todas as quantidades que introduzimos são apresentadas. Observe que, primeiramente, a tabela publica o número de graus de liberdade das somas quadráticas. Como a amostra completa tem n = 400 observações, sabemos que a soma quadrática total *TSS* deve ter 399 graus de liberdade (*n–1*). De maneira semelhante, como os resíduos são calculados com relação às médias dos três grupos, *RSS* apresenta 397 graus de liberdade (*n–3*). Os graus de liberdade da soma quadrática explicada pela variável EnsMed deve ser igual à quantidade de níveis descontada a média geral (*3–1*).

A tabela também publica as somas quadráticas ESS e RSS, a razão dessas somas pelos respectivos graus de liberdade e a estatística F do teste. Comparando o resultado dessa estatística com a distribuição-F, percebemos que *estamos diante de uma diferença estatisticamente significativa com valor-p muito próximo de zero* (i.e., a hipótese nula deverá ser descartada). Porém o teste ainda não indica o que essa diferença significa.

Interpretando as diferenças

Sabendo que não estamos diante de uma casualidade, precisamos finalmente comparar a média da variável escalar nos diversos grupos do nosso modelo. Isso pode ser feito com funções da biblioteca **`dplyr`**:

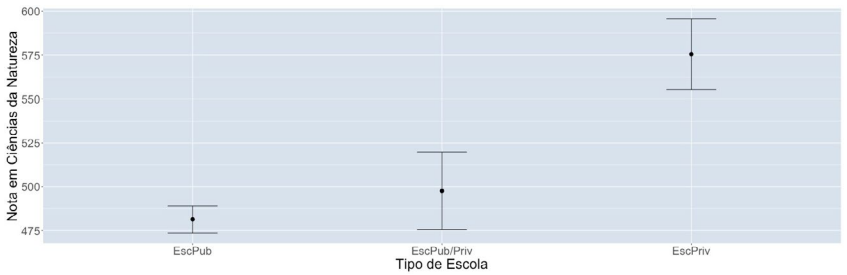
```
library(dplyr)


ENEM %>%
    group_by(EnsMed) %>%
    summarise(mCN  = mean(NotaCN),
              sdCN = sd(NotaCN)/sqrt(n()),
              n    = n())
```

```
## # A tibble: 3 x 4
##   EnsMed       mCN  sdCN    n
## 1 EscPub       481.  3.8  301
## 2 EscPub/Priv  498. 11.1   33
## 3 EscPriv      576. 10.0   66
```

Em princípio, a tabela anterior já permitiria tirar algumas conclusões. Entretanto, as médias da nota em ciências da natureza e seus respectivos desvios padrão podem ser visualizadas com mais clareza em um gráfico. Isso pode ser feito empregando a função **`plotmeans`** da biblioteca **`gplots`** (instale-a, se ainda não o fez). Os intervalos de confiança representados no gráfico 4.6 têm 95% de cobertura:

```
library(gplots)
plotmeans(ENEM$NotaCN ~ ENEM$EnsMed)
```

Gráfico 4.6: Médias de NotaCN por tipo de escola (intervalos de confiança a 95%)



A sobreposição das barras no gráfico 4.6 sugere que não será possível distinguir os efeitos de ter cursado o ensino médio completa ou parcialmente na escola pública. Os estudantes que cursaram o ensino médio completamente na rede privada, por outro lado, apresentam desempenho superior e essa diferença não pode ser atribuída ao acaso.

Uma maneira mais rigorosa de fazer essa avaliação é consultar o resultado da análise de variância usando a função **`summary.lm`**. Ela permite interpretar a análise como se fosse uma regressão linear:

```
summary.lm(fit)

##
## Call:
## aov(formula = ENEM$NotaCN ~ ENEM$EnsMed)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -173.130  -50.212   -9.912   48.518  241.688
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)              481.412      3.991 120.611   <2e-16 ***
## ENEM$EnsMedEscPub/Priv   16.224     12.698   1.278    0.202
## ENEM$EnsMedEscPriv       94.118      9.412  10.000   <2e-16 ***
## ---
## Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
##
## Residual standard error: 69.25 on 397 degrees of freedom
## Multiple R-squared:  0.2012, Adjusted R-squared:  0.1972
## F-statistic:    50 on 2 and 397 DF,  p-value: < 2.2e-16
```

Nessa forma de apresentação dos resultados, cada categoria da variável explicativa tem um coeficiente associado. Eles devem ser interpretados como variações produzidas na variável dependente y por variações nas categorias do modelo. As primeiras categorias de cada variável do modelo são tomadas por referência e não aparecem na tabela (no caso, `EnsMedEscPub`). O valor médio da variável escalar *y* nas categorias de referência corresponde ao `Intercept`.

O grupo de candidatos que estudou parte do ensino médio na escola pública e parte na escola privada (`EscPub/Priv`) tem, em média, *16,224* pontos a mais que o grupo de referência (`EscPub`), mas essa diferença não pode ser considerada estatisticamente significativa (*p=0,202*). O grupo que cursou o ensino médio integralmente na escola privada (`EscPriv`) apresenta, em média, *94,118* pontos a mais em NotaCN e essa diferença deve ser considerada estatisticamente significativa ($p < 2 \cdot 10^{-16}$).

Outra informação publicada é o *coeficiente de determinação* (R-quadrado). Ele é geralmente considerado uma medida da *qualidade do ajuste* e varia sempre entre 0 e 1. O coeficiente de determinação também pode ser interpretado como percentual da dispersão total dos dados explicada pelas categorias do modelo. Na análise de variância, o coeficiente de determinação é definido por:

$$R^2 \equiv \frac{ESS}{TSS}$$

Se o modelo especificasse perfeitamente os valores da variável escalar *y*, então R-quadrado seria igual a 100%. Em situações reais, isso não ocorre e a qualidade do ajuste assume valores intermediários.

Testando as premissas do modelo

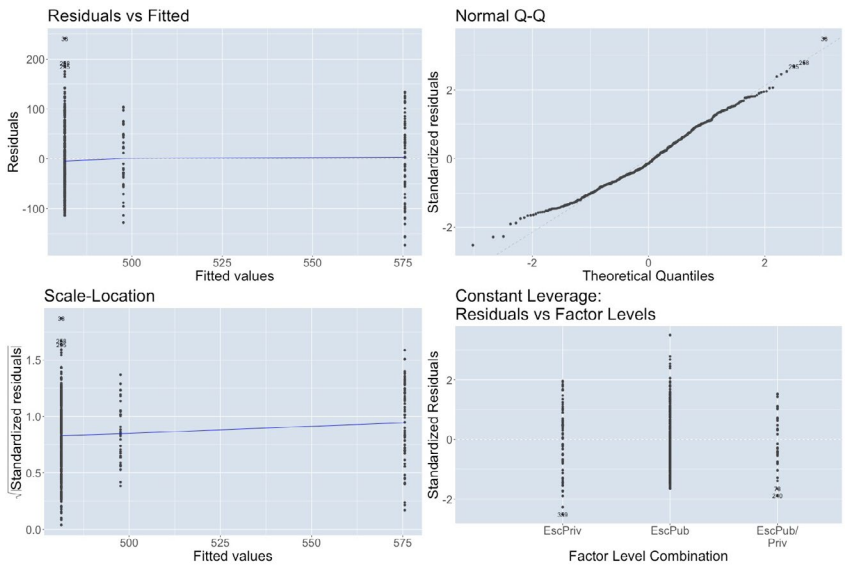
Como qualquer modelo paramétrico, a análise de variância está baseada em premissas que precisam ser testadas. A saber, as premissas são:

1. resíduos independentes;
2. resíduos normalmente distribuídos;
3. variâncias iguais intragrupo;
4. ausência de *outliers* alavancados.

Se inserirmos o modelo diretamente na função **`plot`**, ela gerará os chamados *gráficos diagnósticos*, que permitem avaliar em que medida as premissas do modelo são violadas. Para amostras maiores, preferimos analisar as premissas do modelo por meio de gráficos, pois os testes estatísticos que aprendemos no início desta lição (teste F de Fischer e teste de Shapiro-Wilk) são muito poderosos e quase sempre refutam as hipóteses de normalidade e igualdade das variâncias para amostras da ordem de centenas de observações. Como a análise de variância é capaz de produzir resultados consistentes mesmo quando suas premissas são levemente violadas, importa mais termos uma percepção visual e, a partir daí, decidir o que pode ser feito.

```
plot(fit)
```

Gráficos 4.7-4.10: Diagramas diagnósticos do modelo



Dos gráficos diagnósticos, os mais importantes para testar as premissas da análise de variância são o segundo e o terceiro. O segundo já foi discutido. Trata-se de um gráfico Q-Q e serve para avaliar a normalidade dos resíduos. O resultado esperado é um conjunto de pontos alinhados a uma reta. O terceiro gráfico é utilizado para avaliar a igualdade das variâncias dentro dos grupos. A linha colorida deve ser uma reta aproximadamente horizontal. Uma discussão mais pormenorizada desses gráficos será feita na próxima lição.

Revisando a lição

Nessa lição, você aprendeu a:

1. gerar tabelas com as funções **`group_by`** e **`summarise`**;
2. distinguir testes paramétricos e não paramétricos;
3. avaliar a normalidade de uma amostra com o *gráfico Q-Q* e com o *teste de Shapiro-Wilk*;
4. comparar as variâncias de duas amostras com o *teste F de Fischer*;
5. comparar médias com o *teste t de Student* e com o *teste de Wilcoxon*;
6. produzir uma análise de variância com a função **`aov`**;
7. extrair conclusões com as funções **`summary.aov`** e **`summary.lm`**;
8. comparar médias com gráficos gerados pela função **`plotmeans`**;
9. testar as premissas da análise de variância por meio de gráficos diagnósticos com a função **`plot`**.

Atividades propostas

- Tendo como base a mesma planilha do Enem, teste o efeito da escolaridade dos pais sobre o desempenho dos estudantes na prova de matemática.
- Compare o efeito da renda familiar sobre o desempenho dos estudantes nas quatro áreas de conhecimento do Enem. Para qual delas o efeito da renda é mais determinante?
- Escolha uma planilha que seja do seu interesse e, a partir dela, realize uma análise de variância testando o poder preditivo de uma variável categórica sobre uma variável escalar.

LIÇÃO 5.

CORRELAÇÃO E REGRESSÃO

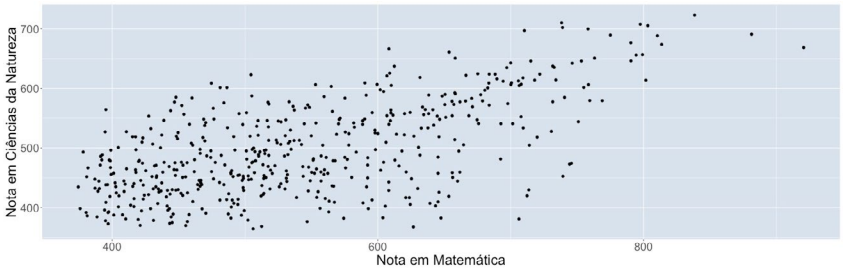
Na lição anterior, aprendemos a testar o poder preditivo de uma variável categórica de dois ou mais níveis sobre o comportamento médio de uma variável escalar. Nesta lição, daremos continuidade a esse debate, explicando como testar a relação preditiva ou associativa entre duas variáveis escalares. Considere novamente o *dataframe* Enem:

```
library(Hmisc)
load("ENEM_exemplo.dat")
ENEM = ENEM[c(1:500), ]
```

A maneira mais simples de visualizar a associação de duas variáveis é gerar um *diagrama de dispersão* (gráfico 5.1). Isso e muitas outras coisas podem ser feitas com a função **`plot`**:

```
plot(ENEM$NotaMT, ENEM$NotaCN)
```

Gráfico 5.1: Dispersão de NotaMT contra NotaCN (intervalos de confiança a 95%)



O resultado do diagrama de dispersão é uma nuvem de pontos. Você pode tentar circular essa nuvem e perceber que seu formato é levemente diagonal. Isso sugere que as notas de ciências da natureza e matemática tenham alguma associação. Ao mesmo tempo, os dados são bastante dispersos, mostrando que essa associação (se existir) *não é perceptível à escala individual*, ou seja, é possível encontrar diversos indivíduos com pontuação baixa em uma prova e desempenho relativamente elevado na outra. De fato, a análise estatística permite perceber relações que muito dificilmente seriam identificadas ao analisar experiências individuais.

Definindo covariância e correlação

A *covariância* entre duas variáveis escalares *x* e *y* é uma medida não padronizada do quanto a mudança em uma está associada à mudança da outra. Ela é definida como a *diferença entre a esperança do produto e o produto das esperanças:*

$$cov(x, y) = E(xy) - E(x)E(y)$$

Alternativamente, pode ser calculada como um somatório do produto dos desvios das variáveis dividido pela quantidade de graus de liberdade:

$$cov(x, y) = \sum_{i=1}^{n} \frac{(x - \overline{x})(y - \overline{y})}{gl}$$

Observe que, se os desvios positivos de uma variável estiverem associados a desvios positivos da outra, a covariância será positiva. Se desvios positivos estiverem associados a desvios negativos, a covariância será negativa. Se os desvios não tiverem relação alguma, a covariância tende a ser nula. No R, podemos calcular a covariância de duas variáveis escalares usando a função **cov**:

```
cov(ENEM$NotaCN, ENEM$NotaMT)

## [1] 5178.429
```

A saber, a covariância depende da escala em que as variáveis foram medidas. Para contornar esse problema e produzir uma medida adimensional de associação, podemos definir a *correlação de Pearson* como a *covariância padronizada*, ou seja, dividimos a covariância de *x* e *y* pelos seus desvios padrão:

$$cor(x, y) = \frac{cov(x, y)}{s_x \cdot s_y}$$

Há vários tipos de correlação além da correlação de Pearson. Elas serão apresentadas mais adiante (cf. lição 14). Até lá, quando eu me referir à correlação sem especificar o tipo, o leitor deve presumir que se trata da correlação de Pearson.

Definida dessa maneira, a correlação é uma quantidade *restrita ao intervalo de –1 a 1* e, portanto, não depende da escala das variáveis. Quando positiva, indica que variações de *x* e *y* têm o mesmo sentido. Quando negativa, indica sentidos opostos. Quando nula, indica que *x* e *y* não estão correlacionados.

A saber, podemos calcular a correlação de duas variáveis escalares usando a função **cor**:

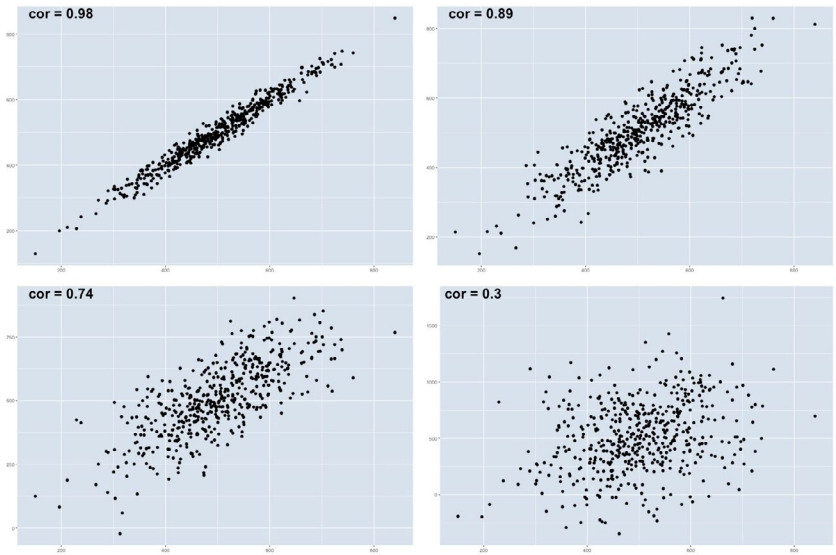
```
cor(ENEM$NotaCN, ENEM$NotaMT)

## [1] 0.620434
```

Gráficos de variáveis correlacionadas

É possível avaliar a correlação entre duas variáveis escalares observando seu diagrama de dispersão. Variáveis altamente correlacionadas são representadas por uma nuvem de pontos diagonal. Quando a correlação linear aumenta, a nuvem fica estreita, aproximando-se de uma reta. Quando a correlação linear diminui, essa nuvem se torna mais dispersa até o ponto em que seus contornos ficam arredondados. Uma nuvem não diagonal representa dados sem correlação.

Ao usar as funções **plot** e **rnorm**, podemos simular diagramas de dispersão para variáveis escalares com diferentes graus de associação visando perceber como a nuvem de pontos muda de formato. O resultado deve ser aproximadamente o seguinte:

Gráficos 5.2-5.5: Diagramas de dispersão com diversas correlações



Matriz de correlações

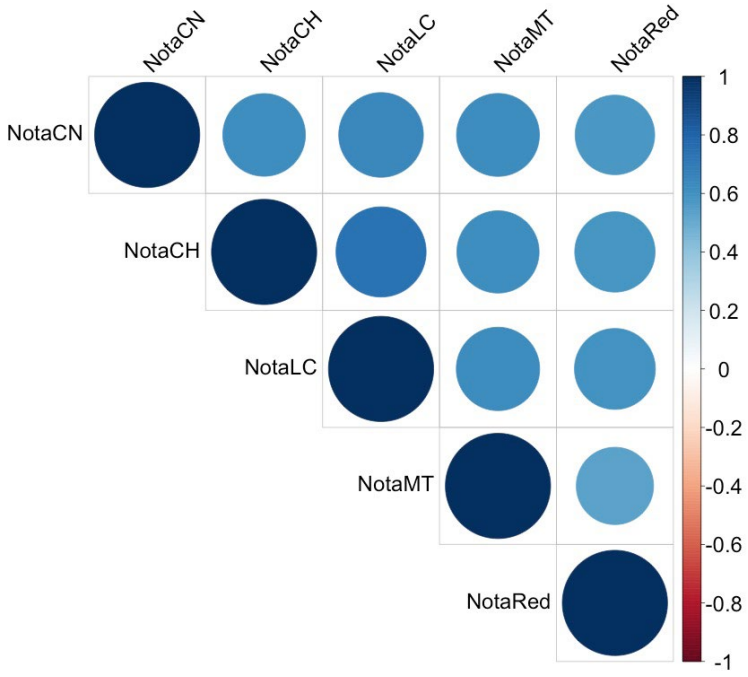
Se você quiser calcular todas as correlações entre pares de um conjunto de variáveis escalares, pode alimentar a função **`cor`** com mais colunas:

```
cor(ENEM[, 4:8])

##         NotaCN NotaCH NotaLC NotaMT NotaRed
## NotaCN    1.00   0.62   0.65   0.62    0.57
## NotaCH    0.62   1.00   0.73   0.61    0.58
## NotaLC    0.65   0.73   1.00   0.63    0.59
## NotaMT    0.62   0.61   0.63   1.00    0.54
## NotaRed   0.57   0.58   0.59   0.54    1.00
```

O resultado é conhecido como *matriz de correlaç*ões. Ela deve ser sempre simétrica e definida positiva, i.e., seu determinante é sempre maior que zero (para saber mais sobre álgebra matricial, cf. apêndice). Uma representação mais amigável e visual dessa matriz de correlações pode ser produzida usando a função **`corrplot`** da biblioteca homônima (instale-a, se ainda não o fez). Essa função deve ser alimentada com a matriz de correlações. A diretiva **`type`** determina se deve ser representada somente a parte superior da matriz (*upper*), inferior (*lower*) ou toda ela (*full*).

```
library(corrplot)
corrplot(cor(ENEM[, 4:8]), type = “upper”)
```

Diagrama 5.1: Representação da matriz de correlações



Interpretação geométrica

O desvio padrão e a correlação têm um comportamento geométrico muito interessante sobre o qual trataremos a seguir. Partindo da definição de variância, é possível demonstrar que a variância da soma pode ser decomposta em uma soma de variâncias da seguinte maneira:

$$s_{x+y}^2 = s_x^2 + s_y^2 + 2s_x s_y cor\left(x, y\right)$$

Observe que a expressão anterior é muito semelhante à *lei dos cossenos* aplicada à *soma de vetores*. Essa semelhança sugere a seguinte interpretação geométrica da correlação entre duas variáveis:

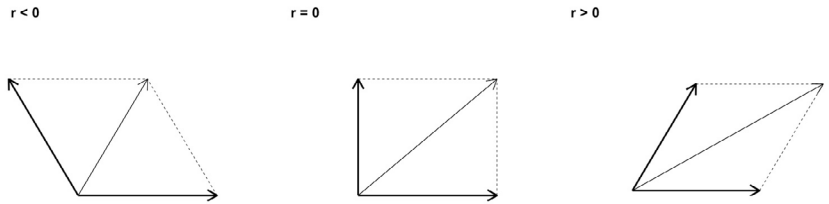
- variáveis escalares *x* e *y* podem ser representadas como vetores em um espaço;
- os módulos desses vetores são iguais aos desvios padrão dessas variáveis;
- a correlação entre variáveis é igual ao cosseno do ângulo entre seus vetores.

$$cor\left(x, y\right) = cos\left(\theta\right)$$

Há duas implicações mais imediatas dessa interpretação. Primeiramente, variáveis *não correlacionadas* são representadas como *vetores ortogonais* no espaço. Observe que vetores opostos são altamente correlacionados (com correlação negativa) enquanto vetores ortogonais representam variáveis sem nenhuma relação entre si. Ao mesmo tempo, qualquer *soma de variáveis* pode ser representada e visualizada como uma *soma vetorial*. A possibilidade de visualização espacial dos dados torna as análises muito mais intuitivas e abre caminho para pensamento de "alto nível". Todos os

teoremas e as ferramentas da álgebra linear podem ser explorados, aumentando nossa capacidade analítica. Voltaremos a essa questão posteriormente.

Diagrama 5.2: Representação vetorial de variáveis correlacionadas



No diagrama 5.2, pares de variáveis e suas respectivas somas são representados como vetores no espaço. Da esquerda para a direita temos: *i)* variáveis com correlação negativa formando um ângulo obtuso, *ii)* variáveis descorrelacionadas formando um ângulo reto; *iii)* variáveis com correlação positiva formando um ângulo agudo.

Testes de significância estatística

O fato de duas variáveis estarem correlacionadas em uma amostra não implica que as populações de dados de onde essas amostras foram extraídas estejam realmente correlacionadas. Para avaliar em que medida as correlações observadas (não) podem ser atribuídas ao acaso, precisamos empregar um teste de significância estatística para as correlações.

Vamos concentrar nossa discussão sobre o teste paramétrico de Pearson (Crawley, 2005). Quando as variáveis do teste são normalmente distribuídas, é possível demonstrar que, sob a hipótese nula, a correlação amostral r, em uma amostra de n elementos, está relacionada à estatística do teste t:

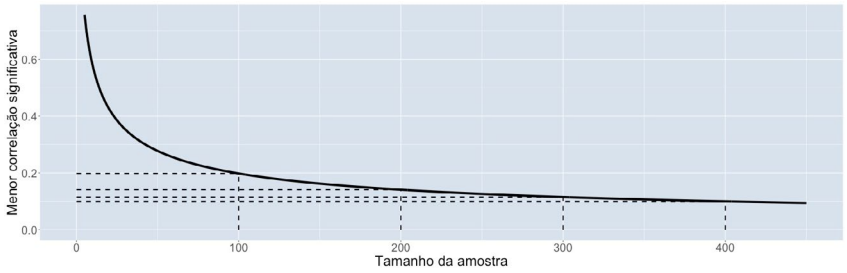
$$t = r\sqrt{\frac{n-2}{1-r^2}}$$

Sabendo que $t > 2$ tende a ser significativo em amostras grandes (i.e., com $n > 30$), podemos inverter a equação anterior e obter a correlação mínima, em função do tamanho da amostra, que não poderá ser atribuída ao acaso:

$$r(n) = \frac{t}{\sqrt{n-2+t^2}}$$

O gráfico 5.6 permite visualizar o comportamento da função *r(n)* para $t = 2$. Como é possível perceber, a sensibilidade do teste de correlação é altamente dependente do tamanho da amostra. Ela aumenta muito quando passamos de 50 para 100 ou 200 observações. Porém, quando já temos uma amostra numerosa, torná-la mais sensível pode ser muito dispendioso.

Gráfico 6.6: Menor correlação significativa por tamanho da amostra (com p<0,05)



O teste de correlação de Pearson descrito aqui pode ser chamado pela função **`cor.test`** a seguir:

```
cor.test(ENEM$NotaCN, ENEM$NotaMT, method = “pearson”)


##
##  Pearson’s product-moment correlation
##
## data:  ENEM$NotaCN and ENEM$NotaMT
## t = 17.654, df = 498, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.5633957 0.6715861
## sample estimates:
##      cor
## 0.620434
```

Como é possível perceber, com valor-p muito próximo de zero, o teste refutou a hipótese nula, i.e., *as notas em ciências da natureza e matemática estão correlacionadas*.

Regressão linear

A regressão linear é outra forma de avaliar a relação entre variáveis escalares. Se a correlação testa uma relação associativa, a regressão testa uma relação preditiva (cf. lição 1). Por agora, importa saber que a regressão linear busca a *reta que mais bem se ajusta aos pontos do diagrama de dispersão*. Enquanto a correlação trata as variáveis de forma simétrica (i.e, a correlação entre *x* e *y* é igual à correlação entre *y* e *x*), o mesmo não ocorre em modelos de regressão. Neles, é muito importante a distinção entre variáveis explicativas (independentes) e variáveis a explicar (dependentes).

Ainda que a análise estatística não seja capaz de estabelecer *causalidade* por si mesma (cf. discussão sobre *design experimental* nas lições 1 e 4), modelos de regressão são particularmente úteis quando o

analista considera, em hipótese, uma relação causal (portanto, assimétrica) entre as variáveis do estudo. A variável independente *x* é tratada como *preditora* da variável dependente *y*. A regressão permite avaliar o poder de predição de uma variável sobre a outra.

Para duas variáveis, a equação de regressão linear é simplesmente a seguinte:

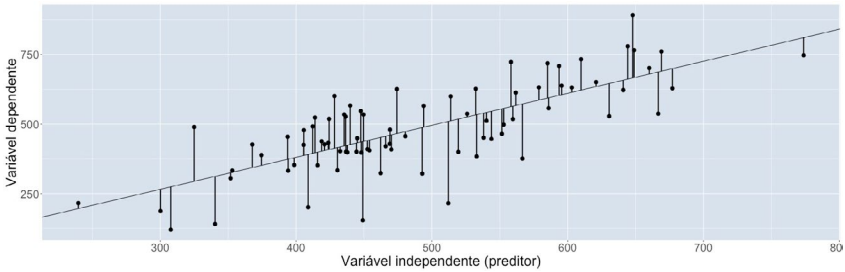
$$\hat{y} = b_1 \cdot x + b_0$$

O modelo anterior tem dois parâmetros ajustáveis: a *declividade* b_1 e o ponto de *interceptação* b_0. As *melhores estimativas* da variável dependente segundo o modelo são representadas por *y-chapéu*. A diferença entre os valores efetivamente observados de *y* e as melhores estimativas é chamada *resíduo*:

$$\epsilon = y - \hat{y}$$

Portanto, resíduos podem ser representados por distâncias verticais dos pontos à equação do modelo no diagrama de dispersão. Os valores estimados dos parâmetros ajustáveis devem *maximizar a verossimilhança* do modelo diante dos dados. Essa máxima verossimilhança, sob certas condições, corresponde à *minimização da soma dos resíduos quadráticos* (Fahrmeir *et al.*, 2013).

Gráfico 5.7: Representação dos resíduos em uma regressão linear



Aplicando a minimização da soma dos resíduos ao quadrado para cada parâmetro do modelo, é possível demonstrar que tais parâmetros podem ser estimados da seguinte maneira:

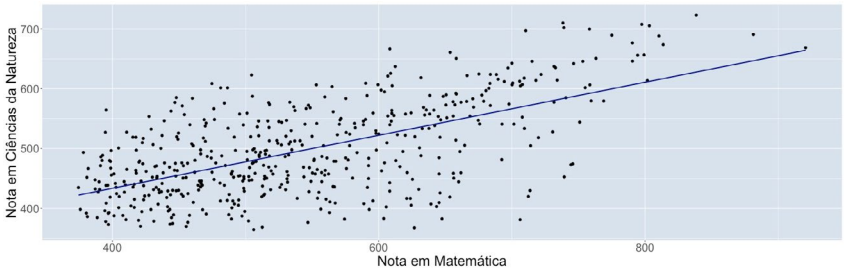
$$b_1 = \frac{cov(x, y)}{s_x^2}$$

$$b_0 = \overline{y} - b_1 \overline{x}$$

Regressões lineares podem ser solicitadas com a função **`lm`** e seus resultados podem ser consultados com a função **`summary`**. A seguir, geramos novamente o diagrama de dispersão com a função **`plot`** e traçamos a reta do ajuste com a função **`abline`**. Observe que a chamada da função **`lm`** é semelhante à análise de variância (cf. lição 4). Nela, usamos **`y ~ x`** para informar ao interpretador que ***x é preditor de y***.

```
x = ENEM$NotaMT
y = ENEM$NotaCN
fit = lm(y ~ x)
plot(x, y)
abline(fit)
summary(fit)
```

Gráfico 5.8: Regressão de ciências da natureza contra matemática (Enem 2018)




```
##
## Call:
## lm(formula = NotaCN ~ NotaMT, data = ENEM)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -188.321  -39.161    2.363   41.905  143.135
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 256.01694   13.93470   18.37   <2e-16 ***
## NotaMT        0.44351    0.02512   17.65   <2e-16 ***
## ---
## Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
##
## Residual standard error: 60.64 on 498 degrees of freedom
## Multiple R-squared:  0.3849, Adjusted R-squared:  0.3837
## F-statistic: 311.7 on 1 and 498 DF,  p-value: < 2.2e-16
```

Com valor-p muito próximo a zero, a regressão linear confirma que a relação entre NotaCN e NotaMT é estatisticamente significativa. Refutada a hipótese nula, a declividade é particularmente importante para a interpretação dos resultados. Ela indica como uma variação em x produz uma variação em y:

$$\delta y = b_1 \cdot \delta x$$

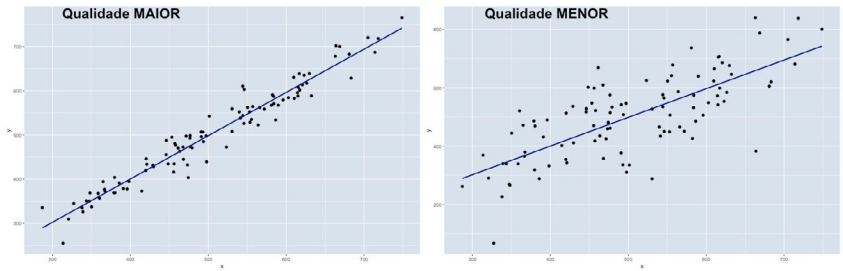
Ou seja, uma distância de 100 pontos na prova de ciências da natureza corresponde, em média, a uma distância de 44 pontos na prova de matemática.

Qualidade do ajuste

Uma vez que a hipótese nula tenha sido descartada e seja possível apontar uma relação entre as variáveis do modelo, surge outra questão importante: *qual é a qualidade do ajuste*? Perceba que qualidade do ajuste e significância estatística são coisas diferentes. Avaliar a significância consiste em decidir descartar ou não a hipótese nula; consiste em saber se a associação entre as variáveis pode ser atribuída ao acaso. Outra questão é, sabendo que estamos diante de uma associação não casual, avaliar o poder explicativo do modelo. Um modelo bem ajustado aos dados deve ter uma grande capacidade preditiva. Em outras palavras, inserindo as variáveis explicativas no modelo devemos ser capazes de predizer, com boa aproximação, os valores assumidos pela variável a explicar.

No diagrama de dispersão, a reta de um modelo bem ajustado aos dados passa próximo à maioria dos pontos do diagrama. Quando o ajuste tem menos qualidade, os pontos do diagrama ficam mais dispersos em torno da reta.

Gráficos 5.9-5.10: Comparando a qualidade do ajuste em dois casos



Tal como na análise de variância, a qualidade do ajuste da regressão linear pode ser avaliada pelo *coeficiente de determinação* (cf. lição 4). Conforme sabemos, esse coeficiente está definido pela razão entre a soma de quadrados explicada (ESS) e a soma de quadrados total (TSS). Porém, no caso da regressão linear, essa definição é equivalente à razão da variância das predições $\hat{y}$ pela variância dos valores efetivamente observados em y:

$$R^2 = \frac{var\left(\hat{y}\right)}{var\left(y\right)}$$

Se as predições do modelo coincidem exatamente com as observações, as variâncias serão iguais e a razão entre elas resultará em 100%. Por outro lado, se os pontos ficam mais dispersos em torno do modelo, a variância das predições será progressivamente menor. No limite em que a capacidade preditiva do modelo é nula, a variância das predições será igual a zero e a razão entre as duas também. Sendo assim, *podemos interpretar o coeficiente de determinação como a fração da variância de y que é explicada pelo modelo*.

Verificação do modelo

Assim como em qualquer modelo paramétrico, os pressupostos que fundamentam a regressão linear precisam ser investigados. A saber, esses são os quatro pressupostos que precisamos verificar (Crawley, 2005; Fahrmeir *et al.*, 2013):

1. linearidade – a relação entre *x e y* é linear, i.e., os resíduos são aleatórios;
2. normalidade – os resíduos padronizados de *y* são normalmente distribuídos;
3. homoscedasticidade (ou igualdade das variâncias) – a variância dos resíduos de *y* não depende de *x*;
4. alavancagem (ing. *leverage*) – a regressão não foi influenciada por *outliers* alavancados.

É importante saber que esses quatro pressupostos quase nunca são verdadeiramente respeitados em situações reais de investigação. Nos casos em que eles são levemente transgredidos, podemos lembrar que modelos de regressão são robustos (i.e., capazes de produzir resultados consistentes mesmo em condições um pouco adversas). Justamente por ser a regressão linear um método robusto, não costumamos fazer testes rigorosos de suas premissas. Por exemplo, sabemos que o teste de Shapiro-Wilk quase sempre concluirá que estamos diante de uma distribuição não normal quando o número de observações for suficientemente elevado. Então não se trata de saber se a distribuição dos resíduos é normal ou não, mas de avaliar quão expressivo é seu desvio da normalidade. Assim, em vez de testar muito rigorosamente as premissas do modelo de regressão, damos preferência às análises gráficas, as quais nos permitem ponderar, a partir de nossa experiência, as situações em que os desvios não comprometem

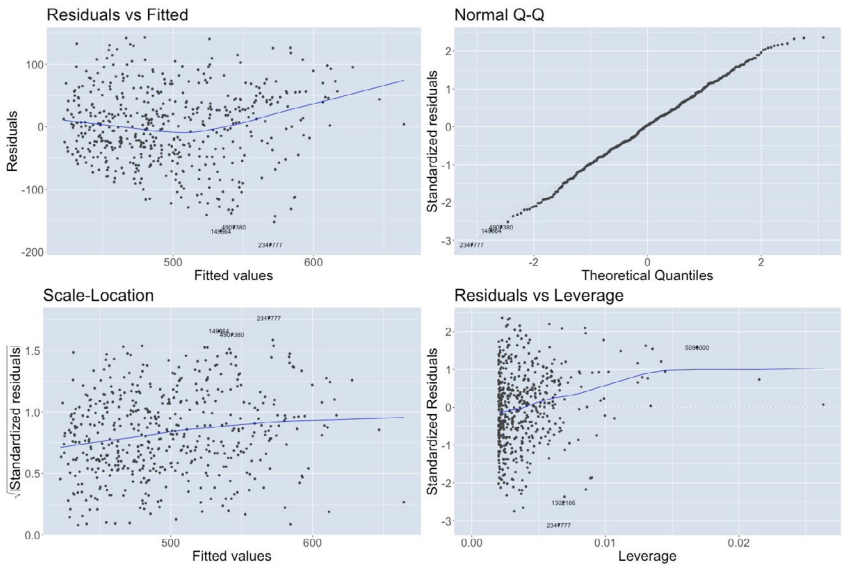
as inferências. O que você jamais pode fazer é ignorar essa etapa de verificação do modelo.

Os quatro gráficos usualmente consultados podem ser produzidos pela função **`plot`** a seguir:

```
plot(fit)
```

Os gráficos estão organizados na ordem das premissas que desejamos avaliar (gráficos 5.11-5.14).

Gráficos 5.11-5.14: Diagramas diagnósticos do modelo



1. O primeiro gráfico (acima e à esquerda) permite avaliar a linearidade da relação entre *x e y.* Como ele apresenta os resíduos em função dos valores ajustados, esperamos uma nuvem descorrelacionada, sem nenhuma tendência ou padrão identificável. A linha colorida representa a média móvel dessa nuvem de dados e deve

ser aproximadamente horizontal. *Linhas curvas ou não horizontais sugerem que os dados não se ajustam bem ao modelo*.

2. O segundo gráfico (acima e à direita) permite avaliar a premissa de normalidade dos resíduos. Ele é o popular diagrama Q-Q (cf. lição 4). Se a distribuição dos resíduos for normal, os pontos do diagrama repousam sobre uma reta. *Linhas curvas sugerem uma distribuição não normal dos resíduos*.
3. O terceiro gráfico (abaixo e à esquerda) permite avaliar a homoscedasticidade. Ele publica a raiz quadrada do módulo dos resíduos contra os valores ajustados. A linha colorida também passa pela média dos valores em cada segmento do gráfico. Ela deve ser aproximadamente horizontal. *Linhas curvas ou muito inclinadas indicam heteroscedasticidade* (i.e., variâncias diferentes).
4. O quarto gráfico (abaixo e à direita) permite avaliar os efeitos de alavancagem (ing. *leverage*). A saber, a alavancagem é uma medida do quanto um ponto é capaz de influenciar, sozinho, o resultado do ajuste. Na medida em que a alavancagem aumenta, os resíduos devem estar aleatoriamente distribuídos. Portanto, *a curva colorida no gráfico de alavancagem deve ser uma reta aproximadamente horizontal*. Além disso, o gráfico de alavancagem publica a chamada "distância de Cook", que nos permite desconfiar da fidedignidade de algumas medidas. *Pontos posicionados além da distância de Cook poderão ser eliminados pelo analista mediante justificativa*.

Os gráficos diagnósticos parecem indicar que o modelo construído até aqui apresenta um desvio da linearidade, pois a linha colorida não é uma reta horizontal no primeiro e no quarto gráfico. Os resíduos dos estudantes com nota mais elevada são todos positivos. Ou seja, todos os estudantes de alto desempenho têm nota maior que a prevista pelo

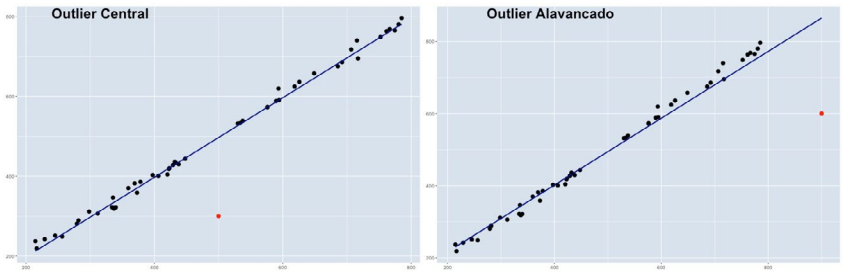
nosso modelo. Esse problema não pode ser ignorado, mas deixo para o leitor imaginar uma maneira de resolvê-lo.

Em geral, problemas identificados nos gráficos diagnósticos podem ser contornados: *i)* aplicando uma transformação aos dados — por exemplo, calculando o quadrado ou o logaritmo de x; ou *ii)* eliminando as linhas da planilha que correspondem aos *outliers* alavancados, caso existam. Intervenções dessa natureza, se bem conduzidas, podem corrigir a falta de linearidade, normalidade e igualdade das variâncias, mas precisarão ser devidamente interpretadas e justificadas.

Interpretando a alavancagem

De todos os gráficos diagnósticos, o único realmente novo até aqui é o diagrama de *alavancagem* e, por essa razão, ele merece algumas palavras. Observar a alavancagem dos pontos é importante porque eles não influenciam a solução da mesma maneira. Alguns pontos são mais influentes que outros e, por isso, são ditos "alavancados". Eles costumam estar posicionados mais à direita ou mais à esquerda no diagrama de dispersão. Os gráficos 5.15 e 5.16 ilustram por que pontos muito afastados se tornam preocupantes quando estão alavancados. No gráfico à esquerda, há um ponto vermelho claramente afastado dos demais. Ele pode ser considerado um *outlier*, mas está posicionado na região central, o que diminui sua alavancagem. Observe que a reta do gráfico 5.15 está bem ajustada a todos os demais pontos do diagrama, mostrando que a presença do ponto vermelho é pouco relevante. Já no diagrama à direita (gráfico 5.16), o ponto vermelho aparece numa região mais alavancada. Como é possível perceber, a regressão linear produziu uma reta que não se ajusta tão bem aos dados e o responsável por isso é o ponto vermelho, um *outlier* alavancado.

Gráficos 5.15-5.16: Outliers centrais e alavancados



Uma visão panorâmica

Nesta lição, exploramos técnicas que permitem avaliar a relação (preditiva ou associativa) entre duas variáveis escalares quaisquer. Para avaliar relações *preditivas* ($x \rightarrow y$, i.e., o efeito de uma variável independente sobre outra, dependente), recomenda-se empregar modelos de regressão. Por outro lado, se desejarmos avaliar *associações recíprocas* entre variáveis escalares (i.e., $x \leftrightarrow y$), faz mais sentido empregar correlações.

A escolha entre métodos preditivos ou associativos (cf. lição 1) fica a critério do analista, mas altera a maneira como os resultados serão interpretados. Métodos preditivos permitem falar sobre o efeito de x sobre y; métodos associativos não. Portanto, quando o analista tem em mente relações unidirecionais do tipo causa-e-efeito, os métodos preditivos são mais recomendados. Quando procura explorar relações estruturantes em um amálgama de variáveis escalares, a correlação é a fermenta mais adequada.

O leitor deve ter observado, também, que há muitas semelhanças entre a análise de variância (lição 4) e a regressão linear que apresentamos aqui. Seus gráficos diagnósticos, por exemplo, são os mesmos. Isso só é possível porque os dois métodos estão baseados nas mesmas premissas. De fato, em lições posteriores, perceberemos que tanto a regressão linear

quanto a análise de variância compõem um modelo linear mais geral (lição 10). A diferença entre as duas é que a análise de variância trabalha com preditores categóricos, enquanto a regressão linear trabalha com preditores escalares.

Revisando a lição

Nessa lição, você aprendeu a:

1. calcular e interpretar covariâncias e correlações (de Pearson) por meio das funções **`cov`** e **`cor`** — com destaque para a interpretação geométrica, que é muito importante;
2. testar a significância estatística da correlação entre duas séries de medidas;
3. produzir e analisar uma regressão linear usando as funções **`lm`**, **`summary`** e **`abline`**;
4. gerar e interpretar gráficos diagnósticos com a função **`plot`**;
5. perceber semelhanças e diferenças (conceituais e de cômputo) entre a regressão linear e a análise das correlações.

Atividades propostas

- Com relação à planilha do Enem, realize uma regressão linear e um teste de correlações com as notas de linguagens e códigos e redação;
- Tente resolver a falta de linearidade de modelo de regressão que apresentamos nesta lição.
- Escolha uma planilha que seja do seu interesse. Relativamente à planilha escolhida, realize uma regressão linear e um teste de correlações com um par de variáveis escalares à sua escolha.

LIÇÃO 6.

TABELAS DE CONTINGÊNCIA

Nas lições anteriores, nós aprendemos ferramentas para investigar:

- relações preditivas de uma variável categórica sobre uma variável escalar – ANOVA (lição 4);
- relações associativas entre variáveis escalares – correlação (lição 5);
- relações preditivas de uma variável escalar sobre outra – regressão linear (lição 5).

Passemos agora às ferramentas para investigar a *associação entre duas variáveis categóricas*. A saber, a forma mais tradicional de cruzar categorias de duas variáveis é a chamada *tabela de contingência*. Em linhas muito gerais, as tabelas de contingência dão continuidade ao que temos feito nas duas últimas lições. Em primeiro lugar, elas proporcionam uma maneira de identificar a relação entre duas variáveis. Contudo, antes de declararmos que as variáveis estão efetivamente associadas, precisamos realizar testes de significância estatística, que consistem em testar a hipótese nula (ou hipótese de não associação).

Por razões que ficarão claras a seguir, vamos tomar 1000 observações do *dataframe* Enem, visando simular uma situação típica do uso de tabelas de contingência:

```
library(Hmisc)
load("ENEM_exemplo.dat")
ENEM = ENEM[c(1:1000), ]
```

Em análises informadas pela sociologia da educação, é usual elaborar modelos estatísticos em que variáveis educacionais (desempenho em avaliações, escalas de interesse) são explicadas por variáveis que representam desigualdades sociais (geralmente codificadas como categorias). A interpretação adequada dos modelos mais elaborados depende também de saber que as variáveis explicativas podem estar fortemente associadas umas às outras. Nesta seção, vamos investigar a associação entre escolaridade do pai e escolaridade da mãe. Entender essa relação será muito útil nas lições seguintes.

Por uma questão de comodidade e visualização, vamos recodificar nossas variáveis de escolaridade para que apresentem menos níveis:

- "Alfb" designa os pais sem escolaridade ou simplesmente alfabetizados;
- "Fund" designa os pais que cursaram até o ensino fundamental (completo ou incompleto);
- "EMed" designa os pais que cursaram até o ensino médio ou técnico (completo ou incompleto);
- "Terc" designa os pais com educação terciária completa ou incompleta (tecnológico, graduação, pós-graduação)

Essa recodificação pode ser feita facilmente com a função **`mapvalues`** da biblioteca **`plyr`**. O resultado pode ser visualizado com a função **`describe`** da biblioteca **`Hmisc`**:

```
library(plyr)
ENEM$EscPai = mapvalues(ENEM$EscPai,
                        from = levels(ENEM$EscPai),
                        to = c("Alfb", "Alfb", "Fund", "Fund", "EMed",
"Terc", "Terc"))

ENEM$EscMae = mapvalues(ENEM$EscMae,
                        from = levels(ENEM$EscMae),
                        to = c("Alfb", "Alfb", "Fund", "Fund", "EMed",
"Terc", "Terc"))

describe(ENEM$EscPai)

## ENEM$EscPai
##        n  missing distinct
##     1000        0        4
##
## Value       Alfb  Fund  EMed  Terc
## Frequency    318   278   260   144
## Proportion 0.318 0.278 0.260 0.144

describe(ENEM$EscMae)

## ENEM$EscMae
##        n  missing distinct
##     1000        0        4
##
## Value       Alfb  Fund  EMed  Terc
## Frequency    203   265   333   199
## Proportion 0.203 0.265 0.333 0.199
```

As quantidades observadas

A forma mais usual de visualizar a associação de duas variáveis categóricas é por meio de uma tabela em que as linhas e colunas representam os níveis de cada variável. Tabelas simples com duas variáveis categóricas podem ser produzidas com a função **`table`**:

```
table(ENEM$EscPai, ENEM$EscMae)

##
##        Alfb Fund EMed Terc
##   Alfb  159   74   65   20
##   Fund   37  135   74   32
##   EMed    5   46  153   56
##   Terc    2   10   41   91
```

Você também pode calcular as ocorrências totais nas linhas e colunas usando as funções **`rowSums`** e **`colSums`**. Esses resultados podem ser adicionados à tabela gerada com a função **`cbind`**, que acrescenta uma coluna, e **`rbind`**, que acrescenta uma linha:

```
tab = table(ENEM$EscPai, ENEM$EscMae)
tab = cbind(tab, TOTAL = rowSums(tab))
tab = rbind(tab, TOTAL = colSums(tab))
tab

##       Alfb Fund EMed Terc TOTAL
## Alfb   159   74   65   20   318
## Fund    37  135   74   32   278
## EMed     5   46  153   56   260
## Terc     2   10   41   91   144
## TOTAL  203  265  333  199  1000
```

Nas tabulações anteriormente apresentadas, os valores publicados serão designados por O_{ij} e representam as *ocorrências observadas* no cruzamento da *i*-ésima linha com a *j*-ésima coluna. Assim sendo, percebemos que, na primeira linha, estão todos os filhos de pai sem escolaridade; na primeira coluna, os filhos de mãe sem escolaridade. Portanto, 159 pessoas têm o pai e a mãe simultaneamente sem escolaridade.

Observe que parece haver um acúmulo de casos junto à diagonal principal da tabela cruzada. Isso pode indicar uma associação das variáveis de tal maneira que pai e mãe tendam a apresentar escolaridades semelhantes. No entanto, antes de fazer essa afirmação, precisamos avaliar a probabilidade com que a associação entre as variáveis da tabela seja imputável ao acaso.

Para construir um teste estatístico que permita estimar o valor-p em questão (cf. lição 1), precisaremos comparar a quantidade de ocorrências observadas em cada célula (O_{ij}) com a quantidade esperada, supondo que as variáveis categóricas sejam independentes. As *quantidades esperadas* sob a hipótese de independência serão representas por E_{ij} . Portanto, o teste estatístico deverá avaliar as distâncias entre as quantidades observadas e esperadas (O_{ij} - E_{ij}) de maneira a estimar a probabilidade de que esse distanciamento tenha sido produzido ao acaso.

A distância do observado ao esperado

Os valores esperados (sob a hipótese nula) em uma tabela de contingência podem ser calculados facilmente. Se as variáveis categóricas da tabela não estiverem associadas, então as escolaridades dos pais são *eventos independentes*, tais como lançamentos sucessivos de uma moeda. De fato, quando jogamos uma moeda duas vezes, o segundo resultado não

depende do primeiro. Se a probabilidade de obter "cara" é de 50%, então a probabilidade de obter "cara" duas vezes é 50% vezes 50%. Portanto, se soubermos estimar as probabilidades separadas de ter uma mãe sem escolaridade e um pai sem escolaridade, a probabilidade de ter os dois nessa posição será o produto dessas duas probabilidades (sob a hipótese nula).

Tais probabilidades separadas podem ser estimadas a partir dos *valores totais* da tabela. Lembre-se que as colunas representam a escolaridade da mãe, 203 mães têm até a alfabetização. Em um total de 1.000 observações, isso permite estimar em 20,3% a probabilidade de ter uma mãe somente alfabetizada. Igualmente, a probabilidade de ter um pai alfabetizado ou sem escolaridade deve ser de aproximadamente 31,8%. Enfim, a melhor estimativa da probabilidade de ter os dois pais nessa posição deve ser igual a 20,3% vezes 31,8%. Esse resultado, multiplicado novamente pelo total geral de observações, produz a quantidade esperada de observações na primeira célula.

Em síntese, a quantidade de ocorrências esperadas em cada célula E_{ij} pode ser obtida a partir dos totais observados da seguinte maneira:

$$E_{ij} = \frac{T_i}{T} \cdot \frac{T_j}{T} \cdot T$$

Na equação anterior, T_i representa o total de ocorrências observadas na *i*-ésima linha; T_j indica o total de ocorrências observadas na *j*-ésima coluna e T representa o total geral. Os valores observados calculados dessa maneira estão disponíveis na tabela a seguir:

```
##           Alfb    Fund    EMed    Terc
##   Alfb  64.554  84.270 105.894  63.282
##   Fund  56.434  73.670  92.574  55.322
##   EMed  52.780  68.900  86.580  51.740
##   Terc  29.232  38.160  47.952  28.656
```

Para que a análise siga com consistência, é recomendável que os valores esperados em cada célula não sejam muito pequenos. Por tradição, a condição para prosseguir com a análise é que o menor valor esperado seja superior a 5 (Crawley, 2005). Na tabela anterior, como o menor valor esperado que calculamos é 28,7, podemos prosseguir normalmente. Se uma das células ficar com valor esperado muito baixo, será preciso recodificar as variáveis (agrupando seus níveis) ou aumentar o número total de observações.

Sabendo os valores esperados em cada célula, é possível calcular a diferença entre o observado e o esperado. Denominamos *resíduo padronizado* a diferença entre a quantidade de casos observados e a quantidade de casos esperados dividida pelo desvio padrão dessa diferença:

$$\varepsilon_{ij} = \frac{O_{ij} - E_{ij}}{\sqrt{E_{ij}}}$$

O resultado dos *resíduos padronizados* ε_{ij} está publicado na tabela a seguir:

```
##             Alfb       Fund       EMed       Terc
##   Alfb 11.7549826 -1.1187523 -3.9739639 -5.4408559
##   Fund -2.5869723  7.1454178 -1.9304605 -3.1355726
##   EMed -6.5767518 -2.7588359  7.1382225  0.5922382
##   Terc -5.0367486 -4.5585683 -1.0039369 11.6462707
```

Os resíduos apresentados geralmente são interpretados como observações de uma variável normalmente distribuída. Isso significa que resíduos superiores a 2,0 (ou inferiores a –2,0) marcam as células em que os valores observados estão significativamente desviados dos valores esperados (cf. lição 3). Há, portanto, uma sobrerrepresentação de casos ao longo da diagonal principal e uma sub-representação nas células mais afastadas.

Isso sugere que, de fato, as variáveis estejam associadas (com $p < 0,05$) de tal maneira que os genitores (pai e mãe) apresentem tipicamente a mesma escolaridade.

Apesar de o resultado do teste já estar bastante evidente, ainda precisamos calcular o valor-p. O somatório dos resíduos padronizados ao quadrado em toda a tabela produz a estatística do *teste chi-quadrado*:

$$\chi^2 = \sum_{ij} \frac{\left(O_{ij} - E_{ij}\right)^2}{E_{ij}}$$

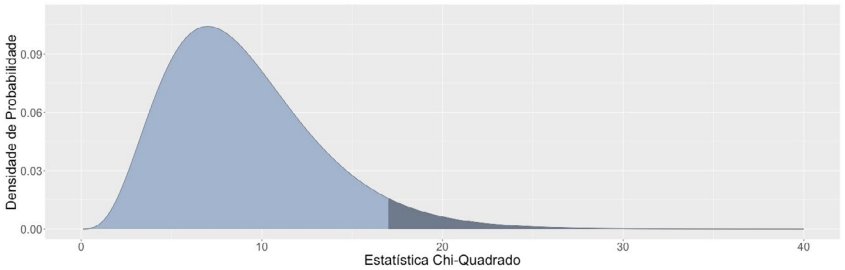
Se a hipótese nula for verdadeira e os resíduos forem normalmente distribuídos, a estatística anterior deve seguir uma distribuição de probabilidades homônima, chamada *distribuição chi-quadrado*. A saber, a quantidade de graus de liberdade dessa estatística pode ser obtida pelo produto da quantidade de linhas (menos 1 unidade) e colunas (menos 1 unidade). Isso ocorre porque, para cada linha e coluna, calculamos valores totais, que foram necessários ao cômputo da estatística do teste.

$$gl = \left(n_c - 1\right)\left(n_l - 1\right)$$

Na equação anterior, gl é a quantidade de graus de liberdade, n_c é a quantidade de colunas e n_l é a quantidade de linhas. No nosso exemplo, a tabela de contingência é 4 x 4 e tem, portanto, 9 graus de liberdade. A estatística chi-quadrado foi estimada em 541,1.

O gráfico 6.1 representa a densidade de probabilidade da distribuição chi-quadrado com 9 graus de liberdade. A região estatisticamente significativa foi colocada em destaque. Como é possível perceber, a estatística do teste está confortavelmente posicionada na região estatisticamente significativa. Portanto a hipótese nula deve ser refutada (com $p < 0,05$).

Gráfico 6.1: Distribuição chi-quadrado com *gl* = *9* (região significativa em destaque)



A tabela de contingência em um comando só

Uma tabela de contingência com todas as informações acima pode ser gerada com a função **`CrossTable`** da biblioteca **`gmodels`** (lembre-se de instalá-la, se ainda não o fez). Por padrão, essa função fornece informações que não precisamos agora e esconde outras informações que precisamos. Isso pode ser contornado com as diretivas da função. Veja o exemplo a seguir:

```
library(gmodels)
CrossTable(ENEM$EscPai, ENEM$EscMae,
           prop.r = F,
           prop.c = F,
           prop.t = F,
           prop.chisq = F,
           expected = T,
           sresid = T,
           format = “SPSS”)

##    Cell Contents
## |-------------------------|
## |                   Count |
## |         Expected Values |
## |            Std Residual |
## |-------------------------|
## Total Observations in Table:  1000
```

```
##
##              | ENEM$EscMae
##  ENEM$EscPai |     Alfb  |     Fund  |     EMed  |     Terc  | Row Total |
## -------------|-----------|-----------|-----------|-----------|-----------|
##         Alfb |      159  |       74  |       65  |       20  |      318  |
##              |   64.554  |   84.270  |  105.894  |   63.282  |           |
##              |   11.755  |   -1.119  |   -3.974  |   -5.441  |           |
## -------------|-----------|-----------|-----------|-----------|-----------|
##         Fund |       37  |      135  |       74  |       32  |      278  |
##              |   56.434  |   73.670  |   92.574  |   55.322  |           |
##              |   -2.587  |    7.145  |   -1.930  |   -3.136  |           |
## -------------|-----------|-----------|-----------|-----------|-----------|
##         EMed |        5  |       46  |      153  |       56  |      260  |
##              |   52.780  |   68.900  |   86.580  |   51.740  |           |
##              |   -6.577  |   -2.759  |    7.138  |    0.592  |           |
## -------------|-----------|-----------|-----------|-----------|-----------|
##         Terc |        2  |       10  |       41  |       91  |      144  |
##              |   29.232  |   38.160  |   47.952  |   28.656  |           |
##              |   -5.037  |   -4.559  |   -1.004  |   11.646  |           |
## -------------|-----------|-----------|-----------|-----------|-----------|
## Column Total |      203  |      265  |      333  |      199  |     1000  |
## -------------|-----------|-----------|-----------|-----------|-----------|
##
##
## Statistics for All Table Factors
##
##
## Pearson's Chi-squared test
## ------------------------------------------------------------
## Chi^2 =  541.0971     d.f. =  9     p =  9.017739e-111
##
##
##
##        Minimum expected frequency: 28.656
```

Nessa tabela de contingência, há três informações em cada célula:

- *Count* designa a contagem de casos observados O_{ij};
- *Expected Values* indica o valor esperado E_{ij};
- *Std Residual* é o resíduo padronizado.

Com um só comando é possível gerar a tabela e perceber que a associação entre escolaridade do pai e da mãe é estatisticamente significativa.

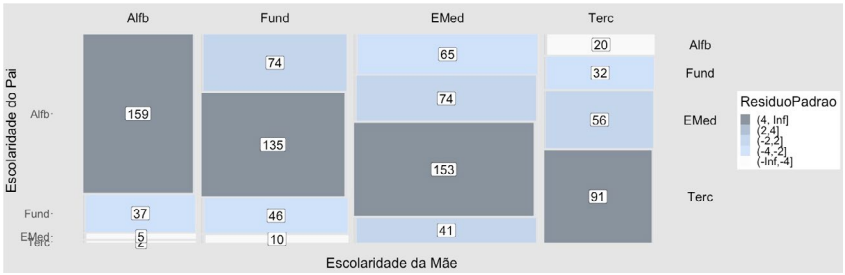
O diagrama mosaico

Em alguns casos, refutada a hipótese nula, é difícil entender a associação entre as variáveis. Para isso, os resíduos padronizados cumprem um papel fundamental. Nas células em que o desvio padronizado é igual ou superior a duas unidades (em módulo), podemos dizer que há uma associação estatisticamente significativa das categorias. Quando o desvio é positivo, há mais observações que o esperado; quando é negativo, há menos observações.

Na tabela de contingência que acabamos de gerar, os desvios padronizados são positivos na diagonal principal. Nas células mais afastadas, os desvios são expressivamente negativos. A distribuição dos desvios pode ser mais facilmente visualizada em um *diagrama mosaico*. Para tanto, é preciso gerar uma tabela com a função **`table`** e inseri-la na função **`mosaicplot`** usando a diretiva **`shade = T`**:

```
tab = table(ENEM$EscPai, ENEM$EscMae)
mosaicplot(tab, shade = T)
```

Diagrama 6.1: Mosaico da tabela de contingência
(escolaridade do pai contra escolaridade da mãe)



Nesse diagrama, cada retângulo representa uma célula da tabela de contingência. As áreas dos retângulos são proporcionais aos valores observados O_{ij} em cada célula. As cores representam os desvios padronizados. As células com desvio positivo marcam casos sobrerrepresentados e as células com desvio negativo, casos sub-representados.

Tudo indica que a união entre pessoas da mesma escolaridade é muito mais provável que a união de pessoas com escolaridades diferentes. Essa associação de discriminantes sociais na constituição de família ou outras formas de relação afetiva duradoura é chamada *endogamia das classes* (Bourdieu, 1984).

Revisando a lição

Nessa lição, você aprendeu a:

1. calcular todos os valores relevantes em uma tabela de contingência usando funções básicas como **`rowSums`**, **`colSums`**, **`cbind`**, **`rbind`**.
2. produzir uma tabela de contingência completa com o comando **`CrossTable`**(`gmodels`);
3. produzir um diagrama de mosaico com a função **`mosaicplot`** do pacote base.

Atividade proposta

- Com relação à planilha do Enem, quais variáveis categóricas estão associadas? Quais não estão? Produza tabelas de contingência para todas as variáveis que considerar relevantes e interprete o resultado.

PARTE III.

MODELAGEM ESTATÍSTICA

Aprender a analisar relações bivariadas é fundamental, mas a maioria das análises mais interessantes supõem construir e testar modelos de muitas variáveis. Diversos métodos de análise multivariada podem ser introduzidos como generalização dos métodos que estudamos até agora. A esse respeito, é importante distinguirmos três grandes grupos de modelos estatísticos tipicamente aplicados à pesquisa social:

1. modelo linear clássico (também chamado modelo linear geral);
2. modelo linear generalizado (GLM – *generalized linear model*);
3. modelo aditivo generalizado (GAM – *generalized additive model*).

A *análise de variância* e a *regressão linear simples* que estudamos nas lições 4 e 5 são versões bivariadas do *modelo linear clássico*, o qual abrange seis métodos estatísticos que costumam ser introduzidos separadamente (quadro III.1). Os nomes historicamente atribuídos a esses métodos variam conforme o tipo e a quantidade de variáveis empregadas. Em todas as suas versões, o modelo linear clássico supõe que a variável dependente seja escalar e normalmente distribuída. Ele também presume que haja uma relação linear direta entre os preditores (escalares ou categóricos) e a variável dependente. Em outras palavras, modelos lineares

clássicos estimam o valor esperado da variável dependente como uma combinação linear dos preditores. Por essa razão (i.e., por envolverem relações estritamente lineares), todas as versões do modelo linear clássico podem ser executadas e interpretadas segundo os mesmos princípios.

Quadro III.1: Versões do modelo linear clássico

	Dependente 1 variável escalar	**Dependentes Diversas variáveis escalares**
Preditores Todos escalares	Regressão simples ou múltipla	Regressão multivariada
Preditores Todos categóricos	Análise de variância (ANOVA)	Análise multivariada de variância (MANOVA)
Preditores Escalares e categóricos	Análise de covariância (ANCOVA)	Analise multivariada de covariância (MANCOVA)

É importante destacar que as relações lineares são amplamente adotadas na pesquisa educacional. Contudo, há situações em que não é possível ajustar um modelo linear clássico aos dados. Por exemplo, se sua variável dependente não for escalar, ela não poderá ser considerada normalmente distribuída sob nenhuma hipótese. Isso ocorrerá, por exemplo, nos casos em que a variável a ser explicada for do tipo binária (aprovado ou reprovado, diplomado ou evadido). Para tratar essas situações, recorremos ao *modelo linear generalizado (GLM)*, que inclui todas as versões do modelo linear clássico como caso particular.

No âmbito no modelo linear generalizado, o método mais popularmente empregado na pesquisa educacional é a regressão logística (cf. lição 9). Ela permite predizer o comportamento de variáveis binárias com base em preditores de qualquer tipo. Versões paramétricas da análise de sobrevivência, que é empregada no estudo da retenção na

educação superior, também podem ser incluídas ao lado dos demais modelos lineares generalizados.

A generalidade do GLM diante do modelo linear clássico está assentada em três razões:

1. *Não há restrição sobre o tipo de variável dependente*. No modelo linear clássico, a variável dependente é sempre escalar, enquanto, no modelo linear generalizado, variáveis categóricas podem ser incluídas como dependentes.
2. *A distribuição dos resíduos não precisa ser normal*. No modelo linear clássico, a distribuição dos resíduos é sempre presumida normal, já o modelo linear generalizado permite incluir variáveis que satisfazem outras distribuições (exponenciais, binomiais).
3. *A relação entre as variáveis não precisa ser estritamente linear*. Evidentemente, o modelo linear generalizado ainda opera com relações lineares. Porém, diversas relações não lineares podem ser linearizadas com as devidas transformações. O GLM incorpora uma *função de ligação*, responsável por associar os valores esperados da resposta aos preditores lineares do modelo.

Por tudo isso, técnicas substancialmente distintas podem ser localizadas no quadro teórico do modelo linear generalizado. Como todos os métodos que citamos aqui são semelhantes em algum sentido, é possível transferir conhecimento entre eles. A maneira como a análise é realizada e interpretada em cada versão do GLM segue os mesmos princípios gerais. Portanto, ao aprender um tipo de modelagem estatística, você terá mais facilidade de aprender todas as demais.

Observe que todos os métodos que constituem o GLM são paramétricos. Ou seja, todos presumem uma relação funcional linear (ou linearizável) entre as variáveis e uma distribuição específica para os resíduos

da variável dependente (mesmo que essa distribuição não seja normal). Quando a condição de linearidade não puder ser assumida, recomenda-se empregar um *modelo aditivo generalizado.* Esse nome é dado a uma classe de modelos que empregam funções suaves (semiparamétricas ou não paramétricas) para predizer o comportamento da variável dependente. Esses modelos têm chance de se ajustar melhor aos dados, mas costumam ser menos interpretáveis. Por ser pouco empregado na pesquisa educacional, o modelo aditivo generalizado não será discutido neste livro.

As próximas lições destinam-se justamente a discutir os modelos lineares mais usuais em pesquisa educacional.

- **Lição 7** (modelo linear clássico): explica como um modelo linear clássico é produzido e interpretado. Ela pode ser considerada uma expansão do que aprendemos sobre análise de variância (lição 4) e regressão linear simples (lição 5).
- **Lição 8** (modelos concorrentes): geralmente o processo de modelagem requer comparar modelos concorrentes, tomando decisões sobre quais variáveis devem ser mantidas ou eliminadas. Esta lição explica as ferramentas que temos à nossa disposição para tomar esse tipo de decisão.
- **Lição 9** (regressão logística): modelos de regressão logística diferem dos modelos clássicos porque sua variável dependente é binária. Eles são importantes para estimar a probabilidade de ocorrência de um evento (e.g., o acesso, a evasão ou a conclusão de um curso de graduação).

Modelos lineares são métodos quantitativos amplamente empregados na pesquisa educacional. Observe que *todos estão baseados em uma demarcação entre variáveis dependentes e independentes* (também chamadas preditores ou regressores). Ainda que o estabelecimento consistente

de uma relação causal ocorra somente por força de *design* experimental ou de interpretação teórica (i.e., jamais por conta da técnica estatística empregada), todas as variações do modelo linear generalizado são particularmente úteis para testar relações causais. Se você deseja testar o "efeito de A sobre B", a ferramenta ideal para sua investigação está nas próximas lições.

LIÇÃO 7.

MODELO LINEAR CLÁSSICO

Diversos fenômenos naturais e sociais podem ser modelados por meio de relações lineares. Ainda que modelos lineares sejam capazes de mascarar relações complexas (Conley, 2012), pesquisas em ciências humanas e sociais (incluída aí a pesquisa em educação) empregam modelos lineares com muito sucesso. De fato, esses modelos permitem produzir descrições muito precisas, orientando o pensamento crítico e a tomada de decisão.

Por exemplo, baseados em dados representativos da população estadunidense, Ready e Wright (2011) construíram um modelo linear para avaliar como os professores alfabetizadores antecipam as habilidades cognitivas dos seus estudantes de maneira a reproduzir preconceitos de classe, cor e gênero. Os autores descobriram que, das diferenças percebidas pelos professores, aproximadamente metade pode ser atribuída às diferenças reais entre grupos sociais. *Controlando os efeitos* da origem social dos estudantes, os autores perceberam também que professores que atuam em contextos periféricos subestimam mais frequentemente as habilidades de seus alunos. Resultados como esse expandem e detalham a crítica usual às categorias do juízo professoral (Bourdieu, 2007b).

Owston, York e Murtha (2013) também empregaram um modelo linear para mapear a percepção de 577 estudantes de graduação de

uma grande universidade estadunidense sobre cursos do tipo *blended learning* (i.e., cursos presenciais com uma porção substancial de atividade remota). Como resultado principal, eles perceberam que a opinião dos estudantes sobre essa estratégia de ensino está intimamente relacionada ao desempenho no curso: estudantes de alto desempenho tendem a avaliar positivamente as atividades remotas enquanto estudantes de baixo desempenho tendem a avaliá-las negativamente. Estudantes de alto desempenho consideram cursos parcialmente remotos mais engajadores, mais convenientes e acreditam aprender mais assim que em cenários presenciais. Como implicação mais imediata, esse resultado revela que a massificação das atividades remotas em instituições de ensino superior pode prejudicar preferencialmente os estudantes de baixo desempenho, que geralmente têm mais dificuldade para serem levados em consideração. De maneira mais ampla, esse resultado polemiza as avaliações de satisfação dos estudantes que não são estratificadas em função do seu desempenho e de outras variáveis relevantes — o que gera a impressão de que um curso foi globalmente bem-sucedido quando, na realidade, ele atendeu às necessidades de uma fração específica do corpo discente.

Após uma revisão ampla de como as desigualdades sociais perante a escola se transformam na medida em que os países expandem o acesso à educação superior, Mont'Alvão Neto e Mont'Alvão (2011) construíram um modelo linear para estimar as chances de um estudante brasileiro concluir o ensino médio e ingressar na universidade em função de algumas variáveis de origem social. Com dados referentes aos anos de 2001 a 2007, eles perceberam o recrudescimento das desigualdades sociais, a persistência e o aumento do impacto da origem social sobre o destino escolar dos estudantes brasileiros. Ainda que as universidades públicas

sejam mais igualitárias, elas têm mostrado um aumento progressivo de sua estratificação — o que corresponde à criação de posições periféricas em que estudantes de origem popular podem ser acomodados sem a perspectiva de usufruir dos privilégios historicamente reservados aos portadores de certificação terciária. Realizando uma regressão linear múltipla com dados do Enem (2012-2019), Lima Junior e Fraga Junior (2021) observaram que a origem social dos estudantes brasileiros é capaz de explicar mais de 35% do desempenho desses estudantes na prova de ciências da natureza. Portanto, uma parte expressiva do sucesso escolar individual está determinada pela posição da família e da escola na estrutura das relações de classe. Ao mesmo tempo, foi possível perceber que o efeito exclusivo da posição social da escola sobre o desempenho dos estudantes é maior que o efeito exclusivo da posição social de suas famílias. Em outras palavras, frequentar uma escola privilegiada é mais importante que nascer em uma família privilegiada. Esses resultados trazem implicações importantes para o enfrentamento da ideologia do mérito na educação em ciências.

Enfim, seria obviamente impossível exemplificar todos os usos dos modelos lineares na pesquisa em educação. Porém, os exemplos citados permitem perceber que esses modelos envolvem testar a relação entre variáveis que possam ser separadas em dependentes e independentes. É essa separação que permite apontar o efeito de uma variável sobre outra ($x \longrightarrow y$). Modelos lineares múltiplos também permitem avaliar o que aconteceria se os efeitos de uma variável fossem controlados. Por exemplo, dado que a origem social e a qualidade do ensino impactam o desempenho escolar, qual seria o efeito de práticas ou instituições de ensino consideradas efetivas quando a origem social é controlada? Perguntas dessa natureza são tipicamente realizadas no contexto dos modelos lineares.

Definição, extração e interpretação do modelo

O modelo linear clássico tem a seguinte forma geral:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon$$

Nela, y é a variável dependente, x_j é a j-ésima variável independente, $beta_j$ é o coeficiente associado a essa variável e épsilon é o resíduo da regressão. Considere que desejamos avaliar o efeito das provas de matemática e linguagens e códigos sobre a nota de ciências da natureza. Após padronizar as variáveis, usamos a função **`lm`** para gerar a seguinte regressão linear:

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

A regressão linear múltipla pode ser chamada pela mesma função **`lm`** da análise bivariada:

```
fit = lm(CN ~ MT + LC)
summary(fit)

##
## Call:
## lm(formula = CN ~ MT + LC)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.04177 -0.51595 -0.00646  0.52473  2.45659
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 6.236e-17  2.340e-02    0.00        1
```

```
## MT          3.295e-01  2.907e-02   11.34   <2e-16 ***
## LC          4.238e-01  2.907e-02   14.58   <2e-16 ***
## ---
## Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
##
## Residual standard error: 0.7399 on 997 degrees of freedom
## Multiple R-squared:  0.4537, Adjusted R-squared:  0.4526
## F-statistic:   414 on 2 and 997 DF,  p-value: < 2.2e-16
```

Em primeiro lugar, para interpretação de qualquer modelo linear, os coeficientes ajustados representam o efeito de uma variável explicativa *quando todas as demais variáveis são controladas*. Do *output* apresentado, observamos que o coeficiente relacionado à prova de matemática vale 0,3295. Em palavras, isso significa que:

- *controlando o efeito das demais variáveis*, cada desvio padrão em matemática (~100 pontos) corresponde a um aumento médio de 0,329 (~32,9 pontos) em ciências da natureza.

Observe que a declaração começa por “controlando o efeito de...”. Essa condição é fundamental para interpretar corretamente os coeficientes de regressão múltipla e não pode ser omitida da declaração do resultado. De fato, em discussões anteriores (cf. lição 5), testamos um modelo similar sem a nota de linguagens e códigos. O leitor deve perceber que o coeficiente que obtivemos aqui para a prova de matemática não é o mesmo. Isso ocorre porque, ao inserir uma variável nova no modelo linear, nós controlamos o efeito dessa variável na estimativa de todos os demais coeficientes. Portanto, os coeficientes da regressão linear múltipla não especificam a relação direta de uma variável x_i com a variável y, mas a relação dessas variáveis quando o efeito de todas as demais foi controlado.

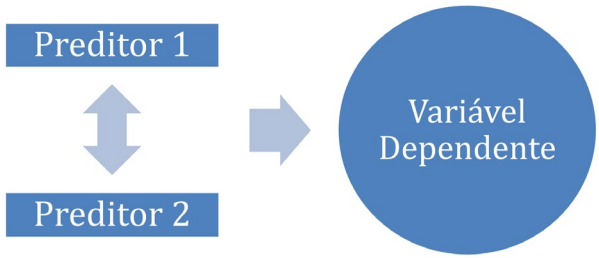
Esse “controle” do efeito de todas as demais variáveis é sublinhado quando o analista acrescenta a cláusula *coeteris paribus* à declaração do

resultado. A saber, essa expressão do latim quer dizer justamente "mantido todo o resto constante", ou "controlando os efeitos das demais variáveis".

Variáveis associadas e qualidade do ajuste

Em modelos multivariados, as variáveis explicativas costumam estar associadas entre si e essa associação precisa ser levada em consideração para interpretarmos corretamente os resultados do modelo. De fato, os coeficientes de regressão representam o efeito de um preditor "controlando" os efeitos de todos os demais. Porém, se os preditores estiverem associados, o que esse controle realmente quer dizer?

Diagrama 7.1: Representação de um modelo com preditores associados



Na seção anterior, aprendemos que os coeficientes de regressão, em vista da cláusula *coeteris paribus*, descrevem *efeitos exclusivos*, ou seja, efeitos que podem ser atribuídos à uma variável quando todas as outras foram controladas. O problema das variáveis associadas é que, como a variação de uma está relacionada à variação da outra, elas não podem ser realmente separadas.

De fato, variáveis associadas apresentam *efeitos comuns* por compartilharem a mesma informação. Tipicamente, isso ocorre quando desejamos avaliar os efeitos de preditores socioeconômicos (renda, ocupação

e escolaridade dos pais) para modelar o desempenho dos estudantes em avaliações educacionais. Porém, isso também pode ocorrer com muitas outras variáveis de interesse educacional. Afinal, medidas de motivação, desempenho, autoeficácia, inteligência estão geralmente associadas.

Como a escolaridade do pai está associada à escolaridade da mãe (cf. lição 6), podemos dizer que essas variáveis carregam, em parte, a mesma informação. Em outras palavras, a escolaridade da mãe já informa, aproximadamente, a escolaridade do pai! Essa situação requer o reconhecimento de que não é realmente possível avaliar o efeito de uma controlando a outra. No entanto, esse impasse pode ser contornado por meio da distinção dos efeitos das variáveis em *exclusivos* e *comuns*.

Para entendermos melhor o que quero dizer aqui, vamos observar o coeficiente de determinação R-quadrado, que pode ser interpretado como uma medida da qualidade do ajuste ou do *poder explicativo* do modelo (cf. lição 5). Considere, agora, a análise de variância (cf. lição 4) em que a escolaridade do pai e da mãe são tomadas como preditores da nota em ciências da natureza. A qualidade do ajuste pode ser extraída por meio da função **`summary.lm`**:

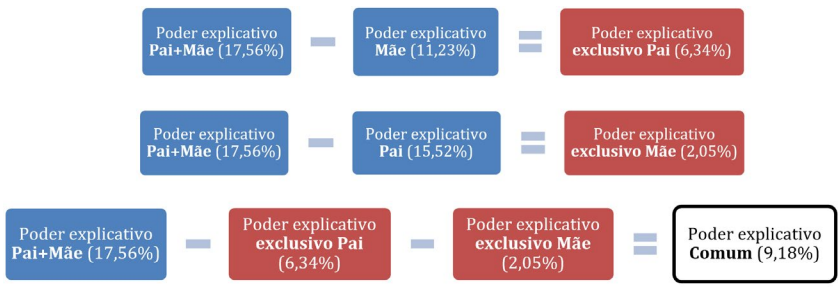
```
fit = aov(NotaCN ~ EscMae + EscPai, data = ENEM)
summary.aov(fit)

##              Df  Sum Sq Mean Sq F value   Pr(>F)
## EscMae        6  620444  103407   22.41  < 2e-16 ***
## EscPai        6  349996   58333   12.64 8.87e-14 ***
## Residuals   987 4553874    4614
## ---
## Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

summary.lm(fit)$r.squared

## [1] 0.1756671
```

Agora, vamos comparar esse resultado (17,57%) com os dois modelos bivariados que podemos gerar:

```
fit = aov(NotaCN ~ EscMae, data = ENEM)
summary.lm(fit)$r.squared## [1] 0.1123115

fit = aov(NotaCN ~ EscPai, data = ENEM)
summary.lm(fit)$r.squared

## [1] 0.1552172
```

Diferentemente dos coeficientes da regressão, o coeficiente de determinação R-quadrado, empregado para avaliar a qualidade do ajuste e o poder explicativo do modelo, tem a propriedade de aditividade (i.e., em certas situações, podemos somá-los e subtraí-los). Assim, faz sentido observar que *o poder explicativo do modelo completo* é menor que a soma das partes. Isso indica a presença de um efeito comum expressivo. Observe, também, que os poderes explicativos exclusivos e comuns podem ser obtidos por operações muito simples (diagrama 7.2):

Diagrama 7.2: Poder explicativo exclusivo e comum das variáveis do modelo

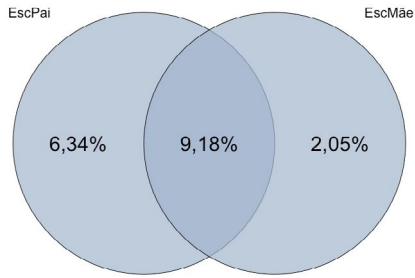


Os resultados dessas operações encontram-se representados no diagrama 7.3, proposto inicialmente por Silveira (1999). Nele, é possível perceber que as *variáveis mutualmente associadas compartilham poder explicativo* quando inseridas no mesmo modelo linear. Ter essa imagem em mente pode ajudar na interpretação correta da cláusula *coeteris paribus*.

Quando duas variáveis estão associadas, nós não conseguimos efetivamente controlá-las, pois a variação de uma produzirá necessariamente uma variação da outra. O que fazemos é *subtrair os efeitos comuns*. Assim, ao estimar o efeito de uma variável "controlando" todas as demais, nós estamos, na verdade, subtraindo parte do poder explicativo que poderia ser razoavelmente atribuído a essa variável. *Os coeficientes de regressão múltipla levam em consideração somente o efeito exclusivo de cada preditor*. Portanto, o "controle" *a posteriori* realizado pela regressão linear não é nada parecido com o controle de variáveis em um *design experimental*, que pretende estabelecer relações causais em sentido estrito (cf. lição 1).

Via de regra, o efeito comum de um conjunto de preditores mutuamente associados é muito mais interessante que seus efeitos individuais. Comparar renda, escolaridade e ocupação dos pais para decidir quem tem maior poder explicativo é, de fato, pouco relevante em vista da possibilidade de empregar todos esses indicadores em conjunto para construir uma indicador global da posição da família na estrutura das relações de classe à semelhança de Bourdieu ao definir o "volume de capital" (Bourdieu, 1984). Em momento oportuno (lição 10 e seguintes), vamos discutir como esse tipo de análise pode ser conduzida.

Diagrama 7.3: Repartição do poder explicativo entre variáveis associadas



O leitor deve atentar para o fato de que muitos marcadores sociais são equivocadamente tratados como se fossem isoláveis. Por exemplo, ao declarar que "o problema do Brasil não é o racismo, mas a pobreza" ou que "a escolaridade da mãe é mais importante que a escolaridade do pai para o desempenho escolar", estamos presumindo que os marcadores sociais possam ser testados em separado, o que é estritamente falso.

Interações entre variáveis explicativas

Associadas ou não, as variáveis explicativas do modelo podem ter *efeitos de interação*. Ou seja, além de seus efeitos principais, elas podem apresentar efeitos combinados, interseccionais. Isso ocorre, por exemplo, quando a combinação de estigmas sociais não corresponde à soma dos efeitos desses estigmas. Meninas pretas e pardas têm despertado muito a atenção dos pesquisadores no que diz respeito ao seu interesse pela ciência e carreiras científicas (Johnson *et al.*, 2011; King; Pringle, 2019; Morton; Parsons, 2018). A evidência indica que os desafios enfrentados por essas meninas são irredutíveis à sobreposição daquilo que se afirma genericamente sobre mulheres (brancas inclusive) e pessoas negras (homens

inclusive). *Esse efeito combinado e interseccional de duas variáveis é o que chamamos de interação*.

A introdução de termos de interação aumenta muito a quantidade de parâmetros ajustáveis do modelo. Por isso, vamos tomar uma quantidade maior de observações da planilha do Enem:

```
load(“ENEM_exemplo.dat”)
ENEM = ENEM[c(1:10000), ]
```

Como proposta, vamos testar um modelo para NotaCN com base na cor do estudante e no tipo de estabelecimento onde ele ou ela cursou o ensino médio. Antes de partir para a análise, investigamos se todas as categorias estão suficientemente ocupadas. Por haver poucos casos, os indígenas e amarelos foram removidos da análise:

```
table(ENEM$Cor, ENEM$EnsMed)

##
##            EscPub EscPub/Priv EscPriv
##   Branca     2273         483    1054
##   Preta      1002         134      86
##   Parda      3706         488     492
##   Amarela     142          27      44
##   Indigena     60           3       6

ENEM = ENEM[ENEM$Cor != “Indigena” & ENEM$Cor != “Amarela”, ]
ENEM = droplevels(ENEM)
```

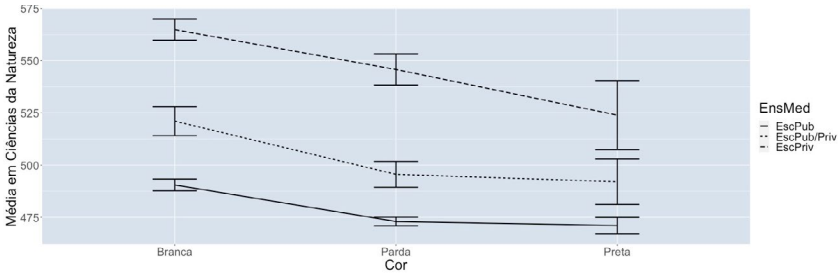
Os efeitos de interação podem ser facilmente visualizados com a função **`interaction.plot`**. Como entrada, ela recebe três vetores que comunicam: *i)* a variável categórica que será representada no eixo *x*;

ii) a variável categórica que será representada com traços no gráfico; *iii)* a variável escalar sobre a qual desejamos saber se há interação.

Se não houver interação, os traços do diagrama serão paralelos.

```
interaction.plot(ENEM$Cor, ENEM$EnsMed, ENEM$NotaCN)
```

Gráfico 7.1: Interação de EnsMed e Cor sobre NotaCN (intervalos de confiança a 95%)



Como é possível perceber, os traços não são paralelos e os termos de interação provavelmente serão significativos. No R, interações são representadas com um sinal de dois pontos (**`Cor:EnsMed`**). O modelo composto por efeitos principais e interações pode ser chamado somando a interação aos efeitos principais (**`Cor + EnsMed + Cor:EnsMed`**) ou, de forma simplificada, usando um asterisco (**`Cor*EnsMed`**):

```
fit = aov(NotaCN ~ Cor * EnsMed, data = ENEM)

summary.aov(fit)

##               Df   Sum Sq Mean Sq F value Pr(>F)
## Cor            2  2585870 1292935 275.471 <2e-16 ***
## EnsMed         2  6599346 3299673 703.023 <2e-16 ***
## Cor:EnsMed     4    46048   11512   2.453 0.0438 *
## Residuals   9709 45569662    4694
## ---
```

```
## Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

summary.lm(fit)

##
## Call:
## aov(formula = NotaCN ~ Cor * EnsMed, data = ENEM)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -472.98  -51.18   -5.98   47.03  306.72
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  490.504      1.437 341.344  < 2e-16 ***
## CorPreta                     -19.456      2.598  -7.489 7.53e-14 ***
## CorParda                     -17.527      1.825  -9.603  < 2e-16 ***
## EnsMedEscPub/Priv             30.540      3.433   8.897  < 2e-16 ***
## EnsMedEscPriv                 74.355      2.553  29.124  < 2e-16 ***
## CorPreta:EnsMedEscPub/Priv    -9.577      7.176  -1.335  0.18203
## CorParda:EnsMedEscPub/Priv    -7.958      4.761  -1.672  0.09464 .
## CorPreta:EnsMedEscPriv       -21.481      8.110  -2.649  0.00809 **
## CorParda:EnsMedEscPriv        -1.547      4.162  -0.372  0.71021
## ---
## Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
##
## Residual standard error: 68.51 on 9709 degrees of freedom
## Multiple R-squared:  0.1685, Adjusted R-squared:  0.1678
## F-statistic: 245.8 on 8 and 9709 DF,  p-value: < 2.2e-16
```

A tabela da análise de variância gerada com **`summary.aov`** mostra que os efeitos principais e a interação entre Cor e EnsMed são estatisticamente significativos. A tabela da regressão linear gerada com **`summary.lm`** mostra como cada categoria afeta o valor médio de NotaCN. Vamos analisar somente as categorias consideradas estatisticamente significativas. Lembre-se que os coeficientes dos níveis de uma variável categórica representam variações com respeito a uma categoria

de base — geralmente decidida pelo interpretador por ordem alfabética e omitida no *output*. No caso, a categoria de base é um cidadão de **`CorBranca`** que cursou **`EnsMedEscPub`**.

Os coeficientes de regressão indicam que o efeito principal da Cor é –19,46 pontos para pretos e –17,53 para pardos (com $p < 0,05$) comparados aos brancos. Pessoas que estudaram parcial ou completamente na rede privada têm um ganho médio de 30,54 e 74,36 pontos respectivamente (com $p < 0,05$) comparados aos estudantes da escola pública, mantido todo o resto constante. Somados aos efeitos principais, observamos que estudantes pretos que estudaram na escola privada atingem, em média, –21.48 pontos, *coeteris paribus*. Somando todos os esses efeitos (principais e interações), é possível perceber que o privilégio de estudar numa escola privada não alcança negros e brancos da mesma maneira.

Sintaxe do modelo

Todas as versões do modelo linear geral empregam uma sintaxe típica bastante intuitiva que especifica como o interpretador deve construir a equação do modelo. Alguns símbolos já foram empregados aqui.

- O *til* ~ separa a variável dependente de seus preditores enquanto o *sinal positivo* + acrescenta termos ao modelo. Portanto, ao escrever $y \sim x1 + x2$ estamos demandando que $x1$ e $x2$ sejam preditores de y.
- O termo de *interceptação* será incluído pelo interpretador automaticamente a menos que o analista o proíba. Além disso, para remover qualquer termo, é empregado sinal negativo. Portanto, ao escrever $y \sim x1 + x2 - 1$, estamos zerando o termo de intercepção, forçando o modelo a passar pela origem.

- Os efeitos principais são representados pelas próprias variáveis, mas as *interações* são separadas por dois pontos. Portanto, ao escrever *y* ~ *x*1 + *x*2 + *x*1:*x*2 estamos solicitando um modelo com interceptação, efeitos principais e interações.
- Efeitos principais e interações podem ser solicitados sinteticamente com um asterisco. Portanto, escrever *y* ~ *x*1**x*2 deve ser equivalente a escrever *y* ~ *x*1 + *x*2 + *x*1:*x*2.

A representação simplificada de termos principais e interações permite gerar outras chamadas. Por exemplo, ***x*1**x*2**x*3** corresponde aos efeitos principais e todas as interações das três variáveis ***x*1 + *x*2 + *x*3 + x1:x2 + x1:x3 + x2:x3 + x1:x2:x3**. Se não desejarmos o último termo de interação, podemos removê-lo na primeira formulação, fazendo ***x*1**x*2**x*3 - *x*1:*x*2:*x*3**. Se escrevermos a soma das variáveis ao quadrado **(*x*1 + *x*2 + *x*3)^2**, isso será interpretado como solicitação de todos os efeitos principais e todas as interações duplas (mas não a interação tripla).

Revisando a lição

Nessa lição, você aprendeu que:

1. os coeficientes de regressão informam os efeitos de um preditor sobre a variável dependente quando todos os demais preditores foram "controlados";
2. esse controle é sublinhado pelo analista na cláusula *coeteris paribus* (i.e., "mantido todo o resto constante"), tipicamente acrescentada aos resultados finais;
3. quando os preditores estão associados, o "controle" das variáveis consiste em subtrair dos preditores seus efeitos comuns,

de maneira que os coeficientes de regressão dependam somente dos efeitos exclusivos de cada preditor;

4. termos de interação são efeitos interseccionais imputáveis aos cruzamentos de duas variáveis.

Atividades propostas

- Para verificar se você compreendeu a diferença entre associação e interação de variáveis, crie um modelo fictício de variáveis explicativas não associadas, mas interagentes.
- Na planilha do Enem ou em outra planilha do seu interesse, escolha duas variáveis preditoras e avalie a) se elas estão associadas e b) se elas apresentam termo de interação estatisticamente significativo sobre uma variável escalar.

LIÇÃO 8.

MODELOS CONCORRENTES

Enquanto os primeiros passos no aprendizado dos métodos quantitativos consistem da realização de testes simples (lições 4 a 6), a situação mais corriqueira da pesquisa em educação em ciências de circulação internacional é a elaboração de modelos complexos com muitas variáveis. O modelo linear clássico (também chamado modelo linear geral) é a primeira ferramenta poderosa para analisar o poder preditivo de um conjunto de variáveis sobre outro. Uma vez que o analista já tenha seus dados coletados e já tenha decidido trabalhar com um modelo linear, seu trabalho consistirá em retirar e acrescentar termos do modelo até chegar em uma formulação com poucos parâmetros e grande poder explicativo. Obviamente, o cuidado com a significância estatística do modelo é somente parte da história. O modelo final deve ser também cientificamente relevante e é esperado que essa preocupação com a relevância conceitual da análise também oriente as escolhas do analista. Conclusões em investigações empíricas nunca vêm exclusivamente dos dados, mas dependem também da maneira como o conhecimento compartilhado pela comunidade orienta as expectativas e escolhas do analista. Essa orientação dá intersubjetividade à análise, tornando-a relevante e inteligível para os demais membros da comunidade.

Do ponto de vista da estatística, a modelagem pode ser vista como um problema de otimização. Elegemos alguns critérios que nos permitem comparar dois ou mais modelos, decidindo qual deles se ajusta melhor aos dados, qual deles é mais plausível. Em seguida, geramos outros modelos (adicionando ou removendo termos) e seguimos testando até que o processo convirja para um modelo que parece ser o melhor entre todos aqueles que puderam ser considerados. Esse é o modelo que levaremos a público para ser discutido e analisado.

Essa seleção entre modelos diferentes pode ser feita manualmente pelo analista, decidindo, caso a caso, quais variáveis serão retidas e quais serão eliminadas. De qualquer maneira, a escolha entre um modelo e outro deve seguir critérios claros, especificados por alguma das métricas de avaliação disponíveis na literatura. Há, ainda, algoritmos que fazem essas escolhas automaticamente baseados em critérios completamente estatísticos. Os leitores mais ingênuos podem achar que a adoção de critérios puramente estatísticos representa uma vantagem, tornando o resultado mais "objetivo", independente do analista. No entanto, ao longo da nossa discussão, você perceberá o quanto essa crença tende a ser um problema.

De partida, a ilusão objetivista precisa ser considerada ingênua porque diversas escolhas são feitas pelo analista o tempo todo. São escolhas: o tamanho da amostra, o modelo linear, as variáveis disponibilizadas no *dataframe*. As métricas de comparação dos modelos também precisam ser escolhidas em meio a diversas opções. Nunca exclusivamente técnicas, as escolhas do analista são escolhas e, se forem feitas de maneira diferente, podem produzir resultados diferentes. A análise quantitativa mais objetiva não é um monólogo dos dados, mas um diálogo no qual a voz do analista importa.

Outro problema do uso de critérios completamente estatísticos é que eles podem gerar modelos pouco relevantes. Algumas variáveis, em vista

do que sabemos sobre elas, precisam ser mantidas no modelo para que ele tenha sentido para a comunidade. Quando o analista executa a modelagem de forma não automática, ele pode arbitrar a relevância conceitual das variáveis em situações que, do ponto de vista da significância estatística, são semelhantes. Porém, quando trabalhamos com modelos muito complexos, com muitas variáveis e muitos termos de interação, ou com muitas observações, essa participação do analista pode tornar a investigação insuportavelmente lenta.

Métricas de avaliação do ajuste

Diversas métricas podem ser utilizadas para avaliar a qualidade do ajuste de um modelo aos dados disponíveis (Crawley, 2005; Fahrmeir *et al*., 2013):

- *R-quadrado*, ou coeficiente de determinação, traduz a proporção da variabilidade da variável dependente explicada pelo modelo. Corresponde à razão entre a soma de quadrados explicada (ESS) e a soma de quadrados total (TSS). *Quanto maior o R-quadrado, melhor o ajuste*. Ele é uma boa medida de qualidade do ajuste *em si*, pois dispensa fazer comparações entre modelos alternativos para ser interpretado. Porém, não é o melhor indicador para comparar modelos concorrentes, pois sempre favorece modelos com mais parâmetros.

$$R^2 = \frac{ESS}{TSS}$$

- *R-quadrado ajustado* é igual ao R-quadrado com uma penalidade para a presença de muitos parâmetros no modelo. Ele também não depende de fazer comparações e funciona como uma boa medida

da qualidade do ajuste *em si*. Porém, a correção introduzida no R-quadrado ajustado pode ser considerada muito pequena e seu uso tende a favorecer modelos com mais variáveis mesmo quando elas não são muito relevantes (Fahrmeir *et al.*, 2013).

$$R^2_{adj} = 1 - \frac{\left(1 - R^2\right)\left(N - 1\right)}{N - p - 1}$$

- *Raiz quadrática média do erro* (RMSE – *Root Mean Square Error*) é uma estimativa do quanto as predições do modelo *y-chapéu* desviam-se dos valores efetivamente observados *y*. Essa estimativa é muito semelhante à variância dos resíduos, com a diferença de que não a dividimos pela respectiva quantidade de graus de liberdade, mas pelo número de observações. Corresponde, portanto, à raiz quadrada da média dos desvios ao quadrado. *Quanto menor a RMSE, melhor o ajuste*. Como ela depende da escala das variáveis, a RMSE é mais adequadamente interpretada na comparação entre modelos semelhantes. Inserir parâmetros ajustáveis no modelo sempre reduz a RMSE. Por isso, ela não é considerada a métrica mais adequada para tomar decisões na modelagem estatística (ainda que seja muito utilizada).

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (y - \hat{y})^2}{n}}$$

- *Média do erro absoluto* (MAE – *Mean Absolute Error*) é uma medida alternativa à RMSE. Ela também avalia o quanto as predições do modelo $\hat{y}$ desviam-se dos valores efetivamente observados y, mas faz isso sem elevar os desvios ao quadrado, sendo, portanto,

menos sensível à presença de *outliers* (pontos muito afastados das predições). *Quanto menor a MAE, melhor o ajuste*. Assim como a RMSE, a MAE depende da escala das variáveis do modelo.

$$MAE = \frac{\sum_{i=1}^{n} | y - \hat{y} |}{n}$$

- *Critério de informação de Akaike* (AIC – *Akaike's Information Criteria*) é definido pelo dobro da diferença entre a quantidade k de parâmetros ajustáveis do modelo e o logaritmo da verossimilhança (ing., *Likelihood*) estimada. A verossimilhança está relacionada à qualidade do ajuste — um modelo mais verossímil está necessariamente mais bem ajustado aos dados. Ao fazer a diferença entre o número de parâmetros e a verossimilhança, essa métrica introduz uma penalidade sobre a inclusão de novas variáveis no modelo, permitindo estimar a probabilidade de dois modelos serem equivalentes no que diz respeito à informação que extraem dos dados. Isso faz com que o AIC oriente o analista de maneira bastante precisa. *Quanto menor o AIC, melhor o ajuste*.

$$AIC = -2log\left(\hat{L}\right) + 2k$$

- *Critério de informação Bayesiano* (BIC – *Bayesian Information Criteria*) é estruturalmente semelhante ao critério de Akaike, mas com uma penalidade muito maior para a inserção de novos parâmetros. É definido como a diferença entre o dobro do logaritmo da verossimilhança e o produto da quantidade de parâmetros ajustáveis *k* e o logaritmo do número de observações. *Quanto menor o BIC, melhor o ajuste*.

$$BIC = -2log\left(\hat{L}\right) + k \cdot log\left(n\right)$$

As decisões do analista podem ser tomadas com a observação de várias dessas métricas ao mesmo tempo ou somente uma delas (AIC ou BIC, por exemplo). A exceção do R-quadrado, todos os critérios colocados aqui devem ser *minimizados* para um ajuste de alta qualidade. O quadro 8.1, a seguir, sintetiza as informações fornecidas nesta seção:

Quadro 8.1: Síntese dos critérios para comparação de modelos

Critério	Características	Deve ser
R-quadrado	Dispensa referência a outros modelos para ser interpretado. Sempre aumenta quando um novo parâmetro é inserido. Não é uma boa medida para comparar modelos.	Maximizado
R-quadrado ajustado	Também dispensa referência a outros modelos para ser interpretado. Inclui uma pequena penalidade para a inserção de novos parâmetros. Tende a favorecer modelos com mais variáveis, mesmo quando elas não são muito relevantes.	Maximizado
RMSE	É definido como a raiz quadrada da média dos resíduos ao quadrado. Permite comparar modelos, mas sempre diminui quando um parâmetro novo é inserido. Ainda que seja amplamente empregado, não é o melhor critério para comparar modelos.	Minimizado
MAE	É definido como a média do módulo dos resíduos. Comparado ao critério RMSE, MAE é menos influenciado por *outliers*, mas sempre diminui quando um novo parâmetro é inserido. Assim como a RMSE, a MAE não é um bom critério para comparar modelos.	Minimizado
AIC	O critério da informação de Akaike é obtido pela diferença entre o dobro do número de parâmetros pelo dobro do logaritmo da verossimilhança estimada. Ele nem sempre diminui quando aumentamos a quantidade de parâmetros do modelo.	Minimizado
BIC	O critério de informação Bayesiano é estruturalmente idêntico ao AIC, mas possui uma penalidade muito maior para a inserção de novos parâmetros (proporcional ao logaritmo da quantidade de elementos da amostra).	Minimizado

Calculando todas as métricas

Quando estamos lidando com poucas variáveis e/ou não estamos interessados em avaliar suas interações, é possível demandar o cômputo das métricas de avaliação do ajuste para todos os modelos possíveis. Assim, a primeira ferramenta que apresentamos aqui é a função **`ols_step_all_possible`** do pacote **`olsrr`** (instale-o, se ainda não o fez). Ela recebe um modelo completo e calcula diversas métricas de avaliação do ajuste para *todos os modelos possíveis* de se derivar por combinações, termo a termo, do modelo completo informado. Como a quantidade de combinações de *n* termos é igual 2^n, essa função pode não ser a mais indicada quando queremos testar muitas variáveis.

Considere como modelo completo: NotaCN ~ Idade + Sexo + Cor + EscPai + EscMae + Renda + EnsMed.

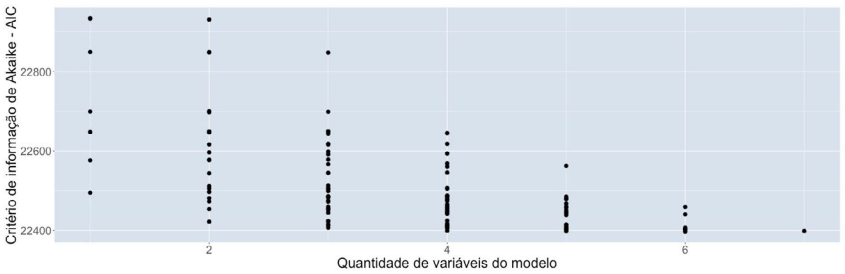
```
library(Hmisc)
load("ENEM_exemplo.dat")
ENEM = ENEM[c(1:2000), ]

library(olsrr)
fit <- lm(NotaCN ~ Idade + Sexo + Cor + EscPai + EscMae + Renda + EnsMed,
data = ENEM)
ols.output <- ols_step_all_possible(fit)
```

Com isso, todas as combinações possíveis das variáveis do modelo foram testadas e diversas métricas da qualidade do ajuste foram calculadas. Digite **`help(ols_step_all_possible)`** para saber como cada uma dessas variáveis corresponde às definições que fornecemos na seção anterior. Tomando o AIC como critério, podemos gerar um gráfico para

visualizar os valores assumidos por AIC em função da quantidade de variáveis retidas no modelo:

```
plot(ols.output$n, ols.output$aic)
```

Gráfico 8.1: AIC para todos os modelos possíveis



Como é possível perceber, um dos modelos apresenta o menor (portanto, melhor) valor de AIC. A seguir, mostramos uma tabela com os dez modelos gerados mais bem ajustados aos dados:

```
##     n                                    modelo      AIC
## 120 6      Idade Cor EscPai EscMae Renda EnsMed 22396.41
## 100 5             Idade Cor EscPai Renda EnsMed 22397.93
## 127 7 Idade Sexo Cor EscPai EscMae Renda EnsMed 22398.33
## 64  4                   Cor EscPai Renda EnsMed 22399.26
## 122 6        Idade Sexo Cor EscPai Renda EnsMed 22399.90
## 99  5            Cor EscPai EscMae Renda EnsMed 22400.80
## 102 5              Sexo Cor EscPai Renda EnsMed 22401.25
## 103 5             Idade Cor EscMae Renda EnsMed 22401.99
## 121 6       Sexo Cor EscPai EscMae Renda EnsMed 22402.78
## 124 6        Idade Sexo Cor EscMae Renda EnsMed 22403.98
```

Poderíamos escolher o primeiro da lista (Idade, Cor, EscPai, EscMae, Renda e EnsMed) e encerrar a análise por aqui. Porém,

é possível que a diferença entre o melhor modelo e os demais não seja estatisticamente significativa. Pode ser também que o segundo ou o terceiro modelo da lista sejam mais pertinentes por razões teóricas. O que faremos nessa situação?

Verossimilhança relativa

Para interpretar a vantagem explicativa de um modelo sobre o outro, podemos considerar a *verossimilhança relativa*, definida como o exponencial da metade da diferença dos valores de AIC:

$$exp\left(\frac{AIC_{menor} - AIC_{maior}}{2}\right)$$

Por construção, a verossimilhança relativa estará sempre entre 0 e 1. Ela pode ser interpretada como um *valor-p* traduzindo a probabilidade de que os dois modelos estejam representando a mesma quantidade de informação presente nos dados. Para que um modelo seja decisivamente descartado diante de outro, é preciso que a verossimilhança relativa entre eles seja muito baixa.

No limite em que o valor-p é igual 0,05, temos que:

$$\begin{aligned} AIC_{menor} - AIC_{maior} &= 2 \cdot ln(0,05) \\ &\approx -5,99 \end{aligned}$$

Portanto, o nível de significância usual estabelece uma distância mínima entre dois valores de AIC para que possamos distinguir decisivamente entre dois modelos. *Diferenças inferiores a 6 nesse critério podem não ser estatisticamente significativas*. Voltando à tabela em que estão

publicados os valores de AIC por modelo, percebemos há diversos modelos com diferença inferior a 6 unidades e que, portanto, o critério AIC não justifica eleger, em definitivo, um modelo em detrimento de todos os outros.

Quando consideramos as incertezas na avaliação da qualidade do ajuste, é usual haver mais de um modelo verossímil. Por essa e outras razões, a modelagem estatística multivariada é considerada um terreno movediço. Quanto mais rigorosos nós somos, mais evidente fica o arbítrio do analista.

Seleção *stepwise* de variáveis

A forma mais comum de gerar modelos com muitas variáveis é recorrer a algoritmos do tipo *stepwise* (i.e., passo-a-passo). Tais algoritmos partem de um modelo vazio (ou de um modelo cheio) e selecionam variáveis para serem inseridas (ou removidas). Em seguida, testam a medida em que essa inclusão (ou exclusão) altera a métrica de avaliação da qualidade do ajuste. Após várias operações, percebendo que não estão ocorrendo alterações significativas na qualidade de ajuste, o algoritmo publica o melhor modelo obtido.

Há três tipos de processo que podem ser implementados por meio desses algoritmos (Crawley, 2005).

1. *Seleção progressiva* (ing. *Forward*): consiste em tomar o modelo vazio como ponto de partida e adicionar progressivamente as variáveis mais relevantes até que não haja ganho expressivo no poder explicativo do modelo.
2. *Seleção regressiva* (ing. *Backward*) ou *eliminação*: consiste em tomar como ponto de partida o modelo cheio (com todas as variáveis e interações que o analista considera) e retirar os termos um a um até que a retirada comprometa o poder explicativo do modelo.

3. *Seleção passo a passo* (ing. *Stepwise*): consiste em combinar os dois métodos. Você começa com o modelo vazio (ou cheio) e adiciona (ou retira) variáveis progressivamente. Porém, como cada variável adicionada ou retirada altera o poder explicativo das demais, talvez seja interessante operar passo a passo.

Com funções da biblioteca **`olsrr`**, conseguimos calcular e comparar as métricas de todos os modelos possíveis! Isso nos permitiu visualizar como os modelos se posicionam uns com relação aos outros. Essa visão panorâmica não é proporcionada por algoritmos do tipo *stepwise*, que se desenrolam como quem caminha no escuro, tateando o chão e as paredes, o que não impede, contudo, que algoritmos *stepwise* recomendem os mesmos modelos que seriam recomendados por uma análise de todos os modelos possíveis.

Apresentamos aqui a função **`stepAIC`** da biblioteca **`MASS`** (instale-a, se ainda não o fez). Ela recebe como *input* um **`modelo completo`**, no qual o analista deve inserir todos os termos (efeitos principais e interações) que deseja considerar para a análise. O analista também deve informar uma direção, que pode ser "*forward*" para seleção progressiva; "*backward*", para seleção regressiva; ou "*both*" para seleção mista. A diretiva **`trace = TRUE`** solicita que o progresso da modelagem seja publicado passo a passo. Se falso, as etapas intermediárias serão omitidas e somente o modelo final será apresentado.

```
full.fit <- lm(NotaCN ~ Idade + Sexo + Cor + EscPai + EscMae + Renda +
EnsMed, data = ENEM)
fit <- stepAIC(full.fit, direction = "both", trace = TRUE)

## Start:  AIC=16720.57
## NotaCN ~ Idade + Sexo + Cor + EscPai + EscMae + Renda + EnsMed
##
```

```
##          Df Sum of Sq     RSS   AIC
## - Sexo    1       350 8237704 16719
## <none>                8237354 16721
## - EscMae  6     56088 8293443 16722
## - Idade   1     26597 8263951 16725
## - EscPai  6     73030 8310385 16726
## - Cor     4     69368 8306722 16729
## - EnsMed  2    192318 8429672 16763
## - Renda  16    392363 8629717 16782
##
## Step:  AIC=16718.66
## NotaCN ~ Idade + Cor + EscPai + EscMae + Renda + EnsMed
##
##          Df Sum of Sq     RSS   AIC
## <none>                8237704 16719
## - EscMae  6     55880 8293585 16720
## + Sexo    1       350 8237354 16721
## - Idade   1     26339 8264044 16723
## - EscPai  6     72713 8310418 16724
## - Cor     4     69445 8307149 16727
## - EnsMed  2    192588 8430292 16761
## - Renda  16    392334 8630038 16780
```

Como resultado, o algoritmo convergiu para o modelo com menor AIC segundo a análise de todos os modelos possíveis.

Testando as hipóteses do modelo

Após testar vários modelos exaustivamente uns contra os outros, ainda é necessário testarmos as premissas do modelo linear clássico (também chamado modelo linear geral). No caso em questão, como ainda estamos usando análises de variância, covariância e regressão linear, todas estão baseadas nos mesmos pressupostos:

1. independência dos resíduos;
2. normalidade dos resíduos;
3. igualdade das variâncias;
4. ausência de *outliers* alavancados.

O usuário pode solicitar visualização dos gráficos diagnósticos e perceber que as premissas do modelo foram aproximadamente respeitadas.

```
plot(fit)
```

Mais informações sobre como interpretar o *output* dessa função podem ser obtidas ao final da lição 5.

Revisando a lição

Nessa lição, você aprendeu a:

1. definir e computar as principais métricas de avaliação da qualidade do ajuste — R-quadrado, R-quadrado ajustado, RMSE, AIC, BIC — com a função **`ols_step_all_possible`**.
2. avaliar a *verossimilhança relativa* entre dois modelos segundo o AIC.
3. realizar regressões do tipo **`stepwise`** com a função **`stepAIC`**.

Atividade proposta

- Tomando como ponto de partida uma planilha com dados educacionais do seu interesse, identifique, entre os modelos mais bem ajustados aos dados, aquele que pareça ser, segundo critérios teóricos claramente especificados, o mais plausível.

LIÇÃO 9.

REGRESSÃO LOGÍSTICA

Modelos lineares clássicos são adequados quando a variável dependente é contínua e aproximadamente normal (Fahrmeir *et al.*, 2013). Há, porém, situações em que desejamos modelar uma *variável de resposta binária*. Um problema educacional bastante atual que se encaixa exatamente nessa situação é o estudo da evasão e da permanência dos estudantes na educação superior (Gilioli, 2016; Lima Junior *et al.*, 2019). De fato, a admissão, a evasão e a diplomação são exemplos de *eventos* que podem ocorrer ou não. Por isso, é conveniente codificá-los como uma variável binária (0 ou 1). Pela sua natureza não escalar, variáveis binárias jamais são normalmente distribuídas e, portanto, seu comportamento não pode ser predito por um modelo linear clássico.

O que geralmente fazemos, nesse caso, é empregar uma das diversas generalizações do modelo linear clássico: a chamada *regressão logística*. Ela permite avaliar o poder preditivo de um conjunto diverso de variáveis (escalares ou não) sobre uma variável binária tipo-evento. Por exemplo, com base em dados da Universidade Federal do Espírito Santo, de 2007 a 2012, Pereira *et al.* (2015) identificaram 21 preditores estatisticamente significativos da ocorrência de diplomação, tais como o coeficiente de rendimento, a quantidade de reprovações, a relação candidato/vaga, a área

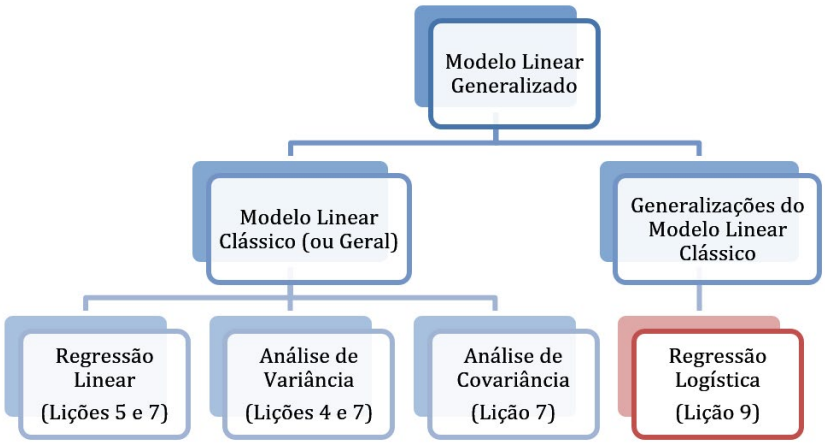
do conhecimento do curso, entre outros. Com base nessas informações e no modelo linear, é possível estimar a chance de diplomação de cada estudante. Em outra investigação, Gottlieb (2018) elaborou uma regressão logística para compreender os preditores da escolha por carreiras científicas entre egressos da educação secundária. Segundo seus resultados, a afirmação de que a ciência é predominantemente escolhida por meninos depende muito da maneira como o analista define as fronteiras entre carreiras científicas e não científicas. No entanto, em todas as demarcações testadas, os construtos da teoria do valor-expectativa (ing. *expectancy-value theory*) foram bons preditores das escolhas profissionais dos meninos brancos, mas menos relevantes para predizer as escolhas profissionais das meninas pretas e pardas.

Enfim, se você deseja avaliar o poder preditivo de um conjunto de variáveis (escalares ou categóricas) sobre uma variável dependente binária, a regressão logística é o método mais recomendado para sua pesquisa. De fato, a quantidade de pesquisas educacionais que emprega esse método de análise é impressionante (Alves; Ortigão; Franco, 2007; Bottia *et al.*, 2018; Dewitt; Archer; Mau, 2016). Neste livro, a apresentação da regressão logística pretende contribuir para que o leitor conheça como o modelo linear geral (cf. lição 7) pode funcionar além dos seus tipos clássicos (Anova, Ancova e regressão). A saber, outro método generalizado que despertou o interesse da pesquisa educacional é a chamada *análise de sobrevivência*. Se a regressão logística tem por objeto a ocorrência de um evento, a análise de sobrevivência permite modelar o *tempo até a ocorrência desse evento*. Para saber mais, remeto o leitor à literatura específica (Colosimo; Giolo, 2006; Lima Junior; Da Silveira; Ostermann, 2011).

Conforme discutimos anteriormente (cf. lição 7), no âmbito do *modelo linear geral*, a relação imediata entre as variáveis dependentes e independentes pode não ser linear. Contudo, tudo o que aprendemos com

a regressão linear clássica pode ser transportado para outras situações se fizermos a escolha conveniente de uma *função de ligação*, i.e., uma função que relacione os valores médios da variável dependente aos seus preditores lineares. A escolha de uma função de ligação não linear geralmente implica impor uma distribuição não normal aos resíduos. Tudo isso pode ser feito de maneira muito intuitiva e usando uma linguagem semelhante à que aprendemos com o modelo linear clássico.

Diagrama 9.1: Posição da regressão logística com relação às demais versões do modelo linear generalizado



Nesta lição, vamos aprender a realizar uma *regressão logística*. Ela é a solução mais usual quando a variável de resposta é binária e as variáveis independentes são escalares ou categóricas (Fahrmeir *et al.*, 2013). Assim como outros modelos lineares, a regressão logística é particularmente útil para abordar situações em que podemos presumir uma relação de predição. Por exemplo, podemos estar interessados em avaliar o efeito do desempenho acadêmico, da modalidade de ingresso, da integração social na probabilidade de evasão dos estudantes. Também podemos estar interessados

em avaliar o efeito da origem social na probabilidade de tentar uma vaga na educação superior ou escolher uma carreira científica e tecnológica.

A função `logit`

Designemos pela letra grega η o *preditor linear* construído a partir de *k* variáveis categóricas ou escalares. A saber, tal preditor linear deve ter a seguinte forma geral:

$$\eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_j x_j + \cdots + \beta_k x_k$$

Designaremos pela letra grega *pi* a *probabilidade condicional* de que um evento ocorra. Em outras palavras, *pi* especifica a probabilidade de ocorrência do evento (y = 1) quando as informações das variáveis *x* foram especificadas. Em um estudo sobre evasão discente, *pi* designaria a probabilidade de um estudante evadir dado, por exemplo, que é um homem ou uma mulher, que obteve aprovação em todas, algumas ou nenhuma disciplina, que ingressou por uma ou outra modalidade de acesso. A probabilidade condicional *pi* pode ser representada dessa maneira:

$$\pi = P(y = 1 \mid x_1, x_2, \cdots, x_k)$$

Por representar uma probabilidade, *pi* está necessariamente limitado ao intervalo de 0 a 1. Por outro lado, o preditor linear *eta* pode assumir qualquer valor na reta real. Portanto, será particularmente inconveniente tentarmos empregar aqui o modelo linear clássico:

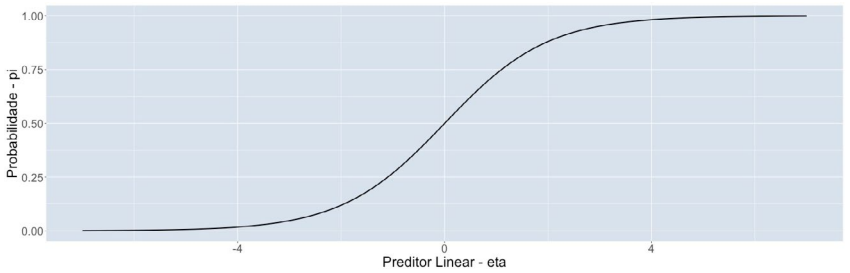
$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

A regressão logística é a solução típica para mapear um preditor linear qualquer em uma quantidade que possa ser interpretada como uma probabilidade (i.e., que esteja confinada ao intervalo de 0 a 1). A *função resposta logit* pode ser definida dessa maneira:

$$\pi = \frac{exp(\eta)}{1 + exp(\eta)}$$

Ao gerar o gráfico dessa função, percebemos que ela cumpre o que promete (uma função monotonicamente crescente que leva toda a reta real no intervalo aberto de 0 a 1).

Gráfico 9.1: Função de resposta `logit`



Podemos inverter a função resposta para obter a chamada *função de ligação logit*:

$$\eta = log\left(\frac{\pi}{1-\pi}\right)$$

Essas duas funções serão muito importantes para orientar a interpretação correta dos parâmetros do modelo na regressão que estamos prestes a realizar.

Interpretando os coeficientes da regressão logística

Definimos a razão entre *pi* e *1 – pi* como a *chance* de ocorrência do evento. Portanto, decorre da função de ligação *logit* que a chance é igual ao exponencial do preditor linear:

$$\frac{\pi}{1-\pi} = exp\left(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k\right)$$

Como o exponencial de uma soma pode ser fatorado em um produto de exponenciais, a relação anterior permite interpretar os coeficientes da regressão de uma maneira intuitiva. Para compreendermos juntos como essa interpretação funciona, considere o *j*-ésimo preditor do modelo e seu respectivo coeficiente $beta_j$. Se essa variável aumentar em 1 unidade, de que maneira esse aumento altera a *chance* de ocorrência do evento? Podemos responder a essa pergunta com o cálculo da razão das chances:

$$\frac{chance\left(x_j + 1\right)}{chance\left(x_j\right)} = \frac{exp\left[\beta_0 + \beta_1 x_1 + \cdots + \beta_j\left(x_j + 1\right) + \cdots + \beta_k x_k\right]}{exp\left[\beta_0 + \beta_1 x_1 + \cdots + \beta_j\left(x_j\right) + \cdots + \beta_k x_k\right]}$$

Fazendo os devidos cancelamentos, percebemos que:

$$chance\left(x_j + 1\right) = exp\left(\beta_j\right) \cdot chance\left(x_j\right)$$

Assim, o *exponencial do coeficiente da regressão indica como ele afeta a chance de ocorrência do evento*. Se o coeficiente beta é tal que seu exponencial é igual a dois, a chance de ocorrência do evento dobra quando a variável associada ao beta ganha uma unidade. Variáveis categóricas

são internamente codificadas com 0 ou 1. Então o efeito da inclusão em uma categoria pode ser facilmente avaliado pelo exponencial de seu respectivo beta. A codificação das variáveis escalares, por outro lado, fica a critério do programador. Logo, é importante codificá-las de maneira que variações unitárias sejam significativas.

Um exemplo prático

A menos da função de ligação e da maneira como ela instrui interpretar os coeficientes do modelo, a regressão logística é conduzida com preocupações muito semelhantes aos modelos lineares clássicos. Precisamos incluir, eliminar e testar a significância estatística das variáveis, bem como avaliar a qualidade do ajuste.

Para experimentarmos, na prática, como funciona a interpretação dos coeficientes, considere os dados da pesquisa *Teaching and Learning International Survey* (TALIS). Trata-se de uma pesquisa internacional liderada pela OCDE que levanta informações sobre as experiências e percepções de professores e diretores de escola no Brasil e no mundo. Ela trata de temas como formação inicial e continuada, carga de trabalho e qualidade de vida, atuação em contextos multiculturais. Seu principal foco está nos anos finais do ensino fundamental (*lower secondary*, ISCED 2). Porém, no Brasil e em outros poucos países, a pesquisa foi estendida para o ensino médio (*upper secondary*, ISCED 3). Os dados referentes aos professores brasileiros de ensino médio foram importados para o R e podem ser recuperados carregando o arquivo “TALIS.dat” (disponível em www.ppgeduc.unb.br/mqpe):

```
library(Hmisc)
load("TALIS.dat")
```

O *dataframe* é enorme e a quantidade de informação relevante é impressionante! Você pode inspecionar as variáveis disponíveis com a função **`describe`**:

```
describe(dados)
```

Para experimentarmos a regressão logística com um problema potencialmente interessante, lançamos a seguinte questão de pesquisa.

- Quais tópicos abordados em ações de desenvolvimento profissional docente têm maior chance de serem percebidos como relevantes pelos professores brasileiros?

Na questão TT3G23, o instrumento pergunta aos professores: "Algum dos seguintes tópicos listados abaixo foi incluído nas suas atividades de desenvolvimento profissional durante os últimos 12 meses?" Os tópicos listados são:

(A) conhecimento e compreensão de minha área de ensino;
(B) competências pedagógicas para lecionar em minha área de ensino;
(C) conhecimento de currículo;
(D) práticas de avaliação de alunos;
(E) habilidades em TIC para o ensino;
(F) comportamento do aluno e gestão da sala de aula;
(G) gestão e administração escoar;
(H) abordagens para o aprendizado individual;
(I) ensino para alunos com necessidades especiais;
(J) ensino em um ambiente multicultural ou multilíngue;
(K) ensino de habilidades intercurriculares (criatividade, pensamento crítico e resolução de problemas);

(L) análise e uso de avaliação dos alunos;

(M) cooperação entre pais/responsáveis e professores;

(N) comunicação com pessoas de diferentes culturas ou países;

(O) outro.

Na questão TT3G25, o mesmo instrumento pergunta aos professores: "Considerando todas as suas atividades de desenvolvimento profissional durante os últimos 12 meses, alguma delas teve um impacto positivo sobre a sua prática de ensino?" A resposta dos professores foi codificada em "sim" ou "não".

Para responder à questão de pesquisa proposta, farei um modelo de regressão logística em que TT3G25 seja a variável dependente e os itens da questão TT3G23 sejam os preditores. Porém, antes de prosseguir, proponho recodificar as variáveis de maneira que o nível de referência seja a resposta "não". Isso tornará a interpretação dos resultados mais intuitiva.

Conforme discutimos anteriormente, na lição 7, preditores categóricos têm um *nível de referência* com relação ao qual todos os demais níveis são contrastados. Tais níveis podem ser listados levando a variável à função **`levels`**. O primeiro da lista funcionará como nível de referência ao ser inserido no modelo linear.

```
levels(dados[,110])

# [1] "Sim" "Não"
```

Se estivéssemos trabalhando com uma variável de cor codificada como "branco", "preto", "pardo", "indígena", nessa ordem, os três últimos níveis ("preto", "pardo", "indígena") seriam contrastados com o primeiro ("branco"), favorecendo comparar grupos minoritários com o grupo privilegiado. Qualquer outra ordem dos níveis tornaria a análise potencialmente

confusa. Da mesma maneira, se estivéssemos trabalhando com uma variável de renda, todos os níveis seriam comparados ao primeiro. Levando isso em consideração, o analista deve decidir se o primeiro nível será de renda mais baixa ou renda mais alta. Portanto, assim como ocorria nos modelos lineares clássicos (cf. lição 7), *o ordenamento dos níveis é muito importante para a maneira como os coeficientes da regressão poderão ser interpretados*.

As variáveis TT3G23x, codificadas em "sim" e "não", informam se cada tópico foi incluído nas atividades formativas do professor nos últimos doze meses. Manter o "sim" como nível de referência corresponderá a avaliar o efeito de exclusão do tópico (em vez de sua inclusão) nas atividades formativas. Igualmente, ao manter o "sim" como nível de referência da questão TT3G25, nós daremos destaque às situações em que os professores *não* reconheceram a efetividade das experiências formativas.

No fundo, a informação é exatamente a mesma. Porém, ter que lidar com várias negativas, tornará nosso resultado mais difícil de compreender. Compare, por exemplo, as duas afirmações a seguir.

- A *não* inclusão do tópico X contribui para que os professores *não* reconheçam a efetividade do desenvolvimento profissional.
- A inclusão do tópico X contribui para que os professores reconheçam a efetividade do desenvolvimento profissional.

Como é possível perceber, embora as duas afirmações carreguem a mesma informação, a segunda é mais fácil de compreender que a primeira. Portanto, para facilitar a interpretação posterior dos coeficientes, vamos redefinir a resposta "não" como nível de referência de todas as variáveis do modelo. Isso pode ser feito empregando a função **`relevel`** e o controlador de fluxo **`for`**:

```
for (j in 110:143) {
    dados[, j] = relevel(dados[, j], “Não”)
}
```

A função **relevel** redefine o nível de referência de cada variável. Como precisamos aplicá-la às colunas da planilha, uma a uma, empregamos o controlador de fluxo **for**. Ele define uma variável, que assumirá uma sequência de valores inteiros especificados pelo analista. No código anterior, o controlador de fluxo faz *j* variar de 110 a 143, executando o código entre chaves para cada um dos valores que j pode assumir. Assim, ele redefine o nível de referência da coluna 110, 111, 112... até a 143.

Feita essa alteração, nós podemos empregar a função **glm** do modelo linear generalizado com a diretiva **family = binomial** para realizar uma regressão logística. Nesse primeiro exemplo, tomamos somente os primeiros três itens da questão TT3G23:

```
fit = glm(TT3G25 ~ TT3G23A + TT3G23B + TT3G23C,
                   data = dados,
                   family = “binomial”)

summary(fit)

##
## Call:
## glm(formula = TT3G25 ~ TT3G23A + TT3G23B + TT3G23C, family = “binomial”,
##     data = dados)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1475   0.4582   0.4582   0.4582   1.1655
##
```

```
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   0.0282     0.1207   0.234 0.815331
## TT3G23ASim    0.5400     0.1606   3.362 0.000775 ***
## TT3G23BSim    1.0117     0.1510   6.702 2.05e-11 ***
## TT3G23CSim    0.6209     0.1379   4.502 6.75e-06 ***
## ---
## Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2186.5  on 2438  degrees of freedom
## Residual deviance: 1978.0  on 2435  degrees of freedom
##   (389 observations deleted due to missingness)
## AIC: 1986
##
## Number of Fisher Scoring iterations: 4
```

Conforme esse modelo, a chance de o professor reconhecer o impacto positivo das ações formativas aumenta quando os tópicos A, B ou C são incluídos. Para avaliar como esses tópicos aumentam a chance de reconhecimento, precisamos calcular os exponenciais dos seus coeficientes fazendo **`exp(coef(fit))`**:

```
exp(coef(fit))

## (Intercept)  TT3G23ASim  TT3G23BSim  TT3G23CSim
##    1.028600    1.716055    2.750401    1.860594
```

Portanto, a chance de reconhecimento do impacto positivo da ação formativa pelo professor é 1,71 vezes maior quando o tópico A é inserido; 2,75 vezes maior quando o tópico B é inserido e 1,86 vezes maior para a inserção do tópico C, *mantido o resto constante*. Como vimos em lições anteriores (cf. lição 7), a cláusula *coeteris paribus* implica que os valores mostrados representam *efeitos exclusivos* (i.e., efeitos de uma variável quando os efeitos de todas as demais foram descontados). De fato, como os cursos de formação tendem a ser montados segundo modelos específicos (Contreras, 2002), é esperado que os tópicos abordados nos cursos estejam associados. Essa associação precisa ser levada em consideração ao interpretar os coeficientes de regressão (cf. lição 7).

Seleção *stepwise* da regressão logística

A seleção *stepwise* de variáveis também pode ser feita na regressão logística, por meio da função **`stepAIC`** disponível na biblioteca **`MASS`**. A biblioteca **`tidyr`** é carregada aqui para fazer funcionar o operador *pipe* %>%, que não está definido na biblioteca de base.

Conforme antecipamos anteriormente, na lição 2, o operador *pipe* funciona como uma tubulação ou um cachimbo, levando a informação da esquerda para a direita. Ele permite escrever funções compostas de maneira mais elegante. Por exemplo, onde escreveríamos **`A(B(x),y)`**, podemos escrever **`B(x) %>% A(y)`**. Dessa maneira, o operador de tubulação torna a leitura do código mais amigável.

```
library(tidyr)
library(MASS)

fit <- glm(TT3G25 ~ .,
             data = na.omit(dados[, c(120:134, 143)]),
             family = binomial) %>%
                 stepAIC(trace = FALSE)

##              Estimate Std..Error Pr...z.. exp.beta sig
## (Intercept)  -0.2547     0.1345   0.0581     0.78
## TT3G23ASim    0.3220     0.1689   0.0565     1.38
## TT3G23BSim    0.8016     0.1600   0.0000     2.23   *
## TT3G23DSim    0.3465     0.1471   0.0185     1.41   *
## TT3G23ESim    0.3835     0.1326   0.0038     1.47   *
## TT3G23GSim    0.3159     0.1795   0.0784     1.37
## TT3G23HSim    0.2216     0.1512   0.1429     1.25
## TT3G23ISim    0.3244     0.1558   0.0374     1.38   *
## TT3G23KSim    0.4967     0.1474   0.0007     1.64   *
## TT3G23OSim    0.4793     0.1585   0.0025     1.61   *
```

O modelo para o qual o algoritmo de seleção convergiu é bem diferente do anterior. Enquanto, no primeiro modelo, os tópicos A, B e C figuravam com grande poder preditivo, a inserção dos demais preditores fez com que o tópico C fosse eliminado e o tópico A perdesse sua significância estatística. De fato, as variáveis eliminadas não são, em si, irrelevantes, mas seu poder explicativo deve estar contemplado nos preditores que foram mantidos. Como sabemos, os coeficientes de regressão registram efeitos exclusivos (i.e., controlados os efeitos de todos os demais preditores). Portanto, os tópicos cujos efeitos exclusivos não podem ser atribuídos ao acaso (com $p<0,05$) são os seguintes:

(B) competências pedagógicas para lecionar em minha área de ensino;

(D) práticas de avaliação de alunos;

(E) habilidades em TIC para o ensino;

(I) ensino para alunos com necessidades especiais;

(K) ensino de habilidades intercurriculares (criatividade, pensamento crítico e resolução de problemas);

(O) outro.

Esses temas em destaque, em sua maioria, sugerem que formações de caráter mais ou menos utilitarista façam sucesso entre os professores brasileiros. A menos do desenvolvimento da criatividade e do pensamento crítico, os professores valorizam o desenvolvimento de competências pedagógicas específicas de seu componente curricular, práticas de avaliação e emprego de tecnologias da informação e comunicação. Portanto, apesar das críticas usuais ao papel dos grandes organismos internacionais no aprofundamento da reforma neoliberal e no caráter tecnicista e conteudista da formação dos professores (cf. Rezende; Ostermann, 2015), a demanda pela racionalidade técnica parece estar instalada no corpo docente brasileiro e não precisa ser imposta por organismos externos para florescer como preferência nacional.

Revisando a lição

Nessa lição, você aprendeu a:

1. realizar uma regressão logística usando a diretiva **`family = binomial`** da função **`glm`**.
2. solicitar uma modelagem logística *stepwise* baseada no critério da informação de Akaike.
3. interpretar os coeficientes do ajuste a partir da função de ligação *logit*.

Atividade proposta

1. Explore a planilha TALIS, identificando outras variáveis binárias interessantes. Que outras conclusões podemos tirar dos dados com as ferramentas que você conhece?

PARTE IV.

ANÁLISE EXPLORATÓRIA MULTIVARIADA

Nas lições anteriores, nós estudamos modelos multivariados que, além de serem sempre lineares em seus parâmetros, fazem uma distinção clara entre variáveis dependentes e independentes. Por essa razão, eles são convenientes para testar relações preditivas, ou relações do tipo causa e efeito. Contudo, em outros cenários, nossas expectativas prévias não nos permitem pensar em termos do "efeito de A sobre B", mas em termos de "associações recíprocas entre A e B". Do ponto de vista da estatística, o pensamento associativo é diferente do preditivo por sua simetria: declarar que A está associado a B é equivalente à declaração de que B está associado a A. Essa simetria requer uma maneira totalmente diferente de abordar os dados.

Algoritmos associativos são muito diferentes daqueles que pretendem testar relações unidirecionais. Nas técnicas de análise bivariada, a correlação de Pearson (lição 5) e a análise de tabelas de contingência (lição 6) são a porta de entrada para testar a associação entre variáveis escalares e categóricas, respectivamente. Já na análise multivariada, há duas grandes famílias de métodos estatísticos que precisam ser identificadas: *i)* análises de componentes principais e *ii)* análises fatoriais. Cada uma dessas famílias tem suas próprias maneiras de testar a associação de variáveis escalares e categóricas.

Ainda que parecidas, essas duas famílias de métodos estatísticos não podem ser confundidas (Mair, 2018). Por outro lado, para distingui-las, é preciso conhecê-las um pouco. De fato, as situações práticas em que podemos aplicar análises fatoriais e de componentes principais são praticamente as mesmas e não há critérios decisivos para preferir uma abordagem à outra. Em geral, análises de componentes principais partem de afirmações menos restritivas, favorecendo que o analista realize um trabalho mais *exploratório* e *descritivo* (em oposição a confirmatório e inferencial). Talvez por essa razão, essas técnicas sejam amplamente empregadas em mineração de dados (*data mining*), aprendizado de máquina (*machine learning*), *Big Data*, e pesquisas sociais do tipo *survey* (i.e., levantamento).

A real diferença entre as análises fatoriais e de componentes principais ficará clara somente após termos nos familiarizado com pelo menos uma delas — por isso, não se preocupe em distingui-las agora. As próximas lições dedicam-se a introduzir a análise de componentes principais e seus métodos relacionados.

- **Lição 10** (análise de componentes principais). Nessa lição, explicamos como identificar padrões de associação expressos em matrizes de covariâncias e correlações — trata-se de uma técnica voltada inicialmente a variáveis escalares, mas extensível a variáveis dicotômicas e ordinais.
- **Lição 11** (análises de correspondência). Muito empregada na interpretação de questionários socioeconômicos, a análise de correspondência (simples, múltipla e conjunta) permite identificar associações estruturais entre variáveis categóricas.
- **Lição 12** (análise de *cluster*). As análises de *cluster* permitem agrupar casos individuais (representados pelas linhas do *dataframe*) em

função da sua similaridade, que pode ser avaliada tanto em termos de variáveis escalares quando categóricas.

Enfim, se você está procurando um método capaz de descrever as múltiplas associações entre variáveis e indivíduos de um *dataframe* sem, com isso, supor relações preditivas ou causais, você provavelmente encontrará o que precisa nas próximas lições.

LIÇÃO 10.

ANÁLISE DE COMPONENTES PRINCIPAIS

A *análise de componentes principais*[1] (ing. *principal component analysis* – PCA) é uma técnica exploratória-descritiva que tem por objetivo gerar *p* variáveis novas (componentes principais) que sejam capazes de descrever, da melhor maneira possível, toda a informação disponível em *k* variáveis originais (chamadas indicadores), com $\boldsymbol{p < k}$. Esse método reduz a quantidade de variáveis necessárias à descrição dos dados em um *dataframe*, facilitando e qualificando a interpretação desses dados. Afinal, é muito mais interessante interpretarmos as relações entre duas componentes principais que considerar todas as relações possíveis entre dezenas de variáveis indicadoras!

Conforme discutido em outra oportunidade (cf. lição 5), existe uma interpretação geométrica que nos permite representar variáveis como

[1] A análise de componentes principais é um método muito simples. Contudo, justamente por essa simplicidade, optei por aprofundar um pouco mais a discussão dos seus fundamentos. Recomendo que, antes dessa lição, todos façam uma breve revisão de álgebra matricial (cf. apêndice) com especial atenção aos *teoremas da decomposição espectral*. Se nada fizer muito sentido, leia o texto sem se importar muito com as equações. O que você não entender agora, poderá entender em outro momento da sua vida. Não tenha pressa.

vetores no espaço. Nesse espaço, cada vetor pode ocupar uma dimensão própria. Em outras palavras, uma planilha com duas variáveis pode ser representada em um espaço bidimensional; três variáveis podem ser representadas por um diagrama tridimensional, mais variáveis demandarão um conjunto maior de dimensões para serem representadas...!

A redução de *k* variáveis indicadoras a *p* componentes principais corresponde a buscar uma quantidade menor de dimensões que seja capaz de descrever, da melhor maneira possível, toda a informação disponível no *dataframe*. Por essa razão, podemos dizer que a análise de componentes principais é um método de *redução dimensional*.

O que é uma redução dimensional?

O primeiro passo para compreender a redução dimensional é recuperar a interpretação geométrica introduzida na lição 5. Segundo essa interpretação, variáveis podem ser representadas por vetores no espaço de tal maneira que o cosseno do ângulo entre dois vetores seja igual à correlação entre as variáveis escalares que eles representam. Essa interpretação geométrica implica que duas variáveis descorrelacionadas serão representadas por vetores ortogonais (i.e., que fazem um ângulo reto entre si); variáveis com correlação positiva serão representadas por vetores que formam ângulo agudo; variáveis com correlação negativa serão representadas por vetores que formam um ângulo obtuso (cf. quadro 10.1). Quanto mais aberto ou mais fechado for o ângulo entre os vetores, mais próximo de −1 ou 1 será a correlação entre as variáveis que eles representam.

Quadro 10.1: Interpretação geométrica das correlações entre variáveis

Correlação entre variáveis	Ângulo entre variáveis
Positiva	Agudo
Nula	Reto
Negativa	Obtuso

Vamos visualizar isso mais concretamente? Pegue duas canetas. Considere que elas sejam a representação vetorial de duas variáveis. Ao unir as canetas pela base, é possível afastar ou aproximar as pontas formando diversos ângulos. Esses ângulos indicarão a correlação entre as variáveis representadas pelas canetas.

Caso queira, você pode testar todos os ângulos possíveis com as canetas em repouso sobre a mesa. Isso só é possível porque dois vetores não definem mais que duas dimensões no espaço. Vamos explorar essa questão um pouco mais profundamente? Pegue três canetas e tente organizá-las de tal modo que os ângulos entre elas sejam sempre retos. Você consegue perceber que essa configuração exige uma terceira dimensão? Seguindo esse raciocínio, um conjunto de *k* variáveis exigirá, inicialmente, um espaço k-dimensional para representar essas variáveis. Devido a essa interpretação geométrica, a redução da quantidade de variáveis necessárias para representar nossos dados é chamada redução dimensional.

Vamos fazer outro exercício. Pegue as três canetas e escolha uma configuração na qual as correlações entre as variáveis que elas representam sejam muito próximas a 1 ou −1. Nesse caso, conforme o quadro 10.1, todos os ângulos serão muito abertos ou muito fechados. Apesar de

três vetores serem capazes de definir 3 dimensões, essa configuração pode ser, com boa aproximação, reduzida a 1 dimensão!

A análise de componentes principais tem como ponto de partida as covariâncias entre todas as variáveis indicadoras. Seu propósito é determinar quantas e quais dimensões são capazes de descrever essas variáveis com boa aproximação. Como é possível perceber do nosso exemplo, a redução dimensional está intimamente relacionada aos padrões de associação entre as variáveis. Se as variáveis indicadoras são fracamente associadas (formando ângulos retos), a redução dimensional não terá muito sucesso. Porém, dependendo dos padrões de associação entre essas variáveis, a redução dimensional produzirá resultados impressionantes!

Exemplos da vida e da literatura

Como é possível perceber, a redução dimensional consiste em reunir, em poucas componentes, informações dispersas em um conjunto numeroso de variáveis muito associadas entre si. Ainda que o procedimento formal seja um pouco abstrato, todos nós lidamos diariamente com informações que só podem ser obtidas por meio de uma redução dimensional. Ou seja, todos lidamos com conceitos que resultam da combinação de indicadores mutuamente associados. Por exemplo, ao distinguir os cidadãos de uma cidade em função da classe, é usual levar em consideração um conjunto mais ou menos heterogêneo de indicadores (renda, escolaridade, ocupação, residência, acesso a bens e serviços, estilos de vida). É justamente por haver uma tendência de associação entre esses indicadores que faz sentido combiná-los no conceito mais abrangente de “classe” (Bourdieu, 1984). Se tais indicadores não estivessem associados, não faria

sentido tratá-los como manifestações específicas de algo mais abrangente. Da mesma maneira, ao classificar eleitores em direita e esquerda, liberais e progressistas, também estamos agrupando pessoas em função de indicadores que tendem a estar associados. De fato, é usual distinguir os cidadãos por seus posicionamentos com respeito ao aborto, à livre expressão sexual, à liberdade de cátedra, à liberdade de imprensa, ao financiamento público da educação. Ainda que esses posicionamentos sejam diferentes uns dos outros, não é surpreendente que estejam associados.

Por exemplo, ao realizar uma análise de componentes principais com informações de estudantes de um curso de geografia e seu professor de geologia, Bezzi (1999) percebeu a presença de uma imagem estereotipada de ciência marcada pela antítese entre física (considerada objetiva e rigorosa) e geociência (considerada subjetiva e aproximada). Em outra pesquisa, Brandsetter, Sandmann e Florian (2017) empregaram a análise de componentes principais para investigar como os estudantes compreendem informações disponíveis em representações visuais da biologia escolar. Em uma pesquisa sobre violência escolar e violência urbana, Garcia-Silva, Lima Junior e Caruso (2021) realizaram uma análise de componentes principais com dados das regiões administrativas do Distrito Federal para perceber que os indicadores de violência escolar não são redutíveis aos indicadores de violência urbana. Em outras palavras: regiões da cidade consideradas violentas podem abrigar escolas muito seguras e escolas consideradas violentas podem estar situadas em regiões consideradas seguras.

Como é possível perceber, o emprego da análise de componentes principais é muito diversificado. Todas as situações em que desejarmos reduzir um conjunto de indicadores escalares a poucas dimensões, a análise de componentes principais será muito relevante.

Definindo as componentes principais

Em síntese, a análise de componentes principais busca responder a seguinte questão de pesquisa: *quantas/quais componentes principais são necessárias para representar aproximadamente toda a informação disponível nas variáveis indicadoras?* Esse método toma como ponto de partida as matrizes de correlação e covariância, pois essas matrizes contêm toda a informação necessária à análise dos padrões de associação entre suas variáveis.

Por definição, as componentes principais *y* são *variáveis mutuamente ortogonais obtidas pela combinação linear das variáveis indicadoras x*. A combinação linear implica que a seguinte relação seja satisfeita (cf. apêndice):

$$\begin{cases} y_1 = a_{11}x_1 + a_{21}x_2 + \ldots + a_{k1}x_k \\ y_2 = a_{12}x_1 + a_{22}x_2 + \ldots + a_{k2}x_k \\ \vdots \\ y_k = a_{1k}x_1 + a_{2k}x_2 + \ldots + a_{kk}x_k \end{cases}$$

No sistema de equações antecedente, os coeficientes *a* são constantes que devem ser determinadas com base nos dados. Essas constantes (também chamadas cargas) permitem identificar os padrões de associação entre as variáveis indicadoras: *variáveis muito associadas entre si apresentarão cargas semelhantes*.

Usando notação matricial (cf., apêndice), o sistema de equações que determina as componentes principais pode ser escrito de maneira mais compacta:

$$\vec{y} = \mathbf{A}^T \vec{x}$$

Nesses termos, as cargas da combinação linear são representadas por uma matriz $\mathbf{A}^T$, que transforma o vetor das variáveis indicadoras $\vec{x}$ em um vetor de variáveis mutuamente ortogonais $\vec{y}$, chamadas componentes principais. É muito importante lembrar que as variáveis resultantes da operação anterior só serão chamadas componentes principais se forem mutuamente ortogonais. A propósito, essa exigência ortogonalidade implica que a covariância entre duas componentes principais seja sempre igual a zero:

$$cov\left(y_i, y_j\right) = 0, \quad \forall i \neq j$$

Em outras palavras, a ortogonalidade das componentes principais implica que sua matriz de covariâncias $\mathbf{\Lambda}_y$ seja diagonal:

$$\mathbf{\Lambda}_y = \begin{bmatrix} s_{y_1}^2 & 0 & \cdots & 0 \\ 0 & s_{y_2}^2 & & 0 \\ \vdots & & \ddots & \\ 0 & 0 & \cdots & s_{y_k}^2 \end{bmatrix}$$

De fato, os elementos fora da diagonal principal da matriz $\mathbf{\Lambda}_y$ devem ser nulos. Afinal, eles representam a covariância entre componentes principais diferentes. Já os elementos da diagonal, como representam a covariância de uma componente principal com ela mesma, resultam nas variâncias dessas componentes (e não serão nulos).

A condição de ortogonalidade permitirá que as componentes principais sejam interpretadas como propriedades não relacionadas entre si. Por exemplo, uma análise que modele o eixo direita-esquerda como ortogonal ao eixo liberal-conservador sugere que a posição de um cidadão em uma dimensão não permite predizer a posição com relação à outra. Se isso estiver correto, deve haver conservadores e liberais (nos costumes) tanto à direita quanto à esquerda do espectro político.

Um problema de diagonalização

O objetivo original da análise é determinar as componentes principais. Porém, percebemos a necessidade de determinar, também, a transformação linear $\mathbf{A}^T$ responsável por gerar essas componentes principais:

$$\vec{y} = \mathbf{A}^T \vec{x}$$

Ao aplicar algum conhecimento de álgebra matricial à equação anterior, é possível demonstrar que:

$$\mathbf{\Lambda}_y = \mathbf{A}^{\mathrm{T}} \mathbf{\Lambda}_x \mathbf{A}$$

Nesta equação, $\mathbf{\Lambda}_x$ é a matriz de covariâncias das variáveis indicadoras (ela pode ser obtida facilmente a partir dos dados). $\mathbf{\Lambda}_y$ é a matriz de covariâncias das componentes principais, que deve ser diagonal. Em outras palavras, a matriz da transformação linear *deve diagonalizar* a matriz das covariâncias dos indicadores.

$$\mathbf{A}^{\mathrm{T}} : \mathbf{\Lambda}_x \rightarrow \mathbf{\Lambda}_y$$

A diagonalização de uma matriz é um dos problemas mais importantes da álgebra linear. Ela está relacionada ao chamado *problema de autovalor e autovetor*. Conceitualmente, a diagonalização pode ser um pouco desafiadora. Porém, do ponto de vista computacional, é uma operação simples. Os autovalores e autovetores (ou vetores e valores próprios) de uma matriz podem ser facilmente obtidos no R pela função **`eigen`**. Para ver o funcionamento, considere o seguinte exemplo:

```
Lambda.x
##      [,1] [,2] [,3]
## [1,]    2    1   -1
## [2,]    1    2    0
## [3,]   -1    0    2
```

Primeiramente, criamos uma matriz A que possa ser interpretada como uma matriz de covariâncias. Essa é a matriz que desejamos diagonalizar. Tanto o resultado da diagonalização quanto a transformação linear responsável por essa diagonalização podem ser obtidos pela função **`eigen`** da biblioteca de base:

```
eigen(Lambda.x)
## eigen() decomposition
## $values
## [1] 3.4142136 2.0000000 0.5857864

## $vectors
##            [,1]      [,2]       [,3]
## [1,] -0.7071068 0.0000000  0.7071068
## [2,] -0.5000000 0.7071068 -0.5000000
## [3,]  0.5000000 0.7071068  0.5000000
```

No *output* acima, `$values` lista os autovalores da matriz Lambda.x, enquanto `$vectors` lista seus respectivos autovetores. Por aplicação direta dos teoremas da decomposição espectral (cf. apêndice), sabemos que: *i)* os autovalores da matriz que desejamos diagonalizar serão iguais à diagonal principal da matriz diagonalizada; *ii)* a matriz dos autovetores será igual à matriz **A** cuja transposta determina a transformação linear.

Essas informações podem ser verificadas fazendo $\mathbf{A}^{\mathrm{T}}\mathbf{\Lambda}_{x}\mathbf{A}$:

```
A = eigen(Lambda.x)$vectors

t(A) %*% Lambda.x %*% A

##               [,1]         [,2]         [,3]
## [1,]  3.414214e+00 0.000000e+00 0.000000e+00
## [2,]  1.110223e-16 2.000000e+00 1.110223e-16
## [3,] -8.326673e-17 1.387779e-16 5.857864e-01
```

Conforme esperado, obtivemos uma matriz aproximadamente diagonal com entradas aproximadamente iguais aos autovalores obtidos.

A estrutura do preconceito escolar

Nesta lição, vamos ilustrar o funcionamento da análise de componentes principais com dados coletados na Pesquisa de Ações Discriminatórias no Âmbito Escolar (www.inep.gov.br). Trata-se de uma pesquisa realizada 2008 com o propósito de investigar a incidência de preconceito e discriminação nas escolas públicas brasileiras — estaduais e municipais.

```
load(file = "discrimina.dat")
```

O *dataframe* **`discrimina`** traz um conjunto selecionado de variáveis. A pesquisa contou com a participação de mais de 15 mil atores escolares (entre estudantes, pais, professores, diretores, secretários, porteiros, merendeiros) distribuídos em 500 escolas públicas brasileiras do ensino fundamental ao médio. As variáveis selecionadas para compor o *dataframe* são respostas a itens do tipo Likert. Nesses itens, os participantes da pesquisa assinalam seu grau de concordância com uma série de afirmações. No caso, elas foram armazenadas como rótulos das variáveis. Os códigos das variáveis no estudo original foram mantidos:

```
library(Hmisc)
label(discrimina)
```

- V0199. “Professor de escola de periferia não precisa ser bem preparado”
- V0210. “Caso exista um homossexual na sala de aula, os pais devem transferir seu filho da escola”
- V0184. “O estudante que entra mais velho na escola tem mais dificuldade para aprender”
- V0152. “Crianças brancas aprendem mais rápido que crianças negras”
- V0196. “Os estudantes surdos deveriam estudar numa escola especial para portadores de deficiências auditivas”
- V0171. “Os brancos merecem trabalhos mais valorizados do que os negros”
- V0197. “Os brancos são superiores aos negros”
- V0216. “Alunos homossexuais deveriam ser afastados da escola”
- V0224. “Existem trabalhos que devem ser realizados apenas por mulheres”
- V0190. “A mulher é mais habilidosa para cuidar da casa”
- V0155. “Os estudantes cegos deveriam estudar numa escola especial para portadores de deficiência visual”
- V0181. “Existem trabalhos que devem ser realizados apenas por homens”

A análise de componentes principais desses dados deve ser capaz de explicitar padrões de associação entre as variáveis mostradas anteriormente. Quanto mais associadas as variáveis estiverem umas às outras, menor será a quantidade de dimensões necessárias para descrevê-las. A depender das relações entre essas variáveis, será possível agrupá-las em conjuntos muito associados em si, mas pouco associados entre si.

A propósito, quais são os seus palpites? Na sua opinião, uma pessoa que concorda com a primeira afirmação provavelmente concordará com quais outras?

Para a análise de componentes principais, precisamos obter uma matriz de correlações. Nesse caso, é tentador fazer:

```
cor(discrimina)
```

No entanto, temos um problema. A correlação de Pearson, calculada pela função **`cor`** (cf. lição 5), está definida para variáveis estritamente escalares e as *respostas a itens tipo-Likert* não formam uma escala propriamente dita (Mair, 2018). Para que uma variável seja considerada escalar, deve ser razoável presumir que intervalos iguais dessa variável representem distâncias iguais no mundo real. Por exemplo, uma pessoa com 1,70 m de altura deve estar equidistante de pessoas com 1,80 m ou 1,60 m. Um estudante que tenha obtido 500 pontos no Enem deve estar equidistante de outros que tenham obtido 520 ou 480 pontos.

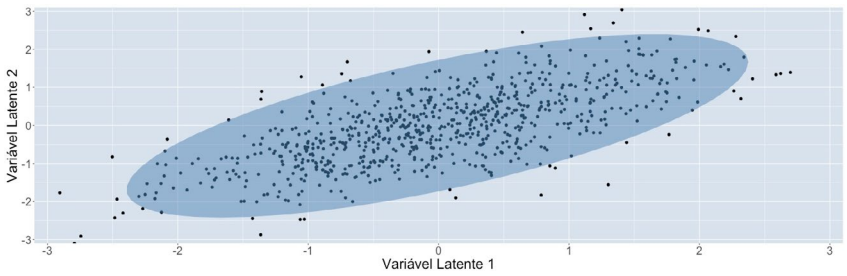
Respostas tipo-Likert são claramente não intervalares. Considere, por exemplo, um item de quatro pontos com os seguintes rótulos: “discordo muito”, “discordo”, “concordo” e “concordo muito”. Essa codificação não permite presumir que a distância entre as categorias seja a mesma. Por exemplo, a distância entre marcar “discordo” e “concordo” pode ser maior que a distância entre “concordo” e “concordo muito”. Por essa razão, não é recomendado calcular a correlação de Pearson entre itens tipo-Likert (Mair, 2018).

Correlações tetracóricas e policóricas

Até o presente momento, ao mencionarmos “correlação” estivemos nos referindo especificamente à correlação de Pearson. Definida como a covariância padronizada (cf. lição 5), ela é adequada para avaliar a associação linear entre variáveis escalares. Porém, para estimar a associação entre variáveis binárias ou ordinais, a correlação de Pearson não é a solução mais adequada. Nesses casos, empregamos as correlações *tetracórica* (para variáveis binárias) ou *policórica* (para variáveis ordinais com mais de dois níveis).

Tais correlações poderão ser interpretadas *como se fossem* correlações de Pearson, mas a maneira de obtê-las é bastante diferente (Mair, 2018). Em primeiro lugar, precisamos supor que os níveis dos *indicadores categóricos manifestos* (i.e., variáveis binárias ou ordinais efetivamente medidas) resultem de *indicadores escalares latentes*, cuja correlação de Pearson nós desejamos conhecer. Se esses indicadores latentes fossem conhecidos, nós seríamos capazes de gerar um diagrama de dispersão tal como no gráfico 10.1.

Gráfico 10.1: Exemplo de diagrama de dispersão entre dois indicadores latentes



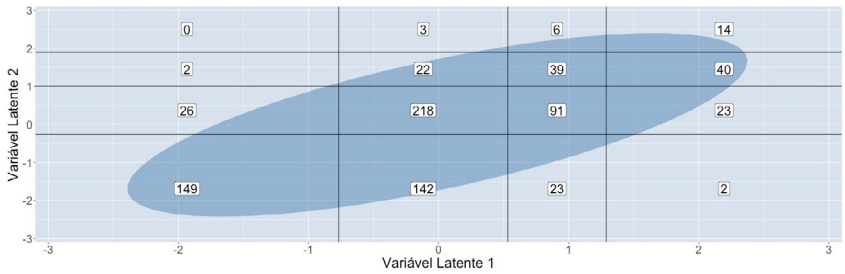
No entanto, dado que os indicadores manifestos não são escalares, não é possível gerar o diagrama de dispersão diretamente. O que podemos fazer é gerar uma tabulação cruzada (cf. lição 6). A tabela 10.1 ilustra o cruzamento de duas variáveis tipo-Likert cuja correlação nós desejamos saber:

Tabela 10.1: Tabulação cruzada das respostas dadas a dois itens tipo-Likert

	Discordo Muito	Discordo	Concordo	Concordo Muito
Concordo Muito	0	1	12	14
Concordo	3	33	41	31
Discordo	43	204	85	26
Discordo Muito	128	157	21	1

O algoritmo que calcula as correlações tetracórica e policórica tem as tabulações cruzadas das variáveis como ponto de partida. Ele supõe que as respostas dadas pelos participantes aos itens resultem diretamente dos valores assumidos pelos indicadores latentes. Isso é mais fácil de entender observando o gráfico 10.2.

Gráfico 10.2: Exemplo de como a correlação policórica é obtida



O gráfico 10.2 mostra o mapeamento da tabulação cruzada (tabela 10.1) no diagrama de dispersão hipotético (gráfico 10.1). O indicador latente 1 (no eixo horizontal) é a representação escalar do grau de concordância dos sujeitos com relação ao primeiro item tipo-Likert. Como é possível perceber, estamos supondo que o grau de concordância varie continuamente no interior de cada nível. Ou seja, entre os que disseram discordar muito do item 1, há pessoas que discordam mais, outras menos. Porém, quando o grau de concordância ultrapassa um limiar (nesse exemplo, aproximadamente igual a – 0,8), as respostas dadas ao item saltam para o próximo nível. De fato, com base nesse modelo, é possível especificar todos os pontos de virada em cada item do questionário. Esses valores de transição são designados pela letra grega *tau*.

A partir dos valores de tau e da tabela de contingência, é possível estimar como deve ser a correlação entre os indicadores latentes. Tais correlações são designadas pela letra grega *rho*. Quanto maior for o acúmulo de casos ao longo da diagonal da tabela, maior será a correlação. Dessa maneira, é possível obter uma medida de associação das variáveis em uma tabela de contingência que pode ser interpretada *como se fosse* uma correlação de Pearson. Evidentemente, esse procedimento só faz sentido para indicadores binários ou ordinais. Indicadores categóricos não ordinais (e.g., etnia, nacionalidade, ocupação, local de residência, nome da escola) precisam ser tratados de outra maneira (cf. lição 11). O procedimento descrito aqui é chamado policórico porque há vários núcleos de informação envolvidos (cf. gráfico 10.2). Quando os indicadores manifestos são variáveis binárias, a tabulação cruzada produz quatro núcleos e o método passa a ser chamado tetracórico.

Obtendo a matriz das correlações entre itens tipo-Likert

Para calcular as correlações policóricas, basta empregar a função **`polychoric`** da biblioteca *psych*. O resultado é uma estrutura de dados que contém uma matriz de correlações *rho* e uma tabela com os valores *tau* empregados na extração dessa matriz.

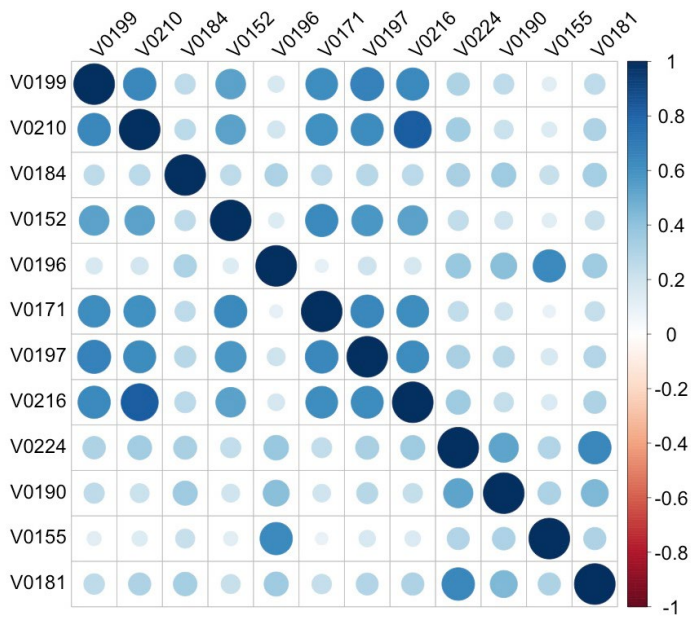
```
library(psych)}

corr = polychoric(discrimina)
corr

## Call: polychoric(x = discrimina)
## Polychoric correlations
##       V0199 V0210 V0184 V0152 V0196 V0171 V0197 V0216 V0224 V0190 V0155
V0181
## V0199 1.00
## V0210 0.64  1.00
## V0184 0.25  0.26  1.00
## V0152 0.54  0.53  0.26  1.00
## V0196 0.17  0.19  0.31  0.15  1.00
## V0171 0.61  0.61  0.25  0.63  0.11  1.00
## V0197 0.67  0.63  0.28  0.58  0.20  0.65  1.00
## V0216 0.63  0.83  0.27  0.53  0.18  0.61  0.62  1.00
## V0224 0.31  0.35  0.32  0.25  0.37  0.25  0.33  0.36  1.00
## V0190 0.25  0.22  0.36  0.20  0.41  0.19  0.27  0.23  0.52  1.00
## V0155 0.13  0.14  0.22  0.13  0.63  0.10  0.17  0.15  0.30  0.31  1.00
## V0181 0.26  0.30  0.34  0.22  0.35  0.23  0.29  0.31  0.64  0.44  0.30
1.00
##
##  with tau of
##           1     2      3
## V0199  0.80  1.29  1.631
## V0210  0.83  1.42  1.760
## V0184 -0.34  0.24  1.085
## V0152  0.79  1.55  1.973
```

```
## V0196 -0.59 -0.19  0.332
## V0171  0.94  1.54  1.821
## V0197  0.77  1.30  1.725
## V0216  0.80  1.42  1.757
## V0224 -0.15  0.25  0.767
## V0190 -0.60 -0.23  0.281
## V0155 -0.85 -0.49 -0.028
## V0181 -0.28  0.14  0.713
```

A matriz de correlações pode ser visualizada empregando a mesma função **`corrplot`** conforme veremos no diagrama 10.5.

```
library(corrplot)
corrplot(corr$rho, type = "full", tl.col = "black")
```

Diagrama 10.5: Representação gráfica da matriz de correlações



Conforme já foi antecipado, essa matriz de correlações contém todas as informações necessárias para realizar a análise. Contudo, por se tratar de um exemplo da vida real, o padrão de associação entre os itens não é evidente.

Quantas dimensões devemos reter?

Em qualquer método de redução dimensional, o analista precisa *decidir* qual é a menor quantidade de dimensões necessárias para representar a informação disponível da melhor maneira possível. Trabalhar com poucas dimensões torna os dados mais interpretáveis, mas aumenta a perda de informação. Portanto, estamos sempre diante de um *trade-off*, uma situação em que precisamos escolher uma coisa em detrimento de outra. Nesta seção, vamos discutir alguns dos critérios que podem orientar o analista a decidir quantas componentes principais serão retidas e quantas serão descartadas. Com experiência, o leitor perceberá que esses critérios geralmente deixam uma margem saudável de negociação dentro da qual o analista pode arbitrar em favor do que lhe parece ser conceitualmente mais consistente.

Para decidir quantas dimensões precisamos reter, é usual inspecionar os *autovalores* da matriz Lambda-x, pois, como eles são os elementos da diagonal principal de Lambda-y' eles informam a variância das componentes principais. Essa informação pode ser obtida acrescentando **`$values`** ao *output* da função **`eigen`**.

```
PCA$values

##  [1] 5.0464303 2.0940492 0.9725321 0.7623058 0.6048275 0.5537262 0.4361243
##  [8] 0.3655395 0.3452595 0.3309566 0.3182188 0.1700302
```

Por padrão, os *autovalores* são sempre dispostos em ordem decrescente. A primeira componente tem, portanto, o maior autovalor. A última, o menor. Como esses autovalores informam a variância de cada componente principal, eles podem ser considerados uma *medida da quantidade de informação original representada*. Assim, componentes principais com maior variância carregam uma quantidade maior de informação.

Além disso, por serem variâncias, os autovalores da matriz Lambda-x podem ser somados uns aos outros, produzindo a *variância total dos dados*:

```
sum(PCA$values)

##  [1] 12
```

A saber, a variância total explicada por todas as componentes principais será sempre igual à soma das variâncias dos indicadores. Portanto, para cada componente principal que descartamos, abrimos mão de parte da informação disponível. Contudo, como as últimas componentes têm variância muito pequena, a informação perdida é compensada pelo ganho em simplicidade e interpretabilidade.

Observe, também, que a variância total resultou igual à quantidade de variáveis. Isso ocorre porque a matriz **`corr$rho`** que inserimos na função **`eigen`** é, na verdade, uma *matriz de correlações*. Como a correlação de qualquer variável consigo mesma é igual a 1 unidade (cf. lição 5), a soma das correlações de k indicadores resultará sempre igual a k. Por outro lado, se tivéssemos alimentado a função **`eigen`** com uma matriz de covariâncias não padronizadas, essa igualdade não seria verdadeira.

Em qualquer situação, calcular o percentual de informação representada em cada componente principal é bastante simples. Basta dividir cada um de seus autovalores pela soma de todos eles:

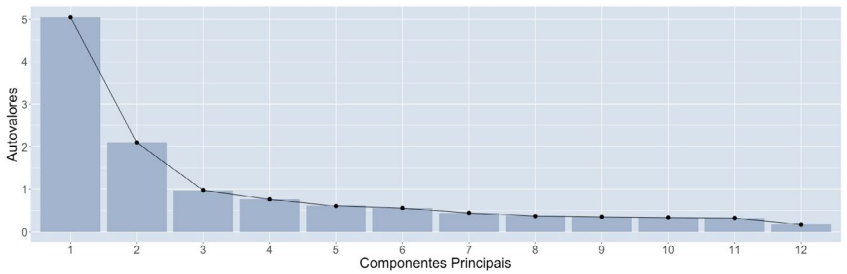
```
PCA$values/sum(PCA$values)

##  [1] 0.42053586 0.17450410 0.08104434 0.06352549 0.05040229 0.04614385
##  [7] 0.03634369 0.03046162 0.02877162 0.02757972 0.02651823 0.01416918
```

Acima, fica evidente o quanto a maior parte da informação está concentrada nas primeiras componentes principais. A grande maioria das componentes descreve menos de 5% da informação disponível e, por isso, podem ser eliminadas sem problema algum. Contudo, ainda nos resta alguma dúvida sobre a quantidade de dimensões que devem ser retidas. Devemos manter somente a primeira? As duas primeiras? As três primeiras?

Para tomar essa decisão, é usual gerar um gráfico dos autovalores (ou dos percentuais da variância explicada). Esse gráfico costuma ser chamado *scree plot* (gráfico 10.3):

```
barplot(PCA$values)
```

Gráfico 10.3: *Scree plot* dos itens de discriminação escolar



A palavra "*scree*" designa a sedimentação de detritos junto à base de uma formação rochosa. A intenção do analista ao gerar um *scree plot* é, portanto, distinguir duas regiões: na região de "rocha", os pontos estão claramente mais elevados; na região de "detrito", as diferenças de altura

são pequenas, lembrando o terreno sedimentar aos pés da montanha. Ao analista, recomenda-se reter a "rocha" e desprezar os detritos. No gráfico 10.3, todas as componentes principais a partir da terceira estão mais ou menos à mesma altura e, por isso, podem ser interpretadas como *scree*. As duas primeiras componentes principais estão mais elevadas com relação às demais. Por isso, devem ser retidas.

Esse critério foi introduzido por Raymond Catell em 1966. Apesar de ser criticado pela grande liberdade de interpretação que ele confere ao analista, o critério de Catell é provavelmente o mais empregado para arbitrar uma redução dimensional. Como critério alternativo, podemos seguir a recomendação de Kaiser, que consiste em descartar todas as componentes principais que carreguem menos informação que um item carregaria. Quando a análise é realizada com a matriz de correlações, esse critério consiste em eliminar todas as componentes principais com autovalor menor que 1 unidade. Para nossa sorte (isso nem sempre ocorre), os dois critérios recomendam manter as duas primeiras componentes principais.

Há, é claro, o critério da interpretabilidade. Não adianta muito reter uma quantidade de componentes principais que não conseguiremos interpretar. Portanto, é importante que o analista leve em consideração a possiblidade de interpretação das componentes ao decidir quantas devem ser retidas e quantas devem ser eliminadas.

Em síntese, os principais critérios para arbitrar uma redução dimensional são os seguintes:

1. *critério de Catell* – observe o *scree* plot e retenha as componentes principais cujos autovalores estão visivelmente acima dos seguintes.
2. *critério de Kaiser* – elimine as componentes principais em que os autovalores carreguem menos informação que um item carregaria.

3. *critério de interpretabilidade* – a melhor escolha é aquela que, observados os critérios puramente estatísticos, você é capaz de interpretar.

Levando os três critérios em consideração, minha decisão é por manter duas componentes principais. Juntas, elas representam 59,5% da variância total dos dados.

Investigando os autovetores

Na seção anterior, aprendemos alguns critérios e ferramentas capazes de orientar nossa decisão sobre quantas componentes principais precisamos reter para descrever os dados em um *dataframe*. Fizemos isso examinando os *autovalores* da matriz de covariâncias Lambxa-x com a ajuda de um diagrama chamado *scree plot*.

A partir de agora, estamos em condições de interpretar as componentes principais. Isso será feito analisando a transformação linear que gerou essas componentes:

$$\vec{y} = \mathbf{A}^T \vec{x}$$

Como já sabemos, a matriz **A** é igual à matriz dos *autovetores* de Lambda-x. Portanto, **A** pode ser obtida fazendo:

```
PCA$vectors
```

Na matriz gerada por esse comando, as *colunas representam os autovetores*. Ou seja, a j-ésima coluna descreve a composição da j-ésima componente principal. Como decidimos reter as duas primeiras componentes, somente as duas primeiras colunas da matriz A devem ser consideradas:

```
PCA$vectors[,c(1,2)]
```

Tabela 10.2: Relação entre componentes principais e variáveis indicadoras

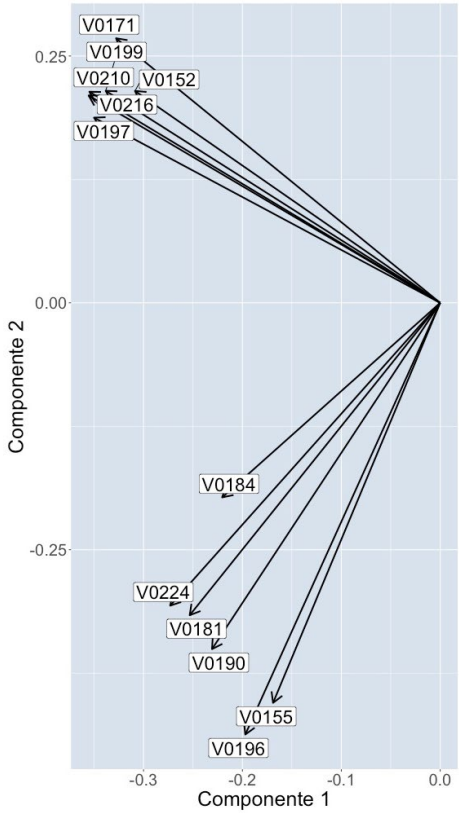
	Componente 01	Componente 02
Indicador 1	-0.3382653	0.2147261
Indicador 2	-0.3547669	0.2137785
Indicador 3	-0.2204875	-0.1970209
Indicador 4	-0.3086623	0.2145830
Indicador 5	-0.1971138	-0.4371468
Indicador 6	-0.3277785	0.2679174
Indicador 7	-0.3501153	0.1875723
Indicador 8	-0.3555505	0.2099962
Indicador 9	-0.2730433	-0.3066519
Indicador 10	-0.2307418	-0.3506348
Indicador 11	-0.1690405	-0.4050934
Indicador 12	-0.2533206	-0.3163681

Essas colunas determinam o peso de cada indicador na composição das duas primeiras componentes principais. Em princípio, devemos distinguir os indicadores de acordo com suas cargas. *Indicadores mutuamente associados devem apresentar cargas semelhantes*. De fato, a maneira mais simples de fazer isso é gerando um diagrama de dispersão dos dados anteriores:

```
plot(PCA$vectors[,c(1,2)], type = “n”, asp=1)
text(PCA$vectors[,c(1,2)], labels = PCA$names)
```

O resultado está publicado no gráfico 10.4. Como é possível perceber, há claramente dois grupos de itens altamente correlacionados em si, mas pouco correlacionados entre si (i.e., aproximadamente ortogonais).

Gráfico 10.4: Cargas das variáveis nas componentes principais



No *primeiro grupo* (acima) estão as seguintes afirmações:

- V0152. “Crianças brancas aprendem mais rápido que crianças negras”
- V0171. “Os brancos merecem trabalhos mais valorizados do que os negros”
- V0197. “Os brancos são superiores aos negros”
- V0199. “Professor de escola de periferia não precisa ser bem preparado”
- V0210. “Caso exista um homossexual na sala de aula, os pais devem transferir seu filho da escola”
- V0216. “Alunos homossexuais deveriam ser afastados da escola”

No *segundo grupo* (abaixo) estão as demais afirmações:

- V0155. “Os estudantes cegos deveriam estudar numa escola especial para portadores de deficiência visual”
- V0181. “Existem trabalhos que devem ser realizados apenas por homens”
- V0184. “O estudante que entra mais velho na escola tem mais dificuldade para aprender”
- V0190. “A mulher é mais habilidosa para cuidar da casa”
- V0196. “Os estudantes surdos deveriam estudar numa escola especial para portadores de deficiências auditivas”
- V0224. “Existem trabalhos que devem ser realizados apenas por mulheres”

Ao observar o conteúdo dos itens, percebemos que, no primeiro grupo, foram agrupados os temas “racismo” e “homofobia”. A possibilidade de agrupar esses itens informa que, nas escolas públicas brasileiras, as pessoas mais homofóbicas são provavelmente as mais racistas. O segundo grupo difere do primeiro não só pelo tema (gênero, idade e deficiência), mas pela sua formulação. Enquanto os primeiros itens tratam abertamente de superioridade, merecimento e banimento, construindo um tipo de narrativa do ódio; os itens do segundo grupo têm uma formulação mais “amena” e sua declaração não seria tipificada como crime. Uma análise das estatísticas descritivas desses itens também revela que a quantidade de pessoas concordando com os itens do segundo grupo é substancialmente maior:

```
Hmisc::describe(discrimina)
```

Sobre as bibliotecas disponíveis

Nesta lição, realizamos a análise de componentes principais de maneira relativamente ousada para um texto introdutório: levando a matriz

de correlações à função **eigen**, da biblioteca de base. Do ponto de vista de cômputo, as operações que fizemos são simples. Porém, o entendimento dessas operações pode levar algum tempo para se sedimentar. É provável que o leitor precise ler esta lição várias vezes, retornando à lição 5, ao apêndice e buscando informações complementares na internet.

Alguns usuários preferem empregar funções especializadas para realizar a análise de componentes principais. Todas têm suas possibilidades e limites. Além das funções mais clássicas (**princomp** e **prcomp**), há as que permitem gerar gráficos de forma mais amigável (**PCA**) e as que permitem incluir diversas operações auxiliares, mas não geram gráficos tão interessantes (**principal**).

Dado que essas funções podem ajudar muito a vida do analista, eu não poderia deixar de dedicar algumas palavras a elas:

- A função **`princomp`** (da biblioteca **`stats`**) faz uma análise de componentes principais calculando a matriz de covariâncias (ou correlações) e levando o resultado na função **`eigen`**. Em outras palavras, ela emprega o *teorema da decomposição espectral de Jordan* (cf. apêndice). Contudo, ela não inclui a correção de viés para covariâncias amostrais, tampouco calcula correlações tetrapolicóricas. Trata-se de uma função obsoleta, preservada por razões históricas. De fato, em todos os contextos nos quais ela pode ser empregada, o analista deve dar preferência à função `prcomp`.
- A função **`prcomp`** (da biblioteca **`stats`**) faz uma análise de componentes principais calculando a matriz de covariâncias (ou correlações) e levando o resultado na função **`svd`**. Em outras palavras, ela emprega o *teorema da decomposição do valor singular*, uma generalização do teorema de Jordan para fatorar matrizes retangulares (cf. Apêndice). Além disso, ela inclui a correção de viés para covariâncias amostrais, mas não

calcula correlações tetrapolicóricas. Por todas essas razões, os resultados obtidos com a função `prcomp` são um pouco mais confiáveis.

- A função **`PCA`** (da biblioteca **`FactoMineR`**) tem todas as vantagens da função *prcomp* (i.e., a correção de viés amostral e o emprego da decomposição do valor singular). Além disso, ela permite distinguir entre variáveis principais e suplementares. Posteriormente discutiremos o que isso significa (cf. lição 11). Contudo, sua principal vantagem está em seus *outputs* gráficos. Com funções da biblioteca **`factoextra`**, é possível gerar *scree plots*, mapas das variáveis, dos indivíduos, e *biplots*. Tais opções gráficas estão cuidadosamente descritas e exemplificadas em *Statistical Tools for High Thought Data Analysis*, STHDA (www.sthda.com). No entanto, assim como as duas funções anteriores, a função **`PCA`** não suporta correlações tetracóricas ou policóricas.
- A função **`principal`** (da biblioteca **`psych`**) é bastante versátil do ponto de vista das opções de cômputo. Ela calcula internamente covariâncias, correlações de Pearson, correlações tetracóricas, policóricas ou mistas, aplicando o teorema da decomposição de Jordan — exatamente como fizemos aqui. Além disso, ela incorpora as rotações da análise fatorial, que serão discutidas posteriormente (cf. lição 14). A documentação da função `principal` é cuidadosa ao explicitar que as componentes rodadas não podem mais ser chamadas componentes principais. Elas são, contudo, muito mais fáceis de interpretar. Infelizmente, os *outputs* gráficos nativos da função `principal` não são tão atraentes quanto os da função `PCA`, mas esse problema pode ser contornado. Usuários mais interessados em psicometria provavelmente encontrarão o que precisam nesta e em outras funções da biblioteca *psych*.

O quadro a seguir resume essas afirmações.

Quadro 10.2: Comparando as funções mais populares para análise de componentes principais

	princomp (stats)	**prcomp** (stats)	**PCA** (FactoMineR)	**principal** (psych)
Emprega decomposição do valor singular?	-	Sim	Sim	-
Inclui correção de viés amostral?	-	Sim	Sim	Sim
Permite incluir variáveis suplementares?	-	-	Sim	-
Apresenta *output* gráfico completo e amigável?	-	-	Sim	-
Permite analisar variáveis binárias e ordinais?	-	-	-	Sim
Inclui rotações das componentes principais?	-	-	-	Sim

Com o passar do tempo, nós vamos ganhando familiaridade com algumas bibliotecas, pois cada função é desenhada para operar em conjunto com outras. Assim, se sua formação está mais ligada à psicometria, provavelmente encontrará tudo que precisa na biblioteca *psych.* Portanto, ao ganhar familiaridade com essa biblioteca, você provavelmente preferirá a função `principal`. Se seu objetivo é trabalhar com mineração de dados, análise de correspondência e análise de *cluster*, os gráficos gerados pela função `PCA`, da biblioteca `FactoMineR`, serão tremendamente intuitivos de interpretar.

Ainda sobre os *outputs* gráficos, a maioria dos métodos de análise emprega a função **`plot`**, da biblioteca de base. Como você deve ter percebido, mesmo os gráficos mais amigáveis tendem a ser um pouco "duros" ou "feios", diferentes daquilo que encontramos em publicações de referência. Porém, já que você conseguiu ler este livro até aqui,

eu gostaria de sugerir os métodos gráficos baseados na biblioteca **`ggplot2`** (www.r-graph-gallery.com). Com eles, o programador tem opções praticamente ilimitadas de edição gráfica!

Diferente da programação estatística do R, a programação gráfica não requer, por parte do analista, a mesma profundidade de entendimento. Portanto, não faz muito sentido discutir as funções dessa biblioteca neste livro. A internet está repleta de exemplos de código que podem ser copiados e adaptados por qualquer pessoa que tenha desenvolvido alguma familiaridade com a linguagem de programação do R. É claro que esse aprendizado tomará tempo e esforço, mas você será capaz de produzir gráficos absolutamente brilhantes! Com esse tipo de controle sobre os *outputs* gráficos, você poderá escolher as funções da sua análise não tanto pelos gráficos que geram, mas pela análise estatística propriamente dita.

Revisitando a lição

Nessa lição, você aprendeu que:

1. a análise de componentes principais é um método exploratório-descritivo de redução dimensional que busca gerar um conjunto de variáveis novas (componentes principais) capaz de descrever sinteticamente a informação disponível em um conjunto de variáveis originais (indicadores);
2. as componentes principais simplificam, ao mesmo tempo, a representação dos indicadores e dos indivíduos, explicitando as relações de associação entre as variáveis indicadoras e as relações de similaridade entre as instâncias individuais;
3. as componentes principais *y* são, por definição, mutuamente ortogonais e resultam da combinação linear das variáveis indicadoras *x*;

4. em consequência da ortogonalidade imposta às componentes principais, é possível extraí-las a partir da *diagonalização da matriz de covariâncias* dos indicadores;
    - a matriz diagonalizada Lambda-y tem os *autovalores* de Lambda-x em sua diagonal principal;
    - a matriz dos *autovetores* de Lambda-x informa a transformação de diagonalização;
5. os autovalores de Lambda-x são as variâncias das componentes principais e podem ser empregados para orientar a decisão do analista sobre quantas dimensões são necessárias para representar a informação disponível no *dataframe*;
6. o *scree plot* é o diagrama que tipicamente orienta essa decisão, pois favorece a comparação visual da quantidade de informação representada em cada componente principal;
7. as cargas dos indicadores podem se visualizadas nos vetores--coluna da matriz dos *autovetores* de Lambda-x;
8. indicadores muito associados são carregados da mesma maneira nas componentes principais;
9. enfim, a análise permite agrupar as variáveis indicadoras em grupos muito associados em si, mas pouco associados entre si.

Atividade proposta

Baixe a planilha completa da Pesquisa de Ações Discriminatórias no Âmbito Escolar (www.inep.gov.br) e tente fazer a mesma análise de correspondência com todas as variáveis. A que conclusões você é capaz de chegar?

LIÇÃO 11.

ANÁLISE DE CORRESPONDÊNCIA

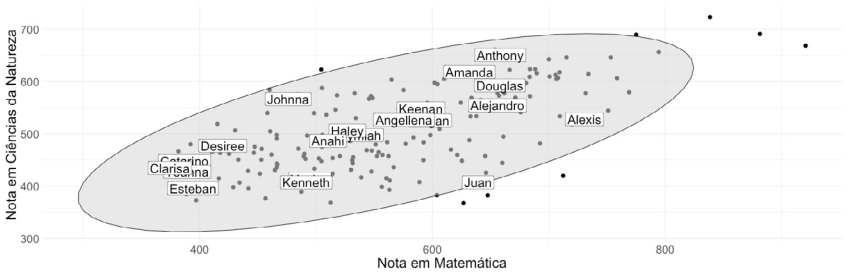
A *análise de correspondência* é um método particularmente útil para representar dados categóricos *como se fossem posições em um mapa* (Greenacre, 2007). Por definição do que significa ser uma variável categórica (cf. lição 1), elas não têm propriedades métricas e, por isso, não podem ser interpretadas como se fossem posições. Ao menos inicialmente, essa restrição impede representações tipo-mapa. Contudo, é justamente esse impedimento que a análise de correspondência permite contornar.

Considere, por exemplo, as capitais dos estados brasileiros. Todas têm posições angulares (latitude e longitude) mais ou menos bem definidas. Portanto, a menos de distorções devidas à esfericidade do planeta, é perfeitamente possível posicionar algumas dessas capitais em um mapa (cf. gráfico 11.1) que permita, entre outras coisas, avaliar visualmente a posição de cada cidade com relação às demais (umas mais ao norte, outras mais ao sul; umas mais a leste, outras mais a oeste).

Gráfico 11.1: Posição de algumas capitais brasileiras

Por analogia, qualquer par de variáveis escalares que tenham a mesma unidade de medida pode ser representado em um gráfico tipo-mapa. Considere, por exemplo, as notas obtidas pelos estudantes na prova de matemática e ciências da natureza do Enem de 2018. Assim como as cidades estão dispersas na superfície do planeta, cada indivíduo está *posicionado* com relação aos demais no diagrama de dispersão (gráfico 11.2). A "nordeste", estão os estudantes com maior desempenho nas duas avaliações, a "sudoeste", os de menor desempenho. Além disso, é possível avaliar as distâncias entre quaisquer dois indivíduos, identificando pares mais afastados ou mais próximos. Todas essas considerações só são possíveis porque os indivíduos foram *posicionados*, uns com relação aos outros, por meio de variáveis com propriedades métricas.

Gráfico 71.2: Mapa das notas em ciências da natureza e matemática



Representações tipo-mapa são, em princípio, impossíveis para variáveis categóricas. Afinal, com base em qual critério nós posicionaríamos brancos, pretos, pardos e indígenas, homens e mulheres, cidadãos com diversos graus de renda e escolaridade? Por definição, mapas exigem posicionar as instâncias individuais, umas com relação às outras, por meio de escalas. De fato, *mapas só serão capazes de representar qualidades se elas puderem ser traduzidas em posições relativas*. Portanto, a representação de variáveis categóricas (sexo, cor, renda, ocupação, residência, grau e tipo de escolaridade) supõe ser possível construir posições relativas a partir dessas categorias. Ou seja, *deve ser possível pensarmos as qualidades das categorias como se fossem posições no espaço*. Em sociologia da educação, esse posicionamento dos indivíduos em escalas de privilégio frente à escola (onde uns estão mais acima ou abaixo, à esquerda ou à direita), é bastante usual (Bourdieu, 1984). *Operacionalizar representações espaciais de variáveis categóricas* é precisamente o desafio que a análise de correspondência pretende resolver.

Uma vez que a intenção geral da análise de correspondência esteja declarada, podemos discutir outras questões ainda abrangentes, mas um pouco mais específicas — tal como *ortogonalidade* e *colinearidade* das variáveis categóricas. Em princípio, ortogonalidade e colinearidade estão definidas para variáveis escalares e resultam da interpretação geométrica dada à correlação

entre essas variáveis (cf. lição 5). Porém, em termos mais gerais, a colinearidade está relacionada à possibilidade de predizer uma variável a partir de outra. Por exemplo, com respeito a uma população de estudantes, se soubermos a escolaridade do pai, poderemos predizer, com boa aproximação, a escolaridade da mãe e a renda familiar (cf. lição 6). Em outras palavras, essas variáveis são aproximadamente colineares. Ou seja, a dimensão do espaço que posiciona os estudantes com relação à escolaridade do pai é provavelmente a mesma que posiciona esses estudantes com relação aos demais marcadores de classe. Contudo, conhecer os indicadores de classe não aumenta nossa chance de predizer o sexo dos estudantes, pois a chance de nascer menino ou menina não depende da classe. Assim, a variável sexo do estudante será provavelmente representada por uma dimensão ortogonal àquela definida pelos indicadores de renda e escolaridade dos pais.

Em termos mais gerais, variáveis associadas tendem a ser aproximadamente colineares enquanto variáveis dissociadas tendem a ser representadas por dimensões ortogonais no espaço. Portanto, a *quantidade de dimensões necessárias* para representar razoavelmente um conjunto de categorias dependerá dos padrões de associação entre essas categorias. Quanto mais associadas forem as variáveis categóricas, menos dimensões serão necessárias para mapeá-las.

Dito tudo isso, fica evidente que o desafio de mapear um conjunto de variáveis categóricas é, também, um problema de *redução dimensional*, completamente análogo ao que vimos na análise de componentes principais (cf. lição 10). De fato, a questão de pesquisa da análise de correspondência pode ser formulada quase exatamente da mesma maneira:

- quantas/quais dimensões são necessárias para representar, da melhor maneira possível, os padrões de associação em um conjunto de variáveis categóricas?

Por diversas razões, podemos dizer que a análise de correspondência é análoga à análise de componentes principais. Ambas são métodos *exploratórios* e *descritivos* (em oposição aos métodos confirmatórios e inferenciais). De fato, ao adotar essas análises, o pesquisador geralmente não pretende testar a plausibilidade de um modelo pré-definido. Pelo contrário, ele aborda os dados munido de poucas premissas, explorando-os livremente e criando condições para que os padrões de associação entre as variáveis sejam explicitados.

Além de serem métodos de redução dimensional, tanto a análise de correspondência quanto a análise de componentes principais podem ser enunciadas como um *problema de diagonalização* e, portanto, podem ser resolvidas por aplicação direta de um dos *teoremas da decomposição espectral* (cf. apêndice). Porém, enquanto a análise de componentes principais diagonaliza a matriz de covariâncias dos dados, a análise de correspondência pode ser calculada pela *diagonalização da matriz dos resíduos padronizados* (cf. lição 6) em uma tabela de contingência (Greenacre, 2007). Portanto, é pela matriz dos resíduos padronizados que nossa viagem começa!

A matriz dos resíduos padronizados

Por praticidade, buscando dar continuidade às discussões anteriores (cf. lição 6), vamos retornar ao questionário socioeconômico do Enem para ilustrar como fazer uma análise de correspondência. Considere, portanto, a tabulação cruzada da escolaridade do pai contra a escolaridade da mãe.

```
library(Hmisc)
load("ENEM_exemplo.dat")
ENEM = ENEM[c(1:10000), ]
```

```
library(plyr)
ENEM$EscPai = mapvalues(ENEM$EscPai,
                  from = levels(ENEM$EscPai),
                  to = c("Alfb", "Alfb", "Fund", "Fund", "EMed", "Terc",
"Terc"))
ENEM$EscMae = mapvalues(ENEM$EscMae,
                  from = levels(ENEM$EscMae),
                  to = c("Alfb", "Alfb", "Fund", "Fund", "EMed", "Terc",
"Terc"))

table(ENEM$EscPai, ENEM$EscMae)

##
##        Alfb Fund EMed Terc
##   Alfb 1483  786  498  129
##   Fund  323 1251  909  324
##   EMed  137  516 1532  688
##   Terc   23   89  390  922
```

Conforme já discutimos (cf. lição 6), a melhor maneira de analisar os padrões de associação de uma tabela de contingência não é observar suas frequências absolutas (listadas anteriormente), mas seus resíduos padronizados (i.e., a diferença padronizada entre os valores constatados e os valores esperados). Isso pode ser calculado facilmente pela função **`CrossTable`** da biblioteca **`gmodels`**.

```
##            Alfb        Fund        EMed       Terc
## Alfb  38.290154   0.7547415 -15.0107641 -19.165031
## Fund  -9.742036  18.7052412  -0.8325586 -10.600164
## EMed -18.001699  -8.8217619  18.6114282   3.914497
## Terc -15.357342 -14.8079346  -3.8603240  36.653313
```

Ainda que os resíduos padronizados em uma tabela de contingência não possam ser interpretados como correlações, o tipo de interpretação que damos aos dois é análogo: resíduos nulos apontam para falta de

associação; resíduos positivos marcam associações de mesmo sentido; resíduos negativos marcam associações de sentido oposto. A partir dessa interpretação, a tabela precedente evidencia que as categorias estão positivamente associadas ao longo da diagonal principal (i.e., a ocorrência de uma torna mais provável a ocorrência da outra) e negativamente associadas nas células afastadas da diagonal.

Conforme já antecipei, o resultado da análise de correspondência pode ser obtido diagonalizando a matriz dos resíduos padronizados em uma tabulação cruzada (Greenacre, 2007). Desde discussões anteriores (cf. lição 10), nós sabemos que a diagonalização de uma matriz pode ser realizada por aplicação direta de um dos teoremas da decomposição espectral (cf. apêndice). Nesta lição, não vamos aprofundar a discussão da álgebra matricial mais do que já fizemos. Em vez disso, vou explicar como a análise de correspondência pode ser realizada e interpretada empregando funções da biblioteca **`ca`** (instale-a, se ainda não o fez). Ao longo do caminho, alguns conceitos-chave serão introduzidos.

Análise de correspondência simples

A análise de correspondência é dita “simples” quando envolve somente duas variáveis categóricas. Para realizá-la, empregamos a função **`ca`** (da biblioteca homônima), que deve ser alimentada com a tabulação cruzada das duas variáveis que desejamos analisar. Essa tabulação é tipicamente obtida por meio da função **`table`**, da biblioteca de base. A função **`ca`**, por sua vez, computa internamente a matriz dos resíduos padronizados e sua respectiva diagonalização por aplicação direta do teorema da decomposição do valor singular (cf. apêndice). A função **`summary`** permite inspecionar o resultado da análise:

```
library(ca)
fit = ca(table(ENEM$EscPai, ENEM$EscMae))
summary(fit)

##
## Principal inertias (eigenvalues):
##
##  dim    value      %   cum%   scree plot
##  1      0.363234  69.9  69.9  *****************
##  2      0.119311  23.0  92.9  ******
##  3      0.036881   7.1 100.0  **
##         -------- -----
##  Total: 0.519426 100.0
##
##
## Rows:
##     name   mass  qlt  inr    k=1 cor ctr    k=2 cor ctr
## 1 | Alfb |  290  996  396 | -795 888 503 |  278 109 187 |
## 2 | Fund |  281  749  107 |  -83  34   5 | -377 714 334 |
## 3 | EMed |  287  763  147 |  406 621 130 | -194 142  91 |
## 4 | Terc |  142  978  349 |  959 723 361 |  570 255 388 |
##
## Columns:
##     name   mass  qlt  inr    k=1 cor ctr    k=2 cor ctr
## 1 | Alfb |  197  993  408 | -951 839 490 |  407 154 274 |
## 2 | Fund |  264  750  125 | -272 302  54 | -332 449 243 |
## 3 | EMed |  333  724  113 |  262 389  63 | -243 335 165 |
## 4 | Terc |  206  983  354 |  832 777 393 |  429 206 318 |
```

Como é possível perceber, o *output* traz, como primeira informação, os *autovalores* da decomposição espectral. Na análise de correspondência, esses autovalores são chamados *inércias principais*. A propósito, inércia é o nome dado, na análise de correspondência, à medida de dispersão dos dados. Portanto, ela cumpre um papel análogo ao cumprido pela variância na análise de componentes principais. Existe uma inércia total que representa a dispersão geral dos dados. Essa inércia total é igual à estatística

chi-quadrado dividida pela quantidade de elementos da amostra (Greenacre, 2007) e pode ser decomposta em parcelas correspondentes a cada dimensão. *Quanto maior for a inércia associada a cada dimensão do espaço, maior é a quantidade de informação representada por essa dimensão*.

Logo, as inércias principais permitem construir um *scree plot* (cf. lição 10), muito importante para orientar a redução dimensional que precisa ser arbitrada pelo analista. Como já sabemos, a redução dimensional está sempre situada em um *trade-off*. De fato, o percentual da inércia total representada pelas dimensões do mapa sempre aumenta quando a quantidade de dimensões aumenta, mas nossa capacidade interpretativa diminui. Sendo assim, ao reter uma quantidade menor de dimensões, o analista abre mão de representar uma parte da informação com o propósito de tornar a análise mais compreensível.

Além dos autovalores e seus respectivos percentuais da inércia total, o *output* publica a versão rudimentar de um *scree plot*. Como já sabemos, a redução dimensional pode ser arbitrada a partir dessas informações com base em alguns critérios (cf. lição 10). Pelo *output*, percebemos que nossa análise possui, no máximo, três dimensões (pois foram publicados somente três autovalores). Portanto, se aplicarmos o critério de Kaiser, cada dimensão retida deve explicar, no mínimo, 33% da inércia total. Esse critério recomendaria reter somente a primeira dimensão. De fato, conforme antecipamos, as variáveis de escolaridade do pai e da mãe tendem a ser aproximadamente colineares, i.e., podem ser suficientemente bem representadas por 1 variável escalar.

Por outro lado, ao olhar o *scree plot*, eu tenho a impressão de que a segunda dimensão se sobressai acima da terceira, devendo ser retida pelo critério de Catell. De qualquer maneira, como precisamos de duas dimensões para gerar, um mapa bidimensional, estou optando por reter

as duas primeiras dimensões. Juntas, elas são capazes de representar incríveis 92,9% da inércia total.

Em seguida, a função `summary` fornece uma tabela de resultados para as categorias de linha e de coluna. Todas as quantidades foram multiplicadas por 1.000. As informações publicadas no *output* são discriminadas no quadro seguinte.

Quadro 11.1: Informações disponíveis no sumário da análise de correspondência

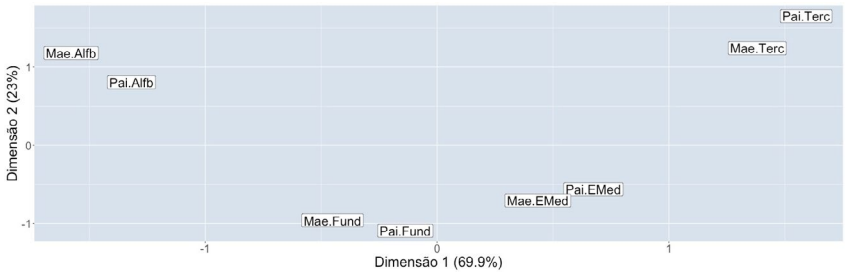
Designação	O que é?	O que significa?
`mass`	A **massa** de cada categoria	A taxa de ocupação de cada nível relativo aos demais da mesma variável.
`qlt`	A **qualidade de exibição**	A qualidade de exibição (em permilagens) de cada categoria no mapa.
`inr`	A **inércia** do ponto	A fração da inércia total (em permilagens) representada por cada categoria.
`k = i`	A **coordenada principal**	A coordenada principal da categoria com relação à i-ésima dimensão da análise multiplicada por 1 mil.
`cor`	A **contribuição relativa**	A contribuição relativa da i-ésima dimensão para a inércia do ponto.
`ctr`	A **contribuição absoluta**	A contribuição absoluta do ponto para a inércia da i-ésima dimensão (em permilagens da inércia principal)

Após inspecionar o resultado da análise com a função `summary`, o mapa que representa especialmente as categorias (gráfico 11.3) pode ser obtido por meio da função **`plot`**. A saber, vários *outputs* gráficos diferentes podem ser gerados na análise de correspondência. Das opções disponíveis, o *mapa simétrico* é o mais popular e é gerado como padrão. Para conhecer as demais opções, digite **`help(plot.ca)`**. Categorias mais associadas tendem a ficar mais próximas no mapa. Porém, é mais correto evitar inferências a partir da distância entre os pontos no mapa simétrico, pois a simetria do mapa só é possível se abrirmos mão de representar

todas as variáveis na mesma escala (Greenacre, 2007). Como alternativa, projetamos as categorias sobre as dimensões do gráfico e damos sentido a essas dimensões. Esse é o procedimento correto.

```
plot(fit)
```

Gráfico 11.3: Mapa simétrico das escolaridades do pai e da mãe (Enem 2018)



Por exemplo, quando percorremos a dimensão 1, que representa 69,9% da inércia dos dados, saímos dos casos de baixa escolaridade para os casos de alta escolaridade. Logo, essa dimensão pode ser interpretada como uma medida da classe social (ao menos no que diz respeito à escolaridade dos pais). A dimensão vertical, por outro lado, não traz informação realmente nova. Ela distingue os setores de escolaridade média e agrega os extremos. Pode ser útil para testar se há privilégios escolares específicos da classe média.

Análise de correspondência múltipla e conjunta

A análise de correspondência múltipla pode ser introduzida como uma análise de correspondência simples aplicada ao que chamamos *matriz de Burt*, uma matriz composta por matrizes de contingência empilhadas (tabela 11.1):

```
fit = mjca(ENEM[, c(3, 15)])
fit$Burt
```

Tabela 11.1: Exemplo de uma matriz de Burt

	Brancos	Pretos	Pardos	Amarelos	Indígenas	EscPub	EscMista	EscPriv
Brancos	3810	0	0	0	0	2273	483	1054
Pretos	0	1222	0	0	0	1002	134	86
Pardos	0	0	4686	0	0	3706	488	492
Amarelos	0	0	0	213	0	142	27	44
Indígenas	0	0	0	0	69	60	3	6
EscPub	2273	1002	3706	142	60	7183	0	0
EscMista	483	134	488	27	3	0	1135	0
EscPriv	1054	86	492	44	6	0	0	1682

A matriz de Burt é uma solução para representar múltiplas tabelas de contingência em uma só. Como é possível perceber, ela é uma matriz quadrada e simétrica que faz o cruzamento simultâneo de todas as categorias da análise. Como consequência, as submatrizes ao longo da diagonal principal de Burt não carregam nenhuma informação sobre a associação das variáveis. Algumas versões da análise de correspondência múltipla são diretamente orientadas a corrigir as distorções produzidas pela diagonal principal da matriz de Burt.

Enfim, a diagonalização dos resíduos padronizados da matriz de Burt pode ser feita automaticamente com a função **`mjca`**, da biblioteca **`ca`**. Ao usar essa função, o usuário precisa escolher uma entre quatro opções de análise:

1. **lambda = "Burt"** fornece a análise de correspondência múltipla baseada na matriz de Burt;
2. **lambda = "*indicator*"** fornece a análise de correspondência múltipla baseada na matriz indicadora (uma alternativa à matriz de Burt);
3. **lambda = "*adjusted*"** fornece uma versão ajustada da análise de correspondência múltipla;
4. **lambda = "JCA"** fornece a análise de correspondência conjunta.

As três primeiras opções atribuem os mesmos valores às coordenadas padronizadas das categorias. Ou seja, os mapas simétricos serão idênticos. As opções "Burt" e "indicator" foram as primeiras a serem desenvolvidas ao longo da história e cumprem um papel importante para introduzirmos a análise de correspondência. No entanto, elas tendem a superestimar a inércia total (e, portanto, subestimar a inércia explicada pelas dimensões da análise). No caso da matriz de Burt, isso fica claro pela presença de várias células que deveriam ser ignoradas em torno da diagonal principal, pois não trazem nenhuma informação sobre a associação das categorias da tabela. Em consequência, os métodos "Burt" e "*indicator*" tendem a subestimar a qualidade dos mapas gerados. Por essas razões, as versões "*adjusted*" e "JCA" são preferíveis às demais, pois corrigem o percentual da inércia explicada.

A solução ajustada ("*adjusted*") é definida como padrão da função *mjca* e realiza a correção da inércia explicada reescalonando a solução da análise de correspondência múltipla para que ela se ajuste melhor aos elementos fora da diagonal principal da matriz de Burt. Ela faz isso preservando as coordenadas das categorias no mapa. Já o algoritmo da análise de correspondência conjunta ("JCA") é iterativo e tem como ponto de partida a análise de correspondência múltipla da matriz de Burt. Feito isso, as submatrizes ao longo da diagonal principal de Burt são substituídas por suas melhores estimativas segundo o modelo. Em seguida, a análise de correspondência é calculada novamente. Esse recálculo não corrige somente a inércia, mas as próprias coordenadas padronizadas, que poderão ser um pouco diferentes. Portanto, pode haver alguma diferença entre os mapas gerados pela análise de correspondência conjunta ("JCA") e os gerados pelos demais métodos de extração ("*Burt*", "*indicator*" e "*adjusted*"). O algoritmo repete o processo até que haja convergência.

A seguir, fazemos uma análise de correspondência múltipla ajustada da renda, escolaridade e ocupação dos pais:

```
fit = mjca(ENEM[, c(9:12, 14)])
summary(fit)

##
## Principal inertias (eigenvalues):
##
##  dim    value      %   cum%   scree plot
##  1      0.356550  69.3  69.3  ******************
##  2      0.103665  20.1  89.4  *****
##  3      0.024109   4.7  94.1  *
##  4      0.002858   0.6  94.7
##  5      0.001393   0.3  95.0
##  6      6.1e-050   0.0  95.0
##  7      9e-06000   0.0  95.0
##  8      2e-06000   0.0  95.0
##  9      1e-06000   0.0  95.0
##  10     00000000   0.0  95.0
##  11     00000000   0.0  95.0
##  12     00000000   0.0  95.0
##  13     00000000   0.0  95.0
##  14     00000000   0.0  95.0
##  15     00000000   0.0  95.0
##         -------- -----
##  Total: 0.514539
##
##
## Columns:
##                              name   mass  qlt  inr     k=1 cor ctr     k=2 cor
## 1  |                EscPai:Alfb |   58  957   31 |   579 700  54 |  -350 257
## 2  |                EscPai:Fund |   56  695   21 |   233 440   9 |   177 255
## 3  |                EscPai:EMed |   57  879   23 |  -188 203   6 |   342 676
## 4  |                EscPai:Terc |   28  993   45 | -1258 930 126 |  -326  63
## 5  |                EscMae:Alfb |   39  946   33 |   651 640  47 |  -450 306
## 6  |                EscMae:Fund |   53  770   23 |   362 728  19 |    86  41
## 7  |                EscMae:EMed |   67  913   20 |   -43  19   0 |   295 894
## 8  |                EscMae:Terc |   41  980   42 | -1014 957 119 |  -157  23
## 9  |           OcupPai:Campones |   46  869   36 |   614 500  48 |  -528 370
## 10 |     OcupPai:TrabSemFormacao |   47  627   23 |   255 381   9 |   205 245
```

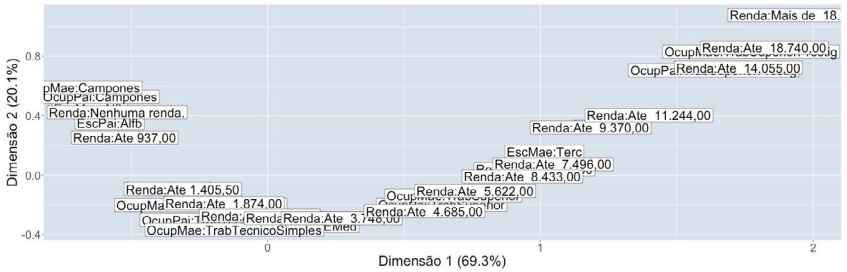
```
## 11 |  OcupPai:TrabTecnicoSimples |   53  746   22 |   152 150   3 |   303 596
## 12 |        OcupPai:TrabSuperior |   40  829   31 |  -637 756  46 |   199  74
## 13 | OcupPai:TrabSuperiorPrestig |   14  896   47 | -1635 755 102 |  -705 141
## 14 |            OcupMae:Campones |   38  869   38 |   688 501  50 |  -590 369
## 15 |     OcupMae:TrabSemFormacao |   90  714   20 |   245 428  15 |   200 286
## 16 |  OcupMae:TrabTecnicoSimples |   15  589   25 |   122  59   1 |   367 530
## 17 |        OcupMae:TrabSuperior |   49  824   33 |  -681 792  63 |   137  32
## 18 | OcupMae:TrabSuperiorPrestig |    9  900   43 | -1772 738  82 |  -832 162
## 19 |        Renda:Nenhuma renda. |    7 1028   26 |   552 648   6 |  -422 380
## 20 |            Renda:Ate 937,00 |   44 1021   27 |   524 828  34 |  -253 194
## 21 |          Renda:Ate 1.405,50 |   48  912   22 |   311 835  13 |    94  77
## 22 |         Renda:Ate  1.874,00 |   19  908   23 |   160 379   1 |   189 529
## 23 |         Renda:Ate  2.342,50 |   18  975   23 |    28  10   0 |   273 965
## 24 |         Renda:Ate  2.811,00 |   10  887   24 |  -136 160   0 |   289 727
## 25 |         Renda:Ate  3.748,00 |   14  926   24 |  -271 433   3 |   290 493
## 26 |         Renda:Ate  4.685,00 |    9  867   26 |  -575 736   9 |   242 131
## 27 |         Renda:Ate  5.622,00 |    8  900   26 |  -761 883  13 |   106  17
## 28 |         Renda:Ate  6.559,00 |    4  906   27 |  -978 904  12 |   -40   1
## 29 |         Renda:Ate  7.496,00 |    3  953   26 | -1046 948   8 |   -73   5
## 30 |         Renda:Ate  8.433,00 |    2  894   26 |  -934 894   6 |     9   0
## 31 |         Renda:Ate  9.370,00 |    2  975   26 | -1184 908   9 |  -321  67
## 32 |        Renda:Ate  11.244,00 |    4  973   28 | -1398 898  20 |  -403  75
## 33 |        Renda:Ate  14.055,00 |    2  939   28 | -1726 798  17 |  -724 140
## 34 |        Renda:Ate  18.740,00 |    2  887   29 | -1820 726  20 |  -859 161
## 35 |    Renda:Mais de  18.740,00 |    3  816   33 | -1969 628  30 | -1077 188
```

Aplicando os critérios de Kaiser e Catell aos autovalores e seu respectivo *scree plot*, percebemos estar diante de uma estrutura claramente bidimensional, com 89,4% da inércia total representada nas duas primeiras dimensões. Observe o poder dessa redução dimensional! Estamos percebendo que duas variáveis escalares são capazes de representar quase integralmente as informações disponíveis em 35 categorias!

O mapa simétrico pode ser produzido facilmente por meio da função **`plot`**:

```
plot(fit)
```

Gráfico 11.4: Mapa simétrico da renda, escolaridade e ocupação do pai e da mãe (Enem 2018)



Como é possível perceber, temos aí a mesma ferradura que vimos no gráfico anterior. A saber, ferraduras nos mapas simétricos podem sugerir tão simplesmente uma *associação linear* entre as variáveis envolvidas. Isso ocorre porque, antes de ser realizada a redução dimensional, as categorias estão posicionadas em uma região do espaço n-dimensional chamada *simplex*. Os limites (mais ou menos triangulares) dessa região fazem com que associações lineares entre as categorias de duas variáveis sejam representadas como curvas (e não tanto como retas). Por isso, a ferradura é um padrão muito comum de se obter ao mapear variáveis que são aproximadamente colineares.

No mapa simétrico que obtivemos agora (gráfico 11.3), as famílias com pouca escolaridade, baixa renda e ocupações mais humildes estão à esquerda. As famílias mais escolarizadas, com alta renda e ocupações prestigiadas estão à direita. Assim, a dimensão 1 (que representa 69,3% da inércia total dos dados) pode ser interpretada como um indicador de classe. A dimensão 2 não traz informação realmente nova, posicionando famílias de classe média (abaixo) e famílias de classe popular e dominante (acima).

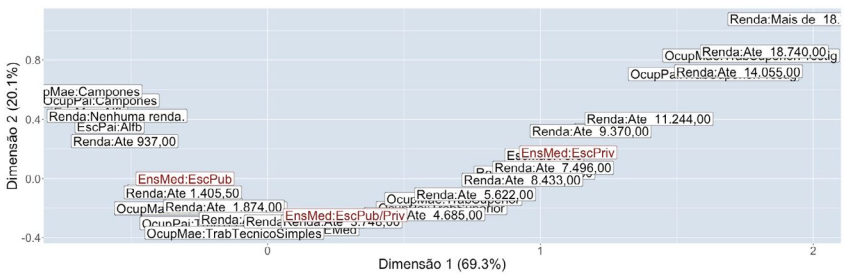
Variáveis suplementares

A análise de correspondência também permite acrescentar *variáveis suplementares*. Elas não participam da determinação da estrutura do espaço, mas são posicionadas junto com as demais. Para trabalhar com variáveis suplementares, basta inclui-las no *dataframe* e especificar, com a diretiva **`supcol`**, as colunas que devem ser consideradas suplementares. O que o algoritmo faz é zerar a massa dessas variáveis. Isso faz com que elas não contribuam para a inércia total dos dados.

No mapa, as variáveis suplementares são geralmente representadas com uma cor diferente. Elas podem ser representadas com um triângulo vazio, enquanto as variáveis ativas (i.e., variáveis empregadas para especificar as dimensões do mapa) são representadas por um triângulo cheio. Observe que, conforme esperado, existe uma aderência das frações de classe mais abastadas à escola privada e das frações de classe popular à escola pública.

```
colnames(ENEM)
fit = mjca(ENEM[, c(9:12, 14, 15)], supcol = c(6))
plot(fit)
```

Gráfico 11.5: Mapa simétrico com uma variável suplementar



A qualidade do mapa

Vimos que a qualidade dos mapas gerados pela análise de correspondência pode ser avaliada pelo percentual da inércia original representada em cada um dos eixos. Vimos, também, que alguns métodos da análise múltipla subestimam esse percentual da inércia explicada e que, por essa razão, damos preferência às análises ajustada ou conjunta. Porém, mesmo empregando o melhor método disponível, é possível que, em alguns casos, os percentuais de inércia sejam baixos. O que isso significa?

Enquanto analista, é importante ter em mente que a possibilidade de representar os dados com poucas variáveis escalares depende fundamentalmente da estrutura própria dos dados e menos de um tratamento feito *a posteriori*. No Brasil, assim como em outros países muito desiguais, a distância entre a renda mais baixa e a mais alta é muito grande. Essa amplitude contribui para que a renda tenha grande aderência aos graus de escolaridade e tipos de ocupação dos cidadãos. Juntando tudo isso à endogamia das classes, percebemos que todas essas cinco variáveis categóricas empregadas aqui podem ser interpretadas como indicadores aproximadamente colineares, descrevendo juntos a posição da família nas relações de classe.

Voltaremos a essa questão mais claramente ao discutir a construção de testes educacionais e a análise fatorial. Porém, é importante ter em mente que a qualidade do mapa depende da possibilidade de representar toda a informação disponível em poucas dimensões, i.e., depende que as informações estejam fortemente associadas a um ou dois construtos básicos (no caso, a "classe"). Se você acrescentar variáveis como "cor", "sexo" e "idade" às análises que fizemos, perceberá que a inércia explicada pelos mapas tende a diminuir bastante. Isso não quer dizer que sua análise está propriamente ruim, mas que a multiplicidade dos dados disponíveis

faz com que muita informação seja perdida ao tentar projetá-los em um mapa bidimensional.

Por outro lado, se você tem algum controle sobre as variáveis categóricas que podem ser representadas (por exemplo, se está construindo um questionário socioeconômico que será submetido a uma análise de correspondência), é estratégico incluir indicadores diferentes, mas mutuamente associados (mesmo que isso torne seu questionário um pouco repetitivo). Isso contribuirá para que o percentual da inércia explicada seja relativamente mais elevado.

Comparando PCA e MCA

Ao longo desta lição, foram enfatizadas as continuidades e semelhanças entre dois métodos da análise exploratória multivariada que nos permitem identificar padrões de associação entre as variáveis de um *dataframe* (Husson; Lê; Pagès, 2010): a análise de componentes principais (PCA, lição 10) e a análise de correspondência (simples, múltipla, conjunta). As continuidades foram enfatizadas por razões didáticas: ideias conectadas costumam ser mais bem assimiladas. Porém, apesar das semelhanças entre esses dois métodos, há diferenças que precisam ser notadas.

A análise de componentes principais foi inicialmente definida para identificar padrões de associação linear entre variáveis escalares. É verdade que as correlações tetrapolicóricas nos permitem generalizar a análise de componentes principais para variáveis categóricas — *desde que essas variáveis possam ser interpretadas como manifestações categóricas de traços latentes linearmente associados*. Isso não ocorrerá a todas as variáveis categóricas. Com variáveis ordinais linearmente associadas (renda familiar, grau de escolaridade), a PCA terá mais sucesso.

Porém, ela tende a falhar com variáveis nominais (nacionalidade, local de residência, ocupação) com padrões de associação mais variados.

A análise de correspondência não tem uma matriz de correlações como ponto de partida, mas uma tabela de contingência. Por isso, ela é capaz de identificar quaisquer padrões de associação (lineares ou não) entre os níveis das variáveis categóricas envolvidas. Ela é, portanto, menos restrita que a PCA no tratamento de variáveis categóricas.

Em suma, o contexto de emprego desses dois métodos está representado no quadro 12.1.

Quadro 12.1: Comparando análise de componentes principais e de correspondência múltipla com respeito às variáveis ativas (i.e. não suplementares) que podem ser incluídas [1]

Variáveis	***Componentes principais***	***Correspondência** múltipla*
Escalares	Sim	-
Categóricas Binárias	Sim	Sim
Categóricas Ordinais	Sim	Sim
Categóricas Nominais	-	Sim

A análise de correspondência na sociologia da educação de Bourdieu

Não poderíamos encerrar esta lição sem dedicar algumas palavras às pesquisas empíricas que empregaram a análise de correspondência

[1] Se, por algum motivo, você precisar analisar um *dataframe* com variáveis escalares e nominais, eu recomendo a análise multivariada com *escalonamento ótimo* (ing. *optimal scaling*), implementada nas funções da biblioteca Gifi (https://cran.r-project.org/web/packages/Gifi/Gifi.pdf). Essas funções permitem realizar, entre outras coisas, uma análise de componentes principais com variáveis escalares e categóricas de todos os tipos. Contudo, como há pouco suporte acessível a iniciantes para as funções dessa biblioteca, eu estou omitindo a análise não linear de componentes principais deste livro.

múltipla. Na educação em ciências, há diversos exemplos (Bächtold; Cross; Munier, 2019; Lima Junior; Fraga Junior *Et Al.*, 2020; Lima Junior; Fraga Junior, 2021; Massi; Carvalho; Giordan, 2020; Monteiro Nascimento *et al.*, 2019). Porém, o exemplo mais influente vem do campo da sociologia geral e da educação. Talvez pela maneira bastante sofisticada de se comunicar ou pela consistência teórica deste autor, muitos leitores se esquecem que todo o argumento de Bourdieu está assentado em evidência empírica (Nogueira; Nogueira, 2009) e o suporte empírico das suas reflexões deve muito à análise de correspondência.

Diversas afirmações bourdieusianas sobre o chamado "espaço social" (Bourdieu, 1996) ganham sentido concreto somente quando conhecemos a análise de correspondência. Por exemplo, a afirmação de que o espaço social é um espaço *multidimensional* de *posições relativas* determinado pela *distribuição desigual* das diversas formas de capital (Bourdieu, 2007a) pode ficar sem o referente adequado até o momento em que nos deparamos com o método de análise discutido nesta lição. Com efeito, nas análises de correspondência e de componentes principais, um "espaço multidimensional" é a maneira como se representa um conjunto de variáveis que não seja completamente colinear. Ou seja, o conhecimento de uma dessas variáveis (e.g., renda) não permite predizer o valor de todas as outras. De fato, se o espaço social é mesmo multidimensional, nem todas as distinções relevantes podem ser reduzidas à desigualdade econômica.

Aqui, o leitor atento deve fazer a seguinte pergunta: o espaço social, em nosso país, é realmente multidimensional? Até que ponto há suporte empírico para afirmar que, no Brasil, o capital cultural é irredutível ao capital econômico? Segundo as análises que fizemos (Lima Junior; Andrade *et al*, 2020; Lima Junior; Fraga Junior, 2021), há uma forte aderência entre graus de escolaridade (indicador de capital cultural) e renda

familiar (indicador de capital econômico). Há, portanto, uma relativa colinearidade entre indicadores que, em teoria, deveriam ser mais ou menos ortogonais. Essa colinearidade, por sua vez, não reflete somente a grande desigualdade social em nosso país, mas a maneira como as variáveis são tipicamente codificadas nos questionários socioeconômicos. Por exemplo, a classificação ocupacional adotada depende, no "fingir dos ovos", do tempo de escolaridade. Ela não permite distinguir entre profissões mais ou menos relacionadas à cultura em cada grau de escolaridade. A codificação empregada não favorece, portanto, a distinção entre uma esquerda cultural e uma direita empresarial explorada nas análises de Bourdieu (1984) e tão significativa no debate político contemporâneo. Sem dúvida, há muita ciência na construção de um questionário — seja ele socioeconômico ou não.

A lista das reflexões relacionadas à análise de correspondência é enorme e não será (nem poderia ser) esgotada aqui. Contudo, não posso deixar de comentar o quanto a análise de correspondência parece ter contribuído para pensar causas sociais segundo a complexidade dos sistemas dinâmicos autorregulados (cf. Dyke, 1999). Como vimos ao longo de todo esse livro, os métodos associativos (e.g., análise de correspondência e de componentes principais) são diferentes dos métodos preditivos (e.g., regressões lineares e análises de variância) pela maneira como separam, ou não, variáveis dependentes e independentes. Essa separação faz com métodos preditivos estabeleçam uma distinção muito clara entre causa e efeito, origem e resultado, processo e produto. Assim como a força resultante produz aceleração, os modelos preditivos favorecem pensar que a pobreza produz fracasso, que o privilégio produz sucesso, que o interesse produz aprendizagem, que reconhecimento produz confiança. O esquema explicativo básico empregado por Bourdieu é completamente diferente e mais compatível com métodos de análise que não fazem esse tipo de separação.

Práticas culturais e escolares não são somente resultado de uma desigualdade anteriormente estabelecida, mas participam da construção da própria desigualdade que as produz. Podemos dizer, por exemplo, que alguns cidadãos vão muito ao museu ou ingressam na universidade federal porque pertencem a um grupo privilegiado. Porém, esses grupos sociais não podem ser identificados, senão pelas suas práticas culturais distintas (e.g., frequência a museus, bibliotecas, teatros, universidades...). Essas práticas, geralmente descritas por variáveis categóricas, aparecem necessariamente duas vezes na análise: no princípio, elas geram os mapas, representações do espaço social; no final, são explicadas pelas posições no espaço que não poderiam ser definidas sem elas.

A rigor, o espaço social, tal como modelado na análise de correspondência, não é anterior às variáveis. As variáveis são, ao mesmo tempo, *determinantes* para as posições no espaço e *determinadas* por essas posições. De fato, o tratamento igualitário das variáveis na análise de correspondência não favorece eleger causas externas para o sucesso e o fracasso escolar. As questões de classe não são, portanto, algo que vem de fora da escola, invadindo-a. Pelo contrário, diferenças de classe são coconstruídas pelas experiências e pelo sucesso ou fracasso na escola. A análise de correspondência favorece, portanto, um esquema explicativo radicalmente diferente da oposição entre infraestrutura e superestrutura (Althusser, 2013), segundo a qual distinções culturais são, em última instância, determinadas por desigualdades econômicas. Nas apropriações mais rasteiras, é dito que o aluno fracassa porque é pobre, porque veio de uma família pouco escolarizada, ou porque o professor não soube reconhecer a legitimidade da cultura da periferia... Essas causalidades muito bem delimitadas são fáceis de reconhecer justamente porque conseguem imputar resultados escolares a atores (estudantes, professores, familiares) ou processos (mérito, motivação, dedicação) bem localizados.

Esse tipo de conclusão causal simples não faz sentido aqui. Aquilo que vale como cultura legítima jamais está dado de maneira irrevogável, tampouco pode ser livremente redefinido por qualquer indivíduo. Isso é assim justamente porque a legitimidade da cultura legítima é reproduzida e transformada nos processos diários de luta por consagração cultural. Tudo isso faz muito sentido dentro da análise de correspondência, pois, nela, *a causalidade só pode existir distribuída na relação recíproca de associação das variáveis* e não como relações preditivas bem localizadas. Enfim, a análise de correspondência favorece pensar a causalidade de maneira mais sistêmica e não local, complexa e autorregulada.

Revisando a lição

Nessa lição, você aprendeu que:

5. a análise de correspondência é análoga à análise de componentes principais e que sua solução corresponde à diagonalização da matriz de resíduos padronizados;
6. a análise de correspondência simples pode ser feita com a função **`ca`** e a análise múltipla, com a função **`mjca`** da biblioteca **`ca`**;
7. os sumários podem ser gerados com a função **`summary`** e permitem decidir quantas dimensões serão retidas na análise por meio de um *scree plot* rudimentar;
8. os mapas simétricos podem ser gerados com a função **`plot`** e são interpretados projetando as categorias sobre os eixos e dando sentido aos eixos;
9. a qualidade do mapa é avaliada pelo percentual de inércia explicada por cada eixo e depende fundamentalmente da estrutura das associações dos dados.

LIÇÃO 12.

ANÁLISE DE *CLUSTER* (OU AGRUPAMENTO)

As associações entre variáveis estão intimamente relacionadas às similaridades interindividuais. Assim como variáveis podem estar mais ou menos *dissociadas*, indivíduos podem ser mais ou menos *dissimilares*. Portanto, é razoável esperar que a estrutura das associações entre variáveis e a estrutura das similaridades interindividuais sejam, em última instância, a mesma coisa. No entanto, uma análise centrada em agrupar os indivíduos em função de sua similaridade pode produzir *insights* muito diferentes de uma análise limitada a investigar a associação entre variáveis.

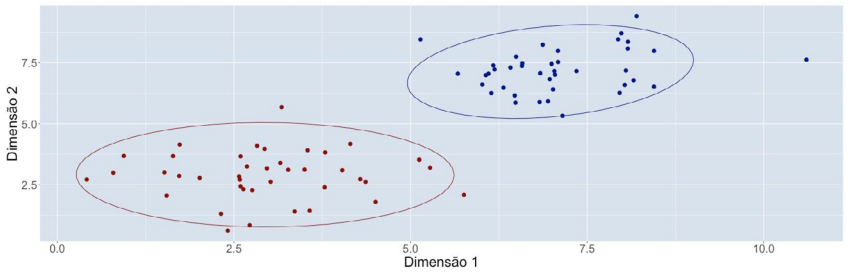
Dando continuidade às técnicas de análise exploratória multivariada, a presente lição dedica-se prioritariamente ao estudo das similaridades interindividuais. De fato, a representação espacial dos indivíduos como uma nuvem de pontos leva quase naturalmente à necessidade de discutirmos critérios de agrupamento desses indivíduos em vista de suas semelhanças. Tais *agrupamentos de instâncias individuais semelhantes* serão denominados *clusters*.

Em algumas investigações, a delimitação dos *clusters* chega a ser mais significativa que a análise das variáveis. Considere, por exemplo,

que tenhamos diversas informações sobre um conjunto de escolas da nossa região (tais como desempenho médio dos estudantes, composição do corpo docente e discente, categoria administrativa, localização). Em vez de explicitar relações entre as variáveis disponíveis, talvez seja mais significativo publicar um mapa que represente a dissimilaridade dessas escolas. Supondo que sejam instituições conhecidas, o leitor talvez tenha ouvido falar delas, talvez tenha até estudado ou trabalhado em algumas dessas escolas. Caso a pesquisa ocorra no interior de uma sala de aula específica, talvez seja relevante agrupar os estudantes em função de suas similaridades antes de realizar uma análise individualizada desses estudantes. Portanto, o mapa das instâncias individuais pode favorecer *insights* que dificilmente seriam produzidos ao considerarmos somente os padrões de associação entre variáveis. De fato, em algumas situações, o leitor atribuirá mais sentido às instâncias individuais da análise que às variáveis empregadas para descrevê-las.

No diagrama de dispersão a seguir (gráfico 12.1), os pontos representam instâncias individuais (que podem ser estudantes, professores, escolas, cidades…). Aqui, elas são posicionadas com base em informações relevantes (que podem ser desempenho escolar, interesse por ciências, renda familiar, capital cultural, custo-aluno, IDH…). Todas as informações relevantes foram reduzidas a duas dimensões (por meio de uma análise de componentes principais ou de correspondência múltipla). No caso hipotético representado no gráfico 12.1, podemos distinguir claramente duas nuvens de indivíduos, ou seja, podemos demarcar, de maneira inequívoca, dois *clusters*: agrupamentos individuais muito semelhantes em si, mas dissemelhantes entre si. Contudo, na maioria das situações de trabalho, as fronteiras entre os *clusters* são mais ambíguas e sua demarcação dependerá muito dos critérios escolhidos pelo analista.

Gráfico 12.1: Representação hipotética de dois clusters no espaço (ou seja, dois agrupamentos de instâncias individuais)



A saber, o desafio principal da *análise de clusters* é delimitar fronteiras que permitam agrupar indivíduos semelhantes e separar os diferentes. Essa delimitação é atingida elegendo uma *medida de dissimilaridade*, capaz posicionar quaisquer dois indivíduos entre si, e um *algoritmo de agrupamento*, capaz de classificar os indivíduos em vista de sua similaridade. Como há inúmeras formas de medir a dissimilaridade e vários algoritmos de agrupamento, o analista tem uma variedade enorme de caminhos possíveis. Na maioria das situações concretas de pesquisa, escolhas igualmente legítimas produzem resultados substancialmente diferentes. Portanto, assim como outros métodos de análise multivariada, a análise de *clusters* é um terreno movediço em que as decisões do analista podem influenciar muito o resultado. Em geral, é possível testar vários métodos e escolher o que parece ser mais convincente (em vista da intuição, da intenção e do conhecimento teórico do analista).

Exemplos da literatura

Na pesquisa educacional e, mais especificamente, na pesquisa em educação em ciências, a análise de *clusters* é muito empregada. Por exemplo, Pegar e Lie (2011) propõem uma análise dos indicadores de interesse

científico dos participantes do Pisa. Essa análise permitiu agrupar os países não tanto em *rankings* ou escalas, mas considerando que os estudantes de cada região podem apresentar diferentes padrões de resposta ao serem questionados sobre seus interesses. O resultado mais importante é a forte preferência relativa pelas ciências da vida e da saúde entre países não europeus, em contraste com a preferência por sistemas físicos e tecnológicos em países da Europa ocidental.

Schmidt, Rosenberg e Beymer (2018), por sua vez, empregaram a análise de *clusters* para tipificar diferentes perfis de engajamento dos estudantes em atividades escolares com respeito a variáveis comportamentais, cognitivas e afetivas. De fato, o engajamento em sala de aula não precisa ser reduzido a uma escala linear. Ou seja, em vez de distinguir os estudantes em mais ou menos engajados, é possível distinguir *de que maneira* esse engajamento ocorre. Segundo os autores (Schmidt; Rosenberg; Beymer, 2018), o engajamento depende muito do tipo de atividade de ensino e as atividades de laboratório proporcionaram experiências de engajamento especialmente polarizadas. Nesse tipo de atividade, os estudantes experimentaram um engajamento total ou universalmente baixo. Em alguns casos, houve um padrão de engajamento descrito como exclusivamente prazeroso (i.e., afetivo, mas não cognitivo nem comportamental).

Outro tema muito importante para a pesquisa em educação em ciências é a delimitação das atividades de ensino que podem ser consideradas efetivamente investigativas. Com esse desafio em vista, Yeh, Jen e Hsu (2012) empregaram a análise de *clusters* para automatizar a revisão de 171 resumos de artigos da área que, publicados entre 1986 e 2010 e indexados na *Web of Science*, faziam menção à investigação científica. De fato, os autores empregaram a análise de *clusters* como uma ferramenta de mineração de dados (ing. *data mining*), permitindo identificar cinco temas

principais (natureza da ciência, construção do conhecimento, habilidade investigativa, investigação explicativa e desenvolvimento profissional) com diferentes graus de coesão. Ao navegar pelas redes dos significados compartilhados por pesquisadores experientes, a revisão de Yeh, Jen e Hsu (2012) pode auxiliar pesquisadores iniciantes a mapear aquilo que a comunidade tem levado em consideração ao delimitar os objetivos educacionais das atividades investigativas. Evidentemente, esse tipo de abordagem exploratória (baseada em mineração de dados) pode operacionalizar grandes revisões da literatura em qualquer área de conhecimento.

Inspirados pelo conceito bourdieusiano de *habitus* (Bourdieu, 1984), Engström e Carlhed (2014) realizaram uma análise de *clusters* com professores de física. Essa análise, fundamentada em descritores de origem social, práticas de ensino e estilos de vida, permitiu identificar três grupos de profissionais: *i)* o gestor do tradicional, *ii)* o inovador tecnológico e *iii)* o desafiador da cidadania. Segundo as autoras, esses perfis correspondem não somente às origens sociais dos professores (famílias com mais capital cultural ou econômico), mas aos seus estilos de vida de uma maneira mais ampla (a maneira como se vestem, os jornais que leem). Os resultados ajudam a pensar o desenvolvimento docente como uma questão de classe.

A diversidade de usos desse método de análise segue. Hoffman *et al*. (2019) empregaram a análise de *clusters* para avaliar as experiências de prazer e sofrimento entre docentes universitários brasileiros e lusitanos. Em outra pesquisa, realizada por da Silva *et al*. (2010), a análise de *clusters* foi empregada para agrupar estudantes de odontologia com relação ao seu desempenho e avaliar os preditores do pertencimento a cada grupo. Em outra pesquisa, a análise de *clusters* foi empregada para agrupar discentes, docentes e técnicos de uma instituição federal de ensino superior com respeito à maneira como eles avaliam essa

instituição (Reis; Silveira; Ferreira, 2010). Resultados apontam que a percepção dos técnicos tende a ser mais negativa que a percepção de professores e estudantes. estudantes. A análise de *clusters* também foi empregada por Garcia-Silva, Lima Junior e Caruso (2022) para agrupar as regiões administrativas do Distrito Federal em função dos seus indicadores de violência escolar, demonstrando que a violência escolar não decorre automaticamente da violência urbana.

Como é possível perceber, a análise de *clusters* tem diversos usos, mas é particularmente útil em pesquisas de caráter descritivo e exploratório, o que permite agrupar as instâncias individuais em vista de sua similaridade. Nesta lição, vou mostrar como realizar e interpretar as duas versões mais populares desse método.

- **Agrupamento hierárquico**, mais adequado à análise de *dataframes* com poucas instâncias individuais; e
- **Método de partição k-médias** (ing., *k-means*), mais adequado ao tratamento de grandes bases de dados.

Realizando uma análise de componentes principais

Para aprendermos como a análise de *clusters* funciona em um cenário concreto da pesquisa educaconal, considere os dados do Enem:

```
library(Hmisc)
load("ENEM_exemplo.dat")
ENEM = ENEM[c(1:100),]
```

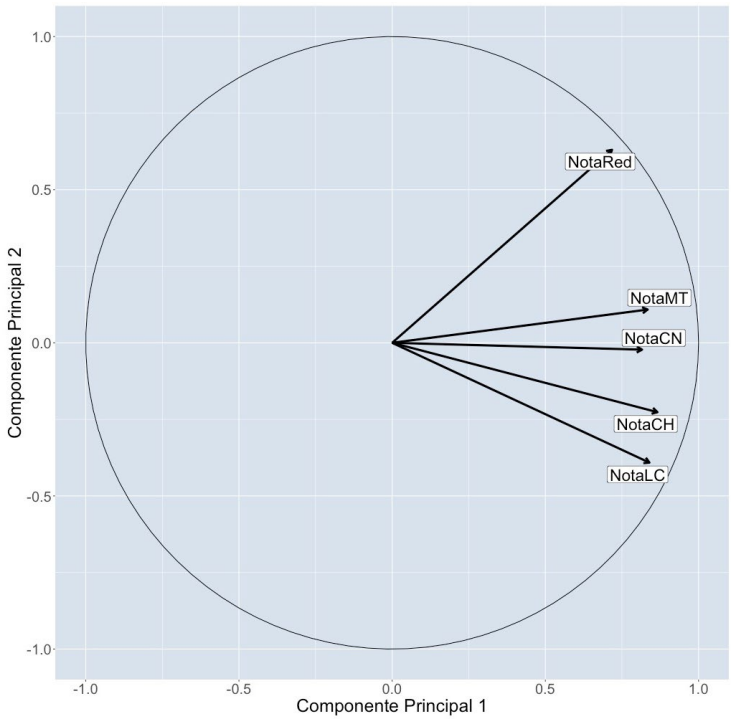
A análise de componentes principais de variáveis escalares pode ser feita com facilidade empregando a função **PCA** da biblioteca **FactoMineR**:

```
library(FactoMineR)
library(factoextra)

fit = PCA(ENEM[,c(4:8)])
summary(fit)
```

Observando o percentual da variância explicada pelas componentes principais, o leitor deve concordar que a estrutura dos dados pode ser descrita por um mapa bidimensional. Afinal, 78,88% da variância está representada nas duas primeiras componentes principais e a terceira componente agrega pouco poder explicativo (segundo os critérios de Kaiser e Catell). O gráfico 12.2 representa a associação das variáveis e nos ajudará a dar sentido às componentes principais.

Por se tratar de uma análise de componentes principais, a associação das variáveis pode ser inferida pelo ângulo entre os vetores no diagrama precedente (cf. lição 10). Ângulos agudos, retos e obtusos indicam, respectivamente, variáveis com associação positiva, nula ou negativa, pois a correlação é aproximadamente igual ao cosseno dos ângulos entre os vetores.

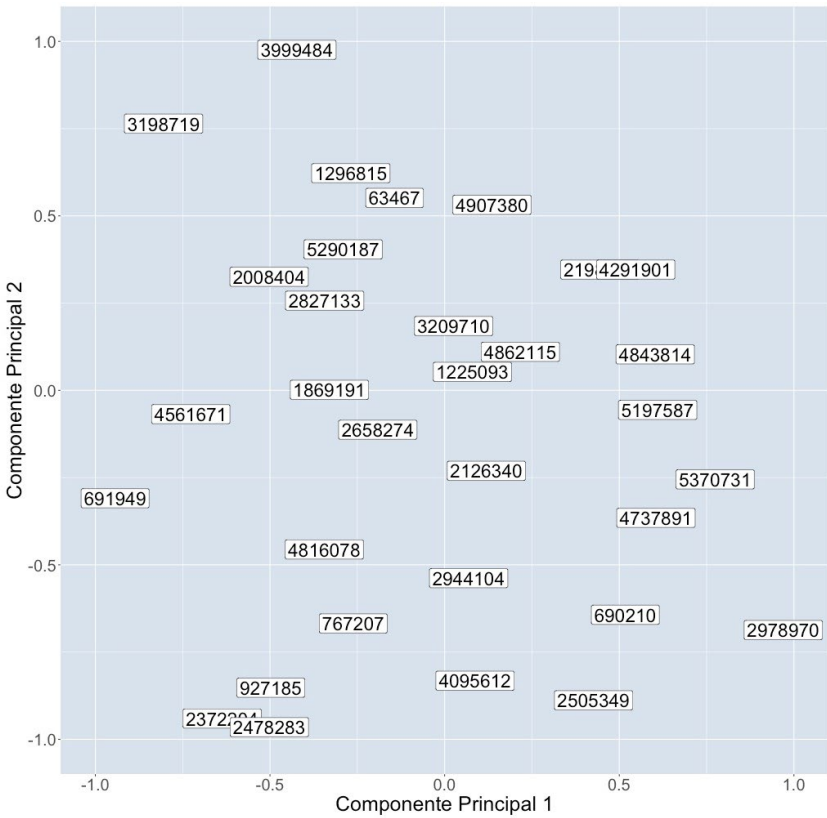
Gráfico 12.2: Mapa das variáveis na análise de componentes principais



Com base no gráfico 12.2 e no *output* da função **`summary`**, podemos dar sentido às componentes principais obtidas. Considere, por exemplo, a primeira componente principal (no eixo das abcissas). Sozinha, ela representa a maior parte da variância dos dados (66,60%). Portanto, ela pode ser interpretada como uma medida global de competência escolar dos estudantes. Quanto maior for a posição de cada estudante com respeito à primeira dimensão, melhor foi seu desempenho global no Enem. Considere, agora, a segunda componente principal. Ela representa uma porção relativamente menor da variância dos dados (12,27%) e distingue a prova de redação das questões de múltipla escolha que compõem o

exame. Como é possível perceber, indivíduos que estiverem mais acima no mapa, apresentaram melhor desempenho na prova de redação e pior desempenho nas demais avaliações. Já os indivíduos posicionados abaixo apresentaram um desempenho relativamente melhor nas questões de múltipla escolha — sobretudo nas provas de linguagens e códigos (NotaLC) e ciências humanas e sociais (NotaCH). O posicionamento dos indivíduos pode ser percebido no gráfico 12.3:

Gráfico 12.3: Mapa dos indivíduos na análise de componentes principais



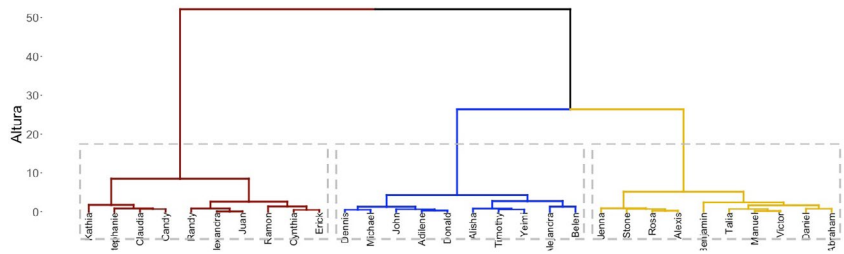
Interpretando um dendograma

Ao observar a nuvem dos casos individuais representados no gráfico 12.3, percebemos que não há agrupamentos inequívocos (como vimos no gráfico 12.1). Essa é a situação usual ao trabalharmos com dados reais e, por isso, a análise de *clusters* pode produzir resultados muito diferentes em vista das escolhas feitas pelo analista. O mais importante é que essas escolhas conduzam a resultados que possam ser interpretados e considerados relevantes para os propósitos da pesquisa.

Em geral, a ferramenta preferida para análise de *clusters* com poucas instâncias individuais é o chamado *agrupamento hierárquico* (ing., *hierarchical clustering*). Ele agrupa progressivamente as instâncias individuais até que todas estejam contidas no mesmo *cluster*. Tais agrupamentos progressivos são representados em uma estrutura tipo-garfo chamada *dendograma* (diagrama 12.1).

Na base do dendograma, as instâncias individuais são nomeadas. Como primeiro passo, o algoritmo de agrupamento hierárquico atribui cada indivíduo a um *cluster* unitário. Em seguida, ele avalia quais são os dois *clusters* mais semelhantes, e os agrupa. Essa etapa de avaliação e agrupamento é repetida até que todos os indivíduos estejam incluídos em um só *cluster*. O dendograma é justamente a representação desse agrupamento progressivo (cf. diagrama 12.1). Na parte de baixo, todos os indivíduos estão em *clusters* separados. Porém, na medida em que subimos, esses *clusters* vão sendo agregados dois a dois a começar pelos mais semelhantes até que todos se encontram em um grupo só.

Diagrama 12.1: Exemplo de dendograma com dados fictícios

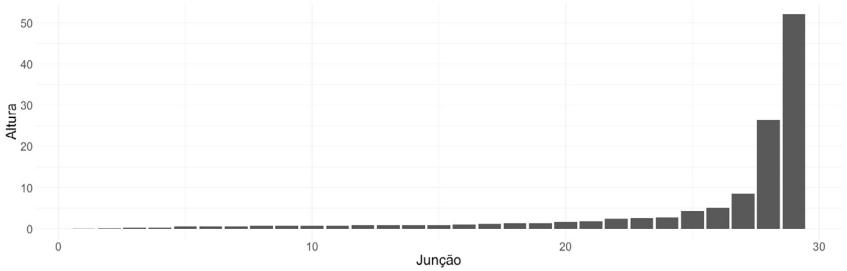


Diferente do que ocorre no mapa de indivíduos da análise de componentes principais, a proximidade entre as instâncias individuais no dendograma nem sempre representa similaridade. Devemos observar a altura em que os indivíduos são unidos. Encontros que ocorrem junto à base indicam que os indivíduos são muito semelhantes. Encontros que ocorrem no alto indicam maior dissimilaridade. Obviamente, o objetivo da análise de *clusters* não é nem seu ponto de partida (com todos os indivíduos separados) nem seu ponto de chegada (quando todos os indivíduos foram agrupados), mas algum lugar intermediário em que uma quantidade pequena de *clusters* é capaz de juntar casos muito semelhantes e separar casos muito diferentes.

Quando as junções começam a ocorrer a uma altura muito elevada (cf. gráfico 12.4), isso indica que indivíduos muito dissimilares estão sendo agregados. Portanto, cabe ao analista escolher a posição do dendograma imediatamente anterior a essas junções. *O corte horizontal imediatamente abaixo das maiores alturas indica, portanto, quantos clusters devem ser retidos*. O gráfico 12.4 ajuda muito nessa decisão. Como é possível perceber, as duas últimas junções ocorrem a uma altura mais elevada. Antes da primeira junção, havia dois *clusters*. Antes da penúltima, três. Em suma, o analista pode contar quantas barras são muito mais elevadas

que as demais e somar 1 unidade. Assim, encontrará a quantidade de *clusters* que devem ser retidos. Observe que esse procedimento é bastante visual. Em algumas situações, a decisão pode ser um pouco ambígua.

Gráfico 12.4: Gráfico de barras das alturas em que ocorrem as junções



A matriz de dissimilaridade

A maneira mais versátil e estável para realizar um agrupamento hierárquico no ambiente R consiste em empregar as funções **`hclust`** e **`get_dist`** das bibliotecas **`stats`** e **`factoextra`**. Ao trabalharem juntas, essas funções cumprem papéis diferentes. A função `get_dist` deve ser avaliada primeiro. Ela calcula as dissimilaridades entre todas as instâncias individuais de um *dataframe* e armazena essa informação em uma matriz. Nessa etapa, o analista tem a oportunidade de escolher qual método será empregado para calcular as dissimilaridades.

Dos métodos disponíveis na função `get_dist`, há dois muito populares (Husson; Lê; Pagès, 2010).

- O **método euclidiano** calcula distâncias usuais, aplicando uma generalização do teorema de Pitágoras. Ele é consistente com a noção de distância empregada na análise de componentes principais e está definido como padrão na função `get_dist`. Apesar da sua

popularidade, por trabalhar com diferenças elevadas ao quadrado, esse método é mais sensível à presença de *outliers*, ou seja, a presença de poucos indivíduos muito discrepantes pode alterar completamente o agrupamento.

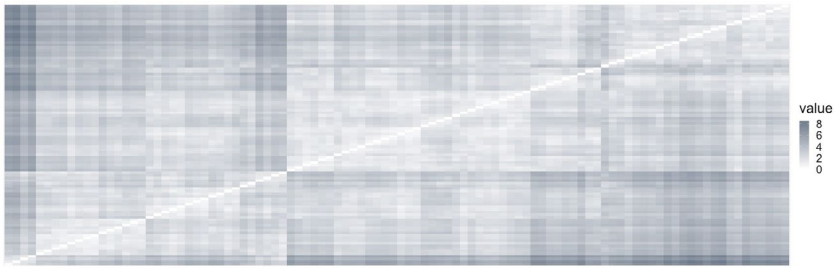
- O **método Manhattan** reduz o peso dos *outliers* calculando a soma dos módulos das diferenças absolutas (sem quadrados nem raízes). Ele recebe esse nome porque corresponde às distâncias percorridas quando precisamos nos deslocar por uma cidade onde todas as ruas são paralelas ou perpendiculares.

Para manter a consistência com a análise de componentes principais que realizamos anteriormente, calcularei a *matriz de dissimilaridade* aplicando o método euclidiano. A matriz de dissimilaridade pode ser visualizada com a função **`fviz_dist`** da biblioteca `factoextra` conforme mostramos a seguir. O resultado está no diagrama 12.2.

```
library(magrittr)

ENEM[,c(4:8)] %>%                       # Dados de ocorrência escolar
  get_dist(method = “euclidian”) %>%    # Calcula matriz de dissimilaridade
  fviz_dist()                           # Gera diagrama representando a matriz
```

Observe que empregamos, aqui, o operador *pipe*: **`%>%`**. Ele simplifica a composição de funções. Portanto, em vez de escrevermos **`A(B(x))`**, escrevemos **`x %>% B() %>% A()`**. Quando estamos compondo muitas funções, esse recurso ajuda a deixar o código mais legível.

Diagrama 12.2: Representação da matriz de dissimilaridade

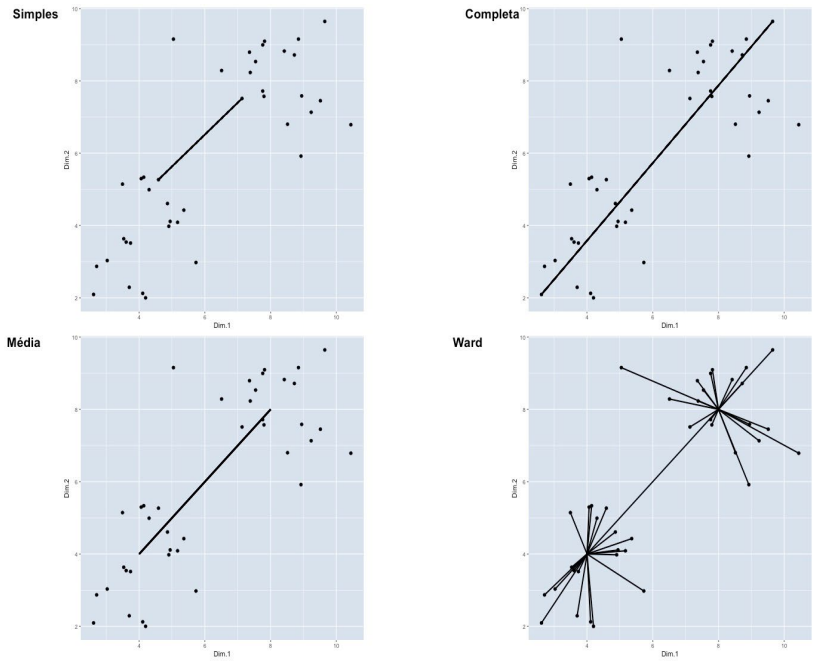


Realizando a análise de agrupamento hierárquico

Uma vez que a matriz de dissimilaridade tenha sido calculada, basta inseri-la na função `hclust` para gerar o agrupamento hierárquico. Porém, essa etapa da análise também requer especificar um método. Os mais populares são os seguintes (Husson; Lê; Pagès, 2010):

- método das **distâncias simples**, que considera a menor distância individual;
- método das **distâncias completas**, que leva em consideração a maior distância individual;
- método das **distâncias médias**, que considera a média das distâncias individuais;
- método de **Ward**, que leva em consideração a inércia entre *clusters* de maneira a minimizar a dispersão dos indivíduos em cada *cluster* ao longo dos sucessivos agrupamentos.

Gráficos 12.5-12.8: Representação dos métodos das distâncias simples, completas, médias e método de Ward



O analista deve saber que há ainda duas versões do critério de Ward na literatura. Uma delas eleva as distâncias euclidianas ao quadrado ("Ward.D2") e a outra não ("Ward.D"). Como nenhum desses critérios é desaconselhável *a priori*, é usual que o analista experimente todas as combinações possíveis até chegar a uma solução que lhe satisfaça… Os leitores que, no fundo, acreditam que fazer ciência é encontrar a única resposta possível a uma pergunta devem estar apavorados com a quantidade de opções e a arbitrariedade dos critérios para escolher uma em detrimento de outra! De qualquer maneira, os critérios de Ward combinados à medida euclidiana de dissimilaridade são os mais empregados.

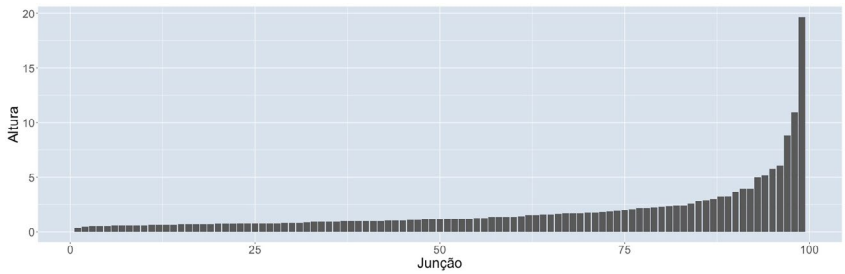
A seguir, levamos novamente as variáveis do Enem na função **`get_dist`** (para calcular distâncias euclidianas) e o resultado dessa operação é levado à função **`hclust`** (para realizar o agrupamento hierárquico segundo o método de Ward):

```
library(cluster)

res.hc = ENEM[,c(4:8)] %>%            # Dados do ENEM
  get_dist(method = "euclidian") %>%  # Matriz de dissimilaridade euclidiana
  hclust(method = "ward.D2")          # Agrupamento hierárquico com o critério de Ward
```

O resultado do agrupamento hierárquico está armazenado na estrutura de dados **`res.hc`**. Você pode fazer **`str(res.hc)`** para inspecionar essa estrutura e perceber que as alturas em que ocorrem as junções estão registradas em **`res.hc$height`**. Isso nos permite gerar o gráfico de barras que desejamos com o seguinte comando:

```
barplot(res.hc$height)
```

Gráfico 12.9: Alturas dos agrupamentos – dados Enem



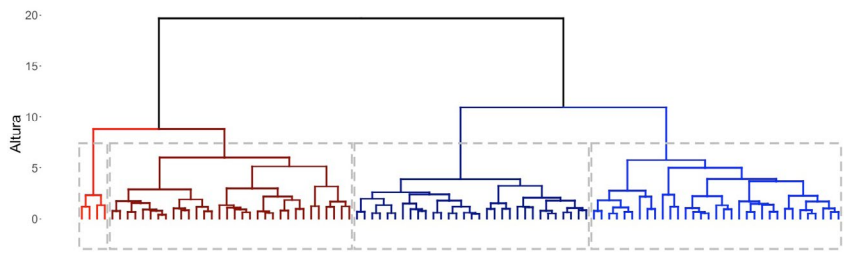
Quando o agrupamento hierárquico for realizado com o critério de Ward e distâncias euclidianas, as alturas podem ser interpretadas como ganhos em inércia após cada junção (Husson; Lê; Pagès, 2010). Como a

inércia é uma medida de dispersão, o aumento da inércia corresponde a *clusters* necessariamente mais dispersos, mais dissimilares em sua composição.

Como regra geral, o analista pode discriminar a quantidade de *clusters* com base na leitura do gráfico das alturas em que ocorrem as últimas junções (gráfico 12.9). Para tomar uma decisão informada, o analista pode evocar um critério semelhante ao proposto por Catell para a análise do *scree plot* (cf. lição 10). Ou seja, recomenda-se prevenir as junções que ocorram em alturas visivelmente mais elevadas que as anteriores. No gráfico 12.9, da direita para a esquerda, os três primeiros degraus estão se sobressaindo. Os demais, nem tanto. Portanto, estou optando por reter somente quatro agrupamentos. Enfim, o dendograma pode ser visualizado com a função **`fviz_dend`** a seguir (diagrama 12.3).

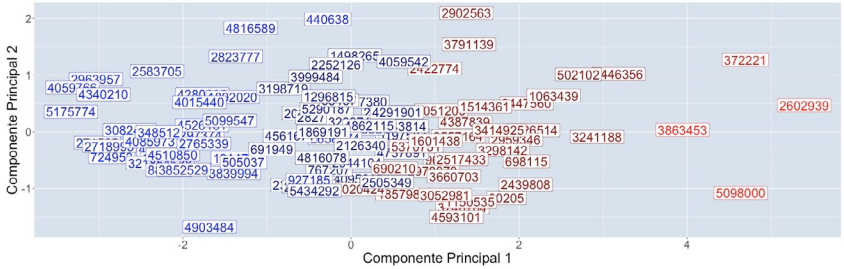
```
fviz_dend(res.hc)
```

Diagrama 12.3: Dendograma de 50 estudantes obtidos aleatoriamente do Enem 2018



O algoritmo recomenda criarmos *quatro clusters*. No gráfico 12.10, percebemos que esses *clusters* distinguem os estudantes em função de seu desempenho global.

Gráfico 12.10: Análise de componentes principais dos indivíduos



Algoritmo de partição (K-means)

Por mais atraentes que sejam os métodos de agrupamento hierárquico, eles enfrentam graves problemas quando o tamanho da amostra aumenta. Em primeiro lugar, são algoritmos muito pesados. Além de agrupar *clusters* dois a dois, cada operação requer calcular e comparar diversas distâncias e escolher a menor. Portanto, agrupar uma base de dados com alguns milhares de casos pode ser demorado. Além disso, as junções detalhadamente representadas no dendograma perdem o sentido quando não podem ser propriamente visualizadas. Um algoritmo mais rápido que produza uma classificação aproximada é muito mais vantajoso nessas situações.

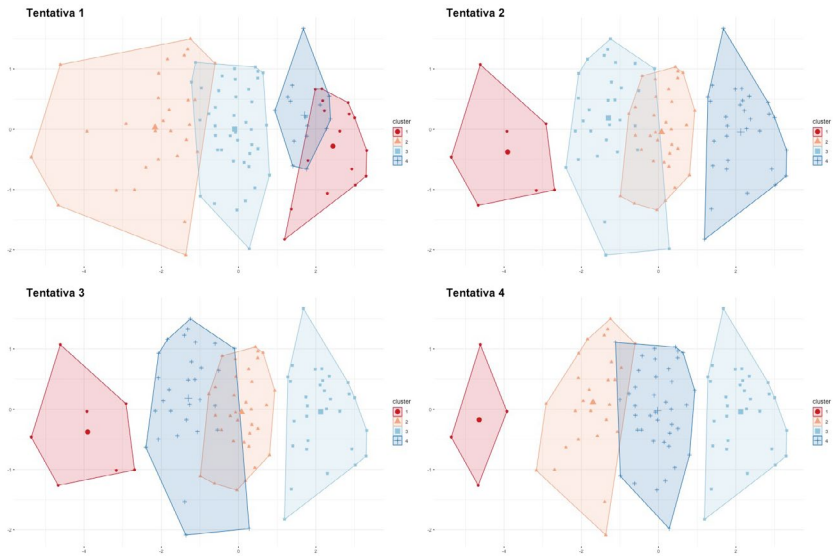
Dos *algoritmos de partição*, o método *k-médias* é o mais popular (Husson; Lê; Pagès, 2010). Ele requer que o analista informe quantos *clusters* deseja no final da sua solução. Seja K a quantidade de *clusters* solicitada pelo analista, o algoritmo posiciona K centroides espalhados no espaço e agrupa os indivíduos ao *cluster* do centroide mais próximo. Uma vez que todos foram devidamente classificados, a posição dos centroides é recalculada e os indivíduos são todos novamente redistribuídos. O algoritmo é interrompido quando os centroides não se movem mais. Em geral, essa convergência é rápida e ocorre em poucas iterações.

O problema da definição *a priori* da quantidade de *clusters* pode ser contornado realizando uma análise hierárquica com uma amostra pequena da base de dados que desejamos classificar. Outro problema é que, como o ponto de partida do algoritmo k-médias é sempre aleatório, não há garantia que a convergência ocorrerá exatamente para a mesma resposta. Porém, quando estamos trabalhando com bases de dados muito numerosas, divergências na classificação de algumas dezenas ou centenas de indivíduos pode ser absorvida como flutuação estatística em torno de uma tendência central.

Para ilustrar como esse o procedimento é simples, extraímos e representamos dois *clusters* com o método k-médias. A mesma extração foi feita em quatro tentativas, para mostrar como os agrupamentos obtidos por esse método estão sujeitos a flutuação estatística:

```
kmeans(ENEM[,c(4:8)], 4) %>%          # Algoritmo de partição k-médias (4 dim).
fviz_cluster(data = ENEM[,c(4:8))     # Gera representação gráfica
```

Gráficos 12.11-12.14: Representação dos clusters produzidos pelo método de partição k-médias



Uma visão geral sobre a análise exploratória multivariada

Nesta lição e nas duas últimas, nós aprendemos alguns métodos *exploratórios* e *descritivos* da análise multivariada. De maneira geral, eles têm como objetivo encontrar padrões de associação entre variáveis (representadas pelas colunas do *dataframe*) e padrões de similaridade entre instâncias individuais (representadas por linhas). Padrões de associação entre variáveis podem ser obtidos pela diagonalização da matriz de covariância (no caso de variáveis escalares linearmente associadas) ou da matriz dos resíduos padronizados (no caso de variáveis categóricas). Essas duas diagonalizações resultam, respectivamente, na *análise de componentes principais* e *de correspondência múltipla*.

A *análise de clusters* que aprendemos aqui pode ser considerada um método exploratório-descritivo complementar especificamente voltado ao tratamento das instâncias individuais. A representação ótima da nuvem de indivíduos já é possível tanto na análise de componentes principais quanto na análise de correspondência. Porém, a análise de *clusters* permite traçar fronteiras que agregam casos semelhantes e separam casos diferentes.

Dos diversos métodos na análise de *clusters*, os dois mais populares são: *i)* o agrupamento hierárquico (com o critério de Ward); e *ii)* o método de partição k-médias. O agrupamento hierárquico é mais ambicioso e fornece mais informações ao analista. Porém, pode ser inviável em amostras muito numerosas. O algoritmo k-médias é a alternativa mais comum para agrupar grandes bases de dados.

Revisando a lição

Enfim, dando continuidade às lições anteriores e finalizando nossa incursão pelos métodos exploratórios e descritivos de análise multivariada, nesta lição você aprendeu a:

1. calcular uma matriz de dissimilaridade por meio da função **`get_dist`** segundo os métodos "euclidian" e "manhattan";
2. gerar uma representação visual da matriz de dissimilaridade com a função **`fviz_dist`**.
3. computar a análise de agrupamento hierárquico com a função **`hclust`** segundo os métodos "*single*", "*complete*", "*avarage*", "Ward.D" e "Ward.D2";
4. gerar um gráfico de barras das alturas em que ocorrem as junções do agrupamento hierárquico com o propósito de decidir quantos agrupamentos devem ser mantidos;
5. gerar uma representação gráfica do dendograma com a função **`fviz_dend`**;
6. computar a análise de agrupamento por partição usando o algoritmo k-médias implementado na função **`kmeans`**;
7. gerar um gráfico desses *clusters* com a função **`fviz_cluster`**.

PARTE V.

ANÁLISES FATORIAIS

Ao longo das últimas seis lições, foram abordadas duas famílias de métodos multivariados. A primeira delas tinha um caráter predominantemente *confirmatório* e *inferencial* (cf. "Parte III"). De fato, o modelo linear geral compreende um conjunto diversificado de métodos (e.g., regressão linear múltipla, análise de variância, análise de covariância, regressão logística) que permitem testar a significância estatística de modelos especificados pelo usuário, bem como ajustar os parâmetros desses modelos. As relações que avaliamos eram claramente *preditivas*, ou seja, os modelos permitiam medir o *tamanho do efeito* das variáveis independentes sobre uma variável dependente. Esses métodos são particularmente interessantes para modelar relações causais imediatas e bem definidas, tais como "o efeito de A sobre B".

A segunda família de métodos abordada tinha caráter predominantemente *exploratório* e *descritivo* (cf. "Parte IV"). De fato, as análises de componentes principais, de correspondência e de *clusters* não estão propriamente voltadas a testar a significância estatística de modelos desenhados pelo analista. Questões como "o efeito de A sobre B" não são colocadas. Ao contrário, esses métodos permitem explicitar *padrões de associação* entre variáveis e *padrões de similaridade* entre instâncias

individuais. Justamente por não fazerem uma distinção clara entre variáveis dependentes e independentes, essa família favorece pensar a causalidade de maneira mais sistêmica, distribuída.

A família de métodos que será tratada nas lições seguintes permite estabelecer uma relação tanto exploratória quanto confirmatória com os dados. Ela permite avaliar padrões de associação entre variáveis por meio de reduções dimensionais e, ao mesmo tempo, testar modelos preditivos. Genericamente designada por *análise fatorial*, essa família está entre as preferidas da pesquisa educacional. Herdada do campo da psicometria, a análise fatorial está inicialmente comprometida com a validação dos processos de medição de algumas propriedades humanas. Por exemplo, um psicólogo pode perguntar ao seu cliente se ele se sente deprimido, assim como um professor pode perguntar ao estudante se ele compreendeu o que foi ensinado. Contudo, se quisermos alguma informação confiável sobre essas propriedades humanas, precisamos construir testes específicos e avaliar criticamente o quanto os dados produzidos por esses testes são confiáveis.

Observe que, até o presente momento, a questão da *validação das medidas educacionais* não foi realmente discutida. Esse talvez seja o mais fundamental de todos os temas em análise quantitativa de dados educacionais. Até agora, nós temos trabalhado com dados coletados por grupos de pesquisa ou instituições sem questionar realmente o quanto esses dados permitem fazer afirmações confiáveis. A saber, essa crítica da qualidade dos dados é importantíssima para orientar a construção de testes educacionais e para sabermos como nos posicionarmos diante das evidências produzidas por esses testes. Afinal, até que ponto a nota dos estudantes na prova de ciências da natureza do Enem (recorrentemente empregada em todos os nossos exemplos) realmente mede a competência científica dos estudantes?

Portanto, o contexto usual de aplicação da análise fatorial é um conjunto de dados produzidos por meio de um teste cuja confiabilidade está sendo avaliada. A ideia básica desse tipo de análise é que as respostas dadas aos itens do teste podem ser interpretadas como se resultassem de *fatores comuns* ou *variáveis latentes* que o teste permite identificar. Assim como outros métodos discutidos até aqui, as análises fatoriais presumem que os itens de um teste (também chamados *variáveis manifestas*) estejam mutuamente associados segundo uma estrutura de relações. Alguns estarão positivamente ou negativamente associados, enquanto outros não terão relação alguma. Esse padrão de associação entre os itens do teste é fundamental para afirmarmos (ou não) a existência de fatores subjacentes às respostas dos participantes.

O leitor perceberá logo que há muita similaridade entre as análises fatoriais e os demais métodos multivariados que discutimos até aqui. De fato, continuaremos falando do teorema espectral, autovalores, autovetores e diagonalizações. Porém, é preciso estar atento às diferenças. Além de algumas especificidades propriamente estatísticas que ainda serão discutidas, as análises fatoriais são tradicionalmente mais empregadas na *validação de escalas psicométricas e educacionais*, enquanto a análise de componentes principais, de correspondência e de *cluster* são mais empregadas em estudos de caráter exploratório, mineração de dados (ing., *data mining*) e aprendizado de máquina (ing., *machine learning*). Em geral, nós não avaliamos a consistência interna de um questionário socioeconômico, mas algum conhecimento de psicometria pode ajudar muito o delineamento desse tipo de instrumento.

Enfim, as próximas lições estão organizadas da seguinte maneira.

- **Lição 13** (testes educacionais): serão discutidos os fundamentos da teoria clássica de testes, que orienta a elaboração e validação de testes educacionais de diversos tipos.

- **Lição 14** (análise fatorial exploratória): tratará de uma ferramenta que pode ser empregada quando não sabemos exatamente quantas dimensões têm o teste nem quais itens compõem cada dimensão. Os fatores produzidos na análise exploratória são geralmente ortogonais. As finalidades e as situações de uso dessa ferramenta são muito semelhantes à análise de componentes principais, por isso, as duas costumam ser confundidas.
- **Lição 15** (análise fatorial confirmatória): tratará de uma ferramenta que pode ser empregada quando o analista deseja testar a plausibilidade de fatores cuja composição provável já está definida *a priori*. Em outras palavras, antes da aplicação do teste, nós sabemos quais itens devem compor cada fator. Os fatores geralmente não são ortogonais e a análise pode ser expandida para testar a plausibilidades de relações preditivas e associativas entre esses fatores. Nesse caso, o método de análise passa a ser chamado *modelagem de equações estruturais*.

Ao lado de modelos baseados na teoria da resposta ao item, a *modelagem de equações estruturais* é um dos métodos mais versáteis e mais empregados na pesquisa educacional. Ela se conecta a tudo o que vimos até aqui. É o único método abordado neste livro que permite incorporar, ao mesmo modelo, testes preditivos e associativos. Ou seja, algumas equações estruturais são do tipo "A está associado a B", enquanto outras são do tipo "A produz um efeito em B". Portanto, a modelagem de equações estruturais permite elaborar modelos complexos em que diversos fatores, cada um deles avaliados por vários itens, mantêm relações de natureza diferente.

LIÇÃO 13.

TESTES EDUCACIONAIS

Até aqui trabalhamos com dados coletados por entidades ou grupos de pesquisa cujos métodos não foram questionados, mas poderiam ter sido. A partir de agora, vamos aprender algumas ferramentas básicas para elaborar nossos próprios testes educacionais e avaliar a qualidade da informação que eles produzem. De fato, o analista deve ser a primeira pessoa a pôr em dúvida a confiabilidade dos dados com os quais trabalha, antecipando-se às críticas que razoavelmente podem ser feitas por seus pares. Em geral, há duas classes de considerações que podemos fazer para estabelecer a confiabilidade dos resultados de um teste educacional: a primeira diz respeito à *relação do dado com a realidade*; a segunda, por sua vez, relaciona-se à *estrutura e consistência interna* do teste. Com essas duas preocupações em mente, é possível desenvolver bons testes educacionais.

Relação com a realidade

O primeiro conjunto de questões está voltado à *relação com a realidade*. Confiar nos dados é, também, confiar que eles medem aquilo que dizem medir. Um teste de inteligência deve medir inteligência (não outra coisa), operacionalizando todos os aspectos relevantes do construto tal

como ele está definido na fundamentação teórica adotada. Ao mesmo tempo, é esperado que um novo teste de inteligência, dentro de certos limites, esteja correlacionado a testes já validados na literatura e outras variáveis que a inteligência deve ser capaz de predizer.

Tanto em ciências naturais quanto em ciências sociais, estamos sempre trabalhando com construtos inventados. Porém, na pesquisa educacional, o caráter teórico-conjectural dos objetos de medição fica muito mais evidente: competências, habilidades, interesses, atitudes, motivações, crenças, disposições são observados na realidade, mas só estão delimitados e só podem ser operacionalizados no âmbito de teorias muito específicas. Geralmente, modelos teóricos diferentes podem delimitar noções básicas de maneira diferente. Portanto *a relação com a realidade supõe sempre uma relação com a teoria*.

Em suma, quando conduzimos nossa pesquisa, precisamos confiar que, dentro de certos limites práticos, *o dado disponível é a melhor representação possível do ente teórico-real que queremos investigar*. Mas como podemos avaliar criticamente essa questão? Um procedimento simples e muito adotado em testes educacionais consiste em submetê-los a pessoas que não estiveram diretamente envolvidas na sua elaboração, mas que são capazes de fazer um julgamento qualificado dos itens. Essa consulta a "especialistas" pode contribuir para aprimorar o teste de diversas maneiras.

Para iniciar a consulta, é necessário ter especificado conceitual e operacionalmente *os objetos que o teste pretende medir*. Por exemplo, em grandes avaliações escolares (Enem, vestibulares), essa especificação costuma ser feita por meio de uma matriz de conteúdos, competências, habilidades. Ao compararem os itens do teste com a matriz, os avaliadores devem ser capazes de identificar se:

1. alguns itens não representam bem aquilo que o teste pretende medir;
2. algumas componentes da matriz não foram devidamente contempladas no teste.

Esse processo costuma ser chamado *validação de conteúdo*.

A propósito, também costumam estar especificadas algumas regras de formato pensadas a partir da maneira como os respondentes tendem a interagir com o teste. Nas questões do Enem, por exemplo, algumas exigências de formato pretendem evitar que alunos menos competentes sejam atraídos para a resposta correta, ou que alunos muito competentes fiquem excessivamente cansados a longo do exame. Um exame muito cansativo (é difícil afirmar que esse não seja o caso do Enem) acaba se tornando mais um teste de resistência e menos um teste de competência. Um exame que induza os estudantes mais habilidosos à resposta errada pode medir mais esperteza e menos conhecimento. Portanto, antes da primeira aplicação, também é necessário avaliar se:

- o formato de apresentação dos itens do teste, em vista daquilo que sabemos sobre o processo de resposta, prejudica que eles meçam o que pretendem medir;

Após a aplicação, é possível avaliar ainda mais a relação dos dados com a realidade. Isso pode ser feito confrontando os dados com outras informações disponíveis e confiáveis. Tais procedimentos estabelecem o que tipicamente é chamado *validade concorrente* e consistem em investigar

1. se os resultados do nosso teste estão correlacionados ao resultado de outro teste consagrado quando os dois são aplicados aos mesmos participantes;
2. em que medida os resultados do teste proposto permitem prever acontecimentos futuros razoavelmente relacionados ao objeto do teste.

O primeiro tipo de avaliação requer aplicar o teste novo junto com um teste antigo às mesmas pessoas. O segundo tipo, por sua vez, requer incorporar outras informações disponíveis sobre os participantes. Por exemplo, após aplicar um teste sobre o interesse por carreiras científicas, é esperado que os itens desse teste tenham algum poder preditivo sobre os percursos profissionais dos estudantes (Archer *et al.*, 2015).

Todas essas evidências podem ser apresentadas como resposta à pergunta: *será que esse teste educacional mede aquilo que pretende medir?*

Estrutura e fidedignidade

O segundo conjunto de questões está voltado à *reprodutibilidade* das medidas. Confiar nos dados é, também, confiar que eles assumiriam valores muito semelhantes se fossem coletados novamente sob as mesmas condições. A pontuação em dois testes de conhecimento equivalentes não pode variar muito entre aplicações sucessivas. Para um intervalo suficientemente curto entre teste e reteste, a posição dos estudantes na escala não pode mudar muito.

Há várias formas de avaliar a estabilidade de uma medida. Todas envolvem a ideia de *repetição*. Para saber como um teste se comportará quando reaplicado sob as mesmas condições, fazemos com que o teste seja, ele mesmo, uma coleção de vários testes. De fato, testes educacionais são muito diferentes de provas tradicionais porque eles são carregados de observações repetidas. As associações entre as partes de um teste permitirão saber quão fidedigno é seu resultado.

A *consistência interna* de um teste depende, portanto, da medida com que seus itens estão mutuamente associados. É claro que essa associação pode estar restrita a certos grupos de itens, ou seja, alguns itens podem estar mutuamente associados dentro de seu grupo, mas pouco associados a itens de

outros grupos. Um exemplo de teste educacional desse tipo é a pesquisa sobre preconceito escolar discutida na lição 10. Nela, era possível distinguir itens mais relacionados ao racismo e homofobia, por um lado, e itens mais relacionados a outros tipos de preconceito (contra a mulher, idosos e pessoas com deficiência). Em casos como esse, dizemos que o teste é *multidimensional*.

Embora possamos ter uma ideia *a priori* da quantidade dimensões de um teste, é somente *a posteriori* (i.e, após a coleta dos dados) que conseguimos avaliar sua estrutura interna e a consistência dos padrões de associação entre os itens. É geralmente este o principal propósito das análises fatoriais: investigar a estrutura e consistência interna de testes multidimensionais.

Validade e validação

Em suma, é preciso considerar que o processo de produção das evidências pode comprometer ou limitar a confiabilidade das conclusões. Em pesquisa educacional, *validade* refere-se à qualidade das inferências, declarações ou decisões baseadas em respostas dadas a um instrumento e *validação* é o processo que pretende produzir evidências capazes de sustentar a adequação, significância e utilidade dessas decisões e inferências (Zumbo; Chan, 2014). Embora seja comum falar em “testes validados”, o que validamos realmente não é o instrumento, mas o processo do qual o instrumento faz parte. Ao validar o processo de coleta e análise dos dados, validamos as conclusões que podem ser tomadas a partir da evidência produzida. Em termos muito práticos, um questionário já validado pode não produzir resultados confiáveis quando aplicado em outra época, a outra população, em outras condições ou com outros propósitos.

Ainda que alguns textos tentem simplificar a discussão, a validação de um teste educacional é um terreno movediço. Não há um protocolo

universal a seguir nem um conjunto de critérios consensuais para garantir a validade das declarações feitas a partir de dados coletados. A validação é um processo argumentativo que não deve estar baseado em uma única fonte de evidência. As diretrizes contemporâneas destacam cinco fontes principais que podem sustentar uma declaração de validade (Zumbo; Chan, 2014):

1. o conteúdo do teste,
2. o processo de resposta aos itens,
3. a estrutura interna do teste,
4. as relações com outras variáveis, e
5. suas consequências.

Observe que todas essas fontes foram discutidas nas duas seções anteriores, mas recorrer a todas não é obrigatório. Por exemplo, muitos pesquisadores em educação valorizam a "validade de face" dos itens. Porém, em alguns casos, pode ser importante que o instrumento dificulte ao participante perceber o que está sendo medido. Nessas situações, um instrumento com grande "validade de face" seria, na verdade, indesejável. É possível também que o analista pretenda questionar os resultados produzidos por instrumentos consagrados na literatura. Nesse caso, não faz muito sentido esperar que o resultado do teste se correlacione com outros. A *estrutura interna do teste* é particularmente importante para o que aprenderemos nesta lição e nas duas seguintes. Ela não é a única fonte de validade que pode ser avaliada quantitativamente, mas costuma empregar métodos específicos.

A teoria clássica de testes

Existem algumas considerações simples que compõem o fundamento da chamada *teoria clássica de testes*. Segundo ela, o valor observado de

uma variável escalar pode ser decomposto em um *valor-verdadeiro* e um *erro* da seguinte maneira (Mair, 2018):

$$x = x_v + \epsilon$$

Tentando colocar em suspenso todos os constrangimentos filosóficos em torno da noção de "valor verdadeiro", podemos reconhecer que ela cumpre aqui uma função interessante: separar a realidade da observação. Ao distinguir entre o valor observado e o "valor verdadeiro", descolamos nossas observações da realidade (mas não completamente), lembrando-nos de que *todo resultado de medição tem uma incerteza associada* (Lima Junior; Da Silveira, 2011).

A ideia aqui é que todas as observações se desviam um pouco daquilo que deveriam ser em condições ideais. Por exemplo, o desempenho de um estudante em avaliações escolares é geralmente traduzido em uma nota. Se o mesmo estudante fizesse um teste equivalente sob as mesmas condições, teria uma nota diferente! Se reaplicássemos um teste educacional várias vezes, teríamos muitas notas diferentes atribuídas ao mesmo estudante, sem que suas competências tenham sofrido qualquer tipo de transformação. Isso ocorre porque *o resultado de um teste educacional nunca depende somente daquilo que ele efetivamente mede, mas de outras circunstâncias que não estão sendo controladas* (tal como a sensação de bem-estar do candidato no dia da avaliação ou a presença de perguntas que, por acaso, ele teve a oportunidade de estudar na semana anterior ou a infelicidade de jamais ter visto). Os efeitos aleatórios dessas circunstâncias somam-se como ruído sobre o resultado que deveríamos obter se as observações não estivessem sujeitas à flutuação.

Dito isso, podemos definir a *fidedignidade* do resultado de um teste como a razão da variância do valor verdadeiro pela variância do valor

observado. Posto que erro e valor verdadeiro devem ser variáveis ortogonais, temos o seguinte:

$$fid \equiv \frac{\sigma_{x_v}^2}{\sigma_x^2} = \frac{\sigma_{x_v}^2}{\sigma_{x_v}^2 + \sigma_\epsilon^2}$$

Como o valor observado é igual ao verdadeiro somado a um ruído, temos que a dispersão do valor observado (denominador) será sempre maior que a dispersão do valor verdadeiro (numerador). Assim, a fidedignidade será sempre uma quantidade positiva entre 0 e 1, podendo ser interpretada como a *fração de verdade em uma observação*. Voltaremos a essa definição posteriormente.

A falácia da atenuação das diferenças

Várias consequências emergem de reconhecermos que o valor observado em uma medida carrega um erro em sua composição. Tanto a maneira fetichista com que as classificações escolares são celebradas quanto a maneira iconoclasta com que são criticadas refletem um desconhecimento sobre a natureza aleatória das medidas. Os primeiros colocados de um vestibular são certamente estudantes muito competentes, mas não podemos garantir que eles sejam os mais competentes. Muito provavelmente, sua nota resulta de competência real acrescida a um favorecimento casual. Na experiência cotidiana, ao participar de processos seletivos muito concorridos, essa casualidade corresponde à sensação de que não basta ser competente: é preciso ter sorte para ser aprovado.

O que podemos dizer sobre os primeiros colocados em uma avaliação escolar, senão que eles tendem a perder sua posição de liderança?

É possível recuperá-la depois e perdê-la novamente… Isso ocorre porque a posição dos indivíduos em uma escala (de competência, interesse, visão de ciência) está *sempre sujeita a flutuação*. Ao reaplicar testes de conhecimento várias vezes, é comum observarmos que os alunos mais competentes tiverem uma queda de desempenho enquanto os alunos classificados como menos competentes tiveram o maior ganho. Em vista da nossa sensibilidade diante das violências praticadas em nome da avaliação escolar e de toda a vaidade que elas mobilizam, é tentador celebrar a queda dos primeiros colocados bem como a ascensão dos últimos. Contudo, *a queda dos primeiros e a ascensão dos últimos é esperada em qualquer medida sequencial onde o erro aleatório tenha um papel expressivo*. Esse comportamento não permite inferir que algo esteja realmente ocorrendo entre os sujeitos.

Vamos criar um modelo hipotético que nos permita entender melhor a *ilusão de atenuação das diferenças* que acabamos de descrever. Considere uma escola de 1.000 estudantes em que a habilidade matemática verdadeira seja normalmente distribuída. Se essa habilidade puder ser representada em uma escala com média 500 e desvio padrão 100, temos:

```
verd = rnorm(1000, mean = 500, sd = 100)
```

Ao aplicar um teste de habilidade matemática, os valores observados serão iguais ao valor verdadeiro somado a um erro (ou ruído) devido às circunstâncias externas que o teste não pretende medir, mas que interferem no resultado. Considere que o erro da medida também seja normalmente distribuído e tenha desvio padrão igual a 50:

```
erro1 = rnorm(1000, mean = 0, sd = 50)
obs1 = verd + erro1
```

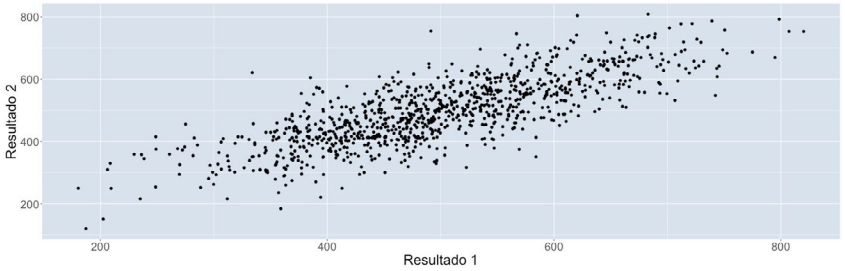
Agora imagine que a escola implementou uma mudança curricular que prometia contribuir para o aprendizado dos estudantes. Após a mudança, os mesmos estudantes foram submetidos a um teste de habilidades matemáticas equivalente ao primeiro. Como se trata de um modelo hipotético, *suponha que a mudança curricular não produziu nenhum efeito verdadeiro*. Nós podemos simular essa situação mantendo iguais os valores verdadeiros nas duas aplicações do teste, mudando somente o componente aleatório da medida:

```
erro2 = rnorm(1000, mean = 0, sd = 50)
obs2 = verd + erro2
```

Você pode gerar um diagrama de dispersão para visualizar como as pontuações dos estudantes nos testes (valores observados) se relacionam quando não há aprendizagem real:

```
df = data.frame(obs1, obs2)
plot(df)
```

Gráfico 13.1: Testes educacionais aplicados sucessivamente



É assim que se parece a dispersão de um pré-teste e pós-teste com ganho estritamente igual a zero. Conforme esperado, o diagrama indica duas medidas correlacionadas. Você pode medir essa correlação e,

em seguida, fazer um teste t para comparar as médias e perceber que, de fato, não é possível identificar qualquer ganho de aprendizagem:

```
cor(df)

##           obs1      obs2
## obs1 1.0000000 0.8060406
## obs2 0.8060406 1.0000000

t.test(df$obs1, df$obs2)

##
##  Welch Two Sample t-test
##
## data:  df$obs1 and df$obs2
## t = 0.40985, df = 1996.8, p-value = 0.682
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -7.893312 12.064072
## sample estimates:
## mean of x mean of y
##  497.8478  495.7625
```

Agora, tente se colocar no lugar do pesquisador. Ele não tem acesso aos valores verdadeiros, somente aos valores observados. Ao comparar as médias, ele percebeu que não houve ganho significativo e que, ao menos em princípio, a reforma curricular foi um grande fracasso… Insistindo um pouco, ele resolve investigar o que está ocorrendo com os estudantes mais fracos e algo lhe chama a atenção:

```
t.test(df$obs1[obs1 < 300], df$obs2[obs1 < 300])

##
##  Welch Two Sample t-test
##
```

```
## data:  df$obs1[obs1 < 300] and df$obs2[obs1 < 300]
## t = -2.7863, df = 50.354, p-value = 0.007498
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -79.96521 -12.97729
## sample estimates:
## mean of x mean of y
##  250.8641  297.3353
```

Entre os estudantes mais fracos (com pontuação inferior a 300 pontos no primeiro teste), parece haver um ganho estatisticamente significativo. Entre os estudantes de mais alto de desempenho, ocorre o contrário: eles parecem ter "desaprendido":

```
t.test(df$obs1[obs1 > 700], df$obs2[obs1 > 700])
```

A conclusão falaciosa consiste em afirmar que está havendo realmente uma *atenuação das diferenças* nessa escola, ou seja, que algo contribuiu para que os estudantes mais fracos aprendessem mais enquanto os estudantes mais competentes desaprendessem… Após algum tempo, seria esperado que todos os estudantes se encontrassem na média e que essa atenuação das diferenças produzisse uma *convergência para a média*. No entanto, tal atenuação espontânea das diferenças nunca ocorreu na história dos sistemas de ensino. Talvez por isso, o analista equivocadamente conclua que aquela mudança curricular, apesar de não ter produzido ganhos em média para todos os alunos, contribuiu para fazer com que os alunos mais fracos ficassem menos fracos e que os mais fortes ficassem menos fortes… Você, que agora ocupa a posição de "Deus" (conhecedor dos valores verdadeiros!), sabe o quanto esse analista está equivocado.

O equívoco consiste em não perceber que a presença de erro na medida educacional cria uma *ilusão de atenuação das diferenças*. Em outras

palavras, em qualquer teste educacional, é esperado que os maiores ganhos ocorram sempre entre aqueles que apresentaram o pior desempenho na primeira aplicação. Essa realidade, contudo, não produz uma convergência para a média, mas uma flutuação aleatória das posições individuais tal qual uma dança das cadeiras. A ascensão dos últimos é compensada pela queda de outros que passam a ocupar o seu lugar (vide gráfico 13.1). Intimamente relacionado à presença de um componente aleatório da medida, o argumento apresentado nesta seção permite suspeitar de diversos resultados que celebram prematuramente a atenuação das diferenças sociais ou escolares. Assim como a mobilidade social dos indivíduos não representa uma ruptura da sociedade de classes, ou o sucesso profissional de algumas mulheres negras não represente a superação do racismo e da misoginia, os ganhos de aprendizagem daqueles que menos aprenderam geralmente não correspondem a uma transformação da estrutura das diferenças escolares — eventualmente, esse ganho sequer indica que houve aprendizagem real.

Enfim, nossas observações sempre terão algum erro, porém, para que nossas conclusões sejam confiáveis, é preciso que os erros sejam controlados. Testes fidedignos são aqueles que produzem aproximadamente os mesmos resultados quando reaplicados. Como nós mesmos geramos os dados do exemplo anterior, podemos calcular a fidedignidade dos testes dividindo a variância verdadeira pela variância observada.

```
100^2/(100^2 + 50^2)

## [1] 0.8
```

Na literatura, uma fidedignidade de 0.8 é considerada boa, atestando que os valores escolhidos para nossa simulação representam uma situação plausível e recorrente da pesquisa educacional.

A elaboração de testes educacionais

Já nas primeiras lições aprendemos que a média de uma série de observações flutua menos que as observações individuais. Isso ocorre porque, quando somamos medidas que contém erros aleatórios, as componentes de erro tendem a se cancelar justamente porque são aleatórias! Portanto, *testes educacionais de qualquer natureza são construídos em torno da ideia de repetição*. Uma vez que você delimitou conceitualmente o objeto do teste, precisa operacionalizá-lo em uma série de itens não idênticos, mas equivalentes. Testes educacionais consistem em perguntar aproximadamente a mesma coisa de maneiras diferentes.

Se você quer avaliar uma habilidade específica, pode fazer várias perguntas (abertas, de certo ou errado, de múltipla escolha) capazes de *discriminar* os estudantes com relação a essa habilidade. Itens bons têm maior poder de discriminação, colocando estudantes habilidosos de um lado e não habilidosos de outro. Para todos, haverá sempre uma possibilidade de erro ou acerto ao acaso, gerando ruído no resultado. Porém, se o teste consiste de vários itens não idênticos, mas equivalentes, *ao computar a média dos acertos, você terá maior precisão na discriminação dos estudantes*.

Também é importante pensar com cuidado no tipo de item do seu teste, pois tipos diferentes trazem desafios e oportunidades também diferentes.

- **Itens de resposta livre** são os mais comuns nas avaliações escolares artesanais, nas quais a quantidade de participantes é de poucas dezenas e a codificação das respostas pode ser feita de forma intuitiva, sem seguir um protocolo rígido. Analisar e codificar respostas abertas de maneira fidedigna não é impossível, mas costuma elevar os custos da pesquisa, podendo impor limites ao tamanho da amostra. A confiabilidade do resultado não dependerá tanto das perguntas feitas (por isso, é fácil fazê-las!),

mas do rigor e da significância do protocolo de classificação das respostas. É usual que o protocolo de classificação das respostas dadas a itens abertos tenha grandes problemas, comprometendo os resultados.

- **Itens de múltipla escolha** são os mais empregados em grandes avaliações educacionais e em questionários socioeconômicos. Bons itens desse tipo podem ser muito difíceis de construir, mas tendem a produzir resultados efetivos.[1]
    - Em avaliações educacionais, deve haver pelo menos uma resposta correta e vários distratores, alternativas equivocadas cuidadosamente plantadas para atrair os estudantes que não alcançaram a melhor compreensão de certo tema (ao mesmo tempo em que afastam os estudantes mais competentes). Tais distratores permitem mobilizar o que sabemos sobre as chamadas concepções alternativas dos estudantes. Permitem, também, gerar diagnósticos, apontando quais equívocos são mais comuns. Por outro lado, quando mal formulados, os distratores podem atrair a resposta dos estudantes mais habilidosos, prejudicando o poder de discriminação do item. Geralmente, as respostas a esse tipo de item são codificadas em uma variável binária, indicando se o estudante marcou ou não a resposta correta.
    - Em questionários socioeconômicos, não há respostas corretas tampouco distratores. Porém, ainda deve haver um cuidado na construção das alternativas. As múltiplas escolhas nesse tipo de questionário devem discriminar os sujeitos sem produzir categorias vazias. Variáveis categóricas com

[1] Para mais informações sobre a produção e a análise de itens de múltipla escolha em avaliações de larga escala, recomendo o livro *Avaliação educacional: fundamentos, metodologia e aplicações*, de Mauro Rabelo (2013).

níveis pouco frequentes tendem a produzir inconsistências na análise e, por isso, pode ser necessário eliminá-las ou recodificá-las ao longo do caminho.

- **Itens de certo ou errado**, por sua simplicidade, são mais fáceis de construir e validar, mas sua interpretação é menos rica. São ainda muito empregados em vestibulares e concursos públicos. Permitem produzir instrumentos com uma maior quantidade de observações, pois o tempo de resposta por item tende a ser menor. A ausência de vários distratores restringe o potencial diagnóstico do instrumento. A ausência de múltiplas escolhas aumenta a chance de acerto ao acaso, mas reduz o custo de produção e validação dos itens.
- **Itens tipo-Likert** são mais empregados para medir variáveis afetivas ou de opinião (e.g., atitudes, interesses, motivação, autoeficácia, crenças religiosas, visões de ciência). Consistem em fazer uma afirmação simples e solicitar que o participante da pesquisa manifeste seu grau de concordância em uma escala numérica que geralmente tem cinco posições: (1) discordo totalmente; (2) discordo parcialmente; (3) indiferente; (4) concordo parcialmente; (5) concordo totalmente. Itens desse tipo são bem fáceis de construir e costumam produzir resultados confiáveis quando alguns cuidados são tomados. Por exemplo, as afirmações devem ser feitas em períodos simples (períodos compostos são desencorajados). Advérbios como "sempre", "nunca", "muito", "pouco" tendem a alterar o padrão de resposta e devem ser empregados somente com o propósito de aumentar o poder de discriminação do item. É possível fazer afirmações negativas com relação ao que se mede, mas negações usando "não" devem ser evitadas para reduzir o ruído por distração do leitor.

A despeito do tipo escolhido, é importante termos em mente que itens em um teste educacional *devem constituir grupos de observações*

repetidas. Sem isso, não será possível avaliar a fidedignidade do teste. Se desejamos quantificar, com itens de certo ou errado, a presença de uma habilidade entre os estudantes, será preciso desenvolver uma bateria de itens voltados à mesma habilidade. Se desejamos quantificar, com itens tipo-Likert, o interesse dos estudantes por carreiras científicas, precisamos construir vários itens diferentes entre si, mas igualmente voltados ao mesmo objeto. Se desejamos medir a posição socioeconômica dos estudantes com itens de múltipla escolha, precisamos elaborar vários itens que, juntos, permitam localizar os respondentes no espaço das diferenças sociais.

Em provas escolares mais artesanais, não é usual pensar dessa maneira. Perguntamos o que desejamos saber uma vez só. Nas provas de física universitária, temos geralmente uma questão aberta para cada capítulo do livro (sem nenhuma preocupação em fazer observações repetidas). Além disso, provas escolares podem ser muito ecléticas, combinando todos os tipos de item que vimos anteriormente (além de muitos outros que não discutimos). Esse ecletismo da forma pode gerar problemas, pois itens de tipo diferente geram variáveis diferentes (binária, categórica, escalar) com poder de discriminação também diferente. Enfim, *testes educacionais e provas escolares não são a mesma coisa*. Do ponto de vista dos critérios de validação de um teste educacional, a maioria das avaliações escolares seria reprovada por falta de confiabilidade.

Alfa de Cronbach

Observe que a fidedignidade, da maneira como está definida, é praticamente uma entidade metafísica, pois depende de conhecermos quantidades inacessíveis em situações concretas de pesquisa (i.e., a variância do valor verdadeiro). Há, portanto, várias tentativas de estimar a fidedignidade de

uma medida. De todas as opções disponíveis, o *coeficiente alfa de Cronbach* pode ser considerado estrela do *rock* da psicometria (Mair, 2018). Ele tem limitações tremendas, mas ainda é o indicador mais empregado por tradição.

Esse coeficiente baseia-se na ideia de que todos os itens de uma bateria podem ser considerados testes paralelos, ou seja, supomos que cada item funcione como um teste educacional, medindo o mesmo objeto da mesma maneira. Portanto, o coeficiente alfa supõe que os itens da bateria sejam *unidimensionais*. Em outras palavras, estamos supondo que as respostas dadas aos itens sejam tão correlacionadas que não possamos delimitar dois ou mais grupos de itens muito associados em si, mas pouco associados entre si. Satisfeita a condição de unidimensionalidade, somos capazes de estimar a variância do valor verdadeiro pela covariância dos itens da bateria de uma maneira muito simples.

Com base na teoria clássica de testes, podemos afirmar que a soma das respostas aos itens da bateria é a melhor representação possível do valor verdadeiro. De fato, se a resposta dada a cada item é composta por valor verdadeiro + erro, ao somar essas respostas, os erros tendem a se cancelar mutuamente (supondo que sejam aleatórios). Designando por x_i o *i*-ésimo item de uma bateria com K componentes, a soma *X* das respostas a todos os itens pode ser escrita da seguinte forma:

$$X = x_1 + x_2 + ... + x_i + ... + x_K$$

Nessa situação, é possível demonstrar que a fidedignidade tem um *limite inferior* que pode ser estimado pelo coeficiente *alfa de Cronbach*:

$$\alpha = \frac{K}{K-1}\left(1 - \frac{\sum_{i=1}^{K}\sigma_{x_i}^2}{\sigma_X^2}\right)$$

Além de estar restrito ao intervalo de 0 a 1, esse indicador é altamente sensível às correlações entre os itens. Um questionário composto por itens altamente correlacionados entre si terá coeficiente alfa próximo a 1. Se os itens forem pouco correlacionados, o coeficiente alfa deve ficar mais próximo de zero. Na literatura, *valores de alfa iguais ou superiores a 0,7 são considerados aceitáveis, enquanto valores superiores a 0,8 são considerados bons ou ótimos*.

Para testar isso em um exemplo, considere a pesquisa TALIS, que levanta diversas informações relevantes sobre o desenvolvimento profissional de professores nos países participantes. Considere especificamente os itens relacionados à questão 7, sobre as razões que levam os professores a escolherem a carreira docente.

```
library(Hmisc)
load(file = "TALIS.dat")
label(dados)[38:44]
```

Os itens são tipo-Likert e talvez não sejam efetivamente testes paralelos. Na próxima lição, aprenderemos a forma adequada de lidar com a bateria de itens apresentada. Por agora, vou apoiar-me nesse exemplo para mostrar como até mesmo uma avaliação inadequada da fidedignidade pode gerar resultados aceitáveis. Ao mesmo tempo, vou mostrar como o coeficiente alfa de Cronbach pode ser facilmente calculado no ambiente do R.

Para variáveis binárias ou escalares, o coeficiente de fidedignidade pode ser obtido usando a função **`alpha`** da biblioteca **`psych`** (instale-a, se ainda não o fez). Como a função `alpha` está definida em outras bibliotecas que podem estar ativas no seu ambiente, vamos sempre chamá-la informando ao interpretador a biblioteca de origem. Isso pode ser feito

escrevendo **`psych::alpha`**. No entanto, antes de aplicar essa função aos dados disponíveis, é preciso transformá-los em variáveis numéricas. Vamos assumir que as respostas dadas às escalas tipo-Likert possam ser interpretadas como se fossem escalas.

```
for (i in 38:44) dados[, i] = as.numeric(dados[, i])
```

Você deve inserir, na função, as colunas do *dataframe* onde estão os itens que você deseja testar.

```
library(psych)
psych::alpha(dados[, c(38:44)])

##
## Reliability analysis
## Call: psych::alpha(x = dados[, c(38:44)])
##
##   raw_alpha std.alpha G6(smc) average_r S/N    ase mean   sd median_r
##       0.78      0.78    0.84      0.33 3.4 0.0063  3.1 0.58     0.17
##
##  lower alpha upper     95% confidence boundaries
## 0.77 0.78 0.79
##
##  Reliability if an item is dropped:
##         raw_alpha std.alpha G6(smc) average_r S/N alpha se var.r med.r
## TT3G07A      0.72      0.73    0.80      0.31 2.7   0.0082 0.059  0.17
## TT3G07B      0.72      0.73    0.79      0.32 2.8   0.0081 0.054  0.17
## TT3G07C      0.71      0.72    0.78      0.30 2.6   0.0085 0.055  0.17
## TT3G07D      0.76      0.76    0.83      0.34 3.1   0.0071 0.081  0.16
## TT3G07E      0.77      0.76    0.82      0.34 3.1   0.0062 0.071  0.17
## TT3G07F      0.77      0.75    0.81      0.34 3.1   0.0062 0.069  0.18
## TT3G07G      0.78      0.77    0.81      0.36 3.3   0.0063 0.060  0.18
##
```

```
##  Item statistics
##            n raw.r std.r r.cor r.drop mean   sd
## TT3G07A 2753  0.77  0.71  0.68   0.64  2.9 0.98
## TT3G07B 2749  0.76  0.70  0.68   0.63  2.7 0.96
## TT3G07C 2732  0.79  0.74  0.72   0.67  2.8 0.98
## TT3G07D 2746  0.67  0.62  0.50   0.48  2.7 1.07
## TT3G07E 2756  0.53  0.61  0.53   0.37  3.6 0.74
## TT3G07F 2753  0.54  0.63  0.56   0.39  3.5 0.76
## TT3G07G 2768  0.46  0.57  0.50   0.34  3.8 0.57
##
## Non missing response frequency for each item
##            1    2    3    4 miss
## TT3G07A 0.13 0.17 0.40 0.31 0.03
## TT3G07B 0.14 0.22 0.42 0.22 0.03
## TT3G07C 0.13 0.20 0.40 0.27 0.03
## TT3G07D 0.18 0.21 0.31 0.30 0.03
## TT3G07E 0.03 0.05 0.25 0.67 0.03
## TT3G07F 0.03 0.06 0.27 0.64 0.03
## TT3G07G 0.02 0.02 0.15 0.81 0.02
```

O *output* da função **`alpha`** tem muitas informações relevantes:

- o valor do coeficiente alfa e seu intervalo de confiança (95%);
- o alfa que obteríamos se cada um dos itens fosse descartado;
- estatísticas dos itens (ocorrências, correlações, médias e desvios padrão);
- frequências relativas de cada uma das respostas em cada item.

Na análise de fidedignidade, é usual descartar *a posteriori* os itens que estão comprometendo a fidedignidade do instrumento. Isso pode ser feito tomando esse *output* como guia. Ele nos informa como seria o alfa se cada um dos itens fosse descartado. Geralmente optamos pelo descarte de um item quando a) a fidedignidade do instrumento será incrementada e b) sua interpretação não será prejudicada com a retirada do item.

No nosso exemplo, o alfa pode ser considerado aceitável (>0,70). Em princípio, é possível que o analista opte por não eliminar nenhum item da bateria.

Conforme discutimos, o coeficiente alfa pode ser considerado um limite inferior para a fidedignidade real do instrumento, supondo que todos os itens sejam efetivamente testes paralelos. Mas será mesmo que podemos assumir essa premissa?

Restrições do coeficiente alfa

A principal limitação do coeficiente alfa é que ele só produz resultados consistentes quando aplicado a um conjunto *unidimensional* de itens. Em outras palavras, todos os itens da bateria para a qual calculamos o alfa devem medir o mesmo construto, devem ser considerados testes paralelos. Calcular o coeficiente diretamente (i.e., antes de testar a dimensionalidade das respostas) é desaconselhável e pode mascarar inconsistências. De fato, na lição seguinte, perceberemos que a bateria de motivações que testamos há pouco é bidimensional. Se você sabe realizar uma análise de componentes principais (lição 9), pode verificar essa afirmação por conta própria, fazendo:

```
library(FactoMineR)
library(factoextra)

fit = PCA(dados[, c(38:44)])
fviz_screeplot(fit, choice = “eigenvalue”)
```

Além de estar restrito a baterias unidimensionais, o coeficiente alfa apresenta outros problemas. Quando *aumentamos a quantidade de itens*, o alfa também tende a aumentar. Esse aumento é tão expressivo que

questionários com dez ou mais itens provavelmente apresentarão um alfa aceitável (>,70) mesmo quando os itens não são sequer unidimensionais — a exemplo da própria bateria da motivação.

Inicialmente, podemos imaginar que um coeficiente alfa mais elevado é sempre preferível. Porém, a avaliação da fidedignidade não é tão simples assim. Coeficientes alfa muito elevados (acima de 0,90) podem resultar de baterias compostas por itens logicamente idênticos. De fato, as repetições do teste educacional devem ser pensadas com muito cuidado. Os itens da mesma bateria devem ser tão diferentes quanto possível, mas equivalentes naquilo que medem. Somente nessas condições, podemos supor que os itens compartilhem o mesmo valor verdadeiro mantendo seus erros descorrelacionados. Se não pudermos supor que os erros dos itens estejam descorrelacionados, todo o argumento que sustenta o alfa de Cronbach "cai por terra".

Outro problema é que baterias muito autocorrelacionadas podem produzir fatores que não se relacionam bem com as outras variáveis da pesquisa (Mair, 2018). Em geral, uma bateria com alfa muito elevado precisará ser revisada para avaliar se o coeficiente não está sendo artificialmente inflado ou se todos as interpretações possíveis do construto foram operacionalizadas.

Por todas essas razões, nós jamais afirmamos a fidedignidade de um teste educacional olhando somente para o valor do coeficiente alfa.

Revisando a lição

Nessa lição, você aprendeu que:

1. a *validação* de um teste educacional é um processo argumentativo para o qual não há protocolo e que, baseada em muitas fontes de evidência, pretende afirmar a qualidade das conclusões que

podemos produzir na análise, antecipando-nos às críticas que razoavelmente podem ser realizadas pelos nossos pares;

2. o valor efetivamente observado em uma medida educacional, no contexto da *teoria clássica de testes*, pode ser interpretado como a soma de um *valor verdadeiro* e um *erro*;
3. a premissa de que valores observados contêm erros aleatórios implica que *testes educacionais* sejam baseados em observações repetidas do mesmo objeto com o propósito de aumentar a fidedignidade da medida;
4. a presença de erros aleatórios nas medidas educacionais também implica que testes fidedignos aplicados sequencialmente tendam a produzir uma *ilusão de atenuação das diferenças* segundo a qual os participantes com menor pontuação em uma escala são justamente os que apresentam os maiores ganhos aparentes na reaplicação do teste mesmo quando não há ganhos verdadeiros;
5. a fidedignidade de um teste pode ser definida como a razão entre a variância populacional do valor verdadeiro e a variância populacional dos valores observados;
6. o limite inferior da fidedignidade, em baterias de itens *unidimensionais*, pode ser estimado pelo alfa de Cronbach, que pode ser obtido pela função **`psych::alpha`**;
7. a fidedignidade de um teste educacional jamais é avaliada somente a partir do coeficiente *alpha* de Cronbach.

LIÇÃO 14.

ANÁLISE FATORIAL EXPLORATÓRIA

Os usos da análise fatorial exploratória na pesquisa em educação são muito variados. Com poucas exceções, essa técnica costuma estar associada ao processo de *validação de testes educacionais*. Por exemplo, Testa *et al.* (2019), baseados em pesquisas sobre como pensam os estudantes, desenvolveram e validaram um instrumento que mede o progresso de universitários no aprendizado de mecânica quântica. Schumm e Bogner (2016) empregaram a análise fatorial exploratória para investigar as motivações científicas em 232 jovens do início do ensino médio. Os dados foram obtidos por meio de instrumentos já validados. Childers e Jones (2017) empregaram a mesma ferramenta para analisar indicadores de percepção, motivação e identidade científica entre estudantes com relação a uma atividade de microscopia remota. Ho *et al.* (2018) empregaram a análise fatorial exploratória para medir a percepção dos profissionais latino-americanos sobre a qualidade da educação física. Batista *et al*. (2013) desenvolveram e validaram uma escala para identificar como docentes e discentes avaliam a instituição de ensino superior em que estudam e trabalham. Rondini, Teixeira Filho e Toledo (2017) avaliaram a estrutura interna de um questionário sobre as concepções homofóbicas de estudantes de ensino médio. Lima Junior *et al*. (2020) aplicaram a

análise fatorial exploratória para validar um questionário sobre evasão e permanência na educação superior.

Para o analista, a experiência de realizar uma análise fatorial exploratória ou uma análise de componentes principais pode ser, à primeira vista, muito semelhante. Ambas são métodos de redução de dimensionalidade que permitem explicitar padrões de associação entre variáveis. Na literatura, é comum que esses dois métodos sejam confundidos. Isso ocorre, por exemplo, quando as componentes principais são chamadas "fatores" (Meli; Lavidas; Koliopoulos, 2018). No entanto, a análise de componentes principais e a análise fatorial exploratória são diferentes na sua formulação, cômputo e interpretação (Fabrigar; Wegener, 2012). A análise de componentes principais favorece uma relação mais descritiva com os dados. Por isso, ela é mais empregada em análises de questionários socioeconômicos, dados demográficos, mineração de dados (ing. *data mining*) e aprendizagem de máquina (ing. *machine learning*). O que todos esses usos têm em comum? Eles não estão muito preocupados com as propriedades psicométricas do instrumento de medida ou, mais especificamente, com a *estrutura e consistência interna* dos dados. Para esses propósitos, as variáveis do *dataframe* podem estar muito correlacionadas como podem não estar. Isso não importa tanto. A eventual falta de associação entre as variáveis não invalida a análise quando seu caráter é exploratório-descritivo.

Testes educacionais, por outro lado, carregam alguns pressupostos que precisam ser testados. Eles não são medidas tão "livres" quanto dados demográficos e socioeconômicos. Por exemplo, eles costumam ser construídos a partir de uma *antecipação da dimensionalidade* dos seus construtos. Em outras palavras, ao desenvolver um teste, o analista precisa antecipar quantas dimensões serão necessárias para medir aquilo

que pretende medir. Se o seu propósito for desenvolver um teste motivacional, quantas motivações diferentes precisam ser levadas em consideração? Ao desenvolver um teste de competências escolares, quantas competências mais ou menos independentes precisam ser reconhecidas? Se estamos falando de um teste de inteligência ou personalidade, quais tipos de inteligência e personalidade podemos distinguir?

Por razões já discutidas (cf. lição 13), a lógica básica dos testes educacionais é a *repetição*. Sendo assim, e partindo de uma quantidade de dimensões esperada, o analista precisa desenvolver uma bateria de itens para cada dimensão. Ao obter os dados resultantes do teste, sua atitude não será meramente descritiva. Ele precisará testar em que medida as dimensões previstas no desenho do instrumento são confirmadas pelas respostas dadas aos itens. Se os itens de uma mesma bateria não estão mutuamente associados, eles provavelmente medem coisas diferentes (ou não medem coisa alguma!). Em bom português, uma bateria de itens não associados não produz resultados confiáveis.

A análise fatorial exploratória é uma ferramenta particularmente adequada para avaliarmos *a estrutura e a consistência interna* de um teste educacional. Conhecemos a estrutura do teste quando sabemos quantas e quais dimensões são necessárias para explicar as respostas dadas aos seus itens. A consistência interna, por outro lado, está dada na intensidade da associação entre os itens em cada dimensão. Quando a estrutura do teste é simples, podemos delimitar *baterias de itens* conceitualmente semelhantes que apresentam respostas muito associadas.

Os itens tipo-Likert, a seguir, ilustram um teste com estrutura multidimensional. Eles foram desenvolvidos para medir as motivações que levam os estudantes a ingressar no curso de graduação em Física (Lima Junior; Fraga Junior; *et al.*, 2020):

a) Quando eu ingressei em Física, eu queria ser pesquisador;
b) Eu não ingressei no curso de Física para ser cientista;
c) Ingressei no curso de Física para me tornar um pesquisador prestigiado;
d) Ingressei em Física acreditando que a carreira de professor me traria satisfação profissional;
e) Quando eu ingressei em Física, eu tinha interesse em ser professor de Física;
f) Na época do ingresso, eu não tinha interesse em ministrar aulas da educação básica;
g) Minha escolha pelo curso de Física levou em consideração a nota de corte para ingresso no curso;
h) A baixa concorrência foi um elemento importante na minha escolha pelo curso de Física;
i) Quando eu entrei no curso de Física, eu já estava interessado em mudar de curso.

A análise fatorial das respostas dadas aos itens acima mostrou que eles podem ser organizados em três baterias internamente consistentes. Não é por acaso que os itens estatisticamente associados também são conceitualmente relacionados. Os três primeiros itens (A-C) manifestam a *motivação para a pesquisa*. Os três seguintes (D-F), a *motivação para a docência*. Os três últimos (G-I) manifestam *motivações dissonantes*. A análise permitiu identificar três fatores internamente consistentes, mas pouco associados entre si (Lima Junior; Fraga Junior *et al.*, 2020), ou seja, o conhecimento de um fator não permite predizer os outros.

Em síntese, podemos perceber que a ideia básica da análise fatorial é supor que as respostas dadas aos itens são produzidas por *fatores latentes* que desejamos conhecer, mas que não podemos acessar diretamente

(Mair, 2018). Adequada ao estudo da estrutura e consistência interna dos testes educacionais, a análise fatorial exploratória permite "descobrir" quantos e quais fatores latentes são necessários para explicar as respostas dadas aos itens, supondo que esses fatores existam. A existência dos fatores é considerada, portanto, anterior às próprias respostas.

Definições do modelo da análise fatorial

Chamaremos *variáveis manifestas* (ou variáveis medidas, ou atributos de superfície) as informações efetivamente coletadas, observadas (Fabrigar; Wegener, 2012). No caso de testes educacionais, as variáveis manifestas são as respostas dadas aos itens. Na análise fatorial, assumimos que tais variáveis podem ser decompostas em duas parcelas.

- Os **fatores comuns** (também chamados variáveis latentes ou atributos internos) são componentes compartilhadas por vários itens. O modelo de análise fatorial propõe que as correlações observadas entre os itens do teste resultem da presença de fatores comuns entre esses itens. O modelo também assume que a quantidade de fatores deve ser menor que a quantidade de variáveis medidas.
- Os **fatores únicos** são responsáveis pela diferença entre as variáveis medidas e o valor esperado em vista dos fatores comuns. Cada variável medida é afetada por um fator único próprio. Como toda a correlação entre itens é explicada pelos fatores comuns, os fatores únicos são assumidos ortogonais entre si e com relação aos fatores comuns.

Diagrama 14.1: Decomposição das variáveis manifestas em fatores comuns e únicos

Portanto, a *variância observada* pode ser dividida em *variância comum* (explicada pelos fatores comuns) e *variância única* (explicada pelos fatores únicos).

$$\sigma^2_{observada} = \sigma^2_{comum} + \sigma^2_{única}$$

A *comunalidade* é definida como a razão entre a variância comum e a variância observada. Quando o poder explicativo dos fatores sobre os itens é elevado, as comunalidades também serão elevadas.

$$Comunalidade \equiv \frac{\sigma^2_{comum}}{\sigma^2_{observada}}$$

O modelo matemático da análise fatorial exploratória

Para cada variável medida, assumimos que seu valor resulte da soma dos efeitos dos fatores comuns e do seu respectivo fator único. Isso pode ser traduzido no seguinte sistema de equações:

$$\begin{cases} x_1 = \lambda_{11} f_1 + \lambda_{12} f_2 + \cdots + \lambda_{1p} f_p + u_1 \\ x_2 = \lambda_{21} f_1 + \lambda_{22} f_2 + \cdots + \lambda_{2p} f_p + u_2 \\ \cdots \\ x_k = \lambda_{k1} f_1 + \lambda_{k2} f_2 + \cdots + \lambda_{kp} f_p + u_k \end{cases}$$

Nesse sistema:

- x designa as k variáveis manifestas (itens do teste);
- f designa os p fatores comuns (com $p < k$);
- u designa os k fatores únicos (um para cada variável manifesta);
- λ designa as *cargas* fatoriais (i.e., a contribuição de cada fator em cada variável manifesta).

As cargas são muito importantes para interpretarmos os fatores comuns. A situação preferível é que cada fator tenha cargas elevadas em um conjunto restrito de variáveis manifestas. Isso permitirá dar sentido ao fator a partir das variáveis. Por exemplo, num teste de preconceito escolar, se um fator tem carga elevada nos itens de racismo e homofobia (cf. lição 10) e carga baixa nos outros itens, esse fator comum pode ser interpretado como uma medida latente do racismo e da homofobia. Por isso, as cargas fatoriais são muito importantes.

Observe que o sistema de equações anterior pode ser representado de forma mais compacta, empregando a notação da álgebra matricial (cf. apêndice):

$$\vec{x} = \boldsymbol{\Lambda}\vec{f} + \vec{u}$$

Nessa notação, x, $\vec{f}$ e $\vec{u}$ são vetores-coluna e $\boldsymbol{\Lambda}$ é uma matriz das cargas fatoriais. Pela ortogonalidade de fatores comuns e únicos, a matriz de covariância das medidas do teste pode ser decomposta nas seguintes parcelas (Mair, 2018):

$$\boldsymbol{\Sigma}_x = \boldsymbol{\Lambda\Phi\Lambda}^T + \boldsymbol{U}$$

Na equação precedente:

- $\boldsymbol{\Sigma}_x$ é a matriz das covariâncias (ou correlações) das k variáveis manifestas;
- $\boldsymbol{\Phi}$ é a matriz de covariâncias dos p fatores comuns;
- $\boldsymbol{U}$ é a matriz de covariâncias dos fatores únicos;
- $\boldsymbol{\Lambda}$ é a matriz retangular das cargas fatoriais.

Como os fatores únicos são mutuamente independentes, a matriz **U** é diagonal. Assumindo que os fatores comuns sejam normalizados (i.e., desvio padrão unitário) e ortogonais, sua matriz de covariância $\boldsymbol{\Phi}$

é a matriz identidade. Portanto, a equação geral do modelo fatorial pode ser reescrita dessa maneira (Mair, 2018):

$$\boldsymbol{\Sigma}_x - \boldsymbol{U} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}^T$$

A diferença de matrizes à esquerda é chamada *matriz de covariância reduzida*. Ela é igual à matriz de covariância dos itens do teste descontada a contribuição dos fatores únicos.

Há vários métodos para conduzir a análise fatorial exploratória. Os primeiros algoritmos a serem desenvolvidos estavam baseados da aplicação dos teoremas da decomposição espectral (cf. apêndice), tal como vimos na análise de componentes principais. Os métodos mais empregados atualmente são diferentes, mas todos, partindo da equação anterior, visam estimar as cargas fatoriais da matriz $\boldsymbol{\Lambda}$ tendo em vista a matriz de covariâncias (ou correlações) reduzida.

O leitor atento deve perceber que o modelo geral da análise de componentes principais (cf. lição 10) e o modelo geral da análise fatorial exploratória que acabamos de apresentar **não** são equivalentes, havendo três diferenças importantes, representadas no quadro 14.1.

Quadro 14.1: Comparando análise fatorial exploratória e de componentes principais

(continua)

	FATORIAL EXPLORATÓRIA	COMPONENTES PRINCIPAIS
FATORES ÚNICOS	Diagonaliza a matriz de covariâncias reduzida, descontando as contribuições dos fatores únicos.	Diagonaliza a matriz de covariâncias tal como ela é, sem levar em consideração a presença de "fatores únicos".
CARGAS FATORIAIS	Os fatores são carregados nos itens, ou seja, a resposta dada aos itens x_i é composta por fatores f. ($x_i = \lambda_{i1} f_1 + \lambda_{i2} f_2 + \ldots$)	Os itens são carregados nas componentes, ou seja, as componentes principais y_i são compostas pelas respostas dadas aos itens x. ($y_i = a_{i1} x_1 + a_{i2} x_2 + \ldots$)

Quadro 14.1: Comparando análise fatorial exploratória e de componentes principais
(conclusão)

	FATORIAL EXPLORATÓRIA	COMPONENTES PRINCIPAIS
REDUÇÃO DIMENSIONAL	Quando o modelo é gerado, a quantidade p de fatores já foi determinada pelo analista.	Quando o modelo é gerado, a quantidade de componentes retidas ainda não foi determinada.

Portanto, como é possível perceber, a análise fatorial tenta descontar o efeito do ruído na composição dos itens ao assumir a presença de fatores únicos. Além disso, ela faz mais imposições aos dados. Presume, por exemplo, que existam p fatores gerando as variáveis. Em termos práticos, antes de calcular as cargas fatoriais, o analista deverá informar quantos fatores devem ser ajustados. Por exemplo, ao demandar o ajuste de 2 fatores comuns, é exatamente isso que será feito — a despeito da dimensionalidade dos dados. Há, é claro, critérios para avaliar a plausiblilidade das escolhas do analista, mas ela será rigorosamente executada pelo programa. Na análise de componentes principais, basta pressupor que as variáveis manifestas talvez estejam associadas umas às outras e, olhando para as componentes, decidir quantas desejamos reter.

Realizando a análise fatorial exploratória

A função **`fa`** da biblioteca **`psych`** é atualmente a mais completa solução do R para realizar a análise fatorial exploratória. Ela permite que o analista faça diversas escolhas importantes. Sendo assim, antes de executar nossa primeira análise fatorial, vou dedicar algumas páginas para discutir as opções que você terá à sua disposição. Na função **`fa`**, o analista deve especificar:

- a *matriz de covariâncias (ou correlações)* a ser empregada;
- a *quantidade de fatores* que devem ser extraídos;

- o *método de extração* das cargas fatoriais;
- o *método de rotação* dos fatores obtidos;
- o *método de estimativa* dos escores fatoriais.

Obviamente, suas escolhas podem alterar substancialmente o resultado.

Escolhendo a matriz de covariâncias

De maneira semelhante ao que ocorria na análise de componentes principais (cf. lição 10), o primeiro passo da análise fatorial é, na verdade, extrair uma matriz de covariâncias ou correlações que represente as associações entre as variáveis manifestas. Essa extração é realizada internamente pela função `fa` por meio da diretiva **"cor"**. Ao analista, são dadas cinco opções de extração.

- A **matriz de covariâncias** (*cor = "cov"*) é recomendada para analisar um *dataframe* composto por variáveis escalares expressas na mesma unidade de medida. A análise fatorial realizada com a matriz de covariâncias dará mais peso aos itens mais dispersos.
- A **matriz de correlações de Pearson** (*cor = "cor"*) é recomendada para analisar um *dataframe* composto por variáveis escalares codificadas de maneira diversa. Ao padronizar todas as variáveis, eventuais diferenças de codificação são corrigidas e todas os itens terão a mesma dispersão.
- A **matriz de correlações tetracóricas** (*cor = "tet"*) é recomendada para analisar um *dataframe* composto por variáveis binárias.[1]

[1] Mais informações sobre como obter correlações tetracóricas estão disponíveis em Mair (2018).

- A **matriz de correlações policóricas** (*cor = "poly"*) é recomendada para analisar um *dataframe* composto por itens ordinais (e.g., tipo-Likert).[2]
- A **matriz de correlações mista** (*cor = "mix"*) combina as três correlações acima segundo cada par de variáveis.

O padrão da função `fa` é realizar a análise fatorial com a matriz de correlações de Pearson. Ou seja, essa será a escolha automática da função caso o analista não se manifeste sobre o tipo de matriz que deseja extrair.

Determinando a quantidade de fatores

A segunda decisão que precisamos ter em mente, ao realizar uma análise fatorial, é a quantidade de dimensões que desejamos reter. Há várias informações que podemos levar em consideração. Algumas podem ser consultadas antes da análise, outras somente após a finalização do modelo. Como a escolha de dimensões depende, também, da nossa possibilidade de interpretá-las, é usual que o analista produza vários modelos (variando dimensões, método de extração e rotação) até chegar em um resultado que seja, ao mesmo tempo, interpretável e estatisticamente plausível.

A função `fa.parallel` produz *scree plots* capazes de orientar o analista na escolha da quantidade de fatores. Essa função informa:

1. o *scree plot da análise de componentes principais*, obtido pelos autovalores da matriz de covariância (ou correlação) das variáveis manifestas;

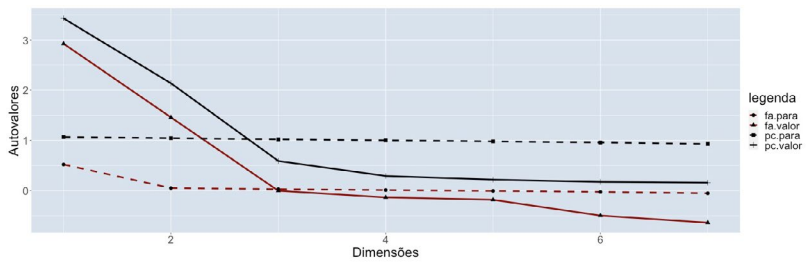
[2] Mais informações sobre como obter correlações policóricas podem ser recuperadas de lições anteriores (cf. Lição 10).

2. o *scree plot da análise fatorial*, obtido pelos autovalores da matriz de covariância (ou correlação) reduzida;
3. o *resultado da análise paralela*, produzido pela comparação das duas informações anteriores com os autovalores de matrizes geradas por método de Monte-Carlo a partir de uma amostra com as mesmas dimensões.

Nesta lição, vamos fazer a análise dos mesmos itens de motivações para a docência que analisamos anteriormente (cf. lição 13). Dado que os itens foram codificados como variáveis categóricas, é preciso transformá-los em variáveis numéricas (isso é feito aplicando a função **`as.numeric`** a cada coluna do *dataframe*). Em seguida, considerando que as variáveis manifestas são respostas a itens tipo-Likert, fazemos **`cor = "poly"`**, que instrui calcular correlações policórica dos itens. Conforme já foi discutido (cf. lição 10), as correlações policóricas são mais adequadas para avaliar os padrões de associação entre variáveis ordinais, em geral, e respostas a itens tipo-Likert, em particular. Para mais informações, faça **`help(fa.parallel)`**:

```
library(Hmisc)
load(file = "TALIS.dat")

library(psych)
for(i in 38:44) dados[,i] = as.numeric(dados[,i])
fa.parallel(dados[, c(38:44)], cor = "poly")
```

Gráfico 14.1: Análise paralela dos itens de motivação



O diagrama da análise paralela (gráfico 14.1) é realmente muito completo. Os dois *scree plots* (linhas contínuas) podem ser interpretados à maneira tradicional, por aplicação dos critérios de Kaiser ou Catell (cf. lição 10). Além disso, há a possibilidade de compará-los com a simulação paralela (linha pontilhada). Ela indica qual seria o *scree plot* de um conjunto de dados completamente descorrelacionado. As dimensões do *scree plot* acima do seu respectivo pontilhado devem ser retidas. No nosso caso, todos os critérios indicam que a bateria de itens é estritamente bidimensional — portanto, o coeficiente alfa de Cronbach que calculamos na lição anterior para os mesmos itens (cf. lição 13) é inconsistente.

Métodos de extração

Na análise fatorial, as cargas são calculadas em duas etapas: extração e rotação. Em cada uma dessas etapas, há várias opções disponíveis e o analista deve fazer escolhas informadas. A *extração* consiste em estimar as cargas a partir da matriz reduzida de correlação (ou covariância). Há vários algoritmos diferentes que permitem fazer essa extração. Alguns são iterativos, outros não. Os tipos mais populares são os seguintes (Fabrigar; Wegener, 2012):

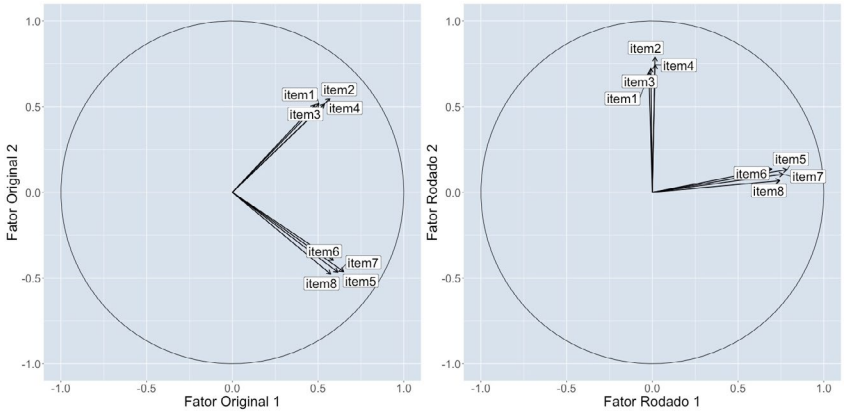
- **análise fatorial principal** tem longa história na análise fatorial e está baseada na extração de autovalores e autovetores da matriz de correlação reduzida, podendo conter iterações ou não.
- **método da máxima verossimilhança** é possivelmente o mais empregado, pois é a única opção de extração implementada na função mais popular da análise fatorial exploratória (`factanal`, da biblioteca `stats`).
- **método dos mínimos resíduos** designa, na verdade, um conjunto de métodos, dos quais o mais tradicional foi implementado como padrão pelos autores da biblioteca `psych`.

Comparados, alguns desses métodos produzirão diferenças muito pequenas nas estatísticas de avaliação da qualidade do ajuste. As cargas extraídas podem ser diferentes na terceira ou quarta casa decimal. Para os propósitos da pesquisa educacional, essa diferença é irrisória! Porém, dependendo da estrutura dos dados e do tamanho da amostra, os resultados com cada método podem ser um pouco diferentes. O método de extração pode ser redefinido pelo usuário empregando a diretiva **`fm`**. A função empregada como padrão costuma ser a mais adequada à maioria das análises. Para mais informações, faça **`help(fa)`**.

Métodos de rotação

A etapa de extração determina as cargas que maximizam o poder explicativo dos fatores sobre os itens segundo os critérios de cada método. Porém, os fatores extraídos são quase sempre desequilibrados. Os primeiros fatores geralmente têm um poder explicativo muito maior e são carregados em muitos itens da bateria, tornando-os inespecíficos, ininterpretáveis. Para que os fatores sejam interpretáveis, é preciso que tenham uma *estrutura simples*. Tal estrutura supõe a possibilidade de agrupar os itens em função dos fatores carregados, ou seja, cada fator deve estar carregado em um conjunto mais ou menos restrito de itens.

Gráficos 14.2-14.3: Exemplo de rotação dos fatores extraídos



O gráfico 14.2 representa dois fatores obtidos após a etapa de extração (i.e., sem rotação). Observe que esses dois fatores têm carga não desprezível em todos os itens. O fator original 1 tem carga sempre positiva. O fator original 2 tem cargas positivas e negativas. Uma situação análoga também ocorre na análise de componentes principais (cf. gráfico 10.4). Porém, se fizermos uma rotação de 45º nos eixos do diagrama, as cargas fatoriais passarão a distinguir os itens em dois grupos bem definidos conforme o gráfico 14.3. O fator rodado 1 tem cargas positivas nos itens de 1 a 4, enquanto o fator rodado 2 possui cargas positivas nos itens de 5 a 8. Assim, esses itens podem ser recuperados para dar sentido aos fatores.

As *rotações* realizadas após a extração, com critérios diferentes, buscam uma estrutura mais simples para as cargas fatoriais. Ao contrário dos diferentes métodos de extração, que podem produzir resultados semelhantes, as rotações tendem a alterar mais dramaticamente as cargas e a interpretação final do modelo. A saber, há dois tipos de rotação:

1. **ortogonais**, que preservam a ortogonalidade postulada para os fatores e devem ser utilizadas se o analista fizer questão que os

fatores sejam interpretados como quantidades descorrelacionadas (seja por razões conceituais ou pela conveniência do tratamento estatístico posterior);

2. **oblíquas**, que não preservam a ortogonalidade dos fatores e, por isso, costumam produzir resultados mais ajustados aos dados.

Das opções de rotação, as ortogonais mais empregadas são *varimax* e *quartimax*. As oblíquas mais conhecidas são *promax* e *oblimin*. Dessas, a rotação oblimin é empregada como padrão na biblioteca **`psych`** — ou seja, se você não especificar nenhuma rotação, a função de análise fatorial aplicará uma rotação *oblimin* antes de publicar os resultados. A propósito, a função **`fa`** permite realizar 16 tipos diferentes de rotação! Todas podem ser especificadas pela diretiva **`rotate`**. Para mais informações, faça **`help(fa)`**.

Interpretando o output da análise fatorial exploratória

Tendo colocado todas as cartas sobre a mesa, proponho fazer uma análise fatorial exploratória das motivações para a docência dos professores brasileiros com as seguintes características:

- correlações policóricas (pois os itens são tipo-Likert);
- dois fatores (conforme indicado pela análise paralela);
- extração pelo método dos mínimos resíduos (padrão da função); e
- rotação oblimin (padrão da função).

Os resultados podem ser visualizados com a função **`print`** a seguir:

```
fit = fa(dados[, c(38:44)], nfactors = 2, cor = “poly”)
print(fit)
```

O *output* apresenta diversas informações. As *cargas fatoriais*, publicadas em primeiro lugar, permitem interpretar o modelo fatorial, dando sentido a cada um dos fatores a partir dos itens que os compõem.

```
## Standardized loadings (pattern matrix) based upon correlation matrix
##           MR1   MR2   h2   u2 com
## TT3G07A  0.87  0.01 0.76 0.24 1.0
## TT3G07B  0.90 -0.04 0.80 0.20 1.0
## TT3G07C  0.93  0.00 0.86 0.14 1.0
## TT3G07D  0.55  0.14 0.35 0.65 1.1
## TT3G07E  0.04  0.81 0.67 0.33 1.0
## TT3G07F  0.04  0.87 0.77 0.23 1.0
## TT3G07G -0.05  0.95 0.88 0.12 1.0
```

Como é possível perceber, o primeiro fator, MR1, está carregado nos itens de A a C (e um pouco no item D). O segundo fator, MR2, está carregado nos itens de E a G. Observando o texto dos itens (com adaptações), podemos dar sentido aos dois fatores:

a) A docência me oferecia uma opção de carreira estável.
b) A docência me dava uma fonte de renda confiável.
c) A docência proporcionava um emprego seguro.
d) O calendário/horário de trabalho se adequava aos compromissos da minha vida pessoal.
e) A docência me permitia influenciar o desenvolvimento de crianças e adolescentes.
f) A docência me permitia ajudar aqueles que são socialmente desfavorecidos.
g) A docência me permitia dar uma contribuição para a sociedade.

Portanto, podemos dizer que o fator MR1 mede as motivações extrínsecas ou individualistas, enquanto o fator MR2 mede as motivações intrínsecas

ou altruístas. Em seguida, o *output* publica informações sobre a *variância explicada* pelo modelo.

```
##                         MR1  MR2
## SS loadings           2.75 2.35
## Proportion Var        0.39 0.34
## Cumulative Var        0.39 0.73
## Proportion Explained  0.54 0.46
## Cumulative Proportion 0.54 1.00
##
##  With factor correlations of
##      MR1  MR2
## MR1 1.00 0.21
## MR2 0.21 1.00
```

Como é possível perceber, nosso modelo bidimensional explica 73% da variância total dos dados, o que pode ser considerado muito bom! A matriz de *correlações dos fatores* mostra que a correlação entre MR1 e MR2 é relativamente fraca, com *r = 0,21*. Em seguida, há várias informações que permitem avaliar a qualidade do ajuste. As mais importantes são o *Tucker-Lewis Index* (TLI) da fidedignidade fatorial e o *Root Mean Square Error of Approximation* (RMSEA).

```
##
## Mean item complexity =  1
## Test of the hypothesis that 2 factors are sufficient.
##
## The degrees of freedom for the null model are  21  and the objective
function was  4.88 0.3 with Chi Square of  13789.95
## The degrees of freedom for the model are 8  and the objective function
was  0.02
##  0.3
## The root mean square of the residuals (RMSR) is  0.01
## The df corrected root mean square of the residuals is  0.01
```

```
##  0.3
## The harmonic number of observations is  2742 with the empirical chi
square  8.54  with prob <  0.38
##  0.3The total number of observations was  2828  with Likelihood Chi
Square =  53.53  with prob <  8.5e-09
##  0.3
## Tucker Lewis Index of factoring reliability =  0.991
## RMSEA index =  0.045  and the 90 % confidence intervals are  0.034 0.057
0.3
## BIC =  -10.05
## Fit based upon off diagonal values = 1
## Measures of factor score adequacy
##                                                  MR1  MR2
## Correlation of (regression) scores with factors   0.97 0.96
## Multiple R square of scores with factors          0.93 0.93
## Minimum correlation of possible factor scores     0.87 0.86
```

Há muita divergência com respeito aos valores aceitáveis das medidas de qualidade do ajuste. Porém, via de regra, os modelos fatoriais são considerados satisfatórios quando apresentam TLI > 0,90 e RMSEA < 0,08 e bons quando TLI > 0,95 e RMSEA < 0,05 (Fabrigar; Wegener, 2012; Mair, 2018). No entanto, é importante lembrar que esses índices não validam automaticamente o modelo fatorial. A validação é um processo que deve empregar vários tipos de evidência (cf. lição 13).

O *Bayesian Information Criteria* (BIC) e o próprio *chi-quadrado* (*chi-square*) são estatísticas dependentes de escala e da dimensão da amostra. Por isso, não há valores de referência, mas elas podem ser empregadas para comparar modelos fatoriais concorrentes.

Extraindo os escores fatoriais

Outra informação importante que resulta da análise é o conjunto dos *valores assumidos pelos fatores* latentes em cada indivíduo. De fato, a essa altura, nós já sabemos que o fator MR1 é uma medida da motivação extrínseca-individualista do professor com relação ao seu ingresso na profissão. Ao mesmo tempo, o fator MR2 pode ser considerado uma medida da sua motivação intrínseca-altruísta. Estimar os valores assumidos por esses fatores em cada indivíduo é fundamental para realizarmos análises posteriores. Esses valores permitiriam, por exemplo, classificar os professores com respeito ao seu grau de altruísmo e individualismo. Poderemos, também, construir modelos de regressão, testando o poder preditivo desses fatores latentes sobre outras variáveis relevantes. Afinal, em que medida os percursos de desenvolvimento profissional estão relacionados aos graus de altruísmo e individualismo identificados na escolha profissional? Todos esses propósitos dependem de podermos salvar, como variáveis novas do *dataframe*, os valores assumidos pelos fatores latentes.

Portanto, é possível que o analista deseje saber qual é o valor assumido por cada fator em cada indivíduo da amostra. Esses valores são chamados *escores fatoriais* e não podem ser obtidos diretamente. Quando executamos a função **`fa`**, todos os escores já foram calculados e salvos na estrutura de dados gerada por essa função (aquela que designamos por "fit"). O método de cômputo dos escores fatoriais deve ser especificado pela diretiva **`scores`** no momento em que executamos a função **`fa`**. Por padrão, a função implementa um método de regressão (**`scores = "regression"`**), mas há outras opções disponíveis. Para mais informações, faça **`help(fa)`**.

Enfim, os escores fatoriais podem ser resgatados por meio de **`fit$scores`**. Você também pode solicitar estatísticas descritivas ao usar a

função **`summary`** ou salvar os escores como colunas novas do seu *dataframe* com a função **`cbind`**:

```
summary(fit$scores)
dados = cbind(dados,fit$scores)
```

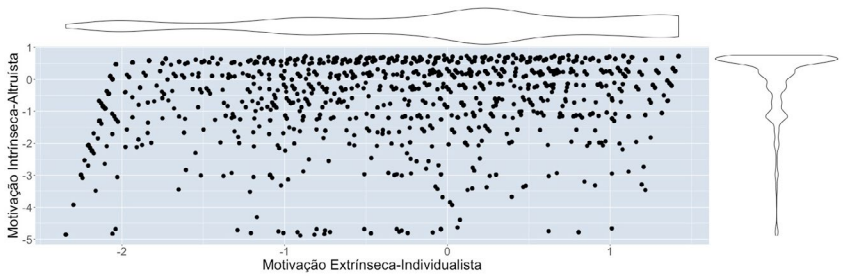
As informações necessárias para avaliar a qualidade dos escores pode ser visualizada ao final do *output* gerado pelo comando **`print(fit)`**:

```
## Measures of factor score adequacy
##                                                 MR1  MR2
## Correlation of (regression) scores with factors   0.97 0.96
## Multiple R square of scores with factors          0.93 0.93
## Minimum correlation of possible factor scores     0.87 0.86
```

Como é possível perceber, os escores fatoriais gerados são confiáveis, pois apresentam alta correlação com seus respectivos fatores. Convencidos da fidedignidade dos escores fatoriais obtidos, podemos gerar um diagrama de dispersão desses dois fatores para ver como eles permitem mapear os indivíduos da amostra:

```
plot(fit$scores)
```

Gráfico 14.4: Diagrama de dispersão dos escores dos fatores latentes



No diagrama gerado (gráfico 14.4), é possível perceber que os escores fatoriais têm distribuições bem distintas. A distribuição das motivações intrínsecas e altruístas (eixo vertical) é claramente assimétrica: a grande maioria dos professores está acumulada na região superior do mapa. Em outras palavras, a maioria dos professores considera ter ingressado na profissão para influenciar o desenvolvimento de crianças e adolescentes, ajudar aqueles que são socialmente desfavorecidos e dar uma contribuição para a sociedade. Essa tendência, observada em uma amostra representativa dos professores brasileiros, indica que a imagem da docência como uma vocação abnegada (Lima Junior, 2018) tem efetiva participação na maneira como esses professores percebem suas escolhas profissionais.

Ao mesmo tempo, as motivações extrínsecas-individualistas estão distribuídas de maneira mais uniforme. Ou seja, adotar uma posição mais utilitária com a profissão (reconhecendo-se motivado por uma carreira estável, uma fonte de renda confiável e um emprego seguro) não concorre com motivações altruístas. Pelo contrário, como vimos no *output* gerado pela função **`print`**, essas duas motivações estão positivamente correlacionadas ($r = 0,21$). Ou seja, se existe alguma associação entre esses dois tipos de motivação, ela tende a ser positiva.

Visualização e interpretação das cargas fatoriais

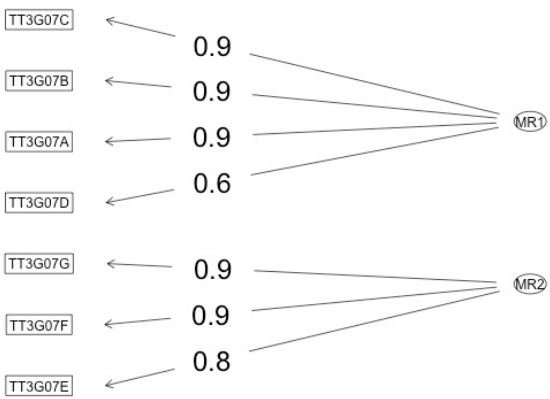
Nem sempre a interpretação dos fatores é tão imediata. Quando a quantidade de variáveis manifestas e de fatores é maior, a atribuição de sentido aos fatores a partir das cargas fatoriais pode ser bem complicada. Sendo assim, é importante conhecermos algumas ferramentas que facilitam a visualização do modelo fatorial. É possível, por exemplo, criar um

diagrama (cf. diagrama 14.2) que relacione os itens (variáveis manifestas) aos fatores comuns de acordo com as cargas. Isso pode ser feito chamando **`fa.diagram`** da biblioteca **`psych`**:

```
fa.diagram(fit)
```

Como já esperávamos, o primeiro fator (MR1) tem cargas elevadas nos itens A, B e C e carga moderada no item D. O segundo fator (MR2) tem cargas fatoriais elevadas nos itens E, F e G.

Diagrama 14.2: Carga fatorial dos itens de motivação

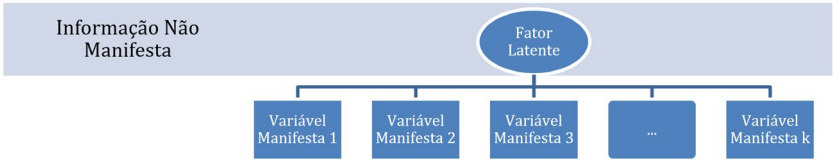


Uma visão geral sobre a análise fatorial exploratória

A ideia básica da análise fatorial é supor que os valores observados (variáveis manifestas) são produzidos por fatores latentes que desejamos conhecer, mas que não podemos acessar diretamente. Assim como um médico consulta um conjunto de indicadores para avaliar a saúde do seu paciente, ou um psicólogo aplica um teste psicométrico para saber se seu cliente sofre de algum transtorno de personalidade, alguns pesquisadores

em educação aplicam testes para levantar informações latentes, não manifestas, com respeito a diversas instâncias individuais (estudantes, professores, gestores, instituições de ensino). Talvez estejamos interessados em avaliar as competências escolares, a relação com o conhecimento, as motivações para estudar, as práticas de ensino, as visões de ciência, a percepção institucional, os interesses profissionais... Todas essas propriedades dos indivíduos são tipicamente inferidas por meio de uma análise fatorial.

Diagrama 14.3: Relação entre fator latente e variáveis manifestas



A plausibilidade de um conjunto de fatores (avaliada pelos índices TLI, RMSEA e outros) está intimamente relacionada à *consistência interna* do instrumento que empregamos para medir tais fatores. Em outras palavras, para que o modelo fatorial resulte plausível, é importante que cada bateria de itens funcione como um conjunto de *testes paralelos*, medindo o mesmo objeto de maneiras diferentes, mas equivalentes (cf. lição 13). A consistência interna de um instrumento aparece na maneira como os fatores são carregados nos itens. Se as baterias forem bem definidas, a análise fatorial exploratória tende a redescobrir o padrão de associações previsto pelo desenvolvedor do instrumento.

Por mais que a análise fatorial exploratória seja um pouco "tateante", é importante lembrarmos que ela costuma ser aplicada a dados que foram necessariamente moldados pela intenção do desenvolvedor do teste. De fato, uma pergunta fundamental que o desenvolvedor do instrumento precisa responder é: quantas dimensões são necessárias para representar

o construto que eu quero medir? Por exemplo, se estamos falando de um teste de inteligência, quantos tipos diferentes de inteligência existem? Se estamos falando de um teste de visões de ciência, quantas visões diferentes podem ser tipificadas? Se pretendemos medir motivações, quantos tipos diferentes de motivação podem ser identificados? A resposta a essas perguntas começa a ser dada no desenvolvimento do teste educacional. A análise não é mais que um passo posterior. Ela é o momento em que a visão do desenvolvedor será confirmada, adaptada ou descartada.

Estou chamando atenção às intenções do desenvolvedor do teste educacional para que o leitor não tenha dúvida de que a estrutura bidimensional limpa que obtivemos nos itens de motivação para a docência do TALIS foi *planejada* para ser assim. A bidimensionalidade identificada não é um acontecimento fortuito. De fato, a literatura está repleta de testes motivacionais organizados em duas baterias (Bøe; Henriksen, 2013; Lee *et al.*, 2016; Mujtaba *et al.*, 2018). De um lado, são colocadas as motivações "intrínsecas", ou seja, motivações que resultam de processos já incorporados cuja origem externa os participantes não são capazes de identificar (e.g., "eu escolhi a carreira científica porque eu gosto"). Essas motivações estão mais relacionadas à ideia de autodeterminação (Glynn *et al.*, 2011). Em oposição a elas, geralmente é possível identificar razões claramente externas que dão uma aparência mais utilitária às escolhas individuais (e.g., "eu escolhi a carreira científica para agradar os meus pais"). Essa segunda classe de motivações é chamada extrínseca e, em geral, não se correlaciona bem com a primeira. Além disso, não é por acaso que os itens de cada bateria estão organizados de maneira ordenada (cf. fator MR1 carregado nos itens A-C e o fator MR2, nos itens E-G). O desenvolvedor sabia muito bem o que estava fazendo! Ao apresentar juntos os itens da mesma bateria, o teste aumenta a correlação e diminui

o ruído das respostas. Tudo isso contribui para que não haja muita dúvida sobre a quantidade de fatores e a interpretação que precisa ser dada a eles.

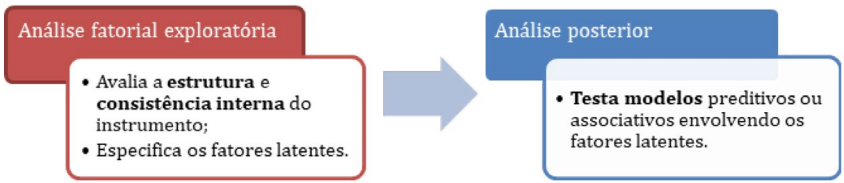
No que concerne à presente lição, muitas ferramentas foram apresentadas. Não será surpreendente se o leitor sentir a necessidade de reler o texto algumas vezes. De maneira geral, vale ressaltar que há uma sequência de perguntas que geralmente precisam ser respondidas em uma análise fatorial exploratória. Tais perguntas foram efetivamente tratadas ao longo no nosso percurso analítico. São elas:

- *Quantos fatores* latentes são necessários para explicar as variáveis manifestas?
    - A resposta pode ser obtida por meio da *análise paralela* (cf. função **`fa.paralel`**).
- Como os fatores podem ser *extraídos*?
    - O analista deve escolher uma matriz de covariância ou correlação, um método de extração e outro de rotação, informando suas escolhas nas diretivas da função **`fa`**.
- Que *sentido* pode ser atribuído a cada um desses fatores?
    - A resposta deve estar baseada na análise das cargas fatoriais (cf. função **`print`**), que são mais facilmente visualizadas no diagrama fatorial (cf. função **`fa.diagram`**).
- Quão *plausível* é o modelo fatorial construído?
    - A resposta está nas medidas de qualidade do ajuste (tais como *TLI* e *RMSEA*).
- Quais são os *valores* assumidos pelos fatores em cada indivíduo?
    - Os escores fatoriais podem ser obtidos fazendo **`fit$scores`**.

Se a análise fatorial exploratória foi realizada com o propósito de avaliar a *estrutura e a consistência interna* de uma medida educacional (cf. lição 13), o trabalho termina por aqui. De fato, vários artigos

não fazem mais que declarar a validade de um processo de medição (Lederman *et al.*, 2002; Özmen; Özdemir, 2019; Tuan; Chin; Shieh, 2005). Contudo, se o analista pretende testar relações preditivas e associativas que envolvem os fatores latentes, talvez seja necessário salvar os escores fatoriais para análise posterior.

Diagrama 14.4: Análise fatorial exploratória empregada como etapa preliminar de uma análise posterior



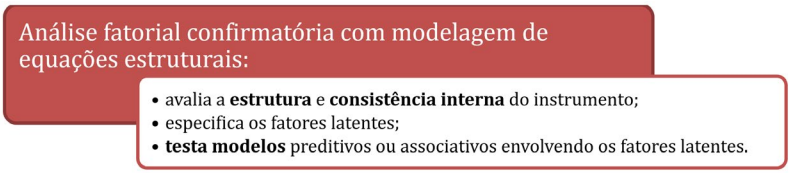
Os escores calculados pela análise fatorial exploratória podem ser empregados com propósitos muito diferentes. Eles podem, por exemplo, gerar diagramas de dispersão (cf. gráfico 14.4), alimentar modelos lineares multivariados (cf. lições 7 a 9) ou mesmo análises de *cluster* (cf. lição 12). Em princípio, todos os métodos que permitem analisar variáveis escalares podem ser empregados após uma análise fatorial. Após validar uma medida das visões de ciência dos estudantes, por exemplo, é possível testar essa medida contra indicadores de religião, sexo, classe, desempenho escolar ou qualquer outra variável que o analista considere um preditor em potencial.

Por outro lado, é importante lembrar que os fatores gerados por uma análise exploratória são mais ou menos flutuantes e o uso posterior dos escores fatoriais requer cautela (Mair, 2018). Dependendo do tamanho da amostra e da quantidade de ruído nos dados, analistas diferentes podem chegar a fatores diferentes fazendo escolhas que, do ponto de vista da análise, são igualmente legítimas (e.g., mudando o método de rotação). Na lição seguinte, conheceremos um método de análise em que os fatores

são mais rigidamente especificados pelo analista e, por isso, não apresentam a mesma arbitrariedade de construção.

A *análise fatorial confirmatória* (cf. lição 15) pertence a uma abordagem mais ampla chamada *modelagem de equações estruturais*. No âmbito da teoria clássica de testes (introduzida na lição 13), a modelagem de equações estruturais é o método multivariado mais empregado para lidar com medidas educacionais sofisticadas. Além de testar a plausibilidade de fatores definidos pelo usuário, a modelagem de equações estruturais permite testar *relações de predição e associação entre os fatores*. Portanto, você não precisará salvar os escores fatoriais para empregá-los em uma análise posterior. O mesmo algoritmo que avalia a consistência interna do instrumento consegue testar relações de predição e associação envolvendo os fatores latentes.

Diagrama 14.5: Análise fatorial confirmatória integrada ao teste de modelos preditivos e associativos envolvendo fatores latentes



O preço que se paga por um método que faz tudo isso? Nós precisaremos abrir mão do caráter exploratório da análise. Em outras palavras, nós daremos preferência à modelagem de equações estruturais quando tivermos uma ideia bem definida de quais itens compõem cada fator e como esses fatores devem se relacionar. Por essa razão, é comum realizar uma análise fatorial exploratória antes de uma análise fatorial confirmatória (Mun *et al.*, 2015). Não que a análise exploratória precise ser confirmada, mas a análise confirmatória requer que o analista tenha uma

boa noção prévia de como os itens de um teste podem ser organizados em baterias internamente consistentes.

Revisando a lição

Em suma, nesta lição, você aprendeu a:

1. avaliar a redução de dimensionalidade por meio da função **`fa.parallel`**;
2. escolher entre métodos de extração e rotação das cargas fatoriais usando diretivas da função **`fa`**;
3. interpretar os fatores a partir de suas cargas fatoriais empregando as funções **`print`** e **`fa.diagram`**;
4. avaliar a plausibilidade do modelo fatorial por meio dos indicadores RMSEA e TLI;
5. salvar os escores fatoriais para análise posterior fazendo **`fit$scores`**;
6. reconhecer as diferenças entre análise fatorial exploratória, confirmatória e análise de componentes principais.

LIÇÃO 15.

ANÁLISE FATORIAL CONFIRMATÓRIA E MODELAGEM *SEM*

Quando nós desenvolvemos e aplicamos um teste educacional pela primeira vez, o mais indicado é empregar a análise fatorial exploratória (cf. lição 14) para aprimorar o teste. Ela permitirá avaliar, com base nas respostas dadas aos itens, quantos e quais fatores são necessários para explicar as respostas dadas pelos participantes da pesquisa. A análise fatorial exploratória também poderá nos ajudar a identificar os itens "fracos", que precisam ser removidos do teste ou desconsiderados. De fato, essa ferramenta analítica é provavelmente a mais empregada para desenvolver, aprimorar e validar medidas educacionais.

No entanto, quando estamos aplicando um teste já aprimorado, nossa atitude frente aos dados produzidos não é mais exploratória. Por exemplo, se os itens do teste apresentam uma resposta estável, o analista deve ser capaz de antecipar quais fatores serão carregados em cada item antes mesmo da aplicação! Com efeito, nos casos em que a estrutura do modelo fatorial é conhecida, todo o aparato exploratório discutido na lição anterior torna-se desnecessário. Nessas situações, a análise fatorial exploratória dará respostas a perguntas que não estamos realmente nos fazendo, e.g., quantos e quais fatores latentes são necessários para explicar as respostas dadas aos itens?

Nas situações em que o teste educacional já foi aplicado e aprimorado algumas vezes, é mais conveniente empregar uma versão não exploratória da análise fatorial, um método que permita testar a plausibilidade de uma estrutura fatorial especificada pelo usuário. Esse método é chamado *análise fatorial confirmatória*. Ao realizar uma análise confirmatória, é importante que o pesquisador saiba delimitar os itens que compõem cada fator. Portanto, é usual realizar uma análise exploratória antes da análise confirmatória — não porque a análise exploratória precise ser confirmada (ela não precisa), mas porque entrar na análise confirmatória com dúvidas sobre a delimitação mais plausível das baterias de itens pode gerar muitas dificuldades ao longo do caminho.

A principal vantagem da análise fatorial confirmatória é que ela pertence a um método mais amplo chamado *modelagem de equações estruturais*, que permite, entre muitas outras coisas, testar relações preditivas e associativas entre os fatores latentes (Mair, 2018). Para imaginarmos como isso funciona na prática, tente formular uma questão de pesquisa envolvendo propriedades psicoeducacionais cuja relação você deseja testar. É importante que algumas dessas propriedades só possam ser avaliadas por meio de testes educacionais (e.g., competências, motivações, interesses, atitudes, crenças, valores, percepções...).

Por exemplo, imagine que desejamos investigar a relação entre *i)* desempenho escolar e *ii)* preferências profissionais. Para dar conta dessa questão, precisamos construir medidas independentes desses dois construtos. Como vimos, não precisamos presumir que eles sejam unidimensionais. Ou seja, é possível modelarmos o "desempenho escolar" e a "escolha profissional" como um conjunto de fatores diferentes, relacionados ou ortogonais (e.g., desempenho em matemática, em linguagens e códigos... preferência por carreiras em artes e humanidades, carreiras

científicas e tecnológicas, carreiras relacionadas à saúde à manutenção do bem-estar). Nosso primeiro passo seria, portanto, construir um teste educacional que permita cobrir todas as dimensões de cada construto, dedicando vários itens a cada dimensão.

Na etapa de aprimoramento do teste, a análise fatorial exploratória deve ser capaz de indicar os itens mais "fracos", i.e., aqueles que não manifestam muito bem os fatores latentes. Além disso, ela deve ajudar a demarcar com mais segurança quais itens formam baterias mais ou menos unidimensionais, ou seja, quais grupos de itens correspondem a cada fator latente. Uma vez que o teste atinja uma *estrutura interna estável*, ou seja, quando for possível presumir que suas reaplicações produzirão aproximadamente os mesmos padrões de associação entre os itens, faz sentido realizarmos uma análise que permita testar a plausibilidade de um modelo fatorial conhecido e, ao mesmo tempo, testar as relações entre os fatores de desempenho escolar e preferência profissional. É nesse momento que recorremos à modelagem de equações estruturais.

Na etapa confirmatória da análise fatorial, é possível construir modelos bastante sofisticados que, além de testar a plausibilidade das baterias de itens, são capazes de testar vários tipos de relação entre os fatores latentes. Todas as possibilidades discutidas no âmbito do modelo linear clássico podem ser retomadas aqui.

Por exemplo, podemos avaliar o efeito de A sobre B, *controlando os efeitos* de C (cf. lição 7). Também é possível construir modelos que incluam *efeitos de interação* (cf. lição 7). Além disso, nossos modelos podem incluir relações de *mediação* (que serão discutidas a seguir). Assim como em outras técnicas de modelagem estatística, é possível comparar modelos concorrentes (cf. lição 8), optando por aqueles que parecem mais plausíveis.

O emprego da modelagem de equações estruturais na pesquisa educacional é muito diverso. Por exemplo, visando avaliar as relações entre habilidades científicas, criatividade científica e variáveis demográficas (grau, idade e sexo), Dikici, Özdemir e Clark (2018) realizaram uma modelagem de equações estruturais com dados de 353 estudantes da educação secundária na Turquia. A análise permitiu testar três modelos diferentes. No modelo mais bem-ajustado aos dados, as variáveis demográficas eram capazes de prever a criatividade científica, direta e indiretamente, por meio de habilidades de processo científico. Em outras palavras, as habilidades do processo científico mediam a relação entre sexo, idade e criatividade científica (relações de mediação serão discutidas posteriormente).

Em uma análise sobre as relações entre crenças epistêmicas, concepções e motivações para aprender ciências, Ho e Liang (2015) realizaram uma modelagem de equações estruturais das respostas dadas por 470 estudantes do ensino médio em Taiwan. Crenças epistêmicas absolutistas dos estudantes mostraram-se relacionadas ao aprendizado baseado em memorização e ao estudo orientado ao desempenho nas provas. Por outro lado, crenças epistêmicas mais sofisticadas estavam relacionadas a aprender ciências como aumento de conhecimento, à aplicação do conhecimento em situações concretas e à compreensão profunda.

Nesta lição, vou mostrar como é possível realizar uma análise fatorial confirmatória e uma modelagem de equações estruturais empregando funções da biblioteca **`lavaan`** e **`semPlot`** (instale-as, se ainda não o fez).

O modelo matemático da análise fatorial confirmatória

A formulação inicial dos modelos de análise fatorial exploratória e confirmatória é idêntica (cf. lição 14). Nos dois modelos, as variáveis

manifestas são consideradas resultado da composição de efeitos fatoriais com efeitos únicos. Empregando notação matricial, essa composição pode ser representada pela seguinte equação:

$$\vec{x} = \boldsymbol{\Lambda}\vec{f} + \vec{u}$$

Nela, $\vec{x}$, $\vec{f}$ e $\vec{u}$ são vetores-coluna que representam as variáveis manifestas os fatores comuns e os fatores únicos, respectivamente. A matriz $\boldsymbol{\Lambda}$ representa as cargas fatoriais. Pela ortogonalidade dos fatores únicos, a matriz de covariância das variáveis manifestas pode ser decomposta nas seguintes parcelas (Mair, 2018):

$$\boldsymbol{\Sigma}_x = \boldsymbol{\Lambda\Phi\Lambda}^T + \boldsymbol{U}$$

Nessa equação,

- $\boldsymbol{\Sigma}_x$ é a matriz das covariâncias (ou correlações) das k variáveis manifestas;
- $\boldsymbol{\Phi}$ é a matriz de covariâncias dos p fatores comuns;
- $\boldsymbol{U}$ é a matriz de covariâncias dos fatores únicos;
- $\boldsymbol{\Lambda}$ é a matriz retangular das cargas fatoriais.

Diferentemente da análise exploratória (cf. lição 14), os fatores definidos pelo usuário na análise confirmatória não são necessariamente ortogonais. Portanto, a matriz das covariâncias não será uma matriz identidade. Por outro lado, ao especificar os itens nos quais cada fator deve ser carregado, o analista impõe, implicitamente, que algumas entradas da matriz das cargas fatoriais sejam iguais a zero. Esses vínculos permitem calcular as cargas fatoriais mesmo sem impor ortogonalidade aos fatores.

Um modelo fatorial simples

Como primeiro exemplo, proponho fazermos uma análise confirmatória dos mesmos itens de motivação para a docência que foram analisados na lição anterior (cf. lição 14). A análise fatorial exploratória desses itens apontou para a existência de dois fatores latentes que podem ser chamados: *i)* motivação individualista e *ii)* motivação altruísta. Comece carregando os dados, recodificando e renomeando as variáveis:

```
library(lavaan)
library(semPlot)

load(file = "TALIS.dat")
for (i in 38:44) dados[, i] = as.numeric(dados[, i])
colnames(dados)[38:44] =
      c("itm7A", "itm7B", "itm7C", "itm7D", "itm7E", "itm7F", "itm7G")
```

Nosso primeiro passo é, portanto, especificar quais itens compõem cada bateria. Isso é feito empregando uma sintaxe muito semelhante à dos modelos lineares (cf. lição 7). Os caracteres "=~" devem ser lidos como "é manifesto por" e são empregados para definir a composição dos fatores latentes em função das variáveis manifestas.

```
modelo = "Indiv =~ itm7A + itm7B + itm7C + itm7D
          Altru =~ itm7E + itm7F + itm7G"
```

No modelo precedente, Indiv e Altru são os nomes dos fatores latentes. A análise fatorial confirmatória pode ser feita com a função **`cfa`** da biblioteca **`lavaan`**. Essa função deve ser alimentada com o modelo fatorial já definido, o nome do *dataframe* e outras informações. A função **`cfa`** também suporta correlações policóricas e tetracóricas. A diretiva **`ordered`**

informa à função quais variáveis do *dataframe* devem ser interpretadas como ordinais. No nosso caso, como todas as variáveis do *dataframe* são tipo-Likert, fazemos **`ordered=colnames(dados)`**. O resultado da análise é armazenado em uma estrutura de dados nomeada pelo analista.

```
fit = cfa(modelo, data = dados, ordered = colnames(dados))
```

Em seguida, empregamos a função **`semPaths`** para gerar um diagrama de caminho (ing., *path diagram*), representação gráfica típica das análises fatoriais confirmatórias e modelos de equações estruturais.

```
semPaths(fit)
```

Diagramas de caminho (cf. diagrama 15.1) são compostos por vértices (ing. *nodes*) e arestas (ing. *edges*). Os vértices representam:

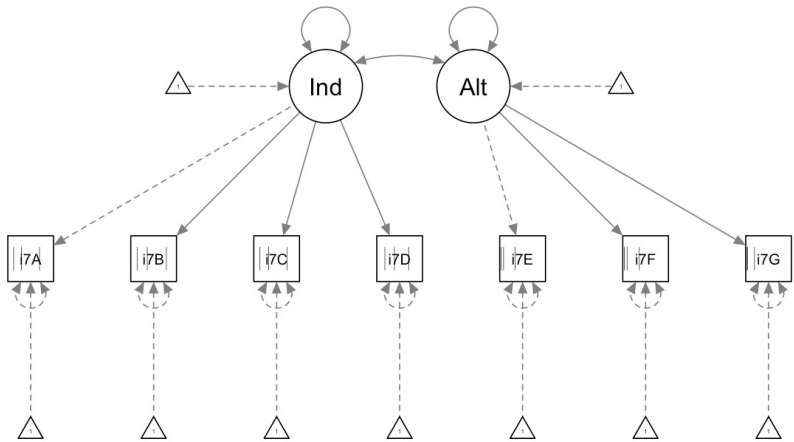
- **variáveis manifestas** (quadrados);
- **variáveis latentes** (círculos); ou
- **constantes** (triângulos).

As arestas, por sua vez, podem representar:

- **relações de predição** (flechas unidirecionais entre vértices diferentes);
- **relações de associação** (flechas bidirecionais entre vértices diferentes);
- **resíduos** (flechas bidirecionais de um vértice contra ele mesmo).

Todos esses elementos podem ser visualizados no diagrama 15.1. Além deles, o leitor deve perceber que algumas arestas são contínuas, já outras são tracejadas. Além disso, no interior dos quadrados, há pequenos traços verticais. Eles registram os limiares (ing. *thresholds*) gerados quando calculamos correlações policóricas (cf. lição 10).

Diagrama 15.1: Diagrama de caminho do modelo fatorial das motivações para o curso de Física



Como bem sabemos, no modelo fatorial, as variáveis latentes (círculos) determinam os valores assumidos pelas variáveis manifestas (quadrados). Por isso, há setas saindo dos círculos para os quadrados (e não o contrário). Como é típico da análise fatorial confirmatória, os fatores latentes são livres para se correlacionar. A seta bidirecional entre os fatores de motivação individualista (**Ind**) e altruísta (**Alt**) indica que a correlação entre esses fatores latentes foi calculada (podendo ser significativa, ou não). Outra coisa que observamos no diagrama de caminho é a presença de um triângulo atribuído a cada variável do modelo, seja ela latente ou manifesta. Esses triângulos representam termos constantes do ajuste (ing., *intercepts*) e, se as variáveis do modelo foram todas padronizadas, essas constantes devem ser identicamente nulas. Por isso, é comum omitir os triângulos do diagrama.

Edição do diagrama de caminho

A função **`semPaths`** permite gerar diagramas de caminho com alta qualidade, controlando diversos parâmetros gráficos (cf. Epskamp, 2015). Por isso, é importante darmos um pouco de atenção às diretivas dessa função. No quadro 15.1, destaco as opções de edição dos diagramas que tendem a ser mais úteis.

A lista completa de diretivas da função **`semPaths`** pode ser obtida fazendo **`help(semPaths)`**. Alguns exemplos mais importantes são apresentados no artigo de Epskamp (2015). Sugiro que o leitor tente todas as opções listadas no quadro 15.1 para ganhar um pouco de familiaridade com a função e ver, na prática, como ela permite controlar o **`output`** gráfico.

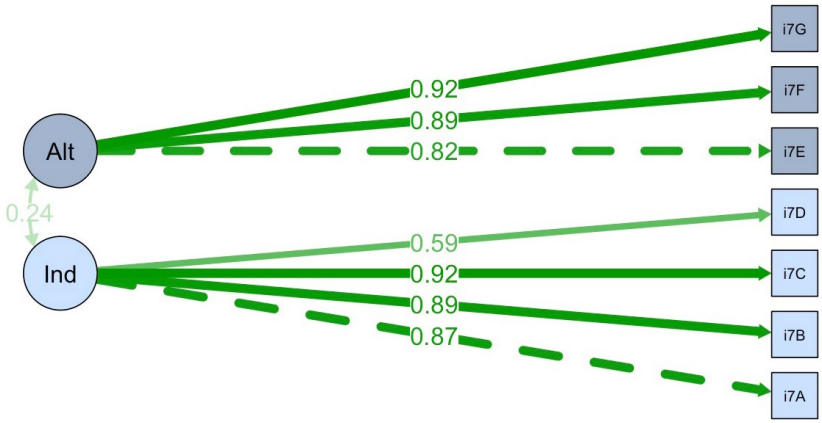
Quadro 15.1: Algumas diretivas da função semPaths

Diretiva	*O que ela faz?*
`what`	*String*. Determina o que será impresso nas arestas do diagrama. Por padrão, todas as arestas são representadas por linhas contínuas. A primeira variável manifesta de cada bateria é conectada ao seu respectivo fator latente por uma linha tracejada. Para que o traço e a espessura das arestas representem os valores dos parâmetros, o programador pode fazer **`what="std"`** (empregando estimativas padronizadas) ou **`what="par"`** (para estimativas não padronizadas).
`whatLabels`	*String*. Determina os rótulos das arestas do diagrama. Por padrão, nenhum rótulo é impresso, mas o analista pode fazer **`whatLabels="par"`** ou **`whatLabels="std"`** para imprimir as estimativas não padronizadas e padronizadas dos parâmetros, respectivamente. A saber, nas estimativas padronizadas, a variância dos fatores latentes é fixa em 1 unidade. Nas estimativas não padronizadas, a variância dos fatores latentes pode assumir qualquer valor. Em compensação, a carga do fator na primeira variável manifesta (aquela registrada com uma linha pontilhada) é fixa em 1 unidade. Portanto, escolher entre representações padronizadas ou não padronizadas equivale a escolher quais parâmetros serão obrigatoriamente unitários.

Diretiva	*O que ela faz?*
`layout`	*String*. Determina como os vértices são dispostos. As principais opções de disposição são em formato de árvore (**`layout="tree"`** ou **`layout="tree2"`**), círculo (**`layout="circle"`** ou **`layout="circle2"`**) e fonte (**`layout="spring"`**)
`intercepts`	*Lógico*. Informa se os termos constantes (vértices triangulares) devem ser impressos.
`residuals`	*Lógico*. Informa se os resíduos devem ser impressos.
`thresholds`	*Lógico*. Informa se os liminares produzidos pelas correlações tetrapolicóricas (cf. lição 10) serão impressos.
`rotation`	*Inteiro*. Aplica rotações retas ao diagrama.
`curvature`	*Escalar*. Quando há mais de dois vértices no mesmo nível, suas ligações podem começar a se sobrepor, problema que pode ser evitado com a alteração do parâmetro de curvatura.
`groups`	*String*. Determina como agrupar os vértices para colori-los. Atribuir a *string* "*lat*" agrupa os vértices em função das variáveis latentes. A *string* "*manifests*" agrupa em função das variáveis manifestas.
`color`	*Vetor*. Especifica quais cores devem ser empregadas para o agrupamento da diretiva *group*.
`structural`	*Lógico*. Se verdadeiro, somente relações estruturais são publicadas.
`label.cex`	*Escalar*. Controla o tamanho dos rótulos nos vértices.
`edge.label.cex`	*Escalar*. Controla o tamanho dos rótulos nas arestas.

Empregando essas diretivas, podemos gerar uma versão mais amigável do diagrama de caminho (cf. diagrama 15.2). Nela, foram publicados valores padronizados. Isso pode ser percebido porque os valores registrados nas linhas pontilhadas não são unitários. Além disso, a cor e espessura das setas reforçam a percepção de quais parâmetros são mais maiores e menores. À semelhança do que vimos na análise exploratória, a carga fatorial no item 07D é menor que as demais, sugerindo que esse item poderá ser desconsiderado na composição do fator de motivações individualistas. Além disso, à semelhança do que ocorreu na análise exploratória (cf. lição 14), a correlação entre os dois fatores latentes é baixa, mas positiva ($r = 0,24$).

Diagrama 15.2: Diagrama de caminho do modelo fatorial das motivações para o curso de Física (editado)



Significância estatística dos parâmetros

Não é por estarem representados no diagrama de caminho que todos os parâmetros são estatisticamente significativos. Dependendo do tamanho da amostra, alguns valores podem ser atribuídos ao acaso. Por exemplo, conforme vimos anteriormente (cf. lição 5), uma correlação $r = 0,24$ pode representar um efeito casual. Portanto, para saber quais valores do diagrama de caminho atingem o nível de significância estatística desejado (com $p < 0,05$), podemos empregar a função **`parameterEstimates`** com a diretiva **`standardized = TRUE`**.

```
parameterEstimates(fit, ci = FALSE, standardized = TRUE)
```

Como é possível perceber, o *output* dessa função apresenta todos os parâmetros do modelo. Alguns foram definidos pelo usuário, mas a maioria é definida internamente. Copio a seguir os parâmetros que precisam ser analisados agora.

```
##      lhs  op   rhs   est    se       z pvalue std.lv std.all std.nox
## 1  Indiv  =~ itm7A 1.000 0.000      NA      0.874   0.874   0.874
## 2  Indiv  =~ itm7B 1.022 0.008 124.147      0  0.893   0.893   0.893
## 3  Indiv  =~ itm7C 1.057 0.008 132.104      0  0.924   0.924   0.924
## 4  Indiv  =~ itm7D 0.673 0.016  42.509      0  0.588   0.588   0.588
## 5  Altru  =~ itm7E 1.000 0.000      NA      0.823   0.823   0.823
## 6  Altru  =~ itm7F 1.080 0.016  66.277      0  0.889   0.889   0.889
## 7  Altru  =~ itm7G 1.123 0.018  62.539      0  0.923   0.923   0.923
## 36 Indiv  ~~ Indiv 0.764 0.010  73.923      0  1.000   1.000   1.000
## 37 Altru  ~~ Altru 0.677 0.018  38.355      0  1.000   1.000   1.000
## 38 Indiv  ~~ Altru 0.175 0.017  10.432      0  0.243   0.243   0.243
```

Conforme já sabemos, os caracteres "=~" devem ser lidos como "é manifesto por" e relacionam fatores latentes às suas variáveis manifestas. Além disso, saiba que os caracteres "~~" indicam covariância. Lembre-se, também, que a covariância de uma variável consigo mesma é igual à sua variância (cf. lição 5).

A coluna **est** informa as estimativas não padronizadas de cada parâmetro (ou seja, as estimativas obtidas quando a carga fatorial na primeira variável manifesta é unitária). Para cada parâmetro, é estimado um erro padrão **se** e uma estatística-**z**, que, à semelhança da estatística-**t** (cf. lição 4), resulta da razão entre o parâmetro e seu respectivo erro padrão. A estatística-z permite calcular o **valor-p** (i.e., a probabilidade de ser verdadeira a hipótese nula desse parâmetro). Quando *pvalue < 0,05*, o parâmetro pode ser considerado estatisticamente significativo.

Em todas as linhas copiadas, os parâmetros do modelo resultaram estatisticamente significativos. Dado o tamanho da amostra, esse resultado não surpreende. Por outro lado, observe que os valores na coluna "est" não coincidem com os valores no diagrama de caminho da seção anterior (diagrama 15.2). Isso ocorre porque aquele diagrama publicou

valores padronizados. Alternativamente, os valores padronizados podem ser conferidos nas colunas **std**. A padronização não altera a significância estatística dos parâmetros.

Avaliação da plausibilidade do modelo

Os indicadores de qualidade do ajuste empregados para avaliar a plausibilidade do modelo fatorial podem ser obtidos por meio da função **`summary`** acrescentando a diretiva **`fit.measures = TRUE`**.

```
summary(fit, fit.measures = TRUE)
```

Na literatura, há diversos valores de referência para avaliar a plausibilidade de modelos fatoriais. Mesmo sendo sempre questionados, esses valores funcionam como um tipo de "regra do dedão" (ing., *thumb rule*), ou seja, eles podem não ser suficientes para remover todas as suspeitas sobre a plausibilidade do modelo fatorial, mas efetivamente orientam as decisões do analista. A saber, os valores recomendados para a análise fatorial confirmatória e a modelagem de equações estruturais são os seguintes (Mair, 2018).

- *Comparative Fit Index,* $CFI \geq 0,95$.
- *Tucker-Lewis Index,* $TLI \geq 0,95$.
- *Root Mean Square Error of Approximation,* $RMSEA \leq 0,05$.
- *RMSEA upper bound* $\leq 0,10$.
- *Standardized Root Mean Square Residual,* $SRMR \leq 0,08$.

Algumas referências consideram aceitável obter $RMSEA \leq 0,08$ ou $TLI \geq 0,90$. De qualquer maneira, é importante saber que esses valores de corte não se justificam com base em princípios universais, mas são recomendações práticas aproximadas.

```
     ## User Model versus Baseline Model:
##
##   Comparative Fit Index (CFI)                    0.999       0.996
##   Tucker-Lewis Index (TLI)                       0.998       0.994
##
##   Robust Comparative Fit Index (CFI)                            NA
##   Robust Tucker-Lewis Index (TLI)                               NA
##
## Root Mean Square Error of Approximation:
##
##   RMSEA                                          0.047       0.057
##   90 Percent confidence interval - lower         0.039       0.048
##   90 Percent confidence interval - upper         0.057       0.066
##   P-value RMSEA <= 0.05                          0.660       0.085
##
##   Robust RMSEA                                                  NA
##   90 Percent confidence interval - lower                        NA
##   90 Percent confidence interval - upper                        NA
##
## Standardized Root Mean Square Residual:
##
##   SRMR                                           0.045       0.045
```

Como é possível perceber, nosso modelo satisfaz todos os padrões de qualidade. Há, contudo, alguma dúvida quanto ao valor-p da hipótese $RMSEA \leq 0,05$. Ou seja, ainda que o RMSEA tenha ficado dentro dos intervalos previstos, a probabilidade de que ele se mantenha nessa região em aplicações posteriores do mesmo teste não é superior a 95% como gostaríamos.

Não há um procedimento-padrão para melhorar a plausibilidade de um modelo fatorial. Se a distância aos padrões de qualidade do ajuste for muito grande, eu recomendaria retornar à análise fatorial exploratória (cf. lição 14). Ela poderá informar, por exemplo, se o problema está na definição das baterias de itens ou se os dados são “ruins” mesmo. De fato,

se a análise fatorial exploratória não produzir um modelo plausível e interpretável após a remoção dos itens mais discrepantes, podemos concluir que a estrutura e consistência interna dos dados está prejudicada. Para esse problema, geralmente não há solução.

Por outro lado, se a distância aos padrões de qualidade não for muito grande (como no nosso exemplo), talvez um modelo mais aceitável possa ser produzido fazendo pequenas alterações no modelo atual. Nas situações em que o modelo fatorial for muito complexo, o analista pode ter dificuldade em identificar qual modificação precisa ser feita para tornar o modelo mais plausível. Para auxiliar essa tomada de decisão, é possível empregar a função **`modindices`** da biblioteca **`lavaan`**.

```
modindices(fit, sort. = TRUE)
```

Essa função publica qual seria a modificação **`mi`** na estatística chi-quadrado caso um vínculo do modelo fosse removido. A remoção de um vínculo corresponde a acrescentar parâmetros ajustáveis ao modelo. Por exemplo, quando geramos um modelo fatorial, nós assumimos, por padrão, que os resíduos das variáveis manifestas são descorrelacionados. No entanto, é possível que alguns itens do questionário estejam associados sem que essa associação esteja representada nos fatores latentes. Situações como essa prejudicam a plausibilidade do modelo como um todo. Nesses casos, pode ser interessante permitir que os resíduos das variáveis manifestas se correlacionem. Decisões dessa natureza podem ser tomadas observando o *output* da função **`modindices`** a seguir:

```
##      lhs op   rhs     mi    epc sepc.lv sepc.all sepc.nox
## 61 Altru =~ itm7D 61.703  0.179   0.147    0.147    0.147
## 77 itm7D ~~ itm7E 34.241  0.141   0.141    0.306    0.306
## 78 itm7D ~~ itm7F 22.578  0.113   0.113    0.305    0.305
```

```
## 59 Altru =~ itm7B 19.696 -0.113  -0.093   -0.093   -0.093
## 57 Indiv =~ itm7G 18.726 -0.111  -0.097   -0.097   -0.097
## 80 itm7E ~~ itm7F 18.726 -0.297  -0.297   -1.140   -1.140
## 72 itm7B ~~ itm7G 16.209 -0.121  -0.121   -0.698   -0.698
## 62 itm7A ~~ itm7B  6.134  0.066   0.066    0.300    0.300
## 55 Indiv =~ itm7E  5.294  0.050   0.044    0.044    0.044
```

Observando os valores da coluna **`mi`**, percebemos que as três recomendações de modificação que mais contribuem para aumentar a plausibilidade do modelo fatorial envolvem o item 7D — justamente aquele que tem carga fatorial relativamente mais baixa que os demais (cf. diagrama 15.2). As sugestões feitas pelo *output* mostrado não fazem muito sentido do ponto de vista conceitual (e.g., ele recomenda incluir o item 7D no fator de motivações altruístas). Por outro lado, essas sugestões indicam que o item 7D está com problemas. Portanto, proponho eliminá-lo.

```
modelo = “Indiv =~ itm7A + itm7B + itm7C
          Altru =~ itm7E + itm7F + itm7G”

fit = cfa(modelo, data = dados, ordered = colnames(dados))
summary(fit, fit.measures = TRUE)
```

Com essa notificação, o modelo fatorial (que já poderia ser considerado plausível) atinge valor-p da hipótese RMSEA < 0,05 superior a 95%. Portanto, é altamente provável que, em aplicações futuras, o modelo fatorial produza resultados consistentes.

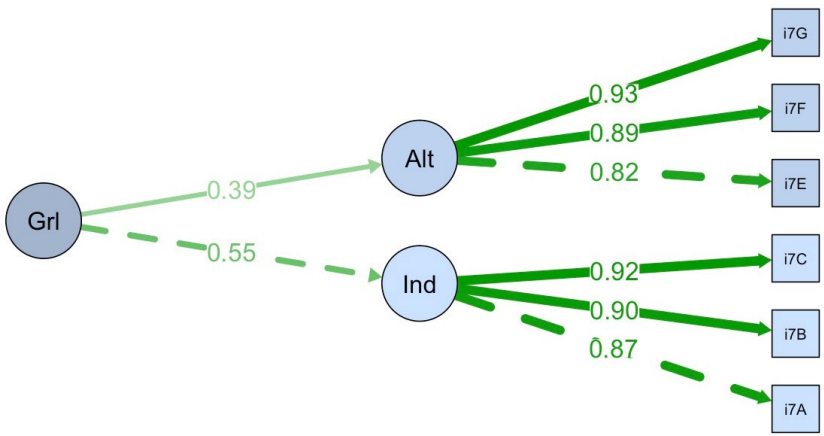
```
## Root Mean Square Error of Approximation:
##
##   RMSEA                                          0.025       0.037
##   90 Percent confidence interval - lower         0.013       0.025
##   90 Percent confidence interval - upper         0.038       0.049
##   P-value RMSEA <= 0.05                          0.999       0.965
```

Modelo fatorial hierárquico

Outra possibilidade interessante da análise fatorial confirmatória é a construção de modelos multinível em que uma variável latente é manifesta por outras variáveis latentes (Mair, 2018). No modelo de motivações para a docência, dado que as motivações individualista e altruísta estão correlacionadas, faz sentido supor que elas sejam explicadas por uma "motivação geral". Tal motivação, anterior às outras duas, pode ser especificada conforme o modelo seguinte. O resultado é plausível e pode ser visualizado no diagrama 15.3.

```
modelo = "Indiv =~ itm7A + itm7B + itm7C
          Altru =~ itm7E + itm7F + itm7G
          Geral =~ Indiv + Altru"

fit = cfa(modelo, data = dados, ordered = colnames(dados))
```

Diagrama 15.3: Exemplo de modelo fatorial hierárquico

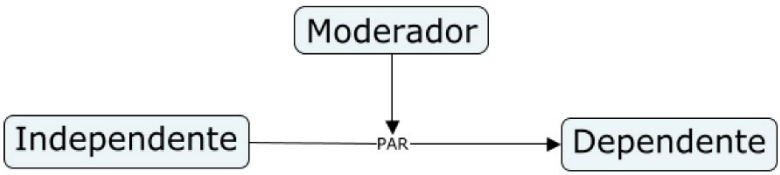


Relações de Moderação e Mediação

Até o presente momento, elaboramos modelos estritamente fatoriais baseados em relações do tipo "é manifesto por". A transição para a modelagem com equações estruturais ocorre quando começamos a acrescentar relações preditivas que envolvam as variáveis manifestas e latentes do modelo. Vários tipos de relação podem ser testados e a maneira de construção dos modelos incorpora todas as considerações que fizemos no âmbito da modelagem linear multivariada. Por exemplo, é possível construir modelos com interação (cf. lição 7). Em lições anteriores, efeitos de interação foram comparados à noção de interseccionalidade desenvolvida pelo feminismo negro, segundo a qual as condições de existência das mulheres negras são irredutíveis à combinação dos efeitos de ser mulher e ser negra. Há particularidades (sobretudo em se tratando de experiências de preconceito, violência e representatividade) que emergem na intersecção desses dois marcadores sociais (Ribeiro; O'dwyer; Heilborn, 2018). Do ponto de vista das relações entre variáveis, ser mulher independe de ser negra. Saber que alguém é mulher não ajuda a predizer sua cor. Saber a cor, não ajuda a predizer sua identidade de gênero. Essas duas variáveis são, portanto, independentes. Contudo seus efeitos são geralmente relacionados.

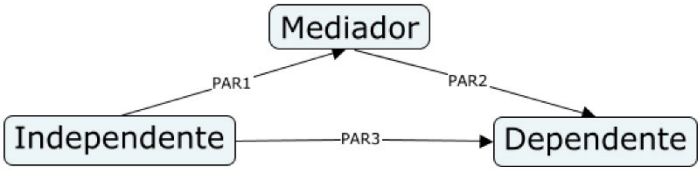
Em termos mais gerais, o parâmetro que descreve relação preditiva entre duas variáveis (dependente e independente) pode ser parcialmente moderado pela presença de uma terceira variável. Nessas situações, dizemos que há uma relação de moderação (cf. figura 15.1).

Figura 15.1: Representação de uma relação de moderação



Em outros casos, é possível que a variável independente tenha um efeito direto e outro indireto sobre a variável dependente. A via direta é computada e representada da maneira usual (cf. figura 15.2). Já a via indireta ocorre quando a variável independente possui efeito sobre uma terceira variável (dita mediadora) e esta, por sua vez, tem um efeito sobre a variável dependente. Quando esse tipo de relação causal ocorre, estamos diante de uma relação de *mediação* (Mair, 2018).

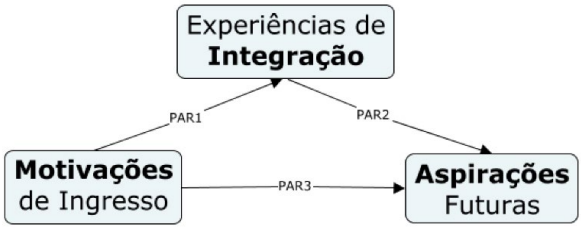
Figura 15.2: Representação de uma relação de mediação



Há vários exemplos de relação de mediação na literatura educacional. No início desta lição, mencionei uma análise em que as habilidades científicas dos estudantes mediavam a relação entre sexo, idade e criatividade científica (Dikici; Özdemir; Clark, 2018). Um modelo bastante influente para pensar as razões institucionais que levam à evasão da educação superior (Lima Junior; Ostermann; Rezende, 2018; Mannan, 2007; Massi; Villani, 2015) foi proposto por Vincent Tinto (1987, 2012). Ele argumenta que os estudantes ingressam nos cursos de graduação com motivações e interesses que, atualizados pelas experiências de integração social e

acadêmica, contribuem para que alguns estudantes decidam abandonar o curso. Essa asserção pode ser representada e testada como um *modelo de mediação* em que as experiências de integração mediam o efeito das motivações de ingresso sobre as aspirações de futuro dos estudantes (cf. figura 15.3).

Figura 15.3: Representação simplificada do modelo de Tinto (1987, 2012)



Com dados de um questionário desenvolvido e aplicado recentemente por Lima Junior *et al.* (2020), dentro de certos limites, podemos testar a plausibilidade do modelo apresentado.

Modelagem de equações estruturais

A modelagem de equações estruturais (ing. *Structural Equation Modeling* – SEM) pode ser realizada com a função `sem` da biblioteca `lavaan`. A sintaxe do modelo é a mesma que aprendemos até agora e está sintetizada no quadro 15.2.

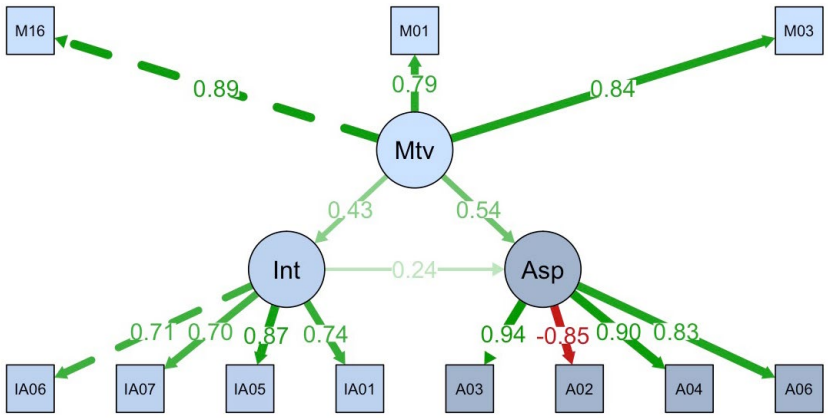
Quadro 15.2: Sintaxe da função `sem`

	O que significa?	O que faz?
=~	"é manifesto por"	Define fatores latentes.
~~	"está associado a"	Inclui termos de covariância (ou variância).
~	"é predito por"	Inclui termos de regressão.

Em um modelo de mediação (cf. figura 15.2), a variável mediadora é predita pela variável independente. Enquanto isso, a variável dependente é predita pelas variáveis independente e mediadora. O código a seguir carrega respostas dadas por estudantes do curso de Física a um questionário baseado no modelo de Tinto. Após realizar uma análise exploratória e testar algumas associações entre as variáveis latentes geradas (Lima Junior; Fraga Junior et al., 2020), proponho o seguinte modelo de equações estruturais:

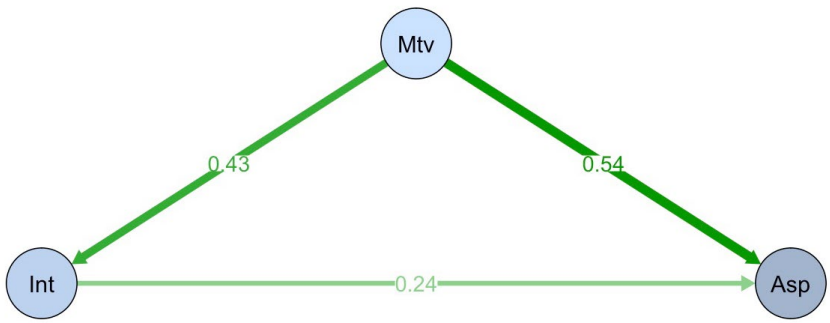
```
rm(list = ls())
load("TINTO.dat")

modelo = "Motiva  =~ Motiva16 + Motiva01 + Motiva03
          Integra =~ IntegraAcad06 + IntegraAcad07 + IntegraAcad05 +
IntegraAcad01
          Aspira  =~ Aspira03 + Aspira02 + Aspira04 + Aspira06
          Aspira   ~ Integra + Motiva
          Integra  ~ Motiva"

fit = sem(modelo, data = dados, ordered = colnames(dados))
```

Diagrama 15.4: Diagrama de caminho do modelo de Tinto



No diagrama 15.4, a motivação **Mtv** pode afetar diretamente as aspirações **Asp** dos estudantes com relação à permanência no curso. Porém, essa motivação também afeta suas experiências de integração **Int**. Ou seja, estudantes que ingressam mais motivados costumam reportar um grau de integração maior. Essa integração, por sua vez, tem poder preditivo sobre as aspirações dos estudantes. No diagrama 15.5, somente as relações estruturais foram representadas com o propósito de explicitar que a integração media o efeito da motivação sobre a permanência.

Diagrama 15.5: Diagrama de caminho do modelo de Tinto (somente relações estruturais)



A significância estatística dos parâmetros publicados no diagrama pode ser investigada fazendo **`parameterEstimates`**.

```
parameterEstimates(fit, ci = FALSE, standardized = TRUE)

##             lhs  op           rhs    est    se       z pvalue std.lv std.all
## 1        Motiva  =~      Motiva16  1.000 0.000      NA     NA  0.888   0.888
## 2        Motiva  =~      Motiva01  0.892 0.079  11.255  0.000  0.792   0.792
## 3        Motiva  =~      Motiva03  0.948 0.067  14.156  0.000  0.842   0.842
## 4       Integra  =~ IntegraAcad06  1.000 0.000      NA     NA  0.714   0.714
## 5       Integra  =~ IntegraAcad07  0.975 0.108   9.031  0.000  0.696   0.696
## 6       Integra  =~ IntegraAcad05  1.226 0.112  10.915  0.000  0.875   0.875
## 7       Integra  =~ IntegraAcad01  1.041 0.097  10.731  0.000  0.743   0.743
```

```
## 8         Aspira  =~     Aspira03  1.000 0.000      NA     NA  0.937   0.937
## 9         Aspira  =~     Aspira02 -0.903 0.039 -23.238  0.000 -0.846  -0.846
## 10        Aspira  =~     Aspira04  0.962 0.033  29.555  0.000  0.901   0.901
## 11        Aspira  =~     Aspira06  0.884 0.042  20.943  0.000  0.828   0.828
## 12        Aspira   ~      Integra  0.318 0.111   2.861  0.004  0.242   0.242
## 13        Aspira   ~       Motiva  0.573 0.088   6.484  0.000  0.543   0.543
## 14       Integra   ~       Motiva  0.348 0.070   4.964  0.000  0.433   0.433
```

Como é possível perceber, todas as cargas fatoriais e todos os parâmetros de regressão resultaram estatisticamente significativos (com $p < 0,05$). A qualidade do modelo pode ser avaliada da seguinte forma:

```
summary(fit, fit.measures = TRUE)

## lavaan 0.6-7 ended normally after 27 iterations
##
##   Estimator                                         DWLS
##   Optimization method                           NLMINB
##   Number of free parameters                         58
##
##                                                   Used       Total
##   Number of observations                           151         160
##
## Model Test User Model:
##                                                Standard      Robust
##   Test Statistic                                 83.504     130.573
##   Degrees of freedom                                 41          41
##   P-value (Chi-square)                            0.000       0.000
##   Scaling correction factor                                   0.716
##   Shift parameter                                            13.888
##        simple second-order correction
##
## Model Test Baseline Model:
##
##   Test statistic                              4442.689    2060.967
##   Degrees of freedom                                55          55
##   P-value                                         0.000       0.000
```

```
##   Scaling correction factor                                  2.187
##
## User Model versus Baseline Model:
##
##   Comparative Fit Index (CFI)                    0.990       0.955
##   Tucker-Lewis Index (TLI)                       0.987       0.940
##
##   Robust Comparative Fit Index (CFI)                            NA
##   Robust Tucker-Lewis Index (TLI)                               NA
##
## Root Mean Square Error of Approximation:
##
##   RMSEA                                          0.083       0.121
##   90 Percent confidence interval - lower         0.057       0.098
##   90 Percent confidence interval - upper         0.109       0.144
##   P-value RMSEA <= 0.05                          0.020       0.000
##
##   Robust RMSEA                                                  NA
##   90 Percent confidence interval - lower                        NA
##   90 Percent confidence interval - upper                        NA
##
## Standardized Root Mean Square Residual:
##
##   SRMR                                           0.081       0.081
```

Enfim, como é possível perceber, o modelo tem algumas limitações. A quantidade de observações é relativamente pequena se comparada à quantidade de parâmetros ajustados. Por outro lado, considerando a maioria das medidas de qualidade do ajuste, a plausibilidade do modelo está entre ótima ($CFI = 0,99,\ TLI = 0,97$), boa ($SRMR \cong 0,08$) e aceitável ($RMSEA \cong 0,08$). De certa maneira, o modelo de Tinto foi corroborado (Lima Junior; Fraga Junior *et al.*, 2020).

Considerações finais

A modelagem de equações estruturais pode ser considerada um dos métodos mais empregados e sofisticados da pesquisa educacional quantitativa. Optei por abordar esse método na última lição justamente porque ele se conecta a quase todos os outros. Incorporando a análise fatorial confirmatória, a modelagem por equações estruturais permite realizar uma redução dimensional. Por isso, algum conhecimento sobre análise de componentes principais (cf. lição 10) e análise de correspondência múltipla (cf. lição 11) pode tornar mais intuitivo o entendimento desse método. Contudo, em contraste com o caráter exploratório-descritivo desses métodos de redução dimensional, a modelagem de equações estruturais permite incorporar relações preditivas ao modelo. Tais relações (do tipo $X \rightarrow Y$), exploradas no âmbito dos modelos lineares multivariados (lições 7 a 9), foram sistematicamente diferenciadas das relações associativas simétricas (do tipo $X \leftrightarrow Y$) que deram origem às análises de correspondência e componentes principais. Nesta lição, de alguma maneira, tudo se conecta.

Para a maioria dos pesquisadores, os métodos bivariados abordados no começo do livro (testes t, *one-way anova*, regressão bivariada, correlação, tabelas de contingência) serão suficientes para gerar respostas adequadas às suas questões de pesquisa. Por isso, não faz sentido começar pelos métodos mais sofisticados. Além disso, o resultado da sua investigação não ganha em qualidade ou confiabilidade se você empregar um método mais sutil. Bons árbitros sabem identificar quando o analista está matando uma mosca com uma bala de canhão e certamente desconfiarão da sua aposta em métodos sofisticados para resolver problemas simples. Na pesquisa educacional, a estatística é um meio, não o fim. Se sua pesquisa envolve

um grande esforço estatístico sem produzir resultados relevantes do ponto de vista educacional, ela está certamente mal formulada.

Ao longo deste livro, dentro dos meus limites, tentei mostrar como todas essas análises podem ser feitas no ambiente de programação do R. Tentei explorar exemplos concretos e simples, mas potencialmente relevantes do ponto de vista da pesquisa educacional. Espero que esses exemplos tenham deixado evidente que os métodos quantitativos não estão comprometidos com ideias positivistas ou conservadoras no campo educacional. De fato, grande parte do pensamento educacional crítico e emancipatório está assentado sobre evidências empíricas levantadas por pesquisas quantitativas clássicas e contemporâneas. No mais, a produção de conhecimento em ciências humanas e sociais segue como sempre foi: um esporte de combate. Espero que a luta que travamos neste livro sirva para que o leitor lute bem suas próprias lutas.

REFERÊNCIAS

ALTHUSSER, Louis. *Ideologias e aparelhos ideológicos de Estado*. Rio de Janeiro: Contraponto, 2013.

ALVES, Fátima; ORTIGÃO, Isabel; FRANCO, Creso. Origem social e risco de repetência: interação raça-capital econômico. *Cadernos de Pesquisa*, v. 37, n. 130, p. 161-180, 2007.

ARCHER, Louise. *et al.* "Science capital": A conceptual, methodological, and empirical argument for extending bourdieusian notions of capital beyond the arts. *Journal of Research in Science Teaching*, v. 52, n. 7, p. 922-948, 2015.

BÄCHTOLD, Manuel; CROSS, David; MUNIER, Valérie. How to Assess and Categorize Teachers' Views of Science? Two Methodological Issues. *Research in Science Education*, 2019.

BATISTA, Marcos Antonio *et al.* Avaliação institucional no ensino superior: construção de escalas para discentes e docentes. *Avaliação*: Revista da Avaliação da Educação Superior (Campinas), v. 18, n. 1, p. 201-218, 2013.

BEZZI, Alfredo. What is this thing called geoscience? Epistemological dimensions elicited with the repertory grid and their implications for scientific literacy. *Science Education*, v. 83, n. 6, p. 675-700, 1999.

BØE, Maria Vetleseter; HENRIKSEN, Ellen Karoline. Love It or Leave It: Norwegian Students' Motivations and Expectations for Postcompulsory Physics. *Science Education*, v. 97, n. 4, p. 550-573, 2013.

BOTTIA, Martha Cecilia *et al.* The role of high school racial composition and opportunities to learn in students' STEM college participation. *Journal of Research in Science Teaching*, v. 55, n. 3, p. 446-473, 2018.

BOURDIEU, Pierre. *A economia das trocas simbólicas*. 6. ed. São Paulo: Perspectiva, 2007a.

BOURDIEU, Pierre. *Distinction*: a social critique of the judgement of taste. London: Routledge, 1984.

BOURDIEU, Pierre. *Escritos de Educação*. Petrópolis: Vozes, 2007b.

BOURDIEU, Pierre. Espaço social e espaço simbólico. *Razões práticas*. Campinas: Papirus, 1996. p. 13–33.

BRANDSTETTER, Miriam; SANDMANN, Angela; FLORIAN, Christine. Understanding pictorial information in biology: students' cognitive activities and visual reading strategies. *International Journal of Science Education*, v. 39, n. 9, p. 1218-1237, 2017.

BUFFLER, Andy *et al.* The development of first year physics students' ideas about measurement in terms of point and set paradigms. *International Journal of Science Education*, v. 23, n. 11, p. 1137-1156, 2001.

BUFFLER, Andy; LUBBEN, Fred; IBRAHIM, Bashirah. The relationship between students' views of the nature of science and their views of the nature of scientific measurement. *International Journal of Science Education*, v. 31, n. 9, p. 1137-1156, 2009.

CHILDERS, Gina; JONES, M. Gail. Learning from a distance: high school students' perceptions of virtual presence, motivation, and science identity during a remote microscopy investigation. *International Journal of Science Education*, v. 39, n. 3, p. 257-273, 2017.

COLOSIMO, Enrico Antônio; GIOLO, Suely Ruiz. *Análise de sobrevivência aplicada*. São Paulo: Edgard Blucher, 2006.

CONLEY, Anne Marie. Patterns of motivation beliefs: combining achievement goal and expectancy-value perspectives. *Journal of Educational Psychology*, v. 104, n. 1, p. 32-47, 2012.

CONTRERAS, José. *Autonomia de professores*. São Paulo: Cortez, 2002.

CRAWLEY, Michael J. *Statistics*: an introduction using R. Hoboken: John Wiley & Sons, 2005.

DA SILVA, Erica Tatiane *et al.* Factors influencing students' performance in a Brazilian dental school. *Brazilian Dental Journal*, v. 21, n. 1, p. 80-86, 2010.

DA SILVEIRA, Fernando Lang. Um exemplo de análise multivariada aplicada à pesquisa quantitativa em ensino de ciências: explicando o desempenho dos candidatos ao concurso vestibular de 1999 da Universidade Federal do Rio Grande do Sul. *Investigações em Ensino de Ciências*, v. 4, n. 2, p. 161-180, 1999.

DEWEY, John. *Experience and education*. Indianapolis: Kappa Delta Pi, 1938.

DEWITT, Jennifer; ARCHER, Louise; MAU, Ada. Dimensions of science capital: exploring its potential for understanding students' science participation. *International Journal of Science Education*, v. 38, n. 16, p. 2431-2449, 2016.

DIKICI, Ayhan; ÖZDEMIR, Gökhan; CLARK, Douglas B. The relationship between demographic variables and scientific creativity: mediating and moderating roles of scientific process skills. *Research in Science Education*, 2018.

DYKE, Chuck. Bourdieuean Dynamics: The American middle-class self-constructs. *In*: SHUSTERMAN, Richard (org.). *Bourdieu: A Critical Reader*. Hoboken: Wiley-Blackwell, 1999. p. 192–213.

ENGSTRÖM, Susanne; CARLHED, Carina. Different habitus: different strategies in teaching physics? Relationships between teachers' social, economic and cultural capital and strategies in teaching physics in upper secondary school. *Cultural Studies of Science Education*, v. 9, n. 3, p. 1-30, 2014.

EPSKAMP, Sacha. semPlot: unified visualizations of structural equation models. *Structural Equation Modeling*, v. 22, n. 3, p. 474-483, 2015.

FABRIGAR, Leandre R.; WEGENER, Duane T. *Exploratory factor analysis*. New York: Oxford University Press, 2012.

FAHRMEIR, Ludwig *et al*. *Regression: models, methods and applications*. Berlin: Springer, 2013.

FREIRE, Paulo. *Pedagogia do oprimido*. São Paulo: Paz e Terra, 1974.

GARCIA-SILVA, Sullyvan; LIMA JUNIOR, Paulo; CARUSO, Haydée. A violência urbana e escolar nas periferias de Brasília. *Educação e Sociedade*, v. 43, e248105, 2022.

GATTI, Bernardete A. Estudos quantitativos em educação. *Educação e Pesquisa*, v. 30, n. 1, p. 11–30, 2004.

GILIOLI, Renato de Sousa Porto. *Evasão em instituições federais de ensino superior no Brasil*: expansão da rede, SISU e desafios. p. 55, 2016.

GLYNN, Shawn M. *et al.* Science motivation questionnaire II: validation with science majors and nonscience majors. *Journal of Research in Science Teaching*, v. 48, n. 10, p. 1159-1176, 2011.

GOTTLIEB, Jessica. STEM career aspirations in Black, Hispanic, and White ninth-grade students. *Journal of Research in Science Teaching*, v. 55, n. 10, p. 1365-1392, 2018.

GREENACRE, Michael. *Correspondence analysis in practice*. New York: Chapman & Hall, 2007.

HÄRDLE, Wolfgang Karl; SIMAR, Léopold. *Applied multivariate statistical analysis*. 4. ed. New York: Springer, 2015.

HEIDEMANN, Leonardo Albuquerque; ARAUJO, Leonardo Albuquerque; VEIT, Eliane Angela. Modelagem didático-científica: integrando atividades experimentais e o processo de modelagem científica no ensino de Física. *Caderno Brasileiro de Ensino de Física*, v. 33, n. 1, p. 3, 2016.

HILLMAN, Susan J. *et al*. K-12 Students' Perceptions of Scientists: Finding a valid measurement and exploring whether exposure to scientists makes an impact. *International Journal of Science Education*, v. 36, n. 15, p. 2580-2595, 2014.

HO, Hsin-Ning Jessie; LIANG, Jyh-Chong. The Relationships Among Scientific Epistemic Beliefs, Conceptions of Learning Science, and Motivation of Learning Science: A study of Taiwan high school students. *International Journal of Science Education*, v. 37, n. 16, p. 2688-2707, 2015.

HO, Walter *et al*. Measuring the perception of quality physical education in Latin American professionals. *Revista Brasileira de Ciencias do Esporte*, v. 40, n. 4, p. 361-369, 2018.

HOFFMANN, Celina *et al*. Prazer e sofrimento no trabalho docente: Brasil e Portugal. *Educação e Pesquisa*, v. 45, p. 0-2, 2019.

HUSSON, François; LÊ, Sébastien; PAGÈS, Jérôme. *Exploratory Multivariate Analysis by Example Using R*. London: CRC Press, 2010.

JOHNSON, Angela *et al.* Authoring identity amidst the treacherous terrain of science: A multiracial feminist examination of the journeys of three women of color in science. *Journal of Research in Science Teaching*, v. 48, n. 4, p. 339-366, 2011.

JOINT COMMITEE FOR GUIDES IN METROLOGY. *Guide for the expression of uncertainty in measurement*. Sèvres: BIPM, 2008a.

JOINT COMMITEE FOR GUIDES IN METROLOGY. *International vocabulary in metrology: basic and general concepts and associated terms*. Sèvres: BIPM, 2008b.

KAYA, Osman Nafiz; YAGER, Robert; DOGAN, Alev. Changes in attitudes towards science - Technology - Society of pre-service science teachers. *Research in Science Education*, v. 39, n. 2, p. 257-279, 2009.

KING, Natalie S.; PRINGLE, Rose M. Black girls speak STEM: Counterstories of informal and formal learning experiences. *Journal of Research in Science Teaching*, v. 56, n. 5, p. 539-569, 2019.

LEDERMAN, Norm G. *et al.* Views of Nature of Science Questionnaire: Toward Valid and Meaningful Assessment of Learners' Conceptions of Nature of Science. *Journal of Research in Science Teaching*, v. 39, n. 6, p. 497-521, 2002.

LEE, Christine S. *et al.* Understanding motivational structures that differentially predict engagement and achievement in middle school science. *International Journal of Science Education*, v. 38, n. 2, p. 192-215, 2016.

LIMA JUNIOR, Paulo *et al.* A integração dos estudantes de periferia no curso de Física: razões institucionais da evasão segundo a origem social. *Ciência & Educação*, v. 26, 2020.

LIMA JUNIOR, Paulo *et al*. Excelência, evasão e experiências de integração dos estudantes de graduação em Física. *Ensaio Pesquisa em Educação em Ciências (Belo Horizonte)*, v. 22, p. 1-23, 2020.

LIMA JUNIOR, Paulo *et al*. Taxas longitudinais de retenção e evasão: uma metodologia para estudo da trajetória dos estudantes na educação superior. *Ensaio: Avaliação e Políticas Públicas em Educação*, v. 27, n. 102, p. 157-178, 2019.

LIMA JUNIOR, Paulo. Trajetórias dos professores de ciências em tempos de proletarização: família e vocação docente. *In*: MASSI, L.; LIMA JUNIOR, P.; BAROLLI, E. (org.). . *Retratos da docência: contextos, saberes e trajetórias*. Araraquara: Letraria, 2018. p. 435-459.

LIMA JUNIOR, Paulo; DA SILVEIRA, Fernando Lang. Discutindo os conceitos de erro e incerteza a partir da tábua de Galton com estudantes de graduação: uma contribuição para a incorporação de novas abordagens da metrologia ao ensino de física superior. *Caderno Brasileiro de Ensino de Física*, v. 28, n. 2, p. 400-422, 2011.

LIMA JUNIOR, Paulo; DA SILVEIRA, Fernando Lang; OSTERMANN, Fernanda. Análise de sobrevivência aplicada ao estudo do fluxo escolar nos cursos de graduação em Física: um exemplo de uma universidade brasileira. *Revista Brasileira de Ensino de Física*, v. 34, n. 1, p. 1403, 2011.

LIMA JUNIOR, Paulo; FRAGA JUNIOR, Jailton Correia. Qual é o efeito da desiguladade social no desempenho em ciências dos estudantes brasileiros? Uma análise do Exame Nacional do Ensino Médio (2012-2019). *Investigações em Ensino de Ciências*, v. 26, n. 1, p. 110, 30 abr. 2021. Disponível em: https://www.if.ufrgs.br/cref/ojs/index.php/ienci/article/view/2057.

LIMA JUNIOR, Paulo; OSTERMANN, Fernanda; REZENDE, Flavia. Marxism in Vygotskian approaches to cultural studies of science education. *Cultural Studies of Science Education*, p. 1-24, 2013.

LIMA JUNIOR, Paulo; OSTERMANN, Fernanda; REZENDE, Flavia. *Razões para desistir: análise sociológica da evasão no curso de Física*. Curitiba: Appris, 2018.

MAIR, Patrick. *Modern Psychometrics with R*. Cham: Springer International Publishing, 2018. (Use R!).

MANNAN, Abdul. Student attrition and academic and social integration : application of Tinto ' s model at the University of Papua New Guinea. *Higher Education*, p. 147–165, 2007.

MASSI, Luciana; CARVALHO, Helena; GIORDAN, Marcelo. Perfil socioformativo dos orientadores, heterogeneidade e hierarquia social na área de ensino da CAPES. *Investigações em Ensino de Ciências*, v. 25, n. 1, p. 421, 2020.

MASSI, Luciana; VILLANI, A. Um caso de contratendência: baixa evasão na licenciatura em química explicada pelas disposições e integrações. *Educação e Pesquisa*, v. 41, n. 4, p. 975-992, 2015.

MAZAS, Beatriz *et al.* Development and validation of a scale to assess students' attitude towards animal welfare. *International Journal of Science Education*, v. 35, n. 11, p. 1775-1799, 2013.

MELI, Kalliopi; LAVIDAS, Konstantinos ; KOLIOPOULOS, Dimitrios. Factors that influence students in choosing Physics Programmes at University Level: the Case of Greece. *Research in Science Education*, p. 1-17, 2018.

MONT'ALVÃO NETO, Arnaldo Lopo; MONT'ALVÃO, Arnaldo. Estratificação educacional no Brasil do século XXI. *DADOS - Revista de Ciências Sociais*, v. 54, n. 2, p. 389-430, 2011.

MONTEIRO NASCIMENTO, Matheus *et al*. Métodos quantitativos interpretativos na educação em ciências: abordagens para análise multivariada de dados. *Revista Brasileira de Pesquisa em Educação em Ciências*, p. 775-800, 2019.

MORTON, Terrell R.; PARSONS, Eileen C. #BlackGirlMagic: the identity conceptualization of Black women in undergraduate STEM education. *Science Education*, v. 102, n. 6, p. 1363-1393, 2018.

MUJTABA, Tamjid *et al.* Students' science attitudes, beliefs, and context: associations with science and chemistry aspirations. *International Journal of Science Education*, v. 40, n. 6, p. 644-667, 2018.

MUN, Kongju *et al*. Korean secondary students' perception of scientific literacy as global citizens: using global scientific literacy questionnaire. *International Journal of Science Education*, v. 37, n. 11, p. 1739-1766, 2015.

NOGUEIRA, Cláudio Marques Martins; NOGUEIRA, Maria Alice. *Bourdieu e a educação*. Belo Horizonte: Autêntica, 2009.

OLSEN, Rolf Vegar; LIE, Svein. Profiles of students' interest in science issues around the world: Analysis of data from pisa 2006. *International Journal of Science Education*, v. 33, n. 1, p. 97-120, 2011.

OWSTON, Ron; YORK, Dennis; MURTHA, Susan. Student perceptions and achievement in a university blended learning strategic initiative. *The Internet and Higher Education*, v. 18, p. 38-46, jul. 2013.

ÖZMEN, Kübra; ÖZDEMIR, Ömer Faruk. Conceptualisation and development of the physics related personal epistemology questionnaire (PPEQ). *International Journal of Science Education*, v. 41, n. 9, p. 1207-1227, 2019.

PEREIRA, Alexandre Severino *et al*. Fatores relevantes no processo de permanência prolongada de discentes nos cursos de graduação presencial: um estudo na Universidade Federal do Espírito Santo. *Ensaio: Avaliação e Políticas Públicas em Educação*, v. 23, n. 89, p. 1015–1039, 2015. Disponível em: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0104-40362015000401015&lang=pt.

RABELO, Mauro. *Avaliação educacional*: fundamentos, metodologia e aplicações no contexto brasileiro. Rio de Janeiro: SBM, 2013.

READY, Douglas D.; WRIGHT, David L. Accuracy and Inaccuracy in Teachers' Perceptions of Young Children's Cognitive Abilities. *American Educational Research Journal*, v. 48, n. 2, p. 335–360, abr. 2011.

REINISCH, Bianca *et al*. Methodical challenges concerning the Draw-A-Scientist Test: a critical view about the assessment and evaluation of learners' conceptions of scientists. *International Journal of Science Education*, v. 39, n. 14, p. 1952-1975, 2017.

REIS, Cisne Zélia Teixeira; SILVEIRA, Suely de Fátima Ramos; FERREIRA, Marco Aurélio Marques. Autoavaliação em uma instituição federal de ensino superior: resultados e implicações. *Avaliação: Revista da Avaliação da Educação Superior (Campinas)*, v. 15, n. 3, p. 109-129, 2010.

REZENDE, Flavia; OSTERMANN, Fernanda. O protagonismo controverso dos mestrados profissionais em ensino de ciências. *Ciência & Educação (Bauru)*, v. 21, n. 3, p. 543–558, set. 2015.

RIBEIRO, Letícia; O'DWYER, Brena; HEILBORN, Maria Luiza. Feminists dilemmas and the possibility of radicalizing democracy in the midst of differences: The Slut Walk of Rio de Janeiro. *Civitas*, v. 18, n. 1, p. 83-99, 2018.

RONDINI, Carina Alexandra; TEIXEIRA FILHO, Fernando Silva; TOLEDO, Lívia Gonsalves. Homophobic conceptions of high school students. *Psicologia USP*, v. 28, n. 1, p. 57-71, 2017.

SCHMIDT, Jennifer A.; ROSENBERG, Jennifer M.; BEYMER, Patrick N. A person-in-context approach to student engagement in science: Examining learning activities and choice. *Journal of Research in Science Teaching*, v. 55, n. 1, p. 19-43, 2018.

SCHUMM, Maximiliane F.; BOGNER, Franz X. Measuring adolescent science motivation. *International Journal of Science Education*, v. 38, n. 3, p. 434-449, 2016.

TESTA, Italo *et al.* Development and validation of a university students' progression in learning quantum mechanics through exploratory factor analysis and Rasch analysis. *International Journal of Science Education*, v. 41, n. 3, p. 388-417, 2019.

TINTO, Vincent. *Completing college: Rethinking institucional action*. Chicago: Chicago University Press, 2012.

TINTO, Vincent. *Leaving college: rethinking the causes and cures of student attrition*. 2. ed. Chicago: Chicago University Press, 1987.

TUAN, Hsiao-Lin; CHIN, Chi-Chin ; SHIEH, Shyang-Horng. The development of a questionnaire to measure students' motivation towards science learning. *International Journal of Science Education*, v. 27, n. 6, p. 639-654, 2005.

VUOLO, José Henrique. *Fundamentos da teoria de erros*. São Paulo: Blucher, 1996.

WAINER, Jacques; MELGUIZO, Tatiana. Políticas de inclusão no ensino superior: avaliação do desempenho dos alunos baseado no Enade de 2012 a 2014. *Educação e Pesquisa*, n. ahead of print, p. 1-15, 2017.

YEH, Yi-Fen; JEN, Tsung-Hau; HSU, Ying-Shao. Major Strands in Scientific Inquiry through Cluster Analysis of Research Abstracts. *International Journal of Science Education*, v. 34, n. 18, p. 2811-2842, 2012.

ZUMBO, Bruno D.; CHAN, Eric K. H. *Validity and Validation in Social, Behavioral, and Health Sciences*. Cham: Springer International Publishing, 2014. v. 54. (Social Indicators Research Series).

APÊNDICE.

ÁLGEBRA MATRICIAL

Esse apêndice serve como uma breve revisão de álgebra matricial. A notação matricial é importante para todas as lições de análise multivariada e alguns teoremas são fundamentais para dar sentido às análises.

Uma **matriz** *n x p* é uma tabela de informações dispostas em *n linhas* e *p colunas*:

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1p} \\ a_{21} & a_{22} & \dots & a_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \dots & a_{np} \end{bmatrix}$$

No ambiente de programação do R, um **vetor** é qualquer sequência de informações do mesmo tipo. Portanto, eles podem ser pensados como se fossem a coluna de uma matriz:

$$\vec{v} = \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{bmatrix}$$

Embora o R permita declarar matrizes e vetores empregando qualquer tipo de variável (e.g., caracteres, valores lógicos), geralmente estamos interessados em matrizes e vetores de entradas numéricas. A propósito, vetores também podem ser representados como segmentos de reta orientados no espaço. Em ciências da natureza, esse tipo de representação geométrica é muito útil para lidar com quantidades orientadas (tais como força e velocidade). Em estatística, a representação geométrica de um vetor será importante para desenvolver um pensamento de alto nível, representando a associação entre variáveis como ângulos entre vetores.

Matrizes especiais

Algumas matrizes recebem nomes especiais. A seguir, definimos o que esses nomes querem dizer e fornecemos exemplos:

1. **matriz transposta** é como chamamos qualquer matriz obtida a partir de outra por uma permutação de suas linhas e colunas. Ela é representada com um T maiúsculo sobrescrito.

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \\ a_{31} & a_{32} \end{bmatrix}$$

$$\therefore \mathbf{A}^T = \begin{bmatrix} a_{11} & a_{21} & a_{31} \\ a_{12} & a_{22} & a_{32} \end{bmatrix}$$

2. **vetor-linha** é obtido pela transposição de um vetor-coluna.

$$\vec{v} = \begin{bmatrix} v_1 \\ v_2 \\ v_n \end{bmatrix}$$

$$\therefore \vec{v}^T = \begin{bmatrix} v_1 & v_2 & v_n \end{bmatrix}$$

3. **matriz quadrada** é como chamamos qualquer matriz que tenha a mesma quantidade de linhas e colunas.

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}$$

4. **matriz diagonal** é qualquer matriz quadrada que seja nula em todos os elementos fora da diagonal principal.

$$\begin{bmatrix} a_{11} & 0 & 0 \\ 0 & a_{22} & 0 \\ 0 & 0 & a_{33} \end{bmatrix}$$

5. **matriz identidade** é uma matriz diagonal com entradas unitárias.

$$1 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

6. **matriz simétrica** é qualquer matriz quadrada que seja igual à sua transposta.

$$\mathbf{A} = \mathbf{A}^T$$

7. **matriz inversa** é uma matriz que, ao multiplicar outra, pela direita ou pela esquerda, produz uma matriz identidade.

$$\mathbf{A}\mathbf{A}^{-1} = \mathbf{A}^{-1}\mathbf{A} = 1$$

8. **matriz ortogonal** é qualquer matriz quadrada que, multiplicada pela direita ou pela esquerda por sua transposta, produz uma matriz identidade. Também podemos dizer que uma matriz ortogonal é tal que sua transposta é igual à sua inversa.

$$\mathbf{A}\mathbf{A}^T = \mathbf{A}^T\mathbf{A} = 1$$

Operações com matrizes

Uma ideia muito importante na álgebra matricial é a possibilidade de interpretarmos matrizes como operadores. Em outras palavras, uma matriz pode representar uma *transformação linear*, levando vetores de um espaço vetorial E1 em vetores de outro espaço E2.

$$\mathbf{A} : E_1 \rightarrow E_2$$

De fato, todas as transformações lineares aplicáveis a vetores no espaço (rotações, inversões, mudanças de escala, projeções) podem ser representadas por matrizes. Contudo, para que isso ocorra, é preciso sermos muito cuidadosos na maneira como nós definimos as operações entre matrizes.

É importante saber que, no ambiente de programação do R, as restrições listadas a seguir não estão automaticamente implementadas. A maneira como o R opera matrizes é tremendamente flexível e algumas operações possíveis não fazem o menor sentido do ponto de vista da álgebra matricial. Essa liberdade, no entanto, não representará um problema na medida em que o usuário souber o que está fazendo.

1. **Adição e subtração**. Duas matrizes podem ser somadas (ou subtraídas) somente quando têm a mesma quantidade de linhas e colunas. Isso permite levar essas operações a cabo em cada célula de maneira independente:

$$\mathbf{A}-\mathbf{B}=\begin{bmatrix} a_{11}-b_{11} & a_{12}-b_{12} & \dots & a_{1p}-b_{1\mathrm{p}} \\ a_{21}-b_{21} & a_{22}-b_{22} & \dots & a_{2p}-b_{2\mathrm{p}} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1}-b_{n1} & a_{n2}-b_{n1} & \dots & a_{np}-b_{n1} \end{bmatrix}$$

No ambiente de programação do R, a restrição sobre as dimensões das matrizes não é sempre respeitada. Por exemplo, ao demandar a soma ou subtração de dois vetores de dimensões diferentes, se a quantidade de entradas em um vetor for um múltiplo inteiro da quantidade de entradas no outro vetor, o R replicará implicitamente o vetor mais curto, publicando o resultado.

```
c(1, 2, 3, 4, 5, 6, 7, 8) + c(10, 20)

## [1] 11 22 13 24 15 26 17 28
```

Esse recurso pode simplificar bastante a programação, mas é importante saber que ele não faz muito sentido do ponto de vista da álgebra matricial.

1. **Multiplicação**. A multiplicação de matrizes talvez seja a operação mais complicada de visualizar. Ela envolve multiplicar e somar diversos elementos das matrizes segundo uma regra específica, que pode ficar mais intuitiva se tomarmos um sistema de equações como ponto de partida.

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 = y_1 \\ a_{21}x_1 + a_{22}x_2 + a_{23}x_3 = y_2 \\ a_{31}x_1 + a_{32}x_2 + a_{33}x_3 = y_3 \end{cases}$$

Considere que desejemos representar essas equações de maneira mais compacta usando a notação matricial. É possível representar os coeficientes a_{ij} do sistema por uma matriz ***A*** e as variáveis x_i e y_i por vetores-coluna. Feito isso, podemos escrever que:

$$\boldsymbol{A}\vec{x} = \vec{y}$$

Na equação anterior, a matriz ***A*** pode ser interpretada como uma transformação linear que leva x em y. Ao mesmo tempo, a multiplicação de matrizes deve satisfazer:

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 \\ a_{21}x_1 + a_{22}x_2 + a_{23}x_3 \\ a_{31}x_1 + a_{32}x_2 + a_{33}x_3 \end{bmatrix}$$

Em termos mais gerais, definimos que o produto de duas matrizes **A** e **B** também é uma matriz e seu elemento da *i*-ésima linha e *j*-ésima coluna pode ser obtido multiplicando e somando os elementos da *i*-ésima linha de ***A*** pela *j*-ésima coluna de ***B***:

$$\{\mathrm{AB}\}_{ij} = \sum\nolimits_k a_{ik} b_{kj}$$

Como é possível perceber, em álgebra matricial, a multiplicação de matrizes só está definida quando a quantidade de colunas da matriz ***A*** é igual à quantidade de linhas da matriz ***B***. Contudo, no ambiente de programação do R, nenhuma dessas exigências está implementada. Multiplicações de matrizes são realizadas de maneira independente em cada célula:

```
A = matrix(data = c(1:9), nrow = 3, ncol = 3)
B = diag(nrow = 3)  # matriz identidade
A * B

##      [,1] [,2] [,3]
## [1,]    1    0    0
## [2,]    0    5    0
## [3,]    0    0    9
```

Para realizar uma multiplicação de matrizes tal como ela está definida em álgebra matricial, é preciso empregar o operador %*%:

```
A %*% B

##      [,1] [,2] [,3]
## [1,]    1    4    7
## [2,]    2    5    8
## [3,]    3    6    9
```

Outras operações simples entre quantidades escalares (e.g., divisão, exponenciação, logaritmo…) não estão propriamente definidas em álgebra matricial. Porém, se demandarmos que o R execute essas operações, ele provavelmente publicará um resultado. A interpretação desse resultado, no entanto, fica a cargo do operador.

```
log(A)

##          [,1]     [,2]     [,3]
## [1,] 0.0000000 1.386294 1.945910
## [2,] 0.6931472 1.609438 2.079442
## [3,] 1.0986123 1.791759 2.197225

B/A

##      [,1] [,2]      [,3]
## [1,]    1  0.0 0.0000000
## [2,]    0  0.2 0.0000000
## [3,]    0  0.0 0.1111111
```

Autovalor e autovetor

Considere uma matriz $\boldsymbol{A}$ quadrada *p x p*. Se um escalar e um vetor satisfazem a seguinte equação:

$$\boldsymbol{A}\vec{v} = \lambda\vec{v}$$

Dizemos que o escalar é um *autovalor* e o vetor é um *autovetor* da matriz $\boldsymbol{A}$. Partindo dessa definição, não é muito difícil chegar às seguintes conclusões:

1. os autovetores indicam as direções do espaço que não são alteradas pela transformação linear $\boldsymbol{A}$;
2. uma matriz quadrada $\boldsymbol{A}$ com dimensões *p x p* deve ter *p* autovalores (nem todos reais, nem todos diferentes) e *p* autovetores correspondentes a esses autovalores;
3. quando $\boldsymbol{A}$ é uma matriz simétrica (i.e., igual à sua própria transposta) e real, todos os seus autovalores serão reais;

No ambiente de programação do R, os autovalores e autovetores de uma matriz podem ser obtidos usando a função **`eigen`**. A seguir, calculamos os autovalores e autovetores de uma matriz de correlações.

```
C


##       x1    x2    x3    x4
## x1  1.00  0.65 -0.02 -0.04
## x2  0.65  1.00  0.00 -0.01
## x3 -0.02  0.00  1.00  0.71
## x4 -0.04 -0.01  0.71  1.00

eigen(C)


## eigen() decomposition
## $values
## [1] 1.7265007 1.6341244 0.3499571 0.2894177
##
## $vectors
##            [,1]       [,2]        [,3]        [,4]
## [1,]  0.3063289 -0.6370389  0.70422094  0.06645947
## [2,]  0.2829616 -0.6484543 -0.70487800 -0.05085992
## [3,] -0.6395629 -0.3017582  0.07161465 -0.70340074
## [4,] -0.6457976 -0.2874550 -0.04572982  0.70584971
```

No *output* anterior, os autovalores são chamados `$values` e são sempre publicados em ordem decrescente. Os autovetores correspondentes aos autovalores são publicados como colunas da matriz `$vectors`.

Decomposições espectrais

Há dois teoremas importantes que permitem fatorar matrizes, i.e., decompô-las em um produto de outras matrizes. Essa decomposição está intimamente relacionada aos autovalores e autovetores da matriz que estamos decompondo (Härdle; Simar, 2015). Apresentaremos aqui somente o primeiro teorema, chamado “teorema da decomposição de Jordan”.

Teorema da decomposição de Jordan

Cada matriz simétrica ***A*** com dimensões *p x p* pode ser escrita como:

$$\mathbf{A} = \mathbf{\Gamma \Lambda \Gamma}^{\mathrm{T}}$$

Em que Lambda é a matriz diagonal dos autovalores de ***A***:

$$\mathbf{\Lambda} = \begin{bmatrix} \lambda_1 & 0 & 0 & \cdots & 0 \\ 0 & \lambda_2 & 0 & & 0 \\ 0 & 0 & \lambda_2 & & 0 \\ \vdots & & & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \lambda_p \end{bmatrix}$$

E Gama é a matriz cujas colunas são autovetores de ***A***:

$$\mathbf{\Gamma} = [\vec{v}_1 \quad \vec{v}_2 \quad \vec{v}_3 \quad \cdots \quad \vec{v}_p]$$

Isso pode ser testado comparando uma matriz de correlações qualquer (pois todas elas são simétricas) com o produto de suas matrizes de autovetores, autovalores e autovetores:

```
C  # matriz de correlações gerada automaticamente

##       x1    x2    x3    x4
## x1  1.00  0.65 -0.02 -0.04
## x2  0.65  1.00  0.00 -0.01
## x3 -0.02  0.00  1.00  0.71
## x4 -0.04 -0.01  0.71  1.00

Lambda = diag(eigen(C)$values)
Gamma = eigen(C)$vectors
Gamma %*% Lambda %*% t(Gamma)  # deve ser igual a C

##       [,1]         [,2]          [,3]  [,4]
## [1,]  1.00  6.50000e-01 -2.000000e-02 -0.04
## [2,]  0.65  1.00000e+00  2.393918e-16 -0.01
## [3,] -0.02  1.31839e-16  1.000000e+00  0.71
## [4,] -0.04 -1.00000e-02  7.100000e-01  1.00
```

A saber, o teorema da decomposição de Jordan pode ser generalizado para fatorar matrizes retangulares. Esse segundo teorema da decomposição espectral é chamado *teorema da decomposição do valor singular* e é amplamente empregado em estatística multivariada.

Os métodos estatísticos são uma ferramenta tão importante quanto negligenciada pela pesquisa educacional brasileira. Este livro está direcionado a todos os pesquisadores em ciências humanas e sociais que desejam empregar métodos quantitativos em suas investigações. Originalmente escrito para um curso de pós-graduação em Educação em Ciências, este livro introduz noções de estatística básica, intermediária e avançada, sem presumir que o leitor domine matemática superior. Além disso, ele ensina como programar na **linguagem R**, a mais poderosa ferramenta de análise computacional da atualidade. Todos os tópicos desta obra estão devidamente exemplificados por análises concretas, que poderão ser experimentadas pelo leitor em seu processo de aprendizagem. Enfim, trata-se de uma referência importante para todos os pesquisadores profissionalmente interessados na análise estatística em ciências humanas e sociais.

Para acessar os materiais complementares a este livro, acesse:
www.perspectivascriticas.com.br